UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

Mathematical modelling of cell fate dynamics in homeostasis

by

Cristina Parigini

MSc

ORCiD: 0000-0002-0468-0432

A thesis for the degree of
Doctor of Philosophy

March 2022

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

Mathematical modelling of cell fate dynamics in homeostasis

by Cristina Parigini

Many biological tissues are not static but continuously renewed through cycles of cell production and cell loss which must be perfectly balanced to maintain the tissue's healthy state, also called homeostasis. The underlying dynamics of cell fate choices in homeostasis are complex and often not well understood. Although an experimental approach is of utmost importance to understand the mechanism regulating cell fate, mathematical modelling of the cell fate dynamics is essential to interpret experimental data. This project develops a framework for studying cell fate dynamics in homeostasis that combines theoretical modelling and numerical simulations given lineage-tracing experimental data. A correct and reliable definition of a cell fate model is a complex task due to the number of unknowns, the scarcity of the data and their uncertainty. Therefore, our approach is to simplify the problem of identifying the lineage hierarchy and the cell proliferation, differentiation and death rates by restricting the search to models compatible with homeostasis and presenting specific tissue-related features. For doing so, we use graph theory, deterministic approximation, stochastic models and Bayesian inference. Based on purely theoretical considerations, this research proves that any homeostatic cell fate model must follow strict rules, requiring self-renewing cells at the apex of the lineage hierarchy and only there. Importantly, self-renewal does not need to be an intrinsic property of a cell type since any cell type located at the apex of a lineage hierarchy may acquire it by interacting with the cell environment. Besides, we showed how stem cells and their self-renewing strategy could be determined based on qualitative features of lineage-tracing experimental data, such as the shape of the clonal size distribution and discrepancies in cell cluster sizes from tissue assays. The developed framework is validated using synthetic data for a study case, the mouse mammary gland, paving the way for future studies where experimental data might be available.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    - Philip Greulich, Benjamin D. MacArthur, Cristina Parigini, and Rubén J. Sánchez-García. Stability and steady state of complex cooperative systems: A diakoptic approach. *Royal Society Open Science*, 6(12), 2019. ISSN 20545703. doi: 10.1098/rsos.191090
    - Cristina Parigini and Philip Greulich. Universality of clonal dynamics poses fundamental limits to identify stem cell self-renewal strategies. *eLife*, 9:1–44, 2020. ISSN 2050084X. doi: 10.7554/eLife.56532
    - Philip Greulich, Benjamin D. MacArthur, Cristina Parigini, and Rubén J. Sánchez-García. Universal principles of lineage architecture and stem cell identity in renewing tissues. *Development (Cambridge)*, 148(11), 2021. ISSN 14779129. doi: 10.1242/DEV.194399

Signed:..... Date:.....

Acknowledgements

I gratefully acknowledge the financial support from the Institute for Life Sciences and the University of Southampton in funding my research project.

Besides, in such difficult times, this project would not have been possible without the support of many people.

First of all, I would like to express my gratitude to my supervisor Dr *Philip Greulich*. Philip, you transmitted your enthusiasm for cell biology, a field entirely new for me. Often, you showed me the mathematical way where I only saw the engineering one. Thank you for your consistent support, encourage and patience. I appreciate that my worldwide movings and growing family made me an atypical PhD student. My sincere thanks also go to my co-supervisor, Prof *Ben MacArthur*. Thank you, Ben, for your valuable advice and constant encouragement. To Prof *Ruben Sanchez Garcia*, Ben and Philip, thank you for the fascinating and enlightening discussions about cell dynamics and network theory. Through these, I gained invaluable knowledge, essential for carrying out my research. I want to thank also Dr *Salah Elias* for the opportunity to work on data analysis, even if things did not turn out as we were hoping to. I am grateful also to Ms Kulvir Bouri, the Graduate School Student Office and iSolutions for their prompt help whenever I needed it.

Of course, I also owe very special thanks to my family and friends. Dear *Sofia*, *Cecilia* and *Anna*, thank you for your immense patience when mum "plays" on the computer instead of playing with you. Dear *Roberto*, thank you for supporting and "sopporting" me every single day. I could not be luckier. Dear *mum* and *dad*, thank you for your unconditional love, no matter how far I am. Dear *Chiara*, *Franci*, *Lisa*, *Alberto*, *Giorgio*, *Brun*, *Eleo* and *Marietto*, there are no words for saying how grateful I feel for having you as friends. Despite the years and the distance, I know I can count on you. Last but not least, thank you *Mose*. You made our Fridays while in the U.K. special, and we miss you a lot.

Finally, I acknowledge the permission to reproduce Figure 1.1 which has been granted by Elsevier, and Figure 1.2 granted by Springer Nature Customer Service Centre GmbH.

Definitions and Abbreviations

α	Kinetic parameter
δ	Local loss rate
γ	Cell death rate
κ	Total transition rate
λ	Cell division rate
μ	Dominant eigenvalue/growth parameter
ω	Cell transition rate
ρ	Cell density
n	Number of cell
r	Cell division outcome probability/symmetric division fraction
T	Cell type
X	Cell state
B	Basal
DNA	DeoxyriboNucleic Acid
DE	Differentially Expressed
(G)IA	(Generalised) Invariant Asymmetry
(G)PA	(Generalised) Population Asymmetry
HVG	High Variable Gene
MaSC	Mammary Stem Cell
LI	Luminal Intermediate
LM	Luminal Mature
LP	Luminal Progenitor
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
RNA	RiboNucleic Acid
SCC	Strongly Connected Component
scRNA-seq	single-cell RNA-sequencing
SNN	Shared Nearest Neighbor
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMI	Unique molecular identifiers

Chapter 1

Introduction

During the development, growth and adult life of a multi-cellular organism, cells continuously proliferate, differentiate and die. Complex dynamics direct the whole process. These dynamics depend on cell states, related, among others, to the protein and RNA levels within the cell, and external signals, such as spatial constraint, biochemical and mechanical signalling. All cells share the same DNA, which contains the complete genetic information of a given organism. Still, the gene expression regulates each cell's functionality as a part of a whole complex system [Alberts et al., 2015].

In tissue development, regeneration, and maintenance, *stem cells* are often defined as cells having the capability of self-renewal. That is the cell's ability to produce more cells identical to itself and recreate a whole functional tissue through proliferation and differentiation [National Institute of Health, 2016]. Stem cells can be classified depending on their degree of *potency*, which is related to the number of differentiated cell types they can produce. Remarkably, a *pluripotent* stem cell has the potential to build an entire adult organism. However, due to the complexity of their identification and classification, there is not a clear and unique definition of a stem cell [Potten and Loeffler, 1990, Shostak, 2006, Vorotelyak et al., 2020].

Identifying stem cells and determining their cell fate choices are also not trivial, and there are still many controversies and open points. As shown in [Blanpain and Simons, 2013], stem cell-specific molecular markers are scarce and rarely linked to function, and stem cells operate in a noisy and dynamic environment. A review of different experimental approaches for identifying stem cells is [Snippert and Clevers, 2011]. Here, it is shown that in-vivo cell lineage tracing, based on a genetic marking of a specific cell to study its progeny, is an effective way of studying actual stem cells in their physiological context. In general, the cell's identity in a given moment depends on its internal state, which can be determined by measuring the gene expression. One emerging approach to transcriptome profiling is *RNA-sequencing* (RNA-seq). Recently,

researchers further exploited RNA-seq approach by applying it to single cells, *single cell RNA-seq* (scRNA-seq). Notably, scRNA-seq overcomes the intrinsic limitation of averaging the expression profiles when analysing pools of cells [Treutlein et al., 2014].

Although the development and use of experimental procedures are of undoubted importance, mathematical modelling is essential for unravelling the *cell fate* dynamics problem, that is, the cell division and differentiation. Nevertheless, experimental methods and conventional statistical methods alone cannot always infer the stem cells' identity and dynamics [Rulands and Simons, 2016]. Within this context, in [Clayton et al., 2007, Doupé et al., 2012, Alcolea et al., 2014] cell lineage data are combined with novel mathematical modelling approaches to investigate cell dynamics, but this type of analysis is tailored for specific cell fate models, e.g. single progenitor cell fate model.

Therefore, the idea behind the present work is to build a framework to determine suitable cell fate dynamics models representing any homeostatic renewing tissue for which experimental data are available. Instead of directly applying a model search and parameter fitting standard methods, we first study the generic features of such dynamics from a mathematical standpoint. Using theoretical and numerical means, we want to exclude a priori models that are not compatible with homeostasis and distinguish classes of models presenting common behaviours which we can easily compare with experimental data. Finally, the derived outcomes are applied to synthetic data representing a study case based on the mouse mammary gland to validate the proposed approach.

This chapter provides an introduction to the modelling of the cell dynamics in adult tissues. More specifically, in Section 1.1 the common self-renewing strategies and their mathematical models are described. The clonal dynamic modelling, which is essential for assessing self-renewal strategies based on experimental data, is discussed in Section 1.2. In Section 1.3, an overview of two promising experimental methods, lineage tracing and single-cell RNA-sequencing, is provided. Section 1.4 presents a study case, the mouse mammary gland. The research aim and objectives are reported in Section 1.5 and the methodology applied in this work, the innovative contribution and the organisation of this thesis are presented in Section 1.6.

1.1 Homeostasis in adult renewing tissues

A *renewing tissue* of an adult individual, also called cycling, is a tissue characterised by a cell turnover, in which there is a balance between cell proliferation and death. This steady-state condition is commonly called *homeostasis*. Adult stem cells are the key players for maintaining and renewing such tissues due to their ability to produce cells through cell division and differentiation persistently [National Institute of Health,

2016]. Thus, for maintaining tissues in a homeostatic state, stem cells must adopt suitable self-renewal strategies, a pattern of fate choices that balances proliferation and differentiation. Any unbalance leading to cell hyper-proliferation is one of the first steps towards cancer development. Thus, understanding and identifying self-renewal strategies have been a fundamental goal of stem cell and cancer biology ever since discovering adult stem cells.

Two patterns, sketched in Figure 1.1, taken from [Simons and Clevers, 2011a], are commonly considered: *Invariant Asymmetry* (IA), also called asymmetric division, and *Population Asymmetry* (PA) [Potten and Loeffler, 1990, Watt and Hogan, 2000, Simons and Clevers, 2011a, Klein and Simons, 2011]. In these models, cells are classified into two cell types that share the same function in the tissue. The *differentiated* cells are cells committed to stop proliferating and eventually die, and the stem cells, also called *progenitor*, are cells that divide without losing their proliferative potential. In the IA model (cases A and B), the equilibrium condition is simply a consequence of the fact that, for every loss of a differentiated cell, a stem cell asymmetrically divides into a stem and a differentiated cell. In this way, the numbers of stem and differentiated cells are maintained. Considering the PA model (cases C and D), each stem cell can either divide symmetrically, asymmetrically or differentiate. In this case, the cell fate is stochastic, and the system reaches an equilibrium at a global level. The regulation of both self-renewing strategies may be internal, that is, from factors within the cells (cases A and C). In other contexts, *niche* factors, which are external to cells, regulate the cell proliferation and differentiation, cases B and D. Over the years, the population asymmetry model has become more relevant since many recent studies have shown the prevalence of this model in many mammalian renewing tissues [Simons and Clevers, 2011a].

It is worth noting that the number of cell types and transitions between them can be much more complicated than that assumed for the IA and PA models, and, in general, it is not clear whether those concepts can be generalised to much more complicated lineages. In this frame, [Greulich and Simons, 2016] proposes a more general mechanism to describe homeostasis in adult renewing tissues. In this work, the authors suggested the existence of a reversible state change, proposed in [Potten and Loeffler, 1990] to explain tissue regeneration, in which generic differentiated and non-proliferative cells are capable of turning back into a proliferative state. This model, called Dynamic Heterogeneity, agrees with experimental data equally well as a more classical model where this reversal rate is not present. However, in contrast to the classical model, the cell population's stability is more robust in the case of a Dynamic Heterogeneity model. As shown in this work and others, such as [Johnston et al., 2007, Sun and Komarova, 2012], homeostasis *robustness* is another crucial topic that requires investigation since, in the absence of proliferation and differentiation control, any stochastic fluctuation potentially disrupts homeostasis.

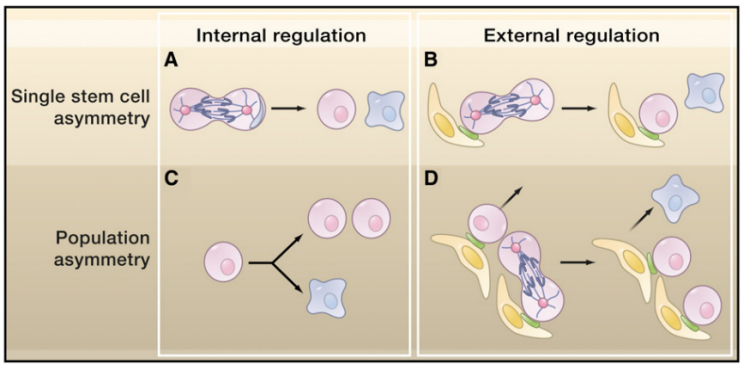


FIGURE 1.1: Sketch of the strategies of stem cell self-renewal taken from [Simons and Clevers, 2011a]. Reprinted from Cell, 145/6, B. D. Simons and H. Clevers, Strategies for homeostatic stem cell self-renewal in adult tissues, Pages No. 851-862, Copyright (2011), with permission from Elsevier. Stem cells are shown in pink, differentiated cells in light blue, and the niche in yellow. In the Invariant Asymmetry pattern (A and B), a stem cell asymmetrically divides into another stem cell and a differentiated cell. In the Population Asymmetry pattern (C and D), symmetric division and cell differentiation are possible stem cell fates. In both patterns, the regulation can be internal to the stem cell (A and C) or external, coming from the environment (B and D).

1.1.1 Mathematical modelling of cell fate dynamics

The two classical stem cell self-renewal strategies described above, the Invariant Asymmetry and the Population Asymmetry, are commonly represented as two cell types systems. These types are the stem cells (S), which can self-renew, i.e. divide without reducing their future potential to divide, and the differentiating cells (D). A multi-type branching process [Haccou et al., 2005], defined by the outcomes of cell divisions, the *cell fate choices*, models both strategies. This parametric model is as follows

$$S \xrightarrow{\lambda} \begin{cases} S+S & \text{with probability } r_S \\ S+D & \text{with probability } 1-r_S-r_D \\ D+D & \text{with probability } r_D \end{cases}, \tag{1.1}$$

in which cells of type S divide with the rate λ. Assuming that divisions events are independent for different cells and that the waiting time between two consecutive events is exponentially distributed with average $1/\lambda$, the process is Markovian. Here, daughter cells configuration $S+S$ corresponds to *symmetric self-renewal division* and $D+D$ to *symmetric differentiation*, while daughter cells of different type, $S+D$, marks an *asymmetric division*. The two self-renewal strategies, IA and PA, are distinguished by the value of the *fractions of symmetric division*, r_S and r_D. The PA model corresponds to any $0 < r_{S,D} \leq 1$ (with $r_S + r_D \leq 1$); the IA model is defined by $r_S = r_D = 0$, i.e. only asymmetric divisions occur. Concerning the D-cells, in the simplest model version, cells of this type are eventually lost with the rate γ, $D \xrightarrow{\gamma} \emptyset$, in which $\emptyset$ corresponds to death, shedding, or emigration of D-cells. Other versions may include

the possibility of limited proliferation of D-cells and a direct differentiation of S-cells, where stem cells differentiate with the rate ω.

To maintain a homeostatic condition, the number of cells must stay constant on average. Thus, deterministic models based on Ordinary Differential Equations (ODEs) for the average cell numbers commonly describe the cell population dynamics. Following this approach, which is usual in modelling biological processes such as chemical reactions [Baker, 2017], the stochastic cells' fate choices are not modelled in detail but averaged to catch the global behaviour of the population.

Considering the stochastic model (1.1) in its simplest version and assuming constant rates, the system of ODEs describing the average number of cells, $\bar{n}_S$ and $\bar{n}_D$, respectively of type S and D, is

$$\begin{cases} \dfrac{d\bar{n}_S}{dt} = \lambda(r_S - r_D)\bar{n}_S \\ \dfrac{d\bar{n}_D}{dt} = \lambda(1 + r_D - r_S)\bar{n}_S - \gamma\bar{n}_D \end{cases} . \tag{1.2}$$

From the first equations, it is clear that, on average, the number of S-cells remains constant when $r_S = r_D$[1]. This means that stem cells following the PA strategy must regulate symmetric self-renewal and differentiation probabilities to be precisely equal, whereas this is trivially assured for the IA model. Assuming that a cell of type D is eventually lost with rate γ, the average total number of D-cells stabilises around the homeostatic value $\bar{n}_D^* = (\lambda/\gamma)\,\bar{n}_S$. This value uniquely depends on the number of stem cells, $\bar{n}_S$ which equals the initial number of stem cells $\bar{n}_{S,0} = \bar{n}_S(t = 0)$. Thus, the (Lyapunov stable) stationary state of total cell numbers $\bar{n} = \bar{n}_S + \bar{n}_D$ is given by

$$\bar{n}^* = \left(1 + \frac{\lambda}{\gamma}\right)\bar{n}_{S,0}. \tag{1.3}$$

Based on Equation (1.3), the process rates λ and γ determine the proportion of type D cells with respect to type S ones. Crucially, there is no difference at tissue level between the IA and PA models, so these self-renewal strategies cannot be distinguished by a deterministic model and by comparison with tissue cell population data only. However, only for the IA model, is the number of stem cells *strictly conserved*, meaning that there is no gain or loss of stem cells.

In general, the applicability of the deterministic approach is widely exploited to model more complex systems (as in the case of several cell types), non-homeostatic scenarios (like cancer stem cells), and internal regulation mechanisms [Ganguly and Puri, 2006, Johnston et al., 2007, Smallbone and Corfe, 2014, Situ and Lei, 2017].

[1] Steady-state where $\bar{n}_S = 0$ is not biologically relevant.

1.2 Cell clonal dynamics

A way to assess self-renewal strategies experimentally is via genetic cell lineage tracing [Kretzschmar and Watt, 2012, Blanpain and Simons, 2013] (details of the experimental background will be provided in Section 1.3.1). By marking single cells with an inheritable genetic marker, each cell's progeny, called a *clone*, which retain that marker, can be traced. The number of cells per clone, also called the *clone size*, is measured, and the statistical frequency distribution of clone sizes, i.e. the *clone size distribution*, is determined. To test the cell fate choice models on that data, one evaluates the models with a single cell as the initial condition and samples the outcome in terms of the final cell numbers, the size of a clone. Importantly, the clonal statistics cannot be based on a deterministic approximation which only describes the average cell numbers without modelling the details of the stochastic process. Hence, in the following sections, we review the typical mathematical modelling of clonal dynamics and the standard Bayesian fitting methodology for estimating the cell fate model parameters.

1.2.1 Mathematical modelling

While population modelling cannot discern between IA and PA self-renewing strategies, as shown in Section 1.1.1, a distinction is evident when we look at the dynamics of single cells and study the clone size distribution, that is, the distribution of its progeny. For the IA model, the number of S-cells is strictly constant, and thus the joint probability distribution, $P(n_S, n_D)$, of both the number of S- and D-cells, respectively indicated as n_S and n_D, is fully determined by the distribution of D-cells, $P(n_D)$. In the basic version of the stochastic model (1.1), where $D \xrightarrow{\gamma} \emptyset$, the probability $P(n_D)$ given a single initial cell of type S is the solution of

$$\frac{dP(n_D)}{dt} = \lambda P(n_D - 1) + \gamma(n_D + 1)P(n_D + 1) - (\lambda + \gamma n_D)\, P(n_D)\,. \tag{1.4}$$

This differential equation, also called the *master equation*[2] [Baker, 2017], corresponds to a simple *birth-and-death* process for which the distribution is Poissonian with mean λ/γ, [Van Kampen, 1981].

Considering the PA model, the master equation is instead given by

$$\begin{aligned}\frac{dP(n_S, n_D)}{dt} =& \lambda\,(r(n_S - 1)P(n_S - 1, n_D) + (1 - 2r)n_S P(n_S, n_D - 1) \\ &+ r(n_S + 1)P(n_S + 1, n_D - 2)) \\ &+ \gamma(n_D + 1)P(n_S, n_D + 1) - (\lambda n_S + \gamma n_D)\, P(n_S, n_D)\,.\end{aligned} \tag{1.5}$$

[2] In general, the master equation is a system of ODEs describing the time evolution of the probability distribution.

In [Antal and Krapivsky, 2010], an exact solution for the distribution of total cell numbers $n = n_S + n_D$ is found when $\lambda = \gamma$ and $r = 1/4$. In general, for different values of λ, γ and r, the long-term distribution is shown to be always Exponential.

Thus, the IA and PA models predict, respectively, a Poisson and an Exponential clone size distribution for long times. These two distributions are fundamentally different, meaning that the self-renewing strategy can easily be distinguished by the shape of the distribution of the clonal data. A series of lineage-tracing experiments confirmed that exponential clone size distributions prevail for most mouse tissues, supporting the PA strategy [Clayton et al., 2007, Lopez-Garcia et al., 2010, Simons and Clevers, 2011b, Doupé et al., 2012, Klein and Simons, 2011].

Another significant feature of the PA model is that the average clone size increases over time, seeming, at first sight, not representative of a homeostatic condition. However, this is consistent with tissues in which a fraction of cell divisions is symmetric. In that case, some clones shrink and eventually die, and others grow, keeping the size of the tissue constant on average. Thus whilst the average size of all the clones, survived and extinct, remains constant, the mean of the surviving clones, which is measured[3], grows. In this context, the *extinction probability* gives a measure of the ratio between the extinct over the total clone number. Its estimation is analytically possible only in some simple models as *single-type* branching process [Haccou et al., 2005]. A generalisation of this approach for generic *multi-type* processes, which describe proliferation and differentiation if there is more than one proliferating cell state, increases the complexity of the model considerably, as shown in [Hautphenne et al., 2013], [Hautphenne, 2015].

Although computationally expensive, numerical simulations of stochastic processes are often used for assessing complex cell fate systems. For example, this is the case when analytical solutions to describe the experimental data are unfeasible, such as multi-type branching processes or non-Markovian processes [Andrews et al., 2009, Paździorek, 2014, Aguilera et al., 2017, Kostiou et al., 2021, Rompolas et al., 2016]. A commonly used method is the Gillespie algorithm, [Gillespie, 1977]. This approach, part of the family of Monte Carlo methods, allows for the estimation of the full clone size distribution, including extinction, [Alcolea et al., 2014, Greulich and Simons, 2016].

1.2.2 Bayesian inference for model fitting

The model fitting aims at finding a set of parameters of the mathematical model that describe the available experimental data. A standard methodology for fitting lineage

[3] A measure of the number of extinct clones might be possible with live imaging techniques (see Section 1.3.1), but this is not always practically feasible.

tracing data based on Bayesian inference is used in several works, such as [Doupé et al., 2012, Alcolea et al., 2014, Frede et al., 2016].

As shown in [Box et al., 1992], the main principle of the Bayesian inference builds on the estimation of the posterior probability, called hereafter the *posterior*. The posterior is a measure of the certainty of a model and a particular set of model parameters, θ, being representative of a set of observed data, $\mathcal{D}$ (i.e. the experiments). The Bayes' theorem gives the posterior probability as

$$P(\theta|\mathcal{D}) = \frac{\mathcal{L}(\mathcal{D}|\theta)}{P(\mathcal{D})} P(\theta), \tag{1.6}$$

in which $\mathcal{L}(\mathcal{D}|\theta)$ is the likelihood function, representing the probability of the observed data given the parameters, $P(\mathcal{D})$ is a normalisation factor and $P(\theta)$ is the prior probability, called hereafter the *prior*, representing known information about the parameters before the data. Based on this approach, the best fit is given by the set of parameters θ^* that maximise the posterior.

A key aspect of Bayesian inference is the proper choice of the prior. The prior has to be dominated by the likelihood function to increase the parameters knowledge by the data. The opposite case is an indication that the observed data are not sufficient to improve the knowledge of the model parameters, resulting in a posterior not so different from the prior. Therefore, when no other information about the model parameters is knownthe prior is just a uniform distribution in the interval in which the likelihood is appreciable (maximum entropy principle, [Cox, 2007]). Instead, specific a priori knowledge of the parameters requires a proper selection of the prior.

Considering now the lineage tracing data fitting, since each observed clone is statistically independent, the formulation of the likelihood is based on a multinomial distribution with a countable number of outcomes [Doupé et al., 2012, Frede et al., 2016]. A generalisation for two-dimensional clonal size distribution is used in [Alcolea et al., 2014]. Although being a commonly used approach, there are practical complexities related to its implementation. Among others, the noise in the numerical estimation of the clonal statistics might result in an erroneous value of the likelihood, and the large variability of the posterior with potentially multiple local minima often requires a large sampling size. An efficient implementation is found, for example, in [Doupé et al., 2012], where the analytical solution of the clone size distribution replaces its estimation based on numerical simulations. Instead, in [Kostiou et al., 2020], the authors propose the use of a Sequential Monte Carlo (SMC) technique coupled with an Approximate Bayesian Computation (ABC) methodology[4]. However, both these approaches are tailored for specific cell fate models of the type (1.1), and their generalisation for more complex cell fates might not be straightforward.

[4]ABC substitutes the classical likelihood computation with a proper distance metrics, as the inter-quantile distances between distributions, or the Kolmogorov–Smirnov (KS) statistic.

1.3 Experimental background

In the following sections, we review two standard experimental methods to assess cell fate dynamics, lineage tracing and single-cell RNA-sequencing, the combination of which is proving to be a promising strategy in the study of stem cells [Kester and van Oudenaarden, 2018].

1.3.1 Lineage tracing

Lineage tracing is an experimental technique aimed at identifying the progeny of a single or a group of cells. Different methods exist, but they all have in common the idea of studying, at different time points, cells that correspond to the progeny of some initially labelled cells. An example for epidermal cells, taken from [Clayton et al., 2007], is shown in Figure 1.2. The design of each specific experiment, which includes the choice of an experimental procedure, defines the cell labelling method, how many cells and which ones are initially labelled. In [Ya-Chieh, 2015], there is a review of different methods, their advantages and limitations, and some examples. Among the key points highlighted in this work, there is the selection of the labelling approach. First of all, knowing the initially marked cells is the basis for studying the time evolution of their progeny. Besides, the cells and their progeny must retain the marker without spreading it over other cells. Importantly, the marker must not affect the cell behaviour.

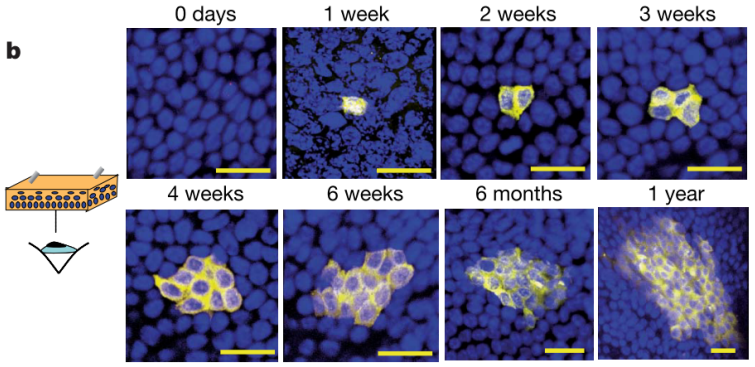


FIGURE 1.2: Example of in-vivo lineage tracing of epidermal progenitor cells showing the time evolution of a clone (cells in yellow), [Clayton et al., 2007]. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature, A single type of progenitor cell maintains normal epidermis, E. Clayton et al., Copyright (2007).

Based on this experimental procedure, a marked clone snapshots the cell proliferation dynamics at single-cell resolution. The evolution of the clone size at different times after labelling gives indirect information about differentiation. Thus, when dealing with *clonal dynamics* assays (in contrast to population-based assays), the challenging

aspect is the initial labelling of the cells. In this case, the initial clone labelling must be well distributed to ensure that single spare cells are marked and that, after some time, different clones do not merge.

One approach to in-vivo cell labelling makes use of genetic markers, also called *reporters* [Debnath et al., 2010], that enable the production of fluorescent proteins that colour the cell when they are expressed. The Cre-lox technology, in which transgenic animal models are used, is a widely adopted method [Sauer, 1998, Kretzschmar and Watt, 2012]. The principle is based on a drug-inducible protein production controlled by specific cell promoters (regulatory sites on DNA that activate or suppress protein production by binding other regulatory molecules) that enable the expression of some colour reporter. A scheme of this process in mammary gland cells is shown in Figure 1.3, taken from [Elias et al., 2017]. In this case, cells containing the CreERT2-IRES-nLacZ cassette (a mobile part of DNA [Hall and Collis, 1995]) express a red tomato reporter initially; the injection of a drug, for example, tamoxifen, TAM, removes the STOP cassette at the site of the colour reporter, switching the cell to the production of a green fluorescent protein. The expression of the reporter is irreversible, and the progeny of a marked cell inherits the marker as well. Thus, this approach overcomes the limitation given by the dilution of the marker in subsequent rounds of cell division, common in other methods. Recent techniques also allow the multi-colour labelling of the clones, as shown in [Livet et al., 2007]. Finally, live imaging techniques that track the cells in situ and real-time are worth mentioning [Ritsma et al., 2014]. The main problems of this type of method are mainly related to the actual complexity and invasiveness of the experimental procedure, e.g. making tissue transparent for imaging, fixing animals for long periods alive, which also have ethical issues.

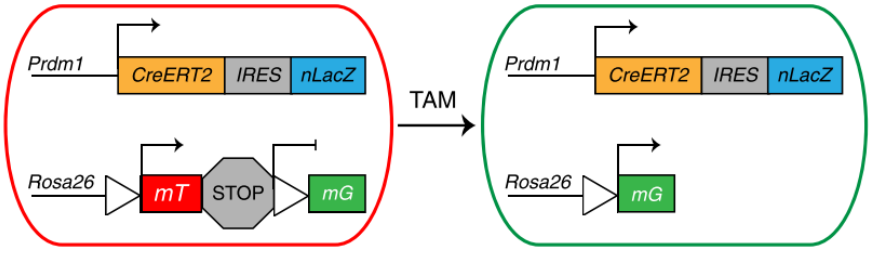


FIGURE 1.3: Sketch of the Cre-recombinase process reprinted from [Elias et al., 2017], Copyright S. Elias et al. (2017), this work is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). Cells containing the CreERT2-IRES-nLacZ cassette express initially a red tomato (mT) reporter (left); the injection of tamoxifen (TAM) removes the STOP cassette at the site of the colour reporter, switching the cell to the production of a green fluorescent protein (mG).

Several approaches use lineage tracing data to map the cell fate in healthy tissues or study early cancer development. In [Elias et al., 2017], for instance, a subpopulation of

mammary gland stem cells are analysed to study their impact on the development of the gland and tissue homeostasis. In [Clayton et al., 2007] and [Rompolas et al., 2016], models for the homeostasis of the epidermis based on the experimental data are proposed. Studies for the gastric epithelial homeostasis are [Barker et al., 2010], and [Leushacke et al., 2013]. In [Frede et al., 2016] and [Alcolea et al., 2014] instead, the growth of oesophageal tumours and hyper-proliferating mutant cells are assessed via lineage tracing experiments. From these examples, it is clear that a statistical approach to the clonal size data, as done, for instance, in [Clayton et al., 2007] and [Alcolea et al., 2014], is essential to provide quantitative information about the cell fate dynamics.

1.3.2 Single-cell RNA-sequencing

RNA-sequencing is a recent experimental method based on the so-called next-generation sequencing techniques. Its goal is to measure the gene expression of groups of cells by quantifying its RNA content [Chu and Corey, 2012],[Wang et al., 2009]. Among the advantages of this method there are accuracy, high resolution and low cost. The single-cell RNA-sequencing (scRNA-seq) applies the same approach at single-cell resolution and is primarily used to infer the cell identity [Treutlein et al., 2014].

In addition to the technical challenges of the method not discussed here (from the single cells isolation to the RNA amplification and measuring), the use of this technique implies working with a high amount of data that have to be stored and analysed [Kolodziejczyk et al., 2015]. The data, downstream of a bioinformatic data processing step[5], consist of the expression levels of a large number of genes in each cell. The data processing identifies the most differently expressed genes, explores the potential correlations and distinguishes cell clusters. This process is even more challenging, considering the noisy environment. Notably, the underlying assumption in this methodology is that cells close to each other in the gene expression space share the same identity. In this frame, clustering and visualisation are crucial for the analyst in post-processing and interpreting the data. Common scRNA-seq data visualisation methods are the gene expression heatmap [Wilkinson and Friendly, 2009] and t-distributed Stochastic Neighbor Embedding (t-SNE) plot [van der Maaten and Hinton, 2008].

However, being an emerging method, a standardised procedure for the scRNA-seq analysis is not yet available. For example, in [Treutlein et al., 2014], principal component analysis and unsupervised hierarchical clustering are employed to classify the epithelial cell populations in the distal lung. Other clustering methods, such as k-means and k-medoids clustering, are used in [Kim et al., 2016] and [Grün et al.,

[5]More specifically, raw data are the number of RNA fragments in each cell. Bioinformatic tools map these fragments with known genes to estimate the mRNA concentration in the cells.

2016]. A completely different approach to RNA-seq data analysis, proposed in [Arai et al., 2020], uses machine learning techniques to determine cells' identity. The idea is to train the model on a known database (e.g. based on cell types markers) and then fit the scRNA-seq output.

Further uses of the single-cell transcriptome data are the derivation of the lineage tree [Grün et al., 2016], the developmental lineage relationships [Pal et al., 2017], and the study of the dynamics of the differentiation [Bach et al., 2017].

1.4 Study Case: the mouse mammary gland

The mammary gland is an interesting case of study as it is an organ that reaches full development after birth, [Inman et al., 2015, Visvader and Stingl, 2014]. In fact, in female individuals during puberty, a branching structure develops through the mammary fat pad forming ducts. These ducts are composed of two layers of cells, called *epithelial* cells: the inner layer is composed of *luminal* cells, the outer one of myoepithelial cells, also called *basal* cells. Besides, during adult life, mammary glands experience changes in their structure and functionality due to pregnancy, lactation, and involution cycles, as sketched in Figure 1.4, which is taken from [Visvader and Stingl, 2014]. During pregnancy and lactation, *alveolar* cells also appear within the luminal compartment (i.e. cells aimed at producing milk). All these features indicate that there are pools of (quiescent) *mammary stem cells* (MaSCs) and progenitors necessary to develop and change the structure of the gland and to sustain homeostasis. Interestingly, homeostasis is a condition encountered whilst the adult maintains virginity (i.e. before the first pregnancy) and during involution stages (i.e. between the end of lactation and the following pregnancy).

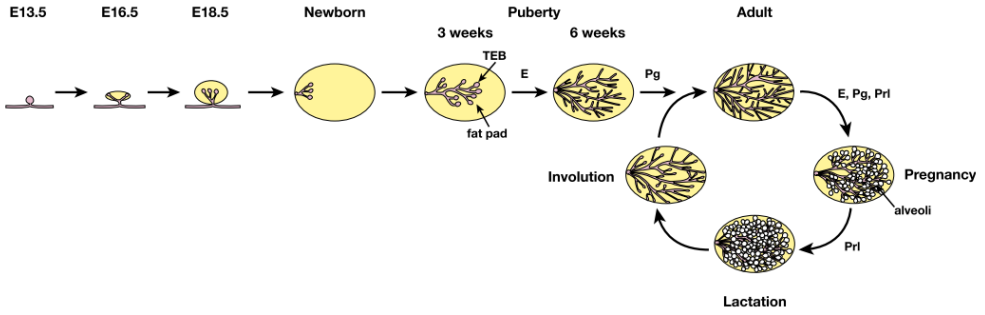


FIGURE 1.4: Developmental stages of the mouse mammary gland, from the embryo (E) to adulthood. Reprinted from [Visvader and Stingl, 2014], Copyright 2014 Visvader and Stingl; Published by Cold Spring Harbor Laboratory Press. This work is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), http://creativecommons.org/licenses/by/4.0/. In adulthood, the tissue undertakes cycles of pregnancy, lactation and involution. Homeostasis is only encountered before the first pregnancy and later, between the involution stage and the successive pregnancy.

Given the complexity of this gland, there are significant discrepancies across studies on cell type markers and, crucially, on cell lineage and hierarchy [Inman et al., 2015]. In works such as [Davis et al., 2016], lineage-tracing experiments suggest that MaSC/progenitors in adulthood are unipotent, meaning that they only contribute to the progeny exclusively of the basal or the luminal compartment. Nevertheless, in this work, it is also stated that rare quiescent bipotent embryonic MaSCs that are not initially labelled may exist (i.e. giving rise to both the luminal and the basal lineages). In contrast, lineage tracing experiment results shown in [Rios et al., 2016] demonstrate that bipotent MaSCs exist in adulthood. Besides, in a recent scRNA-seq study, [Pal et al., 2017], a mixed-lineage basal-like pool of cells has been identified: these cells, part of the basal compartment, are believed to be transient population before committing to the luminal lineage. The cell hierarchy proposed in this work is shown in Figure 1.5. A detailed review of this work and others based on scRNA-sequencing is provided in the next section.

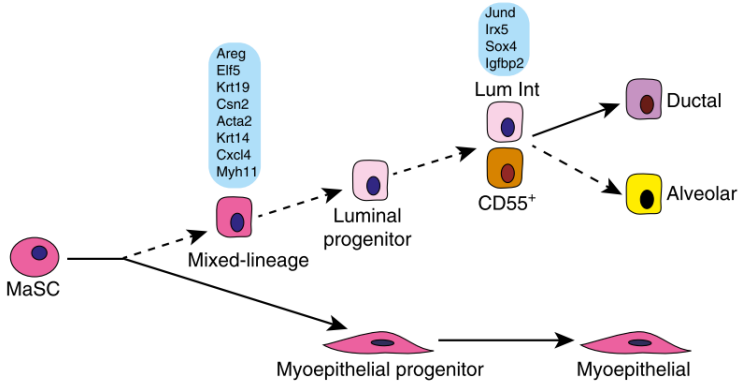


FIGURE 1.5: Mammary gland cells hierarchy proposed in [Pal et al., 2017]. Reprinted from [Pal et al., 2017], Copyright B. Pal et al. (2017), this work is licensed under a Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`). Mammary Stem Cell (MaSC), myoepithelial and mixed-lineage cells are part of the basal compartment; luminal, ductal and alveolar cells form the luminal compartment.

1.4.1 Single-cell RNA-sequencing literature review

Four pieces of work based on the scRNA-seq of the mouse mammary gland were recently published. The main features of the data are summarised in Table 1.1. We observe that the four works differ for the experimental setup (e.g. platform[6] and type of raw data[7]). Additionally, from the review of the methods used for the statistical

[6] Two commonly used platforms are the 10X Genomics Chromium (10X) and the C1 Fluidigm (C1) [Valihrach et al., 2018].

[7] Raw data can be based on *Read Counts* (RC); normalised and filtered RC (NRC); *Fragments Per Kilobase Million* (FPKM), which is a normalised RC that takes into account the length of the genes; and *Unique molecular identifiers* (UMI), which is based on an innovative experimental method aimed at reducing some technical errors and biases in the quantification of the RNA content.

analysis, it is clear that the approach to the analysis of the data varies in each case (e.g. type of normalisation, clustering method, pseudotime[8]). Thus, whilst we leave for later (see Section 5.2.1) a consistent comparison of these data, here, we briefly recall the key finding of each work.

Data	Platform	Type	Stage	Age/time		Cells
[Bach et al., 2017] (GSE 106273)	10X	UMI	Nulliparous	8	week	4223
			Pregnant	14	day	5826
			Lactation	6	day	9319
			Involution	11	day	5642
[Pal et al., 2017][9] (GSE 103275)	10X	RC	Puberty	5	week	5387
			Adult Virgin	10	week	3308
	C1	RC	Newborn	2	week	117
			Puberty	5	week	181
			Adult Virgin	10	week	162
			Pregnant	12.5	day	99
[Sun et al., 2018][10]	C1	FPKM	Adult Virgin	3-4	month	88
			Pregnant	12	day	151
[Giraddi et al., 2018][11] (GSE 111113)	10X	NRC	Embryonic (1)	16	day	690
			Embryonic (2)	18	day	1047
			Newborn	4	day	849
			Adult Virgin	10-16	week	3838

TABLE 1.1: Summary of the main features of the scRNA-seq data available. Rows in grey correspond to the samples that will be examined in Section 5.2.1.

1. In [Bach et al., 2017], it is shown that: i) there is not always a unique marker that uniquely defines a cell identity; ii) the luminal compartment is seen as a continuum in which cells in an intermediate state exist, i.e. between progenitor and mature; iii) pregnancy and lactation have an impact on the luminal progenitors with a shift towards the alveolar phenotype.

2. Key results of the analysis reported in [Pal et al., 2017] are i) a shift in gene expression from a homogeneous basal-like to distinct lineages is detected from pre-puberty to adulthood; ii) *Cd55* is identified as an early progenitor marker; iii) a rare basal mixed-lineage cluster is detected showing expression of both basal and luminal genes, and an intermediate luminal stage is identified in both the adult and puberty samples; iv) the proposed cells lineage, shown in Figure 1.5, is characterised by luminal cells descending from stem cells from the basal compartment.

3. The primary outcomes in [Sun et al., 2018] are summarised as: i) key markers characterising cell clusters are defined; ii) *Cdh5* is identified as a marker for a rare cell subpopulation within a basal compartment which is considered as quiescent Mammary Stem Cell (MaSC); iii) mammary epithelial cells hierarchy is

[8]This analysis is aimed at the reconstruction of cells differentiation trajectories, [Trapnell et al., 2014].

inferred from pseudotime analysis, based on which MaSC, part of the basal compartment, is at the apex of the lineage hierarchy.

4. In [Giraddi et al., 2018], key results are: i); foetal Mammary Stem Cells (fMaSC) are identified; ii) fMaSC define distinct lineages, i.e. basal and luminal, which are separated in adulthood; iii) cell types and signatures are identified in the early mammary epithelial development.

The main discrepancies and open questions include a) the existence of a mixed basal-luminal type of cells; b) the existence of luminal intermediate cells; d) the existence of distinct lineages for luminal and basal cells; and e) the identification of common markers for the definition of the cells' identity[12].

1.5 Research aim and objectives

Provided the complexity and controversy of the research area previously described, this project aims at developing a mathematical framework to study cell fate dynamics, i.e. the cell proliferation and differentiation, to improve the understanding of the homeostasis mechanism in adult renewing tissues. From the biological standpoint, this translates into determining the cell differentiation, proliferation and division outcomes that describe a set of experimental data. Challenges include a) the vast search space, in which there are potentially infinite possible models compatible with data; b) the problem of model overfitting, in which the fitting of noisy data results in an overly complex model that does not represent reality. Therefore, before following a Bayesian inference approach for model fitting, we study these dynamics from a mathematical perspective by theoretical and numerical means to substantially restrict the number of candidate models. As a study case, we will focus on the mouse mammary gland.

For achieving the aim above described, the research project is articulated around the following main objectives.

Obj. 1 To exclude lineage hierarchies not compatible with homeostasis in adult renewing tissues by deriving generic rules that constrain the structure of the dynamical model.

Obj. 2 To identify conditions under which a regulation mechanism gives homeostasis robustness to perturbations and stochastic fluctuations.

[12]Only a few markers are mentioned in all the four articles, e.g. *Krt14* and *Acta2* for the basal cells, and *Krt18* for the luminal cells.

Obj. 3 To distinguish classes of cell fate models characterised by universal features that can be qualitatively compared with experimental data.

Obj. 4 To validate and apply the developed theoretical framework using a specific study case for which experimental data might be available in the future.

1.6 Research approach, innovative contribution and thesis outline

This work is organised into four tasks, as outlined in the following sections and schematised in Figure 1.6. Tasks 1-3 cover the generic theoretical modelling of cell dynamics in homeostasis. Task 4 is specifically related to the study case. Each task is associated with an objective set in Section 1.5.

The results and findings of the research have been published in three journal papers, one as the first author. In particular, [Parigini and Greulich, 2020] was awarded a Doctoral College Research award for the School of Mathematical Sciences, University of Southampton, June 2021. Another journal paper is in preparation (Regulation of homeostasis via crowding feedback: robustness and quasi-dedifferentiation). Ideally, this work will be submitted in one of the following journals: PLOS Computational Biology, Journal of Mathematical Biology, Physical Biology, Bulletin of Mathematical Biology. Lastly, I presented my work in two poster sessions and a seminar.

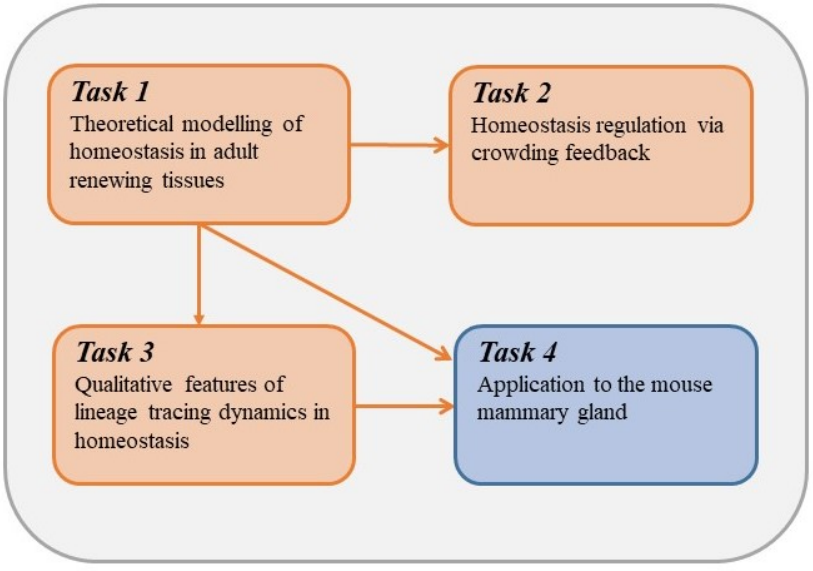


FIGURE 1.6: Organisation of the work. Tasks 1-3 (orange) cover the generic theoretical modelling of cell dynamics in homeostasis; Task 4 (blue) is related to the study case.

1.6.1 Task 1. Theoretical modelling of homeostasis in adult renewing tissues.

Objective 1./Chapter 2.

Approach. Generic cell fate dynamics are modelled as a generalised multi-type branching process for which the classical deterministic formalism is combined with graph theory. Such a generic cell system is seen as a network used to gain insight into the cell fate dynamics' features and determine candidate models for tissue homeostasis. Numerical simulations illustrate the theory developed.

Contribution. A conceptual mathematical model is defined to represent arbitrary cell fate dynamics. Based on that, the conditions required to achieve homeostasis are mathematically derived and then translated into the biological context, proposing a definition of adult stem cells. Such definition strictly relates to a specific hierarchical ordering of the cell types, which is the only one compatible with homeostasis. Consequently, this result significantly simplifies the model fitting problem since all the non-homeostatic cell dynamics can be a priori excluded.
This part of the work is partially based on results published in [Greulich et al., 2019] and [Greulich et al., 2021], which I co-authored. My contribution to [Greulich et al., 2019] is in supporting and verifying the theory developed and its mathematical proof with numerical analyses. Concerning [Greulich et al., 2021], I mainly worked on the mathematical analysis of the crowding feedback modelling (see next section).
Chapter 2 placed these results in a broader context and integrated them with an intuitive view of the developed theory. My analyses were also translated into specific numerical examples practical to illustrate the cell dynamics in different scenarios. These examples were presented in [Greulich et al., 2021] and, more extensively, in this thesis.

1.6.2 Task 2. Homeostasis regulation via crowding feedback.

Objective 2./Chapter 3.

Approach. Starting from Task 1 research outcomes, the cell fate dynamical models are extended to include homeostasis regulation via crowding feedback. From a mathematical standpoint, both theoretical and numerical, a stability analysis of the homeostatic condition is performed. Successively, realistic scenarios in a biological context are used for assessing the robustness of such homeostasis regulation.

Contribution. Under reasonable biological assumptions, cell dynamics regulated via feedback is shown to remain confined around a (dynamic) homeostatic condition. This result is published in [Greulich et al., 2021], where my contribution is the identification of the dynamic homeostatic condition, including its numerical

verification and the illustrative examples reporting.
In Chapter 3, this concept is further assessed, determining a necessary condition and a sufficient one that guarantee a homeostatic state in a strict sense, that is, an asymptotically stable steady state. Homeostasis regulated via crowding feedback is also demonstrated to be robust to perturbations and failures, explaining why the initial stage of cancer development is possibly associated with successive cell mutations. Finally, the stem cell concept derived under Task 1, is generalised showing that stemness naturally arises from the interaction with the environment rather than being an intrinsic property of cells. Based on purely theoretical considerations, the quasi-dedifferentiation mechanism is proposed as an alternative response to the experimentally observed cell dedifferentiation process, activated in case of tissue damage.

1.6.3 Task 3. Qualitative features of lineage tracing dynamics in homeostasis.

Objective 3./Chapter 4.

Approach. The dynamics of lineage tracing in cell fate models, restricting to those fulfilling homeostasis requirements derived under Task 1, are studied. The idea is to derive features of such dynamics that could simplify the definition of the cell fate model via qualitative comparison with experimental data. That translates into stem cell types identification via scRNA-sequencing data and self-renewing strategy via clonal data. The ODE formalism, as done in Task 1, and analytical solutions and numerical simulations of the stochastic process are used for modelling lineage tracing dynamics.

Contribution. For an arbitrary cell fate model, the analytical formulation of the size of cell clusters is derived considering transcriptome data of both lineage-traced cells and tissue samples. The analysis shows that only disconnected stem cell types, defined according to Task 1, justify different clonal and tissue assay measures. A compartment model for studying the clonal statistics is then defined, leading to the generalisation of the common definition of *symmetric* and *asymmetric* divisions, which are the basis of stem cell self-renewing strategies. Successively, models of cell fate dynamics are categorised in two universality classes that predict, under asymptotic conditions, the same clone size distribution, i.e. Exponential or Normal. Building on this, simple rules for identifying the self-renewing pattern from qualitative features of lineage tracing of clones are derived. This analysis highlights the limitation for which models within the same class could not be distinguished only via clonal data in the asymptotic regime, a condition that, in any case, is not always fulfilled in real tissues.

1.6.4 Task 4. Application to the mouse mammary gland.

Objective 4./Chapter 5.

Approach. The research outcomes of Task 1 and Task 3 are combined and applied to a study case, the adult mouse mammary gland. The determination of the cell state network is solved first, followed by a parameter fitting via Bayesian inference methodology. The analysis builds on assumptions supported by literature and synthetic data inspired by actual lineage tracing experiments[13].

Contribution. From the analysis of scRNA-seq literature data for the mouse mammary gland, cell identities and possible relationships among them are identified. Based on this, a cell state network for the study case is defined, making assumptions whenever the literature data were insufficient or in disagreement. Additionally, from a consistent comparison of four published works, suggestions are given about areas where more dedicated experimental work is necessary to resolve open issues. The parameter fitting problem is solved first without imposing homeostasis, showing that although a good fitting of the data might be obtained, such cell fate models are not representative of homeostatic dynamics. Fittings exhibiting very different dynamical behaviour, yet equivalently good, are found when imposing homeostasis, proving the need for additional data. By enriching the dataset, the variability in the fitting parameters is reduced, and, in the case assessed, the self-renewing strategy is determined. Overall, the analysis validated the methodology developed and provided a clear path for future studies based on actual experimental data, thanks to which answers to biological questions can be given.

[13] Inputs from two lineage tracing experiments, under Dr Elias responsibility (Institute for Life Sciences, University of Southampton), were initially expected. Such data were not available when writing this thesis.

Chapter 2

Theoretical modelling of homeostasis in adult renewing tissues

The first objective set in Section 1.6 is to simplify the search for a mathematical model that fits the experimental data by restricting the candidate models to those describing homeostatic dynamics in adult renewing tissues. Therefore, in this chapter, we study cells' dynamics from a theoretical point of view, focusing on the generic features of such dynamics and determining the conditions fulfilled in a homeostatic tissue. For this purpose, starting from a branching process model, we first derive an approximation of the generic tissue dynamics using both dynamical system and graph theory. We then define a cell state network and propose a cell type definition and classification. We also determine the conditions for homeostasis, showing that, based on this framework, the network structure must follow strict rules, requiring self-renewing cells at the apex, and only there, of the lineage hierarchy. Experimental evidence of this feature was already available, but we show that this is the only possible way to achieve homeostasis from a mathematical standpoint. The presented analysis also includes numerical examples instrumental in illustrating the theory developed.

Part of this chapter builds on [Greulich et al., 2019] and [Greulich et al., 2021], which I co-authored (see details in Section 1.6.1).

This chapter is organised as follows: the description of the cells' dynamical model is provided in Section 2.1; homeostasis modelling is discussed in Section 2.2; conclusions are given in Section 2.3.

2.1 Cell dynamics modelling

In this work, we use a continuous-time multi-type branching process to describe the cell fate dynamics [Haccou et al., 2005]. The model is a continuous-time stochastic process characterised by a number m of possible cell states X_i, $i = 1, ..., m$. We define a cell *state* here as a group of cells sharing the same identity, e.g. showing common morphological properties, functional features, protein levels or mRNA expression [Morris, 2019]. In this view, *cell trajectories* consist of transitions between a discrete set of cell states. Although we take a discrete formalism here, the results we will derive also apply to a continuous conceptualisation since continuous processes can always be discretised in a topology-preserving way [Milnor, 2016].

Most generally, cells in a state X_i may be able to divide, producing daughter cells of any cell states X_j and X_k. When $i = j = k$, the cell division represents a simple cell duplication. Furthermore, any cell in state X_i may turn into another state X_j or may be lost, $\emptyset$, through emigration, shedding, or death. Hence, we can study the *cell fate dynamics* based on a generic cell fate model, written as

$$\text{cell division: } X_i \xrightarrow{\lambda_i r_i^{jk}} X_j + X_k \tag{2.1}$$

$$\text{cell state transition: } X_i \xrightarrow{\omega_{ij}} X_j \tag{2.2}$$

$$\text{cell loss: } X_i \xrightarrow{\gamma_i} \emptyset \tag{2.3}$$

for $i, j, k = 1, ..., m$ and in which the *kinetic parameters* λ_i, ω_{ij} and γ_i are the rates of division, transition to state X_j and the loss rates, of cells in state X_i. To determine the outcome of the cell division from the ith state, we also need to specify a set of parameters r_i^{jk} for $j, k = 1, ..., m$ based on which we define $r_i^j = \sum_{k=1}^m (r_i^{jk} + r_i^{kj})/2$ as the probability of having a daughter cell in state X_j. Note that $\sum_{j=1}^m r_i^j = 1$.

Notably, the events depicted in Equations (2.1)-(2.3) are not Markovian as the timing of events is not independent of each other and depends on their history. Cell division, differentiation and death are actually related to the cell cycle [Alberts et al., 2015]. However, neglecting any short-term dynamics and assuming that the dynamics' time scale is longer than the cell cycle's length, we approximate the above process as Markovian, in which the kinetic parameters represent the mean frequency of the events.

Lastly, we observe that cell fate is a highly stochastic process that could depend on the cell environment, for example, through spatial, cell-extrinsic regulation of cell fate [Simons and Clevers, 2011a]. We will assume that the kinetic parameters are constant for modelling homeostasis, which is related to a steady condition. This approximation is valid as long as we remain close enough to this homeostatic state. We will argue later, in Section 2.2.3, why this approximation is adequate to determine necessary

conditions for homeostasis, although it is not possible to address its stability in general without knowing the details of the dynamical system. Hence, we will extend this constant parameter model in Chapter 3, where we will assess a possible homeostasis regulation mechanism.

2.1.1 Deterministic approximation

To study homeostasis, a steady-state of the cell population dynamics, we focus on the dynamics of the mean number of cells. In practice, the stochastic fluctuations due to the random nature of each cell's fate are averaged, and what the cell ensemble does as a whole is studied.

For doing so, we start from the generic branching process (2.1)-(2.3), and indicate as $P(\boldsymbol{n})$ the probability of having $\boldsymbol{n} = \begin{pmatrix} n_1, & n_2, & \dots & n_m \end{pmatrix}$ cells, where m is the number of cell states. Based on this, the first moments $\bar{n}_i = \langle n_i \rangle$, where

$$\langle n_i \rangle = \sum_{\boldsymbol{n}=0}^{\infty} n_i P(\boldsymbol{n}) \text{ for } i = 1, ..., m\,, \tag{2.4}$$

describe the *mean* cell numbers. To compute $P(\boldsymbol{n})$, we need to solve the master equation, a set of ordinary differential equations associated with the generic branching process. However, given the complexity of the problem, in which there is an arbitrary number of cell states and cell state interactions[1], a closed form solution is not possible. Also, alternative approaches based on numerical simulation of the stochastic process would be only applicable to the specific cell fate model considered in the simulations.

Therefore, in analogy to the modelling of chemical reactions (based on the Law of Mass Action [McLean, 1938]), we can derive a system of ODEs that approximate $\bar{\boldsymbol{n}} = \begin{pmatrix} \bar{n}_1, & \bar{n}_2, & , ... \bar{n}_m \end{pmatrix}$ in the limit of large $\bar{\boldsymbol{n}}$ [Baker, 2017][2]. This approach is applicable to our case since the cell numbers in a tissue are generally very high (for example, in the epidermis there are about 75000 cells/mm^2 [Bauer et al., 2001]). This results in

$$\frac{d}{dt}\bar{n}_i = \sum_j \left(\lambda_j 2 r_j^i + \omega_{ji}\right) \bar{n}_j - (\lambda_i + \sum_j \omega_{ij} + \gamma_i)\bar{n}_i\,. \tag{2.5}$$

We recall that $r_i^j = \sum_k (r_i^{jk} + r_i^{kj})/2$ is the probability of having a daughter cell in state X_j produced upon division of a cell in state X_i.

Assuming, for now, a constant parameter model, we can further simplify the problem. In this case, the kinetic parameters do not change over time or as a function of $\bar{\boldsymbol{n}}$, and

[1] Considering the generic model (2.1)-(2.3) with m states, $P(\boldsymbol{n})$ is an m-dimensional multivariate distribution which depends on a large set of parameters. There might be up to m division rates (λ_i), m^3 probability parameters (r_i^{jk}), $m(m-1)$ transition rates (ω_{ij}), and m death rates (γ_i).

[2] According to [Baker, 2017], if low cell numbers are combined with second-order (or higher) terms, then the deterministic approximation deviates from the average of the stochastic process.

the system of ODEs becomes linear. A generalisation of this model will be discussed later, in Section 2.2.3, where the effects of non-linearities in the cell fate dynamics are addressed. Thus, the linear system of ODEs can be written more compactly in terms of the mean cell numbers vector, $\bar{\boldsymbol{n}} = (\bar{n}_1, \bar{n}_2, ..., \bar{n}_m)$, as

$$\frac{d}{dt}\bar{\boldsymbol{n}} = A\bar{\boldsymbol{n}}\,. \tag{2.6}$$

Here A is an $m \times m$ matrix

$$A = \begin{pmatrix} \kappa_{11} - \delta_1 & \kappa_{21} & \kappa_{31} & \cdots \\ \kappa_{12} & \kappa_{22} - \delta_2 & \kappa_{32} & \cdots \\ \kappa_{1m} & \kappa_{2m} & \cdots & \kappa_{mm} - \delta_m \end{pmatrix}, \tag{2.7}$$

in which we define the *total transition rate* as $\kappa_{ij} = \lambda_i 2r_i^j + \omega_{ij}$, combining all transitions from X_i to X_j by cell divisions and direct transitions, and the *local loss rate* as $\delta_i = \lambda_i + \sum_j \omega_{ij} + \gamma_i$.

2.1.2 Cell state network and cell type condensed network

The dynamics of the tissue following the branching process described by Equations (2.1)-(2.3) is described by the linear system of ODEs (2.6). Thus, the tissue behaviour depends only on the properties of the matrix A. In principle, the spectral properties of A determine the tissue's long-term behaviour [Åström and Murray, 2009], but explicitly applying spectral conditions is unwieldy and difficult to interpret biologically. Therefore, we will show here how, based on graph theory, some key features of the dynamics can be derived.

To this aim, we interpret this process as a network, and more specifically, as a directed graph [Bang-Jensen and Gutin, 2007], which gives a more intuitive view of the process. In this model, each graph's nodes correspond to a cell state, and a link from state X_i to X_j exists where a generalised transition, that is, through direct transition, cell division or both, is possible, meaning that $\kappa_{ij} > 0$. The matrix A can be written as the difference of two contributions, K and D, in which K is the matrix composed of the total transition rates, κ_{ij}, and D is a diagonal matrix with the total loss rates, δ_i, i.e. $A = K - D$. Based on this, K is the transpose of the adjacency matrix of the *cell state network*. The matrix K is square, and its rows represent the incoming vertices, i.e. produced cells, and its columns the outgoing ones, i.e. cells leaving. Each matrix element, κ_{ij}, corresponds to a (weighted) graph directed edge connecting two vertices, and its value denotes the link weight (the diagonal terms are self-links). In this view, directed paths of the network, which are sequences of directed links, represent allowed cell state trajectories. Importantly, the matrix K is non negative by definition, i.e. $\kappa_{ij} \geq 0$ for all i, j, and, consequently, A is a Metzler matrix in which the

off-diagonal elements are non-negative, i.e. $a_{ij} \geq 0$ for $i \neq j$. An illustrative example of such a cell state network is shown in Figure 2.1 on the left. Notably, more than one network may result in the same adjacency matrix, K.

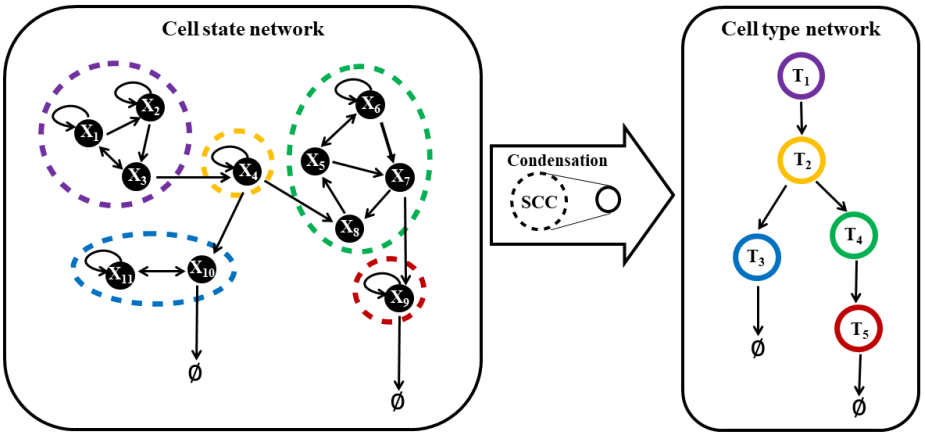


FIGURE 2.1: An illustrative example of a cell state network (left) and the corresponding cell type condensed network (right). In the cell state network, the cell states are represented as nodes and possible transitions between states, through direct transition or cell division, like links. The empty set symbol, $\emptyset$, represents cell loss (via death or emigration). The dashed circles denote the network's Strongly Connected Components (SCCs), each of them including states which are mutually reachable by directed paths. SCCs, representing cell types, correspond to the nodes in the condensed network, and a link between two SCCs exists only if any of their states are connected. This directed network does not have any cycles, and it has a natural hierarchical structure. More specifically, it admits an ordering of the nodes (cell types) $T_1, T_2, \ldots$ such that all transitions respect the ordering, that is, if there is a link, or a trajectory, from T_k to T_l then $k \leq l$. This figure is inspired by those presented in [Greulich et al., 2019, 2021].

Cell identity is commonly associated with the expression of surface markers or clustering from single-cell transcriptome data. From a different perspective, cells can also be classified based on their function in the tissue, which, in the context of cell lineages, is strictly related to their lineage potential. In this frame, we define a *cell type* as a group of cells that have the same lineage potential and vice versa. According to this definition, a cell type is composed of cell states that share the same outgoing trajectories, which means that states of the same cell type must be mutually reachable. From a mathematical point of view, this definition means that a cell type corresponds to a *Strongly Connected Component* (SCC) of the underlying cell state network. An SCC of a directed graph is composed of a set of vertices that any other vertex of the same SCC can reach. Thus, there is always a path connecting any pair of vertices of the SCC (in both directions).

Importantly, any directed network may be uniquely decomposed into SCCs by grouping all strongly connected nodes [Bollobás, 1998]. Based on this, we can

construct a second network, in which SCCs of the cell state network, corresponding to cell types, are the nodes, in the following indicated as T_i, for $i = 1, .., k$, where k is the number of SCCs. In such a network, two cell types are connected, i.e. there is a link between two nodes, if at least one generalised transition from states in one cell type to those in the other exists. Importantly, this connection has a unique direction; otherwise, the two SCCs would form a single SCC. The resulting network, known as the *condensed* network, does not contain cycles (i.e. directed paths from a node to itself) and is, therefore, hierarchical [Cormen et al., 2009]. Hence, we can order the cell types $T_1, T_2, \ldots$, in a way that if there is a trajectory from T_k to T_l, then $k \leq l$. Building on this, we say that T_k is *upstream* of T_l and T_l is *downstream* of T_k. For the illustrative cell state network example shown in Figure 2.1 on the left, the resulting condensation in the *cell type network*, which is representative of the cell lineage, is shown in the same figure on the right. Therefore, this means that cell types are necessarily ordered in a hierarchy, which is commonly observed. Finally, it is reasonable to consider only SCCs either with more than one node or one node with a self-link from a biological standpoint.

2.2 Homeostasis modelling

So far, we have considered the cell state and cell type networks without restricting the tissue's proliferative dynamics. Importantly, cell proliferation and removal must be finely balanced to ensure homeostasis. Thus, homeostasis imposes strong constraints on the dynamics. From a mathematical point of view, homeostasis represents a steady state of the cell population dynamics in which the number of cells of each type stays, on average, constant in the tissue. In this section, we will first determine the conditions required for achieving homeostasis and then assess how homeostasis restricts the possible hierarchy of cell lineages.

Starting from the system of ODEs given by Equation (2.6), we obtain the steady-state by imposing the condition $d\bar{\boldsymbol{n}}/dt = 0$. This corresponds to the solution, $\bar{\boldsymbol{n}}^*$, of the homogeneous linear system of equations

$$A\bar{\boldsymbol{n}} = 0. \tag{2.8}$$

The trivial solution, $\bar{\boldsymbol{n}}^* = 0$, is a solution of the above system, but it is not relevant in this context as we are interested in a steady-state characterised by positive average cell numbers. Therefore, we focus on non-trivial solutions which only exist in a marginally stable system (also known as neutral stability) [Franklin et al., 2001].

We recall that in an asymptotically stable system, the dynamics naturally restore the steady-state after a perturbation, and in an unstable system, any perturbation leads to an unlimited deviation from the steady-state. In contrast, in a marginally stable system, perturbations change the steady-state condition. Therefore, we can

distinguish three distinct long-term behaviours which only depend on the spectral properties of A [Åström and Murray, 2009]:

(a) **Unstable-growing.** If there are at least one or more eigenvalues with a positive real part, or if there is a zero eigenvalue with geometric multiplicity more than one, then the system is unstable, and the number of cells grows infinitely.

(b) **Stable-vanishing.**: If all the eigenvalues have a negative real part, the system is stable around the trivial solution, that is, the extinction. In this case, whatever the initial condition may be, the number of cells naturally decays to zero.

(c) **Marginally stable-homeostatic.** Marginal stability implies that at least one eigenvalue must have zero real part, the geometric multiplicity of eigenvalues with zero real part is equal to their algebraic multiplicity, and all the remaining eigenvalues must have a negative real part. If these conditions are met, then the cell numbers converge to a constant value which is not restored if the number of cells is perturbed. Perturbations instead have the effects of changing the steady-state condition.

Thus, for now, we define homeostasis in a less strict sense via a marginally stable condition, and we will assess later, in Chapter 3, that under realistic conditions (i.e. regulation via feedback), such a marginally stable state effectively turns stable around a non-trivial steady-state condition.

2.2.1 Condition for marginal stability

Considering the above definition of homeostasis, the analysis of the process from a graph theory point of view allows us to determine some constraints on the network structure required to achieve this condition. In particular, we recall from Section 2.1.2 that Equation (2.6) is linear and cooperative, i.e. the off-diagonal elements of matrix A are non-negative, and that the decomposition into the cell state network's SCCs yields an acyclic cell type condensed network that contains SCCs as nodes and directed links between them. Since the adjacency matrix of an SCC is an irreducible matrix [MacCluer, 2000], the block matrix K_i associated with each ith SCC is also irreducible. So is A_i, since matrix transposition and changing the diagonal terms do not affect irreducibility. Applying now a topological reordering of the graph vertices [Cormen et al., 2009], we can write A in a lower triangular block form

$$A = \begin{pmatrix} A_1 & 0 & 0 & 0 & \dots \\ C_{21} & A_2 & 0 & 0 & \dots \\ C_{31} & C_{32} & A_3 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \dots & \dots & \dots & \dots & A_k \end{pmatrix}, \tag{2.9}$$

in which the block C_{ij}, for $i, j = 1, ..., k$ and $j > i$, represents the connection between the ith and jth SCCs. Crucially, this partitioning implies that the spectrum of A corresponds to the union of the eigenvalues of A_i for $i = 1, ..., k$. In this view, the eigenvalues of the global system uniquely depend on those associated with each SCC. Thus, we will first analyse the spectral property of an isolated SCC and then show how the relations between SCCs also play an important role in determining homeostasis.

Focusing now on the ith SCC, we observe that the matrix A_i is an irreducible Metzler matrix and therefore the shifted matrix B_i, defined as $B_i = A_i + d_i I$, in which $d_i = \max(|\text{diag}(A_i)|)$ and I is the identity matrix, is an irreducible non-negative matrix. The Perron-Frobenius theorem [Meyer, 2000] states that for an irreducible non-negative matrix, such as B_i, the following statements hold:

(i) The largest eigenvalue, also called the *dominant eigenvalue*, is real and simple.

(ii) The right and left eigenvectors associated with the dominant eigenvalue, called *dominant eigenvectors*, are strictly positive, i.e. all their components are positive.

Focusing on (i), that means that the dominant eigenvalue of B_i, μ_{Bi}, is real with multiplicity one, and so is the dominant eigenvalue of A_i, which is $\mu_i = \mu_{Bi} - d_i$. Crucially, the value of μ_i determines the long-term dynamics of the isolated ith SCC, which result in an unstable-growing, stable-vanishing or marginally stable-homeostatic behaviour respectively if μ_i is positive, negative or zero. This result leads us to classify a SCC based on the value of its dominant eigenvalue, μ. In particular, we define a SCC as: *super-critical* if $\mu > 0$, *sub-critical* if $\mu < 0$ and *critical* if $\mu = 0$.

In the following, we derive necessary conditions for marginal stability when multiple SCCs are connected. We remark that, in [Greulich et al., 2019], the same conditions were obtained in a slightly different way proving they are not only necessary but also sufficient conditions. However, here, we provide a more intuitive proof. First of all, we consider a system with one or more super-critical SCCs. In this case, the tissue dynamics are unstable since there is at least one eigenvalue with a positive real part. Based on that, in a homeostatic system, all the SCCs must be critical or sub-critical. Assuming that all the SCCs are sub-critical, then the trivial steady-state is the only fixed point of the system, which implies that there must be at least one critical SCC. Thus, the above considerations lead us to derive the following conditions for homeostasis.

(l.i) There must not be not super-critical SCC(s).

(l.ii) There must be at least one critical SCC.

Connecting now all the SCCs and given that the matrix A is in the triangular form (2.9), the steady-state condition of the ith SCC, $\bar{\boldsymbol{n}}_i^*$, only depends on A_i and on the upstream SCCs, that is, all the jth SCC with $j < i$. The steady-state cell number, $\bar{\boldsymbol{n}}_i^*$, is the solution of

$$\sum_{j<i} C_{ij}\bar{\boldsymbol{n}}_j^* + A_i\bar{\boldsymbol{n}}_i = \mathbf{0}\,, \tag{2.10}$$

in which $\bar{\boldsymbol{n}}_j^*$ for $j < i$ is the steady state of the upstream SCCs. Assuming that the ith SCC is critical, then $\mu_i = 0$ and A_i is not invertible since the determinant of A_i is zero. Multiplying Equation (2.10) by the left dominant eigenvector, $\boldsymbol{v}_i$, we obtain a scalar relation which is

$$\boldsymbol{v}_i \sum_{j<i} C_{ij}\bar{\boldsymbol{n}}_j^* + \mu_i \boldsymbol{v}_i \cdot \bar{\boldsymbol{n}}_i = \boldsymbol{v}_i \sum_{j<i} C_{ij}\bar{\boldsymbol{n}}_j^* = 0\,. \tag{2.11}$$

We recall now that C_{ij} for $i = 1, ...k$ and all $j < i$, are non-negative block matrices, the components $\bar{\boldsymbol{n}}_j^*$ are also non-negative by definition and the dominant eigenvector has all positive components (Perron-Frobenius theorem, statement (ii)). Thus, unless C_{ij} is the null matrix, that is, there are no links between the jth and the ith SCCs, then only the trivial steady-state, $\bar{\boldsymbol{n}}_j^* = \mathbf{0}$, fulfils Equation (2.11). Thus, any jth SCC with $j < i$ is either disconnected from the ith SCC or it is trivial.

The crucial implication of this result is that if a critical SCC is downstream of other SCCs, then all the upstream SCCs must be trivial. We can further observe that a critical SCC is marginally stable, and this implies that the dynamics do not restore the initial steady-state if perturbed. In other words, any perturbation in the cell numbers applied to a critical trivial SCC results in a non-trivial critical SCC since this SCC will never go back to being trivial again. Globally, this implies that Equation (2.11) is not fulfilled anymore, meaning that a steady-state does not exist, and it will never be approached. Hence, any trivial SCC upstream of a critical SCC must be sub-critical. This result also means that more critical SCCs can coexist, as long as one is not upstream of the other. Therefore, there must not be any path connecting critical SCCs[3]. Based on the above considerations, we can derive two more requirements for homeostasis, stated below.

(l.iii) If there is any SCC upstream of a critical SCC, it must be trivial.

(l.iv) If there are multiple critical SCCs, there must not be any path connecting them.

The derived necessary (and sufficient [Greulich et al., 2019]) conditions for marginal stability of the dynamical system, (l.i)-(l.iv), are graphically shown in Figure 2.2, where different architectures of the cell type condensed network are classified based on their compatibility with homeostasis requirements. Each circle corresponds to an SCC, which is coloured according to its type, i.e. critical, sub-critical and super-critical; dashed faded lines indicate the trivial SCCs. Whilst networks **a**-**c** (left column) are not

[3]In [Greulich et al., 2019], the proof of this statement is based instead on considerations about the geometric and algebraic multiplicity of the dominant eigenvalue.

compatible with homeostasis, in networks **d-f** all the requirements for homeostasis are met. Focusing on the non-homeostatic examples, we note that networks **a** and **b** respectively break Condition (l.i) and (l.ii); network **c** is not compatible with Condition (l.iv) since, although trivial, there is a critical SCC upstream of another critical SCC.

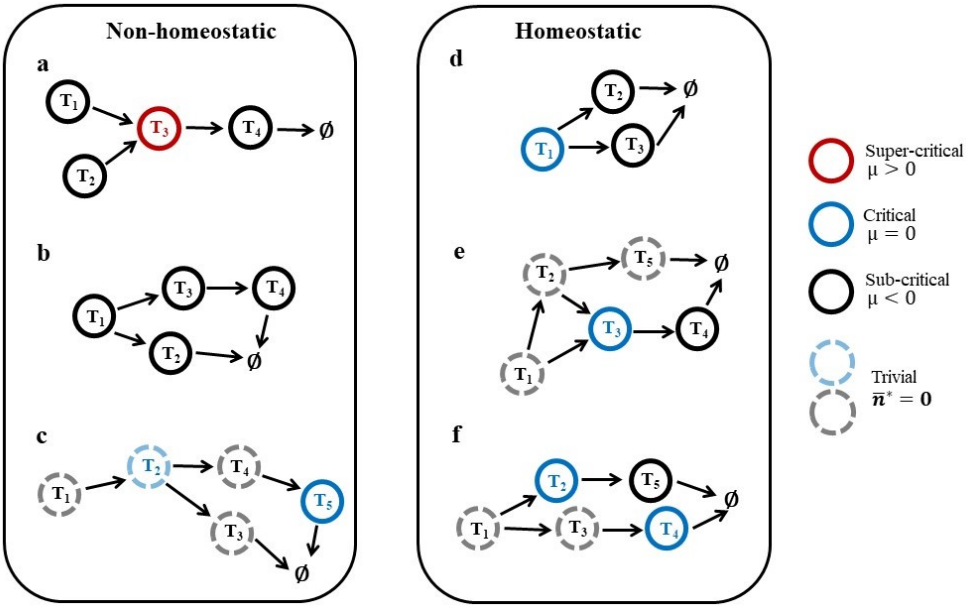


FIGURE 2.2: Illustrative examples of cell type condensed network architectures and compatibility with homeostasis requirements (l.i)-(l.iv). Each circle, corresponding to an SCC, is coloured according to its type; trivial SCCs are indicated with a dashed faded line. The networks **a** and **b** violate respectively Condition (l.i) and (l.ii). In network **c**, there is a (trivial) critical SCC upstream of another critical SCC, which is not compatible with Condition (l.iv). In networks **d-f** all the requirements for homeostasis are met.

2.2.2 Cell type classification and lineage architecture

The results derived so far are purely mathematical. Here, we discuss their implications in the cell type definition and lineage architecture constraints. From the biological perspective, the isolated SCC dynamics represent the cell type's intrinsic long-term proliferative potential since we neglect influx from other cell types. These dynamics only depend on the dominant eigenvalue of the SCC associated with the cell type, μ, which we also call the *growth parameter*. Therefore, we can distinguish three different long-term behaviours for the cell type T_i based on the value of its growth parameter μ_i, which are described below and summarised in Table 2.1.

(a) If $\mu_i > 0$, the expected number of cells of this type increases. Although possible in certain pathologies, such as cancer, this situation is not physiological. We define such cell type as *hyper-proliferating*.

(b) If $\mu_i < 0$, the expected number of cells of type T_i decreases until they vanish. Therefore, all cells of this type and their progeny will eventually differentiate into other types, or die, leaving no cells of type T_i in the tissue. For this reason, we define this cell type as *transient*. However, we observe that an incoming flux of cells sustained over time can prevent this cell type from vanishing.

(c) If $\mu_i = 0$, the expected number of cells remains, on average, constant. Cells of this type maintain their number constant by themselves, without any cell influx from other types. Thus, we define this cell type as *self-renewing*. We note that cells that do not either divide, differentiate, or die (i.e. inert) fall in this class, but they are not considered in this work since they are not part of renewing tissues.

Importantly, we remark that a constant parameter model describes these cell population dynamics, and therefore the growth parameter of each cell type is constant over time. This approximation is acceptable since we expect that homeostasis is associated with a (marginally stable) steady condition. In reality, as we will discuss in Chapter 3, μ may depend on the cellular environment and change over time.

Now that we have classified the cell type based on the growth parameter, we analyse the mathematical Conditions (l.i)-(l.iv) for achieving homeostasis. Condition (l.i) directly relates to hyper-proliferating cells, which are clearly not compatible with homeostasis and therefore excluded hereafter unless specified. Condition (l.ii) is necessary since in the presence of just transient cells (hyper-proliferating ones are excluded based on Condition (l.i)), the tissue cannot self-maintain, and the cell numbers continuously decrease until they vanish. For translating Conditions (l.iii) and (l.iv) in the biological context, we need to consider that any cell influx entering into a self-renewing cell type implies an increase in the number of self-renewing cells which, by definition, maintain their number constant without external contributions. We analyse now all the possible scenarios for a cell type upstream of a self-renewing one: a) a non-trivial self-renewing cell type implies a constant influx of cells; b) a trivial self-renewing cell type gives no cells contribution downstream, but, if perturbed, the contribution becomes non zero, and self-maintains over time; c) cell influx of a non-trivial transient cell type declines until all transient cells disappear; and d) a trivial transient cell type nominally does not produce any cell, condition naturally restored after any perturbation. Although scenario c) might be of interest in tissue development, as later commented, only scenario d) is compatible strictly with homeostasis, which is equivalent to saying that there are no other cell types upstream of a self-renewing cell type.

From the above considerations, an important result for stem cell biology follows, that is *in homeostatic renewing tissues, every self-renewing cell type resides at an apex of a cell lineage hierarchy, and every such lineage has a self-renewing type at its apex*, [Greulich et al., 2021]. An equally important consequence of this result, which is again applicable to

homeostatic renewing tissues, is that the self-renewing cell type has the potential of generating cells in the whole lineage and vice versa (the ability to generate cells in the whole lineage implies self-renewing potential). For this reason, we can identify a self-renewing cell type as an *adult stem cell*. Crucially, this mathematical modelling of homeostasis in renewing tissues leads us to derive these two features (i.e. self-renewing potential and full lineage potential) that characterise the adult stem cells and are commonly observed in real renewing tissues [National Institute of Health, 2016]. We note that multiple apexes of a lineage can, in principle, exist, with different stem cells at each apex, as is conjectured for mouse mammary epithelium lineage, for instance, [Pal et al., 2017]. The remaining cell types must be of the transient type and placed downstream of the adult stem cells. We define them as *committed cells*. This cell population, which would naturally vanish, is maintained by the adult stem cells. The last column of Table 2.1 completes the cell type classification picture based on the growth parameter with the above cell functionality in homeostatic renewing tissues (i.e. adult stem cell or committed cell). In Figure 2.3, we show a typical cell lineage hierarchy.

We finally remark that we discuss some non-homeostatic conditions such as tissue development or regeneration after damage in [Greulich et al., 2021]. In this work, we show that our mathematical model is compatible with the existence of *developing* cells, i.e. transient non-trivial cell types, which are upstream of the adult stem cells, and of *dormant* stem cells.

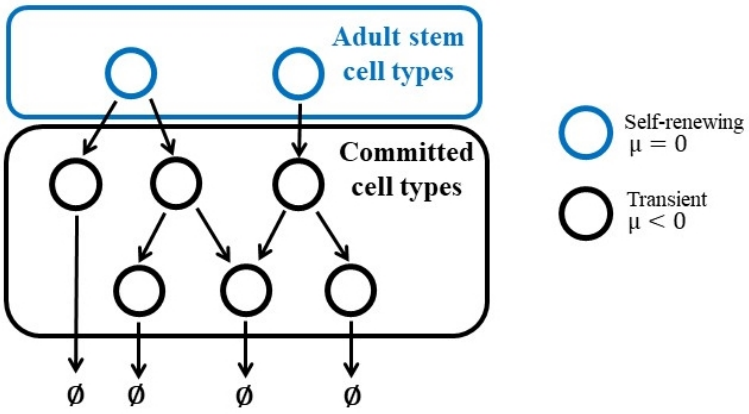


FIGURE 2.3: Illustration of a typical cell lineage tree. Each circle represents a cell type, which comprises a maximal set of mutually reachable cell states, and arrows are possible transitions between cell types. The blue circles represent self-renewing cell types, and the black ones are transient cell types. Crucially, along each homeostatic lineage trajectory, that is, a series of transitions between cell types active in homeostasis, only a single self-renewing cell type can contribute to homeostasis, which we identify as adult stem cells. Therefore, a single stem cell type must be at each apex of the homeostatic lineage. Downstream cell types form the committed types whose progeny is eventually lost. This figure is inspired by that presented in [Greulich et al., 2021].

Math classification		μ	Bio classification	
SCC type	**Long-term dynamics**		**Cell type**	**Cell function**
super-critical	growing	> 0	hyper-proliferating	-
critical	steady	$= 0$	self-renewing	stem
sub-critical	vanishing	< 0	transient	committed

TABLE 2.1: The growth parameter μ of a cell type, corresponding to the dominant eigenvalue of the associated strongly connected component (SCC), determines the classification of the SCC, the corresponding long-term dynamical behaviour of the isolated system, the cell type and its function in homeostatic renewing tissues.

2.2.3 Non-linearity of the cell fate dynamics

So far, we have discussed the cell fate dynamics modelled as a linear system of ODEs and derived a necessary and sufficient condition for a marginally stable homeostatic steady state, requiring self-renewing cells at the apex, and only there, of the lineage tree. In this frame, a fine-tuning of the model parameters is essential for a self-renewing state. This corresponds to a precisely zero growth parameter, a condition biologically implausible without any regulation. In reality, tissue population dynamics are non-linear due to interaction with their local micro-environment and cell signalling. From a mathematical point of view, non-linearity translates into kinetic parameters dependent on the cell numbers, and thus, the dynamics are written as

$$\frac{d}{dt}\bar{\boldsymbol{n}} = A(\boldsymbol{n})\bar{\boldsymbol{n}}\,. \tag{2.12}$$

Notably, a weaker but equally important statement about homeostasis also applies to non-linear systems, formulated as follows. A homeostatic state only exists if the following conditions are satisfied.

(nl.i) There must not be any non-trivial super-critical SCCs.

(nl.ii) There must be at least one non-trivial critical SCC.

(nl.iii) Any SCC upstream of a critical SCC must be trivial.

(nl.iv) There are no directed paths from a non-trivial critical SCC to another.

We discuss below the main differences from the linear case and leave in Appendix A.1 the detailed proof of the above statement, which is taken from [Greulich et al., 2021].

First of all, we observe that homeostasis in a linear system corresponds to a marginally stable steady-state. Instead, in non-linear systems, it corresponds only to a steady-state condition, without any specification about its stability properties. For addressing the stability of the homeostatic steady-state, the underlying regulation

mechanism must be known. Importantly, whilst Conditions (l.i)-(l.iv) are necessary and sufficient conditions for having a marginally stable steady-state in a linear system, Conditions (nl.i)-(nl.iv) are necessary but not sufficient ones for the existence of a steady-state in the non-linear case. That means that non-homeostatic cell fate models where those conditions are satisfied might exist. Finally, we observe that Conditions (nl.i)-(nl.iv) explicitly refer to non-trivial SCCs, whereas the corresponding conditions for the linear system do not. In other words, it is admitted to have, for example, a super-critical SCC if it is trivial since it does not affect the existence of the steady-sate. However, we must consider that in the biological context, a trivial sub-system means that there are no cells of that type in the steady-state, and therefore it is virtually non-existent. Given that, the conditions derived for non-linear systems are equivalent to those for linear ones.

The above considerations imply that the lineage structure depicted in Figure 2.3 is, in general, the only possible one compatible with homeostasis. Nevertheless, homeostasis stability is attained only when the system is adequately regulated. We will assess later, in Chapter 3, these non-linear dynamics when a specific regulation mechanism, the crowding feedback, is included.

2.2.4 Numerical examples of cells' dynamics

This section shows some numerical examples of different cell linear dynamics to illustrate the main results of the homeostasis modelling. Examples of non-linear dynamical models will be analysed instead in Chapter 3.

We first consider two cases of isolated cell type constituted respectively by a single cell state, $m = 1$, and a four cell states, $m = 4$, based on the following models

$$X_1 \xrightarrow{\lambda_1} X_1 + X_1, X_1 \xrightarrow{\gamma_1} \emptyset\,, \tag{2.13}$$

and

$$\begin{aligned}
&X_1 \xrightarrow{\lambda_1} \begin{cases} X_1 + X_1 & \text{with probability } r_1^{11} \\ X_1 + X_4 & \text{with probability } r_1^{14} \end{cases}, X_1 \xrightarrow{\gamma_1} \emptyset, \\
&X_2 \xrightarrow{\lambda_2} X_1 + X_2, X_2 \xrightarrow{\omega_{24}} X_4, X_2 \xrightarrow{\gamma_2} \emptyset, \\
&X_3 \xrightarrow{\omega_{32}} X_2, X_3 \xrightarrow{\gamma_3} \emptyset, \\
&X_4 \xrightarrow{\lambda_4} X_2 + X_3, X_4 \xrightarrow{\omega_{42}} X_2,
\end{aligned} \tag{2.14}$$

in which we arbitrarily chose $r_1^{11} = 0.9$, and $r_1^{14} = 0.1$. The corresponding cell state network is shown in Figure 2.4 (top panels). In the four-state network, we note that cell states are connected such that they form a single strongly connected component.

The corresponding matrix A for these two models is respectively

$$A = \lambda_1 - \gamma_1 \,, \tag{2.15}$$

and

$$A = \begin{pmatrix} \lambda_1(2r_1^1 - 1) - \gamma_1 & \lambda_2 & 0 & 0 \\ 0 & -\omega_{24} - \gamma_2 & \omega_{32} & \omega_{42} + \lambda_4 \\ 0 & 0 & -\omega_{32} - \gamma_3 & \lambda_4 \\ \lambda_1 2r_1^4 & \omega_{24} & 0 & -\lambda_4 - \omega_{42} \end{pmatrix} , \tag{2.16}$$

in which $r_1^1 = (2r_1^{11} + r_1^{14})/2 = 0.95$ and $r_1^4 = r_1^{14}/2 = 0.05$.

For each cell state model, we study the cells' dynamics for different values of the growth parameter $\mu = 0, 0.2, -0.2$, respectively representative of a self-renewing (**SR**), hyper-proliferating (**HP**), and transient (**T**) cell type, conditions achieved by varying the kinetic parameters. While this is straightforward in the single-state case since $\mu = \lambda_1 - \gamma_1$, in the four-state model, we used a stochastic optimisation algorithm (Matlab *ga* function) to find a set of parameters matching the target μ. The corresponding test cases parameters, based on arbitrary units and unitary γ_1, are reported in Table 2.2.

Cell type	μ	$m = 1$		$m = 4$								
		λ_1	γ_1	λ_1	λ_2	λ_4	γ_1	γ_2	γ_3	ω_{24}	ω_{42}	ω_{32}
SR	0	1.00	1.00	1.01	0.76	1.45	1.00	1.20	1.28	0.84	1.78	1.31
HP	0.2	1.20	1.00	1.19	1.24	1.25	1.00	1.26	1.25	1.25	1.25	1.25
T	-0.2	0.80	1.00	0.62	1.08	1.53	1.00	1.29	1.09	1.47	1.28	1.71

TABLE 2.2: Test case model parameters, based on arbitrary units and unitary γ_1, correspond to three values of growth parameter representing respectively a self-renewing (**SR**), $\mu = 0$, hyper-proliferating (**HP**), $\mu = 0.2$, and transient (**T**), $\mu = -0.2$, cell type. Values for the $m = 4$ test case are one solution of a stochastic optimisation problem in which the distance from the target μ is minimised (Matlab *ga* function).

The total cell numbers time evolution, shown in Figure 2.4 (bottom panels), is the results of the integration of the ODEs system (2.5), based the explicit Runge-Kutta Dormand-Prince method (Matlab *ode45*). The initial condition is $\bar{n}^0 = 10^4$ in the single-state model, where the dynamical state is scalar. In the four-state model, instead, the initial condition corresponds to the dominant eigenvector scaled such that the total cell number is equal to 10^4. In the figures, the time is scaled by the reference kinetic parameter $\bar{\alpha} = \gamma_1$ and the total cell numbers by the initial value $\bar{n}^0$. Both cell state networks, $m = 1$ and $m = 4$, show the same dynamical behaviour for each growth parameter value. In particular, the self-renewing case, $\mu = 0$ is a homeostatic case with a total cell number maintained constant and equal to its initial value, whilst the cell number in the hyper-proliferating and the transient cases respectively diverges and vanishes.

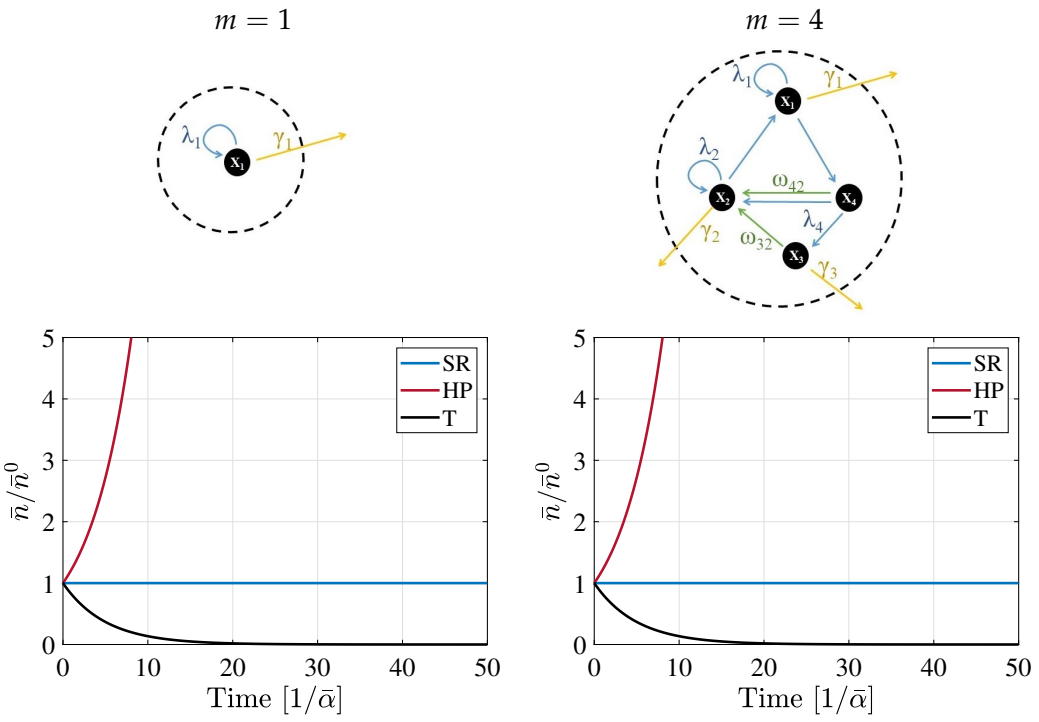


FIGURE 2.4: Cell state networks for $m = 1$ (top-left) and $m = 4$ (top-right) test cases. In the $m = 4$ test case, the four states form a single SCC which, based on the proposed modelling, corresponds to a single cell type. The specific values of kinetic parameters of each network, reported in Table 2.2, correspond to a self-renewing (**SR**), hyper-proliferating (**HP**) and transient (**T**) cell type. The normalised cells' dynamics (bottom panels) refer to the isolated cell type (i.e. without any cell influx). Time is scaled by a reference parameter $\bar{\alpha} = \gamma_1 = 1$ and the total cell number by its initial value. Despite the differences in the cell state networks, the total cell number time evolution only depends on the growth parameter μ, resulting the same in the two test cases. If $\mu = 0$, the cell type presents a self-renewing behaviour where the cell number remains constant; if $\mu > 0$, the cell type is hyper-proliferating, which correspond to a diverging dynamics; and if $\mu < 0$, the cell number of the transient cell type decreases until it completely vanish.

Now that we have shown that independently on the complexity of the cell state network the natural dynamics of a cell type only depend on the growth parameter μ, we focus only on single-state cell types and study how the dynamical behaviour changes when cell types are connected. Thus, we combine k cell types, each of which is based on model (2.13) and characterised by the kinetic parameters reported in Table 2.2 ($m = 1$ columns). In particular, we study the six different configurations shown in Figure 2.2, where cases **a-c** represent non-homeostatic cell type condensed networks and cases **d-f** homeostatic ones. We note that the parameter γ_1 of Table 2.2 is treated as a differentiation rate ω_{ij} connecting two types composed by state i and j (except for the most downstream types of which cells do not differentiate but die). Besides, if a cell type is connected to two cell types (e.g. cases **b-f** of Figure 2.2), then we assume $\omega_{ij} = 0.5\gamma_1$ so that total outgoing cells rate is equal to γ_1.

The integration of the ODEs system (2.5), is again based on the explicit Runge-Kutta Dormand-Prince method (Matlab *ode45*). The initial condition in each non-trivial cell type is randomly chosen such that the total initial cell number is equal to 10^4. For the trivial cell types, the initial cell number is equal to 1, representing a possible perturbation. The time evolution of the total cell number is shown in Figure 2.5 and the details of cell numbers in each cell type in Figure 2.6. The time is scaled by the reference kinetic rate $\bar{\alpha} = \gamma_1 = 1$ and cell number by the initial and the final total cell number, respectively, in the non-homeostatic cases (left panels) and the homeostatic ones (right panels).

We observe that the cells' dynamics agree with the expected ones. In particular, the cell numbers of all the transient cell types tend to vanish in the long term or remain trivial unless there is a non-zero influx of cells from some upstream cell types. Instead, self-renewing cell types are characterised by a constant cell number only if they are at the cell lineage's apex and not connected to other self-renewing cell types (all the homeostatic cases). In particular, concerning case **f**, we note that two separate lineages coexist, one maintained by self-renewing cell type T_2 and one by T_4, which are disconnected. Focusing on case **c**, instead, we observe that the two self-renewing types, the trivial one T_2 and the non-trivial one T_5, are connected. In this case the initial perturbation in the cell number of the trivial cell types, T_1 to T_4, (see zoom in Figure 2.6, bottom-left panel), is only restored in the transient T_1 type, whilst it is maintained in the self-renewing one T_2 and consequently also in the downstream ones, T_3 and T_4, despite being of transient type. Overall, this leads to a small but not negligible increase in the cell number of the self-renewing cell type T_5, which slowly diverge from the homeostatic condition.

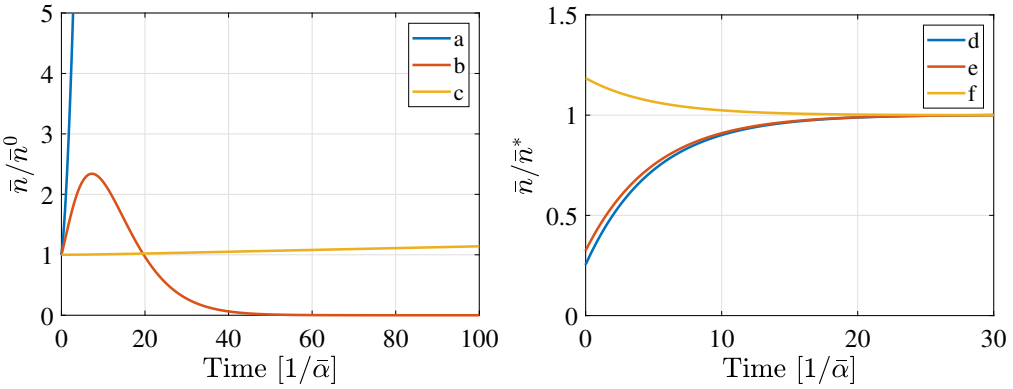


FIGURE 2.5: Cell dynamics corresponding to the cell type networks depicted in Figure 2.2. The evolution of the total cell number is shown as a function of time. The time is scaled by the reference value $\bar{\alpha} = \gamma_1 = 1$ and the total cell number by its initial value in the non-homeostatic cases, **a-c** (left panel), and by its final value in the homeostatic ones, **d-f** (right panel). In the non-homeostatic cases, the cell number diverges or vanishes; in the homeostatic ones, it reaches a constant value in the long term. The details of the cell numbers' evolution for each type are given in Figure 2.6.

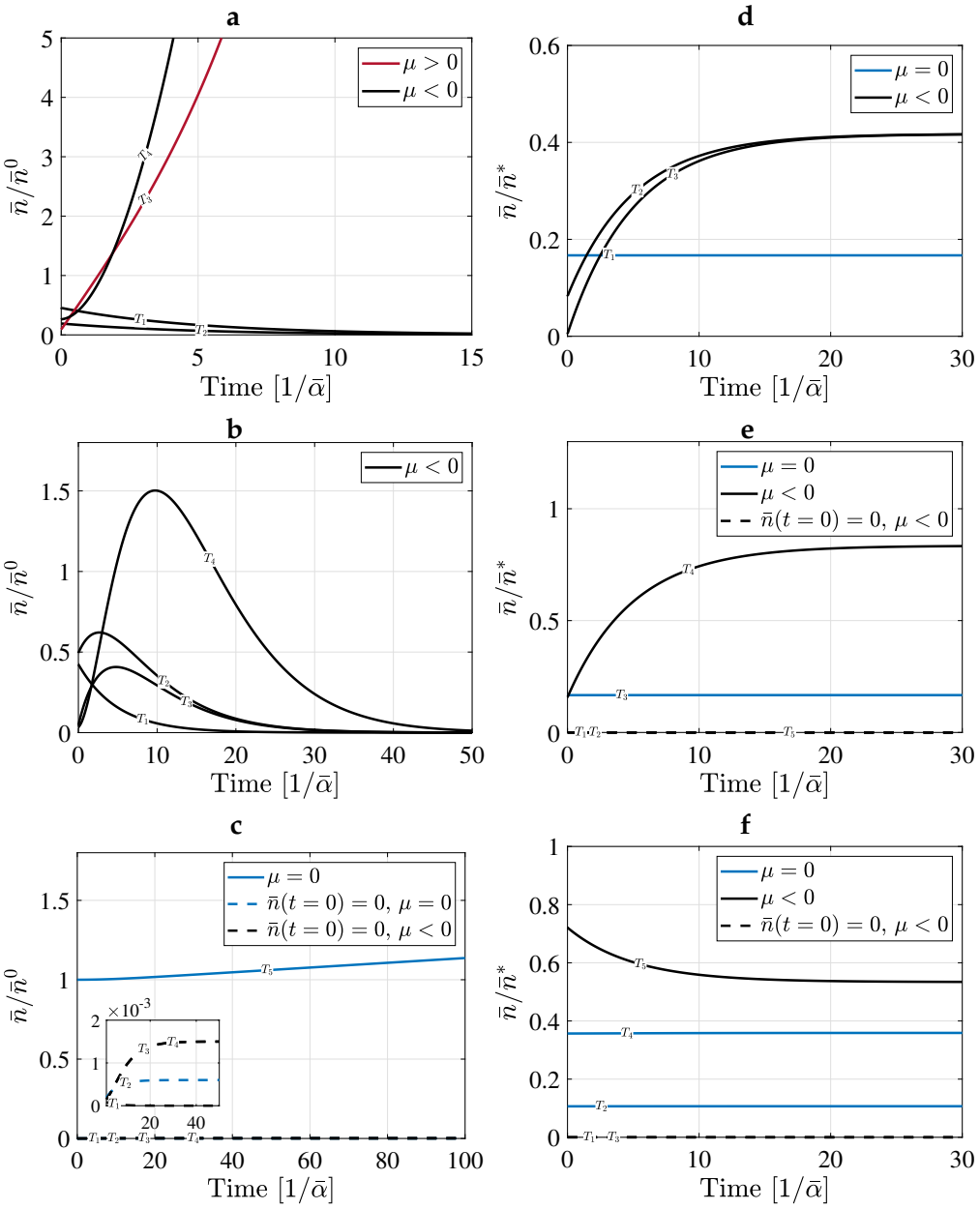


FIGURE 2.6: Details of each cell type number corresponding to the test networks presented in Figure 2.2 which cells' dynamics are shown in Figure 2.5. The time is scaled by a reference value $\bar{\alpha} = \gamma_1 = 1$, and the cell numbers by the initial or final total number, respectively in the non-homeostatic (left panels) and homeostatic (right panels) cases. The colour and style of each curve, labelled with the corresponding cell type number, are consistent with its type (see Figure 2.2). The long-term cells' dynamics of the trivial cell types only depend on the upstream dynamics: if there is no cell influx, the cell type vanishes; otherwise, it presents a constant or diverging trend in case of a constant or increasing cell influx. Self-renewing cell types are characterised by a constant cell number only when there is no connection to other self-renewing types, e.g. **d-f**. In case **c**, instead, the perturbation in the trivial self-renewing cell types T_2 is not restored (shown in the zoom detail) and constantly feed the self-renewing cell type T_5, which shows a slowly diverging behaviour.

2.3 Conclusions

This chapter studied generic tissue population dynamics and derived requirements to achieve homeostatic cell fate models. To this aim, we started from generic modelling of the cell fate, based on a multi-type branching process, and we focused on its deterministic approximation, which describes the average numbers of cells in tissue as a set of Ordinary Differential Equations. We then used graph theory to define a cell state network, providing an intuitive view of the relations between cells in different states. Based on this network, we also proposed a cell type definition as the Strongly Connected Components of such network. In this view, the cell types are formed by the mutually reachable states, meaning that cells of the same type have the potential of renewing themselves. Importantly, cell types are connected in a hierarchical order, forming an acyclic condensed network, called the cell type network, in which cell types are the nodes.

We then analysed the homeostatic state, defined as the non-trivial steady-state condition of the dynamical system. When assuming a constant parameter cell fate model, the dynamics are linear and described by a constant matrix A. Hence, homeostasis is only possible if the system is marginally stable, which only depends on the spectral properties of A. Notably, a topological ordering of the cell states, which is strictly related to the SCC of the cell state network, allowed us to rewrite A in a triangular block form, where the diagonal blocks are irreducible Metzler matrices. Based on these properties of the matrix A, we first classified the SCC into critical, super-critical and sub-critical depending on the sign of the dominant eigenvalue, μ, and then extracted the following condition for marginal stability: (l.i), there must not be any super-critical SCCs; (l.ii) there must be at least one critical SCC; (l.iii) if there is any SCC upstream of a critical SCC, it must be trivial; and (l.iv) if there are multiple critical SCCs, there must not be any path connecting them.

The above mathematical considerations were then translated into the biological context. This resulted in classifying a cell type into self-renewing, hyper-proliferating and transient type, depending on the sign of μ, which assumes the meaning of a cell type growth parameter. Considering that hyper-proliferating cell types are incompatible with homeostasis based on Condition (l.i), this type of cell is excluded a priori in any homeostatic tissue. Besides, from Condition (l.ii)-(l.iv), we derived the important conclusion that in homeostatic renewing tissues, every self-renewing cell type resides at an apex of a cell lineage hierarchy, and every such lineage has a self-renewing type at its apex.

From this result, we defined the adult stem cell as the self-renewing cell type of the cell type condensed network, which must reside at the apex of the lineage tree. The committed cell types are all the other cell types downstream of the adult stem cells, and they must be of transient type. This definition of adult stem cells implies that they

have self-renewing and full lineage potentials, properties commonly used to identify stem cells. This work showed from a mathematical perspective that these properties are deeply coupled in adult renewing homeostatic tissues and that they are the only feasible way to achieve homeostasis. We also showed how, with some limitations, the derived conditions for homeostasis also apply to non-linear systems. Here, the existence of a steady-state requires the same cell type network structure as in linear models. Still, only an adequate regulation mechanism allows for a stable homeostatic state. In the next chapter, we will assess a specific scenario based on the crowding feedback modelling.

We finally proposed some illustrative examples of cell fate dynamics based on numerical integration of the dynamics. We first studied two isolated cell types, showing that the long-term dynamics are determined by their growth parameter only. We then assessed the dynamics in cell type networks with multiple types, confirming that only if Conditions (l.i)-(l.iv) are satisfied homeostasis is achieved.

Based on these results, we can therefore restrict the search of a cell-fate model for the study case, discussed in Chapter 5, into the homeostatic models. Although remaining with an arbitrary number of cell state and states connections to define, models not compatible with homeostasis can be excluded by looking at the structure of the cell state network. Homeostasis requires a particular architecture in the lineage tree, in which stem cells are at the apex and only there.

Chapter 3

Homeostasis regulation via crowding feedback

In the previous chapter, we showed that homeostasis restricts the lineage architecture, requiring, among other conditions, a self-renewing cell type at the apex of the lineage. Only a perfect balance of proliferation and death, which, in a linear system, is mathematically equivalent to marginally stable dynamics, enables the self-renewing capability. However, in the biological context, this is unrealistic without a regulation mechanism since any slight imbalance from this homeostatic condition leads to tissue degeneration with either unlimited growth or shrinking. Hence, in this chapter, we assess a possible regulation mechanism of homeostasis mediated by *crowding feedback*. Based on mathematical modelling, we show that homeostasis modelled as marginally stable dynamics can turn into stable dynamics when such feedback is introduced. Under biologically reasonable assumptions, we also provide a simple condition that allows the dynamics to approach a dynamic long-term self-renewing state, either by converging to or remaining confined around the self-renewing state. Additionally, we derive a more restrictive condition on the feedback functions that guarantees asymptotic stability. Finally, we explore the implications of this regulation mechanism, showing how crowding feedback gives robustness to homeostasis.

The assessment of the crowding feedback dynamic long-term self-renewing state, which is reported in Section 3.2.1, is published in [Greulich et al., 2021], work that I co-authored (see details in Section 1.6.2). We are currently working on another journal article reporting the results shown in the rest of this chapter.

This chapter is organised as follow: the description of the crowding feedback model is provided in Section 3.1; the stability and robustness of homeostasis regulated via feedback are discussed respectively in Section 3.2 and Section 3.3; conclusions are given in Section 3.4.

3.1 Crowding feedback modelling

In Section 2.2.1 we derived the conditions required for having a homeostatic system intended as a marginally stable steady-state. In Section 2.2.3, we showed how these conditions are necessary conditions for the existence of a steady-state in non-linear cell fate dynamics. In particular, a self-renewing cell type, and only that, must be at the apex of a cell lineage. To be self-renewing, the kinetic parameters of this stem cell type (rates of division, cell state transitions, and cell loss) need to be finely tuned to achieve a growth parameter (corresponding to the dominant eigenvalue μ of the dynamical system) of precisely zero. If sustained over an extended period, slight deviations from this value imply a loss of self-renewing capacity. Such fine-tuning is biologically implausible in the absence of a homeostatic control mechanism.

In reality, cells are part of a tissue and may respond to signalling from other cells and environmental factors. A simple example of such regulation is the *crowding feedback*, in which cells sense the density of cells in their niche and respond by adjusting their propensity to divide or differentiate. Experimental evidence of this mechanism indicates that overcrowding results in the acceleration of the cell differentiation [Marinari et al., 2012, Eisenhoffer et al., 2012, Eisenhoffer and Rosenblatt, 2013] or the inhibition of cell proliferation [Puliafito et al., 2012]. Both effects have the consequence of reducing cell density. Other experimental works show, instead, an increase in cell proliferation caused by a reduction in the cell density, obtained, for example, by stretching a tissue [Gudipaty et al., 2017]. Although the mechanisms to mediate the crowding feedback are not always clear, experimental studies on mechanosensing showed that cell overcrowding reduces cell motility and consequently produces a compression on cells that inhibits cell proliferation [Puliafito et al., 2012, Shraiman, 2005]. Another potential mechanism that might result in the form of crowding feedback is the competition for limited growth signalling factors [Kitadate et al., 2019]. More specifically, a decrease in the concentration of some proteins, the growth factor, increases the propensity of cells to differentiate rather than divide, which, in turn, raises the growth factors levels. In the same way, when the concentration is higher than expected, cells tend to divide and consequently lower the growth factors.

We therefore model this situation in which cells are able to sense the cell density and adjust their dynamic behaviour accordingly. In the generic model, given by (2.1)-(2.3), this means that the kinetic parameters λ_i, ω_{ij} and γ_i, for $i, j = 1, 2, ..., m$, depend on the density of cells. In the following, we denote as α_j, $j = 1, 2, ..., m + m^2$, the jth kinetic parameter, independently of whether it is a division, transition or death rate. We also indicate with ρ the cell density, which is defined as the average number of cells per unit of volume, V, that is, $\rho = \bar{n}/V$. Thus, when the jth kinetic parameter increases with the cell density, that is $\partial\alpha_j/\partial\rho > 0$, it exhibits a *positive crowding dependence*; if it

decrease with it, that is $\partial\alpha_j/\partial\rho < 0$, it exhibits a *negative crowding dependence*; it may also neither increase or decrease, that is $\partial\alpha_j/\partial\rho = 0$.

In the previous chapter, we also showed that the dynamics of a cell type is independent of those downstream. We maintain this decoupling by assuming that the kinetic parameters for each x-cell type, T_x, depend only on the density of cells of that type, ρ^x, as representative of the niche. In a more generic case, however, crowding feedback may depend on the total cell density or any combination of cell densities of any type (e.g. through signalling, it might also depend on the densities of another cell type), meaning that the cell population dynamics must be assessed as a whole. Given that, we focus this analysis on the isolated self-renewing cell type at the apex of the lineage hierarchy. We will assess crowding feedback acting on downstream committed cell types later, in Section 3.3.2. In other words, we analyse here the cell dynamics of the upstream critical strongly connected component of the cell state network. We omit hereafter any subscript or superscript referring to this cell type and define the total cells' density of that type as $\rho = \sum_{i\in T}\rho_i$, in which ρ_i is the density of cells in the ith state. Thus, the *crowding feedback* is modelled by assuming that the kinetic parameters are a function of ρ.

We recall now that the dynamics of the average numbers of cells are given by (2.12), where the elements a_{ij} of the matrix A depend on the kinetic parameters, $a_{ij} = a_{ij}(\alpha_1, \alpha_2, ...)$. Considering the crowding feedback modelling, since α_j depends on ρ, the matrix A becomes a function of ρ. Hence the dynamics, written here per unit of volume, are governed by a set of non-linear ordinary differential equations. In vectorial form, this is written as

$$\frac{d}{dt}\boldsymbol{\rho}(t) = A(\rho(t))\,\boldsymbol{\rho}(t)\,, \tag{3.1}$$

in which $\boldsymbol{\rho} = (\rho_1, ..., \rho_m)$ comprises only the cells of the type T (m is the number of states of type T) and A is the corresponding sub-matrix of the complete dynamical system. The dependence $A = A(\rho)$ implies that the elements of the matrix are functions of ρ, and therefore also its dominant eigenvalue, μ, is a function of ρ. Crucially, $\mu = \mu(\rho(t))$ becomes a dynamic quantity. Thus, self-renewal corresponds here to a non-trivial fixed point, $\boldsymbol{\rho}^*$, of the Equation (3.1), for which the dominant eigenvalue of A is zero, that is $\mu(\rho^*) = 0$.

3.2 Stability of homeostasis

In this section, we assess the stability of the fixed point ρ^* of the non-linear dynamical system (3.1) and distinguish two types of behaviour. We first define a *dynamic long-term self-renewing state* as a homeostatic state associated with confined dynamics

(i.e. not vanishing nor diverging) where the steady-state is only an average condition. In contrast, strict homeostasis, assessed later, is related to an asymptotically stable fixed point.

3.2.1 Dynamic long-term self-renewing state

To assess the long-term dynamical behaviour of the tissue regulated via crowding feedback, we introduce a simplifying assumption about the kinetic parameters. In particular, we assume that all the parameters α_j, for $j = 1, 2, ...$, which may depend on ρ, converge for large ρ to some limiting value $\alpha_j^\infty > 0$, i.e. $\alpha_j(\rho) \to \alpha_j^\infty$ for $\rho \to \infty$. Also, they remain finite positive values when $\rho = 0$. This assumption is biologically reasonable as it simply states that cell processes cannot become infinitely fast or slow. Based on this assumption, we will show that a dynamic long-term self-renewing state, that is, a non-constant cell dynamics yet confined, can be achieved if the following condition is satisfied

$$\frac{\partial \mu}{\partial \rho} < 0 \text{ for all } \rho \geq 0\,. \tag{3.2}$$

Considering that

$$\frac{\partial \mu}{\partial \rho} = \sum_j \frac{\partial \mu}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial \rho}\,, \tag{3.3}$$

the importance of Condition (3.2) lies on the fact that whilst, in general, $\partial\mu/\partial\rho$ cannot be measured, the sign of $\partial\alpha_j/\partial\rho$ might be. As an example, an increase of the cell proliferation rate, λ, could be detected after stretching an epithelial tissue [Gudipaty et al., 2017], informing about the negative sign of $\partial\lambda/\partial\rho$. If, for all j, the sign of $\partial\alpha_j/\partial\rho$ is opposite to that of $\partial\mu/\partial\alpha_j$, which can be mathematically determined for a given cell fate model, then Condition (3.2) is met. Therefore, measures of the sign of the dependency of the kinetic parameters with cell density can tell about the stability of the crowding feedback regulation mechanism.

To better understand this condition, we first note that the above relation implies the uniqueness of the non trivial fixed point $\boldsymbol{\rho}^*$. Additionally, it also implies that at the trivial fixed point, $\boldsymbol{\rho}^0 = \mathbf{0}$, the matrix $A^0 = A(\rho^0)$ is characterised by a positive dominant eigenvalue $\mu^0 = \mu(\rho^0) > 0$, whilst for large density values, ρ^∞, the matrix $A^\infty = A(\rho^\infty)$ has a negative dominant eigenvalue $\mu^\infty = \mu(\rho^\infty) < 0$.

For assessing the stability of the fixed points, $\boldsymbol{\rho}^0$ and $\boldsymbol{\rho}^*$, we linearise the dynamics (3.1). To this aim we compute the Jacobian matrix J, whose generic ij-element results

in

$$J_{ij} = \frac{\partial(\frac{d}{dt}\rho_i)}{\partial\rho_j} = \frac{\partial(A\boldsymbol{\rho})_i}{\partial\rho_j} = a_{ij} + \sum_k \frac{\partial a_{ik}}{\partial\rho_j}\rho_k = a_{ij} + \sum_k a'_{ik}\rho_k\,, \tag{3.4}$$

in which $a'_{ik} = \partial a_{ik}/\partial\rho$ (note that $\frac{\partial a_{ik}}{\partial\rho_j} = \frac{\partial a_{ik}}{\partial\rho}\frac{\partial\rho}{\partial\rho_j} = a'_{ik}1$), and evaluate it at the fixed point. In a more compact form we can write

$$J = A + A'\boldsymbol{\rho}\,\mathbf{1}^T\,, \tag{3.5}$$

in which A' is a matrix whose elements are the derivatives of the elements of A with respect to ρ, evaluated at the fixed point.

The stability of the steady-state condition depends on the sign of the real part of the eigenvalues of J, that is, the steady-state is asymptotically stable if the largest real part of the eigenvalues of the Jacobian matrix, μ_J, is negative. Focusing now on the trivial fixed point, we note that the Jacobian matrix, J^0, equals A^0. From this, it follows that the trivial solution is unstable being $\mu_J^0 = \mu^0 > 0$. On the other hand, nothing can be said in general about the stability of $\boldsymbol{\rho}^*$. However, we will see later, in Section 3.2.2, that a simple sign condition on the elements of A' is sufficient to guarantee asymptotic stability. However, we further note that for large ρ, the dynamics can be considered linear and based on A^∞, since the kinetic parameters converge to some constant values. Thus, given that μ^∞ is negative, for such linear dynamics, the direction of the variation of the dynamical state is towards decreasing values.

From these considerations, it follows that the cell density cannot diverge, nor go to zero, and thus Condition (3.2) assures that the tissue cell density will be confined around the self-renewing state, $\boldsymbol{\rho}^*$, defined by $\mu^* = 0$. Therefore, in the long-term, if $\boldsymbol{\rho}^*$ is asymptotically stable, i.e. $\mu_J < 0$, cell densities will converge to this homeostatic state; otherwise, if it is unstable, i.e. $\mu_J > 0$, they will fluctuate or oscillate around it, implying that the steady-state is only *locally unstable*. Notably, this could include a dynamic state, in which cells persistently switch between a hyper-proliferating state ($\mu > 0$) and a declining state ($\mu < 0$), in a way, however, that cell densities remain on average constant in the long-term. Either way, the cell population is long-term self-renewing, resulting in a (dynamic) homeostatic state for the tissue, assuming all other cell types are transient.

To demonstrate the impact of the crowding feedback regulation on tissue dynamics, we study three numerical examples. They are based on the same stochastic network shown in Figure 3.1: this is composed by three states connected by state transitions and cell division, forming a single SCC, meaning that, based on the definition given in Section 2.1.2, cells in these states are of the same type. In order to implement the crowding feedback model, we chose each kinetic parameter $\alpha_i \in \{\lambda_j, d_j, \omega_{jk}\}_{j,k=1,\dots,m}$ being a function of ρ. In particular, we chose a Hill function [Lei et al., 2014] of the

type $\alpha_i(\rho) = c_i + k_i\rho^{n_i}/(K_i^{n_i} + \rho^{n_i})$ in case α_i is an increasing function of ρ (i.e. $\alpha_i' = \partial\alpha_i/\partial\rho > 0$), and $\alpha_i(\rho) = c_i + k_i/(K_i^{n_i} + \rho^{n_i})$ in case of a decreasing function of ρ (i.e. $\alpha_i' < 0$).

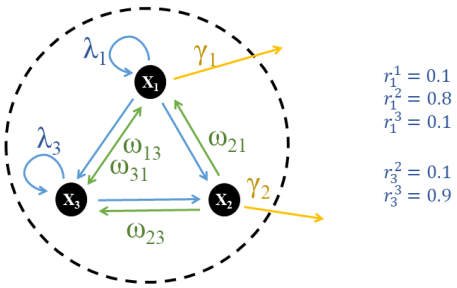


FIGURE 3.1: Cell state network used in the crowding feedback regulation examples. This network is composed of three states connected by state transitions, ω, and cell division λ (the division outcome probability parameters, r_{ij}, are specified on the right). The three states form a single Strongly Connected Component (SCC); based on the definition provided in Section 2.1.2, this means that cells are of the same type. Cells die or exit from the cell type with rate γ. The three test cases, illustrative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics, differs for the parameters describing the feedback regulation which are provided in Table 3.1.

The three test cases differ for the values of the feedback function parameters, which are reported in Table 3.1. These values are the results of a random search, which is detailed in Appendix A.2. More specifically, we chose the test cases based on different values of μ' and μ_J to have different qualitative dynamic behaviours: an asymptotically stable case (**AS**), a locally unstable (**LU**) fulfilling Condition (3.2) and an unstable case (**U**). This is graphically shown in Figure 3.2 (left), where each quadrant is related to one of the three expected behaviours. We remark that cases falling in the (Locally) Unstable quadrant can show an unstable behaviour if Condition (3.2) is not met (i.e. the condition does not apply to the full range of ρ). Thus, whilst for the cases in the Asymptotic Stable and the Unstable quadrants, the expected dynamics is fully determined by the sign of μ' and μ_J, the locally unstable dynamics are just a possible behaviour that depend on each specific dynamical system. Another important remark is that no test case is selected in the quadrant $\mu' > 0$, $\mu_J < 0$. It will be shown later, in Section 3.2.2, that $\mu' < 0$ is a necessary condition for asymptotic stability, thus, with this modelling, that quadrant is not achievable in any way. The dependency of $\mu(\rho)$ in the three cases is shown in Figure 3.2 (right). We note that in general the time unit is arbitrary and therefore omitted. However, for a better comparison of the dynamics, time is scaled by the minimum kinetic rate, $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$, which gives a measure of the timescale of the dynamics, i.e. the smaller the kinetic rate, the longer we need to wait to see events of that type. Consistently, μ' and μ_J are expressed in $\bar{\alpha}$ since they have the same unit as the kinetic rates.

α	**AS**				**LU**				**U**			
	k	K	n	s	k	K	n	s	k	K	n	s
λ_1	0.55	0.77	2	1	26.37	1.17	27	1	0.46	0.70	3	-1
λ_3	77.86	5.80	2	-1	16.62	2.53	2	-1	32.35	3.66	2	-1
γ_1	3.09	0.54	2	1	3.65	0.73	2	1	4.31	0.90	2	-1
γ_2	1.67	0.74	2	1	1.09	0.08	2	1	1.15	0.24	2	-1
ω_{13}	17.52	1.71	2	-1	7.97	0.89	2	-1	5.42	0.46	2	1
ω_{21}	0.30	0.75	10	-1	0.29	0.74	14	-1	0.47	0.81	2	1
ω_{23}	0.26	1.45	5	1	0.04	0.95	104	-1	0.04	0.87	25	-1
ω_{31}	41.72	2.58	2	-1	7.10	0.55	2	1	123.16	4.65	2	-1

TABLE 3.1: Values of the Hill function parameters used to describe the kinetic parameters in case of homeostasis regulation via crowding feedback. The three test cases are illustrative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics. The generic kinetic parameter, α, which is function of the total cell density, ρ, is given by $\alpha(\rho) = c + k\rho^n/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') > 0$ and $\alpha(\rho) = c + k/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') < 0$. A common value $c = 0.05$ is assumed. As detailed in Appendix A.2, these values are computed from α and α' chosen among the results of a random search. The kinetic parameter unit is arbitrary and therefore omitted.

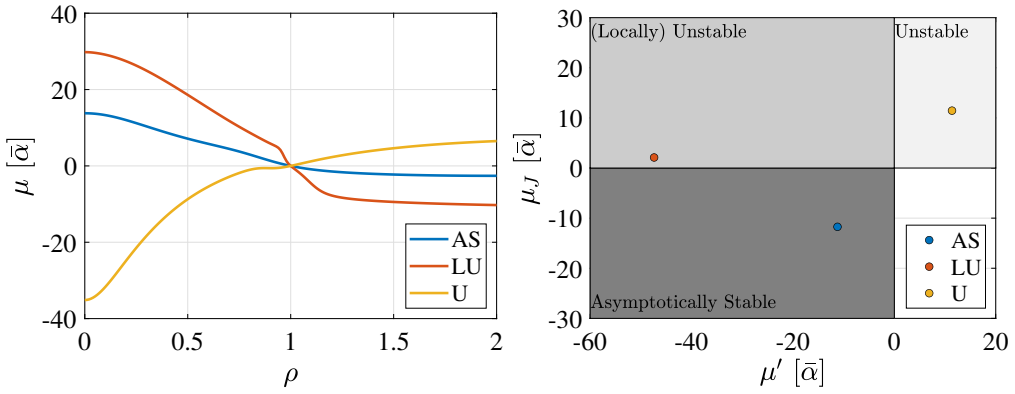


FIGURE 3.2: Dependency of the dominant eigenvalue on cell density, $\mu(\rho)$, (left) and stability parameters, μ_J and $\mu' = \partial\mu/\partial\rho|_{\rho^*}$ (right) for the three test cases, representative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics. The values shown are expressed in $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

We then numerically solved the system of ODEs (3.1) for different initial conditions based on the explicit Runge-Kutta Dormand-Prince method (Matlab *ode45* function). Results are shown in Figure 3.3 as the time evolution of ρ, normalised by the steady-state ρ^*, (left panels), and dominant eigenvalue, μ, (right panels). The first trajectory, indicated in the figure as **H**, is based on an initial condition corresponding to the self-renewing state $\boldsymbol{\rho}^*$; therefore, the system is initially in homeostasis. In the two other simulations, labelled as $\mathbf{P}^-$ and $\mathbf{P}^+$, we apply a perturbation in the initial state. Here, the initial cell densities correspond respectively 0.8 and 1.2 times the expected steady-state $\boldsymbol{\rho}^*$. We observe that, in the **AS**, the effect of the feedback is to compensate for the perturbation in the initial condition so that the system eventually attains a steady-state $\rho \to \rho^*$ (top-left panel) and self-renewal property $\mu \to 0$

(top-right panel) over time. For the **LU** test case (middle panels) instead, the cell density deviates from the steady-state ρ^* but, in the long term, it oscillates with constant amplitude, whereas the average remains constant. As expected, the dominant eigenvalue μ oscillates assuming negative (positive) values for $\rho > \rho^*$ ($\rho < \rho^*$). Finally, in the **U** test case (bottom panels), for any initial condition, the cell densities diverge from the steady-state ρ^*, approaching in one case the trivial condition, which is, in this case, the only stable fixed point, and grows in the others.

Thus, these examples show that for various initial conditions, which might be not self-renewing, if Condition (3.2) is satisfied, the growth parameter approaches or oscillates around the value $\mu = 0$. Consistently, the cell density asymptotically converges to or remains confined around a self-renewing state, corresponding to a (dynamic) homeostatic state. We observe that Condition (3.2) is only a sufficient condition, meaning that a dynamic homeostatic state may exist even when this condition is not fulfilled in the whole range of ρ. In general, due to non-linearities, more complex dynamics with multiple fixed points are also possible. In this case, the specific stability properties of each fixed point must be evaluated to determine the global behaviour of the dynamical system.

3.2.2 Asymptotic self-renewing state

To derive the conditions for a strict homeostatic state, which corresponds to an asymptotically stable fixed point, $\boldsymbol{\rho}^*$, of the non-linear dynamics (3.1), we follow a standard approach and evaluate the eigenvalues of the Jacobian matrix of the linearised system. For clarity, in the following, we indicate as x^* a variable $x(\rho)$ evaluated at the steady state, e.g. $a_{ij}^* = a_{ij}(\rho^*)$, and with x' its partial derivative with respect to ρ, evaluated at the steady state, e.g. $a'_{ij} = \partial a_{ij}/\partial\rho|_{\rho^*}$.

Before linearising the system, we apply a coordinate change through a matrix W, such that the transformed matrix at the fixed point, $\tilde{A}^*$, is in Jordan form. The dynamics in this coordinate system result in

$$\frac{d}{dt}\tilde{\boldsymbol{\rho}}(t) = \tilde{A}(\rho)\,\tilde{\boldsymbol{\rho}}(t), \tag{3.6}$$

in which $\tilde{\rho} = W^{-1}\rho$ and $\tilde{A}(\rho) = W^{-1}A(\rho)W$. The transformation matrix W is constant and does not adjust with ρ, meaning that $\tilde{A}(\rho)$ is not the Jordan normal form for any ρ but only for ρ^*. This transformation results in

$$W = \begin{pmatrix} 1 & 1 & ... \\ w_{21} & w_{22} & ... \\ ... & & \end{pmatrix}, \tag{3.7}$$

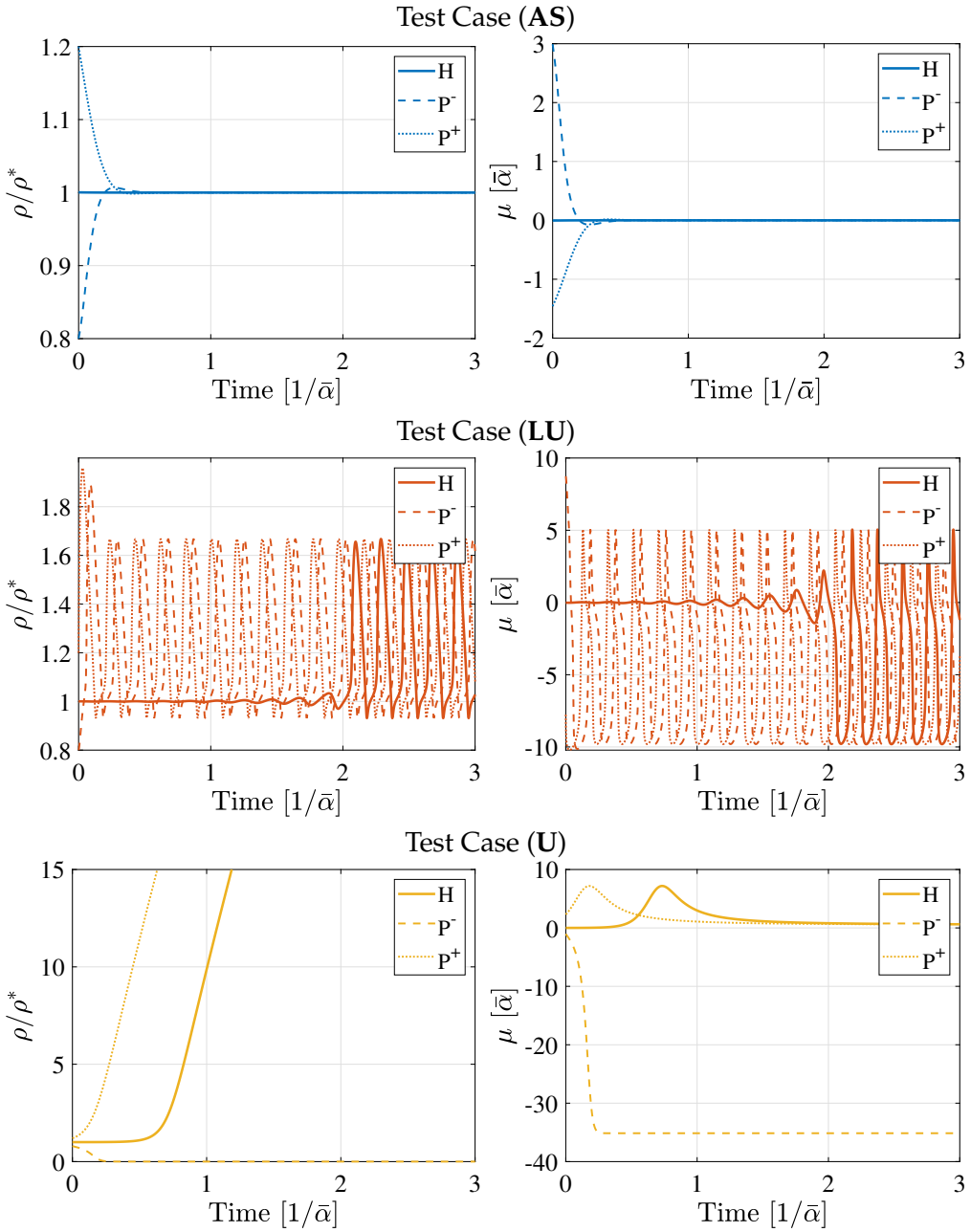


FIGURE 3.3: Cell dynamics based on crowding feedback modelling. The three test cases are representative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics. Line colours are consistent with those used in Figure 3.2. The cell density ρ, normalised by the steady-state ρ^*, (left panels) and the dominant eigenvalue, μ, (right panels), are shown as a function of the time. Time is scaled by the inverse of the smallest rate at the steady-state ρ^*, $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$. For each test case, three sets of initial conditions are tested: **H** corresponds to an initially self-renewing case; $\mathbf{P}^{\pm}$ instead represent two different not self-renewing conditions. As expected, we observe that in the **AS** and **LU** cases, respectively shown in the top and middle panels, the parameters self-adjust over time such that the growth parameter eventually attains $\mu = 0$ or oscillates around this condition. Consistently, the cell density becomes stationary, $\rho \to \rho^*$, or oscillatory with constant amplitude. In the **U** test case (bottom panels), depending on the initial condition, cell density grows or decays to zero, and consistently μ converges to positive or negative values.

which, applied to matrix $A(\rho)$, gives

$$\tilde{A}(\rho) = \begin{pmatrix} \tilde{\mu}(\rho) & (\tilde{\boldsymbol{a}}(\rho))_{1j} \\ (\tilde{\boldsymbol{a}}(\rho))_{i1} & \tilde{A}_{rem}(\rho) \end{pmatrix}, \tag{3.8}$$

in which $\tilde{a}^*_{1j} = 0$ and $\tilde{a}^*_{i1} = 0$ for $i, j = 2, ..., m$, where m is the number of cell states[1]. Importantly, we notice that the transformation W does not affect the eigenvalues of the matrix A, and thus, for any ρ, the dominant eigenvalue of $\tilde{A}$ is the same as that of A, i.e. μ. Furthermore, at the steady state, $\tilde{a}^*_{11} = \tilde{\mu}^* = \mu^* = 0$. Considering that μ is a simple eigenvalue, Theorem 6.3.13(b) of [Horn and Johnson, 1985] holds. This is

$$\frac{\partial \mu}{\partial \tilde{a}_{ij}} = \frac{\bar{v}_i u_j}{\bar{\boldsymbol{v}}\boldsymbol{u}}, \tag{3.9}$$

in which $\boldsymbol{u}$ and $\boldsymbol{v}$ are respectively the right and left eigenvectors, and $\bar{v}$ indicates the conjugate transpose of v (in this particular case however v is real since it is associated to the dominant eigenvalue). Since $\tilde{a}_{11} = \tilde{\mu}$, we can rewrite the left hand term of Equation (3.9) as $\dfrac{\partial \mu}{\partial \tilde{a}_{11}} = \dfrac{\partial \mu}{\partial \tilde{\mu}} = \dfrac{\partial \mu}{\partial \rho}\dfrac{\partial \rho}{\partial \tilde{\mu}}$. At the steady state ρ^*, $\tilde{A}^*$ is in Jordan form, and the left and right dominant eigenvalues, respectively $\boldsymbol{v}^*$ and $\boldsymbol{u}^*$, are zeros except for the first component v_1 and u_1. Thus, the right hand side of Equation (3.9) is equal to 1. This results in $\left.\dfrac{\partial \mu}{\partial \tilde{a}_{11}}\right|_{\rho^*} = \dfrac{\mu'}{\tilde{\mu}'} = 1$, which implies that $\tilde{\mu}' = \mu'$.

The non-linear dynamical system (3.6), close to the steady-state, can be approximated by a linear one, governed by

$$\frac{d}{dt}\Delta\tilde{\boldsymbol{\rho}}(t) = \tilde{J}\,\Delta\tilde{\boldsymbol{\rho}}, \tag{3.10}$$

in which $\Delta\tilde{\boldsymbol{\rho}}$ is the deviation from the steady state $\boldsymbol{\rho}^*$, and $\tilde{J}$ is the Jacobian matrix. The ij-element of this matrix is

$$\tilde{J}_{ij} = \left.\frac{\partial(\frac{d}{dt}\tilde{\rho}_i)}{\partial \tilde{\rho}_j}\right|_{\rho^*} = \left.\frac{\partial(\tilde{A}\tilde{\boldsymbol{\rho}})_i}{\partial \tilde{\rho}_j}\right|_{\rho^*} = \tilde{a}^*_{ij} + \sum_k \left.\frac{\partial \tilde{a}_{ik}}{\partial \tilde{\rho}_j}\right|_{\rho^*} \tilde{\rho}^*_k\,. \tag{3.11}$$

Considering that the term $\dfrac{\partial \tilde{a}_{ik}}{\partial \tilde{\rho}_j} = \dfrac{\partial \tilde{a}_{ik}}{\partial \rho}\dfrac{\partial \rho}{\partial \tilde{\rho}_j} = \tilde{a}'_{ik}\dfrac{\partial \sum_l (W\tilde{\boldsymbol{\rho}})_l}{\partial \tilde{\rho}_j} = \tilde{a}'_{ik}\sum_l w_{lj}$, we can write the generic Jacobian matrix element as

$$\tilde{J}_{ij} = \tilde{a}^*_{ij} + \sum_k \tilde{a}'_{ik}\tilde{\rho}^*_k W_j\,. \tag{3.12}$$

in which $W_j = \sum_l w_{lj}$. We observe now that $\tilde{\rho}^*_k = 0$ for $k > 1$ and $\tilde{\rho}^*_k = \rho^*_1$ for $k = 1$ since $\tilde{\boldsymbol{\rho}}^*$ is the eigenvector associated with the dominant eigenvalue. From this, it

[1] Since A^* is an irreducible Metzler matrix, the dominant eigenvalue μ is real and simple (see Section 2.2.1). Thus, the Jordan block associated to μ is scalar [Horn and Johnson, 1985]

follows that $w_{j1} = \rho_j^* / \rho_1^*$, and therefore $W_1 = \rho^* / \rho_1^*$. Therefore, the Jacobian matrix can be further simplified resulting in

$$\tilde{J} = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{A}^*_{rem} \end{pmatrix} + \begin{pmatrix} \mu' \rho^* & \mu' \rho_1^* W_j \\ (\tilde{\boldsymbol{a}})'_{i1} \rho^* & (\tilde{\boldsymbol{a}})'_{i1} \rho_1^* W_j \end{pmatrix} . \tag{3.13}$$

Importantly, we note that $\tilde{J}$ and J have the same eigenvalues given that

$$\begin{aligned} W^{-1} J W &= W^{-1} A W + W^{-1} A' \boldsymbol{\rho}^* \mathbf{1}^T W = \tilde{A} + W^{-1} A' W \tilde{\boldsymbol{\rho}}^* \mathbf{1}^T W \\ &= \tilde{A} + \tilde{A}' \tilde{\boldsymbol{\rho}}^* \sum_i W_{ji} = \tilde{J} \end{aligned} , \tag{3.14}$$

in which J is the Jacobian matrix in the non-transformed coordinates given by Equation (3.5). Thus, in general, the system's stability depends on the sign of the real part of the eigenvalues of the Jacobian matrix (3.13).

Since it is impossible to derive a generic closed-form solution for any dimension, we first analyse the straightforward one-dimensional case analytically and then the two-dimensional one. These cases correspond respectively to a single-state and a two-state stochastic cell network. Based on that, we derive a sufficient condition for asymptotic stability and verify this condition numerically for higher dimensions. Notably, the derived sufficient condition implies that there might be other configurations for which the condition is not satisfied, yet the solution is asymptotically stable. With the same approach, we also derive a necessary condition for stability.

3.2.2.1 Single-state cell network

In the case of a single-state cell type, the problem is reduced to a scalar one in the cell density ρ. For such dynamics, the matrix A is actually a scalar and it is equal to the dominant eigenvalue μ. Thus, the linearised system around the steady-state, ρ^*, is

$$\tilde{J} = \mu^* + \left. \frac{\partial \mu}{\partial \rho} \right|_{\rho^*} \rho^* = \mu' \rho^* . \tag{3.15}$$

From the above, ρ^*, which is positive, is asymptotically stable if and only if $\mu' < 0$. Thus, in this single-state cell type case, Condition (3.2) for the existence of a (dynamic) homeostatic state assures the asymptotic stability of such steady-state. Importantly, we observe that the here derived necessary and sufficient condition for asymptotic stability is less restrictive than (3.2) as it must hold only at the steady-state and not for any ρ.

3.2.2.2 Two-state cell network

We study now the stability of a generic two-state single-type cell dynamics. We first rewrite the steady state as $\rho_1^* = \beta\rho^*$ and $\rho_2^* = (1-\beta)\rho^*$, where $\rho^* = \rho_1^* + \rho_2^*$ and $\beta = [0,1]$. We also note that since the dominant eigenvalue of A^*, μ^*, is equal to zero, the second eigenvalue, μ_2, must be real and negative[2]. In particular, we can write $\mu_2 = -K$, where $K = \sqrt{(a_{11}^* - a_{22}^*)^2 + 4a_{12}^* a_{21}^*} = -\operatorname{Tr}(A^*)$ is a positive quantity. Based on this parametrisation, we can rewrite the transformation matrix, W, given by Equation (3.7), and the transformed matrix at the steady state, $\tilde{A}^*$, given by Equation (3.8), respectively as

$$W = \begin{pmatrix} 1 & 1 \\ \dfrac{1-\beta}{\beta} & -\dfrac{K}{a_{12}^*} + \dfrac{1-\beta}{\beta} \end{pmatrix}, \tag{3.16}$$

and

$$\tilde{A}^* = \begin{pmatrix} 0 & 0 \\ 0 & -K \end{pmatrix}. \tag{3.17}$$

Thus, the Jacobian matrix in Equation (3.13) results in

$$\tilde{J} = \tilde{A}^* + \tilde{A}' \begin{pmatrix} \rho_1^* \\ 0 \end{pmatrix} \begin{pmatrix} \dfrac{1}{\beta} & \dfrac{1}{\beta} - \dfrac{K}{a_{12}^*} \end{pmatrix} = \begin{pmatrix} \rho\mu' & \rho\mu' \left(1 - \dfrac{K}{a_{12}^*}\beta\right) \\ \rho\tilde{a}_{21}' & -K + \rho\tilde{a}_{21}' \left(1 - \dfrac{K}{a_{12}^*}\beta\right) \end{pmatrix}. \tag{3.18}$$

The eigenvalues of $\tilde{J}$, $\lambda_{1,2}$, are

$$\lambda_{1,2} = 1/2 \left(B \pm \sqrt{4K\rho\mu' + B^2} \right) \tag{3.19}$$

in which $B = -K + \rho((1 - K\beta/a_{12}^*)\tilde{a}_{21}' + \mu')$.

Crucially, we note that if $\mu' > 0$, then $4K\rho\mu' + B^2 > 0$, so the two eigenvalues are real. Also, it holds that $\sqrt{4K\rho\mu' + B^2} > B$, and therefore the dominant eigenvalue of the Jacobian matrix, $\mu_J = \lambda_1$, is always positive. Therefore, a necessary condition for asymptotic stability is that

$$\mu' = \left.\frac{\partial\mu}{\partial\rho}\right|_{\rho^*} < 0\,. \tag{3.20}$$

We note that this is the same condition derived for a single-state system. However, in this case, the condition is necessary but not sufficient for asymptotic stability.

[2] In a two-state cell network, the matrix A^* has two eigenvalues. Since complex eigenvalues are in pairs, if the dominant eigenvalue is real, the other must also be real. Also, if the dominant eigenvalue, which is simple, is zero, then the second one must be negative.

Assuming now that $\mu' < 0$, we note that the term $4K\rho\mu' < 0$, since both K and ρ are positive. Thus, $B < 0$ is a sufficient condition to have two negative eigenvalues[3]. To determine now the sign of B, we note that $-K < 0$, so we focus on the sign of $C = \rho((1 - K\beta/a^*_{12})\tilde{a}'_{21} + \mu')$. We then rewrite the terms μ' and $\tilde{a}'_{21}$ considering that $\tilde{A}' = W^{-1}A'W$, resulting in

$$\mu' = \frac{a^*_{12}}{\beta K} G \left(F + \frac{K}{a^*_{12}} \right) \tag{3.21}$$

and

$$\tilde{a}'_{21} = -\frac{a^*_{12}}{\beta K} GF\,, \tag{3.22}$$

in which $G = a'_{12}(1-\beta) + a'_{11}\beta$ and $F = (-1+\beta)/\beta + (a'_{22}(1-\beta) + \beta a'_{21})/G$. Based on that, $C < 0$ if $\beta(a'_{21} + a'_{11}) + (a'_{22} + a'_{12})(1-\beta) < 0$. This implies that if all the elements of the matrix A' are negative, then the fixed point ρ^* is asymptotically stable. Thus, a sufficient condition for asymptotic stability can be formulated as

$$a'_{ij} = \left.\frac{\partial a_{ij}}{\partial \rho}\right|_{\rho^*} < 0 \text{ for any } i, j\,. \tag{3.23}$$

Importantly, we note that if this condition is satisfied then the necessary condition for asymptotic stability given by (3.20) is automatically fulfilled. In fact, we can write

$$\frac{\partial \mu}{\partial \rho} = \sum_{i,j=1}^{m} \frac{\partial \mu}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial \rho}\,, \tag{3.24}$$

in which m is the number of states. Considering that the dominant left and right eigenvectors of A have non-negative elements, based on Equation (3.9) [Horn and Johnson, 1985], it follows that $\frac{\partial \mu}{da_{ij}} > 0$ for any i, j, which implies that μ' is the sum of negative terms.

In the next section, based on numerical simulations, we will show that both the necessary and the sufficient conditions, given by Equation (3.20) and Equation (3.23), hold for dimensions higher than two. However, it is important to remark that the formulated sufficient condition (3.23), whatever is the dimension of the cell state network, cannot be completely translated into a generic sign condition on the kinetic parameters of the biological process. A direct estimation is possible only for the cell death rate, γ_i, which appears exclusively in the diagonal term a_{ii}. Considering that γ'_i contributes only to a'_{ii} and that $\partial a_{ii}/\partial \gamma_i = -1$, then γ'_i must be positive. Concerning the generic transition rate derivative, ω'_{ij}, it contributes to both the diagonal term a'_{ii} and the off-diagonal one a'_{ji} with opposite sign, since $\partial a_{ii}/\partial \omega_{ij} = -1$ and $\partial a_{ji}/\partial \omega_{ij} = 1$. Lastly, the sign of the division rate derivatives' contributions, λ'_i, to the off-diagonal and the diagonal terms specifically depend on the probability r^i_i and r^j_i

[3]Two cases can be distinguished: a) if $B^2 < -4K\rho\mu'$, then the eigenvalues are complex conjugate, and b) if $B^2 > -4K\rho\mu'$ then the eigenvalues are real. In both cases, if $B < 0$ the real part of the eigenvalues is negative.

and may result in the same or opposite sign. Thus, it is clear that unless the sign of the elements of the matrix A is directly measured, the knowledge of the sign of the kinetic parameter derivative is not sufficient to check the fulfilment of Condition (3.23), but in this case also the values have to be measured. Also, there might be networks where the derived sufficient condition for asymptotic stability can never be satisfied. However, for some specific cell fate configurations, e.g. where feedback is applied only to the death rates, Condition (3.23) for asymptotic stability can be easily verified by measuring only the sign of the derivative of some of the kinetic parameters.

3.2.2.3 Generic m-state cell network

For cell state networks with more than two states, the strategy followed is to test Conditions (3.20) and (3.23) numerically for a large number of random critical cell state networks (i.e. characterised by $\mu = 0$). To this aim, we solve an optimisation problem for each network formulated to find solutions that violate the sufficient or the necessary condition. The procedure followed is detailed in Algorithm 1.

Based on this approach, we tested a large number of cell state networks, $NN = 10^4$, characterised by $m = 3$ and $m = 4$ states. The results are shown in Figure 3.4 in terms of $\mu' = \partial\mu/\partial\rho$ and μ_J. The testing results shown in the left panel, confirm that if all the elements of A' are negative there are no solutions violating the sufficient condition for asymptotic stability (3.23) since all the solutions are characterised by $\mu_J < 0$ and therefore they are asymptotically stable. Testing results shown in the right panel, verify the necessary condition (3.20), as there are no solutions in the quadrant $\mu' > 0$, $\mu_J < 0$.

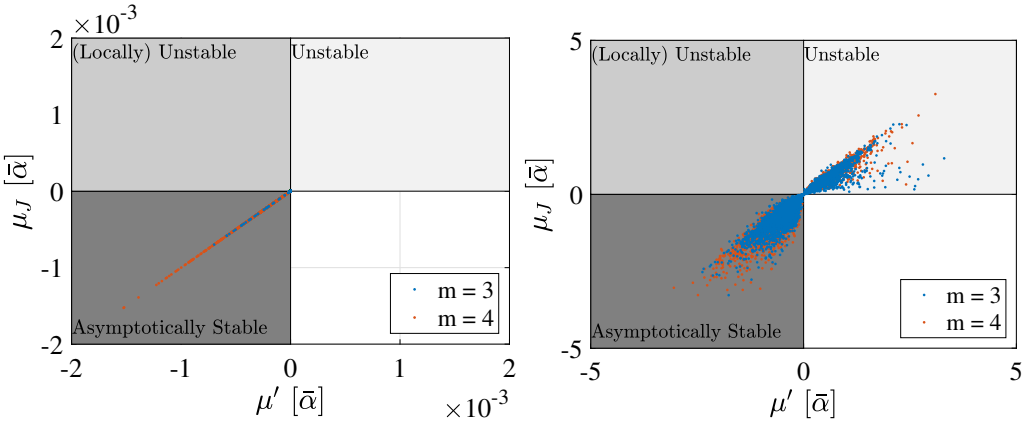


FIGURE 3.4: Numerical tests of the stability conditions for m-state cell networks. For a large number of random systems, the corresponding values $\mu' = \partial\mu/\partial\rho|_{\rho^*}$ and μ_J are optimisation results aimed at violating the tested condition (see Algorithm 1). Concerning the sufficient condition (3.23) (left panel), no solutions can be found in the Unstable and (Locally) Unstable quadrants (when $A' < 0$, asymptotic stability is guaranteed). When testing the necessary condition (3.20) (right panel), solutions characterised by $\mu' > 0$ and $\mu_J < 0$ cannot be found (if $\mu' > 0$, the steady-state is unstable).

Algorithm 1 Optimisation aimed at breaking the sufficient (3.23) and the necessary (3.20) conditions in random dynamical models. The estimation of μ_J is based on the Jacobian matrix J given by Equation (3.5); μ' is instead computed based on Equation (3.24).

Parameter definition: NN, $A_{\min}$, $A_{\max}$, ΔA, $A'_{\min}$, $A'_{\max}$, ϵ_μ;
Variable initialisation: $i = 1$;
while $i \leq NN$ **do**
  /* Definition of a random dynamical system matrix A, with dominant eigenvalue $\mu = 0$ */
  $A_0 = rand([A_{\min}; A_{\max}])$;
  $A^1_{min} = \max(A_0 - \Delta A, A_{min})$, $A^1_{max} = \min(A_0 + \Delta A, A_{max})$;
  $A = \text{argmin}_{A_{min} \leq X \leq A_{max}} |\mu(X)|$; // Local search based on Matlab *fmincon*
  if $|\mu(A)| \leq \epsilon_\mu$ **then**
    /* Single-objective optimisation for testing sufficient condition: can we find $\mu_J > 0$ when the elements of A' are all negative? */
    $A'_S = \text{argmin}_{A'_{min} \leq X \leq 0} - \mu_J(X)$; // Global search based on Matlab *ga* function
    /* Multi-objective optimisation for testing the necessary condition: can we find solutions where $\mu' > 0$ and $\mu_J < 0$? */
    $A'_{Pareto} = \text{argmin}_{A'_{min} \leq X \leq A'_{max}} (\mu_J(X), -\mu'(X))$; // Global multi-objective search based on Matlab *gamultiobj* function
    Extract the solutions, A'_{Pareto^*}, breaking the necessary condition;
    if A'_{Pareto^*} is not empty **then**
      $A'_N = A'_{Pareto^*}$;
    else
      $A'_N = \text{argmin}_{X \in A'_{Pareto}} \sqrt{\mu_J^2(X) + \mu'^2(X)}$; // if solutions breaking the condition are not found, then returns the one closest to the origin ($\mu_J = 0$, $\mu' = 0$)
    end if
    Store A, $\mu(A)$, A'_S, $\mu_J(A, A'_S)$, $\mu'(A'_S)$, A'_N, $\mu_J(A, A'_N)$, $\mu'(A'_N)$;
    i = i+1;
  end if
end while

In principle, the same approach could be applied to systems with more states, i.e. $m > 4$, but this becomes computationally expensive. However, dynamics for higher dimensional systems usually do not show completely new features and are expected to be qualitatively the same as $m = 4$. Given that, we conjecture that the derived conditions hold for any m.

3.3 Robustness of homeostasis

The previous section showed that homeostasis regulation via crowding feedback, under certain conditions, turns the tissue population dynamics stable. However, there are several situations where this regulation mechanism is disrupted. An example is poisoning from drugs, other chemicals or, more in general, environmental cues. In this case, the global behaviour of the tissue might be affected, showing failures or

anomalies in the cell's homeostasis control; also, it could be the case where entire pools of cells die. Another type of disruption is related to cell mutations. In this case, the mutated cell might have a dynamical behaviour that is different from the rest of the tissue, and, in some circumstances, the mutated clone prevails. It is well known that successive mutations are often associated with cancer development [Tomasetti et al., 2013, Colom and Jones, 2016, Rodilla and Fre, 2018].

Therefore, in this section, we want to assess the robustness of the crowding feedback regulation. In general, robustness is a broad concept that sometimes deals with aspects that cannot be mathematically formulated; it also has different, field-related and often controversial interpretations [Nikolov et al., 2007], even within the biological context. In [Greulich and Simons, 2016], for instance, homeostasis's robustness is intended as how a constant parameter dynamical system, which is structurally unstable, becomes stable thanks to the inclusion of the crowding feedback modelling. This assessment is somehow related to the stability analysis shown in Section 3.2 but applied to a particular cell state network. In [Johnston et al., 2007] instead, homeostasis robustness is assessed against feedback mechanism failures in a specific cell dynamical model, representative of the colon crypt.

Given that, we will assess the robustness of homeostasis intended as a) how, in a generic system, homeostasis is maintained after perturbations or complete failure of some of the feedback functions; and b) how a different condition, yet homeostatic, can be achieved if the structure of the cell type condensed network (see definition in Section 2.1) is perturbed, apparently violating the homeostasis requirements in the lineage architecture.

3.3.1 Dysregulation of the feedback mechanism

In Section 3.2, we showed that as long as Condition (3.2) is met, a (dynamic) homeostatic state is guaranteed. This means that if the cell density is perturbed, the (dynamic) homeostatic state is restored. However, in general, there might be factors that disturb the cells' dynamic behaviour. Thus, the system is robust only if it can cope with perturbations and failures in the regulation mechanisms without compromising the homeostatic state. Therefore, in this section, we assess this scenario from a mathematical point of view and show how crowding feedback could compensate for such dysregulations. We will formally address the case of cell-extrinsic factors, i.e. those affecting all the cells in the tissue, and then qualitatively generalise the results to the case of the single-cell mutations.

We first note that $\partial\mu/\partial\rho$ can be written as the sum of different terms, each one related to a specific kinetic parameter α_j. Thus, the stability condition given by Equation (3.2)

results in

$$\frac{\partial \mu}{\partial \rho} = \sum_{j=1}^{L} \frac{\partial \mu}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial \rho} < 0 \text{ for all } \rho \geq 0\,, \tag{3.25}$$

in which L is the total number of kinetic parameters. From this relation, it follows that the feedback applied to a single kinetic parameter α_j can be in principle sufficient to guarantee (dynamic) homeostasis, as long as $\partial\mu/\partial\alpha_j \neq 0$. If this is true, then α_j is a *relevant parameter* of the system, and the (dynamic) homeostatic state can be maintained if the sign of $\partial\alpha_j/\partial\rho$ is opposite of that of $\partial\mu/\partial\alpha_j$. Additionally, we observe that the feedback regulation can still maintain a (dynamic) homeostatic state, despite some relevant parameters playing against stability. This scenario is possible since there might be other terms in the sum (3.25) that overweight these contributions. Thus, for a generic relevant parameter, α_j, the fact that $\dfrac{\partial \mu}{\partial \alpha_j} \dfrac{\partial \alpha_j}{\partial \rho} \geq 0$ does not necessarily compromise homeostasis. Importantly, Condition (3.25) is a sufficient condition, so a (dynamic) homeostatic state might be achieved even if this relation is not fulfilled in the whole range of ρ.

From the above considerations, we conclude that if crowding feedback applies to several relevant parameters, then homeostasis is potentially robust to feedback dysregulation, which may include a simple variation of the feedback function parameters but also perturbation in the feedback functions shape and complete feedback failure (i.e. kinetic parameters remain constant when ρ varies). We also note that the more relevant parameters contributing to stability, the more robust is homeostasis.

We now analyse a numerical example to understand better the impact on homeostasis of the feedback mechanism's dysregulation. The model includes two types of dysregulation: a) variation in the feedback function parameters, where the parameters of the Hill function describing a kinetic rate, $\alpha_i(\rho)$, are perturbed, implying a change in α_i'; and b) failure of the feedback control mechanism, in which the kinetic parameter involved does not adjust with ρ and remains constant. Both types of dysregulation might result, for example, from anomalies in the regulatory pathways.

Considering the same cell state network shown in Figure 3.1, analysed for stability in Section 3.2.1, we focus here on the Asymptotically Stable test case, which is based on the parameters reported in column **AS** of Table 3.1 and dynamics shown in Figure 3.3 (upper panels). Analogous considerations about robustness also apply to the locally unstable case **LU**. Thus, in this homeostatic case, we introduce multiple anomalies in the homeostasis control mechanism and assess their impact on the evolution of the cell dynamics. Two failure scenarios are assessed, as summarised in the $\mathbf{F}_1$ and $\mathbf{F}_2$ columns of Table 3.2 (gray cells indicate the perturbed parameters). More specifically, in the $\mathbf{F}_1$ model, γ_1 is kept constant and equal to its steady-state value in the unperturbed

model. Dysregulation in test case $\mathbf{F}_2$ includes that modelled in $\mathbf{F}_1$, plus failures are applied to parameters ω_{23} and ω_{13}. Whilst ω_{23} is constant and equal to its unperturbed model steady-state value, $\omega_{13}(\rho)$ results in the same steady-state value but an opposite sign of its derivative. We remark that the choice of the model and the dysregulation presented in this example were designed for illustrative purposes and are not related to a specific biological situation.

	AS				$\mathbf{F}_1$				$\mathbf{F}_2$			
α	k	K	n	s	k	K	n	s	k	K	n	s
γ_1	3.09	0.54	2	1	2.44				2.44			
ω_{13}	17.52	1.71	2	-1	17.52	1.71	2	-1	5.99	0.58	2	1
ω_{23}	0.26	1.45	5	1	0.26	1.45	5	1	0.09			

TABLE 3.2: Values of the Hill function parameters used to describe the kinetic parameters in case of crowding feedback dysregulation. The homeostatic unperturbed case (**AS**) corresponds to that analysed for stability in Section 3.2. For the corresponding network and parameters refer to Figure 3.1 and Table 3.1. The generic kinetic parameters, α, is function of the total cell density, ρ, and modelled as $\alpha(\rho) = c + k\rho^n/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') > 0$ and $\alpha(\rho) = c + k/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') < 0$. A common value $c = 0.05$ is assumed. In the dysregulation models, $\mathbf{F}_1$ and $\mathbf{F}_2$, perturbed parameters are highlighted in grey, and a single value indicates a constant model of the type $\alpha(\rho) = \alpha(\rho^*)$. Time unit is arbitrary and therefore omitted.

The cell density time evolution, shown in Figure 3.5 (bottom-left panel), results from the integration of the dynamics based on the explicit Runge-Kutta Dormand-Prince method (Matlab *ode45*) with initial condition corresponding to the homeostatic state. The feedback dysregulation occurs at a time equal to 0. We note that the steady-state is the same as in the unperturbed model, but the dysregulation implies a different dependency $\mu(\rho)$, and thus different values of μ' and μ_J at the steady-state. This is graphically shown in Figure 3.5 (top panels). Concerning the $\mathbf{F}_1$ test case, Condition (3.25) is met in all the range of ρ and the steady-state is asymptotically stable since $\mu_J < 0$. In the $\mathbf{F}_2$ test case instead, μ' and consequently μ_J are positive, so the steady-state is unstable.

Thus, the dysregulation modelled in $\mathbf{F}_1$, although with a degradation of the stability parameters, does not compromise homeostasis. Only including more failures, as in the $\mathbf{F}_2$ test case, the dynamical system behaviour switches to unstable. As confirmed by the time evolution of the cell density shown in Figure 3.5 (bottom-left panel), the tissue approaches a homeostatic state in the first case and grows in the second one. Crucially, depending on the initial conditions and, more in general, on the specific non-linear system, the instability might signify a growing dynamics, as in this case, but also a vanishing one or the convergence to another non-trivial steady-state (a stable one). As shown in the bottom-right panel of the same figure, the homeostatic condition corresponds to a zero dominant eigenvalue of the matrix A ($\mathbf{F}_1$ test case) and the tissue growth to a positive dominant eigenvalue ($\mathbf{F}_2$ test case).

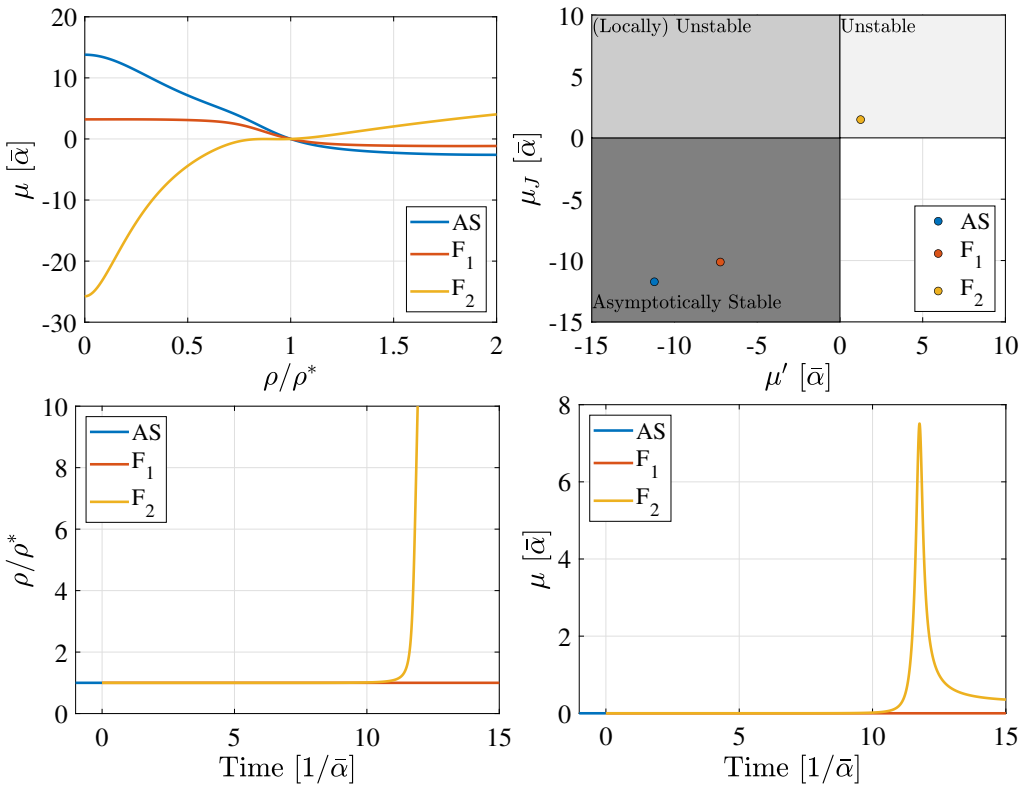


FIGURE 3.5: Feedback dysregulation test cases results. The Asymptotically Stable case, **AS**, analysed for stability in Section 3.2, is modified to include feedback perturbations and failures: these models are indicated as $\mathbf{F}_1$ and $\mathbf{F}_2$. The modelled failure dysregulation does not change the steady-state value but affects the dependency of the dominant eigenvalue on cell density $\mu(\rho)$ (top-left panel) and the stability parameters $\mu' = \partial\mu/\partial\rho|_{\rho^*}$ and μ_J (top-right panel). The expected behaviour based on these stability parameters is consistent with the resulting cell dynamics, which is shown in the bottom panels in terms of the time evolution of the cell density, ρ, normalised by the steady-state, ρ^*, (bottom-left) and the dominant eigenvalue μ (bottom-right). The dysregulation applies at a time equal to 0, and all the simulations start from the homeostatic condition. Whilst model $\mathbf{F}_1$ remains homeostatic, the application of additional failures as in test case $\mathbf{F}_2$ leads the system to an unstable growing condition. Dynamics are scaled by $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

So far, we modelled the feedback dysregulation as acting on a global scale, thus changing the whole tissue's dynamic behaviour. This situation represents a feedback mechanism that is affected by cell-extrinsic signals. If this is the case, any dysregulation applies in the same way to all the cells. However, dysregulation can also act at the single-cell level. This situation can be related to cell mutations. Crucially, cancer's initial development is often associated with the accumulation of successive mutations. As discussed in Section 2.1, the deterministic approximation based on ODEs describes the average behaviour of the cell dynamics but cannot model the single-cell fate choice. Thus, to assess a single cell's impact on tissue dynamics requires stochastic modelling to determine the exact probability of such cell and its progeny eventually becoming extinct. However, as detailed below, we can

extend the derived conclusions to the cell mutation scenario with some careful considerations. A numerical example is instead reported in Appendix A.3.

First, we note that unstable dynamics do not imply the indefinite growth in the cell density. In principle, vanishing dynamics, i.e. converging to a trivial stable steady-state, are unstable as well. Also, considering that the system is non-linear, there might be other (stable) fixed points. Therefore, each model must be specifically assessed, and no general conclusions can be drawn. Furthermore, the mutated clone is always subject to random extinction[4]. Nevertheless, assuming a mutated cell characterised by an unstable growing dynamical model, the mutated clone could at some point prevail over the rest of the tissue. The probability of this occurring might be very low and depends on the system's dynamical behaviour outside the unstable steady-state and on the mutated cells' stochastic fate, i.e. proliferate or die. However, if this is the case, tissue growth will eventually be unavoidable, although the tissue divergence time scale will be much longer than the case where the same dysregulation is applied to all the cells.

3.3.2 Perturbation of the homeostatic lineage architecture

In the previous section, we addressed the case where external factors disrupt homeostasis control in the self-renewing cells of a tissue, yet a robust regulation mechanism maintains homeostasis. However, factors such as injury, poisoning or cell radiation might also affect homeostasis in other ways. An example is when stem cells are depleted from the tissue. In this context, many studies about tissue regeneration after injury report evidence of cell plasticity, which is the cell's ability to change identity. Cell *dedifferentiation* is just an example where differentiated cells return to an undifferentiated state as a response to tissue damage. Lineage tracing experiments confirmed this feature in vivo in several cases [Tata and Rajagopal, 2016, Merrell and Stanger, 2016, Tata et al., 2013, Puri et al., 2015]. In other situations, instead, quiescent stem cells activate as a tissue response to an injury or other external factors.

Therefore, in this section, we study from a mathematical standpoint how the crowding feedback regulates the homeostasis in tissues where the lineage architecture is perturbed. Importantly, we recall that, in Section 2.2, we derived requirements for a homeostatic system, based on which the stem cell type (and only one of this type) must stay at each apex of the cell lineage. Also, multiple lineages can be present in the same tissue, but they must be disconnected. We, therefore, focus on Conditions (nl.iii), (nl.ii) and (nl.iv), which violation could represent the response of a tissue to an injury,

[4]The only exception is the case where the self-renewing strategy is based on invariant asymmetry (see Section 1.2 and its generalisation in Chapter 4). However, in this case, homeostasis regulation is not applicable since the dominant eigenvalue of the dynamical system is equal to one independently of the values of the kinetic parameters.

and the crowding feedback a mechanism to restore homeostasis. However, we do not consider scenarios where condition (nl.i) is not met since the feedback regulation cannot compensate for hyperproliferating cells over a long period. If that is the case, the cells under the feedback regulation will decrease in number and eventually disappear to balance the cell number growth of the hyperproliferating types that, at some point, will prevail.

We consider a generic system composed of several cell types, topologically ordered. Based on the definition of cell type given in Section 2.2 and on the crowding feedback modelling described in Section 3.1, each ith cell type dynamics depend only on itself and the upstream jth cell types, where $j = 1, ..., i-1$. We can further simplify this model considering that the upstream lineage architecture fulfils the homeostasis requirements and that globally the upstream system is at its steady state. Thus, the upstream contribution can be modelled as a constant influx of cells and therefore, we can isolate the dynamics of the ith cell type (we omit hereafter the superscript i), that results in

$$\frac{d}{dt}\boldsymbol{\rho} = A(\rho)\boldsymbol{\rho} + \boldsymbol{u}\,, \tag{3.26}$$

in which $\boldsymbol{u} = \sum_{j=1}^{i-1} \boldsymbol{A}^{ij}(\rho^j)\boldsymbol{\rho}^j$ represents the cell influx, i.e. a constant vector with non negative elements that is function of the cell densities in the cell states of the jth cell type, $\boldsymbol{\rho}^j$, and the connections between cell states of the jth type with those of the ith type, A^{ij}.

We first observe that if $\boldsymbol{u} = \mathbf{0}$, then results about the existence of a steady-state, ρ^*, provided in Section 3.1 apply, meaning that the system at the steady-state must be self-renewing, i.e. with growth parameter $\mu^* = \mu(\rho^*) = 0$. Instead, if $\boldsymbol{u}$ is not zero, then a steady-state, $\boldsymbol{\rho}^*$, exists if $A(\rho^*)$ is invertible, that is $\boldsymbol{\rho}^* = -A(\rho^*)^{-1}\boldsymbol{u}$. Furthermore, such a steady-state only exists if $A(\rho^*)$ has negative eigenvalues[5], which means that the cell type is transient, i.e. with growth parameter $\mu^* < 0$. For assessing the stability of such a steady-state, we need to evaluate the eigenvalues of the Jacobian matrix of the linearised system around it, which is based on Equation (3.4) (or Equation (3.5) in a compact form). The steady-state is stable if they all have real part negative, and thus, if the largest real part of the eigenvalues, μ_J, is negative.

To visualise the two possible perturbed scenarios, we consider a simple homeostatic system sketched in Figure 3.6. The black box encloses the homeostatic system, which comprises a stem cell type (orange), X_S, and a committed cell type (green), X_C. In scenario $\mathbf{D}_1$, an upstream stem cell type, X_Q, is attached to X_S. For instance, this scenario models the activation of a pool of quiescent stem cells. In scenario $\mathbf{D}_2$,

[5] At the steady-state, the Equation (3.26) multiplied by the left dominant eigenvalue v, results in $0 = vA(\rho^*)\boldsymbol{\rho}^* + v\boldsymbol{u} = \mu^* I\boldsymbol{\rho}^* + v\boldsymbol{u}$. Considering that $v\boldsymbol{u}$ is non negative with at least one positive element, then the steady-state ρ^* exists only if $\mu^* < 0$.

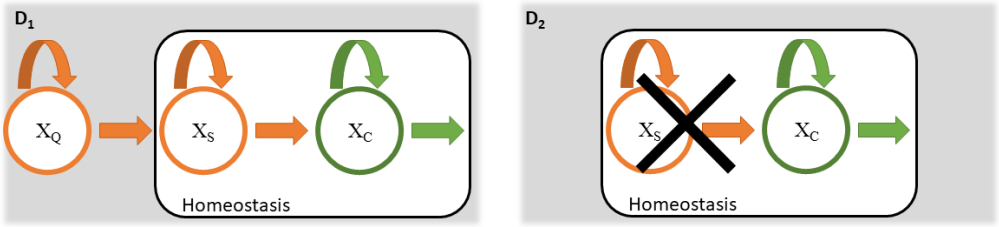


FIGURE 3.6: Sketch of the scenarios analysed to assess the homeostasis robustness against perturbation of the lineage hierarchy. A homeostatic system enclosed in the black box is composed of two cell types: a stem cell type, X_S, (orange) and a committed cell type, X_C, (green). In the unperturbed homeostatic scenario, X_S is self-renewing, that is, characterised by growth parameter at the steady state $\mu^* = 0$, and X_C is transient, with growth parameter at the steady state $\mu^* < 0$. The system is perturbed by adding an upstream self-renewing type, X_Q, in the test case $\mathbf{D}_1$, breaking conditions (nl.iii) and (nl.iv) (left) and removing the stem cell type X_S in the test case $\mathbf{D}_2$, violating requirement (nl.ii) (right).

instead, the stem cell type X_S is removed, and the loss of the stem cell pool might represent the consequence of poisoning that targets a specific cell type or the stem cells removal due to radiation. Although this sketch only includes two cell types, the considerations below apply to more complex models since we assume that additional committed downstream cell types consistently adapt their behaviour.

Now, if homeostasis is not regulated, i.e. A is a constant matrix, then the initial self-renewing cell type, X_s in $\mathbf{D}_1$ (for which $\mu = 0$), perturbed by adding the upstream cell influx, i.e. $\boldsymbol{u} \neq 0$, results in a growing dynamics. Consistently, the initial committed cell type, X_c in $\mathbf{D}_2$ (for which $\mu < 0$), perturbed by removing the upstream cell influx, i.e. $\boldsymbol{u} = 0$, results in vanishing dynamics. Instead, under feedback regulation, cell dynamics are not marginally stable, and there might be two stable fixed points, one corresponding to a self-renewing condition, $\rho^{*(S)}$, achieved when $\boldsymbol{u} = \mathbf{0}$, and the other to a transient one, $\rho^{*(T)}$, applicable in case there is an influx of cells $\boldsymbol{u} \neq \mathbf{0}$. Importantly, whilst $\rho^{*(S)}$ is a feature of the cell type dynamical model, the steady-state $\rho^{*(T)}$ depends also on the particular value of $\boldsymbol{u}$. Thus, in principle, the same cell type is compatible with a steady and a vanishing behaviour, maintaining a homeostatic state with and without cell influx. Crucially, we note that the details of the inner cell state structure (i.e. the cell state network) within the cell type is not of interest, as long as a proper stabilising crowding feedback mechanism regulates it.

The following numerical example illustrates this situation. We focus on the dynamics of a single cell-type, based on the Asymptotically Stable dynamical model (**AS**) analysed for stability in Section 3.2 and robustness to feedback dysregulation in Section 3.3.1. We also choose a constant non-negative $\boldsymbol{u} = \bar{\boldsymbol{u}}$, to model for the cell influx. For such model, two fixed points, $\rho^{*(S)}$ and $\rho^{*(T)}$, exist respectively for $\boldsymbol{u} = \mathbf{0}$

and $\boldsymbol{u} = \bar{\boldsymbol{u}}$. Their properties in terms of $\mu(\rho^*)$, $\mu' = \partial\mu/\partial\rho|_{\rho^{*(S/T)}}$ and μ_J are graphically represented in Figure 3.7. The two steady-states are asymptotically stable since $\mu_J < 0$. Based on this single-cell type model, we study two test cases, $\mathbf{D}_1$ and $\mathbf{D}_2$, as representative of the scenarios depicted in Figure 3.6. In particular, in the $\mathbf{D}_1$ test case, we model the dynamics of X_S where $\bar{\boldsymbol{u}}$ represents the contributions from X_Q. In test case $\mathbf{D}_2$, instead, we model X_C for which the contribution of $\bar{\boldsymbol{u}}$, related to X_S is removed.

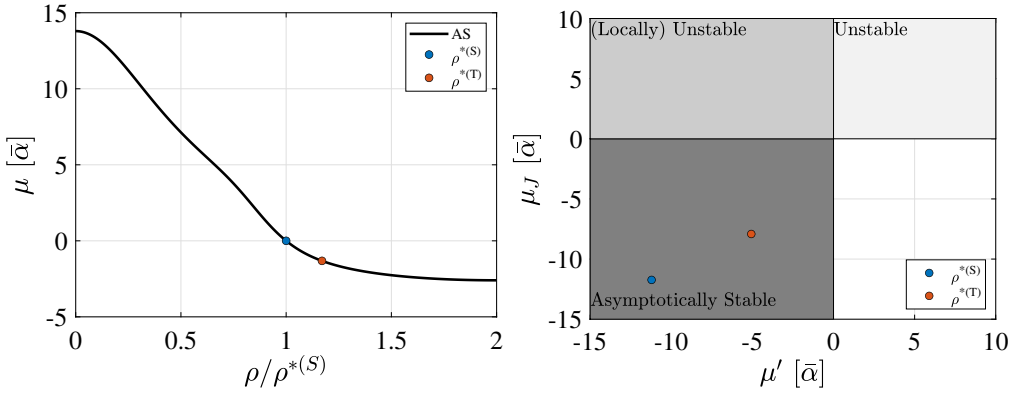


FIGURE 3.7: Dynamic parameters of a single cell type, modelled considering the cell state network and parameters of the Asymptotically Stable test case (**AS**), analysed for stability in Section 3.2 and for robustness to feedback dysregulation in Section 3.3.1 (see Figure 3.1 and Table 3.1). The dynamics of this cell type, based on (3.26), is representative of a self-renewing cell type when $\boldsymbol{u} = \mathbf{0}$, and of a transient one, when $\boldsymbol{u} = \begin{pmatrix}0.02 & 0.07 & 0.06\end{pmatrix}^T$. The dependency of the dominant eigenvalue on cell density, $\mu(\rho)$, (left) and the stability parameters, $\mu' = \partial\mu/\partial\rho|_{\rho^{*(S/T)}}$ and μ_J, (right), show that these two conditions result in a different steady-state, $\rho^{*(S)}$ and $\rho^{*(T)}$, which are both asymptotically stable. Values are shown in $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

The time evolution of the total cell density ρ, normalised by the unperturbed homeostatic value (top panel) and the dominant eigenvalue μ (bottom panel), are shown in Figure 3.8. The figures on the left refer to the $\mathbf{D}_1$ model, whilst those on the right to the $\mathbf{D}_2$. In both cases, the integration of the ODEs for the cell densities dynamical model, given by Equation (3.26), starts from the homeostatic condition, and it is based on the explicit Runge-Kutta Dormand-Prince method (Matlab *ode45* function). At a time equal to 0 we apply the perturbation of the system by switching from $\boldsymbol{u} = 0$ to $\boldsymbol{u} = \bar{\boldsymbol{u}}$ in one case and from $\boldsymbol{u} = \bar{\boldsymbol{u}}$ to $\boldsymbol{u} = \mathbf{0}$ in the other. A different yet stable steady-state is achieved as expected in both cases since the two fixed points of the model, which are sufficiently close, are asymptotically stable.

This result confirms that, under crowding feedback regulation, a cell type might switch from self-renewing to committed if upstream stem cells become active. Conversely, if the self-renewing cell type disappears, the initially committed type might switch to self-renewing, meaning that it actually becomes a stem cell type. We must remark that this switching behaviour does not hold in general since the

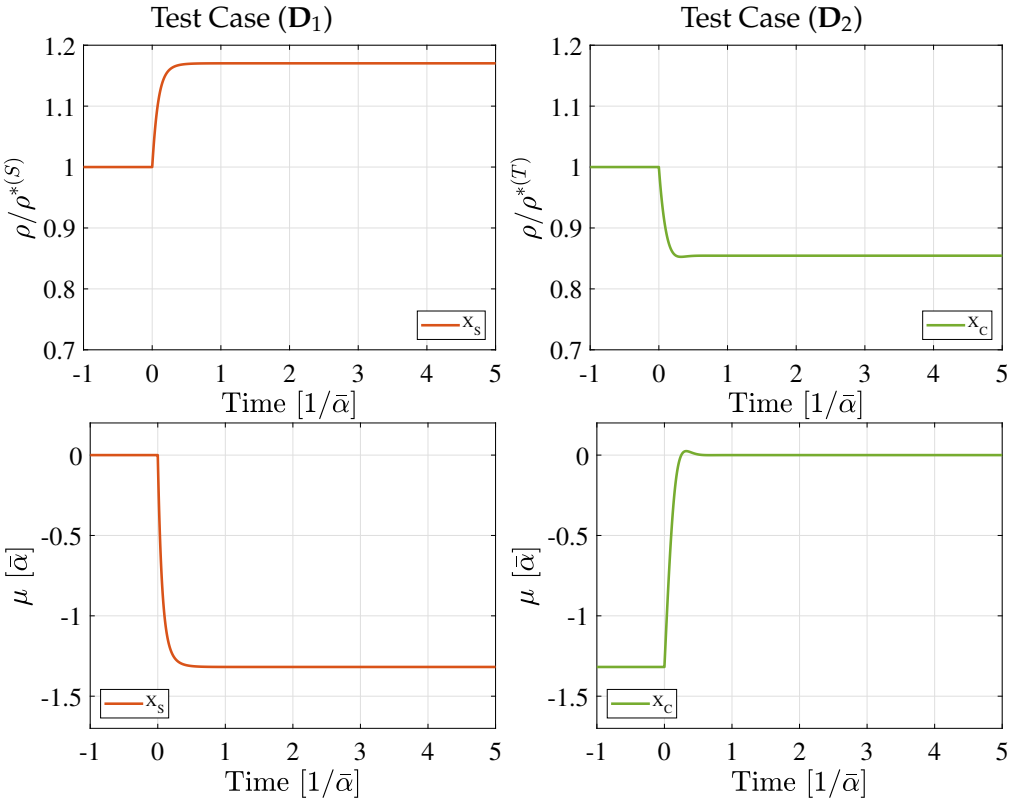


FIGURE 3.8: Lineage architecture perturbation test cases results. The two scenarios modelled are sketched in Figure 3.6 and the details of the stability properties of the dynamical systems are reported in Figure 3.7. The cell dynamics are shown as the time evolution of the cell density, ρ, normalised by the initial steady-state, $\rho^{*(S/T)}$ (top panels), and the dominant eigenvalue (bottom panels). In the $\mathbf{D}_1$ test case (left panels), the dynamical system models the stem cell type, X_S. Initially, the dynamics are based on $\boldsymbol{u} = 0$, implying that the cell type is self-renewing, i.e. $\mu = 0$. At a time equal to 0, an upstream self-renewing cell type, modelled as a constant influx of cells $\boldsymbol{u} = \bar{\boldsymbol{u}}$, is added. As a consequence, X_S switches to a transient cell type where $\mu < 0$. In the $\mathbf{D}_2$ test case (right panels), the opposite case is modelled. Here, the dynamics represent those of an initially committed cell type, X_C, where $\boldsymbol{u} = \bar{\boldsymbol{u}}$ and consequently $\mu < 0$. When the stem cell type X_S is removed, that is, $\boldsymbol{u} = 0$, X_C becomes self-renewing with $\mu = 0$. Dynamics are scaled by $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

dynamics are non-linear. Therefore, each particular system must be specifically assessed.

From a biological point of view, the $\mathbf{D}_2$ scenario, where the tissue reacts to the depletion of stem cells, is of particular interest. If dedifferentiation is a known mechanism for tissue regeneration, here we showed that, in the presence of homeostasis regulation via crowding feedback, a possible alternative response of a tissue to a sudden disruption of the homeostatic condition is to switch the cell type behaviour from transient to self-renewing, without changing the cells' identity. In practice, this means that these cells do not need to change their internal state, i.e. the gene expression, to revert to an undifferentiated self-renewing type and restore the initial homeostatic state. Instead, their internal regulation adjusts the proliferation,

transition and death rates in response to the perturbed environment, and, in doing so, a new stable steady-state is reached. We, therefore, call *quasi-dedifferentiation* this situation, where differentiated cells can sustain homeostasis acquiring self-renewal capability without changing state. At the moment, there is no evidence supporting or discarding quasi-dedifferentiation, which remains a plausible scenario from a mathematical point of view. From an experimental perspective, analysing the cell identities in a tissue achieving homeostasis after completely depleting stem cells could confirm quasi-dedifferentiation.

3.4 Conclusions

This chapter extended the homeostasis model proposed in Chapter 2, including a regulation mechanism. We recall that homeostatic dynamics must respect strict rules in the cell types architecture. That is, the stem cells must stay at the apex of the lineage hierarchy and only there. However, in a constant parameter cell fate model, any slight imbalance in the self-renewing condition leads to an unlimited growth or shrinking of the tissue. Given that this is not plausible in a real biological context, the goal of a homeostasis control mechanism is to guarantee a stable and robust condition in the case of perturbations.

In particular, in this chapter, we assessed the crowding feedback, where cells sense the cells density and adjust their proliferative, differentiation and death potential to maintain homeostasis. In our mathematical model, the kinetic parameters of the cells' dynamics become dependent on the total cell density, ρ. Thus, the linear dynamics describing the average cell numbers, studied in Chapter 2, turns into a non-linear one in the cell densities, provided that the elements of A are a function of ρ (see Equation (3.1)). Consistently, the dominant eigenvalue of A is a dynamic quantity $\mu(\rho)$. Therefore, in this model, homeostasis corresponds to a non-trivial fixed point ρ^* of the non-linear system of ODEs, that is related to the condition $\mu(\rho^*) = 0$.

Based on this model, we first assessed the stability of the steady-state ρ^*, providing a condition on the variation of μ with cell density, $\partial\mu/\partial\rho$, that guarantees the existence of a (dynamic) homeostatic state, (see Equation (3.2)). This condition, which can also be associated with a simple sign criterion on the dependency of the kinetic parameters from the density, ensures that a steady-state is either asymptotically stable, i.e. homeostasis in a strict sense, or only locally unstable, i.e. dynamic homeostasis. In the first case, the feedback regulation restores a homeostatic condition if perturbed. In the latter, the cells density is not constant yet confined around the homeostatic condition.

We then further studied the steady state's stability by linearising the system and analysing the sign of the largest real part of the Jacobian matrix's eigenvalues, μ_J. Based on the analytical derivation of μ_J in single and two-state systems, we derived a

sufficient condition and a necessary one for asymptotic stability. Such conditions were numerically tested in three and four-state systems, suggesting that their applicability also holds in high dimensional cases. Crucially, the necessary condition is related to the variation of the dominant eigenvalue with cell density at the steady-state, μ', which must be negative. This necessary condition is a less restrictive requirement than Condition (3.2) for the existence of a (dynamic) homeostatic state, which requires a negative derivative for any ρ. Concerning the sufficient condition for asymptotic stability, the sign of A element derivatives at the steady-state is involved. Despite the fact that there might be cell fate models that can never meet this condition (e.g. in the case of cell states for which only cell state transitions occur, the corresponding diagonal and off-diagonal elements of A' have, by construction, opposite sign), it remains a helpful criterion to determine tissue dynamics' stability experimentally.

In the last section of this chapter, we discussed the implications of the derived stability conditions, assessing the robustness of homeostasis regulated via crowding feedback to feedback dysregulation and lineage architecture perturbations. We first focused on a single renewing cell type, including multiple feedback failures and perturbations. Since the stability parameter, μ', is the sum of multiple contributions, we showed how the system might remain homeostatic despite some of them playing against stability. Notably, the same conclusion about homeostasis robustness applies if the dysregulation affects a single cell. In this case, dysregulation might represent the cell mutation that is often a critical factor in cancer development. In particular, as long as the system is stable, the mutated cell modified dynamical behaviour does not alter the tissue. However, the steady-state becomes unstable when the cell dysregulation cannot be compensated, and the whole tissue might be affected. Here, the random extinction of the mutated cell plays a significant role, giving the tissue a chance to remain homeostatic if the mutated clone goes extinct by chance.

Lastly, concerning perturbations in the lineage architecture, we demonstrated through an illustrative example how cells of a given type and regulated via crowding feedback switch behaviour, from self-renewing (i.e. stem cell) to transient (i.e. committed) and vice versa, depending just on the influx of cells from the upstream types. Importantly, self-renewal does not need to be an intrinsic property of a cell type since any cell type that is at the apex of a lineage hierarchy may acquire this property by interacting with its environment. Based on these results, we also proposed the quasi-dedifferentiation, a condition where a committed cell type becomes self-renewing after the complete depletion of the pool of stem cells. In this case, the tissue does not turn back to its initial state, but homeostasis can still be maintained. In the context of tissue regeneration after injury, this mechanism is a mathematically plausible alternative response to the known cell dedifferentiation process, where cells change their state to regenerate the tissue.

Chapter 4

Qualitative features of lineage tracing dynamics in homeostasis

In Chapter 2, we have studied the conditions for achieving homeostasis in a generic cell fate model and defined rules based on which we can a priori exclude all the non-homeostatic dynamics. Nevertheless, the cell fate model's definition given experimental data remains a complex task. Therefore, this chapter assesses cell dynamics to identify robust criteria for selecting candidate fitting models directly on qualitative features of experimental data. For this purpose, we consider the same type of experimental data expected for the study case, which are the single-cell transcriptome and clonal statistics based on lineage tracing.

Concerning transcriptome data, we propose a method for detecting disconnected pools of stem cells by comparing tissue samples with lineage tracing ones, where only cells that are the progeny of an initially labelled cell subpopulation are sequenced. Additionally, based on the analysis of clonal statistics, we show that, in homeostatic tissues, models of cell fate dynamics can be categorised into two *universality classes*, whereby models of the same class predict the same clonal statistics under asymptotic conditions. Those classes relate to generalisations of the canonical asymmetric and symmetric stem cell self-renewal strategies, presented in Section 1.1, and are distinguished by a conservation law.

The research outcomes about the self-renewing strategy identification based on clone lineage tracing data, reported in Section 4.3, are published in [Parigini and Greulich, 2020].

This chapter is organised as follows: the modelling of lineage tracing data is described in Section 4.1; the assessment of the stem cell types via trasncriptome data and that of the self-renewing strategy via clonal statistics are reported respectively in Section 4.2 and Section 4.3; conclusions are given in Section 4.4.

4.1 Modelling of lineage-tracing data

In Section 1.3.1, we reviewed the lineage-tracing experiment, a technique used to identify the progeny of single-cells or a specific sub-population of cells. Based on an inheritable genetic marker, which is retained after cell division or state change, lineage tracing allows extracting essential information about cell lineage over time.

To model the lineage-tracing data in a generic tissue, we start from the generic cell dynamics based on a continuous-time stochastic process presented in Chapter 2. This model is characterised by an arbitrary set of cell state, X_i, for $i = 1, ..., m$, having fate choices of the type (2.1)-(2.3). We now recall that at the tissue level, such dynamics are well described by the average numbers of cells, $\bar{n}_i$, and follow the system of Ordinary Differential Equations given by Equation (2.5), which, in compact form, is written as (2.6). More specifically, when modelling the lineage-tracing average dynamics, the system of ODEs is integrated starting from a single cell in the initial labelled cell state. However, as explained later in Section 4.3.1, when analysing clonal data, the random extinction of the clones, which depends on the details of the stochastic process, needs to be taken into account.

We further remember from Chapter 2 that the generic cell fate model, also seen as a cell state network, could be modelled as a cell type condensed network by grouping the states forming the SCCs into nodes, which are associated with a cell type (see Figure 2.1). Each isolated cell type was then classified based on its growth parameter μ, i.e. the dominant eigenvalue of the corresponding SCC in the cell state network, resulting in one of three possible long-term dynamics: growing, vanishing or steady (see Table 2.1). Finally, we showed how this classification, combined with the topological arrangement of the cell types, determines the stability of the global system. Based on this, the conditions for the existence of a homeostatic state are that, at the apex of each lineage, i.e. the condensed cell state network, there must be a self-renewing cell type, i.e. $\mu = 0$, while all SCCs downstream of the former must be transient, i.e. $\mu < 0$, (see Figure 2.3). Since we are interested in modelling cell dynamics in homeostasis, in this chapter, we will only consider homeostatic cell fate models, that is, models fulfilling these rules.

Concerning the non-linearities introduced in the dynamics by homeostatic regulation mechanisms, we first remind that, as shown in Section 2.2.3, the rules for a homeostatic cell fate model derived in the linear case also hold in the non-linear one if the system stays at the steady-state (i.e. if the regulation mechanism is stabilising). In this case, the kinetic parameters are not constant and depend on the total number of cells, and the population dynamics have the form of Equation (3.1) (where we can replace $\boldsymbol{\rho}$ by $\bar{\boldsymbol{n}}$ if there are no variations in the tissue volume). Now, we must consider that the stochastic fluctuations in the number of cells in the traced clones, n, do not affect the total number of cells in the tissue, $\bar{n}$, which remains constant since we

assume to be at the steady-state. In other words, if there is a stabilising regulation mechanism, $\bar{n} = \bar{n}^*$ even if n changes locally. Crucially, at the clonal level, the clonal statistics of spatial models that include cell-extrinsic regulation of the cell fate, such as models of the voter type [Clifford and Sudbury, 1973], are, in the long term, the same as for the corresponding models which do not, i.e. branching processes [Haccou et al., 2005]. An exception is the one-dimensional arrangements of cells[1], as shown in [Bramson and Griffeath, 1980, Klein and Simons, 2011]. Therefore, considering that we are interested in measures in the long-term and focusing only on tissues with two- or three-dimensional arrangements of dividing cells, like epithelial sheets and volumnar tissues, we can model the cell fate dynamics independent of the cell environment. This choice translates into considering the kinetic parameters, $\lambda_i, \omega_{ij}, r_i^{jk}, \gamma_i$ as constants, meaning that feedback can, in effect, be neglected.

4.2 Stem cells type identification via transcriptome data analysis

In Section 1.3, we showed that together with clone lineage tracing, the single-cell RNA-sequencing is typically used in the study of the lineage hierarchy. The scRNA-seq is an innovative experimental technique primarily aimed at identifying cell identities through the clustering analysis of the RNA content in every single cell. The following shows that combining these two experimental techniques could help identify multiple stem cell types based on purely theoretical considerations. We recall that the stem cell types correspond here to disconnected self-renewing SCCs that are at an apex of the lineage hierarchy (see definition in Section 2.1.2).

More specifically, the processing and analysis of scRNA-seq data give, among other results, the identification of the clusters intended as groups of cells sharing the same identity. The *cluster size* corresponds to the number of cells that form a particular cluster. We now assume to have measures of the size of the clusters from samples of cells of the whole tissue and samples containing only lineage traced cells. Concerning the latter, only the progeny of an initially labelled subpopulation of stem cells (via cell-type specific genetic marker) is sequenced. For doing so, we assume to use of the same technology employed in the clonal lineage tracing, that is, the Cre-lox technology described in Section 1.3.1. That means that cells can be sorted based on the expression of the green fluorescent protein so that lineage traced cells are separated from the others. The average cluster size in the lineage traced population is then measured by sequencing only these cells. We observe that, in contrast to clones' lineage tracing, information on individual clones forming the marked cell population

[1] The mean clone size as a function of time also slightly differs for two dimensional systems, but only by a logarithmic pre-factor which approaches a constant for large time [Klein and Simons, 2011].

is not gathered (we only need to distinguish if a cell is part of the lineage traced population or not). Therefore, clones do not necessarily need to be spatially separated or distinguished via other means (e.g. different colouring markers, viral barcodes). However, for clarity in the mathematical development, we will consider that lineage tracing data comes from clones. Despite this being a possibility, it is not strictly required from the experimental point of view.

For a better understanding of how these measures can tell us about the minimum number of stem cell types in a tissue, we first notice that there are only two possible scenarios[2]: a) one where there is only one stem cell type, the one that is initially labelled for tracing; and b) there is at least one stem cell type that is not initially labelled in addition to the labelled one. An illustrative example of these two conditions is sketched in Figure 4.1. Here, only the stem cell type initially labelled is highlighted with a green border. Focusing on b), the black circle represents a stem cell type that is not initially labelled. For simplicity, only one of this type is represented here, but, in principle, there could be more. This type is not connected by any path to the labelled one; otherwise, the rules for homeostasis would be violated. Hence, stem cells of types not initially labelled will never show the green fluorescent marker. Instead, we expect that the progeny of the labelled cell population would show the marker. Therefore, the border of these cell types is shown in green, while the others are shown in black. In this example, we considered just two cell clusters highlighted with a yellow or a red filling. In general, the cell classification from the scRNA-seq analysis can have more resolution, up to one cluster for each cell state. However, rare cells, like stem cells can be, are usually hard to identify.

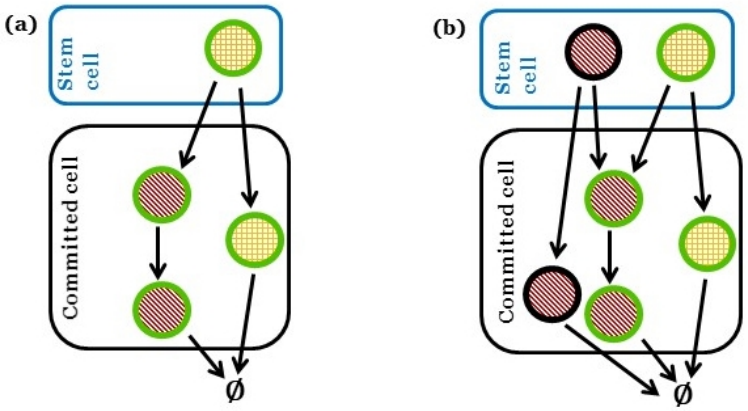


FIGURE 4.1: Examples of two homeostatic cell type networks representative of the two possible scenarios analysed. In (a), there is only one stem cell type, the one that is initially labelled, which is shown with green border. In (b), in addition to the labelled stem cell (green border), there is one stem cell type that is not labelled (black border). The progeny of the labelled cell population, i.e. cell types connected to the green ones, is shown in green. The yellow or red filling indicates the cell cluster.

[2]Here we assume to label only one cell type, but the same reasoning applies even if more than one cell type is initially labelled.

From a mathematical point of view, in a generic cell fate model, assume that we have an arbitrary cluster number, k, and that we have measures of their relative size, $s_i = N_i / \sum_j N_j$ for $i = 1, ..., k$, in which N_i is the number of cells in the ith cluster. We want to compare the measures of cells sampled from the tissue, s_i^* with those of traced cells, $s_{i,s}$. Before doing this, we must recall that the steady-state of the generic non-linear dynamical model given by (2.12) is the same as that of the linear model given by Equation (2.6). By assuming a stabilising regulation mechanism, homeostasis is maintained and therefore, we can apply the linear modelling of the tissue dynamics also to the non-linear case provided that we are only looking at the steady-state (the detailed discussion about the applicability of a constant parameter model is given in Section 4.1).

Focusing now on the tissue data measure, this is related to the steady-state, $\bar{\boldsymbol{n}}^*$, of the dynamical model given by (2.6) (or equivalently Equation (2.12)), that is,

$$s_i^* = \frac{\sum_{j \in I} \bar{n}_j^*}{\bar{n}^*}\,, \tag{4.1}$$

in which I includes all the states associated with the ith cluster and $\bar{n}^*$ is the total average number of cells.

Concerning lineage traced cells, we first note that, if there are a fraction of symmetric divisions, we will only measure the surviving clones[3]. Therefore, $s_{i,s} = \sum_{j \in I} \bar{n}_{j,s} / \bar{n}_s$, in which $\bar{n}_{j,s}$ and $\bar{n}_s$ are the mean numbers of cells in the surviving clones only, respectively in the jth state and the total. However, at any time, the total number of cells in the jth state can be written as $N_j = \bar{n}_{j,s} M_S = \bar{n}_j M$, where M_S and M are respectively the number of the surviving and the total number of clones, and $\bar{n}_j$ is the average number of cells in the jth state in all the clones, including the extinct ones[4]. Hence, $\bar{n}_{j,s} = \bar{n}_j M / M_S$. The same reasoning applies to the total number of cells $N = \bar{n}_s M_S = \bar{n} M$, so that $\bar{n}_s = \bar{n} M / M_S$. Based on this, the relative size of the cluster in the clonal data results in

$$s_{i,s} = \frac{\sum_{j \in I} \bar{n}_j}{\bar{n}} = s_i\,, \tag{4.2}$$

meaning that it does not depend on the level of clonal extinction. We now notice that $\bar{\boldsymbol{n}}$ corresponds to the solution of the dynamical model (2.6) assuming an initial single cell in a state that is part of the self-renewing cell type, i.e. corresponding to the initially

[3] From a practical point of view, it is possible only to measure what is visible in the tissue. Cells that die and, in general, clones that go extinct cannot be measured unless constantly tracked in time.

[4] Omitting for readability the subscript j, assume there are $N(t)$ cells in a given state and the number of cells in M clones are monitored over time. Then, the average number of cells per clone is simply $\bar{n}(t) = N(t)/M$. If a fraction of those clones goes extinct at some point in time, the mean number of cells in the surviving clones only results in $\bar{n}_s(t) = N(t)/M_s(t)$, where $M_s(t)$ is the number of surviving clones only. This is because, by definition, there are no cells in the extinct clones. Thus, at any time, $N(t) = \bar{n}(t) M = \bar{n}_s(t) M_s(t)$.

labelled stem cell type in the experiment (or, equivalently, model (2.12)[5]). Since we are in homeostasis, in the long term, even though $\bar{\boldsymbol{n}}_s$ might increase with time if there are symmetric divisions, $\bar{\boldsymbol{n}}$ converges to a steady-state $\bar{\boldsymbol{n}}^{**}$, and so does the ratio s_i. Therefore, the measure from the the clonal lineage tracing data for a sufficient large time is

$$s_{i,s} \to \frac{\sum_{j\in I} \bar{n}_j^{**}}{\bar{n}^{**}} = s_i^{**} \,. \tag{4.3}$$

We compare now s_i^{**} and s_i^{*}. We observe that $\bar{\boldsymbol{n}}^{*}$ and $\bar{\boldsymbol{n}}^{**}$ are different since they are steady-state conditions of a marginally stable dynamical system related to two different initial conditions. From [Greulich et al., 2019], such steady-state condition has the form

$$\bar{\boldsymbol{n}} = \sum_{r\in I_R} \alpha_r \tilde{\boldsymbol{\phi}}_r \,, \tag{4.4}$$

in which I_R includes all the self-renewing (i.e. critical) SCCs of the system, $\tilde{\boldsymbol{\phi}}_r \geq 0$ is related to the dominant eigenvector associated to the rth critical SCC[6], and the values $\alpha_r \geq 0$ determine the steady-state and specifically depend on the initial conditions. We will hereafter indicate as α_r^{*} and α_r^{**} the values respectively in the tissue and the lineage tracing steady-state.

We observe now that the initial condition in the dynamics of the lineage traced clones is one cell in only one self-renewing SCC, the one related to the initially labelled stem cell subpopulation. Without loss of generality, we will call it $r = 1$. Given that the remaining self-renewing SCC are disconnected from the 1st SCC, then $\alpha_r^{**} = 0$ for $r \in I_R, r \neq 1$. This means that the steady-state given by Equation (4.4) can be written as

$$\bar{\boldsymbol{n}}^{**} = \alpha_1^{**} \tilde{\boldsymbol{\phi}}_1 \,, \tag{4.5}$$

[5]In the non-linear case, since we are modelling the dynamics of the clone embedded in the tissue, the term $A(\boldsymbol{n})$ in Equation (2.12) actually corresponds to $A(\boldsymbol{n}_T)$, where $\boldsymbol{n}_T$ is the cell number in the tissue and not in the clone. In homeostasis, despite the cell number in the clone changes, that in the tissue remains constant and equal to the steady-state, $\boldsymbol{n}_T^{*}$. Hence, $A(\boldsymbol{n}_T) = A(\boldsymbol{n}_T^{*})$ is a constant too.

[6]Considering the partitioning in block as in Equation (2.9), from [Greulich et al., 2019], the steady state associated with block m is

$$\bar{\boldsymbol{n}}_m = \begin{cases} 0 & \text{if} \quad m \text{ is upstream of or disconnected from } I_R \\ \alpha_m \boldsymbol{\phi}_m & \text{if} \quad m \in I_R \\ A_m^{-1}\left[\sum_{r\in I_R} P_{mr}\alpha_r \boldsymbol{\phi}_r\right] & \text{if} \quad m \in I_C \end{cases} ,$$

in which $\boldsymbol{\phi}_i$ is the dominant eigenvector of the ith critical block, and P_{ji} is a combination of the product of off-diagonal blocks C_{ln} and the inverse of the diagonal block A_n as defined in Equation (2.9), that accounts for all the paths from the ith critical SCC to the jth subcritical one (see Theorem 1 in [Greulich et al., 2019]). For the subcritical blocks, we can rearrange the steady state term as $\bar{\boldsymbol{n}}_m = \sum_{r\in I_R} \alpha_r \left[A_m^{-1} P_{mr}\boldsymbol{\phi}_r\right]$. Thus, globally, defining $\tilde{\boldsymbol{\phi}}_r$ as a vector filled with the dominant eigenvector ϕ_r and $A_m^{-1}P_{mr}\boldsymbol{\phi}_r$ in the corresponding components, and zero otherwise, we can write $\bar{\boldsymbol{n}}$ as in Equation (4.4).

with $\alpha_1^{**} > 0$. Substituting Equation (4.5) into Equation (4.3) results in

$$s_i^{**} = \frac{\sum_{j\in I}(\tilde{\boldsymbol{\phi}}_1)_j}{\sum_j(\tilde{\boldsymbol{\phi}}_1)_j}\,. \tag{4.6}$$

Concerning the tissue dynamics steady-state, the values α_r^* for $r \in I_R$ are all positive since we assume to have at least one cell in each self-renewing SCC, and Equation (4.4) can be written as

$$\bar{\boldsymbol{n}}^* = \alpha_1^*\tilde{\boldsymbol{\phi}}_1 + \sum_{r\in I_R, r\neq 1} \alpha_r^*\tilde{\boldsymbol{\phi}}_r\,. \tag{4.7}$$

Substituting now Equation (4.7) into Equation (4.3) results in

$$s_i^* = \frac{\sum_{j\in I}(\alpha_1^*\tilde{\boldsymbol{\phi}}_1 + \sum_{r\in I_R, r\neq 1}\alpha_r^*\tilde{\boldsymbol{\phi}}_r)_j}{\sum_j(\alpha_1^*\tilde{\boldsymbol{\phi}}_1 + \sum_{r\in I_R, r\neq 1}\alpha_r^*\tilde{\boldsymbol{\phi}}_r)_j}\,. \tag{4.8}$$

From Equation (4.8) and Equation (4.6), it is clear that if there is only one critical SCC, this must be $r = 1$, and therefore $s_i^* = s_i^{**}$.

This result implies that if the measures in the tissue and lineage tracing data are different, i.e. $s_i^* \neq s_i^{**}$, then there must be more than one critical SCC. In other words, there must be at least another stem cell type in the tissue in addition to the initially labelled one (corresponding to case (b) in Figure 4.1). Instead, if the measures are the same, nothing can be said, since there are two possibilities (corresponding to cases (a) or (b) in Figure 4.1): i) $r = 1$ is indeed the only self-renewing cell type which might include states with different identities that are interconnected forming a single SCC; ii) other self-renewing cell types contribute to the tissue in the same proportion as the $r = 1$ one, in a way such that the final ratios, given by Equation (4.6) and (4.8), are the same. However, this condition seems somewhat unrealistic without a sophisticated mechanism that balances the different contributions.

4.3 Self-renewing strategy identification via clone lineage tracing

The previous section showed that lineage tracing, when combined with single-cell RNA-seq data analysis, could help determine stem cell types. Here, instead, we focus on the clone lineage tracing. In vivo lineage tracing of clones is commonly used to assess how adult stem cells maintain self-renewing tissues. Based on this, we will show that homeostatic cell fate models, which are those fulfilling the rules derived in Chapter 2, can be categorised into two classes that predict, under asymptotic conditions, sufficiently different clonal statistics. Two fundamental implications

follow. If such conditions are met, we can determine the stem cell self-renewal strategies by a simple qualitative assessment of lineage tracing data. However, models predicting the same, or very similar, clone size distribution cannot be distinguished given the clonal data alone.

4.3.1 Clonal statistics modelling

So far, we have modelled cell fate in homeostasis based on dynamical population models, i.e. based on a system of ordinary differential equations. However, as introduced in Section 1.2, such models are not suitable for studying the clonal dynamics. One first reason is related to the random extinction of the clones, which must be modelled when the stochastic variation in the number of cells is of the same order or larger than the size of the cell population. Also, modelling the average dynamics does not give information about clonal statistics, which is the primary output of the clonal lineage tracing experiment.

Hence, to determine the clone size distribution, we need to model the details of the stochastic process, including clone extinction. In particular, clonal statistics can be estimated by solving the master equation, a coupled system of ODEs describing the time evolution of the probability of the cell numbers [Baker, 2017]. However, a closed-form solution applicable to any cell fate model is not possible. Also, a numerical approach that integrates these differential equations is not appropriate since truncation errors become easily relevant[7]. Thus, whilst the analytical solution of the master equation will be used to gain insight into the cell fate dynamics in some simple models, in general, we will make use of its numerical estimation based on the Gillespie algorithm [Gillespie, 1977], which is detailed in Appendix B.1.1.

We finally recall that in Section 4.1, we justified the use of constant parameters cell fate models when modelling clonal data in homeostatic tissues.

4.3.2 Compartment model of cell fate dynamics

To construct candidate models for clonal dynamics, we recall that in homeostasis, a self-renewing cell type, i.e. $\mu = 0$, is always at an apex of the lineage hierarchy. All other cell types are committed types, i.e. $\mu < 0$. Therefore, taking advantage of the graph theory representation of the generic cell fate model again (see Section 2.1.2), we classify here the SCCs of the condensed network (i.e. the cell types) in two compartments: the self-Renewing compartment ($\mathcal{R}$), which is the SCC at the apex of

[7] Truncating the possible discrete points in the ith state to a maximum of k_i, results in a master equation of size $\prod_{i=1}^{m} k_i + 1$, where m is the number of cell states. Assuming, for example, the same maximum cell number in each state, which means $k_i = k$ for $i = 1, ..., m$, then the master equation consists of $(k+1)^m$ coupled ordinary differential equations.

the lineage tree; and the Committed compartment ($\mathcal{C}$), which consists of all SCCs downstream of the apex SCC. In principle, multiple self-renewing types might coexist within $\mathcal{R}$ as long as they are disconnected. However, only single-cells are labelled in the lineage-tracing experiment, and consequently, only one cell type in this compartment is active. Thus, in the modelling of the clonal dynamics, we assume one self-renewing cell type. Given that, cells in $\mathcal{R}$ can return to any state within the same compartment, and this population maintains itself. Instead, the cell population in $\mathcal{C}$ would vanish without external input since the combined dominant eigenvalue of all those SCCs is negative. Therefore, the progeny of these cells will eventually be lost.

We can thereby classify cells as being of a self-Renewing compartment (R) if their state is within $\mathcal{R}$, and of a Committed compartment (C) if their state is in $\mathcal{C}$. Based on this coarse-grained classification, a generic homeostatic model of clonal dynamics can be represented in terms of compartments $\mathcal{R}$ and $\mathcal{C}$ as

$$\begin{aligned} R \xrightarrow{\lambda_R} & \begin{cases} R+R & \text{with probability } r_{RR} \\ R+C & \text{with probability } 1-r_{RR}-r_{CC}\,, \\ C+C & \text{with probability } r_{CC} \end{cases} \\ R \xrightarrow{\omega_{RC}} C, \qquad & C \xrightarrow{\lambda_C} C+C, \qquad C \xrightarrow{\gamma_C} \emptyset. \end{aligned} \tag{4.9}$$

in which the kinetic parameters $\lambda_{R,C}$, ω_{RC} and γ_C are not constant rates in the Markovian sense, and represent instead the *effective rates* of those events, i.e. the average frequency at which they occur. That means that although there might be many cell states within a compartment whose dynamics are based on a Markovian model, we only observe the total cell numbers in the two compartments. In practice, this represents a hidden Markovian model. The relation between these effective rates and those of the generic model (2.1)-(2.3), will be discussed later, in Section 4.3.4. We also note that the loss events $R \to \emptyset$ are not explicitly modelled, since they can be approximated by a short lived state X_d in $\mathcal{C}$, as $R \to X_d \to \emptyset$. For having a homeostatic condition, it is further required that (i) the R-population remains on average constant, i.e. $\lambda_R r_{RR} = \lambda_R r_{CC} + \omega_{RC}$, and (ii) the loss rate in $\mathcal{C}$ must exceed its proliferation rate, i.e. $\gamma_C > \lambda_C$. In Figure 4.2, we show how, according to the model (4.9), a generic homeostatic cell type network can be condensed into a compartment model of renewing and committed cell states.

The formulation based on renewing and committed states can help us gain insights into the potential behaviours of generic homeostatic cell fate models. In particular, we define the *generalised asymmetric divisions* as events of the type $R \to R + C$, and the *generalised symmetric divisions* as events of the type $R \to R + R$ (symmetric renewal) and $R \to C + C$ (symmetric commitment). With these definitions, we can categorise homeostatic cell fate models into two classes. The *Generalised Invariant Asymmetry* (GIA) models are those which only present generalised asymmetric divisions in the

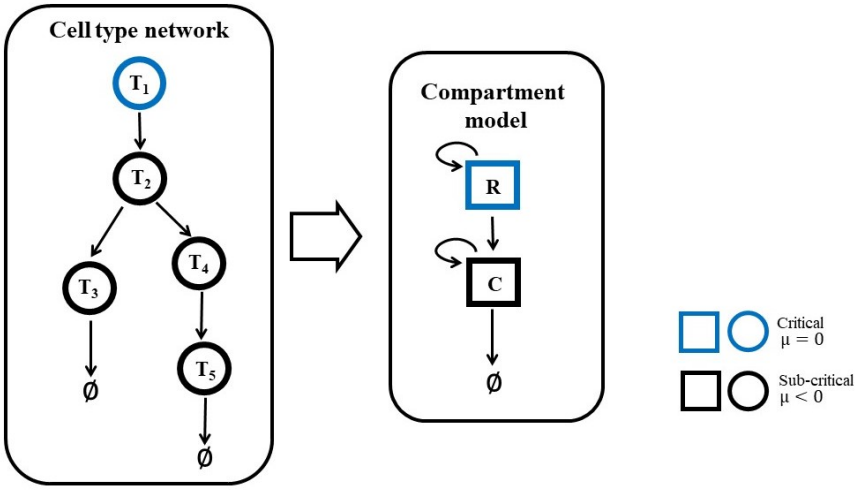


FIGURE 4.2: Illustration of a homeostatic cell type network and its compartment representation, Equation (4.9). The presented cell type network corresponds to the condensation of the cell state network illustrated in Figure 2.1. For a homeostatic network, an SCC with dominant eigenvalue $\mu = 0$ is at the apex, while other SCCs have $\mu < 0$. In the compartment representation, we distinguish the self-Renewing compartment $\mathcal{R}$, consisting of the apex SCC, with $\mu = 0$, and the Committed compartment $\mathcal{C}$ consisting of the remainder, with $\mu < 0$.

renewing compartment so that there are no symmetric divisions nor transitions to the committed compartment. The *Generalised Population Asymmetry* (GPA) are instead models for which such restriction does not hold. We note that a conservation law equivalently characterises the two classes: for the GIA models, the number of cells in $\mathcal{R}$ is strictly conserved, whilst for the GPA models, this conservation law does not hold[8].

Clearly, the classical IA and PA models presented in Chapter 1, and described by Equation (1.1), are part respectively of the GIA and the GPA category. Such models' long-term clone size distribution is the Poisson in the IA case and Exponential in the PA one. Notably, the Dynamic Heterogeneity (DH) model [Greulich and Simons, 2016], which is a model of the type

$$S \xrightarrow{\lambda} S + D, \qquad S \xrightarrow{\omega_S} D, \qquad D \xrightarrow{\omega_D} S, \qquad D \xrightarrow{\gamma} \emptyset, \tag{4.10}$$

despite presenting only asymmetric divisions, falls inside the GPA category, since, in this model, S and D cells form a single SCC at the apex of the lineage hierarchy so that they both are part of $\mathcal{R}$. Therefore, a division $S \to S + D$ in the DH model, which is asymmetric in the conventional sense, corresponds to $R \to R + R$ in the compartment view described by Equation (4.9), and thus it is a generalised symmetric division. Thus, PA and DH models are in the same category (GPA) according to this

[8] Since $\mu = 0$ is necessary for conservation, the only possible conserved cell states in homeostasis are those in $\mathcal{R}$

classification. Crucially, they both predict the same type of clone size distribution, which is an Exponential one.

4.3.3 Two-state Markovian approximation of compartment model

For a better understanding of the clonal dynamics in a generic model, we start from the compartment representation, given by Equation (4.9), and study its Markovian counterpart, as an approximation. Despite not yielding accurate clone size distributions in general, the Markovian counterparts of non-Markovian processes commonly estimate well their limiting distributions.

4.3.3.1 Generalised Invariant Asymmetry models

For the Generalised Invariant Asymmetry (GIA) models, which only feature $R \to R + C$ events between the renewing compartment, $\mathcal{R}$, and the committed compartment, $\mathcal{C}$, a corresponding Markovian model reads,

$$X_1 \xrightarrow{\lambda_1} X_1 + X_2, \qquad X_2 \xrightarrow{\lambda_2} X_2 + X_2, \qquad X_2 \xrightarrow{\gamma} \emptyset, \tag{4.11}$$

in which X_1 represents a single state in $\mathcal{R}$ and X_2 in $\mathcal{C}$. In this model, the number of cells in state X_i is n_i, for $i = 1, 2$ and the analytical solution for the steady state probability distribution, $P(n_1, n_2)$, is derived below.

We first observe that the number of cells in X_1 is conserved, that is, given a single X_1-cell initially, it always remains at $n_1 = 1$. Thus, we only need to study the dynamics of cells in X_2, for which the master equation [Baker, 2017] is given by

$$\begin{aligned} \frac{dP(n_2)}{dt} = &- (\lambda_1 + \lambda_2 n_2 + \gamma n_2)\, P(n_2) \\ &+ (\lambda_1 + \lambda_2 (n_2 - 1))\, P(n_2 - 1) \\ &+ \gamma (n_2 + 1) P(n_2 + 1). \end{aligned} \tag{4.12}$$

Equation (4.12) can be also written as

$$\begin{aligned} \frac{dP(n_2)}{dt} = &- (g(n_2) + r(n_2))\, P(n_2) \\ &+ g(n_2 - 1) P(n_2 - 1) + r(n_2 + 1) P(n_2 + 1), \end{aligned} \tag{4.13}$$

in which $r(n_2) = \gamma n_2$ and $g(n_2) = \lambda_1 + \lambda_2 n_2$.

To derive the steady state distribution, $P^*(n_2)$, corresponding to the solution of $dP(n_2)/dt = 0$, we define the net flux between states n_2 and $n_2 - 1$ as

$$I_{n_2} = r(n_2) P^*(n_2) - g(n_2 - 1) P^*(n_2 - 1)\,. \tag{4.14}$$

This implies that Equation (4.13) results in $I_{n_2+1} - I_{n_2} = 0$ for every n_2 which means that $I_{n_2} = I_0 = r(0)P^*(0) - g(-1)P^*(-1) = 0$. Therefore,

$$P^*(n_2) = \frac{g(n_2-1)}{r(n_2)}P^*(n_2-1) = \prod_{l=0}^{n_2-1}\frac{g(l)}{r(l+1)}P^*(0), \tag{4.15}$$

where $P^*(0)$ is the steady state probability of having zero cells in state X_2. Finally, by applying the conservation of the total probability, $\sum_{n_2=0}^{\infty} P^*(n_2) = 1$, and rearranging the terms we obtain

$$P^*(n_2) = \left(1 - \frac{\lambda_2}{\gamma}\right)^{\lambda_1/\lambda_2}\left(\frac{\lambda_2}{\gamma}\right)^{n_2}\frac{\Gamma\left(\frac{\lambda_1}{\lambda_2}+n_2\right)}{\Gamma(n_2+1)\Gamma\left(\frac{\lambda_1}{\lambda_2}\right)}, \tag{4.16}$$

in which $\Gamma(...)$ is the Gamma function [Abramowitz and Stegun, 1972].

We now define the dimensionless parameters $\hat{\lambda}_1 = \lambda_1/\gamma$ and $\hat{\lambda}_2 = \lambda_2/\gamma$, representing the rescaled division rates respectively for cells in state X_1 and X_2. Equation (4.16) is then rewritten as

$$P^*(n_2) = \left(1-\hat{\lambda}_2\right)^{\hat{\lambda}_1/\hat{\lambda}_2}\hat{\lambda}_2^{n_2}\frac{\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}+n_2\right)}{\Gamma(n_2+1)\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}\right)}. \tag{4.17}$$

It is noted that while $\hat{\lambda}_1$ varies between 0 and ∞, $\hat{\lambda}_2$ is defined between 0 and 1 since the committed compartment is of transient type, according to the classification proposed in Section 2.2.2 (see Table 2.1).

Considering now the mean numbers of cells in each state, indicated respectively as $\bar{n}_1$ and $\bar{n}_2$, they satisfy the system of ODEs.[9]

$$\begin{cases} \dfrac{d\bar{n}_1}{dt} = 0 \\ \dfrac{d\bar{n}_2}{dt} = \lambda_1\bar{n}_1 + (\lambda_2-\gamma)\bar{n}_2 \end{cases}. \tag{4.18}$$

Based on this, the steady state average number of cells is

$$\begin{cases} \bar{n}_1^* = 1 \\ \bar{n}_2^* = \dfrac{\lambda_1}{\gamma-\lambda_2} = \dfrac{\hat{\lambda}_1}{1-\hat{\lambda}_2} \end{cases}. \tag{4.19}$$

[9] Since there is always one cell in X_1, there are not extinct clones.

When the mean number of cells in state X_2 is sufficiently large, i.e. for large $\hat{\lambda}_1$ or in case $\hat{\lambda}_2$ is close to one, the discrete distribution given by Equation (4.17) can be approximated by a continuous probability density function, $P^*(x_2)$, given by

$$P^*(x_2) = (1-\hat{\lambda}_2)^{\frac{\hat{\lambda}_1}{\hat{\lambda}_2}} \hat{\lambda}_2^{\frac{\hat{\lambda}_1 x_2}{(1-\hat{\lambda}_2)}} \frac{\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2} + \frac{\hat{\lambda}_1}{1-\hat{\lambda}_2} x_2\right)}{x_2 \Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}\right) \Gamma\left(\frac{\hat{\lambda}_1}{1-\hat{\lambda}_2} x_2\right)}, \tag{4.20}$$

in which $x_2 = n_2/\bar{n}_2^*$.

The steady-state distribution of the X_2-cells given by Equation (4.17) and, for large mean number by Equation (4.20), exhibits a large variety of shapes. For a clear picture of its variability, we analyse its behaviour in some limiting cases below described. Their full analytical derivation is provided in the Appendix B.2.

(a) In case $\hat{\lambda}_2 \to 0$, that is, when there is no proliferation in the committed compartment, the steady state distribution results in

$$\lim_{\hat{\lambda}_2 \to 0} P^*(n_2) = \frac{\hat{\lambda}_1^{n_2} e^{-\hat{\lambda}_1}}{n_2!} = \text{Poisson}(\hat{\lambda}_1). \tag{4.21}$$

Importantly, this agrees with what we were expecting considering that when $\hat{\lambda}_2 = 0$ the model becomes simply an IA model for which the distribution in n_2 is known to be poissonian (see Section 1.2).

Considering now a large mean value, which is obtained for large $\hat{\lambda}_1$,

$$\lim_{(\hat{\lambda}_2, \hat{\lambda}_1) \to (0,\infty)} P^*(x_2) = \text{Normal}(1, 1/\hat{\lambda}_1). \tag{4.22}$$

(b) When $\hat{\lambda}_2 \to 1$, then

$$\lim_{\hat{\lambda}_2 \to 1} P^*(x_2) = \text{Gamma}(\hat{\lambda}_1, 1/\hat{\lambda}_1), \tag{4.23}$$

that is a Gamma distribution with unitary mean and shape parameter equal to $\hat{\lambda}_1$. Importantly, the Gamma distribution for $\hat{\lambda}_1 \to \infty$ tends to a Normal distribution with unitary mean and variance $1/\hat{\lambda}_1$. For $\hat{\lambda}_1 = 1$, it corresponds instead to an Exponential distribution with unitary mean. We remark that $\hat{\lambda}_2$ cannot be exactly equal to 1, since in this case the committed compartment would be self-renewing, which is incompatible with a homeostatic network.

(c) For large value of $\hat{\lambda}_1$, that is, when $\hat{\lambda}_1 \to \infty$, the probability converges to

$$\lim_{\hat{\lambda}_1 \to \infty} P^*(x_2) \simeq \sqrt{\frac{p}{2\pi}} e^{-1/2p(x_2-1)^2} = \text{Normal}(1, 1/\hat{\lambda}_1), \tag{4.24}$$

that is a Normal distribution with unitary mean and variance $\hat{\lambda}_1$. This result is consistent with those derived in (a) and (b) since the the limiting behaviour of $P^*(x_2)$ for $\hat{\lambda}_2 \to 0$ and $\hat{\lambda}_2 \to 1$ in case of large $\hat{\lambda}_1$ is the same as Equation (4.24), which is applicable for $\hat{\lambda}_1 \to \infty$ and any $\hat{\lambda}_2$.

The above derived asymptotic distributions were compared with the results of numerical simulations of the stochastic process associated with model (4.11) for different values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$. The tested parameters are graphically shown in Figure 4.3 over a contour map showing the expected steady state mean number of cells, $\bar{n}_2^*$. The tested conditions are divided into three groups representing the limiting behaviours discussed above. For each group, the curves from the numerical simulations[10] and the corresponding exact and approximated solutions are summarised in Table 4.1. In general, the analytical solution given by Equation (4.17) and, for large mean by Equation (4.20), its limiting approximations given by Equations (4.21), (4.23) and (4.24) and the numerical simulation of the stochastic process, all agree very well.

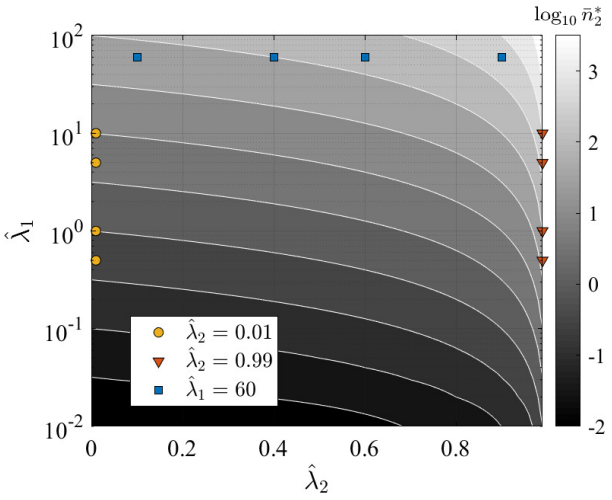
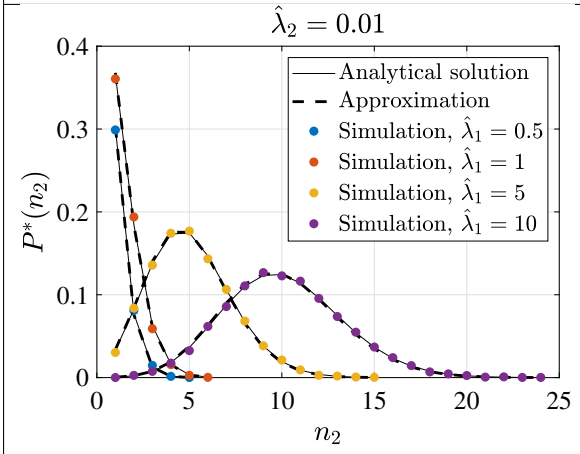


FIGURE 4.3: Parameters for testing the limiting behaviour of the Generalised Invariant Asymmetry Markovian model (4.11). The tested conditions are grouped to represent each approximation of the limiting behaviours for which the steady state distribution is derived: a) $\hat{\lambda}_2 \to 0$; b) $\hat{\lambda}_2 \to 1$; and c) $\hat{\lambda}_1 \to \infty$. The values are shown over the contour map of the expected steady state mean number of cells in state X_2, $\bar{n}_2^*$ as function of $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Tests results are summarised in Table 4.1.

[10] From the numerical simulation of the stochastic process, we computed the distribution at the final simulation time, τ, of the number of cells in state X_2. The final time was chosen here as $\tau = 20/\alpha_{\min}$, where $\alpha_{\min} = \min(\lambda_1, \lambda_2, \gamma)$; this value is well representative of a steady state condition. Furthermore, the kinetic parameters considered are based on a unitary γ (i.e. $\lambda_1 = \hat{\lambda}_1$, $\lambda_2 = \hat{\lambda}_2$ and $\gamma = 1$). The time unit is arbitrary and hence omitted.

(a) $\hat{\lambda}_2 \to 0$



Test cases:

- $\hat{\lambda}_1 = [0.5, 1, 5, 10]$
- $\hat{\lambda}_2 = 0.01$

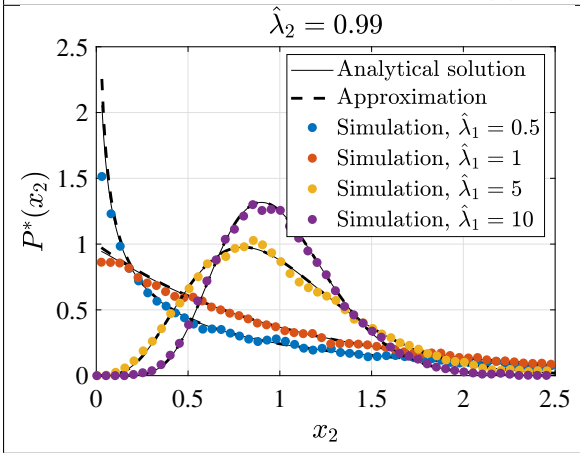
Analytical Solution:

- Equation (4.17)

Approximation:

- Poisson$(\hat{\lambda}_1)$

(b) $\hat{\lambda}_2 \to 1$



Test cases:

- $\hat{\lambda}_1 = [0.5, 1, 5, 10]$
- $\hat{\lambda}_2 = 0.99$

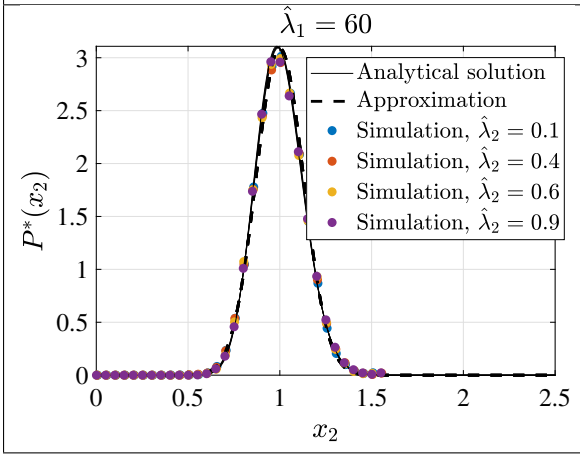
Analytical Solution:

- Equation (4.20)

Approximation:

- Gamma$(\hat{\lambda}_1, 1/\hat{\lambda}_1)$

(c) $\hat{\lambda}_1 \to \infty$



Test cases:

- $\hat{\lambda}_1 = 60$
- $\hat{\lambda}_2 = [0.1, 0.4, 0.6, 0.9]$

Analytical Solution:

- Equation (4.20)

Approximation:

- Normal$(1, 1/\hat{\lambda}_1)$

TABLE 4.1: Summary of the limiting behaviour of the steady state distribution, $P^*(n_2)$, of the the number of cells in state X_2, n_2 (or the continuous counterpart, $P^*(x_2)$, in which $x_2 = n_2/\bar{n}_2^*$) of the Generalised Invariant Asymmetry Markovian model (4.11). The figures compare the results of the numerical simulation of the stochastic process, the corresponding analytical solution and its approximation which are detailed on the right.

4.3.3.2 Generalised Population Asymmetry models

For the Generalised Population Asymmetry (GPA) models, a Markovian approximation in which X_1 represents a single state in $\mathcal{R}$ and X_2 in $\mathcal{C}$ reads,

$$\begin{aligned} X_1 \xrightarrow{\lambda_1} &\begin{cases} X_1 + X_1 & \text{with probability } r_1 \\ X_1 + X_2 & \text{with probability } 1 - r_1 - r_2 \\ X_2 + X_2 & \text{with probability } r_2 \end{cases}, \\ X_1 \xrightarrow{\omega} X_2, \qquad & X_2 \xrightarrow{\lambda_2} X_2 + X_2, \qquad X_2 \xrightarrow{\gamma} \emptyset, \end{aligned} \tag{4.25}$$

whereby for homeostasis to prevail, the conditions $\lambda_1 r_1 = \lambda_1 r_2 + \omega$ and $\lambda_2 < \gamma$ must be satisfied. We note that the dynamics of X_1 are independent of X_2 and thus the number of cells in X_1 in homeostasis satisfies[11]

$$n_1 \xrightarrow{\lambda_1 r_1 n_1} n_1 \pm 1, \tag{4.26}$$

which corresponds to a simple continuous-time branching process with two offspring. It is known that for such model the resulting distribution of cell numbers is Exponential

$$P(n_1) = \frac{1}{\bar{n}_{1,s}} e^{-n_1/\bar{n}_{1,s}}, \tag{4.27}$$

in which $\bar{n}_{1,s} \simeq \lambda_1 r_1 t$ is the mean number of cells in the surviving clones [Haccou et al., 2005]. In the following, we will intuitively show that the rescaled long-term distribution of the total clone size is Exponential as well. The corresponding detailed mathematical analysis is provided in [Parigini and Greulich, 2020].

We first consider that since each surviving cell in X_1 contributes independently to the production of X_2-cells, then $\bar{n}_2 \sim n_{1,s} \sim t$. Given that $n_{1,s} \to \infty$ for $t \to \infty$, then for large times also $\bar{n}_2 \to \infty$. Besides, we note that in the committed compartment, X_2-cells produced according to (4.25) follow a similar fate as those in GIA model (4.11); deviations are due to the simultaneous production of X_2-cells from events of the type $X_1 \to X_2 + X_2$. In particular, in the GIA model, for large λ_1, n_2 tends to a Normal distribution with mean equal to its variance. Crucially, as shown in [Parigini and Greulich, 2020] (Appendix 1), this result also applies to arbitrarily complex self-renewing compartments, i.e. models of the type (4.9), as long as $\lambda_R \to \infty$ or $n_R \to \infty$[12]. Hence, since in this case $n_{1,s} \to \infty$, the distribution of n_2 becomes normally distributed with mean and variance equal to $\bar{n}_2$, where $\bar{n}_2 \to \infty$ for $t \to \infty$.

[11] Based on model (4.25), and provided that there are n_1 cells, the probability of having $n_1 + 1$ cells is $\lambda_1 r_1 n_1$ whilst that of having $n_1 - 1$ cells is $(\lambda_1 r_2 + \omega) n_1$. Considering that the system is homeostatic, $\lambda_1 r_2 + \omega = \lambda_1 r_1$.

[12] In this generic case, the variance of the Normal distribution is not exactly equal but only proportional to the mean. However, if each compartment is composed by a single state, as here, the mean is equal to the variance.

We express now the the number of X_2-cells as a rescaled variable, $x_2 = n_2/\bar{n}_s$, in which $\bar{n}_s$ is the total mean of the surviving clones. For large times the variance of the clone size distribution in x_2 vanishes since $\sigma^2_{x_2} = \sigma^2_{n_2}/\bar{n}^2_s$, where $\sigma^2_{n_2} \sim n_{1,s} \to \infty$ and $n_{1,s} < \bar{n}_s \to \infty$. Hence, x_2 is a random number from a Normal distribution with mean $\bar{x}_2$ and null variance. That is, for a given $x_1 = n_{1,s}/\bar{n}_s$, $x_2 \approx \bar{x}_2 \sim x_1$. Given that, the total rescaled clone size is therefore $x = x_1 + x_2 \sim x_1$, which means that if $P(x_1)$ is Exponential, $P(x)$ is Exponential as well.

4.3.4 Numerical simulation of random cell fate models

In the previous section, we showed that the Markovian representation of the two classes of models, GIA and GPA, presents two different behaviours. In the GIA model, the distribution is characterised by a large variety of shapes, which converges to a Normal distribution under an asymptotic condition. In the GPA model instead, the long-term distribution is always Exponential. Thus, in this section, we analyse the clonal dynamics of random models to check whether this correspondence between model class (i.e. GIA and GPA) and predicted clonal statistics (i.e. Normal and Exponential) holds in general. To this aim, we numerically estimate the clone size distribution for a large number of random stochastic models, implemented via random generation of the parameters λ_i, ω_{ij}, γ_i and r_i^{jk}. Crucially, the generation of these random models is driven by the rules for a homeostatic cell state network derived in Chapter 2. Hence, a random condensed network is built first, with a single self-renewing cell type at the apex of the cell type network. Successively, each cell type is filled with a random cell state network that forms a single SCC characterised by the desired value of μ, i.e. $\mu = 0$ for the self-renewing type and $\mu < 0$ for the committed types. To simulate the clones, we run stochastic simulations based on the Gillespie algorithm [Gillespie, 1977], assuming a Markov process that follows the rules of Equations (2.1)-(2.3). For each model, we run a large number of simulations starting from one cell in the compartment $\mathcal{R}$. In doing so, the cell population in each simulation run represents one clone. Then, we sample their outcomes, the total cell numbers per clone, $n = \sum_i n_i$, to obtain predictions for clonal statistics, namely the frequency distribution of clone sizes and mean clone sizes. The detailed description of the random models' generation and the simulation campaign is given in Appendix B.3.

The simulation results are graphically displayed in Figure 4.4, in terms of evolution of the mean number of surviving clones, $\bar{n}_s$, as function of the time and in Figure 4.5, as final rescaled clone size distribution $P(x)$, in which $x = n/\bar{n}_s$. In the GIA simulations (left panel), the final time τ corresponds to $20/\alpha_{\min}$, in which $\alpha_{\min}$ is the minimum kinetic parameter, $\alpha_{\min} = \min(\lambda_1, ..., \omega_{12}, ..., \gamma_m)$. The results for the GPA models (right panel) correspond instead to the time at 98% clone extinction. In each figure, the grey

shade represents the percentile of all the simulations, and the black lines limit the 5-95 percentile range[13]; the blue lines are instead some selected illustrative cases. As a reference, the standard Exponential distribution is shown in green in the clone size distribution plots.

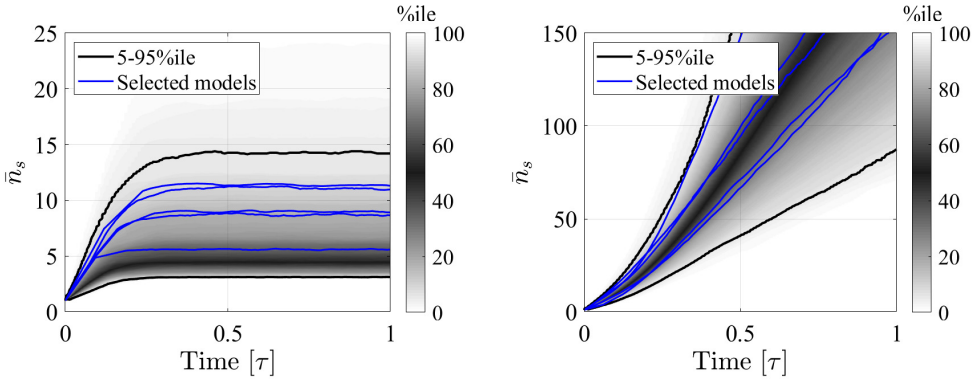


FIGURE 4.4: Simulation results in term of mean size of surviving clones, $\bar{n}_s$, as a function of the time scaled by the final value, τ, for the random GIA models (left), and GPA models (right). The grey shade represents the percentile of all the simulations (black lines limit the 5-95 percentile range); the blue curves correspond to some illustrative selected simulations.

Concerning the mean clone size of surviving clones, presented in Figure 4.5, we note that indeed a common behaviour is seen in each case: whilst for every simulated GIA model, $\bar{n}_s$ saturates at a plateau value, for every GPA model, it steadily increases. This can be understood given that clones in a GPA model can go extinct while those in a GIA model do not. Assuming that there are initially a large number M of clones, the total number of cells is $N = M\,\bar{n}_s$. Since the system is homeostatic, after a sufficient amount of time, it will reach a constant steady-state, N^*, meaning that the mean clone size is $\bar{n}_s = N^*/M$. In case no clones go extinct, as in GIA models, M is constant, and thus $\bar{n}_s$ approaches a constant. Instead, in non-conserved multi-type branching processes, as GPA models are, the clone number M decreases through progressive extinction of clones [Haccou et al., 2005], and therefore $\bar{n}_s$ increases, despite the cell population as a whole staying stationary.

Considering now the clonal statistic results, shown in Figure 4.5, we observe that all simulated GPA models (right panel) predict asymptotically the same rescaled clone size distribution, namely a standard Exponential distribution. For short times and small clone sizes, deviations exist. However, as shown in the next section, where the convergence of the clone size distribution is assessed, these deviations vanish in the large time limit. That means that different models within the GPA class, in the long term, differ only by the mean clone size, a free fit parameter, and consequently, they cannot be distinguished. Instead, regarding the GIA models, shown in Figure 4.5 (left

[13]Simulations for which the final mean is below two and where the final condition is not achieved (due to computational limitations) are omitted. This results in 238 and 571 models, respectively, for the GIA and GPA cases.

panel), we see all kinds of clone size distribution shapes, both peaked distributions and non-peaked ones. Some distributions are even close to an exponential form, and thus, they cannot be distinguished from GPA models. This result is consistent with the Markovian approximation discussed in Section 4.3.3, for which only for large $\hat{\lambda}_R = \lambda_R/\gamma_C$ the clone size distribution tends to a Normal distribution, whilst peaked and non-peaked distributions are found in non-asymptotic conditions. In Section 4.3.5, we will analyse the asymptotic behaviour of these random GIA models.

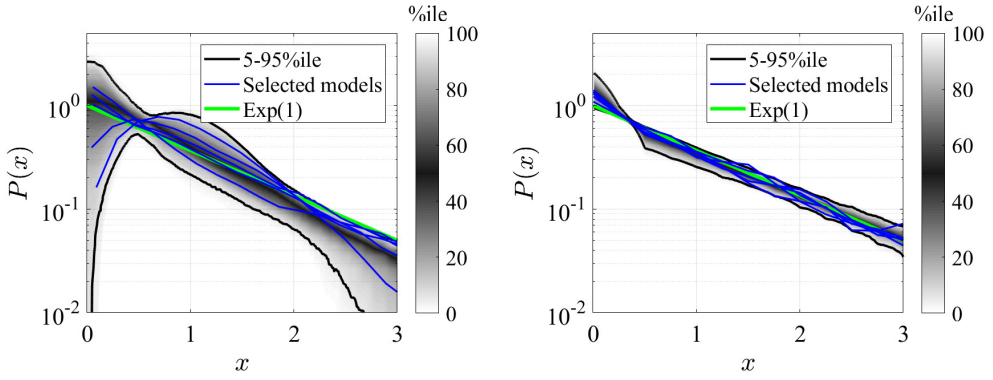


FIGURE 4.5: Simulation results in term of rescaled clone size distribution at the final time τ, $P(x)$, where $x = n/\bar{n}_s$ for the random GIA models (left), and GPA models (right). The grey shade represents the percentile of all the simulations (black lines limit the 5-95 percentile range); the blue curves correspond to some illustrative selected simulations. The Exponential distribution with unitary mean is also shown in green.

4.3.4.1 Convergence of Generalised Population Asymmetry Model

In Figure 4.5, it is shown that GPA models predict asymptotically, for long times, the same rescaled clone size distribution, that is, an Exponential distribution with unitary mean. Here, we analyse how this distribution is approached. In particular, Figure 4.6 reports the 50 percentile of all the distributions in the GPA models at different levels of extinction (which are related to the different time points), showing a gradual convergence to the expected Exponential distribution.

Thus, the GPA models Markov approximation, Equation (4.25), becomes accurate for sufficiently large time, and no significant deviations are observed. This feature also means that the distribution in the long term is independent of the choice of parameters. Only the mean value of surviving clones, $\bar{n}_s$, depends on the parameters, but it does not affect the rescaled distribution, expressed in terms of $x = n/\bar{n}_s$. Therefore, we can conclude that GPA models attain an Exponential clone size distribution for time $t \to \infty$ and abstain from a comprehensive study of different parameter regimes.

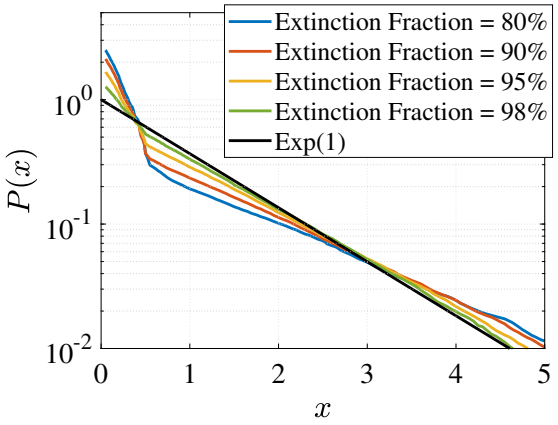


FIGURE 4.6: Convergence of the clone size distribution for increasing extinction fraction (i.e. increasing time) to an Exponential distribution with unitary mean (black line). Each curve represents the 50 percentile of the rescaled distributions $P(x)$, where $x = \bar{n}/\bar{n}_s$, of the GPA random models analysed in Section 4.3.4 (see Figure 4.5).

4.3.5 Analysis of Generalised Invariant Asymmetry models

In contrast to the GPA models, where the total clone size rescaled distribution is independent in the long-term of the model parameters, the choice of parameters becomes relevant in GIA models. We recall from Section 4.3.3 that a generic GIA model can be expressed in terms of the compartments $\mathcal{R}$ and $\mathcal{C}$, where the cell fate dynamics is modelled as Equation (4.9). When the dynamics of compartments are assumed to be Markovian, the steady-state distribution, given by Equation (4.9), shows a variety of shapes depending on the two parameters $\hat{\lambda}_1 = \lambda_1/\gamma$ and $\hat{\lambda}_2 = \lambda_2/\gamma$. Consistently, the clone size distribution in the random model (Figure 4.5, right panel) presents peaked and not peaked profiles depending on the case. Therefore, we first treat the Markovian model assessed in Section 4.3.3.1 as an approximation of more complex GIA cell fate models. Building on this, we identify the corresponding limiting parameters and successively test the behaviour of the random cell fate model when the asymptotic condition is met.

4.3.5.1 Evaluation of the 2-state Markovian approximation in random cell fate models

To evaluate the applicability to the random models of the Markovian approximation discussed in Section 4.3.3.1, we first express the *effective* non-Markovian rates (i.e. the mean frequency of events) of the representation (4.9), $\lambda_{R,C}$ and γ_C, in terms of the original model, (2.1)-(2.3). These rates are computed considering the same steady-state mean number of cells. We therefore rewrite the dynamics of the mean cell numbers,

Equation (2.6), in block form as

$$\begin{cases} \dfrac{d\bar{\mathbf{n}}_R}{dt} = A_{RR}\bar{\mathbf{n}}_R \\ \dfrac{d\bar{\mathbf{n}}_C}{dt} = A_{CR}\bar{\mathbf{n}}_R + A_{CC}\bar{\mathbf{n}}_C \\ \dfrac{d\bar{n}_\emptyset}{dt} = A_{\emptyset C}\bar{\mathbf{n}}_C \end{cases}, \tag{4.28}$$

in which $\bar{\boldsymbol{n}}_{R,C}$ denote the vectors of mean cell numbers of states restricted to compartments $\mathcal{R}, \mathcal{C}$, respectively, and $\bar{n}_\emptyset$ the average number of lost cells (not considered for total cell number counting and homeostasis condition). It is noted that $A_{RC} = \mathbf{0}$, since there cannot be links from $\mathcal{C}$ to $\mathcal{R}$. Also $A_{\emptyset R} = \mathbf{0}$ as we do not consider loss from $\mathcal{R}$, an assumption already discussed in Section 4.3.2.

Hence, summing up all the components in each compartment, $\bar{n}_R = \sum_i (\bar{\mathbf{n}}_R)_i = 1$ and $\bar{n}_C = \sum_i (\bar{\mathbf{n}}_C)_i$, results in

$$\begin{cases} \dfrac{d\bar{n}_R}{dt} = 0 \\ \dfrac{d\bar{n}_C}{dt} = \sum_i (A_{CR}\bar{\mathbf{n}}_R)_i + \sum_i (A_{CC}\bar{\mathbf{n}}_C)_i \\ \dfrac{d\bar{n}_\emptyset}{dt} = A_{\emptyset C}\bar{\mathbf{n}}_C \end{cases}. \tag{4.29}$$

The effective parameters of this 2-state Markovian Approximation (MA^2) are then estimated from the steady state condition $\bar{\mathbf{n}}_x^*$ and $\bar{n}_x^*$, for $x = R, C$, as

$$\lambda_R = \sum_i (A_{CR}\bar{\mathbf{n}}_R^*)_i,\ \gamma_C = \frac{\sum_i (A_{\emptyset C}\bar{\mathbf{n}}_C^*)_i}{\bar{n}_C^*} \text{ and } \lambda_C = \gamma_C - \frac{\lambda_R}{\bar{n}_C^*}. \tag{4.30}$$

We then compared the clone size distribution obtained for the random GIA models with that of the corresponding MA^2, based on model (4.11) with parameters $\hat{\lambda}_1 = \hat{\lambda}_R = \lambda_R/\gamma_C$ and $\hat{\lambda}_2 = \hat{\lambda}_C = \lambda_C/\gamma_C$. The values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ for all the GIA random models are shown in Figure 4.7 (left) over the contour map of the expected mean number of cells in $\mathcal{C}$ (in compartment $\mathcal{R}$ there is always one single cell). In general, $\hat{\lambda}_1$ remains below five and $\hat{\lambda}_2$ is spread between zero and one. As a measure of the error of this approximation, ϵ, we choose the maximum difference between the distributions of the random GIA model and the MA^2 one, relative to the peak value of the random model. For low mean cell numbers, the random GIA model distribution is compared to Equation (4.17); for large mean numbers instead, the rescaled distribution is compared to Equation (4.20). A threshold on the mean cell number equal to ten was chosen to distinguish between these two cases. This relative error ϵ as a function of $\hat{\lambda}_2$ is presented in Figure 4.7 (right), where it is evident that large errors are obtained only for large values of this parameter.

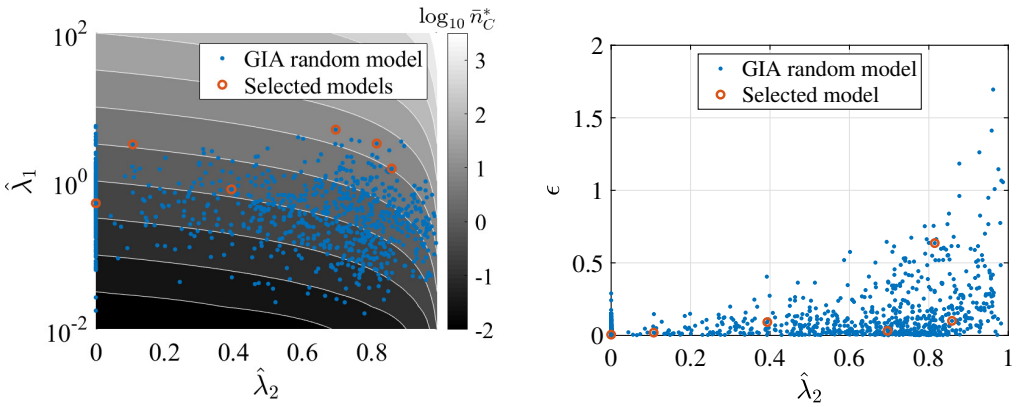


FIGURE 4.7: Effective parameters of the GIA random models $\hat{\lambda}_1 = \hat{\lambda}_R$ and $\hat{\lambda}_2 = \hat{\lambda}_C$, based on Equation (4.30), over the contour map of the expected steady state mean number of committed cells, $\bar{n}_C^*$ (left); relative error of the MA^2 model, ϵ, as function of $\hat{\lambda}_2 = \hat{\lambda}_C$ (right). Some illustrative cases, for which the steady state distribution is shown in Figure 4.8, are highlighted.

Some illustrative cases, representative of different values of $\hat{\lambda}_2$ (indicated as Selected model in Figure 4.7), were chosen and their distribution is shown in Figure 4.8. In these figures, we compare the distributions of a GIA random model, the corresponding MA^2 and, when applicable, the Limiting Approximation (LA) of the MA^2 model for $\hat{\lambda}_2 \to 0$ and $\hat{\lambda}_2 \to 1$. The following considerations are made:

- Two cases for $\hat{\lambda}_2 < 0.2$ are presented in Figure 4.8 (top). Here, the distribution obtained from the random models agrees with the MA^2, which in turn is well approximated by the LA, corresponding to a Poisson distribution. As expected, larger deviations between the MA^2 and the Poisson distribution are noted for increasing values of $\hat{\lambda}_2$. In general, all the random GIA models in this range are well approximated by the MA^2 model.

- Two cases presented in Figure 4.8 (middle) feature $\hat{\lambda}_2 > 0.8$, for which the Gamma distribution is the limiting approximation of the MA^2. In this range of $\hat{\lambda}_2$, the GIA random model distribution in some cases, see, for instance, the left figure presents some deviations with respect to the MA^2. However, globally a good agreement is obtained in most cases (failing ratio, based on a 0.5 maximum error is 21.7%).

- Two cases in an intermediate range $0.2 < \hat{\lambda}_2 < 0.8$ are shown in Figure 4.8 (bottom). Again, the MA^2 is well representative of the distribution of the corresponding GIA random model (failing ratio, based on a 0.5 maximum error is 3.2%). It is noted that for such values of $\hat{\lambda}_2$, a limiting approximation of the MA^2 is not available.

Given that the MA^2 can catch the behaviour of a generic random GIA model in most of the tested cases, it represents a good approximation (global failing ratio, based on a

0.5 maximum error is 6%). In the cases where the MA^2 does not yield a good approximation, the internal structure of the $\mathcal{R}$ and $\mathcal{C}$ compartments become relevant, and subsequent events that affect n_R and n_C become dependent on each other, and thus are non-Markovian.

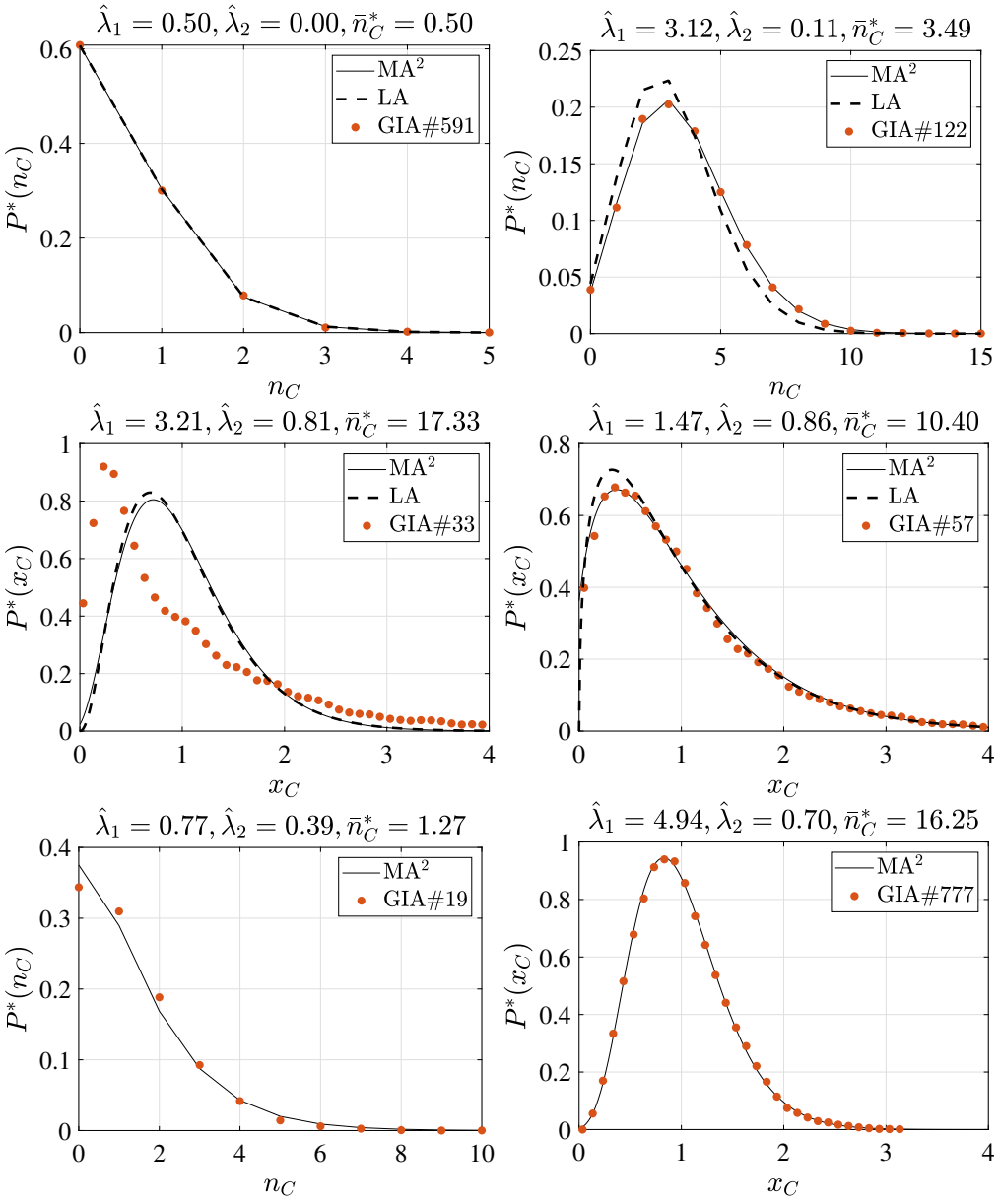


FIGURE 4.8: Steady state distribution $P^*(n_C)$ (or equivalently $P^*(x_C)$) of the number of cells in the committed compartment, $\mathcal{C}$, for some selected GIA random models (see Figure 4.7). The curves are compared with those of the corresponding 2-state Markovian Approximation (MA^2), given by Equation (4.17) (discrete distribution) and Equation (4.20) (continuous distribution). For low $\hat{\lambda}_2$ (top panels) and large $\hat{\lambda}_2$ (middle panels), also the Limiting Approximation (LA), which is respectively Poisson$(\hat{\lambda}_1)$ and Gamma$(\hat{\lambda}_1, 1/\hat{\lambda}_1)$, is shown.

4.3.5.2 Asymptotic behaviour of GIA models

In the previous section, we showed that the dynamics in the random GIA models are consistent with those of the two-state Markovian approximation of the self-Renewing and Committed compartments. Thus, considering that the Markovian GIA model, assessed in Section 4.3.3.1, presents an asymptotic behaviour where clone size tends to a Normal distribution when $\hat{\lambda}_R$ is large, we test here this condition for the GIA random models. Crucially, $\hat{\lambda}_R$ is an intrinsic property of the model, and therefore, the GIA random models previously analysed were modified by changing the kinetic parameters associated with $\mathcal{R}$ to achieve a target value of $\hat{\lambda}_R$. Considering that there are infinite combinations compatible with this condition, we applied a global search method, and more specifically, a Genetic Algorithm [Goldberg, 1989]. Therefore, we set up an optimisation problem, where the kinetic rates are the optimisation variables, and the cost function is the distance of $\hat{\lambda}_R$ from the target.

First, for a single random model, #870, a sensitivity analysis for increasing values of $\hat{\lambda}_R$, spanning from 0.5 to 30, was run, and the resulting clonal size distribution is shown in Figure 4.9. This case clearly shows how an Exponential-like distribution, corresponding to $\hat{\lambda}_R = 0.5$, changes approaching a Normal distribution when $\hat{\lambda}_R = 30$. We consider now all the random models with rates consistent with $\hat{\lambda}_R = 30$. In these cases, the envelope of all the curves[14] and some illustrative profiles are shown in Figure 4.10 (left panel). As a reference, a Normal distribution characterised by unitary mean and variance equal to $1/\hat{\lambda}_R = 1/30$ is also reported. This curve corresponds to the distribution expected in the Markovian model for which $\hat{\lambda}_1 = \hat{\lambda}_R$. Deviations become relevant when, in a random model, the internal structure of the compartments leads to subsequent events that are not independent of each other. These effects alter the variance of the Normal distribution. In fact, Figure 4.10 (right panel) is based on the same simulation results, but in this case, the distribution in each model is rescaled considering both the mean number of cells and its variance (a Normal distribution is a two-parameter distribution).

4.3.5.3 Bimodal distribution of the clone size

Looking in more detail at results presented in the previous section, we note that when taking the limit of large $\hat{\lambda}_R$ also all the kinetic parameters within $\mathcal{R}$ increased as well. What if instead some kinetic parameters in $\mathcal{R}$ do not scale to become large with $\hat{\lambda}_R$? To assess this situation we study a simple test case similar to model (4.11) but containing two states in $\mathcal{R}$, connected via direct state transition. In this model, the self-renewing compartment is composed by states X_1 and X_2. Cells in these states

[14] Simulations for which the final condition (20 times the inverse of the minimum kinetic parameter) is not achieved (due to computational limitations) are omitted, resulting in 922 models.

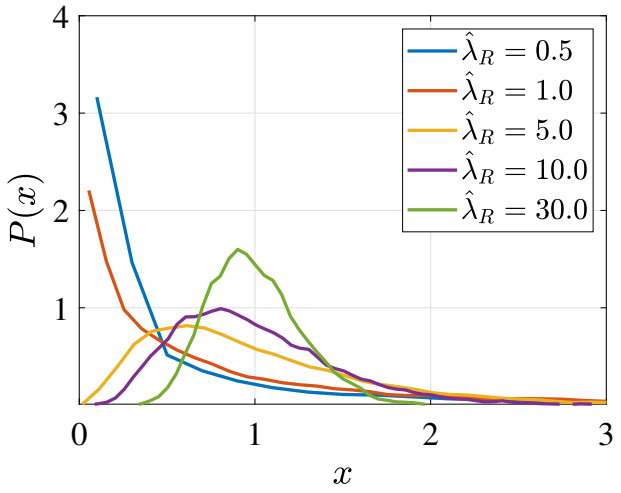


FIGURE 4.9: Sensitivity to parameter $\hat{\lambda}_R$ of the rescaled clone size distribution at the final time τ, $P(x)$, where $x = n/\bar{n}_s$ for an illustrative case, corresponding to the GIA random model #870.

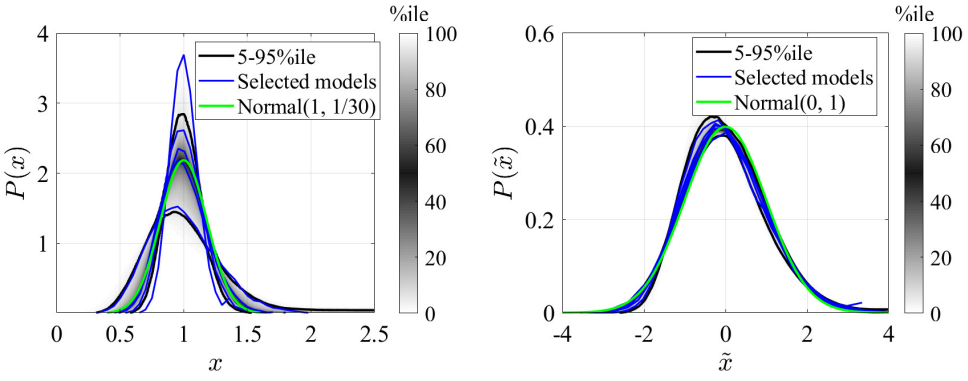


FIGURE 4.10: Simulation results in term of the clone size distribution at the final time τ for the random GIA models when $\hat{\lambda}_R = 30$. The distribution is rescaled by the mean value (left), i.e. $P(x)$, where $x = n/\bar{n}_s$, or by mean and variance (right), i.e. $P(\tilde{x})$, where $\tilde{x} = (n - \bar{n}_s)/\sigma_s$. A reference curve corresponding to a Normal distribution is also shown in green.

divide asymmetrically, that is, one daughter cell remains within the self-renewing compartment while the other enters the committed compartment. They can also change state between X_1 and X_2 (*cell state switching*) while still remaining within $\mathcal{R}$. The committed compartment is composed of a single state, X_3, and cells in this state either duplicate or die. This corresponds to

$$\begin{aligned} &X_1 \xrightarrow{\lambda_1} X_1 + X_3, \qquad X_2 \xrightarrow{\lambda_2} X_2 + X_3, \qquad X_1 \xrightarrow{\omega_{12}} X_2, \qquad X_2 \xrightarrow{\omega_{21}} X_1, \\ &X_3 \xrightarrow{\lambda_3} X_3 + X_3, \qquad X_3 \xrightarrow{\gamma} \emptyset. \end{aligned} \tag{4.31}$$

In this model, the effective parameters as defined in Equation (4.30) results in $\lambda_R = (\lambda_1\omega_{21} + \lambda_2\omega_{12})/(\omega_{12} + \omega_{21})$, $\lambda_C = \lambda_3$ and $\gamma_C = \gamma$. We then scaled the problem by γ_C, defining the following ratios: $\hat{\lambda}_R = \lambda_R/\gamma_C$, $\hat{\omega} = \omega_{12}/\gamma_C$, $a = \lambda_1/\lambda_2$ and

$b = \omega_{12}/\omega_{21}$. In the following, we test this model for different values of a and $\hat{\omega}$ as reported in Table 4.2, while fixing $\hat{\lambda}_R = 30$, which is the main scaling parameter, $\hat{\lambda}_C = 0$ and $b = 1$.

The rescaled distribution of the number of committed cells (i.e. in state X_3), n_C, obtained at the final simulation time τ, is shown in Figure 4.11. A value of τ equal to $20/\alpha_{\min}$ (where $\alpha_{\min}$ is the minimum of all rate parameters) was chosen to assure that the steady-state is reached. We first consider the test cases BM#1 and BM#2 according to Table 4.2. They are characterised by $a = 1$, meaning that they do not feature differences in the division timescales for the two self-renewing states. Both test cases lead to a Normal distribution, independently on the value assumed by $\hat{\omega}$. Test cases BM#3 to BM#7 instead are characterised by $a = 10$, and different orders of magnitude for $\hat{\omega}$ are tested. In these cases, the distribution is Normal until $\hat{\omega} \geq \hat{\lambda}_R/10$ (see cases BM#3 to BM#5). When $\hat{\omega}$ is significantly lower than $\hat{\lambda}_R$, bimodality emerges (see cases BM#6 and BM#7). Finally, looking at the extreme case, BM#7, cells in each self-renewing state, if analysed independently, would result in a Poisson distribution in the committed compartment. These distributions have different mean values, low for the slow-dividing state and large for the fast-dividing one. Globally, this leads to a Bimodal distribution computed as

$$P(n) = \beta \text{Poisson}(\hat{\lambda}_R^{(1)}) + (1 - \beta)\text{Poisson}(\hat{\lambda}_R^{(2)}), \tag{4.32}$$

in which β is the mixing parameter

$$\beta = \frac{\bar{n} - \bar{n}_2}{\bar{n}_1 - \bar{n}_2}, \tag{4.33}$$

and the parameters $\hat{\lambda}_R^{(i)}$ and $\bar{n}_i$ for $i = 1, 2$ correspond to the parameter $\hat{\lambda}_R$ and to the mean number of cells of a system in which the self-renewing compartment would be composed just by state X_i. The total mean number of cells is instead indicated by $\bar{n}$. The bimodal distribution given by Equation (4.32) is indicated as a black dashed-dotted line in Figure 4.11.

Hence, if all rates within $\mathcal{R}$ are large compared to the rates in $\mathcal{C}$, we observe a Normal clone size distribution. However, if the direct transition rates between the states of $\mathcal{R}$ are smaller or equal than γ_C, each dividing state becomes essentially separated from the others. Consequently, such states generate Normal distributions with a mean that is consistent with its cell division rate. Thus, if proliferation rates are very different, we observe a clone size distribution that results from overlaying Normal distributions with different mean, that is, in the case of two dividing cell states, a Bimodal distribution. From a biological point of view, this situation occurs when stem cells can be either in a slow-cycling or a fast-cycling proliferative state (with asymmetric division only), and there is a limited probability of switching state from one state to the other. Given that, we can therefore conclude that GIA models attain a Normal

clone size distribution if all the kinetic parameters within $\mathcal{R}$ are much larger than the inverse lifetime of C-cells, γ_C.

Case	$\hat{\omega}$	λ_1/λ_2
BM#1	30	1
BM#2	0.03	1
BM#3	300	10
BM#4	30	10
BM#5	3	10
BM#6	0.3	10
BM#7	0.03	10

TABLE 4.2: Bimodal clone size distribution test case simulation parameters.

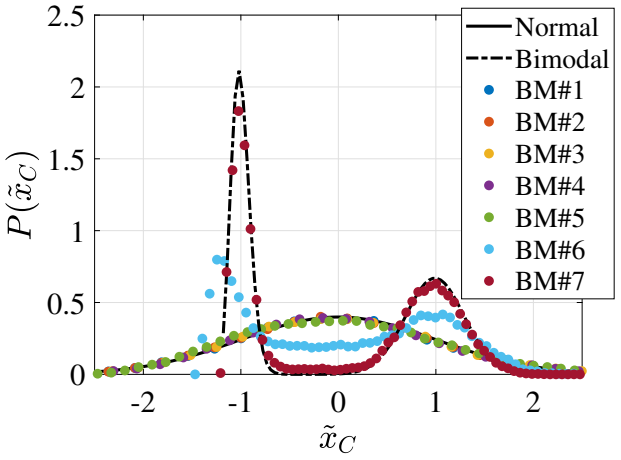


FIGURE 4.11: Clone size distribution of the cell number in the committed compartment, n_C, at the final time for the bimodal test cases. Distributions are rescaled as $P(\tilde{x}_C)$, where $\tilde{x}_c = (n_C - \bar{n}_C)/\sigma_{n_c}$ and σ_{n_c} is the variance of n_C. In addition to the stochastic simulation results, the reference Normal and Bimodal (Equation (4.32)) distributions are also shown.

4.3.6 Universality of cell fate models in homeostasis

So far, we found that the classification in GIA and GPA models mirrors their clonal statistics. Clonal data of models within the GPA class converge with time to an Exponential clone size distribution. For GIA models, instead, distributions can generally vary, but if the rates of divisions and transitions in the $\mathcal{R}$ compartment are much larger than the rate of cell loss, the clone size distribution becomes a Normal distribution. Therefore, in analogy to statistical physics, we categorise them as two *universality classes* [Klein and Simons, 2011], meaning that the details of the model do not affect the scaled outcomes for asymptotic conditions. This is a form of weak convergence of random variables [Billingsley, 1999]. Hence, the universality of the model dynamics implies that effective, simplistic models are often equally accurate to represent experimental data, yet with a higher statistical power due to less free parameters. In other words, given the clonal data in asymptotic conditions, this finding implies that a) we can determine the underlying self-renewing strategy just by looking at the shape of the clonal size distribution, and b) models of cell fate cannot, in general, be distinguished with resolution beyond the R vs C categorisation of cell types.

Although this analysis seems to discourage efforts to unravel details of cell fate dynamics, opportunities remain when the limiting conditions for asymptotic distributions are not fulfilled. Importantly, whilst the asymptotic condition in a GPA scenario is related to the time since the initial labelling of the clones and therefore can always be reached, this is not certain in the GIA case, where it depends instead on the intrinsic rates of cell division, transition and death in the system. In particular, if fast-cycling committed progenitor cells are present together with slow-cycling stem cells, then the condition that the global division rate in $\mathcal{R}$ is much larger than the cell loss rate in $\mathcal{C}$ is not fulfilled. In that case, the details of the model dynamics may affect the shape of the clone size distribution allowing for the distinction between models. Instead, distinguishing between models within the GPA category is more difficult given that the predicted statistics from models of this class always become more similar over time. Short term measurements would, in principle, allow such a distinction. However, the underlying transitions between individual cell states and cell divisions are not truly Markovian in reality. Therefore, the cell fate modelling, which always assumes dynamics between cell states to be Markovian (only the compartment modelling dynamics is not Markovian), is not necessarily a good representation of the actual cell dynamics at short times.

Finally, as shown in the next chapter, caution should be given when an Exponential clone size distribution is observed. This could suggest either a GIA model with high activity of committed progenitor cells or a GPA model. If this is the case, the mean clone size might give hints on the self-renewing strategy. As shown in Section 4.3.4, if in the GPA model the mean clone size keeps increasing with time due to the extinction of part of the clones, in the GIA case, it reaches a plateau. However, limitations on this approach lay on the fact that few time data points in a noisy environment (commonly lineage tracing is based on two or three data time points with a few hundred clones each) might not be enough to determine the trend of the mean clone size properly. Thus, this analysis shows that intrinsic limitations exist for identifying strategies of stem cell self-renewal through clonal data from cell lineage tracing experiments only.

4.4 Conclusion

Cell fate models can be very complex, with many cell sub-types in a tissue. In Chapter 2, we determined the conditions for having a homeostatic system, which requires a well-defined hierarchy of the cell types. Here, instead, we studied how to extract valuable information about the underlying cell fate model from qualitative features of experimental data. Although this analysis does not explicitly consider mechanisms for homeostasis regulation, like that studied in Chapter 3, the derived results apply to most renewing tissues, such as epithelial sheets or volumnar organs, including the mammary gland, which is the study case.

Concerning lineage tracing transcriptome data, we showed how cluster size long-term measures could be used to identify disconnected self-renewing cell types. To this aim, we first derived the relative size of the cell clusters for a generic cell fate model. We then compared the size of the clusters in the tissue and those obtained if only cells that are the progeny of a particular stem cell type are sequenced. Crucially, any difference in the two measures can only be justified by the presence of at least another self-renewing cell type. Thus, if this condition occurs, it is a clear indication that other stem cell types exist in addition to those labelled for lineage tracing. Instead, a different strategy must be followed if those two measures are not distinguishable since this data alone is insufficient to assure that the tissue is maintained only by the labelled stem cell types.

Clonal lineage tracing, instead, could help distinguish the self-renewing strategy. In modelling clonal dynamics, the stochastic details of the underlying dynamics, including the random extinction of the clones, are essential. Therefore, the presented analysis of the clonal statistics builds on numerical simulations of stochastic processes associated with arbitrary complex cell fate models and, whenever possible, their analytic solutions.

For a better insight into the cell fate clonal dynamics, we introduced a new categorisation of cell types, distinguishing between the self-renewing compartment, $\mathcal{R}$, and the committed compartment, $\mathcal{C}$. Cells that are self-renewing (R-cells) retain the potential to remain or return to the apex of the lineage hierarchy, those that are committed (C-cells) are inevitably lost eventually, together with their progeny. By construction, models of this type are compatible with the conditions for homeostasis derived in Chapter 2. According to this categorisation, we classified generic models of cell fate choice as *Generalised Invariant Asymmetry* (GIA), if only generalised asymmetric divisions of the form $R \to R + C$ occur for R-cells, and *Generalised Population Asymmetry* (GPA), when all kind of divisions can occur, as long as gain and loss of R-cells are balanced. Additionally, a conservation law characterises models of the GIA category since the number of R-cells is strictly conserved. Instead, GPA models do not present such a conservation law.

To understand the clonal dynamic behaviour of each model, we first assessed the compartment model Markovian approximation. This is based on a two-state model in which one state is in $\mathcal{R}$, where cells divide with the rate λ_1, and the other in $\mathcal{C}$, in which cells divide or die, respectively with rate λ_2 and γ. The analysis showed that in the GPA model, whatever the rates are, the clonal dynamics converge in the long-term to an Exponential distribution. In the GIA model, instead, the shape of the clonal size distribution might be peaked or non-peaked, depending on the ratio of cell proliferation to cell loss $\hat{\lambda}_1 = \lambda_1/\gamma$. However, by studying the limiting behaviour of this distribution, we identified an asymptotic condition, where, for large values of $\hat{\lambda}_1$, the clone size distribution tends to a Normal distribution.

The analysis of the clonal dynamics in complex random cell fate models resulted in the same correspondence between model class and clone statistics. Whilst in all the GPA cases, no matter the complexity of the model, the clone size distribution converges to an Exponential distribution with time, in the GIA ones, it is characterised by the same variety of shapes, peaked and not peaked, that was identified in the two-state Markovian approximation. Notably, a common feature in all the GIA cases, independently of the shape of the distribution, is that the mean value of the clone size reaches a steady state with time. In contrast, since clones become extinct in the GPA model, the mean value increases with time. Furthermore, by changing the rates in the GIA model, we confirmed the existence of an asymptotic condition for which the clone size distribution tends to a Normal distribution when all the kinetic parameters within $\mathcal{R}$ are much larger than the loss rate in $\mathcal{C}$. Thus, we categorised the GIA and GPA models into two universality classes. For each one of them, a scaling limit exists, i.e. ratio between the kinetic parameters in $\mathcal{R}$ and the loss rate in $\mathcal{C}$ for GIA and time for GPA. In such conditions, all models within a class yield the same rescaled clonal statistics, that is, Normal in GIA models and Exponential in GPA ones.

From a practical perspective, and in view of the application to a study case presented in the next chapter, these results reveal the limitations and strengths of lineage tracing assays. If transcriptome data of lineage traced cells could help determine the cell state network, the clonal statistics is at the basis of the definition of the self-renewing strategy. When the long-term distribution is peaked, there is no doubt that self-renewing is sustained only by generalised asymmetric division. Instead, if the distribution is Exponential, at least a few clonal data points at different time points must be combined to determine the trend of the mean clone size distribution, which increases if the model is within the GPA class whilst it reaches a steady-state if it is in the GIA one. In any case, models of cell fate within the same class, GIA and GPA, cannot generally be distinguished with further resolution beyond the compartment categorisation of cell types if only long-term data is available. If short-term measurements are not applicable in this model due to non-Markovian effects, the combination of long-term and mid-term data could be the key to distinguishing between models.

Chapter 5

Application to the mouse mammary gland

This chapter applies the results derived to a study case, the adult healthy mouse mammary gland. In particular, we will use the information available from the literature to build a sensible cell state network for this study case and synthetic data inspired by actual lineage tracing experiments for the model parameters fitting. In this way, we aim at validating the methodology developed, paving the way for future studies where experimental data might be available.

The presented analysis proves that restricting the search to the homeostatic cell fate models is essential for correctly modelling the dynamics. Furthermore, we show how the combination of mid-term and long-term clonal statistics and the estimation of the time scale of the dynamics are valuable data that could significantly restrict the variability of the fitting parameters and, most importantly, identify the self-renewing strategy.

This chapter is organised as follows: the approach to the definition of a cell fate model for the study case is described in Section 5.1; the cell state network definition is reported in Section 5.2; the generation of synthetic data is described in Section 5.3; the parameter fitting results are shown in Section 5.4; conclusions are given in Section 5.5.

5.1 Approach to cell fate model definition

In Chapter 2, we showed that cell fate models must follow strict rules for being homeostatic. Also, from Chapter 4, qualitative features of the experimental data could help distinguish classes of models applicable to specific scenarios. Despite drastically reducing the possible cell fate models, an infinite number of models with an arbitrary number of cell states and relations remain. Considering the current supercomputing

capabilities, a possible strategy to determine the cell fate model that best fits experimental data is to use innovative and efficient techniques, such as machine learning, to solve model selection and fitting parameter problems. However, we must consider that experimental data is scarce and noisy: for the study case, clone size distribution is supposed to be based on a few hundred clones for just two points in time. These features imply that fitting may suffer from two problems: a) the degeneracy of the model and b) the model overfitting. Concerning a), this means that multiple sets of parameters and models are equivalently good in fitting the data. This problem is even more critical given the outcomes of the analysis of the clonal dynamics presented in Chapter 4, for which a whole class of models, the GPA, converges to the same clone size distribution for long times. Crucially, this problem cannot be entirely resolved by additional data or data of higher quality in the long-term. Model overfitting is instead a well known problem, where an over complex model is the result of matching noisy data very well, at the price of losing the capability of catching the actual behaviour of the system [Claeskens and Hjort, 2008].

Given that, the approach followed is first to address the definition of the structure of the model, i.e. the cell state network, and then to solve the parameter fitting problem. Concerning the identification of potential cell state networks, we can filter out all the non-homeostatic models based on the modelling presented in Chapter 2, and we will use specific knowledge of the study case to define a realistic and meaningful cell fate model. We then apply a Bayesian methodology to determine the best parameters of such a model that match the available data. We are aware that data might not be sufficient to solve the problem completely; however, we stress that we are interested in qualitative features of the dynamics rather than accurate estimations of the rates, and thus we will focus on determining at least the underlying self-renewing strategy.

We remark that this analysis is based on the synthetic data described in Section 5.3. Such data is designed to emulate the expected inputs from two lineage-tracing experiments that were not available in a timeframe compatible with this research project[1]. In particular, one experiment aims at determining the cell identities based on transcriptome analysis of single cells that are the progeny of an initially labelled subpopulation of basal stem cells. In contrast to classical single-cell RNA-sequencing experiments, where all the tissue is processed at a single-cell level, here, only the progeny of the initially labelled subpopulation of cells, which in the case under investigation are stem cells within the basal compartment, are sequenced. Notably, the clustering analysis of this data would allow us to answer the following biologically relevant questions:

Q1 Are there any luminal cells within the labelled basal clones?

[1] One experiment was first cancelled, and the other is not yet completed for reasons outside our control. Among other factors, COVID-19 related delays played a critical role.

Q2 Which cell identities (i.e. cell clusters) are found?

Q3 What is the relative size of each cluster?

Q4 Is the relative size of the cluster consistent with tissue data?

As detailed in the next section, the answers to Q1, Q2 and Q4 are at the basis for building a sensible cell state network for the study case. The answer to Q3 is instead used in the model fitting (see Section 5.4) together with the clone size distribution data. The clonal size distribution data are related to the second lineage tracing experiment, where images of clones are processed, providing clone statistics.

5.2 Cell state network definition

In this section, we describe how to derive a cell fate network for the study case. However, before doing this, we will discuss the main outcomes of the comparison of the available scRNA-seq data that were described in Section 1.4.1. This analysis gives a better insight into the mammary gland tissue and provides biologists valuable inputs for future experiments' design.

5.2.1 Analysis of literature data

In Section 1.4.1, we gave an overview of scRNA-seq data available in the literature. The four pieces of work on scRNA-seq, [Bach et al., 2017, Pal et al., 2017, Sun et al., 2018, Giraddi et al., 2018], provide essential information on cell types, potential lineages, relationships and regulation through the different mammary gland development stages. Nevertheless, there are some discrepancies among their findings, and further research is still needed. This comparison was carried out for training purposes, but it gives insights into the study case's difficulties. Given that, the main idea of this section is to interpret these data using an unbiased approach and give inputs to future experiments aimed at resolving the identified inconsistencies. Furthermore, after the cancellation of the scRNA-seq experiment on lineage tracing data, we used this analysis to support the identification of key cells identity and markers shared among the four sets of data to tentatively extract quantitative information of cell identities during the clone imaging[2].

Considering that the study case is related to homeostasis in adult tissue, we only focus here on the adult virgin samples, and more specifically in the following datasets

[2]More specifically, we propose using antibodies during clone imaging to visualise cells expressing specific proteins. Here, we assume that RNA expression of a particular gene is a proxy of the level of the related protein. In doing so, we were hoping to distinguish the identity of the cells forming the clone.

(details are given in Table 1.1): **Ds#1**, Nulliparous samples (NP-1, NP-2) from [Bach et al., 2017]; **Ds#2**, 10X Adult sample from [Pal et al., 2017]; **Ds#3**, Virgin sample from [Sun et al., 2018]; and **Ds#4**, 10X Adult samples (Adu1) from [Giraddi et al., 2018].

The methodology applied for the data analysis is based on the following steps: 1) gene expression data loading; 2) cell filtering; 3) normalisation; 4) identification of the High Variable Genes (HVG); 5) data scaling; 6) dimensionality reduction; 7) clustering; 8) identification of the Differentially Expressed (DE) genes and 9) visualisation. Whilst we leave in Appendix C.1 the full details of the methodology and data analysis, here we only mention that the presented results are based on the Principal Component Analysis for the dimensionality reduction (point 6) and, for clustering (point 7), the Shared Nearest Neighbors (SNN) algorithm applied to the HVG expression levels in the reduced dimensional space (i.e. after PCA). Concerning the visualisation of the clusters (point 9), they are represented over the t-SNE plots, which are shown in Figure 5.1. For clarity, a consistent naming of the cell identities, which is based on the expression of known genes mainly taken from the reference articles under analysis, is used in all the datasets (and thus, it might be different in the corresponding published work). These are: Basal (B); Luminal Progenitor (LP); Luminal Mature (LM); Luminal Intermediate (LI) and Rare cluster (R). Some particular sub-clusters are labelled with the superscript * as explained below.

Focusing first on the rare clusters, we note that up to six of them are in each dataset. Their relative size is lower than 2.1%, but, despite their small size, they might play an essential role in the tissue (stem cells are usually rare). Three main observations are made.

- **Contamination.** We observe that in both [Bach et al., 2017] and [Giraddi et al., 2018], biological criteria to classify cells as contamination are given; in [Sun et al., 2018] the level of expression of some housekeeping genes is checked, while there are no explicit criteria to exclude contaminating cells in [Pal et al., 2017]. Our analysis includes all the cells, and interesting correspondences between rare clusters are found (see details in Appendix C.1.3.1). The most significant finding is related to clusters R1 of **Ds#3**, cluster R1 of **Ds#1** and clusters R2/R4 of **Ds#4**. They all present a similar level of expression but whilst in [Sun et al., 2018] (**Ds#3**), this cluster is declared to be formed by bipotent mammary stem cells, in [Bach et al., 2017] (**Ds#1**) and [Giraddi et al., 2018] (**Ds#4**) it is considered as contamination and therefore excluded from the analysis. Therefore, it is clear that a more standardised procedure for carrying out the clustering analysis and more experimental work is needed on this topic.

- **Mixed lineage.** In [Pal et al., 2017], mixed-lineage cells are detected in C1 data analysis but not in the 10X data one, which are those analysed here. However, clusters R2 and R3 in **Ds#2** show expression of both basal and luminal key genes.

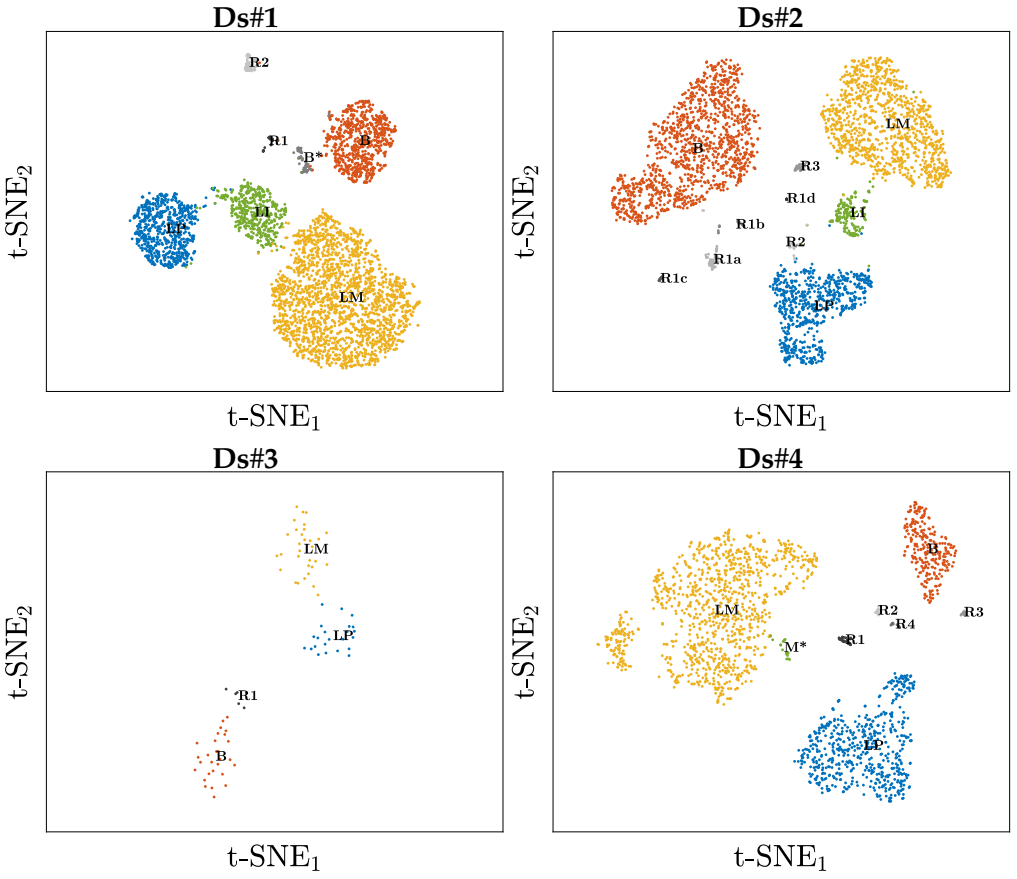


FIGURE 5.1: Clusters visualised as t-SNE plot. Each point corresponds to a single cell which is coloured according to its identity. The cell identity is established by the SNN clustering algorithm applied to the HVG expression levels in the reduced dimensional space, that is, after PCA (details of the methodology are given in Appendix C.1.1). The main clusters are the Basal (B), Luminal Progenitor (LP), Luminal Mature (LM) and Luminal Intermediate (LI); rare (R) or special (*) clusters are coloured in different shades of grey. The four panels correspond to data taken from [Bach et al., 2017] (**Ds#1**), [Pal et al., 2017] (**Ds#2**), [Sun et al., 2018] (**Ds#3**), and [Giraddi et al., 2018] (**Ds#4**); details of each dataset are summarised in Table 1.1.

Consistently, a small but well defined basal sub-cluster detected in the analysis of **Ds#1**, indicated as B*, shows the expression of both basal and luminal key genes. Although this finding suggests that this cluster is potentially a mixed-lineage cluster, further dedicated experimental work is needed to confirm this hypothesis. Another possibility is that it is a cluster containing cell doublets, i.e. two cells incorrectly sequenced together as if they were a single cell, though in [Bach et al., 2017] the presence of doublets in this sample is excluded.

- **Luminal intermediate.** As it will be shown below, in both **Ds#1** and **Ds#2**, a luminal cluster showing expression of both progenitor and mature genes is present and identified as luminal intermediate. From a careful analysis of the t-SNE plot of **Ds#4**, a small sub-cluster of the LM slightly separated from the main cluster in the direction of the LP cluster was identified and following indicated as M*, see Figure 5.1 (bottom-right panel). This cluster presents

expression in genes related to both the LP and LM identities. Thus, this feature suggests that this cluster corresponds to the luminal intermediate cluster of **Ds#1** and **Ds#2**.

Concerning the main clusters, B, LP, LM and LI, the following observations are made.

- **Cluster heterogeneity.** The performed clustering process highlighted a great level of heterogeneity in all of these clusters. Thus, to obtain the results shown in this section, several small clusters were merged based on the analysis of DE genes and the t-SNE plots. Further analysis of the available data should be carried out to highlight correspondences of such sub-clusters in the four datasets.
- **Relative size.** The relative size of the clusters is reported in Table 5.1. This information was primarily extracted to be compared to scRNA-seq of the lineage traced cells and therefore answer to Q4. However, it is apparent that significant differences, up to 24.4% in the basal cluster, are found in the four datasets. Thus, further (experimental) assessments are needed to reduce or explain the variability of these data. For example, the impact of the age of the mice on the size of the clusters could be studied.
- **Shared DE genes.** To answer Q1 and Q2 without the scRNA-seq experiment, we proposed the use of fluorescent tags to tentatively distinguish the identity of the cells in the lineage tracing imaging. In the four published works, the three clusters B, LP and ML were identified; however, there is not a common indication of markers for each cell identity (see Section 1.4.1). Thus, we analysed the DE genes in the available datasets and identified those that are shared in each of the main clusters[3]. The detailed analysis is reported in Appendix C.1.3.2, here, we provide only the list of these genes in Table 5.2. Together with Dr Elias, we then combined this result and additional technical and non-technical information, such as availability of the antibody, known relation gene-marker, and ended up with a list of candidate markers to use during the imaging of the clones. Whilst for the basal cluster, we agreed on the uses of *Acta2* and *Krt14*, which are among those identified here, for the luminal ones, we opted for alternative markers, i.e. *Krt8* and *Elf5* for LP and *Esr1* for LM, given that, according to Dr Elias, they have more potential to distinguish these cell identities. However, at the time of writing this thesis, such experimental data is not available, and thus we have no feedback on the applicability of this methodology. In any case, Table 5.2 provides a starting point for future experiments aimed at identifying key markers that uniquely identify cell identity for this study case.

[3]LI cells show both LP and LM markers, and there are no specific markers for this cluster.

Cluster ID	Cluster size [%]				Δ_{max}
Ds	#1	#2	#3	#4	[%]
B	17.0	37.5	31.8	13.1	24.4
LP	16.4	20.2	26.1	29.6	13.2
LM	50.9	32.0	35.2	51.9	19.9
LI	11.3	4.8	0.0	0.9	11.3
Other	4.5	5.5	6.8	4.5	2.4

TABLE 5.1: Relative cluster size in the four datasets; the last column reports the maximum variability.

B	**LP**	**LM**
Acta2	*Csf3*	*Areg*
Cnn1	*Csn3*	*AW112010*
Cxcl14	*Cst3*	*Cited1*
Krt14	*Cxcl1*	*Cxcl15*
Krt17	*Ltf*	*Fgb*
Tagln	*Mfge8*	*Glul*
Tpm2	*Trf*	*Gpx3*
		Ly6d
		Prlr
		Ptn
		Wfdc2

TABLE 5.2: Differentially expressed genes in the main clusters that are common in the four datasets. LI cells, which are not associated with the expression of specific genes, express both LP and LM genes.

5.2.2 Model definition

Recalling the controversy about the lineage hierarchy in the mouse mammary gland, discussed in Section 1.4, and in light of the analysis reported in the previous section, it is clear that Q1 is a fundamental question for defining the structure of the cell state network. If luminal cells were found, that is a clear indication that the two lineages, the basal and the luminal one, must be connected in homeostasis. This experiment alone does not provide information about the reversibility between states in the two lineages (i.e. cells changing from a basal state to a luminal one and vice-versa), but, at least, it assures a one-way connection, that is, from basal to luminal. Concerning now the answer to Q2, this enables the identification of the model cell states, whereas we assume that cell states correspond to cell identities. Furthermore, in addition to clustering analysis, we might infer transitions between states based on pseudo-time analysis of the data. Finally, the answer to Q4 has fundamental implications in the identification of different pools of stem cells disconnected from the basal one, as demonstrated in Section 4.2. Given that we do not have any information from dedicated experiments, in the following, we will build a cell fate network making assumptions that might not reflect the reality in the study case. However, the above

considerations provide a clear pathway to identify the cell state network once such data becomes available.

First of all, concerning Q1, we assume that there are luminal cells in the clones[4]. In other words, we assume that the basal and the luminal lineages are somehow connected at least in one direction, from basal to luminal states. Regarding Q2, we consider the following cell identities: Basal Stem (BS); Basal Mature (BM); Luminal Progenitor (LP) and Luminal Mature (LM). We do not explicitly include luminal intermediate cells, assuming they are in a transition state with limited proliferation potential. We further assume a clear direction in the lineage, where cells can only change states from stem to progenitor and progenitor to mature states. Despite this not being excluded a priori, there is no evidence in homeostasis of possible reversibility, that is, from mature back to stem. Hence, we consider that each cell identity represents a single cell type, according to the definition from Chapter 2. Furthermore, we note that BS cells are not distinguished from BM in the scRNA-seq clustering analysis presented in Section 5.2.1. Thus, the Basal (B) cluster size includes BS and BM cells. Concerning Q4, according to [Elias et al., 2017, Van Keymeulen et al., 2017, Wang et al., 2017], Luminal Stem cells (LS) exist. What is yet undetermined is whether they are active in homeostasis and feed into the basal pool. Assuming that such a state is disconnected from the BS type, it is not included in the model.

Finally, to keep the model as simple as possible, we assume one state for each cell type and a limited set of connections (i.e. cell state transitions and divisions), as sketched in Figure 5.2. We remark that more complex models could have been designed, such as models with more states, cell types, and connections. However, we intentionally

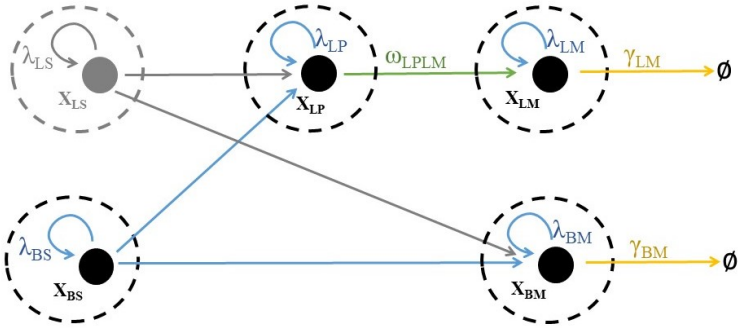


FIGURE 5.2: Cell state network associated with cell fate model (5.1). This model will be used for fitting the synthetic data representative of scRNA-seq and clone lineage tracing. Its design is based on hypothetical answers to Q1-Q4. The network comprises four single-state cell types (i.e. SCC) corresponding to Basal Stem (BS), Basal Mature (BM), Luminal Progenitor (LP) and Luminal Mature (LM) cells. In grey, an additional luminal stem cell type is also shown; however, it will be not considered in the model fitting given that only basal stem cells are labelled in the lineage-tracing experiments.

[4]Dr Elias supports this hypothesis.

choose the simplest model given that it is mainly based on assumptions supported by limited experimental evidence. This choice is also justified by the results presented in Chapter 4, where we showed that simple cell fate models were able to catch very well the global behaviour of more complex models. Furthermore, we want to avoid the degeneracy and the overfitting issues as described in Section 5.1. Given that, the cell state network results in

$$
\begin{aligned}
X_{BS} \xrightarrow{\lambda_{BS}} &\begin{cases} X_{BS} + X_{BS} & \text{with probability } p_{BSBS} \\ X_{BS} + X_{BM} & \text{with probability } p_{BSBM} \\ X_{BS} + X_{LP} & \text{with probability } p_{BSLP} \\ X_{BM} + X_{BM} & \text{with probability } p_{BMBM} \\ X_{LP} + X_{LP} & \text{with probability } p_{LPLP} \end{cases} , \\
X_{BM} \xrightarrow{\lambda_{BM}} X_{BM} + X_{BM}, &\qquad X_{BM} \xrightarrow{\gamma_{BM}} \emptyset, \\
X_{LP} \xrightarrow{\lambda_{LP}} X_{LP} + X_{LP}, &\qquad X_{LP} \xrightarrow{\omega_{LPLM}} X_{LM}, \\
X_{LM} \xrightarrow{\lambda_{LM}} X_{LM} + X_{LM}, &\qquad X_{LM} \xrightarrow{\gamma_{LM}} \emptyset.
\end{aligned}
\tag{5.1}
$$

5.3 Synthetic data generation

In this section, we describe the assumptions and methods used to generate the synthetic data. To emulate the experiments, we assume to have a limited set of data that reflect the constraints in available resources, such as cost and time to carry out the experiments, and ethical aspects, like limiting the number of animals used in the experiments. We also do not consider any source of error such as mouse-to-mouse variability, data acquisition and processing[5].

First of all, to generate the synthetic dataset, we need a cell fate model. For doing that, we assume a lineage hierarchy that, to some extent, is similar to the cell fate network defined in Section 5.2.2, which is composed of stem, progenitor and mature cell types, topologically ordered from stem to mature. However, to introduce the minimum possible bias, we used one of the random models analysed in Chapter 4 and thus, as shown later, there is not a complete correspondence between the two condensed networks (i.e. the fitting cell fate model (5.1) and the synthetic one as a substitute of reality). In particular, we selected model GPA#690[6].

The structure of this model is shown in Figure 5.3, and the numerical values of the kinetic parameters are summarised in Table 5.3. We observe that the cell state network

[5]To model these errors we would introduce further assumptions and approximations.

[6]This network was chosen as it has mid-high complexity, with three SCCs and at least two states in each SCC; it also has a global structure that can be associated with the expected lineage that is, from stem cell to progenitor and mature.

comprises nine cell states, grouped in three SCCs that correspond to three cell types. The apex SCC, formed by two states, is self-renewing; the downstream SCCs, transient, include seven states. To extract the cell clusters size, which would correspond to the analysis of scRNA-seq data, we relate the cell states to the cell identity that are supposed to be distinguished. According to Section 5.2.1, that is, basal, luminal progenitor and luminal mature. For that purpose, we assume that there is no reversibility between the three cell clusters, and thus, we associate each cluster to an SCC. By topological ordering the cell types, we relate the B cluster to the apex SCC, defining the **B-type**, and the LP and LM clusters respectively with the first and second transient downstream ones, i.e. the **LP-type** and the **LM-type**. In this model, there is no distinction between BS and BM cell types, which, in any case, are part of the same cluster in the scRNA-seq data analysis. Also, here, the B-type is directly linked to the LM-type. These two features of the synthetic model imply that the correspondence between this model and the fitting one, given by Equation (5.1), is not perfect, as representative of a realistic scenario.

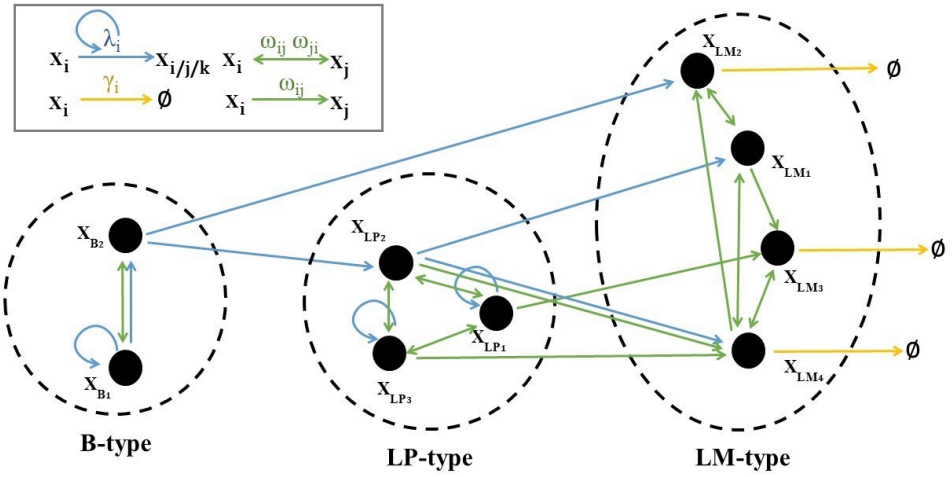


FIGURE 5.3: Cell state model for synthetic data generation. This network corresponds to the random model GPA#690 analysed in Chapter 4; values for the kinetic parameters are reported in Table 5.3. This network is composed of nine cell states, grouped in three SCC: each SCC is associated with a cell cluster/type, and specifically the Basal (B), Luminal Progenitor (LP) and the Luminal Mature (LM).

To generate the synthetic experimental data based on the above-selected cell fate model, we run stochastic simulations using the Gillespie algorithm [Gillespie, 1977] (see details in Appendix B.1.1). Concerning the initial conditions, simulations are initialised with one basal (stem) cell, specifically a cell in state[7] X_{B_2}. The total number of simulated clones is 10^5.

The time points at which the data is extracted are defined consistently with those of the real experiments. Since the rates in the synthetic model are random, we use as a time scale, τ, the time at which the total mean number of cells is $\bar{n}(\tau) = \bar{n}^*(1 - e^{-1})$,

[7] We chose to use the same random initialisation as in the analysis reported in Chapter 4.

B-type		LP-type		LM-type	
λ_{B_1}	2.17	λ_{LP_1}	1.48	$\omega_{LM_1LM_2}$	2.62
λ_{B_2}	1.44	λ_{LP_2}	1.17	$\omega_{LM_1LM_3}$	1.65
$\omega_{B_1B_2}$	2.45	λ_{LP_3}	1.45	$\omega_{LM_1LM_4}$	1.87
$\omega_{B_2B_1}$	1.62	$\omega_{LP_1LP_2}$	1.87	$\omega_{LM_2LM_1}$	1.92
		$\omega_{LP_1LP_3}$	2.36	$\omega_{LM_3LM_4}$	3.05
		$\omega_{LP_1LM_3}$	1.36	$\omega_{LM_4LM_1}$	2.70
		$\omega_{LP_2BS_1}$	2.35	$\omega_{LM_4LM_2}$	2.90
		$\omega_{LP_2LP_3}$	2.99	$\omega_{LM_4LM_3}$	2.06
		$\omega_{LP_2LM_4}$	1.93	γ_{LM_2}	0.33
		$\omega_{LP_3LP_1}$	2.20	γ_{LM_3}	0.79
		$\omega_{LP_3LP_2}$	1.78	γ_{LM_4}	0.95
		$\omega_{LP_3LM_4}$	1.19		

TABLE 5.3: Kinetic parameters of the cell fate model used to generate synthetic data; the structure of the network is shown in Figure 5.3. The values reported, expressed in 1/week, are rescaled to be consistent with the experiment time frame.

where $\bar{n}^*$ is the steady-state value, and assume that in the experiment τ corresponds to four weeks[8]. In Figure 5.4, we show the total mean cell number, $\bar{n}$, normalised by the homeostatic value, $\bar{n}^*$, which represents the average clonal dynamics evolution starting from a single B cell on day zero. We remark that, in principle, these dynamics are unknown and only estimated. The time points at which virtual experimental data are extracted are also indicated, and they are described below.

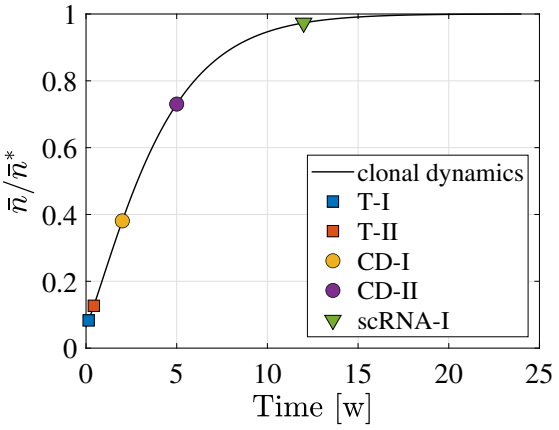


FIGURE 5.4: Virtual experimental data points: **T-I** and **T-II** are data points for testing the experimental setup and they are not considered in the fitting; **CD-I** and **CD-II** are the points at which clonal statistics is expected; **scRNA-I** corresponds to the time at which samples of lineage traced cells are collected for sequencing. The time points are shown over the scaled average clonal dynamics (black line); however, these dynamics are not known in a real scenario but only estimated.

[8] According to Dr Elias, the same time scale measured for luminal clones can also be applied to basal ones. Considering that in [Elias et al., 2017], samples at eight weeks from induction show saturation of clones, as a rough estimation of the turnover time in the tissue, we assume eight to twelve weeks. For what concerns the dynamical model, we expect that the solution converges to the steady-state between two and three τ.

- **T-I and T-II.** On day one, **T-I**, there is a test of the initial labelling of the clones where we expect to see mostly single cells[9]. On day three, **T-II**, we expect the first rounds of divisions. We observe that these points are only used to test the experimental setup and not for extracting quantitative information for the model fitting. Hence, they are only reported for completeness for showing what is done in an actual experiment but not considered hereafter.

- **CD-I and CD-II.** The first set of clonal data, $\mathcal{D}_{CD-I}$, was expected on week two and the second one, $\mathcal{D}_{CD-II}$, on week five. At each of these time points, we extracted from the numerical simulation 200 uncorrelated surviving clones (i.e. with at least one cell), where 200 is representative of the number of clones that might be processed according to Dr Elias. This results in

$$\mathcal{D}_{CD-I} = \{f_n^{(I)}\}\,, \tag{5.2}$$

and

$$\mathcal{D}_{CD-II} = \{f_n^{(II)}\}\,, \tag{5.3}$$

in which $f_n^{(I)}$ and $f_n^{(II)}$ are the relative frequency of clones of size n respectively at time point I and II. The resulting distribution is shown in Figure 5.5. Here, single-cell clones and those in the tail of the distribution are filtered out. We define the tail as formed by clones of size $n > n_{tail}$, where n_{tail} corresponds to the first of three consecutive clone sizes with zero count. That is done to avoid possible overfitting of the model and focus on the main part of the data. The shaded grey area represents the 2σ expected error assuming the number of counts follows a multinomial distribution [Pitman, 1993].

- **scRNA-I.** On week twelve, we expect to have the scRNA-sequencing of lineage traced cells, which provides the third set of data, $\mathcal{D}_{scRNA-I}$. To emulate this experiment, we assume to sequence around 5000 cells that are the progeny of the B cells labelled at time zero. Thus, we extracted from the numerical simulation uncorrelated clones for which the total cell count is consistent with this number, resulting in 112 clones and 4956 cells. Importantly, clones do not have to be distinguished in the real experiment since cells are pulled together for sequencing. The raw data is the count of cells in each of the main clusters, N_x for $x = B, LP, LM$. Based on these counts, we estimate the relative size of each cluster, $s_x = N_x/N$, where $N = \sum_x N_x$. This is

$$\mathcal{D}_{scRNA-I} = \{s_B, s_{LP}, s_{LM}\}\,. \tag{5.4}$$

[9] In clonal dynamics assays, the initial condition is of utmost importance (see Section 1.3.1). Therefore, initial cell labelling must be checked as follow: a) there are enough single cells labelled so that an adequate number of clones remains upon random extinction of some of them; b) labelled cells are sufficiently separated in space to allow for the distinction of the clones in successive time points.

This data and the 2σ variability is shown in Figure 5.6 (left). Similarly to the clonal statistics, the expected variability is estimated considering that the size of each cluster is based on independent counts, and thus it follows a multinomial distribution. Figure 5.6 (right) shows the same data plotted over the time evolution of the average clonal and tissue dynamics (black lines). It is remarked that these dynamics are not known in reality, but they are presented here to confirm the consistency, in the long-term, of the experimental data, average clonal dynamics, and tissue dynamics (see Section 4.2). It is noted that the variability of these measurements is very low. The main reason is that the process of generation of the synthetic data does not consider any animal-to-animal variability (usually at least three animals per time point are considered in an actual experiment), single cell acquisitions or scRNA-seq analysis errors.

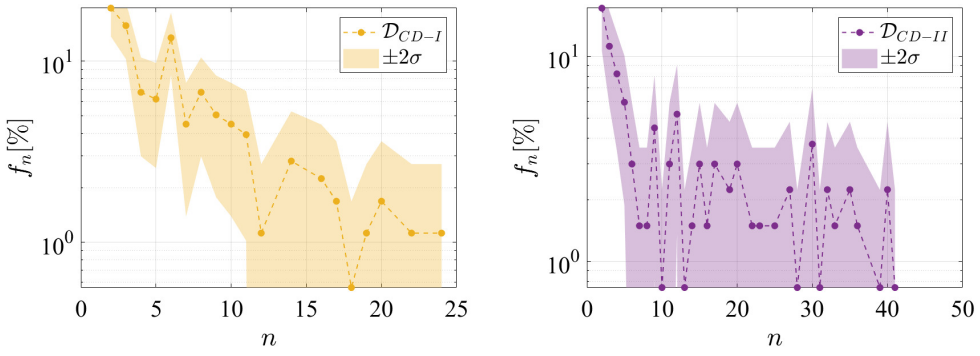


FIGURE 5.5: Clone size distribution in terms of relative frequency, f_n, of the clones at time point **CD-I** (left) and **CD-II** (right). These data correspond to 200 uncorrelated clones, filtered to remove single-cell clones and the distribution tail. In addition to the data, the expected 2σ variability is also shown.

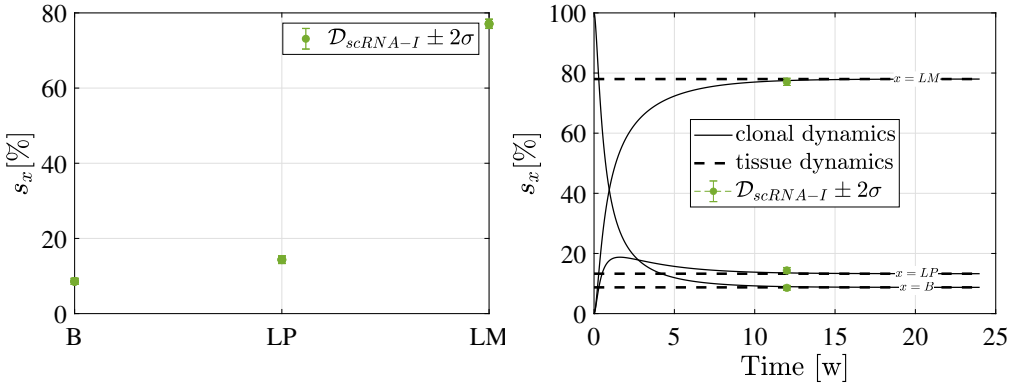


FIGURE 5.6: Relative cluster size, s_x, for $x = B, LP, LM$, of clones at **scRNA-I** based on the emulation of the scRNA-sequencing of lineage tracing data. Data and its 2σ variability (left) is also presented over the average clonal and tissue dynamics (right) as a function of the time (black lines). We observe that these dynamics are not known in a real scenario.

5.4 Cell fate model parameter fitting

In the following, we present the results of fitting the synthetic experimental data presented in the previous section, based on the cell fate model (5.1) described in Section 5.2.2.

5.4.1 Fitting model and strategy

The cell fate model (5.1) has a total of twelve parameters, seven of which are kinetic parameters,

$$\boldsymbol{\alpha} = \Big(\lambda_{BS},\ \ \lambda_{LP},\ \ \lambda_{LM},\ \ \lambda_{BM},\ \ \omega_{LPLM},\ \ \gamma_{BM},\ \ \gamma_{LM}\Big), \tag{5.5}$$

and the remaining five are the probabilities of the outcome of basal stem cells division,

$$\boldsymbol{p} = \Big(p_{BSBS},\ \ p_{BSBM},\ \ p_{BSLP},\ \ p_{BMBM},\ \ p_{LPLP}\Big). \tag{5.6}$$

Considering that $\sum_i p_i = 1$, the model has eleven effective parameters. To implement more easily the homeostasis constraint, we parametrise the problem considering the following relations

$$\begin{cases} p_{BSBS} = r(1+\Delta) \\ p_{BSBM} = (1-2r)p_a \\ p_{BSLP} = (1-2r)(1-p_a) \\ p_{BMBM} = r(1-\Delta)p_s \\ p_{LPLP} = r(1-\Delta)(1-p_s) \\ \lambda_{BM} = \eta_{BM}\gamma_{BM} \\ \lambda_{LP} = \eta_{LP}\omega_{LPLM} \\ \lambda_{LM} = \eta_{LM}\gamma_{LM} \end{cases}. \tag{5.7}$$

Here, $\Delta = [-1;1]$ represents the homeostatic imbalance, and, more specifically, the system is hyperproliferating if $\Delta > 0$, it vanishes if $\Delta < 0$ and it is homeostatic if $\Delta = 0$. The total fraction of asymmetric division is represented by $r = [0;0.5]$; when equal to 0, there are only asymmetric divisions and, in this case, the system is homeostatic for any value of Δ. The values of $p_a = [0;1]$ and $p_s = [0;1]$ determine the outcome respectively of the asymmetric and symmetric cell divisions. Lastly, parameters η_x, for $x =$ BM, LP and LM, are the ratio between the division and the transition or death rates: the system is homeostatic only if $\eta_x < 1$.

For a more efficient fitting, we scale the problem by dividing all the rates in Equation (5.5), α_i, by a reference rate. For doing that, we choose λ_{BS}, i.e. $\tilde{\alpha}_i = \alpha_i/\lambda_{BS}$ and, consistently, we scale the simulation time, t, as $\tilde{t} = t\lambda_{BS}$. Crucially, the solution of the

scaled problem, that is, based on $\tilde{t}$ and $\tilde{\alpha}$, is the same as any other with consistent unscaled rates and time. Thus, the same numerical simulation can be used for any value of λ_{BS}. Given that, the fitting problem is written in terms of $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = \left(\lambda_{BS},\ \ r,\ \ \Delta,\ \ p_a,\ \ p_s,\ \ \eta_{LP},\ \ \tilde{\omega}_{LPLM},\ \ \eta_{LM},\ \ \tilde{\gamma}_{LM},\ \ \eta_{BM},\ \ \tilde{\gamma}_{BM}\right). \tag{5.8}$$

To find the set of parameters that best fit the data, we follow the classical *Maximum A Posteriori* (MAP) approach [Box et al., 1992], which allow us to combine the clonal data, $\mathcal{D}_{CD} = \{\mathcal{D}_{CD-I}, \mathcal{D}_{CD-II}\}$, and the scRNA-seq data, $\mathcal{D}_{scRNA} = \mathcal{D}_{scRNA-I}$, as defined in Section 5.3. We therefore write the posterior, $P(\boldsymbol{\theta}|\mathcal{D})$, where $\mathcal{D} = \{\mathcal{D}_{CD}, \mathcal{D}_{scRNA}\}$, as

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathcal{L}(\mathcal{D}_{CD}|\boldsymbol{\theta})\mathcal{L}(\mathcal{D}_{scRNA}|\boldsymbol{\theta})}{P(\mathcal{D})} P(\boldsymbol{\theta}), \tag{5.9}$$

in which $\mathcal{L}(\mathcal{D}_{CD}|\boldsymbol{\theta})$ and $\mathcal{L}(\mathcal{D}_{scRNA}|\boldsymbol{\theta})$ are respectively the likelihood of the clone size and scRNA-seq data, $P(\mathcal{D})$ is a normalisation factor, and $P(\boldsymbol{\theta})$ is the prior probability. Assuming no prior knowledge of the parameters, we assume a uniform prior in the parameter search space Θ, which means that $P(\boldsymbol{\theta}) = P_\Theta$ is a constant.

Since the clonal size frequency at each time point, $\mathcal{D}_{CD-t}$, with t =I, II, is composed by a set of independent counts, the likelihood function is a multinomial distribution with a countable number of outcomes [Doupé et al., 2012, Frede et al., 2016]

$$\mathcal{L}(\mathcal{D}_{CD-t}|\boldsymbol{\theta}) = \frac{\left(\sum_n f_n^{(t)}\right)!}{\prod_n f_n^{(t)}!} \prod_n p_n^{(t)}(\boldsymbol{\theta})^{f_n^{(t)}}, \tag{5.10}$$

in which $f_n^{(t)}$ is the frequency of having a clone with n cells in the experimental data taken at the time t and $p_n^{(t)}(\boldsymbol{\theta})$ is the probability of having a clone with n cells at the same time point t in the simulation based on the set of parameter $\boldsymbol{\theta}$. Considering that uncorrelated data will be available at different time points, the likelihood of the global set of data $\mathcal{D}_{CD}$ will be the product of Equation (5.10) computed at each time point. In the same way, the likelihood of the scRNA-seq data follows a multinomial distribution, under the assumption that the cell counts are independent. Considering the cluster identities x = B, LP and LM, this is

$$\mathcal{L}(\mathcal{D}_{scRNA}|\boldsymbol{\theta}) = \frac{\left(\sum_x s_x\right)!}{\prod_x s_x!} \prod_x p_x(\boldsymbol{\theta})^{s_x}, \tag{5.11}$$

in which s_x, is the measured relative size of cluster x and $p_x(\boldsymbol{\theta})$ is the probability of a cell to be in cluster x in the simulation based on the set of parameter $\boldsymbol{\theta}$.

Substituting Equations (5.10) and (5.11) into Equation (5.9) and rearranging the terms, results in

$$P(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{K}(\mathcal{D})\mathcal{J}(\boldsymbol{\theta},\mathcal{D}), \tag{5.12}$$

in which

$$\mathcal{K}(\mathcal{D}) = \frac{\left(\sum_n f_n^{(I)}\right)!\left(\sum_n f_n^{(II)}\right)!\left(\sum_x s_x\right)!}{\prod_n f_n^{(I)}!\prod_n f_n^{(II)}!\prod_x s_x!}\frac{P_\Theta}{P(\mathcal{D})} \tag{5.13}$$

is a constant term accounting for all the normalisation factors and the uniform prior, and

$$\mathcal{J}(\boldsymbol{\theta},\mathcal{D}) = f_{CD-I}(\boldsymbol{\theta},\mathcal{D})f_{CD-II}(\boldsymbol{\theta},\mathcal{D})f_{scRNA-I}(\boldsymbol{\theta},\mathcal{D})\,, \tag{5.14}$$

where

$$f_{CD-t}(\boldsymbol{\theta},\mathcal{D}) = \prod_n p_n^{(t)}(\boldsymbol{\theta})^{f_n^{(t)}} \text{ for } t = I, II \tag{5.15}$$

and

$$f_{scRNA-I}(\boldsymbol{\theta},\mathcal{D}) = \prod_x p_x(\boldsymbol{\theta})^{s_x}\,. \tag{5.16}$$

We then define the best fitting parameters, $\boldsymbol{\theta}^*$, as the solution of the unconstrained optimisation problem

$$\max_{\theta\in\Theta} P(\boldsymbol{\theta}|\mathcal{D})\,, \tag{5.17}$$

in which $P(\boldsymbol{\theta}|\mathcal{D})$ is given by Equation (5.12). Since the term $\mathcal{K}(\mathcal{D})$ is constant, and $\mathcal{J}(\boldsymbol{\theta},\mathcal{D})$ might result in very low values spanning several orders of magnitude (it is the product of probabilities), this can be equivalently written as[10]

$$\min_{\theta\in\Theta} -\log_{10}\mathcal{J}(\boldsymbol{\theta},\mathcal{D})\,, \tag{5.18}$$

in which $\mathcal{J}(\boldsymbol{\theta},\mathcal{D})$ is given by Equation (5.14).

We now observe that the MAP problem defined in Equation (5.18), is characterised by a high level of complexity for the following reasons: a) it is a high-dimensional problem, in which the optimisation variable $\boldsymbol{\theta}$, defined in Equation (5.8), has eleven components; b) the estimation of the model probabilities, $p_n^{(I)}(\boldsymbol{\theta})$, $p_n^{(II)}(\boldsymbol{\theta})$ and $p_x(\boldsymbol{\theta})$, is computationally expensive and it is characterised by noise since it is based on numerical simulations of a stochastic process; and c) the optimisation function might have multiple local minima, and the *optimal solution*, defined as the set of model parameters that satisfies the optimisation problem in Equation (5.18), might be not unique. In other words, there might be multiple and potentially infinite optimal solutions.

[10]We set up an equivalent minimisation problem corresponding to the posterior maximisation since numerical methods work on minimising an objective function.

Some high-level considerations on the methodology used to solve the MAP problem are reported below, while details are given in Appendix C.2.1. First of all, we remark that since we use a numerical approach, we will call *optimal fitting(s)* the approximation of the optimal solution(s). To deal with a) and c), we chose to run a global optimisation algorithm multiple times. Among the available options, we selected the Bayesian and the surrogate optimisation algorithms (*bayesopt* and *surrogateopt* functions of Matlab), given their exceptionally high efficiency in approximating the global optimum with a relatively small number of function evaluations. Concerning b), we balanced the number of simulated clones and the noise in the simulation outputs. Since this process does not lead to highly accurate results, we consider as optimal fittings all those cases for which the objective function is within a certain threshold from a reference value. In the case under analysis, this reference value is related to the objective function corresponding to the true model. However, this is generally not known, and the global minimum value found could be used instead. In addition, to improve the algorithm's convergence, we also implemented a smoothing of the distribution tails based on a moving average method[11] (*smooth* function of Matlab). Finally, as before mentioned, to further increase the speed of the fitting process, we take advantage of the time rescaling in the numerical simulation and rewrite the fitting parameter vector $\boldsymbol{\theta}$ in Equation (5.8) as

$$\boldsymbol{\theta} = \left(\lambda_{BS}, \quad \tilde{\boldsymbol{\theta}}\right) . \tag{5.19}$$

For each value of $\tilde{\boldsymbol{\theta}}$, we run a single stochastic simulation considering $\lambda_{BS} = 1$, and then chose the value of λ_{BS} that maximise the objective function in (5.14).

5.4.2 Fitting results

The above-described optimisation problem results in multiple optimal fittings. In this section, we focus on five illustrative cases, which are described and compared below. The details of the optimisation runs and results are reported in Appendix C.2. Crucially, since data is based on a synthetic database, values and profiles for the true model, which, in principle, are not known, are also reported; they are labelled as **T**.

For the selected optimal fittings, the kinetic parameters and division outcome probabilities of the model (see Equation (5.5) and Equation (5.6)) are reported in Table 5.4. In this table, optimisation variables Δ and r, representative of the homeostasis imbalance and the ratio of asymmetric division, are also reported together with the objective function defined by Equation (5.14) (and Equation (5.20),

[11] Considering the relatively low number of clones in the simulation, it is likely that zeros will appear in the tail due to the noise. A clone size characterised by zero frequency in the model, $p(\theta)$, and a non zero frequency in the observation, f_n, results in a zero likelihood ($-\infty$ in logarithmic scale). This situation, which is only due to the noise in the numerical simulations, degrades the likelihood estimation and, therefore, the optimiser's convergence.

which will be introduced later). For clarity, the objective function values are reported in the logarithmic scale and relative to the true model, meaning that negative (positive) values correspond to cases that fit better (worse) than the true model. The clusters size time evolution is shown in Figure 5.7; the clone size distribution is shown in Figure 5.8. Additionally, in Figure 5.9, the mean value extracted from the processed clonal data (left) and of the surviving clones (right) is shown. In Figure 5.10 we report the tissue and the average clonal dynamics based on the integration of the ODEs, given by Equation (2.6), associated with the cell fate model. When applicable, in all these figures, the data and the $\pm 2\sigma$ data variability are shown.

Parameter	**Value**				
Fitting	**NH.1**	**H.1**	**H.2**	**H.3**	**H.1$^+$**
Cell Fate Model Parameters					
λ_{BS} [w^{-1}]	1.589	1.438	1.271	1.015	2.033
λ_{LP} [w^{-1}]	7.350	2.898	4.423	5.294	1.716
ω_{LPLM} [w^{-1}]	7.848	2.899	5.173	5.750	2.297
λ_{LM} [w^{-1}]	0.515	2.661	0.454	8.458	0.172
γ_{LM} [w^{-1}]	1.904	3.092	1.419	9.519	0.592
λ_{BM} [w^{-1}]	1.646	0.139	1.070	3.278	2.562
γ_{BM} [w^{-1}]	7.694	0.240	8.278	3.912	17.922
p_{BSBS}	0.371	0.250	0.472	0.018	0.374
p_{BSBM}	0.101	0.092	0.008	0.100	0.211
p_{BSLP}	0.189	0.408	0.048	0.863	0.042
p_{BMBM}	0.127	0.185	0.010	0.004	0.140
p_{LPLP}	0.212	0.065	0.462	0.014	0.234
Δ	0.046	0.000	0.000	0.000	0.000
r	0.355	0.250	0.472	0.018	0.374
Objective Function					
f_{CD-I}	0.358	0.977	0.265	0.713	-0.139
f_{CD-II}	0.329	-0.727	0.171	0.967	0.186
$f_{scRNA-I}$	-0.512	-0.534	-0.492	-0.548	-0.553
$\mathcal{J}$	0.175	-0.284	-0.057	1.131	-0.506
f_{CD-III}					0.456
f_{TV}					0.335
$\mathcal{J}^+$					0.285

TABLE 5.4: Summary of five illustrative optimal fittings in terms of cell fate model parameters and objective function. The objective function $\mathcal{J}$ and each contribution are defined in Equation (5.14); the final rows correspond to $\mathcal{J}^+$, which will be introduced later in Section 5.4.2.3. Concerning the objective function, for each row x, where $x = f_{CD-I}, f_{CD-II}, ..., \mathcal{J}^+$, values reported are $-\log_{10}(x/x_T)$, in which x_T is the value of x corresponding to the true model. Hence, positive (negative) values mean a fitting that is worse (better) than the true model.

5.4.2.1 Non-homeostatic cell fate models

We first focus on the optimal fitting **NH.1**, which is computed without requiring the cell fate model to be homeostatic. By looking at Figure 5.8 and Figure 5.7, it is clear

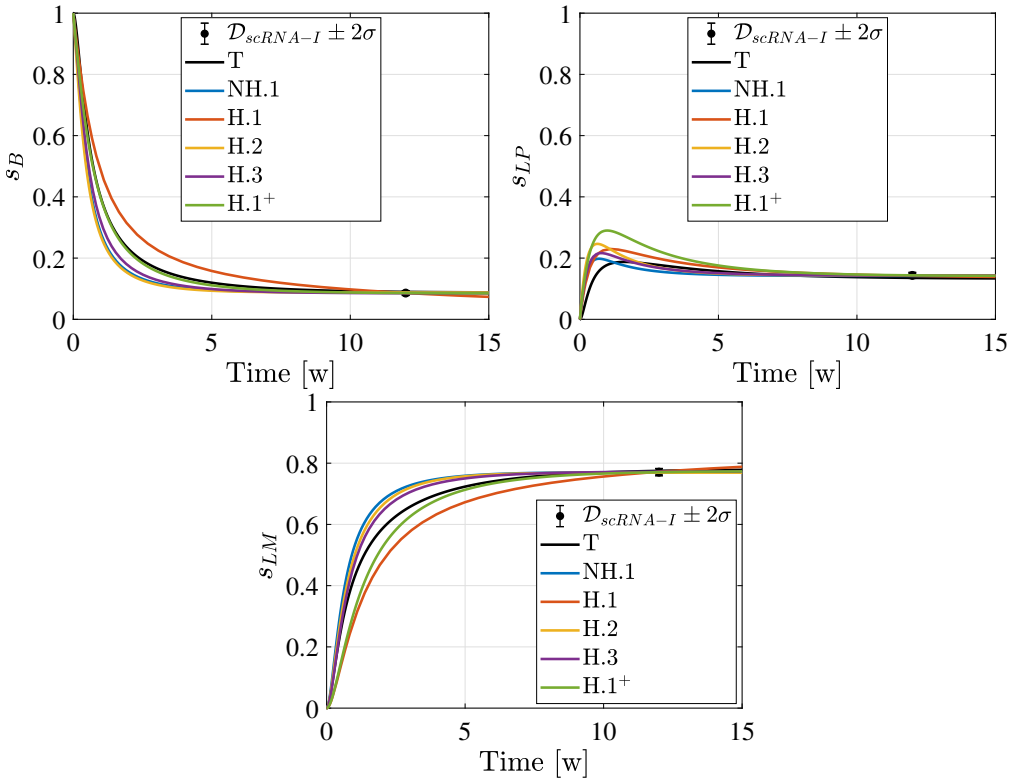


FIGURE 5.7: Time evolution of the clusters size, s_x for $x = B, LP, LM$, for five illustrative optimal fittings (see model parameters in Table 5.4) compared to $\mathcal{D}_{scRNA-I}$. Clusters correspond to basal (top-left), luminal progenitor (top-right) and luminal mature (bottom). Data 2σ variability and the true model curve, labelled as T, are also shown.

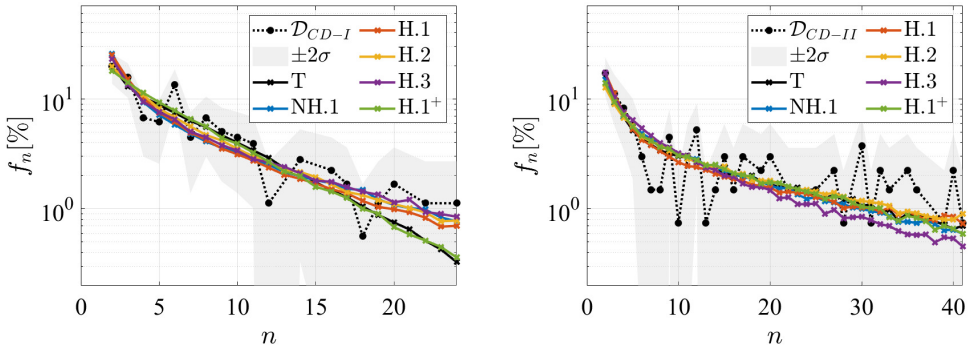


FIGURE 5.8: Profiles of the clone size distribution associated with $\mathcal{D}_{CD-I}$ (left) and $\mathcal{D}_{CD-II}$ (right) for five illustrative optimal fittings (see model parameters in Table 5.4). Data 2σ variability and the clonal statistics for the true model, labelled as T, are also shown.

that this case is a good fit of the available data. This is further confirmed by the fact that the value of the objective function, reported in Table 5.4, is very close to that of the true model. Importantly, a small positive value relative to the true model means that this model is slightly worse than the true one but not too far. However, this fitting is not representative of the actual dynamics since it is not homeostatic, as shown in Figure 5.10 (left). In this particular case, dynamics are globally hyperproliferating and

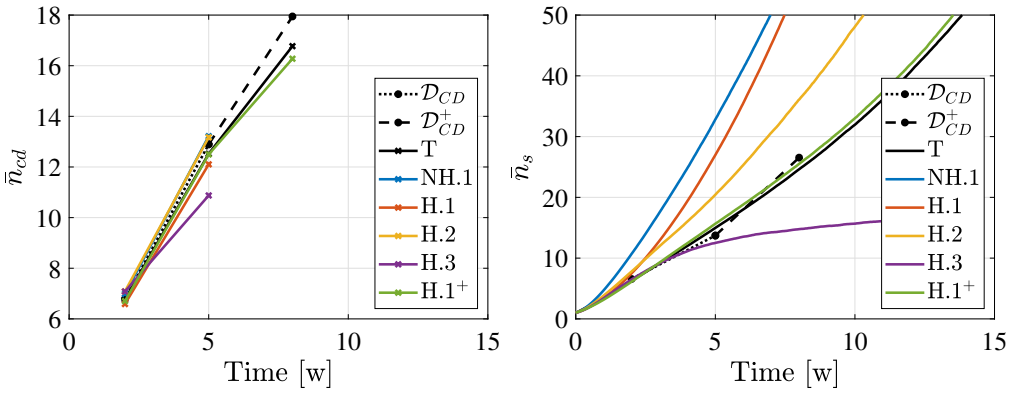


FIGURE 5.9: Mean value of the clone size distribution, $\bar{n}_{cd}$, (left) and mean cell number in the surviving clones, $\bar{n}_s$, as function of the time (right) for the five illustrative optimal fittings (see model parameters in Table 5.4). Values of $\bar{n}_{cd}$ differ from $\bar{n}_s$ since they do not consider single-cell clones and those in the tail of the distribution. The data point indicated as $\mathcal{D}^+_{CD}$ refers to an additional clonal data point that will be discussed in Section 5.4.2.3. Values corresponding to the true model, labelled as T, are also shown.

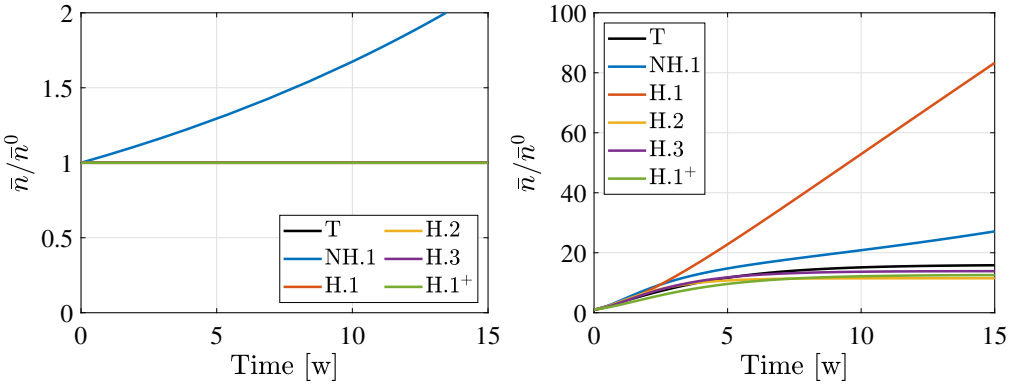


FIGURE 5.10: Mean total cell number evolution as a function of time, based on the integration of the system of ODEs, Equation (2.6), for the five illustrative optimal fittings (see model parameters in Table 5.4). The initial condition, $\bar{\boldsymbol{n}}_0$, is proportional to the dominant eigenvalue (left), as representative of the tissue dynamics, and one B-cell (right) as representative of the average dynamics of labelled clones (neglecting extinction). Curves for the tissue dynamics are overlapped, except for the non-homeostatic case, **NH.1**.

therefore feature a growing total cell number. Vanishing dynamics can fit the model as well, as shown in Appendix C.2.2.

Globally, this result confirms the importance of imposing a priori the homeostasis conditions derived in Chapter 2. In fact, non-homeostatic cell fate models that fit the data might exist and be equally good as the homeostatic ones, hence misleading the search for the correct model.

5.4.2.2 Homeostatic models

We now focus on homeostatic cell fate models obtained by fixing the imbalance parameter, $\Delta = 0$, and asking $\eta_x < 1$, for $x =$ LP, LM and BM. Furthermore, following the results presented in Chapter 4, we visually inspected the shape of the clone size distribution and the time evolution of the mean of surviving clones. In particular, the rescaled clonal data is shown in Figure 5.11 (left), together with a reference exponential distribution with unitary mean. In Figure 5.11 (right), we show the mean of the surviving clones based on both the processed data, that is, removing the single-cell clones and the distribution tail, and the raw data (black line), which include all the clones; here, a point at time zero and $\bar{n} = 1$ is also added as representative of the initially labelled single cells. We observe that clonal data looks exponentially shaped, and the mean of the surviving clones presents an increasing trend. However, the two data points are not in the long-term, and thus, a linearly increasing trend cannot be distinguished from a plateau, which would be observed later. Given that, we deduce that, in this case, it is not possible to a priori exclude any of the two self-renewing strategies and, for this reason, we keep the search bounds of the asymmetric division parameter, r, in the whole range [0; 0.5].

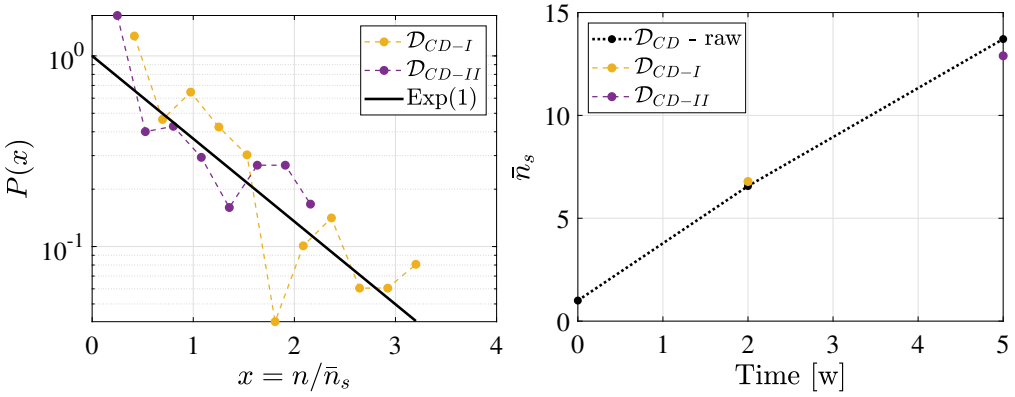


FIGURE 5.11: Rescaled clone size distribution of the data (left) compared with an Exponential distribution with unitary mean (black line). The mean clone size is shown as a function of time (right). The black line, indicated as $\mathcal{D}_{CD}$ - raw, corresponds to the mean of the surviving clones and includes a point at time zero with only single cell clones. The points $\mathcal{D}_{CD-I}$ and $\mathcal{D}_{CD-II}$ are computed based on the processed data where single-cell clones and those in the tail of the distribution are removed.

In Figure 5.12, the results of the optimisation runs are shown in terms of asymmetric division parameter, r and objective function relative to the value for the true model. We must recall that all the points below a certain threshold, which we fixed equal to 2, correspond to optimal fittings of the data. They are highlighted in blue and labelled as **H**. Besides, the points below zero indicate cases that fit the data better than the true model. In this figure, the global optimum fitting, labelled as **H.1**, is characterised by the overall minimum value of the objective function. However, from this figure, it is apparent that despite there is a slight decreasing trend of the objective function with r,

there is no significant difference between models with mostly asymmetric division, i.e. $r \approx 0$, and those with primarily symmetric divisions, i.e. $r \approx 0.5$. Given that, in addition to the global optimal fitting, **H.1**, we choose two other illustrative cases, **H.2** and **H.3**. The case **H.2** is equivalent in terms of the objective function to **H.1**, but it exhibits different dynamics, as shown later. Instead, **H.3** is chosen among the optimal fittings for its low value of r. In this case, the objective function is slightly higher but still reasonably low.

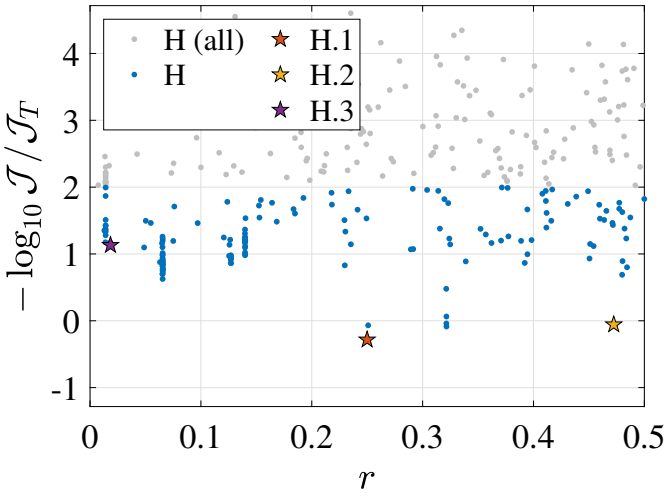


FIGURE 5.12: Value of the objective function, relative to that of the true model, as a function of the ratio of asymmetric divisions. Points highlighted in blue and labelled as **H** are those below a threshold equal to 2, and further analysed in Figure 5.13. The global optimum is indicated as **H.1**. Two other illustrative optimal fittings, **H.2** and **H.3**, are also indicated.

Considering all the optimal fittings **H**, the variability of some model parameters is shown in Figure 5.13. As a reference, the selected illustrative fittings, **H.1-3**, are also shown. We note that there is wide variability in all the parameters shown, but whilst in some cases, there is a clear correlation among them (top panels), in others, model parameters are completely uncorrelated (bottom). Finally, in some cases, a trend is visible but with large dispersion of the parameters (middle panel). Thus, in general, the self-renewing strategy cannot be distinguished by the available data, and the optimal fittings present a large variability in the parameters. This results in very different cell fate models that are equally good fitting of the data.

Focusing now on the selected illustrative cases, as shown in Figure 5.8 and Figure 5.7, they all fit the data very well, including the case **H.3**, which is the one with the highest objective function value and it is representative of a cell fate model with almost exclusively asymmetric division in the self-renewing compartment. Interestingly, considering the mean of the surviving clones, shown in Figure 5.9 (right), the case **H.3** is the closest one to the data among the three analysed here. Thus, this means that even if, formally, **H.3** is not best fitting, the difference in the objective function with

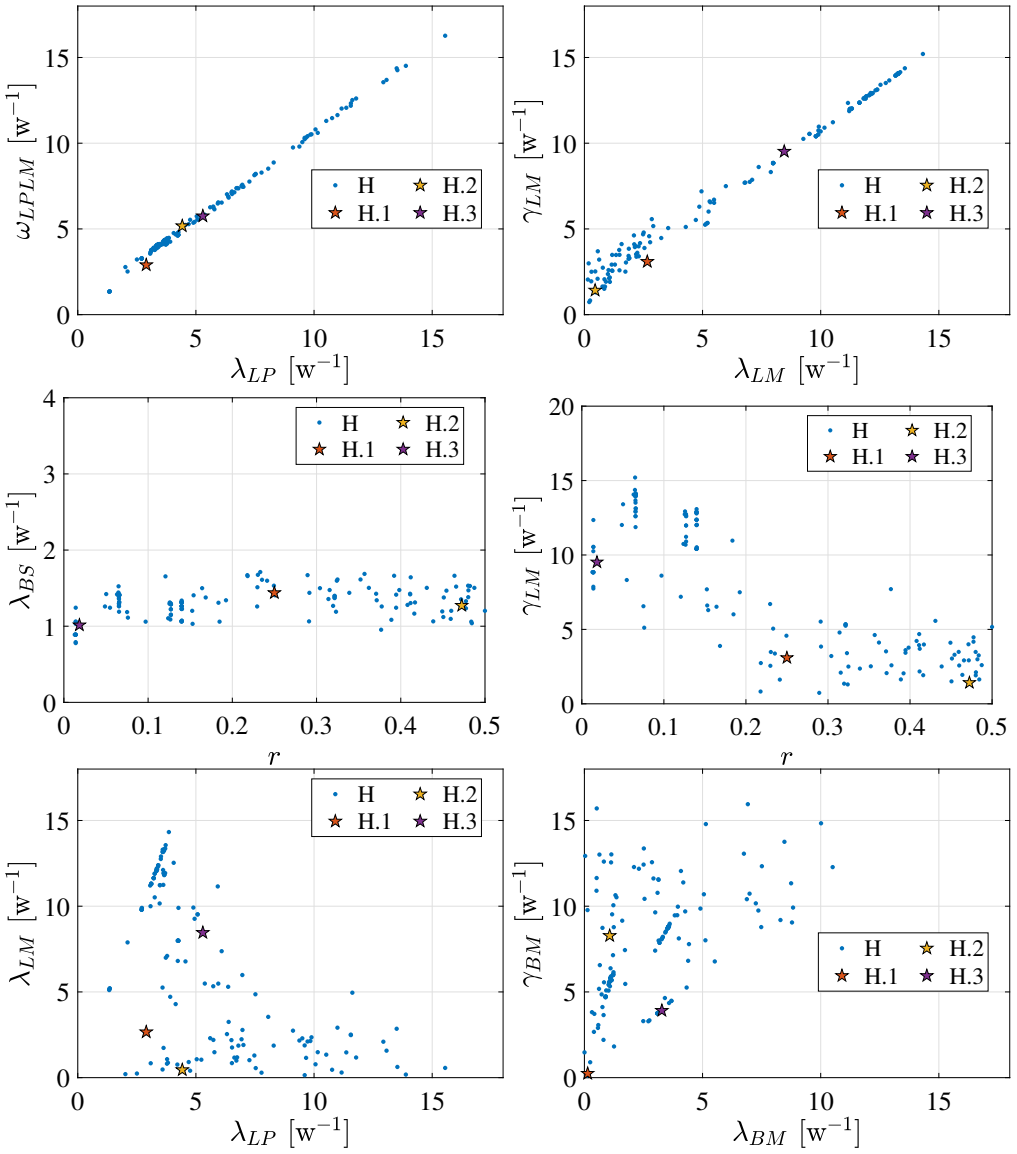


FIGURE 5.13: Model parameters for the optimal fittings, **H**, and the selected illustrative cases, **H.1-H.3**. In most of the cases, parameters present a wide variability; they might be highly correlated (top panels), show a trend but with large dispersion (middle panel) or be completely uncorrelated (bottom panel). We remark that in the bottom-right plot, the half-plane $\lambda_{BM} \geq \gamma_{BM}$ is not reachable since the requirement for homeostasis is applied; in the remaining part of the plane, parameters are uncorrelated.

respect to the other cases is not sufficient to justify the rejection of the asymmetric division strategy.

Furthermore, as expected, the illustrative cases are associated with a homeostatic cell fate model, as demonstrated in Figure 5.10 (left), where all the three curves remain constant and equal to one (note that they are not visible since they are overlapped). However, looking at the average clonal dynamics that is shown in Figure 5.10 (right), we observe that in the time frame analysed, the global optimum fitting, **H.1**, behaves as a non-homeostatic model. This dynamical behaviour is strictly related to the value

of λ_{LP} that is almost equal to ω_{LPLM}, meaning that, for this cell fate model, the luminal progenitor cell type is almost a self-renewing cell type (it cannot be strictly self-renewing unless the system would be non-homeostatic). Consequently, the turnover time associated with this dynamics, estimated to be over 3000 weeks, is considerable compared to the timescale of the experiment (we recall that according to Dr Elias, turnover time is around 8-12 weeks). Instead, the timescale of the dynamics for **H.2** and **H.3** is in line with that of the experiment. Therefore, in a scenario where actual experimental data were used, we probably would have excluded **H.1** (and all those optimal fittings presenting the same structure) by qualitative means, for example, by filtering them out based on the value of the turnover time or by limiting the variability in the η_{LP} parameter. However, since we are using synthetic data, we properly address this problem in the next section by assuming the availability of a measure of the turnover time.

5.4.2.3 Fitting based on an enriched dataset

For the case under analysis, the solutions of the MAP problem result in multiple optimal fittings associated with different dynamics (i.e. degeneracy of the model) and, importantly, data is not sufficient to distinguish the self-renewing strategy. Considering that we are working here on synthetic data, we can easily generate additional data to enhance the dataset and analyse how this affects the fitting.

First of all, we note that to really have a clear idea of how the mean of the surviving clones is evolving, we need a clonal data in the long-term; for that reason, we generate an additional set of clonal data at eight weeks, $\mathcal{D}_{CD-III}$, shown in Figure 5.14. Furthermore, to exclude the dynamics with long timescales, as the optimal fitting **H.1**, we assume to have a measure of the turnover time, t_{TV}, and its expected variability, σ_{TV}, that is, $\mathcal{D}_{TV} = \{t_{TV}, \sigma_{TV}\}$. To include these additional data, indicated as $\mathcal{D}^+$, in the optimisation problem, we define a new objective function as

$$\mathcal{J}^+(\boldsymbol{\theta}, \mathcal{D}, \mathcal{D}^+) = \mathcal{J}(\boldsymbol{\theta}, \mathcal{D}) f_{CD-III}(\boldsymbol{\theta}, \mathcal{D}^+) f_{TV}(\boldsymbol{\theta}, \mathcal{D}^+)\,, \tag{5.20}$$

in which $\mathcal{J}(\boldsymbol{\theta}, \mathcal{D})$ is the objective function as defined by Equation (5.14), $f_{CD-III}(\boldsymbol{\theta}, \mathcal{D}^+) = \prod_n p_n^{(III)}(\boldsymbol{\theta})^{f_n^{(III)}}$ represents the likelihood of the clonal data $\mathcal{D}_{CD-III}$, and $f_{TV}(\boldsymbol{\theta}, \mathcal{D}^+)$ represent the contribution of a prior based on $\mathcal{D}_{TV}$. More specifically, $f_{TV}(\boldsymbol{\theta}, \mathcal{D}^+)$ is related to a Normal distribution with mean t_{TV} and variance σ_{TV}^2.

We analyse now the variability in the asymmetric division parameter, r, of the optimal fittings of the above-described optimisation problem. They are shown in Figure 5.15. As done in the previous section, the optimisation results are shown in terms of the objective function, scaled by the corresponding true model value, as a function of the ratio of symmetric division. The points below a threshold equal to 2 (we used the

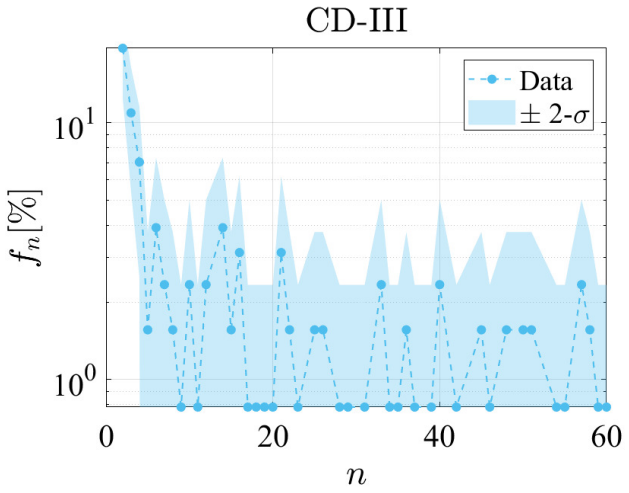


FIGURE 5.14: Clone size distribution in terms of relative frequency, f_n, of the clones at time point **CD-III**. Data correspond to 200 uncorrelated clones, filtered to remove single-cell clones and the tail. In addition to the data, the expected 2σ variability is also shown.

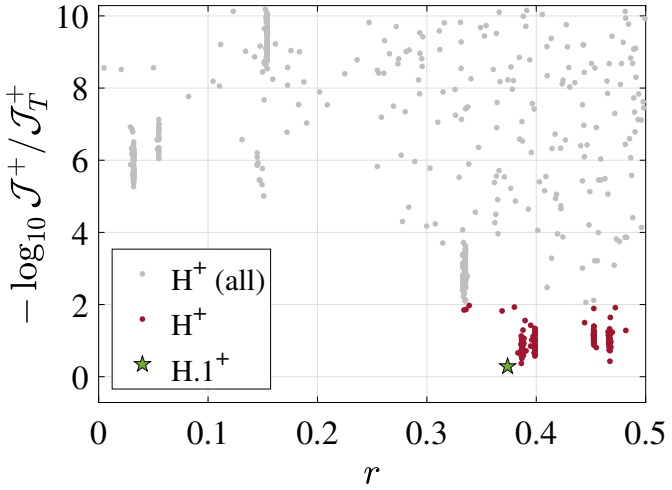


FIGURE 5.15: Value of the objective function, relative to the value for the true model, as a function of the ratio of asymmetric divisions. Points highlighted in red and labelled as $\mathbf{H}^+$ are those below a threshold equal to 2. The global optimum is indicated as $\mathbf{H.1}^+$.

same value as in Section 5.4.2.2) are highlighted in red and labelled as $\mathbf{H}^+$. The global optimum fitting is indicated as $\mathbf{H.1}^+$.

We first note that, in this case, there are no cases that fit the data better than the true model, although $\mathbf{H.1}^+$, and a few others, are very close to that. This observation is consistent with the fact that, in general, adding more data results in a more complex fitting problem. Also, the decreasing trend of the objective function with increasing values of r is much more evident than that shown in Figure 5.12 if only $\mathcal{D}$ is fitted. Crucially, optimal fittings $\mathbf{H}^+$ feature only $r > 0.33$. Therefore, this dataset allows for identifying the self-renewing strategy, which agrees with the true model. As a further

confirmation that the self-renewing strategy is, in this case, based on a population asymmetry pattern, we analyse in Appendix C.2.2 a fitting corresponding to the best one found for low values of r.

Consistently, if we look at the corresponding model parameters, shown in Figure 5.17, the variability of the $\mathbf{H}^+$ optimal fittings is significantly reduced if compared to that of the $\mathbf{H}$ ones. Some exceptions are found in parameters such as γ_{BM} and λ_{BM} (bottom right panel) where a wide variability remains. Analysing the global optimal fitting, $\mathbf{H.1}^+$, we note that this model is a good fitting of the data $\{\mathcal{D}, \mathcal{D}^+\}$, as shown in Figure 5.8, Figure 5.7 and Figure 5.16. Besides, looking at the value of the objective function $\mathcal{J}$, reported in Table 5.4 (i.e. based only on $\mathcal{D}$), this fitting is actually the best one. This observation suggests that adding more data to the fitting problem reduces the dispersion in the parameters and might help the optimiser in converging to good fittings.

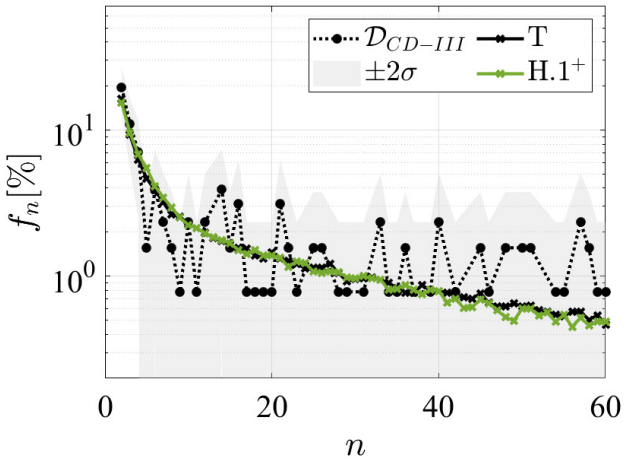


FIGURE 5.16: Clone size distribution associated with $\mathcal{D}_{CD-III}$ for the global optimal fitting $\mathbf{H.1}^+$ (see model parameters in Table 5.4). Data 2σ variability and the clonal statistics for the true model, labelled as T, are also shown.

5.5 Conclusion

This chapter applied the developed framework to assess the cell fate dynamics in a study case. Based on the literature review and synthetic data, we emulated two lineage tracing experiments for the adult healthy mouse mammary gland. One experiment was related to the single-cell RNA-sequencing, the other to clone lineage tracing. Building on this, we validated the developed methodology and defined a pathway for assessing cell fate dynamics whenever such experimental data are available.

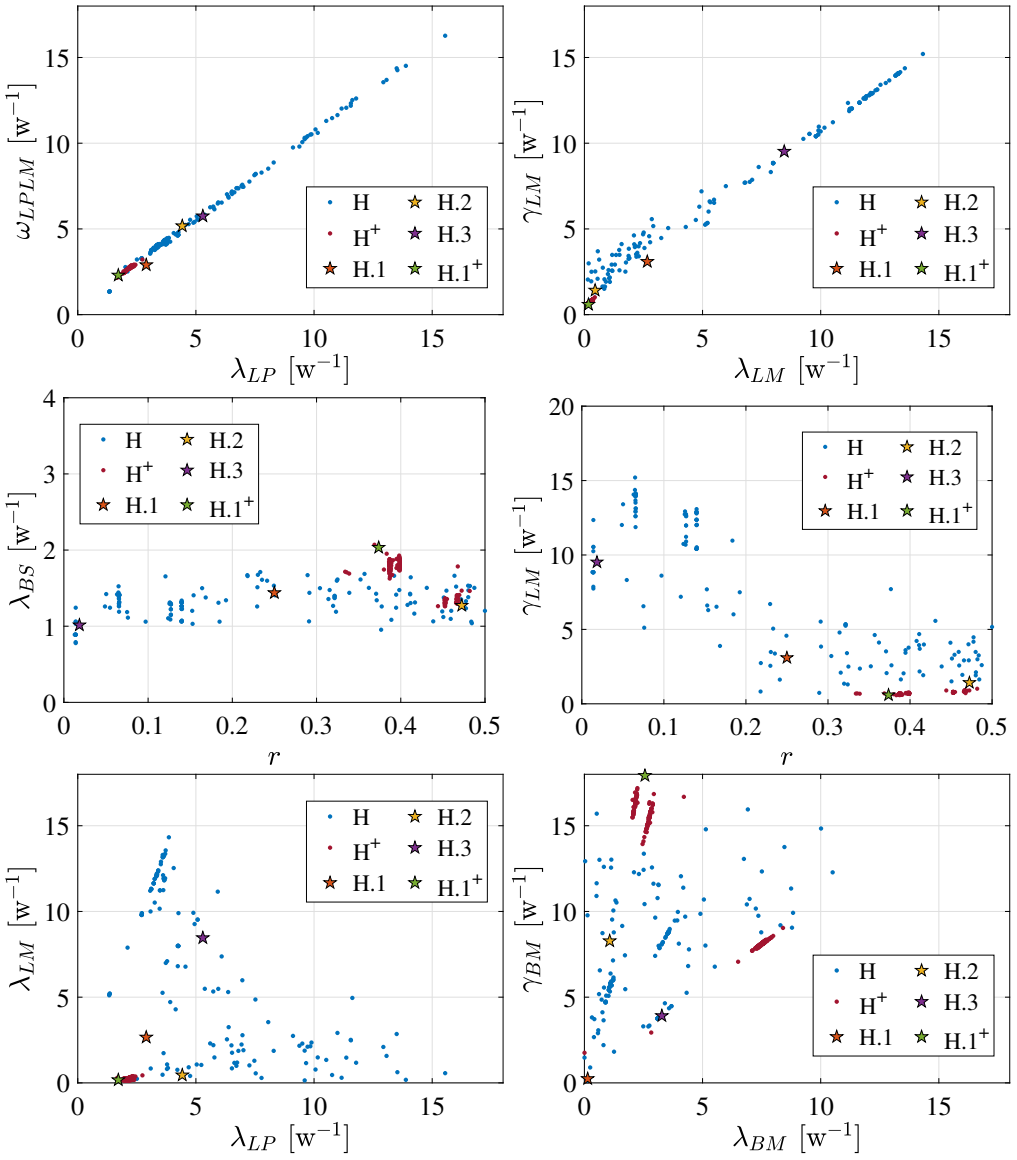


FIGURE 5.17: Comparison of the model parameters shown in Figure 5.13 for the optimal fittings $\mathbf{H}$ and $\mathbf{H}^+$; the selected illustrative cases are also shown. In $\mathbf{H}^+$, the variability of the model parameters is significantly reduced, with the only exception of γ_{BM} and λ_{BM} kinetic parameters (bottom-right panel).

The main idea behind the definition of a cell fate model was to restrict the possible candidate models to those homeostatic (outcome of Chapter 2) and that present qualitative features compatible with the experimental data (outcome of Chapter 4). Although this approach reduces the possible cell fate network configurations and model parameters, there are still infinite possibilities of cell fate modelling. Thus, we first addressed the definition of a sensible cell state network and then applied a standard Bayesian inference approach to the derived specific cell state network.

To determine the cell state network for the study case, we made assumptions supported by the literature whenever possible. In this context, we analysed available scRNA-seq data providing a consistent comparison. In this way, we gained insight

into the study case scenario and highlighted discrepancies and open points that should be addressed by further dedicated experimental work. Based on this, we defined a model with four single-state types for fitting, including two basal (stem and mature) and two luminal (progenitor and mature) cell states. This cell fate model also respects a clear hierarchy in the cell types, where stem cells are linked to progenitors and progenitors to mature ones.

Concerning the model parameters fitting, to introduce the minimum possible bias in the analysis, we generated the synthetic data based on a random cell fate model that was previously analysed in Chapter 4. In particular, we chose a model where the self-renewing strategy is based on a generalised population asymmetry model, i.e. where cells of self-renewing type could divide symmetrically. We successfully set up and solved an optimisation problem in which the objective is to maximise the non-normalised posterior probability of the model parameters given the data. Notably, we showed that non-homeostatic cell fate dynamics fit the data equally well than homeostatic models, meaning that the rules derived in Chapter 2 for achieving homeostasis are essential for adequate modelling of the dynamics. Additionally, several optimisation runs resulted in different fittings and showed that the modelled synthetic data, based on mid-term clonal statistics and scRNA-seq data, is insufficient to determine the self-renewing strategy. Therefore, we proposed and assessed the use of an additional clonal data point in the long term and a measure of the timescale of the dynamics. With this enriched dataset, we significantly reduced the variability in the model parameters and, importantly, successfully determined the self-renewing strategy.

Chapter 6

Conclusion and future work

Many biological tissues are continuously renewed through cycles of cell production and cell loss. Homeostasis is the condition in which cell proliferation and death are correctly balanced to maintain the tissue's healthy state. The underlying dynamics of tissue homeostasis are complex and not always well understood. Importantly, experimental data alone are often insufficient to infer the cell fate model, and thus, mathematical modelling of cell states and cell fate dynamics is essential to understand this mechanism better.

Therefore, this project aimed at improving the understanding of homeostasis in adult renewing tissues by developing a methodology for the mathematical modelling of cell fate dynamics given experimental data. More specifically, the idea was to determine the cell states, their interconnection, and the proliferation, transition and death rates that define a cell fate dynamical model. This problem translates into the model identification and parameters definition, which results in a difficult task. Challenges are related to the number of unknowns, the scarcity and uncertainty in the data, and the noise, features that increase the risk of model overfitting and degeneracy. Therefore, we restricted the search to those models compatible with homeostasis and those presenting specific tissue-related features to simplify the fitting problem substantially. For doing so, we studied the cell fate dynamics using theoretical and numerical means.

As a study case, we focused on the mouse mammary gland, which is of particular interest given its complexity and still unresolved features. Although numerous studies have been carried out, many controversies about the cell identity and the lineage hierarchy characterise this tissue.

The following sections report the conclusions of this work and the innovative contributions in the field; Section 6.2 highlights some limitations and future work.

6.1 Conclusion

To define a cell fate model given experimental data, we set up four objectives. Each of them is addressed in a dedicated chapter of this thesis. The key findings are summarised below.

6.1.1 Objective 1

We recall that the first objective, **Obj. 1** in Section 1.6, was to derive generic rules that constrain the structure of the dynamical model, based on which lineage hierarchies that are not compatible with homeostasis could be excluded a priori. Given that, we started from a generic model of cell fate dynamics based on a generalised multi-type branching process and defined a conceptual mathematical model representing the cell fate dynamics. This model features an arbitrary number of states, cell state transition, division and death.

Based on graph theory, the generic cell fate model was associated with a directed network, the cell state network, giving an intuitive view of the cell states and connections. This representation combined with the deterministic approach was essential for describing the tissue population dynamics in homeostasis. As a first approximation, we considered a linear model, for which homeostasis corresponds to a marginally stable system. By combining the mathematical and biological perspectives, we proposed an alternative definition of a cell type corresponding to the set of mutually reachable cell states, also known as a Strongly Connected Component of the cell state network. In this context, the adult stem cell is defined as a self-renewing cell type, i.e. an SCC characterised by a zero dominant eigenvalue of the associated adjacency matrix, μ. This cell type, if isolated, features marginally stable dynamics. Instead, committed cell types are characterised by $\mu < 0$, thus presenting a vanishing dynamics if no cell influx from other cell types is considered.

Besides the cell type classification based on the value of μ, we proved that any homeostatic cell fate model must follow strict rules, requiring self-renewing cells at each apex, and only there, of the lineage hierarchy. Notably, the applicability of this result was extended to non-linear dynamics, which are a more realistic model of tissue population. In this case, homeostasis is intended as a steady-state condition in which stability cannot be assessed in general, but it relies on the specific dependencies included in the model. Hence, the derived modelling allows excluding cell fate models that do not represent homeostatic dynamics in renewing tissues, a task that is essential in supporting the definition of a proper cell fate model.

6.1.2 Objective 2

Strictly related to the modelling homeostasis, if the lineage architecture requires, among other conditions, a self-renewing cell type at the apex of the lineage, only a perfect balance of proliferation and death enables the self-renewing capability. This condition is unfeasible in a real biological scenario where many environmental and cell-intrinsic factors affect these dynamics. Therefore, as a natural extension of the proposed cell dynamics modelling, we assessed the impact of a regulation mechanism that gives homeostasis robustness to perturbations and stochastic fluctuations. Even though this model has not been directly used in the problem fitting, this assessment is instrumental in justifying a constant parameters model. Notably, since the regulation introduces non-linearities in the cell fate models, we did not address the homeostasis regulation in general but focused on a possible mechanism of homeostasis control mediated by crowding feedback. With this analysis we addressed **Obj. 2** as defined in Section 1.6.

When homeostasis is regulated by crowding feedback, cells sense the density of cells in their local micro-environment and respond by adjusting their propensity to divide and differentiate. Based on mathematical modelling, both theoretical and numerical, we derived a condition under which cell dynamics regulated via feedback remain confined around a (dynamic) homeostatic condition. This condition, which holds under reasonable biological assumptions, also allows the verification of the homeostasis stability based on simple measurements of the sign of the dependency of the model parameters with the cell density.

A further investigation of cell fate dynamics regulated via crowding feedback resulted in the derivation of a necessary condition and a sufficient one for homeostasis in a strict sense, that is, an asymptotic stable steady-state. Whilst the necessary condition is similar to but less restrictive than the condition derived for a (dynamic) homeostatic state, the sufficient condition involves the sign of the dependency of some parameters of the dynamical model from the cell density. In some models, this sufficient condition could never be fulfilled. Still, it is a reliable and straightforward way in which stability might be assessed experimentally.

Furthermore, several situations perturb homeostasis in actual tissues, such as poisoning, diseases, injury, and cell mutations. In these cases, the global behaviour of the tissue might be affected, showing anomalies in the cell's homeostasis regulation. Therefore, after assessing the stability of the crowding feedback mechanism, we explored its robustness in two specific scenarios. First, we modelled feedback perturbations and failures, which apply, for example, in case of poisoning. This analysis showed that the redundancy of the feedback gives robustness in the homeostasis control. This feature means that feedback playing against stability might be compensated by others acting in favour of it. Significantly, this modelling was

qualitatively extended to the case of cell-intrinsic dysregulation, often associated with cell mutation. In this case, the whole tissue is affected and shows unstable dynamics only when the cell dysregulation cannot be compensated, and the mutated clone does not go extinct by chance. The relevance of this finding lies in the fact that it is well known that successive mutations are often associated with the early stage of cancer development.

A second scenario analysed includes perturbations in the lineage architecture that violate the rules for a homeostatic cell fate model. This case is representative of the depletion of the stem cells, for example, through poisoning or radiation, or the activation of quiescent stem cells, as in the tissue response to injury. This analysis showed that in a cell fate model regulated by crowding feedback, the self-renewing ability is a property that a cell type may acquire or lose depending on its position in the lineage hierarchy. In this context, we proposed the quasi-dedifferentiation, a condition for which committed cell types regulate homeostasis by becoming self-renewing if stem cells are completely removed. From a mathematical standpoint, this mechanism is a plausible alternative response to the experimentally observed cell dedifferentiation process, in which cells change their state in response to tissue damage.

6.1.3 Objective 3

Going back to identifying a cell fate model in homeostatic renewing tissues, we were able to discard all the non-homeostatic dynamics based on purely theoretical considerations. Nevertheless, the definition of a suitable cell fate model given some experimental data remained a complex problem. Therefore, with the third objective of this work, **Obj. 3** in Section 1.6, we addressed the classification of cell fate models into broad classes of models presenting specific features that qualitatively compare with experimental data. To this aim, we considered transcriptome data and clonal statistics from lineage tracing experiments, typically used to assess stem cells and self-renewal strategies empirically. In this context, we focused on homeostatic cell fate models that correspond to cell state networks compatible with the rules derived in Chapter 2. Notably, the analysis did not explicitly consider mechanisms for homeostasis regulation, yet the derived results apply to most renewing tissues, including the mammary gland, which is the study case.

First, we proposed sequencing, analysing, and comparing two cell samples, one based on cells from the whole tissue and another based on lineage traced cells, i.e. the progeny of an initially labelled subpopulation of stem cells. More specifically, from the transcriptome data analysis, we aim to cluster cells based on their identity, providing measures of cluster size in the two samples. Based on the mathematical analysis of these scenarios, differences in cluster size of lineage tracing data from tissue data

could only be explained by the existence of at least one self-renewing type in addition to the one labelled for the lineage tracing. In the modelling, this cell type corresponds to a self-renewing SCC of the cell state network, which by definition of the homeostatic network, must be disconnected from the labelled one and must stay at an apex of the lineage hierarchy. Instead, nothing can be said if the clusters size measures in the two samples are the same or not distinguishable given the noise.

Concerning the clone lineage tracing, the presented analysis built on the well-known property for which the Invariant Asymmetry cell fate model, which features only stem cells asymmetric divisions, results in a peaked clonal size distribution, whilst the Population Asymmetry model, which allows for symmetric divisions, results in an Exponential clone size distribution. We, therefore, generalised this concept by analysing the clonal statistics in arbitrary complex cell fate models via analytical modelling and numerical simulations of the stochastic process, including the modelling of the extinction probability. Based on a coarse-grain compartment model formed by the self-renewing and the committed compartments, we classified models of cell fate dynamics in two universal classes as a natural generalisation of the classical Invariant Asymmetry and Population Asymmetry models.

Crucially, we showed that each class of models converges to an identical rescaled clone size distribution. More specifically, the asymptotic regime in which the distributions of the Generalised Population Asymmetry models converge to an Exponential distribution relates to long times. Instead, in the Generalised Invariant Asymmetry models, when all the rates in the self-renewing compartment are much larger than the inverse lifetime of the committed cells, the clonal statistics converge to a Normal distribution. Therefore, if such asymptotic conditions are met, the self-renewing strategy can be determined by looking just at the shape of the clonal size distribution. However, at the same time, the inner details of each compartment cannot be distinguished since cell fate models within the same class predict the same clone size distribution. Crucially, when the asymptotic regime is not fulfilled, as might be in real tissues, helpful information about the self-renewing strategy and model details might be inferred from the evolution of the clones' mean cell number and mid-term clonal data.

6.1.4 Objective 4

To conclude this work, we applied the developed methodology to a specific study case, the healthy adult mouse mammary gland. We reviewed the literature and analysed published data for this biological scenario, for which initially single-cell RNA-sequencing and clone lineage tracing data were expected. We tested the proposed approach by using synthetic data that emulate such experiments. In this way, we defined a validated framework that can be used in future studies to answer

biological questions. Crucially, the developed methodology can be applied not only to the mammary gland but to any other homeostatic renewing tissue where lineage tracing data are available. This part of the work was carried out to fulfil **Obj. 4** detailed in Section 1.6.

To this aim, we first addressed the definition of a cell state network for the study case. For doing so, we combined the rules for homeostasis derived in Chapter 2, which apply in general, and hypothetical qualitative features of the experimental data, following the results presented in Chapter 4. In this way, we derived clear criteria for identifying the cell state network once actual data becomes available. In this frame, an in-depth analysis of published scRNA-seq data was helpful to gain insight into the study case scenario. It also highlighted discrepancies among these works that should be addressed by further dedicated experimental work.

Lastly, we generated the synthetic data and fit the model parameters of the above-derived cell state network by applying a standard Bayesian inference approach. In doing so, we demonstrated that non-homeostatic cell fate models might result in good fittings. This result justifies the need to impose a priori the conditions for homeostasis derived in Chapter 2 and not expect them as an outcome of the fitting. Furthermore, we showed that when the clonal statistic is exponentially shaped, as in the test case analysed, mid-term clonal data alone is insufficient to determine the self-renewing strategy. A long-term clonal data point and the estimation of the timescale of the dynamics were revealed to be suitable additional measures for significantly restricting the variability of the fitting parameters and, importantly, defining the self-renewing strategy.

6.2 Limitations and future work

Besides the achievements summarised in the previous section, some limitations and possible future work is described below.

- This work focuses on homeostasis in adult renewing tissues. An extension of the presented modelling could include some specific non-homeostatic scenarios such as those mentioned in [Greulich et al., 2021], where developing cell types (transient types located upstream of the self-renewing ones) might be present or quiescent stem cell types might change the dynamical behaviour of a homeostatic tissue. Concerning the study case, the mammary gland, these scenarios would be helpful to model cell fate dynamics in different evolution stages, such as puberty and pregnancy. Hyperproliferating cell types could be included as well to assess diseases such as cancer.

- In Chapter 3, we assessed the crowding feedback, a possible homeostasis regulation mechanism. In the proposed model, we assumed that, for each cell type, the kinetic parameters depend on the number of cells of that type. However, another possibility is that these parameters depend on the total cell density (or a specific set of cell types). In that case, the problem must be studied as a whole since dynamics in each SCC are affected by those in the others.

- Concerning the robustness analysis of the homeostasis regulation in the single-cell mutation scenario, we provided a qualitative assessment, in Section 3.3.1 and presented an illustrative case, based on a few stochastic trajectories, in Appendix A.3. However, a detailed assessment aimed at estimating the probability of the extinction of the mutated clone requires an extensive simulation campaign, which, in any case, would remain applicable only to the specific cell fate model assessed.

- The cell fate dynamics, no matter how complex they are, assume a Markovian model for the cell states proliferation, transition and death. However, if short-term clonal measurements were available, this model would not be applicable for the estimation of the clonal statistics, given that the average timescale of the events would be smaller than or of the same order of magnitude as that of the measurements. Therefore, effects like the waiting time between two consecutive rounds of cell division should be included.

- The fitting analysis presented in Section 5.4.2, despite being sufficient to demonstrate the validity of the developed methodology, was not meant to be an exhaustive assessment for which more runs would be required and more solutions analysed. Given that, further work could be done to improve the efficiency of the implementation (e.g. switch to a different simulation environment, exploit a High-Performance Computing System) and consequently the accuracy of the results. Also, a natural extension of this analysis would involve additional sets of synthetic data, for instance, based on cell fate models presenting only asymmetric divisions (i.e. within the GIA model class). The analysis of such additional datasets might confirm the need for long-term clone statistics and measures of tissue turnover time. They might suggest the inclusion of other types of measures as well.

- By having actual experimental data for the study case, we could answer biological questions and help in understanding the cell fate dynamics in the healthy adult mouse mammary gland. More specifically, this includes a) supporting the definition of the cell hierarchy in homeostasis, b) determining the self-renewing strategy for the basal stem cell type, and c) confirming the existence of other stem cell types.

Appendix A

Non-linear cell fate dynamics

A.1 Homeostasis in non-linear dynamic models

In this section we show how Conditions (nl.i)-(nl.iv), discussed in Section 2.2.3, are necessary conditions for having a homeostatic steady-state in a non-linear system. We remark that the below stated Theorem A.1 and proof are part of the Supplemental Material of [Greulich et al., 2021][1].

Theorem A.1. *Consider a dynamical system of the form* $\frac{d}{dt}\bar{\boldsymbol{n}} = A(\bar{\boldsymbol{n}})\bar{\boldsymbol{n}}$*, where* $A(\bar{\boldsymbol{n}})$ *is a Metzler matrix (i.e. with non-negative off-diagonal elements) for all non-negative vectors* $\bar{\boldsymbol{n}} \geq 0$*. If the system has a non-trivial, non-negative steady state,* $\bar{\boldsymbol{n}}^*$*, then the graph* $G(A(\bar{\boldsymbol{n}}^*))$ *satisfies the following three conditions.*

1. *There are no non-trivial super-critical SCC.*
2. *There is at least one non-trivial critical SCC.*
3. *There are no directed paths from a non-trivial critical SCC to another. (Equivalently, there are no other non-trivial critical SCC upstream of any critical SCC.)*

Moreover, any SCC above a critical SCC must be trivial, and every SCC is either trivial or positive.

We remark that Conditions 1, 2 and 3 in Theorem A.1 correspond respectively to Conditions (nl.i), (nl.ii) and (nl.iv) in Section 2.2.3, while Condition (nl.iii) in Section 2.2.3 is part of the last statement of the Theorem A.1.

Let us assume the existence of a non-trivial, non-negative steady state $\bar{\boldsymbol{n}}^*$, that is, $\mathbf{0} \neq \bar{\boldsymbol{n}}^* \geq 0$ and $A\bar{\boldsymbol{n}}^* = \mathbf{0}$, where $A = A(\bar{\boldsymbol{n}}^*)$. We assume a topological ordering of the

[1] Notation is revised here to be consistent with the rest of the thesis.

nodes (see Section 2.2.1) so that the matrix A is a lower triangular matrix and we can write

$$A\boldsymbol{x} = \begin{pmatrix} A_1 & 0 & 0 & 0 & ... \\ C_{21} & A_2 & 0 & 0 & ... \\ C_{31} & C_{32} & A_3 & 0 & ... \\ \vdots & \vdots & \vdots & \ddots & 0 \\ ... & ... & ... & ... & A_h \end{pmatrix} \begin{pmatrix} \bar{\boldsymbol{n}}_1^* \\ \bar{\boldsymbol{n}}_2^* \\ \bar{\boldsymbol{n}}_3^* \\ \vdots \\ \bar{\boldsymbol{n}}_h^* \end{pmatrix} = \mathbf{0}, \tag{A.1}$$

where h is the number of SCC of the corresponding cell state network (see definition in Section 2.1.2), A_k is the block of A associated to the kth SCC, S_k ($1 \leq k \leq h$), C_{kl} encodes the connectivity from S_l to S_k, and we have decomposed the steady state vector as $\bar{\boldsymbol{n}}^* = (\bar{\boldsymbol{n}}_1^*, \bar{\boldsymbol{n}}_2^*, ..., \bar{\boldsymbol{n}}_h^*)^T$. In particular, the kth row gives

$$\sum_{l<k} C_{kl}\bar{\boldsymbol{n}}_l^* + A_k\bar{\boldsymbol{n}}_k^* = \mathbf{0}. \tag{A.2}$$

For convenience, we write $\boldsymbol{y}_k = \sum_{l<k} C_{kl}\bar{\boldsymbol{n}}_l^*$ and rewrite Equation (A.2) as

$$\boldsymbol{y}_k + A_k\bar{\boldsymbol{n}}_k^* = \mathbf{0}. \tag{A.3}$$

As well as for the linear system, each A_k is an irreducible Metzler matrix, so the Perron-Frobenius theorem applies [Arrow, 1989], and thus A_k has a simple real eigenvalue μ_k of maximal real part among all eigenvalues of A_k. Moreover, μ_k has a positive (left) eigenvector $\boldsymbol{v}_k$, that is, $\boldsymbol{v}_k A_k = \mu_k \boldsymbol{v}_k$ and $\boldsymbol{v}_k > 0$. If we multiply Equation (A.3) by the left eigenvector $\boldsymbol{v}_k$ we obtain

$$0 = \boldsymbol{v}_k \cdot \boldsymbol{y}_k + \boldsymbol{v}_k A_k \bar{\boldsymbol{n}}_k^* \implies \boldsymbol{v}_k \cdot \boldsymbol{y}_k = -\mu_k \boldsymbol{v}_k \cdot \bar{\boldsymbol{n}}_k^*. \tag{A.4}$$

In the following, we make repeated use of,

Lemma A.2. *Let $\boldsymbol{z}, \boldsymbol{x} \in \mathbb{R}^n$. If $\boldsymbol{z} > 0$ and $\boldsymbol{x} \geq 0$, then holds: $\boldsymbol{z} \cdot \boldsymbol{x} = 0 \Leftrightarrow \boldsymbol{x} = \mathbf{0}$.*

This holds since $0 = \boldsymbol{z} \cdot \boldsymbol{x} = \sum_i z_i x_i$ requires that all $z_i x_i = 0$, given that all $z_i, x_i \geq 0$. Since $z_i > 0$, this can only be if $x_i = 0$ for all $i = 1, 2, ...$ The direction $\boldsymbol{x} = 0 \Rightarrow \boldsymbol{z} \cdot \boldsymbol{x} = 0$ of the lemma is trivial.

Now we distinguish the cases $\mu_k > 0$, $\mu_k = 0$, and $\mu_k < 0$.

Case $\mu_k > 0$. In this case (a super-critical SCC), $\mu_k > 0$, and $\boldsymbol{v}_k > 0$, $\bar{\boldsymbol{n}}_k^* \geq 0$, so the right hand side of Equation (A.4) is a non-positive number. Since $\boldsymbol{y}_k \geq 0$ and $\boldsymbol{v}_k > 0$, the left-hand side is a non-negative number. Hence both the left- and right-hand-side must be zero. Therefore, according to Lemma A.2, $\bar{\boldsymbol{n}}_k^*$ is zero, that is, the SCC is trivial. Thus, *all super-critical SCC – if there are any – are trivial*, which is **Condition 1**. Moreover, we get from Lemma A.2 that $\boldsymbol{y}_k = \mathbf{0}$ for super-critical SCC, which we will use later.

Case $\mu_k < 0$. In this case (a sub-critical SCC), the matrix A_k is invertible (since all eigenvalues are negative) and we can rearrange Equation (A.3) to

$$\bar{\boldsymbol{n}}_k^* = -A_k^{-1}\boldsymbol{y}_k. \tag{A.5}$$

The matrix $-A_k$ is a non-singular M-matrix (non-positive off-diagonal entries and all eigenvalues with positive real part), hence its inverse is a positive matrix [Meyer and Stadelmaier, 1978]. In particular, $\bar{\boldsymbol{n}}_k^*$ is a positive vector unless $\boldsymbol{y}_k = \boldsymbol{0}$, in which case $\bar{\boldsymbol{n}}_k^* = \boldsymbol{0}$. All in all, *a sub-critical SCC is either trivial or positive*.

Now we can show **Condition 2** by contradiction. Suppose that there are no critical SCC. If A_1 is super-critical, it must be trivial. If A_1 is sub-critical, $\det(A_1) \neq 0$ and $A_1\bar{\boldsymbol{n}}_1^* = \boldsymbol{0}$ (from Equation (A.1)) imply $\bar{\boldsymbol{n}}_1^* = \boldsymbol{0}$. Suppose now that all SCC S_l for $l < k$ are trivial. Then either A_k is super-critical and hence trivial, or A_k is sub-critical and hence also trivial, by Equation (A.5),

$$\bar{\boldsymbol{n}}_k^* = -A_k^{-1}\left(\sum_{l<k} C_{kl}\bar{\boldsymbol{n}}_l^*\right) . \tag{A.6}$$

Hence, if there are no critical SCC, then all $\bar{\boldsymbol{n}}_k^* = \boldsymbol{0}$, and thus $\bar{\boldsymbol{n}}^*$ is vanishing, which is in contradiction to the existence of a non-vanishing steady state $\bar{\boldsymbol{n}}^* \neq 0$.

Case $\mu_k = 0$. In this case (critical SCC), Equation (A.4) becomes, $\boldsymbol{v}_k\boldsymbol{y}_k = 0$. Since $\boldsymbol{y}_k \geq 0$ and $\boldsymbol{v}_k > 0$, Lemma A.2 requires that,

$$\boldsymbol{y}_k = \sum_{l<k} C_{kl}\bar{\boldsymbol{n}}_l^* = \boldsymbol{0}. \tag{A.7}$$

This reduces Equation (A.3) to $A_k\bar{\boldsymbol{n}}_k^* = \boldsymbol{0}$, so either $\bar{\boldsymbol{n}}_k^*$ is zero or it is a non-negative 0-eigenvector of A_k, that is, a positive (Perron-Frobenius) eigenvector of A_k. We conclude that *every critical SCC is either trivial, or positive*.

All in all, we have shown that *every SCC must be either trivial, or positive*. If that is the case, $C_{kl}\bar{\boldsymbol{n}}_l^* = \boldsymbol{0}$ implies $C_{kl} = \boldsymbol{0}$ or $\bar{\boldsymbol{n}}_l^* = \boldsymbol{0}$ (if $\bar{\boldsymbol{n}}_l^*$ is not zero, it must be positive, and the product of a non-negative matrix and a positive vector is zero if and only if the matrix is zero). In particular, $\boldsymbol{y}_k = \boldsymbol{0}$ if and only if $C_{kl} = \boldsymbol{0}$ or $\bar{\boldsymbol{n}}_l^* = \boldsymbol{0}$ for all $l < k$. In words, $\boldsymbol{y}_k$ *is zero if and only if every SCC immediately upstream of* S_k *is trivial*. (We call a SCC S_l immediately upstream of S_k if there is at least one link from a node in S_l to a node in S_k, that is, if $C_{kl} \neq \boldsymbol{0}$, the zero matrix.)

We now revisit the three cases above. If S_k is super-critical, we showed that $\bar{\boldsymbol{n}}_k^* = \boldsymbol{0}$ and that $\boldsymbol{y}_k = \boldsymbol{0}$, that is, *all super-critical SCC must be trivial, and all SCC immediately upstream of a super-critical SCC must also be trivial*. If S_k is sub-critical, we showed that $\bar{\boldsymbol{n}}_k^* = -A_k^{-1}\boldsymbol{y}_k$, with $-A_k^{-1}$ a positive matrix, so that $\bar{\boldsymbol{n}}_k^* = \boldsymbol{0}$ if and only if $\boldsymbol{y}_k = \boldsymbol{0}$. That is, *a sub-critical SCC is trivial if and only if every SCC immediately upstream is trivial,*

otherwise it is positive. Finally, if S_k is critical, we showed that $\boldsymbol{y}_k = \mathbf{0}$, that is, *all SCC immediately above a critical SCC must be trivial*. Since, in addition, any SCC immediately upstream of a super-critical, or a trivial sub-critical, SCC are trivial, we have that *any SCC upstream of a critical SCC must be trivial*. This, in particular, proves **Condition 3** and the last part of the theorem, and we are done.

A.2 Dynamic long-term self-renewing state: test case definition

This section describes the methodology used to define the illustrative examples of the Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics presented in Section 3.2.1.

Based on Section 3.1, we recall that the cell state model of a self-renewing cell type regulated via crowding feedback is described by the non-linear system of equations (corresponding to Equation (3.1) in Section 3.1)

$$\frac{d}{dt}\boldsymbol{\rho}(t) = A(\rho(t))\,\boldsymbol{\rho}(t)\,, \tag{A.8}$$

in which $\boldsymbol{\rho}$ is the vector of cell densities, defined as the average number of cells per unit of volume, V, that is, $\boldsymbol{\rho} = \bar{\boldsymbol{n}}/V$, and $\rho = \sum_i \rho_i$. A non-trivial steady-state, $\boldsymbol{\rho}^*$, corresponds to a condition in which $\mu(\rho^*) = 0$, where $\mu(\rho)$ is the dominant eigenvalue of the matrix $A(\rho)$. For evaluating the linear stability of this steady-state, we can look at the sign of the real part of the eigenvalues of the Jacobian matrix J, which is (corresponding to Equation (3.5) in Section 3.2.1)

$$J = A(\rho^*) + A'\boldsymbol{\rho}^*\,\mathbf{1}^T\,, \tag{A.9}$$

in which A' is a matrix whose elements are the derivative of the elements of A with respect to ρ, evaluated at the fixed point. If the maximum real part of the eigenvalues of J, μ_J, is negative, then the system is asymptotically stable; if μ_J is positive, the steady-state is, at least locally, unstable. Furthermore, in Section 3.2.1, we showed that a dynamic long-term self-renewing state, that is, a non-constant cell dynamics yet confined, can be achieved if the condition (corresponding to Equation (3.2) in Section 3.2.1)

$$\frac{\partial \mu}{d\rho} < 0 \text{ for all } \rho \geq 0\,, \tag{A.10}$$

is satisfied.

Given that, the idea is to find cell fate models in which, depending on the variation of the kinetic parameters with ρ, the same steady state, $\boldsymbol{\rho}^*$ is either Asymptotically Stable

(**AS**), i.e. $\mu_J < 0$, Locally Unstable (**LU**), i.e. $\mu_J > 0$ and Equation (A.10) is satisfied, or Unstable (**U**), i.e $\mu' > 0$. Finding the **LU** case is not straightforward since typically μ_J and μ' have the same sign. Thus to solve this problem, we followed the steps described below.

1. We arbitrarily choose a single-type cell state network, which is shown in Figure 3.1. We indicate as $\boldsymbol{\alpha}$ the kinetic parameters of this cell fate model and $\boldsymbol{\alpha}'$ their derivatives with respect to the total cell density ρ.

2. We focus on the steady-state condition only, and we set up an optimisation problem to find a set of values $\boldsymbol{\alpha}^*$ for which there exists $\boldsymbol{\alpha}'^*$ such that $\mu_J > 0$ and $\mu' < 0$ (i.e. necessary condition for the **LU** test case). The implementation of the objective function is detailed in Algorithm 2. For solving this optimisation problem we use the Particle Swarm Optimisation algorithm, which is a global stochastic optimisation method [Poli et al., 2007]. The values $\boldsymbol{\alpha}^*$ reported in Table A.1 results from an optimisation run in the following search space: $\boldsymbol{\alpha} = [0.2; 4.8]$, $|\boldsymbol{\alpha}'| = [0; 10]$ (we omit here the units since they are arbitrary).

3. The solution found in the previous step, $\boldsymbol{\alpha}^*$, assures that for this model there are conditions in $\boldsymbol{\alpha}'^*$ for which a dynamic long-term self-renewing state might exist (the condition must be fulfilled for any ρ and not only at the steady-state, but this will be checked later). We therefore evaluate a large number, $NN = 10^4$, of random $\boldsymbol{\alpha}'^* = [-10; 10]$ in terms of μ_J and μ'. We note that the sign of the components of $\boldsymbol{\alpha}'^*$ is random, giving both positive and negative values of μ'. From all these cases, shown in Figure A.1, we manually choose the three cases, corresponding to the test cases **AS**, **LU** and **U**, each one associated to a different quadrant (we recall that the quadrant $\mu' > 0$ and $\mu_J < 0$ is not reachable). The corresponding values of $\boldsymbol{\alpha}'^*$ are reported in Table A.1.

4. For each test case and for each kinetic parameter α_i, for $i = 1, ..., 8$, we derived the Hill function parameters corresponding to α_i^* and $\alpha_i'^*$. Hill function is defined as $\alpha(\rho) = c + k\rho^n/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') > 0$ and $\alpha(\rho) = c + k/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') < 0$. Considering there are four unknown parameters and only two values (function and derivative), we solve the problem in k and K for a fixed $c = 0.05$ and choosing the solution with minimum n. The final values of the Hill parameter functions are reported in Section 3.2.1 (see Table 3.1); the profiles of $\alpha(\rho)$ and $\alpha'(\rho)$ are shown in Figure A.2. Finally, for the **LU** case, we checked that $\partial\mu/\partial\rho < 0$ for any value of ρ, that is, the sufficient condition (A.10) for having a dynamic long-term self-renewing state is satisfied.

Algorithm 2 Dynamic long-term self-renewing test case: objective function (see Step 2)

Input parameters: $\boldsymbol{\alpha}_0$, $|\boldsymbol{\alpha}'_0|$;
Variable initialisation: $\Delta\alpha = 0.2$, $Tol = 10^{-8}$;
if $|\mu(\boldsymbol{\alpha}_0)| \geq Tol$ **then**
  Find a homeostatic state, $\boldsymbol{\alpha}^* = \min_{\boldsymbol{\alpha} \in \boldsymbol{\alpha}_0 \pm \Delta\alpha} |\mu(\boldsymbol{\alpha})|$; // Local search based on Matlab fmincon function
else
  $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_0$;
end if
if $|\mu(\boldsymbol{\alpha})| \leq Tol$ **then**
  Determine the sign and the value of $\boldsymbol{\alpha}'$ to assure that $\mu' = \sum_\alpha \partial\mu/\partial\alpha_i \alpha'_i < 0$, i.e. $\alpha'_i = -\text{sign}(\partial\mu/\partial\alpha_i)(|\boldsymbol{\alpha}'_0|)_i$
  Compute the eigenvalue of the Jacobian matrix, $\mu_J(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}')$ // Equation (A.9)
  Assign the value of the objective function, $y = -\mu_J$;
else
  Set the objective function for a non-homeostatic case, $y = \infty$;
end if
return y, μ', μ_J;

α	α^*	α'^*		
		AS	**LU**	**U**
λ_1	0.3940	0.2555	9.1677	-0.7655
λ_3	2.2974	-0.1297	-0.6078	-0.3123
γ_1	2.4403	1.0882	1.6475	-2.6487
γ_2	1.1330	0.7626	0.0142	-2.0475
ω_{13}	4.5110	-2.2722	-4.9956	1.5744
ω_{21}	0.3346	-2.6854	-3.9305	0.2270
ω_{23}	0.0852	0.1520	-3.6489	-0.8496
ω_{31}	5.4836	-1.4154	2.5450	-0.4794

TABLE A.1: Kinetic parameters at the steady-state in the test cases illustrative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics (see details of Step 3). Unit for the kinetic parameters is arbitrary and therefore omitted.

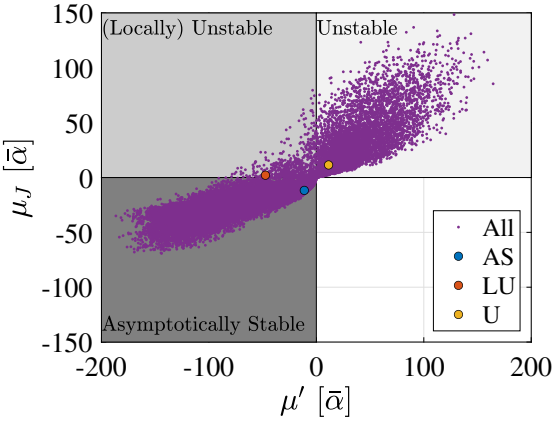


FIGURE A.1: Stability parameters, μ_J and $\mu' = \partial\mu/\partial\rho|_{\rho^*}$, for the kinetic parameters $\boldsymbol{\alpha}^*$ given in Table A.1, and random values $\boldsymbol{\alpha}'^*$ (see details of Step 3). Among these points, we manually choose the three test cases, **AS**, **LU** and **U**, each one associated to a different quadrant.

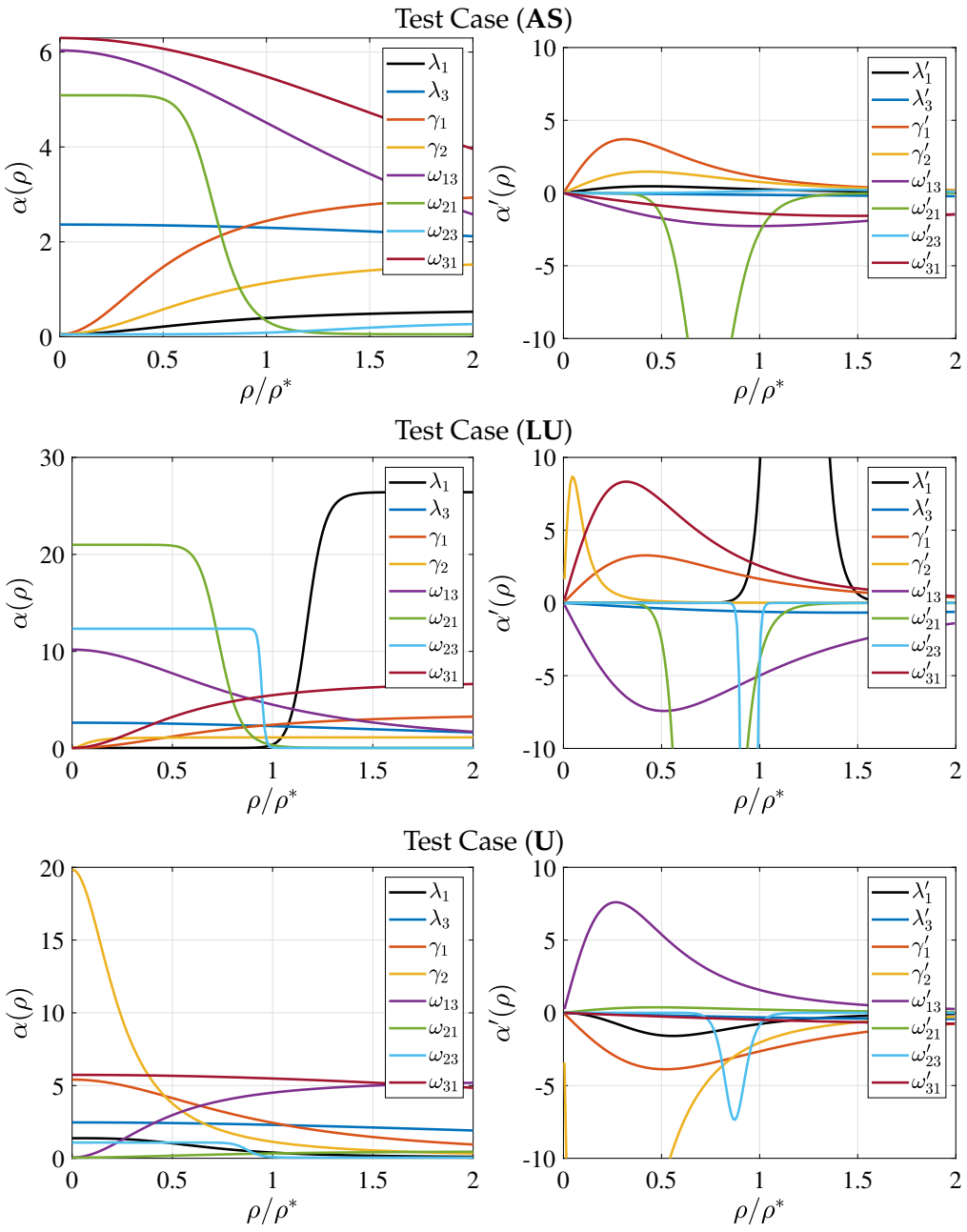


FIGURE A.2: Kinetic parameters (left panels), and their derivative with respect to ρ (right panels) for the test cases representative of an Asymptotically Stable (**AS**), Locally Unstable (**LU**) and Unstable (**U**) dynamics. These parameters, shown as functions of cell density normalised by the steady-state ρ^*, correspond to Hill functions defined as $\alpha(\rho) = c + k\rho^n/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') > 0$ and $\alpha(\rho) = c + k/(K^n + \rho^n)$ when $s = \text{sign}(\alpha') < 0$ (see details in Step 4). Values of the parameter of the Hill function are reported in Table 3.1.

A.3 Single cell mutation test case

In Section 3.3.1, we showed how the crowding feedback is a robust mechanism for regulating homeostasis. Through a numerical example, we showed that dysregulation in the feedback mechanism for some of the kinetic parameters might be compensated by the others, assuring the homeostatic state is maintained (test case $\mathbf{F}_1$). However, when they cannot be compensated entirely, the system deviates from the homeostatic state ($\mathbf{F}_2$). These results apply to a dysregulation acting at the tissue level and thus involving all the cells in the tissue. Here, instead, we assess a single cell dysregulation as representative of cell mutations. Importantly, we will show how mutations in a single cell can affect the whole tissue.

For modelling this scenario, we assume the same cell state network (see Figure 3.1), feedback functions and parameters (see Table 3.1, **AS** test case) and dysregulation (see Table 3.2) as in Section 3.3.1. However, considering that here we need to model the clonal dynamics (dynamics of single cells and their progeny), we will make use of stochastic simulations based on the Gillespie Algorithm, detailed in Appendix B.1.1. Based on this algorithm, we simulate 10^3 uncorrelated trajectories, where each trajectory is a possible realisation of the stochastic process. We chose a total cell number $N_0 = 5000$ as the initial condition (cell density is based on unitary volume). In a real tissue, the number of cells could be a few orders of magnitude larger, but this number is sufficiently large to avoid the extinction of the process in the time scale analysed, so, once rescaled, these dynamics are representative of those in the tissue. Furthermore, our numerical simulator is not suitable to model larger values of N_0, for which a different implementation would be required (see details in Appendix B.1.1). Finally, simulations are stopped when the mutated clone goes extinct, or divergence of the dynamics is detected.

From an implementation point of view, to model the tissue dynamics, including the mutated cell, we consider a cell fate model composed of two disconnected cell state networks: one corresponding to the unperturbed test case **AS**, and the other to the dysregulated one, i.e. $\mathbf{F}_1$ or $\mathbf{F}_2$. The simulation starts with N_0 cells in the **AS** network, distributed in each state proportionally to the expected steady-state distribution in the tissue, and no cells in $\mathbf{F}_x$ network, with $x = 1$ or 2. Thus, since the two networks are disconnected, $\mathbf{F}_x$ remains empty, and the simulation represents the tissue dynamics before the dysregulation. At a time equal to zero, we continued the simulation, moving initially one cell in a random state from the **AS** network to the corresponding state in the $\mathbf{F}_x$ one. This simulation represents the tissue dynamics, including the single mutated cell.

In Figure A.3, we show four illustrative trajectories for test case $\mathbf{F}_1$ (left) and test case $\mathbf{F}_2$ (right). In these cases, the mutated clones go extinct (with different time scales), and tissue dynamics are not affected globally. As expected, divergence is never detected in

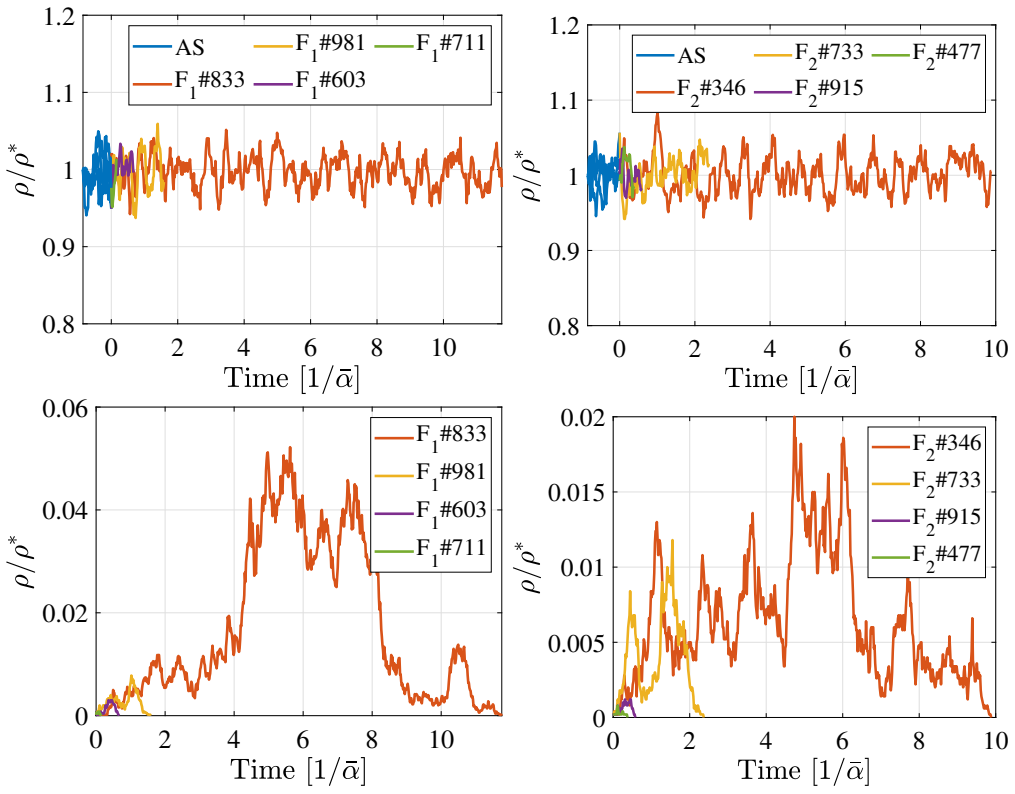


FIGURE A.3: Stochastic dynamics in test case **AS**, and including a single-cell mutated clone based on test case $\mathbf{F}_1$ (left panels) and on test case $\mathbf{F}_2$ (right panels). The total cell density (upper panels) and that of the mutated clone (bottom panels), ρ, is normalised by the homeostatic value, ρ^*, and it is shown as a function of the time. Four illustrative cases are shown; each curve represents a possible realisation of the stochastic process. In all these cases, the mutated clone goes extinct, and the tissue dynamics are globally unaffected. Dynamics are scaled by $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

test case $\mathbf{F}_1$, and all the trajectories present the same behaviour as those shown in Figure A.3. Instead, for the test case $\mathbf{F}_2$, the mutated clone goes extinct in all the trajectories except one. In particular, trajectory #153 results in a growing dynamics as shown in Figure A.4. Here, the mutated clone (black line) grows and eventually prevails, leading all the tissue to diverge.

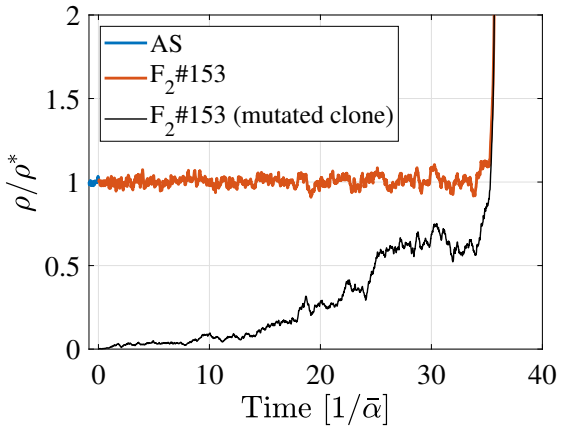


FIGURE A.4: Stochastic dynamics in test case **AS**, and including a single-cell mutated clone based on test case $\mathbf{F}_2$ dysregulation. The total cell density and that of the mutated clone (black line), ρ, is normalised by the homeostatic value, ρ^*, and it is shown as a function of the time. In this case, corresponding to trajectory #153, the mutated clone prevails, and the whole tissue dynamics become unstable. Dynamics are scaled by $\bar{\alpha} = \min_{i,j}\{\lambda_i(\rho^*), \omega_{ij}(\rho^*), \gamma_i(\rho^*)\}$.

Appendix B

Stochastic dynamics modelling

B.1 Implementation of the stochastic simulation

B.1.1 Gillespie algorithm

Throughout this work, all the numerical simulations of the stochastic process associated with a cell fate model are based on the Gillespie algorithm [Gillespie, 1977], which is described in this section. For an m-state process of the type (2.1)-(2.3), we consider a $\tilde{m} = m + 1$ state vector, $\tilde{n} = (n_1, n_2, ..., n_m, n_\emptyset)^T$, in which n_i is the number of cells in the ith state at a given time, and $n_\emptyset$ is the number of dead cells. For each i-cell state, $i = 1, ..., m$, we define a vector $\boldsymbol{\alpha}_i = \left(\omega_{ij}, \quad \lambda_i r_i^{kl}\right)$, for $j, k, l = 1, ..., \tilde{m}$, with $j \neq i$. Provided this, for a generic initial condition $\tilde{n}_0$, the simulation algorithm is detailed in Algorithm 3. The complete process is a trajectory in $\tilde{n}$ as function of the time. In case the initial condition is a single cell, then the trajectory represents a clone. Under the same simulation, N uncorrelated trajectories are run. For this reason, uncorrelated sequences of random numbers u_t and u_x are generated in each trajectory (Step 11 and Step 14 in Algorithm 3).

We remark that the implemented algorithm is suitable for simulating the clonal dynamics for which the number of cells remains relatively low. For large cell numbers, instead, the time increases very slowly (Step 11 in Algorithm 3) since many events can occur in a short time: this has a substantial impact on the efficiency of the algorithm, and alternative approaches (e.g. tau-leap approximation) are usually preferred [Baker, 2017].

B.1.2 Test cases

To verify the implementation of the Gillespie algorithm above described, we set up two test cases based on cell fate models for which analytical results are known. In

Algorithm 3 Gillespie Algorithm

Parameter definition: $\tilde{m}$, $\boldsymbol{\alpha}_i$ (for $i = 1, ..., m$), $\tilde{n}_0$, t_f;
Variable initialisation: $t = 0$; $\tilde{n} = \tilde{n}_0$;
while $t \leq t_f$ **do**
  if $\boldsymbol{\alpha}_i$ is constant **then**
    $\alpha_i = \sum_j (\boldsymbol{\alpha}_i)_j$; // Constant rate model
  else
    $N = \sum_{i=1}^{m} \tilde{n}_i$; // Crowding feedback model (single SCC)
    $\alpha_i = \sum_j (\boldsymbol{\alpha}_i(N))_j$;
  end if
  Evaluation of the total rate as $R = \sum_{i=1}^{m} \alpha_i \tilde{n}_i$;
  Simulation time update, $t = t + R^{-1} \ln\left(\frac{1}{u_t}\right)$, in which u_t is a uniform random number between 0 and 1;
  if $t \leq t_f$ **then**
    Evaluation of the event probability vector, $\boldsymbol{s} = \frac{\begin{pmatrix} \boldsymbol{\alpha}_1 \tilde{n}_1 & \dots & \boldsymbol{\alpha}_m \tilde{n}_m \end{pmatrix}}{R}$; // $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i(N)$ in case of crowding feedback modelling
    An interval between 0 and 1 is split into contiguous sub-intervals each one related to an event of type k with probability s_k. The type of event is selected based on where a uniform random number between 0 and 1, u_n, falls.
    State update consistently to the event selected: if the event is of transition type, with rate ω_{ij}, then $\tilde{n}_i = \tilde{n}_i - 1$ and $\tilde{n}_j = \tilde{n}_j + 1$; if it is of division type, with rate λ_i and r_i^{jk}, then $\tilde{n}_i = \tilde{n}_i - 1$, $\tilde{n}_j = \tilde{n}_j + 1$ and $\tilde{n}_k = \tilde{n}_k + 1$;
  end if
end while

particular, these test cases correspond to the simplest version of the Invariant Asymmetry (IA) and the Population Asymmetry (PA) models (see Section 1.2), that are

$$S \xrightarrow{\lambda} \begin{cases} S + S & \text{Pr. } r \\ S + D & \text{Pr. } 1 - 2r \\ D + D & \text{Pr. } r \end{cases}, D \xrightarrow{\gamma} \emptyset . \tag{B.1}$$

Here, cells of type S represent the stem cells, which divide with the rate λ, and cells of type D are the differentiated cells, which are shed with the rate γ. While in the PA model, the three possible outcomes of the division of a progenitor are controlled by a probability parameter $0 < r \leq 1/2$, in the IA model $r = 0$, meaning that there are strictly asymmetric divisions and the number of S-cells is conserved.

Considering the above model, numerical simulations for the clonal dynamics were run based on the parameters reported in Table B.1. It is noted that the time unit is arbitrary and therefore omitted. Simulations are based on 10^4 and 5×10^4 runs respectively for the **IA** and **PA** test cases. The initial condition is a single stem cell and the final simulation time, indicated as τ, is equal to 10: this value is well representative of a steady state condition (for the **IA** test cases) and at which the total extinction of

the process is not yet achieved (for **PA** test cases only). The resulting clone size distribution at τ is shown in Figure B.1 respectively for test cases **IA** (left) and **PA** (right); in these figures, the numerical simulation results are compared to the expected clone size distribution. In particular, as shown in Section 1.2, for test cases **IA**, we expect a Poisson distribution, $P(n) = \text{Poisson}(\lambda/\gamma)$, shifted by one (i.e. plus the stem cell). For the test cases **PA**, instead, we present the numerical integration of the master equation (1.5), and, for test case PA#1 ($\lambda = \gamma$ and $r = 1/4$), the reference analytic solution provided in [Antal and Krapivsky, 2010].

TABLE B.1: IA and PA test cases simulation parameters

Case	λ	γ	r
IA#1	1.0	1.0	0
IA#2	2.0	1.0	0
IA#3	5.0	1.0	0
PA#1	1.0	1.0	1/4
PA#2	2.0	1.0	1/4
PA#3	2.0	1.0	1/6

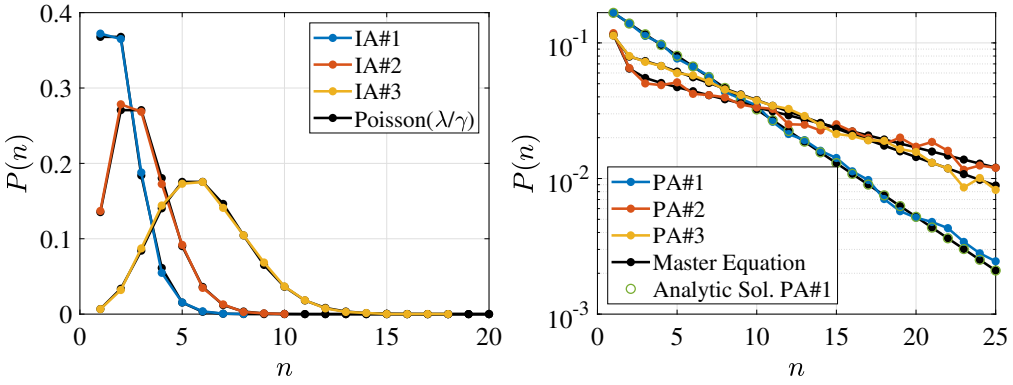


FIGURE B.1: Comparison of the numerical simulation and the reference results for test cases **IA** (left) and **PA** (right). The shown clone size distribution, $P(n)$, is the distribution of the total number of cells n forming the progeny of a single initial cell of type S based on the model (B.1) and parameters reported in Table B.1. For each case, the distribution is shown at the final time, τ, which is well representative of the steady state condition and at which, in test cases **PA**, the total extinction of the process is not yet achieved. The distribution for test case IA#1-3 are compared to the expected Poisson distribution. Test cases PA#1-3 are compared to the solution of the numerical integration of the master equation (1.5) and, for test case PA#1, also to the reference analytic solution from [Antal and Krapivsky, 2010].

B.2 Steady state distribution in GIA Markovian model: limiting behaviour

As shown in Section 4.3.3.1, the steady-state distribution of the X_2-cells in the GIA Markovian model (4.11) is (corresponding to Equation (4.17))

$$P^*(n_2) = \left(1-\hat{\lambda}_2\right)^{\hat{\lambda}_1/\hat{\lambda}_2} \hat{\lambda}_2^{n_2} \frac{\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}+n_2\right)}{\Gamma(n_2+1)\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}\right)}, \tag{B.2}$$

in which $\hat{\lambda}_i = \lambda_i/\gamma$, for $i = 1, 2$. For large mean cell number, the distribution is (corresponding to Equation (4.20))

$$P^*(x_2) = (1-\hat{\lambda}_2)^{\frac{\hat{\lambda}_1}{\hat{\lambda}_2}} \hat{\lambda}_2^{\frac{\hat{\lambda}_1 x_2}{(1-\hat{\lambda}_2)}} \frac{\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}+\frac{\hat{\lambda}_1}{1-\hat{\lambda}_2}x_2\right)}{x_2\Gamma\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}\right)\Gamma\left(\frac{\hat{\lambda}_1}{1-\hat{\lambda}_2}x_2\right)}, \tag{B.3}$$

in which x_2 is the ratio between the number of cells, n_2, and its steady state mean value, $\bar{n}_2^*$.

Now, to better understand the variability of these distributions, we derive here the limiting cases: $\hat{\lambda}_2 \to 0$, $\hat{\lambda}_2 \to 1$ and $\hat{\lambda}_1 \to \infty$. For clarity and readability, we simplify the notation using $p = \hat{\lambda}_1$ and $q = \hat{\lambda}_2$.

(a) $\hat{\lambda}_2 \to 0$

We first analyse the case in which $\hat{\lambda}_2 \to 0$, corresponding, in the simplified notation, to $q \to 0$. In this case, Equation (B.2) can be simplified considering that [Abramowitz and Stegun, 1972]

$$\Gamma(n_2+1) = n_2!\,, \tag{B.4}$$

and that[1]

$$\lim_{q\to 0} \frac{\Gamma\left(\frac{p}{q}+n_2\right)}{\Gamma\left(\frac{p}{q}\right)} \left(\frac{q}{p}\right)^{n_2} = 1\,, \tag{B.5}$$

$$\lim_{q\to 0}(1-q)^{p/q} = e^{-p}\,. \tag{B.6}$$

Thus, the distribution results in

$$\lim_{q\to 0} P^*(n_2) = \frac{p^{n_2}e^{-p}}{n_2!} = \text{Poisson}(p), \tag{B.7}$$

[1] Limits are evaluated using Matlab symbolic toolbox and Mathematica.

that is a Poisson distribution with mean equal to p.

For large mean number of cells, which are obtained for large p (when $q = 0$, then $\bar{n}_2^* = p$), the Poisson distribution tends to a Normal distribution with mean and variance equal to p. Therefore,

$$\lim_{(q,p)\to(0,\infty)} P^*(n_2) = \frac{1}{\sqrt{2\pi p}} e^{-\frac{(n_2-p)^2}{2p}} = \text{Normal}(p,p). \tag{B.8}$$

Rescaling the distribution based on $x_2 = n_2/\bar{n}_2^*$, results in

$$\lim_{(q,p)\to(0,\infty)} P^*(x_2) = \text{Normal}(1, 1/p), \tag{B.9}$$

that is a Normal distribution with unitary mean and variance equal to $1/p$.

(b) $\hat{\lambda}_2 \to 1$

For $\hat{\lambda}_2 \to 1$, that is $q \to 1$, the steady state mean number of cells $\bar{n}_2^* \to \infty$ and Equation (B.3) holds. This equation can be rewritten as

$$P^*(x_2) = q^{p/(1-q)x_2+1} \frac{(1-q)^{p/q}}{q(x_2-1)+1} \frac{\Gamma\left(p\frac{q(x_2-1)+1}{q(1-q)}+1\right)}{\Gamma\left(\frac{p}{q}\right)\Gamma\left(\frac{p}{1-q}x_2+1\right)}. \tag{B.10}$$

If the Stirling's approximation [Abramowitz and Stegun, 1972] is applied

$$\Gamma(z+1) \sim \sqrt{2\pi z}\left(\frac{z}{e}\right)^z, \tag{B.11}$$

we obtain

$$P^*(x_2) = \frac{p^{p/q} e^{-p/q} q^{(q-2p)/(2q)} (q(x_2-1)+1)^{p/(1-q)(x_2-1+1/q)-1/2}}{\Gamma\left(\frac{p}{q}\right) x_2^{x_2 p/(1-q)+1/2}}, \tag{B.12}$$

which is valid for large mean number (in case $q \to 1$, this is true unless $p = 0$).

Considering now that[2]

$$\lim_{q\to 1} \frac{(q(x_2-1)+1)^{p/(1-q)(x_2-1+1/q)-1/2}}{x_2^{x_2 p/(1-q)+1/2}} = e^{p(1-x_2)} x_2^{p-1}, \tag{B.13}$$

it follows that

$$\lim_{q\to 1} P^*(x_2) = \frac{p^p}{\Gamma(p)} x_2^{p-1} e^{-p x_2} = \text{Gamma}(p, 1/p), \tag{B.14}$$

that is a Gamma distribution with unitary mean and shape parameter given by p.

[2] Limit is evaluated using Mathematica.

(c) $\hat{\lambda}_1 \to \infty$

When $\hat{\lambda}_1$, that is p, is large, the mean number of cells is large for any value of q. Thus, Equation (B.12) is valid. By applying the Stirling's approximation also to the term $\Gamma(p/q)$, we obtain

$$P^*(x_2) = \sqrt{\frac{p}{2\pi}} x_2^{-p/(1-q)x_2 - 1/2} (q(x_2-1)+1)^{p/(1-q)(x_2-1+1/q)-1/2}. \tag{B.15}$$

This expression can be also rewritten as

$$P^*(x_2) = \sqrt{\frac{p}{2\pi}} e^K, \tag{B.16}$$

in which

$$\begin{aligned} K = &\frac{p}{1-q}\left(\left(x_2 - 1 + \frac{1}{q}\right)\log(q(x_2-1)+1) - x_2\log(x_2)\right) \\ &- \frac{1}{2}(\log(x_2) + \log(q(x_2-1)+1)). \end{aligned} \tag{B.17}$$

Considering now that p is large, then

$$\begin{aligned} -\frac{1}{2}(\log(x_2) + \log(q(x_2-1)+1)) \ll \\ \frac{p}{1-q}\left(\left(x_2 - 1 + \frac{1}{q}\right)\log(q(x_2-1)+1) - x_2\log(x_2)\right), \end{aligned} \tag{B.18}$$

the term $-1/2(\log(x_2) + \log(q(x_2-1)+1))$ can be neglected. Additionally, for $x_2 \to 1$ the following expansions can be applied

$$\log(q(x_2-1)+1) = \sum_{k=1}^{\infty}\left((-1)^{k+1}\frac{(q(x_2-1))^k}{k}\right), \tag{B.19}$$

and

$$\log(x_2) = \sum_{k=1}^{\infty}\left((-1)^{k+1}\frac{(x_2-1)^k}{k}\right). \tag{B.20}$$

Finally, if we consider that[3]

$$\begin{aligned} \frac{\left(x_2 - 1 + \frac{1}{q}\right)\sum_{k=1}^{\infty}\left((-1)^{k+1}\frac{(q(x_2-1))^k}{k}\right) - x_2\sum_{k=1}^{\infty}\left((-1)^{k+1}\frac{(x_2-1)^k}{k}\right)}{(x_2-1)^2} \\ = -\frac{1}{2(1-q)}, \end{aligned} \tag{B.21}$$

then Equation (B.16) results in

$$\lim_{p\to\infty} P^*(x_2) \simeq \sqrt{\frac{p}{2\pi}} e^{-(p/2)(x_2-1)^2} = \text{Normal}(1, 1/p), \tag{B.22}$$

[3]Expression is simplified using Mathematica.

that is a Normal distribution with unitary mean and variance equal to $1/p$.

B.3 Clonal dynamics for random models

In Section 4.3.4, we presented the results of numerical simulations for a large number of random models. Here we describe how these random models have been generated.

B.3.1 Model Description

We recall from Section 2.1, that cell fate dynamics can be modelled as a continuous-time multi-type branching process [Haccou et al., 2005], that is, a Markov process following the rules of Equations (2.1)-(2.3). Without losing generality, considering an arbitrary number m of cell states, X_i, for $i = 1, ...m$, we model here only two types of events, described as follows.

- **Cell divisions**: a cell in state X_i divides in two cells respectively in state X_j and X_k at a given rate λ_i.

$$X_i \xrightarrow{\lambda_i} X_j + X_k, \qquad i, j, k = 1, ..., m, \tag{B.23}$$

  In this formulation of cell division events, we consider only one possible division outcome upon division of a particular cell state X_i. Nonetheless, multiple division outcomes per state, i.e. based on parameter r_i^{jk} in (2.1), can be implemented by considering additional *metastates*, representing priming of a state X_i towards a certain division outcome option. For example, if in the original model, state X_i has different outcome options, $X_{j_1} + X_{k_1}, X_{j_2} + X_{k_2}, ...$, we can substitute this by, first, transitions from X_i to (new) states $X_{m_1}, X_{m_2}, ...$ and subsequent divisions $X_{m_l} \to X_{j_l} + X_{k_l}$. The validation of the use of metastates to model more complex processes is discussed in detail in Section B.3.2.

- **Direct state transitions**: a cell in state X_i changes to state X_j at a given rate ω_{ij}.

$$X_i \xrightarrow{\omega_{ij}} X_j, \qquad i, j = 1, ..., m \text{ with } \; i \neq j, \tag{B.24}$$

  Importantly, we include cell loss in this scheme by treating it as a transition to an additional special state, called hereafter *death* and denoted by $\emptyset$ (cells in this state do not enter in the counting of the total number of cells). This means that loss rates γ_i in (2.3) correspond here to $\omega_{i\emptyset}$.

These events define a Markov process, which can be represented as a cell state network, as defined in Section 2.1.2. In this view, each node can be related to a cell

state, while the links represent transitions between states via cell divisions and the direct state transitions. We recall that the generic ij-element of the transpose adjacency matrix of the network, K, is $\kappa_{ij} = \lambda_i 2r_i^j + \omega_{ij}$ for $i, j = 1, ..., m$, and it represents the total transition rate, as defined in Equation (2.7). The matrix A, defining the dynamics of the mean number of cells, is related to K since $A = K - \text{diag}(\delta)$, where δ are the total loss rates (see details in Section 2.1.1 and Section 2.1.2). However, in this modelling, the death state $\emptyset$ is treated as a cell state (except it does not enter in the total cell counting) and that the term $r_i^j \leq 1$ is not a continuum value, but instead, it can only take the values $0, 1/2, 1$ depending on the specific outcome of the division of X_i-cells. Notably, more than one stochastic network may result in the same matrix K; therefore, to uniquely define a process, we distinguish a matrix D which describes cell division events (note that this is possible with just a single matrix as there is only one division option per state) and a matrix T which describes direct transition events. The matrix K is the sum of both, $K = D + T$.

B.3.2 Test case: metastate modelling

As argued before, we assume in the random model generation that division of X_i-cells has a unique outcome, $X_i \to X_j + X_k$ given by Equation (B.23). In this way, the stochastic process can be uniquely defined by the two matrices D and T. To accommodate for the possibility of different division outcomes from the same state X_i, as in Equation (1.1) for the modelling of the PA cell fate and as in Equations (2.1)-(2.3), for a generic process, we introduce in the modelling a set of *metastates*. Metastates represent short-lived states that indicate priming for either outcome, from which the cell division outcomes are unique. We will show here that if the metastates are traversed sufficiently quickly, which can be assured by choice of high direct state transition rates in the metastates, this modelling does not lead to significant deviations from the original model (where multiple outcomes for each cell division are possible).

To illustrate this, let us consider the test case described in Appendix B.1.2, which corresponds to model (B.1). Here, instead of having three different outcomes upon division of an S-cell, we define a Metastate model (MS) as

$$\begin{aligned} S &\xrightarrow{\omega_1} M_1,\ M_1 \xrightarrow{\lambda_1} S + S, \\ S &\xrightarrow{\omega_2} M_2,\ M_2 \xrightarrow{\lambda_2} S + D, \\ S &\xrightarrow{\omega_3} M_3,\ M_3 \xrightarrow{\lambda_3} D + D, \\ D &\xrightarrow{\gamma} \emptyset, \end{aligned} \tag{B.25}$$

in which S and D correspond to the same cell type of model (B.1) (i.e. respectively the stem and the differentiated cells), and M_i, for $i = 1, 2, 3$, represent the metastates. These states are temporary states that are used to model each one of the three different

possible division options of the S-cells. The rates λ_i and ω_i, for $i = 1, 2, 3$, are chosen such that the time scales of division and outcome probabilities are the same as in the original model, that is,

$$\frac{\omega_1}{\omega_2} = \frac{r}{1-2r}, \frac{\omega_2}{\omega_3} = \frac{1-2r}{r}, \tag{B.26}$$

$$\frac{1}{(1/\omega_1 + 1/\lambda_1)} = \lambda r, \frac{1}{(1/\omega_2 + 1/\lambda_2)} = \lambda(1-2r), \frac{1}{(1/\omega_3 + 1/\lambda_3)} = \lambda r. \tag{B.27}$$

Equations (B.26) assure that outcome probabilities are the same as in the original model, while Equations (B.27) are needed to have the same total average time between two consecutive events. As there are six unknowns (λ_i and ω_i, for $i = 1, 2, 3$) and only five relations (Equations (B.26) and (B.27)), the following additional equation is added

$$\lambda_1 = \omega_1 \Delta, \tag{B.28}$$

in which Δ is an additional parameter that is used to control how fast cells in metastate M_1 divide. Low values of Δ imply that as soon as an S-cell transits to the metastate M_1, it divides in two S-cells. Globally, this results in

$$\begin{aligned} \omega_1 &= \omega_3 = \lambda r(\Delta + 1)/\Delta\,, \\ \omega_2 &= \lambda(1-2r)(\Delta + 1)/\Delta\,, \\ \lambda_i &= \omega_i \Delta \text{ for } i = 1, 2, 3\,. \end{aligned} \tag{B.29}$$

Numerical simulations for the two models were run and compared, based on the parameters reported in Table B.1, and specifically for the PA#1 and PA#3 test case settings. The arbitrary time unit is omitted. The process rates for the corresponding MS model are computed based on Equation (B.29) and $\Delta = 1/500$. As well as for the **PA** test cases, the initial condition is one cell of type S and the final time, τ, is equal to 10; simulations are based on 5×10^4 trajectories.

In Figure B.2, we show the mean number of cells in the surviving clones and the extinction probability as a function of the time scaled by τ (left) and the clone size distribution at τ (right). Simulations for **MS** test cases, indicated as MS#1,3, agree very well with the corresponding **PA** ones, which justifies our approach to generate the random models.

B.3.3 Generation of Random Models

For testing the behaviour of the clonal dynamics in a generic homeostatic model, we generate a large number of random cell state networks, whereby each network corresponds to a distinct set of parameters $\lambda_1, ..., \lambda_m, \omega_{12}, ..., \omega_{m\emptyset}$ for the stochastic cell fate model defined by (B.23) and (B.24). The strategy followed is based on the key

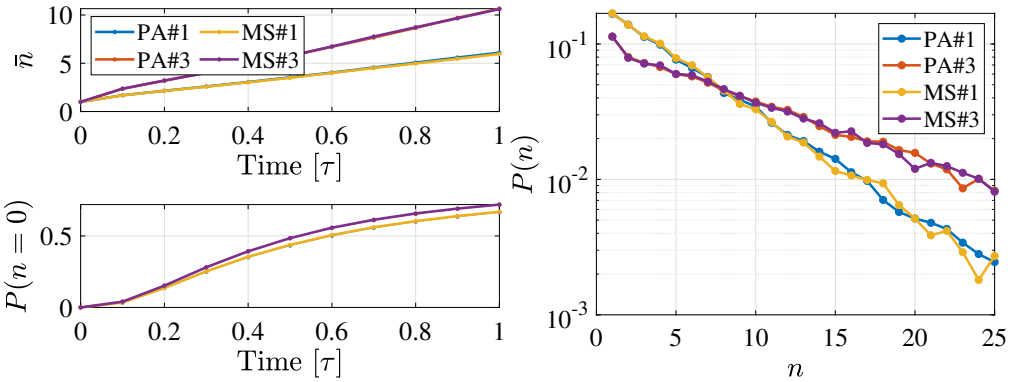


FIGURE B.2: Test case simulation results in terms of mean number of cells in the surviving clones $\bar{n}_s$ and extinction probability $P(n=0)$ as function of time, scaled by the final simulation time τ, and clone size distribution $P(n)$, that is the distribution of the total number of cells n forming the progeny of a single initial stem cell (right). Profiles from the numerical simulation for cases MS#1,3 are compared to the corresponding PA#1,3 test cases which are based on parameters provided in Table B.1 and discussed in Appendix B.1.2.

requirements to achieve homeostasis detailed in Section 2.2.1, which are summarised as: a) each network is composed of Strongly Connected Components (SCCs) that are randomly connected; b) only one SCC, the one at the apex of the network, forms the renewing compartment, $\mathcal{R}$, (i.e. the dominant eigenvalue of A is $\mu = 0$) and all the others form the committed compartment, $\mathcal{C}$, (i.e. they are characterised by a dominant eigenvalue $\mu < 0$).

A two-step process is followed: 1) a large number of (random) SCCs are generated; 2) a condensed network (i.e. corresponding to the cell type condensed network defined in Section 2.1.2) is randomly constructed and filled with randomly picked SCC from Step 1. It is noted that unitary rates are assumed in Step 1. They are successively modified in Step 2 to achieve the desired properties of the dominant eigenvalue μ while ensuring randomness.

We focus now on Step 1, which is the generation of isolated SCCs.

(1.a) The total number of states composing the SCC is defined, indicated as m_S. An additional state is added to represent whatever is outside the SCC. In the current analysis we set $1 \leq m_S \leq 4$.

(1.b) We build all the possible combinations of transition and division matrices separately, indicated hereafter, respectively, with M_T and M_D. These matrices are ordered for increasing number of transitions N_T and divisions N_D. In case GIA networks are generated, the M_D and M_T combinations are filtered, to remain just with those where the division outcome is one cell inside the SCC and one outside the SCC, and where there are only transitions between states within the SCC (i.e. where cell numbers are conserved). From a computational point of view, this process is feasible up to $m_S = 4$.

(1.c) The matrices stored in M_D and M_T are then combined together to form a model (which is completely defined by one matrix in M_D and one in M_T); M_{DT} indicates the pool of possible models. This process is done considering separately each m_S, N_T and N_D. In this step, due to technical limitations given by the high number of possible combinations, if the total number of combinations exceeds 5×10^4, then only 10^4 random matrices from M_D and M_T are combined.

(1.d) Each model in M_{DT} is then processed to check if the corresponding network is an SCC in the first m_S states. If not, then this model is discarded. In case GPA networks are generated, a further check is performed to discard also those models consistent with a GIA network (they cannot be a priori excluded as done in point (1.b) for the GIA ones). These pools of models are indicated as M_{GIA} and M_{GPA} respectively for the GIA and GPA models.

(1.e) For each SCC in M_{GIA} and M_{GPA}, the dominant eigenvalue μ is estimated. By construction, the generated GIA networks are all characterised by $\mu = 0$, while in general, any value can be obtained within M_{GPA}.

(1.f) The SCCs in M_{GPA} are additionally processed to check whether the network is compatible with homeostasis by tuning the rates. Networks satisfying this condition are additionally stored under a new pool of SCCs, called M^*_{GPA}. If not, they are discarded when $\mu > 0$ (i.e. for any combination of rates, the number of cells in these networks is expected to grow).

This process results in three pools of SCCs classified for m_S, N_T and N_D (i.e. number of states, transitions and divisions): 1) M_{GIA} contains GIA models; 2) M^*_{GPA} contains GPA models that can be tuned to have $\mu = 0$ and 3) M_{GPA} contains GPA models characterised by $\mu < 0$ or that can be tuned to meet this condition.

In Step 2, the generation of random networks starting from the individual SCCs is implemented as follows.

(2.a) A number of committed SCCs, N_c, between one and three, is randomly chosen.

(2.b) N_c SCCs are randomly picked from the pool of models M_{GPA}. The selection is done considering equal probability in m_S, N_T and N_D. For each SCC, the unitary rates α (where α stands for any rate λ_i or ω_{ij}) are modified by multiplying them by random numbers (exponentially distributed with mean $\bar{\alpha} = 1$ and minimum $\alpha_m = 0.3$). Additionally, a threshold on the dominant eigenvalue is set, $\mu_{\max} = -1$; if this condition is not satisfied, then the rates are tuned to meet this requirement while maintaining the rates above the minimum.

(2.c) The committed compartment of the condensed network is generated by randomly connecting all the outgoing components of the kth SCC with states in

the lth SCC for $l = k+1, .., N_c$. In this way, the transposed adjacency matrix of the cell state network has a triangular block form of the type

$$K = \begin{pmatrix} K_1 & 0 & 0 & 0 & ... \\ C_{21} & K_2 & 0 & 0 & ... \\ C_{31} & C_{32} & K_3 & 0 & ... \\ \vdots & \vdots & \vdots & \ddots & 0 \\ C_{1\emptyset} & C_{2\emptyset} & ... & ... & 0 \end{pmatrix}. \tag{B.30}$$

The last SCC is forced to be linked to a single death state.

(2.d) With a similar procedure described in point (2.b), two SCCs are randomly picked respectively from the pool of SCCs in M^*_{GPA} and M_{GIA}; the unitary rates are modified (exponentially distributed with mean $\bar{\alpha} = 1$ and minimum $\alpha_m = 0.3$) and, in the GPA case, tuned to meet the condition $\mu = 0$. They represent the renewing part of the network.

(2.e) Two networks, one for the GIA and one for the GPA models, are produced by attaching the selected renewing network upstream the committed one(s). This is done based on an analogous procedure as described in step (2.c).

At the end of this process, we have two networks that differ in the renewing part, one consistent with a GIA model and the other with a GPA one. In total, 2000 networks were constructed.

B.3.4 Simulation campaign

An extensive simulation campaign was run to model the clone dynamics based on the random cell fate models described in the previous section. Since a clone is, by definition, the progeny of a single cell, we choose to initialise the simulation with a single cell in a random state within $\mathcal{R}$. Given the substantial difference in the dynamics of GIA and GPA models, the final time, indicated by τ, is set equal to 20 times the inverse of the minimum process rate, $\alpha_{\min} = \min(\lambda_1, ..., \lambda_m, \omega_{12}, ..., \omega_{m,\emptyset})$, in the GIA models, and to the time at which the fraction of extinct clones reaches 98% in the GPA models[4]. We use 10^3 and 5×10^4 simulations for each GIA and GPA model to determine the clone size distribution. In this way, both models result in the same final number of clones when 98% extinction is considered.

[4]Note that all critical branching processes, as homeostatic clonal dynamics are, will go extinct almost surely at some point in time [Haccou et al., 2005].

Appendix C

Additional analyses for the study case

C.1 Single-cell RNA-sequencing analysis

In this section, we report the work carried out to produce the results shown in Section 5.2.1. The main idea is to provide an unbiased analysis of the available databases described in Section 1.4.1. The methodology used is described in Appendix C.1.1. As a validation of the approach and the implementation, we compared our analysis with the published results in Appendix C.1.2; in the same section, we also evaluate the performance of other methods. Details of the analysis are reported in Appendix C.1.3.

C.1.1 Methodology

As shown in Section 1.4.1, the four works differ for the experimental setup and the type of raw data provided. Additionally, from the review of the methods applied, it is clear that the data analysis approach varies in each case. This includes, for example, the type of normalisation, the clustering method, pseudo-time analysis. Thus, we used the same tool and methods (as far as possible) to analyse the four datasets consistently. Among the available software, such as scran [Lun et al., 2016a] and EdgeR [Robinson et al., 2010, McCarthy et al., 2012]), we chose Seurat (v2.0)[1] [Butler et al., 2018] for its wide range of features and its clear and well documented implementation. Some limitations are related to the reduced number of methods available in some cases. However, it is noted that this limitation can be easily overcome by creating wrappers

[1]This version was the one available at the time of running this comparison (2018). Considering that the scRNA-seq experiment was cancelled in January 2019, we decided not to upgrade the code to later versions.

to external functions as done, for example, in the analysis reported in Appendix C.1.2.5. Importantly, Seurat was not used in any of the articles surveyed.

The main steps for processing scRNA-seq data are summarised below.

1. **Data Loading.** Data are imported from external files (mmt, csv or txt format). The expression level is loaded into a sparse matrix where rows are the genes and columns correspond to cells. Depending on the case, imported data consist of UMI, read counts or normalised read counts. Thresholds in the minimum number of cells expressing a gene, minimum genes per cell and minimum level of expression detected can be applied.

2. **Filtering.** Raw data quality checks exclude poor quality cells from the analysis. Different criteria can be applied, such as the minimum/maximum number of genes or transcripts expressed in each cell or the mitochondrial genes expression percentage. As detailed in the next section, in each database, we use the same filtering criteria used in the related published work.

3. **Normalisation.** Normalisation is a fundamental step in scRNA-seq processing to remove unwanted sources of variation. Among others, the detected gene expression levels are affected by the sequencing depth, the amplification, the efficiency of the reverse transcription, which in turn depends on the technology employed. In [Vallejos et al., 2017] a review of the strategies of existing methods and challenges is provided. This work shows how the normalisation method affects the selection of the high variable genes and how new recent methods explicitly developed for scRNA-seq perform better than those inherited from bulk analyses (i.e. global methods). However, in the version of Seurat we used, the only method available is a global one, and it does not consider variability in the total expression of different cell types and stochasticity in the process of sequencing (i.e. noise). That means that the levels of expression in each cell are scaled by the total expression, and the result is multiplied by a scaling factor (the same for all the cells) and returned in a logarithmic scale[2]. Alternative and more sophisticated methods estimate a cell-specific size factor that corrects its library size. As an example, the sumFactor method [Lun et al., 2016b], which is part of scran package, is used in [Bach et al., 2017]. This method was tested and compared with that used in Seurat, showing that in the datasets under analysis, the selection of the normalisation method has no significant impact in the final clustering of the data (see Appendix C.1.2.5). In two out of four datasets, the data provided are already filtered and normalised. Thus, in these cases, Steps 2 and 3 are skipped.

[2]The normalised expression is shifted by one before applying the logarithmic scale to avoid infinite values when the expression is zero.

4. **High Variable Genes.** Strongly coupled with data normalisation, the identification of the HVG is aimed at distinguishing biological and technical variations in the level of gene expression. In this way, the significant genes are detected, reducing the dimension of the problem appreciably. In [Sham et al., 2018] a comparison of tools and methods is reported. Seurat seems to have overall good performance in this work, although scran is considered the best tool. Once again, as shown in Appendix C.1.2.5, for the datasets under analysis, Seurat and scran results are very similar. Seurat function for detecting HVG requires manual tuning of some threshold parameters based on a visual inspection of the average expression as a function of the average dispersion plot.

5. **Data Scaling.** Considering that different levels of expression are associated with each gene, the aim of data scaling is the generation of a uniform (across genes) dataset. In Seurat, the default scaling is done for each gene by centring the normalised expression around its mean and scaling by the standard deviation.

6. **Dimensionality reduction.** Dimensionality reduction aims at simplifying a dataset by reducing its size without losing key information. Two widely used methods in the scRNA-seq analysis are Principal Component Analysis [James et al., 2014] and the t-distributed Stochastic Neighbor Embedding [van der Maaten and Hinton, 2008]. Our analysis follows a standard approach and combines both methodologies. First, PCA is run based on the HVG. Then, t-SNE estimation (used later for visualisation, see Step 9.) and clustering are performed based on the minimum number of PCs representative of the database.

7. **Clustering.** The clustering is an analysis aimed at organising data in subgroups that present a certain degree of similarity. The similarity between observations is based on a defined metric, such as Euclidean distance and correlation-based distance, which depends on the specific application. In Seurat (v2.0), the algorithm available is the SNN clustering. This method works well with high-dimensional and noisy data, such as scRNA-seq, where clusters have different shapes and densities, [Ertöz et al., 2003, Xu and Su, 2015].

8. **Differentially Expressed genes.** The identification of markers associated with each cluster is based on the analysis of the Differentially Expressed genes. Several different tests to detect DE genes exist. A review and comparison of different methods are provided in [Soneson and Robinson, 2018]. In our analysis, we used the default settings of the Seurat function to identify DE genes.

9. **Visualisation.** This is a fundamental step in the data analysis. Common techniques in scRNA-seq analysis are t-SNE plot and heatmap. The t-SNE plot is a dimensionality reduction (see Step 6) and visualisation technique in which points close in the reduced space correspond to observations close in the original space. The process is based on a stochastic selection of the observations, for

which two similar observations have more probability of being close in the reduced space than two dissimilar ones. Instead, heatmap [Wilkinson and Friendly, 2009] shows the gene expression levels in all the cells as a matrix in which each row corresponds to a gene and each column to a cell (cells are usually reordered and grouped in clusters); the colour indicates the level of expression of each gene in each cell, which, in general, can be absolute or relative to the average value.

C.1.2 Comparison with published results

In this section, we report and compare the outputs of the analysis with the published results in terms of a) the number of cells in each cluster; b) visual comparison of clusters in reduced dimension plot; 3) visual comparison of heatmap and (or) specific gene expression levels. The analysis settings, summarised in Table C.1, were tuned to obtain results as close as possible to the reference ones. However, in most cases, default values are used. We remark that in Ds#2 Pal et al. [2017] both the Puberty (5W) and Adult (P7) samples are analysed.

Parameter	Value			
Ds#(Sample)	1(NP)	2(P7) 2(5W)	3(V)	4(Adu1)
Loading Data				
min.cell*	1	1	1	5
other	-	-	mt-genes excl.	-
Filtering				
Criteria	Bach et al. [2017]	$\leq$ 5% mt-genes	none	none
Normalisation				
Method	Total expression	Total expression	log(FPKM+1)	log(NRC+1)
High Variable Gene				
x.low.cutoff*	0.1	0.1	0.75, 8	0.05
x.high.cutoff*	3	3	8	3
y.cutoff*	[0.5, Inf]	[0.5, Inf]	[0.5, Inf]	[0.5, Inf]
Dimensionality Reduction				
pcs.compute*	50	50	50	50
dims.use*	[1, 10]	[1, 10]	[1, 10]	[1, 10]
tsn_perplexity*	50	50	10	50
Clustering				
k.param*	30	10	10	10
resolution*	0.3	2.0	2.2	0.3

TABLE C.1: Settings for the analysis of the scRNA-sequencing literature data. Values were tuned to obtain results as close as possible to the reference ones, but default values are used in most cases. The superscript * indicates a direct input to a Seurat function.

C.1.2.1 Dataset 1

In [Bach et al., 2017], four main clusters are identified for the nulliparous (NP) stage (it corresponds to the adult virgin): Basal (Bsl); Luminal progenitor (Lp), Hormone sensing progenitor (Hsp) and Hormone sensing differentiated (Hsd). Few cells in the Alveolar progenitor (Avp) and Procr/basal cells (Prc) are detected.

Data analysed in this section are indicated as Ds#1 and correspond to NP-1 and NP-2 samples. They were downloaded from the GEO database (GSE106273). The number of cells after the filtering is 4223 (3.5% are filtered out), which corresponds to the declared number of cells for these two samples. The search for high variable genes detected 906 genes, including the key genes identified in the reference work. The NP-1 and NP-2 samples overlap in the t-SNE plot (not shown), confirming no bias between the two samples.

Our analysis results in six clusters, highlighted in Figure C.1. Two of them are consistent with the exclusion criteria given in the Supplemental Material of this reference, and they are labelled as contaminating cells (Cont). The remaining clusters were easily related to the reference ones identified in the article (based on the declared cell types markers). The number of cells obtained in each cluster is reported in Table C.2 and compared with the declared data. The maximum difference in the relative size of the clusters is around 1%. Additionally, the expression levels and heatmap of key genes shown in Figure C.2 and Figure C.3 are consistent with those shown in [Bach et al., 2017].

Globally, the performed analysis agrees very well with the published one.

		Ref.		**Value**	
Cluster ID	Identity	Number of cells			
		[-]	[%]	[-]	[%]
C13 (C12, C15)	Bsl (Prc)	788	19.1	781	19.1
C6 (C7, C10)	Lp (Avp)	692	16.8	696	17.0
C2 (C1)	Hsp	435	10.6	474	11.6
C4 (C3)	Hsd	2200	53.5	2148	52.4
	Total	4115	100.0	4099	100.0
	Cont	108	-	124	-

TABLE C.2: Summary of clusters, identities and cells number for Ds#1. The column labelled **Ref.** reports the reference values published in [Bach et al., 2017]; the column **Value** corresponds to the analysis reported in this section.

C.1.2.2 Dataset 2

In [Pal et al., 2017], the 10X samples of adult and puberty epithelial cells are analysed independently. In each case, the same six clusters are identified: Basal (Basal);

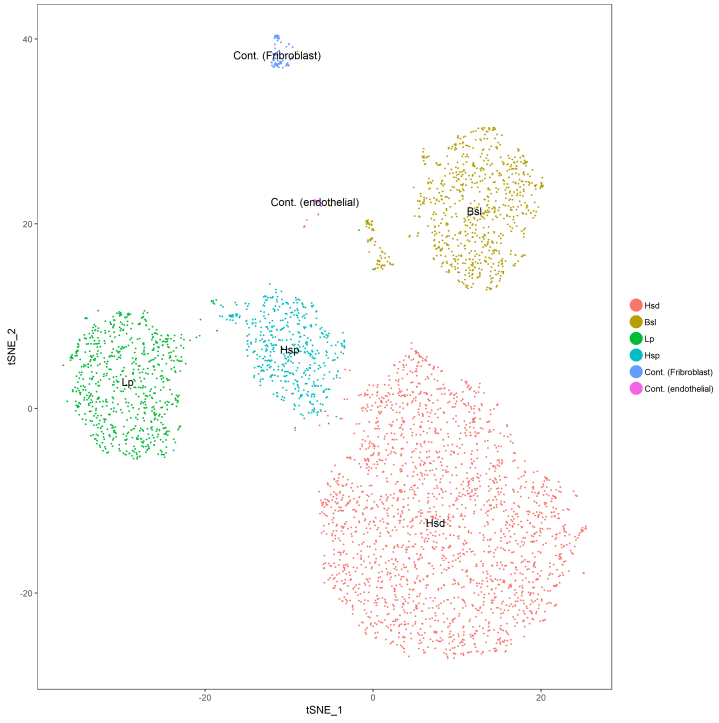


FIGURE C.1: Data analysis (Ds#1) in terms of t-SNE plot, clusters and identity. Each point corresponds to a single cell which is coloured according to its identity.

Luminal Progenitor (LP), Luminal Intermediate (Lum Int), Mature Luminal (ML) and three rare clusters (#5, #6 and #7).

Data presented in this section is taken from GEO database (GSE103275) and correspond to samples P7 (adult) and 5W2 (puberty). The two samples are analysed independently. Although no cells were excluded during the data import and filtering steps in the P7 sample, a slight discrepancy in the total number of cells is detected (less than 0.2%). In the 5W2 sample, the filtering removes 0.4% cells. Concerning the data clustering, a fine-tuning of the cluster resolution was necessary to identify the rare clusters #6 and #7. The initial clustering resulted in many clusters (26 and 27 respectively for the P7 and 5W2 samples). Most of these clusters were subsequently merged by a visual inspection of the t-SNE plot combined with a careful analysis of some known markers. This process leads to the identification of the six reference clusters, as shown in Figure C.4 and Figure C.7 respectively for the P7 and 5W2 samples. The size of each cluster is reported in Table C.3. The difference in the relative estimation of the cluster size is below 1%. For the P7 sample, the expression levels and heatmap of some key genes, shown in Figure C.5 and Figure C.6, agree with the published results.

Again, the performed analysis is consistent with the published one. However, in addition to the discrepancy in the cell number, differences are found in the t-SNE plot for the 5W2 sample. Interestingly, in this case, the clusters #6 and #7 are much closer to the basal cluster than the luminal ones.

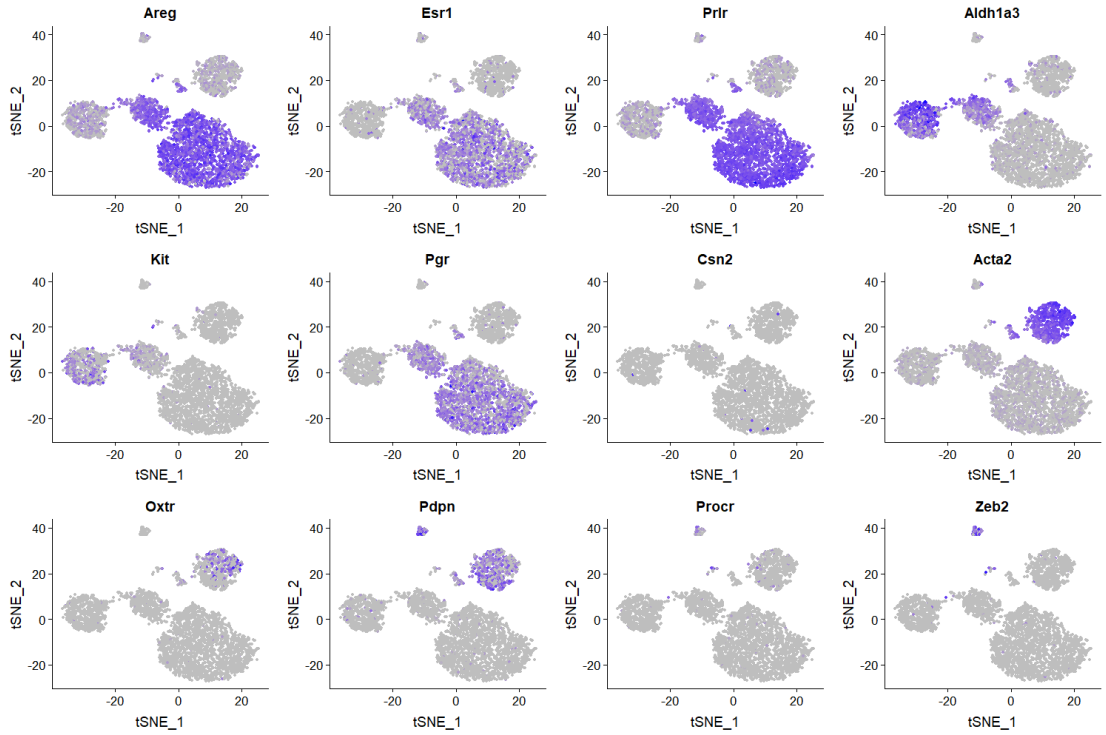


FIGURE C.2: Data analysis (Ds#1) in terms of key genes expression level. In each panel, single cells are coloured according to their level of expression over a t-SNE plot: blue means that a gene is highly expressed and grey corresponds to no expression.

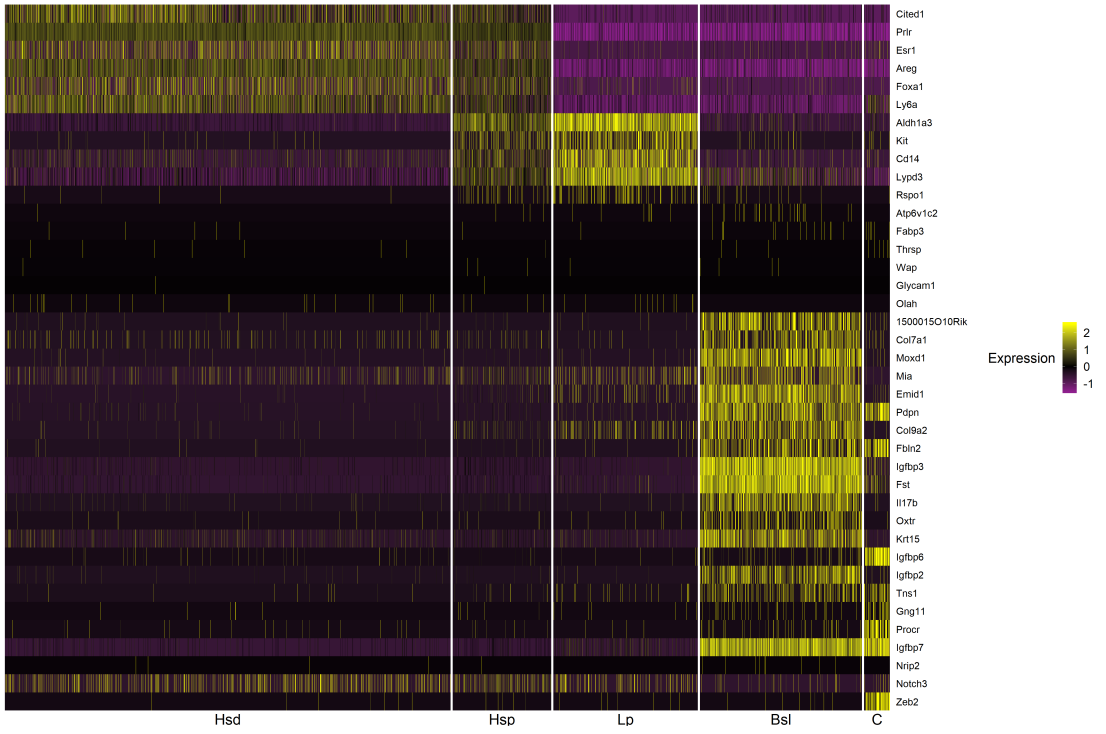


FIGURE C.3: Data analysis (Ds#1) in terms of gene expression heatmap. The heatmap shows the relative gene expression level, whereas each column is associated with a single cell (cells are reordered and grouped in clusters) and rows correspond to genes. In this figure, C corresponds to Contaminating cells.

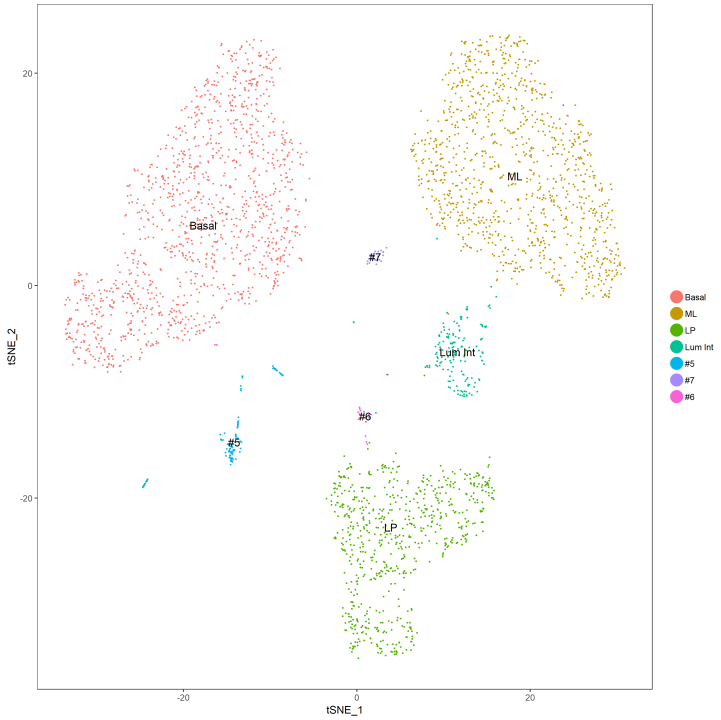


FIGURE C.4: Data analysis result for Ds#2/P7 in terms of t-SNE plot, clusters and identity. Each point corresponds to a single cell which is coloured according to its identity.

	Ref.		**Value**		**Ref.**		**Value**	
Sample	P7				5W2			
Cluster	Number of cells							
	[-]	[%]	[-]	[%]	[-]	[%]	[-]	[%]
Basal	1249	37.8	1239	37.5	1118	20.8	1134	21.1
LP	667	20.2	667	20.2	1760	32.7	1767	32.8
ML	1057	32.0	1058	32.0	2113	39.2	2130	39.6
LI	166	5.0	158	4.8	289	5.4	253	4.7
R5	125	3.8	133	4.0	14	0.3	13	0.2
R6	22	0.7	22	0.7	33	0.6	30	0.6
R7	22	0.7	25	0.8	60	1.1	52	1.0
Total	3308	100.0	3302	100.0	5387	100.0	5379	100.0

TABLE C.3: Summary of clusters, identities and cells number for Ds#2. The columns labelled **Ref.** report the reference values published in [Pal et al., 2017]; the column **Value** corresponds to the analysis reported in this section.

C.1.2.3 Dataset 3

In [Sun et al., 2018] work, seven clusters are identified: C1, C2A, C2B and C3 are classified as basal clusters; C4, C5 and C6 as luminal ones. This classification in basal and luminal cells agrees well with the Fluorescence Activated Cell Sorting (FACS) previously done. Cluster C1 and C2A are mainly composed of cells from the Pregnant samples, so they are not considered in this section. Cluster C2B is the main basal

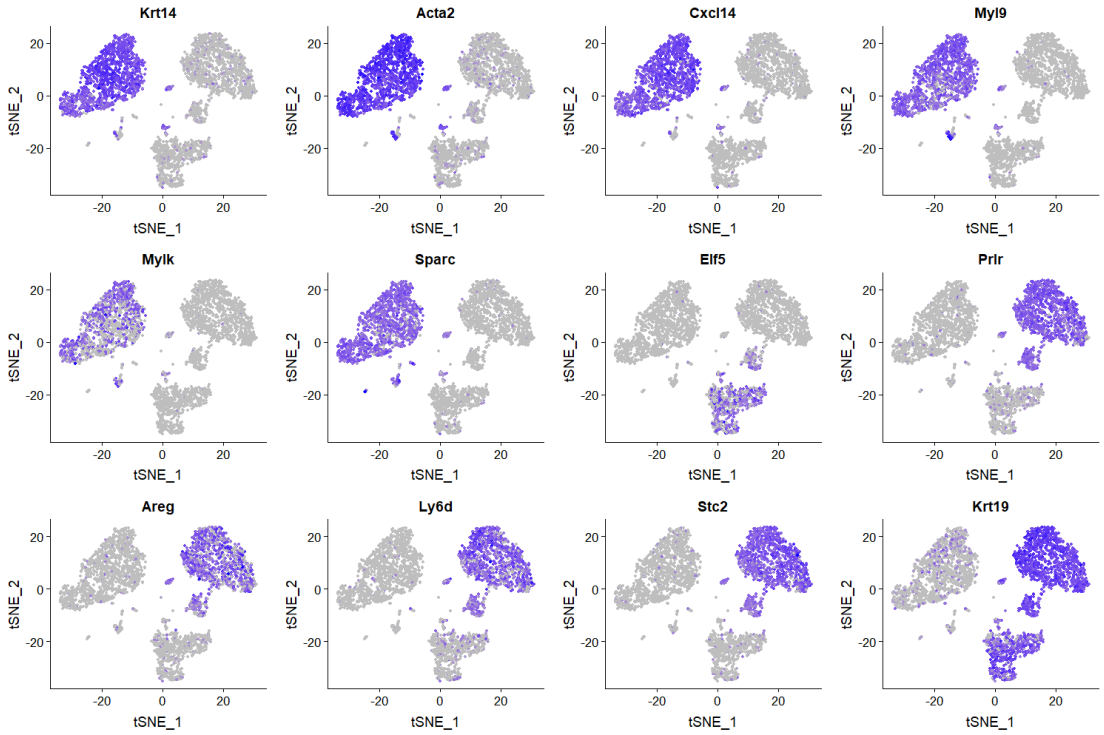


FIGURE C.5: Data analysis (Ds#2/P7) in terms of key genes expression level. In each panel, single cells are coloured according to their level of expression over a t-SNE plot: blue means that a gene is highly expressed and grey corresponds to no expression.

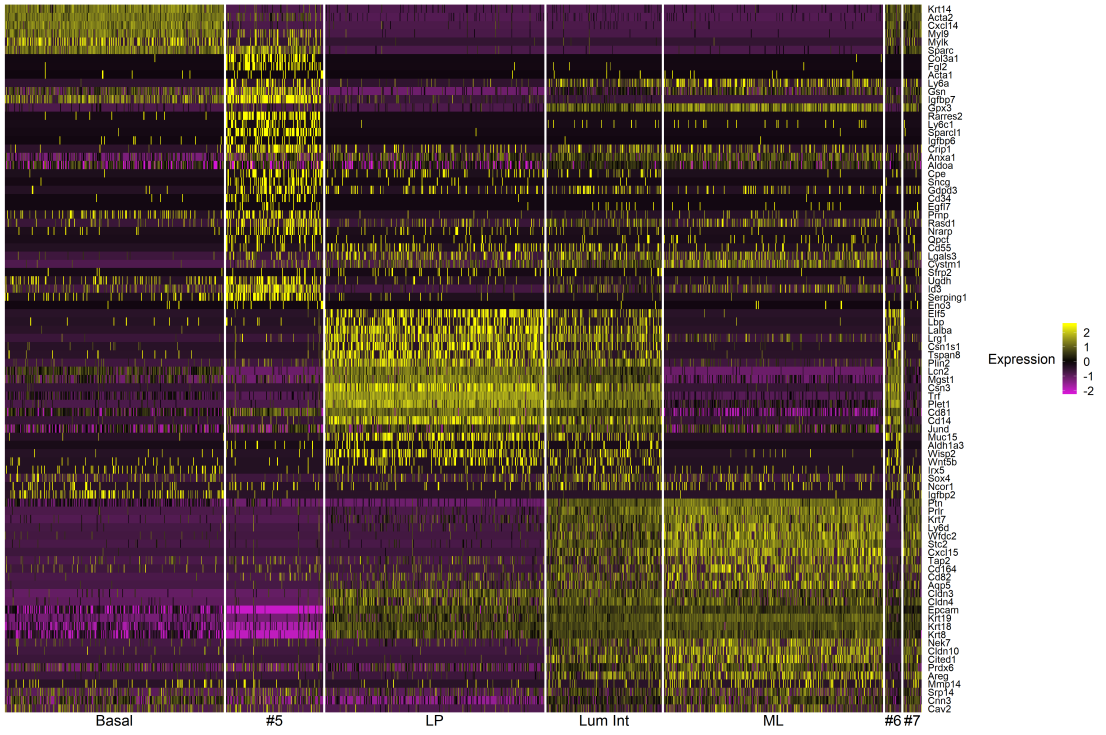


FIGURE C.6: Data analysis (Ds#2/P7) in terms of gene expression heatmap. The heatmap shows the relative gene expression level, whereas each column is associated with a single cell (cells are reordered and grouped in clusters), and rows correspond to genes. In this figure, only 300 randomly picked cells are shown for the large clusters.

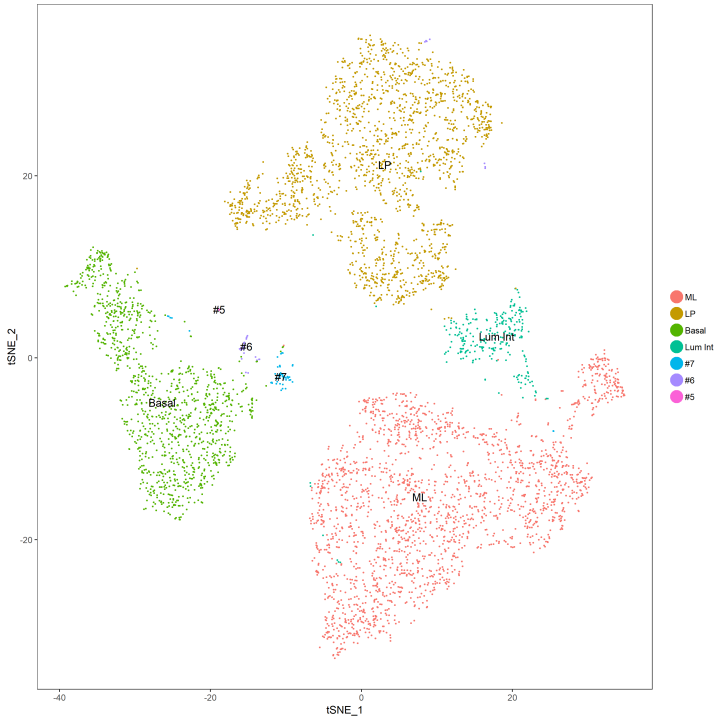


FIGURE C.7: Data analysis result for Ds#2/5W2 in terms of t-SNE plot, clusters and identity. Each point corresponds to a single cell which is coloured according to its identity.

cluster, while C3 is a sub-cluster of basal cells associated with bipotent MaSC cells. Luminal cells in clusters C6 and C5, although not explicitly, can be related to luminal mature due to expression of *Esr1*, *Pgr*, and *Erbb2*; the remaining luminal cluster C4 is just generically associated to luminal cells.

The data analysed in this section is taken from the Supplemental Material of the reference, and it consists of an xlsx-file with the gene counts expressed as FPKM. In this section, a subset of cells is analysed corresponding to the Virgin (V) sample. The number of cells for this sample is 88, and it corresponds to what was declared (we recall that data is already filtered and normalised in this case). The search for high variable genes results in the identification of 2918 high variable genes, which include all of the key genes identified in the reference except for *Inpp5d*. However, it is noted that this gene is expressed only in the Pregnant (P) sample. Clustering analysis results in the identification of seven clusters. These clusters (some were merged) easily related to those declared, except for clusters C4 and C5. These two clusters are similar in terms of expression, and a high clustering resolution was necessary to separate them. Surprisingly, in the published results, in the hierarchical clustering (dendrogram above the heatmap plot presented in [Sun et al., 2018] Figure 2), C5 is under the basal branch and close to the C3 cluster. Instead, in the t-SNE plot (see Figure 3 (A) [Sun et al., 2018]), C5 is between C4 and C6 and partially overlaps with C6. The number of cells in each cluster is reported in Table C.4: difference in the

relative cluster size reaches 10% in C4 and C5. The t-SNE plot for this dataset is shown in Figure C.8. The level of expression of some key markers identified in the paper is shown in Figure C.9 and Figure C.10: they are all consistent with the reference results, with the only exception of *Krt18*. For this gene, the expression level in the present analysis is high in all the luminal cells, while in the published results (Figure S3), almost no expression is detected in clusters C5 and C6.

Overall, the present comparison is good considering that the basal, basal MaSC and mature luminal clusters are correctly identified.

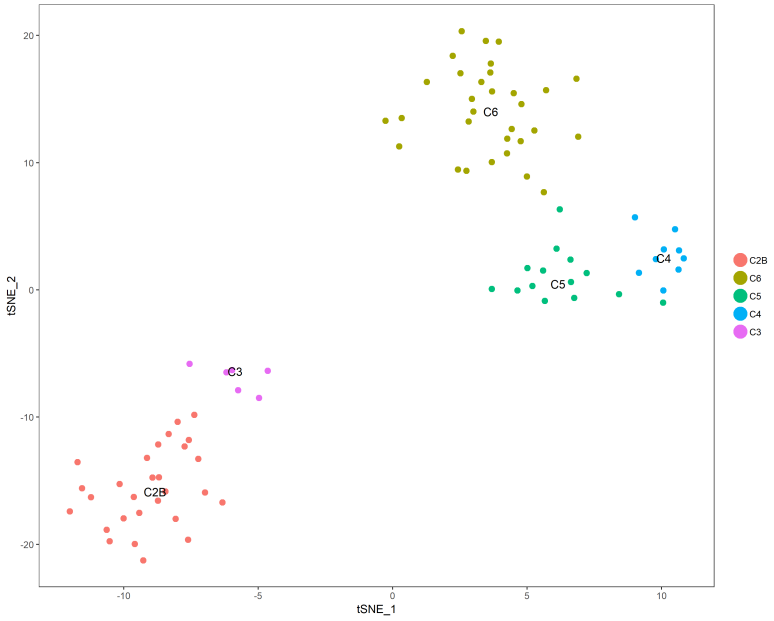


FIGURE C.8: Data analysis result for Ds#3 in terms of t-SNE plot and clusters. Each point corresponds to a single cell which is coloured according to its identity.

Cluster ID	Identity	**Ref.**		**Value**	
		Number of cells			
		[-]	[%]	[-]	[%]
C2B (C1)	Basal	26	29.5	28	31.8
C3	Basal	9	10.2	6	6.8
C4	Luminal	17	19.3	9	10.2
C5	Luminal	5	5.7	14	15.9
C6	Luminal	31	35.2	31	35.2
	Total	88	100.0	88	100.0

TABLE C.4: Summary of clusters, identities and cells number for Ds#3. The column labelled **Ref.** reports the reference values published in [Sun et al., 2018]; the column **Value** corresponds to the analysis reported in this section.

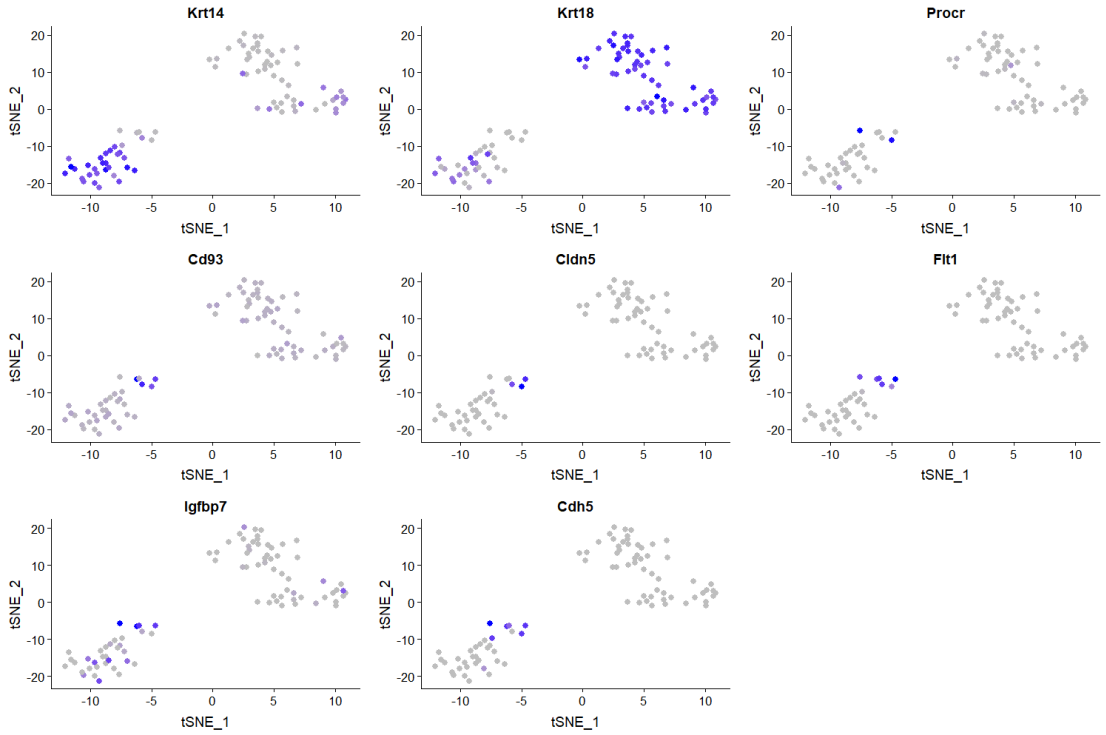


FIGURE C.9: Data analysis (Ds#3) in terms of key genes expression level. In each panel, single cells are coloured according to their level of expression over a t-SNE plot: blue means that a gene is highly expressed and grey corresponds to no expression.

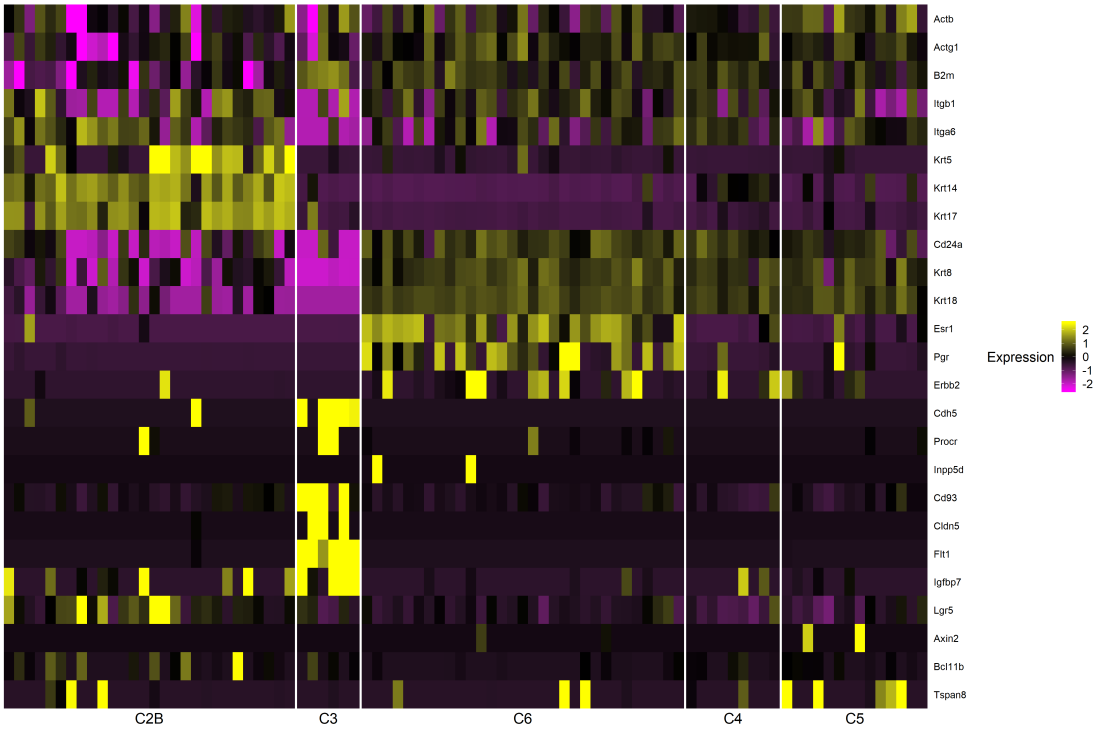


FIGURE C.10: Data analysis (Ds#3) in terms of gene expression heatmap. The heatmap shows the relative gene expression level, whereas each column is associated with a single cell (cells are reordered and grouped in clusters), and rows correspond to genes. In this figure, only 300 randomly picked cells are shown for the large clusters.

C.1.2.4 Dataset 4

The analysis presented in [Giraddi et al., 2018] classifies cells from the Adult sample into three main clusters: basal, luminal mature and alveolar progenitor. An additional cluster, which overlaps with Postnatal and Embryonic samples, is excluded from the analysis as it is associated with contaminating cells. In this work, there is no indication of the number of cells in each cluster. Therefore, only a visual comparison is made in this section.

In this case, the raw data are taken from the GSE111113 series in GEO, and they consist of the filtered and normalised counts. Two samples are available for the adult developmental stage: Adu1 and Adu2. Clusters of cells from the two samples (Adu1 and Adu2) were not overlapping in the t-SNE plot, except for some rare clusters (figure not shown). This is an indication of a potential misalignment in the normalised expression. Therefore, only cells from one sample are analysed here. In particular, sample Adu1 was selected as it is the largest one. The number of cells in this sample is 1979, and 583 high variable genes were detected, including the key genes identified in the reference. Clustering analysis identifies ten clusters, which were analysed and easily related to the declared ones (some of them were merged). The number of cells in each cluster is reported in Table C.5. The smallest and the largest clusters are respectively those for the basal and the mature luminal cells, agreeing with the reference. The t-SNE plot is shown in Figure C.11; the level of expression of some key markers, shown in Figure C.12 and Figure C.13, agrees with the published results.

From a visual inspection of the results, the analysis agrees with the published one. Additionally, the analysis of the Adu2 sample, which is not shown here, gave similar results.

Identity	Number of cells	
	[-]	[%]
Basal	260	13.7
Luminal Mature	1055	55.5
Alveolar Progenitor	586	30.8
Total	1901	100.0
Contaminating	78	-

TABLE C.5: Summary of clusters, identities and cells number for Ds#4. For this database, no reference values are found.

C.1.2.5 Comparison with scran

In this section, we assess the impact of using different methods in the data analysis. In particular, scran package [Lun et al., 2016a] was used, and the results were compared to those obtained using Seurat. More specifically, in the results reported here, the

FIGURE C.11: Data analysis result for Ds#4 in terms of t-SNE plot, clusters and identity. Each point corresponds to a single cell which is coloured according to its identity.

normalisation (when applicable) and HVG methods come from scran package, no PCA is run, and the clustering method (SNN) is directly computed based on the detected high variable genes. The comparison between the results was made in terms of 1) visual comparison of the mean and standard deviation of the normalised expression; 2) common HVG number; 3) the number of cells that are classified in the same cluster. It is noted that the results presented in this section refer to Ds#2/P7 (see Section C.1.2.2), but also the other datasets were analysed.

Concerning the normalisation, the sumFact function (scran) is used based on default settings and considering a pre-clustering step. The average dispersion of the expression in each gene is plotted against the average expression. For Ds#2, this is shown in Figure C.14: although the trend is basically the same, it is apparent that data normalised using scran are characterised by a lower level of dispersion. Improvement in the match can be achieved by tuning the constant scale factor in Seurat (the default value is used here). In Ds#1 instead, the two normalisation methods give almost the same results. In Ds#3 and Ds#4, no normalisation is applied.

When running the HGV function in scran, the following thresholds were used: 0.05 for the adjusted p-value and 0.25 for the biological component of the variance. Based on these settings, the number of genes identified is significantly lower than that obtained using Seurat. In [Sham et al., 2018], instead, an opposite trend is observed (but values for the settings are not specified). The number of HVG detected in the four

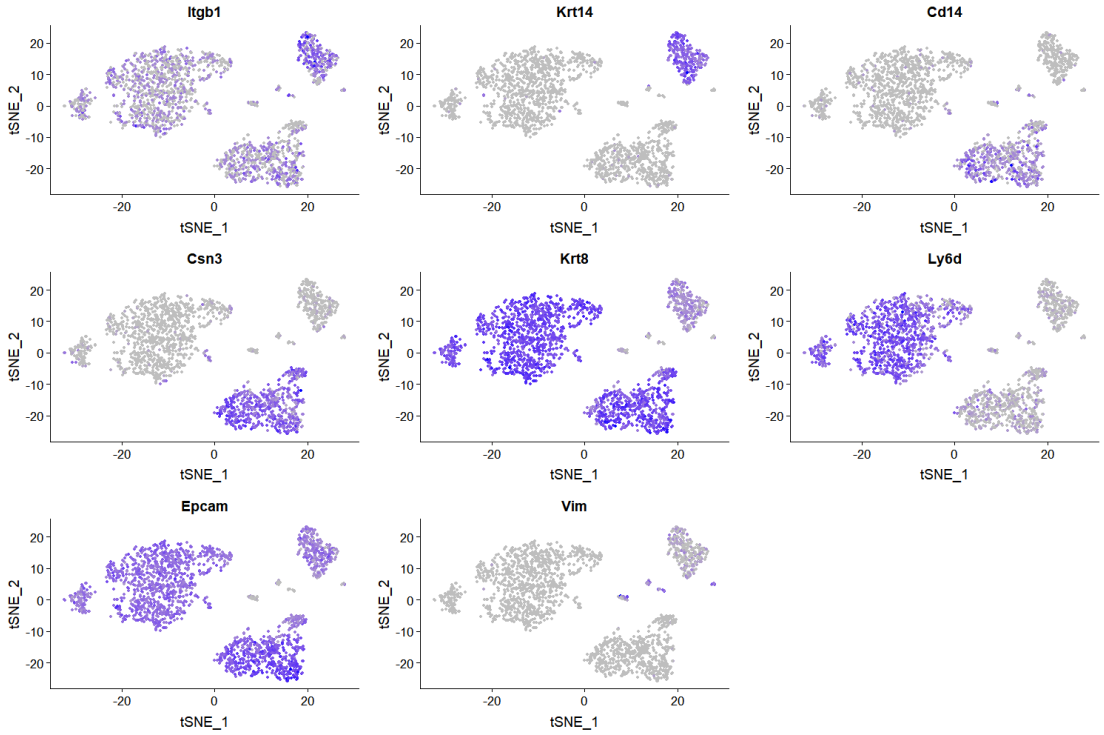


FIGURE C.12: Data analysis (Ds#4) in terms of key genes expression level. In each panel, single cells are coloured according to their level of expression over a t-SNE plot: blue means that a gene is highly expressed and grey corresponds to no expression.

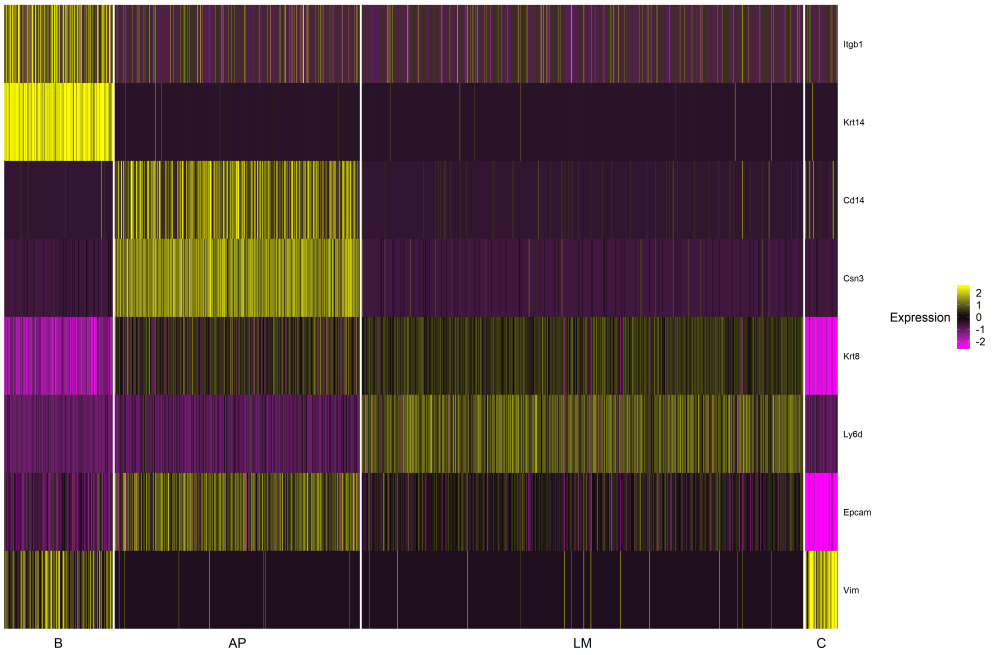


FIGURE C.13: Data analysis (Ds#4) in terms of gene expression heatmap. The heatmap shows the relative gene expression level, whereas each column is associated with a single cell (cells are reordered and grouped in clusters) and rows correspond to genes.

FIGURE C.14: Comparison of normalised expression for Ds#2 computed using scran [Lun et al., 2016a] and Seurat. Each point corresponds to a gene and the plot reports the mean and standard deviation of their expression (in logarithmic scale).

Ds	#1	#2	#3	#4
Seurat	906	1297	2916	583
scran	145	95	366	119
Common	127	58	341	99

TABLE C.6: High Variable Genes (HGV) detected in each database using scran and Seurat.

datasets is reported in Table C.6. Even if the difference is significant, consistent clustering results are always obtained. The worst comparison is obtained in Ds#3, where clusters C4 and C5 were not distinguished by scran (however, Seurat parameters were finely tuned to detect them). The t-SNE plot for Ds#2 is shown in Figure C.15: the clusters seem to be less defined than those resulting from Seurat, shown in Figure C.4 (e.g. part of LI cluster is very close to the LP one). In any case, looking at the number of cells that are clustered consistently using the two approaches, shown in Figure C.16, it is concluded that overall the two methodologies lead to the same cell classification. This consideration applies to the four datasets, where the same identity is given in more than 98% cells.

C.1.3 Database comparison

In this section, we report the detailed comparison of the outputs of the analysis of the four databases. We will focus first on the rare clusters and then on the main ones.

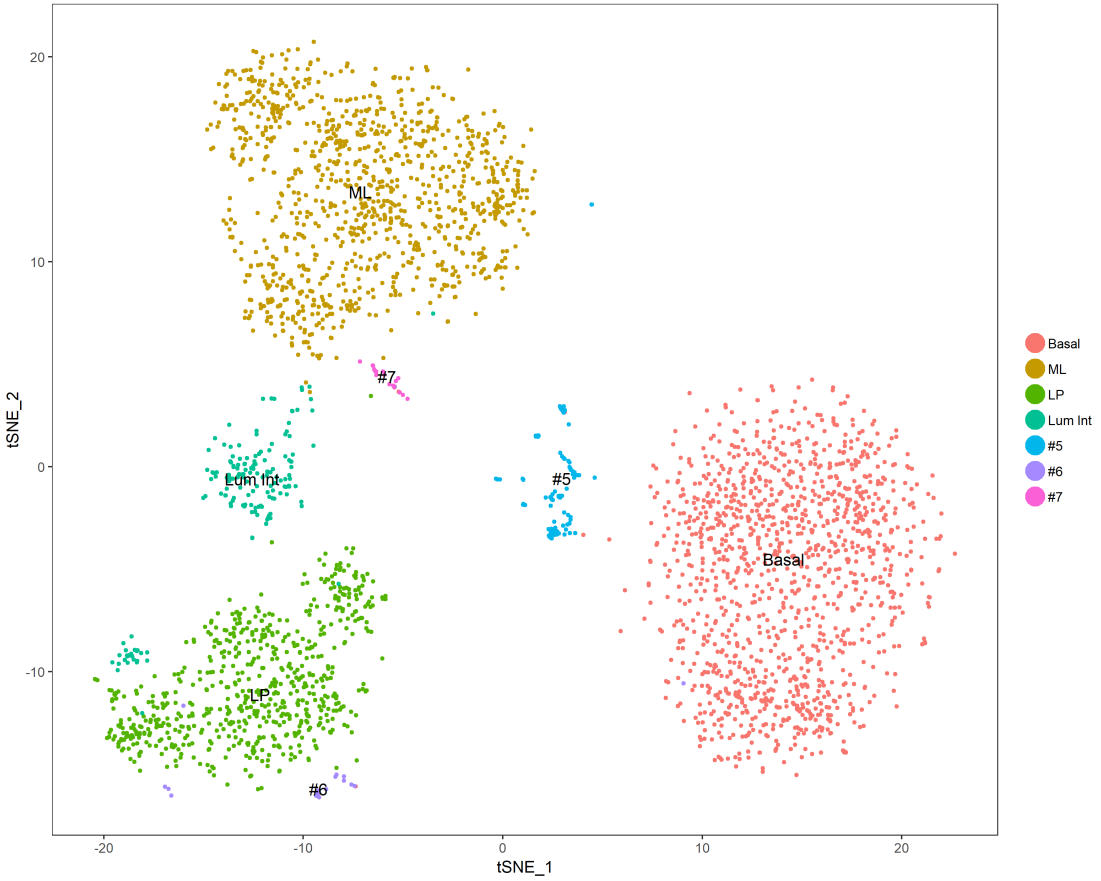


FIGURE C.15: Data analysis result for Ds#2 in terms of t-SNE plot, clustering and cell identity identified using scran. Each point corresponds to a single cell which is coloured according to its identity. This result, obtained using scran package, compares well with the corresponding result obtained with Seurat (see Figure C.4).

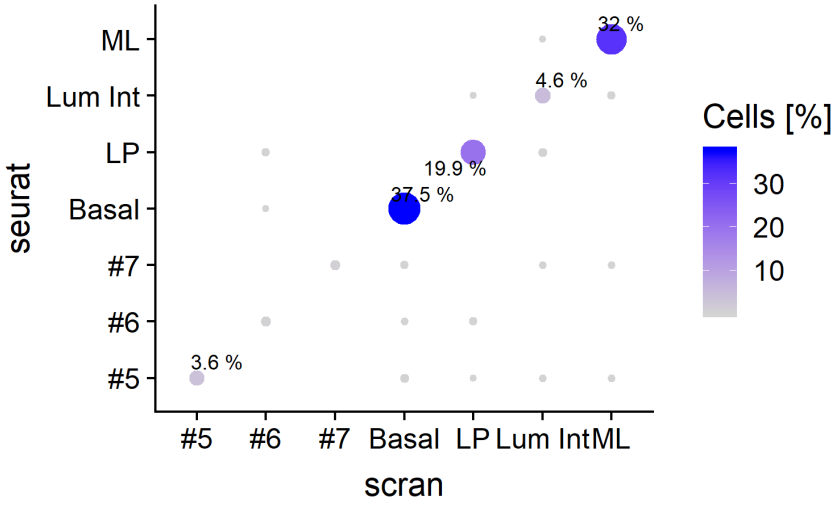


FIGURE C.16: Number of cells, relative to the total number, classified in each cluster by scran and Seurat for Ds#2/P7.

C.1.3.1 Rare clusters

The rare clusters identified in the four datasets were compared based on the criteria extracted from the reference works, which are described below.

- **Cont. (a):** in [Bach et al., 2017] some clusters are considered composed of contamination cells and removed from the analysis based on the expression of the following markers: 1) *Cd52*, *Cd74* and *Cd72* (Immune); 2) *Col3a1*, *Col5a1*, *Col6a1* and *Fn1* (Fibroblast); 3) *Eng*, *S1pr1* and *Emcn* (Endothelial); and 4) *Pdgfrb*, *Cspg4*, *Anpep* and *Des* (Pericyte).
- **Cont. (b):** in [Giraddi et al., 2018] instead, cells are classified as contaminating stroma based on the high expression of *Vim* and low expression of *Epcam* and other non-epithelial markers as: *Col3a1*, *Fap*, *Cd34*, *Cd83*, *Cd86*, *B2m* and *Itgax*.
- **Basal MaSC:** In [Sun et al., 2018], a small cluster of basal cells is declared as bipotent MaSC. For this cluster, key genes are: *Cdh5*, *Procr*, *Cd93*, *Cldn5*, *Flt1*, *Igfbp7* and *Inpp5d*.
- **Other:** Expression level for additional known genes were also checked (based on discussion with Dr Elias): *Tspan8* [Fu et al., 2017], *Lgr5* [Rios et al., 2016], *Sox2* and *Trp63* [Forster et al., 2014].

Heatmaps of the absolute expression level of the above genes are shown from Figure C.17 to Figure C.20. It is noticed that the same genes are represented in the four figures in the same order to allow comparison and, if some particular gene is missing in a figure, it is just because it is not expressed in any cell of the database. Results of the analysis of these figures are summarised in Table C.7. A high degree of similarity is noticed between some clusters: Ds#1-R2, Ds#2-R1c and Ds#4-R3 (hereafter indicated as Group A); and Ds#1-R1, Ds#2-R1a, Ds#3-R1 and Ds#4-R2/R4 (Group B). Notably, clusters in Group B shows both endothelial and MaSC markers. More generally, Cont. (a) and (b) criteria agree quite well, and based on these criteria, some rare clusters from Ds#2 and Ds#3 should be excluded from the analysis. On the other hand, if the Basal MaSC criterion is considered, cells from Ds#1 and Ds#4 should be included. To further study these discrepancies, the expression of a few known genes was checked. Some expression of the basal marker *Trp63* is found only in Ds#1 B*, which can be considered a sub-cluster of the main basal one. This cluster also expresses some *Tspan8*, which is one marker for MaSC, as well as Ds#2-R1d and Ds#2-R6. Cells in Ds#3-R1 express some *Lgr5*, a progenitor marker. In any case, these genes are not particularly highly expressed in these clusters, so that no conclusion can be drawn from them.

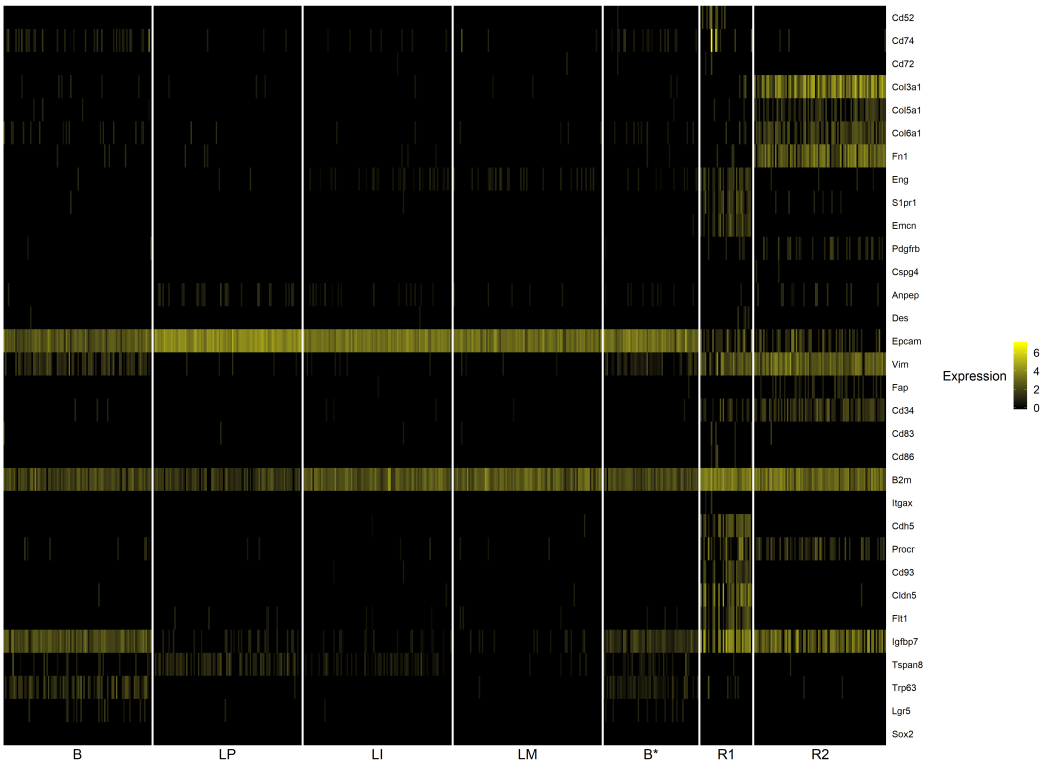


FIGURE C.17: Heatmap of the absolute expression of reference key genes for the rare clusters, Ds#1.

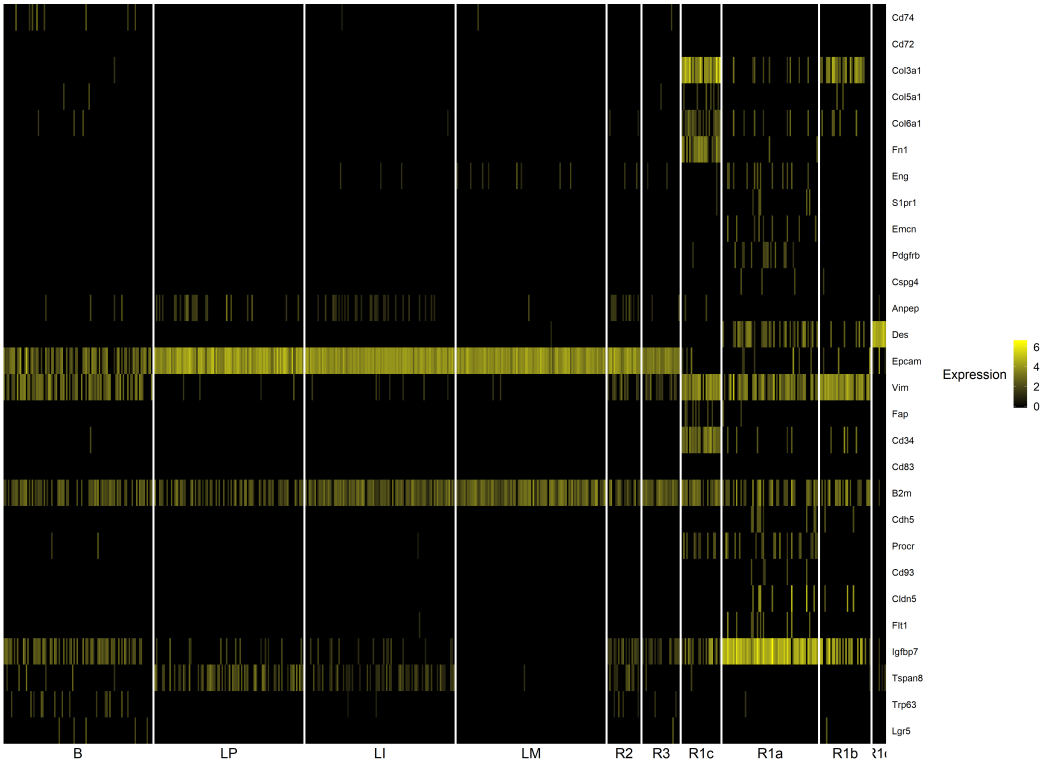


FIGURE C.18: Heatmap of the absolute expression of reference key genes for the rare clusters, Ds#2.

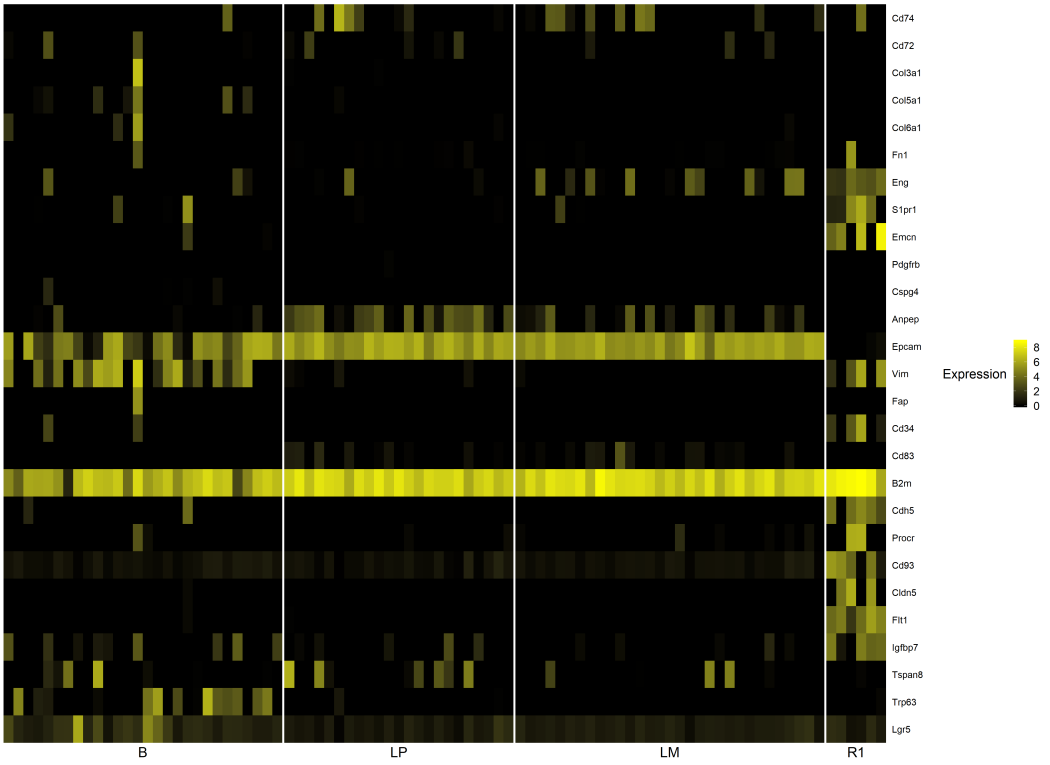


FIGURE C.19: Heatmap of the absolute expression of reference key genes for the rare clusters, Ds#3.

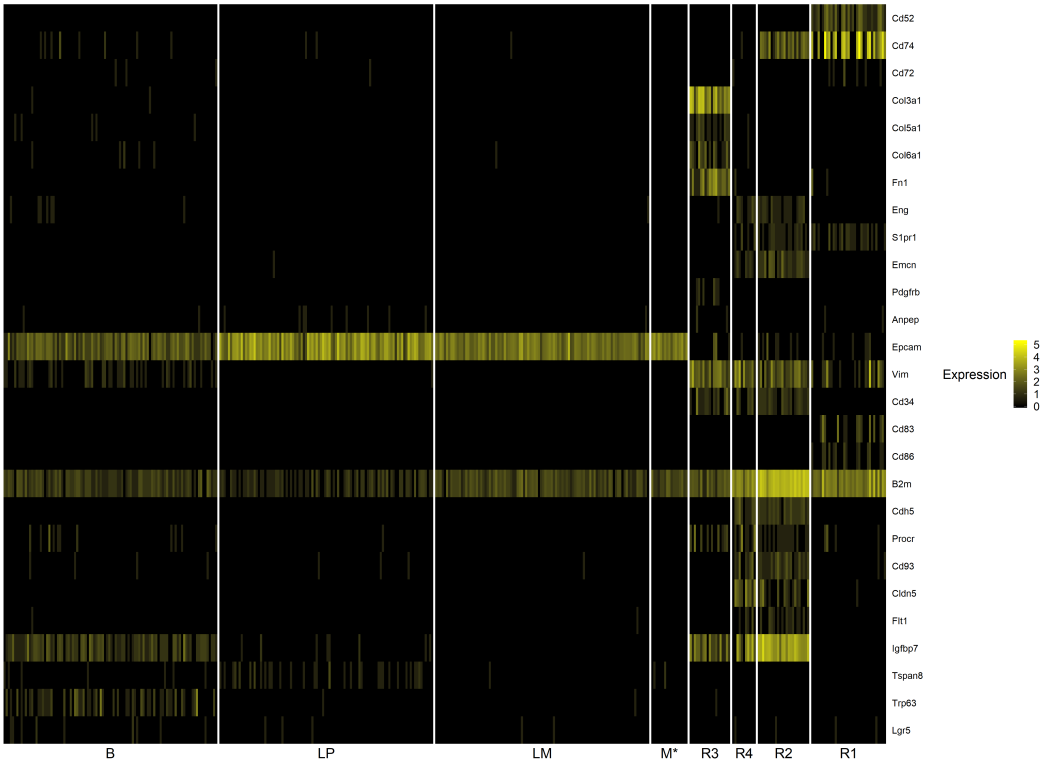


FIGURE C.20: Heatmap of the absolute expression of reference key genes for the rare clusters, Ds#4.

Cluster	Size [%]	Cont. (a)	Cont. (b)	Basal MaSC	Other
Ds#1					
B*	1.5				Trp63/Tspan8
R1	0.8	E	✓	✓	
R2	2.1	F	✓		
Ds#2					
R1a	1.9		✓	✓?	
R1b	1.0		✓		
R1c	0.8	F	✓		
R1d	0.3				Tspan8
R2	0.7				Tspan8
R3	0.8				
Ds#3					
R1	6.8	E	✓	✓	Lgr5
Ds#4					
M*	0.9				
R1	1.8	I?	✓		
R2/R4	1.2	E	✓	✓	
R3	1.0	F	✓		

TABLE C.7: Criteria met by rare clusters in each database; grey rows highlight clusters excluded in the corresponding published work.

To better characterise Group A and B clusters and the remaining rare clusters, an analysis of the DE genes was performed. Heatmaps of the relative expression of the identified genes are shown from Figure C.21 to Figure C.24.

- Group A: 50% of the top DE genes characterising this group is found in at least two datasets, and 20% is common in the three datasets involved (all except Ds#2 where no cells of this type are found). Among these genes, there are *Col3a1* and *Fn1* which are mentioned in both Cont (a) and Cont (b) criteria. Based on this purely biological criterion, the clusters in this group likely represent contaminating cells, and Ds#2-R1c should be removed from the analysis.
- Group B: 20% of the top DE genes is found in at least two datasets, and 4% is common in the four datasets. Among these genes there are: *Emcn*, *Vim*, *Cldn5* and *Igfbp7*, genes involved in three of the criteria above cited (*Epcam* is among the least DE genes in this group). Thus, based on this observation and considering the remaining DE genes, a clear conclusion cannot be drawn regarding this cluster.
- Ds#1-B*: DE genes shown in the heatmap (bottom part of Figure C.21) are the top 20 DE genes expressed in this cluster when compared to the basal cluster. Among the higher DE genes, there are *Krt18* and *Krt19*, which are specific luminal genes. Among the lowest expressed genes, yet expressed if looking at the absolute value, there are typical basal genes such as *Acta2* and *Krt15*. This

finding suggests that this cluster is potentially a mixed-lineage cluster. Another possibility is that it is a cluster containing doublets, and further analyses are required to confirm this hypothesis. It is noted that from the analysis in [Bach et al., 2017] cluster of doublets are not found in this sample but also no mixed-lineage cells.

- Ds#-R2/3: DE genes shown in the heatmap (bottom part of Figure C.22) are the top 20 DE genes expressed in these clusters when compared respectively to the LP and LM ones. Among the highly DE genes, there are basal genes as *Acta2* and *Krt14*, while the lowest DE genes are expressed just at a slightly lower level than the corresponding luminal cluster. Again, these clusters can be associated with mixed-lineage or doublets clusters. However, it is noted that in the related work, [Pal et al., 2017], mixed-lineage cells are found, but only when analysing C1 data. These clusters instead are defined as possible regulatory states and remain undetermined.

- Analogies between Ds#1-B* and Ds#2-R2/3 in terms of expression levels are evident. The analysis of the puberty sample in Ds#2 (reported in Appendix C.1.2.2) results in the two clusters R2 and R3 much closer to the basal cluster, as found in Ds#1. This is remarkably interesting since the mice age in Ds#1 is eight weeks, a value between puberty and the adult samples in Ds#2 (respectively of five and ten weeks). If the existence of this mixed-lineage cluster is confirmed, then the shift towards the luminal phenotype in adulthood would occur after the age of eight weeks.

- Ds#4-M*: cells in this cluster, when compared to the mature luminal ones, are characterised by a higher expression level in genes such as *Wfdc18* and *Csn3*. These genes in the same work are associated with the LP cluster. Thus, M* can be considered the LI cluster that is missing in this database. The size of this cluster is relatively small (0.9%) compared to the LI clusters in Ds#1 and Ds#2 (respectively 11.3% and 4.8%).

C.1.3.2 Main clusters

After the cancellation of the scRNA-seq experiment, we decided, together with our biology collaborator Dr Elias, to extract quantitative information about cell identity from the lineage tracing imaging. The idea was to add antibodies that highlight cells with colour by reacting with specific proteins. Associating a specific protein with a cell identity allows identifying different identities in clone imaging (selecting antibodies that produce different colours). Given that there is not a unique, commonly recognised marker for each cell identity (see Section 1.4.1), we analysed the four databases to identify genes that characterise each cell cluster and that are common in

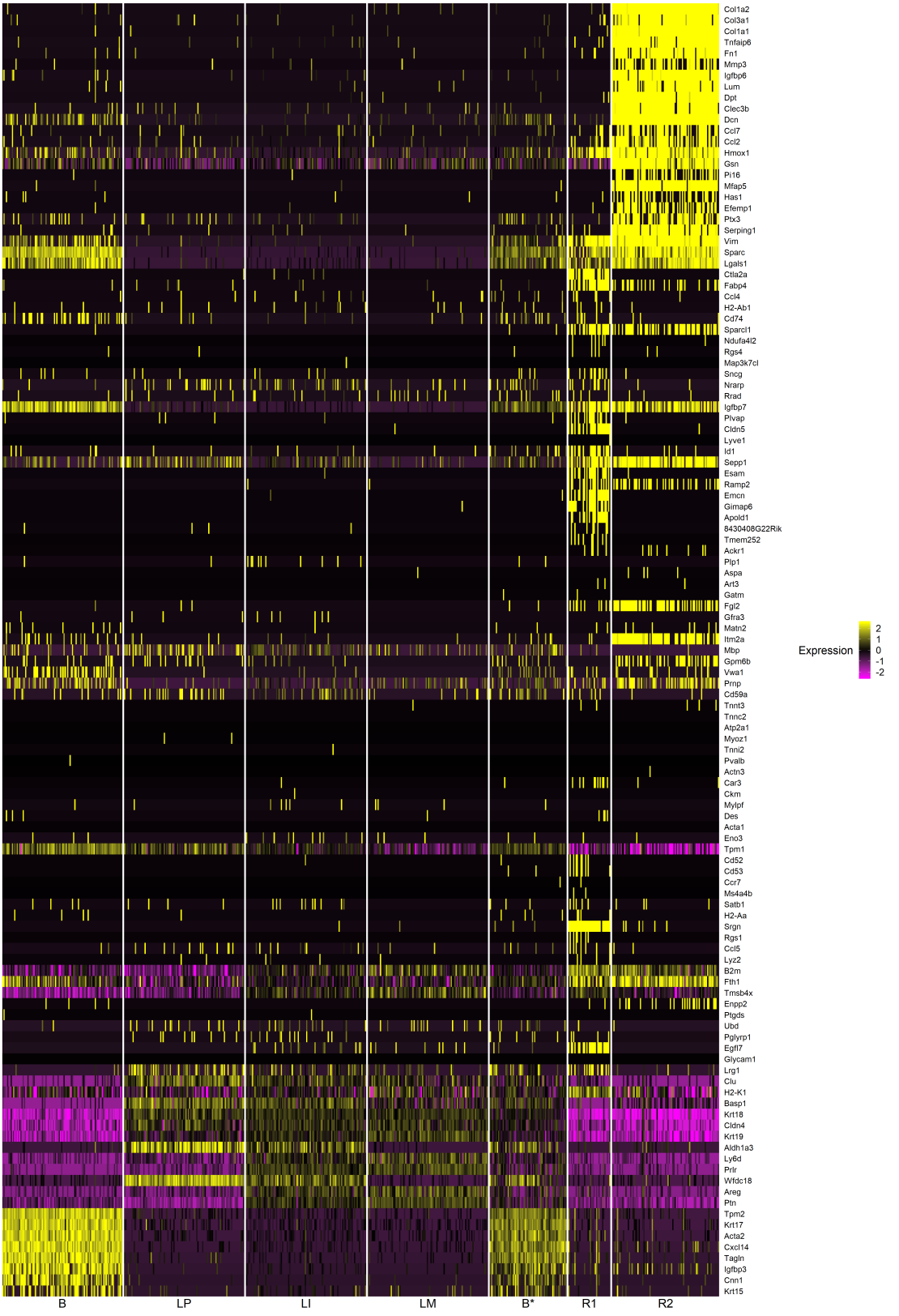


FIGURE C.21: Heatmap of the relative expression of top DE genes, sampling of max 100 cells per cluster, Ds#1.

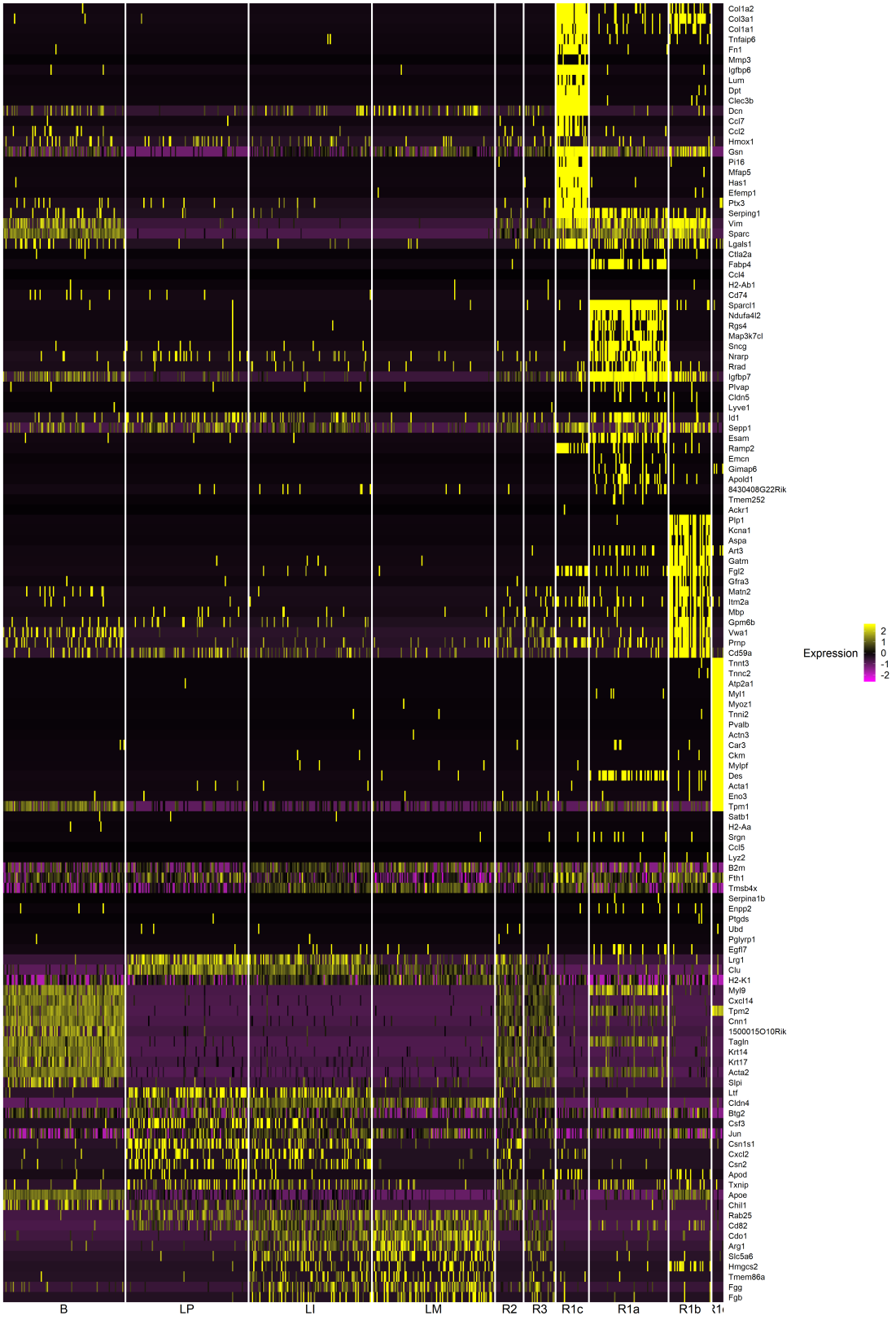


FIGURE C.22: Heatmap of the relative expression of top DE genes, sampling of max 100 cells per cluster, Ds#2.

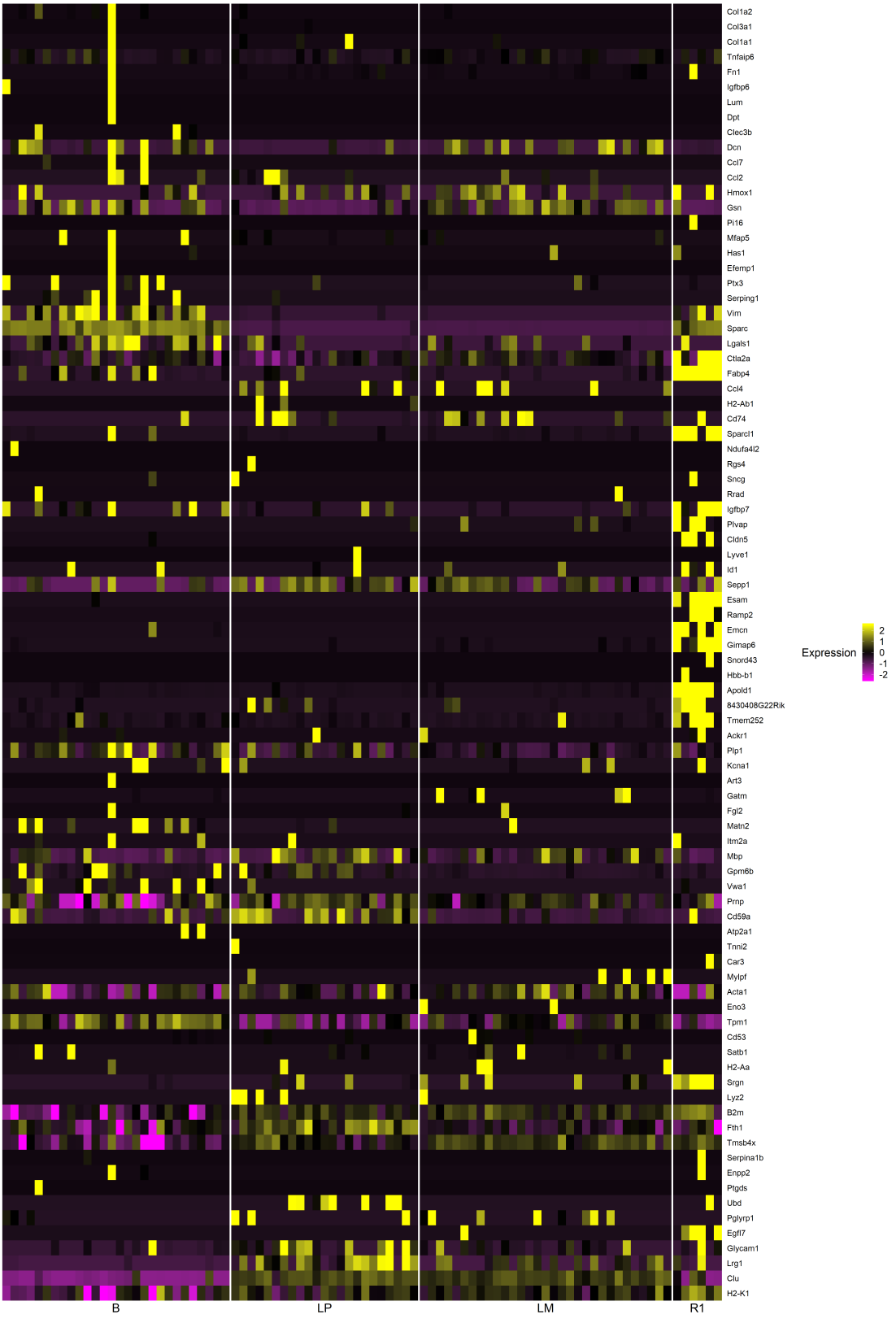


FIGURE C.23: Heatmap of the relative expression of top DE genes, sampling of max 100 cells per cluster, Ds#3.

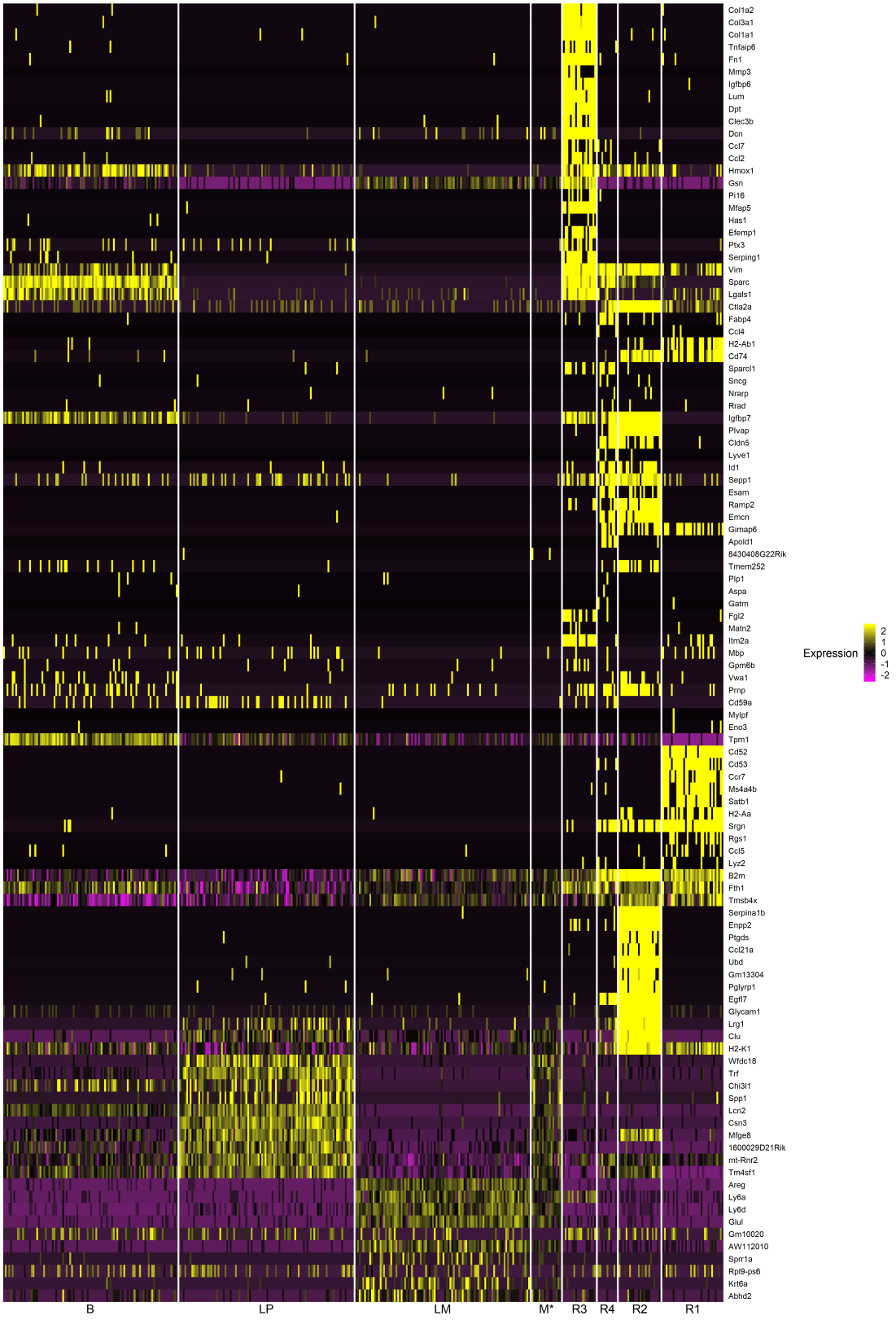


FIGURE C.24: Heatmap of the relative expression of top DE genes, sampling of max 100 cells per cluster, Ds#4.

the four databases. Considering that the order of magnitude of long-term clone size is less than 100 cells, the focus is done on the main clusters only, whose size is above 10%, as shown in Table 5.1. Luminal intermediate cells are also of interest and, therefore, this cluster is considered too. Importantly, we remark that since proteins are detected with the use of specific antibodies, this methodology relies on a strong assumption, that is, mRNA in a cell is a proxy of protein level [Wagner et al., 2016], an assumption that is valid in several cases. Also, other aspects were taken into account in the final choice of the markers (e.g. availability of the antibodies, known marker).

Considering first the LI cluster, its relationship with the LP and LM ones was studied. This assessment led to the identification of a clear correlation between these three clusters, which is shown in Figure C.25 in terms of average fold-change of the DE genes expression level. This plot indicates that if expression increases from LP to LI (positive x-axis), it decreases from LI to LM (negative y-axis), and in the opposite order. Only a few genes are in the first and third quadrant, showing different behaviour. They mainly come from Ds#4 in which the luminal intermediate cluster is relatively small (less than 1%) and, therefore, in some way, it is less representative than the other datasets. In any case, among these genes, there are *Sox4* and *Jund*, two genes mentioned in [Pal et al., 2017] for which a 20% level of expression in LI is higher than in LP and LM. From these considerations and a careful analysis of the heatmaps of the top DE genes in this cluster (not shown here), it is concluded that it is not possible to identify unique markers that characterise the LI cells (and only this). Instead, at least two compatible markers (for example, one colouring the nucleus, the other the cell surface, and proper choice of the colours), one related to LP and one to LM, have to be used.

At this point, three clusters remain that are the basal, luminal progenitor and luminal mature. The DE genes were analysed in each database, and the top differentially expressed genes for each identity were compared among the databases. The number of genes shared in two, three or all four datasets[3] is reported in Table C.8. The list of genes shared in the largest number of databases, i.e. four for the basal cluster and three for luminal ones, are listed in Table 5.2; heatmaps showing the relative expression level of such genes are shown from Figure C.26 to Figure C.29. These heatmaps are based on the same list of genes presented in the same order to allow a direct comparison between the datasets. Notably, among these genes there are known genes such as: *Krt14* and *Acta2* for basal cells, *Csn3* for luminal progenitors and *Areg*, *Cited1* and *Prlr* for the luminal mature cells.

[3]For each cluster (B, LP and LM), we extracted the top 20 highly expressed relative to the other two clusters. We then counted how many genes are common in two, three and all four datasets.

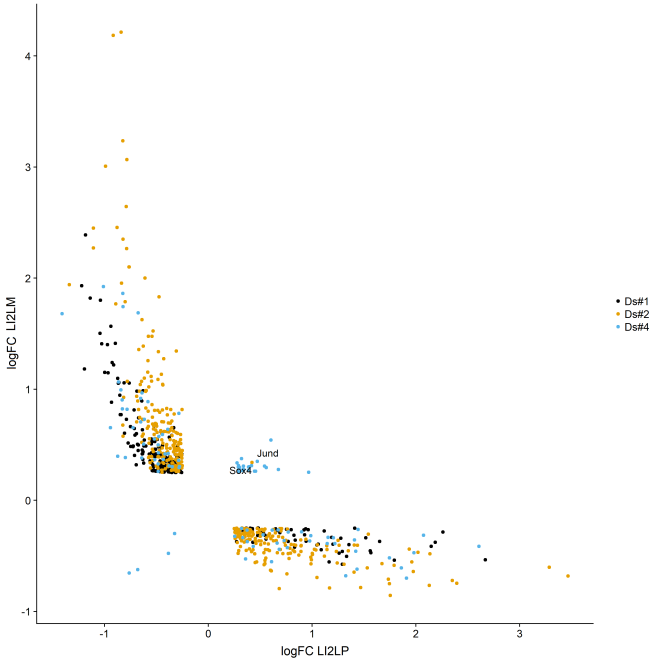


FIGURE C.25: Average fold-change in logarithmic scale of the expression of each gene between Luminal Intermediate and Luminal Progenitor clusters (LI2LP) and between Luminal intermediate and Luminal Mature clusters (LI2LM). Each point corresponds to a gene, which is coloured according to the dataset (Luminal Intermediate cell are not present in Ds#3).

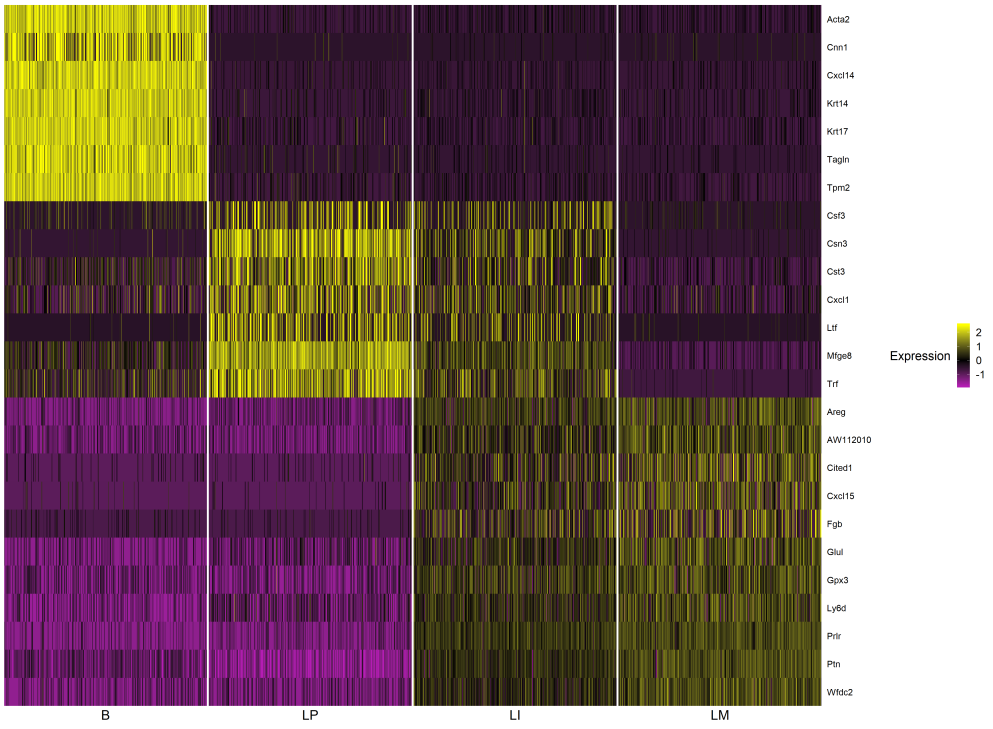


FIGURE C.26: Heatmap of the top shared DE genes listed in Table 5.2, Ds#1.

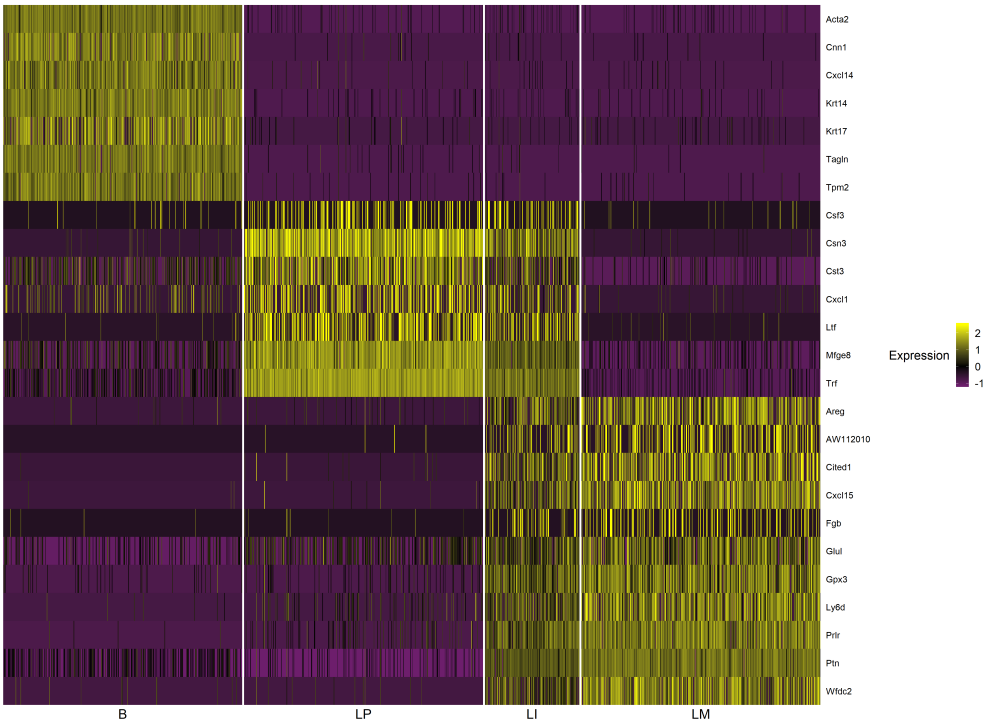


FIGURE C.27: Heatmap of the top shared DE genes listed in Table 5.2, Ds#2.

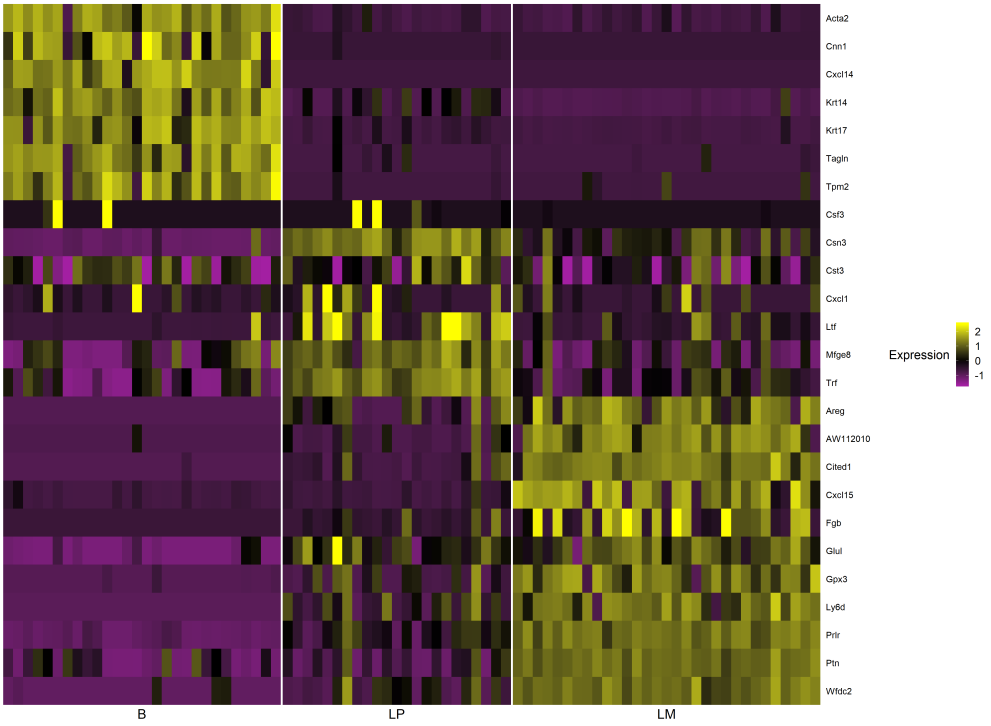


FIGURE C.28: Heatmap of the top shared DE genes listed in Table 5.2, Ds#3.

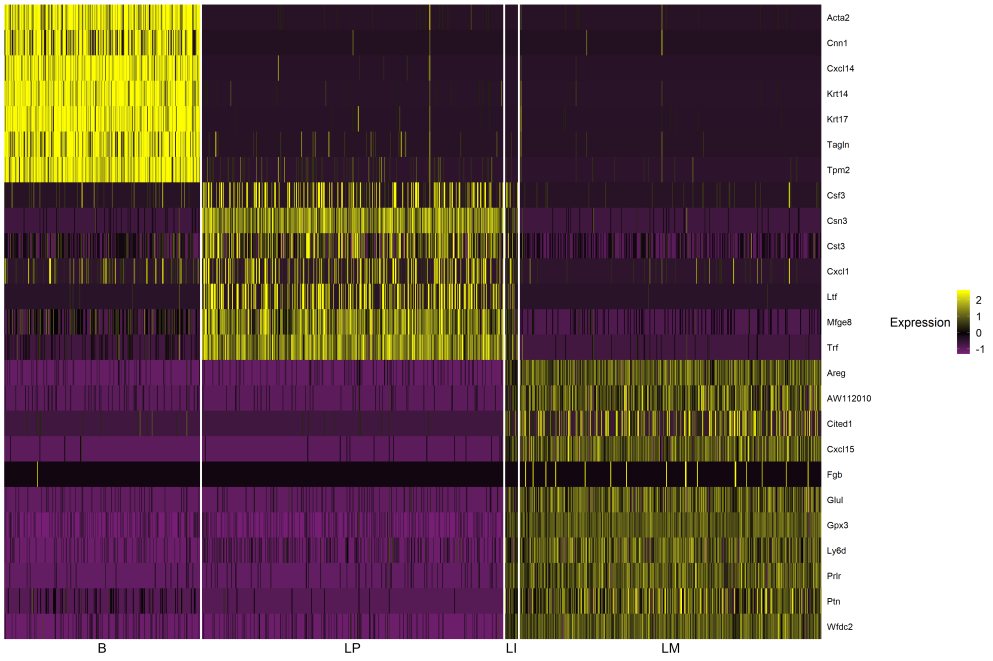


FIGURE C.29: Heatmap of the top shared DE genes listed in Table 5.2, Ds#4.

Cell Identity	Shared genes		
	in 2 Dss	in 3 Dss	in 4 Dss
Basal	19	12	7
Luminal Progenitor	18	7	0
Luminal Mature	24	11	0

TABLE C.8: Number of highly expressed genes in each cluster that are shared in two, three and four datasets.

C.2 Cell fate model parameter fitting

C.2.1 Optimisation runs

The optimisation problem described in Section 5.4 is very complex given the high-dimensional search space and the noisy objective function with potentially multiple local and global minima. Also, it is worth recalling that the stochastic simulations are computationally expensive for a sufficiently accurate clonal statistic estimation. For instance, for the optimal fitting **H.1**, the variability of the objective function in different runs is shown as function of the number of simulated clones, N_c, in Figure C.30 (left). The corresponding estimation of the execution time, Figure C.30 (right), is based on a Matlab implementation run on a laptop[4].

Given that, we follow a two-step approach. First, we simulate a relatively low number of clones, $N_c = 2 \times 10^4$ to 4×10^4, during the optimisation process and store all the optimisation results. Successively, for those cases in which the objective function is

[4]Processor: Intel(R) Core(TM) i7-10510U CPU @ 1.80 GHz 2.30 GHz; RAM 8.00 GB

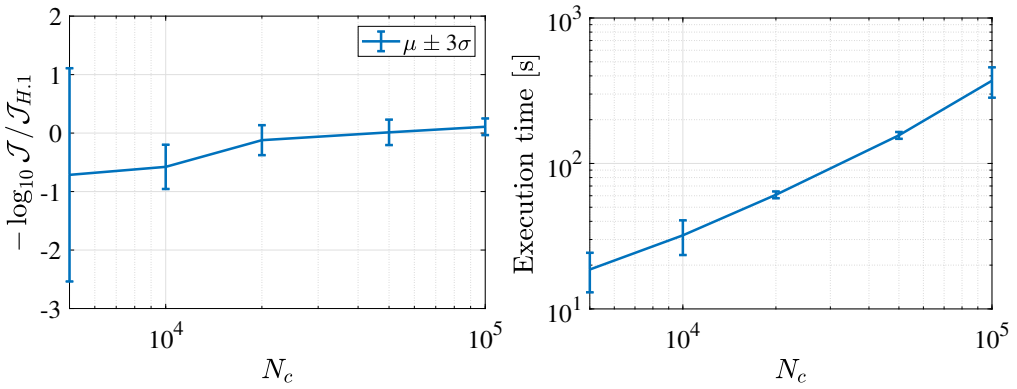


FIGURE C.30: Variability in the objective function (left) and corresponding execution time (right) as a function of the number of simulated clones N_c. Data refers to 20 independent runs for the **H.1** optimal fitting ($\mathcal{J}_{H.1}$ corresponds to objective function value reported in Table 5.4).

below a specific value, we evaluate again the objective function based on simulations with a more significant number of clones, $N_c = 10^5$. This choice was made primarily to speed up the optimisation process and avoid out-of-memory issues associated with using $N_c = 10^5$. Hence, considering that this approach does not give highly accurate results, we consider valid approximations of the optimal solution of the MAP optimisation problem, the cases where the objective function is within a certain threshold from a reference value. These approximations are called optimal fittings. Notably, the threshold was manually chosen by inspection of several fittings. A 5% likelihood ratio is not applicable here since the level of the simulation's noise in the optimisation is high due to the limited number of simulated clones. Thus, these optimal fittings are not optimal in a strict sense.

Additionally, a trade-off between efficiency and accuracy of the optimiser led us to choose, among the several optimisation methods available, the Bayesian and the surrogate optimisation methods (Matlab *Bayesopt* and *surrogateopt* functions). Importantly, we remark that we consider the optimisation solver as a black box, but we briefly review the main features of these methods to justify our choice.

- The Bayesian Optimisation (BO) is extremely efficient in converging to sufficiently accurate fittings for our purposes, considering the objective function's noise level. For instance, our results are based on runs with 500-750 function evaluations. In contrast, the genetic algorithm (Matlab *ga* function), using default settings, would require 200 function evaluations in each iteration for a maximum of 1000 iterations[5]. Additionally, the implementation features parallel computing.

[5]Computation might stop before reaching the maximum iteration number if convergence criteria are met, yet the number of function evaluations is likely to be very large.

- The surrogate optimisation (SO) is another efficient method, which provides accurate fittings at the price of converging, in some cases, to local minima. Even this optimiser features parallel computing.

The results presented in our work are based on runs of BO and SO in the whole search space Θ, which is defined in Table C.9. We also run SO to refine some good fittings locally (SOL); in this case, the search space is set to be $\pm 5\%$ of the global variability, Θ, around the selected case. Overall, the optimisation runs are summarised in Table C.10. All the runs are based on default optimisation settings, except the number of function evaluations. The process of evaluating the objective function with $N_c = 10^5$, includes, for each optimisation run, the global optimum fitting and several sufficiently good fittings found during the optimisation process. The number of these cases, chosen manually, is specified in the last column of Table C.10.

The objective function for all the optimisation runs is shown as function of the ratio of asymmetric division r in Figure C.31. The presented values are based on $N_c = 10^5$ clones. These figures correspond to Figure 5.12 and Figure 5.15 (see Section 5.4.2), but here each point is coloured according to the optimisation method.

Model	Θ											
	θ	λ_{BS}	r	Δ	p_a	p_s	η_{LP}	$\tilde{\omega}_{LP2LM}$	η_{LM}	$\tilde{\gamma}_{LM}$	η_{BM}	$\tilde{\gamma}_{BM}$
NH	θ_{min}	0.04	0	-1	0	0	0	0.1	0	0.1	0	0.1
	θ_{max}	4.11	0.5	1	1	1	3	10	3	10	3	10
H/H^+	θ_{min}	0.04	0	0	0	0	0	0.1	0	0.1	0	0.1
	θ_{max}	4.11	0.5	0	1	1	1	10	1	10	1	10

TABLE C.9: Optimisation search space Θ; all the variables are here dimensionless, with the exception of λ_{BS} which is expressed in [w^{-1}].

Model	Method	Run	Function Evaluation	N_c	Nr. of fitting
NH	BO	1	750	$4\ 10^4$	10
	SOL	2	200	$4\ 10^4$	200
H	BO	2	300	$2\ 10^4$	236
	BO	1	500	$2\ 10^4$	100
	BO	2	750	$2\ 10^4$	45
	SO	2	500	$2\ 10^4$	194
	SOL	3	300	$2\ 10^4$	60
H^+	BO	1	500	$2\ 10^4$	217
	SOL	5	300	$2\ 10^4$	512

TABLE C.10: Summary of the optimisation runs and their settings. The last column indicates the number of fitting selected for being evaluated again based on $N_c = 10^5$.

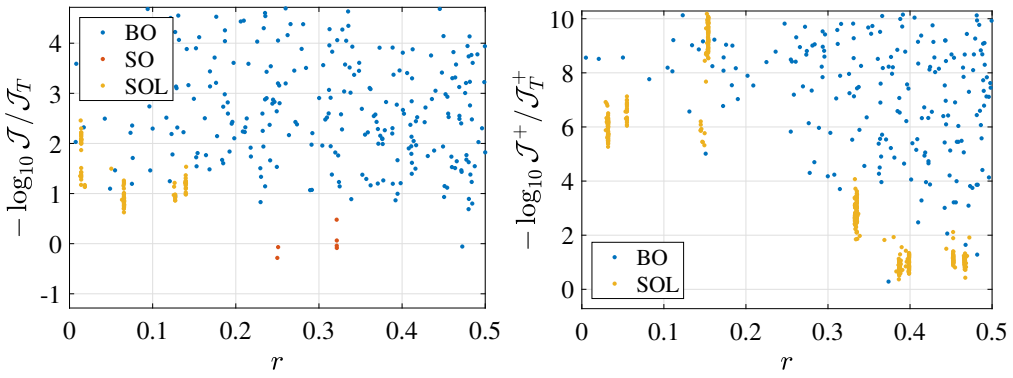


FIGURE C.31: Value of the objective function, relative to that of the true model, as a function of the ratio of asymmetric divisions for the **H** (left) and the $\mathbf{H}^+$ (right) fittings. Points are coloured according to the optimisation method: Bayesian optimisation (BO), surrogate optimisation (SO) and local refinement based on Surrogate optimisation (SOL).

C.2.2 Additional solutions

Two additional fittings are reported in this section. In particular, in Section 5.4.2.1, a non-homeostatic model, **NH.1**, is shown to be a good fitting of the data despite presenting a growing dynamics. Here, we compare this case and another one, **NH.2**, characterised by a vanishing dynamic yet fitting the data sufficiently well. In Section 5.4.2.1, instead, we showed that the self-renewing strategy for the test case analysed is the generalised population asymmetry since the ratio of asymmetric division r is above 0.33 in all the optimal fittings. As a further confirmation of this feature of the cell fate model, we compare here the global optimum fitting, $\mathbf{H.1}^+$, with another model, $\mathbf{H.2}^+$, characterised by mainly asymmetric divisions, i.e. $r \approx 0$, and the minimum objective function in this range of r. For these cases, the model parameters and objective function are reported in Table C.11, the clonal statistics is shown in Figure C.32 and the cluster size evolution in Figure C.33. For comparison, also the values and the profiles of **NH.1** and $\mathbf{H.1}^+$ are shown.

Concerning the non-homeostatic models, although **NH.2** is characterised by an objective function higher than **NH.1**, it fits the data very well. Crucially, the two cases correspond to very different cell fate models, one characterised by a slightly positive homeostatic imbalance and the other by a negative value (the remaining parameters are compliant with a homeostatic model, i.e. $\eta_x < 1$ for $x =$ BM, LP and LM). Consistently, the tissue dynamics grows in **NH.1** and shrinks in **NH.2**, as shown in Figure C.34.

Focusing now on $\mathbf{H.2}^+$, which is not included among the optimal fittings for model $\mathbf{H}^+$, this is representative of a cell fate model with almost completely asymmetric divisions in the self-renewing cell type. In this case, we observe that there is a marked difference in the objective function, $\mathcal{J}^+$. Notably, a significant contribution comes from the fitting of the long-term clonal data, f_{CD-III}, meaning that this last clonal data

point, combined with the other data, is the one that constrains the model to have also symmetric divisions.

Fitting	**NH.1**	**NH.2**	**H.1**$^+$	**H.2**$^+$
Cell Fate Model Parameters				
λ_{BS} [w^{-1}]	1.589	1.654	2.033	1.227
λ_{LP} [w^{-1}]	7.350	8.861	1.716	2.886
ω_{LPLM} [w^{-1}]	7.848	9.938	2.297	3.317
λ_{LM} [w^{-1}]	0.515	14.250	0.172	9.252
γ_{LM} [w^{-1}]	1.904	16.389	0.592	9.885
λ_{BM} [w^{-1}]	1.646	4.225	2.562	0.444
γ_{BM} [w^{-1}]	7.694	14.561	17.922	3.211
p_{BSBS}	0.371	0.213	0.374	0.032
p_{BSBM}	0.101	0.130	0.211	0.255
p_{BSLP}	0.189	0.246	0.042	0.681
p_{BMBM}	0.127	0.125	0.140	0.002
p_{LPLP}	0.212	0.286	0.234	0.030
Δ	0.046	-0.317	0.000	0.000
r	0.355	0.312	0.374	0.032
Objective Function				
f_{CD-I}	0.358	1.612	-0.139	1.299
f_{CD-II}	0.329	-0.087	0.186	0.434
$f_{scRNA-I}$	-0.512	-0.541	-0.553	0.238
$\mathcal{J}$	0.175	0.984	-0.506	1.971
f_{CD-III}			0.456	3.244
f_{TV}			0.335	0.053
$\mathcal{J}^+$			0.285	5.268

TABLE C.11: Summary of some additional illustrative fittings in terms of cell fate model parameters and objective function. The objective function $\mathcal{J}$ and each contribution are defined in Equation (5.14); the final rows correspond to $\mathcal{J}^+$, defined in Equation (5.20). Concerning the objective function, for each row x, where $x = f_{CD-I}, f_{CD-II}, ..., \mathcal{J}^+$, values reported are $-\log_{10}(x/x_T)$, in which x_T is the value of x corresponding to the true model. Hence, positive (negative) values mean a fitting that is worse (better) than the true model. The optimal fitting **NH.1** and **H.1**$^+$ are the same reported in Table 5.4.

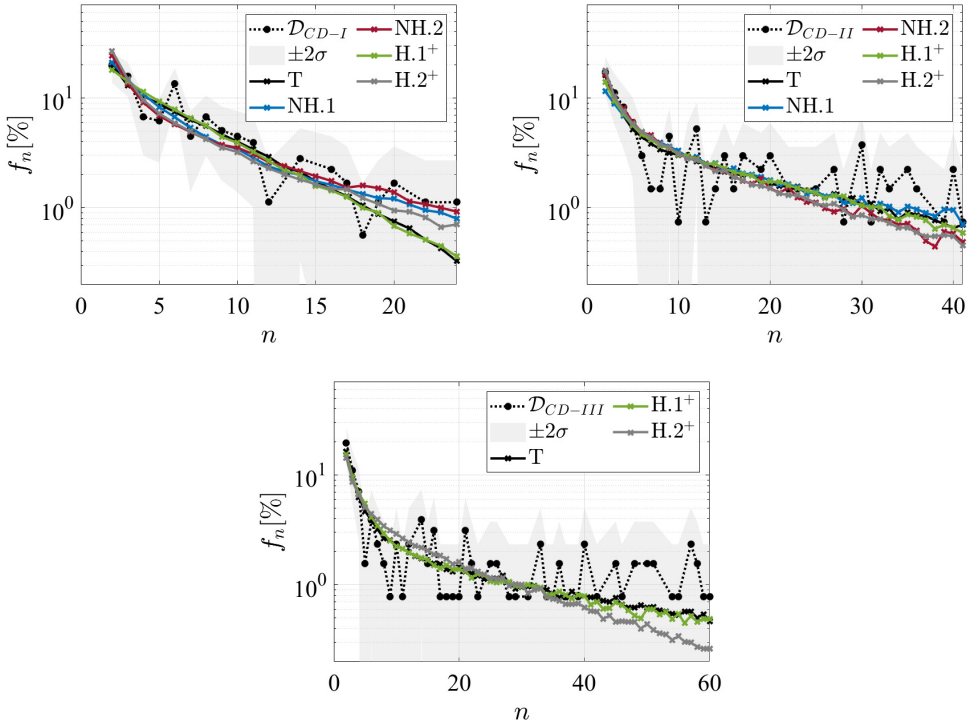


FIGURE C.32: Profiles of the clone size distribution for some illustrative fittings (see model parameters in Table C.11) compared to clonal statistics data, $\mathcal{D}_{CD-I}$ (top-left), $\mathcal{D}_{CD-II}$ (top-right) and $\mathcal{D}_{CD-III}$ (bottom). Data 2σ variability and the clonal statistics for the true model, labelled as T, are also shown. Cases **NH.1** and **H.1**$^+$ are the same reported in Section 5.4.2.

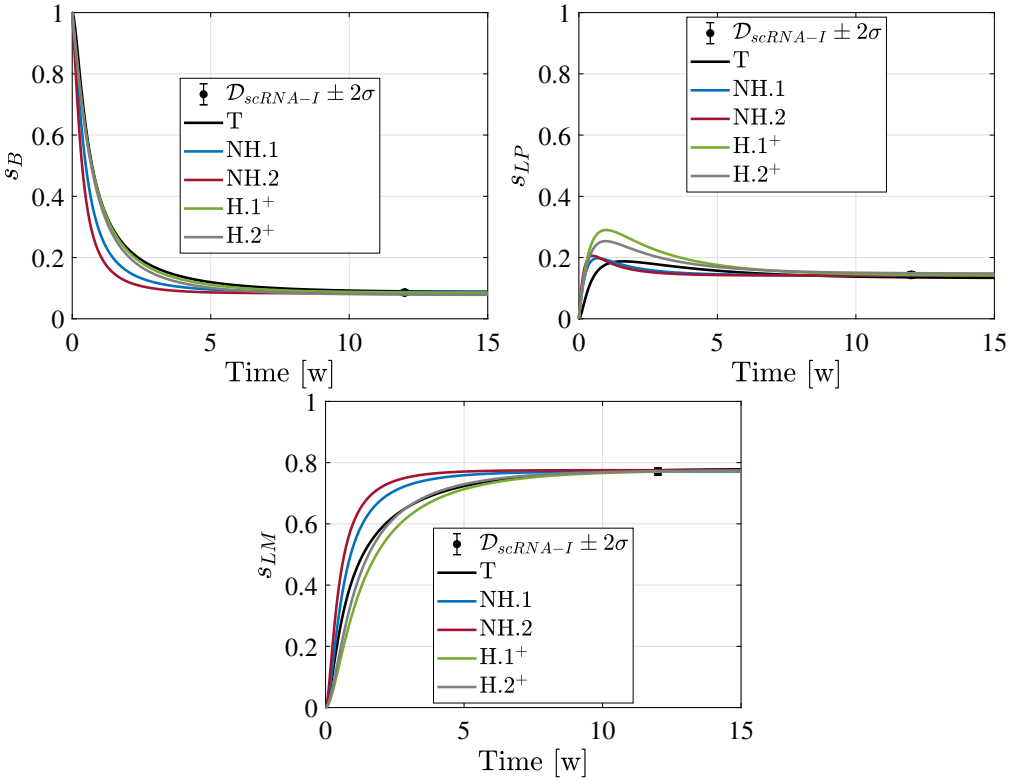


FIGURE C.33: Time evolution of the clusters size for some illustrative fittings (see model parameters in Table C.11), compared to scRNA-seq data $\mathcal{D}_{scRNA-I}$. Clusters correspond to basal (top-left), luminal progenitor (top-right) and luminal mature (bottom). Data 2σ variability and the clonal statistics for the true model, labelled as T, are also shown. Cases **NH.1** and **H.1**$^+$ are the same reported in Section 5.4.2.

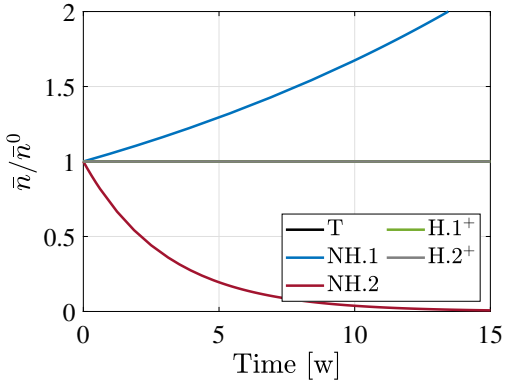


FIGURE C.34: Mean total cell number evolution as a function of time, based on the integration of the system of ODEs, Equation (2.6), for some illustrative fitting (see model parameters in Table C.11). The initial condition, $\bar{\boldsymbol{n}}_0$, is proportional to the dominant eigenvalue, as representative of the tissue dynamics. Cases **NH.1** and **H.1**$^+$ are the same reported in Section 5.4.2.

References

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*. United States Department of Commerce, National Bureau of Standards, 1972. ISBN 0486612724. doi: 10.1115/1.3625776.

Luis U. Aguilera, Christoph Zimmer, and Ursula Kummer. A new efficient approach to fit stochastic models on the basis of high-throughput experimental data using a model of IRF7 gene expression as case study. *BMC Systems Biology*, 11(1), 2017. ISSN 17520509. doi: 10.1186/s12918-017-0406-4.

Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the cell*. Garland Science, 2015. ISBN 1317563751, 9781317563754. doi: 10.3390/ijms161226074.

Maria P. Alcolea, Philip Greulich, Agnieszka Wabik, Julia Frede, Benjamin D. Simons, and Philip H. Jones. Differentiation imbalance in single oesophageal progenitor cells causes clonal immortalization and field change. *Nature Cell Biology*, 16(6): 615–622, 2014. ISSN 14764679. doi: 10.1038/ncb2963.

Steven S. Andrews, Tuan Dinh, and Adam P. Arkin. Stochastic Models of Biological Processes. *Encyclopedia of Complexity and Systems Science*, pages 8730–8749, 2009. ISSN 0387758887. doi: 10.1007/978-0-387-30440-3_524. URL http://link.springer.com/10.1007/978-0-387-30440-3_524.

Tibor Antal and P. L. Krapivsky. Exact solution of a two-type branching process: Clone size distribution in cell division kinetics. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(7), 2010. ISSN 17425468. doi: 10.1088/1742-5468/2010/07/P07028.

Fumio Arai, Patrick S. Stumpf, Yoshiko M. Ikushima, Kentaro Hosokawa, Aline Roch, Matthias P. Lutolf, Toshio Suda, and Ben D. MacArthur. Machine Learning of Hematopoietic Stem Cell Divisions from Paired Daughter Cell Expression Profiles Reveals Effects of Aging on Self-Renewal. *Cell Systems*, 11(6):640–652, 2020. ISSN 24054720. doi: 10.1016/j.cels.2020.11.004. URL https://doi.org/10.1016/j.cels.2020.11.004.

Kenneth J. Arrow. *A "Dynamic" Proof of the Frobenius–Perron Theorem for Metzler Matrices*. Academic Press, 1989. ISBN 978-0-12-058470-3. doi: https://doi.org/10.1016/B978-0-12-058470-3.50009-4. URL https://www.sciencedirect.com/science/article/pii/B9780120584703500094.

Karl J. Åström and Richard M. Murray. *Feedback System, An Introduction for Scientists and Engineers*. Princeton University Press, 2009.

Karsten Bach, Sara Pensa, Marta Grzelak, James Hadfield, David J. Adams, John C. Marioni, and Walid T. Khaled. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-02001-5.

Ruth Baker. Stochastic Modelling of Biological Processes Lecture Notes, 2017.

Jørgen Bang-Jensen and Gregory Z. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer-Verlag New York, 2007. ISBN 9781852336110. doi: 10.1002/stvr.240. URL http://doi.wiley.com/10.1002/stvr.240.

Nick Barker, Meritxell Huch, Pekka Kujala, Marc van de Wetering, Hugo J. Snippert, Johan H. van Es, Toshiro Sato, Daniel E. Stange, Harry Begthel, Maaike van den Born, Esther Danenberg, Stieneke van den Brink, Jeroen Korving, Arie Abo, Peter J. Peters, Nick Wright, Richard Poulsom, and Hans Clevers. Lgr5+veStem Cells Drive Self-Renewal in the Stomach and Build Long-Lived Gastric Units In Vitro. *Cell Stem Cell*, 6(1):25–36, 2010. ISSN 19345909. doi: 10.1016/j.stem.2009.11.013. URL http://dx.doi.org/10.1016/j.stem.2009.11.013.

Jürgen Bauer, Friedrich A. Bahmer, Jürgen Wörl, Winfried Neuhuber, Gerold Schuler, and Manigé Fartasch. A strikingly constant ratio exists between Langerhans cells and other epidermal cells in human skin. A stereologic study using the optical disector method and the confocal laser scanning microscope. *Journal of Investigative Dermatology*, 116(2):313–318, 2001. ISSN 0022202X. doi: 10.1046/j.1523-1747.2001.01247.x.

Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9.

Cédric Blanpain and Benjamin D. Simons. Unravelling stem cell dynamics by lineage tracing. *Nature Reviews Molecular Cell Biology*, 14(8):489–502, 2013. ISSN 14710072. doi: 10.1038/nrm3625. URL http://dx.doi.org/10.1038/nrm3625.

Béla Bollobás. *Modern Graph Theory*. Graduate Texts in Mathematics 184. Springer-Verlag New York, New York, 1 edition, 1998. ISBN 978-0-387-98488-9,978-1-4612-0619-4.

George E. P. Box, George C. Tiao, and D. V. Gokhale. *Bayesian inference in statistical analysis*. John Wiley and Sons, 1992. ISBN 9781118033197. doi: 10.1002/9781118033197.

Maury Bramson and David Griffeath. Asymptotics for interacting particle systems on Zd. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 53(2):183–196, 1980. ISSN 1432-2064. doi: 10.1007/BF01013315. URL https://doi.org/10.1007/BF01013315.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 2018. ISSN 1087-0156. doi: 10.1038/nbt.4096. URL https://www.nature.com/articles/nbt.4096.

Yongjun Chu and David R. Corey. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–4, 2012. ISSN 2159-3345. doi: 10.1089/nat.2012.0367. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3426205&tool=pmcentrez&rendertype=abstract.

Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2008. doi: 10.1017/CBO9780511790485.

Elizabeth Clayton, David P. Doupé, Allon M. Klein, Douglas J. Winton, Benjamin D. Simons, and Philip H. Jones. A single type of progenitor cell maintains normal epidermis. *Nature*, 446(7132):185–189, 2007. ISSN 14764687. doi: 10.1038/nature05574.

Peter Clifford and Aidan Sudbury. A Model for Spatial Conflict. *Biometrika*, 60(3): 581–588, 5 1973. ISSN 00063444. doi: 10.2307/2335008. URL http://www.jstor.org/stable/2335008.

Bartomeu Colom and Philip H. Jones. Clonal analysis of stem cells in differentiation and disease. *Current Opinion in Cell Biology*, 43:14–21, 2016. ISSN 18790410. doi: 10.1016/j.ceb.2016.07.002. URL http://dx.doi.org/10.1016/j.ceb.2016.07.002.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844.

D.R. Cox. *Principles of statistical inference*, volume 404. Cambridge University Press, 2007. ISBN 9780521866736. doi: 10.1007/978-1-59745-530-5_4. URL http://www.ncbi.nlm.nih.gov/pubmed/18450045.

Felicity M. Davis, Bethan Lloyd-Lewis, Olivia B. Harris, Sarah Kozar, Douglas J. Winton, Leila Muresan, and Christine J. Watson. Single-cell lineage tracing in the

mammary gland reveals stochastic clonal dispersion of stem/progenitor cell progeny. *Nature Communications*, 7, 2016. ISSN 20411723. doi: 10.1038/ncomms13053.

Mousumi Debnath, GBKS Prasad, and Prakash Bisen. *Molecular Diagnostics: Promises and Possibilities*. Springer, 2010. ISBN 9789048132607. doi: 10.1007/978-90-481-3261-4_28.

David P. Doupé, Maria P. Alcolea, Amit Roshan, Gen Zhang, Allon M. Klein, Benjamin D. Simons, and Philip H. Jones. A single progenitor population switches behavior to maintain and repair esophageal epithelium. *Science*, 337(6098): 1091–1093, 2012. ISSN 10959203. doi: 10.1126/science.1218835.

George T. Eisenhoffer and Jody Rosenblatt. Bringing balance by force: Live cell extrusion controls epithelial cell numbers. *Trends in Cell Biology*, 23(4):185–192, 2013. ISSN 18793088. doi: 10.1016/j.tcb.2012.11.006.

George T. Eisenhoffer, Patrick D. Loftus, Masaaki Yoshigi, Hideo Otsuna, Chi-Bin Chien, Paul A. Morcos, and Jody Rosenblatt. Crowding induces live cell extrusion to maintain homeostatic cell numbers in epithelia. *Nature*, 484(7395):546–549, 2012. doi: doi:10.1038/nature10999.

Salah Elias, Marc A. Morgan, Elizabeth K. Bikoff, and Elizabeth J. Robertson. Long-lived unipotent Blimp1-positive luminal stem cells drive mammary gland organogenesis throughout adult life. *Nature Communications*, 8(1):1–11, 2017. ISSN 20411723. doi: 10.1038/s41467-017-01971-w. URL http://dx.doi.org/10.1038/s41467-017-01971-w.

Levent Ertöz, Michael Steinbach, Vipin Kumar, Levent Ertoz, Michael Steinbach, and Vipin Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Daniel Barbara and Chandrika Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining (SDM 2003)*, volume 112 of *Proceedings in Applied Mathematics*, pages 47–58. Society for Industrial and Applied Mathematics, 2003. doi: 10.1137/1.9781611972733.5. URL http://www.siam.org/meetings/sdm03/proceedings/sdm03_05.pdf.

Nicole Forster, Srinivas Vinod Saladi, Maaike van Bragt, Mary E. Sfondouris, Frank E. Jones, Zhe Li, and Leif W. Ellisen. Basal cell signaling by p63 controls luminal progenitor function and lactation via NRG1. *Developmental cell*, 28(2):147–160, 1 2014. ISSN 1878-1551. doi: 10.1016/j.devcel.2013.11.019. URL https://www.ncbi.nlm.nih.gov/pubmed/24412575https://www.ncbi.nlm.nih.gov/pmc/PMC3951056/.

Gene F Franklin, David J Powell, and Abbas Emami-Naeini. *Feedback Control of Dynamic Systems*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 4th edition, 2001.

Julia Frede, Philip Greulich, Tibor Nagy, Benjamin D. Simons, and Philip H. Jones. A single dividing cell population with imbalanced fate drives oesophageal tumour growth. *Nature Cell Biology*, 18(9):967–978, 2016. ISSN 14764679. doi: 10.1038/ncb3400.

Nai Yang Fu, Anne C. Rios, Bhupinder Pal, Charity W. Law, Paul Jamieson, Ruijie Liu, François Vaillant, Felicity C. Jackling, Kevin He Liu, Gordon K. Smyth, Geoffrey J. Lindeman, Matthew E. Ritchie, and Jane E. Visvader. Identification of quiescent and spatially restricted mammary stem cells that are hormone responsive. *Nature Cell Biology*, 19:164, 2 2017. URL https://doi.org/10.1038/ncb3471http://10.0.4.14/ncb3471.

R. Ganguly and I. K. Puri. Mathematical model for the cancer stem cell hypothesis. *Cell Proliferation*, 39(1):3–14, 2006. ISSN 0960-7722 (Print) 0960-7722 (Linking). doi: 10.1111/j.1365-2184.2006.00369.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16426418.

Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. doi: 10.1021/j100540a008. URL https://doi.org/10.1021/j100540a008.

Rajshekhar R. Giraddi, Chi Yeh Chung, Richard E. Heinz, Ozlen Balcioglu, Mark Novotny, Christy L. Trejo, Christopher Dravis, Berhane M. Hagos, Elnaz Mirzaei Mehrabad, Luo Wei Rodewald, Jae Y. Hwang, Cheng Fan, Roger Lasken, Katherine E. Varley, Charles M. Perou, Geoffrey M. Wahl, and Benjamin T. Spike. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Reports*, 24(6): 1653–1666, 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.07.025.

David E Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989. ISBN 0201157675.

Philip Greulich and Benjamin D. Simons. Dynamic heterogeneity as a strategy of stem cell self-renewal. *Proceedings of the National Academy of Sciences*, 113(27):7509–7514, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1602779113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1602779113.

Philip Greulich, Benjamin D. MacArthur, Cristina Parigini, and Rubén J. Sánchez-García. Stability and steady state of complex cooperative systems: A diakoptic approach. *Royal Society Open Science*, 6(12), 2019. ISSN 20545703. doi: 10.1098/rsos.191090.

Philip Greulich, Benjamin D. MacArthur, Cristina Parigini, and Rubén J. Sánchez-García. Universal principles of lineage architecture and stem cell identity

in renewing tissues. *Development (Cambridge)*, 148(11), 2021. ISSN 14779129. doi: 10.1242/DEV.194399.

Dominic Grün, Mauro J. Muraro, Jean Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik J. Jansen, Hans Clevers, Eelco J. P. de Koning, and Alexander van Oudenaarden. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19(2):266–277, 2016. ISSN 18759777. doi: 10.1016/j.stem.2016.05.010.

S A Gudipaty, J Lindblom, P D Loftus, M J Redd, K Edes, C F Davey, V Krishnegowda, and J Rosenblatt. Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature*, 543(7643):118–121, 3 2017. ISSN 0028-0836. doi: 10.1038/nature21407.

Patsy Haccou, Peter Jagers, and Vladmir Alekseevich Vatutin. *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, Cambridge, 2005. doi: 10.2277/0521832209. URL http://pure.iiasa.ac.at/7598/.

Ruth M Hall and Christina M Collis. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Molecular Microbiology*, 15(4):593–600, 1995. doi: https://doi.org/10.1111/j.1365-2958.1995.tb02368.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.1995.tb02368.x.

Sophie Hautphenne. Branching processes, 2015.

Sophie Hautphenne, Giang Nguyen, and Guy Latouche. Extinction probability of Branching Processes with Countably Infinitely Many Types. *Advances in Applied Probability*, 45(4):1068–1082, 10 2013. ISSN 00018678. URL http://www.jstor.org/stable/43563325.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 1985. ISBN 9780521548236. doi: 10.1017/cbo9780511810817.

Jamie L. Inman, Claire Robertson, Joni D. Mott, and Mina J. Bissell. Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development*, 142(6):1028–1042, 2015. ISSN 0950-1991. doi: 10.1242/dev.087643. URL http://dev.biologists.org/cgi/doi/10.1242/dev.087643.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

Matthew D. Johnston, Carina M. Edwards, Walter F. Bodmer, Philip K. Maini, and S Jonathan Chapman. Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proceedings of the National Academy of Sciences*,

104(10):4008–4013, 3 2007. ISSN 0027-8424. doi: 10.1073/pnas.0611179104. URL http://www.pnas.org/content/104/10/4008.abstracthttp://www.pnas.org/content/104/10/4008.full.pdf.

Lennart Kester and Alexander van Oudenaarden. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*, 23(2):166–179, 2018. ISSN 18759777. doi: 10.1016/j.stem.2018.04.014. URL https://doi.org/10.1016/j.stem.2018.04.014.

Tae Hee Kim, Assieh Saadatpour, Guoji Guo, Madhurima Saxena, Alessia Cavazza, Niyati Desai, Unmesh Jadhav, Lan Jiang, Miguel N. Rivera, Stuart H. Orkin, Guo Cheng Yuan, and Ramesh A. Shivdasani. Single-Cell Transcript Profiles Reveal Multilineage Priming in Early Progenitors Derived from Lgr5+Intestinal Stem Cells. *Cell Reports*, 16(8):2053–2060, 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.07.056. URL http://dx.doi.org/10.1016/j.celrep.2016.07.056.

Yu Kitadate, David J. Jörg, Moe Tokue, Ayumi Maruyama, Rie Ichikawa, Soken Tsuchiya, Eri Segi-Nishida, Toshinori Nakagawa, Aya Uchida, Chiharu Kimura-Yoshida, Seiya Mizuno, Fumihiro Sugiyama, Takuya Azami, Masatsugu Ema, Chiyo Noda, Satoru Kobayashi, Isao Matsuo, Yoshiakira Kanai, Takashi Nagasawa, Yukihiko Sugimoto, Satoru Takahashi, Benjamin D. Simons, and Shosei Yoshida. Competition for Mitogens Regulates Spermatogenic Stem Cell Homeostasis in an Open Niche. *Cell Stem Cell*, 24(1):79–92, 2019. ISSN 18759777. doi: 10.1016/j.stem.2018.11.013.

Allon M. Klein and Benjamin D. Simons. Universal patterns of stem cell fate in cycling adult tissues. *Development*, 138(15):3103–3111, 2011. ISSN 0950-1991. doi: 10.1242/dev.060103. URL http://dev.biologists.org/cgi/doi/10.1242/dev.060103.

Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620, 2015. ISSN 10974164. doi: 10.1016/j.molcel.2015.04.005. URL http://dx.doi.org/10.1016/j.molcel.2015.04.005.

Vasiliki Kostiou, Huairen Zhang, Michael WJ Hall, Philip Jones, and Benjamin Hall. Methods for analysing lineage tracing datasets. *bioRxiv*, pages 10–12, 2020. doi: 10.1101/2020.01.11.901819.

Vasiliki Kostiou, Huairen Zhang, Michael W. J. Hall, Philip H. Jones, and Benjamin A. Hall. Methods for analysing lineage tracing datasets. *Royal Society Open Science*, 8 (5), 2021. doi: 10.1098/rsos.202231.

Kai Kretzschmar and Fiona M. Watt. Lineage tracing. *Cell*, 148(1-2):33–45, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.01.002. URL http://dx.doi.org/10.1016/j.cell.2012.01.002.

Jinzhi Lei, Simon A. Levin, and Qing Nie. Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. *Proceedings of the National Academy of Sciences*, 111(10), 2014. ISSN 10916490. doi: 10.1073/pnas.1324267111.

Marc Leushacke, Annie Ng, Joerg Galle, Markus Loeffler, and Nick Barker. Lgr5+ Gastric Stem Cells Divide Symmetrically to Effect Epithelial Homeostasis in the Pylorus. *Cell Reports*, 5(2):349–356, 2013. ISSN 22111247. doi: 10.1016/j.celrep.2013.09.025. URL http://dx.doi.org/10.1016/j.celrep.2013.09.025.

Jean Livet, Tamily A. Weissman, Hyuno Kang, Ryan W. Draft, Ju Lu, Robyn A. Bennis, Joshua R. Sanes, and Jeff W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, 2007. ISSN 14764687. doi: 10.1038/nature06293.

Carlos Lopez-Garcia, Allon M. Klein, Benjamin D. Simons, and Douglas J. Winton. Intestinal Stem Cell Replacement Follows a Pattern of Neutral Drift. *Science*, 330 (6005):822–825, 2010. ISSN 0036-8075. doi: 10.1126/science.1196236. URL https://science.sciencemag.org/content/330/6005/822.

Aaron T. L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5: 2122, 10 2016a. ISSN 2046-1402. doi: 10.12688/f1000research.9501.2. URL https://www.ncbi.nlm.nih.gov/pubmed/27909575https://www.ncbi.nlm.nih.gov/pmc/PMC5112579/.

Aaron T.L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, 2016b. ISSN 1474760X. doi: 10.1186/s13059-016-0947-7. URL http://dx.doi.org/10.1186/s13059-016-0947-7.

Charles R MacCluer. The Many Proofs and Applications of Perron's Theorem. *SIAM Review*, 42(3):487–498, 2000.

Eliana Marinari, Aida Mehonic, Scott Curran, Jonathan Gale, Thomas Duke, and Buzz Baum. Live-cell delamination counterbalances epithelial growth to limit tissue overcrowding. *Nature*, 484(7395):542–545, 2012. ISSN 00280836. doi: 10.1038/nature10984.

Davis J. McCarthy, Yunshun. Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 5 2012. ISSN 1362-4962. doi: 10.1093/nar/gks042. URL https://www.ncbi.nlm.nih.gov/pubmed/22287627https://www.ncbi.nlm.nih.gov/pmc/PMC3378882/.

Franklin C. McLean. Application of the Lawof Chemical Equilibrium (Law of Mass Action) to Biological Problem. *Physiological Reviews*, 18(4):495–523, 1938. doi: 10.1152/physrev.1938.18.4.495. URL https://doi.org/10.1152/physrev.1938.18.4.495.

Allyson J. Merrell and Ben Z. Stanger. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style. *Nature Reviews Molecular Cell Biology*, 17 (7):413–425, 2016. ISSN 14710080. doi: 10.1038/nrm.2016.24.

C. D. Meyer and M. W. Stadelmaier. Singular M-matrices and inverse positivity. *Linear Algebra and Its Applications*, 22(C):139–156, 1978. ISSN 00243795. doi: 10.1016/0024-3795(78)90065-4.

Carl D Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000. ISBN 0-89871-454-0.

John Milnor. *Morse Theory. (AM-51)*. Princeton University Press, Princeton, 2016. doi: doi:10.1515/9781400881802. URL https://doi.org/10.1515/9781400881802.

Samantha A. Morris. The evolving concept of cell identity in the single cell era. *Development (Cambridge)*, 146(12):1–5, 2019. ISSN 14779129. doi: 10.1242/dev.169748.

National Institute of Health. Stem Cell Basics, 2016. URL https://stemcells.nih.gov/info/basics.

Svetoslav Nikolov, Elka Yankulova, Olaf Wolkenhauer, and Valko Petrov. Principal difference between stability and structural stability (robustness) as used in systems biology. *Nonlinear Dynamics, Psychology, and Life Sciences*, 11(4):413–433, 2007. ISSN 10900578.

Bhupinder Pal, Yunshun Chen, François Vaillant, Paul Jamieson, Lavinia Gordon, Anne C. Rios, Stephen Wilcox, Naiyang Fu, Kevin He Liu, Felicity C. Jackling, Melissa J. Davis, Geoffrey J. Lindeman, Gordon K. Smyth, and Jane E. Visvader. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nature Communications*, 8(1):1–13, 2017. ISSN 20411723. doi: 10.1038/s41467-017-01560-x. URL http://dx.doi.org/10.1038/s41467-017-01560-x.

Cristina Parigini and Philip Greulich. Universality of clonal dynamics poses fundamental limits to identify stem cell self-renewal strategies. *eLife*, 9:1–44, 2020. ISSN 2050084X. doi: 10.7554/eLife.56532.

Przemysław Rafał Paździorek. Mathematical Model of Stem Cell Differentiation and Tissue Regeneration with Stochastic Noise. *Bulletin of Mathematical Biology*, 76(7): 1642–1669, 2014. ISSN 15229602. doi: 10.1007/s11538-014-9971-5.

Jim Pitman. *Probability*. Springer-Verlag New York, 1993. ISBN 9781461243748 1461243742. URL http://catalog.hathitrust.org/api/volumes/oclc/26932280.html.

Riccardo Poli, James Kennedy, and Tim Blackwell. Particle swarm optimization. *Swarm intelligence*, 1(1):33–57, 2007.

Christopher S. Potten and Markus Loeffler. Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt. *Development*, 110(4): 1001–1020, 1990. ISSN 0950-1991. URL http://www.ncbi.nlm.nih.gov/pubmed/2100251.

Alberto Puliafito, Lars Hufnagel, Pierre Neveu, Sebastian Streichan, Alex Sigal, D. Kuchnir Fygenson, and Boris I. Shraiman. Collective and single cell behavior in epithelial contact inhibition. *Proceedings of the National Academy of Sciences*, 109(3): 739 – 744, 1 2012. ISSN 00225347. doi: 10.1016/j.juro.2012.06.073. URL http://www.pnas.org/content/109/3/739.abstract.

Sapna Puri, Alexandra E. Folias, and Matthias Hebrok. Plasticity and dedifferentiation within the pancreas: Development, homeostasis, and disease. *Cell Stem Cell*, 16(1): 18–31, 2015. ISSN 18759777. doi: 10.1016/j.stem.2014.11.001. URL http://dx.doi.org/10.1016/j.stem.2014.11.001.

Anne C. Rios, Nai Yang Fu, Joseph Cursons, Geoffrey J. Lindeman, and Jane E. Visvader. The complexities and caveats of lineage tracing in the mammary gland. *Breast Cancer Research*, 18(1):1–5, 2016. ISSN 1465542X. doi: 10.1186/s13058-016-0774-5. URL http://dx.doi.org/10.1186/s13058-016-0774-5.

Laila Ritsma, Saskia I.J. J. Ellenbroek, Anoek Zomer, Hugo J. Snippert, Frederic J. De Sauvage, Benjamin D. Simons, Hans Clevers, and Jacco Van Rheenen. Intestinal crypt homeostasis revealed at single stem cell level by in vivo live-imaging. *Nature*, 507(7492):362–365, 2014. ISSN 14764687. doi: 10.1038/nature12972. URL http://dx.doi.org/10.1038/nature12972.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Verónica Rodilla and Silvia Fre. Cellular plasticity of mammary epithelial cells underlies heterogeneity of breast cancer. *Biomedicines*, 6(4):9–12, 2018. ISSN 22279059. doi: 10.3390/biomedicines6040103.

Panteleimon Rompolas, Kailin R. Mesa, Kyogo Kawaguchi, Sangbum Park, David Gonzalez, Samara Brown, Jonathan Boucher, Allon M. Klein, and Valentina Greco. Spatiotemporal coordination of stem cell commitment during epidermal

homeostasis. *Science*, 352(6292):1471–1474, 2016. ISSN 10959203. doi: 10.1126/science.aaf7012.

Steffen Rulands and Benjamin D. Simons. Tracing cellular dynamics in tissue development, maintenance and disease. *Current Opinion in Cell Biology*, 43:38–45, 2016. ISSN 18790410. doi: 10.1016/j.ceb.2016.07.001. URL http://dx.doi.org/10.1016/j.ceb.2016.07.001.

Brian Sauer. Inducible Gene Targeting in Mice Using the Cre/lox System. *Methods*, 14: 381–392, 1998. URL http://ac.els-cdn.com/S104620239890593X/1-s2.0-S104620239890593X-main.pdf?_tid=df26f34e-5c18-11e7-8e63-00000aacb361&acdnat=1498664936_787b2162eeadb0d629e2292aa5ae317b.

Pak Chung Sham, Shun H Yip, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*, 2 2018. doi: 10.1093/bib/bby011. URL https://dx.doi.org/10.1093/bib/bby011.

Stanley Shostak. (Re)defining stem cells. *BioEssays*, 28(3):301–308, 2006. ISSN 02659247. doi: 10.1002/bies.20376.

Boris I. Shraiman. Mechanical feedback as a possible regulator of tissue growth. *Proceedings of the National Academy of Sciences*, 102(9):3318–3323, 2005. ISSN 00278424. doi: 10.1073/pnas.0404782102.

Benjamin D. Simons and Hans Clevers. Strategies for homeostatic stem cell self-renewal in adult tissues. *Cell*, 145(6):851–862, 2011a. ISSN 00928674. doi: 10.1016/j.cell.2011.05.033. URL http://dx.doi.org/10.1016/j.cell.2011.05.033.

Benjamin D. Simons and Hans Clevers. Stem cell self-renewal in intestinal crypt. *Experimental Cell Research*, 317(19):2719–2724, 2011b. ISSN 0014-4827. doi: https://doi.org/10.1016/j.yexcr.2011.07.010. URL https://www.sciencedirect.com/science/article/pii/S0014482711002862.

Qiaojun Situ and Jinzhi Lei. A mathematical model of stem cell regeneration with epigenetic state transitions. *Mathematical Biosciences and Engineering*, 14(5-6): 1379–1397, 2017. ISSN 15510018. doi: 10.3934/mbe.2017071.

Kieran Smallbone and Bernard M. Corfe. A mathematical model of the colon crypt capturing compositional dynamic interactions between cell types. *International Journal of Experimental Pathology*, 95(1):1–7, 2014. ISSN 09599673. doi: 10.1111/iep.12062.

Hugo J. Snippert and Hans Clevers. Tracking adult stem cells. *EMBO Reports*, 12(2): 113–122, 2011. ISSN 1469221X. doi: 10.1038/embor.2010.216. URL http://dx.doi.org/10.1038/embor.2010.216.

Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15:255, 2 2018. URL https://doi.org/10.1038/nmeth.4612http://10.0.4.14/nmeth.4612.

Heng Sun, Zhengqiang Miao, Xin Zhang, Un In Chan, Sek Man Su, Sen Guo, Chris Koon Ho Wong, Xiaoling Xu, and Chu Xia Deng. Single-cell RNA-Seq reveals cell heterogeneity and hierarchy within mouse mammary epithelia. *Journal of Biological Chemistry*, 293(22):8315–8329, 2018. ISSN 1083351X. doi: 10.1074/jbc.RA118.002297.

Zheng Sun and Natialia L. Komarova. Stochastic modeling of stem-cell dynamics with control Zheng. *Math Bioscience*, 240(2):231–240, 2012. doi: 10.1016/j.mbs.2012.08.004.

Purushothama Rao Tata and Jayaraj Rajagopal. Cellular plasticity: 1712 to the present day. *Current Opinion in Cell Biology*, 43:46–54, 2016. ISSN 18790410. doi: 10.1016/j.ceb.2016.07.005. URL http://dx.doi.org/10.1016/j.ceb.2016.07.005.

Purushothama Rao Tata, Hongmei Mou, Ana Pardo-Saganta, Rui Zhao, Mythili Prabhu, Brandon M. Law, Vladimir Vinarsky, Josalyn L. Cho, Sylvie Breton, Amar Sahay, Benjamin D. Medoff, and Jayaraj Rajagopal. Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature*, 503(7475):218–223, 11 2013. ISSN 0028-0836. doi: 10.1038/nature12777. URL http://www.nature.com/articles/nature12777.

Cristian Tomasetti, Bert Vogelstein, and Giovanni Parmigiani. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences*, 110(6):1999–2004, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1221068110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1221068110.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014. ISSN 15461696. doi: 10.1038/nbt.2859. URL http://dx.doi.org/10.1038/nbt.2859.

Barbara Treutlein, Doug G. Brownfield, Angela R. Wu, Norma F. Neff, Gary L. Mantalas, F. Hernan Espinoza, Tushar J. Desai, Mark A. Krasnow, and Stephen R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, 2014. ISSN 14764687. doi: 10.1038/nature13173. URL http://dx.doi.org/10.1038/nature13173.

Lukas Valihrach, Peter Androvic, and Mikael Kubista. Platforms for single-cell collection and analysis. *International Journal of Molecular Sciences*, 19(3):22–24, 2018. ISSN 14220067. doi: 10.3390/ijms19030807.

Catalina A. Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C. Marioni. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods*, 14(6):565–571, 2017. doi: 10.1038/nmeth.4292.Normalizing.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

N.G. G Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library, 1981. ISBN 0444893490. doi: http://dx.doi.org/10.1016/B978-1-85617-567-8.50004-4.

Alexandra Van Keymeulen, Marco Fioramonti, Alessia Centonze, Gaëlle Bouvencourt, Younes Achouri, and Cédric Blanpain. Lineage-Restricted Mammary Stem Cells Sustain the Development, Homeostasis, and Regeneration of the Estrogen Receptor Positive Lineage. *Cell Reports*, 20(7):1525–1532, 2017. ISSN 22111247. doi: 10.1016/j.celrep.2017.07.066.

Jane E. Visvader and John Stingl. Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes and Development*, 28(11): 1143–1158, 2014. ISSN 15495477. doi: 10.1101/gad.242511.114.

Ekaterina Vorotelyak, Andrey Vasiliev, and Vasiliy Terskikh. The Problem of Stem Cell Definition. *Cell and Tissue Biology*, 14(3):169–177, 2020. ISSN 1990-5203. doi: 10.1134/S1990519X20030086. URL https://doi.org/10.1134/S1990519X20030086.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, 2016. ISSN 15461696. doi: 10.1038/nbt.3711. URL http://dx.doi.org/10.1038/nbt.3711.

Chunhui Wang, John R. Christin, Maja H. Oktay, and Wenjun Guo. Lineage-Biased Stem Cells Maintain Estrogen Receptor Positive and Negative Mouse Mammary Luminal Lineages. *Cell Reports*, 18(12):2825–2835, 2017. doi: 10.1016/j.celrep.2017.02.071.

Zhong Wang, Gerstein Mark, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. ISSN 0036-8075. doi: 10.1038/nrg2484.RNA-Seq.

Fiona M. Watt and Brigid L. M. Hogan. Out of Eden: stem cells and their niches. *Science*, 287(February):1427–1430, 2000.

Leland Wilkinson and Michael Friendly. History corner the history of the cluster heat map. *American Statistician*, 63(2):179–184, 2009. ISSN 00031305. doi: 10.1198/tas.2009.0033.

Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btv088.

Hsu Ya-Chieh. Theory and Practice of Lineage Tracing. *Stem Cells*, 33(11):3197–3204, 2015. doi: 10.1002/stem.2123. URL https://stemcellsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/stem.2123.