ABSTRACT

MENNICKE, CHRISTINE VICTORIA. A Data-Driven Framework for Modeling the Neurogenesis-to-Gliogenesis Switch. (Under the direction of Mansoor Haider.)

Embryonic cortical development is driven by a population of cells called radial glial progenitors (RGPs), whose divisions produce the neurons and glial cells that build the cortex. At earlier stages of cortical development, RGPs undergo a period of neuron production called *neurogenesis*. In the later stages of cortical development following neuron production, the RGPs shift to producing glial cells via *gliogenesis*. This shift is known as the *neurogenesis-to-gliogenesis switch*, or NGS. It is vital that the NGS is controlled such that RGPs produce the appropriate quantity of neurons and glia, as disruptions in neural and glial production can cause significant diseases such as multiple sclerosis, epilepsy, and glioma. Thus, being able to describe how RGPs divide and differentiate into neurons and glia during the NGS may give insight into understanding how such diseases progress.

Recently, a cell labeling technique called mosaic analysis with double markers (MADM) was developed, which labels the offspring (neurons or glia) arising from single RGPs with red and green fluorescent markers. Specifically, the red and green labels correspond respectively to cells from the two sublineages that arise after an RGP initially divides. This provides information not only about the total neurons and glia produced (red+green) by individual RGPs, but also about the symmetry or asymmetry of neural and glial production (red versus green). Data that counts the offspring of individual cells is referred to as *clonal data*.

In this thesis, we develop mathematical models and techniques to analyze a set of clonal MADM data gathered in embryonic mouse cortices during the NGS. The data provides the total number of red neurons, red glia, green neurons, and green glia produced by each MADM-labeled RGP, or 'clone.' Our focus is on the glia in this dataset, since neurogenesis has previously been studied using MADM. Thus, our goal is to use mathematical methods to analyze the set of glia per clone and develop hypotheses describing how RGPs divide and differentiate into glia. In particular, clonal level data may exhibit different patterns that denote RGP division mechanisms as *deterministic* or *stochastic*. We identify which patterns are present in the set of glia per clone so that we can define mechanisms that represent RGP behavior during gliogenesis. The work in this thesis thus presents a data-driven approach for identifying rules governing cell division and differentiation during the NGS. These rules are then used to model glial production during cortical development.

First, self-organizing maps (SOMs) are used to compare the sets of clones from each individual mouse (Chapter 2). This comparison is performed to evaluate the consistency of the MADM technique across multiple samples. Following this investigation, we shift our analysis to sets of clones pooled together across multiple mice within the dataset. The focus of the subsequent analysis is on the distribution of clone sizes (total glia per clone) across the entire population of clones. The theory

of branching processes is used to motivate testing the distribution of glia per clone for patterns classified as 'stochastic' (Chapter 3). This produces evidence that a portion of RGPs may behave stochastically, but the other portion may not. We then use clustering and statistical analysis to further delineate the existence of two separate sets of RGPs with different patterns of glial production (Chapter 4). These results lead us to develop a hypothesis that two subpopulations of RGPs exist: NGS-RGPs, which show deterministic patterns of glial production, and G-RGPs, whose patterns of glial production are stochastic. We then test the distribution of glia per clone for deterministic patterns (Chapter 5). The identification of deterministic patterns in the distribution of glia per clone for a subset of the clones, but not all clones combined, further supports our hypothesis of separate subpopulations of NGS-RGPs and G-RGPs. Finally, we use the previous results to propose deterministic and stochastic rules to model NGS-RGP and G-RGP divisions during gliogenesis, respectively (Chapter 6). We simulate sets of clones according to these rules, finding that the simulated clones produce a distribution of glia per clone that is consistent with the true NGS data. These results suggest that a combination of deterministic and stochastic rules describe RGP behavior during the NGS and gliogenesis.

A Data-Driven Framework for Modeling the Neurogenesis-to-Gliogenesis Switch

by
Christine Victoria Mennicke

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2020

APPROVED BY:

Alen Alexanderian	Troy Ghashghaei
Ralph Smith	Mansoor Haider Chair of Advisory Committee

DEDICATION

To my family, friends, and especially Dustin.

BIOGRAPHY

Christine grew up in New Brighton, MN and attended Concordia University, St. Paul, earning a dual degree in Mathematics and Biology in 2011. In 2015, she moved to Raleigh, NC to begin her graduate studies in Applied Mathematics at North Carolina State University. After five long years of work, she still feels that math is v good.

ACKNOWLEDGEMENTS

First, I want to thank my advisor Mansoor Haider for his consistent support over the past five years, including initially encouraging me to apply for the National Science Foundation Graduate Research Fellowship, remaining flexible as I completed outside programs and internships, and providing valuable research and career guidance. I would also like to thank Troy Ghashghaei for helping me understand the data and involving me in his lab.

Additionally, I want to acknowledge the financial support I received from the NCSU Provost Fellowship and the NSF Graduate Research Fellowship. This support had a significant impact in alleviating some of the stresses of graduate school.

Finally, I want to thank my friends, who helped bring me a dose of reality and fun when both were needed. I also want to thank my family for their excitement and cheers (that we could not hear over the Zoom call during my defense). Most importantly, I could not have done this without Dustin, my lifeline who recognized when I needed assistance, helped me untangle my thoughts when I was most confused, and remained supportive and patient over these past years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

CLONAL DATA: DESCRIPTION AND METHODS

1.1 Biological motivation

In this thesis, we study the behavior of progenitor cells that are responsible for building the mammalian cerebral cortex during embryonic development. The timing and biological mechanisms that drive cortical development have been heavily studied, but much remains unclear or unknown. Here, we will describe what is currently known about the progenitor cell population that builds the cerebral cortex as it relates to our application. We will then outline the mathematical techniques that we will use in this thesis to study progenitor cell behavior from data.

1.1.1 Cerebral cortex

The human brain is composed of three major parts: the cerebrum, the cerebellum, and the brain stem. Of these, the cerebrum is the largest and is responsible for higher functions such as speech, reasoning, emotions, and learning [31]. The outer layer of the cerebrum is called the cerebral cortex – we will refer to this as simply the cortex.

Two broad cell types are found in the cortex: glial cells, or glia, and neurons, which are present in an approximately 5:1 ratio, respectively [31]. The combination of these cell types forms what is known as gray matter, while other areas of the brain containing no neuronal cell bodies are known as

white matter [23, 31]. The cortical layer of gray matter is 1.5-4 mm thick in the human brain, and its neurons and glia are arranged in six layers. These layers are formed during embryonic development in an 'inside-out' fashion, with interior/deep layers formed first, followed by superficial layers (Fig. 1.1) [32]. We discuss the formation of the cortex in the following section.

In larger mammals such as humans, the cortex contains folds to increase the surface area available for neuronal connections [43]. The cortex of smaller mammals such as mice is smooth. We can observe this difference by examining cross-sections of human and mouse brains (Fig. 1.2, cortex indicated). The cortex is present in mammals, and its role in higher functions indicates that it is a more recent development, phylogenetically [23]. Importantly, disruption of the production of cortical neurons and glia during development can cause significant diseases such as multiple sclerosis, epilepsy, and glioma [28, 56].

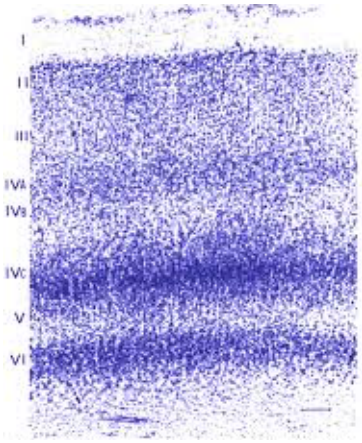


Figure 1.1 A cross-section of the cortex showing cells distributed in six layers. Layer I is the most superficial and develops last, while layer VI is the deepest and develops first. Adapted with permission from "The Primary Visual Cortex," by M. Schmolesky in *Webvision: The Organization of the Retina and Visual System [Internet],* H. Kolb, E. Fernandez, R. Nelson, editors, 2005, Salt Lake City (UT): University of Utah Health Sciences Center. Figure 13, [Nissl stain of the visual cortex reveals the different layers quite clearly.]. Copyright 2020 Webvision.

1.1.2 Cortical development and the neurogenesis-to-gliogenesis switch

At the cellular level, cortical development occurs via a population of cells called radial glial progenitors, or RGPs [38]. RGPs are multipotent progenitor cells, that is, they can undergo cell division to produce multiple cell types: additional RGPs, neurons, and glia. The production of additional RGPs via RGP division is called *proliferation*, and the production of functional cells (neurons and glia) is called *differentiation*. All neurons in the cortex are produced from RGP differentiation, and the glia

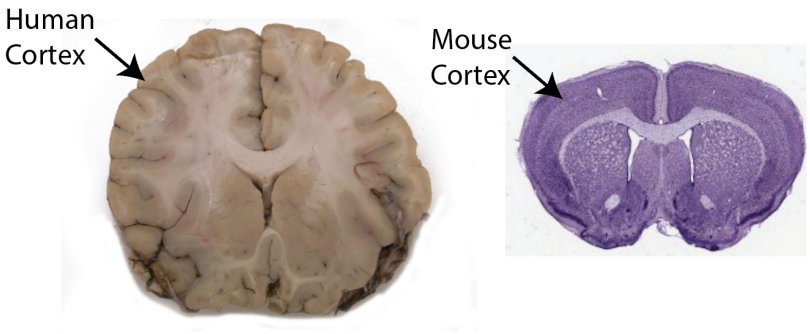


Figure 1.2 Human (left) and mouse (right) cerebral cross-sections. In the human brain, the cortex is identified by the outer layer of gray matter, indicated. Here, the folds which increase the cortical surface area are visible. In the mouse brain, the cortex is also found in the outer layer of the cerebrum. The cortex is not folded in mice, but is smooth. Adapted with permission from "Cerebral edema in a patient following cytoreductive surgery and hyperthermic intraoperative intraperitoneal chemoperfusion" by R.L. Nair, J. Tobias, G. Stemmerman et al., 2006, *World Journal of Surgical Oncology*, 4(85). Copyright 2006, Springer Nature; "Internet-Enabled High-Resolution Brain Mapping and Virtual Microscopy" by S. Mikula, I. Trotts, J. Stone, E.G. Jones, 2007, *NeuroImage*. 35(1):9-15.

produced by RGPs include two cell subtypes, astrocytes and oligodendrocytes [6, 16, 38, 41].

The production of the neurons and glia in the cortex via RGP divisions occurs in a defined series of steps in time and space. This is illustrated in Fig. 1.3 as it occurs in mice. Before the mouse embryo is eleven days old (denoted E11), the population of RGPs is located in a region called the ventricular zone (VZ). The white cell body located in the lower left of Fig. 1.3 denotes an RGP in this region. At this stage of development, the RGPs in the VZ undergo proliferative, symmetric divisions to increase the population size. Around E11, neurogenesis begins, which expands and continues until approximately E16 [22, 41]. Evidence has shown that RGPs differentiate asymmetrically to produce neurons; that is, of the two daughter cells produced from an individual RGP division during neurogenesis, one daughter cell is an RGP and the other is a neuron [16, 26]. Newly born neurons travel from the ventricular zone radially outward to form layers of the cortex, using long extensions of the RGP cell bodies as scaffolding. This process is shown in Fig. 1.3 as the green and orange cells moving upwards between the dates E13 and E16. Over this time frame, the migration of neurons produces the six layers of the cortex, with deeper layers established first, and more superficial layers established later from the newest neurons migrating past the deeper layers [9, 32, 42]. Following neurogenesis, a portion of the RGP population alters its production from neurons to glia [6, 26, 35, 39, 41] in a phenomenon known as the neurogenesis-to-gliogenesis switch, or NGS. RGPs that have begun producing glia will no longer produce neurons, and it is estimated that approximately 10-20% of RGPs become gliogenic during the NGS [16, 41]. In mice, gliogenesis peaks after E16 and continues past birth, which typically occurs at approximately E19. Glia additionally migrate into the six cortical layers and the layer of white matter (WM) immediately below.

The precise molecular, cellular, and physiological mechanisms controlling the behavior of RGP migration, proliferation, and differentiation over the course of neurogenesis and the NGS have been heavily studied, but much remains unknown. It is known, however, that epidermal growth factor receptor, or EGFR, is influential during the NGS. EGFR is a cell receptor located in the surface of RGPs, which accepts the epidermal growth factor (EGF) ligand. Expression of the gene responsible for EGFR production is increased in the developing brain, and EGFR is known to affect gliogenesis [5, 8, 15, 17, 47]. However, its precise role remains unclear, with different studies finding it playing a role in glial production and others finding it important for neurogenesis and neuronal survival [1, 2, 12, 15, 17, 21, 25, 27, 47, 49, 51]. One research aim in this thesis will be to quantify how EGFR expression in RGPs influences the NGS and gliogenesis.

1.2 Mosaic Analysis with Double Markers

To generate data describing RGP behavior during the NGS, a cell labeling technique called mosaic analysis with double markers was utilized. Here, we will describe the general technique and illustrate the type of data it outputs. The specific dataset used to study RGP behavior during the NGS in this thesis will be described in Sec. 1.3.

1.2.1 Method

Mosaic analysis with double markers (MADM) is a cell labeling technique that enables the observation of cell lineages in living organisms (*in vivo*) [55]. The technique can target any population of actively dividing cells in a tissue, but for simplicity and consistency, we will describe the method as acting on radial glial progenitors (RGPs) in the embryonic cortex and tracking their neural and glial offspring. As previously mentioned, RGPs can undergo two types of divisions during the NGS: *proliferation*, dividing to produce two RGPs, and *differentiation*, dividing to produce neurons or glial cells that do not themselves divide further. Differentiation can be *asymmetric*, producing an RGP and a differentiated cell, or *symmetric*, producing two differentiated cells. Figure 1.4a illustrates these division types.

In performing the MADM technique, an RGP is genetically modified at a chosen induction time. This modification activates red and green fluorescent markers in the two daughter cells arising from that RGP, respectively; thus, MADM labeling is only observable after the RGP divides. Any further offspring in the lineages of the two initial daughter cells retain the red and green labeling. The technique operates sparsely, meaning that only a small number of individual cells in a tissue will receive the MADM modification, and thus the number of observable red and green cells will be limited. The level of sparsity or density of cell labeling can be controlled at the induction step of the process. Fig.1.4b shows a simplified depiction of MADM in the NGS, starting with a single RGP

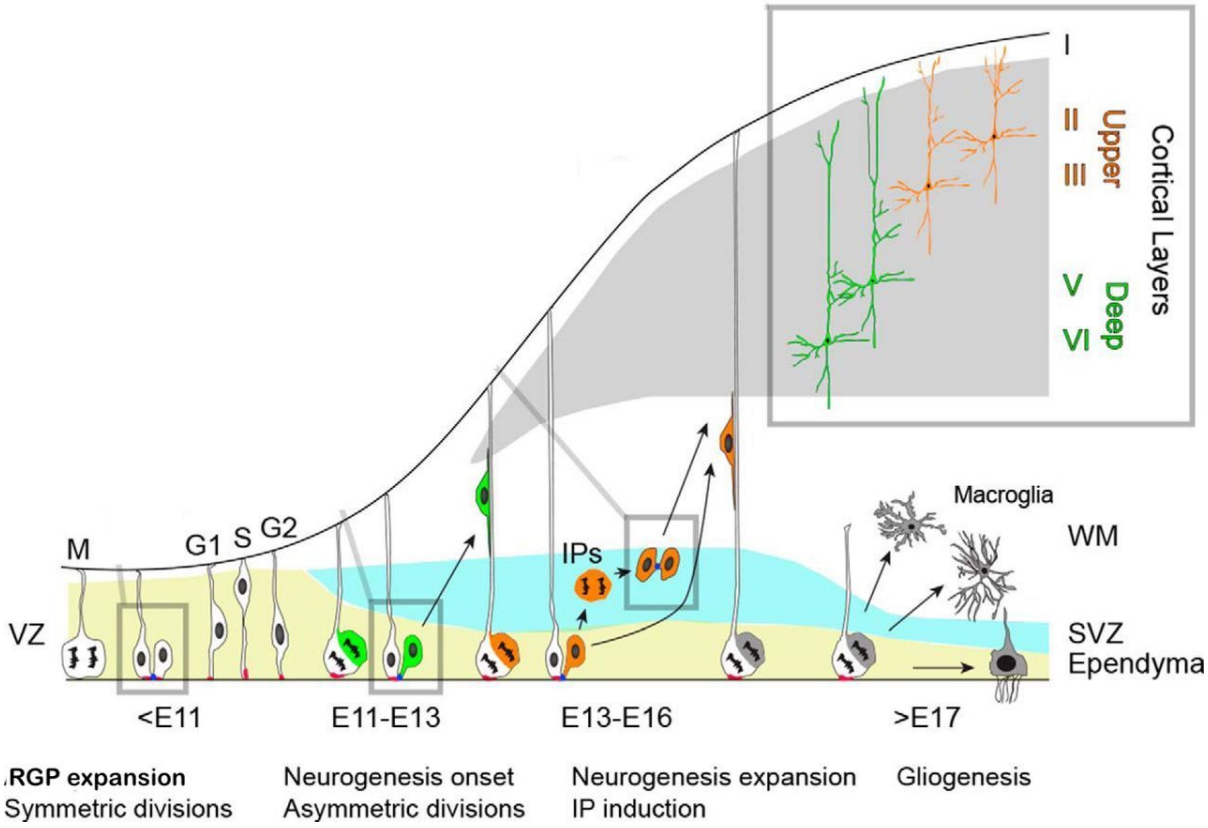


Figure 1.3 Diagram of cortical development from RGP divisions over time and space. The RGPs begin in the ventricular zone (VZ) prior to embryonic day 10, where they proliferate to expand the number of RGPs in the population. Neurogenesis begins around E11, when RGPs begin to differentiate asymmetrically and produce one neuron and one RGP per division. As neurons separate from their ancestor RGP, they travel radially outward along a long extension of the RGP cell body and begin settling into layers. The formation of the six cortical layers occurs with the deepest layers first, followed by more superficial layers. After ≈E16-E17, about 10-20% of RGPs stop neural production and instead produce glia (denoted 'macroglia'). Adapted from "Neural stem cells to cerebral cortex: emerging mechanisms regulating progenitor behavior and productivity" by N. D. Dwyer, B. Chen, S. Chou, S. Hippenmeyer, L. Nguyen, H. T. Ghashghaei, 2016, *Journal of Neuroscience, 36*, p. 11395. Copyright 2016, The Authors.

induced at some initial time in panel 1, continuing with the proliferation of red and green RGPs in panel 2, neurogenesis in panel 3, and the onset of gliogenesis during the NGS in panel 4. Panel 5 shows the completion of the process, where all RGPs have fully differentiated into neurons and glia and no RGPs remain. We can consider this final panel as the point at which the cortex is fully developed and no more neurons and glia are being produced. The sparsity of the MADM-induced RGPs in the tissue results in red and green cells being distinctly identifiable when viewed against unlabeled background cells (depicted as the white cells in the figure).

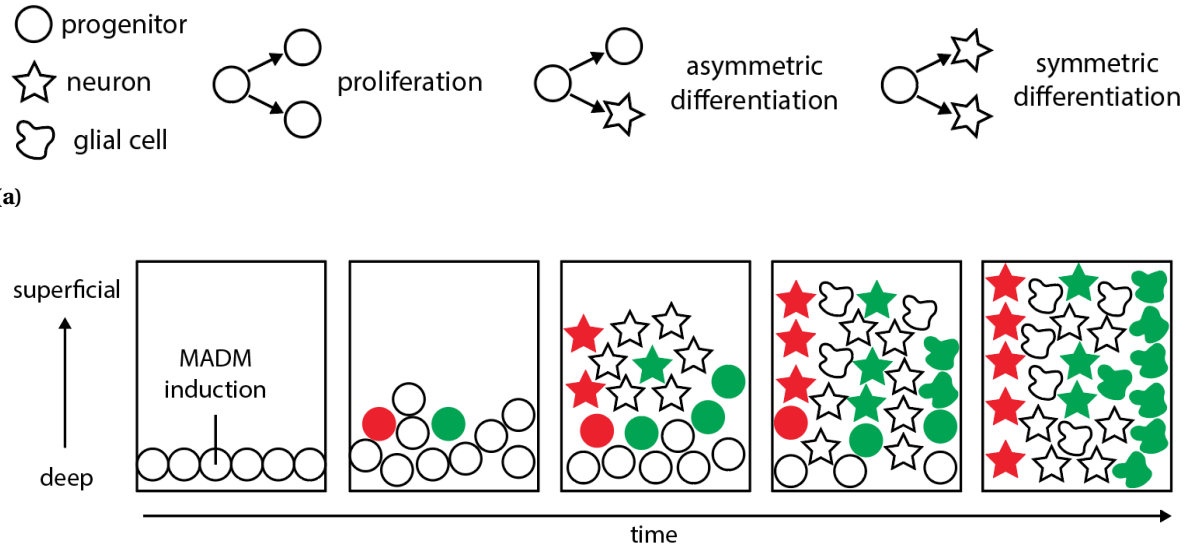


Figure 1.4 (a) Types of divisions an RGP can undergo to form two daughter cells. Producing more RGPs is called proliferation, and producing neurons or glia is called differentiation. (b) Single cell lineage (clone) tracked over the NGS with MADM. The first division of the MADM-induced RGP in Panel 2 initializes red and green labeling in each of its daughter lineages, respectively, which is retained in all neural and glial offspring in the following panels. White cells represent those not labeled with MADM or not arising from the lineages of MADM-induced RGPs. The red and green neurons and glia in Panel 5 are thus observed as arising from the single initial MADM-labeled RGP in Panel 1.

Specifically, the activation of the red and green fluorescent markers is accomplished by targeting genetic recombination in a designated segment of DNA [55]. This targeted recombination can also be designed to simultaneously modify other genes, meaning that the genotypes of the red and green daughter lineages can be distinct from one another. For instance, this may be used to designate the red lineage as 'wild type' cells, having otherwise unaltered DNA, and the green lineage as genetic 'knockout' cells, having a particular gene deleted. This type of experiment results in genetic 'mosaic' organisms, which have a mixture of red wild type and green knockout cells. Comparison of the red

and green lineages can give insight into the particular knocked out gene's function.

1.2.2 Data output

The *in vivo* nature of the MADM technique implies that the data it produces represents cell behavior in the dynamic environment of a developing tissue more accurately than *in vitro* data, which observes the behavior of cells removed from the living tissue [41, 48]. However, since MADM operates on living organisms *in vivo*, the offspring of MADM-labeled cells can only be observed once; the organism must be sacrificed and the tissue segmented to visualize the red and green RGP offspring. Thus, the red and green cells observed represent a single temporal snapshot of the total and types of cells present at the specific time of observation. The data output (counts of total red and green cells) is therefore highly dependent on the chosen observation time as well as the induction time. For instance, the red and green cells in each panel in Fig. 1.4B represent the data that would be gathered if observation occurred at that particular time, given the initial MADM labeling time shown. We see that earlier observations would show more red and green RGPs and fewer differentiated neurons and glia, while observations after the cortex is finished developing (Panel 5) would show only differentiated cells and no RGPs. On the other hand, if MADM induction occurred later in the process, we would not see any of the neurons or glia that already differentiated in previous panels since these cells do not divide and therefore cannot activate the red and green labels. That is, in Fig. 1.4b, if the initial MADM induction occurred in an RGP in Panel 3 and observation occurred in Panel 5, we would not be able to count any of the red and green neurons produced prior to Panel 3, and instead would only count the neurons and glia that appear for the first time in Panels 4 and 5. In summary, the data generated by MADM measures the offspring of RGPs produced the window of time between induction and observation, and it cannot account for offspring arising before induction or after observation.

MADM can produce two types of data as output, depending on the method for counting the red and green cells. First, the total red and green neurons and glia in an area of the tissue can be counted, producing data representing cell densities. In this case, the observed cells are not assumed to arise from the same initial RGP, but would be considered as representing the neural and glial output in the RGP population as a whole. Alternatively, if there is some spatial restriction on the location of a cell's offspring and MADM labeling is sparse enough, it can be assumed that all red and green neurons and glia located near each other during observation arose from one single initial RGP. This gives a type of data called *clonal data*, where each 'clone' in the dataset represents the offspring of an individual RGP in the population. Clonal data can provide particularly valuable insight into the uniformity or variability of cell proliferation and differentiation patterns across the population, since it counts the total and type of cells that each individual RGP produces. That is, in contrast with the first type of MADM data that measures cell densities in an area of tissue, clonal MADM data

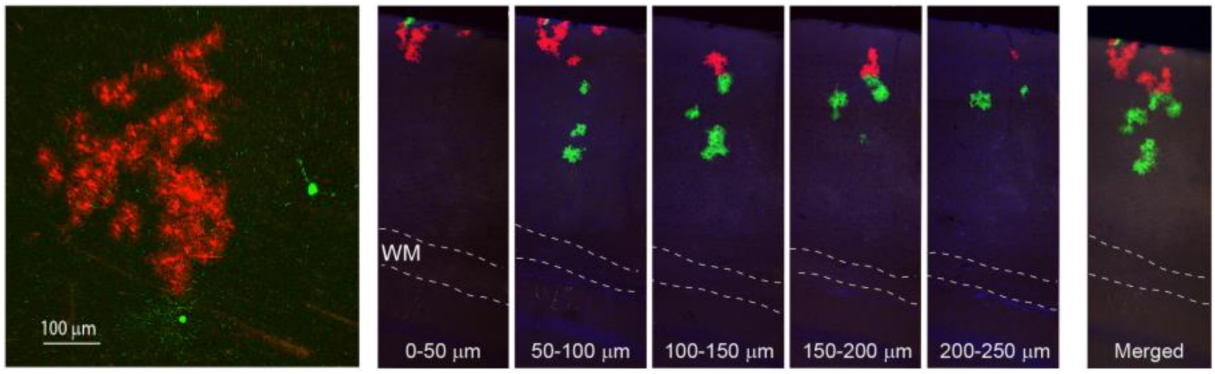
measures RGP offspring at a finer resolution, at the level of individual cells. Our dataset is clonal, and we describe this dataset in the following section.

1.3 Clonal Data from the NGS

The data we consider was generated by the Ghashghaei Lab at NC State (http://tghashghaeilab.org) using MADM induction of RGPs in the embryonic mouse cortex. Mouse embryos received MADM modification at one of three successive embryonic time points: E15.5, E16.5, or E17.5, corresponding to 12PM on the 15th, 16th, and 17th days of gestation, respectively. Additionally, each mouse was designated as being wild type (WT) or knockout (CKO), where WT refers to mice in which both the red and green MADM sublineages had no further genetic modifications, and CKO refers to mice in which the green MADM sublineage had the gene for epidermal growth factor receptor (EGFR) removed. As previously described in Sec. 1.1.2, EGFR is important for the production of glia, and removal of the gene for EGFR suppresses gliogenesis. Clones coming from CKO mice thus tended to exhibit very few green glia and an overproduction of red glia (Fig. 1.5a).

All observations were taken at postnatal date P30 when actively dividing RGPs are no longer present, corresponding to the final panel in Fig. 1.4b. At this time, each mouse cortex was sliced into 50μm thick cross-sections from front to back. For each mouse, the cross-sections were indexed in order and mounted on individual microscope slides, then manually examined for red and green neurons and glia. Cells oriented in vertical columns from deep to superficial cortical layers and located in the same area of the cortex across neighboring cross-sections were considered to have originated from the same initial RGP in a given mouse, due to the outward radial movement of cells coming from individual RGPs (see Sec. 1.1.2). The cells in vertical columns thus gave the total red and green neurons and glia arising from individual RGP lineages (clones). Fig. 1.5b shows the presence of one clone across five separate cortical cross-sections. The total neurons and glia produced by this clone is the total neurons and glia counted across all five cross-sections. We also see that when the cross-sections are merged, the MADM cells appear as one cohesive group. Several clones were counted for each mouse. Table 1.1 summarizes how many mice were labeled under each time and genetic type, as well as how many clones were counted in each mouse. Overall, the full dataset contains a total of 550 clones coming from 27 individual mice.

Each clone is represented by a vector $C \in \mathbb{Z}^{4+}$ corresponding to the total red glia, total red neurons, total green glia, and total green neurons arising from one MADM-labeled RGP. Several other features are recorded for each clone: MADM induction time (E15.5, E16.5, or E17.5), the type of clone (N, G, Mix, representing clones producing only neurons, only glia, or a mix of the two, respectively), the location of the observed cells in the six cortical layers (deep, superficial, or all layers), the genotype regarding the presence or absence of the gene for EGFR in the green MADM sublineage (WT or CKO), and its mouse of origin (indexed 1-18 for WT mice, 1-9 for CKO mice, with ordering from



(a) **(b)**

Figure 1.5 Visualization of MADM cells. (a) Cells in a single CKO clone, showing overproduction of red glia. (b) MADM offspring produced from a single clone across five cortical cross-sections. The merged cross-sections show the red and green cells in the entire clone all together. (Image used with permission from Ghashghaei Lab)

Table 1.1 Total mice and clones per mouse for each MADM induction time (E15.5, E16.5, or E17.5) and genotype (WT or CKO). The mice having the same induction time and genotype are called replicates.

	E15.5 WT	E16.5 WT	E17.5 WT	E15.5 CKO	E16.5 CKO	E17.5 CKO
Mouse 1	14	23	19	16	18	12
Mouse 2	28	27	20	14	25	17
Mouse 3	15	15	15	15	14	26
Mouse 4	22	23	42	–	–	–
Mouse 5	18	12	44	–	–	–
Mouse 6	–	25	8	–	–	–
Mouse 7	–	23	–	–	–	–
Total Clones	97	148	148	45	57	55

Table 1.2 Full features of three clones from the dataset. The values in the columns for Red G, Red N, Green G, and Green N make up the 1×4 vector C for each clone.

Time	Red G	Red N	Green G	Green N	Clone Type	Depth	Genotype	Mouse ID
E15.5	0	6	0	2	N	superficial	WT	1
E16.5	79	3	0	2	Mix	all layers	CKO	5
E17.5	0	0	7	0	G	deep	WT	17

earliest to latest induction times). Table 1.2 shows three examples of clones with varying features from the dataset.

1.4 Mechanisms of clonal behavior

1.4.1 Deterministic versus stochastic

In this thesis, our goal is to use mathematical models and techniques to study clonal data from the NGS and identify what rules of proliferation and differentiation RGPs follow when producing glia. It is also important to define these rules as either *deterministic* or *stochastic*. Under a deterministic mechanism, RGPs have specific predispositions that control the fate of their offspring. On the other hand, RGPs operating under a stochastic mechanism all have equal likelihood of different fates, and the choice of specific fate for each RGP is a random event [29, 45]. For instance, under deterministic rules, we could define a specific 20% of the RGP population as predisposed to producing glia, while under stochastic rules, every individual RGP has a 20% chance of producing glia after neurogenesis (Fig. 1.6). Thus, the two different mechanisms differ in how they control the fate of cells produced by RGPs. Controlling cell fates is important for producing neurons and glia in the proper balance. Hence, if we can determine whether the mechanism of control is deterministic or stochastic, we may gain insight into how that mechanism may be disrupted and lead to diseases characterized by improper neural and glial production.

It is important to note that for our example in Fig. 1.6, we would not be able to distinguish the two rules mathematically from the simple observation that ≈20% of RGPs in the population produce glia. That is, the deterministic and stochastic rules described above can produce the same output at the level of the cell population. However, if the RGP population is perturbed, for instance with a genetic deletion, the effect of that perturbation may be analyzed to characterize the population as deterministic or stochastic. If there is a consistent effect across the entire population, this would point to a stochastic mechanism, whereas if different subsets of RGPs respond differently to the perturbation, it is more likely that these RGPs have different predispositions and are deterministic.

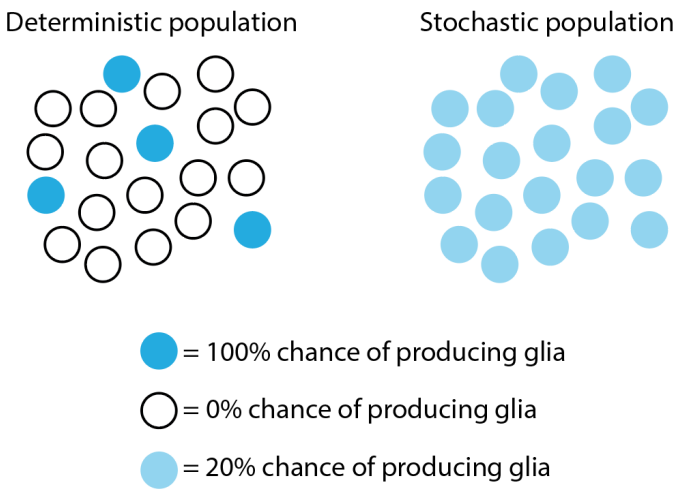


Figure 1.6 Illustration of deterministic versus stochastic fate specification in a progenitor population. In the deterministic case, RGPs are predisposed to specific fates, shown as 20% of the RGPs (4 out of 20) being both capable and guaranteed of producing glia. In the stochastic case, all RGPs have a 20% chance of producing glia. Hence, despite the different mechanisms of fate specification, we would expect 20% of RGPs in the population to produce glia.

It is necessary to have clonal data for this type of experiment, since the response of individual RGPs must be assessed rather than the effect of the genetic knockout on the total population of RGPs. Many studies have used genetic knockouts, cell reprogramming, and cellular responses to external stimuli to classify clonal populations of stem and progenitor cells as having deterministic or stochastic behavior [10, 19, 20, 50, 54].

In particular, it has been observed that deterministic mechanisms are often characterized by asymmetric cell divisions and a pairing or correlation between the behavior of the two daughter lineages of individual cells [16, 34, 44, 54]. Thus, clone sizes and symmetries may be used to classify cell division mechanisms as deterministic. The development of the MADM technique and its ability to delineate daughter lineages by color, as well as its ability to perturb one cell population with a genetic knockout and keep the other unchanged, has resulted in excellent opportunities for analyzing clonal populations to identify deterministic patterns. Recent studies have used MADM modification of progenitor cells in a developing tissue to measure the predictability of each progenitor's offspring, the relationship between the fates of the two red and green daughter lineages, the mobility of progenitor offspring as the tissue develops, the uni- or multipotency of the progenitor population, and the effect of mosaic genetic knockouts [7, 14, 16, 33, 52].

1.4.2 RGP behavior from clonal MADM data

We next highlight two recent studies with the most direct application to the research presented in this thesis. Both studies used the MADM technique to obtain clonal data from RGPs in the developing mouse cortex and attempted to analyze the resulting clones for deterministic or stochastic patterns. However, both focused on neurogenesis rather than gliogenesis and the NGS.

First, in a study by Gao et al. [16], clonal MADM data was collected in the embryonic mouse cortex with induction times between E10.5 and E13.5. These induction times coincided with the bulk of neurogenesis, hence the clones observed predominantly produced large numbers of neurons. Similar to our data, observations were performed between P10-P30 when only neurons and no RGPs appeared, and the analysis thus focused on studying the total capacity for neural output in the clonal population throughout all of cortical development. Clone sizes (total red and green) and symmetries (red versus green) were examined, revealing predictable patterns of clonal sizes and consistent ratios of red to green cells. Thus, it was argued that neurogenesis is a deterministic process. We applied several of these methods to our own data, in particular using Gaussian mixture models to represent clonal output, which we describe in more detail in Chapter 5.

Second, Picco et al. [40] used staggered induction and observation times in different mice for the purpose of gathering time-series clonal data of embryonic neurogenesis, with the goal of constructing a stochastic model of RGP behavior. Fig. 1.7 shows the different MADM induction and observation times used for individual mice in their study. We note that this method of gathering clonal MADM data differs from ours, since clones were observed prior to birth, before the cortex was finished developing. A drawback in this method is that MADM observations performed prior to birth introduce uncertainty in cell counts, since a cell's morphology may not yet be clear enough to indicate whether it is an RGP or a neuron. Their study dealt with this uncertainty by assuming all observed cells were RGPs. However, we deem this an oversimplification that makes the MADM data less informative, particularly when attempting to use clonal output to characterize division behavior as deterministic or stochastic. Additionally, this method requires many mice to generate a significant number of clones since clone sizes are smaller at earlier time points. The cost of such an experiment may not be practical, especially considering the uncertainty in the observed data. Still, the theory of branching processes, which was utilized in this paper for stochastic modeling of the RGP population, is useful for attempting to model our data from the NGS. This method was also used to model neurogenesis in Slater et al. [48]. We explore this method in Chapter 4.

Apart from drawing on the techniques from these two studies, we utilize combinatorial, statistical and machine learning methods to analyze the clonal MADM data and understand the progression of the NGS. First, in Chapter 2, we use self-organizing maps (SOMs) to compare the sets of clones from each individual mouse (Table 1.1). This is performed to explore the MADM technique itself, namely to judge the consistency in the data it produces across different organisms. Following this

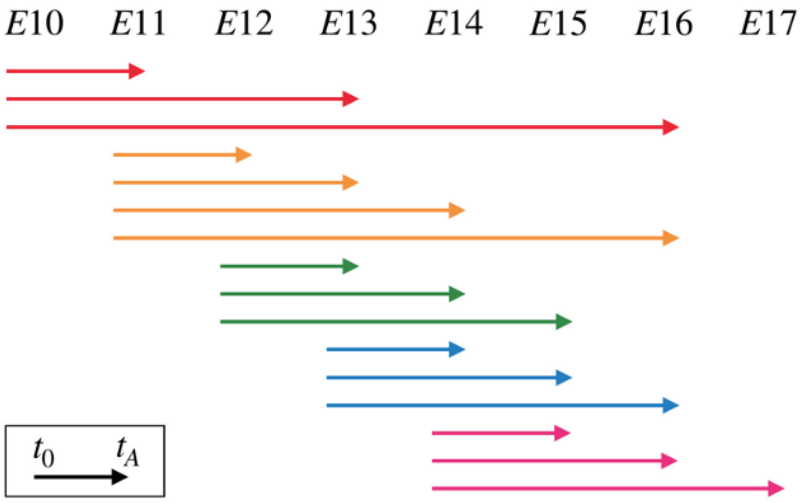


Figure 1.7 Diagram of MADM induction and observation times of RGPs during the NGS for different mice, as performed in the study by Picco et al. (Image source: [40]).

investigation, we shift our analysis to the level of clones pooled together across different mice within the population. The remainder of this thesis deals with clones at this pooled level, thus our analysis is primarily of the distribution of clone sizes (total glia per clone) across the entire population. The analysis we perform is intended to investigate whether clones during the NGS produce glia deterministically or stochastically.

In Chapter 3, we use the theory of branching processes to test the distribution of glia per clone for stochastic patterns. The results of this chapter lead us to hypothesize that two subpopulations of RGPs exist, one of which behaves stochastically. Thus, in Chapter 4, we use clustering and statistical analysis to identify unique behaviors among different subsets of the clones in the data. This work supports the hypothesis of separate subpopulations of RGPs, and indicates that the other subpopulation behaves deterministically. In Chapter 5, we implement the methods from Gao et al. described above, using discrete multi-Gaussian mixture models to test the distribution of glia per clone for the deterministic hallmarks observed during neurogenesis [16]. The work in these three chapters leads us to propose a set of rules describing gliogenesis in which two subpopulations of RGPs exist: NGS-RGPs, which behave deterministically, and G-RGPs, which behave stochastically. Finally, in Chapter 6, we simulate deterministic and stochastic divisions of RGPs according to these rules and test how well our simulated clones agree with the true clonal MADM dataset during the NGS.

CHAPTER

2

MOUSE-LEVEL ANALYSIS: SELF-ORGANIZING MAPS

As described in Sec. 1.3, we consider a set of clonal MADM data gathered during the NGS representing the total neurons and glia produced by individual radial glial progenitors (RGPs). This data set was generated from MADM labeling of RGPs in embryonic mouse cortices, and for each time point considered, multiple mice received MADM labeling: for the WT group, this was 5 mice at E15.5, 7 mice at E16.5, and 6 mice at E17.5 (Table 1.1). It is customary in biological experiments to test multiple genetically identical organisms under the same conditions to help control for the natural variability between different organisms; this is referred to as gathering replicate data. If we assume that apart from their natural variability, a group of mice labeled at the same time point (replicates) are essentially interchangeable, we can think of the clones coming from each mouse to be random samples all drawn from the same distribution despite coming from separate organisms. Under this assumption, it is reasonable to pool together the clones coming from all replicates and analyze these clones as a single dataset. That is, in Table 1.1, we would consider pooling the clones in each column, which come from separate mice having the same induction time and genotype (e.g., E15.5 WT).

However, the MADM technique itself introduces uncertainties in the timing of labeling. As a consequence, the assumption that replicates are at identical developmental stages at the time of MADM labeling may not hold. In this chapter, we discuss the sources of uncertainty in the timing of

MADM labeling, then describe how we can use self-organizing maps (SOMs) to better understand the level of similarity or variability in the clonal data coming from replicate mice.

2.1 Comparing MADM data between mice

2.1.1 Uncertainty of mouse embryo ages

The MADM technique operates on RGPs *in vivo*, in the actual developing mouse embryo, allowing an accurate representation of developmental dynamics. However, its *in vivo* nature prevents us from knowing the precise age of the embryos, and we must instead use an approximated age. An embryo originates when a breeding takes place between mice. Pairs of mice breed overnight, but the actual time of the breeding is unknown since the mice are not observed at all hours. If the observer arrives the next morning to see that a breeding has taken place, the onset time of the embryo, E0, is taken to be at midnight of the previous night. The actual onset time of the embryo could have occurred any time between when the observers leave on the previous day and when they arrive again the following morning, a window of ≈12-16 hours. Thus, two embryos considered to have originated at the same time may be as much as 16 hours apart in age. This implies that if MADM labeling occurs in the two embryos at, say, noon on the 15th day after the breeding takes place (E15.5), both embryos are assumed to be the exact same age at the time of labeling and hence are assumed to be at the same stage of development. In this example, these embryos would be E15.5 replicates in our data.

Since mouse gestation is only ≈ 20 days, and since the clonal population in the cortex rapidly changes during the NGS, a difference in age of 16 hours is enough to produce significantly different clonal distributions between two mice assumed to be replicates. In particular, differences would appear in their proportions of clone types - clones containing only neurons, only glia, or a mix of the two - since the NGS progresses from neurogenesis to gliogenesis over time. We should certainly expect a degree of variability when comparing replicates' clones due to randomness in MADM sampling and natural biological variability. However, large differences in the proportions of clone types between replicate mice more likely indicates uncertainty in the true age of the embryos at the time of MADM induction. We may for instance observe that the clones in a particular E15.5 mouse are more similar to those coming from E16.5 mice, rather than the other E15.5 mice assumed to be replicates, and thus the particular E15.5 mouse may have been older than presumed at the time of MADM labeling.

2.1.2 Dealing with MADM uncertainty in the NGS

Since MADM operates *in vivo*, it is not possible to reduce the uncertainty in embryonic age at the time of MADM labeling by observing the embryo in real time. Instead, under the assumption that same-age mice should be at similar stages in cortical development and thus show similar

behavior at the cellular level, we can consider two mice to have been labeled at approximately the same age if their clones exhibit similar features when analyzed quantitatively. Specifically, as cortical development progresses from E15.5 to E17.5, we would expect a decrease in neurogenic clones and an increase in gliogeneic clones. The percentage of each mouse's clones containing only neurons (%N), only glia (%G), or a mix of both (%Mix) should thus be indicative of MADM induction time. Additionally, clone sizes (total neurons and glia per clone) should distinguish each induction time, since later inductions track only a subset of an RGP's overall proliferation and differentiation behavior.

In Fig. 2.1a, we find the clone type percentages for all clones pooled between replicate mice. We clearly see the expected pattern of %N decreasing and %G increasing over time, and each time point can be distinguished from the others by its clone type percentages. However, if this percentage is calculated for each individual mouse's clones, significantly more variability is observed in the breakdown of %N, %G, %Mix for each replicate mouse as seen in Fig. 2.1b. For instance, Mice 3, 9, and 18 all received MADM induction at different time points, but their clone type percentages are nearly identical. Based on our assumption that the clone type percentages should be characteristic of the true age of the embryo at induction, we may consider the case that these three mice were closer in age at MADM induction than their time point label suggests. Thus, we propose to address the uncertainty in embryonic age at the time of MADM labeling by comparing each mouse's population of clones. We use self-organizing maps to aid clonal comparisons between mice.

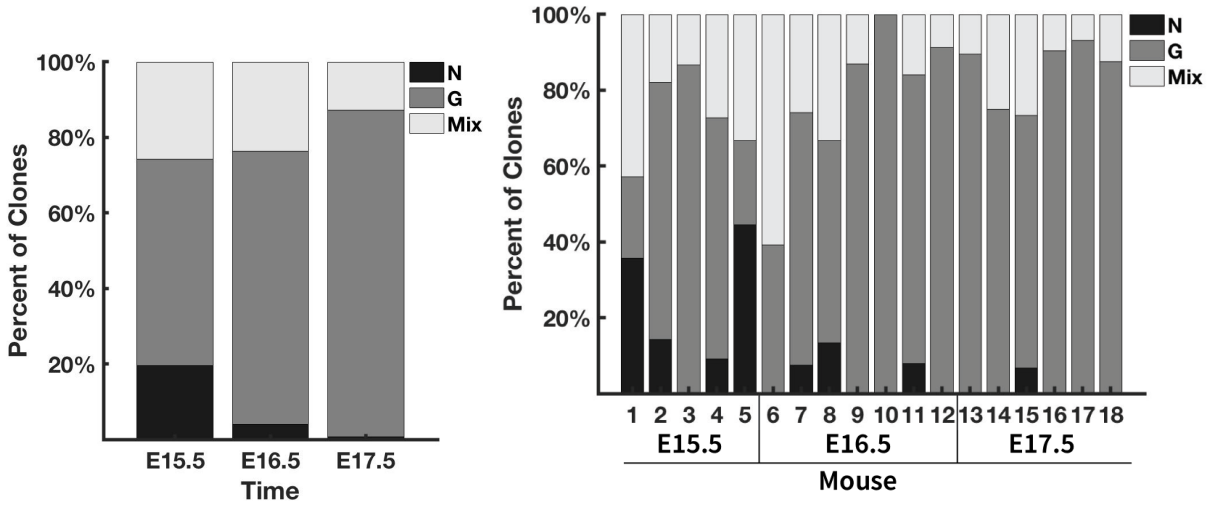


Figure 2.1 Percent of clones containing only neurons (N), only glia (G), or both (Mix). In (a), this percentage is calculated for the clones pooled between all WT mice per time point, whereas in (b), the percentages of clone types are calculated for the clones coming from each individual mouse.

2.2 Self-Organizing Maps

2.2.1 Description of SOM clustering technique

Clustering is a data analysis technique used to partition data into separate groups, or clusters, such that data elements in the same cluster are similar to one another and less similar to data elements in different clusters. Data clustering is a particularly useful method for analyzing or visualizing the relationships present in complex or high-dimensional data, since it provides an easily interpretable way to compare different data elements. Depending on the format of the input data and the information wanted from the partition, different clustering algorithms may be suitable.

We implement clustering using self-organizing maps (SOMs). As with other clustering algorithms, SOMs sort data elements that are similar to one another into clusters together. However, an SOM also provides additional information in the clustering by arranging the clusters into an ordered grid of nodes, typically one- or two-dimensional, such that the data sorted into neighboring nodes is more similar than data sorted into non-neighboring nodes. This makes an SOM grid not only a clustering diagram, but also a form of similarity graph [24]. Additionally, this distinguishes SOMs from the k-means clustering algorithm [30], which places data elements into clusters with no particular ordering relative to one another.

Figure 2.2 illustrates a simple representation of data sorted into a 1×5 SOM. Here, each data element is a 2×1 vector, shown plotted as points in the 2-D data space (xy-plane). The SOM forms five clusters in the data, one for each node of the SOM, and clusters together the elements that are closest in the xy-plane. We outline the data points placed into clusters together in dashed circles. These groups of data points are mapped to the 1×5 grid of nodes in the SOM, such that the clusters closer in the xy-plane are in neighboring nodes. In the figure, the outlines of the clusters in the xy-plane are color-coded according to the node in the SOM to which they belong. We notice that in this example, the x values of the data elements sorted into each node increase monotonically when traversing the nodes in the map from left to right. If we examined the y values of the data elements sorted into each node, we would not see the same monotonic increase. In this example, we would say that the SOM has 'sorted' the data according to the x values. Generally, because the SOM clustering operates on the distances between elements, it partitions the data into separate clusters according to the component(s) of the data with the greatest variance across the set of data elements.

2.2.2 Implementation

The following steps are taken in a one-dimensional SOM with q clusters and R input data elements (vectors) v_r, $r = 1, ..., R$ each of dimension D:

For a 1-D SOM with q clusters, sorting r data vectors v_r each of dimension D:

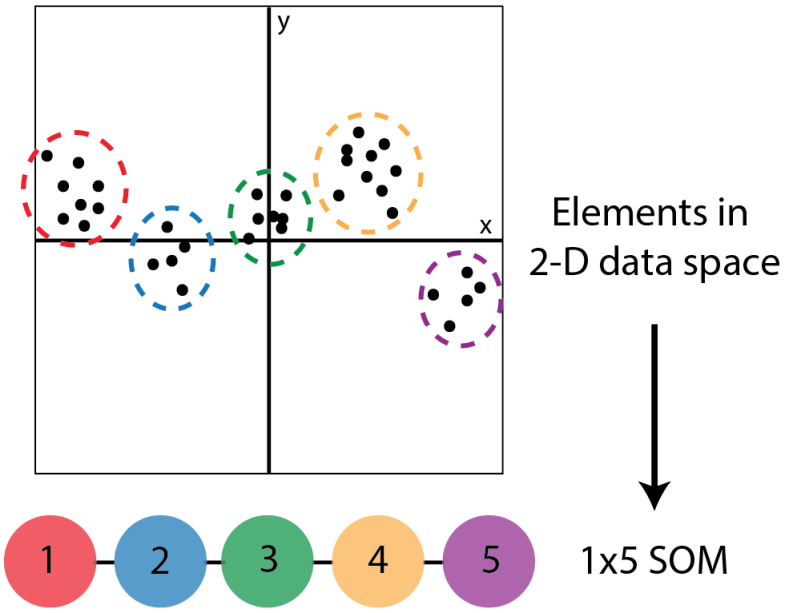


Figure 2.2 Simple representation of data sorted into five clusters in a 1×5 SOM. Elements are clustered together according to their closeness in the data space. The SOM then sorts these clusters into the grid of five nodes so that the data points in neighboring nodes (1 and 2, 2 and 3, etc) are closer in the data space.

1. Initialize each cluster with a weight vector w_j, $j = 1, ..., q$, where each weight has dimension D

2. For each input vector v_r, find its distance from each weight vector

$$d(v_r, w_j) = ||v_r - w_j||_2, \quad j = 1, ..., q$$

   and select the weight w_j^* with the smallest distance to v_r as the 'winner'

3. Update the weight of the winning node w_j^* as

$$w_j^{*\text{new}} = w_j^* + \eta(k)(v_r - w_j^*)$$

   where $\eta(k)$ is the learning rate function at iteration k and depends on an initial learning rate η_0 and learning rate decay parameter a, $0<a<1$:

$$\eta(k) = \eta_0 e^{(-k*a)}.$$

4. For each data vector v_r, find the neighbors w_n of its winning node w_j^* by finding any nodes

within a Gaussian neighborhood w_j^*, where the variance of the Gaussian neighborhood function is

$$\gamma(k) = \gamma_0 e^{(-k*b)}$$

for initial variance $\gamma_0 > 0$ and decay parameter $0 < b < 1$, then update the neighbors according to

$$w_n^{\text{new}} = w_n + \gamma(k)(v_r - w_n)$$

5. Repeat steps 2-4, starting with the updated weights, until some stopping criteria is met, for instance a maximum number of iterations k.

We utilize the MATLAB SOM toolbox in all SOM implementations presented. Other than the data itself, the clustering algorithm takes as input the number of iterations and the parameters a, b, η_0, and γ_0.

2.2.3 Interpreting output

After all data elements are assigned to a cluster, the components of those elements can be visualized for each cluster. For instance, for the clustering in Fig. 2.2, the average x component of the elements in each cluster can be calculated, then visualized in order of clusters 1-5. Since x values increased going from left to right in the SOM geometry, we would observe a positive correlation between x value and cluster number. However, if we calculated the average y component from each cluster and ordered these by clusters 1-5, we would not see a clear correlation between the y value and the cluster number. Thus, visualizing component averages and standard deviations can indicate which components the SOM sorted on, in the case when the input data is higher dimensional and cluster orderings cannot be easily seen by plotting the data itself.

If the data elements have other features which were not used as input to the SOM (that is, they were not used in calculating the distance between data elements during SOM clustering), then the distribution of these features among clusters can be observed as well. If a certain feature is concentrated in clusters at one end of the SOM, this can indicate a correlation between that feature and the component on which the SOM sorted. This correlation can be positive or negative depending on whether the features are concentrated at the end of the SOM with high or low component values.

2.3 Application of SOM to mouse-level NGS data

2.3.1 Data representation for comparing WT mice

To compare mice by their clonal populations, namely by their clone types and sizes, we chose to represent each mouse's clones in a two-dimensional normalized histogram. An example is shown

in Fig. 2.3. In this histogram, eight bins were created in two dimensions to correspond to total neurons and total glia per clone, respectively, resulting in an 8×8 grid pattern for each mouse. One outlier clone was removed from Mouse 6 (E16.5) which contained 164 glia, as the next largest clone contained only 74 glia. All analysis in this chapter was performed on the set of WT mice only.

To help the clustering distinguish between different clone types, the N and G clones (containing only neurons or only glia, respectively) were not placed into bins with Mix clones; the leftmost bin along each axis was designated for clones having only neurons or only glia, and any bins on the interior of those two zero axes corresponded to Mix clones of various sizes and ratios of neurons to glia. To determine the bin width on the interior of the grid, the Freedman-Diaconis rule was used in each dimension: Bin Width=$2\text{IQR}(X)/\sqrt[3]{n}$, where $\text{IQR}(X)$ is the interquartile range of the vector X of total neurons or total glia in all WT mixed clones, and n is the number of WT mixed clones [11]. This gave bin widths of 1.5683 for neurons and 10.6496 for glia, which we rounded to 2 and 11, respectively. Accommodating all clone sizes using these bin widths added seven bins in each direction for the Mix clones, resulting in the final 8×8 binning partition. For a given mouse, the histogram bar heights are calculated by finding the percent of the mouse's clones falling into each bin.

In Fig. 2.3, the tallest bar thus indicates that ≈30% of that mouse's clones fell in the bin corresponding to 0 neurons and between 12-22 glia. We also see that ≈6% of its clones were neuron only and contained 3-4 total neurons. These 8×8 grid patterns were reshaped into 1×64 vectors V_T^j, indexed for each mouse $j = 1, ..., 18$ and its respective MADM induction time T, where T=E15.5 for $j = 1:5$, T=E16.5 for $j = 6:12$, and T=E17.5 for $j = 13:18$. The vectors V_T^j were used as inputs for the SOM to cluster. For simplicity, since each grid pattern/vector represents one of the 18 mice considered, we will generally refer to the SOM operating on 'mice.'

2.3.2 Generation of multiple datasets via sampling

The goal in comparing the mice based on their 8×8 grid patterns was to see how often E15.5 mice were placed in the same cluster as other E15.5 mice, and so on for E16.5 and E17.5. To achieve a balanced comparison between the time points, we considered groups of fifteen mice, with five sampled from each time point. A total of $j = 18$ mice comprise the WT data set, and running through all folds of sampling given five E15.5 mice, seven E16.5 mice, and six E17.5 mice gave a total of $k = \binom{5}{5}\binom{7}{5}\binom{6}{5} = 126$ combinations of mice. We will denote the indicator function

$$\mathbf{1}_{\text{k}}(j) = \begin{cases} 1 & \text{if mouse } j \text{ is in combination } k \\ 0 & \text{if mouse } j \text{ is not in combination } k. \end{cases}$$

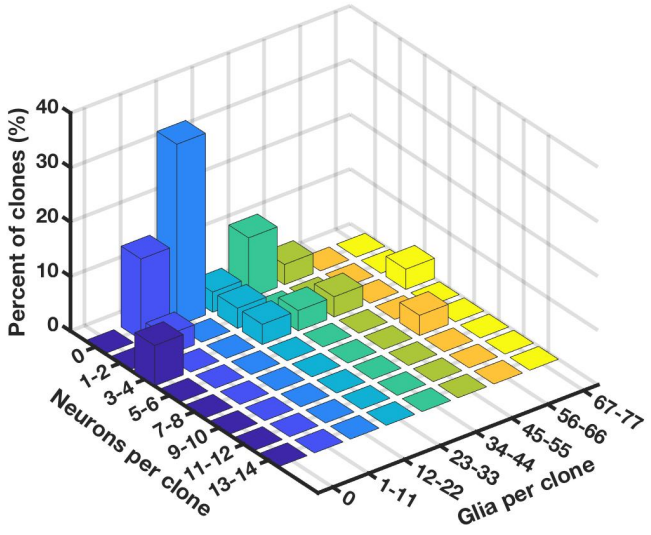


Figure 2.3 Example grid pattern for an E16.5 mouse. An equal number of 8 bins were set in each direction, where the first bin along each axis contains clones with no neurons or no glia. The Mix clones are distributed into the remaining 7 bins in each direction based on their total neurons and glia. For this mouse, the most represented bin is that for 0 neurons and between 12-22 glia, containing ≈30% of its clones.

2.3.3 SOM usage

We opt to cluster our mice via a 1×3 SOM. The implementation of an SOM of this size is useful for several reasons. First, regarding using SOMs themselves, unsupervised learning makes sense given our data because we cannot assume the labeled time point of each mouse is accurate. Instead, we want to find how similar different mice are based on their clonal distributions, regardless of their label. Additionally, an SOM does not require separate sets of training, validation, and testing data like in supervised learning techniques, which would be impractical given a data set with only 18 elements (mice) to compare. Regarding the dimension of the SOM, we would expect the mice to sort into clusters partly based on their %N, %G, and %Mix clones, informed by the bar heights in the grid pattern corresponding to each of these clone types. Thus, based on these percentages shown in Fig. 2.1a, we could expect for instance that the mice sorted clusters 1, 2, and 3 would be ordered from lowest to highest %G clones. Since these percentages correlate with time point, we would expect E15.5, E16.5, and E17.5 mice to sort more often into clusters 1, 2, and 3, respectively.

We note here that an SOM is unbiased in its left-right orientation. Hence, over the $k = 126$ clusterings, the SOM may sometimes designate cluster 1 as representing the group of mice with low %G in cluster 1, but other times this group of mice may be assigned to cluster 3. This difference could make it difficult to compare the results of the $k = 126$ clusterings, since we cannot simply

compare the mice sorted into clusters 1, 2, or 3 in each case. To make it easier to compare the cluster compositions across all clusterings performed, we bias the orientation of the SOM by initializing the weights of clusters 1, 2, and 3 as the centroids of the E15.5, E16.5, and E17.5 mice, respectively. This initialization biases the sorting of mice with a higher percentage of G clones into cluster 3 rather than cluster 1. We can therefore calculate how often E15.5 get sorted into cluster 1 versus clusters 2 or 3 and consider this as a measure of the accuracy of the clustering over the $k = 126$ combinations, and similarly for E16.5 mice in cluster 2 and E17.5 mice in cluster 3.

For each combination $k = 1, ..., 126$ of mice, clustering and analysis was completed using the following steps:

1. Calculate the mean value μ_T of the grid patterns among the five mice from each time point in combination k,
$$\mu_T = \frac{1}{5}\sum_{j=1}^{18} \mathbf{1}_{\mathrm{k}}(j) V_T^j, \quad T = E15.5, E16.5, E17.5 \tag{2.1}$$

2. Initialize the centers of each cluster in a 1×3 SOM as the mean values found from Eqn. 2.1, with $\mu_{E15.5}$, $\mu_{E16.5}$, and $\mu_{E17.5}$ initializing the centers of clusters 1, 2, and 3, respectively. This initializes the orientation of the SOM.

3. Cluster the set of vectors V_T^j, $j \in k$ with the SOM.

4. Record the SOM clustering placement for each mouse $j = 1, ..., 18$ in a 126×3 matrix C_j,
$$C_j(k,l) = \begin{cases} 1 & \text{if mouse } j \text{ is present in combination } k \text{ and is placed in cluster } l,\ l\text{=1, 2, 3} \\ 0 & \text{otherwise .} \end{cases}$$

5. Compare the cluster to which each mouse was assigned to its true time point label, for instance considering a 'correct' clustering as an E15.5 mouse placed in cluster 1, and an 'incorrect' clustering as an E16.5 mouse placed in cluster 1 or cluster 3. Record the SOM performance for each mouse j in the indicator function $\mathbf{1}_{\mathrm{C}}(j,k)$,
$$\mathbf{1}_{\mathrm{C}}(j,k) = \begin{cases} 1 & \text{if mouse } j \text{ is present in combination } k \text{ and is correctly clustered} \\ 0 & \text{otherwise.} \end{cases}$$

6. Compute the fraction of mice out of the fifteen that were classified correctly as the percent accuracy
$$A_k = \frac{1}{15}\sum_{j=1}^{18} \mathbf{1}_{\mathrm{C}}(j,k). \tag{2.2}$$

After repeating these steps over all $k = 126$ combinations of mice, we then calculated the percent of times $\beta_j(l)$ that each mouse j was sorted into each cluster $l = 1, 2, 3$ when included in a combination k,

$$\beta_j(l) = \frac{\sum_{k=1}^{126} \mathbf{1}_{\mathrm{k}}(j) C_j(k,l)}{\sum_{k=1}^{126} \mathbf{1}_{\mathrm{k}}(j)}. \tag{2.3}$$

Additionally, we examined whether the percent accuracy A_k depended on the particular mice included or excluded in a combination k. We focused this analysis on E16.5 mice; none of the E15.5 mice were ever excluded in a combination, and we also found that the accuracies did not vary as significantly when examining each group of combinations excluding a particular E17.5 mouse.

Each particular group of five E16.5 mice was used in $\binom{5}{5}\binom{6}{5} = 6$ combinations out of the total 126, implying that any pair of two specific mice was simultaneously left out of 6 combinations. The average accuracy $\bar{A}(j_n, j_m)$ over all combinations with E16.5 mice j_1 and j_2 left out, $j_1, j_2 = 6, ..., 12$, is thus

$$\bar{A}(j_1, j_2) = \frac{1}{6} \sum_{k=1}^{126} A_k (1 - \mathbf{1}_{\mathrm{k}}(j_n))(1 - \mathbf{1}_{\mathrm{k}}(j_m)) \tag{2.4}$$

with A_k defined as in Eqn. 2.2.

2.4 Results

2.4.1 Accuracy of clustering by time point

Fig. 2.4 shows the percent of clusterings $\beta_j(l)$ in which each mouse was placed in cluster l=1, 2, or 3 as calculated in Eqn. 2.3, which we can classify as 'correct' or 'incorrect': E15.5 mice in Fig. 2.4a are correctly clustered if placed in cluster 1, E16.5 mice in Fig. 2.4b are correctly clustered if placed in cluster 2, and E17.5 mice in Fig. 2.4c are correctly clustered if placed in cluster 3. Each time point's mice were sorted into the correct clusters more often than each of the other clusters. However, the cluster placements of some mice clearly indicate similarity to the mice outside their given time point, which may identify which particular mice were at earlier or later development stages than their presumed age, as previously described in Sec. 2.1.1.

Out of all the time points, the E15.5 mice were most accurately clustered, falling into cluster 1 in 69.69% of all clustering cases. If we examine each mouse separately, we see that Mice 1 and 5 were sorted into cluster 1 in 100% of cases. Mice 3 and 4 were clustered similarly to each other, with placement in cluster 1 for 65.08% and 63.49% of clusterings, respectively, and placement in cluster 2 in the remaining percentage of cases. This could point to the ages of these mice at the time of MADM induction as being slightly older than Mice 1 and 5, somewhere in between the E15.5 and E16.5 ages. Mouse 2 was least clearly clustered, ending up in cluster 1 for 19.84% of cases, cluster 2

for 46.83% of cases, and cluster 3 for 33.33% of cases. This indicates a lack of definition in Mouse 2's grid pattern, which we will examine further in the next section in Fig. 2.6.

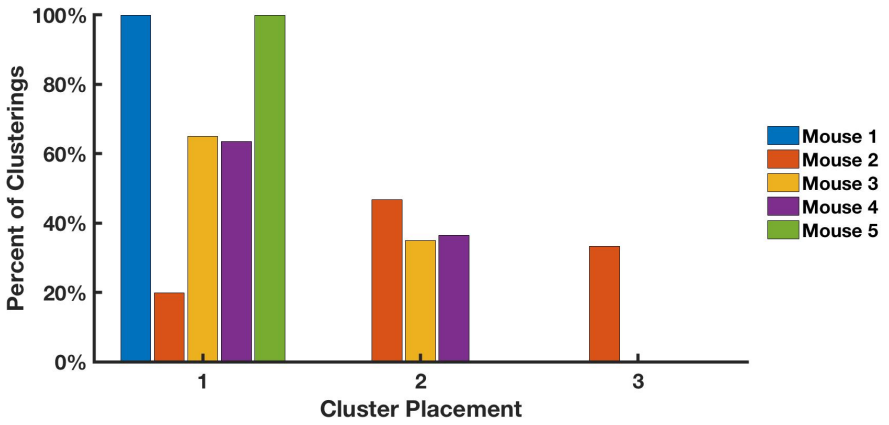
Since E16.5 is the middle time point, there was room for the true ages of these mice at the time of MADM induction to deviate in either direction, being slightly older or younger. This is reflected in the clusterings shown in Fig. 2.4b. Mice 7, 10, and 11 were predominantly placed correctly cluster 2, in respectively 98.89%, 97.78%, and 95.56% of cases, and placed into cluster 3 in all remaining cases. Mouse 8 may have been younger at the time of induction as it was placed in cluster 1 in 100% of cases, and Mouse 6 may have been somewhere between the E15.5 and E16.5 ages at induction time as previously seen with Mice 3 and 4 in E15.5, being in cluster 1 67.78% of the time and cluster 2 for the remainder. Mice 9 and 12 may have been older at MADM induction than their presumed age of E16.5, as these were both placed into cluster 3 for 98.89% of clusterings.

The clusterings of the E17.5 mice are more unexpected. Mice 16 and 17 were correctly placed in cluster 3 for 100% of clusterings. Mice 13 and 14 were in cluster 3 for 64.76% and 67.62% of cases, respectively, and in cluster 2 for the remainder, indicating the age of these mice at induction may have been between E16.5 and E17.5. However, Mice 15 and 18 were placed in cluster 1 in nearly every case – 100% and 93.33% of cases, respectively – corresponding to the E15.5 cluster. It is surprising that two of the six mice would have a large enough deviation in age to place them in the cluster farthest away from their presumed age's cluster. We examine the grid patterns of these mice in more detail in Fig. 2.6 , but this may be an indication that clones at E15.5 and E17.5 follow relatively similar patterns.

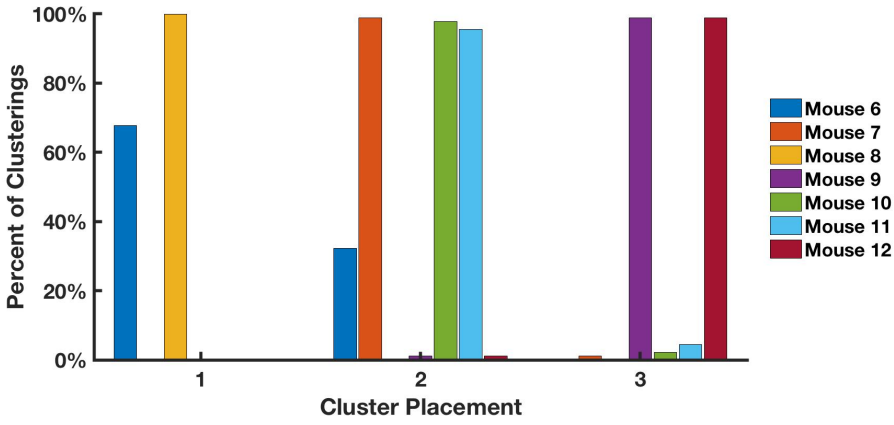
2.4.2 Qualitative comparison of grid patterns for most and least accurately classified

Figure 2.5 shows the grid patterns for the most accurately clustered mice for each time point, while Figure 2.6 shows the mice most often placed in the wrong clusters. From Fig. 2.5, we see that the most accurately classified mice have patterns clearly distinctive from the mice at other time points. The most accurately classified mice at E15.5, 1 and 5, are more dominated by clones containing neurons and smaller numbers of glia. Mice 7 and 12 in E16.5 instead show smaller neuron counts and a wider range of glial counts, and Mice 16 and 17 in E17.5 are dominated by smaller glial only clones. We can think of these patterns as 'representatives' for what is happening in the neurogenesis to gliogenesis switch at their respective time points. This gives a temporal progression of the NGS that is consistent with the presumed ordering of events: neurogenesis finishes first, and neurogenic clones may produce small numbers of glia, then gliogenesis surges as neurogenesis ends.

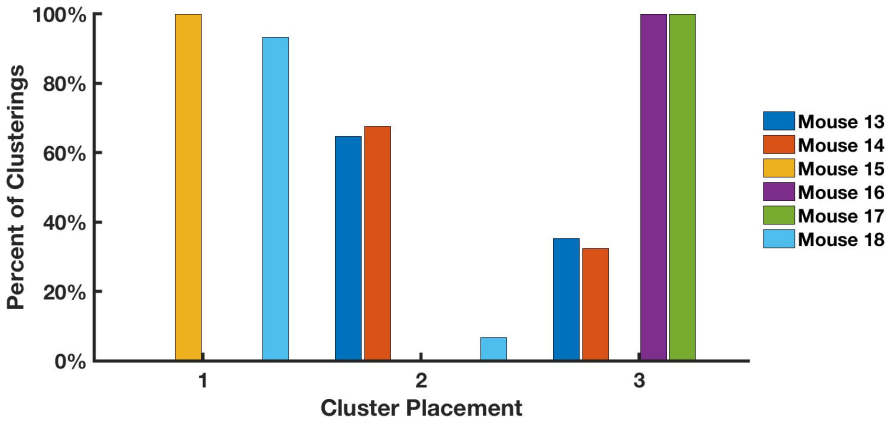
For the least accurately classified mice in Fig. 2.6, we instead see patterns that more closely resemble the representative patterns from other time points in Fig. 2.5. Mouse 2 in E15.5 has clones with small neuron counts and a large number clones of small glial counts, comparable to the representative pattern of E16.5 or E17.5. Mouse 8 in E16.5 was classified as being in the E15.5 cluster,



(a)



(b)



(c)

Figure 2.4 Percent of clusterings in which each mouse was placed in cluster 1, 2, or 3. E15.5 Mice $j = 1-5$ are shown in (a), E16.5 Mice $j = 6-12$ in (b), and E17.5 Mice $j = 13-18$ in (c). Placement of a mouse into clusters 1, 2, or 3 corresponds to a 'classification' of E15.5, E16.5, or E17.5, respectively.

which is consistent with its pattern showing more neurons and fewer glia. In the other direction, Mice 9 and 12 in E16.5 have patterns with no N clones and very high percentages of small G clones, closely resembling the representative pattern of E17.5. For E17.5, Mice 15 and 18 show either N clones or Mix clones containing large numbers of neurons, explaining why these two mice were placed in the E15.5 cluster.

2.4.3 Identifying influential or outlier mice

We can also evaluate the influence that different mice have on the SOM clustering by examining average accuracy of the six clusterings generated by excluding each pair of two mice from E16.5, $\bar{A}(j_n, j_m)$, defined in Eqn. 2.4. These 21 combinations of 2 out of 7 excluded E16.5 mice are shown in Fig. 2.7, and below are the corresponding values of $\bar{A}(j_n, j_m)$, The combinations are presented in order of decreasing average accuracy along with standard deviation bars. Interestingly, six out of the eight most accurate clusterings occurred in combinations that excluded Mouse 12. This particular mouse was one of the least accurately clustered mice as shown in Fig. 2.6, but this also indicates that using Mouse 12 in the initialization of cluster 2's centroid leads to a worse performance of the SOM overall. On the other hand, all five of the least accurate clusterings occurred in combinations that excluded Mouse 7. Mouse 7 was the most accurately classified mouse in E16.5 as previously described, but the fact that its exclusion in the combination of mice leads to the worst accuracy of the SOM indicates that this mouse is particularly influential in distinguishing between the patterns at different time points. Since the presence or absence of these two mice influence the SOM's accuracy, we can assume that Mouse 12 in E16.5 is a poor representation of the true NGS behavior at E16.5, and alternatively, Mouse 7 especially helps delineate NGS behavior at E16.5 from that at E15.5 and E17.5.

Naturally, some variability between mice is expected since the clones measured in each mouse are a random sample of only a small number of the total clones in the cortex. However, we have observed here that the placement of each mouse in a cluster helps us think more accurately about the data as coming from continuous time measurements rather than three completely distinct time points. Additionally, we have seen that particular mice may skew the data intended to represent a certain developmental time. Thus, examining MADM data using clustering is a valuable technique for determining possible outliers in the data, and it can also give insight into the true age of the embryos at the time of MADM induction.

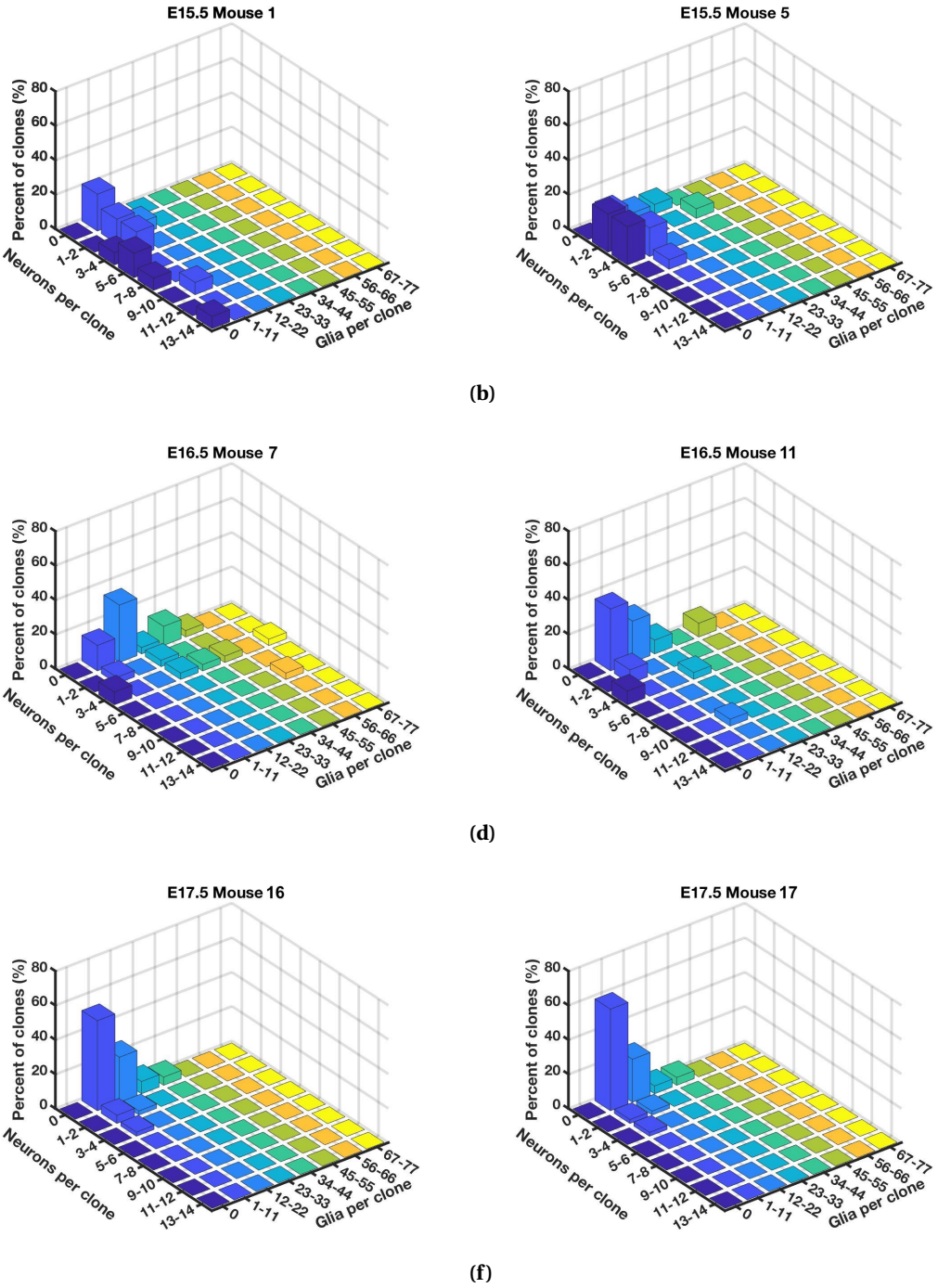


Figure 2.5 Grid patterns corresponding to the most accurately clustered mice from each time point. E15.5 Mice 1 and 5 are shown in (a)-(b), E16.5 Mice 7 and 11 in (c)-(d), and E17.5 Mice 16 and 17 in (e)-(f). The grid patterns are clearly distinct between the different time points and characterize the clonal shift in the NGS over E15.5 to E17.5: neural, to mixed, to glial.

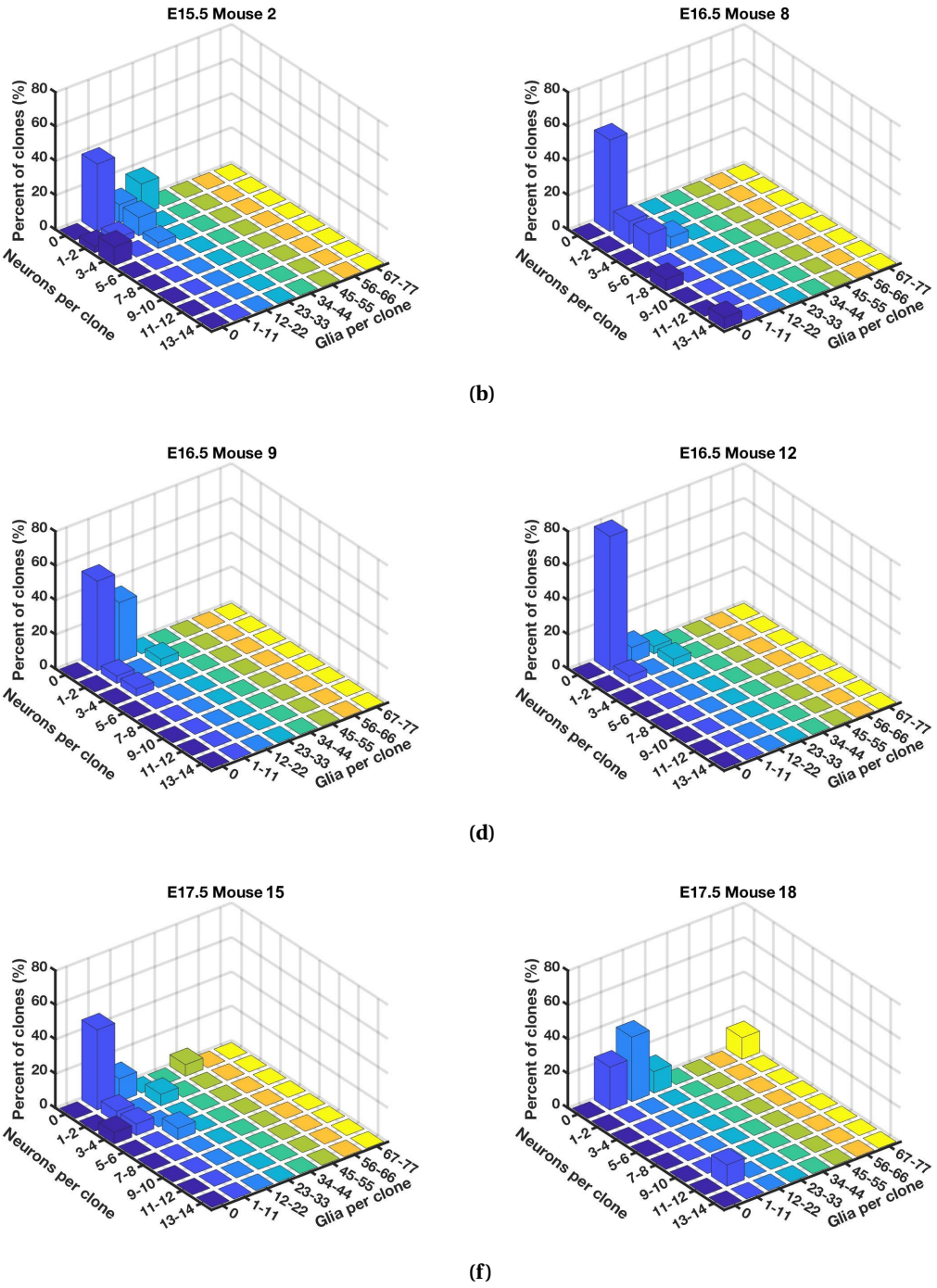


Figure 2.6 Grid patterns for the least accurately clustered mice from each time point. The progression in these grids over time does not accurately describe the progression of neural to mixed to glial clones during the NGS. In (a), E15.5 Mouse 2 has few neurons and mostly small glia. In (b)-(d) E16.5 Mice 8, 9, and 12 have mostly G clones rather than Mix. In (e)-(f) E17.5 Mice 15 and 18 produce more neurons than expected for mice this late in the NGS.

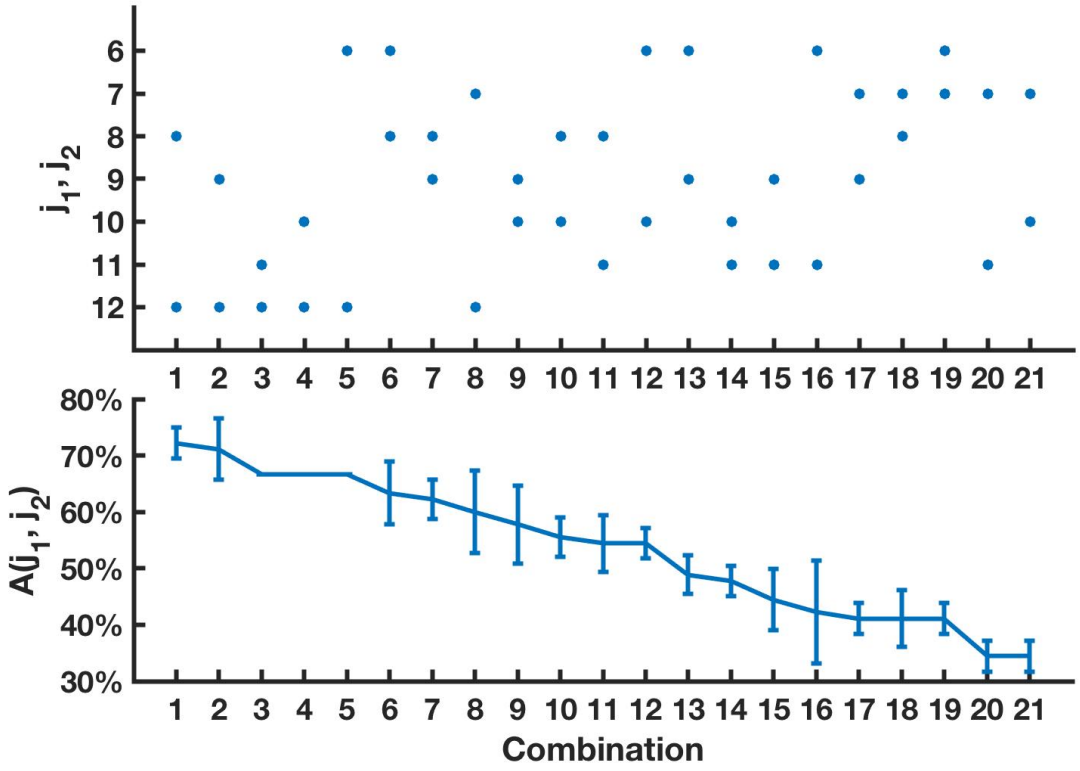


Figure 2.7 Average clustering accuracy over the 21 mouse combinations excluding two specific E16.5 mice. The above diagram indicates which two mice in each combination were excluded, thus combination 1 did not contain mice 8 or 12. Below, we show the mean ± standard deviation of the accuracy of the clusterings performed for combinations of mice that did not include the two mice specified.

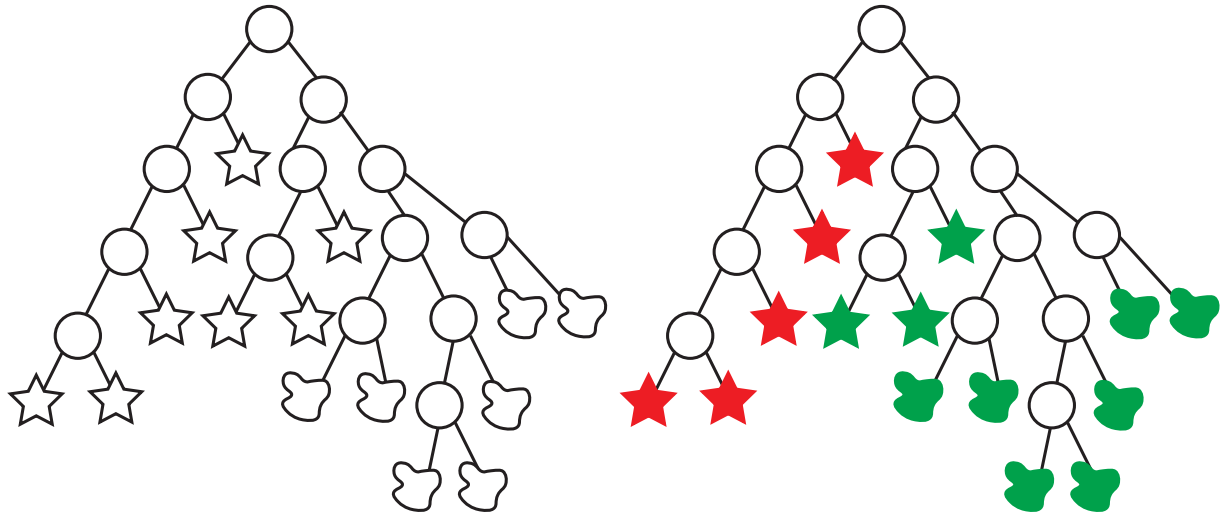
CHAPTER

3

BRANCHING AND TREE MODELS OF CELL DIVISION

In Chapter 1, we described how clonal lineages arise from the divisions of a single progenitor cell (RGP). However, the MADM technique specifically provides only a single temporal snapshot of an RGP's offspring. As an illustration, Fig. 3.1(a) shows an initial RGP undergoing a series of proliferative and differentiating divisions to produce neurons and glia. Our MADM data does not provide any knowledge of the series of divisions, but instead tells us the total number of neurons and glia produced by the RGP's red and green sublineages. The only information we would have for this example lineage is that the RGP produced five red neurons, three green neurons, and seven green glia, which we highlight in Fig. 3.1(b).

Since we want to understand cell division patterns during the NGS, it is natural to ask what can be inferred about the history of cell divisions given only the final number of differentiated cells. In this chapter, we explore this question using two mathematical representations of clonal lineages. First, we use binary trees [13] to estimate the generations of cell division that produce the clone sizes observed during the NGS. Second, we use branching processes [4] to probabilistically model the switch from neurogenesis to gliogenesis and test whether stochastic rules of cell division can reasonably represent the observed distribution of glia per clone in the NGS. Our work here will be used in our final model of glia production in the NGS in Chapter 6.



(a) **(b)**

Figure 3.1 (a) A hypothetical clonal lineage starting with a single RGP (top circle) which divides and differentiates into neurons and glia (b) The same lineage with all terminal cells (neurons and glia) highlighted red and green as would be observed with MADM. The history of divisions producing these neurons and glia as descendants of the single initial RGP would be unknown from the MADM technique.

3.1 Estimating Generations of RGP Division

Given a clone of size l, considering the l differentiated cells all being of the same type, we can establish a simple bound on the number of generations of RGP division h required to produce the clone. We illustrate this in Fig. 3.2 for a clone of size $l = 8$. On the left, we show the maximum generations of division that can produce $l = 8$ terminal cells. Here, the initial RGP has undergone a series of $h-1$ asymmetric differentiating divisions, each of which produce one terminal cell. The final division produces two cells and terminates the division process. The number of terminal cells produced is thus $(h-1)+2 = h+1$. Equating this expression with l and solving for h gives $h = l-1$. On the right, we show the minimum generations required to produce $l = 8$ terminal cells. In contrast with the strict asymmetric divisions in the maximum case, we have strictly symmetric divisions, which double the number of cells present in each generation. The $l = 8$ terminal cells can be produced in $h = 3$ generations, since $\log_2(8) = 3$. Note that this is the maximum number of differentiated cells that can be produced in $h = 3$ generations; if we considered $l = 9$, at least one more round of division would be needed. Generalizing these cases, the rounds of division h required to produce a clone of size l must satisfy

$$\lceil \log_2(l) \rceil \leq h \leq l-1, \tag{3.1}$$

assuming no cell death. For large clone sizes present in the NGS data, the range of possible values for rounds of division h becomes too large to be informative. It would be more useful if we could instead determine how likely a value of h is given a clone size l. To do this, we propose a combinatorial formula generated from representing clonal lineages as binary trees.

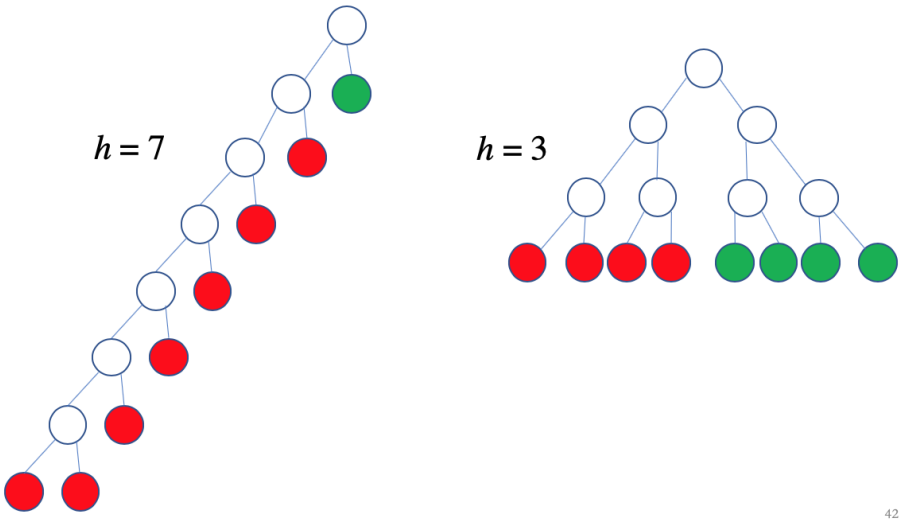


Figure 3.2 Examples of clonal lineages with $l = 8$ differentiated cells (marked red and green) having maximal and minimal heights as determined by Eqn. 3.1. In the left lineage, the maximum number of generations ($h = 7$) occurs, and minimum height ($h = 3$) occurs in the right lineage.

3.1.1 Tree representation

The clonal lineages shown in Fig. 3.1 are represented as a set of nodes and edges, with each node being a cell (RGP, neuron, or glial cell) and each edge connecting cells to their immediate descendants. The topmost RGP is the initial ancestor cell, which divides into two cells in the next generation, which continue to divide in the subsequent generation. All RGPs in the lineage have two direct descendants in the next generation, and all neurons and glia have zero descendants.

This representation therefore makes a clonal lineage into a type of graph known as a rooted full binary tree, defined as a set of nodes and edges that includes a topmost node (the root), and where every node has exactly zero or two direct descendants connected by edges [46]. In the terminology of trees, a *branch node* has two direct descendants, and a *leaf node* has zero direct descendants. The *height* of a tree is the number of edges on the longest path from the root node to a descendant leaf.

In Fig. 3.3, we show that when a clonal lineage is represented as a rooted full binary tree, the initial RGP is the root, additional RGPs in the lineage are branch nodes, the total rounds of division

gives the height of the tree, and the clone size is the number of leaves. Thus, if we can estimate the height of a rooted full binary tree given a certain number of leaves in the tree, we equivalently have an estimate for the division rounds given a certain clone size. For simplicity, we consider clones of only one differentiated cell type; the ordering of neurogenesis and gliogenesis in time implies that clones containing both neurons and glia have additional restrictions on where each cell type could occur in the lineage.

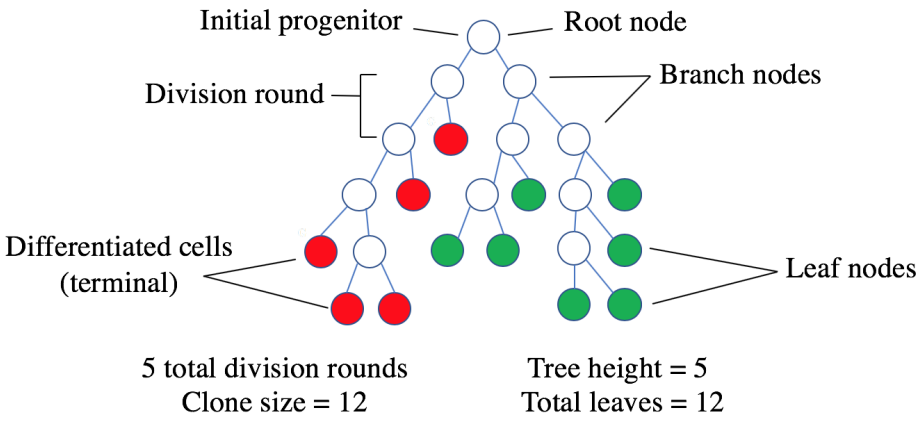


Figure 3.3 RGP lineage with red and green terminal cells and corresponding rooted full binary tree terminology.

For small l, it is easy to draw out all possible configurations of branch and leaf nodes and manually count the tree heights. We introduce the value n_l as the total number of rooted full binary trees with l leaves, which are listed for the trees up to $l = 4$ in Fig. 3.4. All n_l possible tree configurations for a given l are assumed to be equally likely. Next, we introduce the value $G_{h,l}$ as the total rooted full binary trees of height h and leaves l, which is nonzero only for values of h that satisfy Eqn. 3.1. The values of $G_{h,l}$ are additionally listed for each l in Fig. 3.4. Note that

$$n_l = \sum_{h=0}^{l-1} G_{h,l}. \tag{3.2}$$

The values of n_l are known to match the sequence of Catalan numbers, $C_n = (2n)!/((n+1)!n!)$ [18]. Manual counting of the possible tree configurations thus quickly becomes intractable: for $l = 5, 6, 7$ we have $n_5 = 14$, $n_6 = 42$, and $n_7 = 132$. Instead, we draw on an alternative definition of a full binary tree to create a recursion relation for the heights of larger trees.

Definition: A *full binary tree B* is either a single root node, or a root node with two subtrees L and R that are both full binary trees by our previous definition [13].

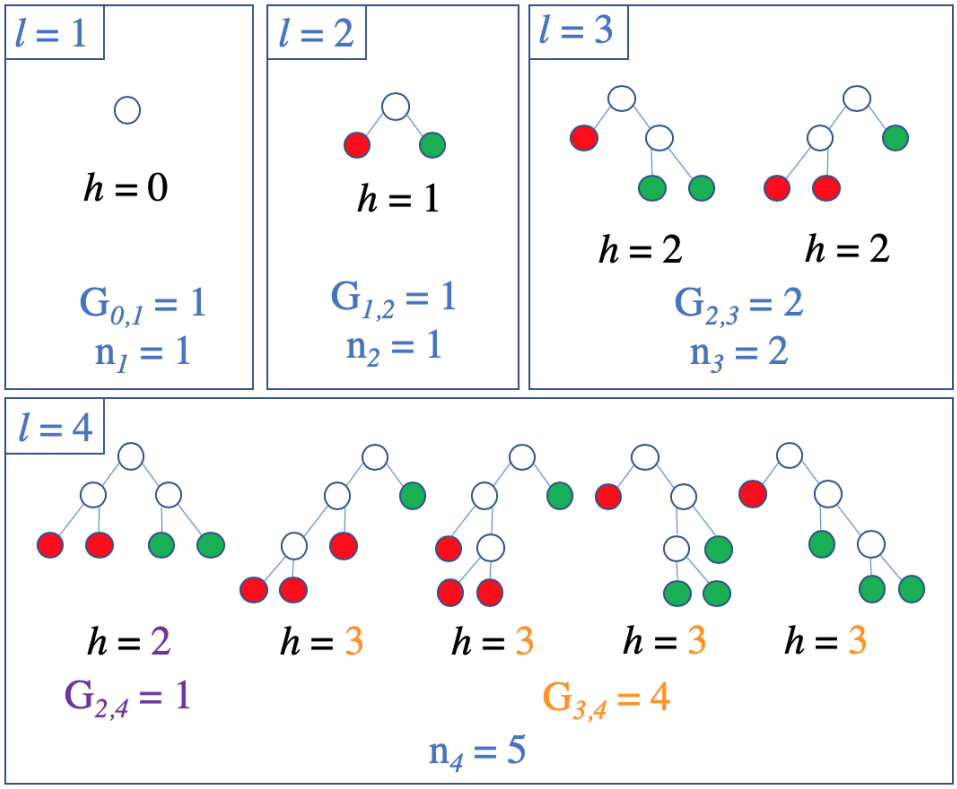


Figure 3.4 Possible lineages for trees with up to $l = 4$ terminal cells. In each case, we denote the number of trees of size l (n_l), the height of each tree (h), and the number of trees having height h and l terminal cells ($G_{h,l}$).

Thus, if we consider a clonal lineage as a full binary tree B with l leaves, we can equivalently consider it to be a root node composed with two full binary trees L and R containing k and $l-k$ leaves, respectively, where $1 \leq k < l$. As an example, we illustrate the possible subtree sizes for a clone with $l = 5$ leaves in Fig. 3.5: either (1,4), (2,3), (3,2), or (4,1). In the (1,4) case, there is one possible configuration of the left subtree L with $k = 1$ leaf, and there are five possible configurations of the right subtree R with $l - k = 4$ leaves, as listed in Fig. 3.4. The total configurations of L and R thus multiply to give five possible configurations of the (1,4) tree. To determine the heights of these five trees, we note that the possible heights of L and R are already known and listed in Fig. 3.4. Since the subtrees are separated from the root by one generation, the height of the full tree B will be

$$\text{height}(B) = 1 + \max(\text{height}(L), \text{height}(R)) \tag{3.3}$$

For the five possible (1,4) tree configurations, the R subtree with 4 leaves has the maximum height in each case. Similarly, for the (2,3) tree configuration, the R subtree with 3 leaves has the maximum height in each of the two possible configurations (Fig. 3.5). We see that the total trees with $l = 5$

leaves is $n_5 = 14$, which can be broken down into $G_{3,5} = 6$ trees with height $h = 3$ and $G_{4,5} = 8$ trees with height $h = 4$. These are the only two possible heights for a tree of $l = 5$ leaves, consistent with Eqn. 3.1.

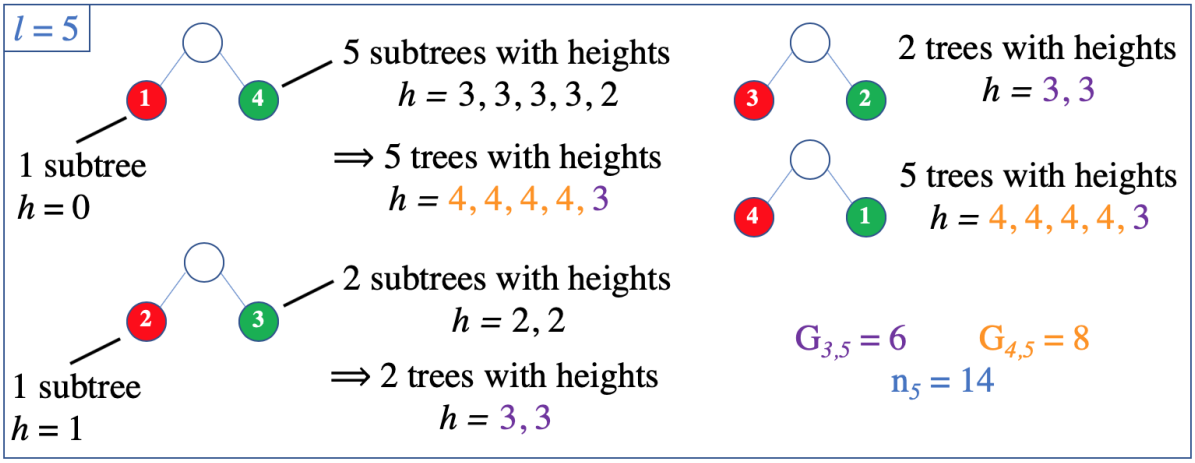


Figure 3.5 Representation of trees with $l = 5$ leaves as two subtrees having k and $l - k$ leaves. For each case of subtree sizes, the number of trees of its size and the heights of those trees are already known, as illustrated in Fig. 3.4. The height of the tree with $l = 5$ leaves can thus be calculated from the subtree heights according to Eqn. 3.3. For $l = 5$, $n_5 = 14$ possible lineages exist, of which 6 have height 3 and 8 have height 4 ($G_{3,5}$ and $G_{4,5}$).

3.1.2 Generation estimate

We can now write a generalized recursive relation for the number of trees $G_{h+1,l}$ having height $h+1$ and leaves l. Eqn. 3.3 implies that if a tree B has height $h+1$, at least one of the subtrees L or R has height h. If that subtree has k leaves, there are $G_{h,k}$ possible configurations of that subtree. The other subtree must therefore have $l-k$ leaves, and its height must be $j \leq h$ to ensure that the height of B is not greater than $h+1$. This subtree thus has $G_{j,l-k}$ possible configurations. Considering each pair of subtree sizes k and $l-k$, $G_{h+1,l}$ is thus represented

$$G_{h+1,l} = \sum_{k=1}^{l-1} G_{h,k} \left[2 \sum_{j=0}^{h} G_{j,l-k} - G_{h,l-k} \right], \tag{3.4}$$

given values of $G_{h,k}$ known for small h and $k < l$. In Eqn. 3.4, the subtrees are ordered so that L has height h, and R has height $j \leq h$. These combinations are doubled to account for the reflected tree, where L and R have heights $j \leq h$ and h, respectively. However, the case of both subtrees having

heights h does not need to be doubled, so we must subtract one of the $G_{h,l-k}$ terms.

We use Eqn. 3.4 to generate the recursive values of $G_{h,l}$ in MATLAB for trees up to size $l = 50$, using initial values shown in Fig. 3.4. As a verification, we confirm that the total trees generated with size l, found using Eqn. 3.2, matches the known sequence of Catalan numbers as expected [18]. For each l, we normalize the values of $G_{h,l}$ to find the conditional probability $p(h|l)$ that a tree with l leaves has height h,

$$p(h|l) = \frac{G_{h,l}}{n_l}. \tag{3.5}$$

The distributions of $p(h|l)$ are shown for a subset of tree sizes l in Fig. 3.6a. To estimate $p(h)$, the probability that h rounds of division occurred in the dataset, we use the law of total probability, which states that the probability of an event A can be calculated as $p(A) = \sum_n p(A|B_n)p(B_n)$ for some finite or countably infinite partition of events B_n. Using Eqn. 3.5, this implies that

$$p(h) = \sum_{l=1}^{50} \frac{G_{h,l}}{n_l} p_l, \tag{3.6}$$

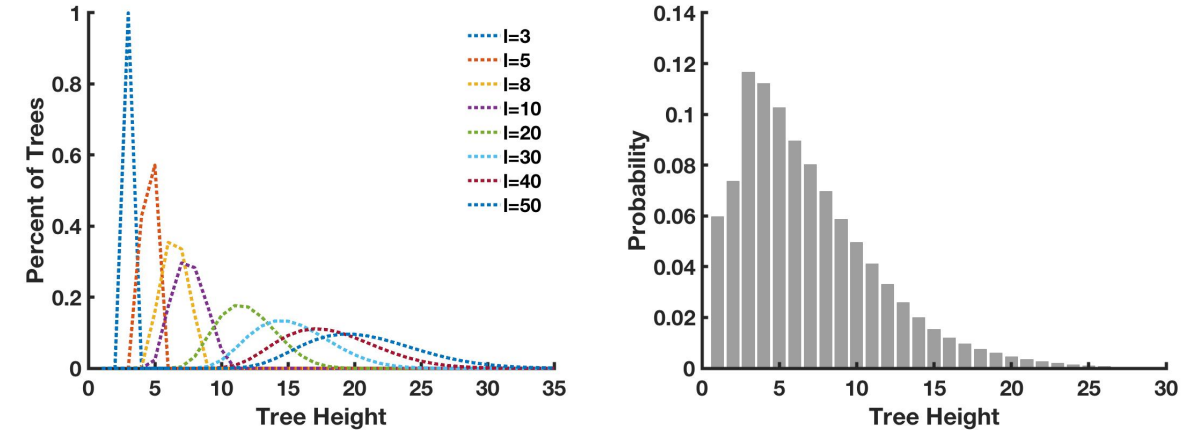
where p_l is the observed frequency of a clone containing l glia in the dataset. Eqn. 3.6 is evaluated and displayed in Fig. 3.6b. The resulting probability distribution of $p(h)$ indicates that ≈95% of lineages are completed in under 15 generations, and that, for $l \leq 50$, very few clones plausibly divide beyond 20 generations. We will therefore set a maximum of 20 generations of division in our future work simulating glial production in the NGS in Chapter 6.

3.2 Branching processes

In the previous section, we illustrated how representing a clonal lineage of size l as a rooted full binary tree with l leaves allows for an estimate of the generations of division h. To do this, we assumed that for a binary tree with l leaves, all possible tree structures were equally likely. This may not be the case if certain types of divisions occur more frequently. For instance, between the two clones of size $l = 8$ shown in Fig. 3.2, the right lineage may be more probable than the left one if progenitors strongly favor symmetric over asymmetric divisions. In this section, we draw on the theory of branching processes to consider how the hypothetical distribution of clone sizes changes if we attach likelihoods to the different types of cell divisions.

3.2.1 Description

Branching processes are a type of stochastic process used to model populations that proliferate, mutate, and/or die over time according to a prescribed set of probabilities. They are useful modeling



(a) (b)

Figure 3.6 (a) Probability distribution $p(h|l)$ of tree heights h for selected values of l, calculated from Eqn. 3.5. The area under the distribution for each l is equal to 1. (b) Probability of a tree of height h calculated from Eqn. 3.6 using the conditional probability distributions shown in (a) and the observed frequency p_l of clones with l glia in the data.

tools in many biological applications including the spread of genetic mutations in a population, cancer proliferation and response to treatment, and, relevant to our work, growing stem and progenitor cell populations [4]. Although branching processes are generally described as acting on 'particles,' we will proceed with describing them specifically as acting on cells.

A basic branching process representing progenitor proliferation and differentiation can be described as follows. The process begins with an initial progenitor cell (the ancestor) at time $t = 0$. We will refer to the ancestor cell as a type 1 cell. After one discrete unit of time, at $t = 1$, the ancestor divides into two daughter cells which may be the same or different types as the ancestor. The cell types of these two offspring are selected randomly according to a set of probabilities $p_{i,j}$, denoting the probability of a type $i = 1,2,...,k$ cell producing offspring of type $j = 1,2,...,k$. The two daughter cells then exist for one discrete unit of time, after which each cell produces two offspring randomly according to $p_{i,j}$. It is important to point out that in the process of progenitor proliferation and differentiation, differentiated cell types are terminal and produce zero offspring with probability 1. Thus, the process continues until no proliferative cells remain. If the probability of differentiation is high enough, the processes is guaranteed to terminate or 'go extinct' in a finite number of time steps, whereas a low probability of differentiation results in the chance that a process continues indefinitely [48]. Fig. 3.7 shows an example of a lineage produced by a process of three cell types: progenitors, neurons, and glia.

The number of cells of each type that are present at time t is a sequence of random vectors $\mathbf{X}_t$, called the state of the process. These values are shown at each time step for each cell type in Fig.

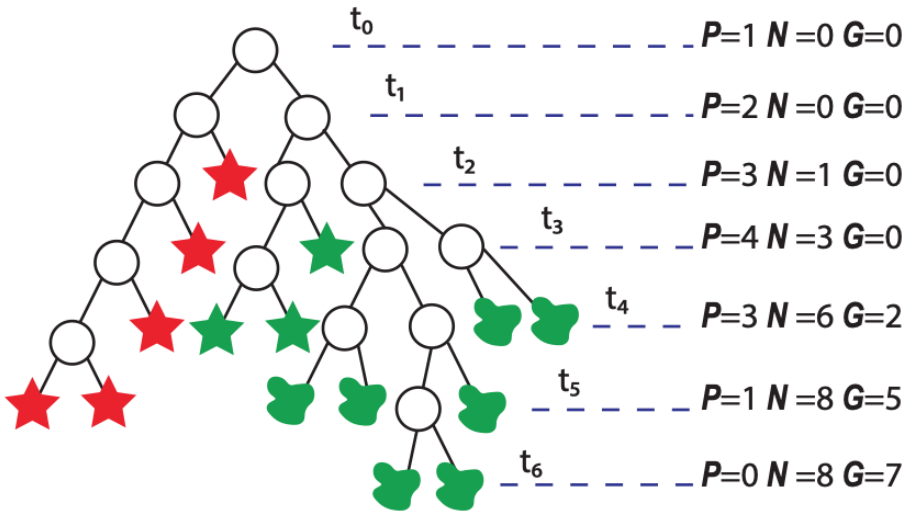


Figure 3.7 Example branching process lineage. The process starts with one progenitor cell P at time t_0, zero neurons N, and zero glia G. The starting state of the process $\mathbf{X}_0$ is thus a vector [1,0,0]. The initial progenitor divides into two progenitors at t_1, hence $\mathbf{X}_1$=[2,0,0]. At successive discrete time points $t_2, t_3, ...$, progenitors divide and differentiate into neurons and glia. The components of the state vector are shown for each time. Note that the number of progenitors present at the end of the process is zero, as all have undergone differentiating divisions.

3.7. All cells behave independently of their ancestors and siblings born in the same generation. A branching process therefore has the Markov property

$$\mathbb{P}[\mathbf{X}_{t+1} = n|\mathbf{X}_t = m_t, \mathbf{X}_{t-1} = m_{t-1}, ..., \mathbf{X}_0 = m_0] = \mathbb{P}[\mathbf{X}_{t+1} = n|\mathbf{X}_t = m_t]. \tag{3.7}$$

Simply stated, the probability of the process having state $\mathbf{X}_{t+1}$ at time $t+1$ depends only on the state of the process at time t and not on any previous states, thus making it 'memoryless' or 'self-similar.' The shift of cells between types in any branching process can be summarized in a transition matrix T where $T_{i,j}$=$p_{i,j}$ as defined previously. T can be used to examine a process's asymptotic behavior and chance of extinction depending on different division probabilities [4].

T is called *reducible* if it can be put into block upper-triangular form by simultaneous row and column permutations. If this cannot be done, T is called *irreducible*. Conceptually, a process being irreducible implies that every cell type in the process can be produced from any other cell type in a finite number of divisions. In the NGS, we consider four cell types: neuron-producing RGPs np, glia-producing RGPs gp, neurons N, and glia G. The potential behaviors of these cell types are as follows:

1. np can self-replicate, differentiate into neurons N, or switch to gliogenesis by producing gp as offspring

2. gp can self-replicate or differentiate into glia G

3. N and G are both terminal and produce no further offspring

With $*$ representing a nonzero probability, our transition matrix would thus have the following upper triangular form

$$\begin{array}{c} \\ np \\ gp \\ N \\ G \end{array} \begin{array}{c} \begin{array}{cccc} np & gp & N & G \end{array} \\ \begin{bmatrix} * & * & * & 0 \\ 0 & * & 0 & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The process is therefore reducible. We can also understand this conceptually by noting that some states in the process cannot be reached from others, for instance, np cells cannot be produced from N cells since N cells do not proliferate.

Two issues arise when attempting to analyze RGP proliferation and differentiation in the NGS using branching processes. First, asymptotic analysis of our problem is not informative since we know the population of RGPs responsible for building the cortex will always 'go extinct' by differentiating into neurons and glia. Second, the theory that exists for branching processes is overwhelmingly applicable to the irreducible case, whereas reducible processes require methods uniquely determined for the particular application [53]. In the next section, we will briefly describe a method used to determine the clone size distribution during neurogenesis, informed by a basic branching process known as a Galton-Watson (GW) process [48]. We then adapt this method to model glial production during the NGS.

3.2.2 Determining Clone Size Distributions Using the Galton-Watson Process : Neurogenesis

In a study by Slater et al. [48], a GW process was formulated to model neurogenesis. The process consisted of two cell types S and N, where type S cells are progenitors that can proliferate or differentiate into neurons, and type N cells are neurons that do not divide. The probabilities of an S cell producing pairs of offspring were defined with parameters p and q, where $P(N,N)=p$, $P(S,S)=q$, and $P(S,N)=P(N,S)=(1-p-q)/2$.

The goal in this study was to determine the probability of a clone of size n (having n neurons), denoted Q_n, given a starting S cell that divides according to some parameters p and q. It was noted that a clone of size 2 could only occur from the initial S cell undergoing an immediate $\{N,N\}$ division. Since the probability of this division occurring is p, the probability Q_2 of a clone of size 2 is also p. A clone of size 3 would require an initial asymmetric $\{S,N\}$ or $\{N,S\}$ division, followed by the S cell produced from this division undergoing an $\{N,N\}$ division to produce two more differentiated

neurons (Fig. 3.8). Thus, considering each cell division as an independent probabilistic event, the probability of the cell divisions are multiplied together to give $Q_3 = p(1-p-q)/2 + p(1-p-q)/2 = p(1-p-q)$. This method of listing possible division patterns mirrors the enumerating of tree sizes we performed in the previous section (Fig. 3.4). As such, we can use recursion to develop a formula for larger clone sizes: the extrapolated pattern gives the recurrence relation

$$Q_n = q\sum_{k=2}^{n-2} Q_k Q_{n-k} + (1-p-q)Q_{n-1}, \quad n \geq 4. \tag{3.8}$$

Here, the probability of a clone of size n is calculated by considering all combinations of subtrees of size k and $n-k$. The first sum considers an initial $\{S,S\}$ division with probability q, where the two subtrees must both have ≥ 2 N cells. The second term considers an initial $\{S,N\}$ or $\{N,S\}$ division, which necessarily has one subtree with only one N cell produced after the first division, and the second subtree must therefore have $n-1$ N cells to sum to a clone of size n.

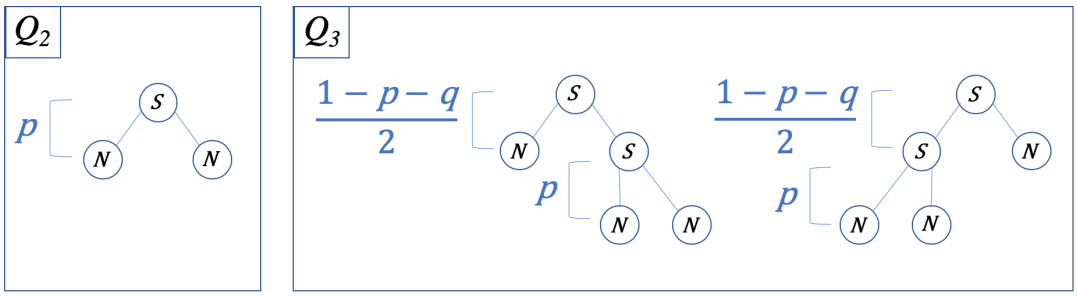


Figure 3.8 Clone size probabilities according to division probabilities. A clone with two neurons (left) arises from a $\{N,N\}$ division, which occurs with probability p. The probability of a clone of size two, Q_2, is thus equal to p. A clone with three neurons (right) would be produced by an asymmetric division, $\{S,N\}$ or $\{N,S\}$, followed by a differentiating $\{N,N\}$ division. Multiplying the probabilities of these two division events implies that the probability of each lineage is $p(1-p-q)/2$. The total probability of a clone with three neurons, Q_3, is the sum of the two clone probabilities, $Q_3 = p(1-p-q)$. Eqn. 3.8 establishes how Q_n is calculated for larger clone sizes n.

Parameters p and q were chosen from observing the frequency of $\{N,N\}$ and $\{S,S\}$ progenitor cell divisions *in vitro*. The probability distribution Q_n was then compared with the frequency distribution of clone sizes in a dataset of neurogenic RGPs and determined to reasonably represent the data by a Chi-square goodness of fit test [37].

3.2.3 Determining Clone Size Distributions Using the Galton-Watson Process: Gliogenesis

As previously described, we consider additional cell types to model RGP behavior in the NGS: neuron-producing RGPs np, glia-producing RGPs gp, neurons N, and glia G. To model the switch into gliogenesis, we will not consider neurons N and only allow np to produce np and gp. Our goal is to determine a recursive formula for glial clone sizes similar to Eqn. 3.8, taking into account that np progenitors must switch to gp progenitors before producing glia. Thus, we must enumerate not only the possible ways of producing glia from gp progenitors, but also of producing gp from np progenitors.

We illustrate in Fig. 3.9 the different divisions that can occur, either with an np or a gp as the ancestor cell. Probability parameters are assigned to each type of division. Using the convention from [48], we denote the probability of np divisions $P(gp,gp) = p_1$, $P(np,np) = q_1$, and $P(np,gp) = P(gp,np) = (1-p_1-q_1)/2$, as the probabilities must sum to 1. Similarly, the probability of gp divisions will be denoted $P(G,G) = p_2$, $P(gp,gp) = q_2$, and $P(G,gp) = P(gp,G) = (1-p_2-q_2)/2$.

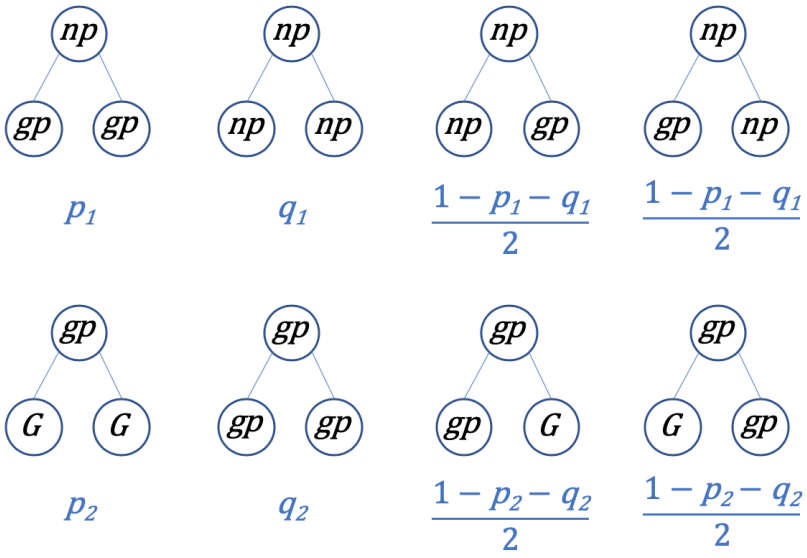


Figure 3.9 Possible clonal divisions during gliogenesis. In each case, the initial cell (neural progenitor np or glial progenitor gp is shown with the two offspring of its division. The probability of each division type occurring is defined with parameters p_1, q_1, p_2, and q_2, where the np and gp probabilities each sum to 1.

To begin, we consider a process we will call Q^1, which starts with a gp cell that divides to produce glia according to probabilities p_2 and q_2. This is actually identical to the process described in Sec. 3.2.2. We will denote the probability of a clone of size n coming from this process as Q^1_n, which is

found similar to Eqn. 3.8 as

$$Q_n^1 = q_2 \sum_{k=2}^{n-2} Q_k^1 Q_{n-k}^1 + (1-p_2-q_2) Q_{n-1}^1, \quad n \geq 4, \tag{3.9}$$

where $Q_2^1 = p_2$ and $Q_3^1 = p_2(1-p_2-q_2)$.

The "1" in the superscript of Q^1 denotes the $m = 1$ initial gp progenitor present in this process. Accordingly, we can define Q^2 as a process beginning with an np cell which immediately undergoes a division into $m = 2$ gp cells. This initial division occurs with probability p_1 as previously defined. Each of the first generation gp offspring then undergo divisions according to the probabilities p_2 and q_2. To determine clone size probabilities Q_n^2, we recall the recursive definition of a full binary tree from Sec. 3.1.1, which states that a full binary tree is composed of left and right subtrees which are each full binary trees. In the case of the Q^2 process, each of those subtrees is formed from a Q^1 process, since each begins with a gp cell (Fig. 3.10). Thus, taking all possible combinations of Q^1 trees that sum to give n cells, the probability Q_n^2 is defined recursively as

$$Q_n^2 = p_1 \sum_{k=2}^{n-2} Q_k^1 Q_{n-k}^1, \quad n \geq 4. \tag{3.10}$$

Note here that the minimum clone size is $n = 4$. Thus, we have a recursive process whose clone size probabilities can be determined from the already known Q_n^1 probabilities.

Having defined the Q^2 process, we can proceed to recursively define processes from other cases of an initial np cell dividing producing m gp cells, $m \geq 3$. In each case, Q_n^m is found by multiplying the probability of the first division, which must be either a $(gp, np)/(np, gp)$ or a (np, np) division since $m \geq 3$, by all combinations of clone sizes possible between its two subtrees. For instance, a Q^2 process has two subtrees which are both Q^1 processes, and a Q^3 process has two subtrees which are a Q^1 and Q^2 tree (Fig. 3.10, one symmetry shown), and so on. A recursive formula can thus be established, similar to Eqn. 3.8,

$$Q_n^m = (1-p_1-q_1) \sum_{k=2}^{n-2} Q_k^{m-1} Q_{n-k}^1 + q_1 \sum_{k=2}^{n-2} \sum_{l=2}^{m-2} Q_k^l Q_{n-k}^{m-l}. \tag{3.11}$$

Here, the first sum considers all combinations of left and right subtrees that can occur after an initial $(gp, np)/(np, gp)$ division, which occurs with probability $1-p_1-q_1$. The double sum considers the possible combinations of left and right subtrees after an initial (np, np) division, which occurs with probability q_1.

Q_n^m is calculated for $n \leq 50$, as $\approx 98\%$ of clones in the NGS dataset have 50 or fewer glia. The maximum value of m we calculated is $m = 16$; we wish to model the switch from np to gp occurring in relatively few divisions, and for $m > 16$, at least four generations of division are required for all

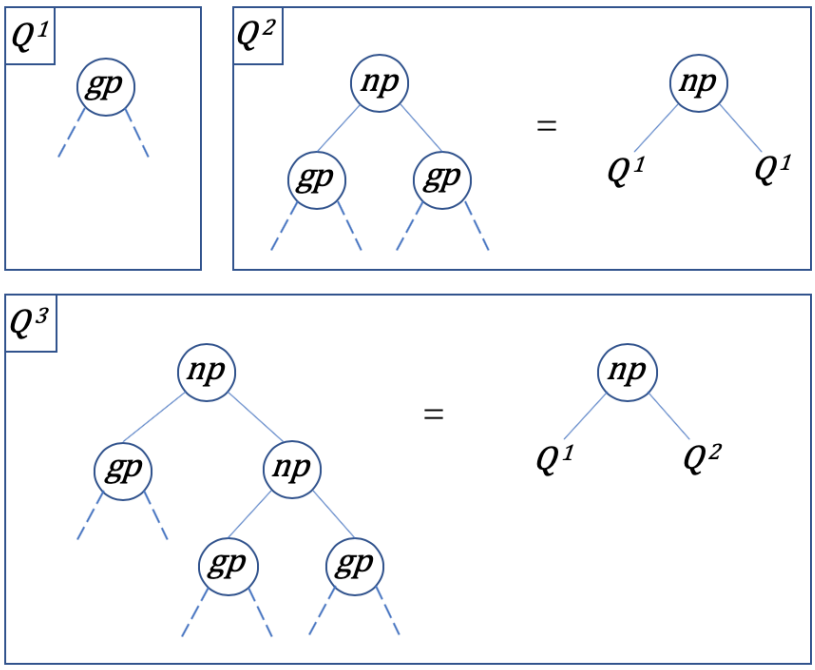


Figure 3.10 Illustration of clonal lineages during gliogenesis built from subtrees. In the upper left, we define the process Q^1, which starts with a gp cell that divides to produce glia according to probabilities p_2 and q_2. Next, we define a Q^2 process, which starts with an np cell and produces two gp cells. Since each root of these two sublineages is a gp cell, each produces glia according to the process Q^1. Similarly, we define Q^3 as the process starting with an np cell and dividing into np and gp cells as shown. The root of the left sublineage of the initial np cell matches the Q^1 process, and the root of the right sublineage matches the Q^2 process as previously defined.

np progenitors to switch to gp. The final probability Q_n of a clone of size n is then found by adding together all Q_n^m terms,

$$Q_n = \sum_{m=2}^{16} Q_n^m, \quad n \geq 4 \tag{3.12}$$

Note that Q_1 is not included in this sum, since we are only considering clones that start with an np cell and undergo a switch to gp at some point; Q_1 was simply used in formulating the recursion. Additionally, this implies that Eqn. 3.12 can only model clones having ≥ 4 differentiated glia.

3.2.4 Evaluation of clone size distribution Q_n

In contrast with the GW process for neurogenesis in [48], we do not have *in vitro* progenitor divisions that we could observe to select division probability parameters. However, we note that Eqn. 3.12 is a multivariate polynomial in p_1, q_1, p_2, and q_2, as a clone's probability of occurring is the product of the probabilities of each of its individual divisions. It is thus possible to find the set of parameters

that minimize the sum of square differences between Q_n and the normalized distribution of glia per clone in the NGS data, which we denote $H(n)$, $n = 4, ..., 50$. We use the native MATLAB function `fmincon` to find parameter values p_1, q_1, p_2, and q_2 that minimize the sum of squares error between $H(n)$ and Q_n. The function `fmincon` allows for constraints on the possible parameter values, thus we use this feature to constrain each parameter to be between 0 and 1. Additionally, we perform the same parameter estimation using the Slater Q_n model for neurogenesis clone sizes. This model, given in Eqn. 3.8, is generated from only two cell types. In this case we would consider the two cell types as glial progenitors gp and glial cells G, which differentiate and proliferate according to probabilities p_2 and q_2. Note that in this case, the minimum clone size is 2 rather than 4, so we compare to $H(n)$, $n = 2, ..., 50$.

Once we find parameter sets for each Q_n model, we use a Chi-square goodness of fit test to determine how well each represents the data. As compared with other statistical goodness of fit tests like the Anderson-Darling or Kolmogorov-Smirnov, the Chi-Square test operates on discrete binned data and is thus preferable for evaluating a statistical distribution's fit to the distribution of glia per clone [37]. The hypotheses for the Chi-Square test are

1. H_0: The data follows the considered distribution.
2. H_1: The data does not follow the considered distribution.

The Chi-square test gives a p-value, which is used to reject the null hypothesis if p is below the confidence level α. We set $\alpha = 0.05$, hence if we perform the Chi-square test for a distribution and find p<0.05, we would conclude that the particular distribution does not represent the data. If p>0.05, then the distribution does sufficiently represent the data.

In Table 3.1, we list the estimated parameter values when fitting each model to each dataset, along with the p-value of the Chi-square test. The fits of each model to the data are additionally shown in Fig. 3.11. We see an interesting result here. When fitting the Slater Q_n model to the distribution $H(i)$ formed from G and Mix data (Fig. 3.11b), the χ^2 p-value is below the $\alpha = 0.05$ significance level. This suggests that we would reject the null hypothesis of the Chi-square test; hence, this dataset does not follow the distribution prescribed by the Slater Q_n models. The χ^2 p-value for fitting the NGS Q_n model to the G and Mix dataset (Fig. 3.11a) is just above $\alpha = 0.05$, which suggests that the G and Mix distribution $H(i)$ may also not follow this model, though we would not reject the null hypothesis at the $\alpha = 0.05$ significance level. However, when fitting both the NGS and Slater Q_n models to the distribution $H(i)$ formed from glial only data (Fig. 3.11c-d, respectively), the χ^2 p-value in each case is well above the null hypothesis rejection level of $\alpha = 0.05$. Thus, it appears that either of these stochastic models could reasonably represent the process that produces clones with only glia and no neurons.

This is an interesting result, as it suggests that there may be some difference in the distribution of glial output between glial only and mixed clones. Currently, it is assumed that all glia come from RGPs

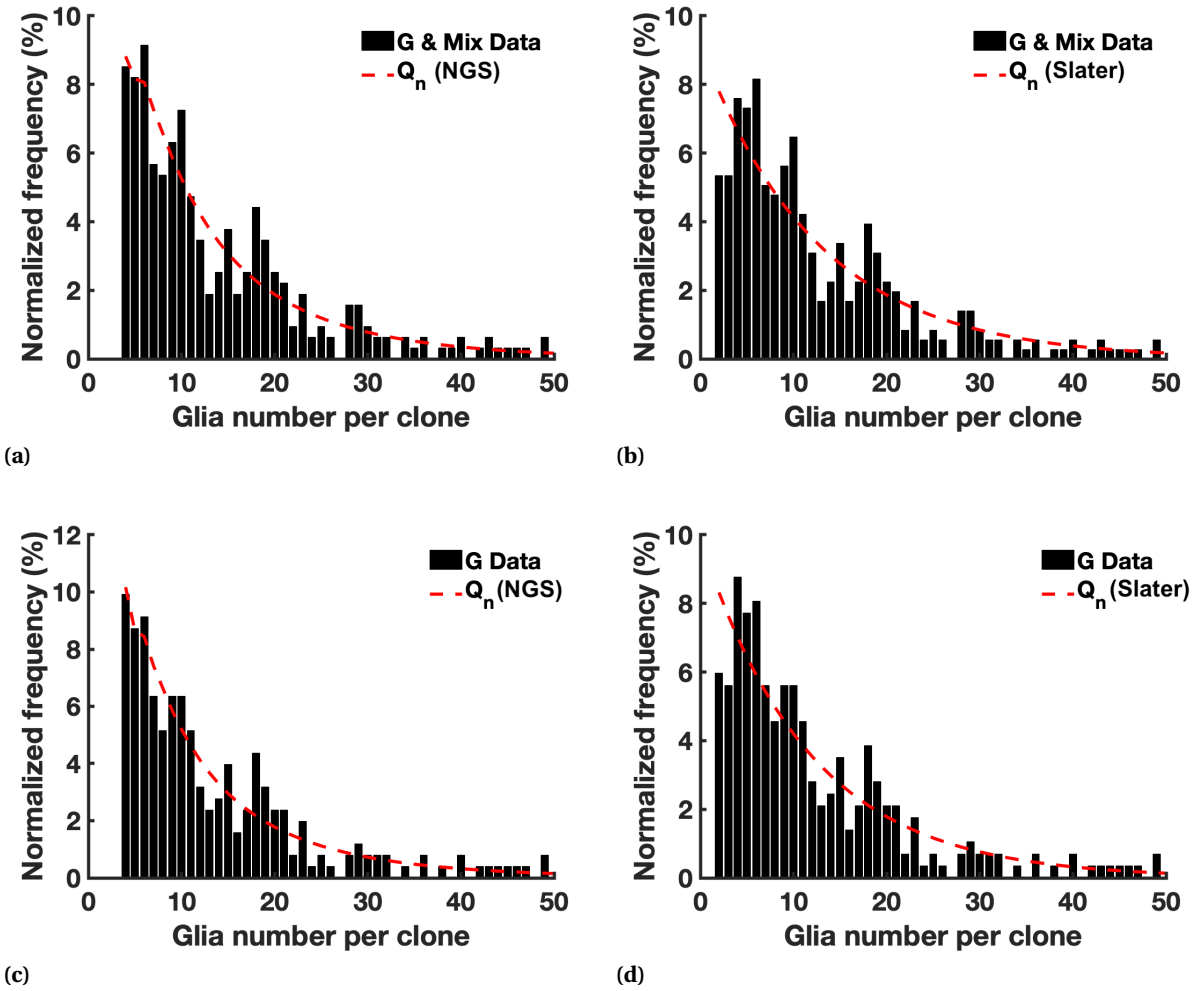


Figure 3.11 Fits of model Q_n to the distribution of total glia per clone in WT G+Mix clones (a)-(b) and WT G clones (c)-(d). Two versions of Q_n are fit to the distributions. In (a) and (c), version (3.12) representing the NGS used. Note that this version of Q_n can only fit clone sizes ≥ 4, and the clone size distribution is normalized accordingly. In (b) and (d) version (3.8) is used, which represented neurogenesis in Slater et al. [48]. Table 3.1 shows the p-values for the Chi-square test evaluating the goodness of fit of each model to each distribution, indicating that only the WT G clones are represented by the model Q_n.

Table 3.1 Parameters p_1, q_1, p_2, and q_2 identified for fitting the model for clone size distributions Q_n to the data of glia per clone in the NGS data. Two forms of Q_n were considered: model (3.8), formulated by Slater et al. for neurogenesis [48] and model (3.12), formulated above for a population of progenitors that undergoes a switch from neurogenesis into gliogenesis (denoted NGS Q_n). Each model was fit to the normalized frequency histogram of glia per clone $H(i)$, formulated either considering all G and Mix clones, or G clones only. The fit of each model to each histogram was evaluated using a Chi-square test [37], for which the p-values are shown. A low p-value (p<0.05) indicates that the particular model does not accurately represent the distribution of glia in the histogram.

$H(i)$ dataset	Model	p_2	q_2	p_1	q_1	χ^2 p-value
G and Mix	NGS Q_n	0.4003	0.1370	0.5377	0.0577	0.0667
G and Mix	Slater Q_n	0.0764	0	-	-	0.0254*
G only	NGS Q_n	0.4240	0.1514	0.5539	0.0667	0.2350
G only	Slater Q_n	0.0820	0	-	-	0.2770

that previously produced neurons. However, if this was the case, we would expect the distribution of glia per clone in mixed and glial only clones to be similar, as the glial only clones would have originated from the same RGPs as mixed clones but were simply labeled at a later time with MADM so that no neurons appear in the lineage. Instead, we see the distribution of glia in glial-only clones sufficiently represented with a stochastic GW model. Furthermore, it is represented sufficiently by the simpler Slater Q_n model given in Eqn. 3.8, while adding mixed clones to the distribution causes the GW model to fail in its representation of the data. It is possible that two separate populations of progenitors exist - one of which produces only glia and behaves stochastically according to a GW model, and the other of which undergoes a switch from neurogenesis into gliogenesis, whose behavior we do not yet know.

3.2.5 Discussion of further modeling

The possibility of different populations of progenitors leads us to consider how the 'mixed' RGPs that switch from neurogenesis into gliogenesis divide and differentiate, and how this differs from the 'glial only' RGPs that follow a stochastic GW process. In the next chapters, we will consider how to distinguish the rules governing the behavior of our two hypothesized RGP populations.

First, we will use statistical analysis and unsupervised machine learning to explore the relationships between clone sizes and features like induction time, location, and clone type, which we cover in Chapter 4. Second, we consider a different method for representing the distribution of glia per clone, formed from a discrete Gaussian mixture model. This type of model was used in a previous study to represent MADM data collected during neurogenesis rather than gliogenesis, and the model represented a specific hypothesis of deterministic clonal behavior during neurogenesis

[16]. We perform this analysis in Chapter 5.

Ultimately, the main principle that we emphasize going forward is that different clonal division rules, whether formulated from a GW process or any other model, affect the distribution of hypothetical clone sizes. Conversely, if we have a clone size distribution measured from data, we can test how well that distribution is represented by different clonal division rules. However, as we have already seen – we failed to reject the Chi-square null hypothesis for both the NGS and Slater Q_n models in their representation of glia per clone in G only clones – two models with differing levels of complexity may both reasonably describe the data. In this case, we chose the simpler Slater Q_n model to represent stochastic gliogenesis. In our remaining analysis, we will similarly attempt to identify rules that are specialized enough to represent the distribution of clone sizes in the NGS better than simpler rules, but that are themselves as simple as possible.

CHAPTER

4

SOM AND STATISTICAL ANALYSIS OF CLONAL DATA

In the previous chapter, we modeled glial production during the NGS using the traditional method of branching processes and compared the model output to clonal MADM data. This analysis showed that the glial production in the subset of clones containing only glia and no neurons (G clones) was reasonably represented by a simple stochastic Galton-Watson branching process with two cell types: glial progenitors gp and glia G. However, this model failed to represent the glia coming from clones with both glia and neurons (Mix clones).

This observation leads us to hypothesize about the presence of two populations of RGPs, where the first subpopulation produces neurons and then switches to producing glia, and the second subpopulation produces only glia. Currently, it is unknown whether this second population of purely gliogenic RGPs exists, and previous studies have suggested that all glia come from previously neurogenic RGPs [16, 41]. Under this assumption, we would say that all glial-producing RGPs undergo a switch from neurogenesis into gliogenesis, and any G clones that we observe in the NGS data were induced with MADM after this switch occurred. Thus, any previously produced neurons would not be observed in the MADM-labeled lineage. We would expect to see similar patterns of glial production between G and Mix clones if this were the case, since G clones would simply be the glial portion of Mix clones. Instead, we observed in Sec. 3.2.4 that the distributions of glia per clone in G and Mix clones were not represented by the same model.

At a glance, we can observe how glial production compares between G and Mix clones. Fig. 4.1 shows the means of the red or green glia produced per clone among different subsets of the data. Recall that in wildtype (WT) mice, the red and green sublineages are genetically identical, while in knockout (CKO) mice, the gene for epidermal growth factor receptor (EGFR) is deleted in the green lineage only. As expected, the CKO clones produce a smaller average number of green glia than WT clones, since gliogenesis is suppressed from the genetic knockout of EGFR. This reduction of green glia in the CKO clones holds for both G and Mix clones. Additionally, the average red glia in CKO clones appears to be increased as compared with the average red glia in WT clones, indicating that the red lineages alter their glial output to compensate for loss of green glia. However, other comparisons are less clear from simply examining the mean glia per clone. For instance, it is difficult to tell whether a difference in glial output exists between G and Mix clones, or whether the G and Mix clones compensate for the loss of green glia differently in CKO mice.

In this chapter, we aim to develop a clearer understanding of how RGP behavior progresses during the NGS. First, in Sec. 4.1 we use self-organizing map (SOM) clustering as in Chapter 2 to observe the broad differences in glial production between clones from different MADM labeling times, and between WT and CKO populations. Second, in Sec. 4.2 we directly compare the glia produced in different subsets of the population using the Wilcoxon Rank-Sum test. These comparisons result in further distinctions between the Mix and G clones, which we incorporate into model construction in Chapters 5 and 6.

4.1 Clonal analysis: Self Organizing Maps

4.1.1 Clone types as an indicator of development stage during the NGS

As time progresses through the switch from neurogenesis to gliogenesis, we would expect the percentage of G clones in each mouse to increase. We observed this phenomenon in Chapter 2 when comparing the WT mice labeled with MADM at E15.5, E16.5, and E17.5, though there was significant variability across different mice having the same time point label. Additionally, recall from Table 1.1 that the number of mice per time point and the number of clones per mouse are variable in the dataset. To allow a more even and robust comparison of the percentage of G clones at different time points, we utilize a clonal subsampling scheme.

4.1.1.1 SOM construction: subsampling clones

We begin by pooling together the clones coming from all mice at a single time point and genotype, resulting in a partition of the data into six groups: WT E15.5, E16.5, E17.5 (97, 148, 148 clones respectively) and CKO E15.5, E16.5, E17.5 (45, 57, 55 clones, respectively). A random sample of n clones was taken from each group; we chose $n = 32$ clones so that the group with the smallest

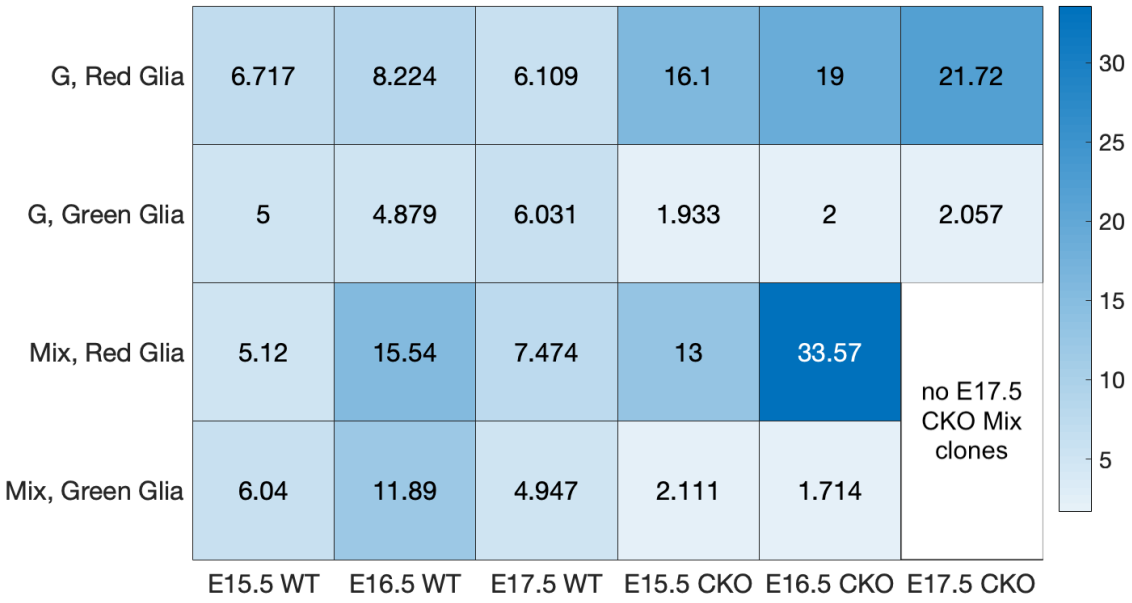


Figure 4.1 Average red glia and green glia per clone in subsets of NGS data.

number of clones, CKO E15.5, would be sampled at roughly 70%. The percentages of N, G, and Mix clones were calculated in each group's sample and recorded in a 3×1 vector. This random sampling and clone type percentage calculation was performed 100 times for each group. Thus, the end result was 600 vectors representing the percentages of N, G, and Mix clones in each sampled set of $n = 32$ clones. Table 4.1 shows five examples of the vector components for %N, %G, and %Mix clones calculated in samples of 32 clones from the group of E15.5 WT mice. To compare the change in G clones over time, we clustered the % G component of the 600 vectors in a 1×5 SOM (see Sec. 2.2) and examined the placement of clones from each group into the five clusters.

Table 4.1 Percent N, G, and Mix clones in five sample sets of $n = 32$ clones from E15.5 WT mice.

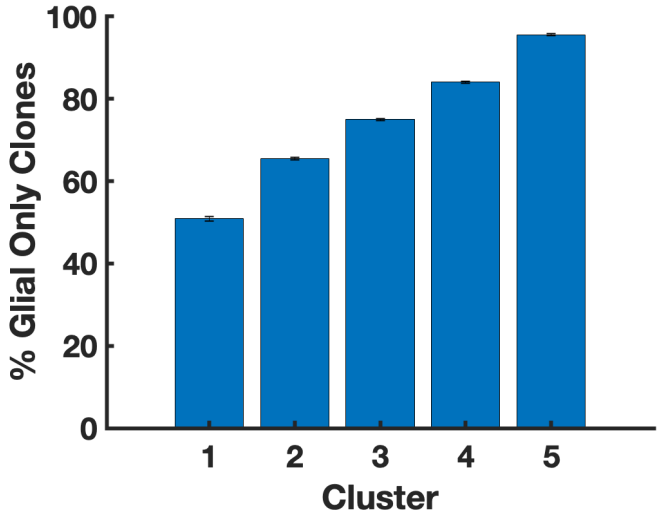
Sample set	1	2	3	4	5
%N	20.59	17.64	23.52	22.06	19.12
%G	50.00	58.82	55.89	52.94	54.41
%Mix	29.41	23.52	20.59	25.00	26.47

4.1.1.2 Results

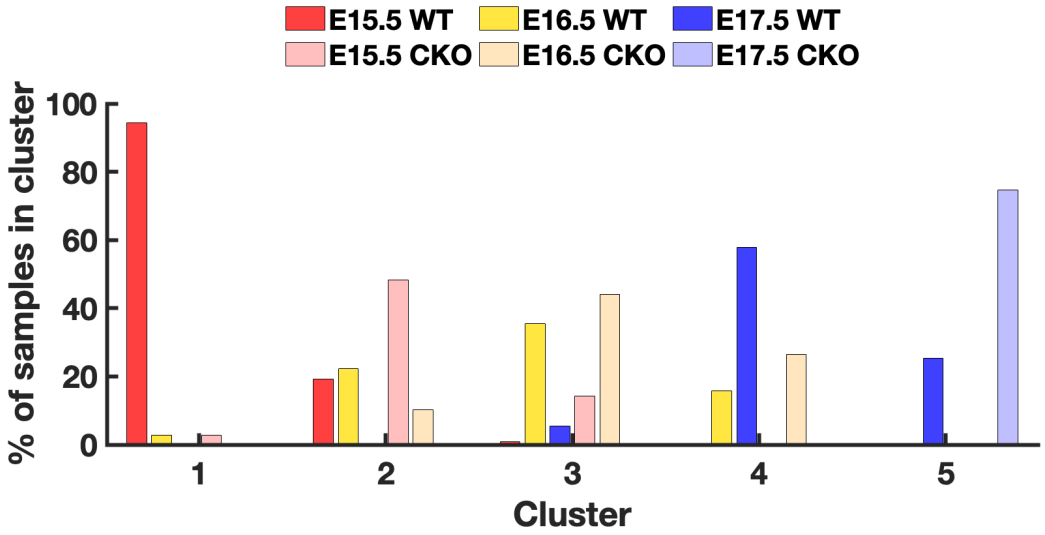
In Fig. 4.2(a), we show the average % G clones among the mice sorted into each cluster. The map has placed clones with lower percentages of G clones into the first cluster, and the percentage increases monotonically when traversing clusters in the map from left to right. Fig. 4.2(b) shows the percentage of samples in each cluster that come from each time point and genotype. Consistent with the previous results in Chapter 2, mice with more G clones come from later time points, and this is the case for both WT and CKO groups. However, there is an interesting shift in the placement of mice coming from the same time point and different genotypes. For each time point, the CKO mice are sorted into clusters farther to the right than the WT mice, indicating that CKO mice have a higher percentage of G clones as compared with their WT counterparts. This is an interesting observation; if we take the dataset of WT clones and simulate the effect of EGFR knockout by deleting all green glia, the result is a decrease in the percentage of G clones, not an increase. As an example, 54.64% of clones in E15.5 WT mice are G clones. Deleting green glia in all clones causes any Mix clones containing all green and no red glia to become N clones, and it also removes any G clones that contained only green and no red glia from the dataset altogether. The percentage of the remaining clones that are G clones is reduced to 44.30%. Thus, the increase in percentage of G clones in CKO mice does not appear to arise directly from the deletion of green glia. Instead, it implies that the knockout of EGFR alters RGP behavior in some way to result in more G clones. We can hypothesize about different mechanisms that could cause G clones to appear in larger numbers in CKO mice, depending on the assumption of one or two populations of RGPs.

First, we consider one population of neurogenic RGPs in which a fraction of the RGPs eventually switch from neurogenesis into gliogenesis (hypothesized to be 10-20% of RGPs [16, 41]), and all glia come from previously neurogenic RGPs. Here, the increased percentage of G clones could occur if the genetic knockout causes RGPs to switch from neurogenesis into gliogenesis earlier. This could occur via a biological mechanism in which RGPs sense the loss of green glia at the start of the NGS and compensate quickly by increasing the rate at which RGPs switch into gliogenesis. The observed increase in the percentage of G clones would thus arise from MADM labeling occurring more often in RGPs that have already switched into gliogenesis. In Fig. 4.3, we illustrate this phenomenon by representing the NGS over E15.5 to E17.5 as a gradient from red to yellow, corresponding to an increase in G clones over time. For the WT case, we show the NGS peaking around E16.5, where fewer G clones are produced prior to E16.5 and more are produced after this time. For the CKO case, we show the NGS occurring earlier, which shifts the gradient of increasing G clones over time to the left. As a result, the CKO case shows more G clones at each time point as compared with the same time point in the WT case. Thus, the increased observation of G clones in CKO mice could occur if the NGS is sped up in response to the EGFR knockout.

Second, we consider the population of RGPs that switches from neurogenesis into gliogenesis as



(a)



(b)

Figure 4.2 Results of clustering the values of %G clones in 600 sample sets of $n = 32$ clones into five clusters using a 1×5 SOM. The number of sample sets sorted into each cluster ranged from 71 to 166. The plot in (a) shows the mean ± SEM of the %G clones value for the sample sets sorted into each cluster. For instance, the sample sets in cluster 1 had mean %G value of 50%. The sample sets sorted into clusters farther to the right in the map contained more G clones. In (b), we show the percentage of samples sorted into each cluster coming from each of the six groups listed. Samples with a lower value of %G clones on the left of the map predominantly came from E15.5 mice, and traversing the map from left to right shows a temporal shift to E16.5 and E17.5 as the average value of %G clones in each cluster increases.

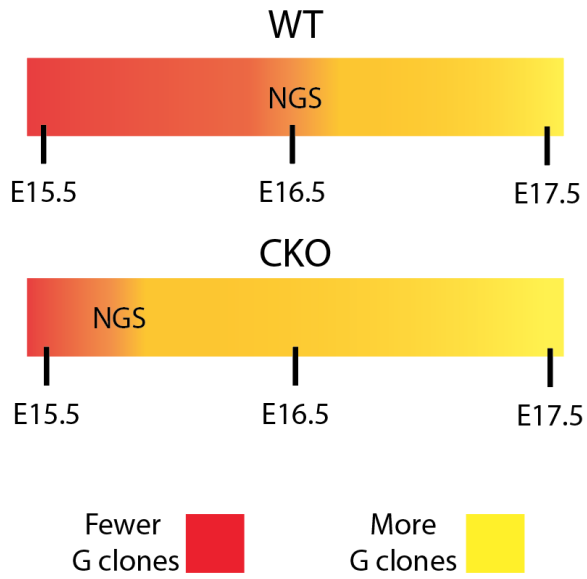


Figure 4.3 Conceptual diagram illustrating the result of an earlier occurrence of the NGS. The gradient from red to yellow shows the increase in G clones in the observed data as the NGS proceeds. If we hypothesize that the NGS is sped up in the CKO case as compared with the WT case and occurs at an earlier time point, then the CKO clones at E16.5 and E17.5 will have increased numbers of G clones as compared with the WT clones from the same times.

well as a second population of RGPs that produce only glia. We will refer to the former population as NGS-RGPs, and the latter as glial-producing RGPs, or G-RGPs. In this case, the increase in G only clones could simply occur if the G-RGPs respond differently than the NGS-RGPs to the deletion of EGFR. For instance, EGFR deletion could cause increased proliferation of G-RGPs, but not in NGS-RGPs, which would increase the proportion of G-RGPs in the entire population and accordingly increase the proportion of G clones. A similar proportional increase in the G-RGP population could occur from an influx of G-RGPs migrating from other regions of the developing cortex; it has been demonstrated that some RGPs migrate tangentially in this manner during cortical development [32].

The hypothesis of one or two populations of RGPs brings us to the question of deterministic versus stochastic mechanisms of RGP behavior during the NGS. Recall from Sec. 1.4.1 that a stochastic mechanism implies that all RGPs would be affected similarly from EGFR knockout, whereas a deterministic one implies that some RGPs may respond differently due to different fate specifications. We can gain support for one of these options by examining the distributions of glia per clone in separate subsets of the data for similarities or differences. Thus, in Sec. 4.2 we will use statistical tests to quantify the clonal response to EGFR knockout, focusing on comparing the glial production of G versus Mix clones.

4.1.2 Glial Production, Wild Type and Knockout

4.1.2.1 SOM construction

Next, we perform SOM clustering using the two-component vectors of total neurons and total glia per clone. For fair and robust comparison between the different time points and genotypes, we again implement a subsampling scheme. As with the previous SOM in Sec. 4.1.1, $n = 32$ clones were sampled from each of the six data groups. The red and green components of each clone were summed to give total neurons and total glia per clone. The resulting 192 2×1 vectors were then clustered with a 1×5 SOM. This sampling was performed 500 times and a 1×5 SOM formulated for each case.

For each SOM, we observed that clones were sorted predominantly based on the total glia per clone, since this component had a wider range than the total neurons component. To allow for comparison between the different clusterings, each SOM was arranged after clustering so that the cluster containing clones with the most glia was on the left of the map and denoted cluster 1, and the average glia per clone decreased in the clusters going from left to right.

4.1.2.2 Results

Fig. 4.4(a) shows the average glia and neurons in the clones sorted into clusters 1-5. Clone size appears correlated with location, with small clones found predominantly in superficial and deep layers, and large clones found among all layers (Fig. 4.4(b)). This makes sense biologically; RGPs produce neurons and glia in layers from deep to superficial over time (see Sec. 1.1.2), so all layer clones would be counting an entire clonal lineage, whereas deep or superficial ones would only be counting a subset of a lineage's cells.

It is interesting then to observe the placement of the six time and genotype groups among the clusters. Cluster 1, having clones with very large numbers of glia that are predominantly found in all layers, is primarily comprised of E16.5 CKO and E17.5 CKO clones. A small number of E16.5 WT clones also fall into this cluster, indicating that an increase in glial output occurs at E16.5 regardless of any genetic perturbation. The E15.5 CKO clones are not seen in Cluster 1, suggesting that any increase in glial output in response to the genetic knockout occurs after E15.5. It may therefore be the case that the naturally occurring increase in glial output at E16.5 is amplified and extended temporally to E17.5 clones in response to the genetic knockout.

However, the WT and CKO clones sorted into Cluster 1 differ by clone type. Within Cluster 1, all of the E16.5 WT clones are Mix clones, thus the natural increase in glial output appears to come from clones switching from neurogenesis into gliogenesis around E16.5. Among the E16.5 CKO clones in Cluster 1, only 38.38% are Mix clones, and the remaining 61.62% are G clones. The E17.5 CKO clones in Cluster 1 are all G clones. Thus, the clones producing large numbers of glia in CKO

mice are predominantly G and not Mix clones, contrasting with WT mice.

Again, this difference between WT and CKO could occur from a mechanism operating on one or two populations of RGPs. If only one population of RGPs existed and the EGFR knockout 'speeds up' the NGS as described in Sec. 4.1.1.2, then the surge in glial production from Mix clones would occur earlier in CKO mice, perhaps at E15.5 instead of E16.5. Accordingly, if these Mix clones with large numbers of glia occur at E15.5, then the glial only portions of their lineage would be observed at E16.5-E17.5. This would explain the large numbers of glia in G CKO clones at these later times. However, we do not actually observe large numbers of glia in E15.5 CKO clones, Mix or G, as no E15.5 clones are located in Cluster 1 in Fig. 4.4(c). On the other hand, if we hypothesize about separate populations of NGS-RGPs and G-RGPs, the surge of glial production in E16.5 and E17.5 CKO G clones could easily be explained by an increase in G-RGP proliferation that occurs after E15.5 and thus does not appear in E15.5 clones. In the next section, we highlight more differences between glial output by time and clone type with the goal of distinguishing the two populations of RGPs.

4.2 Statistical Comparison of Clones

4.2.1 Wilcoxon Rank-Sum test

Next, we perform pairwise comparisons of clonal output between different subsets of the NGS data using the Wilcoxon Rank-Sum test [37]. This method, also known as a Mann-Whitney U test, is an analysis of variance (ANOVA) technique adapted for nonparametric data. That is, the test enables comparing groups of data without the assumption that the data is normally distributed. The test takes as input two sets of data in which the observations in each data set are

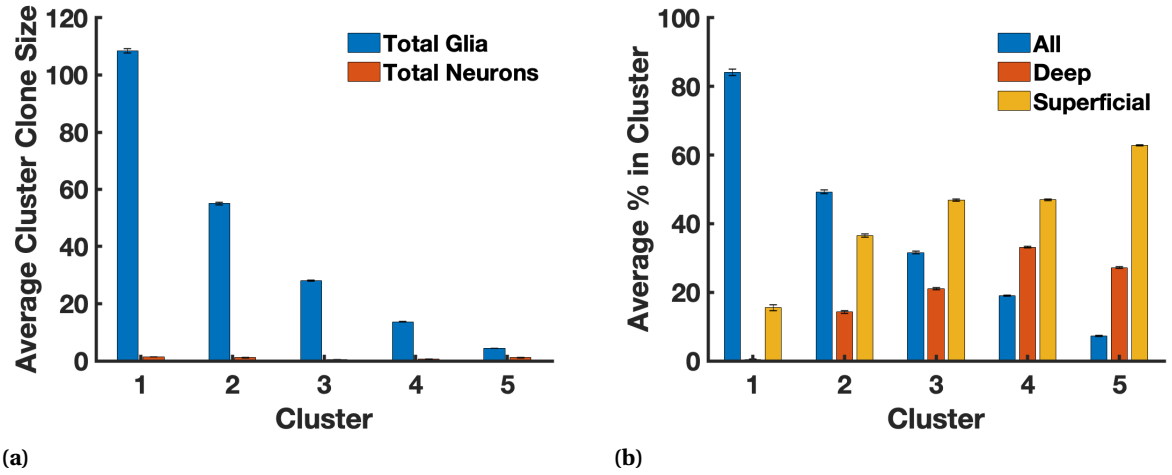
1. independent
2. able to be clearly ordered from smallest to largest.

The two data sets do not need to have the same number of elements, though a large difference in sample size can produce a less reliable result.

The null hypothesis H_0 and alternative hypothesis H_1 for the Wilcoxon Rank-Sum test are as follows:

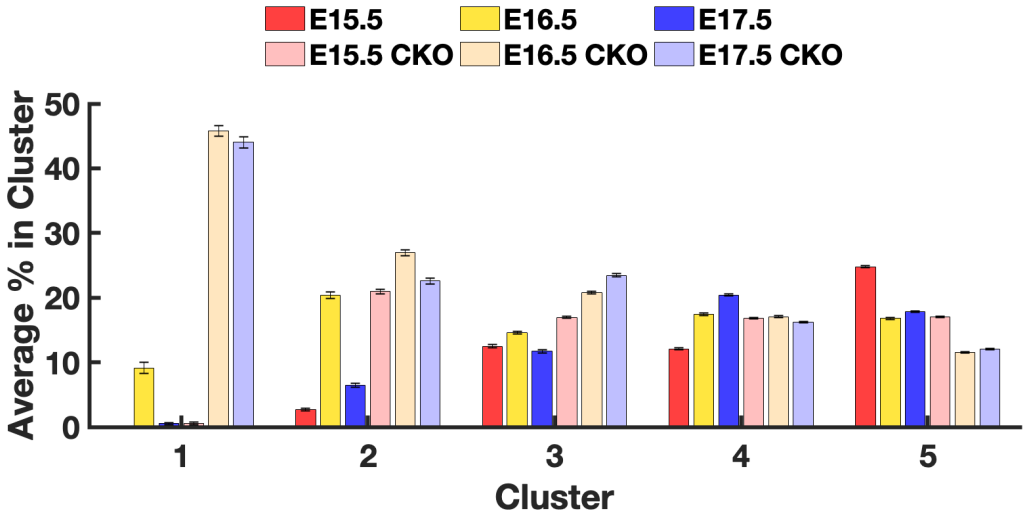
1. H_0: The populations in the two data sets are equal in distribution.
2. H_1: The populations in the two data sets have different distributions.

The test calculates a p value, which indicates rejection of the null hypothesis if p is below a certain significance level α. We consider the standard $\alpha = 0.05$ to determine the rejection of the null hypothesis.



(a)

(b)



(c)

Figure 4.4 Results of clustering the vectors of total glia and total neurons per clone into five clusters via a 1×5 SOM. The average number of clones sorted into each cluster ranged from 8 in cluster 1 to 72 in cluster 4. (a) Mean ± SEM of the components of total glia and total neurons in the clones sorted into each cluster. We see that the SOM sorted clones based on the total glia component, with clones in the clusters on the left of the map having many glia, and clones in the clusters on the right of the map having very few. (b) Percentage of clones sorted into each cluster coming from each of the three cortical tissue locations: all layers, deep, or superficial. All layer clones are highly represented in cluster 1 with large glial clones, whereas deep and superficial clones occur more often in the clusters with small glial clones. (c) Percentage of clones in each cluster by time point and genotype. E16.5 CKO and E17.5 CKO clones are dominant in cluster 1 with large numbers of glia per clone.

In the NGS data we consider, clone sizes are not normally distributed about some mean (see the distribution of glia per clone in Fig. 3.11), but they can be clearly ordered. Additionally, all clones are observed independently, since sampling of RGPs in the cortex occurs randomly. Thus, we satisfy the above assumptions and can use this test to determine whether the production of glia differs by genotype, clone type, and time. The data is considered at the clonal level, with clones pooled among all mice in the respective groups being compared.

4.2.2 Results of comparisons

As a simple introductory example, we compare the sets of red and green glia per clone using the Wilcoxon Rank-Sum test. From Fig. 4.1, we might expect to reject the null hypothesis if we compared the red and green glia in CKO clones, but likely not reject the null hypothesis for the same test in WT clones. Table 4.2 shows the p-values resulting from comparing red to green glia in clones coming from four subsets of the data: G WT, Mix WT, G CKO, and Mix CKO. In both WT cases, $p>0.05$ and we cannot reject the null hypothesis. Thus, WT clones produce a similar distribution of red and green glia with no preference for red over green. Performing the test on red versus green glia in CKO clones results in a rejection of the null hypothesis for both G and Mix clones, confirming that the genetic knockout causes a significant difference in the production of red versus green glia.

Table 4.2 p-values generated from the Wilcoxon Rank-Sum test comparing red and green glia per clone for WT G, WT Mix, CKO G, and CKO Mix clones.

	Red glia vs green glia
G, WT	0.1172
Mix, WT	0.5398
G, CKO	3.3508e-14*
Mix, CKO	1.500e-3*

We perform three more tests to compare the distributions of glia per clone produced in subsets of the population broken down by genotype (WT and CKO), time point (E15.5, E16.5, E17.5), clone type (G and Mix), and in some cases, sublineage (red and green):

Total glia per clone in G versus Mix clones. First, we compare the total glia per clone in the sets of G and Mix clones, broken down by time point and genotype. We note that there are no E17.5 CKO Mix clones, so the comparison to G clones could not be performed for this group. Based on the p-values from the WT comparisons in Table 4.3, G and Mix clones produce similar distributions of glia per clone at E15.5 and at E17.5. At E16.5, the distributions differ, with Mix clones producing on average more glia than G clones. This is consistent with the previous phenomenon discussed in Sec. 4.1.2, where a natural surge in glial production occurs in WT Mix clones at E16.5. Interestingly, in

the E16.5 CKO population, this difference in glial output is not observed between G and Mix clones (p=0.2036).

Table 4.3 p-values generated from the Wilcoxon Rank-Sum test comparing glia per clone in Mix versus G clones. The comparison is done for subsets of clones separated by time and genotype.

	WT G clones vs WT Mix clones	CKO G clones vs CKO Mix clones
E15.5	0.9957	0.5593
E16.5	2.9401e-04*	0.2036
E17.5	0.5748	no Mix clones

Total glia per clone in WT versus CKO clones. Next, we compare the output of glia per clone between the WT and CKO populations. The p-values for the comparison tests between different subsets of WT versus CKO clones are shown in Table 4.4. For G clones, the distribution of glia produced in WT versus CKO clones is significantly different, with p-values well under 0.05; this holds whether we compare the red sublineages only, the green sublineages only, or the total glia per clone. For Mix clones, the genetic knockout appears to influence the distribution of glia in the individual sublineages (p=0.0224 for red, p=0.0672 for green), but the distribution of total glia per clone is not altered between WT and CKO Mix clones (p=0.2956).

Table 4.4 p-values generated from the Wilcoxon Rank-Sum test comparing glial output in WT versus CKO clones. The tests are performed for the sets of red glia, green glia, and total glia.

	Red Glia, WT vs CKO	Green Glia, WT vs CKO	Total Glia, WT vs CKO
G clones	7.5387e-09*	6.7019e-05*	1.2571e-05*
Mix clones	0.0224*	0.0672	0.2956

Total glia per clone by time point. G and Mix clones additionally show a difference when comparing the glia produced from separate MADM induction times. For G clones, the distribution of glia produced is similar between all three time points, both in WT and CKO populations (Table 4.5). WT Mix clones produce a similar distribution of glia at E15.5 and E17.5 (p=0.4766), but the distribution of glia per WT Mix clone differs at E16.5 from the other two time points. Again, this is consistent with the phenomenon of a surge in glial production in E16.5 Mix clones.

Table 4.5 p-values for comparing total glia per clone by time point for WT G, WT Mix, CKO G, and CKO Mix clones.

	E15.5 vs E16.5	E16.5 vs E17.5	E15.5 vs E17.5
G, WT	0.6176	0.7495	0.8160
Mix, WT	0.0031*	0.0208*	0.4766
G, CKO	0.8840	0.6580	0.6285
Mix, CKO	0.1656	no E17.5 CKO Mix	

4.3 Discussion of clonal behaviors in the NGS

The SOM and statistical tests presented in this chapter have shown several trends regarding glial production during the NGS as it relates to time and clone type. We now aim to summarize the observed trends to understand how the NGS proceeds in the WT case, and how the knockout of EGFR affects RGP behavior during the NGS.

4.3.1 NGS in WT and CKO clones

First, we address our hypothesis of one or two separate populations of RGPs with different behavior. In WT clones, we consistently observed a phenomenon in which E16.5 Mix clones produce more glia than both G clones and Mix clones from other time points. The distribution of glia per clone in WT G clones did not differ at different time points (Table 4.5). If G clones all arose as observations of the later part of Mix lineages, we would expect to see a temporal difference in their production of glia as was observed in Mix clones, but this is not the case. Additionally, we observed a different response to EGFR knockout in G and Mix clones, with CKO G clones producing more glia than their WT counterparts, but CKO Mix clones producing a similar distribution of total glia as WT Mix clones, albeit with more red and fewer green glia. As described in Sec. 1.4.1, a different response to genetic perturbation between subsets of a cell population can indicate different deterministic fate specifications in these subsets. Thus, we gain support for the hypothesis that two deterministically different populations of RGPs exist: the NGS-RGPs, which produce neurons and then glia (and result in Mix clones if labeled prior to the switch and G clones if labeled post-switch), and G-RGPs, which only produce glia (and hence can only be G clones).

Under the hypothesis of the existence of these two populations of RGPs, we can describe how the NGS proceeds in the WT case according to the SOM and statistical test results. Between E15.5 and E17.5, NGS-RGPs switch from neurogenesis into gliogenesis. Those that undergo this switch around E16.5 end up producing more glia than those that switch around E15.5 or E17.5. This could be due to increased expression of EGFR at E16.5 promoting gliogenesis as described in Sec. 1.1.2, but other unknown factors could also contribute. The subpopulation of G-RGPs produce consistent numbers

of glia per clone from E15.5 to E17.5, which suggests a steady migration of glia from G-RGPs into the cortex. That is, if all G-RGPs were present and able to be labeled with MADM from the earliest time point, E15.5, then we would expect to see the number of glia per clone decrease in G clones at E16.5 and E17.5, since these G clones would be subsets of the E15.5 G clone lineages. Instead, the distribution of G clones sizes is consistent over time, suggesting that the G clones labeled with MADM at E16.5 and E17.5 are not from the same population as those labeled at E15.5.

We can also describe how the NGS proceeds differently in the CKO case, based on the hypothesis of separate behavior between NGS-RGPs and G-RGPs. NGS-RGPs appear to have a similar surge in glial production at E16.5 as in the WT case, evidenced by the E16.5 CKO Mix clones being clustered with E16.5 WT Mix clones in the clonal SOM in Fig. 4.4. EGFR deletion shifts the production of red and green glia in NGS-RGPs, but leaves the total glia unchanged (Table 4.4). On the other hand, G-RGPs increase their glial output in the CKO case at E16.5 and E17.5, but not at E15.5 (Fig. 4.4). Overall, the total glia produced by G-RGPs is increased in the CKO case (Table 4.4). Furthermore, more G clones exist in the CKO case. Returning to the discussion of two populations of RGPs from Sec. 4.1.1.2, this could be from a migration of G-RGPs from another location in the developing brain, or simply from increased proliferation that increases G-RGP population.

4.3.2 Deterministic versus stochastic: clonal level

The analysis presented in this chapter supports the hypothesis of two deterministically fate-specified subpopulations of NGS-RGPs and G-RGPs, which are distinguished from one another by different patterns of glial production. However, within those two populations, at the clonal level, we can label the patterns of glial production as deterministic or stochastic. These terms as they relate to clonal output are less well-defined in literature than at the level of the population. Summarizing what previous studies have used to define 'deterministic' clonal behavior, we will propose that possible indications of clonally deterministic mechanisms include a predictable, normally distributed number of differentiated cells per clone, a pairing of fates between the sublineages of individual RGPs, and a resistance to fate alteration via genetic reprogramming (see Sec. 1.4.1). Clonally 'stochastic' behavior may be defined by the absence of these deterministic signals.

Comparing the observations for Mix and G clones thus far appears to indicate a clonally deterministic mechanism for NGS-RGPs, but a clonally stochastic mechanism for G-RGPs. Regarding the former group, Mix clones produce similar total glia per clone even after deletion of the gene for EGFR, pointing to a resistance in the NGS-RGP population to the alteration of total glia production. For the G-RGPs, total glia production was increased in response to genetic alteration. Additionally, the distribution of glia per clone in G clones is not normal but was found to be sufficiently represented by a distribution Q_n matching the output of a stochastic Galton-Watson (GW) branching process in Sec. 3.2.4. The similarity of glial production from G clones at successive time points is

further evidence of behavior matching a stochastic GW process; these processes hold to the Markov property (see Eqn. 3.7), which makes them self-similar in time.

Thus, in developing a model for clonal behavior during the NGS, we so far can represent the stochastic behavior of G-RGPs with a GW process. The deterministic behavior of NGS-RGPs remains to be understood. In the next chapter, we test whether a clone size distribution based on a Gaussian mixture model, intended to represent clonally deterministic behavior, can represent the distribution of glia per clone in Mix clones in the NGS data. If so, then we can define a set of deterministic rules for the behavior of NGS-RGPs in Chapter 6.

CHAPTER

5

CLONAL DISTRIBUTION ANALYSIS: GAUSSIAN MIXTURE MODELS

Based on the analysis of clone size distributions in Chapter 3 and the statistical comparison of groups of clones in Chapter 4, we have suggested that two subpopulations of radial glial progenitors (RGPs) exist during cortical development and the neurogenesis-to-gliogenesis switch (NGS). One subpopulation, which we denote G-RGPs, produces glia in a stochastic manner that can be representing using a branching process. This type of model was consistent with the glia produced by clones containing no neurons (G clones). The other subpopulation, denoted NGS-RGPs, is hypothesized to produce glia in a deterministic manner, but we have not yet determined a model that can represent a deterministic distribution of glia per clone. In our MADM dataset, this subpopulation of corresponds to the Mix clones, which produce both neurons and glia. Thus, in this chapter, we aim to test whether a deterministic model is sufficient to represent the distribution of glia per clone in Mix clones. Namely, we test the goodness of fit of a multi-Gaussian mixture model with linearly spaced peaks; this particular model was proposed as a way of describing clonal output from deterministically behaving RGPs during neurogenesis [16].

We begin this chapter, in Sec. 5.1, by highlighting two features of clonal MADM data that can be used as indicators of specific proliferation and differentiation patterns in clonal lineages: symmetry and recursion. Next, we explain how these features were analyzed in clonal MADM data collected during neurogenesis in [16], both at the level of individual clones and at the level of the distribution

of neurons per clone in the population. Lastly, we adapt and expand on their techniques to test whether glia are produced during the NGS deterministically or stochastically in Sections 5.3-5.6. The hypothesized mechanism of clonal behavior that we identify will be simulated in Chapter 6.

5.1 Inferring clonal division history from MADM data

We recall that the MADM data gathered during the NGS does not measure the entire history of each labeled RGP's divisions, but only provides the number of differentiated cells in the developed cortex that are descendants of that RGP. Knowing the precise history of divisions would aid in judging whether RGP expansion and differentiation occur deterministically or stochastically, since we would directly know every cell's fate. Although this precise history is unknown, two features of clonal MADM data can help determine which mechanism most likely drives the proliferative capacity and division history of a clonal population.

5.1.1 Symmetry

First, comparing the red and green sublineages of each MADM clone can give insight into whether specific types of cell divisions correspond to clonal output. We note that the number of differentiated cells in a clone's red and green sublineages indicate what type of division the original MADM labeled RGP underwent (Fig. 5.1). If both sublineages contain one cell, the RGP completed a *symmetric differentiating division*. This case is trivial since there is no uncertainty about what type of division occurred, disregarding the possibility of cell death. More importantly, if one sublineage (red or green) contains only one cell and the other (green or red, respectively) contains more than one, then the initial RGP performed an *asymmetric differentiating division*. If both sublineages contain more than one cell, the initial division was a *symmetric proliferative division*, since the red and green sublineages would each need an RGP at their roots to produce at least two differentiated cells.

We will thus classify clones as 'asymmetric' if one sublineage contains one cell and the other contains more than one, or 'symmetric' if both sublineages contain more than one cell. The clone sizes in the subpopulations of asymmetric and symmetric clones can be examined separately to determine if an RGP's initial division affects the total number of differentiated cells it produces. If a clear pattern in clonal output exists for asymmetric or symmetric clones, we would have evidence of a deterministic relationship between clone size and initial division type.

5.1.2 Recursion

Second, the numbers of differentiated cells in subsequent generations of the lineage are related to one another by recursion. This property was previously used in Sec. 3.1 to develop an estimate for the rounds of division h given a clone size l. In addition, the recursive property of clonal lineages

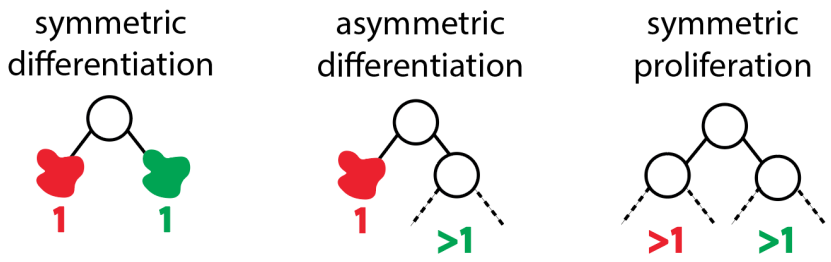


Figure 5.1 Effect of the initial RGP division on the total differentiated cells. A *symmetric differentiation* produces two cells, one red and one green (left). *Asymmetric differentiation* (center) produces one red/green cell and more than one green/red cell. A *symmetric proliferative division* (right) is required to produce greater than one cell in both red and green sublineages.

can enable reconstruction of division histories given a sample of MADM clones at different time points. This is best illustrated if we think of a clonal lineage as a rooted binary tree whose leaves are differentiated cells, as in Sec. 3.1. Recall that the number of leaves descending from a node is the sum of the leaves in the node's left and right subtrees. For MADM data, this translates to the number of cells descending from an initial RGP being equivalent to the sum of its red and green cells. Fig. 5.2 shows two possible scenarios of MADM labeling in an RGP lineage. If the RGP 'a' was labeled with MADM, the observed clone would count 3 red and 10 green glia for a total of 13 cells. However, consider if MADM labeling instead occurred in the RGP 'b'. In this case, the observed clone would have 2 and 1 glia in its respective red and green lineages. If we observed these two clones in the data and knew that one had been labeled with MADM during a later development stage than the other, then we could infer that RGP 'b' producing 2+1=3 glia was the same RGP from the red sublineage of RGP 'a', also producing 3 glia.

Thus, since MADM tracks a random sample of clones at multiple time points, the clones measured at later time points can be thought to represent sublineages descending from clones present at earlier time points. A rooted full binary tree can be reconstructed if the number of leaves in each of its subtrees are provided [36]. Therefore, if a sufficient number of clones are measured at multiple time points, we can consider that these clones provide a representation of a lineage and its sublineages.

By treating a sample of MADM clones collected at subsequent time points as representing sublineages of a global lineage process, we can use the distribution of clone sizes in the sample to infer RGP proliferation and differentiation histories. For instance, we may notice by observing peaks in the clone size frequency distribution that clones of size 5 and 8 are very common. If 13 is also a common clone size, we may assume that the clones producing 13 cells tend to do so by having two sublineages with 5 and 8 cells, respectively.

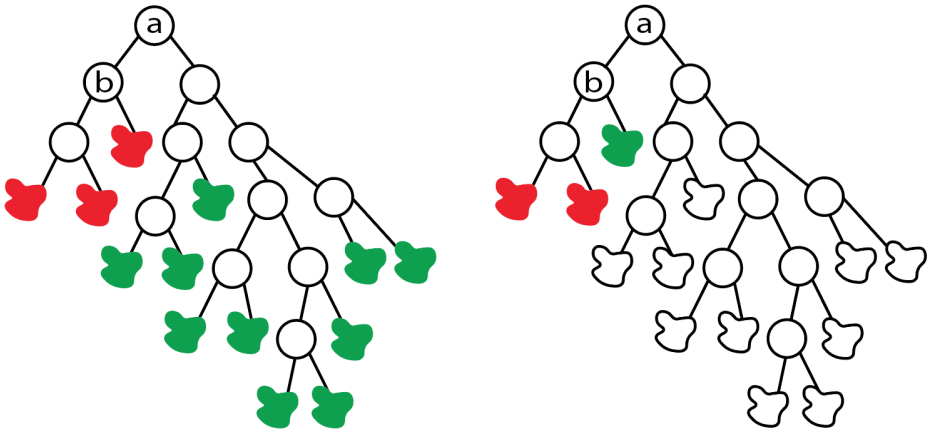


Figure 5.2 Recursion of cell counts from MADM labeling in successive generations of a clonal lineage. Labeling the initial progenitor (a) results in all glia being counted, 3 red and 10 green, while labeling progenitor (b) would only count the 3 left glia as 2 red and 1 green.

5.2 Modeling clonal distributions in neurogenesis

As described in Sec. 1.1.2, the RGP population in the embryonic mouse cortex proliferates through symmetric divisions prior to E10. Experimental observations suggest that upon entering the phase of neurogenesis after E10, RGPs switch to asymmetric divisions, producing one neuron and one self-renewed RGP at each division step [39]. This continues until a final symmetric differentiation into two neurons, a switch into gliogenesis, or the death of the RGP.

Recently, a set of deterministic rules defining RGP behavior during neurogenesis was proposed in [16] based on their analysis of clonal MADM data. The identification of these rules would not have been possible without the MADM technique, as their analysis relied on assumptions regarding clone symmetries and clone size frequencies. In their study, clonal MADM data was gathered in the embryonic mouse cortex at earlier stages of development corresponding to the peak of neurogenesis: E10-E13 (see Fig. 1.3). These clones produced large numbers of neurons, predominantly with no glia, in contrast with the glial-dominant clones gathered with inductions during the NGS from E15.5-E17.5 that we have in our own data.

5.2.1 Asymmetric neural clones

When examining the subset of asymmetric neural clones, it was found that the distribution of neurons per clone was normal with mean μ_0=8.4 and standard deviation σ=2.6 (Fig. 5.3a). This indicated that clones which have begun neurogenesis and are undergoing asymmetric differentiating divisions have a limited capacity to produce neurons, and do not randomly exit the cell cycle, but rather favor producing an average number of ≈ 8 neurons before terminating division. The value of

$\mu_0 = 8.4$ also appeared in the distribution of clone sizes for all clones, asymmetric or not, with peaks in the distribution appearing at integer multiples of μ_0 (Fig. 5.3c).

5.2.2 Symmetric neural clones

Symmetric neurogenic clones were defined in [16] as those having four or more neurons in both red and green sublineages. The symmetric cell counts imply that these clones were labeled with MADM during the proliferative expansion phase, prior to the onset of asymmetric differentiation during the phase of neurogenesis. Accordingly, the total neurons produced by symmetric clones tended to be larger. The largest symmetric clones were the earliest labeled ones at E10, and clone sizes decreased through E11 and E12, implying that E11 and E12 clones were subclones of the same RGP lineages present initially at E10 (Fig. 5.3b). That is, the feature of recursion appears to be present in neurogenic clonal MADM lineages measured at subsequent time points.

Most interestingly, the ratio of the larger to smaller subclones was concentrated between a 1:1 and 2:1 ratio, with an average ratio of ≈ 1.6 [16]. In their analysis, this ratio was taken to indicate that subclones either undergo the same number of proliferative divisions before the onset of asymmetric neurogenesis, which would result in an average 1:1 ratio of neurons in the larger to smaller subclone, or the generations of proliferative divisions may be offset by one, which would result in an average 2:1 ratio of neurons in the larger to smaller subclone. Thus, it was argued that the offspring of proliferating RGPs do not randomly begin asymmetric neurogenic divisions, but instead time their onset of neurogenesis closely in sync with their subclone siblings. This close relationship between neurogenic capacity in subclones was taken as further evidence of deterministic behavior of RGPs during neurogenesis [16]. We additionally point out that maintaining the same average ratio of subclone sizes at all levels of a lineage implies a fractal relationship, which would strongly support a deterministic level of organization governing cell population growth during this stage of cortical development.

5.2.3 Deterministic mechanism and observation in full neural population

Overall, these observations of clone sizes and symmetries were used to hypothesize a two-step deterministic process that all neurogenic RGPs follow. In the proposed process, RGPs first proliferate, favoring symmetric divisions and producing a discrete number of daughter RGPs $R = 1, 2, 3, \ldots$ as offspring. The R RGPs are distributed between the two sublineages in a ratio between 1:1 and 2:1, hence if one sublineage undergoes n proliferative divisions, the other sublineage undergoes $n \pm 1$ proliferative divisions. Each daughter RGP then begins neurogenesis by switching to asymmetric differentiating divisions, producing an average of $\mu_0 \approx 8$ neurons each. Thus, an initial RGP giving rise to R RGP offspring, each producing an average of μ_0 neurons, produces $R\mu_0$ neurons on average. The result of this process is the presence of peaks in the distribution of neurons per clone at μ_0, $2\mu_0$,

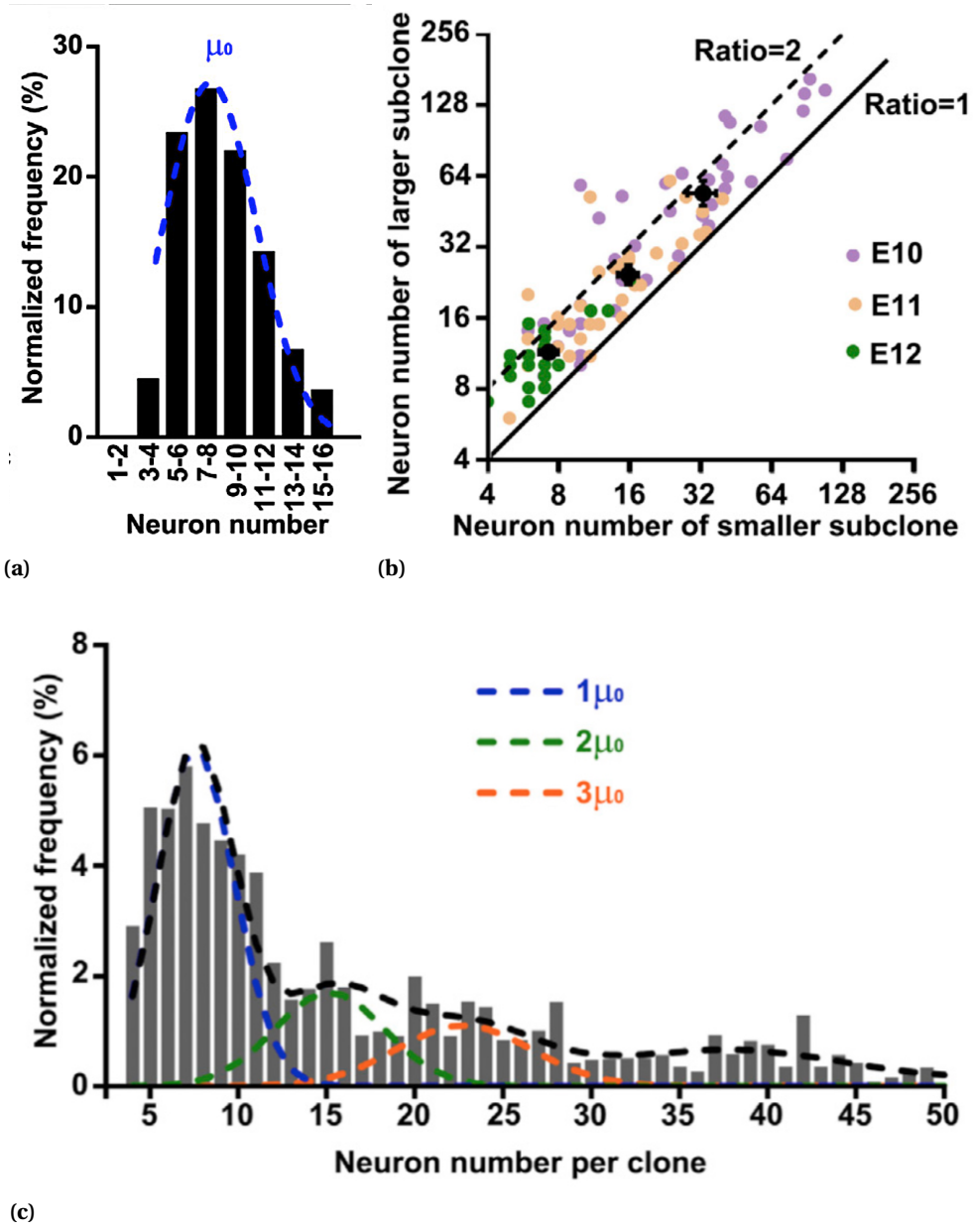


Figure 5.3 (a) Distribution of total neurons per clone among neurogenic clones with asymmetric red and green cell counts from [16]. The distribution is normal with $\mu_0 = 8.4$ and $\sigma = 2.6$. (b) Ratio of larger to smaller subclones in symmetric neural clones from [16]. (c) Distribution of total neurons per clone up to size 50 [16]. The Gaussian mixture model curve fit is shown (black dashed line) with three of the individual Gaussians in the model separately plotted. Reprinted with permission from [16]. Copyright 2014, The Authors. Published by Elsevier Inc.

$3\mu_0$, etc, corresponding to $R = 1, 2, 3$ respectively (Fig. 5.3b, colored dashed lines). The distribution can therefore be represented by a Gaussian mixture model with means located at integer multiples of μ_0. We will describe this type of model in more detail in Section 5.4.

5.3 Clone sizes during the NGS

The success of using clone sizes and symmetries to infer deterministic clonal behavior during neurogenesis suggests that a similar approach may be useful for NGS data. We examine the normalized frequency histograms of total glia per clone for asymmetric and symmetric wild type (WT) G clones to observe whether deterministic clone size patterns appear. Note that this analysis does not include Mix clones, since we want to strictly consider when the production of glia is asymmetric or symmetric in the red and green lineages. We then examine the frequency distribution of glia in Mix clones separately, as well as the distribution of glia in the combined population of G and Mix clones.

5.3.1 Asymmetric G clones

Unlike neurogenesis, gliogenesis is not known to be defined by RGPs undergoing a series of asymmetric divisions. Not surprisingly, the distribution of total glia per clone in WT G clones having asymmetric glial counts shows a different pattern than the distribution of asymmetric neural clones. Fig. 5.4a shows the distribution of total glia in asymmetric WT G clones. As described in the previous section, the Gaussian shape of the asymmetric neural clone size distribution (Fig. 5.3a) pointed to a limited, deterministic capacity for neuron production in asymmetrically differentiating RGPs. For asymmetric glial clones, the range of sizes in the distribution indicates a greater capacity for glial production after an asymmetric division occurs. This suggests that the clonal rules governing glial production in WT G clones do not match the deterministic rules for neurogenesis.

It is also unlikely that this range of sizes would be the result of a series of sequential asymmetric divisions. During neurogenesis, RGPs require ≈13-19 hours to undergo one asymmetric differentiating division, producing one neuron [9]. If this cell cycle length is similar for asymmetric gliogenic differentiation, producing 50 glia with a series of asymmetric divisions could take up to 40 days, far longer than the time frame of the NGS. Instead, it is more likely that for asymmetric glial clones, symmetric proliferative divisions occurred following the initial asymmetric division; symmetric proliferative cell cycles are shorter and double the cell population with each division, producing more cells in a shorter amount of time than asymmetric divisions [9]. Thus, we do not have evidence that an initial asymmetric glial RGP division begins a series of asymmetric divisions, constrains the RGPs gliogenic output, or results in a deterministically predictable number of glia.

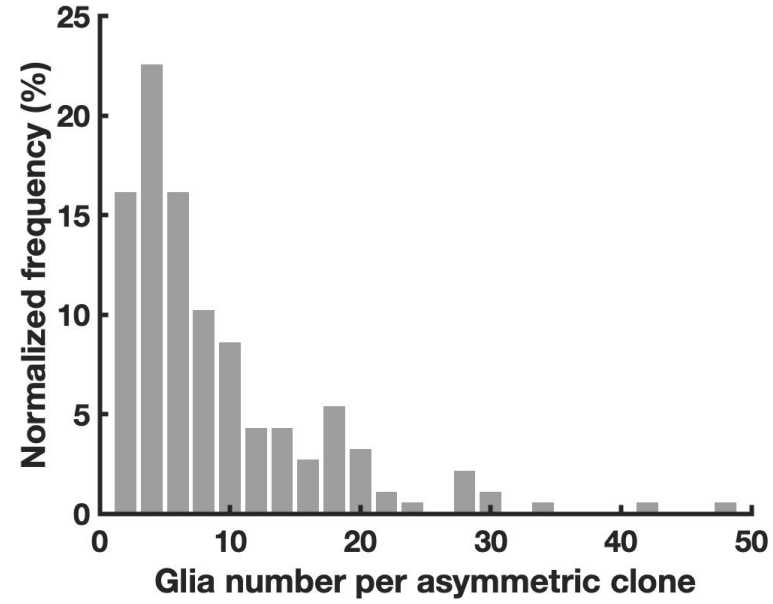
5.3.2 Symmetric G clones

The sizes and subclone ratios in symmetric WT G clones also differ from those for neurogenesis. Fig. 5.4b shows all symmetric WT G clones having ≥ 4 glia in both red and green sublineages. First, we notice that clone sizes do not decrease progressively going from earliest (E15.5) to latest (E17.5) MADM labeling time, suggesting that not all clones labeled at E16.5 and E17.5 were descendants from those present in the population at E15.5. This violates the notion that NGS clones are tracked recursively with MADM, in contrast with the clones observed during neurogenesis in [16]. That is, a migration of RGPs into the cortex from a different area of the developing brain may occur after E15.5, and these previously unobservable RGPs receive the MADM labeling at E16.5 and E17.5. This mechanism of migrating RGPs was previously proposed in Sec. 4.3.1.

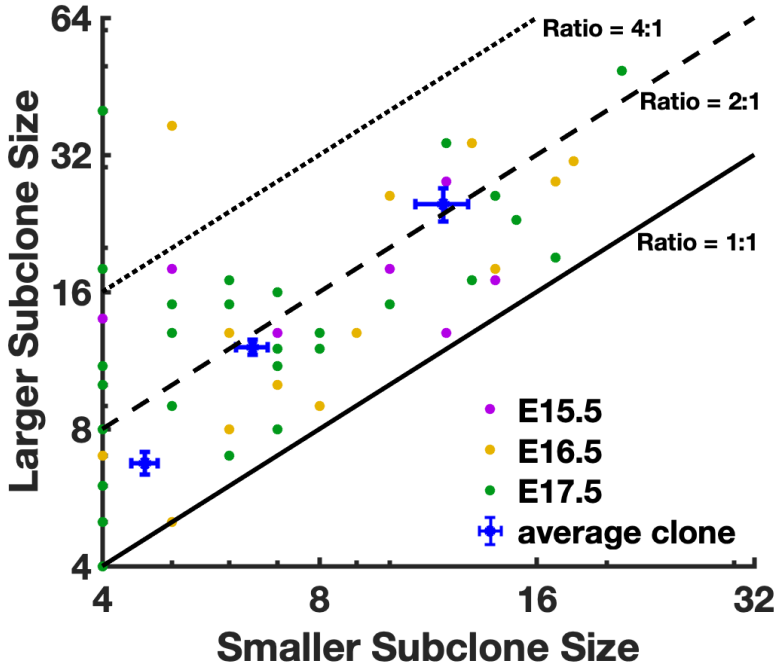
The ratios of glia in the larger to smaller subclone are also more variable than in neurogenesis. Ratios are less concentrated between 1:1 and 2:1, with 30.65% of symmetric clones having a greater than 2:1 subclone ratio. To examine the average larger to smaller subclone ratios by clone size as in Fig. 5.3b, we note that in neurogenesis it was possible to use time point as a proxy for clone size, since the total neurons per clone decreased predictably over time. Since the symmetric WT G clones do not decrease similarly over time, we binned them into three equally sized groups by their total glia (1-14, 15-22, and 23-70 glia, respectively) and calculated average smaller and larger subclones in each group. The average for each group is shown in Fig. 5.4b in blue, with error bars representing the standard error of the mean. The ratio of larger to smaller subclone in these averages does not appear to be conserved as clone sizes increase as was seen in neurogenesis in Fig. 5.3b. Instead, larger clones are more imbalanced between their larger and smaller subclones. Unlike in neurogenesis, this does not give evidence that the differentiation fates between glial subclones are connected, and it does not reflect a self-similar fractal pattern of symmetries.

5.3.3 Mix clones

Fig. 5.5 shows the distribution of glia per clone in Mix WT clones across all three MADM time points. Only 73 out of the 393 WT clones are in the Mix group, so the overall shape of their distribution is not immediately clear. However, it appears that the distribution may have peaks at different clone sizes, perhaps occurring around 10, 18, and 28 glia per clone. It is therefore possible that this distribution could be described by a Gaussian mixture model with linearly spaced peaks, as previously shown for neurogenesis in Fig. 5.3c. If so, we would gain support for the hypothesis of deterministic glial production in Mix clones. On the other hand, if the peaks are merely artifacts of having a relatively small, randomly sampled dataset, we would not expect any sort of pattern to the location of the peaks. Thus, in Sec. 5.6.2, we will test whether the peaks in this distribution appear at linearly spaced locations or not.



(a)



(b)

Figure 5.4 Total glia per clone in subsets of NGS MADM data. (a) Glia per clone in asymmetric clones, having ≤ 1 red (or green) glia and ≥ 2 green (or red) glia. (b) Glia per clone in symmetric clones with ≥ 4 glia in both the red and green lineages, delineated by time. Average smaller and larger subclones among clones divided into three groups by size (1-14, 15-22, and 23-70 glia, respectively) are shown ± SEM error bars.

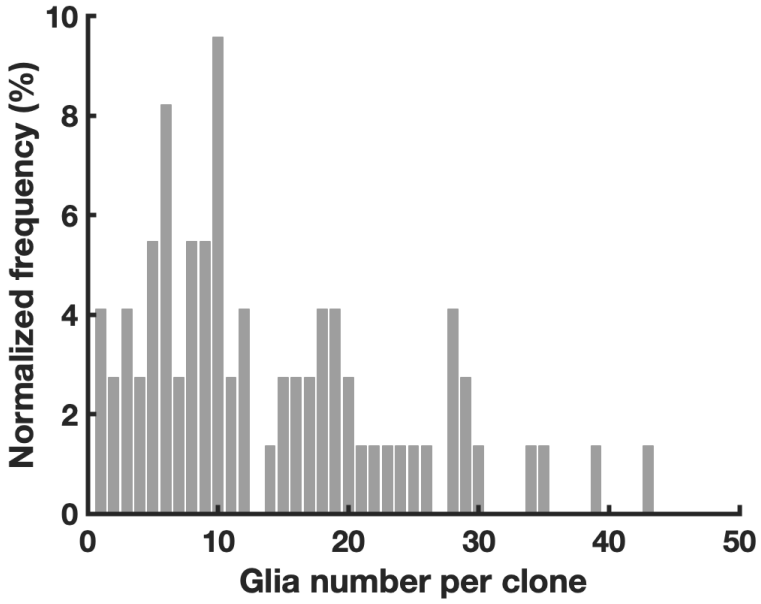


Figure 5.5 Normalized frequency distribution of glia per clone in Mix WT clones with a maximum of 50 glia. The number of Mix WT clones in the data, in all combined time points, is 73.

5.3.4 Full glial population, G and Mix

In Secs. 5.3.1 and 5.3.2, we observed that G clones do not show the same deterministic signatures in asymmetric and symmetric glial production as were observed for the production of neurons from RGPs during neurogenesis in [16]. This is consistent with our hypothesis thus far that G clones behave according to stochastic rules rather than deterministic. In Sec. 5.3.3, examining the distribution of glia per clone in Mix clones indicated that a linearly spaced multi-Gaussian distribution may describe how Mix clones produce glia, which would be evidence of deterministic rules. Again, this is consistent with our hypothesis thus far.

Interestingly, the normalized frequency distribution of total glia per WT clone (G and Mix combined) up to size 50 appears to have several peaks at various clone sizes (Fig. 5.6a). This is also the case for CKO clones in Fig. 5.6b. Previously, in Sec. 3.2.4, we demonstrated that the distribution of glia G clones was consistent with the distribution of glia that would be produced via a stochastic branching process. It is therefore possible that the peaks appearing in the combined distribution appear solely from the contribution of Mix clones. For instance, if Mix clones produce a linearly spaced multi-Gaussian distribution of glia and G clones do not, and we combine the two sets of clones, the G clones may not completely obscure the peaks from the Mix clones. They may, however, change how the peaks appear in the combined distribution. For instance, the G clone distribution may add enough 'noise' to the Mix distribution so that the peaks in the combined distribution

cannot be determined to be linearly spaced. Thus, we will test whether or not the peaks in the combined distribution of G and Mix clones shown in Fig. 5.6a can be represented with a multi-Gaussian model with linearly spaced peaks. To conclude, we will discuss possible deterministic or stochastic clonal behaviors by comparing how well the linearly spaced multi-Gaussian model represented the distribution of G and Mix clones, and how well it represented the distribution of Mix clones only.

We note that gathering clonal data is a costly and time-intensive process since each mouse must be sacrificed to observe the clones, and each clone's cells must be counted manually. In our analysis of clone size distributions in this chapter, we operate on the assumption that the $N = 73$ WT Mix, $N = 359$ WT Mix+G clones, and $N = 134$ CKO Mix+G clones are large enough samples to be able to estimate the locations of the peaks in their clone size distributions. By comparison, the analysis of clone size distributions in [16] used a sample size of $N = 192$.

5.4 Multi-Gaussian models for clone size distribution in gliogenesis

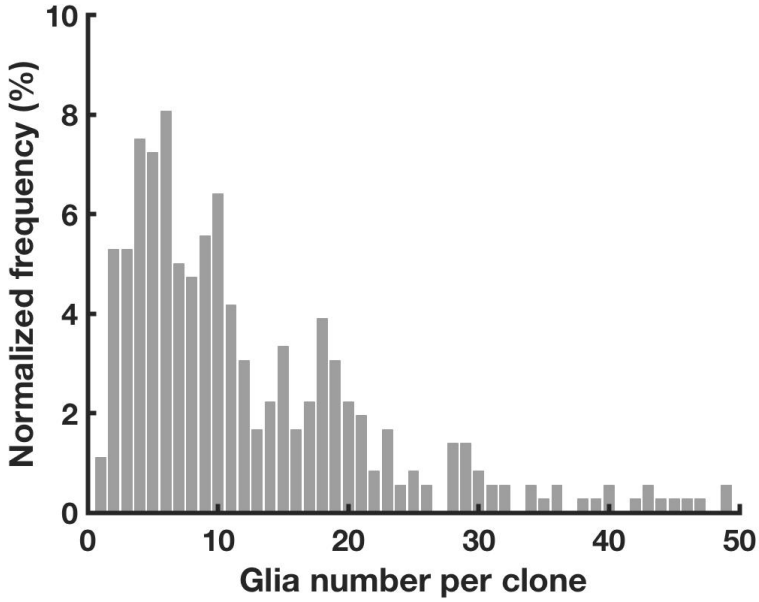
The distribution of neurons per clone for sizes $i = 1, ..., 50$ in [16] was represented with a Gaussian mixture model of the form

$$M_{CON}(i|\vec{p}) = \sum_{j=1}^{k} \beta_j^2 N(i|j\mu_0, \sigma_j^2). \tag{5.1}$$

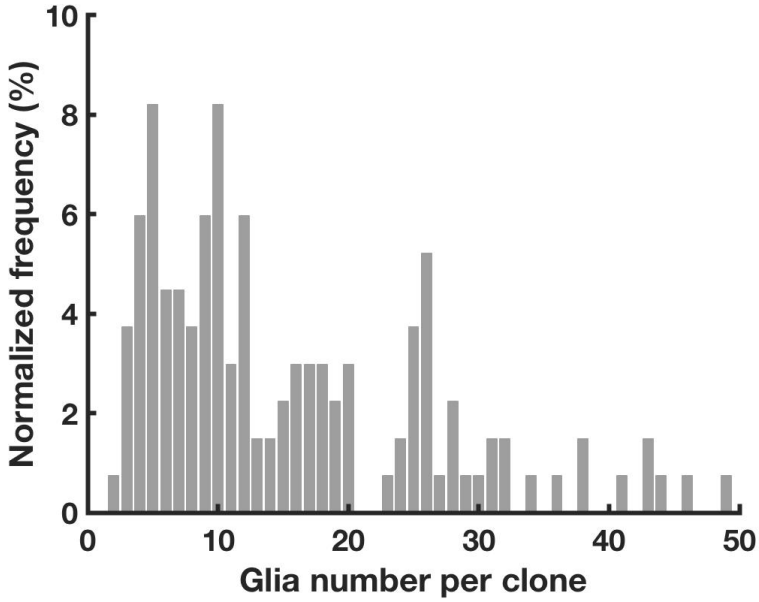
Here, k is the number of Gaussians in the model, $\vec{p}$ is the vector of parameters $[\beta_j^2, \mu_0, \sigma_j^2]$, where for each $j = 1:k$, $\beta_j^2 < 1$ is a positive weight, and $N(i|j\mu_0, \sigma_j^2)$ is a normal distribution evaluated at integers i given mean $j\mu_0$ and variance σ_j^2. The form of model (5.1) was considered specifically to represent the proposed mechanism of neurogenesis described in Sec. 5.2, with $j\mu_0$ constraining the means of the Gaussian peaks to be at integer multiples of a deterministic mean μ_0. A set of parameters $\vec{p}^*$ was identified in [16] using least squares minimization and $k = 5$ Gaussians in the model. We note however that none of the parameter values in $\vec{p}^*$ were reported, and no other models were tested to compare their fit to the distribution of neurons per clone in [16]. Rather, the observation that model (5.1) evaluated at $\vec{p}^*$ (Fig. 5.3b, black dashed line) produced a qualitatively reasonable fit to the distribution of neurons per clone was taken as evidence supporting their proposed hypothesis of deterministic clonal divisions during neurogenesis.

In our study, we wish to specifically use Gaussian mixture models to test whether the locations of the peaks in the distribution of glia per clone are evenly spaced or not. Thus, we compare two potential model distributions: (5.1) and

$$M_{UNCON}(i|\vec{p}) = \sum_{j=1}^{k} \beta_j^2 N(i|\mu_j, \sigma_j^2), \tag{5.2}$$



(a)



(b)

Figure 5.6 Distribution of total glia per clone in NGS MADM data for the populations of (a) WT Mix+G clones (b) CKO Mix+G clones. These two groups consist of $N = 359$ and $N = 134$ clones, respectively.

which sets no constraint on the locations of the Gaussian means. In the next section, we describe the methods used to compare models (5.1) and (5.2).

5.5 Parameter estimation and model evaluation criteria

For the WT and CKO datasets separately, we found the density histogram $H(i)$ of glial clone sizes. As with the previous neurogenesis study [16], we considered clone sizes $i = 1, ..., 50$ when calculating $H(i)$; only 2.18% of WT clones (8 clones in total) contained more than 50 glia, and these clones were sparsely distributed, having between 52 and 164 glia. The distributions $H(i)$ are shown in Fig. 5.6a for the WT data and Fig. 5.6b for the CKO data.

To fit models (5.1) and (5.2) to the distributions of glia per clone, we again used `fmincon`. The weights β_j^2 were constrained to be between 0 and 1 to ensure positive weightings of the Gaussians. Variances σ_j^2 were constrained to be between 0 and 6 to limit the overlap between Gaussians and better distinguish individual peaks, and means μ_j were free to fall anywhere between 0 and 50. We sought to find parameter sets $\vec{p}_{CON}$ and $\vec{p}_{UNCON}$ that minimized the sum of squares error between $H(i)$ and the respective model:

$$SSE_{CON} = \min_{\vec{p}_{CON}} \sum_{i=1}^{50} (M_{CON}(i|\vec{p}_{CON}) - H(i))^2 \tag{5.3}$$

$$SSE_{UNCON} = \min_{\vec{p}_{UNCON}} \sum_{i=1}^{50} (M_{UNCON}(i|\vec{p}_{UNCON}) - H(i))^2. \tag{5.4}$$

The parameters σ_j^2 and β_j^2 were initialized at the values 4 and 0.2, respectively, for all j. The value of μ_0 for the constrained model (5.1) was initialized at 5 by visual inspection of the location of the first peak in Fig. 5.3a. For the unconstrained model (5.2), the initial values of μ_j were set as the clone size with the greatest frequency in bins 1 through 6 for $j = 1$, bins 7 through 12 for $j = 2$, and so on. That is, for the WT distribution shown in Fig. 5.3a, these initial values were μ_1=6 and μ_2=10.

5.5.1 F-test for model comparison

For identical k values, (5.1) and (5.2) are *nested* , with M_{CON} using a proper subset of the parameters used for M_{UNCON} [3]. That is, M_{CON} is a special case of M_{UNCON} in which the μ_j parameters are linearly spaced. M_{CON} is referred to as the *reduced* model as compared with M_{UNCON}, referred to as the *complete* or *full* model.

Nested models can be compared using an F-test [3], which evaluates whether adding extra parameters improves the model's representation of the data significantly enough to outweigh the cost of the model's increased complexity. The null and alternate hypotheses in the F-test are

1. H_0: The complete model does not produce a better fit to the data than the reduced model.

2. H_1: The complete model produces a significantly better fit to the data than the reduced model.

[3]. To evaluate the null hypothesis, the F-test calculates an F-statistic from the fits of two nested models to the data distribution. For (5.1) and (5.2), this is

$$F = \frac{(SSE_{CON} - SSE_{UNCON})/(p_{UNCON} - p_{CON})}{SSE_{UNCON}/(n - p_{UNCON})} \tag{5.5}$$

where p_{UNCON} and p_{CON} are the total number of parameters in the respective models, n is the number of evaluation points of the model ($n = 50$, from evaluating the model at clone sizes 1 though 50) and SSE_{CON} and SSE_{UNCON} are the sum of squared errors calculated according to (5.3) and (5.4) [3]. The value of the F-statistic is used to calculate a p-value. If the p-value is below a significance level α, commonly $\alpha = 0.05$, then the null hypothesis is rejected, and the complete model is considered to be a better representation of the data. That is, if p<0.05, the model with more parameters is better than the model with fewer.

Models (5.1) and (5.2) represent two hypotheses themselves: that the peaks in the distribution are evenly spaced or that they are not. Thus, failing to reject the null hypothesis in the F-test would indicate that a model with evenly spaced peaks sufficiently represents the data distribution. On the other hand, rejection of the null hypothesis would suggest that the peaks in the distribution of glia per clone are unevenly spaced and that the clonal mechanism producing those peaks in clone sizes is different from that observed during neurogenesis.

5.5.2 F-test for selection of number of Gaussians k

To directly compare the performance of models (5.1) and (5.2) in their fit to a distribution of clonal data, it is reasonable to set the same number of Gaussians k for each model. Here, we discuss the selection of k using an F-test. For either individual model, increasing k creates a sequence of nested models; for instance, model (5.2) with K Gaussians can be transformed into a model with $K-1$ Gaussians if one of the Gaussian coefficients β_j is set to zero. Thus, we may use the F-test to determine whether a complete model with $k = K$ Gaussians is a better representation of the data than a reduced model with $k = K-1$. The simplest complete model (smallest value of k) that produces an improved fit over the reduced model will be selected to avoid overfitting to the distribution with too many Gaussian peaks.

We opt to use the fits of model M_{UNCON} (5.2) to select k, then set the same value of k for M_{CON}. The values tested were $k = 3, 4, 5, 6, 7$; these values are selected from visual inspection of the number of possible peaks in the glial distribution in Fig. 5.6a and for consistency with the $k = 5$ Gaussians used in the neurogenesis distribution fit from Fig. 5.3b. An F-test (see Sec. 5.5.1) is performed on

models with successive k values, and we select the lowest value of k for which the null hypothesis cannot be rejected.

5.5.3 Subsampling method

In the previous section, we detailed a method for evaluating the fits of two nested models to a set of data. However, the data we used to fit these models is the distribution of a random sample of clone sizes. This raises the question of whether the result of the F-test comparing models (5.1) and (5.2) would be the same if we had a different random sample of clones. A more robust comparison method would be to evaluate the fit of both models to multiple samples of clones. To achieve multiple comparisons of models (5.1) and (5.2), we implemented a subsampling scheme on the set of WT Mix+G clones and the set of CKO Mix+G clones. Subsampling was not performed for the WT Mix dataset due to the relatively smaller number of clones (73) in this group.

To implement the subsampling scheme, the clones in the set were split into ten equal groups. The groups were then cycled through in ten folds, leaving one group out of the data at a time, and a density histogram was formulated for the remaining 90% of clones in each case. Models (5.1) and (5.2) were fit to each subsample's histogram, and SSEs (5.3) and (5.4) were found, giving a set of ten errors for each model after completing all sample folds. These two error sets were compared against each other with ANOVA [37] to test whether either model produced a significantly lower mean error.

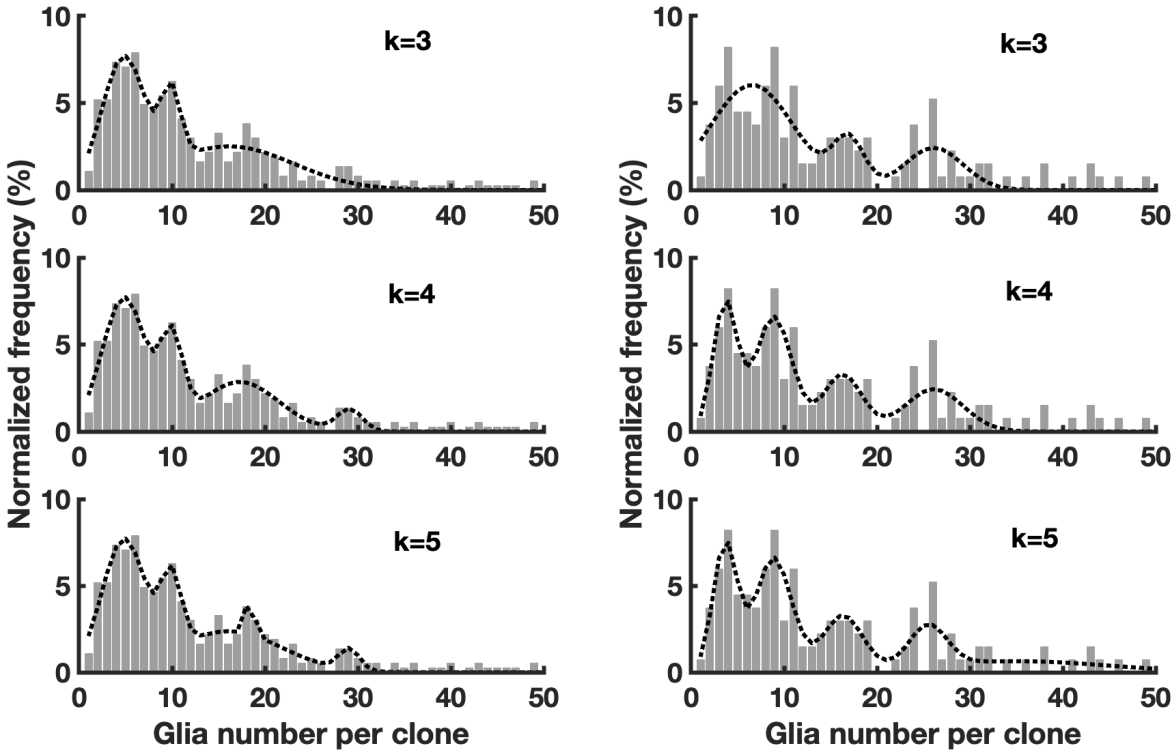
5.6 Results

5.6.1 k selection

Here, we review the results from selecting the number of Gaussians k in the mixture models (5.1) and (5.2) using the F-test (Sec. 5.5.1). The p-values for each test performed are listed in Table 5.1. For the WT Mix data, p>0.05 when comparing the model M_{UNCON} with $k=3$ and $k=4$ Gaussians, indicating that $k=3$ provides a sufficient fit over $k=4$. Since we aimed to select the smallest value of k, the remaining successive comparisons are not relevant. Thus, when we fit the models M_{CON} and M_{UNCON} to the WT Mix clones in Sec. 5.6.2, we select $k=3$ for the number of Gaussians in models.

When testing the fit of M_{UNCON} using successive k values to the WT Mix+G data, the p-value is less than 0.05 for $k=3$ to $k=4$. We therefore reject the null hypothesis of this F-test and conclude that the model's representation of the data is improved by using $k=4$ Gaussians instead of $k=3$. We see from Fig. 5.7a that using k=4 in the model discerns more of the peaks in the data as compared with k=3. The p-value going from $k=4$ to $k=5$ is greater than 0.05, causing us to fail to reject the null hypothesis. Thus, increasing k to 5 does not add useful information in the model, and $k=4$ Gaussians is sufficient. In Fig. 5.7a, this may be apparent in the fit for $k=5$ 'overfitting' the

peaks in the data as compared with the fit for $k = 4$. The F-test on the CKO Mix+G data showed a similar result, where the first p-value greater than 0.05 occurred when changing from $k = 4$ to $k = 5$. (p=0.1074). We therefore select $k = 4$ Gaussians when fitting the models M_{CON} and M_{UNCON} to the Mix+G clones for WT or CKO in Sec. 5.6.3.



(a) **(b)**

Figure 5.7 Fits of model (5.2) using k=3, 4, and 5 to the distribution of glia per clone in (a) WT (b) CKO.

5.6.2 Evaluation of Multi-Gaussian models, Mix clones

Fig. 5.8 shows the fit of models (5.1) (M_{CON}, constrained means) and (5.2) (M_{UNCON}, unconstrained means) using $k = 3$ Gaussians to the distribution of glia per clone in WT Mix clones. The two models show remarkable agreement in their fits to the data, as the blue and red curves overlap almost perfectly in Fig. 5.8. Additionally, the parameter values for the Gaussian means are similar. The p-value for the F-test comparing the two models is 0.9302. Since the p-value is greater than $\alpha = 0.05$, we cannot reject the null hypothesis of the F-test, and we conclude that the reduced model M_{CON}

Table 5.1 p-values for F test comparing the performance of the unconstrained Gaussian mixture model M_{UNCON} (5.2) using successive numbers of Gaussians k. The p-values below $\alpha = 0.05$ are denoted with an asterisk. The lowest value of k for which the p-value is above 0.05 was selected, thus $k = 3$ was chosen for number of Gaussians in (5.2) and (5.1) when fitting these models to WT Mix clones, and $k = 4$ was chosen for the model fits to WT Mix+G and CKO Mix+G clones.

Comparison	WT Mix clones	WT Mix+G clones	CKO Mix+G clones
k=3 to k=4 Gaussians	0.5034	0.0353*	0.0012*
k=4 to k=5 Gaussians	0.04270*	0.1096	0.1074
k=5 to k=6 Gaussians	1.000	0.2337	0.2211
k=6 to k=7 Gaussians	0.5229	0.5229	0.5345

Table 5.2 p-values for F-test comparing fits of constrained and unconstrained Gaussian models to WT Mix+G clones and CKO Mix+G clones. The fits are shown in Fig. 5.9. In both cases, p<0.05, thus the complete model M_{UNCON} with unconstrained Gaussian means better represents the distributions.

Test	p value
F-test, WT Mix+G clones, M_{CON} vs M_{UNCON}	0.0339*
F-test, CKO Mix+G clones, M_{CON} vs M_{UNCON}	0.0012*

represents the distribution better than the complete model M_{UNCON}.

5.6.3 Evaluation of Multi-Gaussian models, Mix+G clones

5.6.3.1 F-test

The fits of models (5.1) (M_{CON}, constrained means) and (5.2) (M_{UNCON}, unconstrained means) using $k = 4$ Gaussians are shown in Fig. 5.9a for the WT Mix+G data and Fig.5.9b for the CKO Mix+G data. Visually, both models performed similarly when fitting the portion of the distribution of glia per clone for clones smaller than ≈12 glia. However, the two models deviated from one another for larger glial sizes. Because of this close agreement, it is important to consider whether the unconstrained model produces a significantly better fit than the constrained model given the increase in the number of parameters. Table 5.2 gives the p-values from the F-test of these two models and their fit to the WT and CKO data sets. For both data sets, the p-value is significant at the $\alpha = 0.05$ level, indicating that the complete model M_{UNCON} with unconstrained means represents the distribution better than M_{CON}. This contrasts with the result for WT Mix clones in Sec. 5.6.2.

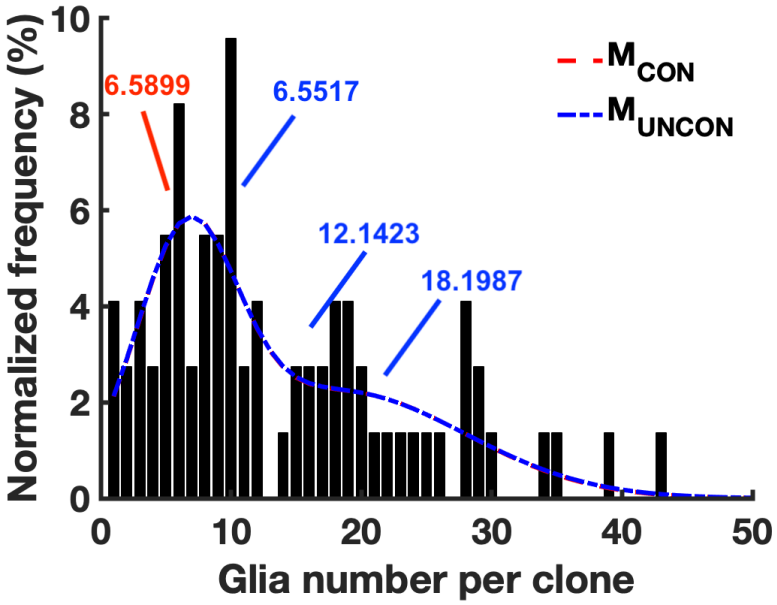
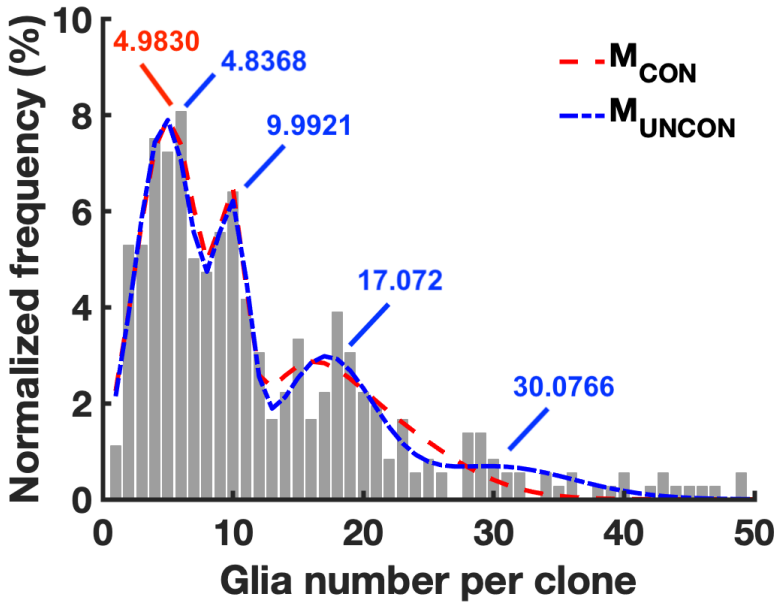
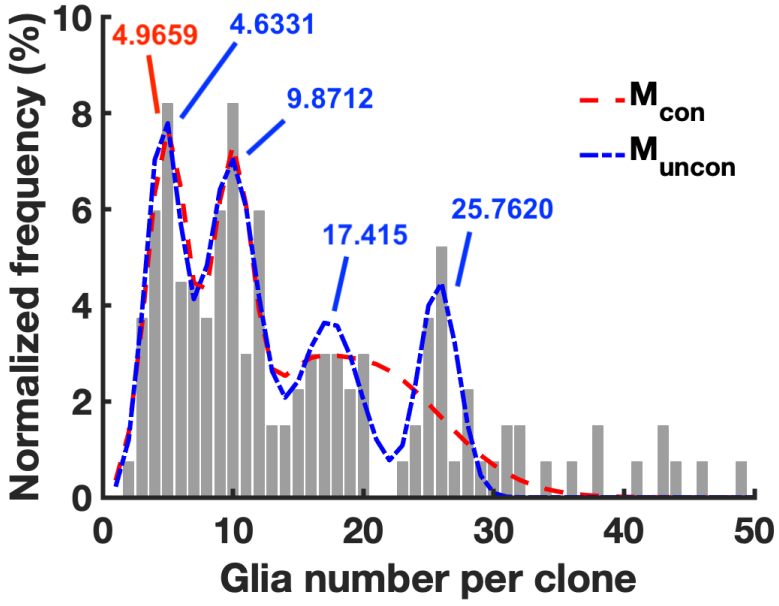


Figure 5.8 Fit of models (5.1) (M_{CON}, constrained means) and (5.2) (M_{UNCON}, unconstrained means) using $k = 3$ Gaussians to the distribution of glia per clone in WT Mix clones. The two models produce a nearly identical fit to the data and have similar parameter values for the Gaussian means, shown. The F-test p-value from comparing the fits of these two distributions was p=0.9302, indicating that the simpler model M_{CON} is sufficient to represent the data.



(a)



(b)

Figure 5.9 Fits of models (5.1) and (5.2) to the distribution of (a) WT Mix+G clones (b) CKO Mix+G clones. Locations of the four Gaussian means for model (5.2) and the initial mean μ_0 for model (5.1) are shown.

5.6.3.2 Sampling error ANOVA

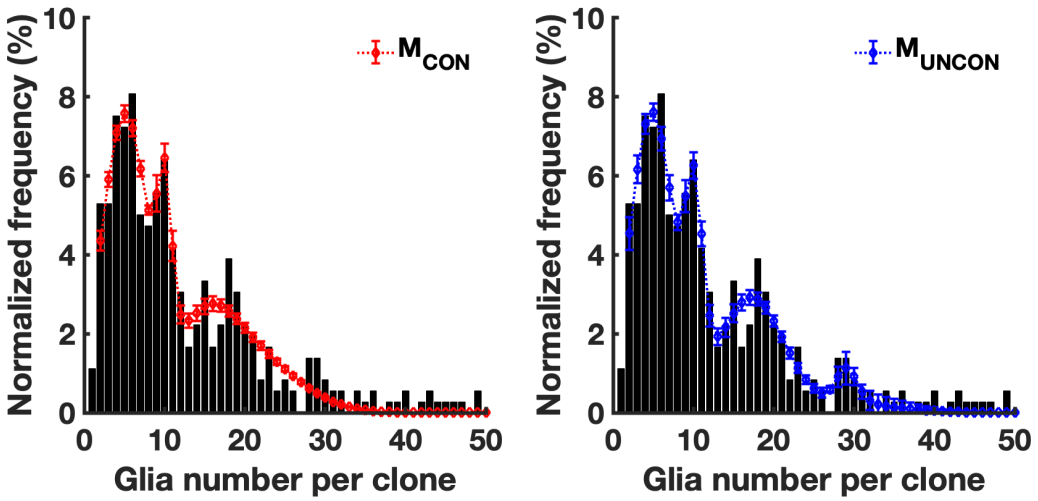
The mean and standard deviation of the ten subsampling fits of the constrained mean model (5.1) and unconstrained mean model (5.2), as described in Sec. 5.5.3, are shown in Fig. 5.10a for WT data and Fig. 5.10b for CKO data.

The average of the fitting errors for each model, calculated over each sampled fit from Eqns. 5.3-5.4, is shown in Fig. 5.11a-b (denoted J in the figure), with error bars denoting the standard error of the mean. Performing ANOVA on the sets of errors indicated a significant difference between models (5.1) and (5.2) at the $\alpha = 0.05$ level for the WT data (p=0.0044), but not for the CKO data (p=0.6279). That is, for the WT data, the set errors of the fit of M_{UNCON} were significantly lower than those for M_{CON}, and we conclude that the unconstrained model performed better than the constrained model. This is consistent with the result of the F-test for fitting the models once to the full set of WT Mix+G clones in Sec. 5.6.3.1. Additionally, performing an F-test for each of the ten fits resulted in a p-value greater than 0.05 in only two cases, and the p-values were relatively low in these cases (p=0.0774 and p=0.1493). Thus, in eight out of ten cases, M_{UNCON} produced a statistically significantly better fit to the subsample distribution than M_{CON}, suggesting that this result is robust over multiple samplings of clones.

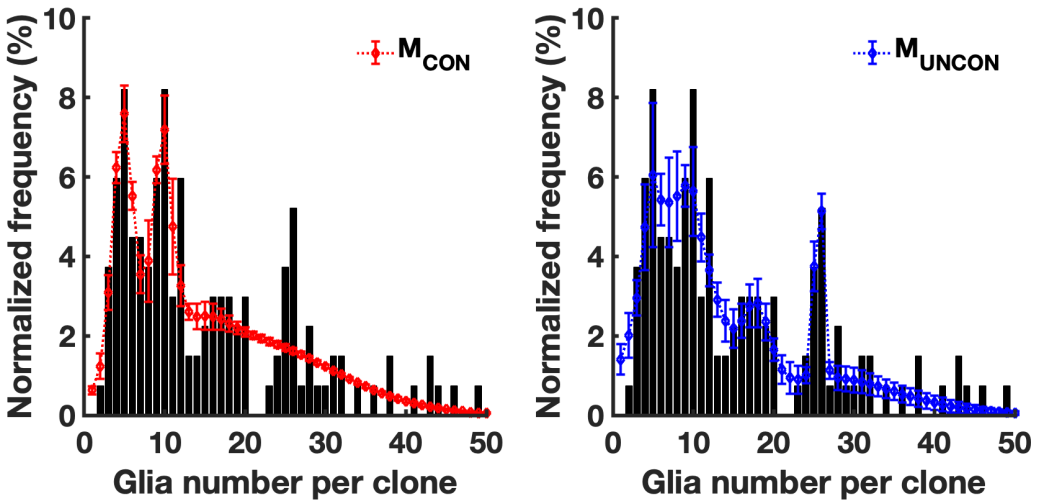
For CKO data, the sets of fitting errors were similar, thus M_{UNCON} did not perform better than M_{CON} over the ten subsampled fits. Interestingly, this contrasts with the F-test result for CKO Mix+G data in Sec. 5.6.3.1. However, we must note here that the locations of the unconstrained means were not consistent across the ten subsample fits to the CKO data. Table 5.3 shows the average parameter values for μ_1-μ_4 over the ten fits, but we observed that over different samples of the CKO data, one of two subsets of parameter values tended to occur. We show these two subsets in Table 5.3. When we performed an F-test for the model fits to each individual sample, we noticed that the p-value was low when the parameters from the first column were used, whereas it was high when the parameters from the second column were used (average p-values of 0.0349 and 0.8111, respectively). The F-test result is thus not robust over multiple samplings of clones, which is likely a result of the smaller number of CKO Mix+G clones (134) in comparison with WT Mix+G (359). We are therefore cautious about drawing conclusions from this subsampling scheme for the CKO clones, and it may still be the case that M_{UNCON} represents the distribution of CKO clones better than M_{CON}.

5.7 Discussion of clonal division rules

We now discuss what clonal division rules may be suggested by the results of the multi-Gaussian model fits to the WT Mix and WT Mix+G datasets. Our focus here is WT clones, since the conclusions from analyzing CKO clones in this chapter were not clear. Discussion of CKO clones is left for Sec. 6.6.



(a)



(b)

Figure 5.10 Average and standard deviation of ten fits of models (5.1) (red) and (5.2) (blue) to distribution of glia per clone formed from sampling 90% of clones in (a) WT Mix+G clones (b) CKO Mix+G clones. The average locations of the Gaussian means in each case are listed in Table 5.3.

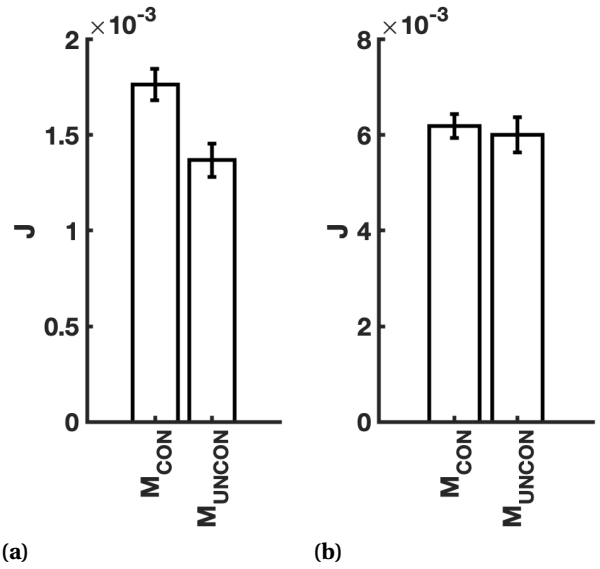


Figure 5.11 Average sum of squared error over the ten fits of models (5.1) and (5.2) to (a) WT (b) CKO. Error bars indicate standard error of the mean. For WT, model (5.2) had significantly lower fitting error (p=0.0044), but for CKO, the errors were not significantly different (p=0.6279).

Table 5.3 Mean values of Gaussian means μ_1-μ_4 over ten subsampling fits of models (5.1) and (5.2) to WT Mix+G clones and CKO Mix+G clones, as shown in Fig. 5.10. The parameter values for μ_1-μ_4 were generally consistent over the ten subsample fits with the exception of the CKO, M_{UNCON} case, which tended to produce one of two different subsets of μ_1-μ_4.

	WT, M_{CON}	WT, M_{UNCON}	CKO, M_{CON}	CKO, M_{UNCON}	CKO, M_{UNCON} subsets	
μ_1	5.0004	4.7726	4.8283	6.7870	4.7444	7.6624
μ_2	10.0008	9.99	9.6566	15.3533	9.5321	17.8482
μ_3	15.0012	15.17	14.4849	23.1357	17.4425	25.5756
μ_4	20.0016	18.29	19.3132	28.3192	25.5950	29.4868

First, we found a pattern of linearly spaced peaks in the distribution of WT Mix clones. This indicates that the deterministic rules of neural production in Gao et al. [16] may also be applied to glia production in Mix clones. Previously, we hypothesized that Mix clones originate from a subset of RGPs which we called NGS-RGPs. Our observations of glial production from Mix clones in Chapters 3 and 4 led us to also hypothesize that the NGS-RGPs produce glia deterministically. Thus, the multi-Gaussian analysis in this chapter has further supported the hypothesis of deterministic NGS-RGP behavior. Furthermore, since the multi-Gaussian distribution model was directly associated with a set of clonal division rules in [16], we can simulate clones that follow this set of rules. We define these rules and perform simulations of clones in Chapter 6.

Second, we found that the peaks in the distribution of WT Mix+G clones were not linearly spaced. Thus, the combined distribution of WT Mix+G clones cannot be labeled deterministic, suggesting that this combined population does not follow deterministic rules of cell division despite the deterministic patterns in the WT Mix clones. Additionally, our previous work in Sec. 3.2.4 indicated that G clones produce a distribution of glia consistent with a stochastic branching process, rather than a multi-Gaussian.

In summary, we have supported the hypothesis that two subpopulations of RGPs exist during the NGS: deterministically behaving NGS-RGPs, and stochastically behaving G-RGPs. The former group produces a set of glia whose distribution can be represented by a multi-Gaussian model with linearly spaced peaks M_{CON} (5.1), and the latter produces a glial distribution according to a stochastic Galton-Watson process, which we defined as Q_n in Eqn. 3.8 in Sec. 3.2.2. Thus, combining the sets of Mix and G clones into one distribution would combine the distributions M_{CON} and Q_n. It is possible that combining these two separate distributions obscures the ability to identify either one. That is, the distribution of glia in Mix+G clones may be a combination of these two distributions, but as we saw in Secs. 3.2.4 and 5.6.3, models Q_n and M_{CON} each failed to represent this clone size distribution. We test this idea in Chapter 6 by simulating two sets of gliogenic clones, deterministic NGS-RGPs and stochastic G-RGPs, and comparing their combined distribution of clone sizes to the models Q_n, M_{CON}, and M_{UNCON}.

CHAPTER

6

SIMULATION OF CLONAL DIVISIONS

6.1 Introduction

In the previous chapters, we have presented several methods of analyzing clonal data for deterministic and stochastic patterns. In Chapter 3, we used the theory of branching processes to formulate a probability distribution Q_n (see Eqn. 3.8 in Sec. 3.2.3), which represented the probability of a clone producing n glia from a series of stochastic division events. When comparing this distribution to the clonal NGS data, we demonstrated that this distribution sufficiently represented the glial output from G clones, but that it failed to represent the glial output from G and Mix clones pooled together. This suggested that G and Mix clones are produced during the NGS under two different sets of rules, and that stochastic rules may drive the production of G clones.

In Chapter 4, we aimed to verify the difference in glial production between G and Mix clones, particularly focusing on how these two sets of clones behaved differently across the E15.5, E16.5, and E17.5 time points as well as across the WT and CKO genotypes. This was accomplished using unbiased clustering (SOMs in Sec. 4.1) and statistical comparisons between subsets of clones (Wilcoxon rank-sum test in Sec. 4.2). The results from this analysis further supported the hypothesis that G and Mix clones do not produce glia according to the same set of rules during the NGS. Furthermore, we noticed that the distribution of total glia per clone in Mix clones was unchanged in response to suppression of gliogenesis in the green MADM sublineage via EGFR knockout. As discussed in Sec. 4.3.2, this was a potential sign that Mix clones may produce glia in a deterministic

manner.

In Chapter 5, we tested whether the distribution of clone sizes in the NGS could be represented with a multi-Gaussian distribution with linearly spaced peaks. A clone size distribution with linearly spaced peaks was hypothesized to correspond to a deterministic mechanism of division in a study of neurogenesis [16]. We denoted this distribution M_{CON} (see Eqn. 5.1 in Sec. 5.4) and compared its ability to represent the distribution of glia per clone in the NGS data to a multi-Gaussian distribution without linearly spaced peaks, M_{UNCON} (Eqn. 5.2, Sec. 5.4). It was found that the combined distribution of glia per clone in G and Mix clones could not be represented by the deterministic distribution M_{CON}, with M_{UNCON} producing a significantly better fit to the data. However, the distribution of glia per clone in Mix clones was sufficiently described by M_{CON}, providing further support for the hypothesis of deterministic glial production from Mix clones.

We posited that two subpopulations of RGPs with separate behaviors were responsible for these separate distributions of glia per clone. First, there is the subpopulation which we denoted NGS-RGPs, which produce neurons and then glia over the course of the NGS and correspond to Mix clones in the data. Second, there is the subpopulation denoted G-RGPs, which only produce glia and no neurons. The G clones in the data are assumed to come mostly from G-RGPs, though we note that G clones can also originate from MADM labeling of NGS-RGPs after the switch from neurogenesis into gliogenesis has occurred. The distribution of glia per clone from Mix clones showed evidence of deterministic patterns, so we predict that NGS-RGPs produce glia according to a deterministic set of division rules. On the other hand, the distribution of glia per clone from G clones pointed to stochastic patterns, so we predict that G-RGPs follow a stochastic set of division rules when producing glia.

In this chapter, we aim to test these predictions by establishing deterministic and stochastic rules of cell division that can be used to simulate clones arising from NGS-RGPs and G-RGPs, respectively. We then use these rules to simulate sets of clones and test whether the resulting distributions of glia per simulated clone can be represented by the distribution corresponding to its intended rules, either M_{CON} for the deterministic NGS-RGP rules, or Q_n for the stochastic G-RGP rules. Additionally, we simulate a scenario in which a portion of the clones follows deterministic rules, and the other portion follows stochastic rules. This is intended to simulate our hypothesis in which both NGS-RGPs and G-RGPs are present during cortical development and the NGS. We test whether this distribution of clones simulated using a combination of deterministic and stochastic rules can be represented by M_{CON} or Q_n, or if the unconstrained multi-Gaussian model M_{UNCON} produces the best fit to the data instead. That is, we aim to test whether combining a set of clones that were simulated from two separate rules into a single distribution obscures the ability to identify either set of rules, as was observed when we attempted to fit M_{CON} and Q_n to the combined distribution of glia in G and Mix clones in Sec. 5.6.3.

Ultimately, the analysis in this chapter is intended to more robustly establish the validity of

defining deterministic or stochastic mechanisms given only the distribution of glia per clone. In previous chapters, we have searched the distributions of glia per clone for the signatures we assume would be present in a deterministic or stochastic mechanism of glial production. We have thus assumed that if different rules govern glial production at the level of clonal divisions, then this difference should be discernible when examining the distribution of glia per clone. That is, we have held that a difference in distribution implies a difference in clonal division rules. Here, we reverse this analysis by simulating clones according to different rules that are known to be deterministic or stochastic, then testing whether the resulting distributions of glia per clone match the model intended to represent each mechanism, either M_{CON} or Q_n. If we show that the correct models match the simulated mechanism, we have a stronger case that our analysis of deterministic and stochastic patterns in the distribution of glia per clone from the NGS data does in fact support the hypothesis that these patterns occur at the clonal level.

6.2 Branching process construction

6.2.1 NGS-RGP subpopulation

To simulate glial production from NGS-RGPs using a branching process, we must establish division rules that can produce a set of clones with a multi-Gaussian distribution of glia. Previously, in Sec. 5.2.3, we discussed the deterministic clonal rules that were hypothesized to produce a multi-Gaussian distribution of neurons per clone during neurogenesis in [16]. These rules defined a two-stage process: in the first stage, RGPs proliferate to produce more RGPs, and in the second stage, RGPs switch to asymmetric differentiating divisions to produce neurons. Thus, we can use a similar two-stage framework to define rules that NGS-RGPs follow. We do so by creating a branching process with three cell types: p, which represent proliferating NGS-RGPs, dp, which represent differentiating NGS-RGPs, and G, which represent differentiated glia. In the branching process, p and dp cells divide into two offspring at each discrete time step, and the cell types of the offspring are determined according to a probability distribution. G cells are terminal and do not divide. The rules for p and dp divisions are defined as follows.

Cell type p We base p divisions on the basic GW process, with the addition of a decay parameter. A p progenitor divides to produce a set of two offspring: $\{p,p\}$, $\{p,dp\}$, $\{dp,p\}$, or $\{dp,dp\}$. The probability of producing $\{p,p\}$ or $\{dp,dp\}$ at the first division are the parameters ρ_1 and ρ_2, respectively. This division occurs at the time $t = 1$. At each successive time step $t = 2,3,...$, these probabilities are updated according to a decay parameter d, $0 < d < 1$, such that the probability of a

p division producing a set of offspring at time t is

$$P(\{p,p\},t) = \rho_1 d^{t-1} \tag{6.1}$$

$$P(\{dp,dp\},t) = \rho_2 + \rho_1(1-d^{t-1}) \tag{6.2}$$

$$P(\{p,dp\}) = P(\{dp,p\}) = \frac{1-\rho_1-\rho_2}{2}. \tag{6.3}$$

Thus, the probability of producing more p progenitors decreases at each time step, and the probability of creating dp progenitors absorbs the difference. Note that this is a departure from traditional branching processes, where a cell type's division probabilities do not change over time. We opted to update the probabilities at each step using the decay parameter d so that the ρ_1 could begin at a large value and promote early proliferation, while avoiding the possibility of simulating clones that continue to proliferate and never terminate. For simplicity, the asymmetric probabilities $P(\{p,dp\}) = P(\{dp,p\})$ remain constant in time through the simulation. We discuss the calibration of parameters ρ_1, ρ_2, and d in Sec. 6.3.

Cell type dp. In the model, p cells produce dp cells, hence dp cells may originate at times $t = 2,3,...$ and begin dividing according to their respective rules from this time of onset. For easier indexing, we will refer to the divisions of dp cells as occurring at time steps relative to this onset, $t_2 = 1,2,....$ At each discrete time step t_2, a dp cell undergoes an asymmetric division to produce offspring $\{dp,G\}$ or $\{G,dp\}$, where G is a glial cell. Glia do not divide further, while the dp cell produced as offspring continues divide asymmetrically at subsequent time points until a final $\{G,G\}$ division. The probability of a final $\{G,G\}$ division occurring at time point t_2^* is modeled by a discrete version of a truncated normal distribution, defined to be

$$P(t_2 = t_2^*) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(t_2^*+1-\mu)^2}{2\sigma^2}\right)}. \tag{6.4}$$

This mode of differentiation was chosen based on the distribution of asymmetric neurogenic clone sizes in [16]. We do not actually know if differentiating NGS-RGPs produce glia via asymmetric divisions. To determine this, we would need to be able to compare the symmetries of red and green glia production from NGS-RGPs while they are undergoing differentiation. However, we can only say for certain that Mix clones arise from NGS-RGPs, which produce neurons prior to producing glia and thus are not undergoing differentiation into glia from the onset of MADM labeling. However, since our goal is to simply simulate clones that, when differentiating, produce a Gaussian distribution of clone sizes, the symmetry of the clones in the simulation will not matter for our analysis.

We discuss the selection of parameters μ and σ in Sec. 6.3.2. For the discrete normal distribution model in Eqn. 6.4, we note that $P(t_2 < 1) > 0$, whereas we can only simulate clones for values of $t_2 \geq 1$. To avoid sampling values of t_2 that are less than 1 in the simulation, we evaluate the probability distribution shown in Eqn. 6.4 at $t_2 = 1,...,20$ at the selected values of μ and σ, then normalize

the distribution. We note that a maximum of $t_2 = 20$ was chosen since $P(t_2 > 20)$ is very small for the parameters μ and σ that we determine. Thus, we only allow t_2 to be between 1 and 20 in our simulation.

Lastly, we define the probability parameter β. With probability $\beta > 0$, the simulation begins with a type p cell, and with probability $1-\beta$, the simulation begins with a type dp cell. Thus, the minimum number of differentiated cells possible in a simulated clone is two, which would arise from a clone beginning the simulation as a dp progenitor immediately undergoing a symmetric differentiating division at $t_2 = 1$. No cell death was considered in the model.

The simulation ends when no p or dp cells remain, all having differentiated into G cells. For computational efficiency, a maximum of $t = 20$ rounds of proliferative division was set, based on the cell division estimates in Sec. 3.1.2; in the case that any progenitors p remained after $t = 20$, these progenitors were terminated. We note that this was a very rare occurrence given our calibrated parameter values. The final output of the simulation is the total glia produced in each clone's two sublineages, with the sum being the total glia produced per clone. Fig. 6.1 shows an example of a simulated lineage, which would produce an output of 5 glia in the left sublineage and 10 in the right, for a total of 15 glia in the clone.

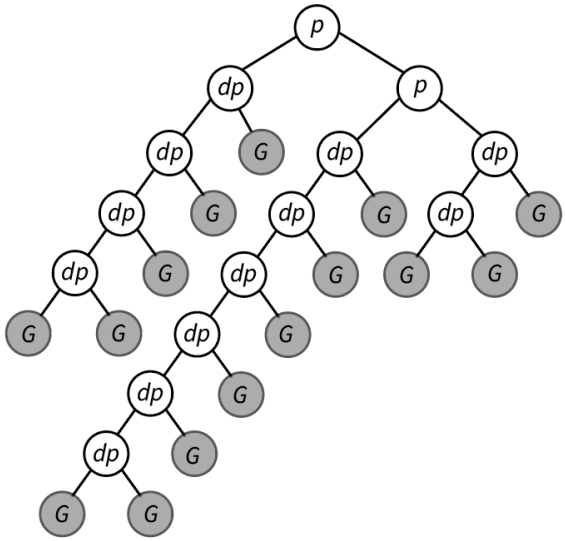


Figure 6.1 Possible clonal lineage generated by the model of glial production from NGS-RGPs, producing 15 glia. Here, the clone begins as a progenitor p, and when a dp progenitor is generated after a division, glia are laid down in successive asymmetric divisions.

6.2.2 G-RGP subpopulation

The G-RGP subpopulation divides stochastically and can thus be simulated using a basic GW branching process. Thus, we can simulate glial production in this population with a model identical to that for neurogenesis from [48], which we previously discussed in Sec. 3.2.2. Here, the simulation begins with a progenitor p, which divides into two offspring at discrete time steps according to probabilities fixed in time:

$$P(\{p,p\}) = \xi_1 \tag{6.5}$$

$$P(\{G,G\}) = \xi_2 \tag{6.6}$$

$$P(\{p,G\}) = P(\{G,p\}) = \frac{1-\xi_1-\xi_2}{2}. \tag{6.7}$$

The simulation ends when all p cells have differentiated into G cells.

6.3 Parameter Selection and Estimation

We simulate four scenarios using the above models for NGS-RGP and G-RGP behavior. First, since the NGS-RGP model was informed by a proposed deterministic model for neurogenesis, we simulate clones using this model to compare to the clonal MADM data from neurogenesis collected in the study by Gao et al. [16]. Second, we use the same model to simulate deterministic glial production in NGS-RGPs. For these two simulations, we estimate the model parameters using the clonal MADM data from neurogenesis and the NGS, respectively, using methods proposed in [16]. We describe these parameter estimation methods in detail in Sec. 6.3.1-6.3.2. Third, we use the G-RGP model to simulate stochastic glial production during the NGS. The parameters for this simulation are determined by fitting the distribution Q_n to the distribution of glia per clone as performed in Sec. 3.2.4, which we briefly review in Sec. 6.3.3. Lastly, we simulate a scenario in which half of the population of clones follows the NGS-RGP model, and the other half follows the G-RGP model. For this simulation, we use the Mix clones to estimate the parameters for the NGS-RGP group, and the G clones to estimate the parameters for the G-RGP group, which we explain in Sec. 6.3.4.

6.3.1 Scenario 1: Neurogenesis, NGS-RGP model

Simulating neurogenesis according to the NGS-RGP model requires the following parameters: β, the proportion of clones beginning the simulation as p cells, ρ_1 and ρ_2, the initial probabilities of a $\{p,p\}$ or $\{dp,dp\}$ division, respectively, and d, the decay rate of ρ_1 at each discrete time step. Additionally, the normal distribution parameters μ and σ are required for the asymmetric differentiation step.

As previously shown in Fig. 5.3, the parameters $\mu = 8.4$ and $\sigma = 2.6$ were already found in the

neurogenesis study [16]. The parameters β, ρ_1, ρ_2, and d require further calibration to simulate clones comparable to the data coming from neurogenesis. These parameters control the number of rounds of division a p cell performs before producing dp offspring. Thus, we wish to calibrate the parameters so that the simulated p cells undergo approximately the same number of proliferative divisions as were observed to occur during neurogenesis in [16].

Naturally, since clonal data is collected *in vivo* using MADM, their study was not able to directly measure the number of progenitor divisions before beginning asymmetric neurogenesis. However, the rounds of division were estimated using simplifying assumptions. First, it was argued from observing clone symmetries during neurogenesis that the progenitors arising from the red and green sublineages of an initial MADM-labeled progenitor begin asymmetric differentiation within one generation of division of one another. By this argument, progenitor expansion is mostly symmetric prior to differentiation, and thus the number of progenitors present doubles at each round of division. If each differentiating progenitor produces $\approx \mu = 8.4$ neurons, the rounds of proliferative division n required to produce a clone with m total neurons can therefore be approximated by the relation

$$n = \log_2\left(\frac{m}{\mu}\right). \tag{6.8}$$

In [16], the rounds of division n were calculated according to Eqn. 6.8 for each neural clone, then binned into a normalized frequency distribution with bin widths of 0.2. These binned frequencies calculated for the neurogenisis data from time points E10-E12 are shown as dots in Fig. 6.2, along with a smoothed curve.

We can use these generation estimates to calibrate the parameters in our neurogenesis simulation. We focus on attempting to simulate clones whose rounds of division follow the green curve generated for the data from E12, which we denote $F(n)$; these later clones predominantly produce ≤ 50 neurons, which we have emphasized previously in the multi-Gaussian model fits. $F(n)$ was digitally extracted for n values corresponding to the bin centers, $n = 0.1, 0.3, ..., 4.9$.

The extracted values of $F(n)$ were used to initialize parameters for β, ρ_1, ρ_2, and d. First, the function $F(n)$ for $n < 1$ represents the progenitors that were labeled with MADM while undergoing asymmetric divisions. In relation to the model, this is the portion of progenitors beginning as dp cells, 1-β. We thus find 1-β=$\sum_{n<1} F(n)$=0.3874 and set β =0.6126. The parameter ρ_2, denoting the probability of a p cell undergoing a $\{dp, dp\}$ division at the first generation, corresponds to the occurrence of $n = 1$ rounds of proliferative division. We find $\sum_{1<n<2} F(n)$=0.4366 and normalize this value according to the percent of clones starting the simulation as p cells, β, giving ρ_2 = 0.4366/0.6126=0.7127. To remain consistent with the previously stated assumption that progenitors favor symmetric divisions during proliferation, we set the probabilities of asymmetric $\{p, dp\}$ or $\{dp, p\}$ divisions at a small value, 0.1 each. Since the probabilities must sum to 1, this gave our initial value for the final parameter $\rho_1 = 0.7127 - 0.1 - 0.1$=0.0873. The value of the decay parameter

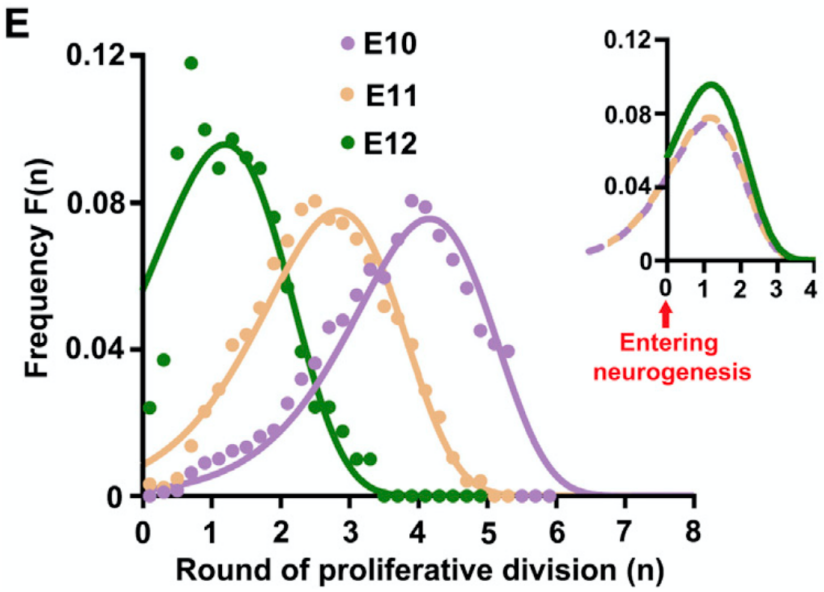


Figure 6.2 Estimation of proliferative division rounds n during neurogenesis from [16]. Dots represent the binned frequencies of division rounds estimated using Eqn. 6.8, with the input m being the total neurons per clone from the neurogenesis data in [16]. A smoothed curve was fit to the frequencies of n for each MADM time point, E10-E12. It was observed that the estimated rounds of proliferative division decreased when going from earliest (E10) to latest (E12) MADM time point. Reprinted with permission from [16]. Copyright 2014, The Authors. Published by Elsevier Inc.

d was assumed to be high so that rounds of proliferative division would not go beyond $n = 4$ as shown in the green curve from Fig. 6.2. We thus chose an initial value of $d = 0.7$.

Using these parameters, a set of 5000 clones was simulated, and the rounds of division n for each clone was found according to Eqn. 6.8. We found that 5000 clones was a sufficient sample size to stabilize the distribution of simulated clone sizes. The division rounds were binned as was done for $F(n)$ and smoothed, creating a binned distribution $\tilde{S}(n)$, $n = 0.1, 0.3, ..., 4.9$ (Eqn. 6.10). To tune the parameters and create a distribution of division rounds $\tilde{S}(n)$ that more closely represents the data $F(n)$, we implemented least squares minimization of the difference between $\tilde{S}(n)$ and $F(n)$ using the MATLAB function `fmincon`, with parameters initialized at their values above. It is important to note that we should not expect these parameters to be unique. For instance, similar distributions of $\tilde{S}(n)$ may be created by increasing the initial proliferation rate ρ_1 while also increasing the decay rate d. However, our goal is simply to find a set of parameters that can plausibly represent the observed neurogenesis data, which can be done using our informed initial parameter estimates.

6.3.2 Scenario 2: Deterministic gliogenesis, NGS-RGP model

As in the previous section, modeling gliogenesis using the deterministic NGS-RGP model requires parameters β, ρ_1, ρ_2, d, μ, and σ as shown in Eqns. 6.1-6.4. We began with setting the normal distribution parameters as the first Gaussian mean and standard deviation from fitting the constrained multi-Gaussian model to the distribution of glia per clone as shown in Fig. 5.9, giving μ=4.9020 and σ=2.4043. The remaining parameters β, ρ_1, ρ_2, and d, were then initialized using the same process described in Sec. 6.3.1. That is, the number of glia m in each G and Mix clone from the NGS data was used to estimate the rounds of proliferative division n by Eqn. 6.8. The values of n were then binned into a frequency distribution $F(n)$. The initial value for the parameter β was set as 1-$\sum_{n<1} F(n)$, and the probability ρ_2 was set as $\sum_{1<n<2} F(n)/\beta$. The asymmetric division probabilities were again initialized at a small value, 0.1, and the decay parameter d was initialized at 0.7. This set of initialized parameters was then tuned using the same method described in Sec. 6.3.1.

6.3.3 Scenario 3: Gliogenesis, G-RGP model

The population of glia coming from G-RGPs was simulated following the Galton-Watson model with division probabilities defined in Eqns. 6.5-6.7. A recursive expression for the probability distribution of a clone of size n following this model was previously discussed in Sec. 3.2.2, and is represented as Q_n in Eqn. 3.8. The same model can be used for G-RGPs by substituting in $P(\{p,p\}) = \xi_1$ and $P(\{G,G\}) = \xi_2$ for the parameters q and p, respectively. Thus, we have $Q_2 = \xi_2$, $Q_3 = \xi_2(1-\xi_2-\xi_1)$, and

$$Q_n = \xi_1 \sum_{k=2}^{n-2} Q_k Q_{n-k} + (1-\xi_2-\xi_1) Q_{n-1}, \quad n \geq 4. \tag{6.9}$$

We found the parameters ξ_1 and ξ_2 that minimize the sum of squares difference between Q_n and the distribution of glia in G and Mix clones using `fmincon` in MATLAB.

6.3.4 Scenario 4: Gliogenesis, combination NGS-RGP and G-RGP models

As previously mentioned, the simulation of a combined population of NGS-RGPs and G-RGPs is performed by simulating each subpopulation according to its specific rules. Thus, the parameters for each subpopulation's simulation are determined using the same methods put forth in Secs. 6.3.2-6.3.3, with the exception that the parameters are calibrated to the data coming from the separated groups of Mix or G clones, respectively. That is, in scenarios 2 and 3, we assumed that all simulated clones obey the same model of gliogenesis, either deterministic or stochastic respectively. The parameters for simulating these scenarios were thus calibrated using the set of glia per clone in all clones, Mix and G. Instead, in this final scenario, we use the Mix population of clones to calibrate the parameters of the deterministic model, following the procedure detailed in Sec. 6.3.2, and the G population of clones to calibrate the stochastic model parameters, using the procedure from Secs. 6.3.3.

6.4 Evaluation of Simulations

For scenarios 1-4 described in Sec. 6.3, we simulate populations of N total clones using the respective rules and calibrated parameters. We set $N = 359$, which is equal to the number of clones having $\leq$50 glia in the NGS WT data. In scenario 4, we simulate $N = 180$ clones from each subpopulation for a total of 360 clones in the combined population. From each simulated population, we form the normalized frequency histogram of glia per clone, $H(i)$, $i = 1, ..., 50$.

For the frequency histograms formed for scenarios 1-3, we test the goodness of fit of the model that is intended to represent the simulated rules. Thus, for $H(i)$ in scenarios 1 and 2, which model a deterministic mechanism, we test whether M_{CON} sufficiently represents the distribution as compared with M_{UNCON}. This is done with the methods used to evaluate these two models in Chapter 5, which we recall here: first, we use an F-test to compare the nested models as described in Sec. 5.5.1, and second, we run the simulations multiple times and compare the sets of fitting errors between $H(i)$ and each model over the multiple runs. This is similar to what was done in the subsampling scheme described in Sec. 5.5.3. However, instead of simulating one population of clones and fitting M_{CON} and M_{UNCON} to the distribution of multiple subsamples, we simply simulate a new population of $N = 359$ clones in each run. For the clones simulated with a stochastic mechanism in scenario 3, we use a Chi-Square test to evaluate whether Q_n from Eqn. 6.9 accurately represents the distribution of simulated clones $H(i)$.

Finally, for the clones simulated in scenario 4 with a combination of deterministic and stochastic

behaviors, we use each of the methods above to test how well the distribution of clone sizes is represented by each model: M_{CON}, M_{UNCON}, or Q_n.

6.5 Results

6.5.1 Neurogenesis Model

6.5.1.1 Parameters

We calculated the division rounds n for 5000 clones simulated according to the initialized parameters β, ρ_1, ρ_2, and d using Eqn. 6.8. The values of n were then binned into a normalized frequency histogram $S(n)$. Based on the procedure described in [16], a smoothed distribution $\tilde{S}(n)$ was created using a moving window of two points to the left and right of $S(n)$ so that

$$\tilde{S}(n) = \frac{1}{5} \sum_{k=n-0.4}^{n+0.4} S(k). \quad n = 0.1, 0.3, ...4.9 \tag{6.10}$$

Fig. 6.3a shows $\tilde{S}(n)$ (dots) computed using the initial parameters chosen in Sec. 6.3.1 in comparison with the curve $F(n)$ representing neurogenesis division rounds from [16]. Our initial parameter choices appear to get relatively close to representing $F(n)$, indicating a good performance from our method of initialization. Fig. 6.3b shows the distribution of $\tilde{S}(n)$ following parameter tuning, with estimated parameters β=0.4232, ρ_1=0.0977, ρ_2=0.5174, and d=0.7585. These parameters produce a better overall fit to $F(n)$.

6.5.1.2 Simulations

Using the tuned parameters, ten samples of 359 clones were generated and the frequency histogram of clone sizes $H(i)$, $i = 1, ..., 50$ was created for each sample. For each sample, the multi-Gaussian models (5.1) and (5.2) with constrained and unconstrained means, respectively, were fit to the distribution $H(i)$. For this neurogenesis simulation, $k = 5$ Gaussians were used in the models to match the model in [16]. An F-test was run to generate a p-value comparing the performance of the two models, and the sum of squares error J was also recorded in each case.

Fig. 6.4a shows a realization of $H(i)$ along with the average and standard deviation of the ten model fits M_{CON} and M_{UNCON}. Visually, both models perform similarly, and we found that out of the ten p-values generated from the F-test, only two were less than 0.05. This indicates that clones following the deterministic division rules described in [16] most likely produce a frequency distribution that can be sufficiently represented with a multi-Gaussian curve with linearly spaced means. Additionally, we see observe that the mean ±SEM of J for each model over the ten fits were not significantly different (ANOVA, p=0.4462). These results are expected, as they support the case

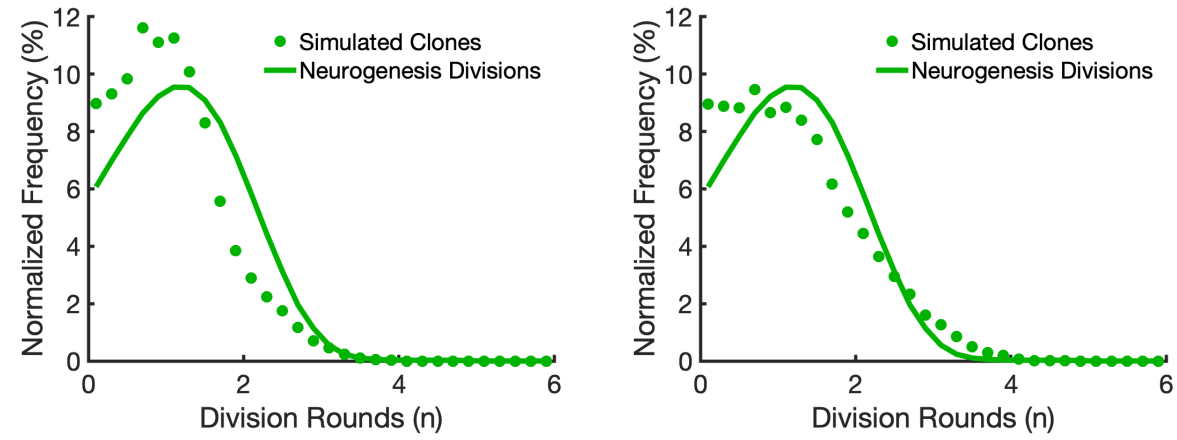


Figure 6.3 Calibration of the proliferation parameters for simulating neurogenesis. The dots represent $\tilde{S}(n)$, the distribution of division rounds n computed in Eqn. 6.10. The smooth curve was manually extracted from the green curve in Fig. 6.2 and represents the distribution of division rounds in neurogenesis estimated in [16]. In (a), we show $\tilde{S}(n)$ for the initialized parameter values given in Sec. 6.3.1. In (b), we show $\tilde{S}(n)$ for the tuned parameter set β=0.4232, ρ_1=0.0977, ρ_2=0.5174, and d=0.7585.

that deterministic rules at the clonal level can translate to identifiable patterns in the resulting distribution of clone sizes.

6.5.2 Deterministic gliogenesis model

6.5.2.1 Parameters

We selected parameters $\mu = 4.9830$ and $\sigma = 2.5212$ based on the first Gaussian mean and standard deviation found from fitting M_{CON} to the WT Mix+G data in Fig. 5.9. The proliferation parameters $\beta = 0.3717$, ρ_1=0.4093, ρ_2=0.4418, and d=0.4600 were then estimated using the procedure described in Sec. 6.3.1. Fig. 6.5 shows the division rounds $\tilde{S}(n)$ from 5000 clones simulated using the tuned parameters, compared with the distribution of $F(n)$ found for glial clones using Eqn. 6.8 with $\mu = 4.9830$.

6.5.2.2 Simulations

We show a realization of a clonal distribution $H(i)$ from $N = 359$ clones simulated using the neurogenesis model with gliogenesis parameters in Fig. 6.6a, along with the mean and standard deviation of the ten fits for each model. Clearly, using different parameters alters the shape of the distribution of clone sizes as compared with Fig. 6.4a. The fitting errors J in Fig. 6.6b are not significantly different (ANOVA, p=0.4104). The F-test comparing the fits of models M_{CON} and M_{UNCON} gave a p-value

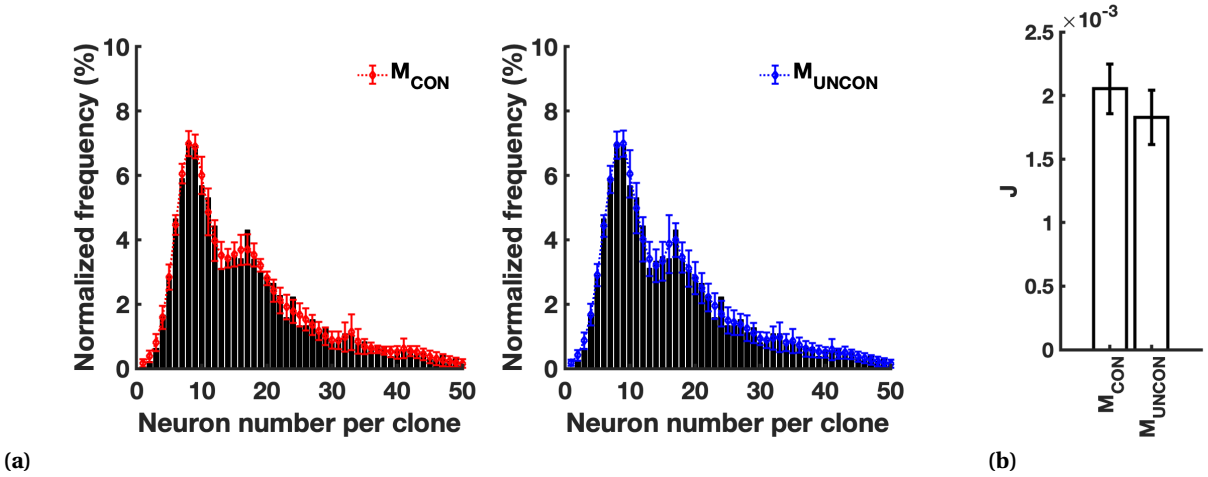


Figure 6.4 Results of neurogenesis simulation with deterministic NGS-RGP model. In (a), we show an example distribution of neurons per clone from a set of $N = 359$ clones simulated according to the NGS-RGP rules. Ten such distributions were created from sets of simulated clones, and the M_{CON} and M_{UNCON} models were fit to the distribution in each case. The red and blue curves show the average and standard deviation of the distributions M_{CON} and M_{UNCON} over the ten samples. The average ±SEM of the ten fitting errors J of each model to the distribution is shown in (b). These errors were not found to be significantly different (ANOVA, p=0.4462).

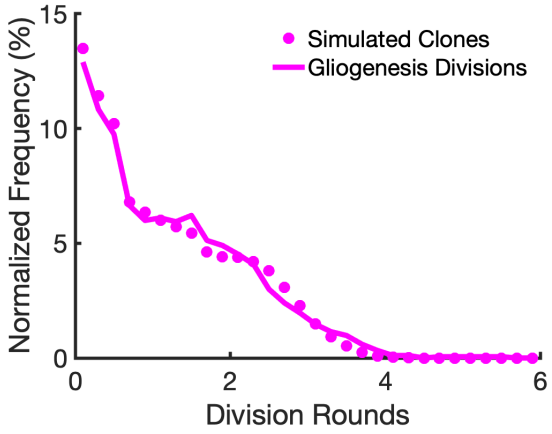


Figure 6.5 Calibration of the proliferation parameters for simulating gliogenesis under NGS-RGP rules. The curve represents the distribution of division rounds occurring in the data set of G and Mix clones, estimated using Eqn. 6.8 with $\mu = 4.9830$. The dots represent $\tilde{S}(n)$, the distribution of division rounds n computed in Eqn. 6.10 from a set of clones simulated in the NGS-RGP model with parameters $\beta = 0.3717$, ρ_1=0.4093, ρ_2=0.4418, and d=0.4600.

<0.05 in one out of ten cases. Again, this indicates that it is likely that the multi-Gaussian model with linearly spaced means sufficiently represents the clonal distribution.

It is worth noting that we have reached the same conclusion here and in Sec. 6.5.1 when using the same deterministic model with a different sets of parameters. Thus, our result appears to be robust and not dependent on selecting a particular set of division probabilities.

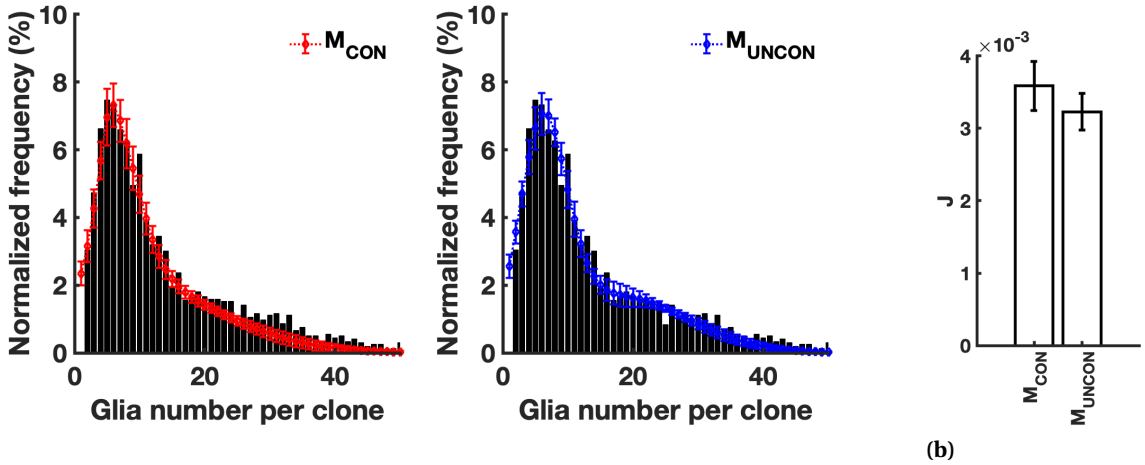


Figure 6.6 Results of gliogenesis simulation with deterministic NGS-RGP model. In (a), we show an example distribution of neurons per clone from a set of $N = 359$ clones simulated according to the NGS-RGP rules. Ten such distributions were created from sets of simulated clones, and the M_{CON} and M_{UNCON} models were fit to the distribution in each case. The red and blue curves show the average and standard deviation of the distributions M_{CON} and M_{UNCON} over the ten samples. The average ±SEM of the ten fitting errors J of each model to the distribution is shown in (b). These errors were not found to be significantly different (ANOVA, p=0.4104).

6.5.3 Stochastic gliogenesis model

6.5.3.1 Parameters

Previously, we determined the parameters that fit Q_n to the distribution of glia in G and Mix clones, shown in Table 3.1 of Sec. 3.2.4 (in the row for G and Mix, Slater). Since the distribution Q_n directly represents a stochastic GW process, we can use these parameters in the simulation. Thus, we set the parameters for the stochastic G-RGP model as $\xi_1 = 0$ and $\xi_2 = 0.0764$.

6.5.3.2 Simulations

Fig. 6.7 shows the frequency distribution of glia per clone $H(i)$ from $N = 359$ clones generated from the G-RGP model, along with the Q_n distribution fit to the simulated data. We ran ten such groups of clones, and for each group, we performed a Chi-square goodness of fit test for the Q_n distribution (see Sec. 3.2.4). In 9 out of 10 cases, the p-value from the test was above $\alpha = 0.05$, thus we could not reject the null hypothesis of the Chi-square test in these cases, indicating that the model Q_n sufficiently represented the simulated clones.

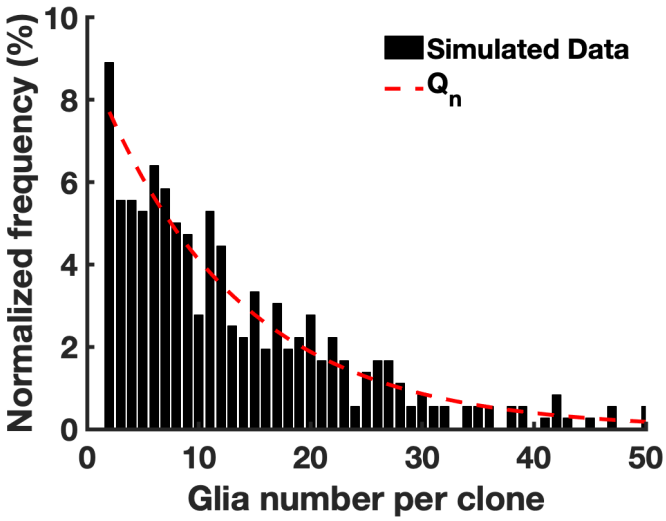


Figure 6.7 Distribution of clone sizes for a set of $N = 359$ clones simulated using the stochastic G-RGP model (histogram). The dashed red line shows the fit of the Q_n distribution to this particular set of clones. Q_n was judged to accurately represent the distribution of simulated clones in 9 out of 10 cases.

6.5.4 Combination gliogenesis model: deterministic and stochastic

6.5.4.1 Parameters

As in the previous section, the parameters for modeling the stochastic G-RGP portion of the simulated clones were already determined, shown in Table 3.1 of Sec. 3.2.4 (in the row for G and Mix, Slater). Hence, $\xi_1 = 0$ and $\xi_2 = 0.0820$. For the parameters governing the glia arising from mixed clones, we proceed as before by first setting the normal distribution parameters $\mu = 6.5899$ and $\sigma = 3.7997$ from the first Gaussian mean and standard deviation found when fitting M_{CON} to the distribution of glia in mixed clones in Sec. 5.6.2. The remaining parameters are calibrated as described in Sec. 6.3.1. Fig. 6.8b shows the distribution of division rounds $\tilde{S}(n)$ from 5000 clones simulated

using the found parameters β=0.2190, ρ_1=0.5356, ρ_2=0.4469, and d=0.4278. The curve represents $F(n)$ calculated from Eqn. 6.8 for clone sizes m coming from mixed clones and $\mu = 6.5899$.

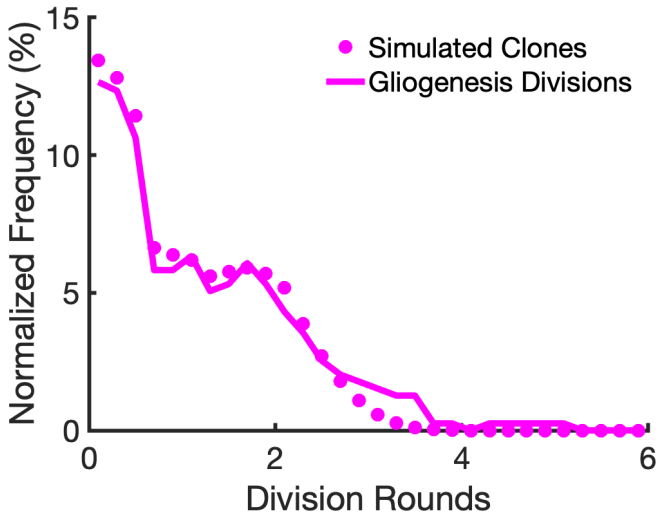


Figure 6.8 Calibration of the proliferation parameters for simulating deterministic gliogenesis in Mix clones under NGS-RGP rules. The curve represents the distribution of division rounds occurring in the data set of Mix clones, estimated using Eqn. 6.8 with $\mu = 6.5899$. The dots represent $\tilde{S}(n)$, the distribution of division rounds n computed in Eqn. 6.10 from a set of clones simulated in the NGS-RGP model with parameters β=0.2190, ρ_1=0.5356, ρ_2=0.4469, and d=0.4278.

6.5.4.2 Simulations

A sample simulation of 360 clones from the gliogenesis model is shown in Fig. 6.9, with average and standard deviations of the fits of multi-Gaussian models M_{CON} and M_{UNCON} with $k = 4$ over ten simulations. 180 clones came from each population. We see some evidence of the multi-Gaussian peaks coming from the mixed clones that follow the neurogenesis model, but more of an overall shape similar to the distribution formed from a Galton-Watson process as in Fig. 6.7. Out of the ten F-tests performed comparing models M_{CON} and M_{UNCON} to the simulated distribution of glia per clone, eight out of ten showed p<0.05. Thus, in these cases, M_{UNCON} represented the distribution better. The set of fitting errors J was also significantly lower for M_{UNCON} (ANOVA, p=4.6052e-04).

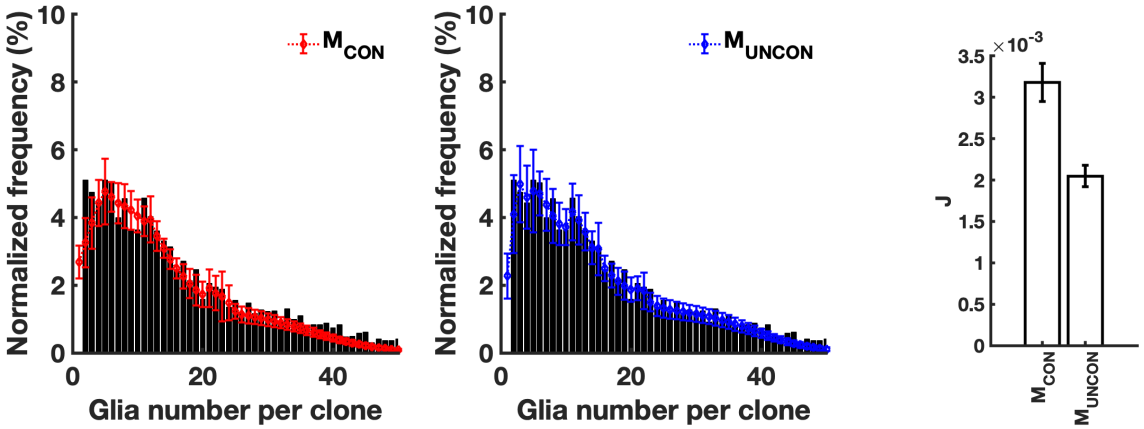
Additionally, we fit the Q_n distribution to each distribution of simulated clones. One realization of this is shown in Fig. 6.10. We performed Chi-square test (Sec. 3.2.4) to evaluate the fit of Q_n to each of the ten simulated groups of clones. In only one out of ten cases, the Chi-square p value was below 0.05. Thus, in most cases, we cannot reject the null hypothesis of the Chi-square test,

indicating that Q_n accurately represents this combined distribution of clones.

This is an interesting result; for this simulation of clones with 50% following deterministic NGS-RGP rules and 50% following stochastic G-RGP rules, the combined distribution of clones can be represented by the stochastic model Q_n. This result does not match the previous result from fitting Q_n, M_{CON}, and M_{UNCON} to the distribution of Mix+G clone sizes in the true NGS data, where Q_n and M_{CON} both failed to represent the distribution. As an experiment, we test different percentages of G-RGPs in the simulation population, from 10% to 50%. Table 6.1 lists several values found from evaluating the fits of Q_n, M_{CON}, and M_{UNCON} to distributions of clones simulated with these percentages of G-RGP inclusion. First, the average initial mean for M_{CON} is listed, as well as all values of $\mu_1-\mu_4$ for the M_{UNCON} fit. We note that in general, the values found for $\mu_1-\mu_3$ for M_{UNCON} fall in a range similar to the values $\mu_1-\mu_3$ when fitting M_{UNCON} to the true data in Table 5.3 (4-5 for μ_1, 9-10 for μ_2, and 15-18 for μ_3). Thus, we appear to have reasonably identified rules that can produce a distribution of simulated clones similar to the data. Below the values of $\mu_1-\mu_4$ in Table 6.1 are the p-values found from comparing the sets of fitting errors between M_{CON} and M_{UNCON} using ANOVA. Lastly, we list the percent of cases in which the p-value for the Chi-square test, evaluating the fit of Q_n to the simulated clones, was below 0.05.

From the values listed in Table 6.1, we can see that simulating groups of clones with varying percentages of G-RGPs alters the observed means as well as the results of the statistical tests. At lower % G-RGPs, the ANOVA p-value is high, indicating that M_{CON} accurately represents these simulated distributions. On the other hand, with a lower % G-RGPs, the p-values in the Chi-square test are frequently >0.05, hence Q_n does not represent the simulated distribution. This makes sense; since the G-RGPs follow the Q_n distribution and NGS-RGPs follow the M_{CON} distribution, a low percentage of G-RGPs and a high percentage of NGS-RGPs in the simulated population should result in a distribution matching M_{CON} and not Q_n. We also see from the table that as the % G-RGPs increases, the results of these statistical tests are reversed, with the ANOVA p-value being low and the %p<0.05 in the Chi test also decreasing.

We recall that when fitting Q_n, M_{CON}, and M_{UNCON} to the true Mix+G NGS data, we found a small p-value for the error comparison using ANOVA (p=0.0044, Sec. 5.6.3) and a small Chi-square p-value (p=0.0254, Sec. 3.2.4). Thus, to match the results from the data, we would want a low ANOVA p-value, and a higher percentage of Chi-square p-values below 0.05. Among the simulations with different inclusion levels of G-RGPs in Table 6.1, it appears that simulating around 30% of the clones as G-RGPs produces a result most consistent with the true data. We can therefore hypothesize that about 70% of the RGP population present during the neurogenesis-to-gliogenesis switch are NGS-RGPs, while 30% are G-RGPs. This hypothesis could potentially be explored in future biological experiments on clones during the NGS.



(a) (b)

Figure 6.9 Results of gliogenesis simulation with clones coming from a combination of the deterministic NGS-RGP and stochastic G-RGP models. In (a), we show an example distribution of glia per clone from a set of $N = 360$ clones, where half were simulated under each model. Ten such distributions were created from sets of simulated clones, and the M_{CON} and M_{UNCON} models were fit to the distribution in each case. The red and blue curves show the average and standard deviation of the distributions M_{CON} and M_{UNCON} over the ten samples. The average ±SEM of the ten fitting errors J of each model to the distribution is shown in (b). These errors were found to be significantly different (ANOVA, p=4.6052e-04), hence M_{CON} failed to represent the distribution of simulated clones.

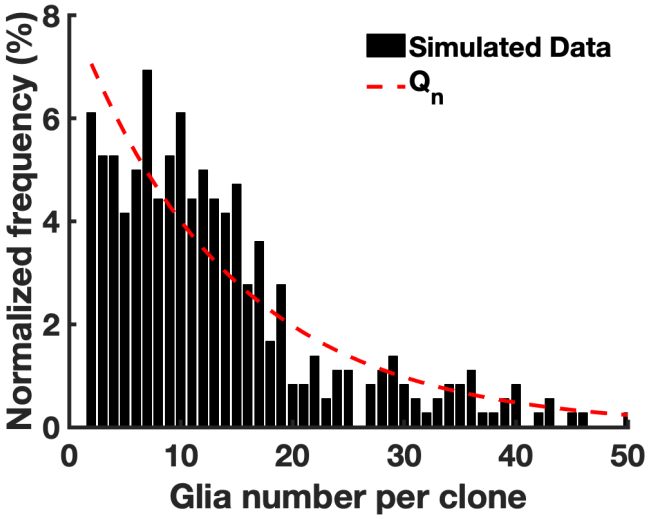


Figure 6.10 Fit of the Q_n distribution to the distribution of glia per clone in the simulated population of half G-RGPs and half NGS-RGPs. Q_n was judged to accurately represent the distribution of simulated clones in 9 out of 10 cases.

Table 6.1 Clonal simulations with varying percent inclusion of G-RGPs. Ten samples were simulated for each percent inclusion, and models (6.9), (5.1), and (5.2) were fit to the distribution of glia per clone for each simulation. For the Gaussian models (5.1) and (5.2), the average values of $\mu_1 - \mu_4$ over the ten simulations for each % G-RGPs are shown. The ten sets of residual errors from fitting these two models were compared with ANOVA, and the p-value for this comparison is shown. For the stochastic model (6.9), the fit for each simulated sample was evaluated with a Chi-square test. The percentage of the ten Chi-square p-values below $\alpha = 0.05$ for each % G-RGP inclusion level is shown.

% G-RGPs	10	20	30	40	50
μ_1, M_{CON}	8.0800	8.5452	6.8747	6.3138	6.1226
μ_1, M_{UNCON}	5.8618	5.0065	3.6433	4.2157	3.6080
μ_2, M_{UNCON}	9.1842	9.4269	9.1615	10.2648	9.6180
μ_3, M_{UNCON}	16.6824	15.7339	16.7188	18.8568	15.2574
μ_4, M_{UNCON}	30.8426	31.1692	27.8096	31.0609	27.5572
p-value, ANOVA	0.7409	0.9850	0.0224	0.0060	4.6052e-04
% p<0.05, Chi	80%	60%	30%	0%	10%

6.6 Discussion and conclusions

In this thesis, we have used several mathematical models and methods to identify mechanisms of radial glial progenitor (RGP) proliferation and differentiation during cortical development and the neurogenesis-to-gliogenesis switch (NGS). These methods have been driven by a set of clonal data gathered using mosaic analysis with double markers (MADM). This technique creates a unique type of data measuring the total cells arising from the two daughter lineages of individual cells, namely, the neurons and glia arising from the divisions of RGPs (clones). Patterns present in the distributions of glia per clone can indicate that the RGPs produce glia deterministically or stochastically.

Based on our analysis of the data with branching processes (Chapter 3), clustering methods and statistical tests (Chapter 4), and Gaussian mixture models (Chapter 5), we proposed a set of rules that RGPs follow while dividing and differentiating into glia during the NGS. These rules defined two subpopulations of RGPs: deterministic NGS-RGPs and stochastic G-RGPs. We demonstrated in this chapter that simulating glial production in a population composed of a combination of these two subpopulations could produce clone size distributions with similar features as the true data. That is, the same type of model M_{UNCON} represented both the simulated data and the true NGS data better than two simpler models M_{CON} and Q_n. These two simpler models represented strictly deterministic and stochastic mechanisms of glial production, respectively. Thus, our results imply that glial production during the NGS is neither strictly deterministic nor strictly stochastic, and suggests that a combination of behaviors may be present. Additionally, this result depended on the relative proportion of NGS-RGPs and G-RGPs in the simulated population (Table 6.1). We can

hypothesize that the percentage of G-RGPs (30%) which created a clone size distribution best fit by M_{UNCON} over the other two models is the percentage of G-RGPs present in the real population of RGPs during the neurogenesis-to-gliogenesis switch. Thus, our final hypothesis, which we have developed from mathematical analysis of this clonal dataset, is that the NGS and gliogenesis can be characterized by a mixture of deterministic and stochastic RGP behaviors, and that approximately 30% of the RGPs produce glia in a stochastic manner.

In formulating this hypothesis, a few features in the data were not addressed. First, our analysis of the knockout (CKO) clones was limited, since this dataset was smaller than the WT dataset. For instance, we were unable to definitively identify deterministic or stochastic patterns in the distribution of glia per CKO clone in Sec. 5.6.3. We thus did not simulate any mechanisms of cell division that RGPs may follow in response to the EGFR knockout. However, we did observe patterns in the CKO clones that allow us to hypothesize about the response of RGPs to EGFR knockout. In Sec. 4.3, we found that Mix clones produced a similar distribution of total glia in both WT and CKO clones, while the average glia produced in G clones was increased. By our hypothesis of two RGP populations, this would indicate that G-RGPs produce more total glia and 'overcompensate' for the loss of green glia in response to EGFR knockout, but that NGS-RGPs do not.

Second, we did not incorporate the spatial portion of the clonal data into our simulation. It was observed in Fig. 4.4 that the broad location of clones in the six layers of the cortex (superficial, deep, or all layers) was correlated with clone size and potentially with genotype (WT or CKO). However, it was challenging to compare the clone size distributions in the subsets of data separated by location, since the subsets of deep and all layer clones contained much fewer clones than the subset of superficial clones. Attempting to partition these groups further and see how location was associated with other features, such as development time (E15.5, E16.5, or E17.5) or clone type (N, G, or Mix), resulted in even smaller subsets. These small subsets could be compared using statistical tests as in Chapter 4, but their clone size distributions did not contain enough data points to produce identifiable deterministic or stochastic patterns. Thus, we were not able to identify specific rules of clonal behavior that were associated with location, and we left this feature out of our simulation framework. More clonal data could be gathered to increase the sizes of the data subsets by location, but we note that MADM data collection is highly time intensive and costly.

Lastly, our simulations did not incorporate the separate red and green cell counts for each simulated clone, but instead focused on the total glia per clone. Thus, our simulated clones did not have any rules so that their red and green lineages matched the patterns of red/green symmetry and asymmetry shown in Sec. 5.3. Part of the issue encountered with comparing the red and green lineages in the NGS clones was that a significant portion of the G clones (>50%) contained only red or only green glia. However, this in and of itself is an interesting observation in the NGS data. If a clone produces only red or only green glia, this implies that the other lineage either died or was not located where it could be observed. This could perhaps tie into the previously mentioned hypothesis

that G-RGPs migrate during the NGS (Sec. 4.3). Overall, the complexity of developing rules that would control the symmetries of simulated clones was deemed beyond the scope of this thesis. If a set of rules was developed, we would want to incorporate a dependence on EGFR, which would enable us to simulate the effect of EGFR knockout on glial production. We would likely require data with a larger number of the CKO clones to more definitively establish these rules. Thus, there is ample opportunity to expand on the mathematical analysis of NGS clones presented in this thesis. However, since the methods we used were data-driven, any further work leading to a robust set of extended mathematical models would require additional data.

BIBLIOGRAPHY

[1] Aguirre, A., Dupree, J., Mangin, J. and Gallo, V. "A functional role for EGFR signaling in myelination and remyelination". *Nat Neurosci* **10** (2007), pp. 990–1002.

[2] Aguirre, A., Rizvi, T., Ratner, N. and Gallo, V. "Overexpression of the epidermal growth factor receptor confers migratory properties to nonmigratory postnatal neural progenitors". *J Neurosci* **25** (2005), pp. 11092–11106.

[3] Allen, M. *Understanding Regression Analysis*. Springer, 1997.

[4] Axelrod, D. and Kimmel, M. *Branching Processes in Biology*. Springer, 2015.

[5] Bachoo, R., Maher, E., Ligon, K., Sharpless, N., Chan, S., You, M., Tang, Y., DeFrances, J., Stover, E., Weissleder, R., Rowitch, D., Louis, D. and DePinho, R. "Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis". *Cancer Cell* **1** (2002), pp. 269–277.

[6] Beattie, R. and Hippenmeyer, S. "Mechanisms of radial glia progenitor cell lineage progression". *FEBS Letters* **24** (2017), pp. 3993–4008.

[7] Boije, H., Rulands, S., Dudczig, S., Simons, B. and Harris, W. "The independent probabilistic firing of transcription factors: a paradigm for clonal variability in the zebrafish retina". *Dev Cell* **34** (2015), pp. 532–543.

[8] Burrows, R., Wancio, D., Levitt, P. and Lillien, L. "Response diversity and the timing of progenitor cell maturation are regulated by developmental changes in EGFR expression in the cortex". *Neuron* **19** (1997), pp. 251–267.

[9] Calegari, F., Haubensak, W., Haffner, C. and Huttner, W. "Selective lengthening of the cell cycle in the neurogenic subpopulation of neural progenitor cells during mouse brain development". *The Journal of Neuroscience* **25** (2005), pp. 6533–6538.

[10] Crane, G., Taghizadeh, R. and Sherley, J. "In vitro evidence for differentiation resistance by distributed stem cells during deterministic asymmetric self-renewal". *Journal of Stem Cell Research and Medicine* **2** (2017), pp. 1–8.

[11] D., F. and Diaconis, P. "On the histogram as a density estimator: L2 theory". *Probability Theory and Related Fields* **57** (1981), pp. 453–476.

[12] Doetsch, F., Petreanu, L., Caille, I., Garcia-Verdugo, J. and Alvarez-Buylla, A. "EGF converts transit-amplifying neurogenic precursors in the adult brain into multipotent stem cells". *Neuron* **36** (2002), pp. 1021–1034.

[13] Drmota, M. *Random Trees: An Interplay between Combinatorics and Probability*. Springer-Verlag/Wien, 2009.

[14] Espinosa, J. and Luo, L. “Timing neurogenesis and differentiation: insights from quantitative clonal analyses of cerebellar granule cells”. *J Neurosci* **28** (2008), pp. 2301–2312.

[15] Galvez-Contreras, A., Quinones-Hinojosa, A. and Gonzalez-Perez, O. “The role of EGFR and ErbB family related proteins in the oligodendrocyte specification in germinal niches of the adult mammalian brain”. *Front Cell Neurosci* **7** (2013), p. 258.

[16] Gao, P., Postiglione, M., Krieger, T., Hernandez, L., Wang, C., Han, Z., Streicher, C., Papusheva, E., Insolera, R. and Chugh, K. “Deterministic progenitor behavior and unitary production of neurons in the neocortex”. *Cell* **159** (2014), pp. 775–788.

[17] Gonzalez-Perez, O., Romero-Rodriguez, R., Soriano-Navarro, M., Garcia-Verdugo, J. and Alvarez-Buylla, A. “Epidermal growth factor induces the progeny of subventricular zone type B cells to migrate and differentiate into oligodendrocytes”. *Stem Cells* **27** (2009), pp. 2032–2043.

[18] Grimaldi, R. *Fibonacci and Catalan Numbers: An Introduction*. Wiley and Sons, 2012.

[19] Guo, S. e. a. “Nonstochastic reprogramming from a privileged somatic cell state”. *Cell* **156** (2014), pp. 649–662.

[20] Hanna, J., Saha, K., Pando, B., Zon, J. van, Lengner, C., Creyghton, M., Oudenaarden, A. van and Jaenisch, R. “Direct cell reprogramming is a stochastic process amenable to acceleration”. *Nature* **462** (2009), pp. 595–601.

[21] Ivkovic, S., Canoll, P. and Goldman, J. “Constitutive EGFR signaling in oligodendrocyte progenitors leads to diffuse hyperplasia in postnatal white matter”. *J Neurosci* **28** (2008), pp. 914–922.

[22] Kirischuk, S., Sinning, A., Blanquie, O., Yang, J., Luhmann, H. and Kilb, W. “Modulation of neocortical development by early neuronal activity: physiology and pathophysiology”. *Front Cell Neurosci* **11** (2017).

[23] Kohonen, H. *Textbook of Neuroanatomy*. Philadelphia: Lippincott, 1963.

[24] Kohonen, T. *Self-Organizing Neural Networks: Recent Advances and Applications*. Physica-Verlag HD, 2002.

[25] Kornblum, H., Hussain, R., Wiesen, J., Miettinen, P., Zurcher, S., Chow, K., Derynck, R. and Werb, Z. “Abnormal astrocyte development and neuronal death in mice lacking the epidermal growth factor receptor”. *J Neurosci* **53** (1998), pp. 697–717.

[26] Kriegstein, A. and Alvarez-Buylla, A. “The glial nature of embryonic and adult neural stem cells”. *Annu Rev Neurosci* **32** (2009), pp. 149–184.

[27] Kuhn, H., Winkler, J., Kempermann, G., Thal, L. and Gage, F. “Epidermal growth factor and fibroblast growth factor-2 have different effects on neural progenitors in the adult rat brain”. *J Neurosci* **17** (1997), pp. 5820–5829.

[28] Ligon, K., Fancy, S., Franklin, R. and Rowitch, D. “Olig gene function in CNS development and disease”. *Glia* **54** (2006), pp. 1–10.

[29] Macken, C. and Perelson, A. *Stem Cell Proliferation and Differentiation: A Multitype Branching Process Model*. Berlin, Springer-Verlag, 1988.

[30] MacQueen, J. “Some methods for classification and analysis of multivariate observations”. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** (1967), pp. 281–297.

[31] McClintic, J. *Human Anatomy*. St. Louis: Mosby, 1983.

[32] Meyer, G. “Genetic control of neuronal migrations in human cortical development”. *Advances in anatomy, embryology, and cell biology* **189** (2007).

[33] Mihalas, A. and Hevner, R. “Clonal analysis reveals laminar fate multipotency and daughter cell apoptosis of mouse cortical intermediate progenitors”. *Development* **145** (2018).

[34] Mizrahi, S., Sandler, O., Lande-Diner, L., Balaban, N. and Simon, I. “Distinguishing between stochasticity and determinism: examples from cell cycle duration variability”. *BioEssays* **38** (2015).

[35] Molofsky, A. and Deneen, B. “Astrocyte development: a guide for the perplexed”. *Glia* **8** (2015), pp. 1320–1329.

[36] Ng, M. P. and Wormald, N. C. “Reconstruction of rooted trees from subtrees”. *Discrete Applied Mathematics* **69** (1996), pp. 19–31.

[37] *NIST/SEMATECH e-Handbook of Statistical Methods*. URL: `http://www.itl.nist.gov/div898/handbook/`.

[38] Noctor, S., Flint, A., Weissman, T., Dammerman, R. and Kriegstein, A. “Mechanisms of radial glia progenitor cell lineage progression”. *Nature* **6821** (2001), pp. 714–720.

[39] Noctor, S., Martinez-Cerdeno, V., Ivic, L. and Kriegstein, A. “Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases”. *Nature Neuroscience* **7** (2004), pp. 136–144.

[40] Picco, N., Hippenmeyer, S., Rodarte, J., Streicher, C., Molnar, Z., Maini, P. and Woolley, T. “A mathematical insight into cell labeling experiments for clonal analysis”. *Journal of Anatomy* **235** (2019), pp. 686–696.

[41] Qian, X., Shen, Q., Goderie, S., He, W., Capela, A., Davis, A. and Temple, S. “Timing of CNS cell generation: a programmed sequence of neuron and glial cell production from isolated murine cortical stem cells”. *Neuron* **28** (2000), pp. 69–80.

[42] Rakic, P. “Mode of cell migration to the superficial layers of fetal monkey neocortex”. *Nature* **1** (1972), pp. 61–83.

[43] Rakic, P. “Evolution of the neocortex: perspective from developmental biology”. *Nat Rev Neurosci* **10** (2009), pp. 724–735.

[44] Reynolds, J., Coles, M., Lythe, G. and Molina-Paris, C. “Deterministic and stochastic naive T cell population dynamics: symmetric and asymmetric cell division”. *Dynamical Systems* **27** (2012), pp. 75–103.

[45] Ro, S. and Rannala, B. “Methylation patterns and mathematical models reveal dynamics of stem cell turnover in the human colon”. *Proc of the Nat Acad of Sci* **19** (2001).

[46] Rosen, K. *Handbook of Discrete and Combinatorial Mathematics*. New York: Chapman and Hall/CRC, 2017.

[47] Sibilia, M., Steinbach, J., Stingl, L., Aguzzi, A. and Wagner, E. “A strain-independent postnatal neurodegeneration in mice lacking the EGF receptor”. *EMBO J* **17** (1998), pp. 719–731.

[48] Slater, J., Landman, K., Hughes, B., Shen, Q. and Temple, S. “Cell lineage tree models of neurogenesis”. *Journal of Theoretical Biology* **256** (2009), pp. 164–179.

[49] Sun, Y., Goderie, S. and Temple, S. “Asymmetric distribution of EGFR receptor during mitosis generates diverse CNS progenitor cells”. *Neuron* **45** (2005), pp. 873–886.

[50] Theunissen, T. and Jaenisch, R. “Molecular control of induced pluripotency”. *Cell Stem Cell* **6** (2014), pp. 720–734.

[51] Wagner, B., Natarajan, A., Grunaug, S., Kroismayr, R., Wagner, E. and Sibilia, M. “Neuronal survival depends on EGFR signaling in cortical but not midbrain astrocytes”. *EMBO J* **25** (2006), pp. 752–762.

[52] Wong, S., Scott, E., Mu, W., Guo, X., Borgenheimer, E. and Freeman, M. e. a. “In vivo clonal analysis reveals spatiotemporal regulation of thalamic nucleogenesis”. *PLos Biol* **16** (2008).

[53] Yakovlev, A., Stoimenova, V. and Yanev, N. “Branching processes as models of progenitor cell populations and estimation of the offspring distributions”. *Journal of the American Statistical Association* **103** (2008), pp. 1357–1366.

[54] Yunusova, A., Fishman, V., Vasiliev, G. and Battulin, N. “Deterministic versus stochastic model of reprogramming: new evidence from cellular barcoding technique”. *Open Biol* **7** (2017).

[55] Zong, H., Espinosa, J., Su, H., Muzumdar, M. and Luo, L. “Mosaic analysis with double markers in mice”. *Cell* **121** (2005), pp. 479–492.

[56] Zonouzi, M., Scafidi, J., Li, P., McEllin, B., Edwards, J., Dupree, J., Harvey, L., Sun, D., Hubner, C., Cull-Candy, S. and Farrant M. ad Gallo, V. “GABAergic regulation of cerebellar NG2 cell development is altered in perinatal white matter injury”. *Nat Neurosci* **18** (2015), pp. 674–682.