Breaking out of the cycle: Including quiescence in cell cycle classification

Samantha A. O'Connor[1], Leonor Garcia [2], Anoop P. Patel [3,4], Jean-Philippe Hugnot [2], Patrick J. Paddison [5], Christopher L. Plaisier [1,*]

[1] School of Biological and Health Systems Engineering, Arizona State University, Tempe AZ, USA.
[2] Institut de Génomique Fonctionnelle, Université de Montpellier, CNRS, INSERM, 141 rue de la Cardonille, 34091, Montpellier, France.
[3] Brotman-Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA.
[4] Department of Neurosurgery, Preston Robert Tisch Brain Tumor Center, Duke University, Durham, NC, USA.
[5] Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle WA, USA
* To whom correspondence should be addressed: Christopher Plaisier, E-mail: plaisier@asu.edu, Phone: (480) 965-6832, Address: P.O. Box 879709, Tempe, AZ 87287-9709

Abstract

Single-cell transcriptomics has unveiled a vast landscape of cellular heterogeneity in which the cell cycle is a significant component. We trained a high-resolution cell cycle classifier (ccAFv2) using single cell RNA-seq (scRNA-seq) characterized human neural stem cells. The ccAFv2 classifies six cell cycle states (G1, Late G1, S, S/G2, G2/M, and M/Early G1) and a quiescent-like G0 state, and it incorporates a tunable parameter to filter out less certain classifications. The ccAFv2 classifier performed better than or equivalent to other state-of-the-art methods even while classifying more cell cycle states, including G0. We showcased the versatility of ccAFv2 by successfully applying it to classify cells, nuclei, and spatial transcriptomics data in humans and mice, using various normalization methods and gene identifiers. We provide methods to regress the cell cycle expression patterns out of single cell or nuclei data to uncover underlying biological signals. The classifier can be used either as an R package integrated with Seurat (https://github.com/plaisier-lab/ccafv2_R) or a PyPI package integrated with scanpy (https://pypi.org/project/ccAFv2/). We proved that ccAFv2 has enhanced accuracy, flexibility, and adaptability across various experimental conditions, establishing ccAFv2 as a powerful tool for dissecting complex biological systems, unraveling cellular heterogeneity, and deciphering the molecular mechanisms by which proliferation and quiescence affect cellular processes.

Introduction

Single-cell RNA sequencing (scRNA-seq) is a robust method for dissecting the transcriptional states of individual cells obtained from specific conditions. These cellular transcriptional states are influenced by various biological signals, including cell type and the phase of the cell cycle. The cell cycle is a tightly regulated and intricately coordinated biological process that orchestrates the division of a cell into two daughter cells. Adult stem cell populations often reside in a quiescent G0 state outside of the cell cycle, reactivating only upon receiving appropriate signals to divide (Doetsch 2003; Obernier et al. 2018). Current state of the art methods to predict cell cycle states based on scRNA-seq transcriptome profiles lump G0 cells with G1 cells (Hao et al. 2021; Zheng et al. 2022; Schwabe et al. 2020; Liu et al. 2017; Hsiao et al. 2020; Scialdone et al. 2015). The grouping of G0 with G1 fails to recognize the clear differences in expression patterns and quiescent phenotype displayed by G0 cells, making them readily distinguishable from G1 cells (O'Connor et al. 2021). Previously, we used scRNA-seq of U5 human Neural Stem Cells (U5-hNSCs; Davis and Temple 1994; Johe et al. 1996) grown *in vitro* to discern seven cell cycle states including a quiescent-like G0 state (O'Connor et al. 2021). An Artificial Neural Network (ANN) (Ma and Pellegrini 2020) classifier named the cell cycle ASU/Fred Hutch (ccAF) was trained to predict these seven cell cycle states in cells from new datasets (O'Connor et al. 2021). In those studies, the ccAF classifier was applied to a host of neuroepithelial derived cells characterized by scRNA-seq, including glioblastoma patient tumor cells. The underlying software packages for constructing ANNs (TensorFlow and Keras) have been significantly improved and we hypothesized that reimplementation of the ccAF classifier would significantly improve classifier performance and provide likelihoods for each classification, a feature not available in the original ccAF implementation.

In addition to the advancements in ANN methodology, numerous new scRNA-seq studies have been conducted on neuroepithelial derived cell types. Especially useful for assessing the quality of the *in vitro* hNSC trained classifier are the 94,297 *in vivo* hNSCs (Zeng et al. 2023) from fetal tissue 3 to 12 weeks post conception. An atlas of the developing human spinal cord (Zhang et al. 2021) will be used to determine whether the classifier can be applied to both single cell and single nuclei RNA-seq (scRNA-seq and snRNA-seq). An atlas of adult neurogenesis in the ventricular-subventricular zone (V-SVZ) (Cebrian-Silla et al. 2021) will be used to demonstrate that the classifier can be applied to mouse cells. Additionally, this allows comparisons to be made between the cell cycle proportions of cell types from adult mouse neurogenesis in the V-SVZ (Cebrian-Silla et al. 2021) and the developing human telencephalon (Nowakowski et al. 2017). Two studies of quiescent neural stem cells will be crucial for demonstrating the identity of the G0 cell state (Llorens-Bobadilla et al. 2015; Dulken et al. 2017). Additionally, we collected scRNA-seq for two IDH mutant low-grade glioma (LGG) cell lines in conditions with and without growth factors. This will allow us to gain insights into the performance of the classifier when confronted with a higher proportion of non-cycling cells. We will also apply the classifier to *in vivo* glioblastoma tumor cells and *in vitro* glioblastoma tumor derived cancer stem cells that were not included in the previous ccAF classifier studies (Couturier et al. 2020a). Finally, we will apply the classifier to a spatial-transcriptomics (ST-seq) study of eight tissue slices from a human fetus four weeks post-conception to demonstrate that the classifier can provide information about proliferation spatially across a developing organism (Zeng et al. 2023). These neuroepithelial derived datasets offer a robust foundation for rigorously testing and validating the improved ccAF version 2 (ccAFv2) classifier, showcasing its versatility across species, single cells and nuclei, and a diverse array of neuroepithelial cell types.

We implement the ccAFv2 cell cycle classifier ANN in these studies using improved machine learning software packages. We also provide an R- and Python-based cell cycle classifier that

seamlessly integrates into the Seurat or Scanpy analysis workflows. Refining the underlying model and utilizing training data processed with modern quality control and normalization techniques improves prediction accuracy. Additionally, incorporating a likelihood threshold parameter gives users the flexibility to fine-tune classification according to the specific characteristics of their data. We compare ccAFv2 to existing state-of-the-art cell cycle classifiers and show it delivers improved performance while classifying more cell cycle states, including being the only classifier for G0 states. We illustrate that the cell cycle states identified by ccAFv2 can be used to remove cell cycle expression from single cell and nuclei transcriptomic data to uncover underlying biological signals, mirroring the methodology previously employed in ccSeurat (Luecken and Theis 2019). We apply ccAFv2 to various human and mouse neurodevelopment datasets, neural and glioma stem cell populations, and gliomas to show its generalizability across neuroepithelial-derived cell types. Finally, we show that ccAFv2 can lead to novel insights about the spatial organization of proliferation and quiescence.

Results

Implementation of neural network classifier for ccAFv2

We implemented the core algorithm of ccAFv2 to take advantage of significant improvements in machine learning tools that should improve classifier performance and provide likelihoods for each predicted cell cycle classification. The ccAFv2 core algorithm is broken up into two steps. First, the input data is run through the artificial neural network (ANN) to compute likelihoods for each class (i.e., Neural G0, G1, Late G1, S, S/G2, G2/M or M/Early G1; **Figure 1A**-**B**). The underlying ANN for ccAFv2 starts with a dense input layer connected to two hidden layers that connect to a softmax output layer (**Figure 1A**). Overfitting in the ANN is mitigated by dropout regularization via two dropout layers. The first dropout layer is positioned between the first and second hidden layers and the second dropout layer is between the second hidden layer and the softmax output (**Figure 1A**, Xie et al. 2019). Second, the likelihoods calculated by the ANN for each cell cycle state are used to determine which state should be assigned for each cell (**Figure 1B**). The cell cycle state with the maximum likelihood is identified and if the likelihood is greater than or equal to the likelihood threshold then the state is returned. Otherwise, if the maximum likelihood is less than the likelihood threshold a state of "Unknown" is returned. These improvements to the core ANN of ccAFv2 will be rigorously tested in the subsequent sections.

Training the ccAFv2 classifier

The training data for ccAFv2 is comprised of scRNA-seq from actively dividing U5 human neural stem cells (U5-hNSCs) cultured *in vitro* (O'Connor et al. 2021). The U5-hNSCs were cultivated from the telencephalon of a human fetus 8 weeks post-conception (Bressan et al. 2017). We previously identified 7 transcriptional states in the U5-hNSCs that were mapped to cell cycle states (i.e., Neural G0, G1, Late G1, S, S/G2, G2/M, and M/Early G1; O'Connor et al. 2021). The U5-hNSC scRNA-seq data were reanalyzed using current quality control and normalization methods which resulted in 2,962 good quality single-cell transcriptome profiles. The U5-hNSC scRNA-seq profiles, along with the previously established cell cycle labels (O'Connor et al. 2021), were the training data used to develop the new ccAFv2 cell cycle classifier.

We compared the newly implemented ccAFv2 classifier against four distinct classification methods: support vector machine with rejection (SVMrej), random forest (RF), scRNA-seq optimized K-nearest neighbor (KNN), and ACTINN (Ma and Pellegrini 2020) which was used to build ccAF (O'Connor et al. 2021). The training dataset for all classifiers consisted of the pre-processed U5-hNSC scRNA-seq subset to the 861 genes up-regulated in cell cycle states (log2FC $\geq$ 0.25, adjusted p-value $\leq$ 0.05; **Supplementary Table S1**). We applied 10-fold cross-validation (CV) for each classification method (**Figure 1C**) and observed that ccAFv2 exhibited significantly improved F1 scores for each cell cycle state compared to other classification methods (p-values $\leq 2.8 \times 10^{-6}$; **Figure 1D**), establishing it as the most accurate cell cycle classifier overall. The accuracy of ccAFv2 when applied to U5-hNSCs was 88.4%, and the main difference when compared to ccAF was an improvement in Late G1 cell predictions (**Supplementary Figure S1A**-**B**). The overall error rate for ccAFv2 was 3.3%, which is a considerable improvement from the 18.4% of ccAF (O'Connor et al. 2021). The reimplementation of the ANN for the ccAFv2 classifier has significantly improved its performance across all cell cycle states, providing a robust foundation for further optimization and comprehensive characterization of its capabilities.

Optimizing the number of neurons in hidden layers

A crucial factor in optimizing the parameters of the ccAFv2 ANN was determining the ideal number of neurons in each hidden layer. We conducted a systematic comparison of 18 different combinations for the number of neurons in the two hidden layers (first hidden layer: ranging from 200 to 700 neurons, and second hidden layer: ranging from 100 to 400 neurons). The

optimal combination was determined by having the highest average F1-score and Adjusted Mutual Information (AMI) score using ccSeurat as the reference (**Figure 1E**; **Supplementary Table S2**). We found that configuring the ccAFv2 ANN with 600 neurons in the first hidden layer and 200 in the second hidden layer yielded the largest average F1 score and second largest AMI score (**Figure 1E**). This specific parameterization has been assigned for the hidden layers of the ccAFv2 ANN, and all prior and subsequent ccAFv2 classifications use this parameterization.

Comparison with existing cell cycle classifiers

An important means to test the performance of ccAFv2 is to compare it to existing state-of-the-art methods for cell cycle state classification. We evaluated the following methods: ccAF (O'Connor et al. 2021), ccSeurat (Hao et al. 2021), tricycle (Zheng et al. 2022), Revelio/SchwabeCC (Schwabe et al. 2020; Zheng et al. 2022), reCAT (Liu et al. 2017), peco (Hsiao et al. 2020), and cyclone (Scialdone et al. 2015). Each tool predicts a different subset of cell cycle phases, uses a different classification algorithm, was trained on different data, and requires different input genes and data formats (**Supplementary Table S3**). First, we made comparisons using AMI scores computed by comparing the predicted labels for U5-hNSCs from each classifier to the reference cell labels from O'Connor et al., 2021 (**Figure 2A-B**). The AMI score is impacted by the number of cell cycle states in the reference, seven cell cycle states from O'Connor et al., 2021, and the number of states predicted by each algorithm. We used simulation studies to define the expected range of AMI scores that correspond to specific levels of similarity to the reference given the number of cell cycle states in the reference and the classifier. The ccAFv2 classifier surpassed all other classifiers in AMI score (**Figure 2A**) and had slightly greater than 90% similarity to the reference. The next best classifiers were ccSeurat and ccAF which had AMI scores that had less than 90% but greater than 80% similarity to the reference (**Figure 2A**). A limitation of making comparisons using the U5-hNSCs is that the ccAFv2 was trained using these data, and thus it was vitally important to extend comparisons to an independent dataset.

Next, we applied ccAFv2 alongside the other state-of-the-art cell cycle classification methods to cells derived from a glioblastoma patient tumor (BT322; Couturier et al. 2020b) (**Figure 2C-D**). These cells were not used in the training of ccAFv2, thereby providing an independent dataset to unbiasedly compare the efficacy of the cell cycle prediction algorithms. We chose the ccSeurat labels as the reference because true cell cycle state labels do not exist for this dataset and because ccSeurat performed well in the U5-hNSC comparisons (**Figure 2A**). The largest AMIs were observed for tricycle, reCAT, and ccAFv2; all of which correspond to just below 90% similarity to the reference (**Figure 2C**). These results demonstrate that ccAFv2 delivers at least equivalent performance when compared to contemporary state-of-the-art cell cycle classifiers, while providing the highest resolution of cell cycle state predictions including a quiescent-like G0 state.

Defining an appropriate classification likelihood threshold

The improved ccAFv2 classifier calculates likelihoods for each cell cycle state which can be used to determine the most likely state and to assess the quality of the classification. We hypothesized that applying a likelihood threshold to ccAFv2 classifications would ensure reliability and confidence in predicted cell cycle states by setting classifications for cells with less certainty to an "Unknown" state. We explored the range of possible likelihood thresholds on 94,297 hNSCs collected from human fetal tissue at 3 to 12 weeks post-conception (Zeng et al. 2023); **Figure 3**). Application of ccAFv2 to the *in vivo* fetal hNSCs was found to differ by week stage (**Figure 3B**). The amount of Neural G0 cells from the *in vitro* U5-hNSCs, derived from fetal brain tissue at 8 weeks post-conception, (**Figure 3A**) matches closely to the *in vivo* hNSCs at eight weeks post-conception (**Figure 3B**). Moreover, the expression patterns of cyclins

between the *in vitro* and *in vivo* hNSCs eight weeks post-conception were strikingly similar (**Figure 3C-D**). In both hNSC populations, CCNE2 exhibited its peak expression during the S phase, while CCNA2 showed highest expression levels during the S/G2 and G2/M phases, and CCNB1 displayed elevated expression in G2/M phase cells. Notably, the lowest expression of the key regulator of cell cycle progression, CCND1, was observed in the Neural G0 state. These similarities suggests that the *in vivo* hNSCs will be useful in validating the efficacy of the ccAFv2 classifier and for choosing a default likelihood threshold.

Next, we tested ccAFv2 likelihood thresholds ranging from 0.0 to 0.9 in increments of 0.1 (**Supplementary Figure S2**). The calculated cell cycle state likelihood was required to be greater than or equal to the threshold, otherwise an "Unknown" state was returned (**Figure 1B)**. Each likelihood cutoff was assessed using the percentage of cells predicted and an AMI score with ccSeurat cell cycle states as a reference. As the likelihood threshold increases the number of cells predicted decreases and the AMI scores increase (**Figure 3E**, **Supplementary Figure S2**, **Supplementary Table S4**). In other words, the removal of less certain classifications improves the accuracy of the overall classifications (**Figure 3E**). Next, we further demonstrated that the increase in AMI resulted from the specific removal of cells which had low classification likelihoods, by comparing it to the random removal of an equivalent number of cells (representative analysis for 3 weeks post-conception is shown in **Figure 3F**). The randomly removed cells do not increase the AMI (**Figure 3F**), only the selected removal of cells with low likelihoods were able to increase the AMI. We found that the median AMI scores calculated with likelihood thresholds of 0.4 to 0.9 were significantly higher than the median AMI scores of the randomly removed cells (**Figure 3F, Supplementary Table S4**), which indicates that the likelihood cutoffs of greater than or equal to 0.4 improve classification accuracy. We selected the likelihood threshold of greater than or equal to 0.5 because it signifies a minimum of 50% certainty in the classified cell cycle state. Additionally, greater than 90% of *in vivo* hNSCs could be assigned a cell cycle state with a likelihood threshold of 0.5 (**Figure 3E**). Thus, the threshold of 0.5 was set as the default for ccAFv2 and used in subsequent analyses, except where noted. We also provide users with the flexibility to adjust the likelihood threshold parameter in ccAFv2, allowing them to adapt the classifier's operation to suit the unique characteristics of their dataset.

Effect of missing gene expression values on ccAFv2

A known limitation of scRNA-seq is that dropouts are common. A dropout occurs when lowly to moderately expressed transcripts are detected in one cell but are not detected in another cell of the same cell type (Qiu 2020). The ccAFv2 classifier uses the expression of 861 genes to predict cell cycle states. We hypothesized that dropouts could be simulated by randomly removing defined percentages of genes and that this would provide a reasonable approximation of the influence of missing genes on the accuracy of ccAFv2's cell cycle state classifications. We evaluated the consequences of these simulated gene dropouts on the classifier error rate, AMI, and the number of cells predicted (**Supplementary Figure S3-4, Supplementary Table S5**). As described earlier, the median error rate of applying ccAFv2 to U5-hNSCs was 3.3% with 99% of the input genes (99% is used to allow for cross-validation). Removing 20% of the ccAFv2 input genes yielded a smaller median error rate (12.2%) than the original ccAF error rate with all the input genes (18.4%), underscoring the improved performance of the new model. Removing 40% of ccAFv2 input genes led to a 29% median error rate, and 96% of cells were predicted (**Supplementary Table S5**). The error rate was the most affected by the loss of input genes (**Supplementary Figure S3A)** and the median percentage of cells predicted remained above 80% even when 70% of the input gene list was removed (**Supplementary Table S5**), suggesting that increasing ccAFv2's likelihood threshold (>0.5) might be required to ensure the quality of predictions for datasets with greater than 20% missing ccAFv2 input genes. Indeed, the error rate for removing 20% of ccAFv2 input genes decreased from 12.2% median error rate

at 0.5 likelihood threshold to 9.9% with a 0.7 likelihood threshold (**Supplementary Figure S3B, Supplementary Table S5**) and 6.2% with a 0.9 likelihood threshold (**Supplementary Figure S3C, Supplementary Table S5**). Thus, losing 20% of input genes led to four times the error rate, and the increased error rate can be mitigated in part by increasing the likelihood threshold.

Capturing the effect of growth factors on cellular proliferation

Growth factors are used to increase cellular proliferation *in vitro*, and we characterized the transcriptomes of LGG cells (grade 2 astrocytoma and grade 3 oligodendroglioma) with and without the application of growth factors (**Figure 4**). For this analysis we tested the impact of adjusting the ccAFv2 likelihood threshold across a range of values 0 to 0.9 (**Supplementary Figure S5A**-**B**). Increasing the likelihood threshold values from 0.4 to 0.9 led to an increased proportion of "Unknown" classifications in the samples without growth factors, which is consistent with the known effect of growth factors to stimulate proliferation and the cell cycle. The increased proportion of "Unknown" cells may correspond to new growth factor starvation state(s) not included in ccAFv2 classification states. Additionally, the S, S/G2, and G2/M cell cycle states were disproportionately removed as the likelihood threshold increased (**Supplementary Figure S5A**-**B**). We then set the likelihood threshold to 0.9 and observed that the cells grown with growth factors form clusters of cell cycle state labels, outlining the expected progression of cell cycle phases (G1 → S → S/G2 → G2/M → M/Early G1; **Figure 4A**, **C**, **D**, & **F**). Conversely, cells grown without growth factors exhibit a more dispersed distribution of cell cycle state labels (**Figure 4B**, **C**, **E**, & **F**). The ability to change the likelihood threshold of ccAFv2 allows us to observe the biological impact of adding growth factors to LGG cells and demonstrates what to expect when the cell cycle is not the main transcriptional signal in cells.

Removing cell cycle expression signatures

The cell cycle generates a strong transcriptional signature that can obscure other less robust transcriptional signatures of interest. Previous studies have shown that statistical methods can effectively remove cell cycle transcriptional signatures and that the residual transcriptional variance can be used to study less robust transcriptional signatures of interest (Luecken and Theis 2019). We showcase successful removal of the cell cycle transcriptional signatures for the U5-hNSCs, and LGG cells. First, each cell cycle state's average expression of marker genes is computed for every single cell or nuclei. Then, these average cell cycle expression patterns are regressed out of the dataset during normalization. The ccSeurat regression method uses only the S and G2/M cell cycle states, so we first tested regression with the S and G2/M cell cycle states from ccAFv2. We found that the ccAFv2 marker gene derived average cell cycle expression patterns could mitigate cell cycle transcriptional signatures as effectively as ccSeurat (empirical p-value > 0.05, **Supplementary Table S6**, **Supplementary Figure S6**). Additionally, we found that incorporating Late G1, S, S/G2, G2M, and M/Early G1 was also quite effective and led to a more robust homogenization of the cell cycle states based on PCA plots (**Supplementary Table S6**, **Supplementary Figure S6**). This approach enables researchers to dissect complex gene expression patterns and uncover novel insights into cellular processes beyond the cell cycle.

Classifying neuroepithelial-derived cells in humans and mice

In a previous study (O'Connor et al. 2021) we applied the ccAF classifier to cells from the developing human telencephalon (Nowakowski et al. 2017). We applied ccAFv2 to these same cells and compared the ccAF and ccAFv2 predicted cell cycle proportions. We observed that the Neural G0 state was less frequent in all cell types for ccAFv2 relative to ccAF (**Figure 5B**; **Supplementary Table S7**). The Neural G0 state was distinctly less frequent in the neuronal cell types. For ccAF Neural G0 made up most of the cell cycle states for EN-PFC and EN-V1, but in ccAFv2 these two cell types classified primarily as G1 (**Figure 5B**). The glial cell types had the

largest Neural G0 subpopulations (**Figure 5B**). These differences for the ccAFv2 cell cycle classifications suggest Neural G0 might be capturing a glial rather than neuronal G0-like cell state.

We also applied ccAFv2 to cells from the ventricular-subventricular zone (V-SVZ) of the adult mouse brain (Cebrian-Silla et al. 2021), a location known to contain neural stem and precursor cells in the adult brain (Lim and Alvarez-Buylla 2016). The adult mouse V-SVZ validates observations from the developing human telencephalon (**Figure 5E**). In the V-SVZ the glial cells tended to have larger Neural G0 subpopulations, neuronal cell types tended to have less Neural G0, and microglial had the smallest amount Neural G0 (**Figure 5E**). The results are remarkably similar given the differences between species, developmental state, and anatomical origins. These findings illustrate that the ccAFv2 classifier can be applied to cells originating from both humans and mice.

Classifying quiescent-like neural stem cells

Previously we validated the Neural G0 state using two independent *in vivo* scRNA-seq profiling studies of NSCs from adult neurogenesis in the subventricular zone (SVZ) that used fluorescence activated cell sorting (FACS) to sort out quiescent and activated NSCs (Llorens-Bobadilla et al. 2015; Dulken et al. 2017). We applied ccAFv2 to these same cells and compared the ccAF and ccAFv2 predicted Neural G0 subpopulations. The proportion of cells classified as Neural G0 decreased for ccAFv2 in the quiescent NSCs (qNSCs) and was replaced by more G1, S/G2, and a small amount of G2/M (**Figure 5D, Supplementary Figure S7**). For Llorens-Bobadilla et al., 2015 the active NSCs 1 (aNSC1) were more highly enriched with S phase cells, and aNSC2 were enriched with S/G2 and G2/M. A similar trend was observed for the Dulken et al., 2017 dataset. Overall, the qNSCs are enriched with quiescent-like Neural G0 cells, and the aNSCs are at some stage of the cell cycle (**Figure 5D**). These results continue to validate our assertion that Neural G0 represents a quiescent-like cell state, and that ccAFv2 can accurately classify this quiescent-like G0 state.

Accurate classification of cells and nuclei

Tissues in single-cell studies can be processed into cells for scRNA-seq or nuclei for snRNA-seq (**Figure 5A**). Both methods are commonly used and have advantages and limitations (Slyper et al. 2020). Thus, it is important to demonstrate whether ccAFv2, which is trained on cells, can accurately classify cell cycle states for single nuclei. We employed the Zhang et al., 2021 dataset which characterized developing human spinal cord tissue from five developmental time points using both scRNA-seq and snRNA-seq from the same experimental conditions (Zhang et al. 2021). The proportions of cells in each cell cycle state are remarkably similar between scRNA-seq and snRNA-seq from the same condition (**Figure 5C**), illustrating the versatility of the ccAFv2 classifier in effectively analyzing both scRNA-seq and snRNA-seq profiles.

Mapping proliferation onto tissue through spatial transcriptomics

ScRNA-seq and snRNA-seq provide valuable information about the transcriptional states of cells and nuclei, but they do not provide information about spatial localization of these states within tissues. Spatial transcriptomics captures the transcriptional activity of a single-cell or a region containing a small number of cells at a position within an intact tissue. We applied ccAFv2 to spatial transcriptomics of eight slices from a human fetus four weeks post-conception (Zeng et al. 2023). Then, we compared the proportions of predicted cell cycle states from scRNA-seq of all cells from a human fetus four weeks post-conception to the proportions from the spatial transcriptomics slices (**Figure 6A**). The amount of Neural G0 and G1 are similar, but there are more S phase cells in scRNA-seq and more Late G1 in spatial transcriptomics (**Figure 6A**). The proportions of cell cycle states are consistent across the spatial transcriptomics slices

(**Figure 6A**). Additionally, the spatial distribution of cell cycle states can be visualized by superimposing them onto the H&E stained image (**Figure 6B-C**). Alternatively, the likelihoods for individual cell cycle states are more informative due to the multiple cells underlying each spot on the 10X Visium spatial transcriptomics platform. Thus we have also overlaid the likelihoods onto the H&E stained fetal tissue (**Figure 6B-J**). Specifically, we observe that Neural G0 is more centrally located (**Figure 6D**), and G2/M is resolved to a region behind the head where the neural tube is developing (**Figure 6I**). These results demonstrate that ccAFv2 can be applied to spatial transcriptomics, and that it can lead to novel insights about the spatial organization of proliferation.

Discussion

We designed ccAFv2 to use transcriptomic data to accurately classify cell cycle states and a quiescent-like G0 state for single cells or nuclei. The performance of the updated classifier was superior to its predecessor and demonstrated comparable or better performance than other state-of-the-art cell cycle classifiers. The ccAFv2 classifies cells into a broader range of cell cycle states than the contemporary state-of-the-art cell cycle classifiers (Hao et al. 2021; Zheng et al. 2022; Schwabe et al. 2020; Liu et al. 2017; Hsiao et al. 2020; Scialdone et al. 2015), and it is the only classifier to include a G0 state. Moreover, ccAFv2 features a tunable parameter to filter out less certain classifications. We showcased the versatility of ccAFv2 by successfully applying it to classify cells, nuclei, and spatial transcriptomics data in humans and mice, using various normalization methods and gene identifiers. The classifier can be used either as an R package integrated with Seurat (https://github.com/plaisier-lab/ccafv2_R) or a PyPI package integrated with scanpy (https://pypi.org/project/ccAF/). We proved that ccAFv2 has enhanced accuracy, flexibility, and adaptability across various experimental conditions, establishing ccAFv2 as a powerful tool for exploring cell cycle dynamics in diverse biological contexts.

The ccAFv2 classifier will be most helpful in biological contexts where the cell cycle is active. We utilized atlases of developing human and mouse brain and spinal cord because proliferation is essential in developing organisms (Soufi and Dalton 2016; Pauklin and Vallier 2013). Evidence is building to show that cell fate decisions are tightly coupled to cell cycle events and machinery (Pauklin and Vallier 2013). In healthy adult organisms, proliferation plays critical roles in several processes: maintenance of stem cell populations (Harada et al. 2021), clonal expansion of both innate and adaptive immune cells (Adams et al. 2020), and germ cell meiosis, encompassing oogenesis (Bukovsky et al. 2005) and spermatogenesis (Guo et al. 2018). Defects in cell cycle machinery or regulation can lead to runaway proliferation characteristic of cancer (Hanahan and Weinberg 2000), or the lack of proliferation of crucial cell types can lead to neurodegenerative disorders (Joseph et al. 2020). Cell cycle classification would benefit any *in vitro*, *in vivo*, or *ex vivo* studies of proliferating cells. On the other hand, we provide methods to regress the cell cycle expression patterns out of single cell or nuclei data to uncover underlying biological signals. Overall, incorporating cell cycle states into single-cell and nuclei studies enhances our ability to dissect complex biological systems, unravel cellular heterogeneity, and decipher the molecular mechanisms by which proliferation affects cellular processes.

The studies reported here demonstrate that ccAFv2 performs well for neuroepithelial-derived and glioma cell types. The expression signatures for the primary cell cycle states should translate to other cell types, especially Late G1, S, S/G2, G2/M, and M/Early G1. However, it is unknown whether Neural G0 will translate beyond neuroepithelial-derived cells. In prior work, we showed that Neural G0 was an active transcriptional signature executed by a subpopulation of U5-hNSCs (O'Connor et al. 2021). We also discovered fewer Neural G0 cells in the mesenchymal subpopulations of cells from glioblastoma patient tumors (O'Connor et al. 2021), which could either mean that the Neural G0 does not capture the mesenchymal G0 or that there are fewer G0 cells in the mesenchymal subtype. At this point, it is still unclear which hypothesis is correct, and future studies are needed to resolve this question. Additional support for novel G0 states was observed in the growth factor deprived LGG cells, where an increased proportion of "Unknown" cells was detected, hinting at novel quiescent-like state(s) missing from ccAFv2. Other studies have identified multiple G0 states in a single cell type that are invoked in response to different stimuli (e.g., spontaneous loss of mitogenic factors, serum starvation, drug treatment, etc.) (Stallaert et al. 2022). Thus, we find it very likely that additional G0 states with distinct transcriptional signatures will be identified. The ccAFv2 ANN and its associated training

software are fully equipped to integrate these additional G0 states. Future studies that extend the cell cycle classifier to include novel G0 states holds immense potential for advancing our understanding of quiescence in biological systems. By leveraging advanced computational methods, high-throughput technologies, and interdisciplinary approaches, researchers can unravel the complexities of cellular dormancy and pave the way for innovative strategies to manipulate quiescent cell behavior to improve health and combat disease.

Methods

Cell culture

For the “no growth factors” condition, cells from LGG glioma neurospheres (LGG275, BT237) were dissociated, seeded into two poly-D-lysine and laminin coated T25 cm^2 flasks at a density of 40,000 cells/cm^2, and cultured for 4 days using medium without growth factors. For the “with growth factors” condition, cells were cultured for 4 days as neurospheres with EGF and FGF2 at 10µg/L and heparin at 2mg/L in PolyHEMA coated flasks, and medium was replaced 1 day before single cell sequencing.

scRNA-seq library preparation and sequencing

Cells were dissociated and single cell suspensions loaded onto the Chromium controller (10x Genomics, Pleasanton, CA) to generate single-cell Gel Beads-in-Emulsion (GEMs). The single-cell RNA-Seq libraries were prepared using the Chromium Next GEM Single Cell 3’ Reagent Kits V3.1 (Dual Index, P/N 1000268, 10x Genomics). Briefly, reverse transcription was performed at 53°C for 45 min followed by incubation at 85°C for 5 min. GEMs were then broken and the single-stranded cDNAs were cleaned up with DynaBeads MyOne Silane Beads (Thermo Fisher Scientific; P/N 37002D). The cDNAs were PCR amplified, cleaned up with SPRIselect beads (SPRI P/N B23318), fragmented, end-repaired, A-tailed, and size-selected with SPRIselect beads. Indexed adapters were ligated and cleaned up with SPRIselect beads. The resulting DNA fragments were PCR amplified and size selected with SPRIselect beads. The size distribution of the resulting libraries was monitored using a Fragment Analyzer (Agilent Technologies, Santa Clara, CA, USA) and the libraries were quantified using the KAPA Library quantification kit (Roche, Basel, Switzerland). The libraries were denatured with NaOH, neutralized with Tris-HCl, and diluted to 150 pM. Clustering and sequencing were performed on a NovaSeq 6000 (Illumina, San Diego, CA, USA) using the paired-end 28-90 nt protocol on one lane of an SP flow cell and on one lane of an S4 flow cell. Sequencing data can be accessed from NCBI SRA. Both library preparation and sequencing were performed at the Montpellier GenomiX facility (MGX) in Montpellier, France.

Data analysis

Image analyses and base calling were performed using the NovaSeq Control Software and the Real-Time Analysis component (Illumina). Demultiplexing was performed using the 10x Genomics software Cellranger mkfastq (v7.1.0), a wrapper of Illumina's bcl2fastq (v2.20). The quality of the raw data was assessed using FastQC (v0.11.9) from the Babraham Institute and the Illumina software SAV (Sequencing Analysis Viewer). FastqScreen (v0.15.1) was used to identify potential contamination. Alignment, gene expression quantification and statistical analysis were performed using Cell Ranger count with the human's transcriptome (GRCh38). To discard ambient RNA falsely identified as cells, Cell Ranger count was run a second time with the option --force-cells to force the number of cells to detect. Cell Ranger aggr was used to combine each sample results into one single analysis. Cell Ranger output files can be accessed from NCBI GEO at GSE263796.

scRNA-seq, snRNA-seq, and ST-seq neuroepithelial datasets

In total 42 scRNA-seq, 11 snRNA-seq, and 8 ST-seq datasets were processed and employed in the studies used to characterize the ccAFv2 classifier. Detailed descriptions of the source, quality control, processing, normalization of each dataset can be found in the **Supplementary Material**.

Implementation of the ccAFv2 ANN model

The core algorithm of the ccAFv2 is a fully connected artificial neural network (ANN) implemented using the Keras API (v2.12.0) that employs TensorFlow (v2.12.0) to construct ANNs. A fully connected ANN model was developed to classify the cell cycle state of single cells (**Figure 1A**). The input for a single cell is expression for the 861 most highly variable genes ($\log_2$(FC) > 0.25; p value adj < 0.05) from O'Connor et al., 2021. A dense input layer takes in the expression of the 861 and is fully connected to the first hidden layer comprised of 600 neurons. The first hidden layer is fully connected to the second hidden layer comprised of 200 neurons which then connects to the output layer of seven neurons (one for each cell cycle class: Neural G0, G1, Late G1, S, S/G2, G2/M, and M/Early G1). A SoftMax regression function in the output layer is used to compute the likelihood for each class. Overfitting in the ANN is prevented through the incorporation of two dropout layers using a dropout rate of 50%. The first dropout layer is positioned between the first and second hidden layers and the second dropout layer is between the second hidden layer and the output layer (**Figure 1A**, Xie et al. 2019). Neuron activation functions were modeled using the Rectified Linear Unit (ReLU) function. The loss function for the ccAFv2 ANN was categorial cross-entropy and Stochastic Gradient Descent (SGD) was used to optimize the learning. The predicted class for a single cell is identified as the highest likelihood exceeding the specified threshold (**Figure 1B**). By default, the threshold is set at 0.5 and can be adjusted within the range of 0 to 1. If a cell's likelihood falls below the threshold it is classified as "Unknown."

Training the ccAFv2 ANN classifier

The ccAFv2 ANN model was trained on the 2,692 cells and 861 genes from the U5-hNSCs dataset (O'Connor et al. 2021) using the labels from O'Connor et al., 2021. The training process encompassed ten epochs repeated five times consecutively. In each epoch, the training data was randomly partitioned into 80% for training and 20% for testing, with the testing subset held out to assess training accuracy.

Comparing ccAFv2 to other classification methods

Classifiers were trained using the scRNA-seq gene expression of 2,962 cells with 861 genes and cell cycle labels from the U5-hNSCs. The ccAFv2 classifier was tested against: (i) support vector machine with reject option (SVMrej; classification cutoff ≥ 0.7), a general-purpose classifier from the Scikit-learn library; (ii) random forest (RF), another general-purpose classifier from the Scikit-learn library; (iii) k-nearest neighbor (KNN) from the scanpy ingest method (Wolf et al., 2018); and (iv) neural network (NN) ACTINN (Ma & Pellegrini, 2020). Classifier performance was determined using F1 scores computed for each cell cycle state. Ten-fold cross-validation with an 80% training and 20% testing split was used to determine the variance of F1 scores for each cell cycle state from each classifier. A Student's t-test was used determine if the mean of the F1 scores were significantly lower than ccAFv2.

Optimizing the number of neurons in hidden layers

The configuration of neurons in the two hidden layers is designed to reduce the number of neurons at each layer from the 861 input genes down to the 7 cell cycle states. In total, 18 ccAFv2 models were trained using the U5-hNSCs dataset to determine the optimal number of neurons for these hidden layers. This involved testing at increments of 100 the number of neurons in the first hidden layer within the range of 200 to 700 neurons and in the second hidden layer within the range of 100 to 400 neurons. For comparisons the F1 scores were computed for each cell cycle state. Each model was also tested on pre-processed scRNA-seq data of glioma stem cells (BT322, BT324, BT326, BT333, BT363, BT368) and tumor cells (BT363, BT368) from Couturier et al. 2022, along with Grade 2 Astrocytoma (LGG275) (Augustus et al. 2021). For these datasets, Adjusted Mutual Information (AMI) scores, with the reference labels derived from ccSeurat calls, and the number of cells predicted were calculated using the AMI function from the aricode package in R. Barcodes with an “Unknown” ccAFv2 label were removed before metrics were calculated.

Comparing ccAFv2 to existing cell cycle classifiers

The performance of ccAFv2 was compared with existing cell cycle state classifiers: ccAF (v1) (O’Connor et al. 2021), Seurat (Hao et al. 2021), Tricycle (Zheng et al. 2022), SchwabeCC (Schwabe et al. 2020; Zheng et al. 2022), reCAT (Liu et al. 2017), Peco (Hsiao et al. 2020) and Cyclone (Scialdone et al. 2015). Each classifier was applied to the U5-hNSC and BT322 scRNA-seq datasets. Data was prepared as required to run each classifier method. The quality of predicted cell cycle states for each classification method was determined by computing the AMI score relative to reference cell cycle states. Ten-fold cross-validation with a 20% hold-out testing set was used to determine the variance of AMI scores for each cell cycle state from each classification method. For the U5-hNSCs the cell labels from O’Connor et al., 2021 were used as the reference cell cycle states for computing AMI scores. For the BT322 glioma stem cells the ccSeurat predicted cell cycle states were used as the reference for computing AMI scores. Cells with “Unknown” labels were excluded when computing AMI scores. The median AMI scores were tabulated and plotted against the number of predicted states for each classifier. Representative cell cycle state predictions for each classification method were also visualized as UMAPs.

Because each classifier predicts different numbers of cell cycle states (3 – 8 cell cycle states) it was necessary to use simulated datasets to determine the range of AMI scores that correspond to specific amounts of similarity to the reference. Predicted cell cycle states with 3 to 8 states were simulated that contained specific 0 to 100% similarity to a simulated reference, at 10% increments. The average AMI from 100 simulated cell cycle classifications was computed for each specific amount of similarity to a simulated reference and plotted as a guide to assess the quality between classification methods with different numbers of cell cycle states.

Finding the optimal likelihood threshold

A neuroepithelial dataset of *in vivo* hNSCs from fetal tissue at 3 to 12 weeks post-conception from Zeng et al., 2023, that was independent of the ccAFv2 training data, was used to

determine the optimal likelihood threshold. Random sub-sampling of 90% of cells for each timepoint was used to determine the variance of the classifications and ccAFv2 was applied with likelihood thresholds ranging from 0.0 to 0.9 by increments of 0.1. For each iteration metrics were collected including the number of cells predicted, and an AMI score computed using ccSeurat cell cycle states as the reference. Cells with “Unknown” labels were excluded when computing AMI scores. Metrics were not computed when 20 or fewer cells were predicted. Student’s t-tests were used to compare AMIs computed at each examined likelihood threshold with those derived from a likelihood threshold of 0.0, which is equivalent to not using a likelihood threshold, and a significant difference was considered a p-value ≤ 0.05. A baseline for comparison was provided by random removal of an equivalent percentage of cells that were classified as “Unknown” for each likelihood threshold, and an AMI was computed with the remaining cells. Student’s t-tests were used to compare AMIs of the likelihood thresholded and random removal at each likelihood threshold, and a significant difference was considered a p-value ≤ 0.05.

Determining the sensitivity of ccAFv2 to missing genes

Sensitivity analysis was conducted on the U5-hNSC dataset by randomly excluding a defined percentage of classifier genes (1-90%) and applying the ccAFv2 classifier. Each percentage of classifier genes was subsampled ten times and for each iteration the metrics error rate and percentage of cells predicted were recorded.

Regressing out cell cycle transcriptional signatures using ccAFv2 marker genes

The average expression from the marker genes for each cell cycle state (**Supplementary Table S6**) was computed using the AddModuleScore function in Seurat. The S and G2/M or Late G1, S, S/G2, G2/M, M/Early G1 module scores were regressed out in the SCTransform function in Seurat. The variance explained by the first principal component of the marker genes was used as a metric for co-expression of the cell cycle transcriptional signatures. Empirical p-values were calculated by comparing the observed variance explained to the variance explained of 1,000 randomly sampled gene sets of the same size. Significantly regressing out the cell cycle transcriptional signature was determined by a reduction in the variance explained that made the empirical p-value non-significant (>0.05).

Application of ccAFv2 to neuroepithelial scRNA-seq and snRNA-seq profiling studies

To maximize overlap with the ccAFv2 input genes, we enabled the option to apply SCTransform (do_sctransform) for SCTransformed datasets. The species ('human' or 'mouse') and gene ID ('ensembl' or 'symbol') options were configured based on the specifications of each dataset. Predicted cell cycle states were collected from each dataset and integrated with meta information.

Appplication of ccAFv2 to ST-seq data

The ccAFv2 classifier was applied to spatial transcriptomics (ST-seq) data obtained from tissue slices of a fetus at four weeks post-conception from Zeng et al., 2023.

R and Python package for ccAFv2

The ccAFv2 classifier has been implemented as an R package (https://github.com/plaisier-lab/ccafv2_R) that can be installed and used as part of a Seurat workflow, and works for both Seurat version 4 and 5 (**Supplementary Figure S8**). Due to differences in the Seurat v5 SCTransform function it was necessary to set the vst.flavor equal to “v1” to make it equivalent to Seurat v4.3.0.1, and leaving the vst.flavor as the default in v5 leads to only small differences (**Supplementary Figure S8**). For the Seurat v5.0.2 the matrixStats package was required to be v1.1.0. Additionally, the ccAFv2 classifier has been implemented as a Python PyPI installed package (https://pypi.org/project/ccAF/) that can be installed and used as part of a scanpy workflow. It should be noted that SCTransform normalization is the suggested method for preparing data that will be classified by ccAFv2, and as of now there is no SCTransform option in scanpy.

Data Access

The four LGG scRNA-seq samples generated for these studies have been deposited in GEO under the accession GSE263796. All the data in used in our analyses are available on Zenodo (https://doi.org/10.5281/zenodo.10963137). We also provide all code on github.com (https://github.com/plaisier-lab/ccAFv2) and Docker images on DockerHub that were used to run all analyses (https://hub.docker.com/r/cplaisier/ccafv2_extra and https://hub.docker.com/r/cplaisier/ccnn).

R package for ccAFv2

We have developed an R package that can be installed using devtools from github. The instructions for installation and usage can be found on github: https://github.com/plaisier-lab/ccafv2_R

Python package for ccAFv2

We have also developed an Python package that can be installed using pip from PyPi. The instructions for installation and usage can be found on PyPi and github: https://pypi.org/project/ccAFv2/ and https://github.com/plaisier-lab/ccAFv2_py

Docker images for ccAFv2

We also provide Docker images that include all dependencies and ccAFv2 preinstalled to make the package more user friendly. Please see the github repositories for information about how to get, run, and use the Docker images.

- R package:
    - Seruat v4: https://hub.docker.com/r/cplaisier/ccafv2_seurat4
    - Seurat v5: https://hub.docker.com/r/cplaisier/ccafv2_seurat5
- Python package: https://hub.docker.com/r/cplaisier/ccafv2_py

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the following grants: NINDS/NIH (R01NS119650: A.P., P.P., C.L.P.) and (R01NS123038: CLP); NCI/NIH (R01CA190957; R21CA170722; P30CA15704: P.P.); DoD Translational New Investigator Award (CA100735: P.P.); and the Pew Biomedical Scholars Program (P.P.). The BT237 cell line was provided by Keith Ligon from the Dana-Farber Cancer Institute Center for Patient Derived Models. Some components of Figure 5 were made using Biorender.com.

Author contributions

Project conception and experimental design were carried out by CLP, SAO, PJP, and AP. Implementation of ccAFv2 and the R and Python packages was performed jointly by SAO and CLP. ScRNA-seq of LGG primary cell lines was performed by LG and J-PH. Single cell/nuclei dataset curation was performed by SAO under the supervision of CLP. Testing of ccAFv2 was performed by SAO under the supervision of CLP. ST-seq dataset curation and testing was performed by CLP. The manuscript was written by SAO and CLP with input from all authors.

References

Adams NM, Grassmann S, Sun JC. 2020. Clonal expansion of innate and adaptive lymphocytes. *Nat Rev Immunol* **20**: 694–707.

Augustus M, Pineau D, Aimond F, Azar S, Lecca D, Scamps F, Muxel S, Darlix A, Ritchie W, Gozé C, et al. 2021. Identification of CRYAB+ KCNN3+ SOX9+ Astrocyte-Like and EGFR+ PDGFRA+ OLIG1+ Oligodendrocyte-Like Tumoral Cells in Diffuse IDH1-Mutant Gliomas and Implication of NOTCH1 Signalling in Their Genesis. *Cancers* **13**: 2107.

Bressan RB, Dewari PS, Kalantzaki M, Gangoso E, Matjusaitis M, Garcia-Diaz C, Blin C, Grant V, Bulstrode H, Gogolok S, et al. 2017. Efficient CRISPR/Cas9-assisted gene targeting enables rapid and precise genetic manipulation of mammalian neural stem cells. *Dev Camb Engl* **144**: 635–648.

Bukovsky A, Caudle MR, Svetlikova M, Wimalasena J, Ayala ME, Dominguez R. 2005. Oogenesis in adult mammals, including humans. *Endocrine* **26**: 301–316.

Cebrian-Silla A, Nascimento MA, Redmond SA, Mansky B, Wu D, Obernier K, Romero Rodriguez R, Gonzalez-Granero S, García-Verdugo JM, Lim DA, et al. 2021. Single-cell analysis of the ventricular-subventricular zone reveals signatures of dorsal and ventral adult neurogenesis eds. J.G. Gleeson, M.E. Bronner, and D. Van der Kooy. *eLife* **10**: e67436.

Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, Allache R, Baig S, Yan X, Bourgey M, et al. 2020a. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat Commun* **11**: 3406.

Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, Allache R, Baig S, Yan X, Bourgey M, et al. 2020b. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat Commun* **11**: 3406.

Davis AA, Temple S. 1994. A self-renewing multipotential stem cell in embryonic rat cerebral cortex. *Nature* **372**: 263–266.

Doetsch F. 2003. A niche for adult neural stem cells. *Curr Opin Genet Dev* **13**: 543–550.

Dulken BW, Leeman DS, Boutet SC, Hebestreit K, Brunet A. 2017. Single-Cell Transcriptomic Analysis Defines Heterogeneity and Transcriptional Dynamics in the Adult Neural Stem Cell Lineage. *Cell Rep* **18**: 777–790.

Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, Guo Y, Takei Y, Yun J, Cai L, et al. 2018. The adult human testis transcriptional cell atlas. *Cell Res* **28**: 1141–1157.

Hanahan D, Weinberg RA. 2000. The Hallmarks of Cancer. *Cell* **100**: 57–70.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-3587.e29.

Harada Y, Yamada M, Imayoshi I, Kageyama R, Suzuki Y, Kuniya T, Furutachi S, Kawaguchi D, Gotoh Y. 2021. Cell cycle arrest determines adult neural stem cell ontogeny by an embryonic Notch-nonoscillatory Hey1 module. *Nat Commun* **12**: 6562.

Hsiao CJ, Tung P, Blischak JD, Burnett JE, Barr KA, Dey KK, Stephens M, Gilad Y. 2020. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res* **30**: 611–621.

Johe KK, Hazel TG, Muller T, Dugich-Djordjevic MM, McKay RD. 1996. Single factors direct the differentiation of stem cells from the fetal and adult central nervous system. *Genes Dev* **10**: 3129–3140.

Joseph C, Mangani AS, Gupta V, Chitranshi N, Shen T, Dheer Y, KB D, Mirzaei M, You Y, Graham SL, et al. 2020. Cell Cycle Deficits in Neurodegenerative Disorders: Uncovering Molecular Mechanisms to Drive Innovative Therapeutic Development. *Aging Dis* **11**: 946–966.

Lim DA, Alvarez-Buylla A. 2016. The Adult Ventricular–Subventricular Zone (V-SVZ) and Olfactory Bulb (OB) Neurogenesis. *Cold Spring Harb Perspect Biol* **8**: a018820.

Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T. 2017. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun* **8**: 22.

Llorens-Bobadilla E, Zhao S, Baser A, Saiz-Castro G, Zwadlo K, Martin-Villalba A. 2015. Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell* **17**: 329–340.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**: e8746.

Ma F, Pellegrini M. 2020. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinforma Oxf Engl* **36**: 533–538.

Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, Haeussler M, Sandoval-Espinosa C, Liu SJ, Velmeshev D, et al. 2017. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**: 1318–1323.

Obernier K, Cebrian-Silla A, Thomson M, Parraguez JI, Anderson R, Guinto C, Rodriguez JR, Garcia-Verdugo J-M, Alvarez-Buylla A. 2018. Adult neurogenesis is sustained by symmetric self-renewal and differentiation. *Cell Stem Cell* **22**: 221-234.e8.

O'Connor SA, Feldman HM, Arora S, Hoellerbauer P, Toledo CM, Corrin P, Carter L, Kufeld M, Bolouri H, Basom R, et al. 2021. Neural G0: a quiescent-like state found in neuroepithelial-derived cells and glioma. *Mol Syst Biol* **17**: e9522.

Pauklin S, Vallier L. 2013. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell* **155**: 135–147.

Qiu P. 2020. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun* **11**: 1169.

Schwabe D, Formichetti S, Junker JP, Falcke M, Rajewsky N. 2020. The transcriptome dynamics of single cells during the cell cycle. *Mol Syst Biol* **16**: e9946.

Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F. 2015. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**: 54–61.

Slyper M, Porter CBM, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, Smillie C, Smith-Rosario G, Wu J, Dionne D, et al. 2020. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat Med* **26**: 792–802.

Soufi A, Dalton S. 2016. Cycling through developmental decisions: how cell cycle dynamics control pluripotency, differentiation and reprogramming. *Dev Camb Engl* **143**: 4301–4311.

Stallaert W, Taylor SR, Kedziora KM, Taylor CD, Sobon HK, Young CL, Limas JC, Varblow Holloway J, Johnson MS, Cook JG, et al. 2022. The molecular architecture of cell cycle arrest. *Mol Syst Biol* **18**: e11087.

Xie P, Gao M, Wang C, Zhang J, Noel P, Yang C, Von Hoff D, Han H, Zhang MQ, Lin W. 2019. SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res* **47**: e48.

Zeng B, Liu Z, Lu Y, Zhong S, Qin S, Huang L, Zeng Y, Li Z, Dong H, Shi Y, et al. 2023. The single-cell and spatial transcriptional landscape of human gastrulation and early brain development. *Cell Stem Cell* **30**: 851-866.e7.

Zhang Q, Wu X, Fan Y, Jiang P, Zhao Y, Yang Y, Han S, Xu B, Chen B, Han J, et al. 2021. Single-cell analysis reveals dynamic changes of neural cells in developing human spinal cord. *EMBO Rep* **22**: e52728.

Zheng SC, Stein-O'Brien G, Augustin JJ, Slosberg J, Carosso GA, Winer B, Shin G, Bjornsson HT, Goff LA, Hansen KD. 2022. Universal prediction of cell-cycle position using transfer learning. *Genome Biol* **23**: 41.

Figure Legends

Figure 1. Implementing and testing the ccAFv2 classifier. **A.** The design of the Artificial Neural Network (ANN) implemented for the ccAFv2. Expr. = expression, ReLU = Rectified Linear Units. **B.** Method designed to determine the predicted class from the likelihoods generated by running expression data from a single cell through the ccAFv2 ANN. **C.** Pipeline for the quality control, training, and evaluation of the ccAFv2 classifier. Data is quality controlled and normalized using scTransform. Then the data is split 80% for training and 20% for testing using 10-fold cross validation. For each iteration three steps are performed: 1) the ANN is trained using the training dataset, 2) then the trained classifier is used to predict on the test set, and 3) the predicted labels are compared to the true labels to evaluate the accuracy of the classifier. **D.** Comparison of five different classification methods using F1 scores (a metric that integrates precision and recall, and has a maximum value of 1), from the 10-fold cross validation analysis of training on the U5-hNSCs. The F1 scores are computed for each cell cycle state from each of the 10 testing datasets. **E.** Determining the optimal number of neurons in each hidden layer using average U5-hNSC F1 score across cell cycle states on the x-axis, and the average AMI score across the remaining datasets (U5-hNSCs; glioma stem cells: BT322, BT324, BT326, BT333, BT363, BT368; tumor cells: BT363, BT368; and Grade 2 Astrocytoma: LGG275). Each combination of hidden layer neurons is labeled using: number of hidden layer one neurons / number of hidden layer two neurons. The chosen optimal configuration of 600 hidden layer 1 neurons and 200 hidden layer 2 neurons (600 / 200) is denoted in red.

Figure 2. Comparing the performance of ccAFv2 to existing cell cycle state classifiers. **A.** Median AMI score for each cell cycle classifier's predictions of the U5-hNSCs relative to the O'Connor et al., 2021 cell cycle states is plotted against the number of cell cycle states predicted by the classifier. The average similarity to the reference given the number of cell cycle states predicted by the classifier was computed and is plotted at 10 percent intervals to facilitate comparison between classifiers with differing numbers of predicted cell cycle states. **B.** Overlay of representative cell cycle state predictions on the U5-hNSCs UMAP. **C.** Median AMI score for each cell cycle classifier's predictions of the glioma stem cell line BT322 relative to the ccSeurat cell cycle states is plotted against the number of cell cycle states predicted by the classifier. The average similarity to the reference given the number of cell cycle states predicted by the classifier was computed and is plotted at 10 percent intervals to facilitate comparison between classifiers with differing numbers of predicted cell cycle states. **D.** Overlay of representative cell cycle state predictions on the BT322 UMAP.

Figure 3. Application of ccAFv2 to *in vivo* hNSCs and determining the optimal likelihood threshold. **A.** Proportion of cell cycle states in U5-hNSCs which were grown *in vitro* and were derived from a fetus at 8 PCW for both ccAFv2 and ccSeurat. **B.** Proportion of cell cycle states of hNSCs extracted from 3 – 12 PCW fetal tissue for both ccAFv2 and ccSeurat (Zeng et al, 2023). **C**. Histograms of cyclin expression in the *in vitro* cultured U5-hNSCs grouped by cell cycle phase. The cell cycle phase with the highest expression for each cyclin is denoted by an arrowhead and line colored in the same color. **D.** Histograms of cyclin expression in the *in vivo* hNSCs from 8 PCW (Zeng et al, 2023) defined cell cycle phases. The cell cycle phase with the

highest expression for each cyclin is denoted by an arrowhead and line colored in the same color. **E.** Testing different likelihood thresholds 0.0 to 0.9 using AMI score and percent of cells predicted as the metrics. Dashed red line indicates 90 percent of cells were predicted, and red dot indicates significantly improved AMI score due to applying threshold. **F.** Comparison of likelihood threshold application to random removal of the same number of cell predictions for in vivo hNSCs 3 PCW. Metric used for assessment is the AMI score. Likelihood thresholds start at 0.3 on the x-axis because AMI values at likelihood thresholds 0 to 0.3 are the same. ns = not significant.

Figure 4. Application of ccAFv2 to LGGs with and without growth factors. Overlay of ccAFv2 predictions on Grade 2 astrocytoma (LGG275) with growth factors (**A**) and without growth factors (**B**). **C.** Summary of the proportion of cell cycle states with and without growth factors. Overlay of ccAFv2 predictions on Grade 3 oligodendroglioma (BT237) with growth factors (**D**) and without growth factors (**E**). **F**. Summary of the proportion of cell cycle states with and without growth factors.

Figure 5. Application of ccAFv2 to single cells and nuclei from human and mice. **A.** Summary schematic of data ccAFv2 can be applied to and suggested data preparation. **B.** Proportion of cells assigned to each cell cycle state for scRNA-seq data from the developing human telencephalon (Nowakowski et al, 2017). **C.** Proportions of cell cycle states from scRNA-seq (C) and snRNA-seq (N) from spinal, cervical, lumbar, and thoracic regions from the developing human spinal cord at 8, 10, 11, 20, and 23 PCW (Zhang et al, 2021). **D.** Proportions of cell cycle states from scRNA-seq from GLAST and PROM1 flow-sorted cells from the subventricular zone (SVZ) of mice (Llorens-Bobadilla et al, 2015), and EGFR, GFAP, and PROM1 flow-sorted cells from the subventricular zone (SVZ) of adult mice (Dulken et al, 2017). qNSC1 = dormant quiescent neural stem cell, qNSC2 = primed-quiescent neural stem cell, aNSC1 = active neural stem cell, aNSC2 = actively dividing neural stem cell. qNSC = quiescent neural stem cell, aNSC = active neural stem cell. **E.** Proportions of cell cycle states from scRNA-seq from the ventricular (V)-SVZ of the adult mouse brain (Cebrian-Silla et al, 2021). Prog. = progenitors, Inh = inhibitor, Ex = excitatory, NSPCs = neural stem/progenitor cells, IntProg. = intermediate progenitor cells.

Figure 6. Application of ccAFv2 to spatial transcriptomics data from a fetus at 4 PCW. **A.** Summary of the proportions of cell cycle states from scRNA-seq (sc) at 4 PCW and 8 slices (s1 – s8) from a 4 PCW fetus. **B.** H&E staining for tissue slice s1. **C.** Spatial overlay of ccAFv2 predictions from slice s1. **D**–**J.** Likelihoods for each of the cell cycle states spatially overlayed onto slice s1. The magnitude of the likelihood indicates the probability that a cell with that cell cycle state underlies that spot of the spatial array.

Figures

Figure 1.

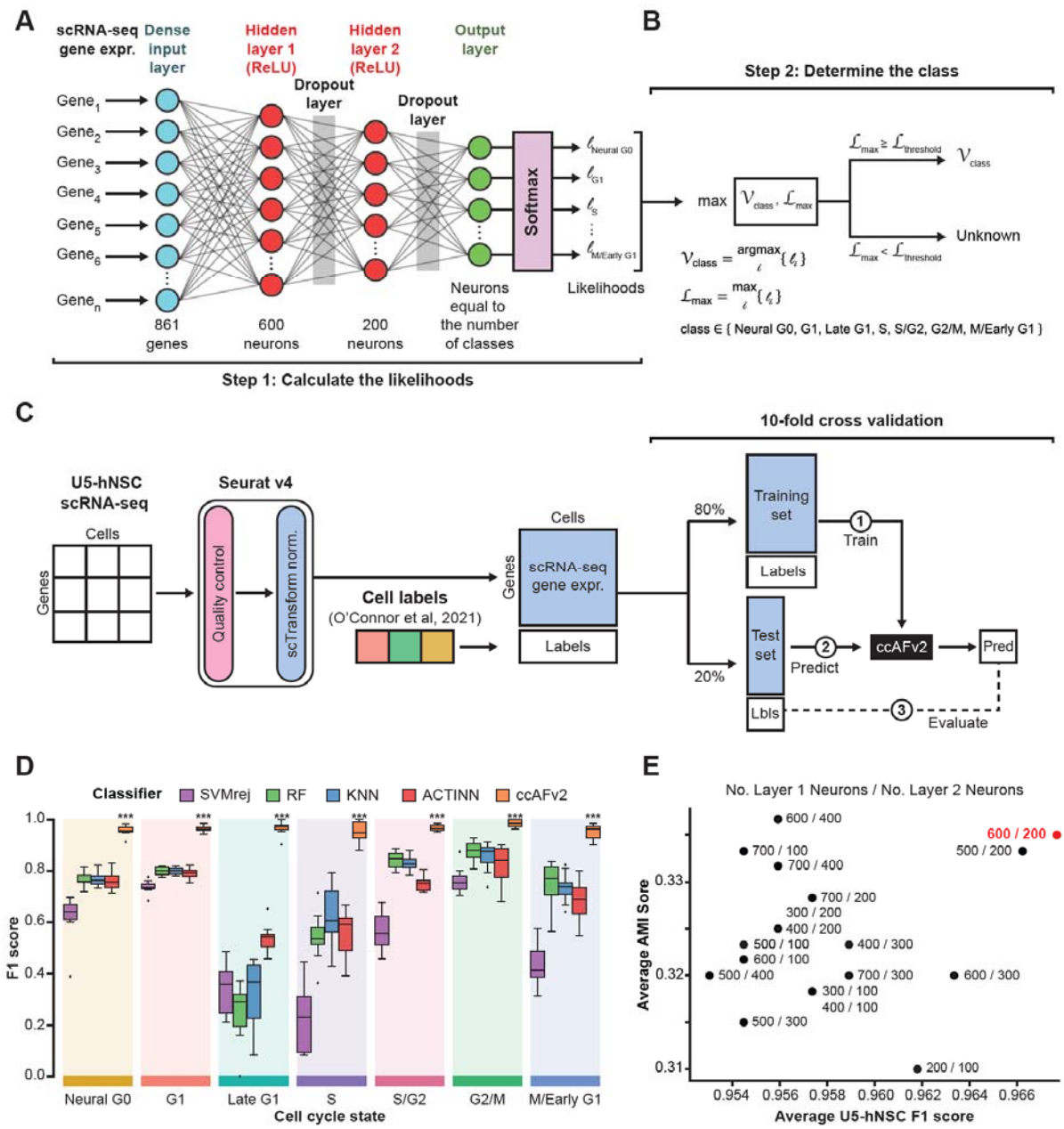
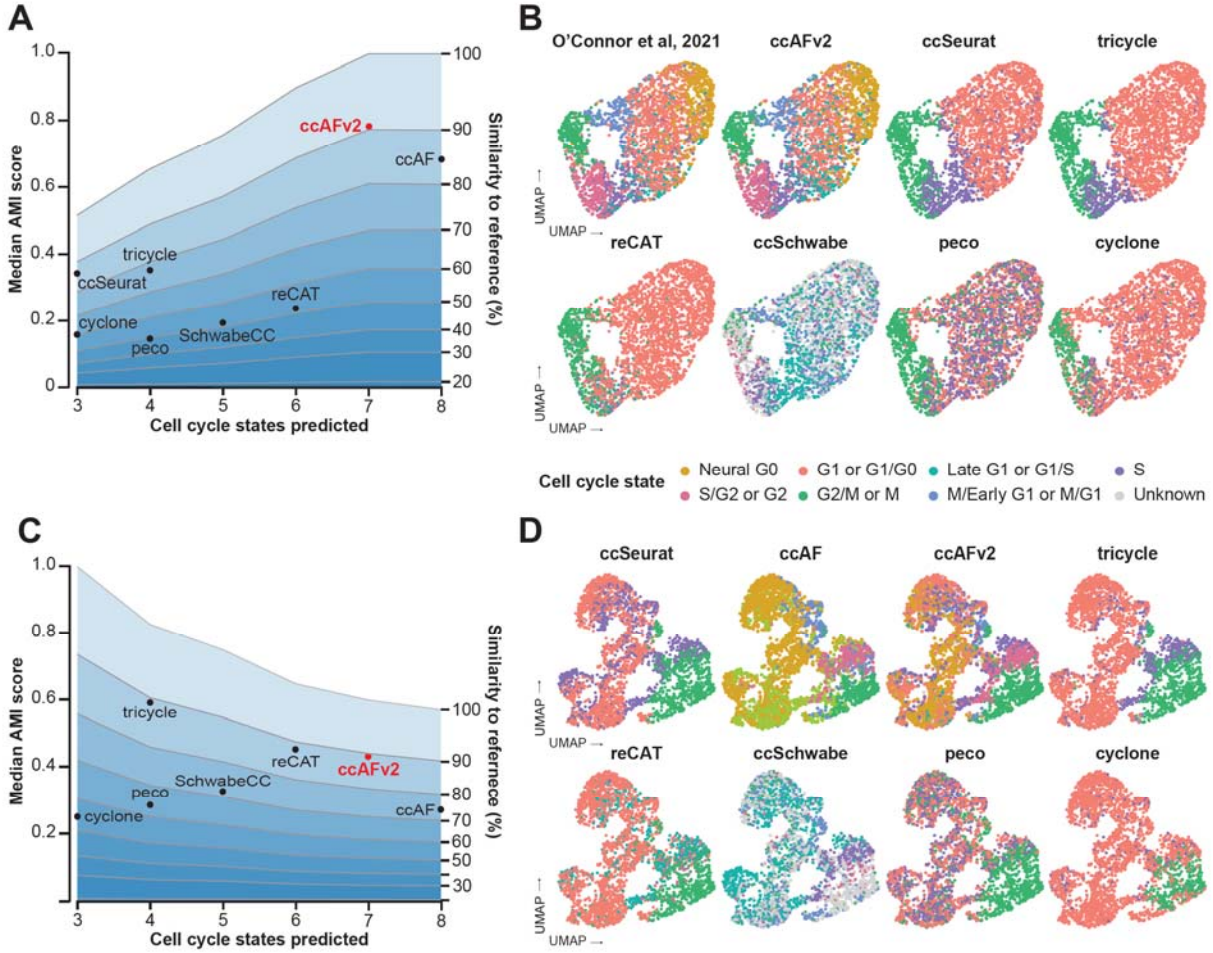
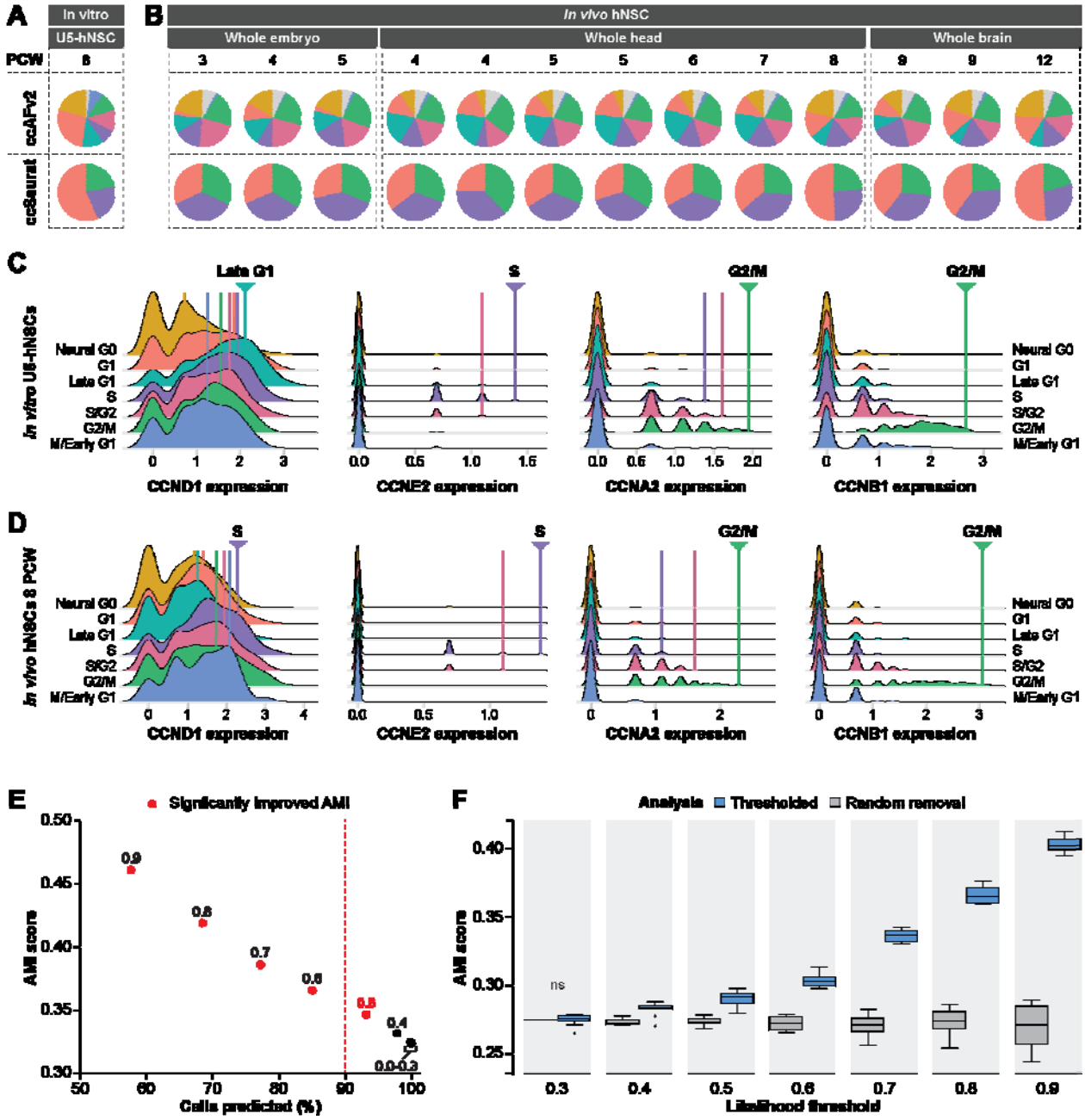


Figure 2.


A
Median AMI score
Similarity to reference (%)
Cell cycle states predicted
ccAFv2
ccAF
tricycle
ccSeurat
reCAT
cyclone
SchwabeCC
peco
B
O'Connor et al, 2021
ccAFv2
ccSeurat
tricycle
reCAT
ccSchwabe
peco
cyclone
UMAP
Cell cycle state
Neural G0
G1 or G1/G0
Late G1 or G1/S
S
S/G2 or G2
G2/M or M
M/Early G1 or M/G1
Unknown
C
Median AMI score
Similarity to refernece (%)
Cell cycle states predicted
tricycle
reCAT
ccAFv2
SchwabeCC
peco
cyclone
ccAF
D
ccSeurat
ccAF
ccAFv2
tricycle
reCAT
ccSchwabe
peco
cyclone
UMAP

825 **Figure 3.**



826

Figure 4.

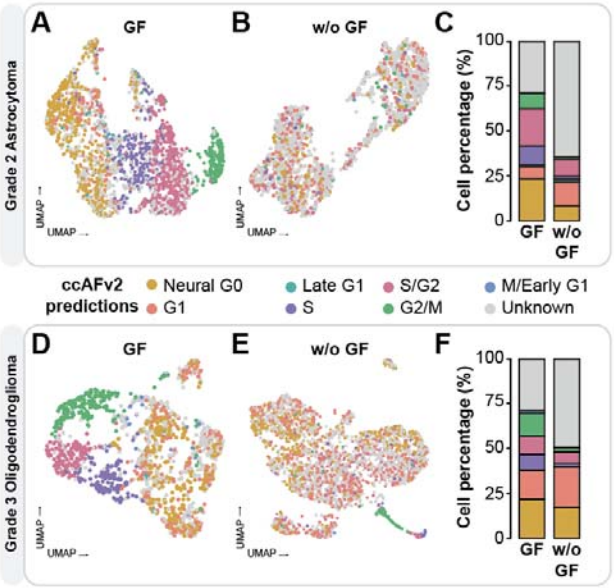

A
GF
B
w/o GF
C
Grade 2 Astrocytoma
Cell percentage (%)
100
75
50
25
0
GF
w/o GF
UMAP
ccAFv2 predictions
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Unknown
D
E
F
Grade 3 Oligodendroglioma

Figure 5.

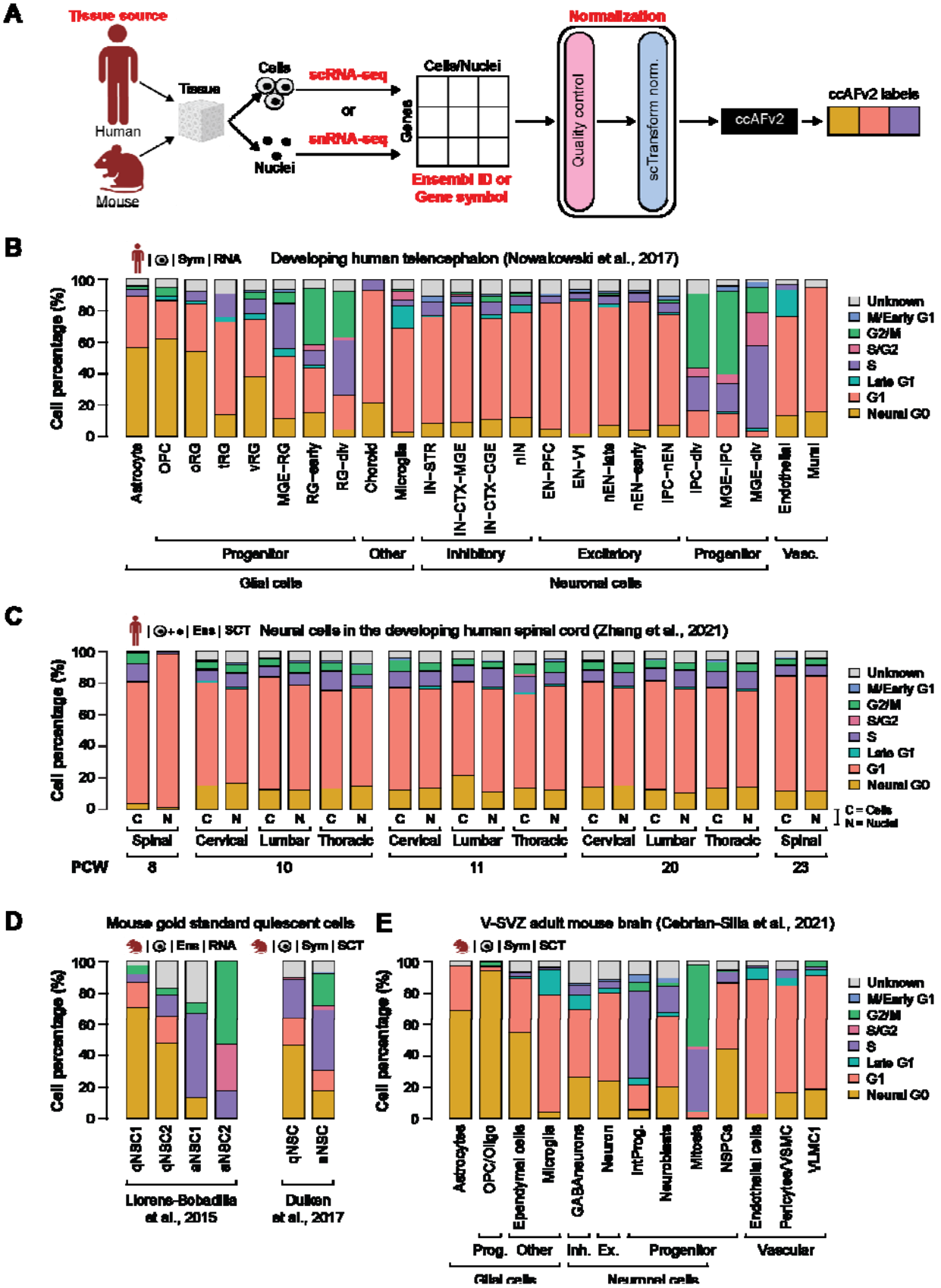
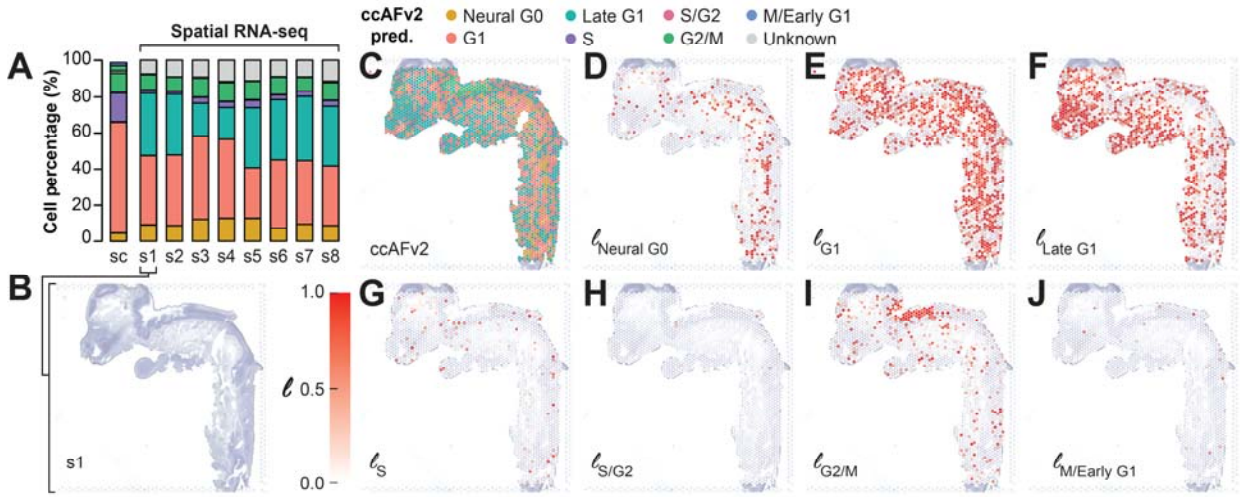
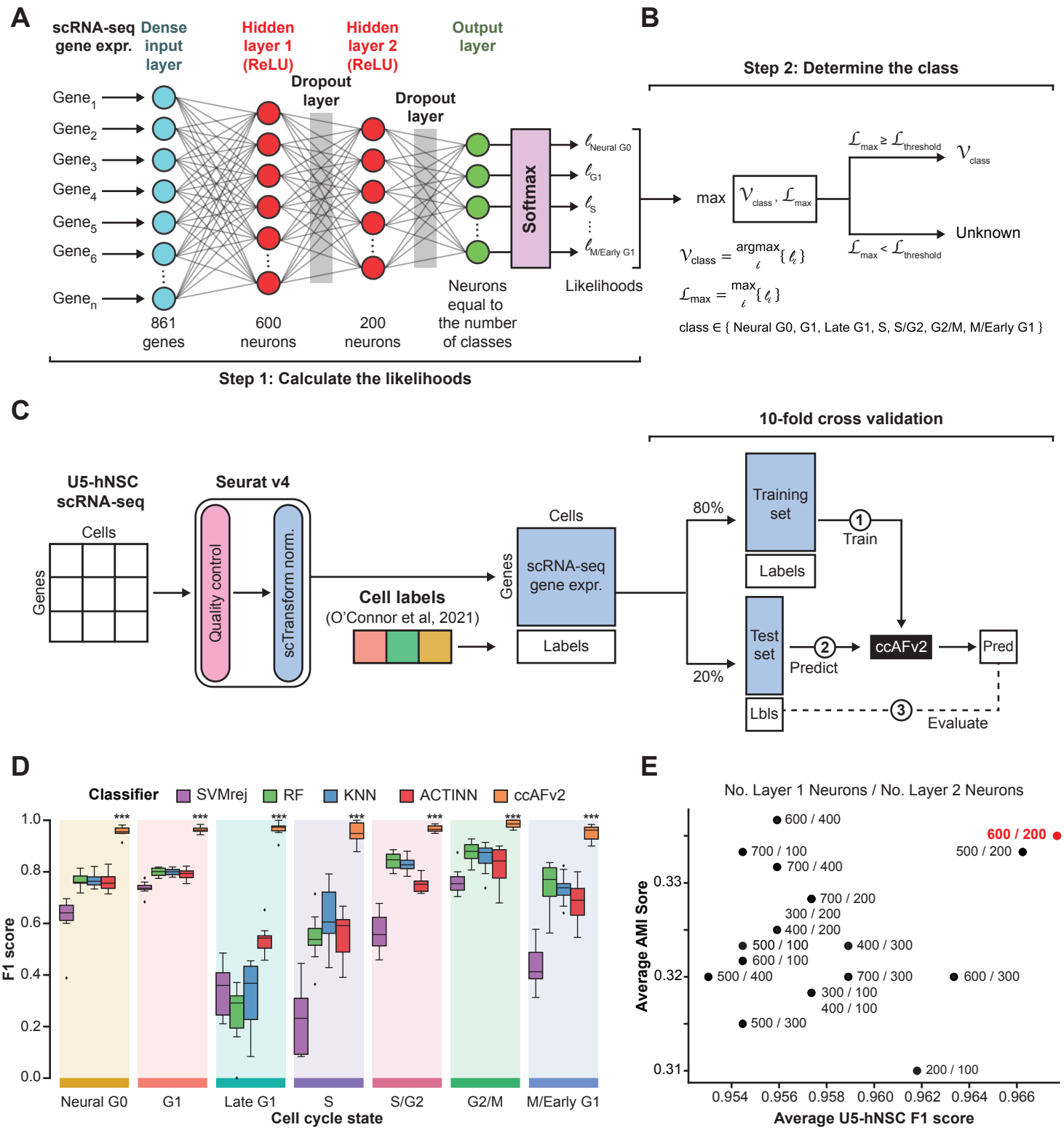
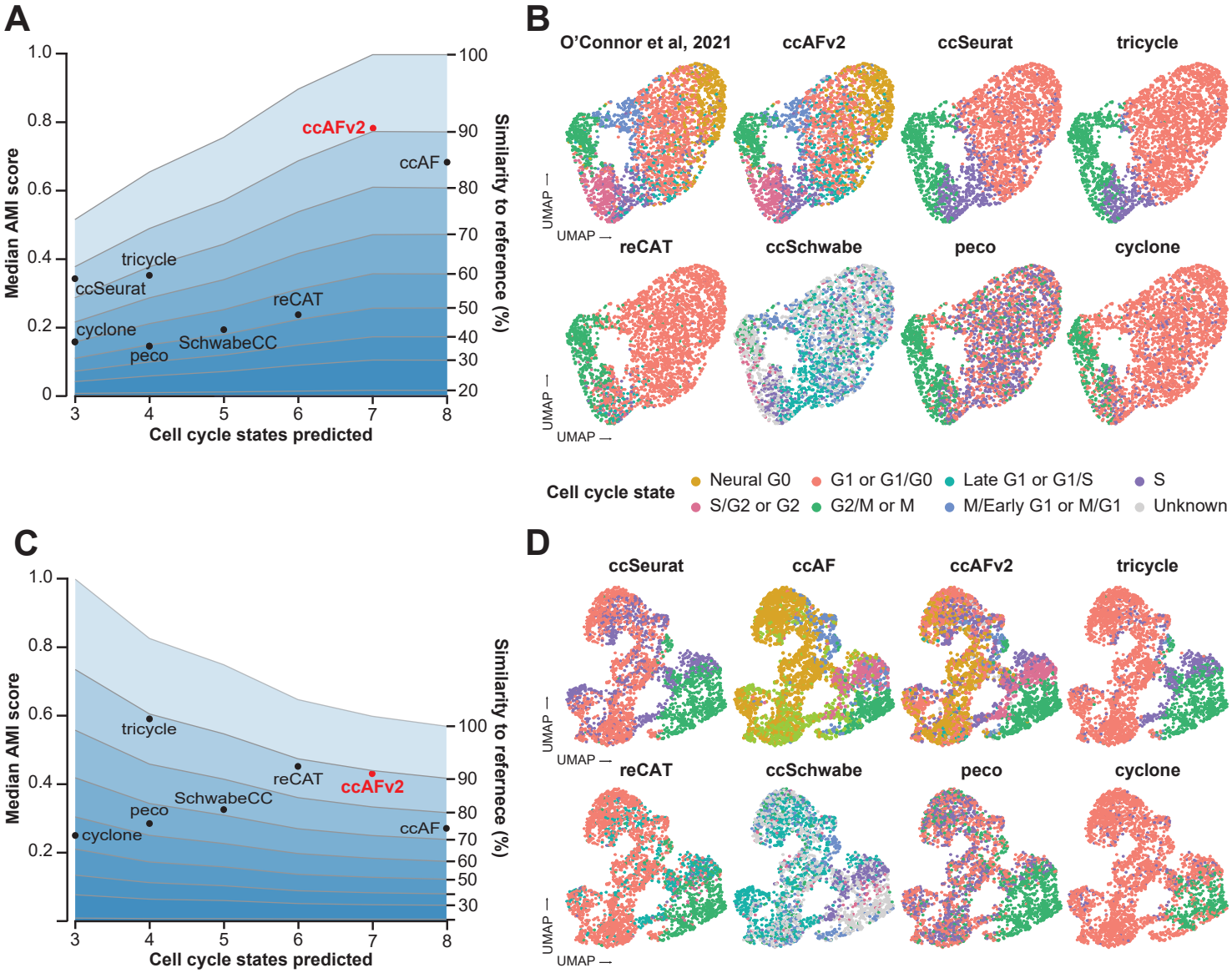
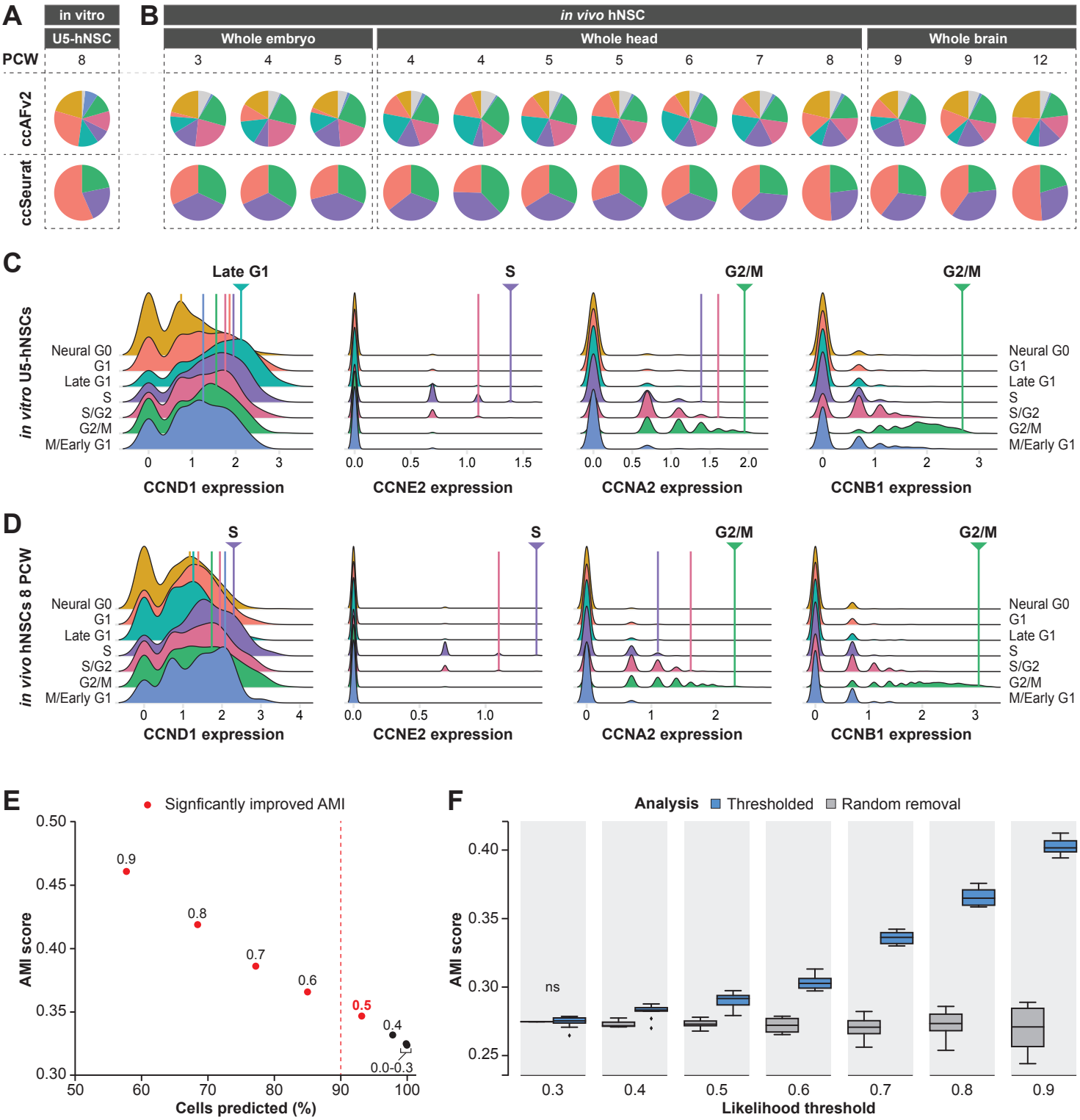

A
Tissue source
Human
Mouse
Tissue
Cells
Nuclei
scRNA-seq
or
snRNA-seq
Cells/Nuclei
Genes
Ensembl ID or Gene symbol
Normalization
Quality control
scTransform norm.
ccAFv2
ccAFv2 labels
B
Sym | RNA
Developing human telencephalon (Nowakowski et al., 2017)
Cell percentage (%)
Unknown
M/Early G1
G2/M
S/G2
S
Late G1
G1
Neural G0
Astrocyte
OPC
oRG
tRG
vRG
MGE-RG
RG-early
RG-div
Choroid
Microglia
IN-STR
IN-CTX-MGE
IN-CTX-CGE
nIN
EN-PFC
EN-V1
nEN-late
nEN-early
IPC-nEN
IPC-div
MGE-IPC
MGE-div
Endothelial
Mural
Progenitor
Other
Inhibitory
Excitatory
Progenitor
Vasc.
Glial cells
Neuronal cells
C
Ens | SCT
Neural cells in the developing human spinal cord (Zhang et al., 2021)
C = Cells
N = Nuclei
Spinal
Cervical
Lumbar
Thoracic
PCW
8
10
11
20
23
D
Mouse gold standard quiescent cells
Ens | RNA
Sym | SCT
qNSC1
qNSC2
aNSC1
aNSC2
qNSC
aNSC
Llorens-Bobadilla et al., 2015
Dulken et al., 2017
E
V-SVZ adult mouse brain (Cebrian-Silla et al., 2021)
Sym | SCT
Astrocytes
OPC/Oligo
Ependymal cells
Microglia
GABAneurons
Neuron
IntProg.
Neuroblasts
Mitosis
NSPCs
Endothelial cells
Pericytes/VSMC
V-LMC1
Prog.
Other
Inh.
Ex.
Progenitor
Vascular
Glial cells
Neuronal cells

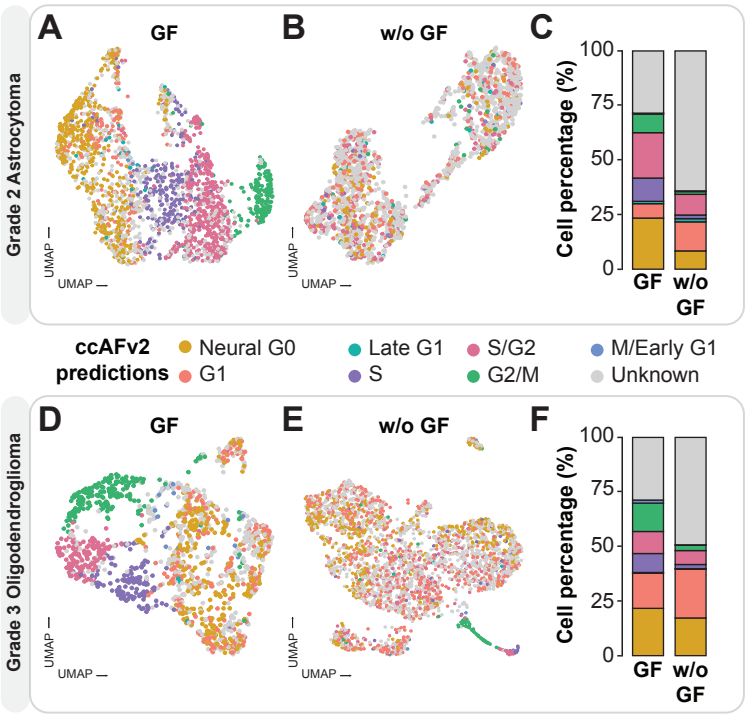
Figure 6.

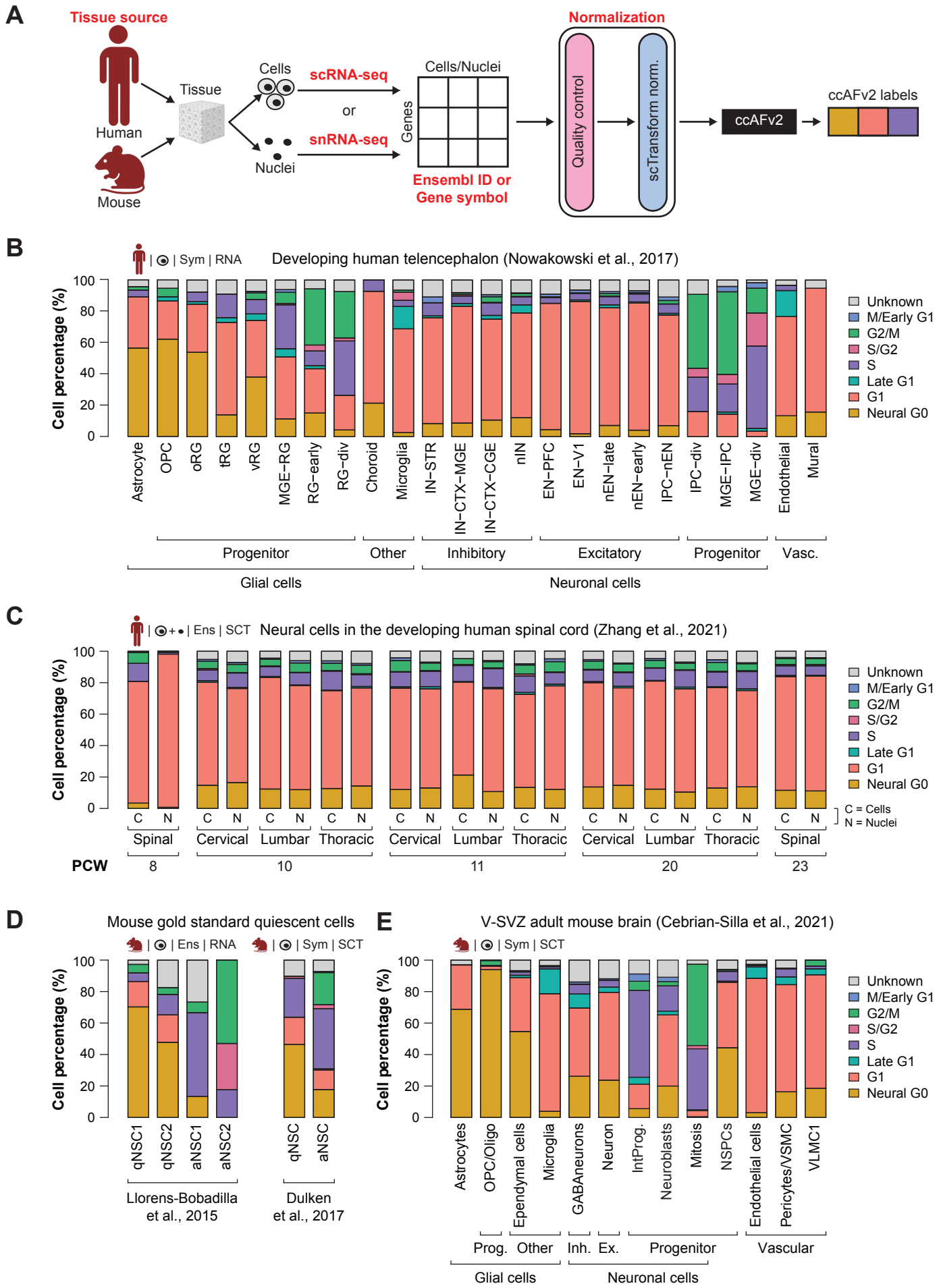

A
Spatial RNA-seq
Cell percentage (%)
100
80
60
40
20
0
sc
s1
s2
s3
s4
s5
s6
s7
s8
ccAFv2 pred.
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Unknown
B
s1
1.0
0.5
0.0
C
ccAFv2
D
ℓNeural G0
E
ℓG1
F
ℓLate G1
G
ℓS
H
ℓS/G2
I
ℓG2/M
J
ℓM/Early G1

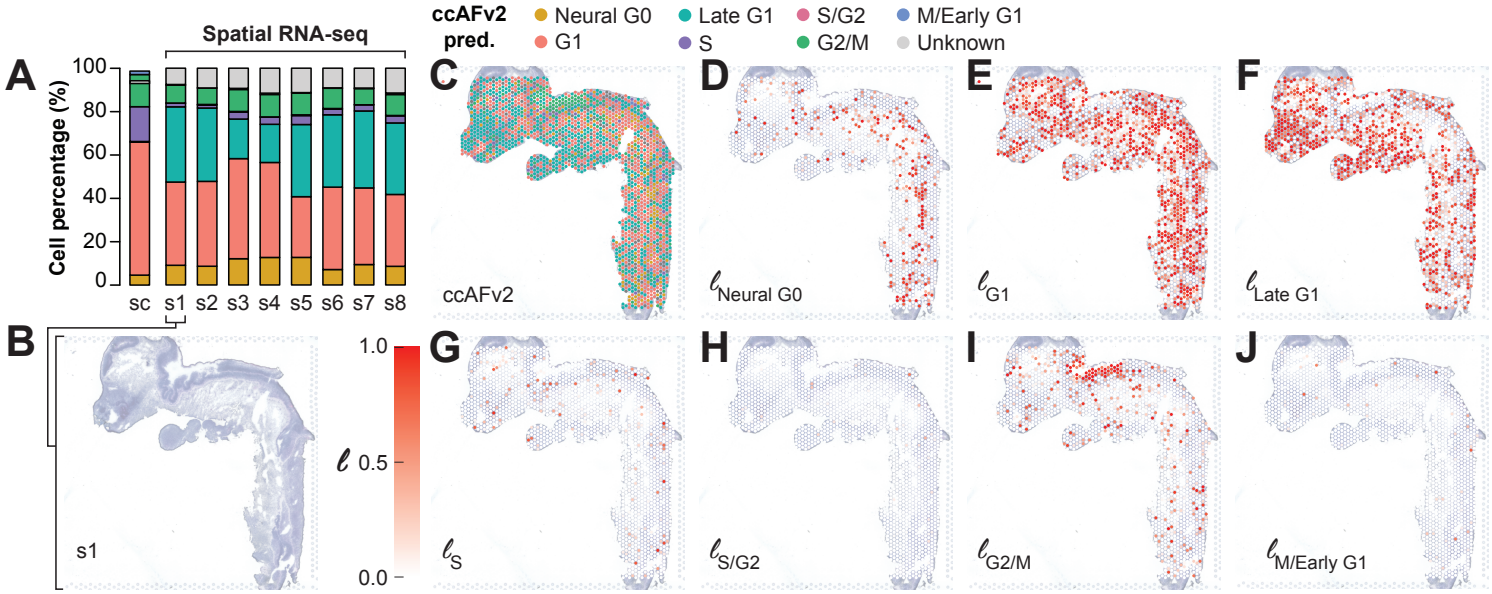

A
scRNA-seq gene expr.
Dense input layer
Hidden layer 1 (ReLU)
Dropout layer
Hidden layer 2 (ReLU)
Dropout layer
Output layer
Gene1
Gene2
Gene3
Gene4
Gene5
Genen
Softmax
ℓNeural G0
ℓG1
ℓS
ℓM/Early G1
861 genes
600 neurons
200 neurons
Neurons equal to the number of classes
Likelihoods
Step 1: Calculate the likelihoods
B
Step 2: Determine the class
max
𝒱class , ℒmax
ℒmax ≥ ℒthreshold
𝒱class
ℒmax < ℒthreshold
Unknown
𝒱class = argmaxi{ℓi}
ℒmax = maxi{ℓi}
class ∈ { Neural G0, G1, Late G1, S, S/G2, G2/M, M/Early G1 }
C
10-fold cross validation
U5-hNSC scRNA-seq
Cells
Genes
Seurat v4
Quality control
scTransform norm.
Cell labels
(O'Connor et al, 2021)
Cells
Genes
scRNA-seq gene expr.
Labels
80%
20%
Training set
Labels
1
Train
Test set
Lbls
2
Predict
ccAFv2
Pred
3
Evaluate
D
Classifier
SVMrej
RF
KNN
ACTINN
ccAFv2
F1 score
1.0
0.8
0.6
0.4
0.2
0.0
***
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Cell cycle state
E
No. Layer 1 Neurons / No. Layer 2 Neurons
Average AMI Sore
0.33
0.32
0.31
600 / 400
600 / 200
700 / 100
500 / 200
700 / 400
700 / 200
300 / 200
400 / 200
500 / 100
400 / 300
600 / 100
500 / 400
700 / 300
600 / 300
300 / 100
400 / 100
500 / 300
200 / 100
0.954
0.956
0.958
0.960
0.962
0.964
0.966
Average U5-hNSC F1 score


A
Median AMI score
Similarity to reference (%)
Cell cycle states predicted
ccAFv2
ccAF
tricycle
ccSeurat
reCAT
cyclone
peco
SchwabeCC
B
O'Connor et al, 2021
ccAFv2
ccSeurat
tricycle
reCAT
ccSchwabe
peco
cyclone
UMAP
Cell cycle state
Neural G0
G1 or G1/G0
Late G1 or G1/S
S
S/G2 or G2
G2/M or M
M/Early G1 or M/G1
Unknown
C
Similarity to refernece (%)
D
ccSeurat
ccAF
ccAFv2
tricycle
reCAT
ccSchwabe
peco
cyclone


A
in vitro
U5-hNSC
PCW
8
ccAFv2
ccSeurat
B
in vivo hNSC
Whole embryo
Whole head
Whole brain
3
4
5
4
4
5
5
6
7
8
9
9
12
C
in vitro U5-hNSCs
Late G1
S
G2/M
G2/M
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
CCND1 expression
CCNE2 expression
CCNA2 expression
CCNB1 expression
D
in vivo hNSCs 8 PCW
S
S
G2/M
G2/M
CCND1 expression
CCNE2 expression
CCNA2 expression
CCNB1 expression
E
Signficantly improved AMI
AMI score
Cells predicted (%)
0.9
0.8
0.7
0.6
0.5
0.4
0.0-0.3
F
Analysis
Thresholded
Random removal
AMI score
ns
Likelihood threshold


Grade 2 Astrocytoma
A
GF
B
w/o GF
C
Cell percentage (%)
100
75
50
25
0
GF
w/o GF
UMAP
UMAP
ccAFv2 predictions
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Unknown
Grade 3 Oligodendroglioma
D
GF
E
w/o GF
F
Cell percentage (%)
100
75
50
25
0
GF
w/o GF
UMAP
UMAP


A
Tissue source
Human
Mouse
Tissue
Cells
Nuclei
scRNA-seq
or
snRNA-seq
Cells/Nuclei
Genes
Ensembl ID or
Gene symbol
Normalization
Quality control
scTransform norm.
ccAFv2
ccAFv2 labels
B
| Sym | RNA
Developing human telencephalon (Nowakowski et al., 2017)
Cell percentage (%)
Astrocyte
OPC
oRG
tRG
vRG
MGE−RG
RG−early
RG−div
Choroid
Microglia
IN−STR
IN−CTX−MGE
IN−CTX−CGE
nIN
EN−PFC
EN−V1
nEN−late
nEN−early
IPC−nEN
IPC−div
MGE−IPC
MGE−div
Endothelial
Mural
Progenitor
Other
Inhibitory
Excitatory
Progenitor
Vasc.
Glial cells
Neuronal cells
Unknown
M/Early G1
G2/M
S/G2
S
Late G1
G1
Neural G0
C
| Ens | SCT
Neural cells in the developing human spinal cord (Zhang et al., 2021)
Cell percentage (%)
C = Cells
N = Nuclei
Spinal
Cervical
Lumbar
Thoracic
PCW
8
10
11
20
23
D
Mouse gold standard quiescent cells
| Ens | RNA
| Sym | SCT
Cell percentage (%)
qNSC1
qNSC2
aNSC1
aNSC2
qNSC
aNSC
Llorens-Bobadilla
et al., 2015
Dulken
et al., 2017
E
V-SVZ adult mouse brain (Cebrian-Silla et al., 2021)
| Sym | SCT
Cell percentage (%)
Astrocytes
OPC/Oligo
Ependymal cells
Microglia
GABAneurons
Neuron
IntProg.
Neuroblasts
Mitosis
NSPCs
Endothelial cells
Pericytes/VSMC
VLMC1
Prog.
Other
Inh.
Ex.
Progenitor
Vascular
Glial cells
Neuronal cells


A
Spatial RNA-seq
Cell percentage (%)
100
80
60
40
20
0
sc
s1
s2
s3
s4
s5
s6
s7
s8
ccAFv2 pred.
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Unknown
B
s1
1.0
ℓ 0.5
0.0
C
ccAFv2
D
ℓNeural G0
E
ℓG1
F
ℓLate G1
G
ℓS
H
ℓS/G2
I
ℓG2/M
J
ℓM/Early G1