Statistical biophysics models of immune repertoire dynamics

Meriem Bensouda Koraichi

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Statistical biophysics models of immune repertoire dynamics.

Soutenue par

Meriem Bensouda Koraichi

Le 23 Septembre 2022

École doctorale n°564

Physique en Île-de-France

Spécialité

Physique

Composition du jury :

Grant Lythe University of Leeds	*Président du jury et Rapporteur*
Andrew J. Yates Columbia University	*Rapporteur*
Chase Broedersz VU Amsterdam	*Examinateur*
Anne-Florence BITBOL EPFL, Lausanne	*Examinatrice*
Aleksandra M WALCZAK CNRS - Ecole Normale Supérieure	*Directrice de thèse*
Thierry MORA CNRS - Ecole Normale Supérieure	*Co-directeur de thèse*

ENS | PSL

La maturité de l'homme est d'avoir retrouvé le sérieux qu'on avait au jeu quand on était enfant.

Alain Damasio, *La Horde du Contrevent.*

Acknowledgements

I would like to express my deep gratitude to my supervisors, Aleksandra Walczak and Thierry Mora, for their invaluable guidance, support, and encouragement throughout my Ph.D. journey. Their expertise and insights have been invaluable in shaping the direction and focus of my research, and I am grateful for their patience and support in helping me to overcome the many challenges that arose along the way. Pursuing a Ph.D. during the Covid pandemic was not easy and so I am very thankful of the way the communication and research were maintained during these critical times. Thank you for making me understand what is research and that it is not supposed to be easy. I have learned so much during these three years and not only about T-cells and statistical physics.

I would also like to thank the members of my supervisory committee, Andrew Yates and Grant Lythe, for their valuable feedback and insights on my work. Their guidance and support have been instrumental in helping me to deliver my Ph.D. manuscript.

I would like to extend my appreciation to my colleagues and friends at ENS Paris and abroad: Victor, Maria, Giulia, Fede, Nati, Giulio, Thomas, Paul, Antonio, Carlos, Max, Barbara, Xiaowen, Andrea, Huy, Francesco, Thomas, Antonio, Paul, Maria-Francesca, Jacopo who have provided a supportive and inspiring environment for my research. This was an era of my life I will not forget thanks to them. I am glad I made friends for life, traveling around the world with them to talk about science and build memories together. I am also very glad to have met very valuable and impacting persons that influenced the way I am thinking about myself and my research.

Finally, I would like to thank my family and friends for their constant love and support throughout my Ph.D. journey. Their encouragement and belief in me have been a source of strength and inspiration, and I am forever grateful for their presence in my life.

Publications

During the three last years I have spent working on my Ph.D. research, I dedicated most of my time to trying to understand the dynamics of T-cell receptor repertoires. This thesis is based on several submitted and published articles. My primary research and input are shown in chapters 3 and 6. These two parts have been submitted for publication, and the preprint can be found in Bensouda Koraichi et al. (2022) and Bensouda Koraichi et al. (2021).

At the beginning of my Ph.D., I had the pleasure of collaborating with Thomas Dupic, with whom we developed the *Immprint* method, published in Dupic et al. (2021). The associated results are presented in chapter 4 and use some of the results of chapter 3.

Also, at the beginning of my Ph.D., which coincided with the beginning of the Covid pandemic, I had the opportunity to help understand T-cell responses to this new virus by analyzing biological data using the methods I was developing, and that led to Bensouda Koraichi et al. (2021). The complete study was published in Minervina et al. (2021).

[1] Bensouda Koraichi M, et al., Inferring the T-cells repertoire dynamics of healthy individuals, *biorXiv* December 2022 10.1101/2022.05.01.490247

[2] Dupic T, Bensouda Koraichi M et al., Immune fingerprinting through repertoire similarity, *PLoS Genetics* 17 (1), e1009301

[3] Bensouda Koraichi M, et al. NoisET: Noise learning and Expansion detection of T-cell receptors with Python, *arXiv* July 17th, 2022 10.48550/arXiv.2102.03568

[4] Minervina AA, et al, Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection, *eLife* 10, e63502

CONTENTS

1 Introduction 9

2 The T-cell receptor repertoire 13

2.1 Immunology background 13
2.1.1 Innate and adaptive immune system 13
2.1.2 T-cell recognition of foreign antigens 14
2.1.3 Naive and memory pool 15
2.1.4 Homeostatic maintenance of the immune system 16
2.1.5 Origin of the diversity of the TCR repertoire - V(D)J recombination 17

2.2 High throughput repertoire sequencing 20
2.2.1 Different Repertoire Sequencing technologies 20
2.2.2 Noise in TCR RepSeq data 23

3 Technical background 27

3.1 Motivation 27

3.2 Stochastic population dynamics 28
3.2.1 Birth and death processes - the neutral model 28

3.2.2 Population dynamics of TCR repertoires 33

3.3 Inference tools 38

3.3.1 Bayesian inference 38
3.3.2 Maximum Likelihood Estimation - MLE 39
3.3.3 The likelihood of the model 40
3.3.4 Hidden variables, continuous hidden-Markov Model 41
3.3.5 Computing errors on parameter estimations 45

4 Inferring T-cell repertoire dynamics from healthy individuals 47

4.1 Introduction and motivation 47

4.2 Results 49

4.2.1 Longitudinal sampling of TCR repertoires of healthy individuals 49
4.2.2 Mathematical model of stochastic clonal dynamics 51
4.2.3 Mathematical model of stochastic clonal dynamics 52
4.2.4 Model inference 55
4.2.5 Validation of the inference methods on synthetic data 57
4.2.6 Analysis of repertoires 58

4.1 Discussion 62

4.2 Methods 65

4.2.1 Longitudinal data 65
4.2.2 Naive inference 65
4.2.3 Full inference 66
4.2.4 Synthetic data 67
4.2.5 Comparison to the VDJdb database 68
4.2.6 Code availability 68

4.3 Supplementary figure 69

5 Immune fingerprinting through repertoire similarity 75

5.1 Introduction and motivation 75

5.2 Methods 82

5.2.1 Datasets & Pre-processing 82
5.2.2 Discrimination scores 82
5.2.3 Estimating mean scores from RepSeq datasets 83
5.2.4 Error rate estimates 85
5.2.5 Modeling the evolution of autologous scores 86

5.3 Supplementary figures 87

6 NoisET 93

6.1 Introduction 93

6.2 Model 95

6.2.1 Modeling experimental noise 95
6.2.2 Detecting responding TCR clones 99

6.3 Features 100

6.3.1 Detecting the peak moment of the response 100
6.3.2 Learning the noise model 101
6.3.3 Detecting responding clones 101
6.3.4 Generating trajectories 101
6.3.5 Diversity estimator 102

6.4 Applications of NoisET 102

6.5 Comparison with existing software 103

6.6 Discussion 104

7 Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection 107

7.1 Introduction 108

7.2 Results 109

7.2.1 Longitudinal tracking of TCR repertoires of COVID-19 patients 109
7.2.2 Two waves of T-cell clone response 111
7.2.3 Memory formation and pre-existing memory 112
7.2.4 Validation by MHC tetramer-staining assay 113
7.2.5 TCR sequence motifs of responding clones 114
7.2.6 Mapping TCR motifs to SARS-CoV-2 epitopes 114
7.2.7 Validation of CD4+ COVID-19 HLA-restricted specificity by cohort association analysis 117

7.3 Discussion 118

7.4 Methods 119

7.4.1 Donors and blood samples 119
7.4.2 SARS-CoV-2 S-RBD domain specific ELISA 119
7.4.3 Isolation of PBMCs and T cell subpopulations 120
7.4.4 TCR library preparation and sequencing 120
7.4.5 TCR repertoire data analysis 120
7.4.6 Data availability 122
7.4.7 Supplementary figures 122

8 Conclusion 131

8.1 Main contributions of this thesis 131

8.2 Future research directions 132

CHAPTER

1

INTRODUCTION

The immune system of humans and animals is constantly triggered by endogenous (deficient cells) or exogenous (infectious agents) stimuli. Its primary role is to distinguish self-molecules from non-self molecules to eliminate both endogenous and exogenous threats. The space of possible target molecules is diverse, and the network of interactions between antigens and immune receptors consists of a bipartite network for which cross-interactions exist. Seeing this system as a physicist, I will introduce general and brief notions of biology needed to understand the immunological questions discussed in this thesis. Fully understanding how the human immune system functions requires a multi-scale approach from the chemical and protein level (for receptors-antigen interactions, protein signaling path, energy barriers for biochemical reactions), via the cellular level, the population of cells interacting and tissues responses to stimuli, to networks of people at the level of with epidemiology and propagation of virus and pathogens within a human population.

In this manuscript, I give tools to model the immune system and use current data to understand fundamental characteristics of the immune system at the scale of the repertoire. In the following, I will focus on working with the number of cells (abundance) sharing the same membrane specific T cell receptor. Each number of cells associated with a specific label (which can be a phenotypic property such as the cell type or the cell's ability to interact with the environment thanks to its receptor) is a dynamic variable. In this thesis, our system of interest will be the T-cell receptor (TCR) repertoire which corresponds to the list of all T-cell receptors as-

sociated with the number of cells having this particular receptor (clonal abundance).

The goal of the work described in this manuscript is to understand the dynamics underlying the T-cell receptor repertoire as a whole system **thanks to data**, taking into account the abundances (or populations) of each receptor, which I want to predict their dynamical evolution. T-cell receptor populations are changing because of their interaction with the environment. These can be interactions with self-peptides or foreign peptides presented to the TCR on MHC-antigen complexes or interactions with signaling proteins that enable communication between cells belonging to one specific pool of the immune system. This signaling can underlie response after an acute stimulus such as a vaccine or a virus, or chronic stimulations due to a particular virus or malignant cells leading to cancer. It is crucial to understand which tool to use to learn from data about how the immune system responds to TCR and individual stimuli to tackle medical questions such as the efficiency of a vaccine or drug design. In this thesis, I model the dynamics of a large number of variables, the N kinds (or species) of TCR present in one person, and the causes of the time variations of these variables. This problem can be mapped to a population dynamics question for which species (TCR) interact with an environment, such as predators interact with prey in ecology theory. As the system of interest is large, of the order of $10^9 - 10^{10}$ TCR species (or clones) with an even larger number of sources of interactions (or antigens), it is not convenient to write deterministic equations to model the trajectory of each TCR clone in this context. My strategy as a statistical physicist is to translate this lack of information into mathematical noise and write stochastic equations instead of deterministic equations to tackle the lack of knowledge about this extensive and complex system.

The TCR repertoire can be seen as an example of a high-dimensional personalized biomarker, the hallmark of future precision medicine. Learning robust models from High-Throughput Repertoire Sequencing data is still complicated. This thesis uses Bayesian inference to extract information about T-cell receptor repertoire abundances.

In chapter 3, I introduce the work discussed in the preprint Bensouda Koraichi et al. (2022). In this chapter, we first study the neutral dynamics that drive the immune system without acute stimuli in healthy individuals. Taking advantage of longitudinal High Throughput TCR Repertoire Sequencing (RepSeq), we quantify the experimental noise to disentangle fundamental changes in TCR clonal abundances between two-time points separated by years. Using Bayesian inference, we learn from data the parameters that describe the TCR repertoire dynamics in healthy people of different ages. We find that data is consistent with the stochastic popu-

lation dynamics we consider, that we can interpret these models biologically, and that they can be used to study and quantify the turnover of a person TCR repertoire.

In chapter 4, I explain the work behind *Immprint*, a method designed in collaboration with Thomas Dupic, for which we use TCR RepSeq abundances data to design a classifier to distinguish two individuals thanks to their TCR repertoire identity ("*Immprint*") and test its stability with time. This study was published as Dupic et al. (2021). Immune repertoires provide a unique fingerprint reflecting the immune history of individuals. This chapter shows that this information is personal and can be used to identify people from repertoires of just a few thousand T-cells. The tool uses an information-theoretic measure of repertoire similarity to classify pairs of repertoire samples from the same versus different individuals. We tested the classifier on published data to discriminate individuals with great accuracy, including homozygous twins, by computing false positive and false negative rates $< 10^{-6}$. We also tested the method's robustness to acute infections and possible changes in the TCR repertoire information with time. To do so, we used the results presented in chapter 3.

The core of the work put forward in this thesis is the manipulation and extraction of TCR RepSeq data thanks to the Bayesian and probabilistic approach inspired by statistical physics. This novel approach and techniques, already studied and exposed in Puelma Touzel et al. (2020) are reviewed in the article building chapter 4. I have implemented these techniques and extended them to several applications in software called *NoisET* (Noise learning and Expansion detection of T-cell receptors with Python). As physicists studying biology and researchers tackling multiple disciplines, we must make the models we design easy to use to optimize the number of applications the method can have. This is the goal of *NoisET*. High-throughput sequencing of T-cell receptors makes it possible to track TCR immune repertoires across time, in different tissues, in acute and chronic diseases, and in healthy individuals. However, quantitative comparison between repertoires is confounded by variability in the read count of each receptor clonotype due to sampling, library preparation, and expression noise. *NoisET* is an easy-to-use python package implementing previous and new methods to account for biological and experimental noise to pre-process longitudinal TCR RepSeq data, infer experimental noise from replicates samples, and generate synthetic RepSeq data and its possible dynamics thanks to the model shown in chapter 3, estimate the diversity of TCR immune repertoires and detect responding clones to acute stimuli. We test the package on different repertoire sequencing technologies and datasets. We review how such approaches have been used to identify responding clonotypes in vaccination and disease data.

In early 2020, at the beginning of the pandemic caused by the SARS-CoV-2 coronavirus, I contributed to an extensive collaboration study to improve our global understanding of TCR responses to the new SARS-Cov-2 virus. I improved and optimized the methods behind the previously introduced *NoisET* in chapter 5. Chapter 6 is an analysis of the TCR repertoire response to an acute stimulus and gives a good biological background on TCR repertoire immunology analysis, published in Minervina et al. (2021). This analysis has enabled a better understanding of the diversity of resulting immune memory, the dynamics, and cross-reactivity of the SARS-CoV-2-specific T cell response.

CHAPTER

2

THE T-CELL RECEPTOR REPERTOIRE

2.1. Immunology background

This section covers some basic notions about the human adaptive immune system. The explanation will be comprehensive enough to understand the biological processes we model in this manuscript. For a broader and more complete overview of the immune system, see Murphy et al. (2007).

2.1.1 Innate and adaptive immune system

The immune system of mammals and humans is divided into three defensive lines: the epithelial barrier, innate immunity, and adaptive immunity. The epithelial barriers, such as the skin, and the lining of your digestive tract, are physico-chemical obstacles that prevent infectious agents from invading the body. When a microbe overcomes these barriers, it activates several defense mechanisms immediately after the infection: collectively termed innate immunity. Several innate mechanisms exist anti-microbial peptides, complement activation, neutrophils, natural killer cells, and phagocytosis. This response takes place from minutes to hours after the infection. In most situations, this response is sufficient to stop the disease if the microbe persists and avoids the innate immunity. An adaptive immune response involving T-cells and B-cells responses will take over: this is th main focus in this thesis. This response will be effective after several days as cell divisions are required for the adaptive immune system to occur. This last defensive line is specific to vertebrates. The innate immune system is not only the first line of defense but also triggers the adaptive

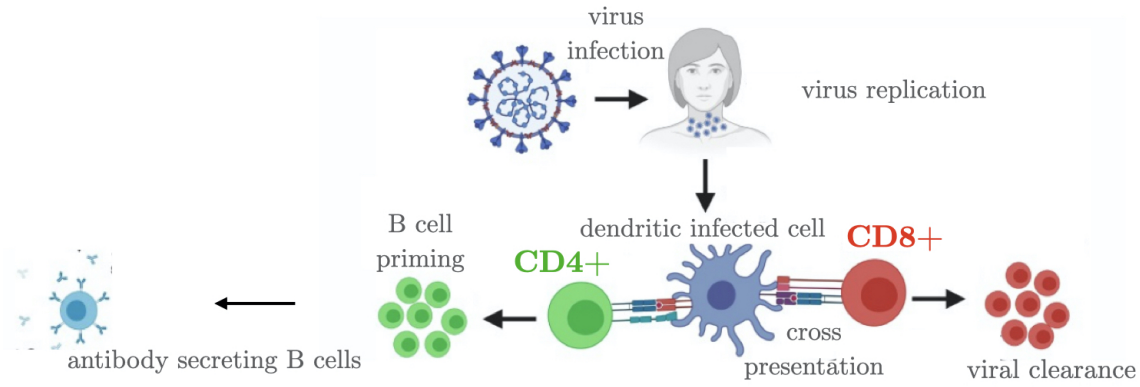

Figure 2.1: When recognizing malignant antigen peptides, the T-cell repertoire acts through two kinds of cells: helper T cells carrying the CD4+ surface protein and cytotoxic T cells carrying the CD8+ surface protein. Here we give the example of how the T-cell immune system responds to virus infection: helper T-cells will send signals to other actors of the immune system, such as B-cells and other T-cells, to clear the virus, and killer T-cells will eradicate cells that the virus has contaminated. Sketch adapted from Gutierrez et al. (2020).

immune system since interaction pathways exist between these two systems. In this manuscript, we will focus only on one of the leading actors of the adaptive immune system, the T-cells.

2.1.2 T-cell recognition of foreign antigens

T-cells have a receptor made up of two proteins to sense the environment surrounding them: the α and β chains. Each T-cell has a specific $\alpha\beta$ pair expressed on its surface: the T-cell receptor (TCR). Each human cell presents major histocompatibility complex (MHC) proteins coded by histocompatibility leukocyte antigen proteins (HLA) alleles, whose role is to present intra-cellular peptides to TCR. When a cell is infected by a virus or has a malfunction, specific T-cell receptors will bind to the MHC-antigen-peptide to trigger an immune response consisting of an expansion of immune cells responsible for eradicating the malfunction. TCR specificity is cross-reactive: several kinds of TCR can recognize a given antigen, and, conversely, several antigens can be detected by one shown TCR. Quantifying and predicting computationally the binding affinities of the formation of the MHC-peptide complex and the TCR-complex represent a vast research topic (Glanville et al. (2017); Fischer et al. (2020); Montemurro et al. (2020); Isacchini et al. (2021); Wu et al. (2021); Shugay et al. (2018); Bravi et al. (2021b,a)) very useful in the field of drug discovery and vaccine design. TCR recognition is not the only sensor of T-cell activation toward an antigen. Signaling proteins such as cytokines activate cell division and death by responding to antigen detection.

One can distinguish two categories of T-cells when looking for actors of an immune response. Two classes of major histocompatibility complex (MHC) exist, class I and class II. T-cells presenting a co-receptor CD4 are called T-helpers because they act via other cells. They bind with peptides associated with MHC class II and initiate B-cell activation (other actors of the adaptive immune system Murphy et al. (2007)). T-cells presenting a co-receptor CD8 will bind with peptides linked to MHC class I. T CD8 cells are called T-killers as they are in charge of eliminating infected cells. The sketch in Fig. 2.1 summarizes the modes of action of CD4 and CD8 T-cells.

Each stimulus will provoke a unique dynamic response of the T-cell immune system. An immune answer consists of expanding the abundances of stimulus-specific T-cell receptors some days after the infection, followed by a contraction of these abundances. These various dynamics are different for most infections, and unveiling them for diverse T-cells subtypes such as CD4/CD8 ones is a question that will be tackled in this manuscript. This question is addressed in the context of the Yellow Fever vaccine, Yellow-Fever double vaccinations, and SARS-CoV2 mild infection. For example, for the Yellow Fever vaccine, it has been known for several years that the peak of the T-cell immune system answer is fifteen days after the introduction of the vaccine (Miller et al. (2008)).

Two keywords are necessary for the understanding of this manuscript: TCR **clone** and TCR repertoire. A TCR **clone** is the collection of all T-cells presenting the same $\alpha\beta$ chain protein. The TCR **repertoire** of an individual is the list of all TCR clones this individual possesses, associated with the number of cells populating each clone. We will call this number abundance or size in this manuscript. The normalized TCR clonal population will be called clone **frequency** and correspond to the TCR population divided by the total number of T-cells. We can define the TCR repertoire **diversity** as its richness, i.e., the number of different TCR clones composing it.

2.1.3 Naive and memory pool

T-cells can be of different types, each of which has a particular role in the response mechanism of the human body to the proliferation of a malignant antigen. Naive T-cells are T cells that have differentiated in the thymus and have successfully undergone the positive and negative processes of central selection in the thymus. Selected naive T-cells should not be too reactive to self-antigens. Still, they should have the potential to bind efficiently against non-self antigens, which makes them a good discriminator between self and non-self-antigens. After selection, naive T-cells are released into the peripheral system composed of the blood, the lymphatic system, and different human body tissues. Naive T-cells have not encountered their cog-

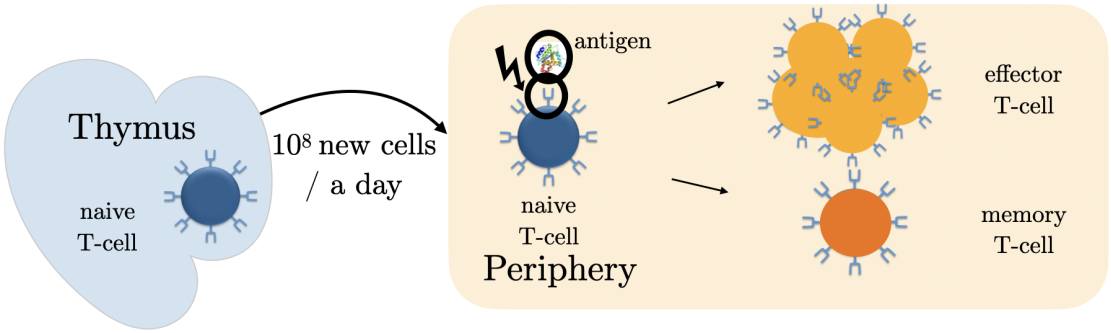

Figure 2.2: Naive T-cells, produced in the thymus following the VDJ recombination process and selection, are released in the somatic periphery. The daily number of new naive T-cells entering the periphery has been estimated to be 10^8 (Yates (2014)). Thanks to their receptors, naive T-cells can sense protein or other malignant cells in the somatic environment and are activated and divided into effector T-cells or memory T-cells.

nate antigen within the periphery. Like every T-cell, naive T-cells, thanks to their receptors, sense protein or other cells present in the physical environment and are activated and divide into effector T-cells (to erase the current antigenic threat) or into memory T-cells (to optimize the future answer to a similar antigenic invasion) (as pictured in Fig. 2.2).
Let us emphasize that naive and memory cells are very different cells with different proteins on their surface, and a different homeostatic ensures separate division and apoptosis dynamics. Consequently, a memory and a naive T-cell have a different lifetime as they have different biological fates. A lifetime of memory cells and naive cells, as well as their ability to divide and answer to an antigenic signal as a function of time, are questions very much related to the field of immunosenescence or how aging impacts the efficiency of the immune response. These questions were broadly tackled quantitatively in Zhang et al. (2021a); Goronzy and Weyand (2013). Still, no quantitative tools exploiting the availability of T-cell repertoire data have been designed to answer these questions. **This will be one of the ambitions of this manuscript.**

2.1.4 Homeostatic maintenance of the immune system

The formation of the naive T-cells repertoire has been quantitatively modeled in different studies in Dessalles et al. (2022); Gaimann et al. (2020); **?**); Greef et al. (2019). This statistical analysis of the distributions of naive T-cells and the modeling of the naive pool seems to lead to three conclusions:

- the naive pool is very diverse. It is composed of cells having a longer lifetime than memory ones to compensate for thymus involution at the beginning of

adulthood (Yates (2014)),

- naive cells generally have small somatic clone abundances, and naive somatic cells with large abundance are likely to be established early in the development of the T-cell repertoire, whose properties are mainly shaped early in life (Pogorelyy et al. (2017)),
- it is still difficult to assess diversity, and to quantify statistical properties of the naive T-cell repertoire, as data are still lacking information to draw robust conclusions.

Regarding the dynamic evolution of abundances of memory T-cells, it has been shown that memory T cells accumulate throughout life. A study discribing the different aspects of the maintenance, compartmentalization, and homeostasis of the human memory T-cell pool can be found in Farber et al. (2014).

Mechanisms that help the homeostatic maintenance of the naive and memory T-cell pools, ensure their diversity, specificity, and healthy metabolism are still unknown. In this manuscript, we propose models and inference tools to help quantify the various changes occurring during multiple vaccinations, multiple infections, chronic diseases, or simply in a healthy state to improve our general knowledge about T-cell repertoire dynamics and distinguish healthy from nonhealthy behaviors.
Naive and memory T-cells exist in two subsets that differ by their role in eliminating a threatening antigen: helper T cells (CD4+) and cytotoxic T cells (CD8+), introduced in section 2.1.2. As shown in multiple studies such as by Rane et al. (2022); De Boer et al. (2003), CD4 and CD8 have different TCR clonal abundances statistics, which may underlie differences in their dynamic behaviors, which we will quantify in this manuscript in the presence of known stimuli and in the absence of known infections to explore the long-timescale dynamics of both T-cell subsets.

2.1.5 Origin of the diversity of the TCR repertoire - V(D)J recombination

The TCR repertoire relies on many TCR species (or clones). Estimating the diversity of this system has been an important question tackled by quantitative immunologists Qi et al. (2014); Lythe et al. (2015); **?**); Robins et al. (2009); Chao and Bunge (2002); Laydon et al. (2015); Qi et al. (2014); Arstila et al. (1999). Unfortunately, because of the inherent statistical properties of the TCR repertoire clonal abundances that we will review later, data have not been informative enough to draw robust diversity estimates. The number of unique clones has been quantified to be between 10^8 and 10^{10} (Lythe et al. (2015); **?**). In this manuscript, we will give a method to estimate this diversity despite data sparseness, making the most of understanding statistical measures of TCR data frequency distributions.

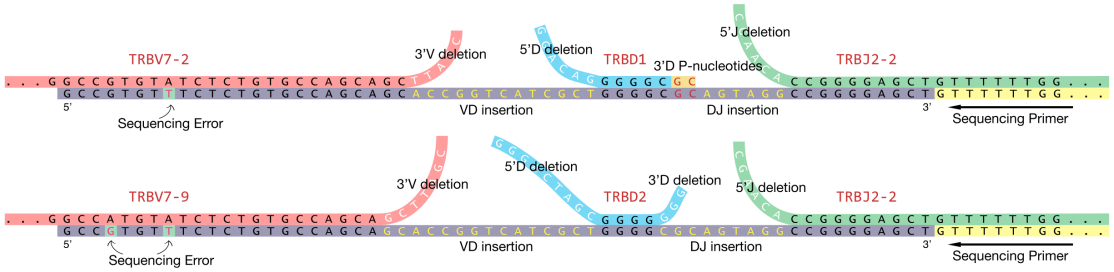

Figure 2.3: From a CDR3 β chain sequence generated from an experiment, several VDJ recombination scenarios are possible. Here two scenarios are shown for the CDR3 sequence in grey: one choice of V gene aligned in two different ways (TRBV7-2 or TRBV7-9), two different choices of d genes (TRBD1 or TRBD2) and one choice of J gene (TRBJ2-2). At top, insVD=13, insDJ=6, delV =5, delJ = 6, del5'D=6, and del3'D = -2 and at bottom, insVD=15, insDJ=9, delV =7, del5'D=9, and del3'D = 3. As the grey sequence is coming from data, you can see that the gene alignements can be not optimal with the presence of sequencing errors. Besides, here, the location of the sequencing primer is given to indicate the start of the CDR3 from J gene sequence. Reproduced from Murugan et al. (2012).

Antigen receptors are proteins that must be encoded as genes in the DNA. Humans have if the order of 10^4 protein-coding genes. Directly encoding the whole diversity of immune receptor genes in each genome makes it impossible for the DNA to fit in a cell nucleus of size $\approx 5\ \mu m$. To explain the high diversity of TCR α and β protein sequences in each individual, we will introduce the V(D)J recombination process, discovered by Susumu Tonegawa (Nobel prize for Physiology or Medicine in 1987).

T-cell receptors are composed of an α and a β chain encoded by separate genes stochastically generated by the thymus's V(D)J recombination process. Each chain is then generated from the combinatorial concatenation of two (V and J for the α chain) or three segments (V as Variable, D as Diversity, and J as Joining for the β chain), picked at random from germline template genes. This biological process is called thymopoiesis — the production of thymocytes, which later become T cells. It includes the random process of choosing among 47 V genes and 61 J functional TRA (α chain) genes and among 48 V, 3D, and 12J functional TRB (β chain) genes. The combinatorics of templates resulting from choosing a V, D, and J gene typically results in $\approx 10^3$ different possible receptors.

But actually, this large diversity comes mainly from random nucleotide insertions and deletions at the VD and DJ junctions for β chains and VJ junction for α chains. The germline gene usage is highly non-uniform due to differences in gene copy numbers, DNA conformation, and processive excision dynamics during recom-

bination. The biases imply that some recombination events are more likely than others. In addition, specific recombination events can lead to the same nucleotide sequence, and many nucleotide sequences can lead to the same amino-acid sequence. The stochastic nature of TCR CDR3 proteins generation increases the hypothetical diversity number of TCR to $\approx 10^{40}$ for each chain and $\approx 10^{60}$ for each $\alpha\beta$ TCR pair for every individual (Marcou et al. (2018)).

Fig. 2.4 gives the different steps of both α and β generation. Each T-cell has two chromosomes coding for V, D, and J genes. The recombination processes described in the previous paragraph may lead to the creation of out-of-frame sequences with codon stops, for example, in the middle of the sequence or nucleotide sequences with a length that is not a multiple of three. The recombination of the two chromosomes is sequential for β and parallel for α, as described in Fig. 2.4. First, the β chain is rearranged on one of the chromosomes before the α chain. It is expressed along with a non-recombined template gene of the α chain on the cell's surface to be checked for function. If the recombination event's first β chain leads to a non-functional receptor, the β chain is rearranged on the second chromosome. If this step gives a functional β chain, this T-cell divides and expands a few times before α chain recombination starts. At this point, the candidate $\alpha\beta$ chains resulting from the process in Fig. 2.4 must pass thymus selection before joining the periphery. In most of the thesis, I will define a TCR clone by one of its two chains, β only or α only.

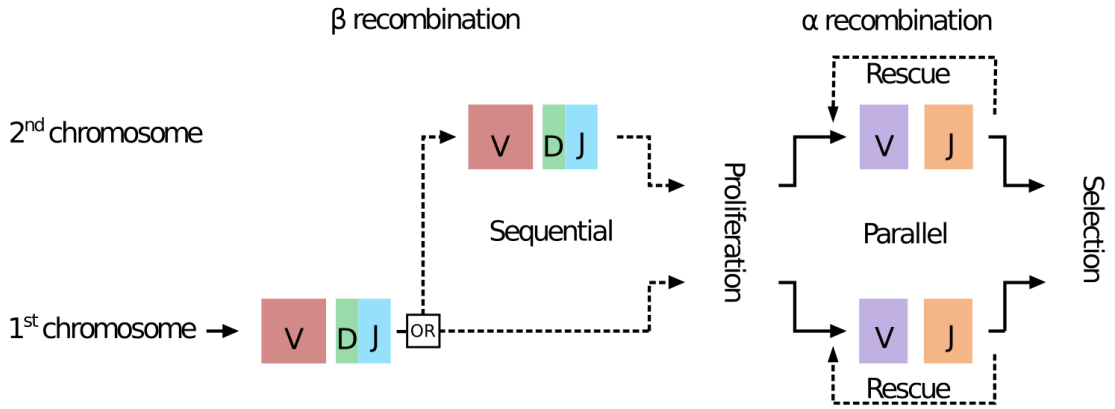

Figure 2.4: VDJ recombination process: different optional paths can lead to the generation of an $\alpha\beta$ TCR on the surface of a T-cell. β chain rearrangement occurs before α chain one. β chain protein functionality is first tested in a sequential way after one or two attempts on each of the chromosomes before a parallel α chain generation happens. The resulting TCR $\alpha\beta$ needs then to pass thymus selection to be released in the periphery and detect threatening antigen peptides. Reproduced from Dupic et al. (2018).

2.2. High throughput repertoire sequencing

2.2.1 Different Repertoire Sequencing technologies

One of the main challenges tackled in this thesis is using new experimental data to design data-driven techniques to extract stochastic population dynamic models of T-cell receptor repertoires. In the last years, many technologies have been developed to produce higher and higher quality data. Understanding how the data is made and what they are is of important for the statistical physicist trying to model the fundamental dynamic behavior of the TCR repertoire. In this section, I will give an overview of the main advancements in TCR repertoire data generation, called Repertoire Sequencing data (RepSeq) (Weinstein et al. (2009); Robins et al. (2009); Boyd et al. (2009); Benichou et al. (2012); Six et al. (2013); Robins (2013); Georgiou et al. (2014); Heather et al. (2017); Minervina et al. (2019b); Rubelt et al. (2017)), and give a non-exhaustive list of experimental sources of noise in the data (Heather et al. (2017); Barennes et al. (2020)). Accurately capturing the TCR repertoire sequence counts presents a significant challenge. Each RepSeq sample is a complex multistep protocol for which each step may impact the RepSeq data and interpretation. The novelty of immune sequencing comes from the recent (in early 2010) rapid development of sequencing techniques and their reduction in costs.

The different steps of the RepSeq data protocol are summarized in Fig. 2.5.

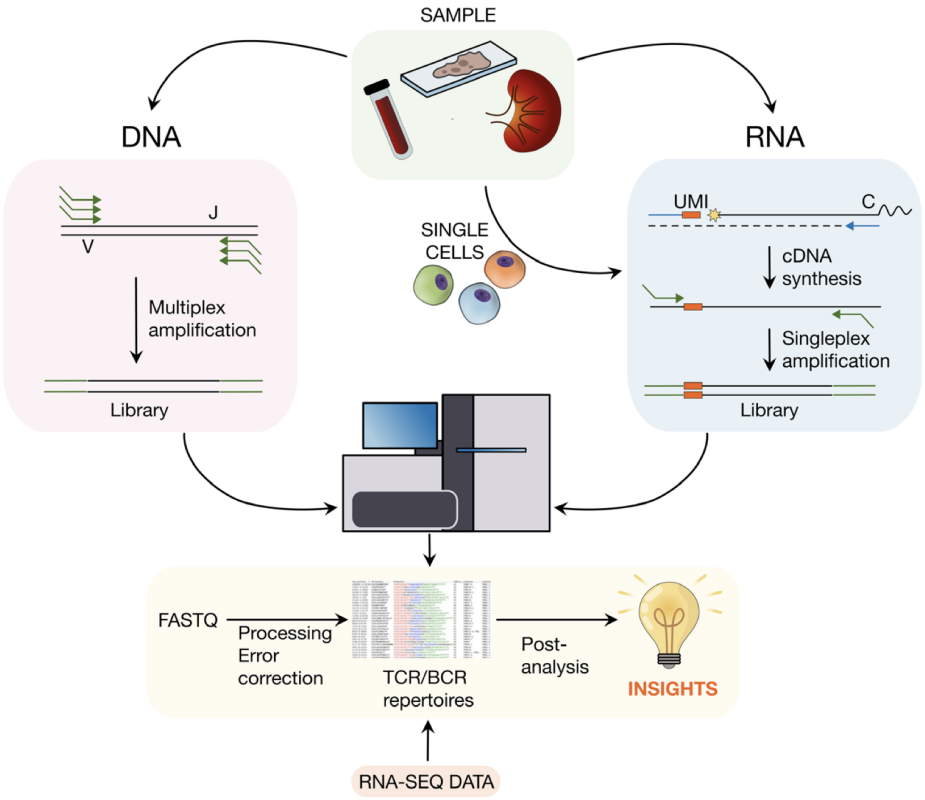

Figure 2.5: After T-cells are sorted from blood samples, DNA- or RNA-based sequencing techniques using either multiplex amplification (for DNA) or Unique Molecular Identifiers (UMIs (for RNA) followed by cDNA synthesis (because we can only sequence DNA material) provide FASTQ files, or list of sequences that need to be processed for error correction. The result is called Repertoire Sequencing (RepSeq) data, as it gives the list of TCR nucleotide and amino-acid sequences and their counts in the analyzed sample. Reproduced from Minervina et al. (2019a).

TR chain	Method	Replicability	Reliability	Sensitivity	Cost per sample	Controls & standards	Format type	fastq data availability
TRA	RACE-1	7	4	4	~230	-	lab protocol	YES
	RACE-1_U	4	5	4	~230	UMI	lab protocol	YES
	RACE-2	5	4	5	230-280	-	service or kit	YES
	RACE-2_U	4	5	5	230-280	UMI	service or kit	YES
	RACE-3	3	2	3	~150	-	kit	YES
	RACE-4	5	6	4	~150	-	lab protocol	YES
	RACE-5	2	3	3	~300	-	lab protocol	YES
TRB	mPCR-1	3	3	3	~350-550*	synthetic TCRs	service or kit	NO
	mPCR-2	6	7	7	~230	-	lab protocol	YES
	mPCR-3	5	5	3	~350-550*	-	service or kit	YES
	RACE-1	6	5	4	~230	-	lab protocol	YES
	RACE-1_U	4	6	5	~230	UMI	lab protocol	YES
	RACE-2	6	6	6	230-280	-	service or kit	YES
	RACE-2_U	6	6	7	230-280	UMI	service or kit	YES
	RACE-3	2	2	3	~150	-	kit	YES
	RACE-4	3	5	4	~150	-	lab protocol	YES

Figure 2.6: After T-cells being sorted from blood samples, DNA- or RNA-based sequencing techniques using either multiplex amplification (for DNA) or Unique Molecular Identifiers (UMIs (for RNA) followed by cDNA synthesis (because we can only sequence DNA material) provide FASTQ files, or list of TCR sequences that need to be processed for error correction. The result is called Repertoire Sequencing (RepSeq) data, as it gives the list of TCR nucleotide and amino-acid sequences and their counts in the analyzed sample. Reproduced from Barennes et al. (2020).

T-cells corresponding to Peripheral blood mononuclear cells (PBMC) are sorted in blood samples. Then, TCR genetic information, which can be the DNA or RNA of each cell, is sequenced using different technologies such as Illumina, and lists of TCR associated with their abundances are generated.

Methods can be classified as DNA- or RNA-based. One RNA-based method uses multiplex PCR (mPCR) with panels of V and J primers and short single-stranded nucleic acids to initiate DNA synthesis. The DNA binding sequence of the primer has to be specifically chosen, which is done using a method called basic local alignment search tool (BLAST) that scans the DNA and finds specific and unique regions for the primer to bind. Another RNA based method is the rapid amplification of cDNA-ends by PCR (RACE-PCR) with the possibility of using unique molecular identifiers (UMI) to limit PCR amplification bias and sequencing errors. Most of the data-set I will introduce and use in this thesis were generated using UMI to determine PCR amplification noise. Unique molecular identifiers are used to mark each T-cell receptor molecule with a unique barcode that can be used to correct sequencing errors and amplification bias. Immune receptor sequences can also be extracted from bulk RNA-Seq data.

Each method has potential advantages and limitations summarized in Fig. 2.6 (Barennes et al. (2020)). DNA-based techniques are believed to be more quantitative and can be used in situations where RNA quality may not be guaranteed. In contrast, RNA-based methods are more sensitive because of the presence of multiple

mRNA copies per cell and can also be more precise as UMI can be combined with the technique. Each method imposes its methodological imprint on the repertoire profile. This is a crucial point to keep in mind when analyzing repertoire data. The experimental noise is technology-dependent, so each data set coming from different RepSeq techniques should be tackled with a new perspective. Fig. 2.6 gives the pros and cons of nine other RepSeq methods for library preparation and sequencing either on RACE-PCR (RACE-1 to RACE-6) or on multiplex-PCR. This comparison is extracted from Barennes et al. (2020).

2.2.2 Noise in TCR RepSeq data

I define noise as the difference in TCR sequences and abundances between RepSeq data and the actual immune repertoire of an individual. Noise can have various sources.

- Experimental noise: stochasticity in reverse transcription, and PCR amplification.
- Sampling noise.
- Detection of rare clones in a blood sample because of the universal and intrinsic power-law distribution of TCR abundances.
- Intrinsic biological noise: mRNA expression.

Noise from the biotechnical experiments

There are acceptable TCRseq methods based on either DNA or RNA input, and in both cases, the number of materials impacts both diversity and the detection of rare clones. The availability of raw data is crucial in allowing reliable and reproducible in-depth analysis of TCR repertoires. Depending on the technology and the different steps of the protocol, TCR read counts are unreliable and need to be used carefully. The sources of PCR noise are described in detail in section 2.2.1.

Detection of rare clones

When dealing with TCR clone abundances, the variable that is commonly used in the analysis of TCR repertoire is not the abundance of the TCR, which will be denoted by $\hat{n}$ in this manuscript, but the normalized associated variable - the empirical frequency defined as $\hat{f} = \hat{n}/N_{\text{reads}}$. In the third chapter dedicated to inference techniques from RepSeq data, I will show that the error between $\hat{f}$ and the actual value of the clone frequency f depends on the value of f. It is important to assess this error to be able to work with frequencies of small clones after defining what a small clone is. For small clones, the value of $\hat{f}$ and f are believed to depend also on the sequencing protocol.

Numerous studies show evidence that TCR frequencies in RepSeq samples follow

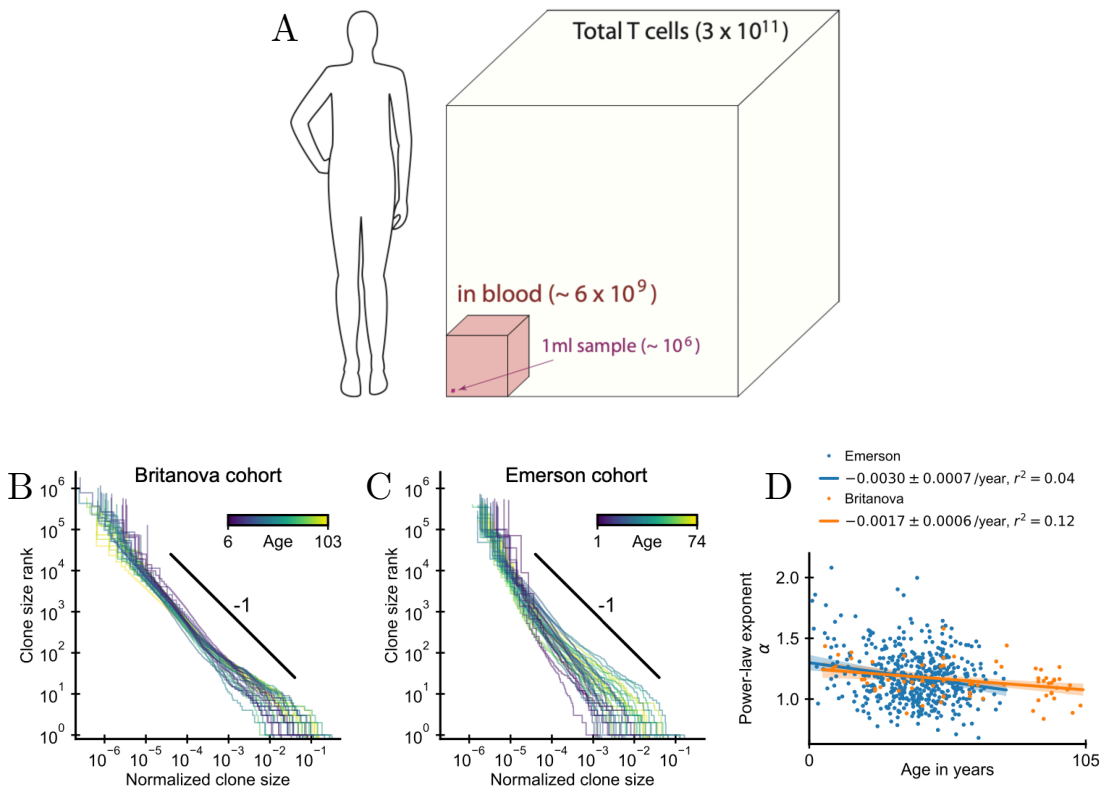

Figure 2.7: **A**.Among T-cells constituting an individual entire TCR repertoire, only blood cells, constituting $\approx 10\%$ of the immune repertoire can be sampled and sequenced. And among this 10%, we are sampling 1 mL of blood out of 5L, so approximately 10^6 cells, which represents 0.001% of the TCR repertoire. Reproduced from **?**. **B-D**.TCR frequency distributions from RepSeq data : 600 people of Emerson cohort (from Adaptive Biotechnologies study in Emerson et al. (2017), from DNA-based sequencing technology) and 170 people in Britanova cohort (from RNA-based sequencing technology, Britanova et al. (2016)). The power-law of these distributions seem to be independent of the age of the individual for which TCR repertoire is analyzed. Reproduced from Gaimann et al. (2020).

a power-law distribution with a very conserved exponent between individuals. As shown in figure Fig 2.7) published in Gaimann et al. (2020), this power-law exponent does not seem to vary subsequently among the 600 people in the cohort from Emerson et al. (2017) and the 170 people of the cohort from Britanova et al. (2014). Additionally, the exponent does not seem to depend on the individual's age or the RepSeq protocol used to produce the data, as depicted in figure Fig 2.7. The form of the TCR abundance distributions does not depend on the sample size. From this power law, one can start assessing the quality of the data. Indeed, depending on the RepSeq protocole, the distributions may change for small clones, which are more affected by PCR error. As a direct effect of subsampling and the inherent form of the TCR frequency distribution, small clones – with a normalized abundance of $1/N_{\text{reads}}$ (N_{reads} being the size of the sample) – have very low probability to be sampled twice in two PBMCs biological replicates produced the same day.

Sampling noise

Another issue when dealing with RepSeq samples and the abundance of clones is sample noise. Most current approaches quantify repertoire properties using measurement statistics that are limited to what is observed in the RepSeq sample, rather than what is expressed in the individual. Only a tiny fraction of the total number of clones is captured in the samples, as depicted in Fig 2.7. Among T-cells constituting an individual's entire TCR repertoire, only blood cells comprising $\approx 1\%$ of the immune repertoire can be sampled and sequenced. And among this 1%, we are testing 1 mL of blood out of 5L, so approximately 10^6 cells represent 0.001% of the TCR repertoire. Some extrapolation is needed to capture the complete information about TCR clone abundances in the data. To tackle this issue in this manuscript, we model the experimental noise using different probabilistic strategies.

Intrinsic biological noise

mRNA is expressed in transcriptional bursts from inheriting noisy promoters (Elowitz et al. (2002); Ozbudak et al. (2002); Cai et al. (2006); Taniguchi et al. (2010); Hornos et al. (2005)). This stochastic transcription leads to long-tailed mRNA abundance distributions, which get translated into TCR abundances in RepSeq experiments. We model this inherent biological noise including long-tailed distributions in our noise model.

CHAPTER

3

TECHNICAL BACKGROUND

3.1. Motivation

In this chapter, I will introduce the mathematical tools needed to model and infer stochastic population dynamics of T-cell receptor repertoires from actual data. I am presenting essential tools to describe stochastic variables and their dynamics to model the abundances of cells that compose, the TCR human repertoire. I introduce descriptions such as the master equation and Fokker-Planck and Langevin equations on the example of the neutral birth and death process. I also provide some simple models of T-cell dynamic predictions in the presence or absence of a known stimulus. In the second part of the chapter, I will introduce Bayesian tools to analyze data to understand the biological processes modeled using stochastic models. In the following, I give a general overview of Bayesian inference, general concepts to study longitudinal TCR RepSeq dynamics, and among them, tools to understand the statistical analysis I have performed in the work presented in this manuscript. In practice, in the results that will be presented in the future chapters, I always have preferred the simplest model, taking the minimum amount of data possible as inputs. I have learned that searching for meaningful biological results instead of mathematical beauty in the models should always be the best choice. The best statistical model to unveil the information in data does not need to be refined.

3.2. Stochastic population dynamics

Stochastic processes model complex systems subject to noise: the trajectory of a random variable describing the physical state as a function of time. Two main approaches are used when dealing with stochastic processes: **Langevin equations** introduce the trajectories of the variable of interest stochastically, and **master equations** with its continuous form, **Fokker-Planck equations**, show the dynamic evolution of the probability distribution describing the variable of interest. To describe the TCR temporal evolution of clone abundance - (or normalized abundance - frequency), I will use Langevin equations or the Fokker-Planck equations. I will describe how to go from one formalism to the other. Once Langevin or Fokker-Planck equations have been defined to respect the assumptions that translate the interactions between agents of the immune repertoire and the environment, propagators should be computed from one of these formalisms. The propagator describes the probability of a random variable at a later timepoint given its earlier value. Basic notions about stochastic processes for physical modeling of biological phenomena can be found in Gardiner (2009).

3.2.1 Birth and death processes - the neutral model

I will first describe the neutral birth and death process. Neutral birth and death mean the cell divides and dies without interacting with the environment independent of its metabolism or age. These phenomena are always present in the background when looking at cell abundances, even if they can be neglected if stronger physical interactions trigger cellular divisions or apoptosis.

Master equation for the birth and death process

First, let us define the master equation for a general discrete process modeling cell abundance:

$$\frac{\partial P(n,t)}{\partial t} = \sum_{n'=1}^{\infty} W(n|n')P(n',t) - \sum_{n'=1}^{\infty} W(n'|n)P(n,t), \tag{3.1}$$

with n the abundance of a TCR species (or clone), $P(n,t)$, the probability for the TCR clone to have a somatic abundance of n cells at time t, $W(n|n')$, the transition probability of going from abundance n' to abundance n. The variables n and n' describe the cell numbers and are integer and discrete random variables. This is what we call a discrete process.

$W(n|n')$, the transition probability has the following definition:

$$W(n|n')\Delta t = \lim_{\Delta t \to 0} P(n, t+\Delta t|n', t). \tag{3.2}$$

Looking at the Markov process (for which knowledge of the present determines

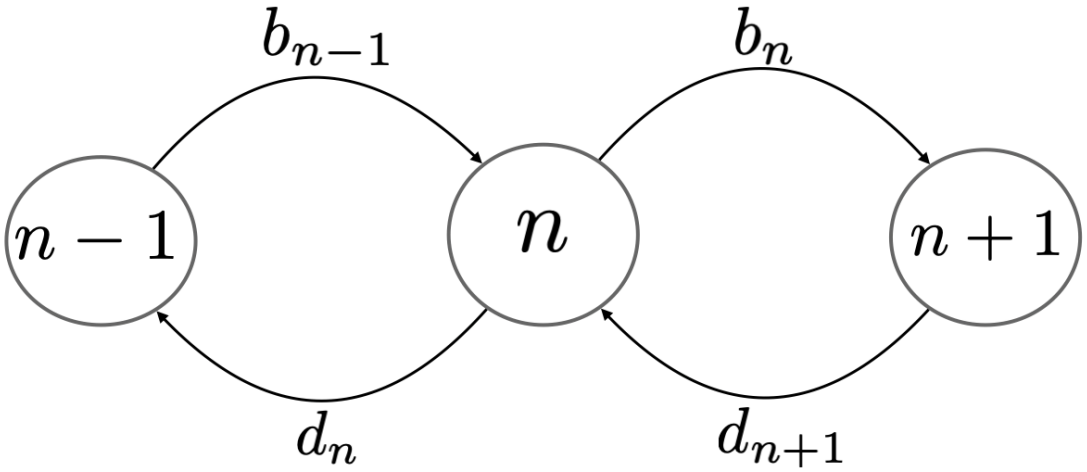

Figure 3.1: Graph-visualization of a birth-death model. For each TCR species having a abundance of n cells, this number can jump to $n+1$ with probability b_n, depending on the variable n and decrease of one increment of one to $n-1$ with probability $n-1$. Solving the process returns to explicitly define the transition probabilities b_n and d_n, with $b_1 = b$ and $d_1 = d$, biological parameters for each cell.

the future) displayed in Fig 3.1, we can write the corresponding master equation followed by every TCR abundance n as the following:

$$\frac{\partial P(n,t)}{\partial t} = d_{n+1}P(n+1,t) + b_{n-1}P(n-1,t) - (b_n + d_n)P(n,t), \tag{3.3}$$

with b_n, the transition probability for producing one additional T cell of the TCR clone, and d_n, the transition probability for one T cell to die for a TCR abundance of n T cells. Birth and death rates are independent of the TCR clone identity. This process depends only on the knowledge of a TCR abundance n at time t. To recover the full neutral dynamics of the TCR repertoire, we can compute the joint probability, assuming that abundances of each clonal TCR are independent variables:

$$P(n^1, n^2, .., n^N, t) = \prod_{i=1}^{N} P(n^i, t), \tag{3.4}$$

with N, the total number of clones of the repertoire, and t is time.

General continuous derivation of the master equation.

From the master equation displayed in Eq. 3.1, we can derive the Fokker-Planck equation, a continuous approach to the master equation. Let us consider the continuous version of the master equation. To do so, we assume that the variable n, describing TCR clone abundance, describes a continuous state and is not constrained to integers. Eq. 3.1 leads to the following equation:

$$\frac{\partial P(n,t)}{\partial t} = \int_{n'=-\infty}^{\infty} W(n|n')P(n',t)dn' - \int_{n'=-\infty}^{\infty} W(n'|n)P(n,t)dn'. \tag{3.5}$$

To simplify the former equation, we apply the following change of variable $r = n - n'$, which transforms $W(n|n') \equiv W(n',r)$ and $W(n|n') = W(n-r,r)$. We approximate both terms of Eq. 3.5 under the assumption of a small jump r, such as $n >> r$:

$$\int_{n'=-\infty}^{\infty} W(n|n')P(n',t)dn' = \int_{r=-\infty}^{\infty} W(n-r,r)P(n-r,t)dr, \tag{3.6}$$

The Taylor expansion of $W(n-r,r)P(n-r,t)$ gives:

$$W(n-r,r)P(n-r,t) = W(n,r)P(n,t) + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} r^k \frac{\partial^k}{\partial n} \left[W(n,r)P(n,t)\right]. \tag{3.7}$$

After using Eq. 3.7 in Eq. 3.5 and defining the coefficients $a_k(n)$ such as :

$$a_k(n) = \int_{r=-\infty}^{\infty} r^k W(n,r)dr, \tag{3.8}$$

we recover the so-called **Kramers-Moyal** expression:

$$\frac{\partial P(n,t)}{\partial t} = \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \frac{\partial^k}{\partial n} \left[a_k(n)P(n,t)\right]. \tag{3.9}$$

Expanding the Kramers-Moyal expression to order 2, we derive the Fokker-Planck equation associated to the general Master-Equation in Eq. 3.1:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}[a_1(n)P(n,t)] + \frac{1}{2}\frac{\partial^2}{\partial n}[a_2(n)P(n,t)] \tag{3.10}$$

In what follows, we will call $a_1(n) \equiv A(n)$ and $a_2(n) \equiv B(n)$ respectively, the **drift** and the **diffusion** coefficients of the stochastic process defining the dynamic evolution of the variable n. One of the goals developed in this manuscript is to understand and express these two parameters $A(n)$ and $B(n)$ in the context of the stochastic population dynamics of T-cell receptor repertoires.

Deriving the Fokker-Planck equation for the birth and death process.

In this section, we are going to derive the Fokker-Planck equation associated to the master equation in Eq. 3.3 with no source term. To do so we introduce the operator Ξ applied to a function f, $\Xi f(m) = f(m+1)$ and the inverse operation, $\Xi^{-1} f(m) = f(m-1)$. $f(m+1)$ and $f(m-1)$ can be approximated using a Taylor expansion under the assumption that $m >> 1$:

$$f(m+1) = \left[\sum_{0}^{\infty} \frac{1}{k!}(\frac{\partial^k}{\partial m})^k\right] f(m), \tag{3.11}$$

and

$$f(m-1) = \left[\sum_{0}^{\infty} \frac{(-1)^k}{k!}\frac{\partial^k}{\partial m}\right] f(m). \tag{3.12}$$

We can now define explicitly general approximations of both operators Ξ and Ξ^{-1}:

$$\Xi = 1 + \frac{\partial}{\partial m} + \frac{1}{2}\frac{\partial^2}{\partial m} + o(m^2), \tag{3.13}$$

$$\Xi^{-1} = 1 - \frac{\partial}{\partial m} + \frac{1}{2}\frac{\partial^2}{\partial m} + o(m^2). \tag{3.14}$$

Now that we have all the tools to simplify and express Ξ and Ξ^{-1} operators, we can use them to write Eq. 3.3:

$$\frac{\partial P(n,t)}{\partial t} = \Xi d_n P(n,t) + \Xi^{-1} b_n P(n,t) - \left[b_n + d_n\right] P(n,t). \tag{3.15}$$

Using the Taylor expansions of both Ξ and Ξ^{-1} operators, we can derive the Fokker-Planck equation associated to the birth and death process:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}[(b_n - d_n)P(n,t)] + \frac{1}{2}\frac{\partial^2}{\partial n^2}[(b_n + d_n)P(n,t)]. \tag{3.16}$$

From Eq. 3.16, we can identify both drift and diffusion parameters of the birth and death process that were introduced in Eq. 3.33, the drift coefficient $A(n) = b_n - d_n$ and the diffusion coefficient $B(n) = b_n + d_n$.

Langevin Equation

The Langevin equation is a differential equation of a random variable whose solution is a random function. Each solution of the Langevin equation represents a different random trajectory. The **Langevin** equation associated with a stochastic process is another way (besides the Fokker-Planck equation) to define the stochastic trajectory described by the variable n, such as the abundance of one TCR clone. As we want to know the trajectories of all N (number of TCR clones in one individual) n variables describing the dynamic evolution of the TCR repertoire, we need to write N possibly coupled SDE or Langevin equations. The Langevin equation can be decomposed into a deterministic force that undergoes the variables n and stochastic forces, called noise. Insights about the noise properties can be found using time correlations of the variable n.

Usually, the Langevin equation is used to describe a system with slow degrees of freedom coupled with fast degrees of freedom whose effect is not taken into account

explicitly but implicitly using a random force. It is often easier to derive a Langevin equation from the microscopic description of single trajectories in the system (if we know the associated noise) and then to extract global information about the probability density. Following the above illustrated simple intuition, the Langevin equation for a microscopic trajectory $n(t)$ is the following:

$$\frac{dn(t)}{dt} = a(n(t), t) + b(n(t), t)\xi(t), \tag{3.17}$$

with $\xi(t)$ white noise process with the Ito convention (Gardiner (2009)), with the following characteristics: $\langle \xi(t) \rangle = 0$ and $\langle \xi(t) \rangle \langle \xi(t') \rangle = \delta(t - t')$, $b(n(t), t)$ is a deterministic function measuring the sensitivity of n to the standard external noise. If the noise has a non-zero mean, it produces a drift over time, and we can treat the drift as part of the deterministic function $a(n(t), t)$. This description of the stochastic process is less intuitive than the Master-Equation, as the Langevin description Eq. 3.17 does not tell us the origin of the randomness in the noise term. It does not tell us if the noise is intrinsic to the system or is due to some external environment. Solving stochastic population dynamics of T cells in the absence or presence of a stimulus, we have already mentioned previously that the noise is coming from the intrinsic randomness of a T cell to divide or die (birth and death noise) and from interactions with the environment.

From the Langevin formalism to the Fokker Planck Equation

In the general derivation of the Fokker-Planck equation from the master equation of a stochastic process, we have introduced the drift $A(n)$ and diffusion $B(n)$ coefficients associated with the Fokker-Planck definition of a stochastic process. As we have defined the Langevin formalism in the previous section, it is useful to be able to link these two parameters $A(n)$ and $B(n)$ to $a(n(t), t)$ and $b(n(t), t)$, in Eq. 3.17.

$$a(n(t), t) = -A(n, t) \tag{3.18}$$

$$b(n(t), t) = \sqrt{B(n, t)} \tag{3.19}$$

These formulae can be derived using Itô rules of stochastic calculus (Gardiner (2009)).

Going back to our birth and death model, we can use the Langevin formalism to define it using Eq. 3.18 and Eq. 3.19:

$$\frac{dn}{dt} = (b - d)n + \sqrt{(b + d)n}\xi(t) \tag{3.20}$$

Trajectory of the total number of cells of the TCR repertoire

Let us consider the total number of T cells $N_{cells} = \sum_{i=1}^{N} n^i$ describing the abundances of the N TCR clones that composed the TCR repertoire. We can write its trajectory thanks to Eq. 3.22, taking into account the thymus source term:

$$\frac{dN_{cells}}{dt} = (b-d)N_{cells} + \sqrt{(b+d)}\sum_{i=1}^{N} n^i(t)\xi(t) + SN_0, \tag{3.21}$$

with S, the rate of introduction of new clones, and N_0 is the abundance at which they are introduced to new clones. We assume that N_0 is a constant, but this number is a function of time and the clone.

If we neglect the birth and death noise here and study the deterministic trajectory of the size of the TCR repertoire N_{cells}, we recover the steady state value of the repertoire size N_∞:

$$N_\infty = \frac{SN_0}{d-b}. \tag{3.22}$$

The constant introduction of new clones (or species) and, new cells ensures a negative growth rate $b < d$ and the existence of a steady state.

3.2.2 Population dynamics of TCR repertoires

A review gathering the state of the art on the topic has been written by Desponds et al. (2021), where multiple models of abundance dynamics of immune repertoire have been introduced in De Boer and Perelson (1994); De Boer et al. (2003); De Boer and Perelson (2013); Lythe et al. (2015); Mayer et al. (2019). Most of these models are built on the same equations based on clonal proliferation into effector and memory cells triggered by strong antigenic recognition. These signals are most of the time of pathogenic origin, for which each TCR clone is not sensitive in the same way. The cell abundance n of TCR clones is no longer controlled by a neutral birth and death model defined previously but depends mainly on the ability of a T-cell receptor (TCR) to bind different peptides. Consequently, the abundance dynamics describing $n(t)$ must depend on the TCR binding properties to antigens present at time t.

Let us take a set of N clones with abundances $(n^1(t), n^2(t), ..., n^N(t))$ describing the TCR repertoire of an individual and a set of M antigens at concentrations $(a^1(t), a^2(t), ..., a^M(t))$. We can define the interaction matrix K of size $N \times M$, with its elements K_{ij} describing the binding affinities between one TCR i and one antigen j. The dynamics of each clone is described by division and death, which occur with Poisson rates that depend on a receptor-specific antigenic stimulus s_i which is a function of K, $(n^1(t), n^2(t), ..., n^N(t))$ and $(a^1(t), a^2(t), ..., a^M(t))$. This is

a birth-death model for which birth and death rates depends on s_i: $b(s_i)$ and $d(s_i)$. For each TCR:

$$s_i = \sum_{j=1}^{M} K_{ij} F_j a_j, \tag{3.23}$$

where F_j quantifies availabilities of antigens for TCR i and model competition: the more clones are specific to antigen j, the less available it will be. It is important to notice that there is no interaction between different species (different TCR) for this model class. To keep the explanations of this manuscript simple, we approximate F_j to 1 and we neglect the stochastic nature of division and death yielding the following Langevin equation (from Eq. 3.22) for each TCR:

$$\frac{dn_i}{dt} = \left(b + \sum_{j=1}^{M} K_{ij} a_j - d \right) n_i, \tag{3.24}$$

with $b(s_i) = b + \sum_{j=1}^{M} K_{ij} a_j$ and $d(s_i) = d$. Approximately the number of T cells composing the TCR repertoire to be constant on average, we add a source term modeling thymic input. Mathematically this is done through a source rate S of the number of clones per time unit with an introduction size N_0 drawn from a Poisson distribution. The stochastic process introduced in Eq. 3.24 has also adsorbing boundary conditions. For the system size not to reach infinity, we should also describe clonal extinction with absorbing boundary conditions. When the clone abundance n_i reaches 0 at time t, for all time τ for which $\tau \leq t$, $n_i(\tau) = 0$.

As the work presented in this thesis is data-driven, we want to exploit new available RepSeq data introduced in chapter 2 to understand how such models work in the presence or absence of stimuli. Unfortunately, data does not allow us to learn the form of the large random matrix K. To simplify Eq. 3.24 into a simpler model for which the parameters can be learned from available data, we introduce the notion of fitness $\mathcal{F}$. Fitness is a term coming from evolution quantifying the ability of a species to reproduce efficiently.

As introduced in Desponds et al. (2016), we can write Eq. 3.24, using the fitness definition:

$$\frac{dn_i}{dt} = \mathcal{F}_i(t) n_i(t), \tag{3.25}$$

with $\mathcal{F}(t) = \left(b + \sum_{j=1}^{M} K_{ij} a_j - d \right)$. $\mathcal{F}(t)$ also follows a stochastic process. In this thesis, we study the abundance dynamics of TCR clones in the absence or presence of a strong stimulus. As we assume that we have no a priori knowledge of the trajectories of antigen concentrations and that possible bursts of antigen concentration can be neglected for long-time studies, we write $\mathcal{F}(t)$ as a Brownian motion:

$$\frac{d\mathcal{F}_i}{dt} = f_0 + \frac{1}{\sqrt{\theta}}\eta_i(t), \tag{3.26}$$

with $\eta_i(t)$ a white noise of average $\langle \eta_i(t) \rangle = 0$ and time correlations $\langle \eta_i(t)\eta_i(t') \rangle = \delta(t - t')$, $f_0(t - t_0)$ the average value of the fitness $\mathcal{F}$ and θ the time scale of the fluctuations of $\mathcal{F}$.

Geometric Brownian motion

In our model, we consider that each clone interacts with the environment that exhibits a force that does not depend on the value of the variable n_i at time t. The fitness follows a Brownian motion, and we can substitute Eq. 3.26 in Eq. 3.25:

$$\frac{dn_i}{dt} = \left(f_0 + \frac{1}{\sqrt{\theta}}\eta_i(t) \right) n_i(t). \tag{3.27}$$

Performing Ito's change of variable, we obtain the following process for the log-abundance:

$$\frac{d \ln n_i(t)}{dt} = -\frac{1}{\tau} + \frac{1}{\sqrt{\theta}}\eta_i(t), \tag{3.28}$$

with $f_0 = -1/\tau + 1/2\theta$.

For each TCR clone i:

$$n_i(t) = n_i(t_0) \exp\left(-\frac{t}{\tau} + \frac{1}{\sqrt{\theta}} W_i(t) \right), \tag{3.29}$$

with W_i, the so-called Wiener process is equivalent to a Brownian motion, and t_0 is the initial time we start looking at the process.

We can then assume that:

$$\ln \frac{n_i(t)}{n_i(t_0)} \sim \mathcal{N}\left(-\frac{(t - t_0)}{\tau}, \frac{(t - t_0)}{\theta} \right), \tag{3.30}$$

with $\mathcal{N}$, the normal distribution of average $-\frac{(t-t_0)}{\tau}$ and variance $\frac{(t-t_0)}{\theta}$. In such settings, each clone abundance grows exponentially following the equation:

$$n_i(t) = n_i(t_0) \exp\left(-\frac{t}{\tau} \right) \tag{3.31}$$

with probability one if we observe the clonotype abundance for a long time. The expectation value of each clonotype abundance grows exponentially following the equation:

$$\langle n_i(t) \rangle = \langle n_i(t_0) \rangle \exp\left(-\frac{(t - t_0)}{\tau} + \frac{(t - t_0)}{2\theta} \right). \tag{3.32}$$

These previous statements can give us insights in modeling turnover time-scales

of TCR repertoires. τ can be considered to describe TCR repertoire *diversity - richness* rate of turnover (i.e the typical time for which clonotypes are replaced). The second timescale displayed in Eq.3.32 is $(2\tau\theta)/(2\theta-\tau)$: it is the typical timescale for which the total number of cells $N_{cells}(t)$ decays. Indeed $N_{cells}(t)$ is driven by the value of the mean $\langle N_{cells}(t)\rangle$ dynamics that can be computed thanks to Eq. 3.32 .

Steady-state distribution

To compare with empirical TCR clonal abundance distributions, we should search for the steady state solution of the previously defined geometric Brownian motion for which we add a source term. The Fokker-Planck equation associated to the geometric Brownian motion with a source term is then the following:

$$\partial_t \rho(\ln n, t) = \frac{1}{\tau}\partial_{\ln n}\rho(\ln n, t) + \frac{1}{2\theta}\partial^2_{\ln n}\rho(\ln n, t) + s(\ln n), \tag{3.33}$$

where $s(\ln n)$ is the source term describing the size of newly introduced clones, S is the introductory rate of thymus export, and N_0 is the regular introductory size of new clones. Assuming a constant initial size, $s(\ln n) = S\delta(\ln n - \ln N_0)$, we can compute the solution of this FP equation.

We compute the steady-state $\rho(\ln n)$ distribution solution of Eq. 3.33, with the absorbing condition: $\rho(\ln n = 0) = \rho(n = 1) = 0$. When a clonal abundance reaches an abundance of one, it goes extinct forever.

$$\rho(\ln n) = \begin{cases} \tau S(1 - n^{-\alpha}) & \text{if } \ln n < \ln N_0 \\ \tau S n^{-\alpha}(N_0^{\alpha} - 1) & \text{if } \ln n > \ln N_0, \end{cases} \tag{3.34}$$

with $\alpha = \frac{2\theta}{\tau}$. Performing a change of variable, we are able to compute the clonotype size distributions:

$$\rho(n) = \begin{cases} \frac{\tau S}{n}(1 - n^{-\alpha}) & if\ n < N_0 \\ \tau S(N_0^{\alpha} - 1)n^{-\alpha-1} & if\ n > N_0 \end{cases} \tag{3.35}$$

We obtain the scaling $\rho(n) \propto n^{-\alpha-1}$. This scaling observation was found in numerous data-sets from different studies in Weinstein et al. (2009); Britanova et al. (2016); Emerson et al. (2017). As shown in chapter 2, the power-law exponent $-\alpha$ is found to be conserved among more than 1000 individuals and close to 1.

Using these distributions, we can compute the total number of cells $N_{cells-ss}$ and of clonotypes $N_{clones-ss}$ at steady-states as functions of the parameters, τ, α, and S.

$$N_{cells-ss} = \int_0^{\infty} d\ln n\ \rho(\ln n)n = \frac{2S(N_0 - 1)\tau\theta}{2\theta - \tau} \tag{3.36}$$

$$N_{clones-ss} = \int_0^{\infty} d\ln n \rho(\ln n) = \tau S \ln N_0 \tag{3.37}$$

Rates of extinction and introduction of clonotypes

To understand better the steady-state of our system and for simulations purposes, it is useful to compute the extinction rate from the steady state values of $N_{cells-ss}$, $N_{clones-ss}$.

$$\begin{cases} \frac{dN_{clones}}{dt} = rN_{clones} + S \\ \frac{dN_{cells}}{dt} = \left(-\frac{1}{\tau} + \frac{1}{2\theta}\right) N_{cells} + SN_0, \end{cases} \tag{3.38}$$

where

$$r = -\frac{1}{\tau \ln N_0} \tag{3.39}$$

When simulating repertoire dynamics and sampling clone sizes from these distributions, we use the number of clonotypes that have a size smaller or larger than the introductory size N_0 ($N_{clones\ N<N_0}$ and $N_{clones\ N>N_0}$). We are also interested in computing the number of cells that belong to the clones with a size smaller or larger than the introduction one ($N_{cells\ N<N_0}$ and $N_{cells\ N>N_0}$):

$$N_{cells\ N<N_0} = \tau S(N_0 - 1) - \frac{\tau S}{1-\alpha} N_0(N_0^{-\alpha} - 1), \tag{3.40}$$

$$N_{cells\ N>N_0} = \frac{\tau S}{1-\alpha} N_0(N_0^{-\alpha} - 1), \tag{3.41}$$

$$N_{clones\ N<N_0} = \tau S \ln N_0 - \frac{\tau S}{\alpha}(1 - N_0^{-\alpha}), \tag{3.42}$$

$$N_{clones\ N>N_0} = \frac{\tau S}{\alpha}(1 - N_0^{-\alpha}). \tag{3.43}$$

Propagator of the dynamic

To generate synthetic data and build a model to predict stochastic population dynamics of TCR repertoires, we need to determine the propagator of Eq. 3.33.

The solution of Eq.3.33 is simply the propagator associated with a geometric Brownian motion:

$$G(\ln n_2, t_2 | \ln n_1, t_1) = \sqrt{\frac{\theta}{2\pi\Delta t}} e^{-\frac{\theta\left(\log n_2 - \log n_1 - \frac{\Delta t}{\tau}\right)^2}{2\Delta t}}. \tag{3.44}$$

3.3. Inference tools

Thanks to a breakthrough in TCR RepSeq sequencing described in chapter 2, it is possible to make the most of the availability of new and numerous RepSeq data to learn about TCR repertoire dynamics. One of the goals of this thesis is to be able to infer models introduced in the previous section in the context of TCR response to an acute stimulus over a time scale of the immune response (a few days) or to understand the stochastic population dynamics of such a system over a few years. The goal is to understand the *physical* forces shaping the TCR immune repertoire for a healthy individual. We use Bayesian (or probabilistic) modeling to learn dynamic models from data. The fundamental notions behind inference are introduced in this section. For more details about Bayesian inference and its applications for biological data, one can read Bishop (2007), and Durbin et al. (1998).

3.3.1 Bayesian inference

Bayesian statistical inference uses many observations found in data to learn a probability distribution that describes a physical phenomenon most accurately within a chosen class of models. The candidates for the probability distribution can be inspired by physical or mathematical modeling from the natural and experimental laws shaping the data. The input is the data (or observations) and the parameters describing the probability distribution we want to decipher. The goal is then to find a procedure to compute the most suitable values of the parameters to explain and interpret the data faithfully.

Using Bayesian statistics to understand a system is equivalent to using probabilities as a measure of our belief about the studied phenomenon, in contrast to the frequentist approach for which probabilities are defined as occurrence frequencies of random variables. The degree of belief of an event x is then translated into the probability $P(x)$. This degree can be conditioned by the accuracy of another random variable y to be known or verified: $P(x|y)$. The relation between these two beliefs can be established thanks to Bayes theorem.

Coming back to our original problem of modeling a physical phenomenon thanks to data and probability theory, we can define the probability definition $P(\mathcal{D}|\theta)$, with $\mathcal{D}$, the data, or observations contained in the sample. These observations can be written mathematically as vectors of dimensions depending on the nature of the data. In this manuscript, observations are denoted as $\hat{\mathcal{N}}$. They define the abundances of every TCR clone present in the analyzed biological RepSeq samples at T multiple times. This is what is called a longitudinal data set. The matrix $\hat{\mathbb{N}}$ is therefore of size $N_{obs} \times T$, with N_{obs} the number of clones we analyze. Each observation i would be in this case the clone i abundance trajectory $\hat{\mathcal{N}}_i = \left(\hat{n}^i_{t_1}, \hat{n}^i_{t_2}, ..., \hat{n}^i_{t_T}\right)$.

θ represents the parameters of the distribution to learn. θ is also a vector of multiple entries, for which the dimension depends on the number of parameters contained in the statistical model to optimize from the observations.
If we write Bayes theorem with these probabilities, we have:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}, \tag{3.45}$$

with $P(\mathcal{D}|\theta)$ the likelihood of the model (the probability to observe $\mathcal{D}$ given the values of the parameters, $P(\theta)$, the prior knowledge on the parameters without having any information about the data, $P(\theta|\mathcal{D})$, the posterior distribution. $P(\theta|\mathcal{D})$ is the best distribution of guesses one can make about the parameter distributions using the knowledge contained in data. $P(\mathcal{D})$ is the evidence distribution, i.e., the probability to generate this data integrating (or summing) over all possible value of θ.
$P(\mathcal{D})$ can be recovered by summing over all possible values that can θ take:

$$P(\mathcal{D}) = \int d\theta P(\mathcal{D}, \theta) \tag{3.46}$$

3.3.2 Maximum Likelihood Estimation - MLE

The goal of Bayesian inference is to uncover the parameters θ^{true} describing the best the inquire physical phenomenon thanks to partial information contained in the observations $\mathcal{D}$. To obtain values for θ, we want to maximize θ over the posterior distribution $P(\theta|\mathcal{D})$ which is a function of θ and takes observations as inputs. In this case, θ^* is called the maximum a posterior estimate (MAP) and is defined:

$$\theta^* = \underset{\theta}{\text{argmax}}\ P(\theta|\mathcal{D}) = \underset{\theta}{\text{argmax}}\ \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}, \tag{3.47}$$

with θ^*, the result of the Bayesian inference task.

In the limit of large numbers of observations N_{obs}, we should recover that:

$$\theta^{\text{true}} = \lim_{N_{obs}\to\infty} \theta^*(N_{obs}). \tag{3.48}$$

MAP estimation is closely linked to maximum likelihood estimation (MLE), for which we maximize the likelihood $P(\mathcal{D}|\theta)$ vs θ to obtain θ^*:

$$\theta^*_{MLE} = \underset{\theta}{\text{argmax}}\ P(\mathcal{D}|\theta). \tag{3.49}$$

MLE can be seen as a particular case of MAP estimation for which there is no prior knowledge on the model ($P(\theta)$ is a constant). If no prior information is known, $P(\theta)$ describes a uniform distribution and does not depend on the value of θ. As the evidence does not depend on the value of θ neither, $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)$ (Eq. 3.45).

In this manuscript, all parameter estimations using Bayesian inference are done using MLE.

3.3.3 The likelihood of the model

As explained in the previous subsection 3.3.2, we will maximize the likelihood of a model in this manuscript every time we want to learn a model from data $\mathcal{D}$. The first challenge is to analytically express the global likelihood of the entire data sample from single observations. In this manuscript, we consider that every observation is drawn independently from the others, which enables us to express the global likelihood from the likelihoods of single observations:

$$P(\hat{\mathbb{N}}|\theta) = \prod_{i}^{N_{obs}} P(\hat{\mathcal{N}}_i|\theta). \tag{3.50}$$

MLE can be done maximizing the log-likelihood quantity:

$$\theta^*_{MLE} = \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N_{obs}} \ln\left(P(\hat{\mathcal{N}}_i|\theta)\right). \tag{3.51}$$

In our applications, we will constrain the values each $\hat{\mathcal{N}}^i$ can take: $\hat{\mathcal{N}}^i \in \text{cond}$ The new likelihood to optimize is slightly modified, adding a new normalization term to the initial probability:

$$P(\hat{\mathbb{N}}|\theta, \hat{\mathcal{N}} \in \text{cond}) = \frac{P(\hat{\mathbb{N}}, \hat{\mathcal{N}} \in \text{cond}|\theta)}{P(\hat{\mathcal{N}} \in \text{cond}|\theta)}. \tag{3.52}$$

In practice, we expressed this conditioned MLE as

$$\theta^*_{MLE} = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i}^{N_{obs}} \ln\left(P(\hat{\mathcal{N}}_i|\theta)\right) - N_{obs} \ln\left(P(\hat{\mathcal{N}}^i \in \text{cond}|\theta)\right) \right]. \tag{3.53}$$

There are two main challenges when doing MLE of such a system. On the one hand, one needs to be able to compute numerically and efficiently the single log-likelihood for each observation which requires a lot of numerical computations on the other hand. One needs to optimize a multivariate function whose convexity is not a priori known. One must reach a global minimum to have a reliable estimate of θ from data.

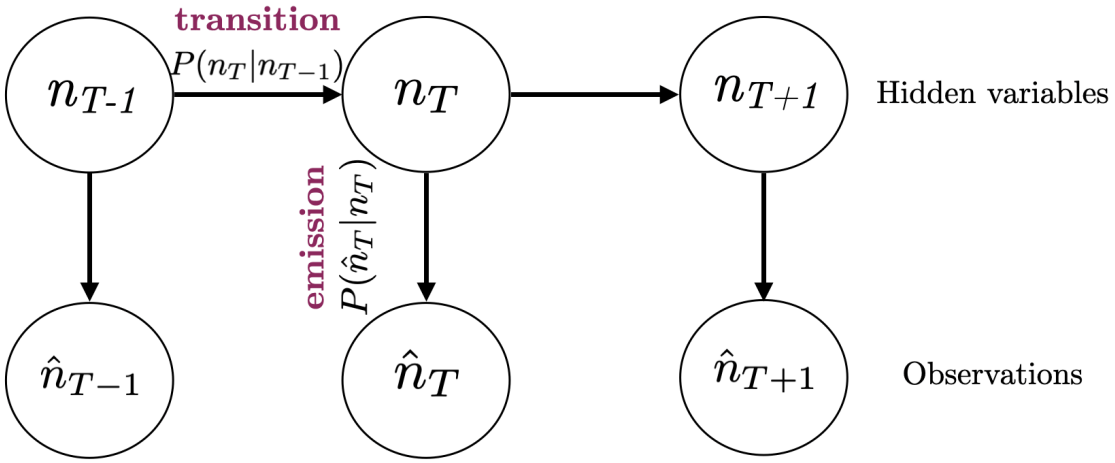

Figure 3.2: Graph-visualization of a hidden-Markov model for which the hidden states are the true clone abundances n_T, and the observations are the number of experimental RepSeq reads for the associated clones, $\hat{n}_T$.

3.3.4 Hidden variables, continuous hidden-Markov Model

Let us return to our problem of interest: inferring TCR repertoire dynamics. $\mathcal{D}$ is the data we use in our Bayesian framework to understand how TCR abundances vary with time. As mentioned in the first chapter, data and observations describing the best TCR abundances are generated using different RepSeq technologies. Counting T cells belonging to the same clones can be done differently depending on the technique. Because of the difference between empirical TCR clone abundances and the actual ones due to sampling (chapter 2) and the intrinsically noisy way to produce data, data $\mathcal{D} = \hat{\mathbb{N}} = \left(\hat{\mathcal{N}}_1, ..., \hat{\mathcal{N}}_{N_{obs}}\right)$, empirical abundances found in a RepSeq sample are a priori insufficient to proceed to Bayesian inference of a model . To calculate the likelihood, we need to marginalize out the latent variable $\mathbb{N} = (\mathcal{N}_1, ..., \mathcal{N}_{N_{obs}})$, which are the actual abundances, or hidden variables. The complete likelihood $P(\hat{\mathbb{N}}, \mathbb{N}|\theta)$ is analytically tractable, and learning this distribution is the goal of the statistical inference task. We will use $\mathbb{F} = (\mathcal{F}_1, ..., \mathcal{F}_{N_{obs}}) = \left(\mathcal{N}_1/N_{cells}, ..., \mathcal{F}_{N_{obs}/N_{cells}}\right)$ or $\mathbb{N}$ depending on the application.

Hidden-Markov model for TCR abundance dynamics

A hidden Markov model is a Bayesian probabilistic model representation to learn sequence data known to follow a Markov process. As defined in 3.2.1, a Markov process is a memoryless process for which we have $P\left(n_{t_T}|n_{t_1}, n_{t_2}, ..., n_{t_{T-1}}\right) = P\left(n_{t_T}|n_{t_{T-1}}\right)$. We make the assumption that the true TCR abundances follow Markov dynamics as a first approximation. The true abundances $\mathcal{N}$ are the unobservable hidden states or variables and follow a Markov process. $P\left(n_{t_T}|n_{t_{T-1}}\right)$ is called the *transition probability* which transcribes the physical dynamic in the inference model. For

simplicity and consistency with the rest of the manuscript, we are going to use the continuous variables TCR frequencies f which are the normalized TCR abundances n: $f = n/N_{cell}$, N_{cells} the total number of T cells in the repertoire. The observations $\hat{\mathcal{N}}$ are emitted by the true variables $\mathcal{F}_T = (n_{t_1}/N_{cell}, n_{t_2}/N_{cell}, ..., n_{t_T}/N_{cell})$. The generation process from the true values of frequencies $\mathcal{F}_T$ and the observations $\hat{\mathcal{N}}_T$ contained in the data are encoded in the *emission probabibilty* $P(\hat{n}|f)$ at every time point of the process. As a tool of visualization, the hidden-Markov model is represented in Fig. 3.2. Solving this hidden-Markov model means finding the transition and emission probabilities as well as the complete likelihood $P(\hat{\mathbb{N}}, \mathbb{N}|\theta)$. The complete likelihood is

$$P(\hat{\mathcal{N}}, \mathcal{F}|\theta) = \rho(f_1) \prod_{j=1}^{T-1} P(\hat{n_j}|f_j) P(f_{j+1}|f_j, \theta), \tag{3.54}$$

with f_{min}, the theoretical lower-bound of the TCR frequency values. The likelihood of the observations consists of integrating the complete likelihood $P(\hat{\mathcal{N}}, \mathcal{N}|\theta)$ over all range values:

$$P(\hat{\mathcal{N}}|\theta) = \int_{f_{min}}^{1} ... \int_{f_{min}}^{1} df_1 ... df_T \rho(f_1) \prod_{j=1}^{T-1} P(\hat{n_j}|f_j) P(f_{j+1}|f_j, \theta). \tag{3.55}$$

Computing $P(\hat{\mathcal{N}}, \mathcal{F}|\theta)$ and $P(\hat{\mathcal{N}}|\theta)$ can be a challenge to realize numerically. Let us remember that statistical inference aims to find the best suitable θ parameters (with our observations) that are the MLE θ^{MLE}. One can compute the function and optimize it numerically using tools inspired by gradient descent to maximize the likelihood or the log-likelihood. But there are more efficient numerical techniques to compute the likelihood of the model. Also, knowing that we have reached a global maximum over the optimization step is not always clear. There are some solutions to overcome this when the likelihood is not easy to represent and is not systematically concave.

Compute the likelihood using the forward algorithm

As mentioned in the previous paragraph, one needs to find a way to compute numerically the log-likelihood or the likelihood of the model in a reasonable amount of time despite the numerous integrations. One trick is to use the forward algorithm, a dynamic programming algorithm based on recursive relations between probability of the observed sequences up to and including f_t

$$g_t(f_t) = P(\hat{\mathcal{N}} = (\hat{n}_1, \hat{n}_2, ..., \hat{n}_t\ , f_t), \tag{3.56}$$

which translates into:

$$g_t(f_t) = \int_{f_{min}} df_1..df_{t-1}\rho(f_1)P(\hat{n}_1|f_1)\prod_{j=2}^{t} P(\hat{n}_j|f_j)P(f_j|f_{j-1},\theta). \tag{3.57}$$

We have the recursive relation :

$$g_{t+1}(f_{t+1}) = \int_{f_{min}} df_t g_t(f_t)P(\hat{n}_{t+1}|f_{t+1})P(f_{t+1}|f_t,\theta). \tag{3.58}$$

The last step, $T-1$, will return the likelihood of the model:

$$P(\hat{\mathcal{N}}|\theta) = \int_{f_{min}} df_T g_T(f_T). \tag{3.59}$$

Before going into the details of the EM algorithm, Let us define two important notions of statistical inference: the Shannon entropy $S[p]$ associated with the probability distribution p

$$S[p(x)] = -\int p(x)\ln[p(x)], \tag{3.60}$$

and the Kullback-Leibler $KL[p||q]$ divergence between two probability distributions p and q

$$KL[p(x)||q(x)] = \int p(x)\ln\frac{p(x)}{q(x)}. \tag{3.61}$$

The Kullback-Leibler divergence measures how the probability distribution p differs from another probability distribution q. KL divergence has two important properties: it is always positive, and if $p = q$, the KL is equal to 0. KL divergence is not a distance as it is not symmetric.

Let us come back to the EM algorithm. When dealing with hidden variables, the (EM) algorithm is an excellent tool for complicated likelihood optimization. EM is an optimization method ensuring maximizing the log-likelihood quickly and precisely. We introduce an auxiliary probability distribution q over the hidden variables $\mathcal{F}$. We can write the following equality:

$$\ln P(\hat{\mathcal{N}}|\theta) = \int d\mathcal{F}q(\mathcal{F})\ln\frac{P\left(\hat{\mathcal{N}},\mathcal{F}|\theta\right)}{q(\mathcal{F})} - \int d\mathcal{F}q(\mathcal{F})\ln\frac{P\left(\mathcal{F}|\hat{\mathcal{N}},\theta)\right)}{q(\mathcal{F})}. \tag{3.62}$$

It is simple to derive the previous equality using $\ln P(\hat{\mathcal{N}}|\theta) = \ln P(\hat{\mathcal{N}},\mathcal{F}|\theta) - \ln P(\mathcal{F}|\hat{\mathcal{N}},\theta)$, Bayes formulae and $\int d\mathcal{F}q(\mathcal{F}) = 1$.
In practice, the EM algorithm is an iterative process with multiple steps ensuring convergence of the log-likelihood toward a global maximum. Going from step t to step $t+1$, we update the value of θ from θ^t to θ^{t+1} until a chosen criteria for this optimization procedure is respected. To do so, the most common strategy is to

chose the function q to be equal to $P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)$ and $P\left(\hat{\mathcal{N}},\mathcal{F}|\theta\right) = P\left(\hat{\mathcal{N}},\mathcal{F}|\theta^{t+1}\right)$ in order to have a log-likelihood function $\mathcal{L}(\theta^{t+1}|\theta^t)$ to maximize with respect to θ^{t+1}, at every step :

$$\mathcal{L}(\theta^{t+1}|\theta^t) = \int d\mathcal{F}P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)\ln P\left(\hat{\mathcal{N}},\mathcal{F}|\theta^{t+1}\right) - \int d\mathcal{F}P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)\ln\frac{P\left(\mathcal{F}|\hat{\mathcal{N}},\theta^{t+1}\right)}{P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)}$$
$$+\int d\mathcal{F}P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)\ln P(\mathcal{F}|\hat{\mathcal{N}},\theta^t). \tag{3.63}$$

with θ^t and θ^{t+1} the values of the parameters at steps t and $t+1$ of the algorithm.

We recognize in Eq. 3.63, the Kullback-Leibler divergence, $KL[P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)||P(\mathcal{F}|\hat{\mathcal{N}},\theta^{t+1})]$ and the Shannon entropy $S[P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)]$. We also define the $\mathcal{Q}-$function as $\mathcal{Q}(\theta^{t+1},\theta^t) = \int d\mathcal{F}P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)\ln P\left(\hat{\mathcal{N}},\mathcal{F}|\theta^{t+1}\right)$ which can be interpreted as the average quantity of the log likelihood $\ln\left[P(\hat{\mathcal{N}},\mathcal{F}|\theta)\right]$ weighted by the posterior of the hidden variable at the previous step t, for which $\theta = \theta^t$,

$$\mathcal{L}(\theta^{t+1}|\theta^t) = \mathcal{Q}(\theta^{t+1},\theta^t) - KL[P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)||P(\mathcal{F}|\hat{\mathcal{N}},\theta^{t+1})] + S[P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)]. \tag{3.64}$$

Because the Kullback-Leibler divergence $KL[P(\mathcal{F}|\hat{\mathcal{N}},\theta^t)||P(\mathcal{F}|\hat{\mathcal{N}},\theta^{t+1})] \geq 0$, the quantity $\mathcal{Q}+S$ in Eq. 3.64 (which is negative) is a lower bound on the log-likelihood function $\mathcal{L}(\theta^{t+1}|\theta^t)$. The quantity S does not depend on the variable θ^{t+1}. We conclude then that maximizing the log-likelihood $\mathcal{L}(\theta^{t+1}|\theta^t)$ with respect to θ^{t+1} is the same as maximizing $\mathcal{Q}(\theta^{t+1},\theta^t)$ with respect to θ^{t+1}.
The EM algorithm is a 2-stage iterative optimization procedure for finding maximum likelihood estimates. The expectation E-step consists of evaluating the $\mathcal{Q}$ function for a given value of θ^t with respect to θ^{t+1}. Computing the E-step is not always straightforward and can be done using multiple strategies, for instance, simulation-based techniques Bishop (2007). The M-step consists of maximizing the $\mathcal{Q}$ function with respect to θ^{t+1} to update the value of θ. These two steps are then repeated until convergence of the $\mathcal{Q}$ function.

In this manuscript, I have not used the forward algorithm or the Expectation-Maximization algorithm to compute the likelihood of the model in the following chapters, where I proceed only with pairs of time points. Still, these tools helped me better understand data and try a model that has not resulted in convincing results.

3.3.5 Computing errors on parameter estimations

When the log-likelihood optimization is achieved, and we obtained a numerical value, it is important to assess the error on the estimation of θ^{*MLE}. Here we use θ^{*MLE} and not θ^{*MAP} as we have a flat prior $P(\theta)$. If the posterior $P(\theta|\mathcal{D})$ can be computed, the model should give us θ^{MLE} (in the case of no prior $P(\theta)$). The error should be proportional to the variance associated with this distribution and inversely proportional to the number of observations (if we remember Eq. 3.48). Let us develop in θ^{*MLE} to second order the log-likelihood – $\ln P(\mathcal{D}|\theta) = \mathcal{L}(\theta)$ – first for a multivariate form (θ is a vector or several parameters):

$$\mathcal{L}(\theta) = \mathcal{L}(\theta^{*MLE}) + \frac{1}{2}(\theta - \theta^{*MLE})^{\mathrm{T}}\mathbb{H}(\theta^{*MLE})(\theta - \theta^{*MLE}) + o(||\theta - \theta^{*MLE}||^2), \quad (3.65)$$

with $\mathbb{H}$, the Hessian matrix associated with the log-likelihood function $\mathcal{L}$. If the optimization has been correctly performed, we have that $\nabla\mathcal{L}|_{\theta^{*MLE}} = 0$, and that is why this term does not appear in the previous equation Eq. 3.65. From this result, we have that:

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta) \propto \exp^{-\frac{1}{2}(\theta-\theta^{*MLE})^{\mathrm{T}}\mathbb{H}(\theta^{*MLE})(\theta-\theta^{*MLE})}, \quad (3.66)$$

at second order. Then computing the Hessian matrix of the log-likelihood and evaluating it at θ^{*MLE} gives us the Covariance Matrix of the Gaussian distribution defining $P(\theta|\mathcal{D})$, and so errors on all parameters composing the vector θ.

The entries in the column vector are $\mathbb{X} = (X_1, X_2, ..., X_T)$, with $\forall t \in [1, T]$, X_t a random variable. We define the entries of the covariance matrix (l, m) $\mathbb{C}_{XX}$ as

$$\mathbb{C}_{X_l X_m} = E[(X_l - E[X_l])(X_m - E[X_m])], \quad (3.67)$$

where E denotes the expectation of a random variable.

CHAPTER

4

INFERRING T-CELL REPERTOIRE DYNAMICS FROM HEALTHY INDIVIDUALS

In this chapter, I introduce the work discussed in the preprint Bensouda Koraichi et al. (2022), which is the core research of this thesis. We study the neutral dynamics that drive the immune system without acute and intense stimuli for healthy individuals. This analysis focuses on data-driven approaches taking advantage of longitudinal immune repertoire sequencing (RepSeq) data availability. We quantify the experimental noise and learn the long-timescale dynamics of TCR repertoires for healthy people. Applying Bayesian inference, we infer parameters that enable us to quantify for the first time the turnover dynamics of TCR repertoires as a function of the age of the individual.

4.1. Introduction and motivation

The adaptive immune system protects us from many infections including those caused by pathogenic challenges that did not exist when we were born. This amazing plasticity is encoded, in part, in a diverse repertoire of T cells carrying surface receptors capable of recognizing different antigens, which trigger an immune response. About 10^8 new T cells are estimated to be generated and enter the periphery in human adults every day Yates (2014); Bains et al. (2009b), where they undergo specific proliferation due to antigen stimulation but also non-specific divisions Jameson (2002); Dowling and Hodgkin (2009) and death. These processes together result in clone sizes of different T cells that differ over a few orders of magnitude, forming

long tailed distributions Desponds et al. (2016); Mora and Walczak (2019b). The total number of different T cell clones is estimated between 10^8 and 10^{10} Qi et al. (2014); Lythe et al. (2016); Mora and Walczak (2019a). Qualitatively describing the T cell clonal dynamics in the periphery is important for predicting long- as well as short-term immune response and to understand the maintenance of immune memory.

A lot of effort has been put into describing antigen specific response and memory formation De Boer et al. (2003); Kedzierska et al. (2012); Mayer et al. (2019). At any given timepoint the majority of the T cell repertoire is not always directly involved in fighting the current antigenic challenge. Yet, processes such as homeostasis Jameson (2002) and unspecific signals in both naive and memory subrepertoires result in frequency changes of background clones. Many first-principles models of naive T cells dynamics have been proposed to study the balance between thymic output and peripheral proliferation and death Bains et al. (2009b,a); Dowling and Hodgkin (2009); de Greef et al. (2020); Dessalles et al. (2022). The role of competition for antigens between T cells has been pointed out De Boer and Perelson (1994), as well as the effect of cross-reactivity Dash et al. (2017) (the ability for one T cell to recognize different antigens), the relative size of a primary versus secondary response to similar antigens Mayer et al. (2019), or the effect of heritable changes affecting the homeostatic rate of thymic exports Johnson et al. (2012). These studies highlight the importance for the naive repertoire of clonal expansions that are not necessarily linked to specific challenges. While these models were instrumental in advancing our understanding of bulk repertoire dynamics, and allowed for the interpretation of deuterated water and bromide staining experiments that describe cellular lifetimes De Boer and Perelson (2013), the class of models that are consistent with the data is still large and unexplored.

Thanks to advances in immune repertoire sequencing (RepSeq) Lindau and Robins (2017); Davis and Boyd (2019); Minervina et al. (2019b), dynamical models can now be assessed directly against repertoire data at the clonal level. RepSeq experiments isolate and sequence the T cell receptors (TCR) in a blood sample of individuals. By counting reads with the same TCR sequence, one can estimate the frequency of the corresponding clone (defined as the set of cells with the same receptor) in blood. Even single repertoire snapshots can be informative about the dynamics: the distribution of clone sizes follows a power law Burgos and Moreno-Tovar (1996); Koch et al. (2018); Naumov et al. (2003); Zarnitsyna et al. (2013); Mora and Walczak (2019b), in accord with proposed models of stochastic growth and death Desponds et al. (2016). Taking samples from the same individual at different time points allows for tracking the evolution of TCR clone sizes in time. The longitudinal experiments that have been performed in healthy donors Britanova et al. (2016); Chu et al. (2019) suggest that the repertoire is relatively stable over

years.

Our main goal in this article is to characterize the dynamics of the unstimulated background repertoire. We use an inverse approach to learn models of stochastic TCR clonal dynamics directly from data, collecting human TCR RepSeq datasets where we could identify at least two time points between which there was no reported specific acute antigenic stimulation Britanova et al. (2016); Sycheva et al. (2018); Pogorelyy et al. (2018c); Chu et al. (2019); Minervina et al. (2020). A key aspect of our method is the treatment of experimental noise, which confounds naive analyses of stochastic time traces. The method first quantifies both the sampling and natural biological noise thanks to replicate RepSeq experiments Puelma Touzel et al. (2020); Bensouda Koraichi et al. (2021) , and then infers the parameters of a stochastic dynamical model to describe the trajectories of each TCR clone population in a healthy individual, i.e. who did not have medical conditions or known infections during the sampled interval. We explicitly show how correcting for noise allows us to robustly learn the underlying dynamics. A recent study Gaimann et al. (2020) has investigated the formation of the T cell repertoire during development and its maintenance into adulthood. Here we focus on healthy adult repertoires that are already shaped during the first years of an individual's life and ask how they evolve and get renewed. We extract *clonal* (and not cellular) turnover time scales, and describe how these time scales depend on the person's age. Characterizing these baseline dynamics is an important step towards interpreting TCR dynamics in the presence of antigenic stimuli.

4.2. Results

4.2.1 Longitudinal sampling of TCR repertoires of healthy individuals

T-cell repertoires are large ecosystems in which each species is a clone of T cells carrying the same TCR i formed by a unique pair of α and β chains. The dynamics of this system is characterized by the time course of the number of cells carrying each receptor, $n_i(t)$. This number can be accessed indirectly through TCR repertoire sequencing (RepSeq), obtained by sequencing the TCR of small samples of peripheral blood mononuclear cells (PBMC), giving us a read count $\hat{n}_i(t)$ for a given chain at different timepoints (Fig. S1A). Because the two chains are not paired in the data, from here on we define clones as collections of cells having the same α or β chain, which we will refer to as clonotypes. This approximation is justified by the low occurence of TCRs that share one chain but not the other Minervina et al. (2020).

We collected repertoire data from 9 individuals P1-P9, aged 18-57, sampled at various time points from one month up to 3 years apart, with and without biological

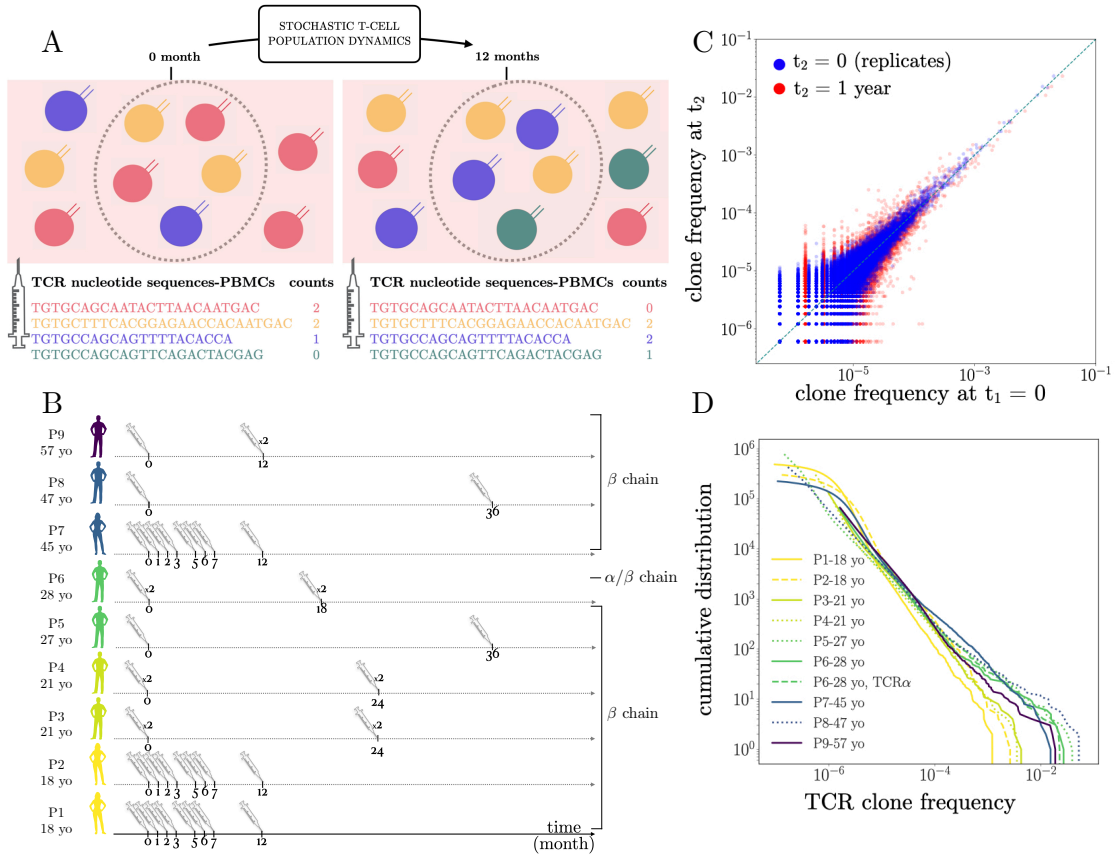

Figure 4.1: **Longituinal tracking of T-cell repertoires. A.** Experimental workflow. PBMC from a healthy individual are extracted at two timepoints, and their TCR repertoire sequenced, yielding lists of clonotypes with count numbers corresponding to the number of individual measurements (or reads). The way in which the two sampled repertoires has changed between the two timepoints is predicted by a stochastic model of the dynamics of T cell clones. **B.** Summary of the TCR α and β repertoire data used in this study. 9 individuals from 5 studies, aged 18-57, male and female, were included. When available, replicate experiments are annotated with $\times 2$. Datasets were produced using two different sequencing technologies based on cDNA and gDNA. **C.** Typical scatter plot of frequencies of TCR clones in two samples from the same individual P9. Blue: two biological replicates obtained on the same day show the effect of experimental nosie. Red: 2 samples taken 1 year apart show a larger spread, resulting from a combination of real changes and noise. The goal of the analysis is to disentangle real changes from the noise. **D.** Cumulative distributions of TCR frequencies, which follow a universal power law in all samples and donors, with exponent ≈ 1.

replicates. β chain repertoires were sequenced for all samples, and α chains only for individual P6. The properties of the datasets, including their number of clones N_c, total read counts $N_r = \sum_{i=1}^{N_c} \hat{n}_i$, age, library preparation (from genomic DNA or from mRNA), and chain, are summarized in Fig. S1B and in Table S1. Because P3, P4, P6, and P9 were included in a vaccination study, they had received a shot of the YFV 17D yellow fever vaccine (P3, P4, P6) or of the influenza vaccine (P9) 45 days prior to the first time point, after the decay of their T-cell response, so we assume that the dynamics of vaccine-specific T does not affect much our analysis of the global repertoire.

A major challenge when analyzing RepSeq data is that the measured abundances $\hat{n}_i(t)$ only provide a noisy reflection of the true ones $n_i(t)$. Observed differences between datasets thus result from a combination of the repertoire dynamics and biological and experimental noise. The magnitude of that noise can be assessed by comparing the normalized clonotype frequencies $\hat{f} = \frac{\hat{n}}{N_r}$ between two biological replicates obtained at the same time point in the same individual (Fig. S1C, blue dots). By contrast, comparing those frequencies between two timepoints separated by one year (Fig. S1C, red dots) show a larger dispersion, and a slight overall decrease of clonotype frequencies. Our goal is to measure this difference quantitatively.

Another difficulty arises from the observation that clonotype frequencies are highly heterogeneous, with their distribution following a power law $P(\hat{n}) \propto \hat{n}^{-1-\alpha}$ spanning no less than 4 orders of magnitude, with an exponent $\alpha \approx 1$ which is largely invariant across individuals and timepoints (Fig. S1D), as previously reported Mora and Walczak (2019b); Desponds et al. (2016); Gaimann et al. (2020). This implies that most clonotypes have very low abundance and are thus particularly subject to sampling and experimental noise.

4.2.2 Mathematical model of stochastic clonal dynamics

The dynamics of T cell clones is driven by the proliferation and death of cells belonging to them. In addition, new clones with their distinct TCR are continually produced and released by the thymus, although the rate of thymic exports decays rapidly with age Yates (2014). Cell division, death, and introduction of new clones constitute the basis of our model (Fig. 4.2A). Cell division may be caused by antigen stimulation (both self and foreign) or by cytokine and growth factors, and cells die by lack of stimulation or by apoptotic signals. Even in the absence of strong and chronic antigenic stimuli, T cell clonotype abundances display stochastic trajectories due to either weak stimulation, repertoire homeostasis and demographic fluctuations. In addition, individuals may get mild infections over the course of months and years. Since these events are numerous and unknown, we model them by an effectively random net growth rate (divisions minus deaths). It can be shown Desponds et al. (2016); Altan-Bonnet et al. (2020) that on time scales much longer

than the typical resolution time of infections, each clonotype size $n_i(t)$ may then be modeled by a geometric Brownian motion (GBM). Its evolution is governed by an effective mean net growth, to which random fluctuations are added to account for bursts of proliferation and decay:

$$\frac{d\ln n_i(t)}{dt} = -\tau^{-1} + \theta^{-1/2}\eta_i(t), \tag{4.1}$$

where $\eta_i(t)$ is a clonotype-specific white noise of zero mean and unit amplitude. Note that the mean growth rate of clones, $-\tau^{-1}$, is typically negative. On average, each clone should decay to make room for new thymic exports, because of homeostatic pressures that control the total number of cells. In this interpretation, τ is the typical decay and turnover time of each clone, which would evolve with time as $n_i(t) = n_i(0)e^{-t/\tau}$ in the absence of fluctuations. But recall that this is just an average—many clones do not decay, but instead undergo episodes of large growth and decay, as illustrated by simulations of (4.2) in Fig. 4.2B. The typical amplitude of these fluctuations grows with time as $\sqrt{t/\theta}$ (dashed lines). Thus, θ may be interpreted as the typical time it takes for a clone to rise or decay above or below the typical behaviour by one log-unit.

In addition to being biological motivated, the proposed dynamics have the desirable property that, in the presence of a constant rate of thymic exports, the distribution of clone sizes is predicted to evolve in time towards a perfect power law, $P(n) \propto n^{-1-\alpha}$, with exponent $\alpha = 2\theta/\tau$ given by twice the ratio of the two time scales of the model Desponds et al. (2016). This is illustrated in Fig. 4.2C on simulated repertoires at steady state, and agrees well with the empirical distributions of Fig. S1D.

Our goal is to capture the parameters of these dynamics that is informative about the repertoire turnover timescales, while constraining the experimentally observed clone size frequency distribution. Our approach assumes that on the timescales of the analysis we do not observe signals of strong and specific antigenic stimulation. It also ignores potential dependences on the size of the clone, which could be mediated by phenotypic differences between clones. This last assumption will be revisited later.

4.2.3 Mathematical model of stochastic clonal dynamics

The dynamics of T cell clones is driven by the proliferation and death of cells belonging to them. In addition, new clones with their distinct TCR are continually produced and released by the thymus, although the rate of thymic exports decays rapidly with age Yates (2014). Cell division, death, and introduction of new clones constitute the basis of our model (Fig. 4.2A).

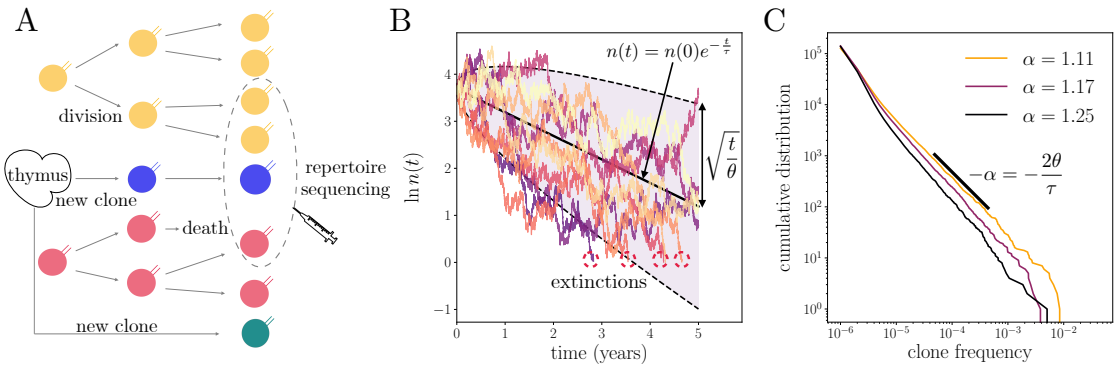

Figure 4.2: **Stochastic model of repertoire dynamics. A.** T cells are introduced in the peripheral immune system by thymic export, providing a source of new TCR clones. T cells belonging to a specific TCR clone (labeled by their color) divide and die depending on their interactions with the antigenic environment, increasing or reducing the abundance of its TCR in the repertoire. This process is modeled by a geometric Brownian motion **B.** Example traces of TCR abundances simulated from the model Eq. 4.2 with $n(0) = 40$, with $\tau = 2$ years and $\theta = 1.11$ year. Clones that reach abundance < 1 go extinct (red circles). The typical trend is for clones to decay exponentially with time scale τ (black solid line). Stochastic events of growth and decay account for a broad variability of individual traces, whose magnitude grows as $\sqrt{t/\theta}$ with time (shaded area) in logarithmic scale. **C.** Cumulative frequencies distributions of synthetic TCR clone abundances. The model predicts a power law of exponent $\alpha = 2\theta/\tau$. Different values of τ and θ were used to lead to different values of the exponent α. Parameters: $\tau = 2$ years, $N_{\text{cell}} = 10^{10}$, $n_0 = 40$.

Cell division may be caused by antigen stimulation (both self and foreign) or by cytokine and growth factors, and cells die by lack of stimulation or by apoptotic signals. Even in the absence of strong and chronic antigenic stimuli, T cell clonotype abundances display stochastic trajectories due to either weak stimulation, repertoire homeostasis and demographic fluctuations. In addition, individuals may get mild infections over the course of months and years. Since these events are numerous and unknown, we model them by an effectively random net growth rate (divisions minus deaths). It can be shown Desponds et al. (2016); Altan-Bonnet et al. (2020) that on time scales much longer than the typical resolution time of infections, each clonotype size $n_i(t)$ may then be modeled by a geometric Brownian motion (GBM). Its evolution is governed by an effective mean net growth, to which random fluctuations are added to account for bursts of proliferation and decay:

$$\frac{d\ln n_i(t)}{dt} = -\tau^{-1} + \theta^{-1/2}\eta_i(t), \tag{4.2}$$

where $\eta_i(t)$ is a clonotype-specific white noise of zero mean and unit amplitude. Note that the mean growth rate of clones, $-\tau^{-1}$, is typically negative. On average, each clone should decay to make room for new thymic exports, because of homeostatic pressures that control the total number of cells. In this interpretation, τ is the typical decay and turnover time of each clone, which would evolve with time as $n_i(t) = n_i(0)e^{-t/\tau}$ in the absence of fluctuations. But recall that this is just an average—many clones do not decay, but instead undergo episodes of large growth and decay, as illustrated by simulations of (4.2) in Fig. 4.2B. The typical amplitude of these fluctuations grows with time as $\sqrt{t/\theta}$ (dashed lines). Thus, θ may be interpreted as the typical time it takes for a clone to rise or decay above or below the typical behaviour by one log-unit.

In addition to being biological motivated, the proposed dynamics have the desirable property that, in the presence of a constant rate of thymic exports, the distribution of clone sizes is predicted to evolve in time towards a perfect power law, $P(n) \propto n^{-1-\alpha}$, with exponent $\alpha = 2\theta/\tau$ given by twice the ratio of the two time scales of the model Desponds et al. (2016). This is illustrated in Fig. 4.2C on simulated repertoires at steady state, and agrees well with the empirical distributions of Fig. S1D.

Our goal is to capture the parameters of these dynamics that is informative about the repertoire turnover timescales, while constraining the experimentally observed clone size frequency distribution. Our approach assumes that on the timescales of the analysis we do not observe signals of strong and specific antigenic stimulation. It also ignores potential dependences on the size of the clone, which could be mediated

by phenotypic differences between clones. This last assumption will be revisited later.

4.2.4 Model inference

We estimate the parameters (θ, τ) of the dynamics in Eq. (4.2) from the observed clonotype abundance trajectories using a Bayesian approach for the posterior distribution of parameters given the data:

$$(\tau^*, \theta^*) = \arg\max_{\tau,\theta} \prod_{i=1}^{N_c} P(\hat{n}_i(t_1), \hat{n}_i(t_2)|\tau, \theta), \tag{4.3}$$

where t_1 and t_2 are the times of the two samples.

We use two methods to learn the model parameters: *naive inference* and *full inference*. The naive inference assumes the empirical abundances faithfully represents the real clonal abundances n_i through a simple proportionality rule, $\hat{n}_i \approx (N_r/N_{\text{cell}})n_i$, where $N_{\text{cell}} = \sum_i n_i$ is the total number of T cells in the body. In practice, we work with clonotype frequencies $\hat{f}_i = \hat{n}_i/N_r$, and $f_i = n_i/N_{\text{cell}}$, so that this assumption becomes $\hat{f}_i = f_i$. Further assuming that the total number of cells N_{cell} is approximately constant in time at steady state, $\hat{f}_i$ is then governed by the same equation (4.2) as n_i. We take advantage of the closed solution available for the propagator associated to the GBM, (4.5), to maximize the log-likelihood (see Methods). This maximization is equivalent to plotting a histogram of the change in log-frequencies between the two timepoints, and simply read off τ^{-1} and θ^{-1} as the negative mean and the variance of the distribution divided by $t = t_2 - t_1$ (Fig. S2A), consistent with their biological interpretation.

The *full inference* incorporates the fact that the observed clonotype abundances are contaminated by biological (mRNA expression) and experimental noise sources (sequencing errors, stochastic PCR amplification, and sampling), which means they do not correspond exactly to the clonotype abundances. To give a sense of just the sampling issue, a PBMC sample of $\sim 1\,\text{mL}$ contains about 1 million cells, yielding about 1 million reads. By comparison, the organism contains of the order of 10^{11} T cells. TCR clonotype frequencies are thus extrapolated from observing a fraction $10^6/10^{11} \approx 10^{-5}$, or 0.001%, of the whole repertoire Mora and Walczak (2019a). In addition, not all cells are captured, and each cell may be represented by multiple reads, either through sequencing of multiple mRNA from the same cell, or from PCR amplication, depending on the context. To address these sources of uncertainty, in the full inference approach we introduce an error model Puelma Touzel et al. (2020) relating observed frequencies $\hat{f}$ to their true value f probabilistically through the transfer function $P(\hat{f}|f)$ (Fig. S2B). We use the previously introduced a software tool, NoisET Bensouda Koraichi et al. (2021), which learns such a noise model from

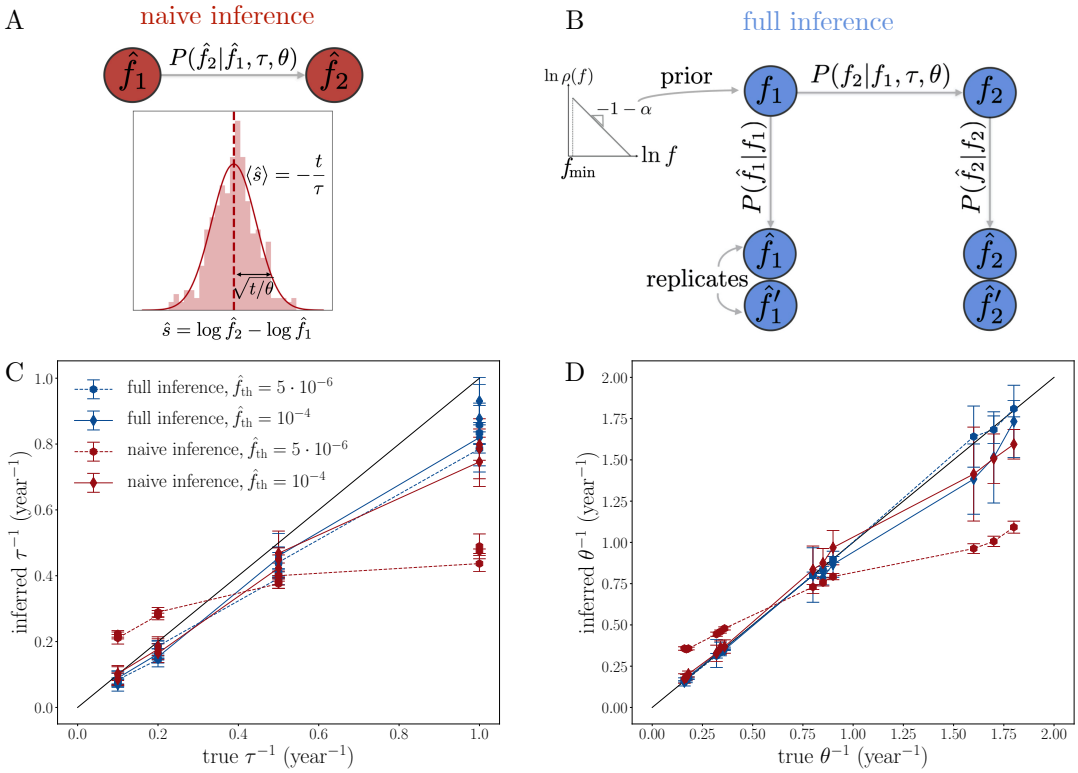

Figure 4.3: **Inferring dynamical parameters from data. A.** Naive inference. The empirical clonotype frequencies in the RepSeq samples at the two time points, $\hat{f}_1$ and $\hat{f}_2$, are treated as the true ones, f_1 and f_2. We estimate the two parameters τ and θ from the average and standard deviation over all observed TCR clones of the log-fold change in frequency between two-time points, $\log(\hat{f}_2/\hat{f}_1)$, which the model predicts is distributed normally. The distribution of $\hat{s}$ is shown in light red, and the Gaussian fit in solid red. **B.** Full inference. The empirical frequencies $\hat{f}$ are modeled as noisy read-outs of the true ones f, through a probabilistic noise model. First, the noise model $P(\hat{f}|f)$ is inferred from replicate experiments such as shown in Fig. S1C. The inference procedure also learns the distribution of frequencies $\rho(f)$, assumed to follow a power law with adjustable exponent α and minimal frequency $f_{\min}$. Second, using the noise model, the parameters of the dynamical propagator $P(f_2|f_1,\tau,\theta)$ are inferred from two timepoints, where f_1 and f_2 are treated as latent variables and $\hat{f}_1$, $\hat{f}_2$ as observables, using a Maximum likelihood estimator. **C-D.** Validation of naive and full inference models on synthetic data. Model parameters: $t_2 - t_1 = 2$ years, $\tau = 1, 2, 5, 10$ years, $\alpha = 1.11, 1.18, 1.25$, with all 12 combinations tested; number of cells $N_c = 10^{10}$; initial clone size $n_0 = 40$; the other parameters (number of clones, thymic output) are deduced assuming steady state (see Methods). Sampling model: number of sampled reads $N_r = 10^6$; noise model parameters $a = 0.7$ and $b = 1.1$. Error bars are standard deviations over 10 simulations.

replicate RepSeq experiments (see Methods).

We applied NoisET to individuals P3, P4, P6 and P9 for whom replicates were available. The noise model assumes that the read count $\hat{n}$ of each clone is drawn from a negative binomial distribution, whose variance grows with the frequency as $\mathrm{Var}(\hat{n}) = fN_r + a(fN_r)^b$, with two learnable parameters a, b. In addition, since true frequencies are unknown, we assume as a prior that frequencies are distributed according to a power law $\rho(f) \propto f^{-1-\alpha}$ with a cut-off $f > f_{\min}$, with α and $f_{\min}$ another two parameters. These parameters are reported for all individuals and time points in Fig. S1.

Once the noise model has been learned using NoisET, the likelihood of the data is computed by summing over the latent variables $f_1 = f_i(t_1)$ and $f_2 = f_i(t_2)$:

$$P(\hat{n}_i(t_1), \hat{n}_i(t_2)|\tau, \theta) = \iint_{f_{\min}}^{1} df_1 df_2 \, \rho(f_1) P(f_2|f_1, \tau, \theta) \times P(\hat{f}_1|f_2) P(\hat{f}_2|f_2), \tag{4.4}$$

where $P(f_2|f_1; \tau, \theta)$ is the propagator of the geometric Brownian motion Eq. 4.2, and $\hat{f}_j = \hat{n}_i(t_j)/N_r(t_j)$, $j = 1, 2$.

To explore the dependence of the τ and θ parameters on the frequency of clonotypes, and to eliminate clones that are not seen at both timepoints, we can generalize the formulas above to include only clonotypes with frequencies larger than a specific threshold $\hat{f}_{\mathrm{th}}$, which modifies the normalization of the maximum likelihood estimator (see Methods).

4.2.5 Validation of the inference methods on synthetic data

We first test the naive and full model inference on simulated RepSeq samples. We simulate 10^{10} cells corresponding to $\sim 10^8$ synthetic longitudinal trajectories designed to mimic as closely as possible the features of the real repertoire data at time points two years apart. The initial size $n_i(t_1)$ of each clone is drawn from the steady state distribution of the GBM with a constant source (see Methods). Then Eq. 4.2 is simulated between times t_1 and t_2 with an extinction condition when $n_i < 1$, and with a source of new clones whose rate of introcution is matched to the mean extinction rate (see Methods). We varied the two timescales of the model, τ and θ, from months to years, while keeping $\alpha = 2\theta/\tau$ within the observed experimental range 1.1 – 1.25 Gaimann et al. (2020).

We model experimental sampling using a negative binomial distribution with variance parameters $a = 0.7$ and $b = 1.1$. Sequencing depth was set to $N_r = 10^6$ reads at both time points (we checked that asymmetric numbers of reads at each timepoint did not affect the results, see Fig. S2), resulting in $\sim 10^5$ sampled distinct clonotypes. For each set of parameters we generated 10 longitudinal datasets to assess errors. We then performed the naive and full inference methods on these

datasets, restricted to clones with $\hat{f}_1 \geq \hat{f}_{\text{th}}$ and $\hat{f}_2 > 0$, and compared the inferred values of τ and θ to the true ones (Fig. S2C-D).

While the full inference (blue points) works for all values of the parameters and frequency threshold, the naive inference (red points) performs poorly for large values of the parameters. Increasing the cutoff frequency to $f_{\text{th}} = 10^{-4}$ improves the naive inference by limiting the effect of the sampling noise, which is relatively smaller in large clones. For lower values of the threshold, the more numerous small clones dominate the inference, yielding an erroneous estimate. However, since the naive inference does not require replicates or a noise model, and is faster to implement, it provides a practical solution for learning τ and θ for large clones.

4.2.6 Analysis of repertoires

We applied the full inference to longitudinal data sets of healthy individual TCR repertoires presented in Fig. S1 for which replicates were available, focusing on large enough clones ($\hat{f}_1 \geq f_{\text{th}} = 10^{-5}$). With this cutoff we limit experimental noise and focus mainly on memory clones, since large clones are more likely to be have arisen from expansion and belong to the memory pool de Greef et al. (2020). For all individuals, the inferred values of τ and θ fall close to a line defined by $\tau \approx 2\theta$ (Fig. 4.4A) corresponding to a predicted exponent of $\alpha \approx 2\theta/\tau = 1$ in the power law of the clone size distribution. This result is in agreement with empirical observations of Fig. S1D. A more refined comparison of the predicted exponent, $2\theta/\tau$, with the one directly inferred from the distribution of clone sizes, α, gives consistent but noisy results (Fig. 4.4B), primarily because of the narrow range of values of α ($0.9-1.2$) and the small number of individuals. We note that the two inferred values of α use completely independent pieces of information, namely the clone size distribution in one case, and the dynamics of clone sizes in the other.

Since our approach is probabilistic, it provides as a byproduct the posterior distribution of the fold change of individual clones (see Methods). The average of this posterior over clones agrees very well with the model propagator (4.5) (Fig. S3), validating its consistency with the data.

The turnover time τ increases sharply with age, from a few years at age 21 to ~ 50 years at age 57 (Fig. 4.4C). Since the ratio $2\theta/\tau$ is constrained to be ≈ 1, this implies that the amplitude of the stochastic stimulations, θ^{-1}, decreases with age. The TCR repertoire is more dynamic, with faster turnover, for young individuals, who also have a larger rate of introduction of new TCR clones from the thymus than older individuals. At the same time, a turnover time of ~ 20 years at the age of 40 suggests that the repertoires of adults remain dynamic despite greatly reduced thymic output.

For individual P6, both TCRα and TCRβ RepSeq samples were available. We recover very similar dynamic parameters for both receptor chains (Fig. 4.4A-B).

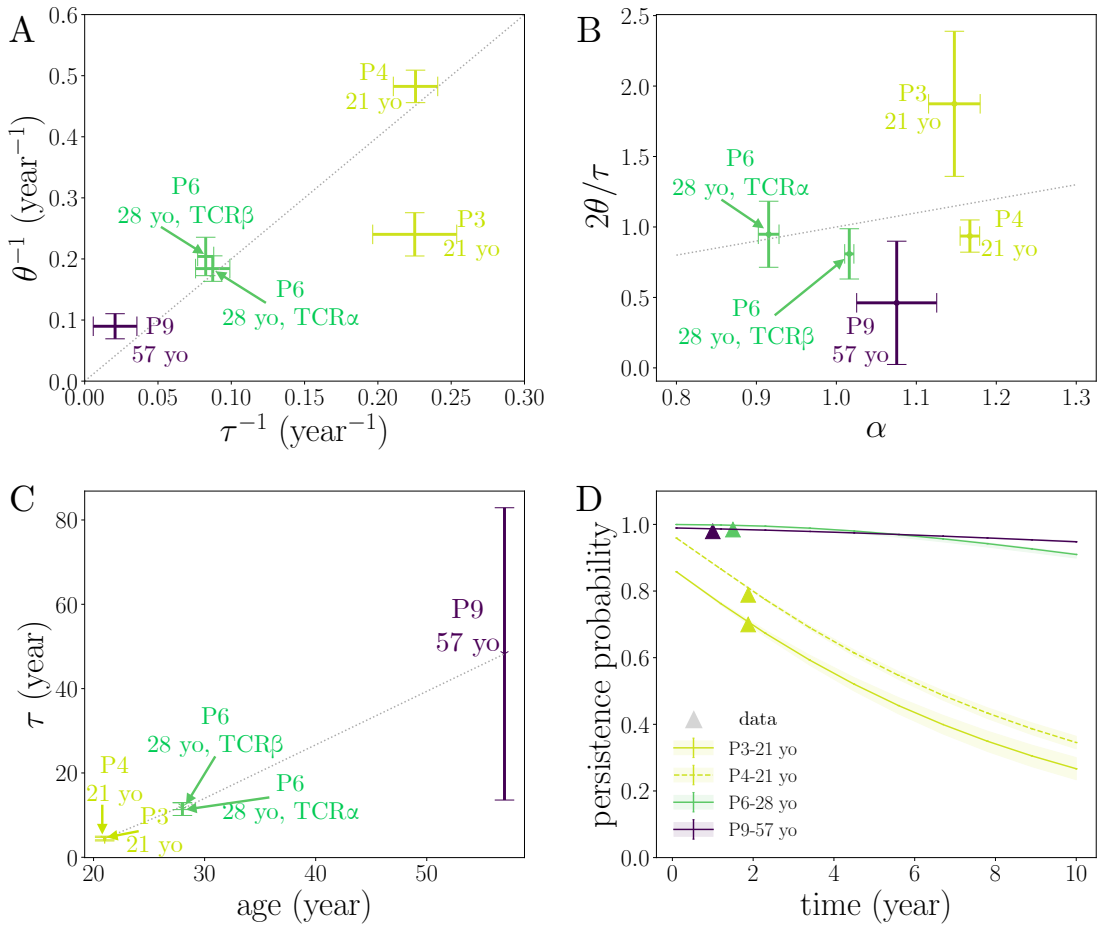

Figure 4.4: **Dynamical parameters of healthy TCR repertoires. A.** Typical decay rate τ^{-1}, and inverse fluctuation amplitude θ^{-1} for the 5 donors for whom replicates were available, as obtained using the full inference procedure with $f_{\text{th}} = 10^{-5}$. All donors but one are consistent with the relation $2\theta = \tau$, corresponding to $\alpha = 1$. Error bars are standard deviations over all combinations of the replicates at each time point. **B.** Direct test of the prediction $\alpha = 2\theta/\tau$. Most values of α fall close to one, allowing for only a narrow range of tested values. Error bars on α show standard deviations across time points. **C.** Turnover parameter τ as a function of donor age. **D.** Probability for a clone detected at some timepoint with frequency $\hat{f} \geq 10^{-5}$ to be detected again at a later time point (with the experimental dataset size). Symbols are empirical estimates. The model predictions show excellent agreement. Error bars in **B.-D.** are propagated from **A.**

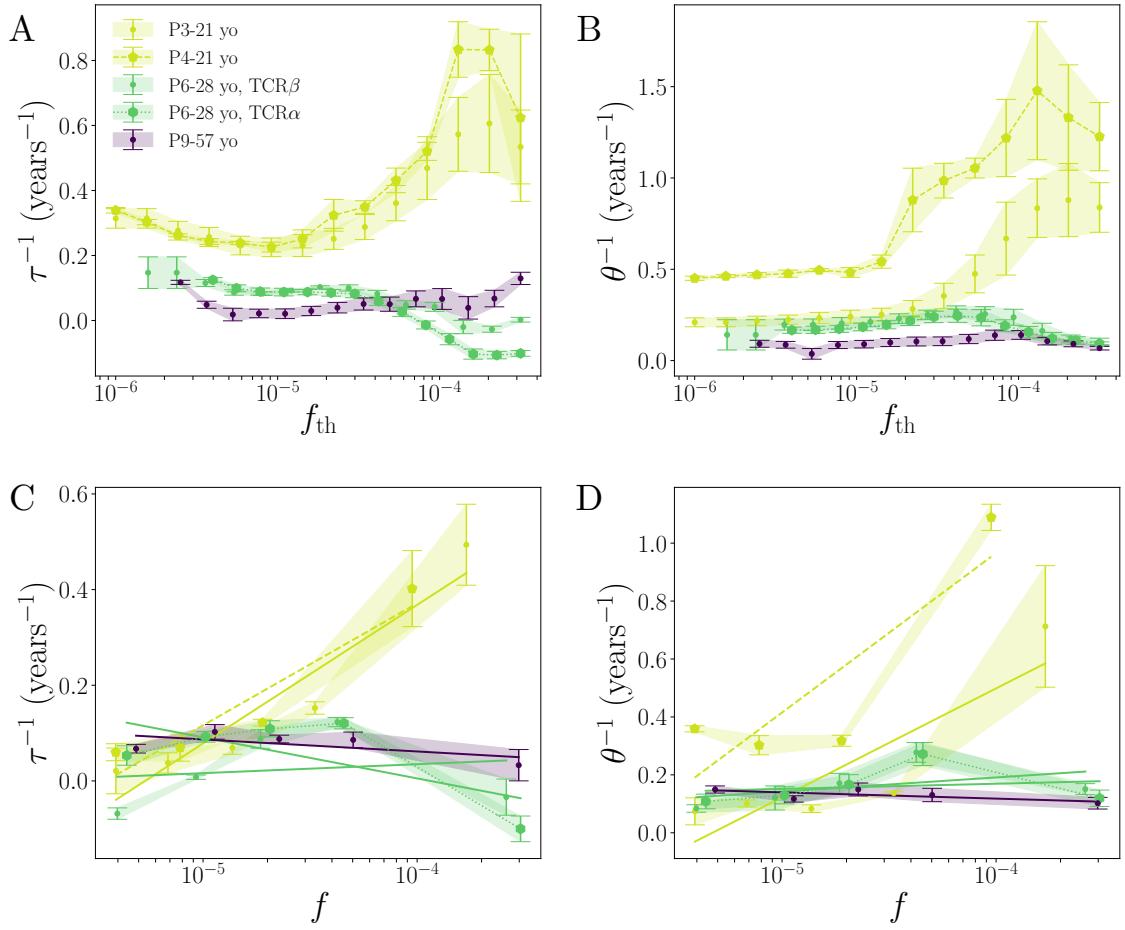

Figure 4.5: **Clonal dynamics are frequency dependent. A-B.** Results of the full inference as a function of the minimal frequency threshold $\hat{f}_{\rm th}$ for τ^{-1} and θ^{-1}. **C-D.** Dynamical parameters as a function of clone frequency. The inference was performed on separate subsets of clones sorted by their frequency in intervals $n_{\rm min} < \hat{n} \leq n_{\rm max}$, with $n_{\rm min,max}$ consecutive numbers in $(2, 5, 10, 20, 100, \infty)$. Error bars are estimated as in Fig. 4.

This justifies our hypothesis that the bulk sequencing of single chains captures well the dynamics of $\alpha\beta$ clonotypes.

For comparison, we also applied the naive inference procedure, which allows us to include all 9 patients even when replicates are not available. This inference generally gave much larger rates τ^{-1} and θ^{-1} (Fig. S4A), suggesting confounding effects of the noise on both parameters (reversion to the mean for τ^{-1}, and larger variance for θ^{-1}). Indeed, results obtained for a larger value of the frequency threshold ($f_{\text{th}} = 10^{-4}$, Fig. S4B) gave smaller values, and in better agreement with the age dependence.

To ask whether the clonal dynamics depended on the cell type, we separately analyzed the longitudinally sampled CD4 and CD8 repertoires of P6, the only individual for which such data were available. The clone size distribution of CD4 falls off with a larger exponent than that of CD8, meaning that its largest expanded clones are relatively smaller (Fig. S5 A), as already noted Britanova et al. (2016). We then applied the naive inference procedure with $f_{\text{th}} = 10^{-4}$ (since we did not have replicates for the CD4 and CD8 repertoires). The inference (Fig. S5B) reveals that CD4 clones turn over more slowly than CD8 cells (smaller τ^{-1}), but also have much smaller fluctuations in their sizes (smaller θ^{-1}). This result is consistent with a shorter tail of large clones and a larger α in CD4 than in CD8 (Fig. S5C).

The inference results can be used to predict the persistence of clones, whose turnover has been discussed in the context of aging Britanova et al. (2016); Chu et al. (2019); Gaimann et al. (2020). For a given individual, we define persistence as the probability that a clone initially observed at frequency $\geq \hat{f}_{\text{th}} = 10^{-5}$ is re-sampled at a later time. This probability strongly depends on the dynamics of turnover of the TCR repertoire, and therefore on the age of the individual, as well as on the time interval between the two samples (Fig. 4.4D). We can estimate this persistence probability directly from data, and compare it to the predictions of our inferred dynamics, showing excellent agreement. This analysis show that even moderately large clones persist for many years and even decades in older individuals.

Our model assumes that clones have unique trajectories, but that the statistical properties of these trajectories are uniform. However, because of their distinct histories and phenotypical compositions, clones may differ in those dynamical properties. To investigate that possibility, we asked whether the inferred time scales τ and θ depended on the value of the clone size threshold $\hat{f}_{\text{th}}$. Low values of $\hat{f}_{\text{th}}$ mean that all clones are taken into account in the inference, while high values mean that we focus on the largest clones only. We found that the values of both τ^{-1} and θ^{-1} increase with the threshold (Fig. 4.5A and B) for P3, P4, and P9, suggesting that large clones tend to be more dynamic, with a faster turnover. Therefore, while the overall trends reported in Fig. 4.4 are still correct, these results imply that the model should be revisited to allow for frequency-dependent dynamics.

To measure the frequency dependence of the dynamic parameters more finely,

we separately inferred τ and θ for clones sorted into contiguous intervals according to their initial count $\hat{n}(t_1)$. Both time scales showed an approximately linear dependency on the logarithm of the initial frequency (Fig. 4.5 C and D). This confirms the observation that large clones tend to have faster dynamics than small ones, especially in younger individuals. The two time scales τ and θ vary in concert, so that their ratio remains approximately constant across frequencies (Fig. S6). As we have argued before, this ratio is linked to the power-law exponent of the distribution of clone sizes at steady state. This exponent can be read off as the slope of that distribution on a double logarithmic scale, which it is consistently observed to be constant in the data (Fig. S1D). Finally, we applied the naive inference procedure to learn the frequency-dependent dynamics of clones in all 9 individuals. As expected, this inference yielded more noisy and less stable results than the full inference, especially at low frequencies for which noise is largest (Fig. S7). The dependence of the inferred parameters on clonal frequency vary across individuals, but confirm a picture in which older individuals have more stable large clones.

4.1. Discussion

The sizes of T cell clones change constantly throughout the lifetime of an individual, not only due to specific stimulation. We used data sampled on timescales of the order of a year from individuals that did not undergo any strong identified antigenic stimulations to learn the repertoire turnover dynamics. These dynamics include both random unstimulated T cell proliferation and death as well as asymptomatic, or weakly symptomatic antigenic stimulation. We showed that a geometric Brownian motion correctly captures the clone dynamics. This model imposes strict relations that link the exponent of the TCR clone size distribution at steady state, α, with the parameters of the dynamics, τ and θ. We showed that, for all individuals, we were able to predict the measured exponent $\alpha \approx 1$ from the inferred dynamical parameters, suggesting that geometric Brownian motion is a good description of the process. Although the actual timescales vary between individuals, with much younger individuals having much faster clone turnover dynamics than older individuals, their ratio is fixed. Indeed, as already noted on a larger cohort of individuals in Ref. Gaimann et al. (2020), the exponent of the power-law distribution of clone sizes does not depend on age.

The source of the faster turnover in younger individuals is not explained by our analysis. It can be linked to a larger thymic output rate Yates (2014), imposing a faster turnover. It could also be linked to more a rapid formation of new immune memories at a young age. We did not attempt to separately learn the dynamics of memory and naive pools, since we did not have sorted longitudinal data for which

abundance information could be trusted. While it is sometimes assumed that larger clones have a memory phenotype because they must have expanded, a recent study in mice has shown that naive clones can be large as well de Greef et al. (2020). It will be interesting to perform a separate analysis of careful sorted naive and memory repertoires in the future using the method described here, especially for individuals of different ages.

More generally, we expect clonal dynamics to be linked to the cellular phenotype, as our preliminary analysis showed for CD4 and CD8 cells. Phenotypes can be characterized with increasing resolution using single-cell expression data Pai and Satpathy (2021), which also provides paired TCR information Valkiers et al. (2022). Future work combining longitudinal sampling with single-cell techniques could help explore the relationship between neutral clonal dynamics and cell type. Additionally, we know that TCR with similar sequences form clusters that often respond to similar stimulants Dash et al. (2017); Glanville et al. (2017), and methods are being developped to annotate repertoire with cluster membership Mayer-Blackwell et al. (2021) or specificity Gielis et al. (2019); Montemurro et al. (2020); Sidhom et al. (2021); Springer et al. (2021); Zhang et al. (2021b). As these annotation become comprehensive, one will be able to study the dynamics of specificity clusters, and to assess the persistence of specific immune memories across different immune challenges.

Our current model is based on two effective parameters that describe the timescale for clone turnover, τ, and the timescale of random changes, θ. Two major assumptions underlie this model. First, it assumes that antigenic stimulation happens repeatedly on short time scales, so that its cumulated effect on longer time scales look like random fluctuations of the net growth rate. Testing this assumption would require longer time traces of the clonal dynamics, to look for memory effects in the clonal growth rates. Second, it assumes that dynamical properties do not depend on the clone size. As observed in Fig. 4.5, this assumption is only partially verified, with clear violations for 2 of the youngest donors, in which the larger clones display much faster dynamics than the smaller ones. The longidutinal analysis of larger cohorts with a broad age distribution would be required to investigate this effect in detail.

The turnover time scales we infer range from a few years to 50 years, depending on the age of the individual. It has been shown that even sparsely sampled T-cell repertoires can provide a fingerprint that uniquely identifies individuals Dupic et al. (2021). The stability of this immune fingerprint is guaranteed for tens of years, provided that the turnover rate is of the order of years or more, as we showed here.

Direct measurements of T cell lifetimes using heavy water De Boer et al. (2003) give lifetimes of months for memory cells, to a few years for naive cells. These estimates are consistent with our findings: our time scale τ is linked to the inverse of

the net growth rate of the *clone*, which results from the balance between cell proliferation and death, while experiments based on heavy water measure the turnover of individual *cells*. For instance, memory cells are short lived, but also divide rapidly to compensate for death, so that the size of memory clones remains stable. One may also want to compare our estimate with the previously reported persistence time of clonotypes believed to be of fetal origin, ≈ 37 years Pogorelyy et al. (2017). This persistence time is not directly comparable to τ, which is the decay rate of the *abundance* of each clone, but it is similar to the characteristic decay of the persistence probability (Fig. 4.4D), which may be slower. Another caveat is that fetal clonotypes are also primarily naive and take up only a few percent of the repertoire, so that they may not be representative of the overall properties of the clonal dynamics.

Our work was possible because we were able to calibrate the noise using replicate samples. However, replicates are not always available. In this case, the dynamics can still be learned for large clones: we showed using simulations that above a certain frequency threshold, the sampling error becomes small and we can use empirical observations to learn TCR repertoire dynamics directly from read counts. This allows us to correctly estimate the dynamics of large clones without a noise model, if the clones sizes are large at both time points. However, since the repertoire is described by a power law distribution, the role of small clones is far from negligible. An alternative to replicates may be to use close-by timepoints (relative to the time scales of the dynamics) as surrogate replicates. While we had such time points separated by one month for P1, P2 and P7, we did not attempt a full inference on these samples: we did not manage learn a reliable noise model for these donors, because we lacked both the raw sequencing reads and details about the processing procedure (PCR amplification, error correction, etc). In particular, unlike uniquely barcoded cDNA sequencing, PCR amplification of gDNA used for these donors inflates rare clonotypes (as suggested by the low-frequency plateau in the clone size distributions, see Fig. S1D), potentially confounding the analysis.

One of the main conclusions of our work is that repertoires are very dynamic systems, with clone frequencies changing by orders of magnitude on timescales of years, even in the absence of strong known stimulation. This observation challenges our ability to identify responding clonotypes to direct immune stimulation, such as vaccination or diseases. This work builds the ground for inference procedures that not only correct for experimental and biological noise but also for the natural repertoire dynamics. The methods we designed are general and can be used on larger cohorts of individuals presenting different health status, age, and immunodeficiencies features. They provide a promising tool to better understand the maintenance and efficiency of T-cells, enabling to quantify immunosenescence Zhang et al. (2021a), which plays an important role in vaccines performance and cancer research.k

4.2. Methods

4.2.1 Longitudinal data

The datasets analyzed in this study are summarized in Table S1, along with accession number and links to databases.

Data was collected from 4 different studies, which uses two different techniques for repertoire sequencing. Data from Britanova et al. (2016); Pogorelyy et al. (2018c); Sycheva et al. (2018); Minervina et al. (2020) was generated by sequencing TCR mRNA of PBMCs from healthy individuals, while data from Chu et al. (2019) was obtained by directly sequencing genomic DNA (gDNA), as described in detail in each original study.

Briefly, mRNA sequencing was done through cDNA synthesis with template switch allowing for the addition of a unique molecular identifier (UMI), followed by 2-step PCR amplification of the TCR loci (alpha and/or beta), multiplexing, sequencing on an Illumina platform, and processing using the MiXCR software package Bolotin et al. (2015), to obtain lists of clonotypes (V and J segments and Complementarity Determining Region 3 nucleotide sequence) corrected for UMI multiplicity and sequencing errors. gDNA sequencing was done by extracting genomic DNA and performing multiplex PCR to amplify the TCR beta gene before sequencing on an Illumina HiSeq system. Raw data processing was performed using closed software. Since the raw data is not available, we used the processed data provided on the ImmuneAccess platform.

4.2.2 Naive inference

The *naive inference* method directly uses the observed TCR clonal frequencies to learn τ and θ parameters, assuming that they represent exactly the true frequencies: $f = \hat{f} = \hat{n}/N_r$. We aim here at maximizing directly the log-likelihood $\mathcal{L}(\tau, \theta) = \log \mathbb{P}(\{(f_i(t_1), f_i(t_2))\})|\tau, \theta)$, which can be expressed by integrating Eq.4.2:

$$\begin{aligned} &\mathbb{P}(\{(\ln f_i(t_1), \ln f_i(t_2))\})|\tau, \theta) \\ &= \prod_i^{N_c} G(\ln f_i(t_2)|\ln f_i(t_1); \tau, \theta) P(\ln f_i(t_1)), \end{aligned} \tag{4.5}$$

where

$$G(x|y;\tau,\theta) = \sqrt{\frac{\theta}{2\pi\Delta t}} \exp -\frac{\theta(x - y - \Delta t\tau^{-1})^2}{2\Delta t} \tag{4.6}$$

is the propagator of the Brownian motion, $\Delta t = t_2 - t_1$ the time interval between the two time points, and where we have assumed that N_{cell} is a constant of time. Maximizing the log-likelihood with respect to τ and θ is equivalent to doing linear regression of $\ln f(t_2) - \ln f(t_1)$ against Δt.

4.2.3 Full inference

Using same-day replicates at time t_j, we jointly learn the parameters $(\alpha, f_{\min}, a, b)$ of the clone-size distribution $\rho(f) = Cf^{-1-\alpha}$ (for $f_{\min} \leq f \leq 1$), and the noise model $P(\hat{n}|f) = \text{NegBin}(\hat{n}; N_r f, N_r f + a(N_r f)^b)$ using the NoiSET software Bensouda Koraichi et al. (2021), where $\text{NegBin}(n; x, \sigma)$ is a negative binomial of mean x and variance σ. The learned parameters are reported in Fig. S1.

We then learn the parameters of the dynamics by maximizing the likelihood of samples taken at two different time points, using the noise model to account for the discrepancy between true frequencies and sequence counts: For one clone, the full model likelihood reads

$$\begin{aligned}&\mathbb{P}(\hat{n}_i(t_1) = \hat{n}_1, \hat{n}_i(t_2) = \hat{n}_2)|\tau, \theta) = \\ &\int_{[f_{\min},1]^2} df_1 \rho(f_1) \frac{df_2}{f_2} G(\ln f_2 | \ln f_1; \tau, \theta) P(\hat{n}_1|f_1) P(\hat{n}_2|f_2).\end{aligned} \tag{4.7}$$

where the noise models are specific to each time point.

The maximum likelihood estimator is given by :

$$(\tau^*, \theta^*) = \underset{(\tau,\theta)}{\text{argmax}} \prod_{i=1}^{N_c} \frac{\mathbb{P}(\hat{n}_i(t_1), \hat{n}_i(t_2)|\tau, \theta)}{\mathbb{P}(\hat{n}_i(t_1) \geq N_r f_{\text{th}}, \hat{n}_i(t_2) > 0|\tau, \theta)}, \tag{4.8}$$

where the denominator accounts for the condition that the clone be included in the analysis: $\hat{f}_i(t) \geq f_{\text{th}}$ and $\hat{n}_2 > 0$. The choice to impose a threshold on the first time point is justified by the fact we are learning the forward propagator of the dynamics, which is conditioned on the initial frequency. Typically around 50-70% of clones above threshold on the first time point remain above threshold in the second time point. Likewise, similar percentages (40-80%) of clones above threshold in the second time point were also seen in the first time point. This loss is expected since the frequencies follow stochastic trajectories, many of which are likely to cross the threshold between the two time points.

The persistence probability of Fig. 4D is linked to that normalization and is computed as:

$$P_{\text{pers}}(\tau, \theta) = \frac{\mathbb{P}(\hat{n}_i(t_1) \geq N_r f_{\text{th}}, \hat{n}_i(t_2) > 0|\tau, \theta)}{\mathbb{P}(\hat{n}_i(t_1) \geq N_r f_{\text{th}}|\tau, \theta)}. \tag{4.9}$$

Once the model is learned, the posterior distribution of fold changes $s_i \equiv \ln f_i(t_2) - \ln f_i(t_1)$ of each clone i is computed through

$$\begin{aligned}&\mathbb{P}(s_i = s|\hat{n}_1, \hat{n}_2, \tau^*, \theta^*) = \\ &\frac{\int_{f_{\min}}^1 df_1 \rho(f_1) G(\ln f_1 + s | \ln f_1; \tau^*, \theta^*) P(\hat{n}_1|f_1) P(\hat{n}_2|f_1 e^s)}{\mathbb{P}(\hat{n}_1, \hat{n}_2|\tau^*, \theta^*)}.\end{aligned} \tag{4.10}$$

The overall posterior distribution over all clones (solid lines in Fig. S3) is then given

by $\mathbb{P}_{\text{post}}(s) = (1/N_c)\sum_{i=1}^{N_c}\mathbb{P}(s_i = s|\hat{n}_1, \hat{n}_2, \tau^*, \theta^*)$. The prior distribution (dashed line), by contrast, is directly given by $G(\ln f_1 + s, \ln f_1|\tau^*, \theta^*)$, which is independent of f_1.

When doing inference in each frequency bin, the product in (6.4) runs over clones that fall in the bin, and the normalization in the denominator is replaced by the probability to observe $\hat{n}_i(t_1)$ in the bin of interest, and $\hat{n}_i(t_2) > 0$. The maximization is performed using the `minimize` function from the Scipy package, with the Sequential Least Squares Programming (SLSQP) method Virtanen et al. (2020) with parameters `tol=1e-8` and `maxiter=300` and initial condition $\tau = 2, \theta = .5$ and constraint $\theta^{-1} > 10^{-3}$.

4.2.4 Synthetic data

Synthetic data was generated by simulating Eq. 4.2 with a source term producing new clones with rate S at initial size $n = n_0 = 40$, and an absorbing boundary condition at $n = 1$. We work with the $x = \ln n$ variable for convenience. The simulation is initialized at steady state, which can be computed analytically Desponds et al. (2016); Mora and Walczak (2019a). The analytical solution gives us the expected number of cells and clones as a function of the model parameters: $N_{\text{cell}} = S(n_0 - 1)/(\tau^{-1} - \theta^{-1}/2)$, and $N_c = S\tau \ln n_0$. Fixing the number of cells to $N_{\text{cell}} = 10^{10}$, we then compute the number of clones necessary to achieve that size, $N_c = N_{\text{cell}}(1 - \tau/2\theta)\ln n_0/(n_0 - 1)$. We then draw the size $n_i(t_1) = e^{x_i(t_1)}$ of each clone $i = 1, \ldots, N_c$ from the continuous steady state distribution Desponds et al. (2016):

$$\rho_x(x) = \begin{cases} S\tau\,(1 - e^{-\alpha x}) & \text{if } x \leq x_0 \equiv \ln n_0 \\ S\tau e^{-\alpha x}\,(e^{\alpha x_0} - 1) & \text{if } x > x_0, \end{cases} \tag{4.11}$$

with $\alpha = 2\theta/\tau$.

Then the evolution of each clone from time t_1 to $t_2 = t_1 + \Delta t$ is determined by the modified propagator with absorbing boundary condition at $x = 0$:

$$G_{\text{abs}}(x|y) = G(x|y) - e^{-\alpha y}G(x| - y) \tag{4.12}$$

where $G(x|y)$ is defined in (4.6). In practice, we kill clone i with probability $1 - P_{\text{surv}}(x_i(t_1)) \equiv 1 - \int_0^\infty dx G_{\text{abs}}(x|x_i(t_1))$, which can be expressed in terms of error functions. Otherwise, its new log-size $x_i(t_2)$ is drawn from the distribution $G_{\text{abs}}(x|x_i(t_1))/P_{\text{surv}}(x_i(t_1))$.

In addition, new clones are introduced during Δt. We draw their number from a Poisson distribution of mean $S\Delta t$, and their introduction times t from a uniform distribution in the interval $[t_1, t_2]$. Then their dynamics are drawn in the same way as for the initial clones, but with $\Delta t = t_2 - t$ instead of $t_2 - t_1$.

Once the abundances $(n_i(t_1) = e^{x_i(t_1)}, n_i(t_2) = e^{x_i(t_2)})$ have been determined, the

number of reads $\hat{n}_i(t_1), \hat{n}_i(t_2)$ from each time point is drawn from a negative binomial distribution of mean $\langle \hat{n}_i(t_1) \rangle = N_r n_i(t_1)/N_{\text{cell}}$ and variance $\langle \hat{n}_i(t_1) \rangle + a \langle \hat{n}_i(t_1) \rangle^b$, and likewise for $\hat{n}_i(t_2)$, with $N_r = 10^6$, $a = 0.7$ and $b = 1.1$.

4.2.5 Comparison to the VDJdb database

We downloaded the 2022-03-30 release of VDJdb Bagaev et al. (2020a) at `https://github.com/antigenomics/vdjdb-db/releases/download/2022-03-30/vdjdb-2022-03-30.zip`, and restricted our search to CDR3s associated to antigens from the following species: CMV, InfluenzaA, EBV, SARS-CoV-2, HIV-1, HCV, YFV, HTLV-1, DENV1, DENV3/4, HIV, HSV-2, M.tuberculosis, DENV2, HCoV-HKU1, HPV, MCPyV, StreptomycesKanamyceticus, E.Coli, HIV1, HHV, PseudomonasAeruginosa, PseudomonasFluorescens, SaccharomycesCerevisiae, SelaginellaMoellendorffii, totalling 65616 CDR3 amino-acid sequences.

4.2.6 Code availability

All scripts to produce the figures can be found at `https://github.com/statbiophys/Inferring_TCR_repertoire_dynamics/`.

4.3. Supplementary figure

ID	age	sex	# clones	# reads	Tech	DOI ref
P1	18-19 yo	female	4.08 - 5.11·10^5	10.8 - 25.1·10^6	gDNA	Chu et al. (2019)
P2	18-19 yo	female	1.55 - 2.93·10^5	10.5 - 20.7·10^6	gDNA	Chu et al. (2019)
P3	21-23 yo	male	2.04 - 6.44·10^5	0.23 - 1.03·10^6	RNA	Pogorelyy et al. (2018c)
P4	21-23 yo	male	1.9 - 10.06·10^5	0.28 - 1.79·10^6	RNA	Pogorelyy et al. (2018c)
P5	27-30 yo	male	7.6 - 18.12·10^5	1.53 - 6.94·10^6	RNA	Britanova et al. (2016)
P6	28-29 yo	male	2.62 - 6.38·10^5	0.56 - 1.69·10^6	RNA	Minervina et al. (2020)
P7	45-46 yo	female	1.93 - 2.42·10^5	8.86 - 22.8·10^6	gDNA	Chu et al. (2019)
P8	47-50 yo	male	1.38 - 9.54·10^5	1.53 - 5.93·10^6	RNA	Britanova et al. (2016)
P9	57-58 yo	male	3.25 - 7.29·10^5	0.62 - 1.64·10^6	RNA	Sycheva et al. (2018)

Table 4.1: Summary of the repertoire samples and individuals used in this study.

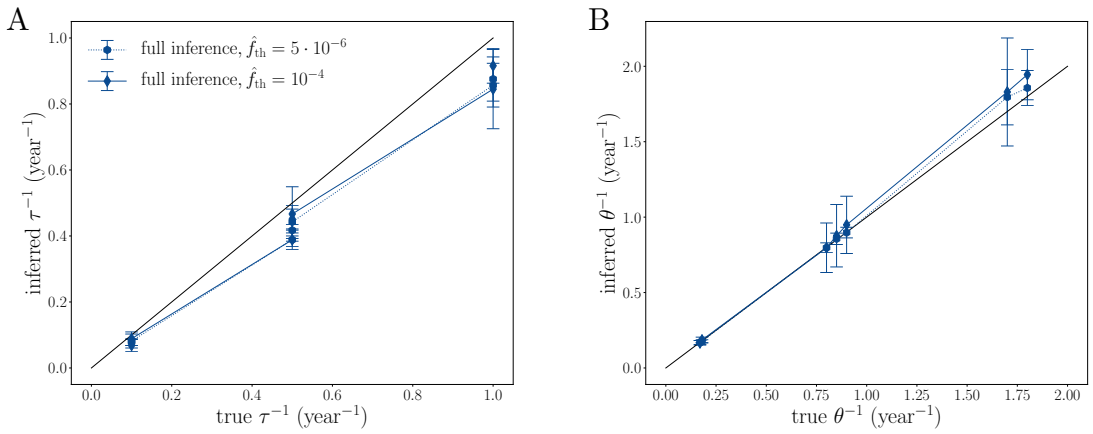

Figure S1: Validation of the inference method for time point with very different number of reads. Parameters: all 9 combinations of $\tau^{-1} = (0.1, 0.5, 1)$ year^{-1} and $\alpha = (1.11, 1.17, 1.25)$. The number of reads are $N_r(t_1) = 10^6$ and $N_r(t_2) = 10^5$.

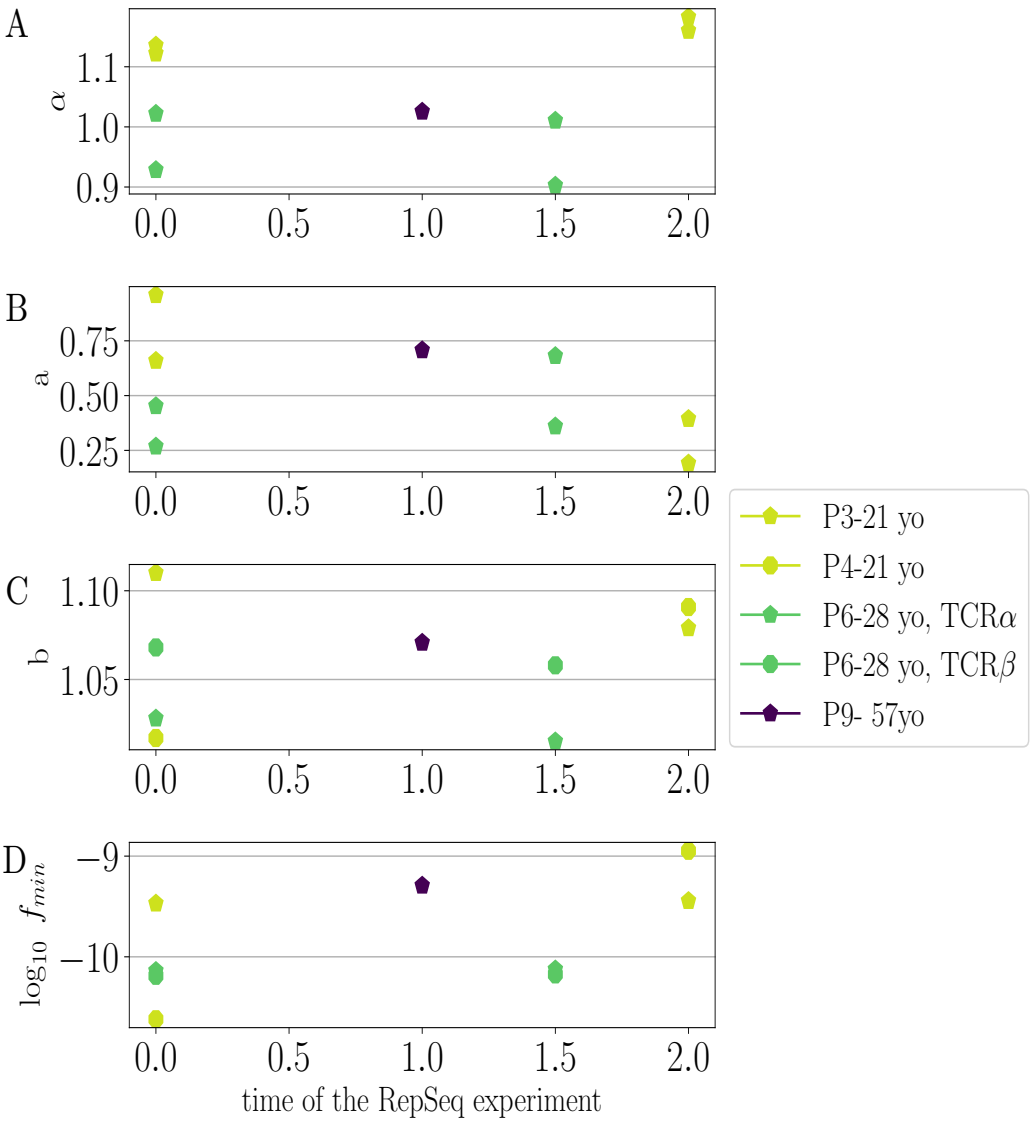

Table 4.2: Parameters of the null model, which characterize both the noise model (a and b) and the prior power law distribution of frequencies (α,$f_{\min}$). The x-axis represents time in years since the first sample was taken for each individual. Only samples for which duplicates are available are shown.

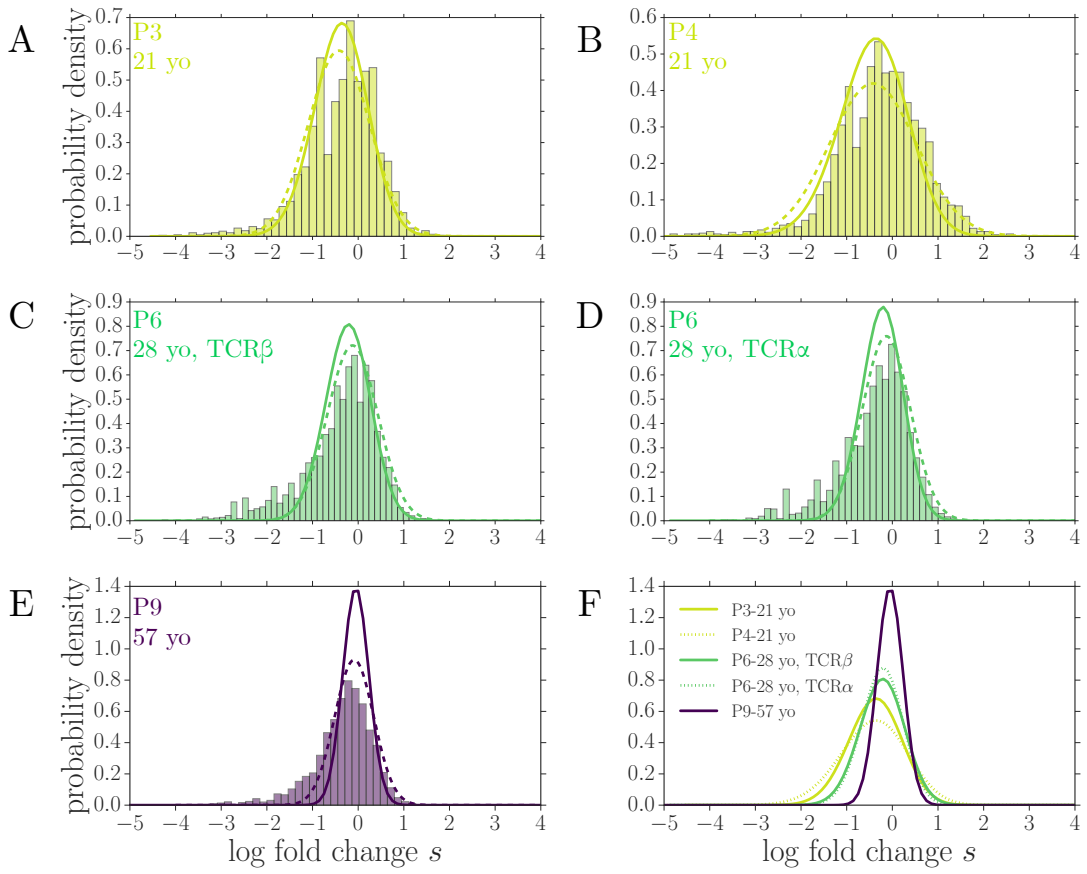

Figure S2: Distribution of log-fold changes s. A-E. For each indivudal and chain we compare the naive distribution $s \approx \ln(\hat{f}_2/\hat{f}_1)$ (histogram bars), the prior distribution $G(\ln f_1 + s, \ln f_1|\tau^*, \theta^*)$ with the inferred parameters, and the posterior distribution $(1/N_c)\sum_i \mathbb{P}(s|\hat{n}_i(t_1), \hat{n}_i(t_2); \tau^*, \theta^*)$. While the naive distribution has an excess of low values due to small number errors, the prior and posterior distribution agree well. **F.** Comparison of posteriors across individuals, showing how both the average decay and its spread decrease with age.

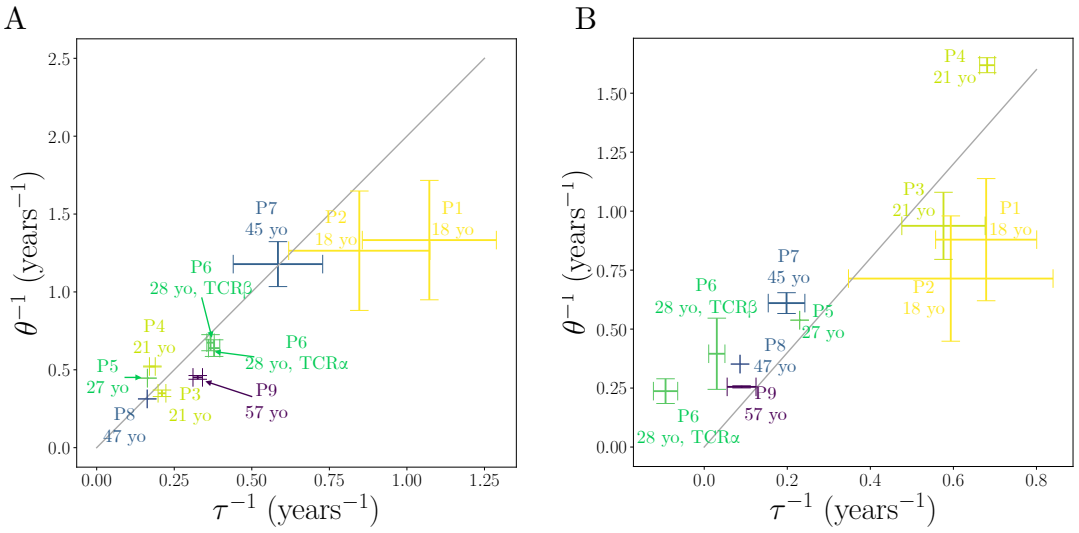

Figure S3: Naive inference of the dynamical parameters on all individuals, with **(A)** $f_{\text{th}} = 10^{-5}$ and **(B)** $f_{\text{th}} = 10^{-4}$.

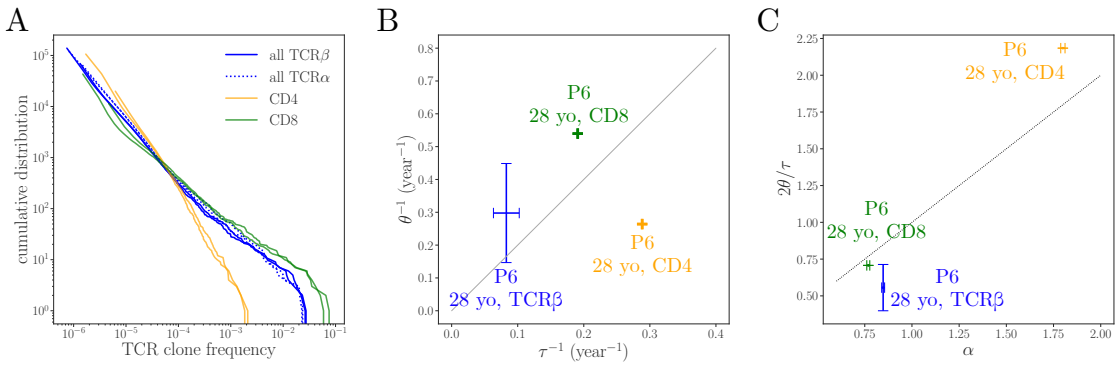

Figure S4: Comparison for the CD4 and CD8 repertoire dynamics in P6. **A.** Clone size distrition in the bulk (alpha and beta chains) and in the CD4 and CD8 beta chain repertoires. The CD4 repertoire has a shorter tail, corresponding to a larger exponent α. **B.** Prior and posterior distributions of the log-fold change, as in Fig. S3. **C.** Inferred parameters for each repertoire. **D.** Model prediction ($2\theta/\tau$) vs measured power-law exponent α. The smaller amplitude of frequency fluctuations θ^{-1} in the CD4 repertoire is consistent with its shorter tail of large clones.

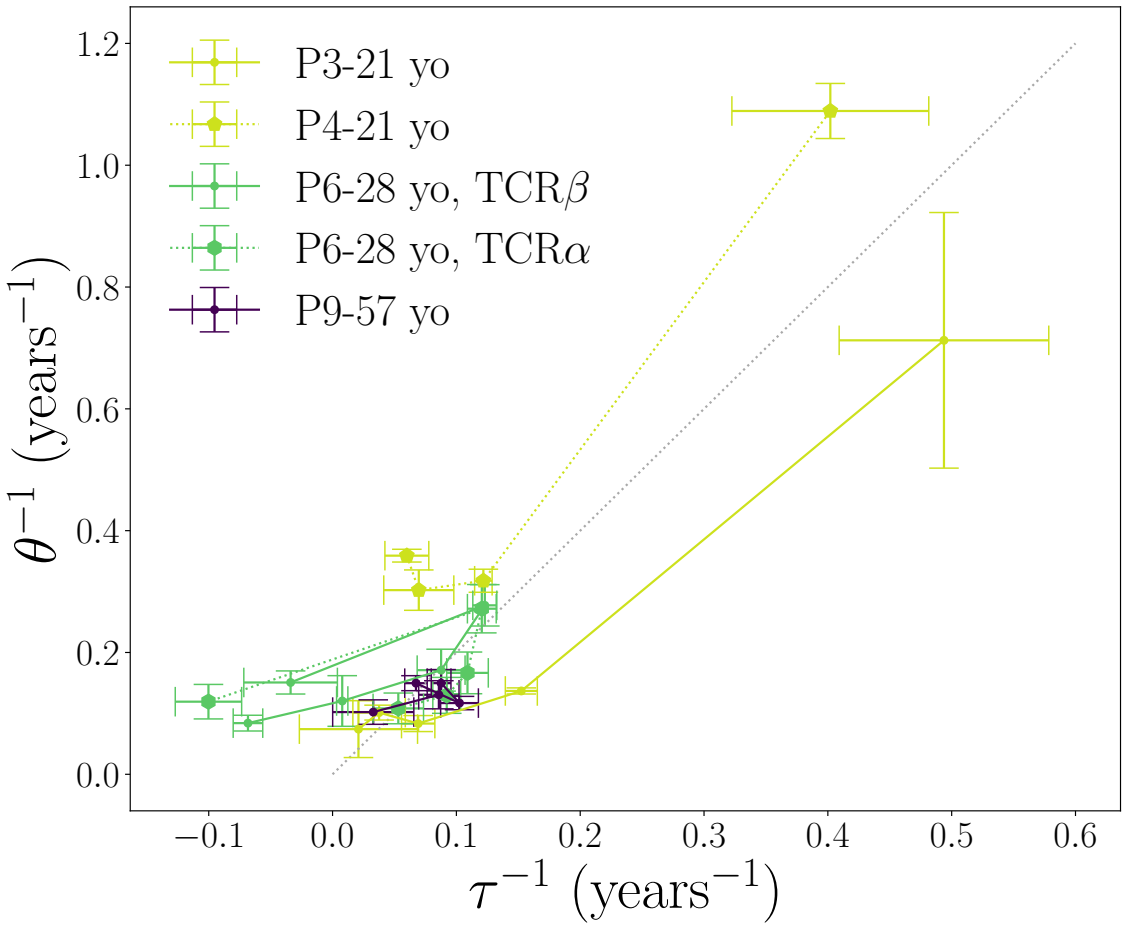

Figure S5: Inferred values of τ and θ for different individuals and frequencies. Frequency intervals and datapoints are the same as in Fig. 5 C-D.

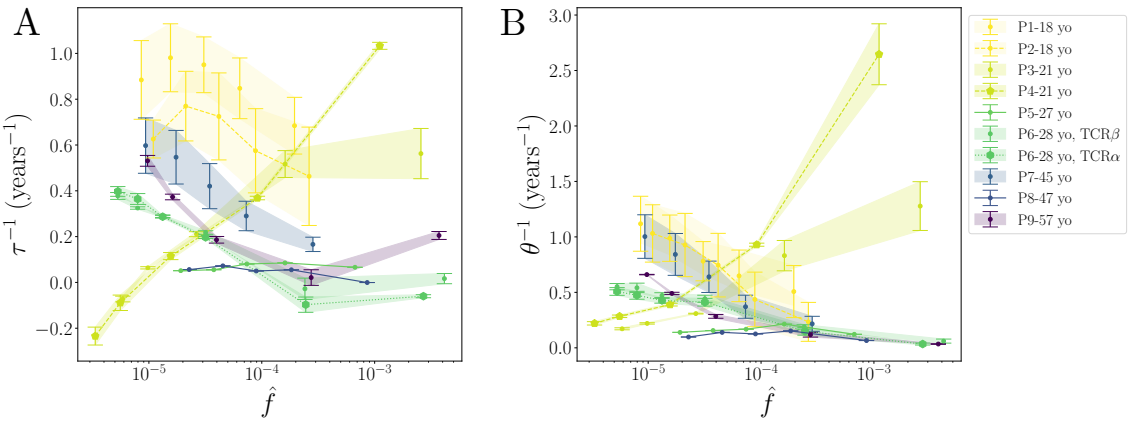

Figure S6: The inference was performed on separate subsets of clones sorted by their frequency in intervals $n_{\min} < \hat{n} \leq n_{\max}$, with $n_{\min,\max}$ consecutive numbers in $(3, 5, 7, 15, 10, 1000, \infty)$ for P3, P4, P6, P9, and $(100, 200, 400, 800, 2000, \infty)$ for P1, P2, P5, P7 and P8.

CHAPTER

5

IMMUNE FINGERPRINTING THROUGH REPERTOIRE SIMILARITY

In chapter 5, I am explaining the work behind *Immprint*, published in Dupic et al. (2021), a method designed in collaboration with Thomas Dupic, for which I helped to address manipulate TCR RepSeq abundances data to conceive a classifier tool to distinguish two individuals thanks to their TCR repertoire identity ("*Immprint*"). Thomas Dupic was the leading scientist in the research published in this paper, and my other main contribution here was to test the method's robustness to acute infections and possible changes in the TCR repertoire information with time. To do so, we exploited data analysis I had already performed on actual data in chapter 4, and use the stochastic population dynamics model also shown in chapter 4 to generate immune repertoire trajectories to validate our tool on synthetic data. My actual work here was to understand basal dynamics of abundances of T cells in the time scale of several years to be able to generate synthetic data and test our *Immprint* scores and their robustness versus time. I also helped Thomas Dupic conceiving the classifier thanks to my previous work working with T-cell abundances on this kind of data.

5.1. Introduction and motivation

Personalized medicine is a frequent promise of next-generation sequencing. These high-throughput and low-cost sequencing technologies hold the potential of tailored treatment for each individual. However, progress comes with privacy concerns.

Genome sequences cannot be anonymized: a genetic fingerprint is in itself enough to fully identify an individual, with the rare exception of monozygotic twins. The privacy risks brought by these pseudonymized genomes have been highlighted by multiple studies Homer et al. (2008); Naveed et al. (2015); Sweeney et al. (2013), and the approach is now routinely used by law enforcement. Sequencing experiments that focus on a limited number of expressed genes should be less prone to these concerns. However, as we will show, B- and T-cell receptor (BCR and TCR) genes are an exception to this rule.

BCR and TCR are randomly generated through somatic recombination Hozumi and Tonegawa (1976), and the fate of each B- or T-cell clone depends on the environment and immune history. The immune repertoire, defined as the set of BCR or TCR expressed in an individual, has been hailed a faithful, personalized medical record, and repertoire sequencing (RepSeq) as a poential tool of choice in personalized medicine Robins (2013); Attaf et al. (2015); Woodsworth et al. (2013); Bradley and Thomas (2019); Davis and Boyd (2019). In this report we show that each person's repertoire is truly unique. We describe how, from small quantities of blood (blood spot or heel prick), one can extract enough information to uniquely identify an individual, providing an immune fingerprint, which we call "Immprint".

Given two samples of peripheral blood respectively containing M_1 and M_2 T cells, we want to distinguish between two hypothetical scenarios: either the two samples come from the same individual ("autologous" scenario), or they were obtained from two different individuals ("heterologous" scenario), see Fig. S1a.

TCR are formed by two protein chains α and β. They each present a region of high somatic variability, labeled CDR3α and CDR3β, randomly generated during the recombination process. These regions are coded by short sequences (around 50 nucleotides), which are captured by RepSeq experiments. The two chains are usually not sequenced together so that the pairing information between α and β is lost. Most experiments focus on the β chain, and we will focus on that chain, but the results are largely independent of this choice. CDR3β sequences are very diverse, with more than 10^{40} possible sequences **?**. For comparison, a human TCR repertoire is composed of 10^8 to 10^{10} unique clonotypes Qi et al. (2014); Lythe et al. (2015). As a result, most of the sequences found in a repertoire are "private".

To discriminate between the autologous and heterologous scenarios, one can count the number of nucleotide receptor sequences, $\mathcal{S}$, shared between the two samples. Samples coming from the same individual should have more receptors in common because T-cells are organized in clones of cells carrying the same TCR. By contrast, $\mathcal{S}$ should be low in pairs of samples from different individuals, in which sharing is due to rare convergent recombinations. Appropriately setting a threshold to jointly minimize the rates of false positives and false negatives (Fig. S1b), we can use $\mathcal{S}$ as a classifier to distinguish autologous from heterologous samples.

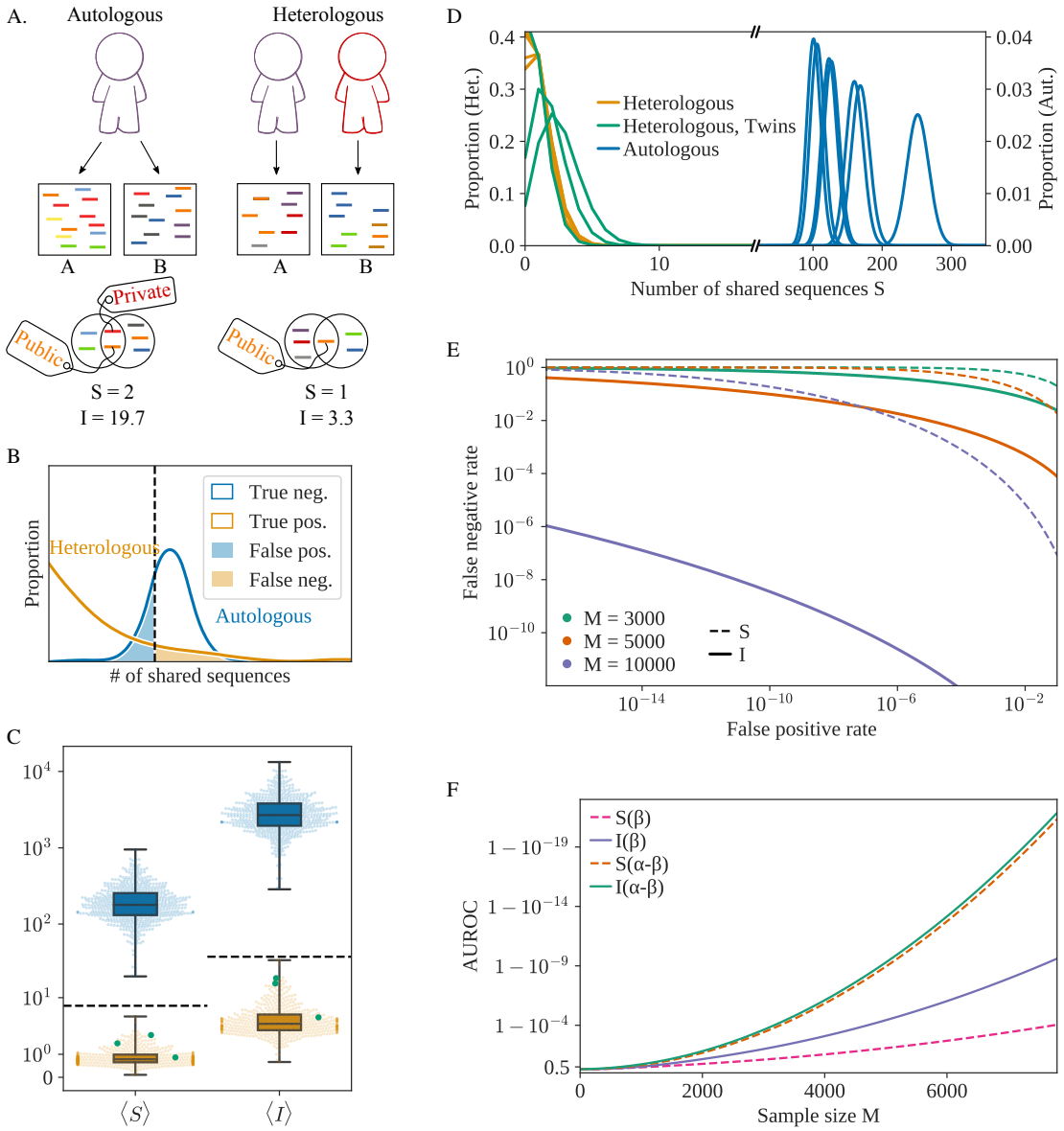

Figure S1: a) The two samples A and B can either originate from the same individual (autologous) or two different individuals (heterologous). In both scenarios, sequences can be shared between the two samples, but their quantity and quality vary. b) Schematic representation of the distribution of the $\mathcal{S}$ or $\mathcal{I}$ scores in both scenarios. The dashed vertical line represents the threshold value. c) Expected value of $\mathcal{S}$ and $\mathcal{I}$ for different pairs of samples, sampled from the same individual (in blue) or different ones (orange). Red dots represent samples extracted from pairs of twins. The dashed lines represents the theoretical upper bound (see Methods) for both $\mathcal{S}$ and $\mathcal{I}$ ($\gamma = 12$). d) Distribution of $\mathcal{S}$ in both scenarios (orange heterologous, blue autologous) for different pairs of samples, $M = 5000$. The distributions in red correspond to a pair of samples extracted from twins. e) Detection Error Trade-off (DET) graph for both summary statistics and different sample sizes M. $\mathcal{I}$ ($\gamma = 12$) outperforms $\mathcal{S}$ in all scenarios. f) AUROC (Area Under Receiver Operating Characteristic), as a function of M. The AUROC is a traditional measure of the quality of a binary classifier (a score closer to one indicates a better classifier). The results are shown for $\mathcal{S}$ and $\mathcal{I}$ both in the default case (only the β chain considered) or for the full (α-β) receptor.

The $\mathcal{S}$ score can be improved upon by exploiting the fact that some receptors are much more likely than others to be generated during V(D)J-recombination, with variations in generation probability (P_{gen}, Murugan et al. (2012); Marcou et al. (2018); Sethna et al. (2020)) spanning 15 orders of magnitude. Public sequences (with high P_{gen}) are likely to be found in multiple individuals Venturi et al. (2008), while rare sequences (low P_{gen}) are unlikely to be shared by different individuals, and thus provide strong evidence for the autologous scenario when found in both samples. To account for this information, we define the score:

$$\mathcal{I} = \sum_{\text{shared } s} \left[\ln\left(1/P_{\text{gen}}(s)\right) - \gamma\right], \tag{5.1}$$

which accounts for Shannon's "surprise" $\ln(1/P_{\text{gen}})$—a measure of unexpectedness—associated with each shared sequence s, so that rare shared sequences count more than public ones. The constant γ depends on the repertoire's clonal structure and is set to 12 in the following (see Methods for an information-theoretic derivation). P_{gen} is computed using models previously trained on data from multiple individuals Marcou et al. (2018). Small differences reported between the P_{gen} of distinct individuals justify the use of a universal model Sethna et al. (2020).

We tested the classifiers based on the $\mathcal{S}$ and $\mathcal{I}$ scores on TCRβ RepSeq datasets from 656 individuals Emerson et al. (2017). Sequences were downsampled to mimic experiments where $M_1 = M_2 = M$ cells were analyzed (including a procedure to correct for the limited diversity of the sampled repertoire relative to the full repertoire, see Methods). Similar results may be obtained when M_1 and M_2 are different (see Methods). In Fig. S1c, we plot the mean value of $\mathcal{S}$ (over many draws of $M = 5000$ receptors) for each individual (autologous scenario, in blue) and between pairs of different individuals (heterologous scenario, in orange). The two scenarios are clearly discernable under both scores. This result holds for pairs of monozygotic twins obtained from a distinct dataset Pogorelyy et al. (2018a) (pink dots), consistent with previous reports that twins differ almost as much in their repertoires as unrelated individuals Zvyagin et al. (2014); Pogorelyy et al. (2018a); Tanno et al. (2020). Heterologous scores (orange dots) vary little, and may be bounded from above by a theoretical prediction (dashed line) based on a model of recombination Elhanati et al. (2018) (see Methods). On the other hand, autologous scores (blue dots) show several orders of magnitude of variability across individuals. These variations stem from the clonal structure of the repertoire, and correlates with measures of diversity, which is known to vary a lot between individuals and correlates with age Britanova et al. (2016), serological status, and infectious disease history Sylwester et al. (2005); Khan et al. (2002). To explore the worst case scenario of discriminability, hereafter we will focus on the individual with the lowest autologous $\mathcal{S}$ found in the dataset.

The sampling process introduces an additional source of variability within each

individual. Two samples of blood from the same individual do not contain the exact same receptors, and the values of $\mathcal{S}$ and $\mathcal{I}$ is expected to vary between replicates. Example of distributions for $\mathcal{S}$ between different pairs of replicates in the same (blue) and in different individual are given in Fig. S1d. The distribution of $\mathcal{S}$ is well-approximated by a Poisson distribution, while $\mathcal{I}$ follows approximately a compound distribution of a normal and Poisson distributions (see Methods for details). Armed with these statistical models of variations, we can predict upper bounds for the false negative and false positive rates. As seen from the detection error trade-off (DET) graph Fig. S1e, the Immprint classifier performs very well for a few thousand receptors with an advantage for $\mathcal{I}$.

With $10,000$ cells, corresponding to ~ 10 μL of blood, Immprint may simultaneously achieve a false positive rate of $< 10^{-16}$ and false negative rate of $< 10^{-6}$, allowing for the near-certain identification of an individual in pairwise comparisons against the world population $\sim 10^{10}$. When a large reference repertoire has been collected ($M_1 = 1,000,000$, corresponding from $\sim$ 1mL of blood), an individual can be identified with just 100 cells.

The AUROC estimator (Area Under the Curve of the Receiver Operating Characteristic), a typical measure of a binary classifier performance, can be used to score the quality of the classifier with a number between 0.5 (chance) and 1 (perfect classification). The $\mathcal{I}$ score outperforms the $\mathcal{S}$ score (Fig. S1f), particularly above moderate sample sizes ($M \approx 5000$). Both scores can be readily generalized to the case of paired receptors TCR$\alpha\beta$, when the pairing of the two chains is available (through single-cell sequencing Dash et al. (2011); Redmond et al. (2016); Grigaityte et al. (2017) or computational pairing Howie et al. (2015)), using $P_{\text{gen}}(\alpha, \beta) = P_{\text{gen}}(\alpha) \times P_{\text{gen}}(\beta)$ Dupic et al. (2018) for the generation probability of the full TCR. Because coincidental sharing of both chains is substantially rarer than with the β chain alone, using the paired chain information greatly improves the classifier.

The previous results used samples obtained at the same time. However, immune repertoires are not static: interaction with pathogens and natural aging modify their composition. The evolution of clonal frequencies will decrease Immprint's reliability with time, especially if the individual has experienced immune challenges in the meantime.

To study the effect of short-term infections, we analyzed an experiment where 6 individuals were vaccinated with the yellow fever vaccine, which is regarded as a good model of acute infection, and their immune system was monitored regularly through blood draws Pogorelyy et al. (2018c). We observe an only moderate drop in $\mathcal{S}$ caused by vaccination (Fig. S2a). This is consistent with the fact that infections lead to the strong expansion of only a limited number of clones, while the rest of the immune system stays stable DeWitt et al. (2015); Wolf et al. (2018); Qi et al.

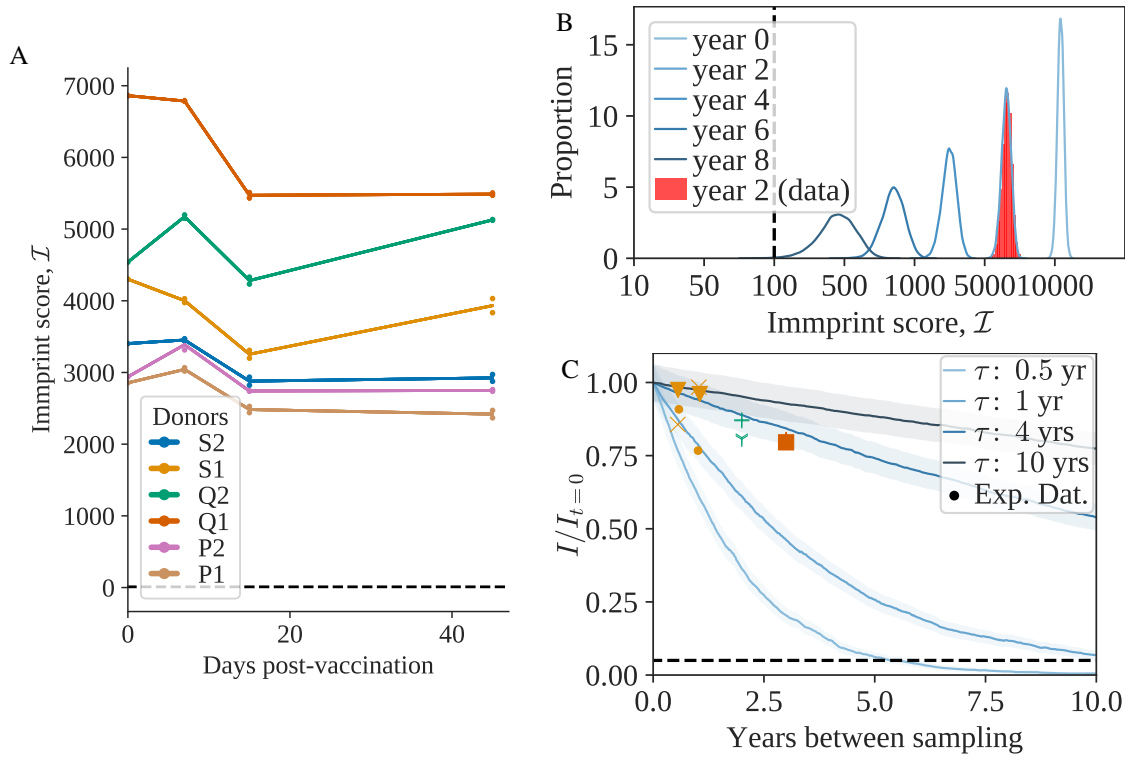

Figure S2: a) Evolution of $\mathcal{S}$ during vaccination, between a sample taken at day 0 (vaccination date) and at a later timepoint. Each color represents a different individual. Each pair timepoint/individual has two biological replicates. The dashed line represents the threshold value. b) Evolution of $\mathcal{S}$ between a sample taken at year 0 and a later timepoint. Blue histograms show theoretical estimates, and the red histogram corresponds to a real dataset. c) Evolution of the (normalized) mean of $\mathcal{S}$ as a function of time for different values of the turnover rate τ. The dashed line represents the threshold value divided by the smallest value of $\mathcal{S}_{t=0}$ in the data.

(2016); **?**. While other types of infections, auto-immune diseases, and cancers may affect Immprint in more substantial ways, our result suggests that it is relatively robust to changes in condition.

We then asked how stable Immprint is over long times. Addressing this issue is hampered by the lack of longitudinal datasets over long periods, so we turn to mathematical models Borghans and De Boer (2007); Thomas-Vaslin et al. (2008); Desponds et al. (2016); Lythe et al. (2015); Greef et al. (2019) to describe the dynamics of the repertoire. Following the model of fluctuating growth rate described in Ref. Desponds et al. (2016), we define two typical evolutionary timescales for the immune system: τ, the typical turnover rate of T-cell clones, and θ, which represents the typical time for a clonotype to grow or shrink by a factor two as its growth rate fluctuates. The model predicts a power-law distribution for the clone-size distribution, with exponent $-1-\tau/2\theta$. This exponent has been experimentally measured to be ≈ -2, which leaves us with a single parameter τ, and $\theta = \tau/2$. An example of simulated evolution of Immprint with time is shown in Fig. S2b. The highlighted histogram represents a data point at two years obtained from Chu et al. (2019). While a fit is possible for this specific individual, the τ parameter is not universal, and we expect it to vary between individuals, especially as a function of age. In Fig. S2c we explore a range of reasonable values for the clone turn-over rate τ (from 6 months to 10 years), and their effect on the stability of Immprint. We observe that for most individuals, bar exceptional events, Immprint should conserve its accuracy for years or even decades.

In summary, we demonstrated that the T-cells present in small blood samples provide a somatic and long-lived barcode of human individuality, which is robust to immune challenges and stable over time. Unlike genome sequencing, repertoire sequencing can discriminate monozygotic twins with the same accuracy as unrelated individuals. However, a person's unique immune fingerprint can be completely wiped out by a hematopoietic stem cell transplant Buhler et al. (2020). Immprint is implemented in a python package and webapp (see Methods) allowing the user to determine the autologous or heterologous origin of a pair of repertoires. Beyond identifying individuals, the tool could be used to check for contamination or labelling errors between samples containing TCR information. The repertoire information used by Immprint can be garnered not only from RepSeq experiments, but also from RNA-Seq experiments, which contain thousands of immune receptor transcripts Li et al. (2017); Bolotin et al. (2017). Relatively small samples of immune repertoires are enough to uniquely identify an individual even among twins, with potential forensics applications. At the same time, unlike genetic data from genomic or mRNA sequencing, Immprint provides no information about kin relationships, very much like classical fingerprints, and avoids privacy concerns about disclosing genetic information shared with non consenting relatives.

5.2. Methods

5.2.1 Datasets & Pre-processing

We use four independant RepSeq datasets in this study: *(i)* genomic DNA from Peripheral blood mononuclear cells (PBMCs) from 656 healthy donors Emerson et al. (2017); *(ii)* cDNA of PBMCs sampled from three pairs of twins, before and after a yellow-fever vaccination Pogorelyy et al. (2018c); *(iii), (iv)* two longitudinal studies of healthy adults Chu et al. (2019); Britanova et al. (2016) .

CDR3 nucleotide sequences were extracted with MIGEC Shugay et al. (2014) (for the second dataset) coupled with MiXCR Bolotin et al. (2015). We also extract the frequency of reads from the three datasets. The non-productive sequences were discarded (out-of-frame, non-functional V gene, or presence of a stop codon). The generation probability (P_{gen}) was computed using OLGA Sethna et al. (2019), with the default TCRβ model. The frequency of each clone was estimated through the number of reads, which we use as an imperfect proxy for the number of cells.

5.2.2 Discrimination scores

To discriminate between the autologous and heterologous scenarios, we introduce a log-likelihood ratio test between the two possibilities:

$$\mathcal{I} = \sum_s \ln \frac{P(y_1(s), y_2(s)|\text{autologous})}{P(y_1(s), y_2(s)|\text{heterologous})}, \tag{5.2}$$

where $y_1(s) = 1$ if the sequence s is found in sample 1, and 0 otherwise; likewise $y_2(s) = 1$ if s is in sample 2. The sum runs over all potential sequences s, including unseen ones. To be present in a sample, a sequence s first has to be present in the repertoire. This occurs with probability $1 - (1 - p(s))^{N_c}$, where N_c is the total number of clonotypes in the repertoire, and $p(s)$ is the probability of occurence of sequence s (resulting from generation and selection, see below). Second, it must be picked in a sample of size M, with probability $1 - (1 - f)^M \approx Mf$ (assuming $Mf \ll 1$) depending on its frequency f, which is distributed according to the clone size distribution $\rho(f)$. We checked that $f(s)$ and $P_{\text{gen}}(s)$ were not correlated). Then one can write

$$P(y_1(s) = 1, y_2(s) = 1 \mid \text{autologous}) \approx \left(1 - e^{-N_c p(s)}\right) M_1 M_2 \int df\, \rho(f)\, f^2, \tag{5.3}$$

$$P(y_1(s) = 1, y_2(s) = 0 \mid \text{autologous}) \approx \left(1 - e^{-N_c p(s)}\right) \frac{M_1}{N_c} \text{ and } 1 \leftrightarrow 2, \tag{5.4}$$

$$P(y_1(s) = 0, y_2(s) = 0 \mid \text{autologous}) \approx 1 - \left(1 - e^{-N_c p(s)}\right) \frac{M_1 + M_2}{N_c}, \tag{5.5}$$

where we've used $\int df\, \rho(f)\, f = 1/N_c$. For the heterologous case the probability factorizes as:

$$P(y_1(s), y_2(s) \mid \text{heterologous}) = P_1(y_1(s))P_2(y_2(s)), \tag{5.6}$$

with

$$P_a(y_a(s) = 1) \approx \left(1 - e^{-N_c p(s)}\right) \frac{M_a}{N_c}, \quad a = 1, 2. \tag{5.7}$$

Since only the term $y_1(s) = y_2(s) = 1$ (shared sequences) is different between the autologous and heterologous cases, we obtain:

$$\mathcal{I} = \sum_{\text{shared } s} \left[\ln(N_c^2 \langle f^2 \rangle) - \ln\left(1 - e^{-N_c p(s)}\right)\right]. \tag{5.8}$$

Further assuming $N_c p(s) \ll 1$, and $p(s) = P_{\text{gen}}(s) q^{-1}$ (where q accounts for selection Elhanati et al. (2018) and $P_{\text{gen}}(s)$ is the probability of sequence generation Marcou et al. (2018)), the score simplifies to Eq. 5.1, with $\gamma = -\ln(q N_c \langle f^2 \rangle) = \ln(q^{-1} \langle f \rangle / \langle f^2 \rangle)$. The factor γ depends on unknown parameters of the model, but can be estimated assuming a power-law for the clone size distribution Touzel et al. (2020), $\rho(f) \propto f^{-2}$ extending from $f = 10^{-11}$ to $f = 0.01$, and $q = 0.01$ Elhanati et al. (2018), yielding $\gamma \approx 12.24$. Alternatively we optimized γ to minimize the AUROC, yielding $\gamma \approx 15$. Since performance degrades quickly for larger values, we conservatively set $\gamma = 12$.

5.2.3 Estimating mean scores from RepSeq datasets

To estimate the autologous $\mathcal{S}$ and $\mathcal{I}$ of two samples of size M_1 and M_2 in the absence of true replicates, we computed their expected values from a single dataset containing N reads, from which two random subsamples of sizes M_1 and M_2 were taken. The mean value of $\mathcal{S}$ is equal to $\langle \mathcal{S} \rangle = \sum_s (1 - (1 - f(s))^{M_1})(1 - (1 - f(s))^{M_2})$, where $f(s)$ is the true (and unknown) frequency of sequence s. A naive estimate of $\langle mS \rangle$ may be obtained by repeatly resampling subsets of sizes M_1 and M_2 from the observed repertoire, calculate $\mathcal{S}$ for each draw, and average. One get the same result by replacing $f(s)$ by $\hat{f}_s = n(s)/N$ in the previous formula, where $n(s)$ is the number of s reads in the full dataset, and $N = \sum_s n(s)$. However, this naive estimate leads to a systematic overestimate of the sharing (visible when compared with biological replicates, simply because this procedure overestimates the probability of resampling rare sequences, in particular singletons whose true frequency may be much lower that $1/N$. A similar bias occurs when computing $\mathcal{I}$. To correct for this bias, we look for a function $h(n)$ that satisfies for all f:

$$\langle h(n) \rangle \equiv \sum_n \binom{N}{n} f^n (1 - f)^{N-n} h(n) = (1 - (1 - f)^{M_1})\ (1 - (1 - f)^{M_2}), \tag{5.9}$$

so that $\langle \mathcal{S} \rangle$ and $\langle \mathcal{I} \rangle$ can be well approximated by:

$$\langle \mathcal{S} \rangle \approx \sum_s h(n(s)), \tag{5.10}$$

$$\langle \mathcal{I} \rangle \approx -\sum_s h(n(s)) \left[\ln(1/P_{\text{gen}}(s)) - \gamma \right]. \tag{5.11}$$

Expanding the right-hand side of Eq. 5.9 into 4 terms, we find that $h(n) = 1 - g_{M_1}(n) - g_{M_1}(n) + g_{M_1+M_2}(n)$ satisfies Eq. 5.9 provided that:

$$\sum_n \binom{N}{n} f^n (1-f)^{N-n} g_M(n) = (1-f)^M. \tag{5.12}$$

Under the change of variable $x = f/(1-f)$, the expression becomes:

$$\sum_n \binom{N}{n} x^n g_M(n) = (1+x)^{N-M} = \sum_n \binom{N-M}{n} x^n. \tag{5.13}$$

Identifying the polynomial coefficients in x^n on both sides yields:

$$g_M(n) = \binom{N-M}{n} \Big/ \binom{N}{n}. \tag{5.14}$$

These corrected estimates agree with the direct estimates using biological replicates).

Similarly, $\langle \mathcal{S} \rangle$ and $\langle \mathcal{I} \rangle$ in heterologous samples can be estimated using:

$$\langle \mathcal{S} \rangle \approx \sum_s [1 - g_{M_1}(n(s))][1 - g_{M_2}(n'(s))], \tag{5.15}$$

$$\langle \mathcal{I} \rangle \approx \sum_s [1 - g_{M_1}(n(s))][1 - g_{M_2}(n'(s))] \left[\ln(1/P_{\text{gen}}(s)) - \gamma \right]. \tag{5.16}$$

where $n(s)$ and $n'(s)$ are the empirical counts of sequence s in the two samples.

Theoretical upper bound on heterologous scores

When the two samples were extracted from two different people (heterologous scenario), we can use the universality of the recombination process to give upper bounds on the values of $\mathcal{S}$ and $\mathcal{I}$. These bounds are represented by the dashed lines in FigS1c). If two samples of respectively M_1 and M_2 unique sequences are extracted from two different individuals, the number of shared sequences between them is given by Elhanati et al. (2018):

$$\langle \mathcal{S} \rangle_{\text{heterologous}} \leq \sum_s \left(1 - (1-p(s))^{M_1}\right) \left(1 - (1-p(s))^{M_2}\right) \lessapprox M_1 M_2 \sum_s p(s)^2 = M_1 M_2 \langle p(s) \rangle. \tag{5.17}$$

$p(s)$ is the probability of finding a sequence s in the blood. Following Elhanati et al. (2018), we make the approximation $p(s) = P_{\text{gen}}(s)q^{-1}$, where the $q = 0.01$ factor is the probability that a generated sequences passes selection. Then $\langle p(s)\rangle$ can be estimated from the mean over generated sequences. Similarly, we can estimate $\mathcal{I}$ as

$$\langle \mathcal{I}\rangle_{\text{heterologous}} \lessapprox M_1 M_2 \sum_s p(s)^2 \left[\ln\left(1/P_{\text{gen}}(s)\right) - \gamma\right] = -M_1 M_2 \langle p(s)[\gamma + \ln(qp(s))]\rangle, \tag{5.18}$$

which is also estimated from the mean over generated sequences.

5.2.4 Error rate estimates

To make the quantitative predictions shown in Fig. S1, we need to constrain the tail behavior of the distributions of $\mathcal{S}$ and $\mathcal{I}$, for the two scenarios.

The $\mathcal{S}$ statistic can be rewritten as a sum of Bernouilli variables over all possible sequences, each with a parameter corresponding to its probability of being present in both samples, either in the autologous or the heterologous case. Therefore $\mathcal{S}$ is a Poisson binomial distribution, a sum of independent Bernouilli variables with potentially different parameters. The variance and tails of that distribution are bounded by those of the Poisson distribution with the same mean, denoted by m_a for the autologous case, and m_h for the heterologous case).

Thanks to that inequality, the rates of false negatives and false positives for a given threshold r are bounded by:

$$P(\mathcal{S} < r|\text{autologous}) \leq Q(r+1, m_a), \qquad P(\mathcal{S} > r|\text{heterologous}) \leq 1 - Q(r+1, m_h), \tag{5.19}$$

where Q is the regularized gamma function, which appears in the cumulative distribution function of the Poisson distribution. The mean autologous score m_a is estimated from experimental data: we use the smallest value of $\langle \mathcal{S}\rangle$ in the Emerson dataset and Eq. 5.10. To compute m_h, we use the semi-theoretical prediction made using the universality of the recombination process, Eq. 5.17.

Similarly, $\mathcal{I}$ can be viewed as a sum of $\mathcal{S}$ independent random variables, all following the distribution of $\ln(1/P_{\text{gen}}) - \gamma$. However, this distribution differs in the two scenarios. Sequences shared between more than one donor have an higher probability of being generated, their $\ln(P_{\text{gen}})$ distribution has higher mean and smaller variance.

The sum is composed of a relatively large number of variables in most realistic scenarios. Hence, we rely on the central limit theorem to approximate it by a normal distribution, of mean and variance proportional to $\mathcal{S}$. Explicitly:

$$P(\mathcal{I} < r|\text{autologous}) = \frac{1}{2}\sum_{\mathcal{S}=0}^{\infty} \frac{(m_a)^{\mathcal{S}}\, e^{-m_a}}{\mathcal{S}!} \left(1 + \text{erf}\left(\frac{r - \mathcal{S}\langle \ln(1/P_{\text{gen}}) - \gamma\rangle}{\sqrt{2\mathcal{S}\text{Var}[\ln(1/P_{\text{gen}}) - \gamma]}}\right)\right), \tag{5.20}$$

$$P(\mathcal{I} > r|\text{heterologous}) = \frac{1}{2}\sum_{\mathcal{S}=0}^{\infty} \frac{(m_h)^{\mathcal{S}}\, e^{-m_h}}{\mathcal{S}!} \left(1 - \text{erf}\left(\frac{r - \mathcal{S}\langle \ln(1/P_{\text{gen}}) - \gamma\rangle_{\text{shared}}}{\sqrt{2S\text{Var}[\ln(1/P_{\text{gen}}) - \gamma]_{\text{shared}}}}\right)\right). \tag{5.21}$$

The AUROC are computed based on these estimates, by numerically integrating the true positive rate $P(\mathcal{S}, \mathcal{I} < r|\text{heterologous})$ with respect to the false negative rate $P(\mathcal{S}, \mathcal{I} < r|\text{autologous})$ as the threshold r is varied.

5.2.5 Modeling the evolution of autologous scores

We use the model of Ref. Desponds et al. (2016) to describe the dynamics of individual T- or B-cell clone frequencies f under a fluctuating growth rate reflecting the changing state of the environment and the random nature of immune stimuli:

$$\frac{df}{dt} = \left[-\frac{1}{\tau} + \frac{1}{2\theta} + \frac{1}{\sqrt{\theta}}\eta(t)\right] f(t), \tag{5.22}$$

where $\eta(t)$ is a Gaussian white noise with $\langle \eta(t)\rangle = 0$ and $\langle \eta(t)\eta(t')\rangle = \delta(t - t')$.

With the change of variable $x = \ln(f)$, these dynamics simplify to a simple Brownian motion in log-frequency: $\partial_t x = -\tau^{-1} + \theta^{-1/2}\eta(t)$. In that equation, τ appears as the decay rate of the frequency, while θ is the timescale of the noise, interpreted as the typical time it takes for the frequency to rise or fall by a logarithmic unit owing to fluctuations. Considering a large population of clone, each with their independent frequency evolving according to Eq. 5.22, and a source term at small f corresponding to thymic exports, one can show that the steady-state probability density function of f follows a power-law Desponds et al. (2016), $\rho(f) \propto f^{-\alpha}$, with exponent $\alpha = 1 + 2\theta/\tau$. α was empirically found to be ≈ 2 in a wide variety of immune repertoires Weinstein et al. (2009); **?**); Oakes et al. (2017); Touzel et al. (2020), implying $2\theta \approx \tau$. The turn-over time τ is unknown, and was varied from 1/2 year to 10 years in the simulations.

We simulated the evolution of human TRB repertoires by starting with the empirical values of the frequencies of each observed clones, $f(s, 0) = \hat{f}(s, 0) = n(s, 0)/N$ from the analysed datasets. A sample of size M was randomly selected with respect to these frequencies, and the frequencies of the clones captured in that sample were then evolved with a time-step of 2 days using Euler-Maruyama's method, which is exact in the case of Brownian motion. Clones with frequencies falling below 10^{-11} (corresponding to a single cell in the organism) were removed. At each time $t > 0$, we measured the mean value of $\mathcal{S}$ with the formula $\sum_s (1 - (1 -$

$f(s,t))^M$) where the sum runs over the sequences captured in the initial sample.

5.3. Supplementary figures

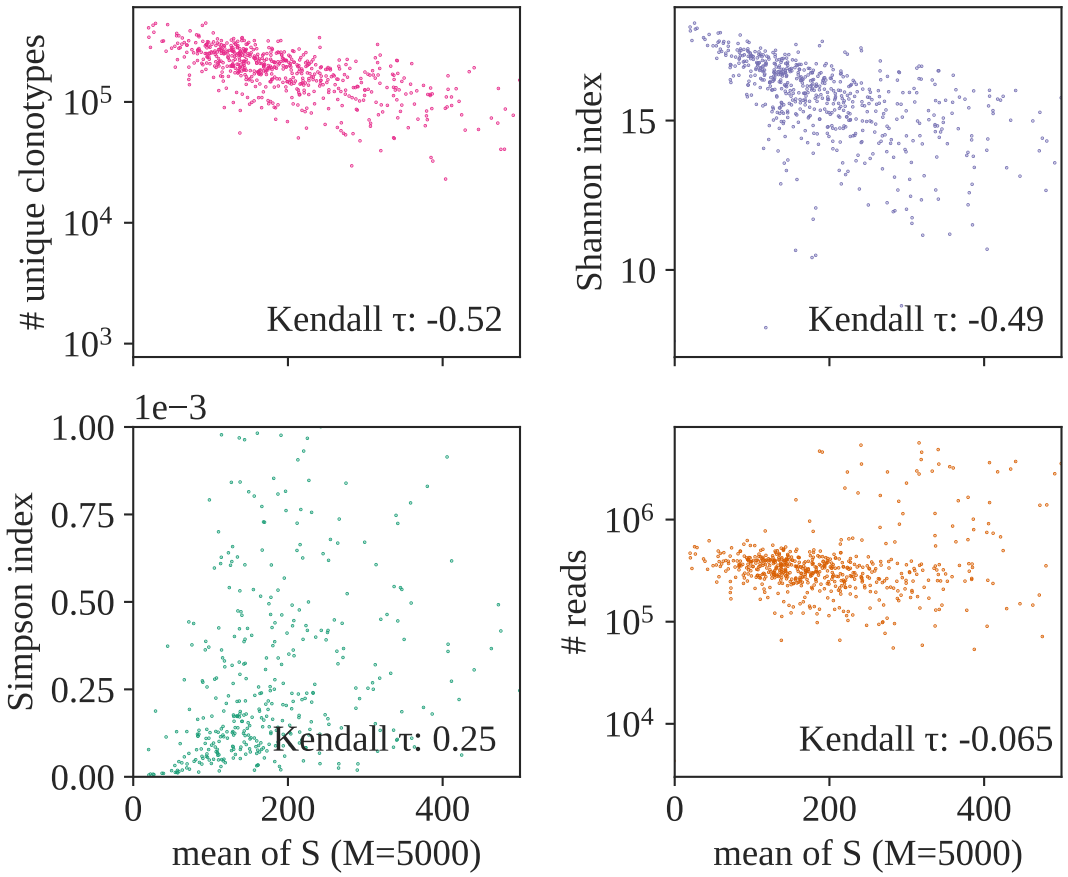

Figure S1: Comparison between the mean of $\mathcal{S}$ (autologous case), and three common diversity measures: the number of unique sequences found in the dataset (top left), the Shannon index, $-\sum \hat{f}_s \ln \hat{f}_s$ (top right), the Simpson index (bottom left), and the total number of reads in each datasets (bottom right). All the diversity measures show a strong correlation with $\mathcal{S}$, but the correlation with the sequencing depth is low.

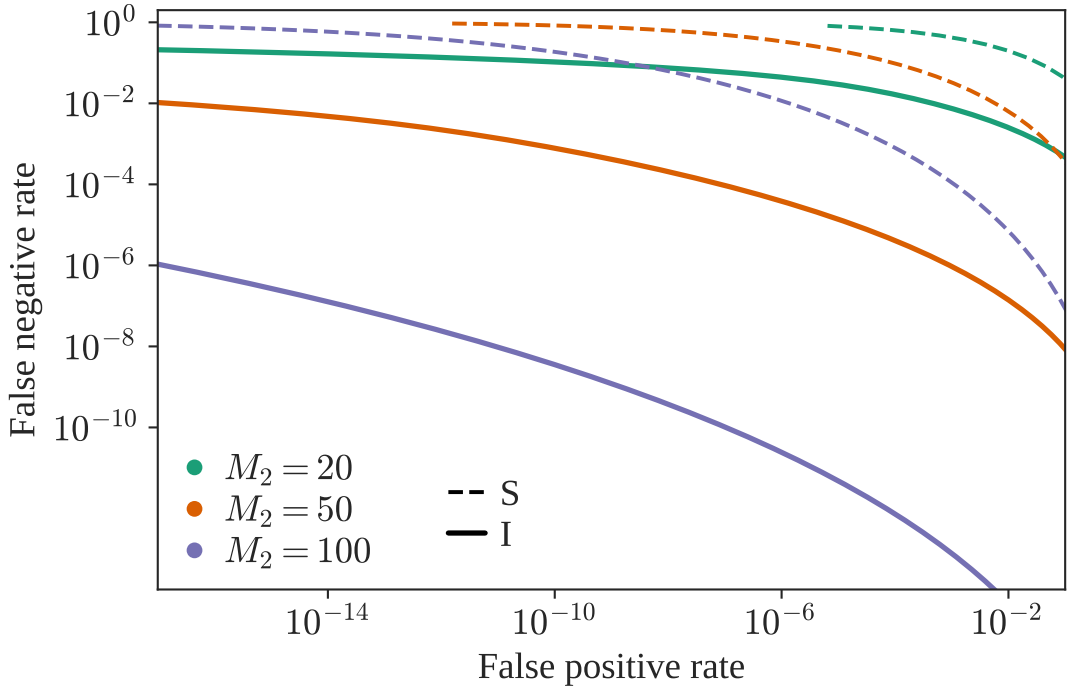

Figure S2: Detection Error Trade-off (DET) graph for both summary statistics, between a large sample (full dataset, $M_1 = 10^6$) and a smaller one, of size $M_2 = M$.

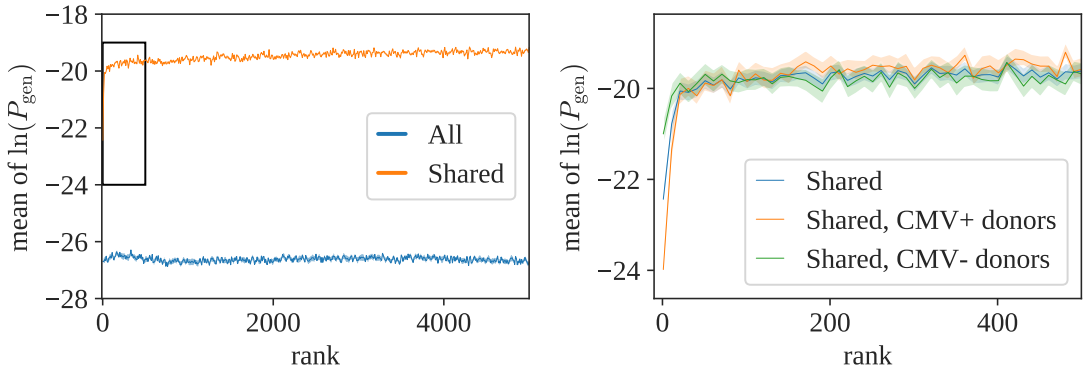

Figure S3: Left: Mean value of P_{gen} as a function of the rank of the clonotype, for generic sequences (blue) and sequences shared between more than two donors (orange). The mean stays flat indicating that the probability of being generated does not generally depend on the clonotype size. There is an exception (black rectangle), shown as a close-up on the right panel. The top twenty clones, when shared between donors, have a smaller probability of being generated than expected by chance. This difference is likely to be driven by convergent selection against common pathogens, since CMV positive donors show a more prononced effect than CMV negative ones.

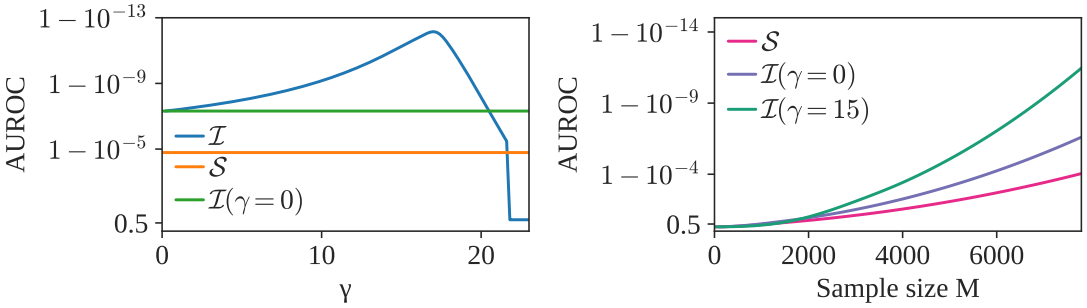

Figure S4: Left panel: AUROC (Area Under Receiver Operating Characteristic) of $\mathcal{I}$, as a function of γ ($M = M_1 = M_2 = 5000$). We observe an optimum near $\gamma = 15$. Right panel: AUROC as a function of M, for $\mathcal{S}$, $\mathcal{I}(\gamma = 0)$, and $\mathcal{I}(\gamma = 15)$.

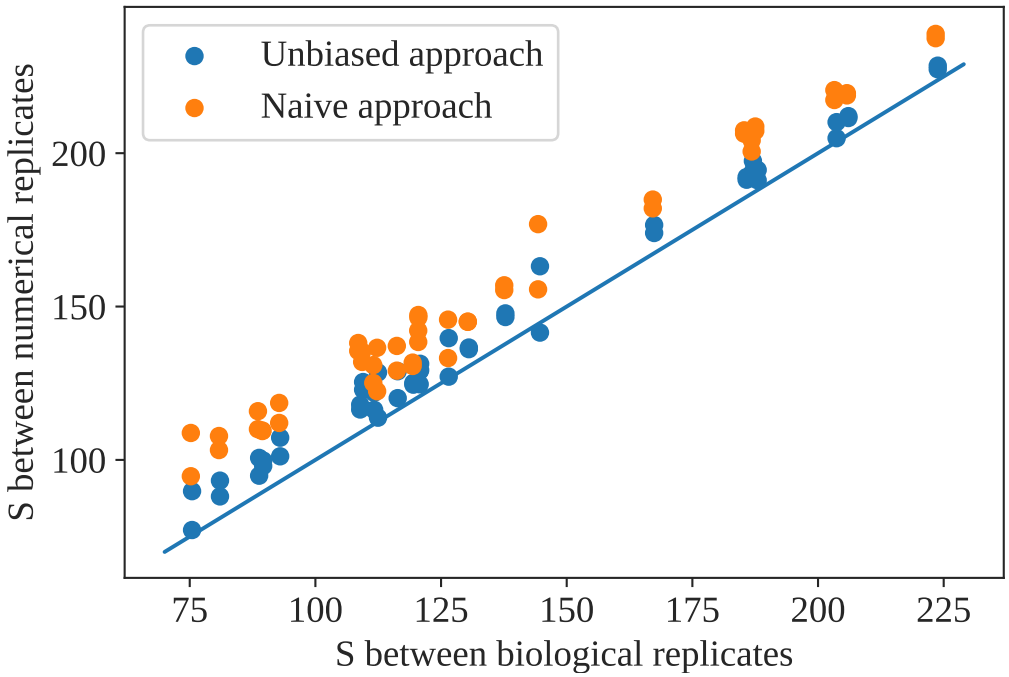

Figure S5: Naive and corrected estimates of the autologous $\mathcal{S}$ from single datasets, versus its values computed using true biological replicates from Ref. Pogorelyy et al. (2018a).

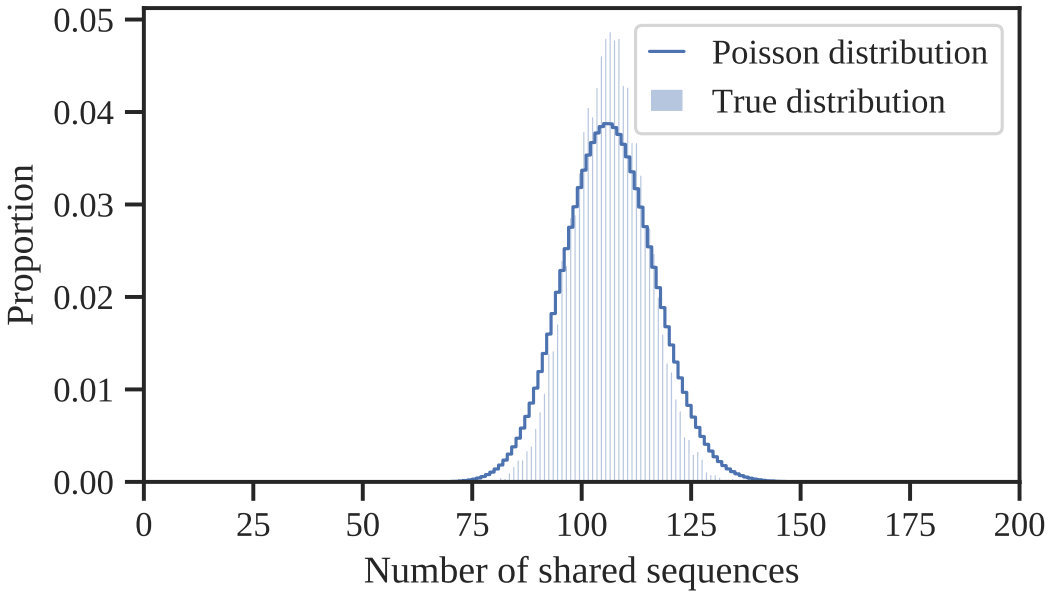

Figure S6: Comparison between the distribution of $\mathcal{S}$ obtained by computationally and repeatedly downsampling a single repertoire from Ref. Emerson et al. (2017) with $M = 5,000$ (histogram), and a Poisson distribution of the same mean (full line).

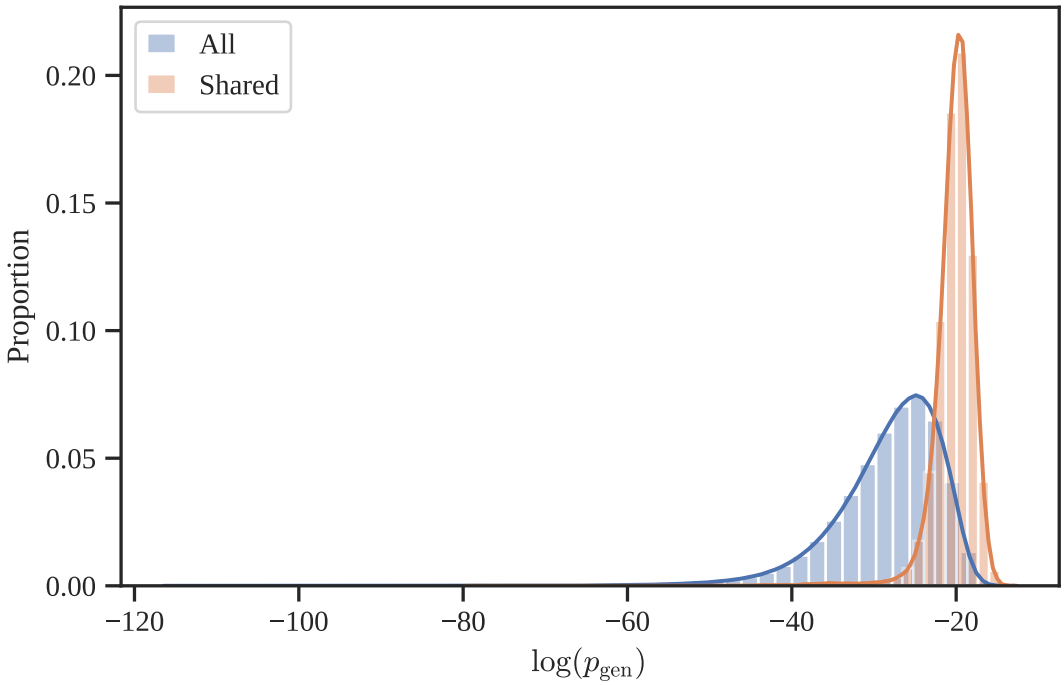

Figure S7: Distribution of $\ln(P_{\text{gen}})$ for generic sequences, and for sequences shared between heterologous samples.

CHAPTER

6

NOISET

In chapter 6, I explain through the designing of a software which are the different analyses to perform on longitudinal TCR RepSeq data when wanting to decipher the dynamics of T-cell abundances. My starting point was first to build a python library to use easily for future needs the methods developed in Puelma Touzel et al. (2020) to infer sampling and experimental noise from RepSeq data and use this knowledge to detect significant expansion or contraction of TCR clonal abundances between two-time points. I also develop in the chapter the improvements I achieved for these two tasks and new features to use for anyone who is working with this kind of data. My work here was to understand the different probabilistic models that one can use to assess experimental noise (and sampling noise) described in the introduction of this manuscript. Model selection here is not developed but can be done regarding each practical case: one can compare the likelihood of the model, or the BIC score if models do not have same number of parameters. Another important point is the computation time, which can be a reason to prefer a model to another.

6.1. Introduction

Cells of the adaptive immune system, T and B lymphocytes, recognize molecules foreign to our body and protect us against pathogenic threats. These cells also have the ability to eliminate cells that harbor anomalies, such as cancer cells. Lymphocytes perform this discrimination task between potentially dangerous and normally functioning “self" molecules using specialized receptors on their surface that constantly

sample and bind molecules in our organisms. Each cell has one type of receptor and the system relies on a large diversity of a repertoire of different receptors expressed on over 10^9 different B or T cells to protect the organism against infections Robins et al. (2009); **?**); Lythe et al. (2015); DeWitt et al. (2016).

The composition of the repertoire contains information about past infections and conditions. Reading this information requires quantitatively understanding the natural repertoire dynamics. Upon recognition of a pathogenic molecule, the recognizing cell proliferates making many cells with the same receptor, forming a clone, which enables fast infection clearance. New cells are constantly produced and introduced into this diverse repertoire. Additionally to specific stimulation, cells also undergo random divisions. Each cell has a finite lifetime and clones can go extinct if all the cells of that clone die. Together these processes define a natural dynamics of the repertoire, which leads to a constantly changing set of different cells present at different frequencies.

High-Throughput Repertoire Sequencing (RepSeq) of T and B cell receptors (TCR and BCR) Weinstein et al. (2009); Robins et al. (2009); Boyd et al. (2009); Benichou et al. (2012); Six et al. (2013); Robins (2013); Georgiou et al. (2014); Heather et al. (2017); Minervina et al. (2019b); Rubelt et al. (2017) enables us to study the dynamics of lymphocytes at the resolution of single clones, by comparing their concentrations across timepoints or conditions. To detect biologically relevant clones, one must be able to distinguish true differences in clone frequencies from experimental noise. This variability has three sources. First, laboratories use various sequencing and sample preparation protocols using either gDNA or cDNA (with or without unique molecular identifiers), with different outcomes in terms of amplification bias and errors Heather et al. (2017); Barennes et al. (2020). This makes it difficult to reliably estimate TCR or BCR clonal frequencies from sequence counts. Secondly, in the case of cDNA based sequencing, these uncertainties are not solely due to different sample preparation but have a more fundamental, biological source. mRNA is produced in bursts Elowitz et al. (2002); Ozbudak et al. (2002); Cai et al. (2006); Taniguchi et al. (2010); Hornos et al. (2005), which adds a natural longtailed noise to the sequencing read distribution. Thirdly, one must translate immune information contained in a few milliliters of blood to the whole repertoire. To describe these sources of variability, one needs a probabilistic approach.

Puelma Touzel et al. Puelma Touzel et al. (2020) developed a statistical model to identify responding clones using sequence counts in longitudinal RepSeq data. This model captures features of a repertoire response to a single, strong perturbation (e.g. yellow fever vaccination), giving rise to a fast transient response dynamics. The method was proposed as an alternative to commonly used tests such as Fisher's exact test Balachandran et al. (2017) or beta binomial models Rytlewski et al. (2019). Its main innovation is to account for the different sources of biological and

experimental noise in the clone count measurements in a Bayesian way, allowing for a more reliable detection of expanded or contracted clones.

Here we briefly review the ideas behind the method that calibrates the noise and we introduce NoisET (Noise sampling learning and Expansion detection of TCRs), an easy-to-use python package that implements this method and extends it to datasets of diverse origin describing the clonal repertoire response to acute infections. We also review several applications of this approach.

6.2. Model

In order to correctly identify expanding or contracting clonotypes, whether after direct antigenic stimulation or due to random cell division and death, we need to correctly separate biological and experimental noise from the lymphocyte dynamics. The main idea behind the Bayesian probabilistic modeling method implemented in the NoisET software is learning probabilistic distributions describing sampling and experimental noise from empirical frequencies of TCR counts in biological replicate samples from the same individual. In this section we introduce the two types of models implemented in NoisET: the noise model and the response model.

6.2.1 Modeling experimental noise

TCR sequencing (TCRseq) methods, depending if they are based on DNA or RNA input, produce data with different characteristics. For example, RNA-based methods allow for the usage of unique molecular identifiers (UMI) to limit PCR amplification bias and sequencing errors. Non-UMI methods are better in capturing rare clones which motivates their frequent usage Barennes et al. (2020). During this first step, NoisET learns at the same time the exponent of the underlying power-law TCR frequency distribution, $\rho(f)$ Weinstein et al. (2009) and the parameters of error model between the empirical abundance of one specific TCR clone $\hat{n}$ and its true frequency f: $P(\hat{n}|f)$. NoisET has also the power to learn these distributions constraining the size of the clones we want to take into account for the analysis.

For each TCR clone, the likelihood to sample $\hat{n}$ reads from the first biological replicate and $\hat{n}'$ reads from the second biological replicate is:

$$\mathbb{P}(\hat{n}, \hat{n}'|\Theta) = \int_{f_{\min}}^{1} df \rho(f|\Theta) P(\hat{n}|f, \Theta) P(\hat{n}'|f, \Theta), \tag{6.1}$$

where Θ are the parameters of the noise model which define the error model $P(\hat{n}|f)$, $f_{\min}$ corresponds to the minimum clonal frequency for each individual, and $\rho(f)$, which is the clonal frequency prior known to be a power-law distribution $\propto f^{\alpha}$ Weinstein et al. (2009); Mora and Walczak (2019a). NoisET learns the parameters of the noise model Θ by maximizing the log-likelihood of the observed TCR counts,

$\hat{n}, \hat{n}'$, from the two biological replicates:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N_{obs}} \mathbb{P}(\hat{n}_i, \hat{n}'_i|\Theta). \tag{6.2}$$

Since in RepSeq samples we only partially sample an individual's repertoire, the likelihood in Eq. 6.1 needs to be modified. We condition the likelihood on observing that specific clone in at least one of the two replicates: $\mathbb{P}(\hat{n}+\hat{n}'>0)$. The modified likelihood becomes $\mathbb{P}(\hat{n},\hat{n}'|\hat{n}+\hat{n}'>0) = \mathbb{P}(\hat{n},\hat{n}',\hat{n}+\hat{n}'>0)/\mathbb{P}(\hat{n}+\hat{n}'>0)$. We can also choose to learn the noise model only on clones having a size larger than a certain threshold. In this case the likelihood in Eq. 6.1 becomes: $\mathbb{P}(\hat{n},\hat{n}'|\hat{n}>\hat{n}_{th},\hat{n}'>\hat{n}_{th}) = \mathbb{P}(\hat{n},\hat{n}',\hat{n}>\hat{n}_{th},\hat{n}'>\hat{n}_{th})/\mathbb{P}(\hat{n}>\hat{n}_{th},\hat{n}'>\hat{n}_{th})$.

To take into account different possible sources of noise due to the various RepSeq method, NoisET gives the choice of three different probabilistic distributions to learn the biological and experimental noise in the measured TCR abundances, $P(\hat{n}|f)$:

- The Poisson distribution, $P(\hat{n}|f) = \text{Poisson}(fN_r)$. In this case, the noise parameters Θ are the exponent α of the clone-size distribution $\rho(f) = Cf^{\alpha}$ and the minimum clonal frequency in Eq. 6.1, $f_{\min}$. N_r is the total number of reads in the sample.
- The negative binomial distribution: $P(\hat{n}|f) = \text{NegBin}(\hat{n}; N_r f, N_r f + a(N_r f)^b)$, where $\text{NegBin}(n; x, \sigma)$ is a negative binomial of mean x and variance σ. In this case, $\Theta = (\alpha, a, b, f_{\min})$ with α, $f_{\min}$ being the same parameters as described for the Poisson distribution, and a and b, the parameters of the negative binomial distribution.
- The negative binomial combined with a Poisson distribution: $P(\hat{n}|f) = \sum_{m_i}^{\infty} P(\hat{n}|m_i)P(m_i|f)$, with $P(m_i|f) = \text{NegBin}(m_i; fM, fM + a(fM)^b)$ and $P(\hat{n}|m_i) = \text{Poisson}(m_i N_r/M)$. A clone of size f appears in a sample containing M T-cells on average as fM cells. To account for over dispersion, the number of cells associated to a specific clone is m and follows a negative binomial of mean fM and variance $fM + a(fM)^b$. For each clone the empirical abundance read in the biological sample is distributed according to a Poisson distribution with mean mN_r/M. For this model $\Theta = (\alpha, a, b, M, f_{\min})$.

While the mathematical framework is the same, when applied to identifying expanding clonotypes NoisET uses noise parameters inferred at both time points, contrary to the approach taken in Pogorelyy et al. (2018c) and Puelma Touzel et al. (2020). Experimental conditions at both time points can vary and it is important to use both sets of parameters Θ, to have the correct form of $P(\hat{n}|f,t_1)$ and $P(\hat{n}|f,t_2)$. The exponent of the power-law α and $f_{\min}$ in Eq. 6.1 are the learnt values inferred at the time point for which sequencing depth is the larger.

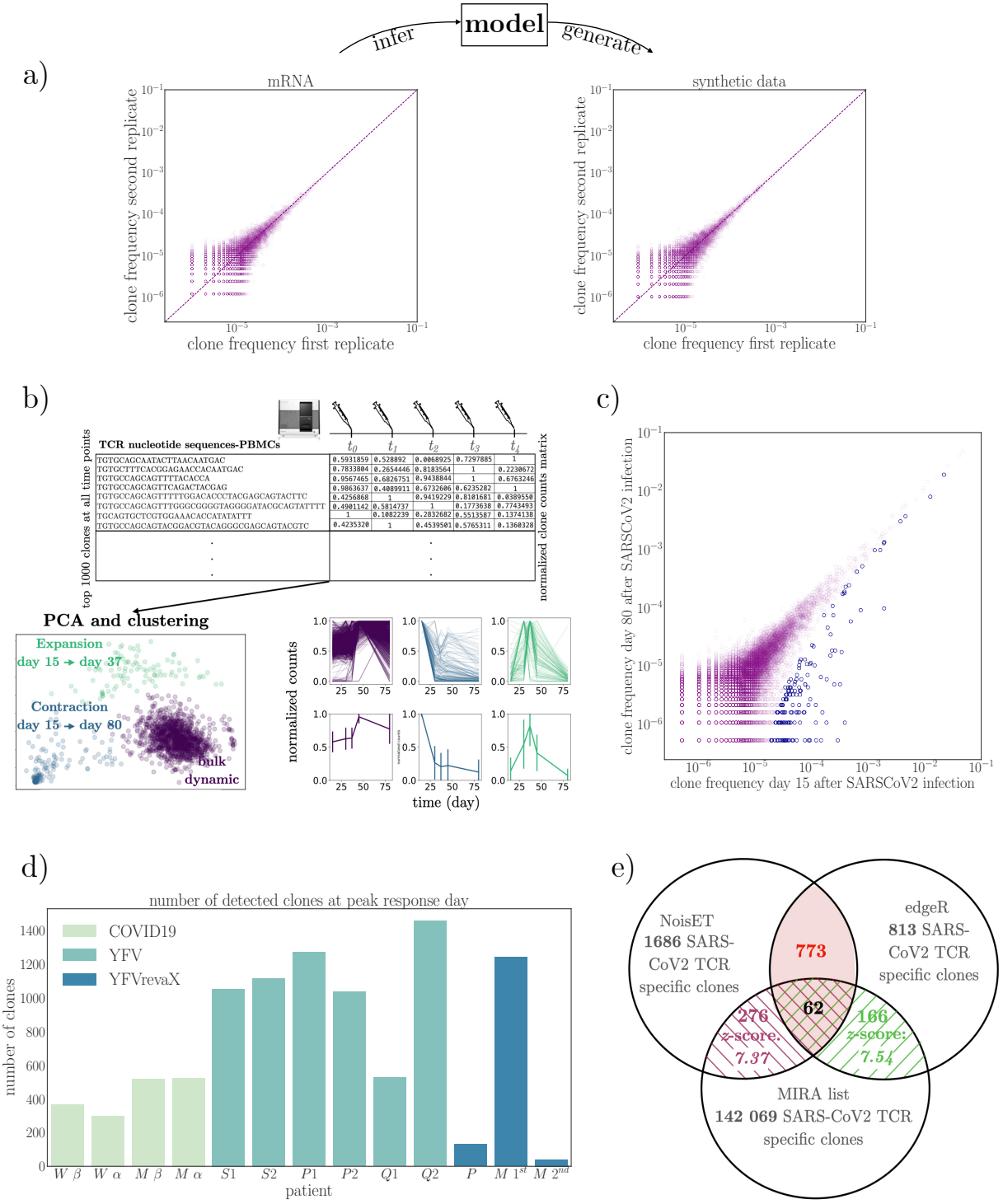
a)
infer
model
generate
mRNA
synthetic data
clone frequency second replicate
clone frequency first replicate
b)
TCR nucleotide sequences-PBMCs
top 1000 clones at all time points
normalized clone counts matrix
PCA and clustering
Expansion
day 15 → day 37
Contraction
day 15 → day 80
bulk
dynamic
normalized counts
time (day)
c)
clone frequency day 80 after SARSCoV2 infection
clone frequency day 15 after SARSCoV2 infection
d)
number of detected clones at peak response day
COVID19
YFV
YFVrevaX
number of clones
patient
e)
NoisET
1686 SARS-CoV2 TCR specific clones
773
edgeR
813 SARS-CoV2 TCR specific clones
62
276
z-score: 7.37
166
z-score: 7.54
MIRA list
142 069 SARS-CoV2 TCR specific clones

Figure S1:

Figure S1: **(a)** Scatter plots of sequence counts from two biological replicates from Pogorelyy et al. (2018a) (left). NoisET learns a statistical model of sequence frequencies and observed counts from these data (here with negative binomial sampling noise model), which can then be used to generate realistic synthetic data (right). **(b)** PCA (Principal Component Analysis) performed on the matrix composed of the normalized clone counts of the top 1000 clones present at every time point of the longitudinal dataset. The clustering of the data projected on the two first principal components enables us to understand different kinds of dynamics for three clusters of clones here (bottom left). The number of clusters can be adjusted in NoisET and should be tested. In this example, this pre-analysis of the longitudinal dataset enabled us to find a significant contracting dynamical trend between day 15 to day 85 and a significant expansion trend between day 15 and day 37 following a mild COVID-19 infection Minervina et al. (2021) (bottom left). The top plots show the individual trajectories in each trend, the bottom plots the average with standard variation error bars. **(c)** Scatter plot of contracted clones from day 15 to day 85 after a mild COVID-19 infection Minervina et al. (2021). Clones detected as contracting by NoisET are shown in blue. **(d)** The number of responding clones detected by NoisET (using a two step noise model) for 3 studies: donors M and W (with both α and β TCR chains) in response to a SARS-CoV-2 infection between days 15 and 85 post infection Minervina et al. (2021); 6 twin donors (S1 through Q2, only β chain) between days 0 and 15 following yellow-fever vaccination Pogorelyy et al. (2018a); and yellow-fever first (M) and second vaccination (M and P) Minervina et al. (2020). **(e)** Venn diagram showing the overlap between the number of called responding TCR clones by both NoisET and edgeR after a mild COVID-19 infection Minervina et al. (2021). The $z-$scores and $p-$values of the common clones found by both methods and the MIRA data base. Plots **a-c** are standard NoisET output.

6.2.2 Detecting responding TCR clones

To account for differential expression after antigenic stimulation, NoisET implements the approach of previous work Pogorelyy et al. (2018a); Puelma Touzel et al. (2020) that introduced a selection factor s defined as the log-fold change between a clone's frequency at time t_1 $f(t_1) = f$, and the frequency at time t_2, $f(t_2) = fe^s$. A prior is assumed over the variable s, $P(s|\gamma, \bar{s}) = \gamma \exp(-|s|/\bar{s})/(2\bar{s}) + (1-\gamma)\delta(s)$, with $0 \leq \gamma \leq 1$ the fraction of responding clones and $\bar{s} > 0$, their typical effect size. The likelihood associated to observing a clone with empirical abundances $\hat{n}_1$ at time t_1 and $\hat{n}_2$ at time t_2 integrating the prior knowledge over the log-fold change s is the following:

$$\begin{aligned}&\mathbb{P}(\hat{n}_i(t_1) = \hat{n}_1, \hat{n}_i(t_2) = \hat{n}_2)|\gamma, \bar{s}) = \\ &\int\int df_1 \rho(f_1) ds P(s|\gamma, \bar{s}) P(\hat{n}_1|f_1) P(\hat{n}_2|f_1 e^s).\end{aligned} \tag{6.3}$$

The parameters $(\gamma, \bar{s})$, are learned by maximizing the likelihood of the count pair data taken at two given time points:

$$(\gamma^*, \bar{s}^*) = \underset{(\gamma,\bar{s})}{\text{argmax}} \prod_{i=1}^{N_{obs}} \frac{\mathbb{P}(\hat{n}_i(t_1), \hat{n}_i(t_2)|\gamma, \bar{s}, \Theta(t_1), \Theta(t_2))}{\mathcal{Z}(\gamma, \bar{s})}, \tag{6.4}$$

with $\mathcal{Z}(\gamma, \bar{s})$ a normalization factor accounting for the probability to observe TCR clone counts in both analyzed samples and $\Theta(t_1), \Theta(t_2)$ the noise parameters learned at both time points t_1 and t_2 with NoisET. These two parameters were then used to compute the posterior $\mathbb{P}(s|\hat{n}_i(t_1), \hat{n}_i(t_2))$:

$$\mathbb{P}(s|\hat{n}_1, \hat{n}_2) = \frac{\mathbb{P}(\hat{n}_1, \hat{n}_2|s, \gamma, \bar{s}) P(s|\gamma, \bar{s})}{\mathbb{P}(\hat{n}_1, \hat{n}_2))}. \tag{6.5}$$

The knowledge of the log-fold change posterior (6.5) is used to discriminate expanded or contracted clones from the bulk between t_1 and t_2. In analogy with p-values, we define $p = \mathbb{P}(s \leq 0|\hat{n}_1, \hat{n}_2, \gamma, \bar{s}, \Theta(t_1), \Theta(t_2))$, the probability corresponding to the null hypothesis of no expansion. If $p <$ threshold, the clone is classified as expanded. When looking at contraction, we use the same method reversing times t_1 and t_2 and looking at significant expansions from t_2 to t_1. The value of the threshold can be chosen by the user. In all the results presented in this review, the threshold was set to 0.05, however no threshold was applied when identifying contracting clones. Another threshold on the median of the $\mathbb{P}(s|\hat{n}_1, \hat{n}_2)$ distribution can be applied to select for clones that are greatly expanded.

The output of NoisET detection of responding clones is the list of statistical properties of the true log-fold change variable s called according to the posterior $P(s|\hat{n}_1, t_1, \hat{n}_2, t_2)$ learned from data after learning the noise and differential model

Feature	Description
$s_{1,\text{low}}$	$\int_{-\infty}^{s_{1,\text{low}}} \mathbb{P}(s\|\hat{n}_1,\hat{n}_2)ds = 0.025$
$s_{2,\text{med}}$	$\int_{-\infty}^{s_{2,\text{med}}} \mathbb{P}(s\|\hat{n}_1,\hat{n}_2)ds = 0.5$
$s_{3,\text{high}}$	$\int_{-\infty}^{s_{3,\text{high}}} \mathbb{P}(s\|\hat{n}_1,\hat{n}_2)ds = 0.975$
$s_{\max}$	$\underset{(s)}{\text{argmax}}\mathbb{P}(s\|\hat{n}_1,\hat{n}_2)$
$\bar{s}$	$\int_{-\infty}^{+\infty} s\mathbb{P}(s\|\hat{n}_1,\hat{n}_2)ds$
$1-\mathbb{P}(s>0)$	$\int_{-\infty}^{+\infty} s\mathbb{P}(s\|\hat{n}_1,\hat{n}_2)ds$

Table S1: Mathematical definition of statistical properties of the hidden variable s, the log-fold change of counts of a given clone, computed from the posterior distribution $P(s|\hat{n}_1,t_1,\hat{n}_2,t_2)$, learned from the noise and differential model. The output of NoisET when detecting significantly expanded clones consists of the list of clones that are detected to have respectively increased or decreased in term of abundance associated with these specific s characteristics.

(Eq. 6.2,6.4). These statistics are mathematically defined in Table S1 and are the values of s that defines the first quantile $s_{1,\text{low}}$, the median of the posterior $s_{2,\text{med}}$, the value of s that defines the third quantile $s_{3,\text{high}}$, the mode of the posterior $s_{\max}$, the average of the posterior $\bar{s}$ and and the p-value like value defined as $P(s\leq 0|\hat{n}_1,\hat{n}_2)$.

6.3. Features

NoisET has two main functions: (1) inference of a statistical null model of sequence counts and variability, using replicate RepSeq experiments, as described by the models presented in section II A; (2) detection of responding clones to a stimulus by comparison of two repertoires taken at two timepoints, as described by the models presented in section II A. The second function requires a noise model, which is given as an output of the first function. Both functions require two lists of sequence counts associated to each TCR or BCR present in the repertoires: from replicate experiments for the first function (Fig. 1a left), and from repertoires before and after the stimulus for the second function (Fig. 1a right). In addition, NoisET has features for detecting the time points to be compared, to simulate natural immune repertoire dynamics, and to estimate diversity.

All functions are described in a README and notebooks available on the Github repository (`https://github.com/statbiophys/NoisET`). A tutorial explains the different functions of NoisET.

6.3.1 Detecting the peak moment of the response

When more than two time points are available, and when the timescales of the dynamical response of the TCR repertoire to an acute infection are not known, it is difficult to know which pairs of time points in longitudinal data can be informative about responding clonotypes. A method based on Principal Component Analysis

(PCA) of longitudinal trajectories was first used in Minervina et al. (2020, 2021) to identify the peak of the response (Fig 1b). It uses the first two PCA components of the 1000 most abundant TCR clonotype frequencies normalized by their maximum post-infection values. The clustered trajectories identify different modes of clonal abundance dynamics. NoisET includes a feature for performing this PCA on trajectories as a preliminary step to pick the best timepoints.

6.3.2 Learning the noise model

When learning a noise model from replicates, the user must pick the type of noise model, which describes how the sequence count in the RepSeq sample depends probabilistically on its true frequency in the blood. Choices are: a Poisson distribution, a negative binomial distribution, or a two-step model Puelma Touzel et al. (2020). Once the parameters have been learned (Maximum Likelihood Estimation optimization algorithm), a generation tool can be applied to qualitatively check the agreement between data and model for replicates (Fig 1a). We also successfully learned a null model from gDNA data Rytlewski et al. (2019), which is included in the package example notebook.

6.3.3 Detecting responding clones

To use the second function to detect responding clonotypes, the user provides, in addition to the two datasets to be compared, two sets of experimental noise parameters learned at both times using the first function. When replicates are not available for each time point or donor, a common null model may be used for both timepoints. This should be done with caution, since even if both samples are produced with the same technology for the same donor, the sequencing depth and distribution of clone frequencies may vary between timepoints. Finally the user provides two thresholds: one for the posterior probability above which a clone is labeled as responding, and one for the median log-fold frequency difference above which detection is allowed. The output is a CSV file containing a table of putative responding clones. The result is illustrated in Fig. 1c, which shows contracted clones (purple points) detected from day 15 to day 85 from a mild COVID-19 infection Minervina et al. (2021).

Compared with software introduced in Ref. Puelma Touzel et al. (2020), NoisET allows for conditioning on TCR clones sizes in the analysis, and for using a Poisson or negative binomial distribution for the experimental noise model.

6.3.4 Generating trajectories

Using NoisET, one can also generate *in-silico* RepSeq samples, and their neutral dynamics following the stochastic population dynamics developed in Desponds et al. (2016), and in Bensouda Koraichi et al. (2022). The function takes as input the noise model method (negative binomial or Poisson), the noise model parameters at both

time points, the number of reads at both time points, the duration of the simulations, and the values of τ and θ describing the global stochastic population dynamics. The neutral dynamics for each clone is defined by $\frac{dn}{dt} = \left[-\frac{1}{\tau} + \frac{1}{2\theta} + \frac{1}{\sqrt{\theta}}\eta(t)\right] n(t)$, with $n(t)$ – the true somatic abundance for a clone belonging to an individual repertoire.

6.3.5 Diversity estimator

Learning the noise model is also helpful for computing diversity estimates which are known to be sensitive to sampling noise Mora and Walczak (2019a). NoisET includes a diversity estimator $D_0 = N_{\text{obs}}/(1 - P(\hat{n} = 0, \hat{n}' = 0))$, with N_{obs} the number of clones observed in both replicates used to learn the experimental noise, and $P(\hat{n} = 0, \hat{n}' = 0)$ the learned fraction of non-sampled clones from the repertoire. This value is expected to be close to 1. Evidently, the larger N_{obs}, the deeper the sequencing is and so the diversity estimate is expected to be more trustworthy, assuming comparable quality of data generation.

6.4. Applications of NoisET

The method on which NoisET is based has been applied in two published studies identifying clones involved in yellow fever vaccination Pogorelyy et al. (2018a) and SARS-CoV2 responses Minervina et al. (2021). In both cases, the analysis was performed on longitudinal TCR RepSeq cDNA data sets and from several different time-points, we were able to identify the peak of the response (expansion or contraction) thanks to the trajectory PCA method Minervina et al. (2020) now encoded in NoisET. Fig. 1d reports the number of responding clonotypes detected by NoisET applied to these datasets, as well as to data from a secondary Yellow-Fever vaccination study Minervina et al. (2020).

In the yellow fever vaccination study, TCR repertoires of three pairs of identical twins were sequenced Pogorelyy et al. (2018a). In each donor, 600 to 1700 responding TCR clones were identified. The TCR response was highly personalized even among twins. Analyzing the clonotypes the method called responding, we were able to show that while the responding TCRs were mostly private, they could be well-predicted using a classifier based on sequence similarity. Using the a posteriori distribution, different types of dynamics were found in different TCR subsets: CD4+ cells contract faster than CD8+ cells.

TCR cDNA-based repertoire response identified groups of CD4+ and CD8+ T cell clones that contract after recovery ($\sim$ 15 days after the onset of symptoms) from a SARS-CoV-2 infection Minervina et al. (2021). A secondary response peak of the response was identified $\sim$ 40 days after the onset of symptoms. This secondary peak was also seen in other SARS-CoV2 studies Weiskopf et al. (2020), however it did not correspond to known tetramer probes. Analyzing repertoire data for the same

individuals taken a year and two years before the SARS-CoV2 infection, we showed that T-cell clones detected as reacting to SARS-Cov-2 were present one year before the SARS-Cov-2 infection. A network analysis revealed that these pre-existing cells that could confer immunity were specific to a SARS-CoV2 epitope with a one amino acid mutation compared to a common cold coronavirus. This observation raised the question of the correlation between the presence of cross-reactive T cells before infection and mildness of the disease. The detected reactive T-cell clones were also found in memory subpopulations at least three months after the infection.

As mentioned in section III D the noise learning feature of NoisET has also been used to learn the natural dynamics of TCR repertoires based on gDNA and cDNA data in the absence of direct antigenic stimulation Bensouda Koraichi et al. (2022). This study considered the TCRβ repertoires of 9 people and showed that the dynamics of all people, regardless of age is constrained by the power law exponent of the frequency distribution. The exponent itself is given by the ratio of the deterministic turnover timescale and the stochastic noise timescale. The reproducibility of this ratio is a very strong constraint, not directly encoded by the model but learned from the statistics of the data, that implies strong amplitudes of environmental antigenic fluctuations compared to the mean fitness of lymphocyte clones. This parameter regime translates into a very susceptible dynamical system since the mean of the clone size distribution diverges. This property allows the repertoire to maintain a large number of cells, even if the source disappears or becomes very small. While the ratio is constrained, the repertoire turnover timescale shows a strong dependence on the age of the individual, with clear signatures of ageing in the physical sense: turnover timescales grow linearly with the biological age of the individual from ~ 10 years for 20-year olds to $\sim$ 40 years for 60-year olds. This timescale gives us an estimate of how likely we are to find clones in the repertoire after a certain number of years, depending on the person's age.

6.5. Comparison with existing software

The need to characterise experimental noise has been well recognized in the sequencing community. EdgeR is a package used to analyse a variety of data produced with HTS (High Throughput Sequencing) that includes read counts Yunshun, Chen and Aaron, Lun and Davis, McCarthy and Xiaobei, Zhou and Mark, Robinson and Gordon, Smyth (2017). This software has been mostly used for differential gene expression analysis, differential splicing and bisulfite sequencing. Applied to lymphocyte repertoires, the EdgeR package enables using statistical tests to identify TCR clones expanded after an acute infection.

We compare EdgeR and NoisET detected clones assumed to respond to SARS-Cov2 antigen, based on TCR data from Adaptive Biotechnologies (`https://clients.`

adaptivebiotech.com/pub/covid-2020). We can validate the responding clones using the MIRA dataset from the same group for which the reactivity to SARS-Cov2 antigen was validated experimentally https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418734/. To count the overlap between the responding clonotypes called by each software and the TCR MIRA database, we used the AtrieGC software (https://github.com/mbensouda/NoisET_tutorial/), which enables to rapidly compare two lists of amino-acids. In order to ascribe statistical significance to our results, we compare the numbers of overlapping TCRs called by EdgeR and NoisET to overlapping TCRs between lists of 1000 randomly sampled clones from the experimental samples and the MIRA list. Given the mean and standard deviation of overlapping clones, we quantify the performance of the two softwares using a z-score. The conclusion is drawn in a Venn diagram in Fig.S1e. For this specific task of recognizing SARS-Cov2 TCR clones NoisET (p-value of 5.10^{-13}) performs similarly to edgeR (p-value of 2.410^{-14}) with the benefits of better understanding of the data, better knowledge of the log-fold change statistics and the possibility to generate synthetic data. We note that the MIRA database is non exhaustive so both NoisET and edgeR may have called truly responding SARS-Cov2 TCR clones that are not included in the MIRA database.

6.6. Discussion

High-throughput sequencing of immune repertoires is poised to revolutionize systems immunology as well as precision medicine. In particular, there is a growing interest in identifying T-cell receptors that respond to acute infections and vaccine challenges, based on experiments that probe repertoires before and after an antigenic challenge. Due to experimental and biological noise, identifying the response simply based on differences in counts before and after the challenge is not reliable. The commonly used solution is to prune these estimates using statistical tests, which are not tailored to account for these specific sources of noise.

In our previous work, we provided a computational method that accounts for the different biological and experimental sources of noise in the clone count measurements in a Bayesian way, allowing for a more reliable detection of expanded or contracted clones. However, while the proof-of-principle algorithm explored the applicability of the method, it did not provide a user-friendly tool, which limits its wide use by the community of immunologists and clinicians. Here, we described a new computational tool, NoisET (https://github.com/statbiophys/NoisET), a python package with a command-line interface that implements the method for characterizing the noise and identifies statistically significant responding clones. The tool is applicable to datasets of diverse origin describing the clonal repertoire response to acute infections and non-stimulated long-term dynamics.

NoisET is designed as an easy-to-use package to learn the noisy statistics of sequence counts and to detect responding clones to a stimulus as reliably as possible. It captures the experimental and biological noise for both RNAseq and gDNAseq replicate technologies. Although the package has been tested on diverse datasets, choosing and using the adequate statistical null model should be done with caution.

Among the different types of noise models offered, the negative binomial noise model is recommended to start the analysis as its running time is shorter than the two step model, while retaining the ability to account for arbitrary noise amplitudes. So far, NoisET has been used to study the short time scale dynamics for acute infections, but could also be used to compare bulk repertoires with selected repertoires derived from functional or cultured assays Balachandran et al. (2017). For longer time scales, the dynamics of lymphocyte populations should be modeled to best describe slow global repertoire changes that cannot be attributed to a single stimulus Desponds et al. (2016); Bensouda Koraichi et al. (2022).

The Bayesian approach encoded in NoisET results in a more reliable way to account for uncertainty than statistical estimates that are also less interpretable. The detection of responding clones based on the fold change of empirical abundances was not optimal without a robust interpretation of the details of the noise model. Errors in noise identification also propagate to erroneous calling of clonotypes.

From a more general perspective, NoisET and the methods behind it combine many years of the study of gene expression noise Elowitz et al. (2002); Ozbudak et al. (2002); Cai et al. (2006); Taniguchi et al. (2010); Hornos et al. (2005). NoisET strongly exploits the intermittency of mRNA production and the heterogeneity of mRNA counts in individual cells.

As we briefly discussed, NoisET has been applied to identify SARS-CoV-2-specific T-cell receptors and in the future can be used to study and understand the heterogeneity of SARS-Cov-2 vaccine response. It has potential application uncovering responding T-cell receptors to acute infections and vaccine response.

While the method is generally applicable to T cells and B cells Altan-Bonnet et al. (2020); Chakraborty and Košmrlj (2010), due to the somatic hypermutations occurring in B cells upon proliferation, care must be taken when preparing B cell data input and interpreting the model. One possibility is to collapse the sequences into lineages and consider the dynamics of a lineage in the periphery. However, while this is a reasonable first approximation, more work is needed to correctly account for the complexity of B cell repertoires. For this reason we discuss existing applications to T cells. Nevertheless, the conceptual ideas behind noise calibration as implemented in the NoisET software apply.

More broadly, the noise inference using the first module of NoisET has also been used to learn the natural dynamics of T cell repertoires in the absence of specific antigenic stimulation Bensouda Koraichi et al. (2022). For all individuals studied

we found a universal contraint on the dynamics, which translated into a susceptible dynamical system that can easily maintain a large number of diverse cells. If the same type of dynamics holds for coarse-grained B cell repertoires, which remains to be seen, it would point to universal laws that constrain clone size distributions and govern repertoire dynamics.

CHAPTER

7

LONGITUDINAL HIGH-THROUGHPUT TCR REPERTOIRE PROFILING REVEALS THE DYNAMICS OF T CELL MEMORY FORMATION AFTER MILD COVID-19 INFECTION

In April 2020, at the beginning of the pandemic caused by the SARS-CoV-2 coronavirus, I contributed to an extensive collaboration study to improve our global understanding of TCR responses to the new SARS-Cov-2 virus. I improved and optimized the methods behind the previously introduced *NoisET* in chapter 5. My contribution to this analysis was the detection of expanding and contracting clones from longitudinal RepSeq data of two people having contracted SARS-CoV-2 with mild symptoms. Chapter 6 is a comprehensive analysis of the TCR repertoire response to an acute stimulus and gives a good biological background on TCR repertoire immunology analysis, published in Minervina et al. (2021). I am a co-author of this article and all the other analyses were done by other scientists.

Chapter 7. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19

7.1. Introduction

COVID-19 is a global pandemic caused by the novel SARS-CoV-2 betacoronavirus Vabret et al. (2020). T cells are crucial for clearing respiratory viral infections and providing long-term immune memory Schmidt and Varga (2018); Swain et al. (2012). Two major subsets of T cells participate in the immune response to viral infection in different ways: activated CD8+ T cells directly kill infected cells, while subpopulations of CD4+ T cells produce signaling molecules that regulate myeloid cell behaviour, drive and support CD8 response and the formation of long-term CD8 memory, and participate in the selection and affinity maturation of antigen specific B-cells, ultimately leading to the generation of neutralizing antibodies. In SARS-1 survivors, antigen-specific memory T cells were detected up to 11 years after the initial infection, when viral-specific antibodies were undetectable Ng et al. (2016); Oh et al. (2011). The T cell response was shown to be critical for protection in SARS-1-infected mice Zhao et al. (2010). Patients with X-linked agammaglobulinemia, a genetic disorder associated with lack of B cells, have been reported to recover from symptomatic COVID-19 Quinti et al. (2020); Soresina et al. (2020), suggesting that in some cases T cells are sufficient for viral clearance. Theravajan et al. showed that activated CD8+HLA-DR+CD38+ T cells in a mild case of COVID-19 significantly expand following symptom onset, reaching their peak frequency of 12% of CD8+ T cells on day 9 after symptom onset, and contract thereafter Thevarajan et al. (2020). Given the average time of 5 days from infection to the onset of symptoms Bi et al. (2020), the dynamics and magnitude of T cell response to SARS-CoV-2 is similar to that observed after immunization with live vaccines Miller et al. (2008). SARS-CoV-2-specific T cells were detected in COVID-19 survivors by activation following stimulation with SARS-CoV-2 proteins Ni et al. (2020), or by viral protein-derived peptide pools Weiskopf et al. (2020); Braun et al. (2020); Snyder et al. (2020); Le Bert et al. (2020); Meckiff et al. (2020); Bacher et al. (2020); Peng et al. (2020). Some of the T cells activated by peptide stimulation were shown to have a memory phenotype Weiskopf et al. (2020); Le Bert et al. (2020); Mateus et al. (2020), and some potentially cross-reactive CD4+ T cells were found in healthy donors Braun et al. (2020); Grifoni et al. (2020); Sekine et al. (2020); Bacher et al. (2020).

T cells recognise short pathogen-derived peptides presented on the cell surface of the Major Histocompatibility Complex (MHC) using hypervariable T cell receptors (TCR). TCR repertoire sequencing allows for the quantitative tracking of T cell clones in time, as they go through the expansion and contraction phases of the response. It was previously shown that quantitative longitudinal TCR sequencing is able to identify antigen-specific expanding and contracting T cells in response to

yellow fever vaccination with high sensitivity and specificity Minervina et al. (2020); Pogorelyy et al. (2018c); DeWitt et al. (2015). Not only clonal expansion but also significant contraction from the peak of the response are distinctive traits of T cell clones specific to the virus Pogorelyy et al. (2018c).

In this study we use longitudinal TCRalpha and TCRbeta repertoire sequencing to quantitatively track T cell clones that significantly expand and contract (after recovery from a mild COVID-19 infection), and determine their phenotype. We reveal the dynamics and the phenotype of the memory cells formed after infection, identify pre-existing T cell memory clones participating in the response, and describe public TCR sequence motifs of SARS-CoV-2-reactive clones, suggesting a response to immunodominant epitopes.

7.2. Results

7.2.1 Longitudinal tracking of TCR repertoires of COVID-19 patients

In the middle of March (day 0) donor W female and donor M (male, both healthy young adults), returned to their home country from the one of the centers of the COVID-19 outbreak in Europe at the time. Upon arrival, according to local regulations, they were put into strict self-quarantine for 14 days. On day 3 of self-isolation both developed low grade fever, fatigue and myalgia, which lasted 4 days and was followed by a temporary loss of smell for donor M. On days 15, 30, 37, 45 and 85 we collected peripheral blood samples from both donors (Fig. 1a). The presence of IgG and IgM SARS-CoV-2 specific antibodies in the plasma was measured at all timepoints using SARS-CoV-2 S-RBD domain specific ELISA (Fig. S1). From each blood sample we isolated PBMCs (peripheral blood mononuclear cells, in two biological replicates), CD4+, and CD8+ T cells. Additionally, on days 30, 45 and 85 we isolated four T cell memory subpopulations (Fig. S2): Effector Memory (EM: CCR7-CD45RA-), Effector Memory with CD45RA re-expression (EMRA: CCR7-CD45RA+), Central Memory (CM: CCR7+CD45RA-), and Stem Cell-like Memory (SCM: CCR7+CD45RA+CD95+). From all samples we isolated RNA and performed TCRalpha and TCRbeta repertoire sequencing as previously described Pogorelyy et al. (2017). For both donors, TCRalpha and TCRbeta repertoires were obtained for other projects one and two years prior to infection. Additionally, TCR repertoires of multiple samples for donor M – including sorted memory subpopulations – are available from a published longitudinal TCR sequencing study after yellow fever vaccination (donor M1 samples in Minervina et al. (2020)).

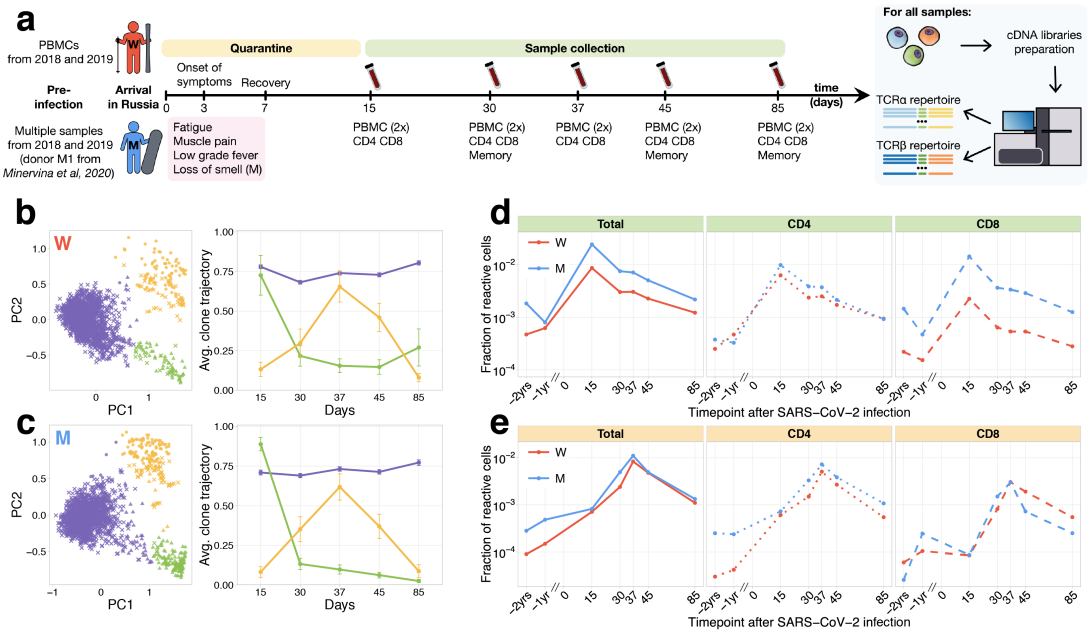

Figure S1: **Longitudinal tracking of T cell clones after mild COVID-19. a, Study design.** Peripheral blood of two donors was sampled longitudinally on days 15, 30, 37, 45, 85 after arrival in Russia. At each timepoint, we evaluated SARS-CoV-2-specific antibodies using ELISA (Fig. S1) and isolated PBMCs in two biological replicates. Additionally, CD4+ and CD8+ cells were isolated from a separate portion of blood, and EM, CM, EMRA, SCM memory subpopulations were FACS sorted on days 30, 45 and 85. For each sample we sequenced TCRalpha and TCRbeta repertoires. For both donors pre-infection PBMC repertoires were sampled in 2018 and 2019 for other projects. **b,c, PCA of clonal temporal trajectories identifies three groups of clones with distinctive behaviours.** Left: first two principal components of the 1000 most abundant TCRbeta clonotype frequencies normalized by maximum value for each clonotype in PBMC at post-infection timepoints. Color indicates hierarchical clustering results of principal components; symbol indicates if clonotype was called as significantly contracted from day 15 to day 85 (triangles), or expanded from day 15 to day 37 (circles) by both edgeR and NoisET (Fig. S5 shows overlap between clonal trajectory clusters and edgeR/NoisET hits). Right: each curve shows the average ± 2.96 SE of normalized clonal frequencies from each cluster. **Contracting (d) and expanding (e) clones include both CD4+ and CD8+ T cells, and are less abundant in pre-infection repertoires.** T cell clones significantly contracted from day 15 to day 85 **(d)** and significantly expanded from day 15 to day 37 **(e)** were identified in both donors. The fraction of contracting **(d)** and expanding **(e)** TCRbeta clonotypes in the total repertoire (calculated as the sum of frequencies of these clonotypes in the second PBMC replicate at a given timepoint) and corresponding to the fraction of responding cells of all T cells) is plotted in log-scale for all reactive clones (left), reactive clones with the CD4 (middle) and CD8 (right) phenotypes. Similar dynamics were observed in TCRalpha repertoires (Fig. S3), and for significantly expanded/contracted clones identified with the NoisET Bayesian differential expansion statistical model (alone) (Fig. S4).

7.2.2 Two waves of T-cell clone response

From previously described activated T cell dynamics for SARS-CoV-2 Thevarajan et al. (2020), and immunization with live vaccines Miller et al. (2008), the peak of the T cell expansion is expected around day 15 post-infection, and responding T cells significantly contract (afterwards). However, Weiskopf et al. Weiskopf et al. (2020) reports an increase of SARS-CoV-2-reactive T cells at later timepoints, peaking in some donors after 30 days following symptom onset. To identify groups of T cell clones with similar dynamics in an unbiased way, we used Principal Component Analysis (PCA) in the space of T cell clonal trajectories (Fig. 1b and c). This exploratory data analysis method allows us to visualize major trends in the dynamics of abundant TCR clonotypes (occuring within top 1000 on any post-infection timepoints) between multiple timepoints.

In both donors, and in both TCRalpha and TCRbeta repertoires, we identified three clusters of clones with distinct dynamics. The first cluster (Fig. 1bc, purple) corresponded to abundant TCR clonotypes which had constant concentrations across timepoints, the second cluster (Fig. 1bc, green) showed contraction dynamics from day 15 to day 85, and the third cluster (Fig. 1bc, yellow), showed an unexpected clonal expansion from day 15 with a peak on day 37 followed by contraction. The clustering and dynamics are similar in both donors and are reproduced in TCRbeta (Fig. 1bc) and TCRalpha (Fig. S3ab) repertoires. We next used edgeR, a software for differential gene expression analysis Robinson et al. (2010) and NoisET, a Bayesian differential expansion model Puelma Touzel et al. (2020), to specifically detect changes in clonotype concentration between pairs of timepoints in a statistically reliable way and without limiting the analysis to the most abundant clonotypes. Both NoisET and edgeR use biological replicate samples collected at each timepoint to train a noise model for sequence counts. Results for the two models were similar (Fig. S4) and we conservatively defined as expanded or contracted the clonotypes that were called by both models simultaneously. We identified 291 TCRalpha and 295 TCRbeta clonotypes in donor W, and 607 TCRalpha and 616 TCRbeta in donor M significantly contracted from day 15 to day 85 (largely overlapping with cluster 2 of clonal trajectories, Fig. S5). 176 TCRalpha and 278 TCRbeta for donor W, and 293 TCRalpha and 427 TCRbeta clonotypes for donor M were significantly expanded from day 15 to 37 (corresponding to cluster 3 of clonal trajectories).

Note that, to identify putatively SARS-CoV-2 reactive clones, we only used post-infection timepoints, so that our analysis can be reproduced in other patients and studies where pre-infection timepoints are unavailable. However, tracking the identified responding clones back to pre-infection timepoints reveals strong clonal

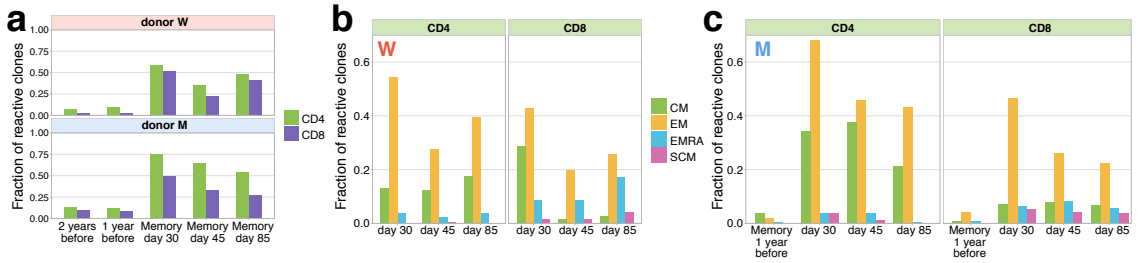

Figure S2: **Memory phenotypes of responding clonotypes contracting after day 15. a, A large fraction of contracting clonotypes is identified in T cell memory subsets after infection.** Bars show the fraction of contracting CD4+ and CD8+ TCRbeta clonotypes present in 2-year; 1-year pre-infection PBMC; in at least one of memory subpopulation sampled on day 30, day 37 and day 85 post infection. **b,c Responding clones are found in different memory subsets.** Fraction of CD4+ (left panels) and CD8+ (right panels) contracting clones of donor W **(b)** and M **(c)** that were identified in each memory subpopulation repertoire at each timepoint. For both donors, CD4+ clonotypes were found predominantly in Central Memory (CM) and Effector Memory (EM), while CD8+ T cells were enriched in EMRA compartment.**c,** For donor M, CD4+ contracting clonotypes are also identified in memory subsets 1 year before the infection, with a bias towards the CM subpopulation and a group of CD8+ clones is found in the pre-infection EM subpopulation.

expansions from pre- to post-infection (Fig. 1de, Fig. S3cd). For brevity, we further refer to clonotypes significantly contracted from day 15 to 85 as *contracting* clones and clonotypes significantly expanding from day 15 to 37 as *expanding* clones. Contracting clones corresponded to 2.5% and 0.9% of T cells on day 15 post-infection, expanding clones reached 1.1% and 0.8% on day 37 for donors M and W respectively (Fig. 1de, left). This magnitude of the T cell response is of the same order of magnitude as previously observed after live yellow fever vaccine immunization of donor M (6.7% T cells on day 15 post-vaccination). For each contracting and expanding clone we determined their CD4/CD8 phenotype using separately sequenced repertoires of CD4+ and CD8+ subpopulations (see Methods). Both CD4+ and CD8+ subsets participated actively in the response (Fig. 1de). Interestingly, clonotypes expanding after day 15 were significantly biased towards the CD4+ phenotype, while contracting clones had balanced CD4/CD8 phenotype fractions in both donors (Fisher exact test, $p < 0.01$ for both donors).

7.2.3 Memory formation and pre-existing memory

On days 30, 45 and 85 we identified both contracting (Fig. 2a-c) and expanding (Fig. S6a-c) T cell clones in the memory subpopulations of peripheral blood. Both CD4+ and CD8+ responding clones were found in the CM and EM subsets, however

CD4+ were more biased towards CM (with exception of donor W day 30 timepoint, where a considerable fraction of CD8+ clones were found in CM), and CD8+ clones more represented in the EMRA subset. A small number of both CD4+ and CD8+ responding clonotypes were also identified in the SCM subpopulation, which was previously shown to be a long-lived T cell memory subset Fuertes Marraco et al. (2015). Note that we sequenced more cells from PBMC than from the memory subpopulations (Table S1), so that some low-abundant responding T cell clones are not sampled in the memory subpopulations. Intriguingly, a number of responding CD4+ clones, and fewer CD8+ clones, were also represented in the repertoires of both donors 1 and 2 years before the infection. Pre-existing clones were expanded after infection, and contracted afterwards for both donors (Fig. S7). For donor M, for whom we had previously sequenced memory subpopulations before the infection Minervina et al. (2020), we were able to identify pre-existing SARS-CoV-2-reactive CD4+ clones in the CM subpopulation 1 year before the infection and a group of CD8+ clones in the pre-infection EM subpopulation. Interestingly, on day 30 after infection the majority of pre-infection CM clones were detected in the EM subpopulation, suggesting recent T cell activation and a switch of the phenotype from memory to effector. These clones might represent memory T cells cross-reactive for other infections, e.g. other human coronaviruses.

A search for TCRbeta amino acid sequences of responding clones in VDJdb Bagaev et al. (2020b) — a database of TCRs with known specificities — resulted in essentially no overlap with TCRs not specific for SARS-CoV-2 epitopes: only two clonotypes matched. One match corresponded to the CMV (cytomegalovirus) epitope presented by the HLA-A*03 MHC allele, which is absent in both donors (Table S2), and a second match was for Influenza A virus epitope presented by HLA-A*02 allele. The absence of matches suggests that contracting and expanding clones are unlikely to be specific for immunodominant epitopes of common pathogens covered in VDJdb. We next asked if we could map specificites of our responding clones to SARS-CoV-2 epitopes.

7.2.4 Validation by MHC tetramer-staining assay

On day 25 post-infection donor M participated in study by Shomuradova et al. Shomuradova et al. (2020) (as donor p1434), where his CD8+ T cells were stained with HLA-A*02:01-YLQPRTFLL MHC-I tetramer. TCRalpha and TCRbeta of FACS-sorted tetramer-positive cells were sequenced and deposited to VDJdb (see Shomuradova et al. (2020) for the experimental details). We matched these tetramer-specific TCR sequences to our longitudinal dataset (Fig. 3a for TCRbeta and Fig. S8 for TCRalpha). We found that their frequencies were very low on pre-infection

timepoints and monotonically decreased from their peak on day 15 ($7.1 \cdot 10^{-4}$ fraction of bulk TCRbeta repertoire) to day 85 ($1.3 \cdot 10^{-5}$ fraction), in close analogy to our contracting clone set. Among the tetramer positive clones that were abundant on day 15 (with bulk frequency $> 10^{-5}$), 17 out of 18 or TCRbetas and 12 out of 15 TCRalphas were independently identified as contracting by our method.

7.2.5 TCR sequence motifs of responding clones

It was previously shown that TCRs recognising the same antigens frequently have highly similar TCR sequences Dash et al. (2017); Glanville et al. (2017). To identify motifs in TCR amino acid sequences, we plotted similarity networks for significantly contracted (Fig. 3bc, Fig. 4ab) and expanded (Fig. S9b-e) clonotypes. The number of edges in all similarity networks except CD8+ expanding clones was significantly larger than would expected by randomly sampling the same number of clonotypes from the corresponding repertoire (Fig. 3d and Fig. S9a). In both donors we found clusters of highly similar clones in both CD4+ and CD8+ subsets for expanding and contracting clonotypes. Clusters were largely donor-specific, as expected, since our donors have dissimilar HLA alleles (SI Table 1) and thus each is likely to present a non-overlapping set of T cell antigens. The largest cluster, described by the motif TRAV35-CAGXNYGGSQGNLIF-TRAJ42, was identified in donor M's CD4+ contracting alpha chains. Clones from this cluster constituted 16.3% of all of donor M's CD4+ responding cells on day 15, suggesting a response to an immunodominant CD4+ epitope in the SARS-CoV-2 proteome. The high similarity of the TCR sequences of responding clones in this cluster allowed us to independently identify motifs from donor M's CD4 alpha contracting clones using the ALICE algorithm Pogorelyy et al. (2018b) (Fig. S10). While the time dependent methods (Fig. 1) identify abundant clones, the ALICE approach is complementary to both edgeR and NoisET as it identifies clusters of T cells with similar sequences independently of their individual abundances.

7.2.6 Mapping TCR motifs to SARS-CoV-2 epitopes

In CD8+ T cells, 3 clusters of highly similar TCRbeta clonotypes in donor M and one cluster of TCRalpha clonotypes correspond to YLQPRTFLL-tetramer-specific TCR sequences described above. To map additional specificities for CD8+ TCRbetas, we used a large set of SARS-CoV-2-peptide specific TCRbeta sequences from Snyder et al. (2020) obtained using Multiplex Identification of Antigen-specific T cell Receptors Assay (MIRA) with combinatorial peptide pools Klinger et al. (2015). For each responding CD8+ TCRbeta we searched for the identical or highly similar (same VJ combination, up to one mismatch in CDR3aa) TCRbeta sequences specific for given SARS-CoV-2 peptides. A TCRbeta sequence from our set was considered

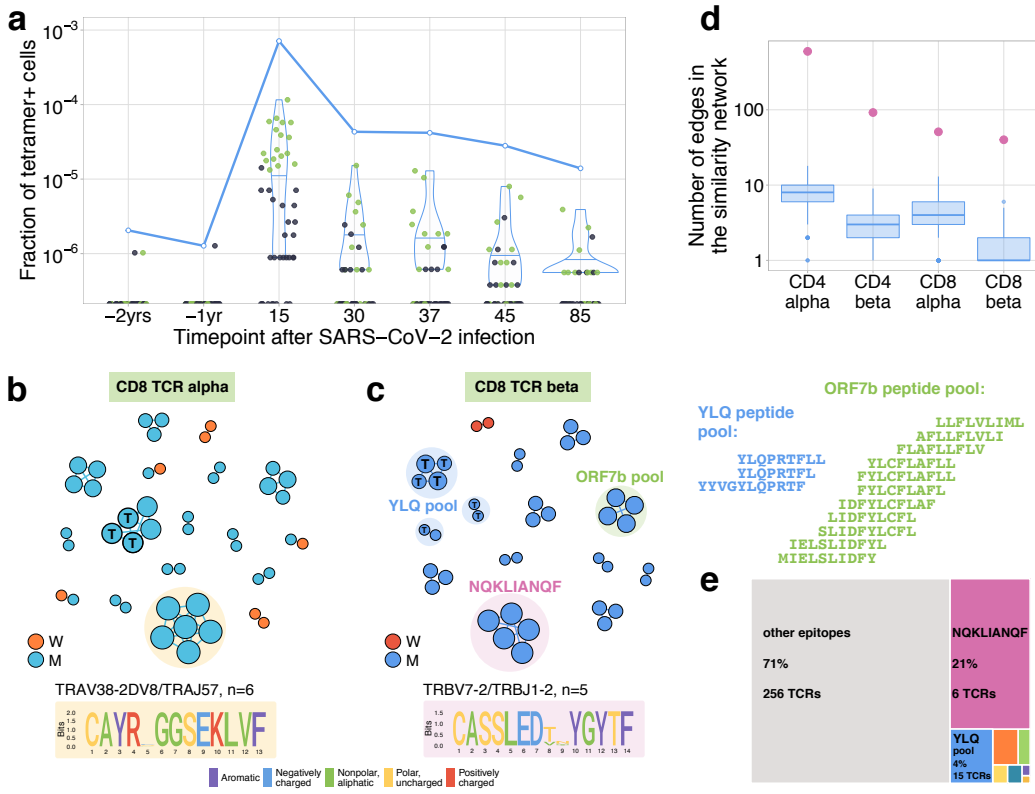

Figure S3: **a, SARS-CoV-2-specific TCRs are independently identified by clonal contraction.** Each dot corresponds to the frequency of HLA-A*02:01-YLQPRTFLL-tetramer specific TCRbeta clonotype in bulk repertoire of donor M (donor p1434 from Shomuradova et al. (2020)) at each timepoint. Green dots correspond to clonotypes independently identified as contracting in our longitudinal dataset. Blue line shows the cumulative frequency of tetramer specific TCRbeta clonotypes. **b, c Analysis of TCR amino acid sequences of contracting CD8+ clones reveals distinctive motifs.** For each set of CD8alpha, and CD8beta contracted clonotypes, we constructed a separate similarity network. Each vertex in the similarity network corresponds to a contracting clonotype. An edge indicates 2 or less amino acid mismatches in the CDR3 region, and identical V and J segments. Networks are plotted separately for CD8alpha **(b)** and CD8beta **(c)** contracting clonotypes. Clonotypes without neighbours are not shown. Sequence logos corresponding to the largest clusters are shown under the corresponding network plots. 'T' on vertices indicate TCRbeta clonotypes confirmed by HLA-A*02:01-YLQPRTFLL tetramer staining. Shaded colored circles (**(c)**) indicate clonotypes with large number of matches to CD8+ TCRs recognising SARS-CoV-2 peptides pools from ref. Snyder et al. (2020) (MIRA peptide dataset). Lists of peptides in YLQ and ORF7b peptide pools are shown on the right. **d, Sequence convergence among contracting clonotypes.** The number of edges in each group is shown by pink dots and is compared to the distribution of that number in 1000 random samples of the same size from the relevant repertoires at day 15 (blue boxplots). **e,** Fraction of TCRbeta clonotypes with matches in the MIRA dataset (coloured rectangles) out of all responding CD8+ TCRbeta clonotypes in donor M on day 15.

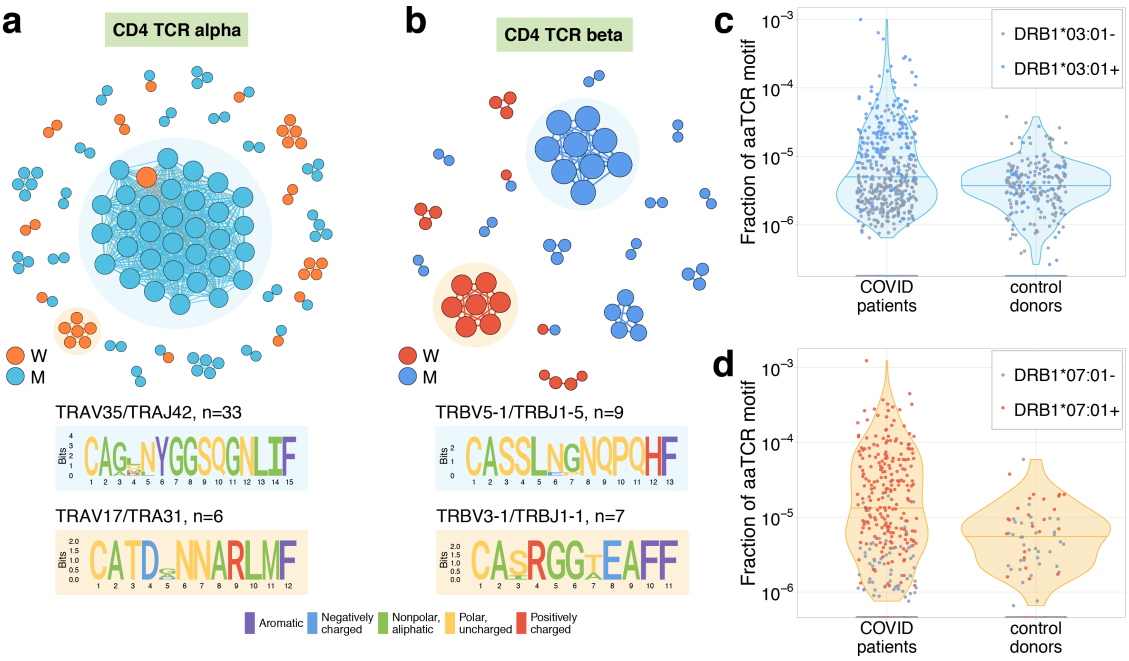

Figure S4: **a, Analysis of TCR amino acid sequences of CD4+ contracting clones reveal distinctive motifs.** Each vertex in the similarity network corresponds to a contracting clonotype. An edge indicates 2 or less amino acid mismatches in the CDR3 region (and identical V and J segments). Networks are plotted separately for CD4alpha **(a)**, CD4beta **(b)**, contracting clonotypes. Clonotypes without neighbours are not shown. Sequence logos corresponding to the largest clusters are shown under the corresponding network plots. **c, d,** Clonotypes forming the two largest motifs are significantly more clonally expanded ($p<0.001$, one sided t-test) in a cohort of COVID-19 patients Snyder et al. (2020) than in a cohort of control donors Emerson et al. (2017). Each dot corresponds to the total frequency of clonotypes from motifs shaded on **(b)** in the TCRbeta repertoire of a given donor. Colored dots show donors predicted to share HLA-DRB1*07:01 allele with donor W **(c)**, or HLA-DRB1*03:01-DQB1*02:01 haplotype with donor M **(d)**.

mapped to a given peptide if it had at least two highly similar TCRbeta sequences specific for this peptide in the MIRA experiment. This procedure yielded unambiguous matches for 32 CD8+ TCRbetas — just one clonotype was paired to two peptide pools (Table S3). The vast majority of matches to MIRA corresponded to groups of contracting clones. As expected, we found that all clusters corresponding to HLA-A*02:01-YLQPRTFLL MHC-I tetramer-specific TCRs were matched to the peptide pool YLQPRTFL,YLQPRTFLL,YYVGYLQPRTF in the MIRA dataset. Another large group of matches corresponded to the HLA-B*15:01-restricted Stamatakis et al. (2020) NQKLIANQF epitope. Interestingly, clonotypes corresponging to this cluster together made up 21% of the CD8+ immune response on day 15, suggesting immunodominance of this epitope. Two TCRbeta clonotypes mapped to this epitope were identified in Effector Memory subset one year before the infection, suggesting potential cross-reactive response. We speculate, that this response might be initially triggered by NQKLIANAF, a homologous HLA-B*15:01 epitope from HKU1 or OC43, common human betacoronaviruses. To predict potential pairings between TCRalpha and TCRbeta motifs, we used a method of alpha/beta clonal trajectory matching described in Minervina et al. (2020) (see Methods for details). We found consistent pairing between one of the motifs in TCRalpha to the largest motif in TCRbeta T cells, which is associated to HLA-B*15:01-NQKLIANQF.

7.2.7 Validation of CD4+ COVID-19 HLA-restricted specificity by cohort association analysis

At the time of writing, no data on TCR sequences specific to MHC-II class epitopes exist to map specificities of CD4+ T-cells in a similar way as we did with MIRA-specific TCRs. However, a recently published database of 1414 bulk TCRbeta repertoires from COVID-19 patients allowed us to confirm the SARS-CoV-2 specificity of contracting clones indirectly. Public TCRbeta sequences that can recognize SARS-CoV-2 epitopes are expected to be clonally expanded and thus sampled more frequently in the repertoires of COVID-19 patients than in control donors. In Fig. 4cd we show that the total frequency of TCRbeta sequences forming the largest cluster in donor M (Fig. 4c) and donor W (Fig. 4d) is significantly larger in the COVID-19 cohort than in the healthy donor cohort from ref. Emerson et al. (2017), suggesting antigen-dependent clonal expansion. We hypothesized that the difference between control and COVID-19 donors in motif abundance should be even larger if we restrict the analysis to donors sharing the HLA allele that presents the epitope. Unfortunately, HLA-typing information is not yet available for the COVID-19 cohort. However, using sets of HLA-associated TCRbeta sequences from ref. DeWitt et al. (2018), we could build a simple classifier to predict the HLA alleles of donors from both the control and COVID-19 cohorts exploiting the presence of TCRbeta

sequences associated with certain HLA alleles (see Methods for details). We found that the CD4+ TCRbeta motif from donor W occurs preferentially in donors predicted to have DRB1*07:01 allele, while the motif from donor M appears to be associated with HLA-DRB1*03:01-DQB1*02:01 haplotype. The frequency of sequences corresponding to these motifs can then be used to identify SARS-CoV-2 infected donors with matching HLA alleles (Fig. S11).

7.3. Discussion

Using longitudinal repertoire sequencing, we identified a group of CD4+ and CD8+ T cell clones that contract after recovery from a SARS-CoV-2 infection. Our response timelines agree with T cell dynamics reported by Theravajan et al. Thevarajan et al. (2020) for mild COVID-19, as well as with dynamics of T cell response to live vaccines Miller et al. (2008). We further mapped the specificities of contracting CD8+ T cells using sequences of SARS-CoV-2 specific T cells identified with tetramer staining in the same donor, and as well as the large set of SARS-CoV-2 peptide stimulated TCRbeta sequences from ref. Snyder et al. (2020). For large CD4+ TCRbeta motifs we show strong association with COVID-19 by analysing the occurence patterns and frequencies of these sequences in a large cohort of COVID-19 patients.

Surprisingly, in both donors we also identified a group of predominantly CD4+ clonotypes which expanded from day 15 to day 37 after the infection. One possible explanation for this second wave of expansion is the priming of CD4+ T cells by antigen-specific B-cells, but there might be other mechanisms such as the migration of SARS-CoV-2 specific T cells from lymphoid organs or bystander activation of non-SARS-CoV-2 specific T cells. It is also possible that later expanding T cells are triggered by another infection, simultaneously and asymptomatically occurring in both donors around day 30. In contrast with the first wave of response identified by contracting clones, for now we do not have confirmation that this second wave of expansion corresponds to SARS-CoV-2 specific T cells. Accumulation of TCR sequences for CD4+ SARS-CoV-2 epitope specific T cells may further address this question.

We showed that a large fraction of putatively SARS-CoV-2 reactive T cell clones are later found in memory subpopulations and remain there at least 3 months after infection. Importantly, some of responding clones are found in long-lived stem cell-like (SCM) memory subset, as also reported for SARS-CoV-2 convalescent patients in ref. Sekine et al. (2020). A subset of CD4+ clones were identified in pre-infection central memory subsets, and a subset of CD8+ T cells were found in effector memory. Among these are CD8+ clones recognising NQKLIANQF, an

immunodominant HLA-B*15:01 restricted SARS-CoV-2 epitope, for which homologous epitope differing by 1 aa mismatch exists in common human betacoronaviruses.

The presence of SARS-CoV-2 cross-reactive CD4+ T cells in healthy individuals was recently demonstrated Braun et al. (2020); Grifoni et al. (2020); Le Bert et al. (2020); Meckiff et al. (2020); Bacher et al. (2020); Peng et al. (2020). Our data further suggests that cross-reactive CD4+ and CD8+ T cells can participate in the response *in vivo*. It is interesting to ask if the presence of cross-reactive T cells before infection is linked to the mildness of the disease (with predicted HLA-B*15:01 cross-reactive epitope described above as a good starting point). Larger studies with cohorts of severe and mild cases with pre-infection timepoints are needed to address this question.

7.4. Methods

7.4.1 Donors and blood samples

Peripheral blood samples from two young healthy adult volunteers, donor W (female) and donor M (male) were collected with written informed consent in a certified diagnostics laboratory. Both donors gave written informed consent to participate in the study under the declaration of Helsinki. HLA alleles of both donors (Table S2) were determined by an in-house cDNA high-throughput sequencing method.

7.4.2 SARS-CoV-2 S-RBD domain specific ELISA

An ELISA assay kit developed by the National Research Centre for Hematology was used for detection of anti-S-RBD IgG according to the manufacturer's protocol. The relative IgG level (OD/CO) was calculated by dividing the OD (optical density) values by the mean OD value of the cut-off positive control serum supplied with the Kit (CO). OD values of d37, d45 and d85 samples for donor M exceeded the limit of linearity for the Kit. In order to properly compare the relative IgG levels between d30, d37, d45 and d85, these samples were diluted 1:400 instead of 1:100, the ratios d37:d30 and d45:d30 and d85:d30 were calculated and used to calculate the relative IgG level of d37, d45 and d85 by multiplying d30 OD/CO value by the corresponding ratio. Relative anti-S-RBD IgM level was calculated using the same protocol with anti-human IgM-HRP conjugated secondary antibody. Since the control cut-off serum for IgM was not available from the Kit, on Fig. S1b. we show OD values for nine biobanked pre-pandemic serum samples from healthy donors.

7.4.3 Isolation of PBMCs and T cell subpopulations

PBMCs were isolated with the Ficoll-Paque density gradient centrifugation protocol. CD4+ and CD8+ T cells were isolated from PBMCs with Dynabeads CD4+ and CD8+ positive selection kits (Invitrogen) respectively. For isolation of EM, EMRA, CM and SCM memory subpopulations we stained PBMCs with the following antibody mix: anti-CD3-FITC (UCHT1, eBioscience), anti-CD45RA-eFluor450 (HI100, eBioscience), anti-CCR7-APC (3D12, eBioscience), anti-CD95-PE (DX2, eBioscience). Cell sorting was performed on FACS Aria III, all four isolated subpopulations were lysed with Trizol reagent immediately after sorting.

7.4.4 TCR library preparation and sequencing

TCRalpha and TCRbeta cDNA libraries preparation was performed as previously described in Pogorelyy et al. (2017). RNA was isolated from each sample using Trizol reagent according to the manufacturer's instructions. A universal primer binding site, sample barcode and unique molecular identifier (UMI) sequences were introduced using the 5'RACE technology with TCRalpha and TCRbeta constant segment specific primers for cDNA synthesis. cDNA libraries were amplified in two PCR steps, with introduction of the second sample barcode and Illumina TruSeq adapter sequences at the second PCR step. Libraries were sequenced using the Illumina NovaSeq platform (2x150bp read length).

7.4.5 TCR repertoire data analysis

Raw data preprocessing. Raw sequencing data was demultiplexed and UMI guided consensuses were built using migec v.1.2.7 Shugay et al. (2014). Resulting UMI consensuses were aligned to V and J genomic templates of the TRA and TRB locus and assembled into clonotypes with mixcr v.2.1.11 Bolotin et al. (2015). See Table S1 for the number of cells, UMIs and unique clonotypes for each sample.

Identification of clonotypes with active dynamics. Principal component analysis (PCA) of clonal trajectories was performed as described before Minervina et al. (2020). First we selected clones which were present among the top 1000 abundant in any of post-infection PBMC repertoires, including biological replicates, i.e. considered clone abundant if it was found within top 1000 most abundant clonotypes in at least one of the replicate samples at one timepoint. Next, for each such abundant clone we calculated its frequency at each post-infection timepoint and divided this frequency by the maximum frequency of this clone for normalization. Then we performed PCA on the resulting normalized clonal trajectory matrix and identified three clusters of trajectories with hierarchical clustering with average linkage, using

Euclidean distances between trajectories.
We identify statistically significant contractions and expansions with edgeR as previously described Pogorelyy et al. (2018c), using FDR adjusted $p < 0.01$ and $\log_2$ fold change threshold of 1. NoisET implements the Bayesian detection method described in Puelma Touzel et al. (2020). Briefly, a two-step noise model accounting for cell sampling and expression noise is inferred from replicates, and a second model of expansion is learned from the two timepoints to be compared. The procedure outputs the posterior probability of expansion or contraction, and the median estimated $\log_2$ fold change, whose thresholds are set to 0.05 and 1 respectively.

Mapping of COVID-19 associated TCRs to the MIRA database. TCRbeta sequences from T cells specific for SARS-CoV-2 peptide pools MIRA (ImmuneCODE release 2) were downloaded from `https://clients.adaptivebiotech.com/pub/covid-2020`. V and J genomic templates were aligned to TCR nucleotide sequences from the MIRA database using mixcr 2.1.11. We consider a TCRbeta from MIRA matched to a TCRbeta from our data, if it had the same V and J and at most one mismatch in CDR3 amino acid sequence. We consider a TCRbeta sequence mapped to an epitope if it has at least two identical or highly similar (same V, J and up to one mismatch in CDR3 amino acid sequence) TCRbeta clonotypes reactive for this epitope in the MIRA database.

Computational alpha/beta pairing by clonal trajectories. Computational alpha/beta pairing was performed as described in Minervina et al. (2020). For each TCRbeta we determine the TCRalpha with the closest clonal trajectory (Tables S3 and S5). We observe no stringent pairings between TCRbeta and TCRbeta motifs with exception of two contracting CD8 TCRbeta clusters: TRBV7-2/TRBJ1-2 NQKLIANQF-associated clones from donor M paired to TRAV21/TRAJ40 alphas from the same cluster (CASSLEDTNYGYTF - CAVHSSGTYKYIF and CASSLEDTIYGYTF - CAALTSGTYKYIF), and TRBV7-9/TRBJ2-3 beta cluster paired to largest alpha cluster (CASSPTGRGRTDTQYF - CAYRSGGSEKLVF and CASSPTGRGGTDTQYF - CAYRRPGGEKLTF).

Computational prediction of HLA-types. To predict HLA-types from TCR repertoires of COVID-19 cohort we used sets of HLA-associated TCR sequences from DeWitt et al. (2018). We use TCRbeta repertoires of 666 donors from cohort from Emerson et al. (2017), for which HLA-typing information is available in ref. DeWitt et al. (2018) as a training set to fit logistic regression model, where presence or absense of given HLA-allele is an outcome, and the number of allele-associated sequences in repertoire, as well as the total number of unique sequences in the repertoire, are the predictors. A separate logistic regression model was fitted for

each set of HLA-associated sequences from ref. DeWitt et al. (2018), and then used to predict the probability p that a donor from the COVID-19 cohort has this allele. Donors with $p < 0.2$ were considered negative for a given allele.

7.4.6 Data availability

Raw sequencing data are deposited to the Short Read Archive (SRA) accession: PRJNA633317. Processed TCRalpha and TCRbeta repertoire datasets, resulting repertoires of SARS-CoV-2-reactive clones, and raw data preprocessing instructions can be accessed from: `https://github.com/pogorely/Minervina_COVID`.

7.4.7 Supplementary figures

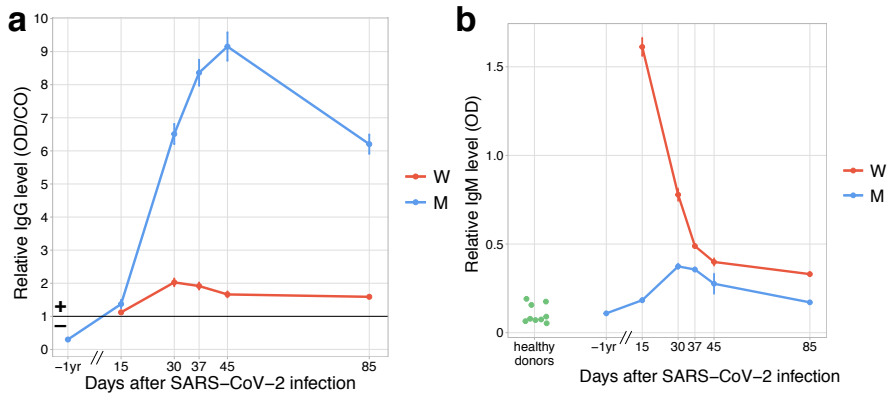

Figure 1 - Figure supplement 1. **Both donors developed anti-SARS-CoV-2 IgG and IgM responses by day 15 post infection. a,** The relative level of SARS-CoV-2 S-RBD domain specific IgG (y-axis) is plotted against time. Solid black line shows the threshold for positive testing. **b,** Relative IgM levels in donors M and W are shown over time. Relative IgM levels for pre-pandemic serum samples from healthy donors are shown on the left (green dots).

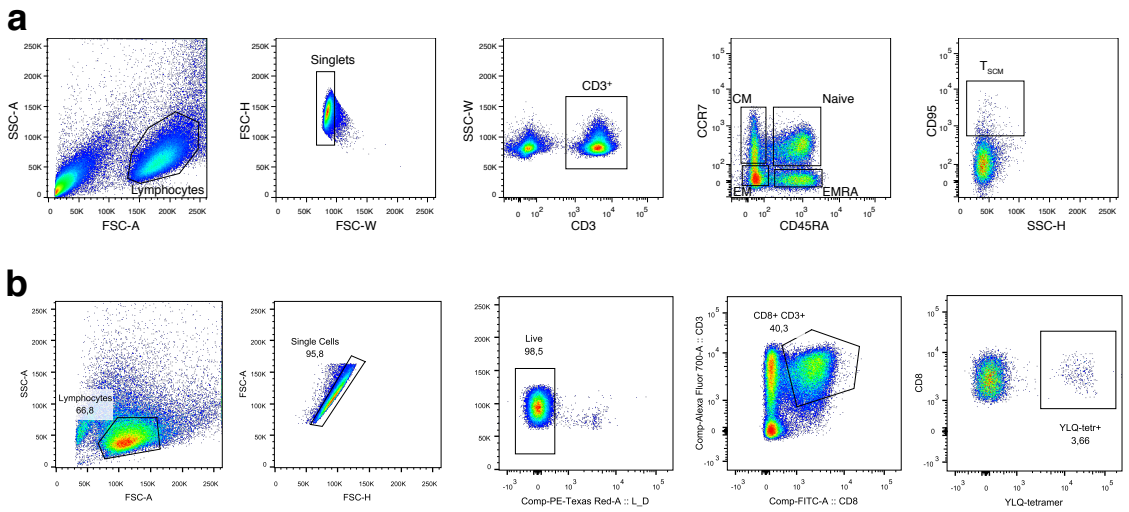

Figure 1 - Figure supplement 2. **a, Memory subpopulation gating strategy.** Three populations of memory T cells: EM, CM and EMRA are defined by CCR7 and CD45RA markers, SCM are distinguished from naive CCR7+ CD45RA+ T cells by CD95 expression. **b, HLA-A*02:01-YLQPRTFLL subpopulation gating strategy.** On day 25 post-infection donor M participated in study by Shomuradova et. al Shomuradova et al. (2020) (as donor p1434), where his T cells were *in vitro* expanded and stained with HLA-A*02:01-YLQPRTFLL tetramer, TCRalpha and TCRbeta repertoires were sequenced and resulting sequences deposited to VDJdb. See Shomuradova et al. (2020) for experimental details.

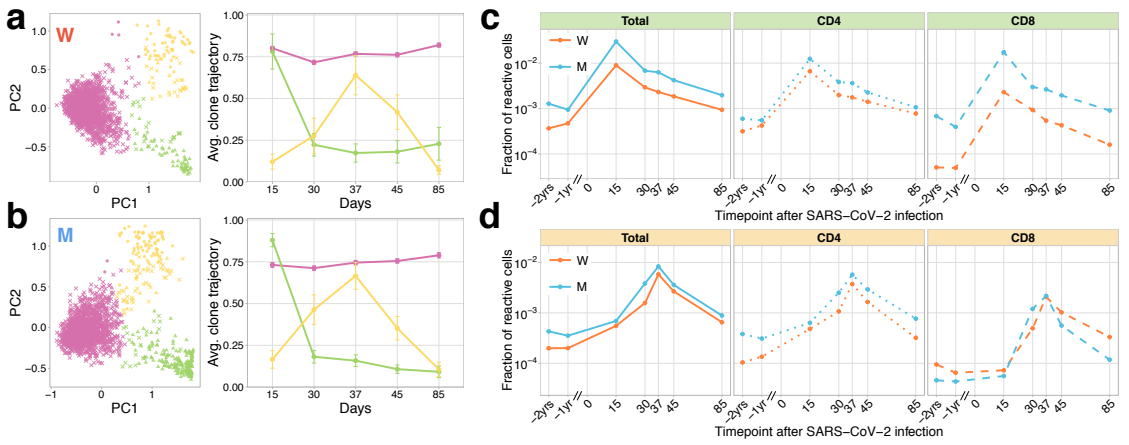

Figure 1 - Figure supplement 3. **Longitudinal tracking of T cell clones after mild COVID-19 with TCRalpha repertoire sequencing. a,b, PCA of clonal temporal trajectories identifies three groups of clones with distinctive behaviours.** Left: first two principal components of the 1000 most abundant TCRalpha clonotype frequencies normalized by maximum value for each clonotype in PBMC at post-infection timepoints. Color indicates hierarchical clustering results of principal components; symbol indicates if clonotype was called as significantly contracted from day 15 to day 85 (triangles), or expanded from day 15 to day 37 (circles) by both edgeR and NoisET (Fig. 1 suppl. 5 shows overlap between clonal trajectory clusters and edgeR/NoisET hits). Right: each curve shows the average ± 2.96 SE of normalized clonal frequencies from each cluster. **Contracting (c) and expanding (d) clones include both CD4+ and CD8+ T cells, and are less abundant in pre-infection repertoires.** T cell clones significantly contracted from day 15 to day 85 **(c)** and significantly expanded from day 15 to day 37 **(d)** were identified in both donors. The fraction of contracting **(c)** and expanding **(d)** TCRalpha clonotypes in the total repertoire (calculated as the sum of frequencies of these clonotypes in the second PBMC replicate at a given timepoint and corresponding to the fraction of responding cells of all T cells) is plotted in log-scale for all reactive clones (left), reactive clones with the CD4 (middle) and CD8 (right) phenotypes.

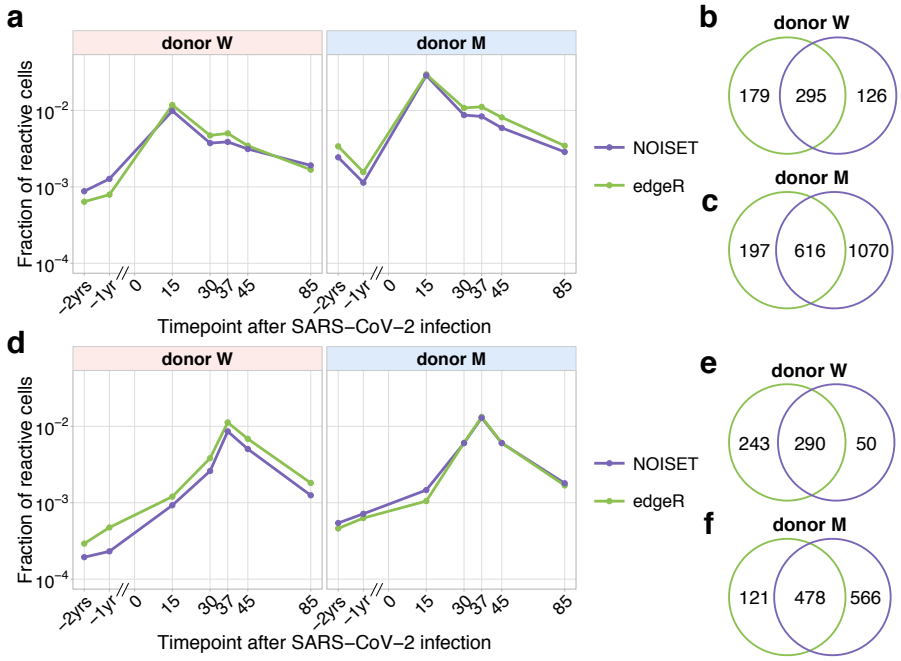

Figure 1 - Figure supplement 4. **Comparison of edgeR and NoisET clonal expansion detection procedures.** The fraction (plotted in the log-scale) of contracting **(a)** and expanding **(d)** TCRbeta clonotypes in the total repertoire was estimated using subsets of expanded and contracted clones called by edgeR (green) and NoisET (purple) models. Overlaps for contracted clones **(b,c)** and expanded clones **(e,f)** identified by both models are shown on the right.

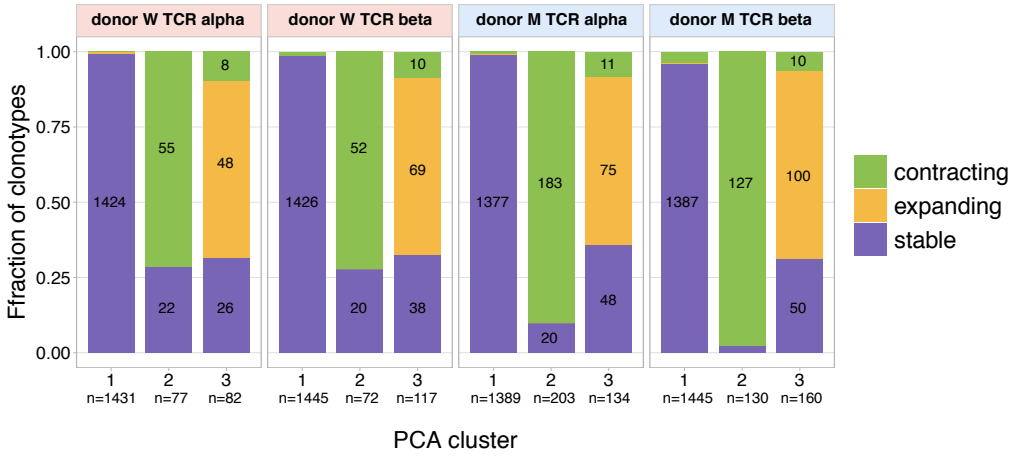

Figure 1 - Figure supplement 5. **The overlap between clusters of clonal trajectories identified by PCA and groups of expanding/contracting clones identified with edgeR/NoisET.** For each cluster of clonal trajectories identified on Fig. 1bc. and Fig. 1 suppl. 3ab we show overlap with groups of significantly (called by edgeR and NoisET simultaneously) expanding clonotypes from day 15 to day 37 in yellow, significantly contracting clonotypes from day 15 to day 85 in green, other clonotypes are shown in purple ("stable" clonotypes which were not called significant by edgeR and NoisET simultaneously). Bar heights show fraction of abundant clonotypes in PCA cluster overlapping with expanding/contracting/non-significant groups called by edgeR/NoisET, raw number of overlapping clonotypes is shown inside bars.

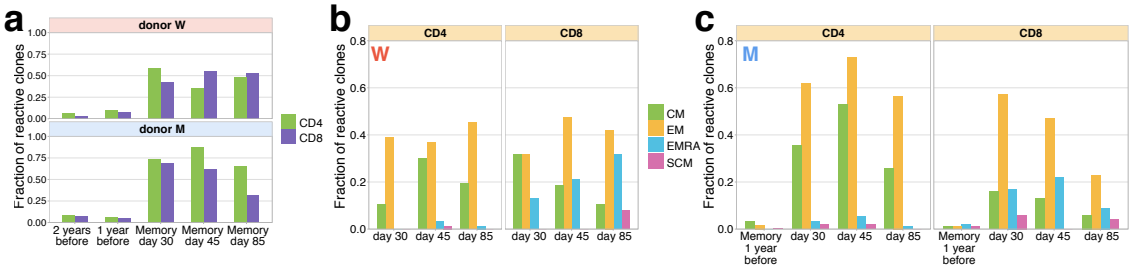

Figure 2 - Figure supplement 1. **Memory phenotypes of responding clonotypes expanding from day 15 to day 37. a, A fraction of expanding clonotypes is identified in T cell memory subsets after infection.** Bars show the fraction of expanding CD4+ and CD8+ TCRbeta clonotypes present in 2-year; 1-year pre-infection PBMC; and in at least one of memory subpopulation sampled on day 30 and day 37 post infection. **b, Responding clones are found in different memory subsets.** For both W **(b)** and M **(c)** donors, CD4+ clonotypes were found predominantly in Central Memory (CM) and Effector Memory (EM), while CD8+ T cells were also present in EMRA.

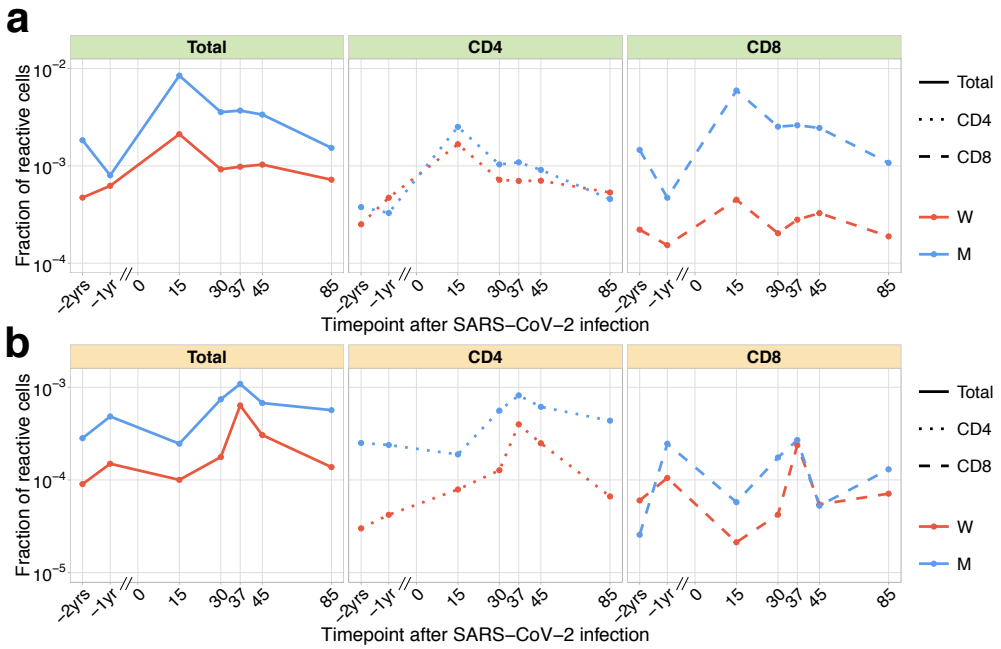

Figure 2 - Figure supplement 2. **Dynamics of pre-existing SARS-CoV-2 responding clones.** The fraction of pre-existing (identified in -1 yr and/or -2 yr timepoint pre-infection) contracting **(a)** and expanding **(b)** TCRbeta clonotypes in the total repertoire (corresponding to the fraction of responding cells of all T cells) is plotted in log-scale for all reactive clones (left), reactive clones with the CD4 (middle) and the CD8 phenotype (right).

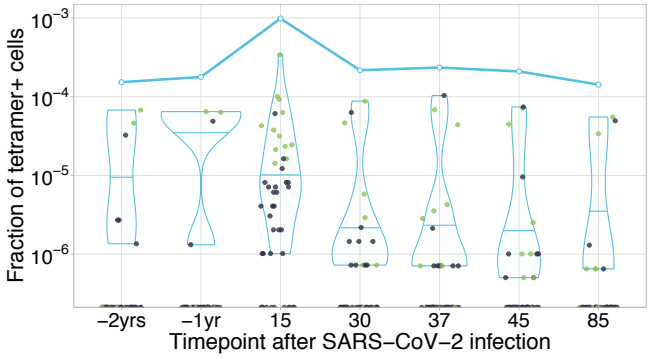

Figure 3 - Figure supplement 1. **HLA-A*02:01-YLQPRTFLL-specific TCRs are independently identified by clonal contraction.** Each dot corresponds to the frequency of HLA-A*02:01-YLQPRTFLL-tetramer specific TCRalpha in bulk repertoire from donor M (donor p1434 from Shomuradova et al. (2020)) at given timepoint (an estimate of fraction of tetramer+ cells of all CD3+ cells). Green dots correspond to clonotypes independently identified as contracting in our longitudinal dataset. Blue line shows cumulative frequency of HLA-A*02:01-YLQPRTFLL-tetramer specific TCRalpha clonotypes.

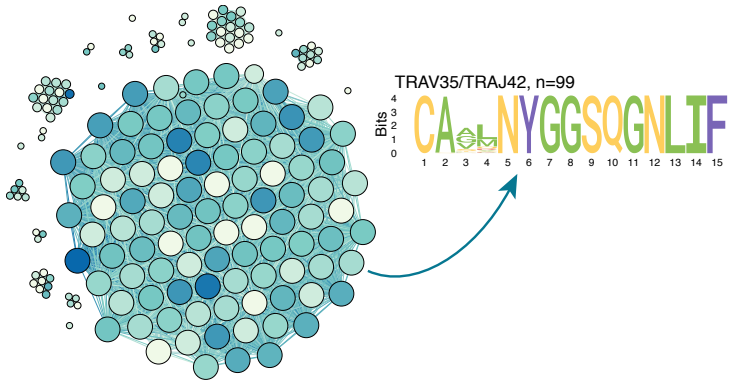

Figure 3 - Figure supplement 2. **ALICE algorithm output for TCRalpha PBMC repertoire of donor M on day 15.** Similarity network shows ALICE hits (clones in repertoire with more neighbours than expected by chance), which differ by 2 mismatches or less in TCRalpha amino acid sequence. Darker colors indicate larger frequency of clone in the repertoire, vertex size indicates degree. The majority (54%, 99/183) of hits identified by the algorithm correspond to a single large TRAV35/TRAJ42 cluster of CD4+ contracting clones also seen on Fig. 4a.

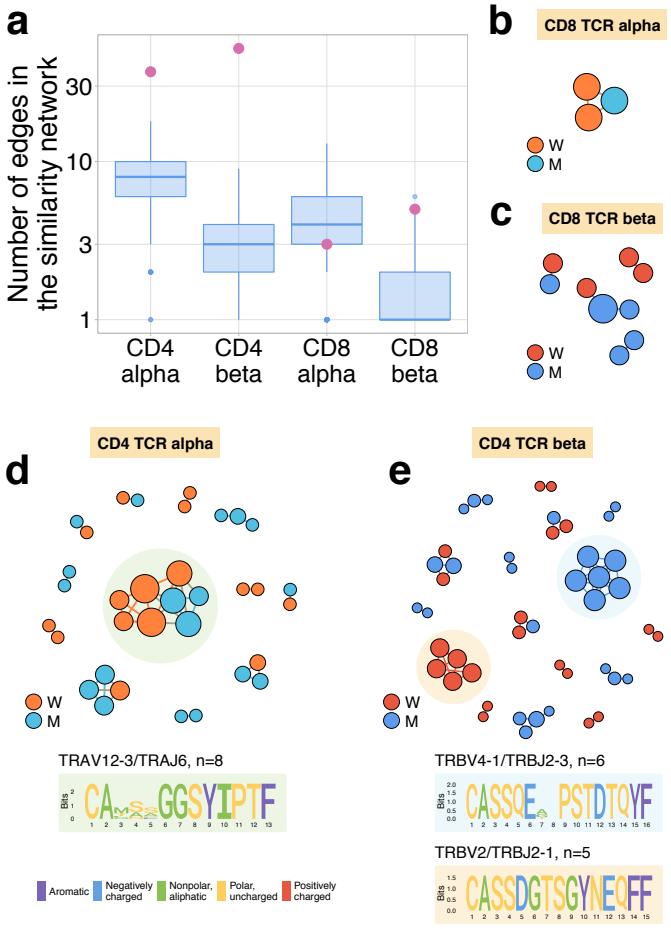

Figure 4 - Figure supplement 1. **a, Expanding CD4+ (but not CD8+) clonotypes show unexpected TCRalpha and TCRbeta sequence convergence.** For each set of CD4alpha, CD4beta, CD8alpha and CD8beta expanded clonotypes, we constructed separate similarity networks. Each vertex in the similarity network corresponds to an expanding clonotype. An edge indicates 2 or less amino acid mismatches in the CDR3 region, and identical V and J segments. The number of edges in each group is shown by pink dots and is compared to the distribution of that number in 1000 random samples of the same size from the relevant repertoires at day 37 (blue boxplots). **b, c, d, e Analysis of TCR amino acid sequences of expanding clones reveal distinctive motifs.** Networks are plotted separately for CD8alpha **(b)** and CD8beta **(c)** CD4alpha **(d)** and CD4beta **(e)** expanding clonotypes. Clonotypes without neighbours are not shown. Sequence logos corresponding to the largest clusters are shown under the corresponding network plots.

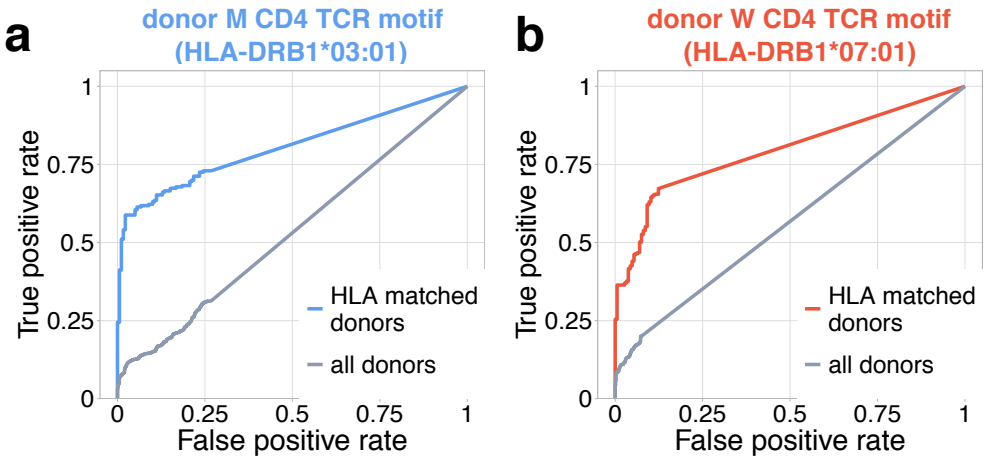

Figure 4 - Figure supplement 2. **Identification of COVID-19 patients by frequency of TCR motifs from contracting CD4+ clones from donors M (a) and W (b).** Receiver Operating Characteristic (ROC) curves for classifying TCRbeta repertoires from COVID cohort vs control by cumulative frequency of clones from CD4beta motifs. Blue curve shows ROC curve (area under the ROC or AUROC=0.8) for the classification of control and COVID donors predicted to be DRB1*03:01-DQB1*02:01 haplotype-positive with motif from donor M. Red curve show ROC curve (AUROC=0.79) for classification of control and COVID donors predicted to be DRB1*07:01-positive using motif from donor W. Grey ROC curves show classifier performance on all donors, irrespective of HLA allele matching (AUROC=0.53 for (**a**), AUROC=0.57 for (**b**)).

CHAPTER

8

CONCLUSION

8.1. Main contributions of this thesis

High-throughput sequencing experiments have shown that in all repertoire sequencing data sets, clonotype sizes follow a power-law distribution with an exponent close to -1 (chapter 1, chapter 3), with low variability between individuals. This result motivated the study of a universal population dynamic process for T-cell receptor repertoires dynamics for healthy individuals (Bensouda Koraichi et al. (2022)).

In a first-order Markov model, we have demonstrated that the logarithms of the abundances of different TCR clones follow geometric Brownian motion trajectories (whose stationary distribution is a power law). I extracted characteristic time scales of cell renewal and clonal renewal of different T-cell receptor repertoires from this model. New RepSeq data allow us to confront the model with reality. It is essential to understand data to extract information when working with it. It is, therefore, necessary to be aware of biotechnological progress and the different stages of the protocols leading to data generation.

The major challenge I had to tackle was considering the complex experimental and biological noise. Using Bayesian inference methods, I could infer the noise distribution parameters robustly. I then used all available data of healthy longitudinal repertoires (from 2011 to 2020) created by different laboratories with different experimental techniques to learn TCR population dynamics parameters. Inference of the dynamic with data of other individuals from other studies, ages, and sexes leads us to learn parameters following the stationary well-known power-law clonal

distribution. I was also able to learn characteristic time scales of T-cell immune repertoires and predict turnover rates of the T-cells repertoire. These turnover rates are highly dependent on the age of the analyzed individual and can give us insights into the diversity maintenance of the ecosystem with implications for immunosenescence questions. These results are giving us a powerful tool to quantify, for the first time, neutral dynamics of the TCR repertoire of healthy individuals. Understanding the dynamics of TCR clones when people are healthy can help us develop biomarkers of TCR repertoire deficiency and develop tools to improve the detection of efficient TCR against viruses, chronic diseases, and cancers, for which precise tools are needed.

The previously mentioned dynamic model describing T cell repertoires led me to pursue the Imprint project, published in PLoS Genetics Dupic et al. (2021), for which we show that a drop of blood combined with sequencing of the immune repertoire is sufficient to identify any person, even monozygous twins. We have proposed to use the repertoires as a versatile and innovative fingerprinting and identification tool that opens the debate on data collection, confidentiality, and ethics in this field.

Following the Sars-Cov2 pandemic, I developed the NoisET python software Bensouda Koraichi et al. (2021). This software allows for the use of experimental noise analysis methods of longitudinal samples of T cell repertoires to detect as precisely as possible the clonal families that have expanded between two given dates. It also offers valuable ways to facilitate the work of everyone who wants to analyze TCR clone dynamics from longitudinal RepSeq data. I have studied with this tool the Sars-Cov2 data, whose results have been published in eLife, Minervina et al. (2021). We have shown that hosts develop an active immune memory to the virus by identifying the families of clones involved and that the pre-existing immune memory existed in these people. A large community of scientists can use NoisET software to detect T-cell immune responses to vaccination. It could greatly help to study the heterogeneity of reactions to the SARS-Cov2 vaccine.

8.2. Future research directions

Antigen-induced clonal expansion is the main characteristic of an adaptive immune response, and measuring clonotype size is often a critical component of diagnosis. However, it is not accessible to precisely assess this number because of the already mentioned power-law clonal distributions that make evaluating count observations in the data and the biological noise challenging. Also, it can be helpful in longitudinal studies to evaluate neutral population dynamics to detect significant clonal expansions or contractions. My expertise in these tasks motivates me to integrate

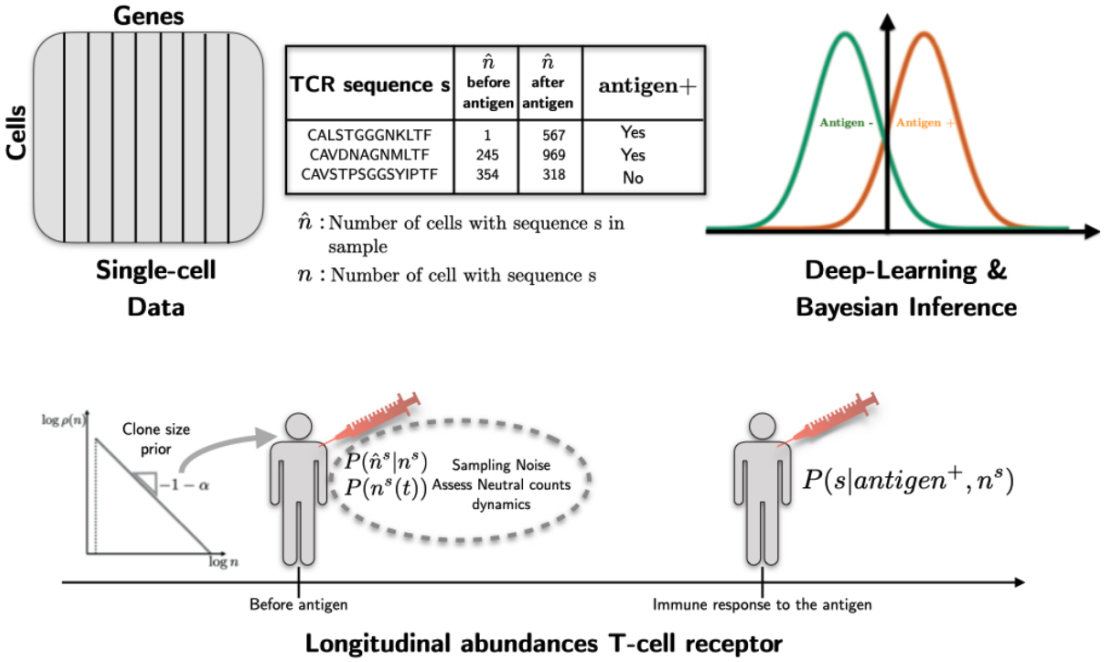

Figure S1: Use the knowledge learnt from this thesis about the experimental and biological noise monitoring the experimental counts of TCR to improve existing methods of TCR-epitope recognition thanks to RepSeq data.

this modeling to properly analyze correlations between properties encoded in TCR amino-acid sequences (where the biochemistry is hidden) that would have expanded because of the presence of a specific antigen.

Recently, it has become possible to simultaneously assay T-cell specificity concerning large sets of antigens (10x Genomics, 2019) and the T-cell receptor (TCR) sequence in high-throughput single-cell experiments. Recent advances in single-cell RNA-seq technologies allow for the detection of rare subpopulations that play important roles in host-pathogen interactions. Therefore, it would be great to gather all the TCR sequence information, T-cell specificity, gene expression in single-cell experiments, and T-cell abundances in a longitudinal data set to predict TCR-antigen pairing. Information contained in single-cell experiments reported on single-cell investigations is very rich and can give us clues about the functioning of a responsive T-cell. Taking advantage of new upcoming data, I want to learn distributions of TCR sequences that would be responsive to a specific antigen using single-cell data, sequence counts, and the usage of advanced generative protein models. This task is one of the most challenging ones in modern immunology. An outcome can significantly progress in finding more effective new treatments using protein design and drug discovery. The ability to accurately predict T-cell activation upon antigen recognition is still unsuccessful because of the lack of training data and adequate statistical models.

BIBLIOGRAPHY

Altan-Bonnet, G., T. Mora, and A. M. Walczak
2020. Quantitative immunology for physicists. *Physics Reports*, 849.

Arstila, T. P., A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky
1999. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, 286(5441):958–961.

Attaf, M., E. Huseby, and A. K. Sewell
2015. $\alpha\beta$ T cell receptors as predictors of health and disease. *Cellular & molecular immunology*, 12(4):391–399. Edition: 2015/01/26 Publisher: Nature Publishing Group.

Bacher, P., E. Rosati, D. Esser, G. Rios Martini, C. Saggau, E. Schiminsky, J. Dargvainiene, I. Schr&oumlder, I. Wieters, F. Eberhardt, H. Neb, Y. Khodamoradi, M. Sonntagbauer, M. J. Vehreschild, C. Conrad, F. Tran, P. Rosenstiel, R. Markewitz, K.-P. Wandinger, J. Rybniker, M. Kochanek, F. Leypoldt, O. A. Cornely, P. Koehler, A. Franke, and A. Scheffold
2020. Pre-existing T cell memory as a risk factor for severe 1 COVID-19 in the elderly. preprint, Allergy and Immunology.

Bagaev, D. V., R. M. A. Vroomans, J. Samir, U. Stervbo, C. Rius, G. Dolton, A. Greenshields-Watson, M. Attaf, E. S. Egorov, I. V. Zvyagin, N. Babel, D. K. Cole, A. J. Godkin, A. K. Sewell, C. Kesmir, D. M. Chudakov, F. Luciani, and M. Shugay
2020a. VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062.

Bagaev, D. V., R. M. A. Vroomans, J. Samir, U. Stervbo, C. Rius, G. Dolton, A. Greenshields-Watson, M. Attaf, E. S. Egorov, I. V. Zvyagin, N. Babel, D. K. Cole, A. J. Godkin, A. K. Sewell, C. Kesmir, D. M. Chudakov, F. Luciani, and M. Shugay
2020b. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062.

Bains, I., R. Antia, R. Callard, and A. J. Yates
2009a. Quantifying the development of the peripheral naive CD4+ T-cell pool in humans. *Blood*, 113(22):5480–5487.

Bains, I., R. Thiébaut, A. J. Yates, and R. Callard
2009b. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *Journal of immunology (Baltimore, Md. : 1950)*, 183(7):4329–4336.

Balachandran, V. P., , M. Łuksza, J. N. Zhao, V. Makarov, J. A. Moral, R. Remark, B. Herbst, G. Askan, U. Bhanot, Y. Senbabaoglu, D. K. Wells, C. I. O. Cary, O. Grbovic-Huezo, M. Attiyeh, B. Medina, J. Zhang, J. Loo, J. Saglimbeni, M. Abu-Akeel, R. Zappasodi, N. Riaz, M. Smoragiewicz, Z. L. Kelley, O. Basturk, M. Gönen, A. J. Levine, P. J. Allen, D. T. Fearon, M. Merad, S. Gnjatic, C. A. Iacobuzio-Donahue, J. D. Wolchok, R. P. DeMatteo, T. A. Chan, B. D. Greenbaum, T. Merghoub, and S. D. Leach
2017. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature*, 551(7681):512–516.

Barennes, P., V. Quiniou, M. Shugay, E. S. Egorov, A. N. Davydov, D. M. Chudakov, I. Uddin, M. Ismail, T. Oakes, B. Chain, A. Eugster, K. Kashofer, P. P. Rainer, S. Darko, A. Ransier, D. C. Douek, D. Klatzmann, and E. Mariotti-Ferrandiz
2020. Benchmarking of t cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology*.

Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni
2012. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191.

Bensouda Koraichi, M., S. Ferri, A. M. Walczak, and T. Mora
2022. Inferring the t-cells repertoire dynamics of healthy individuals. *bioRxiv*.

Bensouda Koraichi, M., M. P. Touzel, T. Mora, and A. M. Walczak
2021. NoisET: Noise learning and Expansion detection of T-cell receptors with Python. *arXiv:2102.03568 [q-bio]*.

Bi, Q., Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, and T. Feng
2020. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases*, P. S1473309920302875.

Bishop, C. M.
2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edition. Springer.

Bolotin, D. A., S. Poslavsky, A. N. Davydov, F. E. Frenkel, L. Fanchi, O. I. Zolotareva, S. Hemmers, E. V. Putintseva, A. S. Obraztova, and M. Shugay
2017. Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology*, 35(10):908–911.

Bolotin, D. A., S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, and D. M. Chudakov
2015. MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5):380–381.

Borghans, J. A. M. and R. J. De Boer
2007. Quantification of T-cell dynamics: From telomeres to DNA labeling. *Immunological Reviews*, 216(1):35–47.

Boyd, S. D., E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. Z. Fire
2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing. *Sci Transl Med*, 1(12):12ra23.

Bradley, P. and P. G. Thomas
2019. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annual Review of Immunology*, 37(1):547–570.

Braun, J., L. Loyal, M. Frentsch, D. Wendisch, P. Georg, F. Kurth, S. Hippenstiel, M. Dingeldey, B. Kruse, F. Fauchere, E. Baysal, M. Mangold, L. Henze, R. Lauster, M. Mall, K. Beyer, J. Roehmel, J. Schmitz, S. Miltenyi, M. A. Mueller, M. Witzenrath, N. Suttorp, F. Kern, U. Reimer, H. Wenschuh, C. Drosten, V. M. Corman, C. Giesecke-Thiel, L.-E. Sander, and A. Thiel
2020. Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. preprint, Infectious Diseases (except HIV/AIDS).

Bravi, B., V. P. Balachandran, B. D. Greenbaum, A. M. Walczak, T. Mora, R. Monasson, and S. Cocco
2021a. Probing T-cell response by sequence-based probabilistic modeling. *PLOS Computational Biology*, 17(9):e1009297.

Bravi, B., J. Tubiana, S. Cocco, R. Monasson, T. Mora, and A. M. Walczak
2021b. RBM-MHC: A Semi-Supervised Machine-Learning Method for Sample-Specific Prediction of Antigen Presentation by HLA-I Alleles. *Cell Systems*, 12(2):195–202.e9.

Britanova, O. V., E. V. Putintseva, M. Shugay, E. M. Merzlyak, M. A. Turchaninova, D. B. Staroverov, D. A. Bolotin, S. Lukyanov, E. A. Bogdanova, I. Z. Mamedov, Y. B. Lebedev, and D. M. Chudakov
2014. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *The Journal of Immunology*, 192(6):2689–2698.

Britanova, O. V., M. Shugay, E. M. Merzlyak, D. B. Staroverov, E. V. Putintseva, M. A. Turchaninova, I. Z. Mamedov, M. V. Pogorelyy, D. A. Bolotin, M. Izraelson, A. N. Davydov, E. S. Egorov, S. A. Kasatskaya, D. V. Rebrikov, S. Lukyanov, and D. M. Chudakov
2016. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *The Journal of Immunology*, 196(12):5005–5013.

Buhler, S., F. Bettens, C. Dantin, S. Ferrari-Lacraz, M. Ansari, A.-C. Mamez, S. Masouridi-Levrat, Y. Chalandon, and J. Villard
2020. Genetic T-cell receptor diversity at 1 year following allogeneic hematopoietic stem cell transplantation. *Leukemia*, 34(5):1422–1432.

Burgos, J. D. and P. Moreno-Tovar
1996. Zipf-scaling behavior in the immune system. *Biosystems*, 39(3):227–232.

Cai, L., N. Friedman, and X. S. Xie
2006. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362.

Chakraborty, A. K. and A. Košmrlj
2010. Statistical mechanical concepts in immunology. *Annual Review of Physical Chemistry*, 61(1):283–303. PMID: 20367082.

Chao, A. and J. Bunge
2002. Estimating the number of species in a stochastic abundance model. *Biometrics*, 58(3):531–539.

Chu, N. D., H. S. Bi, R. O. Emerson, A. M. Sherwood, M. E. Birnbaum, H. S. Robins, and E. J. Alm
2019. Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunology*, 20(1):19.

Dash, P., A. J. Fiore-gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, P. G. Thomas, N. L. L. Gruta, P. Bradley, and P. G. Thomas
2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547:89–93.

Dash, P., J. L. McClaren, T. H. Oguin, W. Rothwell, B. Todd, M. Y. Morris, J. Becksfort, C. Reynolds, S. A. Brown, P. C. Doherty, and P. G. Thomas
2011. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *Journal of Clinical Investigation*, 121(1):288–295.

Davis, M. M. and S. D. Boyd
2019. Recent progress in the analysis of $\alpha\beta$ T cell and B cell receptor repertoires. *Current Opinion in Immunology*, 59:109–114.

De Boer, R. J., D. Homann, and A. S. Perelson
2003. Different Dynamics of CD4 $^+$ and CD8 $^+$ T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection. *The Journal of Immunology*, 171(8):3928–3935.

De Boer, R. J. and A. S. Perelson
1994. T Cell Repertoires and Competitive Exclusion. *Journal of Theoretical Biology*, 169(4):375–390.

De Boer, R. J. and A. S. Perelson
2013. Quantifying T lymphocyte turnover. *Journal of Theoretical Biology*, 327:45–87.

de Greef, P. C., T. Oakes, B. Gerritsen, M. Ismail, J. M. Heather, R. Hermsen, B. Chain, and R. J. de Boer
2020. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife*, 9:e49900.

Desponds, J., A. Mayer, T. Mora, and A. M. Walczak
2021. *Population Dynamics of Immune Repertoires*, Pp. 203–221. Cham: Springer International Publishing.

Desponds, J., T. Mora, and A. M. Walczak
2016. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences*, 113(2):274–279.

Dessalles, R., Y. Pan, M. Xia, D. Maestrini, M. R. D'Orsogna, and T. Chou
2022. How Naive T-Cell Clone Counts Are Shaped By Heterogeneous Thymic Output and Homeostatic Proliferation. *Frontiers in Immunology*, 12.

DeWitt, W. S., R. O. Emerson, P. Lindau, M. Vignali, T. M. Snyder, C. Desmarais, C. Sanders, H. Utsugi, E. H. Warren, J. McElrath, K. W. Makar, A. Wald, and H. S. Robins
2015. Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection. *Journal of Virology*, 249(February):JVI.03474–14.

DeWitt, W. S., P. Lindau, T. M. Snyder, A. M. Sherwood, M. Vignali, C. S. Carlson, P. D. Greenberg, N. Duerkopp, R. O. Emerson, and H. S. Robins
2016. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PloS one*, 11(8):e0160853.

DeWitt, W. S., A. Smith, G. Schoch, J. A. Hansen, F. A. Matsen, and P. H. Bradley
2018. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *bioRxiv*, P. 313106.

Dowling, M. R. and P. D. Hodgkin
2009. Modelling naive T-cell homeostasis: Consequences of heritable cellular lifespan during ageing. *Immunology & Cell Biology*, 87(6):445–456.

Dupic, T., M. Bensouda Koraichi, A. A. Minervina, M. V. Pogorelyy, T. Mora, and A. M. Walczak
2021. Immune fingerprinting through repertoire similarity. *PLOS Genetics*, 17(1):e1009301.

Dupic, T., Q. Marcou, A. M. Walczak, and T. Mora
2018. Genesis of the $\alpha\beta$ T-cell receptor. *arXiv:1806.11030*.

Durbin, R., D. Richard, R. Durbin, S. Eddy, R. Eddy, S. Eddy, A. Krogh, K. Anders, and G. Mitchison
1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Biological Sequence Analysis: Probalistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Elhanati, Y., Z. Sethna, C. G. Callan, T. Mora, and A. M. Walczak
2018. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol Rev*, 284:167–179.

Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain
2002. Stochastic gene expression in a single cell". *Science*, 297(5584):1183–1186.

Emerson, R. O., W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, M. Rieder, and H. S. Robins
2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49(5):659–665.

Farber, D. L., N. A. Yudanin, and N. P. Restifo
2014. Human memory T cells: generation, compartmentalization and homeostasis. *Nat. Rev. Immunol.*, 14(1):24–35.

Fischer, D. S., Y. Wu, B. Schubert, and F. J. Theis
2020. Predicting antigen specificity of single T cells based on TCR CDR 3 regions. *Molecular Systems Biology*, 16(8).

Fuertes Marraco, S. A., C. Soneson, L. Cagnon, P. O. Gannon, M. Allard, S. Abed Maillard, N. Montandon, N. Rufer, S. Waldvogel, M. Delorenzi, and D. E. Speiser
2015. Long-lasting stem cell-like memory CD8+ T cells with a naïve-like profile upon yellow fever vaccination. *Science Translational Medicine*, 7(282):282ra48.

Gaimann, M. U., M. Nguyen, J. Desponds, and A. Mayer
2020. Early life imprints the hierarchy of T cell clone sizes. *eLife*, 9:e61639.

Gardiner, C. W.
2009. *Stochastic methods: A handbook for the natural and social sciences. Berlin: Springer.*

Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake
2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, 32(2):158–68.

Gielis, S., P. Moris, W. Bittremieux, N. De Neuter, B. Ogunjimi, K. Laukens, and P. Meysman
2019. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology*, 10.

Glanville, J., H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis
2017. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98.

Goronzy, J. J. and C. M. Weyand
2013. Understanding immunosenescence to improve responses to vaccines. *Nat. Immunol.*, 14(5):428–436.

Greef, P. C. D., T. Oakes, B. Gerritsen, M. Ismail, M. James, R. Hermsen, B. Chain, and R. J. D. Boer
2019. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *bioRxiv:691501*.

Grifoni, A., D. Weiskopf, S. I. Ramirez, J. Mateus, J. M. Dan, C. R. Moderbacher, S. A. Rawlings, A. Sutherland, L. Premkumar, R. S. Jadi, D. Marrama, A. M. de Silva, A. Frazier, A. Carlin, J. A. Greenbaum, B. Peters, F. Krammer, D. M. Smith, S. Crotty, and A. Sette
2020. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*.

Grigaityte, K., J. A. Carter, S. J. Goldfless, E. W. Jeffery, J. Ronald, Y. Jiang, D. Koppstein, A. W. Briggs, G. M. Church, and G. S. Atwal
2017. Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire. *bioRxiv:213462*.

Gutierrez, L., J. Beckford, and H. Alachkar
2020. Deciphering the TCR repertoire to solve the COVID-19 mystery. *Trends Pharmacol. Sci.*, 41(8):518–530.

Heather, J. M., M. Ismail, T. Oakes, and B. Chain
2017. High-throughput sequencing of the t-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*, P. bbw138.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig
2008. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLOS Genetics*, 4(8):e1000167.

Hornos, J. E. M., D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes
2005. Self-regulating gene: An exact solution. *Phys. Rev. E*, 72:051907.

Howie, B., A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins
2015. High-throughput pairing of T cell receptor α and β sequences. *Science Translational Medicine*, 7(301):301ra131.

Hozumi, N. and S. Tonegawa
1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences*, 73(10):3628–3632.

Isacchini, G., A. M. Walczak, T. Mora, and A. Nourmohammad
2021. Deep generative selection models of T and B cell receptor repertoires with soNNia. arXiv:2011.03112 [q-bio].

Jameson, S. C.
2002. Maintaining the norm: T-cell homeostasis. *Nature Reviews Immunology*, 2(8):547–556.

Johnson, P. L. F., A. J. Yates, J. J. Goronzy, and R. Antia
2012. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. *Proceedings of the National Academy of Sciences*, 109(52):21432–21437.

Kedzierska, K., S. A. Valkenburg, P. C. Doherty, M. P. Davenport, and V. Venturi
2012. Use it or lose it: Establishment and persistence of T cell memory. *Frontiers in Immunology*, 3.

Khan, N., N. Shariff, M. Cobbold, R. Bruton, J. a. Ainsworth, A. J. Sinclair, L. Nayak, and P. a. H. Moss
2002. Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality in healthy elderly individuals. *Journal of immunology (Baltimore, Md. : 1950)*, 169(4):1984–1992.

Klinger, M., F. Pepin, J. Wilkins, T. Asbury, T. Wittkop, J. Zheng, M. Moorhead, and M. Faham
2015. Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLOS ONE*, 10(10):e0141561.

Koch, H., D. Starenki, S. J. Cooper, R. M. Myers, and Q. Li
2018. powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire. *PLOS Computational Biology*, 14(11):e1006571.

Laydon, D. J., C. R. M. Bangham, and B. Asquith
2015. Estimating t-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1675):20140291.

Le Bert, N., A. T. Tan, K. Kunasegaran, C. Y. L. Tham, M. Hafezi, A. Chia, M. H. Y. Chng, M. Lin, N. Tan, M. Linster, W. N. Chia, M. I.-C. Chen, L.-F. Wang, E. E. Ooi, S. Kalimuddin, P. A. Tambyah, J. G.-H. Low, Y.-J. Tan, and A. Bertoletti
2020. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*, 584(7821):457–462.

Li, B., T. Li, B. Wang, R. Dou, J. Zhang, J. S. Liu, and X. S. Liu
2017. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. *Nature Genetics*, 49(4):482–483.

Lindau, P. and H. S. Robins
2017. Advances and applications of immune receptor sequencing in systems immunology. *Current Opinion in Systems Biology*, 1:62–68.

Lythe, G., R. E. Callard, R. Hoare, and C. Molina-París
2015. How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224.

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

Marcou, Q., T. Mora, and A. M. Walczak
2018. High-throughput immune repertoire analysis with IGoR. *Nature Communications*, 9(1):561.

Mateus, J., A. Grifoni, A. Tarke, J. Sidney, S. I. Ramirez, J. M. Dan, Z. C. Burger, S. A. Rawlings, D. M. Smith, E. Phillips, S. Mallal, M. Lammers, P. Rubiro, L. Quiambao, A. Sutherland, E. D. Yu, R. da Silva Antunes, J. Greenbaum, A. Frazier, A. J. Markmann, L. Premkumar, A. de Silva, B. Peters, S. Crotty, A. Sette, and D. Weiskopf
2020. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science*, P. eabd3871.

Mayer, A., Y. Zhang, A. S. Perelson, and N. S. Wingreen
2019. Regulation of T cell expansion by antigen presentation dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 116(13):5914–5919.

Mayer-Blackwell, K., S. Schattgen, L. Cohen-Lavi, J. C. Crawford, A. Souquette, J. A. Gaevert, T. Hertz, P. G. Thomas, P. G. Bradley, and A. Fiore-Gartland
2021. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife*, 10:e68605.

Meckiff, B. J., C. Ramírez-Suástegui, V. Fajardo, S. J. Chee, A. Kusnadi, H. Simon, S. Eschweiler, A. Grifoni, E. Pelosi, D. Weiskopf, A. Sette, F. Ay, G. Seumois, C. H. Ottensmeier, and P. Vijayanand
2020. Imbalance of Regulatory and Cytotoxic SARS-CoV-2-Reactive CD4+ T Cells in COVID-19. *Cell*, 183(5):1340–1353.e16.

Miller, J. D., R. G. van der Most, R. S. Akondy, J. T. Glidewell, S. Albott, D. Masopust, K. Murali-Krishna, P. L. Mahar, S. Edupuganti, S. Lalor, S. Germon, C. Del Rio, M. J. Mulligan, S. I. Staprans, J. D. Altman, M. B. Feinberg, and R. Ahmed
2008. Human effector and memory cd8+ t cell responses to smallpox and yellow fever vaccines. *Immunity*, 28(5):710–722.

Minervina, A., M. Pogorelyy, and I. Mamedov
2019a. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transplant International*, 32(11):1111–1123.

Minervina, A., M. Pogorelyy, and I. Mamedov
2019b. TCR and BCR repertoire profiling in adaptive immunity. *Transplant International*, Pp. 0–2.

Minervina, A. A., E. A. Komech, A. Titov, M. B. Koraichi, E. Rosati, I. Z. Mamedov, A. Franke, G. A. Efimov, D. M. Chudakov, T. Mora, A. M. Walczak, Y. B. Lebedev, and M. V. Pogorelyy
2021. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of t-cell memory formation after mild COVID-19 infection. *eLife*, 10.

Minervina, A. A., M. V. Pogorelyy, E. A. Komech, V. K. Karnaukhov, P. Bacher, E. Rosati, A. Franke, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, and A. M. Walczak
2020. Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife*, 9:e53704.

Montemurro, A., V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters, L. E. Jessen, and M. Nielsen
2020. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR and sequence data. P. 13.

Mora, T. and A. M. Walczak
2019a. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology*, 18:104–110.

Mora, T. and A. M. Walczak
2019b. Quantifying lymphocyte receptor diversity. In *Systems Immunology*. CRC Press.

Murphy, K., P. Travers, and M. Walport
2007. *Janeway's Immunology*, 7 edition edition. Garland Science.

Murugan, A., T. Mora, A. M. Walczak, and C. G. Callan
2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166.

Naumov, Y. N., E. N. Naumova, K. T. Hogan, L. K. Selin, and J. Gorski
2003. 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.

Naveed, M., E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang
2015. Privacy in the Genomic Era. *ACM computing surveys*, 48(1).

Ng, O.-W., A. Chia, A. T. Tan, R. S. Jadi, H. N. Leong, A. Bertoletti, and Y.-J. Tan
2016. Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. *Vaccine*, 34(17):2008–2014.

Ni, L., F. Ye, M.-L. Cheng, Y. Feng, Y.-Q. Deng, H. Zhao, P. Wei, J. Ge, M. Gou, X. Li, L. Sun, T. Cao, P. Wang, C. Zhou, R. Zhang, P. Liang, H. Guo, X. Wang, C.-F. Qin, F. Chen, and C. Dong
2020. Detection of SARS-CoV-2-Specific Humoral and Cellular Immunity in COVID-19 Convalescent Individuals. *Immunity*, P. S1074761320301813.

Oakes, T., J. M. Heather, K. Best, R. Byng-Maddick, C. Husovsky, M. Ismail, K. Joshi, G. Maxwell, M. Noursadeghi, N. Riddell, T. Ruehl, C. T. Turner, I. Uddin, and B. Chain
2017. Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Frontiers in Immunology*, 8(OCT):1–17.

Oh, H.-L. J., A. Chia, C. X. L. Chang, H. N. Leong, K. L. Ling, G. M. Grotenbreg, A. J. Gehring, Y. J. Tan, and A. Bertoletti
2011. Engineering T Cells Specific for a Dominant Severe Acute Respiratory

Syndrome Coronavirus CD8 T Cell Epitope. *Journal of Virology*, 85(20):10464–10471.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden
2002. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73.

Pai, J. A. and A. T. Satpathy
2021. High-throughput and single-cell T cell receptor sequencing technologies. *Nature Methods*, 18(8):881–892.

Peng, Y., A. J. Mentzer, G. Liu, X. Yao, Z. Yin, D. Dong, W. Dejnirattisai, T. Rostron, P. Supasa, C. Liu, C. Lopez-Camacho, J. Slon-campos, Y. Zhao, D. Stuart, G. Paeson, J. Grimes, F. Antson, O. W. Bayfield, D. E. Hawkins, D.-S. Ker, L. Turtle, K. Subramaniam, P. Thomson, P. Zhang, C. Dold, J. Ratcliff, P. Simmonds, T. de Silva, P. Sopp, D. Wellington, U. Rajapaksa, Y.-L. Chen, M. Salio, G. Napolitani, W. Paes, P. Borrow, B. Kessler, J. W. Fry, N. F. Schwabe, M. G. Semple, K. J. Baillie, S. Moore, P. J. Openshaw, A. Ansari, S. Dunachie, E. Barnes, J. Frater, G. Kerr, P. Goulder, T. Lockett, R. Levin, Oxford Immunology Network Covid-19 Response T cell Consortium, R. J. Cornall, C. Conlon, P. Klenerman, A. McMichael, G. Screaton, J. Mongkolsapaya, J. C. Knight, G. Ogg, and T. Dong
2020. Broad and strong memory CD4 $^{+}$ and CD8 $^{+}$ T cells induced by SARS-CoV-2 in UK convalescent COVID-19 patients. preprint, Immunology.

Pogorelyy, M. V., Y. Elhanati, Q. Marcou, A. L. Sycheva, E. A. Komech, V. I. Nazarov, O. V. Britanova, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, and A. M. Walczak
2017. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLOS Computational Biology*, 13(7):e1005572.

Pogorelyy, M. V., A. A. Minervina, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, and A. M. Walczak
2018a. Method for identification of condition-associated public antigen receptor sequences. *eLife*, 7(D):1–13.

Pogorelyy, M. V., A. A. Minervina, M. Shugay, D. M. Chudakov, Y. B. Lebedev, T. Mora, and A. M. Walczak
2018b. Detecting T-cell receptors involved in immune responses from single repertoire snapshots. *arXiv:1807.08833*.

Pogorelyy, M. V., A. A. Minervina, M. P. Touzel, A. L. Sycheva, E. A. Komech, E. I. Kovalenko, G. G. Karganova, E. S. Egorov, A. Y. Komkov, D. M. Chudakov,

I. Z. Mamedov, T. Mora, A. M. Walczak, and Y. B. Lebedev
2018c. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences*, P. 201809642.

Puelma Touzel, M., A. M. Walczak, and T. Mora
2020. Inferring the immune response from repertoire sequencing. *PLOS Computational Biology*, 16(4):e1007873.

Qi, Q., M. M. Cavanagh, S. Le Saux, H. NamKoong, C. Kim, E. Turgano, Y. Liu, C. Wang, S. Mackey, G. E. Swan, C. L. Dekker, R. A. Olshen, S. D. Boyd, C. M. Weyand, L. Tian, and J. J. Goronzy
2016. Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Science Translational Medicine*, 8(332):332ra46–332ra46.

Qi, Q., Y. Liu, Y. Cheng, J. Glanville, D. Zhang, J.-Y. Lee, R. a. Olshen, C. M. Weyand, S. D. Boyd, and J. J. Goronzy
2014. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 111(36):13139–44.

Quinti, I., V. Lougaris, C. Milito, F. Cinetto, A. Pecoraro, I. Mezzaroma, C. M. Mastroianni, O. Turriziani, M. P. Bondioni, M. Filippini, A. Soresina, G. Spadaro, C. Agostini, R. Carsetti, and A. Plebani
2020. A possible role for B cells in COVID-19? Lesson from patients with agammaglobulinemia. *Journal of Allergy and Clinical Immunology*, P. S0091674920305571.

Rane, S., T. Hogan, E. Lee, B. Seddon, and A. J. Yates
2022. Defining the dynamics of naive cd4 and cd8 t cells across the mouse lifespan. *bioRxiv*.

Redmond, D., A. Poran, and O. Elemento
2016. Single-cell TCRseq: Paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Medicine*, 8(1):80.

Robins, H.
2013. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, 25(5):646–52.

Robins, H. S., P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson
2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alpha-beta T cells. *Blood*, 114(19):4099–4107.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth
2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140.

Rubelt, F., C. E. Busse, S. Ahmad, C. Bukhari, J.-p. Bürckert, E. Mariotti-ferrandiz, L. G. Cowell, C. T. Watson, N. Marthandan, W. J. Faison, U. Hershberg, U. Laserson, B. D. Corrie, M. M. Davis, B. Peters, and M.-p. Lefranc
2017. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature immunology*, 18(12):1274–1278.

Rytlewski, J., S. Deng, T. Xie, C. Davis, H. Robins, E. Yusko, and J. Bienkowska
2019. Model to improve specificity for identification of clinically-relevant expanded t cells in peripheral blood. *PLOS ONE*, 14(3):e0213684.

Schmidt, M. E. and S. M. Varga
2018. The CD8 T Cell Response to Respiratory Virus Infections. *Frontiers in Immunology*, 9:678.

Sekine, T., A. Perez-Potti, O. Rivera-Ballesteros, K. Strålin, J.-B. Gorin, A. Olsson, S. Llewellyn-Lacey, H. Kamal, G. Bogdanovic, S. Muschiol, D. J. Wullimann, T. Kammann, J. Emgård, T. Parrot, E. Folkesson, O. Rooyackers, L. I. Eriksson, J.-I. Henter, A. Sönnerborg, T. Allander, J. Albert, M. Nielsen, J. Klingström, S. Gredmark-Russ, N. K. Björkström, J. K. Sandberg, D. A. Price, H.-G. Ljunggren, S. Aleman, and M. Buggert
2020. Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *Cell*, P. S0092867420310084.

Sethna, Z., Y. Elhanati, C. Callan, A. M. Walczak, and T. Mora
2019. OLGA: Fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981.

Sethna, Z., G. Isacchini, T. Dupic, T. Mora, A. M. Walczak, and Y. Elhanati
2020. Population variability in the generation and thymic selection of T-cell repertoires. *arXiv:2001.02843*, Pp. 1–17.

Shomuradova, A. S., M. S. Vagida, S. A. Sheetikov, K. V. Zornikova, D. Kiryukhin, A. Titov, I. O. Peshkova, A. Khmelevskaya, D. V. Dianov, M. Malasheva, A. Shmelev, Y. Serdyuk, D. V. Bagaev, A. Pivnyuk, D. S. Shcherbinin, A. V. Maleeva, N. T. Shakirova, A. Pilunov, D. B. Malko, E. G. Khamaganova, B. Biderman, A. Ivanov, M. Shugay, and G. A. Efimov
2020. SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T cell receptors. *Immunity*, P. S1074761320304696.

Shugay, M., D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, A. V. Eliseev, E. Van Dyk, P. Dash, M. Attaf, C. Rius, K. Ladell, J. E. McLaren, K. K. Matthews, E. B. Clemens, D. C. Douek, F. Luciani, D. Van Baarle, K. Kedzierska, C. Kesmir, P. G. Thomas, D. A. Price, A. K. Sewell, and D. M. Chudakov
2018. VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 46(D1):D419–D427.

Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov
2014. Towards error-free profiling of immune repertoires. *Nature Methods*, 11(6):653–655.

Sidhom, J.-W., H. B. Larman, D. M. Pardoll, and A. S. Baras
2021. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, 12(1):1605.

Six, A., M. E. Mariotti-Ferrandiz, W. Chaara, S. Magadan, H.-P. P. Pham, M.-P. P. Lefranc, T. Mora, V. Thomas-Vaslin, A. M. Walczak, P. Boudinot, E. Mariotti-Ferrandiz, W. Chaara, S. Magadan, H.-P. P. Pham, M.-P. P. Lefranc, T. Mora, V. Thomas-Vaslin, A. M. Walczak, and P. Boudinot
2013. The past, present and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Frontiers in Immunology*, 4(413):413.

Snyder, T. M., R. M. Gittelman, M. Klinger, D. H. May, E. J. Osborne, R. Taniguchi, H. J. Zahid, I. M. Kaplan, J. N. Dines, M. N. Noakes, R. Pandya, X. Chen, S. Elasady, E. Svejnoha, P. Ebert, M. W. Pesesky, P. De Almeida, H. O'Donnell, Q. DeGottardi, G. Keitany, J. Lu, A. Vong, R. Elyanow, P. Fields, J. Greissl, L. Baldo, S. Semprini, C. Cerchione, M. Mazza, O. M. Delmonte, K. Dobbs, G. Carreño-Tarragona, S. Barrio, L. Imberti, A. Sottini, E. Quiros-Roldan, C. Rossi, A. Biondi, L. R. Bettini, M. D'Angio, P. Bonfanti, M. F. Tompkins, C. Alba, C. Dalgard, V. Sambri, G. Martinelli, J. D. Goldman, J. R. Heath, H. C. Su, L. D. Notarangelo, J. Martinez-Lopez, J. M. Carlson, and H. S. Robins
2020. Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. preprint, Infectious Diseases (except HIV/AIDS).

Soresina, A., D. Moratto, M. Chiarini, C. Paolillo, G. Baresi, E. Focà, M. Bezzi, B. Baronio, M. Giacomelli, and R. Badolato
2020. Two X[U+2010]linked agammaglobulinemia patients develop pneumonia as

COVID[U+2010]19 manifestation but recover. *Pediatric Allergy and Immunology*, P. pai.13263.

Springer, I., N. Tickotsky, and Y. Louzoun
2021. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, 12.

Stamatakis, G., M. Samiotaki, A. Mpakali, G. Panayotou, and E. Stratikos
2020. Generation of SARS-CoV-2 S1 spike glycoprotein putative antigenic epitopes in vitro by intracellular aminopeptidases. preprint, Biochemistry.

Swain, S. L., K. K. McKinstry, and T. M. Strutt
2012. Expanding roles for CD4+ T cells in immunity to viruses. *Nature Reviews Immunology*, 12(2):136–148.

Sweeney, L., A. Abu, and J. Winn
2013. Identifying Participants in the Personal Genome Project by Name. SSRN Scholarly Paper ID 2257732, Social Science Research Network, Rochester, NY.

Sycheva, A. L., M. V. Pogorelyy, E. A. Komech, A. A. Minervina, I. V. Zvyagin, D. B. Staroverov, D. M. Chudakov, Y. B. Lebedev, and I. Z. Mamedov
2018. Quantitative profiling reveals minor changes of T cell receptor repertoire in response to subunit inactivated influenza vaccine. *Vaccine*, 36(12):1599–1605.

Sylwester, A. W., B. L. Mitchell, J. B. Edgar, C. Taormina, C. Pelte, F. Ruchti, P. R. Sleath, K. H. Grabstein, N. A. Hosken, F. Kern, J. A. Nelson, and L. J. Picker
2005. Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. *Journal of Experimental Medicine*, 202(5):673–685.

Taniguchi, Y., P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie
2010. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538.

Tanno, H., T. M. Gould, J. R. McDaniel, W. Cao, Y. Tanno, R. E. Durrett, D. Park, S. J. Cate, W. H. Hildebrand, C. L. Dekker, L. Tian, C. M. Weyand, G. Georgiou, and J. J. Goronzy
2020. Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proceedings of the National Academy of Sciences*, 117(1):532–540.

Thevarajan, I., T. H. O. Nguyen, M. Koutsakos, J. Druce, L. Caly, C. E. van de Sandt, X. Jia, S. Nicholson, M. Catton, B. Cowie, S. Y. C. Tong, S. R. Lewin, and K. Kedzierska
2020. Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe COVID-19. *Nature Medicine*, 26(4):453–455.

Thomas-Vaslin, V., H. K. Altes, R. J. de Boer, and D. Klatzmann
2008. Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. *The Journal of Immunology*, 180(4):2240–2250.

Touzel, M. P., A. M. Walczak, and T. Mora
2020. Inferring the immune response from repertoire sequencing. *PLoS Computational Biology*, 16(4):1–21.

Vabret, N., G. J. Britton, C. Gruber, S. Hegde, J. Kim, M. Kuksin, R. Levantovsky, L. Malle, A. Moreira, M. D. Park, L. Pia, E. Risson, M. Saffern, B. Salomé, M. E. Selvan, M. P. Spindler, J. Tan, V. van der Heide, J. K. Gregory, K. Alexandropoulos, N. Bhardwaj, B. D. Brown, B. Greenbaum, Z. H. Gümüş, D. Homann, A. Horowitz, A. O. Kamphorst, M. A. Curotto de Lafaille, S. Mehandru, M. Merad, and R. M. Samstein
2020. Immunology of COVID-19: current state of the science. *Immunity*, P. S1074761320301837.

Valkiers, S., N. de Vrij, S. Gielis, S. Verbandt, B. Ogunjimi, K. Laukens, and P. Meysman
2022. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics*, 5:100009.

Venturi, V., H. Y. Chin, D. A. Price, D. C. Douek, and M. P. Davenport
2008. The Role of Production Frequency in the Sharing of Simian Immunodeficiency Virus-Specific CD8+ TCRs between Macaques. *The Journal of Immunology*, 181(4):2597–2609.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt
2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Weinstein, J. A., N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake
2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810.

Weiskopf, D., K. S. Schmitz, M. P. Raadsen, A. Grifoni, N. M. Okba, H. Endeman, J. P. van den Akker, R. Molenkamp, M. P. Koopmans, E. C. van Gorp, B. L. Haagmans, R. L. de Swart, A. Sette, and R. D. de Vries
2020. Phenotype of sars-cov-2-specific t-cells in covid-19 patients with acute respiratory distress syndrome. *medRxiv*.

Wolf, K., T. Hether, P. Gilchuk, A. Kumar, A. Rajeh, C. Schiebout, J. Maybruck, R. M. Buller, T. H. Ahn, S. Joyce, and R. J. DiPaolo
2018. Identifying and Tracking Low-Frequency Virus-Specific TCR Clonotypes Using High-Throughput Sequencing. *Cell Reports*, 25(9):2369–2378.e4.

Woodsworth, D. J., M. Castellarin, and R. a. Holt
2013. Sequence analysis of T-cell repertoires in health and disease. *Genome medicine*, 5(10):98.

Wu, K., E. Y. Kathryn, D. Bence, B. Julia A., X. Yu, E. Takeshi, S. Ansuman, C. Howard Y., and Z. James
2021. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. P. 39.

Yates, A. J.
2014. Theories and quantification of thymic selection. *Frontiers in Immunology*, 5(FEB):13.

Yunshun, Chen and Aaron, Lun and Davis, McCarthy and Xiaobei, Zhou and Mark, Robinson and Gordon, Smyth
2017. edger.

Zarnitsyna, V. I., B. D. Evavold, L. N. Schoettle, J. N. Blattman, and R. Antia
2013. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in Immunology*, 4(DEC):485.

Zhang, H., C. M. Weyand, and J. J. Goronzy
2021a. Hallmarks of the aging T-cell system. *The FEBS Journal*, 288(24):7123–7142.

Zhang, W., P. G. Hawkins, J. He, N. T. Gupta, J. Liu, G. Choonoo, S. W. Jeong, C. R. Chen, A. Dhanik, M. Dillon, R. Deering, L. E. Macdonald, G. Thurston, and G. S. Atwal
2021b. A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances*, 7(20):eabf5835.

Zhao, J., J. Zhao, and S. Perlman

2010. T Cell Responses Are Required for Protection from Clinical Disease and for Virus Clearance in Severe Acute Respiratory Syndrome Coronavirus-Infected Mice. *Journal of Virology*, 84(18):9318–9325.

Zvyagin, I. V., M. V. Pogorelyy, M. E. Ivanova, E. A. Komech, M. Shugay, D. A. Bolotin, A. A. Shelenkov, A. A. Kurnosov, D. B. Staroverov, D. M. Chudakov, Y. B. Lebedev, and I. Z. Mamedov

2014. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 111(16):5980–5985.

RÉSUMÉ

Le système immunitaire adaptatif comprend une diversité de lymphocytes T capables de reconnaître un large éventail d'antigènes. La spécificité de chaque cellule T pour les antigènes est déterminée par ses récepteurs (TCR), qui forment un répertoire de milliards de récepteurs uniques chez chaque individu. Les cellules T possèdent un récepteur capable de reconnaître l'agent pathogène ou la cellule maligne que l'organisme veut éliminer et se multiplient suite à cette reconnaissance. Ces événements se produisent non seulement après un stimulus aigu, mais aussi en permanence, car le système immunitaire interagit constamment avec l'environnement extérieur. En utilisant la dynamique stochastique des populations pour faire correspondre le répertoire de lymphocytes T à un système écologique, et des outils issus de la théorie de l'inférence bayésienne, nous construisons et confrontons des modèles biophysiques à des données réelles. Nous tirons parti de la disponibilité de nouvelles données dû aux progrès majeurs du séquençage à haut débit et donc, de la génération de données sur le répertoire de lymphocytes T qui contient beaucoup d'informations à déchiffrer. Grâce à ces méthodes, le résultat est multiple, nous sommes capables d'interpréter les données et le bruit qui les sous-tend, de comprendre la dynamique du répertoire de lymphocytes T en présence ou en l'absence d'un stimulus fort sur une échelle de temps courte ou longue, de caractériser la spécificité de chaque répertoire TCR avec le temps et de développer des outils qui peuvent être utilisés par tout immunologiste quantitatif intéressé par l'étude de la dynamique du répertoire de lymphotcytes T.

MOTS CLÉS

immunologie, Biophysique statistique, Inférence, Dynamique de populations

ABSTRACT

The adaptive immune system includes a diversity of T cells capable of recognizing a wide range of antigens. The specificity of each T cell for antigens is determined by its T cell receptors (TCRs), which together form a repertoire of billions of unique receptors in each individual. The T cells have a receptor capable of recognizing the pathogen or malignant cell the body wants to eliminate. T-cells multiply following this recognition. These events happen after an acute stimulus and continuously as the immune system interacts permanently with the external environment. Using stochastic population dynamics to match the TCR repertoire system to an ecological system and tools from Bayesian inference theory, we build and confront biophysical models to actual data. We take advantage of the availability of new data due to significant advances in high-throughput sequencing. Thanks to these methods, we can interpret data and the noise behind it, understand the TCR repertoire dynamics of people in the presence or absence of a strong stimulus on a short or long time scale, characterize the specificity of each TCR repertoire with time and develop tools that can be used by any quantitative immunologist interested in studying TCR repertoire dynamics.

KEYWORDS

immunology, Statistical biophysics, Inference, Population dynamics