Classifying cell cycle states and a quiescent-like G0 state using single-cell transcriptomics

Samantha A. O'Connor[1], Leonor Garcia [2], Anoop P. Patel [3,4], Benjamin B. Bartelle [1], 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 (qG0), 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 demonstrate that the ccAFv2 classifier is generalizable across cell types and all three germ layers by applying it to developing fetal cells. 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 or a PyPI package integrated with scanpy. 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 that include actively dividing cells. Particularly valuable for assessing the quality and generalizability of the classifier is an atlas of 245,906 cells from 15 different cell types, spanning all three germ layers, derived from human fetal tissue 3 to 12 weeks post-conception (Zeng et al. 2023). A second atlas of developing human spinal cord (Zhang et al. 2021) will be used to evaluate 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. 2020). Finally, we will apply the classifier to a high-resolution spatial-transcriptomics (ST-seq) study of a mouse embryo at 15.5 weeks post-conception (E15.5) allowed us to resolve canonical biological and morphological phenomena for the developmental stage. These 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 generalizability across cell types from all germ layers.

The goal of this research is to develop an improved cell cycle classifier using current state of the art machine learning technology. We aim to demonstrate that the classifier outperforms existing models and generalizes well across various cell types, library preparation methods (scRNA-seq, snRNA-seq, ST-seq), gene annotations, and normalization techniques. Lastly, we aim to provide a classifier with a more user-friendly interface to facilitate its application in future studies.

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 (**Supplemental Figure S1A**). 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; **Supplemental Table S1**). We applied 10-fold cross-validation (CV) for each classification method (**Supplemental Figure S1A**) 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 1C**), 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 (**Supplemental Figure S1B**-**C**). 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) across U5-hNSCs (O'Connor et al. 2021), a low grade glioma stem cell line (LGG275), six glioma stem cell lines (BT322, BT324, BT326, BT333, BT363, and BT368; Couturier et al. 2020), and two glioma tumors (BT363 and BT368; Couturier et al. 2020). The optimal combination was determined by having the highest average F1-score and Adjusted Mutual Information (AMI) score using ccSeurat as the reference (**Figure 1D**; **Supplemental Table S2**). We chose to employ the ccSeurat classifier (Butler et al. 2018) to predict the reference labels because true cell cycle state labels do not exist for all datasets. The ccSeurat classifier was chosen for three reasons: 1) it is the de facto standard method for cell cycle classification currently, 2) it performs well when applied to many different datasets, and 3) it uses a totally different underlying algorithm to classify cell cycle state than ccAFv2. 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 1D**). This specific parameterization has been assigned for the hidden layers of the ccAFv2 ANN, and all prior and subsequent ccAFv2 classifications use this parameterization.

Most important features for classifying ccAFv2 states

After optimizing the training of the ccAFv2 ANN, it is sensible to determine which features are most essential for classifying each of the seven states. We computed feature importance by permuting one of the 861 genes in the U5-hNSCs dataset and asking what impact that had on the likelihoods for each of the seven states. Randomizing the expression of an important feature for classifying a ccAFv2 state would lead to reductions in the states likelihood for cells known to be of this state. Thus, it is crucial that the dataset used for feature importance have cell cycle labels, which is why the U5-hNSCs were used for feature importance analyses (**Figure 1E**). We report the top 15 most important genes for each of seven ccAFv2 states (**Figure 1F-L**).

Eleven of the most important genes for classifying the Neural G0 state (**Figure 1F**) were also marker genes of Neural G0 in the U5-hNSCs. The first most important gene for classifying the G1 state (**Figure 1G**) was HMGN2, and in prior studies over-expression of HMGN2 in osteosarcoma cells led to significantly higher number of cells in G0/G1 (Liang et al. 2015). The top two most important genes for the classifying the Late G1 state (**Figure 1H**) include two Immediate-Early Genes (IEGs) CCN1 and CCN2 (also known as CYR61 and CTGF, respectively) which are known to be induced rapidly after initiation of cell cycle progression by many factors (Tullai et al. 2007). The top four most important genes for classifying the S state (**Figure 1I**) include three genes required for DNA replication during S phase (CLSPN, GINS2, and PCNA) and the cyclin associated with S phase (CCNE2). The top four most important genes for classifying the S/G2 state (**Figure 1J**) are all histones, specifically one H4 histone and multiple H1 histones isoforms that enable the condensation of nucleosomes into chromatin. The top five most important genes for classifying the G2/M state (**Figure 1K**) include a gene involved in keeping sister chromatids from separating (PTTG1), and two genes involved in kinetochore and centromere maintenance and function (CENPA, HMMR; Maxwell et al. 2005). Additionally, the ninth most important gene for classifying G2/M is CCNB1 the cyclin that peaks in mitosis, and MKI67 which is an established marker of cell proliferation (Scholzen and Gerdes 2000). Finally, the top three most important genes for classifying the M/Early G1 state (**Figure 1L**) are a microtubule component protein TUBA1B, a microtubule associated protein STMN1, and a component of the chromosome passage protein complex (CPC) which is essential for sister chromatid alignment and segregation during mitosis and cytokinesis (Vong et al. 2005). The functions of the key genes for classifying each state align well with the molecular processes of each cell cycle state, supporting the conclusion that the identified classes in U5-hNSCs reflect the underlying biology of the cell cycle.

Next, we evaluated the expression of important genes for each ccAFv2 state in an independent dataset, the *in vivo* hNSCs collected from whole fetal brain at 9 weeks post-conception (PCW 9 R1; Zeng et al. 2023). This allows us to assess the generalizability of these genes as key markers across novel datasets, providing insight into their broader applicability and robustness. The expression of important genes for all ccAFv2 states were expressed strongly in the state they marked, except for the important genes for the G1 state (**Supplemental Figure S1D**). In our prior study it was difficult to identify markers for G1 phase cells, and so the lack of translation for important genes for the G1 state is not surprising. The overall translation of the important genes to an independent dataset supports the hypothesis that ccAFv2 and the marker genes from U5-hNSCs are generalizable.

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 , 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 (**Supplemental Table S3**). We applied ccAFv2 alongside the other state-of-the-art cell cycle classification methods to *in vivo* hNSCs collected from whole human fetal brain at PCW 9 R1 (**Figure 2A**-**B**; Zeng et al. 2023). These cells represent an independent dataset for an unbiased comparison of the cell cycle prediction algorithms. The hNSCs from Zeng et al., 2023 were also chosen for their similarity to the U5-hNSCs and their added real-world relevance, as they were collected *in vivo*. We chose to employ the ccSeurat classifier (Butler et al. 2018) to predict the reference labels for classifier comparison for the reasons described above. The AMI score is impacted by the number of cell cycle states in the reference (i.e., three cell cycle states in ccSeurat), and the number of states predicted by each algorithm, e.g., seven cell cycle states in ccAFv2. 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 being tested. The highest AMI was observed for tricycle, showing an 80% similarity to the reference (**Figure 2A**). This result aligns with the UMAP colorization, indicating a strong match within classifiers that predicted a comparable number of classes to ccSeurat (**Figure 2B**). reCAT and ccAFv2, predicting six and seven cell cycle states, respectively, achieved the next highest AMI scores, both demonstrating over 70% similarity to the reference (**Figure 2A**). Notably, ccAFv2 identified an S/G2 cluster of cells positioned between the S and G2/M cells classified by ccSeurat and tricycle, which is biologically plausible (**Figure 2B**). Additionally, while Neural G0 cells are intermixed with G1 and Late G1 cells within the proliferating cell population on the left side of the UMAP, the right side reveals a distinct cluster of Neural G0 cells (**Figure 2B**). This suggests the presence of a quiescent population in these normal human neural stem cells that is not detectable by the ccSeurat, tricycle, or reCAT classifiers.

We also applied ccAFv2 alongside the other cell cycle classification methods to cells derived from a glioblastoma (GBM) patient tumor (BT322; Couturier et al. 2020; **Figure 2C**-**D**). GBM patient tumors are characterized by both quiescent and proliferating subpopulations (Tejero et al. 2019) making them ideal datasets for evaluating and comparing different cell cycle classification methods. We used the ccSeurat labels as the reference because true cell cycle state labels do not exist for this dataset. Like the *in vivo* PCW 9 R1 hNSCs, 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 (**Figure 2D**).

***In vivo* cyclin expression and maker genes validate ccAFv2 cell cycle states**

We explored the distribution of cell cycle states in 94,297 hNSCs collected from human fetal tissue at 3-12 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* (O'Connor et al, 2021) and 5,575 *in vivo* hNSCs from PCW 9 R1 were strikingly similar (**Figure 3C**-**G**). 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 (**Figure 3G**). Notably, the highest expression of the key regulator of cell cycle progression, CCND1, was observed in the Late G1 state (**Figure 3G**).

Additionally, we identified ccAFv2 marker genes that corresponded to cell cycle state markers in the PCW 9 R1 hNSCs. Differentially expressed genes for each cell cycle state were identified, and only those overlapping with the ccAFv2 marker gene lists were reported (**Figure 3H**). Genes important to the ccAFv2 classifier were enriched among the translatable marker genes for PCW 9 R1 (**Figure 3H**). The expression patterns of these translatable marker genes were consistent with those observed in O'Connor et al., 2021. The exclusive or semi-exclusive expression of these markers in adjacent cell cycle states strongly supports the presence of high-resolution ccAFv2 clusters in hNSCs *in vivo*. Furthermore, the biological function of the translatable marker genes for each ccAFv2 cell cycle state validates the biological basis of the ccAFv2 clusters, providing further evidence of their relevance.

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 the 94,297 hNSCs collected by Zeng et al., 2023.

We tested ccAFv2 likelihood thresholds ranging from 0.0 to 0.9 in increments of 0.1 (**Supplemental 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 3I**; **Supplemental Figure S2**; **Supplemental Table S4**). In other words, the removal of less certain classifications improves the accuracy of the overall classifications (**Figure 3I**). 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 9 weeks post-conception is shown in **Figure 3J**). The randomly removed cells do not increase the AMI (**Figure 3J**), 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 3J**; **Supplemental 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 3I**). 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). Factors affecting dropouts include the number of sequencing reads from each cell and the complexity of the cell's transcriptome. The ccAFv2 classifier uses the expression of 861 genes to predict cell cycle states. We hypothesized that dropouts could be simulated by randomly setting the expression of a defined percentage of genes to zero 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 (**Supplemental Figure S3-4**; **Supplemental Table S5-6**). 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). Missing information for 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. Introducing missing information for 40% of ccAFv2 input genes led to a 29% median error rate, and 96% of cells were predicted (**Supplemental Table S5**). The error rate was the most affected by the introduction of missing information (**Supplemental Figure S3A)** and the median percentage of cells predicted remained above 80% even when 70% of the input gene list was set to missing (**Supplemental Table S5**). When breaking down the error rate by cell cycle state, we observed that S and M/Early G1 had the highest error rates as missing information increased (**Supplemental Figure S3B**). However, the number of cells predicted remained relatively consistent across all states despite the increasing in missing data (**Supplemental Figure S3C**). The increase in error rate without a concomitant decrease in the number of cells predicted suggests that raising 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 introducing 20% missing information decreased from 12.2% median error rate at 0.5 likelihood threshold to 9.9% with a 0.7 likelihood threshold (**Supplemental Figure S3D**; **Supplemental Table S5**) and 6.2% with a 0.9 likelihood threshold (**Supplemental Figure S3E**; **Supplemental Table S5**). Thus, introducing 20% missing information led to four times the error rate, and the increased error rate can be mitigated in part by increasing the likelihood threshold. Increasing the likelihood threshold decreases the error rate by removing classifications for cells where the missing information has degraded the confidence in the prediction. By removing predictions with less confidence, the error rate decreases, but the overall number of cells classified with cell cycle states decreases. Testing the impact of increasing the likelihood threshold on the number of predicted cells can be quite insightful for choosing an appropriate likelihood threshold (**Supplemental Figure S3F**). Careful consideration of the balance between minimizing errors and retaining enough cells classified cells for downstream studies is essential.

Neural G0 is generalizable across germ layers

An important consideration for using ccAFv2 is whether it's quiescent-like G0 state can capture G0 cells in cell types beyond neuroepithelial cells. In Zeng et al., 2023, they profiled single cells from human fetal tissues, representing all three germ layers (endoderm, mesoderm, and ectoderm; **Figure 4A**). We applied ccAFv2 to 245,906 cells from the atlas and found that the Neural G0 state was detectable in every cell type and germ layer (**Figure 4B**). Next, we

identified marker genes for the Neural G0 classified cells from each cell type, and cross-comparison revealed 21 marker genes shared by at least seven different cell types (**Figure C** & **D**). All the genes but EPB41L4A-AS1, MEG3, VIM, and WSB1 were ccAFv2 marker genes for Neural G0 (**Figure 4C**). Notably, eight of the genes (GPM6B, MOXD1, PRCP, PCSK1N, PTPRZ1, S100B, SCRG1, and SOX4) ranked among the top 15 most important features for predicting the Neural G0 state (**Figure 1F** and **Figure 4C**). The relevance of the important features across the cell types and germ layers demonstrates the generalizability of Neural G0 to quiescent-like G0 states across germ layers. Thus, we propose renaming the Neural G0 state in ccAFv2 to 'quiescent-like G0' or 'qG0', as it more accurately reflects the generalizable nature of the quiescent-like G0 cells it classifies.

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 5**). For this analysis we tested the impact of adjusting the ccAFv2 likelihood threshold across a range of values 0 to 0.9 (**Supplemental 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 (**Supplemental 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 $\rightarrow$ S $\rightarrow$ S/G2 $\rightarrow$ G2/M $\rightarrow$ M/Early G1; **Figure 5A**, **C**, **D**, & **F**). Conversely, cells grown without growth factors exhibit a more dispersed distribution of cell cycle state labels (**Figure 5B**, **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, **Supplemental Table S7**; **Supplemental 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 (**Supplemental Table S7**; **Supplemental 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

It was crucial for ccAFv2 to be highly user-friendly, ensuring researchers can easily apply it across a wide range of datasets. The model was designed to accept inputs for tissue source, data type, and gene identifier, eliminating the need for manual data conversion (**Figure 6A**). 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 qG0 state was less frequent in all cell types for ccAFv2 relative to ccAF (**Figure 6B**; **Supplemental Table S8**). The qG0 state was distinctly less frequent in the neuronal cell types. For ccAF qG0 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 6B**). The glial cell types had the largest qG0 subpopulations (**Figure 6B**).

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 6C**). In the V-SVZ the glial cells tended to have larger qG0 subpopulations, neuronal cell types tended to have less qG0, and microglial had the smallest amount qG0 (**Figure 6C**). 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 qG0 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 qG0 subpopulations. Overall, the qNSCs are enriched with quiescent-like qG0 cells, and the aNSCs are at some stage of the cell cycle (**Figure 6D**). The proportion of cells classified as qG0 decreased for ccAFv2 in the quiescent NSCs (qNSCs) and was replaced by more G1, S/G2, and a small amount of G2/M (**Figure 6D**; **Supplemental 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. Additionally, we used the G0 arrest signature from Wiecek et al., 2023 to validate the qG0 state in ccAFv2 (Wiecek et al. 2023). We found significant enrichment of the G0 arrest signature (i.e., QuieScore G0) within the U5-hNSC qG0 and G1 states (**Supplemental Figure S8B**). These results continue to validate our assertion that qG0 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 6A**). 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 6E**), 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 without contextual information, relating these states with previously described biology can be challenging. 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, offering structural information that can be referenced to anatomical atlases and established histology. We applied ccAFv2 to the highest resolution sequencing-based spatial transcriptomic dataset currently available, derived from a slice of a mouse E15.5 embryo binned to 8µm, achieving near cellular resolution (**Figure 7A**-**B**). Despite rapid development at late prenatal stages, qG0 classified spots were highly prominent, particularly across the midbrain region. The proliferation marker Mki67 was minimally expressed in these G0 enriched areas and, in the brain, highlighted the described stem cell niches of the lower, medial, and caudal ganglionic eminences (LGE, MGE, CGE), along with the stem cell migratory paths from these regions along the sub-ventricular zone (SVZ; **Figure 7C**; Kriegstein and Noctor 2004).

From E12.5 to E17.5 the mouse cortex develops in a well-defined layers (**Figure 7D**). Applying ccAFv2 to the cortex captured the layered patterning of cell cycle states that fit with our current model of cortical development (**Figure 7E**-**Q**). Histology showed an outer layer of dermis (skin) marked by Krt5 expression (**Supplemental Figure S9A**-**B**), covering the developing skull identified by Col1a1 expression (**Supplemental Figure S9C**), followed by densely packed cells of the brain (**Figure 7D**). Precursors of excitatory neurons migrate along and divide in the SVZ with asymmetric division specifically occurring within the ventricular zone (VZ). A much smaller population of inhibitory neuronal precursors migrate and divide along the intermediate zone (IZ) and medial zone (MZ) before invading the cortical plate (CP) and differentiating into their neuronal sub-type. By E15.5, the CP is already well populated with post-mitotic neurons that previously migrated from the SVZ from E12.5-E14.5 and will form layers IV-VI of the adult cortex. Upon reaching the border of the CP and MZ, neural stem cells receive maturation factors from glial cell types including Reln (**Figure 7F**), with canonically post-mitotic neurons marked by Tbr1 (**Figure 7G**; Englund et al. 2005). This post-mitotic region classifies as qG0 by ccAFv2 (**Figure 7E** & **K**). Similarly, Mki67 is sparse within the CP, but highly active in the SVZ and VZ (**Figure 7J**). Intermediate progenitor cells (IPCs) in the SVZ, marked by Eomes (Tbr2), undergo symmetric division before radiating outward (**Figure 7G**). Within the VZ radial glia migrate further inward before asymmetric division, with newly divided IPCs radiating back up to the SVZ and outward to populate the cortex. These events are captured by ccAFv2 as a single enriched S and S/G2 band marking the SVZ (**Figure 7N** & **O**). We also observed two bands containing G2/M classified spots, corresponding to regions of symmetric division in the SVZ and asymmetric division in the VZ (**Figure 7P**). Cells committed to differentiation, immediately migrate outward along and the few spots that classify as M/Early G1 were almost entirely in the IZ and above (**Figure 7Q**). These results demonstrate that ccAFv2 can be effectively applied to spatial transcriptomics of the developing cortex, accurately recapitulating known biological insights into the spatial organization of proliferative activity.

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 includes a quiescent-like 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 embryos and fetuses 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 the quiescent-like G0 state (qG0) in ccAFv2 is detectable across all three germ layers in developing human fetal cells and provided a list of putative marker genes that are common across cell types. This corroborates our previous findings that qG0 was an active transcriptional signature executed by a subpopulation of U5-hNSCs (O'Connor et al. 2021). However, 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 **Supplemental 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.

Computing feature importance

Feature importance for all 861 ccAFv2 features was determined by permuting each feature’s expression, running ccAFv2 with the permuted expression matrix, and comparing the likelihoods of all cells for a specific ccAFv2 state to the unpermuted likelihoods of the same cells. The average difference in likelihood was computed for each feature in each ccAFv2 state. A negative average difference in likelihood indicates that a feature was important, and the most negative features are the most important.

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 PCW 9 R1 (Zeng et al, 2023) and BT322 (Couturier et al. 2020b) 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 both datasets, 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.

Cell cycle state validation using hNSCs (PCW 9 R1)

We used the hNSCs collected from whole fetal brain at nine weeks post-conception replicate 1 (PCW 9 R1) to validate the cell cycle states assigned by ccAFv2. After quality control (**Supplemental Table S11**) and normalization with sctransform, 5575 cells were classified into distinct cell cycle states using ccAFv2. We selected five key markers of cell cycle states: CCND1 (Late G1), CCNE2 (S), CCNA2 (G2/M), CCNB1 (G2/M), and CDK1 (G2/M) to assess the expression patterns associated with these phases. The average expression levels of the genes were calculated and visualized using violin plots, which were grouped according to the cell cycle states predicted by ccAFv2. In addition, we monitored the dynamic changes in the average expression of each key marker as cells transitioned between different cell cycle states. Student's t-tests were used to determine if the marker expression was significantly different at each cell cycle state compared to the G1 state. Finally, relative expression levels of top marker genes for each cell cycle state were identified using FindAllMarkers() and visualized using a heatmap, with cells grouped by cell cycle state.

Comparison of Neural G0 state with G0 arrest signature using QuieScore

We applied the QuieScore algorithm (https://github.com/dkornai/QuieScore) to the U5-hNSCs using the cancer type parameter of "LGG". The G0 cells were identified by a q_score_raw of greater than 3. We evaluated the similarity between the QuieScore-identified G0 cells with the ccAFv2-identified Neural G0 cells using hypergeometric enrichment analysis.

Determining the sensitivity of ccAFv2 to missing genes

Sensitivity analysis was conducted on the U5-hNSC dataset by randomly setting a defined percentage of classifier genes (1-90%) to zero 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.

Demonstrating the generalizability of ccAFv2

The 245,906 human fetal cells 3 to 12 weeks post conception (Zeng et al., 2023) encompassing fifteen cell types that represent all three germ layers were classified by ccAFv2. Positive marker genes for the Neural G0 cells were identified for each cell type using the FindAllMarkers function (log2 fold change ≥ 0.25; adjusted p-value ≤ 0.05). The Neural G0 markers were tabulated among each dataset and across all datasets to identify common Neural G0 marker genes.

Regressing out cell cycle transcriptional signatures using ccAFv2 marker genes

The average expression from the marker genes for each cell cycle state (**Supplemental Table S7**) 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.

Application of ccAFv2 to ST-seq data

We downloaded the transcriptome profiles for a 5 μm section of a male C57BL/6 mouse embryo taken from an FFPE tissue block obtained from Charles River Laboratories that was made public by 10X Genomics (https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-embryo). The 10X Visium HD Gene Expression Library preparation kit afforded a resolution of 2 μm^2 spots and details about sample preparation and library performance and QC can be found on the 10X website linked above. In Seurat the 2 μm^2 spots were binned into 8 μm^2 bins, the data log normalized, and ccAFv2 was applied to predict cell cycle states for each spot. Expression of key genes was plotted using the normalized and scaled values.

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 (**Supplemental Figure S10**). 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 (**Supplemental Figure S10**). 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

All raw and processed sequencing data for the four LGG scRNA-seq generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE263796.

All other data used in our analyses are available on Zenodo (https://zenodo.org/doi/10.5281/zenodo.10963136). 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 a 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. Interpretation of ST-seq results was performed by BBB and CLP. The manuscript was written by SAO, BBB, 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.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**: 411–420.

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. 2020. 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.

Englund C, Fink A, Lau C, Pham D, Daza RAM, Bulfone A, Kowalczyk T, Hevner RF. 2005. Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. *J Neurosci Off J Soc Neurosci* **25**: 247–251.

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.

831 Harada Y, Yamada M, Imayoshi I, Kageyama R, Suzuki Y, Kuniya T, Furutachi S, Kawaguchi D, Gotoh Y.
832 2021. Cell cycle arrest determines adult neural stem cell ontogeny by an embryonic Notch-
833 nonoscillatory Hey1 module. *Nat Commun* **12**: 6562.

834 Hsiao CJ, Tung P, Blischak JD, Burnett JE, Barr KA, Dey KK, Stephens M, Gilad Y. 2020. Characterizing and
835 inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res* **30**: 611–
836 621.

837 Johe KK, Hazel TG, Muller T, Dugich-Djordjevic MM, McKay RD. 1996. Single factors direct the
838 differentiation of stem cells from the fetal and adult central nervous system. *Genes Dev* **10**:
839 3129–3140.

840 Joseph C, Mangani AS, Gupta V, Chitranshi N, Shen T, Dheer Y, KB D, Mirzaei M, You Y, Graham SL, et al.
841 2020. Cell Cycle Deficits in Neurodegenerative Disorders: Uncovering Molecular Mechanisms to
842 Drive Innovative Therapeutic Development. *Aging Dis* **11**: 946–966.

843 Kriegstein AR, Noctor SC. 2004. Patterns of neuronal migration in the embryonic cortex. *Trends Neurosci*
844 **27**: 392–399.

845 Liang G, Xu E, Yang C, Zhang C, Sheng X, Zhou X. 2015. Nucleosome-binding protein HMGN2 exhibits
846 antitumor activity in human SaO2 and U2-OS osteosarcoma cell lines. *Oncol Rep* **33**: 1300–1306.

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

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

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

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

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

858 Maxwell CA, Keats JJ, Belch AR, Pilarski LM, Reiman T. 2005. Receptor for hyaluronan-mediated motility
859 correlates with centrosome abnormalities in multiple myeloma and maintains mitotic integrity.
860 *Cancer Res* **65**: 850–860.

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

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

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

870 Pauklin S, Vallier L. 2013. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell* **155**:
871 135–147.

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

873 Scholzen T, Gerdes J. 2000. The Ki-67 protein: from the known and the unknown. *J Cell Physiol* **182**: 311–
874 322.

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

877 Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F.
878 2015. Computational assignment of cell-cycle stage from single-cell transcriptome data.
879 *Methods* **85**: 54–61.

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

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

885 Stallaert W, Taylor SR, Kedziora KM, Taylor CD, Sobon HK, Young CL, Limas JC, Varblow Holloway J,
886 Johnson MS, Cook JG, et al. 2022. The molecular architecture of cell cycle arrest. *Mol Syst Biol*
887 **18**: e11087.

888 Tejero R, Huang Y, Katsyv I, Kluge M, Lin J-Y, Tome-Garcia J, Daviaud N, Wang Y, Zhang B, Tsankova NM,
889 et al. 2019. Gene signatures of quiescent glioblastoma cells reveal mesenchymal shift and
890 interactions with niche microenvironment. *EBioMedicine* **42**: 252–269.

891 Tullai JW, Schaffer ME, Mullenbrock S, Sholder G, Kasif S, Cooper GM. 2007. Immediate-early and
892 delayed primary response genes are distinct in function and genomic architecture. *J Biol Chem*
893 **282**: 23981–23995.

894 Vong QP, Cao K, Li HY, Iglesias PA, Zheng Y. 2005. Chromosome alignment and segregation regulated by
895 ubiquitination of survivin. *Science* **310**: 1499–1504.

896 Wiecek AJ, Cutty SJ, Kornai D, Parreno-Centeno M, Gourmet LE, Tagliazucchi GM, Jacobson DH, Zhang P,
897 Xiong L, Bond GL, et al. 2023. Genomic hallmarks and therapeutic implications of G0 cell cycle
898 arrest in cancer. *Genome Biol* **24**: 128.

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

902 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
903 spatial transcriptional landscape of human gastrulation and early brain development. *Cell Stem*
904 *Cell* **30**: 851-866.e7.

905 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
906 reveals dynamic changes of neural cells in developing human spinal cord. *EMBO Rep* **22**: e52728.

907 Zheng SC, Stein-O'Brien G, Augustin JJ, Slosberg J, Carosso GA, Winer B, Shin G, Bjornsson HT, Goff LA,
908 Hansen KD. 2022. Universal prediction of cell-cycle position using transfer learning. *Genome Biol*
909 **23**: 41.

910

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.** 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. **D.** 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. **E.** UMAP of U5-hNSCs with cells colored by the labels from O'Connor et al., 2021. **F-L.** The top 15 most important features for the ccAFv2 classifier were identified based on the mean change (Δ) in likelihood after permuting each feature's expression. A negative mean change in likelihood indicates that the feature increased the likelihood of predicting a ccAFv2 state.

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 hNSCs from a whole fetal brain at 9 weeks post conception (PCW 9 R1; Zeng et al., 2023) 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 was computed, based on the number of cell cycle states in the reference and predicted by the classifier, and were 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 hNSCs from a whole fetal brain at PCW 9 R1. **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. Again, average similarity to the reference was computed based on the number of cell cycle states in the reference and predicted by the classifier and were 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 tumor cells of BT322.

Figure 3. Application of ccAFv2 to *in vivo* hNSCs from fetal tissue 3 to 12 weeks post conception. **A.** Proportions of cell cycle states in U5-hNSCs which were grown *in vitro* and were derived from a human fetus at 8 PCW for both ccAFv2 and ccSeurat. **B.** Proportions of cell cycle states of hNSCs extracted from 3 to 12 PCW fetal tissue for both ccAFv2 and ccSeurat (Zeng et al, 2023). **C-F.** Distribution of cyclin expression in the *in vivo* hNSCs from a whole human fetal brain at PCW 9 R1 grouped by cell cycle phase. **G.** Mean expression of cyclins across the ccAFv2 cell cycle phases in cells from a whole human fetal brain at PCW 9 R1. Red

points denote the ccAFv2 cell cycle state with the highest average expression. Gene expression levels at each cell cycle state were compared to those in G1 cells using Student's t-test (**** indicates p ≤ 0.0001). **H.** Expression of ccAFv2 marker genes for each cell cycle state in hNSCs from a human whole fetal brain at PCW 9 R1. Important genes names are denoted in dark red. **I.** 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. **J.** Comparison of likelihood threshold application to random removal of the same number of cell predictions for *in vivo* hNSCs from a human whole fetal brain at PCW 9 R1. 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. AMI scores at each likelihood threshold were compared using Student's t-test (**** indicates p ≤ 0.0001). Rep. = biological replicate.

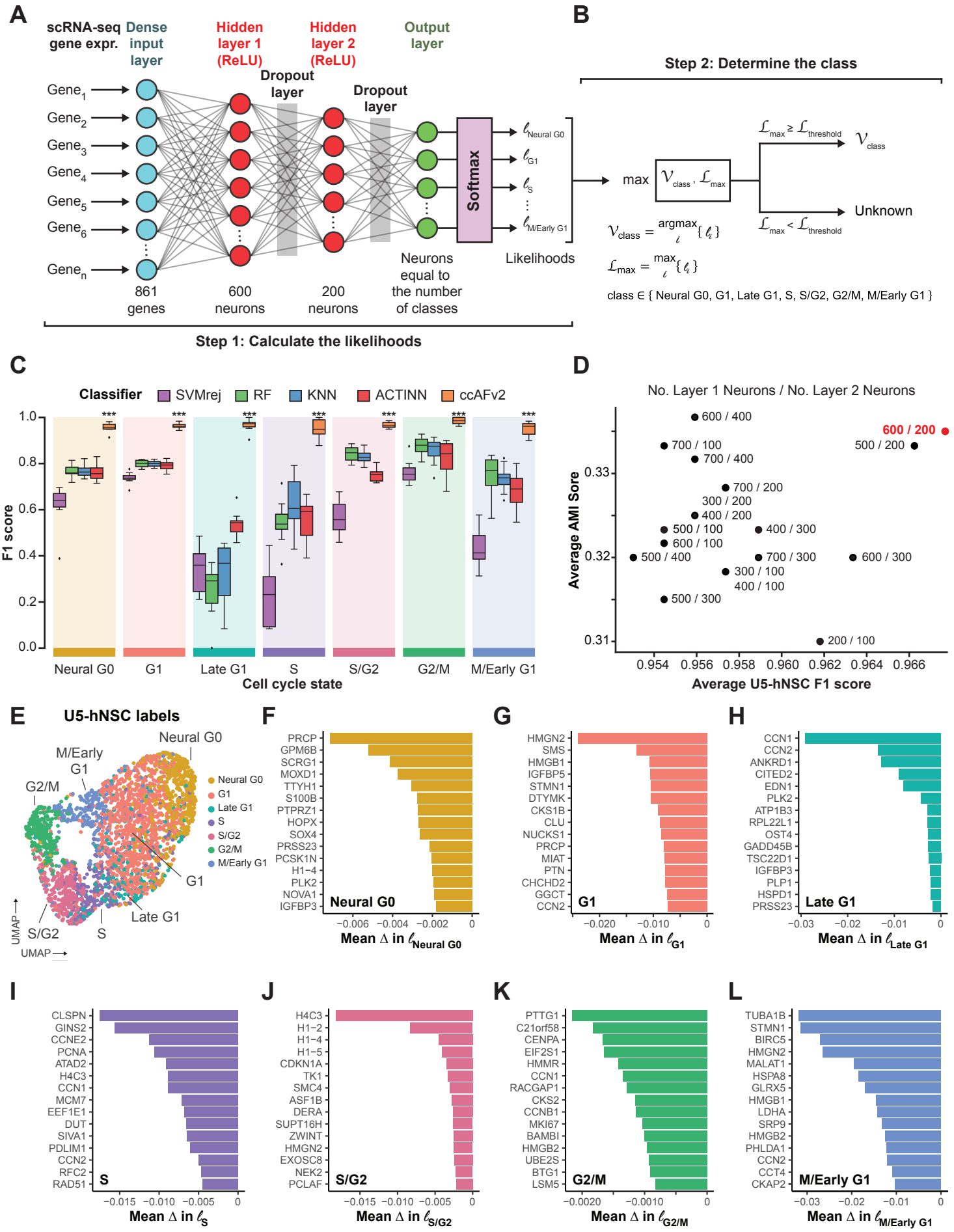
Figure 4. Application of ccAFv2 to the transcriptomes of 245,906 single cells from human fetuses aged 3 to 12 PCW. **A.** The 15 different cell types included in the analysis encompass all three germ layers. For each cell type the number of cells is given. **B.** Percentage of each ccAFv2 predicted state for each cell type. **C.** Genes found to be markers of Neural G0 cells in at least 7 cell types overlap with ccAFv2 Neural G0 marker genes (orange) and the top 15 most important features (dark red; from **Figure 1F**). **D.** For each cell type the number of marker genes discovered for Neural G0 cells is given. Next, a matrix was created to display, for each gene identified as a marker in at least seven cell types, which specific cell types it marked (dark grey). Above the matrix, bar plots show the number of cell types and the specific germ layers in which each gene served as a marker.

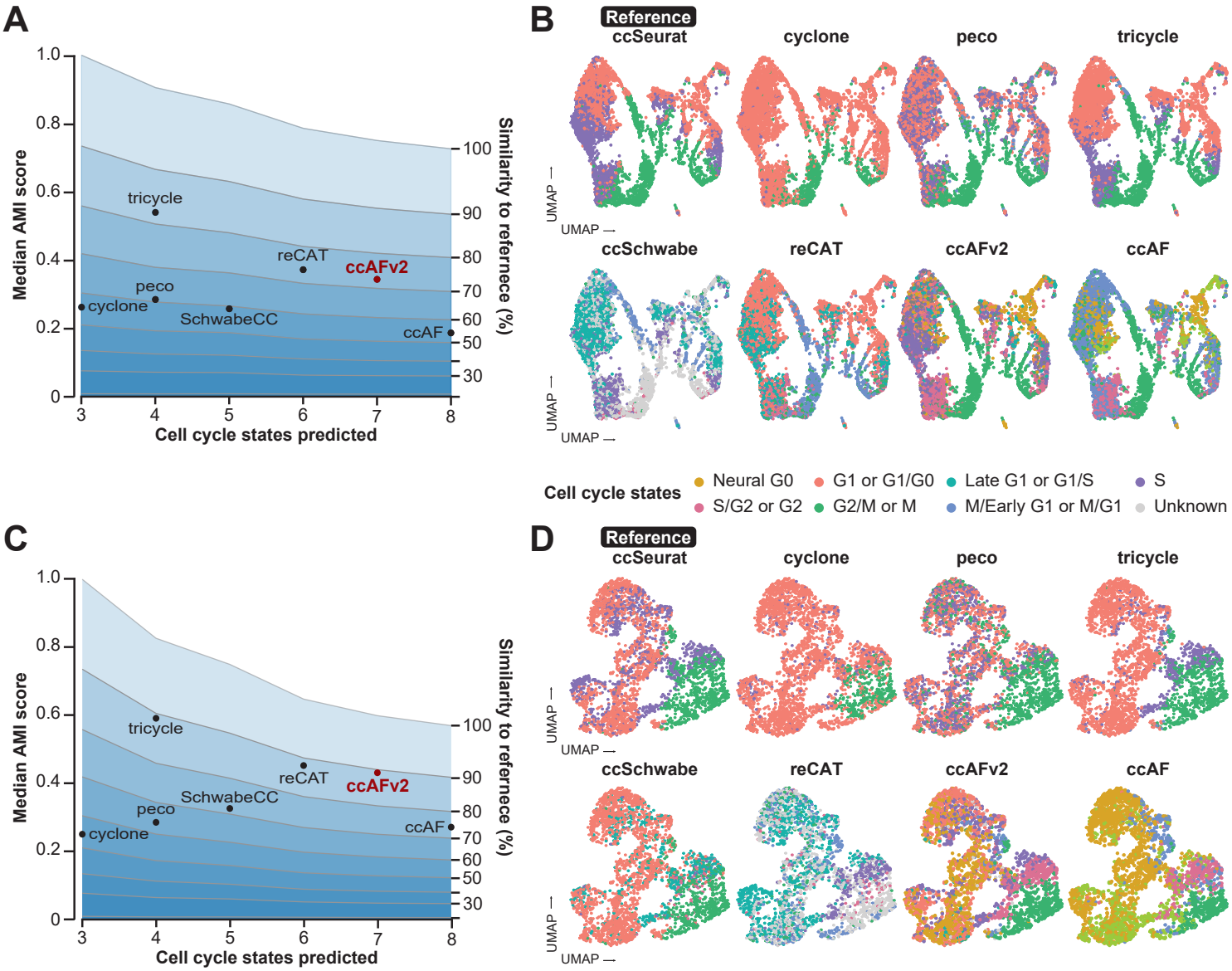
Figure 5. Application of ccAFv2 to LGGs with and without growth factors. Overlay of ccAFv2 predictions on Grade 2 astrocytoma (LGG275) with growth factors (**A**) and without growth factors (**B**). **C.** Summary of the proportions of cell cycle states with and without growth factors. Overlay of ccAFv2 predictions on Grade 3 oligodendroglioma (BT237) with growth factors (**D**) and without growth factors (**E**). **F**. Summary of the proportions of cell cycle states with and without growth factors.

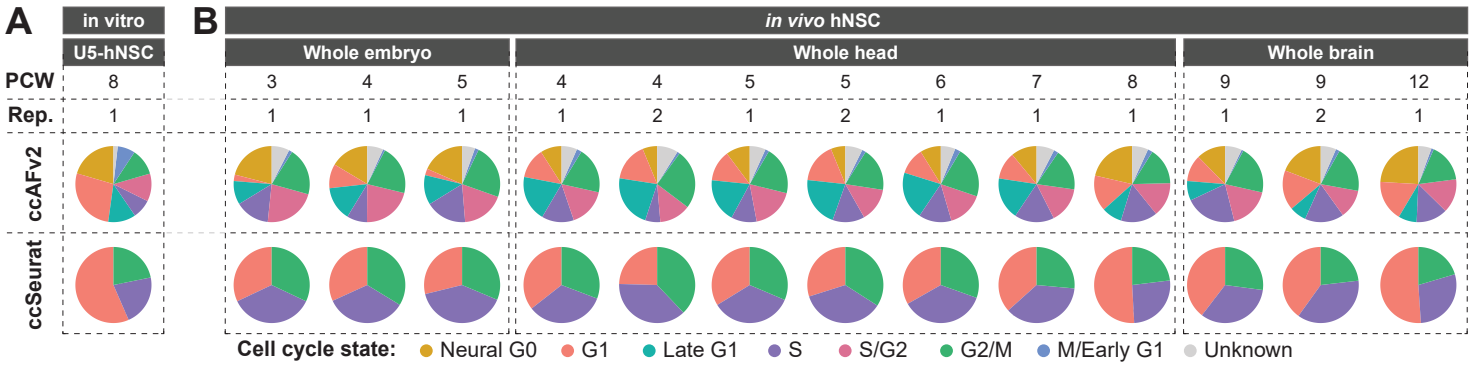
Figure 6. Application of ccAFv2 to single cells and nuclei from human and mice. **A.** Summary schematic of data ccAFv2 can be applied to and suggested data preparation. **B.** Proportion of cells assigned to each cell cycle state for scRNA-seq data from the developing human telencephalon (Nowakowski et al, 2017). **C.** Proportions of cell cycle states from scRNA-seq 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. **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 (C) and snRNA-seq (N) from spinal, cervical,

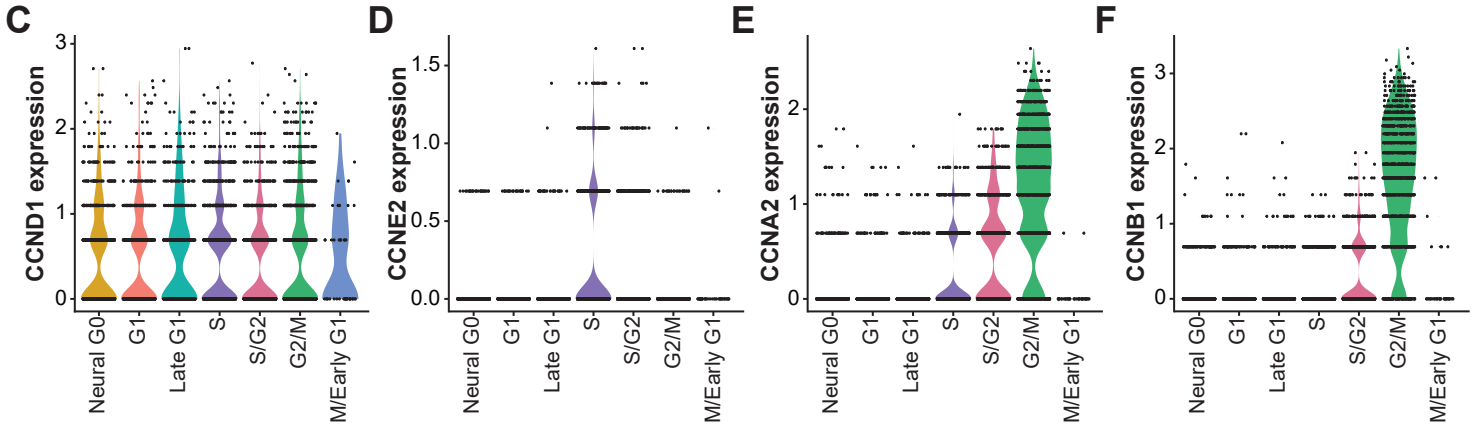
lumbar, and thoracic regions from the developing human spinal cord at 8, 10, 11, 20, and 23 PCW (Zhang et al, 2021).

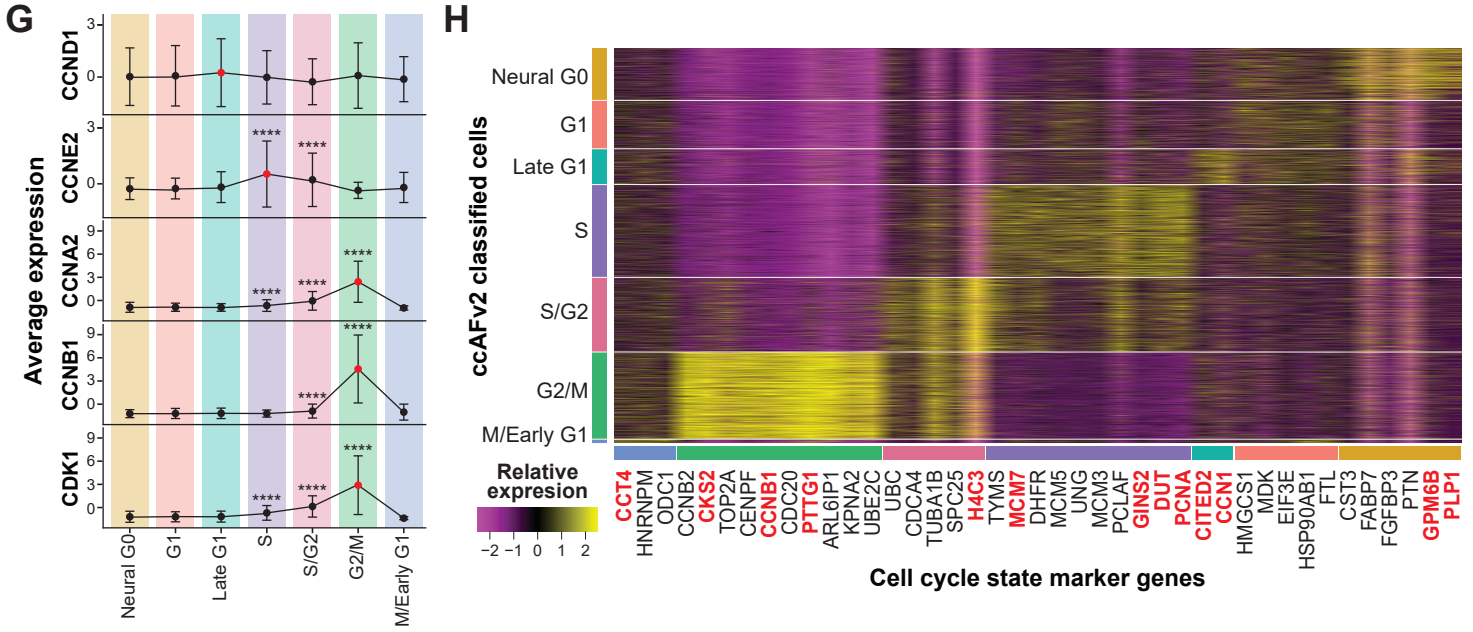
Figure 7. Application of ccAFv2 to spatial transcriptomics data from a male C57BL/6 mouse embryo at E15.5. **A.** H&E staining for the whole embryo. **B**. Spatial overlay of the predicted ccAFv2 states onto the whole embryo. **C.** Spatial expression of the cell cycle marker gene Mki67 for the whole embryo. The black boxes in panels **A** through **C** indicate the region of the developing cortex that was magnified in panels **D** through **Q**. The developmental regions of the developing cortex are denoted on the side: Dermis = developing skin, Skull = developing skull, CP/MZ = cortical plate and marginal zone, IZ = intermediate zone, SVZ = subventricular zone, VZ = ventricular zone. **D.** H&E staining for the developing embryo cortex. **E.** Spatial overlay of the predicted ccAFv2 states onto the developing embryo cortex. **F**-**I.** Expression of key marker genes describing the developmental regions in the developing embryo cortex. **J.** Spatial expression of the cell cycle marker gene Mki67 in the developing embryo cortex. **K**-**Q.** Likelihoods for each of the cell cycle states spatially overlayed onto the developing embryo cortex. The magnitude of the likelihood indicates the probability that a cell with that cell cycle state underlies that spot of the spatial array.

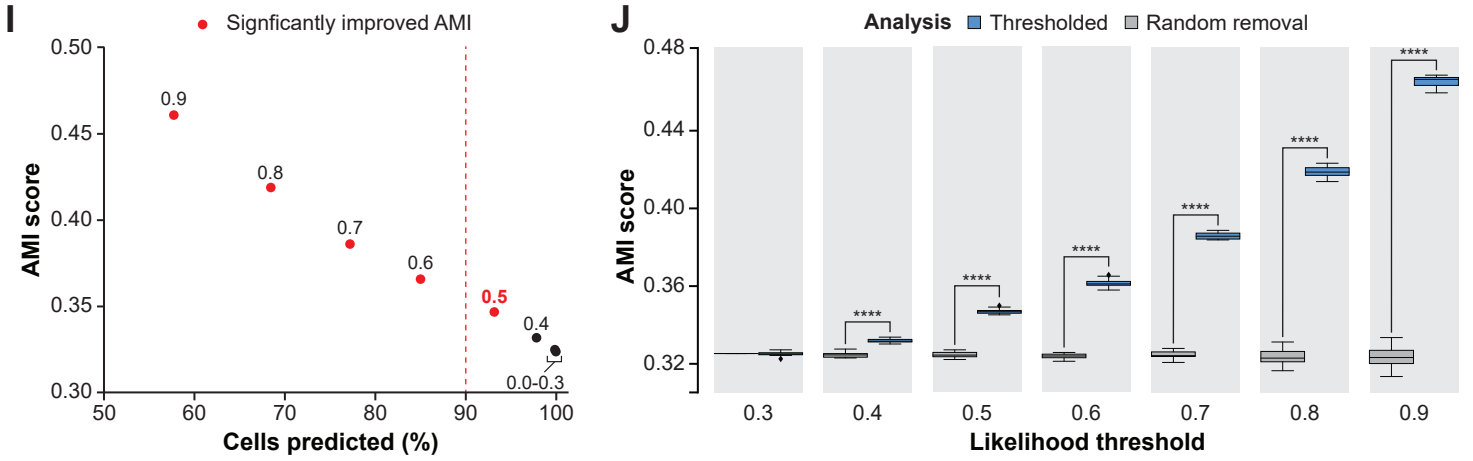

A
scRNA-seq gene expr.
Dense input layer
Hidden layer 1 (ReLU)
Dropout layer
Hidden layer 2 (ReLU)
Dropout layer
Output layer
Softmax
Likelihoods
861 genes
600 neurons
200 neurons
Neurons equal to the number of classes
Step 1: Calculate the likelihoods
B
Step 2: Determine the class
max
Unknown
class ∈ { Neural G0, G1, Late G1, S, S/G2, G2/M, M/Early G1 }
C
Classifier
SVMrej
RF
KNN
ACTINN
ccAFv2
F1 score
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Cell cycle state
D
No. Layer 1 Neurons / No. Layer 2 Neurons
Average AMI Score
Average U5-hNSC F1 score
600 / 200
E
U5-hNSC labels
Neural G0
M/Early G1
G2/M
G1
Late G1
S/G2
S
UMAP
F
Neural G0
Mean Δ in ℓ Neural G0
G
G1
Mean Δ in ℓ G1
H
Late G1
Mean Δ in ℓ Late G1
I
S
Mean Δ in ℓ S
J
S/G2
Mean Δ in ℓ S/G2
K
G2/M
Mean Δ in ℓ G2/M
L
M/Early G1
Mean Δ in ℓ M/Early G1

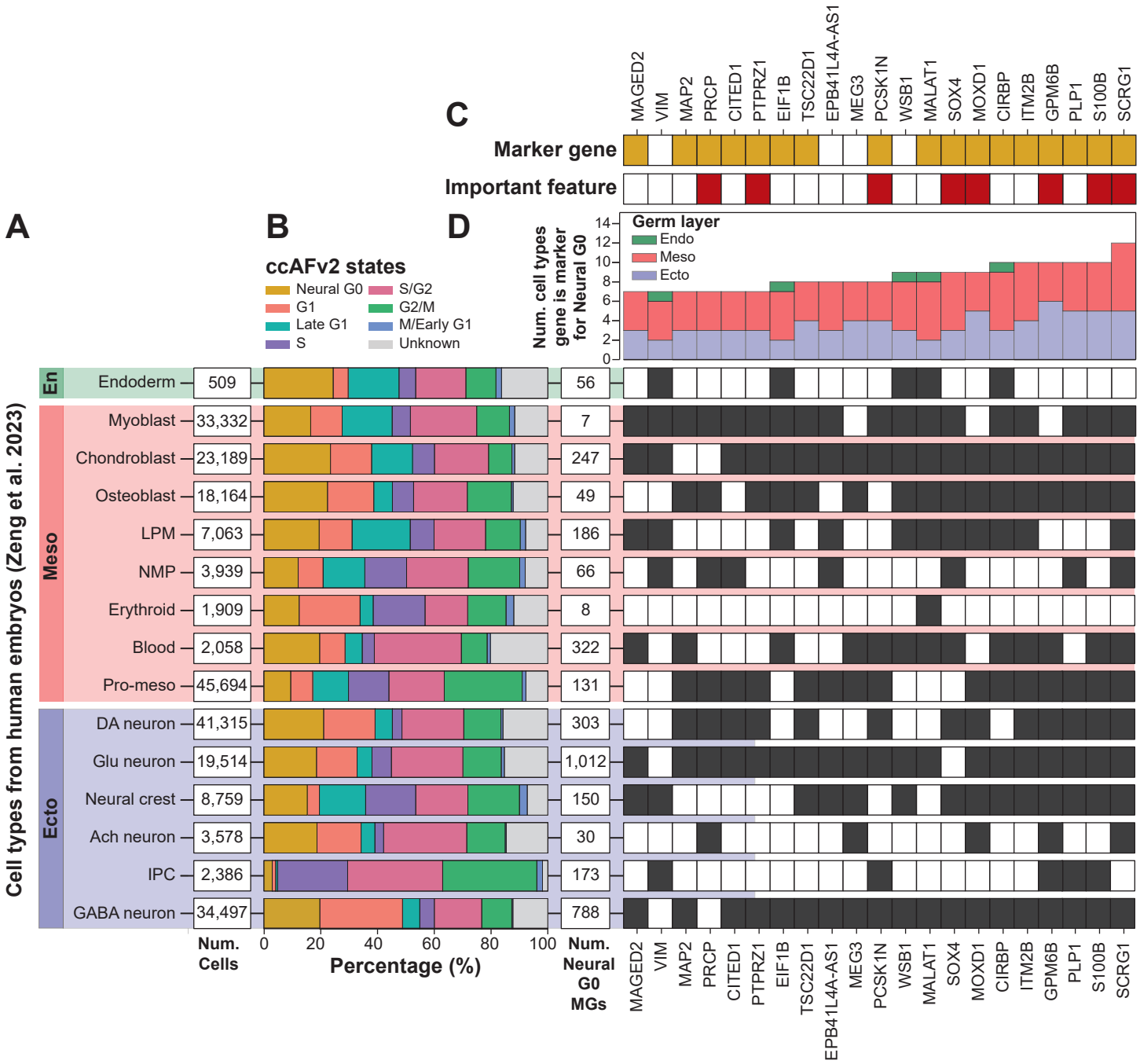

A
Median AMI score
Similarity to refernece (%)
Cell cycle states predicted
tricycle
reCAT
ccAFv2
peco
cyclone
SchwabeCC
ccAF
B
Reference
ccSeurat
cyclone
peco
tricycle
ccSchwabe
reCAT
ccAFv2
ccAF
UMAP
Cell cycle states
Neural G0
G1 or G1/G0
Late G1 or G1/S
S
S/G2 or G2
G2/M or M
M/Early G1 or M/G1
Unknown
C
D

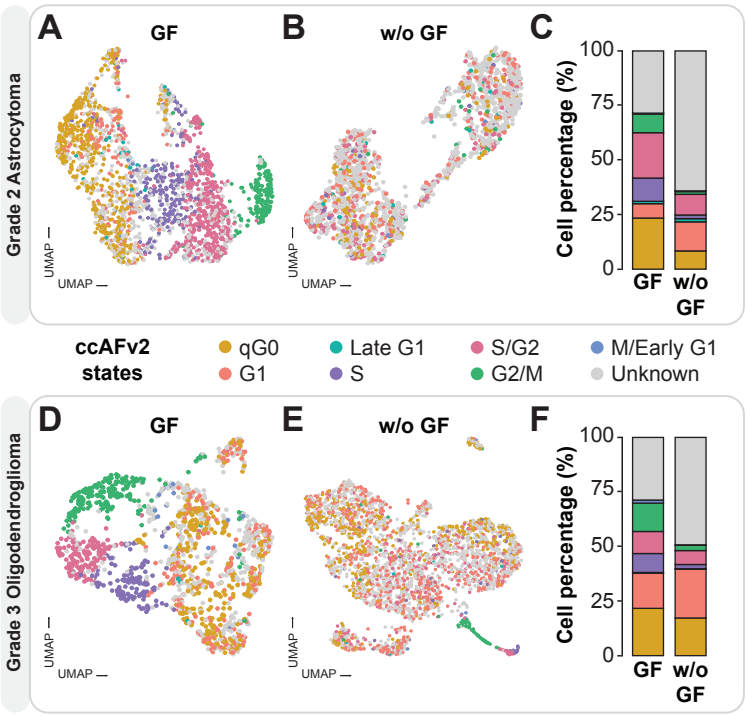

A
in vitro
U5-hNSC
B
in vivo hNSC
Whole embryo
Whole head
Whole brain
PCW
Rep.
ccAFv2
ccSeurat
Cell cycle state: Neural G0 G1 Late G1 S S/G2 G2/M M/Early G1 Unknown

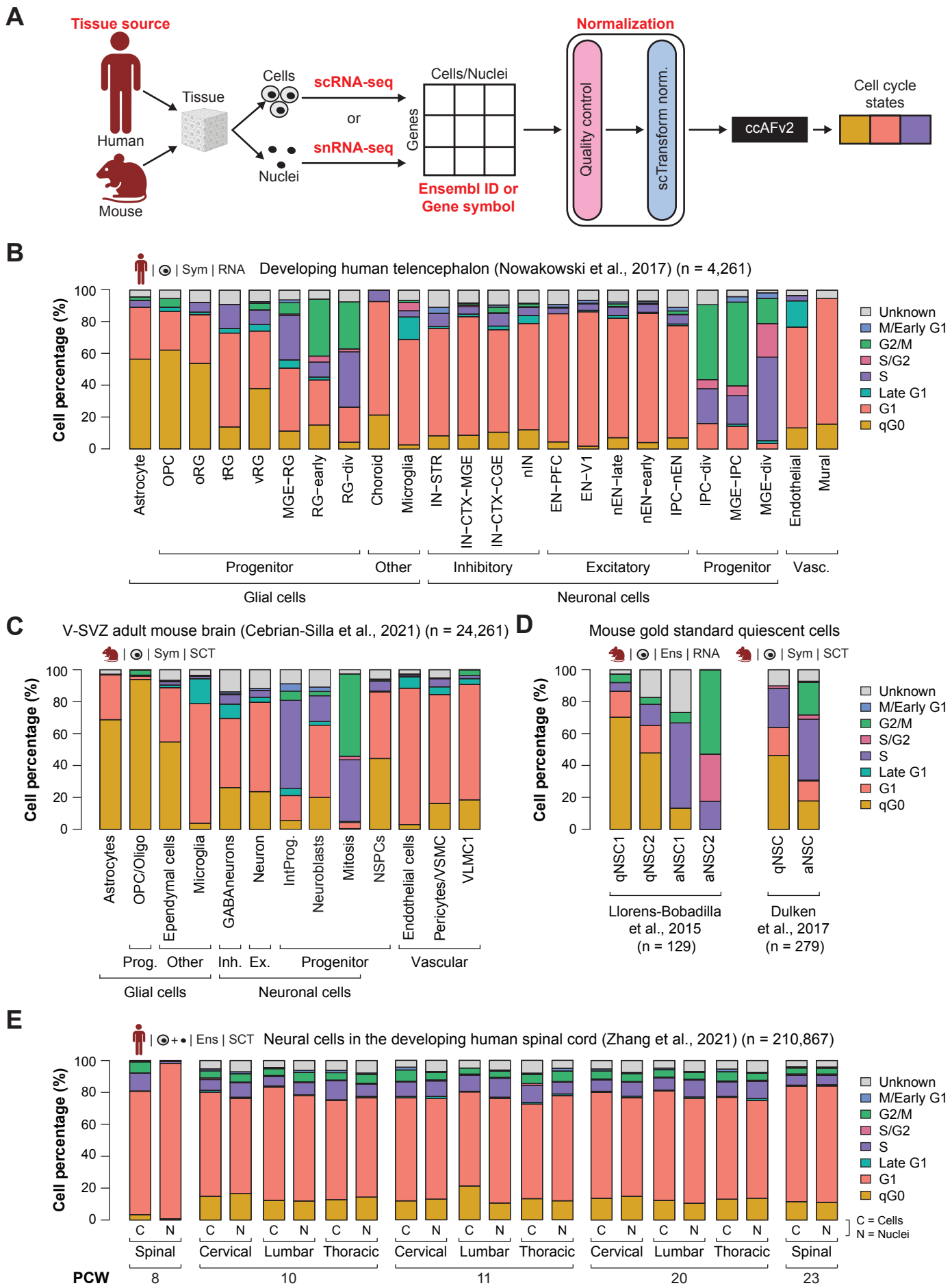

C
CCND1 expression
D
CCNE2 expression
E
CCNA2 expression
F
CCNB1 expression
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1

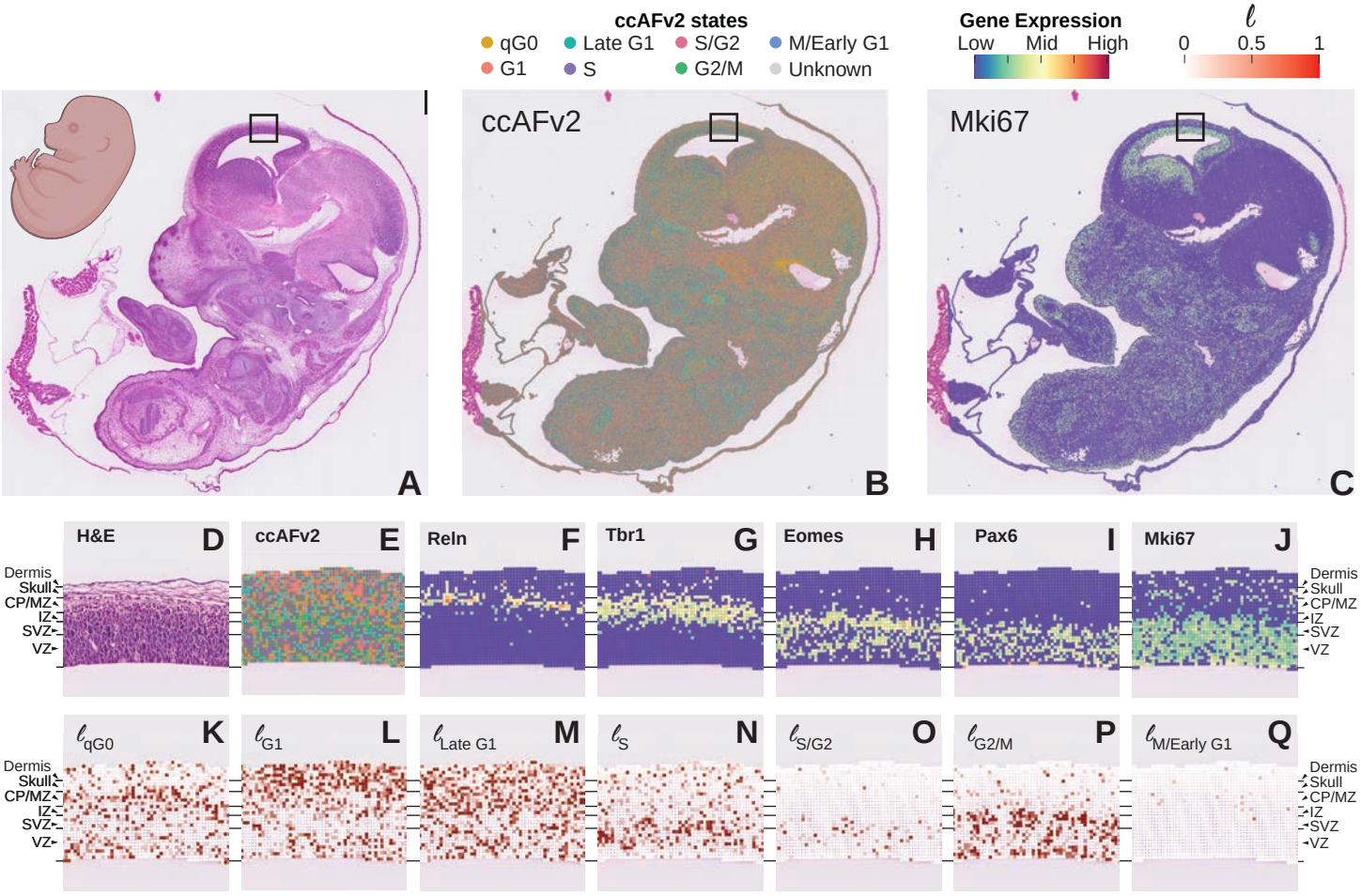

G
Average expression
CCND1
CCNE2
CCNA2
CCNB1
CDK1
H
ccAFv2 classified cells
Relative expression
CCT4 HNRNPM ODC1 CCNB2 CKS2 TOP2A CENPF CCNB1 CDC20 PTTG1 ARL6IP1 KPNA2 UBE2C UBC CDCA4 TUBA1B SPC25 H4C3 TYMS MCM7 DHFR MCM5 UNG MCM3 PCLAF GINS2 DUT PCNA CITED2 CCN1 HMGCS1 MDK EIF3E HSP90AB1 FTL CST3 FABP7 FGFBP3 PTN GPM6B PLP1
Cell cycle state marker genes


I
Signficantly improved AMI
AMI score
Cells predicted (%)
J
Analysis Thresholded Random removal
AMI score
Likelihood threshold


A
B
C
D
ccAFv2 states
Neural G0
G1
Late G1
S
S/G2
G2/M
M/Early G1
Unknown
Marker gene
Important feature
Germ layer
Endo
Meso
Ecto
Num. cell types gene is marker for Neural G0
Cell types from human embryos (Zeng et al. 2023)
En
Meso
Ecto
Endoderm 509 56
Myoblast 33,332 7
Chondroblast 23,189 247
Osteoblast 18,164 49
LPM 7,063 186
NMP 3,939 66
Erythroid 1,909 8
Blood 2,058 322
Pro-meso 45,694 131
DA neuron 41,315 303
Glu neuron 19,514 1,012
Neural crest 8,759 150
Ach neuron 3,578 30
IPC 2,386 173
GABA neuron 34,497 788
Num. Cells
Percentage (%)
Num. Neural G0 MGs
MAGED2
VIM
MAP2
PRCP
CITED1
PTPRZ1
EIF1B
TSC22D1
EPB41L4A-AS1
MEG3
PCSK1N
WSB1
MALAT1
SOX4
MOXD1
CIRBP
ITM2B
GPM6B
PLP1
S100B
SCRG1


Grade 2 Astrocytoma
A
GF
B
w/o GF
C
Cell percentage (%)
100
75
50
25
0
GF
w/o GF
UMAP
UMAP
ccAFv2 states
qG0
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
Cell cycle
states
B
Sym | RNA
Developing human telencephalon (Nowakowski et al., 2017) (n = 4,261)
Cell percentage (%)
Unknown
M/Early G1
G2/M
S/G2
S
Late G1
G1
qG0
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
V-SVZ adult mouse brain (Cebrian-Silla et al., 2021) (n = 24,261)
Sym | SCT
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
D
Mouse gold standard quiescent cells
Ens | RNA
Sym | SCT
qNSC1
qNSC2
aNSC1
aNSC2
qNSC
aNSC
Llorens-Bobadilla
et al., 2015
(n = 129)
Dulken
et al., 2017
(n = 279)
E
Ens | SCT
Neural cells in the developing human spinal cord (Zhang et al., 2021) (n = 210,867)
C = Cells
N = Nuclei
Spinal
Cervical
Lumbar
Thoracic
PCW
8
10
11
20
23


ccAFv2 states
qG0
Late G1
S/G2
M/Early G1
G1
S
G2/M
Unknown
Gene Expression
Low
Mid
High
ℓ
0
0.5
1
ccAFv2
Mki67
A
B
C
H&E
D
ccAFv2
E
Reln
F
Tbr1
G
Eomes
H
Pax6
I
Mki67
J
Dermis
Skull
CP/MZ
IZ
SVZ
VZ
ℓqG0
K
ℓG1
L
ℓLate G1
M
ℓS
N
ℓS/G2
O
ℓG2/M
P
ℓM/Early G1
Q