Evaluation of CD8$^+$ T-cell numbers and repertoires

Marco Ferrarini

Department of Applied Mathematics

The University of Leeds

A thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

April 2018

Related Publications

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. Different parts of Chapter 3 have been published in the two following publications:

- M. Ferrarini, C. Molina-París, and G. Lythe. Sampling from T cell receptor repertoires, 67:79. Springer International Publishing, Cham, 2017. [61]
- P. Gonçalves, M. Ferrarini, C. Molina-París, G. Lythe, F. Vasseur, A. Lim, B. Rocha, and O. Azogui. A new mechanism shapes the naïve $CD8^+$ t cell repertoire: The selection for full diversity. Molecular immunology, 85:66, 2017. [71]

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

"Problems cannot be solved with the same mind set that created them."
Albert Einstein

Acknowledgements

I would like to acknowledge to the European Union for the economical support for this work. In particular, to the Quantitative T cell Immunology (QuanTI) Initial Training Network (ITN) for having trained me in Systems and Computational Immunology and giving me the incredible opportunity to learn from great experimental and theoretical scientists.

To my supervisors, Grant Lythe, Carmen Molina-París and Martín López-García, for their patience, guidance and support, without which this entire work would have not been possible.

To my whole research group, which warmly welcomed me from the very first day, and encouraged me all the way throughout these years.

To INSERM and Pasteur Institutes in Paris, in particular Benedita Rocha and Pedro Gonçalves, for providing the data considered in this thesis, and for the their precious help in understanding important immunological concepts.

To Stuart Barber, from the University of Leeds, for his precious help with the statistical analysis.

To all the QuanTI PIs and fellows, for the incredible experiences lived together.

To my parents, Claudio Ferrarini and Valeria Barletta, to my brother, Luca Ferrarini, and to my grandmother Maria Assunta Belmonte, for their incomparable love and support. I will never forget what you have done for me.

Thank you!

Abstract

This thesis tackles the problem of T-cell receptor (TCR) diversity, from two different points of view. On one hand, the observed TCR diversity is studied from a mathematical perspective, concentrating on the probability of a sample to reproduce a certain percentage of the total TCR diversity. On the other hand, biological samples are considered, focusing on statistical analysis of the observed VDJ gene segments. To conclude, a stochastic model is developed to explore the population dynamics of a simulated TCR repertoire. Computer simulations complete this multidisciplinary approach, helping verifying the different mathematical theories behind the stochastic models.

Contents

List of Figures

Chapter 1

Biological Introduction

Immunology is the study of the different mechanisms of defense of the body against infections. The disease-causing agents can be divided into viruses, bacteria, parasites and fungi. Our body is constantly in contact with millions of them but only few represent a real threat for it. The human body has three possible layers of defense against pathogens: physical and chemical barriers, the innate immune system and the adaptive immune system. Physical and chemical barriers, such as skin and mucosal epithelial lining of the airways and gut, prevent pathogens from entering the body [113, 80]. When these barriers are not sufficient for a certain pathogen, the innate immune system comes into play.

Before going further into the details of the innate part of the immune system, it is worth clarifying that the immune system has four general main functions. The first one, called immunological recognition, defines the very first step of defense after the physical barriers have been overcome [28]. This task requires the ability of recognizing an infection in the shortest possible time. The second important task includes all the so called immune effector functions, such as the T-cell and B-cell activities. As for the third task, the immune system has to have the capacity of self-regulation. This function is extremely important and, in case of failure, it could bring to autoimmune diseases, allergies or over reactions to pathogens [76]. Immunological memory is the last of the four tasks: once the body has been exposed to a certain pathogen, a class of cells of the immune system, called memory cells, are generated and remain in the circulation and tissues patrolling the environment inside the body [1]. These cells will be the first reacting to a second possible entrance of the same pathogen in the body, generating a much stronger and faster response than the first one [113].

The innate immune system (or non-specific immune system) [3, 123], mainly composed by cells that recognize and respond to pathogens in a generic way such as macrophages [15], is able to set up a quick response (within hours) to microbial infections. The main goal is to delay the growth of pathogen numbers in the body as much as possible, while the adaptive immune response gets ready for action. This process is initiated when antigens,

that is molecules or microbial components capable of inducing an immune response, reach the secondary lymphoid organs such as lymph nodes, tonsils and spleen. Dendritic cells [18], usually present in those tissues that are in contact with the environment such as skin, play a critical role in the immune system activation. They are the main carriers of these antigens from peripheral tissues to secondary lymphoid organs, where antigens will be presented to T-cells and B-cells [19]. Macrophages, the mature form of monocytes (bone marrow derived cells), are also a fundamental part of the innate response. Their role is to kill microbes ingesting them [72]. During this process, macrophages secrete chemokines and general cytokines [96], small molecules that cause inflammation and attract cells of the adaptive immune system [114]. It is in fact this last mentioned property that gives the name to chemokines, *chemo*tactic cyto*kines*. It refers to the ability to induce chemotaxis (the movement of an organism in response to a chemical stimulus) in nearby responsive cells. In this way, cells like monocytes or neutrophils are recruited from the bloodstream into the infected tissue.

The inflammation process has now begun. Some pathogens have, unfortunately for us, evolved in such a way to be able to evade these first two barriers. When this happens, an action far more specific than the ones before is evoked: the activation of the adaptive immune system. Its action is highly specific for a particular pathogen and takes longer (maybe days) to be started. The main steps are the recognition of specific *non-self* antigens among millions of *self* antigens, the generation of a response that is tailored to the specific pathogen and the development of immunological memory [113, 23]. This capillary and specific response is obtained thanks to specialized antigen receptors present on the surface of the adaptive immune system cells [51]. Billions of cells are part of this system, generating a vast repertoire of different antigen receptors and allowing the body to respond to ideally every possible pathogen it could be exposed to [113, 172, 46]. Lymphocytes are the main cells of the adaptive response, grouped in B-cells, T-cells, and natural killer cells (NK cells) [10].

As previously said, the effectiveness of this response relies on the accuracy with which the antigen is presented to these cells. Basically all cells in the body are able to present a certain antigen on their cell surface, thanks to a biological complex called MHC (Major Histocompatibility Complex) [87]. Some cells are more specialized than others in the presentation process, being equipped with specific co-stimulatory ligands that can be recognized by the co-stimulatory receptors on the surface of the T-cells [98]. These specialized cells are mainly B-cells and dendritic cells, and are known by the name of antigen presenting cells (APCs). Next section will describe the T-cells, the main focus of this thesis.

1.1 T cells

T cells are a type of lymphocyte that play a central role in the adaptive immune system response. They can be distinguished from other lymphocytes, such as B-cells and natural killer cells (NK cells), by the presence of a T-cell receptor (TCR) on their cell surface, a molecule responsible for recognizing antigens [152]. The co-evolution of TCR and MHC complexes is the basis of an impressive recognition system, thanks to which minor changes in the binding site (between MHC complex and the peptide) may lead to many different coordinated T-cell responses [37].

Their name "T-cells" is due to their maturation process in the thymus [4, 113], an organ in the upper chest. In fact, like B cells, they derive from multi-potent hematopoietic stem cells in the bone marrow and, only after, they migrate to the thymus via the blood. Here they undergo a strict selection process with thymic cells, shaping the mature T-cell repertoire in the body. At this stage, they enter the bloodstream as mature naïve T cells. Their main task is now to circulate through the peripheral lymphoid tissues looking for their corresponding antigens. From these encounters the adaptive immune response will be initiated against the particular pathogen.

There are two different groups of T cells, each with a distinct function. The first group represents $CD4^+$ T cells. Their name is due to the expression of the CD4 (Cluster of Differentiation 4) glycoprotein on their cell surface [174]. The second one is constituted by $CD8^+$ T cells, expressing the CD8 (Cluster of Differentiation 8) glycoprotein [113, 127]. $CD4^+$ T cells can be further split into helper T cells T_H [122], T follicular helper cells (T_{FH}) [39] and regulatory T cells (T_{reg}) [158, 22]. T helper cells (T_H cells) assist (therefore the name "helper") other white blood cells during different immunologic processes, such as the maturation of B cells and activation of cytotoxic T-cells and macrophages. They can be further subdivided in different functional classes, mainly T_H1, T_H2 and T_H17. The former, T_H1, produce a particular kind of cytokines called interferon gamma (IFNγ), capable to activate macrophages [60]. T_H2 produce other cytokines, *IL-4*, *IL-5* and *IL-13*, helping the recruitment of eosinophils and basophils [169]. T_H17, so called from the cytokine *IL-17* that they produce, induce the arrival of neutrophils to the sites of infection. T_{FH} cells focus on helping B-cells in the lymphoid follicles, while T_{reg} cells have the important function of suppressing other T-cells responses when needed, helping to prevent negative outcomes such as autoimmune diseases.

The activation of $CD4^+$ T cells begins with the presentation and recognition processes of peptide antigens between TCRs and MHC class II molecules, molecules expressed on the surface of APCs [55]. Thanks to this, $CD4^+$ T cells become activated starting to divide rapidly and to secrete cytokines that regulate or assist the active immune response [4, 127]. Cytotoxic T-cells (CTLs or T-killer cells) are involved in the destruction of virally infected

cells, tumor cells or cells that have been exposed to DNA damage for different reasons [20]. In order for the TCR to bind to the class I MHC molecule, a molecule expressed on the surface of nearly all host cells, the former must be accompanied by a glycoprotein called CD8, which is able to bind to the constant region of the class I MHC molecule. Once being exposed to infected or nonfunctional host cells, CTLs release the cytotoxins that enhance apoptosis (programmed cell death) in the infected cell.

Thanks to molecules secreted by other T cells called regulatory T cells, the $CD8^+$ cells can be inactivated to an anergic state, which prevents autoimmune diseases [4, 127]. It is worth noticing that, as explained above, TCR is incomplete by itself. Its need to encounter the stimulating ligand through the help of another cell is the main difference between a TCR and antibodies, that can bind to antigens in absence of other structures [37]. Under normal conditions of absence of any infection, the majority of the lymphocytes circulate in the body (blood and limph) as small inactive cells with few cytoplasmatic organelles. In this form they are referred to as naïve lymphocytes and all together they create the naïve repertoire. A lymphocyte is then activated, by binding to an APC through the TCR [65, 67]. Following activation, T cells undergo clonal expansions and their daughter cells start differentiating into different functional classes [112, 33].

1.2 T-cells and repertoire development

As already outlined before, T cells develop from progenitors (same progenitors of the B cells) that derive from pluripotent hematopoietic stem cells in the bone marrow. Their maturation journey starts with the migration of these progenitors to the thymus, the organ situated in the upper anterior thorax, above the heart. It is lobulated on its surface and each of these lobules contains cortical and medullary regions. Once the progenitors arrive in the thymus, they receive a strong signal from stromal cells directly through one of their receptor called Notch1. This is the sign for the progenitors, to switch on specific genes that induce the commitment to undergo the T-cell lineage rather than the B-cell one. Once in the thymus, T-cell precursors start differentiating for up to a week before undergoing a phase of massive proliferation. These developing thymocytes have to pass through different steps before being able to leave the thymus. The first stage for this population is called double negative (DN), reflecting the absence of both CD4 and CD8 molecules on their cell surface [114].

This phase can be further subdivided into four different passages, known as DN1, DN2, DN3 and DN4. DN1 cells express the CD44 glycoprotein, typically used to track early T-cell development. At this point the genes encoding for both chains of the TCR are still in the germline configuration. Following with the maturation process, these lymphocytes start expressing the alpha chain of the IL-2 receptor, known as CD25. This is step DN2,

while the DN3 step is characterized by the decrease of CD44 expression. The somatic recombination of the beta chain of the TCR starts in the DN2 phase and continues until the DN3 one. Here the beta chain is coupled with a pre-T-cell alpha chain (pTα), forming the pre-TCR. This immature TCR pairs with the CD3 (cluster of differentiation 3) protein complex, inducing cell proliferation, the end of the β-chain rearrangement and the expression of both CD8 and CD4. This step is the interface between DN4 and the first step of the double-positive (DP) phase [69].

From this stage we have two different lineages: γ/δ and α/β T cells. The lineage α/β represents the majority of the lymphocytes in the body and it is the only one presenting the CD4 or CD8 molecules. The DP stage is characterized by cells that enlarge and start dividing, reaching a following resting state after some divisions. The rearrangement of the alpha chain of the TCR takes place now. These resting cells express low levels of TCR and are now tested for their ability to recognize self-peptide: self-MHC complexes. Only the ones that recognize these self-complexes are positively selected, and go on to mature and express high levels of TCR. At the same time, they stop the expression of either CD4 or CD8, becoming single-positive T cells (SP) [53].

The other hard examination for these lymphocytes is the so called negative selection, in which all the T-cells responding to self-peptides with high affinity are eliminated. This is the main process that avoids possible autoimmune diseases. It has been clear for the past 20 years that only 5% of the total DP lymphocytes are actually able to survive this double-check, maturing as single positive and entering the blood stream [69, 162]. In these studies it was found that nearly 90% of developing thymocites die for neglect (process killing those T cells which would not be functional due to their inability to bind MHC), while a further 5% die for deletion. When we add this small survival percentage to the thymic involution due to aging, it is clear that much more effort is needed in order to understand the functional relationship between diversity and immune robustness.

An interesting fact is that thymic involution cannot be considered, by itself, the cause of observed loss in diversity of the repertoire. In particular, a study [82] argues the importance of future experimental capabilities, mathematical modeling and data analysis in unraveling the interconnections among thymic involution and clonal expansions due to virus infections or genetic mutations. On this matter, another study found challenging results [130]. A part from confirming the fact that thymic involution does not imply a low diversity in the repertoire, they found that age had an important impact on the inequality of clonal sizes. In particular, the results were indicating an uneven homeostatic proliferation in elderly individuals, where this unevenness was not related to clonal expansions in the memory sub-population. Most remarkably, the authors clearly state their opinion on the difference between human and animals repertoires, comparing their results to those

of a similar study [52] :"Conclusions for the human repertoire from animal models are unreliable because the size of the T-cell compartment and mechanisms and kinetics of T-cell homeostasis are fundamentally different in humans and mice". The author of this manuscript fully agrees with this particular point of view.

Many studies have been focusing on the impact of positive and negative selection on repertoire diversity. Some of them, have particularly focused on the impact that the self-peptides sampling process deployed by APCs has on the repertoire diversity both in the thymus and in the periphery [88]. In particular, these authors recall the importance of maintaining the well established affinity model to understand selection in the thymus [2], which proposes that selection outcome is established by the affinity of the TCR for a pMHC complex, but they also emphasize how important it would be, for a coherent model of thymocyte selection, to consider also the spatial and temporal aspects of self recognition in different micro-environments within the thymus itself. The molecular mechanism that distinguishes positive and negative selection and their impact on repertoire diversity remains nowadays a complex system not fully understood, and it is striking to think that nearly 20 years ago we were already starting thinking about this unresolved problem [144].

The problem of repertoire diversity in the periphery has been studied for a long time, and a "flight for survival" was suggested for lymphocytes in the periphery [64]. The authors discuss the importance of considering a continuous selective pressure throughout the entire lymphocyte life story, due to the constant need to acquire selective advantage on their competitors. The importance of different life stages is also underlined, as well as the presence of specific survival niches to which different lymphocytes belong in different stages of their cell differentiation.

1.3 Biological terminology

- **AIRE:** The autoimmune regulator (AIRE) is a protein encoded by the AIRE gene in humans. It is a transcription factor mainly present in a part of the thymus called medulla and it controls the mechanism underlying the prevention of autoimmune diseases. T cells recognise epitopes presented on a MHC molecule complex, and those T cells that attack the body's own proteins are eliminated in the thymus. The main goal of AIRE is to induce transcription of a wide selection self genes that creates proteins which a T cell could only usually encounter in peripheral tissues, creating what has been defined as an "immunological self-shadow" in the thymus [11]. These proteins, called "tissue-specific self-antigens" (TSAs), are then expressed by medullary thymic epithelial cells (mTECs) or general stromal cells, and T cells that respond to those proteins are eliminated through cell death (apoptosis). This is the reason why AIRE it is thought to drive negative selection [114].

- **Antibody:** Also known as immunoglobulin (Ig), it is a protein produced by plasma cells, a type of white blood cells. They recognize a unique part of the foreign target, called antigen [99]. Antibodies can either be secreted from the cell, having a soluble form, or they can be bound to the B cell surface. In this case, they are known as B cell receptors (BCRs)[114].

- **Antigen:** An antigen (Ag) is any molecule that serves as target for the T or B cell receptors and antibodies. The name is an abbreviation of **anti**body **gen**erator [114].

- **Apoptosis:** Apoptosis is the process of programmed cell death (PCD). It is a complex set of activation mechanisms that, once started, inevitably lead to cell death. Apoptosis should not be confused with necrosis, that is the death of a cell caused by external factors [58].

- **Artemis:** One of the key enzymes involved in V(D)J recombination process [50]. This is the process by which T cell or B cell receptors are created by recombining gene segments known as variable (V), diversity (D) and joining (J). The joining of a V and D segment starts with the RAG (recombination activating gene) nuclease cutting both DNA strands besides the V segment and the D segment. A hairpin structure is formed at the two remaining ends, called the coding ends. Here is when Artemis nuclease comes into play, together with the DNA-dependent protein kinase (DNAPK), binding to these DNA ends and making a single cut near the center of the hairpin. Further processing is applied to the exposed 3' termini, mainly deletion and addition of nucleotides, before the V and D segments are ligated to restore the integrity of the chromosome. The exact cutting point for Artemis is variable and this variability, combined with random nucleotide deletion and addition, is the source of extreme diversity in the resulting antibody or T cell receptor genes [114].

- **ATP hydrolysis:** Reaction in which chemical energy is released from the high-energy phosphoanhydride bonds in adenosine triphosphate (ATP) where it is usually stored. This reaction produces mechanical energy. Adenosine diphosphate (ADP) and an inorganic phosphate, orthophosphate (P_i), are the two main products. A further hydrolyzation process can then occur on the ADP to produce energy, adenosine monophosphate (AMP), and another orthophosphate (P_i) [105].

- **Autophosphorylation:** Modification of proteins occurring after the translation process. A phosphate group is added to serine, threonine or tyrosine residues within protein kinases, normally to regulate the catalytic activity [149].

- **Blunt end:** Blunt end refers to the simplest DNA end of a double stranded molecule. In a blunt-ended molecule both strands terminate in a base pair, leaving no overhangs or unpaired bases.

- **DNA-PK:** DNA-dependent protein kinase is an enzyme very important for the V(D)J recombination process. DNA-PKcs is the catalytic subunit of DNA-PK. The second component is the enzyme Ku [73].

- **Hairpin:** Also known as stem-loop, it is an intramolecular structure. It consists of a single-strand DNA (or just a RNA molecule) in which two complimentary regions come together, forming a double-helix that ends with a non-pairing base sequences, leaving the end open in a loop [43].

- **In-frame:** See ORF and Reading frame.

- **Ku:** Ku is an important enzyme for the V(D)J recombination process. It is a heterodimer of two polypeptides, Ku70 (XRCC6) and Ku80 (XRCC5). In humans, Ku forms a complex with the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) to form the full DNA-dependent protein kinase, DNA-PK [110].

- **Locus:** In genetics, a locus is the specific location of a gene, DNA sequence, or position on a chromosome.

- **ORF:** An open reading frame (ORF) is the part of a reading frame that has the potential to code for a protein or peptide. It is a continuous stretch of DNA typically starting with the a methionine sequence (ATG), and ending with a stop codon (TAA, TAG or TGA in most genomes).

- **RAGs:** The recombination-activating genes (RAGs) encode enzymes that play an important role in the VDJ recombination process. There are two recombination-activating gene products known as RAG-1 and RAG-2, whose cellular expression is restricted to lymphocytes during their developmental stages [139, 114].

- **Reading frame:** A reading frame is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets. In particular, these triplets are called codons if they represent amino acids or stop signals during translation.

- **RSSs:** Recombination signal sequences (RSSs) are short stretches of DNA flanking the V, D and J gene segments of the V(D)J recombination process. They are composed of seven conserved nucleotides (a heptamer) that reside next to the gene

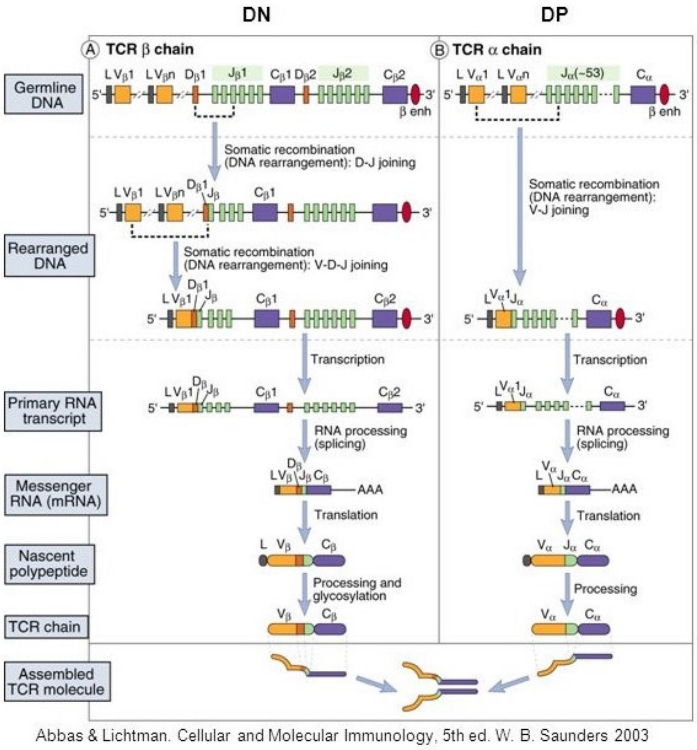

Figure 1.1: Sequential rearrangement of TCR $\alpha\beta$ genes.

encoding sequence followed by a spacer (containing either 12 or 23 variable nucleotides) followed by a conserved nonamer (9 base pairs). Only a pair of dissimilar spacer RSSs are efficiently recombined (i.e. one with a spacer of 12 nucleotides will be recombined with one that has a spacer containing 23 nucleotides). This is known as the 12/23 rule of recombination (or the one-turn/two-turn rule) [21, 114].

- **TdT:** Terminal deoxynucleotidyl transferase (TdT) is a specialized DNA polymerase heavily involved in the V(D)J recombination process. It adds N-nucleotides to the V,D, and J exons during gene recombination, enabling the phenomenon of junctional diversity [25].

- **Transcription factor:** Protein that binds to specific DNA sequences, thereby controlling the rate of transcription of genetic information from DNA to messenger RNA [34].

The reader might want to refer to Figure 1.1 to better understand the following sections.

1.4 V(D)J recombination

The adaptive immune system has to constantly cope with millions of possible pathogens and, in order to recognize them all, it requires an enormous diversity in its lymphocytes

repertoire. This high level of diversity would not be possible without a unique genetic mechanism called V(D)J recombination process, also known as somatic recombination.

The entire process, occurring only in developing lymphocytes during their early stage of maturation, aims at rearranging different gene segments and it occurs in both B cells and T cells, generating a wide repertoire of antibodies/immunoglobulins (Igs) and T cell receptors (TCRs) respectively. The question of whether the recognition system in T cells and B cells differs and, if so, how, has been present from the very beginning of the studies on V(D)J recombination [42].

The lymphatic system of a human being comprises lymphatic organs, lymphatic vessels and the circulating lymph. Thymus and bone marrow are known to be the two primary lymphoid organs. They represent, respectively, the maturation centers for T-cells and B-cells. As previously said, somatic recombination occurs in these organs and its ultimate results are new antigen-binding regions of Igs and TCRs, allowing for the recognition of antigens from nearly all pathogens.

This incredible process was first discovered in 1987 by Susumu Tonegawa, Ph.D. who has then been awarded the Nobel Prize in Physiology or Medicine [159].

1.4.1 Recombination in T cell receptors

T-cell receptors (TCRs) are heterodimers built of two different protein chains. In humans, the great majority of TCRs consist of an alpha (α) and a beta (β) chain, while only a minority is built from two different chains: gamma (γ) and delta (δ). We focus here on α/β TCRs, although the recombination process for the γ/δ ones is very similar. Both TCR α and β chains consist of a variable (V) protein region (V_α and V_β) and a constant (C) region (C_α and C_β) [114].

The DNA sequence of the human V_α region (TCRα locus, chromosome 14) contains $\sim$ 70 variable (V) segments and 61 joining (J) segments, followed by a single constant (C) gene which contains separate exons for the different parts of the constant region of the protein chain. The human TCRβ locus (chromosome 7) has a different organization: 52 V segments, 1 diversity (D) segment ($D_\beta 1$), 6 J segments ($J_\beta 1$), 1 constant gene, 1 D segment ($D_\beta 2$), 7 J segments ($J_\beta 2$) and another final constant gene [127]. Different from humans, the mice TCRβ locus (chromosome 6) contains 35 V segments, 2 D segments and 12 J segments [95]. The mice V_α region contains instead $\sim$ 132 V segments and $\sim$ 60 J segments [66, 95].

The variable regions in both chains contain three hyper-variable regions, also called complementarity determining regions (CDR1, CDR2, CDR3). The first two regions are only encoded by V segments, while the CDR3 region is the main focus of the entire V(D)J recombination [166]. The recombination process starts with the joining of a D segment to

a J segment in the β chain, which can involve either the joining of the $D_\beta 1$ gene segment to one of the six $J_\beta 1$ segments or the joining of the $D_\beta 2$ gene segment to one of the seven $J_\beta 2$ segments. The next step is the joining of a V segment to the newly formed DJ complex, followed by deletion of all other gene segments among them. At this stage, the incorporation of the constant domain gene ($V_\beta - D_\beta - J_\beta - C_\beta$) occurs, followed by the synthetisation of the primary transcript. Transcription of the mRNA brings to the full length protein for the TCRβ chain. The α-chain undergoes the same process, differing only in the lack of the D segments. The β- and α- chains are then assembled, resulting in the formation of the $\alpha\beta$-TCR that is expressed on the majority of T lymphosytes. In the next section we will focus on the main joining process that gives rise to any D-J, V-DJ or V-J coupling.

1.4.2 V(D)J recombination process

The V(D)J recombination process starts with the binding of the recombination activating gene 1 and 2 enzymes (RAG1 and RAG2) to a recombination signal sequence (RSS) flanking a coding gene segment (V, D, or J). These 2 genes were first discovered in the late 80s [142] but it was not clear from the very beginning whether these genes could only encode tissue-specific components of the main structure governing the somatic recombination process, the V(D)J recombinase [7]. The recognition of RSS by the RAG complex is fairly straightforward given the very conservative shape of RSSs. There are three important elements that help this recognition: a heptamer of seven conserved nucleotides, a spacer region of 12 or 23 basepairs in length, and a nonamer of nine conserved nucleotides. These consensus heptamer and nonamer are highly conserved (CACAGTG and ACAAAAACC). On the other hand, the spacer region is highly variable but with a highly conserved length [38].

Gene segments that have to be recombined are usually adjacent to RSSs of different spacer lengths, that is one has a "12RSS" and one has a "23RSS" [160], following what is well known as the 12/23 Rule. Once two RAG complexes have bound to two different RSS, the two complexes are brought together and, once close to each other, they create a single-strand notch in the DNA between the two first bases of the two RSSs (just before the heptamers) and the respective coding segments attached to these two RSSs (e.g. V and D segments) [114]. The presence of these breaks introduced at these junctures was demonstrated early during the studies on somatic recombination [140] but a proper understanding on the type of notches was not immediately clear [7].

Two different DNA ends are thus created: a hairpin (stem-loop) on the coding segment and a blunt end on the signal segment [143]. The two double stranded breaks at the blunt end are then ligated together by the action of a heterodymer protein called Ku

(Ku70:Ku80) in association with X-ray repair cross-complementing protein 4 (XRCC4), producing a circular piece of DNA containing the material between the coding segments, known as signal joint. It is still not clear whether this signal joint is discarded or then reused in different ways. On the hairpin side, the coding ends are processed prior to their ligation, leading eventually to junctional diversity [126].

A Ku protein also binds on this side, followed by the DNA-dependent protein kinase (DNA-PK) and Artemis nuclease complex. Artemis usually has an exonuclease activity, but it can gain endonuclease activity once bound to the DNA-PK complex, enabling the opening of the hairpins. Artemis is activated and opens the coding end hairpins [101]. If the cleavage is in the center of the hairpin, a blunt end is created. If the cleavage is not centered, the result is an overhang of extrabases on one of the two ends. These bases are called palindromic (P) nucleotides.

Next, DNA ligase IV, XRCC4, Ku, and DNA-PK align the DNA ends (from the two different gene segments) and recruit the template-independent DNA polymerase (TdT) in order to add, in a 5' to 3' direction, non-templated (N) nucleotides to the coding end. This enzyme was immediately designated, from its very discovery, as one (if not the only one) of the tissue-specific enzymes capable of modifying the V(D)J junctions [6, 7]. Although it is thought that the addition is random, there have been signs of TdT exhibiting a G/C preference for the added nucleotides [126].

As last stage, exonucleolytic activity takes place, removing bases to adjust the process and finally pair the two ends, eventually ligated by DNA ligase IV in association with XRCC4. The result of this process is a highly variable TCR binding region, allowing the adaptive immune response to be almost always ready for novel pathogens. The entire process requires a high amount of energy, which needs to be strictly regulated and controlled. As it is now clear, somatic recombination is a highly complex process involving highly specific procedures.

This mechanism is essential to our survival, being the only process capable of creating such a diversity against the nearly infinite pathogens possibilities. Being so specific, it has, unfortunately, the potential to generate aberrant DNA damage in developing lymphocites [138]. Although the study of these negative effects is not the goal of this thesis, we thought it was necessary to at least cite the problem for the awareness of the reader [173, 125].

1.5 Studies on T-cell repertoire diversity

Numerous studies examined the T-cell repertoire diversity during the last 30 years, many of them focusing on the impact of viral infections on the diversity itself [124, 69, 151, 89]. In particular, some studies focused on the stability of the diversity of the TCR repertoire during and following a viral infection [97]; the authors showed how the Lymphocytic

Choriomeningitis Virus (LCMV) induced repertoire changes following the clearance of the viral antigens. The authors also realized the T-cell repertoire skewing properties of LCMV infections at the memory level. Another interesting result was the generation, by genetically identical mice, of different T cell responses to the same peptides. Other authors studied the dynamics of the $CD8^+$ TCR repertoire in response to LCMV infection, identifying differences of the order of $10^1 - 10^2$ in the expansion of the T-cell population in response to two different peptides of the same virus [44]. Thanks to the usage of mathematical models, they show how this difference in expansion could be due to an actual difference in proliferation rates or in the proliferation period. Following studies aimed at determining the main differences among acute and chronic LCMV infections [9], where mathematical modelling was able to detect the immunodominance effect caused by chronic infection, as previously showed by [171]. The impact of different epitopes on $CD8^+$ T-cell response gave even more insights [141], indicating that the magnitude of the response might be influenced by the epitope specificity but the same does not seem to hold for the TCR β chain repertoire diversity (in mice). In fact, the β-repertoires showed very little difference in response to three different infections. The main goal of the authors was to address the question on whether or not it would be possible to modify the diversity and specificity of the $CD8^+$ repertoire by changing the vector used to deliver the particular epitope taken into consideration.

A different aspect of the T-cell repertoire diversity has also been studied during the years: the bias present at the very first steps of the TCR repertoire creation, that is the unbalanced process of the somatic recombination. An important result was shown [136], and for the first time it was clarified that the V(D)J recombination process is biased towards some specific V-D-J combinations, contrarily to what was believed until that moment. The authors proved this striking result thanks to an analysis of the size of the V-D-J overlap in unrelated adult humans. As previously described, the positive and negative selections play a main role in shaping the repertoire. It becomes even more important when considering the possibility that T cells could escape negative selection even though presenting the potential to express two different functional TCR α chains [41]. The results from this study came from a new PCR technique for the simultaneous analysis of both α and β chains from single cells experiments, developed by the same authors. The authors discuss the importance of these findings, suggesting the possibility of triggering autoimmunity as a consequence of infection, in case one of these special double-TCR T cells were to be the respondent to some infection. Other authors have discussed the problem of autoimmunity related to the cross-reactivity concept [120]. The focus was on the possibility, for an incompletely deleted naïve T cell population specific for a tissue-restricted self peptide, to be triggered by systemic production of self-peptides and

cause autoimmune problems. Studies on $CD8^+$ and $CD8^-$ TCR repertoire in troutes were carried with some interesting results [32]: different regulatory patterns for the diversity of TCR β chains by $CD8^+$ and $CD8^-$ in trout and mammals could exist. The authors in fact found out that the CDR3 region for the different β chains had much more regular profiles after the viral infection, suggesting a mechanism for which the infection itself lead to multiple expansions of $CD8^-$ T cell clonotypes, possibly reducing the importance of the large peaks otherwise observed in the uninfected troutes. The authors argue that precisely this finding could favor the idea of different impact of infections on $CD8^-$ and $CD8^+$, in both trout and mice.

TCR diversity was also studied in relation to T-cell subsets, such as effector, central/effector memory and naïve cells, in both $CD8^+$ and $CD4^+$ populations [27]. TCR expansions were found in the effector subpopulations as opposed to naïve or memory ones. Following these findings, the authors argued the importance of including subset analysis in TCR repertoire studies and suggest the idea of a more polyclonal-oriented antigen-driven expansion in long-lived T central memory cells as opposed to a more oligoclonal expansion in short-lived T effector cells. The importance of T-cell subsets was also the focus of other authors [119]. T-cell precursor frequencies were found not to be correlated with immunodominance hierarchies induced by pathogen presentation; public clonotypes were found to be rare in the precursor pool while the memory pool presented narrower TCR repertoire diversity. A broad analysis of CDR3β from healthy mice was carried out in 2014 [103]. An extensive number of publicly shared sequences was found and it was suggested that despite the random generation process of TCR repertoires, a sort of uniformity was present in the mice repertoires' diversity, suggesting ongoing selection tends to modify the initial ranomized diversity. A different study suggested instead a possible connection between T-cell precursors and immunodominance hierarchies [132]: in fact the authors suggest that the age-related modifications in T-cell immunodominace hierarchies may be driven by changes in numbers of T-cell precursors.

A deeper insight into the mechanism of somatic recombination and TCR repertoire diversity was given by a study aiming at assessing the impact of individual genetic factors on the immune diversity [175]. Exciting discoveries emerged from this study: first, the overlap between TCR repertoires of monozygous twins was not that different to the overlap of unrelated individuals. Second, and possibly even more striking, the authors discuss results in which the TCR V genes choice for recombination in the thymus is strictly related to genetic traits, as already shown in a previous study [109], while the choice of TCR J genes seems to be completely random. They also discuss the preference, in subsequent selection in the thymus, of some α J segments as opposed to β J segments.

The potential of bioinformatics to detect the immunological status of a patient was studied in a very interesting piece of work [75]. Here the authors introduce a bioinformatics framework working on Hill-based diversity profiles enabling quantification of immunological information enclosed in immune repertoires. A broad range of immunological states such as healthy or transplantation recipient were able to be predicted with high accuracy, instilling the idea that, in the near future, repertoire profiling could help recovering a great amount of what they define as immunodiagnostic fingerprints. In a similar way, other authors reviewed different biotechnological methodologies (mainly high-throughput sequencing HTC) for the study of TCR repertoire diversity, highlighting the profound effect of these techniques on our knowledge of the immune system dynamics during health or diseases [78]. Going back on the topic of the importance of bioinformatics on immune repertoire analyses, it is worth mentioning a recently developed computational model called TraCeR [154]. This method is able to reconstruct full-length, paired TCR sequences from single-cell analysis.

The first analysis of both the naïve and the epitope-specific TCR $\alpha\beta$ repertoires was developed only recently [40]. The authors argue the scarce and not accurate information carried on by studies of subsets of cells or of single TCR chains. They claim in fact that TCRα usage is at least as diverse as TCRβ usage. Before concluding this section, we would like to give a quick overview of the different clonal size estimations that have been given during the years, in order to give a taste of the difficulty of this research area. From bulk studies on mouse, TCRβ uniquess was said to be around 10% [30], 28% [31], 55% [128] and 68% [131] of the total sequences analyzed. In humans the situation is not much better, as the number of distinct clonotype classes was estimated to be between 10^6 and 2×10^7 [13, 130, 135].

So far, we have given an overview of the general understanding of the biology behind those immunological complex systems that somatic recombination and diversity maintenance are. More could be said, especially from the point of view of the quantitative efforts made during the last 30 years to broaden the knowledge in this area, but we will leave these themes for the next chapter, in a specific section dedicated exactly to the evolution of quantitative methods in this research area.

Chapter 2

Mathematical Introduction

This section focuses on the definition of the basic probability concepts generally used in this thesis. Firstly we will introduce the general concepts of a **probability space**, which will help defining a general random variable X, together with some of their main properties, such as its expected value $\mathbb{E}(X)$, its variance $\mathrm{Var}(X)$ and probability generating function $\phi_X(z)$. We will then focus on some important general properties of the pgf, before introducing different examples of both discrete and continuous distributions. Finally, we will move the attention to stochastic processes and Markov chains, introducing basic results.

2.1 Probability spaces

In probability theory **probability space** $(\Omega, \mathcal{F}, \mathcal{P})$ is defined as a mathematical construct [77]. The first part of a probability space is the **sample space** Ω, defined as the set of all possible outcomes of the considered random process. We define now the other two parts: $\mathcal{F}$ and $\mathcal{P}$.

2.1.1 The set of events $\mathcal{F}$

Let Ω and 2^Ω be respectively a set and its power set, where the power set 2^Ω is defined as the set of all possible subsets of Ω including the empty set $\emptyset$ and Ω itself. The **set of events** $\mathcal{F} \in 2^\Omega$ is mathematically defined as a σ-algebra [77], that is a subset of 2^Ω satisfying three important properties:

- $\Omega \in \mathcal{F}$;
- $\mathcal{F}$ is said to be closed under complementation, that is if $f \in \mathcal{F}$, then $\Omega \setminus f \in \mathcal{F}$;
- $\mathcal{F}$ is said to be closed under countable unions, that is if $f_1, f_2, f_3, \cdots \in \mathcal{F}$, then $f = f_1 \cup f_2 \cup f_3 \cup \cdots \in \mathcal{F}$.

It is important to notice that the term **event** must be seen as a set of zero, one or multiple outcomes, that is a subset of the sample space.

2.1.2 The probability measure $\mathcal{P}$

Let us consider a set of events $\mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. A probability measure $\mathcal{P}$ is defined as a real-valued function $\mathcal{P}(f)$ on $\mathcal{F}$ satisfying two main requirements:

- $\mathcal{P}(f) \in [0, 1]$, with $\mathcal{P}(\emptyset) = 0$ and $\mathcal{P}(\Omega) = 1$;
- Countable additivity property i.e., for all countable collections $\{f_k\}$ of pairwise disjoints sets, $\mathcal{P}\left(\bigcup_{k \in K} f_k\right) = \sum_{k \in K} \mathcal{P}(f_k)$.

2.2 Random variables

A real-valued random variable X is a real-valued function $X : \Omega \to \mathbb{R} = (-\infty, \infty)$. In this probability space, the probability $\mathcal{P}(X \leq k)$ represents the probability of the set of outcomes $\{\omega \in \Omega : X(\omega) \leq k\}$. From now on $\mathcal{P}(X \leq k)$ will be indicated by $\Pr(X \leq k)$. Real-valued random variables can be divided in two distinct cathegories: discrete and continuous random variables. Discrete random variables may only take a countable number of values, such as $\{0, 1, 2, \cdots\}$, while a continuous random variable takes an uncountable (infinite) number of possible values. Examples of both kinds will be given in the next sections. We are now defining the cumulative distribution function for a generic random variable, which will allow us to define the probability mass function for a discrete random variable and the probability density function for a continuous random variable. We will also introduce the expected value, variance of a general random variable, distinguishing between the discrete and the continuous case. Finally, we will introduce the concept of probability generating function for a discrete random variable.

2.2.1 Cumulative distribution function (cdf)

Given a random variable $X : \Omega \to E (E \subseteq \mathbb{R})$, the cumulative distribution function (cdf) $F_X(x) : E \to [0, 1]$ for the random variable X is the function defined as

$$F_X(x) = \Pr(X \leq x) = \Pr(\{\omega \in \Omega : X(\omega) \leq x\}).$$

2.2.2 Probability mass function (pmf)

Given a discrete random variable $X : \Omega \to E (E \subseteq \mathbb{R})$, the probability mass function (pmf) $f_X(x) : E \to [0, 1]$ for the random variable X is the function defined as

$$f_X(x) = \Pr(X = x) = \Pr(\{\omega \in \Omega : X(\omega) = x\}).$$

2.2.3 Probability density function (pdf)

Given a continuous random variable X with cdf F_X, and assuming the existence of a non-negative, integrable function $f_X(x) : E \to [0, \infty)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy,$$

then the function $f_X(x)$ is called the probability density function (pdf) for the random variable X.

2.2.4 Independent random variables

Consider two continuous random variables X and Y, with probability density functions $f_X(x)$ and $f_Y(y)$ respectively. Let their joint probability distribution be $f(x, y)$. The random variables X and Y are said to be independent if and only if

$$f(x, y) = f_X(x)f_Y(y) \quad \forall (x, y) \in \Omega_x \times \Omega_y.$$

If we consider discrete random variables, the same holds considering probability mass functions instead of probability density functions.

2.2.5 Expected value and variance

Given a *discrete random variable* X taking values $x_1, x_2, x_3, \cdots$ with probabilities $p_1, p_2, p_3, \cdots$, its expected value $\mathbb{E}(X)$ is defined as

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} p_i x_i,$$

while its variance $\mathrm{Var}(X)$ is defined as

$$\mathrm{Var}(X) = \sum_{i=1}^{\infty} p_i (x_i - \mathbb{E}(X))^2.$$

Given a *continuous random variable* X taking values in $A \subseteq \mathbb{R}$ with pdf $f_X(x)$, its expected value $\mathbb{E}(X)$ is defined as

$$\mathbb{E}(X) = \int_A x f_X(x)dx,$$

while its variance $\mathrm{Var}(X)$ is defined as

$$\mathrm{Var}(X) = \int_A (x - \mathbb{E}(X))^2 f_X(x)dx.$$

2.2.6 Conditional probability

Consider two random variables X and Y. The conditional probability that $X = x$ given that $Y = y$ is defined as

$$\Pr(X = x|Y = y) = \frac{\Pr(X = x \text{ and } Y = y)}{\Pr(Y = y)}.$$

2.2.7 Probability generating function (pgf)

Let $X \geq 0$ be a discrete random variable with probability mass function $f_X(x)$. The probability generating function (pgf) of X is defined as

$$\phi_X(z) = \mathbb{E}(z^X) = \sum_{x=0}^{\infty} f_X(x) z^x. \tag{2.1}$$

Some important basic properties of pgf are now given. The reader is directed to any university textbook on probability theory for the proof of these properties. A good reference book for stochastic processes applied to biology where these properties can be found is [5]. The first property is the relation between the pmf $f_X(x)$ and the pgf, given by

$$f_X(k) = \Pr(X = k) = \frac{\phi_X^{(k)}(0)}{k!}, \tag{2.2}$$

where $k!$ indicates the factorial of k while, for a general function $g(x)$, $g^{(k)}(0)$ indicates the k^{th} derivative of the function g calculated at $x = 0$.

The second important property is

$$\phi_X(1^-) = \sum_{x=0}^{\infty} f(x) = 1, \tag{2.3}$$

where $\phi_X(1^-) = lim_{z \to 1^-} \phi_X(z)$ and z is going to 1 from below.

A third important property of the pgf is as follows:

$$\mathbb{E}(X) = \phi_X^{(1)}(1^-). \tag{2.4}$$

Finally, we cite the following useful property: let $X_1, X_2, \cdots, X_n$ be a set of independent random variables. Define $S_n = \sum_{i=1}^{n} X_n$. The pgf of S_n is

$$\phi_{S_n}(z) = \phi_{X_1}(z)\phi_{X_2}(z) \cdots \phi_{X_n}(z). \tag{2.5}$$

2.3 Discrete random variables

In this section we give a quick overview of the main discrete distributions that will be used in different chapters of this manuscript. We briefly recall that a discrete random variable takes values on a finite (or countable) list of possible values with certain probabilities described by the probability mass function.

2.3.1 Bernoulli distribution

Probability distribution of a random variable X taking value 1 with probability $p \in (0, 1)$ and value 0 with probability $q = 1 - p$. We have

- $\Pr(X = k) = \begin{cases} p & \text{for } k = 1 \\ q & \text{for } k = 0 \end{cases}$,
- $\mathbb{E}(X) = p$,
- $\mathrm{Var}(X) = pq$,
- $\phi_X(z) = q + pz$.

2.3.2 Binomial distribution

Probability distribution of a random variable X representing the number of successes in a sequence of n independent Bernoulli trials, where each trial has a probability p of success. The special case $n = 1$ represents the Bernoulli distribution. We have

- $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$,
- $\mathbb{E}(X) = np$,
- $\mathrm{Var}(X) = np(1-p)$,
- $\phi_X(z) = (1 - p + pz)^n$.

2.3.3 Geometric distribution

Probability distribution of the number X of Bernoulli trials needed to get the first success, where each trial has a probability p of success. We have

- $\Pr(X = k) = p(1-p)^{k-1}$, k = 1, 2, ...
- $\mathbb{E}(X) = \dfrac{1}{p}$,
- $\mathrm{Var}(X) = \dfrac{1-p}{p^2}$,
- $\phi_X(z) = \dfrac{pz}{1-(1-p)z}$.

2.3.4 Poisson distribution

Probability distribution of the number X of events occurring in a fixed interval of time, knowing that these events occur independently and with a fixed average rate λ. We have

- $\Pr(X = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$, k = 0, 1, 2, ...
- $\mathbb{E}(X) = \lambda$,
- $\mathrm{Var}(X) = \lambda$,
- $\phi_X(z) = e^{\lambda(z-1)}$.

2.3.5 Logarithmic distribution

Probability distribution originally used for the modelling of species abundance. The only parameter is p. We have

- $\Pr(X = k) = \dfrac{-1}{\log(1-p)}\dfrac{p^k}{k}$, k = 1, 2, ...
- $\mathbb{E}(X) = \dfrac{-1}{\log(1-p)}\dfrac{1}{1-p}$,
- $\mathrm{Var}(X) = -p\dfrac{p+\log(1-p)}{(1-p)^2\log^2(1-p)}$,
- $\phi_X(z) = \dfrac{\log(1-pz)}{\log(1-p)}$.

2.3.6 Hypergeometric distribution

Probability distribution of the number of successes X in n draws without replacement, where the draws are taken from a finite population of size N that contains exactly K successes. It is similar to the Binomial distribution, with the exception that the draws are taken without replacement. We have

- $\Pr(X = k) = \dfrac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, k = 0, ... min(n,k)
- $\mathbb{E}(X) = \dfrac{nK}{N}$,
- $\mathrm{Var}(X) = \dfrac{nK}{N}\dfrac{N-n}{N-1}\dfrac{N-K}{N}$,
- $\phi_X(z) = \dfrac{\binom{N-K}{n}\ {}_2F_1\left(-n, -K; N-K-n+1; z\right)}{\binom{N}{n}}$,

where ${}_2F_1\left(-n, -K; N-K-n+1; z\right)$ represents the ordinary hypergeometric function.

2.4 Continuous random variables

In this section we give a quick overview of the main continuous distributions that will be used in different chapters of this manuscript. We briefly recall that a continuous random variable takes values on an uncountable list of possible values with certain probabilities described by the probability density function.

2.4.1 Exponential distribution

Probability distribution describing the time between events in a process where events occur in a continuous way, independently from each other, and at a constant average rate (known as Poisson process). It can be seen as the continuous analogue of the geometric distribution. We have

- $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$
- $\mathbb{E}(X) = \dfrac{1}{\lambda}$,
- $Var(X) = \dfrac{1}{\lambda^2}$.

2.4.2 Gamma distribution

This probability distribution can be parametrized by a shape parameter α and a rate parameter β. Different parametrizations are possible. We have

- $f(x) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)}$, $x \geq 0$
- $\mathbb{E}(X) = \dfrac{\alpha}{\beta}$,
- $Var(X) = \dfrac{\alpha}{\beta^2}$,

where $\Gamma(z)$ represents the Gamma function defined as

$$\Gamma(z) = \begin{cases} (z-1)! \text{ for } z \text{ positive integer number} \\ \\ \displaystyle\int_0^\infty t^{z-1} e^{-t} \mathrm{d}t \text{ for } z \text{ complex number with positive real part} \end{cases}$$

2.4.3 Beta distribution

This probability distribution is parametrized by two shape parameters α and β. We have

- $f(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1} e^{-x\beta}}{B(\alpha, \beta)}$, $0 \leq x \leq 1$

- $\mathbb{E}(X) = \dfrac{\alpha}{\alpha+\beta}$,
- $Var(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$,

where $B(x,y)$ represents the Beta function defined as

$$B(x,y) = \begin{cases} \dfrac{(x-1)!(y-1)!}{(x+y-1)!} & \text{for } x,y \text{ positive integer numbers} \\ \\ \displaystyle\int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t & \text{for } x,y \text{ complex numbers with positive real part} \end{cases}$$

2.5 Stochastic processes

Consider a subset T of $[0,\infty)$. Consider a family of random variables $\{X_t(s) : t \in T, s \in \Omega\}$. This family is called a stochastic process. Depending on whether T is countable or uncountable, and depending on whether $X_t(s)$ are discrete or continuous random variables, $\{X_t(s) : t \in T, s \in \Omega\}$ needs different techniques to be analyzed. For a countable T, we have a discrete-time stochastic process; when instead we have an uncountable T, the process is defined as continuous-time stochastic process. The literature and knowledge on stochastic processes are very broad but, for the purposes of this thesis, the following sections will only briefly introduce a special class of stochastic processes, the Markov processes.

2.5.1 Markov processes and Markov chains

Briefly speaking, a stochastic process has the Markov property if its future behaviour depends on the present state only, and not on all the previous ones. In particular we are interested in expressing this property for a discrete-time stochastic process where $T = \{0,1,2,\ldots\}$. We say that the process $\{X_t\}_{t\in T}$ has the Markov property if

$$\Pr(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \Pr(X_t = x_t | X_{t-1} = x_{t-1}).$$

In particular, the probabilities $\Pr(X_t = x_t | X_{t-1} = x_{t-1})$ are called transition probabilities. Given the set T being discrete, we define this Markov process as discrete-time Markov chain. The natural extension of this property to the continuous-time case can be stated in the following way. Consider T uncountable. A continuous-time stochastic process $\{X_t\}_{t\in T}$ with space of states $\mathbb{S}$ is a continuous-time Markov chain if

$$\Pr(X(t) = j | X(t_n) = i_n, \ldots, X(t_1) = i_1) = \Pr(X(t) = j | X(t_n) = i_n)$$

where $0 \leq t_1 \leq \cdots \leq t_n \leq t$ is any non-decreasing sequence of $n+1$ times and $i_1, \ldots, i_n, j \in S$ are any $n+1$ states in the space of states, for any integer $n \geq 1$.

2.6 Mathematical analyses of repertoire diversity

Species diversity has been the focus of a broad amount of work for more than 70 years now. A great amount of literature has been produced during these years, but an overview of this literature is not the goal of this section. Here we aim at reviewing mathematical models and species diversity similarity measures, as well as statistical methods, applied to the study of T-cell receptor repertoire diversity. An enormous effort has been made by scientists from all over the world in the last 20 years to deepen the knowledge of the mathematical bases of V(D)J recombination process, TCR repertoire diversity, immune sampling diversity and immunenodiversity-related problems. We here review the main steps of this journey, trying to maintain a certain historical continuity.

One of the first contributions in this sense focused on the development of a probabilistic model trying to explain the connection between repertoire diversity, self antigens and foreign antigens [46]. The authors basically address the question of how diverse the repertoire should be in order to be able to recognize all the theoretically possible pathogens. The importance of the diversity of self antigens that the immune system needs to avoid reactivity with is established, showing that the number of these particular peptides is the main driver of repertoire diversity, rather than the number of foreign antigens. A few years later, similar work was carried out focusing specifically on the immunological feature of cross-reactivity [106]. In this study, the necessity of a wide cross-reactive ability of T cells were analysed, leading to the well known problem of the lymphoid system of an hypothetical mouse with one clonotype for each possible MHC-associated peptide: such a mouse, even considering the best case in which it had only one cell in each clone, would need a lymphoid system 100 times larger than the mouse body itself. In 1999, interesting work was done on affinity-driven TCR repertoire selection and the problem of alloreactivity [53], the problem of T cells responding to foreign MHC entering the body, for example following an organ transplant [116]. The authors simplify the structural complexity of a TCR to a digit string representation, claiming the advantage of such a formalism, over previously proposed bit strings approaches [45], in order to control the resolution of affinity distribution.

A different interesting problem was also studied in those years: the requirement for regulatory T cells (Tregs) [107]. The existence of Tregs had already been experimentally established, but the reasons behind their presence in the immune system were not clear. Apart from preventing autoimmunity problems, the authors suggest that Tregs may be a subset directly generated in the thymus with the goal of controlling inflammatory responses

induced by enteric organisms. In 2003 a very interesting article was published, suggesting the possibility for the repertoire to have fractal properties [118]. The main striking result concerned the possibility to describe not only the population of ranked clonotypes, but also the clonal frequencies of many different subsets of the repertoire, by a power law-like distribution.

A quantitative model of thymic selection, involving structural differences within the thymus, was developed in those years [59]; the main goal of the model was to estimate the fractions of T cells positively (and negatively) selected in both main areas of the thymus, namely the cortex and the medulla. Results indicated that the majority of the thymocytes die due to neglect selection in the cortex, and that the negative selection might happen with a higher probability in the medulla. During those years, another important research area was being explored: mathematical methods to compare the diversity of samples drawn from a given TCR repertoire. Facing this problem means facing the same problems that mathematical ecology had been facing (and was still facing) in trying to estimate species diversities from selected samples, but in a much more hostile mathematical environment. In fact, as the reader will be able to appreciate in this manuscript, the number of lymphocytes in a human body is of the order of 10^{12} while the estimated number of different clonotype classes is approximately 10^8. Established methods for comparing sample diversity were discussed in a work of 2007 [164]. In particular, one method was discussed (applied in this manuscript in Section 4.3.2): the use of a non-parametric statistical test called randomization test, based on a test statistic called Simpson's diversity index.

Stochastic models were used to study TCR repertoire diversity [153]. The authors built a stochastic model incorporating competition for survival signals among competitors. In particular, they built on the concept of niche overlap previously introduced by [47, 48, 49] and defined the concept of robust repertoire as one in which the loss of some classes wouldn't affect its capabilities that much, but at the same time with the least possible overlap in the coverage of the epitope space. The main result was, for a robust repertoire, the biological property for which the majority of the clonotype classes clusters around a mean value of niche overlap $\nu = 1$. The authors avoided to include other types of competition in the model, such as the non-TCR specific one for cytokines previously studied in [26], arguing that this non-TCR specific competition impact the total number of T cells, and not repertoire diversity, as discussed in [104].

Different Poisson abundance models (PAMs) were presented in [145], where the authors discussed the statistical incompleteness of previous studies on TCR diversity estimates comparing the repertoires of T-cell subsets [62, 170, 92]. In particular, the PAMs were used to study the clonal size distributions of mice with limited TCR diversity. The exponential distribution is among those theoretical distributions obtained as possible clonal

size distributions. Although we believe that discrete distribution would better fit the modelling of clonal size distributions, an interesting point of discussion was proposed in 1957: the broken stick model of [102]. In particular, the authors argue that, even though the exponential distribution can be derived by this model, this model itself might not be well representative for the TCR repertoire given the time-dependent changes of thymic micro-environments, described as structural niche, together with signal niche as described above. To overcome this problem, they recall a similar model, the sequential broken stick model [155]. This model would be a better representative of the time dependent structural properties of the thymic niches, and would bring to the hypothesis of a Lognormal clonal size distribution. The application of the analytical solution to the unseen species problem given by [63] to the problem of TCR diversity was followed by [135], with the main result being a much higher diversity ($\sim$ 4-fold) than what reported in the previous studies [13].

A mathematical model taking into account both intra- and inter-clonal competition was developed to evaluate the different impacts of TCR-specific and TCR-nonspecific regulatory signals over T cells coming from a transplanted thymus in patients affected by DiGeorge Anomaly [35]. The main result of this stochastic model was the little importance of TCR-specific regulatory signals if a homeostatic case was considered. As previously said, there are many statistical measures developed by ecological studies on species richness or species overlap. Many of these non-parametric models have been used in TCR repertoire diversity with many limitations and very little insight. To possibly overcome these problems, a parametric model, based on a multivariate Poisson-lognormal distribution, was suggested and tested on transgenic mice populations [134]. The particular idea of Poisson-lognormal was based on previous studies indicating that mixtures of Lognormal distributions might be good estimation of clonal size distributions [145]. A follow up on similar ideas has been made some years later (2013) by [74], where the observed receptor counts were modelled by a multivariate Poisson abundance mixture (mPAM). The new idea proposed was a Bayesian parameter fitting model not based, as in previous studies, on the conditional posterior likelihood (conditioned on the number of observed species) but on the complete one, showing this technique to be more effective in modelling TCR count data. On a very different line of reasoning, a work for the statistical inference of the generation probability of TCRs was carried out [115]. The authors built a model capable of predicting the probability of being generated, for a given CDR3 region, by the primitive recombination process. The typical CDR3 sequence could be produced by something like 30 different recombination events, suggesting that a deterministic approach would cause great systematic biases and correlation in the model which could not be overcome. In particular, they focused on nonproductive CDR3 sequences, allowing the description of the generation probability of the CDR3 sequences before any kind of functional selection.

A remarkable result is the importance of insertions in the enhancement of the diversity level: around 60% of the total diversity is due to insertions.

Following the interesting idea of 2003 explained above, some authors described the fractal properties of the human TCR repertoire in both health and after stem cell transplantation conditions [108]. In particular, they find a diminished (but still present) fractal distribution of TCR gene segments in patients after stem cell transplantation.

In 2013, a comprehensive overview of the existing technological and mathematical analysis for repertoire diversity studies was carried out by [148]. From the mathematical point of view, the authors drew a sketch of the recent mathematical efforts to unravel TCR repertoire diversity, in particularly the ones inspired by the advances in sequencing technologies such as [135, 168]. They cite anyway some recent works on population dynamics based on differential equations [129, 12], while for the various systems-biology approaches to signal processing and population survival, they refer to [68, 8].

A fascinating biological hypothesis, especially for its simplicity, accounting for the great difference between potential and actual TCR repertoire diversity was named as "evolutionary sloppiness" by [172]. The authors suggests that the theoretical potential diversity of the thymus could be attributed to a simple fact: reducing the amount of potential diversity could require much more energy-consuming check in the recombination process, energy that evolution simply decided not to waste. The authors also argue for the idea that cross-reactivity in response to different pathogens is indeed a rare event in naïve repertoires, but that it can become much less than rare in repertoires subject to successive infections. A different attempt to study the immune system has been made by bioinformatics. In particular, a review on immunological profiling and computational tools to analyse high-dimensional data was published in 2014 [86]. The main goal of the paper is to highlight the need of a more general framework able to integrate different data sets, and they achieve this goal while describing some of the main analysis and visualization tools for systems immunology.

The different shortcomings of recent mathematical methods for the diversity estimation that had been applied to TCR studies were discussed again in 2015 [93]. One of the main problems brought to light is the well-known "unseen species" problem. Some estimators were shown to be biased by sample sizes, as well as by what they define as the problem of "under-sampling". Parametric statistical methods were argued to be based on the need of an *a priori* frequency distribution which we actually do not know. In response to this, the authors developed a new estimator (DivE) which does not require any *a priori* assumption. DivE had been previously tested, against five different non-parametric estimators, on three independent datasets [94]. DivE was proved to be much more efficient in estimating diversity with different sample sizes. In particular, the other estimators

tended to increase the estimated diversity with the increase of sample sizes, while DivE was able to maintain an accurate estimate for all datasets. On the same line of [115], a second paper was published to infer somatic recombination processes of B cells [57]. The resulting similarities with the previous work on T cells were expected to be very common, given the existence of a unique underlying recombination process for both T and B cells. A very interesting aspect of these studies is given by the authors' suggestion for a sort of evolutionary adaptation of the generation process, due to the main results indicating that sequences with higher probability to be produced are also the ones with higher probability to pass the selection process in the thymus. Besides the many similarities, there was an important difference between T- and B-cell repertoires: the former are much less diverse than the latter ones, due to the lower number of insertions in the T-cell recombination machinery.

A different study, defined by its authors as the first of its kind, focused on in-depth analysis of $CD8^+$ T-cell repertoire at the single level lineages [54]. In particular the authors aimed at a deeper understanding of the dynamics of $CD8^+$ T-cell repertoire upon vaccination of a particular attenuated yellow fever virus vaccine. High-throughput data and an algorithm based on the well-known Fisher exact statistical test were the basis for this study. Thanks to these techniques, the authors were able to identify 2000 different clones, 12% of which was then detected in the long-term memory compartment. Different authors also focused on long-term maintenance of human T cells, with a particular focus to the naïve repertoire [157]. The study aimed at assessing the spatial distribution of diversity in different lymphoid tissue sites. Results revealed an interesting tendency of the repertoire diversity: for individuals forty years old (or older), site-specific clonal expansions were detected and a minimal overlap among the different lymphoid tissues. These main results were suggested by the use of Simpson's diversity index, Shannon entropy and Jensen-Shannon divergence measure. In 2016 two papers were published, proving once more the increasing importance of bioinformatics and data-driven modelling in solving immunological problems [83, 165]. From the creation of a statistical method (Recon), based on maximum-likelihood theory, to the perspectives emerging from a particular workshop of the National Institute of Allergy and Infectious Diseases, the importance of computational data-driven modelling which enhance the quality of immune studies is becoming clearer and clearer.

A different mathematical approach to the study of clonotype diversity in the repertoire was used by [14]. The authors apply the idea of stochastic descriptors to a previous work [153], aiming at a deeper understanding of the survival probability distribution of a single clonotype emigrating from the thymus into the repertoire. In particular, two possible fates are shown, the first being extinction of the clone in the short-term in case of a too hostile

environment, while the second representing its long-term survival in the periphery. The authors also showed that the probability distribution of the maximum size obtained by the clonotype in the second case is bi-modal.

Finally, to conclude this journey through mathematics applied to the study of TCR repertoire diversity, and to give the reader the possibility to have a more general glance at the broad use of mathematics in immunology, it is important and necessary to cite the beautifully-written, latest review on the area of mathematical immunology [56]. Even though its focus on TCR diversity is very restricted, citing only a few of the latest contributions in this area such as [153, 17, 100], this review range over an incredible number of mathematical techniques, from agent-based modelling to eco-immunology, passing through Gillespie and Monte-Carlo algorithms, besides cellular automata, ordinary, partially and stochastic differential equations (ODEs, PDEs and SDEs), sensitivity analysis and principal component analysis (PCA). Model validation and parameter estimation are deeply discussed, and the literature review is organized by levels, from the molecular to the population one. As a mathematician and PhD student in systems immunology, I believe this review can be easily defined as a goldmine for whomever is researching in the area of mathematical and systems immunology.

Chapter 3

Mathematics for T-cell sampling

3.1 Abstract

Modern next generation sequencing (NGS) technologies allow us to sequence DNA or RNA from single cells. In particular, single-cell sequencing techniques enable the study of the DNA or RNA sequences of T-cell receptors (TCRs) from a sample, one cell at a time. The upscaling of fundamental properties from the sample to the whole repertoire remains one of the biggest mathematical challenges in systems immunology. This chapter focuses on the distribution of number of repeats of any particular TCR clonotype in a sample of T cells, trying to give some insights on the true clonal size distribution of a repertoire. We compute the mean number of T-cell extractions needed to find a repeat with a given probability. We give insights on the mathematical relation that binds the clonal size distribution in a repertoire with the one observed in a sample. Equal clonal sizes in the repertoire is the first hypothesis that we consider, although not biologically relevant. We then consider a different case where the number of T cells per clonotype class in the repertoire is a random variable with a geometric, Poisson or logarithmic distribution. A repertoire in which a small fraction of clones are expanded is also considered.

3.2 Introduction

The number of T cells circulating in an adult human body has been estimated to be approximately 4×10^{11} [81]. Each one of these T cells is able to express on its cell surface something like 30,000 T-cell receptors (TCRs), usually all being a clone of each other [163]. In the thymus, T cells are constantly accurately selected, based on their ability to bind with self-peptides expressed in association with major histocompatibility complex molecules (self-pMHC) [137], [81], [16], [161]. More details regarding this selection process can be found in Section 1.2. If we could arrange all the T cells in classes based on their TCR, the result would be a repertoire of classes, each one defined as "clonotype class". T

cells follow two different paths for peptide recognition depending on whether they belong to the $CD8^+$ type or to the $CD4^+$ one. The former type recognises peptides bound to MHC class I molecules while the second one, $CD4^+$ type, recognises peptides bound to MHC class II molecules [163], [150], [114]. Once TCR clonotypes have been defined, some very important questions naturally come up, such as how many TCR clonotypes are actually present in a human, mouse or other mammal immune systems? How many T cells does a clonotype class maintain? And what kind of clonal size distribution does a body maitain? [91], [24], [36], [156]. Nowadays, sequencing technologies do not allow us to directly answer these questions. Both chains (α and β) of each single TCR are created by a complex semi-random process called somatic recombination or *VDJ recombination* (see Section 1.4). Some estimates of the total possible number of different TCRs that could, in principle, be produced by this process are about 10^{15} [146], [121], [172], [115]. However, it is well known the paper showing how these estimates could never be achieved, as 10^{15} T cells would weigh about 500 kg [106]. We define here the number of distinct TCR clonotypes, N, as the total number of T cells divided by the mean number of cells per clonotype class. Equivalently, N can be seen as the product of the rate of release of new clonotypes from the thymus to the periphery, θ, times the mean lifetime of a clonotype in the periphery [100]. Some lower limits on the number of distinct TCR β chains in the repertoire have been given in the literature 4×10^6 [13], [85], [135], [168]. If each TCR β chain can combine with 25 α chains, then the number of possible distinct clonotypes in a human immune system should be at least 10^8 [130]. Even though direct estimates of TCR diversity have been made by PCR amplification of mRNA from pools of cells, it is necessary to keep in mind that numbers of mRNA vary from cell to cell and PCR amplification may depend on the particular TCR sequence. This problem makes PCR amplification a questionable technique for the measurement of clonal size distributions. A possibility to overcome this problem, therefore avoiding TCR-related biases, is given by single-cell measurements, where PCR and sequencing are performed on one cell at a time. However, these experiments are very expensive, allowing us to sequence only hundreds of cells from a single individual. This is the point where mathematical analyses can help, allowing estimates of diversity based on samples of small sizes [164], [135], [145], [93]. We consider the number of T cells per clonotype class to be a random variable with a given distribution. Once a random sample has been extracted from the repertoire, we want to study the random variable representing the number of copies, actually found in the sample, of a particular TCR. We approach this problem by making use of probability generating functions. The absence of most of the clonotypes in a small sample it is not surprising, given the enormous size of the repertoire itself. Besides, the majority of clonotypes that are

in the sample present no repeats. We show the subtle relationship between the observed distribution of clonal sizes in the sample and the theoretical distribution in the repertoire.

3.3 Sampling from a repertoire

What can be deduced from a sample of m cells taken from a repertoire of T cells if the total number of cells in the repertoire, S, is very large? Let us begin by describing the structure of the repertoire, which is divided into N subsets, called TCR clonotypes. Denote by n_i the number of cells of a clonotype labeled i. The index i runs from 1 to N, and $\sum_i n_i = S$ (see Figure 3.1). Typically S is known, but N and the n_i are not.

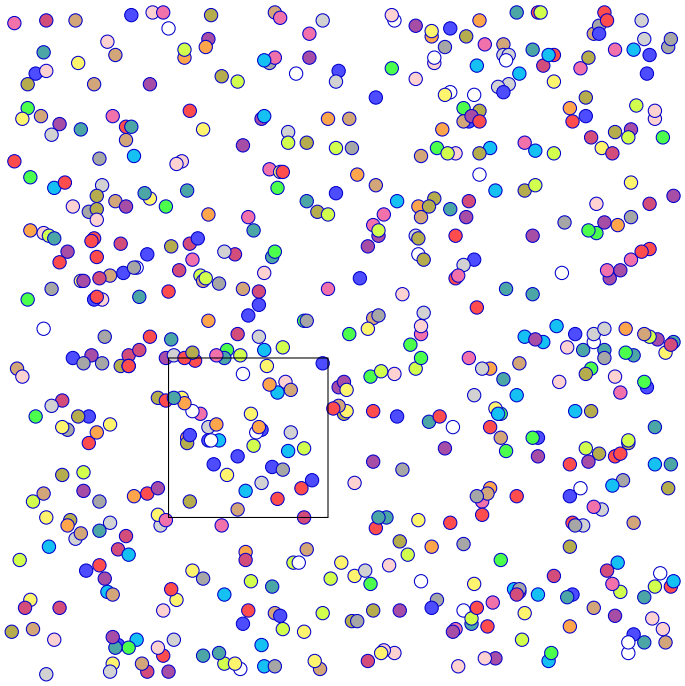

Figure 3.1: The repertoire contains S cells, divided up into N TCR clonotypes. Here, cells are represented by small coloured circles, a TCR clonotype is the set of cells of one colour, and a random sample of cells is represented by those circles inside the black square.

When the number of cells in the sample, m, is much smaller than the number of cells in the repertoire, S, and much smaller even than the number of TCR clonotypes, N, it is not obvious how to draw direct conclusions. On the other hand, some mathematical simplifications can be made. Let us consider one TCR clonotype, with label i. If $mn_i \ll S$ then, instead of the full expressions involving multivariate hypergeometric distributions, we can use the binomial approximations that

1. the probability that none of the m cells in the sample are of clonotype i is $\left(1 - \frac{n_i}{S}\right)^m$

2. the probability that exactly one of the m cells in the sample is of clonotype i is $m\frac{n_i}{S}\left(1-\frac{n_i}{S}\right)^{m-1}$.

3. the probability that exactly two of the m cells in the sample are of clonotype i is $\frac{1}{2}m(m-1)\frac{n_i}{S}\frac{n_i-1}{S}\left(1-\frac{n_i}{S}\right)^{m-2}$.

It is worth noting that here we use the quantity $n_i(n_i-1)$ rather then n_i^2 (as it would be expected from a binomial distribution) due to the small general values of n_i. In fact, this is trying to reproduce the sampling without replacement of the hypergeometric distribution that is not present in the binomial one. If $m \gg 1$ but $mn_i/S \ll 1$ then we can approximate the last expression by r_i, where

$$r_i = \frac{1}{2}\left(\frac{m}{S}\right)^2 n_i(n_i-1). \tag{3.1}$$

We say there is a repeat in the sample if two (or more) of the m cells are of the same clonotype. Let us consider a group of M identified clonotypes in the repertoire, with numbers of cells $n_1, n_2, \ldots, n_M$. How many repeats, of clonotypes in this group, will we see in our sample? If

$$r_i \ll 1 \qquad \forall i = 1, \ldots, M\ ,$$

so that the occurrences of repeats in distinct clonotypes can be taken as independent events, then

$$\mathbb{E}(\text{number of repeats of identified clonotypes}) = \sum_{i=1}^{M} r_i = \frac{1}{2}\left(\frac{m}{S}\right)^2 \sum_{i=1}^{M} n_i(n_i-1).$$

That is,

$$\mathbb{E}(\text{number of repeats of identified clonotypes}) = \frac{1}{2}\frac{m^2}{S^2} M\ \mathbb{E}(n_i(n_i-1))\ , \tag{3.2}$$

where the expectation is taken over the M clonotypes:

$$\mathbb{E}(n_i(n_i-1)) = M^{-1}\sum_{i=1}^{M} n_i(n_i-1)\ .$$

Repertoires can be constructed and sampled inside a simple computer program, where each clonotype is assigned a label i and values of n_i are assigned according to a probability distribution. We have constructed repertoires with uniform, geometric, Poisson, logarithmic and heterogeneous distributions of clonal sizes to verify the conclusions presented here.

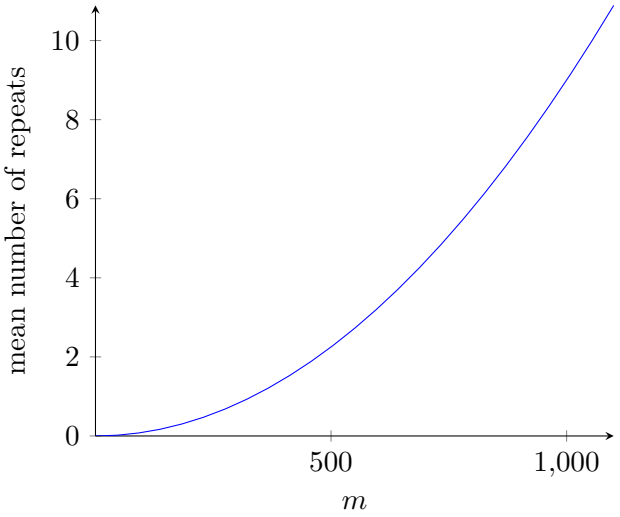

Figure 3.2: Mean number of repeats as a function of the number of cells in the sample, from a repertoire of $N = 10^5$ clonotypes and a geometric distribution of clonal sizes, with $\bar{n} = 10$.

3.4 The mean number of repeats

To find the mean number of repeats of *any clonotype from the repertoire* in the sample, we put $M = N$ in (3.2) and write $S = N\,\mathbb{E}(n_i)$, to obtain

$$\mathbb{E}(\text{number of repeats}) = \sum_{i=1}^{N} r_i = \frac{m^2}{2N} \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} \,. \tag{3.3}$$

The expression (3.3) is the product of the factor $\frac{m^2}{2N}$, that does not depend on the distribution of clonal sizes, and the factor $\frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2}$, that does. The latter can be written

$$\frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} = \frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} - \frac{1}{\mathbb{E}(n_i)} \,.$$

- If $n_i = \bar{n}$ for every i then

$$\frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} = 1 \qquad \text{and} \qquad \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} = 1 - \frac{1}{\bar{n}} \,.$$

- If n_i has a geometric distribution with mean $\bar{n}$ (that is, $\Pr(n_i \le k) = 1 - \left(1 - \frac{1}{\bar{n}}\right)^k$, $k = 1, 2, \ldots$) then

$$\frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} = 2 - \frac{1}{\bar{n}} \qquad \text{and} \qquad \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} = 2\left(1 - \frac{1}{\bar{n}}\right) \,.$$

See Figure 3.2.

3.5 Number of draws to find the first repeat

Let us consider the probability of finding no repeats in a sample of m cells. With r_i defined in (3.1), we approximate by $1 - r_i$ the probability that fewer than two cells of clonotype i are found in the sample, so that

$$\Pr(\text{no repeat in a sample of } m \text{ cells}) = \prod_{i=1}^{N} (1 - r_i) \ . \tag{3.4}$$

We can then write

$$\log\left(\Pr(\text{no repeat in a sample of } m \text{ cells})\right) = \sum_{i=1}^{N} \log(1 - r_i) \tag{3.5}$$

$$\simeq -\sum_{i=1}^{N} r_i \ , \tag{3.6}$$

assuming $r_i \ll 1$ for every i and applying Taylor approximation $\log(1 + x) = x + \mathcal{O}(x^2)$. Thus, we have

$$\Pr(\text{no repeat in a sample of } m \text{ cells}) = \exp\left(-\lambda\right) \ , \tag{3.7}$$

where

$$\lambda = \sum_{i=1}^{N} r_i = \frac{m^2}{2N} \frac{\mathbb{E}(n_i(n_i - 1))}{\mathbb{E}(n_i)^2} \tag{3.8}$$

is the mean number of repeats in a sample of m cells, as seen in Eq. 3.3.

How many cells do we need to sample in order to have a 50 percent chance of finding a repeat? Let this number be $m_{0.5}$. This implies $\lambda = \log 2$. Then, using (3.8),

$$m_{0.5}^2 = \frac{\mathbb{E}(n_i)^2}{\mathbb{E}(n_i(n_i - 1))} \, 2N \log 2 \ . \tag{3.9}$$

In the simplest case, when all clonotypes have the same number of cells, $\bar{n}$, we find $\Pr(\text{no repeats}) = \exp(-\frac{m^2}{2N}(1 - \frac{1}{\bar{n}}))$ and

$$m_{0.5} = \sqrt{\left(\frac{2N \log 2}{1 - \frac{1}{\bar{n}}}\right)} \ .$$

When the distribution of the number of cells per clonotype is geometric with mean $\bar{n}$, and the desired number is

$$m_{0.5} = \sqrt{\left(\frac{N \log 2}{1 - \frac{1}{\bar{n}}}\right)} \ .$$

See Figure 3.3. For a more general case, if we want to answer the question on how many cells we need to sample in order to have a certain probability p of finding a repeat, we should use the relationship

$$1 - p = e^{-\lambda} \quad \Rightarrow \quad \lambda = \log\left(\frac{1}{1 - p}\right) .$$

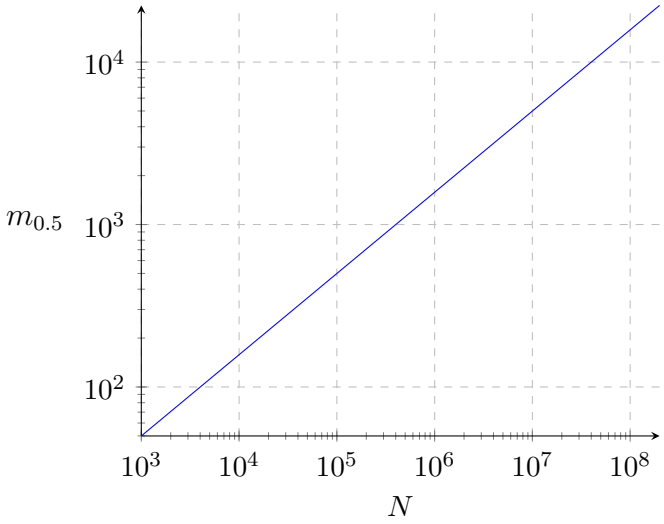

Figure 3.3: Mean number of cells that need to be sampled in order to have a 50 percent chance of one repeat, from a repertoire of N clonotypes and a geometric distribution of clonal sizes, with $\bar{n} = 10$.

3.6 Poisson distribution of number of repeats in a sample

Let k be the total number of repeats in a sample of m cells. For example, if 96 sequences are found once and 2 sequences are found twice, then $m = 100$ and $k = 2$. We have already seen $\Pr(k = 0)$ in (3.7). Let us consider the case $k = 1$:

$$\Pr(k = 1) = \sum_{i=1}^{N} r_i \prod_{\substack{j=1 \\ j \neq i}}^{N} (1 - r_j).$$

If $r_i \ll 1$ for every i then, from (3.4) and (3.7), we have $\displaystyle\prod_{\substack{j=1 \\ j \neq i}}^{N} (1 - r_j) \simeq \prod_{i=1}^{N} (1 - r_i) = e^{-\lambda}$ and, from (3.8),

$$\Pr(k = 1) = \lambda e^{-\lambda} .$$

The same argument works for all $k \ll m$, so that the number of repeats in a sample has approximately a Poisson distribution:

$$\Pr(\text{number of repeats is } k) = \frac{\lambda^k}{k!} e^{-\lambda} .$$

3.7 Estimating the size of the repertoire from one repeat

Suppose there is one repeat in a sample of m cells. We then use (3.8) to estimate N. Putting $\lambda = 1$ and, assuming a geometric distribution of clonal sizes, we conclude that

$$N = m^2 \, \frac{n - 1}{n} \simeq m^2 .$$

If we find one repeat per 100 cells, we estimate the number of distinct clonotype classes to be around 10^4. If we find one repeat per 1000 cells, we estimate that the size of the repertoire is 10^6. In practice, the estimate m^2 is likely to be conservative, because any clonal expansion will increase the number of observed repeats.

3.8 The observed distribution of clonal sizes

In this section, our goal is to find the probability distribution of the number of instances of k copies of a TCR in a random sample of m cells. Firstly, consider the point of view of one cell in the total of S cells in the repertoire. The probability, which we denote q, that this cell is one of the m cells in the sample is

$$q = \frac{\text{combinations of } S-1 \text{ elements in } m-1 \text{ places}}{\text{combinations of } S \text{ elements in } m \text{ places}} = \frac{\binom{S-1}{m-1}}{\binom{S}{m}} = \frac{m}{S}. \tag{3.10}$$

Next, let us define the Bernoulli random variable B:

$$\Pr(B=0) = 1-q \quad \text{and} \quad \Pr(B=1) = q, \qquad \text{where } q = \frac{m}{S}\ .$$

The probability generating function of B is

$$\phi_B(z) = \sum_{k=0}^{\infty} \Pr(B=k) z^k = 1 - q + qz\ . \tag{3.11}$$

If n_i is the number of cells of a clonotype labeled i, then the number of cells of type i in the sample is the random variable Y_i, which can be written

$$Y_i = B_1 + \cdots + B_{n_i}\ , \tag{3.12}$$

where B_j, $j = 1, \ldots, n_i$ are random variables with the same distribution as B. With the approximation that the B_j are independent random variables, the probability generating function of Y_i is (See Sec. 2.2.7)

$$\phi_{Y_i}(z) = \phi_B(z)^{n_i} = (1 - q + qz)^{n_i}. \tag{3.13}$$

In Appendix A we show that the random variable Y_i can be approximated as a binomial distribution with parameters n_i and $q = m/S$. Here we use though a slightly different approximation, for which the random variable Y_i is approximated with a binomial distribution with parameters n_i and $q = m/S$. It is easy to simulate these distributions and check that they overlap for large enough values of the parameter S. Now, let Y be the number of copies of a randomly-chosen clonotype found in the sample of m cells, which

can be 0 or any integer greater than 0 (up to $\min\{m, n_i\}$). To find the probability distribution of Y, we must take the distribution of values of n_i into account. Suppose that the probability generating function of the random variable n_i is $\phi_n(z)$. Then

$$\phi_Y(z) = \sum_k \Pr(n_i = k)(1 - q + qz)^k = \phi_n(1 - q + qz). \tag{3.14}$$

Let us recall the Taylor expansion of a function $f(x)$ around the point $x = 0$, defined as

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$$

where $f^{(k)}(0)$ represents the k^{th} derivative of the function calculated in $x = 0$. Therefore, the Taylor expansion of the general $\phi_Y(z)$ is

$$\phi_Y(z) = \sum_{k=0}^{\infty} \frac{\phi_Y^{(k)}(0)}{k!} z^k.$$

By definition of probability generating function, we can write

$$\phi_Y(z) = \sum_{k=0}^{\infty} \Pr(Y = k) z^k,$$

which results in

$$\Pr(Y = k) = \frac{\phi_Y^{(k)}(0)}{k!}. \tag{3.15}$$

It is important to notice that the random variable Y describes the probability, for a general clonotype in the repertoire, of being present ($k \geq 1$) or not being present ($k = 0$) in the sample. What is interesting, especially from an experimental point of view, is to rescale these probabilities so that they can describe the observed clonal size distribution. In other words, we want to focus on the probability that a certain clonotype is present in the sample with a certain number of copies $k \geq 1$, given that this clonotype has been seen in the sample. Let us define $p_k = \Pr(Y = k)$ and $s_k = \Pr(Y = k | Y \neq 0)$. It follows that, for $k \geq 1$, we have

$$s_k = \frac{\Pr(Y \neq 0 | Y = k \geq 1) \Pr(Y = k)}{\Pr(Y \neq 0)} = \frac{\Pr(Y = k)}{\Pr(Y \neq 0)} = \frac{p_k}{1 - p_0}. \tag{3.16}$$

These probabilities represent the observed distribution of clonal sizes; that is, the histogram that is obtained by plotting number of TCRs versus number of cells in the sample. We can consider them as the probabilities of the random variable Y^{obs}, defined as the number of copies of a randomly-chosen clonotype actually seen in the sample. It is also of interest to try to say something about the probability generating function of Y^{obs}. Dividing $\phi_Y(z)$ by $1 - p_0$ we obtain

$$\frac{\phi_Y(z)}{1 - p_0} = \frac{p_0}{1 - p_0} + \frac{p_1}{1 - p_0} z + \frac{p_2}{1 - p_0} z^2 + \cdots = s_0 + s_1 z + s_2 z^2 + \cdots$$

where s_k are defined in (3.16) for $k \geq 1$. What we want though, for our random variable Y^{obs}, is $s_0 = 0$. Therefore, we obtain

$$\phi_{Y^{obs}}(z) = \frac{\phi_Y(z)}{1-p_0} - s_0 = \frac{\phi_n(1-q+qz) - \phi_n(1-q)}{1-\phi_n(1-q)}. \tag{3.17}$$

Since $\phi_n(1-q)$ is a constant, this last result shows that the distribution of a generic clonotype class in the repertoire is transferred, with different parameters, to the distribution of the same clonotype class in the sample. We will validate this general result in the following sections based on specific cases of clonal size distributions. In particular, the general property $\mathbb{E}(X) = \left[\phi_X^{(1)}(z)\right]_{z=1}$ for any general r.v. X brings us to

$$\mathbb{E}(Y^{obs}) = \left[\phi_{Y^{obs}}^{(1)}(z)\right]_{z=1} = \left[\frac{\phi_n^{(1)}(1-q+qz)}{1-\phi_n(1-q)}\right]_{z=1} = \frac{q\mathbb{E}(n_i)}{1-\phi_n(1-q)}. \tag{3.18}$$

Let us try to give now a different interpretation of (3.15). Recalling (3.14), we can prove that

$$\phi_Y^{(k)}(z) = q^k\ \phi_n^{(k)}(1-q+qz).$$

In particular, we obtain

$$\phi_Y^{(k)}(0) = q^k\ \phi_n^{(k)}(1-q).$$

This gives us a different interpretation of (3.15)

$$\Pr(Y=k) = \frac{q^k \phi_n^{(k)}(1-q)}{k!}, \tag{3.19}$$

in particular when we are dealing with $q \ll 1$. In fact, in this case, we can write

$$\phi_n^{(k)}(1-q) \simeq \phi_n^{(k)}(1) = \mathbb{E}\left[(n_i)_k\right],$$

where $\mathbb{E}\left[(n_i)_k\right]$ represents the k^{th} factorial moment of the random variable n_i. In order to check the validity of the last equation, let us consider a general random variable X. Its probability generating function $\phi_X(z) = \mathbb{E}(z^X)$ can be differentiated k times obtaining

$$\phi_X^{(k)}(z) = \mathbb{E}[X(X-1)\dots(X-k+1)z^{X-k}],$$

resulting in

$$\phi_X^{(k)}(1) = \mathbb{E}[X(X-1)\dots(X-k+1)] = \mathbb{E}\left[(n_i)_k\right]. \tag{3.20}$$

To conclude, (3.19) can be written as a function of the raw moments of the random variable n_i as

$$\Pr(Y=k) \simeq \frac{q^k \sum_{t=0}^{k} S(k,t)\mathbb{E}(n_i^t)}{k!}, \tag{3.21}$$

where the coefficients $S(k,t)$ represent the Stirling numbers of the first kind, that is the coefficients in the expansion

$$(x)_n = \sum_{k=0}^{n} S(n,k)x^k,$$

where $(x)_n$ denotes the falling factorial $(x)_n = x(x-1)(x-2)\dots(x-n+1)$.

Using (3.19), we can then observe the following relationship

$$\frac{s_{k+1}}{s_k} = \frac{p_{k+1}}{p_k} \simeq \frac{q}{k+1}\frac{\phi_n^{(k+1)}(1-q)}{\phi_n^k(1-q)} \simeq \frac{q}{k+1}\frac{\mathbb{E}\left[(n_i)_{k+1}\right]}{\mathbb{E}\left[(n_i)_k\right]}. \tag{3.22}$$

In particular, (3.22) can be further simplified in cases such as Poisson and geometric distributions for n_i. In fact, (3.22) would become $\frac{q}{k+1}\lambda$ and $\frac{q}{k+1}(k+1)(n-1)$ for the two cases respectively. As a last observation, we can use (3.19) to obtain

$$s_k = \frac{q^k \phi_n^{(k)}(1-q)}{k!}[1-\phi_n(1-q)]^{-1}. \tag{3.23}$$

3.9 Homogeneous cases

In this section we will consider a consistant clonal size distribution in the repertoire. We analyze four different distributions: constant, geometric, Poisson and logarithmic clonal size distribution. The constant case is somehow the basic type of possible discrete clonal size distributions, although unrealistic. The Poisson distribution has been chosen for its well-known ability in representing count data. The geometric and logarithmic distributions were chosen as being different discrete distributions from the Poisson one, and also for a different reason. In a work published in 2011 [111], the authors firstly consider a general birth and death process and discuss two possible approximations for the limiting conditional probability distribution (LCD) of the process, deriving them from a previous work [117]. Defining λ_n and μ_n as the birth and death rates of the process, the first approximation is given by

$$\begin{cases} \pi_1^{(1)} = \frac{1}{1+\sum_{n=2}^{\infty}\frac{\lambda_1\lambda_2\dots\lambda_{n-1}}{\mu_2\mu_3\dots\mu_n}}, \\ \\ \pi_n^{(1)} = \frac{\lambda_1\lambda_2\dots\lambda_{n-1}}{\mu_2\mu_3\dots\mu_n}\pi_1^{(1)} \quad \text{for} \quad n \geq 2. \end{cases} \tag{3.24}$$

The second approximation is given instead by

$$\begin{cases} \pi_1^{(2)} = \frac{1}{1+\sum_{n=2}^{\infty}\frac{\lambda_1\lambda_2\dots\lambda_{n-1}}{\mu_1\mu_2\dots\mu_{n-1}}}, \\ \\ \pi_n^{(2)} = \frac{\lambda_1\lambda_2\dots\lambda_{n-1}}{\mu_1\mu_3\dots\mu_{n-1}}\pi_1^{(2)} \quad \text{for} \quad n \geq 2. \end{cases} \tag{3.25}$$

The authors apply then these approximations to a specific birth and death process. Here we focus on the process with birth and death rates $\lambda_n = \lambda n^k$ and $\mu_n = \mu n^k$. It is easy to show that the first approximation for these rates can be written as

$$\begin{cases} \pi_1^{(1)} = \left[Li_k(\lambda/\mu)\right]^{-1}, \\ \\ \pi_n^{(1)} = (\lambda/\mu)^{n-1}\left[n^k Li_k(\lambda/\mu)\right]^{-1} \quad \text{for} \quad n \geq 2, \end{cases} \tag{3.26}$$

where $Li_k(z)$ represents the polylogarithmic function of order k defined as

$$Li_k(z) = \sum_{i=1}^{+\infty} \frac{z^i}{i^k}.$$

The second approximation can be written as

$$\begin{cases} \pi_1^{(2)} = 1 - (\lambda/\mu), \\ \\ \pi_n^{(2)} = (\lambda/\mu)^{n-1}[1 - (\lambda/\mu)] \quad \text{for} \quad n \geq 2. \end{cases} \tag{3.27}$$

As can be seen, the second approximation suggests that a birth and death process with rates of the kind $\lambda_n = \lambda n^k$ and $\mu_n = \mu n^k$ would reach an LCD closely related to geometric distribution, which could be seen as the distribution of the clonal size of a particular clonotype subject to the previously explained process. From here, and from the relation between the logarithmic distribution and the polylogarithmic one, we decided to explore geometric and logarithmic clonal size distributions.

3.9.1 Constant clonal size distribution

Let us consider a repertoire of N different clonotype classes and let us suppose that $n_i \equiv n \; \forall i \in \{1, 2, \dots N\}$. It follows that

$$\phi_n(z) = z^n, \tag{3.28}$$

so that

$$\phi_Y(z) = (1 - q + qz)^n. \tag{3.29}$$

Considering the known property of probability generating functions, for which

$$\mathbb{E}(Y) = \phi_Y^{'}(1), \tag{3.30}$$

we obtain that $\mathbb{E}(Y) = nq$. Moreover, applying (3.15) we obtain

$$\Pr(Y = k) = \begin{cases} (1-q)^n, & \text{for } k = 0 \\ \\ \dfrac{(n)_k q^k (1-q)^{n-k}}{k!} = \dbinom{n}{k} q^k (1-q)^{n-k}, & \text{for } k \geq 1 \end{cases}$$

where $(n)_k = n(n-1)\dots(n-k+1)$.

3.9.2 Geometric clonal size distribution

Let us now consider the case in which n_i has a geometric distribution with mean n. It is worth noting that we have to consider the geometric distribution with support on the set

$\{k = 1, 2, 3, \ldots\}$ given that we are working on clones which are present in the repertoire (that is $k > 0$). Thus

$$\Pr(n_i = k) = \frac{1}{n}(1 - \frac{1}{n})^{k-1}.$$

It follows that

$$\phi_n(z) = \frac{z}{n - (n-1)z}, \tag{3.31}$$

so that

$$\phi_Y(z) = \frac{1 - q + qz}{n - (n-1)(1 - q + qz)} = \frac{1 - q + qz}{1 + (n-1)q(1-z)}. \tag{3.32}$$

Applying (3.15) we obtain

$$\Pr(Y = k) = \begin{cases} \dfrac{1-q}{1+q(n-1)}, & \text{for } k = 0 \\ \dfrac{(n-1)^{k-1} n q^k}{(1+q(n-1))^{k+1}}, & \text{for } k \geq 1 \end{cases}$$

The following recursive formula holds

$$p_k = p_1 \gamma^{k-1} \qquad k \geq 1, \tag{3.33}$$

where

$$\gamma = \frac{(n-1)q}{1+q(n-1)}. \tag{3.34}$$

To understand the distribution of Y, let us factorize p_1 as

$$p_1 = \theta p,$$

where $p = (1 + q(n-1))^{-1}$ and

$$\theta = \frac{nq}{1+q(n-1)}.$$

Therefore $\gamma = 1 - p$, resulting in

$$p_k = \theta p (1-p)^{k-1} \qquad k \geq 1. \tag{3.35}$$

Being $q < 1$, it follows that $\theta < 1$. An important observation comes from (3.30) for which, in this case too, the expected value of the random variable Y is $\mathbb{E}(Y) = nq$.

Applying (3.16), we obtain

$$s_k = \frac{\theta p (1-p)^{k-1}}{1 - p_0} = p(1-p)^{k-1}. \tag{3.36}$$

We conclude that the observed distribution of clonal sizes (the histogram that is obtained by plotting number of TCRs versus number of cells) is geometric with parameter p, resulting in a mean of $1 + (n-1)q$. This validate, as previously discussed, the relation (3.17),

that is the property for which the distribution of a generic clonotype class in the repertoire is transferred, with different parameters, to the same clonotype class found in the sample. It is worth noting the following relation:

$$s_{k+1} = \gamma s_k. \tag{3.37}$$

This is a particular property of all positive geometric distributions. See Figure 3.4 for a particular case of this relation. For both constant and geometric unimodal cases, see Figure 3.5.

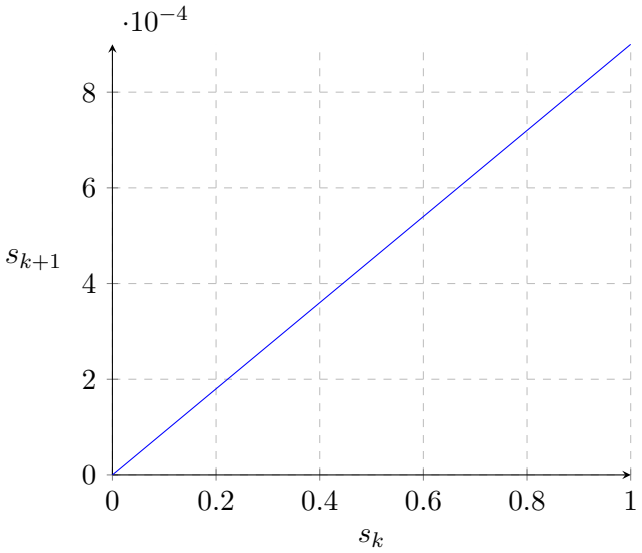

Figure 3.4: Relation between s_{k+1} and s_k, with $q = 0.0001$ and $n = 10$. See (3.37).

3.9.3 Poisson clonal size distribution

Let us now consider the case in which n_i has a Zero-Truncated Poisson distribution (ZTP) with mean n. It is worth noting that we have to consider this distribution because of its support on the set $\{k = 1, 2, 3, ...\}$ given that we are working on clones which are present in the repertoire (that is a Poisson distribution with $k > 0$). Thus, we have

$$\Pr(n_i = k) = \frac{\lambda^k}{(e^k - 1)\, k!}.$$

We need the expected value of the ZTP(λ) to be equal to n:

$$n = \frac{\lambda e^\lambda}{e^\lambda - 1}. \tag{3.38}$$

In order to find an expression for our parameter λ, we rearrange (3.38) and we multiply both sides by e^{-n}, in order to obtain

$$-ne^{-n} = e^{\lambda - n}(\lambda - n),$$

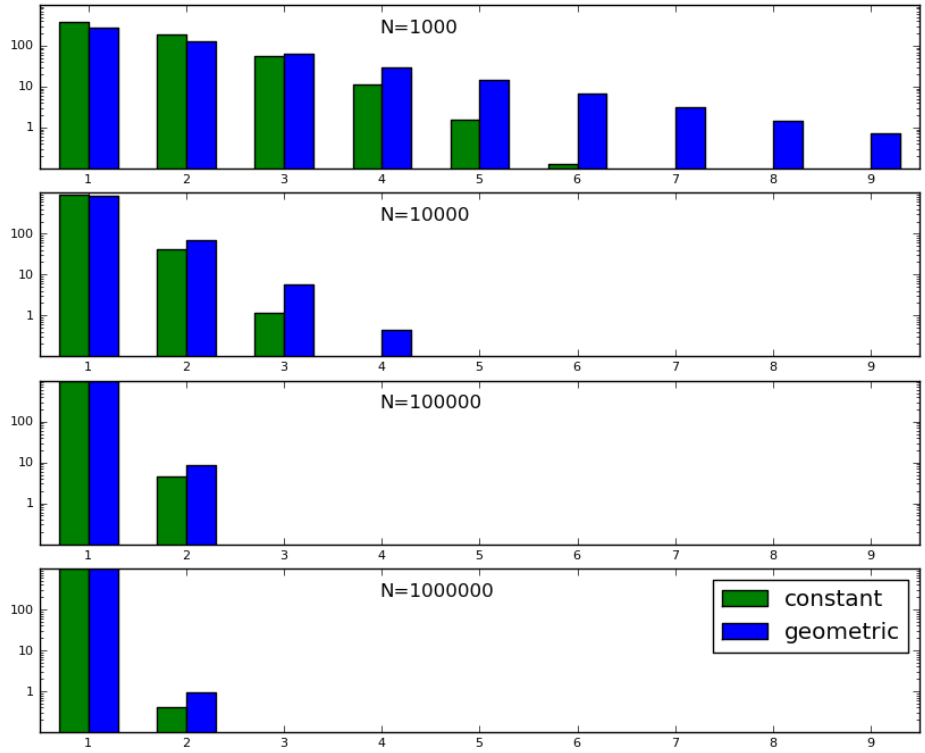

Figure 3.5: Observed clonal size distribution in a sample of 1000 cells, from repertoires containing different numbers of clones, N. A "constant" repertoire means that there are 10 cells of each clonotype. In a "geometric" repertoire, the number of cells in each clonotype is drawn from a geometric distribution with mean 10.

that in turn gives the solution

$$\lambda = n + W\left(-\frac{n}{e^n}\right), \tag{3.39}$$

where $W(\cdot)$ represents the Lambert W function, defined as the set of functions for which the solution of $ze^z = f(z)$ is $x = W(f(z))$. In particular, here we have

$$e^{\lambda - n}(\lambda - n) = f(\lambda - n), \quad \text{with} \quad f(\lambda - n) = -ne^{-n}.$$

It is important to notice that for $-\frac{1}{e} < x < 0$ the function $W(\cdot)$ is double-valued taking two possible branches: $W_0(\cdot)$ or $W_{-1}(\cdot)$. Therefore, we need to understand which one to use so that we can obtain uniqueness for the solution (3.39). For $-\frac{1}{e} < x < 0$, we have $-1 < W_0(\cdot) < 0$ and $-\infty < W_{-1}(\cdot) < -1$. Substituting (not inside the W function) (3.38) in (3.39), we obtain

$$W\left(-\frac{n}{e^n}\right) = -\frac{\lambda}{e^\lambda - 1}.$$

For $\lambda > 0$, we have $-1 < -\dfrac{\lambda}{e^{\lambda}-1} < 0$, indicating $W_0(\cdot)$ as the right branch to be chosen.

Thus,

$$\Pr(n_i = k) = \frac{\lambda^k}{k!\ (e^k - 1)}, \tag{3.40}$$

and

$$\phi_{n_i}(z) = \frac{e^{\lambda z} - 1}{e^{\lambda} - 1}, \tag{3.41}$$

with λ given by (3.39). We can then find $\phi_Y(z)$ a

$$\phi_Y(z) = \frac{e^{\lambda(1-q+qz)} - 1}{e^{\lambda} - 1}. \tag{3.42}$$

It follows that

$$\Pr(Y = 0) = \frac{e^{\lambda(1-q)} - 1}{e^{\lambda} - 1}$$

and

$$\Pr(Y = k) \equiv p_k = \frac{\lambda^k q^k e^{\lambda(1-q)}}{k!\ (e^{\lambda} - 1)} \quad k = 0, 1, 2, ...$$

The observed distribution is then defined by

$$s_k = \frac{(\lambda q)^k}{k!\ (e^{\lambda q} - 1)} \quad k = 1, 2, ... \tag{3.43}$$

showing that the observed distribution is also a ZTP distribution with parameter $\lambda^{'} = q\lambda$. This confirms the previously shown relation (3.17). It is worth noting the following relation:

$$s_{k+1} = \frac{\lambda q}{k+1} s_k. \tag{3.44}$$

This is a particular property of all ZTP distributions.

3.9.4 Logarithmic clonal size distribution

Let us now consider the case in which n_i has a logarithmic distribution with mean n. Thus, we need the expected value of the Log(p) to be equal to n:

$$-\frac{1}{\ln(1-p)} \frac{p}{1-p} = n. \tag{3.45}$$

In order to find an expression for our parameter p, we define $x = 1 - p$ and we rearrange (3.45), obtaining

$$x \ln(x) + \frac{1-x}{n} = 0.$$

This gives

$$\frac{1 + nx\left[\ln(x) - \frac{1}{n}\right]}{n} = 0,$$

which in turns can be seen as

$$x \ln(xe^{-\frac{1}{n}}) = -\frac{1}{n}.$$

Multiplying both sides by $e^{-\frac{1}{n}}$, we obtain

$$xe^{-\frac{1}{n}}\ln(xe^{-\frac{1}{n}}) = -\frac{1}{n}e^{-\frac{1}{n}},$$

that can be written as

$$e^{\ln(xe^{-\frac{1}{n}})}\ln(xe^{-\frac{1}{n}}) = -\frac{1}{n}e^{-\frac{1}{n}}.$$

The solution of this equation is given by

$$\ln(xe^{-\frac{1}{n}}) = W\left(-\frac{1}{n}e^{-\frac{1}{n}}\right),$$

where $W(\cdot)$ represents the Lambert W function, defined as the function for which the solution of $xe^x = a$ is $x = W(a)$. In particular, we obtain

$$x = e^{\frac{1}{n}+W\left(-\frac{1}{n}e^{-\frac{1}{n}}\right)},$$

that gives us the final relation

$$p = 1 - e^{\frac{1}{n}+W\left(-\frac{1}{n}e^{-\frac{1}{n}}\right)}. \tag{3.46}$$

As previously stated in Section 3.9.3, for $-\frac{1}{e} < x < 0$ the function $W(\cdot)$ is double-valued taking two possible branches: $W_0(\cdot)$ or $W_{-1}(\cdot)$. Therefore, we need to understand which one to use so that we can obtain uniqueness for the solution (3.46). For $-\frac{1}{e} < x < 0$, we have $-1 < W_0(\cdot) < 0$ and $-\infty < W_{-1}(\cdot) < -1$. Substituting (not inside the W function) (3.45) in (3.46), we obtain

$$W\left(-\frac{1}{n}e^{-\frac{1}{n}}\right) = \frac{\ln(1-p)}{p}.$$

For $0 < p < 1$, we have $\frac{\ln(1-p)}{p} < -1$, indicating $W_{-1}(\cdot)$ as the right branch to be chosen. Thus,

$$\Pr(n_i = k) = \frac{-1}{\ln(1-p)}\frac{p^k}{k}, \tag{3.47}$$

and

$$\phi_{n_i}(z) = \frac{\ln(1-pz)}{\ln(1-p)}, \tag{3.48}$$

with p given by (3.46). Using (3.23), we can then find

$$s_k = \frac{-1}{\ln(1-a)}\frac{a^k}{k} \text{ with } a = \frac{pq}{1-p(1-q)} \quad k = 1, 2, ... \tag{3.49}$$

showing that the observed distribution is also a logarithmic distribution with parameter $a = \frac{pq}{1-p(1-q)}$. This confirms the previously shown relation (3.17). It is worth noting the following relation:

$$s_{k+1} = \frac{ak}{k+1}s_k. \tag{3.50}$$

This is a particular property of all logarithmic distributions.

3.10 Heterogeneous cases: expansion of a subset of the repertoire

In this section we assume that a fraction $f \ll 1$ of clones undergoes expansion following an immune response to a certain infection. We call E the expanded part and U the unexpanded one. We consider two special cases: constant and geometric repertoire.

3.10.1 Constant clonal size distribution: expansion case

Let us consider the case in which the random variable n_i of T cells of type i is

$$n_i \equiv \begin{cases} n, & \text{if clone } i \text{ is in the unexpanded part } U \\ \\ n\alpha, & \text{if clone } i \text{ is in expanded part } E. \end{cases}$$

Recalling (3.14), we need to focus on finding the probability generating function ϕ_n of the random variable n_i. By definition,

$$\phi_n(z) = \sum_{k=0}^{\infty} \Pr(n_i = k) z^k. \tag{3.51}$$

Thus we have to find $\Pr(n_i = k)$. From the theory we have that

$$\Pr(n_i = k) = \Pr(n_i = k \,|\, i \in E) \Pr(i \in E) \; + \; \Pr(n_i = k \,|\, i \in U) \Pr(i \in U).$$

In our case, we have

$$\Pr(n_i = k \mid i \in D) = \begin{cases} 1, & \text{for } \{D = E \text{ and } k = n\alpha\} \text{ or for } \{D = U \text{ and } k = n\} \\ \\ 0, & \text{otherwise} \end{cases}$$

and

$$\Pr(i \in D) = \begin{cases} f, & \text{for } D = E \\ \\ 1 - f, & \text{for } D = U. \end{cases}$$

At this point, we can write (3.51) as

$$\phi_n(z) = (1 - f) z^n + f z^{n\alpha}. \tag{3.52}$$

Therefore, the probability generating function of the random variable Y is

$$\phi_Y(z) = \phi_n(1 - q + qz) = (1 - f)(1 - q + qz)^n + f(1 - q + qz)^{n\alpha}. \tag{3.53}$$

Following the same procedure as in Section 3.9.2, we can find that

$$p_k = \binom{n}{k}(1-f)q^k(1-q)^{n-k} + \binom{n\alpha}{k} f q^k (1-q)^{n\alpha-k} \qquad k \geq 0 \tag{3.54}$$

and

$$s_k = \frac{p_k}{1-p_0} \qquad k \geq 1.$$

The $\mathbb{E}(Y)$ can be easily found following (3.30), obtaining $\mathbb{E}(Y) = nq[1+(\alpha-1)f]$. Therefore,

$$\mathbb{E}(Y^{obs}) = \frac{nq(1+(\alpha-1)f)}{1-p_0}. \tag{3.55}$$

3.10.2 Geometric clonal size distribuion: expansion case

We consider the case in which the random variable n_i of T cells of type i is:

$$n_i \sim \begin{cases} \text{Geom}(\frac{1}{n}), & \text{if clone } i \text{ is in the unexpanded part } U \\ \\ \text{Geom}(\frac{1}{n\alpha}), & \text{if clone } i \text{ is in expanded part } E. \end{cases}$$

We now follow the same procedure as in Section 3.10.1. In this case, we have

$$\Pr(n_i = k \mid i \in D) = \begin{cases} \left(1-\frac{1}{n\alpha}\right)^{k-1}\left(\frac{1}{n\alpha}\right), & \text{for } D = E \\ \\ \left(1-\frac{1}{n}\right)^{k-1}\left(\frac{1}{n}\right), & \text{for } D = U \end{cases}$$

and

$$\Pr(i \in D) = \begin{cases} f, & \text{for } D = E \\ \\ 1-f, & \text{for } D = U. \end{cases}$$

Therefore we obtain

$$\Pr(n_i = k) = f\left(1-\frac{1}{n\alpha}\right)^{k-1}\left(\frac{1}{n\alpha}\right) + (1-f)\left(1-\frac{1}{n}\right)^{k-1}\left(\frac{1}{n}\right).$$

At this point, we have

$$\phi_n(z) = f\sum_{k=1}^{\infty}\left(1-\frac{1}{n\alpha}\right)^{k-1}\left(\frac{1}{n\alpha}\right) z^k + (1-f)\sum_{k=1}^{\infty}\left(1-\frac{1}{n}\right)^{k-1}\left(\frac{1}{n}\right) z^k.$$

We see now that

$$\phi_n(z) = f\,\phi_e(z) + (1-f)\,\phi_u(z), \tag{3.56}$$

where e and u are random variables with a geometric distribution with parameters $\frac{1}{n\alpha}$ and $\frac{1}{n}$ respectively.

Therefore, the probability generating function of the random variable Y is

$$\phi_Y(z) = \phi_n(1-q+qz) = f\ \phi_e(1-q+qz)\ +\ (1-f)\ \phi_u(1-q+qz) \tag{3.57}$$

where

$$\phi_e(1-q+qz) = \frac{1-q+qz}{n\alpha-(n\alpha-1)(1-q+qz)} = \frac{1-q+qz}{1-q(n\alpha-1)(z-1)}$$

and

$$\phi_u(1-q+qz) = \frac{1-q+qz}{n-(n-1)(1-q+qz)} = \frac{1-q+qz}{1-q(n-1)(z-1)}.$$

Following the main ideas of Section 3.9.2, we find

$$\Pr(Y=k) = \begin{cases} f\ \dfrac{1-q}{1+q(n\alpha-1)}\ +\ (1-f)\ \dfrac{1-q}{1+q(n-1)}, & \text{for } k=0 \\ f\ \dfrac{(n\alpha-1)^{k-1}\ n\alpha\ q^k}{(1+q(n\alpha-1))^{k+1}}\ +\ (1-f)\ \dfrac{(n-1)^{k-1}\ n\ q^k}{(1+q(n-1))^{k+1}}, & \text{for } k\geq 1. \end{cases}$$

We can find s_k as defined in (3.16). Moreover, applying (3.30), we find

$$\mathbb{E}(Y) = nq(1+(\alpha-1)f). \tag{3.58}$$

Therefore,

$$\mathbb{E}(Y^{obs}) = \frac{nq(1+(\alpha-1)f)}{1-p_0}. \tag{3.59}$$

Case $\mathbf{q \ll 1}$

Let us consider the case in which $S \gg m$, that is $q \ll 1$. In this case, excluding all the parts proportional to q^2, the quantity $1-p_0$ can be approximated by

$$1-p_0\ \simeq\ \frac{1+q(n(\alpha+1)-2)-[f(1-q)+fq(n-1)+(1-f)(1-q)q(1-f)(n\alpha-1)]}{1+q(n(\alpha+1)-2)},$$

which turns out to be

$$1-p_0\ \simeq\ \frac{nq(1+(\alpha-1)f)}{1+q(n(\alpha+1)-2)}.$$

Therefore,

$$\mathbb{E}(Y^{obs})\ \simeq\ 1+q(n(\alpha+1)-2)\ \simeq\ 1+qn\alpha. \tag{3.60}$$

It is interesting to notice that this quantity does not depend on the fraction f of expanded clones in the repertoire. In fact, this quantity is a good estimation of $\mathbb{E}(Y^{obs})$ for $f \geq 0.1$. See Figure 3.6.

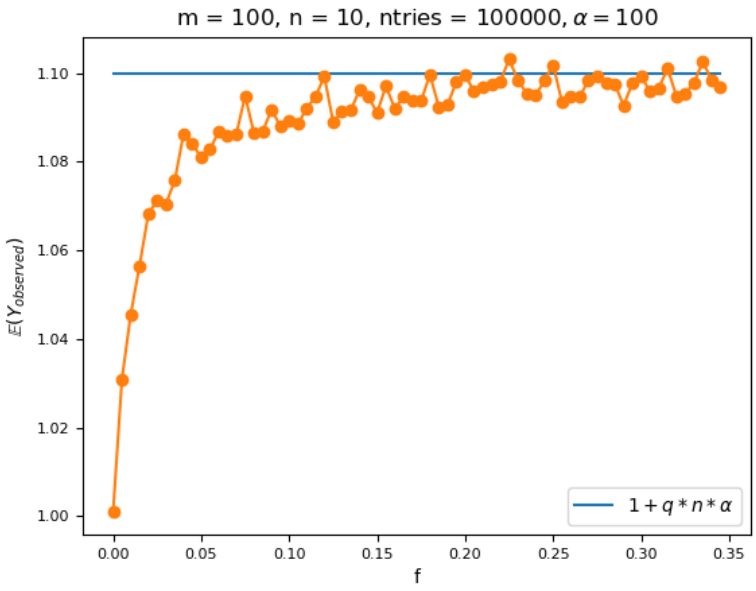

Figure 3.6: Expansion case with geometric clonal size distributions. Independence of $\mathbb{E}(Y^{obs})$ from f for $f \geq 0.1$. $S = 10^6$.

3.11 Analysis of the TCRβ repertoire of naive CD8$^+$ T cells

Suppose the repertoire, from which cells are sampled, contains a total of S cells that are shared among N TCRβ clones. That is, N is equal to the total number of distinct TCRβ sequences in the repertoire.

We use the letter i to denote a clone in the repertoire that consists of n_i cells. Thus $i = 1, 2, \ldots, N$ and $n_1 + n_2 + \cdots + n_N = S$. The mean clonal size is denoted by $\bar{n}$. It is equal to S/N, the mean number of cells per clone. Three types of hypothesis are as follows:

(i) that each individual clone has the same number of cells;

(ii) that the clonal sizes follow a simple geometric distribution, where the probability of finding clones with small size is higher than that of finding large clones;

(iii) that there are two types of clones in the repertoire, the majority of clones made up of only a few cells, and a small minority of clones that contain many cells.

Suppose that a sample of m cells is taken and the TCRβ of each of the cells is sequenced. We define q (as in Section 3.8) to be the probability that one cell, randomly-chosen from the total of S cells, is part of the sample of size m:

$$q = \frac{m}{S}.$$

Let m_0 be the number of distinct sequences found in the sample, and let m_1 be the number of sequences found only once in the sample. If m_k, $k = 1, 2, \cdots$ is the number of sequences found once, twice, ... then

$$m_0 = m_1 + m_2 + m_3 + \cdots \qquad \text{and} \qquad m = m_1 + 2m_2 + 3m_3 + \cdots$$

3.11.1 The mean clonal size of the CD8$^+$ GP33$^+$ subset

Here we analyse the single cell data from our QuanTI collaborators on the CD8$^+$ GP33$^+$ subset, where GP33$^+$ is a specific LCMV (Lymphocytic Choriomeningitis Virus) epitope. The value of S is the total number of GP33$^+$ cells, estimated to be 441 (BM) or 2293 (SP+LN). Thus, with sample sizes m between 94 and 271, the value of q is between 0.04 and 0.12. Hypothesis (i) is not consistent with the data: if $n = 1$ then m is always equal to m_0; if $n = 2$ or larger, the predicted values of the ratio $\frac{m}{m_0}$ are larger than those observed. In fact, defining Y^{obs} as in Section 3.8, we can write

$$\mathbb{E}(Y^{\text{obs}}) = \sum_{k=1}^{\infty} k \Pr(Y^{\text{obs}} = k) \approx \sum_{k=1}^{\infty} k \frac{m_k}{m_0} = \frac{m}{m_0}.$$

That is, the mean of the ratio $\dfrac{m}{m_0}$ is $\dfrac{nq}{nq + \frac{1}{2}n(n-1)q^2}$. Thus, we consider hypothesis (ii).

Geometric

We first consider the geometric distribution of values of number of cells per clone, n_i. The statement that n_i has a geometric distribution with mean $\bar{n}$ is that

$$\Pr(n_i = k) = \frac{1}{\bar{n}} \left(1 - \frac{1}{\bar{n}}\right)^{k-1}, \qquad k = 1, 2, \ldots.$$

Note that $\bar{n} \geq 1$. The fraction of clones that consist of only one cell is

$$\Pr(n_i = 1) = \frac{1}{\bar{n}}.$$

If the distribution of values of n_i is geometric, then the distribution of the number of copies of each Tcrb sequence found in a sample of m cells is also geometric, with mean equal to $1 + (\bar{n} - 1)q$ (as shown in Section 3.9.2). Defining Y^{obs} as in Section 3.8, we can write

$$\mathbb{E}(Y^{\text{obs}}) = \sum_{k=1}^{\infty} k \Pr(Y^{\text{obs}} = k) = \sum_{k=1}^{\infty} k \frac{m_k}{m_0} = \frac{m}{m_0}.$$

That is, the mean of the ratio $\frac{m}{m_0}$ is $1 + (\bar{n} - 1)q$. Because the values of S and m are known, we obtain one estimate of $\bar{n}$ from each measured value of m_0:

$$\bar{n} = 1 + S\left(\frac{1}{m_0} - \frac{1}{m}\right). \tag{3.61}$$

We use (3.61) to estimate $\bar{n}$ in the GP33$^+$ repertoire, where $S = 2300$. There are five independent measurements, summarised:

Mouse 5: 271 sequences, 268 unique, so estimate $\bar{n} = 1.09$.

Mouse 6: 188 sequences, 186 unique, so estimate $\bar{n} = 1.13$.

Mouse 7: 128 sequences, 127 unique, so estimate $\bar{n} = 1.14$.

Mouse 10: 244 sequences, 240 unique, so estimate $\bar{n} = 1.16$.

Mouse 11: 165 sequences, 165 unique, so estimate $\bar{n} = 1.00$.

The mean of the estimated values of $\bar{n}$ is 1.10, with standard deviation 0.05.

Poisson

We next consider the hypothesis that the number of cells per clone, in the repertoire, has a Poisson distribution. The statement that n_i has a positive Poisson distribution with mean $\bar{n}$ is that

$$\Pr(n_i = k) = \frac{1}{e^\lambda - 1}\frac{\lambda^k}{k!} \qquad k = 1, 2, \ldots, \quad \text{and } \bar{n} = \frac{\lambda e^\lambda}{e^\lambda - 1}. \tag{3.62}$$

In this case, the distribution of the number of copies y_i of each TCR sequence found in a sample of m cells is also positive Poisson, with

$$\Pr(y_i = k) = \frac{1}{e^{\lambda q} - 1}\frac{(\lambda q)^k}{k!} \qquad k = 1, 2, \ldots.$$

The mean value of $\frac{m}{m_0}$ is $\frac{\lambda q e^{\lambda q}}{e^{\lambda q}-1}$ which, because $\lambda q \ll 1$, we can write as $\frac{m}{m_0} = 1 + \frac{1}{2}\lambda q + \frac{1}{4}(\lambda q)^2 + \cdots$ .. Retaining up to first order in λq,

$$\lambda = 2S\left(\frac{1}{m_0} - \frac{1}{m}\right). \tag{3.63}$$

For each mouse, we estimate λ using (3.63), then calculate $\bar{n}$ using (3.62).

Mouse 5: 271 sequences, 268 unique, so estimate $\bar{n} = 1.10$.

Mouse 6: 188 sequences, 186 unique, so estimate $\bar{n} = 1.14$.

Mouse 7: 128 sequences, 127 unique, so estimate $\bar{n} = 1.15$.

Mouse 10: 244 sequences, 240 unique, so estimate $\bar{n} = 1.17$.

Mouse 11: 165 sequences, 165 unique, so estimate $\bar{n} = 1.00$.

The mean of the estimated values of $\bar{n}$ is 1.11, with standard deviation 0.05.

We also estimate $\bar{n}$ from one sample of the GP33$^+$CD44$^+$ repertoire, where $S = 441$. Mouse 12: 94 sequences, 93 unique, so estimate $\bar{n} = 1.05$.

Thus the two cases, based on distributions of different shapes, give similar estimates of the mean number of cells per clone, corresponding to a repertoire in which $91\% \pm 4\%$ of clones consist of one cell only.

3.12 Discussion

The study of TCRs repertoire is of great importance nowadays. A very diverse repertoire could imply a much higher probability, for the immune system, of being able to properly react to certain foreign peptides. As discussed in the biological introduction, single-cell sequencing techniques allow us to obtain small samples from the totality of the repertoire of a particular individual. As it has been shown in this chapter, estimates of TCR diversity depend on the clonal size distribution in the repertoire, though small samples allow the simplifying approximation that random variables describing quantities of interest, such as the numbers of cells of different types in the sample, are independent. We have proved that the probability generating function of the distribution of clonal sizes in the sample can be seen as the composition of that of a Bernoulli random variable (that takes values 0 or 1) and that of the true distribution of clonal sizes in the repertoire that is being sampled from. In particular, we have expressed the relation between the clonal size distribution in the repertoire and the one in the sample for different distribution cases. Our work is motivated by studies of the repertoire of T cells in humans and mice. As previously discussed in the introduction, in this type of experiment, where mRNA is extracted from a pool of cells, it is difficult to obtain statistics of the number of cells of each clonotype (abundance data) that is free from biases. Single-cell TCR sequencing can eliminate biases but can, at present, only be carried out on a few hundred cells from one individual. This is, though, a great opportunity to apply mathematical techniques to such an important research area related to human health.

Chapter 4

VDJ recombination & Data analysis

4.1 Abstract and Introduction

Single-cell sequencing techniques probably represent the most reliable way to study the DNA or RNA sequences of T-cell receptors (TCRs) from a sample, one cell at a time. In the previous chapter we focused on the relation between the true clonal size distribution (i.e., the distribution of the repertoire) and the observed clonal size distribution (i.e., the distribution of the sample). This chapter focuses on a slightly different question, which can be expressed as follows: given the observed clonal size of a specific TCR clonotype class, what can we deduced about the true clonal size distribution of that particular clonotype class? As we will see, this question will provide new insights on the study of the diversity of a repertoire.

4.2 Data

The data considered in this study are part of a broader study on TCR diversity on naïve and LCMV (lymphocytic choriomeningitis virus) infected mice, of which only the naïve part has been published so far [71]. Samples from 10 specific-pathogen-free (SPF) B6 mice were analysed; five were not infected (naïve) and labelled as BA1, BA2, BA3, BA4 and BA5; the other five mice were subdivided in two subgroups, two being only immunized, labelled EF1 and EF2, and three being infected with LCMV and labelled EF3, EF4 and EF5. In particular, EF1 and EF2 were immunized with the epitope GP33 (from here their alternative label GP33i1 and GP33i2). The mice were obtained from breeding colonies at the Centre de Distribution, Typage et Archivage (CDTA, Orleans, France), and all the experiments were performed in accordance with the National European Commission guidelines for the care and handling of laboratory animals and were approved by the site

etichal review committee. In this section we present the data related only to V and J distributions and only related to GP33-specific repertoire. For each mouse, both 2D and 3D plots were created representing the V-J distribution. Tables with the actual data can also be found in Appendix B.

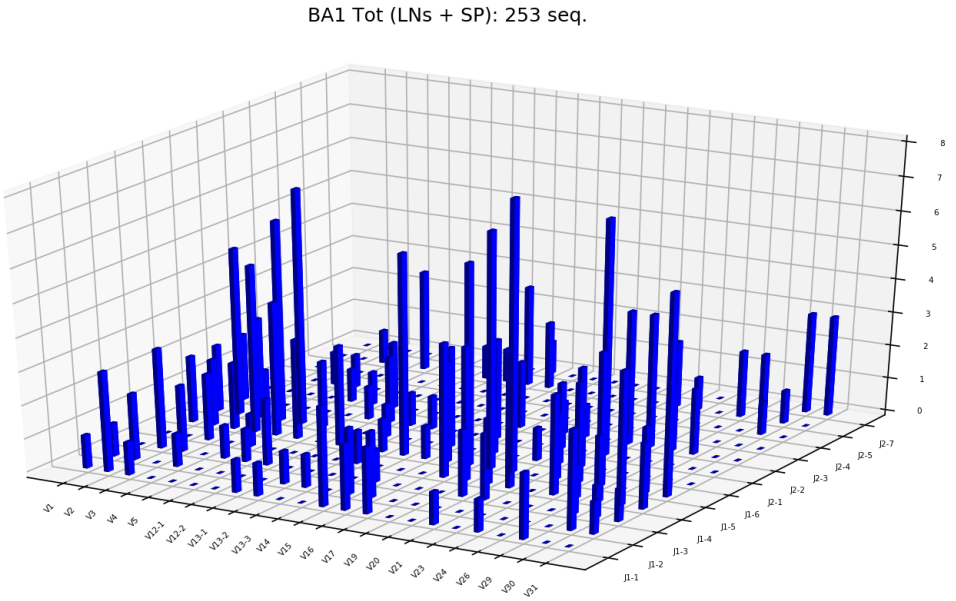

Figure 4.1: V-J plot for the uninfected mouse BA1.

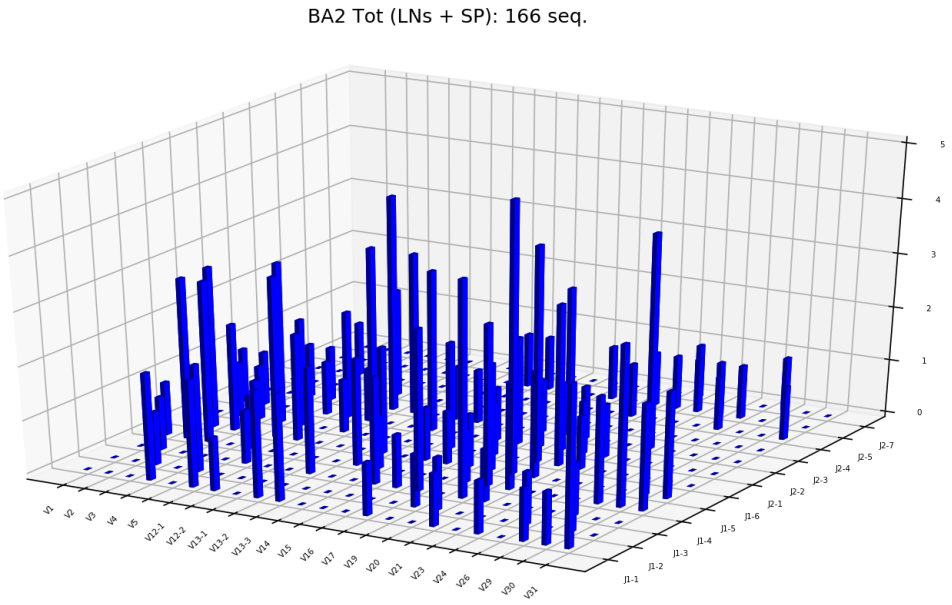

Figure 4.2: V-J plot for the uninfected mouse BA2.

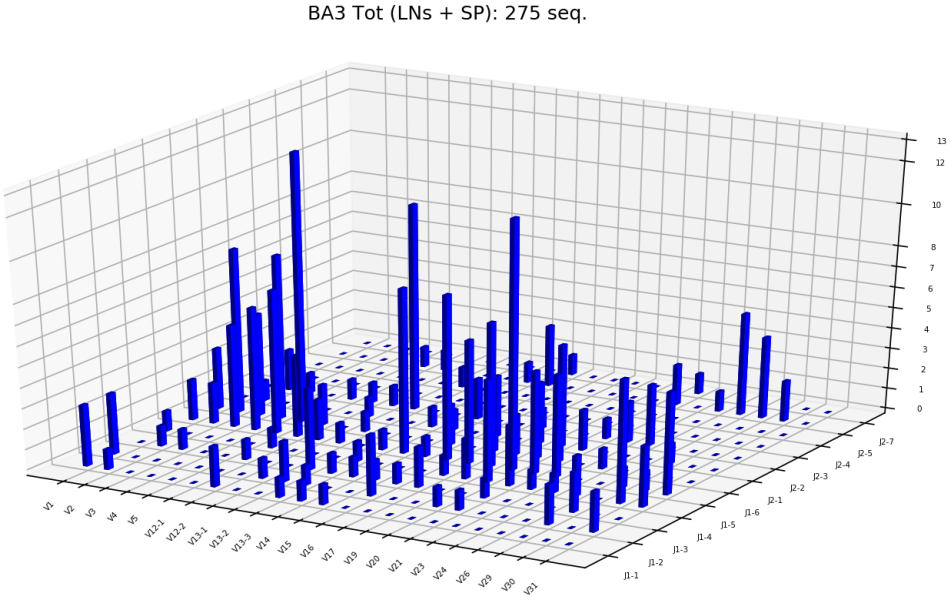

Figure 4.3: V-J plot for the uninfected mouse BA3.

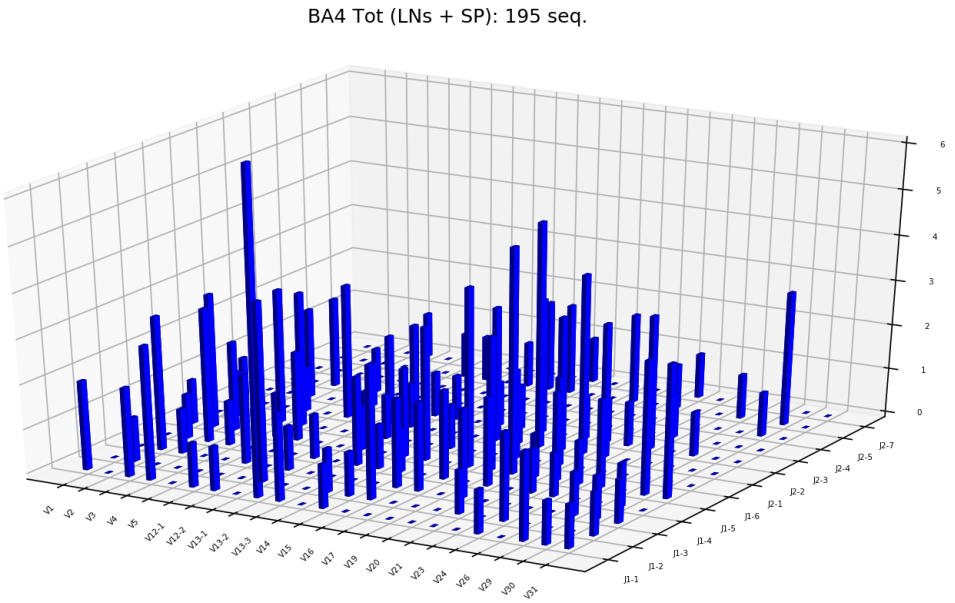

Figure 4.4: V-J plot for the uninfected mouse BA4.

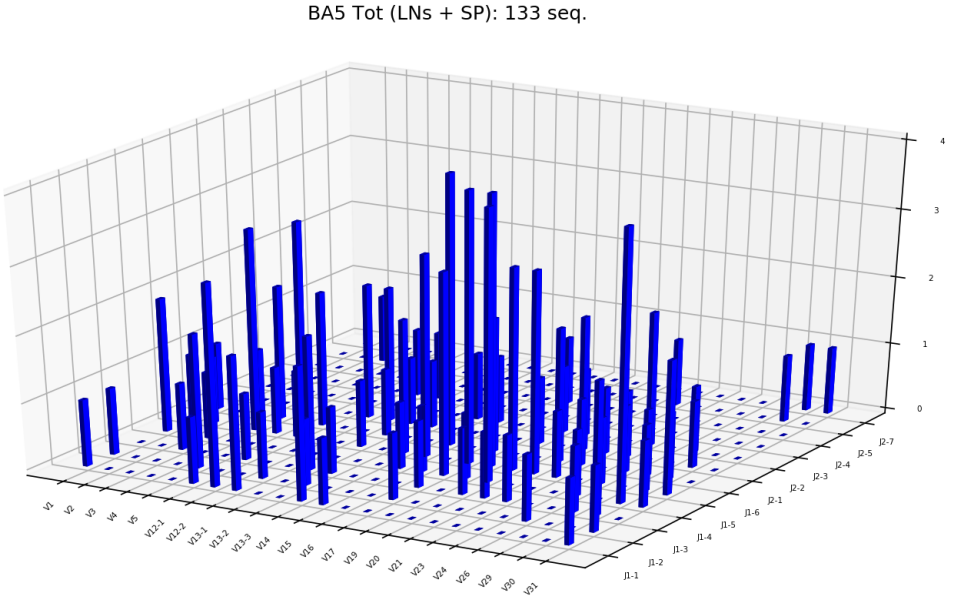

Figure 4.5: V-J plot for the uninfected mouse BA5.

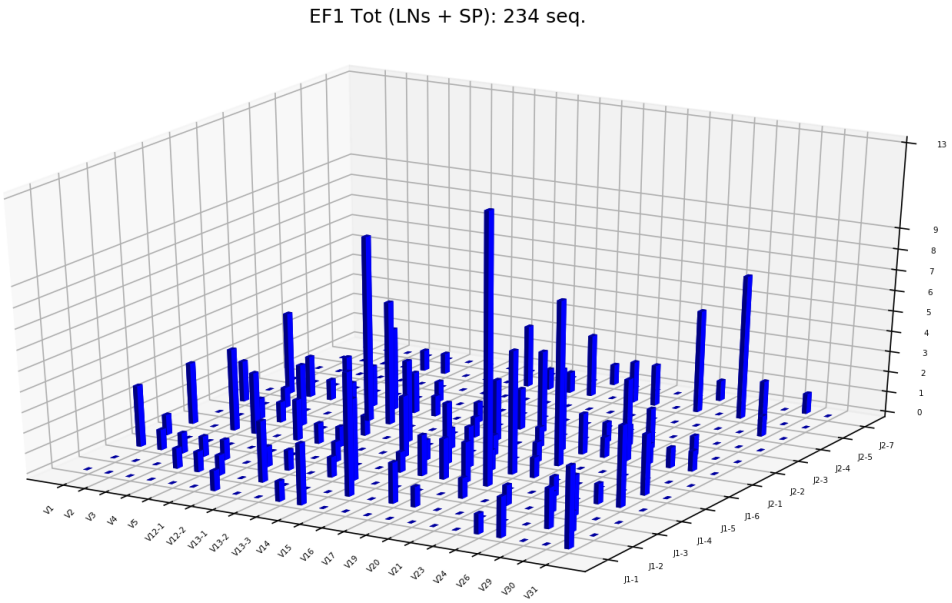

Figure 4.6: V-J plot for the infected mouse EF1.

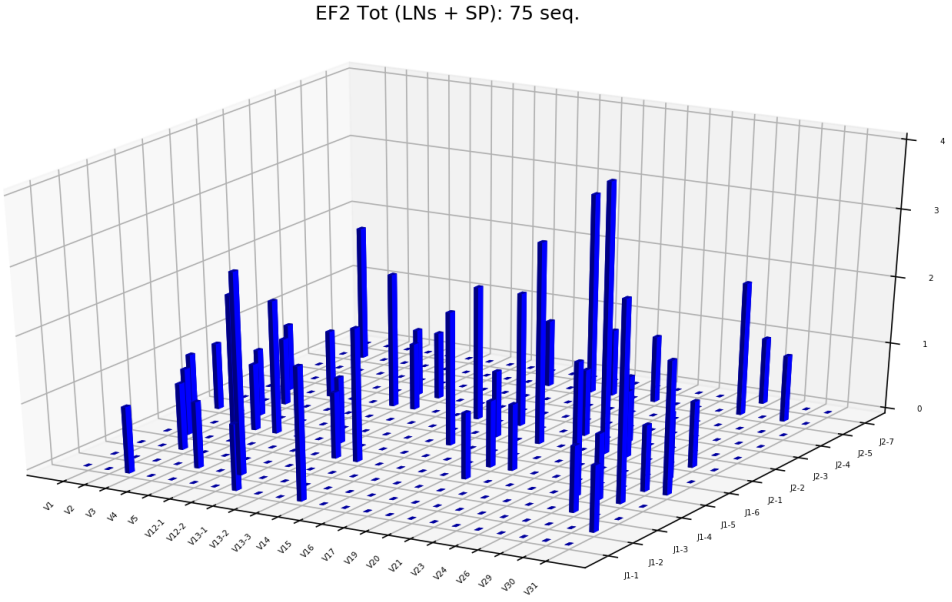

Figure 4.7: V-J plot for the infected mouse EF2.

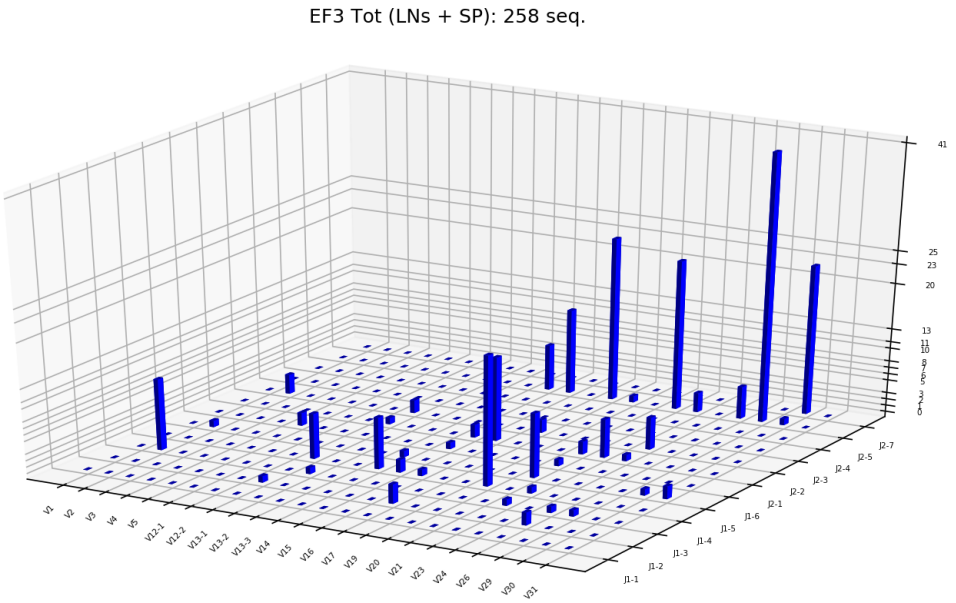

Figure 4.8: V-J plot for the infected mouse EF3.

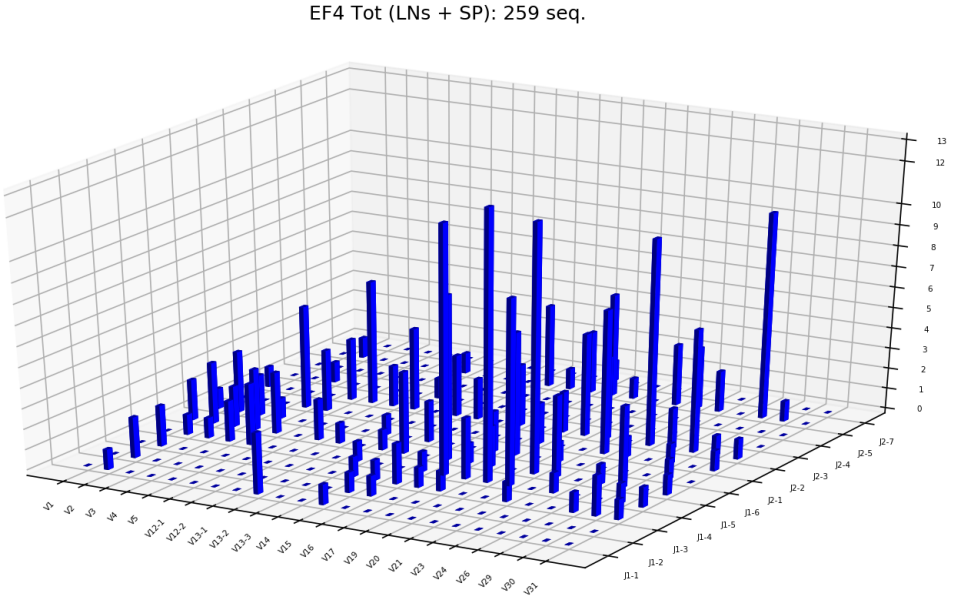

Figure 4.9: V-J plot for the infected mouse EF4.

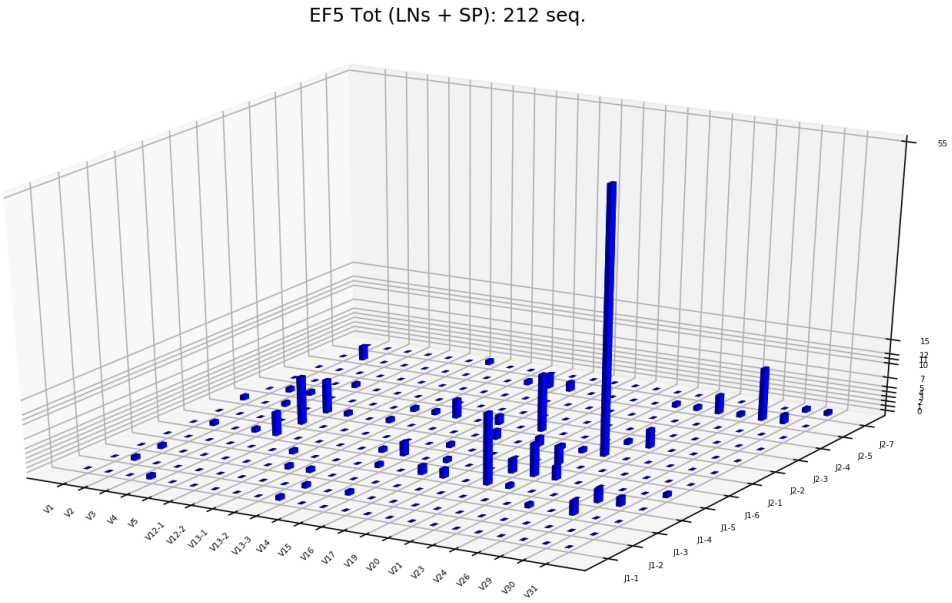

Figure 4.10: V-J plot for the infected mouse EF5.

4.3 Statistics

This statistical section will serve as a reference for the entire chapter in which we will try to answer some biological questions related to diversity inter (and intra) groups of infected and uninfected mice. The analyses will focus on the V-J repertoire of the data collected from our QuanTI collaborators Prof. Benedita Rocha and Dr. Pedro Filipe Fernandes Goncalves [71]. The goal will be to apply some of the techniques of [164], in particular the technique known as Randomization Test, to the data. Other analyses for public and private V-J repertoires will also follow.

4.3.1 Statistical terms

We recall here the definition of two important statistical terms:

- **test statistic:** A test statistic is a single measure of some attribute of a sample (i.e. a statistic) and it's used in statistical hypothesis testing. The main idea behind the test statistic is to summarize the data to a single value, that will be ultimately used to perform an hypothesis test. The test statistic, together with the p-value associated to the hypothesis test, allow to determine whether to accept or reject the null hypothesis (H_0) of the hypothesis test.

- **p-value:** The p-value is defined as the probability of obtaining a result equal to or more extreme than the result that was actually observed. It is a function of the observed sample and it is used for testing a statistical hypothesis. Before the test is performed, a threshold value is chosen, called the significance level of the test, traditionally 5% or 1%. A small p-value (≤ 0.05) indicates strong evidence against the null hypothesis, that is the data suggest to reject the null hypothesis. A large p-value (> 0.05) indicates weak evidence against the null hypothesis, that is the data are not sufficient to reject the null hypothesis.

4.3.2 Randomization Test

The Randomization test is a statistical test to determine the significance of some observed test statistic aimed at assessing a particular hypothesis. The assessment is achieved thanks to the generation of a distribution of the test statistic assuming the null hypothesis. The proportion of the distribution that is at least as extreme in absolute value as the observed test statistic is then determined. This proportion is a good estimate of the p-value and represents the probability that the observed test statistic could have been achieved by the distribution of the test statistic under the null hypothesis. See [164] for an illustration of the methodology of the randomization tests (in particular Fig. 4.11).

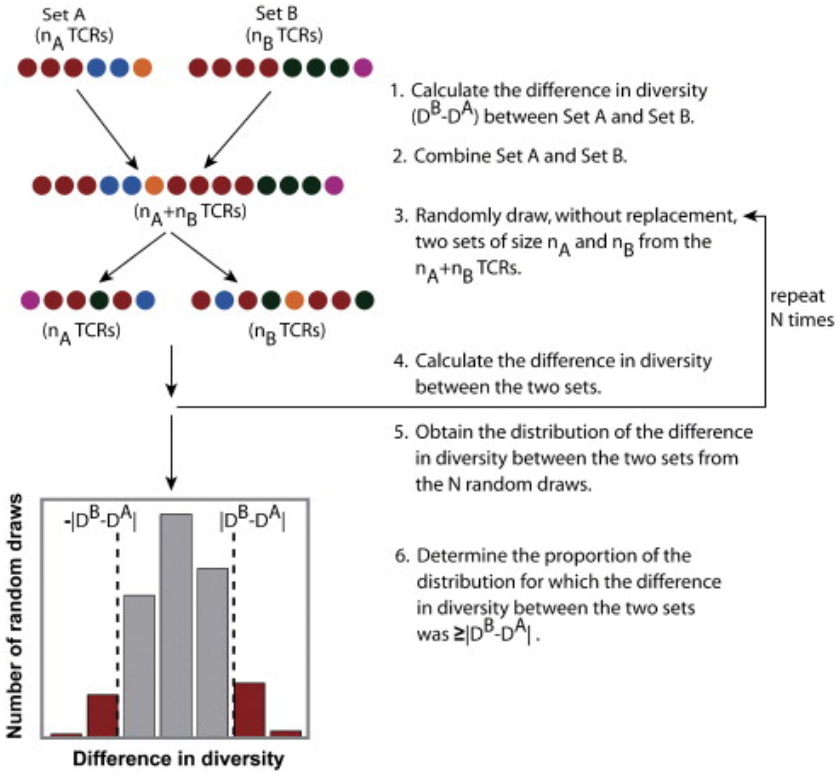

Figure 4.11: The randomization test for comparing the diversity of TCR samples. A schematic of the randomization test method for determining the statistical significance of the difference in a diversity measure ($D^B - D^A$) between two TCR sets, A and B. The method involves first pooling all sequences from Set A and Set B and then randomly drawing two new sets (of the same sizes as the original Set A and Set B), and calculating the difference in diversity that arose from this random sampling. This procedure is repeated multiple times (i.e.: repeat steps 3 and 4 multiple times) to obtain a distribution of the difference in diversity measures assuming the null hypothesis that both samples are drawn from the same distribution. The p-value for the difference in diversity ($D^B - D^A$) is the proportion of the distribution (highlighted in red) from the random draws for which the difference in diversity was greater than that observed experimentally (i.e.: ($D^B - D^A$)).

4.3.3 Simpson's diversity index

The Simpson's diversity index is a measure of the degree of concentration of a sample when individuals are classified into types, and it is defined as

$$D_S = 1 - \frac{\sum_{i=1}^{c} n_i(n_i - 1)}{n(n-1)} \,, \tag{4.1}$$

where n_i is the number of individuals of the i-th type (or class) in the sample, c is the number of different types in the sample, and n is the total number of individuals in the sample. This index ranges between 0 (all individuals of the sample belong to the same type) and 1 (each individual in the sample belongs to a different type or class) representing minimal and maximal diversity (within the sample) respectively [147].

4.3.4 Jaccard distance

The Jaccard distance between two sets A and B is a measure of overlapping dissimilarity between sample sets. It is defined as

$$d_J(A, B) = 1 - J(A, B) \,,$$

where $J(A, B)$ is the Jaccard index and it is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \,.$$

This index ranges between 0 (the two sets share everything) and 1 (the two set share nothing) representing minimal and maximal dissimilarity respectively [79, 133].

4.3.5 Wilcoxon-Mann-Whitney U test

The Wilcoxon-Mann-Whitney U test is a non-parametric statistical test, that is a statistical hypothesis testing test which make no *a priori* assumptions on the variables taken into consideration. As opposed to parametric tests, the parameters are present but they are determined by the data rather than by the model assumptions. Consider two different independent samples X and Y. Consider also generic observations x from sample X and y from sample Y. The hypotheses associated to the test are as follows:

- H_0: $\Pr(x > y) = \Pr(y > x)$;
- H_1: $\Pr(x > y) \neq \Pr(y > x)$.

The test is based upon a statistics called U which, under H_0 and considering samples with size greater or equal than 20, can be approximated with a normal distribution. For smaller samples, tables for the exact distribution of U exist. To compute the statistic U, an easy and intuitive way is as follows:

- Consider an observation x and define n_x the number of times an observation in sample X is greater than any other observation y in sample Y;
- Start with $n_x = 0$ and increase n_x of 1 each time x is greater than y, while increase it of 0.5 each time x meets a tie;
- Define $U_X = \sum_{x \in X} n_x$;
- Define U_Y with the contrary procedure.

4.3.6 Pearson's χ^2 test

The χ^2 test is is a statistical test in which the sampling distribution of the test statistic follows a chi-squared distribution when the null hypothesis is true. It is commonly used to assess independence of unpaired observations of two different random variables X and Y, where in general the random variables are expressed in contingency tables. Before giving the mains steps for the computation of this statistical test, we define the Pearson's χ^2 test statistic (which asymptotically approaches the χ^2 distribution) as follows:

$$\chi^2 = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}},$$

where

$$e_{i,j} = \frac{\left(\sum_j n_{i,j}\right)\left(\sum_i n_{i,j}\right)}{\sum_{i,j} n_{i,j}}$$

represents the expected frequency of type (i, j) if the occurrences were distributed randomly over the contingency table. The test is structured in 5 different steps, as follows:

- Compute the Pearson's χ^2 test statistic as specified above;
- Determine the number of degrees of freedom, $df = (r - 1) \times (c - 1)$, where r and c represent the number of rows and columns in the contingency table respectively;
- Define the confidence level for the test result (i.e. significance level of the test);
- Compare the χ^2 (computed in the first point) with the tabulated values of the χ^2 distribution (critical value) with degrees of freedom computed in the second point and with selected level of confidence chosen in the third point;
- If the test statistic exceeds the critical value of the fourth point, reject the null hypothesis H_0 for which the two random variables X and Y are independent.

4.4 Js and Vs frequency plots

We show here the plots for the fraction of all V and J gene segments in the data for both naïve, infected and immunized mice. Single V and J plots per mouse can be found with greater resolution in Appendix B. Figure 4.12 shows a clear preference for V_{13-1}, V_{13-2} and V_{13-3} in naïve mice. It also shows a preference for J_2 genes rather than for J_1 genes. Figure 4.13 shows instead the importance of V_{29} for infected and immunized mice. Different exploratory plots for the marginal distributions are shown in Figures 4.14, 4.15, 4.16 and 4.17, while plots for the joint distributions are shown in Figures 4.18 to 4.27. It is important to notice that the results for naïve mice are in concordance with previous studies, for both V [84] and J [29] segments.

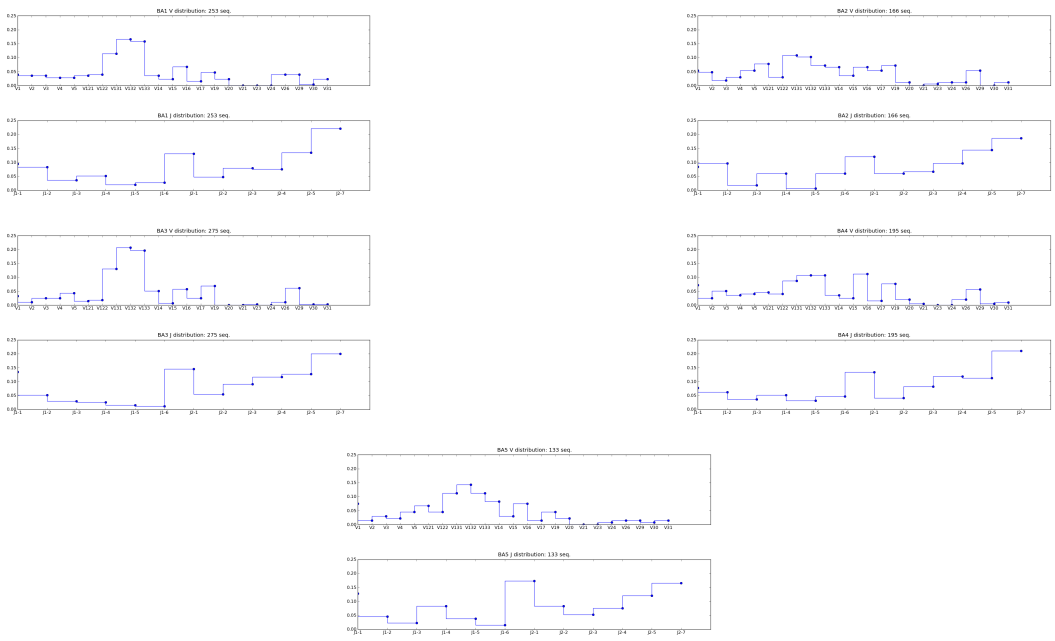

Figure 4.12: V-J frequency plots for the naïve mice BAs.

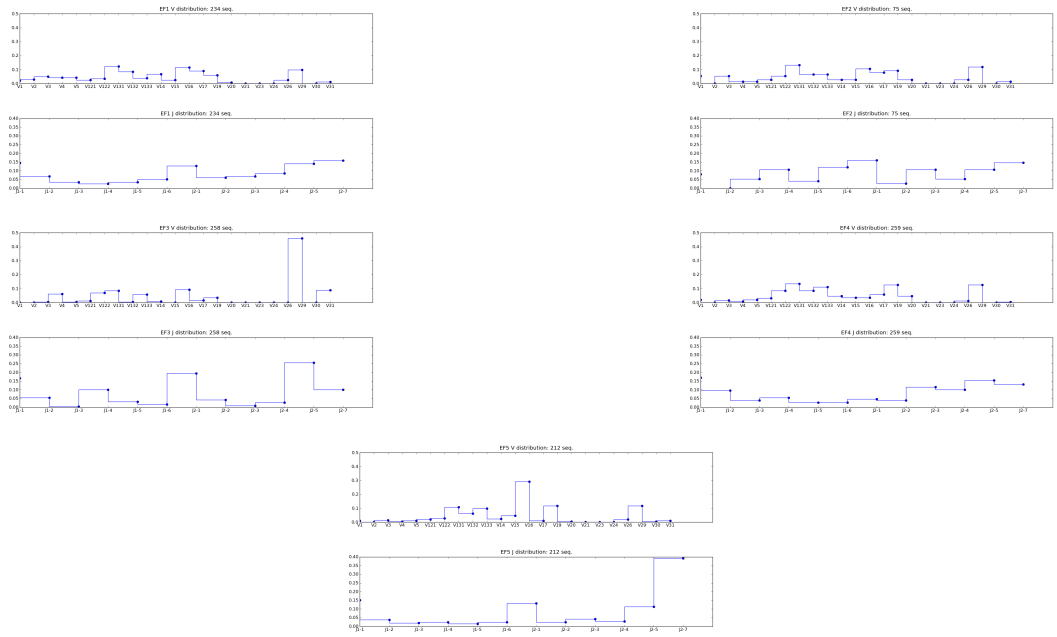

Figure 4.13: V-J frequency plots for the infected mice EFs.

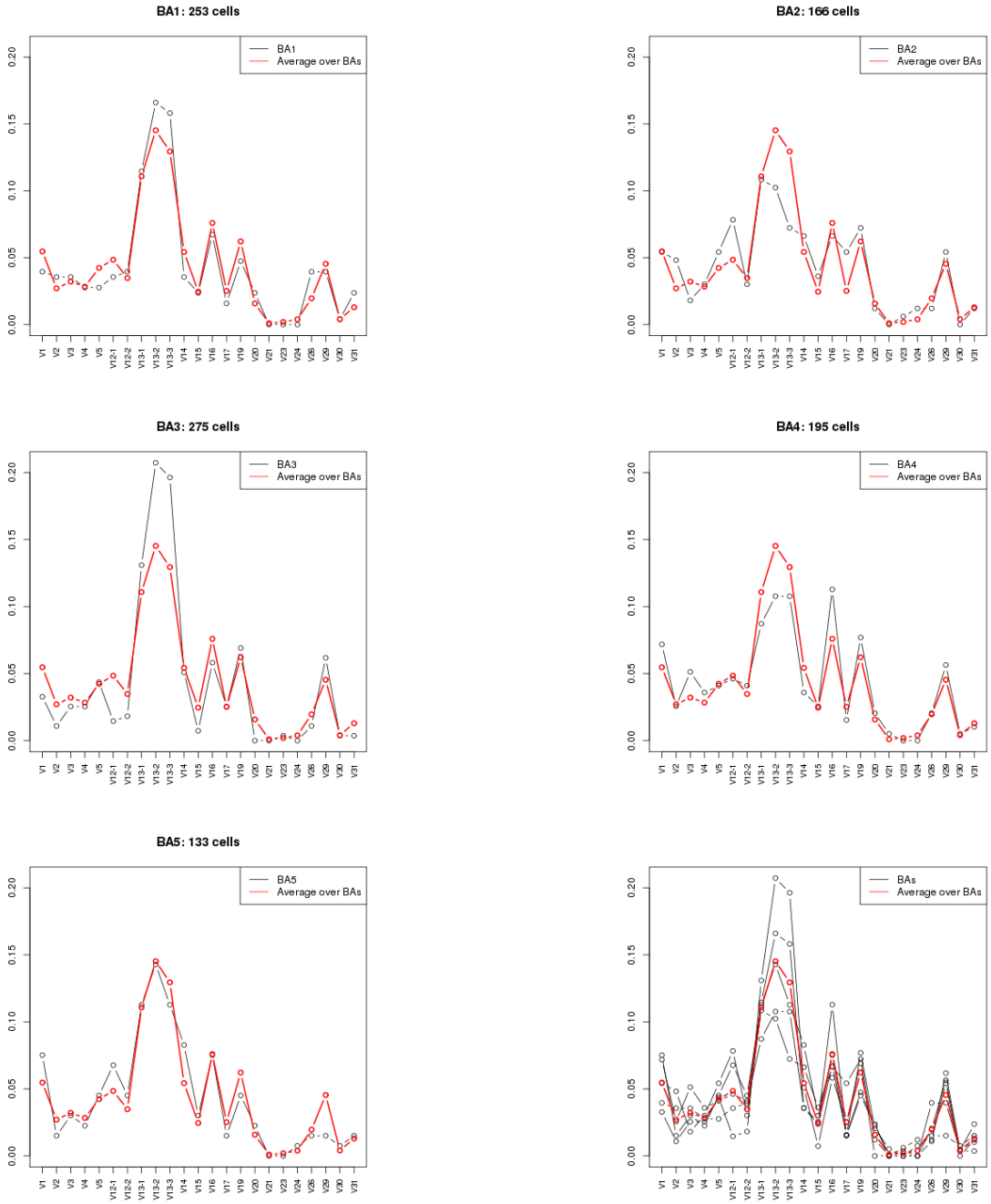

Figure 4.14: V frequency plots for each naïve mouse compared to average frequencies of naïve mice.

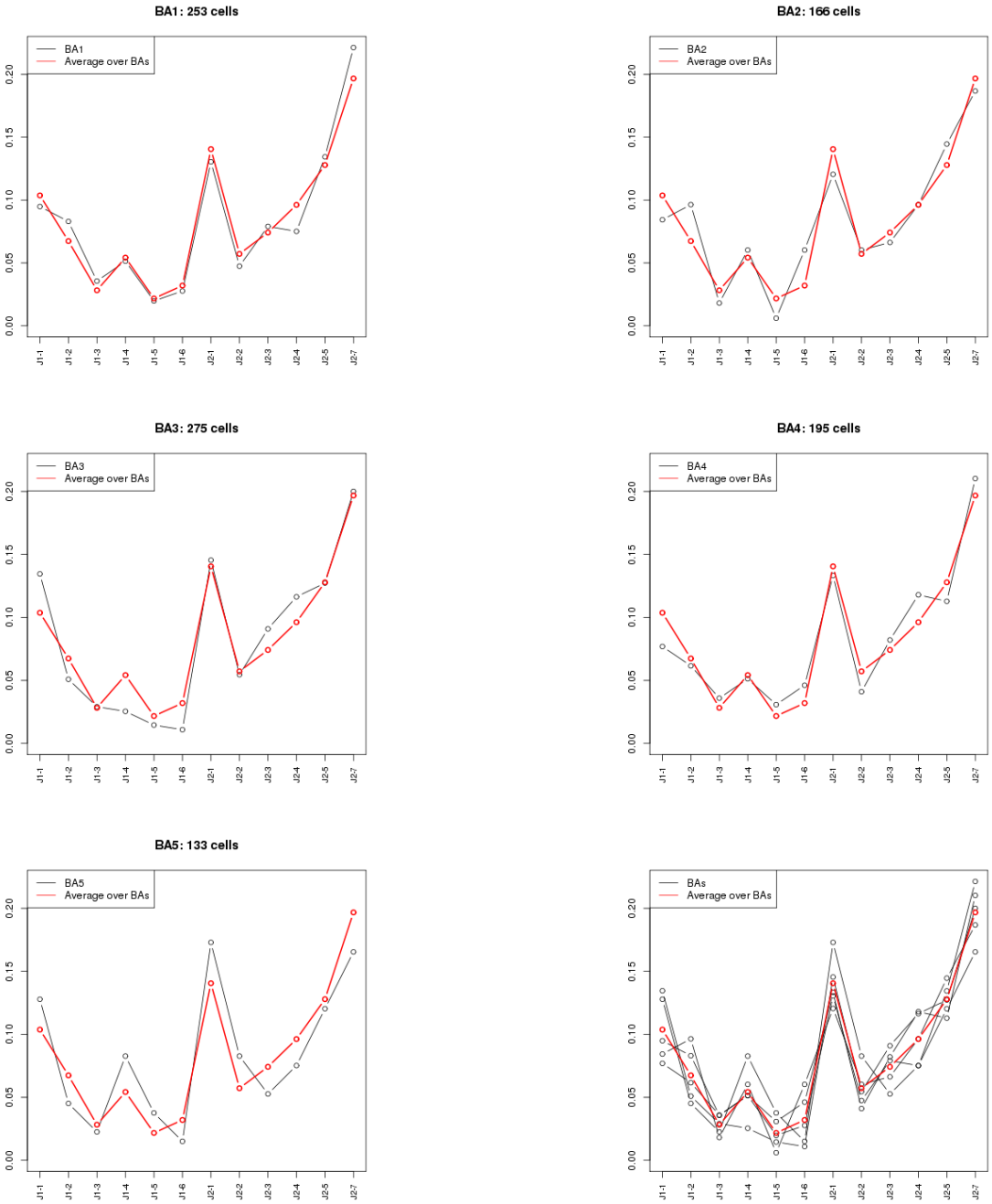

Figure 4.15: J frequency plots for each naïve mouse compared to average frequencies of naïve mice.

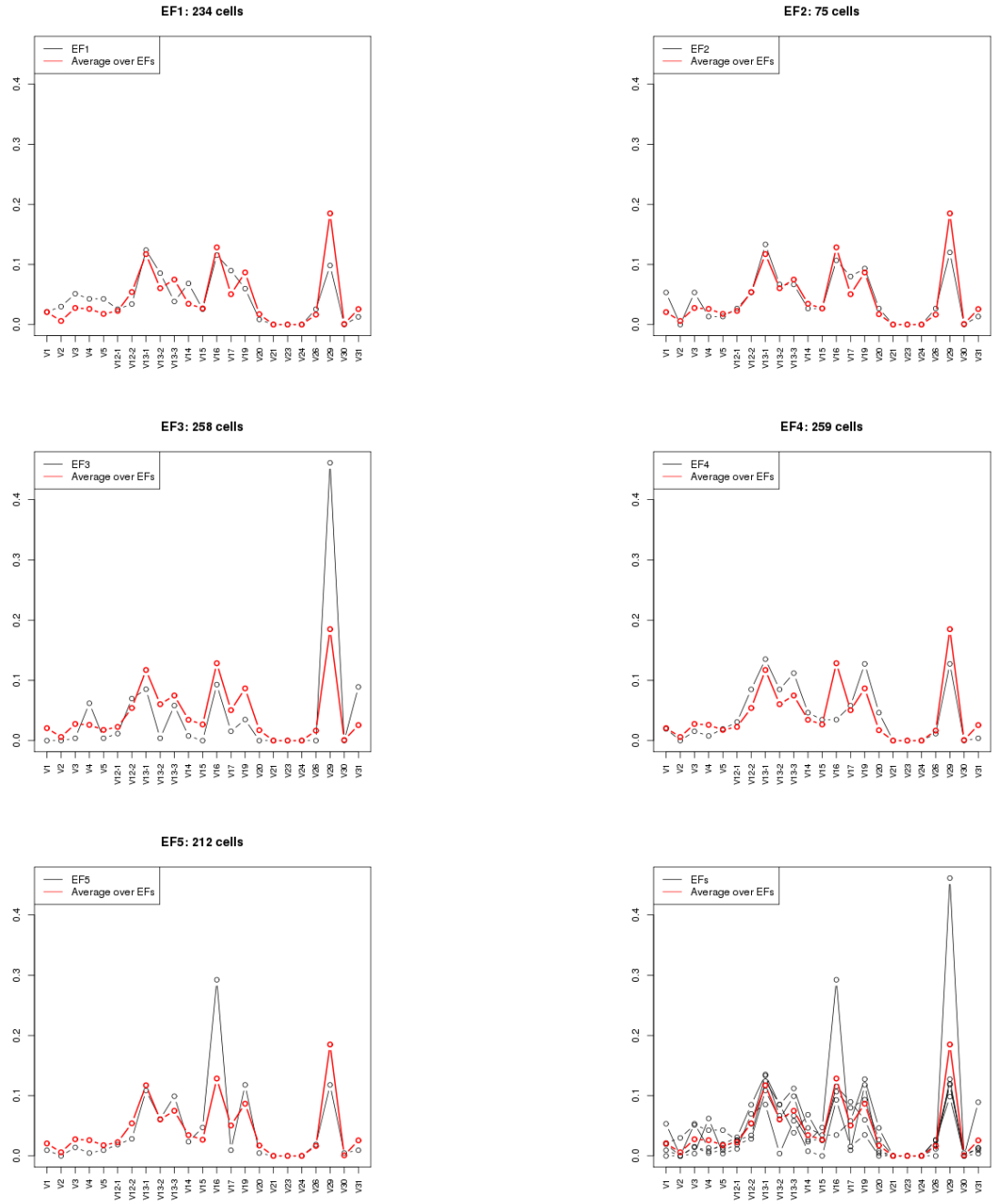

Figure 4.16: V frequency plots for each infected mouse compared to average frequencies of infected mice.

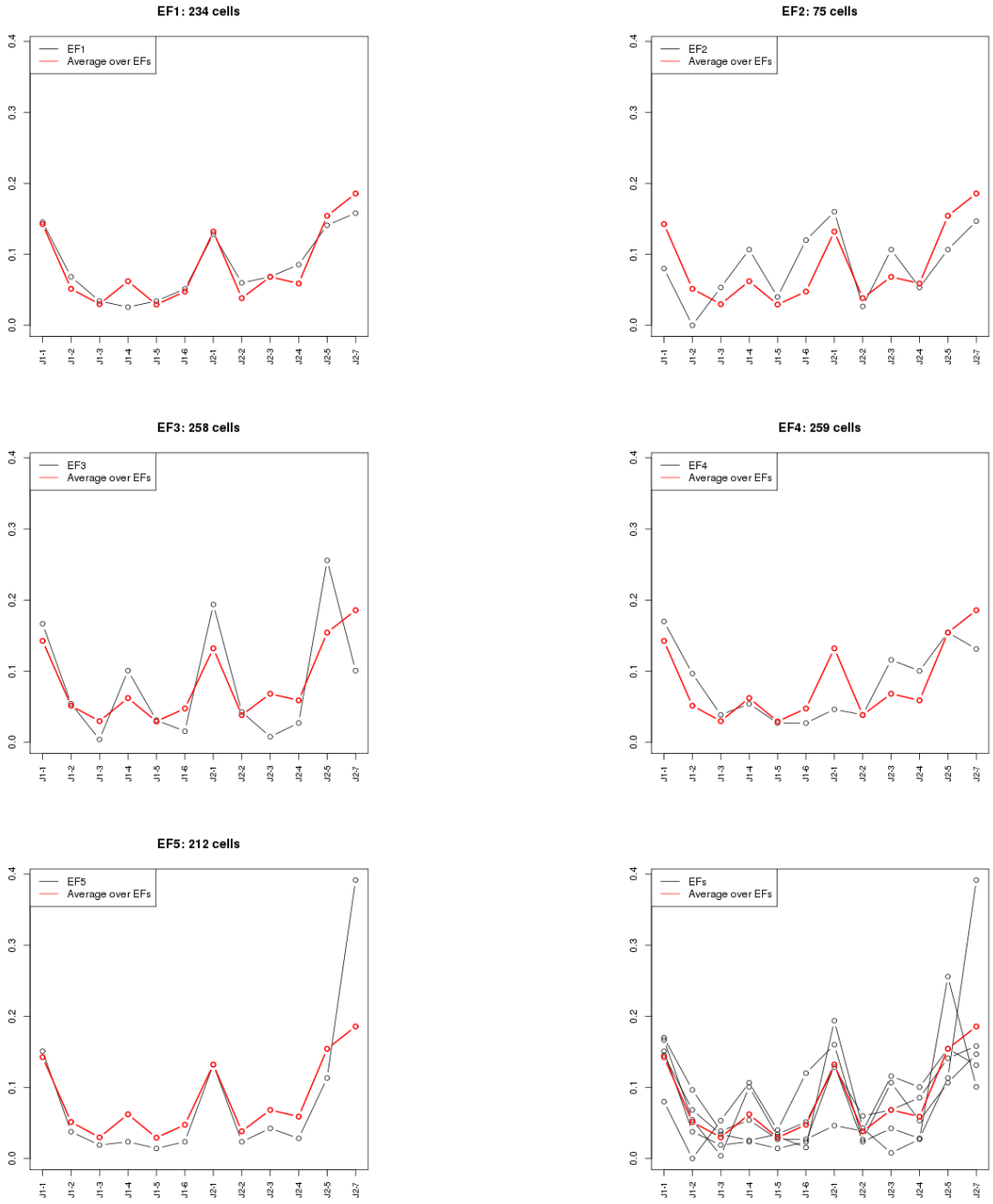

Figure 4.17: J frequency plots for each infected mouse compared to average frequencies of infected mice.

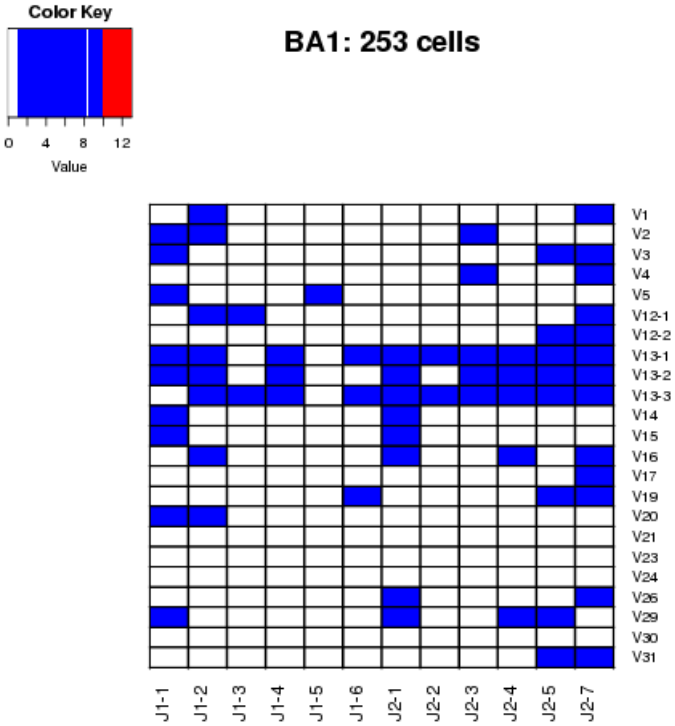

Figure 4.18: VJ plot for the naïve mouse BA1.

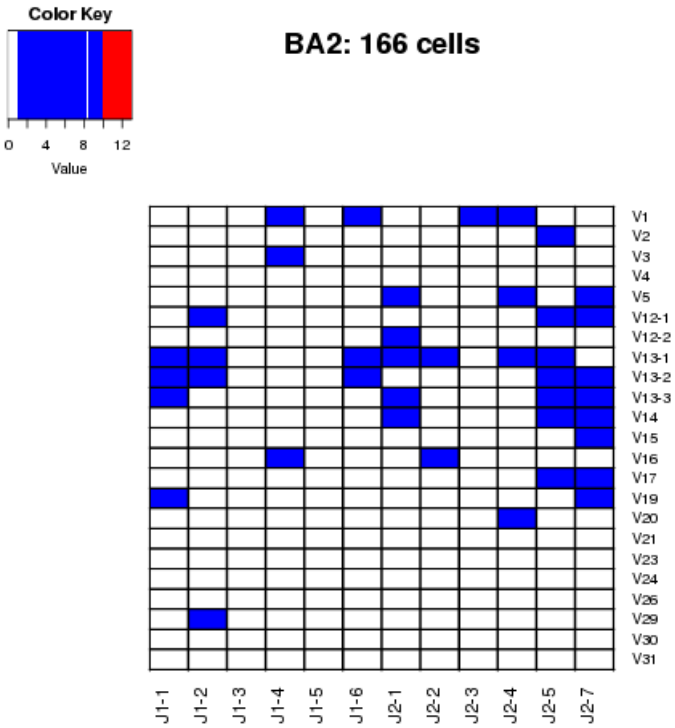

Figure 4.19: VJ plot for the naïve mouse BA2.

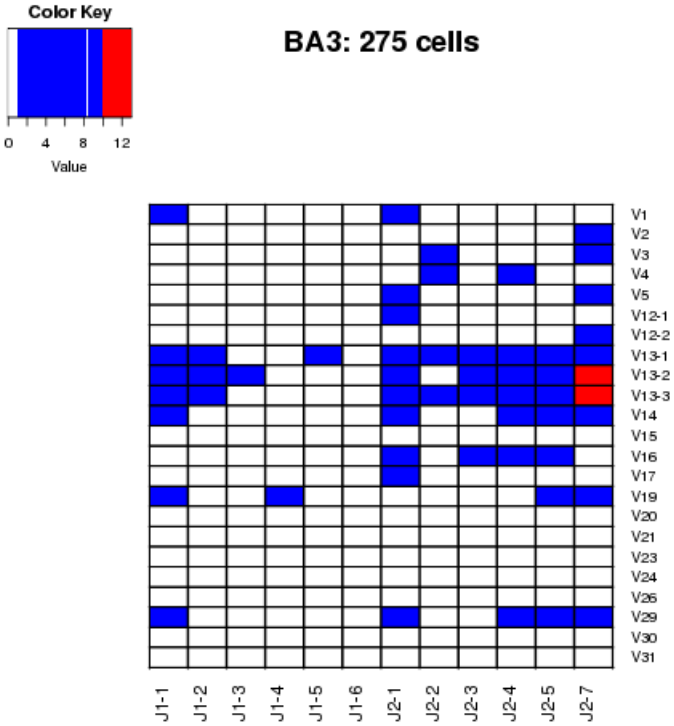

Figure 4.20: VJ plot for the naïve mouse BA3.

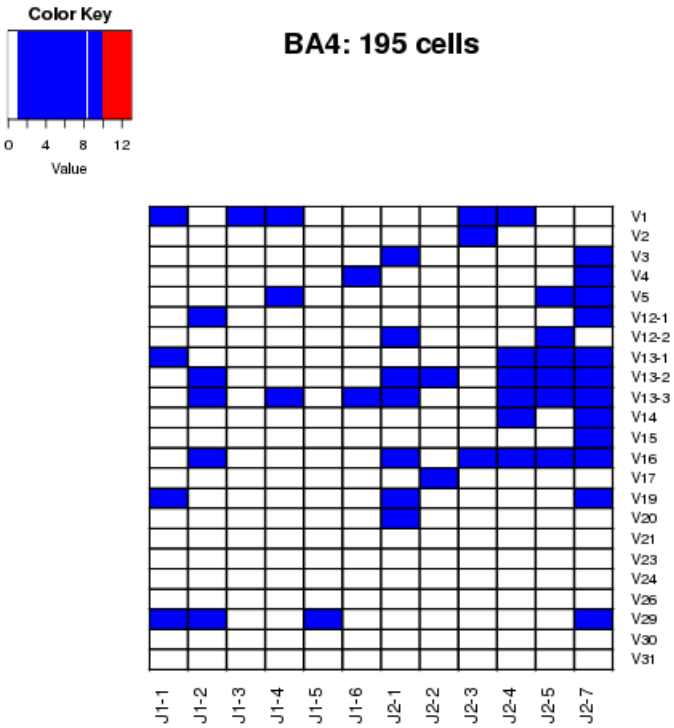

Figure 4.21: VJ plot for the naïve mouse BA4.

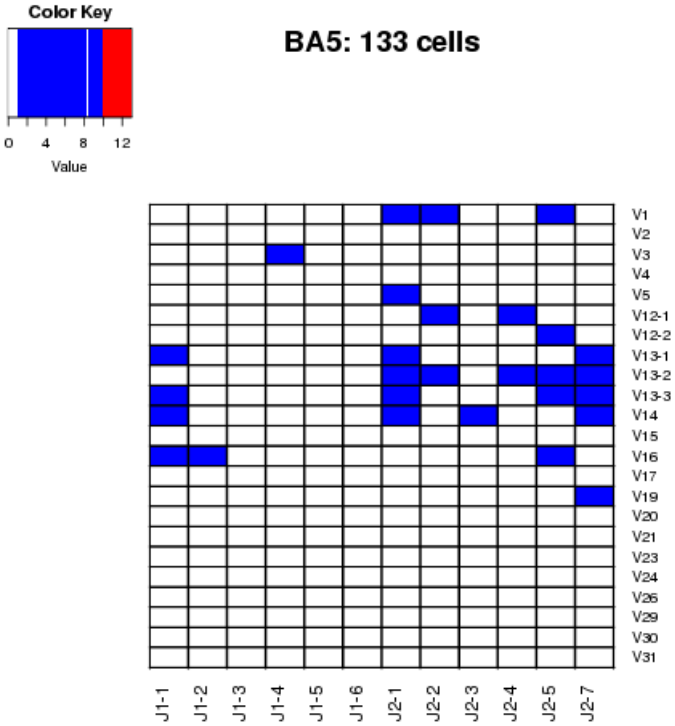

Figure 4.22: VJ plot for the naïve mouse BA5.

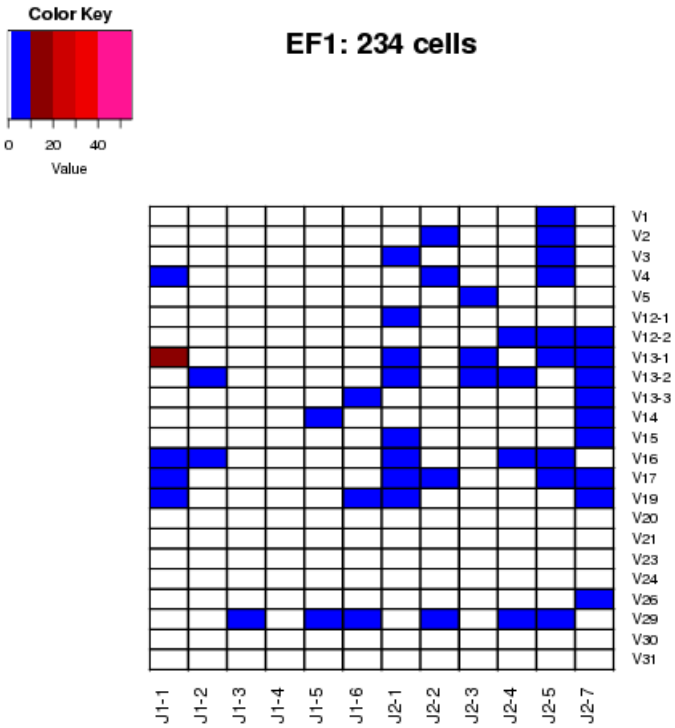

Figure 4.23: VJ plot for the infected mouse EF1.

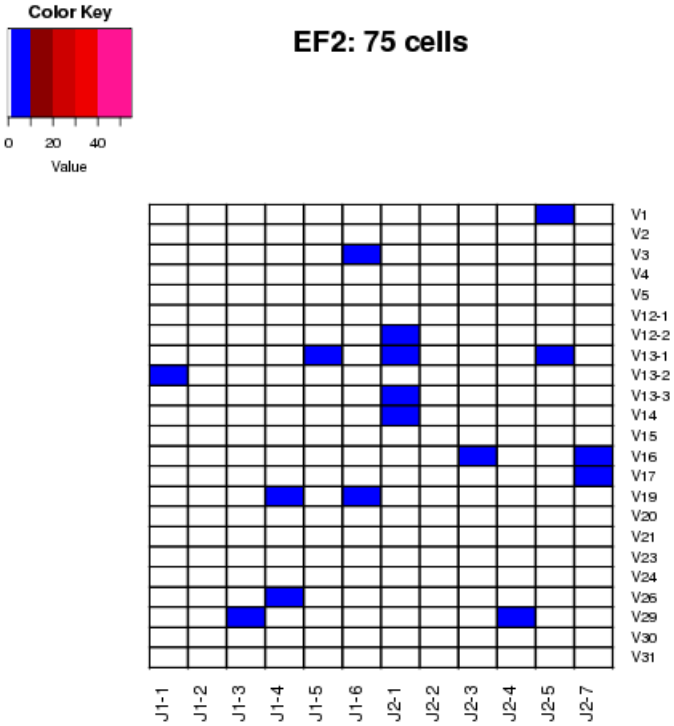

Figure 4.24: VJ plot for the infected mouse EF2.

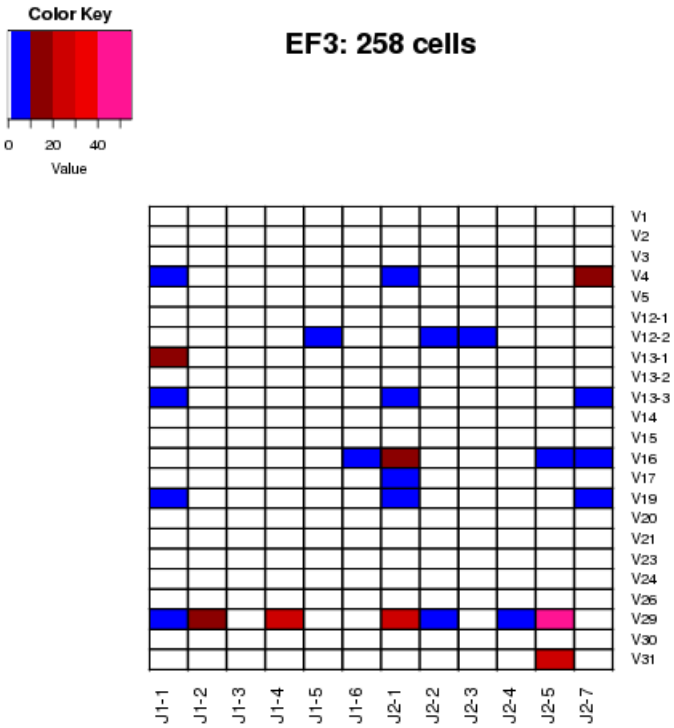

Figure 4.25: VJ plot for the infected mouse EF3.

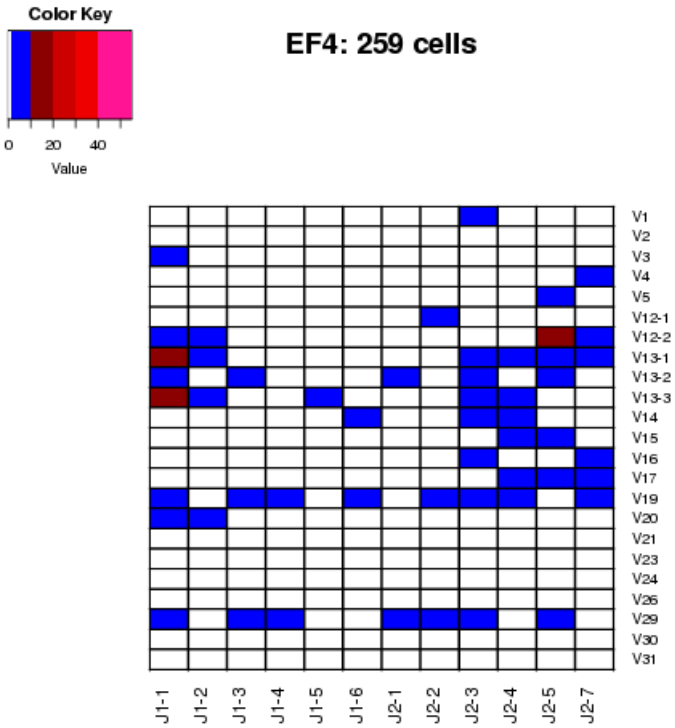

Figure 4.26: VJ plot for the infected mouse EF4.

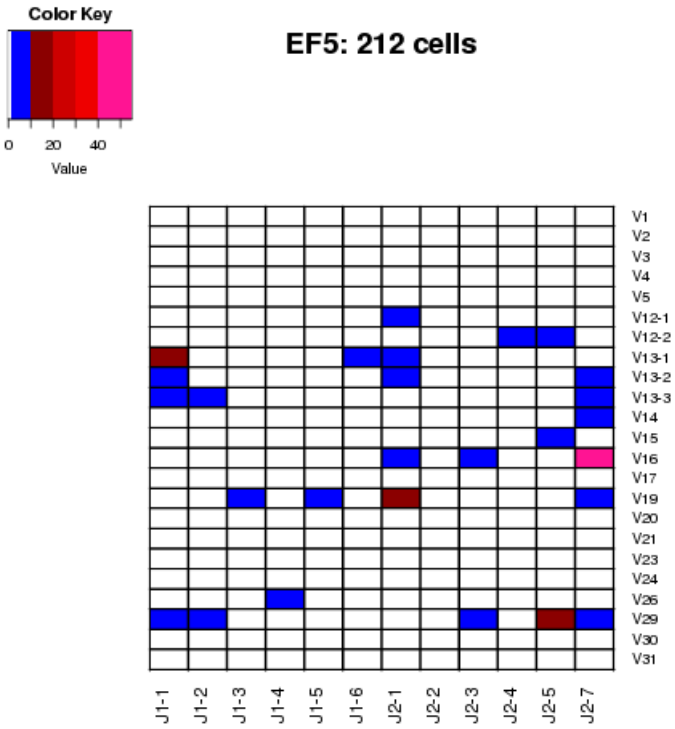

Figure 4.27: VJ plot for the infected mouse EF5.

4.5 Js and Vs Simpson's diversity

This section presents the plots for Simpson's indices for naïve and infected mice. It is worth noting that the two immunized mice (EF1 and EF2) have higher diversity index with respect to the infected ones, probably suggesting that infection induce a higher skewness in the diversity distribution with respect to immunization. Figure 4.28 indicates a minimum in the Vs diversity of mouse EF3. This agrees with Figure 4.13, where the V distribution in mouse EF3 is shown to be have low diversity. More interestingly, Figure 4.28 shows how important the diversity of the J component is for the general diversity of VJs, as it can be seen in the case of mouse BA3. Here, the V diversity is low with respect to the other mice, but a value of J diversity similar to the other mice implies a level of total VJ diversity comparable to that of other mice.

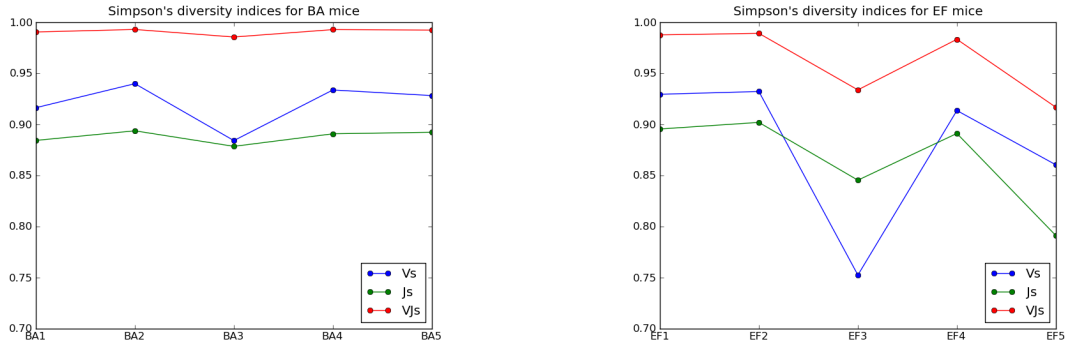

Figure 4.28: V-J-VJ Simpson's indices for naïve and infected mice.

4.5.1 Wilcoxon-Mann-Whitney U test

Three Wilcoxon-Mann-Whitney U tests were performed on the data representing naïve and infected mice. The first test was performed between the two sets

- $S_{BA,V} = \{D_{BA1,V}, D_{BA2,V}, D_{BA3,V}, D_{BA4,V}, D_{BA5,V}\}$
- $S_{EF,V} = \{D_{EF1,V}, D_{EF2,V}, D_{EF3,V}, D_{EF4,V}, D_{EF5,V}\}$,

where the generic $D_{BAn,V}$ represents the Simpson's diversity index of the naïve mouse BAn with respect to the V segments distribution. Similar reasoning holds for $D_{EFn,V}$. The second test was performed between the two sets

- $S_{BA,J} = \{D_{BA1,J}, D_{BA2,J}, D_{BA3,J}, D_{BA4,J}, D_{BA5,J}\}$
- $S_{EF,J} = \{D_{EF1,J}, D_{EF2,J}, D_{EF3,J}, D_{EF4,J}, D_{EF5,J}\}$,

where the generic $D_{BAn,J}$ represents the Simpson's diversity index of the naïve mouse BAn with respect to the J segments distribution. Similar reasoning holds for $D_{EFn,J}$. The third and last test was performed between the two sets

- $S_{BA,VJ} = \{D_{BA1,VJ}, D_{BA2,VJ}, D_{BA3,VJ}, D_{BA4,VJ}, D_{BA5,VJ}\}$
- $S_{EF,VJ} = \{D_{EF1,VJ}, D_{EF2,VJ}, D_{EF3,VJ}, D_{EF4,VJ}, D_{EF5,VJ}\}$,

where the generic $D_{BAn,VJ}$ represents the Simpson's diversity index of the naïve mouse BAn with respect to the VJ combination distribution. Similar reasoning holds for $D_{EFn,VJ}$. The result are shown here:

- BAV and EFV: U = 18, p-value = 0.3095
- BAJ and EFJ: U = 12, p-value = 1
- BAVJ and EFVJ: U = 23, p-value = 0.03175

Given the high p-values for the first two tests, we cannot consider the results as conclusive. The third test though, has a p-value lower than 0.05, suggesting that Simpson's diversity index is statistically higher in naïve mice with respect to infected mice.

4.6 Randomization tests for VJ's diversity

This section focuses on the distribution of diversity, using Simpson's diversity index as a test statistic. In other words, the study of this section focuses on how much a single mouse CD8+ TCR V-J repertoire differs from a flat CD8+ TCR V-J repertoire (case Simpson's index equal to 1) and how much these differences vary among mice and within groups. It is important to understand that we are not focusing on the actual V-J profile of a repertoire. This means that two different V-J profiles could have the same Simpson's diversity index. As an example, let us imagine that mouse X has the same V-J repertoire of mouse Y except for the fact that the number of sequences of class $\{V_1J_{1-1}\}$ and the number of sequences of class $\{V_1J_{1-2}\}$ are inverted. In this case, the two mice would have two different V-J profiles but the same Simpson's diversity index. Thus, this section focuses on the "amount" of V-J diversity.

Randomization tests based on Simpson's index were applied to each possible pair of mice (within the naïve group and within the infected group), resulting in a series of p-values plotted in Figure 4.29. The tested hypotheses are

H_0 : the two samples come from the same diversity distribution

H_1 : the two samples do not come from the same clonal size diversity distribution.

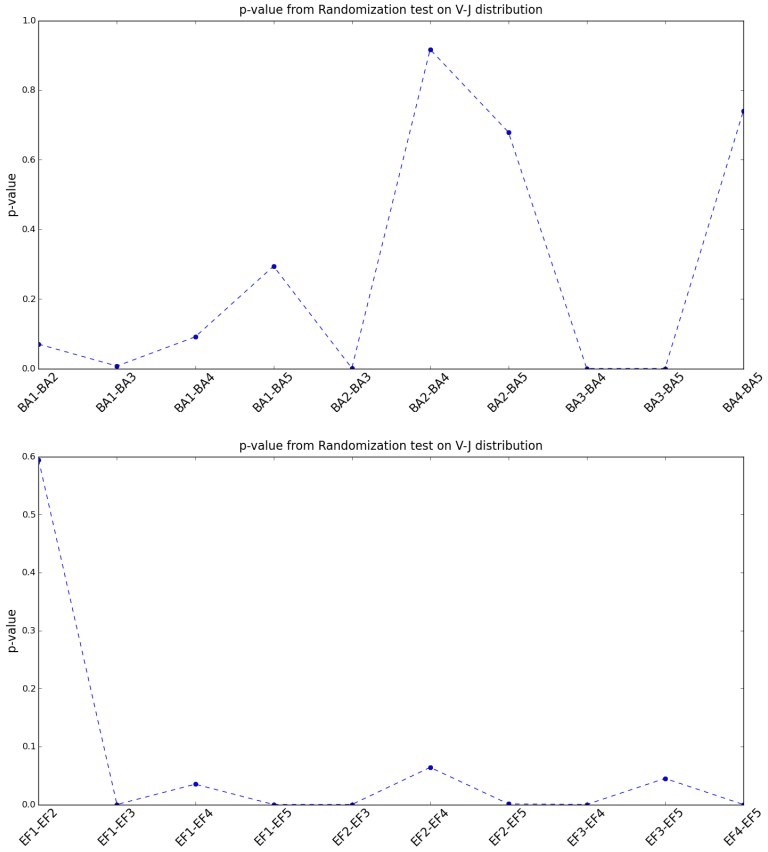

Figure 4.29: p-values for the randomization test based on Simpson's index, for both naïve and infected mice.

Figure 4.29 implies that we can exclude the null hypothesis (i.e. coming from the same diversity distribution) for the pairs of mice BA3-BA1, BA3-BA2, BA3-BA4, BA3-BA5, EF1-EF3, EF1-EF4, EF1-EF5, EF2-EF3, EF2-EF5, EF3-EF4, EF3-EF5. The EF1-EF2 pair is particularly interesting cause it represents the two previously immunized mice; see Tables 4.1 and 4.2.

BA1	X				
BA2	0.07011	X			
BA3	0.00578	0.00031	X		
BA4	0.08240	0.91210	0.00003	X	
BA5	0.29003	0.68246	0.00323	0.74695	X
	BA1	BA2	BA3	BA4	BA5

Table 4.1: p-values for the randomization test (10^5 simulations) based on Simpson's index for naïve mice.

EF1	X				
EF2	0.59052	X			
EF3	0.00000	0.00004	X		
EF4	0.04211	0.07309	0.00000	X	
EF5	0.00000	0.00211	0.03837	0.0000	X
	EF1	EF2	EF3	EF4	EF5

Table 4.2: p-values for the randomization test (10^5 simulations) based on Simpson's index for infected mice.

4.6.1 χ^2 test

Two Pearson's χ^2 tests were performed on the data. The first test was performed between the two categorical variables $X_1 = Vsegments$ and $Y_1 =$ naïve mice, with a number of degrees of freedom of $(23-1)\times(5-1)$. The second test was performed between the two categorical variables $X_2 = Jsegments$ and $Y_2 =$ naïve mice, with a number of degrees of freedom of $(12-1)\times(5-1)$. We show here the results

- Result for test 1: $\chi^2 = 126.06$, df $= 88$, p-value $= 0.004872$
- Result for test 2: $\chi^2 = 47.655$, df $= 44$, p-value $= 0.3264$.

Considering a confidence level of 0.05, we see that the results from test 1 allow us to reject H_0, that is the two variables X_1 and Y_1 are not independent. Test 2 does not allow to reject H_0 instead. The same procedure was applied to immunised/infected mice. The results were

- Result for test 1: $\chi^2 =$ NaN, df $= 88$, p-value $=$ NA
- Result for test 2: $\chi^2 = 224.94$, df $= 44$, p-value $< 2.2e-16$.

Considering a confidence level of 0.05, we see that the results from test 2 allow us to reject H_0, that is the two variables X_2 and $Y_2 =$ infected mice are not independent. Test 1 did not work because three of the columns of the contingency table (precisely the columns related to V21, V23 and V24) were full of 0s (each line represents a particular mouse), not allowing the computation of the $e_{i,j}$ described in 4.3.6. To overcome this issue, and to try to understand the relationship between the non-zero columns and the mice, the three columns were removed from the contingency table and the test was repeated (Test 3). The result was

- Result for test 3: $\chi^2 = 456.19$, df $= 76$, p-value $< 2.2e-16$.

This last result shows how the different V segments are not independent on the different immunized/infected mice.

4.7 Public & Private VJ repertoire

A map of shared and not shared V-J repertoires among naïve and infected mice has been computed and plotted. Results are shown in Fig. 4.30 and Fig. 4.31.

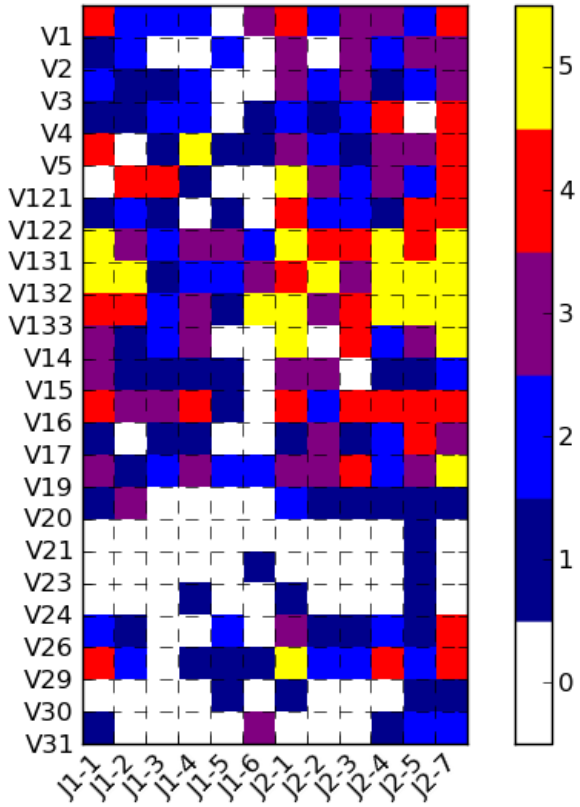

Figure 4.30: V-J repertoires sharing plot for naive mice.

The Jaccard distance was applied in order to vizualize the diversity in public and private V-J repertoires. It is immediately clear from the analysis that the average Jaccard distance for naïve mice is lower than the one for infected or immunised ones, suggesting that infected or immunized mice share less among each other than the naïve ones. See Figures 4.32 and 4.33.

4.8 Sample and repertoire frequencies

In this section we find a general result on the relation between frequencies in the repertoire and frequencies in a sample. We recall Chapter 3 and, in particular, (3.10)

$$q = \frac{\text{combinations of } S-1 \text{ elements in } m-1 \text{ places}}{\text{combinations of } S \text{ elements in } m \text{ places}} = \frac{\binom{S-1}{m-1}}{\binom{S}{m}} = \frac{m}{S}. \tag{4.2}$$

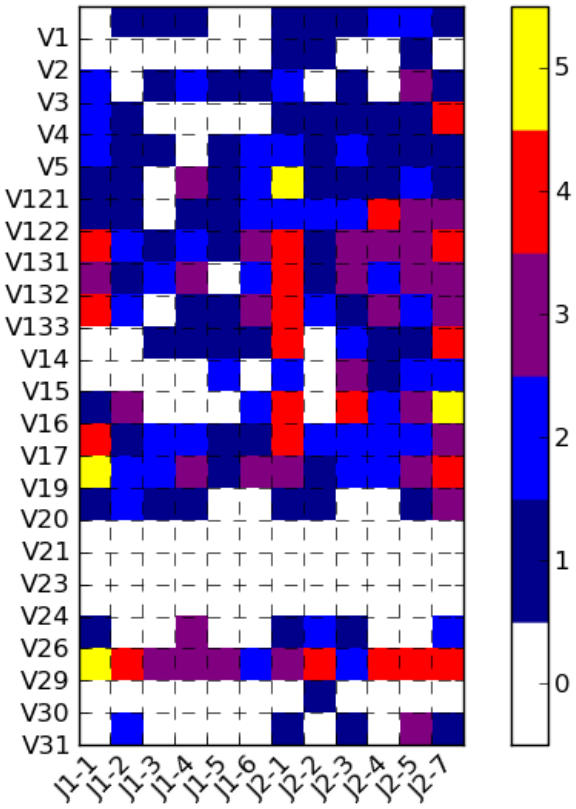

Figure 4.31: V-J repertoires sharing plot for infected mice.

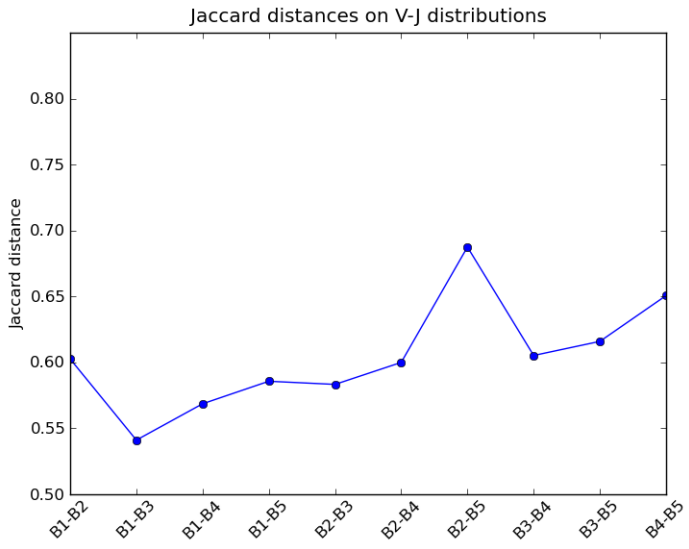

Figure 4.32: Jaccard indices among naïve mice.

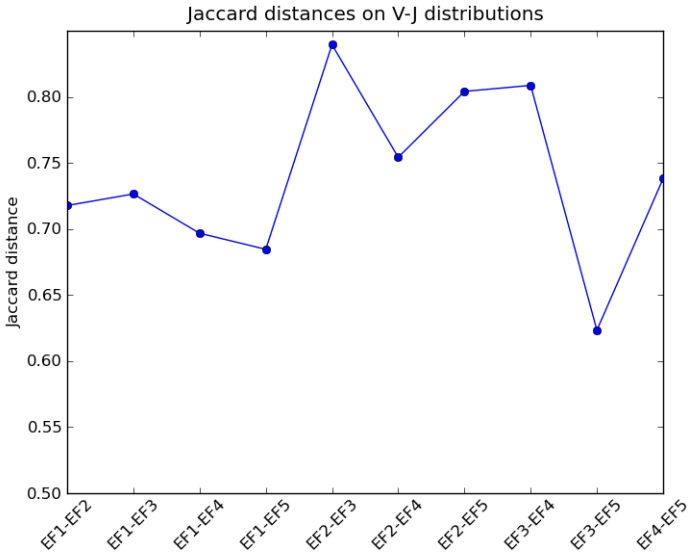

Figure 4.33: Jaccard indices among infected mice.

and (3.13)

$$\phi_{Y_i}(z) = (1 - q + qz)^{n_i}, \tag{4.3}$$

where n_i is the number of T cells in the clonotype class i in the repertoire. If we define $p_k = \Pr(Y_i = k)$, we have $\phi_{Y_i}(z) = p_0 + p_1 z + p_2 z^2 + \cdots$ and if we want to consider Y_i^{obs}, that is the number of observed T cells of clonotype i in the sample, then we have to consider

$$\Pr(Y_i^{obs} = k) = \Pr(Y_i = k | Y_i \neq 0) = \frac{\Pr(Y_i = k)}{\Pr(Y_i \neq 0)} = \frac{p_k}{1 - p_0}. \tag{4.4}$$

If we divide $\phi_{Y_i}(z)$ by $1 - p_0$ though, what we obtain is not exactly $\phi_{Y_i^{obs}}(z)$. In fact, we obtain

$$\frac{\phi_{Y_i}(z)}{1 - p_0} = \frac{p_0}{1 - p_0} + \frac{p_1}{1 - p_0} z + \frac{p_2}{1 - p_0} z^2 + \cdots$$

Defining $q_k = \dfrac{p_k}{1 - p_0}$, we have that $q_0 \neq 0$, which is not what we want, given that we expect $q_0 = \Pr(Y_i^{obs} = 0) = 0$. Therefore, we need

$$\phi_{Y_i^{obs}}(z) = \frac{\phi_{Y_i}(z)}{1 - p_0} - q_0 = \frac{(1 - q + qz)^{n_i} - (1 - q)^{n_i}}{1 - (1 - q)^{n_i}}. \tag{4.5}$$

Defining $f_i = n_i / S$ as the frequency of clonotype i in the repertoire, it is easy to compute

$$\mathbb{E}(Y_i^{obs}) = \frac{\mathbb{E}(Y_i)}{1 - p_0} = \frac{n_i q}{1 - (1 - q)^{n_i}} = \frac{m f_i}{1 - \left(\frac{S - m}{S}\right)^{S f_i}}, \tag{4.6}$$

and

$$\mathrm{Var}(Y_i^{obs}) = \frac{\mathrm{Var}(Y_i)}{(1 - p_0)^2} = \frac{n_i q (1 - q)}{[1 - (1 - q)^{n_i}]^2} = \frac{m f_i (S - m)}{S \left[1 - \left(\frac{S - m}{S}\right)^{S f_i}\right]^2}. \tag{4.7}$$

Defining g_i as the frequency of clonotype i in a sample of m cells, we can write $\mathbb{E}(Y_i^{obs}) = m\mathbb{E}(g_i)$. Therefore, from (4.6) we obtain

$$\mathbb{E}(g_i) = \frac{f_i}{1 - \left(\frac{S-m}{S}\right)^{Sf_i}}. \tag{4.8}$$

4.8.1 General solution of (4.8)

Equation (4.8) can be written as

$$\mathbb{E}(g_i) = \frac{f_i}{1 - \alpha^{f_i}} \quad \text{where} \quad \alpha = (1-q)^S. \tag{4.9}$$

This allows us to write

$$\alpha^{f_i} = 1 - \frac{1}{\mathbb{E}(g_i)} f_i. \tag{4.10}$$

We focus now on the solution of the general equation $p^{ax+b} = cx + d$, where $p > 0$ and $a, c \neq 0$. Using the substitution $-t = ax + \frac{ad}{c}$, which can be seen as $x = -\frac{d}{c} - \frac{t}{a}$, this equation can be transformed into

$$tp^t = R = -\frac{a}{c} p^{b - \frac{ad}{c}}.$$

This can be seen as

$$te^{t\ln(p)} = R \quad \text{and, therefore,} \quad t\ln(p)e^{t\ln(p)} = R\ln(p),$$

which gives

$$t = \frac{W(R\ln(p))}{\ln(p)} \quad \text{where } W \text{ represents the Lambert } W \text{ function.}$$

Thus, we have the general solution

$$x = -\frac{W\left(-\frac{a\ln(p)}{c} p^{b-\frac{ad}{c}}\right)}{a\ln(p)} - \frac{d}{c}. \tag{4.11}$$

We can now see (4.10) as a special case of this general case, where $p = \alpha$, $a = d = 1$, $b = 0$ and $c = -\mathbb{E}(g_i)^{-1}$. Thus, we obtain the solution

$$f_i = \mathbb{E}(g_i) - \frac{W\left(\mathbb{E}(g_i)\ln(\alpha)\alpha^{\mathbb{E}(g_i)}\right)}{\ln(\alpha)}. \tag{4.12}$$

This solution, as can also be seen in (4.10), strongly depends on the value of $\mathbb{E}(g_i)$ or f_i. In fact, no matter how small α could be, if raised to the f_i (See Eq. (4.10)) or to the $\mathbb{E}(g_i)$ (See Eq. (4.12)), it can still range between 0 and 1. For example, $\alpha = 10^{-20}$ can still become $\alpha^{0.001} = 0.95$. On the other hand, $\alpha = 10^{-20}$ becomes $\alpha^{0.1} = 0.01$. For this reason, in the next section we try to give some approximation for f_i.

4.8.2 Approximation of (4.8)

Using Laurent expansion at $f_i \simeq 0$, we can approximate (4.8) as

$$\mathbb{E}(g_i) \simeq -\frac{1}{S \ln\left(1-\frac{m}{S}\right)} + \frac{f_i}{2} + \mathcal{O}(f_i^2), \tag{4.13}$$

which gives us

$$f_i \simeq 2\left(\mathbb{E}(g_i) + \frac{1}{S \ln\left(\frac{S-m}{S}\right)}\right). \tag{4.14}$$

Applying Taylor expansion, we obtain

$$\frac{1}{\ln\left(1-\frac{m}{S}\right)} \simeq -\frac{S}{m} + \mathcal{O}(1/2), \tag{4.15}$$

and therefore

$$\mathbb{E}(g_i) \simeq \frac{1}{m} + \frac{f_i}{2}. \tag{4.16}$$

On the other side, it is clear from (4.9) that if f_i is not small enough, then we have $\alpha^{f_i} \simeq 0$ and, therefore,

$$\mathbb{E}(g_i) \simeq f_i. \tag{4.17}$$

It is interesting to notice that (4.16) and (4.17) have only one point in common, which is $f_i = \frac{2}{m}$. This value of f_i represents also the point in which both (4.16) and (4.17) have maximum distance from the real trajectory (4.8). This distance (or error) is the difference between (4.8) in $f_i = \frac{2}{m}$ and (4.17) in $f_i = \frac{2}{m}$:

$$\frac{\frac{2}{m}}{1-(1-q)^{\frac{2}{q}}} - \frac{2}{m} = \frac{2}{m}\left[\frac{(1-q)^{\frac{2}{q}}}{1-(1-q)^{\frac{2}{q}}}\right].$$

Using Taylor approximation, we can write

$$\frac{(1-q)^{\frac{2}{q}}}{1-(1-q)^{\frac{2}{q}}} \simeq \frac{1}{e^2-1} + \mathcal{O}(q).$$

This allows us to give a definitive answer to our approximation problem:

- For $f_i \in \left(0, \frac{2}{m}\right]$, we use (4.16) to approximate (4.8);
- For $f_i \in \left(\frac{2}{m}, 1\right]$, we use (4.17);
- The maximum error produced is $err_i = \frac{0.3}{m}$ at frequency $f_i = \frac{2}{m}$.

Thus, we have

$$f_i \simeq \begin{cases} 2\left(\mathbb{E}(g_i) - \frac{1}{m}\right) & \text{for } \mathbb{E}(g_i) \in \left(\frac{1}{m}, \frac{2}{m}\right] \\ \mathbb{E}(g_i) & \text{for } \mathbb{E}(g_i) \in \left(\frac{2}{m}, 1\right]. \end{cases} \tag{4.18}$$

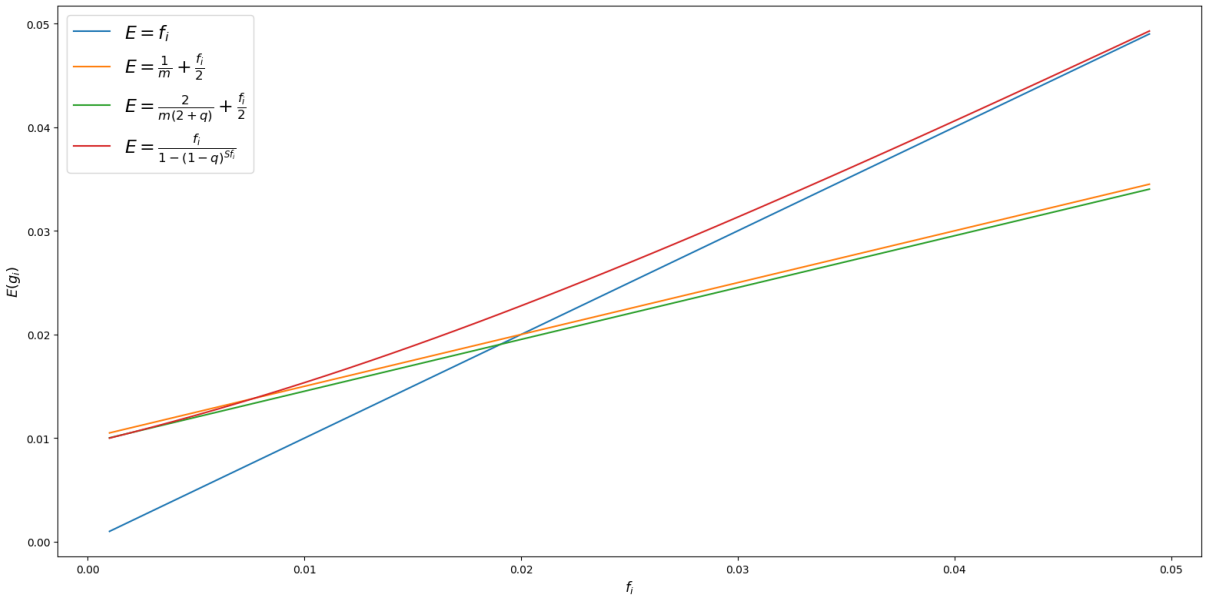

Figure 4.34: Test of goodness of (4.17), (4.16) and (4.20) for (4.8). Parameters are S = 1000 and m = 100.

As we can see from (4.18), the case $\mathbb{E}(g_i) = \frac{1}{m}$ would give us no information at all about f_i. The reason behind this comes from the approximation (4.15). Therefore, we consider a better Taylor approximation

$$-\frac{1}{S\ln\left(1-\frac{m}{S}\right)} \simeq -\frac{1}{S}\frac{1}{\left[-q-\frac{q^2}{2}\right]} = \frac{2}{m(2+q)}. \tag{4.19}$$

Thus, we can write

$$\mathbb{E}(g_i) \simeq \frac{2}{m(2+q)} + \frac{f_i}{2}, \tag{4.20}$$

eventually obtaining

$$f_i \simeq \begin{cases} 2\left(\mathbb{E}(g_i) - \frac{2}{m(2+q)}\right) & \text{for } \mathbb{E}(g_i) \in \left[\frac{1}{m}, \frac{2}{m}\right] \\ \mathbb{E}(g_i) & \text{for } \mathbb{E}(g_i) \in \left(\frac{2}{m}, 1\right]. \end{cases} \tag{4.21}$$

In Figure 4.34 we can see a plot of these results.

4.8.3 Approximation of $\mathrm{Var}(g_i)$

Similar steps can be done to approximate $\mathrm{Var}(Y_i^{obs})$. In fact, applying Taylor expansion to Eq. (4.7), we have

$$\mathrm{Var}(Y_i^{obs}) \simeq \frac{m(S-m)}{S}\left[\frac{1}{f_i S^2 \ln^2\left(1-\frac{m}{S}\right)} - \frac{1}{S\ln\left(1-\frac{m}{S}\right)} + \mathcal{O}(f_i)\right]. \tag{4.22}$$

If we consider now (4.15) and the other Laurent expansion

$$\frac{1}{\ln^2\left(1-\frac{m}{S}\right)} \simeq \frac{S^2}{m^2} - \frac{S}{m} + \mathcal{O}(1/12),$$

we obtain

$$\mathrm{Var}(Y_i^{obs}) \simeq \frac{(S-m)}{S}\left[\frac{1}{f_i m} - \frac{1}{S f_i} + 1\right], \tag{4.23}$$

and

$$\mathrm{Var}(g_i) \simeq \frac{(S-m)}{m^2 S}\left[\frac{1}{f_i m} - \frac{1}{S f_i} + 1\right] := \alpha(f_i), \tag{4.24}$$

which in turn gives

$$\sigma_{g_i} \simeq \sqrt{\alpha(f_i)}. \tag{4.25}$$

4.8.4 Standard error of $\bar{g}_i$

Let us imagine to extract K independent samples and to observe the quantities $Y_i^{(1)}$, $Y_i^{(2)}$, $\cdots, Y_i^{(K)}$, where $Y_i^{(j)}$ represents the number of T cells of clonotype i found in the j^{th} sample of size m. Let us define the sample mean of $\{Y_i^{(j)}, j = 1, 2, \cdots, K\}$ as $\bar{Y}_i$. In the same way, we define the sample mean of $\{g_i^{(j)}, j = 1, 2, \cdots, K\}$ as $\bar{g}_i$.

Given the relations expressed in (4.21), we want now to focus on the standard error of $\bar{g}_i$. We want to do this to understand how much $\bar{g}_i$ differs from $\mathbb{E}(g_i)$ with a small number of samples K and, in turn, how good is the estimate in (4.14).

Let us now focus on the standard error of $\bar{g}_i$. We know from statistical theory that

$$\mathrm{SE}_{\bar{g}_i} = \frac{s}{\sqrt{K}}, \quad \text{where} \quad s = \sqrt{\frac{1}{K-1}\sum_{j=1}^{K}\left(g_i^{(j)} - \bar{g}_i\right)^2}. \tag{4.26}$$

We know that $\mathrm{SE}_{\bar{g}_i}$ is an estimate of how far the sample mean $\bar{g}_i$ is likely to be from the population mean $\mathbb{E}(g_i)$, giving us an idea of the goodness of using $\bar{g}_i$ in place of the general $\mathbb{E}(g_i)$ in (4.18).

We would like now to find an upper bound for $\mathrm{SE}_{\bar{g}_i}$ which could be expressed as a function of the only variable K. The following steps show our upper bound:

- $K\bar{g}_i = \sum_{k=i}^{K} g_i^{(k)}$;
- $\bar{g}_i - g_i^{(j)} = \frac{1}{K}\sum_{k\neq j} g_i^{(k)} - \frac{(K-1)g_i^{(j)}}{K} \leq \frac{K-1}{K}\left[1 - g_i^{(j)}\right] \leq \frac{K-1}{K}$;
- $\left|g_i^{(j)} - \bar{g}_i\right| \leq \frac{K-1}{K}\ \forall j \in \{1, 2, \cdots, K\}$;
- $\left(g_i^{(j)} - \bar{g}_i\right)^2 \leq \frac{(K-1)^2}{K^2}\ \forall j \in \{1, 2, \cdots, K\}$;
- $\sqrt{\sum_{j=1}^{K}\left(g_i^{(j)} - \bar{g}_i\right)^2} < \frac{K-1}{\sqrt{K}}$;

which gives us the upper bound

$$\mathrm{SE}_{\bar{g}_i} < \frac{\sqrt{K-1}}{K}. \tag{4.27}$$

Figure 4.35 shows numerical simulations of frequencies in the repertoire and their relative frequencies in the samples (average values over 100 samples). Other simulations, showing sampling from different kind of clonal size distributions, are shown in Figures 4.36-4.53. In particular, for each kind of distribution (e.g., geometric distribution with mean 3), three different plots are shown, for three different sampling and plotting procedures: (i) extraction of one single sample and plotting of all of the classes in the sample, (ii) extraction of five samples and plotting of all of the classes in the five samples, and (iii) extraction of five samples and plotting of the classes that are present in all of the samples. The third kind is the reason why some plots do not show any point. Each figure also plots (4.8) ± (4.25).

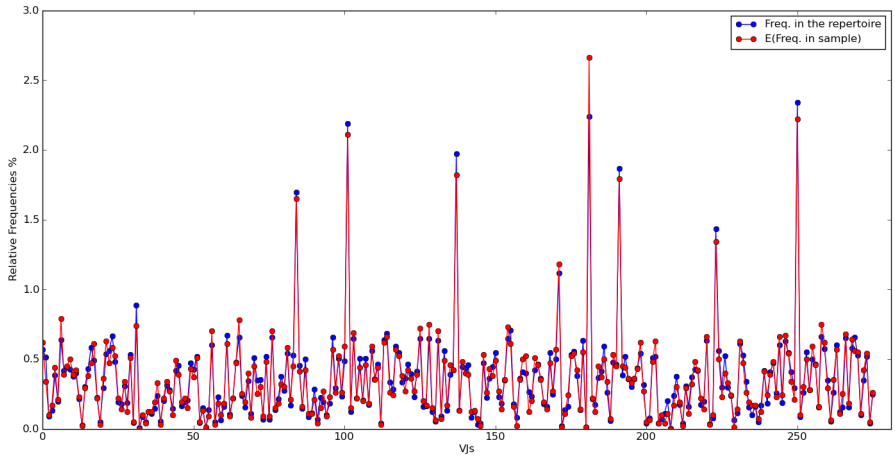

Figure 4.35: Simulation of frequencies in the repertoire and related frequences of observed classes in the sample. Average values over 100 samples.

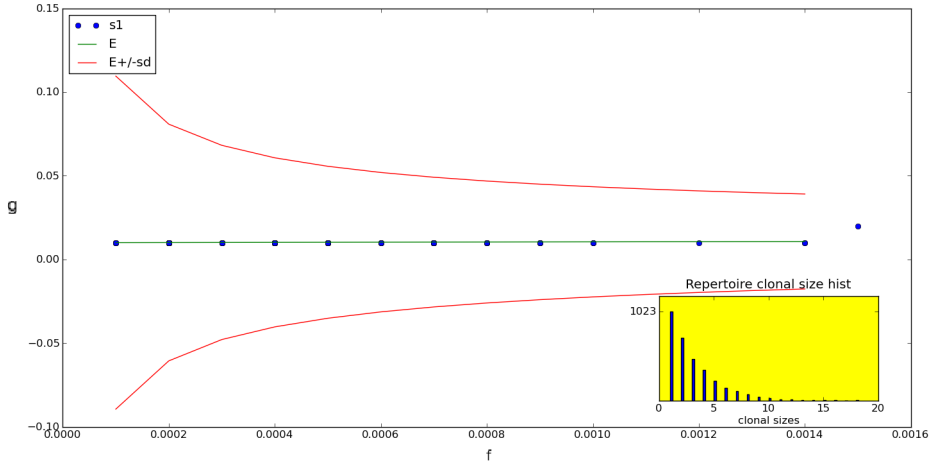

Figure 4.36: Geometric repertoire with mean 3. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 3250$.

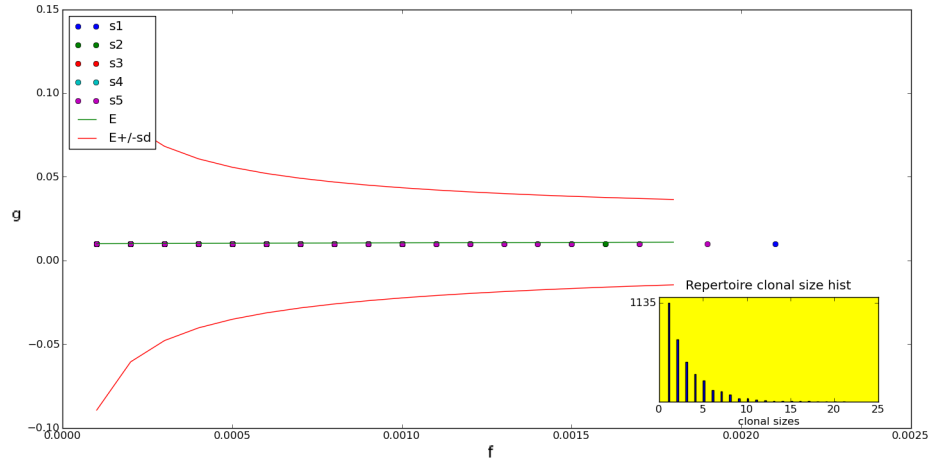

Figure 4.37: Geometric repertoire with mean 3. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 3307$.

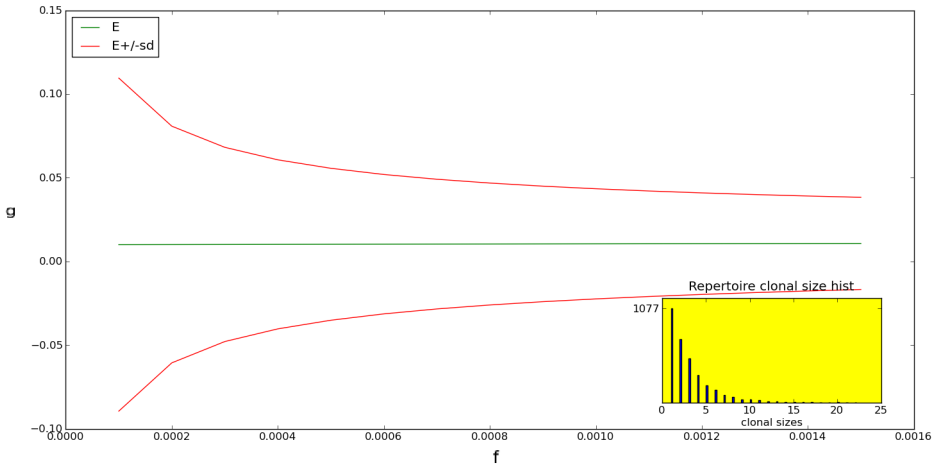

Figure 4.38: Geometric repertoire with mean 3. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 3274$.

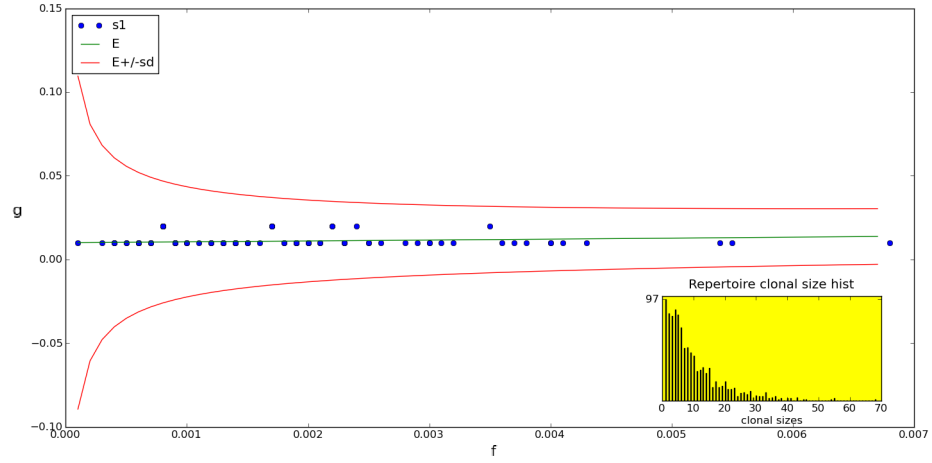

Figure 4.39: Geometric repertoire with mean 10. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 1029$.

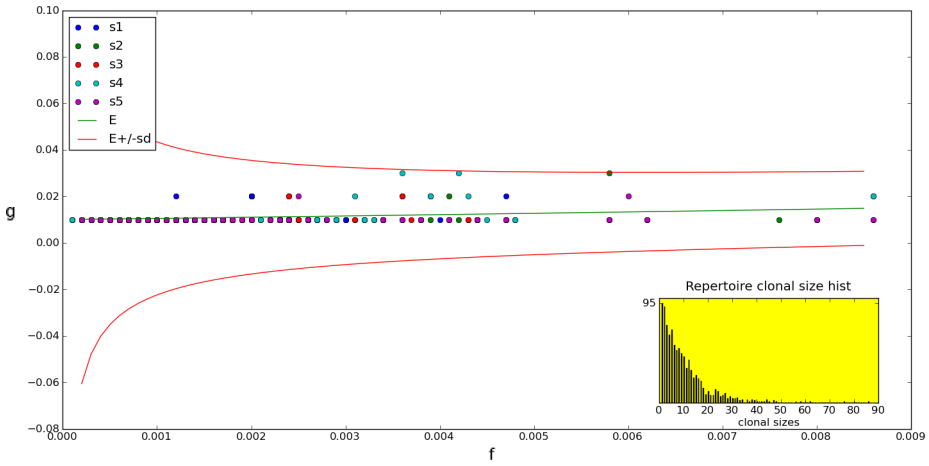

Figure 4.40: Geometric repertoire with mean 10. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 991$.

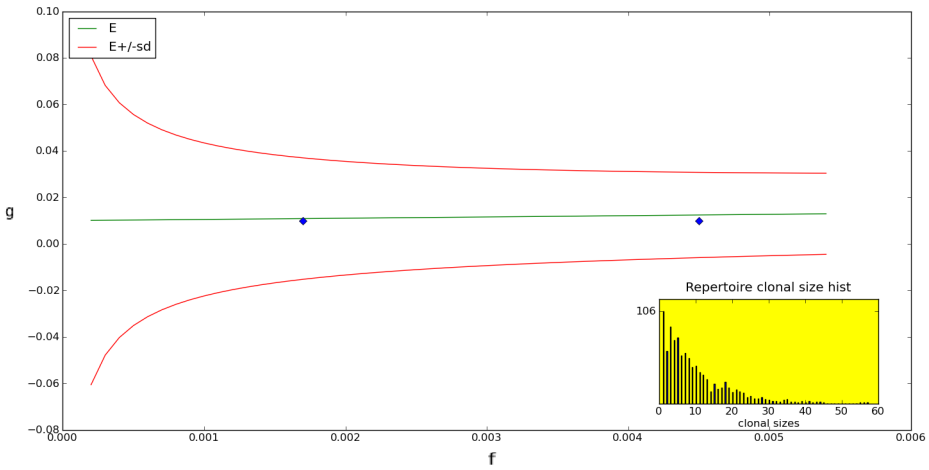

Figure 4.41: Geometric repertoire with mean 10. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 1002$.

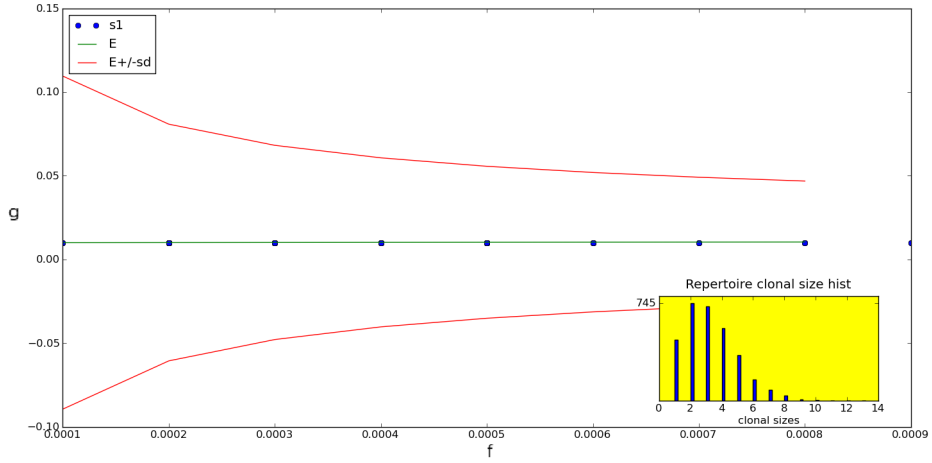

Figure 4.42: Poisson repertoire with mean 3. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 3118$.

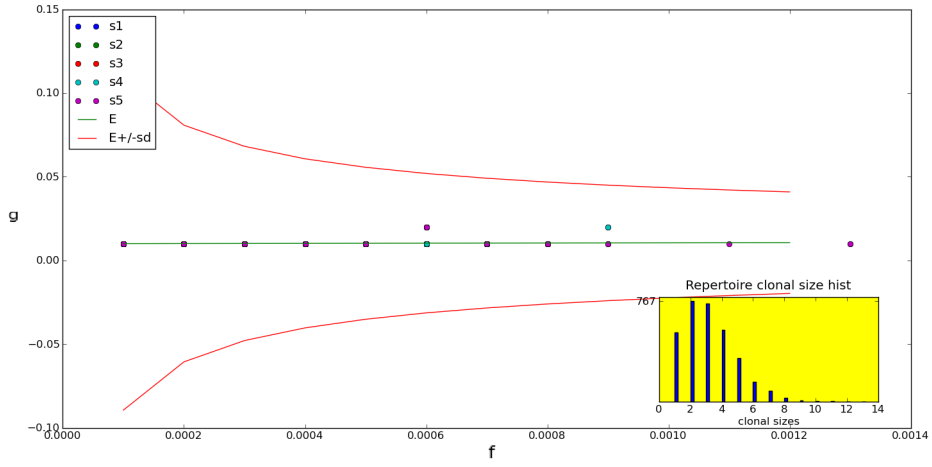

Figure 4.43: Poisson repertoire with mean 3. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 3202$.

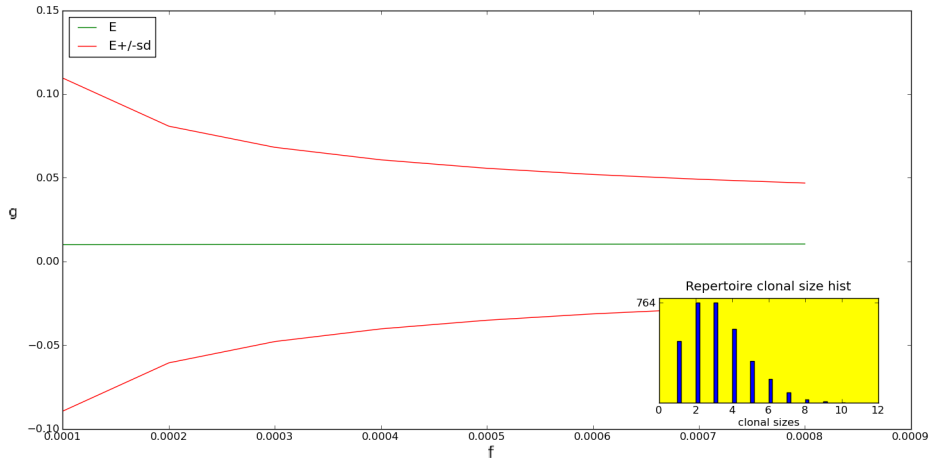

Figure 4.44: Poisson repertoire with mean 3. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 3166$.

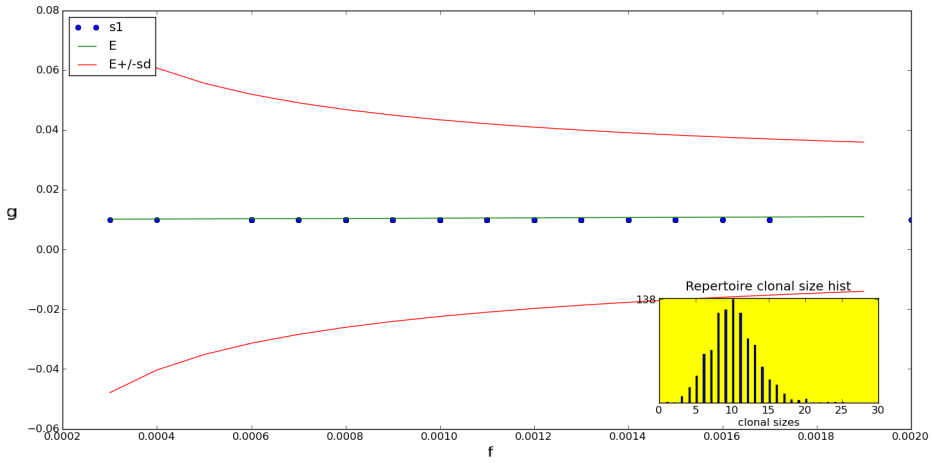

Figure 4.45: Poisson repertoire with mean 10. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 995$.

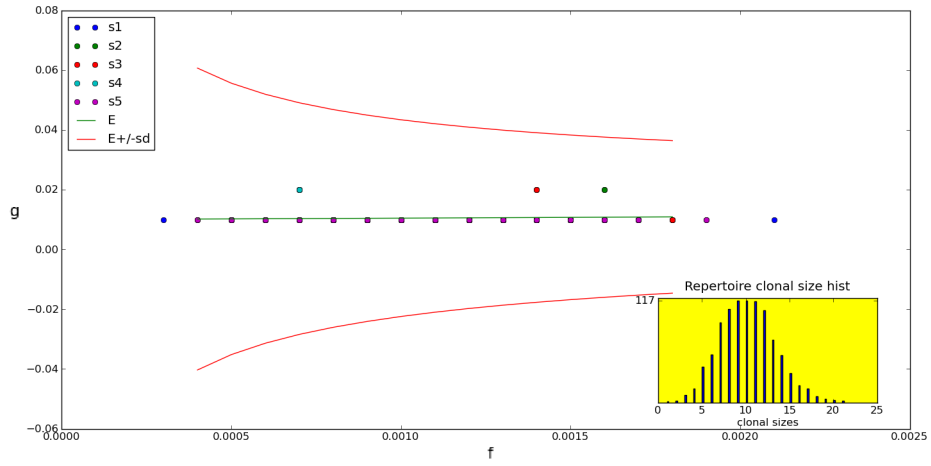

Figure 4.46: Poisson repertoire with mean 10. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 991$.

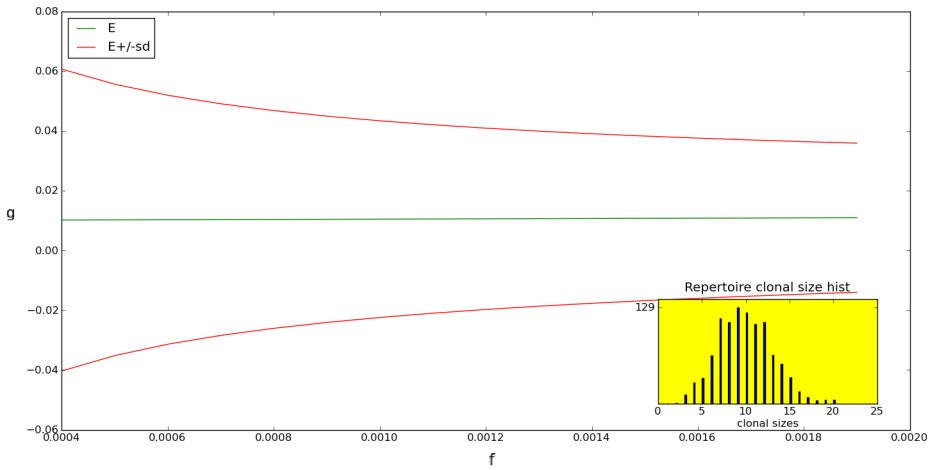

Figure 4.47: Poisson repertoire with mean 10. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 1019$.

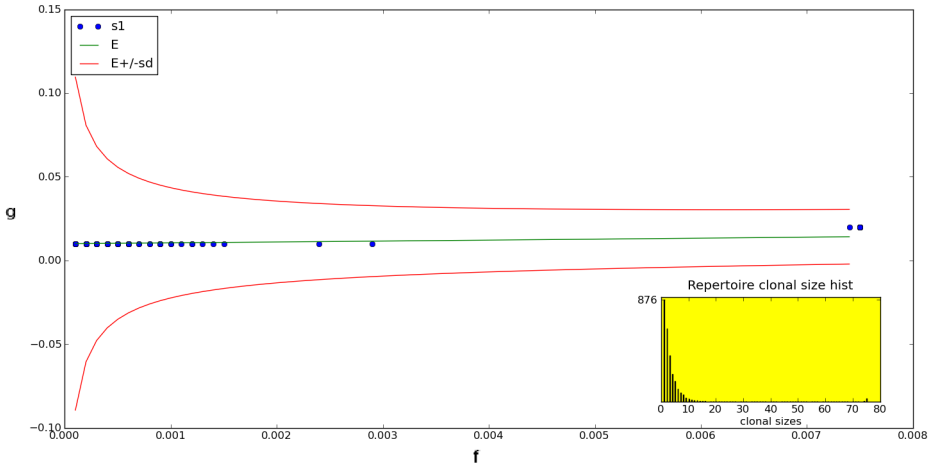

Figure 4.48: Heterogeneous repertoire: unexpanded part geometric with mean 3 and expanded part (0.01 of total clones) constant with mean 75. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 2688$.

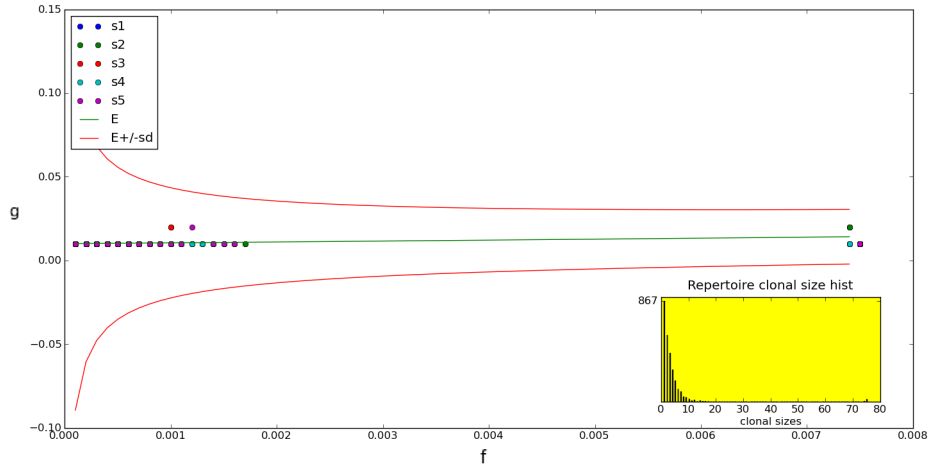

Figure 4.49: Heterogeneous repertoire: unexpanded part geometric with mean 3 and expanded part (0.01 of total clones) constant with mean 75. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 2687$.

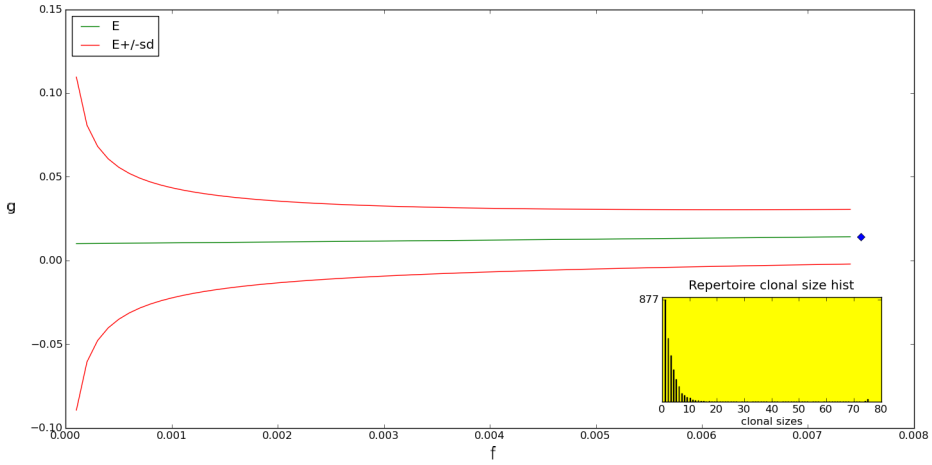

Figure 4.50: Heterogeneous repertoire: unexpanded part geometric with mean 3 and expanded part (0.01 of total clones) constant with mean 75. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 2677$.

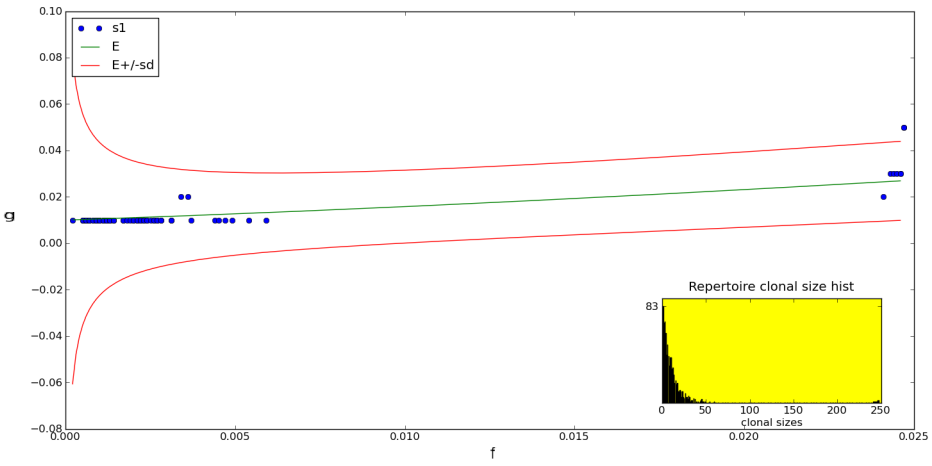

Figure 4.51: Heterogeneous repertoire: unexpanded part geometric with mean 10 and expanded part (0.01 of total clones) constant with mean 250. One sample of size 100 is taken. Parameters: $S = 10^4$, $N = 806$.

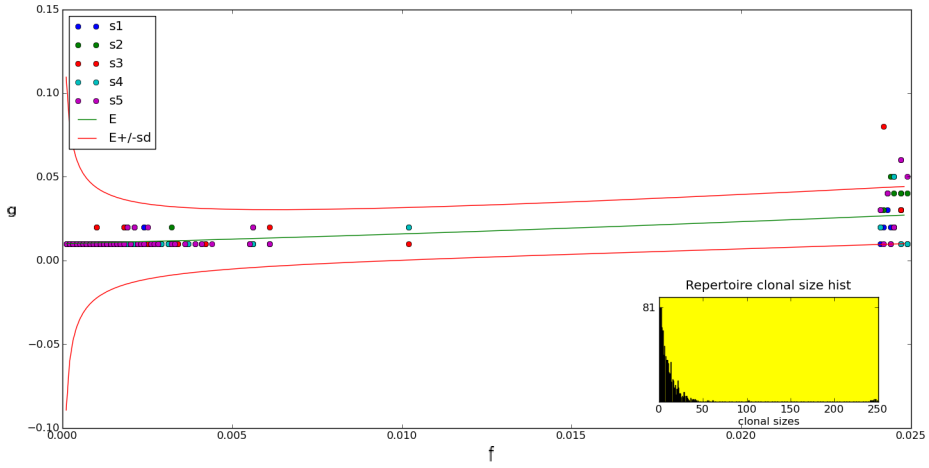

Figure 4.52: Heterogeneous repertoire: unexpanded part geometric with mean 10 and expanded part (0.01 of total clones) constant with mean 250. Five samples of size 100 are taken. Parameters: $S = 10^4$, $N = 806$.

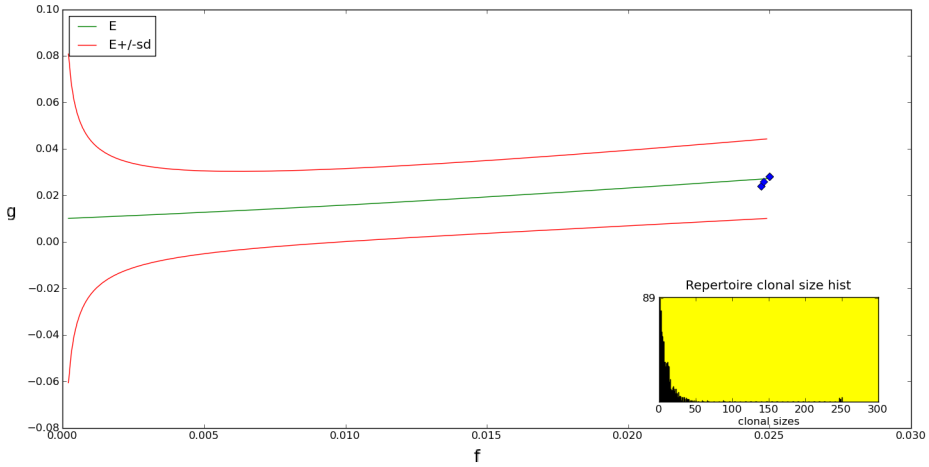

Figure 4.53: Heterogeneous repertoire: unexpanded part geometric with mean 10 and expanded part (0.01 of total clones) constant with mean 10*25. Five samples of size 100 are taken and only the common classes (common to all samples) are displayed. Parameters: $S = 10^4$, $N = 807$.

4.9 Data frequencies and implications on frequencies in repertoire

This section presents the frequencies for all V and J segments in the 10 different mice. If the samples were of the same sizes, the reader could directly apply the techniques developed in Section 4.8 to obtain an estimate of the respective frequencies in the repertoire. In fact, the expected value (4.8), representing the mean of the random variable Y_i^{obs}, has to be considered over samples of the same size m. Our data do not follow this criterion, as we have different sample sizes for each mouse. There might be a way out of this situation though, considering the following reasoning. Define m_1 and m_2 as two different sample sizes, and S as the repertoire size. As long as $m_1 \approx m_2$ and $m_1, m_2 \ll S$, we can easily verify (even just by simulations) that a binomial distribution with parameters n_i and m_1/S would not be distinguishable from another binomial distribution with parameters n_i and m_2/S. This is in fact our current situation, as we have different similar sample sizes m_j for $j = 1, \ldots, 5$ and a binomial distribution for the random variable Y_i, representing the number of T cells of type i in the sample. Similar reasoning holds for the random variable Y_i^{obs}, represented by the zero-truncated Y_i. For this reason, we apply here the analyses of Section 4.8 to obtain some estimates of the V and J segments in the GP33-specific repertoire of naïve, immunized and infected mice. We start by plotting the frequencies in Figures 4.54 to 4.61. See Appendix C for more details.

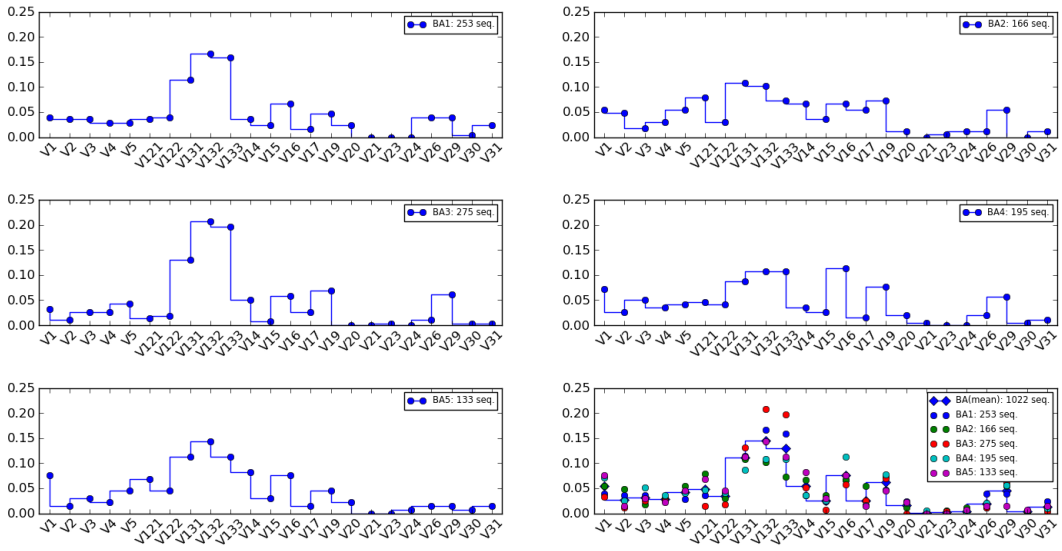

Figure 4.54: V frequencies for naïve mice.

In order to apply the techniques of Section 4.8, we need to use a unique value for the sample size, which we define as the mean over the different values. In particular, we will

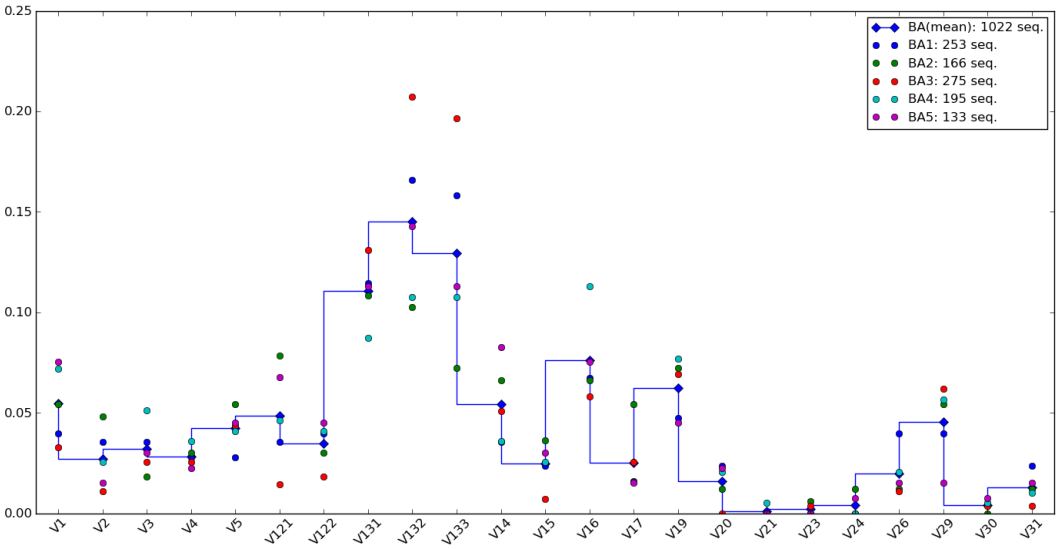

Figure 4.55: V frequencies for naïve mice and mean V frequency.

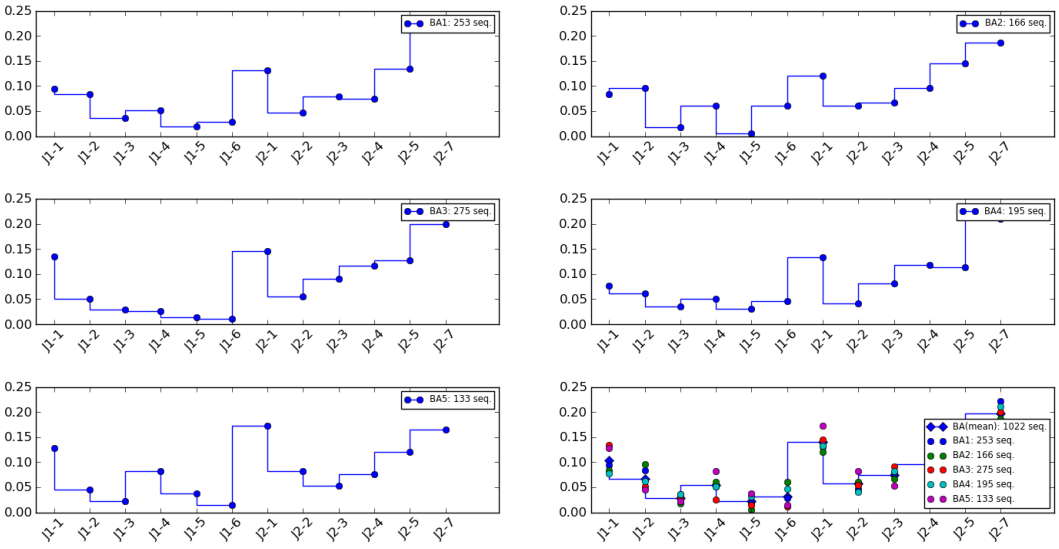

Figure 4.56: J frequencies for naïve mice.

use $m_{BA} = 205$ and $m_{EF} = 208$ for the naïve and infected respectively.

Mean frequencies for both V and J segments are displayed in Tables 4.3 and 4.4, subdivided in the two groups of mice (naïve and immunized/infected). It can be immediately seen that the V segments V_{21} V_{23} and V_{24} are never found in immunized/infected mice. All the other average values are above zero, wrongly suggesting that we could substitute these values in (4.21) in place of $\mathbb{E}(g_i)$ to approximate f_i. In fact, g_i represents the frequency of

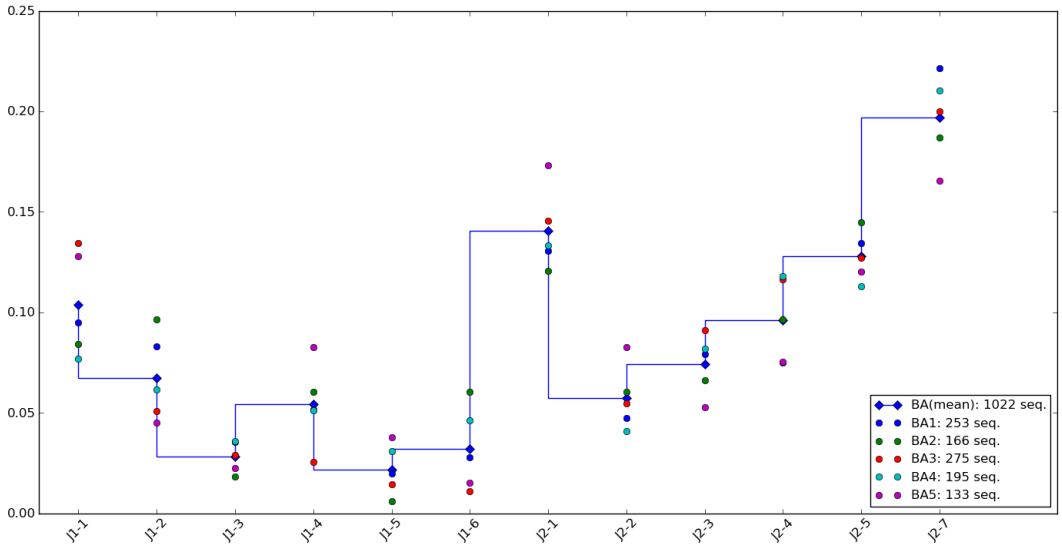

Figure 4.57: J frequencies for naïve mice and mean J frequency.

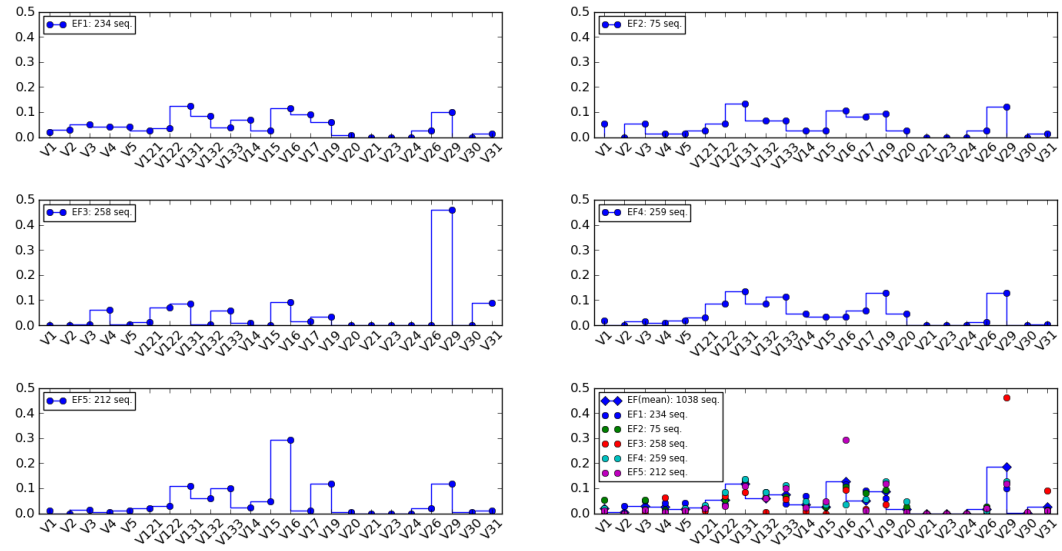

Figure 4.58: V frequencies for immunized and infected mice.

a particular type i (V or J particular segment in our case) which is actually observed in the samples. Therefore, before applying (4.21), we need to understand if and which are the V and J segments absent in one or more mice. In these cases, we won't be able to estimate the respective frequency in the repertoire. Of course, one could apply the same equation only to those mice with actually these particular V or J segments missing from the other mice, but we decided not to do it here, given the low number of available samples.

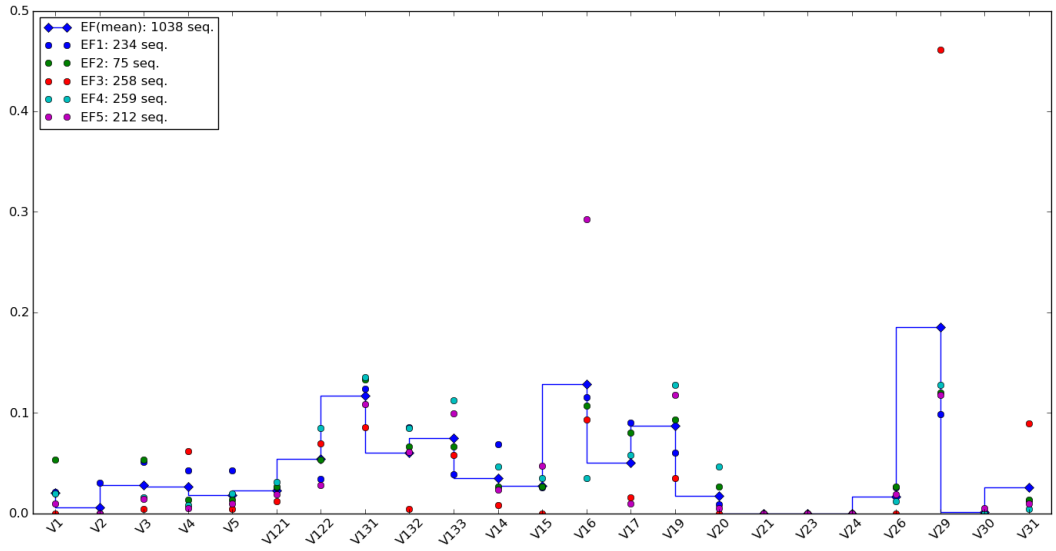

Figure 4.59: V frequencies for immunized and infected mice and mean V frequency.

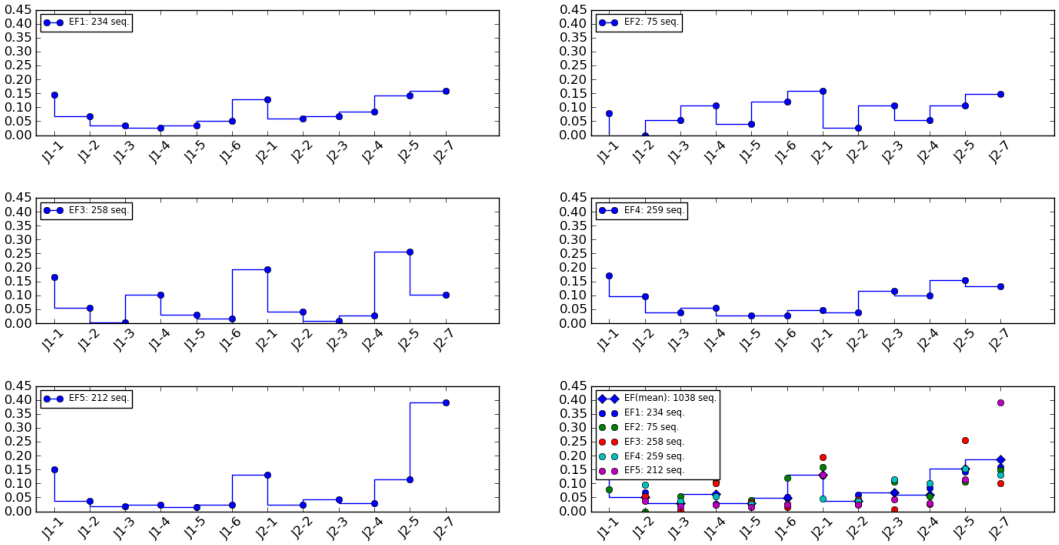

Figure 4.60: J frequencies for immunized and infected mice.

Looking at the data, we realize that all J segments are present in all mice. Regarding the V segments, here the result of what is missing for each mouse:

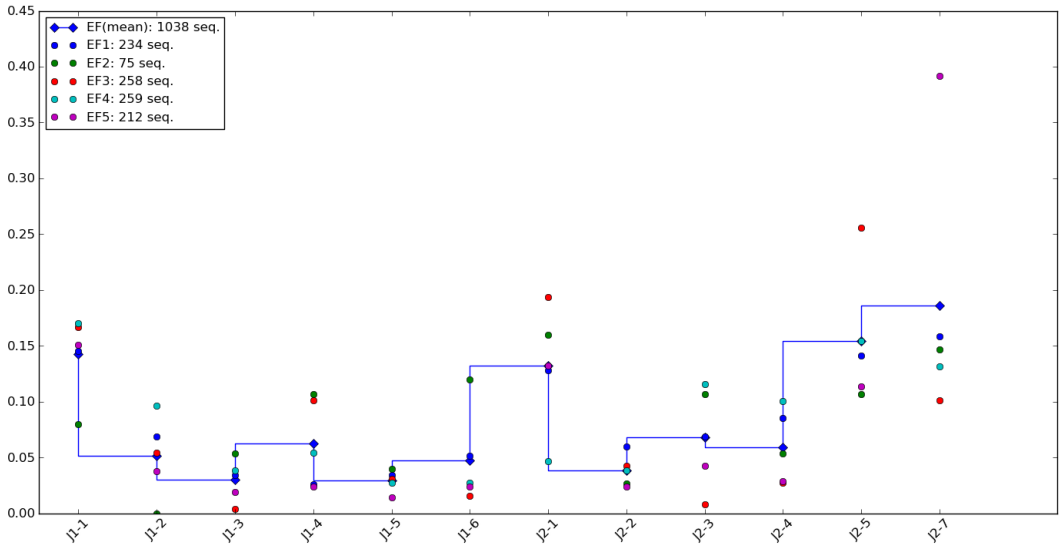

Figure 4.61: J frequencies for immunized and infected mice and mean J frequency.

V segment	BA1	BA2	BA3	BA4	BA5	EF1	EF2	EF3	EF4	EF5
V_1								X		
V_2							X	X	X	X
V_{15}								X		
V_{20}			X					X		
V_{21}	X	X	X		X	X	X	X	X	X
V_{23}	X			X	X	X	X	X	X	X
V_{24}	X		X	X		X	X	X	X	X
V_{26}								X		
V_{30}						X	X	X	X	

We now give an example of how we would apply (4.21) to the described experimental data, and in particular to the estimation of the frequency of the V_1 gene segment in the repertoire. As previously shown, the V_1 gene segment is present in all 5 samples from the 5 naïve mice. Therefore, its average frequency (i.e., $\mathbb{E}(g_i) = 0.0547$) shown in Table 4.3 in the BAs column is free from interference due to mice without that particular gene segment. Considering $m_{BA} = 205$, we obtain $[1/m_{BA}, 2/m_{BA}] = [0.0048, 0.0097]$. $\mathbb{E}(g_i)$ falls out of this interval, therefore suggesting that we should use the second part of (4.21). Thus, we believe that the percentage of V_1 gene segments in the whole repertoire of the naïve mice is around 5.47%.

	BAs	EFs
V_1	$5.47 \cdot 10^{-2}$	$2.07 \cdot 10^{-2}$
V_2	$2.71 \cdot 10^{-2}$	$5.98 \cdot 10^{-3}$
V_3	$3.21 \cdot 10^{-2}$	$2.76 \cdot 10^{-2}$
V_4	$2.83 \cdot 10^{-2}$	$2.61 \cdot 10^{-2}$
V_5	$4.23 \cdot 10^{-2}$	$1.77 \cdot 10^{-2}$
V_{12-1}	$4.85 \cdot 10^{-2}$	$2.27 \cdot 10^{-2}$
V_{12-2}	$3.48 \cdot 10^{-2}$	$5.41 \cdot 10^{-2}$
V_{13-1}	0.11	0.12
V_{13-2}	0.15	$6.05 \cdot 10^{-2}$
V_{13-3}	0.13	$7.49 \cdot 10^{-2}$
V_{14}	$5.43 \cdot 10^{-2}$	$3.45 \cdot 10^{-2}$
V_{15}	$2.46 \cdot 10^{-2}$	$2.68 \cdot 10^{-2}$
V_{16}	$7.59 \cdot 10^{-2}$	0.13
V_{17}	$2.52 \cdot 10^{-2}$	$5.05 \cdot 10^{-2}$
V_{19}	$6.22 \cdot 10^{-2}$	$8.67 \cdot 10^{-2}$
V_{20}	$1.58 \cdot 10^{-2}$	$1.73 \cdot 10^{-2}$
V_{21}	$1.03 \cdot 10^{-3}$	0
V_{23}	$1.93 \cdot 10^{-3}$	0
V_{24}	$3.91 \cdot 10^{-3}$	0
V_{26}	$1.96 \cdot 10^{-2}$	$1.66 \cdot 10^{-2}$
V_{29}	$4.54 \cdot 10^{-2}$	0.18
V_{30}	$4.05 \cdot 10^{-3}$	$9.43 \cdot 10^{-4}$
V_{31}	$1.29 \cdot 10^{-2}$	$2.57 \cdot 10^{-2}$

Table 4.3: V means over the five naïve and five immunized/infected mice, that is including mice without some V genes.

	BAs	EFs
J_{1-1}	0.1	0.14
J_{1-2}	$6.74 \cdot 10^{-2}$	$5.14 \cdot 10^{-2}$
J_{1-3}	$2.82 \cdot 10^{-2}$	$2.98 \cdot 10^{-2}$
J_{1-4}	$5.42 \cdot 10^{-2}$	$6.21 \cdot 10^{-2}$
J_{1-5}	$2.17 \cdot 10^{-2}$	$2.93 \cdot 10^{-2}$
J_{1-6}	$3.2 \cdot 10^{-2}$	$4.75 \cdot 10^{-2}$
J_{2-1}	0.14	0.13
J_{2-2}	$5.72 \cdot 10^{-2}$	$3.83 \cdot 10^{-2}$
J_{2-3}	$7.42 \cdot 10^{-2}$	$6.82 \cdot 10^{-2}$
J_{2-4}	$9.62 \cdot 10^{-2}$	$5.89 \cdot 10^{-2}$
J_{2-5}	0.13	0.15
J_{2-7}	0.2	0.19

Table 4.4: J means over the five naïve and five immunized/infected mice, that is including mice without some J genes.

4.10 Discussion

This chapter focused on exploratory analyses of the data described in Section 4.2 and on some probabilistic results connecting a particular frequency in the repertoire with frequencies in different samples. A clear preference for V_{13-1}, V_{13-2} and V_{13-3} in naïve mice was shown in Figure 4.12. The same figure also shows a preference for J_2 genes in naïve mice with respect to J_1 genes. This result can be also seen in Figure 4.30. The importance of V_{19} for infected mice is shown in Figure 4.13. Interestingly, Figure 4.28 shows how important the diversity of the J component is for the general diversity of VJs. The V diversity of mouse BA3 is low with respect to the other mice, but a value of J diversity similar to the other mice implies a level of total VJ diversity comparable to all the mice. Figure 4.29 shows the p-values for the randomization test on different mice. The data exclude the hypothesis of similar diversity distribution for all the couples of mice a part from the couples BA2-BA3, BA3-BA4, BA3-BA5, EF1-EF2, and EF2-EF4. The last pair is particularly interesting, representing the two previously immunized mice. This could indicate that these two mice have reached a similar diversity distribution due to immunization. A higher level of clonal sharing among naïve mice rather than among infected mice is shown in Figures 4.32 and 4.33. The major innovative point of this chapter is represented by (4.21). The importance of this formula depends on the kind of data we would like to use it for. In fact, $\mathbb{E}(g_i)$ represents the average of g_i taken over all those samples where the clonotype i was actually observed. This means that to estimate the frequency f_i of a clonotype with this equation, we should first observe the clonotype in all of our samples (or at least 2 samples). This could be very challenging for a general clonotype but could become much easier if we considered V, J or even VJ classes, rather than clonotype classes. It is worth noting that the formula is perfectly able to work at different levels of diversity, clonotype classes included, although the currently available single-cell technologies are not developed yet to produce enough frequency data for clonotype classes. To properly understand the level of diversity on which our formula can work nowadays, see Figure 4.62.

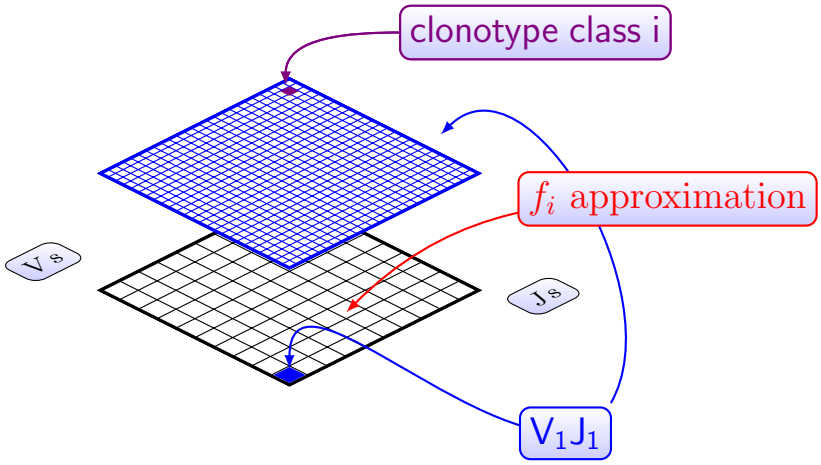

Figure 4.62: Example of the level of diversity (VJs) where (4.21) could work rather than at the clonotype class level.

Chapter 5

Markov chains and TCR repertoire renewal

5.1 Abstract and Introduction

Let us consider a TCR repertoire at a given time t with N distinct clonotype classes. It is well known that the thymus is constantly producing new clonotypes, while in the periphery a certain diversity is maintained due to a balanced birth and death process based on competition among classes for biological signal. The biological questions we address concern the time evolution of diversity in the repertoire. In particular, we first explore (i) the random variable describing the average time at which a given percentage of the original N clonotype classes have disappeared from the repertoire due to competition or natural death, (ii) the size of the repertoire at such time, and (iii) the maximum repertoire diversity achieved in this time interval. We believe that these questions are of foremost importance in order to understand the real value of sampling from a repertoire and try to estimate its diversity at a given point in time. These three points will be described as specific stochastic descriptors, in connection to the Markov model that will be explained in the following section.

5.2 Mathematical model

We present here an unidimensional continuous-time Markov chain (CTMC) representing the dynamical process of competition among clonotype classes. In particular, we consider the CTMC $\mathcal{X} = \{X(t) : t \geq 0\}$ defined over the space of states $\mathbb{S} = \{0, 1, 2, \ldots\}$ and with initial condition $X(0) = N$, where $X(t)$ represents the number of distinct clonotype classes present in the repertoire at time $t \geq 0$. The birth rate λ_n represents thymic production of new clonotypes in the repertoire, and it is described by a constant thymic output rate $\lambda_n \equiv \theta$, as defined in [100]. The extinction of a given clonotype, in reality dependent on reception of survival stimuli from the environment and on competition among classes [64], is represented here by a rate μ_n of transition of process $\mathcal{X}$ from state n to state $n-1$; see Figure 5.1. In the following sections, we propose two different choices for μ_n, denoted $\mu_n^{(1)}$ and $\mu_n^{(2)}$, which incorporate clonotype competition in two different ways. According to the choice of μ_n, we will label the process $\mathcal{X}$ as $\mathcal{X}_{(1)}$ or $\mathcal{X}_{(2)}$.

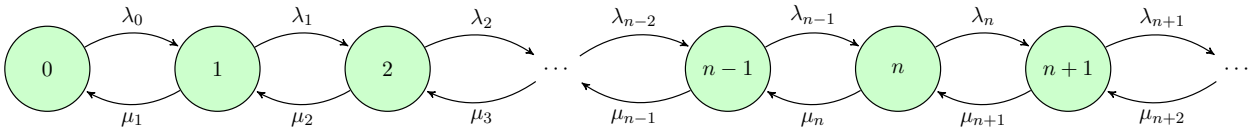

Figure 5.1: Continuous-time birth-and-death process $\mathcal{X}$.

5.2.1 Implicit competition

We consider the process $\mathcal{X}_{(1)}$ with a linear death rate $\mu_n^{(1)} = \tilde{\mu} n$, where $\tilde{\mu}^{-1}$ is defined as the average survival time of a clonotype in the repertoire, where clonotypes are assumed to act independently from each other. This average survival time was previously evaluated in [100] as

$$T(\alpha, n_\theta) = \frac{1}{\alpha\mu}(\gamma_E - e^{\alpha n_\theta} \cdot \mathrm{Ei}(-\alpha n_\theta) + \log(\alpha n_\theta)), \tag{5.1}$$

where γ_E represents the Euler-Mascheroni constant, $\mathrm{Ei}(x)$ represents the exponential integral defined as

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt$$

and $\alpha = \dfrac{\theta n_\theta}{\gamma M}$ represents the strength of the thymic production relative to the peripheral division. Moreover, the main assumptions in [100] are that: (i) each clonotype comes out of the thymus at a rate θ and a fixed size n_θ, and (ii) the environment is populated by M distinct self pMHC subsets. The parameter μ in (5.1) represents the single cell death rate, while γ represents the single cell division rate. The dependence of (5.1) on parameters n_θ, M, γ, θ, μ is shown in Appendix D.

We are therefore expressing competition among clonotypes intrinsically through the rate $\tilde{\mu}$. $T(\alpha, n_\theta)$ is in fact defined as the mean time until the stochastic process $\mathbf{X_t}$, where $\mathbf{X_t}$ is defined as the diffusion process on the real line approximating the number of T cells $n_i(t)$ of clonotype class i. The process $\mathbf{X_t}$ satisfies the stochastic differential equation (15) in [100]

$$d\mathbf{X_t} = -\alpha\mu\mathbf{X_t}d\mathbf{X_t} + \sqrt{2\mu\mathbf{X_t}}d\mathbf{W_t},$$

where $\mathbf{W_t}$ represents a Wiener process. The term $-\alpha\mu$ in this stochastic differential equation is the term including competition among clonotypes in the general scenario, therefore inducing implicit competition in the time $T(\alpha, n_\theta)$.

Thus, we set $\tilde{\mu}^{-1} = T(\alpha, n_\theta)$, where the values of the parameters in (5.1) will be chosen for our numerical results according to Table 1 in [100], and to the rescaling process described in Section 5.7, with $\mu_n = \mu_n^{(1)} = \tilde{\mu}n$ for process $\mathcal{X}_{(1)}$.

5.2.2 Explicit competition

We consider here a second alternative for the choice of μ_n, defining the process $\mathcal{X}_{(2)}$ with death rate $\mu_n = \mu_n^{(2)} = n(\beta_1 + \beta_2 p(n-1))$. In process $\mathcal{X}_{(2)}$, parameter $\beta_1 n$ represents a linear contribution, while parameter $\beta_2 p n^2$ is used to model clonotype competition in a similar way to Mathematical Ecology models [90]. In particular, the parameter $p = 10^{-6}$ is defined in [100] as the probability that any given self pMHC is recognised by a randomly-selected T-cell clonotype. This quadratic term can be seen as the environmental pressure a single clonotype is subject from that fraction p of clonotypes with which it competes. In order to give biological meaning to the parameters β_1 and β_2, we need to find two different equations relating β_1 and β_2. We start by recalling that in Section 5.2.1 the death rate was defined as $\mu_n^{(1)} = \tilde{\mu}n$, where $\tilde{\mu} = [T(\alpha, n_\theta)]^{-1}$. We notice that the authors in [100] compute the time $T(\alpha, n_\theta)$ as the average extinction time of a single clonotype in a repertoire of an average number of clonotypes N^*. Thus, we assume $\mu_n^{(1)}$ to be equal to $\mu_n^{(2)}$ for the particular case $n = N^*$. This allows us to write the first equation

$$\mu_{N^*}^{(1)} = \mu_{N^*}^{(2)} \quad \Rightarrow \quad \beta_2 = \frac{\tilde{\mu} - \beta_1}{p(N^* - 1)}. \tag{5.2}$$

We choose $\tilde{\mu} > \beta_1$ so that $\beta_2 > 0$. This condition comes naturally from the definitions of β_2 and β_1. In fact they represent the environmental pressure that a clonotype is subject to when it belongs to an environment with multiple and no competing clonotypes respectively. We need now to find a second equation to pair with (5.2). We notice $\mu_1^{(2)} = \beta_1$, meaning that β_1 represents the death rate of a single clonotype subject to no environmental pressure. We consider the birth and death process $\mathcal{C} = \{C(t) : t \geq 0\}$ representing the number of T cells belonging to this clonotype subject to no competition, where the death rate is μ (0.5 year^{-1} for human, 1 month^{-1} for mouse), the birth rate is γ (10 year^{-1} for human,

10^{-4} month^{-1} for mouse) and the initial state is $C(0) = n_\theta$. The parameters μ, γ and n_θ were defined in [100]. Define T_{0,n_θ} as the time until absorption of process $\mathcal{C}$. Therefore we can assume $\beta_1 = [T_{0,n_\theta}]^{-1}$. It is clear that, focusing on mice, $T_{0,n_\theta} < +\infty$ as $\mu > \gamma$. The same does not hold for humans as $\mu < \gamma$. Define $\eta_{0,n_\theta} = \mathbb{E}\left[T_{0,n_\theta} | T_{0,n_\theta} < +\infty\right]$. Therefore for this case we choose

$$\beta_1 = \frac{\Pr(T_{0,n_\theta} < +\infty)}{\eta_{0,n_\theta}}. \tag{5.3}$$

It is possible to analyse η_{0,n_θ} with a first step argument, considering a maximum number of T cells S belonging to a clonotype class; see Gillespie simulations for η_{0,n_θ} in Figure 5.2, and Figure 5.3 for simulations of $\Pr(T_{0,n_\theta} < +\infty)$. Figure 5.4 shows the dependence of the parameter β_2 on β_1.

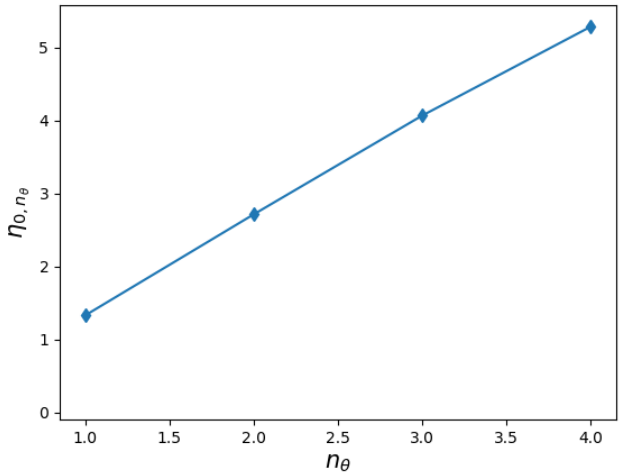

Figure 5.2: Gillespie simulations of η_{0,n_θ}. Parameters in accordance with Section 5.7: $\mu = 0.5$ year^{-1}, $\gamma = 1.25$ year^{-1}, and maximum number of T cells allowed in a clonotype class $S = 1000$. Number of simulations $= 10^5$.

5.3 Certainty of first visit to state 0 in finite mean time

In this section we focus on proving that process $\mathcal{X}_{(2)}$ visits the state 0 with probability 1 and in finite mean time. Similar arguments apply to process $\mathcal{X}_{(1)}$ and are here omitted. To this goal, we consider the state 0 to be an absorbing state; see Figure 5.5.

Thus, process $\mathcal{X}_{(2)}$ can be seen as a birth-and-death process defined on $S = \{0\} \cup \mathcal{C}$, with $\mathcal{C} = \{1, 2, \ldots\}$ and 0 being the absorbing state; see Chapter 6 in [5]. Our goal is therefore to prove that for any initial state x, $\alpha(x) = \lim_{t \to +\infty} \Pr(X(t) = 0 | X(0) = x) = 1$. This, accordingly to Theorem 6.2 of [5], occurs if and only if

$$\sum_{k=1}^{+\infty} \frac{\mu_1 \mu_2 \cdots \mu_k}{\lambda_1 \lambda_2 \cdots \lambda_k} = +\infty,$$

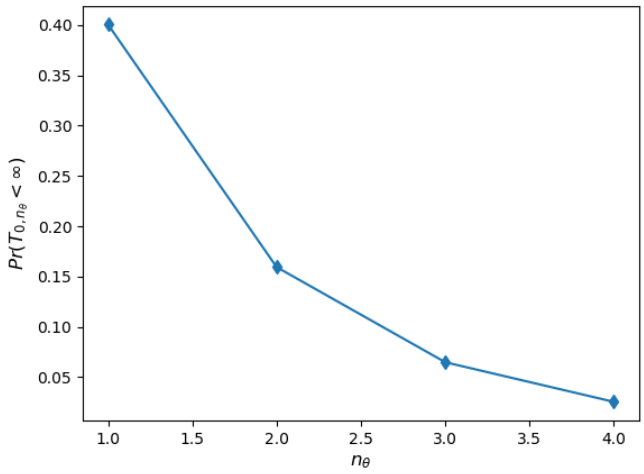

Figure 5.3: Gillespie simulations of $\Pr(T_{0,n_\theta} < +\infty)$. Parameters in accordance to Section 5.7: $\mu = 0.5$ year^{-1}, $\gamma = 1.25$ year^{-1}, and maximum number of T cells allowed in a clonotype class $S = 1000$. Number of simulations $= 10^5$.

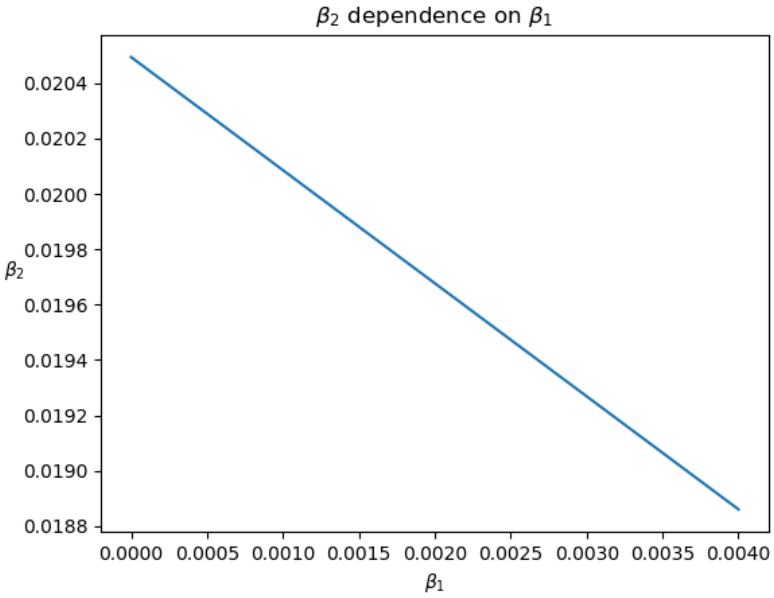

Figure 5.4: Plot of (5.2). Parameters in accordance to Section 5.7: $\mu = 0.5$ year^{-1}, $\gamma = 1.25$ year^{-1}, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $p = 0.05$, and $N^* = 50$. Number of simulations $= 10^5$.

where $\lambda_k \equiv \theta$ and $\mu_k = k(\beta_1 + \beta_2 pk)$ for $k \in \mathcal{C}$. We have

$$\sum_{k=1}^{+\infty} \frac{\mu_1 \mu_2 \cdots \mu_k}{\lambda_1 \lambda_2 \cdots \lambda_k} = \sum_{k=1}^{+\infty} \frac{(\beta_1 + \beta_2 p)(2\beta_1 + 4\beta_2 p)(3\beta_1 + 9\beta_2 p) \cdots (k\beta_1 + k^2 \beta_2 p)}{\theta^k}$$

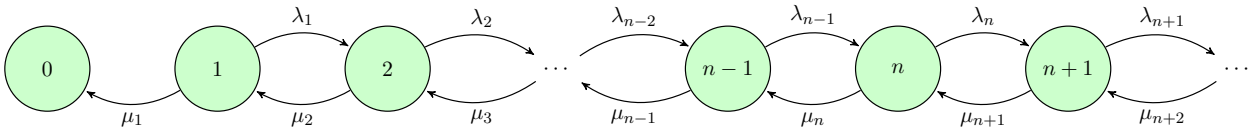

Figure 5.5: Continuous-time birth-and-death process $\mathcal{X}$.

which is bounded by the case $\beta_2 p = 0$, that is

$$\sum_{k=1}^{+\infty} \frac{\mu_1 \mu_2 \cdots \mu_k}{\lambda_1 \lambda_2 \cdots \lambda_k} \geq \sum_{k=1}^{+\infty} \frac{\beta_1^k k!}{\theta^k} = +\infty.$$

The last equality holds because $\forall x \in \mathbb{R}^+$, $\exists k \in \mathbb{N}$ such that $k! > x^k$. Thus, $\alpha(x) = 1\ \forall x \in S$. We note here that $\alpha(x)$ can be re-expressed as $\alpha(x) = \Pr(T_x < +\infty)$, where T_x represents the time until absorption for the initial state $X(0) = x$. We can prove that $\mathbb{E}[T_x] < +\infty$ by considering Theorem 6.3 of [5]. In particular, $\mathbb{E}[T_x] < +\infty$ if and only if

$$\sum_{k=2}^{+\infty} \frac{\lambda_1 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} < +\infty.$$

We have

$$\begin{aligned}\sum_{k=2}^{+\infty} \frac{\lambda_1 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} &= \sum_{k=2}^{+\infty} \frac{\theta^{k-1}}{(\beta_1 + \beta_2 p)(2\beta_1 + 4\beta_2 p)(3\beta_1 + 9\beta_2 p) \cdots (k\beta_1 + k^2 \beta_2 p)} \\ &\leq \sum_{k=2}^{+\infty} \frac{\theta^{k-1}}{\beta_1^k k!} < +\infty,\end{aligned}$$

so that $\mathbb{E}(T_x) < +\infty$.

5.4 Time $T_N(A)$ from N to A original clonotypes in the repertoire

Our interest here is to analyse the random variable $T_N(A)$ representing the time to reach for the first time a number $A < X(0) = N$ of original clonotypes in the repertoire. The main reason behind the study of this stochastic descriptor is the quest to understand the timings of regenerative capabilities of a repertoire. We follow here a first step argument. In order to analyse the random variable $T_N(A)$, we need to keep track of the original clonotypes as the stochastic process evolves. This is due to the fact that $T_N(A)$ is not the time of first visit to state A; that is, $T_N(A) \neq \inf\{t \geq 0 : X(t) = A\}$, since when process $\mathcal{X}$ reaches state A, the A clonotypes remaining in the repertoire are not necessarily among the original ones (some of them might be new ones as a result of thymic output). Thus we consider an auxiliary random variable $Y(t)$ and an augmented process $\mathcal{X}^{aug} = \{(X(t), Y(t)) : t \geq 0\}$ defined on $\mathbb{S}^{aug} = \{(n, m) : m \in \{0, 1, 2, \ldots, X(0)\},\ n \geq m\}$, where $Y(t)$ amounts to the

number of original clonotypes in the repertoire at time $t \geq 0$, which constantly decreases. Thus we have $T_N(A) = \inf\{t \geq 0 : Y(t) = A\}$. For process $\mathcal{X}^{aug}$ we define $\mu_{n,m}^{(X)}$ and $\mu_{n,m}^{(Y)}$ as the death rates of new and original clonotypes, respectively, accordingly to the original rate μ_n; see Figure 5.6. To note that, even though $X(t)$ represents the total number of clonotypes (original + newly created), we nevertheless use here $\mu_{n,m}^{(X)}$ to express the death rate of only the newly created clones. We do this with the only aim to ease the notation.

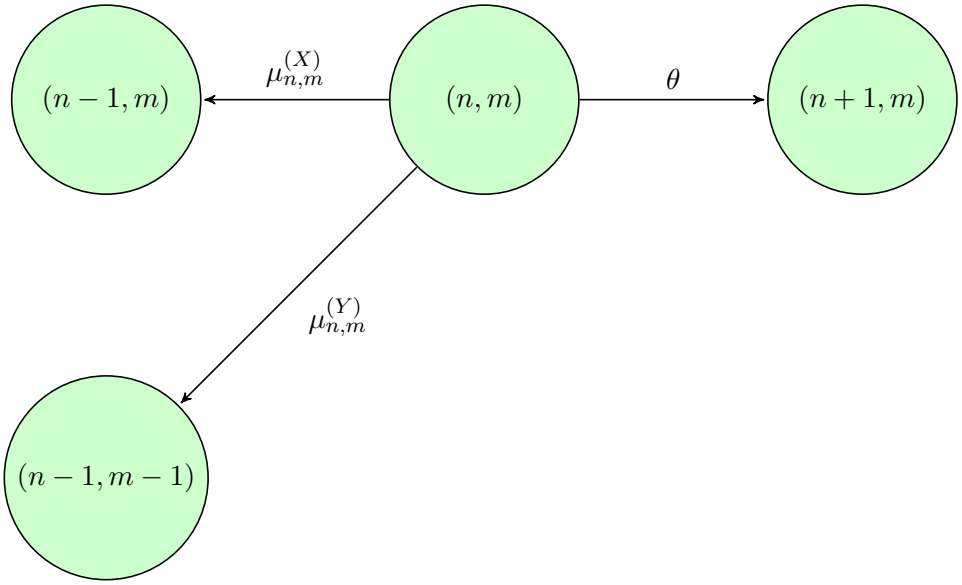

Figure 5.6: Transitions diagram for bivariate continuous-time birth-and-death process $\mathcal{X}^{aug}$.

5.4.1 Implicit competition

We consider here $\mu_{n,m}^{(X)} = \mu_{n,m}^{(X,1)} = \tilde{\mu}(n-m)$ and $\mu_{n,m}^{(Y)} = \mu_{n,m}^{(Y,1)} = \tilde{\mu}m$, so that $\mu_{n,m}^{(X,1)} + \mu_{n,m}^{(Y,1)} = \mu_n^{(1)}$, leading to the analysis of process $\mathcal{X}_{(1)}^{aug}$. These rates are directly obtained by assuming clonotypes going to extinction at a common rate $\tilde{\mu}$ in an independent fashion. We recall that $Y(t)$ is just an auxiliary variable keeping track of original clonotypes without affecting the dynamics of $X(t)$, so that both variables go necessarily to extinction with probability one in mean finite time, according to Section 5.3; see Figure 5.7. Moreover, since every clonotype behaves independently, the process $\mathcal{Y} = \{Y(t) : t \geq 0\}$ defined on $\{0, 1, 2, \ldots, X(0)\}$, is a pure-death linear process with death rate $\tilde{\mu}m$; see Figure 5.8.

Thus, $T_N(A)$ can be analysed by noting that $F_{T_N(A)}(t) = \Pr(T_N(A) \leq t) = \Pr(Y(t) \leq A)$ for any initial state $X(0) = Y(0) = N$. Moreover, from Section 6.4.2 of [5], we can write

$$p_k(t) = \Pr(Y(t) = k | Y(0) = N) = \binom{N}{k} e^{-k\tilde{\mu}t}(1 - e^{-\tilde{\mu}t})^{N-k}. \tag{5.4}$$

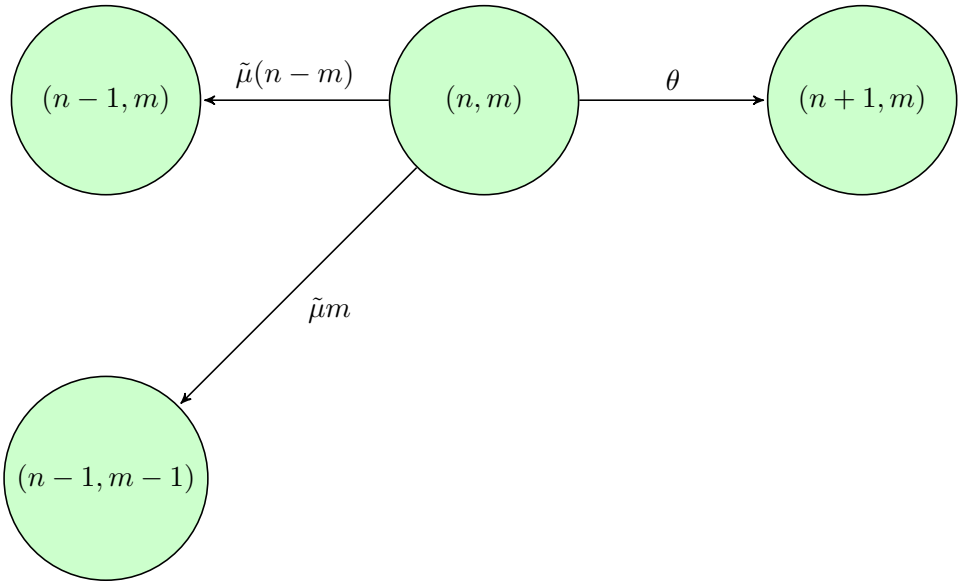

Figure 5.7: Bivariate continuous-time birth-and-death process $\mathcal{X}_{(1)}^{aug}$ with $\mu_{n,m}^{(X)} = \mu_{n,m}^{(X,1)} = \tilde{\mu}(n-m)$ and $\mu_{n,m}^{(Y)} = \mu_{n,m}^{(Y,1)} = \tilde{\mu}m$.

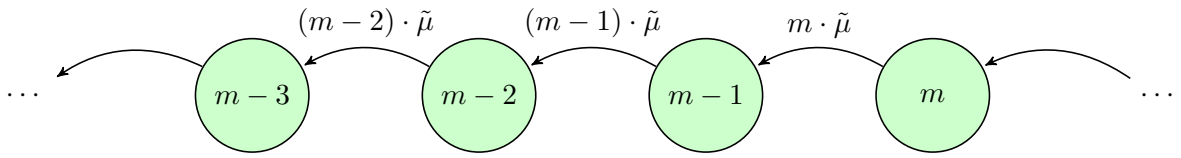

Figure 5.8: Continuous-time pure-death process $\mathcal{Y}$, representing the death of original clonotypes.

This equation can be seen as the different possible ways of choosing k surviving clonotypes at time t out of the initial N, where a clonotype survives until time t with probability $e^{-\tilde{\mu}t}$, leading to the binomial formula in (5.4). Given (5.4), we have

$$F_{T_N(A)}(t) = \Pr(Y(t) \leq A) = \sum_{k=0}^{A} p_k(t) = \sum_{k=0}^{A} \binom{N}{k} e^{-k\tilde{\mu}t}(1-e^{-\tilde{\mu}t})^{N-k}.$$

We would like to find a closed form for $F_{T_N(A)}(t)$, in order to compute the density function

$$f_T(t) = \frac{d}{dt} F_{T_N(A)}(t).$$

We follow now arguments of Section (3-7) in [167] in order to prove the last equality of

$$1 - F_{T_N(A)}(t) = 1 - \sum_{k=0}^{A} p_k(t) = \sum_{k=A+1}^{N} p_k(t) = I_{e^{-\tilde{\mu}t}}(A+1, N-A), \tag{5.5}$$

where $I_x(a,b) = B(x;a,b)/B(a,b)$ represents the regularized incomplete beta function,

with

$$B(x;a,b) = \int_0^x s^{a-1}(1-s)^{b-1}ds$$

and

$$B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

To prove the last equality of (5.5), we prove a more general case represented by Eq. (3-3) of Section 3-7 of [167], that is

$$\sum_{k=A+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = \frac{\int_0^p y^A (1-y)^{N-A-1}dy}{\int_0^1 y^A (1-y)^{N-A-1}dy} = I_p(A+1, N-A). \tag{5.6}$$

We start by defining

$$Q_{A+1} = \int_0^1 y^A (1-y)^{N-A-1}dy, \tag{5.7}$$

$$S_{A+1} = \int_0^p y^A (1-y)^{N-A-1}dy. \tag{5.8}$$

Recalling the Beta function

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \tag{5.9}$$

we have $Q_{A+1} = B(A+1, N-A)$. The Beta function verifies

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{5.10}$$

where $\Gamma(t)$ is the Gamma function, defined as

$$\Gamma(t) = \int_0^{+\infty} s^{t-1}e^{-s}ds \tag{5.11}$$

or, for any integer n, as

$$\Gamma(n) = (n-1)!\,. \tag{5.12}$$

Thus, we can write

$$Q_{A+1} = \frac{A!(N-A-1)!}{N!} = \left[(A+1)\binom{N}{A+1}\right]^{-1}. \tag{5.13}$$

We focus now on finding an expression for S_{A+1}. The following steps could also be applied in order to find (5.13). We recall the rule of integration by parts, that is

$$\int_a^b u\ dv = [uv]_a^b - \int_a^b v\ du \tag{5.14}$$

where $u = u(x)$ and $v = v(x)$ are functions of the variable x. Thus, using (5.14) and defining the function

$$q(x) = \frac{p^x(1-p)^{N-x}}{x}, \tag{5.15}$$

we can write

$$S_{A+1} = q(A+1) + \left(\frac{N-A-1}{A+1}\right) S_{A+2}. \tag{5.16}$$

This relationship can be written for the general case

$$S_a = q(a) + \left(\frac{n-a}{a}\right) S_{a+1} \quad \text{for} \quad a = A+1, A+2, \cdots, N-1 \tag{5.17}$$

and, for the case $a = N$, it becomes

$$S_N = q(N). \tag{5.18}$$

This brings us to the final expression

$$\begin{aligned} S_{A+1} = q(A+1) + \frac{N-A-1}{A+1} q(A+2) + \frac{(N-A-1)(N-(A+2))}{A+1(A+2)} q(A+2) + \cdots \\ + \frac{(N-A-1)(N-(A+1))\cdots(N-(N-1))}{A+1(A+2)\cdots(N-1)} q(N). \end{aligned} \tag{5.19}$$

Is is now sufficient to divide each single term of (5.19) by (5.13) to obtain

$$\sum_{k=A+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = \frac{\int_0^p y^A (1-y)^{N-A-1} dy}{\int_0^1 y^A (1-y)^{N-A-1} dy}. \tag{5.20}$$

This gives the relationship

$$\sum_{k=A+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = I_p(A+1, N-A). \tag{5.21}$$

Thus, using (5.5), we obtain

$$F_{T_N(A)}(t) = I_{1-e^{-\tilde{\mu}t}}(N-A, A+1), \tag{5.22}$$

using the property of the regularised incomplete beta function for which

$$I_x(a,b) = 1 - I_{1-x}(b,a). \tag{5.23}$$

To prove this property, we recall that $I_x(a,b) = B(x;a,b)/B(a,b)$, where

$$B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1} dt$$

and

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{(a-1)!(b-1)!}{(a+b-1)!},$$

and we notice that $B(a,b) = B(b,a)$ for obvious properties of the factorials. Therefore, in order to prove (5.23), we need to show that

$$\frac{1}{B(a,b)} \left[\int_0^x t^{a-1}(1-t)^{b-1} dt + \int_0^{1-x} q^{b-1}(1-q)^{a-1} dq \right] = 1.$$

We change the variable in the second integral, introducing the new variable $t = 1 - q$, which gives $dt = -dq$. Thus, we obtain

$$\int_0^{1-x} q^{b-1}(1-q)^{a-1}dq = -\int_1^x t^{a-1}(1-t)^{b-1}dt.$$

Therefore we have

$$\begin{aligned}\frac{1}{B(a,b)}\left[\int_0^x t^{a-1}(1-t)^{b-1}dt - \int_1^x t^{a-1}(1-t)^{b-1}dt\right] &= \frac{1}{B(a,b)}\left[\int_0^1 t^{a-1}(1-t)^{b-1}dt\right] \\ &= 1.\end{aligned}$$

This proves (5.23). Thus, from (5.22) we obtain

$$F_{T_N(A)}(t) = (N-A)\binom{N}{A}\int_0^{1-e^{-\tilde{\mu}t}} s^{N-A-1}(1-s)^A ds, \tag{5.24}$$

and the density function can be obtained as

$$f_{T_N(A)}(t) = \frac{d}{dt}F_{T_N(A)}(t).$$

In order to compute this derivative, we need to state the general form of the Leibniz integral rule:

$$\frac{d}{dt}\left(\int_{a(t)}^{b(t)} g(t,s)ds\right) = g(t,b(t))\cdot b'(t) - g(t,a(t))\cdot a'(t) + \int_{a(t)}^{b(t)}\frac{\partial}{\partial t}g(t,s)ds. \tag{5.25}$$

In our case, we have

$$a(t) \equiv 0, \;\; b(t) = 1 - e^{-\tilde{\mu}t} \;\text{ and }\; g(t,s) \equiv g(s) = (N-A)\binom{N}{A}s^{N-A-1}(1-s)^A.$$

Thus, we obtain

$$\begin{aligned} f_{T_N(A)}(t) &= g(1-e^{-\tilde{\mu}t})\cdot\frac{d}{dt}(1-e^{-\tilde{\mu}t}) \\ &= (N-A)\binom{N}{A}(1-e^{-\tilde{\mu}t})^{N-A-1}(e^{-\tilde{\mu}t})^A\cdot\tilde{\mu}e^{-\tilde{\mu}t} \\ &= (N-A)\binom{N}{A}(1-e^{-\tilde{\mu}t})^N(e^{\tilde{\mu}t}-1)^{-(A+1)}\tilde{\mu}. \end{aligned} \tag{5.26}$$

We check that the area under $f_{T_N(A)}(t)$ is actually 1, as it should be for a probability density function, that is

$$\int_0^{+\infty} f_{T_N(A)}(t) = (N-A)\binom{N}{A}\tilde{\mu}\int_0^{+\infty}\frac{(e^{\tilde{\mu}t}-1)^{N-A-1}}{e^{N\tilde{\mu}t}} = 1.$$

We introduce the new variable $p = e^{\tilde{\mu}t}$, which gives $dp = \tilde{\mu}e^{\tilde{\mu}t}dt$. Thus we can write

$$\int_0^{+\infty} f_{T_N(A)}(t) = (N-A)\binom{N}{A}\int_1^{+\infty}\frac{(p-1)^{K_1}}{p^{K_2}}dp,$$

where $K_1 = N - A - 1$ and $K_2 = N + 1$. We recall the definition of Gamma function

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx = (z-1)!$$

where the first equality holds for any z complex number, while the second equality holds only if z is a positive integer. We also recall a particular way of defining the Beta function

$$B(x,y) = \int_0^{+\infty} \frac{t^{x-1}}{(1+t)^{x+y}}.$$

With the change of variable $t = p - 1$ we can write

$$\int_1^{+\infty} \frac{(p-1)^{K_1}}{p^{K_2}} dp = \int_0^{+\infty} \frac{t^{K_1}}{(1+t)^{K_2}} dt = B(K_1 + 1, K_2 - K_1 - 1).$$

Recalling that

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

we have

$$\int_1^{+\infty} \frac{(p-1)^{K_1}}{p^{K_2}} dp = \frac{\Gamma(K_1+1)\Gamma(K_2-K_1-1)}{\Gamma(K_2)}.$$

Thus we obtain

$$\int_0^{+\infty} f_{T_N(A)}(t) = 1.$$

When considering numerical results in Section 5.7, computation of the function $f_{T_N(A)}(t)$ from (5.26) is practically limited for numerical and computational reasons. In these cases, we can consider an approximation $\tilde{f}_{T_N(A)}(t)$ for $f_{T_N(A)}(t) = e^{\log(f_{T_N(A)}(t))}$ by using the approximation

$$\begin{aligned}\log(f_{T_N(A)}(t)) \simeq N\log\left(\frac{N}{N-A}\right) + A\log\left(\frac{N-A}{A}\right) + \log(N-A) + (N-A-1)\log\left(e^{\tilde{\mu}t} - 1\right) \\ - N\tilde{\mu}t + \log(\tilde{\mu}) + \frac{1}{2}\log\left(\frac{N}{2\pi A(N-A)}\right). \qquad (5.27)\end{aligned}$$

To prove that this approximation holds, we start recalling the Stirling's approximation for large values of n

$$n! \simeq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

so that

$$\log(n!) \simeq \frac{1}{2}\log(2\pi n) + n\log(n) - n$$

and, consequently,

$$\begin{aligned}\log\left(\binom{m}{n}\right) &\simeq m\log(m) - m + \frac{1}{2}\log(2\pi m) - n\log(n) + n - \frac{1}{2}\log(2\pi n) \\ &\quad - (m-n)\log(m-n) + (m-n) - \frac{1}{2}\log(2\pi(m-n)) \\ &= m\log(m) - n\log(n) - (m-n)\log(m-n) + \frac{1}{2}\log\left(\frac{m}{2\pi n(m-n)}\right).\end{aligned}$$

Thus we can write

$$\begin{aligned}\log(f_{T_N(A)}(t)) \simeq \log(N-A) + N\log(N) - A\log(A) - (N-A)\log(N-A) \\ + \frac{1}{2}\log\left(\frac{N}{2\pi A(N-A)}\right) + (N-A-1)\log\left(1-e^{-\tilde{\mu}t}\right) \\ + (A+1)\log(e^{-\tilde{\mu}t}) + \log(\tilde{\mu}).\end{aligned}$$

This can be written as

$$\begin{aligned}\log(f_{T_N(A)}(t)) \simeq N\log\left(\frac{N}{N-A}\right) + A\log\left(\frac{N-A}{A}\right) + \log(N-A) + \frac{1}{2}\log\left(\frac{N}{2\pi A(N-A)}\right) \\ + (N-A-1)\log\left(e^{\tilde{\mu}t}-1\right) - N\tilde{\mu}t + \log(\tilde{\mu}).\end{aligned}$$

This concludes the proof. See Figure 5.9 where $\tilde{f}_{T_N(A)}(t)$ is plotted for $N = 50$ and different values of A. See Figure 5.10 for the counterparts obtained by Gillespie simulations of the process.

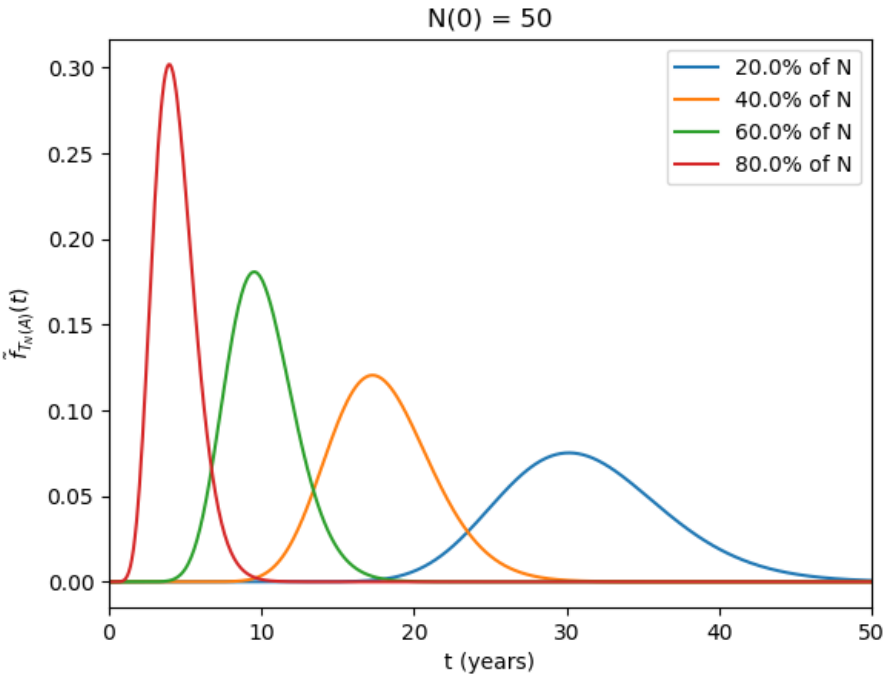

Figure 5.9: Plot of $\tilde{f}_{T_N(A)}(t)$ vs t, for process $\mathcal{X}_{(2)}$ an parameter values $\theta = 2.5$ years^{-1}, $\gamma = 1.25$ years^{-1}, $\mu = 0.5$ years^{-1}, $M = 200$, and $n_\theta = 4$. Different colours correspond to different values of A.

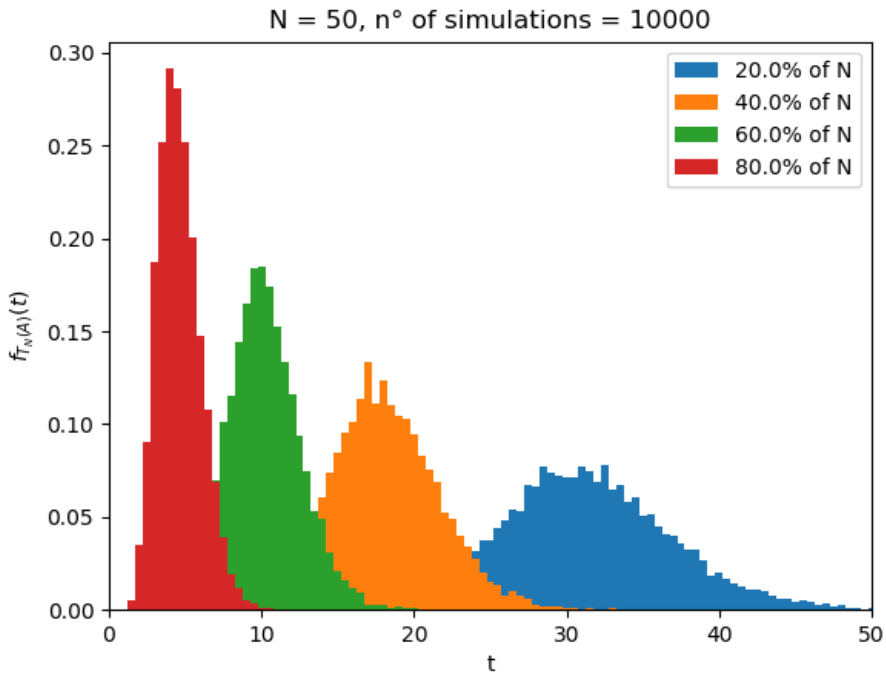

Figure 5.10: Approximations of $f_{T_N(A)}(t)$ obtained from 10^4 Gillespie simulations of process $\mathcal{X}_{(2)}$, and parameter values $\theta = 2.5$ years^{-1}, $\gamma = 1.25$ years^{-1}, $\mu = 0.5$ years^{-1}, $M = 200$ and $n_\theta = 4$. Different colours correspond to different values of A.

Using (5.26) we find the moment generating function of the random variable $T_N(A)$ as

$$M_{T_N(A)}(s) = \mathbb{E}\left[e^{sT_N(A)}\right] = \int_0^{+\infty} e^{ts} f_{T_N(A)}(t)dt$$
$$= (N-A)\binom{N}{A}\tilde{\mu}\int_0^{+\infty}(e^{\tilde{\mu}t}-1)^{N-A-1}e^{t(s-\tilde{\mu}N)}dt. \tag{5.28}$$

Defining the new variable $p = \exp\{\tilde{\mu}t\}$, we have $\mathrm{d}p = \tilde{\mu}\exp\{\tilde{\mu}t\}\ \mathrm{d}t$ and $t = \log(p)/\tilde{\mu}$. We can therefore rewrite (5.28) as

$$M_{T_N(A)}(s) = (N-A)\binom{N}{A}\int_1^{+\infty} p^{k_1}(p-1)^{k_2}\ \mathrm{d}p, \tag{5.29}$$

with

$$k_1 = \frac{s-\tilde{\mu}(N+1)}{\tilde{\mu}},$$
$$k_2 = N-A-1.$$

To find an explicit solution for (5.29), we need to find a solution for the integra

$$\int_1^{+\infty} p^{k_1}(p-1)^{k_2}\ \mathrm{d}p. \tag{5.30}$$

To this aim, we apply integration by parts with

$$\mathrm{d}u = p^{k_1}\ \mathrm{d}p \ \Rightarrow\ u = \frac{p^{k_1+1}}{k_1+1},$$
$$v = (p-1)^{k_2} \ \Rightarrow\ \mathrm{d}v = k_2(p-1)^{k_2-1}\ \mathrm{d}p.$$

Thus we obtain the recursion

$$\int_1^{+\infty} p^{k_1}(p-1)^{k_2}\ \mathrm{d}p = \left[\frac{p^{k_1+1}(p-1)^{k_2}}{(k_1+1)}\right]_1^{+\infty} - \frac{k_2}{(k_1+1)}\int_1^{+\infty} p^{k_1+1}(p-1)^{k_2-1}\ \mathrm{d}p. \tag{5.31}$$

We can now apply (5.31) recursively to its own right side of the equation. We apply it (k_2-1) times until we reach, as part of the right hand side, the integral

$$\int_1^{+\infty} p^{k_1+k_2}(p-1)^0\ \mathrm{d}p,$$

which can be computed as $\left[\frac{p^{k_1+k_2+1}}{k_1+k_2+1}\right]_1^{+\infty}$. Therefore we obtain the solution

$$\int_1^{+\infty} p_1^k(p-1)^{k_2}\ \mathrm{d}p = \left[\frac{p^{k_1+1}(p-1)^{k_2}}{(k_1+1)}\right]_1^{+\infty} + \sum_{k=2}^{k_2+1}(-1)^{k-1}\left[\frac{p^{k_1+k}(p-1)^{k_2-(k-1)}\prod_{i=0}^{k-2}(k_2-i)}{\prod_{i=1}^{k}(k_1+i)}\right]_1^{+\infty}. \tag{5.32}$$

We can therefore apply (5.32) to (5.29) finding

$$M_{T_N(A)}(s) = (N-A)\binom{N}{A}\left[\frac{p^{k_1+1}(p-1)^{k_2}}{(k_1+1)} + \sum_{k=2}^{k_2+1}(-1)^{k-1}\frac{p^{k_1+k}(p-1)^{k_2-(k-1)}\prod_{i=0}^{k-2}(k_2-i)}{\prod_{i=1}^{k}(k_1+i)}\right]_1^{+\infty}. \tag{5.33}$$

We note that the highest power of p in (5.33) is $k_1 + k_2 + 1$, which gives the condition for the existence of $M_{T_N(A)}(s)$, that is

$$k_1 + k_2 + 1 < 0 \;\Rightarrow\; s < \tilde{\mu}(A+1).$$

We also note that, due to this condition, all the elements in (5.33) evaluated in $p = +\infty$ tend to 0. Evaluating in $p = 1$, everything results in being 0 because of the elements $(p-1)^{k_2-k+1}$, except for $k = k_2 + 1$. Thus we can write

$$\begin{aligned} M_{T_N(A)}(s) &= (N-A)\binom{N}{A}(-1)^{k_2+1}\frac{\prod_{i=0}^{k_2-1}(k_2-i)}{\prod_{i=1}^{k_2+1}(k_1+i)} \\ &= (N-A)\binom{N}{A}(-1)^{k_2+1}B(k_1+1,k_2+1), \end{aligned} \tag{5.34}$$

where

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

represents the Beta function. It is easy to prove that $M_{T_N(A)}(0) = 1$ as should be for a moment generating function. Higher moments for the random variable $T_N(A)$ can be computed as

$$\mathbb{E}\left[T_N(A)^k\right] = \frac{d^k}{ds^k}M_{T_N(A)}(s)\bigg|_{s=0}.$$

However, computation of these moments can also be carried out if we note that $T_N(A)$ is the time to absorption of a pure-death process formed by a sequence of independent exponentially distributed times, corresponding to death events. Define $T_{i,j}$ the time for the process to go from state i to state j. We can then write $T_N(A) = T_{N,N-1} + T_{N-1,N-2} + \cdots + T_{A+1,A}$, so that $\mathbb{E}\left[T_N(A)\right] = \mathbb{E}(T_{N,N-1}) + \mathbb{E}(T_{N-1,N-2}) + \cdots + \mathbb{E}(T_{A+1,A})$. As previously said, we know that $T_{i,i-1}$ is distributed as an exponential random variable with parameter $i\tilde{\mu}$. We recall that, if Z is an exponential random variable with parameter λ, its expected value is λ^{-1}. This implies (see Figure 5.11)

$$\mathbb{E}(T_N(A)) = \frac{1}{\tilde{\mu}}\sum_{i=A+1}^{N}\frac{1}{i} = \frac{1}{\tilde{\mu}}\left(\frac{1}{A+1} + \frac{1}{A+2} + \cdots + \frac{1}{N}\right). \tag{5.35}$$

As shown in Section (6.7) of [5], we can write

$$\mathbb{E}(T_N(A)) \approx \frac{1}{\tilde{\mu}}\log\left(\frac{N}{A}\right).$$

In a similar way, recalling that if Z is an exponential random variable with parameter λ, then its variance is λ^{-2}, we can write

$$\mathrm{Var}(T_N(A)) = \frac{1}{\tilde{\mu}^2}\sum_{i=A+1}^{N}\frac{1}{i^2} = \frac{1}{\tilde{\mu}^2}\left(\frac{1}{(A+1)^2} + \frac{1}{(A+2)^2} + \cdots + \frac{1}{N^2}\right). \tag{5.36}$$

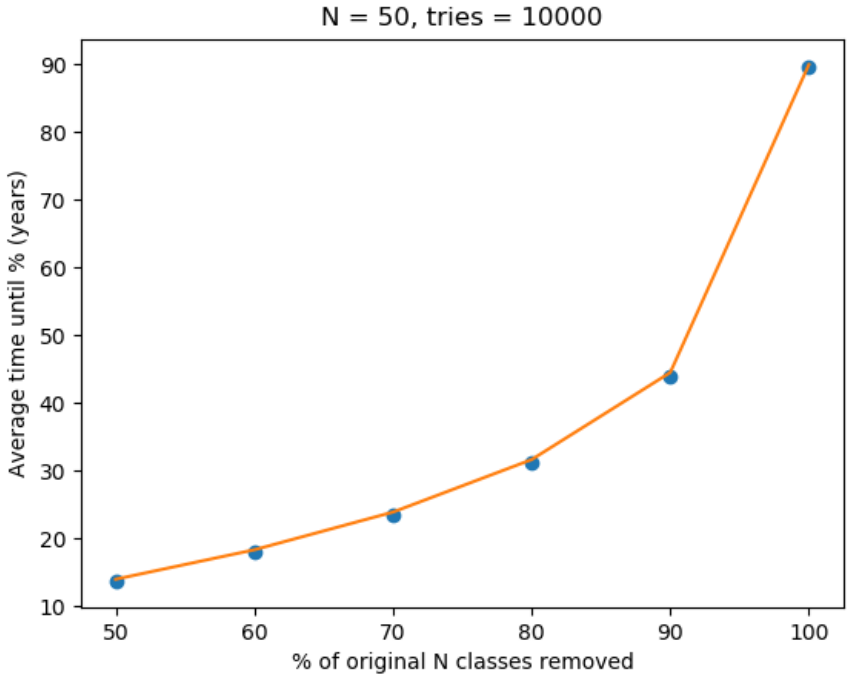

Figure 5.11: Plot of (5.35) (orange) and simulations of death process (blue). Time until absorption (years) as a function of the % of original clones removed from the repertoire. Parameters: $\theta = 2.5$ years^{-1}, $\gamma = 1.25$ years^{-1}, $\mu = 0.5$ years^{-1}, $M = 200$, $n_\theta = 4$ and $N = 50$.

Let us focus now on the general higher moment $\mathbb{E}\left[T_N(A)^k\right]$. We want now to prove that

$$\mathbb{E}\left[T_N(A)^k\right] = \frac{k!}{\tilde{\mu}^k} \sum_{k_1+k_2+\ldots+k_{N-A}=k} \left[\frac{1}{N^{k_1}(N-1)^{k_2}\cdots(A+1)^{k_{N-A}}}\right],$$

where the sum is taken over all possible combinations of non-negative integer indices $k_1, k_2, \ldots, k_{N-A}$ such that $\sum_{j=1}^{N-A} k_j = k$. Let us consider the process $\mathcal{Y}_{(1)}$. We recall that the random variable $T_N(A)$, representing the time to reach for the first time a number $A < X(0)$ of original clonotypes in the repertoire, can be written as $T_N(A) = T_{N,N-1} + T_{N-1,N-2} + \cdots + T_{A+1,A}$, where $T_{i,j}$ amounts to the time spent by the process $\mathcal{Y}_{(1)}$ to go from state i to state j. We are interested in computing the higher moments $\mathbb{E}\left[T_N(A)^k\right]$ of the random variable $T_N(A)$. We write $\mathbb{E}\left[T_N(A)^k\right] = \mathbb{E}\left[\left(T_{N,N-1} + T_{N-1,N-2} + \cdots + T_{A+1,A}\right)^k\right]$. In order to find a formula for the higher moments of $T_N(A)$, we first state an important theorem of combinatorics: the *multinomial theorem*. For any positive integer $m > 0$ and any integer $n \geq 0$, and considering m terms $x_1, x_2, \ldots, x_m$, we can write

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{k_1+k_2+\ldots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{j=1}^{m} x_j^{k_j},$$

where the sum is taken over all possible combinations of integer indices $k_1, k_2, \ldots, k_m \geq 0$ such that $\sum_{i=1}^{m} k_i = k$ and

$$\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}$$

is called multinomial coefficient, so that the case $m = 2$ represents the famous binomial theorem. We can therefore write

$$\mathbb{E}\left[T_N(A)^k\right] = \mathbb{E}\left[\sum_{k_1+k_2+\ldots+k_{N-A}=k} \binom{k}{k_1, k_2, \ldots, k_{N-A}} T_{N,N-1}^{k_1} T_{N-1,N-2}^{k_2} \cdots T_{A+1,A}^{k_{N-A}}\right],$$

and linearity properties of the expected value can transform this last equality as follows

$$\mathbb{E}\left[T_N(A)^k\right] = \sum_{k_1+k_2+\ldots+k_{N-A}=k} \left[\binom{k}{k_1, k_2, \ldots, k_{N-A}} \mathbb{E}\left[T_{N,N-1}^{k_1} T_{N-1,N-2}^{k_2} \cdots T_{A+1,A}^{k_{N-A}}\right]\right].$$

We recall that $T_{i,i-1}$ are independent exponentially distributed random variables with parameter $i\tilde{\mu}$, so that $\mathbb{E}\left[T_{i,i-1}^k\right] = \frac{k!}{(i\tilde{\mu})^k}$ and

$$\mathbb{E}\left[T_N(A)^k\right] = \sum_{k_1+k_2+\ldots+k_{N-A}=k} \left[\binom{k}{k_1, k_2, \ldots, k_{N-A}} \frac{k_1!}{(N\tilde{\mu})^{k_1}} \frac{k_2!}{((N-1)\tilde{\mu})^{k_2}} \cdots \frac{k_{N-A}!}{((A+1)\tilde{\mu})^{k_{N-A}}}\right].$$

It follows that

$$\mathbb{E}\left[T_N(A)^k\right] = \frac{k!}{\tilde{\mu}^k} \sum_{k_1+k_2+\ldots+k_{N-A}=k} \left[\frac{1}{N^{k_1}(N-1)^{k_2} \cdots (A+1)^{k_{N-A}}}\right].$$

5.4.2 Explicit competition

We consider here $\mu_{n,m}^{(X)} = \mu_{n,m}^{(X,2)} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu_{n,m}^{(Y)} = \mu_{n,m}^{(Y,2)} = m(\beta_1 + \beta_2 pn)$, so that $\mu_{n,m}^{(X,2)} + \mu_{n,m}^{(Y,2)} = \mu_n^{(2)}$, leading to the process $\mathcal{X}_{(2)}^{aug}$; see Figure 5.12.

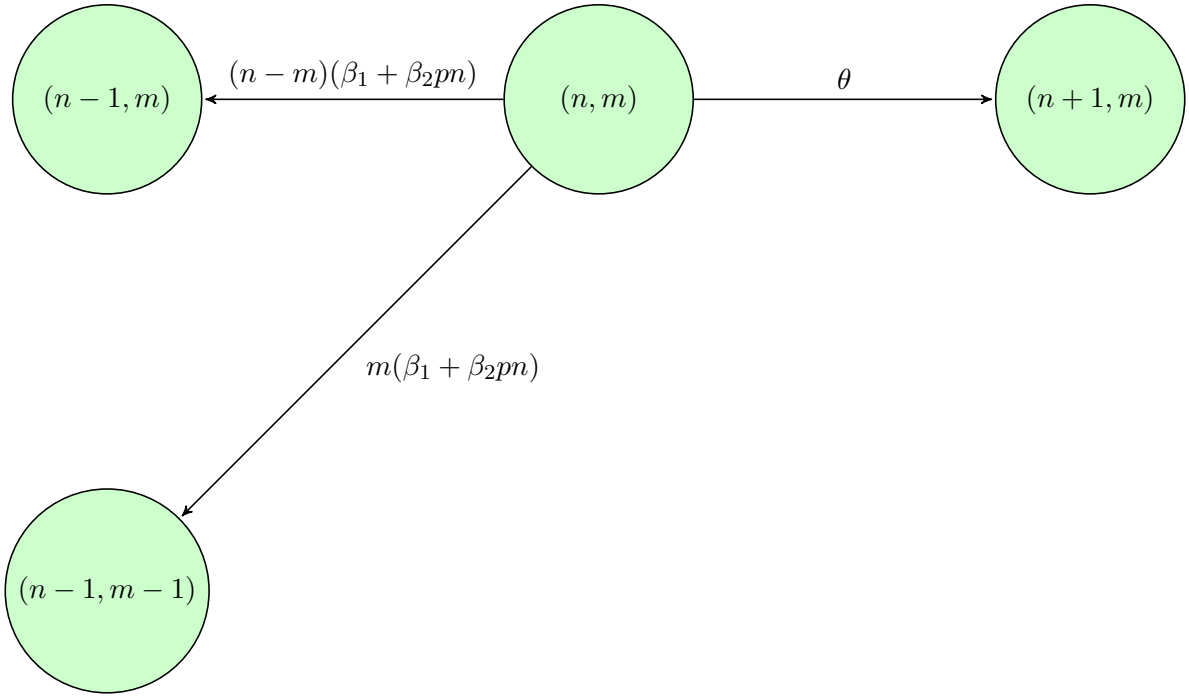

Figure 5.12: Bivariate continuous-time birth-and-death process $\mathcal{X}_{(2)}^{aug}$ with $\mu_{n,m}^{(X)} = \mu_{n,m}^{(X,2)} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu_{n,m}^{(Y)} = \mu_{n,m}^{(Y,2)} = m(\beta_1 + \beta_2 pn)$.

Unlike in the previous section, the process $\mathcal{Y}_{(2)} = \{Y(t) : t \geq 0\}$ cannot be considered here as a pure-death linear process, since the death rate $\mu_{n,m}^{(Y,2)}$ depends on both n and m. Therefore, in order to find the expected value $\mathbb{E}(T_N(A))$ of the time $T_N(A)$, the same arguments cannot be applied and a different method has to be followed. We start by considering the process $\mathcal{X}_{(2)}^{aug}$ defined over the space of states $\mathbb{S} = \{(n,m) : m \in \{0,1,\ldots,X(0)\}, n \geq m\}$ and with initial conditions $(X(0),Y(0)) = u$. Once defined u, and defining $\mathbb{A} \subset \mathbb{S}$ as a set of states reachable from initial state u, we define $T_u(\mathbb{A})$ as the time at which the process reaches $\mathbb{A}$. A system of equations for the expression of the expected value $\tau_u = \mathbb{E}(T_u)$ is shown. In order to give the general result for the expected value τ_u, we define $q_{u',u''}$ as the transition rate from state u' to state u''. We use the notation $u \to u'$ to define the event describing the first step of the process, from initial state u to the second state u'. Thus,

$$\tau_u = \sum_{u' \in \mathbb{S}} \left[\mathbb{E}(T_u | u \to u') \cdot \Pr(u \to u') \right], \tag{5.37}$$

where (see Eq. (5.12) of [5])

$$\Pr(u \to u') = \frac{q_{u,u'}}{\sum_{u'' \in \mathbb{S}} q_{u,u''}}.$$

In order to fully understand (5.37) we need to focus on its first factors $\mathbb{E}(T_u | u \to u')$. The random variable $T_u | u \to u'$ can be written as

$$T_u | u \to u' = (T_{u'} + t_{u \to u'}) | u \to u' = (T_{u'}) + (t_{u \to u'} | u \to u'),$$

where the random variable $t_{u \to u'}$ represents the time for the process to go from state u to state u'. The last equality stands because of the independence of the random variable $T_{u'}$ from the past event $\{u \to u'\}$, that is because of the Markov property of the process. Being $\mathcal{X}^{aug}_{(2)}$ a continuous-time Markov chain, the random variable $t_{u \to u'}$ takes non-negative real values and has an exponential distribution with parameter $q_{u,u'}$, that is

$$t_{u \to u'} \sim \text{Exp}(q_{u,u'}).$$

We need though to focus on the random variable $\Psi = t_{u \to u'} | u \to u'$, which is representing the random variable $t_{u \to u'}$ knowing that the first movement of the process is $u \to u'$. This can be read as $\Psi = \min_{u'' \in \mathbb{S}}(t_{u \to u''})$, where the minimum is taken over all the possible u'' first movements from the initial state u. Thus, from the properties of exponential random variables, it follows

$$\Psi \sim \quad \text{Exp}\left(\sum_{u''} q_{u,u''}\right).$$

We can now write

$$\mathbb{E}(T_u | u \to u') = \mathbb{E}(T_{u'} + \Psi) = \tau_{u'} + \frac{1}{\sum_{u'' \in S} q_{u,u''}},$$

leading us to

$$\tau_u = \mathbb{E}(T_u) = \sum_{u' \in S} \left[\left(\tau_{u'} + \frac{1}{\sum_{u'' \in S} q_{u,u''}} \right) \cdot \frac{q_{u,u'}}{\sum_{u'' \in S} q_{u,u''}} \right] \tag{5.38}$$

Eq. (5.38) can now be applied to the process $\mathcal{X}^{aug}_{(2)}$ with $\mu^{(X)}_{n,m} = \mu^{(X,2)}_{n,m} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,2)}_{n,m} = m(\beta_1 + \beta_2 pn)$. We consider $u = (n, m)$ as the starting point of the process and we define the set of states $\mathbb{A} = \{(n, m) : m \equiv A\} \subset \mathbb{S}$ that the augmented process $\mathcal{X}^{aug}_{(2)}$ has to reach. There are three possible u' states reachable from u with one step: $(n-1, m)$, $(n+1, m)$ and $(n-1, m-1)$. The three transition rates are, respectively, $(\beta_1 + \beta_2 pn)(n - m)$, θ and $m(\beta_1 + \beta_2 pn)$, giving

$$\sum_{u'' \in \mathbb{S}} q_{u,u''} = n(\beta_1 + \beta_2 pn) + \theta.$$

Thus, applying (5.38) we obtain

$$\tau_{(n,m)} = \frac{\theta \cdot \tau_{(n+1,m)} + (\beta_1 + \beta_2 pn)(n-m) \cdot \tau_{(n-1,m)} + m(\beta_1 + \beta_2 pn) \cdot \tau_{(n-1,m-1)} + 1}{n(\beta_1 + \beta_2 pn) + \theta}. \tag{5.39}$$

To find explicit solutions for the recursive equation (5.39), we start focusing on the particular states $\{(n,m) : m = A+1\}$. In order to solve the (infinite) system of equations given by (5.39), we consider a maximum number of total clonotype classes M that the random variable $X(t)$ cannot exceed, so that (5.39) becomes a finite system of equations represented by the equality

$$\begin{aligned}\tau_{(n,A+1)} = &\frac{\delta_{n<M} \cdot \theta \cdot \tau_{(n+1,A+1)} + (\beta_1 + \beta_2 pn)(n-(A+1)) \cdot \tau_{(n-1,m)}}{n(\beta_1 + \beta_2 pn) + \delta_{n<M} \cdot \theta} \\ &+ \frac{(A+1)(\beta_1 + \beta_2 pn) \cdot \tau_{(n-1,A)} + 1}{n(\beta_1 + \beta_2 pn) + \delta_{n<M} \cdot \theta},\end{aligned} \tag{5.40}$$

where δ_D is defined as

$$\delta_D = \begin{cases} 1 & \text{if } D \text{ is satisfied,} \\ 0 & \text{if not.} \end{cases}$$

We define the functions

$$\begin{aligned} v^{(2)}(n) &= n(\beta_1 + n\beta_2 p) + \delta_{n<M} \cdot \theta \\ g^{(2)}(n,m) &= (\beta_1 + n\beta_2 p)(n-m). \end{aligned}$$

Note that the notation $v^{(2)}(n)$ and $g^{(2)}(n,m)$ is due to the usage of the second kind of death rates $\mu^{(X)}_{n,m} = \mu^{(X,2)}_{n,m} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,2)}_{n,m} = m(\beta_1 + \beta_2 pn)$ in this section. For the same reason, in the following sections we will also use the notation $v^{(1)}(n)$ and $g^{(1)}(n,m)$ when dealing with the cases $\mu^{(X)}_{n,m} = \mu^{(X,1)}_{n,m} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,1)}_{n,m} = m(\beta_1 + \beta_2 pn)$. Definitions $v^{(2)}(n)$ and $g^{(2)}(n,m)$, together with (5.40), give

$$\begin{aligned}\tau_{(M,A+1)} &= \frac{1}{M(\beta_1 + M\beta_2 p)} + \frac{(\beta_1 + M\beta_2 p)(M-(A+1))}{M(\beta_1 + M\beta_2 p)} \cdot \tau_{(M-1,A+1)} \\ &= \frac{1}{v^{(2)}(M)} + \frac{g^{(2)}(M, A+1)}{v^{(2)}(M)} \cdot \tau_{(M-1,A+1)}.\end{aligned} \tag{5.41}$$

Thus, $\tau_{(M,A+1)}$ can be written as

$$\tau_{(M,A+1)} = \frac{a_M}{b_M} + \frac{c_M}{b_M} \cdot \tau_{(M-1,A+1)},$$

where $a_M = 1$; $b_M = v^{(2)}(M)$; $c_M = g^{(2)}(M, A+1)$. Let us focus now on $\tau_{(M-1,A+1)}$. Following similar arguments as for $\tau_{(M,A+1)}$, we have

$$\begin{aligned}\tau_{(M-1,A+1)} &= \frac{\theta}{v^{(2)}(M-1)} \cdot \tau_{(M,A+1)} + \frac{g^{(2)}(M-1, A+1)}{v^{(2)}(M-1)} \cdot \tau_{(M-2,A+1)} + \frac{1}{v^{(2)}(M-1)} \\ &= \frac{\theta + v^{(2)}(M)}{v^{(2)}(M-1)v^{(2)}(M)} + \frac{\theta g^{(2)}(M,y) \cdot \tau_{(M-1,A+1)}}{v^{(2)}(M-1)v^{(2)}(M)} + \frac{g^{(2)}(M-1, A+1) \cdot \tau_{(M-2,A+1)}}{v^{(2)}(M-1)},\end{aligned} \tag{5.42}$$

where the last equality is obtained by replacing (5.41) in the first equality. We can now rearrange (5.42) to obtain

$$\tau_{(M-1,A+1)} = \frac{\theta + v^{(2)}(M)}{v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)} + \frac{v^{(2)}(M)g^{(2)}(M-1, A+1) \cdot \tau_{(M-2,A+1)}}{v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)}, \tag{5.43}$$

so that

$$\tau_{(M-1,A+1)} = \frac{a_{M-1}}{b_{M-1}} + \frac{c_{M-1}}{b_{M-1}} \cdot \tau_{(M-2,A+1)},$$

where $a_{M-1} = \theta + v^{(2)}(M)$; $b_{M-1} = v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)$; $c_{M-1} = v^{(2)}(M)g^{(2)}(M-1, A+1)$. Let us focus now on $\tau_{(M-2,A+1)}$. We have

$$\begin{aligned}
\tau_{(M-2,A+1)} &= \frac{\theta \cdot \tau_{(M-1,A+1)}}{v^{(2)}(M-2)} + \frac{g^{(2)}(M-2, A+1) \cdot \tau_{(M-3,A+1)}}{v^{(2)}(M-2)} + \frac{1}{v^{(2)}(M-2)} \\
&= \frac{\theta[\theta + v^{(2)}(M)]}{v^{(2)}(M-2)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)]} \\
&\quad + \frac{\theta v^{(2)}(M)g^{(2)}(M-1, A+1)}{v^{(2)}(M-2)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)]} \cdot \tau_{(M-2,A+1)} \\
&\quad + \frac{g^{(2)}(M-2, A+1)}{v^{(2)}(M-2)} \cdot \tau_{(M-3,A+1)} + \frac{1}{v^{(2)}(M-2)},
\end{aligned} \tag{5.44}$$

where the last equality is obtained by replacing (5.43) in the first equality. We can now rearrange (5.44) to obtain

$$\begin{aligned}
\tau_{(M-2,A+1)} = &\frac{\theta[\theta + v^{(2)}(M)] + [v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)]}{v^{(2)}(M-2)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)] - \theta v^{(2)}(M)g^{(2)}(M-1, A+1)} \\
&+ \frac{g^{(2)}(M-2, A+1)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)] \cdot \tau_{(M-3,A+1)}}{v^{(2)}(M-2)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)] - \theta v^{(2)}(M)g^{(2)}(M-1, A+1)}.
\end{aligned}$$

so that

$$\tau_{(M-2,A+1)} = \frac{a_{M-2}}{b_{M-2}} + \frac{c_{M-2}}{b_{M-2}} \cdot \tau_{(M-3,A+1)},$$

where

$$\begin{aligned}
a_{M-2} &= \theta[\theta + v^{(2)}(M)] + [v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)], \\
b_{M-2} &= v^{(2)}(M-2)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)] - \theta v^{(2)}(M)g^{(2)}(M-1, A+1), \\
c_{M-2} &= g^{(2)}(M-2, A+1)[v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M, A+1)].
\end{aligned}$$

Noticing the two first sets of relations

$$\begin{aligned}
a_{M-1} &= b_M + \theta a_M, \\
b_{M-1} &= b_M v^{(2)}(M-1) - \theta c_M, \\
c_{M-1} &= b_M g^{(2)}(M-1, A+1),
\end{aligned}$$

and

$$\begin{aligned} a_{M-2} &= b_{M-1} + \theta a_{M-1}, \\ b_{M-2} &= b_{M-1} v^{(2)}(M-2) - \theta c_{M-1}, \\ c_{M-2} &= b_{M-1} g^{(2)}(M-2, A+1), \end{aligned}$$

we can write the general recursive relations

$$\begin{aligned} a_{M-k} &= b_{M-k+1} + \theta a_{M-k+1}, \\ b_{M-k} &= b_{M-k+1} v^{(2)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(2)}(M-k, A+1), \end{aligned} \tag{5.45}$$

with initial values $a_M = 1$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M, A+1)$ and

$$\tau_{(M-k,A+1)} = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot \tau_{(M-k-1,A+1)}. \tag{5.46}$$

In order to obtain explicit solution for the general values $\tau_{(n,A+1)}$, an algorithm has to be followed. Before showing the different steps of the algorithm, it is worth noticing that (5.39) and (5.46) give, respectively, the two equations of the following system

$$\begin{cases} \tau_{(A+1,A+1)} = \dfrac{\theta}{v^{(2)}(A+1)} \tau_{(A+2,A+1)} + \dfrac{1}{v^{(2)}(A+1)}, \\ \tau_{(A+2,A+1)} = \dfrac{a_{A+2}}{b_{A+2}} + \dfrac{c_{A+2}}{b_{A+2}} \tau_{(A+1,A+1)}, \end{cases}$$

which gives

$$\begin{cases} \tau_{(A+1,A+1)} = \dfrac{\theta a_{A+2} + b_{A+2}}{b_{A+2} v^{(2)}(A+1) - \theta c_{A+2}}, \\ \tau_{(A+2,A+1)} = \dfrac{a_{A+2} v^{(2)}(A+1) + c_{A+2}}{b_{A+2} v^{(2)}(A+1) - \theta c_{A+2}}. \end{cases}$$

Given these values $\tau_{(A+1,A+1)}$ and $\tau_{(A+2,A+1)}$ dependent on a_{A+2}, b_{A+2} and c_{A+2}, we can now give the steps of the algorithm:

- Start with $a_M = 1$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M, A+1)$;
- Use the recursive relations (5.45) to find a_{A+2}, b_{A+2} and c_{A+2};
- Find $\tau_{(A+1,A+1)}$ and $\tau_{(A+2,A+1)}$ as explained above;
- Use (5.46) to find all the different values of $\tau_{(n,A+1)}$, up to $\tau_{(M,A+1)}$.

We focus now on the more general states $\{(n,m) : m \in \{A+2, A+3, \cdots, N\}\}$. We recall the functions $v^{(2)}(n)$ and $g^{(2)}(n,m)$ and we define the new function $q(n,m)$ as

$$\begin{aligned} v^{(2)}(n) &= n(\beta_1 + n\beta_2 p) + \delta_{n<M} \cdot \theta \\ g^{(2)}(n,m) &= (\beta_1 + n\beta_2 p)(n-m) \\ q(n,m) &= 1 + m(\beta_1 + n\beta_2 p) \cdot \tau_{(n-1,m-1)}. \end{aligned}$$

Note that for the function $q(n,m)$ we do not use any index as this function will be used only in this case for this particular section. These definitions, together with (5.39), give

$$\tau_{(M,m)} = \frac{q(M,m)}{v^{(2)}(M)} + \frac{g^{(2)}(M,m)}{v^{(2)}(M)} \cdot \tau_{(M-1,m)}. \tag{5.47}$$

With the same techniques used for computing (5.43), we can find

$$\tau_{(M-1,m)} = \frac{\theta q(M,m) + v^{(2)}(M) q(M-1,m)}{v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M,m)} + \frac{v^{(2)}(M) g^{(2)}(M-1,m)}{v^{(2)}(M-1)v^{(2)}(M) - \theta g^{(2)}(M,m)} \cdot \tau_{(M-2,m)},$$

obtaining the relationships

$$\begin{aligned} a_{M-1} &= b_M q(M-1,m) + \theta a_M, \\ b_{M-1} &= b_M v^{(2)}(M-1) - \theta c_M, \\ c_{M-1} &= b_M g^{(2)}(M-1,m), \end{aligned}$$

and, recursively, the general ones

$$\begin{aligned} a_{M-k} &= b_{M-k+1} q(M-k,m) + \theta a_{M-k+1}, \\ b_{M-k} &= b_{M-k+1} v^{(2)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(2)}(M-k,m), \end{aligned} \tag{5.48}$$

with $a_M = q(M,m)$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M,m)$ and

$$\tau_{(M-k,m)} = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot \tau_{(M-k-1,m)}. \tag{5.49}$$

In order to obtain explicit solution for the general $\tau_{(n,m)}$, similar steps have to be followed as for the particular case $m = A+1$. Thus, we notice that (5.39) and (5.49) give, respectively, the two equations of the following system

$$\begin{cases} \tau_{(m,m)} = \dfrac{\theta}{v^{(2)}(m)} \tau_{(m+1,m)} + \dfrac{q(m,m)}{v^{(2)}(m)}, \\ \tau_{(m+1,m)} = \dfrac{a_{m+1}}{b_{m+1}} + \dfrac{c_{m+1}}{b_{m+1}} \tau_{(m,m)}, \end{cases}$$

which gives

$$\begin{cases} \tau_{(m,m)} = \dfrac{\theta a_{m+1} + b_{m+1} q(m,m)}{b_{m+1} v^{(2)}(m) - \theta c_{m+1}}, \\ \\ \tau_{(m+1,m)} = \dfrac{a_{m+1} v^{(2)}(m) + c_{m+1} q(m,m)}{b_{m+1} v^{(2)}(m) - \theta c_{m+1}}. \end{cases}$$

Given these two values $\tau_{(m,m)}$ and $\tau_{(m+1,m)}$ dependent on a_{m+1}, b_{m+1} and c_{m+1}, the steps of the algorithm are as follows:

- Start with $a_M = q(M,m)$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M,m)$;
- Use the recursive relations (5.48) to find a_{m+1}, b_{m+1} and c_{m+1};
- Find $\tau_{(m,m)}$ and $\tau_{(m+1,m)}$ as explained above;
- Use (5.49) to find all the different values of $\tau_{(n,m)}$, up to $\tau_{(M,m)}$.

The reader can find simulations for this algorithm and for a Gillespie code representing the same biological process in Figures 5.38 and 5.39 respectively. The hitting times are also plotted for a specific initial state as function of both the β_1 and $\beta_2 p$ variables, as a heat map in Figure 5.40.

5.5 Size of the repertoire at time $T_N(A)$

We recall the definition of our augmented process $\mathcal{X}^{\text{aug}} = \{(X(t), Y(t)) : t \geq 0\}$ defined on the space of states $\mathbb{S}^{aug} = \{(n,m) : m \in \{0,1,2,\ldots,X(0)\},\ n \geq m\}$. We also recall the definition of $T_N(A)$ as the time when, for the first time, the process $\mathcal{X}^{\text{aug}}$ reaches the space of states $\mathbb{A} = \{(n,m) : m = A\}$, representing thus the first time until only $A < X(0)$ of the original clonotypes remain in the repertoire. The aim of this section is to analyse the probability $p_{(n,m)}(\bar{n}, A) = \Pr(X(T_N(A)) = \bar{n} | (X(0), Y(0)) = (n,m))$ that the size $X(t)$ of the repertoire at time $T_N(A)$ equals a particular value $\bar{n} \geq A$. As for the previous section, we consider a maximum number of total clonotype classes M that the random variable $X(t)$ cannot exceed. We believe this stochastic descriptor represents a significant aspect of the internal dynamics of a repertoire, as it has the capability to describe the probability distribution of the size of the renewed part of the repertoire at time $T_N(A)$. The last section focused on the use of a first step argument to find an equation for the stochastic descriptor $T_N(A)$. In the following sections a similar argument will be applied to the study of the probabilities $p_{(n,m)}(\bar{n}, A)$, known as *hitting probabilities*.

5.5.1 Implicit competition

Analytical analysis

We consider the auxiliary random variable $Z(t) = X(t) - Y(t)$ and the process $\mathcal{Z} = \{Z(t) : t \geq 0\}$ defined on $\mathbb{S}_z = \{0, 1, 2, \ldots\}$, where $Z(t)$ amounts to the number of newly generated clonotypes in the repertoire at time $t \geq 0$. We note that $\mathcal{Z}$ can be seen as a pure-death linear process with immigration, with death rate $\mu_z = \tilde{\mu} z$ and immigration rate θ. We want to analyse the probability, for the process $\mathcal{X}^{\text{aug}}$, to be in the state $(z + A, A)$ at time $T_N(A)$. This is equivalent to analyse the probability that the process $\mathcal{Z}$ is in state z at time $T_N(A)$. Thus we want to analyse the following probabilities

$$\Pr\left(Z(T_N(A)) = z\right) = \Pr\left(X(T_N(A)) = z + A\right), \quad z = 0, 1, 2, ...$$

Given the independence between process $\mathcal{Y} = \{Y(t) : t \geq 0\}$ defined on $\{0, 1, 2, \ldots, X(0)\}$ and process $\mathcal{Z}$, and given the probability density function $f_{T_N(A)}(t)$ of the random variable $T_N(A)$ given by (5.26), we can write

$$\Pr\left(Z(T_N(A)) = z\right) = \int_0^{+\infty} p_z(t) \cdot f_{T_N(A)}(t)\, dt, \tag{5.50}$$

where $p_z(t) = \Pr(Z(t) = z)$. Therefore, we need to find $p_z(t)$. We start writing the forward Kolmogorov equations for process $\mathcal{Z}$ as

$$\begin{cases} \dfrac{dp_z(t)}{dt} = \theta p_{z-1}(t) + \tilde{\mu}(z+1)p_{z+1}(t) - (\theta + \tilde{\mu} z)p_z(t), \quad z = 1, 2, \ldots, \\ \dfrac{dp_0(t)}{dt} = \tilde{\mu} p_1(t) - \theta p_0(t). \end{cases}$$

From these equations, we find now the partial differential equation for the probability generating function $\phi_Z(s, t)$ of the random variable $Z(t)$, following the arguments of Chapter 6 in [5]. We multiply the differential equations by s^z and sum over z. Then

$$\begin{aligned} \frac{\partial \phi_Z(s,t)}{\partial t} &= \theta \sum_{z=2}^{+\infty} p_{z-1}(t) s^z + \tilde{\mu} \sum_{z=0}^{+\infty} (z+1) p_{z+1}(t) s^z - \theta \sum_{z=1}^{+\infty} p_z(t) s^z - \tilde{\mu} \sum_{z=1}^{+\infty} z p_z(t) s^z \\ &= \theta \sum_{z=1}^{+\infty} p_z(t) s^{z+1} + \tilde{\mu} \sum_{z=1}^{+\infty} z p_z(t) s^{z-1} - \theta \sum_{z=1}^{+\infty} p_z(t) s^z - \tilde{\mu} \sum_{z=1}^{+\infty} z p_z(t) s^z \\ &= \theta s \sum_{z=1}^{+\infty} p_z(t) s^z + \tilde{\mu} \sum_{z=1}^{+\infty} z p_z(t) s^{z-1} - \theta \sum_{z=1}^{+\infty} p_z(t) s^z - \tilde{\mu} s \sum_{z=1}^{+\infty} z p_z(t) s^{z-1}. \end{aligned}$$

Thus we have

$$\begin{cases} \dfrac{\partial \phi_Z(s,t)}{\partial t} = \tilde{\mu}(1-s)\dfrac{\partial \phi_Z(s,t)}{\partial s} + \theta(s-1)\phi_Z(s,t), \\ \phi_Z(s,0) = s^{Z(0)} = 1. \end{cases} \tag{5.51}$$

We notice that this equation agrees with arguments in Section (6.4.4) of [5]. In fact, it would be sufficient to consider the partial differential equation for the moment generating

function for the special case $\lambda = 0$ and using the substitution $\theta = \log(s)$ in order to find our equation (Note that θ in the equation from [5] is not representing our immigration rate θ, but just the independent variable of the moment generating function). We use the method of characteristics to find a solution $\phi_Z(s,t)$ for the general initial condition $Z(0) = Z_0$. We will then apply it to our specific case where $Z_0 = 0$. We start by rewriting the system as follows

$$\begin{cases} \tilde{\mu}(s-1)\dfrac{\partial \phi_Z(s,t)}{\partial s} + \dfrac{\partial \phi_Z(s,t)}{\partial t} + \theta(1-s)\phi_Z(s,t) = 0, \\ \phi_Z(s,0) = s^{Z_0}. \end{cases}$$

We write the ODEs system

$$\begin{cases} ds/dw = \tilde{\mu}(s-1), \\ dt/dw = 1, \\ d\phi_Z(s,t)/dw = \theta(s-1)\phi_Z(s,t). \end{cases}$$

We solve the first two ODEs of the system, finding

$$\begin{cases} s(w) = c_1 e^{\tilde{\mu}w} + 1, \\ t(w) = c_2 + w. \end{cases} \tag{5.52}$$

We consider the characteristic line $(s(w), t(w))$, setting $(s(0), t(0)) = (s_0, 0)$. Therefore, substituting in (5.52), we obtain $c_1 = s_0 - 1$ and $c_2 = 0$. We can now find

$$\begin{cases} s(s_0, w) = (s_0 - 1)e^{\tilde{\mu}w} + 1, \\ t(s_0, w) = w. \end{cases} \tag{5.53}$$

Using the initial condition $\phi_Z(s(s_0,w), t(s_0,w) = 0) = s^{Z_0}$ together with (5.53), we can write the initial condition as $\phi_Z(s(s_0,w), t(s_0,w) = 0) = s_0^{Z_0}$. Thus

$$\begin{cases} d\phi_Z(s_0,w)/dw = \theta(s(s_0,w)-1)\phi_Z(s_0,w) = \theta(s_0-1)\phi_Z(s_0,w), \\ \phi_Z(s_0,0) = s_0^{Z_0}. \end{cases}$$

Therefore we have

$$\begin{cases} \phi_Z(s_0,w) = c_3 e^{\frac{\theta(s_0-1)e^{\tilde{\mu}w}}{\tilde{\mu}}}, \\ \phi_Z(s_0,0) = s_0^{Z_0}, \end{cases}$$

which gives $c_3 = s_0^{Z_0} e^{-\frac{\theta(s_0-1)}{\tilde{\mu}}}$. We now use (5.53) to find

$$\begin{cases} s_0(s,t) = (s-1)e^{-\tilde{\mu}t} + 1, \\ w(s,t) = t. \end{cases}$$

Finally, we obtain

$$\phi_Z(s,t) = c_3(s_0(s,t), w(s,t)) \cdot e^{\frac{\theta(s_0(s,t)-1)e^{\tilde{\mu}w(s,t)}}{\tilde{\mu}}},$$

which gives

$$\phi_Z(s,t) = \left[1 + (s-1)e^{-\tilde{\mu}t}\right]^{Z_0} \cdot \exp\left\{\frac{\theta(s-1)}{\tilde{\mu}}\left(1 - e^{-\tilde{\mu}t}\right)\right\}. \tag{5.54}$$

As previously said, we now apply this general solution to our particular case, where $Z(0) = Z_0 = 0$. Thus we have

$$\phi_Z(s,t) = \exp\left\{\frac{\theta(s-1)}{\tilde{\mu}}\left(1 - e^{-\tilde{\mu}t}\right)\right\}. \tag{5.55}$$

Recalling the general property of the probability generating function for which

$$p_z(t) = \Pr(Z(t) = z) = \frac{1}{z!}\left.\frac{\partial^z \phi_Z(s,t)}{\partial s^z}\right|_{s=0},$$

we can write

$$p_z(t) = \frac{1}{z!}\left(\frac{(1 - e^{-\tilde{\mu}t})\theta}{\tilde{\mu}}\right)^z e^{-\frac{(1-e^{-\tilde{\mu}t})\theta}{\tilde{\mu}}}. \tag{5.56}$$

We substitute $p_z(t)$ in (5.50) to obtain

$$\Pr\left(Z(T_N(A)) = z\right) = \int_0^{+\infty} \frac{1}{z!}\left(\frac{(1 - e^{-\tilde{\mu}t})\theta}{\tilde{\mu}}\right)^z e^{-\frac{(1-e^{-\tilde{\mu}t})\theta}{\tilde{\mu}}}(N-A)\binom{N}{A}\frac{(1 - e^{-\tilde{\mu}t})^N}{(e^{\tilde{\mu}t} - 1)^{A+1}}\tilde{\mu}\ dt.$$

This can be rewritten as

$$\Pr\left(Z(T_N(A)) = z\right) = \Omega \int_0^{+\infty} e^{-\frac{(1-e^{-\tilde{\mu}t})\theta}{\tilde{\mu}}}\frac{\left(1 - e^{-\tilde{\mu}t}\right)^{N+z}}{(e^{\tilde{\mu}t} - 1)^{A+1}}\ dt, \tag{5.57}$$

where

$$\Omega = \frac{\theta^z(N-A)\binom{N}{A}\tilde{\mu}}{z!\ \tilde{\mu}^z}.$$

We will now find an explicit solution for (5.57). First we define the new variable $p = \exp\{\tilde{\mu}t\}$, which implies $\mathrm{d}p = \tilde{\mu}\exp\{\tilde{\mu}t\}\ \mathrm{d}t$, $t = \log(p)/\tilde{\mu}$ and

$$\Pr\left(Z(T_N(A)) = z\right) = \Omega_1 \int_1^{+\infty}\left(e^{a\frac{p-1}{p}}\right)\frac{(p-1)^b}{p^c}\ \mathrm{d}p, \tag{5.58}$$

where

$$\begin{aligned}
\Omega_1 &= \frac{\theta^z(N-A)\binom{N}{A}}{z!\ \tilde{\mu}^z},\\
a &= -\frac{\theta}{\tilde{\mu}},\\
b &= N - A - 1 - z,\\
c &= N + z + 1.
\end{aligned}$$

Following similar steps, we define the new variable $q = \frac{p-1}{p}$, which implies $\mathrm{d}q = \frac{\mathrm{d}p}{p^2}$, $p = \frac{1}{1-q}$ and

$$\Pr\left(Z(T_N(A)) = z\right) = \Omega_1 \int_0^1 e^{aq} q^b (1-q)^A\ \mathrm{d}q. \tag{5.59}$$

Solving the integral on the right hand side of (5.59) is possible but requires quite a few steps. Therefore we focus now on the solution of this integral, coming back only later to the explicit solution of (5.59). Using integration by parts with

$$u = q^b(1-q)^A \text{ and}$$
$$\mathrm{d}v = e^{aq}\mathrm{d}q,$$

we obtain the recursive formula

$$\int_0^1 e^{aq}q^b(1-q)^A \,\mathrm{d}q = \frac{1}{a}\left[e^{aq}q^b(1-q)^A\right]_0^1 - \frac{b}{a}\int_0^1 e^{aq}q^{b-1}(1-q)^A \,\mathrm{d}q + \frac{A}{a}\int_0^1 e^{aq}q^b(1-q)^{A-1} \,\mathrm{d}q. \tag{5.60}$$

It is fundamental at this point to notice that

$$\left[e^{aq}q^b(1-q)^A\right]_0^1 \neq 0 \text{ only in the cases } b = 0 \text{ or } A = 0.$$

For simplicity of notation, define the function

$$f(b, A) = \left[e^{aq}q^b(1-q)^A\right]_0^1.$$

Thus, (5.60) becomes

$$\int_0^1 f(b, A) \,\mathrm{d}q = \frac{1}{a}\left[f(b, A)\right]_0^1 - \frac{b}{a}\int_0^1 f(b-1, A) \,\mathrm{d}q + \frac{A}{a}\int_0^1 f(b, A-1) \,\mathrm{d}q, \tag{5.61}$$

where

$$\left[f(b, A)\right]_0^1 \neq 0 \text{ only in the cases } b = 0 \text{ or } A = 0. \tag{5.62}$$

Applying recursively (5.61) to itself and eliminating all the different $[f(b-x, A-y)]_0^1$ for $x \neq b$ and $y \neq A$, it can be proved that each remaining item is of the kind

$$(-1)^x \binom{x+y}{x} \frac{b^{\underline{x}} A^{\underline{y}}}{a^{x+y}} \int_0^1 f(b-x, A-y) \,\mathrm{d}q,$$

where $n^{\underline{k}} = n(n-1)(n-2)\ldots(n-k+1)$. Given (5.62), the only integrals we are interested in are for $y = A$ and $x = b$ that is, respectively,

$$(-1)^x \binom{x+A}{x} \frac{b^{\underline{x}} A!}{a^{x+A}} \int_0^1 f(b-x, 0) \,\mathrm{d}q,$$

$$(-1)^b \binom{b+y}{b} \frac{b! A^{\underline{y}}}{a^{b+y}} \int_0^1 f(0, A-y) \,\mathrm{d}q.$$

It can also be proved that all these integrals have to be taken into consideration, that is for $x \in \{0, 1, \ldots, b\}$ and $y \in \{0, 1, \ldots, A\}$, excluding only the case $(x, y) = (0, 0)$ which does not exist. Therefore, in order to solve (5.59), we need now to focus on the two integrals

$$I_1(x) = \int_0^1 e^{aq}q^{b-x} \,\mathrm{d}q,$$

$$I_2(y) = \int_0^1 e^{aq}(1-q)^{A-y} \,\mathrm{d}q.$$

First we notice that, defining the new variable $w = 1 - q$, we have

$$I_2(y) = e^a \int_0^1 e^{\tilde{a}t} t^{A-y} \, \mathrm{d}t,$$

with $\tilde{a} = -a$. Thus, to solve both I_1 and I_2, we focus on finding a solution for the generic integral

$$\int_0^1 e^{cu} u^n \, \mathrm{d}u.$$

Integrating by parts, it is easy to show that

$$\int_0^1 e^{cu} u^n \, \mathrm{d}u = \left[e^{cu} \sum_{i=0}^{n} (-1)^i \frac{n! \, u^{n-i}}{(n-i)! \, c^{i+1}} \right]_0^1 = e^c \sum_{i=0}^{n} (-1)^i \frac{n!}{(n-i)! \, c^{i+1}}.$$

Therefore, we can write

$$I_1(x) = e^a \sum_{i=0}^{b-x} (-1)^i \frac{(b-x)!}{(b-x-i)! \, a^{i+1}},$$

$$I_2(y) = e^a e^{\tilde{a}} \sum_{i=0}^{A-y} (-1)^i \frac{(A-y)!}{(A-y-i)! \, \tilde{a}^{i+1}} = \sum_{i=0}^{A-y} -\frac{(A-y)!}{(A-y-i)! \, a^{i+1}}.$$

We are now finally able to write an explicit formula for $\Pr\left(Z(T_N(A)) = z\right)$. In fact we have

$$\Pr\left(Z(T_N(A)) = z\right) = \Omega_1 \int_0^1 f(b, A) \, \mathrm{d}q,$$

with

$$\Omega_1 = \frac{\theta^z (N-A) \binom{N}{A}}{z! \, \tilde{\mu}^z}, \quad \text{and}$$

$$\int_0^1 f(b, A) \, \mathrm{d}q = \sum_{x=0}^{b-1} (-1)^x \binom{x+A}{x} \frac{b^{\underline{x}} A!}{a^{x+A}} \, I_1(x) + \sum_{y=0}^{A-1} (-1)^b \binom{b+y}{b} \frac{b! A^{\underline{y}}}{a^{b+y}} \, I_2(y),$$

where

$$I_1 = e^a \sum_{i=0}^{b-x} (-1)^i \frac{(b-x)!}{(b-x-i)! \, a^{i+1}}$$

$$I_2 = \sum_{i=0}^{A-y} -\frac{(A-y)!}{(A-y-i)! \, a^{i+1}}.$$

The next section follows the first step argument analysis to find similar analytical results.

First step argument analysis

This section focuses on finding a formula for the probabilities $p_{(n,m)}(\bar{n}, A)$, when we consider the death rates $\mu^{(X)}_{n,m} = \mu^{(X,1)}_{n,m} = \tilde{\mu}(n-m)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,1)}_{n,m} = \tilde{\mu}m$ for the augmented process $\mathcal{X}^{\text{aug}}$. We recall that these probabilities are defined as

$$p_{(n,m)}(\bar{n}, A) = \Pr(\bar{u} = (\bar{n}, A) | u = (n, m)), \tag{5.63}$$

where u represents the initial state $(X(0), Y(0))$ and $\bar{u}$ represents the state hit by the process $\mathcal{X}^{\text{aug}}$ at time $T_N(A)$. The first step argument analysis gives

$$p_{(n,m)}(\bar{n}, A) = \frac{\delta_{n<M} \cdot \theta \cdot p_{(n+1,m)}(\bar{n}, A) + g^{(1)}(n, m) \cdot p_{(n-1,m)}(\bar{n}, A) + r^{(1)}(n, m)}{v^{(1)}(n)}, \tag{5.64}$$

where we define the functions $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $r^{(1)}(n, m)$ as

$$\begin{aligned} v^{(1)}(n) &= n\tilde{\mu} + \delta_{n<M} \cdot \theta, \\ g^{(1)}(n, m) &= (n - m)\tilde{\mu}, \\ r^{(1)}(n, m) &= m\tilde{\mu} \cdot p_{(n-1,m-1)}(\bar{n}, A). \end{aligned}$$

Note that the notations $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $r^{(1)}(n, m)$ are due to the usage of the first kind of death rates $\mu^{(X)}_{n,m} = \mu^{(X,1)}_{n,m} = (n - m)\tilde{\mu}$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,1)}_{n,m} = m\tilde{\mu}$ in this section. For the same reason, in the following section we will use the notation $v^{(2)}(n)$, $g^{(2)}(n, m)$ and $r^{(2)}(n, m)$ when dealing with the cases $\mu^{(X)}_{n,m} = \mu^{(X,2)}_{n,m} = (n - m)(\beta_1 + \beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,2)}_{n,m} = m(\beta_1 + \beta_2 pn)$. In order to find explicit solutions for the recursive equation (5.64), we first analyse the particular space of states $\{(n, m) : m = A + 1\}$. The functions $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $r^{(1)}(n, m)$, together with (5.64) give

$$p_{(M,A+1)}(\bar{n}, A) = \frac{r^{(1)}(M, A + 1)}{v^{(1)}(M)} + \frac{g^{(1)}(M, A + 1)}{v^{(1)}(M)} p_{(M-1,A+1)}(\bar{n}, A), \tag{5.65}$$

where $r^{(1)}(M, A + 1) = (A + 1)\tilde{\mu} \cdot p_{(M-1,A)}(\bar{n}, A)$ and

$$p_{(M-1,A)}(\bar{n}, A) = \begin{cases} 1 & \text{if } \bar{n} = M - 1, \\ 0 & \text{otherwise.}, \end{cases}$$

Let us define

$$a_M = r^{(1)}(M, A + 1);\ b_M = v^{(1)}(M);\ c_M = g^{(1)}(M, A + 1),$$

so that $p_{(M,A+1)}(\bar{n}, A)$ can be written as

$$p_{(M,A+1)}(\bar{n}, A) = \frac{a_M}{b_M} + \frac{c_M}{b_M} \cdot p_{(M-1,A+1)}(\bar{n}, A). \tag{5.66}$$

We continue with $p_{(M-1,A+1)}(\bar{n}, A)$. As for $p_{(M,A+1)}(\bar{n}, A)$, it can be written

$$p_{(M-1,A+1)}(\bar{n},A) = \frac{\theta\cdot p_{(M,A+1)}(\bar{n},A)}{v^{(1)}(M-1)} + \frac{g^{(1)}(M-1,A+1)\cdot p_{(M-2,A+1)}(\bar{n},A)}{v^{(1)}(M-1)} + \frac{r^{(1)}(M-1,A+1)}{v^{(1)}(M-1)}$$
$$= \frac{\theta a_M + b_M r^{(1)}(M-1,A+1)}{b_M v^{(1)}(M-1)-\theta c_M} + \frac{b_M g^{(1)}(M-1,A+1)}{b_M v^{(1)}(M-1)-\theta c_M}\cdot p_{(M-2,A+1)}(\bar{n},A), \tag{5.67}$$

where the last equality is obtained by replacing (5.66) in the first equality. We define

$$\begin{aligned} a_{M-1} &= \theta a_M + b_M r^{(1)}(M-1,A+1),\\ b_{M-1} &= b_M v^{(1)}(M-1) - \theta c_M,\\ c_{M-1} &= b_M g^{(1)}(M-1,A+1), \end{aligned}$$

so that $p_{(M-1,A+1)}(\bar{n},A)$ can be written as

$$p_{(M-1,A+1)}(\bar{n},A) = \frac{a_{M-1}}{b_{M-1}} + \frac{c_{M-1}}{b_{M-1}}\cdot p_{(M-2,A+1)}(\bar{n},A).$$

Thus, the general recursive relations

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} r^{(1)}(M-k,A+1),\\ b_{M-k} &= b_{M-k+1} v^{(1)}(M-k) - \theta c_{M-k+1},\\ c_{M-k} &= b_{M-k+1} g^{(1)}(M-k,A+1), \end{aligned} \tag{5.68}$$

with $a_M = r^{(1)}(M,A+1)$, $b_M = v^{(1)}(M)$, $c_M = g^{(1)}(M,A+1)$ and

$$p_{(M-k,A+1)}(\bar{n},A) = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}}\cdot p_{(M-k-1,A+1)}(\bar{n},A). \tag{5.69}$$

In order to obtain explicit solution for the general $p_{(n,A+1)}(\bar{n},A)$, an algorithm has to be followed, as previously done for the study of the stochastic descriptor $T_N(A)$. Before showing the different steps of the algorithm, we notice that (5.64) and (5.69) give, respectively, the two equations of the following system

$$\begin{cases} p_{(A+1,A+1)}(\bar{n},A) = \dfrac{\theta}{v^{(1)}(A+1)} p_{(A+2,A+1)}(\bar{n},A) \;+\; \dfrac{r^{(1)}(A+1,A+1)}{v^{(1)}(A+1)},\\[2ex] p_{(A+2,A+1)}(\bar{n},A) = \dfrac{a_{A+2}}{b_{A+2}} + \dfrac{c_{A+2}}{b_{A+2}} p_{(A+1,A+1)}(\bar{n},A), \end{cases}$$

which gives

$$\begin{cases} p_{(A+1,A+1)}(\bar{n},A) = \dfrac{\theta a_{A+2} + b_{A+2} r^{(1)}(A+1,A+1)}{b_{A+2} v^{(1)}(A+1) - \theta c_{A+2}},\\[2ex] p_{(A+2,A+1)}(\bar{n},A) = \dfrac{a_{A+2} v^{(1)}(A+1) + c_{A+2} r^{(1)}(A+1,A+1)}{b_{A+2} v^{(1)}(A+1) - \theta c_{A+2}}. \end{cases}$$

Given the two values $p_{(A+1,A+1)}(\bar{n},A)$ and $p_{(A+2,A+1)}(\bar{n},A)$ dependent on a_{A+2}, b_{A+2} and c_{A+2}, the steps of the algorithm is as follows:

- Start with $a_M = r^{(1)}(M, A+1)$, $b_M = v^{(1)}(M)$, $c_M = g^{(1)}(M, A+1)$;
- Use the recursive relations (5.68) to find a_{A+2}, b_{A+2} and c_{A+2};
- Find $p_{(A+1,A+1)}(\bar{n}, A)$ and $p_{(A+2,A+1)}(\bar{n}, A)$;
- Use (5.69) to find all the different values of $p_{(n,A+1)}(\bar{n}, A)$, up to $p_{(M,A+1)}(\bar{n}, A)$.

The same steps are now applied to the more general states $\{(n, m) : m \in \{A+2, A+3, \cdots, N\}\}$. We recall the functions $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $r^{(1)}(n, m)$ as

$$\begin{aligned} v^{(1)}(n) &= n\tilde{\mu} + \delta_{n<M} \cdot \theta, \\ g^{(1)}(n, m) &= (n - m)\tilde{\mu}, \\ r^{(1)}(n, m) &= m\tilde{\mu} \cdot p_{(n-1,m-1)}(\bar{n}, A). \end{aligned}$$

These definitions, together with (5.64), give

$$p_{(M,m)}(\bar{m}, A) = \frac{r^{(1)}(M, m)}{v^{(1)}(M)} + \frac{g^{(1)}(M, m)}{v^{(1)}(M)} p_{(M-1,m)}(\bar{n}, A). \tag{5.70}$$

With the same techniques used for computing (5.67), we can find

$$p_{(M-1,m)}(\bar{n}, A) = \frac{\theta \cdot p_{(M,m)}(\bar{n}, A)}{v^{(1)}(M-1)} + \frac{g^{(1)}(M-1, m)}{v^{(1)}(M-1)} \cdot p_{(M-2,m)}(\bar{n}, A) + \frac{r^{(1)}(M-1, m)}{v^{(1)}(M-1)}, \tag{5.71}$$

obtaining the relations

$$\begin{aligned} a_{M-1} &= \theta a_M + b_M r^{(1)}(M-1, m), \\ b_{M-1} &= b_M v^{(1)}(M-1) - \theta c_M, \\ c_{M-1} &= b_M g^{(1)}(M-1, m), \end{aligned}$$

and, recursively, the general ones

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} r^{(1)}(M-k, m), \\ b_{M-k} &= b_{M-k+1} v^{(1)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(1)}(M-k, m), \end{aligned} \tag{5.72}$$

with $a_M = r^{(1)}(M, m)$, $b_M = v^{(1)}(M)$, $c_M = g^{(1)}(M, m)$ and

$$p_{(M-k,m)}(\bar{n}, A) = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot p_{(M-k-1,m)}(\bar{n}, A). \tag{5.73}$$

To obtain explicit solution for the general $p_{(n,m)}(\bar{n}, A)$, we notice once again that (5.64) and (5.73) give, respectively, the two equations of the following system

$$\begin{cases} p_{(m,m)}(\bar{n}, A) = \dfrac{\theta}{v^{(1)}(m)} p_{(m+1,m)}(\bar{n}, A) + \dfrac{r^{(1)}(m, m)}{v^{(1)}(m)}, \\ p_{(m+1,m)}(\bar{n}, A) = \dfrac{a_{m+1}}{b_{m+1}} + \dfrac{c_{m+1}}{b_{m+1}} p_{(m,m)}(\bar{n}, A), \end{cases}$$

which gives

$$\begin{cases} p_{(m,m)}(\bar{n}, A) = \dfrac{\theta a_{m+1} + b_{m+1} r^{(1)}(m,m)}{b_{m+1} v^{(1)}(m) - \theta c_{m+1}}, \\ \\ p_{(m+1,m)}(\bar{n}, A) = \dfrac{a_{m+1} v^{(1)}(m) + c_{m+1} r^{(1)}(m,m)}{b_{m+1} v^{(1)}(m) - \theta c_{m+1}}. \end{cases}$$

Given these two values $p_{(m,m)}(\bar{n}, A)$ and $p_{(m+1,m)}(\bar{n}, A)$ dependent on a_{m+1}, b_{m+1} and c_{m+1}, the algorithm is as follows:

- Start with $a_M = r^{(1)}(M, m)$, $b_M = v^{(1)}(M)$, $c_M = g^{(1)}(M, m)$;
- Use the recursive relations (5.72) to find a_{m+1}, b_{m+1} and c_{m+1};
- Find $p_{(m,m)}(\bar{n}, A)$ and $p_{(m+1,m)}(\bar{n}, A)$;
- Use (5.73) to find all the different values of $p_{(n,m)}(\bar{n}, A)$, up to $p_{(M,m)}(\bar{n}, A)$.

Simulations of this algorithm and numerical results of a Gillespie code representing the same biological process can be found in Figures 5.13 and 5.14 respectively. The hitting probabilities of a particular final state are also plotted for a specific initial state as function of the variables n_θ, θ, γ, M_c and μ, as a heat map in Figures 5.15 to 5.24.

5.5.2 Explicit competition

This section focuses on finding a formula for the probabilities $p_{(n,m)}(\bar{n}, A)$, when we consider the death rates $\mu^{(X)}_{n,m} = \mu^{(X,2)}_{n,m} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,2)}_{n,m} = m(\beta_1 + \beta_2 pn)$ for the augmented process $\mathcal{X}^{\text{aug}}$. As for the previous section, probabilities are defined as

$$p_{(n,m)}(\bar{n}, A) = \Pr(\bar{u} = (\bar{n}, A) | u = (n, m)), \tag{5.74}$$

where u represents the initial state $(X(0), Y(0))$ and $\bar{u}$ represents the first state hit by the process $\mathcal{X}^{\text{aug}}$ at time $T_N(A)$. Following a first step argument we obtain the equation

$$p_{(n,m)}(\bar{n}, A) = \frac{\delta_{n<M} \cdot \theta \cdot p_{(n+1,m)}(\bar{n}, A) + g^{(2)}(n,m) \cdot p_{(n-1,m)}(\bar{n}, A) + r^{(2)}(n,m)}{v^{(2)}(n)}, \tag{5.75}$$

where we define the functions $v^{(2)}(n)$, $g^{(2)}(n, m)$ and $r^{(2)}(n, m)$ as

$$\begin{aligned} v^{(2)}(n) &= n(\beta_1 + n\beta_2 p) + \delta_{n<M} \cdot \theta, \\ g^{(2)}(n,m) &= (\beta_1 + n\beta_2 p)(n-m), \\ r^{(2)}(n,m) &= m(\beta_1 + n\beta_2 p) \cdot p_{(n-1,m-1)}(\bar{n}, A). \end{aligned}$$

In order to find explicit solutions for the recursive equation (5.75), we focus on the particular space of states $\{(n,m) : m = A+1\}$. The functions $v^{(2)}(n)$, $g^{(2)}(n,m)$ and $r^{(2)}(n,m)$, together with (5.75) give

$$p_{(M,A+1)}(\bar{n},A) = \frac{r^{(2)}(M,A+1)}{v^{(2)}(M)} + \frac{g^{(2)}(M,A+1)}{v^{(2)}(M)} p_{(M-1,A+1)}(\bar{n},A), \tag{5.76}$$

where $r^{(2)}(M,A+1) = (A+1)(\beta_1 + \beta_2 pM) \cdot p_{(M-1,A)}(\bar{n},A)$ and

$$p_{(M-1,A)}(\bar{n},A) = \begin{cases} 1 & \text{if } \bar{n} = M-1, \\ 0 & \text{otherwise.} \end{cases}$$

Let us define

$$a_M = r^{(2)}(M,A+1);\ b_M = v^{(2)}(M);\ c_M = g^{(2)}(M,A+1),$$

so that $p_{(M,A+1)}(\bar{n},A)$ can be written as

$$p_{(M,A+1)}(\bar{n},A) = \frac{a_M}{b_M} + \frac{c_M}{b_M} \cdot p_{(M-1,A+1)}(\bar{n},A). \tag{5.77}$$

We consider $p_{(M-1,A+1)}(\bar{n},A)$. As for $p_{(M,A+1)}(\bar{n},A)$, it can be written

$$\begin{aligned} p_{(M-1,A+1)}(\bar{n},A) = {} & \frac{\theta \cdot p_{(M,A+1)}(\bar{n},A)}{v^{(2)}(M-1)} + \frac{r^{(2)}(M-1,A+1)}{v^{(2)}(M-1)} \\ & + \frac{g^{(2)}(M-1,A+1)}{v^{(2)}(M-1)} \cdot p_{(M-2,A+1)}(\bar{n},A), \end{aligned} \tag{5.78}$$

that is

$$\begin{aligned} p_{(M-1,A+1)}(\bar{n},A) = {} & \frac{\theta a_M + b_M r^{(2)}(M-1,A+1)}{b_M v^{(2)}(M-1) - \theta c_M} \\ & + \frac{b_M g^{(2)}(M-1,A+1)}{b_M v^{(2)}(M-1) - \theta c_M} \cdot p_{(M-2,A+1)}(\bar{n},A), \end{aligned} \tag{5.79}$$

where the equality is obtained by replacing (5.77) in (5.78). We define

$$\begin{aligned} a_{M-1} &= \theta a_M + b_M r^{(2)}(M-1,A+1), \\ b_{M-1} &= b_M v^{(2)}(M-1) - \theta c_M, \\ c_{M-1} &= b_M g^{(2)}(M-1,A+1), \end{aligned}$$

so that $p_{(M-1,A+1)}(\bar{n},A)$ can be written as

$$p_{(M-1,A+1)}(\bar{n},A) = \frac{a_{M-1}}{b_{M-1}} + \frac{c_{M-1}}{b_{M-1}} \cdot p_{(M-2,A+1)}(\bar{n},A).$$

The general recursive relations follow

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} r^{(2)}(M-k,A+1), \\ b_{M-k} &= b_{M-k+1} v^{(2)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(2)}(M-k,A+1), \end{aligned} \tag{5.80}$$

with $a_M = r(M, A+1)$, $b_M = v(M)$, $c_M = g(M, A+1)$ and

$$p_{(M-k,A+1)}(\bar{n}, A) = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot p_{(M-k-1,A+1)}(\bar{n}, A). \tag{5.81}$$

We notice that (5.75) and (5.81) give, respectively, the two equations of the following system

$$\begin{cases} p_{(A+1,A+1)}(\bar{n}, A) = \dfrac{\theta}{v^{(2)}(A+1)} p_{(A+2,A+1)}(\bar{n}, A) \ + \ \dfrac{r^{(2)}(A+1, A+1)}{v^{(2)}(A+1)}, \\ p_{(A+2,A+1)}(\bar{n}, A) = \dfrac{a_{A+2}}{b_{A+2}} + \dfrac{c_{A+2}}{b_{A+2}} p_{(A+1,A+1)}(\bar{n}, A), \end{cases}$$

which gives

$$\begin{cases} p_{(A+1,A+1)}(\bar{n}, A) = \dfrac{\theta a_{A+2} + b_{A+2} r^{(2)}(A+1, A+1)}{b_{A+2} v^{(2)}(A+1) - \theta c_{A+2}}, \\ p_{(A+2,A+1)}(\bar{n}, A) = \dfrac{a_{A+2} v^{(2)}(A+1) + c_{A+2} r^{(2)}(A+1, A+1)}{b_{A+2} v^{(2)}(A+1) - \theta c_{A+2}}. \end{cases}$$

These two values $p_{(A+1,A+1)}(\bar{n}, A)$ and $p_{(A+2,A+1)}(\bar{n}, A)$, dependent on a_{A+2}, b_{A+2} and c_{A+2}, are now used as part of the following algorithm:

- Start with $a_M = r^{(2)}(M, A+1)$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M, A+1)$;
- Use the recursive relations (5.45) to find a_{A+2}, b_{A+2} and c_{A+2};
- Find $p_{(A+1,A+1)}(\bar{n}, A)$ and $p_{(A+2,A+1)}(\bar{n}, A)$;
- Use (5.81) to find all the different values of $p_{(n,A+1)}(\bar{n}, A)$, up to $p_{(M,A+1)}(\bar{n}, A)$.

We focus now on the general space of states $\{(n, m) : m \in \{A+2, A+3, \cdots, N\}\}$, recalling the functions $v^{(2)}(n)$, $g^{(2)}(n, m)$ and $r^{(2)}(n, m)$

$$\begin{aligned} v^{(2)}(n) &= n(\beta_1 + n\beta_2 p) + \delta_{n<M} \cdot \theta, \\ g^{(2)}(n, m) &= (\beta_1 + n\beta_2 p)(n - m), \\ r^{(2)}(n, m) &= m(\beta_1 + n\beta_2 p) \cdot p_{(n-1,m-1)}(\bar{n}, A). \end{aligned}$$

We have from (5.75)

$$p_{(M,m)}(\bar{n}, A) = \frac{r^{(2)}(M, m)}{v^{(2)}(M)} + \frac{g^{(2)}(M, m)}{v^{(2)}(M)} p_{(M-1,m)}(\bar{n}, A). \tag{5.82}$$

With the same techniques used for computing (5.79), we can find

$$p_{(M-1,m)}(\bar{n}, A) = \frac{\theta \cdot p_{(M,m)}(\bar{n}, A)}{v^{(2)}(M-1)} + \frac{g^{(2)}(M-1, m)}{v^{(2)}(M-1)} \cdot p_{(M-2,m)}(\bar{n}, A) + \frac{r^{(2)}(M-1, m)}{v^{(2)}(M-1)} \tag{5.83}$$

obtaining the relations

$$\begin{aligned} a_{M-1} &= \theta a_M + b_M r^{(2)}(M-1,m), \\ b_{M-1} &= b_M v^{(2)}(M-1) - \theta c_M, \\ c_{M-1} &= b_M g^{(2)}(M-1,m), \end{aligned}$$

and, recursively, the general ones

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} r^{(2)}(M-k,m), \\ b_{M-k} &= b_{M-k+1} v^{(2)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(2)}(M-k,m), \end{aligned} \tag{5.84}$$

with $a_M = r^{(2)}(M,m)$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M,m)$ and

$$p_{(M-k,m)}(\bar{n},A) = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot p_{(M-k-1,m)}(\bar{n},A). \tag{5.85}$$

Equations (5.75) and (5.85) give, respectively, the two equations of the following system

$$\begin{cases} p_{(m,m)}(\bar{n},A) = \dfrac{\theta}{v^{(2)}(m)} p_{(m+1,m)}(\bar{n},A) \; + \; \dfrac{r^{(2)}(m,m)}{v^{(2)}(m)}, \\ \\ p_{(m+1,m)}(\bar{n},A) = \dfrac{a_{m+1}}{b_{m+1}} + \dfrac{c_{m+1}}{b_{m+1}} p_{(m,m)}(\bar{n},A), \end{cases}$$

which gives

$$\begin{cases} p_{(m,m)}(\bar{n},A) = \dfrac{\theta a_{m+1} + b_{m+1} r^{(2)}(m,m)}{b_{m+1} v^{(2)}(m) - \theta c_{m+1}}, \\ \\ p_{(m+1,m)}(\bar{n},A) = \dfrac{a_{m+1} v^{(2)}(m) + c_{m+1} r^{(2)}(m,m)}{b_{m+1} v^{(2)}(m) - \theta c_{m+1}}. \end{cases}$$

The final algorithm follows:

- Start with $a_M = r^{(2)}(M,m)$, $b_M = v^{(2)}(M)$, $c_M = g^{(2)}(M,m)$;
- Use the recursive relations (5.84) to find a_{m+1}, b_{m+1} and c_{m+1};
- Find $p_{(m,m)}(\bar{n},A)$ and $p_{(n+1,m)}(\bar{n},A)$;
- Use (5.85) to find all the different values of $p_{(n,m)}(\bar{n},A)$, up to $p_{(M,m)}(\bar{n},A)$.

Simulations of this algorithm and numerical results of a Gillespie code representing the same biological process can be found in Figures 5.41 and 5.42 respectively. The hitting probabilities of a particular final state are also plotted for a specific initial state as function of both the β_1 and $\beta_2 p$ variables, as a heat map in Figure 5.43.

5.6 Maximum repertoire diversity in $[0, T_N(A)]$

Given the augmented process $\mathfrak{X}^{\text{aug}} = \{(X(t), Y(t)) : t \geq 0\}$ defined on the space of states $\mathbb{S}^{aug} = \{(n, m) : m \in \{0, 1, 2, \ldots, X(0)\},\ n \geq m\}$, we recall the definition of $T_N(A)$ as the time when, for the first time, the process $\mathfrak{X}^{\text{aug}}$ reaches the space of states $\mathbb{A} = \{(n, m) : m = A\}$. Our goal here is to derive the distribution of the maximum value of total clonotypes $X^{\max} = \max\{X(t) : t \in [0, T_N(A)]\}$ that the augmented process $\mathfrak{X}^{\text{aug}} = \{(X(t), Y(t)) : t \geq 0\}$, defined on the space of states $\mathbb{S}^{aug} = \{(n, m) : m \in \{0, 1, 2, \ldots, X(0)\},\ n \geq m\}$, reaches during the period of time $[0, T_N(A)]$. In particular, we analyse the probabilities $\phi_{(n,m)} = \Pr(X^{\max} \geq M | (X(0), Y(0)) = (n, m)\})$ for the stochastic descriptor $X^{\max}$ to reach (and maybe overpass) a certain value $M > X(0)$. We believe this descriptor is an important measure of the renewal dynamics of the repertoire, as able to describe the overpopulation effects of a population subject to renewal. Such a descriptor has been previously studied in the literature in relation to a generic two-species competition process [70]. Other authors apply similar steps to the study of the repertoire dynamics, not considering though the renewal problem we are focusing on [14]. This descriptor is also relevant from a technical point for the section on numerical results. In fact, it is only through the analyses of these probabilities that we can obtain a good idea of which value of M to use for the simulations of the two previously discussed stochastic descriptors.

5.6.1 Implicit competition

We find a formula for the probabilities $\phi_{(n,m)}$, when we consider the death rates $\mu^{(X)}_{n,m} = \mu^{(X,1)}_{n,m} = \tilde{\mu}(n - m)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,1)}_{n,m} = \tilde{\mu}m$ for the augmented process $\mathfrak{X}^{\text{aug}}$. We recall that these probabilities are defined as

$$\phi_{(n,m)} = \Pr(X^{\max} \geq M | (X(0), Y(0)) = (n, m)\}). \tag{5.86}$$

Using the first step argument as explained in the previous sections, we write

$$\phi_{(n,m)} = \frac{\delta_{n<M} \cdot \theta \cdot \phi_{(n+1,m)} + g^{(1)}(n, m) \cdot \phi_{(n-1,m)} + w^{(1)}(n, m)}{v^{(1)}(n)}, \tag{5.87}$$

where the functions $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $w^{(1)}(n, m)$ are defined as

$$\begin{aligned} v^{(1)}(n) &= n\tilde{\mu} + \delta_{n<M} \cdot \theta \\ g^{(1)}(n, m) &= (n - m)\tilde{\mu} \\ w^{(1)}(n, m) &= m\tilde{\mu} \cdot \phi_{(n-1,m-1)}. \end{aligned}$$

Note that the notations $v^{(1)}(n)$, $g^{(1)}(n, m)$ and $w^{(1)}(n, m)$ are due to the use of the first kind of death rates $\mu^{(X)}_{n,m} = \mu^{(X,1)}_{n,m} = (n - m)\tilde{\mu}$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,1)}_{n,m} = m\tilde{\mu}$ in this section.

For the same reason, in the following section we will use the notation $v^{(2)}(n)$, $g^{(2)}(n,m)$ and $w^{(2)}(n,m)$ when dealing with the cases $\mu_{n,m}^{(X)} = \mu_{n,m}^{(X,2)} = (n-m)(\beta_1 + \beta_2 pn)$ and $\mu_{n,m}^{(Y)} = \mu_{n,m}^{(Y,2)} = m(\beta_1 + \beta_2 pn)$. The boundary conditions for this stochastic descriptor are

$$\phi_{(M,m)} = 1 \ \ \forall m \in [A, N], \tag{5.88}$$

$$\phi_{(n,A)} = \begin{cases} 1 & \text{if } n = M, \\ 0 & \text{otherwise.} \end{cases} \tag{5.89}$$

In order to find explicit solutions for the recursive equation (5.87), we focus on the general space of states $\{(X(t), Y(t)) : Y(t) = m, \ m \in \{A+1, A+2, \dots Y(0)\}\}$, as shown in the previous section. Following the same steps of the previous sections, the functions $v(n)$, $g(n,m)$ and $w(n,m)$, together with (5.87), give (for $k \geq 1$)

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} w^{(1)}(M-k, m), \\ b_{M-k} &= b_{M-k+1} v^{(1)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(1)}(M-k, m), \end{aligned} \tag{5.90}$$

with $a_{M-1} = \theta + w^{(1)}(M-1, m)$, $b_{M-1} = v^{(1)}(M-1)$, $c_{M-1} = g^{(1)}(M-1, m)$ and

$$\phi_{(M-k,m)} = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}} \cdot \phi_{(M-k-1,m)}. \tag{5.91}$$

In order to obtain explicit solution for the general $\phi(n,m)$, an algorithm has to be followed. Equations (5.87) and (5.91) give, respectively, the two equations of the following system

$$\begin{cases} \phi_{(m,m)} = \dfrac{\theta}{v^{(1)}(m)} \phi_{(m+1,m)} + \dfrac{w^{(1)}(m,m)}{v^{(1)}(m)}, \\ \phi_{(m+1,m)} = \dfrac{a_{m+1}}{b_{m+1}} + \dfrac{c_{m+1}}{b_{m+1}} \phi_{(m,m)}, \end{cases}$$

which gives

$$\begin{cases} \phi_{(m,m)} = \dfrac{\theta a_{m+1} + b_{m+1} w^{(1)}(m,m)}{b_{m+1} v^{(1)}(m) - \theta c_{m+1}}, \\ \phi_{(m+1,m)} = \dfrac{a_{m+1} v^{(1)}(m) + c_{m+1} w^{(1)}(m,m)}{b_{m+1} v^{(1)}(m) - \theta c_{m+1}}. \end{cases}$$

Given these two values $\phi_{(m,m)}$ and $\phi_{(m+1,m)}$ dependent on a_{m+1}, b_{m+1} and c_{m+1}, we can now give the steps of the algorithm as follows:

- Start with $a_{M-1} = \theta + w^{(1)}(M-1, m)$, $b_{M-1} = v^{(1)}(M-1)$, $c_{M-1} = g^{(1)}(M-1, m)$;
- Use the recursive relations (5.90) to find a_{m+1}, b_{m+1} and c_{m+1};
- Find $\phi_{(m,m)}$ and $\phi_{(m+1,m)}$;

- Use (5.91) to find all the different values of $\phi_{(n,m)}$, up to $\phi_{(M-1,m)}$.

The reader can find simulations for this algorithm and for a Gillespie code representing the same biological process in Figures 5.25, 5.26 and 5.27 respectively. The hitting probabilities of a particular final state are also plotted for a specific initial state as function of the variables n_θ, θ, γ, M_c and μ, as a heat map in Figures 5.28 to 5.37.

5.6.2 Explicit competition

We find a formula for the probabilities $\phi_{(n,m)}$, when we consider the death rates $\mu^{(X)}_{n,m} = \mu^{(X,2)}_{n,m} = (n-m)(\beta_1+\beta_2 pn)$ and $\mu^{(Y)}_{n,m} = \mu^{(Y,2)}_{n,m} = m(\beta_1+\beta_2 pn)$ for the augmented process $\mathcal{X}^{\text{aug}}$. We recall that these probabilities are defined as

$$\phi_{(n,m)} = \Pr(X^{\max} \geq M | (X(0),Y(0)) = (n,m)\}). \tag{5.92}$$

Using the first step argument as explained in the previous sections, we can write

$$\phi_{(n,m)} = \frac{\delta_{n<M}\cdot\theta\cdot\phi_{(n+1,m)} + g^{(2)}(n,m)\cdot\phi_{(n-1,m)} + w^{(2)}(n,m)}{v^{(2)}(n)}, \tag{5.93}$$

where the functions $v^{(2)}(n)$, $g^{(2)}(n,m)$ and $w^{(2)}(n,m)$ are defined as

$$\begin{aligned} v^{(2)}(n) &= n(\beta_1+n\beta_2 p) + \delta_{n<M}\cdot\theta, \\ g^{(2)}(n,m) &= (\beta_1+n\beta_2 p)(n-m), \\ w^{(2)}(n,m) &= m(\beta_1+n\beta_2 p)\cdot\phi_{(n-1,m-1)}. \end{aligned}$$

The boundary conditions for this stochastic descriptor in this process are

$$\phi_{(M,m)} = 1 \ \ \forall m \in [A,N], \tag{5.94}$$

$$\phi_{(n,A)} = \begin{cases} 1 & \text{if } n = M, \\ 0 & \text{otherwise.} \end{cases} \tag{5.95}$$

In order to find explicit solutions for the recursive equation (5.93), we focus, as shown in the previous section, on the general space of states $\{(X(t),Y(t)) : Y(t) \equiv m\}$. The functions $v^{(2)}(n)$, $g^{(2)}(n,m)$ and $w^{(2)}(n,m)$, together with (5.93) give (for $k \geq 1$)

$$\begin{aligned} a_{M-k} &= \theta a_{M-k+1} + b_{M-k+1} w^{(2)}(M-k,m), \\ b_{M-k} &= b_{M-k+1} v^{(2)}(M-k) - \theta c_{M-k+1}, \\ c_{M-k} &= b_{M-k+1} g^{(2)}(M-k,m), \end{aligned} \tag{5.96}$$

with $a_{M-1} = \theta + w^{(2)}(M-1,m)$, $b_{M-1} = v^{(2)}(M-1)$, $c_{M-1} = g^{(2)}(M-1,m)$ and

$$\phi_{(M-k,m)} = \frac{a_{M-k}}{b_{M-k}} + \frac{c_{M-k}}{b_{M-k}}\cdot\phi_{(M-k-1,m)}. \tag{5.97}$$

Equations (5.93) and (5.97) give, respectively, the two equations of the following system

$$\begin{cases} \phi_{(m,m)} = \dfrac{\theta}{v^{(2)}(m)}\phi_{(m+1,m)} + \dfrac{w^{(2)}(m,m)}{v^{(2)}(m)}, \\ \phi_{(m+1,m)} = \dfrac{a_{m+1}}{b_{m+1}} + \dfrac{c_{m+1}}{b_{m+1}}\phi_{(m,m)}, \end{cases}$$

which gives

$$\begin{cases} \phi_{(m,m)} = \dfrac{\theta a_{m+1} + b_{m+1}w^{(2)}(m,m)}{b_{m+1}v^{(2)}(m) - \theta c_{m+1}}, \\ \phi_{(m+1,m)} = \dfrac{a_{m+1}v^{(2)}(m) + c_{m+1}w^{(2)}(m,m)}{b_{m+1}v^{(2)}(m) - \theta c_{m+1}}. \end{cases}$$

Given these two values $\phi_{(m,m)}$ and $\phi_{(m+1,m)}$ dependent on a_{m+1}, b_{m+1} and c_{m+1}, we can now give the steps of the algorithm as follows:

- Start with $a_{M-1} = \theta + w^{(2)}(M-1,m)$, $b_{M-1} = v^{(2)}(M-1)$, $c_{M-1} = g^{(2)}(M-1,m)$;
- Use the recursive relations (5.96) to find a_{m+1}, b_{m+1} and c_{m+1};
- Find $\phi_{(m,m)}$ and $\phi_{(m+1,m)}$;
- Use (5.97) to find all the different values of $\phi_{(n,m)}$, up to $\phi_{(M-1,m)}$.

The reader can find simulations for this algorithm and for a Gillespie code representing the same biological process in Figures 5.44, 5.45 and 5.46 respectively. The hitting probabilities of a particular final state are also plotted for a specific initial state as function of both the β_1 and $\beta_2 p$ variables, as a heat map in Figure 5.47.

5.7 Numerical results

In this section we display results from small-scale numerical simulations. We decided to rescale the system with different parameters than those chosen in [100] for computational reasons. We explain here the reasoning behind the rescaling procedure, and the invariances maintained from the clonotype perspective. Let us consider a repertoire, in homeostasis conditions, made of $N(0) = 50$ initial clonotype classes. The average clonal size equals 10 T cells. Thus, we consider a total of 500 initial T cells. The number of T cells of any new clonotype coming out of the thymus is $n_\theta = 4$. The mean cell death rate is maintained to $\mu = 0.5$ year^{-1}, as in [100]. The total number of self pMHC subsets is chosen following the relation $M_c = 4N(0) = 200$. This relation is kept from the simulations of Figure 7 in [100]. It is important to notice that we changed the notation of the total number of self pMHC subsets, from M to M_c, since in our notation M represents the maximum number of possible clonotype classes in the simulated repertoire. This parameter is important for the realization of the first step argument analysis, as explained in (5.40). We set $pM_c = 10$, to maintain an invariance from the clonotype perspective, since pM_c is the mean number of self pMHC recognised by a single TCR clonotype. This gives p=0.05, which also gives $pN(0) = 2.5$, that is the mean number of TCR clonotypes that are able to recongnise a given self pMHC. As explained in [100], we consider to biological case in which the division of peripheral cells gives the dominant contribution (as it is for adult humans), therefore approximating the mean number of total T cells in the system with $(\gamma M_c)/\mu$. Thus we have $(\gamma M_c)/\mu = 500$, giving $\gamma = 1.25$ year^{-1}. We extrapolate the relation $50\theta n_\theta = (\gamma M_c)/\mu$. Thus we obtain $\theta = 2.5$ year^{-1}. Sensitivity analysis will be done on the parameter θ, as well as on the other parameters. To conclude, we note that the average number of clonotype classes (i.e., 50) equals $\theta/\tilde{\mu}$, being $\tilde{\mu} = 0.05$ year^{-1} (see definition of $\tilde{\mu}$ explained in Section 5.2.1). This is coherent with the fact that, in homeostasis conditions, the average number of clonotype classes should be around $\theta/\tilde{\mu}$.

5.7.1 Implicit competition

This section presents the numerical results for the stochastic model with implicit competition, as described in Section 5.2.1. Results are shown for both the stochastic descriptors defining (i) the size of the repertoire at time $T_N(A)$ and (ii) the maximum repertoire diversity in $[0, T_N(A)]$. Solutions of the first step argument methods of Section 5.5.1 and Section 5.6.1 are shown for different initial states, together with Gillespie simulations which confirm the correctness of the first step argument methods. Sensitivity analysis is also performed over the different paramenters θ, n_θ, M_c, γ and μ.

Size of repertoire at time $T_N(A)$

We present here the results for the stochastic descriptor defining the size of the repertoire at time $T_N(A)$. Parameters are chosen in accordance to Section 5.7. Figure represents the solution of the system described in Section 5.5.1. The hitting state $(35, 25)$ and the different initial states $\{(x, y) : x \in \{48, \ldots, 60\}, y \in \{47, \ldots, 50\}\}$ have been chosen just as examples. Numerical (Gillespie) simulations of the same process are shown in Figure 5.14, with hitting state $(35, 25)$ and initial states $\{(x, 50) : x \in \{50, \ldots, 60\}\}$. Sensitivity analyses are then presented in Figures 5.15 - 5.24.

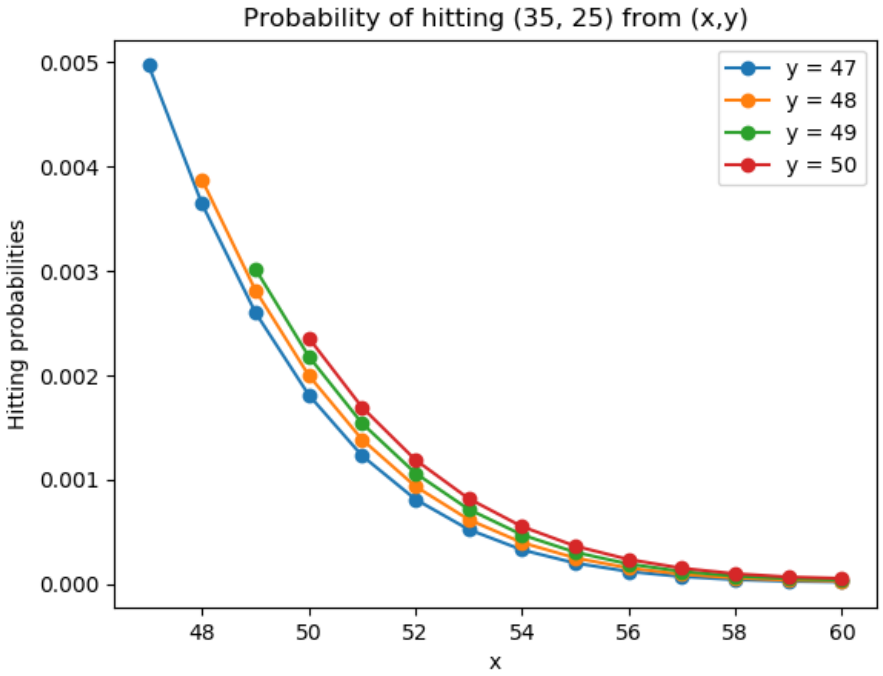

Figure 5.13: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $M_c = 200$, $\gamma = 1.25$ year^{-1} and $\mu = 0.5$ year^{-1}. Different colours represent different values of y. The hitting probabilities, for a given y, are plotted as functions of x.

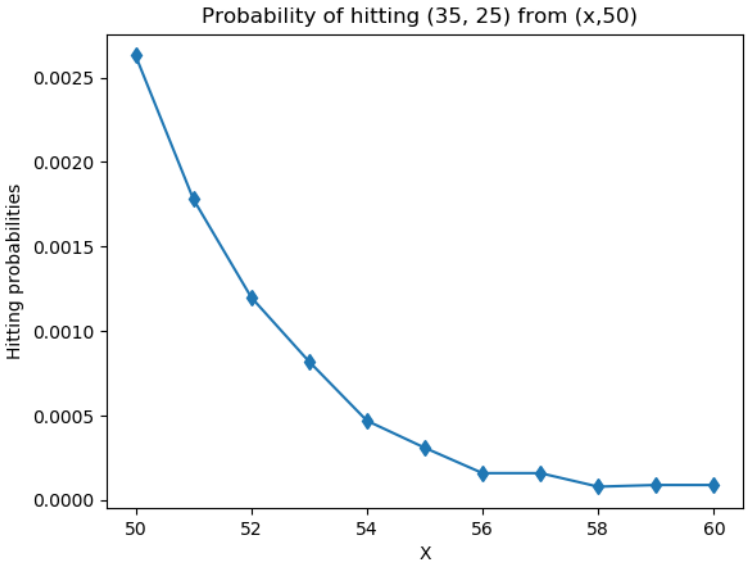

Figure 5.14: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $M_c = 200$, $\gamma = 1.25$ year^{-1} and $\mu = 0.5$ year^{-1}. The plot represents the hitting probabilities of state $(35, 25)$ from different states $(x, 50)$. Number of simulations $= 10^5$.

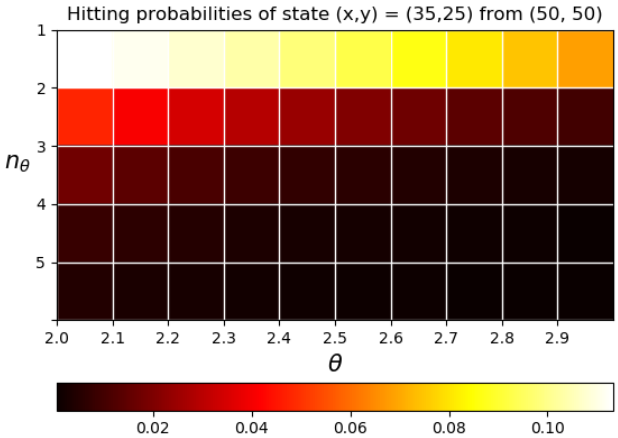

Figure 5.15: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both n_θ and θ variables. Fixed parameters are $M_c = 200$, $\gamma = 1.25$ year^{-1} and $\mu = 0.5$ year^{-1}.

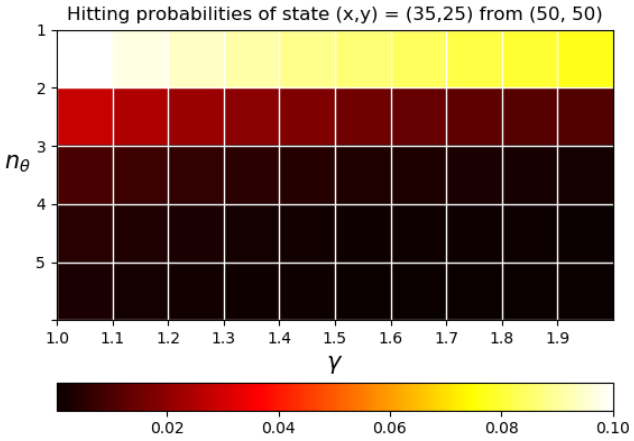

Figure 5.16: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both n_θ and γ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $M_c = 200$ and $\mu = 0.5$ year^{-1}.

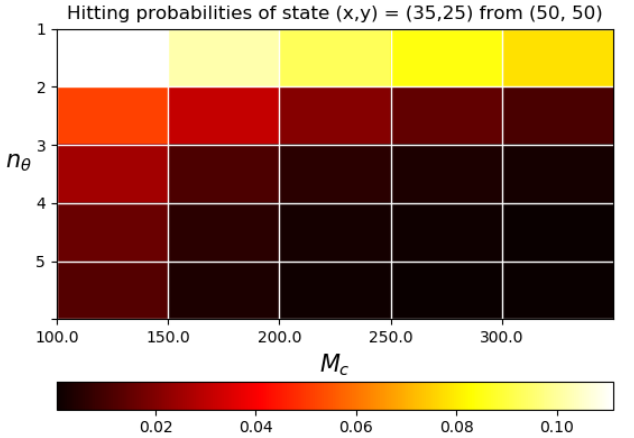

Figure 5.17: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both n_θ and M_c variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $\gamma = 1.25$ year^{-1} and $\mu = 0.5$ year^{-1}.

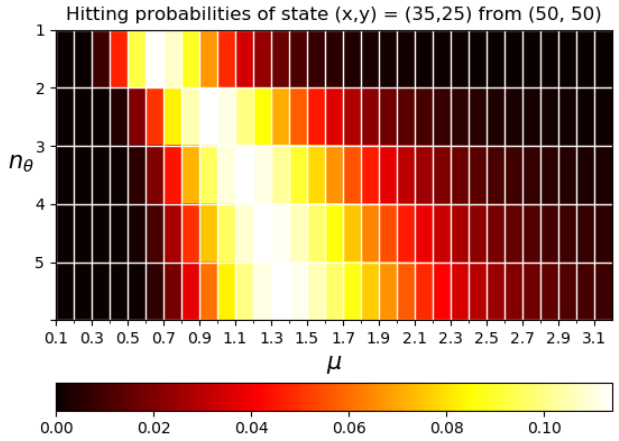

Figure 5.18: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both n_θ and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $M_c = 200$ and $\gamma = 1.25$ year^{-1}.

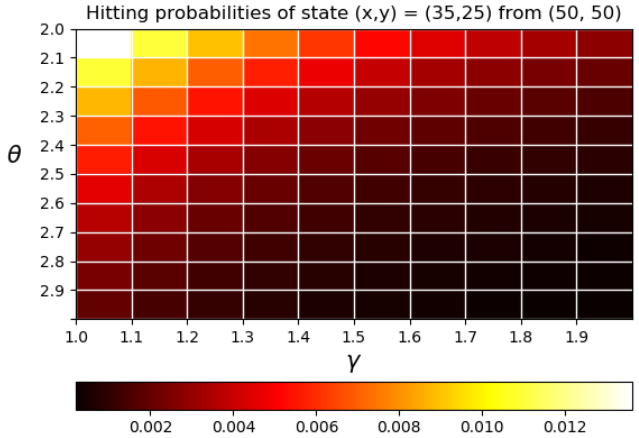

Figure 5.19: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both θ and γ variables. Fixed parameters are $n_\theta = 4$, $M_c = 200$ and $\mu = 0.5$ year^{-1}.

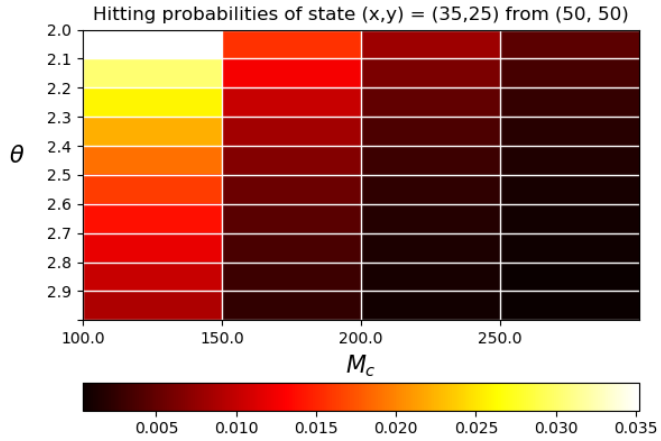

Figure 5.20: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both θ and M_c variables. Fixed parameters are $n_\theta = 4$, $\gamma = 1.25$ year^{-1} and $\mu = 0.5$ year^{-1}.

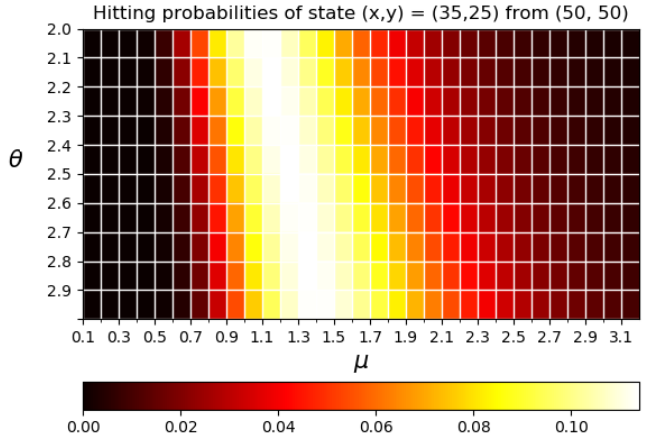

Figure 5.21: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both θ and μ variables. Fixed parameters are $n_\theta = 4$, $M_c = 200$ and $\gamma = 1.25$ year^{-1}.

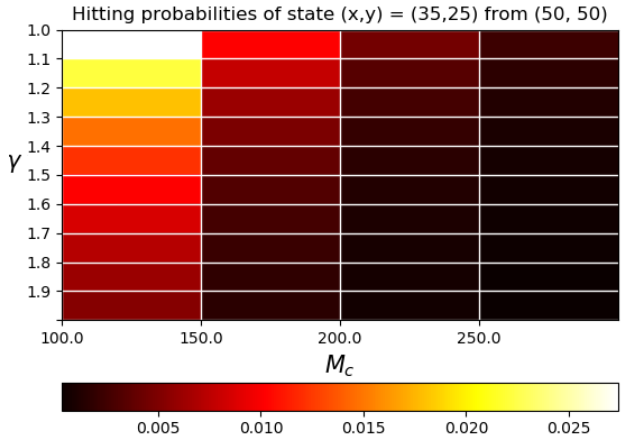

Figure 5.22: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both γ and M_c variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $\mu = 0.5$ year^{-1}.

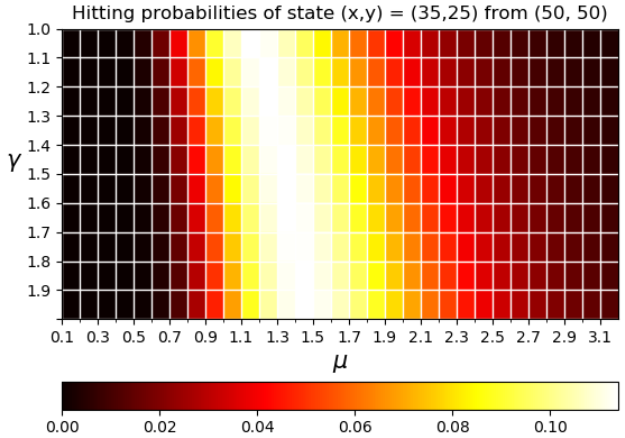

Figure 5.23: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both γ and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $M_c = 200$.

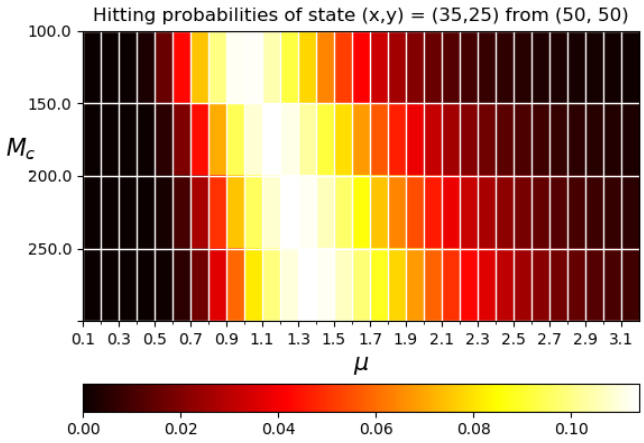

Figure 5.24: The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$ as a function of both M_c and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $\gamma = 1.25$ year^{-1}.

Maximum repertoire diversity in $[0, T_N(A)]$

We present here the results for the stochastic descriptor defining the maximum repertoire diversity in the time interval $[0, T_N(A)]$. Parameters are chosen in accordance to Section 5.7. Figure represents the solution of the system described in Section 5.6.1, that is the $Pr(X^{max} > \bar{x})$. The maximum diversity value $\bar{x} = 52$ to be exceeded, and the different initial states $\{(x, y) : x \in \{46, \ldots, 52\}, y \in \{46, \ldots, 50\}\}$ have been chosen just as examples. Figure 5.26 represents the $Pr(X^{max} = x)$, while Figure 5.27 represents its Gillespie counterpart. Sensitivity analyses are then presented in Figures 5.28 - 5.37.

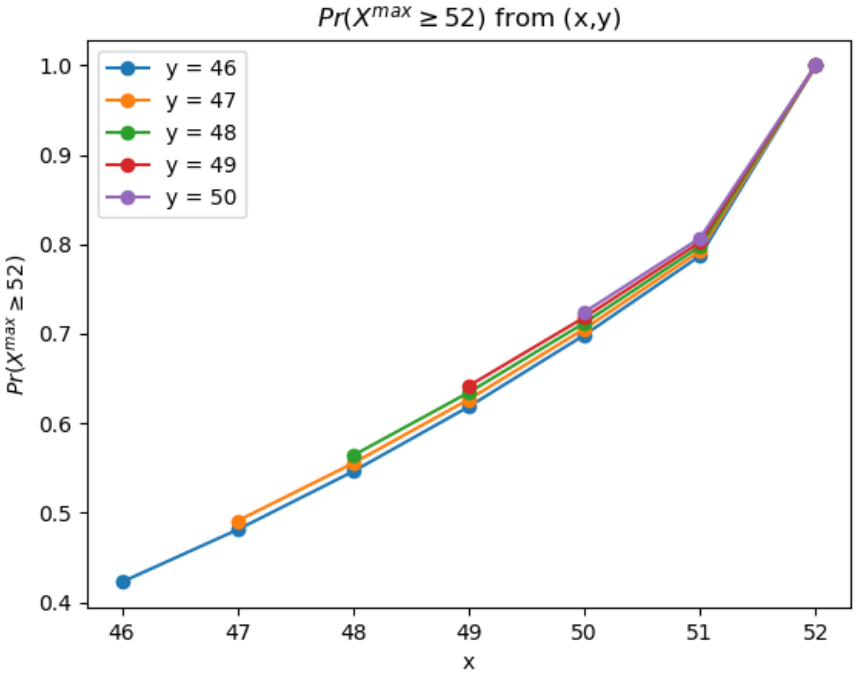

Figure 5.25: Parameters: $A = 25$, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $\gamma = 1.25$ year^{-1}, $M_c = 200$ and $\mu = 0.5$ year^{-1}. Different colours represent different values of y. The hitting probabilities, for a given y, are plotted as functions of x.

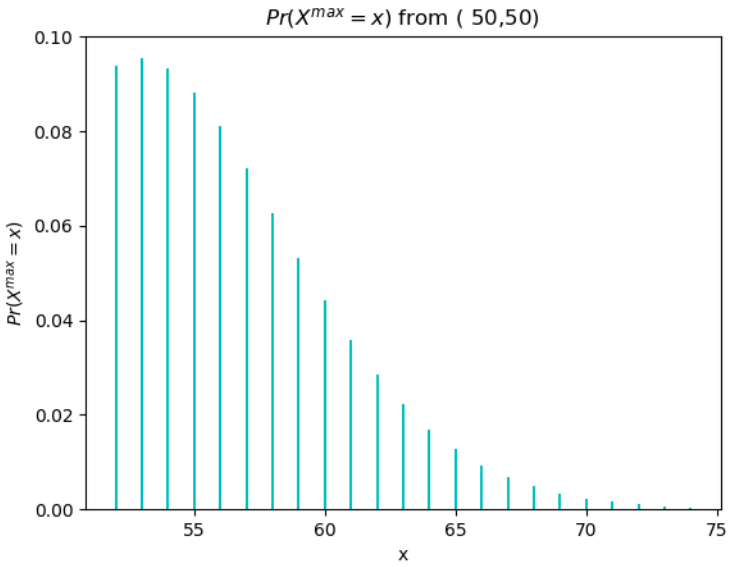

Figure 5.26: Parameters: $A = 25$, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $\gamma = 1.25$ year^{-1}, $M_c = 200$, $\mu = 0.5$ year^{-1} and initial state $(50, 50)$. The plot represents the probabilities for $X^{\max}$ being equal to x from the initial state $(50, 50)$.

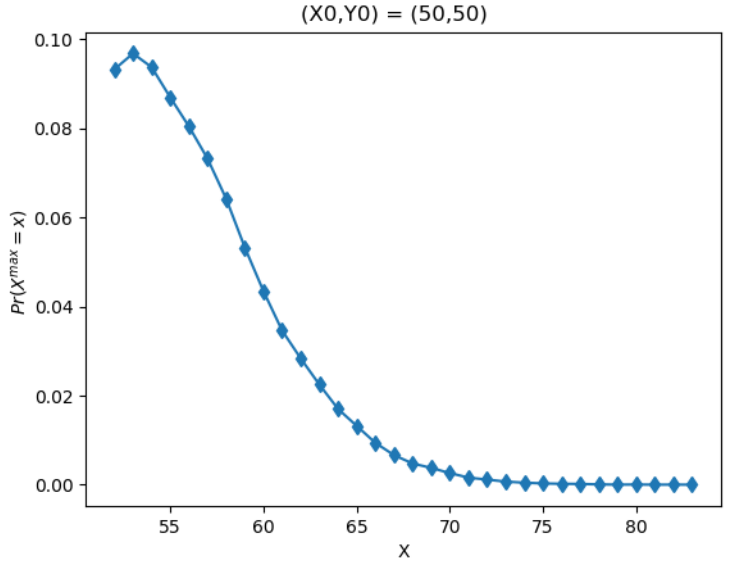

Figure 5.27: Parameters: $A = 25$, $\theta = 2.5$ year^{-1}, $n_\theta = 4$, $\gamma = 1.25$ year^{-1}, $M_c = 200$, $\mu = 0.5$ year^{-1} and initial state $(50, 50)$. The plot represents the probabilities (from Gillespie algorithm) for $X^{\max}$ being equal to x from the initial state $(50, 50)$. Number of simulations $= 10^5$.

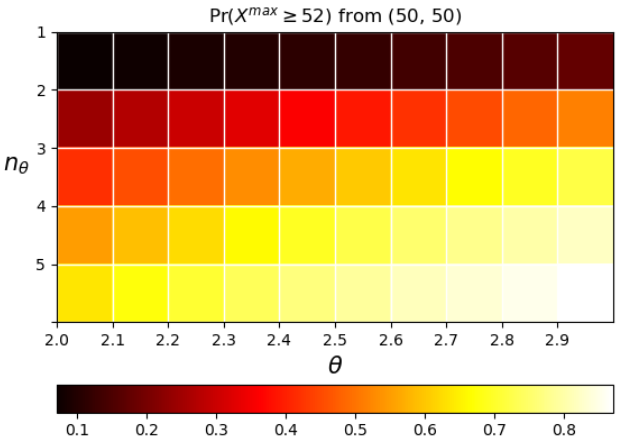

Figure 5.28: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both n_θ and θ variables. Fixed parameters are $M_c = 200$, $\mu = 0.5$ year^{-1} and $\gamma = 1.25$ year^{-1}.

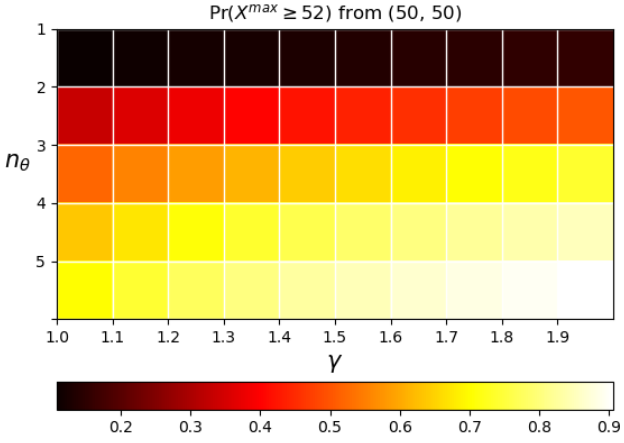

Figure 5.29: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both n_θ and γ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $M_c = 200$ and $\mu = 0.5$ year^{-1}.

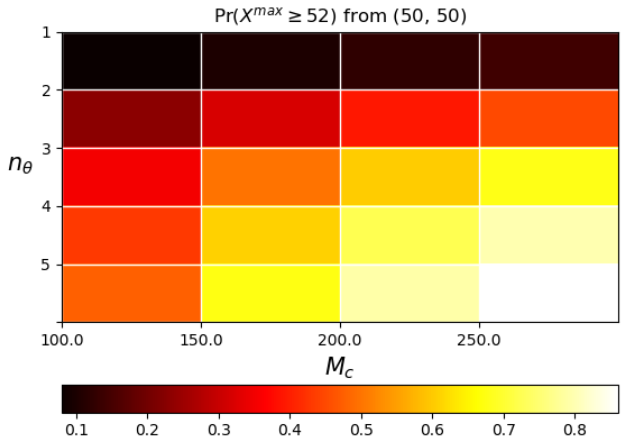

Figure 5.30: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both n_θ and M_c variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $\mu = 0.5$ year^{-1} and $\gamma = 1.25$ year^{-1}.

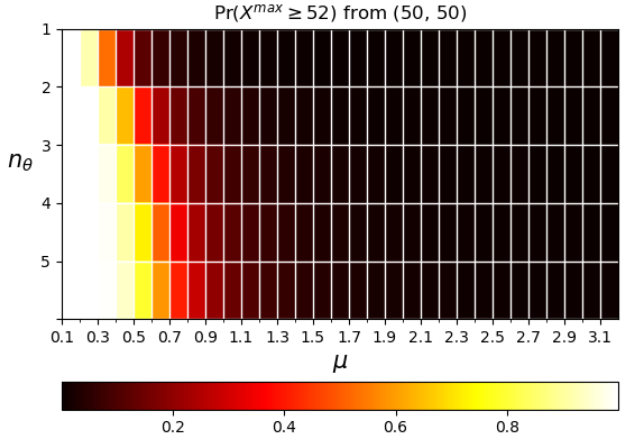

Figure 5.31: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both n_θ and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $M_c = 200$ and $\gamma = 1.25$ year^{-1}.

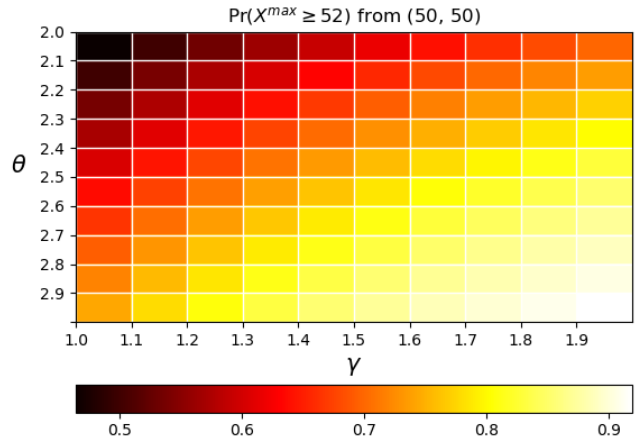

Figure 5.32: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both θ and γ variables. Fixed parameters are $n_\theta = 4$, $M_c = 200$ and $\mu = 0.5$ year^{-1}.

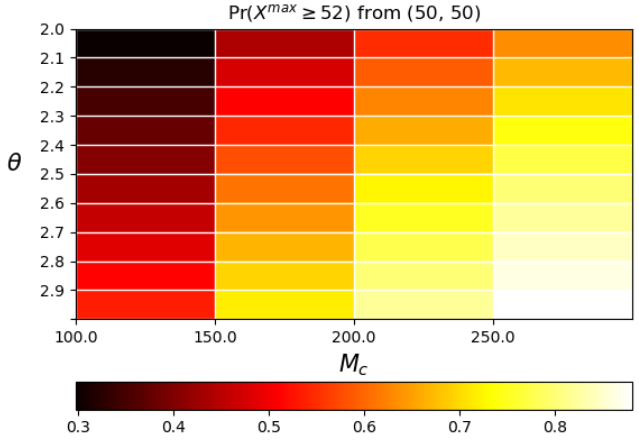

Figure 5.33: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both θ and M_c variables. Fixed parameters are $n_\theta = 4$, $\mu = 0.5$ year^{-1} and $\gamma = 1.25$ year^{-1}.

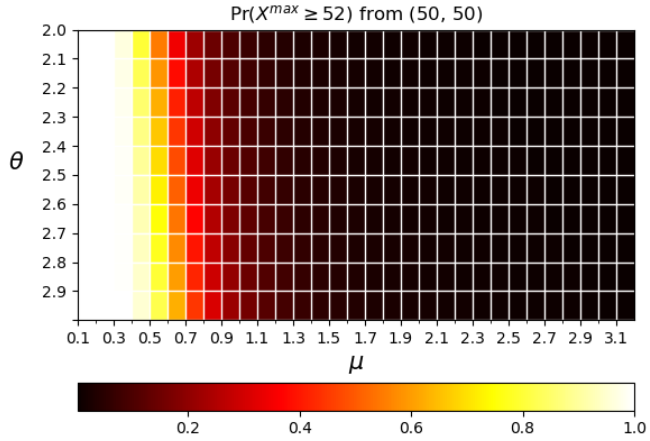

Figure 5.34: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both θ and μ variables. Fixed parameters are $n_\theta = 4$, $M_c = 200$ and $\gamma = 1.25$ year^{-1}.

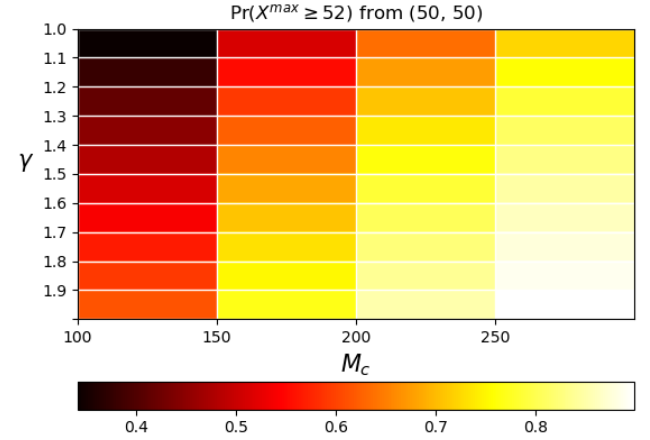

Figure 5.35: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both γ and M_c variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $\mu = 0.5$ year^{-1}.

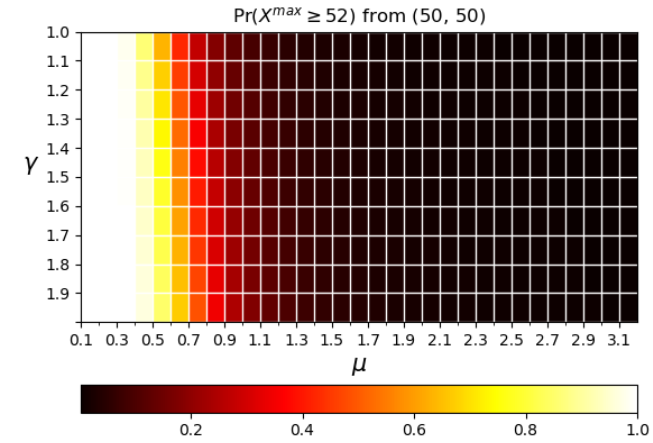

Figure 5.36: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both γ and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $M_c = 200$.

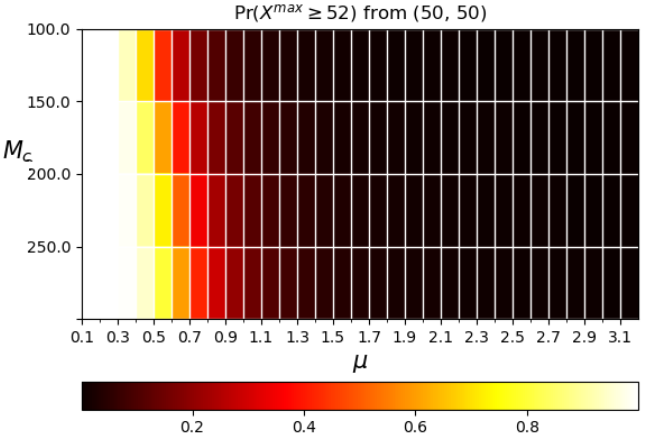

Figure 5.37: The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both M_c and μ variables. Fixed parameters are $\theta = 2.5$ year^{-1}, $n_\theta = 4$ and $\gamma = 1.25$ year^{-1}.

5.7.2 Explicit competition

This section presents the numerical results for the stochastic model with implicit competition, as described in Section 5.2.2. Results are shown for the three stochastic descriptors defining (i) the time $T_N(A)$ to reach for the first time a number $A < X(0)$ of original clonotypes in the repertoire, (ii) the size of the repertoire at time $T_N(A)$ and (iii) the maximum repertoire diversity in $[0, T_N(A)]$. Solutions of the first step argument methods of Sections 5.4.2, 5.5.2 and 5.6.2 are shown for different initial states, together with Gillespie simulations which confirm the correctness of the first step argument methods. Sensitivity analysis is also performed over the parameter β_1 and the product $p\beta_2$. The choice to vary the product $p\beta_2$ instead of the single parameter β_2 comes from the importance of the product itself, defined as the environmental pressure due to clonotype competition in Section 5.2.2.

Time $T_N(A)$ to reach for the first time a number $A < X(0)$ of original clonotypes in the repertoire

We present here the results for the stochastic descriptor defining the time $T_N(A)$ to reach for the first time a number $A < X(0)$ of original clonotypes in the repertoire. Parameters are chosen in accordance to Section 5.7. Figure 5.38 represents the solution of the system described in Section 5.4.2. The value $A = 25$, as well as the initial states $\{(x, y) : x \in \{48, \dots, 60\}, y \in \{47, \dots, 50\}\}$ have been chosen just as examples. Figure 5.39 represents the Gillespie counterpart of Figure 5.38, for the particular value $y = 50$. Sensitivity analyses are then presented in Figure 5.40.

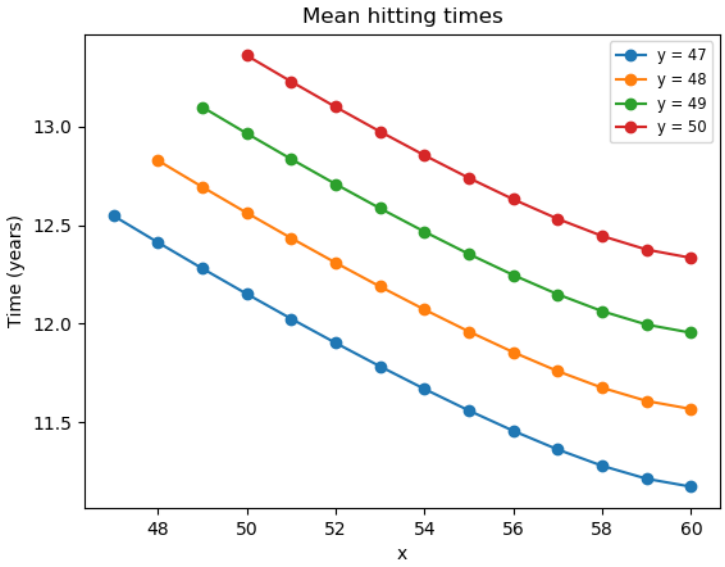

Figure 5.38: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting times of level A, that is one of the general states (x,25), from the different states (x, y). Different colours represent different values of y. The hitting probabilities, for a given y, are plotted as functions of x.

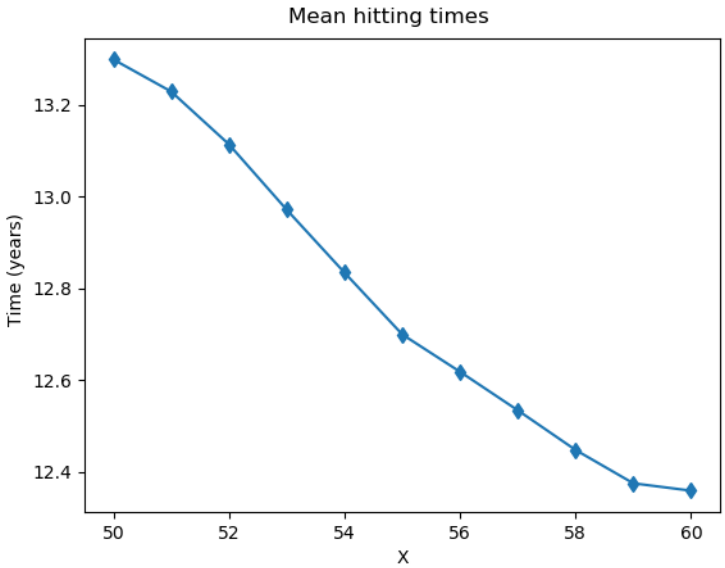

Figure 5.39: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting times (from Gillespie algorithm) of level A, that is one of the general states (x,25), from the initial state $(50, 50)$. Number of simulations $= 10^5$.

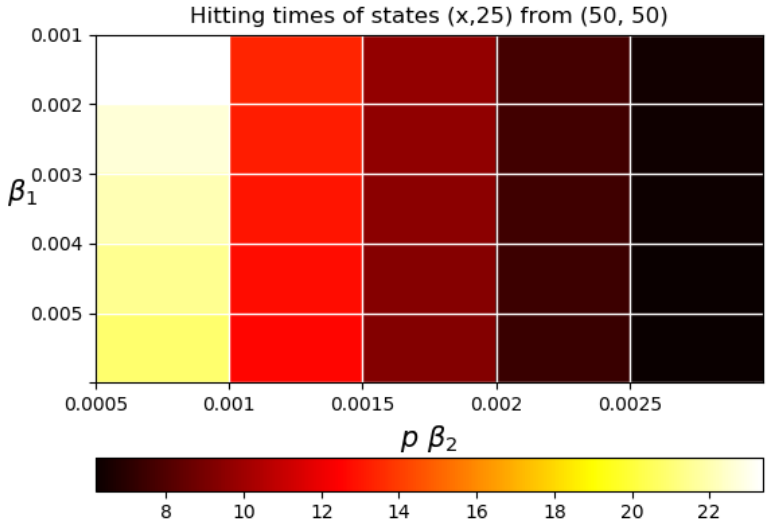

Figure 5.40: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting times of level A, that is one of the general states (x,25), from the initial state $(50, 50)$ as a function of both β_1 and $\beta_2 p$ variables.

Size of repertoire at time $T_N(A)$

We present here the results for the stochastic descriptor defining the size of repertoire at time $T_N(A)$. Parameters are chosen in accordance to Section 5.7. Figure 5.41 represents the solution of the system described in Section 5.5.2. The hitting state $(35, 25)$, as well as the initial states $\{(x, y) : x \in \{46, \dots, 60\}, y \in \{46 \dots, 50\}\}$ have been chosen just as examples. Figure 5.42 represents the Gillespie counterpart of Figure 5.41, for the particular value $y = 50$. Sensitivity analyses are then presented in Figure 5.43.

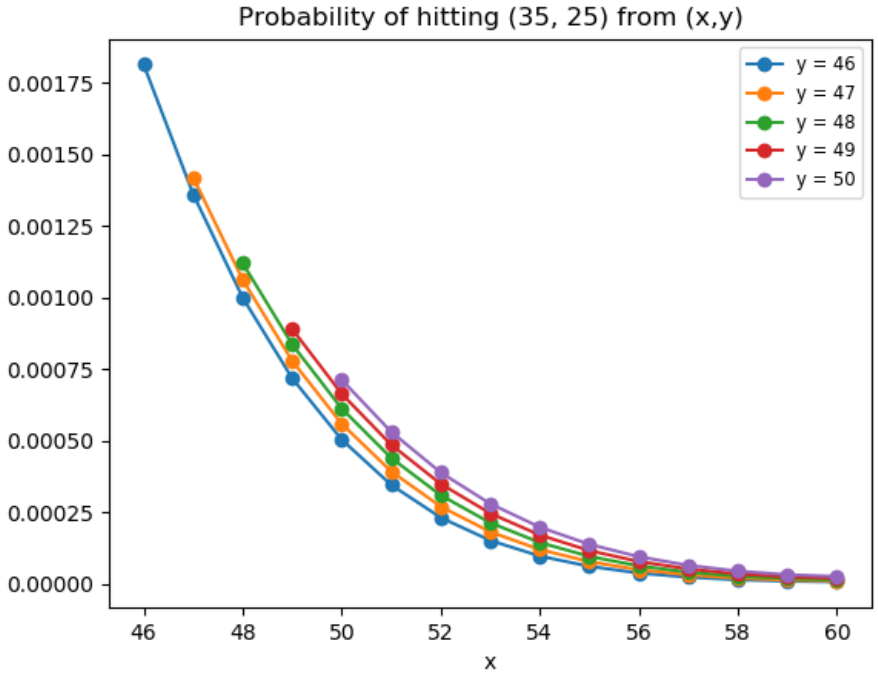

Figure 5.41: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting probabilities of state $(35, 25)$ from the different states (x, y). Different colours represent different values of y. The hitting probabilities, for a given y, are plotted as functions of x.

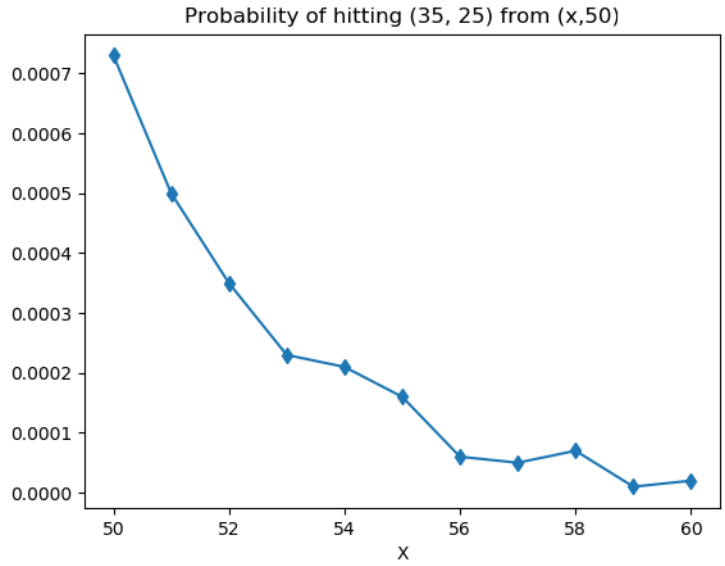

Figure 5.42: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting probabilities of state $(35, 25)$ from the initial state $(50, 50)$. Number of simulations $= 10^5$.

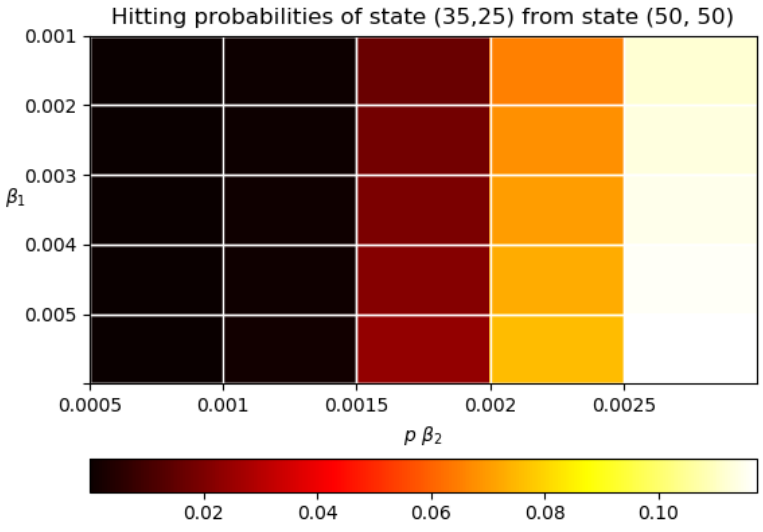

Figure 5.43: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the hitting probabilities of state $(50, 25)$ from the initial state $(50, 50)$ as a function of both β_1 and $\beta_2 p$ variables.

Maximum repertoire diversity in $[0, T_N(A)]$

We present here the results for the stochastic descriptor defining the maximum repertoire diversity in the time interval $[0, T_N(A)]$. Parameters are chosen in accordance to Section 5.7. Figure 5.44 represents the solution of the system described in Section 5.6.2, that is the $Pr(X^{max} > \bar{x})$. The maximum diversity value $\bar{x} = 52$ to be exceeded, and the different initial states $\{(x, y) : x \in \{46, \dots, 52\}, y \in \{46, \dots, 50\}\}$ have been chosen just as examples. Figure 5.45 represents the $Pr(X^{max} = x)$, while Figure 5.46 represents its Gillespie counterpart. Sensitivity analyses are then presented in Figure 5.47.

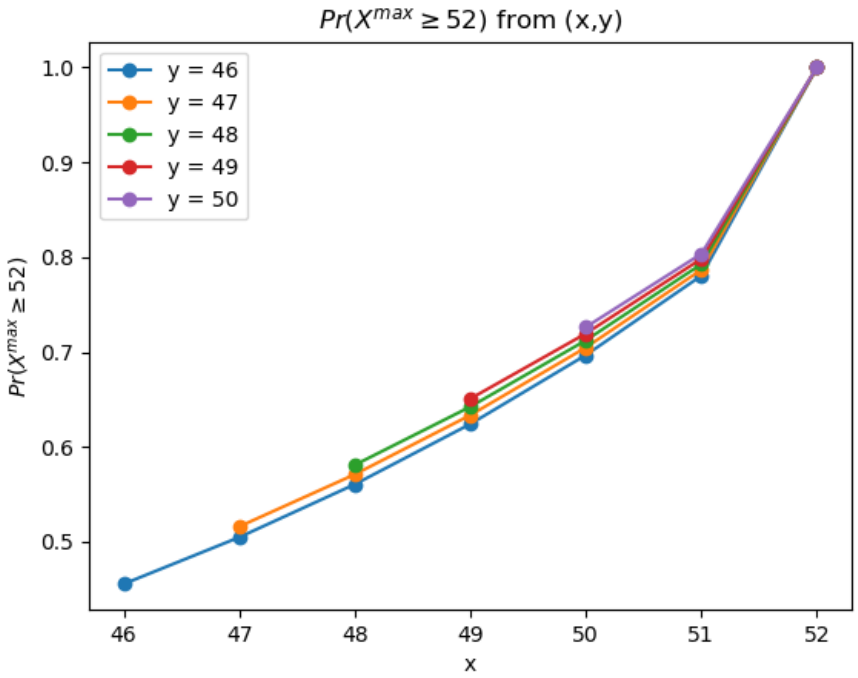

Figure 5.44: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. Different colours represent different values of y. The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state (x, y).

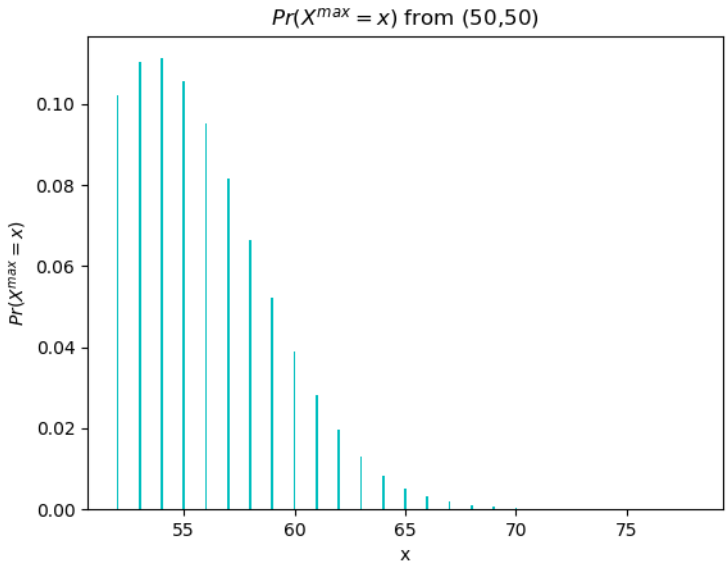

Figure 5.45: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the probabilities for $X^{\max}$ being equal to x from the initial state $(50, 50)$.

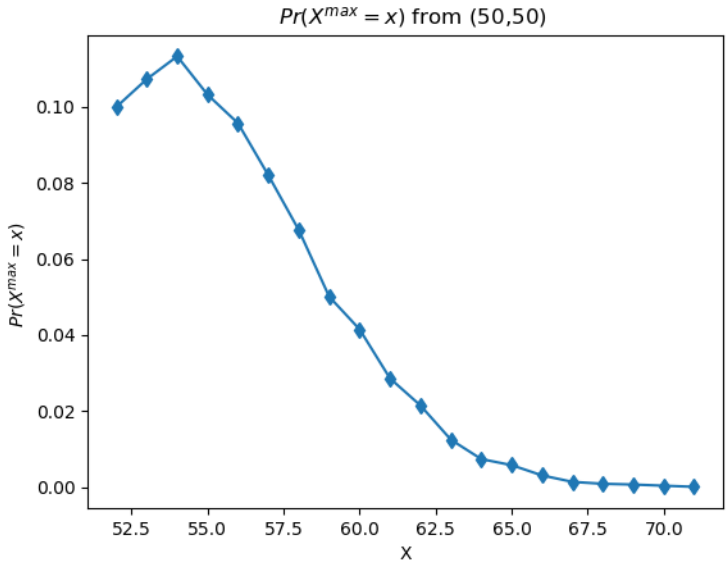

Figure 5.46: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the probabilities (Gillespie simulations) for $X^{\max}$ being equal to x from the initial state $(50, 50)$. Number of simulations $= 10^5$.

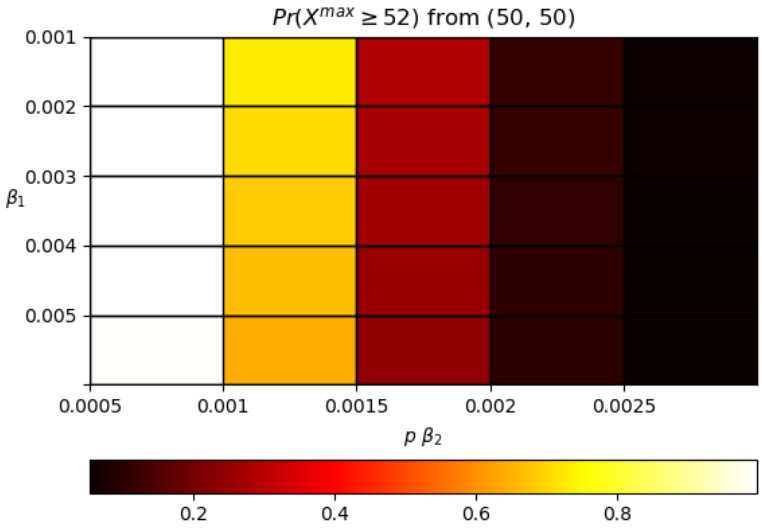

Figure 5.47: Parameters: $A = 25$, $M = 60$, $\theta = 2.5$ year^{-1}, $\beta_1 = 0.004$ year^{-1}, $\beta_2 = 0.02$ year^{-1} and $p = 0.05$. The plot represents the probabilities for $X^{\max}$ being greater or equal to 52 from the initial state $(50, 50)$, as a function of both β_1 and $\beta_2 p$ variables.

5.8 Discussion

This chapter focused on the biological problem concerning the evolution of diversity in a $CD4^+$ T-cell repertoire. In particular, we follow the dynamics of both original and new clonotype classes. The study of these dynamics is of foremost importance to evaluate the relevance of sampling from a repertoire and trying to estimate its diversity at a given point in time. To this aim, we built a continuous-time Markov chain (CTMC) representing the stochastic processes of competition and natural death of the original and new TCR clonotypes. We then decided to further consider two different types of competitions: implicit and explicit competition. The implicit case considers a death rate $\mu_n^{(1)} = \tilde{\mu}n$, where n represents the number of different clonotype classes, and $\tilde{\mu}$ represents the average survival time of a clonotype in the repertoire, incorporating competition within its very definition, as explained in Section 5.2.1. The explicit case considers a death rate $\mu_n^{(2)} = n(\beta_1+\beta_2 p(n-1))$, where n represents the number of different clonotype classes, p represents the probability that any given self pMHC is recognised by a randomly-selected TCR clonotype, β_1 represents the average survival time of a clonotype in the repertoire, and β_2 represents the strength of competition among clonotypes competing for the same self pMHCs, as explained in Section 5.2.2. Three different stochastic descriptors were studied: (i) the time $T_N(A)$ needed for the N original clonotype classes to become A (with $A < N$), (ii) the size of the repertoire at time $T_N(A)$, and (iii) the maximum repertoire diversity achieved in the time interval $[0, T_N(A)]$.

The first descriptor was analysed analytically for the implicit case, and its density function was computed, together with the moment generating function. The results were seen to be very similar to those from the explicit case, which was analysed by the first step argument analysis. Figure 5.11 and Figures 5.38–5.39 show the average time until half of the original clonotype classes are lost, starting with $N(0) = 50$ clonotypes. The time is around 13.5 years for both implicit and explicit cases.

The second descriptor was analysed with the first step argument technique for both implicit and explicit cases, showing slightly different results. Figures 5.13–5.14 and 5.41–5.42 show the probability of hitting the state $(35, 25)$ from the state $(50, 50)$. The probabilities are 0.0025 and 0.00075 for the implicit and explicit case respectively.

The third descriptor was also analysed with the first step argument technique for both implicit and explicit cases, showing very similar results. Figures 5.25 and 5.44 show the probability for the maximum number of distinct clonotypes to be greater than 52, starting with $N(0) = 50$ clonotypes. The probability is around 0.75 for both cases.

Small-scale simulations are of course far from reality, representing a simplified version of the dynamics of the homeostasis conditions of the immune system. Nevertheless, we believe these simulations carry two important results. First, they show the power of the

first step argument for the analysis of stochastic descriptors which cannot be analytically studied. This is proven by the accuracy of the analyses compared with the Gillespie simulations. Second, these results show that within a life-time period, the repertoire of an individual could easily evolve in a way that it could become very different from the starting one. This enhances the importance of our first question: up to what point does it really make sense to have an estimate of the actual repertoire diversity of an individual at a specific point in time?

Chapter 6

Conclusions

The immune system is a complex machine, whose mechanisms are indispensable to life itself. Its outstanding power is given by the unique ability of recognising virtually almost all the possible existing pathogens. This singular characteristic is maintained, among others, thanks to an incredibly broad army of distinct T-cell classes, characterised by different T-cell receptors (TCRs), and able to recognise different overlapping groups of pathogens. Understanding the biological mechanisms behind the creation of the immune system diversity is one of the central aspects of current research in immunobiology.

The present work tried to shed some light on different facets of this crucial aspect, producing some interesting quantitative results and suggesting future research areas in mathematical immunology. Data analysis and stochastic models were the main drivers of the thesis, together with statistical and probabilistic analysis.

Starting from the probabilistic side of the subject, Chapter 3 studied the relation between the total (unmeasurable) TCR repertoire diversity, and the observable diversity, shown by small biological samples. The probability generating function (PGF) of the distribution of the observed clonal sizes is proven to be the composition of two PGFs of two distinct random variables: (i) a Bernoulli random variable and (ii) the random variable of the true clonal size distribution in the repertoire that is being sampled from. In particular, different clonal size distributions were studied and their sample distributions were analysed. The expected number of TCR repeats in a sample, and the number of draws to find the first repeat are also studied, and analytical formulae are given. These techniques were then used to estimate the clonal size distribution of a subset of the mice immune system ($GP33^+$ repertoire), following different distribution hypotheses.

Chapter 4 targeted the TCR diversity problem from a data analysis point of view, focusing on experimental data gathered from a broader study on TCR diversity in naïve and LCMV (lymphocytic choriomeningitis virus) infected mice. The analyses centered on the distributions of the V and J gene segments of the TCR β chains. Statistical tests such as

Wilcoxon-Mann-Whitney U test, Randomization test, and χ^2 test were conducted. Simpson's diversity index, which was used to quantify genetic diversity among mice, resulted to be statistically higher in the naïve mice than in the infected ones. Though, this result only concerned the diversity at the V-J pair level, and was not found for the single V or J diversities. The clonal sharing was also studied, trying to distinguish public from private VJ classes, in both naïve and infected mice. The chapter continued with the development of a mathematical relation among the observed (g_i) and unobserved (f_i) frequency of a given TCR clonotype class (i) of the repertoire. The chapter is concluded with a quick application of this relation to some of the data taken into consideration: the frequency of the V_1 gene segment is suggested to be around 5.47% in the naïve repertoire of a mice.

Chapter 5 focused on the evolution of diversity in a $CD4^+$ T-cell repertoire. A continuous-time Markov chain was built, to follow the dynamics of TCR clonotype classes of a simulated repertoire. Two different cases for inter clonal competition were considered: implicit and explicit competition. The first case considered the competition intrinsically defined into the constant $\tilde{\mu}$ within the definition of the death rate of each clonotype class ($\mu_n^{(1)} = \tilde{\mu}n$), where n represented the number of different clonotype classes in the repertoire. The second case considers the competition in an explicit way, through a second death rate $\mu_n^{(2)} = n(\beta_1 + \beta_2 p(n-1))$. Both cases were analysed and three different stochastic descriptors were taken into consideration: (i) the time $T_N(A)$ needed for the N original clonotype classes to become A (with $A < N$), (ii) the size of the repertoire at time $T_N(A)$, and (iii) the maximum repertoire diversity achieved in the time interval $[0, T_N(A)]$. Small-scale simulations were displayed, showing strong similarities between the two cases for both the first and third stochastic descriptor. The driving question of the chapter was about the importance of a time-point estimate of the actual TCR diversity, taking into consideration the renewal process. The simulations, although representing an overly simplified reality, suggest the existance of a renewal process that should be taken into consideration when studying the TCR repertoire diversity.

The study of TCR repertoire diversity is of foremost importance for immunology today. Its complexity is such that even the smallest step forward in knowledge could have incredibly positive repercussions for human and animal health. Difficulties are found at each level of the research, from genetics up to population studies, from the biological to the modelling point of view. This work tackled some of these problems from a quantitative perspective, and the author is well aware that the path to a final understanding of TCR diversity remains an uphill battle. With this in mind, the author looks at this thesis as a small step on an ambitious path, with the hope that somehow, someday science will eventually unravel this unbelievable mystery of the immune system.

Appendix A

Binomial approximation of hypergeometric distribution

Let's define D_i as the clonotype class i for $i \in 1, 2, ..., N$. Thus the repertoire will be $D = \bigcup_{i=1}^{N} D_i$. Let n_i be the number of T cells in the clonotype class i, that is $n_i = |D_i|$. Let $S = \sum_{i=1}^{N} n_i$ be the total number of T cells in the repertoire. Now let us extract a sample $X = (X_1, X_2, ..., X_m)$ of size m, where X_j is the j^{th} extracted T cell. Define Y_i as the number of T cells of type i in the sample X, for $i \in 1, 2, ..., N$. Note that $\sum_{i=1}^{N} Y_i = m$. The distribution of $(Y_1, Y_2, ..., Y_N)$ is called multivariate hypergeometric distribution with parameters $(S, (n_1, ..., n_N), m)$. In particular, the marginal distributions are described by

$$\Pr(Y_i = y) = \frac{\binom{n_i}{y}\ \binom{S-n_i}{m-y}}{\binom{S}{m}} \quad \text{for } y \in 0, 1, ..., m. \tag{A.1}$$

Defining $\Pr(Y_i = y)$ as p_i, we can rewrite this equation as

$$\begin{aligned} p_i &= \binom{n_i}{y} \frac{m!}{(m-y)!} \frac{(S-m)!}{S!} \frac{(S-n_i)!}{(S-n_i-(m-y))!} \\ &= \binom{n_i}{y} \frac{m(m-1)\cdots(m-(y-1))}{S(S-1)\cdots(S-(m-1))} \prod_{k=0}^{m-y-1} (S-n_i-k) \end{aligned} \tag{A.2}$$

We now assume $n_i \ll m$, which in turn implies $y \ll m$, obtaining

$$p_i \approx \binom{n_i}{y} \frac{m^y}{S^y S^{m-y}} \prod_{k=0}^{m} (S-n_i-k). \tag{A.3}$$

We also assume that $n_i \ll S$, therefore approximating p_i as

$$p_i \approx \binom{n_i}{y} q^y \frac{1}{S^{m-y}} (S)(S-1)(S-2)\cdots(S-m). \tag{A.4}$$

We now realize that (A.4) can be rewritten as

$$p_i \approx \binom{n_i}{y} q^y \frac{1}{S^{m-y}} \, (S-m+m)(S-m+(m-1))(S-m+(m-2)) \cdots (S-m)), \quad \text{(A.5)}$$

and then approximated by

$$p_i \approx \binom{n_i}{y} q^y \frac{1}{S^{m-y}} \, \left((S-m)+\frac{m}{2}\right)^{m-y}. \quad \text{(A.6)}$$

Assuming $m \ll S$, we obtain

$$p_i \approx \binom{n_i}{y} q^y \frac{(S-m)^{m-y}}{S^{m-y}} = \binom{n_i}{y} q^y (1-q)^{m-y}. \quad \text{(A.7)}$$

Appendix B

V-J data

Tables B.1-B.10 show, for each individual mouse, the number of T cells with a particular combination of V and J genes.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	1	3	1	0	0	0	0	1	1	0	0	3
V_2	2	2	0	0	1	0	1	0	2	0	0	1
V_3	2	0	1	0	0	0	0	1	1	0	2	2
V_4	0	0	0	0	0	0	0	0	3	1	0	3
V_5	2	0	1	1	2	0	0	0	0	1	0	0
V_{12-1}	0	2	2	0	0	0	1	1	1	0	0	2
V_{12-2}	0	1	0	0	0	0	1	0	0	0	4	4
V_{13-1}	2	4	0	3	1	2	4	2	2	2	2	5
V_{13-2}	4	3	0	3	0	1	6	1	3	6	7	8
V_{13-3}	0	2	3	4	0	2	6	2	5	3	6	7
V_{14}	2	0	0	1	0	0	3	0	1	0	1	1
V_{15}	3	0	0	0	0	0	2	1	0	0	0	0
V_{16}	1	2	1	1	0	0	3	0	0	2	0	7
V_{17}	0	0	0	0	0	0	0	0	0	1	1	2
V_{19}	0	0	0	0	0	2	0	1	1	0	4	4
V_{20}	2	2	0	0	0	0	1	1	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	1	0	0	0	1	0	3	0	0	1	1	3
V_{29}	2	0	0	0	0	0	2	1	0	2	2	1
V_{30}	0	0	0	0	0	0	0	0	0	0	1	0
V_{31}	0	0	0	0	0	0	0	0	0	0	3	3

Table B.1: V-J distribution of the 253 T cells of mouse BA1.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	2	0	2	1	0	2	2	0	0
V_2	0	1	0	0	1	0	1	0	1	1	3	0
V_3	0	1	0	2	0	0	0	0	0	0	0	0
V_4	0	1	1	0	0	0	1	0	1	0	0	1
V_5	0	0	0	1	0	0	2	0	0	2	1	3
V_{12-1}	0	4	1	0	0	0	1	0	0	1	3	3
V_{12-2}	0	1	0	0	0	0	2	1	0	0	0	1
V_{13-1}	3	2	0	0	0	2	3	2	0	3	2	1
V_{13-2}	3	2	0	0	0	2	0	1	1	1	2	5
V_{13-3}	2	0	0	0	0	1	2	0	1	1	3	2
V_{14}	0	1	0	0	0	0	2	0	1	0	4	3
V_{15}	1	0	0	0	0	0	1	1	0	0	1	2
V_{16}	1	0	0	3	0	0	1	2	1	1	1	1
V_{17}	0	0	0	0	0	0	0	1	1	0	3	4
V_{19}	3	0	1	1	0	0	1	1	1	0	0	4
V_{20}	0	0	0	0	0	0	0	0	0	2	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	1	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	1	0	0	0	1	0
V_{26}	0	1	0	0	0	0	0	0	0	0	0	1
V_{29}	1	2	0	1	0	1	1	1	1	1	0	0
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	1	0	0	0	1	0	0

Table B.2: V-J distribution of the 166 T cells of mouse BA2.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	3	1	0	0	0	0	2	0	0	1	1	1
V_2	0	0	0	0	0	0	0	0	0	0	0	3
V_3	0	0	0	0	0	0	1	2	1	0	0	3
V_4	0	0	1	1	0	0	0	2	0	2	0	1
V_5	1	0	0	1	0	0	4	1	1	1	1	2
V_{12-1}	0	0	1	0	0	0	2	0	0	1	0	0
V_{12-2}	0	0	1	0	0	0	1	0	0	0	1	2
V_{13-1}	5	3	1	1	2	0	6	3	2	2	5	6
V_{13-2}	7	4	2	1	0	1	8	1	8	6	7	12
V_{13-3}	5	5	0	0	1	1	5	3	8	5	8	13
V_{14}	2	0	1	0	0	0	2	0	1	2	2	4
V_{15}	0	1	0	0	0	0	0	1	0	0	0	0
V_{16}	1	0	0	1	1	0	3	0	3	5	2	0
V_{17}	0	0	1	0	0	0	2	1	0	1	1	1
V_{19}	10	0	0	2	0	0	1	0	0	1	2	3
V_{20}	0	0	0	0	0	0	0	0	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	1	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	0	0	1	1	0	0	0	1
V_{29}	3	0	0	0	0	0	2	0	1	5	4	2
V_{30}	0	0	0	0	0	0	0	0	0	0	0	1
V_{31}	0	0	0	0	0	1	0	0	0	0	0	0

Table B.3: V-J distribution of the 275 T cells of mouse BA3.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	2	0	2	3	0	1	1	0	2	2	0	1
V_2	0	0	0	0	0	0	1	0	2	1	1	0
V_3	1	0	0	0	0	0	4	0	0	1	1	3
V_4	0	0	0	1	0	2	0	0	0	1	0	3
V_5	1	0	0	2	0	1	0	0	0	0	2	2
V_{12-1}	0	2	0	0	0	0	1	1	1	0	1	3
V_{12-2}	1	0	0	0	1	0	2	1	1	0	2	0
V_{13-1}	2	0	1	1	0	0	1	0	1	3	2	6
V_{13-2}	1	2	0	0	1	0	2	3	0	4	3	5
V_{13-3}	1	2	1	2	0	3	3	0	1	2	3	3
V_{14}	0	0	1	1	0	0	1	0	0	2	0	2
V_{15}	0	0	0	0	0	0	1	0	0	1	0	3
V_{16}	0	2	1	0	0	0	3	1	5	3	4	3
V_{17}	0	0	0	0	0	0	0	2	0	0	1	0
V_{19}	2	1	1	0	1	1	3	0	1	1	1	3
V_{20}	0	1	0	0	0	0	2	0	1	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	1	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	1	0	0	0	0	0	0	0	1	1	0	1
V_{29}	2	2	0	0	2	0	1	0	0	1	0	3
V_{30}	0	0	0	0	1	0	0	0	0	0	0	0
V_{31}	1	0	0	0	0	1	0	0	0	0	0	0

Table B.4: V-J distribution of the 195 T cells of mouse BA4.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	1	0	0	0	0	1	3	2	0	0	2	1
V_2	0	0	0	0	0	0	0	0	0	0	1	1
V_3	0	0	0	2	0	0	1	0	1	0	0	0
V_4	1	0	0	0	0	0	1	0	0	1	0	0
V_5	1	0	0	1	0	0	2	1	0	0	0	1
V_{12-1}	0	1	1	1	0	0	1	2	0	2	0	1
V_{12-2}	0	0	0	0	0	0	0	0	1	1	3	1
V_{13-1}	4	0	0	0	1	0	4	1	1	1	0	3
V_{13-2}	1	1	0	0	1	0	2	3	0	4	4	3
V_{13-3}	3	1	0	1	0	1	2	1	0	1	2	3
V_{14}	2	0	0	1	0	0	4	0	2	0	0	2
V_{15}	1	0	1	1	1	0	0	0	0	0	0	0
V_{16}	2	2	1	1	0	0	0	0	1	0	2	1
V_{17}	1	0	0	1	0	0	0	0	0	0	0	0
V_{19}	0	0	0	1	1	0	0	1	1	0	0	2
V_{20}	0	1	0	0	0	0	0	0	0	0	1	1
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	1	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	1	0	1	0	0	0	0	0
V_{29}	0	0	0	0	0	0	1	0	0	0	0	1
V_{30}	0	0	0	0	0	0	1	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	1	1

Table B.5: V-J distribution of the 133 T cells of mouse BA5.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	0	0	0	1	0	0	1	3	0
V_2	0	0	0	0	0	0	1	2	0	0	4	0
V_3	0	0	1	1	1	0	3	0	0	0	6	0
V_4	2	1	0	0	0	0	0	2	1	0	3	1
V_5	1	1	1	0	1	1	0	1	4	0	0	0
V_{12-1}	0	1	0	1	0	1	2	0	0	1	0	0
V_{12-2}	0	0	0	0	0	1	0	0	1	2	2	2
V_{13-1}	13	0	0	1	0	1	4	0	3	0	4	3
V_{13-2}	0	2	1	1	0	0	3	1	3	2	1	6
V_{13-3}	1	0	0	0	0	3	0	0	0	1	1	3
V_{14}	0	0	1	0	4	0	1	0	1	0	1	8
V_{15}	0	0	0	0	0	0	2	0	1	0	0	3
V_{16}	9	6	0	0	0	1	3	0	1	4	2	1
V_{17}	4	1	1	1	0	0	4	2	1	1	2	4
V_{19}	2	1	0	1	0	2	4	0	0	1	1	2
V_{20}	0	1	0	0	0	0	0	0	0	0	0	1
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	1	0	0	0	0	0	1	1	0	0	0	3
V_{29}	1	1	3	0	2	2	0	5	0	7	2	0
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	1	0	0	0	0	1	0	0	0	1	0

Table B.6: V-J distribution of the 234 T cells of mouse EF1.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	1	0	0	0	0	1	0	0	2	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	0	0	0	1	0	3	0	0	0	0	0	0
V_4	0	0	0	0	0	0	0	0	0	1	0	0
V_5	1	0	0	0	0	0	0	0	0	0	0	0
V_{12-1}	0	0	0	0	0	0	1	0	0	0	1	0
V_{12-2}	0	0	0	0	0	1	2	0	0	0	0	1
V_{13-1}	0	0	0	0	2	1	3	0	1	0	2	1
V_{13-2}	2	0	0	1	0	0	0	0	0	0	1	1
V_{13-3}	0	0	0	0	0	1	2	1	0	1	0	0
V_{14}	0	0	0	0	0	0	2	0	0	0	0	0
V_{15}	0	0	0	0	1	0	0	0	1	0	0	0
V_{16}	0	0	0	0	0	0	1	0	3	0	0	4
V_{17}	1	0	0	1	0	0	1	0	1	0	0	2
V_{19}	1	0	0	2	0	2	0	0	1	0	1	0
V_{20}	0	0	0	0	0	0	0	0	0	0	1	1
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	2	0	0	0	0	0	0	0	0
V_{29}	1	0	3	1	0	1	0	0	0	2	0	1
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	1	0	0	0

Table B.7: V-J distribution of the 75 T cells of mouse EF2.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	0	0	0	0	0	0	0	0	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	0	0	0	0	0	0	1	0	0	0	0	0
V_4	3	0	0	0	0	0	2	0	0	0	0	11
V_5	0	0	0	0	0	0	1	0	0	0	0	0
V_{12-1}	0	0	0	1	0	1	1	0	0	0	0	0
V_{12-2}	0	0	0	0	7	0	0	8	2	1	0	0
V_{13-1}	20	0	1	0	0	0	0	0	0	1	0	0
V_{13-2}	0	0	0	0	0	0	1	0	0	0	0	0
V_{13-3}	10	0	0	0	0	1	2	0	0	0	0	2
V_{14}	0	0	0	0	0	0	1	0	0	0	0	1
V_{15}	0	0	0	0	0	0	0	0	0	0	0	0
V_{16}	0	1	0	0	0	2	13	0	0	0	2	6
V_{17}	1	0	0	0	0	0	3	0	0	0	0	0
V_{19}	2	0	0	0	0	0	2	0	0	0	0	5
V_{20}	0	0	0	0	0	0	0	0	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	0	0	0	0	0	0	0	0
V_{29}	7	13	0	25	1	0	23	3	0	5	41	1
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	23	0

Table B.8: V-J distribution of the 258 T cells of mouse EF3.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	1	0	0	0	0	0	0	3	0	0	1
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	2	0	0	0	0	0	0	0	0	0	1	1
V_4	0	0	0	0	0	0	0	0	0	0	0	2
V_5	0	0	0	0	0	0	0	0	1	1	2	1
V_{12-1}	1	0	0	1	0	0	1	2	1	0	1	1
V_{12-2}	2	3	0	0	0	0	1	0	0	1	12	3
V_{13-1}	13	9	0	1	0	0	1	1	2	3	2	3
V_{13-2}	3	0	2	1	0	1	4	0	8	1	2	0
V_{13-3}	12	4	0	1	4	0	1	1	3	2	1	0
V_{14}	0	0	0	1	0	2	0	0	2	6	0	1
V_{15}	0	0	0	0	1	0	0	0	0	5	3	0
V_{16}	0	0	0	0	0	0	0	0	2	0	0	7
V_{17}	1	0	1	0	1	1	0	0	0	3	6	2
V_{19}	4	0	3	2	0	3	0	2	5	4	0	10
V_{20}	2	6	1	1	0	0	1	0	0	0	0	1
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	1	0	0	0	1	1	0	0	0
V_{29}	4	1	3	5	1	0	3	3	2	0	10	1
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	1	0	0	0	0	0	0	0	0	0	0

Table B.9: V-J distribution of the 259 T cells of mouse EF4.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	1	0	0	0	0	0	1	0	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	1	0	0	0	0	0	0	0	1	0	1	0
V_4	0	0	0	0	0	0	0	0	0	0	0	1
V_5	0	0	0	0	0	1	1	0	0	0	0	0
V_{12-1}	0	0	0	0	1	0	3	0	0	0	0	0
V_{12-2}	0	0	0	1	0	0	0	1	0	2	2	0
V_{13-1}	15	1	0	0	0	3	2	0	0	1	0	1
V_{13-2}	5	0	0	0	0	1	3	0	1	0	0	3
V_{13-3}	7	3	0	0	0	0	1	0	0	0	0	10
V_{14}	0	0	0	0	0	0	1	0	0	0	0	4
V_{15}	0	0	0	0	0	0	1	0	1	0	7	1
V_{16}	0	1	0	0	0	0	2	0	2	1	1	55
V_{17}	0	0	0	0	0	0	1	1	0	0	0	0
V_{19}	1	1	4	0	2	0	12	0	0	0	1	4
V_{20}	0	0	0	0	0	0	0	1	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	3	0	0	0	0	0	0	0	1
V_{29}	3	2	0	0	0	0	1	1	4	1	11	2
V_{30}	0	0	0	0	0	0	0	1	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	1	1

Table B.10: V-J distribution of the 212 T cells of mouse EF5.

VJ frequency plots

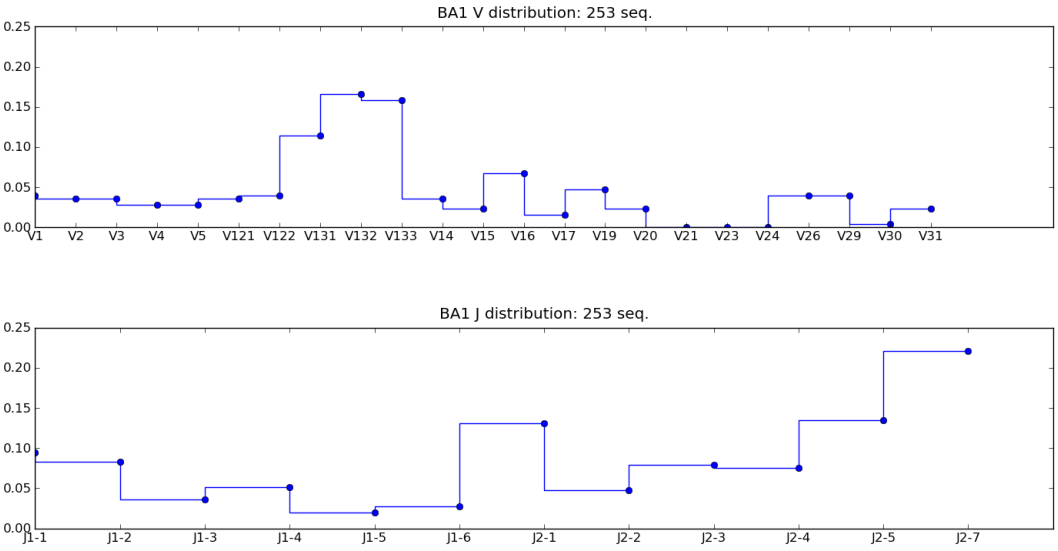

Figure B.1: V-J frequency plots for the naïve mouse BA1.

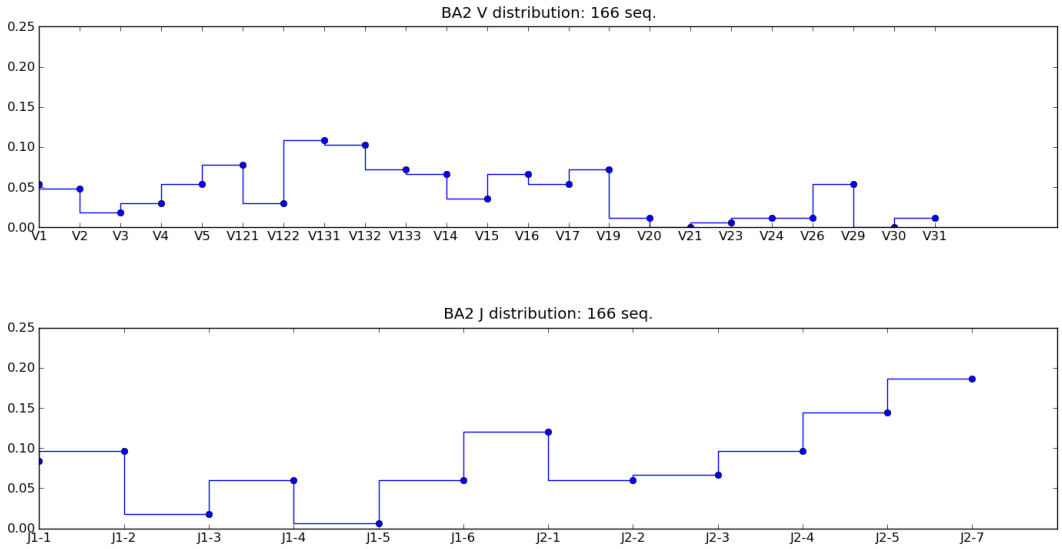

Figure B.2: V-J frequency plots for the naïve mouse BA2.

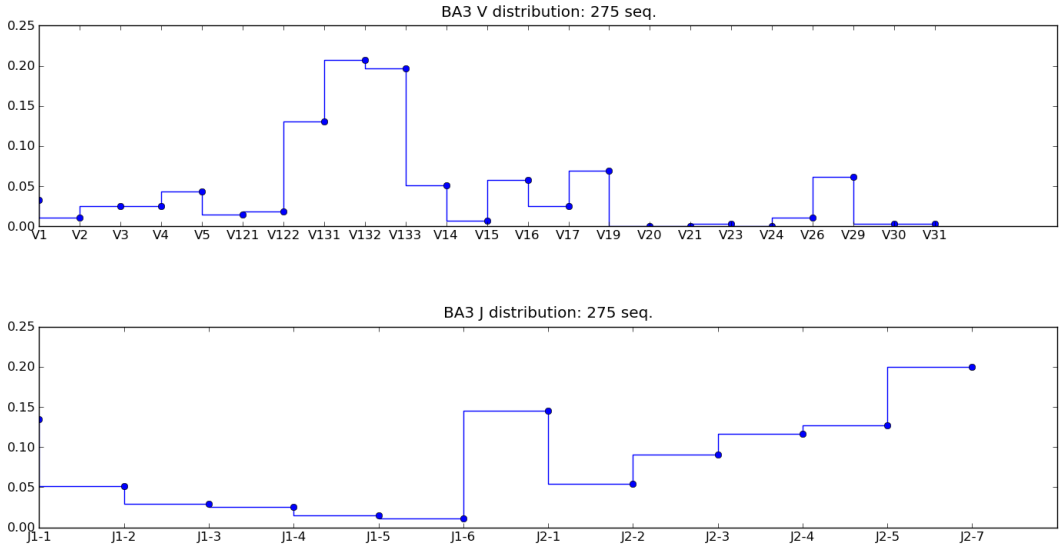

Figure B.3: V-J frequency plots for the naïve mouse BA3.

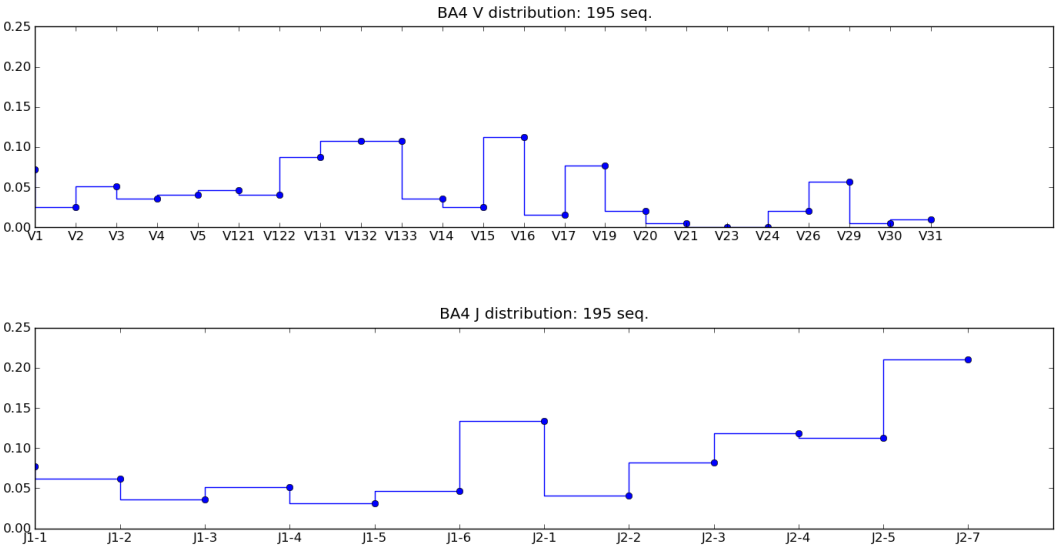

Figure B.4: V-J frequency plots for the naïve mouse BA4.

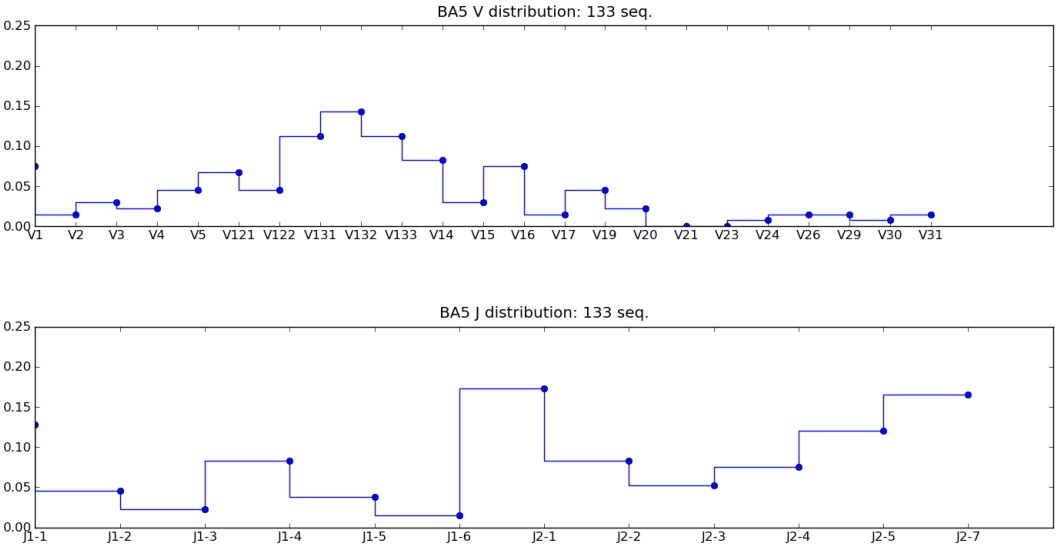

Figure B.5: V-J frequency plots for the naïve mouse BA5.

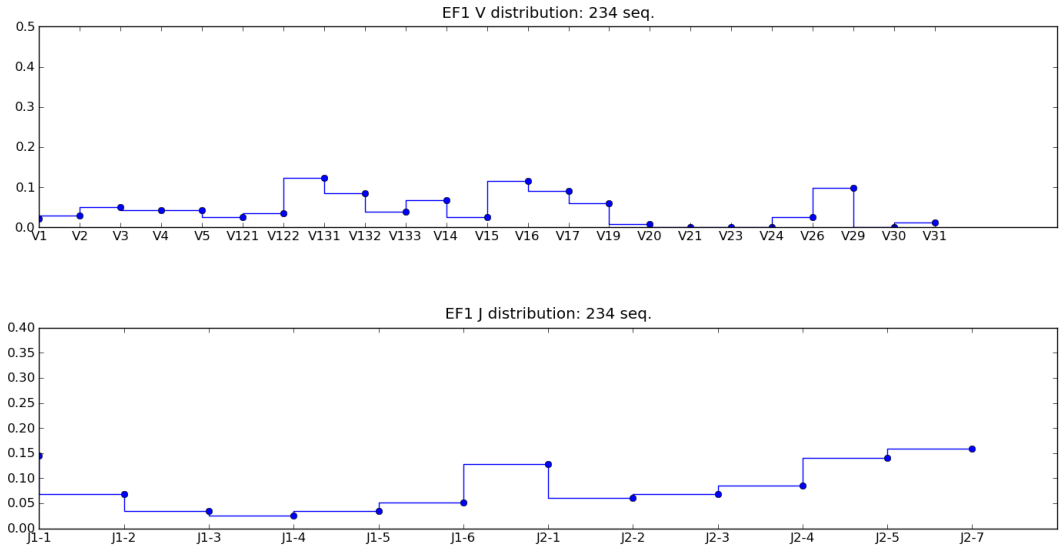

Figure B.6: V-J frequency plots for the infected mouse EF1.

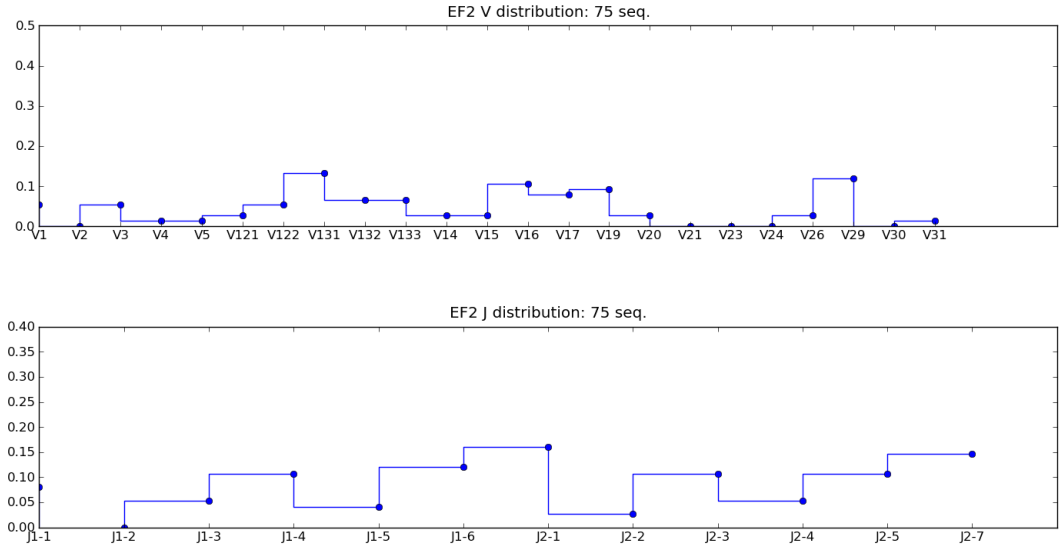

Figure B.7: V-J frequency plots for the infected mouse EF2.

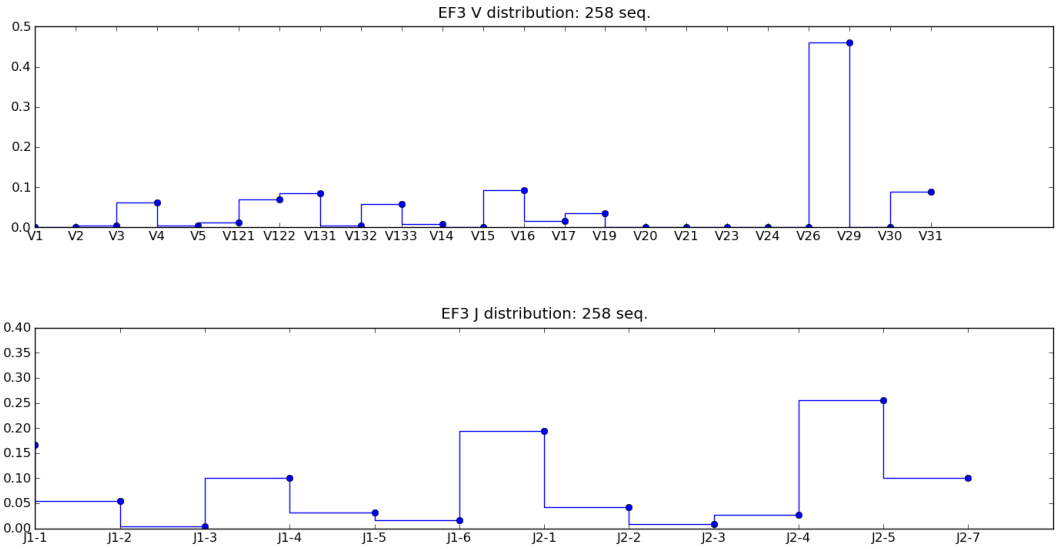

Figure B.8: V-J frequency plots for the infected mouse EF3.

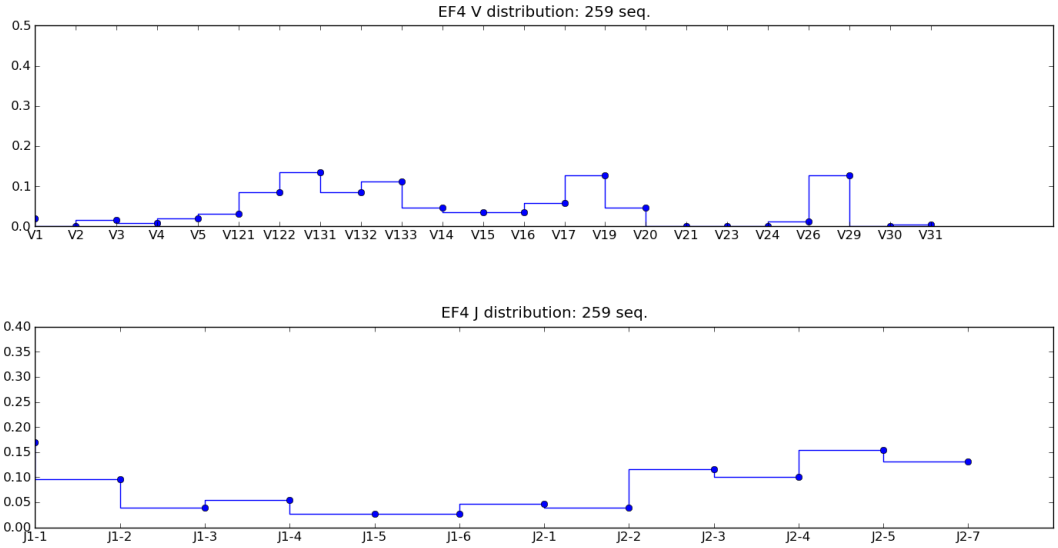

Figure B.9: V-J frequency plots for the infected mouse EF4.

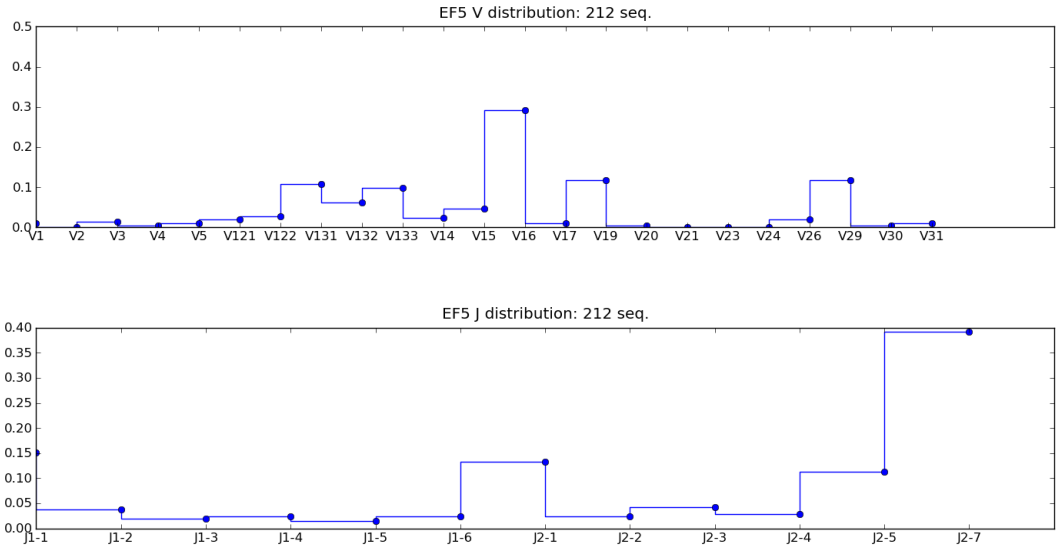

Figure B.10: V-J frequency plots for the infected mouse EF5.

Appendix C

VJ frequencies

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	$4 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	0	0	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	0	$1.2 \cdot 10^{-2}$
V_2	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	0	0	$4 \cdot 10^{-3}$	0	$4 \cdot 10^{-3}$	0	$7.9 \cdot 10^{-3}$	0	0	$4 \cdot 10^{-3}$
V_3	$7.9 \cdot 10^{-3}$	0	$4 \cdot 10^{-3}$	0	0	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$
V_4	0	0	0	0	0	0	0	0	$1.2 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$
V_5	$7.9 \cdot 10^{-3}$	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	0	0	0	0	$4 \cdot 10^{-3}$	0	0
V_{12-1}	0	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	0	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	0	$7.9 \cdot 10^{-3}$
V_{12-2}	0	$4 \cdot 10^{-3}$	0	0	0	0	$4 \cdot 10^{-3}$	0	0	0	$1.6 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$
V_{13-1}	$7.9 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$	0	$1.2 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$2 \cdot 10^{-2}$
V_{13-2}	$1.6 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	$1.2 \cdot 10^{-2}$	0	$4 \cdot 10^{-3}$	$2.4 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$2.8 \cdot 10^{-2}$	$3.2 \cdot 10^{-2}$
V_{13-3}	0	$7.9 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	0	$7.9 \cdot 10^{-3}$	$2.4 \cdot 10^{-2}$	$7.9 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$2.8 \cdot 10^{-2}$
V_{14}	$7.9 \cdot 10^{-3}$	0	0	$4 \cdot 10^{-3}$	0	0	$1.2 \cdot 10^{-2}$	0	$4 \cdot 10^{-3}$	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$
V_{15}	$1.2 \cdot 10^{-2}$	0	0	0	0	0	$7.9 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	0	0	0
V_{16}	$4 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	0	$1.2 \cdot 10^{-2}$	0	0	$7.9 \cdot 10^{-3}$	0	$2.8 \cdot 10^{-2}$
V_{17}	0	0	0	0	0	0	0	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$
V_{19}	0	0	0	0	0	$7.9 \cdot 10^{-3}$	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	$1.6 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$
V_{20}	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	0	0	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	$4 \cdot 10^{-3}$	0	0	0	$4 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$	0	0	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$
V_{29}	$7.9 \cdot 10^{-3}$	0	0	0	0	0	$7.9 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	0	$7.9 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	$4 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	0	0	0	0	$4 \cdot 10^{-3}$	0
V_{31}	0	0	0	0	0	0	0	0	0	0	$1.2 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$

Table C.1: Table for BA1.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	$1.2 \cdot 10^{-2}$	0	$1.2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	0
V_2	0	$6 \cdot 10^{-3}$	0	0	$6 \cdot 10^{-3}$	0	$6 \cdot 10^{-3}$	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	0
V_3	0	$6 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$	0	0	0	0	0	0	0	0
V_4	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	0	0	$6 \cdot 10^{-3}$	0	$6 \cdot 10^{-3}$	0	0	$6 \cdot 10^{-3}$
V_5	0	0	0	$6 \cdot 10^{-3}$	0	0	$1.2 \cdot 10^{-2}$	0	0	$1.2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$
V_{12-1}	0	$2.4 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	0	0	0	$6 \cdot 10^{-3}$	0	0	$6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$
V_{12-2}	0	$6 \cdot 10^{-3}$	0	0	0	0	$1.2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	0	0	0	$6 \cdot 10^{-3}$
V_{13-1}	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	0	0	$1.2 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$
V_{13-2}	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	0	0	$1.2 \cdot 10^{-2}$	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$3 \cdot 10^{-2}$
V_{13-3}	$1.2 \cdot 10^{-2}$	0	0	0	0	$6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$
V_{14}	0	$6 \cdot 10^{-3}$	0	0	0	0	$1.2 \cdot 10^{-2}$	0	$6 \cdot 10^{-3}$	0	$2.4 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$
V_{15}	$6 \cdot 10^{-3}$	0	0	0	0	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	0	$6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$
V_{16}	$6 \cdot 10^{-3}$	0	0	$1.8 \cdot 10^{-2}$	0	0	$6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$
V_{17}	0	0	0	0	0	0	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	$1.8 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$
V_{19}	$1.8 \cdot 10^{-2}$	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	0	$2.4 \cdot 10^{-2}$
V_{20}	0	0	0	0	0	0	0	0	0	$1.2 \cdot 10^{-2}$	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	$6 \cdot 10^{-3}$	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	$6 \cdot 10^{-3}$	0	0	0	$6 \cdot 10^{-3}$	0
V_{26}	0	$6 \cdot 10^{-3}$	0	0	0	0	0	0	0	0	0	$6 \cdot 10^{-3}$
V_{29}	$6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	0	$6 \cdot 10^{-3}$	0	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	0	0
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	$6 \cdot 10^{-3}$	0	0	0	$6 \cdot 10^{-3}$	0	0

Table C.2: Table for BA2.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	$1.1 \cdot 10^{-2}$	$3.6 \cdot 10^{-3}$	0	0	0	0	$7.3 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$
V_2	0	0	0	0	0	0	0	0	0	0	0	$1.1 \cdot 10^{-2}$
V_3	0	0	0	0	0	0	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	0	$1.1 \cdot 10^{-2}$
V_4	0	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	0	0	$7.3 \cdot 10^{-3}$	0	$7.3 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$
V_5	$3.6 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	0	0	$1.5 \cdot 10^{-2}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$
V_{12-1}	0	0	$3.6 \cdot 10^{-3}$	0	0	0	$7.3 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	0	0
V_{12-2}	0	0	$3.6 \cdot 10^{-3}$	0	0	0	$3.6 \cdot 10^{-3}$	0	0	0	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$
V_{13-1}	$1.8 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	0	$2.2 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$7.3 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$
V_{13-2}	$2.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$7.3 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$	$2.9 \cdot 10^{-2}$	$3.6 \cdot 10^{-3}$	$2.9 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	$4.4 \cdot 10^{-2}$
V_{13-3}	$1.8 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	0	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$2.9 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	$2.9 \cdot 10^{-2}$	$4.7 \cdot 10^{-2}$
V_{14}	$7.3 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$	0	0	0	$7.3 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$
V_{15}	0	$3.6 \cdot 10^{-3}$	0	0	0	0	0	$3.6 \cdot 10^{-3}$	0	0	0	0
V_{16}	$3.6 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	$1.1 \cdot 10^{-2}$	0	$1.1 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	$7.3 \cdot 10^{-3}$	0
V_{17}	0	0	$3.6 \cdot 10^{-3}$	0	0	0	$7.3 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$
V_{19}	$3.6 \cdot 10^{-2}$	0	0	$7.3 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	0	0	$3.6 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$
V_{20}	0	0	0	0	0	0	0	0	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	$3.6 \cdot 10^{-3}$	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	0	0	$3.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	0	0	0	$3.6 \cdot 10^{-3}$
V_{29}	$1.1 \cdot 10^{-2}$	0	0	0	0	0	$7.3 \cdot 10^{-3}$	0	$3.6 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$7.3 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	0	0	0	0	0	$3.6 \cdot 10^{-3}$
V_{31}	0	0	0	0	0	$3.6 \cdot 10^{-3}$	0	0	0	0	0	0

Table C.3: Table for BA3.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	$1 \cdot 10^{-2}$	0	$1 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$
V_2	0	0	0	0	0	0	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0
V_3	$5.1 \cdot 10^{-3}$	0	0	0	0	0	$2.1 \cdot 10^{-2}$	0	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$
V_4	0	0	0	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	0	0	0	$5.1 \cdot 10^{-3}$	0	$1.5 \cdot 10^{-2}$
V_5	$5.1 \cdot 10^{-3}$	0	0	$1 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	0	0	0	0	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
V_{12-1}	0	$1 \cdot 10^{-2}$	0	0	0	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$
V_{12-2}	$5.1 \cdot 10^{-3}$	0	0	0	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	0
V_{13-1}	$1 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	0	$5.1 \cdot 10^{-3}$	0	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$3.1 \cdot 10^{-2}$
V_{13-2}	$5.1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	0	0	$5.1 \cdot 10^{-3}$	0	$1 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	$2.1 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$
V_{13-3}	$5.1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	0	$1.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$
V_{14}	0	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	0	$5.1 \cdot 10^{-3}$	0	0	$1 \cdot 10^{-2}$	0	$1 \cdot 10^{-2}$
V_{15}	0	0	0	0	0	0	$5.1 \cdot 10^{-3}$	0	0	$5.1 \cdot 10^{-3}$	0	$1.5 \cdot 10^{-2}$
V_{16}	0	$1 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	0	0	0	$1.5 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$2.1 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$
V_{17}	0	0	0	0	0	0	0	$1 \cdot 10^{-2}$	0	0	$5.1 \cdot 10^{-3}$	0
V_{19}	$1 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$
V_{20}	0	$5.1 \cdot 10^{-3}$	0	0	0	0	$1 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	$5.1 \cdot 10^{-3}$	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	$5.1 \cdot 10^{-3}$	0	0	0	0	0	0	0	$5.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	0	$5.1 \cdot 10^{-3}$
V_{29}	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0	0	$1 \cdot 10^{-2}$	0	$5.1 \cdot 10^{-3}$	0	0	$5.1 \cdot 10^{-3}$	0	$1.5 \cdot 10^{-2}$
V_{30}	0	0	0	0	$5.1 \cdot 10^{-3}$	0	0	0	0	0	0	0
V_{31}	$5.1 \cdot 10^{-3}$	0	0	0	0	$5.1 \cdot 10^{-3}$	0	0	0	0	0	0

Table C.4: Table for BA4.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	$7.5 \cdot 10^{-3}$	0	0	0	0	$7.5 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	0	$1.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$
V_2	0	0	0	0	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$
V_3	0	0	0	$1.5 \cdot 10^{-2}$	0	0	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	0	0	0
V_4	$7.5 \cdot 10^{-3}$	0	0	0	0	0	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	0	0
V_5	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	0	0	$1.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	0	0	0	$7.5 \cdot 10^{-3}$
V_{12-1}	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	0	$1.5 \cdot 10^{-2}$	0	$7.5 \cdot 10^{-3}$
V_{12-2}	0	0	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$
V_{13-1}	$3 \cdot 10^{-2}$	0	0	0	$7.5 \cdot 10^{-3}$	0	$3 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	$2.3 \cdot 10^{-2}$
V_{13-2}	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	0	$1.5 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	0	$3 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$
V_{13-3}	$2.3 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$
V_{14}	$1.5 \cdot 10^{-2}$	0	0	$7.5 \cdot 10^{-3}$	0	0	$3 \cdot 10^{-2}$	0	$1.5 \cdot 10^{-2}$	0	0	$1.5 \cdot 10^{-2}$
V_{15}	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	0	0	0	0	0
V_{16}	$1.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	0	0	$7.5 \cdot 10^{-3}$	0	$1.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$
V_{17}	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	0	0	0	0	0	0	0	0
V_{19}	0	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	0	0	$1.5 \cdot 10^{-2}$
V_{20}	0	$7.5 \cdot 10^{-3}$	0	0	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	$7.5 \cdot 10^{-3}$	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	$7.5 \cdot 10^{-3}$	0	$7.5 \cdot 10^{-3}$	0	0	0	0	0
V_{29}	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	0	0	0	0	$7.5 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	$7.5 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$

Table C.5: Table for BA5.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	0	0	0	$4.3 \cdot 10^{-3}$	0	0	$4.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	0
V_2	0	0	0	0	0	0	$4.3 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	0	0	$1.7 \cdot 10^{-2}$	0
V_3	0	0	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	$1.3 \cdot 10^{-2}$	0	0	0	$2.6 \cdot 10^{-2}$	0
V_4	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	0	0	0	0	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	$1.3 \cdot 10^{-2}$	$4.3 \cdot 10^{-3}$
V_5	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	0	0	0
V_{12-1}	0	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	0	0	$4.3 \cdot 10^{-3}$	0	0
V_{12-2}	0	0	0	0	0	$4.3 \cdot 10^{-3}$	0	0	$4.3 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$
V_{13-1}	$5.6 \cdot 10^{-2}$	0	0	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0	$1.7 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$
V_{13-2}	0	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	0	$1.3 \cdot 10^{-2}$	$4.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$
V_{13-3}	$4.3 \cdot 10^{-3}$	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0	0	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$
V_{14}	0	0	$4.3 \cdot 10^{-3}$	0	$1.7 \cdot 10^{-2}$	0	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	$3.4 \cdot 10^{-2}$
V_{15}	0	0	0	0	0	0	$8.5 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	0	0	$1.3 \cdot 10^{-2}$
V_{16}	$3.8 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	0	0	0	$4.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	0	$4.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$
V_{17}	$1.7 \cdot 10^{-2}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	0	$1.7 \cdot 10^{-2}$	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$
V_{19}	$8.5 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	$4.3 \cdot 10^{-3}$	0	$8.5 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	0	0	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$
V_{20}	0	$4.3 \cdot 10^{-3}$	0	0	0	0	0	0	0	0	0	$4.3 \cdot 10^{-3}$
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	$4.3 \cdot 10^{-3}$	0	0	0	0	0	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	0	0	0	$1.3 \cdot 10^{-2}$
V_{29}	$4.3 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	0	$8.5 \cdot 10^{-3}$	$8.5 \cdot 10^{-3}$	0	$2.1 \cdot 10^{-2}$	0	$3 \cdot 10^{-2}$	$8.5 \cdot 10^{-3}$	0
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	$4.3 \cdot 10^{-3}$	0	0	0	0	$4.3 \cdot 10^{-3}$	0	0	0	$4.3 \cdot 10^{-3}$	0

Table C.6: Table for EF1.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	$1.3 \cdot 10^{-2}$	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0	$2.7 \cdot 10^{-2}$	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	0	0	0	$1.3 \cdot 10^{-2}$	0	$4 \cdot 10^{-2}$	0	0	0	0	0	0
V_4	0	0	0	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0
V_5	$1.3 \cdot 10^{-2}$	0	0	0	0	0	0	0	0	0	0	0
V_{12-1}	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0	0	$1.3 \cdot 10^{-2}$	0
V_{12-2}	0	0	0	0	0	$1.3 \cdot 10^{-2}$	$2.7 \cdot 10^{-2}$	0	0	0	0	$1.3 \cdot 10^{-2}$
V_{13-1}	0	0	0	0	$2.7 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0	$2.7 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$
V_{13-2}	$2.7 \cdot 10^{-2}$	0	0	$1.3 \cdot 10^{-2}$	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$
V_{13-3}	0	0	0	0	0	$1.3 \cdot 10^{-2}$	$2.7 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0	0
V_{14}	0	0	0	0	0	0	$2.7 \cdot 10^{-2}$	0	0	0	0	0
V_{15}	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0	0	$1.3 \cdot 10^{-2}$	0	0	0
V_{16}	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	0	$4 \cdot 10^{-2}$	0	0	$5.3 \cdot 10^{-2}$
V_{17}	$1.3 \cdot 10^{-2}$	0	0	$1.3 \cdot 10^{-2}$	0	0	$1.3 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0	0	$2.7 \cdot 10^{-2}$
V_{19}	$1.3 \cdot 10^{-2}$	0	0	$2.7 \cdot 10^{-2}$	0	$2.7 \cdot 10^{-2}$	0	0	$1.3 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0
V_{20}	0	0	0	0	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	$2.7 \cdot 10^{-2}$	0	0	0	0	0	0	0	0
V_{29}	$1.3 \cdot 10^{-2}$	0	$4 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$	0	0	0	$2.7 \cdot 10^{-2}$	0	$1.3 \cdot 10^{-2}$
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	$1.3 \cdot 10^{-2}$	0	0	0

Table C.7: Table for EF2.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	0	0	0	0	0	0	0	0	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0
V_4	$1.2 \cdot 10^{-2}$	0	0	0	0	0	$7.8 \cdot 10^{-3}$	0	0	0	0	$4.3 \cdot 10^{-2}$
V_5	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0
V_{12-1}	0	0	0	$3.9 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	0	0	0	0
V_{12-2}	0	0	0	0	$2.7 \cdot 10^{-2}$	0	0	$3.1 \cdot 10^{-2}$	$7.8 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	0
V_{13-1}	$7.8 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0
V_{13-2}	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0
V_{13-3}	$3.9 \cdot 10^{-2}$	0	0	0	0	$3.9 \cdot 10^{-3}$	$7.8 \cdot 10^{-3}$	0	0	0	0	$7.8 \cdot 10^{-3}$
V_{14}	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	$3.9 \cdot 10^{-3}$
V_{15}	0	0	0	0	0	0	0	0	0	0	0	0
V_{16}	0	$3.9 \cdot 10^{-3}$	0	0	0	$7.8 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	0	0	0	$7.8 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$
V_{17}	$3.9 \cdot 10^{-3}$	0	0	0	0	0	$1.2 \cdot 10^{-2}$	0	0	0	0	0
V_{19}	$7.8 \cdot 10^{-3}$	0	0	0	0	0	$7.8 \cdot 10^{-3}$	0	0	0	0	$1.9 \cdot 10^{-2}$
V_{20}	0	0	0	0	0	0	0	0	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	0	0	0	0	0	0	0	0	0
V_{29}	$2.7 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	0	$9.7 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$	0	$8.9 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0	$1.9 \cdot 10^{-2}$	$1.6 \cdot 10^{-1}$	$3.9 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	$8.9 \cdot 10^{-2}$	0

Table C.8: Table for EF3.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0	0	$1.2 \cdot 10^{-2}$	0	0	$3.9 \cdot 10^{-3}$
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	$7.7 \cdot 10^{-3}$	0	0	0	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$
V_4	0	0	0	0	0	0	0	0	0	0	0	$7.7 \cdot 10^{-3}$
V_5	0	0	0	0	0	0	0	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	$7.7 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$
V_{12-1}	$3.9 \cdot 10^{-3}$	0	0	$3.9 \cdot 10^{-3}$	0	0	$3.9 \cdot 10^{-3}$	$7.7 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$
V_{12-2}	$7.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	$3.9 \cdot 10^{-3}$	$4.6 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$
V_{13-1}	$5 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-3}$	0	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	$7.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$7.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$
V_{13-2}	$1.2 \cdot 10^{-2}$	0	$7.7 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	0	$3.1 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$	$7.7 \cdot 10^{-3}$	0
V_{13-3}	$4.6 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$7.7 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0
V_{14}	0	0	0	$3.9 \cdot 10^{-3}$	0	$7.7 \cdot 10^{-3}$	0	0	$7.7 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-3}$
V_{15}	0	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	$1.9 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	0
V_{16}	0	0	0	0	0	0	0	0	$7.7 \cdot 10^{-3}$	0	0	$2.7 \cdot 10^{-2}$
V_{17}	$3.9 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	0	0	$1.2 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	$7.7 \cdot 10^{-3}$
V_{19}	$1.5 \cdot 10^{-2}$	0	$1.2 \cdot 10^{-2}$	$7.7 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$	0	$7.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	0	$3.9 \cdot 10^{-2}$
V_{20}	$7.7 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	0	$3.9 \cdot 10^{-3}$	0	0	0	0	$3.9 \cdot 10^{-3}$
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	$3.9 \cdot 10^{-3}$	0	0	0	$3.9 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	0	0	0
V_{29}	$1.5 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$1.9 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$	0	$1.2 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$7.7 \cdot 10^{-3}$	0	$3.9 \cdot 10^{-2}$	$3.9 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	0	0	0	0	0	0
V_{31}	0	$3.9 \cdot 10^{-3}$	0	0	0	0	0	0	0	0	0	0

Table C.9: Table for EF4.

	J_{1-1}	J_{1-2}	J_{1-3}	J_{1-4}	J_{1-5}	J_{1-6}	J_{2-1}	J_{2-2}	J_{2-3}	J_{2-4}	J_{2-5}	J_{2-7}
V_1	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	0
V_2	0	0	0	0	0	0	0	0	0	0	0	0
V_3	$4.7 \cdot 10^{-3}$	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	$4.7 \cdot 10^{-3}$	0
V_4	0	0	0	0	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$
V_5	0	0	0	0	0	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	0	0	0	0	0
V_{12-1}	0	0	0	0	$4.7 \cdot 10^{-3}$	0	$1.4 \cdot 10^{-2}$	0	0	0	0	0
V_{12-2}	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	$4.7 \cdot 10^{-3}$	0	$9.4 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$	0
V_{13-1}	$7.1 \cdot 10^{-2}$	$4.7 \cdot 10^{-3}$	0	0	0	$1.4 \cdot 10^{-2}$	$9.4 \cdot 10^{-3}$	0	0	$4.7 \cdot 10^{-3}$	0	$4.7 \cdot 10^{-3}$
V_{13-2}	$2.4 \cdot 10^{-2}$	0	0	0	0	$4.7 \cdot 10^{-3}$	$1.4 \cdot 10^{-2}$	0	$4.7 \cdot 10^{-3}$	0	0	$1.4 \cdot 10^{-2}$
V_{13-3}	$3.3 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	0	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	0	$4.7 \cdot 10^{-2}$
V_{14}	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	0	$1.9 \cdot 10^{-2}$
V_{15}	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	$4.7 \cdot 10^{-3}$	0	$3.3 \cdot 10^{-2}$	$4.7 \cdot 10^{-3}$
V_{16}	0	$4.7 \cdot 10^{-3}$	0	0	0	0	$9.4 \cdot 10^{-3}$	0	$9.4 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	$2.6 \cdot 10^{-1}$
V_{17}	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	0	0	0	0
V_{19}	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-2}$	0	$9.4 \cdot 10^{-3}$	0	$5.7 \cdot 10^{-2}$	0	0	0	$4.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-2}$
V_{20}	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	0
V_{21}	0	0	0	0	0	0	0	0	0	0	0	0
V_{23}	0	0	0	0	0	0	0	0	0	0	0	0
V_{24}	0	0	0	0	0	0	0	0	0	0	0	0
V_{26}	0	0	0	$1.4 \cdot 10^{-2}$	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$
V_{29}	$1.4 \cdot 10^{-2}$	$9.4 \cdot 10^{-3}$	0	0	0	0	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-2}$	$4.7 \cdot 10^{-3}$	$5.2 \cdot 10^{-2}$	$9.4 \cdot 10^{-3}$
V_{30}	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	0	0	0	0
V_{31}	0	0	0	0	0	0	0	0	0	0	$4.7 \cdot 10^{-3}$	$4.7 \cdot 10^{-3}$

Table C.10: Table for EF5.

Appendix D

Dependance of $T(\alpha, n_\theta)$ on its parameters

We focus here on the dependance of

$$T(\alpha, n_\theta) = \frac{1}{\alpha\mu}(\gamma_E - e^{\alpha n_\theta} \cdot \mathrm{Ei}(-\alpha n_\theta) + \log(\alpha n_\theta)), \tag{D.1}$$

on the parameters n_θ, M, γ, θ and μ. Figures D.1 to D.5 show the dependence of the equation on the different parameters, while Figures D.6 to D.10 represent the elasticity of (D.1) with respect to the different parameters, where the elasticity $Ef(a)$ of a function f in the point $x = a$ is defined as

$$Ef(a) = \frac{a}{f(a)} f'(a),$$

with $f'(a)$ representing the derivative of the function f calculated in the point $x = a$.

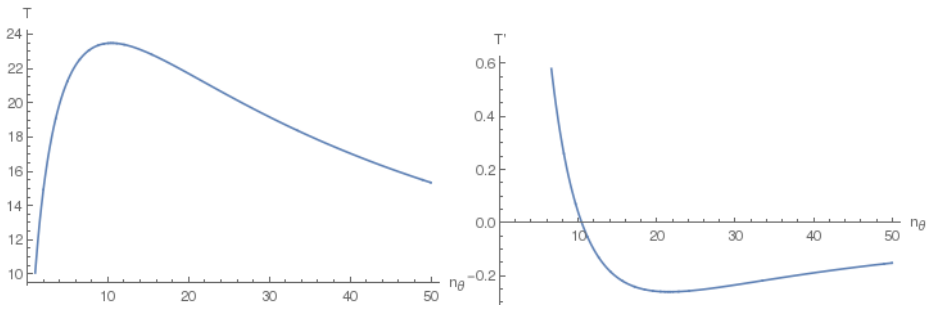

Figure D.1: Parameters: $\theta = 10^9$, $\gamma = 10$, $\mu = 0.5$, $M = 10^{10}$.

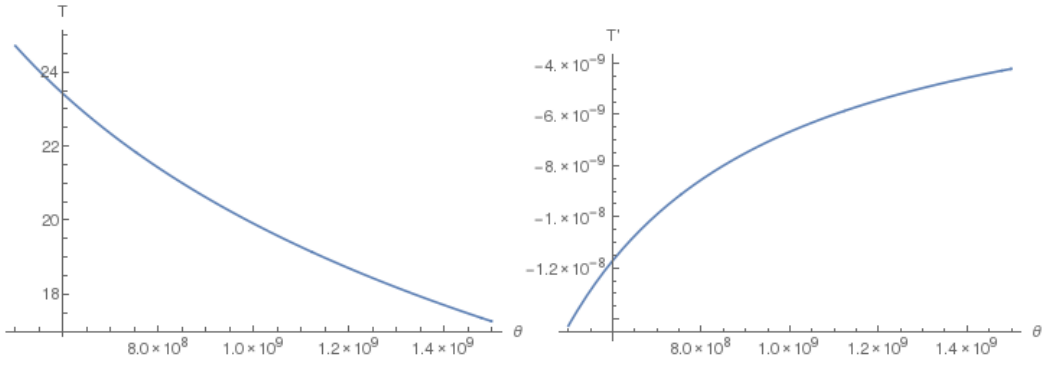

Figure D.2: Parameters: $n_\theta = 4$, $\gamma = 10$, $\mu = 0.5$, $M = 10^{10}$.

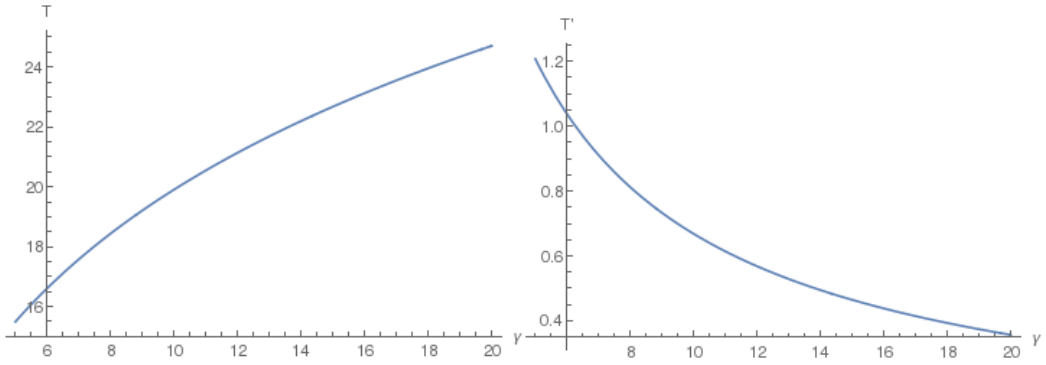

Figure D.3: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\mu = 0.5$, $M = 10^{10}$.

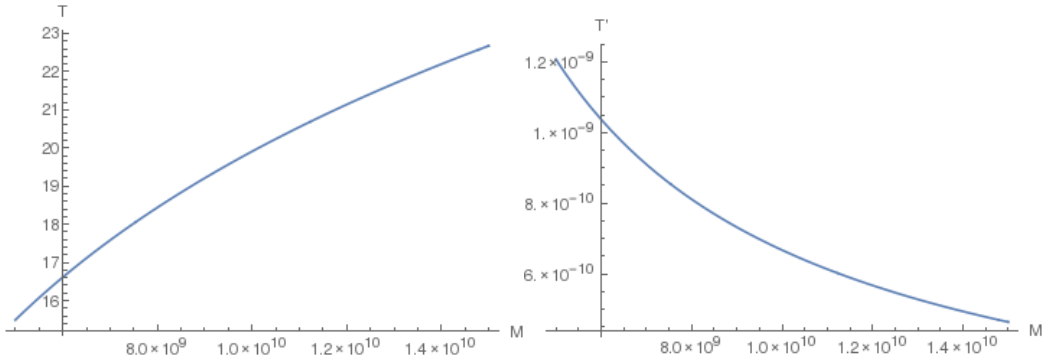

Figure D.4: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\gamma = 10$, $\mu = 0.5$.

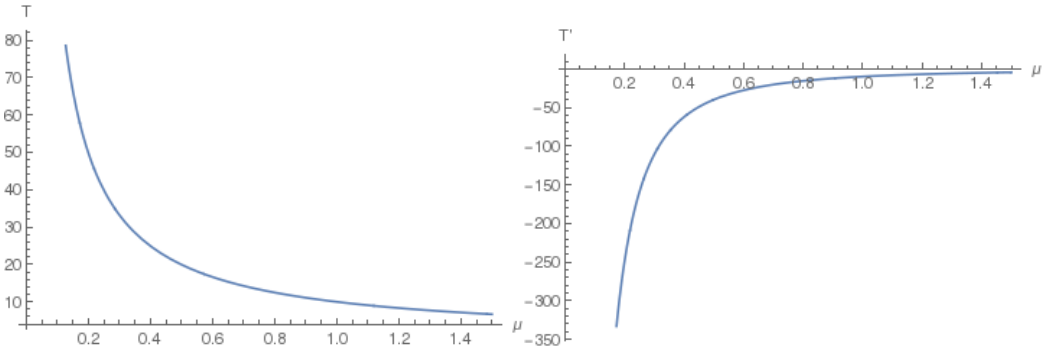

Figure D.5: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\gamma = 10$, $M = 10^{10}$.

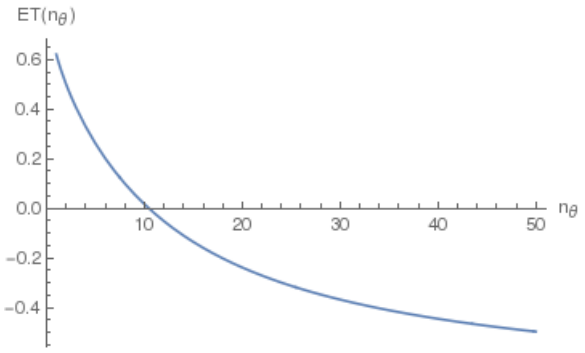

Figure D.6: Parameters: $\theta = 10^9$, $\gamma = 10$, $\mu = 0.5$, $M = 10^{10}$.

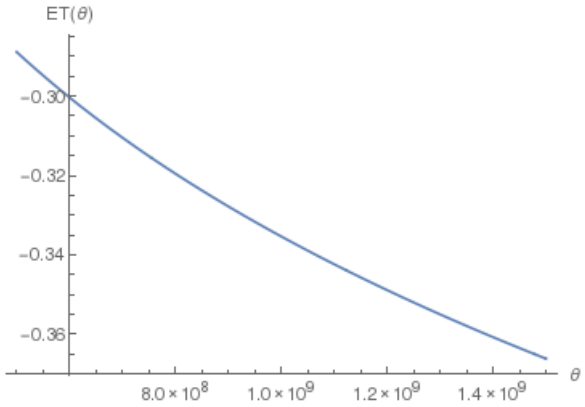

Figure D.7: Parameters: $n_\theta = 4$, $\gamma = 10$, $\mu = 0.5$, $M = 10^{10}$.

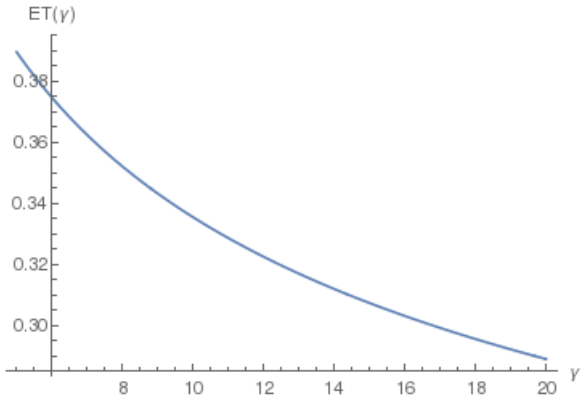

Figure D.8: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\mu = 0.5$, $M = 10^{10}$.

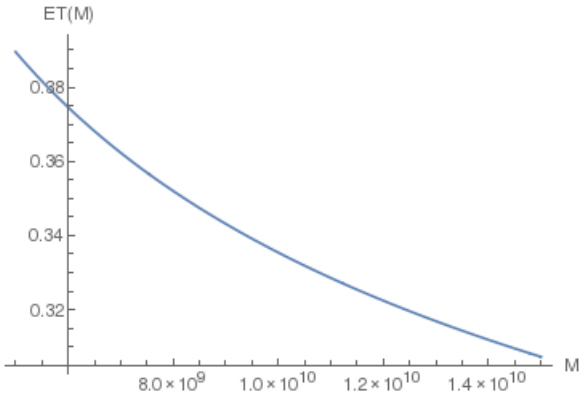

Figure D.9: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\gamma = 10$, $\mu = 0.5$.

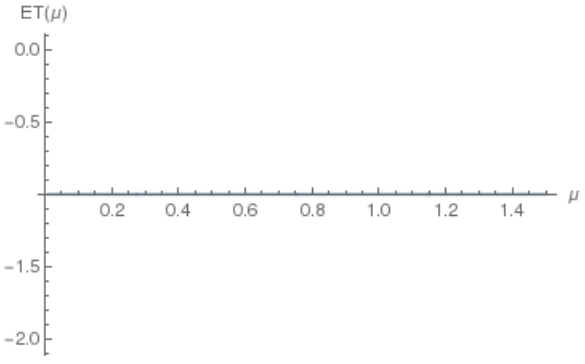

Figure D.10: Parameters: $n_\theta = 4$, $\theta = 10^9$, $\gamma = 10$, $M = 10^{10}$.

Bibliography

[1] Rafi Ahmed and David Gray. Immunological memory and protective immunity: understanding their relation. *Science*, 272(5258):54, 1996. 1

[2] S Munir Alam, Paul J Travers, Jay L Wung, Wade Nasholds, et al. T-cell-receptor affinity and thymocyte positive selection. *Nature*, 381(6583):616, 1996. 6

[3] R Alan, B Ezekowitz, and Jules Hoffmann. Innate immunity: the blossoming of innate immunity. *Current Opinion in Immunology*, 10(1):9–11, 1998. 1

[4] B Alberts et al. Molecular biology of the Cell in Cell 4th, 2002. 3, 4

[5] Linda JS Allen. *An introduction to stochastic processes with applications to biology*. CRC Press, 2010. 20, 110, 112, 113, 122, 126, 132, 133

[6] Frederick W Alt and David Baltimore. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three d-jh fusions. *Proceedings of the National Academy of Sciences*, 79(13):4118–4122, 1982. 12

[7] Frederick W Alt, Eugene M Oltz, Faith Young, James Gorman, Guillermo Taccioli, and Jianzhu Chen. VDJ recombination. *Immunology today*, 13(8):306–314, 1992. 11, 12

[8] Grégoire Altan-Bonnet and Thierry Emonet. Systems immunology: a primer for biophysicists. *Rev Microbiol*, 4:577–87, 2007. 28

[9] Christian L Althaus, Vitaly V Ganusov, and Rob J De Boer. Dynamics of CD8+ T cell responses during acute and chronic lymphocytic choriomeningitis virus infection. *The Journal of Immunology*, 179(5):2944–2951, 2007. 13

[10] A Altman, T Mustelin, and KM Coggeshall. T lymphocyte activation: a biological model of signal transduction. *CRC critical reviews in immunology*, 10(4):347–391, 1990. 2

[11] Mark S Anderson, Emily S Venanzi, Ludger Klein, Zhibin Chen, Stuart P Berzins, Shannon J Turley, Harald Von Boehmer, Roderick Bronson, Andrée Dierich, Christophe Benoist, et al. Projection of an immunological self shadow within the thymus by the AIRE protein. *Science*, 298(5597):1395–1401, 2002. 6

[12] Rustom Antia, Vitaly V Ganusov, and Rafi Ahmed. The role of models in understanding CD8+ T-cell memory. *Nature Reviews Immunology*, 5(2):101–111, 2005. 28

[13] T Petteri Arstila, Armanda Casrouge, Véronique Baron, Jos Even, Jean Kanellopoulos, and Philippe Kourilsky. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, 286(5441):958–961, 1999. 15, 27, 32

[14] Jesús R Artalejo, Antonio Gómez-Corral, M López-García, and C Molina-París. Stochastic descriptors to study the fate and potential of naive T cell clonotypes in the periphery. *Journal of Mathematical Biology*, pages 1–36, 2016. 29, 144

[15] Reto Asmis. Monocytes and macrophages: A fresh look at functional and phenotypic diversity, 2016. 1

[16] Iren Bains, Rustom Antia, Robin Callard, and Andrew J Yates. Quantifying the development of the peripheral naive CD4+ T-cell pool in humans. *Blood*, 113(22):5480–5487, 2009. 31

[17] Irina Baltcheva, Ellen Veel, Thomas Volman, Dan Koning, Anja Brouwer, Jean-Yves Le Boudec, Kiki Tesselaar, Rob J de Boer, and José AM Borghans. A generalized mathematical model to estimate T- and B-cell receptor diversities using AmpliCot. *Biophysical journal*, 103(5):999–1010, 2012. 30

[18] Jacques Banchereau, Francine Briere, Christophe Caux, Jean Davoust, Serge Lebecque, Yong-Jun Liu, Bali Pulendran, and Karolina Palucka. Immunobiology of dendritic cells. *Annual review of immunology*, 18(1):767–811, 2000. 2

[19] Jacques Banchereau and Ralph M Steinman. Dendritic cells and the control of immunity. *Nature*, 392(6673):245–252, 1998. 2

[20] Michele Barry and R Chris Bleackley. Cytotoxic T lymphocytes: all roads lead to death. *Nature Reviews Immunology*, 2(6):401–409, 2002. 4

[21] Craig H Bassing, Frederick W Alt, Maureen M Hughes, Margaux D'auteuil, Tara D Wehrly, Barbara B Woodman, Frank Gärtner, J Michael White, Laurie Davidson, and Barry P Sleckman. Recombination signal sequences restrict chromosomal V (D) J recombination beyond the 12/23 rule. *Nature*, 405(6786):583–586, 2000. 9

[22] Stefan Beissert, Agatha Schwarz, and Thomas Schwarz. Regulatory T cells. *Journal of investigative dermatology*, 126(1):15–24, 2006. 3

[23] Lucy Bird. T cells: Memory cells need more (not less) antigen. *Nature Reviews Immunology*, 14(3):139–139, 2014. 2

[24] Joseph N Blattman, Rustom Antia, David JD Sourdive, Xiaochi Wang, Susan M Kaech, Kaja Murali-Krishna, John D Altman, and Rafi Ahmed. Estimating the precursor frequency of naive antigen-specific CD8+ T cells. *The Journal of experimental medicine*, 195(5):657–664, 2002. 32

[25] Frederick J Bollum. 5. terminal deoxynucleotidyl transferase. *The enzymes*, 10:145–171, 1974. 9

[26] Onur Boyman, Jared F Purton, Charles D Surh, and Jonathan Sprent. Cytokines and T-cell homeostasis. *Current opinion in immunology*, 19(3):320–326, 2007. 26

[27] Isabell Bretschneider, Michael J Clemente, Christian Meisel, Manuel Guerreiro, Mathias Streitz, Werner Hopfenmüller, Jaroslav P Maciejewski, Marcin W Wlodarski, and Hans-Dieter Volk. Discrimination of T-cell subsets and T-cell receptor repertoire distribution. *Immunologic research*, 58(1):20–27, 2014. 14

[28] Frank M Burnet. Immunological recognition of self. *Science*, 133(3449):307–311, 1961. 1

[29] S Candeias, Caroline Waltzinger, Christophe Benoist, and Diane Mathis. The V beta 17+ T cell repertoire: skewed J beta usage after thymic selection; dissimilar CDR3s in CD4+ versus CD8+ cells. *Journal of Experimental Medicine*, 174(5):989–1000, 1991. 66

[30] Alison J Carey, Donald T Gracias, Jillian L Thayer, Alina C Boesteanu, Ogan K Kumova, Yvonne M Mueller, Jennifer L Hope, Joseph A Fraietta, David BH van Zessen, and Peter D Katsikis. Rapid evolution of the CD8+ TCR repertoire in neonatal mice. *The Journal of Immunology*, 196(6):2602–2613, 2016. 15

[31] Armanda Casrouge, Emmanuel Beaudoing, Sophie Dalle, Christophe Pannetier, Jean Kanellopoulos, and Philippe Kourilsky. Size estimate of the $\alpha\beta$ TCR repertoire of naive mouse splenocytes. *The Journal of Immunology*, 164(11):5782–5787, 2000. 15

[32] Rosario Castro, Fumio Takizawa, Wahiba Chaara, Aurélie Lunazzi, Thi Huong Dang, Bernd Koellner, Edwige Quillet, Adrien Six, Uwe Fischer, and Pierre

Boudinot. Contrasted TCRβ diversity of CD8+ and CD8- T cells in rainbow trout. *PloS one*, 8(4):e60175, 2013. 14

[33] Susanna Celli, Zacarias Garcia, and Philippe Bousso. CD4 T cells integrate signals delivered during successive DC encounters in vivo. *The Journal of experimental medicine*, 202(9):1271–1278, 2005. 4

[34] Kevin Chen and Nikolaus Rajewsky. The evolution of gene regulation by transcription factors and microrna. *Nature Reviews Genetics*, 8(2):93–103, 2007. 9

[35] Stanca M Ciupe, Blythe H Devlin, M Louise Markert, and Thomas B Kepler. The dynamics of T-cell receptor repertoire diversity following thymus transplantation for digeorge anomaly. *PLoS Comput Biol*, 5(6):e1000396, 2009. 27

[36] Stanca M Ciupe, Blythe H Devlin, Mary Louise Markert, and Thomas B Kepler. Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. *BMC immunology*, 14(1):1, 2013. 32

[37] Eric T Clambey, Bennett Davenport, John W Kappler, Philippa Marrack, and Dirk Homann. Molecules in medicine mini review: the $\alpha\beta$ T cell receptor. *Journal of Molecular Medicine*, 92(7):735–741, 2014. 3, 4

[38] Lindsay G Cowell, Marco Davila, Dale Ramsden, and Garnett Kelsoe. Computational tools for understanding sequence variability in recombination signals. *Immunological reviews*, 200(1):57–69, 2004. 11

[39] Shane Crotty. Follicular helper CD4 T cells (tfh). *Annual review of immunology*, 29:621–663, 2011. 3

[40] Tania Cukalac, Wan-Ting Kan, Pradyot Dash, Jing Guan, Kylie M Quinn, Stephanie Gras, Paul G Thomas, and Nicole L La Gruta. Paired TCR$\alpha\beta$ analysis of virus-specific CD8+ T cells exposes diversity in a previously defined narrowrepertoire. *Immunology and cell biology*, 93(9):804–814, 2015. 15

[41] Pradyot Dash, Jennifer L McClaren, Thomas H Oguin, William Rothwell, Brandon Todd, Melissa Y Morris, Jared Becksfort, Cory Reynolds, Scott A Brown, Peter C Doherty, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *The Journal of clinical investigation*, 121(1):288–295, 2011. 13

[42] Mark M Davis and Pamela J Bjorkman. T-cell antigen receptor genes and t-cell recognition. 1988. 10

[43] Angus Davison and David RF Leach. Two-base DNA hairpin-loop structures in vivo. *Nucleic acids research*, 22(21):4361–4363, 1994. 8

[44] Rob J De Boer, Mihaela Oprea, Rustom Antia, Kaja Murali-Krishna, Rafi Ahmed, and Alan S Perelson. Recruitment times, proliferation, and apoptosis rates during the CD8+ T-cell response to lymphocytic choriomeningitis virus. *Journal of virology*, 75(22):10663–10669, 2001. 13

[45] Rob J De Boer and Alan S Perelson. Size and connectivity as emergent properties of a developing immune network. *Journal of theoretical biology*, 149(3):381–424, 1991. 25

[46] Rob J De Boer and Alan S Perelson. How diverse should the immune system be? *Proceedings of the Royal Society of London B: Biological Sciences*, 252(1335):171–175, 1993. 2, 25

[47] Rob J De Boer and Alan S Perelson. T cell repertoires and competitive exclusion. *Journal of theoretical biology*, 169(4):375–390, 1994. 26

[48] Rob J De Boer and Alan S Perelson. Towards a general function describing T-cell proliferation. *Journal of theoretical biology*, 175(4):567–576, 1995. 26

[49] Rob J De Boer and Alan S Perelson. Competitive control of the self-renewing T-cell repertoire. *International immunology*, 9(5):779–790, 1997. 26

[50] Jean-Pierre de Villartay. V (d) j recombination deficiencies. In *V (D) J Recombination*, pages 46–58. Springer, 2009. 7

[51] Peter J Delves, Seamus J Martin, Dennis R Burton, and Ivan M Roitt. *Roitt's essential immunology*. John Wiley & Sons, 2016. 2

[52] Ineke den Braber, Tendai Mugwagwa, Nienke Vrisekoop, Liset Westera, Ramona Mögling, Anne Bregje de Boer, Neeltje Willems, Elise HR Schrijver, Gerrit Spierenburg, Koos Gaiser, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36(2):288–297, 2012. 6

[53] Vincent Detours, RAMIT MEHR, and ALAN S PERELSON. A quantitative theory of affinity-driven T cell repertoire selection. *Journal of theoretical biology*, 200(4):389–403, 1999. 5, 25

[54] William S DeWitt, Ryan O Emerson, Paul Lindau, Marissa Vignali, Thomas M Snyder, Cindy Desmarais, Catherine Sanders, Heidi Utsugi, Edus H Warren, Juliana McElrath, et al. Dynamics of the cytotoxic T cell response to a model of acute viral infection. *Journal of virology*, 89(8):4517–4526, 2015. 29

[55] Carolyn Doyle and Jack L Strominger. Interaction between CD4 and class II MHC molecules mediates cell adhesion. 1987. 3

[56] Raluca Eftimie, Joseph J Gillard, and Doreen A Cantrell. Mathematical models for immunology: Current state of the art and future research directions. *Bulletin of Mathematical Biology*, 78(10):2091–2134, 2016. 30

[57] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Inferring processes underlying B-cell repertoire diversity. *Phil. Trans. R. Soc. B*, 370(1676):20140243, 2015. 29

[58] Susan Elmore. Apoptosis: a review of programmed cell death. *Toxicologic pathology*, 35(4):495–516, 2007. 7

[59] Jose Faro, Santiago Velasco, África González-Fernández, and Antonio Bandeira. The impact of thymic antigen diversity on the size of the selected T cell repertoire. *The Journal of Immunology*, 172(4):2247–2255, 2004. 26

[60] Michael A Farrar and Robert D Schreiber. The molecular cell biology of interferon-gamma and its receptor. *Annual review of immunology*, 11(1):571–611, 1993. 3

[61] Marco Ferrarini, Carmen Molina-París, and Grant Lythe. *Sampling from T Cell Receptor Repertoires*, pages 67–79. Springer International Publishing, Cham, 2017. 2

[62] Cristina Ferreira, Yogesh Singh, Anna L Furmanski, F Susan Wong, Oliver A Garden, and Julian Dyson. Non-obese diabetic mice select a low-diversity repertoire of natural regulatory T cells. *Proceedings of the National Academy of Sciences*, 106(20):8320–8325, 2009. 26

[63] Ronald Aylmer Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943. 27

[64] Antonio A Freitas and Benedita Rocha. Population biology of lymphocytes: the flight for survival. *Annual review of immunology*, 18(1):83–111, 2000. 6, 108

[65] K Christopher Garcia, Luc Teyton, and Ian A Wilson. Structural basis of T cell recognition. *Annual review of immunology*, 17(1):369–397, 1999. 4

[66] Raphael Genolet, Brian J Stevenson, Laurent Farinelli, Magne Østerås, and Immanuel F Luescher. Highly diverse TCRα chain repertoire of pre-immune CD8+ T cells reveals new insights in gene recombination. *The EMBO journal*, 31(7):1666–1678, 2012. 10

[67] Ronald N Germain. MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. *Cell*, 76(2):287–299, 1994. 4

[68] Ronald N Germain, Martin Meier-Schellersheim, Aleksandra Nita-Lazar, and Iain DC Fraser. Systems biology in immunology: a computational modeling perspective. *Annual review of immunology*, 29:527–585, 2011. 28

[69] Ananda W Goldrath and Michael J Bevan. Selecting and maintaining a diverse T-cell repertoire. *Nature*, 402(6759):255–262, 1999. 5, 12

[70] Antonio Gómez-Corral and M López García. Extinction times and size of the surviving species in a two-species competition process. *Journal of mathematical biology*, 64(1-2):255–289, 2012. 144

[71] P Gonçalves, M Ferrarini, C Molina-Paris, G Lythe, F Vasseur, A Lim, B Rocha, and O Azogui. A new mechanism shapes the naïve CD8+ T cell repertoire: The selection for full diversity. *Molecular immunology*, 85:66, 2017. 2, 55, 62

[72] Siamon Gordon. Alternative activation of macrophages. *Nature reviews immunology*, 3(1):23–35, 2003. 2

[73] Tanya M Gottlieb and Stephen P Jackson. The DNA-dependent protein kinase: requirement for DNA ends and association with Ku antigen. *Cell*, 72(1):131–142, 1993. 8

[74] Joshua Greene, Marc R Birtwistle, Leszek Ignatowicz, and Grzegorz A Rempala. Bayesian multivariate poisson abundance models for T-cell receptor data. *Journal of theoretical biology*, 326:1–10, 2013. 27

[75] Victor Greiff, Pooja Bhat, Skylar C Cook, Ulrike Menzel, Wenjing Kang, and Sai T Reddy. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome medicine*, 7(1):49, 2015. 15

[76] Jean-Gerard Guillet, Ming-Zong Lai, Thomas J Briner, Soren Buus, and Alessandro Sette. Immunological self, nonself discrimination. *Science*, 235:865–871, 1987. 1

[77] A. Gut. *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York, 2012. 17

[78] Dongni Hou, Cuicui Chen, Eric John Seely, Shujing Chen, and Yuanlin Song. High-throughput sequencing-based immune repertoire study during infectious disease. *Frontiers in Immunology*, 7, 2016. 15

[79] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50, 1912. 64

[80] Charles A Janeway, Paul Travers, MJ Walport, Mark J Shlomchik, et al. *Immunobiology: the immune system in health and disease*, volume 2. Churchill Livingstone, 2001. 1

[81] Marc K Jenkins, H Hamlet Chu, James B McLachlan, and James 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, 2009. 31

[82] Philip L. F. Johnson, Jrg J. Goronzy, and Rustom Antia. A population biological approach to understanding the maintenance and loss of the T-cell repertoire during aging. *Immunology*, 142(2):167–175, 2014. 5

[83] Joseph Kaplinsky and Ramy Arnaout. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nature Communications*, 7, 2016. 29

[84] Tomohiro Kato, Satoshi Suzuki, Hiroko Sasakawa, Kayo Masuko, Yoko Ikeda, Kusuki Nishioka, and Kazuhiko Yamamoto. Comparison of the Jβ gene usage among different T cell receptor Vβ families in spleens of C57BL/6 mice. *European journal of immunology*, 24(10):2410–2414, 1994. 66

[85] Can Keşmir, José AM Borghans, and Rob J de Boer. Diversity of human $\alpha\beta$ T cell receptors. *Science*, 288(5469):1135–1135, 2000. 32

[86] Brian A Kidd, Lauren A Peters, Eric E Schadt, and Joel T Dudley. Unifying immunology with informatics and multiscale biology. *Nature immunology*, 15(2):118–127, 2014. 28

[87] Jan Klein et al. *Natural history of the major histocompatibility complex*. Wiley, 1986. 2

[88] Ludger Klein, Bruno Kyewski, Paul M Allen, and Kristin A Hogquist. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, 14(6):377–391, 2014. 6

[89] Andrej Košmrlj, Elizabeth L Read, Ying Qi, Todd M Allen, Marcus Altfeld, Steven G Deeks, Florencia Pereyra, Mary Carrington, Bruce D Walker, and Arup K Chakraborty. Effects of thymic selection of the T-cell repertoire on HLA class [thinsp] i-associated control of HIV infection. *Nature*, 465(7296):350–354, 2010. 12

[90] Amaury Lambert et al. The branching process with logistic growth. *The Annals of Applied Probability*, 15(2):1506–1535, 2005. 109

[91] RE Langman and M Cohn. The ET (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protecton. *Molecular immunology*, 24(7):675–697, 1987. 32

[92] Stephanie K Lathrop, Nicole A Santacruz, Dominic Pham, Jingqin Luo, and Chyi-Song Hsieh. Antigen-specific peripheral shaping of the natural regulatory T cell population. *Journal of Experimental Medicine*, 205(13):3105–3117, 2008. 26

[93] Daniel J Laydon, Charles RM Bangham, and Becca Asquith. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Phil. Trans. R. Soc. B*, 370(1675):20140291, 2015. 28, 32

[94] Daniel J Laydon, Anat Melamed, Aaron Sim, Nicolas A Gillet, Kathleen Sim, Sam Darko, J Simon Kroll, Daniel C Douek, David A Price, Charles RM Bangham, et al. Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput Biol*, 10(6):e1003646, 2014. 28

[95] Marie-Paule Lefranc, Veronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Geraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jerome Lane, et al. IMGT®, the international immunogenetics information system®. *Nucleic acids research*, 37(suppl 1):D1006–D1012, 2009. 10

[96] W Conrad Liles and Wesley C Van Voorhis. Review: nomenclature and biologic significance of cytokines involved in inflammation and the host immune response. *Journal of Infectious Diseases*, 172(6):1573–1580, 1995. 2

[97] Meei Yun Lin and Raymond M Welsh. Stability and diversity of T cell receptor repertoire usage during lymphocytic choriomeningitis virus infection of mice. *Journal of Experimental Medicine*, 188(11):1993–2005, 1998. 12

[98] Peter S Linsley and Jeffrey A Ledbetter. The role of the CD28 receptor during T cell responses to antigen. *Annual review of immunology*, 11(1):191–212, 1993. 2

[99] Gary W Litman, Jonathan P Rast, Michael J Shamblott, Robert N Haire, Michele Hulst, William Roess, Ronda T Litman, Kristin R Hinds-Frey, Anna Zilch, and CT Amemiya. Phylogenetic diversification of immunoglobulin genes and the antibody repertoire. *Molecular biology and evolution*, 10(1):60–72, 1993. 7

[100] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-París. How many TCR clonotypes does a body maintain? *Journal of theoretical biology*, 389:214–224, 2016. 30, 32, 108, 109, 110, 148

[101] Yunmei Ma, Klaus Schwarz, and Michael R Lieber. The Artemis: DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps. *DNA repair*, 4(7):845–851, 2005. 12

[102] Robert H MacArthur. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences*, 43(3):293–295, 1957. 27

[103] Asaf Madi, Eric Shifrut, Shlomit Reich-Zeliger, Hilah Gal, Katharine Best, Wilfred Ndifon, Benjamin Chain, Irun R Cohen, and Nir Friedman. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome research*, 24(10):1603–1612, 2014. 14

[104] Vinay S Mahajan, Ilya B Leskov, Jian Zhu Chen, et al. Homeostasis of T-cell diversity. *Cell Mol Immunol*, 2(1):1–10, 2005. 26

[105] Koscak Maruyama. The discovery of adenosine triphosphate and the establishment of its structure. *Journal of the History of Biology*, 24(1):145–154, 1991. 7

[106] Don Mason. A very high level of crossreactivity is an essential feature of the T cell receptor. *Immunology today*, 19(9):395–404, 1998. 25, 32

[107] Don Mason. Some quantitative aspects of T-cell repertoire selection: the requirement for regulatory T cells. *Immunological reviews*, 182(1):80–88, 2001. 25

[108] Jeremy Meier, Catherine Roberts, Kassi Avent, Allison Hazlett, Jennifer Berrie, Kyle Payne, David Hamm, Cindy Desmarais, Catherine Sanders, Kevin T Hogan, et al. Fractal organization of the human T cell repertoire in health and after stem cell transplantation. *Biology of Blood and Marrow Transplantation*, 19(3):366–377, 2013. 28

[109] J Joseph Melenhorst, Matthew DH Lay, David A Price, Sharon D Adams, Josette Zeilah, Edgardo Sosa, Nancy F Hensel, Dean Follmann, Daniel C Douek, Miles P Davenport, et al. Contribution of TCR-β locus and HLA to the shape of the mature human Vβ repertoire. *The Journal of Immunology*, 180(10):6484–6489, 2008. 14

[110] T Mimori and John Avery Hardin. Mechanism of interaction between Ku protein and DNA. *Journal of Biological Chemistry*, 261(22):10375–10379, 1986. 8

[111] Carmen Molina-París, Emily Stirk, Katie Quinn, and Grant Lythe. Continuous-time birth and death processes: diversity maintenance of naïve T cells in the periphery. In *Mathematical Models and Immune Cell Biology*, pages 171–186. Springer, 2011. 41

[112] Daniel L Mueller, Marc K Jenkins, and Ronald H Schwartz. Clonal expansion versus functional clonal inactivation: a costimulatory signalling pathway determines the outcome of T cell antigen receptor occupancy. *Annual review of immunology*, 7(1):445–480, 1989. 4

[113] Kenneth Murphy. *Janeway's immunobiology*. Garland Science, 2011. 1, 2, 3

[114] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland Science, 2016. 2, 4, 6, 7, 8, 9, 10, 11, 32

[115] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012. 27, 29, 32

[116] ZA Nagy. Alloreactivity: an old puzzle revisited. *Scandinavian journal of immunology*, 75(5):463–470, 2012. 25

[117] Ingemar Nåsell. Extinction and quasi-stationarity in the verhulst logistic model. *Journal of Theoretical Biology*, 211(1):11–27, 2001. 41

[118] Yuri N Naumov, Elena N Naumova, Kevin T Hogan, Liisa K Selin, and Jack Gorski. A fractal clonotype distribution in the CD8+ memory T cell repertoire could optimize potential for immune responses. *The Journal of Immunology*, 170(8):3994–4001, 2003. 26

[119] Michelle A Neller, Kristin Ladell, James E McLaren, Katherine K Matthews, Emma Gostick, Johanne M Pentier, Garry Dolton, Andrea JA Schauenburg, Dan Koning, Ana Isabel CA Fontaine Costa, et al. Naive CD8+ T-cell precursors display structured TCR repertoires and composite antigen-driven selection dynamics. *Immunology and cell biology*, 93(7):625–633, 2015. 14

[120] Ryan W Nelson, Daniel Beisang, Noah J Tubo, Thamotharampillai Dileepan, Darin L Wiesner, Kirsten Nielsen, Marcel Wüthrich, Bruce S Klein, Dmitri I Kotov, Justin A Spanier, et al. T cell receptor cross-reactivity between similar foreign and self peptides influences naive cell population size and autoimmunity. *Immunity*, 42(1):95–107, 2015. 13

[121] Janko Nikolich-Žugich, Mark K Slifka, and Ilhem Messaoudi. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–132, 2004. 32

[122] Randolph Noelle and E Charles Snow. T helper cells. *Current opinion in immunology*, 4(3):333–337, 1992. 3

[123] Ed Palmer. Negative selectionclearing out the bad apples from the T-cell repertoire. *Nature Reviews Immunology*, 3(5):383–391, 2003. 1

[124] Christophe Pannetier, Jos Even, and Philippe Kourilsky. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunology today*, 16(4):176–181, 1995. 12

[125] Elli Papaemmanuil, Inmaculada Rapado, Yilong Li, Nicola E Potter, David C Wedge, Jose Tubio, Ludmil B Alexandrov, Peter Van Loo, Susanna L Cooke, John Marshall, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature genetics*, 46(2):116–125, 2014. 12

[126] William E Paul. Fundamental immunology, pub, 1993. 12

[127] William E Paul. The immune system: an introduction. *Fundamental Immunology*, 3, 1993. 3, 4, 10

[128] Laetitia Peaudecerf, Sara Lemos, Alessia Galgano, Gerald Krenn, Florence Vasseur, James P Di Santo, Sophie Ezine, and Benedita Rocha. Thymocytes may persist and differentiate without any input from bone marrow progenitors. *Journal of Experimental Medicine*, 209(8):1401–1408, 2012. 15

[129] Alan S Perelson. Modelling viral and immune system dynamics. *Nature Reviews Immunology*, 2(1):28–36, 2002. 28

[130] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36):13139–13144, 2014. 5, 15, 32

[131] Máire F Quigley, Hui Yee Greenaway, Vanessa Venturi, Ross Lindsay, Kylie M Quinn, Robert A Seder, Daniel C Douek, Miles P Davenport, and David A Price. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proceedings of the National Academy of Sciences*, 107(45):19414–19419, 2010. 15

[132] Kylie M Quinn, Sophie G Zaloumis, Tania Cukalac, Wan-Ting Kan, Xavier YX Sng, Michiko Mirams, Katherine A Watson, James M McCaw, Peter C Doherty, Paul G Thomas, et al. Heightened self-reactivity associated with selective survival, but not expansion, of naïve virus-specific CD8+ T cells in aged mice. *Proceedings of the National Academy of Sciences*, 113(5):1333–1338, 2016. 14

[133] Raimundo Real and Juan M Vargas. The probabilistic basis of Jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996. 64

[134] Grzegorz A Rempala, Michał Seweryn, and Leszek Ignatowicz. Model for comparative analysis of antigen receptor repertoires. *Journal of theoretical biology*, 269(1):1–15, 2011. 27

[135] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19):4099–4107, 2009. 15, 27, 28, 32

[136] Harlan S Robins, Santosh K Srivastava, Paulo V Campregher, Cameron J Turtle, Jessica Andriesen, Stanley R Riddell, Christopher S Carlson, and Edus H Warren. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Science translational medicine*, 2(47):47ra64–47ra64, 2010. 13

[137] Benedita Rocha and Harald von Boehmer. Peripheral selection of the T cell repertoire. *Science*, 251(4998):1225–1228, 1991. 31

[138] David B Roth. V (d) j recombination: mechanism, errors, and fidelity. *Microbiology spectrum*, 2(6), 2014. 12

[139] David B Roth and Nancy L Craig. VDJ recombination: a transposase goes to work. *Cell*, 94(4):411–414, 1998. 8

[140] David B Roth, Pamela B Nakajima, Joseph P Menetski, Melvin J Bosma, and Martin Gellert. V (d) j recombination in mouse thymocytes: double-strand breaks near T cell receptor δ rearrangement signals. *Cell*, 69(1):41–53, 1992. 11

[141] Brian D Rudd, Vanessa Venturi, Megan J Smithey, Sing Sing Way, Miles P Davenport, and Janko Nikolich-Žugich. Diversity of the CD8+ T cell repertoire elicited against an immunodominant epitope does not depend on the context of infection. *The Journal of Immunology*, 184(6):2958–2965, 2010. 13

[142] David G Schatz, Marjorie A Oettinger, and David Baltimore. The V (D) J recombination activating gene, Rag-1. *Cell*, 59(6):1035–1048, 1989. 11

[143] David G Schatz and Patrick C Swanson. V (d) j recombination: mechanisms of initiation. *Annual review of genetics*, 45:167–202, 2011. 11

[144] Eric Sebzda, Sanjeev Mariathasan, Toshiaki Ohteki, Russell Jones, Martin F Bachmann, and Pamela S Ohashi. Selection of the T cell repertoire. *Annual review of immunology*, 17(1):829–874, 1999. 6

[145] Nuno Sepúlveda, Carlos Daniel Paulino, and Jorge Carneiro. Estimation of T-cell repertoire diversity and clonal size distribution by poisson abundance models. *Journal of immunological methods*, 353(1):124–137, 2010. 26, 27, 32

[146] Andrew K Sewell. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, 2012. 32

[147] Edward H Simpson. Measurement of diversity. *Nature*, 1949. 64

[148] Adrien Six, Maria Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M Walczak, and Pierre Boudinot. The past, present, and future of immune repertoire biology–the rise of next-generation repertoire analysis. *Frontiers in immunology*, 4, 2013. 28

[149] Jeffrey A Smith, Sharron H Francis, and Jackie D Corbin. Autophosphorylation: a salient feature of protein kinases. In *Reversible Protein Phosphorylation in Cell Regulation*, pages 51–70. Springer, 1993. 7

[150] George D Snell. Studies in histocompatibility. *Science*, 213(4504):172–178, 1981. 32

[151] David JD Sourdive, Kaja Murali-Krishna, John D Altman, Allan J Zajac, Jason K Whitmire, Christophe Pannetier, Philippe Kourilsky, Brian Evavold, Alessandro Sette, and Rafi Ahmed. Conserved t cell receptor repertoire in primary and memory CD8 T cell responses to an acute viral infection. *Journal of Experimental Medicine*, 188(1):71–82, 1998. 12

[152] Timothy K Starr, Stephen C Jameson, and Kristin A Hogquist. Positive and negative selection of T cells. *Annual review of immunology*, 21(1):139–176, 2003. 3

[153] Emily R Stirk, Carmen Molina-París, and Hugo A van den Berg. Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of theoretical biology*, 255(2):237–249, 2008. 26, 29, 30

[154] Michael JT Stubbington, Tapio Lönnberg, Valentina Proserpio, Simon Clare, Anneliese O Speak, Gordon Dougan, and Sarah A Teichmann. T cell fate and clonality inference from single-cell transcriptomes. *Nature methods*, 2016. 15

[155] George Sugihara. Minimal community structure: an explanation of species abundance patterns. *The American Naturalist*, 116(6):770–787, 1980. 27

[156] Niclas Thomas, Katharine Best, Mattia Cinelli, Shlomit Reich-Zeliger, Hilah Gal, Eric Shifrut, Asaf Madi, Nir Friedman, John Shawe-Taylor, and Benny Chain. Tracking global changes induced in the CD4 T cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, page btu523, 2014. 32

[157] Joseph JC Thome, Boris Grinshpun, Brahma V Kumar, Masaru Kubota, Yoshiaki Ohmura, Harvey Lerner, Gregory D Sempowski, Yufeng Shen, and Donna L Farber. Long-term maintenance of human naïve T cells through in situ homeostasis in lymphoid tissue sites. *Science Immunology*, 1(6):eaah6506, 2016. 29

[158] Claire Thompson and Fiona Powrie. Regulatory T cells. *Current opinion in pharmacology*, 4(4):408–414, 2004. 3

[159] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983. 10

[160] Dik C van Gent, Dale A Ramsden, and Martin Gellert. The RAG1 and RAG2 proteins establish the 12/23 rule in V (D) J recombination. *Cell*, 85(1):107–113, 1996. 11

[161] François Van Laethem, Anastasia N Tikhonova, and Alfred Singer. MHC restriction is imposed on a diverse T cell receptor repertoire by CD4 and CD8 co-receptors during thymic selection. *Trends in immunology*, 33(9):437–441, 2012. 31

[162] Joost PM van Meerwijk, Samuel Marguerat, Rosemary K Lees, Ronald N Germain, BJ Fowlkes, and H Robson MacDonald. Quantitative impact of thymic clonal deletion on the T cell repertoire. *The Journal of experimental medicine*, 185(3):377–384, 1997. 5

[163] Rajat Varma. TCR triggering by the pMHC complex: valency, affinity, and dynamics. *Sci Signal*, 1:21, 2008. 31, 32

[164] Vanessa Venturi, Katherine Kedzierska, Stephen J Turner, Peter C Doherty, and Miles P Davenport. Methods for comparing the diversity of samples of the T cell receptor repertoire. *Journal of immunological methods*, 321(1):182–195, 2007. 26, 32, 62

[165] Yoram Vodovotz, Ashley Xia, Elizabeth L Read, Josep Bassaganya-Riera, David A Hafler, Eduardo Sontag, Jin Wang, John S Tsang, Judy D Day, Steven H Kleinstein, et al. Solving immunology? *Trends in Immunology*, 2016. 29

[166] Harald von Boehmer. Selection of the T-cell repertoire: receptor-controlled checkpoints in T-cell development. *Advances in immunology*, 84:201–238, 2004. 10

[167] George P George P Wadsworth and Joseph G Bryan. Introduction to probability and random variables. Technical report, 1960. 114, 115

[168] René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome research*, 21(5):790–797, 2011. 28, 32

[169] Robert L Weber and VJ Iacono. The cytokines: a review of interleukins. *Periodontal clinical investigations: official publication of the Northeastern Society of Periodontists*, 19(1):17, 1997. 3

[170] Jamie Wong, Reinhard Obst, Margarida Correia-Neves, Grigoriy Losyev, Diane Mathis, and Christophe Benoist. Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory CD4+ T cells. *The Journal of Immunology*, 178(11):7032–7041, 2007. 26

[171] Allan J Zajac, Joseph N Blattman, Kaja Murali-Krishna, David JD Sourdive, M Suresh, John D Altman, and Rafi Ahmed. Viral immune evasion due to persistence of activated T cells without effector function. *Journal of Experimental Medicine*, 188(12):2205–2213, 1998. 13

[172] Veronika I Zarnitsyna, Brian D Evavold, Louis N Schoettle, Joseph N Blattman, and Rustom Antia. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in immunology*, 4, 2013. 2, 28, 32

[173] Jinghui Zhang, Li Ding, Linda Holmfeldt, Gang Wu, Sue L Heatley, Debbie Payne-Turner, John Easton, Xiang Chen, Jianmin Wang, Michael Rusch, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, 481(7380):157–163, 2012. 12

[174] Jinfang Zhu, Hidehiro Yamane, and William E Paul. Differentiation of effector CD4 T cell populations. *Annual review of immunology*, 28:445–489, 2009. 3

[175] Ivan V Zvyagin, Mikhail V Pogorelyy, Marina E Ivanova, Ekaterina A Komech, Mikhail Shugay, Dmitry A Bolotin, Andrey A Shelenkov, Alexey A Kurnosov, Dmitriy B Staroverov, Dmitriy M Chudakov, et al. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 111(16):5980–5985, 2014. 14