Ollscoil
Mhá Nuad
Ollscoil na hÉireann
Má Nuad

A mathematical framework for clonal data analysis

Author:
Giulio Prevedello

Supervisor:
Prof. Ken Duffy

A thesis submitted in fulfilment of the requirements
for the Ph.D. degree in Applied Mathematics

at the

Hamilton Institute
Maynooth University,
Maynooth, Co. Kildare,
Ireland
Department head: Prof. Ken Duffy

October 2018

Summary

This dissertation reports on the development of the mathematical and statistical framework that was necessary for the analysis of data from a novel single-cell assay designed to address questions in fundamental biology. Many biological systems function by generating new cells from activated ancestors through cellular division. To investigate such systems, a high throughput experimental protocol was recently developed that marks initial cells so that their cellular offspring, the number of rounds of division from their ancestor, and their phenotype can be determined. The clonal data that result from this technique, however, are characterised by familial associations that impede their analysis using classical quantitative tools, necessitating the development of a new mathematical framework where suitable statistics are formulated that take these complex dependencies into account. The design, development and implementation of that framework, as well as inferences made from its use, are the subject of the present thesis.

Acknowledgements

I would like to thank Prof. Ken Duffy, for all the time, patience, and effort shown to me during his supervision. Achieving this thesis was made possible by his invaluable contributions and ideas. Learning from his reasoning has improved me as a researcher and as a person.

I would like to thank my Australian collaborators at Walter Eliza Hall Institute (WEHI) Andrey, Julia, Miles, Jie, and Su, for the time I enjoyed working with them, and for the friendliness they manifested. A special mention goes to Prof. Phil Hodgkin, for his commitment to science that inspired me to pursue research with more and more determination.

The collaboration with the WEHI team has been possible thanks to the European Union's Marie Skłodowska-Curie Actions, that provided the funding for my work within the Quantitative T-cell Immunology (QuanTI) network project. I am grateful to all the people I met as part of QuanTI: Prof. Carmen Molina-Paris and Prof. Grant Lythe, for the organisation of our several meetings; the senior partners, for the expertise they provided; last but not least, all the fellows, for the unforgettable moments we shared.

I would like to thank Prof. Muriel Médard and her group members for welcoming my office mates and me in their laboratory.

I would like to thank my office mates Aisling, Alex, Gianfelice, Harry, Mark, and Sarah for providing moral support and for putting up with my usual ranting.

I would like to thank every person I met within the four walls of the Hamilton Institute, past and present. A special mention goes to Rosemary and Kate, for their infinite help, their welcoming cheer and the kindness they reserve to everyone in the institute.

I would like to thank Prof. Mark Dukes, Alex, Stefania and all those that spent their time reviewing this thesis and giving precious advice for its improvement.

I would like to thank my friends and social companions: Alessandro, the (extended) members of COER, Karl and Yasmine, Niall, Riccardo, and all those people that I have not met as often as I wish.

I would like to thank my mother, my father and my brother for the unconditional support I received from them to sustain me in my life choices.

Finally, I would like to thank Lara who stayed by my side all these years, with incredible effort and limitless patience, motivating me through her example. I owe a lot to her.

Contents

Chapter 1

Questions in quantitative immunology

1.1 Abstract

In this chapter, we provide the minimal background in immunology that is required to motivate our work. First, we outline the main features of the immune system, focusing on how a particular cellular subgroup develops and functions, since it will be central in the following chapters. These cells possess the ability to generate a population of heterogeneous cell types through subsequent divisions, starting from a smaller pool of progenitors that are homogeneous in appearance. To explain this phenomenon, researchers have investigated different mechanisms, that drive the division process of the cells and depend on the system conditions in which these proliferate. From the literature, we report recent findings and the experimental methods employed to cast light on this subject. The reader with a solid knowledge on immunology can skip to Section 1.7, where we provide a summary for the thesis.

1.2 Characteristics of the immune system

The immune system is the part of the body that defends its host by foreign pathogens (from the ancient Greek $\pi\acute{\alpha}\theta o\varsigma$ pathos "suffering" and -$\gamma\varepsilon\nu\acute{\eta}\varsigma$ -genēs "generator of"). This system is organized in primary lymphoid organs, including the bone marrow and the thymus, where immune cells develop, and secondary lymphoid organs, such as the spleen and the lymph nodes, where immunity is initiated and orchestrated.

The hallmarks of the immune system response are:

Recognition (Owen et al., 2013). As a pathogen breaks into the body, overcoming the first physical line of defences, such as skin or mucous membranes, cells of the immune system need way to detect the threat. Their main method to interact with the foreign organism is through a specific set of receptor molecules that are present on their cellular membrane. Any ligand capable of binding with these receptors is called an antigen, a general term to refer to molecules that can trigger an immune response, which includes whole pathogens or their fragments, or the products they secrete.

Activation and response (Owen et al., 2013). Upon pathogen recognition, the immune cells are activated and initiate their program mounting the immune response. The resulting population of cells is diverse, as some engage with the elimination of the pathogen or virally infected cells (cellular immunity), while others release soluble proteins to recruit and regulate the whole response (humoral immunity).

Tolerance (Murphy and Weaver, 2016). In order to avoid targeting cells of the host itself, the immune system must be capable of self and non-self discrimination. The current consensus is that tolerance is organised in layers and achieved through different checkpoints. The central tolerance mechanism in the thymus eliminates cells that would otherwise be activated by molecules of the host body. To avoid encounters with cells that have survived central tolerance deletion, some organs (for example the pancreas) restrict access to the immune system. If self-reaction occurs, peripheral tolerance can suppress the response of the reacting cells, through internal controls or inhibitory regulation from other immune cells.

Dysfunctions of these mechanisms may lead to severe consequences for the health of the host body (Owen et al., 2013):

Hypersensitivity. An overreaction to an antigenic molecules that would not cause any harm. Allergies are a manifestation of this effect.

Autoimmune diseases. Incorrect identification of self tissue as non-self may lead to disease that are organ-specific (e.g. multiple sclerosis) or systemic (e.g. rheumatoid arthritis).

Immunodeficiencies. These occur when the immune system fails to mount an adequate response against a pathogen. Depending on their cause, immunodeficiencies are separated in primary, due to genetic inheritance or developmental defects, and secondary, if resulting from external agents or infections.

The immune system is distinguished into two parts that, although functionally different, cooperate and interact together to provide defence against a wide range of infectious agents:

Innate immune system (Owen et al., 2013). Being one of the earliest form of protection in evolutionary terms, this system is found in all multicellular plant and animals. Most cells from the innate immune system present Pattern Recognition Receptors (PRRs) that recognise Pathogen-Associated Molecular Patterns (PAMPs), chemical structures common to several pathogenic microbes. Upon activation through their PRRs, cells from the innate immune system react with a quick response that is initiated within minutes or hours (see Fig. 1.1). The specificity of PRRs ensures perfect self/non-self discrimination, since PAMPs are not present in the healthy host body.

Adaptive immune system (Owen et al., 2013). This system arose in jawed vertebrates as a result of evolution. An adaptive immune cell is characterised by the presence of a receptor that binds to very small class of antigens, with a stronger or weaker interaction that depends on the binding affinity. This biological recognition is based on reciprocity of receptor-antigen shapes, which, in this system, restricts the interaction of a given receptor to a small set antigens that are specific for it. The detection of a multitude of antigens is then achieved by an army of cells, thus covering a wide range of specificities.

Upon activation, cells from the adaptive system that have never experienced antigen interaction (called "naive" cells), remain at rest for days, as opposed to the strategy of the innate immune system, whose cells reaction triggers within hours. After this period of latency, activated cells undergo extensive mitotic division generating a population of clones that all express the same receptor and are capable of recognising the cells infected by the pathogen they are specific for. In this way, the infected cell can be eliminated and, if the immune response is adequate, the pathogen is eradicated (see Fig. 1.1).

The principle that the protection against an antigen is achieved through a small subset of specific cells that expands into a larger population of clones, was postulated by F. M. Burnet in his clonal selection theory (Burnet, 1957). In that paper, he was the first to propose that specificity in the adaptive immune system was achieved by stochastic generation of antigen receptors. The implication of his idea were furtherer explored in Burnet (1959). Revised over the time, clonal selection theory is currently the accepted paradigm of adaptive immunity (Hodgkin et al., 2007).

Once the pathogen is eliminated, most of the cells produced from an adaptive response die, leaving behind a small number of cells that will serve as long lived memory in case the same antigen enters the host body again. Upon subsequent encounter, memory cells generate an immune response of greater number and more quickly than their naive progenitors (Murphy and Weaver, 2016), thus resulting in more effective protection

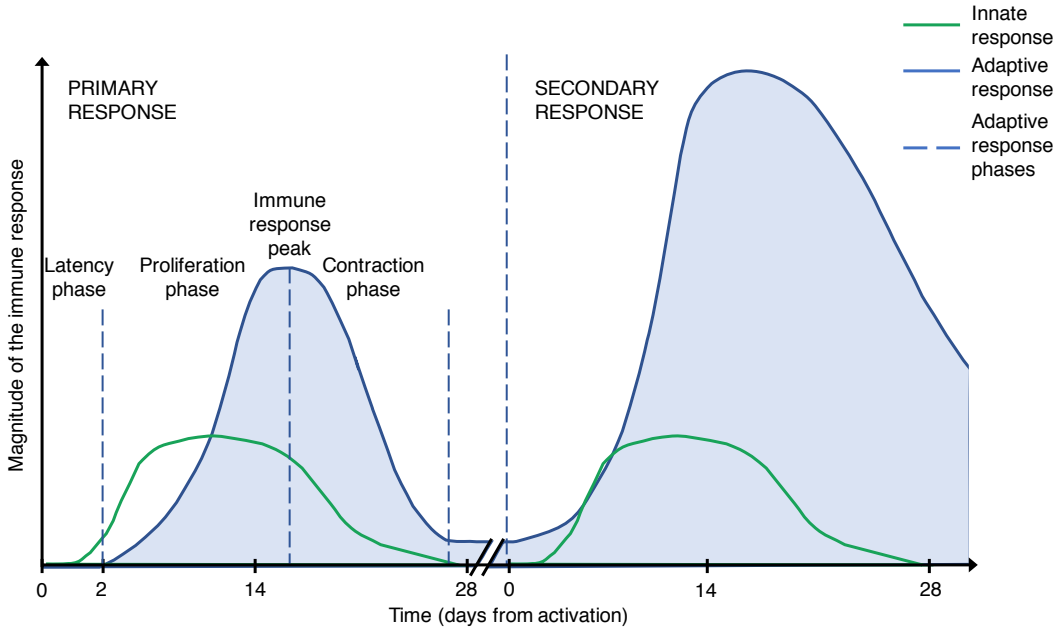

FIGURE 1.1: **Immune response phases.** [Adapted from Fig. 1 of Kaech and Cui (2012) and Fig. 1-8 of Owen et al. (2013)] Upon activation (0 time-points), the cells from the innate immune system generate a response whose magnitude follows the same evolution (green line) for both primary and subsequent antigen encounters (on the left and right side of the plot, respectively). On the other hand, the cells from the adaptive immune system, that experience the antigen for the first time, undergo a phase of latency, after which they commence division. During this proliferation phase, the magnitude of the response (blue line) grows exponentially until the pathogen is eradicated. Around this time, the magnitude of the response reaches its peak that, in general, is higher than the one achieved from an innate response. Afterwards, a contraction phase begins, where most of the newly generated adaptive system cells die. A subset of these, larger in size than the initial number of responding naive progenitors, survives to provide protection for future encounters of the same antigen. If this event occurs, the subsequent response follows a pattern similar to the first, but generating a peak of greater size, without latency phase.

(see Fig. 1.1). This feature of persistent memory is a fundamental characteristic of the adaptive immune system.

An important cellular subgroup of the immune system are the lymphocytes, so called for being the major subset of the cell found in the lymphatic system. Lymphocytes include Natural Killer (NK) cells from the innate immune system, and both B and T cells from the adaptive immune system (Murphy and Weaver, 2016). The generation of randomised receptor, which are antigen specific, occurs in B and T cells through the somatic rearrangement mechanism called V(D)J-recombination, an important discovery accomplished by S. Tonegawa and co-workers (Tonegawa, 1983).

From this programme, the potential repertoire of different B-cell receptors (BCRs) and T-cell receptors (TCRs) is vast. Concerning the latter, the number of possible TCRs produced is of about 10^{15} (Davis and Bjorkman, 1988; Nikolich-Žugich et al.,

2004; Sewell, 2012; Lythe et al., 2016), which is so large that presenting even one T cell for each TCR would be physically impossible for any human body (Lythe et al., 2016). In fact, the sets of cells with same TCR have been estimated to be between 10^6 and 10^8 (Qi et al., 2014; Lythe et al., 2016). Still, the TCR repertoire is so diverse that two samples from the same individual present only a small portion of overlapping repertoire (Heather et al., 2017). This poses important challenges in the estimation of TCR repertoire, to the point that specific bioinformatic tools have been specifically developed to analyse TCR sequencing data (Yu et al., 2015; Heather et al., 2017). The interest in this problem is also explained as a precise method to determine the antigen protection in single individuals, such as their TCR repertoire, would open the door to personalised medical treatments.

In the next sections, we narrow in, outlining the development and function of T cells, in order to provide the context for the research questions related, that will be addressed in the following chapters.

1.3 Naive T-cell development

Hematopoietic stem cells (HSCs) are found in the bone marrow. This type of cells is capable of renovating its own compartment (stemness property) and generating all the cells from the blood and the immune system (multipotent property). Descendants of the HSCs are the common lymphoid progenitors (CLPs), from which B and T cells derive (Murphy and Weaver, 2016).

Once a CLP commits to the B-cell lineage, it becomes a pro-B cell that remains in the bone marrow and, ultimately, develops into a B cell. During this process, the maturing cell acquires a specific receptor by V(D)J recombination (Murphy and Weaver, 2016). Although the "B" from B cell conveniently refers to the place where B cells develop, i.e. the bone marrow, in fact the name originated from the Bursa of Fabricius, a lymphoid organ in birds where these cells were first discovered to mature (Cooper et al., 1965; Gitlin and Nussenzweig, 2015).

Similarly, the "T" of T cell is defined after the initial letter of the thymus, the organ where the offspring of CLPs that committed to the T-cell lineage migrate to as T-cell precursor and complete their development into T cells. It is in the thymus where the maturing T cell, called thymocyte, undergoes V(D)J recombination that results in the expression of a randomized TCR for a specific antigen (Murphy and Weaver, 2016).

In the thymic cortex, the thymocyte is selected for their TCR's capacity to engage the host Major Histocompatibility Complex (MHC) molecules. These molecules are a

class of membrane proteins that can form complexes with an antigen to display it on the surface of a cell, thus referred to as antigen-presenting cell (APC). The antigen so exhibited can be a peptide either derived from a pathogenic agent or produced from the host body, hence termed self-antigen in this second case (Murphy and Weaver, 2016).

To ensure MHC interaction from a newly generated TCR, the thymocyte undergoes a process called positive selection, eliminating all T cells under development whose TCR does not bind with sufficient strength to the MHC of the host body. If the TCR-MHC interaction is too strong, instead, the maturing T cell bearing that TCR is also eliminated from the thymus to preserve self-tolerance, through another mechanism termed negative selection. In general, MHC proteins differ between individuals, in fact the discrimination of antigen/self-antigen is host-dependent. This is a major cause of transplant rejection, occurring when the cell from a donor body is not recognised as "self" (Murphy and Weaver, 2016).

Among the relevant T-cell output from the thymus, we mention T helper (T_H) cells, regulatory T cells (T_{REG}) and cytotoxic T lymphocytes (CTL). Characteristic of the first two subgroups is the expression of Cluster Differentiation (CD) 4 on their membrane surface, also said to be CD4 positive ($CD4^+$). CTLs are CD4 negative ($CD4^-$), but have positive expression of CD8 instead, and for this reason are referred to as $CD8^+$ T cells (Owen et al., 2013).

As these cells mature and exit the thymus, they join the naive T cell pool, waiting to come in contact with the antigen they are specific to (Murphy and Weaver, 2016). In the following section we focus on $CD8^+$ T cells and provide some details about their immune response following antigen recognition.

1.4 $CD8^+$ T-cell response

The average human body is estimated to have 3.72×10^{13} total cells (Bianconi et al., 2013). Of these, an order of 10^{11} consists of T cells (Jenkins et al., 2010). After maturation, these lymphocytes exit the thymus and circulate in the secondary lymphoid organs, checking other cells for the presence of the antigen specific to their TCR. In a typical scenario, an APC, such as a dendritic cell, enters the draining lymph node local to an ongoing infection while displaying on its surface a foreign antigenic molecule in complex with the MHC. In the case where a naive T cell bearing the specific TCR encounters such APC, the TCR binds with the antigen-MHC complex and transmits a signal, i.e. any event inducing a cell state change, that activates the T cell (activation signal) and triggers its subsequent immune response. This discrimination in

favour of cells that are antigen-specific is referred to as clonal selection (see Fig. 1.2). Additionally, T cells can also recognise costimulatory signal B7, a class of proteins found on some APC, ligand to the CD28 receptor found on the T cell surface. B7-CD28 interaction provides an additional stimulus that enhances the T cell response (Murphy and Weaver, 2016).

Measurements in humans are uncommon, whereas the mouse is a standard scientific model. In this system, the activated T cell, after a latency period from 30 to 50 hours (h) (Hawkins et al., 2007a), undergoes mitosis producing two daughters. In turn, these two cells and their offspring divide, for a total number spanning from 5 to 20 rounds (Duffy and Hodgkin, 2012) with each subsequent division taking approximately 10 to 13 h (Dowling et al., 2014), thus generating a family of clonal duplicates. All the T cells recruited from the same antigen undertake this pattern of proliferation and, together, they build the immune response. As it grows, the resulting population differentiates into two functionally distinct classes of effector and memory cells (see Fig. 1.2).

Effector cells travel to the site of infection, where they may interact with cytokines, a wide class of proteins, which includes interleukins (ILs), secreted by other cells of the immune system cell fighting the pathogen to amplify the effector efficacy (Owen et al., 2013). Among the ILs, we mention IL-2 which is a signal for expansion and differentiation of T cells, originally called T-cell growth factor (Smith and Cantrell, 1985). In particular, when $CD8^+$ T cells are activated, they are capable of IL-2 autocrine production, namely they can release IL-2 and sense their own product to enhance their immune response (Owen et al., 2013). Once the place of infection is reached, T cells detect the diseased cells, which present the antigen on their surface. Upon interaction, the T cell injects molecules into the infected cell inducing their apoptosis (see Fig. 1.2), that is a mechanism of programmed cell death, thus eliminating it (Murphy and Weaver, 2016). If the overall immune response produced is effective, the disease is ultimately eradicated.

At this point, the population of specific T cells has reached its peak and the contraction phase begins (see Fig. 1.2), when most of the T lymphocytes die, leaving behind 5-10% of the population, which consists of the memory pool (Schumacher et al., 2010; Kaech and Cui, 2012). These **memory cells** remain in the secondary lymphoid organs where they maintain their number by homoeostatic equilibrium to provide long-lasting protection in case the pathogen they are specific for returns (Surh and Sprent, 2008). Upon reinfection, these T cells behave similarly as their naive progenitors but their reaction is more effective, as they commence proliferating in a shorter time after pathogen exposure and generate an immune response of greater magnitude in terms of cell number (Murphy and Weaver, 2016).

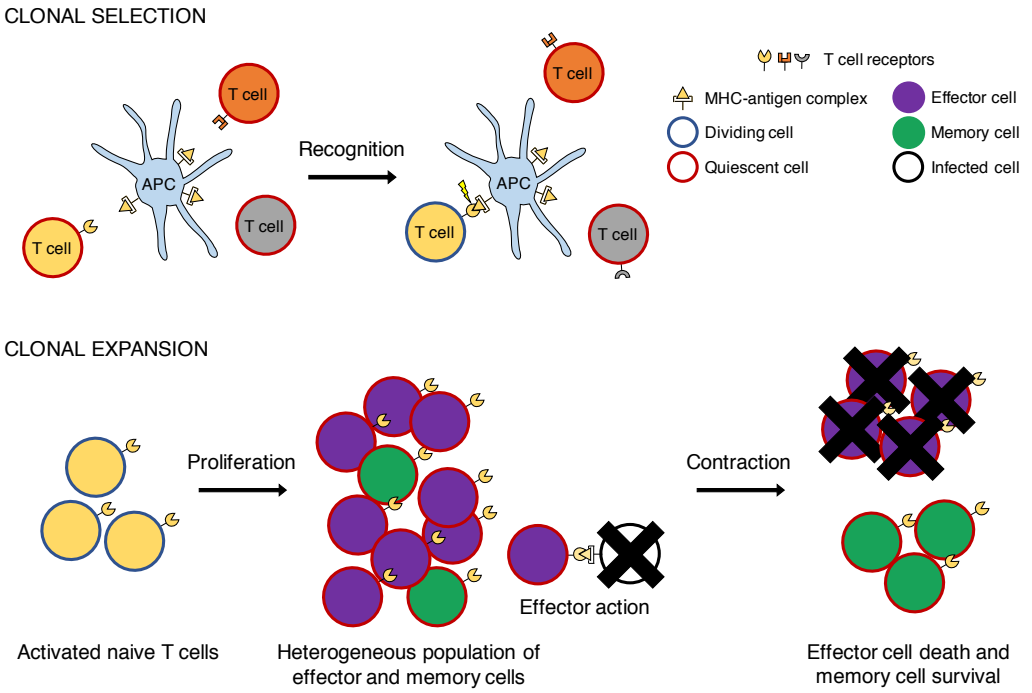

FIGURE 1.2: **Clonal selection and clonal expansion.** (Top panel) When an Antigen Presenting Cell (APC) enters the lymph node, it activates only the T cells that can recognize the antigen, bound to the Major Histocompatibility Complex (MHC) molecules on the APC surface, in a process called clonal selection. These cells are selected because they bear the same T-Cell Receptor (TCR) that is specific for such antigen, thus termed clones with respect to their TCR equivalence. (Bottom panel) As the antigen-specific T-cell clones are activated through the process of clonal selection, they enter in a proliferating phase where they divide and increase their total number. The result of this clonal expansion is a larger population of T cells that are heterogeneous with respect to their function. Some cells have effector properties, that will search for cells infected by the antigenic pathogen to eliminate them. When the threat is eradicated, a contraction phase occurs, where effector cells die, leaving behind a smaller pool of the cells that, instead, acquired memory function. These cells will protect the host in case of subsequent antigen exposure, generating an immune response that is more effective that the one from the first encounter.

Similarly to T cells, B cells are involved with a pattern of activation, proliferation and differentiation into effector and memory. Effector B cells, however, are referred to as plasma cells and produce antibodies that boost and regulate the elimination of the pathogen, thus participating to the humoral, rather than cellular, immune response. Another distinct feature of B cells is that a subgroup of these may aggregate in the lymph nodes to form a germinal centre, where proliferating cells can alter their genes in a process called somatic hypermutation, thus leading to affinity maturation. This latter process consists in the selection for the mutated B cells presenting higher affinity with the antigen that triggered the immune reaction.

While the T-cell immunity has been extensively studied at the population level, there are still many open questions concerning how a single naive T cell contributes to the

immunisation and what are the main drivers of such contribution. In fact, unveiling these mechanisms is essential to predict how the T cell behaves upon different stimulations and conditions, and to reprogram its processes for medical purposes. Advances on this topic have already lead to new therapeutic solutions, particularly against cancer (Webster and Mentzer, 2014; Webster, 2014; Johnson et al., 2015; Johnson and Sosman, 2015; June et al., 2015; Schumacher et al., 2015; Schumacher and Schreiber, 2015; Rosenberg and Restifo, 2015; Verdegaal et al., 2016; Zacharakis et al., 2018).

Chapter 2 of this dissertation is a study on the activation and proliferation of T cells focusing, in particular, on how apparently homogeneous naive T cells develop into a population of cells that are heterogeneous in function and familial size, even under controlled conditions. In the next section, we report the latest results concerning this, yet unexplained, heterogeneity, that are seminal for our work.

1.5 The problem of heterogeneity

In a mouse, there are reported to be about 6×10^7 naive T cells (Jenkins et al., 2010), 10^2 of which are estimated to be specific to a given antigen (Schumacher et al., 2010). This yields about 200 cells that can provide an adequate immune response to counter an infection, making the chances of TCR-antigen interaction remarkably unlikely (Reiner and Adams, 2014). Notwithstanding that, the immune response consistently generates an heterogeneous population of lymphocytes, capable of effector and memory functions. How these rare cells achieve this consistent heterogeneity remains a main conundrum in contemporary immunological research. In particular, it is unclear how each naive progenitor participates to the overall immunity and what are the cellular program that drives the diversity in its offspring. Possible sources of heterogeneity have been considered:

- Differences in the early progenitors (Lemaître et al., 2013; Rohr et al., 2014);
- Lineage priming (Hawkins et al., 2009; Duffy and Hodgkin, 2012; Gerlach et al., 2013);
- Division-dependent processes during proliferation (Gett and Hodgkin, 1998; Jenkins et al., 2008; Schlub et al., 2009; Kinjyo et al., 2015);
- Asymmetric inheritance of molecules, from mother to daughters, upon mitosis, either deterministically or stochastically programmed (Chang et al., 2007; Reiner and Adams, 2014);

- Environmental niches providing various signals, each with possibly different impact on the cell fate, depending on the location (Plumlee et al., 2013; Rohr et al., 2014).

Among the mitosis-linked mechanisms, asymmetric cell division (ACD) has been one of the most studied (Oliaro et al., 2010; Chang et al., 2011; Barnett et al., 2012; King et al., 2012; Arsenio et al., 2014; Metz et al., 2015; Pollizzi et al., 2016; Verbist et al., 2016; Yassin and Russell, 2016). ACD was first proposed in Chang et al. (2007) as the key mechanism that enables the generation from each single cell of memory and effector cells. In that study, it was suggested that T cell activation from an APC induces an asymmetric segregation of molecules having the contact point with the APC as polar cue. Thus, upon ACD, the two daughter cells would inherit molecules in different quantities, with the sibling generated from the side of the interaction with the APC (proximal cell) resulting in a profile that is consistent with the memory lineage, while the other sibling (distal cell) would presents a profile closer to the effector trait.

In their Opinion paper, Reiner and Adams (2014) argued for ACD, as scarcity of antigen specific cells requires a deterministic program in order to guarantee robust and heterogeneous response, which could not be left to aleatory events. Of viewpoint different than Reiner and Adams (2014), Hodgkin's group suggested instead that stochastic mechanisms may be consistent with experimental observations (Hodgkin et al., 2014). They argued that, although subjected to variations, stochastic programs can achieve a consistent outcome thanks to probabilistic properties, notably the law of large numbers, already for a number of T cell recruited as "large" as 20. Thus, the stochastic variability would ensure a diverse cellular commitment, even in absence of specific environmental inputs. Moreover, *in vivo* experiments, using mice lacking of polarity protein Scribble, supported that ACD is not necessary for T cells to mount a robust immune response (Hawkins et al., 2013).

Since immune response diversity is achieved even in *in vitro* systems (Hodgkin et al., 2014), which are characterised by simpler conditions than an *in vivo* response, it is reasonable to undertake an experimental approach that is reductionist, that is unravelling the problem of heterogeneity using *in vitro* protocols where more input signals can be manually controlled.

As a result of their studies following this method of investigation (Gett and Hodgkin, 1998, 2000; Deenick et al., 2003; Tangye et al., 2003; Hasbold et al., 2004; Hawkins et al., 2007b; Hodgkin, 2007; Hommel and Hodgkin, 2007; Turner et al., 2008; Hawkins et al., 2009), Hodgkin's group developed a stochastic, mathematical model, the Cyton Model, of lymphocytes proliferation (Hawkins et al., 2007a). This model consists of different

time variables, one per cellular fate (e.g. death, and division), that compete within each cell in a stochastic manner, to determine the lymphocyte commitment to a certain action. Such mechanism is referred to as stochastic competition. Under this paradigm, the response heterogeneity is explained through the variabilities of these timers, whose distributions are the result of cellular programs and environmental signals. Thus, different system's conditions push the odds in favour of a particular outcome (Duffy and Hodgkin, 2012; Hodgkin et al., 2014).

While an analysis of the mean-population size per generation is provided in Hawkins et al. (2007a), in Subramanian et al. (2008) extrapolations to higher moments were determined by framing the Cyton Model as a non-standard branching process in which lifetimes and offspring numbers are correlated. The use of branching processes theory (Harris, 1964), provides a framework that is employed in many biological models (Kimmel and Axelrod, 2002), although, in general, it requires the assumption that each cell behaves independently of the others. Since strong correlation has been observed between related cells in time lapse microscopy experiments (Hawkins et al., 2009; Markham et al., 2010; Duffy et al., 2012; Dowling et al., 2014), the model was furtherer adapted to account for dependencies among cells in the same generation, that equals the number of divisions from the progenitor (Duffy and Subramanian, 2009). Stochastic competition from the Cyton model recapitulates the sibling correlations, as shown by Duffy et al. (2012) with time lapse microscopy of *in vitro* B cells.

The technique of time lapse microscopy consists of a sequence of microscope images taken at a distance of minutes from one another, so that its accelerated reproduction results in a video of the process. Applied to lymphocytes, this method can track offspring from the same progenitor that progresses through the generations, recording birth, division, and death times of the cells and their parental relations. In combination with fluorescent markers, that bind with determined molecules to allow the recording of cellular phenotypic information, time lapse microscopy enables the observation of time and type of lineage commitment. Additional data can be obtained using cells from genetically modified mice. Of note, we mention the Fluorescent, Ubiquitination-based Cell Cycle Indicator (FUCCI) system (Sakaue-Sawano et al., 2008), which is a reporter for cell cycle stages. Cells from FUCCI mouse express red fluorescent protein when in resting phase (G_0 and G_1), while a green fluorescent protein is produced during the cycling phase (S, G_2 and M), indicating that the cell is preparing to undergo mitosis.

Although time lapse microscopy provides data that are not obtainable from other methods, this technique has limits. If the capture rate of the frames is too low, similar events may not be distinguished from others, such as in the case when two co-cultured cells divide between two consecutive frame-shots thus losing the familial separation of their

progeny. If the frame rate is too high, instead, it may negatively impact on the cellular survival due to photobleaching. Moreover, tracking is easily lost due to high motility of cells (Zaretsky et al., 2012; Polonsky et al., 2016) or the formation of three-dimensional structures occurs, which is a feature common to several systems. For example, proliferating B cells, under certain conditions, are characterised by homotypic adhesion (Klaus et al., 1994), that is the formation of clusters between cells from the same type, which can hide some cells from the microscope view. This was the case in Duffy et al. (2012), where the experimental protocol was adjusted by sorting only one cell from generation 0, 2, 4 or 6 per culture well, thus limiting the analysis of siblings in generations 1, 3, 5 and 7, as a maximum of two cells could be tracked in the same time and place, due to cellular adhesion.

Although, quantitative measurements from time lapse microscopy are possible, intravitally, for *in vivo* systems at the cost of a complex set-up with a restricted recording area and time (Hawkins et al., 2016). To cope with these difficulties and the remarkable processing workload required (which may consist in the manual annotation of the frames collected), several software solutions for automatised cell tracking have been developed (Rieger et al., 2009; Kan et al., 2011; Pham et al., 2013; Shimoni et al., 2013; Chakravorty et al., 2014; Mankowski et al., 2015). Still, thanks to time lapse microscopy that synchrony in activated lymphocyte families was shown for the first time (Hawkins et al., 2009; Markham et al., 2010; Duffy et al., 2012; Dowling et al., 2014).

Other methods for clonal labelling, that allows the tracking even for subsequent antigen encounters, were employed by Gerlach et al. (2013) and Buchholz et al. (2013) to study $CD8^+$ T cells *in vivo*. Gerlach et al. (2013) employed a cellular DNA-barcoding technique to create T-cell progenitors with different genetic tags that are passed on to their progeny. This enabled the separation of each of the families and the estimation of their clonal size through barcode sequencing. Buchholz et al. (2013), instead, used cells from mice that were genetically modified to allow for the expression of congenic markers, up to eight distinguishable combinations. By adoptive cell transfer of eight cells, one from each strain, into a ninth genetically different mouse, it was possible to track the size of their progenies by sampling from the recipient mouse at a given time after the immune response. Both these studies reported that the number of cells per family is heterogeneous, as few "giant" clones composed the majority of the immune response.

Moreover, in Buchholz et al. (2013) the analysis was extended to study the emergence of effector and memory cellular subsets within the proliferating population, suggesting that the memory compartment arises earlier than the effector one. Opposite conclusions

were reached in another study, which supports the hypothesis that memory cells appears after the effector ones (Kinjyo et al., 2015). This controversy points out that there is no consensus yet concerning how and when T cells differentiate to build the memory and effector pools.

The difficulties arise because, contrary to a B cell, the functional class of a T cell cannot be precisely determined from its phenotype, which consists in a set of measurable traits, including concentration of surface markers and transcription factors. The latter are proteins that promote or inhibit gene expression by binding to the associated to the gene DNA. Among these, we mention: Killer-cell-lectin-like-receptor-G1 (KLRG1) (Voehringer et al., 2001), T box transcription factor (T-bet) (Xin et al., 2016), and PR domain zinc finger protein 1 (Blimp-1) (Kallies et al., 2009; Rutishauser et al., 2009; Xin et al., 2016), that are highly expressed in effector cells; Eomesodermin (Eomes) (Banerjee et al., 2010), L-selectin (CD62L) (Sallusto et al., 1999; Buchholz et al., 2013; Gerlach et al., 2013), CD27 (Hikono et al., 2007), CXCR3 (Groom and Luster, 2011), and Bach2 (Sidwell and Kallies, 2016), that are highly expressed in memory cells.

Depending on the combined expression levels of these proteins, numerous phenotype categorisations for effector and memory cells have been defined, but their value is considered more semantic rather than functional (Mahnke et al., 2013). Some groups proposed that T-cell differentiation occurs as a slow and gradual commitment, which can be regulated by the initial stimulation strength and environmental cues (Kaech et al., 2002; Sallusto et al., 2004; Gerlach et al., 2011; Kaech and Cui, 2012), and that may be related to the activity of key transcription factor, such as Interferon-regulatory factor IRF4 (Man and Kallies, 2015).

Unravelling the problem of differentiation and heterogeneity in $CD8^+$ T cells is furtherer hindered, since multiple sources of diversifications have been identified to impact the phenotype "before, during, and after the first T cell division" (Lemaître et al., 2013), thus confounding the contribution to heterogeneity from each stage. To overcome this difficulty, one way is to study this problem on a reduced model where signal delivery is highly controlled, so as to render the investigation of heterogeneity more tractable by separating the early phases of T-cell immunity.

This methodology was employed in the experiments analysed in the present dissertation, where we focus on the effects of activation and costimulatory signals to heterogeneity. For this reason, in the next section, we provide a brief summary of the "Two-signal theory" for lymphocyte activation, reviewed in Baxter and Hodgkin (2002), and discuss the findings of Marchingo et al. (2014) on this topic, which are seminal to our work.

1.6 Theories of T-cell activation and flow cytometry

Under Burnet's clonal selection theory, one of the keys for immunity specificity and tolerance resides in the mechanism for lymphocyte activation. Late models for activation, i.e. Janeway (1989) and Matzinger (1994), postulated that two stimuli are required to trigger the immune response of naive lymphocytes: the antigenic signal, and a costimulatory one as delivered, for example, by B7 molecules upon contact with APC (Baxter and Hodgkin, 2002). Additional costimuli, such as inflammatory signals, could also be provided from cytokines, which are typically localised near a site of infection, and have been shown to impact the proportion of naive cells that commit at least one division (Voisinne et al., 2015). This two-signal theory of activation, however, has been contradicted from experimental evidence where *in vitro* T cells proliferated when cultured with anti-CD3, a ligand for CD3 (Gett and Hodgkin, 2000). In fact, CD3 is a protein complex that associates with the TCR and participates to the transduction of the antigenic signal.

A different mechanism for activation was presented in Heinzel et al. (2017), where it was shown that stimulatory signals (antigenic, costimulatory, or inflammatory) induce the production of Myc, a cell-cycle-regulating protein, that acts on lymphocytes as a license for division, as long as its expression level is above a certain threshold. If initial stimulation is strong, Myc surpasses this threshold and the progenitor starts dividing. Over the time, Myc levels degrade at a given rate, the threshold is ultimately reached, and the offspring undergo quiescence. If, instead, initial stimulation is weak, Myc levels do not go beyond the threshold and the progenitor does not proliferate. Under this paradigm, the antigenic signal alone would suffice to initiate proliferation if the stimulus it provides is strong enough, which is the case when an antigen binds to its cognate TCR with high affinity (Hommel and Hodgkin, 2007).

The first quantitative study to investigate the effect of stimuli combination to clonal expansion was published by Marchingo et al. (2014). In that paper, *in vitro* and *in vivo* evidence showed that the integration of different signals impacts linearly on the statistic for the expansion of a population of $CD8^+$ T cells, taking into account the generation of the cells and thus termed mean division number. In fact, the overall progression of the population under analysis could be predicted by the sum of the contributions to the expansion of each stimulus present in the system condition. This result is central for the biological investigation of Chapter 2, where we study if the additivity of signals also emerges at the level of the clones. In the following, we detail the part of the experimental methods from Marchingo et al. (2014) that is relevant in the future chapters.

To reduce the possible variables, the cells used in that study were obtained from OT-I mice (Hogquist et al., 1994), which are genetically engineered so that all their T cells possess the same TCR, specific for the ovalbumin (OVA) peptide. In particular, OVA can present different variants, such as SIINFEKL (N4) and SIIQFEKL (Q4) that bind to the TCR from the OT-I with high and low affinity, respectively.

T lymphocytes were analysed with the technique of flow cytometry (Owen et al., 2013), a technology that allows the separation and measurement thousands of cells per second. As a single cell is collected and isolated, it is focused in a fluid stream that flows past laser beams. The cell is then shot by the lasers, scattering their light, which can then be recorded as a spectrum profile of various intensities. Based on this information, a flow cytometer can be programmed to sort cells with given specifications, in which case it is referred to as a Fluorescence Activated Cell Sorter (FACS), so to provide one (or many) subpopulation of cells in output.

In this way, several characteristics of a cell can be analysed. For example, the forward scatter (FSC) and the side scatter (SSC) of the light are respectively associated with cell size and granularity. Moreover, to measure other properties of the cell or the expression level of its molecules, it is possible to use particular fluorescence markers that emit light when excited by a laser, thus providing a measure for their related molecule.

For example, mice can be engineered so that their cells express these markers. This is the case for the FUCCI mouse (Sakaue-Sawano et al., 2008), the reporter for cell cycle, whose cells produce a red protein during the resting phase, while a green protein is expressed while cycling.

Other markers can be added in the system as ligand for specific receptors. For example, expression levels of CD25, a component of the membrane receptor for IL-2, can be measured through its ligand, anti-CD25, when this is coupled with a fluorescent protein.

Furthermore, with experimental manipulation, cells may also incorporate fluorescent molecules which could serve to label a particular subpopulation, or to flag a given feature. We mention 5-(and -6)-carboxyfluorescein diacetate succinimidyl ester (CFSE), a green fluorescent protein which can be cultured with cells so that they imbibe it and whose dilution across divisions has been used to report the generation in which a cell is found (Lyons and Parish, 1994; Gett and Hodgkin, 2000). Before the first division occurs, CFSE dye is added to the culture wells and is accumulated by the cells therein. From these, a subpopulation is sorted through the FACS machine for bearing a given concentration of CFSE and is then re-cultured. As about half of the dye concentration is inherited to the daughter cells from their mother, cells can be harvested at a later

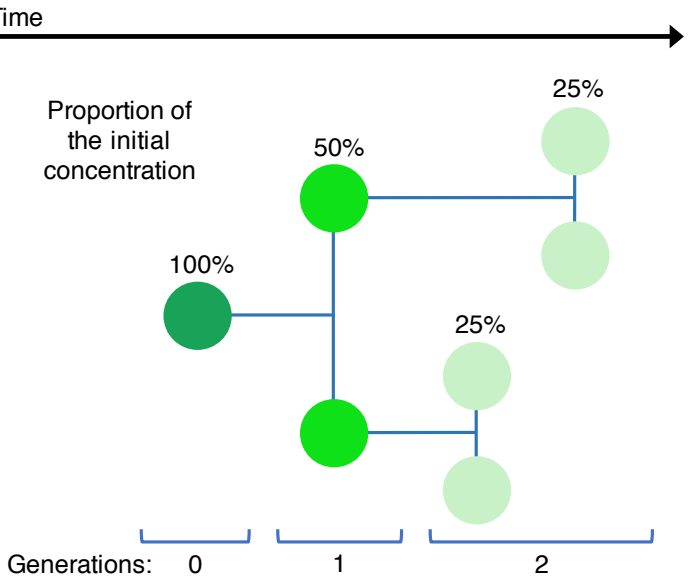

FIGURE 1.3: **Determining cellular generation by CFSE dilution method.** Provided the progenitor cell with a known quantity of fluorescence protein, such as the CFSE, the concentration inherited to the offspring is halved upon every division. This mechanism enables the identification of the generation in the cells recovered, by the amount of fluorescent protein they express.

time and analysed via flow cytometry to measure their CFSE expression level. In this way, their generation can be derived from how many times the CFSE concentration has been halved from the progenitor (see Fig. 1.3).

The determination of generations also involves the process of gating. This step consists in visualising the marker profile of the cells pooled, as recorded by the instruments, and set the values that isolate the different clusters of cells (e.g. those in generation zero, one, etc.), which are drawn to separate the peaks in the distribution of the expression level.

1.7 Thesis outline

In this thesis we will analyse the effect of stimulatory signals, alone and combined, on $CD8^+$ T-cell proliferation, in order to study the mechanisms of activation and division in T lymphocytes, and the properties of their clonal families.

In Chapter 2, we question the hypothesis of additive signal integration at the level of the clones, that was experimentally shown at the population level in Marchingo et al. (2014). To do so, we utilise data from a new clonal multiplex assay, which is based on flow cytometry and multiple dyes dilution so to enable the recording of cellular

generation and familial membership with high-throughput. The work in this Chapter was published in Nature Communications (Marchingo, Prevedello et al., 2016).

To analyse the data in output from this method, in Chapter 3 we introduce a mathematical framework for the structure of clones and study the effect of sampling on the recovered data. As the cells from the same clone are arranged in a family-tree, which is an object from the graph theory, we define a new operation of tree addition to represent the signal integration of the clones. The work in this Chapter is a mathematical expansion of the version used in Marchingo, Prevedello et al., (2016).

A novel statistic, and hypothesis tests based on it, is designed in Chapter 4 to assess the dependence between the contribution, to clonal expansion, from different costimulatory signals, and was put to use in Marchingo, Prevedello et al., (2016). This problem reduces to comparing the equality in distribution between the sums of independent random variables. From this setting, a statistic, based on the operation of discrete convolution, emerges as a consequence of a nonparametric, maximum likelihood approach. The development of this statistical procedure has been submitted for publication in a statistics journal.

Finally, in Chapter 5, the multiplex clonal assay, from Chapter 2, is expanded to include the measurement of expression level from single-cell markers. We complement the experimental protocol with the statistical methodology, based on permutation tests, to evaluate the presence of significant structures in the data, such as clonal, generational or environmental membership. The work in this Chapter was published in the Journal of Immunology (Horton, Prevedello et al., 2018).

Chapter 2

Independent signal integration regulates T-cell clonal division fate

2.1 Abstract

When stimulated with antigenic and costimulatory signals, T cells undergo a typical response pattern of activation, expansion, quiescence and contraction. Recent studies have shown that the heterogeneity observed at the population level is composed by clonal families of highly different size, thus suggesting the importance of single-cell studies. In this chapter, we present the work from Marchingo, Prevedello et al., (2016) accomplished in collaboration with our partners in Prof. Philip Hodgkin's lab at Walter Eliza Hall Institute (WEHI), where a novel protocol for clonal data was implemented. From the paper, one major experiment, in the main text, and its repeat, in the supplementary information, were performed with such a method under controlled stimulation environment. Here we report their findings alongside each other focusing on the data analysis which we had the greatest contribution to, while complementary experimental details are deferred to Appendix A. From these experiments, cells from the same clone stopped dividing in the same or two consecutive generations, leading to an heterogeneous burst size at the population level that was determined by inter-clonal variability of the generation reached by the clone. Comparing the familial expansion achieved under different signals, we found that the costimulatory contributions were integrated independently of each other. This evidence suggests that substantial heterogeneity can be traced back to progenitor antigen interaction or costimulatory receptors variability.

2.2 Introduction

To provide protection against pathogens, a portion of the T-cell pool that is specific to the threat is activated and expands into a larger population of effector cells that kills the infected cells (Moon et al., 2007). This population is created by the contribution of each rare pathogen-specific clone and, under a given stimulation, the total population size is broadly reproducible. Although the response is robust, recent studies have shown that individual families present highly heterogeneous properties such as clonal size or phenotype (Stemberger et al., 2007; Gerlach et al., 2010; Buchholz et al., 2013; Gerlach et al., 2013; Plumlee et al., 2013; Tubo et al., 2013). This suggests that the analysis at the single-cell level is necessary to cast light on the underling mechanism of T-cell regulation.

In this chapter, the question is addressed concerning what is the main driver of cellular heterogeneity, as at present it is unclear if the main factor is a stochastic or deterministic programme or whether it is inherited from the ancestor cell or, instead, triggered in each offspring cell (Rohr et al., 2014). To investigate the fundamental biology in the absence of confounding factors, we employ an highly reduced system where activation and co-stimulatory signals delivery is controlled and endogenous signals are blocked.

Previous studies showed that T cells with T-cell receptors (TCRs) specific for the same antigen produce an heterogeneous population of cells and become quiescent within a wide range of generations (Zehn et al., 2009; Marchingo et al., 2014; Starbeck-Miller et al., 2014), even under highly controlled, homogeneous stimulation (Marchingo et al., 2014). In recent results, our collaborators investigated how such heterogeneity arises at the clonal level through the analysis of Division Destiny (DD) (Turner et al., 2008; Hawkins et al., 2009; Marchingo et al., 2014), a term coined to indicate the generation in which a lymphocyte returns to quiescence after mitogenic stimulus.

To explain how the reported heterogeneity in DD arises at the population level, we consider two distinct clonal hypotheses (Fig. 2.1a): DD is determined by highly concordant families that cease to divide in a narrow range of generations (Fig. 2.1b top panel); DD is the result of discordant families whose cells undergo quiescence in very diverse generations (Fig. 2.1b bottom panel). The first scenario would suggest that DD is a clonal feature inherited and conserved through the daughter cells from their initial naive ancestor. The second scenario may instead emerge from asymmetric cell division (ACD) mechanism (Chang et al., 2007; Reiner and Adams, 2014) or by stochastic regulation (Duffy and Hodgkin, 2012; Hodgkin et al., 2014). As indicated in Fig. 2.1, studies with population data cannot identify if families are concordant or discordant and therefore provide a deeper intuition on how clonal expansions is regulated.

Whichever the generational profile type is, it must ultimately agree with the program that is initiated in the founder cell by activation stimuli. As the kind and the strength of certain stimulatory signals regulate the expansion, Marchingo et al. (2014) showed that the combination of signals determines a population DD distribution with a mean and a variance that is the sum of each signal contribution, thus visualised in Fig. 2.1c. This result suggests the hypothesis that signals are integrated independently of one another into the clone's expansion programme, which leads to the problem of how the DD profiles from different stimulatory contributions should be interlaced together (Fig. 2.1d). We reason that, under independent integration of two signals, the family tree generated by their combination must be regular if the families generated by each signal alone is concordant as well (Fig. 2.1d, top panel). If one of these trees is not concordant, it is unclear which tree should result from their addition and what the notion of addition should be used (Fig. 2.1d, bottom panel).

In the following sections, we address how single clones participates to the population DD and how signal integration occurs consistently with the clonal DD distribution. We investigate these questions using a novel multiplex clonal division-tracking assay, developed with our collaborators, which employs a combination of division tracking dyes to identify the generation and the familial membership of each cell. Our analysis of the data from this new method, applied to $CD8^+$ T-cell cultures by our partners, indicates that under controlled conditions these clones present strongly concordant generations when quiescent, so that clone-to-clone variation is the key determinant of population heterogeneous DD. This supports familial programming and is inconsistent with the most widespread mathematical models, based on branching processes, which assume related cells behave independently (Harris, 1964; Kimmel and Axelrod, 2002).

To address the question of whether costimulatory signals have a stochastic effect that is independently integrated during T-cell activation, we develop a new operation for the interlacement of family trees, a mathematical exposition of which is postponed to Section 3.4 in Chapter 3. We report that, even if a consequence of the addition of stochastic contributions, the population-level outcome is reproducible and the resulting heterogeneity can be substantially explained by stochastic antigen interaction and initial receptor sensitivity.

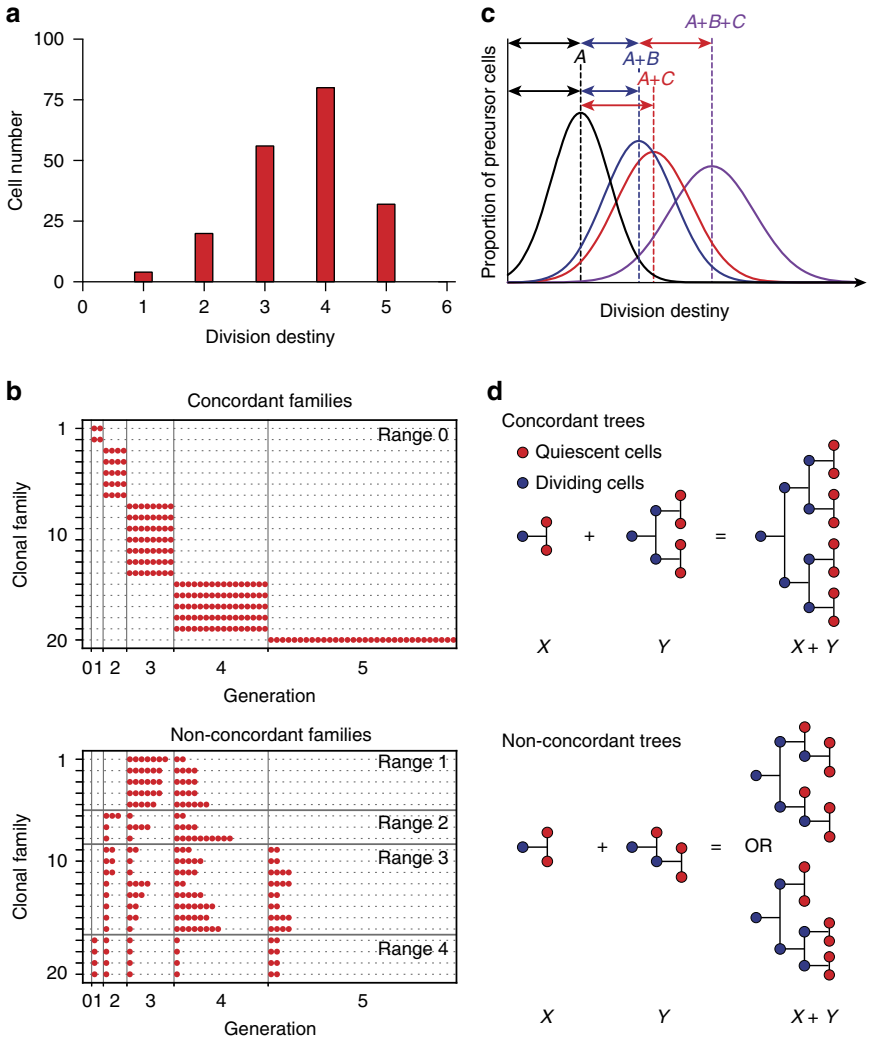

FIGURE 2.1: **How is T-cell division destiny (DD) regulated at a clonal level?** [Corresponding to Figure 1 from Marchingo, Prevedello et al., (2016)] Hypothetical data. (a) When apparently identical T cells are stimulated, they proliferate to different extents, resulting in the population of progeny cells returning to quiescence across multiple generations. (b) Two distinct clonal family DD behaviours are consistent with the data in (a): a highly concordant clonal DD (top panel) or a highly discordant family DD (bottom panel). Each row represents a single clone, with circles showing progeny cells reaching DD per generation. Range is the difference between maximum and minimum generation number. (c) Signals affecting T-cell DD have been shown to add together at the population level (Marchingo et al., 2014). (d) If signal effects are independent, addition of concordant trees must result in a tree that is also concordant (top panel). If the families are discordant, it is unclear which tree should result from their addition (bottom panel and Fig. 2.9).

2.3 Results

2.3.1 A novel multiplex assay to measure clonal division

In order to study the regulation of T-cell DD at the clonal level, our collaborators employed an *in vitro* experimental method that can identify the stage of proliferation from several naive T cells. In this section, we report the experimental protocol as our partners implemented at WEHI laboratories. They labelled lymphocytes with distinct combinations of division tracking dyes 5-(and 6)-carboxyfluorescein diacetate succinimidyl ester (CFSE), CellTrace Violet (CTV) and Cell Proliferation Dye eFluor670 (CPD) at different concentrations as illustrated in Fig. 2.2a,b (Lyons and Parish, 1994; Quah and Parish, 2012), enabling the co-culture of 10 different clones in the same well. This multiplex assay was used in combination with the same *in vitro* stimulatory conditions that was previously applied to study DD at the population level (Marchingo et al., 2014).

OT-I $CD8^+$ T cells, which recognize SIINFEKL (N4) peptide presented on H2Kb, and deficient for the pro-apoptotic protein Bim (OT-I/$Bcl2l11^{-/-}$) were purified, labelled with the division tracking dye multiplex and stimulated by peptide self-presentation in the presence of anti-mouse Interleukin-2 (IL-2) blocking antibody (clone S4B6; Fig. 2.2a-c). Bim-deficiency enhanced cell survival without altering DD and addition of anti-mouse IL-2 blocking antibody limited the autocrine IL-2 present in the culture, allowing T cells to reach DD within the range of division tracking dyes (Marchingo et al., 2014).

After 26 h (just prior to the first division) cells were sorted so that a single stimulated but undivided cell, identified by high forward scatter (FSC) fluorescence and undiluted division tracking dye, from each fluorescently distinct population was seeded per well. Cells were returned to culture, without peptide but with S4B6, until analysis by flow cytometry at 54, 62 and 72 h post stimulation (Fig. 2.2d-e), capturing times when most cells were reaching DD without considerable cell death having occurred (Marchingo et al., 2014).

Clonal family division fate from each labelling configuration was identified using the fluorescent gating scheme outlined in Fig. 2.2f. Small cell size was used to indicate return to quiescence, as previously defined by correlation with a cell cycle reporter of G_0 (Hawkins et al., 2009; Marchingo et al., 2014; Kinjyo et al., 2015); see Fig. A.1 from Appendix A Section A.1.6 for further details.

Data from the multiplex clonal assay for OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells, as described, are shown later in Fig. 2.3a-c and Fig. 2.4. These cells were stimulated by N4,

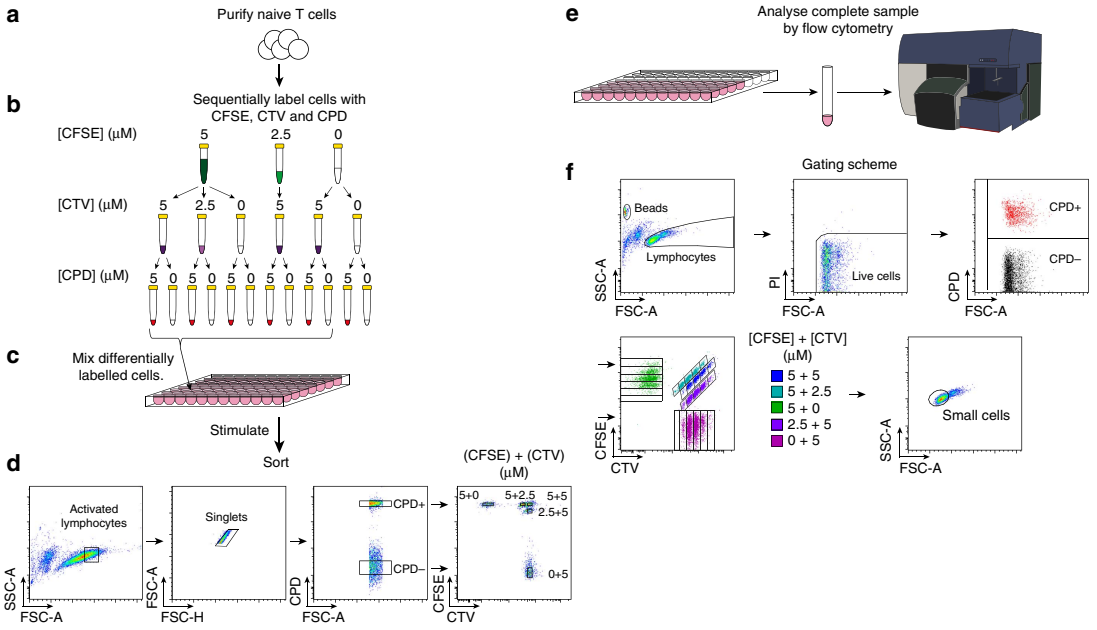

FIGURE 2.2: **A novel high-throughput clonal assay to measure T-cell division fate.** [Corresponding to Figure 2 from Marchingo, Prevedello et al., (2016)] (a) OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were purified and (b) labelled sequentially with different combinations and concentrations of CFSE, CTV and CPD. (c) T cells from the 10 different labelling configurations indicated were mixed together and stimulated with N4 peptide $\pm\alpha$CD28 (2 μg ml^{-1}). Between 500 and 2,000 cells from each of all 12 labelling configurations were also cultured separately to use as compensation and gating controls. (d) Just prior to first division (26 h) a single cell per labelling configuration from each stimulation condition was sorted into new wells and cultured $\pm$hIL-2 (1 U ml^{-1}). Thus there were four stimulation conditions in total: N4-only, N4+αCD28, N4+IL-2, N4+αCD28+IL-2. (e) 7,500 beads and propidium iodide (PI) were added per well before analysis to estimate sample recovery and detect dead cells. Cells were sampled at 54, 62 and 72 h post stimulation and analysed through flow cytometry. (f) Gates for data analysis were created using control populations at each time-point then applied to the clonal samples.

anti-CD28 (αCD28) and IL-2, the latter added as human IL-2 (hIL-2) to overcome blocking by S4B6 present in the culture wells. In particular, S4B6 antibody is added in the culture wells to exclude the contribution to DD of autocrine IL-2, i.e. sensed by the same cell that produces it. The release of this cytokine by activated CD8^{+} T cells would confound the precise assessment of the inflammatory signal contribution to the expansion, initially added in known quantity as hIL-2. All these molecules (N4, αCD28, hIL-2 and S4B6) are provided in precise quantities (see Section A.1.4 in Appendix A) allowing the measurement and the comparison of their contribution to the clonal expansion.

A culture of 500-2,000 cells served as population control with cells from 8 division dye labelling combinations (Fig. 2.3a) that were harvested 72h after their stimulation. This time point is chosen so that most cells have returned to quiescence, before the contraction phase is of major impact (Marchingo et al., 2014). Here the use of

OT-I/$Bcl2l11^{-/-}$ system is of essence, as T cells from these mice have reduced pro-apoptotic molecule Bim and present enhanced survival time without other functions being affected from the disabling of the *Bcl2l*11 gene (Marchingo et al., 2014), thus increasing the chance of finding living cells in DD at later times. The control population in Fig. 2.3a was used to determine the generation-gate from each one of the dye-combinations, as summarised in Fig. 2.2. The control gating was then overlaid on the data from single co-cultured wells harvested at the same time (Fig. 2.3b) in order to determine the generation and the clonal membership of the recovered cells. The aim of this procedure is to obtain data at the clonal level which we subsequently analyse to investigate familial concordance under specific culture conditions.

2.3.2 T-cell proliferation is synchronous

We analyse the clones from Fig. 2.3c to assess their proliferation synchronicity. From the initial culture of 224 clones, 171 (i.e. 76%) were recovered across three time points, and were found having at least one cell. From 42% of these, all the daughter cells were sampled. The recovery proportions from the multiplex assay are comparable to those from time-lapse microscopy method of non-adherent cells (Hawkins et al., 2009). Also, the distributions of cells across generations between clonal progeny and population control are similar (Fig. 2.5).

Most of the collected T-cell families were concordant (Fig. 2.3c, Fig. 2.4). Clones sampled at 54h presented many cells that were still dividing (blue), but when harvested at 62h and 72h, the majority of the cells were in a quiescent state (red). Across time points, 85% of the clones were found with their offspring in the same generation, while in the remaining families their cells are situated in contiguous generations (Fig. 2.3c). Mixed-type families were detected with cells in the same generation or with dividing cells in the generation previous to the quiescent cells, indicating that a possible cause of discordance may arise from small differences in division times from an otherwise synchronous clone (as in Fig. 2.1b upper panel). Synchronous proliferation was also observed when the antigenic signal persisted in the culture (Fig. 2.6). The strong concordance is consistent with past experiments, where sibling and cousin cells were analysed for their division times (Hawkins et al., 2009; Duffy et al., 2012; Dowling et al., 2014; Kinjyo et al., 2015).

2.3.3 Division fate is concordant in response to different stimuli

As previously shown in Marchingo et al. (2014), the present T-cell system displays a population level DD that is highly affected by the kind and the combination of

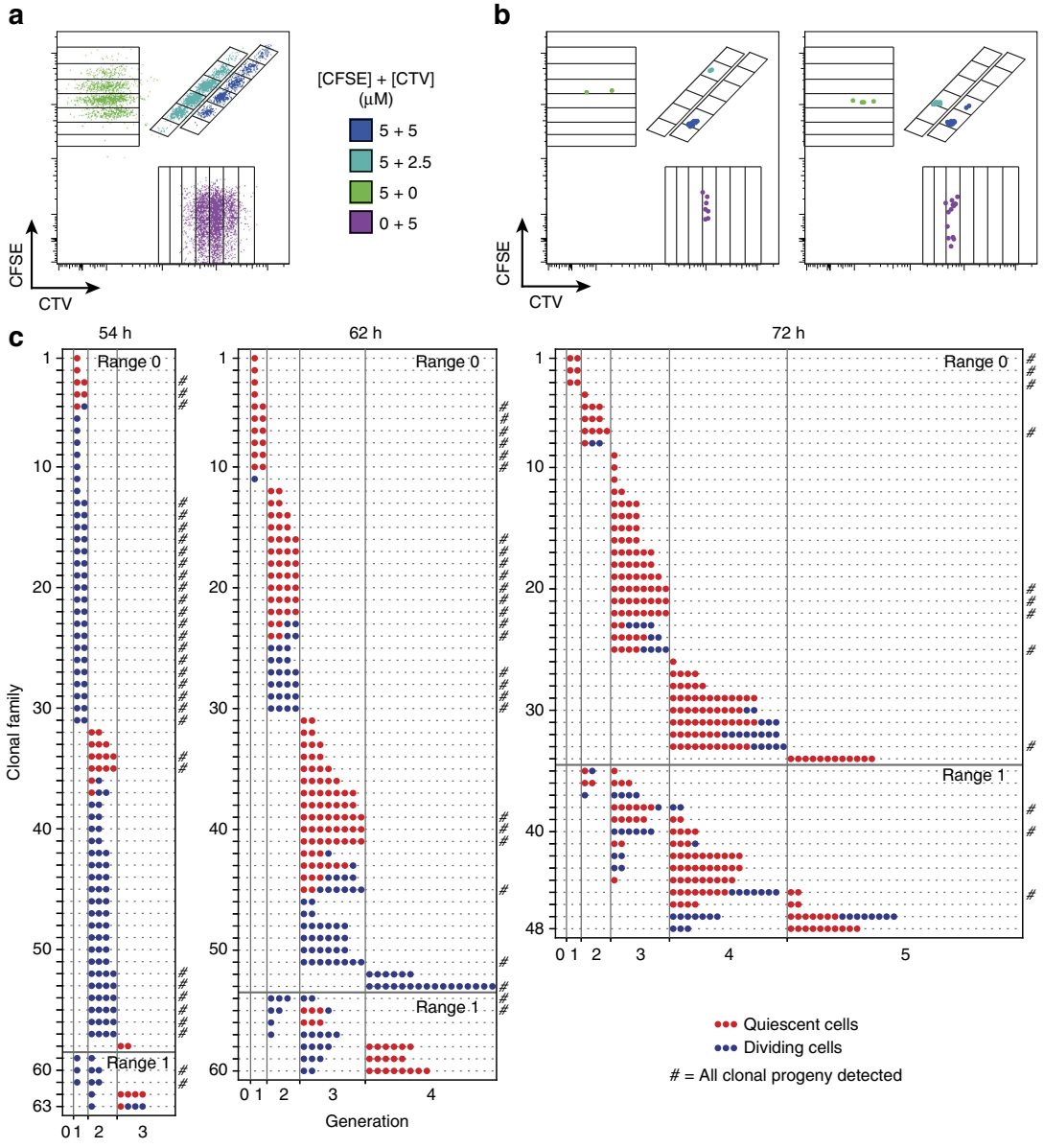

FIGURE 2.3: **Clonal T-cell family proliferation is synchronized.** [Corresponding to Figure 3 from Marchingo, Prevedello et al., (2016)] OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were processed, stimulated and analysed as described in Fig. 2.2. All cultures contained S4B6 (25 mg ml^{-1}). (a) N4+αCD28+IL-2 stimulated population of control cells labelled with CTV and CFSE to distinguish generation number for four distinct labelling configurations. Cells were separated into CPD^+ and CPD^-, allowing division tracking in eight populations per well. (b) Examples of clonal progeny detected in individual wells. Example data shown for 72 h time point. (c) Generation number of progeny cells detected from individual clonal families at each time point from N4+αCD28+IL-2 stimulation condition. Progeny cells were classified as quiescent based upon small cell size (refer to Section A.1.6). Range is the difference between maximum and minimum generation number. The symbol # at the end of a line denotes clones where all progeny cells were detected. Founder cell input after sorting was 80, 80 and 64 for data from 54, 62 and 72 h respectively. Results from a second independent experiment are shown in Fig. 2.4.

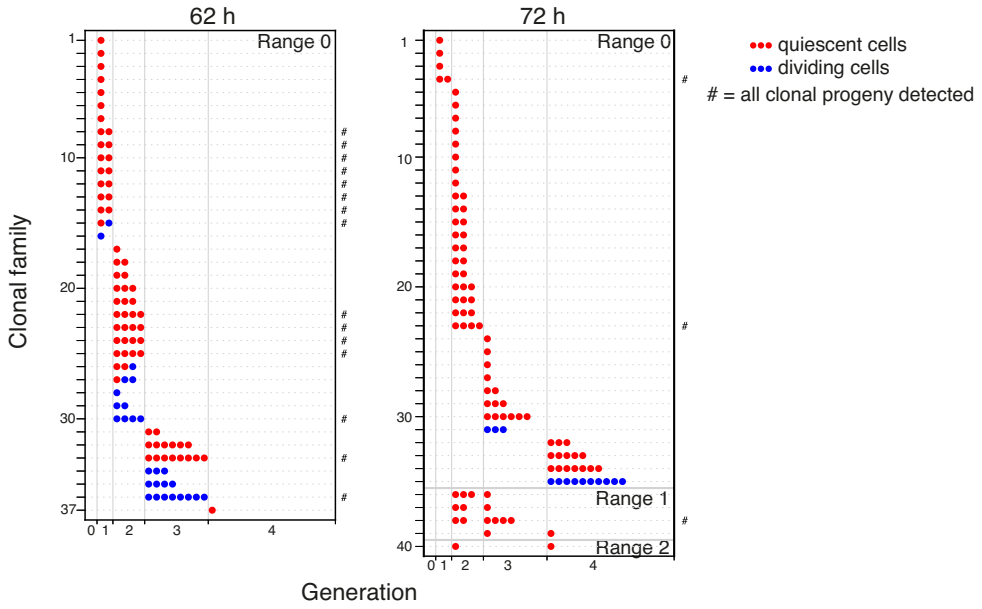

FIGURE 2.4: **Clonal T-cell family proliferation is synchronized, experimental repeat.** [Corresponding to Supplementary Figure 3 from Marchingo, Prevedello et al., (2016)] Repeat of the experience as described in Fig. 2.3 with the exception that cells were analysed only at 62 and 72 h. Founder cell input after sorting was 168 and 168 for data from 62 and 72 hours respectively. Note that the lower clonal recovery in this experiment is likely attributable to a longer time spent out of culture during cell sorting, which reduces clone viability.

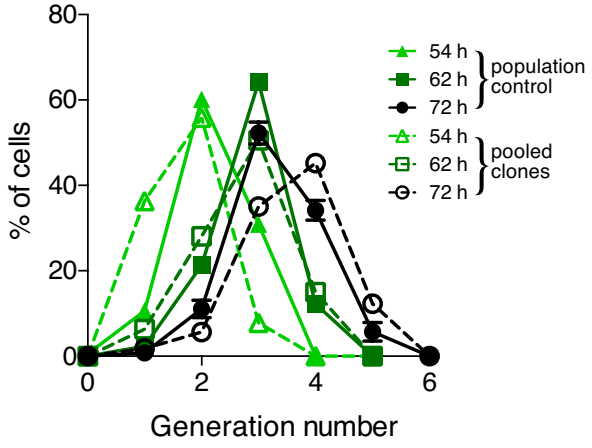

FIGURE 2.5: **Pooled clonal proliferation data recapitulates population response.** [Corresponding to Supplementary Figure 4 from Marchingo, Prevedello et al., (2016)] OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were isolated and stimulated with N4 peptide (0.01 μg ml^{-1}), αCD28 (2 μg ml^{-1}) and hIL-2 (1 U ml^{-1}) in the presence of S4B6 (25 μg ml^{-1}) as described in Fig. 2.2a-e and gated as outlined in Fig. 2.2f. At each time point the total progeny cell number detected per generation was pooled for all clones, the percentage of progeny cells per generation calculated (dotted lines) and compared to the percentage cells per generation detected in the population control (solid lines) (based on CTV dilution in 0 μM CFSE + 5 μM CTV ± 5 μM CPD labelling conditions). Graphs are representative of 2 independent experiments. Mean ± s.e.m. from triplicate culture wells.

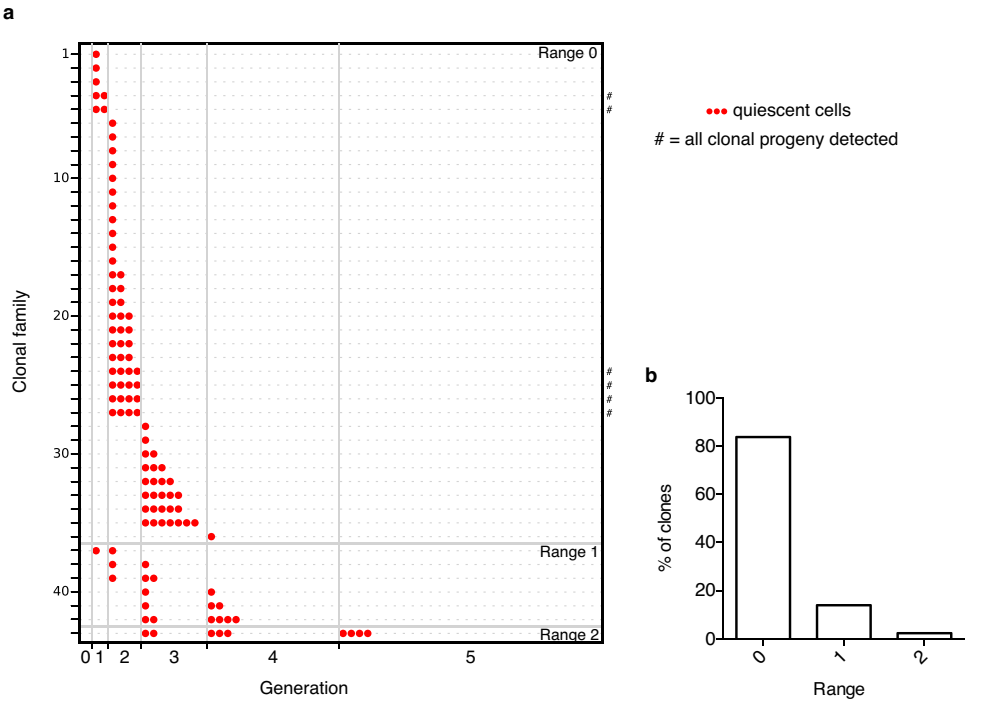

FIGURE 2.6: **Clonal T cell division destiny is concordant when peptide persists.** [Corresponding to Supplementary Figure 5 from Marchingo, Prevedello et al., (2016)] OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were isolated and labelled with division tracking dye combinations as outlined in Fig. 2.2a-c. Cells were stimulated with N4 peptide (0.01 μg ml^{-1}) in the presence of S4B6 (25 μg ml^{-1}) for 50.5, 62.5 or 72.5 h before analysis by flow cytometry as outlined in Fig. 2.2e. Cells were gated according to Fig. 2.2f only including data from the clones that were CPD^+, due to background autofluorescence into the CTV and CFSE channels by the unlabelled cells. (a) The generation in which progeny cells were detected, from clones in which all the progeny cells were quiescent. (b) Percentage of clones vs. range (i.e. maxDD − minDD). Data from one experiment.

stimulatory signals received. To explore this effect, the multiplex assay was applied to different conditions: N4 alone, N4 + αCD28, N4 + IL-2, and N4 + αCD28 + IL-2, where IL-2 was added as hIL-2. In order to investigate the DD at the clonal level for each condition, data from quiescent-only families were pooled across 54, 62 and 72 h time points, thus excluding clones recovered with at least one dividing cell for not being fully expanded (Fig. 2.7a, Fig. 2.8a). Moreover, families composed of one cell in generation 0 were also excluded as their lack of division could have been the result of stimulatory failure.

To quantify the degree of clonal concordance, we defined the difference between maximum and the minimum clonal DD and termed this measure range (see 3.2.3 in Chapter 3). For perfectly concordant families the range is zero, and null range was detected for most of the observed clones (2.7b and Fig. 2.8b).

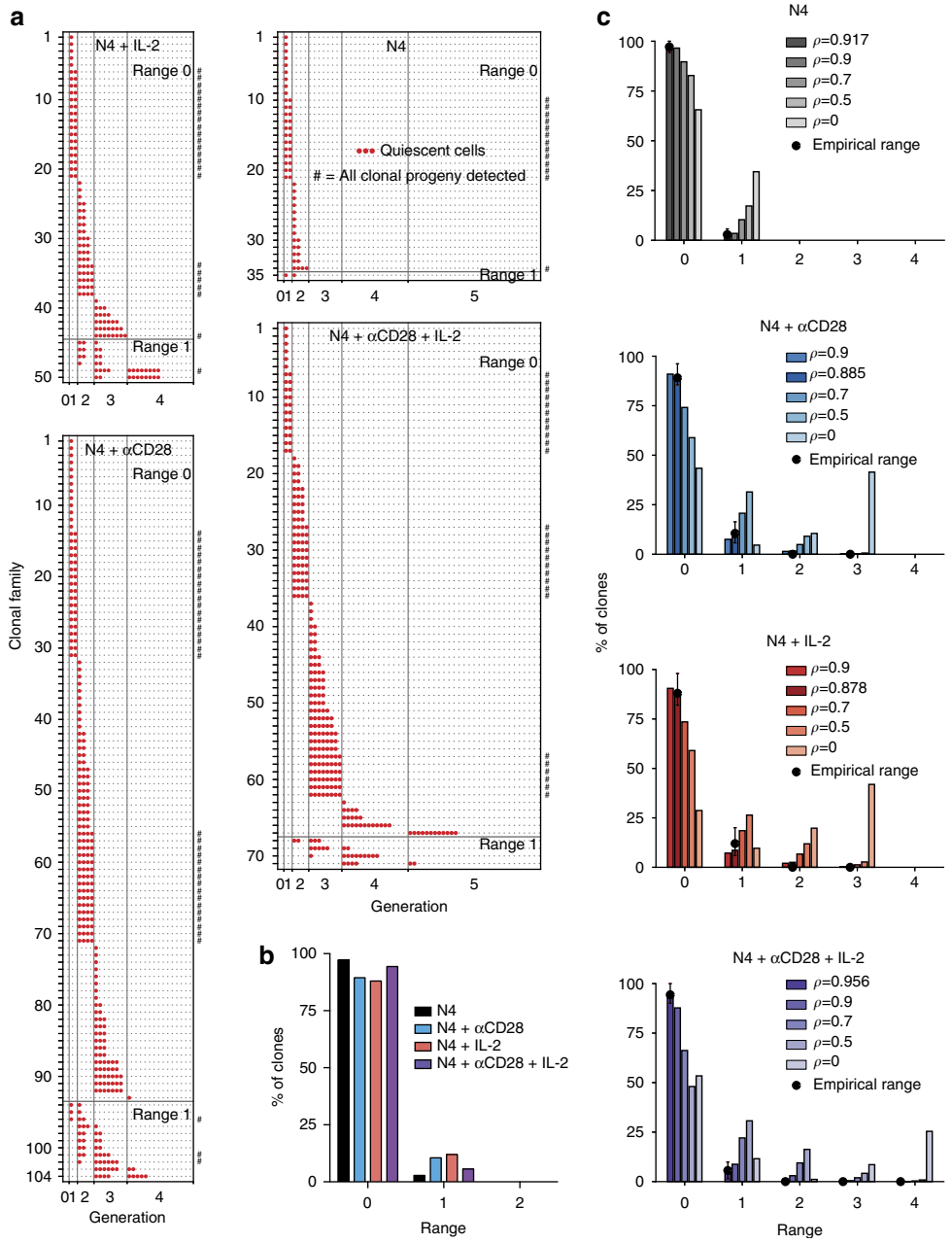

FIGURE 2.7: **Clonal family DD is highly concordant.** [Corresponding to Figure 4 from Marchingo, Prevedello et al., (2016)] OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells labelled with a division tracking dye multiplex were stimulated with N4 peptide ±αCD28 (2 μg ml^{-1}) for 26 h, sorted for one clone per labelling configuration per new well then cultured ±hIL-2 (1 U ml^{-1}) as described in Fig. 2.2. All cultures contained S4B6 (25 μg ml^{-1}). (a) Generation number in which progeny cells reached DD. Data pooled from 54, 62 and 72 h from families where all detected progeny were quiescent. Founder cell input after sorting was 96, 224, 96 and 224 for N4, N4+αCD28, N4+IL-2 and N4+αCD28+IL-2, respectively. (b) Percentage of clones with concordant (range = 0) or discordant (range > 0) DD. (c) To quantitatively question the level of familial correlation in DD required to explain the range data in (b), a mathematical model was constructed and parametrized by the data and pairwise correlation, ρ, in DD fate (see Section 3.3.2 in Chapter 3). The empirical distribution of range is shown for each condition (black dots within 95% confidence intervals, see Section 2.3.4), in addition to the range distribution for different values of ρ, including the per-condition best-fit. Results from a second independent experiment are shown in Fig. 2.8.

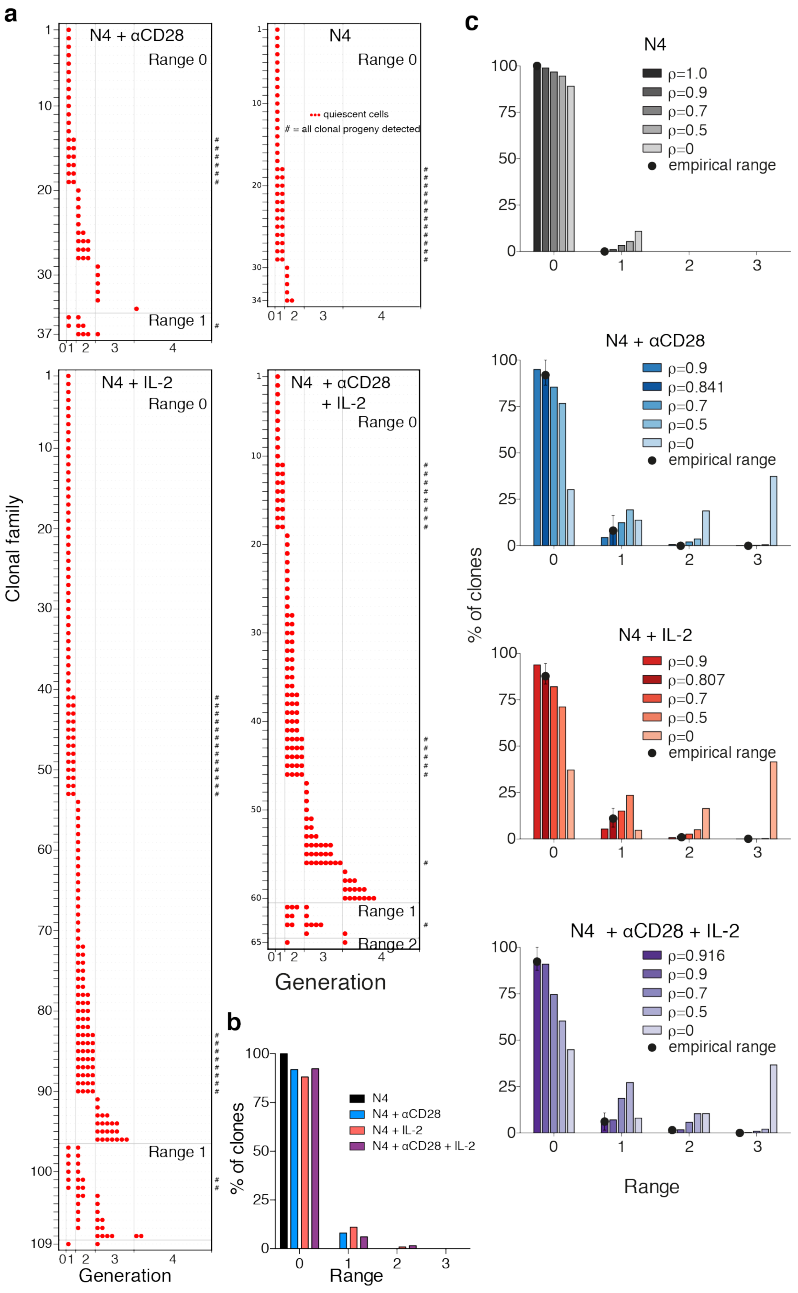

FIGURE 2.8: **Clonal family DD is highly concordant, experimental repeat** [Corresponding to Supplementary Figure 6 from Marchingo, Prevedello et al., (2016)] Repeat of the experience as described in Fig. 2.7, with the exception that cells were analysed only at 62 and 72 h. Founder cell input after sorting was 168 for all conditions. Condition N4+αCD28+IL-2 is pooled from families in Fig. 2.4 where all detected progeny were quiescent.

2.3.4 Clonal family DD is concordant

T cells originated from the same clone looked highly correlated in their DD (Fig. 2.7b). This evidence, however, was influenced by partial recovery as a discordant family may look concordant if some of its cells are not sampled. To take this effect into account for DD analysis, we developed a stochastic model of quiescence cell range, introducing a parameter for the correlation between cells within a family. We defer to Section 3.3.2 in Chapter 3 for a formal description of such model.

For a fixed conditions, families were selected if all their cells had undergone DD. From this pool, the proportion of cohort that reached generation i and divided to generation $i+1$ was estimated as

$$\hat{p}_i = \frac{\sum_{j \geq i+1} 2^{-j} n_j}{\sum_{j \geq i} 2^{-j} n_j}, \tag{2.1}$$

where n_i is the number of cells in generation i from the pool.

Assuming that each cell reverts to quiescence independently of the others, then the number of cells that divides to generation $i+1$ from i would be distributed as a binomial with parameter $\hat{p}_i$. To include the possibility that cell fate is correlated within the family, the binomial variable is replaced by a beta-binomial with probability of progression $\hat{p}_i$ and intra-class correlation $\rho \in [0, 1]$. The latter parameter modulates the extent of dependence across cells when deciding whether dividing to the next generation or becoming quiescent. In particular, for $\rho = 0$, we recapitulate the binomial behaviour described above. For $\rho = 1$, instead, cells in the same generation share a common fate and all divides, with probability $\hat{p}_i$, or all reach DD, with probability $1 - \hat{p}_i$. In this setting, the proportion of clonal cohort that divides from one generation to the subsequent, i.e. $\{\hat{p}_i \colon i = 0, \ldots, 6\}$, does not depend on the intra-clonal correlation ρ.

From the parameters $\{\hat{p}_i \colon i = 0, \ldots, 6\}$ and ρ, for each condition, we computed the distribution of DD configuration of a clonal family before its recovery. Cells must then be sampled, independently of their fate, with probability that is estimated either from the average proportion of beads recovered per well (Fig. 2.7c) or the average per well volume collected (Fig. 2.8c). In this way, we derived the distribution of the range for a sampled clone (see Section 3.3.2), from which the maximum likelihood estimate $\hat{\rho}$ of ρ was calculated.

The resulting correlation fits were high across all conditions, as $\hat{\rho} \geq 0.8$ in all cases, thus indicating that a mechanism where DD is regulated in each cell, independently of their clone, may be inappropriate. Moreover, the distributions from these estimate were comparable with the empirical distributions of the range data (Fig. 2.7c) and were included within their 95% confidence intervals (CIs). In order to produce these

intervals, we used the following standard bootstrap technique to create asymmetric 95% confidence intervals (CIs) (Beran, 1984).

For a given estimator $\hat{\theta}$ for a statistic $\theta \in \mathbb{R}$ of the data, this method consists in generating bootstrap datasets, by sampling the original data with replacement $K = 10,000$ times, and calculate the statistic of interest $\hat{\theta}$ on each bootstrapped dataset, to obtain $\{\hat{\theta}^*_1, \ldots, \hat{\theta}^*_K\}$. Then, the bias corrected bootstrap CI for $\hat{\theta}$ is defined as

$$\left[2\hat{\theta} - \theta^*_{(u)}, 2\hat{\theta} - \theta^*_{(l)}\right] \tag{2.2}$$

where $u = \lceil K\, 0.975 \rceil$ and $l = \lfloor K\, 0.025 \rfloor$ (Efron and Tibshirani, 1993). The procedure was applied in Fig. 2.7 and 2.8 to create CIs for θ being the probability mass function of range calculated in $k = 0, \ldots, 4$. We used the same technique in Fig. 2.10 and 2.11, where an additional step was necessary to calculate the CIs, as θ is the cumulative distribution function of the sum of two random variables X and Y, such as maxDD from (N4) and (N4+αCD28+IL-2). In this case, each bootstrap iteration consisted of sampling with replacement data from X and, independently, data from Y then following the same rationale as above.

2.3.5 Stimuli effects on clonal division fate add independently

Data in Fig. 2.7, from the multiplex assay with selected stimulatory signals, shows that T cells present a concordant clonal DD that is inherited from the founder to the offspring. We now investigate how different co-stimulations regulate such concordant DD. As reported in Marchingo et al. (2014), the conditions under analysis, when combined, displayed a linear additive effect to the mean and variance of the population DD: this suggests that the integration of each costimulatory signal may contribute independently to the clonal DD distribution.

To test this prediction quantitatively, we defined a novel class of operations between two family trees, that will be detailed in Section 3.4 of Chapter 3. As a consequence of the analysis therein, the DD configuration from the resulting can be deduced from the number of quiescence cells per generations of the trees that are combined (Fig. 2.1d, Fig. 2.9), regardless the operation, from the new class, considered. This property enabled us to study signal integration using multiplexed data, which provides the necessary clonal information of quiescence cell count per generation. Such a notion of tree addition is particularly suitable to study clones expanded under more than one mitogenic signal, as no assumption is made concerning the order of each stimulus contribution.

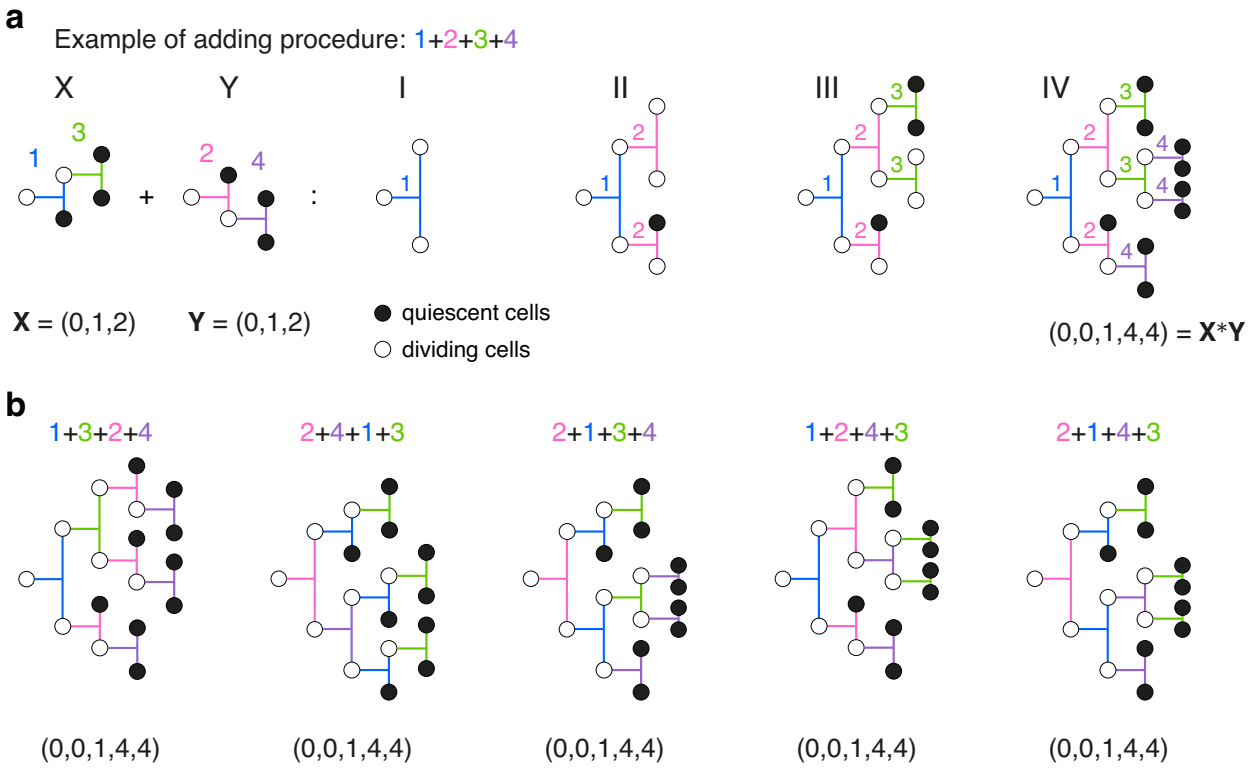

FIGURE 2.9: **Combinatorial summation of discordant family trees.** [Corresponding to Supplementary Figure 1 from Marchingo, Prevedello et al., (2016)] Example of two clonal family trees X and Y that contain one progeny cell reaching DD in generation 1 and two progeny cells reaching DD in generation 2, generating the vectors $\mathbf{X}$=(0,1,2) and $\mathbf{Y}$=(0,1,2). The open circles represent cells that will go on to divide further. The black circles represent cells that have reached DD. (a) Explicit construction of one possible addition that gives rise to one appropriate summed tree. Each clonal family tree can be broken up into multiple subtrees as specified by the different numbers and colours above (e.g. X is split into 1 and 3, blue and green respectively). The arrangement of these subtrees indicates the order of the contributions from X and Y when merged. Such order cannot violate the configuration from the original trees (e.g. subtree 1 must not descend from a copy of subtree 3, as only 3 descends from 1 in their original tree X). In the example, I-IV show the iterative appending of subtrees $1, 2, 3, 4$ in this order. All other possible rearrangements of these contributions are shown in (b). Strikingly, irrespective of the arrangement, the division in which progeny reach DD is unchanged in the consequent tree and produces a vector that is the discrete convolution between $\mathbf{X}$ and $\mathbf{Y}$, namely $\mathbf{X}*\mathbf{Y}$=(0,0,1,4,4). A formal derivation of this result is postponed to Section 3.4 of Chapter 3.

To describe the clonal participation to the overall expansion under each condition, we plotted the empirical cumulative distribution functions of maximum DD (maxDD) and cohort normalised mean DD (mDD) in Fig. 2.10a and b, respectively, whose formal definition will be postponed to Chapter 3. These measures of clonal expansion (see Section 3.2.3) are linear with respect to the operators for trees combination (see Section 3.4.3). Therefore, under independent signal integration, the distribution of maxDD or mDD, for trees expanded under two stimuli, would be characterised by the convolution of the distributions relative to each single stimulus.

Analogously, if the contributions of αCD28 and IL-2 on DD configuration were independent, even if one or the other were correlated to the contribution of N4, the convolution of the distributions of maxDD (or mDD) on (N4 + αCD28) and (N4 + IL-2) would correspond with the convolution of the distributions generated from (N4) and (N4 + αCD28 + IL-2). This similarity is illustrated in Fig. 2.10 and Fig. 2.11.

If the distributions to be compared result from convolution, the requirements for standard testing procedures are violated. To overcome this problem, we tested the hypothesis of statistical independence between αCD28 and IL-2 contributions developing a new testing procedure, whose details we defer to Chapter 4. The p-values so produced corresponded to 0.399 and 0.377 for maxDD and mDD, respectively, and the hypothesis of independent additivity of αCD28 and IL-2 was not rejected.

2.3.6 Signal sensitivity regulates clonal family DD

The findings of clonal concordance in DD and independent signal integration indicates that a major decision is taken by the mother cell and passed along the offspring to ultimately shape the whole family. We question if the inherited features can be explained by signal perception. In particular, we reason that if the signal receptor levels are uncorrelated then, at the clonal level, stimuli contributions should be independent. Otherwise, we would expect that receptor correlation would lead to dependent signal integration, which is not supported by our findings. Under the conditions previously considered (Figs. 2.2, 2.3, 2.7 and 2.10), CD28 and interleukin-2 receptor chain alpha (IL-2Rα) levels were found relatively uniform, with a Spearman's correlation of 0.16 (Fig. 2.12). This agreed with our hypothesis of costimuli independence.

Subsequently, we investigated whether difference in the mother cell's receptor levels correlate with the clonal size at quiescence. To this end, we measured mDD regulations from naive OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells that were sorted into high and low CD28 level populations, noted $CD28^{hi}$ and $CD28^{lo}$ respectively (Fig. 2.13a). Little expansion was observed from sorting when no αCD28 was added to the culture (open circles, Fig. 2.13b-d). With the inclusion of such costimulatory signal, we recorded a 50% increase in the population size (closed circles, Fig. 2.13b,c) and an increase in mDD up to approximatively 0.6 relative to the $CD28^{lo}$ population (black arrow, Fig. 2.13c,d). We conclude that initial differences in CD28 receptor levels contributes to the variation of clonal DD.

A different behaviour was observed when naive T cells were sorted for IL-2Rα high (IL-2Rα^{hi}) and low (IL-2Rα^{lo}) levels prior to first division (Fig. 2.13e): the sorting

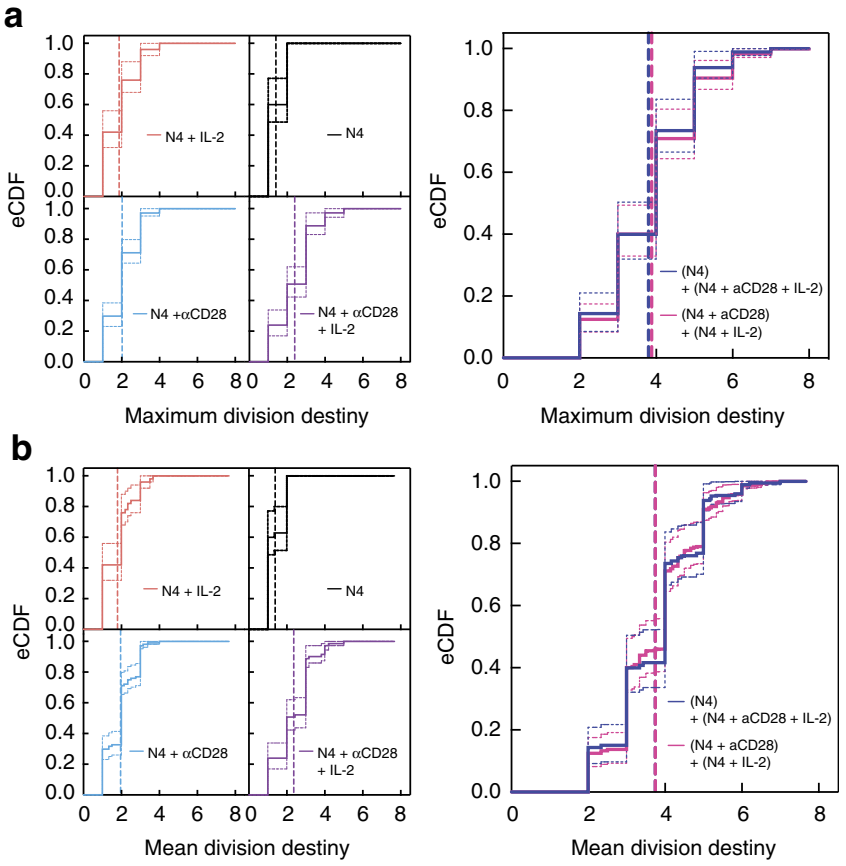

FIGURE 2.10: **Stimuli effects on DD sum independently at the level of the clonal family tree.** [Corresponding to Figure 5 from Marchingo, Prevedello et al., (2016)] From the data as in Figure 2.8, empirical cumulative distribution functions (eCDF) for measures of clonal expansion, (a) maximum DD (maxDD) and (b) mean DD (mDD), are plotted for each individual stimulation condition (left panel). To test clonal signal addition, the convoluted distribution of the statistics from (N4)+(N4+αCD28+IL-2) and (N4+αCD28) +(N4+IL-2) were compared (right panel and Section 3.4). Vertical dashed lines represent mean of the pooled clones. Dotted lines show 95% confidence intervals (see Section 2.3.4). A non-standard χ^2-test of independence (see Chapter 4) was not rejected for either maxDD (p-value = 0.399) or mDD (p-value = 0.377). Results from a second independent experiment are shown in Fig. 2.11.

for IL-2Rα induced a 0.73 increase in mDD irrespective of IL-2 addition to the culture (Fig. 2.13f-h).

It was previously shown that IL-2Rα is regulated by the TCR stimulation strength (Zehn et al., 2009; Wensveen et al., 2010; Gottschalk et al., 2012). Therefore, we reason that sorting for IL-2Rα is a proxy for a discrimination based on intrinsic TCR stimulation strength, which itself impacts on DD (Marchingo et al., 2014). Given that mDD was extended by IL-2 presence, but left unaffected by initial IL-2Rα expression, indicates that the inflammatory signal integration was not entirely explained on this multi-subunit receptor. Previous studies showed that IL-2Rα level exceeds the IL-2Rβ and γ chains level, whose compound is necessary for the IL-2 transmission (Smith

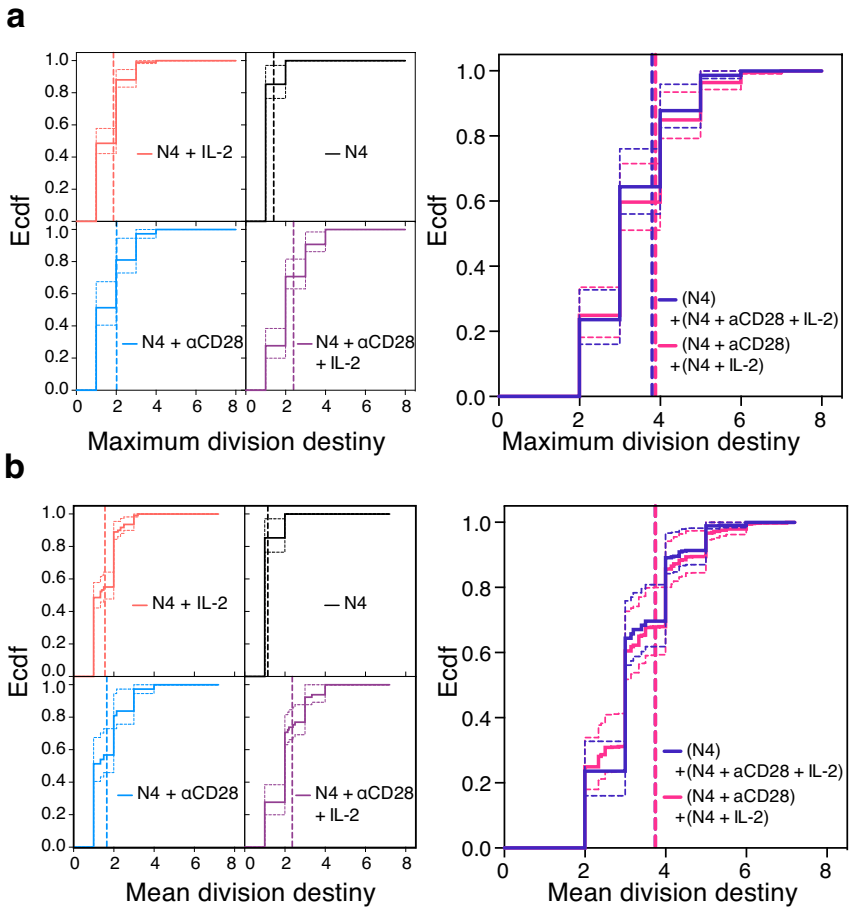

FIGURE 2.11: **Stimuli effects on DD sum independently at the level of the clonal family tree, experimental repeat.** [Corresponding to Supplementary Figure 7 from Marchingo, Prevedello et al., (2016)] Procedure described in 2.10 for data as in Fig. 2.8. A non-standard χ^2-square test for independence (see Chapter 4) was not rejected for either maxDD (p-value = 0.613) or mDD (p-value = 0.6).

and Cantrell, 1985; Feinerman et al., 2010). From these findings, we reason that IL-2 integration is independent of other co-receptor and the independence not affected by IL-2Rα expression. In particular, the dependence of signal integration with the receptor levels is conditioned by the extent of signal sensitivity ultimately achieved.

Fig. 2.14 recapitulates the effect that sorting for CD28 and IL-2Rα has on DD expansion. OT-I/*Bcl2l11*$^{-/-}$ CD8$^+$ T cells were selected for CD28$^{\text{hi}}$ expression (top 20%) and stimulated by N4 self-presentation. At 26 h the population was sorted for IL-2Rα^{hi} (top 35%), cultivated with the addition of hIL-2 and harvested at 72 h for the comparison with the unsorted control population. As expected, the DD variance decreased from 1.3 in the unsorted pool to 0.67 in the sorted one.

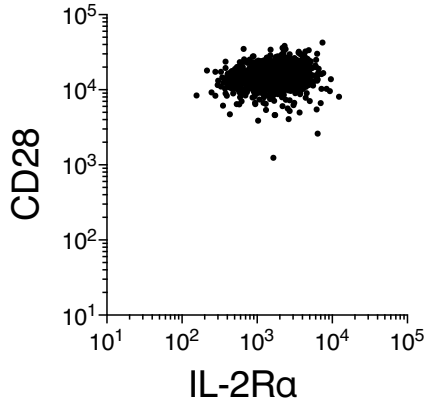

FIGURE 2.12: **CD28 and IL-2Rα levels prior to the first division are relatively uniform.** [Corresponding to Supplementary Figure 8 from Marchingo, Prevedello et al., (2016)] CD28 and IL-2Rα expression of activated CTV labelled OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells stimulated for 24 h with N4 peptide (0.01 μg ml^{-1}) in the presence of αCD28 (2 μg ml^{-1}), S4B6, (25 μg ml^{-1}) and hIL-2 (1 U ml^{-1}). Representative of duplicate culture wells from 3 independent experiments.

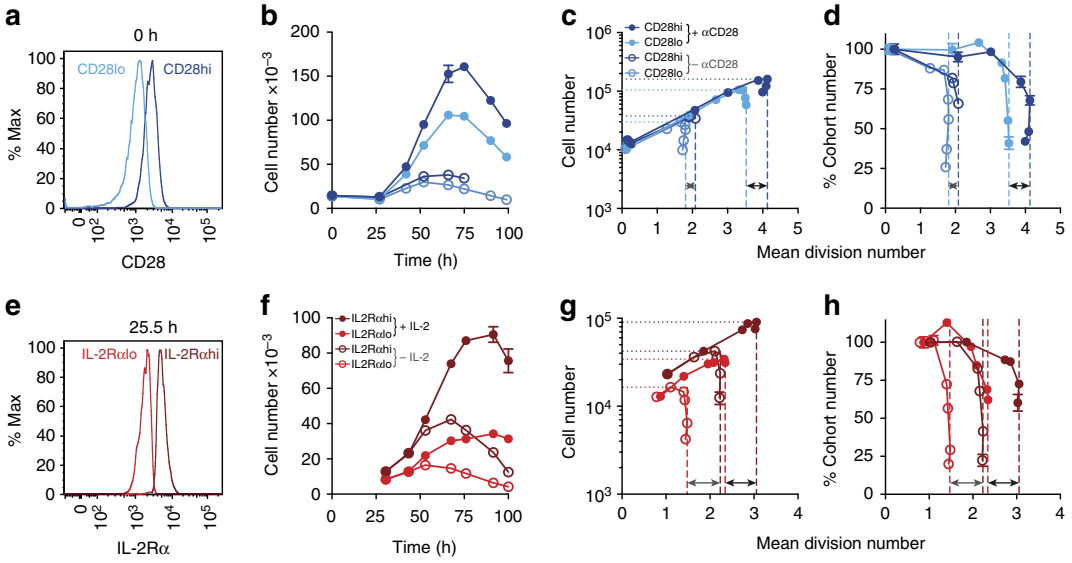

FIGURE 2.13: **Inter-clonal variation in DD is regulated by receptor sensitivity and clonal experience.** [Corresponding to Figure 6 from Marchingo, Prevedello et al., (2016)] (a-d) Naive CTV-labelled OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were sorted into (a) CD28 high and low expressing populations and stimulated with N4 peptide $\pm\alpha$CD28 agonist antibody (20 μg ml^{-1}). All cultures contained S4B6 (25 μg ml^{-1}). Cell number versus (b) time and (c) mean division number (MDN) were measured (see A.1.7). (d) An estimation of the percentage of the starting cells whose progeny are contributing to the response at that time point, calculated by removing the effect of cell expansion (percentage cohort number, see A.1.8) versus MDN. (e-h) Naive CTV-labelled OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were stimulated with N4 peptide and αCD28 (2 μg ml^{-1}) for 25 h then sorted for (e) IL-2Rα high or low expression. Cells were placed back into culture $\pm$hIL-2 (3.16 U ml^{-1}) and cell number versus (f) time and (g) MDN were measured and (h) percentage cohort number versus MDN calculated. Arrows indicate the difference in mDD between populations when no additional ligand (grey) or ligand at the specified concentration (black) was added to the culture. Representative of two (a-d) or three (e-h) independent experiments. Mean $\pm$ s.e.m. of triplicate culture wells.

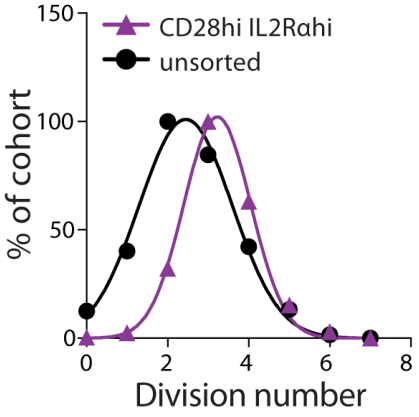

FIGURE 2.14: **Variation in DD is reduced when restricted to a constrained receptor range.** [Corresponding to Supplementary Figure 9 from Marchingo, Prevedello et al., (2016)] Naive CTV labelled OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were sorted for CD28$^{\text{hi}}$ expression (top 20%) and stimulated with N4 peptide + αCD28 (2 μg ml^{-1}). After 26 h cells were sorted for IL-2Rαhi expression (top 35%) and placed back in culture with hIL-2 (3.16 U ml^{-1}). After 75 h cell proliferation of sorted cells and unsorted control was compared. Cohort number vs. division number fitted with normal distributions. Data from one experiment. Mean ± s.e.m. of triplicate culture wells.

2.4 Discussion

Given the small number of cells in the immune system that are specific to any one antigen, their clones undergo extensive mitotic divisions in order to mount a robust response. At the peak of their expansion, *in vivo* naive CD8^{+} T-cells can divide up to 15-20 times (Butz and Bevan, 1998; Murali-Krishna et al., 1998; De Boer et al., 2003) and commit to different classes, such as memory or effector cells, that share the same specificity to the antigen that initially activated the founder cell. Previous studies, using single clone tracking methods, showed that highly heterogeneous family size and similar cell specialisation arose *in vivo* from identical precursors cells (Buchholz et al., 2013; Gerlach et al., 2013). In order to explain the heterogeneity of a T-cell population, other research teams have proposed mechanism of asymmetric cell division (ACD) that takes place at the early stages of the expansion phase (Chang et al., 2007; Reiner and Adams, 2014; Buchholz et al., 2016). Given the growing interest surrounding how extrinsic and intrinsic factors generate diversity from a pool of similar cells, we analyse T cells in a controlled *in vitro* environment to study their external and programmed regulation.

To this end, our collaborators implemented a novel multiplex clonal assay that allowed the tracking of division fate up to 6-7 generations from cells of the same progeny and identical TCR, thanks to the combination of division tracking dyes and flow cytometry technology. Together, we analysed the output from this method to study the clonal

contribution to population DD, as the generation in which cells stop dividing is highly variable (Turner et al., 2008; Hawkins et al., 2009; Marchingo et al., 2014). Under different stimuli, we observed that cells from the same clone turn quiescent in the same or adjacent generation, thus indicating inter-clonal variation as the main cause for population DD heterogeneity. Results for T cells are analogous to those for B cells, where time-lapse microscopy technique showed DD fate in related cells was highly correlated when stimulated by the toll-like receptor agonist CpG DNA (Hawkins et al., 2009).

To explain these findings we hypothesise that, when stimulated, resting T or B cells undergo several rounds of divisions until destiny is reached, which is programmed and inherited from the founder. While this possibility accounts for strong concordant families, the discordance that arises must be explained. Previous studies have analysed the effect of extrinsic signals on T-cell fate (Mescher et al., 2006; Chen and Flies, 2013; Buchholz et al., 2016). In particular, costimuli provided by the contact with antigen presenting cell (APC) promotes asymmetric division in early generations (Chang et al., 2007; King et al., 2012), and variation in APC or antigen exposure can shape the fate of a single familial branch. The multiplex assay, presented in this chapter, is ideal for the study of the effect of extrinsic signals in a controlled environment, as it avoids the technical difficulties of time-lapse microscopy that are crucial in some systems such as with APC or stromal cells stimulation.

With the multiplex method we investigated the effects of costimulatory and cytokine signals to T-cell DD. Our analysis necessitated the development of a new mathematical framework for the concatenation of family trees to study signal additivity (see Section 3.4). Using a novel statistical test (see Chapter 4) we probed the hypothesis that the contribution to DD of CD28 was not affected by IL-2 and vice versa. Together, these techniques support the view that in T cells signal integration occurs as a linear and independent sum of stochastic contributions from costimulatory and cytokine signals. These conclusions indicated that a DD variability is an inherited feature that may be traced back to differences in molecular determinants of founder cells.

Thanks to the controlled system in consideration, it was possible to investigate if costimulatory receptors CD28 and IL-2Rα were responsible of the early DD programming. In naive OT-I/$Bcl2l11^{-/-}$ $CD8^{+}$ T cells, CD28 is approximatively log-normally distributed and higher levels induce a stronger signal reception that leads to an increased DD expansion. To analyse the role of cytokine IL-2 in T-cell DD, we eliminated autocrine IL-2 in the culture to avoid local influence that may add uncontrolled variations to the system. As for CD28, we studied the effect on clonal DD of IL-2Rα initial levels, which is highly variable under N4 antigen stimulation. We found that this receptor

was not the limiting factor for the transmission of the IL-2 signal. Previous studies have shown that IL-2Rα expression is correlated with stimulation strength (Zehn et al., 2009; Wensveen et al., 2010; Gottschalk et al., 2012). If this stimulation were sufficiently weak, the reception of autocrine IL-2 could lead to local effects that may increase the clonal DD variation. More *in vitro* studies are required to interrogate such hypothesis, with subsequent *in vivo* experiments to determine its functional significance.

There are other receptors that affect the progression through division of T-cell clones, such as CD27, IL12, IL-4, IL-2 and IL-6 (Curtsinger and Mescher, 2010; Marchingo et al., 2014; Starbeck-Miller et al., 2014; Voisinne et al., 2015). Assuming that these receptors are variable from cell to cell and expressed independently to one another, the resulting population would present heterogeneous potentials with different behaviours under the same condition. We speculate that the diversity of receptor combinations in the initial pool would ultimately display a consistent behaviour, given that the stochastic drivers of receptor heterogeneity are reproducible. Consequence of this would be a population with heterogeneous DD fate and in various state, such as effector or memory, which was found to be related with division progression (Gett and Hodgkin, 1998; Schlub et al., 2009). Further investigation combining index sorting and multiplex assay methods together would cast light on these effects due to different receptor levels.

Given the remarkable concordance from the controlled *in vitro* experiments we analysed, a question arises concerning how much heterogeneity is determined by inter-clonal variation of concordant clones, rather than extrinsic factors, during an *in vivo* infection of $CD8^+$ T cell. Current *in vivo* experiments cannot distinguish between these two sources of variation, but *in vivo* data can be interrogated to test whether a programmed and concordant DD is admissible. Thus, the DD distribution from Marchingo et al. (2014) *in vivo* population experiments was estimated to infer the relative familial size distribution under the hypothesis of concordant clonal DD. As a result, a small number of large clones were responsible for most of the total response size, an outcome that is quantitatively similar to *in vivo* studies from Buchholz et al. (2013) and Gerlach et al. (2013) (Fig. 2.15a).

To achieve this comparison, data for the burst size of individual OT-I $CD8^+$ T-cell clones at response peak during a Listeria monocytogenes-OVA infection was obtained from previous clonal studies from Gerlach and colleagues (data from Fig. 2.1c, clones distinguished using genetic barcoding technology from Gerlach et al. (2013)) and Buchholz and colleagues (data from Fig. 2.1e, clones distinguished using a congenic marker matrix from Buchholz et al. (2013)). This was compared to a previously published population time course of OT-I/FucciRG $CD8^+$ T cells responding to an HKx31-OVA influenza infection (Fig. 2.1d,e from Marchingo et al. (2014)). The DD distribution for

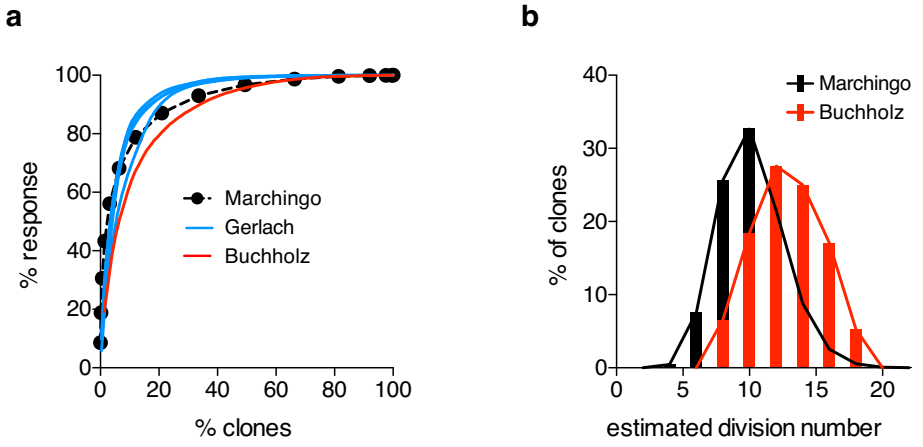

FIGURE 2.15: **Concordant T-cell division destiny is consistent with previous *in vivo* population and clonal response data.** [Corresponding to Supplementary Figure 10 from Marchingo, Prevedello et al., (2016)] (a) The clonal contribution to T cell response magnitude was predicted for population response data (black circles) by assuming the clonal progeny exhibited concordant DD and followed the DD distribution estimated previously by Cyton fitting (Marchingo et al., 2014). This was compared to the contribution to response magnitude of individual OT-I $CD8^+$ T-cell clones from Gerlach and colleagues (blue lines, each line shows data from an individual mouse) and Buchholz and colleagues (red line). (b) By assuming that DD was clonally concordant the DD distribution generated by individual clonal families in Buchholz et al. (2013) was estimated from the response magnitude (see Section A.1.9 in Appendix A) and compared to the DD distribution estimated by Cyton fitting to the population data in Marchingo et al. (2014). Due to the read-count threshold that was applied to remove background noise in Gerlach et al. (2013), data on small families is also lost. As a result it is not possible to estimate the full response DD distribution for this data.

the OT-I/FucciRG $CD8^+$ T cells in this experiment was estimated by mathematical fitting using the Cyton model. The inferred DD distribution of Marchingo et al. (2014) was also compared with Buchholz et al. (2013), which were both distributed over 10-15 generations (Fig. 2.15b).

Although these comparisons are consistent with the hypothesis of clonally regulated DD, more investigations are required to confirm *in vivo* intrinsic concordance of DD under a wider range of conditions, the interplay with extrinsic signals and the consequences for cellular differentiation to effector or memory state. The findings discussed so far are of fundamental importance to understand clonal regulation of T cells and will impact the design of predictive models for T-cell response. Given the growing importance of anti-cancer therapies that expand and reinvigorate highly clonal *in vitro* and *in vivo* T-cell responses (Restifo et al., 2012; June et al., 2015), this better understanding of the fundamental nature and source of the variation in burst-size will facilitate rational optimization of T-cell manipulation.

Chapter 3

Mathematical methods for multiplex clonal assay data analysis

3.1 Abstract

In this chapter we present the mathematical framework we developed for the analysis of the multiplex clonal assay output. We motivate the mathematical assumptions by referring to the biological and experimental constraints reported in Chapter 2. First, we introduce the data structures for clones and multiplexed data, i.e. family trees and family vectors respectively. Then, we formalise the statistics for clonal progression and study how these are affected by partial recovery of the cells within a clone. In order to probe linear signal integration (as in Section 2.3.5 of Chapter 2), we define a novel class of operations between family trees that describes the action of two stimuli linearly contributing to the clonal expansion when simultaneously provided. When two family trees are transformed by any operation from this class, the resulting tree presents a family vector that is the discrete convolution of the family vectors relative to the initial family trees. As a consequence, we set the discrete convolution as the all-encompassing representative for the rooted trees operations considered. Since the statistics for clonal progression are linear with respect to discrete convolution, we set the basis to assess the hypothesis that signal integration is independent using a statistical test for equality in distribution of two distinct sums of independent and discrete random variables that is derived in Chapter 4.

3.2 Structures for clonal data

3.2.1 Family trees

In Chapter 1 we discussed how various experimental protocols provide different informations concerning a particular system under analysis. We mentioned methods that provide the clonal size or a proxy for it, exploiting adoptive transfer of labelled single cells (Buchholz et al., 2013) or cellular barcoding technique (Gerlach et al., 2013), or population cell number, through cohort experiments (King et al., 2012; Lemaître et al., 2013; Marchingo et al., 2014; Kinjyo et al., 2015; Heinzel et al., 2017). Time-lapse microscopy (Hawkins et al., 2009; Zaretsky et al., 2012; Dowling et al., 2014) is a technique that, *in vitro*, allows the recording of the entire clonal history, that consists in division and death times plus familial relation of each cell (see Fig. 3.1 top), although it typically requires a substantial processing workload, to the point that software solutions for automatised cell tracking have been developed (Rieger et al., 2009; Kan et al., 2011; Pham et al., 2013; Shimoni et al., 2013; Chakravorty et al., 2014; Mankowski et al., 2015). Because of the amount of details that can be recovered, time-lapse microscopy provides remarkable information at the single-cell level and serves as benchmark for other methods' output.

Although, time-lapse microscopy presents some limitations. For example, tracking is lost when cells cannot be distinguished from one frame to the next, which occurs in case of high motility or when a cell, adjacent to another one, divides. Also, the formation of three-dimensional blocks, that would impede the detection of some cells, is an hindrance that concerns systems where cells adhere to each other or that must be cultured with other cell types. We mention Duffy et al. (2012) as an example where an experimental protocol was designed *ad hoc* to film the proliferation of B cells affected by cellular adhesion.

The multiplex clonal assay, as described in Chapter 2, overcomes these physical limitations as cells are separated when analysed through flow cytometry, which then enables the recording of clonal membership, generation and state (quiescent or dividing) from all the observed cells at a fixed time (see Fig. 3.1 middle).

Later, in Chapter 5, we shall see that multiple cell surface makers can also be measured simultaneously, allowing the statistical testing of independence of these markers, or the phenotypes derived from them, with clonal or environmental structures and the visualisation of division pattern in the first generation of cells. Still, one of the main drawbacks is the impossibility of recovering the information concerning familial relationships.

As a clonal family is a collection of the single initial progenitor plus its offspring cells and the genealogical relationships they establish, it is naturally represented by a rooted tree (see Fig. 3.1 bottom), which is an element from graph theory whose formal definition we report here adapted from Rosen (2011). We forewarn that this notation does not account for any temporal information, concerning, for example, the occurrence and duration of division events. In fact, we study a clonal family only as the generational configuration of its cells at a fixed time, either when quiescent (see Chapter 2) or during the expansion phase (see Chapter 5).

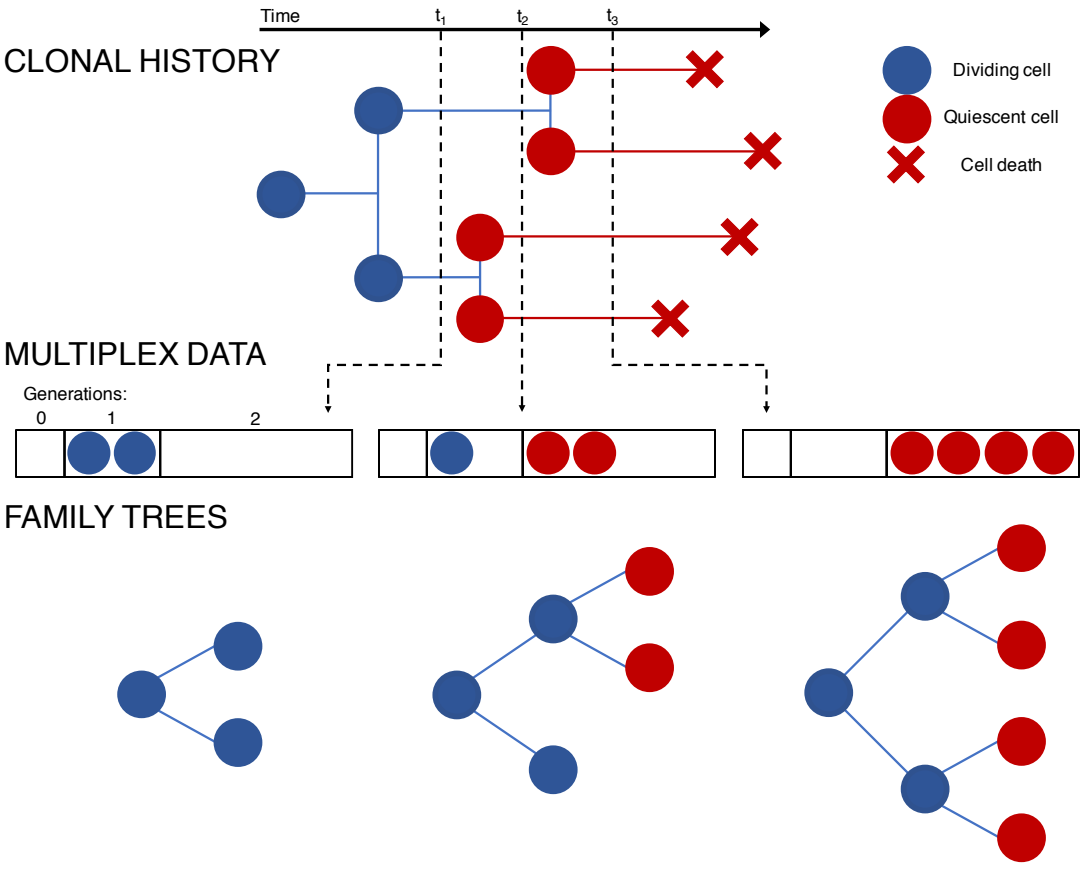

FIGURE 3.1: **Genealogical structure of a clonal history at a given time.** (Top panel) Starting from a progenitor cell, a clonal family develops with time (from left to right) generating cells that have different lifespans (horizontal lines length) before dividing (blue circles), turning quiescence (red circles) or ultimately dying (red crosses). (Middle panel) Data from the multiplex clonal assay described in Chapter 2 are a sample of the evolving clone at a given moment. Given three time points $0 < t_1 < t_2 < t_3$, the data recovered from the clone in top panel are: a dividing clone with two cells in generation 1 (t_1, left); a mixed-type clone with one dividing cell in generation 1 and two quiescent in generation 2 (t_2, centre); a quiescent clone with four cells in generation 2 (t_3, right). (Bottom panel) Below each sample from the middle panel, the genealogical structure achieved by the recovery time point, relative to the sample, is summarised into a rooted tree, whose root corresponds to the initial progenitor and is oriented from left to right mirroring the temporal progression.

Definition 3.1 (Graph)**.** A graph $G = (V, E)$ consists of V, a non empty set of vertices

(or nodes) and E, a set of edges. Each edge has either one or two vertices associated with it, called its endpoints. An edge is said to connect its endpoints and the vertices associated to such endpoints are said adjacent (or neighbours).

Definition 3.2 (Subgraph)**.** A subgraph of a graph $G = (V, E)$ is a graph $H = (W, F)$, where $W \subseteq V$ and $F \subseteq E$.

Definition 3.3 (Simple graph)**.** A graph in which each edge connects two different vertices and where no two edges connect the same pair of vertices is called a simple graph.

Definition 3.4 (Directed/Undirected graph)**.** A graph $G = (V, E)$ is said to be directed if each edge is associated with an ordered pair of vertices. A graph $G = (V, E)$ is said to be undirected if each edge is associated with an unordered pair of vertices.

Definition 3.5 (Path and circuit)**.** Let n be a nonnegative integer and G an undirected graph. A path of length n from u to v in G is a sequence of n edges $e_1, \ldots, e_n$ of G for which there exists a sequence $x_0 = u, x_1, \ldots, x_{n-1}, x_n = v$ of vertices such that e_i has, for $i = 1, \ldots, n$, the endpoints x_{i-1} and x_i. When the graph is simple, we denote this path by its vertex sequence $x_0, x_1, \ldots, x_n$ (because listing these vertices uniquely determines the path). The path is a circuit if it begins and ends at the same vertex, that is, if $u = v$, and has length greater than zero. The path or circuit is said to pass through the vertices $x_1, x_2, \ldots, x_{n-1}$ or traverse the edges $e_1, e_2, \ldots, e_n$. A path or circuit is simple if it does not contain the same edge more than once.

Definition 3.6 (Connectivity)**.** An undirected graph is called connected if there is a path between every pair of distinct vertices of the graph. An undirected graph that is not connected is called disconnected. We say that we disconnect a graph when we remove vertices or edges, or both, to produce a disconnected subgraph.

Definition 3.7 (Tree)**.** A tree is a connected undirected graph with no simple circuits. In particular, for any two vertices v, w in a tree there exists one and only one path having v and w as endpoints.

Definition 3.8 (Rooted tree)**.** A rooted tree is a tree in which one vertex has been designated as the root. A vertex different than the root that is connected to one and only one other node is called leaf. A vertex that is neither root or leaf is referred to as internal. If the rooted tree is degenerate, presenting only one the root node, then we set the convention that such node is also a leaf. Given two vertices v, w of a rooted tree, w is said descendent of v if the only path from the root to w passes through v. In particular, w is called a child (or daughter) of v if w is also adjacent to v. Given a rooted tree T and a vertex v of T, the generation of v in T is defined as length of the only path from the root of T to v.

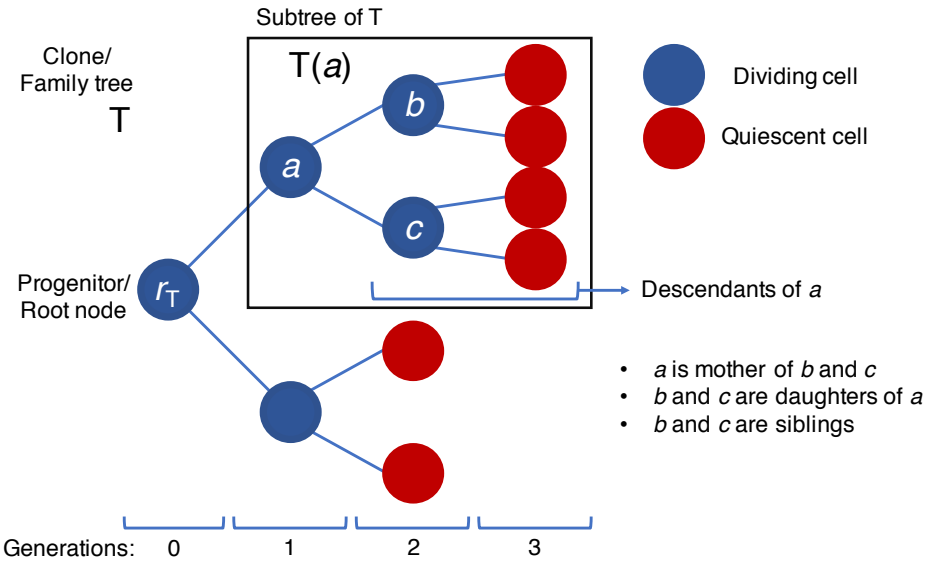

FIGURE 3.2: **Family tree for the representation of a clone.** A full binary rooted tree T displays the familial configuration of a clonal family generated from a single progenitor r_T, the root node in generation 0. Every cell is represented by a vertex whose colour indicates if the cell ultimately divides (blue) or becomes quiescent (red). Every edge connects two nodes if their cells establish a mother-daughter relationship. In particular, the mother cell is always in previous generation than its daughter cell. Moreover, no cells in the same generation (siblings) can be linked by an edge. The descendants of a cell is the collection its daughters and all their subsequent offspring. Thus, the subtree rooted in a, i.e. T(a) within the box, is the subgraph generated by a and all its descendants together connected.

Definition 3.9 (Subtree of a rooted tree)**.** Given a tree T and a node v of T, the subtree of T rooted in v is the subgraph of T obtained removing all nodes other than v and its descendants and all the edges with at least one endpoint connecting a removed node. We denote $T(v)$ the subtree so defined.

As illustrated in Fig. 3.2, the representation of a clonal family through a rooted tree, is achieved by identifying each cell with a different node and setting the ancestor cell's node as the root. Subsequently, the graph edges are determined so as to connect each pair of nodes that are related to a mother-daughter pair of cells. We call such graph a family tree.

In a family tree, the generation of a node is the length of the only path from the root to such a node. This value indicates the number of mitotic division undergone from the progenitor to obtain the cell relative to the node in question. The root-to-node paths induce a natural orientation in a rooted trees placing mother cells closer to the root node, previous in generation than their descendants. In particular, the left-to-right orientation reflects the temporal direction (see Fig. 3.1). Moreover, as each cell has either two daughters or none, a family tree belongs to a special class of trees called full binary tree.

Definition 3.10 (Full binary tree)**.** A binary tree is a rooted tree such that its root vertex is connected to a maximum of two nodes and every internal node has no more than 2 children. In particular, for a binary tree, the maximum number of nodes in generation $k \geq 0$ attainable is 2^k. A binary tree is full if its root vertex is connected to either two or no nodes and every internal vertex has exactly 2 children.

Definition 3.11 (Family tree)**.** For any fixed $k \in \mathbb{N}$, $\mathcal{T}_k$ is the set of full binary trees with nodes non-exceeding generation k. Any element of $\mathcal{T}_k$ is referred to as family tree.

The binary nature of family trees has important ramifications, from cohort normalisation to cell recovery quantification, which will be evident in the following sections.

3.2.2 Family vectors

In order to describe data from the experimental protocol presented in Chapter 2, we need to find an alternative representation other than rooted trees, since familial relationships cannot be recorded by the multiplexed clonal assay. To this end, we associate each clone to the vector of the counts of cells at a given time, which we call a family vector. A family vector is an element $v \in \mathbb{N}_0^k$ such that v_i is the number of cells found in generation i, where by $\mathbb{N}_0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$ we indicate the set of nonnegative integers. As for the binary property of family trees (Section 3.2.1), family vectors can be characterised by the feature that every cell gives birth to either two or no cells.

Definition 3.12 (Family vectors)**.** Let $k \geq 0$. For any element $v \in \mathbb{N}_0^{k+1}$, the cohort correction of v, or say that v is cohort corrected, is

$$\mathrm{cc}(v) = (v_0, 2^{-1}v_1, \ldots, 2^{-k}v_k) \in \mathbb{R}_0^{k+1} \tag{3.1}$$

and the cohort number of v is

$$\mathrm{cn}(v) = \sum_{i=0}^{k} 2^{-i} v_i = \sum_{i=0}^{k} \mathrm{cc}(v)_i. \tag{3.2}$$

Then

$$\mathcal{V}_k = \left\{ v \in \mathbb{N}_0^{k+1} \colon \mathrm{cn}(v) = 1 \right\} \tag{3.3}$$

is the set of family vectors with maximum generation k.

Of note, for a fixed $k \in \mathbb{N}_0$, any full binary tree $T \in \mathcal{T}_k$ (i.e. a family tree) can be mapped to the vector $v \in \mathcal{V}_k$ such that v_i is the number of leaves in generation i of T, see Fig. 3.3.

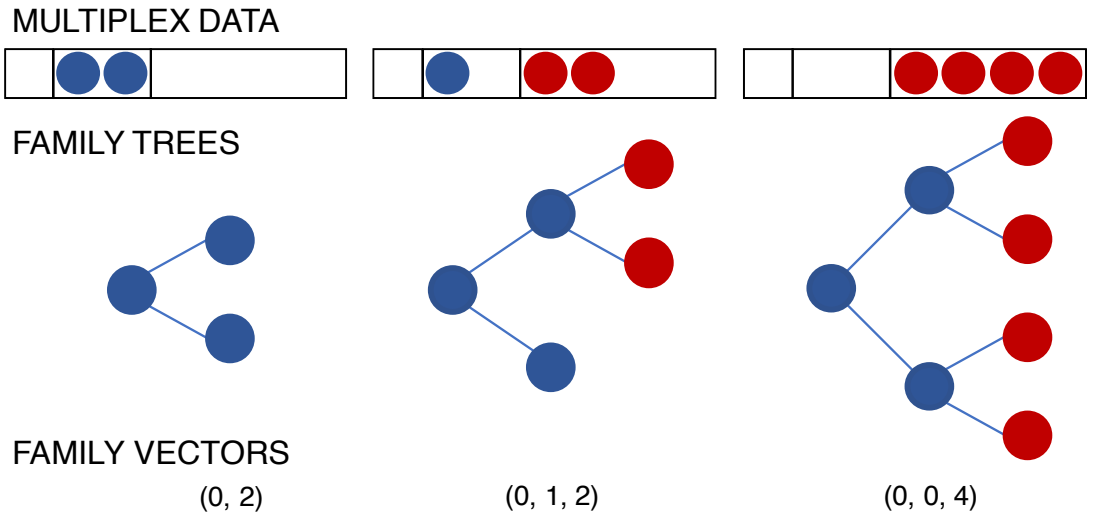

FIGURE 3.3: **Family tree and family vector relation with multiplex clonal assay data.** Given the multiplex data and the relative family trees as in Fig. 3.1, the family vectors associated to them are $(0,2)$ (left), $(0,1,2)$ (centre) and $(0,0,4)$ (right), whose entries are the count of leaves per generation.

Proposition 3.13 (Projection of family trees into family vectors)**.** *Let $k \in \mathbb{N}_0$ and $T \in \mathcal{T}_k$ a family tree. Given the vector $v \in \mathbb{N}_0^{k+1}$ such that v_i is the number of leaves in generation i of T, for $i = 1, \dots, k$, then*

$$\mathrm{cn}(v) = 1 \tag{3.4}$$

and $v \in \mathcal{V}_k$.

Proof. We recall that, for fixed $i \in \{0, \dots, k\}$, 2^i is a sharp upper bound for the number of leaves in generation i, by Definition 3.10 of a binary tree. Hence, introducing the variables w_i, for the number of cells in generation i that undergo division, and u_i, for the number of cells in generation i that could have descended from cells that stopped dividing in earlier generations, we deduce that

$$2^i - u_i = v_i + w_i. \tag{3.5}$$

For $i > 0$ and $j \leq i-1$, let $u_{i,j}$ be defined as the number of cells in generation i that could have been generated by cells that, instead, stopped dividing in generations $j, \dots, i-1$, then $u_i = u_{i,0}$. In particular, since T is a full binary tree, $u_{i,i-1} = 2v_{i-1}$ as each cell that stopped dividing in generation $i-1$ would have contributed with 2 cells in generation i. Analogously, $u_{1,0} = 2v_0$. Finally, from the recursive relation

$$u_{i,i-2} = u_{i,i-1} + u_{i-1,i-2}, \tag{3.6}$$

we obtain

$$u_i = \sum_{j=0}^{i-1} 2^{i-j} v_j. \tag{3.7}$$

By substitution of u_i in (3.5) we find

$$w_i = 2^i - u_i - v_i = 2^i - \sum_{j=0}^{i} 2^{i-j} v_j. \tag{3.8}$$

As $T \in \mathcal{T}_k$, then no cell divides in generation k and $w_k = 0$, leading to

$$2^k = \sum_{j=0}^{k} 2^{k-j} v_j. \tag{3.9}$$

Dividing by 2^k, we deduce that $\mathrm{cn}(v) = 1$. □

Proposition 3.13 justifies the constraint $\mathrm{cn}(v) = 1$ required in the definition of a family vector $v \in \mathcal{V}_k$, for any $k \in \mathbb{N}_0$. Moreover, this result evidences that the map from $\mathcal{T}_k$ to $\mathcal{V}_k$ of the counting of the leaves per generation is a surjection, thus enabling the extension to family vectors of properties from the family trees.

3.2.3 Statistics of progression

We introduce several statistics that summarize the extent of clonal progression as measured from family vectors. To be consistent with the analysis in Chapter 2, we use the language of Division Destiny (DD), but, mathematically, it is not necessary that the family vectors correspond to final expansion of clones (see Fig. 3.3).

Definition 3.14 (Expansion statistics)**.** Given $k \geq 0$ and $v \in \mathbb{N}_0^{k+1}$ such that $v \neq (0, \ldots, 0) = v_0$, we define: the mean division destiny,

$$\mathrm{mDD}(v) = \frac{\sum_{i=0}^{k} i 2^{-i} v_i}{\sum_{i=0}^{k} 2^{-i} v_i}; \tag{3.10}$$

the maximum division destiny,

$$\mathrm{maxDD}(v) = \max\big\{i \in \{0, \ldots, k\} \colon v_i \neq 0\big\}; \tag{3.11}$$

the minimum division destiny,

$$\mathrm{minDD}(v) = \min\big\{i \in \{0, \ldots, k\} \colon v_i \neq 0\big\}; \tag{3.12}$$

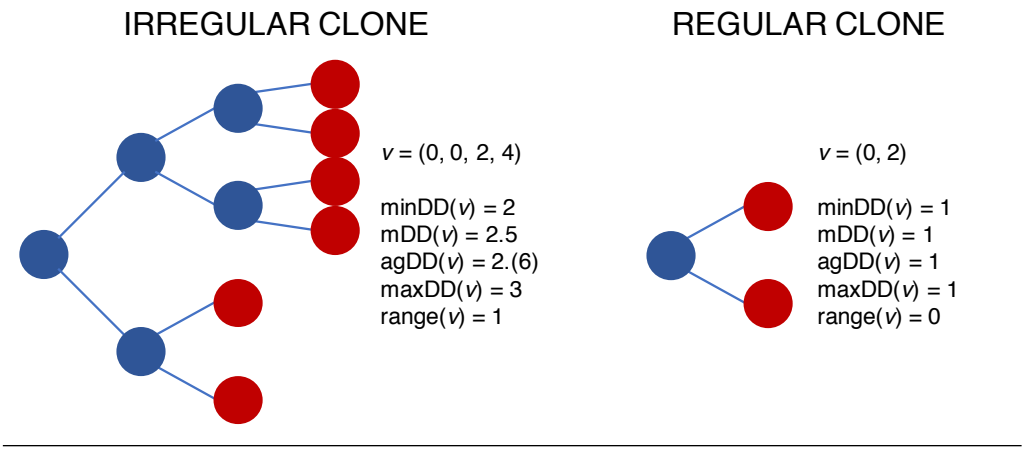

FIGURE 3.4: **DD statistics.** Example of DD statistics on irregular (left, range > 0) and regular clones (range, range $= 0$), with associated family vector v equal to $(0, 0, 2, 4)$ and $(0, 2)$, respectively.

and the average-generation division destiny,

$$\mathrm{agDD}(v) = \frac{\sum_{i=0}^{k} i v_i}{\sum_{i=0}^{k} v_i}. \tag{3.13}$$

We adopt the convention that $\mathrm{mDD}(v_0) = \mathrm{maxDD}(v_0) = \mathrm{minDD}(v_0) = \mathrm{agDD}(v_0) = 0$.

See Fig. 3.4 as a visual example of the statistic introduced above. Note that, for any $k \geq 0$, these functions are well-defined for elements other than $\mathcal{V}_k$. This will be useful in Section 3.3, when we will introduce the set family vectors for clones that are partially recovered.

While maxDD and minDD are intuitive mathematical descriptions for the expansion of a clone, in terms of maximal and minimal generation reached, mDD and agDD are two averages of the generations that were considered in recent publications. In particular, agDD (Turner et al., 2008; Weber et al., 2016) weights each generation on the proportion of cells that falls in it, thus considering all cells as equivalent. Instead, mDD (Hommel and Hodgkin, 2007; Marchingo et al., 2014) bases its weighting on the proportion of cohort vector, so to evaluate a cell in generation i equivalent to two cells in generations $i + 1$, for $i \geq 0$, that is accounting for a cell's generative potential.

Simple inequalities relate the four statistics of progression defined above.

Proposition 3.15 (Order of expansion statistics)**.** *Let $k \geq 0$, $v \in \mathbb{N}_0^{k+1}$. Then*

$$\mathrm{minDD}(v) \leq \mathrm{mDD}(v) \leq \mathrm{agDD}(v) \leq \mathrm{maxDD}(v). \tag{3.14}$$

Proof. The relation is true for the degenerate case $v = (0, \ldots, 0) \in \mathbb{N}_0^{k+1}$, for any $k \geq 0$. For $v \neq (0, \ldots, 0)$, the first and third inequalities result from

$$\text{minDD}(v) = \frac{\sum_{i=0}^{k} \text{minDD}(v) 2^{-i} v_i}{\sum_{i=0}^{k} 2^{-i} v_i} \leq \text{mDD}(v) \tag{3.15}$$

$$\text{agDD}(v) \leq \frac{\sum_{i=0}^{k} \text{maxDD}(v) v_i}{\sum_{i=0}^{k} v_i} = \text{maxDD}(v) \tag{3.16}$$

We prove the second inequality by induction on k. For $k = 0$, it is trivial, since $\text{mDD}(v) = \text{agDD}(v) = 0$. So, assuming the relation holds true for k, we obtain that

$$\sum_{i=0}^{k} i 2^{-i} v_i \sum_{j=0}^{k} v_j \leq \sum_{j=0}^{k} j v_j \sum_{i=0}^{k} 2^{-i} v_i. \tag{3.17}$$

We can then compare

$$\begin{aligned} \text{mDD}(v) \sum_{j=0}^{k+1} v_j \sum_{i=0}^{k+1} 2^{-i} v_i = &\sum_{i=0}^{k} i 2^{-i} v_i \sum_{j=0}^{k} v_j \\ &+ (k+1) 2^{-(k+1)} v_{k+1} \sum_{j=0}^{k} v_j + v_{k+1} \sum_{i=0}^{k} i 2^{-i} v_i + (k+1) 2^{-(k+1)} v_{k+1}^2 \end{aligned} \tag{3.18}$$

with

$$\begin{aligned} \text{agDD}(v) \sum_{j=0}^{k+1} v_j \sum_{i=0}^{k+1} 2^{-i} v_i = &\sum_{j=0}^{k} j v_j \sum_{i=0}^{k} 2^{-i} v_i \\ &+ (k+1) v_{k+1} \sum_{i=0}^{k} 2^{-i} v_i + 2^{-(k+1)} v_{k+1} \sum_{j=0}^{k} v_j + (k+1) 2^{-(k+1)} v_{k+1}^2. \end{aligned} \tag{3.19}$$

Thus, subtracting $(k+1)2^{-(k+1)} v_{k+1}^2$ on both (3.18) and (3.19) and recalling (3.17), it suffices to show that

$$\begin{aligned} v_{k+1} \left((k+1) 2^{-(k+1)} \sum_{j=0}^{k} v_j + \sum_{i=0}^{k} i 2^{-i} v_i \right) \\ \leq v_{k+1} \left((k+1) \sum_{i=0}^{k} 2^{-i} v_i + 2^{-(k+1)} \sum_{i=0}^{k} j v_j \right), \end{aligned} \tag{3.20}$$

or, equivalently,

$$\sum_{i=0}^{k} v_i \left((k+1) 2^{-(k+1)} + i 2^{-i} \right) \leq \sum_{i=0}^{k} v_i \left((k+1) 2^{-i} + i 2^{-(k+1)} \right), \tag{3.21}$$

which holds true since, for every $i = 0, \dots, k$, $v_i \geq 0$ and

$$(k+1)2^{-(k+1)} + i2^{-i} \leq (k+1)2^{-i} + i2^{-(k+1)}, \tag{3.22}$$

since

$$(k+1-i)2^{-(k+1)} \leq (k+1-i)2^{-i}. \tag{3.23}$$

□

We now introduce one last description for the configuration of a clone in DD.

Definition 3.16 (Clonal range)**.** Let $k \geq 0$, $v \in \mathbb{N}_0^{k+1}$. The range of v is

$$\text{range}(v) = \text{maxDD}(v) - \text{minDD}(v) \geq 0. \tag{3.24}$$

If $\text{range}(v) = 0$ we say that v, and the clone associated to it, is regular, and irregular otherwise.

The statistic of range enables the quantification of the generations span in which all the cells of a clone are found. Fig. 3.4 offers a visual example of range computation. By Proposition 3.15, all DD statistics must coincide on every family vector v that is regular, as $\text{range}(v) = 0$ implies $\text{maxDD}(v) = \text{minDD}(v)$.

While these are descriptions of clonal properties in general, range was introduced in Chapter 2 to investigate the division patterns underwent by the expanding clone (see Fig. 2.1b). If cells were highly correlated and stop dividing in the same generation, the clone produced would be regular with null range. On the other hand, if cells underwent division independently of each other, the resulting family would be irregular, with cells that revert to quiescence over several generations. As multiplexed data consist of sampled clones (see Fig. 2.2 and Section A.1.6), the recorded range may be reduced (e.g. in Fig. 3.6), thus biasing the division pattern in favour of the correlated cells scenario. To quantify the impact of partial recovery, in the next section we define a simple parametric model that describes the pattern of division through the generations, to calculate the range distribution while taking cell loss into account.

3.3 The effect of sampling on the clonal range

3.3.1 Sampled family vectors

In the experimental practice, researchers often must deal with incomplete data recovery. To design the multiplex clonal assay presented in Chapter 2, our collaborators had to

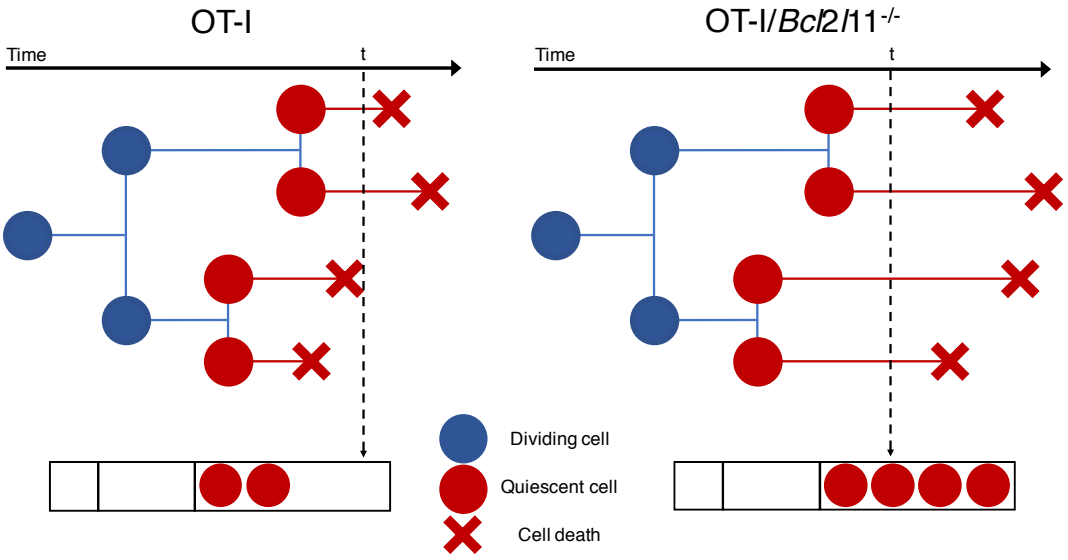

FIGURE 3.5: **Enhanced cell survival increases sample recovery.** If a clonal family is sampled at a given time $t > 0$ of its development from the OT-I mice, some quiescent cells may already be dead (left). As knocking-out the *Bcl2l11* gene from the OT-I system increases the survival time of the cells, the time window in which all quiescent cells can be recovered becomes larger, greatly reducing the data loss by cell death.

deal with data loss due to partial recovery of samples or cell death, which mainly occurs soon after the expansion phase. In order to avoid missing cells by apoptosis, our collaborators employed the OT-I/$Bcl2l11^{-/-}$ mice (see Section A.1.1 from Appendix A), which is transgenically modified for enhanced cell survival (Fig. 3.5).

As a consequence, we can ascribe incomplete recovery solely to experimental sampling, and account for its effect in a model if recovery ratio is available. This parameter is then estimated adding a certain number of calibrating beads to the culture wells, and calculating the proportion of these extracted (see Section A.1.5). If the beads are not included in the experiment, the estimate can be substituted by the percentage of volume collected. To study sampling effects, we need to extend the set of family vectors so to include partial recovery in the data structure.

Definition 3.17 (Sampled family vectors)**.** Given $k \geq 0$,

$$\mathcal{S}_k = \left\{ v \in \mathbb{N}_0^{k+1} \colon \operatorname{cn}(v) \leq 1 \right\}, \tag{3.25}$$

is the set of sampled family vectors whose generation do not exceed k.

Thus the data from a multiplex clonal assay method take value in $\mathcal{S}_k$, defined in (3.25), for an appropriate $k \geq 0$. For example, $k = 2$ in Fig. 3.6.

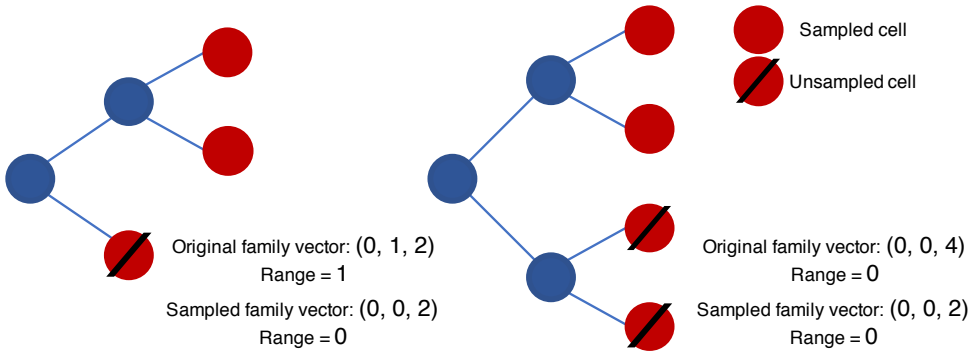

FIGURE 3.6: **Sampled family vectors.** Example of sampling effect on irregular (left, range $>$ 0) and regular clones (right, range $= 0$). If one cell is lost (barred red dot) in generation 1 from the family vector $(0, 1, 2)$ (left) the resulting vector becomes $(0, 0, 2)$, the same as when two cells in generation 2 are lost from the family vector $(0, 0, 4)$ (right). In particular, the range in the former clone is reduced to 0. This example shows that it is not always possible to trace back the original family vector from its sampled version, as the same sampled family vector may result from the partial recovery of different clones.

3.3.2 Beta-binomial model

From the multiplexed experiments on CD8$^+$ T cells whose data are summarised in Fig. 2.7 and 2.8 in Chapter 2, clones stimulated in a controlled system were found regular in most cases, whereas only a minority were mildly irregular with range equal to 1. This evidence suggested that homogeneity in DD may be due to some features inherited from the progenitor cell to its descendants, envisaging a correlated behaviour, between related cells, about whether to divide or to become quiescent (see Fig. 2.1b). To investigate whether the observed homogeneity was a consequence of sampling that reduces familial range, we propose a minimalistic model for the clonal expansion, realised by correlated divisions through the generations, and subsequent sampling.

Suppose the experimental data consist of $N > 0$ sampled family vectors, $z_1, \ldots, z_N \in \mathcal{S}_k$, with $k \geq 0$. Each of these vectors must have been sampled from a fully expanded clone whose associated family vector is in $\mathcal{V}_k$. Let $w \in \mathcal{V}_k$ be one of such vectors. To describe the development of w, we introduce the parameters $p_i \in [0, 1]$, with $i = 0, \ldots, k-1$ of the probability for a cell in generation i to divide to generation $i+1$, thus $1 - p_i$ is the probability for that cell to become quiescent.

Let u_i denote the number of cells that reached generation i, the decision whether to divide or not, may be concerted within the clone. We quantify the extent of such coordination with the parameter $\rho \in [0, 1]$ and model the number of cells that divide to generation i with a beta-binomial random variable $\beta(u_i, a_i, b_i)$, where $a_i = p_i(1-\rho)/\rho$

and $b_i = (1 - p_i)(1 - \rho)/\rho$. In fact, a_i and b_i are defined as solution to the equations

$$p_i = \frac{a_i}{a_i + b_i} \tag{3.26}$$

$$\rho = \frac{1}{1 + a_i + b_i}, \tag{3.27}$$

so that ρ is defined as the intraclass correlation of the cells, whose marginal distribution is binomial with parameter p_i. With this description, when $\rho = 0$ each cell behaves independently from the others, whereas for $\rho = 1$ all cells in one generation either divide or become quiescent. In any case, we assume that these decisions are independent across generations.

Since a clone starts with one cell in generation zero, and all dividing cells give birth to two others in the subsequent generation, then $u_0, \ldots, u_k$ are defined from w through the recursive relation

$$u_0 = 1 \tag{3.28}$$

$$u_{i+1} = 2(u_i - w_i) \quad \text{for} \quad i = 0, \ldots, k-1. \tag{3.29}$$

With some manipulation, for every $i = 1, \ldots, k$ we find

$$u_i = 2^i - \sum_{j=0}^{i-1} 2^{i-j} w_j = 2^i (1 - \sum_{j=0}^{i-1} 2^{-j} w_j). \tag{3.30}$$

This implies the existence of a solution such that $u_0, \ldots, u_{k-1} \geq 0$ and $u_k = 0$, since $w \in \mathcal{V}_k$ and $\mathrm{cn}(w) = \sum_{j=0}^{k} 2^{-j} w_j = 1$. If W denotes the distribution of a family vector generated from the model described above, then the probability of W taking value $w \in \mathcal{V}_k$ is

$$\begin{aligned} &\mathbb{P}(W = w) \\ &= \mathbb{P}(1 - \beta(1, a_0, b_0) = w_0, \ldots, u_{k-1} - \beta(u_{k-1}, a_{k-1}, b_{k-1}) = w_{k-1}) \\ &= \prod_{i=0}^{k-1} \mathbb{P}(u_i - \beta(u_i, a_i, b_i) = w_i) = \prod_{i=0}^{k-1} \mathbb{P}(\beta(u_i, b_i, a_i) = w_i) \end{aligned} \tag{3.31}$$

where the last relation is possible thanks to the properties of beta-binomial distributions. Cellular development under the beta-binomial model is illustrated in Fig. 3.7.

Next, we introduce the recovery proportion and assume that each cell of $w \in \mathcal{V}_k$ is independently sampled with probability r. As a consequence, the number of cells sampled in generation $i = 0, \ldots, k$ is distributed as $\mathrm{B}(w_i, r)$, a binomial variable with w_i trials and success probability r. Let V be the random variable for the sampled family vectors. Then, the conditional probability of observing the sample $v \in \mathcal{S}_k$ given

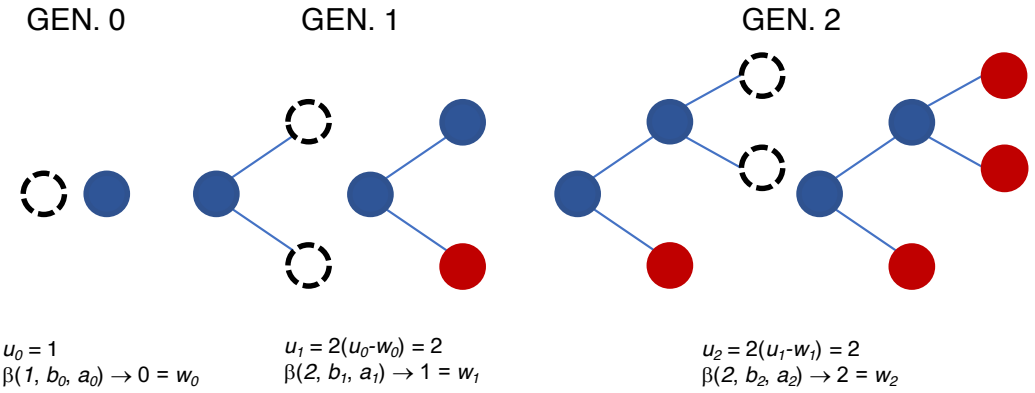

FIGURE 3.7: **Cellular expansion under the beta-binomial model.** Example of clonal development, starting from $u_0 = 1$, the one progenitor cell in generation 0 (left) that progress through the generations (from left to right) until all cells have reverted to quiescence (red dots), instead of dividing (blue dots). Given $i = \{0, 1, 2\}$, u_i is the number of cells born in generation i. Of these, the number of cells that revert to quiescence is w_i, which is the outcome of a beta-binomial distribution $\beta(u_i, b_i, a_i)$ where $a_i = p_i(1-\rho)/\rho$, $b_i = (1-p_i)(1-\rho)/\rho$. p_i and ρ are the parameters of the beta-binomial model.

$w \in \mathcal{V}_k$, the family vector subjected to sampling, is

$$\begin{aligned}
&\mathbb{P}(V = v | W = w) \\
&= \mathbb{P}\left(\mathrm{B}(w_0, r) = v_0, \ldots, \mathrm{B}(w_k, r) = v_k \middle| \sum_{i=0}^{k} \mathrm{B}(w_i, r) > 0, W = w \right) \\
&= \frac{\left(\prod_{i=0}^{k} \binom{w_i}{v_i} \right) r^{\sum_{i=0}^{k} v_i} (1-r)^{\sum_{i=0}^{k} w_i - v_i}}{1 - (1-r)^{\sum_{i=0}^{k} w_i}} \mathbb{1}_{C(w,v)},
\end{aligned} \tag{3.32}$$

where $C(w, v)$ is the shorthand for the event $\{v_0 \leq w_0; \ldots; v_k \leq w_k; \sum_{i=0}^{k} v_i > 0\}$. Note that the probability above is conditioned on the event that at least one cell is recovered from the clone w (of probability $1 - (1-r)^{\sum_{i=0}^{k} w_i}$). In fact, the observation of $\{V = (0, \ldots, 0)\}$ is censored, since a clone is not reported as sampled unless one of its cells is recovered (Fig. 3.8).

Putting (3.31) and (3.32) together, we finally derive the distribution for V by the law of total probabilities, that is

$$\mathbb{P}(V = v) = \sum_{w \in \mathcal{V}_k} \mathbb{P}(W = w)\, \mathbb{P}(V = v | W = w)\,, \tag{3.33}$$

which depends on the parameters $r, p_0, \ldots, p_k, \rho$. First, $r, p_0, \ldots, p_k$ are inferred by method of moments. As such, the sample proportion r is estimated by $\hat{r}$ as the average, across culture wells, of the percentage of beads or volume recovered which are part of the experimental method for this specific purpose. For any fixed $i = 0, \ldots, k$, p_i is the

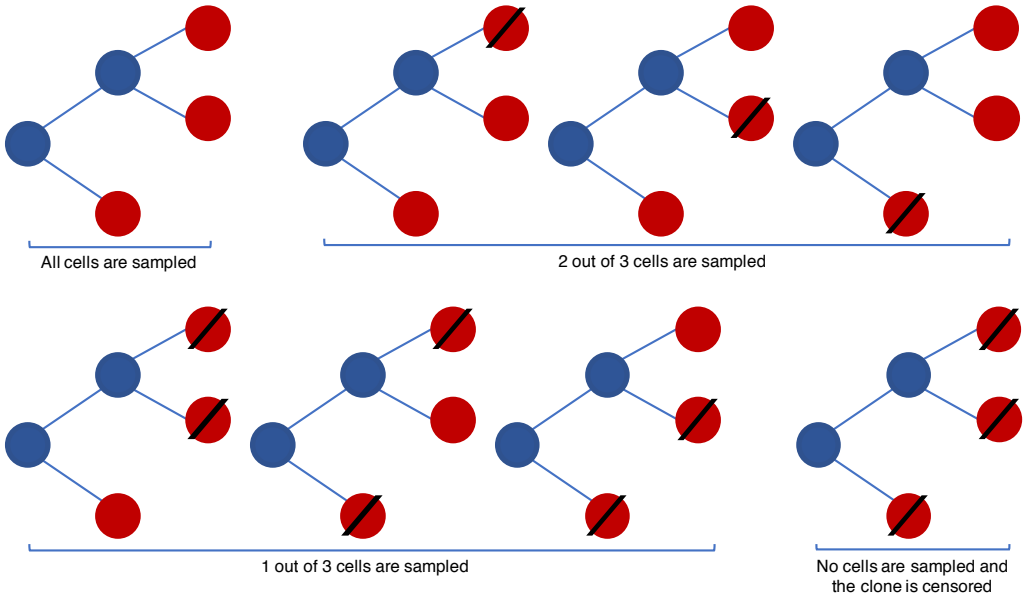

FIGURE 3.8: **Sampling effect on family vectors.** Given the clone with underlying family vector $(0, 1, 2)$ (top left), the number of possible sampled family vectors derived from it are: one, if all cells are sampled (top left); three, if two cells are sampled (top right); three, if one cell is sampled (bottom left). If none of the cells are sampled, the clone is not recovered and the observation of sampled family vector $(0, 0, 0)$ is censored.

probability of that a cell in generation i divides to generation $i+1$. Thus we calculate the estimate $\hat{p}_i$ of p_i from the sampled clones $z_1, \ldots, z_N$, as the expected proportion of cells that divide to generation $i+1$, out of those that reached generation i, namely

$$\hat{p}_i = \frac{\sum_{l=1}^{N} \sum_{j=i+1}^{k} 2^{i-j} z_{lj}}{\sum_{l=1}^{N} \sum_{j=i}^{k} 2^{i-j} z_{lj}}. \tag{3.34}$$

Finally, ρ is estimated by the $\hat{\rho}$ that maximises the log-likelihood of the clonal range data. That is

$$\begin{aligned}\hat{\rho} &= \underset{\rho\in[0,1]}{\operatorname{argmax}} \sum_{n=1}^{N} \log \mathbb{P}(\operatorname{range}(V) = \operatorname{range}(v_n)) \\ &= \underset{\rho\in[0,1]}{\operatorname{argmax}} \sum_{n=1}^{N} \log \left(\sum_{w\in\mathcal{V}_k} c(w, \operatorname{range}(v_n)) \prod_{i=0}^{k-1} \mathbb{P}\left(\beta\left(u_i, (1-\hat{p}_i)\frac{1-\rho}{\rho}, \hat{p}_i \frac{1-\rho}{\rho}\right) = w_i \right) \right)\end{aligned} \tag{3.35}$$

with

$$c(w, j) = \sum_{v\in\mathcal{S}_k} \mathbb{1}_{\{\operatorname{range}(v)=j\}} \mathbb{1}_{C(w,v)} \frac{\left(\prod_{i=0}^{k} \binom{w_i}{v_i}\right) \hat{r}^{\sum_{i=0}^{k} v_i} (1-\hat{r})^{\sum_{i=0}^{k} w_i - v_i}}{1 - (1-\hat{r})^{\sum_{i=0}^{k} w_i}} \tag{3.36}$$

for every $j = 0, \ldots, k-1$.

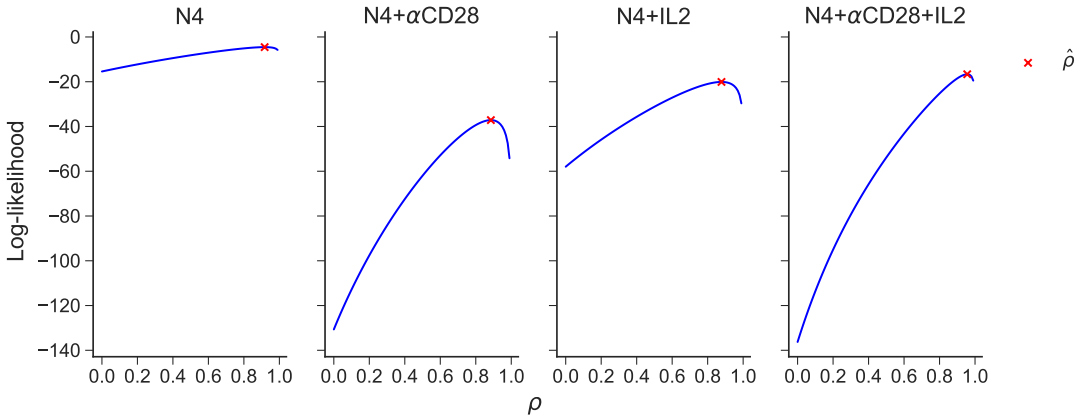

FIGURE 3.9: **Likelihood function of clonal range under the beta-binomial model.** Log-likelihood of parameter $\rho \in [0, 1]$ is plotted for range distributions under conditions N4, N4+αCD28, N4+IL2 and N4+αCD28+IL2 (from left to right), for multiplex data as in Fig. 2.7, assuming they progressed as in the beta-binomial model. Red crosses indicate the best fit solutions of $\hat{\rho}$ (3.35).

For $k \leq 6$, the cardinality of $\mathcal{V}_k$, written $|\mathcal{V}_k|$, is small enough so that the summation over all family vectors $w \in \mathcal{V}_k$ is computationally feasible. The optimisation can then be achieved using Python software's standard tools (i.e. the "scipy.optimize.minimize" function from Scipy library version 0.19.1) and this is the case for the data acquired in Chapter 2. As evidenced in Fig. 3.9, $\hat{\rho}$ is well determined for the multiplexed data under analysis.

As shown in Fig. 2.7c of Chapter 2, $\hat{\rho}$ offers a fit distribution that recapitulates range empirical distribution accurately. In particular, $\hat{\rho} \geq 0.8$ for all four stimulatory conditions tested, supporting the view that cell divisions are highly correlated within generations and the observed clonal regularity is not a mere result of sampling, but a biological feature inherited along the family. As a consequence, clonal expansion appears to be programmed by the progenitor cell at the moment of activation, raising questions on how such programmed mechanism occurs. In this regards, in Section 2.3.5 we investigated how stimulatory signals are integrated by the initial cell thus, in the next section, we present the mathematical framework that enables such study.

3.4 Rooted tree operations

3.4.1 Motivation

In the recent study from our collaborators, *in vitro* and *in vivo* experiments showed that stimulatory signals (i.e. costimulus αCD28 and cytokine IL2) contributed linearly to the average expansion of a population of T cells (Marchingo et al., 2014) with mean

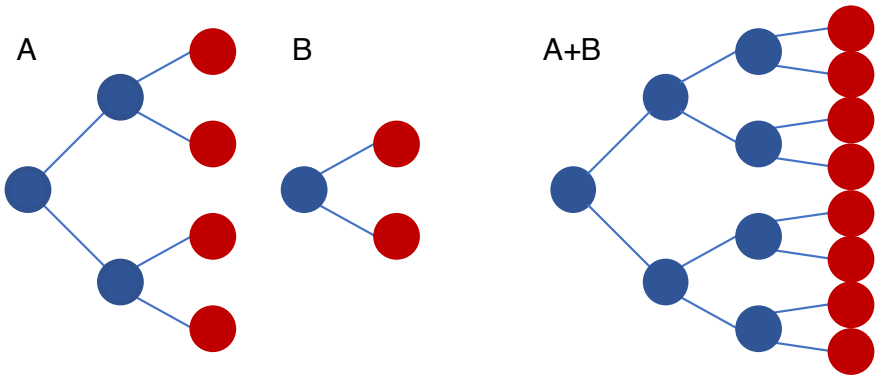

FIGURE 3.10: **Linear integration, regular case.** Two regular rooted trees A, with all leaves in generation 2 (left), and B, with all leaves in generation 1 (centre), are associated to clonal families developed under two distinct stimulatory conditions. If these stimuli were provided together and linearly integrated, the resulting rooted tree $A+B$ (right) would be regular with all leaves in generation 3, as the linear sum of the contributions from A and B, singularly, to the expansion. For example, $A+B$ can be defined by appending copies of B to the leaves of A or vice versa.

and variance that were both additive. In Chapter 2, using multiplex clonal assay data, we addressed the same question at the single cell level. In the following, we describe the mathematics that enabled this analysis. The main goal is to build a framework for rooted tree addition that allows the comparison between family trees stimulated by different signals.

To illustrate the complication we encounter, consider two family trees relative to clones receiving stimulation A and B respectively, plus a third one, generated by those stimuli combined, namely $A+B$. How should we "linearly" merge the trees A and B together to compare them with $A+B$? As an answer, we introduce a novel operation to describe tree addition, where the linear property is defined with respect to the number of divisions realised.

Suppose the family trees generated under stimuli A and B are regular, with the first stimulus providing two rounds of division and the second only one. Then, if the stimuli combined were linearly integrated, the family tree of $A+B$ would be regular as well, presenting three rounds of divisions, as in Figure 3.10. Many operations between the trees would fulfil this requirement. For example, we mention the appending of copies of B to each terminal node of A, or vice versa.

The difficulties arise when the trees of A and B are irregular, as in Fig. 3.11. In this instance, the appending of copies of one tree to the end of the other would not even be commutative. Furthermore, any partition of the initial trees and the procedural appending of the consequent subtrees that does not violate the original order (i.e. no mother cell copy can be descendant from a copy of its progeny), would produce several

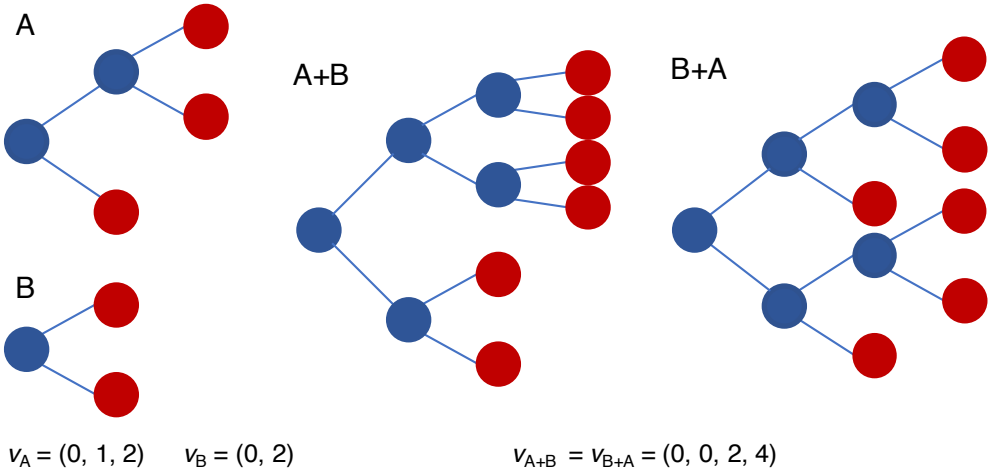

FIGURE 3.11: **Linear integration, irregular case.** Two rooted trees A, irregular (top left), and B, regular (bottom left), are associated to two clonal families developed under two distinct stimulatory conditions. To the right, the rooted tree $A + B$ (respectively $B + A$) results from appending copies of B (respectively A) to the leaves of A (respectively B), the same operation proposed in Fig. 3.10 that respected linearity in the regular case. Since A is irregular, this operation is not commutative as $A+B$ is different from $B+A$. However, both these trees present $(0, 0, 2, 4)$ for the family vectors v_{A+B} and v_{B+A}, which results from the discrete convolution of v_A and v_B, the family vectors associated to A and B ($(0, 1, 2)$ and $(0, 2)$, respectively).

notions of tree addition that respect the linearity observed in the regular case (see Fig. 2.9).

Strikingly, one common trait emerges from this class of operations: the family vector relative to the resulting tree is equal to the discrete convolution between the family vectors of the initial trees irrespective of the interlacement choice, which may produce trees that are not isomorphic as in the example of Fig. 3.11. This property will result from Theorem 3.27, justifying the choice of discrete convolution as representative for the class of tree operations that satisfy linearity and commutativity, even for irregular families, and whose result does not violate the original order of the combined trees. As a consequence, to combine two clones, only their family vectors are required, which is the information provided by the multiplex clonal assay.

This rationale will not provide us with a specific procedure of trees combination. However, while this lack of uniqueness seems inconvenient, it reflects the absence of knowledge concerning the order –if any exists– for the integration of stimulatory impulses: for example, if the choice of tree operation were the appending of copies of the second tree B to the terminal nodes of the first A, as previously mentioned, this would have the biological implication that signal B is integrated only after A stimulation is depleted.

3.4.2 Definition

The aim of this section is to formalise the class of operations on rooted trees described in the previous paragraph and show their equivalence in terms of the resulting family vectors. To do so we define a novel generating function associated to a rooted tree that is based on its leaves count and distance from the root, thus termed the leaf-depth generating function, and exploit the framework of generating functions (Wilf, 1990; Flajolet and Sedgewick, 2009). Subsequently, we introduce two basic operations, of tree appending (Fig. 3.12) and subtree removal (Fig. 3.13), and determine their action on the leaf-depth generating functions relative to the trees being transformed. We will then proceed with the definition of sequential appending and tree insertion. These satisfy the property of linear integration for regular trees (as in Fig. 3.10) and respect the original order of the trees they operate on. Finally, we show that the leaf-depth generating functions produced from these operations are the same.

Definition 3.18 (Leaf-depth generating function from a rooted tree)**.** Let $A = (V, E)$ be a rooted tree. Let $L(A) \subseteq V$ be the set of leaves of A and, for every $v \in V$, let $g(A, v) \in \mathbb{N}_0$ indicate the generation of v in A. Then, the leaf-depth generating function associated to A is the polynomial $G_A \in \mathbb{N}_0[x]$, with integer coefficients and variable x, defined as

$$G_A(x) = \sum_{v \in L(A)} x^{g(A,v)} = \sum_{i \geq 0} a_i x^i \tag{3.37}$$

where a is the vector of leaves count of A, that is

$$a_i = \big|\{v \in L(A) \colon g(A, v) = i\}\big|, \tag{3.38}$$

for every $i \geq 0$, with $|S|$ denoting the number of elements, or cardinality, of any set S.

Note that a family vector, as introduced in Section 3.2.2, is a vector of leaves count for a full binary tree.

Definition 3.19 (Tree appending)**.** Given $A = (V_A, E_A)$ and $B = (V_B, E_B)$, two trees rooted at $r_A \in V_A$ and $r_B \in V_B$, respectively, with $V_A \cap V_B = \emptyset$ and $v \in L(A)$ a leaf node of A, we define the appending of B to A at v as the rooted tree $\mathcal{A}(A, B, v) = (V', E')$ with root r_A, such that

$$\begin{aligned} V' &= V_A \cup V_B \setminus \{r_B\} \\ E' &= E_A \cup E_B \cup R' \setminus R, \end{aligned} \tag{3.39}$$

where $R \subseteq E_B$ is the set of edges with endpoint in r_B and R' is the set of edges connecting v to each neighbour of r_B.

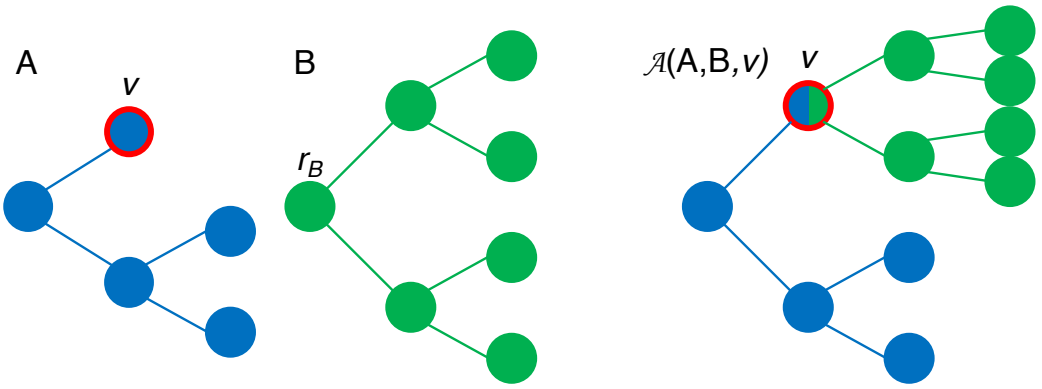

FIGURE 3.12: **Tree appending.** $\mathcal{A}(A, B, v)$ (right, blue and green) is the rooted tree obtained by appending the tree B, with root at r_B, (centre, green) to the rooted tree A (left, blue) at the node v (red outline).

Definition 3.20 (Subtree removal)**.** Given a rooted tree $A = (V_A, E_A)$ with root at $r_A \in V_A$, $w \in V_A$ a node of A and $A(w) = (V_w, E_w)$ the subtree of A rooted at w, we define $A \setminus A(w) = (V', E')$ as the tree rooted at r_A obtained by the removal of $A(w)$ from A, as

$$\begin{aligned} V' &= V_A \setminus V_w \cup \{w\} \\ E' &= E_A \setminus E_w. \end{aligned} \tag{3.40}$$

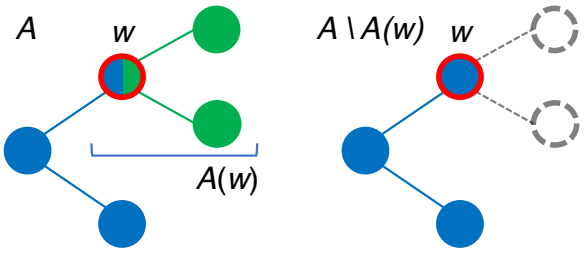

FIGURE 3.13: **Subtree removal.** $A \setminus A(w)$ (right, blue) is the rooted tree obtained by removing $A(w)$ (left, green), the subtree of A rooted at w (red outline), from A (left, blue and green).

The following lemma shows the action, on the leaf-depth generating functions, induced by tree appending and subtree removal.

Lemma 3.21 (Leaf-depth generating function from tree appending and removal)**.** *Let A and B be rooted trees with leaf-depth generating functions G_A and G_B and let $v \in L(A)$. The leaf-depth generating function of the tree obtained appending B to A at v, namely $C = \mathcal{A}(A, B, v)$, is*

$$G_C(x) = G_A(x) + x^{g(A,v)}(G_B(x) - 1). \tag{3.41}$$

In particular, given w a node of A, then the leaf-depth generating function of the tree obtained removing the subtree $A(w)$ from A, namely $D = A \setminus A(w)$, is

$$G_D(x) = G_A(x) - x^{g(A,w)}(G_{A(w)}(x) - 1). \tag{3.42}$$

Proof. To prove (3.41), we reason on the number and position of the leaves in C. In fact, the leaf v in generation $g(A, v)$ is removed, while each leaf $w \in L(B) \subseteq L(C)$ is added, with a path from the root r_A, in C, of length $g(C, w) = g(A, v) + g(B, w)$. This implies

$$G_C(x) = G_A(x) - x^{g(A,v)} + \sum_{w \in L(B)} x^{g(A,v)+g(B,w)} = G_A(x) + x^{g(A,v)}(G_B(x) - 1). \tag{3.43}$$

To show (3.42), it suffices to notice that $A = \mathcal{A}(D, A(w), w)$ and apply (3.41). □

We can now define the sequential appending of copies of one tree to the terminal nodes of another (Fig. 3.14) as a simple iteration of the tree appending operation.

Definition 3.22 (Sequential appending)**.** Let A be a tree rooted at r_A with a finite number n of leaves and let B be another rooted tree. Given any indexing of the leaves of A, namely $L(A) = \{v_i\}_{i=1}^n$, the sequential appending of copies of B to A is defined as

$$\mathcal{I}(A, B, L(A)) = \mathcal{A}(\cdots \mathcal{A}(\mathcal{A}(A, B_1, v_1), B_2, v_2), \cdots, B_n, v_n). \tag{3.44}$$

This is the tree, rooted at r_A, obtained by appending n different copies $B_1, \ldots, B_n$ of B to A. If the number of leaves is countably infinite then, given any indexing of the leaves $L(A) = \{v_i\}_{i \geq 1}$, $\mathcal{I}(A, B, L(A))$ is defined as the limit, for n large, of the sequence $\mathcal{A}(\cdots \mathcal{A}(\mathcal{A}(A, B_1, v_1), B_2, v_2), \cdots, B_n, v_n)$.

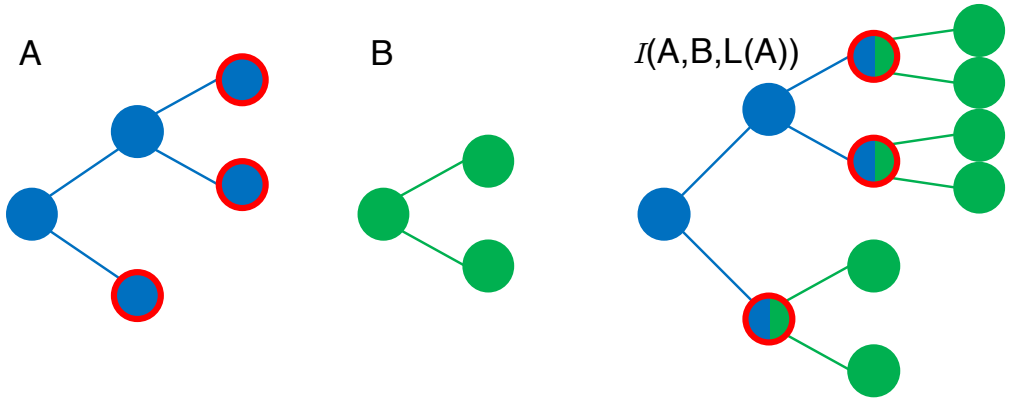

FIGURE 3.14: **Sequential appending.** $\mathcal{I}(A, B, L(A))$ (right, blue and green) is the rooted tree obtained by appending copies of the rooted tree B (centre, green) to the rooted tree A (left, blue) at its leaves $L(A)$ (red outline).

Note that choosing an indexing different to $\{v_i\}_{i\geq 1}$ does not modify the final result, as every appending operation in one node does not affect the appending operations of others. It is also necessary to introduce the sequence $B_1 = (V_{B_1}, E_{B_1}), \ldots, B_n = (V_{B_n}, E_{B_n})$ of copies of $B = (V_B, E_B)$, so that $V_{B_i} \cap V_{B_j} = \emptyset$ for every $i, j \in \{1, \ldots, n\}$ and $i \neq j$, and each tree appending iteration adds a different tree, in accordance with (3.39). In particular, it holds that

$$G_{B_i}(x) = G_B(x), \tag{3.45}$$

as B_i is isomorphic to B for every $i \geq 0$.

If B is the degenerate rooted tree, namely $V_B = \{r_B\}$ with r_B the root and $E_B = \emptyset$, then B is the identity element for the operation of tree appending, that is

$$\mathcal{A}(A, B, v) = A \tag{3.46}$$

for every $v \in L(A)$. Therefore, the degenerate tree is also the identity element for the sequential appending, that is

$$\mathcal{I}(A, B, L(A)) = A, \tag{3.47}$$

and, in particular, $G_B(x) = 1$.

The leaf-depth generating function from the sequential appending is immediately derived through the tree appending properties.

Proposition 3.23 (Leaf-depth generating function from sequential appending)**.** *Given two rooted trees A and B, the leaf-depth generating function associated to $C = \mathcal{I}(A, B, L(A))$ is*

$$G_C(x) = G_A(x)G_B(x). \tag{3.48}$$

In particular, if $G_A(x) = \sum_{k\geq 0} a_k x^k$ and $G_B(x) = \sum_{k\geq 0} b_k x^k$, then

$$G_A(x)G_B(x) = \sum_{k\geq 0} (a * b)_k x^k, \tag{3.49}$$

where $(a*b)$ indicated the discrete convolution of a and b, namely $(a*b)_k = \sum_{i=0}^{k} a_i b_{k-i}$

Proof. Given the indexing $L(A) = \{v_i\}_{i\geq 1}$, let

$$C_n = \mathcal{A}(\cdots \mathcal{A}(\mathcal{A}(A, B_1, v_1), B_2, v_2) \cdots, B_n, v_n) \tag{3.50}$$

for every $n \geq 1$. Then, by Lemma 3.21, we obtain the recursive relation

$$\begin{aligned} G_{C_1}(x) &= G_A(x) + x^{g(A,v_1)}(G_B(x) - 1) \\ &\vdots \\ G_{C_n}(x) &= G_{C_{n-1}}(x) + x^{g(A,v_n)}(G_B(x) - 1). \end{aligned} \tag{3.51}$$

This implies

$$G_{C_n}(x) = G_A(x) + \sum_{i=1}^{n} x^{g(A,v_i)}(G_B(x) - 1), \tag{3.52}$$

from which we deduce

$$\begin{aligned} G_C(x) &= G_A(x) + \sum_{i \geq 1} x^{g(A,v_i)}(G_B(x) - 1) \\ &= G_A(x) + G_A(x)(G_B(x) - 1) = G_A(x)G_B(x). \end{aligned} \tag{3.53}$$

The fact that $G_A(x)G_B(x) = \sum_{k\geq 0}(a * b)_k x^k$ is deduced by the relations between discrete convolution and polynomial multiplication. □

To define the operation of sequential tree insertion which extends the sequential appending, first we need to identify the set of nodes where a tree can be inserted without violating linearity property (Fig. 3.10) and the original structure of the two tree combined (Fig. 3.15).

Definition 3.24 (Leaves partition)**.** Let $A = (V_A, E_A)$ be a rooted tree. We call $P \subseteq V_A$ a leaves partition of A, if each leaf of A is a descendent of one and only one node in P. Additionally, we call A_P the tree rooted at r_A that is obtained removing each subtree A_v from A, for $v \in P$.

Definition 3.25 (Tree insertion)**.** Let A and B be two rooted trees. Given P, a leaves partition of A such that $|P| = n$ and any indexing of P, namely $P = \{v_i\}_{i=1}^n$, the sequential insertion of copies of B into A at P is defined as

$$\mathcal{I}(A, B, P) = \mathcal{I}(\cdots \mathcal{I}(\mathcal{I}(A_P, Z_1, v_1), Z_2, v_2) \cdots, Z_n, v_n), \tag{3.54}$$

where $B_1, \ldots, B_n$ are different copies of B and $Z_i = \mathcal{I}(B_i, A_{v_i}, L(B_i))$ for $i = 1, \ldots, n$. If $|P|$ is countably infinite and given any indexing $P = \{v_i\}_{i\geq 1}$, then $\mathcal{I}(A, B, P)$ is defined as the limit, for n large, of the sequence $\mathcal{I}(\cdots \mathcal{I}(\mathcal{I}(A_P, Z_1, v_1), Z_2, v_2) \cdots, Z_n, v_n)$.

We highlight that each leaf of A is required to descend from at least one node of the leaves partition P so that the insertion of copies of a tree B into A at the nodes of P respects the property of linearity. Furthermore, to ensure that the original order of the

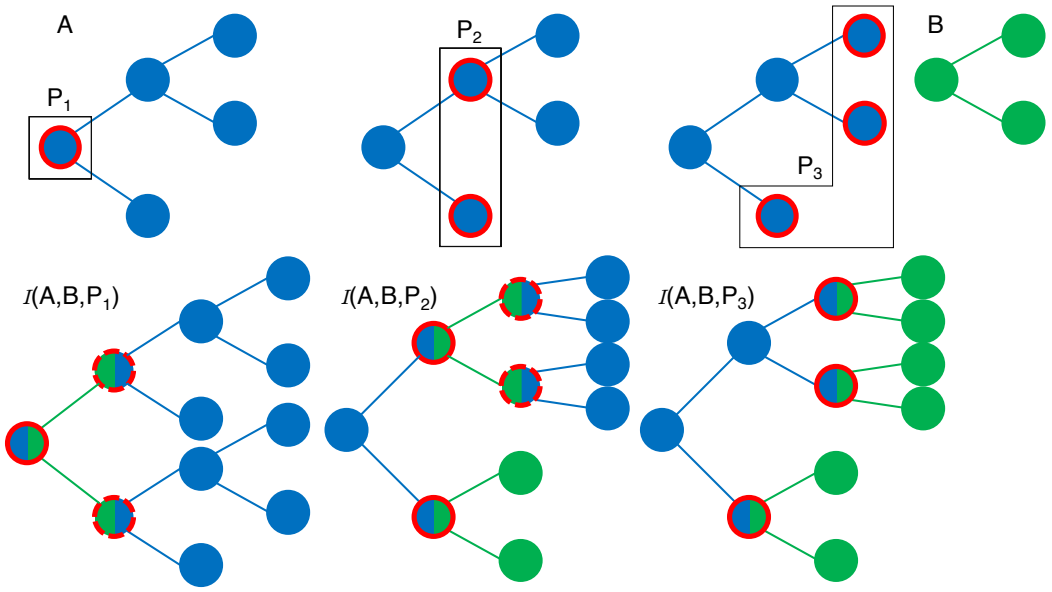

FIGURE 3.15: **Leaves partition and tree insertion.** For the rooted tree A (top, blue), all possible leaves partitions P_1, P_2 and P_3 are shown (nodes within a box). For each of these partitions, copies of the rooted tree B (top, green) are inserted into A at the nodes of the partition (full red outline), resulting in $\mathcal{I}(A, B, P_1)$, $\mathcal{I}(A, B, P_2)$ and $\mathcal{I}(A, B, P_3)$ (bottom, blue and green). For P_1 and P_2, at least one non-degenerate subtree of A is rooted at a node of the partition, so every such subtree must be sequentially appended at the leaves of the copy of B (dashed red outline) that is inserted in the root node of the subtree.

trees A and B is preserved, it is necessary that each leaf of A descend from only one node of P. Otherwise new familial relationships may be established between copies of nodes that were not present in their original trees. Both cases are illustrated in Fig. 3.16. In particular, sequential appending is included by setting $P = L(A)$. Similarly as for that operation, tree insertion does not change with different indexing of the leaves partition P. Furthermore, the degenerate tree $B = (\{r_B\}, \emptyset)$ is the identity element for tree insertion, that is

$$\mathcal{I}(A, B, P) = A, \tag{3.55}$$

as in (3.47), for any P leaves partition of A.

From the definition of leaves partition, we have the following useful property for the leaf-depth generating function.

Corollary 3.26 (Leaf-depth generating function arising from a leaves partition)**.** *Let A be a rooted tree and P a partition of its leaves. Then*

$$G_A(x) = \sum_{v \in P} x^{g(A,v)} G_{A_v}(x). \tag{3.56}$$

Proof. The statement is deduced by applying (3.41) from Lemma 3.21 and noting that A results from the appending of A_v to v for every $v \in P$. □

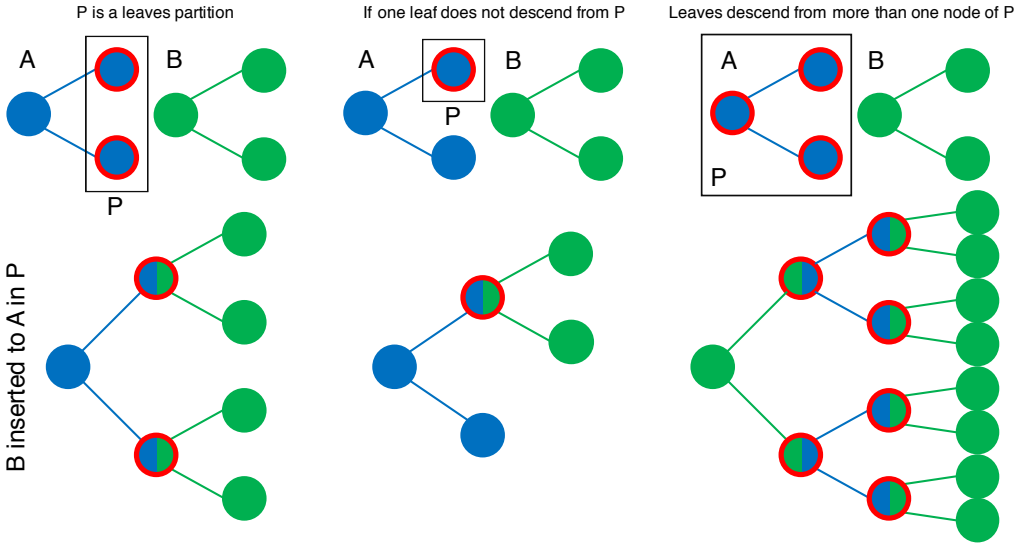

FIGURE 3.16: **Leaves partition requirements.** The regular rooted tree B (top, green) is inserted into the regular rooted tree A (top, blue) at P (red outline nodes within a square). If P is a leaves partition (left) the resulting rooted tree (bottom left) respects linearity for the regular case as in Fig. 3.10. If at least one leaf node of A does not descend from P (centre), the rooted tree obtained (bottom centre) does not satisfy the linearity property. If a node of P descends from another node of P (right), the insertion produces a rooted tree (bottom right) that does not comply with linearity and also violates the original arrangement of the trees. To illustrate the latter fact, note that the leaves from the resulting tree, which are copies of the leaves of B, descend from the leaves in generation 1, that are copies of the leaves of B. As this familial relationship was absent in the original tree B (each leaf of B is not descendants of the other), the order of B is not respected by the operation considered.

We are ready to prove the main result of this section, which shows how the operation of tree insertion produces trees that share the same leaf-depth generating function, irrespective of the choice for the leaves partition P.

Theorem 3.27 (Leaf-depth generating function from tree insertion)**.** *Let A and B be two rooted trees and let P be a leaves partition of A. Then the leaf-depth generating function associated to $C = \mathcal{I}(A, B, P)$ is*

$$G_C(x) = G_A(x)G_B(x). \tag{3.57}$$

In particular, G_C does not depend on the choice of P.

Proof. Given the indexing $P = \{v_i\}_{i\geq 1}$, let

$$C_n = \mathcal{I}(\cdots \mathcal{I}(\mathcal{I}(A_P, Z_1, v_1), Z_2, v_2) \cdots, Z_n, v_n) \tag{3.58}$$

for every $n \geq 1$, with $Z_i = \mathcal{I}(B_i, A_{v_i}, L(B_i))$ for $i \geq 1$. Then, by Lemma 3.21 and Proposition 3.23, we obtain the recursive relation

$$\begin{aligned} G_{C_1}(x) &= G_{A_P}(x) + x^{g(A_P,v_1)}(G(Z_1) - 1) \\ &= G_{A_P}(x) + x^{g(A_P,v_1)}(G_B(x)G_{A_{v_1}}(x) - 1) \\ &\vdots \\ G_{C_n}(x) &= G_{C_{n-1}}(x) + x^{g(A_P,v_n)}(G_B(x)G_{A_{v_n}}(x) - 1). \end{aligned} \tag{3.59}$$

This implies

$$G_{C_n}(x) = G_{A_P}(x) + \sum_{i=1}^{n} x^{g(A_P,v_i)}(G_B(x)G_{A_{v_i}}(x) - 1), \tag{3.60}$$

from which we deduce

$$G_C(x) = G_{A_P}(x) - \sum_{i\geq 1} x^{g(A_P,v_i)} + G_B(x) \sum_{i\geq 1} x^{g(A_P,v_i)} G_{A_{v_i}}(x) = G_A(x)G_B(x), \tag{3.61}$$

where the last equality holds thanks to the definitions of leaves partition and by Corollary 3.26. □

As tree insertion makes use of tree appending in its definition, specifically for $Z_i = \mathcal{I}(B_i, A_{v_i}, L(B_i))$ with $i = 1, \ldots, n$, it allows for greater generalisation in a similar way as we extended tree insertion from tree appending. In fact, given A, B, B_i for $i \geq 1$ and $P = \{v_i\}_{i\geq 1}$ as in Theorem 3.27 and let, for each $i \geq 1$, P_{v_i} be a leaves partition of B. Then each Z_i in Definition 3.25 can be replaced by

$$Z_i = \mathcal{I}(B_i, A_{v_i}, P_{v_i}), \tag{3.62}$$

leading to a more nuanced definition of tree insertion, that we write as $\mathcal{I}(A, B, \{P, (P_v)_{v\in P}\})$ with abuse of notation. This new operation first demands the insertion of A_{v_i} into B_i at P_{v_i}, to obtain Z_i, and the subsequent insertion of Z_i into A_P at P. As we are replacing $Z_i = \mathcal{I}(B_i, A_{v_i}, L(B))$ with $Z_i = \mathcal{I}(B_i, A_{v_i}, P_{v_i})$, that have identical leaf-depth generating functions by Theorem 3.27, the leaf-depth generating function for $\mathcal{I}(A, B, \{P, (P_v)_{v\in P}\})$ would still be $G_A(x)G_B(x)$. This line of reasoning can be repeated indefinitely, by replacing each $\mathcal{I}(B_i, A_{v_i}, P_{v_i})$ with $\mathcal{I}(B_i, A_{v_i}, \{P_{v_i}, (P_{v_i w})_{w\in P_{v_i}}\})$, where $P_{v_i w}$ is now a leaves partition of A_{v_i} for every $w \in P_{v_i}$ and $i \geq 0$, and so on.

In conclusion, any combination of two rooted trees A and B, based on tree insertion operations, results in a tree whose leaf-depth generating function is $G_A(x)G_B(x) = \sum_{k\geq 0}(a*b)_k x^k$, thus supporting our claim that discrete convolution is the encompassing

notion of addition that we initially sought. In the next section, we will examine the relation of this operation with the statistics of clonal expansion presented in Section 3.2.3.

3.4.3 Linearity of expansion statistics

As discrete convolution was appointed to satisfy linearity for regular trees addition, with respect to the number of divisions, we expect that linearity also arises for expansion statistics from Section 3.2.3, at least in in the regular case. Actually, this is holds true even for family vectors that are irregular.

Proposition 3.28 (Linearity of expansion statistics with discrete convolution)**.** *Given $k, h \geq 1$ and two vectors $a \in \mathbb{N}_0^k$, $b \in \mathbb{N}_0^h$, then*

$$\begin{aligned} \mathrm{mDD}(a * b) &= \mathrm{mDD}(a) + \mathrm{mDD}(b), \\ \mathrm{maxDD}(a * b) &= \mathrm{maxDD}(a) + \mathrm{maxDD}(b), \\ \mathrm{minDD}(a * b) &= \mathrm{minDD}(a) + \mathrm{minDD}(b), \\ \mathrm{agDD}(a * b) &= \mathrm{agDD}(a) + \mathrm{agDD}(b). \end{aligned} \tag{3.63}$$

Proof. We prove the relations in order. For mDD, we first define $G_A(x) = \sum_{i=0}^k a_i x^i$, $G_B(x) = \sum_{j=0}^h b_j x^j$ and $C = \mathcal{I}(A, B, L(A))$, so that $G_C(x) = G_A(x)G_B(x)$ (by Proposition 3.23) and $G'_C(x) = G'_A(x)G_B(x) + G_A(x)G'_B(x)$. We can now write

$$\mathrm{mDD}(a) = \frac{\sum_{i=0}^k i 2^{-i} a_i}{\sum_{i=0}^k 2^{-i} a_i} = \frac{G'_A(2^{-1})}{2G_A(2^{-1})},\ \mathrm{mDD}(b) = \frac{\sum_{j=0}^h j 2^{-j} b_j}{\sum_{j=0}^h 2^{-j} b_j} = \frac{G'_B(2^{-1})}{2G_B(2^{-1})} \tag{3.64}$$

and

$$\begin{aligned} \mathrm{mDD}(a * b) &= \frac{G'_C(2^{-1})}{2G_C(2^{-1})} = \frac{G'_A(2^{-1})G_B(2^{-1}) + G_A(2^{-1})G'_B(2^{-1})}{2G_A(2^{-1})G_B(2^{-1})} \\ &= \frac{\sum_{i=0}^k i 2^{-i} a_i}{\sum_{i=0}^k 2^{-i} a_i} = \frac{G'_A(2^{-1})}{2G_A(2^{-1})} + \frac{G'_B(2^{-1})}{2G_B(2^{-1})} = \mathrm{mDD}(a) + \mathrm{mDD}(b) \end{aligned} \tag{3.65}$$

to conclude.

For maxDD, let $m_a = \mathrm{maxDD}(a)$, $m_b = \mathrm{maxDD}(b)$ and $m_* = \mathrm{maxDD}(a * b)$. First note that

$$(a * b)_{m_a + m_b} = \sum_{i=0}^k \sum_{j=0}^h a_i b_j \mathbb{1}_{\{i+j=l\}} \geq a_{m_a} b_{m_b} > 0 \tag{3.66}$$

by definition of m_a and m_b, thus $m_* \geq k_1 + k_2$. Let $u \geq 0$ be such that $m_* = m_a + m_b + u$, then as $(a * b)_{m_*} = (a * b)_{m_a + m_b + u} > 0$ there exist i, j such that $a_i > 0$, $b_j > 0$ and

$i+j = m_a+m_b+u$. But, by definition of m_a and m_b, it holds that $i \leq m_a$ and $j \geq m_b$, therefore $0 \leq m_a+m_b+u \leq m_a+m_b$, implying $u = 0$ and $m_* = m_a+m_b$. Linearity for minDD and agDD are deduced following the same reasoning as for maxDD and mDD, respectively, using $\mathrm{agDD}(a) = G'_A(1)G_A(1)^{-1}$ for the latter. □

3.4.4 Consequences

We now have acquired all the elements to formalise the initial problem from Section 3.4.1, where we sought to question whether linear contribution of stimulatory signals for the clonal expansion arises at the single cell level.

Recollecting the setting from Chapter 2, we have three signals, that are N4 (antigenic), αCD28 (costimulatory) and IL2 (cytokine), and cell count vector data for clones stimulated with N4, N4+αCD28, N4+IL2 and N4+αCD28+IL2. We assume these vector are observation from the random variables V_{N4}, $V_{\mathrm{N4}+\alpha\mathrm{CD28}}$, $V_{\mathrm{N4+IL2}}$ and $V_{\mathrm{N4}+\alpha\mathrm{CD28+IL2}}$ respectively. We want to verify that stimuli αCD28 and IL2 contribute independently to the expansion, as hypothesised from Marchingo et al. (2014). In particular, we suppose that αCD28 is independent with N4, while IL2 is not, due to the documented correlations between antigen signal strength and IL2 receptor expression (Zehn et al., 2009; Wensveen et al., 2010; Gottschalk et al., 2012). If the integration of αCD28 with N4 or with N4+IL2 occur independently, then their contribution must be linear at the level of family trees, that is

$$V_{\mathrm{N4}+\alpha\mathrm{CD28}} \sim V_{\mathrm{N4}} * V_{\alpha\mathrm{CD28}} \tag{3.67}$$

and

$$V_{\mathrm{N4}+\alpha\mathrm{CD28+IL2}} \sim V_{\mathrm{N4+IL2}} * V_{\alpha\mathrm{CD28}}, \tag{3.68}$$

where the symbol $\sim$ refers to equality in distribution. Using maxDD as a description of the expansion and its linear property with respect to discrete convolution $*$ (by Proposition 3.28), we deduce that

$$\begin{aligned} &\mathrm{maxDD}(V_{\mathrm{N4}}) + \mathrm{maxDD}(V_{\mathrm{N4}+\alpha\mathrm{CD28+IL2}}) \\ &\qquad \sim \mathrm{maxDD}(V_{\mathrm{N4}}) + \mathrm{maxDD}(V_{\mathrm{N4+IL2}}) + \mathrm{maxDD}(V_{\alpha\mathrm{CD28}}) \\ &\qquad \sim \mathrm{maxDD}(V_{\mathrm{N4}+\alpha\mathrm{CD28}}) + \mathrm{maxDD}(V_{\mathrm{N4+IL2}}). \end{aligned} \tag{3.69}$$

Of note, the variables in the first and last terms of this relation are observables of the multiplex assay method. In addition, the strong clonal regularity (Section 2.3.4), supported from the beta-binomial model fits (Section 3.3.2), assures that deviations

of maxDD distribution on the sampled family vectors from the unsampled ones are negligible.

Now that the question of linearity of expansion effects of the stimuli αCD28 and IL2 is formally expressed (3.69), we seek to test it statistically. This problem can be formulated as

$$\mathbf{H}_0 \colon X_1 + X_2 \sim Y_1 + Y_2, \tag{3.70}$$

with X_1, X_2, Y_1, Y_2 independent, integer random variables in place of $\mathrm{maxDD}(V_{\mathrm{N4}})$, $\mathrm{maxDD}(V_{\mathrm{N4+}\alpha\mathrm{CD28+IL2}})$, $\mathrm{maxDD}(V_{\mathrm{N4+}\alpha\mathrm{CD28}})$, $\mathrm{maxDD}(V_{\mathrm{N4+IL2}})$ respectively. In the following chapter, we develop a statistical testing procedure for a larger class of null hypothesis, which includes (3.70). This was applied to obtain the p-values presented in Fig. 2.10 and 2.11 of Chapter 2.

Chapter 4

Testing for the sum of discrete and independent random variables

4.1 Abstract

In order to study how stimulatory signals αCD28 and IL-2 are integrated by naive CD8$^+$ T cells to enhance their expansion (Section 2.3.5 of Chapter 2), we interrogated the output from the new multiplex clonal assay (Section 2.3.1 of Chapter 2) through our novel framework of family vectors and rooted tree addition (Chapter 3). In Section 3.4.4 of Chapter 3, we showed that the hypothesis of independent signal integration can be assessed by statistically testing for equality in distribution between two sums of discrete and independent random variables. Since the experimental data available (Fig. 2.7 and 2.8 of Chapter 2) consisted of unequal sample size for each addend variable of the sum, the computation of classical χ^2 statistics (e.g. Pearson, 1900), which would not include all observations, results in loss of power, especially when samples are small. As an alternative, the nonparametric maximum likelihood estimator for the distribution of the sum of discrete and independent random variables, named convolution statistic, is proposed and its limiting normal covariance matrix defined. To challenge the null hypothesis of equality in distribution, the generalised Wald's method (Moore, 1977) is applied to define a testing statistic asymptotically distributed as a χ^2 with as many degrees of freedom as the rank of such covariance matrix. Rank analysis also reveals a connection with the roots of the probability generating functions associated to the addend variables of the sum. A simulation study is performed to compare the convolution test with Pearson's χ^2, and to provide usage guidelines.

4.2 Introduction

We examine the problem of testing the null hypothesis of equality in distribution, denoted $\sim$, for two linear models with distinct observables, that is

$$\mathbf{H}_0\colon a_0 + \sum_{i=1}^{k} a_i A_i \sim b_0 + \sum_{i=1}^{h} b_i B_i. \tag{4.1}$$

We assume that the random variables $A_1, \ldots, A_k, B_1, \ldots, B_h$ are bounded, independent, of possibly different distribution and all take values in a real lattice $\Lambda(\zeta) = \{\zeta u\colon u \in \mathbb{Z}\}$ for some $\zeta \in \mathbb{R}$ and that $a_0, b_0 \in \Lambda(\zeta)$, $a_1, \ldots, a_k, b_1, \ldots, b_h \in \mathbb{Z}$, the set of integers.

Equality in distribution between random variables can be tested by using statistics such as Pearson's χ^2 (Pearson, 1900) or the more general power-divergence family (Cressie and Read, 1984). The computation of these statistics, however, assumes that the number of observations of each of $A_1, \ldots, A_k$ (and $B_1, \ldots, B_h$) are equal. If that is not the case, it would seem that the data sets must be truncated for application of those methods, which could prove wasteful if samples come in unequal counts and their collection is costly or laborious.

For example, consider a problem in meta-analysis, where two studies are described by linear models with distinct independent variables, and we wish to test for equality in distribution between these models as in (4.1). In the simplest case, for $k = 2$, $h = 1$ with $\mathbf{H}_0\colon A_1 + A_2 \sim B_1$, the independent variables observed are n_1 distributed as A_1, n_2 as A_2 and n_3 as B_1, respectively noted $\{A_{11}, \ldots, A_{1n_1}\}$, $\{A_{21}, \ldots, A_{2n_1}\}$ and $\{B_{11}, \ldots, B_{1n_3}\}$, with $n_1, n_2, n_3 \in \mathbb{N}$. This scenario may arise because the independent variables are grouped differently in the studies (e.g., A_1 occurrences of event E_1, A_2 occurrences of event E_2, B_1 occurrences of any event E_1 or E_2) or because the model choice is different (e.g., model one is $A_1 + A_2$ and model two is $B_1 \sim f(A_1, A_2)$ for a given function f). Then, to test $\mathbf{H}_0$, Pearson's χ^2 could be computed using $B_1, \ldots, B_{n_3}$ and the data from A_1 and A_2 paired as, for example, $\{A_{11} + A_{21}, \ldots, A_{1m} + A_{2m}\}$ with $m = n_1 = n_2$, so that these m variables are independent and identically distributed as $A_1 + A_2$ to comply with Pearson's statistic assumptions. If observation sizes are unequal, e.g. $n_1 > n_2$, then $n_1 - n_2 > 0$ variables from $\{A_{11}, \ldots, A_{1n_1}\}$ could be excluded from the calculation of $\{A_{11} + A_{21}, \ldots, A_{1m} + A_{2m}\}$, now with $m = \min(n_1, n_2)$. But any pairing or variables exclusion are two choices that, either arbitrarily or randomly determined, influences the outcome of the test.

Using data from the multiplex clonal assay (Fig. 2.7 and 2.8 of Chapter 2), we seek to statistically test the hypothesis that the expansion impetus of two stimuli were integrated independently by cells when the signals were provided together, which had been

hypothesised in a previously published study (Marchingo et al., 2014). The experimental data obtained for Marchingo, Prevedello et al., (2016) was costly to produce, both in terms of manpower and reagents, and inherently came with distinct numbers of observations of all variables. Thus we sought to develop a statistical test that utilised all available data. The resulting test may prove useful in other fields, such as medicine for efficacy evaluation of combination therapies (e.g. Wolchok et al., 2013), which is a topic of growing interest (Editorial, 2017).

We first show that the null hypothesis in (4.1) is equivalent to one without the scalar multipliers, $\sum_{i=1}^{k} X_i \sim \sum_{i=1}^{h} Y_i$, which simplifies notation (Lemma 4.1). To obtain a test statistic that utilises all data and therefore outperforms methods that require equal sized data sets, in Section 4.3 we study the maximum likelihood estimator (MLE) for the probability mass vector (PMV) of $\sum_{i=1}^{k} X_i$. This transpires to be the discrete convolution of the empirical probability mass vector (EPMV) of each variable $X_1, \ldots, X_k$ and so we refer to it as the "convolution statistic" (Proposition 4.2).

We then derive the asymptotic distribution of the convolution statistic and build a testing procedure for both goodness-of-fit and equality in distribution versions (Proposition 4.3), leveraging the generalised Wald's method. This technique was introduced in Moore's work (Moore and Spruill, 1975; Moore, 1977; Mihalko and Moore, 1980; Moore, 1982), as an extension of Wald's method (Wald, 1943), to build χ^2 tests for statistics that are asymptotically normal distributed with a singular covariance matrix. It was subsequently adjusted in Hadi and Wells (1990), whose version we employ. Such methodology found applications in the fields of econometrics (Vuong, 1987; Andrews, 1987, 1988; Wilson and Koehler, 1991), biology (Zhang, 1999; Marchingo, Prevedello et al., 2016) and statistical theory (Tyler, 1981; Drost, 1989; Voinov et al., 2008).

In Section 4.4, we investigate the rank from the covariance matrix asymptotic of the convolution statistic (Theorem 4.7 and Corollary 4.8), which is the central problem for the derivation of a testing procedure through the generalised Wald's framework. Interestingly, such rank is related to the roots of the probability generating functions of $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_h$ (Lemma 4.6).

Finally, in Section 4.5 we provide simulated performance analysis for the convolution statistic against Person's χ^2, and in Section 4.6 we discuss the guidelines for its application.

4.3 Convolution statistic

To derive a statistic for the testing of $\mathbf{H}_0\colon a_0+\sum_{i=1}^{k} a_iA_i \sim b_0+\sum_{i=1}^{h} b_iB_i$, we begin by showing that this null hypothesis is equivalent to another in which the variables have finite, positive integer support and no parameters $a_0,\dots,a_k,b_0,\dots,b_h$ are present. As a consequence, we will work in this new setting as it facilitates the definition of the convolution statistic for the testing of $\mathbf{H}_0$, especially in regard to a simpler notation.

Lemma 4.1 (Null hypothesis simplification)**.** *Let $A_1,\dots,A_k$, $B_1,\dots,B_h$ be a sequence of finite and independent random variables that map the probability space $(\Omega,\mathcal{F},\mathbb{P})$ into a lattice $\Lambda(\zeta)$ with $\zeta\in\mathbb{R}$. Given the null hypothesis*

$$\mathbf{H}_0\colon a_0+\sum_{i=1}^{k} a_iA_i \sim b_0+\sum_{i=1}^{h} b_iB_i, \tag{4.2}$$

with $a_0,b_0\in\Lambda(\zeta)$, $a_1,\dots,a_k,b_1,\dots,b_h\in\mathbb{Z}$, there exists a sequence of positive, finite and independent random variables $X_1,\dots,X_k$, $Y_1,\dots,Y_h$ from the probability space $(\Omega,\mathcal{F},\mathbb{P})$ into $\{0,\dots,r_l\}$, respectively, with $r_l\in\mathbb{N}$ for $l=1,\dots,k+h$, $\mathbb{P}(X_i=0)>0$ for $i=1,\dots,k$, $\mathbb{P}(Y_j=0)>0$ for $j=1,\dots,h$, and such that

$$\mathbf{H}_0\colon \sum_{i=1}^{k} X_i \sim \sum_{i=1}^{h} Y_i, \tag{4.3}$$

is equivalent to (4.2)*.*

Proof. Without loss of generality, it is possible to shift from the lattice $\Lambda(\zeta)$ to the set of integers $\mathbb{Z}$ through the natural isomorphism $\phi\colon\Lambda(\zeta)\to\mathbb{Z}$, $\phi(\zeta u)=u$ for every $u\in\mathbb{Z}$. Using this function, we define $A'_i=a_i\phi(A_i)$ and $B'_j=b_j\phi(B_j)$ for $i=1,\dots,k$, $j=1,\dots,h$, so as to account for the multiplicative constants $a_1,\dots,a_k,b_1,\dots,b_h$ in the variables $A'_1,\dots,A'_k,B'_1,\dots,B'_h$ mapping Ω to $\mathbb{N}_0$, and reduce (4.2) into

$$\mathbf{H}_0\colon \phi(a_0)+\sum_{i=1}^{k} A'_i \sim \phi(b_0)+\sum_{i=1}^{h} B'_i. \tag{4.4}$$

Given $\tau_i=\min\{j\colon\mathbb{P}(A'_i=j)>0\}$ for $i=1,\dots,k$ and $\tau_{k+i}=\min\{j\colon\mathbb{P}(B'_i=j)>0\}$ for $i=1,\dots,h$, that are well defined since the variables $A'_1,\dots,A'_k,B'_1,\dots,B'_h$ are assumed finite, we rewrite (4.4) as

$$\mathbf{H}_0\colon \phi(a_0)+\sum_{i=1}^{k}\tau_i+\sum_{i=1}^{k}(A'_i-\tau_i) \sim \phi(b_0)+\sum_{i=1}^{h}\tau_{k+i}+\sum_{i=1}^{h}(B'_i-\tau_{k+i}). \tag{4.5}$$

For the null hypothesis (4.5) to be true, $\phi(a_0) + \sum_{i=1}^{k} \tau_i = \phi(b_0) + \sum_{i=1}^{h} \tau_{k+i}$ must hold. Otherwise, for example, if $\phi(a_0) + \sum_{i=1}^{k} \tau_i < \phi(b_0) + \sum_{i=1}^{h} \tau_{k+i}$, by definition of $\tau_1, \dots, \tau_{k+h}$ we would have

$$\begin{aligned} 0 &= \mathbb{P}\left(\phi(b_0) + \sum_{i=1}^{h} \tau_{k+i} + \sum_{i=1}^{h} (B'_i - \tau_{k+i}) = \phi(a_0) + \sum_{i=1}^{k} \tau_i\right) \\ &= \mathbb{P}\left(\phi(a_0) + \sum_{i=1}^{k} \tau_i + \sum_{i=1}^{k} (A'_i - \tau_i) = \phi(a_0) + \sum_{i=1}^{k} \tau_i\right) \geq \prod_{i=1}^{k} \mathbb{P}\left(A'_i = \tau_i\right) > 0, \end{aligned}$$

that is impossible. Therefore (4.5) is equivalent to

$$\mathbf{H}_0 \colon \sum_{i=1}^{k} (A'_i - \tau_i) \sim \sum_{i=1}^{h} (B'_i - \tau_{k+i}),$$

which, in turn, can be reduced to the form (4.3) by defining $X_i = A'_i - \tau_i$ for $i = 1, \dots, k$ and $Y_i = B'_i - \tau_{k+i}$ for $i = 1, \dots, h$ thus accounting for the subtraction of the constants in the distribution of X_i and Y_i. As a consequence, the support of X_i is $\{0, \dots, r_i\}$ for some positive integer $r_i \in \mathbb{N}$ and

$$\mathbb{P}(X_i = 0) > 0 \tag{4.6}$$

for every $i = 1, \dots, k$. The same applies to $Y_1, \dots, Y_h$. □

As a result of Lemma 4.1, we need only to consider $\mathbf{H}_0$ stated in equation (4.3). Thus given a sequence of $k \geq 2$ integer and independent random variables $X_1, \dots, X_k$ we write, for fixed $i \in \{1, \dots, k\}$, that $X_i \sim x_i \in \Delta^{r_i}$ with

$$\Delta^{r_i} = \left\{ v = (v_0, \dots, v_{r_i}) \in \mathbb{R}^{r_i+1} \colon v_0, v_{r_i} \in (0,1);\, v_j \geq 0, j = 1, \dots, r_i - 1; \sum_{j=1}^{r_i} v_j = 1 \right\}$$

to indicate that X_i takes values in $\{0, \dots, r_i\} \subseteq \mathbb{N}_0$, with $r_i > 0$, and is distributed with PMV $x_i = (x_{i0}, \dots, x_{ir_i})$, that is $\mathbb{P}(X_i = j) = x_{ij}$ for $j = 0, \dots, r_i$.

We remark that $x_{i0}, x_{ir_i} \in (0,1)$, for every $i = 1, \dots, k$, are assumed to avoid degenerate cases, without loss of generality. In fact, $x_{i0} > 0$ descends from (4.6). Moreover, given any $t \leq k$ such that $x_{tr_t} = \mathbb{P}(X_t = r_t) = 0$, there exists $\tau = \max\{j \colon x_{tj} > 0\} < r_t$, so that X_t can be replaced by $\tilde{X}_t \sim \tilde{x}_t = (\tilde{x}_{t0}, \dots, \tilde{x}_{t\tau}) \in \Delta^{\tau}$, with $\tilde{x}_{tj} = x_{tj}$ for $j = 0, \dots, \tau$. Lastly, the constraint $x_{i0}, x_{ir_i} < 1$ for every $i = 1, \dots, k$ ensures that X_i is not a constant value.

From now on, with these assumptions and notation, for the null hypothesis of goodness-of-fit test we consider

$$\mathbf{H}_0 \colon \sum_{i=1}^{k} X_i \sim z \tag{4.7}$$

with $s = \sum_{i=1}^{k} r_i$ and $z \in \Delta^s$. By independence, the sum of $X_1, \ldots, X_k$ is distributed as the discrete convolution, denoted $*$, of their PMVs, that is

$$\sum_{i=1}^{k} X_i \sim x_1 * \ldots * x_k,$$

where, for any two vectors $v = (v_0, \ldots, v_a) \in \mathbb{R}^{a+1}$, $w = (w_0, \ldots, w_b) \in \mathbb{R}^{b+1}$ with $a, b > 0$, we have $v * w \in \mathbb{R}^{a+b+1}$ and $(v * w)_i = \sum_{j=0}^{a} \sum_{l=0}^{b} v_j w_l \delta_{j+l,i}$ with $\delta_{i,j} = 1$ if $i = j$ and being null otherwise. In particular, the null hypothesis for the goodness-of-fit-test (4.7) is equivalent to

$$\mathbf{H}_0 \colon x_1 * \ldots * x_k = z.$$

Our first goal is to determine a statistic to test (4.7), which will subsequently be extended to assess the equality in distribution between $\sum_{i=1}^{k} X_i$ and $\sum_{i=1}^{h} Y_i$, where $Y_1 \sim y_1 \in \Delta^{r_{k+1}}, \ldots, Y_h \sim y_h \in \Delta^{r_{k+h}}$ are other $h \geq 1$ independent random variables, with $s = \sum_{i=1}^{k} r_i = \sum_{i=1}^{h} r_{k+i}$, namely

$$\mathbf{H}_0 \colon \sum_{i=1}^{k} X_i \sim \sum_{i=1}^{h} Y_i, \tag{4.8}$$

or, equivalently,

$$\mathbf{H}_0 \colon x_1 * \ldots * x_k = y_1 * \ldots * y_h.$$

When the PMVs x_i for $i = 1, \ldots, k$ and y_j for $j = 1, \ldots, h$ are unknown, care must be taken to define the test statistics for (4.7) and (4.8) based only on available information. In this regard, the data consist of the observation of n_i independent random variables $\{X_{i1}, \ldots, X_{in_i}\}$ identically distributed as X_i for $i = 1, \ldots, k$ and n_{k+i} independent random variables $\{Y_{i1}, \ldots, Y_{in_{k+i}}\}$ identically distributed as Y_i for $i = 1, \ldots, h$. Following a nonparametric approach, we fix $i \in \{1, \ldots, k\}$ and define $\hat{x}_{in_i} \in \Delta^{r_i}$ the MLE of x_i, that is

$$(\hat{x}_{in_i})_u = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{1}_{\{X_{ij}=u\}}$$

for $u = 0, \ldots, r_i$, with $\mathbb{1}_A$ being the indicator function of the event A. In particular, by the multivariate central limit theorem (Serfling, 1980), $\sqrt{n_i}(\hat{x}_{in_i} - x_i)$ is asymptotically distributed as a centred normal random variable with covariance $\Sigma(x_i)$ for n_i large,

namely $\sqrt{n_i}(\hat{x}_{in_i} - x_i) \sim_{n_i \to \infty} \mathcal{N}(\Sigma(x_i))$, where, for any PMV $v = (v_0, \ldots, v_a) \in \Delta^a$ and $a \geq 0$, we define $\Sigma(v) \in \mathbb{R}^{a+1} \times \mathbb{R}^{a+1}$ such that $(\Sigma(v))_{ij} = v_i \delta_{i,j} - v_i v_j$ for $i, j = 0, \ldots, a$. With this notation, we derive the MLE for the distribution of $\sum_{i=1}^k X_i$.

Proposition 4.2 (MLE for a sum of independent random variables)**.** *Given $k \geq 2$, let $\{X_{i1}, \ldots, X_{in_i}\}$ be $n_i \in \mathbb{N}$ random variables independent and identically distributed as $X_i \sim x_i \in \Delta^{r_i}$ with $r_i \in \mathbb{N}$, for $i = 1, \ldots, k$. Set $s = \sum_{i=1}^k r_i$. The MLE for the PMV $x_1 * \ldots * x_k \in \Delta^s$ of $\sum_{i=1}^k X_i$ is $\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k} \in \Delta^s$, defined as*

$$(\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k})_u = \left(\prod_{j=1}^k n_j \right)^{-1} \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} \mathbb{1}_{\{\sum_{j=1}^k X_{ji_j} = u\}},$$

for every $u = 0, \ldots, s$.

Proof. Noting $a_{ij} \in \{0, \ldots, r_i\}$ the independent sample from X_{ij} for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, the MLE of $x_1 * \ldots * x_k$ is the element $\theta \in \Delta^s$ that maximises

$$\begin{aligned} &\mathbb{P}(X_{11} = a_{11}, \ldots, X_{1n_1} = a_{1n_1}, \ldots, X_{k1} = a_{k1}, \ldots, X_{kn_k} = a_{kn_k} | \theta) \\ &\qquad = \prod_{i=1}^k \mathbb{P}(X_{i1} = a_{i1}, \ldots, X_{in_i} = a_{in_i} | \theta)\,. \end{aligned}$$

On the right hand side, for fixed $i \in \{1, \ldots, k\}$, $\mathbb{P}(X_{i1} = a_{i1}, \ldots, X_{in_i} = a_{in_i} | \theta)$ achieves maximum value for any $\theta = \theta_1 * \ldots * \theta_k$ such that $\theta_i = \hat{x}_{in_i}$, with $\theta_j \in \Delta^{r_j}$ for any j. In particular $\theta = \hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k}$ maximises all factors, hence the whole product. □

Of note, Proposition 4.2 shows that the MLE for $x_1 * \ldots * x_k$ is calculated using all observables X_{ij} for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$, which may not be the case for Person's statistic, as explained in Section 4.2, if at least one sample size of $n_1, \ldots, n_k$ differs from another.

We introduce additional notation for what follows: A^+, A', $\mathrm{Ker}(A)$, $\mathrm{nul}(A)$, $\mathrm{rk}(A)$ for the Moore-Penrose inverse, transpose, kernel, nullity and rank of a matrix A, respectively (Horn and Johnson, 1986; Hagen et al., 2000); $\chi^2(s)$ to indicate the χ^2 distribution with $s > 0$ degrees of freedom; $T^{b+1}(v) \in \mathbb{R}^{b+1} \times \mathbb{R}^{a+b+1}$ for the matrix of the discrete convolution between $v \in \mathbb{R}^{a+1}$ and any $b+1$-dimensional vector, i.e. $T^{b+1}(v)w = v * w \in \mathbb{R}^{a+b+1}$ with $w \in \mathbb{R}^{b+1}$, given $a, b \geq 0$. We write $T(v)$ without explicit domain dimension if this is stated or clear from the context.

We are now ready to determine the asymptotic behaviour of $\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k}$, which follows from an application of the delta method as well as properties of quadratic transformation of asymptotically multivariate normal vectors (Serfling, 1980). In order

for the MLE $\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k}$ to converge to $x_1 * \ldots * x_k$ it is necessary that the sample sizes $n_1, \ldots, n_k$ grow with proportional rates. For this reason, from now on, we set

$$m = \min(n_1, \ldots, n_{k+h}) \text{ and assume } c_i = \lim_{m\to\infty} \frac{m}{n_i}$$

is finite and positive for every $i = 1, \ldots, k+h$.

Proposition 4.3 (Asymptotic normality of convolutions)**.** *Under the null hypothesis* (4.7)*, that* $x_1 * \ldots * x_k = z$*, it holds that*

$$V_m = \sqrt{m}\left(\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k} - z\right) \underset{m\to\infty}{\sim} \mathcal{N}(\Psi) \tag{4.9}$$

and

$$V_m' \Psi^+ V_m \underset{m\to\infty}{\sim} \chi^2\left(\mathrm{rk}(\Psi)\right) \tag{4.10}$$

where $\Psi = \sum_{i=1}^{k} c_i T(x_{(i)}) \Sigma(x_i) T(x_{(i)})'$ *and* $x_{(i)} = x_1 * \ldots * x_{i-1} * x_{i+1} * \ldots * x_k$ *for* $i = 1, \ldots, k$*. Alternatively, under the null hypothesis* (4.8)*, that* $x_1 * \ldots * x_k = y_1 * \ldots * y_h$*, it holds that*

$$W_m = \sqrt{m}\left(\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k} - \hat{y}_{1n_{k+1}} * \ldots * \hat{y}_{hn_{k+h}}\right) \underset{m\to\infty}{\sim} \mathcal{N}(\Psi + \Xi) \tag{4.11}$$

and

$$W_m' (\Psi + \Xi)^+ W_m \underset{m\to\infty}{\sim} \chi^2\left(\mathrm{rk}(\Psi + \Xi)\right) \tag{4.12}$$

where $\Xi = \sum_{i=1}^{h} c_{k+i} T(y_{(i)}) \Sigma(y_i) T(y_{(i)})'$ *and* $y_{(i)} = y_1 * \ldots * y_{i-1} * y_{i+1} * \ldots * y_k$ *for* $i = 1, \ldots, h$*.*

We remark that expressions (4.10) and (4.12) require the knowledge of Ψ and $\Psi + \Xi$, but these may, in general, be unknown. Thus we take advantage of the generalised Wald's method (Moore, 1977), which shows how to construct χ^2 tests from consistent estimators of the covariance matrices such as Ψ and $\Psi + \Xi$. We recall here Moore (1977, Theorem 2) which will serve as backbone for the subsequent results.

Proposition 4.4 (Generalised Wald's method (Moore, 1977, Theorem 2))**.** *Suppose a sequence of estimators* $\{\hat{\theta}_m\}_{m\geq 1}$ *of a parameter* $\theta_0 \in \mathbb{R}^d$*, with* $d > 0$*, is such that*

$$\sqrt{m}\left(\hat{\theta}_m - \theta_0\right) \underset{m\to\infty}{\sim} \mathcal{N}(\Sigma)$$

with $\mathrm{rk}(\Sigma) \leq d$*. Noted* $\{B_m\}_{m\geq 1}$ *a sequence of d-dimensional square matrices such that* $B_m \sim_{m\to\infty} B$ *with* B *generalised-inverse of* Σ*, then*

$$m\left(\hat{\theta}_m - \theta_0\right)' B_m \left(\hat{\theta}_m - \theta_0\right) \underset{m\to\infty}{\sim} \chi^2(\mathrm{rk}(\Sigma)).$$

Since the entries of

$$\hat{\Psi}_m = \sum_{i=1}^{k} c_i T(\hat{x}_{(i)n_i}) \Sigma(\hat{x}_{in_i}) T(\hat{x}_{(i)n_i})'$$

are continuous functions of $x_1, \ldots, x_k$, whose consistent estimators are $\hat{x}_{1n_1}, \ldots, \hat{x}_{kn_k}$ respectively, then $\hat{\Psi}_m$ is a consistent estimator of Ψ and similarly

$$\hat{\Xi}_m = \sum_{i=1}^{h} c_{k+i} T(\hat{y}_{(i)n_{k+i}}) \Sigma(\hat{y}_{in_{k+i}}) T(\hat{y}_{(i)n_{k+i}})'$$

for Ξ.

Note that Proposition 4.4 cannot be directly applied to (4.9) by setting $B_m = \hat{\Psi}_m^+$, as $\hat{\Psi}_m^+$ may not be a consistent estimator of Ψ^+. Given a sequence of consistent estimators $\{A_m\}_{m\geq 1}$ for a matrix A of finite dimensions, then $\{A_m^+\}_{m\geq 1}$ is a sequence of consistent estimators for A^+ if and only if $\mathrm{rk}(A_m) = \mathrm{rk}(A)$ for m large (Nashed, 1976). In particular, as the rank is a lower-semicontinuous operator on the space of finite dimensional matrices, then only $\mathrm{rk}(\hat{\Psi}_m) \geq \mathrm{rk}(\Psi)$ is guaranteed as m tends to infinity.

If the limiting rank is known, consistency is ensured by applying the Eckart-Young-Mirsky theorem (Eckart and Young, 1936), which is the solution to the basic low rank approximation of a finite dimensional matrix (Markovsky, 2012). To this end, given any d-dimensional symmetric matrix $A \in \mathbb{R}^d \times \mathbb{R}^d$ with $d \in \mathbb{N}$ and its eigen-decomposition $A = P'\Lambda P$, with $0 < \mathrm{rk}(A) \leq d$, Λ diagonal matrix of the decreasing eigenvalues and P orthogonal matrix, then for any $0 < r \leq \mathrm{rk}(A)$ we define a rank-r matrix that approximates A (in light of the Eckart-Young-Mirsky theorem) as $A^r = (D^r P)'\Lambda^r D^r P \in \mathbb{R}^d \times \mathbb{R}^d$, where D^r is a $\mathbb{R}^r \times \mathbb{R}^d$ matrix with 1 at the diagonal and 0 elsewhere and Λ^r is the $\mathbb{R}^r \times \mathbb{R}^r$ diagonal matrix of the largest r eigenvalues of Σ. In particular, A^r may not be unique, as for the case when the r^{th} and $r+1^{\text{th}}$ eigenvalues are equal. The following result is found in Hadi and Wells (1990, Theorem 2.3) for the generalised inverses and we report it here for the case of Moore-Penrose inverses.

Proposition 4.5 (Rank approximation (Hadi and Wells, 1990, Theorem 2.3))**.** *Suppose a sequence of centred random variables $\{U_m\}_{m\geq 1} \in \mathbb{R}^d$ is asymptotically distributed as $\mathcal{N}(\Sigma)$ for m large, with $0 < \mathrm{rk}(\Sigma) \leq d$ where $d > 0$. Let $\{\hat{\Sigma}_m\}_{m\geq 1}$ be a sequence of square matrices that are consistent estimators of Σ, then for every $0 < r \leq \mathrm{rk}(\Sigma)$*

$$U_m'(\hat{\Sigma}_m^r)^+ U_m \underset{m\to\infty}{\sim} \chi^2(r),$$

where $\hat{\Sigma}_m^r$ is a rank-r approximation of $\hat{\Sigma}_m$.

Proposition 4.5 highlights the central role of the rank of Σ, which will be derived in the next section for $\Sigma = \Psi$ and $\Sigma = \Psi + \Xi$. In general, the determination of $\mathrm{rk}(\Sigma)$ may be a difficult problem that depends on the structure of the Σ under consideration, and this limitation may explain why an otherwise flexible tool such as the generalised Wald's method from Proposition 4.4 is not more widely employed. But, if the rank is known, Proposition 4.5 provides a method for statistical testing the null hypotheses, such as $\mathbf{H}_0\colon\ x_1 * \ldots * x_k = z$ and $\mathbf{H}_0\colon\ x_1 * \ldots * x_k = y_1 * \ldots * y_h$, under which Σ is not invertible. This result also assures a solution if only a lower bound of the rank is given, at the cost of statistical power. Furthermore, the exclusion of smaller eigenvalues may still be necessary to achieve numerical stability when calculating the pseudo-inverse of $\hat{\Sigma}_m$, as due to Proposition 4.5, the effect of that truncation can be accounted for in the statistic formulation.

4.4 Determining the covariance matrix rank

In this section, we investigate the rank of Ψ and $\Psi + \Xi$, the covariance matrices from (4.9) and (4.11) of Proposition 4.3, in order to derive the number of degrees of freedom from the limiting statistics for the goodness-of-fit (4.10) and equality in distribution (4.12) tests. Focusing on Ψ, we begin by showing that $c_i T(x_{(i)})\Sigma(x_i)T(x_{(i)})'$ is a positive semidefinite matrix for any fixed $i \in \{1, \ldots, k\}$. In fact, since $\Sigma(x_i)$ is positive semidefinite, for every $v \in \mathbb{R}^{s+1}$

$$c_i v' T(x_{(i)})\Sigma(x_i)T(x_{(i)})'v = c_i w'\Sigma(x_i)w \geq 0, \tag{4.13}$$

where $w = T(x_{(i)})'v$. Additionally, we deduce from Horn and Johnson (1986, Observation 7.1.3) that given A and B two positive semidefinite matrices with the same dimensions, then

$$\mathrm{Ker}(A + B) = \mathrm{Ker}(A) \cap \mathrm{Ker}(B)\,. \tag{4.14}$$

Taken together, (4.13) and (4.14) imply

$$\mathrm{Ker}(\Psi) = \bigcap_{i=1}^{k} \mathrm{Ker}\big(T(x_{(i)})\Sigma(x_i)T(x_{(i)})'\big)\,. \tag{4.15}$$

Using kernel properties, we write

$$\begin{aligned}\mathrm{Ker}\big(T(x_{(i)})\Sigma(x_i)T(x_{(i)})'\big) &= \mathrm{Ker}\big(\Sigma(x_i)T(x_{(i)})'\big) \\ &= \mathrm{Ker}\big(T(x_{(i)})'\big) \oplus \{v \in \mathbb{R}^{s+1}\colon T(x_{(i)})'v \in \mathrm{Ker}(\Sigma(x_i))\}\end{aligned} \tag{4.16}$$

where the first relation descends from $\Sigma(x_i)$ and $T(x_{(i)})\Sigma(x_i)T(x_{(i)})'$ being positive semidefinite matrices and $\oplus$ represents the direct sum operation.

Let $i \in \{1,\dots,k\}$ be fixed and let $L_i = \{l\colon x_{il} = 0\}$ be the set of indexes of the null entries of $x_i \in \Delta^{r_i}$. In general, the kernel of $\Sigma(x_i)$ is generated by the r_i+1-dimensional all-ones vector 1_{r_i} and the canonical vectors $e_l^i = (e_{l0}^i,\dots,e_{lr_i}^i) \in \mathbb{R}^{r_i+1}$ for every $l \in L_i$, where $e_{lu}^i = \delta_{l,u}$ for $u = 0,\dots,r_i$, that is $\mathrm{Ker}(\Sigma(x_i)) = \langle 1_{r_i}\rangle \oplus E_i$, where $E_i = \langle\{e_l^i\colon l \in L_i\}\rangle$.

Denoting $r_{(i)} = \sum_{j\neq i}^k r_j$, we can expand $T(x_{(i)})$ into

$$T(x_{(i)}) = \begin{bmatrix} x_{(i)0} & 0 & \dots & 0 \\ x_{(i)1} & x_{(i)0} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ x_{(i)r_{(i)}} & x_{(i)r_{(i)}-1} & \ddots & x_{(i)0} \\ 0 & x_{(i)r_{(i)}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & x_{(i)r_{(i)}} \end{bmatrix},$$

from which we deduce $T(x_{(i)})'1_s = 1_{r_i}$, for every $x_{(i)} \in \Delta^{r_{(i)}}$, and in particular $1_s \notin \mathrm{Ker}\big(T(x_{(i)})'\big)$.

To achieve an explicit formulation for the rank of Ψ (and analogously for $\Psi+\Xi$), in Theorem 4.7 we will assume that $x_i \in \Delta_{\mathrm{Int}}^{r_i} \subset \Delta^{r_i}$, defined as

$$\Delta_{\mathrm{Int}}^{r_i} = \{v = (v_0,\dots,v_{r_i}) \in \mathbb{R}^{r_i+1}\colon \sum_{j=1}^{r_i} v_j = 1; 0 < v_j < 1, j = 0,\dots,r_i\}$$

for every $i = 1,\dots,k$. This ensures that $E_i = \emptyset$, so that $\mathrm{Ker}(\Sigma(x_i)) = \langle 1_{r_i}\rangle$. Under this hypothesis, from (4.15) and (4.16) we deduce that

$$\mathrm{Ker}(\Psi) = \bigcap_{i=1}^k \mathrm{Ker}\big(T(x_{(i)})\Sigma(x_i)T(x_{(i)})'\big) = \langle 1_s\rangle \oplus \bigcap_{i=1}^k \mathrm{Ker}\big(T(x_{(i)})'\big)\,, \tag{4.17}$$

and, with the same reasoning applied to $\Psi+\Xi$, it follows that

$$\mathrm{Ker}(\Psi+\Xi) = \langle 1_s\rangle \oplus \Big(\big(\bigcap_{i=1}^k \mathrm{Ker}\big(T(x_{(i)})'\big)\big) \cap \big(\bigcap_{j=1}^h \mathrm{Ker}\big(T(y_{(j)})'\big)\big) \Big). \tag{4.18}$$

In the following Lemma we show how

$$\bigcap_{i=1}^{k} \operatorname{Ker}\big(T(x_{(i)})'\big) \tag{4.19}$$

and

$$\left(\bigcap_{i=1}^{k} \operatorname{Ker}\big(T(x_{(i)})'\big)\right) \cap \left(\bigcap_{j=1}^{h} \operatorname{Ker}\big(T(y_{(j)})'\big)\right) \tag{4.20}$$

depend on the roots in common between the probability generating functions of the random variables $X_1, \ldots, X_k, Y_1, \ldots, Y_h$. This result will be achieved in full generality without restrictions for the PMVs, that is with $x_1 \in \Delta^{r_1}, \ldots, x_k \in \Delta^{r_k}, y_1 \in \Delta^{r_{k+1}}, \ldots, y_h \in \Delta^{r_{k+h}}$. We first provide some insight into this connection by considering the case $k = 2$

$$\operatorname{Ker}\big(T(x_{(1)})'\big) \cap \operatorname{Ker}\big(T(x_{(2)})'\big) = \operatorname{Ker}\left(\begin{bmatrix} T(x_2) & T(x_1) \end{bmatrix}'\right).$$

In fact,

$$\begin{bmatrix} T(x_2) & T(x_1) \end{bmatrix} = \begin{bmatrix} x_{20} & 0 & \ldots & 0 & x_{10} & 0 & \ldots & 0 \\ x_{21} & x_{20} & \ddots & \vdots & x_{11} & x_{10} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & 0 \\ x_{2r_2} & x_{2r_2-1} & \ddots & x_{20} & x_{1r_1} & x_{1r_1-1} & \ddots & x_{10} \\ 0 & x_{2r_2} & \ddots & \vdots & 0 & x_{1r_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & x_{2r_2} & 0 & \ldots & 0 & x_{1r_1} \end{bmatrix}$$

belongs to $\mathbb{R}^{r_1+r_2+2} \times \mathbb{R}^{r_1+r_2+1}$ and has the same structure, with different dimensions, of a Sylvester matrix (Markovsky, 2012), whose nullity is the degree of the polynomial from the greatest common divisor of the probability generating functions associated to x_1 and x_2.

To formalise the connection with PMVs and polynomials, we introduce the bijection $\varphi\colon \cup_{a\geq 0} \{u = (u_0, \ldots, u_a) \in \mathbb{R}^{a+1}\colon \sum_{i=0}^{a} u_i = 1; u_a \neq 0\} \to \{u(t) \in \mathbb{R}[t]\colon u(1) = 1\}$ that, for any $a \geq 0$, maps a vector $v = (v_0, \ldots, v_a) \in \{u = (u_0, \ldots, u_a) \in \mathbb{R}^{a+1}\colon \sum_{i=0}^{a} u_i = 1; u_a \neq 0\}$ to the polynomial $\varphi(v)(t) = \sum_{i=0}^{a} v_i t^i \in \mathbb{R}[t]$ of degree $\deg \varphi(v) = a$ with coefficients v. In particular, this map transforms the convolution of vectors into the product of polynomials: given $w \in \{u = (u_0, \ldots, u_b) \in \mathbb{R}^{b+1}\colon \sum_{i=0}^{b} u_i = 1; u_b \neq 0\}$,

for any $b \geq 0$, we have

$$\varphi(v * w)(t) = \sum_{i=0}^{a} (v * w)_i t^i = \varphi(v)(t)\varphi(w)(t).$$

The map φ allows the extension of the notion of greatest common divisor between any two polynomials $\gcd(\varphi(v),\ \varphi(w))$ to their related vectors v, w. This is achieved by establishing $\gcd(u_1(t), u_2(t)) \in \{u(t) \in \mathbb{R}[t] \colon u(1) = 1\}$ for any $u_1(t), u_2(t) \in \{u(t) \in \mathbb{R}[t] \colon u(1) = 1\}$, so that the greatest common divisor is uniquely defined, and by setting, for any $v, w \in \cup_{a \geq 0}\{u = (u_0, \ldots, u_a) \in \mathbb{R}^{a+1} \colon \sum_{i=0}^{a} u_i = 1; u_a \neq 0\}$,

$$\begin{aligned} \gcd(v, w) &= \varphi^{-1}(\gcd(\varphi(v)(t), \varphi(w)(t))) \\ &\in \{u = (u_0, \ldots, u_{r_g}) \in \mathbb{R}^{r_g+1} \colon \sum_{i=0}^{r_g} u_i = 1; u_{r_g} \neq 0\} \end{aligned}$$

with $r_g = \deg \gcd(\varphi(v)(t), \varphi(w)(t)) \geq 0$. In particular, we say the vectors v, w are coprime if and only if $r_g = 0$. Following the same logic, we import the concept of least common multiple between v and w, denoted $\operatorname{lcm}(v, w)$, and the property of divisibility between vectors. Of note, with the notation above, the probability generating functions of $x_1, \ldots, x_k, y_1, \ldots, y_h$ are, respectively, $\varphi(x_1), \ldots, \varphi(x_k), \varphi(y_1), \ldots, \varphi(y_h)$.

We now establish the relation between the kernels of (4.19), (4.20) and the greatest common divisors $g_k = \gcd(x_{(1)}, \ldots, x_{(k)})$ and $\bar{g}_h = \gcd(y_{(1)}, \ldots, y_{(h)})$.

Lemma 4.6 (Kernels from gcd of PMVs)**.** *Let $k \geq 2$ and $x_1 \in \Delta^{r_1}, \ldots, x_k \in \Delta^{r_k}$. Given $g_k = \gcd(x_{(1)}, \ldots, x_{(k)}) \in \mathbb{R}^{r_{g_k}+1}$, it holds that $T(g_k)' \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{g_k} + 1} \times \mathbb{R}^{\sum_{i=1}^{k} r_i + 1}$ and*

$$\bigcap_{i=1}^{k} \operatorname{Ker}\big(T(x_{(i)})'\big) = \operatorname{Ker}\big(T(g_k)'\big)\,. \tag{4.21}$$

Additionally, let $h \geq 1$ and $y_1 \in \Delta^{r_{k+1}}, \ldots, y_h \in \Delta^{r_{k+h}}$. Given $\bar{g}_h = \gcd(y_{(1)}, \ldots, y_{(h)}) \in \mathbb{R}^{r_{\bar{g}_h}+1}$ and $\tilde{g} = \gcd(g_k, \bar{g}_h) \in \mathbb{R}^{r_{\tilde{g}}+1}$, it holds that $T(\bar{g}_h)' \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{\bar{g}_h} + 1} \times \mathbb{R}^{\sum_{i=1}^{k} r_i + 1}$, $T(\tilde{g})' \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{\tilde{g}} + 1} \times \mathbb{R}^{\sum_{i=1}^{k} r_i + 1}$ and

$$\bigcap_{i=1}^{k} \operatorname{Ker}\big(T(x_{(i)})'\big) \cap \bigcap_{j=1}^{h} \operatorname{Ker}\big(T(y_{(j)})'\big) = \operatorname{Ker}\left(\begin{bmatrix} T(g_k)' \\ T(\bar{g}_h)' \end{bmatrix}\right) = \operatorname{Ker}\big(T(\tilde{g})'\big) \tag{4.22}$$

Proof. We first prove (4.21) by induction on k. For $k = 2$ we have to show that

$$\operatorname{Ker}\big(T(x_{(1)})'\big) \cap \operatorname{Ker}\big(T(x_{(2)})'\big) = \operatorname{Ker}\left(\begin{bmatrix} T(x_{(1)})' \\ T(x_{(2)})' \end{bmatrix}\right) = \operatorname{Ker}\big(T(g_2)'\big)\,,$$

where $g_2 = \gcd(x_{(1)}, x_{(2)}) = \gcd(x_2, x_1) \in \mathbb{R}^{r_{g_2}+1}$ and $r_{g_2} \geq 0$. By definition of g_2, there exist two coprime vectors $z_1 \in \mathbb{R}^{r_1 - r_{g_2}+1}$, $z_2 \in \mathbb{R}^{r_2 - r_{g_2}+1}$ such that $x_{(1)} = z_1 * g_2$ and $x_{(2)} = z_2 * g_2$ so that, by composition of discrete convolution, we can write

$$\begin{bmatrix} T(x_{(1)})' \\ T(x_{(2)})' \end{bmatrix} = \begin{bmatrix} T(z_1)' \\ T(z_2)' \end{bmatrix} T(g_2)',$$

where $T(g_2)' \in \mathbb{R}^{r_1+r_2-r_{g_2}+1} \times \mathbb{R}^{r_1+r_2+1}$, $T(z_1)' \in \mathbb{R}^{r_2+1} \times \mathbb{R}^{r_1+r_2-r_{g_2}+1}$ and $T(z_2)' \in \mathbb{R}^{r_1+1} \times \mathbb{R}^{r_1+r_2-r_{g_2}+1}$. As the number of columns of $\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix}'$ is lesser than the number of rows, since $r_1 + r_2 - r_{g_2} + 1 \leq r_1 + r_2 + 2 \Leftrightarrow r_{g_2} + 1 \geq 0$, to prove (4.4), it suffices to show that $\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix}'$ is of full rank $r_1 + r_2 - r_{g_2} + 1$. Thus, from rank-nullity theorem,

$$\begin{aligned} r_1 + r_2 - r_{g_2} + 1 = \mathrm{rk}\left(\begin{bmatrix} T(z_1)' \\ T(z_2)' \end{bmatrix} \right) &= \mathrm{rk}\Big(\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix} \Big) \\ &= r_1 + r_2 + 2 - \mathrm{nul}\Big(\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix} \Big) \end{aligned}$$

holds if and only if $\mathrm{nul}\Big(\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix} \Big) = r_{g_2} + 1$. The latter is true by coprimeness between z_1 and z_2 and by matrix dimensionality, since the solutions (a, b) with $a \in \mathbb{R}^{r_1+1}$, $b \in \mathbb{R}^{r_2+1}$, to the homogeneous system of equations $T(z_1)a + T(z_2)b = z_1 * a + z_2 * b = 0$, are characterised by $a = z_2 * u$, $b = -z_1 * u$ with $u \in \mathbb{R}^{r_{g_2}+1}$ vector of free parameters. Assuming (4.21) for k, we now prove the case $k+1$ to conclude. By inductive step and associative property of convolution, we can write

$$\begin{aligned} \bigcap_{i=1}^{k+1} \mathrm{Ker}\big(T(x_{(i)})'\big) &= \mathrm{Ker}\left(\begin{bmatrix} T(x_{(1)})' \\ \vdots \\ T(x_{(k+1)})' \end{bmatrix} \right) \\ &= \mathrm{Ker}\left(\begin{bmatrix} T(g_k)'T(x_{k+1})' \\ T(x_{(k+1)})' \end{bmatrix} \right) = \mathrm{Ker}\left(\begin{bmatrix} T(g_k * x_{k+1})' \\ T(x_{(k+1)})' \end{bmatrix} \right). \end{aligned} \tag{4.23}$$

Thus, by the definition of $g_{k+1} = \gcd(x_{(1)}, \ldots, x_{(k+1)}) \in \mathbb{R}^{r_{g_{k+1}}+1}$, the properties of gcd lead to $g_{k+1} = \gcd(g_k * x_{k+1}, x_1 * \ldots * x_k) = g_k \gcd(x_{k+1}, u_{k+1})$, where $u_{k+1} \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{g_k}+1}$ is such that $u_{k+1} * g_{k+1} = x_1 * \ldots * x_k$. In particular, we deduce

$$r_{g_{k+1}} \geq r_{g_k}. \tag{4.24}$$

As in the previous step, we introduce $z_1 \in \mathbb{R}^{r_{g_k} + r_{k+1} - r_{g_{k+1}}+1}$ and $z_2 \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{g_{k+1}}+1}$ coprime vectors such that $z_1 * g_{k+1} = g_k * x_{k+1}$ and $z_2 * g_{k+1} = x_1 * \ldots * x_k$. Thus,

resuming (4.23),

$$\bigcap_{i=1}^{k+1} \operatorname{Ker}\big(T(x_{(i)})'\big) = \operatorname{Ker}\left(\begin{bmatrix} T(z_1)' \\ T(z_2)' \end{bmatrix} T(g_{k+1})'\right) = \operatorname{Ker}\big(T(g_{k+1})'\big)$$

where last equality is analogous as for the case $k = 2$, given that the number of columns of $\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix}' \in \mathbb{R}^{\sum_{i=1}^{k+1} r_i - r_{g_k} + 2} \times \mathbb{R}^{\sum_{i=1}^{k+1} r_i - r_{g_{k+1}} + 1}$ is lesser than the number of rows, that is

$$\sum_{i=1}^{k+1} r_i - r_{g_{k+1}} + 1 \leq \sum_{i=1}^{k+1} r_i - r_{g_k} + 2 \quad \Leftrightarrow \quad r_{g_{k+1}} + 1 \geq r_{g_k},$$

which holds true by (4.24). To prove (4.22), we note that the first equation therein is established by (4.21), thus only the second equivalence shall be proved. Once more, we define z_1 and z_2 such that such that $\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix}' T(\tilde{g})' = \begin{bmatrix} T(g_k) & T(\bar{g}_h) \end{bmatrix}'$ with $T(z_1)' \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{g_k} + 1} \times \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{\tilde{g}} + 1}, T(z_2)' \in \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{\bar{g}_h} + 1} \times \mathbb{R}^{\sum_{i=1}^{k} r_i - r_{\tilde{g}} + 1}$. Again, the conclusion is verified by checking that, in the matrix $\begin{bmatrix} T(z_1) & T(z_2) \end{bmatrix}'$, there are less rows than columns, namely

$$\sum_{i=1}^{k} r_i + r_{\tilde{g}} + 1 \geq r_{g_k} + r_{\bar{g}_h},$$

which holds true since $g_k * \bar{g}_h = \gcd(g_k, \bar{g}_h) * \operatorname{lcm}(g_k, \bar{g}_h) = \tilde{g} * \operatorname{lcm}(g_k, \bar{g}_h)$ and $\operatorname{lcm}(g_k, \bar{g}_h)$ is a divisor of $z = x_1 * \ldots * x_k = y_1 * \ldots * y_h$ (as polynomials), so $\deg \operatorname{lcm}(g_k, \bar{g}_h) \leq \sum_{i=1}^{k} r_i$. □

As consequence of Lemma 4.6, we can finally determine the rank of the covariance matrices Ψ and $\Psi + \Xi$ assuming $x_1 \in \Delta_{\text{Int}}^{r_1}, \ldots, x_k \in \Delta_{\text{Int}}^{r_k}, y_1 \in \Delta_{\text{Int}}^{r_{k+1}}, \ldots, y_h \in \Delta_{\text{Int}}^{r_{k+h}}$, in order to calculate the number of degrees of freedom of the limiting χ^2 distribution in (4.10) and (4.12) from Proposition 4.3 for this case.

Theorem 4.7 (Covariance matrix rank)**.** *Under the assumptions of Proposition 4.3 and the notations of Lemma 4.6 and given $x_1 \in \Delta_{\text{Int}}^{r_1}, \ldots, x_k \in \Delta_{\text{Int}}^{r_k}, y_1 \in \Delta_{\text{Int}}^{r_{k+1}}, \ldots, y_h \in \Delta_{\text{Int}}^{r_{k+h}}$, it follows that*

$$\operatorname{Ker}(\Psi) = \langle 1_s \rangle \oplus \operatorname{Ker}\big(T(g_k)'\big) \tag{4.25}$$

and

$$\operatorname{Ker}(\Psi + \Xi) = \langle 1_s \rangle \oplus \operatorname{Ker}\big(T(\tilde{g})'\big)\,. \tag{4.26}$$

In particular, with s defined immediately prior to equation (4.8)*,*

$$\operatorname{rk}(\Psi) = s - r_{g_k} \tag{4.27}$$

and

$$\mathrm{rk}(\Psi + \Xi) = s - r_{\tilde{g}}. \tag{4.28}$$

Proof. The relations (4.25) and (4.26) derive as applications of Lemma 4.6 to (4.17) and (4.18), respectively. Lastly, (4.27) and (4.28) follow from (4.25) and (4.26), respectively, through rank-nullity properties of linear transformations from a finite-dimensional domain. □

In the case where $x_1, \ldots, x_k, y_1, \ldots, y_h$ are coprime, that is when their probability generating functions $\varphi(x_1), \ldots, \varphi(x_k), \varphi(y_1), \ldots, \varphi(y_h)$ have no root in common, we can simplify Theorem 4.7 as follows.

Corollary 4.8 (Rank from the coprime case)**.** *Under the assumptions of Theorem 4.7, if $x_1, \ldots, x_k, y_1, \ldots, y_h$ are coprime vectors, then $\mathrm{rk}(\Psi) = \mathrm{rk}(\Psi + \Xi) = s$, where s is defined immediately prior to equation* (4.8).

Proof. This follows from Theorem 4.7, as $\mathrm{Ker}(T(g_k)') = \mathrm{Ker}(T(\tilde{g})') = \{(0, \ldots, 0)\} \subseteq \mathbb{R}^{s+1}$ by coprimeness. □

Taken together, Lemma 4.6 and Theorem 4.7 serve as example of how to determine the covariance matrix rank upon application of the generalised Wald's framework. Moreover, Theorem 4.7 and Corollary 4.8 offer a way to study estimators such as $f(\hat{x}_{1n_1} * \ldots * \hat{x}_{kn_k})$ of $f(x_1 * \ldots * x_k)$ through the application of the delta method (Serfling, 1980), for all $f \colon \mathbb{R}^{s+1} \to \mathbb{R}^m$, with $m \in \mathbb{N}$, such that the differential of f in $x_1 * \ldots * x_k$ is not null.

4.5 Power comparison

We evaluate the performances of the convolution test in terms of type I error and power (1 minus type II error), which are the proportion of rejections with significance level $\alpha = 0.05$ under the null and alternative hypothesis respectively, setting Pearson's χ^2 test as the benchmark. To do so, we simulate the smallest parametrised model that enables the investigation of how samples size, degrees of freedom reduction, and observables distribution affect the convolution test, using different parameters choices. It also allows the transition from the null to alternative hypotheses by modulating a single parameter.

We consider $k = 2$, $h = 1$ and X_1, X_2 are two Bernoulli random variables with parameters $p, q \in (0, 1)$ so that $x_1 = (1 - p, p)$, $x_2 = (1 - q, q)$ and $x_1, x_2 \in \Delta^1_{\mathrm{Int}}$. With

$a = pq + \sqrt{pq(1-p)(1-q)}$, we define for $\rho \in [0,1]$

$$\begin{aligned} z(\rho) &= (1-\rho)x_1 * x_2 + \rho(1-a, 0, a) \\ &= \big((1-\rho)(1-p)(1-q)+\rho(1-a), (1-\rho)(p+q-2pq), (1-\rho)pq+\rho a\big), \end{aligned} \tag{4.29}$$

where a is defined so that $z(\rho) = (z(\rho)_0, z(\rho)_1, z(\rho)_2) \in \Delta^2_{\text{Int}}$ is the PMV for the distribution of $Z_1 + Z_2$, where Z_1 and Z_2 are two Bernoulli random variables with parameter p and q, respectively, and ρ is their correlation. The null hypothesis for the goodness-of-fit (GF) test is

$$\mathbf{H}_0 : X_1 + X_2 \sim z(0)$$

and we set a family of alternative hypotheses parametrized over $\rho \in (0,1]$ as

$$\mathbf{H}_1^\rho : X_1 + X_2 \nsim z(\rho).$$

Similarly, the null and the alternative hypotheses of the test for equality in distribution (ED) are defined as

$$\begin{aligned} &\mathbf{H}_0\colon X_1 + X_2 \sim Y_1 \quad \text{with} \quad Y_1 \sim z(0) \quad \text{and} \\ &\mathbf{H}_1^\rho\colon X_1 + X_2 \nsim Y_1 \quad \text{with} \quad Y_1 \sim z(\rho),\ \rho \in (0,1]. \end{aligned}$$

To facilitate the following discussion, we set the sample sizes n_1, n_2 and n_3 for X_1, X_2 and Y_1, respectively, so that $n_1, n_2 \leq n_3$ and $m = \min(n_1, n_2, n_3) = \min(n_1, n_2)$. Since we are interested in the comparison between the convolution and Pearson's χ^2 statistics, we need to calculate the latter even in the case of unequal sample size, i.e. when $n_1 \neq n_2$. Thus, we define

$$\begin{aligned} P_m^{\text{GF}} &= \sum_{j=0}^{2} \frac{\big(\sum_{i=1}^{m} \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} - m z(\rho)_j\big)^2}{m z(\rho)_j} \quad \text{and} \\ P_m^{\text{ED}} &= \sum_{j=0}^{2} \left(\frac{\big(\frac{n_3}{m+n_3}\sum_{i=1}^{m} \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} - \frac{m}{m+n_3}\sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}}\big)^2}{\frac{m}{m+n_3}\big(\sum_{i=1}^{m} \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} + \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}}\big)} \right. \\ &\qquad \left. + \frac{\big(\frac{m}{m+n_3}\sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}} - \frac{n_3}{m+n_3}\sum_{i=1}^{m} \mathbb{1}_{\{X_{1i}+X_{2i}=j\}}\big)^2}{\frac{n_3}{m+n_3}\big(\sum_{i=1}^{m} \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} + \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}}\big)} \right) \end{aligned}$$

for Pearson's goodness-of-fit and equality in distribution testing statistic, respectively. In particular, $n_1 + n_2 - 2m$ observations will not be used in the computation of P_m^{GF} and P_m^{ED}.

We define the convolution statistic with fixed rank $r = 1, 2$ from the notation in Proposition 4.3 and Proposition 4.5 as $V'_m(\hat{\Psi}^r_m)^+ V_m$ and $W'_m((\hat{\Psi}_m + \hat{\Xi}_m)^r)^+ W_m$. In the case where the n_1 and n_2 observations from the random variables X_1 and X_2, respectively,

are all equal, the sample covariance matrix is null, i.e. $\hat{\Psi}_m = 0$, and $(\hat{\Psi}_m^r)^+$ is not well defined. In this scenario, Pearson's P_m^{GF} can still be calculated. As we aim to compare the power gain over Pearson's procedures, we calculate the convolution statistics for goodness-of-fit test as $V_m'(\hat{\Psi}_m^r)^+V_m$, where well defined, otherwise we set it to P_m^{GF}. With the same reasoning for the equality in distribution case, for the following simulations we define the convolution statistic as

$$\begin{aligned} C_{rm}^{\mathrm{GF}} &= V_m'(\hat{\Psi}_m^r)^+V_m(1 - \mathbb{1}_{\{\hat{\Psi}_m=0\}}) + P_m^{\mathrm{GF}}\mathbb{1}_{\{\hat{\Psi}_m=0\}} \quad \text{and} \\ C_{rm}^{\mathrm{ED}} &= W_m'((\hat{\Psi}_m + \hat{\Xi}_m)^r)^+W_m(1 - \mathbb{1}_{\{\hat{\Psi}_m+\hat{\Xi}_m=0\}}) + P_m^{\mathrm{ED}}\mathbb{1}_{\{\hat{\Psi}_m+\hat{\Xi}_m=0\}}, \end{aligned}$$

for goodness-of-fit and equality in distribution tests, respectively. Note that $\lim_{m\to\infty} C_{rm}^{\mathrm{GF}} \sim \chi^2(r)$, since $\hat{\Psi}_m = 0$ if and only if $\hat{x}_{1n_1}, \hat{x}_{2n_2} \in \{(1,0),(0,1)\}$, but, for $j = 1, 2$, $\lim_{m\to\infty} \mathbb{P}\left(\hat{x}_{jn_j} \in \{(1,0),(0,1)\}\right) = 0$. From analogous reasoning, also $\lim_{m\to\infty} C_{rm}^{\mathrm{ED}} \sim \chi^2(r)$ holds true.

Moreover, in order not to confound the comparative analysis, we do not reduce the limiting χ^2 degrees of freedom in the case where a positive eigenvalue of $\hat{\Psi}_m$, or $\hat{\Psi}_m+\hat{\Xi}_m$, is set to 0 for being smaller than $10^{-\epsilon}$ (here $\epsilon = 15$ is the machine precision from Python's floating point number in Numpy 1.13.1).

Finally, to assess whether deviations from the limit of the convolution statistic are due to the estimate of the covariance matrix pseudo-inverse, we introduce

$$Z_{rm}^{\mathrm{GF}} = V_m'\Psi^+V_m \quad \text{and} \quad Z_{rm}^{\mathrm{ED}} = W_m'(\Psi + \Xi)^+W_m$$

for $r = 1, 2$, which are the convolution statistics calculated with the true covariance matrices.

For all these statistics, we evaluate the proportion of hypothesis rejection for the significance level $\alpha = 0.05$ by Monte Carlo approximation over $L = 100{,}000$ independent instances of the data. That is, given $S_r(t) = \mathbb{P}\left(\chi^2(r) \geq t\right)$ the survival function for a χ^2 distribution with r degrees of freedom, the proportions of rejections for P_m^{GF}, C_{rm}^{GF} and Z_{rm}^{GF}, respectively, are

$$\frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_2(P_{ml}^{\mathrm{GF}})<\alpha\}}, \quad \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_r(C_{rml}^{\mathrm{GF}})<\alpha\}}, \quad \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_r(Z_{rml}^{\mathrm{GF}})<\alpha\}}, \tag{4.30}$$

where $\{P_{m1}^{\mathrm{GF}}, \ldots, P_{mL}^{\mathrm{GF}}\}$, $\{C_{rm1}^{\mathrm{GF}}, \ldots, C_{rmL}^{\mathrm{GF}}\}$ and $\{Z_{rm1}^{\mathrm{GF}}, \ldots, Z_{rmL}^{\mathrm{GF}}\}$ are L independent simulation from the statistics for the goodness-of-fit testing P_m^{GF}, C_r^{GF}, Z_{rm}^{GF}, respectively. Similarly, we set the proportion of rejections for their equality in distribution

counterparts as

$$\frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_2(P_{ml}^{\mathrm{ED}})<\alpha\}},\quad \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_r(C_{rml}^{\mathrm{ED}})<\alpha\}},\quad \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{\{S_r(Z_{rml}^{\mathrm{ED}})<\alpha\}}, \tag{4.31}$$

for P_m^{ED}, C_{rm}^{ED} and Z_{rm}^{ED}, respectively.

In Fig. 4.1, we report the statistical power under $\mathbf{H}_0$ and a range of alternative hypotheses $\mathbf{H}_1^{\rho}$. We implement these comparisons for small and large samples with respect to $m = \min(n_1, n_2, n_3)$, and with equal and unequal sizes.

To comply with the rule-of-thumb recommendation for Pearson's χ^2 statistic application (Cressie and Read, 1984), the requirement for the expected frequencies $m(x_1 * x_2)_u \geq 1$ for the categories $u = 0, 1, 2$, must be met when m observations are sampled from the distribution of $X_1 + X_2$. Thus, we select two cases for the parameters (p, q) so that, under small samples $m = 10$, all three constraints from the rule-of-thumb are satisfied when $(p, q) = (0.3, 0.8)$, while only one, i.e. $m(x_1 * x_2)_1 \geq 1$, holds for $(p, q) = (0.1, 0.9)$.

We select these parameters in order to check whether the convolution statistic offers a better alternative over Pearson's χ^2, under cases favourable to P_m^{GF} and P_m^{ED}, when $(p, q) = (0.3, 0.8)$ and sample sizes are equal, or unfavourable, when $n_1 + n_2 - 2m > 0$ observations are excluded and the rule-of-thumb is violated.

For the small sample cases when $(p, q) = (0.3, 0.8)$, C_{2m}^{GF} provides better power over P_m^{GF}, and the latter over C_{1m}^{GF}, but C_{2m}^{GF} shows a proportion of rejections that is above α under $\mathbf{H}_0$ (Fig. 4.1a, top left and middle left panels). When $(p, q) = (0.1, 0.9)$, C_{1m}^{GF} and C_{2m}^{GF} have similar behaviour which outperforms P_m^{GF} (Fig. 4.1a, top right and middle right panels). Equivalent conclusions are inferred for the equality in distribution testing statistic counterparts (Fig. 4.1b, top and middle panels). For large samples, under $\mathbf{H}_0$, the proportion of rejections becomes closer to α for C_{2m}^{GF} (Fig. 4.1a, bottom panels) and it coincides for C_{2m}^{ED} (Fig. 4.1b, bottom panels); in terms of power, convolution statistics outperform Pearson's χ^2, with the exception of C_{1m}^{GF} and C_{1m}^{ED} when $(p, q) = (0.3, 0.8)$ (Fig. 4.1a-b, bottom left panels).

In Fig. 4.2 we illustrate the convergence of the rejection rate for m large to the significance level α under $\mathbf{H}_0$ and the power convergence under the alternative hypothesis $\mathbf{H}_1^{0.25}$. When testing for goodness-of-fit, C_{2m}^{GF} shows the highest rejection proportion which leads to good power under the alternative hypothesis (Fig. 4.2b), but a slow convergence to α under $\mathbf{H}_0$, reaching peaks of rejection up to 2α (Fig. 4.2a). C_{1m}^{GF} and P_m^{GF}, instead, have similar behaviours better than C_{2m}^{GF}, with P_m^{GF} outperforming C_{1m}^{GF} in its most favourable case ($(p, q) = (0.3, 0.8)$, bottom left panels from Fig. 4.2a,b)

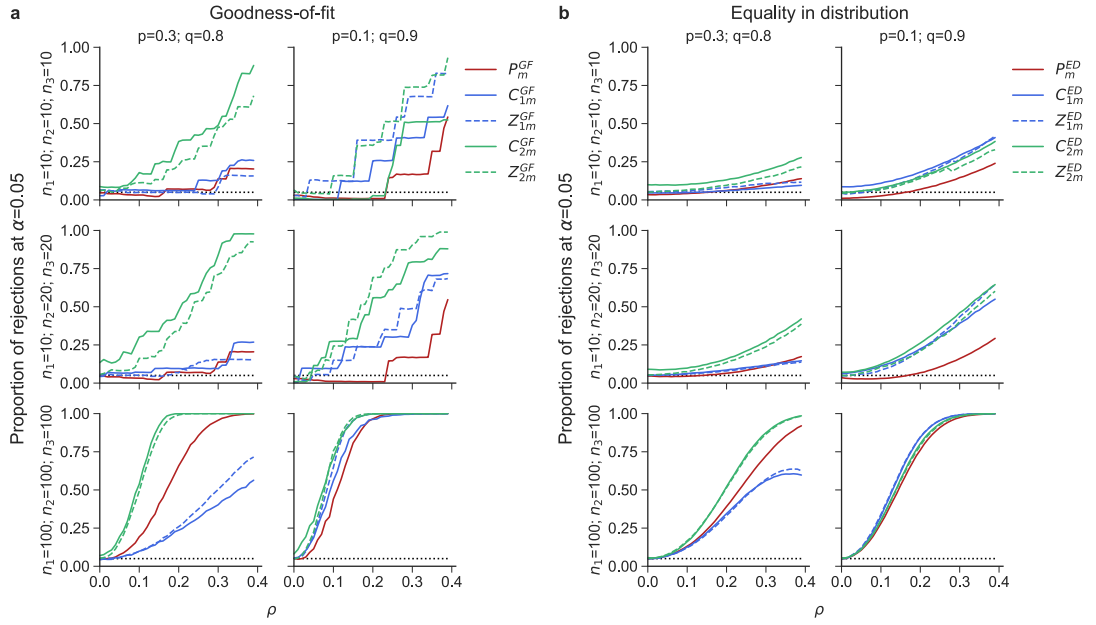

FIGURE 4.1: **Power comparison.** Solid lines correspond to Pearson's χ^2 statistics P_m^{GF} and P_m^{ED} (red), convolution statistic with rank 1 C_{1m}^{GF} and C_{1m}^{ED} (blue) and with rank 2 C_{2m}^{GF} and C_{2m}^{ED} (green). Dashed lines indicate convolution statistics calculated from the true covariance matrix, approximated to rank 1 Z_{1m}^{GF} and Z_{1m}^{ED} (blue) and to rank 2 Z_{2m}^{GF} and Z_{2m}^{ED} (green). These statistics are sorted for those testing goodness-of-fit, $\mathbf{H}_0 : X_1 + X_2 \sim z(0)$ against $\mathbf{H}_1^\rho : X_1 + X_2 \nsim z(\rho)$ (grid a), and equality in distribution, $\mathbf{H}_0 \colon X_1 + X_2 \sim Y_1$ against $\mathbf{H}_1^\rho \colon X_1 + X_2 \nsim Y_1$ (grid b), where $X_1 \sim (1-p, p)$, $X_2 \sim (1-q, q)$ and $Y_1 \sim z(\rho)$ as from (4.29). Distinct choices of (p, q) are used for each column of one grid ($(0.3, 0.8)$ left, $(0.1, 0.9)$ right). Different sample sizes (n_1, n_2, n_3), for X_1, X_2 and Y_1 respectively, are employed for each row within a grid (small sample sizes that are equal $(10, 10, 10)$ or unequal $(10, 20, 20)$ and equally large sample sizes $(100, 100, 100)$ from top to bottom). The proportion of rejections (see (4.30) and (4.31)) is plotted as function of the parameter $\rho = 0, 0.01, \ldots, 0.4$, so as to indicate the power under the null hypothesis $\mathbf{H}_0$, when $\rho = 0$, and under the alternative hypothesis $\mathbf{H}_1^\rho$, when $\rho > 0$. A dotted black horizontal line is depicted at $\alpha = 0.05$, the nominal rejection level for $\mathbf{H}_0$

while the converse holds in the other cases. When testing for equality in distribution, results for C_{1m}^{ED} and P_m^{ED} are similar (Fig. 4.2c,d), while C_{2m}^{ED} presents a much faster convergence under $\mathbf{H}_0$ than its goodness-of-fit counterpart, as it approaches α already at $m = 100$ (Fig. 4.2c).

Together, Figs. 4.1 and 4.2 suggest a tendency of the convolution statistics to attain a more anti-conservative behaviour (type I error higher than α), while for Person's χ^2 statistic this is more conservative (type I error lower than α).

Lastly, in Fig. 4.3 we analyse the convolution statistic in the case of a covariance matrix that is near a reduced rank form, when its smallest positive eigenvalue approaches zero. Equivalently, this situation occurs if the roots of the PMVs for X_1 and X_2 are close, i.e. $(p-1)/p - (q-1)/q$ becomes null. To this end, we fix $q \in (0, 1)$, so that $\mathrm{rk}(\Psi) = 1$

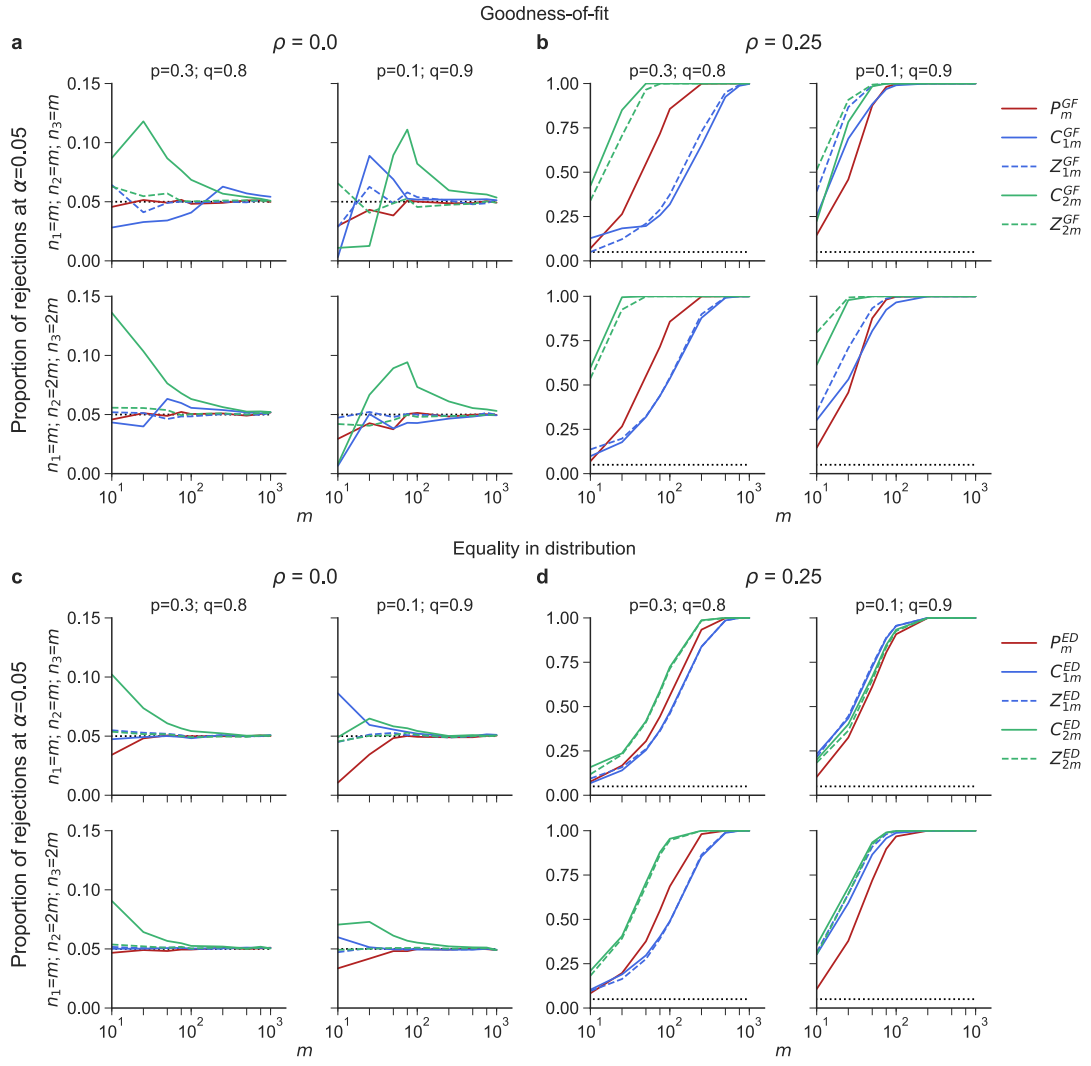

FIGURE 4.2: **Speed of convergence comparison.** Solid lines correspond to Pearson's χ^2 statistics P_m^{GF} and P_m^{ED} (red), convolution statistic with rank 1 C_{1m}^{GF} and C_{1m}^{ED} (blue) and with rank 2 C_{2m}^{GF} and C_{2m}^{ED} (green). Dashed lines indicate convolution statistics calculated from the true covariance matrix, approximated to rank 1 Z_{1m}^{GF} and Z_{1m}^{ED} (blue) and to rank 2 Z_{2m}^{GF} and Z_{2m}^{ED} (green). These statistics are sorted for those testing goodness-of-fit, $\mathbf{H}_0 : X_1 + X_2 \sim z(0)$ (grid a) against $\mathbf{H}_1^{0.25} : X_1 + X_2 \nsim z(0.25)$ (grid b), and equality in distribution, $\mathbf{H}_0 \colon X_1 + X_2 \sim Y_1$ (grid c) against $\mathbf{H}_1^{0.25} \colon X_1 + X_2 \nsim Y_1$ (grid d), where $X_1 \sim (1-p, p)$, $X_2 \sim (1-q, q)$ and $Y_1 \sim z(\rho)$ as from (4.29), with $\rho = 0, 0.25$. Distinct proportions of sample sizes (n_1, n_2, n_3), for X_1, X_2 and Y_1 respectively, are used for each row of the grids: equal $n_1 = n_2 = n_3$ (top row) or unequal $2n_1 = n_2 = n_3$ (bottom row). Different choices of (p, q) are used for each column of one grid ($(0.3, 0.8)$ left, $(0.1, 0.9)$ right). The proportion of rejections (see (4.30) and (4.31)) is plotted as a function of the minimum sample size $m = \min(n_1, n_2, n_3) = 10, 25, 50, 75, 100, 250, 500, 750, 1000$. A dotted black horizontal line is depicted at $\alpha = 0.05$, the nominal rejection level for $\mathbf{H}_0$

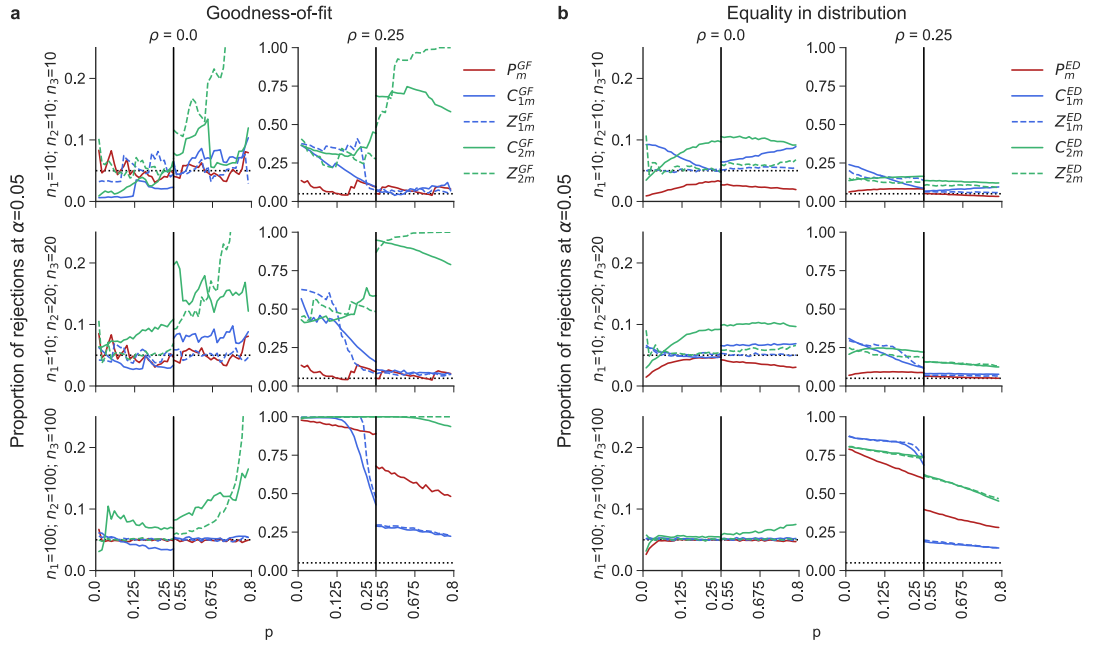

FIGURE 4.3: **Proportion of rejections comparison as the smallest positive eigenvalue of Ψ tends to zero.** Solid lines correspond to Pearson's χ^2 statistics P_m^{GF} and P_m^{ED} (red), convolution statistic with rank 1 C_{1m}^{GF} and C_{1m}^{ED} (blue) and with rank 2 C_{2m}^{GF} and C_{2m}^{ED} (green). Dashed lines indicate convolution statistics calculated from the true covariance matrix, approximated to rank 1 Z_{1m}^{GF} and Z_{1m}^{ED} (blue) and to rank 2 Z_{2m}^{GF} and Z_{2m}^{ED} (green). These statistics are sorted for those testing goodness-of-fit, $\mathbf{H}_0 : X_1 + X_2 \sim z(0)$ (gird a, left column) against $\mathbf{H}_1^{0.25} : X_1 + X_2 \nsim z(0.25)$ (grid a, right column), and equality in distribution, $\mathbf{H}_0 \colon X_1 + X_2 \sim Y_1$ (gird b, left column) against $\mathbf{H}_1^{0.25} \colon X_1 + X_2 \nsim Y_1$ (gird b, right column), where $X_1 \sim (1-p, p)$, $X_2 \sim (1-q, q)$ and $Y_1 \sim z(\rho)$ as from (4.29), with $\rho = 0, 0.25$. Different sample sizes (n_1, n_2, n_3), for X_1, X_2 and Y_1 respectively, are employed for each row within a grid $((10, 10, 10)$, $(10, 20, 20)$, $(100, 100, 100)$ from top to bottom). With q fixed to 0.8, the proportion of rejections (see (4.30) and (4.31)) is plotted as a function of p for values far from q ($p = 0.01, 0.02, \ldots, 0.25$, left side of the plot) or close to it ($p = 0.55, 0.56, \ldots, 0.79$, right side of the plot). A dotted black horizontal line is depicted at $\alpha = 0.05$, the nominal rejection level for $\mathbf{H}_0$

if $p = q$ (but still $\mathrm{rk}(\Psi + \Xi) = 2$), and we compare the proportion of rejections when the roots are well separated or close each other, under both $\mathbf{H}_0$ and $\mathbf{H}_1^{0.25}$.

For small samples, we observe that C_{1m}^{GF} and P_m^{GF} present similar performance under $\mathbf{H}_0$ (Fig. 4.3a, left panels), while P_m^{ED} and C_{1m}^{ED} are, respectively, conservative and anti-conservative (Fig. 4.3b, left panels). The power under $\mathbf{H}_1^{0.25}$ favours the use of C_{1m}^{GF} over P_m^{GF} in the small sample setting, but, for large samples, P_m^{GF} outperforms C_{1m}^{GF} when p is near q (Fig. 4.3a, right panels). As the spread between p and q widens, eventually C_{1m}^{GF} performs better than P_m^{GF}. This commentary is also true for the equality in distribution statistic counterparts (Fig. 4.3b, right panels).

As predicted from Figs. 4.1 and 4.2 analyses, the behaviour of C_{2m}^{GF} is of higher proportion of rejections. Furthermore, due to $\mathrm{rk}(\Psi) = 1$, as p approaches q, C_{2m}^{GF} undertakes

a dramatic deviation from the nominal rejection proportion α, indicative of the consistency failure of $(\hat{\Psi}^2_m)^+$ in the estimation of Ψ^+ when $p = q$ (Fig. 4.3a, bottom left panels). For large samples, we see that C^{ED}_{2m} is more powerful than C^{ED}_{1m}, since $\mathrm{rk}(\Psi + \Xi) = 2$, and also more powerful than P^{ED}_m when the roots are close (Fig. 4.3b, bottom panels).

The discrepancy between C^{GF}_{2m} and C^{ED}_{2m} behaviours highlights the merit for the rank analysis of Section 4.4 and shows the danger of setting the same degrees of freedom for the convolution as for Pearson's χ^2 statistics without careful consideration.

4.6 Discussion

Providing a thorough comparison between convolution statistic and Pearson's χ^2 would be very laborious for scenarios more complex than the one discussed above. In particular, the combination of cases rapidly increases when selecting different parameters for: small and large samples, of either equal or unequal size; alternative hypotheses; spread between PMVs' roots; number of random variables to be summed. Based on the results from the previous section, however, we propose the following as general indications:

- When the number of samples is large and PMV's roots are distinct, the convolution statistic C^{GF}_{rm} with $r = \mathrm{rk}(\Psi)$, or C^{ED}_{rm} with $r = \mathrm{rk}(\Psi + \Xi)$, provides the best power;

- When samples are small or some PMV's roots are close, C^{GF}_{rm} for $r < \mathrm{rk}(\Psi)$, or C^{ED}_{rm} for $r < \mathrm{rk}(\Psi + \Xi)$ provides a good compromise between type I and type II errors control;

- Pearson's χ^2 is recommended over the convolution statistic approach only for small samples in which the rule-of-thumb is not violated, especially when the type I error is allowed to be smaller than the nominal level α and type II error is considered of secondary importance.

Ultimately, it is desirable to run the comparison over any scenario of interest, particularly to evaluate the optimal extent of rank reduction for the convolution statistic. Figs. 4.1–4.3 also show that the convolution statistics Z^{GF}_{rm} and Z^{ED}_{rm}, calculated with the exact covariance matrix (dashed lines), would provide the best results in most situations considered. Thus, we speculate that any improvement on the estimation of the covariance matrix pseudo-inverse would lead to an even better power.

Chapter 5

Multiplexed division tracking dyes for clonal lineage tracing

5.1 Abstract

In this chapter, we present the work from Horton, Prevedello et al., (2018)[1] accomplished in collaboration with our partners in Prof. Philip Hodgkin's lab at Walter Eliza Hall Institute (WEHI). As activated lymphocytes undergo several rounds of division and result in a population of cells with diverse traits and functions, distinguishing between intrinsic and extrinsic sources of such heterogeneity is difficult with the current experimental techniques and would benefit from new, more practical methods. To this end, the high-throughput procedure based on the multiplex clonal assay from Chapter 2 is improved to include the measure of cellular expression levels, together with clonal and generational information. The output data are analysed through statistical techniques that account for complex dependence associations between clonally related cells. The method is illustrated by studying the *in vitro* activation of murine $CD8^+$ T-cell cultures. This approach has broad utility as it can be applied to other *in vitro* culture systems and, potentially, *in vivo*.

5.2 Introduction

Determining the contribution of asymmetric cell division, intercellular communication, quorum sensing, lineage priming, and autonomous programming to clonal cell fate is a key focus of immunology and many other fields of biology (Snippert et al., 2010;

[1]Miles B. Horton and Giulio Prevedello equally contributed to this work.

Buchholz et al., 2013; Gerlach et al., 2013; Perié et al., 2015; Yu et al., 2016; Heinzel et al., 2018). However, progress has been impeded by the low throughput and laborious nature of common lineage tracing and fate mapping approaches such as time lapse microscopy. Recently introduced technologies such as retroviral barcoding (Gerlach et al., 2013; Naik et al., 2013), CRISPR-induced heritable genetic lesions (McKenna et al., 2016), and the development of fluorescent lineage reporters (Livet et al., 2007; Tas et al., 2016; Yu et al., 2016) have contributed to improved throughput in lineage tracing experiments, and thus revealed important discoveries into the emergence of diverse cell types across multiple systems. Despite this success, such methods remain highly resource-dependent and time-consuming. Furthermore, these methods typically lack information regarding clonal division progression, an important source of information in understanding the mechanisms that drive cell fate decisions (Hodgkin et al., 1996; Bird et al., 1998; Gett and Hodgkin, 1998; Tangye et al., 2003; Jenkins et al., 2008; Kueh et al., 2013; Kinjyo et al., 2015).

Although many cellular processes across multiple systems have demonstrated an association between cell state transitions and division (Bird et al., 1998; Gett and Hodgkin, 1998; Kueh et al., 2013; Bernitz et al., 2016; Kueh et al., 2016; Polonsky et al., 2016), thorough examination of these associations across the progeny of expanding single cell lineages has, to date, been limited. A fast, easy, high-throughput method that, for individual clones, simultaneously measured division progression as well as cell state, in the form of marker and/or fluorescent reporter expression, could therefore significantly contribute to progress in this field. In this chapter, we introduce such a method, utilising multiplexed division tracking dyes in combination with flow cytometry-based phenotyping.

Earlier variants of the dye-multiplexing approach have been applied to high-throughput cytotoxicity assays (Quah and Parish, 2012), analysis of clonal division progression (Marchingo, Prevedello et al., 2016), and identification of distinct co-cultured cell populations (Voisinne et al., 2015). Here, we demonstrate that the utility of this method can be significantly extended by integrating phenotypic information with proliferation-based lineage tracing and by provision of the statistical tools necessary for data interrogation.

5.3 Multiplexing division tracking dyes

The premise of the method is to label the cells under consideration with distinct combinations and concentrations of division tracking dyes, generating multiple unique fluorescence profiles (Fig. 5.1a). After labelling, cells are sorted according to their fluorescence

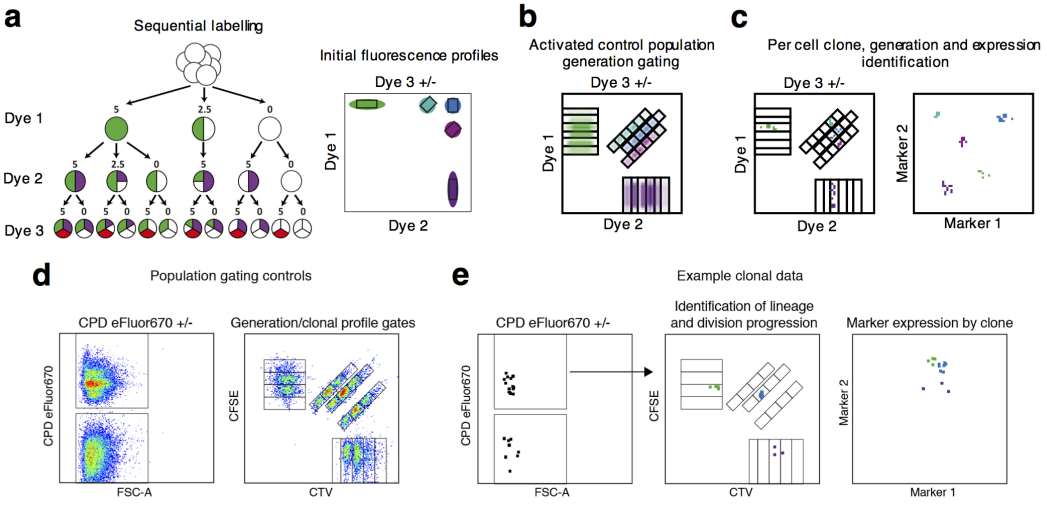

FIGURE 5.1: **Dye labelling strategy to generate multiple unique fluorescence profiles.** [Corresponding to Figure 1 from Horton, Prevedello et al., (2018)] Protocol schematic. (a) Cells of interest are sequentially labelled with combinations of division diluting fluorescent dyes to generate distinct fluorescence profiles. Numbers depict micromolar dye concentration. (b) For each profile, the proliferation of bulk populations is used to identify generation-determining gates. (c) Single representatives from each profile are sorted and placed in the system of interest. On harvest, FACS measurement reveals clonal membership, cell division number, and phenotype. (d) Example data of bulk population controls used to set lineage and proliferation gates. Cells are first separated into CPD+ and CPD- and the combinations of CFSE and CTV are used to define 5 distinct fluorescence signatures for both populations. (e) Example data of an individual well showing the implementation of the gating strategy used in (d). Shown are cells first gated on CPD+ and clones are then identified using control-generated gates and their division progression and marker expression is analysed.

profile and placed in a system of interest, such as an *in vitro* or *in vivo* environment. In concert, bulk populations of labelled cells are used to identify generation-determining gates for each unique profile (Fig. 5.1b). On recovery at a later time, the lineage-membership, generation-number, and phenotypic state of each cell can be determined by flow cytometry (Fig. 5.1c).

As a number of division-diluting dyes with distinct fluorescent spectra are commercially available, the number of the combinatorially-created distinguishable profiles generated can be optimized for the system of interest. Furthermore, the use of division tracking dyes to monitor clonal lineages is a significant feature of this approach, enabling simultaneous measurement of phenotypic changes and clonal division progression.

5.4 Tracing fluorescently-labelled CD8$^+$ T-cell clonal progeny

For illustration of the method, we analysed the *in vitro* differentiation of stimulated, purified murine CD8$^+$ T cells at the level of individual clones. Upon activation, CD8$^+$

T cells generate substantial population-level heterogeneity, which is underpinned by a significant familial component (Buchholz et al., 2013; Gerlach et al., 2013; Lemaître et al., 2013; Plumlee et al., 2013; Marchingo, Prevedello et al., 2016). In this section, the experimental protocol for the multiplex clonal assay is reported as designed by our WEHI partners M. B. Horton, J. M. Marchingo, J. H. S. Zhou, S. Heinzel, P. D. Hodgkin and my supervisor K. R. Duffy[2], and also implemented from M. B. Horton. Additional details of the protocol are deferred to Appendix A. Subsequent sections will cover our contribution to the visualisation and the statistical analysis of the data produced from this method.

Purified murine $CD8^+$ T cells were labelled with three division-tracking dyes, CFSE, CTV and CPD, resulting in 10 distinct combinations, and then stimulated with anti-CD3, anti-CD28 and rhIL-2. Anti-mouse IL-2 blocking antibody was also added to remove the effect of any endogenous production (Deenick et al., 2003; Marchingo et al., 2014). After 24 hours, just prior to their first division (Marchingo et al., 2014), a single founder cell from each of the fluorescence profiles was sorted and mixed into each of 29 tissue culture wells, allowing analysis of up to 10 distinct, co-cultured clonal families per well. In parallel cultures, cells from each fluorescence signature were sorted into new tissue culture plates. In all cases the cells were maintained in the same stimulatory conditions as during the initial activation period.

Sixty hours after initial stimulation, cells were harvested and analysed for division progression (Fig. 5.1d) and expression of CD8, CD62L and CD25 by flow cytometry (Fig. 5.1e). A known number of beads was added to each well to enable estimation of sample recovery. Pooled across wells, 156 clonal families constituting a total of 865 cells spread over 4 generations were recovered. The resulting data is presented in Fig. 5.2 and permits the concurrent visualisation of clonal lineage, marker expression level and division progression.

Additional independent experiments were performed using the same stimulation conditions described in Fig. 5.2 and analysed at different time-points (Fig. 5.3), as well as experiments utilising $CD8^+$ T cells from distinct transgenic and reporter mice under different culture conditions. Data from these further experiments are provided in Fig. 5.4 and 5.5, respectively including the clonal expression of the transcription factor Blimp-1 and the chemokine receptor CXCR3. For additional information concerning the experimental methods, we defer the reader to Section A.2 of Appendix A.

Upon visualisation it is clear that, complementary to previously demonstrated division synchrony (Marchingo, Prevedello et al., 2016), clones display substantial familial homogeneity. For each marker, CD8, CD62L and CD25, the overall distribution in

[2]K. R. Duffy, S. Heinzel and P. D. Hodgkin share senior authorship.

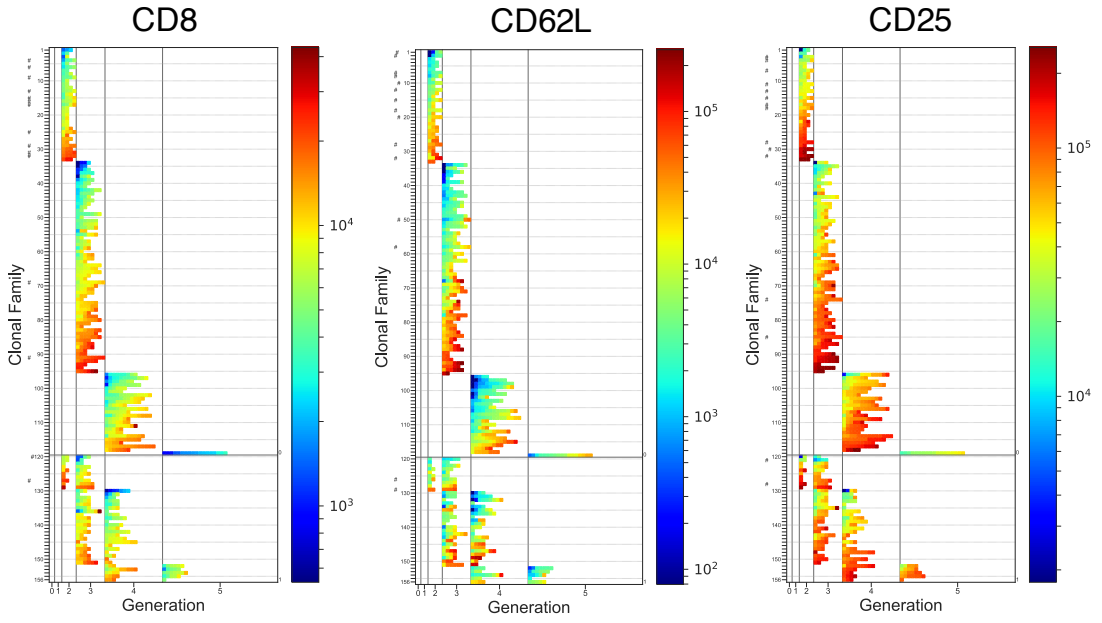

FIGURE 5.2: **Simultaneous visualisation of marker expression, division progression and clonal lineage membership in activated CD8$^+$ T cells.** [Corresponding to Figure 2 from Horton, Prevedello et al., (2018)] Purified C57BL/6 CD8$^+$ T cells were sequentially dye labelled with CFSE, CTV and CPD, resulting in 10 unique profiles (Fig. 5.1). These cells were stimulated with anti-CD3 (10 μg ml^{-1}), anti-CD28 (2 μg ml^{-1}) and rhIL-2 (31.6 U ml^{-1}) for 24h in the presence of an anti-mouse IL-2 blocking antibody clone S4B6 (25 μg ml^{-1}). Single cells from each of the 10 combinations were sorted and pooled into each of 29 individual wells followed by culture for a further 36h. Generation number and fluorescence intensity of CD8 (APC-Cy7), CD62L (PE) and CD25 (PE-Cy7) expression were determined by flow cytometry 60h post-stimulation. Image displays data pooled from all wells. Vertical column bins represent generation numbers, rows represent clonal families and data points represent cells. Cell colour indicates marker fluorescence intensity according to the provided legend. Clones whose cells were found in the same generation are ordered first, followed by clones whose cells were found in adjacent generations, and are rank ordered within groups by geometric mean fluorescence. '♯' denotes fully recovered clones - those for whom every cell is measured. Of 300 clones initially seeded, 156 families were detected with at least 2 members, yielding a recovery of 52%.

expression level varies over one-to-two orders of magnitude across the CD8$^+$ T-cell population, whereas the intraclonal distribution of expression is far narrower. This suggests that, under these stimulation conditions, a key source of phenotypic heterogeneity across a population of CD8$^+$ T cells early after *in vitro* activation is underpinned by intraclonal concordance and interclonal variation.

5.5 Statistical tools for the analysis of phenotypic clonal data

The data produced by the assay has an unusual structure that necessitates careful consideration for statistical hypothesis testing. The primary concern is that clones

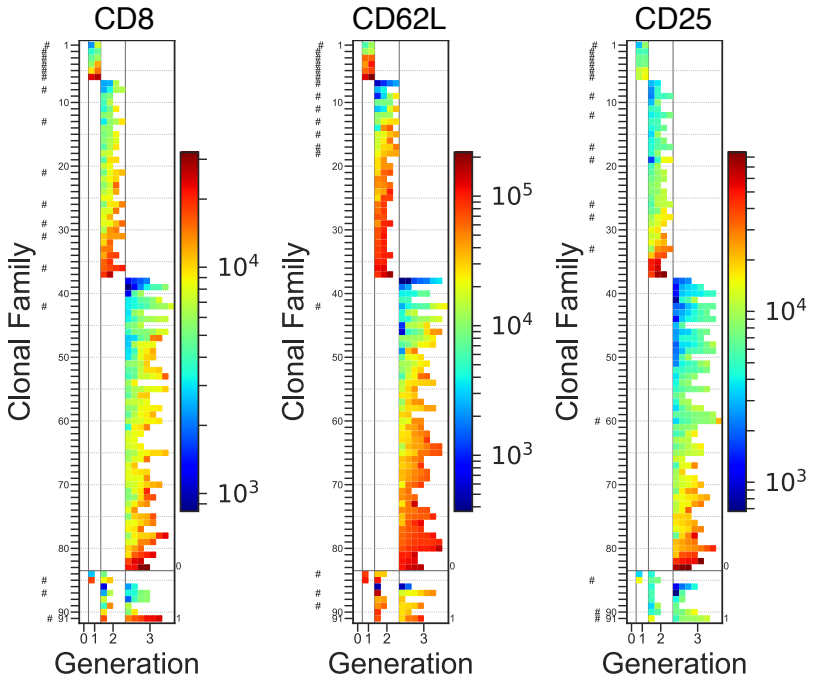

FIGURE 5.3: **Multiplexing tracking dyes to determine clonal membership, generation number and cell phenotype is reproducible across time points.** [Corresponding to Supplemental Figure 1 from Horton, Prevedello et al., (2018)] Purified murine $CD8^+$ T cells were processed analogously as in Fig. 5.2, and multiplex dye labelled with 5-(and 6)-carboxyfluorescein diacetate succinimidyl ester (CFSE), CellTrace Violet (CTV) and Cell Proliferation dye eFluor670 (CPD), resulting in 10 profiles. These cells were stimulated with anti-CD3 (10 μg ml^{-1}), anti-CD28 (2 μg ml^{-1}) and rhIL-2 (31.6 U ml^{-1}) for 24h in the presence of the anti-mouse IL-2 blocking antibody clone S4B6 (25 μg ml^{-1}). Single cells from each of the 10 combinations were sorted and pooled into 20 individual wells followed by culture for a further 27h without additional anti-CD3 stimulation. Generation number and fluorescence intensity of CD8 (APC-Cy7), CD62L (PE) and CD25 (PE-Cy7) expression were determined by flow cytometry 51h post-stimulation. Of 160 clones initially seeded, 91 families were detected with at least 2 members, yielding a recovery of 56.9%.

consist of a relatively small number of cells so that statistical tests based on asymptotic results may be inappropriate. A secondary concern is that the proportion of each clone recovered from a single captive environment (culture well, animal, etc.) can result in a systemic, rather than biological, statistical coupling between co-habiting clones that must be circumvented. Thus, to complement the experimental method implemented by our collaborators, we developed a choice of simple-to-implement, non-asymptotic permutation tests (Lehmann and Romano, 2005) to interrogate the data (Fig. 5.6a) for a range of null hypotheses. Before their application, the principles of this statistical procedure are outlined in Section 5.5.1.

A natural exploratory statistic based on the label-permuted data can also be plotted, providing visual cues as to the likely outcome of such tests (Fig. 5.6b).

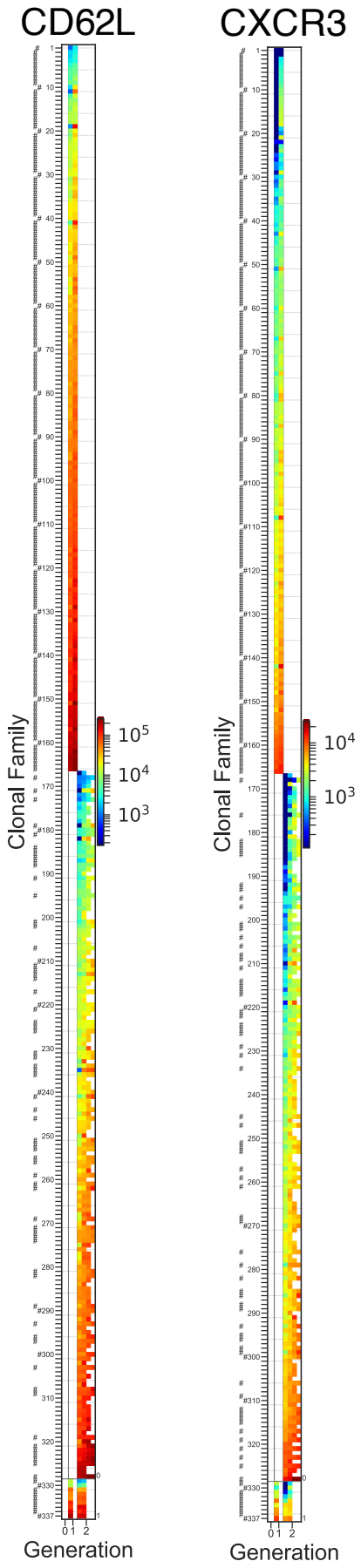

FIGURE 5.4: **Multiplexed tracking dyes to determine clonal membership and generation number can be adjusted to assess distinct components of cell phenotype.** [Corresponding to Supplemental Figure 2 from Horton, Prevedello et al., (2018)] Purified OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were multiplex dye labelled with CFSE, CTV and CPD, resulting in 10 profiles. These cells were stimulated with N4 peptide (0.01 μg ml^{-1}) and IL-4 (1,000 U ml^{-1}) for 24h in the presence of the anti-mouse IL-2 blocking antibody clone S4B6 (25 μg ml^{-1}). Single cells from each of the 10 combinations were sorted and pooled into 59 individual wells followed by culture for a further 22h. Generation number and fluorescence intensity of CXCR3 (PE-Cy7) and CD62L (PE) expression were determined by flow cytometry 46h post-stimulation. Image displays data pooled from all wells. Vertical column bins represent generation numbers, rows represent clonal families and data points represent cells. Cell colour indicates marker fluorescence intensity according to the provided legend. Clones whose cells were found in the same generation are ordered first, followed by clones whose cells were found in adjacent generations, and are rank ordered within groups by geometric mean fluorescence. '♯' denotes fully recovered clones. Of 600 clones initially seeded, 337 families were detected with at least 2 members, yielding a recovery of 56.2%.

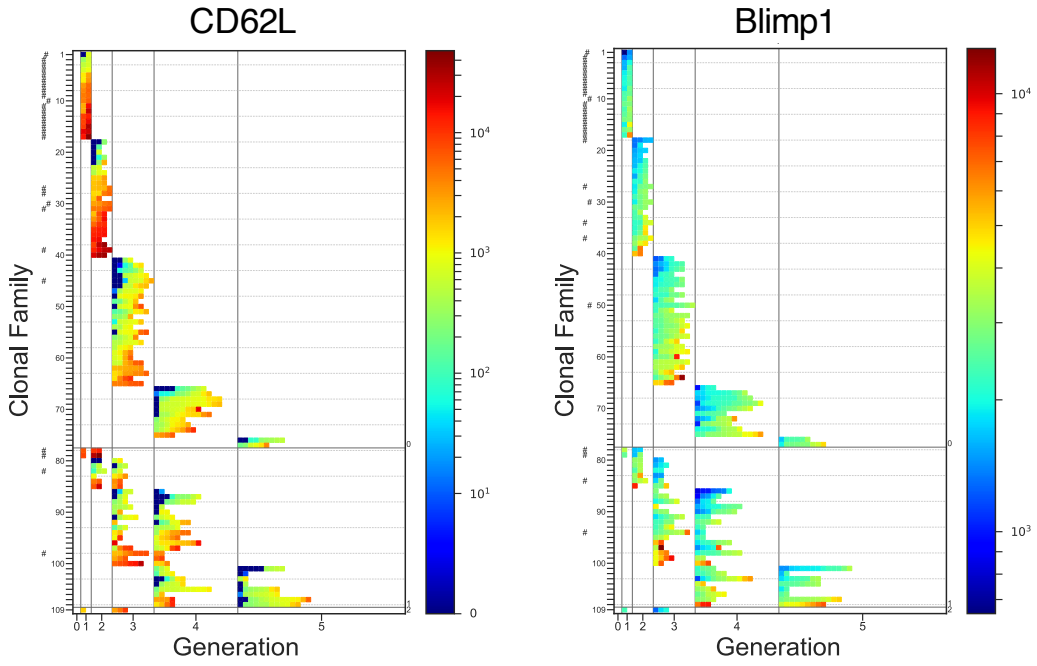

FIGURE 5.5: **Distinct tracking dyes can be multiplexed to allow for compatibility with fluorescent reporters.** [Corresponding to Supplemental Figure 3 from Horton, Prevedello et al., (2018)] Purified Blimp$^{gfp/+}$ CD8^{+} T cells were multiplex dye labelled with CellTrace Yellow (CTY), CTV and CPD, resulting in 6 profiles. These cells were stimulated with anti-CD3 (10 μg ml^{-1}), rhIL-2 (31.6 U ml^{-1}) and mIL-12 (10 ng ml^{-1}) for 24h in the presence of the anti-mouse IL-2 blocking antibody clone S4B6 (25 μg ml^{-1}). Single cells from each of the 6 combinations were sorted and pooled into 49 individual wells followed by culture for a further 42h. Generation number and fluorescence intensity of CD62L (APC-Cy7) and Blimp1 (GFP-reporter) expression were determined by flow cytometry 66h post-stimulation. Image displays data pooled from all wells. Vertical column bins represent generation numbers, rows represent clonal families and data points represent cells. Cell colour indicates marker fluorescence intensity according to the provided legend. Clones whose cells were found in the same generation are ordered first, followed by clones whose cells were found in adjacent generations, and are rank ordered within groups by geometric mean fluorescence. '♯' denotes fully recovered clones. Of 360 clones initially seeded, 109 families were detected with at least 2 members, yielding a recovery of 30.3%.

For illustration, the statistical pipeline described in Fig. 5.6c-e was applied to the data shown in Fig. 5.2. Fig. 5.6f plots the CD62L expression levels of all 865 cells pooled, as well as fractionated per-well (i.e. per-environment) and per-clone, where the latter two are rank ordered from highest mean geometric fluorescence to lowest. Fig. 5.6g plots the difference between the rank ordered geometric mean fluorescence of the data and the geometric mean fluorescence of the label-reassigned data, averaged over reassignments, as well as 95% confidence intervals under the null hypothesis. For these data, the per-well statistic consistently lies within the confidence intervals, while the per-clone statistic falls far outside.

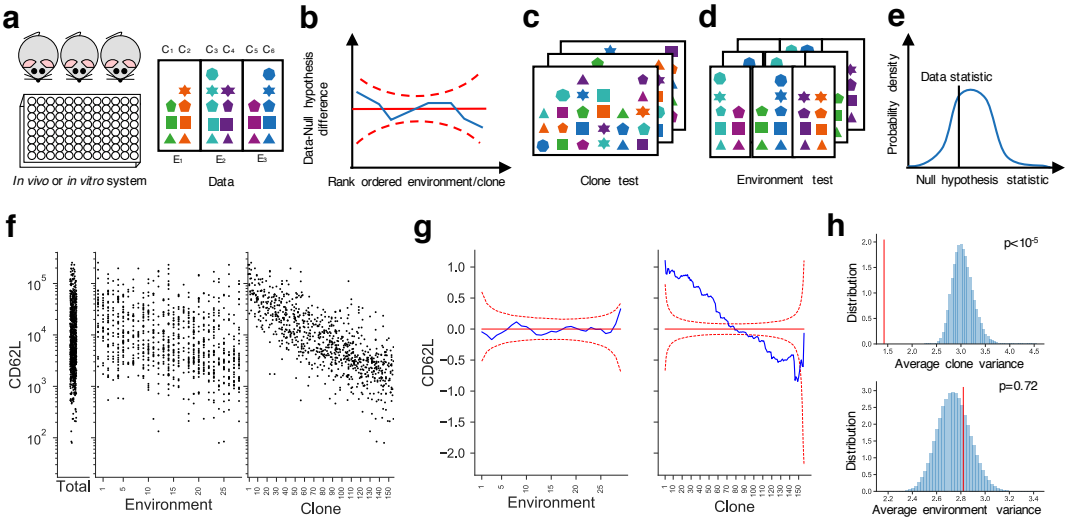

FIGURE 5.6: **Testing for independence of phenotype and clonal membership or environment.** [Corresponding to Figure 3 from Horton, Prevedello et al., (2018)] (a-e) Statistical schematic. (a) Multiplex data are collected from the *in vitro* or *in vivo* system and fractioned in distinct clones (C_1, C_2, ...) or environments (E_1, E_2, ...), e.g. wells or animals. (b) For a given per-clone or per-environment statistic, the data are rank-ordered (blue line), and compared against the 95% confidence intervals of the centred distribution obtained from the rank-ordered of the permuted datasets (red lines). (c) To test the null hypothesis that the expression level of cells is independent of generation, clone and environment, cell-to-clone labels are permuted. (d) To test the null hypothesis that the expression levels, expansion and recovery of clones are independent of the environment, clone-to-environment labels are permuted. The resulting p-value for both (c-d) is the proportion of testing statistics, calculated on the permuted datasets, as extreme as observed for the true assignment (see Section 5.5.1). (f-h) Example data with CD62L expression levels from Fig. 5.1d. (f) The data are pooled, and fractionated by environment (i.e. well) and clone, and rank-ordered from highest-to-lowest geometric mean. (g) The blue line is the difference between the rank ordered true data and the mean label-permuted data. Dashed red lines indicate 95% confidence intervals under the null hypothesis that expression is independent of label, as in (b). (h) The vertical red line indicates the location of the data statistic and, with a null hypothesis as in (c) (top panel) or (d) (bottom panel), the histogram shows the density of the same statistic determined for 250,000 uniformly-at-random permuted assignments of cell-to-clone (top panel) or clone-to-environment (bottom panel), with the lower one-sided p-value being the fraction whose statistic were smaller than for the true data (see equation (5.3) of Section 5.5.1).

5.5.1 Principles of permutation test procedures

Given a data set of $n \in \mathbb{N}$ ordered observations $D = (Z_1, Z_2, \ldots, Z_n)$, a permutation π of them is a reassignment of the labels of the individual datum, $i \to \pi(i)$, to create the reordered data set $D_\pi = (Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(n)})$. For example, if $\pi(i) = n + 1 - i$ for all $i = 1, \ldots, n$, then $D_\pi = (Z_n, Z_{n-1}, \ldots, Z_1)$ is the original data but in reverse order.

The principle of permutation testing is to evaluate a real valued statistic on the recorded data, denoted $T(D) \in \mathbb{R}$, where the statistic depends on the data order. If, for a given null hypothesis $\mathbf{H}_0$, a collection of data permutations Q can be characterised so that

all data reorderings $\{D_\pi\}_{\pi \in Q}$ are equally likely under $\mathbf{H}_0$, then $T(D)$ can be compared with the distribution of the statistic computed on the reordered datasets $T(D_\pi)_{\pi \in Q}$. In particular, denoting by $|Q|$ the number of elements in a set Q, the proportion of permutations that lead to a statistic that is lower than that observed for the true data order is the lower p-value

$$p_l = \frac{|\{\pi \in Q \colon T(D) \geq T(D_\pi)\}|}{|Q|}, \tag{5.1}$$

while the proportion of permutations that lead to a statistic that is higher than that observed for the true data order is the upper p-value

$$p_u = \frac{|\{\pi \in Q \colon T(D) \leq T(D_\pi)\}|}{|Q|}, \tag{5.2}$$

To realize a permutation test successfully, it is important that the collection of allowed permutations accurately describe the null hypothesis, and that the test statistic tends to deviate from the true statistic if the null hypothesis is not true.

For many tests, the number of possible permutations $|Q|$ is too large for $T(D_\pi)$ to be computed for every permutation $\pi \in Q$. For example, for data with n interchangeable elements under a null hypothesis, there are n factorial, permutations, which grows faster than exponentially in n. Thus it is common to use Monte Carlo methods to estimate p_l and p_u. This achieved by drawing a large number, $B \in \mathbb{N}$, of samples from Q uniformly at random and then making empirical estimates of the p-values p_l (5.1) and p_u (5.2), respectively as

$$\hat{p}_l^B = \frac{1 + \sum_{i=1}^{B} \mathbb{1}_{\{T(D) \geq T(D_{\pi_i})\}}}{B+1} \tag{5.3}$$

and

$$\hat{p}_u^B = \frac{1 + \sum_{i=1}^{B} \mathbb{1}_{\{T(D) \leq T(D_{\pi_i})\}}}{B+1}.$$

5.5.2 Implementation of permutation test procedures

The data from the multiplex clonal assay can be defined as a sequence of four elements

$$D = \Big((x_i, g_i, c_i, e(c_i)) \Big)_{i=1}^{N}, \tag{5.4}$$

where N is the total number of cells recovered and the i^{th} cell is encoded by the expression level $x_i \in \mathbb{R}$, its generation $g_i \in \mathcal{G} = \{0, \ldots, G\}$, its familial label $c_i \in \mathcal{C} = \{1, \ldots, M\}$, and the environment label associated to its family $e(c_i) \in \mathcal{E} = \{1, \ldots, E\}$, with $G, M, E \in \mathbb{N}$. Depending on the hypothesis to be tested, a statistic T on these data can be computed and compared with the distribution of the same statistic for a collection of permutations, as previously explained (Section 5.5.1).

An all-encompassing hypothesis would posit that each cell's expression level is independent of its clone, generation, and environment. Thus, in Fig. 5.6c, we tested for the null hypothesis $\mathbf{H}_0$ that every cell's fluorescence is equal in distribution irrespective of generation, clone and environment. Our test statistic was the per-clone variance in fluorescence averaged across all clones, that is

$$T(D) = \frac{1}{M} \sum_{a \in \mathcal{C}} \sigma^2\Big((x_i,\, i = 1, \ldots, N \colon c_i = a)\Big), \tag{5.5}$$

where

$$\sigma^2(A) = \frac{1}{|A| - 1} \sum_{z \in A} \left(z - |A|^{-1} \sum_{z' \in A} z' \right)^2 \tag{5.6}$$

is the sample variance over a finite sequence A of values in $\mathbb{R}$. The set Q of allowable permutations was all possible reordering of cell labels, resulting in cells being reassigned amongst clones and environments. Formally, an element $\pi \in Q$ was defined such that

$$\pi : D \mapsto D_\pi = \Big((x_{\tilde{\pi}(i)}, g_i, c_i, e(c_i)) \Big)_{i=1}^{N}, \tag{5.7}$$

with $\tilde{\pi}$ permutation of $\{1, \ldots, N\}$. In total there were 865 cells, then there were 865 factorial allowable permutations, requiring Monte Carlo methods to compute the one-sided test p-values as described in Section 5.5.1.

In order to challenge the null hypothesis that the expression levels of clones, rather than cells, are independent of their environment, in Fig. 5.6d we set $\mathbf{H}_0$ such that each clone's expansion, recovery and fluorescence levels are equal in distribution across environments. We defined the statistic to be the per-environment variance in fluorescence averaged across environments, that is

$$T(D) = \frac{1}{E} \sum_{a \in \mathcal{E}} \sigma^2\big((x_i,\, i = 1, \ldots, N \colon e(c_i) = a)\big). \tag{5.8}$$

There, permutations (i.e. Q) were all possible relabelings of the environment label of clones, effectively swapping whole clones across environments. This was achieved by

defining every $\pi \in Q$ such that

$$\pi : D \mapsto D_\pi = \big((x_i, g_i, c_i, e(\tilde{\pi}(c_i)))\big)_{i=1}^{N}, \tag{5.9}$$

with $\tilde{\pi}$ permutation of $\{1, \ldots, M\}$. The resulting p-value for these, and all other permutation tests, were the proportion of permuted assignments that resulted in a statistic that was at least as extreme as for the true assignment (Fig. 5.6e).

The application of these two testing procedures to the CD62L data from Fig. 5.2, demonstrated strong evidence that the expression of this cell surface receptor depends on clone ($p < 10^{-5}$), but no evidence of per-environment dependence ($p = 0.56$), for this system.

To challenge more nuanced hypotheses, a similar procedure can be used in conjunction with suitable restrictions on the class of allowed reassignments. For example, if one suspected that recovery of clones were environment-dependent but still wished to challenge if clonal expression was independent of the environment, one cannot arbitrarily reassign clones amongst environments as the test described in Fig. 5.6d could fail due to correlations in the level of clone recovery rather than any inherent biological environmental dependence. Instead, the desired test can be achieved by restricting reassignments across environments only to clones that are fully recovered (i.e. for which every expected cell is measured) and have the same generation structure.

In Fig. 5.7a we tested the null hypothesis that, regardless of the environment in which they are found, each clone's fluorescence levels are equal in distribution for clones at the same developmental stage (i.e. for clones that have the same number of cells in each generation). As in the previous test, the statistic T was the average per-environment variance in fluorescence (5.8). What had changed was that not all permutations of clones were allowable. Instead we first identified all families in which all cells were measured. Following the notation from Section 3.2.2 of Chapter 3, this requirement was verified if $v(c_i) \in \mathcal{S}_G \subseteq \mathbb{N}_0^{G+1}$, the sampled family vector associated to the i^{th} clone c_i, was such that its cohort number is equal to one, i.e. $\text{cn}(v) = 1$ and $v(c_i) \in \mathcal{V}_G$. Among the clones whose all cells were recovered, those that were characterised by the same family vector (i.e. generation profile) were interchangeable under the null hypothesis, and swapping these formed the basis of the permutations in Q. This worked as these clones were conditioned to not be subject to sampling bias. Therefore, the allowed transformation of the data was identified, in this case, as the set of maps $\pi \in Q$ such that

$$\pi : D \mapsto D_\pi = \big((x_i, g_i, c_i, e(\tilde{\pi}(c_i)))\big)_{i=1}^{N}, \tag{5.10}$$

with

$$\tilde{\pi}(c) = \begin{cases} \tilde{\pi}_w(c) & \text{if } v(c) = w \\ c & \text{otherwise} \end{cases}, \tag{5.11}$$

for any $\tilde{\pi}_w$ permutation of the set $\{c, c = 1, \ldots, M \colon v(c) = w\}$, where w is any family vector in $\mathcal{V}_G$, that is $\mathrm{cc}(w) = 1$.

Similar approaches can be used to test several alternate and restricted hypotheses regarding other dependencies of clonal progression and cellular phenotype.

In Fig. 5.7b we tested the null hypothesis that fluorescence levels are equal in distribution between cells from the same environment and generation, irrespective of their clone membership. The test statistic was the average per-clone variance (5.5) as in Fig. 5.6c, but again not all permutations of cell labels were allowed. Instead cells were only permuted with other cells of the same generation. To this end, a possible data rearrangement was identified by $\pi \in Q$ such that

$$\pi : D \mapsto D_\pi = \left((x_{\tilde{\pi}(i)}, g_i, c_i, e(c_i))\right)_{i=1}^{N}, \tag{5.12}$$

with

$$\tilde{\pi}(i) = \begin{cases} \tilde{\pi}_{g,e}(i) & \text{if } g_i = g \text{ and } e(c_i) = e \\ i & \text{otherwise} \end{cases}, \tag{5.13}$$

where $\tilde{\pi}_{g,e}$ is any permutation of the set $\{i,\, i = 1, \ldots, N \colon g_i = g, e(c_i) = e\}$, for any choice of $g \in \{0, \ldots, G\}$ and $e \in \{1, \ldots, E\}$, .

Finally, in Fig. 5.7c we tested the null hypothesis that clonal expansion and recovery are equal in distribution across different environments. The statistic was the average per-environment variance in clone size, that is

$$T(D) = \frac{1}{E} \sum_{a \in \mathcal{E}} \sigma^2 \left(\left(\sum_{i=1}^{N} \mathbb{1}_{\{c_i = c\}},\, c \leq M \colon e(c_i) = a \right) \right). \tag{5.14}$$

The collection of permutations Q was the swapping of clones across environments, such as in (5.9).

In all cases described, we reported the lower p-value approximated via Monte Carlo with $B = 250,000$ permutations, that is $\hat{p}_l^B$ from equation (5.3). This tests whether the data has lower average variance, and thus has greater within-group (i.e. clone or environment) relatedness, than one would expect under one of the null hypotheses.

Analysis for CD8 and CD25, equivalent as for CD62L, is presented in Fig. 5.8.

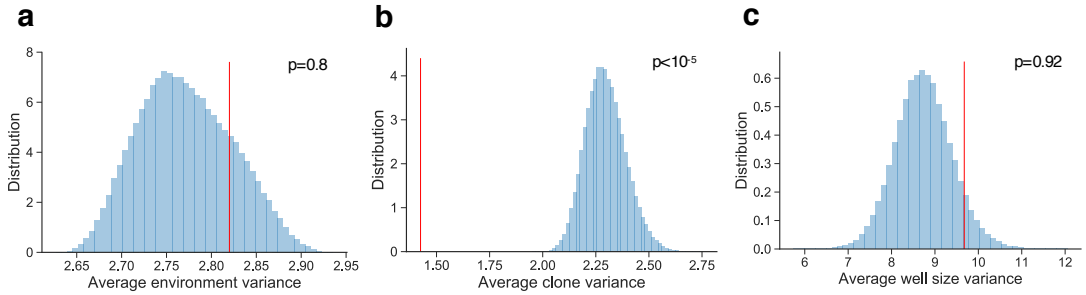

FIGURE 5.7: **Testing null hypotheses of independence.** [Corresponding to Supplemental Figure 4 from Horton, Prevedello et al., (2018)] Analysis of the CD62L data presented in Fig. 5.2. (a) To test the null hypothesis that each cell's fluorescence is independent of its membership of an environment, but potentially dependent on its generation, while also being cognizant that the sampling of clones in the same environment may lead to a coupling in their recovery, data permutation is restricted to clones that have the same generational structure and for whom all cells in each clone are measured (see Section 5.5.2). (b) To test the null hypothesis that each cell's fluorescence is independent of its clonal membership, but potentially dependent upon its generation, data permutation is restricted to cells across clones within the same generation (see Section 5.5.2). (c) To test for the null hypothesis that clonal expansion and recovery are independent of environmental membership, clones are permuted across environments (see Section 5.5.2). The vertical red line indicates the location of the data statistic of the originally ordered data. The histogram shows the density of the same statistic determined for 250,000 uniformly-at-random permutations of the data. The lower one-sided p-value is depicted in legend (see equation (5.3) of Section 5.5.1), resulting in rejection of the hypothesis in b and non-rejection of the hypotheses in a and c with a significance level of 0.05.

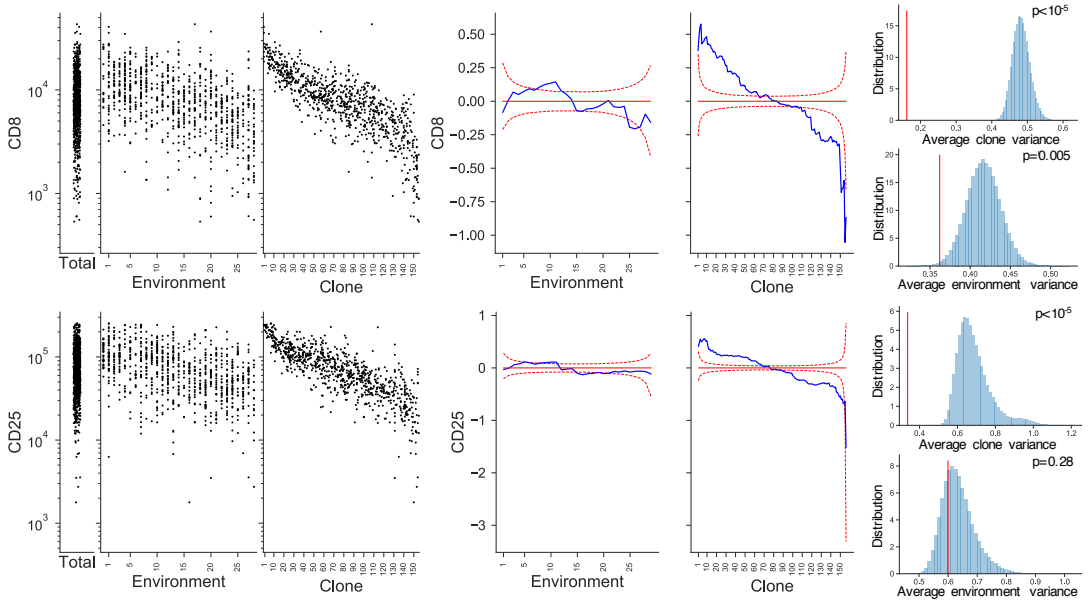

FIGURE 5.8: **Visualization, and testing for environment and clone independence.** [Corresponding to Supplemental Figure 5 from Horton, Prevedello et al., (2018)] Analogous analysis as in Fig. 5.6f-h for the CD8 (top row) and CD25 (bottom row) data in Fig. 5.2.

5.6 Analysis of first generation siblings for patterns of phenotypic inheritance

As this method enables identification of clonal progression and phenotypic expression, it allows for the direct measurement of asymmetric expression amongst sibling cells after the first division following stimulation (Fig. 5.9). Asymmetric Cell Division (ACD) is a key driver of cellular diversity during development (Knoblich, 2008) that has been implicated in mature stem cell systems (Morrison and Kimble, 2006) as well as the adaptive immune response (Chang et al., 2007; Barnett et al., 2012; Hawkins et al., 2013; Arsenio et al., 2014).

In order to determine if ACD has occurred, it is necessary to identify cells that are siblings and to measure properties of each. This is typically challenging as generating statistically meaningful numbers of sibling cell pairs is highly time-consuming by conventional methods, such as fixed-image microscopy or live filming, but is made much more attainable with the multiplex assay. On plotting the expression levels of siblings, one anticipates distinct patterns (Fig. 5.9 upper panels) dependent on whether the underlying biology was: ACD with identifiable sibling polarity, achievable by specific ligand-receptor labelling (Pasqual et al., 2018) or asymmetrically segregating endocytosed fluorescent beads (Thaunat et al., 2012), ACD with undetermined sibling polarity; if there were no inheritance; or if there were symmetric inheritance.

For illustration, we repeated the experimental setup described in Fig. 5.2 using plate-bound CD3 in the presence of anti-CD28 and rhIL-2, but harvested cells 42 hours post-stimulation to observe more clonal families after only one division event. 178 clonal families with two or more members were recovered, totalling 427 cells (see Fig. 5.10). Clonal and environmental statistical analysis, analogous as from Fig. 5.6 and 5.8, is also reported in Fig. 5.11, leading to similar outcome. Of these data, 96 clones consisted of two sibling-cells in generation one allowing us to examine their expression relationships. Plots for each of CD8, CD25 and CD62L are provided in Fig. 5.9 lower panels and are redolent of Fig. 5.9 upper right panel, indicating highly symmetric divisions for this system under these stimulation conditions.

5.7 Discussion

The clonal basis of T-cell activation and subsequent emergence of phenotypic heterogeneity is an important focus in furthering our understanding of lymphocyte biology (Rohr et al., 2014; Buchholz et al., 2016; Polonsky et al., 2016). Using example data

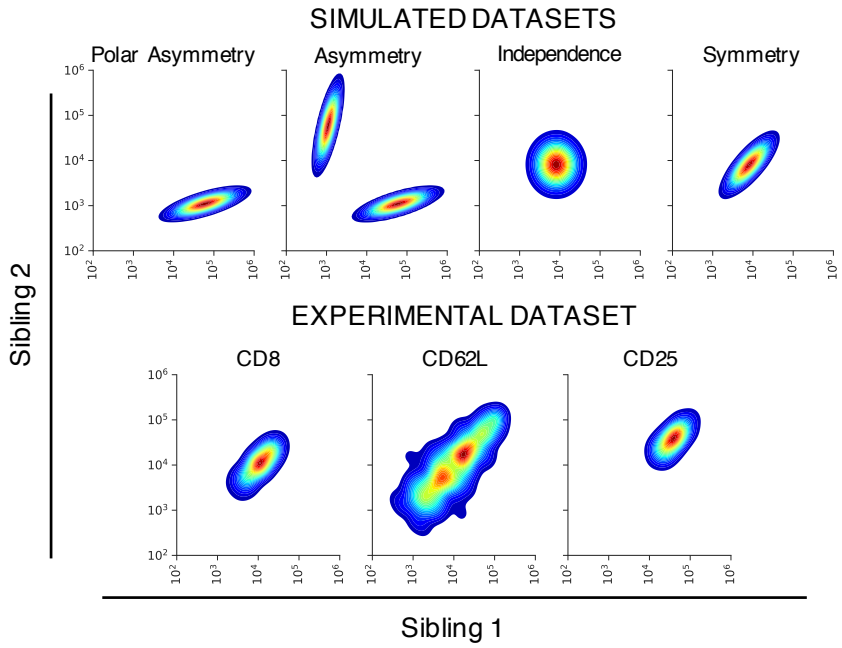

FIGURE 5.9: **First division siblings reveal symmetrical pattern of inheritance for marker expression.** [Corresponding to Figure 4 from Horton, Prevedello et al., (2018)] (Upper panels) Asymmetric versus symmetric cell division sketch. Expression levels of a given phenotypic marker across sibling cells assuming asymmetry with measurable polarity, asymmetry without measurable polarity, independence and symmetric inheritance (from left to right). (Lower panels) Experimental data as in Fig. 5.7, from the same system as in Fig. 5.2 but harvested at 42h, showing expression levels (fluorescence intensities) of CD8, CD25 and CD62L (from left to right) for 96 first generation siblings, resembling symmetric inheritance (right upper panel).

sets we have demonstrated the utility of combining multiplexed division tracking dyes with single cell sorting and conventional flow cytometry-based phenotyping to analyse the clonal lineage properties of $CD8^+$ T cells with simplicity and high-throughput.

Using this method, we observed a striking and significant concordance in marker expression amongst the progeny of single T-cell clones after standard *in vitro* culture. These data imply that activated founder $CD8^+$ T cells have the potential to pass on a heritable, phenotype-determining program to their progeny through multiple rounds of cell division. The nature of this program, and how it is preserved to such a precise degree through numerous repetitions of the cell cycle, is unknown. The relative contribution of shared heritable fate determinants, as seen here *in vitro*, and the imposition of lineage branching points by, for example asymmetric cell division, or a chance encounter with a cytokine, will require further experiments tracing cells during ongoing immune responses *in vivo*.

A key advantage of the method is the ability to undertake direct measurement of sibling phenotype generated after the first division following stimulation. As the system

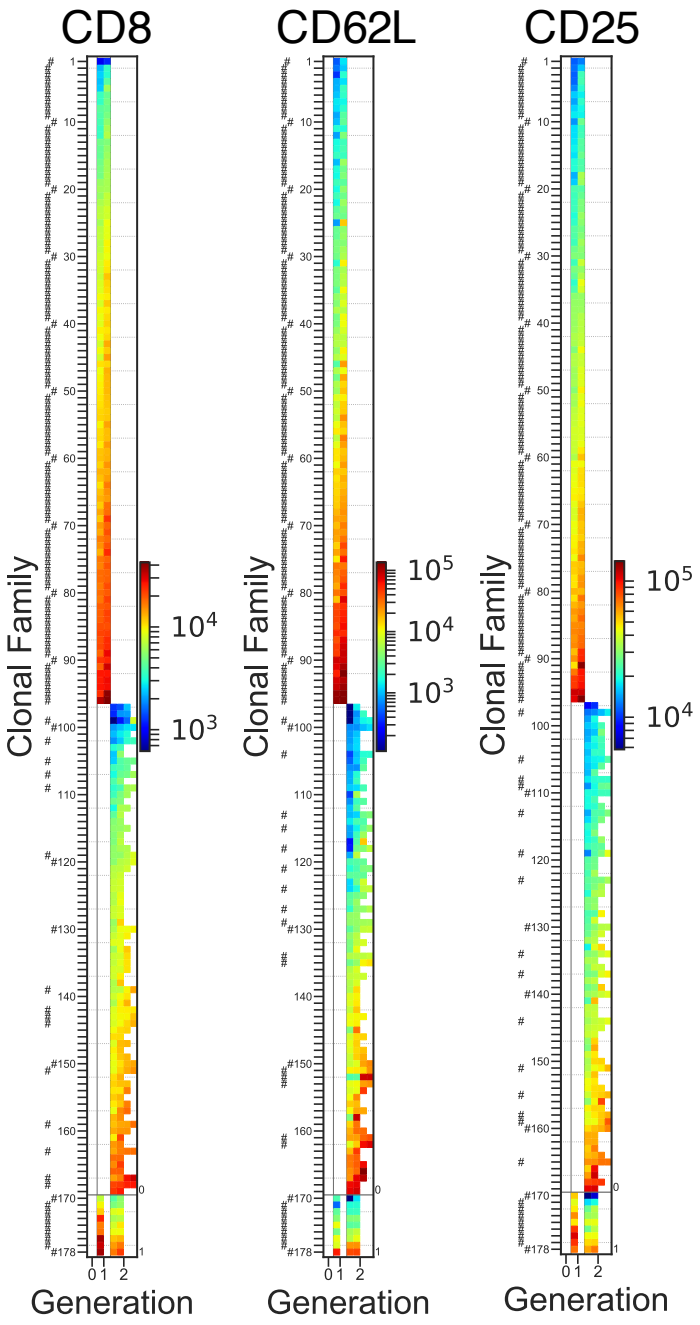

FIGURE 5.10: **Multiplexed tracking dyes to determine clonal membership, generation number and cell phenotype can identify clones as early as the first division.** [Corresponding to Supplemental Figure 6 from Horton, Prevedello et al., (2018)] Purified murine $CD8^+$ T cells were processed analogously as in Fig. 5.2, and multiplex dye labelled with CFSE, CTV and CPD, resulting in 10 profiles. These cells were stimulated with anti-CD3 (10 μg ml^{-1}), anti-CD28 (2 μg ml^{-1}) and rhIL-2 (31.6 U ml^{-1}) for 24h in the presence of the anti-mouse IL-2 blocking antibody clone S4B6 (25 μg ml^{-1}). Single cells from each of the 10 combinations were sorted and pooled into 30 individual wells followed by culture for a further 18h. Generation number and fluorescence intensity of CD8 (APC-Cy7), CD62L (PE) and CD25 (PE-Cy7) expression were determined by flow cytometry 42h post-stimulation. Of 300 clones initially seeded, 178 families were detected with at least 2 members, yielding a recovery of 59.3%.

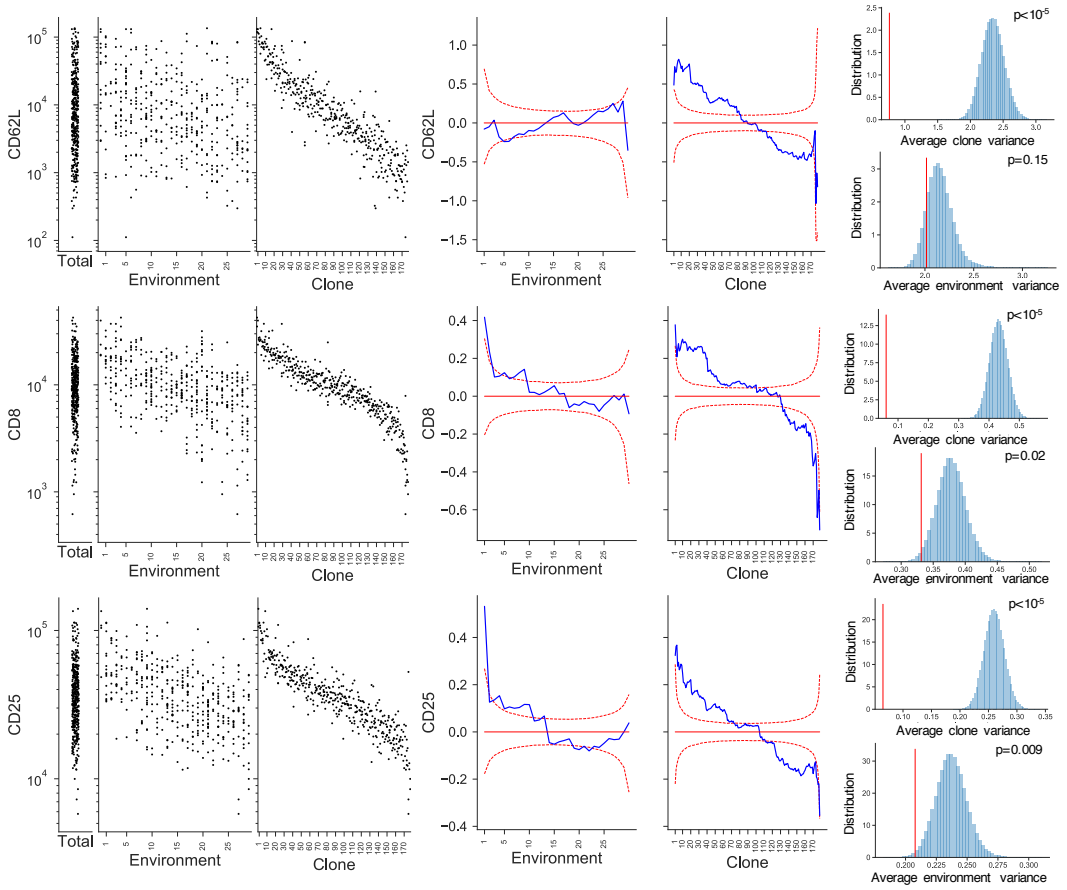

FIGURE 5.11: **Visualization and testing of environment and clonal independence.** [Corresponding to Supplemental Figure 7 from Horton, Prevedello et al., (2018)] Analogous analysis as in Fig. 5.6f-h and 5.8 for the data presented in Fig. 5.10.

can identify siblings in the presence of other accessory cells it will be possible to systematically investigate how manipulation of the activation conditions affects the fate of each sibling in a pair. For example, it has been suggested that the synapse that forms between a dendritic cell and a T cell provides polarity cues for an asymmetric division and that this cue is further enhanced by the affinity of interaction (Plumlee et al., 2013; Rohr et al., 2014; Polonsky et al., 2016). The method is well suited to systematically measure how such culture and stimulation variables affect concordance and fate in first generation siblings and later generation relatives.

Existing lineage tracing technologies have contributed significantly to the understanding of the clonal basis of many biological processes. Those methods, however, have a number of caveats that leave important aspects of biological systems unmeasured. Measuring division progression, as enabled by the approach described here, ameliorates some of these shortcomings and allows the development of the customised statistical methodology presented here alongside the clonal data. These tools provide prospective users with a robust means of assessing the relative impact of clonal lineage and

environmental influence on cell fate selection.

Perhaps the most significant advantage of this method is its ease of implementation. By making use of affordable, commercially available reagents and widely accessibly technology, any researcher with access to flow cytometry services can easily apply this method to study clonal dynamics in their system of interest. Therefore, while we have illustrated the method here for *in vitro* T cell systems, we believe it will find wide use including application to *in vivo* cell tracing systems, although this is not yet validated. This method does not require genetic manipulation, cell infection, or breeding of fluorescent or congenic reporter systems. It can identify lineages of adherent cells *in vivo*, or *in vitro* within complex cultures that include additional cell types, provided they are labelled and/or identified using compatible cell-specific markers. Consequently, it is broadly applicable and well suited to address questions of expansion and differentiation at the level of clones.

Appendix A

Experimental systems implemented by collaborators

In this appendix, we provide information concerning the experimental design, as performed by our collaborators, for completeness. In Section A.1, we present the additional methods required for the multiplex assay from Chapter 2 as implemented by J. M. Marchingo in Marchingo, Prevedello et al., (2016). In Section A.2, we report the technical details for the experiments of Chapter 5 from Horton, Prevedello et al., (2018), performed by M. B. Horton.

A.1 Marchingo, Prevedello et al., (2016)

A.1.1 Mice

OT-I/*Bcl2l11*$^{-/-}$ and OT-I/FucciRG mice (Marchingo et al., 2014) were bred and maintained under specific pathogen-free conditions in the WEHI animal facilities (Parkville, Victoria, Australia) and used between 6-10 weeks of age. OT-I/FucciRG mice were bred from the red (R) FucciG1-#639 and green (G) FucciS/G2/M-#492 mouse lines. All experiments were performed under the approval of the WEHI Animal Ethics Committee.

A.1.2 CD8$^+$ T-cell purification

CD8$^+$ T cells were isolated from mouse lymph nodes and spleens by negative selection using EasySep Mouse CD8$^+$ T-cell Isolation kit (StemCell Technologies) according to

the manufacturer's protocols. Enrichment of OT-I $CD8^+$ T-cells was confirmed by flow cytometry with a yield of 90-95% $CD8^+V\alpha2^+$ lymphocytes.

A.1.3 Labelling with division tracking dyes

OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were labelled with the indicated combinations and concentrations of the division tracking dyes CTV, CFSE (both Invitrogen) and CPD (eBioscience) in PBS (WEHI media) containing 0.1% BSA (Sigma) (PBS 0.1% BSA) at a density of $\leq 10^7$ cells ml^{-1} at 37 °C for 20, 10 and 10 min, respectively. The reaction was quenched by washing with 2 ml ice-cold RPMI 5% FCS.

A.1.4 *In vitro* cell culture

Complete tissue culture medium was RPMI 1640 medium supplemented with 10% FCS, non-essential amino acids, 1 mM Sodium-pyruvate, 10 mM HEPES, 2 mM GlutaMAX, 100 U ml^{-1} Penicillin, 100 μg ml^{-1} Streptomycin (all Invitrogen) and 50 μM 2β-mercaptoethanol (Sigma). OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were stimulated with 0.01 μg ml^{-1} SIINFEKL (N4) peptide (Auspep) in 96 well round-bottomed plates by self-presentation at a density of 10,000 cells per well in 200 μl complete tissue culture medium, as described previously (Marchingo et al., 2014).

All cultures contained 25 μg ml^{-1} of anti-mouse IL-2 monoclonal antibody (supernatant from hybridoma cell line S4B6, WEHI monoclonal antibody facility) that blocks the activity of mouse IL-2 *in vitro* but does not recognize human IL-2 (hIL-2) (Marchingo et al., 2014). Recombinant hIL-2 (Peprotech) and anti-CD28 (clone 37.51, WEHI monoclonal antibody facility) were added to cultures where indicated. Cells were incubated in a humidified environment at 37 °C in 5% CO_2.

A.1.5 Cell sorting and flow cytometry

Cell sorting was performed on a FACSAria W or L (BD Biosciences) cell sorter. For IL-2Rα and CD28 level sorting, cells were labelled with anti-CD28-PECy7 (clone 37.51, eBioscience) or anti-CD25-FITC (clone 7D4, BD). Flow cytometry was performed on a FACSCanto II or LSRFortessa X-20 cytometer (both BD Biosciences). Data were analysed using FlowJo software (Treestar). A known number of beads (Rainbow calibration particles, BD Biosciences) and propidium iodide (0.2 μg ml^{-1}, Sigma) was added to samples immediately prior to analysis. The ratio of beads to live cells was used to estimate the absolute cell number.

The following monoclonal antibodies were used for the detection of cell surface markers: anti-CD25-PECy7, or -APC (clone PC61, BD Biosciences) anti-CD28-PECy7 (clone 37.51, eBioscience). Staining was performed in PBS containing 0.1% BSA and 0.1% sodium azide (Sigma). In Fig. 2.12, activated cells were defined as the 50% of cells with the highest FSC fluorescence. Spearman's correlation was calculated using Matlab 2011a's "corr" function.

A.1.6 High-throughput clonal multiplex assay to measure DD

Naive OT-I/$Bcl2l11^{-/-}$ $CD8^+$ T cells were purified and sequentially labelled with CFSE (5, 2.5, 0 μM), CTV (5, 2.5, 0 μM) and CDP (5, 0 μM) (Fig. 2.2a,b). After the population labelling controls were plated, cells from the 10 labelling combinations indicated in Fig. 2.2c were pooled together and 10,000 cells were added per well for stimulation with N4 peptide in the presence of S4B6, either with or without anti-CD28 (2 μg ml^{-1}, Fig. 2.2c). After 26 h (just prior to the first division), cells were sorted so that a single stimulated (estimated based upon high FSC fluorescence) but undivided cell from each fluorescently distinct population was sorted into each "sample" well of 96-well round-bottomed plates (Fig. 2.2d). Cells were cultured in the presence of S4B6 either with or without hIL-2 (1 U ml^{-1}). Four different stimulation combinations were monitored: N4, N4 + anti-CD28, N4 + IL-2 and N4 + anti-CD28 + IL-2. Cells were collected for analysis by flow cytometry at 54, 62 and 72 h post stimulation. At each analysis time point, 7,500 beads were added to measure sample recovery (>90% of the sample for >90% of the tubes in the experiment shown), and PI (0.2 μg ml^{-1}) for dead cell exclusion. Samples were carefully transferred into 5 ml polystyrene tubes and entire sample was analysed (Fig. 2.2e).

For data analysis, gates were set using single label configuration population controls then applied to clonal data as shown in Fig. 2.2f. Briefly, lymphocytes were identified from FSC/side scatter (SSC) profiles and dead cells excluded using PI. Cells were divided into CPD^- and CPD^+ then division gating for each labelling configuration was performed on CFSE versus CTV dot plots. Finally cells were gated as "small" or "not small" from FSC/SSC profiles to classify cells as quiescent or dividing respectively. Small cell size has previously been demonstrated to be a good surrogate of lymphocyte quiescence (Hawkins et al., 2009; Marchingo et al., 2014; Kinjyo et al., 2015). We further demonstrated this with independent experiments using OT-I/Fucci cell cycle reporter mice, in which cells accumulate the FucciRed reporter when they have reverted to a quiescent state, or express the FucciGreen reporter when progressing through the S/G2/M phases of the cell cycle (Sakaue-Sawano et al., 2008; Tomura et al., 2013; Marchingo et al., 2014; Dowling et al., 2014). OT-I/FucciRG $CD8^+$ T

cells were stimulated in similar conditions to those used in the clonal studies and cell size and Fucci reporter fluorescence measured across several time points where the cells were reaching DD to estimate the accuracy of small cell gates to classify cellular quiescence (Fig. A.1). Forward-scatter side scatter profiles were used to set small cell gates (Fig. A.1, left columns) then Fucci fluorescence used to gauge the proportion of incorrectly classified cells (that is, FucciG$^+$ cells that fell within the "small" gate and FucciR$^+$ quiescent cells that fell within the "large" gate). In this example, 3.3%, 2.9% and 4.7% of cells were incorrectly classified by size-based gating at the 50.5, 66 and 73 h time-points, respectively (Fig. A.1). Extrapolating these error rates to the data shown in Fig. 2.3 we can estimate that the quiescence status of 165 of the 171 clones in this example has been correctly classified. Collectively, along with previous findings these results indicate that small cell size is an accurate method to estimate cellular quiescence in these studies.

A.1.7 Population DD measurements

The definition and methods by which DD can be estimated during a population response have been published previously (Marchingo et al., 2014). Briefly, the cohort number (an estimate of the number of starting cells whose progeny are contributing to the response at a time point) was calculated by dividing the cell number per division by 2^i, where i is the cell's generation. The population mean division number (MDN) was calculated as the arithmetic mean of the cohort numbers at each time point. Assuming little death, the MDN will increase in time, plateauing at the point where the cells reach DD. Thus, the population total expansion was estimated as the maximum MDN measured over all the time points.

A.1.8 Estimating clonal contributions to *in vivo* population DD

To calculate the percentage contribution of clonal families to the magnitude of the total response (Fig. 2.15), it was assumed that all progeny cells would adopt a concordant DD. The probability of a clone reaching DD in a given division was obtained from Cyton fitting to the OT-I/FucciRG CD8$^+$ T cell *in vivo* influenza (HKx31-OVA) infection data in Fig. 2.1d,e and Supplementary Table 1 from Marchingo et al. (2014). The discretised probability function was multiplied by the mean initial cell number for the two experiments ($N_0 = 1,808$, Supplementary Table 1 from Marchingo et al. (2014)) and only divisions that contained at least one clone were used to determine clonal family contribution to response magnitude (that is, divisions from 4 to 19). The probabilities in each division were normalized so that the discretised probability distribution (f_i) for

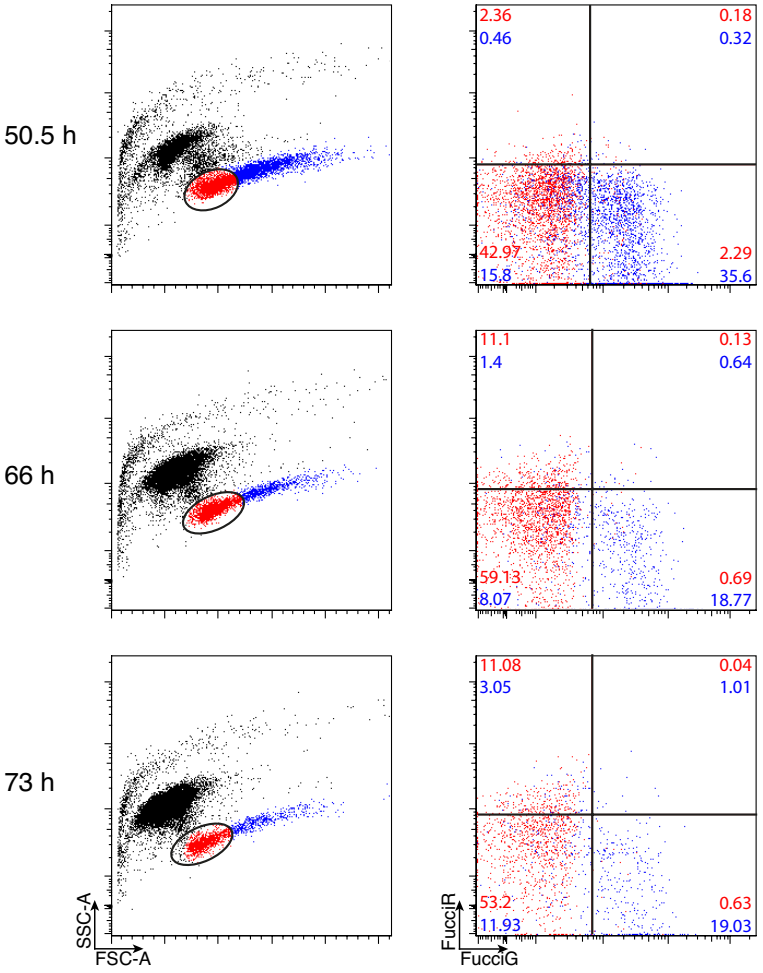

FIGURE A.1: **Small cell size is an accurate estimator of quiescence.** [Corresponding to Supplementary Figure 2 from Marchingo, Prevedello et al., (2016)] CTV labelled OT-I/FucciR$^+$G$^+$ CD8$^+$ T cells stimulated with N4 peptide (0.01 μg ml^{-1}) and cultured with S4B6 (25 μg ml^{-1}) and hIL-2 (1 U ml^{-1}). FSC-A vs SSC-A profiles were used to visually determine "small" cell (red) and "large" cell (blue) gates (left column, black dots show dead cells and debris). Subsequently FucciRed vs. FucciGreen fluorescence was used to assess the frequency with which gating upon cell size incorrectly classifies FucciG$^+$ as small and FucciR$^+$ cells as large (right column, numbers are the average percentage of the live cell population from technical replicates). Representative of triplicate culture wells from 2 independent experiments.

$i \in \{4, \ldots, 19\}$ summed to 1. The number of progeny cells produced by clones reaching DD in division i was corrected to reintroduce the effects of cell expansion as follows:

$$N_i^{\text{qui}} = N_0 f_i 2^i. \tag{A.1}$$

The percentage contribution to the total response magnitude of progeny cells reaching DD in each division $i \in [4, 19]$ was calculated as follows:

$$100 \frac{N_i^{\text{qui}}}{\sum_{j=4}^{19} N_j^{\text{qui}}}. \tag{A.2}$$

This was then plotted as a cumulative function against the percentage of the total clones that generated these cells.

A.1.9 Inference of DD distribution from *in vivo* clonal studies

The percentage contribution of clonal families to the magnitude of the total response was obtained directly from Buchholz et al. (2013). To estimate the DD distribution for this *in vivo* clonal data, we assumed that *in vivo* DD was concordant and that minimal cell death had occurred at the time point measured. We estimated the DD as $\log_2(N)$, where N is the total number of progeny cells detected per clone. Clonal DD was rounded up to the next integer value and binned for every second division.

A.2 Horton, Prevedello et al., (2018)

A.2.1 Mice

The three murine strains, wild-type C57BL/6, ovalbumin specific OT-I/$Bcl2l11^{-/-}$ Marchingo, Prevedello et al., 2016, and $\text{Blimp}^{\text{gfp/+}}$ (Kallies et al., 2004) mice were maintained under specific pathogen-free conditions in the Walter and Eliza Hall Institute (WEHI) animal facilities and were used at 6-10 weeks of age. All experiments were performed under the approval of the WEHI Animal Ethics Committee.

A.2.2 $CD8^+$ T-cell purification

Spleens and lymph nodes were homogenised through a 70 μM cell strainer to generate single cell suspensions. $CD8^+$ T cells were isolated by negative selection using EasySep Mouse $CD8^+$ T cell Isolation Kit according to manufacturer protocol (StemCell Technologies).

A.2.3 Sequential labelling protocol using CFSE, CTV and CPD

CFSE label. Purified CD8$^+$ T cells were resuspended in sterile phosphate buffered saline containing 0.1% bovine serum albumin (PBS 0.1% BSA) and labelled with either 5 μM, 2.5 μM or 0 μM CFSE (Invitrogen) at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 10 minutes and washed twice with 10 ml ice-cold RPMI-1640 10% FCS.

CTV label. Cells were resuspended in PBS 0.1% BSA and those labelled with 5 μM CFSE were further labelled with either 5 μM, 2.5 μM or 0 μM CTV (Invitrogen). Cells labelled with 2.5 μM CFSE were labelled with 5 μM CTV. Cells labelled with 0 μM CFSE were labelled with either 5 μM or 0 μM CTV. All labelling performed at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 20 minutes and all cells were washed twice with 10 ml ice-cold RPMI-1640 10% FCS.

CPD label. Cells were resuspended in PBS 0.1% BSA and labelled with either 5 μM or 0 μM CPD eFluor670 (eBioscience) at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 10 minutes and washed once with 10 ml ice-cold RPMI-1640 10% FCS and once with tissue culture medium.

A.2.4 Sequential labelling protocol using CTY, CTV and CPD

CTY label. Purified CD8$^+$ T cells were resuspended in PBS 0.1% BSA and labelled with either 10 μM or 0 μM CTY (Invitrogen) at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 20 minutes and washed twice with 10 ml ice-cold RPMI-1640 10% FCS.

CTV label. Cells were resuspended in PBS 0.1% BSA and labelled with either 5 μM or 0 μM CTV at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 20 minutes and washed twice with 10 ml ice-cold RPMI-1640 10% FCS.

CPD label. Cells were resuspended in PBS 0.1% BSA and labelled with either 5 μM or 0 μM CPD at a density of $\leq 2\times107$ cells ml^{-1} at 37 °C for 10 minutes and washed once with 10 ml ice-cold RPMI-1640 10% FCS and once with tissue culture medium.

A.2.5 *In vitro* cell culture

Tissue culture medium was RPMI-1640 with 10% FCS, 1 mM sodium-pyruvate, 2 mM GlutaMAX, 10 mM HEPES, 100 U ml^{-1} Penicillin, 100 μg ml^{-1} Streptomycin (all Invitrogen) and 50 μM 2β-mercaptoethanol (Sigma). Purified C57BL/6 CD8$^+$ T cells were stimulated with 10 μg ml^{-1} plate-bound anti-CD3 monoclonal antibody in flat-bottomed 96 well plates (WEHI monoclonal antibody facility, clone 145-2C11). For

some experiments OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were stimulated with 0.01 μg ml^{-1} SIINFEKL (N4) peptide (Auspep). The use of Bim-deficient T cells enhances survival *in vitro* but does not affect proliferative or phenotypic behaviours (Prlic and Bevan, 2008; Marchingo et al., 2014). These cells were stimulated in round-bottomed 96 well plates at a density of 20,000 cells per well. This protocol leads to the self-presentation of peptide by T cells and is used as a minimal culture system to enable the addition of further costimulatory signals (Denton et al., 2011; Marchingo et al., 2014; Marchingo, Prevedello et al., 2016; Heinzel et al., 2017).

Cells were cultured in 200 μL of tissue culture medium in the presence of 25 μg ml^{-1} anti-mouse IL-2 monoclonal antibody (WEHI monoclonal antibody facility, clone S4B6) that inhibits the activity of mouse IL-2 but does not act on rhIL-2 (Deenick et al., 2003). RhIL-2 (Peprotech), anti-CD28 (WEHI monoclonal antibody facility, clone 37.51), mouse IL-4 (purified from baculovirus transfected Sf21 insect cells) and mouse IL-12 (Miltenyi Biotec, 130-096-707) were added to cultures where indicated. Cells were incubated at 37 °C in 5% CO_2.

A.2.6 Stimulation and sorting

Purified C57BL/6 and OT-I/*Bcl2l11*$^{-/-}$ CD8^{+} T cells were sequentially labelled with CFSE, CTV and CPD. The uniquely labelled cell populations were mixed (except for the unlabelled and CPD-only labelled controls) and stimulated under conditions indicated. After 22-26 hours, prior to the first division, cells from across multiple wells stimulated under the same conditions were pooled and sorted according to their distinct fluorescence profiles into new wells such that each well contained a single cell from each unique labelling profile, with the exception of cells labelled with only 5 μM CPD or unlabelled cells. Wells contained the same conditions under which the cells were initially stimulated. Bulk population controls were also sorted into new wells, with 1,000 cells from each labelling profile sorted into separate wells, in addition to 100 cells from each population sorted into the same well. This gave bulk populations of each fluorescence profile both separately and in combination. This included cells labelled with 5 μM CPD only and unlabelled cells.

Purified Blimp$^{gfp/+}$ CD8^{+} T cells were sequentially labelled with CTY, CTV and CPD and stimulated with plate-bound anti-CD3 (10 μg ml^{-1}), rhIL-2 (31.6 U ml^{-1}) and mIL-12 (10 ng ml^{-1}) in the presence of S4B6 (25 μg ml^{-1}) and were subsequently cultured and sorted according to the same criteria as above. Sorting was performed on either a BD Biosciences FACSAria III or a BD Biosciences Influx.

A.2.7 Antibody staining, flow cytometry and analysis

At time points indicated cells were stained on ice with indicated antibodies used at the following concentrations; 1:400 dilution anti-CD8-APCCy7 (BD Biosciences clone 53-6.7), 1:1600 dilution anti-CD62L-PE (BD Biosciences clone MEL-14), 1:1600 anti-CD62L-APCCy7 (BD Biosciences clone MEL-14), 1:400 dilution anti-CD25-PECy7 (BD Biosciences clone PC61) and 1:3200 dilution anti-CXCR3-PECy7 (eBioscience). 104 beads (Rainbow calibration particles BD Biosciences) and 0.2 μg ml^{-1} propidium iodide (PI, Sigma) was also added to samples prior to analysis. An antibody staining mix containing all relevant antibodies along with beads and PI was prepared for each experiment. Antibody staining mix was added at staggered time points ($\sim$2-3 minutes apart) to each sample in the 96-well culture plates and later transferred to 5 ml polystyrene tubes such that each sample was stained for as close to 30 minutes as possible prior to immediate acquisition of as much of the sample as possible (duration of acquisition lasted $\sim$2-3 minutes per sample).

Analysis was performed on a BD Biosciences LSRFortessa-X20. Gates were set using labelled bulk population controls and these were then applied to clonal data. Live lymphocytes were identified using forward and side scatter and PI exclusion. Cells were separated into CPD+ and CPD- populations and division gates were identified for each labelling profile on CFSE versus CTV for C57BL/6 and OT-I/$Bcl2l11^{-/-}$, or CTY versus CTV for Blimp$^{gfp/+}$. Clonal families were identified and the division numbers and expression levels of surface markers or Blimp$^{gfp/+}$ of each cell was enumerated and exported for data visualization and further analysis.

Bibliography

D. W. K. Andrews. Asymptotic results for generalized Wald tests. *Economic Theory*, 3(3):348–358, 1987.

D. W. K. Andrews. Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics*, 37(1):135–156, 1988.

J. Arsenio, B. Kakaradov, P. J. Metz, S. H. Kim, G. W. Yeo, and J. T. Chang. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. *Nature Immunology*, 15(4):365–372, 2014.

A. Banerjee, S. M. Gordon, A. M. Intlekofer, M. A. Paley, E. C. Mooney, T. Lindsten, E. J. Wherry, and S. L. Reiner. Cutting edge: The transcription factor eomesodermin enables cd8+ t cells to compete for the memory cell niche. *The Journal of Immunology*, 185(9):4988–4992, 2010.

B. E. Barnett, M. L. Ciocca, R. Goenka, L. G. Barnett, J. Wu, T. M. Laufer, J. K. Burkhardt, M. P. Cancro, and S. L. Reiner. Asymmetric B cell division in the germinal center reaction. *Science*, 335(6066):342–344, 2012.

A. G. Baxter and P. D. Hodgkin. Activation rules: the two-signal theories of immune activation. *Nature Reviews Immunology*, 2(6):439, 2002.

R. Beran. Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 86:14–30, 1984.

J. M. Bernitz, H. S. Kim, B. MacArthur, H. Sieburg, and K. Moore. Hematopoietic stem cells count and remember self-renewal divisions. *Cell*, 167(5):1296–1309, 2016.

E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider. An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6):463–471, 2013.

J. J. Bird, D. R. Brown, A. C. Mullen, N. H. Moskowitz, M. A. Mahowald, J. R. Sider, T. F. Gajewski, C.-R. Wang, and S. L. Reiner. Helper T cell differentiation is controlled by the cell cycle. *Immunity*, 9(2):229–237, 1998.

V. R. Buchholz, T. N. Schumacher, and D. H. Busch. T cell fate at the single-cell level. *Annual Review of Immunology*, 34:65–92, 2016.

V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer, and D. H. Busch. Disparate individual fates compose robust CD8+ T cell immunity. *Science*, 340(6132):630–635, 2013.

F. M. Burnet. *The clonal selection theory of acquired immunity.* Cambridge University Press, Cambridge, England, 1959.

F. M. Burnet. A modification of Jerne's theory of antibody production using the concept of clonal selection. *The Australian Journal of Science*, 20(3):67–69, 1957.

E. A. Butz and M. J. Bevan. Massive expansion of antigen-specific CD8+ T cells during an acute virus infection. *Immunity*, 8(2):167–175, 1998.

R. Chakravorty, D. Rawlinson, A. Zhang, J. Markham, M. R. Dowling, C. Wellard, J. H. S. Zhou, and P. D. Hodgkin. Labour-efficient in vitro lymphocyte population tracking and fate prediction using automation and manual review. *PloS ONE*, 9(1): e83251, 2014.

J. T. Chang, V. R. Palanivel, I. Kinjyo, F. Schambach, A. M. Intlekofer, A. Banerjee, S. A. Longworth, K. E. Vinup, P. Mrass, J. Oliaro, N. Killeen, J. S. Orange, S. M. Russell, W. Weninger, and S. L. Reiner. Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science*, 315(5819):1687–1691, 2007.

J. T. Chang, M. L. Ciocca, I. Kinjyo, V. R. Palanivel, C. E. McClurkin, C. S. DeJong, E. C. Mooney, J. S. Kim, N. C. Steinel, J. Oliaro, C. C. Yin, B. I. Florea, H. S. Overkleeft, L. J. Berg, S. M. Russell, G. A. Koretzky, M. S. Jordan, and S. L. Reiner. Asymmetric proteasome segregation as a mechanism for unequal partitioning of the transcription factor T-bet during T lymphocyte division. *Immunity*, 34(4):492–504, 2011.

L. Chen and D. B. Flies. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nature Reviews Immunology*, 13(4):227–242, 2013.

M. D. Cooper, R. D. Peterson, and R. A. Good. Delineation of the thymic and bursal lymphoid systems in the chicken. *Nature*, 205(4967):143, 1965.

N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 46(3):440–464, 1984.

J. M. Curtsinger and M. F. Mescher. Inflammatory cytokines as a third signal for T cell activation. *Current Opinion in Immunology*, 22(3):333–340, 2010.

M. M. Davis and P. J. Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402, 1988.

R. J. De Boer, D. Homann, and A. S. Perelson. Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *The Journal of Immunology*, 171(8):3928–3935, 2003.

E. K. Deenick, A. V. Gett, and P. D. Hodgkin. Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. *The Journal of Immunology*, 170(10):4963–4972, 2003.

A. E. Denton, R. Wesselingh, S. Gras, C. Guillonneau, M. R. Olson, J. D. Mintern, W. Zeng, D. C. Jackson, J. Rossjohn, P. D. Hodgkin, P. C. Doherty, and S. J. Turner. Affinity thresholds for naive CD8+ CTL activation by peptides and engineered influenza A viruses. *The Journal of Immunology*, 187(11):5733–5744, 2011.

M. R. Dowling, A. Kan, S. Heinzel, J. H. S. Zhou, J. M. Marchingo, C. J. Wellard, J. F. Markham, and P. D. Hodgkin. Stretched cell cycle model for proliferating lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6377–6382, 2014.

F. C. Drost. Generalized chi-square goodness-of-fit tests for location-scale models when the number of classes tends to infinity. *Annals of Statistics*, 17(3):1285–1300, 1989.

K. R. Duffy and P. D. Hodgkin. Intracellular competition for fates in the immune system. *Trends in Cell Biology*, 22(9):457–464, 2012.

K. R. Duffy, C. J. Wellard, J. F. Markham, J. H. S. Zhou, R. Holmberg, E. D. Hawkins, J. Hasbold, M. R. Dowling, and P. D. Hodgkin. Activation-induced B cell fates are selected by intracellular stochastic competition. *Science*, 335(6066):338–341, 2012.

K. R. Duffy and V. G. Subramanian. On the impact of correlation between collaterally consanguineous cells on lymphocyte population dynamics. *Journal of Mathematical Biology*, 59(2):255–285, 2009.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Editorial. Rationalizing combination therapies. *Nature Medicine*, 23:1113, 10 2017.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, USA, 1st edition, 1993.

O. Feinerman, G. Jentsch, K. E. Tkach, J. W. Coward, M. M. Hathorn, M. W. Sneddon, T. Emonet, K. A. Smith, and G. Altan-Bonnet. Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. *Molecular Systems Biology*, 6(1):437, 2010.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, UK, 1st edition, 2009.

C. Gerlach, J. C. Rohr, L. Perie, N. van Rooij, J. W. van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, R. J. de Boer, and T. N. Schumacher. Heterogeneous differentiation patterns of individual CD8+ T cells. *Science*, 340(6132): 635–639, 2013.

C. Gerlach, J. W. Van Heijst, E. Swart, D. Sie, N. Armstrong, R. M. Kerkhoven, D. Zehn, M. J. Bevan, K. Schepers, and T. N. Schumacher. One naive t cell, multiple fates in cd8+ t cell differentiation. *Journal of Experimental Medicine*, 207(6):1235–1246, 2010.

C. Gerlach, J. W. van Heijst, and T. N. Schumacher. The descent of memory T cells. *Annals of the New York Academy of Sciences*, 1217(1):139–153, 2011.

A. V. Gett and P. D. Hodgkin. Cell division regulates the T cell cytokine repertoire, revealing a mechanism underlying immune class regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16):9488–9493, 1998.

A. V. Gett and P. D. Hodgkin. A cellular calculus for signal integration by T cells. *Nature Immunology*, 1(3):239–244, 2000.

A. D. Gitlin and M. C. Nussenzweig. Immunology: fifty years of B lymphocytes. *Nature*, 517(7533):139–141, 2015.

R. A. Gottschalk, M. M. Hathorn, H. Beuneu, E. Corse, M. L. Dustin, G. Altan-Bonnet, and J. P. Allison. Distinct influences of peptide-MHC quality and quantity on *in vivo* T-cell responses. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3):881–886, 2012.

J. R. Groom and A. D. Luster. CXCR3 in T cell function. *Experimental Cell Research*, 317(5):620–631, 2011.

A. S. Hadi and M. T. Wells. A note on generalized Wald's method. *Metrika*, 37(1): 309–315, 1990.

R. Hagen, S. Roch, and B. Silbermann. *C* - Algebras and Numerical Analysis*. CRC Press, New York, USA, 2000.

T. E. Harris. *The Theory of Branching Processes*. Springer, Berlin, Germany, 1st edition, 1964.

J. Hasbold, L. M. Corcoran, D. M. Tarlinton, S. G. Tangye, and P. D. Hodgkin. Evidence from the generation of immunoglobulin G-secreting cells that stochastic mechanisms regulate lymphocyte differentiation. *Nature Immunology*, 5(1):55, 2004.

E. D. Hawkins, M. L. Turner, M. R. Dowling, C. van Gend, and P. D. Hodgkin. A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):5032–5037, 2007a.

E. D. Hawkins, J. F. Markham, L. P. McGuinness, and P. D. Hodgkin. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proceedings of the National Academy of Sciences of the United States of America*, 106(32):13457–13462, 2009.

E. D. Hawkins, M. Hommel, M. L. Turner, F. L. Battye, J. F. Markham, and P. D. Hodgkin. Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nature Protocols*, 2(9):2057–2067, 2007b.

E. D. Hawkins, J. Oliaro, A. Kallies, G. T. Belz, A. Filby, T. Hogan, N. Haynes, K. M. Ramsbottom, V. Van Ham, T. Kinwell, B. Seddon, D. Davies, D. Tarlinton, A. M. Lew, P. O. Humbert, and S. M. Russell. Regulation of asymmetric cell division and polarity by Scribble is not required for humoral immunity. *Nature Communications*, 4:1801, 2013.

E. D. Hawkins, D. Duarte, O. Akinduro, R. A. Khorshed, D. Passaro, M. Nowicka, L. Straszkowski, M. K. Scott, S. Rothery, N. Ruivo, K. Foster, M. Waibel, R. W. Johnstone, S. J. Harrison, D. A. Westerman, H. Quach, J. Gribben, M. D. Robinson, L. E. Purton, D. Bonnet, and C. Lo Celso. T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments. *Nature*, 538(7626):518–522, 2016.

J. M. Heather, M. Ismail, T. Oakes, and B. Chain. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*, 1: 12, 2017.

S. Heinzel, T. B. Giang, A. Kan, J. M. Marchingo, B. K. Lye, L. M. Corcoran, and P. D. Hodgkin. A Myc-dependent division timer complements a cell-death timer to regulate T cell and B cell responses. *Nature Immunology*, 18(1):96, 2017.

S. Heinzel, J. M. Marchingo, M. B. Horton, and P. D. Hodgkin. The regulation of lymphocyte activation and proliferation. *Current Opinion in Immunology*, 51:32–38, 2018.

H. Hikono, J. E. Kohlmeier, S. Takamura, S. T. Wittmer, A. D. Roberts, and D. L. Woodland. Activation phenotype, rather than central- or effector-memory phenotype, predicts the recall efficacy of memory CD8+ T cells. *The Journal of Experimental Medicine*, 204(7):1625–1636, 2007.

P. D. Hodgkin. A probabilistic view of immunology: drawing parallels with physics. *Immunology and Cell Biology*, 85(4):295–299, 2007.

P. D. Hodgkin, J.-H. Lee, and A. B. Lyons. B cell differentiation and isotype switching is related to division cycle number. *The Journal of Experimental Medicine*, 184(1): 277–281, 1996.

P. D. Hodgkin, W. R. Heath, and A. G. Baxter. The clonal selection theory: 50 years since the revolution. *Nature Immunology*, 8(10):1019–1026, 2007.

P. D. Hodgkin, M. R. Dowling, and K. R. Duffy. Why the immune system takes its chances with randomness. *Nature Reviews Immunology*, 14(10):711–711, 2014.

K. Hogquist, S. Jameson, W. Heath, J. Howard, M. Bevan, and F. Carbone. T cell receptor antagonist peptides induce positive selection. *Cell*, 76(1):17–27, 1994.

M. Hommel and P. D. Hodgkin. TCR affinity promotes CD8+ T cell expansion by regulating survival. *The Journal of Immunology*, 179(4):2250–2260, 2007.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, USA, 1st edition, 1986.

M. B. Horton, G. Prevedello, J. M. Marchingo, J. H. S. Zhou, K. R. Duffy, S. Heinzel, and P. D. Hodgkin. Multiplexed division tracking dyes for proliferation-based clonal lineage tracing. *The Journal of Immunology*, 2018. Advance online publication. doi: 10.4049/jimmunol.1800481.

C. A. Janeway. Approaching the asymptote? Evolution and revolution in immunology. In *Cold Spring Harbor Yymposia on Quantitative Biology*, volume 54, pages 1–13. Cold Spring Harbor Laboratory Press, 1989.

M. K. Jenkins, H. H. Chu, J. B. McLachlan, and J. J. Moon. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. *Annual Review of Immunology*, 28:275–294, 2010.

M. R. Jenkins, J. Mintern, N. L. La Gruta, K. Kedzierska, P. C. Doherty, and S. J. Turner. Cell cycle-related acquisition of cytotoxic mediators defines the progressive differentiation to effector status for virus-specific CD8+ T cells. *The Journal of Immunology*, 181(6):3818–3822, 2008.

D. B. Johnson and J. A. Sosman. Therapeutic advances and treatment options in metastatic melanoma. *The Journal of the American Medical Association Oncology*, 1(3):380–386, 2015.

D. B. Johnson, C. Peng, and J. A. Sosman. Nivolumab in melanoma: latest evidence and clinical potential. *Therapeutic Advances in Medical Oncology*, 7(2):97–106, 2015.

C. H. June, S. R. Riddell, and T. N. Schumacher. Adoptive cellular therapy: a race to the finish line. *Science Translational Medicine*, 7:280ps7–280ps7, 2015.

S. M. Kaech and W. Cui. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nature Reviews Immunology*, 12(11):749–761, 2012.

S. M. Kaech, E. J. Wherry, and R. Ahmed. Effector and memory T-cell differentiation: implications for vaccine development. *Nature Reviews Immunology*, 2(4):251–262, 2002.

A. Kallies, J. Hasbold, D. M. Tarlinton, W. Dietrich, L. M. Corcoran, P. D. Hodgkin, and S. L. Nutt. Plasma cell ontogeny defined by quantitative changes in Blimp-1 expression. *The Journal of Experimental Medicine*, 200:967–77, 2004.

A. Kallies, A. Xin, G. T. Belz, and S. L. Nutt. Blimp-1 transcription factor is required for the differentiation of effector CD8+ T cells and memory responses. *Immunity*, 31(2):283–295, 2009.

A. Kan, R. Chakravorty, J. Bailey, C. Leckie, J. Markham, and M. R. Dowling. Automated and semi-automated cell tracking: addressing portability challenges. *Journal of Microscopy*, 244(2):194–213, 2011.

M. Kimmel and D. E. Axelrod. *Branching Processes in Biology*. Springer, New York, USA, 1st edition, 2002.

C. G. King, S. Koehli, B. Hausmann, M. Schmaler, D. Zehn, and E. Palmer. T cell affinity regulates asymmetric division, effector cell differentiation, and tissue pathology. *Immunity*, 37(4):709–720, 2012.

I. Kinjyo, J. Qin, S. Y. Tan, C. J. Wellard, P. Mrass, W. Ritchie, A. Doi, L. L. Cavanagh, M. Tomura, A. Sakaue-Sawano, O. Kanagawa, A. Miyawaki, P. D. Hodgkin, and W. Weninger. Real-time tracking of cell cycle progression during CD8+ effector and memory T-cell differentiation. *Nature Communications*, 6:6301, 2015.

G. G. Klaus, M. Holman, and J. Hasbold. Properties of mouse CD40: the role of homotypic adhesion in the activation of B cells via CD40. *European Journal of Immunology*, 24(11):2714–2719, 1994.

J. A. Knoblich. Mechanisms of asymmetric stem cell division. *Cell*, 132(4):583–597, 2008.

H. Y. Kueh, A. Champhekar, S. L. Nutt, M. B. Elowitz, and E. V. Rothenberg. Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. *Science*, 341(6146):670–673, 2013.

H. Y. Kueh, M. A. Yui, K. K. Ng, S. S. Pease, J. A. Zhang, S. S. Damle, G. Freedman, S. Siu, I. D. Bernstein, M. B. Elowitz, and E. V. Rothenberg. Asynchronous combinatorial action of four regulatory factors activates Bcl11b for T cell commitment. *Nature Immunology*, 17(8):956, 2016.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, USA, 3rd edition, 2005.

F. Lemaître, H. D. Moreau, L. Vedele, and P. Bousso. Phenotypic CD8+ T cell diversification occurs before, during, and after the first T cell division. *The Journal of Immunology*, 191(4):1578–1585, 2013.

J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56, 2007.

A. B. Lyons and C. R. Parish. Determination of lymphocyte division by flow cytometry. *Journal of Immunological Methods*, 171(1):131–137, 1994.

G. Lythe, R. E. Callard, R. L. Hoare, and C. Molina-París. How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224, 2016.

Y. D. Mahnke, T. M. Brodie, F. Sallusto, M. Roederer, and E. Lugli. The who's who of T-cell differentiation: human memory T-cell subsets. *European Journal of Immunology*, 43(11):2797–2809, 2013.

K. Man and A. Kallies. Synchronizing transcriptional control of T cell metabolism and function. *Nature Reviews Immunology*, 15(9):574, 2015.

W. C. Mankowski, G. Konica, M. R. Winter, F. Chen, C. Maus, R. Merkle, U. Klingmüller, T. Höfer, A. Kan, S. Heinzel, S. Oostindie, P. D. Hodgkin, and A. R. Cohen. Multi-modal segmentation for quantifying fluorescent cell cycle indicators throughout clonal development. In *Bioimage Informatics Conference*, 2015.

J. M. Marchingo, G. Prevedello, A. Kan, S. Heinzel, P. D. Hodgkin, and K. R. Duffy. T-cell stimuli independently sum to regulate an inherited clonal division fate. *Nature Communications*, 7:13540, 2016.

J. M. Marchingo, A. Kan, R. M. Sutherland, K. R. Duffy, C. J. Wellard, G. T. Belz, A. M. Lew, M. R. Dowling, S. Heinzel, and P. D. Hodgkin. Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science*, 346 (6213):1123–1127, 2014.

J. F. Markham, C. J. Wellard, E. D. Hawkins, K. R. Duffy, and P. D. Hodgkin. A minimum of two distinct heritable factors are required to explain correlation structures in proliferating lymphocytes. *Journal of The Royal Society Interface*, 7(48): 1049–1059, 2010.

I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, Cambridge, UK, 1st edition, 2012.

P. Matzinger. Tolerance, danger, and the extended family. *Annual Review of Immunology*, 12(1):991–1045, 1994.

A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, 2016.

M. F. Mescher, J. M. Curtsinger, P. Agarwal, K. A. Casey, M. Gerner, C. D. Hammerbeck, F. Popescu, and Z. Xiao. Signals required for programming effector and memory development by CD8+ T cells. *Immunological Reviews*, 211(1):81–92, 2006.

P. J. Metz, J. Arsenio, B. Kakaradov, S. H. Kim, K. A. Remedios, K. Oakley, K. Akimoto, S. Ohno, G. W. Yeo, and J. T. Chang. Regulation of asymmetric division and CD8+ T lymphocyte fate specification by protein kinase Cζ and protein kinase Cλ/ι. *The Journal of Immunology*, 194(5):2249–2259, 2015.

D. P. Mihalko and D. S. Moore. Chi-square tests of fit for type II censored data. *Annals of Statistics*, 8(3):625–644, 1980.

J. J. Moon, H. H. Chu, M. Pepper, S. J. McSorley, S. C. Jameson, R. M. Kedl, and M. K. Jenkins. Naive CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity*, 27(2):203–213, 2007.

D. S. Moore. Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72(357):131–137, 1977.

D. S. Moore. The effect of dependence on chi squared tests of fit. *Annals of Statistics*, 10(4):1163–1171, 1982.

D. S. Moore and M. C. Spruill. Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics*, 3(3):599–616, 1975.

S. J. Morrison and J. Kimble. Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature*, 441(7097):1068, 2006.

K. Murali-Krishna, J. D. Altman, M. Suresh, D. J. Sourdive, A. J. Zajac, J. D. Miller, J. Slansky, and R. Ahmed. Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. *Immunity*, 8(2):177–187, 1998.

K. Murphy and C. Weaver. *Janeway's Immunobiology*. Garland Science, New York, USA, 9th edition, 2016.

S. H. Naik, L. Perié, E. Swart, C. Gerlach, N. van Rooij, R. J. de Boer, and T. N. Schumacher. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, 496(7444):229, 2013.

M. Z. Nashed. Perturbations and approximations for generalized inverses and linear operator equations. In M. Z. Nashed, editor, *Generalized Inverses and Applications*, pages 325–396. Academic Press, New York, USA, 1st edition, 1976.

J. Nikolich-Žugich, M. K. Slifka, and I. Messaoudi. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–132, 2004.

J. Oliaro, V. Van Ham, F. Sacirbegovic, A. Pasam, Z. Bomzon, K. Pham, M. J. Ludford-Menting, N. J. Waterhouse, M. Bots, E. D. Hawkins, S. V. Watt, L. A. Cluse, C. J. P. Clarke, D. J. Izon, J. T. Chang, N. Thompson, M. Gu, R. W. Johnstone, M. J. Smyth, P. O. Humbert, S. L. Reiner, and S. M. Russell. Asymmetric cell division of T cells upon antigen presentation uses multiple conserved mechanisms. *The Journal of Immunology*, 185(1):367–375, 2010.

J. A. Owen, J. Punt, and S. A. Stranford. *Kuby Immunology*. WH Freeman, New York, USA, 7th edition, 2013.

G. Pasqual, A. Chudnovskiy, J. M. Tas, M. Agudelo, L. D. Schweitzer, A. Cui, N. Hacohen, and G. D. Victora. Monitoring T cell–dendritic cell interactions in vivo by intercellular enzymatic labelling. *Nature*, 553:496–500, 2018.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, and T. N. Schumacher. The branching point in erythro-myeloid differentiation. *Cell*, 163(7):1655–1662, 2015.

K. Pham, R. Shimoni, M. J. Ludford-Menting, C. J. Nowell, P. Lobachevsky, Z. Bomzon, M. Gu, T. P. Speed, C. J. McGlade, and S. M. Russell. Divergent lymphocyte signalling revealed by a powerful new tool for analysis of time-lapse microscopy. *Immunology & Cell Biology*, 91(1):70–81, 2013.

C. R. Plumlee, B. S. Sheridan, B. B. Cicek, and L. Lefrançois. Environmental cues dictate the fate of individual CD8+ T cells responding to infection. *Immunity*, 39 (2):347–356, 2013.

K. N. Pollizzi, I.-H. Sun, C. H. Patel, Y.-C. Lo, M.-H. Oh, A. T. Waickman, A. J. Tam, R. L. Blosser, J. Wen, G. M. Delgoffe, and J. D. Powell. Asymmetric inheritance of mTORC1 kinase activity during division dictates CD8+ T cell differentiation. *Nature Immunology*, 17(6):704, 2016.

M. Polonsky, B. Chain, and N. Friedman. Clonal expansion under the microscope: studying lymphocyte activation and differentiation using live-cell imaging. *Immunology & Cell Biology*, 94(3):242–249, 2016.

M. Prlic and M. J. Bevan. Exploring regulatory mechanisms of CD8+ T cell contraction. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43):16689–16694, 2008.

Q. Qi, Y. Liu, Y. Cheng, J. Glanville, D. Zhang, J.-Y. Lee, R. A. Olshen, C. M. Weyand, S. D. Boyd, and J. J. Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 111(36):13139–13144, 2014.

B. J. C. Quah and C. R. Parish. New and improved methods for measuring lymphocyte proliferation *in vitro* and *in vivo* using CFSE-like fluorescent dyes. *Journal of Immunological Methods*, 379(1-2):1–14, 2012.

S. L. Reiner and W. C. Adams. Lymphocyte fate specification as a deterministic but highly plastic process. *Nature Reviews Immunology*, 14(10):699–704, 2014.

N. P. Restifo, M. E. Dudley, and S. A. Rosenberg. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nature Reviews Immunology*, 12(4):269–281, 2012.

M. A. Rieger, P. S. Hoppe, B. M. Smejkal, A. C. Eitelhuber, and T. Schroeder. Hematopoietic cytokines can instruct lineage choice. *Science*, 325(5937):217–218, 2009.

J. C. Rohr, C. Gerlach, L. Kok, and T. N. Schumacher. Single cell behavior in T cell differentiation. *Trends in Immunology*, 35:170–177, 2014.

K. H. Rosen. *Discrete Mathematics and Its Applications*. McGraw-Hill Higher Education, New York, USA, 7th edition, 2011.

S. A. Rosenberg and N. P. Restifo. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, 348(6230):62–68, 2015.

R. L. Rutishauser, G. A. Martins, S. Kalachikov, A. Chandele, I. A. Parish, E. Meffre, J. Jacob, K. Calame, and S. M. Kaech. Transcriptional repressor Blimp-1 promotes CD8+ T cell terminal differentiation and represses the acquisition of central memory T cell properties. *Immunity*, 31(2):296–308, 2009.

A. Sakaue-Sawano, H. Kurokawa, T. Morimura, A. Hanyu, H. Hama, H. Osawa, S. Kashiwagi, K. Fukami, T. Miyata, H. Miyoshi, T. Imamura, M. Ogawa, H. Masai, and A. Miyawaki. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132(3):487–498, 2008.

F. Sallusto, D. Lenig, R. Förster, M. Lipp, and A. Lanzavecchia. Two subsets of memory t lymphocytes with distinct homing potentials and effector functions. *Nature*, 401 (6754):708, 1999.

F. Sallusto, J. Geginat, and A. Lanzavecchia. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annual Review of Immunology*, 22:745–763, 2004.

T. E. Schlub, V. Venturi, K. Kedzierska, C. Wellard, P. C. Doherty, S. J. Turner, R. M. Ribeiro, P. D. Hodgkin, and M. P. Davenport. Division-linked differentiation can account for CD8+ T-cell phenotype *in vivo*. *European Journal of Immunology*, 39 (1):67–77, 2009.

T. N. M. Schumacher, C. Gerlach, and J. W. J. van Heijst. Mapping the life histories of T cells. *Nature Reviews Immunology*, 10(9):621, 2010.

T. N. Schumacher and R. D. Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74, 2015.

T. N. Schumacher, C. Kesmir, and M. M. van Buuren. Biomarkers in cancer immunotherapy. *Cancer Cell*, 27(1):12–14, 2015.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, USA, 1st edition, 1980.

A. K. Sewell. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9): 669–677, 2012.

R. Shimoni, K. Pham, M. Yassin, M. Gu, and S. M. Russell. TACTICS, an interactive platform for customized high-content bioimaging analysis. *Bioinformatics*, 29(6): 817–818, 2013.

T. Sidwell and A. Kallies. Bach2 is required for B cell and T cell memory differentiation. *Nature Immunology*, 17(7):744, 2016.

K. A. Smith and D. A. Cantrell. Interleukin 2 regulates its own receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 82(3):864–868, 1985.

H. J. Snippert, L. G. Van Der Flier, T. Sato, J. H. Van Es, M. Van Den Born, C. Kroon-Veenboer, N. Barker, A. M. Klein, J. Van Rheenen, B. D. Simons, and H. Clevers. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell*, 143(1):134–144, 2010.

G. R. Starbeck-Miller, H. Xue, and J. T. Harty. IL-12 and type I interferon prolong the division of activated CD8 T cells by maintaining high-affinity IL-2 signalling *in vivo*. *Journal of Experimental Medicine*, 211(1):105–120, 2014.

C. Stemberger, K. M. Huster, M. Koffler, F. Anderl, M. Schiemann, H. Wagner, and D. H. Busch. A single naive CD8+ T cell precursor can develop into diverse effector and memory subsets. *Immunity*, 27(6):985–997, 2007.

V. G. Subramanian, K. R. Duffy, M. L. Turner, and P. D. Hodgkin. Determining the expected variability of immune responses using the cyton model. *Journal of Mathematical Biology*, 56(6):861–892, 2008.

C. D. Surh and J. Sprent. Homeostasis of naive and memory T cells. *Immunity*, 29(6): 848–862, 2008.

S. G. Tangye, D. T. Avery, and P. D. Hodgkin. A division-linked mechanism for the rapid generation of ig-secreting cells from human memory B cells. *The Journal of Immunology*, 170(1):261–269, 2003.

J. M. J. Tas, L. Mesin, G. Pasqual, S. Targ, J. T. Jacobsen, Y. M. Mano, C. S. Chen, J.-C. Weill, C.-A. Reynaud, E. P. Browne, M. Meyer-Hermann, and G. D. Victora. Visualizing antibody affinity maturation in germinal centers. *Science*, 351(6277): 1048–1054, 2016.

O. Thaunat, A. G. Granja, P. Barral, A. Filby, B. Montaner, L. Collinson, N. Martinez-Martin, N. E. Harwood, A. Bruckbauer, and F. D. Batista. Asymmetric segregation of polarized antigen on B cell division shapes presentation capacity. *Science*, 335 (6067):475–479, 2012.

M. Tomura, A. Sakaue-Sawano, Y. Mori, M. Takase-Utsugi, A. Hata, K. Ohtawa, O. Kanagawa, and A. Miyawaki. Contrasting quiescent G0 phase with mitotic cell cycling in the mouse immune system. *PLoS ONE*, 8(9):e73801, 2013.

S. Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983.

N. J. Tubo, A. J. Pagán, J. J. Taylor, R. W. Nelson, J. L. Linehan, J. M. Ertelt, E. S. Huseby, S. S. Way, and M. K. Jenkins. Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell*, 153(4):785–796, 2013.

M. L. Turner, E. D. Hawkins, and P. D. Hodgkin. Quantitative regulation of B cell division destiny by signal strength. *The Journal of Immunology*, 181(1):374–382, 2008.

D. E. Tyler. Asymptotic inference for eigenvectors. *Annals of Statistics*, 9(4):725–736, 1981.

K. C. Verbist, C. S. Guy, S. Milasta, S. Liedmann, M. M. Kamiński, R. Wang, and D. R. Green. Metabolic maintenance of cell asymmetry following division in activated T lymphocytes. *Nature*, 532(7599):389, 2016.

E. M. E. Verdegaal, N. F. C. C. de Miranda, M. Visser, T. Harryvan, M. M. van Buuren, R. S. Andersen, S. R. Hadrup, C. E. van der Minne, R. Schotte, H. Spits, J. B. A. G. Haanen, E. H. W. Kapiteijn, T. N. Schumacher, and S. H. van der Burg. Neoantigen landscape dynamics during human melanoma-T cell interactions. *Nature*, 536(7614): 91–5, 2016.

D. Voehringer, C. Blaser, P. Brawand, D. H. Raulet, T. Hanke, and H. Pircher. Viral infections induce abundant numbers of senescent CD8 T cells. *The Journal of Immunology*, 167(9):4838–4843, 2001.

V. Voinov, A. Roza, and N. Pya. Recent achievements in modified chi-squared goodness-of-fit testing. In F. Vonta, M. Nikulin, N. Limnios, and C. Huber-Carol, editors, *Statistical Models and Methods for Biomedical and Technical Systems*. Springer, 2008.

G. Voisinne, B. G. Nixon, A. Melbinger, G. Gasteiger, M. Vergassola, and G. Altan-Bonnet. T cells integrate local and global cues to discriminate between structurally similar antigens. *Cell Reports*, 11(8):1208–1219, 2015.

Q. H. Vuong. Generalized inverses and asymptotic properties of Wald tests. *Economics Letters*, 24(4):343–347, 1987.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3): 426–482, 1943.

T. S. Weber, L. Perié, and K. R. Duffy. Inferring average generation via division-linked labeling. *Journal of Mathematical Biology*, 73(2):491–523, 2016.

R. M. Webster. The immune checkpoint inhibitors: where are we now? *Nature Reviews Drug Discovery*, 13(12):883–884, 2014.

R. M. Webster and S. E. Mentzer. The malignant melanoma landscape. *Nature Reviews Drug Discovery*, 13(7):491–492, 2014.

F. M. Wensveen, K. P. van Gisbergen, I. A. Derks, C. Gerlach, T. N. Schumacher, R. A. van Lier, and E. Eldering. Apoptosis threshold set by Noxa and Mcl-1 after T cell activation regulates competitive selection of high-affinity clones. *Immunity*, 32 (6):754–765, 2010.

H. S. Wilf. *Generatingfunctionology*. Academic Press, New York, USA, 2nd edition, 1990.

J. R. Wilson and K. J. Koehler. Hierarchical models for cross-classified overdispersed multinomial data. *Journal of Business and Economic Statistics*, 9(1):103–110, 1991.

J. D. Wolchok, H. Kluger, M. K. Callahan, M. A. Postow, N. A. Rizvi, A. M. Lesokhin, N. H. Segal, C. E. Ariyan, R. Gordon, K. Reed, M. M. Burke, A. Caldwell, S. A. Kronenberg, B. U. Agunwamba, X. Zhang, I. Lowy, H. D. Inzunza, W. Feely, C. E. Horak, Q. Hong, A. J. Korman, J. M. Wigginton, A. Gupta, and M. Sznol. Nivolumab plus Ipilimumab in advanced melanoma. *The New England Journal of Medicine*, 369 (2):122–133, 2013.

A. Xin, F. Masson, Y. Liao, S. Preston, T. Guan, R. Gloury, M. Olshansky, J. X. Lin, P. Li, T. P. Speed, G. K. Smyth, M. Ernst, W. J. Leonard, M. Pellegrini, S. M. Kaech, S. L. Nutt, W. Shi, G. T. Belz, and A. Kallies. A molecular threshold for effector CD8+ T cell differentiation controlled by transcription factors Blimp-1 and T-bet. *Nature Immunology*, 17(4):422–432, 2016.

M. Yassin and S. M. Russell. Polarity and asymmetric cell division in the control of lymphocyte fate decisions and function. *Current Opinion in Immunology*, 39: 143–149, 2016.

V. W. Yu, R. Z. Yusuf, T. Oki, J. Wu, B. Saez, X. Wang, C. Cook, N. Baryawno, M. J. Ziller, E. Lee, H. Gu, A. Meissner, C. P. Lin, P. V. Kharchenko, and D. T.

Scadden. Epigenetic memory underlies cell-autonomous heterogeneous behavior of hematopoietic stem cells. *Cell*, 167(5):1310–1322.e1317, 2016.

Y. Yu, R. Ceredig, and C. Seoighe. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Research*, 44(4):e31–e31, 2015.

N. Zacharakis, H. Chinnasamy, M. Black, H. Xu, Y.-C. Lu, Z. Zheng, A. Pasetto, M. Langhan, T. Shelton, T. Prickett, J. Gartner, L. Jia, K. Trebska-McGowan, R. P. Somerville, P. F. Robbins, S. A. Rosenberg, S. L. Goff, and S. A. Feldman. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nature Medicine*, 24(6):724–730, 2018.

I. Zaretsky, M. Polonsky, E. Shifrut, S. Reich-Zeliger, Y. Antebi, G. Aidelberg, N. Waysbort, and N. Friedman. Monitoring the dynamics of primary T cell activation and differentiation using long term live cell imaging in microwell arrays. *Lab on a Chip*, 12(23):5007–5015, 2012.

D. Zehn, S. Y. Lee, and M. J. Bevan. Complete but curtailed T-cell response to very low-affinity antigen. *Nature*, 458(7235):211–214, 2009.

B. Zhang. A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 86(3):531–539, 1999.