Designing signaling environments to steer transcriptional diversity in neural progenitor cell populations

Jong H. Park[1,2], Tiffany Tsou[1,2], Paul Rivaud[1,2], Matt Thomson[1,2,3], and Sisi Chen[1,2,3]

[1]Division of Biology and Biological Engineering, California Institute of Technology. Pasadena, California, 91125, USA.

[2]Beckman Center for Single-cell Profiling and Engineering. Pasadena, California, 91125, USA.

[3]correspondence to: sisichen@caltech.edu, mthomson@caltech.edu

Abstract

Stem cell populations within developing embryos are diverse, composed of many different subpopulations of cells with varying developmental potential. The structure of stem cell populations in cell culture remains poorly understood and presents a barrier to differentiating stem cells for therapeutic applications. In this paper we develop a framework for controlling the architecture of stem cell populations in cell culture using high-throughput single cell mRNA-seq and computational analysis. We find that the transcriptional diversity of neural stem cell populations collapses in cell culture. Cell populations are depleted of committed neuron progenitor cells and become dominated by a single pre-astrocytic cell population. By analyzing the response of neural stem cell populations to forty distinct signaling conditions, we demonstrate that signaling environments can restructure cell populations by modulating the relative abundance of pre-astrocyte and pre-neuron subpopulations according to a simple linear code. One specific combination of BMP4, EGF, and FGF2 ligands switches the default population balance such that 70% of cells correspond to the committed neurons. Our work demonstrates that single-cell RNA-seq can be applied to modulate the diversity of in vitro stem cell populations providing a new strategy for population-level stem cell control.

Highlights

- Natural progenitor diversity in the brain collapses during in vitro culture to a single progenitor type
- Loss of progenitor diversity alters fate potential of cells during differentiation
- Large scale single-cell signaling screen identifies signals that reshape population structure towards neuronal cell types
- Signals regulate population structure according to a simple log-linear model

Introduction

Stem cells remain an exciting resource for regenerative medicine and tissue replacement. Work over many decades has shown that a wide range of stem cell populations, including embryonic stem cells and neural stem cells, can be extracted from an embryo and then expanded and differentiated in cell culture using applied signaling molecules [1]. Expansion of cells in culture provides a strategy for generating a wide range of functional cell types from a small initial pool of precursors that could be used to replace or repair damaged tissue in the brain, heart, and other organs that do not have intrinsic regenerative capacity.

However, despite the promise, the control of stem cell populations in cell culture, outside of the embryonic environment, has proved to be difficult and limited in many cases. Practically, it has been challenging to generate homogenous populations of functionally differentiated cells from stem cell precursors [2]. Although, for example, we can isolate and grow cells from the developing brain that express stem cell marker genes, the extracted cells often fail to differentiate into functional neurons both in vitro [3] or upon transplantation [4] limiting the use of these cells for treating neurodegenerative diseases. In other cases, differentiation occurs in cell culture but is highly variable, so that the resulting cell populations contain many different types of cells [2]. While the nature and source of variability is poorly understood, variability has led to practical bottlenecks in research and costly failures in clinical trials [5, 4, 6].

The practical challenges that complicate stem cell therapies raise fundamental questions about the generation and maintenance of stem cell populations in the embryo and the subsequent dynamics of cell populations within cell culture signaling environments. The combinatorial signaling environment within mammalian embryos generates a broad spectrum of different stem cell populations [7, 8, 9] that each generate specific differentiated cell-types. For instance, the mammalian brain develops through the coordinated proliferation and differentiation of multiple subpopulations of stem cells, which range in commitment and propensity for distinct differentiated neural cell fates [7]. Clonal tracking of neural stem cells has revealed a clonal bias in astrocytic versus neuronal fate[10], and suggests that propensity for distinct layers of the cortex is based on combinatorial patterns of transcription factor expression[11, 12]. Despite the importance of stem cell population diversity in the developing embryo, we do not know how natural progenitor diversity changes within cell culture signaling environments or how to shape population diversity to amplify stem cell types of practical interest.

In this paper, we apply single cell mRNA-seq to demonstrate a new approach to programming stem cell fate by controlling the balance of stem cell subpopulations primed for different lineages. By comparing stem cell populations from the developing mouse brain to neural stem cells growing in culture, we discover that the natural population diversity of neural stem cells collapses in culture. In culture, the critical pre-neuron progenitor cell subpopulations that generate neurons in the brain are lost. By exposing neural stem cells to forty different signaling environments, we show

that signaling environments can be designed using a simple log-linear model to alter the balance of pre-neuron and pre-astrocyte cell populations enabling amplification of the committed neurons that are lost under canonical conditions. Our work points to population structure and dynamics as playing an important and previously hidden role in impacting the outcome of stem cell differentiation. Broadly, we demonstrate an experimental and computational strategy for designing signaling environments that can restructure cell populations to overcome challenges implicit in stem cell therapies.

Results

Neural stem cell populations experience collapse in transcriptional diversity in cell culture

We wanted to understand how the diversity of neural stem cell populations changes when cells are extracted from the embryonic brain and placed in cell culture signaling environments. Therefore, we applied single cell mRNA-seq to create a map of the stem cell populations contained in the developing mouse brain, and then compared the underlying transcriptional diversity to neural stem cells grown in cell culture. Specifically, we profiled 23,438 cells at three time points of embryonic development (embryonic day 18 subventricular zone, postnatal day 4, and Adult Cortex) (Fig1a) and compared these cells to populations of subventricular zone (SVZ) stem cells dissociated from the ventricular zone of the E18 mouse embryonic cortex and cultured with EGF and FGF2 proteins. To analyze the data, we used a probabilistic modeling platform, PopAlign, [13] to build, align, and compare Gaussian mixture models (GMMs) of the embryonic and cell culture populations (Fig 1b). Cell type markers for mature differentiated cell types (Fig 1a - inset) are used to type subpopulations in the adult cortex (Fig 1b), which provide a reference for assessing differentiation potential. For each sample, the resulting PopAlign models were low error both qualitatively (SI Figure 1) and quantitatively (SI Figure 2) and so provided a computational substrate for performing quantitative comparison between the embryonic and in vitro cells.

Our computational analysis revealed that embryonic day 18 (E18) brain contains a broad range of neuronal and astrocyte progenitor cells but that population diversity collapses in cell culture, so that cultured cell populations become dominated by a single pre-astrocyte cell state. Specifically, in the E18 SVZ, we identified three major categories of cells as the centroids of the GMMs identified by PopAlign (Fig 1b, 1c), committed neurons, pre-neuronal stem cells, and pre-astrocytic stem cells, as 71.4%, 13.8%, and 12.4% of the population respectively. Committed neurons (CNs), the dominant cell populations, expressed the canonical neuron markers Tubb3 and Map2 (Fig. 1d). The pre-neuronal progenitors (NPs) expressed stem cell markers Sox2, Pax6, and the neuronal transcription factor Neurog2, while the pre-astrocyte progenitors (APs) expressed glial transcription factor Sox9[14] in addition to Sox2 and Pax6 (Fig. 1d).

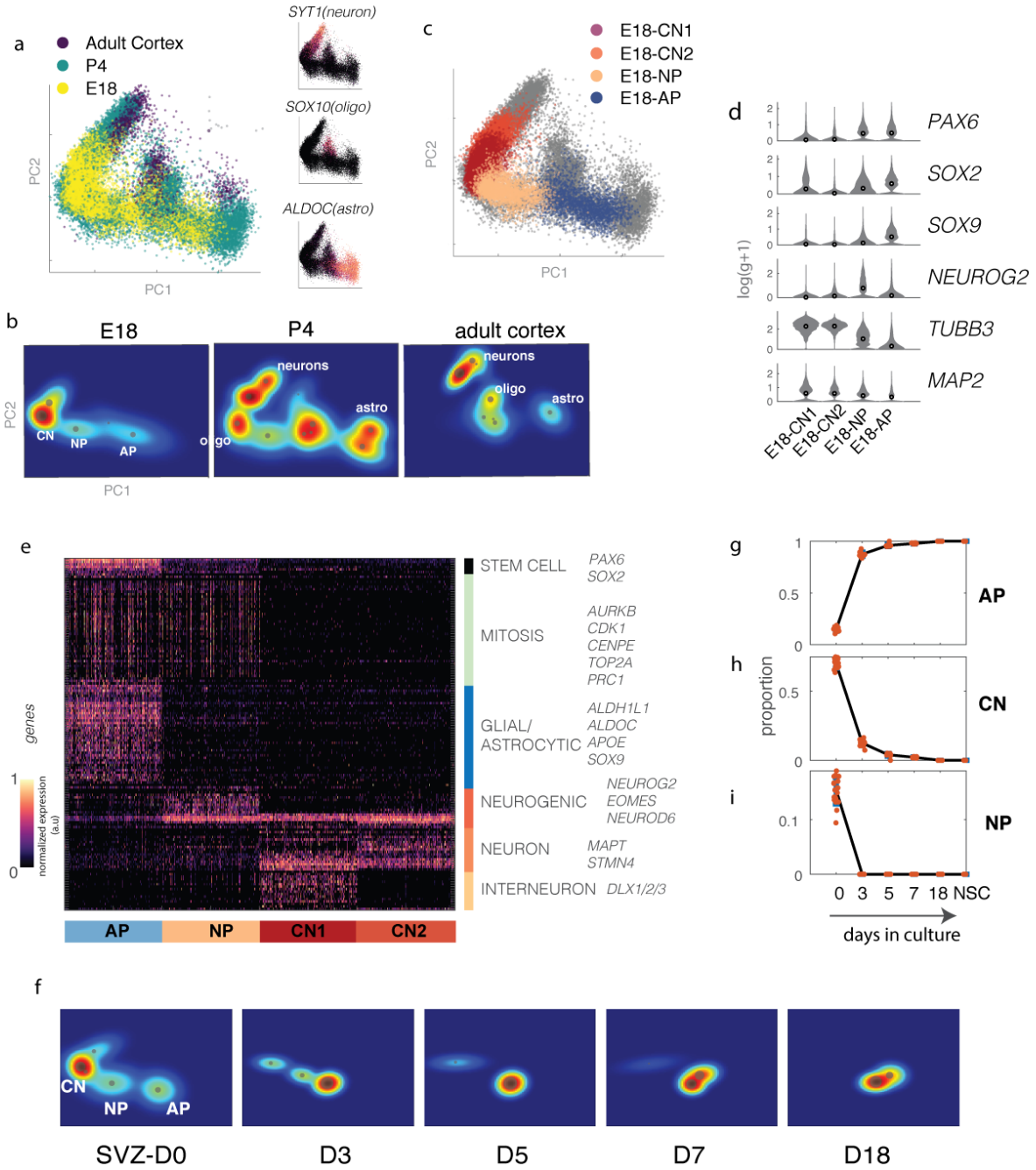

a
Adult Cortex
P4
E18
PC2
PC1
SYT1(neuron)
SOX10(oligo)
ALDOC(astro)
c
E18-CN1
E18-CN2
E18-NP
E18-AP
PC2
PC1
d
PAX6
SOX2
SOX9
NEUROG2
TUBB3
MAP2
log(g+1)
E18-CN1
E18-CN2
E18-NP
E18-AP
b
E18
P4
adult cortex
neurons
astro
oligo
CN
NP
AP
PC2
PC1
e
genes
normalized expression (a.u)
STEM CELL
PAX6
SOX2
MITOSIS
AURKB
CDK1
CENPE
TOP2A
PRC1
GLIAL/ ASTROCYTIC
ALDH1L1
ALDOC
APOE
SOX9
NEUROGENIC
NEUROG2
EOMES
NEUROD6
NEURON
MAPT
STMN4
INTERNEURON
DLX1/2/3
AP
NP
CN1
CN2
g
h
i
proportion
AP
CN
NP
0 3 5 7 18 NSC
days in culture
f
CN
NP
AP
SVZ-D0
D3
D5
D7
D18

Figure 1: Transcriptional diversity of embryonic neural stem cell populations collapses in vitro. (a) PCA scatter plots of cells (n=23,438) at three developmental timepoints (E18, P4, and Adult Cortex). Inset graphs: mature cell type gene markers indicate terminally differentiated neurons, astrocytes, and oligodendrocytes. (b) 2D PCA renderings of GMM population model built for each timepoint. Each GMM describes a probability distribution of cell populations within a 10D PCA space. The mean gene expression state (Gaussian mean: μ) is displayed as a dot scaled by population abundance. (c) Scatter plot of E18 cells colored by subpopulation class shown in (b). (d) Violin plots of marker genes across E18 subpopulations, with mean denoted by a circle. (e) Gene expression heatmap of the four identified cell subpopulations from E18 brain. Each cell type is composed of expression of one of six indicated gene expression programs. AP expresses stem cell genes, mitotic genes, as well as a glial/astrocytic gene block. NP expresses stem cell genes, mitotic genes, as well as a neurogenic gene program. CNs express neuron and interneuron programs. (f) Renderings of population models in a timecourse of in vitro culture showing diversity collapse over 18 days to AP population. (g-i) Quantification of proportions across timecourse for each major subpopulation type (g) AP, (h) CN (both types are combined) and (i) NP. Proportion values are bootstrapped over multiple model building runs to average out potential variability in model fitting. For each timepoint, we built 20 distinct models (red points), and aligned subpopulations to reference populations in a reference SVZ-D0 model to associate abundance values.

Each identified cell population also expressed blocks of lineage specific genes, supporting the idea that the identified cell populations could exhibit distinct differentiation potential. Both the pre-neuronal and pre-astrocyte progenitor cells expressed a base gene expression program of stem cell-associated genes including Glast(SLC1A3), Vim, Fabp7 (Blbp), and Dbi. These genes are known to be glial and reflect the radial-glial origin of neural stem cells in the brain [15]. However, in addition to this base stem cell program, pre-neuronal progenitors (NPs) specifically expressed a group of 13 neurogenic genes including neuron-specific TFs such as Tbrs (Eomes) Neurod6, and Neurog2. (Fig. 1e) . Pre-astrocytic progenitors (APs) expressed a large block of 45 genes associated with mature astrocytes, including Aldh1l1, a mature astrocytic marker [16], and Glul (GS), an astrocyte-specific glutamine synthetase. While all progenitor cells expressed the core radial glial program, the astrocyte and neurogenic programs are mutually exclusive within the progenitor populations suggesting that they might repress each other and foreshadow downstream fate. Committed neurons (CN) expressed neuron specific gene expression programs, but not the stem cell program or the mitotic gene expression program (Fig. 1e).

When placed into the cell culture signaling environment with EGF and FGF2, the diversity of the progenitor population dramatically shifted undergoing a qualitative loss of the neural progenitor (NP) and committed neuron (CN) subpopulations within five days of culture (Fig. 1f). By day five of culture, the cell population had become dominated by the pre-astrocytic progenitor (AP) cells (96% of cells at day 5 but 18% at day 0) (Fig. 1g) and the CN cells have been lost (4% day 5 vs 66% day 0) (Fig 1h). Further, pre-neuronal progenitors (NPs) diminish from 16% to 0% at day 5 (Fig 1i). In this way, although the developing brain at E18 contains a diverse population of both pre-neuronal and pre-astrocytic progenitors, only the pre-astrocytic progenitors are expanded in EGF/FGF2 cell culture.

Decreased population diversity is associated with limited differentiation potential

The collapsed pre-astrocytic populations derived from the SVZ and grown in EGF and FGF2 for five days (Fig 2a) had limited differentiation potential. Upon differentiation in serum-containing media, the cells generated uniform cultures of astrocytes with no neurons generating no density in the neuronal branch of a PCA fate map (Fig 2b). The fate map was constructed using populations at postnatal day 4 (P4), which temporally corresponds to 5 days of in vitro differentiation. Differentiated populations quantitatively align most closely to the P4 astrocytic population, and produce matching gene expression signatures (Fig 2c). However, under the same conditions, cells extracted from SVZ tissue (Fig 2d) and immediately differentiated could generate mixed cultures of neurons and astrocytes based on the fate map (Fig 2e), and gene expression profiles (Fig. 2f). Although we cannot definitively conclude that the AP cells lack neuronal potential under all possible contexts, our results show that they do not generate neurons in conditions that support neuronal differentiation from a broader range of progenitor cell types from the brain. Thus these findings suggest that the collapse in progenitor cell diversity during cell culture limits the capacity of the population to generate neurons, and imply that increasing the progenitor diversity could expand the neurogenic potential of stem cells in vitro.

Combinatorial signaling environments modulate balance between pre-astrocytic progenitors and committed neurons

While EGF and FGF2 at fixed concentration represent just a single signaling condition, the embryonic brain contains combinatorial gradients of signaling molecules. We hypothesized that EGF and FGF2 likely select for the pre-astrocytic progenitor (AP) cells, thus restricting the differentiation potential of the cell population. Therefore, we asked whether we could stabilize a more diverse progenitor population by employing additional signaling factors. We performed a signaling screen where we exposed the SVZ stem cell populations across a range of different signaling conditions and analyzed the impact of signaling combinations on population structure using single-cell transcriptional profiling.

We selected signaling cues from the EGF, FGF, Wnt, PDGF, and BMP families whose receptors are all present in the developing brain (SI Figure 3) and are known to be associated with diverse functions such as proliferation (EGF,FGF2 [17], Wnt pathway activation via small molecule agonist CHIR, PDGF-AA [18]), neuronal differentiation, (WNT, [19] PDGF), astrocytic differentiation(BMP4 [20]), and the maintenance of quiescence [BMP4 [21]]. We also included two immune signaling cytokines, GM-CSF and IFN-gamma as outgroups that are not normally active during neural development, but whose receptors are present in our cells (SI Figure 3).

We combined all signals with EGF/FGF2 in varying doses ranging from 20ng/mL, a high concen-

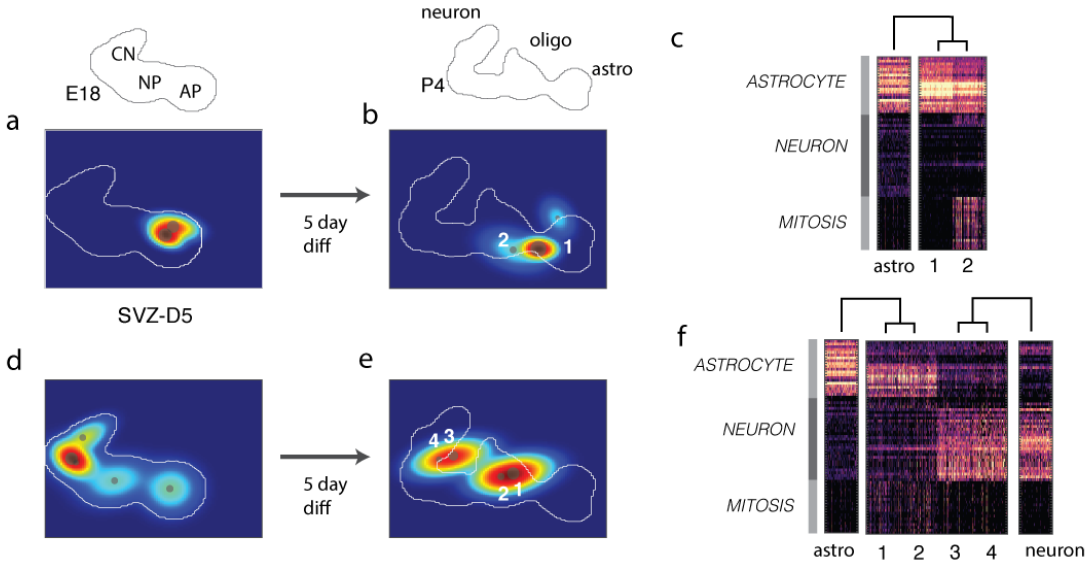


Figure 2: Decreased progenitor diversity is associated with limited differentiation potential. (a) PCA rendering of cell populations extracted from the SVZ and cultured for 5 days in EGF/FGF and (b) after differentiating in serum containing media for 5 days. Above: cartoon maps showing relative positions of populations from corresponding stages of natural brain development development (E18 and P4). (c) Gene expression heatmaps of population centroids from differentiated cells, compared to aligned populations from the in vivo brain at P4. Astrocytic, mitotic, and neuronal gene expression programs are indicated in the key. (d) Model renderings of cells dissociated directly from the SVZ and (e) their differentiated progeny after 5 days of in vitro differentiation. (i) Gene expression heat-map comparing the centroids of populations from (e) to their aligned in vivo populations.

tration that is the standard culture condition, down to 0.8 ng/mL, a low concentration that we found was necessary to maintain cell survival. Signals were added immediately to cells after SVZ dissociation. Cells were profiled after 5 days of exposure, and the Clicktags multiplexing platform [22] applied for sample labeling. We profiled an average of 500 cells per signaling condition leading to a data set with 18,700 cells, and applied PopAlign to dissect and analyze population diversity and to compare identified cell populations to those in the E18 brain.

We found that signaling conditions could shift the balance between pre-neuronal and pre-astrocyte cell populations. Across all signaling conditions, we identified three major subpopulations of progenitor cells that aligned to the three cell populations we previously identified (Figure 1) in the E18 mouse brain: committed neuroblasts (CN), pre-astrocytic progenitors (AP) and pre-neuronal progenitors (NP) (Fig 3a, SI Fig 4). Individual signals altered the relative abundances of the pre-neuronal and pre-astrocyte subpopulations, with the subset of conditions including BMP showing particularly strong impact. Specifically, BMP4 shifted the population structure towards more pre-neuronal cells (CN - pink), which could be observed qualitatively in terms of increasing density in the CN population (Fig 3b) and quantitatively in a phase-diagram (Fig 3c). Increasing the concentration of EGF/FGF2 had the opposite effect, shifting the population away from CN, towards pre-astrocytic progenitors (AP) (Fig 3b, 3d). CHIR boosted the proportions of pre-astrocytic progenitors (AP) (SI Fig 5a), while having no effect on pre-neuronal CNs (SI Figure 5b). Other signals such as FGF9, PDGF-AA, GM-CSF, and IFNG predominantly reduced the proportions of pre-astrocytic progenitors (AP) (SI Fig 5c) and boosted pre-neuronal CNs (SI Figure 5c).

Simple log-linear model enables design of signaling environments to achieve target population structure

The combinatorial impact of the signaling environment on the relative abundance of the three underlying progenitor cell populations could be captured by a simple log-linear model. To capture the relationship between signaling inputs and population size, we consider a simple model where the relative growth rates of specific cell populations respond independently to signaling inputs (Fig 4a). In this model, the log of the relative proportion of population (e.g. CN/NP) depends on a linear combination of signaling molecule concentrations (Fig 4b). The coefficients A, can be interpreted as the difference in relative growth rate between population i versus reference population k for each signal j. Critically, Aij can be encode both positive and negative values and can reveal antagonistic relationships between signals (Fig 4c). For instance, BMP4 and EGF compete to modulate the relative population sizes of both CN and AP cells (relative to the NP population) (Fig 4c). Model parameters reveal that each additional unit of BMP4 increases the relative proportion of CNs by 4.4x , while EGF/FGF reverses this trend, scaling down the relative CN proportion by 0.4x (Fig 4c).

The simple log-linear model was sufficient to recreate observed population abundances across all experiments. The accuracy of the model is both qualitative and quantitative; Model-generated pre-

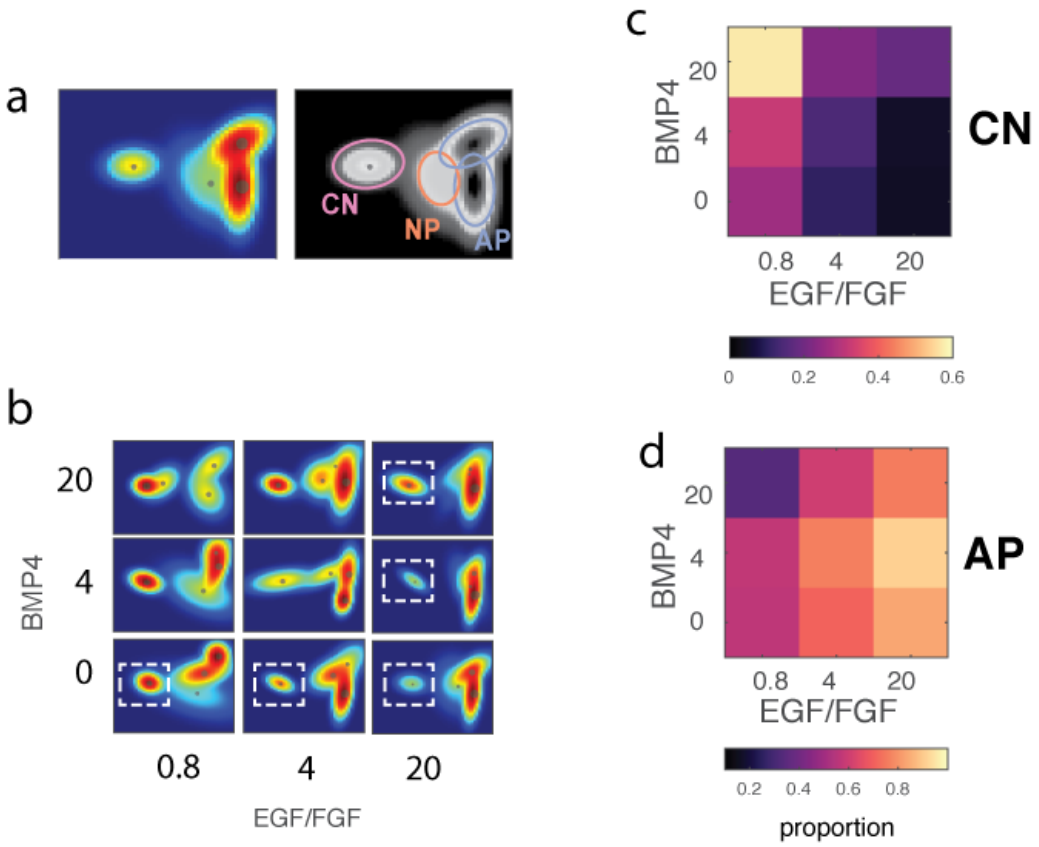


Figure 3: Signals reshape global population diversity. (a) 2D PCA rendering of common model built using single-cell data sampled across 40-condition signaling screen Identified subpopulations align to the main subtypes from the in vivo SVZ-D0 population: AP, CN, NP. (b) Renderings show population shifts in response to varying EGF/FGF and BMP4. The density of the CN population, highlighted in dashed white box, increases with increasing BMP4, and decreases with increasing EGF/FGF. (c) Heatmap of CN proportions with varying concentrations of BMP4 and EGF/FGF. Increasing BMP4 increases the CN proportion, while increasing EGF diminishes it. (d) Heatmap of AP proportions, showing AP proportions increasing with EGF/FGF, and decreasing with BMP4.

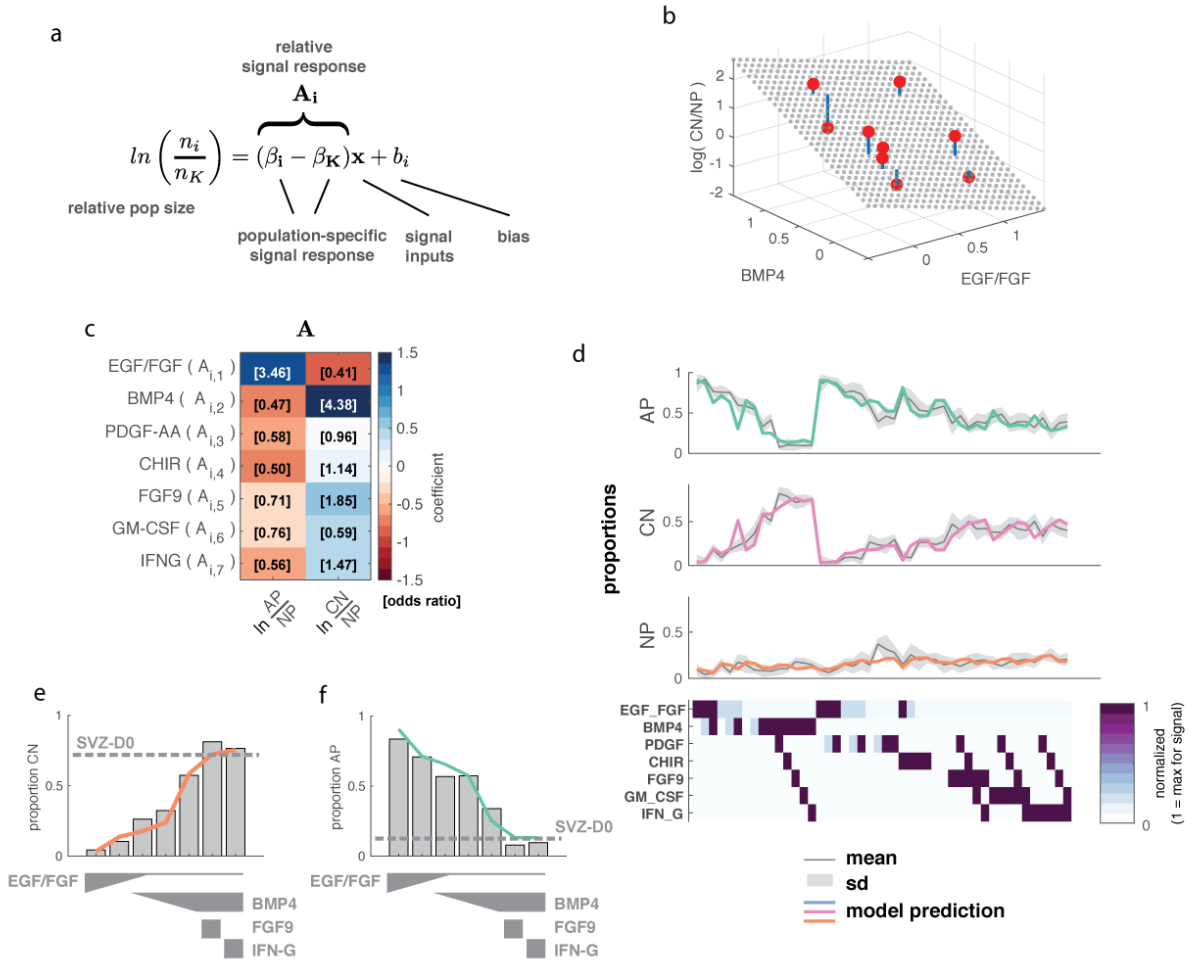


Figure 4: Population structure responds to signaling input according to a simple log-linear model. (a) Model equation relating relative population size (n_i/n_k) to a linear function of signaling inputs, x. Coefficients Ai can be considered the relative growth rate of population i to population k. The bias encodes the natural difference in growth rates in the absence of signal (x=0). (b) The relative proportion of CN to AP populations (logged) responds linearly to signaling inputs. Grey dots: model output, red dots: observed proportions, blue line: error between observed data and prediction. (c) Coefficient matrix A linking relative population ratio (columns) to signals (rows). The exponent of the coefficient ($e^(A_{i,j})$) is denoted in brackets and reports how much the ratio n_i/n_k goes up for each unit increase in signal. (d) Observed and predicted (colored) subpopulation proportions reported for each experimental sample (column). The standard deviation is calculated from bootstrapping samples from experimental data (n=100 cells each, iterations=20). Heatmap represents signaling concentration used within the column. (e-f) Model (colored lines) predicts combinations of signals that push CN and AP proportions towards target levels, such as that seen in original tissue (dotted line - SVZ-D0). Note: Signal concentrations across all panels in this figure is normalized so that 1 unit equals the maximum concentration (all 20 ng/mL except CHIR, whose maximum concentration is 3 uM).

dictions match patterns seen in the measured subpopulation abundances (Fig 4d). Additionally, the RMS error between the model-predicted proportions and the observed proportions is small, averaging only 5.58% of the observed value. The accuracy of the simple log-linear model suggests that additive impact of signaling molecules on cell-type specific proliferation rates might be sufficient to explain population size changes observed across the complex combinatorial signaling conditions considered in our screen (Methods).

Using the idea of a linear signaling code, we were able to design combinatorial signaling conditions to achieve desired population structures. For instance, although none of the signaling combination we tested could completely recreate the transcriptional diversity of the SVZ tissue, some conditions could approximate it by suppressing unwanted pre-astrocytic populations and boosting desired pre-neuronal populations. The original SVZ tissue was composed of ¿ 70% CNs (Fig 4e-dotted line). By reducing EGF/FGF2 alone, we boosted the population of CNs to 26% of the population (Fig 4e). Adding BMP4 boosts the population percentage to 56%. Finally, adding another CN-boosting and AP-reducing signal such as FGF9 or IFN-G can boost the CNs again up to ¿ 80%, while also reducing the number of pre-astrocytic progenitors to 4-6% (Fig 4g). Globally, the conditions satisfying these constraints (low EGF/FGF2, high BMP4, and an additional AP-reducing signal) produced populations of cells that most closely matched the original SVZ tissue, based on a statistical metric which calculates the divergence of the two models (SI Fig 5 - Ranking).

Broadly, our results demonstrate that the structure of neural progenitor cell populations can be steered through the modulation of a relatively simple combinatorial signaling environment. In addition to altering the broad population structure, we also note that the signaling environment had an impact on the detailed transcriptional state of specific cell populations. In addition to altering population structure by boosting the CN population, BMP4 had the biggest impact on the transcriptional state of both CN (Fig 5a) and AP (Fig 5b) cells, resulting in the most differentially expressed genes. BMP4 altered the gene expression program of the AP cells inducing a neuronal gene expression program in these astrocyte progenitor cells. The gene expression program contained genes specific to neuron populations in different layers of the cortex [23] (Fig 5c), and are not induced by other signals. This result suggests that BMP4 might be able to induce the neuronal target gene program in the background of the pre-astrocytic stem cell-type.

Discussion

The embryo naturally regulates distinct pools of multipotent stem and progenitor cells to generate a diversity of mature cell types in the correct proportions. Changes in the population diversity of lineage-primed progenitors can determine the relative abundance of neurons versus astrocytes during brain development, or, in the immune system, the relative abundance of lymphoid versus myeloid cells in disease or aging [9]. The ability to manipulate the relative proportions of distinct stem cell and progenitor subpopulations provides an important route to controlling the differenti-

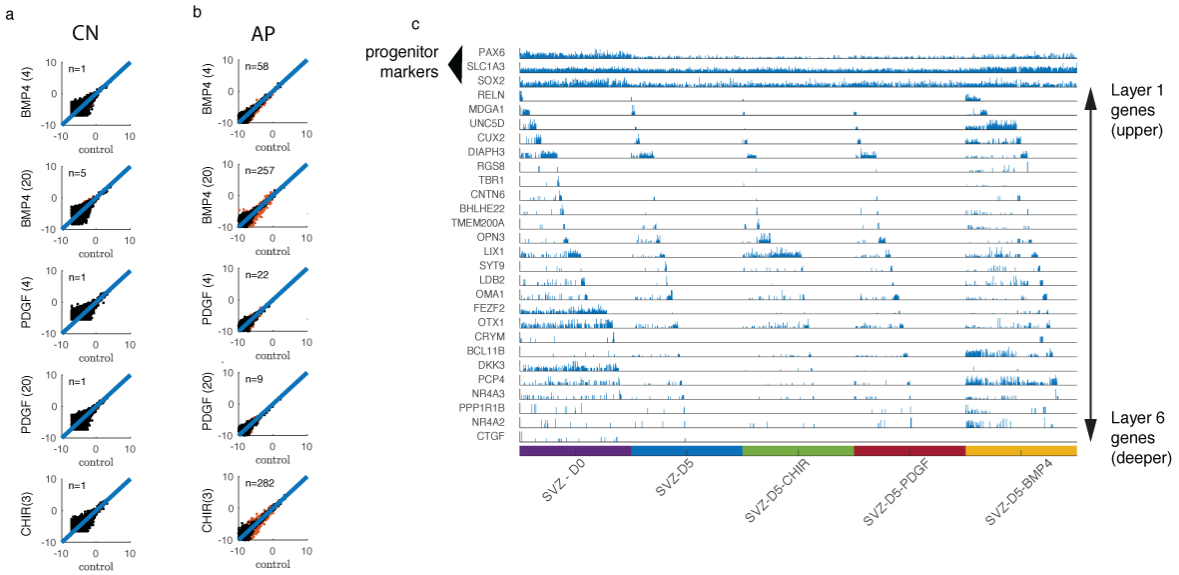


Figure 5: BMP and CHIR impact AP populations but not CN. Scatter plot of gene means for both (a) CN and (b) AP populations in response to BMP4, CHIR , and PDGF. Gene means for control population (20ng/mL EGF/FGF, 5 days) are plotted on the x-axis. Red points denote genes which have log2(fold change) ¿ 1, and statistically significant p-values (T-test, Benjamini-Hochberg corrected, alpha=0.01). Gene means are normalized by total transcript count per cell and rescaled. Number of statistically significant genes reported on upper left. (c) Gene expression plot of layer-specific genes [**?**] in AP cells across samples. Gene levels (y-axis) is individually normalized (rescaled by max), and 500 random cells are displayed for each sample. All populations maintain expression of progenitor markers Pax6, Sox2, and SLC1A3(Glast), but only BMP4 reinvigorates expression of layer-specific genes broadly.

ation potential of any stem cell population and also to modulating the state of tissues in the body during development and disease. In this work, we show that signaling ligands can directly steer the population structure of stem and progenitor cells isolated from the embryonic mouse brain. We identify conditions that amplify specific pre-neuronal and pre-astrocytic populations. In future work, it will be interesting to explore whether additional factors allow the expansion of other cell types including proliferative pre-neuronal stem cells.

Our approach has implications for understanding the behavior of stem cell populations broadly. Cells in the embryo are exposed to combinations of signaling molecules generated by morphogen gradients. An important question is to determine how stem cells can decode combinatorial signaling environments to select their fate and proliferation rate. Our findings point towards a potentially simple signaling code in which different signals linearly regulate the proliferation rate of distinct transcriptionally primed stem cell populations. Interestingly, few signals in our screen created cells with entirely new transcriptional signatures (only BMP4 and CHIR), suggesting that transcriptional patterns within these subtypes could be hardwired by epigenetics and difficult to alter. Future work using dynamical measurements that can accurately reflect subpopulation state will be important for determining the possibility of direct transition between subpopulations with different transcriptional lineage priming.

Our work emphasizes the role of local signaling environments in generating and maintaining population diversity both in the embryo and in artificial cell culture contexts. The embryo maintains population diversity naturally, by controlling signaling environments on cellular length scales to allow divergent cell-types to be generated in close spatial proximity [24]. The population diversity of the embryo is thus directly enabled by the self-organization of local combinatorial signaling environments. An important challenge for cell engineering is to control combinatorial signaling environments outside of the body with similar precision. In engineering applications, signals have traditionally been applied globally, to an entire cell culture, and therefore, might be fundamentally limited in their ability to control the generation of different cell types in close spatial proximity. Organoid based approaches solve the signaling control problem by generating signals locally through cell-cell interactions [25] It might also be possible to engineer synthetic signaling environments where signals are modulated on cellular length scales to generate cell populations where different cell-types can emerge through distinct, micron-scale, local signaling environments.

Broadly, our approach can be applied to study the combinatorial impact of a wide range of signaling molecules on heterogeneous populations of stem and progenitor cells. The key technical innovations we apply are single-cell mRNA-seq, population level gene expression analysis and computational modeling. We anticipate that this strategy can be applied to uncover the underlying signaling code for engineering the population structure of heterogeneous multicellular constructs across both synthetic biology and tissue engineering applications.

References

[1] A. L. Perrier, V. Tabar, T. Barberi, M. E. Rubio, J. Bruses, N. Topf, N. L. Harrison, and L. Studer, "Derivation of midbrain dopamine neurons from human embryonic stem cells," *Proceedings of the National Academy of Sciences*, vol. 101, no. 34, pp. 12543–12548, 2004.

[2] R. Tsunemoto, S. Lee, A. Szűcs, P. Chubukov, I. Sokolova, J. W. Blanchard, K. T. Eade, J. Bruggemann, C. Wu, A. Torkamani, *et al.*, "Diverse reprogramming codes for neuronal identity," *Nature*, vol. 557, no. 7705, p. 375, 2018.

[3] V. Balasubramaniyan, A. De Haas, R. Bakels, A. Koper, H. Boddeke, and J. Copray, "Functionally deficient neuronal differentiation of mouse embryonic neural stem cells in vitro," *Neuroscience research*, vol. 49, no. 2, pp. 261–265, 2004.

[4] S. E. Marsh, S. T. Yeung, M. Torres, L. Lau, J. L. Davis, E. S. Monuki, W. W. Poon, and M. Blurton-Jones, "Hucns-sc human nscs fail to differentiate, form ectopic clusters, and provide no cognitive benefits in a transgenic model of alzheimer's disease," *Stem cell reports*, vol. 8, no. 2, pp. 235–248, 2017.

[5] A. J. Anderson, K. M. Piltti, M. J. Hooshmand, R. A. Nishi, and B. J. Cummings, "Preclinical efficacy failure of human cns-derived stem cells for use in the pathway study of cervical spinal cord injury," *Stem Cell Reports*, vol. 8, no. 2, pp. 249–263, 2017.

[6] S. Temple and L. Studer, "Lessons learned from pioneering neural stem cell studies," *Stem cell reports*, vol. 8, no. 2, pp. 191–193, 2017.

[7] S. J. Franco and U. Müller, "Shaping our minds: stem and progenitor cell diversity in the mammalian neocortex," *Neuron*, vol. 77, no. 1, pp. 19–34, 2013.

[8] R. A. Nimmo, G. E. May, and T. Enver, "Primed and ready: understanding lineage commitment through single cell analysis," *Trends in cell biology*, vol. 25, no. 8, pp. 459–467, 2015.

[9] C. E. Muller-Sieburg, H. B. Sieburg, J. M. Bernitz, and G. Cattarossi, "Stem cell heterogeneity: implications for aging and regenerative medicine," *Blood, The Journal of the American Society of Hematology*, vol. 119, no. 17, pp. 3900–3907, 2012.

[10] M. A. Bonaguidi, M. A. Wheeler, J. S. Shapiro, R. P. Stadel, G. J. Sun, G.-l. Ming, and H. Song, "In vivo clonal analysis reveals self-renewing and multipotent adult neural stem cell characteristics," *Cell*, vol. 145, no. 7, pp. 1142–1155, 2011.

[11] S. Lodato and P. Arlotta, "Generating neuronal diversity in the mammalian cerebral cortex," *Annual review of cell and developmental biology*, vol. 31, pp. 699–720, 2015.

[12] L. Telley, G. Agirman, J. Prados, S. Fievre, P. Oberst, I. Vitali, L. Nguyen, A. Dayer, and D. Jabaudon, "Single-cell transcriptional dynamics and origins of neuronal diversity in the developing mouse neocortex," *bioRxiv*, p. 409458, 2018.

[13] S. Chen, J. H. Park, P. Rivaud, E. Charles, J. Haliburton, F. Pichiorri, and M. Thomson, "Dissecting heterogeneous cell-populations across signaling and disease conditions with popalign," *bioRxiv*, p. 421354, 2018.

[14] S. Martini, K. Bernoth, H. Main, G. D. C. Ortega, U. Lendahl, U. Just, and R. Schwanbeck, "A critical role for sox9 in notch-induced astrogliogenesis and stem cell maintenance," *Stem Cells*, vol. 31, no. 4, pp. 741–751, 2013.

[15] S. C. Noctor, A. C. Flint, T. A. Weissman, W. S. Wong, B. K. Clinton, and A. R. Kriegstein, "Dividing precursor cells of the embryonic cortical ventricular zone have morphological and molecular characteristics of radial glia," *Journal of Neuroscience*, vol. 22, no. 8, pp. 3161–3173, 2002.

[16] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, *et al.*, "A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function," *Journal of Neuroscience*, vol. 28, no. 1, pp. 264–278, 2008.

[17] H. G. Kuhn, J. Winkler, G. Kempermann, L. J. Thal, and F. H. Gage, "Epidermal growth factor and fibroblast growth factor-2 have different effects on neural progenitors in the adult rat brain," *Journal of Neuroscience*, vol. 17, no. 15, pp. 5820–5829, 1997.

[18] K. Funa and M. Sasahara, "The roles of pdgf in development and during neurogenesis in the normal and diseased nervous system," *Journal of Neuroimmune Pharmacology*, vol. 9, no. 2, pp. 168–181, 2014.

[19] M. Y. S. Kalani, S. H. Cheshier, B. J. Cord, S. R. Bababeygy, H. Vogel, I. L. Weissman, T. D. Palmer, and R. Nusse, "Wnt-mediated self-renewal of neural stem/progenitor cells," *Proceedings of the National Academy of Sciences*, vol. 105, no. 44, pp. 16970–16975, 2008.

[20] M. A. Bonaguidi, T. McGuire, M. Hu, L. Kan, J. Samanta, and J. A. Kessler, "Lif and bmp signaling generate separate and discrete types of gfap-expressing cells," *Development*, vol. 132, no. 24, pp. 5503–5514, 2005.

[21] D. A. Lim, A. D. Tramontin, J. M. Trevejo, D. G. Herrera, J. M. García-Verdugo, and A. Alvarez-Buylla, "Noggin antagonizes bmp signaling to create a niche for adult neurogenesis," *Neuron*, vol. 28, no. 3, pp. 713–726, 2000.

[22] J. Gehring, J. H. Park, S. Chen, M. Thomson, and L. Pachter, "Highly multiplexed single-cell rna-seq for defining cell population and transcriptional spaces," *bioRxiv*, p. 315333, 2018.

[23] B. J. Molyneaux, P. Arlotta, J. R. Menezes, and J. D. Macklis, "Neuronal subtype specification in the cerebral cortex," *Nature reviews neuroscience*, vol. 8, no. 6, p. 427, 2007.

[24] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, *et al.*, "Transcriptome-scale super-resolved imaging in tissues by rna seqfish+," *Nature*, vol. 568, no. 7751, p. 235, 2019.

[25] A. Warmflash, B. Sorre, F. Etoc, E. D. Siggia, and A. H. Brivanlou, "A method to recapitulate early embryonic spatial patterning in human embryonic stem cells," *Nature methods*, vol. 11, no. 8, p. 847, 2014.

Acknowledgements Allan Pool-Hermann, Carlos Lois, David Schaffer, Elliott Robinson, Eric Chow, Chris McGinnis, Jase Gehring, Sandy Nandagopal, Michael Elowitz, Elisha Mackay, Thomson Lab, Beckman Institute Single-cell Profiling and Engineering Center (SPEC).

Supplementary Information:
Designing signaling environments to steer transcriptional diversity in neural progenitor cell populations

Jong H. Park[1,2], Tiffany Tsou[1,2], Paul Rivaud[1,2], Matt Thomson[1,2,3], and Sisi Chen[1,2,3]

[1]Division of Biology and Biological Engineering, California Institute of Technology. Pasadena, California, 91125, USA.

[2]Beckman Center for Single-cell Profiling and Engineering. Pasadena, California, 91125, USA.

[3]correspondence to: sisichen@caltech.edu, mthomson@caltech.edu

Methods

Mouse brain tissue dissociation Tissue from mouse brain was purchased from Brainbits or acquired from collaborators (Allan Pool, Oka Lab, Caltech, Elisha McKay, Gradinaru Lab), and digested using 6mg of Papain (Brainbits) at 2mg/mL in Hibernate E medium(Brainbits) for 30 minutes at 30C. After incubation, tissue was triturated for 1 minute using a silzanized pasteur pipet. Dissociation medium containing the disociated cells were centrifuged for 1 minute at 200 x g and the cell pellet was either resuspended in media for further culture or resuspended in 1x PBS + 0.04% BSA for single-cell profiling through the 10X Genomics platform.

SVZ cell derivation Neural progenitors are dissociated from the combined cortex, hippocampus and ventricular zone of E18 mice (Brainbits). After dissociation and resuspension of the cells in NPgrow, cells were counted and plated onto ultra-low attachment 100mm culture plates (Corning) at one million cells per plate in 10mL of NPgrow, which contains 20 ng/mL EGF and FGF2, and supplemented with 1% Penicillin-Streptomycin. Cells were cultured at 37C and 5% CO2 and supplemented medium was changed every 2 to 3 days.

Combinatorial signaling screen We applied combinatorial signals to cells dissociated from the SVZ region of the E18 mouse brain to understand how signals regulate the population structure and transcriptional state of progenitor cells. We used NPGrow media which did not contain EGF/FGF, and added signaling factors to the final concentrations indicated in Table 1. For EGF, FGF2, BMP4, PDGF-AA, and FGF9, the maximum concentration used was 20 ng/mL. For IFN-G and GM-CSF, the maximum concentration used was 3 ng/mL, and for CHIR99021, the maximum concentration used was 3 uM. After cells were resuspended in media, they were cultured for 5 days before dissociation for clicktag multiplexing and single-cell profiling.

Experimental multiplexing using Clicktags Cultured cells were dissociated in Accutase into a

single cell suspension. Cells were then fixed in ice-cold 80% methanol. Prior to cell labeling, two methyltetrazine (MTZ) activated barcoding oligos were combined for each labeling sample, NHS-trans-cyclooctene (NHS-TCO) was then added and left to reaction for 5 minutes. The previously fixed cells were then mixed thoroughly with the barcode oligo solution and incubated for 30 minutes at room temperature. Labeling reaction was then quenched with the addition of Tris-HCl and methyltetrazine-DBCO. After the quenching step, cells from all labeling samples were pooled together along with the addition of two volumes of PBS-BSA then centrifuged at 800 x g for 5 minutes followed by two more cycles of PBS-BSA washes and centrifugation steps. After the final centrifugation step, the cell pellet was resuspended in 100uL of PBS-BSA and counted on a hemocytometer. Cells were then loaded onto 3-4 lanes, targeting 10,000 cells per lane, of the Chromium Controller and processed as recommended by the 10x Genomics v2 Single Cell 3' reagent kit protocol until completion of the cDNA amplification PCR step. At this step, SPRI size-selection was used to separate the barcode oligos from the cDNA which were then processed separately. We used a dual-tagging approach, so that each experimental sample gets two clicktag barcodes.

Barcode demultiplexing Filtered cell barcodes from the aligned gene expression (Cellranger) is used as a whitelist to parse clicktags data from fastqs. Tag sequences extracted from fastqs are aligned to a list of known clicktag sequences, allowing for 1 mismatch (Levenshtein distance ¡=1). Counts for clicktags are accumulated and stored in a cell barcodes by tags matrix. To threshold the data, we computed a count distribution for each individual clicktag, and used Otsu's method to obtain optimal thresholds for each clicktag sequence. Each cell is thresholded for each barcode independently, and cells which are have the correct barcode pair combinations are retained.
Data normalization Gene expression values are normalized according to the PopAlign paper[1]. Briefly, gene expression values, g_i, are first rescaled by the total number of transcripts per cell to account for droplet-to-droplet variation in transcript capture. Normalized values are then log-transformed after multiplication by a scaling factor β.

$$g_i' = \log(\beta \frac{g_i}{\sum_i^n g_i} + 1) \tag{1}$$

The scaling factor β is chosen in order to achieve a smooth transition in the distribution of transformed g_i' when raw g_i values step from 0 to 1. We found that by setting β to be roughly the median number of total transcript counts in a cell (3000) generally yields good results.

Probabilistic models and alignment To compare single-cell datasets, we construct probabilistic models that estimate the underlying joint probability distribution of cells within gene expression space. Probabilistic models are useful because they are a highly-compressed, conceptually meaningful representations of single-cell data, and their sub-components can be quantitatively compared using well-defined statistical metrics. In this work, the probabilistic model-based comparisons allow us to automatically align in vitro-grown cells to cells from the natural brain using statistical divergence, and rank and compare entire populations of cells using a log-likelihood ratio metric.

Models are constructed as described in the PopAlign paper [1]. Briefly, single-cell gene expression

data is first normalized as described above. The genes are then filtered and only highly variable genes (e.g. supra-Poisson) are kept.

We then compress gene expression data into a set of m features, which are gene expression programs extracted from the data using matrix factorization techniques. In this work, we use the top 10- PCA features for analysis. Using these features, we construct a compressed representation of each cell as a vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_m)$ of m feature coefficients, which weight the magnitude of gene expression programs within a given cell.

For each sample we build a mixture model that factors the population into a set of distinct subpopulations, each represented by a distinct mixture component (e.g. a Gaussian probability density):

$$P(\boldsymbol{c}) = \sum_{j=1}^{l} w_j \, \phi_j(\boldsymbol{c}) \tag{2}$$
$$\text{where } \phi_j(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$$

Aligning probabilistic models To determine which subpopulations to compare between samples, we perform an alignment between models based on statistical 'closeness'. We denote the population to be compared as the 'reference' population, and the other populations as 'test' populations. For each mixture component in the 'test' population, we can find the closest mixture component within the reference set by minimizing the Jeffrey Divergence, a symmetrized version of the KL-divergence:

$$\arg\min_j D_{\text{JD}} \left(\phi_i^{\text{test}}(\boldsymbol{c}) \parallel \phi_j^{\text{ref}}(\boldsymbol{c})\right), \tag{3}$$

Query In samples where the number of cells is too small to build a standalone model, we use an existing model to classify cells into different subpopulations. For the combinatorial signaling screen, some samples had roughly 100 cells, which would not adequately sample the feature space. For these experiments, we query each datapoint c_i against a control model built from the reference population. The control model consists of a collection of mixture components $\{\phi_j(\boldsymbol{c})\}$. We can then assign a class Y_i for each specific cell - $\boldsymbol{c_i}$:

$$Y_i = \arg\max_j \left(\phi_i(\boldsymbol{c_i})\right)$$

All cells in each sample receive a class label Y, these class labels can then be used to compute sample-specific subpopulation parameters - μ_j, Σ_j and w_j. The resulting abundances w_j are used in a simple log-linear model to learn how population structure is steered by signals.

Mathematical model of population response to signal concentrations

Changing the signaling environment is a common strategy for the developing embryo to alter its population structure in accordance with developmental needs. Understanding how complex combinations of signals interact to regulate the relative abundance of populations within a complex mixture is critical for uncovering design principles that allow us to program populations of distinct stem and progenitor cell populations.

In our experimental system, we can apply many signaling concentrations on dissociated cells from the developing brain to understand how the populations are affected by single signals or their combinations. Many of the signals modulate the relative proportions of specific subpopulations within our heterogeneous mixture. For instance, we found that increasing EGF/FGF increases the proportion of pre-astrocytic progenitor cell type (AP), while decreasing the proportion of the committed neuronal cell type (CN).

We wanted to see if a simple model could describe the effects observed in the data, so we developed a model that encoded only a linear dependence on the concentrations of signaling molecules. Prior work has shown that in individual cell types, combinations of signals can be integrated by cellular signal processing in potentially complex ways, including ratiometric detection [2], or hierarchically [3]. However, additive responses to signaling combinations have been shown to be more common [2], so we believed a linear model could be reasonably capture the population-level restructuring we observed in our signal-response data.

We start with an exponential growth model for a specific subpopulation, i:

$$n_i(t = T) = n_{i,0}e^{rT} \tag{4}$$

where n_1 is the number of cells in population i at $t = T$, $n_{i,0}$, is the starting population at $t = 0$, and r is the growth rate. We explicitly relate the growth rate r_iT to a linear function of signals:

$$r_iT = \boldsymbol{\beta_i x} + c_i \tag{5}$$

where $\boldsymbol{\beta_i} = (\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,j})$ are the coefficients linking the signaling cues to the proliferation rate, $\boldsymbol{x} = (x_1, x_2 \ldots, x_j)^T$ are the concentrations of signaling cues, and b_i is a bias term that describes the growth rate of population i in the absence of all signals. Thus, we can express the growth model for a specific subpopulation of cells in terms of a linear response to signaling inputs:

$$ln\left(\frac{n_{i,T}}{n_{i,0}}\right) = \boldsymbol{\beta_i x} + c_i \tag{6}$$

However, we are not considering only a single population of cells, but many subpopulations that make up an entire heterogeneous mixture. Thus we can reformulate this model in terms of the rel-

ative population size of each population to a reference population, k. We normalize the proportion by directly dividing n_1/n_k and taking the log to generate the following relationship:

$$\begin{aligned} ln\left(\frac{n_{i,T}}{n_{k,T}}\right) &= (\boldsymbol{\beta_i} - \boldsymbol{\beta_k})\boldsymbol{x} + c_i - c_k + ln(\frac{n_{i,0}}{n_{k,0}}) \\ ln\left(\frac{n_{i,T}}{n_{k,T}}\right) &= \boldsymbol{A_i}\boldsymbol{x} + b_i \end{aligned} \tag{7}$$

In this way, the relative ratios n_i/n_k are also dependent on a linear combination of signals. A_i can be seen as the relative growth rate, while b_i encapsulates both the difference in starting populations as well as differences in the basal growth rates in the absence of signals.

For k classes of cells, there are only k-1 degrees of freedom (given a constant total population size), so we can write k-1 equations. In our experiment, we have k=3 classes of cells (yielding 2 sets of equations), and j=7 signals, resulting in a set of 14 parameters to be fit. In total we have experimental data from 40 signaling combinations, which generates an overdetermined set of equations. To fit these equations, we use mldivide in the Matlab package, which uses QR decomposition to find a least squares solution.

The model can be used to directly calculate the single population probabilities by constraining the sum to be 1. By rearranging the following equation,

$$n_{tot} = \sum_{i=1}^{k} n_i \tag{8}$$

we get the following relationship between n_k and the relative ratios $\frac{n_i}{n_k}$:

$$n_k = \frac{n_{tot}}{\frac{n_1}{n_k} + \frac{n_2}{n_k} + \cdots + 1} \tag{9}$$

Using this formula, we can calculate n_i's for any arbitrary population size. However, when we constrain $n_{tot} = 1$, we essentially force n_i's to behave like probabilities, p_i. In this way, we can use the model to produce predictions of the subpopulation-specific probabilities.

References

[1] S. Chen, J. H. Park, P. Rivaud, E. Charles, J. Haliburton, F. Pichiorri, and M. Thomson, "Dissecting heterogeneous cell-populations across signaling and disease conditions with popalign," *bioRxiv*, p. 421354, 2018.

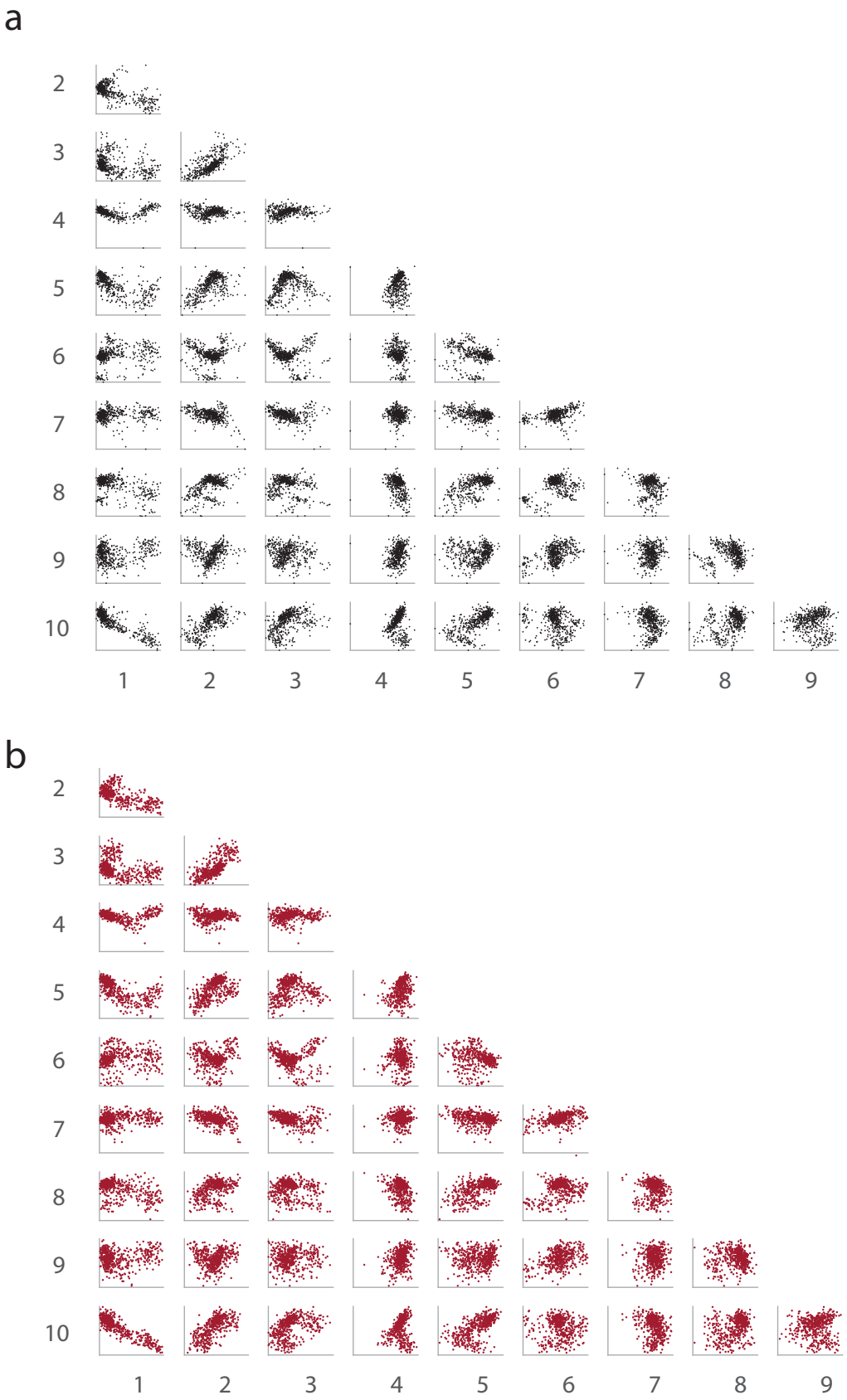
[2] Y. E. Antebi, J. M. Linton, H. Klumpe, B. Bintu, M. Gong, C. Su, R. McCardell, and M. B. Elowitz, "Combinatorial signal perception in the bmp pathway," *Cell*, vol. 170, no. 6, pp. 1184–1196, 2017.

[3] S. Chen, A. W. Bremer, O. J. Scheideler, Y. S. Na, M. E. Todhunter, S. Hsiao, P. R. Bomdica, M. M. Maharbiz, Z. J. Gartner, and D. V. Schaffer, "Interrogating cellular fate decisions with high-throughput arrays of multiplexed cellular communities," *Nature communications*, vol. 7, p. 10309, 2016.

Table 1: Datasets

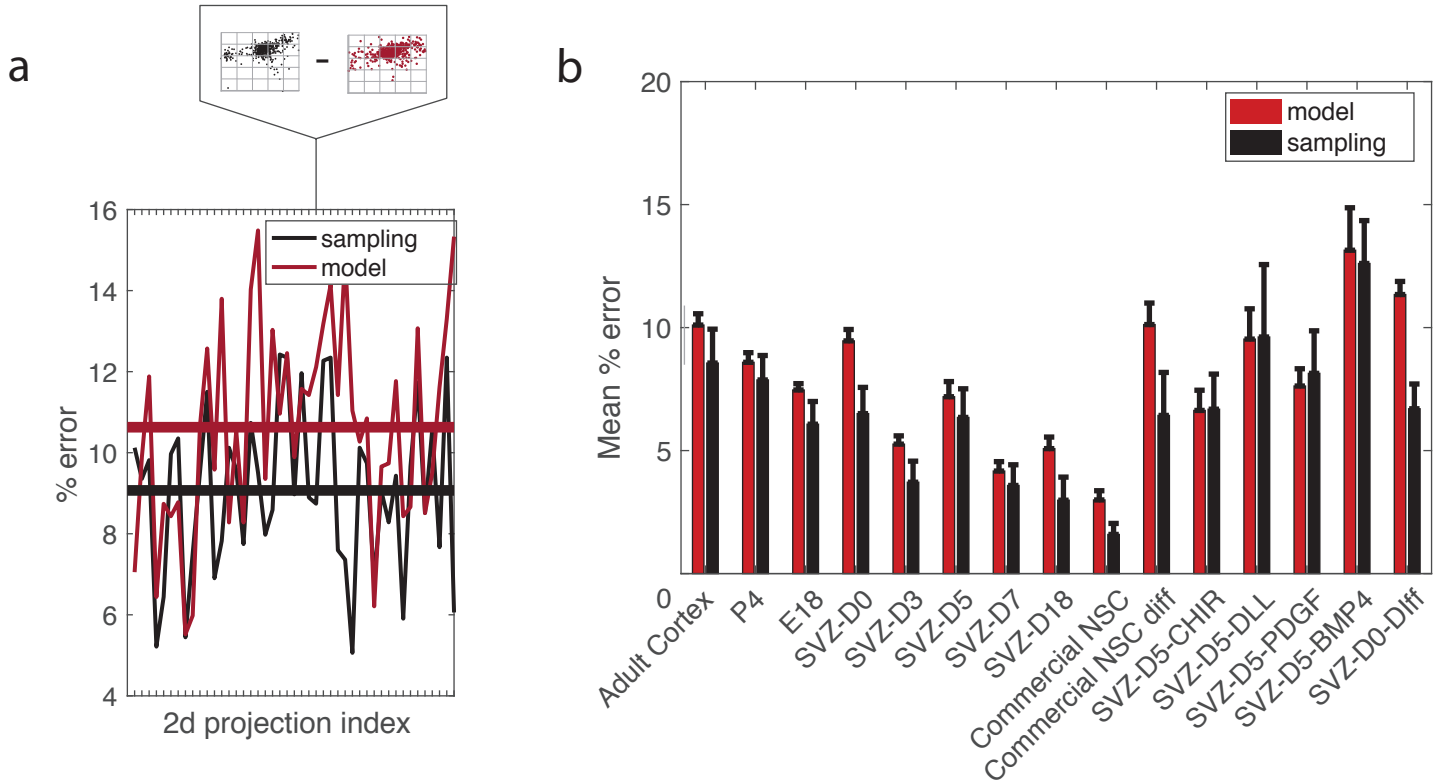
Name	Description	# cells
P4	Whole brain from P4 mice (Brainbits)	12015
AdCo	Cortex from 8-week old mouse (Oka lab)	2657
SVZ-D0	combined cortex, hippocampus, and ventricular zone (Brainbits)	5030
SVZ-D3	3 days in Npgrow	9862
SVZ-D5	5 days in Npgrow	1125
SVZ-D7	7 days in Npgrow	3820
SVZ-D18	18 days in Npgrow	6256
SVZ-D0-DIFF	5 days in co-culture media	3719
SVZ-D5-DIFF	5 days in NPgrow, then 5 days in co-culture media	3505
SVZ-D5	5 days in NPgrow	1125
CHIR-D5	5 days in NPgrow + CHIR (3 uM)	997
PDGF-D5	5 days in NPgrow + PDGF (20ng/mL)	903
BMP4-D5	5 days in NPgrow + BMP4 (20ng/mL)	715
NPC-MULT-1	5 days in NPgrow (minus EGF/FGF) plus supplemental factors	8691
NPC-MULT-2	5 days in NPgrow (minus EGF/FGF) plus supplemental factors	10020

SI Figure 1: Model matches experimental data qualitatively



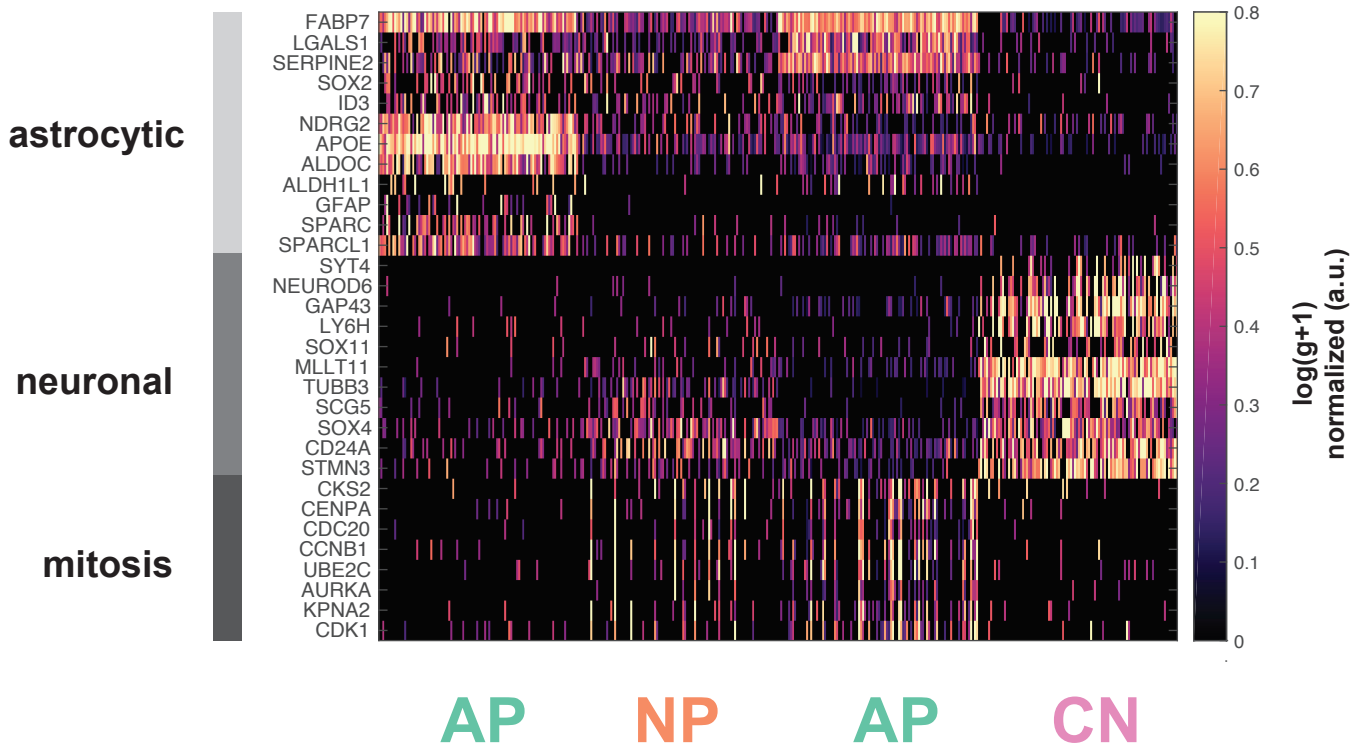
SI Figure 1: Array of 2D scatterplots of experimental data for the SVZ-D0 sample. (a) and model-simulated data (b) For each plot tile, the PC number is displayed on the axis. Corresponding plots have the same x- and y- axis scaling.

SI Figure 2: Model matches experimental data quantitatively



SI Figure 2: Quantitative comparison of model error for all samples (a) % error for each 2D projection, calculated across a 5x5 array of 2D bins (see Methods). Model error is in red and sampling error (calculated between training and validation data) is in black. Mean % error across all 2D projections is denoted as thick horizontal line. (b) Mean % error across all 2D projections is compared to data sampling error for each dataset. Error bars show the standard deviation calculated acoss 20 separate model building and data resampling attempts.

SI Fig 4: Gene expression heatmap of populations from combinatorial signaling screen

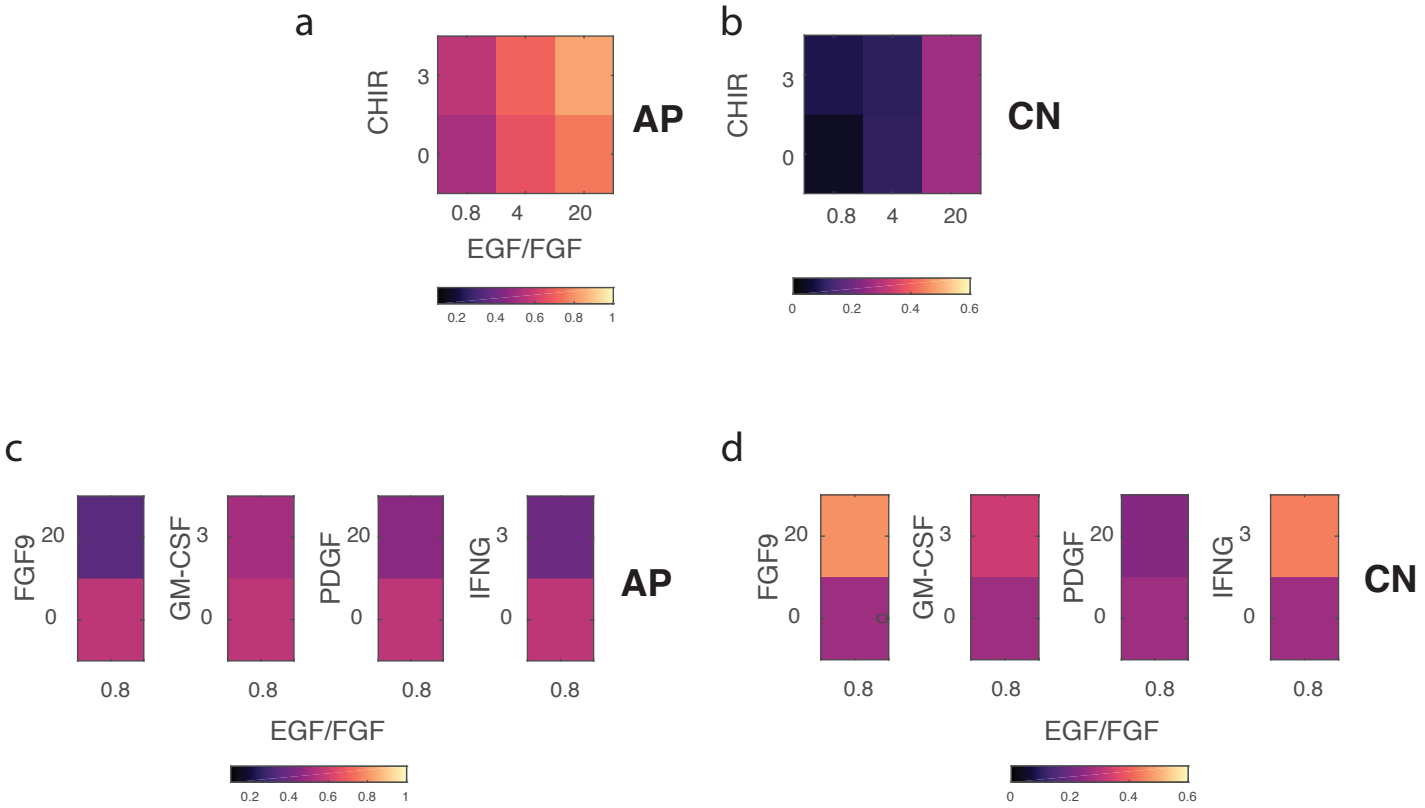


SI Figure 4: Gene expression heatmap of AP, NP, and CN cells from the combinatorial signaling screen. Gene expression programs are denoted on the right. Gene counts are normalized and logged (see Methods), and then rescaled by the gene maximum in each row. Displayed values are saturated at 0.8 to improve visual clarity.

SI Figure 3: Signaling pathways are active in cells from SVZ-D0.

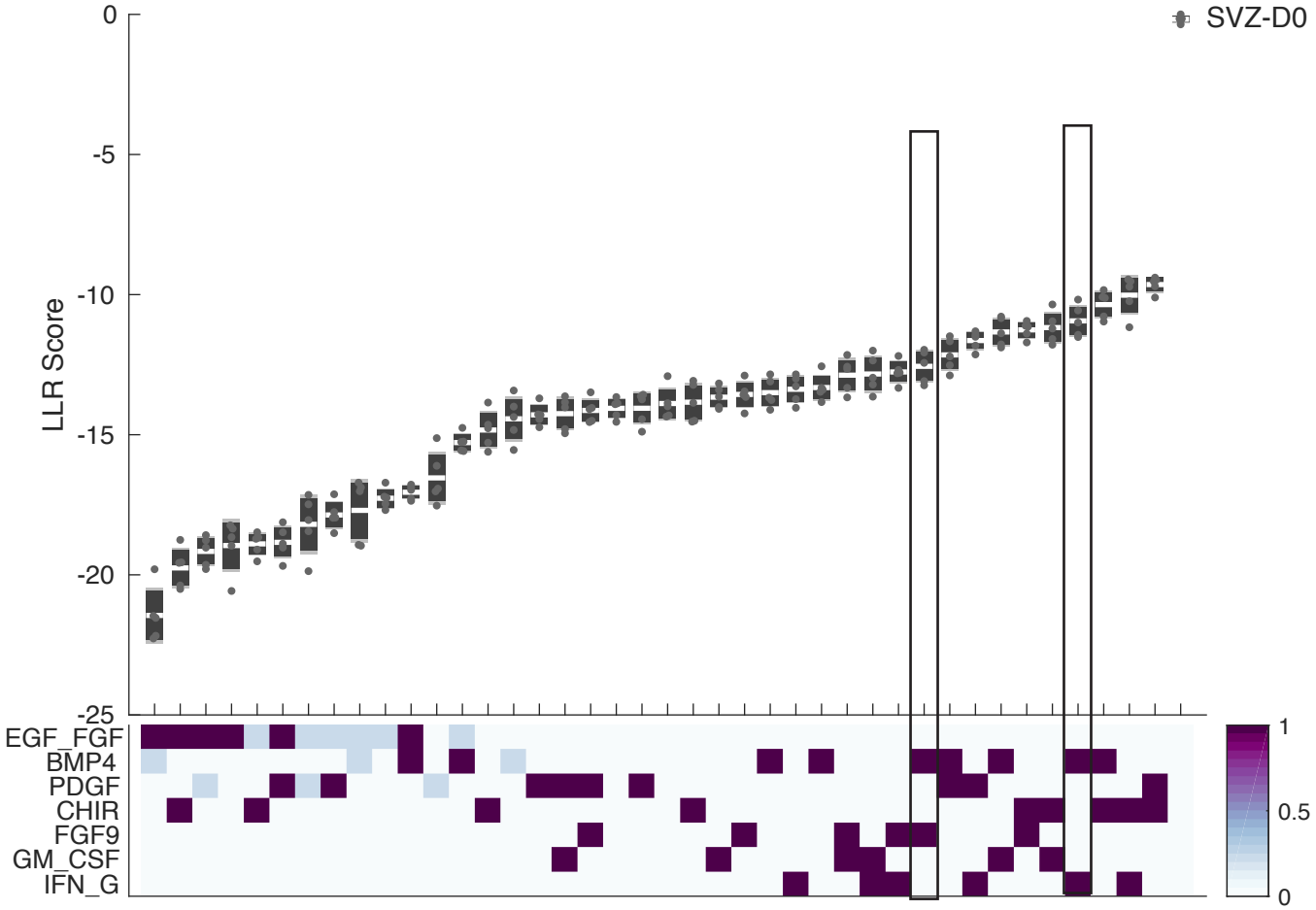
SI Figure 3: Signaling pathways (BMP, PDGF, EGF, FGF, WNT, and Notch) are active in cells from the ventricular zone at E18. (a) t-SNE plot of VZ-D0 cells separated into distinct subpopulations. (b) Normalized gene expression values for receptors of the aforementioned signaling pathways. (c) Normalized gene expression values for ligands of the aforementioned signaling pathways. All data are log-normalized and rescaled from 0 to 1.

SI Figure 5: Heatmaps of AP and CN population proportions in response to additional signals.



SI Figure 5: Heatmaps pf AP and CN population proportions in response to additional signals.The proportion of (a) AP population and (b) CN populations in response to CHIR and EGF/FGF. The proportions of (c) AP and (d) CN populations in response to FGF9, GM-CSF, PDGF, IFN-G, at low EGF/FGF concentrations (0.8 ng/mL).

SI Fig 6. Ranking of signaling sceen populations relative to in vivo SVZ-D0 population



SI Figure 6. Ranking of samples relative to SVZ-D0 control (a) Log-likelihood ratio (LLR) scores for each sample against the control population. The more a population of cells diverges from the SVZ-D0 sample, the lower its LLR score. The two conditions designed in Fig 4e-f are highlighed in boxes. LLR scores are bootstrapped by building 20 models for each signaling ondition and then calculating the LLR relative to the SVZ-D0 model, and using 100 cells are sampled from each signaling condition. White line: mean. Dark gray: sem. Light gray: std. Outliers are individually plotted.