Probabilistic approaches to the adaptive immune repertoire : a data-driven approach

Quentin Marcou

UNIVERSITÉ PARIS DESCARTES
ED474 "Frontières du Vivant"

Probabilistic approaches to the adaptive immune repertoire

A data-driven approach

Par Quentin Marcou

Thèse de doctorat de Biophysique
préparée au Laboratoire de Physique Théorique,
École Normale Supérieure

Présentée et soutenue publiquement le 28 septembre 2017

Composition du jury :

Benjamin CHAIN	rapporteur - University College London
Martin WEIGT	rapporteur - Université Paris 6 Pierre et Marie Curie
Anne-Florence BITBOL	examinatrice - Université Paris 6 Pierre et Marie Curie
Elizabeth MACINTYRE	examinatrice - Université Paris 5 Descartes
François AMBLARD	examinateur - Ulsan National Institute of Science and Technology
Aleksandra WALCZAK	directrice de thèse - École Normale Supérieure
Thierry MORA	co-encadrant - École Normale Supérieure

PROBABILISTIC APPROACHES TO THE ADAPTIVE IMMUNE REPERTOIRE

A data-driven approach

QUENTIN MARCOU
Laboratoire de Physique Théorique
École Normale Supérieure

September 28, 2017

Quentin Marcou: *Probabilistic approaches to the adaptive immune repertoire,* A data-driven approach, © September 28, 2017

À René, moustache érudite qui n'aura pas vu la fin de ce périple...

ABSTRACT

An individual's adaptive immune system needs to face repeated challenges of a constantly evolving environment with a virtually infinite number of threats. To achieve this task, the adaptive immune system relies on large diversity of B-cells and T-cells, each carrying a unique receptor specific to a small number of pathogens. These receptors are initially randomly built through the process of V(D)J recombination. This initial generated diversity is then narrowed down by a step of functional selection based on the receptors' folding properties and their ability to recognize self antigens. Upon recognition of a pathogen the B-cell will divide and its offsprings will undergo several rounds of successive somatic hypermutations and selection in an evolutionary process called affinity maturation.

This work presents principled probabilistic approaches to infer the probability distribution underlying the recombination and somatic hypermutation processes from high throughput sequencing data using IGoR - a flexible software developed throughout the course of this PhD. IGoR has been developed as a versatile research tool and can encode a variety of models of different biological complexity to allow researchers in the field to characterize evermore precisely immune receptor repertoires. To motivate this data-driven approach we demonstrate that IGoR outperforms existing tools in accuracy and estimate the sample sizes needed for reliable repertoire characterization. Finally, using obtained model predictions, we show potential applications of these methods by demonstrating that homozygous twins share T-cells through cord blood, that the public core of the T cell repertoire is formed in the pre-natal period and finally estimate naive T cell clone lifetimes in human.

RÉSUMÉ

Le système immunitaire de chaque individu doit faire face à des agressions répétées d'un environnement en constante évolution, constituant ainsi un nombre de menaces virtuellement infini. Afin de mener ce rôle à bien, le système immunitaire adaptatif s'appuie sur une énorme diversité de lymphocytes T et B. Chacune de ces cellules exhibe à sa surface un récepteur unique, créé aléatoirement via le processus de recombinaison V(D)J, et spécifique à un petit nombre de pathogènes seulement. La diversité initiale générée lors de ce processus de recombinaison est ensuite réduite par une étape de sélection fonctionnelle basée sur les propriétés de repliement du récepteur ainsi que ses capacités à interragir avec des proteines du soi. Pour les cellules B, cette diversité peut être à nouveau étendue après rencontre d'un pathogène lors du processus de maturation d'affinité durant lequel le récepteur subit des cycles successifs d'hypermutation et sélection.

Ces travaux présentent des approches probabilistes visant à inferrer les distributions de probabilités sous-tendant les processus de recombinaison et d'hypermutation à partir de données de séquençage haut débit. Ces approches ont donné naissance à IGoR, un logiciel polyvalent dont les performances dépassent celles des outils existants. En utilisant les modèles obtenus comme base, je présenterai comment ces derniers peuvent être utilisés afin d'étudier le vieillissement et évolution du répertoire immunitaire, la présence d'emprunte parentale lors de la recombinaison V(D)J ou encore pour démontrer que les jumeaux échangent des lymphocytes au cours de la vie fœtale.

PUBLICATIONS

[2] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. "Inferring processes underlying B-cell repertoire diversity." In: *Phil. Trans. R. Soc. B* 370.1676 (2015), p. 20140243.

[1] Yuval Elhanati, Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. "repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data." In: *Bioinformatics* 32.13 (2016), pp. 1943–1951.

[3] Quentin Marcou, Irit Carmi-Levy, Coline Trichot, Vassili Soumelis, Thierry Mora, and Aleksandra Walczak. "A qualitative model for the integration of conflicting exogenous and endogenous signals by dendritic cells." In: *bioRxiv* (2016), p. 065706.

[4] Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. "IGoR: a tool for high-throughput immune repertoire analysis." In: *arXiv preprint arXiv:1705.08246* (2017).

[5] Mikhail V Pogorelyy et al. "Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires." In: *PLoS Computational Biology* 13.7 (2017).

The IGoR software is available at `https://bitbucket.org/qmarcou/igor/`

REMERCIEMENTS

Je tiens tout d'abord à remercier mes encadrants Aleksandra Walczak et Thierry Mora (bien que ma phobie administrative ne lui aie pas officiellement accordé le titre de co-directeur), pour leur encadrement sur plus de trois années. Merci de m'avoir accepté initialement pour mon stage de master et de m'avoir introduit à l'univers de la biophysique et plus généralement de la physique théorique. Bien qu'initialement résolu à ne pas poursuivre en thèse vous aurez remis en question chacun de mes choix d'orientation, sans que je ne l'ai jamais regretté. Merci de votre confiance malgré mon profil atypique, de votre implication au quotidien et vos enseignements. Merci de m'avoir donné l'opportunité d'intégrer un milieu si stimulant et dynamique, d'assister à de nombreux cours, à Paris ou ailleurs, contribuant à réduire un peu mon ignorance. Peut être saurais-je même un jour ce que le mot "quantique" cache réellement. Même si à la sortie de cette thèse mon chemin s'éloigne à nouveau du parcours typique menant à la recherche académique, cette expérience m'aura convaincu que c'est ce à quoi j'aspire, et j'espère que dans les méandres de mon cheminement dans cette direction nos chemins se croiseront aussi souvent que possible.

Je tiens dans un second temps à remercier mes rapporteurs Martin Weigt et Benny Chain pour avoir accepté de laisser la lecture de ce manuscrit ternir leur été ensoleillé. Je remercie aussi Anne-Florence Bitbol et Elizabeth Macintyre d'avoir accepté d'être membres de mon jury. Enfin, je remercie plus spécialement François Amblard pour avoir accepté de faire partie de mon jury, d'avoir suivi mes travaux sur les trois années en compagnie de Martin Weigt, mais surtout pour ses enseignements, visions et conseils précieux bien que souvent déroutants.

Car les travaux présentés dans cet ouvrage sont loin d'être le fruit d'un travail solitaire je tiens à remercier l'ensemble de mes collaborateurs. De l'autre côté de l'atlantique Curt Callan Jr et Zachary Sethna. De l'autre côté de l'Oural Misha Pogorelyy pour sa créativité et pour m'avoir fait découvrir le spectre insoupçonné des textures de pommes. Un merci tout particulier à Yuval Elhanati, mon compagnon dans la malédiction des données adversariales et autres bugs en tout genre, pour avoir patiemment répondu à mes innombrables questions et m'avoir tant appris.

Durant ces trois années j'ai eu la chance d'appartenir à un large groupe de personnalités hétéroclites et j'aimerai donc remercier dans le désordre: Andreas, mon compagnon de thèse, pour sa capacité à étendre le temps et à son coup de fourchette qui aura souvent prolongé nos pauses café et déjeuner de plusieurs heures, Jonathan pour son engouement permanent, son rire si communicatif, nos tête a tête californiens et nos projets révolutionnaires, Rhys pour son humour si subtil, Max pour être le visage souriant au dessus de mon écran, son attitude calme et toujours positive, avec qui j'aurai maintes fois disséqué et refait le monde avant l'avènement des machines. Merci à Paulina, malgré

nos désaccords sur la pratique du ski, pour son énergie et nos séances de psychanalyse. Bien que sa présence dans mon bureau coïncide probablement avec un gouffre de productivité, merci à Dario pour ses conseils de vie, son flot de paroles intarissable, et son idéal du rap français en Ferrari. Merci à Christophe qui aura enduré maintes fois la faim et dont les invitations maudites auront malheureusement contribué à sa rareté. Merci à Huy l'homme orchestre et ses histoires extravagantes et Jacopo que je n'aurai cotoyé que trop brièvement.

Merci à Viviane Sebille et Sandrine Patachhini qui par leur implication et leur efficacité auront réussi à rendre la bureaucratie un peu moins effrayante.

Merci à l'ensemble des enseignants qui auront accepté ma présence à leur cours pour petit à petit combler mes lacunes.

Merci à l'ensemble des personnes du département qui auront rendu les déjeuner et ballades dans les couloirs plus attrayant Alice, Marco, Fabio, Lorenzo, Elizabeta, Marylou et enfin merci à Louis pour sa passion du croustillant et son amour du thé.

Merci au programme Frontières du Vivant pour être un peu plus qu'une coquille administrative. Merci à Paul, Lise, Marion, Bérengère, Frances, Carlos et Jean pour leur compagnie de la grisaille parisienne au ciel d'azur.

Le parcours non linéaire me menant à écrire ces lignes est le fruit d'une longue liste de rencontres qui m'auront guidées pas à pas dans mon cheminement personnel. Je tiens donc à remercier Pascal Silberzan et l'ensemble de son équipe, Olivier Cochet, Simon Garcia , Maxime Deforet, Hanna Yevick et Guillaume Duclos pour m'avoir si chaleureusement accueilli pour ma première expérience de recherche malgré mon ignorance complète. Bien qu'ayant lentement dérivé vers des disciplines plus théoriques, je n'aurai pu espérer meilleur environnement et cette expérience aura été décisive sur mes choix futurs. Car sans leur vision désintéressée, cette thèse n'existerait pas, je remercie l'ensemble des enseignants du cursus Médecine-Sciences et plus particulièrement Jean-Claude Chottard et Philippe Ascher qui se seront battus corps et âmes sur des décennies pour la création et maintient de ces doubles cursus. Je ne saurait trop remercier M. Abehsira, un enseignant passionné et passionnant qui, peut être à son insu, aura su éveiller en moi la passion des mathématiques et paradoxalement de la physique. Sa fièvre pour sa discipline a rallumé la curiosité sans laquelle probablement aucune de ces rencontres ne se seraient produites.

Pour leur compagnie sur ces trois années, merci à mes collocataires Nicolas, Pierre, Thomas, Benjamin et Émilien pour nos dynamiques quasi-burlesque qui auront rendues mes soirées moins mornes et la décontraction (trop?) facile. Merci à Timothée qui aura partagé la plupart de mes vacances et qui malgré mes nombreux refus continue à me proposer des sommets. Merci à Sixtine qui, parfois sans le réaliser, m'aura soutenu dans les mètres les plus difficiles. Parce que la liste détaillée serait peut être trop longue pour ce manuscrit, merci à tous les autres qui auront été là au cours de ces années. Merci à ma famille que je n'aurai pas beaucoup vu et particulièrement à ma grand mère gardienne de mon lieu d'ermitage où ces travaux ont été écrits.

Enfin je voudrais remercier tous les acteurs du monde du libre sans lesquels j'entretiendrai moins de débats idéologiques enflammés et sans qui les travaux de cette thèse seraient loin d'être ce qu'ils sont.

CONTENTS

BIOLOGICAL ACRONYMS

AID	Activation Induced cytidine Deaminase ... 20
APC	Antigen Presenting Cell ... 11
BCR	B cell receptor ... xv
BER	Base excision repair ... 98
BNAb	broadly neutralizing antibody ... 18
CD	Cluster of differentiation ... 16
CDR	Complementarity-Determining Region ... 13
FACS	Fluorescence-activated cell sorting ... 25
FR	Framework Region ... 13
HIV	Human Immunodeficiency Virus ... 18
HLA	Human leukocyte antigen ... 21
HSC	Hematopoietic Stem Cell ... 10
Ig	Immunoglobulin ... 13
MALT	mucosa-associated lymphoid tissue ... 10
MHC	Major Histocompatibility Complexes ... 21
MMR	Mismatch repair ... 98
NGS	Next Generation Sequencing ... 25
PAMP	pathogen-associated molecular pattern ... 10
PCR	Polymerase Chain Reaction ... 25
PRR	Pattern Recognition Receptor ... 10
RSS	Recombination Signal Sequence ... 14
SHM	Somatic Hypermutation ... 19
SNP	Single Nucleotide Variant ... 46
TCR	T cell receptor ... 11
TdT	Terminal deoxynucleotidyl transferase ... 15
UMI	Unique Molecular Identifier ... 26

MATHEMATICAL ACRONYMS

EM	Expectation-Maximization ... 29
HMM	Hidden Markov Model ... 37
MAP	Maximum a posteriori ... 29
ML	Maximum likelihood ... 29

Part I

INTRODUCTION

INTRODUCTION

Entanglement of mathematics, physics and biology (or medicine) dates from immemorial time. Eminent figures of this multidisciplinarity are embodied by Middle-Age Arab physicians such as Avicenna or Averroes. Throughout the ages many scientists have routinely crossed barriers imposed by these disciplines, revolutionizing one field for the purpose of another. Isaac Newton invented calculus to solve the equations derived from the laws of motion. Ronald Fisher constructed modern statistics concurrently with population genetics. However, it is only recently that such disciplinary wanderings have been baptized, as *inderdisciplinary science*.

Physicists have been puzzled by the functioning of biological systems for a long time. In his short essay *What is life?* [151], Schrödinger wondered how living systems can remain out of equilibrium and reliably function despite small number and thermal fluctuations. Due to the lack of suitable measurements, his considerations remained however very abstract. During the last decades the appearance of quantitative methods in biology, and biologists' zeal to dissect and describe every actor of life at various scales have depicted formidably complex interacting systems. These investigations have produced results that appealed to physicist eager to understand the quantitative laws governing biological systems. This large endeavor of physicist and the growing interest of biologists for quantitative predictive theories have consolidated the bridge between physics and biology that is now known as *biophysics*.

More recently, the growing throughput of quantitative methods have enabled to probe complex biological systems. The adaptive immune system of jawed vertebrates, the subject of study for the work presented in this manuscript, is one of these systems.

Across his lifetime an individual will face repeated challenges from a virtually infinite number of threats or pathogens. Because these threats are so diverse and evolve faster than the lifetime of the individual, providing a specific defense against each is a tour de force that the adaptive immune system manages to achieve. In order to be able to fight specifically each pathogen, the adaptive immune system relies on a set of tremendously diverse antigen receptors carried by T and B-cells. Each T or B-cell carries one receptor able to recognize specifically a small number of antigens. This diversity is initially generated through a stochastic germline DNA editing process called V(D)J recombination. Because of the stochastic nature of this process, receptors that are reactive against the host can emerge. In order to avoid self destruction, this initially created diversity is then narrowed down by a step of functional selection against recognition of the individual's self-proteins. The ensemble of receptors that has passed this selection is called the repertoire. Upon encounter of their cognate antigen some of these random receptors will be further randomly edited by somatic hypermutations and evolve towards larger affinity for the targeted antigen in a process called affinity maturation.

Because of its complexity and stochastic foundations, its connections with population genetics and epidemiology, theoretical immunology is an active field that has drawn interest of many physicist. Some of them have been appealed by theoretical considerations such as the fraction of antigenic environment an immune cell can react to [126], the existence of idiotypic networks[1] [127], the optimal organization for an adaptive immune system [104], knowing whether an immune system organization achieves optimal performance for its environment [105] or immune systems links with defenses of computer networks [75]. Others were appealed by more applied considerations aiming at building descriptive models whose predictions could help cure diseases such as HIV [125, 184].

The work presented in this manuscript belongs to this second class and aims at designing a general statistical framework to describe the recombination, selection and hypermutation processes. The empirical use of vaccination [53], monoclonal antibody treatments [136] and more recently cancer immunotherapy [152] are already successful clinical achievements. However, such techniques are only using fractions of the immune system's capabilities, the full understanding of the adaptive immune system formation and dynamics remains a cornerstone for personalized medicine. The advent of high throughput repertoire sequencing providing a snapshot of an individual's adaptive immune system, promises to revolutionize personalized medicine by providing new statistical diagnostic tools for biology and medicine. The state of one's repertoire could be used to infer an individual's past and present immune challenges, and their susceptibility to future infections or diseases. However, because of the adaptive immune system's formidable complexity and stochastic nature interpreting this data is challenging and should rely on the understanding of the rules governing the system.

What should be the scale for these rules? Shall we model the recombination machinery along with all its molecular constituents and dynamics to capture an individual's repertoire statistics? In his 1972 paper *More is different* [4], Anderson argued that one does not need to model physical systems from their most fundamental constituents because, climbing the ladder of complexity, some microscopic details will become irrelevant as one see the emergence of new macroscopic properties. His argument mostly relies on the irrelevance of such an approach, while some more fundamental ideas [37, 67] would suggest that building biology from fundamental physical constituent might simply be doomed to fail.

At the other extreme of the constructionist scale lies machine learning. The past years have seen the explosion of computational power and an ever growing amount of data. Conjugated with the advent of deep neural networks [90] some have been tempted to call it the end of theory and the scientific method altogether [3]. Such techniques have recently been used in physical and biological systems [34, 202] to investigate hard problems such as many body local-

1 Because a large variety of receptors are created and cannot be tested as belonging to the self it was thought that two receptors could bind to each other. Such binding would provide an "antigenic" stimulation and all these interactions would regulate the composition of lymphocyte populations.

ization [29], detecting phases of matter and their associated phase transitions [30, 123], or inference of selection in population genetics [158]. Despite their very strong predictive power the actual features learned by these methods are not understood and might simply not correspond to a sensible representation of the object to characterize [68, 122].

I believe that taken together, the limitations of the two approaches justify the intermediate data-driven approach we adopt in the work presented in this manuscript. Provided current biological knowledge we will build simplified and interpretable statistical models to infer V(D)J recombination and hypermutation rules, and increase their complexity only when they do not recapitulate correctly some data statistics. In general, we wish to delineate which traits are universal or individual specific to understand whether the differences of efficiency of different individuals' immune systems can be attributed to physical parameters or stochastic fluctuations.

The rest of the dissertation is organized as follows:

- Chapter 1 introduces the functioning of the immune system in vertebrates with strong emphasis on the adaptive immune system. The current knowledge about T and B-cell roles and their interactions, the V(D)J recombination process, and initial functional selection will be summarized before introducing modern sequencing techniques that will constitute the basis of our modeling work.

- The following chapter, Chapter 2, introduces the mathematical tools and concepts that are used or useful to understand the work presented in the manuscript. The end of the chapter presents the challenges and achievements of repertoire sequencing analysis along with the already existing bioinformatic tools my work relates to.

- Chapter 3 presents a probabilistic assignment approach to characterize and infer V(D)J recombination rules. Because different types of data might exhibit different peculiarities we made our general method available through IGoR a versatile software tool.

- Chapter 4 introduces how from these models of V(D)J recombination we can extract information about an individual's haplotype and estimate the recombination rescue probability.

- In chapter 5 I will show how, combining these models with models of somatic selection, we tackle the notion of shared or public receptors, and how our data-driven approach led us to demonstrate that twins exchange immune receptors in utero and that such early immune cells are long lived.

- Chapter 6 presents work aiming at statistically describing the somatic hypermutation process. After reviewing current knowledge and work, I will present an independent site targeting model and its shortcomings. Using our probabilistic assignment approach I will show that hypermutation cluster and call for better models.

- Finally, in Part iv I will summarize our findings and propose future research directions.

Part II

BACKGROUND, CONCEPTS AND METHODS.

In this part of the manuscript I will introduce notions of immunology and mathematical tools that will be used in the following parts.

The first chapter presents an overiew of the immune system with a strong emphasis on the adaptive immune system.

The second chapter presents an overview of the methods on which my work was built.

1 AN OVERVIEW OF THE ADAPTIVE IMMUNE SYSTEM

1.1 INTRODUCTION

The birth of *immunology* (from Latin *immunis*, meaning exempt) is often attributed to Edward Jenner coining the term *vaccination* for the inoculation of cowpox (vaccinia) as a protection for smallpox in 1796. However ideas about non-mystical disease mechanisms and natural or acquired immunity are documented since ancient Greece. It is remarkable how tight the relationship between smallpox and immunology is. Indeed, the Hippocratic school with the general humor theory described diseases as a quantitative imbalance among humors (blood, yellow bile, black bile and phlegm). Galen of Pergamon further introduced the notion of possible qualitative changes in these humors, such that smallpox was for long described as the result of blood fermentation [161]. Later, in a treaty on smallpox and measles [138] the Arab physician Abu Bekr Mohammed ibn Zakariya al-Razi (880–932 AD, also known as Rhazes) along with a precise description of the disease expressed his belief in the existence of a long lasting acquired immunity to smallpox. He proposed that smallpox was due to an excess moisture of the blood that would be expelled through the fluid contained in the pustules. Thus in agreement with humor theory, stating that the blood dries with age, only young people could suffer from the disease and immunity would be acquired with age or upon previous contraction of the disease. While those descriptions depict diseases as an individual purely internal dysregulation, in 1546 Girolamo Fracastoro proposed that the disease would be transmitted through small seeds (*seminaria*) and would be transmissible from a person to another [161]. All those seeds would have particular affinity and would in turn only affect subsets of animals and plants, thus providing a first basis for natural (or innate) immunity.

While these tentative theories are conceptually interesting they did not provide much insights (or at least accurate ones) into possible ways of treating diseases. At the same time in Asia and Middle East more practical solutions were used and inoculation of ground smallpox pustules was used to prevent future infection. In the early 18th century such practices were brought to the attention of Western medicine, and before risking these procedures on noble's children, led to the conduction on prisoners and orphans of the first immunological clinical trial. Extending this procedure, Edward Jenner using cowpox for protecting against smallpox proved the existence of cross-immunity.

The rise of modern bacteriology in the 1870s, with prominent figures such as Louis Pasteur and Robert Koch, provided the etiologic agents responsible for diseases and enabled in vitro and in vivo experimentations. From then the field of immunology, along with pathology, bacteriology and medicine, started blossoming. Over the last hundred years, much progress has been made and countless names would need to be cited to reach the evermore precise and rich

view we have nowadays. Our vision of the role has widened such that we now talk about antigen to merely be any substance potentially recognized by the immune system such as proteins, polysaccharides or even metals [114].

Despite this, much remains to learn about the vertebrate adaptive immune system concerning its generation and dynamics as a whole system, problems that we partially try to address in this work. In the remainder of this chapter I will try to give a brief overview of our current knowledge with particular emphasis on concepts relevant to understand the framework and results presented throughout this manuscript.

1.2 INNATE AND ADAPTIVE IMMUNE SYSTEMS

Vertebrate immune systems are classically described in two parts: the innate immune system and the adaptive immune system. Both systems' responses depend upon the activities of white blood cells or leukocytes, and while the inclusion in either one of different cell types or effectors might be fuzzy, most cells actively participating in the immune system derive from the same pluripotent progenitor Hematopoietic Stem Cell (HSC), develop and then potentially mature in the bone marrow. [1] Once mature, those cells migrate through the blood and a dedicated transport system called the lymphatic system. The lymphatic system drains the extra cellular fluid along with immune cells from tissues, forming lymph, towards peripheral lymphoid organs and eventually back to the vascular system. These lymphoid organs comprising the spleen, lymph nodes and mucosa-associated lymphoid tissues (MALTs) [2] are the site of the adaptive immune system's activation by the innate immune system.

The innate immune system is responsible for natural immunity. Although comprising cells whose primary role is not immune defense, the first line of the innate immune system are anatomical and chemical barriers. Those epithelial barriers, such as skin or gut lumen or even the brain blood barrier [3] , must be first breached by foreign pathogens in order to harm the individual and potentially trigger an immunological response. Once breached, pathogens will face sentinel cells dedicated to the innate immune system such as dendritic cells, macrophages and neutrophils. These censor cells will initiate the immune response through secretion of inflammatory [4] mediators (or cytokines) and chemo-attractants (or chemokines) upon recognition of pathogenic threats through a limited set of invariant innate recognition receptors. Those Pattern Recognition Receptors (PRRs) target common pathogenic signal known as pathogen-associated molecular patterns (PAMPs), such as lipopolysaccharides (LPS) contained in bacterial membranes or byproducts of pathogenic damages

1 Some effector cells of the innate immune system, such as the tissue-resident macrophages in the brain (microglia) are generated during embryonic life from the yolk sack or fetal liver

2 MALTs are immunological structures sitting directly at mucosal anatomical barriers. Well studied examples in the gut are Peyer's patches, however similar structures can be found in e.g the nasal and bronchus mucosa.

3 The brain blood barrier is so tight and selective that even antibodies and most antibiotics cannot pass it, making brain infections extremely severe. In order to enable brain infection clearance brain specific immune cells such as microglia reside there from early development.

4 Inflammation is the result of an increased blood vessel epithelium permeability, leading to a net flow of fluid, proteins and cells from blood to the extra-cellular medium.

such as ATP in the extra cellular medium. The extent of the innate immune system activation is a balance between pro and anti inflammatory signals [102], leading individual cells to secrete mediators damaging pathogens or engage in more direct actions such as phagocytosis. While some cells such as neutrophils phagocyte pathogens simply for pathogen extermination, Antigen Presenting Cells (APCs) comprising mostly macrophages and dendritic cells serve other purposes. Indeed, after phagocytosis (or macropinocytosis) activated APCs migrate towards the lymphoid organs where they relay infection information and activate the adaptive immune system. [5]

The adaptive immune system comprises antigen-specific lymphocytes, namely B and T lymphocytes. Both derive from the lymphoid lineage a differentiation of HSCs. Contrary to innate immune cells each B and T lymphocyte carry one specific receptor [6] whose sequence is not contained in the individual germinal DNA and is randomly produced through the process of V(D)J recombination. BCRs are formed by the same genes that encode antibodies as shall be explained in later sections [7] , while T cell receptors (TCRs) have a slightly different structure and function. After the long journey from precursor lymphoid cell to functional naive lymphocyte, each lymphocyte divides according to some external stimulus as described by clonal selection theory [27]. Because the gene rearrangement process irreversibly edits the lymphocyte's DNA, all its progeny inherit the same receptor. The ensemble of lymphocytes deriving from the same ancestor, carrying the same random receptor[8], is called a clone - a concept that we shall discuss at length in the next sections and later in chapter 5. Young mature lymphocytes continually recirculate between peripheral lymphoid organs to which pathogenic antigens are brought by APCs. Those lymphocytes that have not yet been activated by one of their cognate antigen are known as naive lymphocytes; those that have met their antigen, after a proliferation step, differentiate further into fully functional effector lymphocytes. A unique feature of the adaptive immune system is its capacity of generating immunological memory, so that having been exposed once to an infectious agent, an individual will make a faster and stronger response against any subsequent exposure to it.

T and B-cells specificities are presented in sections 1.5 p.21 and 1.4 p.18 respectively.

The processes through which each receptor is randomly created, matured and selected are described in sections 1.3.2 and 1.6.

This short presentation emphasizes the role of the innate immune system as a trigger for the adaptive one thus hiding many roles of the former and the formidable mesh of interactions between these two systems. Indeed the innate immune system can to some extent clear pathogens independently, but is also a major downstream effector of the adaptive response for which the adaptive immune system acts as a targeting aid. Those two systems act in complete synergy and the boundary in between them is fuzzy. The work presented in this manuscript dealing exclusively with the adaptive immune system, the pur-

Some examples of effector of the innate immune system roles will be given in the next sections.

5 Note that I have left out of this description numerous actors of the innate immune system such as natural killers lymphocytes, innate lymphoid cells, basophils, eosinophils, mast cells and finally the complement system that I will briefly present in section 1.4.3 p.19.

6 Each lymphocyte carries only one sort of random receptor but carries many copies of it on its membrane.

7 BCRs are thus sometimes referenced as membrane or surface immunoglobulins, however for the sake of simplicity we will not use these denominations in this manuscript.

8 Or a slight variation of it for hypermutated BCRs

pose of this emphasis is to pass one message to the reader: without the innate immune system the adaptive immune system's response cannot be elicited. This was realized during the first half of the 20^{th} century, by observing that purified antigens were not sufficient to elicit a specific immune response and that adding bacterial materials as adjuvant[9] (or "helper") enabled this response [45, 55]. Along with this concept I will finally insist on four major differences between these systems. First their timescales of action: the innate adaptive immune system gets activated and acts within hours, while the time needed for an APC to migrate (or a soluble antigen to diffuse) from the afferent lymphatic to the lymph node, activate lymphocytes, and the activated lymphocytes migrate through the efferent lymphatic to the blood and then towards the site of infection is 4-6 days. Second, as previously mentioned the adaptive immune system relies on a large diversity of receptors, each specific to a few antigens, while the innate immune system relies on a few receptors targeting generic pathogenic signatures. Lastly, with the ability to produce new receptors[10] and as mentioned earlier the ability to form an immunological memory the adaptive immune system can be trained while the innate immune system is static.

1.3 ADAPTIVE IMMUNE SYSTEM RECEPTORS

As soluble molecules antibodies were easier to study than membrane bound receptors that are BCRs and TCRs. Their architecture being extremely similar (appart for slight specific differences outlined in the next section) detailing their structure will give the reader a good understanding of how those membrane receptors work. This section focuses on commonly described human and murine immune receptors, however in nature exception is the rule and, through the course of evolution, other species such as camelids or sharks have acquired slight variations that will not be discussed in this manuscript.

1.3.1 *General structure*

TCRs and BCRs (or antibodies) are composed of respectively one or two heterodimers formed by two polypeptidic chains: one of lesser diversity (respectively α or light chain) and one of greater diversity (respectively β and heavy chain). Each of these chains are independent random products of the germline DNA editing process of V(D)J recombination that I present in section 1.3.2. Each chain contains a constant region (C_α, C_L, C_β, C_H) where the disulfide bonds necessary to assemble the heterodimer will be formed (Fig. 1.1). These constant regions are regions anchoring TCRs and BCRs to the cellular membrane and take on a functional importance for different antibody classes as explained in section 1.4.5. Each chain also contains a variable region (not surprisingly respectively named V_α, V_L, V_β and V_H). The variability created by the recombination process is however not constant over the full variable region. While the peptide variability remains rather low in most regions, most likely due to fold-

9 Finding good adjuvants (such as aluminium salts) is still a challenge for vaccination.
10 Although this production goes down with age.

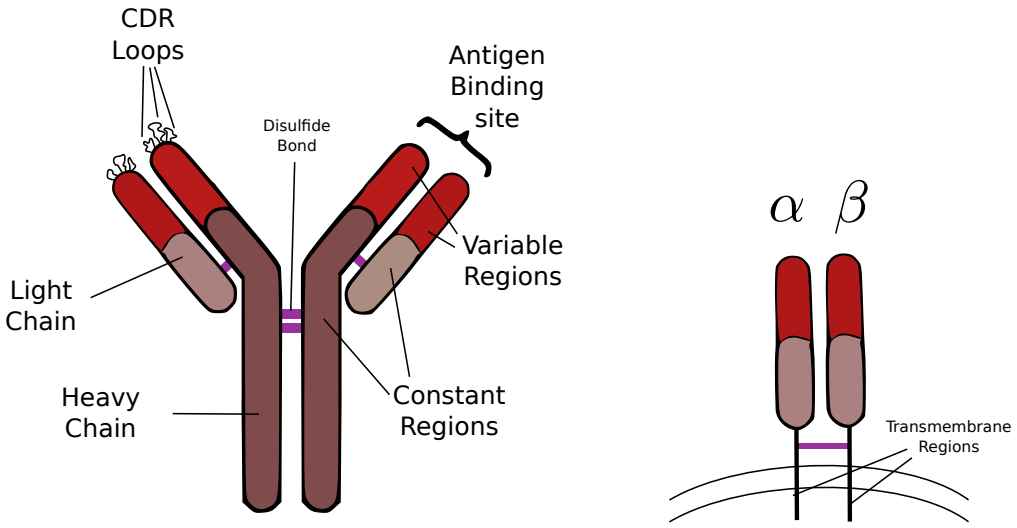

Figure 1.1: **Schematic representation of antibodies and TCRs.** The antibody (left) is a homodimer of heterodimers connected through disulfide bonds (purple). TCRs (right) are smaller heterodimers anchored in the cell membrane.

ing and stability constraints, a few regions (3 in most cases), spanning less than tens of residues exhibit high variability. From X-Ray crystallography the low variability or Framework Regions (FRs) are known to each form a beta sheet, together assembling in a beta sandwich. The three hypervariable regions or Complementarity-Determining Regions (CDRs) (CDR1, CDR2 and CDR3) form free flexible neighboring loops connecting the beta strands. The six flexible hypervariable structures of the two chains are neighboring in the heterodimer and together form the antigen binding site (see Fig. 1.1). This antigen binding site binds specifically subparts of one or a few antigens. Each antigen subpart recognized by a TCR, a BCR or an antibody is called an epitope. The multiplicity of possibly recognized epitopes is referred to as cross-reactivity.

B and T-cells receptors recognize different types of epitopes. Their specificities are detailed in sections 1.4.1 p.18 and 1.5.1 p.21.

As mentioned earlier BCRs and antibodies are homodimers formed of two copies of a heavy and light chain heterodimer. The heavy chain is much larger than the light[11], due to a large constant region. Disulfide bonds are formed between the heavy chains' constant regions and form the homodimer. The end product is a Y (Fig. 1.1) shaped protein with three globular regions of comparable sizes. Arms of the Y (containing the light chain) are flexible and bind to antigens[12]. The trunk of the Y has functional importance to define the Ig class and thus the role of the antibody as well as the BCR anchor point in the B-cell's membrane.

See section 1.4.5 for Ig classes and roles

TCRs on the other hand are composed of only one α :β heterodimer. Those two chain have roughly the same size[13] such that the heterodimer much ressembles an arm of the Y shaped Ig. Slight variation of the folding makes the antigen

11 50kDa against 25kDa

12 These flexible arms allow the Immunoglobulin (Ig) to bind more efficiently antigens, and for several soluble Igs to bind the same antigen with less steric constrains forming structure called haptens.

13 Around 40kDa

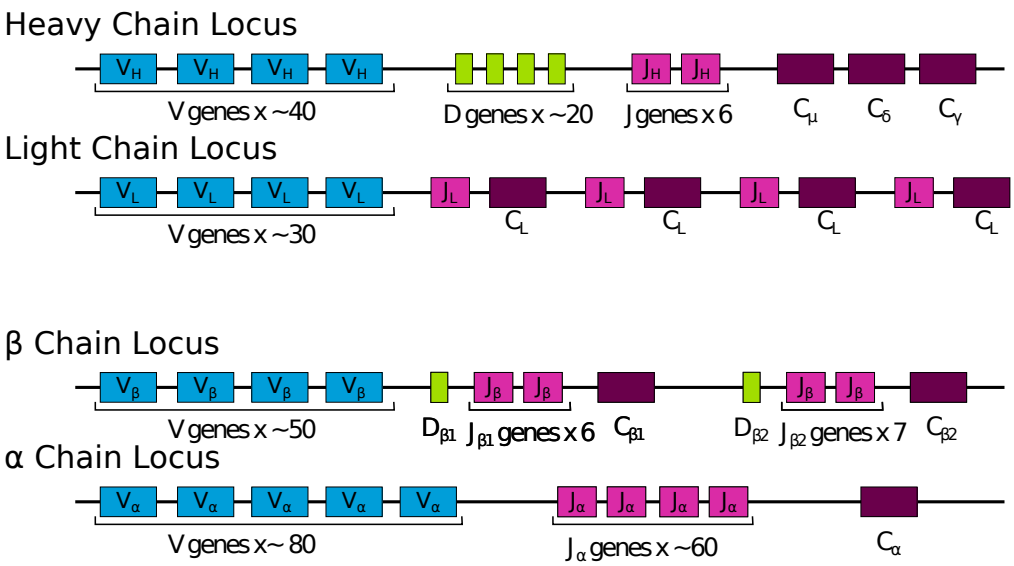

Figure 1.2: **Recombination locus organization for the different chains.**

binding end hypervariable loops less flexible than its Ig counter part which is of functional importance.

1.3.2 *V(D)J recombination*

As previously mentioned adaptive immune receptors are not directly encoded in the genome but arise from the stochastic germline DNA editing process of V(D)J recombination. This process involves recombination of three or four gene families called: V for *Variability* genes, D for *Diversity* genes [14] , J for *Joining* genes and C for *Constant* genes. For V, D and J gene classes several genes are initially present and a fully recombined receptor is formed by a combination of one of each of these genes. The V gene choice fully determines the CDR1 and CDR2 regions while the most hypervariable CDR3 region is encoded by the combination of a V, a D (for heavy and β) and J gene. The C gene encodes the constant region of the receptor. Several C genes might be present and can carry different functions for the receptor (see section 1.4.5 for Igs example).

1.3.2.1 *General recombination mechanism*

In order to ensure the recombination of a V(D)J triplet, DNA rearrangements are guided by conserved noncoding DNA sequences, called Recombination Signal Sequences (RSSs), that are adjacent to the recombination spots. At each recombination spot (at the beginning or end of a V, D or J gene), two RSS are present: a nonamer and a heptamer and are separated by a variable spacer sequence of either 12 or 23 base pairs[15] . The length of the spacer sequence determines whether two recombination spots can be joined, such that RSSs

14 D genes are only present in heavy and β chains and are responsible for the extra diversity encoded in those chains.

15 12 nucleotides correspond two one DNA double helix turn while 23 corresponds to two turns

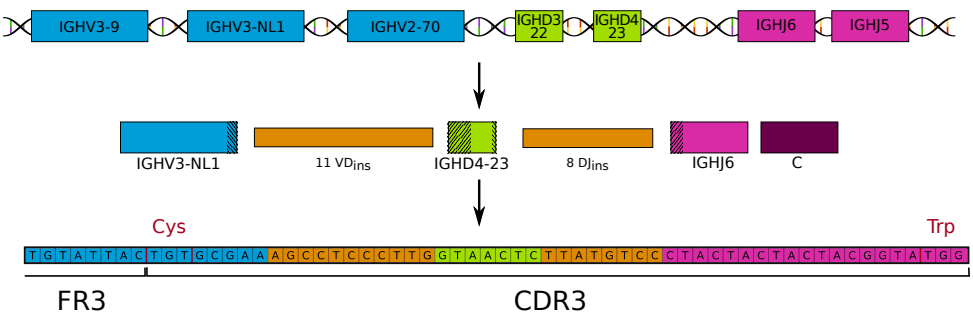

Figure 1.3: **V(D)J recombination in Ig heavy chain.** V(D)J recombination proceeds by joining randomly selected segments (V, D, and J segments in the case of IGH). Each segments gets trimmed at its ends (hashed areas), and a varying number of nontemplated insertions are added between them (orange). The resulting sequence encodes the CDR3 region contained between conserved Cystein and Tryptophan.

with a 12bp spacer can only recombine with a 23bp one. This is known as the 12/23 rule.

A violation of this rule for BCRs will be discussed in section 3.7

During the recombination process the V(D)J recombinase enzyme complex [16] first brings two corresponding RSSs together forming a DNA loop and excise the RSSs and all remaining genes potentially contained in the loop. This creates two DNA ends with double strand break. At each of these ends a DNA hairpin is then created by joining the two last complementary base pairs. Once the hairpins created the Artemis enzyme will create a single strand DNA break at a random position, thus creating a palindromic sequence on one of the strands. These extra palindromic insertions are called P-nucleotides.

Although the precise mechanism is unknown, Artemis is also thought to have an exonuclease activity and delete a random number of nucleotides on the free single strand sequence possibly further than the P nucleotide range thus effectively deleting base pairs somatically encoded in the genomic template. This mechanism is supported by the fact that P nucleotides are not observed together with deletions [115]. These deletions constitute an extra source of diversity for the resulting receptor sequence.

Contribution to the diversity of each recombination element is discussed in section 3.4

Finally, random non templated insertions (also called N nucleotides) can be added on each free single strand end by the Terminal deoxynucleotidyl transferase (TdT) enzyme. The two resulting single strand sequences will then join upon a sequence of complementary nucleotides. Extra nucleotides beyond the paired region are then excised and remaining single stranded regions are complemented to form a double stranded DNA molecule.

1.3.2.2 *Heavy and β chain recombination*

Lymphocytes arise from progenitors cells with intact chromosomes and evolve through many intermediary maturation steps until reaching the mature lymphocyte state. Those steps are well documented and well characterized by the

16 The V(D)J recombinase complex contains lymphocyte specific enzymes RAG-1 and 2 and a number of ubiquitous proteins involved in DNA double strand break repair.

presence or absence of a number of cell surface proteins called Cluster of differentiations (CDs) that will be overlooked here. These proteins serve as a general classification system for immunophenotyping.

Both B and T-cells start by recombining the DJ junction of their large diversity chain. At this stage the recombination is thought to occur on both chromosomes at the same time. Once the DJ junction recombined, the VD junction of only one of the two chromosomes is recombined, while the other stay untouched. This is known as allelic exclusion. Currently, it is unknown how allelic exclusion and the recombination timing of DJ and VD recombinations are imposed.

Once the full chain recombined, it is mounted on a surface receptor with a surrogate lesser diversity chain somatically encoded in the genome. These are called pre-BCR and pre-TCR for B and T-cells respectively. At this point no ligand binding testing is carried out, and dimerization of those receptors on the cell surface will carry an intracellular message to suspend expression of the V(D)J recombinase and carry on the development further. This step is thus simply testing for the ability of the recombined chain to fold and interact with a templated lesser diversity chain.

Transduction of the dimerization signal will stop the recombination process and stimulate the pre-lymphocyte proliferation. However chances of obtaining a non functional chain are high, for instance a frame shift in the CDR3 region occurs $\sim 2/3$ of recombination attempts. The expression of the recombination enzyme complex will then trigger the recombination of the VD junction of the so far untouched chromosome[17]. Upon failure of the second recombination the lymphocyte engages in apoptosis.

Estimating the frequency of this rescue mechanism is the topic of section 4.2

1.3.2.3 *Light and α chain recombination*

After the replication step each lymphocyte with an identical β or heavy chain will recombine the second lower diversity chain separately, and will in turn obtain a different receptor.

Due to the absence of the D gene cluster, recombination of the second chain is a one step process joining the V and J region. While light chains exhibit allelic exclusion and recombine one chromosome at a time, alpha chains do not and are known to recombine both loci concomitantly. Due to their similar organization, both α and light chains can recombine several times on the same loci. Indeed, upon the joining of a V and a J gene all genes in between will be excised. However all 5′ most V and 3′ most J genes are still intact and can recombine thus excising the previously joined VJ couple.

As for the recombination timing, the recombination stopping criterion is also thought to be different between B and T-cells. Upon recombination of the light chain, the full BCR receptor is expressed on the B-cell surface. As for pre-BCR, the newly recombined BCR provides a ligand independent tonic basal signal

17 Actually due to the organization of the TCRβ locus, failed β recombinations can sometimes get a second recombination attempt. Indeed, if the first recombination involved the $D_{\beta 1}$-$J_{\beta 1}$ cluster, the second one with its own constant region has not been excised and can still recombine with a V gene upstream of the one involved in the first recombination that has not been excised during the VD joining.

indicating that a functional protein can be expressed on the cell surface. This signal temporarily prevents further recombination of the light chain locus. This process remains independent of any ligand binding solely assessing folding and assembling capabilities of the resulting BCR. Further recombination of the light chain, called receptor editing, could be re-induced if the receptor is found to be reactive to self antigens as described in section 1.6.1. Such tonic signaling has not been described for T-cells, such that it is thought that both α chains loci will keep recombining until a positive selection signal based on ligand binding capabilities is delivered (see section 1.6.2). In practice T-cells could exhibit two different receptors with identical β and different α chains, however it is unlikely that both have a functional role[18].

1.3.2.4 *Non productive sequences*

There exist many reasons why a recombined sequence may be non-functional (beyond ligand binding capabilities) meaning producing a correctly folded receptor. The protein could be a truncated protein, or use a gene segment known to produce non-functional rearrangements or even exhibit an amino acid in the CDR3 region destabilizing the full protein structure. This ability is tested by producing pre-lymphocyte receptors after β or heavy chain recombination, or by integration of BCR tonic signals for light chain functional testing. However mapping the sequence to a folding state for a protein is a hard problem and we are therefore incapable of predicting if a given sequence produces a functional protein. Still, for a few obvious cases we are able to call sequences non-productive. Because it reflects almost entirely the recombination product, the CDR3 region is determinant in assessing the productivity of a sequence. This CDR3 is generally defined between a conserved cystein in the V gene and a conserved phenylalanin, for TCRs, or tryptophan, for BCRs, in the J gene. In the rest of this manuscript we will consider all sequences with a frameshift[19] or a stop codon in the CDR3 junction to be non productive or non coding. We will denote the ensemble of the remaining sequences as productive sequences.

We will assume non productive sequences to be non functional, and thus not contributing to clonal selection. In order to pass initial selection, each mature lymphocyte must carry at least one functional receptor. Thus all non functional sequences that can be observed through immune repertoire sequencing are subject to random selection due to the necessary functional sequence carried on the second chromosome. In chapters 3 and 6 we will use these non selected non productive sequences to build models capturing the V(D)J recombination and the somatic hypermutation process statistics.

On the other hand productive sequences are not necessarily functional and trying to predict functionality is a general problem addressed briefly mentioned in chapter 5. The term coding instead of productive might thus sometime be preferred to lift this ambiguity. Still, a majority of productive sequences

18 Having two receptors with a functional role would first require that both can be expressed and form a surface receptor, one of the two to be kept by positive selection and finally none of them to be deleted by negative selection.

19 Rearrangements whose CDR3 regions contains a number of nucleotides that is not multiple of 3.

in repertoire sequencing data should be functional since each sampled mature lymphocyte must carry at least one functional receptor. This subject will be discussed in more details in section 4.2.

1.4 B-CELLS, SURFACE RECEPTORS AND ANTIBODIES

From the lymphoid progenitor to the mature lymphocyte state B-cell stay and develop in the bone marrow. They will only leave it once they complete maturation and selection against recognition of self-antigens, as detailed in section 1.6 p.23. This section briefly presents how their receptors bind to ligands and their role in the immune response.

1.4.1 *BCRs and antibodies ligands'*

BCRs and antibodies generally recognize small molecules such as vitamins, or only a small region of the exterior surface of a large molecule such as a polysaccharide or protein. These molecules can be freely diffusing soluble molecules or membrane bound ones such as membrane constituents of bacteria or viral capsides.

Binding operates through non covalent interactions (electrostatic, dipole-dipole, Van der Waals or hydrophobic entropic forces) between the epitope atoms and amino acids of the light and heavy chain CDRs. Because they interact with intact proteins, BCRs and antibodies cannot access the hydrophobic core of globular or membrane proteins and only interact with their hydrophilic shell. This is of particular relevance for e.g Human Immunodeficiency Virus (HIV) infection control by broadly neutralizing antibodys (BNAbs) [192]. These antibodies have the ability to bind a hidden conserved region of an HIV capside receptor. It has been reported that those antibodies tend to have a long CDR3 loop region whose flexibility could help reaching hidden conserved subparts of the antigen [192].

1.4.2 *Plasmocytes*

Upon binding of a B lymphocyte's BCR to one of its cognate antigen and reception of additional signals from helper cells[20], the B-cell proliferates and differentiates into its corresponding effector cell called a plasmocyte or a plasma cell.

These cells migrate from the secondary lymphoid organs to the bone marrow. These cells are characterized by a large Golgi apparatus indicating an important proteo-synthesis activity. Since they are terminally differentiated cells, most of the plasma cell activity consists of producing antibodies that will flow in the blood and finally to the infection sites.

20 Upon recognition of an antigen by their BCR B-cells can internalize the pathogen and mount it on MHC class II in order to gather additional signals from follicular helper T-cells introduced in section 1.5.2 p.22

1.4.3 *Role of antibodies*

Antibodies are found in blood, extra cellular fluids and lumen of organs communicating with the outside world. Because body fluids were once known as humors, immunity mediated by antibodies is known as humoral immunity. There exist different kinds of antibodies detailed in the subsection1.4.5, each specialized in one of the following tasks:

- The most direct effect of antibodies is neutralization. Through binding directly to an antigenic molecule the antibody can block the antigen's function. This is particularly important for preventing viruses to enter their target cells. By binding the surface antigen enabling entrance in the cell the antibody can block the infection. Neutralization is also a major defense against toxic or poisonous hazards. Toxins such as the ones secreted by some bacterias are usually composed of two chains, one with toxic activity the other enabling to enter desired cells. By binding the latter one antibodies can prevent damages. Antibodies fulfilling such a neutralizing role are called neutralizing antibodies. The currently widely studied HIV BNAbs are part of this class. The term broad stem for their ability to neutralize HIV despite its constant evolution by targeting a structurally conserved region of a surface antigen.

- By coating the surface of pathogens that are self replicating such as bacteria, antibodies can act as targeting flags for innate immune cells with receptors binding constant regions of the antibodies. This coating process is known as opsonization.

- Finally by coating pathogens' surfaces antibodies can also trigger the activation of an important ingredient of the innate immune system: the complement. The complement system is an acellular system composed of a set of $\sim$ 30 soluble proteins found in the blood and extra cellular fluids. In the absence of pathogens most of these proteins are inactive and their activation comes from a cascade of proteolytic activity cleaving inactivating subparts of the proteins. Once activated the complement system can diffuse and promote inflammation by recruiting phagocytic cells, coating pathogens' surfaces and facilitate pathogens phagocytose by innate immune cells or directly disrupting the pathogens' membrane through formation of membrane-attack complexes.

1.4.4 *Affinity maturation and somatic hypermutations*

This section only gives a very brief outline of the affinity maturation process, and will only skim through the Somatic Hypermutation (SHM) process. A much more detailed summary of the current knowledge on SHMs is made as an introduction for chapter 6.

As described above, the primary B-cell response promotes cell division and for some, further differentiation into plasmocytes effector cells. However, some

of the newly divided cells will migrate within the lymph node[21] and keep dividing to form a structure known as a germinal center.. This structure comprises some B-cells, follicular dendritic cells, macrophages and follicular helper T-cells.

In those germinal centers B-cells start producing the Activation Induced cytidine Deaminase (AID) hypermutating enzyme. Because of this enzyme, B-cells accumulate random mutations at each cell division in their BCR variable region with a rate of $\sim 10^{-3}$ mutations per base pair per division. B-cells are then selected on their ability to recognize the antigen presented by follicular dendritic cells and present it to the helper T-cells, that will in turn provide signals preventing apoptosis and promoting division. Thus, B-cells bearing mutations destabilizing the receptor structure and preventing antigen binding will go through apoptosis (purifying or negative selection). For the ones still able to bind the antigen they will compete for follicular helper T-cell division signals, such that B-cells with receptors of higher affinity will carry an evolutionary advantage (positive selection).

The full process of division, mutation and selection is what is called affinity maturation. Overall, it is an accelerated evolution process aiming at providing B-cells with BCRs evermore affine for a given antigen. The full phylogeny deriving from an initial B-cell is called a clone.

1.4.5 *Ig classes and class switch*

On top of SHMs, the AID enzyme presented in the previous section serves another purpose. By creating double strand DNA breaks it allows to change the constant region lying on the 3' side of the J gene cluster (see Fig. 1.2 p. 14) and thus the function of the Ig encoded in a B-cell. Their function is dictated both by the downstream effector mechanisms their constant region triggers and the specific transporters carrying them to their specific sites of action. This process changing the expressed Ig type class is called class switch..

There exist five main Ig classes:

- immunoglobulin M (IgM) and immunoglobulin D (IgD). Respectively encoded by the C_μ and C_δ constant regions, lying right after the J genes cluster. These two types of Ig are concomitantly expressed as membrane receptors by alternative mRNA splicing. Since they lie right after the J gene cluster, these are the Igs that are expressed by B-cells before any class switch and thus before encountering of their cognate antigens. These are also the antibodies secreted by the primary plasma cells upon the first encounter of an antigen. Being produced before any affinity maturation, they generally have lower affinity than other Ig types. Still, IgM can form pentamers allowing strong binding on antigens with repeated epitopes such as bacterial membrane polysaccharides. Their primary role is to activate the complement. The role of IgD on the other hand is still poorly understood.

21 Together with their associated follicular helper T-cells

- immunoglobulins G (IgGs) come in four subsets (IgG1, IgG2, IgG3 and IgG4) respectively encoded by the $C_{\gamma 1}$, $C_{\gamma 2}$, $C_{\gamma 3}$, and $C_{\gamma 4}$ regions. Taken together IgGs are quite generalist and act through neutralization, opsonization or complement activation. They are found freely diffusing in the blood and extra vascular fluids. Some IgGs can also cross the placenta barrier so the mother can provide protection to the foetus during embryonic and early life. IgGs generally have a long plasmatic lifetime and are found in abundance in the blood.
- immunoglobulin A (IgA) are encoded by the C_{α} constant region. Together with IgG they are the predominant antibody class. Upon dimerization IgAs can be transported through epithelial barriers by specific transporters and thus be secreted in hollow organ lumens. There they mostly act by neutralization of pathogens and exogenous toxins. As a complement of placental IgGs, some IgAs are present in the mother's milk to transfer a temporary immunity to the newborn's gut.
- immunoglobulin E (IgE) encoded by the C_{ϵ} constant region. IgE antibodies are present only at very low levels in the blood and trigger mast cells activation. These are mostly involved in expulsion (and allergic) reactions such as sneezing, coughing or vomiting.

Class switch is a definitive change of the B-cell somatic DNA. It is triggered by follicular helper T-cells signal. The cocktail of cytokines expressed by follicular helper T-cells in germinal centers directs the choice of the class switch identity.

1.5 T-cells and their receptors

T-cells accomplish almost all their development in the thymus. Different maturations stages correspond to occupancy of different zones of the highly organized organ that is the thymus. There they will acquire their unique receptor whose functionality will be tested against self antigens. This process will be more thoroughly described in section 1.6 while this section will focus on describing TCR functioning along with the different T-cell subpopulations and their respective roles in the adaptive immune system.

1.5.1 *Ligands*

While BCRs recognize parts of soluble or membrane bound full proteins, TCR epitopes are fragments of antigens presented to the T-cell by other cells through an adapter protein. These proteins fragments are generated by degradation of exogenous, or the cells own, proteins, and then mounted inside the dedicated groove of the adapter protein, stabilizing its structure, before the antigen-adapter complex can be expressed on the cell's membrane. These transmembrane adapters are widely known as Major Histocompatibility Complexess (MHCs)[22].

22 For humans these are also referred to as Human leukocyte antigens (HLAs)

There exist two major classes of MHC molecules differing both by the type of peptide they can present and the cell types expressing them. Each individual possesses a combination of several MHC alleles of each class from maternal and paternal origin. Since not all MHC within a class can present the same protein fragments the large number of alleles allows one individual to have a larger epitope coverage . Class I MHCs bind short peptides of 8–10 amino acids and are ubiquitously expressed among nucleated cells. This class of MHC can be seen as a way for the adaptive immune system to constantly monitor the internal state of a cell and detect anomalies such as viral infection or cancerous protein expression. Class II MHCs on the other hand can bind peptides of various length greater than 13 amino acids and are expressed only by APCs.

As will be detailed in section 1.6 functional TCRs thus need to be able to bind to an MHC molecule and then recognize specifically the presented peptide. Specific co-receptors help the TCR bind to the MHC complex away from the peptide binding site. Their impact on T-cells role will be detailed in the next subsection.

Since T-cells recognize degraded protein fragments,they can access peptides hidden in the hydrophobic core of globular proteins that antobodies cannot access.

1.5.2 *Cytotoxic, helper and regulatory T-cells*

The two main classes of T-cells express either a cell-surface co-receptor protein called CD8 or another called CD4. As aforementioned, these co-receptor bind to subparts of the MHC complex. While CD8 only binds MHC class I, CD4 only binds MHC class II. The commitment to either receptor expression is made after full recombination of a TCR upon the positive selection step, and thus solely depends on the recombination product affinity for either MHC class molecule.

T-cell initial selection is presented in more details in section 1.6.2 p.24.

All CD8 mature T-cells are called cytotoxic T-cells. Their role is to kill cells to which their TCR binds by triggering their apoptosis (programmed cell death). Since CD8 promotes binding to MHC class I, that is ubiquitously expressed by nucleated cells, CD8 T-cells role is thus to kill all cells with abnormal (non self) proteic content and are thus of primary importance in fighting viral or intracellular infections or even cancerous cells.

CD4 T-cells are further subdivided in several subfamilies however their general role is mostly indirect through secretion of cytokines regulating activity of other immune cells. Subfamilies comprise:

- T_H1, T_H2 and T_H17 helper T-cells regulate the innate immune system activity. T_H1 mostly regulates macrophages, while T_H2 controls eosinophils, mast cells and basophils and T_H17 neutrophils.
- T_{FH} follicular helper T-cells regulating B-cell activity and affinity maturation in the lymph nodes
- T_{reg} regulatory T-cells. While other CD4 T-cells promote the immune response, T_{reg}s generally dampen it.

1.5.3 $\alpha : \beta$ *and* $\gamma : \delta$ *receptors*

So far we have only described T-cells bearing $\alpha : \beta$ receptors however there exist a second class of T-cells with different receptors that we briefly mention in chapter 3's conclusion. These $\gamma : \delta$ receptors originate from different recombination gene clusters. To this day it seems these receptors are not restricted to recognize ligands presented by the MHC, but their precise ligands remain unclear and so do the mechanisms controlling the commitment to the $\gamma : \delta$ or $\alpha : \beta$ lineages. Still, most $\gamma : \delta$ appear to lie in epithelial mucosal tissue and seem to be an intermediate level between the innate and adaptive immune systems.

1.6 INITIAL AND PERIPHERAL SELECTION

While being able to face virtually any pathogen armed with the tremendous diversity generated through the recombination process might already seem like a challenge, the task is even more extraordinary provided the constraints of self-antigen avoidance (autoimmunity) and finite resources (i.e finite total number of lymphocytes). As exposed in the following subsections this first constraint is dealt with by eliminating lymphocytes whose receptors bind strongly to self antigens in a process of purifying or negative selection. The second one should aim at keeping only the cells with most useful receptors. This is accomplished in several steps throughout the lymphocyte's life. First as mentioned in section 1.3.2, recombination products that cannot produce a viable receptor on the cell's membrane are negatively selected. Upon success to this functionality test, the receptor might then be tested for its ability to bind its cognate ligands and if so be positively selected. Finally, through competition for finite amounts of survival and division signals the population dynamics of peripheral selection will select the most useful receptors.

1.6.1 *B-cells central and peripheral tolerance*

B-cells development mostly implements negative selection. As briefly mentioned in section 1.3.2.3 p.16, upon recombination of a light chain a full BCR can be mounted on the cell membrane. Once on the membrane, the receptor is exposed to antigens expressed by the bone marrow cells or freely diffusing antigens produced elsewhere in the body. Binding of the receptor on a molecule at this stage is a signal of auto-reactivity, and will cause the B-cell to recombine again its light chain or die. Cells that do not react to any pathogen exit the marrow and migrate to the periphery.

At this point a number of self-antigens might not have been sampled by the developing B-cell, either due to their low concentration in the marrow or because of their tissue specific expression. Before the B-cell final maturation in the spleen, binding to an antigen while circulating will also result in the death of the lymphocyte. Because of the limited amount of maturing follicles in the spleen, newly created B-cell face harsh competition to enter them and thus spend some time sampling peripheral antigens.

1.6.2 *T-cell thymic selection*

Because they cannot directly bind antigens and need MHCs as adapter proteins, T-cells' selection process differs from B-cells'. A this stage immature T-cells express both CD4 and CD8 co-receptors. Without a signal of binding of its TCR to an MHC:self-antigen complex the immature T-cell will go trough apoptosis. This process is known as positive selection. On the other, hand upon too strong binding to one of those complexes the immature T-cell will be considered auto-reactive, and thus engage apoptosis. This in turn is called negative selection. T-cell thymic selection is thus a subtle balance between the ability to generally bind an MHC but not bind a specific MHC:self-antigen as illustrated in Fig. 1.4. This process, together with T cell activation, has been widely studied both at the molecular and the population scale [52], and the details of this decision are still not clear.

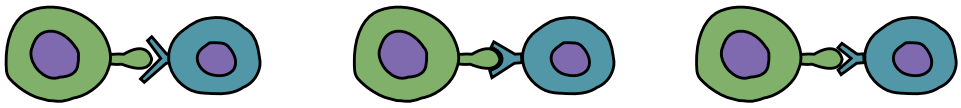

Figure 1.4: **Schematic for thymic selection.** Lymphocytes binding too weakly to the self MHC-peptides (left) engage in apoptosis implementing positive selection. Lymphoctyes binding too strongly (middle) also undergo apoptosis through negative selection. The obtained naive population binds but not too much self MHC:peptides complexes (right).

1.6.3 *Peripheral selection*

Once mature, lymphocyte populations compete for antigen resources and proliferation. This was initially described by the term clonal selection. Some recent theoretical work [42, 43, 96] has tried to describe clonal dynamics, however interpreting them to e.g detect clonal expansion upon vaccination trials remains a major challenge.

1.7 NAIVE AND MEMORY REPERTOIRES

Mature lymphocytes that have passed the different initial selection steps reside in the secondary lymphoid organs where they wait for activation. These mature lymphocytes that have not yet met their cognate antigen are termed naive. The ensemble of naive T or B lymphocytes constitute the naive repertoire.

Upon the encounter of its target antigen the naive lymphocyte divides and most of its progeny become effector cells. The remaining will constitute a pool of long lived and slowly dividing lymphocytes constituting the memory repertoire. This pool of memory cells is more numerous and easier to activate than the pool of naive lymphocytes. The presence of this memory repertoire is the basis of immunological memory and allows for a faster and stronger response upon successive encounters of the same pathogen. The stronger response is

due both to the amount of memory cells present and to their efficiency. Memory B-cells for instance can be created after several rounds of affinity maturation, such that class switched antibodies with sharpened affinity can readily be secreted upon secondary detection of an antigen. The long lifetime of these memory cells can provide to an individual a lifelong protection against a pathogen.

1.8 TOOLS TO STUDY THE ADAPTIVE IMMUNE SYSTEM

In this section I give a brief description of experimental techniques that may be useful to understand the work and results presented throughout the manuscript.

1.8.1 *Flow Cytometry*

As introduced in section 1.5.2 presence or absence of some membrane CD protein helps differentiating different stages of developpment, different populations and naive/memory repertoire. Flow cytometry typically uses fluorescent tags to count the amount of a given receptor on each cell membrane. Several tags [23] can be used to perform a multidiemnsional population analysis. A large field of immunology aims at performing immunophenotyping and divide functional cell classes according to their surface markers. The recent development of mass cytometry (CyTOF), relying on heavy metals as tags in mass spectrometry assays, enables large dimensional analysis of surface markers without fluorescence spectrum overlap restriction and promises to provide ever fine grained population description.

Such technique coupled with microfluidic droplets can be also used to isolate populations and perform cell sorting on the fly. This technique is referred to as Fluorescence-activated cell sorting (FACS). This is how sequencing datasets can be broken into naive/memory or even CD4/CD8 for T-cells.

1.8.2 *Immune repertoire sequencing*

The last decade has seen the advent of high-throughput repertoire sequencing with Next Generation Sequencing (NGS) techniques. These techniques allow to sequence millions of BCRs or TCRs around the CDR3 region containing most of the V(D)J recombination information and thus most of the repertoire's diversity.

A typical sequencing workflow consists of a few steps: genetic information extraction from the desired lymphocyte population, library preparation, Polymerase Chain Reaction (PCR) amplification[24] and sequencing. An example sequencing protocol used in Ref. [129] can be found in appendix B.1.

23 Up to non overlapping fluorescence spectra of the probes

24 PCR allows to duplicate genetic information. Starting from double stranded DNA, a first denaturation step is performed to obtain single stranded DNA. Using a small complementary sequence of each strand or *primer*, the genetic information is copied with a DNA polymerase synthesizing the rest of the complementary strand. This procedure is iterative and allows in theory to double the number of copies of a sequence at each cycle.

There currently exist two competing technologies using either genomic DNA or messenger RNA (reviewed in Refs. [28, 63]) as starting material

- DNA sequencing [88, 140, 189] uses the stable chromosomic information. Because productive and non productive rearrangements are equally efficiently sequenced and each cell contains the same amount of genetic material DNA sequencing should be extremely advantageous for unbiased clone size estimation (i.e number of cells carrying the same receptor). However, because the genetic information initially only exists in one copy many PCR cycles will be needed for sequencing. Since intronic regions are still present in DNA, this PCR relies on specific primers for the different V and J genes. Because PCR primers have different efficiencies and because the PCR amplification is itself a noisy process [9], in practice, DNA sequencing does not provide accurate sequence counts statistics despite its theoretical advantages.

- RNA sequencing [100, 124, 181] technologies on the other hand rely on the expressed genetic material. This introduces clear biases in count statistics as cells containing more mRNA will be overrepresented. Such a bias could potentially come from differential expression (e.g with allelic exclusion) or mRNA stability, and has been shown to strongly affect non productive rearrangements [28]. Although the starting amount of genetic material is larger, RNA-Seq technologies still require PCR amplification after retro-transcription in cDNA. Because, intronic regions have been excised during transcription a unique primer in the C region can be used to alleviate primer amplification biases. As for the inherent PCR bias, technologies using unique random molecular barcodes [83, 84], or Unique Molecular Identifiers (UMIs), attached to each cDNA molecule have been developed in order to track the number of times each molecule has been duplicated during the PCR process. Using these barcodes allows to efficiently correct for amplification biases. Because retrotranscription is not needed, adaptation of this technique to DNA sequencing would require to introduce UMIs in a first PCR replication step[155], along with an inherent primer bias, and has to this date not been done. Overall with molecular barcoding RNA techniques provide more accurate count statistics than their DNA counterparts, despite the inherent expression bias.

While this section only outlines the basic principles behind repertoire sequencing there exist many different methods varying in their depth, quality or read length are in constant development. One special case to mention is paired end sequencing allowing to obtain much longer reads ($\sim$ 200nt) than single read sequencing ($\sim$ 100nt) with lower error rates, however with the downside of using a second sequencing primer in the V region with possible subsequent primer bias. Such approaches however are still limited to study one chain and not the full receptor. The last years have seen the apparition of techniques to sequence paired receptors (α β or light-heavy) either through biochemical pairing [40] or statistical pairing [76, 91], an exciting development to study full receptor function.

Although I have clearly emphasized high throughput techniques targeting the hypervariable region some repertoire information have also been obtained using single cell RNA-Seq [26, 168], providing information on the whole cell transcriptional activity and paired receptor chains. Another recent development to mention is the assembly of immune receptor sequences from whole genome shotgun sequencing [15].

Because of the unprecedented insights in the immune repertoire global composition they offer, repertoire sequencing techniques promise to answer many long standing immunological questions. To mention just a few, Rep-Seq has already been used to try and assess repertoire overlap and its random or genetic basis [181, 206], recombination machinery statistics [115] and development [137, 154], initial selection traits [49], affinity maturation diversification [38, 195] and selection [194], repertoire diversity [140], and its links to aging [23] and diseases [183], dynamics of response to acute [59, 60, 77, 169] or chronic [73, 180, 193] infections. Although not directly sequencing the full repertoire, other approaches aiming at understanding the sequence to function mapping [1, 39, 203] are also an important development of immune receptor sequencing. However, interpretation of this wealth of data is arduous, and our ability to use it depends on the development of complex statistical and computational pipelines briefly presented in the next chapter.

2

INFERENCE, BIOINFORMATICS AND IMMUNE REPERTOIRE SEQUENCING

In this chapter I will introduce the quantitative methods I have used in this thesis. The first section presents the Expectation-Maximization (EM) algorithm and variants that we use in our probabilistic framework. Section 2.2 presents various models used in bioinformatics that we shall use or discuss within this manuscript. In the last section I will review challenges and existing solutions for repertoire sequences motivating the probabilistic approach we have adopted in this work. As a complement to this chapter, appendix A presents basic notions of optimization, information theory and bayesian statistics underlying the presented work.

2.1 INCOMPLETE DATA AND THE EXPECTATION-MAXIMIZATION ALGORITHM

Some problems come naturally with incomplete-data (degrees of freedom that are not or cannot be measured). In the case of Hidden Markov Models, as we will discuss in section 2.2.2.2, latent variables can be introduced on purpose to circumvent some model limitations. Solving problems that contain hidden variables can be done with the help of the Expectation-Maximization (EM) algorithm [41]. The EM algorithm allows one to perform Maximum likelihood (ML)[1] estimation of parameters given a statistical model by iteratively alternating between *Expectation* and *Maximization* steps. Because the V(D)J recombination machinery is degenerate and mutations cannot be assessed without the knowledge of the ancestor sequence the work presented in this manuscript naturally falls in the class of incomplete-data problems and EM will be at the center of this work. I will start this section by presenting a derivation of the EM algorithm and discussing its uses. I will then present extensions of this algorithm and finally propose a new stochastic variant of the algorithm.

2.1.1 *Derivation and use*

Let's assume we observe a dataset D of N independently and identically distributed observations x_n. Each of these observations is the result of a set of hidden variables z_n, with probability $P(x_n|z_n, \hat{\theta})$, distributed according to the distribution $P(z_n|\hat{\theta})$. The true parameter set $\hat{\theta}$ parameterizing these two distributions is unknown a priori and our goal is to estimate it using ML estimation.

1 Some variants can also be used to perform Maximum a posteriori (MAP) estimation [106], however we will not use this property in this work, implicitly assuming a uniform prior over the space of parameters.

Provided these ingredients the likelihood of a single observation x_n for an arbitrary set of parameters θ is

$$\begin{aligned}\mathcal{L}(x_n,\theta) = P(\boldsymbol{x}_n|\theta) &= \sum_{z_n} P(\boldsymbol{x}_n, z_n|\theta) \\ &= \sum_{z_n} P(z_n|\theta)P(\boldsymbol{x}_n|z_n,\theta). \end{aligned} \tag{2.1}$$

From Eq. 2.1 it is straightforward to compute the total likelihood of the dataset for a set of parameter θ

$$\mathcal{L}(D,\theta) = \prod_{\boldsymbol{x}_n} P(\boldsymbol{x}_n|\theta) = \prod_{\boldsymbol{x}_n}\sum_{z_n} P(z_n|\theta)P(\boldsymbol{x}_n|z_n,\theta). \tag{2.2}$$

For a large state space of hidden variable this calculation is hard (when not intractable) thus classical optimization techniques requiring many evaluations of the function before convergence might be extremely computationally expensive for directly maximizing this likelihood. After deriving correctness of the EM algorithm I will discuss its use and advantages compared to other convex optimization methods.

From the initial guess of the set of parameters θ one wishes to update these parameters to another set of parameters θ'. From Bayes formula the updated likelihood is $P(\boldsymbol{x}_n|\theta') = P(\boldsymbol{x}_n, z_n|\theta')/P(z_n|\boldsymbol{x}_n,\theta')$ and by computing the expectation of this likelihood over hidden variables with the current set of parameters θ on both sides we obtain

$$\begin{aligned}\sum_{z_n} P(z_n|\boldsymbol{x}_n,\theta)\ln P(\boldsymbol{x}_n|\theta') &= \sum_{z_n} P(z_n|\boldsymbol{x}_n,\theta)[\ln P(\boldsymbol{x}_n, z_n|\theta') - \ln P(z_n|\boldsymbol{x}_n,\theta')] \\ \iff \quad \ln P(\boldsymbol{x}_n|\theta) &= q(\theta'|\theta,\boldsymbol{x}_n) + h(\theta'|\theta,\boldsymbol{x}_n), \end{aligned} \tag{2.3}$$

where we have used $\sum_{z_n} P(z_n|\boldsymbol{x}_n,\theta) = 1$ and have defined

$$q(\theta'|\theta,\boldsymbol{x}_n) = \sum_{z_n} P(z_n|\boldsymbol{x}_n,\theta)\ln P(\boldsymbol{x}_n, z_n|\theta') \tag{2.4}$$

$$h(\theta'|\theta,\boldsymbol{x}_n) = -\sum_{z_n} P(z_n|\boldsymbol{x}_n,\theta)\ln P(z_n|\boldsymbol{x}_n,\theta'). \tag{2.5}$$

The difference between the log-likelihood $\ln\mathcal{L}(D,\theta)$ between the current set of parameters θ and the candidate new parameters θ' reads:

$$\begin{aligned}\ln\mathcal{L}(D,\theta') - \ln\mathcal{L}(D,\theta) &= \sum_{\boldsymbol{x}_n} q(\theta'|\theta,\boldsymbol{x}_n) - q(\theta|\theta,\boldsymbol{x}_n) + h(\theta'|\theta,\boldsymbol{x}_n) - h(\theta|\theta,\boldsymbol{x}_n) \\ &\geqslant \sum_{\boldsymbol{x}_n} q(\theta'|\theta,\boldsymbol{x}_n) - q(\theta|\theta,\boldsymbol{x}_n) \\ &\geqslant Q(\theta'|\theta) - Q(\theta|\theta), \end{aligned} \tag{2.6}$$

where $Q(\theta'|\theta) = \sum_{x_n} q(\theta'|\theta, x_n)$ and where we have used Gibbs inequality (Eq. A.19):

$$h(\theta'|\theta, x_n) - h(\theta|\theta, x_n) = \sum_{z_n} P(z_n|x_n, \theta) \ln \frac{P(z_n|x_n, \theta)}{P(z_n|x_n, \theta')} \geqslant 0. \tag{2.7}$$

This inequality ensures that maximizing the quenched average of the joint likelihood or "pseudo-log-likelihood" $Q(\theta'|\theta)$ over θ' increases the total likelihood by at least the same amount. The Expectation-Maximization scheme updates θ by doing such a maximization, and repeating the procedure iteratively. This guarantees linear [106] convergence of the algorithm to a local maximum of the likelihood.

In the rest of the manuscript $Q(\theta'|\theta)$ will always denote this quenched average likelihood or pseudo likelihood in the context of EM.

The EM algorithm thus finds the global optimum of the likelihood provided the likelihood function is convex. This task could also be carried through the use of convex optimization methods (see section A.1). What are the advantages of EM? In fact EM and gradient methods are tightly connected and it can be shown that the convergence speed of EM varies with the amount of missing information contained in the latent variables [147]. While for problems with small amounts of missing information[2] EM converges in a Newton-like fashion it can prove worse than direct gradient methods for problems with the fraction of missing information approaching unity. For well conditioned cases it thus offers fast convergence without the cost of Newton's Hessian inversion, quickly prohibitive for large numbers of parameters. Moreover, although one still has to optimize the Q function the quenched average allows to decouple model components for composite models such as Bayesian Networks, and thus perform independent optimizations in subspaces of lower dimensionality. Finally, in practice EM is a parameter free optimization thus preventing potential ill conditioning problems.

2.1.2 *Accelerating EM*

In this section I will briefly introduce possible accelerations of the EM scheme. Although some solutions involving switching (based on local estimations of the amount of missing information) between EM and gradient or Newton like methods to improve convergence exist [147] I will focus on techniques directly improving EM itself.

2.1.2.1 *Generalized EM*

Although initially formulated with a maximization step Eq. 2.6 implies that any marginal improvement of the pseudo likelihood yields improvement of the log-likelihood. For settings where the maximization step is computationally more costly than the expectation one, one can slightly improve the pseudo likelihood without carrying the full maximization, e.g using only one Newton step with line search. Such variants are known as Generalized EM or GEM.

2 E.g a mixture of well separated gaussians

2.1.2.2 *Sparse and stochastic EM*

Sparse EM use will be discussed in 3.2.6 p.57

On the other hand if the expectation step is the limiting one another strategy is to perform an approximate version of it. Indeed, the EM algorithm can be viewed as a maximization-maximization procedure by rewriting it in terms of a variational free energy (up to sign inversion) [119]:

$$F_n(R_n, \theta) \equiv \mathbb{E}_{R_n}[\ln P(x_n, z_n|\theta)] + H(R) \tag{2.8}$$

$$= -D_{KL}(R_n(z_n) \,||\, p(z_n|x_n, \theta)) + \ln \mathcal{L}(x_n, \theta), \tag{2.9}$$

with $R_n(z_n)$ an arbitrary probability distribution over hidden variables. From Eq. 2.9 one can show that if $F_n(R, \theta)$ has a local maximum at (R_n^*, θ^*) then $\mathcal{L}(x_n, \theta)$ also has a local maximum at θ^*. The likelihood of a data point $\mathcal{L}(x_n, \theta)$ can then be maximized by a coordinate ascent procedure alternatively maximizing $F_n(R, \theta)$ with respect to R_n and θ while keeping the other parameter set fixed. This first maximization step is carried by setting $R_n(z_n) = p(z_n|x_n, \theta)$ and is simply the expectation step. However as for generalized EM, on does not have to directly maximize F_n regarding R_n since any decrease in $D_{KL}(R_n(z_n) \,||\, p(z_n|x_n, \theta))$ allows us to iterate the coordinate ascent.

For very large state space of z_n for which the exact expectation step might be intractable, this formulation justifies approaches with inexact expectation steps through Monte Carlo integration (Stochastic EM), or only looking at subsets of the most likely hidden variables when a majority of them have negligible contributions (Sparse EM)[3].

2.1.2.3 *Incremental EM*

In the previous section I have presented a view of EM justifying approximations to the expectation step for a single observation x_n. Building on this construction we can write the variational free energy for the full dataset

$$F(R_1, \ldots, R_n, \theta) = \sum_{n=1}^{N} F_n(R_n, \theta) \tag{2.10}$$

$$= \ln \mathcal{L}(D, \theta) - \sum_{n=1}^{N} D_{KL}(R_n \,||\, p(z_n|x_n, \theta)). \tag{2.11}$$

Maximizing F regarding all R_n at the same time thus entails performing the exact expectation step over the whole data set. However, using the previous coordinate ascent view, the R_n being independent dimensions, maximizing F regarding only one or a subset of them increases F and allows us to iterate the coordinate ascent in θ.

Such a view justifies performing the expectation step on small observation batches [119, 121] in order to perform the maximization step more frequently and use more quickly the newly acquired information, leading to faster convergence. Assuming the pseudo likelihood Q can be formulated through a vector

3 This approach however does not guarantee to find the ML estimate, and gets arbitrary closer with the number of hidden states taken into account.

of sufficient statistics $\mathbf{s}^{(t)}(D) = \sum_n s_n^{(t)}(x_n)$ summarizing the inferential import of the complete data for the current set of parameters, an incremental version of the EM algorithm can be formulated as:

1. **E Step:** Chose a random subset $\mathbf{x}'$ of N' observations
   a) **for each** x_n in $\mathbf{x}'$ compute $s_n^{(t)}(x_n)$ and update $\mathbf{s}^{(t)}(D) = \mathbf{s}^{(t-1)}(D) - \sum_{x_n \in \mathbf{x}'} s_n^{(t-1)}(x_n) + \sum_{x_n \in \mathbf{x}'} s_n^{(t)}(x_n)$
   b) **for each** x_n not in $\mathbf{x}'$, $s_n^{(t)} = s_n^{(t-1)}(x_n)$ (do nothing)
2. **M Step:** Maximize $Q(\theta^t | \mathbf{s}^{(t)})$ with respect to θ^t

Such a design is exact and will lead to the ML estimate of the parameters θ. However if N is large this algorithm will still take time to forget out-dated sufficient statistics and storage of individual observations s_n^t might be prohibitive. An approximate incremental version with exponentially decaying memory can then be used where $\mathbf{s}$ is updated by

$$\mathbf{s}^{(t)} = \gamma \mathbf{s}^{(t-1)} + s_n^{(t)}, \tag{2.12}$$

with $\gamma \in]0, 1[$. These dynamics resemble a lot the stochastic gradient with momentum mentioned in section A.1.5 and the same proposed adaptive sampling strategy could thus be used to improve EM's convergence.

With modern parallel computing architectures one can increase linearly the computation speed, however the rules to synchronize and learn parameters in a stochastic setting are not obvious and should be a matter of caution [200].

2.2 BIOINFORMATIC APPROACHES TO SEQUENCE ANNOTATION

2.2.1 *Pairwise alignments and probabilistic interpretation*

Ever since the advent of DNA, RNA or protein sequencing techniques, assessing whether two complete or two pieces of sequences are related, has been a much sought after question. Phylogenist have tried to order the tree of life assessing the homology of two sequences, geneticists have tried to map RNA sequences to the underlying DNA substrate creating them and understand alternative splicing, or assemble genomes from sequence fragments by finding sequence overlaps. Finally, on their side immunologist have been interested in assigning the correct V, D and J segments used to create a particular receptor. In this section I will first introduce the principles and scoring scheme of pairwise alignments and how these can be implemented to study immune repertoires.

2.2.1.1 *Principle*

Pairwise alignments rely on a local probabilistic assessment of the relationship between nucleotides of two sequences [47]. Let's consider two sequences $\mathbf{x}$ and $\mathbf{y}$, which for simplicity we will assume to be vectors of same length L such that

$\mathbf{x}, \mathbf{y} \in (A, C, T, G)^L$. The aim of the pairwise alignment is to assess whether the two sequences are related (hypothesis R) or unrelated ($\bar{R}$).

Given $P(n)$ the probability of observing nucleotide n, one can construct the likelihood of observing $\mathbf{x}$ and $\mathbf{y}$ from unrelated sources as:

$$P(\mathbf{x}, \mathbf{y}|\bar{R}) = \prod_{i=1}^{L} P(x_i)P(y_i). \tag{2.13}$$

Now assuming we know $P(n, m|R)$ the probability of observing nucleotides n and m given that they descend from the same unknown parent nucleotide, we can write the likelihood for the two sequences to descend from the same ancestor:

$$P(\mathbf{x}, \mathbf{y}|R) = \prod_{i=1}^{L} P(x_i, y_i|R). \tag{2.14}$$

From this we can define the alignment score $S(\mathbf{x}, \mathbf{y})$ as the log-odds of the sequences to be related:

$$S(\mathbf{x}, \mathbf{y}) = \log\left(\frac{P(\mathbf{x}, \mathbf{y}|R)}{P(\mathbf{x}, \mathbf{y}|\bar{R})}\right). \tag{2.15}$$

Using such a logarithmic score allows us to break the global alignment of the two sequence problem into simpler subproblems that are the alignments of individual nucleotides:

$$S(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{L} s(x_i, y_i) = \sum_{i=1}^{L} \frac{P(x_i, y_i|R)}{P(x_i)P(y_i)}. \tag{2.16}$$

The $\mathbf{s}$ matrix is referred to as the substitution matrix. Given this matrix one can compute the alignment score of $\mathbf{x}$ and $\mathbf{y}$. However the choice of the parameters contained in this matrix is not obvious to fulfill our probabilistic interpretation and the choice of a good matrix will be discussed in the next section.

So far we have considered the case of a global alignment of two sequences of identical length. This is however an idealized case. For instance, finding a homologous sequence in a full genome entails finding the best global alignment between two sequences of very different sizes. In high throughput repertoire sequencing, since genomic templates undergo deletions it entails finding the best local alignment between the read and the genomic template. Section 2.2.1.3 details a dynamic programming algorithm solving this problem. Finally, evolutionary processes, SHMs, and sequencing machines can introduce insertions and deletions in sequences. These are modeled as gaps. Gaps can also be given a probabilistic interpretation.

Let's assume we know the probability of a gap of length l to occur and that this probability is a function $g(l)$ solely depending on the gap length. Assuming the inserted nucleotides are independent from the gap length and

are randomly drawn from the null distribution used to simulate the unrelated model:

$$P(gap) = g(l) \prod_{i \text{ in gap}} P(x_i). \tag{2.17}$$

Since we would like to know whether this sequence is a gap introduced after the common ancestor or if it is a random sequence, and thus compute the log-odds for a gap:

$$\gamma(gap) = \log\left(\left(\frac{P(gap)}{P(g\bar{a}p)}\right)\right) = \log\left(\frac{g(l)\prod_{i \text{ in gap}} P(x_i)}{\prod_{i \text{ in gap}} P(x_i)}\right) = \log g(l). \tag{2.18}$$

Note that assuming the probability of a deletion of length l follows the same function $g(l)$ we obtain directly the same result. Since the hidden common ancestor is usually unknown it is natural to chose such a convention when it is impossible to assess whether nucleotides were deleted from one sequence or inserted in the other. As for substitution parameters, the choice of the function g is left to parametrize the alignment score. The most common choice is to assume the gap length follows a geometric distribution, albeit other alternatives such as powerlaws have been proposed [199]. Such a geometric distribution again allows to have a simple linear or affine[4] additive score to model gaps allowing dynamic programming approaches to find the best alignment between two sequences.

2.2.1.2 *Substitution parameters estimation*

I emphasized the elegant probabilistic foundation of pairwise sequence alignments with a clear purpose: I want to stress that these methods encode a probabilistic model that might not be well suited for V(D)J recombination analysis.

Alignment of each genomic template allows to assess whether each nucleotide of the read is likely to originate from the template or another source. This source can be: random insertions, another gene class (such as D or J when aligning V genes) or another allele or gene of the same gene family[5]. It is thus unclear how to build a correct "null" model as we have done earlier for the traditional alignment examples discussed. Traditional substitution matrices such as *NUC4.4* were computed to assess phylogenetic data and encode evolutionary pressures on sequence changes. However this bias due to evolutionary pressures is irrelevant for V(D)J annotation. Finally, through their dynamic programming approach, alignments use only local information of the sequence while long range correlations might arise from the recombination

4 The affine scoring scheme $\gamma(gap) = -d - l \times e$ takes a gap opening penalty $d > 0$ and a gap extension penalty $e > 0$, where l is the length of the sequence.

5 For genes of the same family the sequence alignment score is used to chose the best gene candidate. However this is biasing the question, answering "What is the longest sequence I can align?", instead of "What is the sequence best explaining this recombination product?". Here again long range correlations might provide information.

process. In chapter 3 I will show to which extent modeling these long range correlations can improve recombination scenario assignment results.

Despite these limitations pairwise sequence alignments remain useful to extract coarse grained features of the V(D)J recombination process. In the framework presented in chapter 3 we will thus use pre-alignments as general guides to position[6] putative genomic templates on the sequencing read, assuming that such a positioning is "easy" to obtain and weakly sensitive to the underlying alignments parameters.

2.2.1.3 *Smith-Waterman local alignment*

The Smith-Waterman alignments allows one to find the best aligning subsequences of two sequences $\mathbf{x}$ and $\mathbf{y}$ accounting for possible gaps with a linear penalty. It is a dynamic programming approach allowing one to find one or several best local alignments with a complexity proportional to the product of the length of the two sequences ($O(L^2)$ for sequences of comparable length). The algorithm proceeds by iteratively filling an alignment score matrix $\mathbf{S}$ starting from $(i,j) = (0,0)$ with the following recursion rule:

$$S(i,j) = \max \begin{cases} 0 & \text{begin/end of alignment} \\ S(i-1,j-1) + s(x_i, y_j) & \text{match/mismatch} \\ S(i-1,j) - e & \text{insertion in } \mathbf{x} \\ S(i,j-1) - e & \text{insertion in } \mathbf{y}. \end{cases} \tag{2.19}$$

The same recursion without the possibility to begin/end (by setting $S(i,j) = 0$) a new local alignment enforces global alignment of the two sequences and is known as the Needleman-Wunsch algorithm. The best local alignment is obtained by starting from the position (i,j) of the matrix $\mathbf{S}$ with highest score and backtracking following the path used to fill the matrix until the beginning of the alignment ($S(i,j) = 0$). Finding the best global alignment via the Needleman-Wunsch algorithm also involves starting from the position with the best score which is now constrained to the last row or column.

2.2.1.4 *Hybrid strategy*

In order to identify potential V, D and J gene ancestors for a given sequencing read, one can use similar sequence alignment techniques. Since the D gene undergoes deletions from both sides we really seek a local alignment of the D gene on the read, and the Smith-Waterman (SW) algorithm is well suited. However V and J genes can only be deleted from one side, while the other side should fully align to the sequencing read. This calls for a hybrid strategy between global and local alignment.

6 Disregarding the score, and asking "If those sequences were to be related, how would they align best?".

Such a strategy can be implemented by removing the possibility to start/begin alignments and still starting backtracking from the position with the highest score, regardless of its location in the matrix[7].

2.2.2 *Markov Chains and Hidden Markov models*

So far the described inference algorithms considered identically distributed independent observations. However many real world observations come as ordered sequential data (such as time in speech recognition) or could be viewed as such (such as biological sequences). Markov chains are examples of such ordered sequential data and will be briefly introduced in the first subsection. Some of the presented notions will be used in chapter 3. The rest of the section focuses on the Hidden Markov Model (HMM) construction and related algorithms. Such models are widely used in biology [47], and, as will be discussed in section 2.3, more specifically in the field of our interest that is V(D)J recombination scenario assignment. Although the work presented in this thesis does not use HMMs I emphasize their importance because many software tools encode them, and a good understanding of their power and weaknesses will be useful to the reader to assess the work presented in chapter 3 and 6.

2.2.2.1 *Markov chains*

Markov processes are memoryless stochastic processes satisfying the Markov property:

$$P(x_{t+1}|x_t, x_{t-1}, \ldots, x_0) = P(x_{t+1}|x_t). \tag{2.20}$$

where x_t is the state of the system at time t and x_0 the initial condition. This memoryless property is often summarized saying that the next state of the system only depends on its current state .

Markov chains are stochastic processes with either a discrete state space or a discrete index set[8] satisfying this property. For the focus of this manuscript only discrete state space with discrete time chains will be discussed. Such processes can be summarized by a transition matrix $\mathbf{T}$ whose entries satisfy $P(x_{t+1} = j|x_t = i) = T_{ij}$[9] and a probability distribution over states $\boldsymbol{\pi}_t$, a row vector with entries $\pi_t^i = P(x_t = i)$. The dynamics of the stochastic process are then given by:

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t \mathbf{T}. \tag{2.21}$$

7 This strategy does not directly apply to J genes since it allows for deletions on the 5′ side instead of 3′ side. By reverting the genomic J and the read sequence this strategy can however be directly applied.

8 The index set is often time in physical processes, however as we will see in the next subsection it can be other sequential series such as position along a DNA sequence

9 Since the transition probability from state i to all other states must be 1, the sum over rows of $\mathbf{T}$ must be equal to one. This defines a right stochastic matrix. The problem could also be formulated with a left stochastic matrix with sum over columns equal 1. In that case the probability distribution over states is given by a column vector.

For time homogeneous Markov chains (whose transition matrix does not vary over time) one can easily compute a unique steady state distribution $\boldsymbol{\pi}^*$ provided the chain is irreducible (any state can be reached by any other state) and all its states are positive recurrent[10]. By definition, such a distribution $\boldsymbol{\pi}^*$ satisfies $\boldsymbol{\pi}^*\mathbf{T} = \boldsymbol{\pi}^*$ and is thus the row eigenvector of the transition matrix $\mathbf{T}$ associated with eigenvalue 1.

Provided the steady state distribution of the Markov chain one can compute a proxy (assuming steady state) for the entropy of a chain of length t. In general, for any stochastic process, the entropy rate H_r is the average information per unit time produced by the stochastic process such that:

$$H_r(x) = \lim_{t\to\infty} \frac{1}{t} H(x_1, \ldots, x_t), \tag{2.22}$$

where $H(x_1, \ldots, x_t)$ is the entropy of the sequence. Knowing the Markov chain has a stationary distribution, the entropy rate is independent of the initial distribution and converges to:

$$H_r(x) = -\sum_{ij} \pi_i^* T_{ij} \log T_{ij}. \tag{2.23}$$

Assuming the initial distribution π_0 is equal to the stationary one (or that the convergence time is small compared to t) the entropy of the Markov chain sequence $S(X, t)$ at time t is:

$$H(x, t) \simeq S(\pi_0) + (t-1)H_r(x). \tag{2.24}$$

2.2.2.2 *Hidden Markov Models (HMMs)*

Sequential or ordered data refers to any process whose current state only depends on past states (not future ones) and for which, without loss of generality, one could capture the full sequence probability by writing:

$$P(x_1, \ldots, x_N) = \prod_{n=1}^{N} P(x_n | x_1, \ldots, x_{n-1}). \tag{2.25}$$

Although exact, such a design is obviously not tractable since it would require learning ever larger sets of parameters with increasing sequence length N. Most sequential processes have finite memory and one could imagine restricting the learning to a conditional distribution on the M latest ancestors. Although viable this approach would still suffer from a large number of parameters to be learned, increasing exponentially with M. To circumvent this issue, one approach is to reduce the number of parameters by assuming some parametrized distribution for x_n and some relationship between parameters of the current distribution and those of the M^{th} last ancestors . Another approach is to add a layer of hidden variables.

Problems involving hidden variables have already been discussed within the scope of Expectation-Maximization (EM) in section 2.1. Hidden Markov

10 Starting from state i the chain will return to state i within finite time with probability 1.

Models (HMMs) [14] are a class of models describing sequential visible data x_n through a layer of latent variables z_n following a Markov Chain. An HMM can be generally defined as follows:

$$P(x_1,\dots,x_N,z_1,\dots,z_N) = P(z_1)\left(\prod_{n=2}^{N} P(z_n|z_{n-1})\right)\left(\prod_{n=1}^{N} P(x_n|z_n)\right), \quad (2.26)$$

where $P(z_n|z_{n-1})$ is the transition matrix[11] for the Markov chain and $P(x_n|z_n)$ the emission probability is the observed variables' dependency on the latent ones. Note that both x and z could be multidimensional and that there is no requirement for x and z to have equal dimension. In fact, x could be continuous. In a simple example the emission probability follows a Gaussian distribution parametrized by z and the resulting observed variable x follows a Gaussian mixture distribution. The emission probability $p(x|z)$ thus encodes possibly complicated non linear operations on top of a random process making the framework quite general.

An most interesting property of HMMs is that the resulting memory encoded for x_n need not be Markovian. From Eq. 2.26 and using Bayes formula, one can rewrite:

$$\begin{aligned} P(x_n|x_1,\dots,x_{n-1}) &= \frac{P(x_1,\dots,x_n)}{P(x_1,\dots,x_{n-1})} \\ &= \frac{\sum_{z_1,\dots,z_n} P(x_1,\dots,x_n,z_1,\dots,z_n)}{\sum_{z_1,\dots,z_{n-1}} P(x_1,\dots,x_{n-1},z_1,\dots,z_{n-1})}. \end{aligned} \quad (2.27)$$

This simple manipulation outlines the fact that $P(x_n|x_1,\dots,x_{n-1})$ cannot generally be reduced to $P(x_n|x_{n-1})$ and that a well designed HMM can in principle describe a general sequential process as introduced in Eq. 2.25. In the end, by introducing latent variables we are able to leviate the growth in the number of parameters needed to retain memory in a brute force design and can record long range correlations by learning a transition matrix and the emission probability distribution parameters.

A simple example can provide useful intuition. Let's consider an HMM for a dice game. In this dice game, the player throws two six-sided dice in a black box and a croupier announces the sum of numbers on the upper face of each dice. In this construct we thus have a two dimensional hidden variable z standing for the result of each dice whose value is converted to an observed variable x through a deterministic linear operation (the emission probability $p(z|x)$ is thus a delta Dirac peaked distribution). Now let's relax one of these assumptions and under the assumption that casinos are evil assume the croupier can lie and add ± 1 with equal probability to the score of each throw with probability q[12]. We now have constructed an emission probability distribution encoding a linear operation and a random process. However this problem still

11 Note that for large state spaces the transition matrix could be sparse and could be abandoned for a graph oriented approach.

12 Note that in this case, anytime a 0 or 13 is announced the croupier lies with probability 1 and one could trivially learn the parameter q simply using this data. However we could easily design another model with a more clever croupier and remove this possibility.

does not require to be solved by an HMM, since each throw is independent and data are not really sequential. In order to refine our toy HMM let's now change the rules of the game such that instead of throwing the dice the player has to give a slight flick to each die such that the dice can only roll on one of the side faces or stand still but cannot roll twice and display the face previously facing the ground as a result. Now our toy HMM is complete and our hidden variable follows a reducible Markov chain since the next result of each dice clearly depends and only depends on its present state. Given this physical HMM implementation and its parameters could one try and predict the most likely hidden dice results z underlying the announced score x? If the dice were to be biased could one learn their bias along with that of the croupier? These two questions can be respectively answered using two well known algorithms, the Viterbi and Baum-Welch algorithms, each succintly presented in the next subsections.

2.2.2.3 *Viterbi algorithm*

The aim of the Viterbi algorithm is to find the most likely path through an HMM given a sequence of observed variables $(x_1, \dots, x_N)$. We look for the path maximizing $P(x_1, \dots, x_N, z_1, \dots, z_N)$ over hidden variables states. The Viterbi algorithm much ressembles the transfer matrix approach for a one dimensional Ising spin chain at zero temperature, and thus only one accessible state, with each spin under an individual local field [6]. The spins stand for the hidden variables and the observed variables impose a local field for each spin.

Following Eq. 2.26 the algorithm functions as follows:

- compute $P(x_1, z_1) = P(z_1)P(x_1|z_1)$ for each possible state of z_1 where $P(z_1)$ is the initial state distribution.
- compute $P(x_1, x_2, z_1, z_2)$ for each couple of states (z_1, z_2). Since we are interested in finding only the most likely path, any transition from a state z_1 to a given state z_2 that is not locally the most likely cannot be part of the most likely path and shall not be remembered. We shall thus record only the K values $\max_{z_1} P(x_1, x_2, z_1, z_2)$ for each state z_2, where K is the number of states for any node z_n. These values will be used for the next iteration to compute the most likely path leading to any state z_3. In order to further be able to backtrack the most likely path, we shall store the K links between each state z_2 and its most likely ancestor.
- iterate until reaching the final node z_N by computing iteratively at each node $\max_{z_1,\dots,z_{n-1}} P(x_1, \dots, x_n, z_1, \dots, z_n)$ for each K possible states and record the K links pointing towards the most likely ancestor for each state of the current node z_n.
- once the last node is reached look for the state ending the most likely path and start backtracking.

Overall this approach provides the most likely path with $O(K^2N)$ computing complexity and $O(NK)$ memory requirement. This approach can be extended to extract the M most likely paths in parallel with the same computing

complexity or sequentially with $O(MK^2)$ computing complexity, both with $O(NMK)$ memory requirements [153].

Note that finding the most probable sequence of latent states is not the same as that of finding the set of states that are individually the most probable, since such a sequence might not even be viable if the Markov chain is reducible.

2.2.2.4 *Forward, Backward and Baum-Welch algorithms*

The Forward algorithm allows to compute $\alpha(z_n) \equiv P(x_1, \ldots, x_n, z_n)$ and is very similar to the Viterbi algorithm although instead of recording only the most likely transition leading to one state of the node z_n, all possible path leading to this node are summed. The computing complexity is thus comparable with $O(K^2N)$. Running the forward algorithm until the last node N is thus similar to using the transfer matrix approach to compute the partition function of the formerly described Ising spin chain with arbitrary temperature. The forward algorithm can be used to predict the most likely next observable x_{N+1} given the full history.

The Backward algorithm, sometimes referred to as *smoothing*, allows to compute $\beta(z_n) \equiv P(x_{n+1}, \ldots, x_N | z_n)$. Together with the Forward algorithm it can be used to compute the probability of a symbol x_n to come from a given hidden state z_n:

$$\begin{aligned} P(z_n | x_1, \ldots, x_N) &= \frac{P(x_1, \ldots, x_n, z_n) P(x_{n+1}, \ldots, x_N | x_1, \ldots, x_n, z_n)}{P(x_1, \ldots, x_N)} \\ &= \frac{\alpha(z_n)\beta(z_n)}{\sum_{z_N} \alpha(z_N)}. \end{aligned} \tag{2.28}$$

The Baum-Welch algorithm is the formulation of EM in the context of an HMM. Skipping the derivations, the expectation step consists in computing the posterior one point $P(z_n | x_1, \ldots, x_N)$ and two points $P(z_{n-1}, z_n | x_1, \ldots, x_N)$ marginal probabilities of the hidden state node z_n. Both these quantities can be computed from $\alpha(z_n)$ and $\beta(z_n)$ (Eq. 2.28), obtained respectively by the forward and backward algorithms. Computing these two quantities from $\alpha(z_n)$ and $\beta(z_n)$ is usually referred to as the Forward-Backward algorithm. The pseudo-likelihood $Q(\theta, \theta')$[13] can then be maximized. Iterating the E and M steps will lead to a maximum likelihood estimation of the HMM parameters (initial state distribution, transition matrix, and emission probabilities).

2.2.3 *Bayesian Networks*

Bayesian networks are a class of graphical models encoding conditional dependencies between random variables through a directed acyclic graph [14]. Graphical models are generally useful as their representation provides intuition and their correct implementation allows flexibility in model design. More-

13 Please note that so far no mention of the HMM parameters θ was made although all quantities related to the Viterbi, Forward and Backward algorithm are conditioned on θ. This conditioning has been omitted for clarity of exposition.

over, Bayesian networks exhibit interesting factorization properties regarding the inference of the parameters governing the random variables e.g through the use of the EM algorithm as shall be used in 3 p.51 or to compute quantities such as its entropy (section 2.2.3.2). The following subsections present formal definition of Bayesian Networks and factorization for entropy computation. Since a large amount of my work has been dedicated to implement a flexible software for V(D)J recombination statistics assessment through a Bayesian Network, I will use notations favoring intuition for the work presented in chapter 3.

2.2.3.1 *Definition*

Bayesian networks are encoded as directed acyclic graphs, whose vertices $i = 1, \ldots, K$ label individual random variables E_i. Note that we shall not make any formal distinction between a node and the variable to which it corresponds but will simply use the same symbol to refer to both. Dependence of the random variable E_j upon E_i is encoded, in the adjacency matrix $\boldsymbol{\nu}$, by a directed edge between E_i and E_j, denoted $\nu_{ij} = 1$ (while $\nu_{ij} = 0$ means no direct dependence). The set of parents of E_i, i.e. processes on which E_i depends directly, is denoted by $\mathcal{P}_i = \{j | \nu_{ji} = 1\}$.

Using these definitions we can, generally and irrespectively of the assumed form of the underlying model, write the joint probability of a complete scenario $\mathbf{E} = (E_1, \ldots, E_K)$ as:

$$P(\mathbf{E}|\theta) = \prod_{i=1}^{K} P(E_i | \{E_j\}_{j \in \mathcal{P}_i}, \theta), \tag{2.29}$$

with θ the parameter set parametrizing individual nodes distributions. Note that there is no constraint on the actual form of the distributions underlying the different random variables, such as whether they are discrete or continuous.

2.2.3.2 *Cross Entropy*

Since both the entropy and the Kullback-Leibler divergence between two distributions can be computed once one knows how to compute the cross entropy $H(\theta_1 \,\|\, \theta_2) = \sum_x P(x|\theta_1) \ln P(x|\theta_2)$ between the distributions for the two sets of parameters θ_1 and θ_2, we focus here on the computation of $H(\theta_1, \theta_2)$.

For Bayesian networks, the cross-entropy can be divided into subparts for each model component or node,

$$H(\theta_1 \,\|\, \theta_2) = \sum_{i=1}^{K} H_i(\theta_1 \,\|\, \theta_2), \tag{2.30}$$

with

$$H_i(\theta_1 \,\|\, \theta_2) = \sum_{\mathbf{E}} P(\mathbf{E}|\theta_1) \ln P(E_i | \{E_j\}_{j \in \mathcal{P}_i}, \theta_2). \tag{2.31}$$

To calculate this sum, one does not need to sum over all possible scenarios $\mathbf{E}$, but only over combinations of processes that affect E_i directly or indirectly. Let us call $A_i \subset \{1, \dots, K\}$ the set of indices affecting process i. These are defined as the "ancestors" of i in the acyclic graph, i.e. indices j such that there exists a lineage from j to i, $(i_1 = i, i_2, \dots, i_k = j)$ with $i_{\ell+1} \in \mathcal{P}_{i_\ell}$ (note that A_i includes i itself as a 0th order ancestor). Then the previous sum can be reduced to a sum over the processes in A only:

$$H_i(\theta_1 \,\|\, \theta_2) = \sum_{\mathbf{E}_{A_i}} \left[\prod_{j \in A_i} P(E_j | \{E_{j'}\}_{j' \in \mathcal{P}_j}, \theta_1) \right] \ln P(E_i | \{E_j\}_{j \in \mathcal{P}_i}, \theta_2). \tag{2.32}$$

where $\mathbf{E}_{A_i}$ denotes the subvector of elements of $\mathbf{E}$ with indices in A. Estimating the cross entropy for an event E_i requires exponential time in the number of ancestors of that node. Fortunately, in the models considered in this work the set of ancestors are small and obtaining the cross entropy is easy for every event.

2.3 Existing methods for Rep-Seq analysis

Analysis of repertoire sequencing data is challenging in many aspects. The degeneracy of the V(D)J recombination and hypermutation processes already naturally make the assignment of a recombination scenario ambiguous, with e.g identification of the incorporated D gene when it has undergone many deletions. During repertoire sequencing, PCR and sequencing further introduce errors. Short read length not covering the whole V and J region further introduce uncertainty as e.g analogous TCR V genes only differ by a few nucleotides. Finally, the wealth of data obtained, despite bringing unprecedented analysis opportunities, remains a major computational challenge.

Throughout the last few years a large bioinformatics endeavor has tried to propose solutions to tackle this analysis load, resulting in a variety of softwares, T or B-cell specific or generalist, each addressing a particular issue. As reviewed in Ref. [72] it is conceptually useful to separate low level processing methods aiming at preprocessing raw sequencing data for further data analysis from higher level methods aiming at extracting biological information. One should however bear in mind that this division is artificial, as preprocessing potentially influences the obtained biological information and accurate modeling of biological and sequencing processes should improve preprocessing. This simplifying hypothesis remains however for now necessary, as no tool can address all major conceptual issues of Rep-Seq. In the rest of this section I will expose the three major computational challenges of Rep-Seq analysis, their implications in low and high level data processing, existing solutions and interconnections. The last subsection will give a brief overview of higher level methods building on them.

2.3.1 *Error correction, clustering and clonal inference*

As mentioned earlier PCR and sequencing are error prone procedures. Introduced errors produce reads not corresponding to real receptor sequences, possibly biasing further diversity estimates, count statistics, or simply sequences themselves. Correcting for these errors by aggregating sequences originating from the same clone is thus of biological importance.

Such processing is sometimes carried by sequencing companies using proprietary software [72], for which reproducibility and detection of error correction artifacts are an issue. However, in general such software rely on a few simple approaches (reviewed in [72]):

- filtering of sequences with a low sequencing quality (Phred) score[14]
- clustering sequences based on pairwise similarity or distance, such as the Hamming or Levenshtein distance between sequences. Such an approach alone could however bias the resulting sequence statistics, for example sequences with lower number of insertions might be more prone to be clustered as nucleotide diversity is larger in the inserted region than in the genomic parts.
- removal of rare reads as they might come from late PCR or sequencing errors. This however artificially reduces the observed sequence diversity as sequence counts distribution are known to exhibit long tails [110, 112].

One of the most efficient error correction scheme relies on the mRNA molecular barcoding strategy previously described in section 1.8.2. The diversity of both the molecular barcode and V(D)J recombination makes very unlikely the pairing of similar UMI for two sequences that also resemble each other, and error correction simply consists in clustering similar sequences with similar UMIs. Despite being very effective, this strategy cannot correct errors that occurred during retro-transcription or early PCR steps. Refinement of the naive strategies involve using assigned V(D)J genomic templates to detect errors in the assigned genomic parts and perform clustering based on the CDR3 region (pRESTO [176], IMSEQ [87]). Some recent methods,such as MiGEC [160] or RTCR [64], model PCR error bias to further refine error detection even in barcoded data.

For memory B-cells, the clustering procedure is complicated by the fact that several distinct receptors may originate from a common ancestor due to affinity maturation. Reconstructing B-cells phylogenies is of biological interest to study disease evolution or quantify selection during affinity maturation. A wealth of software have been developed to address B-cells clustering using tailored distances [195], raw V(D)J annotation [15, 22] or more refined probabilistic approaches [79, 134]. Because hypermutated sequences in a clone might differ by only a single nucleotide and PCR errors are introduced in a branching process, also producing phylogenies, it is clear that error correction and clonal inference strategies might interfere.

14 This score is an output of the sequencing machine based on the likelihood of calling the incorrect nucleotide given the observed light spectrum upon nucleotide identification.

In the work presented in this manuscript we will rely on the provided clustered data for DNA TCRβ and heavy chains, and use a simple clustering based on alignments for the analyzed RNA TCRα and β .

2.3.2 *V(D)J annotation*

V(D)J annotation is probably the most prolific software field for repertoire sequencing analysis. The approaches relying on different algorithmic concepts differ in accuracy and speed by orders of magnitudes.

A number of assignment software tools rely on sequence alignment for V(D)J annotation. Some rely on the Smith-Waterman algorithm described in section 2.2.1.3 such as IMGT-V-QUEST [65]. However because the use of this algorithm is computationally demanding a majority of alignment software use the much faster BLAST [2][15] approach (IgBLAST [198], IMonitor [204], MiXCR [15]). Even faster methods focusing on small tags specific to each genomic templates can also be used. Such methods are implemented in Vidjil [46], LymAnalyzer [201], TCRklass [197], JOINSOLVER [165] or Decombinator [172] that implements an Aho-Corasick algorithm similar to the UNIX *grep* command. Although very efficient, these approaches will yield poorer results when there is only small portions of a gene that can be observed.

Another variety of software encode HMMs (RepGenHMM [48], iHMMunealign [58], Partis [135], SoDa2 [113]). All rely on pre-alignment processing and differ in their graph structure. Assignment is generally carried out using the Viterbi algorithm. The major difference between these algorithms is in their graph structure (i.e their statistical model assumptions) and the way their parameters (transition and emission probabilities) are estimated. As for the alignment based softwares some of them (iHMMune-align and SoDa2) rely on ad-hoc parameters while the others (Partis and RepGenHMM) are designed to be trained and learn parameters directly on provided datasets. Such a data driven approach is the focus of the work presented in this manuscript, however HMM model formulation is restrictive and cannot natively include long range correlations.

Alternatively more general methods can be used, using the same Bayesian framework as HMMs, by direct enumeration of possible recombination scenarios [115]. This direct enumeration, although computationally costly, does not suffer any model restriction thus enabling precise modeling of the biological process that is V(D)J recombination. The work presented in this manuscript generalizes this approach to provide IGoR (Inference and Generation of Repertoires) - a modular software that can encode models of arbitrary biological complexity. The general framework and applications to V(D)J annotation are presented in chapter 3, while chapter 6 illustrates how this general framework can be used to implement and infer a context dependent hypermutation model, a feat that is impossible for alignment and HMM based methods.

15 BLAST alignments consist of finding small identical regions between the query and reference sequences and elongate the alignment from there.

2.3.3 *Genomic templates inference*

The third cornerstone of accurate repertoire sequence analysis is the use of correct genomic templates for V(D)J annotation. Because erroneous or missing alleles can bias annotations and error or hypermutation assessment, inferring the appropriate alleles for each individual is primordial.

A first effort to standardize and centralize reference genomic templates annotation for different species was made with IMGT [92]. This manually annotated database has been successful and is the genomic base of many V(D)J annotation tools. However not all species have been annotated and not all allelic variants and Single Nucleotide Variants (SNPs) can be reported while the high variability and copy number variations of e.g V genes have been reported several times [18, 57, 185]. Because of the number of existing variants and their inference from short reads it is hard to constitute a complete and precise database [188].

In order to address this issue several tools have been specifically developed to infer dataset specific variants through alignment [198] or phylogenetic algorithms [56, 205].

2.3.4 *Other high level computations*

In the previous sections I have introduced the computational pillars supporting high level biological predictions based on repertoire sequencing. The following details a few examples of higher level computations embodying the successes of the dawn of repertoire sequencing analysis.

Borrowing diversity measures from ecology the large amount of sequences produced by repertoire sequencing has been used to estimate the diversity of an individual's repertoire, as reviewed in Ref. [110].

Despite the large diversity created by the V(D)J recombination process, the same clones can often be found to react against the same pathogen in different individuals. Using high throughput sequencing Venturi and colleagues have shown that such "public" response could arise simply from convergent recombination [179, 181], while N. Friedman's group suggested a link with sequence abundance and selection for self-associated antigens [98, 118].

The estimation of recombination statistics is both useful for V(D)J annotation and biological interest. As carried out in Refs. [48, 50, 115], using non-productive sequences one can extract raw V(D)J recombination statistics provided unbiased V(D)J recombination scenario exploration. Such statistics constitute a baseline and can be compared between individuals and receptor chains to delineate universal from specific V(D)J components. Building on this work, Ref. [49] proposed a framework to quantify selection on immune receptors. Altogether, these frameworks allow to estimate the potential diversity of the generated repertoire and compute the expected number or shared clones between individuals as discussed in chapter 5. Interestingly, this work also hints that the recombination machinery might have evolved to be biased towards the production of sequences that will be selected upon and antici-

pate selection. However it remains a challenge to relate these selection traits to physical constraints.

Similarly, predicting clonotypes that will respond to an infection based on their sequence remains far from our reach. Still significant progress has been made to predict the infection status of an individual based on its full repertoire. Using machine learning techniques B. Chain's group in Refs. [35, 173] managed to capture signatures of past infections and predict with great accuracy whether a repertoire has been exposed to a disease or not. Again, such approaches are still far from providing an understanding of the underlying physical processes but are still encouraging as they promise that repertoire sequencing data contains such information.

In the work presented in this manuscript we will build on IGoR's general statistical framework presented in chapter 3 to address biological questions such as the existence of BCR rearrangements incorporating several D genes. In chapter 4 we will use the inferred gene usage to reconstruct the chromosome organization and evaluate the probability of rescue upon failure of a first recombination attempt. Finally, in chapter 6 we will show that somatic hypermutations introduced during affinity maturation cluster.

Part III

A STUDY OF THE ADAPTIVE IMMUNE SYSTEM

This part contains the results of the research conducted during this PhD. It is largely based on a number of publications that are clearly indicated. I additionnally include text parts adressing more preliminary or unpublished work.

3 THE V(D)J RECOMBINATION PROCESS

Most of this chapter has been submitted for publication in Ref. [103].

3.1 INTRODUCTION

The adaptive immune system recognizes pathogens by binding their antigens to specific surface receptors expressed on T and B cells. The recent advent of high throughput immune repertoire sequencing (RepSeq) [63, 162, 186, 191] gives us direct insight into the diversity of B-cell and T-cell receptor (BCR and TCR) repertoires with great potential to change the way we diagnose, treat and prevent immune system related disorders. A growing number of algorithms and software tools have been designed to address the new challenges of RepSeq, in particular sequence analysis, germline assignment and clone construction [15, 25, 46, 70, 135, 172]. However, each receptor sequence can be generated in a large number of ways, or "scenarios," through recombination of genomic segments, insertions and deletions and hypermutations. Standard assignments introduce systematic errors when describing this inherently stochastic process. Quantitatively characterizing the diversity and the biases of these mechanisms remains a challenge for understanding adaptive immunity and applying RepSeq for diagnostics.

We present a flexible computational method and software tool, IGoR (Inference and Generation of Repertoires), that processes raw immune sequence reads from any source (cDNA or gDNA) and learns unbiased statistics of V(D)J recombination and somatic hypermutations. Using these statistics, for each sequence IGoR outputs a whole list of potential recombination and hypermutation scenarios, with their corresponding likelihoods. IGoR's performance at identifying the correct scenario is 2.5 times better than current state-of-the-art methods. IGoR used as a sequence generator produces an arbitrary number of randomly rearranged sequences with the same statistics as in the dataset.

This section details our general framework and how IGoR models the recombination machinery. Some sections will outline how the output information can be used to answer a few biological questions. Details about handling of hypermutations will be discussed in chapter 6.

3.2 METHODS

3.2.1 *Probabilistic assignment of recombination scenarios*

V(D)J recombination selects two or three segments (V and J for TCR α and BCR lights chains; V, D, and J for TCR β and BCR heavy chains) from a library of germline genes, and assembles them while deleting base pairs and inserting other non-templated ones at the junctions (Fig. 3.1a). B cell receptors can fur-

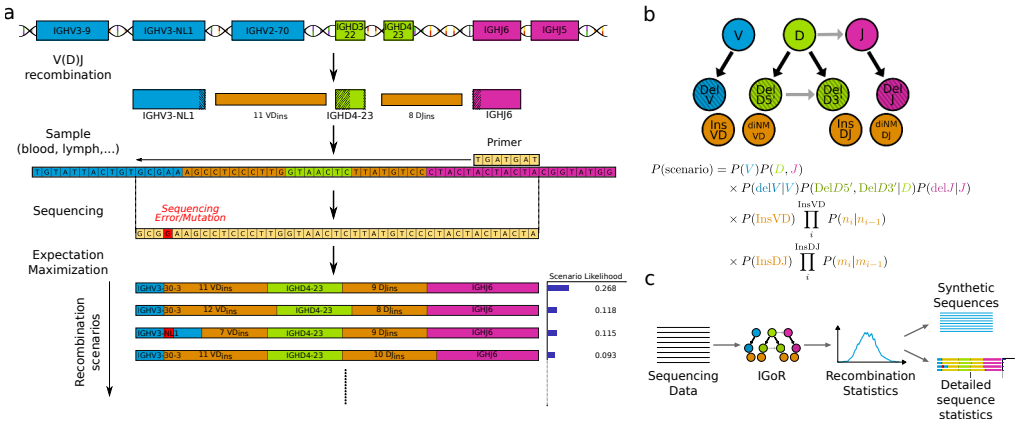

Figure 3.1: **IGoR's pipeline for sequence analysis. (a)** V(D)J recombination proceeds by joining randomly selected segments (V, D, and J segments in the case of IGH). Each segments gets trimmed at its ends (hashed areas), and a varying number of nontemplated insertions are added between them (orange). Hypermutations (in the case of B cells) or sequencing errors (in red) further enhance diversity. IGoR lists putative recombination scenarios consistent with the observed sequence, and weighs them according to their likelihood. **(b)** The likelihood of each scenario is computed using a Bayesian network of dependencies between the recombination features (V, D, J segment choices, insertions and deletions), as illustrated here for the human TRB locus. Architectures for TRA and IGH are described in Online Methods. **(c)** IGoR's pipeline includes three modes. In the learning mode, IGoR learns recombination statistics from data sequences. In the analysis mode, IGoR outputs detailed recombination scenario statistics for each sequence. In the generation mode, IGoR produces synthetic sequences with specified recombination statistics.

ther diversify through somatic hypermutations during affinity maturation. The recombination process is degenerate, as the same sequence can be generated in many different ways [179]. IGoR starts by listing the possible recombination and hypermutation scenarios leading to an observed sequence in the dataset. It then assigns probability weights reflecting the likelihood of these scenarios. As the example in Fig. 3.1a shows, explored scenarios can be very different yet have comparable contributions to the sequence likelihood. Since exploring all possible scenarios would be computationally too costly, IGoR restricts its exploration to the reasonably likely ones. Scenario exploration takes from 1 ms up to less than a second per sequence on a single CPU core, depending on the chain (see full distributions of runtimes in Fig. C.1). Different recombination architectures and dependencies can be configured within IGoR by specifying dependencies between elementary events (gene choices, deletions, insertions, hypermutations) through an acyclic directed graph, or Bayesian network, as illustrated in Fig. 3.1b for the case of TCR β chains (see section 3.2.2 for the other used structures).

IGoR functions according to three modes: learning, analysis, and generation (Fig. 3.1c). In the learning mode, IGoR infers the recombination statistics of large datasets of sequences using a Sparse Expectation-Maximization algorithm (see section 3.2.6). In the analysis mode, IGoR assigns recombination events to sequences in a probabilistic way, by outputing the most likely scenar-

ios ranked by their probabilities, as well as the overall generation probability of the sequence. In the generation mode, IGoR outputs random sequences with specified statistics, e.g. learned from real datasets.

In the next section we give the particular model structures used in this study. We then give a more general definition applicable to other general types of recombination products.

3.2.2 *Models for TRA, TRB and IGH*

We define a probabilistic model for each type of chain (e.g. α, β, heavy, light) that describes the probability of each recombination event $\mathbf{E}$ by the probabilities of the known elements of the recombination subprocess (gene choice, insertions, deletions at each of the junctions etc) for each chain, and assumes only the minimum correlations between the subprocesses needed to explain the correlations observed in the data. We model insertions as a Markov chain (the identity of an inserted nucleotide only depends on the previously inserted one) with a nonparametric length distribution [50, 115, 129]. For each insertion site (X= VD and DJ for β and heavy chains and X=VJ for α and light chains) we infer the probability of observing a non-templated sequence of a given length, $P(\text{insX})$, and the transition matrices $P_{VJ}(n_i|n_{i-1})$, $P_{VD}(n_i|n_{i-1})$, $P_{DJ}(m_i|m_{i-1})$ giving the probability of inserting a given nucleotide as a function of the identity of previous one. For each gene we infer the probability of the number of deletions conditioned on the gene identity, e.g. $P(\text{delV}|V)$ for deletions from the V gene. We model templated palindromic insertions as negative deletions [50, 115]. The D gene is very short and may get fully deleted. This introduces correlations between the deletions on both sides of the original D gene template. We account for these correlations by inferring the joint probability $P(\text{delD}l, \text{delD}r|D)$. We treat every allele as a different gene [50] and infer the joint gene usage $P(V, D, J)$ for β and heavy chains, and $P(V, J)$ for α and light chains, to be able to capture correlations between segment usage.

For TCR α chains or BCR light chains, the probability of a recombination event $\mathbf{E} = (V, J, \text{delV}, \text{delJ}, \text{insVJ})$ is:

$$\begin{aligned} P^{\alpha/L}_{\text{recomb}}(\mathbf{E}) = {} & P(V, J)P(\text{delV}|V)P(\text{delJ}|J) \\ & \times P(\text{insVJ}) \prod_{i}^{\text{insVJ}} P_{VJ}(n_i|n_{i-1}). \end{aligned} \tag{3.1}$$

Similarly, the probability $P^{\beta/h}_{\text{recomb}}(\mathbf{E})$ of a recombination event

$$\mathbf{E} = (V, D, J, \text{delV}, \text{delD}l, \text{delD}r, \text{delJ}, \text{insVD}, \text{insDJ})$$

for a TCRβ or BCR heavy chain is:

$$\begin{aligned} P^{\beta/H}_{recomb}(\mathbf{E}) = {} & P(V,D,J)P(delV|V) \\ & \times P(insVD)P(delDl, delDr|D) \\ & \times P(insDJ)P(delJ|J) \\ & \times \prod_{i}^{insVD} P_{VD}(n_i|n_{i-1}) \prod_{i}^{insDJ} P_{DJ}(m_i|m_{i-1}). \end{aligned} \tag{3.2}$$

In the case of TRB, gene usage is further factorized as $P(V,D,J) = P(V)P(D,J)$.

These models are similar to those used in Refs. [50, 115, 129]. The conditional dependencies were introduced so as to reproduce the mutual information computed between the different recombination events on real sequencing data.

3.2.3 *General model formulation*

The definition and properties of Bayesian Networks are given in section 2.2.3 p.41

IGoR is designed in a modular way so the user can define arbitrary model forms. The models are Bayesian networks encoded as directed acyclic graphs, whose vertices $i = 1, \ldots, K$ label individual recombination subprocesses E_i (V, D, J choices, deletions, etc. in the examples above).

As introduced in section 2.2.3.1 we can write the probability of a recombination scenario $\mathbf{E} = (E_1, \ldots, E_K)$ as:

$$P_{recomb}(\mathbf{E}|\theta) = \prod_{i=1}^{K} P(E_i|\{E_j\}_{j\in\mathcal{P}_i}, \theta), \tag{3.3}$$

where θ denotes the underlying model parameters (i.e. probability distributions of gene choice, insertions at a given junction, and deletions from a given gene in the studied examples) and $\mathcal{P}_i$ the set of parents of the event indexed by i.

Each recombination scenario $\mathbf{E}$ leads to a unique sequence $\hat{\mathbf{S}}(\mathbf{E}) = (\hat{S}_1, \ldots, \hat{S}_L)$, $\hat{S}_i(E) \in \{A, C, G, T\}$ (in the following we often write $\mathbf{S}$ for $\hat{\mathbf{S}}(\mathbf{E})$ for brevity). However, in order to produce a given sequence $\mathbf{S}$ several scenarios might be equivalent, and we can write the probability of generating a given sequence as:

$$P_{gen}(\mathbf{S}|\theta) = \sum_{\mathbf{E}|\hat{\mathbf{S}}(\mathbf{E})=\mathbf{S}} P_{recomb}(\mathbf{E}|\theta). \tag{3.4}$$

The above description only holds to assess the generation probability of a pure product of recombination and does not account for sequencing errors or hypermutations. Note that, since longer reads allow for more reliable determination of V and J gene segments, P_{gen} depends in general on read length: shorter reads can be created in more ways than longer reads, leading to larger P_{gen}.

3.2.4 *Errors and hypermutations*

Sequencing is inherently noisy and introduces nucleotide substitutions. In addition, BCRs can accumulate hypermutations, which can be mathematically treated in the same way as errors. For the sake of clarity, we distinguish between the sequencing read $\mathbf{R}$ and the original sequence $\mathbf{S}$ resulting from recombination, as defined above. For simplicity we ignore insertion and deletion errors, so that $\mathbf{R}$ and $\mathbf{S}$ are of the same length L.

We define our error model as deviations from the initial recombination event (through sequencing errors or somatic hypermutations) such that $P_{\text{err}}(\mathbf{R}|\mathbf{S},\theta)$ is the probability of observing the sequencing read $\mathbf{R}$ given the recombination product $\mathbf{S}$. Since the recombination scenario $\mathbf{E}$ completely determines $\mathbf{S}$, $P_{\text{err}}(\mathbf{R}|\mathbf{S},\theta) = P_{\text{err}}(\mathbf{R}|\mathbf{E},\theta)$, and we use these two notations interchangeably. The dependence on θ reflects the fact that θ also includes the parameters of the error or hypermutation model.

We write the joint probability of producing a given sequence $\mathbf{S}$ and observing a given read $\mathbf{R}$ as:

$$P(\mathbf{R},\mathbf{S}|\theta) = P_{\text{gen}}(\mathbf{S}|\theta)P_{\text{err}}(\mathbf{R}|\mathbf{S},\theta). \tag{3.5}$$

Summing over all possible recombination products, the likelihood of a sequencing read is:

$$\begin{aligned} P_{\text{read}}(\mathbf{R}|\theta) &= \sum_{\mathbf{S}} P(\mathbf{R},\mathbf{S}|\theta) \\ &= \sum_{E} P_{\text{recomb}}(E|\theta)P_{\text{err}}(\mathbf{R}|\mathbf{E},\theta), \end{aligned} \tag{3.6}$$

and the total likelihood of the model given a dataset of reads $(\mathbf{R}^1,\ldots,\mathbf{R}^N)$ is given by:

$$\mathcal{L}_{\text{total}}(\theta) = \prod_{a=1}^{N} P_{\text{read}}(\mathbf{R}^a|\theta). \tag{3.7}$$

3.2.5 *Maximum likelihood estimate*

The recombination machinery is degenerate, as several scenarios of recombination and hypermutations can lead to the same sequence, and the recombination scenario $\mathbf{E}$ from which the sequencing read $\mathbf{R}$ comes from is in general unknown. As previously introduced, the Expectation-Maximization algorithm is a commonly used algorithm that maximizes the likelihood of models with hidden variables given the data. In this section we derive the update rules for our class of models.

A derivation and justification of the Expectation-Maximization algorithm is given in section 2.1 p.29

3.2.5.1 *Optimizing the recombination model*

The pseudo-log-likelihood can be broken up in two independent terms, $Q(\theta'|\theta) = Q_{\text{recomb}}(\theta'|\theta) + Q_{\text{err}}(\theta'|\theta)$, respectively corresponding to the recombination model and the error or hypermutation model:

$$Q_{\text{recomb}}(\theta'|\theta) = \sum_{a=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a, \theta) \ln P_{\text{recomb}}(\mathbf{E}|\theta'). \tag{3.8}$$

$$Q_{\text{err}}(\theta'|\theta) = \sum_{a=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a, \theta) \ln P_{\text{err}}(\mathbf{R}|\mathbf{E}, \theta'). \tag{3.9}$$

In order to maximize the pseudo-log-likelihood of the recombination model we need to maximize $Q_{\text{recomb}}(\theta'|\theta)$ with respect to every model component contained in the parameter set θ', $P'(E_i|\{E_j\}_{j\in\mathcal{P}_i})$. We impose normalization using Lagrange multipliers, λ_i, and define:

$$\hat{Q}_{\text{recomb}}(\theta'|\theta) = Q_{\text{recomb}}(\theta'|\theta) + \sum_i \lambda_i \left[1 - \sum_{E_i} P'(E_i|\{E_j\}_{j\in\mathcal{P}_i})\right]. \tag{3.10}$$

Taking the functional derivative of $\hat{Q}_{\text{recomb}}(\theta^*|\theta)$ with respect to the model parameter we get:

$$\frac{\partial \hat{Q}_{\text{recomb}}(\theta'|\theta)}{\partial P'(E_i|\{E_j\}_{j\in\mathcal{P}_i})} = \sum_{a=1}^{N} \sum_{\mathbf{E}'} \delta_{E_i,E_i'} \frac{P(\mathbf{E}'|\mathbf{R}^a, \theta)}{P'(E_i|\{E_j\}_{j\in\mathcal{P}_i})} + \lambda_i. \tag{3.11}$$

Setting this derivative to zero gives:

$$P'(E_i|\{E_j\}_{j\in\mathcal{P}_i}) = \frac{1}{N} \sum_{a=1}^{N} \sum_{\mathbf{E}'} \delta_{E_i,E_i'} P(\mathbf{E}'|\mathbf{R}^a, \theta), \tag{3.12}$$

where the Lagrange parameter $\lambda_i = N$ ensures normalization. In other words the modified log-likelihood is maximized by using an update rule that equates the probability of a realization of a recombination event to its posterior frequency.

3.2.5.2 *Optimizing the independent single nucleotide sequencing error model*

The independent single nucleotide error model is the simplest instance of an error model, where each nucleotide of the read has a probability r to be missequenced as one of the three other nucleotides with equal probability. For this model we have

$$P_{\text{err}}(\mathbf{R}|\mathbf{S}, \theta) = \left(\frac{r}{3}\right)^{N_{\text{err}}} (1-r)^{L)-N_{\text{err}}(\mathbf{R},\mathbf{S})}. \tag{3.13}$$

where $N_{err}(\mathbf{R}, \mathbf{S})$ the number of mismatches between $\mathbf{R}$ and $\mathbf{S}$, and L the number of error-prone base pairs. We compute the derivative of the modified log-likelihood of the error model with respect to r^* as:

$$\frac{dQ_{err}(\theta'|\theta)}{dr'} = \sum_{a=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a, \theta) \left(\frac{N_{err}(\mathbf{R}^a, \hat{\mathbf{S}}(\mathbf{E}))}{r'} - \frac{L(\mathbf{R}^a, \mathbf{E}) - N_{err}(\mathbf{R}^a, \hat{\mathbf{S}}(\mathbf{E}))}{1 - r'} \right). \tag{3.14}$$

Setting this derivative to zero yields:

$$r' = \frac{\sum_{a=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a, \theta) N_{err}(\mathbf{R}^a, \hat{\mathbf{S}}(\mathbf{E}))}{\sum_{a=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a, \theta) L(\mathbf{R}^a, \mathbf{E})}, \tag{3.15}$$

where $L(\mathbf{R}^a, \mathbf{E})$ is the number of potentially erroneous nucleotides in read a. For simplicity we ignore errors and hypermutations in the insertion part of the sequence, as they are almost indistinguishable from unmutated random insertions, and accounting for them would imply summing over an exponentially large number of scenarios. As a result, L in the above formula is not the read length, but rather the number of genomic nucleotides in each scenario, which depends on the scenario $\mathbf{E}$ as well as on the sequence read.

3.2.6 *Pruning the tree of scenarios*

Since enumerating all possible scenarios for each sequence is not tractable, we used a heuristic method for reducing their numbers. Exploring all possible scenarios is equivalent to exploring all the terminal leafs of a tree. Our heuristic is to prune all branches that do not contribute substantially to the likelihood of the read. To do this we implement a Sparse Expectation Maximization algorithm as previously motivated in section 2.1.2.2. Due to the acyclicity of the directed graph underlining the Bayesian network, there exists a topological sorting of the events constituting a partially ordered set (we will assume in the following that the indices of the different events E_i respect this ordering). IGoR processes event realizations according to this order corresponding to different layers of depth in the tree. To discard irrelevant branches (containing negligible scenarios) IGoR computes at each depth k (with $0 \leqslant k < K$) an upper bound on the probability of the currently explored scenario:

$$\frac{\prod_{0 \leqslant i \leqslant k} P(E_i, \mathbf{R}|\{E_j\}_{j \in \mathcal{P}_i}, \theta) \prod_{k < i < K} \max_{e_i} P(E_i, \mathbf{R}|\theta)}{\max_{\mathbf{E} \in \mathcal{E}} P(\mathbf{E}, \mathbf{R}|\theta)} > \varepsilon, \tag{3.16}$$

where $\mathcal{E}$ is the set of already fully explored scenarios, and $0 \leqslant \varepsilon \leqslant 1$ is a tunable parameter setting the precision of the sparsity approximation. While $\varepsilon = 0$ will explore every possible scenario and perform an exact Expectation step, $\varepsilon = 1$ will explore only scenarios more likely that any scenario already explored.

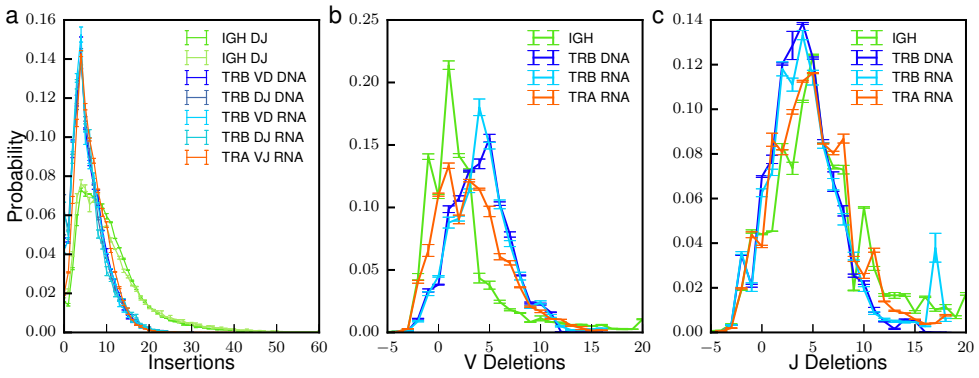

Figure 3.2: **IGoR infers reproducible recombination statistics. (a)** Distribution of the number of insertions at the junctions of recombined genes: IGH at the VD and DJ junctions from DNA data [88], TRB at the VD and DJ junction from both DNA [140] and mRNA data [129], and TRA at the VJ insertion site from mRNA data [129]. **(b)**,**(c)**. Average distribution of the number of deletions across **(b)** V and **(c)** J genes. Negative deletions correspond to palindromic insertions (P nucleotides), e.g. -2 means 2 P-nucleotides. The inferred distributions are robust to the choice of individuals, genetic material (mRNA or DNA) and sequencing technology. Error bars show 1 standard deviation across individuals.

Although Eq. 3.16 captures the essence behind our tree pruning approach, in practice IGoR uses more information than a simple upper probability bound. By picking two gene choice realizations, imposing the identity and position of these specific V and J genes, we explicitly impose the total nucleotide length of event realizations between those V and J genes (number of insertions, deletions, D gene length, ...). When computing the probability upper-bounds IGoR computes the upper probability bound for a given junction length between two event realizations, and uses this refined bound to efficiently prune the tree of scenarios.

3.3 PARAMETERS LEARNED ON SEQUENCING DATA

We used IGoR's learning mode to infer the accurate statistics of V(D)J recombination from four datasets comprised of unique sequences of non-productive rearrangements of three different chains, sequenced either at the levels of mRNA (TCRα chain or TRA, and TCRβ chain or TRB [129]) or DNA (TRB [140], BCR heavy chain or IGH from naive cells [88]), generalizing earlier methods [48, 50, 115]. Restricting to nonproductive unique sequences allowed us to avoid biases introduced by functional selection. The Expectation-Maximization algorithm converged within a few iterations (see Fig. C.2 for convergence of parameters, and Fig. C.3 for the case of IGH).

The same TRB insertion and deletion distributions were inferred regardless of the individual, laboratory of origin, or sequencing protocol, and of whether DNA [140] (light blue distributions in Fig. 3.2) or mRNA [129] (dark blue) was used. By contrast, V and J gene usage varied moderately but significantly across individuals, and even more across sequencing technologies, suggest-

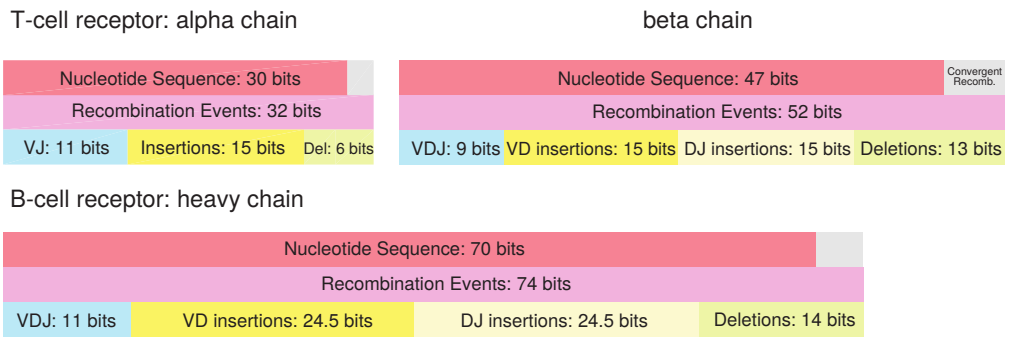

Figure 3.3: **Recombination entropy for TCRα ,TCRβ and BCR heavy chains.** The total recombination entropy (purple) can be decomposed into individual contributions of the recombination events. This figure decomposes contributions of the choice of the V(D)J genes (blue), the number and identity of non templated insertions (yellow), and number of deletions (light green). The sequence entropy (red) is slightly smaller than the recombination entropy because several recombination events can lead to the same sequence (convergent recombination, in gray). Adapted from Ref. [110] with authors' permission.

ing possible primer-dependent biases (Fig. C.4, see also Fig. 3.7 for IGH D-J gene usage). Insertions at the TRA V-J junction, and at the TRB V-D and D-J junctions have similar distributions (Fig.3.2a), as previously reported [48]. IGH have significantly more insertions at the junctions than TCRs, consistent with previous observations [50].

3.4 RECOMBINATION ENTROPY

IGoR's recombination models are encoded by Bayesian networks. As such, recombination entropy can be computed as explained in section 2.2.3.2. For most recombination elements representing categorical distributions this computation is straightforward. The dinucleotide Markov model encodes a Markov chain whose length is dictated by the insertion length distribution and its entropy can be approximated as presented in section 2.2.2.1. The cross entropy of an inserted region of length insVJ (or insVD, or insDJ) for two sets of parameters $\theta 1$ and $\theta 2$ is given by

$$h(\text{insVJ}, \theta_1, \theta_2) = \sum_{\mathbf{n}} P(\mathbf{n}, \theta_1) \ln P(\mathbf{n}, \theta_2) \tag{3.17}$$

$$= \sum_{n_1} P_s(n_1|\theta_1) \ln P_s(n_1|\theta_2) \tag{3.18}$$

$$+ (\text{insVJ} - 1) \sum_{n_1, n_2} P_s(n_1|\theta_1) P(n_2|n_1, \theta_1) \ln P(n_2|n_1, \theta_2), \tag{3.19}$$

where $\mathbf{n} = (n_1, \ldots, n_{\text{insVJ}})$ is the inserted sequence, and $P_s(n_1, \theta)$ is the stationary distribution of the Markov chain of insertions.

Although not necessarily conditioned on insertion length the dinucleotide model functionally depends on the number of insertions. The cross entropy

for a dinucleotide model, once averaged over possible lengths is then given by

$$H_{\text{VJ insertions}}(\theta_1 \,\|\, \theta_2) = \sum_{\mathbf{E}_B} \left[\prod_{j \in B} P(E_j | \{E_{j'}\}_{j' \in \mathcal{P}_i}, \theta_1) \right] h(\text{insVJ}, \theta_1, \theta_2), \quad (3.20)$$

where $B \subset \{1, \ldots, K\}$ is the subset of processes affecting either insVJ or $\mathbf{n}$, excluding insVJ itself.

The partition of the entropic contributions of the inferred model components for the different receptor chains is shown in Fig. 3.3. As previously reported in Refs. [48, 50, 110, 115] non templated insertions are responsible for a large part of the recombination entropy, dominating the combinatorial diversity generated by the choice of genomic templates.

Note that the entropy of IGoR's model stands for the recombination scenario entropy. The recombination machinery being degenerate several scenarios may lead to the same resulting sequence. Sequences entropy cannot be computed in closed form and must be approximated through Monte-Carlo sampling

$$H_S(\theta) = \frac{1}{Z} \sum_{\mathbf{S}} P_{gen}(\mathbf{S}) \ln P_{gen}(\mathbf{S}) - \ln Z, \quad (3.21)$$

where $Z = \sum_{\mathbf{S}} P_{gen}(\mathbf{S})$ is a normalization for finite sampling of $\mathbf{S}$ from the inferred distribution.

The sequence entropy corresponding to the different receptor chains is displayed in Fig. 3.3. Assuming TCR α and β recombinations to be independent, the total nucleotide TCR diversity is $\sim$ 77bits corresponding to $\sim 10^{23}$ equiprobable sequences. Because the generated sequences are not equiprobable the potential number of sequences is actually greater. However this number clearly indicates that an individual's composed of $\sim 10^{13}$ [12] cells repertoire is only a small sample of a large statistical ensemble. The difference between sequences and recombination entropy corresponds to convergent recombination scenarios entropy. The importance of this convergent recombination entropy for V(D)J recombination scenario assignment will be discussed in section 3.6.

3.5 CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATE

We then validated the learning algorithm on synthetic datasets. Sequences were generated in batches of 10^3 to 10^5 by IGoR with a variable error rate, using statistics inferred from 60bp DNA TRB data. IGoR's learning algorithm was then run on these raw sequences, and the resulting statistics compared to the known ground truth. We found that the inference was highly accurate for datasets of 10^5 sequences and an error rate set to its typical experimental value, 10^{-3} (Fig. 3.4a and b), and was not affected by overfitting. However, not all high-throughput sequencing datasets reach this depth, especially when restricted to unique non-productive sequences. In addition, hypermutation rates in BCRs, which IGoR treats in the same way as errors, can reach 1-10%. To as-

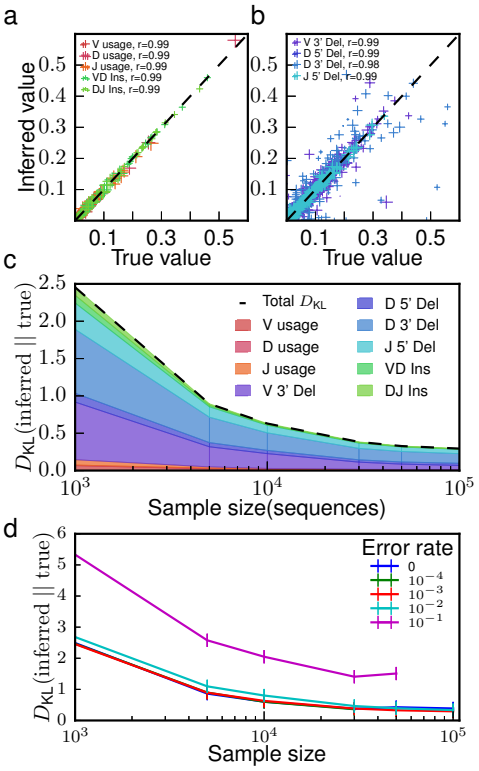

Figure 3.4: **Validation on synthetic data.** Short synthetic reads of recombined TRB or IGH sequences were generated with known recombination statistics, and given to IGoR as input to re-infer these statistics. Inference with 10^5 sequences and a typical sequencing error rate of 10^{-3} gives excellent agreement for (**a**) gene usage and insertion statistics and (**b**) deletion statistics (Pearson's r for deletions is calculated on the joint statistics of gene usage and deletion number; cross size scales with gene usage). (**c**) Discrepancy between true and inferred values of the recombination statistics, measured by the Kullback-Leibler divergence, as a function of the number of unique sequences in the sample, and decomposed according to the features of the recombination scenario. (**d**) Same as (**c**), for increasing rates of sequencing errors or of hypermutations.

sess how these limitations affect accuracy, we calculated the Kullback-Leibler divergence (a non-parametric measure of difference between probability distributions defined in section A.2.2) between the true distributions and the inferred ones, for varying sizes of datasets and error rates. For an error rate of 10^{-3}, $\sim$ 5000 unique out-of-frame sequences (which can be obtained from less than 2ml of blood with current mRNA sequencing technologies [129]) were sufficient to learn an accurate model of TRB (Fig. 3.4c), with the majority of the estimation error due to deletion profiles (which account for the majority of parameters). Increasing the error rate has little effect up to rates of 10^{-2}, but significantly degrades accuracy for typical hypermutation rates, 10^{-1} (Fig. 3.4d), with the gene usage distribution affected the most (Fig. C.5). This suggests that the recombination statistics of BCRs should be inferred using sequences from naive, non hypermutated cells (as we did in Fig. 3.2).

3.6 THE "ASSIGNMENT" PROBLEM

3.6.1 *Analysis of scenario degeneracy*

By considering all possible recombination scenarios for each sequence, our approach departs significantly from most existing methods, whose goal is to find the most likely one. To assess how often the most plausible scenario is the correct one, we analyzed synthetic sequences for which the generation scenario is known. For each generated sequence, we used IGoR's analysis mode to enumerate the set of scenarios that were consistent with the nucleotide sequence, and ranked them according to their likelihood. Fig. 3.5a shows the distribution of the rank of the true recombination scenario for TRB and IGH synthetic data. The maximum-likelihood scenario is not the correct one in 72% of IGH sequences and 85% of 60bp TRB sequences. The distributions have long tails, meaning that a substantial fraction of sequences have a very large recombination degeneracy.

We then estimated how many scenarios, ranked from most likely to least likely, were needed to explain a given fraction f of the total sequence likelihood. The distributions of this number across 100,000 generated sequences are shown in Fig. 3.5b for various values of f (see Fig. C.6 for the equivalent plot for TRB data). To enumerate the correct scenario with $f = 95\%$ confidence requires to include at least 30 to 50 scenarios. This analysis indicates that many scenarios need to be considered to correctly characterize the generation process.

IGoR outputs the probability of generation of the processed sequences, by summing the probabilities of all their possible scenarios, which deterministic assignment methods cannot do. It was shown that this generation probability was predictive of sharing properties between healthy individuals [115, 129] as will be discussed in chapter 5. This functionality could be used as a useful indicator of convergent recombination in studies attempting to identify antigen-specific or auto-immune related sequences from large clinical datasets.

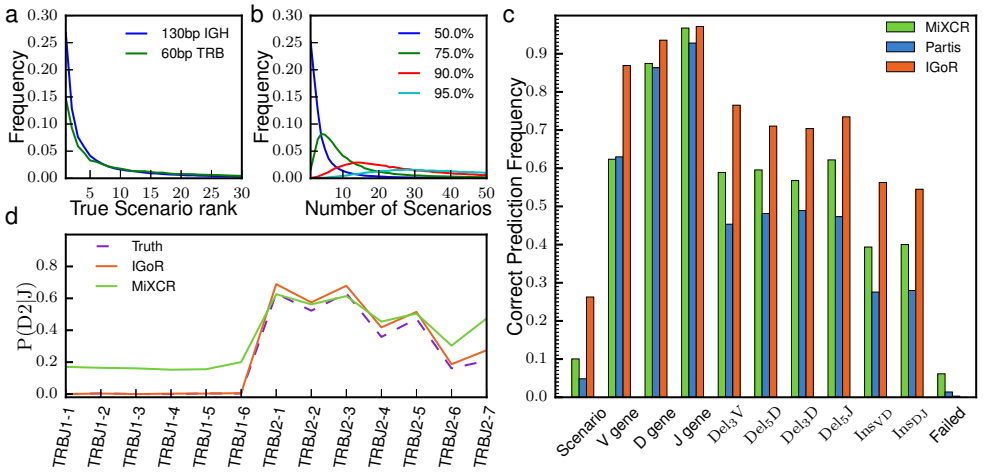

Figure 3.5: **Probabilistic analysis of putative recombination scenarios and comparison to existing methods.** Synthetic 130-bp reads of recombined IGH sequences and 60-bp reads of TRB sequences were generated with a $5 \cdot 10^{-3}$ error rate, and processed for analysis by IGoR and two existing methods, MiXCR [15] and Partis [135]. IGoR ranks putative scenarios by descending order of likelihood. (**a**) Distribution of the rank of the true scenario as called by IGoR. Note that the best-ranked (maximum-likelihood) scenario is the correct one in less than 30% of cases. (**b**) Distribution of the number of scenarios that need to be enumerated (from most to least likely) to include the true scenario with 50% (blue), 75% (green), 90% (red), or 95% (cyan) confidence. (**c**) Frequency with which IGoR, MiXCR and Partis call the correct scenario of recombination as the most likely one ('scenario'), as well as each separate feature of the scenario ('V gene,' etc.). 'Failed' corresponds to sequences for which the algorithm did not output an assignment. (**c**) Usage frequency of TRB D gene conditioned on the J gene, inferred by the IGoR and MiXCR (Partis does not handle TCR sequences). IGoR recovers the physiological exclusion between D2 and J1, while MiXCR does not.

3.6.2 *Comparison to other methods*

We compared our method to two representative state-of-the-art algorithms: MiXCR [15], an efficient assignment tool that finds the best matching germline genes through deterministic alignment, and Partis [135], a BCR-specific tool encoding an HMM that uses maximum likelihood to find the most plausible scenario. 130 base-pair IGH sequences were synthetized in silico from a data-inferred model using IGoR's generation mode. We then assigned recombination scenarios using MiXCR, Partis and IGoR, and compared them to the true scenarios with which sequences were generated. In IGoR's and Partis' case, the model parameters were learned from the generated dataset to mimick the analysis of real data. Fig. 3.5c shows the performance of the three methods in assigning the correct scenario of recombination. IGoR performs about 2.5 times better than MiXCR and Partis in predicting the complete recombination scenario, as well as each of its individual components. Note that Partis does not include palindromic insertions, which both IGoR and MiXCR treat by appending a short palindromic sequence at the end of each germline segment; restricting the analysis to sequences generated without palindromic insertions makes Partis' performance comparable to that of MiXCR (Fig. C.7).

Next, we compared the recombination statistics learned by the three methods to the true statistics used to generate the data. For MiXCR and Partis, we built the distribution of recombination events assigned to each sequence, while for IGoR these distributions were inferred using Expectation-Maximization, as explained before. All three methods yield similar statistics for V and J gene usage and deletion profiles (see Fig. C.8). However, the dependency between D an J usage in TRB is correctly captured by IGoR but not by the other methods (Fig. 3.5d). TRB D and J genes are organised in two clusters, one containing D1 followed by genes of the J1 family, the other containing D2 followed by genes of the J2 family (see Fig. 1.2 p. 14). Because of this organisation, D2 cannot be recombined with genes from the J1 family [114]. MiXCR assigns 20% of impossible D2-J1 recombination events to sequences (note that Partis does not process TCRs). By constrast, IGoR correcly learns the rule by assigning zero frequency to these impossible D-J pairs. The same results are obtained directly on real data (see Fig. C.9). Finally, IGoR accurately reconstructs the distribution of insertions, while the other methods systematically overestimate the probability of zero insertions (Fig. C.8a and b).

3.7 DOUBLE DS INSERTION AND UNIVERSAL INSERTION DISTRIBUTION

Section 3.3 showed that TCRβ VD, DJ and TCRα VD insertion profiles are identical while BCR heavy chains VD and DJ insertions profile are broader. Although these longer junctional regions have already been reported, the same TdT enzyme [114] introducing non templated insertions acts at every loci. It is thus not clear why BCRs would exhibit broader insertion distribution.

Several studies have already reported the existence of recombinations events in BCR heavy chains containing multiple tandem D genes in the junctional region [21, 85, 88, 148], thus violating the 12-23 recombination rule. However,

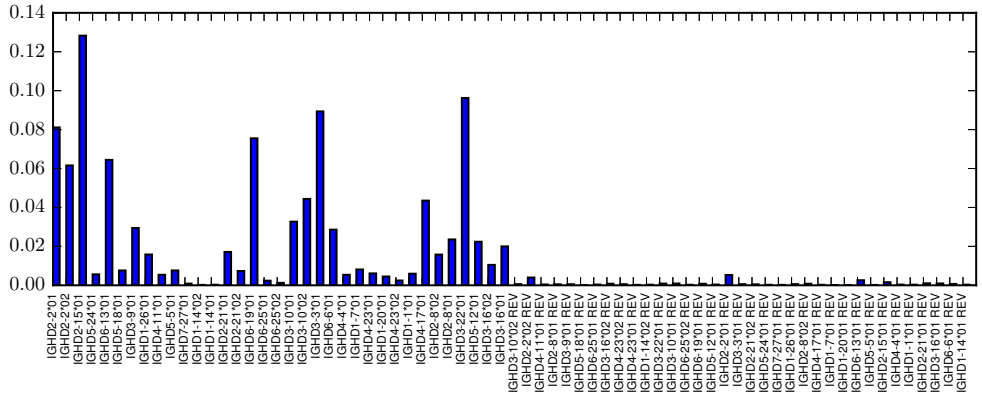

Figure 3.6: **BCR reversed complement Ds usage.** By appending the reversed complement of each D gene to the list of D genes we have tested the occurrence of reversed Ds during the VDJ recombination process. We can see that although some reversed complement Ds can be observed the effect is minor.

because of D shortness and deletions it is challenging to distinguish those rearrangements from random insertions. To test this, we computed the frequency with which one could deterministically align (with the Smith-Waterman algorithm) two non-overlapping Ds over at least 10 consecutive nucleotides, between the best V and best J alignments in BCR heavy and TCR beta chains sequencing data. We then compared these results with predictions from IGoR's synthetic sequences generated with models allowing for a single D segment learned on the very same datasets. We found 5 times more double-D assignments in IGH data than in the control, validating the findings of [88]. In contrast, the same analysis performed on TRB showed no significant presence of tandem Ds. Future versions of IGoR should include the possibility of including multiple D rearrangements and possibly uncover the same universal insertion distribution for all loci.

Using IGoR we learned a recombination model including the possibility of reversed D gene usage. Overall we found that only 3 reversed D genes appeared for a total of 1% of recombined sequences (Fig. 3.6) and conclude that if existing incorporation of reversed Ds is a minor feature. Inspired by TCR DJ association we also checked whether a similar pairing could be observed for BCRs (Fig. 3.7), and although we do find some correlations, no clear pattern as for TCRs could be observed. Such correlations have already been reported, and it has been hypothesized that they originate from the distances separating the D and J gene [81]. A more in depth analysis of these correlations could give precious insights on the recombination biophysical process.

3.8 Probability of generation

From the inferred models of recombination we are can sample the distribution of probability of generation (Fig. 3.8). As already described in Refs. [48, 50, 115] these distribution span many orders of magnitude and should serve as a null model for over representation of some sequences. As we will discuss in

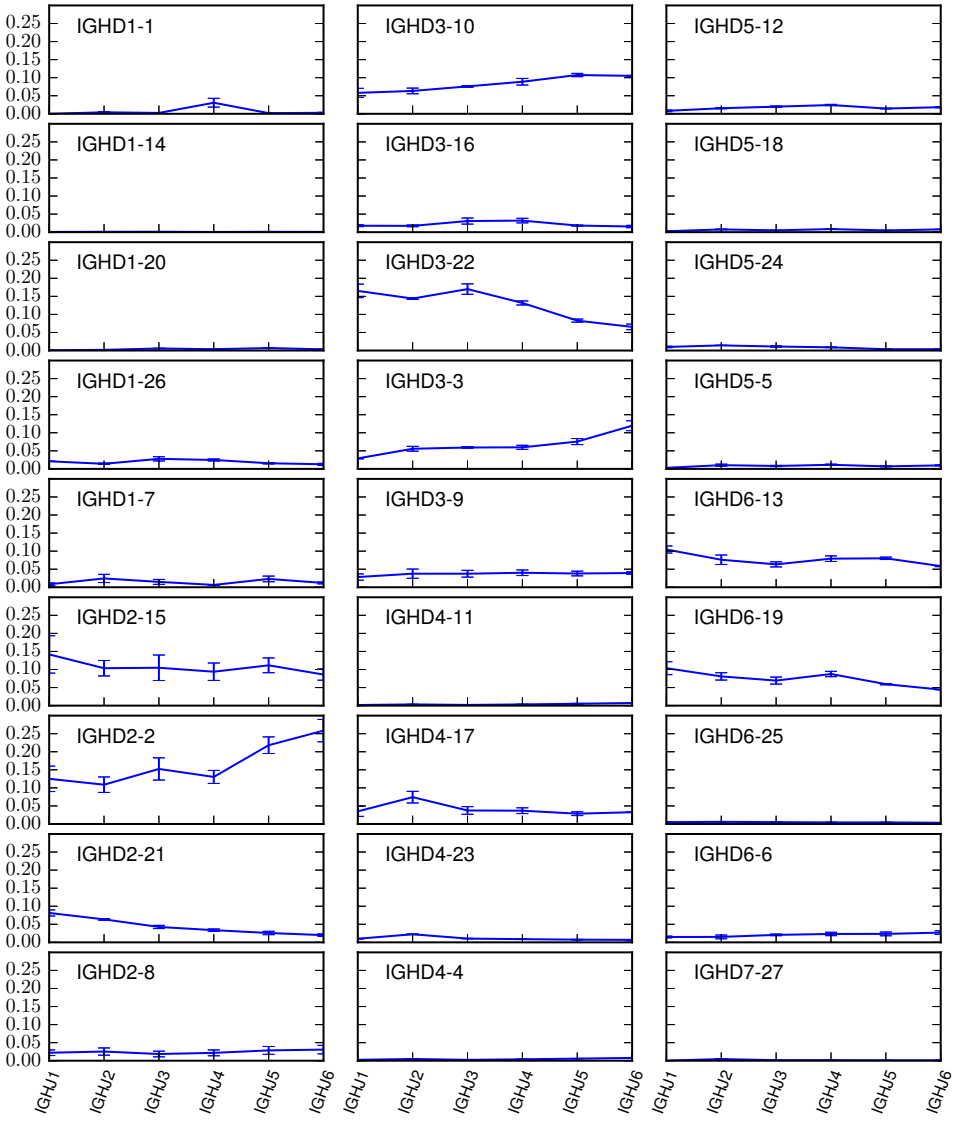

Figure 3.7: **BCR D-J association.** As we have shown the D,J pairing rule for TCRs in Fig. 3.5d, we plot P(D|J) for each pair. Unlike TCRs, BCRs do not seem to exhibit such a clear coupling.

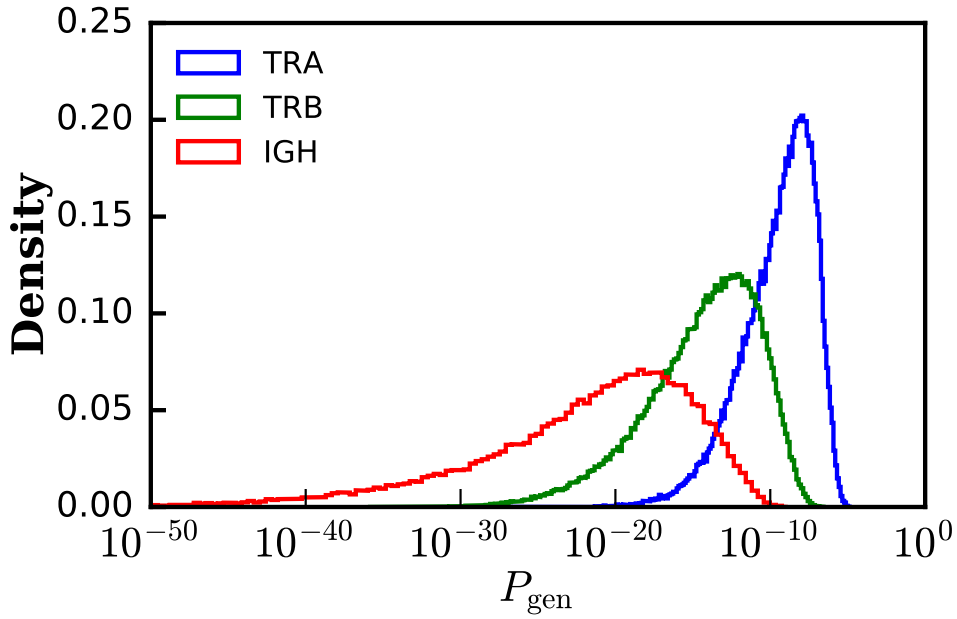

Figure 3.8: $\mathrm{P}_{\mathbf{gen}}$ **distribution for the different chains.** By randomly generating sequences according to the inferred models of recombination we can reproduce the probability of generation for TCRα (blue), TCRβ (green) and BCR heavy (red) chains

chapter 5 these distributions are also good predictors for the number of shared sequences between two samples, and we shall use them as a null hypothesis for over sharing between twins.

IGoR can in principle calculate the generation probability of any sequence. However, highly hypermutated sequences pose an additional challenge because the ancestral (unmutated) recombined sequence itself is sometimes not known with certainty. Indeed, although the probability of generation of a sequence without errors or hypermutations is well defined (section 3.2.3), computing the probability of generation of a mutated sequence[1], before mutations occurred, is strictly speaking not possible because that sequence is not know with certainty. However, we can compute a good approximation for it, and we can also calculate its distribution across sequences.

To approximate $\mathrm{P}_{\mathrm{gen}}(\mathbf{S})$ from a noisy or hypermutated sequence $\mathbf{R}$, we take its geometric average weighted by the probability of the recombination product $\mathbf{S}$:

$$\ln \mathrm{P}^{*}_{\mathrm{gen}}(\mathbf{R}) \approx \sum_{\mathbf{E}} \mathrm{P}(\mathbf{E}|\mathbf{R},\theta) \ln \mathrm{P}_{\mathrm{gen}}(\hat{\mathbf{S}}(\mathbf{E}),\theta), \tag{3.22}$$

with $\mathrm{P}(\mathbf{E}|\mathbf{R},\theta) = \mathrm{P}_{\mathrm{recomb}}(\mathbf{E},\theta)\mathrm{P}_{\mathrm{err}}(\mathbf{R}|\hat{\mathbf{S}}(\mathbf{E}),\theta)/\mathrm{P}_{\mathrm{read}}(\mathbf{R},\theta)$. Alternatively, one can take the generation probability of the most likely recombination product:

$$\mathrm{P}^{*}_{\mathrm{gen}}(\mathbf{R}) \approx \mathrm{P}_{\mathrm{gen}}(\mathbf{S}^{*},\theta), \tag{3.23}$$

1 Or to a lesser extent sequences from an error-prone sequencing experiment.

where $\mathbf{S}^* = \text{argmax}_{\mathbf{S}} P(\mathbf{S}|\mathbf{R}, \theta)$.

Using synthetic data, we checked the performance of these two estimators for the generation probability of individual sequences and observe that it is well predicted by this method ($r = 0.97$, see Fig. 3.9).

The distribution $\rho(x)$ of the log-probabilities of generation, $x = \log P_{\text{gen}}$, can be computed from data using:

$$\rho(x) = \frac{1}{N} \sum_{\alpha=1}^{N} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}, \theta) \delta \left[x - \ln P_{\text{gen}}(\hat{\mathbf{S}}(\mathbf{E}), \theta) \right] . \tag{3.24}$$

Note that unlike estimates for single sequences, this expression should become exact in the limit of $N \to \infty$. Using the same synthetic sequences as before we show that the generation probability distribution is accurately reproduced (see Fig. 3.10).

The precision of these estimation however relies on the correctness of the error/mutation model at hand.

This chapter outlined our probabilistic framework for V(D)J annotation and its general software implementation IGoR. Although we demonstrated its functions on human TCRα , β and BCR heavy chains, IGoR's flexible structure makes it applicable to any variable lymphocyte receptor (TCR or immunoglobulin) and species for which genomic data is available. Unlike HMMs based methods (e.g. [48, 135]), it can include a wide array of possible dependencies between the recombination events. As we have illustrated modeling these correlation, and more generally accurately modeling the actual recombination process is of importance for V(D)J assignment as our method outperforms existing ones. IGoR's model can also be adapted to handle unusual or incomplete rearrangements (D-J rearrangments, DD2/DD3 rearrangements in TCR δ chains, hybrid TRA/TRD recombinations, etc.).

Although the learning procedure could be carried on any sequence dataset, we used non-productive sequences in order to access the raw V(D)J recombination statistics and potentially some of its biophysical parameters as we will discuss in the next chapter. This allows us to derive meaningful quantities such as the recombination entropy and sequences' probability of generation. The ability to generate sequences mimicking the recombination process is also of importance both for benchmarking and providing null model datasets for sequences that did not undergo selection as we will use in chapter 5.

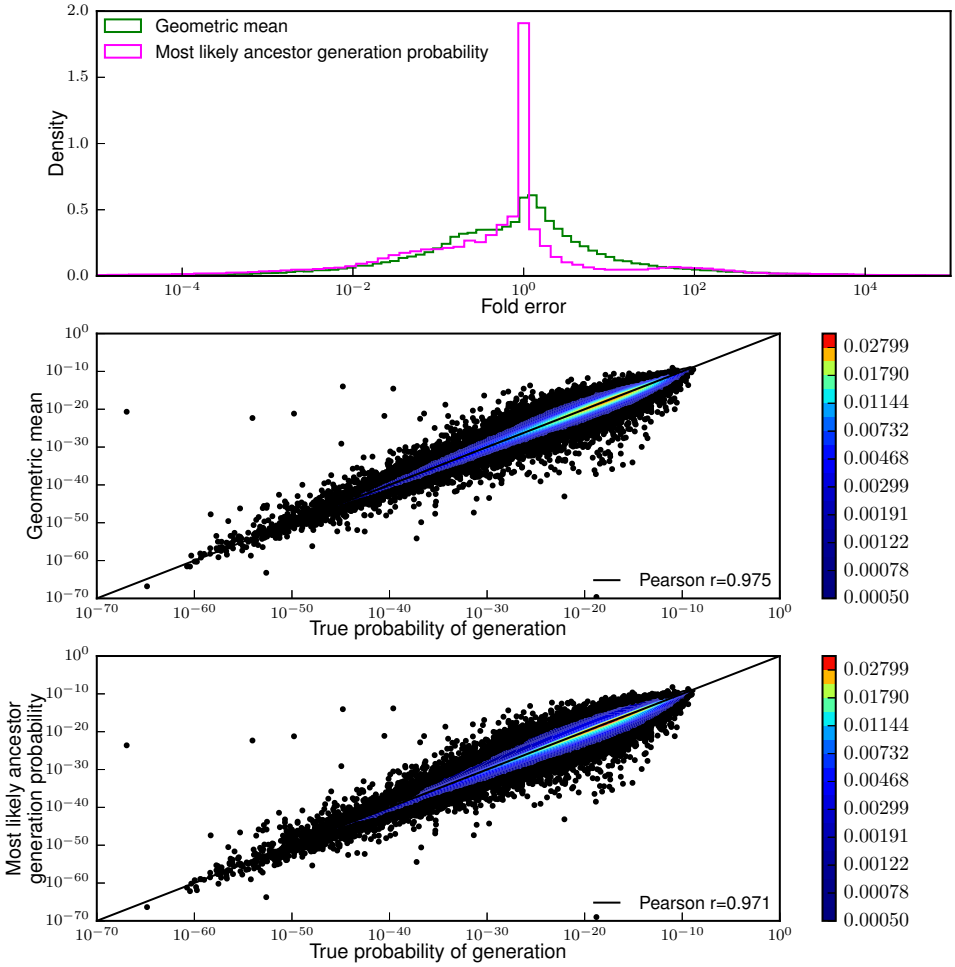

Figure 3.9: **Sequence probability of generation estimation** By generating synthetic 130bp BCR sequences from an inferred recombination model without errors we were able to compute their probability of generation P_{gen} (see SI 3.2.4). We further introduced errors in those sequences, errors whose statistics correspond to an inferred hypermutation model and computed an estimate for the probability of generation of the unmutated ancestor. We propose two different estimators: $\overline{P_{gen}}$ a geometric average of putative ancestors probability of generation weighted by it's posterior probability (green and middle) and $P_{gen}(\underset{S}{argmax} P(\mathbf{S}|\mathbf{r}))$ the probability of generation of the most likely ancestor (pink and bottom). Note that due to convergent recombination the most likely ancestor does not necessarily correspond to the sequence implied by the most likely scenario. Thus these two estimates can only be made thanks to direct exploration of recombination scenarios. Both estimators show almost perfect correlation despite the error distribution of most likely ancestor probability of generation being non symmetric.

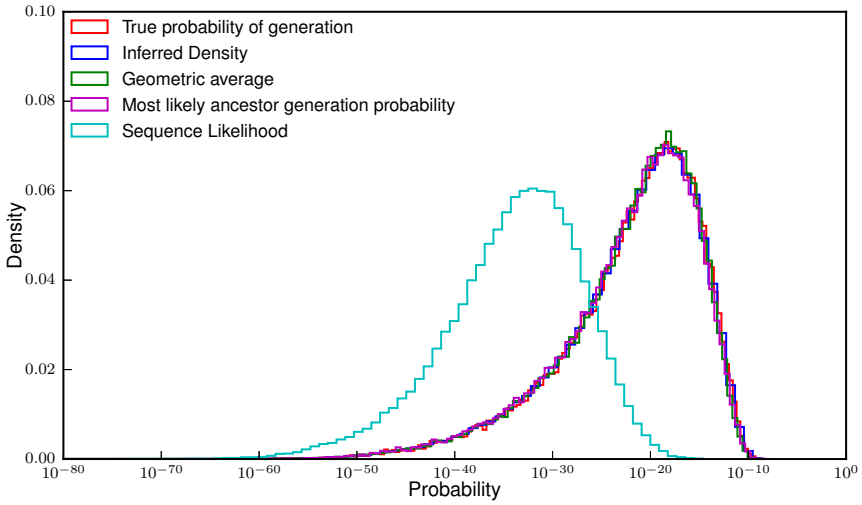

Figure 3.10: **Density of the probability of generation of sequences** We plot the distribution of probability of generation obtained from different estimators against the true distribution of generation probabilities. The true probability of generation, the geometric average and the probability of generation of the most likely ancestor are presented in Fig. 3.9's caption. The inferred density (blue) is a histogram of each sequence putative ancestors probability of generation weighted by it's posterior probability. We also plot the distribution of sequence likelihoods, that could be obtained by other methods (e.g forward algorithm) and show that it greatly differs from the distribution of generation probability.

4

ALLELIC EXCLUSION AND RECOMBINATION RESCUE

The recombination models presented in the previous chapter provide direct insights to the recombination process and unbiased sequence statistics. The obtained models and their predictions also contain information that can readily be used to extract hidden larger scale biological features with little effort. This short chapter presents two such examples, first showing how from the inferred recombination models we infer chromosome organization and relative usage and second how we use model prediction to evaluate the recombination rescue probability.

4.1 BUILDING CHROMOSOMIC ASSOCIATION

The method described in this section has been published as part of Ref. [50].

As presented in section 1.3.2 and confirmed by the results obtained in section 3.6.2, the V(D)J recombination machinery introduces long range correlations in the gene usage statistics within a single chromosome. This only explains correlations at the level of genes and not at the level of allele identity. The recombination process only allows for gene recombination within a single chromosome, so such correlations can be attributed to the assignment of a given allele to one of the two chromosomes.

By treating every allele as different genes and learning the $P(V, D, J)$ probability for producing a VDJ triplet, we can exploit observed correlations to build the underlying chromosome organization. From the organization of the different alleles on two different chromosomes, some V-D, D-J and V-J allele associations are impossible because the recombination machinery works on one chromosome or the other at a given time, never on both at the same time. Given our probabilistic approach, this should be reflected in a lower probability for inappropriate V-D-J triplets involving alleles of different chromosomes in the inferred joint $P(V, D, J)$ probability. For instance, since rearrangements happen on a single chromosome, the probability of recombining a heterozygous V allele with a heterozygous D allele on different chromosomes should be zero, up to assignment errors (Fig. 4.1). We exploit this fact to reconstruct the chromosomal organization as follows: each gene with two alleles is assigned a two state variable: each gene is marked as either heterozygous or homozygous. At this point, based on the initial list of genomic templates, each gene that has at least two candidate alleles is marked as heterozygous. An iterative procedure described below re-assigns the homo/heterozygosity parameters.

If the gene is marked as heterozygous, each allele is assigned to one of the two chromosomes or marked as erroneous, with the constraint that two alleles of the same gene cannot lie on the same chromosome and that each chromosome must be assigned an allele. If the gene is marked as homozygous, one

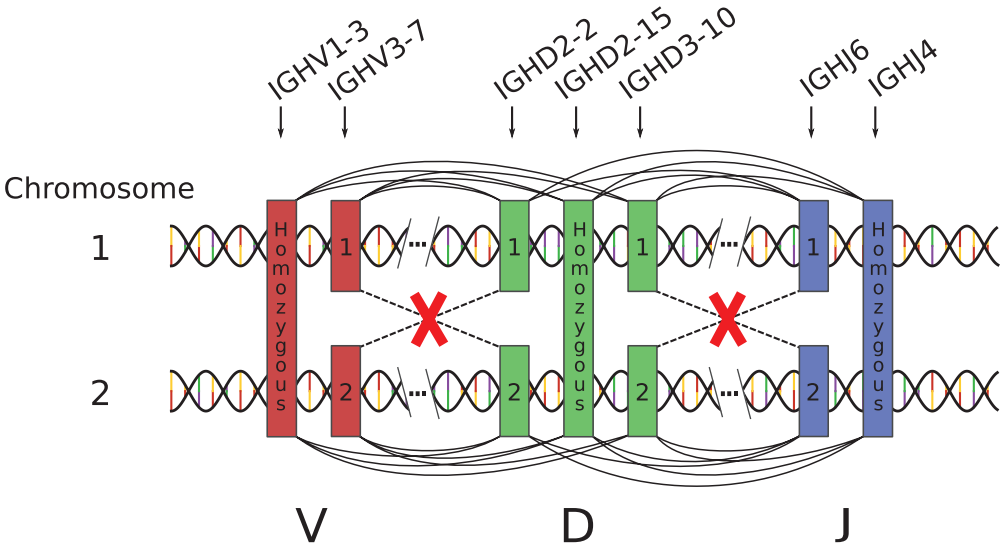

Figure 4.1: **The organization of heterozygous genes into chromo- somes can be probabilistically determined.** Every recombina- tion event ties together a V, a D, and a J gene, as indicated by the arcs drawn above and below the two chromosomes. Links that recombine alleles on different chromosomes are forbidden (red crosses). Our method gives the probability P (V, D, J) of all possible linkages between three genes (distinguishing be- tween alleles of the same gene), but does not address how the various alleles are grouped on chromosomes. We find the best chromosomal segregation by minimizing the sum of all terms in P (V, D, J) that contain forbidden links (red crosses).

of the alleles is "real", while the other is erroneous (again with the constraint that the two alleles of a given gene must be in two different states - real or erroneous.). Finding the chromosomal organization entails doing a search to find the values of these parameters that minimize the net probability (derived from the $P(V, D, J)$ distribution) of recombination scenarios involving V, D or J alleles that do not lie on the same chromosome.

In practice, all genes with two alleles are initially taken to be heterozygous and all alleles are assigned randomly to a chromosome (or erroneous state for genes with more than two candidate alleles). After initialization, a gene is chosen at random and the probability of scenarios violating the chromosomal organization is computed for the five possible states of the two alleles of this gene (heterozygous - chromosome 1, heterozygous - chromosome 2 , heterozygous - erroneous allele, homozygous - real allele, homozygous - erroneous allele) given by the previously defined two and three state variables. A change in the assignment of these parameters is accepted only if it decreases the probability of erroneous recombination events. This step is iteratively repeated until no further change is possible, thus implementing a simple hill-climbing algorithm [146]. This procedure is ensured to converge to a local minimum. Repetitions of this procedure starting from randomly chosen initial states always converge to the same final state, and we conclude that only one global minimum exists.

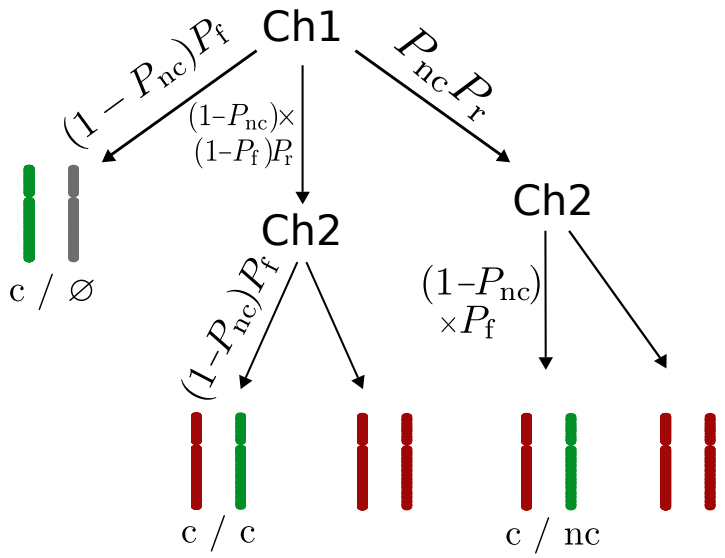

Figure 4.2: **Simplified diagram for recombination outcomes.** This is valid for chains with allelic exclusion

Because we tacitly impose diploidy, this procedure can be used to clean up genomic information for an individual by removing spurious "erroneous" alleles as in Ref. [50].

One question of interest is the relative usage of both chromosomes for V(D)J recombination and the existence of a parental imprinting[1] to allelic exclusion. With the approach described above we were able to compute this relative gene usage for BCR heavy chains. In this way we found a chromosomal organization for the two individuals that accounted for about 90% of all sequences. We can also evaluate the usage probability of the two chromosomes identified using this procedure. For both individuals, it was consistent with equal usage probability between the two chromosomes, within errors.

4.2 RESCUE PROBABILITY

This work is for the moment unpublished in the hope to gather more precise measurements with different datasets.

As introduced in section 1.3.2.2, upon failure of recombining the heavy or β loci and assembling a pre-receptor on the cell surface the immature lymphocyte can be rescued and attempt a second recombination on the untouched chromosome[2]. It is however unknown whether this rescue is systematic, and if not, how frequent it is. In this section we discuss a simple calculation to estimate this frequency.

So far we have treated coding and non-coding sequences as separate datasets in order to infer raw V(D)J recombination statistics from the latter one. How-

1 The epigenetic phenomenon by which genes are expressed in a parent of origin specific manner.
2 For the β chain there exist a possibility to be for a second recombination on the same chromosome depending the DJ cluster involved in the first one. We will however assume that this is a feature effectively learned by our models.

ever, both arise from the same sequencing experiment and as DNA sequencing is not sensitive to allelic exclusion it should output coding and non-coding sequences with the same efficiency[3]. The relative fraction of sequences of each sequencing experiment should thus reflect the efficiency of lymphocyte development to produce functional sequences in one or two attempts.

In order to extract this information we write a simple model for each recombination outcome as summarized in Fig. 4.2. This model depends on three parameters:

- P_{nc} the probability for a recombination shot to produce a non-coding sequence. This parameter can be readily estimated by generating sequences from our inferred model. This way we estimate to 27.1% the chance of obtaining a non coding TCR β chains.
- P_r the rescue probability or frequency at which a second attempt of recombination is made on the second chromosome after failure of the first. We assume this probability to be a scalar independent of the previous recombination product.
- P_f the probability of a sequence to be functional given that it is an apparently coding sequence. Because heavy and β chains are not initially tested for their ligand binding abilities, this quantity only reflects the ability of a sequence to form a functional folded pre-receptor. Here again for the sake of simplicity we will assume this probability to be a scalar while it is clear that it depends on the recombination product.

From these three parameters and the decision tree for lymphocyte fate in Fig. 4.2 we can write the expected fraction of non-productive sequences F_{nc} in a sequencing experiment as

$$F_{nc} = \frac{P_{nc}(1-P_{nc})P_rP_f}{(1-P_{nc})P_f + 2(1-P_{nc})^2(1-P_f)P_fP_r + 2P_{nc}(1-P_{nc})P_rP_f} \tag{4.1}$$

$$= \frac{P_{nc}P_r}{1+2(1-P_{nc})(1-P_f)P_r + 2(1-P_{nc})P_r}. \tag{4.2}$$

By drawing a colormap of F_{nc} (Fig. 4.3) as a function of the two unknown parameters P_r and P_f we observe that the value of P_f only weakly influence the expected fraction of non-coding sequences[4]. By overlaying the fraction of non coding sequences observed in DNA sequencing experiments as contour lines we can estimate the probability of rescue. This back of the envelope calculation would then suggest that the recombination rescue, far from being systematic, only occurs with $\sim 35\%$ chance for TCR β chains.

Because of the simplicity of the model this indirect estimate remains imprecise. With the recent development of statistically paired sequences [76, 91] we should be able to access a more direct measure for P_r by directly measuring the relative fraction of recombination end products shown in Fig. 4.2. Sadly, such experiments have been carried with RNA sequencing technologies sensitive to

This work using statistically paired sequences is currently under investigation in collaboration with T. Dupic.

3 Up to primer PCR amplification bias.
4 This could have been guessed from Eq. 4.2 as P_f is a dominated term in the denominator

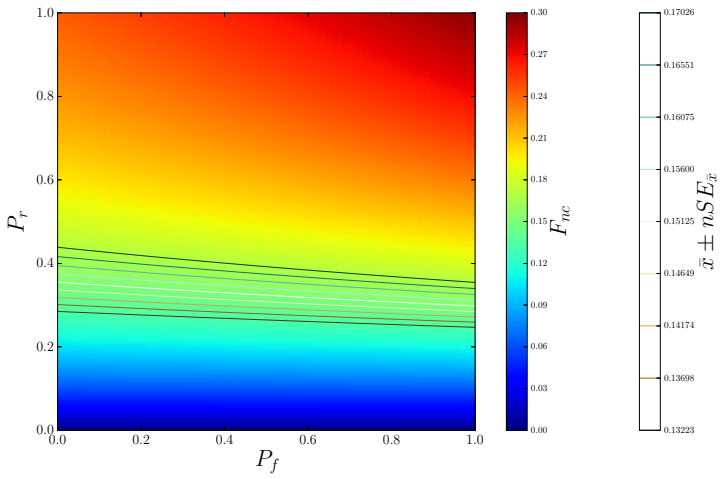

Figure 4.3: **Fraction of non coding sequences in a naive TCR dataset.**

allelic exclusion thus making the direct estimates incorrect without a proper statistical correction.

By reconstructing the chromosomic organization from joint gene usage probabilities we have shown that there is no preferential allelic exclusion for one chromosome or the other. A similar analysis for haplotype inference was conducted in Ref. [82], however relying on assignments that did not model long range correlations. The authors report frequent copy number variations in contradiction with our diploidy hypothesis and confirm the high variability in genomic information between individuals. This assumption could however be easily relaxed and a proper inference of genomic templates contained in the dataset should be conducted.

The estimation of the probability of rescue on the other hand did not seem to have been previously addressed by the community, and suggest that recombination rescue is far from being systematic. As the presented methodology is rather simple it shall be extended to other receptor chains exhibiting allelic exclusion such as BCR heavy chains.

5

INTER-INDIVIDUAL RECEPTOR SHARING.

Is an individual's ability to fight against pathogens tied to the identity of the precise set of clones constituting the individual's repertoire? As mentioned in 1.3.1 one receptor can recognize different antigens. Conversely the same antigen can be recognized by different receptors with different strength such that $20-200$ out of $\sim 10^7$ TCRs can bind the same MHC-peptide dimer [109].

While theoretical studies [104] incorporating these ingredients would suggest that the precise set of receptors is not important and that the repertoire is organized as a whole, many studies have reported "public" clonotypes shared among several individuals either in health [177, 181, 206] or disease [51, 179]. Following these theories, could these clones be shared by pure chance?

Our ability to answer this question depends greatly on our ability to quantify clonal diversity at the sequence ensemble, individual and sequencing sample levels. A number of different diversity measures (Shannon entropy, Simpson index, species richness, Chao1 [32] and 2 [33], DivE [89]) deriving from ecology have been used to quantify lymphocyte repertoires diversity from clone abundance data in sequencing experiments. All these estimations, reviewed in Ref. [110], are related to Rényi entropy and put the accent on different parts of the clone abundance distribution [111]. More importantly, Ref. [110] demonstrates the limit of these estimations on the finite amount of data exhibiting fat-tailed distributions such as power laws in repertoire sequencing data.

By learning the probability distribution underlying the V(D)J recombination process we obtain an ensemble description of unselected sequences and reduce the rare clone sampling issue[1]. We can thus compute the Simpson index of this ensemble, related to the potential diversity of V(D)J recombination, that is the probability of two independently recombined sequences to be identical by chance:

$$\langle P_{gen} \rangle_S = \sum_S P_{gen}^2(\mathbf{S}), \tag{5.1}$$

where $P_{gen}(\mathbf{S})$ is the probability of generation of sequence $\mathbf{S}$ as defined in section 3.2.3. Note that this calculation cannot be performed in closed form and is estimated via Monte Carlo sampling. Assuming cell proliferation and peripheral selection completely dominate clone abundance distributions[2], counts only reflect the frequency at which a lymphocyte functional receptor can bind to its cognate antigens. Non productive sequence counts would then also reflect functional receptor fitness and shall be discarded to retain only unique sequence information. From our recombination model, the number of shared

1 Still, under-sampling remains an issue especially when recombination statistics seem to be linked to clone abundance as suggested by the work presented in this chapter.

2 Although again, some data presented in this chapter suggest that time of generation and homeostatic state of the repertoire could play a role in abundance of species.

unique non productive sequences between two datasets should thus be predicted by $M_{pre} = |N_1|.|N_2|\langle P_{gen}\rangle_S$ where $|N_1|$ and $|N_2|$ are the number of unique non productive sequences contained in each dataset.

Provided a V(D)J recombination model one is able to compute $P_{gen}(\mathbf{S})$ the probability to generate a sequence $\mathbf{S}$ from the recombination machinery. As explained in chapter 3, this is good descriptor for sequences that did not undergo selection such as non productive sequences. Productive sequences on the other hand, have gone through several steps of functional selection (folding, central and peripheral selection) biasing their statistics that no longer represent the raw recombination. We shall call $P_{post}(\mathbf{S})$ the obtained distribution. Biologically, this distribution should accurately describe naive functional lymphocyte receptor statistics.

In Ref. [49] Elhanati and collaborators compute a generic selection factor $Q(\mathbf{S}) \geqslant 0$ per sequence defined as

$$Q(\mathbf{S}) = \frac{P_{post}(\mathbf{S})}{P_{gen}(\mathbf{S})}. \tag{5.2}$$

They propose a simple decomposition for this selection factor into

$$Q(\mathbf{S}) = Q(\boldsymbol{\tau}, V, J) = \frac{1}{Z} q_{VJ} q_L \prod_{i=0}^{L} q_{i,L}(\tau_i), \tag{5.3}$$

where $\boldsymbol{\tau}$ denotes the amino acid CDR3 of sequence $\mathbf{S}$ and L its length. The Z constant ensures normalization. The q_{VJ} coefficient is a selection factor for the joint usage of a pair of V and J genes, q_L a selection factor for the CDR3 length. Finally, $q_{i,L}(\tau_i)$ is a selection factor for the identity of the amino acid at position i for a CDR3 of length L. Because CDR1 and CDR2 loops are encoded in the V gene sequence, this model incorporates all regions responsible for antigen binding.

The parameters are inferred by comparing statistics of "productive" [3] sequences randomly generated from the recombination model and naive productive sequences from a sequencing experiment. Note that by using naive productive sequences it is implicitly assumed that $Q(\mathbf{S}) = 0$ for non productive sequences. Because V and J genes cannot be unambiguously assigned to sequencing reads, they are treated as hidden variables and the ML estimate for $Q(\mathbf{S})$ parameters is obtained via the EM algorithm.

From these selection models we can compute a Simpson index, similar to the one presented above for non-productive sequences, for the potential diversity of sequences post selection. While, as previously discussed in 3.3, the recombination statistics inferred on different individuals are almost identical, inferred selection factors may vary slightly. The predicted number of shared productive sequences is

$$M_{post} = |N_1|.|N_2| \sum_S P^{(1)}_{post} P^{(2)}_{post}, \tag{5.4}$$

3 Or at least not obviously non productive, as discussed in section 1.3.2.4 p.17

where $|N_1|$ and $|N_2|$ are now the number of productive sequences in the dataset, and $P_{post}^{(i)}$ the resulting post selection distribution with the selection model inferred on individual i[4]. As for P_{gen}, P_{post} cannot be computed in closed form and M_{post} can readily be approximated via Monte-Carlo sampling

$$M_{post} = \frac{|N_1|.|N_2|}{|S_1|.|S_2|} \sum_{\mathbf{s} \in S_1 \bigcap S_2} Q^{(1)}(\mathbf{s}) Q^{(2)}(\mathbf{s}), \tag{5.5}$$

where S_1 and S_2 are sets of respectively $|S_1|$ and $|S_1|$ sequences drawn from an inferred recombination model.

These estimators have proven to be accurate predictors [49, 115] of the observed number of shared sequences between individuals[5] and represent good validation of our models.

This chapter recapitulates how we used these tools to study how persisting fetal clonotypes influence repertoire overlap among twins and unrelated individuals. *It has been published in Ref. [129].*

5.1 ABSTRACT

The diversity of T-cell receptors recognizing foreign pathogens is generated through a highly stochastic recombination process, making the independent production of the same sequence rare. Yet unrelated individuals do share receptors, which together constitute a "public" repertoire of abundant clonotypes. The TCR repertoire is initially formed prenatally, when the enzyme inserting random nucleotides is downregulated, producing a limited diversity subset. By statistically analyzing deep sequencing T-cell repertoire data from twins, unrelated individuals of various ages, and cord blood, we show that T-cell clones generated before birth persist and maintain high abundances in adult organisms for decades, slowly decaying with age. Our results suggest that large, low-diversity public clones are created during pre-natal life, and survive over long periods, providing the basis of the public repertoire.

5.2 INTRODUCTION

The adaptive immune system relies on the diversity of T-cell repertoires to protect us from many possible pathogenic threats. Each T cell expresses on its surface many copies of a unique T-cell receptor (TCR), which engages with antigenic peptides – from self or foreign proteins – presented by other cells through their Major Histocompatibility Complex (MHC) molecules. The binding strength between the TCR and the peptide-MHC complex, which is typically weak for self peptides, and strong for some foreign peptides, is a major factor in determining the onset of an immune response. Since each TCR is only

4 Note that without assumption on universal V(D)J recombination statistics we can also learn private models for it. This is what will be done in the rest of this chapter in order to discard any genetic basis for twin receptor sharing.

5 Up to a multiplicative factor for some sequencing strategies.

specific to a small fractions of the possible peptides, the body needs to maintain a very large diversity of TCRs to be able to recognize any possible foreign peptide from pathogens. Understanding how this diversity is generated, and how it develops and matures with age, is thus paramount to understanding adaptive immunity.

TCR diversity is produced by the V(D)J recombination machinery which generates the repertoire *de novo* in each individual. Repertoire diversity is encoded not only in the set of specific receptors expressed in a given individual, but also in their relative abundances – the number of T-cells expressing each unique TCR – which can differ by orders of magnitude. These differences are in part due to antigenic stimulation (infection, vaccination), implying that clones increase their sizes in response to common or recurring infections. Despite this great diversity, different individuals—regardless of their degree of relatedness—do express a subset of the exact same receptors, called the *public* repertoire [179]. This overlap is often interpreted as the convergence of individual repertoire evolutions in response to common antigenic challenges [98]. Indeed, some public TCRs are known to recognize common pathogens such as the cytomegalovirus (CMV) or the Epstein-Barr virus (EBV) [107]. However, this interpretation is challenged by the fact that these two properties—large differences in clone sizes and public repertoires—are also observed in naive repertoires, for which antigenic stimulation is not expected to be important [109, 120].

An alternative explanation for public clones, which does not invoke convergent repertoire evolution, is that both abundant and public receptors are more likely to be produced by rearrangement, and just occur by coincidence [179, 181]. This idea is backed by some compelling evidence. First, the amount of clonotype sharing between pairs of individuals can be accurately predicted in both naive and memory pools from statistical models of sequence generation [49]. Second, the likelihood that a clonotype sequence is shared by individuals has been reported to correlate with its abundance [181, 206]. However the origin of this correlation remains elusive. In addition, public clonotypes often have few or no randomly inserted N nucleotides, which limits their diversity [181]. Terminal deoxynucleotidyl transferase (TdT), the enzyme responsible for N insertions, is inactive in invariant T-cell subsets [178] and in some fetal T-cell clones. These subsets could contribute to the emergence of the public repertoire. Another confounding factor is the ageing of repertoires, and the concomitant loss of diversity, which is expected to affect the structure of clonal abundances as well as the repertoire's sharing properties. How do all these effects shape the structure and diversity of TCR repertoires, and control their functional capabilities? Here we propose and test the hypothesis that a sizeable fraction of public clonotypes are created before birth. These clonotypes have low diversity because of reduced TdT activity, making them more likely to be shared among unrelated invididuals. Their large abundances, due to reduced homeostatic pressures in the early stages of repertoire development, allow them to survive over long periods.

5.3 RESULTS

5.3.1 *Clonotype sharing between individuals*

We first examined in detail the question of clonotype sharing between individuals. Each TCR is a heterodimer made of two chains encoded by two distinct genes. Each gene is formed in the thymus by assembling together two or three gene templates from a finite set of germline segments – V and J segments for the α chain, and V, D and J segments for the β chain. In addition to the large diversity created by the combinatorial choice of germline segments, further diversity is produced by randomly deleting base pairs from the joining ends of the segments, and by inserting random non-templated (N) base-pairs at each junction. Each chain forms three loops, called Complementarity Determining Regions (CDR), which come in contact with the peptide-MHC complex during recognition. The first two loops, CDR1 and CDR2, are encoded in the germline V gene and are thought to interact mostly with the MHC. By contrast, the CDR3 concentrates most of the diversity, as it covers the junctions between the germline segments. The CDR3 interacts with the peptide directly, and is thus believed to play the biggest role in the recognition of foreign peptides.

After recombination, receptors are tested and selected for function and lack of auto-reactivity. The recombination mechanism frequently produces non-functional (also called nonproductive) receptor sequences, typically containing frameshifts or stop-codons. If the recombination result of the first chromosome is non productive, the second chromosome will recombine. In case this second recombination is successful, the cell will contain two recombined genes—one productive and one nonproductive. To avoid confounding effects due to convergent selection (both selection in thymus and clonal expansion in response to infection), we first focused on out-of-frame receptor sequences, which are non-productive and hence must result from these first unsuccessful recombination events. Because the cells that contain them owe their selection and survival to the productive gene on the second chromosome, these out-of-frame sequences give us direct insight into the raw V(D)J recombination process [115, 141], free of clonal selection effects. The number of shared clonotypes between two sets of clonotypes, or clonesets, is approximately proportional to the product of the cloneset sizes [115, 159, 206]. We call the ratio of the two the normalized sharing number. In the regime of rare convergent recombination, this number is equal to the probability that two independent recombination events give the same sequence; it is thus independent of the cloneset sizes, and provides an appropriate measure of sharing for comparing different pairs of datasets with different sequencing depths. Under the assumption that sharing occurs by pure chance, only due to convergent recombination, this number can be predicted using data-driven generative probabilistic models of V(D)J recombination accounting for the frequencies of the assembled V, D, and J gene segments and the probabilities of insertions and deletions between them [48, 49, 103, 115]. We can estimate sharing either of the entire nucleotide chain (alpha or beta), or of the CDR3.

5.3.2 *Twins share more clonotypes than unrelated individuals*

Genetically identical individuals may be expected to have more similar recombination statistics due to similar recombination enzyme biases [66, 71, 131, 144, 182, 206], and therefore share more sequences. To assess these genetic effects, we looked at the sharing of TCR alpha and beta-chain receptor repertoires between three pairs of monozygous twins (6 individuals). We synthesized cDNA libraries of TCR alpha and beta chains from the donors' peripheral blood mononuclear cells and sequenced them on the Illumina HiSeq platform (see Fig. B.1 p. 136 and section B.1 p.127). For each pair of individuals, the normalized number of shared out-of-frame alpha sequences was compared to the prediction from the recombination model trained on the out-of-frame repertoire of each individual, as shown in Fig. 5.1 (see also Fig. B.2 p. 137 for similar results on sharing of CDR3 nucleotide sequences). Sharing in unrelated individuals (the 12 non-twin pairs among 6 individuals, black circles) was well predicted by the model (Pearson's $R = 0.976$), up to a constant multiplicative factor of 2.07, probably due to differences in effective cloneset sizes. While twins did share more sequences than unrelated individuals (the 3 twin pairs, red circles), this excess could not be explained by their recombination process being more similar. The model prediction was obtained by generating nucleotide sequences from models inferred using each individual's cloneset as input [48, 103], mirroring their specific recombination statistics (see section B.1 p.127). The normalized sharing number departed significantly from the model prediction only in twins, calling for another explanation than coincidence in that case. The same result was obtained for beta out-of-frame CDR3 nucleotide sequences (Fig. B.3 p. 138), although less markedly because of a lower signal-to-noise ratio due to smaller numbers of shared sequences. Most of beta out-of-frame nucleotide sequences shared among the highest-sharing twin pair associated with CD8 CD45RO+ (memory) phenotype in both individuals. This observation is surprising, because the non-functionality of these sequences excludes convergent selection as an explanation for it (see S1 Text for details).

We then examined the sharing of in-frame nucleotide CDR3 sequences. Most of in-frame sequences are functional, and have passed thymic and peripheral selection. Since these selection steps involve genetically-encoded HLA types (the type of MHC that cells express) and are therefore expected to be similar in related individuals, we wondered whether the functional repertoires of twins also displayed excess sharing. Remarkably, we found some excess sharing in the in-frame beta repertoire (Fig. B.4 p. 139), but none in the in-frame alpha repertoire (Fig. B.5 p. 140). However, the failure to observe excess sharing in this last case can be explained by the much higher expected number of shared nucleotide sequences in the alpha in-frame repertoire (due to both in-frame sequences being more numerous than out-of-frame ones, and to the lower diversity of alpha chains compared to beta chains) which could mask this excess in twins (see S1 Text).

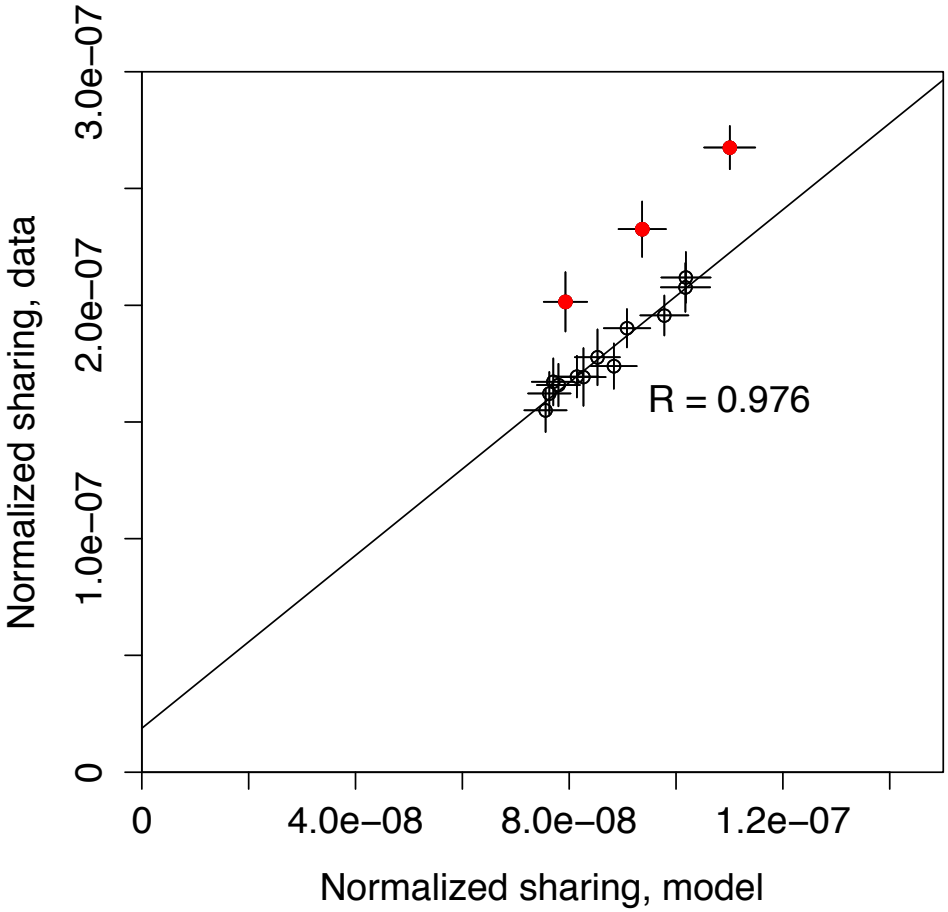

Figure 5.1: **TCR out-of-frame repertoire sharing in monozygous twins is higher than in unrelated individuals, or than predicted by stochastic models of recombination.** The number of shared out-of-frame alpha TCR clonotypes between all 15 pairs among 6 donors consisting of 3 twin pairs (ordinate) is compared to the model prediction (abscissa). To be able to compare pairs of datasets of different sizes, the sharing number was normalized by the product of the cloneset sizes. The three outstanding red circles represent the twin pairs, while the black circles refer to the 12 pairs of unrelated individuals among the 6 twins. The model prediction is based on a generative stochastic model of VJ recombination [48, 103], inferred separately for each donor to account for differences between individuals. It agrees well with the data from unrelated individuals up to a common multiplicative factor, but systematically underestimates sharing in twins. Error bars show one standard deviation.

5.3.3 *Low generation probabilities of excess shared clonotypes between twins suggest in utero T cell trafficking*

To investigate the origin of excess sharing between twins, we looked at the statistical properties of shared alpha out-of-frame nucleotide sequences from Fig. 5.1. Shared clonotypes between non-twins, which happen by coincidence, should have a higher probability P_{gen} to have been produced by V(D)J rearrangement compared to non-shared clonotypes. Indeed, the distribution of P_{gen} among shared sequences, plotted in Fig. 5.2, can be calculated from the probabilistic model of generation (blue curve), and the prediction agrees very well with the data between non-twins (red curves). By contrast, shared sequences between twins deviate from the prediction (green curve), especially in the tail of low-probability sequences, but are consistent with a mixture of $18 \pm 3\%$ of regular sequences (black curve), and the rest of coincidentally shared sequences (blue curve). These numbers agree well with the excess sharing in twins, which amounts to $17\% \pm 3\%$ of non-coincidentally shared sequences, as estimated from Fig. 5.1. Nucleotide sequences shared between twins also have higher numbers of insertions and are therefore longer than those shared between unrelated individuals or according to the model (Fig. B.6 p. 141, $p = 2 \cdot 10^{-8}$, two-sided t-test) – a trend that is even more pronounced in memory cells (Fig. B.7 p. 142, $p < 10^{-16}$). Note these observations about recombination probabilities and the number of insertions are related: sequences with many insertions each have a low generation probability because of the multiplicity of inserted nucleotides.

Taken together, these observations support the existence of another source of shared sequences than coincidence in twins. Since the sharing of cord blood between twins is the only natural instance when the immune systems of two individuals share cells, we propose that the increased sharing of private TCRs between identical twins dates back to the sharing of cord blood cells, and that these shared clones persist into late age. This persistence of fetal clonotypes could be due to the long lifetime of the exchanged naive clones. Alternatively, long persistence could be achieved by the independent transition to memory of the shared clones in both twins.

5.3.4 *Sequences with no N insertions are enriched among abundant naive clonotypes in cord blood and in young adults*

To verify the hypothesis that clones formed during fetal life persist over long periods, we now turn to the analysis of data from unrelated individuals. We characterized the in-frame beta-chain repertoire of human cord blood and also three healthy non-twin adult donors of different ages (see Materials and methods and S1 Text). One feature of the rearranged chains is the number of insertions at the junctions between the gene segments (VD and DJ in the case of beta chains). We ranked beta TCR clonotypes from human cord blood data by decreasing abundances and plotted the mean number of insertions (inferred iteratively and averaged over groups of 3000 clonotypes, see S1 Text), as a function of this abundance rank (Fig. 5.3A). The most abundant clones in

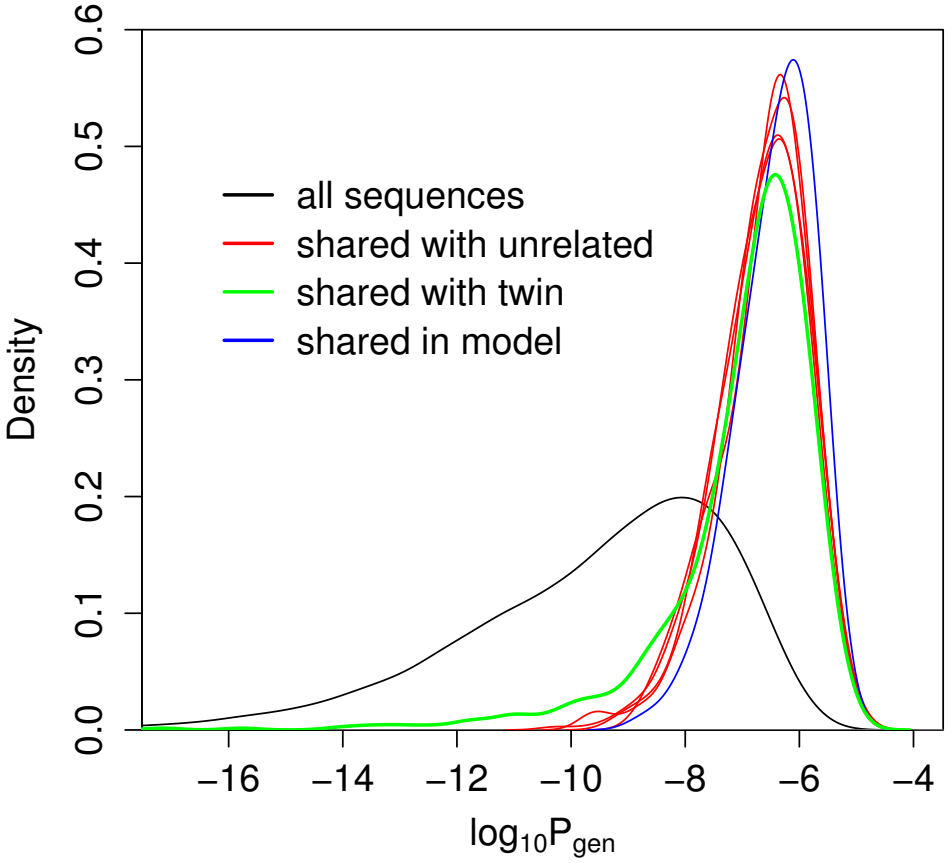

Figure 5.2: **TCR nucleotide sequences shared between twins are statistically different from sequences shared between unrelated individuals.** Distribution of $\log_{10} P_{gen}$, with P_{gen} the probability that a sequence is generated by the VJ recombination process, for shared out-of-frame TCR alpha clonotypes between one individual and the other five. While the distribution of shared sequences between unrelated individuals (red curves) is well explained by coincidental convergent recombination as predicted by our stochastic model (blue), sequences shared between two twins (green) have an excess of low probability sequences: 31 sequences with $\log_{10} P_{gen} < -10$. For comparison the distribution of P_{gen} in regular (not necessarily shared) sequences is shown in black.

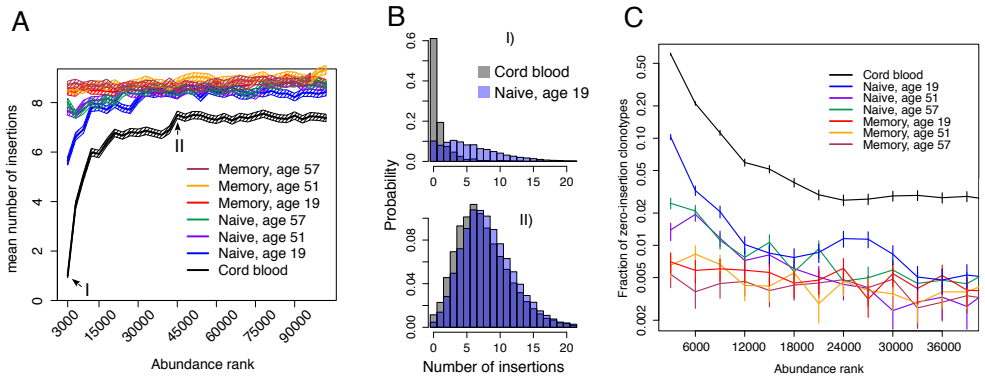

Figure 5.3: **The number of inserted nucleotides in in-frame TCR beta clonotypes depends on their abundance. A**. Mean numbers of insertions were obtained by analysing groups of 3000 sequences of decreasing abundance. Clonotypes from the cord blood (black) show a strong dependence on abundance, with high-abundance clones having much fewer insertions than low-abundance ones. Clonotypes in a young adult naive repertoire (blue) show a similar but less marked trend. Naive clonotypes in older adults (violet and green) show an even weaker trend. Adult memory samples of all ages show no dependence at all (red, yellow and maroon). Error bars show 2 standard errors. **B**. Probability distributions of the number of insertions in two rank classes, for young naive and cord-blood samples (ranks 1-3000 on top, ranks 45001-48000 on bottom). For high-ranking sequences, the probability of having zero insertions is high both for adult naive and cord blood samples. For middle-ranking sequences, the probability of 0 insertions is much lower, and the distributions are similar between adult naive and cord-blood samples. **C**. Fraction of clonotypes with zero insertions for different abundance classes. Error bars show one standard deviation. We present the analysis for independently published cord blood donors and different bin sizes in Fig. B.11 and Fig. B.10 p. 145 respectively.

cord blood had markedly smaller numbers of insertions (black line). The naive repertoire of a young adult (blue line) showed a much weaker dependence on abundance than the cord blood repertoire, but followed a similar trend. The dependence was even further reduced in older adults (purple and green lines). Interestingly, the number of insertions in the beta chains of the adult memory repertoire (red, orange and maroon lines) did not depend of the abundance of these cells. This observation can be explained by the resetting of the size of memory clones following an infection, erasing features of the abundance distribution inherited from fetal life. Looking more closely into the distribution of the number of insertions (Fig. 5.3B) reveals that low mean numbers of insertions are associated with an enrichment in clonotypes with zero insertions. Accordingly, the fraction of naive zero-insertion sequences generally decreased with abundance rank (Fig. 5.3C), with again a stronger dependency in cord blood and young adults. Fewer numbers of insertions in the cord blood are expected because TdT, the enzyme responsible for random insertions, is initially strongly downregulated in prenatal development [7, 62]. This enrichment in low-insertion sequences persists and shows weak signatures in the adult naive repertoire, suggesting long lifetimes of cord blood clonotypes (although not necessarily of individual cells).

5.3.5 *Abundant clonotypes with no N insertions decay slowly with age, but faster than the attrition of the naive cell pool*

The enrichment of zero-insertion sequences in large clonotypes of young people, relative to the baseline of zero-insertion clonotypes produced in adulthood, can be used to verify the hypothesis of long lived fetal clonotypes originating from the cord blood. Analysing publicly available TCR beta repertoire data from individuals of different ages [23, 24], we observed a slow decay of abundant zero-insertion clonotypes in the unpartitioned repertoire (memory plus naive) with age, with decay rate of 0.027 ± 0.009 yr^{-1}, or a characteristic time of 37 years (Fig. 5.4). However, the excess of abundant TdT- clonotypes of fetal origin only affects naive cells (Fig. 5.3A), whose relative fraction in the repertoire is also known to decrease with time [23]. To assess the importance of this confounding effect, we fit an exponential decay model for the percentage of naive cells measured in same donors using flow cytometry (see S3 Table) and found a characteristic decay rate of 0.015 ± 0.002 yr^{-1}, or a decay time of 67 years. The red curve in Fig. 5.4, which shows the expected decay of zero-insertion clonotypes if it had been solely caused by the decay of the naive pool, does not agree with the data. Although the decay of naive cells within the top 2000 clonotypes could in principle be faster than in the overall T-cell population, we did not observed such an effect in the three individuals for which we have data partitioned into memory and naive clonotypes (see S1 Text I.G). Therefore, the attrition of the naive pool alone cannot explain the decrease of zero-insertion clonotypes, which we attribute instead to the progressive extinction of clones of fetal origin combined with their gradual replacement by newly generated naive cells. This is consistent with the hypothesis that excess clonotype sharing between twins is enabled by long-lived naive cells, but

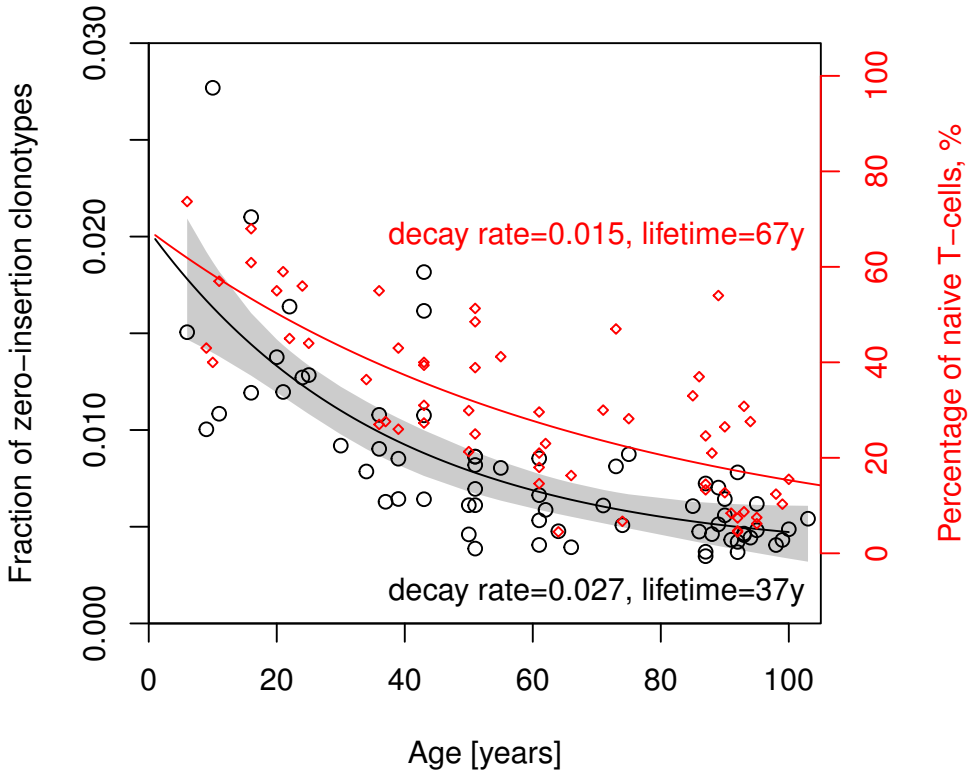

Figure 5.4: **Lifetime of abundant in-frame TCR beta clonotypes with zero insertions.** The fraction of zero-insertion clonotypes among the 2000 most abundant clonotypes in the unpartitioned repertoire as a function of age (black circles) is well fitted by an exponentially decaying function of time (black curve). This decay is faster than would be predicted from the decay of the naive compartment alone (red curve), indicating a slow decay of zero-insertion clonotypes of fetal origin. Red diamonds show percentage of naive T-cells measured using flow cytometry (see [23] for details). Scale of red axis was chosen so that the two decay curves start at the same point at age 0, and have the same long-time limit. We present the analysis for different bin sizes in Fig. B.10

does not exclude the possibility that this excess sharing can be supported by memory cells as well.

5.3.6 *Clonotypes with zero N insertions quantitatively explain the relation between clonotype abundance and sharing between unrelated individuals*

We have shown that abundant clones are enriched with zero-insertion sequences, both in the cord blood and in the adult naive repertoire. Zero-insertion clonotypes (regardless of their origin) are most likely to be shared by convergent recombination than regular sequences, because they are more likely to be generated due to reduced diversity. What are the implications of this observation for sharing between unrelated individuals? Since zero-insertion sequences are overrepresented among abundant clonotypes (Fig. 5.3), we predict that abundant out-of-frame clones are more likely to be shared.

To make our prediction quantitative, we built a mixture model of the out-of-frame alpha repertoire (see S1 Text for details). We assumed that clonotypes of a given abundance C are made up of a certain fraction $F(C)$ of TdT-, zero-insertion clonotypes, and a complementary fraction $1 - F(C)$ of regular TdT+ clonotypes. Because TdT+ clonotypes may also have no insertions, the fraction of the TdT+ and TdT- sets had to be learned in a self-consistent manner. To learn these fractions, for each abundance class C we directly quantified the fraction $F_0(C)$ of sequences in the data that are consistent with zero insertions (i.e. can be entirely matched to the germline segments). Because non-templated nucleotides can coincide with the template, and also because TdT+ cells may have no insertions, $F_0(C)$ is not equal to $F(C)$. However they are linearly related, so that it is enough for a model to agree with the data in terms of $F_0(C)$ to also guarantee agreement in terms of $F(C)$. We generated a large number of nucleotide alpha out-of-frame sequences using our recombination model, and separated them into two groups: those that are consistent with no insertions (group A), and the others (group B). For each abundance class C, we created articifical datasets made of a fraction $F_0(C)$ of sequences from group A, and a fraction $1 - F_0(C)$ from group B, where we recall that $F_0(C)$ is estimated from the data. We then repeated the sharing analysis in these artificial datasets in the same way as in the real datasets. The model accurately predicts the normalized sharing number of out-of-frame alpha-chain CDR3s as a function of clonotype abundance (Fig. 5.5), up to the common multiplicative factor of 1.7 by which the non-mixture model generally underestimates CDR3 sharing (see Fig. B.2 p. 137). Thus, the enhanced sharing of high-abundance clonotypes is entirely attributable to their higher propensity to have no insertions, making them more likely to be shared by chance.

5.4 Discussion

We found that adult twins present an interesting case of microchimerism in the adaptive immune system: shared rare TCR variants that recombined before birth survive for decades in their repertoires. We have also shown that adult naive repertoires, but not memory repertoires, have similar zero-insertion TCR

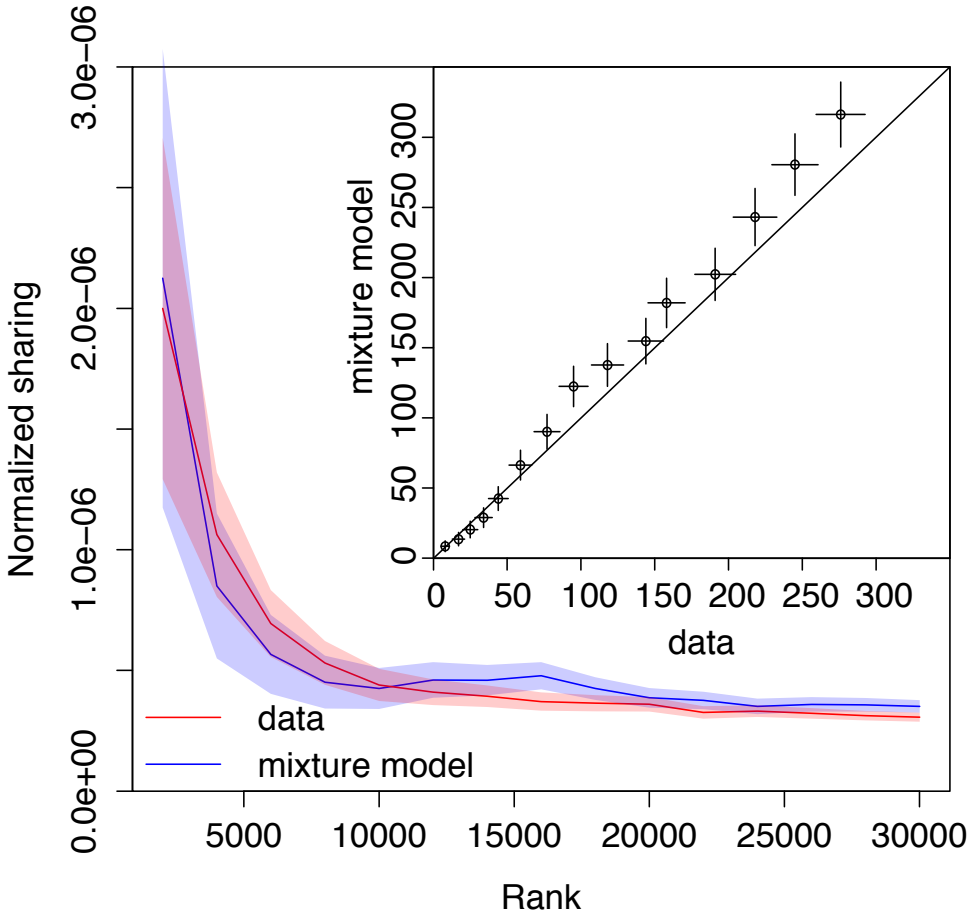

Figure 5.5: **Sharing of alpha out-of-frame TCR clonotypes as a function of clonal abundance.** The normalized number of shared out-of-frame alpha CDR3 nucleotide sequences between two individuals is showed as a function of clonotype abundance (e.g. normalized sharing for 2000 most abundant clones from both repertoires, 4000 most abundant, etc.), and compared to the amount of sharing that would be expected by chance (blue curve), taking into account the variable fraction of zero-insertion clonotypes as a function of their abundance. Data and predictions show excellent quantitative agreement (inset), with one fitting parameter. Error bars show one standard deviation.

clones distributions as cord blood repertoires. With age, the clone size distribution of naive adult repertoire becomes more similar to that of the memory repertoire. We hypothesize that this similarity between adult naive and cord blood repertoires is maintained by long lived fetal clones. Our results on the biological trafficking of T cells in twins are robust to possible experimental artefacts. First, our framework relies on the accurate counting of TCR cDNA sequences using unique molecular identifiers [84]. To exclude the possibility of contamination during the PCR and sequencing process, we double barcoded each cDNA library. To further exclude the possibility of early contamination of the blood samples, we performed replicate experiments at different times using different library preparation protocols. Comparison of repertoire overlaps from such replicate experiments for the same set of twins shows no difference and rules out experimental contamination as a confounding effect (see B.1 p.127). We also observed the same effects in previously and independently collected datasets [206], further excluding the possibility of experimental artefacts (Fig. B.8 p. 143). This reproducibility also suggests that the majority of out-of-frame sequences are not sequencing errors. Additional evidence for this fact comes from the different fractions of out-of-frame sequences observed in alpha and beta chains in TCR cDNA sequencing data, 13 and 3 percents respectively [206]– both of these fractions are much higher than the indel rate for the illumina platform [16, 149]. Our conclusions rely on a variety of data sources, and make extensive use of statistical analysis. As it is not yet possible to collect data from the same donors over many years, statistical evidence such as the amount of sharing in twins, or the amount of zero-insertion clonotypes versus age, is needed to investigate the evolution of repertoires over decades.

Cord blood sharing between twin embryos could have important implications on twin immunity: they could share and respond with private clonotypes, which would otherwise not be likely to be produced independently. This could possibly include sharing of malignant [54, 171, 190] or autoimmune clones, leading to disease in both individuals. In very rare cases such transfusion could also occur between dizigotic twins, leading to chimerism [13]. Anastomoses between monochorionic twin placentas are very common (more than 85 percent of uncomplicated pregnancies [95]), however the amount of exchanged blood may vary, and in some extreme cases it even leads to adverse outcomes such as twin-to-twin transfusion syndrome [94]. These effects could possibly affect the initial number of in-utero shared clonotypes. This mechanism of sequence sharing is very different from sharing by convergent recombination [181], because it also implies the sharing of the second TCR chain and of the cell phenotype. Paired repertoires studies, which combine alpha and beta chains together [76, 174], could be used to track clones shared between twins more precisely, and distinguish them from convergently recombined ones.

Our results suggest two mechanisms with opposite effects on the sharing of clonotypes in twins as a function of the number of insertions. On the one hand, we have argued in Figs. 5.1 and 5.2 that clonotypes shared through direct cell exchange should have a 'normal' number of insertions, because they are not due to random convergent recombination (which favors low numbers

of insertions). On the other hand, we have shown in Fig. 5.3 that cord blood cells are enriched in zero-insertion clonotypes, suggesting that clones shared in utero should be enriched in clonotypes with no of few insertions. Which one of these two effects dominate? TdT is suppressed in human embryos mostly in the first trimester of pregnancy [62]. Since TdT is active in the later trimesters the majority of the cord blood repertoire consists of clones with non-zero insertion numbers [137] similarly to the regular TdT+ post-natal clones. We show that the insertion distribution for non-abundant clones in cord blood closely resembles the insertion distribution observed in adults, with most clonotypes having insertions (see Fig. 5.3B II). Such clonotypes could be exchanged in utero between twins, and easily identified as shared clonotypes with low P_{gen}. Our theory predicts that twins should also exchange zero-insertion clonotypes, which are abundant in cord blood. However these shared clonotypes are indistinguishable from clonotypes shared by convergent recombination, which are also likely to have zero insertions. Therefore, the higher abundance of zero-insertion clonotypes in cord blood relative to mature repertoires does not contradict the observed sharing of high-insertion clonotypes due to cord blood exchange.

We have also showed that some of the clonotypes transferred in utero have the CD45RO+ phenotype, typical of central memory cells. It is possible that the longevity of these clones is connected with their memory status acquired early in life. To test this hypothesis, one would need to perform deep sequencing of purely sorted naive T-cells from adult twins and repeat the analysis presented in this paper. The transition from naïve to memory is also associated with clonal expansion, so it is possible that, within the in utero transfer hypothesis, the most easily detectable clonotypes shared between twins come from the memory population simply due to sampling effects. At the same time, the results plotted in Fig. 5.3 suggest that naïve clonotypes may also be long lived. Thus, clonotypes transferred in utero in twins could be either of naive or memory origin.

Our conclusion that fetal clonotypes are long-lived is based on the analysis of over-abundant zero-insertion clonotypes. Invariant T-cells, MAIT (Mucosal-Associated Invariant T-cells) and iNKT (Invariant Natural Killer T-cells) are intrinsically insertion-less, have restricted VJ usage for alpha chain, and are often abundant. These cells are produced in adulthood and could in principle constitute a substantial fraction of our zero-insertion dataset, confounding our analysis. Since our abundant zero-insertion clonotypes have a very diverse usage of VJ genes, we can exclude that the majority of them are from invariant T-cells, although we did identify a small number of such invariant TCR alpha chain clonotypes, see S1 Text. An alternative explanation of the skewed zero-insertion clone size distribution of naive repertoires (see Fig. 5.3A) is the existence of previously unknown subset of insertionless T-cells characterized by large proliferation activity, which would be produced in adulthood and make up the most abundant clones of the naive repertoire. To support this hypothesis, one would need to further assume that the production of these cells decays with age, to be consistent with the observations of Fig. 5.4. Another related possibility is that insertionless clonotypes are generally favored by

thymic selection, again in a age-dependent manner. However, in-frame clonotypes have been reported to be only moderately enriched (by less than 20%) in zero-insertion sequences relative to out-of-frame sequences (see Ref. [49], Fig. 5.3E-F), meaning that thymic selection does not substantially favor zero-insertion clonotypes on average.

Our current data clearly shows that clonotypes that originated in the cord blood tend to be among the most abundant in the naive repertoire, but we cannot unambiguously point to the source of this effect. One possibility is convergent recombination [132, 181]: high clonotypes abundances could be due to the accumulation of multiple convergent recombination events made more likely by the limited recombination diversity during fetal development. However, we observed clonotypes with low generative probabilities among the most abundant clones in the cord blood repertoire, and also clonotypes with high generation probability among the least abundant clones. We conclude that convergent recombination alone could not predict cord blood clone frequencies. An alternative explanation is that these clones have had more time to expand than others. Fetal cells come from different precursors, and mature in a different environment (the fetal liver), than post-natal cells [108]. *In vitro* experiments have shown that fetal T-cells have a different proliferation potential than post-fetal cells [150]. Additionally, a vacant ecological niche effect may play a role. When these clones first appeared, the repertoire had not reached its carrying capacity set by homeostatic regulation, leaving room for future expansion. These clones may have initially filled the repertoire, later to be gradually replaced by post-fetal clonotypes. Consequently, fetal clones, including those whose TCR was recombined with no TdT, would be expected to have larger sizes. Quantitative TCR repertoire profiling (preferably with the use of unique molecular identifiers for accurate data normalization and error correction), performed for species with no TdT activity in the embryo, such as mice, as well as novel cell lineage tracking techniques [116] could be used to investigate the detailed dynamics of fetal clones. This large initial expansion of fetal clones could protect them from later extinction. This would suggest that the estimated 37-year lifetime of zero-insertion fetal clonotypes could be longer than that of regular clones produced after birth.

Sharing of beta TCRs has previously been shown to decrease with age [23]. Depletion of fetal clonotypes, which are more likely to be shared, could contribute to this phenomenon. Our results also predict that the excess sharing of clonotypes between twins due to the trafficking of fetal cells should decrease with age. In general, the observed abundance of large zero-insertion clonotypes and their persistence through significant part of our life should have important consequences for the adaptive immunity regulation both in pre- and post-fetal period. Interestingly, transgenic mice with induced fetal TdT expression showed impaired antibody response to certain bacterial pathogens, suggesting an important functional role of the low-diversity fetal repertoire in immune competence [7]. We could speculate that the primary target of these cells might be common pathogens with a long history of coevolution with humans, such as CMV and EBV.

Lastly, our general framework for analyzing the overlap between different repertoires has far-reaching practical implications for the tracking of T-cell clonotypes in the clinic. In particular, the analysis of overlap between pre- and post-treatment repertoires using probabilistic characteristics of clonotypes sharing could help determine the host or donor origin of clonotypes after hematopoietic stem cell transplantation (HSCT), and also increase reliability of malignant clones identification in minimal residual disease follow-up.

5.5 MATERIALS AND METHODS

For a more detailed description of experimental and data analysis procedures see S1 Text Materials and Methods.

NGS library preparation. RNA was isolated from the PBMC of healthy Caucasian donors: 3 pairs of female monozygotic twins (aged 23, 23 and 25 years old), 19 year old and 57 year old males, a 51 year old female and cord blood from a female newborn. CD4+ and CD8+ populations were isolated using CD4+ and CD8+ T-cell positive isolation kits (Invitrogen), CD45RO+ and naive cells were isolated from PBMC using CD45RO+ enrichment and human naive T-cell isolation kits (Myltenyi) respectively. cDNA of TCR alpha and beta chain was synthesized and sequenced on the Illumina HiSeq platform (see Fig. B.1 p. 136). for library preparation technique, Table **??** p.147 for the oligonucleotides used, Table **??** p.148 for all samples and numbers of sequencing reads).

Raw data processing. Raw data processing and data analysis were performed using published open-source software tools: MiGEC (`https://github.com/mikessh/migec`), MiXCR (`https://github.com/milaboratory/mixcr/`), tcR (`https://github.com/imminfo/tcr`) and repgenHHM (`https://bitbucket.org/yuvalel/repgenhmm/downloads`). We processed raw sequencing data with MiGEC [160] to extract unique molecular identifiers and we used MiXCR [15] to determine the CDR3 position. All raw data is available online on our server (see S1 Text Methods E. for the links) and also in Short Read Archive (SRP078490).

Data analysis. Recombination models for beta and alpha chains were inferred using an EM-algorithm as described in [48, 103, 115], using the repgenHHM [48] and IGoR[103] software tools, selection models were inferred as described in [49]. The shared clonotype analysis was performed using the tcR package [117] and R statistical programming language [133]. To predict the number of shared out-of-frame clonotypes we generated random sequences using the recombination model parameters inferred separately for each individual in the previous step. We then filtered out-of-frame clonotypes and calculated the number of shared sequences between these simulated datasets using the tcR package.

To predict the number of shared in-frame clonotypes we also generated random sequences with recombination model parameters, filtered in-frame sequences and calculated the Q selection factors for each CDR3 amino acid sequence using selection models inferred separately for each individual. The

number of shared sequences in the simulated in-frame datasets was reweighted by the Q factors as:

$$\frac{1}{|S_1| \cdot |S_2|} \sum_{s \in S_1 \cap S_2} Q^{(1)}(s) Q^{(2)}(s), \tag{5.6}$$

where S_1, and S_2 are two synthetic sequence samples drawn from two models $P^{(1)}_{gen}$,$P^{(2)}_{gen}$ learned separately from the out-of-frame sequences of the two individuals, and $Q^{(1)}(s)$, $Q^{(2)}(s)$ are selection factors learned separately from these two individuals' in-frame sequences. $|S_1|$ and $|S_2|$ denote the size of the two samples. The sum runs over sequences s found in both samples.

To estimate the distribution of the number of inserted nucleotides for different subsets of the repertoire (Fig. 5.3 and Fig. 5.4), we used the same EM-algorithm when inferring the full repertoire models. To minimize the noise due to small subset sizes, we only learned the insertion distribution and took all other model parameters to be the same as in the previously inferred model in [115].

To fit the exponent decay of the ageing data we used the nlm2 R package. The data used in these fits is given in S3 Table. Fitting an exponentially decaying curve to the fraction Z of zero-insertion clonotypes in the 2000 most abundant clones as a function of age T (Fig. 5.4):

$$Z \approx c + a \exp(-bT), \tag{5.7}$$

we found $c = 0.00363 \pm 0.00154$, $b = 0.0272 \pm 0.0091$ yr^{-1}, and $a = 0.016696 \pm = 0.00188$.

Fitting an analogous model for the attrition of the naive T-cell pool, *i.e.* the fraction N of naive T-cells as identified using flow cytometry (see [23] for details),

$$N \approx a' \exp(-b'T). \tag{5.8}$$

we obtained $a' = 0.68 \pm 0.054$ and $b' = 0.01485 \pm 0.0018$ yr^{-1}.

6 SOMATIC HYPERMUTATIONS

Most of the results presented in this chapter have been submitted for publication in Ref. [103].

Section 1.4.4 briefly introduced the affinity maturation process, during which B cell clones diversify and evolve to create more and more specific receptors for a given antigen. This chapter focuses on understanding the statistical rules governing this diversification arising from Somatic Hypermutations (SHMs).

6.1 INTRODUCTION

Somatic Hypermutations (SHMs) are introduced by the AID hypermutating enzyme and elements of the constitutive DNA repair machinery. Functioning of this process is puzzling at different scales.

First, at the global scale, how does the mutating complex find the correct loci to mutate? Several studies have shown that the transcriptional activity regulated by promoters [128], remote regulatory elements [10, 143] or the chromatin state (methylation, acetylation) controls the overall mutation rate such that genes with expression comparable to the Ig loci will exhibit similar mutation rate [5]. This lack of specificity is known to promote lymphomagenesis [167] by accidental edition of oncogenes and is thus of clinical interest.

Second, at the local scale, what are the mutation rules and what makes a nucleotide more prone to mutation than its neighbor? These questions have been partially answered by numerous studies either from a mechanistic or statistical point of view. The rest of the section will review the current state of knowledge from these complementary approaches. The next sections will present some work investigating an independent site targeting model for SHMs and its possible improvements.

6.1.1 *Mechanistic models*

Many reviews found in Ref. [31, 44, 78, 166, 167] aggregate current experimental knowledge with different mechanisms proposal. I here summarize what seems agreed upon.

AID binds single stranded DNA, most likely upon opening of the double DNA strand by the Pol. II RNA polymerase complex [86]. Upon binding AID catalyzes deamination of deoxycytidine (C) to deoxyuridine (U) consequently transforming C:G pairs into U:G mismatched pairs. From then, the most supported DNA-based hypermutation model proposes three alternative pathways for somatic mutations:

- if the mismatch is not detected by the DNA repair machinery, it will be fixed by DNA replication upon cellular division. A daughter cell will

then inherit a T:A base pair, while the other will inherit the original C:G pair.

- the newly created uracil is excised through Base excision repair (BER). This involves the uracil-DNA glycosylase (UNG) enzyme, cleaving the uracil base and leaving an abasic site (i.e a DNA base without any purine or pyrimidine). Upon cellular division, the abasic site will make the replication machinery stall and depending on the polymerase and other factors might introduce transversions or transitions. This mechanism could also be a source of insertions and deletions.

- the U:G mismatch is recognized by MSH2/MSH6 mismatch recognition heterodimer. This Mismatch repair (MMR) pathway would trigger a patch DNA synthesis process with an error prone DNA polymerase (Pol. η). Because the two previous mechanisms only explain mutations at C:G pairs, the MMR mechanism is the only explanation for A:T pairs mutation accumulation.

Conversely, a less supported RNA and not DNA based mechanism [166], suggests that mutations are accumulated via retrotranscription and integration of mutated cDNA in the locus.

The mechanism by which AID acts on given regions and limits its range of action is not understood. Ref. [167] proposes a halting mechanism for Pol. II. Not all Pol. II complexes would be associated with AID and would thus be able to keep fully transcribing the BCR. The ones associated with AID would halt at some random positions stopping RNA transcription. This model would explain a finite range from transcription initiation for mutations, but remains however highly hypothetical, and it is not clear whether purifying selection, in some experiments, would not be an alternative candidate explanation for reduced mutability outside variable regions.

Finally, cytidine deamination in the switch regions created by AID would also lead to DNA double strand breaks triggering class switch recombination. Although not studied in details in this work SHMs are known to introduce insertions and deletions on top of point mutations with a frequency that remains unknown [17].

6.1.2 *Statistical models*

Parallel to the molecular and structural biologist endeavor to explain SHMs mechanism geneticist and bioinformaticians have studied statistical models for predicting per base mutability and provide a neutral model for SHM targeting. In order to access raw SHM statistics without selection biasing two approaches have been explored, both aiming at providing context dependent mutation models:

- as for evolutionary biology, synonymous mutations provide mutation statistics in principle free of selection[1]. Ref. [196] constructs a penta-

1 Although one could imagine that a base change could change the RNA secondary structure leading to a less stable mRNA.

nucleotide context dependent mutation model from such synonymous mutations from long V_H sequences. Because only half of possible pentamers can be observed from synonymous mutations, the authors inferred the remaining ones by averaging over related observed pentamers.

- because non productive rearrangements also undergo somatic hypermutations, their statistics should provide unbiased relative mutation frequencies. In Ref. [157], the authors use V_H and J_H from non productive rearrangements in different species to build di- and tri-nucleotide context mutation models from low throughput sequencing experiments. More recently, Ref. [38] built a penta-nucleotide model from long V_H genes sequences in rearrangements engineered not to be productive in transgenic mice.

Such models include large sets of parameters, exponentially large in the context size, and are prone to over-fitting as a large context size quickly completely specify the position on the gene. Building these models still require proper assignment of the underlying unmutated gene ancestor. Because the V gene is long and easier to identify with certainty, most of these approaches have focused on building models solely on V gene. However, as we are interested in extracting the physical parameters of the hypermutation process, we seek a universal description that would also describe observed D and J gene mutation rates. In this section we relax the full context dependence assumption by using a independent site model, allowing us to probe various context sizes while keeping the number of parameters small.

6.2 INDEPENDENT SITE MUTATION MODEL

To study patterns of SHMs in BCR expressed by memory B cells, we included into IGoR the possibility to infer a sequence-dependent hypermutation rate. The probability of error or mutation at a given position on the nucleotide sequence is assumed to depend on its immediate n-mer context (see Fig. 6.1a), through the logistic transformation of an additive score computed using a Position Weight Matrix (PWM), similar to binding energy motifs used to describe DNA binding sites [8].

6.2.1 *Model definition*

The hypermutation model assumes the following form for the probability of hypermutations:

$$P_{err}(\mathbf{R}|\mathbf{S}) = \prod_{x, S_x \neq R_x} \frac{P_{mut}(S_{x-m}, \ldots, S_{x+m})}{3} \prod_{x, S_x = R_x} [1 - P_{mut}(S_{x-m}, \ldots, S_{x+m})], \tag{6.1}$$

with

$$\frac{P_{mut}(\boldsymbol{\pi})}{1-P_{mut}(\boldsymbol{\pi})} = \mu \exp\left(\sum_{i=-m}^{m} e_i(\pi_i)\right), \tag{6.2}$$

where $(\pi_{-m},\ldots,\pi_m) = (S_{x-m},\ldots,S_{x+m})$ is the sequence context of the original recombination product around a hypermutation at position x. The parameters $e_i(N)$, the position-weight matrix, and μ, the overall mutation rate, are part of the parameter set θ. In order to lift the degeneracy of the model we impose that $\sum_{N=A,C,G,T} e_i(N) = 0$ at every position i.

The pseudo-log-likelihood of the hypermutation model reads:

$$Q_{err}(\theta'|\theta) = \sum_{a=1}^{M} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a,\theta) \sum_{x=1}^{L} \left[\delta_{S_x,R_x} \ln \frac{1}{1+r'(\mathbf{S},x)} + (1-\delta_{S_x,R_x}) \ln \frac{r'(S,x)/3}{(1+r'(\mathbf{S},x))}\right], \tag{6.3}$$

where $r'(S,x) = r'(S_{x-m},\ldots,S_{x+m}) = \mu' \exp\left(\sum_{i=-m}^{m} e'_i(S_{x+i})\right)$. It can be rewritten as:

$$Q_{err}(\theta'|\theta) = \sum_{\boldsymbol{\pi}} \left[\left(\ln(\mu'/3) + \sum_{i=0}^{N} e'_i(\pi_i)\right) N_{mut}(\boldsymbol{\pi}) - \ln\left(1+\mu' \exp\left(\sum_{i=1}^{N} e'(\pi_i)\right)\right) N_{bg}(\boldsymbol{\pi})\right], \tag{6.4}$$

where

$$N_{bg}(\boldsymbol{\pi}) = \sum_{a=1}^{M} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a,\theta) \sum_{x=1}^{L} \prod_{i=-m}^{m} \delta_{S_{x+i},\pi_i} \tag{6.5}$$

$$N_{mut}(\boldsymbol{\pi}) = \sum_{a=1}^{M} \sum_{\mathbf{E}} P(\mathbf{E}|\mathbf{R}^a,\theta) \sum_{x=1}^{L} (1-\delta_{S_x,R_x}) \prod_{i=-m}^{m} \delta_{S_{x+i},\pi_i}. \tag{6.6}$$

During the Expectation step, we compute these two quantities for each (2m+1)-mer and then maximize Q_{err} at each step of the Expectation-Maximization scheme using Newton's method with a backtracking line search. To impose $\sum_\sigma e_i(\sigma) = 0$ we remove one parameter per position i by setting for one nucleotide, $e_i(N) = -\sum_{\sigma\neq N} e_i(\sigma)$.

We can then compute the entries of the gradient vector $\mathbf{J}$ (of size $3(2m+1)+1$):

$$\frac{\partial Q_{err}(\theta'|\theta)}{\partial \mu'} = \sum_{\boldsymbol{\pi}} \left(\frac{N_{mut}(\boldsymbol{\pi})}{\mu'} - N_{bg}(\boldsymbol{\pi}) \frac{r'(\boldsymbol{\pi})}{\mu'(1+r'(\boldsymbol{\pi}))}\right), \tag{6.7}$$

$$\frac{\partial Q_{err}(\theta'|\theta)}{\partial e'_i(\sigma)} = \sum_{\boldsymbol{\pi}} (\delta_{\pi_i,\sigma} - \delta_{\pi_i,N}) \left[N_{mut}(\boldsymbol{\pi}) - N_{bg}(\boldsymbol{\pi}) \frac{r'(\boldsymbol{\pi})}{1+r'(\boldsymbol{\pi})}\right], \tag{6.8}$$

along with the Hessian matrix $\mathbf{H}$ entries:

$$\frac{\partial^2 Q_{err}(\theta'|\theta)}{\partial \mu'^2} = \sum_{\boldsymbol{\pi}} \left(N_{bg}(\boldsymbol{\pi}) \frac{r'(\boldsymbol{\pi})^2}{\mu'^2 (1 + r'(\boldsymbol{\pi}))^2} - \frac{N_{mut}(\boldsymbol{\pi})}{\mu'^2} \right), \tag{6.9}$$

$$\frac{\partial^2 Q_{err}(\theta'|\theta)}{\partial \mu' \partial e'_i(\sigma)} = \sum_{\boldsymbol{\pi}} (\delta_{\pi_i,N} - \delta_{\pi_i,\sigma}) N_{bg}(\boldsymbol{\pi}) \frac{r'(\boldsymbol{\pi})}{\mu'(1 + r'(\boldsymbol{\pi}))^2}, \tag{6.10}$$

$$\frac{\partial^2 Q_{err}(\theta'|\theta)}{\partial e'_i(\sigma) \partial e'_j(\sigma')} = \sum_{\boldsymbol{\pi}} (\delta_{\pi_i,N} - \delta_{\pi_i,\sigma})(\delta_{\pi_j,N} - \delta_{\pi_j,\sigma'}) N_{bg}(\boldsymbol{\pi}) \frac{r'(\boldsymbol{\pi})}{(1 + r'(\boldsymbol{\pi}))^2}. \tag{6.11}$$

For each step of Newton's method we find the step direction by solving $\mathbf{H}\Delta\theta' = -\mathbf{J}$ and we gradually refine the step size based on the Armijo-Goldstein condition. These operations are iteratively repeated until the pseudo-log-likelihood of the error model for a given Maximization step of the EM framework is maximized.

6.2.2 *Results*

We ran IGoR on memory out-of-frame IGH sequences from Ref. [88] to learn 7-mer PWMs, as well as overall mutation rates (the geometric mean of the mutation rate over all possible 7-mers), while fixing the recombination statistics to those previously learned from naive sequences, using Expectation Maximization. IGoR's probabilistic framework handles the degeneracy of sequence origin caused by convergent combinations of gene choices and hypermutations. The learning procedure differs crucially from Ref. [50], where the hypermutation rate was uniform. Three distinct PWMs were learned for V, D, and J templated regions (Fig. 6.1b). To validate our PWM and mutation rate learning algorithm, we generated synthetic data with hypermutations according to the model learned from the real dataset, and re-learned its parameters using IGoR, finding excellent agreement (Fig. C.10).

The PWM prediction for the position-dependent probability of hypermutations correlated well with that actually observed in the sequences ($r = 0.7$ for V genes, see Fig. 6.1c and Fig. 6.2). PWMs were very reproducible across the two tested individuals ($r = 0.98$, Fig. C.11), indicating that the inference procedure is robust to the individual history of infections, and pointing to the universal nature of the SHM mechanism. By constrast, the inferred overall mutation rate differred by a two-fold factor between the two individuals, probably owing to differences in age, past infections, or lifestyle (Fig. C.11). The motifs we found recapitulate previously reported hotspot motifs (positive values of the PWM) for every gene, including WRCY (or WRCH [142]) and WA [11, 157] (W = A or T, Y = C or T, R = G or A; mutated position underlined), as well as cold-spot motifs albeit to a lesser extend (SYC, where S = C, G) [20]. In all three motifs, C and G are generally underrepresented, except for the mutated position in V and D genes where T is less mutated than others. We assessed the robustness of the model to n-mer length by learning PWMs of sizes ranging from 3 to 9 (Fig. 6.3). The contributions of each relative position did not change sub-

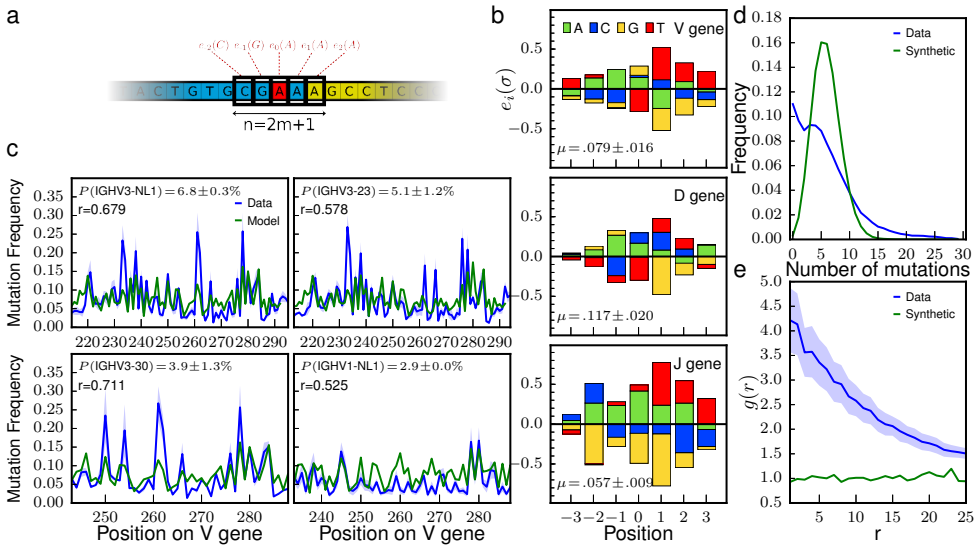

Figure 6.1: **Hypermutation landscape.** (**a**) Position-Weight Matrix (PWM) model for predicting hypermutation hotspots in IGH. Each nucleotide σ at position i within $\pm m$ of the hypermutation site (in red) has an additive contribution $e_i(\sigma)$ to the hypermutation log odd (Eq. 6.2). The PWM is learned by Expectation-Maximization from the out-of-frame sequences of memory B cells. (**b**) Comparison between the observed mutation rate per nucleotide and its prediction by the PWM model, as a function of position along the V segment, for the four most frequent V genes. Pearson correlation coefficient ρ and gene usage are given for each. (**c**) PWMs inferred from the V, D, and J genes. (**d**) Distribution of the number of mutations in each sequence. Data sequences have a broader distribution than predicted by the model (as computed from generating synthetic sequences and mutations with a data-inferred 7-mer PWM model). (**e**) Spatial co-localization index $g(r)$, measuring the overrepresentation of pairs of hypermutations at genomic distance r from each other. Synthetic sequences have $g(r) \approx 1$ by construction (green).

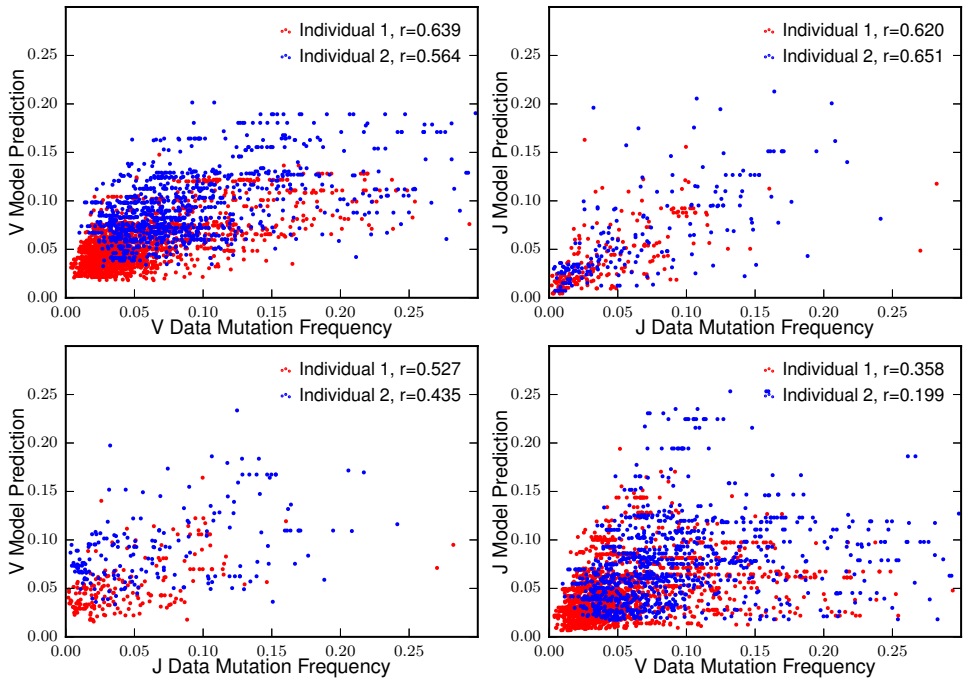

Figure 6.2: **Prediction of the mutation frequencies on real data.** By direct exploration of recombination scenarios we recorded the posterior mutation frequency per individual base pairs on V and J genomic templates and compare it to the independent 7-mer model. We plot a scatter for base pairs that have been observed at least 2000 times on a 100 000 sequences dataset, for which we can compute a reliable mutation frequency, and the mutation frequency predicted by our model. The two top panels show good predictive power for the gene on which the model was learned. However the two bottom panels show a lesser ability to predict the correct mutation frequencies on the whole locus, hence suggesting that differences observed in inferred position weight matrices (Fig. 6.3) are of biological relevance.

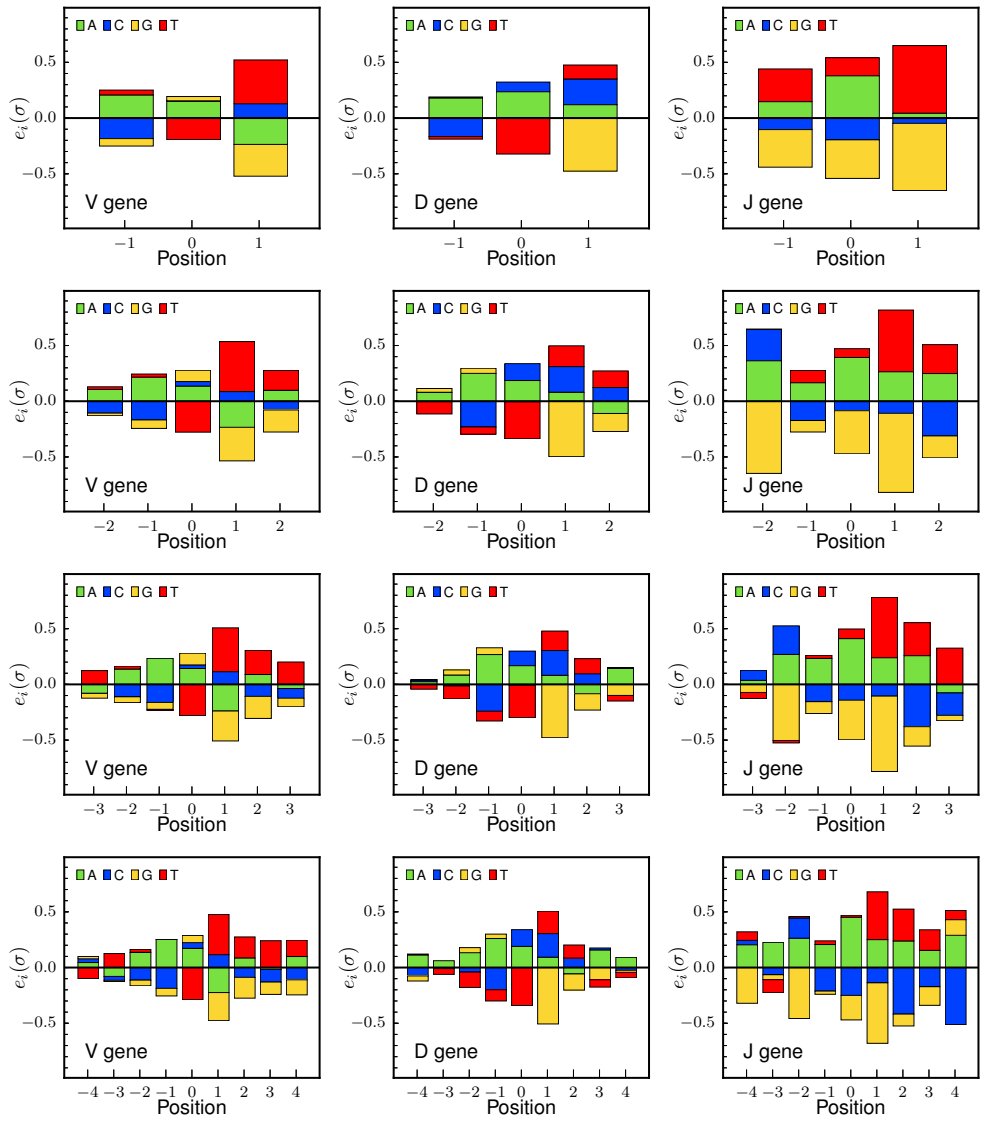

Figure 6.3: **Context logo for different context sizes on the three different genes.** We inferred position weight matrices for different n-mer sizes for V, D and J. With increasing n-mer sizes, side contributions do not vanish.

stantially as a function of context length. Positions at least up to 4 nucleotides away from the mutation locus contribute to the motif. This could mean that the context dependence is broad, or alternatively that the motif model is indirectly capturing non-contextual effects. Overall, the inferred PWMs give both a more detailed and more nuanced view of the rules that govern hotspot positions, and cannot be reduced to a few easily describable motifs.

Fig. 6.1b shows that the motifs differ substantially between V, D, and J genes. V-learned PWMs only moderately predict J-gene hypermutation rates ($r = 0.5$ versus $r = 0.7$ for V-gene rates), and J-learned PWMs predict V-gene rates even worse ($r = 0.24$, see Fig. 6.2). This disagreement indicates that predictions purely based on context-dependent motifs are insufficient to explain all of the variability in hypermutation probabilities, and that other mechanisms must be at play. The overall mutation rate was also different between germline genes, consistent with reports that the chromatin state affects hypermutation rates [31, 78, 166].

6.3 MUTATION ORDERING

As the context of different mutations might overlap the unknown order in which these mutation appear matters. In theory each mutation ordering corresponds to a hypermutation scenario and one should sum over all these scenarios. However the number of these scenarios increases exponentially with the number of mutations, and would thus quickly becomes intractable. Because only neighboring mutations with overlapping context interfere, the actual number of scenarios to explore only increases exponentially with the number of mutations with overlapping contexts. Overall, summing over all these scenarios is similar to finding all Hamiltonian paths[2] of the connected components[3] of a graph whose vertices correspond to mutated positions and edges are drawn for mutated positions whose distance on the sequence is smaller than the context size n. Although finding a Hamiltonian path in a generic graph is in principle an NP-complete problem, the particular structure of the described mutation graph is suitable to use a dynamic programming approach and sum efficiently over mutation scenario orderings.

However, taking into account the mutation ordering would only be necessary if we observed that our ability to infer the mutation PWM is affected. In practice our synthetic mutated sequences were generated taking mutation order into account and we observe that the naive strategy described in section 6.2.1, always using the germline sequence as a baseline, is sufficient to correctly re-infer the hypermutation model (see Fig. C.10) for our sizes of contexts and the considered $\sim 10\%$ mutational load .

2 A path going through each vertex of the graph exactly once.
3 A subgraph in which a path exist between any two vertices.

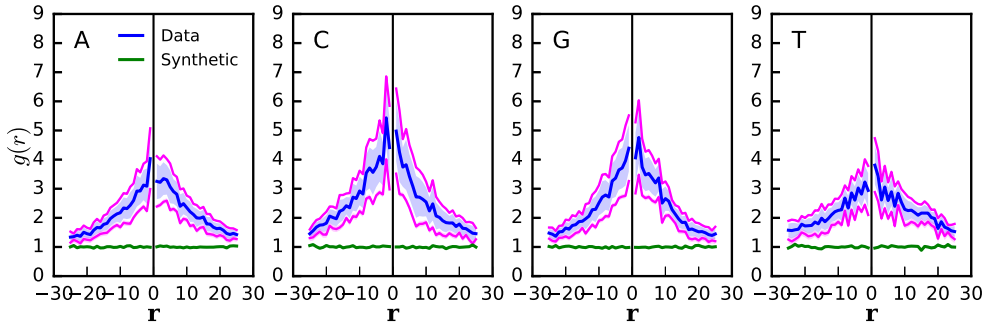

Figure 6.4: **Asymmetric mutation spatial-co-localization function.** By conditioning on the identity (top left) of the central ($r = 0$) nucleotide the symmetry of the radial distribution function is broken. Each figure shows the averaged (blue) mutation enrichment over two individuals (magenta). As a control the same quantity was computed on synthetic sequences with hypermutations distributed according to an inferred independent site mutation model (green).

6.4 BEYOND POISSON PROCESS

We then used the inferred PWM within IGoR to probabilistically call putative hypermutations in sequences. We first examined the distribution of the number of mutations in a sequence (Fig. 6.1d). The empirical distribution (red) is more skewed and has a longer tail than would be expected by assuming independent hypermutations in each sequence, as predicted by generating randomly hypermutated sequences with the inferred PWM (blue). This observation is consistent with the fact that different B cells have undergone a variable number of cycles of affinity maturation, resulting in differences in effective hypermutation rates.

6.5 SPATIAL CORRELATION

We asked whether hypermutations co-localized within the same sequence, by calculating the enrichment, or radial distribution function $g(r)$, of hypermutations at two positions as a function of their genomic distance (Fig. 6.1e)

$$g(r) = \frac{1}{N_r} \sum_{V;(i,j)\in C_V(r)} \frac{f(i,j,V)}{f(i,V)f(j,V)}, \tag{6.12}$$

where $f(i,V)$ and $f(i,j,V)$ are the frequencies of hypermutations at position i, and at both positions i and j, respectively, calculated from individual scenario statistics weighted by their posterior probabilities. $C_V(r)$ is the set of pairs of positions separated by r that were observed a large enough number of times in gene V, and $N_r = \sum_V |C_V(r)|$.

While this enrichment is 1 in synthetic sequences (since our model assumes that hypermutations are independent of each other), real data shows up to a 4-fold enrichment of hypermutations at nearby positions. This difference is consistent with the fact that AID can cause repairs of DNA over large re-

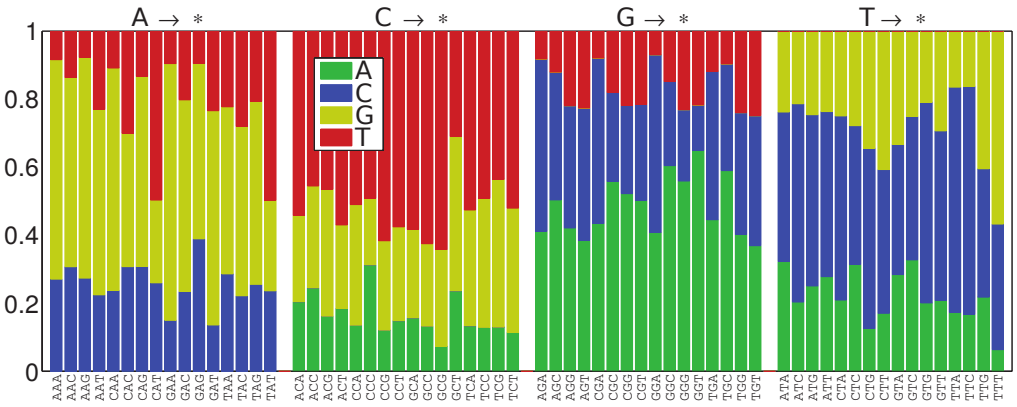

Figure 6.5: **Observed context dependence of substitution statistics in SHMs.** Substitution probabilities to the different bases as stacked columns vs. the local trimer context, grouped by the central base. Substitution is not uniform, depending primarily on the base being mutated, but varying with the context. This figure has been published in Ref. [50].

gions [175]. The typical distance at which the co-localization enrichment index decays gives an estimate for the length of these correlated regions of hypermutations, about 15 base pairs.

The radial distribution function in Eq. 6.12 is by construction a symmetric object. This symmetry can be broken by conditioning on the identity of the central nucleotide ($r = 0$) as shown in Fig. 6.4. While again the enrichment is always 1 in synthetic sequences we observe different enrichments in data depending on the central nucleotide such that C shows greater enrichment than A, G and T. From the mechanisms reviewed in section 3.7 this enrichment remains counter intuitive provided that C is the initial lesion and can be mutated without patch repair, the only proposed mechanism that would lead to co-occurrence of mutations. Strikingly, the patterns observed for the two individuals are extremely similar suggesting that this measure captures biophysical features of the hypermutation process. The apparent symmetry is reassuring regarding the abundance of insertions or deletions that would create an over enrichment on $r > 0$ part.

6.6 SUBSTITUTION STATISTICS

The model described in the previous sections only explains the preferential targeting of hypermutations from the nucleotide context, and assumes that the central nucleotide can mutate to any of other three nucleotides with equal probability. However preferential substitutions have already been reported [167]. Although we do not directly model them, the substitution statistics can be extracted from the individual scenarios statistics. Fig. 6.5 shows a clear dependence on the identity of the mutated base, with additional context-dependent variability from the local trimer sequence. The clear preference for C $\rightarrow$ T and corresponding G $\rightarrow$ A transitions are in agreement with the unrepaired uracyl replication.

In this chapter we have investigated an independent site context dependent hypermutation model built within IGoR's statistical framework. We inferred the parameters associated to this model on non-productive hypermutated data separately on V, J and for the first time on D, showing that our probabilistic treatment allows to lift the uncertainty even on D nucleotides. The inferred PWMs on a single gene were found to be very reproducible among individuals as previously reported [38]. However the inferred parameters were not reproducible among genes, questioning the ability of a context model to capture the biophysical process underlying SHMs. Because of the form of the model, this could arise if we were missing some dependencies captured by a full Nmer model and if the Nmer background was very different among genes. However, by constructing a full 5-mer model from the posterior number of times each background has been observed mutated or not (Eqs. 6.5 and 6.6) we also observe much weaker inter-gene ($r = 0.43$) than inter-individual ($r = 0.93$) agreement, confirming our first interpretation.

Using IGoR's ability to aggregate individual recombination scenario statistics we also showed that hypermutations cluster confirming once more that context dependent models cannot fully capture the SHM process. Overall, this analysis calls for a better modeling of SHMs, inspired by the known molecular processes, and confirms the need for a generic tool as IGoR to handle arbitrary complex probabilistic models.

Part IV

CONCLUSIONS AND OUTLOOKS.

CONCLUSIONS AND OUTLOOKS

In this manuscript I have introduced principled probabilistic approaches to describe immune receptor repertoire formation from high throughput sequencing experiments.

In chapter 3 I have presented IGoR, a statistical framework to study V(D)J recombination. By treating alignments of immune receptors to the germline probabilistically [115], IGoR corrects for systematic biases in the estimate of V(D)J recombination statistics, and predicts recombination scenarios more accurately than previous methods. Its detailed analysis of recombination scenarios further reveals that, even with a perfect estimator, the scenario is incorrectly called in more than 70% of sequences, suggesting caution when interpreting results from deterministic assignments. All models presented in this work are readily available with IGoR to allow researchers from the field to annotate their sequences, compute their generation probability and generate synthetic datasets. IGoR's modular design is a baseline for future research and an invitation for researchers in the field to help characterize new species, types of rearrangements, and refine our comprehension of V(D)J recombination and, more generally, immune repertoires formation.

Such refinements should be data-driven, and guided by the observation of discrepancies between real sequencing data statistics and synthetically generated ones. An example of such a discrepancy were the tandem Ds observed in BCR rearrangements, in section 3.7, found using deterministic alignments. This observation also outlines that simple methods, such as alignments, are useful to exhibit differences between two datasets, however proper quantification of the processes creating these differences must rely on a complete statistical treatment. In this spirit, further work on IGoR's BCR heavy chain model should be carried out, allowing for multiple D gene inclusion upon recombination. Beyond simply better characterizing the V(D)J recombination process in BCR heavy chains, understanding how frequent such rearrangements are is of clinical interest as most reported BNAbs in HIV controlling patients exhibit unusually long CDR3 regions [192]. Our ability to compute the probability of generation of the unmutated ancestors of such antibodies could thus be useful to design a vaccine maximizing the probability of a host response.

IGoR's modularity goes beyond its model definition as the full implementation has been designed to ensure evolvability and usability for new challenging data types such as paired receptor chain data. By making IGoR a fully open source platform, we hope to gather the community around the development of this research tool and allow the possibility to combine it with already existing software to allow seamless analysis of repertoire sequence of any technological origin.

As discussed in section 2.3 V(D)J annotation is only one of the three pillars of high throughput repertoire sequencing analysis. Because error detection, V(D)J annotation and genomic template inference are interconnected the need

for a tool addressing simultaneously these issues is great. With IGoR we hope to provide a good starting base to build a framework for a concurrent principled probabilistic handling of these issues. Such a framework will have a high computational cost. With the increasing popularity of repertoire sequencing, larger sequencing depth and new sequencing techniques the amount of data to treat quickly grows and keeping such a framework relatively fast in calculation time is a real challenge. For this purpose, future development of IGoR will soon implement modern stochastic optimization techniques as proposed in section 2.1.2.3. Such approaches will allow to considerably speed up the learning phase and keep the model inference tractable for regular computers despite the increasing dataset size and without subsampling drawbacks. Another development direction is the coupling of the already existing Sparse EM algorithm to a dynamic programming approach for scenario exploration. As presented in section 3.2.6 exhaustive scenario exploration is equivalent to traversing all terminal leaves of a tree. Because some events are functionally independent, the same operations might be carried several times without change. These functional dependences define a directed graph whose connected components can be separately explored as subtrees of the initial scenario tree.

Beyond these algorithmic considerations, I discussed the use of IGoR to answer concrete biological questions. Chapter 4 briefly outlined how simple predictions and parameters of these models could be used in this regard, by first showing that there is no parental imprinting for V(D)J recombination using the inferred joint VDJ usage probability. This discovery emphasizes again the importance of modeling the long ranged correlations induced by V(D)J recombination. Second, we used IGoR's inferred model statistics to estimate for the probability of rescue upon failure of the recombination process to produce a valid receptor. This question had, to our knowledge, not been addressed by the community. This measurement remains however a crude estimate. The analysis of statistically paired-sequence data with productive and non-productive rearrangement pairing is a potential lead to improve this estimate.

These two applications clearly show that some physical parameters of the V(D)J recombination are captured by our models. A detailed study of the different model components should be carried to relate them to molecular processes. For instance, we learn deletion profiles for each gene, while a single exonuclease enzyme is involved. Finding the sequence determinants responsible for the different deletion profiles would be an interesting research direction. Similarly, we model insertions as a Markov chain filling the junction from one gene to the other. A much more complicated process has been described at the molecular level (see section 1.3.2) and it would be of interest to assess whether our insertion model captures correctly inserted regions statistics, refine it, and relate the inferred parameters to the actual functioning of TdT.

While these predictions are based on single repertoire sequencing experiments, Chapter 5 addressed the much more delicate question of sequencing experiment comparison. From the inferred recombination model we predicted the number of clones shared by chance between two individuals. We then showed that the excess sharing observed in adult twins could not be explained simply by their similar genome, and that this excess was due to long lived

clones exchanged in utero. Armed with this proof of clonal persistence we then estimated the lifetime of TCR clones created before birth to be tens of years. Because clonotypes of fetal origin seem to be the largest clones this results suggest that one should be cautious upon inferring a recombination model on a small sequence sample, in which fetal clones would be overrepresented.

Finally, in chapter 6 we used IGoR's framework to encode a context dependent SHM model. Because of its spatial extension such a model could not be encoded using an HMM. Within this probabilistic framework we were able to infer PWMs corresponding to the hypermutation model on V, D and J. Using IGoR's ability to aggregate detailed recombination scenario statistics we showed that SHMs cluster. Together with the different PWMs obtained on the three genomic templates this suggests that simple motif models [38, 157], despite its good predictivity, cannot capture essential biophysical features behind the SHM process.

A future development will be to combine detailed V(D)J annotation statistics and existing biological theories about the molecular process described in section 6.1.1 to build a successful neutral SHM model taking into account preferential targeting, substitutions and possible insertion and deletion. Obtaining such a model is of primary importance as it is the first brick to construct a null model for affinity maturation in germinal centers and further quantify selection using known phylogenies. Recent experimental developments managed to track B-cell clones evolution in situ [170] and provide example phylogenies for which the clonal relation is certain. Such data would also constitute invaluable benchmarks for clonal reconstruction methods, although the throughput of these methods is however limited.

However, such tracking techniques cannot be applied in humans for clinical use, since the clonal relationship is unknown in the repertoire bulk sequencing experiment. As introduced in section 2.3 assessing the clonal relationship of sequences is a hard task and is an active field of research [80, 134, 195]. Fully solving the clonal inference also entails inferring the phylogenetic relations within clones. This problem is hard, as the unmutated ancestor is unknown and the large span of generation probabilities may play a role to find its identity. It is also a formidable theoretical problem as SHMs are context dependent and correlated, thus violating assumptions of existing phylogenetic methods [74]. As SHMs might also accumulate outside the Ig loci, the use of single cell RNA-seq techniques could simplify lineages reconstruction with the help of non Ig loci mutations.

A problem only overviewed in this manuscript is somatic selection. As briefly mentioned in chapter 5 the framework set up in Ref. [49] accounts for multiple selection layers from mRNA stability or receptor folding to peripheral selection and competition for antigens. It was recently shown in mice that such selection models inferred on blood or thymus extracted sequences exhibited the same selection traits, pointing to the inability of such models to capture peripheral selection [154]. The similarity of models obtained in different individuals also suggest that traits obtained by these models mostly capture general features such as folding constraints. Decomposing selection into its individual processes would allow to delineate individual from univer-

sal selection pressures. Because not all coding sequences are actually productive, one could extend current selection models and model coding sequences as coming from a mixture of selection traits: a set of selection traits for non-folding receptors and thus non productive sequences and a set for productive sequences and further functional central and peripheral selection. Such inference could be facilitated using cells containing two coding sequences from statistically paired sequences or the possibly different mRNA expression distribution between productive and non productive sequences. By isolating folding constraints, testable predictions on the ability of receptors to produce a pre-receptor could further be experimentally tested, similarly to WW protein domains [164].

Part V

APPENDIX

A

INTRODUCTION TO OPTIMIZATION, INFORMATION THEORY AND BAYESIAN STATISTICS.

A.1 OPTIMIZATION

Optimization denotes a set of mathematical tools aimed at finding the extrema of an objective function (or functional) f with respect to its parameters.

Most optimization methods are formulated in terms of minimization problems, for which f is denoted as the cost or lost function. In turn, any maximization problem can be translated into a minimization one by a simple transform $\hat{f} = -f$. There are many classes of optimization problems among which problems involving functions with only one optimum[1] define of convex optimization [19]. For such problems we generally seek to find the root of the function's derivative (when it exists) either analytically or numerically.

This section briefly presents convex optimization problems and some numerical methods to solve them. A particular emphasis will be put at the end on stochastic methods and a potential new stochastic algorithm due to their ever growing interest in large scale machine learning.

A.1.1 *Convex problems*

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if:

$$\forall (x_1, x_2) \in (\mathbb{R}^d)^2, \quad \alpha \in [0, 1], \quad f(\alpha x_1 + (1-\alpha)x_2) \leqslant \alpha f(x_1) + (1-\alpha) f(x_2). \tag{A.1}$$

This geometric definition makes no assumption on the function's properties and simply states than any line drawn between two points of the function is fully contained in the epigraph (the set of points above the function) of this function. Assuming the function is differentiable once, an equivalent definition is

$$\forall (x_1, x_2) \in (\mathbb{R}^d)^2, \quad f(x_1) \geqslant f(x_2) + f'(x_2)^T.(x_1 - x_2), \tag{A.2}$$

where f' is the function's derivative. Geometrically this inequality states that a convex function always lies above its tangents. Assuming the function is twice differentiable a third definition is

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad f''(x) \geqslant 0 \quad \text{(positive semi-definite Hessian)}, \tag{A.3}$$

1 There might be set of neighboring different values however all leading to the same value of the objective function, such as for a constant loss function.

where $f''(x)$ is the function's second derivative regarding x. The first definition in Eq. A.1 can be generalized to an arbitrary number of points and is known as Jensen's inequality[2]

$$\lambda_i \in \mathbb{R}, \quad \sum_i \lambda_i = 1, \quad x_i \in \mathbb{R}^d, \quad f(\sum_i \lambda_i x_i) \leqslant \sum_i \lambda_i f(x_i). \tag{A.4}$$

As previously mentioned convex functions are of much interest in optimization since convexity ensures that all local minimas are also global minimas. This property is useful when one tries to mathematically prove convergence of an optimization scheme. However, in practice it is often hard to prove rigorously whether an optimization problem is convex or not (even when it is). Still, since convex optimization methods are intuitive and easy to implement, one might try and use them on possibly non-convex problems. For such non-convex problems, upon trying different initialization parameters one might obtain different optimal points thus calling for the use of more sophisticated non convex optimization methods. This is the pragmatic approach we have adopted within the frame of this work, for which we only used convex optimization techniques, testing our assumption using different initializing conditions.

A.1.2 *Equality constraints*

Some problems can be subject to various equality constraints $g(\mathbf{x}) = c$ such as normalization constraints upon inferring a probability distribution. Such constraints can be incorporated into the cost function using Lagrange multipliers

$$\hat{f}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - c), \tag{A.5}$$

where $\hat{f}$ is the new cost function and λ the Lagrange multiplier. This can be generalized to an arbitrary number of constraints. Note that adding a Lagrange multiplier effectively adds a dimension to the optimization problem introducing possible issues with saddle points solutions that are not solutions of the original problem. When suitable it is thus preferable to absorb the equality constraint by eliminating one dimension.

A.1.3 *Gradient descent*

Optimization is a very active field and across the years many methods and refinements have been proposed to solve convex optimization problems. Here I will only present the two simplest first and second order methods around which many algorithms are built: gradient descent and Newton's method. These methods rely on the ability to compute analytically respectively first and second derivatives of the objective function. For non differentiable functions or functions whose derivative cannot be computed in closed form there

2 This inequality is of use in probability theory since it provides $f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)]$

exist zeroth order methods relying on finite difference estimation. However since none of these methods have been used in this work they shall not be discussed.

Starting from a position x_0 one wises to find $x^* \equiv \text{argmin}_x f(x)$, the global minimizer of the loss function f. To find it, the most naive approach would be to make small steps always in the direction of largest decrease in the objective function's value. This is what gradient descent achieves using the gradient $\nabla f(x)$ with the following recursion

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n), \tag{A.6}$$

where γ_n is the step size. This step size can be constant, vary as a function of n or be computed at each time step using a line search procedure depending on the setting of the problem. Line search refinement provides fastest convergence, however it requires many extra calculations of the function's value and might prove computationally expensive in e.g large scale machine learning. Otherwise specific sequences of decreasing gamma values ensure convergence. The constant step size on the other hand is mostly used for stochastic approximation methods as discussed in section A.1.5. The recursion is stopped upon finding a value of the gradient with a norm lower than an a priori set threshold.

A.1.4 *Newton Raphson methods*

Newton-Raphson (often simply called the Newton method) is an iterative procedure initially designed to find roots $f(x) = 0$ of a differentiable function $f(x)$. At each step a linear approximation of the function is made, such that the next step leads to the intercept between the tangent and the $x = 0$ axis

$$f'(x_n)(x_{n+1} - x_n) + f(x_n) = 0. \tag{A.7}$$

This method finds a natural application in convex optimization for which we seek to find the unique extrema of a function, and thus the root of its derivative. The recursion then becomes

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_{n+1})}. \tag{A.8}$$

In a more general setting of a multidimensional space Eq. A.8 generalizes to

$$\mathbf{H}\Delta\mathbf{x}_n = -\mathbf{J}, \tag{A.9}$$

where $\mathbf{J}$ is the Jacobian vector and $\mathbf{H}$ the Hessian matrix.

The method is prone to overshoot for some classes of functions, for which convergence can be obtained using a backtracking line search. Starting with a relatively large step size along the search direction (given upon solving Eq. A.9), the step size is iteratively reduced until finding a decrease of the cost function corresponding to the decrease expected by the value of the lo-

cal gradient. Such conditions are known as Armijo-Goldstein conditions. Note that such an approach does not aim at finding the best candidate point on the search line but rather a good starting point from which Newton's method can be further iterated.

As a stopping criterion the magnitude of the difference between the objective function at the current point and the minima of the quadratic approximation of the function at this point can be used [19].

The major drawback of Newton's method is the need to compute and invert the Hessian matrix whose size grows quadratically with dimensionality. Quasi-Newton methods have thus been developed to circumvent these issues. However for very high dimensional problems gradients methods remains more tractable.

A.1.5 *Stochastic Optimization*

Sometimes the exact evaluation of the objective function is either not possible or computationally too demanding. This is for instance the case for machine learning applications over extremely large datasets.

At each iterate n one can make an inexact evaluation $f_n(x_n)$ of the objective function $f(x)$ such that

$$\mathbb{E}[f_n(x_n)] = f(x_n). \tag{A.10}$$

The stochastic gradient descent (SGD) [139] method uses noisy evaluations of the gradient $\nabla f_n(x_n)$ on small random data batch at each iteration such that

$$x_{n+1} = x_n - \gamma_n \nabla f_n(x_n). \tag{A.11}$$

These dynamics define an Ornstein-Uhlenbeck process [61] in the space of parameters, whose steady state distribution can be computed in some simple cases such as constant step size γ_n [101]. This steady state distribution is Gaussian and a traditional way of estimating the optimal point is by averaging [130, 145]

$$\hat{x}_n = \frac{1}{n} \sum_n x_n. \tag{A.12}$$

In most cases the batch size remains constant during the optimization and there exist a computational trade-off between how fast one can update the parameters (inversely proportional to the batch size) and the noise in the estimate (also inversely proportional to the batch size). For some settings, the optimal batch size for fastest convergence of $\hat{x}_n$ can be calculated [101]. However, such calculations are based on the steady state distribution of the Ornstein-Uhlenbeck process and do not take into account how fast the algorithm converges to the steady state distribution. Intuitively, this convergence speed should be related to how fast the algorithm reaches the neighborhood of the solution.

A possible refinement of such algorithms would thus be to adopt an adaptive batch size. Starting from a small batch size would allow to move quickly into the region of interest, and gradually increasing it to a desired value would allow to converge faster to the solution. Because the successive parameters estimate x_n reach a steady state distribution one could increase the batch size everytime stationarity is reached. A possible test for reaching stationarity would be to detect a decrease in the variance[3] of the parameters value. Such technique would also be applicable in stochastic gradient approaches with momentum.

The algorithm proposed in this paragraph has not been tested and its development and analysis would be an interesting research direction.

A.2 BASICS OF INFORMATION THEORY

In his seminal 1948 paper *A Mathematical Theory of Communication* [156] Claude Shannon paved the way for the birth of information theory. Its initial aim was to quantify the amount of information that needed to be transfered through a communication channel to convey a message. However to achieve this goal a proper definition of information content was needed. The next subsections will briefly present some information theoretic quantities starting with Shannon's entropy that we will simply call entropy.

A.2.1 *Entropy*

The information content [97] of the outcome x of a random variable whose probability is given by the distribution $P(x)$ is defined by

$$h(x) = \log \frac{1}{P(x)}. \tag{A.13}$$

An intuitive justification for it would be the following: consider a random event with two possible outcomes a and b. If we have a strong belief that outcome a is very likely and b is very unlikely, acquiring knowledge of outcome a would only provide a slight "confirmation" information, thus low information content. However acquiring knowledge of outcome b is surprising and challenges our belief. In that sense b provides more information. Note the use of surprise or uncertainty to denote information. Upon a random event uncertainty on the outcome is lost and the same amount of information is gained. Uncertainty thus relates to events yet to be observed and is transformed into information upon data acquisition.

Shannon's entropy [36, 97, 156], $H(x)$, is defined as the average information content over the probability distribution:

$$H(x) = -\mathbb{E}_x[\log P(x)] = -\sum_x P(x) \log P(x), \tag{A.14}$$

where $\mathbb{E}_x$ denotes the expectation over x.

3 The Ornstein-Uhlenbeck process underlying potential is given by the loss function, the deterministic motion increases the computed variance over iterates. Upon reaching the steady state distribution the computed variance will start decreasing.

Actually Shannon's entropy is the *only* function fulfilling the following desirable conditions for a measure of information:

- $H(x) \geqslant 0$ information cannot be negative[4]
- Entropy is additive. The entropy of a process is the sum of individual entropies of it constitutive subprocesses, such that $H(x, y) = H(x) + H(y)$ if x and y are independent random variables.
- $H(x) = 0$ if and only if the process is deterministic
- The maximum entropy (or maximally uncertain) distribution corresponds to the uniform one for which all outcomes are equally likely when no additional constraint is imposed. Moreover, the entropy is a convex function and there exist only one maximum entropy distribution for a given set of constraints.

Entropy is defined up to a multiplicative constant, encoded by the choice of the logarithm basis. Base 2 logarithm or *bits* is the most usual unit for entropy as computers work with binary switches. This can also be interpreted as the minimum number of dichotomic operations (yes/no questions) to perform to answer a question where all answers are equiprobable.

From this definition result the definition of the joint entropy between two random variables x and y governed by the joint probability distribution $P(x, y)$

$$H(x, y) = -\sum_{x,y} P(x, y) \log P(x, y), \tag{A.15}$$

and conditional entropy

$$H(y|x) = \mathbb{E}_x[-\sum_{y} P(y|x) \log P(y|x)] = \sum_{x} P(x)[-\sum_{y} P(y|x) \log P(y|x)]. \tag{A.16}$$

These definitions naturally satisfy the additivity property of entropy

$$H(x, y) = H(x) + H(y|x) \qquad \text{(chain rule)}. \tag{A.17}$$

The entropy of a process is thus the sum of the entropy of its subprocesses.

A.2.2 *Kullback-Leibler Divergence and Cross Entropy*

The relative entropy or Kullback-Leibler divergence is a non parametric measure of dissimilarity between two probability distributions P and Q. Often denoted $D_{KL}(P \,||\, Q)$ (of P with respect to Q) it is defined as

$$D_{KL}(P \,||\, Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}, \tag{A.18}$$

4 Even a liar provides information provided we know he is lying

with $D_{KL}(P \,||\, Q) = \infty$ if the value x occurs with non zero probability in P and is assigned zero probability in Q.

Gibbs inequality (following Jensen's inequality (Eq. A.4) for convex functions) establishes

$$D_{KL}(P \,||\, Q) \geqslant 0, \tag{A.19}$$

with equality if and only if $P = Q$. Although sometimes called 'KL distance', note that it is not a symmetric object nor does it fulfill the triangular inequality.

The Kullback-Leibler divergence $D_{KL}(P \,||\, Q)$ is a measure of the inefficiency of assuming that random events come from the distribution Q when the true distribution is P. As briefly explained in the previous section, knowing the true distribution P of the random variable, one could construct a code with average description length $H(P)$. If, instead, one used the code for a distribution Q, one would need

$$\begin{aligned} H(P \,||\, Q) &= H(P) + D_{KL}(P \,||\, Q) && \text{(A.20)} \\ &= \sum_x P(x) \log Q(x), && \text{(A.21)} \end{aligned}$$

bits on average to describe the random variable. This quantity $H(P \,||\, Q)$ is called cross entropy of P with respect to Q.

A.2.3 *Mutual Information*

Another natural quantity one can derive using the definition of entropy is mutual information

$$\begin{aligned} I(x, y) &= H(x) - H(x|y) && \text{(A.22)} \\ &= H(y) - H(y|x) && \text{(A.23)} \\ &= D_{KL}(P(x, y) \,||\, P(x)P(y)). && \text{(A.24)} \end{aligned}$$

It corresponds to the amount of information gained (or uncertainty lost) about x upon knowing y (or vice versa). As Eq. A.24 suggests it can also be interpreted as the inefficiency of assuming independent variables to describe the joint process.

While correlation coefficients such as Spearman's or Pearson's ρ assume respectively monotonic or linear relationship between two variables mutual information is much more general and can be used for characterizing dependencies between the variables in any probability distribution. However as a drawback, while computing these correlations coefficients on a set of data points is straightforward computing the mutual information will require binning or fitting a correct distribution to the dataset.

A.3 BAYESIAN APPROACHES AND INFERENCE

Estimation of parameters from data is a central task in any research field ranging from physics, chemistry, biology, medicine or even sociology. This task arises in simple operations ranging from estimating the mean of a sample, to building a predictive linear model via linear regression. In section A.1 I have introduced techniques for solving optimization problems, i.e finding the parameters estimate minimizing a provided objective loss function. Although presented as a purely analytical problem, the value of the objective function may depend on a set of data. This is, for instance, the case for linear regression with least squared error and more generally in any machine learning algorithm. Although the same loss function will always provide the same solution, there is an infinite space of such functions that will provide an infinity of different answers. Each instance of these loss function carries implicit assumptions. This naturally brings the following question: is there a principled way for deriving a proper loss function or estimator?

While frequentist approaches aim at assessing the performance of a given estimator (i.e a given loss function) on any conceivable data and consider probabilities as limiting frequencies with infinite data, Bayesian inference provides a framework for combining in a mathematical model all the observed data and the a priori information or belief one has about the studied problem to provide an estimator. By taking a probabilistic model Bayesian approaches make explicit (subjective) assumptions and an estimator (or loss function) corresponding to these assumptions can be extracted. In this section I will briefly present the Bayesian inference framework and reasoning.

A.3.1 *Posterior, prior and likelihood*

Let's suppose we wish to study a dataset $\mathbf{D}$ of observations. In order to study it we assume a mathematical model, arbitrarily broad or precise, encapsulated in a global hypothesis H. This model is parametrized by a set of parameters θ. These three ingredients are necessary and sufficient for a Bayesian approach and their role can be decomposed using Bayes theorem

$$\underbrace{P(\theta|\mathbf{D},H)}_{\text{Posterior}} = \frac{\overbrace{P(\mathbf{D}|\theta,H)}^{\text{Likelihood}}\overbrace{P(\theta|H)}^{\text{Prior}}}{\underbrace{P(\mathbf{D}|H)}_{\text{Evidence}}}, \tag{A.25}$$

where the Ps are the different conditional probability distributions. The left hand side of the equation or *posterior* summarizes all our knowledge about the data and parameters. It depends on the *likelihood* function, summarizing information provided by the data, weighted by the *prior* summarizing our subjective belief for the value of a set of parameters. Evidence does not depend on

the parameters we are interested in, and is thus a normalization constant that we shall denote z[5].

As mentioned previously Bayesian approaches are claimed subjective approaches in which our belief is encoded in the choice of the model and our a priori belief in the model parameters value is encoded in the prior. Setting of this prior can be done with different aims:

- Assuming the prior is not important, set a uniform prior[6] or prior conjugate with respect to the likelihood for easier tractability of the problem.
- Prior is known and thus set to the known value.
- Subjective Bayesian approach where the prior is set by an "expert".
- Objective Bayesian: chose the least informative prior given the model. This can be achieved for instance by choosing the prior minimizing mutual information between posterior and prior. Such priors are known as non informative or Jeffrey's priors.

Previously I have made the dependence on the model assumptions H clear to outline that subjectivity of any Bayesian approach not only depends on the prior but also on the less disputed model assumptions. In fact, these model assumptions could be fully encoded in the prior as delta peaked distributions in the functional space to select e.g a family of distributions for our model. Assumptions deterministically encoded in the prior can range from any scale of assumption such as assuming i.i.d observations, a family of distributions or the precise number of components in a mixture model. In a sense a strength of Bayesian approaches is to make these subjective assumptions explicit. An even bigger strength is that these assumptions need not be binary and can remain fuzzy. In the context of model selection (e.g a family of distributions) one can also compare the posterior probability of different models. While frequentist predictions rely on a hard set of assumptions, a Bayesian approach could combine predictions of different models weighted by their posterior probability as a prediction [7] [97].

A.3.2 *Maximum a posteriori and Maximum likelihood*

From this simple definition how can we objectively design an estimator for the parameters θ?

5 The choice of the letter z is not fortuitous and shall remind the partition function of a physical system. Just as the partition function, the evidence is usually hard to compute as it involves integrating over the space of parameters such that $z = \int P(\mathbf{D}|\theta, H)P(\theta|H)d\theta$.

6 Uniform priors can however be problematic as they are not invariant over some transformations. A simple illustration of this is the learning of the probability of success p of a Bernouilli trial on which we impose a flat prior $P(p) = c$ expressing our lack of knowledge about p. Say now that for some obscure reason the log odds ratio $r = \log(p/(1-p))$ is easier to use than p for our inference. The resulting prior for r with imposing a flat prior on p is then $P(r) = e^r/(1+e^r)^2$ is not flat although there exist a bijection between r and p [187].

7 Although if predictions of the different models are very different it is not clear if a linear combination of these predictions would be a good predictor. Answering such a question is the task of frequentist approaches.

It may seen natural that the best set of parameters should be the one maximizing the posterior probability. However the obtained MAP estimator is not a priori invariant under all desired transformations and discards much of the posterior information. In theory, Bayes estimators are derived as estimators minimizing the expectation of some loss function over the posterior distribution. When this loss function is the mean squared error the corresponding Bayes estimator is the mean of the posterior. Still, under some regularity conditions the posterior distribution is approximately normal and the MAP estimator would thus be optimal too [187]. Provided a flat prior, from Eq. A.25 using the ML estimator is the same as performing MAP estimation. Under some regularity conditions ML is thus also optimal and unbiased. When such regularity conditions are not fulfilled ML will remain asymptotically unbiased. Because of these properties and its easier implementation we will thus use ML to perform parameter estimation in the work presented in this manuscript.

B

PERSISTING FETAL CLONOTYPES INFLUENCE THE STRUCTURE AND OVERLAP OF ADULT HUMAN T CELL RECEPTOR REPERTOIRES

B.1 SUPPLEMENTARY MATERIALS AND METHODS

B.1.1 *Blood samples*

Blood samples were collected from 3 pairs of monozygotic twin female donors, 23 (donors S1 and S2), 23 (donors P1 and P2) and 25 (donors Q1 and Q2) years old respectively. The individuals in each twin pair lived together for most of their lives, they were also tested for absence of dangerous infections before working with their blood (e.g. Hep C, HIV, syphilis). We also collected blood from two 19 and 57 year old male donors, along with a 51 year old female donor for memory and naive T-cells isolation, and a cord blood sample from a female newborn. All donors were healthy Caucasians, blood samples were collected with informed consent, and local ethical committee approval. The genetic identity of the twins was checked using polymorphic Alu insertion genotyping [99].

PBMCs were isolated from 12 ml of blood using Ficoll-Paque (Paneco, Russia) density gradient centrifugation. One third of the isolated PBMCs was used for total RNA isolation with the Trizol reagent (Invitrogen, USA) according to the manufacturer's protocol. Other cells were used for CD4, CD8 and CD45RO+ T-cells isolation.

B.1.2 *CD4, CD8, 45RO+ T-cell isolation*

CD4 and CD8 T-cells were isolated from PBMCs using the CD4+ and CD8+ positive selection kit (Invitrogen, USA) according to the manufacturer's protocol. CD8 T-cells were isolated from CD4 depleted samples to maximize the cell yield. 45RO+ cells were extracted using human CD45RO microbeads (Myltenyi, USA). Naive T-cells were isolated with the CD8+ T-cell naive isolation kit (Myltenyi, USA) according to the manufacturer's protocol without the final CD8 enrichment step.

Total RNA was immediately extracted from the isolated cells using the Trizol reagent (Invitrogen).

B.1.3 *TCR α and TCR β cDNA library preparation*

The library preparation protocol was adapted from [100] with modifications. The cDNA first strand was produced from the total RNA using the SmartScribe kit (Clontech, USA) and universal primers specific for the C-segment (see Fig. B.1A). Custom cap-switching oligonucleotides with unique molecular iden-

tifiers (UMI) and sample barcodes were used to introduce the universal primer binding site to the 3′ end of the cDNA molecules (see Fig. B.1 B). Each tube contained 500 ng of total RNA (corresponding to approximately 500000 PBMCs), 1x SmartScribe buffer, dNTP (1 mM each), 10pcmol of BCuniR4vvshort and TRACR2 primers (see Table S1 for sequences) and 1 µl of SmartScribe reverse transcriptase. 5mkg of the total RNA was used for the cDNA synthesis for each sample (10 tubes per sample, corresponding to approximately 5000000 PBMCs). The cDNA synthesis product was treated (45 min, 37C) with 1 µl of 5u/µl UDG (NEB, USA) to digest the cap-switching oligonucleotide and purified with the Quigen PCR purification kit. After the cDNA synthesis two steps of PCR amplification were used to amplify the cDNA and also introduce Illumina TruSeq adapters as well as the second sample barcode. After both steps the PCR product was purified using the Quigen PCR purification kit according to the manufacturer's protocol. The first PCR step (see Fig. B.1C) consists of 16 cycles of: 94 C for 20 sec, 60C for 15 sec, 72C for 60 sec. Each tube contained (total reaction volume 15 µl) 1x Q5 polymerase buffer (NEB), 5 pmol of Sm1msq and RPbcj1, RPbcj2, RPacj primers, dNTP(0.125 mM each) and 0.15 µl of Q5 polymerase. Then 1 µl of the purified PCR product was used for the second amplification step (see Fig. B.1D) consisting of 12 cycles of: 94C 20 sec, 60C 15 sec, 72C 40 sec. Each tube contained (total reaction volume 25 µl): 1x Q5 polymerase buffer, 5 pmol of Smoutmsq and Il-bcj-ind or Il-acj-ind primers (with sample specific indices, for beta and alpha libraries respectively, one primer per sample), dNTP(0.125 mM each) and 0.25 µl of Q5 polymerase. Size selection for 500-800bp fragments of the purified PCR product was performed using electrophoresis in 1% agarose gel.

B.1.4 *Next Generation Sequencing*

cDNA libraries were sequenced on the Illumina HiSeq platform (2x100nt). Custom sequencing primer sequences are listed in Table S1. The total numbers of sequencing reads are shown in Table S2.

B.1.5 *Raw data preprocessing*

All raw datasets used in this study are available online. For details about the donors see SI Materials and Methods Section A.

Twin TCR alpha chain sequences (3 identical twin pairs):
https://files.pub.cdr3.net/pogorely/HtSyudY2lkJ78TgzUKEshYUj4/alpha.tar

Twin TCR beta chain sequences (3 identical twin pairs):
https://files.pub.cdr3.net/pogorely/HtSyudY2lkJ78TgzUKEshYUj4/beta.tar

Memory and naive cells TCR beta sequences for three donors aged 19, 51 and 57, and an unsorted cord blood sample:
https://files.pub.cdr3.net/pogorely/HtSyudY2lkJ78TgzUKEshYUj4/mem_naive_cord.tar

Sample sheet containing barcode sequences and filenames of the samples:

https://docs.google.com/spreadsheets/d/1YTBXYP8ITpaVkUx46s_DtfBlZfvIu6UdGjcde-csMy4

Sequencing data from individuals of different ages used in Fig. 5.4 is publicly available in the SRA:

http://www.ncbi.nlm.nih.gov/sra/PRJNA316572

Raw sequencing data files were preprocessed with MiGEC [160], sequencing reads were clustered by unique molecular identifiers (UMI). UMIs with less than two reads were discarded to reduce the number of erroneous sequences. Then sequences were processed with MiXCR [15] to determine the CDR3 position and nucleotide sequence. For the numbers of UMIs after filtering see Table S2.

B.1.6 *Learning recombination statistics*

We built a generative model that describes the probability of generation of recombined sequences, following the theoretical framework described in [48, 103, 115]. The generation probability for each sequence is calculated as the sum over all recombination scenarios r that can produce that sequence, $P_{gen}(\text{sequence}) = \sum_r P_{rearr}(r)$. For TCR alpha chains the model assumes the following factorized form for a recombination scenario defined by the choice of genes (V and J), $P(V, J)$, deletions (delV and delJ), $P(\text{delV}|V)$ and $P(\text{delJ}|J)$ and insertions (ins), $P(\text{ins})$:

$$P^{\alpha}_{rearr}(r) = P(V, J)P(\text{delV}|V)P(\text{delJ}|J)P(\text{ins}). \tag{B.1}$$

The parameters of the models, the different probabilities in the factorized formula, were inferred by maximizing the likelihood of the observed out-of-frame sequences given the model, using Expectation-Maximization [115]. For alpha chains, the model was reformulated as a Hidden Markov Model, and the parameters were learned efficiently using a Baum-Welch algorithm, as described in [48].

For beta chains, the model describes probabilities for V, D and J choices, with possible deletions and insertions at each of the two junctions:

$$\begin{aligned} P^{\beta}_{rearr}(r) \quad = \quad & P(V, D, J)P(\text{delV}|V)P(\text{insVD}) \\ & \times P(\text{delDl}, \text{delDr}|D)P(\text{insDJ})P(\text{delJ}|J) \end{aligned} \tag{B.2}$$

The parameters for the beta chain model were inferred directly using the Expectation-Maximization algorithm, by enumerating all possible recombination scenarios that can produce each sequence, using the procedure described in [103, 115].

This procedure allows us to learn the features of the recombination statistics with great accuracy, in particular the distribution of number of insertions at the junctions, even though the recombination events themselves cannot be unambiguously be determined for each sequence because of convergent recombination.

B.1.7 *Distribution of insertions for each beta chains abundance class*

We applied the procedure described in the previous section separately for each abundance class of the beta-chain sequences. However, given the small size of the datasets (2000 or 3000 sequences), we did not learn the full model for each class. Instead, we used a previously inferred universal beta-chain recombination model [115] for the V,D,J gene usages and their deletion profiles, and we learned the insertion distributions ($P(\text{insVD})$ and $P(\text{insDJ})$) for each class separately, while keeping the other parameters constant. The distribution of insertions thus inferred are used to plot the results of Figs. 5.3 and 5.4 of the main text.

It should be noted that the effect size depends on the bin size. We replicated our analysis with different bin sizes, to show that the effect is still present (see Fig. B.10). Larger bins lead to lower effect sizes, but also to lower errors, so the significance of the difference in number of insertions between abundant and non-abundant clones is robust to the choice of the bin size.

To show that our results are not specific to certain donors, we reproduced our results shown on Fig. 5.3A for 7 additional published cord blood repertoires from [24], see Fig. B.11. All mean insertion distributions in all samples follow the same trend as the one presented on Fig. 5.3.

We also show how abundance varies with ranks inside each sample presented on Fig. 5.3A on Fig. B.12. Memory clones are typically more abundant than naive clones in same the individual, as was previously described [181]. The high frequencies of the few most abundant naive clones could be explained by contamination with memory compartment on the magnetic column. More accurate naive-memory separation method could potentially enhance the effect seen in Fig. 5.3A.

In Fig. 5.4 we show the decay of zero-insertion clonotypes from the 2000 most abundant clones in unsorted TCR repertoires from a published dataset of donors of various ages [24]. We hypothesise that the observed decay is due not only to the decay of naive pool, but also to the decay of fetal clones within the naive pool. However, a possible dramatic difference in the naive-memory partition of these abundant clones could confound this effect. To exclude this possibility, we estimated the naive-memory composition of 2000 most abundant clones from the unpartitioned, naive, and memory datasets of the three donors presented on Fig. 5.3A, who are of different ages. We attribute a clonotype from the unpartitoned dataset to the memory pool if the rank of this clone in the memory dataset was higher than in the naive one. We show that the ratio of naive to memory clonotypes in the 2000 most abundant clones is similar among all 3 donors, and is not decaying significantly with age: 1159 memory to 767 naive for the 19 year old donor (74 clones have undetermined phenotype), 1313 memory to 686 naive for the 57 year old donor (1 clone has undetermined phenotype) , and 1128 memory to 858 naive (14 clones have undetermined phenotype) for the 51 year old donor.

B.1.8 *Inference of selection factors*

In-frame sequences statistically differ from out-of-frame sequences (besides their frameshift), because in-frame sequences are functional and have passed thymic selection. For each sequence we defined a selection factor Q as the ratio of the probability of observing the sequence in the in-frame set, to the probability of recombining the sequence according to out-of-frame statistics (as inferred above). The overal selection factor Q is assumed to be the product of several independent factors q acting on the CDR3 length L and on the identity of amino acid a_i at each position i of the CDR3 [49]:

$$Q \propto q_L \prod_{i=1}^{L} q_{i;L}(a_i) \tag{B.3}$$

The parameters were inferred by maximizing the likelihood with gradient ascent, as described in [49].

B.1.9 *Data analysis*

Analysis of the shared clonotypes was performed using the R statistical programming language [133] and the tcR package [117].

B.1.10 *Out-of-frame sharing prediction*

To predict sharing for each individual, we generated sequences using our recombination model P_{gen} (alpha or beta), with individually inferred model parameters. Normalized sharing of the TCR sequences between two clonesets is defined as the number of the same unique TCR nucleotide sequences observed in both of them, divided by the product of the total numbers of unique TCR nucleotide sequences in the two datasets.

We calculated sharing of either whole chains, or of their CDR3, defined as the sub-sequence going from the conserved cystein at the end of the V region, to the conserved phenylalanine in the J region.

The alpha chain results for whole-chain sharing are plotted in the main text in Fig. 5.1, and the data shows good agreement with the model. The results for CDR3 sharing are shown in Fig. B.2. The model systematically underestimates the normalized sharing by a common multiplicative factor of 1.7 for non-twins, with a Pearson correlation coefficient of 0.8 between the data and the model prediction. Absolute numbers of shared CDR3 sequences for alpha chains varied from 400 to 1200.

For beta chain sequences, the prediction of out-of-frame sharing is more difficult because of the low numbers of out-of-frame sequences in the RNA data, which, combined to a lower mean P_{gen}, results in a much lower number of shared out-of-frame sequences. We also identified and removed from the dataset 26 out-of-frame sequences shared between more than two individuals. These sequences are likely to arise due to reproducible aligner errors or technology artifacts – some of them contained intronic sequences, etc. Absolute

numbers of shared beta CDR3 sequences varied from 0 to 82. Nevertheless, the number of shared beta out-of-frame CDR3 sequences for twins exceeded the model prediction (see Fig. B.3), confirming our hypothesis of biological contamination during pregnancy.

B.1.11 *In-frame sharing prediction*

To accurately predict the normalized sharing number for in-frame nucleotide clonotypes, we generated sequences from P_{gen} as we did for out-of-frame sequences, but weighted them by their selection factor Q to account for thymic selection. The predicted normalized sharing number was then calculated as:

$$\frac{1}{|S_1| \cdot |S_2|} \sum_{s \in S_1 \cap S_2} Q^{(1)}(s) Q^{(2)}(s), \tag{B.4}$$

where S_1, and S_2 are two synthetic sequence samples drawn from two models $P_{gen}^{(1)}$,$P_{gen}^{(2)}$ individually learned from the out-of-frame sequences of two individuals, and $Q^{(1)}(s)$, $Q^{(2)}(s)$ are selection factors learned individually from these two individuals' in-frame sequences. $|S_1|$ and $|S_2|$ denote the size of the two samples. The sum runs over sequences s found in both samples.

For both the beta and the alpha chains, the prediction agrees very well with the data (Fig. B.4 and Fig. B.5). For the beta chain, twins share more CDR3 sequences than non-twin pairs, while no such effect was observed for the alpha chain sequences. This fact could be explained by the much higher number of clonotypes shared due to convergent recombination in the alpha in-frame dataset than in the beta in-frame and alpha and beta out-of-frame datasets. Excess of shared CDR3 nucleotide sequences due to biological contamination in twins is lower than the amount of convergent recombination noise in the alpha in-frame shared CDR3 nucleotide sequences. Absolute numbers of shared in-frame CDR3 sequences for alpha chains varied from 30000-50000 sequences depending of the pair, and 5000-9000 for beta chains.

B.1.12 *Mixed model inference*

We hypothesized that the larger amount of zero insertion clonotypes is responsible for the increase in sharing between the most abundant clonotypes of the out-of-frame repertoires of unrelated individuals. To test this hypothesis, we constructed a mixture model for each abundance class, each class containing 2000 clonotypes ranked by decreasing abundance.

We assume that abundance class C contains a fraction $F(C)$ of clonotypes generated with zero insertions, and $1 - F(C)$ of regular clonotypes. Obtaining $F(C)$ is not straightforward because regular clonotypes can also zero insertions. In addition, the number of insertions cannot be determined with certainty – for example, a deletion followed by an insertion matching the germline sequence can be wrongly interpreted as a case of no insertions.

To circumvent this problem, we determine for each abundance class a simpler quantity to estimate, namely the fraction $F_0(C)$ of clonotypes that are con-

sistent with zero insertions, *i.e.* that can be entirely matched to the germline genes. Because of the reasons outlined above, $F_0(C)$ is *not* equal to $F(C)$. However, $F_0(C)$ is a linear function of $F(C)$, $F_0(C) = A + BF(C)$. Therefore, if we can generate synthetic sequences such that their $F_0(C)$ agrees with data, then we are guaranteed that their $F(C)$ will coincide with the data as well, even if we do not know the explicit mixing parameters $F(C)$.

To obtain this mixture, we generated many sequences from our recombination model P_{gen}. To determine which generated sequences were consistent with zero insertions, we aligned them to all possible V and J genomic templates. We then separated out the sequences consistent with zero insertions from the others, and created, for each abundance class C, and artificial dataset with a fraction $F_0(C)$ of such sequences, and $1 - F_0(C)$ of the other sequences (not consistent with zero insertions), where $F_0(C)$ is given by the data.

We then calculated normalized sharing in the synthetic data by including an increasing number of abundance classes, starting with the most abundant ones, and compared to data in Fig. 5.5.

B.2 Supplementary results

B.2.1 *Distinctive properties of shared clonotypes between twins*

Shared clonotypes in unrelated individuals appear in the process of convergent recombination. Sequences with a higher P_{gen} are thus more likely to be shared, and we can calculate accurately the distribution of P_{gen} among shared sequences (see Fig. 5.2). We observe that sequences shared between twins violate this prediction, consistent with our hypothesis that some of these sequences are due to biological contamination. To confirm this, we used a sequence feature that is negatively correlated with P_{gen} [115]: the number of insertions in the CDR3 region. The number of insertions in CDR3 sequences shared between unrelated individuals was indeed lower (Fig. B.6) than the mean number of insertions in non-shared sequences. However, the mean number of insertions in sequences shared between twins (black boxes) is higher than in unrelated individuals, $p = 1.83 \cdot 10^{-8}$, two-sided t-test. The same and even stronger effect is observed for memory (CD45RO+) cells, $p < 10^{-16}$, two-sided t-test (Fig. B.7).

Our theory also predicts that twins should have an excess of zero-insertion shared clonotypes, relative to non-twins. To check for this, we compared the normalized sharing number of zero-insertion out-of-frame clonotypes in the data and according to the model (see Fig. B.9). Although we observe higher sharing numbers in twins, this effect is made non-significant by high levels of noise. Since zero-insertion clonotypes have low diversity, these normalized sharing numbers are much higher than their generic counterpart of Fig. 5.1. In other words, convergent recombination is much more likely, masking the effects of fetal contamination.

Finally, the mean clone size of low-probability ($P_{gen} < 10^{-10}$), twin-shared sequences from Fig. 5.2, 8.8 ± 0.7, is significantly larger than that of generic low-

probability ($P_{gen} < 10^{-10}$) clones from that individual, 1.83 ± 0.013, providing another evidence of their fetal origin.

B.2.2 *The phenotype of beta chain out-of-frame shared clonotypes*

Two individuals displayed the most prominent excess of shared beta out-of-frame sequences. Since the model prediction for the number of shared sequences is close to zero we suppose that most of these shared sequences did not arise due to convergent recombination. These out-of-frame clones bear a second functional allele (otherwise they would have been filtered by selection in a thymus), and they also should have either the CD4 or the CD8 phenotype. To attribute these clonotypes a phenotype we separately sequenced CD4, CD8 and CD45RO positive subsets for the two donors and searched for the 84 out-of-frame CDR3s shared between the unpartitioned out-of-frame repertoires. 44 CDR3s were found in the CD8 subsets of both individuals, and only 5 sequences were found in the CD4 subsets of both individuals. 25 out of the 44 CD8 and 3 out of the 5 sequences were also found in the 45RO+ compartment. Only 3 sequences were mapped discordantly (e.g. CD4 in one twin and CD8 in the second twin), and 2 sequences were absent from the CD4, CD8 and CD45RO compartments of both individuals. For the other 32 sequences the CD4/CD8 status could be determined only for one individual (most probably due to the sequencing depth limitations). In case of convergent recombination it is unlikely that shared nonproductive sequences would have the same phenotype in different donors. The phenotypic study thus confirms the biological contamination hypothesis.

B.2.3 *Our results are reproducible using previously published data*

We tested the robustness of our results on previously published twin data from [206]. We observed the same excess of low-probability shared sequences in twins compared to unrelated individuals as in Fig. 5.2 (see Fig. B.8).These data also allowed us to control for possible experimental contamination. One of the twin pairs that participated in the present study was sequenced three years ago, using a different technology described in [206], excluding the possibility of any contamination between the old and new samples. Out of 84 beta out-of-frame clonotypes shared between two new twin samples, 59 were also shared between the new sample of one twin, and the old sample of the second twin. Therefore the out-of-frame sequences shared between the twins are reproducible and could not be result of experimental contamination with PCR-products or RNA.

B.2.4 *Invariant T-cell alpha clonotypes in the data*

It was previously shown that mucosal-associated invariant T-cells (MAIT) and natural killer T-cells (NKT) have an invariant alpha chain with very low diversity [69]. Specific V-J combinations are chosen (TRAV10/TRAJ18 for NKTs and

TRAV1-2/TRAJ33 for MAIT) and no nucleotides are inserted in the recombination process of these clonotypes. To see whether these clonotypes could potentially confound our analysis, we searched for published NKT and MAIT sequences in our datasets. 25 out of the 27 known MAIT sequences were found in the datasets at least once (21 out of them in the all six individuals), and 8 out of the 13 known NKT sequences (2 of them in the all six individuals). MAIT and NKT sequences are present in our data, but only a few shared sequences could be explained by them, so we do not exclude MAIT and NKT alpha sequences from the analysis. The majority of shared zero insertion sequences could thus not be attributed to known MAIT or NKT subsets.

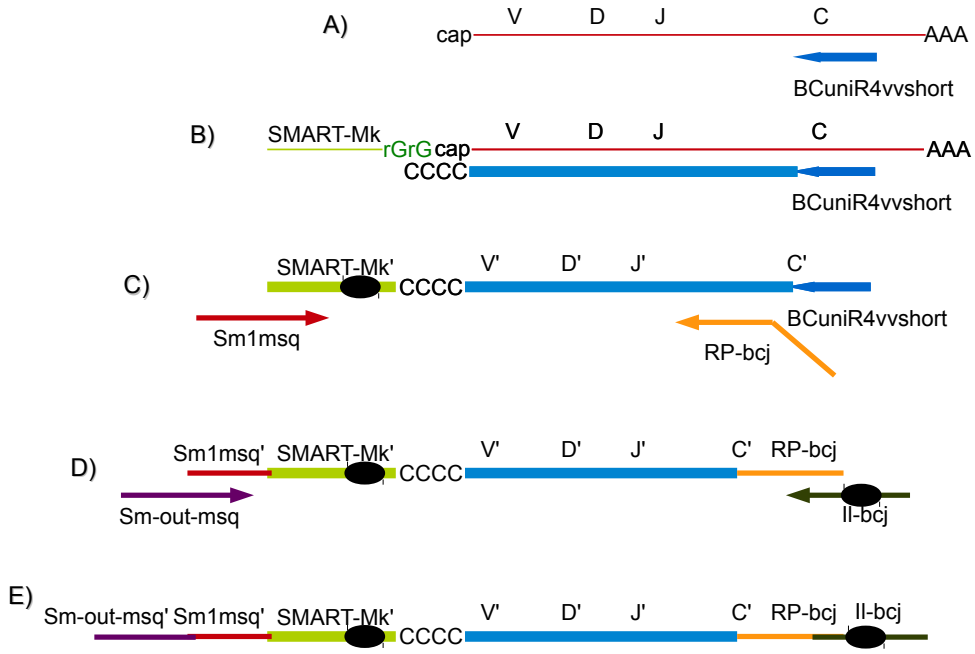

Figure B.1: **Library preparation protocol.** A) cDNA first strand synthesis for alpha and beta chains starts from specific primers in the C-segment conserved region. B) The template switching effect was used to introduce a universal primer binding site to the 3′cDNA end. The SMART-Mk sequence contains a sample barcode (black ellipse) for contamination control. C) and D) In two subsequent PCR steps we introduce the TruSeq adapter sequences along with Illumina sample barcodes (black ellipse). E) The resulting cDNA molecule is double barcoded, contains a Unique Molecular Identifier (UMI) and is suitable for direct sequencing on the Illumina HiSeq platform with the custom primers.

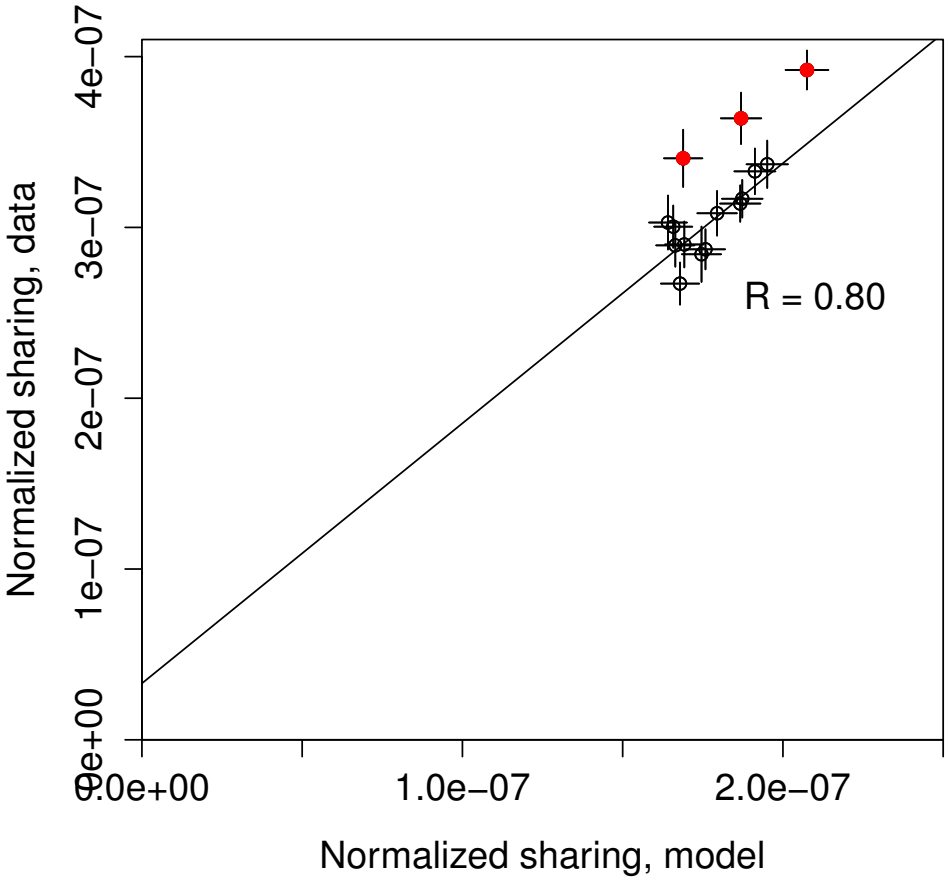

Figure B.2: Number of shared out-of-frame alpha TCR CDR3 clonotypes reported between all 15 pairs of 6 donors consisting of 3 twin pairs (ordinate) compared to the model prediction (abscissa). To be able to compare datasets of different sizes, the sharing number was normalized by the product of the two cloneset sizes. The outlying three red circles represent the twin pairs, while the black circles refer to pairs of unrelated individuals. Error bars show one standard deviation. The diagonal line is a linear fit for unrelated individuals, of slope 1.7.

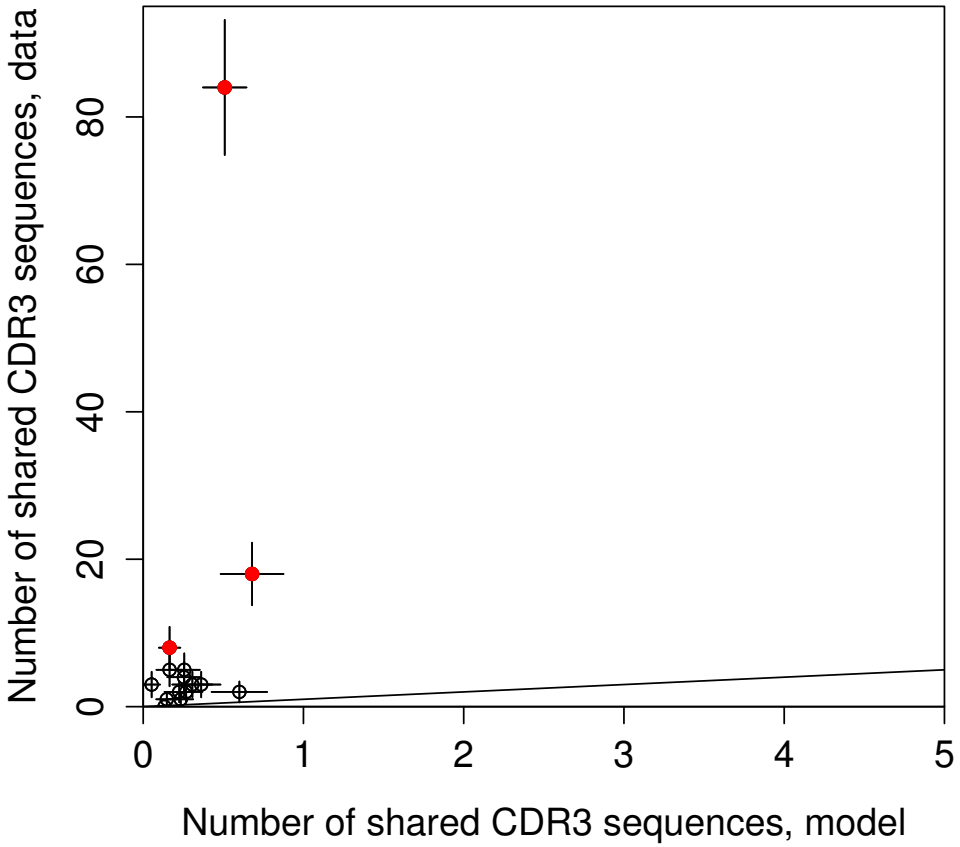

Figure B.3: Number of shared out-frame beta TCR CDR3 clonotypes reported between all 15 pairs of 6 donors consisting of 3 twin pairs (ordinate) compared to the model prediction (abscissa). The three outlying red circles represent the twin pairs, while the black circles refer to pairs of unrelated individuals. Error bars show one standard deviation.

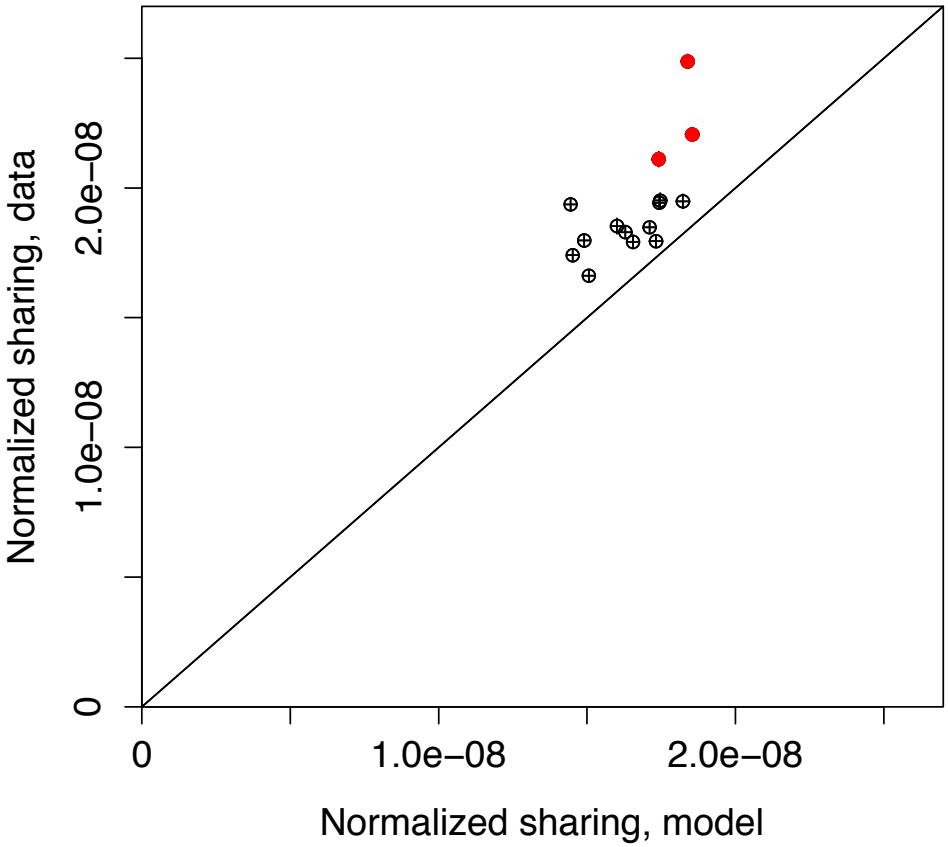

Figure B.4: Number of shared in-frame beta TCR CDR3 clonotypes reported between all 15 pairs of 6 donors consisting of 3 twin pairs (ordinate) compared to the model prediction (abscissa). To be able to compare datasets of different sizes, the sharing number was normalized by the product of the two clone-set sizes. The three outlying red circles represent the twin pairs, while the black circles refer to pairs of unrelated individuals. Diagonal is equality line. Error bars show one standard deviation.

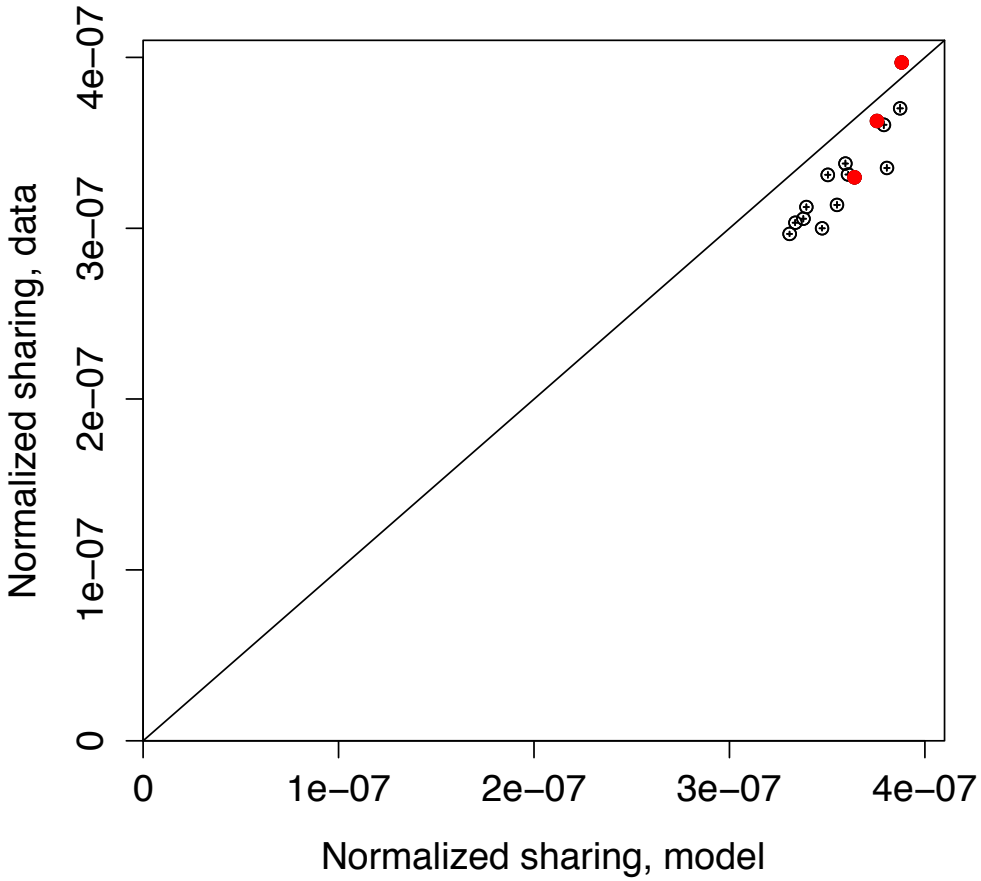

Figure B.5: Number of shared in-frame alpha TCR CDR3 clonotypes reported between all 15 pairs of 6 donors consisting of 3 twin pairs (ordinate) compared to the model prediction (abscissa). To be able to compare datasets of different sizes, the sharing number was normalized by the product of the two clone-set sizes. The three red circles represent the twin pairs, while the black circles refer to pairs of unrelated individuals. Diagonal is equality line.

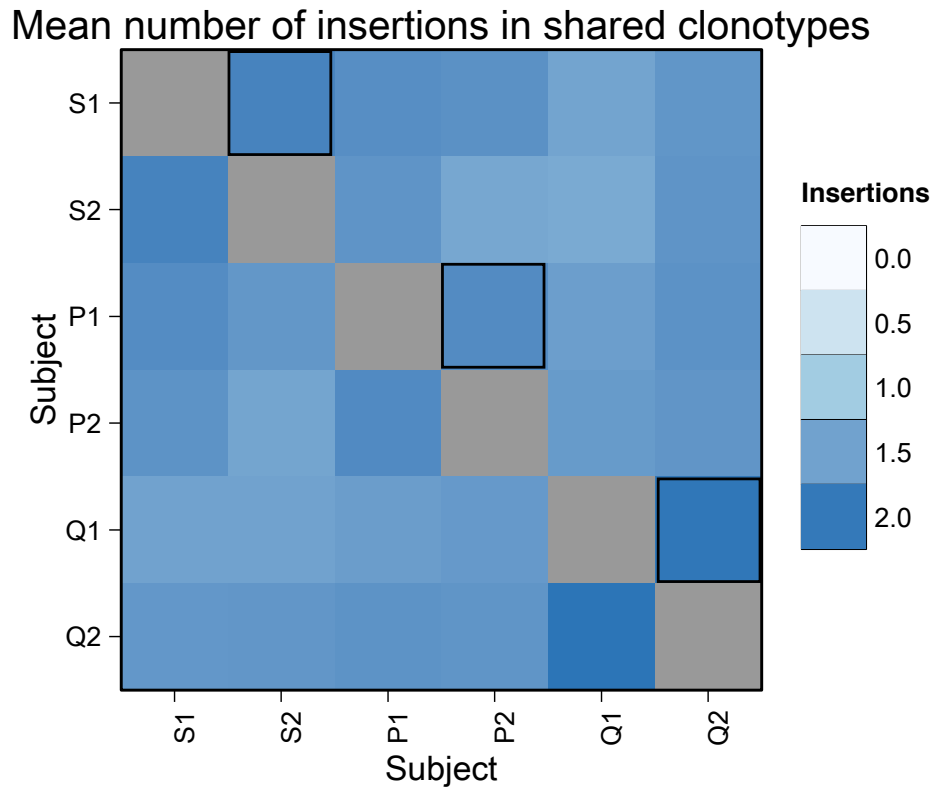

Figure B.6: Mean number of insertions in shared sequences in alpha out-of-frame repertoires.

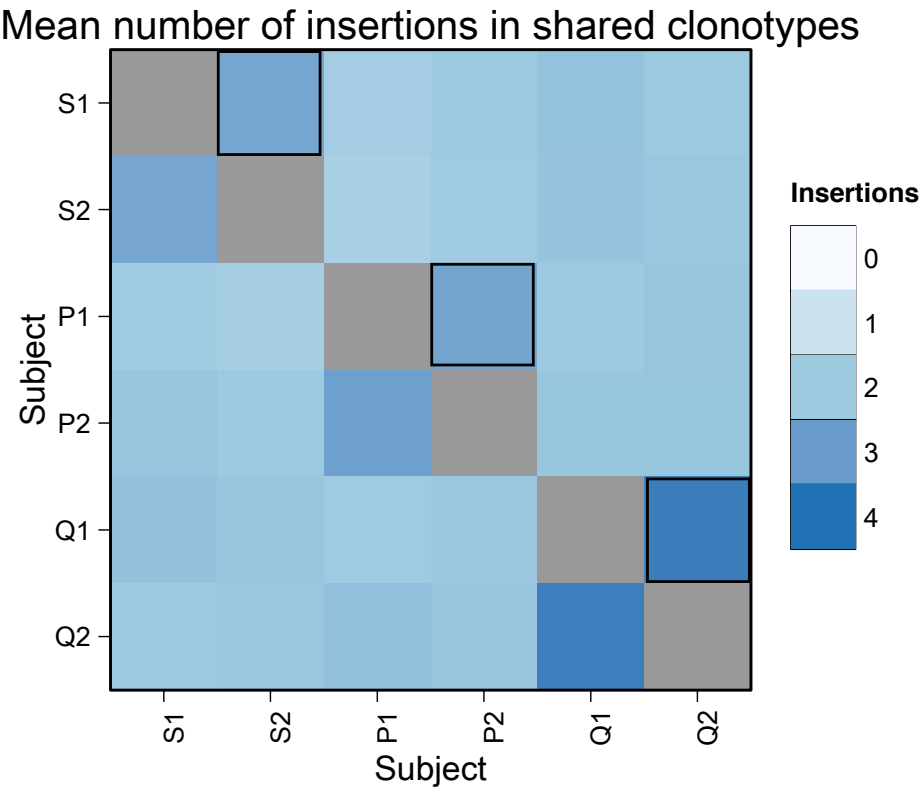

Figure B.7: Mean number of insertions in shared sequences in alpha out-of-frame repertoires of CD45RO+ (memory) cells.

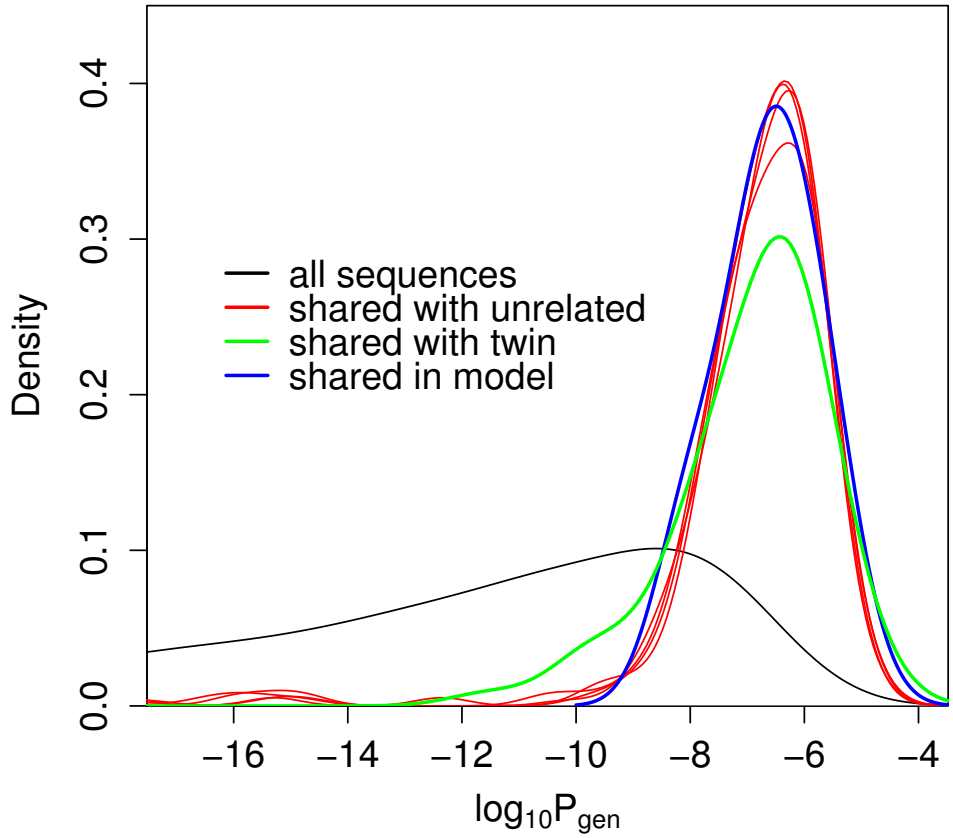

Figure B.8: **Reproducibility of our results using previously published data.** Distribution of P_{gen} – the probability that a sequence is generated by the VJ recombination process – for shared out-of-frame TCR alpha clonotypes between individual A_1 from [206] and the other five individuals. While the distribution of shared sequences between unrelated individuals (red curves) is well explained by coincidental convergent recombination as predicted by our stochastic model (blue curve), sequences shared between two twins (green curve) have an excess of low probability sequences: 68 sequences with $\log_{10} P_{gen} < -10$. For comparison the distribution of P_{gen} in regular (not necessarily shared) sequences is shown in black.

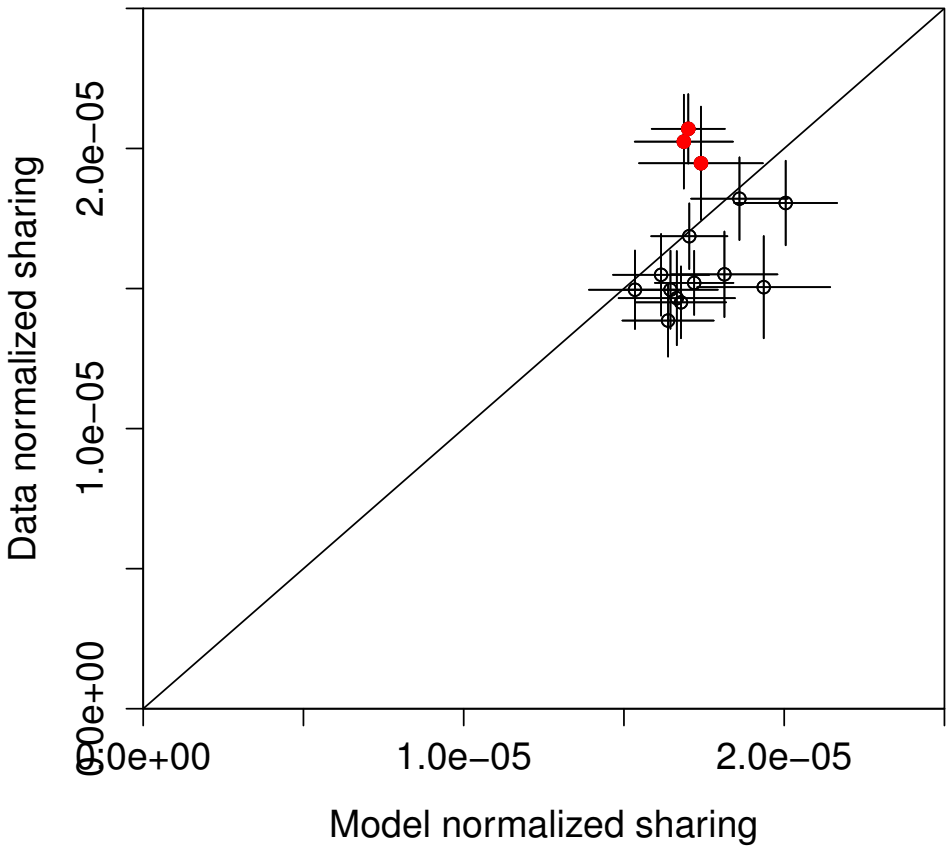

Figure B.9: **Normalized sharing of out-of-frame zero insertion clonotypes.** Number of shared out-frame alpha zero insertion TCR CDR3 clonotypes reported between all 15 pairs of 6 donors consisting of 3 twin pairs (ordinate) compared to the model prediction (abscissa). The three red circles represent the twin pairs, while the black circles refer to pairs of unrelated individuals. Diagonal is equality line. Error bars show one standard deviation.

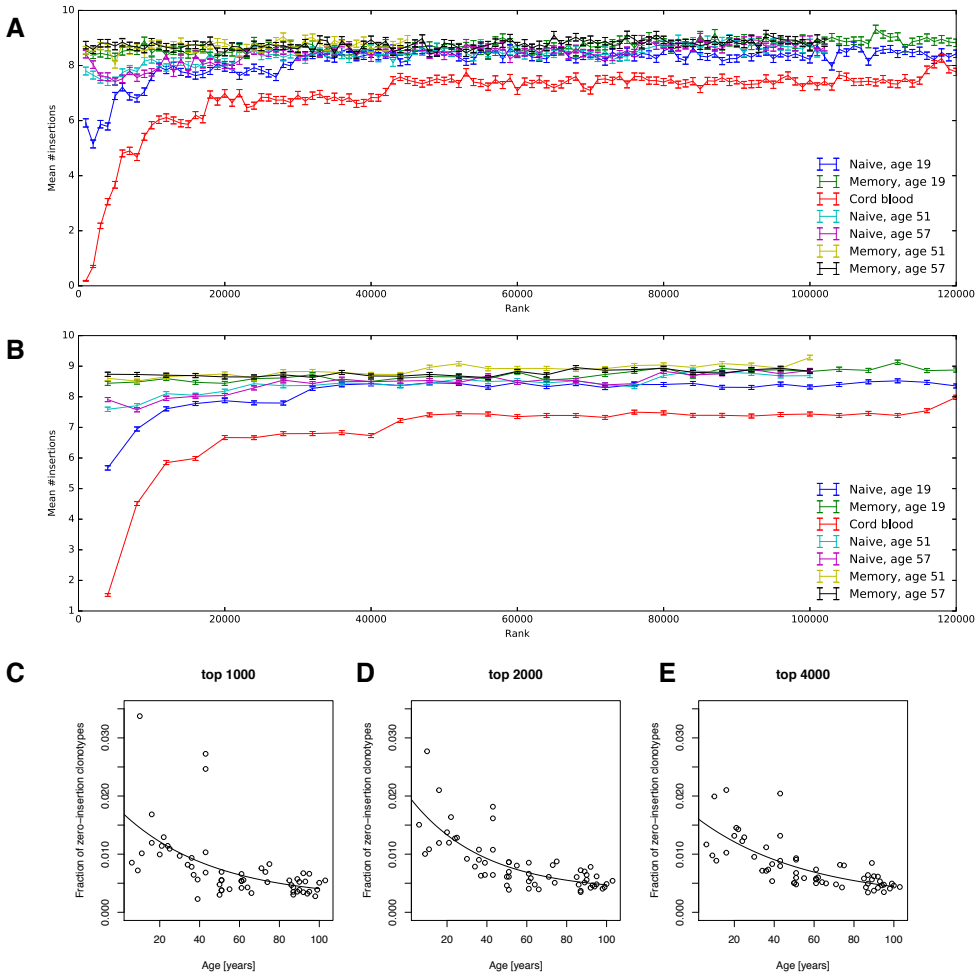

Figure B.10: **Dependence of mean insertion number on rank holds for different bin sizes.** Mean numbers of insertions were obtained by analysing subsequent groups of 1000 (A) and 4000 (B) sequences of decreasing abundances, as in Fig. 5.3A from the main text. (C,D,E) are results for ageing datasets reproduced for the top 1000, 2000 and 4000 clonotypes. Solid lines are independent fits to exponential decays (see main text Methods). Decay rate parameters for top 1000 and top 4000 clones are 0.0218 yr^{-1} and 0.0184 yr^{-1} respectively, within one standard error of the estimate for the top 2000 clones, 0.0272 ± 0.0091 yr^{-1}.

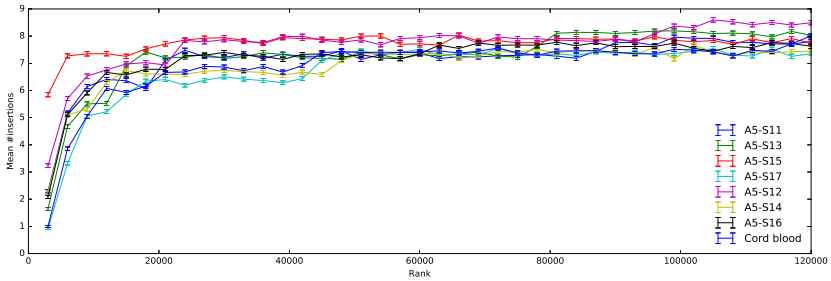

Figure B.11: **The dependence between clone abundance and mean insertion number is robust across cord blood donors.** Mean numbers of insertions were obtained by analysing groups of 3000 sequences of decreasing abundances as in Fig. 5.3A, for 7 independent published cord blood samples from [24]. A similar decreasing trend is observed for all samples.

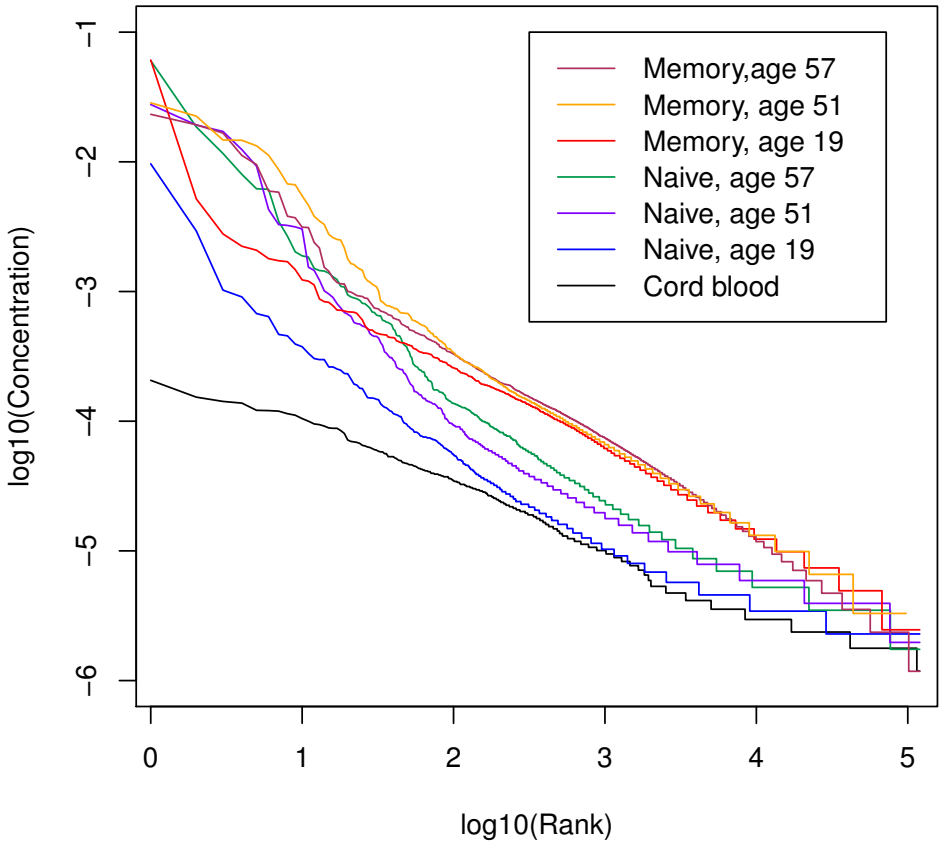

Figure B.12: **Rank-abundance dependencies.** Here we show the dependence of the clone abundance on its abundance rank in samples from Fig. 5.3A. Memory clones are typically larger than the naive and cord blood clones of same rank, possibly due to the history of clonal expansions.

SMART-Mk cap-switching oligonucleotides	
MK-108	CAGUGGUAUCAACGCAGAGUACNNNNNNUAATGCUNNNNNNUCTT(rG)(rG)(rG)(rG)
MK-248	CAGUGGUAUCAACGCAGAGUACNNNNUNNTGGCANNUNNNNNNUCTT(rG)(rG)(rG)(r
MK-253	CAGUGGUAUCAACGCAGAGUACNNNNUNNTTATGNNUNNNNNNUCTT(rG)(rG)(rG)(rG
MK-103	CAGUGGUAUCAACGCAGAGUACNNNNNNUAACGGUNNNNNNUCTT(rG)(rG)(rG)(rG)
MK-257	CAGUGGUAUCAACGCAGAGUACNNNNUNNTTGCGNNUNNNNNNUCTT(rG)(rG)(rG)(r
MK-143	CAGUGGUAUCAACGCAGAGUACNNNNNNUCAGATUNNNNNNUCTT(rG)(rG)(rG)(rG)
MK-135	CAGUGGUAUCAACGCAGAGUACNNNNNNUATGCAUNNNNNNUCTT(rG)(rG)(rG)(rG)
MK-227	CAGUGGUAUCAACGCAGAGUACNNNNUNNTAACCNNUNNNNNNUCTT(rG)(rG)(rG)(r
cDNA synthesis primers	
BC_uni_R4vvshort	TGGAGTCATTGA
TRAC_R2	ACACATCAGAATCCTTACTTTG
PCR I step primers	
Sm1msq	GAGATCTACACGAGTCAGCAGTGGTATCAACGCAG
RPbcj1	CGACTCAGATTGGTACACCTTGTTCAGGTCCTC
RPbcj2	CGACTCAGATTGGTACACGTTTTTCAGGTCCTC
RPacj	CGACTCAAGTGTGTGGGTCAGGGTTCTGGATAT
PCR II step primers	XXXXXX stands for the Truseq index
Sm-out-msq	AATGATACGGCGACCACCGAGATCTACACGAGTCA
Il-bcj-indX	CAAGCAGAAGACGGCATACGAGATXXXXXXCGACTCAGATTGGTAC
Il-acj-indX	CAAGCAGAAGACGGCATACGAGATXXXXXXCGACTCAAGTGTGTGG
Custom sequencing primers	
IL-AIRP	ATATCCAGAACCCTGACCCACACACTTGAGTCG
IL-IRP-b1	GAGGACCTGAAAAACGTGTACCAATCTGAGTCG
IL-IRP-b2	GAGGACCTGAACAAGGTGTACCAATCTGAGTCG
IL-RP1-msq	ACACGAGTCAGCAGTGGTATCAACGCAGAGTAC
IL-RP2-b1	CGACTCAGATTGGTACACGTTTTTCAGGTCCTC
IL-RP2-b2	CGACTCAGATTGGTACACCTTGTTCAGGTCCTC
IL-ARP2	CGACTCAAGTGTGTGGGTCAGGGTTCTGGATAT

Table B.1: List of primers used

Alpha chain			
Sample_id	Number of reads	Number of UMI	Number of unique CDR3nuc
P1_CD4	6566952	430915	248457
P1_CD8	4620425	378044	162607
P1_unpart	9571058	574439	348419
P1_45RO	4099026	431529	173883
P2_CD4	4269624	941176	432476
P2_CD8	4040615	561437	204094
P2_unpart	8213565	873546	471850
P2_45RO	4608991	653326	228429
Q1_CD4	3894188	653649	277621
Q1_CD8	3201067	589757	147918
Q1_unpart	8360990	1091786	456024
Q1_45RO	3587344	687916	201218
Q2_CD4	3877893	828573	315922
Q2_CD8	3880048	825539	158954
Q2_unpart	9159719	1215155	473672
Q2_45RO	3890664	834828	224276
S1_CD4	4655514	734158	360161
S1_CD8	1009038	219433	105232
S1_unpart	3191701	621723	351923
S1_45RO	4977466	495057	189739
S2_CD4	11727155	761495	348109
S2_CD8	12436797	468345	190534
S2_unpart	11135704	610105	336177
S2_45RO	9064981	633362	228579
Beta chain			
Sample_id	Number of reads	Number of UMI	Number of unique CDR3nuc
P1_CD4	3757755	759270	235040
P1_CD8	3565384	517737	204963
P1_unpart	7429601	955106	444708
P1_45RO	4036708	695379	195023
P2_CD4	3042278	449048	475545
P2_CD8	3438238	477696	241048
P2_unpart	8144134	817306	624074
P2_45RO	4598733	578663	249001
Q1_CD4	3694288	673037	386005
Q1_CD8	4586088	758201	237511
Q1_unpart	6511237	1060251	581114
Q1_45RO	3171012	664732	216879
Q2_CD4	3066472	605062	351640
Q2_CD8	3389029	691438	174552
Q2_unpart	7256515	1241753	644594
Q2_45RO	3110044	667997	214628
S1_CD4	3510759	722883	423689
S1_CD8	3162597	489393	248236
S1_unpart	7019324	1181194	673755
S1_45RO	3363725	574876	218185
S2_CD4	4034384	717023	410283
S2_CD8	4267632	546529	258832
S2_unpart	7093628	875357	521882
S2_45RO	2848644	526765	209807
Memory_aged19	7486248	424156	149292

Sample id	fraction of o ins in top 2000	Naive,%	Age, years
A2-i132	0.015056135255	73.7	6
A2-i131	0.010037196444	43	9
A2-i136	0.027691639038	40	10
A2-i129	0.0108412940125	57	11
A2-i134	0.0210075545075	68	16
A2-i133	0.0119257041822	60.9	16
A4-i194	0.013765206508	55	20
A4-i195	0.0119673129492	59	21
A4-i191	0.01637900271	45	22
A4-i192	0.012716977224	56	24
A4-i189	0.012839842368	44	25
A6-I2010b	0.0091925381272	NA	30
A3-i110	0.0078554903232	36.4	34
A3-i101	0.0107838068688	55	36
A4-i101	0.0090257537105	27	36
A4-i102	0.00628983345724	27.6	37
A3-i107	0.00851643362094	43	39
A4-i107	0.0064344051544	26	39
A3-i106	0.016159136094	39.4	43
A3-i102	0.0107591339774	27.3	43
A4-i110	0.018164859228	40	43
A4-i106	0.00642081990976	31	43
A5-S23	0.0046042762969	21.3	50
A5-S24	0.0061143105585	29.9	50
A6-I160	0.008621670788	38.9	51
A5-S21	0.0086245934928	51.3	51
A6-I2150b	0.00819076572358	NA	51
A5-S22	0.00695571384444	48.5	51
A6-I150	0.0061129801278	NA	51
A5-S20	0.00387005779589	25	51
A5-S19	0.0080402564192	41.2	55
A4-i185	0.0085319088075	29.6	61
A4-i186	0.00532914538306	14.6	61
A4-i184	0.00405847825812	21	61
A4-i188	0.00663226556694	18	61
A4-i128	0.0058717051432	23	62
A4-i125	0.00476704046791	4.5	64
A4-i124	0.00394006128853	16.3	66
A2-i141	0.0060990185169	30	71
A2-i140	0.0081195988401	47	73
A2-i138	0.00507840452028	6.7	74
A2-i139	0.008749966888	28.2	75
A4-i122	0.00606575047668	33	85
A3-i145	0.004749303571	37	86
A4-i132	0.0034771649962	14.5	87
A4-i183	0.00723588404502	24.6	87
A3-i150	0.0037069726895	13.3	87
A6-I2140b	0.0046188525124	21	88
A5-S10	0.007023235658	NA	89
A4-i118	0.00512286685575	54	89
A4-i127	0.005589445878	12.7	90
A5-S9	0.00642820638494	26.5	90
A6-I2110b	0.00432554146357	8.4	91
A5-S8	0.00421932231855	4.5	92
A5-S7	0.0078096377085	4.7	92
A6-I2100b	0.00368734455504	7.4	92
A6-I2080b	0.0045677109953	8.7	93
A5-S4	0.0046450251048	30.8	93
A6-I2070b	0.0044350512973	27.6	94
A6-I2060b	0.0061812657375	6.2	95
A6-I2050b	0.00481739413682	7.5	95
A5-S3	0.0040549739527	12.4	98
A6-I2040b	0.00431740407138	10.3	99
A5-S2	0.00486991171424	15.5	100
A5-S1	0.00541415235339	NA	103

Table B.3: Ageing data used for Fig. 4 and exponential decay fits. Percentage of the naive T-cells defined using flow cytometry, see [23] for details.

C

SUPPLEMENTARY MATERIAL FROM *IGOR: A TOOL FOR HIGH-THROUGHPUT IMMUNE REPERTOIRE ANALYSIS*

C.1 SUPPLEMENTARY INFORMATION

C.1.1 *Data and software*

C.1.1.1 *Genomic templates*

We used custom genomic templates derived from the IMGT database [93]. TCR alpha V and J genomic templates were taken from the IMGT human database. For TCR beta V, D and J genes we used curated genomic templates from [115]. BCR heavy chain V, D and J genes were taken from the customized genomic templates used in [50]. For software comparison we used default genomic templates provided with Partis and MiXCR.

C.1.1.2 *Alignments*

Initial alignments to germline genes were performed using the Smith-Waterman algorithm [163], with scores of 5 for matching base pairs, -14 for mismatches, and a 50 gap penalty. Alignments with a score below the following gene dependent threshold were discarded: 50 for TRBV, 0 for TRBD, 10 for TRBJ, 20 for TRAV, 10 for TRAJ, 50 for IGHV, 40 for IGHD, 10 for IGHJ. We also discarded alignments whose score fell below the maximum alignment score (found for this read and segment type), minus the following variable range: 55 for TRBV, 35 for TRBD, 10 for TRBJ, 55 for IGHV, 20 for IGHJ.

The alignment offset (the index of the nucleotide on the read to which the first letter of the undeleted genomic template is aligned) was constrained depending on known primer locations on the J gene.

C.1.2 *Generating synthetic sequences*

Synthetic sequences are generated by randomly drawing scenarios of recombination from the probability distribution in Eq. 3.1 or 3.2. In order to fit the data, the resulting sequences are then cut to mimic the sequencing process (e.g. fixed starting point and fixed read length).

C.1.3 *Comparison to other software*

We benchmarked our method against MiXCR 2.0.2 [15] – a commonly used deterministic alignment method. We used the MiXCR sequence assignment to compute the frequency of gene usage, insertion length, deletions and obtain the distributions shown in Fig. C.8. We also compared to Partis [135] – a recent HMM based model of recombination. Since Partis uses a Viterbi learning

algorithm, we used the most likely assignments it outputs to compute the corresponding probability distribution shown in Fig. C.8. Since Partis is designed to handle BCRs we assessed its performance on the BCR dataset only.

C.2 SUPPLEMENTARY FIGURES

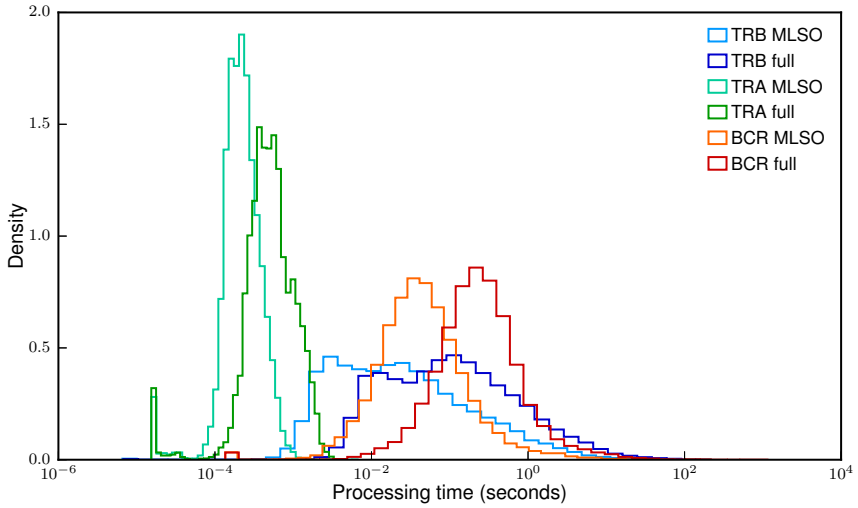

Figure C.1: **Distribution of the processing time per sequence.** Shows the processing time for finding the Most Likely Scenario Only (MLSO) and to evaluate all scenarios (full) for the different chains. Histograms were computed on 20000 sequences for each chain on a single core of an Intel(R) Xeon(R) CPU E5-2680 v3 2.50GHz processor running code compiled with gcc (Debian 4.9.2-10). We benchmarked IGoR's performance for evaluating possible recombination scenarios on real data sequences used to infer the models presented in the main text. We used 60bp TCR β sequences for benchmarking since the difficulty for finding the correct V and J for alignment is higher. Finding the Most Likely Scenario Only(MLSO) is on average 3× faster than evaluating all possible scenarios. Restricting possible scenarios to deterministically assigned V and J genes is on average 6× faster(data not shown).

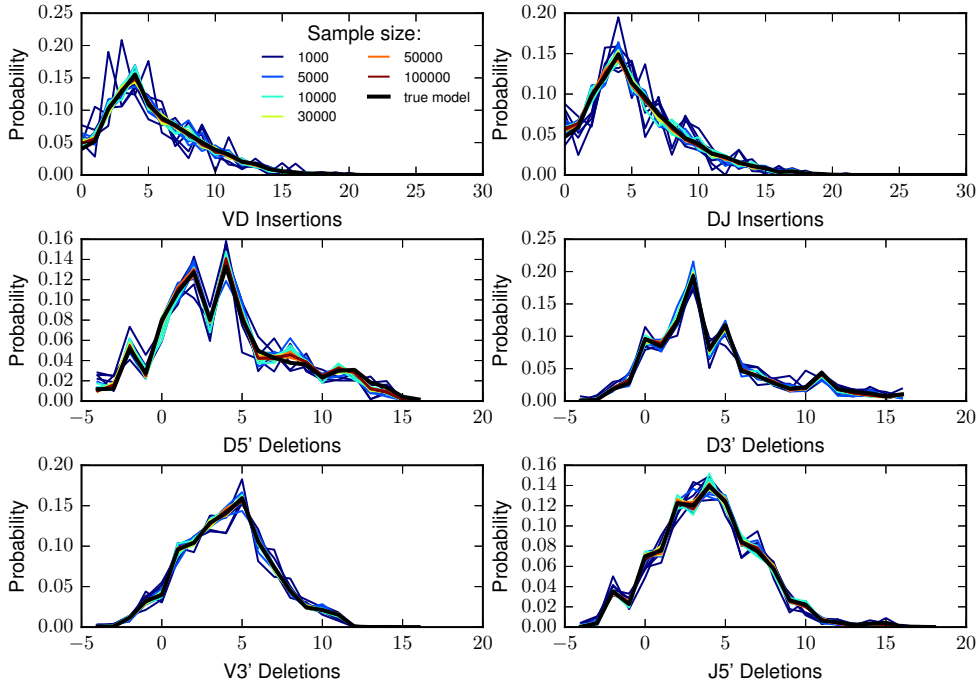

Figure C.2: **Tested on simulated data with a known underlying model Igor converges to the true distribution for different error rates.** We show insertion and deletion distributions obtained from 60bp TCR generated samples of various sizes and with various error rates, to underline qualitative differences hidden by the Kullback-Leibler divergence shown in Fig. 3.4 and Fig. C.5.

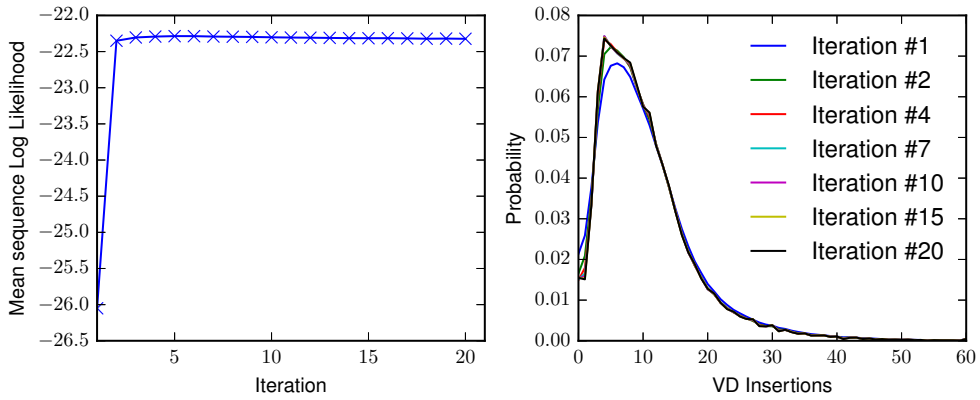

Figure C.3: **Convergence of IGoR for a naive BCR dataset. A.** The mean log likelihood per sequence increases and quickly plateau, thus reaching the maximum likelihood estimate of the parameters. **B.** Convergence of the distribution is shown with the example of the distribution of number of VD insertions.

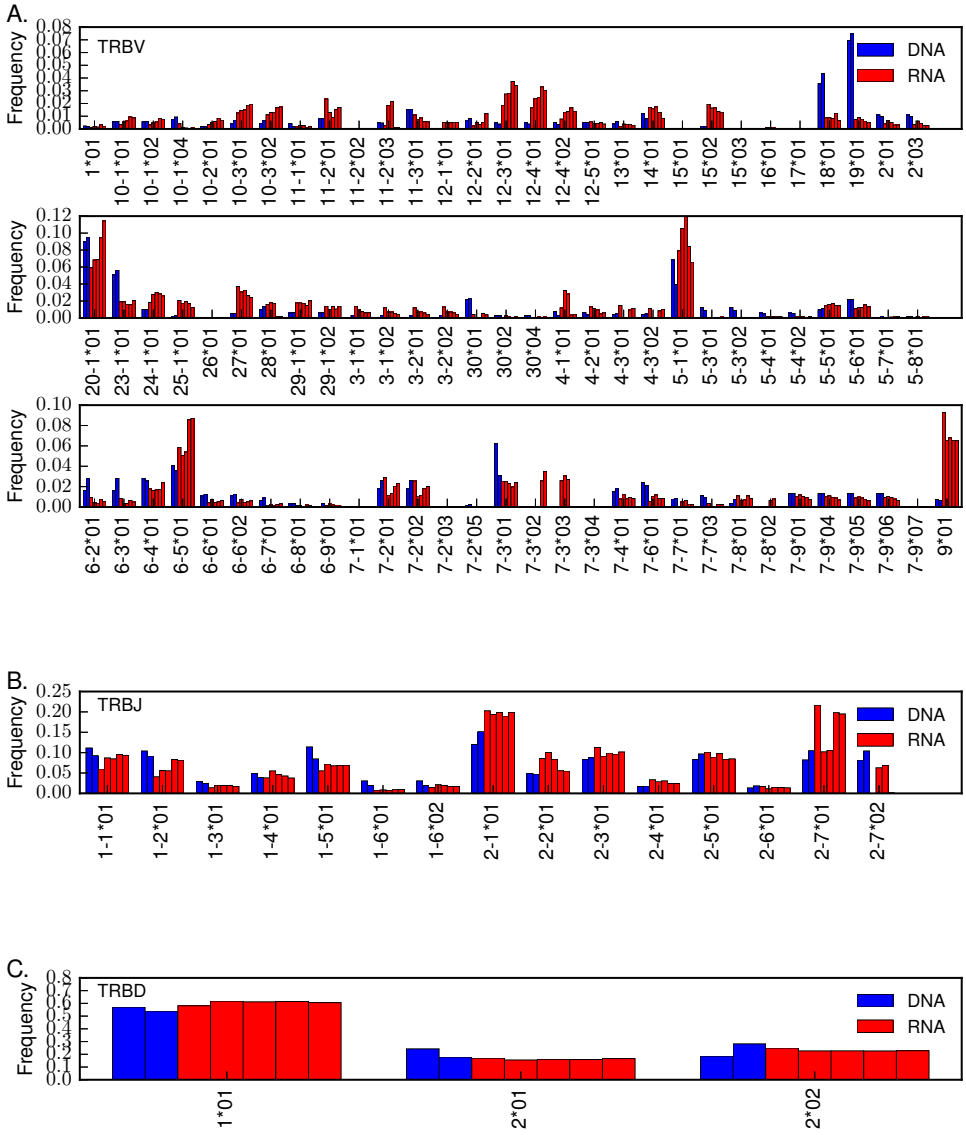

Figure C.4: **Gene usage in TRB mRNA vs DNA data.** We plot the marginal gene usage averaged over conditional dependencies for V, D and J genes respectively inferred using IGoR from mRNA 100bp (red) and DNA 60bp (blue) technology datasets. We observe a higher inter-method than inter-individual variability.

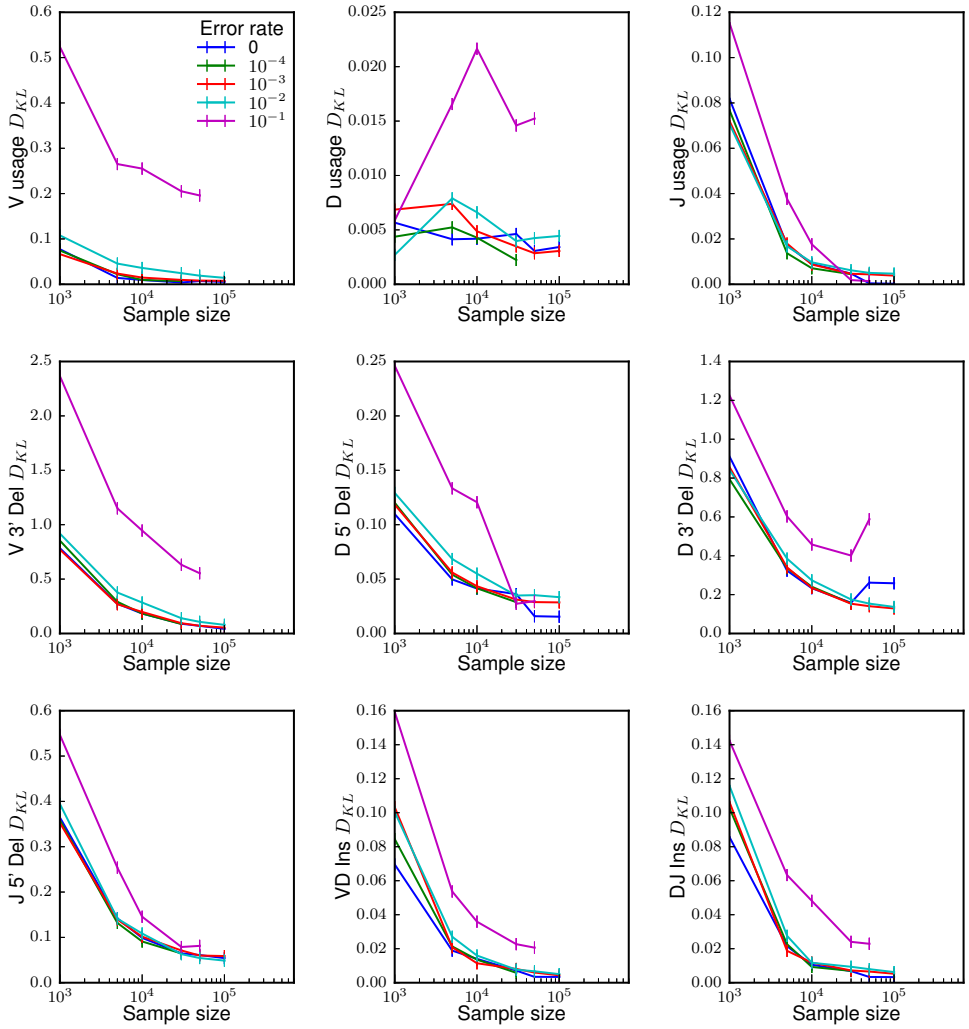

Figure C.5: **Synthetic sampling D_{KL} breakup** Kullback-Leibler divergence (D_{KL}(inferred $\|$ true)) in bits between models inferred on various sample sizes of sequences with various error rates and the ground truth decomposed for the different model components. All components reach a small divergence value for sufficiently large sample sizes.

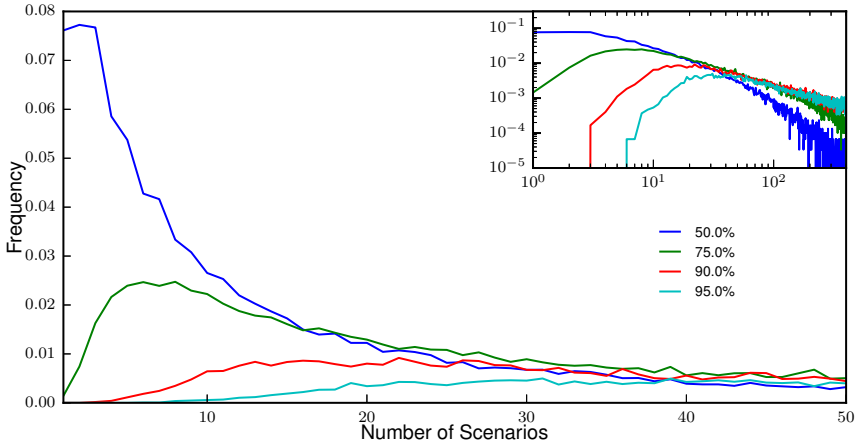

Figure C.6: **A probabilistic assignment approach is crucial for TCRs.** Equivalent of main text Fig. 3.5b for 30000 60bp TCRs. This figure shows the distribution of the number of scenarios that need to be enumerated (from most to least likely) to include the true scenario with 50% (blue), 75% (green), 90% (red), or 95% (cyan) confidence. The shorter read length compared to 130bp BCRs entail a higher uncertainty on the V gene identity, for which a higher number of scenarios must be considered.

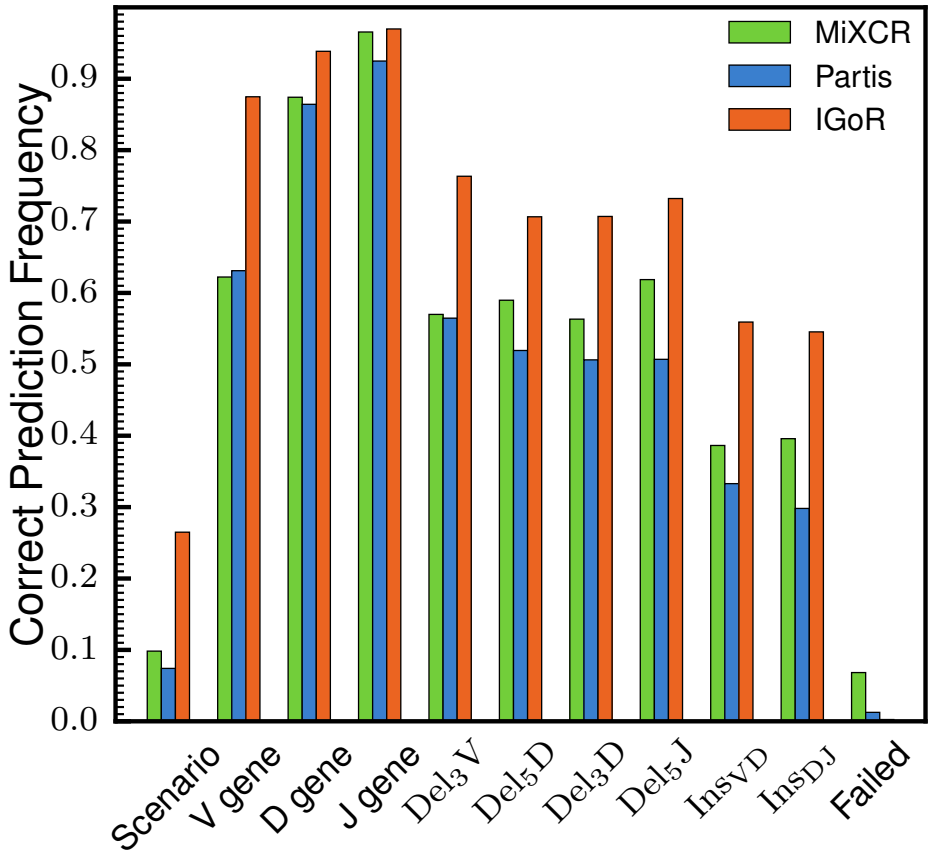

Figure C.7: **Assignment performance on sequences without palindromic insertions** We have shown in main text Fig. 3.5c the ability of MiXCR, Partis and IGoR to predict the correct scenario of recombination. Since Partis does not model palindromic insertions we here present the performance of the three software one sequences that were generated without any. Although Partis' prediction is improved and reaches 7.4% close to MiXCR's 9.8% accuracy, both remain lower than IGoR's 26.5% correct predictions.

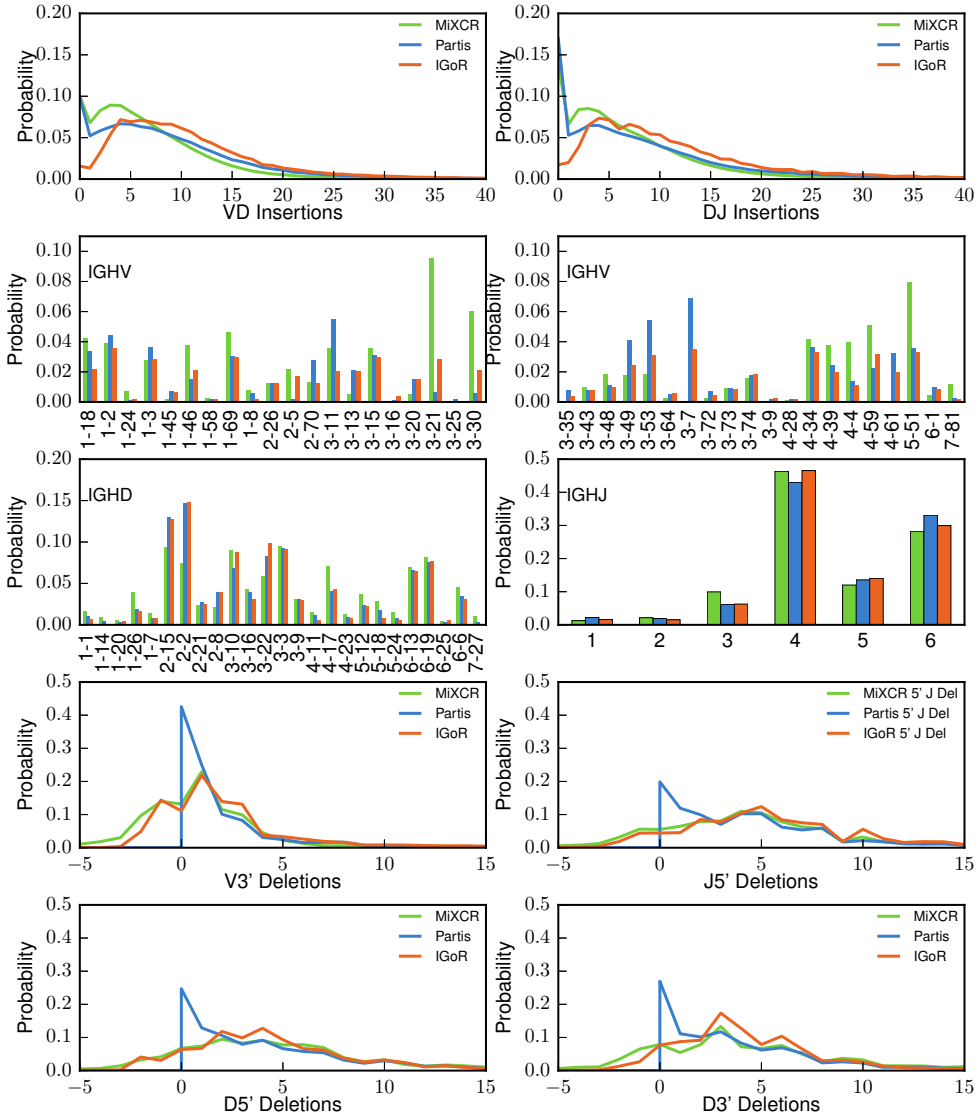

Figure C.8: **Comparison of distributions obtained from different softwares.** MiXCR performing deterministic alignments and Partis Viterbi learning we used the output assignments to compute the corresponding recombination statistics. We plot them along with IGoR's distribution obtained from our maximum likelihood approach. Note that for ease of presentation we show distributions averaged over conditional dependences. From the two top panels we observe that Partis and MiXCR overestimate the frequency of low number of non templated insertions. Gene usage is mostly consistent between methods. In the four bottom panels, negative number of deletions denote palindromic insertions. We observe that the three methods obtain qualitatively different marginal distribution of number of deletions.

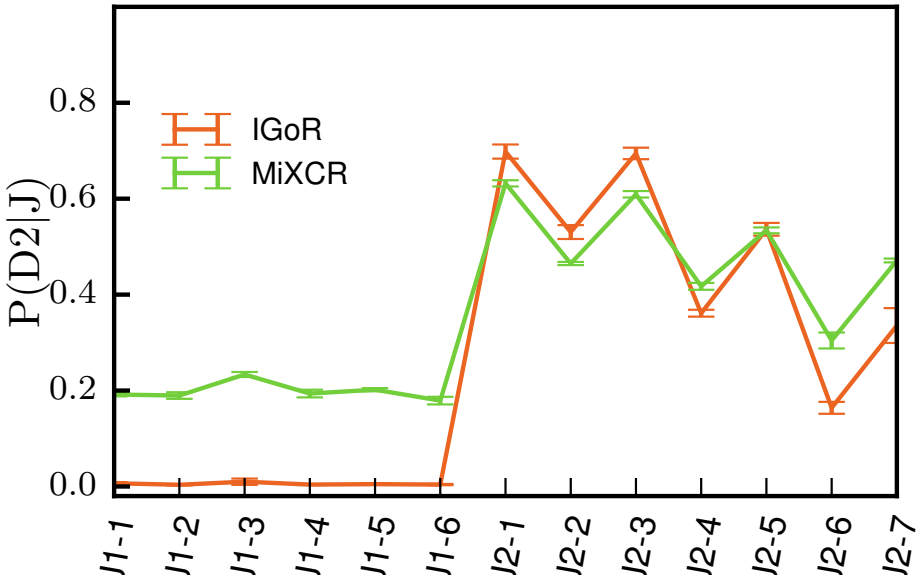

Figure C.9: **Data TRB D2-J association.** As we have shown the D,J pairing rule for TCRs on synthetic data in main text Fig. 3.5D, we plot here the distributions P(D2|J) obtained on real 100bp TCR mRNA data for IGoR and MiXCR. Again, IGoR is able to capture the physiological exclusion between D2 and J1 while MiXCR is not.

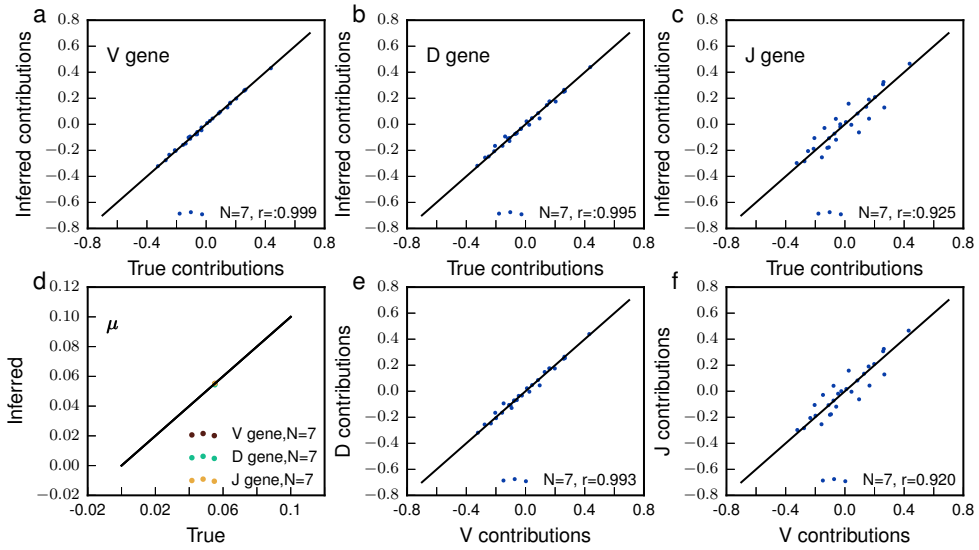

Figure C.10: **Inference of the 7mer hypermutation model on synthetic sequences.** In order to assess the validity of our method we generate synthetic BCRs sequences from a heavy chain model learned on naive data sequences. We then generate Poisson distributed errors on the sequences by simulating mutations at each base pair by a Bernouilli process according to the hypermutation model learned on the V gene of memory sequences. We then cut the sequences in 130bp reads in order to mimic real data sequences. The results of this experiment shows that the model can be perfectly inferred on V and D genes while the scatter on J gene is higher. This can be explained by the limited number of n-mers observed on J gene since sequences were cut to mimic sequencing from a primer in the J.

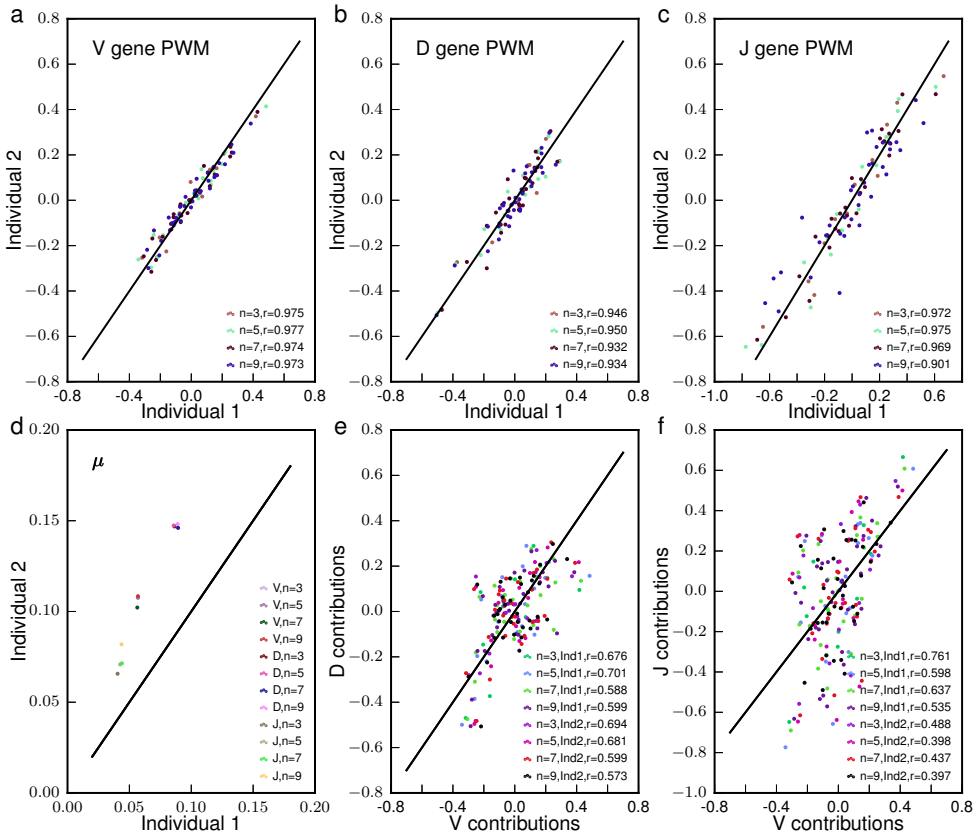

Figure C.11: **Inference of the hypermutation model on real non productive memory BCR sequences. a**, **b** and **c** compare the position weight matrices inferred on respectively V, D and J genes for different n-mer length. For all sizes and gene the inferred contributions are extremely reproducible from an individual to the other. **d** Comparison of the overall mutational load on different individuals and gene for different n-mer size. This overall mutational load varies from individual to individual and within the locus. **e** and **f** Comparison between contributions inferred on different genes. We observe weaker inter gene correlations than the one observed for inter individual contributions.

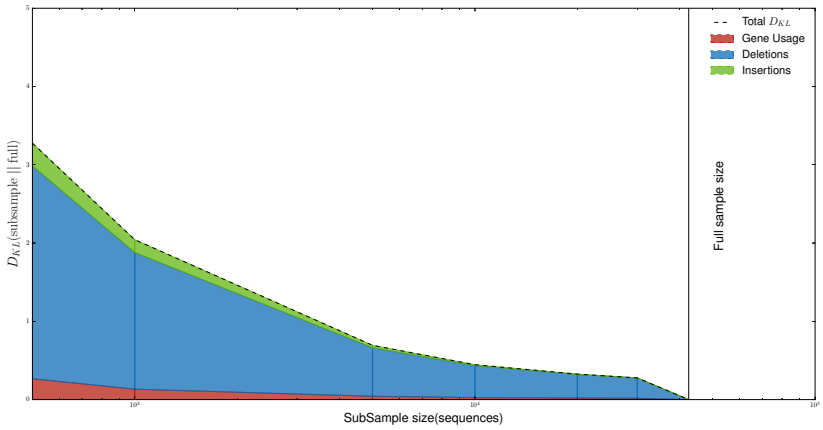

Figure C.12: **Bootstrap** Kullback-Leibler divergence (D_{KL}(subsample || full)) in bits between the model inferred on the full data sample and models inferred on various subsamples sizes.

BIBLIOGRAPHY

[1] Rhys M Adams, Thierry Mora, Aleksandra M Walczak, and Justin B Kinney. "Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves." In: *Elife* 5 (2016), e23156.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. "Basic local alignment search tool." In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[3] Chris Anderson. "The end of theory: The data deluge makes the scientific method obsolete." In: *Wired magazine* 16.7 (2008), pp. 16–07.

[4] Philip W Anderson et al. "More is different." In: *Science* 177.4047 (1972), pp. 393–396.

[5] Jürgen Bachl, Chris Carlson, Vanessa Gray-Schopfer, Mark Dessing, and Carina Olsson. "Increased transcription levels induce higher mutation rates in a hypermutating cell line." In: *The Journal of Immunology* 166.8 (2001), pp. 5051–5057.

[6] Rodney J Baxter. *Exactly solved models in statistical mechanics*. Elsevier, 2016.

[7] C L Benedict, S Gilfillan, T H Thai, and J F Kearney. "Terminal deoxynucleotidyl transferase and repertoire development." In: *Immunol. Rev.* 175.4 (2000), pp. 150–157. ISSN: 0105-2896. DOI: 10.1034/j.1600-065X.2000.017518.x.

[8] Otto G Berg and Peter H von Hippel. "Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters." In: *Journal of molecular biology* 193.4 (1987), pp. 723–743.

[9] Katharine Best, Theres Oakes, James M Heather, John Shawe-Taylor, and Benny Chain. "Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding." In: *Scientific reports* 5 (2015).

[10] Alexander G Betz, César Milstein, Africa González-Fernández, Richard Pannell, Tammy Larson, and Michael S Neuberger. "Elements regulating somatic hypermutation of an immunoglobulin κ gene: critical role for the intron enhancer/matrix attachment region." In: *Cell* 77.2 (1994), pp. 239–248.

[11] Alexander G Betz, Cristina Rada, Richard Pannell, Cesar Milstein, and Michael S Neuberger. "Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots." In: *Proceedings of the National Academy of Sciences* 90.6 (1993), pp. 2385–2388.

[12] Eva Bianconi et al. "An estimation of the number of cells in the human body." In: *Annals of human biology* 40.6 (2013), pp. 463–471.

[13] Valerie Biran et al. "A long-term competent chimeric immune system in a dizygotic dichorionic twin." In: *Pediatrics* 128 (2011), e458–e463. ISSN: 0031-4005. DOI: 10.1542/peds.2010-3557.

[14] Christopher M Bishop. "Pattern recognition." In: *Machine Learning* 128 (2006), pp. 1–58.

[15] Dmitriy A Bolotin et al. "MiXCR: software for comprehensive adaptive immunity profiling." In: *Nature methods* 12.5 (2015), pp. 380–381.

[16] Dmitry a Bolotin et al. "Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms." In: *European Journal of Immunology* 42.11 (Nov. 2012), pp. 3073–3083. ISSN: 00142980. DOI: 10.1002/eji.201242517. URL: http://www.ncbi.nlm.nih.gov/pubmed/22806588%20http://doi.wiley.com/10.1002/eji.201242517.

[17] Peter M Bowers et al. "Nucleotide insertions and deletions complement point mutations to massively expand the diversity created by somatic hypermutation of antibodies." In: *Journal of Biological Chemistry* 289.48 (2014), pp. 33557–33567.

[18] Scott D Boyd et al. "Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements." In: *The Journal of Immunology* 184.12 (2010), pp. 6986–6992.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[20] Ronda Bransteitter, Phuong Pham, Peter Calabrese, and Myron F Goodman. "Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase." In: *Journal of Biological Chemistry* 279.49 (2004), pp. 51612–51621.

[21] Bryan S Briney, Jordan R Willis, Mark D Hicar, James W Thomas, and James E Crowe. "Frequency and genetic characterization of V (DD) J recombinants in the human peripheral blood antibody repertoire." In: *Immunology* 137.1 (2012), pp. 56–64.

[22] Bryan Briney, Khoa Le, Jiang Zhu, and Dennis R Burton. "Clonify: unseeded antibody lineage assignment from next-generation sequencing data." In: *Scientific reports* 6 (2016).

[23] O. V. Britanova et al. "Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling." In: *The Journal of Immunology* 192 (2014), pp. 2689–2698. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1302064. URL: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1302064.

[24] O. V. Britanova et al. "Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians." In: *J. Immunol.* (2016). ISSN: 0022-1767. DOI: 10.4049/jimmunol.1600005. URL: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1600005.

[25] Xavier Brochet, Marie Paule Lefranc, and Véronique Giudicelli. "IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis." In: *Nucleic Acids Res.* 36.Web Server issue (2008), pp. 503–508. ISSN: 13624962. DOI: 10.1093/nar/gkn316.

[26] Florian Buettner et al. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells." In: *Nature biotechnology* 33.2 (2015), pp. 155–160.

[27] Frank Macfarlane Burnet et al. "A modification of Jerne's theory of antibody production using the concept of clonal selection." In: *Australian J. Sci.* 20.3 (1957), pp. 67–9.

[28] Jorg JA Calis and Brad R Rosenberg. "Characterizing immune repertoires by high throughput sequencing: strategies and applications." In: *Trends in immunology* 35.12 (2014), pp. 581–590.

[29] Giuseppe Carleo and Matthias Troyer. "Solving the quantum many-body problem with artificial neural networks." In: *Science* 355.6325 (2017), pp. 602–606.

[30] Juan Carrasquilla and Roger G Melko. "Machine learning phases of matter." In: *Nature Physics* (2017).

[31] Vivek Chandra, Alexandra Bortnick, and Cornelis Murre. "AID targeting: old mysteries and new challenges." en. In: *Trends in Immunology* 36.9 (Sept. 2015), pp. 527–535. ISSN: 14714906. DOI: 10.1016/j.it.2015.07.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S1471490615001702 (visited on 02/03/2017).

[32] Anne Chao. "Nonparametric estimation of the number of classes in a population." In: *Scandinavian Journal of statistics* (1984), pp. 265–270.

[33] Anne Chao and John Bunge. "Estimating the number of species in a stochastic abundance model." In: *Biometrics* 58.3 (2002), pp. 531–539.

[34] Travers Ching et al. "Opportunities And Obstacles For Deep Learning In Biology And Medicine." In: *bioRxiv* (2017), p. 142760.

[35] Mattia Cinelli et al. "Feature selection using a one dimensional na¨ve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires." In: *Bioinformatics* 33.7 (2017), pp. 951–955.

[36] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[37] Toby S Cubitt, David Perez-Garcia, and Michael M Wolf. "Undecidability of the spectral gap." In: *Nature* 528.7581 (2015), pp. 207–211.

[38] Ang Cui et al. "A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data." en. In: *The Journal of Immunology* 197.9 (Nov. 2016), pp. 3566–3574. ISSN: 0022-1767, 1550-6606. DOI: 10.4049/jimmunol.1502263. URL: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1502263 (visited on 02/03/2017).

[39] Pradyot Dash et al. "Quantifiable predictive features define epitope-specific T cell receptor repertoires." In: *Nature* (2017).

[40] Brandon J DeKosky et al. "High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire." In: *Nature biotechnology* 31.2 (2013), pp. 166–169.

[41] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm." In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[42] Jonathan Desponds, Andreas Mayer, Thierry Mora, and Aleksandra M Walczak. "Population dynamics of immune repertoires." In: *arXiv preprint arXiv:1703.00226* (2017).

[43] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. "Fluctuating fitness shapes the clone-size distribution of immune repertoires." In: *Proceedings of the National Academy of Sciences* 113.2 (2016), pp. 274–279.

[44] Javier M Di Noia and Michael S Neuberger. "Molecular mechanisms of antibody somatic hypermutation." In: *Annu. Rev. Biochem.* 76 (2007), pp. 1–22.

[45] L Dienes and EW Schoenheit. "The reproduction of tuberculin hypersensitiveness in guinea pigs with various protein substances." In: *Am. Rev. Tuberc* 20 (1929), p. 92.

[46] Marc Duez, Mathieu Giraud, Ryan Herbert, Tatiana Rocher, Mikaël Salson, and Florian Thonier. "Vidjil: A web platform for analysis of high-Throughput repertoire sequencing." In: *PLoS One* 11.11 (2016), pp. 1–12. ISSN: 19326203. DOI: 10.1371/journal.pone.0166126.

[47] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[48] Yuval Elhanati, Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. "repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data." In: *Bioinformatics* In press (2016). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw112. arXiv: 1511.00107.

[49] Yuval Elhanati, Anand Murugan, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. "Quantifying selection in immune receptor repertoires." In: *Proc. Natl. Acad. Sci.* 111.27 (July 2014), pp. 9875–9880. ISSN: 1091-6490. DOI: 10.1073/pnas.1409572111. URL: http://arxiv.org/abs/1404.4956$%5Cbackslash$nhttp://www.ncbi.nlm.nih.gov/pubmed/24941953%20http://www.pnas.org/content/111/27/9875.abstract.

[50] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. "Inferring processes underlying B-cell repertoire diversity." In: *Phil. Trans. R. Soc. B* 370.1676 (2015), p. 20140243. URL: http://rstb.royalsocietypublishing.org/content/370/1676/20140243.abstract (visited on 02/03/2017).

[51] Ryan O Emerson et al. "Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire." In: *Nature Genetics* 49.5 (2017), pp. 659–665.

[52] Ofer Feinerman, Joël Veiga, Jeffrey R Dorfman, Ronald N Germain, and Grégoire Altan-Bonnet. "Variability and robustness in T cell activation from regulated heterogeneity in protein levels." In: *Science* 321.5892 (2008), p. 1081.

[53] Frank Fenner, Donald Ainslie Henderson, Isao Arita, Zdenek Jezek, Ivan D Ladnyi, et al. *Smallpox and its eradication*. Vol. 6. World Health Organization Geneva, 1988.

[54] A M Ford et al. "In utero rearrangements in the trithorax-related oncogene in infant leukaemias." In: *Nature* 363.6427 (1993), pp. 358–360. ISSN: 0028-0836. DOI: 10.1038/363358a0.

[55] Jules Freund and Katherine McDermott. "Sensitization to horse serum by means of adjuvants." In: *Experimental Biology and Medicine* 49.4 (1942), pp. 548–553.

[56] Simon DW Frost, Ben Murrell, AS Md Mukarram Hossain, Gregg J Silverman, and Sergei L Kosakovsky Pond. "Assigning and visualizing germline genes in antibody repertoires." In: *Phil. Trans. R. Soc. B* 370.1676 (2015), p. 20140240.

[57] Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, and Steven H Kleinstein. "Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles." In: *Proceedings of the National Academy of Sciences* 112.8 (2015), E862–E870.

[58] Bruno A Gaëta, Harald R Malming, Katherine JL Jackson, Michael E Bain, Patrick Wilson, and Andrew M Collins. "iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences." In: *Bioinformatics* 23.13 (2007), pp. 1580–1587.

[59] Jacob D Galson, Andrew J Pollard, Johannes Trück, and Dominic F Kelly. "Studying the antibody repertoire after vaccination: practical applications." In: *Trends in immunology* 35.7 (2014), pp. 319–331.

[60] Jacob D Galson et al. "B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation." In: *Genome medicine* 8.1 (2016), p. 68.

[61] Crispin W.. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer-Verlag, 1994.

[62] J F George and H W Schroeder. "Developmental regulation of D beta reading frame and junctional diversity in T cell receptor-beta transcripts from human thymus." In: *Journal of immunology (Baltimore, Md. : 1950)* 148.4 (1992), pp. 1230–9. ISSN: 0022-1767. URL: http://www.ncbi.nlm.nih.gov/pubmed/1310710.

[63] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. "The promise and challenge of high-throughput sequencing of the antibody repertoire." In: *Nat. Biotechnol.* 32.2 (2014), pp. 158–68. ISSN: 1546-1696. DOI: 10.1038/nbt.2782. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/24441474.

[64] Bram Gerritsen, Aridaman Pandit, Arno C Andeweg, and Rob J De Boer. "RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data." In: *Bioinformatics* 32.20 (2016), pp. 3098–3106.

[65] Veronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. "IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V–J and V–D–J rearrangement analysis." In: *Nucleic acids research* 32.suppl_2 (2004), W435–W440.

[66] Jacob Glanville, Tracy C Kuo, H Von Büdingen, Lin Guey, Jan Berka, and Purnima D Sundar. "Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation." In: *Proc. Natl. Acad. Sci.* 108.50 (2011), pp. 20066–20071. DOI: 10.1073/pnas.1107498108.

[67] Kurt Gödel. "Some basic theorems on the foundations of mathematics and their implications." In: *Collected works* 3 (1951), pp. 304–323.

[68] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." In: *arXiv preprint arXiv:1412.6572* (2014).

[69] Hui Yee Greenaway, Benedict Ng, David A Price, Daniel C Douek, Miles P Davenport, and Vanessa Venturi. "NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination." In: *Immunobiology* 218.2 (Feb. 2013), pp. 213–24. ISSN: 1878-3279. DOI: 10.1016/j.imbio.2012.04.003. URL: http://www.ncbi.nlm.nih.gov/pubmed/22647874.

[70] Namita T. Gupta, Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. "Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data." In: *Bioinformatics* 31.20 (2015), pp. 3356–3358. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv359.

[71] G E Hawes, L Struyk, and P J ven den Elsen. "Differential usage of T cell receptor V gene segments in CD4+ and CD8+ subsets of T lymphocytes in monozygotic twins1." In: *Journal of Immunology* 150 (1993), pp. 2033–2045.

[72] James M Heather, Mazlina Ismail, Theres Oakes, and Benny Chain. "High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities." In: *Briefings in bioinformatics* (2017), bbw138.

[73] James M Heather et al. "Dynamic perturbations of the T-cell receptor repertoire in chronic HIV infection and following antiretroviral therapy." In: *Frontiers in immunology* 6 (2015).

[74] Kenneth B Hoehn, Gerton Lunter, and Oliver G Pybus. "A phylogenetic codon substitution model for antibody lineages." In: *Genetics* 206.1 (2017), pp. 417–427.

[75] Steven A Hofmeyr and Stephanie Forrest. "Immunity by design: An artificial immune system." In: *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 2*. Morgan Kaufmann Publishers Inc. 1999, pp. 1289–1296.

[76] Bryan Howie et al. "High-throughput pairing of T cell receptor and sequences." In: *Science Translational Medicine* 7.301 (2015), 301ra131–301ra131. URL: http://stm.sciencemag.org/content/7/301/301ra131.short (visited on 02/03/2017).

[77] Ning Jiang et al. "Lineage structure of the human antibody repertoire in response to influenza vaccination." In: *Science translational medicine* 5.171 (2013), 171ra19–171ra19.

[78] Amy L Kenter, Satyendra Kumar, Robert Wuerffel, and Fernando Griggera. "AID hits the jackpot when missing the target." en. In: *Current Opinion in Immunology* 39 (Apr. 2016), pp. 96–102. ISSN: 09527915. DOI: 10.1016/j.coi.2016.01.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S0952791516000200 (visited on 02/03/2017).

[79] Thomas B Kepler. "Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors." In: *F1000Research* 2 (2013).

[80] Thomas B Kepler et al. "Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation." In: *Frontiers in immunology* 5 (2014).

[81] Marie J Kidd, Katherine JL Jackson, Scott D Boyd, and Andrew M Collins. "DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes." In: *The Journal of Immunology* 196.3 (2016), pp. 1158–1164.

[82] Marie J Kidd et al. "The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements." In: *The Journal of Immunology* 188.3 (2012), pp. 1333–1340.

[83] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W Kinzler, and Bert Vogelstein. "Detection and quantification of rare mutations with massively parallel sequencing." In: *Proceedings of the National Academy of Sciences* 108.23 (2011), pp. 9530–9535.

[84] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, and Martin Bonke. "Counting absolute numbers of molecules using unique molecular identifiers." In: *Nature methods* 9.1 (2011), pp. 1–5. DOI: 10.1038/nMeth.1778. URL: http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.1778.html.

[85] Kimberly D. Klonowski, Laura L. Primiano, and Marc Monestier. "Atypical VH-D-JH rearrangements in newborn autoimmune MRL mice." In: *The Journal of Immunology* 162.3 (1999), pp. 1566–1572. URL: http://www.jimmunol.org/content/162/3/1566.short (visited on 02/03/2017).

[86] Prashant Kodgire, Priyanka Mukkawar, Justin A North, Michael G Poirier, and Ursula Storb. "Nucleosome stability dramatically impacts the targeting of somatic hypermutation." In: *Molecular and cellular biology* 32.10 (2012), pp. 2030–2040.

[87] Leon Kuchenbecker et al. "IMSEQ—a fast and error aware approach to immunogenetic sequence analysis." In: *Bioinformatics* 31.18 (2015), pp. 2963–2971.

[88] K. Larimore, M. W. McCormick, H. S. Robins, and P. D. Greenberg. "Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing." en. In: *The Journal of Immunology* 189.6 (Sept. 2012), pp. 3221–3230. ISSN: 0022-1767, 1550-6606. DOI: 10.4049/jimmunol.1201303. URL: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1201303 (visited on 02/03/2017).

[89] Daniel J Laydon et al. "Quantification of HTLV-1 clonality and TCR diversity." In: *PLoS computational biology* 10.6 (2014), e1003646.

[90] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444.

[91] Edward S Lee, Paul G Thomas, Jeff E Mold, and Andrew J Yates. "Identifying T Cell Receptors from High-Throughput Sequencing: Dealing with Promiscuity in TCRα and TCRβ Pairing." In: *PLoS computational biology* 13.1 (2017), e1005313.

[92] Marie-Paule Lefranc et al. "IMGT, the international ImMunoGeneTics database." In: *Nucleic acids research* 27.1 (1999), pp. 209–212.

[93] Marie-Paule Lefranc et al. "IMGT®, the international ImMunoGeneTics information system®." In: *Nucleic acids research* 37.suppl 1 (2009), pp. D1006–D1012.

[94] Liesbeth Lewi, Jan Deprest, and Kurt Hecher. "The vascular anastomoses in monochorionic twin pregnancies and their clinical consequences." In: *American Journal of Obstetrics and Gynecology* 208.1 (2013), pp. 19–30. ISSN: 00029378. DOI: 10.1016/j.ajog.2012.09.025. URL: http://dx.doi.org/10.1016/j.ajog.2012.09.025.

[95] Enrico Lopriore et al. "Accurate and simple evaluation of vascular anastomoses in monochorionic placenta using colored dye." In: *Journal of visualized experiments : JoVE* 55 (2011), e3208. ISSN: 1940-087X. DOI: 10.3791/3208. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3230184%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[96] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-Par´s. "How many TCR clonotypes does a body maintain?" In: *Journal of theoretical biology* 389 (2016), pp. 214–224.

[97] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[98] Asaf Madi et al. "T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity." In: *Genome Res.* 24.10 (Oct. 2014), pp. 1603–12. ISSN: 1549-5469. DOI: 10.1101/gr.170753.113. URL: http://www.ncbi.nlm.nih.gov/pubmed/25024161.

[99] Ilgar Z Mamedov, Irina A Shagina, Marya A Kurnikova, Sergey N Novozhilov, Dmitry A Shagin, and Yury B Lebedev. "A new set of markers for human identification based on 32 polymorphic Alu insertions." In: *European journal of human genetics : EJHG* 18.7 (July 2010), pp. 808–14. ISSN: 1476-5438. DOI: 10.1038/ejhg.2010.22. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2987352%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[100] Ilgar Z Mamedov et al. "Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling." In: *Frontiers in immunology* 4 (2013).

[101] Stephan Mandt, Matthew D Hoffman, and David M Blei. "Stochastic Gradient Descent as Approximate Bayesian Inference." In: *arXiv preprint arXiv:1704.04289* (2017).

[102] Quentin Marcou, Irit Carmi-Levy, Coline Trichot, Vassili Soumelis, Thierry Mora, and Aleksandra M Walczak. "A model for the integration of conflicting exogenous and endogenous signals by dendritic cells." In: *arXiv preprint arXiv:1607.07244* (2016).

[103] Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. "IGoR: a tool for high-throughput immune repertoire analysis." In: *arXiv preprint arXiv:1705.08246* (2017).

[104] Andreas Mayer, Vijay Balasubramanian, Thierry Mora, and Aleksandra M Walczak. "How a well-adapted immune system is organized." In: *Proceedings of the National Academy of Sciences* 112.19 (2015), pp. 5950–5955.

[105] Andreas Mayer, Thierry Mora, Olivier Rivoire, and Aleksandra M Walczak. "Diversity of immune strategies explained by adaptation to pathogen statistics." In: *Proceedings of the National Academy of Sciences* (2016), p. 201600663.

[106] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[107] John J Miles, Daniel C Douek, and David A Price. "Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination." In: *Immunol. Cell Biol.* 89.3 (2011), pp. 375–387. ISSN: 0818-9641. DOI: 10.1038/icb.2010.139.

[108] Jeff E Mold et al. "Fetal and adult hematopoietic stem cells give rise to distinct T cell lineages in humans." In: *Science* 330.6011 (2010), pp. 1695–1699. ISSN: 0036-8075. DOI: 10.1126/science.1196509.

[109] James J Moon et al. "Naive CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude." In: *Immunity* 27.2 (2007), pp. 203–213.

[110] Thierry Mora and Aleksandra Walczak. "Quantifying lymphocyte receptor diversity." In: *arXiv preprint arXiv:1604.00487* (2016).

[111] Thierry Mora and Aleksandra M Walczak. "Rényi entropy, abundance distribution, and the equivalence of ensembles." In: *Physical Review E* 93.5 (2016), p. 052418.

[112] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. "Maximum entropy models for antibody diversity." In: *Proceedings of the National Academy of Sciences* 107.12 (2010), pp. 5405–5410.

[113] Supriya Munshaw and Thomas B Kepler. "SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements." In: *Bioinformatics* 26.7 (2010), pp. 867–872.

[114] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. 9th. Garland Science, 2016.

[115] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan. "Statistical inference of the generation probability of T-cell receptors from sequence repertoires." In: *Proc. Natl. Acad. Sci.* 109.40 (Oct. 2012), pp. 16161–16166. ISSN: 1091-6490. DOI: 10.1073/pnas.1212755109. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3479580%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[116] Shalin H Naik et al. "Diverse and heritable lineage imprinting of early haematopoietic progenitors." In: *Nature* 496.7444 (Apr. 2013), pp. 229–232. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature12013%2010.1038/nature12013%20http://www.nature.com/nature/journal/v496/n7444/abs/nature12013.html%7B%5C#%7Dsupplementary-information.

[117] Vadim I Nazarov et al. "tcR: an R package for T cell receptor repertoire advanced data analysis." In: *BMC bioinformatics* 16.1 (Jan. 2015), p. 175. ISSN: 1471-2105. URL: http://www.biomedcentral.com/1471-2105/16/175.

[118] Wilfred Ndifon et al. "Chromatin conformation governs T-cell receptor Jβ gene segment usage." In: *Proceedings of the National Academy of Sciences* 109.39 (2012), pp. 15865–15870.

[119] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants." In: *Learning in graphical models*. Springer, 1998, pp. 355–368.

[120] Michelle A Neller et al. "Naive CD8+ T-cell precursors display structured TCR repertoires and composite antigen-driven selection dynamics." In: *Immunol. Cell Biol.* 93.October 2014 (2015), pp. 1–9. ISSN: 0818-9641. DOI: 10.1038/icb.2015.17. URL: http://dx.doi.org/10.1038/icb.2015.17$%5Cbackslash$nhttp://www.nature.com/doifinder/10.1038/icb.2015.17.

[121] Shu-Kay Ng and Geoffrey J McLachlan. "On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures." In: *Statistics and Computing* 13.1 (2003), pp. 45–55.

[122] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 427–436.

[123] Evert PL van Nieuwenburg, Ye-Hua Liu, and Sebastian D Huber. "Learning phase transitions by confusion." In: *Nature Physics* 13.5 (2017), pp. 435–439.

[124] T Oakes, AL Popple, J Williams, RJ Dearman, I Kimber, and B Chain. "Analysis of antigen-specific responses by high-throughput sequencing of the T cell receptor repertoire." In: *Immunology* 143 (2014), p. 63.

[125] Alan S Perelson and Patrick W Nelson. "Mathematical analysis of HIV-1 dynamics in vivo." In: *SIAM review* 41.1 (1999), pp. 3–44.

[126] Alan S Perelson and George F Oster. "Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination." In: *Journal of theoretical biology* 81.4 (1979), pp. 645–670.

[127] Alan S Perelson and Gérard Weisbuch. *Theoretical and experimental insights into immunology*. Vol. 66. Springer Science & Business Media, 2013.

[128] Andrew Peters and Ursula Storb. "Somatic hypermutation of immunoglobulin genes is linked to transcription initiation." In: *Immunity* 4.1 (1996), pp. 57–65.

[129] Mikhail V Pogorelyy et al. "Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires." In: *PLoS Computational Biology* 13.7 (2017).

[130] Boris T Polyak and Anatoli B Juditsky. "Acceleration of stochastic approximation by averaging." In: *SIAM Journal on Control and Optimization* 30.4 (1992), pp. 838–855.

[131] Qian Qi et al. "Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination." In: *Science Translational Medicine* 8.332 (Mar. 2016), 332ra46–332ra46. ISSN: 1946-6234. DOI: 10.1126/scitranslmed.aaf1725. URL: http://stm.sciencemag.org/cgi/doi/10.1126/scitranslmed.aaf1725.

[132] Máire F Quigley et al. "Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.45 (2010), pp. 19414–9. ISSN: 1091-6490. DOI: 10.1073/pnas.1010586107. URL: http://www.pnas.org/content/107/45/19414.short.

[133] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: http://www.R-project.org/.

[134] Duncan K Ralph and Frederick A Matsen IV. "Likelihood-Based Inference of B Cell Clonal Families." In: *PLoS computational biology* 12.10 (2016), e1005086.

[135] Duncan K. Ralph and Frederick A. Matsen. "Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation." en. In: *PLOS Computational Biology* 12.1 (Jan. 2016). Ed. by Bjoern Peters, e1004409. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004409. URL: http://dx.plos.org/10.1371/journal.pcbi.1004409 (visited on 02/03/2017).

[136] HP Rang, M Maureen Dale, JM Ritter, and PK Moore. "Pharmacology Churchill Livingstone." In: *New York* (2003), pp. 3–4.

[137] Erez Rechavi et al. "Timely and spatially regulated maturation of B and T cell repertoire during human fetal development." In: *Science translational medicine* 7.276 (2015), 276ra25. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aaa0072. URL: http://www.ncbi.nlm.nih.gov/pubmed/25717098.

[138] Rhazes. *A Treatise on the Small-pox and Measles*. Trans. by William Alexander Greenhill. Syndenham Society, 1848.

[139] Herbert Robbins and Sutton Monro. "A stochastic approximation method." In: *The annals of mathematical statistics* (1951), pp. 400–407.

[140] Harlan S Robins et al. "Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells." In: *Blood* 114.19 (2009), pp. 4099–4107.

[141] Harlan S. Robins et al. "Overlap and effective size of the human CD8+ T cell receptor repertoire." In: *Sci. Transl. Med.* 2.47 (Sept. 2010), 47ra64. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.3001442. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3212437%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[142] Igor B Rogozin and Marilyn Diaz. "Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process." In: *The Journal of Immunology* 172.6 (2004), pp. 3382–3384.

[143] Pauline Rouaud et al. "The IgH 3 regulatory region controls somatic hypermutation in germinal center B cells." In: *Journal of Experimental Medicine* 210.8 (2013), pp. 1501–1507.

[144] Florian Rubelt et al. "Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells." In: *Nature communications* 7 (2016), p. 11112. ISSN: 2041-1723. DOI: 10.1038/ncomms11112. URL: http://www.ncbi.nlm.nih.gov/pubmed/27005435.

[145] David Ruppert. *Efficient estimations from a slowly convergent Robbins-Monro process*. Tech. rep. Cornell University Operations Research and Industrial Engineering, 1988.

[146] Stuart Russell, Peter Norvig, and Artificial Intelligence. "A modern approach." In: *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25 (1995), p. 27.

[147] Ruslan Salakhutdinov, Sam T Roweis, and Zoubin Ghahramani. "Optimization with EM and expectation-conjugate-gradient." In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 672–679.

[148] I Sanz. "Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions." In: *the Journal of Immunology* 147.5 (1991), pp. 1720–1729.

[149] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. "Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data." In: *BMC Bioinformatics* 17.1 (2016), p. 125. ISSN: 1471-2105. DOI: 10.1186/s12859-016-0976-y. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4787001%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[150] Stefan O. Schönland et al. "Homeostatic control of T-cell generation in neonates." In: *Blood* 102.4 (2003), pp. 1428–1434. ISSN: 00064971. DOI: 10.1182/blood-2002-11-3591.

[151] Erwin Schrödinger. *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press, 1992.

[152] Ton N Schumacher and Robert D Schreiber. "Neoantigens in cancer immunotherapy." In: *Science* 348.6230 (2015), pp. 69–74.

[153] Nambirajan Seshadri and C-EW Sundberg. "List Viterbi decoding algorithms with applications." In: *IEEE transactions on communications* 42.234 (1994), pp. 313–323.

[154] Zachary Sethna et al. "Insights into immune system development and function from mouse T-cell repertoires." In: *Proceedings of the National Academy of Sciences* 114.9 (2017), pp. 2253–2258.

[155] Dmitriy A Shagin et al. "Application of nonsense-mediated primer exclusion (NOPE) for preparation of unique molecular barcoded libraries." In: *BMC genomics* 18.1 (2017), p. 440.

[156] CE Shannon. "A mathematical theory of communication, bell System technical Journal 27: 379-423 and 623–656." In: *Mathematical Reviews (MathSciNet): MR10, 133e* (1948).

[157] Gary S Shapiro, Katja Aviszus, David Ikle, and Lawrence J Wysocki. "Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition." In: *The Journal of Immunology* 163.1 (1999), pp. 259–268.

[158] Sara Sheehan and Yun S Song. "Deep learning for population genetic inference." In: *PLoS computational biology* 12.3 (2016), e1004845.

[159] Mikhail Shugay, Dmitriy A Bolotin, Ekaterina V Putintseva, Mikhail V Pogorelyy, Ilgar Z Mamedov, and Dmitriy M Chudakov. "Huge overlap of individual TCR beta repertoires." In: *Frontiers in Immunology* 4.466 (2013). ISSN: 1664-3224. DOI: 10.3389/fimmu.2013.00466. URL: http://www.frontiersin.org/t_cell_biology/10.3389/fimmu.2013.00466/full.

[160] Mikhail Shugay et al. "Towards error-free profiling of immune repertoires." In: *Nature methods* 11.6 (2014), pp. 653–655.

[161] Arthur M Silverstein. *A history of immunology*. Academic Press, 2009.

[162] Adrien Six et al. "The past, present and future of immune repertoire biology - the rise of next-generation repertoire analysis." In: *Front. Immunol.* 4.413 (Jan. 2013), p. 413. ISSN: 1664-3224. DOI: 10.3389/fimmu.2013.00413. URL: http://www.frontiersin.org/t%7B%5C_%7Dcell%7B%5C_%7Dbiology/10.3389/fimmu.2013.00413/abstract%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3841818%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[163] T F Smith and M S Waterman. "Identification of Common Molecular Subsequences." In: *J. Mol. Biol.* 147 (1981), pp. 195–197.

[164] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, et al. "Evolutionary information for specifying a protein fold." In: *Nature* 437.7058 (2005), p. 512.

[165] M Margarida Souto-Carneiro, Nancy S Longo, Daniel E Russ, Hongwei Sun, and Peter E Lipsky. "Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER." In: *The Journal of Immunology* 172.11 (2004), pp. 6790–6802.

[166] Edward J Steele. "Somatic hypermutation in immunity and cancer: critical analysis of strand-biased and codon-context mutation signatures." In: *DNA repair* 45 (2016), pp. 1–24.

[167] Ursula Storb. "Why does somatic hypermutation by AID require transcription of its target genes." In: *Adv Immunol* 122 (2014), pp. 253–77.

[168] Michael JT Stubbington et al. "T cell fate and clonality inference from single-cell transcriptomes." In: *Nature methods* 13.4 (2016), pp. 329–332.

[169] Yuxin Sun et al. "specificity, Privacy, and Degeneracy in the cD4 T cell receptor repertoire Following immunization." In: *Frontiers in Immunology* 8 (2017).

[170] Jeroen MJ Tas et al. "Visualizing antibody affinity maturation in germinal centers." In: *Science* (2016), aad3439.

[171] O Teuffel, D R Betts, M Dettling, R Schaub, B W Schäfer, and F K Niggli. "Prenatal origin of separate evolution of leukemia in identical twins." In: *Leukemia* 18.10 (2004), pp. 1624–1629. ISSN: 0887-6924. DOI: 10.1038/sj.leu.2403462.

[172] Niclas Thomas, James Heather, Wilfred Ndifon, John Shawe-Taylor, and Benjamin Chain. "Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine." In: *Bioinformatics* 29.5 (2013), pp. 542–550.

[173] Niclas Thomas et al. "Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence." In: *Bioinformatics* 30.22 (2014), pp. 3181–3188.

[174] Maria a. Turchaninova et al. "Pairing of T-cell receptor chains via emulsion PCR." In: *Eur. J. Immunol.* 43.9 (2013), pp. 2507–2515. ISSN: 00142980. DOI: 10.1002/eji.201343453.

[175] S. Unniraman and D. G. Schatz. "Strand-Biased Spreading of Mutations During Somatic Hypermutation." In: *Science (80-.).* 317.5842 (2007), pp. 1227–1230. ISSN: 0036-8075. DOI: 10.1126/science.1145065. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1172133%7B%%7D5Cnhttp://www.sciencemag.org/cgi/doi/10.1126/science.1145065%7B%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/19628857.

[176] Jason A Vander Heiden et al. "pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires." In: *Bioinformatics* 30.13 (2014), pp. 1930–1932.

[177] Vanessa Venturi, Katherine Kedzierska, Stephen J Turner, Peter C Doherty, and Miles P Davenport. "Methods for comparing the diversity of samples of the T cell receptor repertoire." In: *Journal of immunological methods* 321.1 (2007), pp. 182–195.

[178] Vanessa Venturi, Brian D Rudd, and Miles P Davenport. "Specificity, promiscuity, and precursor frequency in immunoreceptors." In: *Current Opinion in Immunology* (July 2013), pp. 1–7. ISSN: 09527915. DOI: 10.1016/j.coi.2013.07.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0952791513001052.

[179] Vanessa Venturi et al. "Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination." In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18691–18696.

[180] Vanessa Venturi et al. "TCR β-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV." In: *The Journal of Immunology* 181.11 (2008), pp. 7853–7862.

[181] Vanessa Venturi et al. "A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing." In: *The Journal of Immunology* 186.7 (2011), pp. 4285–4294.

[182] Chen Wang et al. "B-cell repertoire responses to varicella-zoster vaccination in human identical twins." In: *Proc. Natl. Acad. Sci.* 112.2 (Dec. 2014), pp. 500–505. ISSN: 1091-6490. DOI: 10.1073/pnas.1415875112. URL: http://www.ncbi.nlm.nih.gov/pubmed/25535378.

[183] George C Wang, Pradyot Dash, Jonathan A McCullers, Peter C Doherty, and Paul G Thomas. "T cell receptor αβ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection." In: *Science translational medicine* 4.128 (2012), 128ra42–128ra42.

[184] Shenshen Wang et al. "Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies." In: *Cell* 160.4 (2015), pp. 785–797.

[185] Yan Wang et al. "Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants." In: *Immunogenetics* 63.5 (2011), pp. 259–265.

[186] Edus H. Warren, Frederick a. Matsen, and Jeffrey Chou. "High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology." In: *Blood* 122.1 (2013), pp. 19–22. ISSN: 15280020. DOI: 10.1182/blood-2013-03-453142.

[187] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[188] Corey T Watson et al. "Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data"." In: *The Journal of Immunology* 198.9 (2017), pp. 3371–3373.

[189] Joshua A Weinstein, Ning Jiang, Richard A White, Daniel S Fisher, and Stephen R Quake. "High-throughput sequencing of the zebrafish antibody repertoire." In: *Science* 324.5928 (2009), pp. 807–810.

[190] J L Wiemels et al. "Prenatal origin of acute lymphoblastic leukemia in children." In: *Lancet* 354 (1999), pp. 1499–1503.

[191] Daniel J Woodsworth, Mauro Castellarin, and Robert a Holt. "Sequence analysis of T-cell repertoires in health and disease." In: *Genome Med.* 5.10 (2013), p. 98. ISSN: 1756-994X. DOI: 10.1186/gm502. URL: http://www.ncbi.nlm.nih.gov/pubmed/24172704.

[192] Xueling Wu et al. "Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1." In: *Science* 329.5993 (2010), pp. 856–861.

[193] Xueling Wu et al. "Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection." In: *Cell* 161.3 (2015), pp. 470–485.

[194] G. Yaari, M. Uduman, and S. H. Kleinstein. "Quantifying selection in high-throughput Immunoglobulin sequencing data sets." en. In: *Nucleic Acids Research* 40.17 (Sept. 2012), e134–e134. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks457. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks457 (visited on 02/03/2017).

[195] Gur Yaari, Jennifer IC Benichou, Jason A Vander Heiden, Steven H Kleinstein, and Yoram Louzoun. "The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales." In: *Phil. Trans. R. Soc. B* 370.1676 (2015), p. 20140242.

[196] Gur Yaari et al. "Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data." In: *Frontiers in Immunology* 4 (2013). ISSN: 1664-3224. DOI: 10.3389/fimmu.2013.00358. URL: http://journal.frontiersin.org/article/10.3389/fimmu.2013.00358/abstract (visited on 02/03/2017).

[197] Xi Yang et al. "TCRklass: a new k-string–based algorithm for human and mouse TCR repertoire characterization." In: *The Journal of Immunology* 194.1 (2015), pp. 446–454.

[198] Jian Ye, Ning Ma, Thomas L Madden, and James M Ostell. "IgBLAST: an immunoglobulin variable domain sequence analysis tool." In: *Nucleic acids research* 41.W1 (2013), W34–W40.

[199] Edouard Yeramian and Edouard Debonneuil. "Probabilistic sequence alignments: realistic models with efficient algorithms." In: *Physical review letters* 98.7 (2007), p. 078101.

[200] Jiangtao Yin, Yanfeng Zhang, and Lixin Gao. "Accelerating expectation-maximization algorithms with frequent updates." In: *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*. IEEE. 2012, pp. 275–283.

[201] Yaxuan Yu, Rhodri Ceredig, and Cathal Seoighe. "LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins." In: *Nucleic acids research* 44.4 (2015), e31–e31.

[202] Lenka Zdeborová. "Machine learning: New tool in the box." In: *Nature Physics* 13.5 (2017), pp. 420–421.

[203] Shu-Qi Zhang et al. "Direct measurement of T cell receptor affinity and sequence from na¨ve antiviral T cells." In: *Science translational medicine* 8.341 (2016), 341ra77–341ra77.

[204] Wei Zhang et al. "IMonitor: a robust pipeline for TCR and BCR repertoire analysis." In: *Genetics* 201.2 (2015), pp. 459–472.

[205] Wei Zhang et al. "IMPre: an accurate and efficient software for prediction of T-and B-cell receptor germline genes and alleles from rearranged repertoire data." In: *Frontiers in immunology* 7 (2016).

[206] Ivan V Zvyagin et al. "Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing." In: *Proc. Natl. Acad. Sci. U. S. A.* 111.16 (Apr. 2014), pp. 5980–5. ISSN: 1091-6490. DOI: 10.1073/pnas.1319389111. URL: http://www.ncbi.nlm.nih.gov/pubmed/24711416.

INDEX

Abstract

An individual's adaptive immune system needs to face repeated challenges of a constantly evolving environment with a virtually infinite number of threats. To achieve this task, the adaptive immune system relies on large diversity of B-cells and T-cells, each carrying a unique receptor specific to a small number of pathogens. These receptors are initially randomly built through the process of V(D)J recombination. This initial generated diversity is then narrowed down by a step of functional selection based on the receptors' folding properties and their ability to recognize self antigens. Upon recognition of a pathogen the B-cell will divide and its offsprings will undergo several rounds of successive somatic hypermutations and selection in an evolutionary process called affinity maturation.

This work presents principled probabilistic approaches to infer the probability distribution underlying the recombination and somatic hypermutation processes from high throughput sequencing data using IGoR - a flexible software developed throughout the course of this PhD. IGoR has been developed as a versatile research tool and can encode a variety of models of different biological complexity to allow researchers in the field to characterize evermore precisely immune receptor repertoires. To motivate this data-driven approach we demonstrate that IGoR outperforms existing tools in accuracy and estimate the sample sizes needed for reliable repertoire characterization. Finally, using obtained model predictions, we show potential applications of these methods by demonstrating that homozygous twins share T-cells through cord blood, that the public core of the T cell repertoire is formed in the pre-natal period and finally estimate naive T cell clone lifetimes in human.

Keywords

biophysics, immunology, inférence, V(D)J recombination, hypermutations

Résumé

Le système immunitaire de chaque individu doit faire face à des agressions répétées d'un environnement en constante évolution, constituant ainsi un nombre de menaces virtuellement infini. Afin de mener ce rôle à bien, le système immunitaire adaptatif s'appuie sur une énorme diversité de lymphocytes T et B. Chacune de ces cellules exhibe à sa surface un récepteur unique, créé aléatoirement via le processus de recombinaison V(D)J, et spécifique à un petit nombre de pathogènes seulement. La diversité initiale générée lors de ce processus de recombinaison est ensuite réduite par une étape de sélection fonctionnelle basée sur les propriétés de repliement du récepteur ainsi que ses capacités à interragir avec des proteines du soi. Pour les cellules B, cette diversité peut être à nouveau étendue après rencontre d'un pathogène lors du processus de maturation d'affinité durant lequel le récepteur subit des cycles successifs d'hypermutation et sélection.

Ces travaux présentent des approches probabilistes visant à inferrer les distributions de probabilités sous-tendant les processus de recombinaison et d'hypermutation à partir de données de séquençage haut débit. Ces approches ont donné naissance à IGoR, un logiciel polyvalent dont les performances dépassent celles des outils existants. En utilisant les modèles obtenus comme base, je présenterai comment ces derniers peuvent être utilisés afin d'étudier le vieillissement et évolution du répertoire immunitaire, la présence d'emprunte parentale lors de la recombinaison V(D)J ou encore pour démontrer que les jumeaux échangent des lymphocytes au cours de la vie fœtale.

Mots Clés

biophysique, immunologie, inférence, recombinaison V(D)J, hypermutations

UNIVERSITÉ PARIS DESCARTES

USPC Université Sorbonne Paris Cité