Immune repertoire sequencing with application to infectious disease

Paul Louis Lindau

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Philip D. Greenberg, Chair

Harlan S. Robins

Michael Emerman

Program authorized to offer degree:

Molecular and Cellular Biology

University of Washington

Abstract

Immune repertoire sequencing with application to infectious disease

Paul Louis Lindau

Chair of the Supervisory Committee:

Professor Philip Greenberg

Immunology

B and T lymphocytes are the cellular effectors of the adaptive immune system and perform the learning and recall functions that are the basis of immunological memory. Immunoglobulin and T cell receptors enable the immune system to recognize an enormous set of exogenous and endogenous antigens. The vast majority of B and T cell clones possess a single unique heterodimeric receptor with a highly diverse binding domain generated through ordered somatic gene rearrangements known as V(D)J recombination. The combination of different (Variable, Diversity and Joining) gene segments, insertion and deletion of non-templated nucleotides at the junctions between segments and pairing of heavy and light chains control the specificity of recognition. As a result, the diversity of B and T cell receptors in the body determine an individual's ability to respond to new and previously seen antigens. In this work, I present a database of B cell receptor sequences that unites experimental and computational techniques to accurately estimate the richness of the naïve and memory B cell repertoires. I also present two studies exploring the diversity of the T cell repertoire in acute and chronic

viral infections. To study the diversity of the T cell repertoire in response to an acute infection, I combine live-attenuated yellow fever virus vaccination and T cell repertoire sequencing to identify and track vaccine responsive clones longitudinally. Lastly, I explore the hypothesis that chronic cytomegalovirus infection in the elderly compromises immune function by reducing $CD8^+$ T cell repertoire diversity. Together these studies further our understanding of the relationship between immune repertoire diversity and immune function.

Table of Contents

1 Acknowledgements

First, I would like to thank my advisor, Harlan Robins. Harlan has been incredibly supportive of both my personal and professional goals. He took a risk in allowing me to learn computational biology without any prior experience and I am very grateful. Harlan's mentorship has enabled me to become a better scientist and develop the skills to become an independent investigator.

Secondly, I would like to thank my committee, Phil Greenberg, Michael Emerman, Jesse Bloom and Marshall Horwitz. Together they provided valuable insights and guidance on all of my projects and helped me develop my oral and written presentation skills. I am thankful for the time an energy that each member dedicated to my development.

The computational biology department at the Hutch has been very supportive. I especially want to thank Melissa Alvendia. Without her help, none of my projects would have succeeded. The MCB and MSTP have also been supportive through the years. I am grateful to Maia Low and Marcie Buckner for their assistance and willingness to answer my many questions.

I would also thank Charles Chan, Lauren Ehrlich, Agnieszka Czechowicz, Rob Negrin, Rob Lowsky and Irv Weissman. Without their support, I would not be in the UW MSTP. To this day they continue to provide me with advice and guidance.

Lastly, I want to thank my family, fiancée and friends. Their constant and unwavering support has inspired me throughout this process. I am incredibly grateful for the generosity of my parents whose sacrifices have enabled me to achieve my goals. I am grateful to my sister, Rachel, for her encouragement and emotional support. Thank you to my fiancée, Maren, for her sacrifices, support, guidance and encouragement. I also want to express my gratitude to my friends for providing me with welcome distractions from my work. Thank you all.

This thesis is dedicated to my grandparents.

2 Introduction

2.1 Summary

B and T cells of the adaptive immune system are critical for protecting the body from pathogenic infections. Each B and T cell possesses a highly specific antigen receptor formed from somatic recombination of different gene segments(*1-4*). The diverse collection of B and T cells enable the adaptive immune system to recognize innumerable pathogens(*5, 6*). A major advance has been the development of methods to sequence antigen receptors at high-throughput. This technology has allowed us to explore the influence of immune repertoire diversity on overall immune function. Here I present a study on the diversity of the B cell repertoire in healthy adults as well as studies examining the diversity of the T cell repertoire in acute and chronic viral infections. I discuss the assembly of B and T cell receptors and the role of diversity in adaptive immunity. Lastly, I describe T cell responses to viral infections with a focus on the diversity of the antigen-specific T cell repertoire. This work has implications for using the adaptive immune repertoire to diagnose and test for infectious diseases.

2.2 Background

Adaptive immunity

The adaptive immune system originated in jawless vertebrates where organs and tissues with large numbers of lymphocyte-like cells have been identified(*7*). These cells bear receptors containing leucine-rich repeats assembled by differential recombination, which

recognize bacterial and blood cell antigens. In jawed vertebrates, including mammals, the adaptive immune system is composed of B and T cells that possess antigen receptors produced through the somatic recombination of gene segments(*8*). The richness of this repertoire of receptors enables the adaptive immune system to recognize previously unseen antigens. Another advantage of adaptive immunity is formation and encoding of antigen-specific immunological memories. After the resolution of an adaptive immune response, memory cells are produced that allow for the rapid elimination of previously encountered pathogens, forming the basis for vaccination(*9, 10*). In humans, a functionally heterogeneous collection of naïve and memory B and T cell subsets coupled with a highly diverse repertoire of antigen receptors provides immunity against innumerable pathogens(*11-14*). As antigen recognition is necessary for B and T cell function, this work focuses on human B and T cell receptor repertoires.

B cells

The primary function of B cells is to secrete antibodies, soluble molecules that bind and neutralize foreign antigens. In addition, B cells are critical for the initiation of T cell responses and the regulation of inflammation(*15*). There are two types of B cells, B 1 and B 2, which are classified based on the type of antibody produced and anatomic location. B 1 B cells are rare, reside in the peritoneal and pleural cavities and responsible for the production of “natural antibodies” which accumulate in the absence of infection(*16*). These antibodies are of the modest affinity immunoglobulin M (IgM) isotype and tend to be polyreactive and anti-microbial. The majority of B cells are of the conventional B 2 type, which circulate through the blood and reside in the follicles of the

lymph nodes and spleen(*17*). Mature naïve B cells respond to T cell dependent foreign antigens by either directly differentiating into antibody-secreting plasmablasts or entering into germinal center (GC) reactions. The GC reaction results in the generation of higher affinity antibodies with more diverse functions. After resolution of a primary immune response, immunological memory is encoded in GC-derived memory B cells, which possess the capacity to rapidly differentiate into both higher affinity memory B cells and antibody-secreting plasma cells upon subsequent exposure to antigen(*18*).

The B cell or immunoglobulin receptor

The B cell receptor (BCR) heterodimer is formed from the immunoglobulin heavy (IgH) and immunoglobulin light (IgL) chains. The *IgH* locus is located on chromosome 14 and comprised of 51 functional variable (V), 25 diversity (D) and 6 joining (J) gene segments(*19*). Beginning in the bone marrow at the pro-B cell stage, the RAG endonuclease initiates ordered somatic recombination of V, D and J genes at the *IgH* locus(*20*). RAG recognizes conserved recombination signal sequences between a pair of coding gene segments and cleaves the intervening DNA resulting in a double-stranded break (DSB). The repair of this DSB by classical nonhomologous end joining and the insertion of non-templated (N) nucleotides by terminal deoxynucleotidyl transferase (TdT) at V/D and D/J junctions yield a heavy chain(*21*). Productive in-frame IgH rearrangements capable of pairing with surrogate light chains enable a pro-B cell to develop into a pre-B cell and undergo light chain rearrangement. The two *IgL* loci, *Ig*κ and *Ig*λ, are located on chromosome 2 and 22 respectively and each is comprised of approximately 30-40 V and 4-5 J gene segments. Using a similar mechanism to the IgH

locus, the light chain is formed from VJ somatic rearrangement, however, if the rearrangement is non-productive or cannot pair with the IgH then successive Igκ or Igλ VJ rearrangements can occur until a productive BCR is assembled(*22*). Immature B cells exit the bone marrow expressing a membrane bound BCR of the IgM and IgD isotype.

Immunoglobulin modifications

Upon encountering cognate antigen in secondary lymphoid organs, naïve B cells become activated and gain the ability to produce a secreted antibody. In the context of a T cell dependent immune response, concurrent T and B cell activation leads to the formation of germinal centers(*23*). Germinal centers are the sites of somatic hypermutation (SHM), a process that further diversifies the BCR in order to increase affinity for antigen. SHM occurs on both the IgH and IgL loci, primarily at the three hypervariable complementarity determining regions (CDRs) and is facilitated by activation-induced cytidine deaminase (AID)(*24*). This process results in nucleotide substitutions in the BCR by deaminating cytidines to uridines in single-stranded DNA and leveraging the error-prone (U/G) mismatch repair mechanisms. If the nucleotide substitution increases the antigen binding affinity of the BCR, it is positively selected. B cell affinity maturation is essential to the production of high affinity antibodies against numerous pathogens(*25*).

In addition to SHM, class switch recombination (CSR) frequently occurs after antigen dependent B cell activation. CSR replaces the constant region exon of the IgH with another to tailor the antigen elimination function of the secreted antibody to the type of pathogen. In humans, the *IgH* locus contains a set of nine constant region exons, or isotypes, comprising five functionally distinct classes(Table 1)(*26*). AID catalyzes CSR

by inducing DSBs in the switch introns upstream of IgH constant region exons(*27*). These DSBs are fused by deletional end-joining, which removes the intervening DNA resulting

Isotype Class	Structure	Primary Function
IgA	monomer or dimer	Pathogen neutralization at mucosal surfaces
IgD	monomer	None known
IgE	monomer	Activation of mast cells, basophils and eosinophils
IgG	monomer	Pathogen neutralization in the blood
IgM	pentamer	Opsonisation and complement fixation

Table 1: Antibody Isotype Classes and Functions(26)

in a permanent change of the antibody isotype.

T cells

T cells primarily function to eliminate infected cells and orchestrate innate and adaptive immune responses. The majority of T cells are separated into functionally distinct lineages by CD4 or CD8 co-receptor expression(*28*). $CD4^+$ T cells recognize peptide antigens 13-25 amino acids in length derived from endocytosed molecules displayed on major histocompatibility complex (MHC) II proteins by dendritic cells and B cells(*3*). Depending upon the type of infection and inflammatory environment, $CD4^+$ T cells are polarized into different subsets with highly specialized functions(*29*). These include the recruitment and activation of phagocytes as well as promoting the production of particular antibody isotypes. In contrast to $CD4^+$ T cells, $CD8^+$ T cells recognize peptides 8-10 amino acids in length derived from intracellular proteins and displayed on MHC I, which is present on all nucleated cells. $CD8^+$ T cells function to eliminate the replication of intracellular pathogens by secreting effector cytokines such as interferon-γ (IFN-γ), and producing perforin and granzyme B, which induce apoptosis of infected cells(*30*). $CD8^+$ T cells are especially important in viral infections, where they curtail the spread of virus

by eradicating infected host cells. Depending on the class of pathogen, $CD4^+$ and $CD8^+$ T cells form antigen-specific memory cells after the resolution of an infection with the

Memory T cell Type	Phenotype	Infection type	Cytokines
$CD8^+$ Effector Memory	$CD45RO^+CCR7^-$	Intracellular virus and bacteria	IFN-γ, IL-4, IL-5
$CD8^+$ Central Memory	$CD45RO^+CCR7^+$	Intracellular virus and bacteria	IL-2
$CD4^+$ T_H1	$CD45RO^+CXCR3^+CCR4^-CCR6^-$	Intracellular virus and bacteria	IFN-γ
$CD4^+$ T_H2	$CD45RO^+CXCR3^-CCR4^+CCR6^-$	Paracites and venoms	IL-4, IL-5, IL-13
$CD4^+$ T_H17	$CD45RO^+CXCR3^-CCR4^+CCR6^+$	Fungi and extracellular bacteria	IL-17, IL-22

Table 2: Memory T cell types and functions(12, 31)

capability to rapidly proliferate when a pathogen is reencountered(Table 2)(*31*).

The T cell receptor

The T cell receptor (TCR) is a membrane-bound heterodimer composed of either α and β or γ and δ chains. Roughly 90% of T cells in the periphery bear an $\alpha\beta$ TCR. Similar to the *IgH* locus the TCRβ (*Tcrb*) and TCRδ (*Tcrd*) chain loci contain V, D and J gene segments and are located on chromosomes 7 and 14, respectively. The *Tcrb* locus is composed of 40-48 V, 2 D and 12-13 J gene segments while the *Tcrd* locus is composed of 7-8 V, 3 D and 4 J gene segments. The TCRα (*Tcra*) and TCRγ (*Tcrg*) chain loci are similar to the IGL and located on chromosomes 14 and 7, respectively. The *Tcra* locus is comprised of 45-47 V and 50 J gene segments and the *Tcrg* locus is comprised of 4-6 V and 5 J gene segments(*32*). Beginning in the thymus at the $CD4^-$ $CD8^-$ $CD44^+$ $CD25^-$ double negative 2 (DN2) developmental stage, the *Tcrb*, *Tcrd* and *Tcrg* loci begin to rearrange employing a recombination mechanism identical to that of B cells. The signal strength delivered via the TCR at the $CD4^-$ $CD8^-$ $CD44^-$ $CD25^+$ double negative 3 (DN3) stage of a cell expressing a productively rearranged $\gamma\delta$ TCR or a TCRβ chain paired with the invariant pre-Tα chain determine lineage commitment(*33, 34*). Following $\alpha\beta$ T cell

lineage commitment, the *Tcra* locus rearranges at the $CD4^+$ $CD8^+$ double positive (DP) developmental stage. Similar to the IgL, the *Tcra* locus can undergo successive rearrangements until a productive TCRα chain is generated. After a functional $\alpha\beta$ TCR is produced, binding to MHC and self-peptide determine the fitness of a T cell clone to exit the thymus and enter the peripheral circulation(*35*).

B and T cell receptor repertoires

The nearly random combination of V, D and J gene segments coupled with the random insertion and deletion of nucleotides at the junctions of joined segments is capable of generating more than 10^{13} unique BCRs or TCRs(*36*). The result of this extraordinary genetic diversity is that each B or T cell completing development and entering the periphery bears a single unique antigen receptor. Consequently, B and T cells are derived from clones that are defined by the nucleotide sequence of their BCR or TCR. The B or T cell repertoire is the number of distinct B or T cell clones in a particular tissue or sample(*37*). Homeostatic proliferation and antigen encounter result in B and T cell clonal expansions, which alter the frequencies of clones in the repertoire(*38, 39*). In addition, age-related changes and turnover of B and T cells modifies the clonal composition of the repertoire over time(*40-42*).

High-throughput B and T cell receptor sequencing (Immunosequencing)

High-throughput sequencing enables the resolution of millions of BCR or TCR sequences in parallel. In order to resolve individual B and T cell clonotypes, the hypervariable third complementarity determining region (CDR3) of the receptor is sequenced. This section

of the BCR or TCR is sufficiently diverse to serve as a molecular fingerprint for a B or T cell clone(*43*). The CDR3 is formed at the V(D)J junction; therefore, PCR primers specific to all V and J gene segments surrounding the CDR3 are used to amplify the genomic DNA(*44-46*) or reverse transcribed mRNA(*47-49*) of rearranged B and T cell receptor genes. The nucleotide sequences of this library of amplified BCRs or TCRs is then determined using next generation sequencing methods. For all my work, genomic DNA was amplified because it provides a more quantitative estimate of the number clonotypes in a sample. Developments in immunosequencing include the use of a synthetic immune repertoire to control PCR amplification bias(*50*) and the use of unique molecular identifiers (UMIs) to quantitatively estimate the number of B or T cells sequenced(*48*). The major disadvantage of immunosequencing is that it is restricted to a single chain of the BCR or TCR heterodimer. Specifically, in a single experiment either the *Tcra, Tcrb, Tcrg, Tcrd, IgH* or *IgL* can be sequenced. More recently paired BCR and TCR sequencing have been developed albeit with lower throughput compared to single chain sequencing(*51-54*).

Repertoire diversity

The diversity of clones in the both the naïve and memory B and T cell repertoires is hypothesized to be an indicator of immunological fitness(*55-57*). Specifically, a more diverse repertoire is thought to positively correlate with the probability of mounting an adaptive immune response against a pathogen. As an extreme example, individuals with compromised adaptive immune repertoires due to genetic defects in immune receptor rearrangement are at risk for a wide variety of infections(*58-60*). However, obtaining

accurate estimates of B or T cell repertoire diversity is challenging due to the large number of cells that must be sampled(*61, 62*) and the technical difficulties in obtaining unbiased estimates of clonal frequencies(*63*). In chapter 3, we develop a novel multi-replicate sequencing method to approximate digital clone counts and estimate the diversity of conventional naïve and memory B cell repertoires in the peripheral blood of healthy adults.

T cell responses to acute viral infection

In the setting of an acute viral infection, the adaptive immune response is initiated in the lymph node (LN) draining the infected tissue. Activated dendritic cells (DCs) carrying viral antigens migrate from the infected tissue through the lymphatic circulation and enter the LN parenchyma using CCR7/CCL21 based chemotaxis(*64*). In the LN, these cells either transfer viral particles to LN resident DCs or directly present virus-derived peptides to T cells. The inflammatory environment of a virally infected tissue triggers migratory DCs to secrete interleukin 12 (IL-12) and interleukin 18 (IL-18), which instruct activated effector T cells to produce antiviral cytokines like IFN-γ(*65, 66*). Activated antigen-specific $CD4^+$ and $CD8^+$ T cell clones differentiate to yield a population of effector T cells that exit the lymph node, enter the peripheral circulation and home to virally infected tissue(*67, 68*). Both the diversity and frequency of viral antigen-specific effector T cell clones are associated with viral control(*56, 69-71*). In chapter 4, we develop a novel computational approach to identify virus reactive effector T cell clones and examine the diversity of the $CD8^+$ T cell repertoire in response to an acute viral infection.

T cell responses to chronic viral infection

In contract to acute viral infections, where the immune response subsides after the virus is cleared from the body, chronic viral infections result in persistent and lifelong adaptive immune responses. Viruses that elicit chronic immune responses have cycles of dormancy and reactivation and include cytomegalovirus (CMV), hepatitis C virus (HCV), herpes simplex virus (HSV), varicella-zoster virus (VZV) and human immunodeficiency virus (HIV). Because the immune system is unable to eradicate the virus, T cells continuously respond to viral antigens in order to prevent reactivation. Over time, these T cells become exhausted and dysfunctional potentially leading to a decrease in overall immune function(*72*). In chronic viral infections, anti-viral $CD4^+$ T cells exhibit decreased IFN-γ and tumor necrosis factor (TNF) production and fail to proliferate after continued exposure to cognate antigen(*73*). Similarly, $CD8^+$ T cells gradually lose the ability to produce TNF, IFN-γ, interleukin 2 (IL-2) and perforin leading to a reduction in anti-viral functions(*74, 75*). In addition, the sustained stimulation of $CD8^+$ T cells with viral antigens leads to massive clonal expansions, which are thought to occur with a proportionate loss of T cell clones from the repertoire(*41, 76*). This contraction of the T cell repertoire is hypothesized to increase susceptibility to new infections in elderly individuals harboring chronic viruses. In chapter 5, we examine the impact of long-term CMV seropositivity on $CD8^+$ T cell repertoire composition and diversity in the elderly.

Conclusion

The adaptive immune system is essential to protect the body from infection and disease. A highly diverse collection of BCRs and TCRs ensures that the adaptive immune system

is capable of recognizing an extraordinary breadth of antigens. However, changes in the composition and diversity of the B or T cell repertoire contribute to the control as well as the development and severity of disease. The development of high-throughput BCR and TCR sequencing have allowed us to explore this link. The work presented here unites experimental and computational techniques to examine the diversity of the adaptive immune repertoire in health and disease.

2.3 Organization of Thesis

In this thesis, I describe the application of high-throughput B and T cell receptor sequencing to understand the relationship between immune repertoire diversity and overall immune health.

The third chapter describes a database of millions of IgH sequences from healthy adults. In this study, we developed a novel high-throughput method to approximate digital cell counting based on multi-replicate sequencing. The ability to accurately count clones enabled us to apply a maximum likelihood method to estimate the clonal diversity of naïve and memory B cell repertoires. In addition, we developed a set of tools to characterize general properties of the B cell repertoire including V gene usage and SHM motifs.

The fourth chapter describes the dynamics of the $CD8^+$ T cell repertoire during acute infection using yellow fever virus (YFV) vaccination. YFV vaccination is an excellent model of acute viral infection because the live-attenuated virus retains the ability to replicate in host cells, inducing a potent adaptive immune response(*77-79*). We preformed high-throughput TCR sequencing on $CD8^+$ T cells at distinct phases of the immune response against the vaccine and developed a computational method to identify

vaccine reactive clones. This approach enabled us to determine fate of all clones responding to the vaccine and provided an estimate for the diversity of a $CD8^+$ T cell response to an acute viral infection.

The fifth chapter examines the association between CMV infection and contraction of the $CD8^+$ T cell repertoire in elderly adults. Lifelong CMV infection is hypothesized to cause a reduction in naïve T cell repertoire diversity that exacerbates age-related declines in overall immune function. We preformed high-throughput TCR sequencing on naïve and memory $CD8^+$ T cell subsets in CMV seropositive and seronegative elderly adults. We compare differences in the overlap, distribution and diversity of T cell clones in each subset to characterize the effects of CMV on the T cell repertoire.

The final chapter of this thesis discusses key conclusions from this work and future directions.

3 A Public Database of Memory and Naive B Cell Receptor Sequences

3.1 Abstract

The vast diversity of B cell receptors (BCR) and secreted antibodies enables the recognition of, and response to, a wide range of epitopes, but this diversity has also limited our understanding of humoral immunity. We present a public database of more than 37 million unique BCR sequences from three healthy adult donors that is many fold deeper than any existing resource, together with a set of online tools designed to facilitate the visualization and analysis of the annotated data. We estimate the clonal diversity of the naive and memory B cell repertoires of healthy individuals and provide a set of examples that illustrate the utility of the database, including several views of the basic properties of immunoglobulin heavy chain sequences, such as rearrangement length, subunit usage, and somatic hypermutation positions and dynamics.

3.2 Introduction

The diverse B cell repertoire of a healthy individual allows the recognition of a wide range of antigenic epitopes, resulting in a robust adaptive humoral immune response against pathogens. The vast majority of B lymphocytes express a single unique B cell antigen receptor (BCR), a heterodimeric protein complex composed of a heavy and a light immunoglobulin chain, each of which contains a highly diverse antigen-binding domain. The human immunoglobulin heavy chain (IgH) locus comprises approximately one megabase of chromosome 14, and contains at least 51 functional variable (V) region genes, 25 diversity (D) genes and 6 joining (J) genes that undergo a series of

recombination events to assemble a functional heavy chain[1–3]. This recombination process creates a vast array of antigen-binding receptors through the random assortment of different V, D, and J segments (combinatorial diversity), and the insertion of non-templated (N) and palindromic nucleotides (P) at the junctions between V/D and D/J segments (junctional diversity). Productive in-frame VDJ rearrangements result in a functional heavy chain and lead to a permanent alteration of the genomic DNA sequence of a B cell, defining it as a clone. Similarly, the human immunoglobulin light chain κ and λ loci occupy approximately one megabase on chromosomes 2 and 22, respectively, and contain 30–40 V and 4–5 J segments that can recombine to generate a light chain that is assembled with the heavy chain to form a functional receptor, jointly determining the specificity of recognition[3].

This initial BCR repertoire created in naive B cells through combinatorial and junctional diversity increases upon antigen encounter through the process of somatic hypermutation (SHM), which is mediated by activation-induced cytidine deaminase (AID)[4]. As a result, single base substitutions and occasional insertions or deletions occur throughout the rearranged BCR genes, generating a BCR with increased affinity for its antigen [5, 6]. Our understanding of SHM is limited by the relatively small number of BCR sequences from antigen-experienced B cells that have been available until recently.

The clonal diversity of the human BCR repertoire has been difficult to estimate. Early studies relied on extrapolation from the relatively small number of sequences obtained through low-throughput methods such as immunoscope or traditional Sanger-based sequencing (reviewed in [7, 8]). In recent years, high-throughput sequencing (HTS)

methods have considerably increased the number of unique BCR sequences available to the scientific community. However, most of the sequences generated to date are not readily available in a centralized and curated database — the most widely used resource of immune loci (International ImMunoGeneTics, or IMGT) currently contains approximately 50,000 rearranged human IgH sequences[9]. On the other hand, several other large datasets are publicly available: for example, the National Center for Bio-technology Information (NCBI) Sequence Read Database (SRA, http://www.ncbi.nlm.nih.gov/sra)) includes 454 pyrosequencing data from HIV-1 neutralizing antibodies from the Vaccine Research Center (SRP02639) and antibodies generated in response to influenza vaccination from dbGaP (SRP029381), as well as Illumina sequencing data from healthy donor repertoires from BioProject (SRP037774). In addition to this, a number of publications in the last few years have made considerable numbers of BCR sequences available to the scientific community[10–24]. As a consequence of this recent surge in the number of B cell sequences available, centralized data- base and complex data processing and visualization tools are needed to analyze, visualize and interpret these large datasets of immune sequences.

Immunosequencing of the TCR and BCR repertoires has greatly improved our understanding of B and T cell biology[25], leading to the refinement and modification of B and T cell development models[26–29]. In addition, these data have resulted in multiple clinical advances. For example, immunosequencing has resulted in clinical tests for diagnosis and monitoring of minimal residual disease for lymphoid malignancies[23, 30], has guided the discovery of neutralizing antibodies against HIV[31], has been used to dissect the role of T cells in autoimmunity[32, 33] vaccination[34] and transplant[35, 36],

and to better understand the role of infiltrating T lymphocytes in ovarian cancer[37], melanoma[38] and glioblastoma[39].

Here, we present a public resource of more than 37 million unique immunoglobulin heavy chain (IgH) sequences resulting from the digital amplification and sequencing of the most variable region of the IgH gene from 10 million naive and 10 million memory B cells each from three healthy adult donors, using the immunoSEQ platform[18, 27, 40]. In addition, we have created a suite of software tools that facilitates the visualization and analysis of these data. Using many barcoded replicates for each sample, our method approximates single-molecule sequencing of BCRs at high-throughput, thus ensuring a faithful quantitative representation of nearly all clones present in the biological sample. Besides describing the study design, the specifics of the sequencing technology employed, and the resulting data set, we illustrate the use of the weB based tools developed to enable visualization and analysis of these data.

Finally, to further demonstrate the utility of this resource, we explore a few of the many potential biological questions that can be addressed through our data set: (1) we explored and compared the clonal diversity of naive and memory BCR repertoires at an hitherto unprecedented level of sequencing depth; (2) we confirmed V gene family usage patterns in healthy subjects using a bias-free approach; (3) we examined variations in the length of the third Complementarity Determining Region (CDR3) in naive and memory B cell populations; (4) we analyzed SHM within the steady-state BCR repertoire; and (5) we deconvoluted patterns of SHM substitutions in V genes for naive and memory cells.

3.3 Materials and Methods

Sample source and B cell isolation procedure

Whole blood samples were collected from three 25-40 year-old Caucasian males participating in a study of healthy human volunteers under approval of the Fred Hutchinson Cancer Research Center Institutional Review Board. The donors did not report any infections or vaccinations in the 6 months previous to sample collection. All donors provided written informed consent. All samples were processed less than 2 hours after venipuncture. Peripheral blood mononuclear cells were separated from 400 mL of whole blood by Ficoll (GE Healthcare) gradient density centrifugation at 400g and 22°C. Next, total B cells were enriched from PBMCs using CD19 MicroBeads and the autoMACS Pro Separator (Miltenyi Biotec). B cells were then stained with anti-CD19APC, anti-CD3FITC, anti-CD27PE, anti-IgM-APC750, and anti- IgD-PECy7 (all from BD BioSciences) and sorted using the BD FACS Aria II with FACSDiva v6.1.3 software (BD BioSciences). Naive ($CD19^+$, $CD27^-$, $CD3^-$, IgM^+, IgD^+) and memory ($CD19^+$, $CD27^+$, $CD3^-$) B cells were sorted to a purity of 97% or greater. Sort purity was assessed by passing a small sample of each sorted population back through the flow cytometer. We note that this memory B cell sort contains all $CD27^+$ B cells, including both class-switched and IgM memory B cells. Representative flow cytometry plots of CD27 versus IgD expression on gated $CD19^+$ B cells and CD27 versus IgM expression on gated $CD19^+CD27^+$ B cells are shown in S1A and S1B Fig, respectively.

Sorted B cell populations were pelleted at 300g at 4°C, and finally flash frozen in liquid nitrogen before being stored at -80°C. Genomic DNA was purified from sorted B cell populations using the QIAmp DNA Blood Mini Kit (Qiagen). Genomic DNA was

normalized and the equivalent of 50,000 cells was dispensed each of 188 wells of 96-well plates.

PCR amplification and reduction of PCR bias

To amplify the CDR3 region of IgH, we used a 2-PCR reaction approach as previously described [23]. Briefly, the first step consists of a multiplex PCR that uses gene specific V-forward and J- reverse primers that bind to 47 V and 6 J functional genes as well as many of the pseudogenes for both V and J. The primers are designed for perfect complementarity to the germline V and J gene targets. In addition, the final five nucleotides of each primer were selected so as to bind to sequences that are much less likely to be affected by SHM[41]. The second PCR adds Illumina adaptor sequences and well-specific barcodes, for a total of 31 cycles of amplification.

Despite efforts to achieve consistent melting temperatures (T_m) between all the V and all the J primers, there is a wide variation in amplification efficiency. To remove this bias, we created a synthetic set of IgH receptors with universal flanking sequences that allow for direct sequencing on the Illumina platform[40]. The synthetic genes include all V-J combinations labeled with barcodes that allow for the ready identification of each template. This synthetic immune system is sequenced directly to precisely determine the abundance of each template. Then, multiplex PCR amplification with the V and J gene primers is performed on the synthetic pool and the resulting DNA is also sequenced. Comparing the known starting abundances with the resulting amplified sequences, we are able to assess the relative amplification efficiency of each V and J primer. We then modify the concentration of the primers that over and under amplify. The process is

iterated several times until the majority of the bias is removed. We have shown that the results of this process are robust to variations in the length, GC-content, and overall abundance of the template.

Resolution of nucleotide sequences

To measure the amount of nucleotide assignment error in our analysis, we randomly selected molecules from the PCR amplified library of IgH receptor sequences and sequenced them at a depth of at least 10 times the starting template quantity. In other words, since each well contained approximately 50,000 B cells, we aimed to sequence at least 500,000 molecules from each PCR library. This ensured that, even with some amplification variation and random sampling error, multiple copies of each template would be sequenced. Due to the very low error rate in Illumina sequencing (~.1%), the number of errors in a 130-basepair sequence is roughly distributed as $k_{error} \sim Bin(n = 130, p = .001)$, from which we compute $Pr(k_{error} = 0) \cong .88$, and $Pr(k_{error} = 1 \mid k_{error} > 0) \cong .94$. Thus, ~90% of all our templates result in no PCR or sequencing errors. Of the remaining ~10%, the large majority contain a single error. Given that these errors are not systematic, any particular error is almost always unique. Thus, we are able to readily correct these errors by identifying reads present once in the data set that differ by a single nucleotide from a sequence present multiple times and collapsing them into the predominant clone. Additionally, since memory samples were found to have many more clones present in multiple wells, error correction was performed on data aggregated from all wells of a given sample. This ensures consistent consensus sequence assignment across wells. In terms

of the diversity inference described below, this method of collapsing errors across wells is intrinsically conservative.

Germline annotation of nucleotide sequences and SHM detection

The CDR3 region was identified according to the standard previously determined by the IMGT collaboration[9]. Identification of the V, D, and J gene segments was performed using a scored alignment across a definition list of all known V, D, and J gene and allele members from IMGT. The most likely assignments (allowing for ties for similar gene sequences) for each gene segment were then added to the sequence reads as their germline annotation. Somatic hyper- mutation was calculated over just the V gene segment, based on sequence variations from the assigned germline gene/allele match.

Estimation of repertoire diversity from replicate occupancy data

To estimate clonal diversity, we derived an extension of an established sampling model in ecology and corpus linguistics: the Poisson abundance model[42–44]. This allows the construction of a likelihood function for replicate occupancy data parameterized by the richness and abundance distribution of the repertoire. Briefly, we synthesized the combinatorial probability of the replicate occupancy of a clone conditioned on sample abundance, with the Poisson abundance model of sample abundance conditioned on repertoire parameters. Analytically marginalizing over sample abundance as a latent variable, we formed the desired likelihood function and deployed tandem numerical and analytical optimizations facilitated by an asymptotic approximation for large richness. The full mathematical derivation and computational validation of this model can be found in the Supporting Information (S1 Method).

3.4 Results and Discussion

Immunosequencing of naive and memory B cells

In healthy adults, $CD19^+$ B cells comprise 7–11% of lymphocytes circulating in peripheral blood[45]. This population is dominated by naive B cells, which correspond roughly to 65% of all peripheral B cells, while memory B cells account for about 30% of all circulating B cells[45]. To faithfully capture the breadth of the B cell repertoire, we isolated naive (N, $CD19^+$ $CD27^-$ IgD^+ IgM^+) and memory (M, $CD19^+$ $CD27^+$) B cells from 400 mL of peripheral blood obtained from each of 3 healthy adult donors (D1, D2 and D3)[46]. Additionally, in order to estimate the reproducibility of the approach, we included two biological replicates of the naive B cell sample from Donor 1 (i.e. D1-Na and D1-Nb).

These samples yielded 2–4 x 10^7 naive B cells and 1.5–2 x 10^7 memory B cells at greater than 97% purity from each donor. Considering that the approximately 5 L of peripheral blood of healthy adults is estimated to contain on average 6.5 x 10^8 naive B cells and 3.0 x 10^8 memory B cells[45], we calculate that by using a 400 mL sample, we captured 3.1–6.1% of the naive and 5–6.7% of the memory B cells circulating in peripheral blood, respectively.

Next, we sequenced a segment of the immunoglobulin heavy chain (IgH) gene from the naive and memory B cell populations purified from each donor that includes CDR3[18]. Since the CDR3 rearranges somatically during B cell development, the resulting sequences can be used to define unique B cell clones, in the sense of

descendants from a common naïve B cell; however, somatic hypermutation means that even among mature B cells that share a CDR3 by common descent, there can be additional sequence differences in e.g. the CDR1 and CDR2 regions.

In brief, for each of the samples, we extracted genomic DNA and we dispensed an amount corresponding to ~10^7 naive or memory B cells into 188 wells of two 96-well plates (the remaining wells were used for controls). This resulted in the allocation of the equivalent of approximately 50,000 cells per well (Fig 1A). We then performed a two-step PCR, including a multiplex step that uses V and J-specific primers to amplify a region of the IgH gene, followed by a second amplification that adds unique well-specific barcodes and Illumina adaptors. Next, we used a HiSeq instrument to sequence a 130 nt-long segment of the IgH gene that includes the CDR3[18]. This approach enabled us to sample the naive and memory repertoires of B cells of three healthy individuals to a depth much greater than other studies.

The value of the resulting dataset depends both on the accuracy of the IgH nucleotide sequences and the quantitation of the abundance of each B cell clone. Importantly, there are two major obstacles that hinder the quantitative immunosequencing of IgH genes. The first challenge, which is shared by other immune genes such as those encoding for T cell receptors, arises from the process of gene rearrangement and the resulting intrinsic diversity of both types of immune receptors. The second challenge, unique to B cells, results from the additional level of divergence from the genomic sequence generated by SHM in antigen-experienced cells. Our approach to address these challenges is described in the Material and Methods section, and our analytical

approach is described in detail in the S1 Method included in the Supporting Information section.

In brief, we used a digital counting method that yields counts of clones based on their presence or absence in each of the 188 wells, as diagrammed in Fig 1B. Quantitative accuracy is achieved by inclusively sequencing the receptors in each uniquely-barcoded well. We aimed for a minimum of 10-fold coverage of each BCR molecule in each well and achieved an effective coverage that ranged from 8 to 12 average reads per template in the different samples. We also analyzed the distribution of the number of unique productive BCRs over the 188 wells for each sample, as shown in S2 Fig. Most of the samples had an average of 40,000 unique productive rearrangements per well, with the exception of the naive sample from Subject 2, which had a lower number of unique productive rearrangements per well.

Our method is binary, since we only consider presence or absence of each sequence in each well, and robust against a wide range of amplification efficiencies. The sequences in each well are identifiable by the presence of the unique barcode assigned to that well, and thus we report an “occupancy” value for each BCR sequence, which corresponds to the number of wells it was observed in. Clones with abundance in the repertoire of less than 1:1,000,000 B cells (i.e. the vast majority of all B cell clones) will rarely be present more than once in any well. Therefore, for most B cells, their sample abundance will be equal to the number of wells they are observed in. We determined that the vast majority of clones have an occupancy value equal to 1 (Fig 2A). Since multiple cells of the same clone are unlikely to appear in any given well, this strongly implies that

a single cell out of the initial 10^7 expressed that particular BCR sequence. As occupancy increases, this metric becomes a decreasingly precise (and increasingly negatively biased) estimator for sample abundance, since the incidence of multiple occurrences of a given clone in a single well becomes more probable.

Diversity of the naive and memory B cell receptor repertoires

We first compared the overlap between the naive and memory B cell repertoires of the three donors studied (Table 1). For this analysis, we only considered exact sequence matches.

For each sample obtained from each of the donors (D1-Na, D1-Nb and D1-M; D2-N and D2-M; and D3-N and D3-M), the table indicates the pairwise overlap between repertoires, computed as the fraction of the unique sequences for each sample in the rows labeled to the left that are also found (with no mismatches allowed) in the each of the samples listed in the columns. The color gradient of the cells indicates the degree of overlap, with higher overlaps indicating a darker shade of red.

Due to the intrinsically large size and diversity of the B cell repertoire, the overall overlap between samples is small. However, as expected, it is higher between the two independent replicates of the naive repertoire of Donor 1 than between those of different donors. Also, the naive and memory B cell populations of each donor are more similar to each other than to those of different donors.

Next, for each sequence present in the data we computed the maximum well occupancy among all samples (a measure of clonal abundance), and also the number of

subjects the sequence was observed in. S3 Fig shows the distribution of maximum occupancy among sequences found in only 1 subject, in any two subjects, and in all three subjects. We observe that shared sequences (those present in two or three subjects) tend to have higher maximum occupancy. This could be the result of shared memory cells resulting from common pathogen exposures among subjects, or alternatively, the consequence of recurrent generation of high- probability V(D)J recombinations that are identical by state but not by descent in different individuals.

We also estimated the clonal diversity of the repertoires – i.e. the number of distinct somatically rearranged receptors present in each repertoire and their relative abundances–which defines the search space available for immune recognition and is therefore essential for the quantitative characterization of the BCR repertoire. For each sample, we inferred two diversity indices: richness, defined as the number of distinct clones, and clonality, a measure of abundance uniformity that ranges from 0 (maximally uniform) to 1 (most disparate, or clonally dominated; see the Materials and Methods and S1 Method for a detailed description of these indices). Fig 2B shows the maximum likelihood estimates of clonal diversity. Using either diversity metric, the samples cluster distinctly by cell type, and these results were consistent across individuals. As expected, our results indicate that memory clones have more disparate repertoire abundances (higher clonality) than naive clones, and that naive clones are extremely diverse.

Our replicate PCR well methodology accurately assesses the abundance of nearly all B cell clones in each sample. A small number of memory clones are present at high frequency, and thus are found in all or nearly all of the replicate PCR wells. This is

expected to cause negative bias in the clonality inferences for the memory populations. Despite this conservative bias, the memory and naive populations cluster distinctly.

The inferred richness of the naive B cell repertoire is of a similar magnitude to the expected abundance of naive B cells in the peripheral blood (~$1x10^9$)[45], suggesting that the typical naive clone does not undergo proliferation prior to antigen encounter. In contrast, the richness of the memory B cell population is consistent with each clone undergoing several divisions on average. The relatively higher clonality observed for memory cells as compared to naive cells indicates that a small percentage of these clones experience significant proliferation. Our conclusion that the typical naive B cell clone undergoes no proliferation prior to antigen encounter raises questions regarding previous calculations that suggested that naive B cells in the peripheral blood of adults undergo approximately 1.9 cycles of homeostatic proliferation on average [47]. However, it is important to point out that the study by Van Zelm et al. uses an indirect method of estimating the replication history based on deletion circles, and that, unlike our approach, it does not have the ability to resolve distinct clones. On the other hand, we do not measure replication history and instead calculate it from the diversity metric and estimates of the number of B cells in the periphery reported in the literature. Thus, both sets of results are not directly comparable and do not necessarily contradict each other.

In summary, our data confirm that the naive repertoire of a healthy adult is extremely rich, and thus suggests that the typical naive B cell clone undergoes no proliferation prior to antigen encounter, while we observe that memory B cell clones undergo several cycles of division on average. Future studies will focus on mining this

extremely deeply-sequenced data to further understand ongoing maturation of clones within the memory compartment at steady state. The assay also has the potential to determine whether the different subsets of cells contained in the memory compartment (i.e. switched memory cells, unswitched memory cells, as well as any plasmablasts or plasma cells present due to ongoing immune responses) possess different distributions of mutation rates.

Examples of possible explorations of this dataset

To demonstrate that our data are accurate and of high quality, we made use of these tools to answer several fundamental questions about the B cell repertoire in healthy individuals. In addition to the clonal diversity inferences described above, we provide a set of four examples that illustrate the utility of the data set and the related analysis tools. For each of these examples, we created a dashboard in the immunoSEQ Analyzer workspace (http://adaptivebiotech.com/link/publicBCellResource) so that the analysis of each example and the accompanying visualizations that follow can be reproduced by the user.

Example 1: Characterization of IGHV family and gene usage. The IGH V locus contains over 50 functional genes (depending on the individual's haplotype) that are classified into 7 families based on nucleotide sequence homology[48]. Each gene segment has a certain likelihood of undergoing rearrangement and being incorporated into a mature immunoglobulin molecule, and in addition the process of negative selection of immature B cells further restricts V gene segment use, resulting in an unequal representation of V gene families in the naive B cell repertoire. Similarly, the positive

selection of naive B cells to populate the memory compartment results in variations in V gene segment representation[13, 49].

Traditionally, standard measurements of TCR usage in T cells have utilized PCR-based V beta spectratyping (reviewed in reference [7]), but no equivalent approach exists for the analysis of V gene usage in B cells. However, recent immunosequencing approaches have begun to shed light on B cell gene usage[18, 19]. To assess the broad similarities and differences in gene usage between the naive and memory B cell repertoires, we compared the IGHV family and gene usage in naive and memory B cells in three healthy donors (Fig 3). In agreement with previous reports, we found that the IGHV3 gene family is utilized most commonly in both repertoires[49, 50]. Moreover, we observed that, in these subjects, IGHV3-48 is the most commonly used V gene in the naive repertoire followed by IGHV3-30 or IGHV3-64, two genes that are indistinguishable over the region covered by the sequence reads. In the memory repertoire, IGHV3-23 is used most commonly, followed by IGHV3-48. We found that the second most commonly expressed gene family in the naive repertoire of these subjects corresponds to IGHV1, followed by IGHV4. In contrast, the memory repertoire has equivalent representation of the IGHV1 and IGHV4 gene families. At the gene specific level, we observed a decrease in the relative frequency of IGHV1-69 and IGHV1-18 within the IGHV1 family in memory com- pared to naive B cells, consistent with previous studies[13]. Taken together, these data, which were obtained from a single experiment, reproduce observations from several previously published studies [13, 18, 49–51], validating the utility of this dataset.

Example 2. Measurement of CDR3 length distribution. The immunoglobulin CDR3 is the most important determinant of antibody-antigen recognition[52, 53]. Its length varies

mostly due to recombination and can also change slightly from SHM. Therefore, we compared the CDR3 length distribution of the naive and memory repertoires to understand both the limits and flexibility of the antigen-binding capacity of B cells. We found the average CDR3 length in the naive B cell repertoire to be 48 nucleotides, while the memory B cells had, on average, a CDR3 length of 45 nucleotides (Fig 4). Unproductive CDR3 sequences have an even longer average size (~60 nt) than that seen for productive sequences in naive or memory cells. These two facts suggest that, while the B cell recombination process generates long and highly diverse CDR3 regions, functional clones that become part of the memory repertoire are biased towards shorter CDR3 sequences. In addition, we observe that there is a greater variability in CDR3 length in naive cells compared to memory cells, suggesting that the naive repertoire has the potential to bind a wider range of antigens than are actually encountered by the donors in this study. These data agree with previous findings [13, 18, 54], further confirming the validity of our dataset.

Example 3: Assessment of purity of flow-cytometry sorted cell populations. Since SHM occurs during antigen-induced maturation, a naive B cell is characterized by the absence of substitutions in its germline V gene[5]. Thus, to examine the purity of our sorted B cell populations, we determined the rate of substitutions in the V genes of the naive and memory B cell repertoires (Fig 5). Approximately 95% of sorted naive B cells displayed no V gene substitutions, and had low clonal abundances, which are typical of naive cells. In contrast, memory B cells harbored an average of 3-4 substitutions per 100 nt in the V genes, and additionally displayed a much broader range of clonal abundances,

as expected of antigen-experienced B cells. Taken together, these analyses suggest that our method accurately and faithfully captures the circulating B cell populations.

Example 4: Analysis of somatic hypermutation in memory B cells. Affinity maturation, including somatic hypermutation and class switching, is critical to the production of functional antibodies[55–60]. We were able to easily define somatic hypermutation sites by identifying variations from germline sequences within the sequenced region of the V gene. While a certain number of single nucleotide variations in the V gene may result from inherited SNPs, a review of the V gene sequences observed in naive cells in the same individual makes it easy to exclude this possibility in most cases.

After identifying likely somatically hypermutated residues in the V gene segments, we created a set of tools to view these data for all genes and samples over the sequenced V gene region. Fig 6 shows an example of the resulting data for gene IGHV1-69. Our analysis and visualization tools allow a clear visualization of SNPs and transition/transversion rates (top panel), as well as overall SHM rates by position (middle panel) gleaned from our very deep sampling of memory B cell sequence data. In addition, several reported hotspot and coldspot AID targeting motifs[61] can be evaluated (bottom panel). The most frequently reported hotspot motif (most generally described as GYW/WRC on the two strands[6]) accounts for many of the observed positions with high SHM levels, while some nucleotides that display SHM, such as nucleotide 267 in several V genes including V01-69 and V03-23, are not part of a known hot- spot motif. It is possible that mutations of this position, which flanks the CDR3, might have increased functional importance for improved antibody binding, despite the absence of known AID-targeting motifs.

3.5 Conclusions

In this study, we provide the research community with an accurate and rich dataset of BCRs, as well as a set of straightforward tools to enable its in-depth study. By combining flow cytometry purification of peripheral B cells with high-throughput immunosequencing of 10 million naive and 10 million memory B cells from each of three healthy adult donors, we generated a BCR sequence library containing more than 37 million unique BCR sequences. Whereas some of the currently existing databases, such as IMGT[9], contain a large number of curated IgH sequences from many individuals, this method allowed us to probe the B cell repertoire of a small number of individuals at an unprecedented depth. In parallel, we developed set of tools tailored to analyze and visualize the resulting data set, which can be accessed from http:// adaptivebiotech.com/pub/robins-bcell-2016 (please follow the 'Advanced Visualizations' link).

As an example of the utility of our dataset, we assessed a fundamental property of the BCR repertoires, i.e. their clonal diversity. To do this, we approximated high throughput digital cell counting using a multi-replicate experimental design, and we inferred the clonal diversity of the memory and naive BCR repertoires of three healthy adults using a novel likelihood model.

To further illustrate the utility of these data and the associated tools, we present several other examples that assess general properties of B cell repertoires that have been previously investigated at a smaller scale, including V gene family usage patterns; the length of CDR3 regions; the numbers of SHM substitutions, and the patterns and types

of SHM in naive and memory B cells. Importantly, our observations match previous reports and thus confirm the robustness of our dataset.

Finally, the many-replicate experimental design employed in this study, in which each of the 188 PCR wells corresponds to a replicate sample, constitutes a sample abundance probe robust to the inherent stochasticity of PCR amplification. Moreover, this approach represents a crucial quantitative advance over previous sequencing studies of antigen receptor repertoire diversity, which have been limited by either poor quantitation or by the lower throughput of single-cell methods[27, 62, 63]. We expect that these data will be used by other experts in the field of immunology to address additional fundamental questions about BCR development and in vivo antigen binding in humans.

3.6 Data Availability

Access to the data set resulting from the experiments described in this study (both at the well level and at the sample level), as well as a link to the tools we developed to enable the analyses presented herein, can be found at http://adaptivebiotech.com/pub/robins-bcell-2016. We have also assigned a unique identifier to this dataset: http://doi.org/10.21417/B71018. The immuno- SEQ Analyzer interface includes several tools that can be used to perform further analyses of the data. The “Advanced Visualization” link found in the landing page for this dataset enables access to Fig 2 to Fig 6 in this study, and each of them is followed by a set of interactive dashboards that allow viewing different aspects the data, such as Occupancy (data underlying Fig 2), VDJ tools (data underlying Fig 3), CDR3 tools (data underlying Fig 4), Substitutions tools (data underlying Fig 5), and SHM tools (data underlying Fig 6). Most

dashboards include a sample selection option: data are coded by sample type (naive vs. memory) and for each of the three donors stud- ied (including the two repeats for the naive sample from donor 1). Several of the dashboards include filters that allow viewing subsets of the data (e.g. sequences for productive vs. non-pro- ductive rearrangements, out-of-frame sequences or sequences with STOP codons). The code for the tools developed for the analysis can be downloaded from the Public B cell dataset code link.

Finally, the full dataset can also be downloaded from the Public B cell dataset link, and Dryad Digital Repository at http://datadryad.org/resource/doi:10.5061/dryad.35ks2.

3.7 Figures and Tables

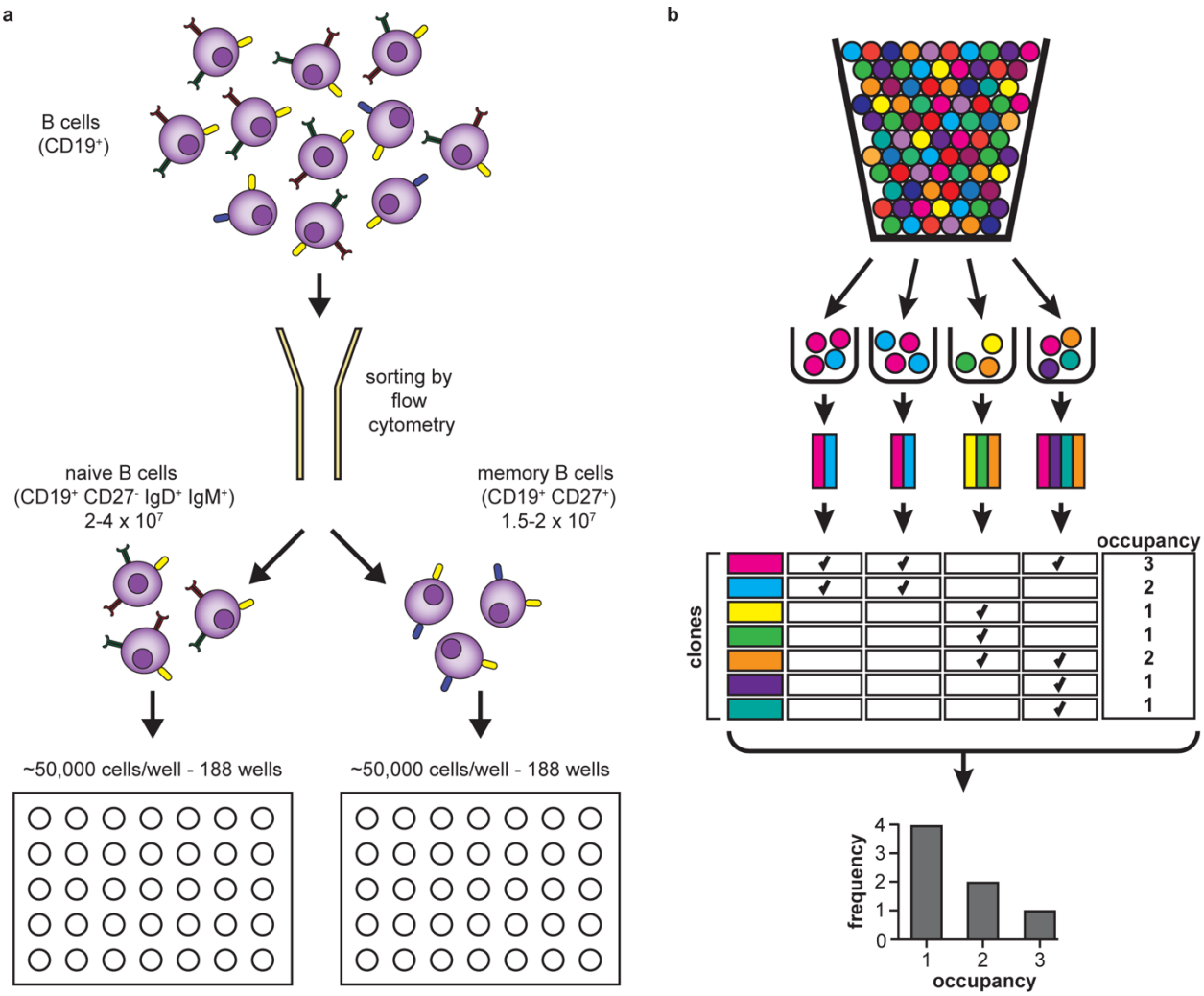


Fig 1. Experimental and informatic design. (a) Peripheral blood samples from three healthy donors were sorted using flow cytometry to isolate naive ($CD19^+$ $CD27^-$ IgD^+ IgM^+) and memory ($CD19^+$ $CD27^+$) B cells. For each sample, approximately 10^7 cells were distributed into two 96-well plates (i.e., into 188 wells, resulting in ~50,000 cells per well), and processed by immunosequencing. (b) Schematic of the 'urn sampling' quantitation method. Cells are represented by colored balls, with each color indicating a different clone identity. Each ball (cell) is randomly allocated to a sample bin (well). Occupancy is calculated after censoring count information, and thus is expressed as presence or absence. The majority of clones are present in

just one out of 188 wells, indicating that they were almost certainly represented by a single cell in the original sample.

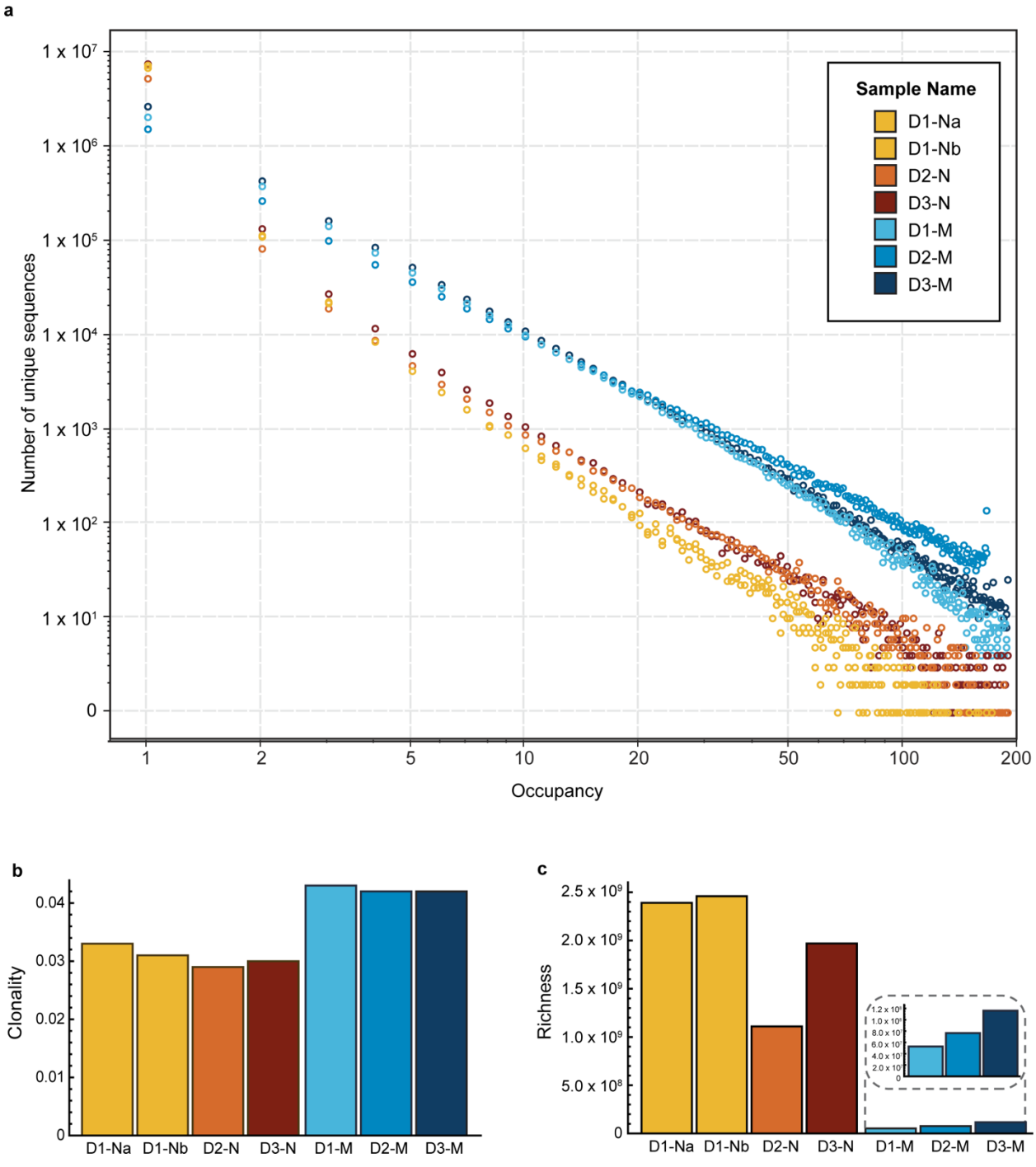


Fig 2. Inference of diversity in the naive and memory B cell repertoires. (a) The graph shows the distribution of unique sequences, as the number of unique sequences (y-axis) versus their occupancy (x-axis) for the naive (orange) and memory (blue) samples for the three donors (D1, D2 and D3, including two technical replicates for the naive sample from Donor 1). The vast majority of the sequences have occupancy of 1. (b) Clonality index for all samples. (c) Richness index for all samples. While the clonality index is higher for memory samples, the richness index is higher for the naive samples.

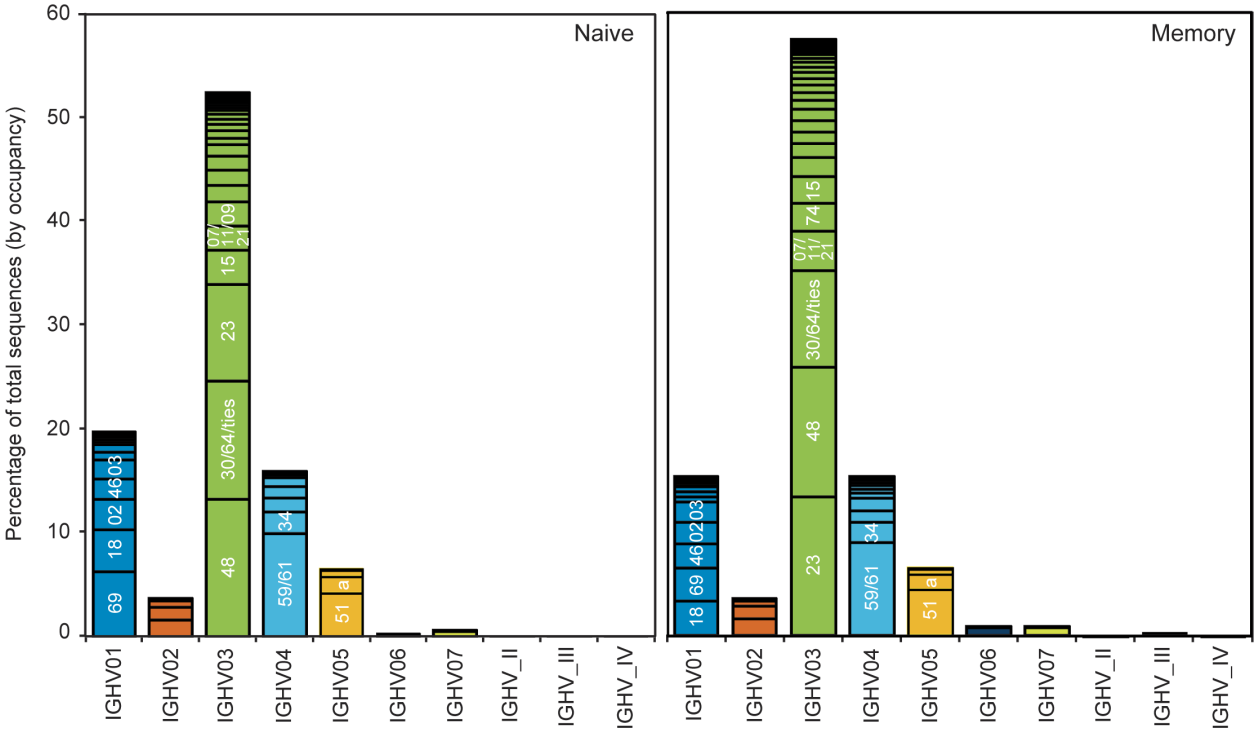


Fig 3. V family and V gene usage patterns. The histograms show the relative percent of total sequences (by occupancy) for each of the IGHV families (as shown under the graphs), for the naive (left panel) and memory (right panel) samples, aggregated for the three donors. Within each family, discrete bands represent each of the individual genes. The most abundant genes within each family are indicated (e.g., 69 in IGHV01 refers to the gene IGHV01- 69). Overall, memory samples contain fewer IGHV01 and more IGHV03 family sequences than naive samples, with some gene-level differences evident as well.

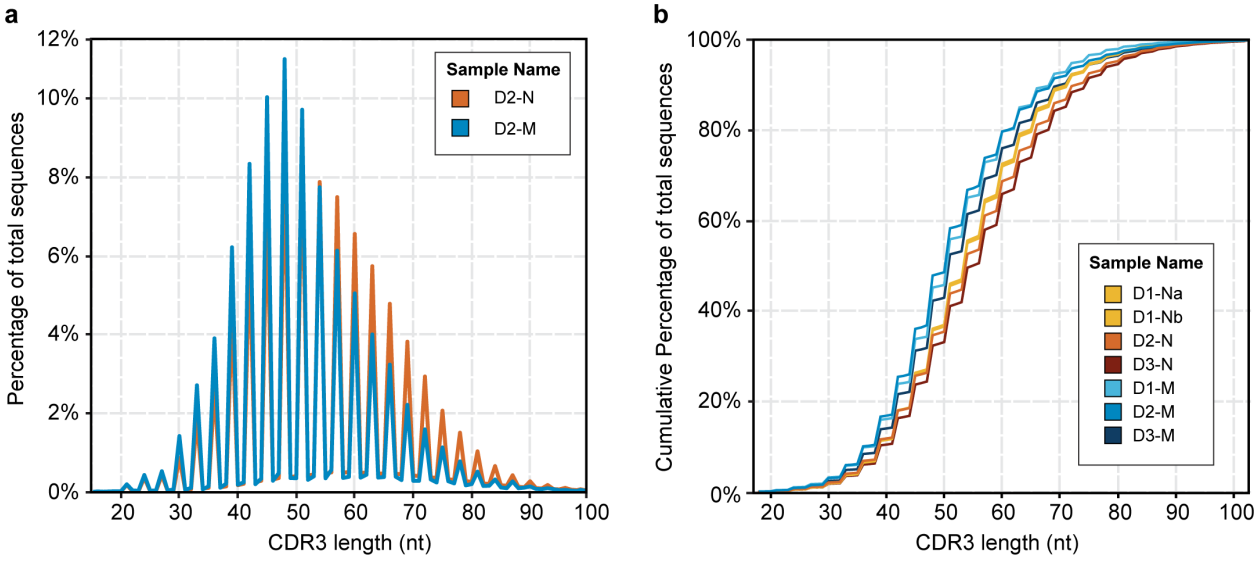


Fig 4. Comparison of CDR3 lengths in naive versus memory B cell samples. (a) The graph shows the normalized percentage of total sequences for the naive (orange) and memory B cells (blue) from donor D2. (b) The graph shows the cumulative percentage of total sequences at a given CDR3 length for all naive and memory samples, as indicated in the inset. The technical replicates for donor D1 overlap closely and are not distinguishable in this figure. The memory repertoire is consistently 3 nucleotides (or 1 amino acid) shorter than the naive repertoire at the same cumulative frequency.

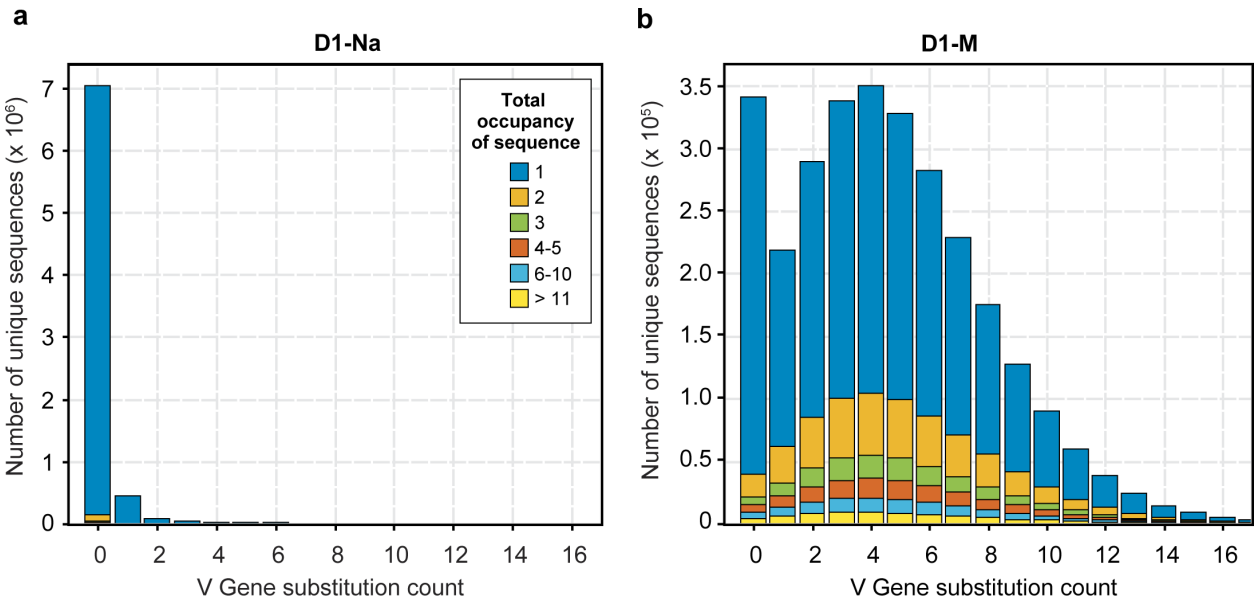


Fig 5. Comparison of Somatic Hyper Mutation in paired naive and memory B cell samples from the same donor. The figure shows data for the naive (a) and memory sample (b) from Donor 1, which is representative of all three donors. The x-axis corresponds to the number of substitutions differing from the germline V gene sequence, and the y-axis indicates the number of unique sequences that display that number of substitutions. The colors indicate different total well occupancies, with blue indicating singletons present in just one well, and the other colors showing progressively higher well occupancy, as indicated in the figure. The majority of the sequences in the naive B cell sample have 0 substitutions and correspond to low abundance clones observed in a single well (blue). In contrast, the memory B cell sample from the same individual shows a much broader distribution of substitutions, as well as many more sequences with occupancy greater than 1.

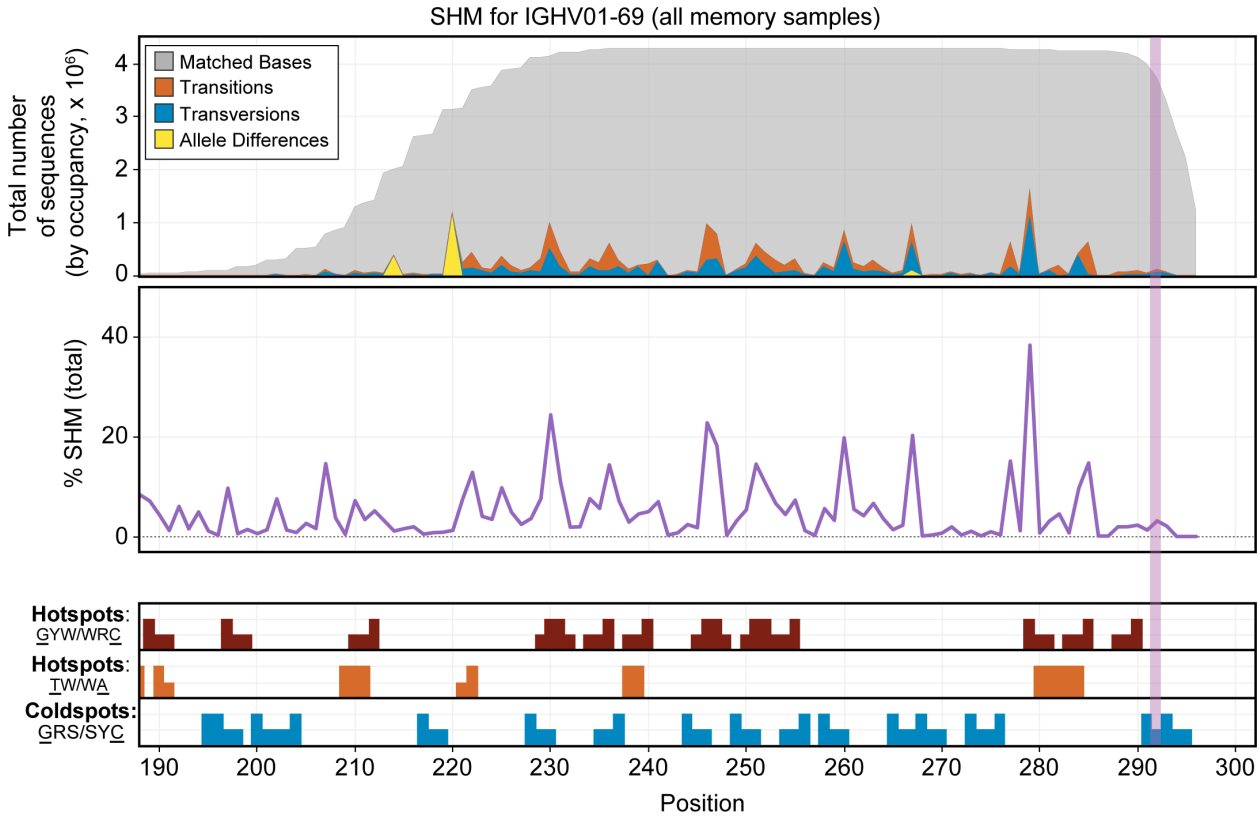


Fig 6. Somatic hypermutation pattern observed over the sequenced region of the IGHV01-69 gene. The figure includes combined data from the memory B cell population for all 3 donors.

The top panel shows the total distribution of sequenced bases by occupancy for the primary allele of IGHV01- 69. Nucleotides that match the germline sequence are displayed in gray. Transitions are shown in orange and transversions in blue. Allelic differences, which are also seen in the naive samples, are indicated in yellow. The vertical dotted line marks the average start of the CDR3 region. The middle panel shows the normalized percentage SHM by base for this gene across the memory B cell samples for all three donors. The bottom panel shows suspected SHM hotspot (red and orange bars) and coldspot (blue bars) motifs present in the sequence of this gene over the region assayed. Positions with higher bars indicate bases targeted within the motif (underlined in the legend to the left). The GYW/WRC pattern (red) explains most of the significant sites of SHM for this gene, but some spots of high mutation are not captured by the displayed motifs. In the data viewer, this view can be generated for any V gene and for any combination of data sets.

Table 1. Overlap among the naive and memory repertoires of the three donors.

		1			2		3	
		Na	Nb	M	N	M	N	M
1	Na		4.18E-03	7.87E-04	4.41E-04	8.34E-05	6.96E-04	1.07E-04
	Nb	4.55E-03		8.21E-04	4.44E-04	8.42E-05	6.92E-04	1.08E-04
	M	1.75E-03	1.68E-03		5.42E-05	9.48E-06	8.90E-05	2.03E-05
2	N	6.55E-04	6.06E-04	3.61E-05		2.38E-03	7.06E-04	1.15E-04
	M	2.42E-04	2.25E-04	1.24E-05	4.67E-03		2.65E-04	5.43E-05
3	N	6.66E-04	6.09E-04	3.83E-05	4.56E-04	8.74E-05		4.07E-03
	M	1.94E-04	1.80E-04	1.65E-05	1.40E-04	3.38E-05	7.70E-03	

3.8 Supplementary Information

Supplementary Methods

Supplementary Method:
Replicate immunosequencing as a robust probe of antigen receptor repertoire diversity

I. INTRODUCTION

Previous approaches to antigen receptor repertoire diversity estimation redeploy methods developed in the ecology and corpus linguistics literature to estimate species diversity and vocabulary size (see review [1]), respectively. Specifically, Poisson abundance models, with both parametric and nonparametric estimators, are used. Although conceptually erroneous, mark-recapture formulae have also been applied [2]. Antigen receptor repertoires more closely achieve the idealizations of these models than the their original applications; populations are very large and well-mixed, and detection probabilities are homogeneous. However, studies suffer from limitations in sequencing data that blunt sophisticated computational approaches.

Robins et al. [3] assessed T cell receptor (TCR) richness from high-throughput immunosequencing data using a nonparametric empirical Bayes method requiring divergent series regularization [4, 5] (a substantially improved regularization technique, applied to estimating the molecular complexity of PCR libraries, is advanced in [6]). However, the sequencing read count assigned to each unique TCR (after error correction) was associated with its clonal abundance in the sample. This introduces noise and bias, since each single template is stochastically amplified by PCR. Although this high-throughput study captured the diversity of a realistic biological sample, inference of repertoire richness was problematic due to limited quantitation of sample abundance for each clone.

Rempala et al. [7, 8] employed a likelihood model and posterior inference for mouse TCR richness using single-cell sequencing to quantitate sample abundance of T cell clones. Although this approach allows for precise quantitation of sample abundance, it is so low throughput (one cell per well on a 96-well plate) that diversity estimation was only possible for transgenic mice engineered to have dramatically limited TCR diversity. Although quantitatively principled, severe experimental limitations restricted the study to less biologically relevant repertoires.

II. EXPERIMENTAL DESIGN

In the present study a high-throughput and quantitatively robust (albeit indirect) probe of B cell clone sample abundance was devised. B cells from three adults were sorted into memory and naive populations (with two naive replicates for subject 1), each with $\sim 10^7$ cells (Fig. 1a, main text). Extracted DNA from each sample was evenly partitioned into 188 PCR replicates for amplification and uniquely barcoded for immunosequencing of the rearranged IgH locus [9], identifying clones by unique CDR3 sequence in their B cell receptor. Instead of relying on sequencing read counts to estimate a clone's sample abundance, we use its occupancy - the number of replicates it is observed in. In the regime of small occupancies, this approximates digital cell counting - a clone observed in only one replicate almost surely has a sample abundance of one cell. For larger sample abundances, co-occupancies become more probable, so occupancy increasingly underestimates abundance. Clones with sample abundance much larger than the number of replicates will saturate, appearing in all replicates.

To address possible template quantity variation across replicates and non-detection effects, we selected the subset of 150 replicates for each sample having minimum variance in the number of unique clones. Removing replicates with outlying allocations of cells or underperforming amplification is necessary to avoid breaking exchange symmetries invoked in our model.

III. MODEL

We advance a combinatorial extension of a well-studied model of sample abundance, enabling application to occupancy data. After introducing a parameterization of this extended model, a maximum likelihood diversity estimation is introduced, validated with simulations, and applied to BCR repertoire occupancy data to infer both richness (the number of clonal species) and an index of relative diversity (evenness of clone abundances).

A. Poisson abundance model of replicate occupancy

As is canon in the ecology and corpus linguistics literature, we begin by modeling sampling from a diverse population as a superposition of homogeneous Poisson processes. A mixing measure, $\mu(\lambda)$, characterizes the distribution of Poisson rates over all categories (B cell clones, as identified by productively rearranged IgH CDR3 segment, in our case). Since a clone's Poisson rate, λ, is given by its repertoire fraction times the number of cells sampled, $\mu(\lambda)$ is tantamount to the repertoire clonal abundance distribution. Homogeneity entails the approximation that the repertoire is effectively an infinite reservoir (or is being sampled with replacement), such that the data is not sensitive to depletion of the population fractions of the sampled clones. An equivalent urn model samples with replacement from a finite urn with an unknown number of ball colors, or without replacement with an urn with an infinite number of balls and

specified fractional abundances for each color. The total number of balls (cells) sampled from the urn (repertoire) is taken to be a Poisson sample from a multinomial population. The marginal distributions of sample cellular abundance, j, of each clone are then independently and identically distributed as

$$p\left(j|\mu(\lambda)\right) = \int_0^\infty d\mu(\lambda)\frac{\lambda^j e^{-\lambda}}{j!}$$

To model replicate occupancy, we assume that each sampled cell is randomly assigned to one of L possible replicates with equal probability (Fig. 1b, main text). Because the sample material was partitioned equally among the L replicates, it is not strictly correct to assume that each cell is assigned to a replicate independently. However, for large samples this approximation is very accurate. If the sample contains N cells, then under this model the number of cells in each replicate is binomially distributed with N trials and success probability $1/L$. The coefficient of variation is $\sqrt{(L-1)/N}$. The number of replicates used in this study was 150, and about 10 million cells were sequenced for all samples, leading to a coefficient of variation of about 0.004.

The distribution of a clone's replicate occupancy, i, conditioned on sample abundance, j, is then determined combinatorially as

$$q\left(i|j\right) = \frac{\binom{L}{i} i! \left\{ {j \atop i} \right\}}{L^j}.$$

This is simply the ratio of the number of ways to partition j cells into i out of L replicates, divided by the total number of ways to allocate j cells among L replicates. $\left\{ {j \atop i} \right\}$ denote Stirling numbers of the second kind, which count the number of ways to partition j distinguishable objects into i indistinguishable nonempty subsets.

Marginalizing over the hidden sample abundance gives the distribution of each clone's occupancy as

$$\begin{aligned} r\left(i|\mu(\lambda)\right) &= \sum_{j=0}^{\infty} q\left(i|j\right) p\left(j|\mu(\lambda)\right) \qquad (1) \\ &= \sum_{j=0}^{\infty} \frac{\binom{L}{i} i! \left\{ {j \atop i} \right\}}{L^j} \int_0^\infty d\mu(\lambda) \lambda^j e^{-\lambda}/j! \\ &= \binom{L}{i} i! \int_0^\infty d\mu(\lambda) e^{-\lambda} \sum_{j=0}^{\infty} \frac{\left\{ {j \atop i} \right\}}{j!} \left(\frac{\lambda}{L}\right)^j \\ &= \binom{L}{i} \int_0^\infty d\mu(\lambda) e^{-\lambda} \left(e^{\frac{\lambda}{L}} - 1\right)^i, \end{aligned}$$

where we have exchanged the order of integration and summation, and identified the sum as a well-studied exponential generating function for the Stirling numbers ([10], p.83). In formal power series notation, $\left\{ {j \atop i} \right\} = j! \left[z^j\right] \left((e^z - 1)^i / i!\right)$. We consider a finite-dimensional subspace of measures, $\mu_\theta(\lambda)$, parameterized by the vector θ, and thus write (1) as

$$r_\theta(i) = \binom{L}{i} \int_0^\infty d\mu_\theta(\lambda) e^{-\lambda} \left(e^{\frac{\lambda}{L}} - 1\right)^i. \qquad (2)$$

For a repertoire with clonal diversity S, the sample occupancy of each clone is drawn from distribution (2). Let $l_1, l_2, \ldots, l_S$ denote the replicate occupancies of S labelled clones. For a very diverse repertoire and a limited sample, many clones will not be sampled, and thus have occupancy zero (the *missing species*). Due to exchangeability of the clone labels, it is sufficient to consider the frequencies of nonzero occupancies, defined by the vector indicator random variable $o = (o_1, o_2, \ldots, o_L)$, with $o_i = |\{c \in \{1, 2, \ldots, S\} : l_c = i\}|$ (the number of clones occupying exactly i replicates).

We may write a multinomial likelihood function as

$$\mathcal{L}\left(\theta, S|o\right) = \frac{S!}{(S-s)!} r_\theta(0)^{S-s} \prod_{i=1}^{L} \frac{r_\theta(i)^{o_i}}{o_i!} \qquad (3)$$

where $s = \sum_{i=1}^{L} o_i$ is the sample diversity. There are $S - s$ missing species.

B. Parameterization

Antigen receptor repertoires have been observed to follow Zipf's law [11]: the logarithms of the frequencies of clones are inversely proportional to the logarithms of their ranks by frequency. As a continuous analog of this discrete power law behavior, we make the parametric ansatz $d\mu(\lambda) \propto \lambda^{\gamma-1} \exp\left(-\frac{\lambda_a}{\lambda} - \frac{\lambda}{\lambda_b}\right) d\lambda$. The exponential factors cut off scaling behavior from below and above, and correspond to minimum and maximum abundances in the repertoire. This distribution, properly normalized, is the generalized inverse Gaussian [12]. For Poisson abundance models, the parameters λ_a and λ_b are strongly asymptotically correlated in the likelihood for fixed γ [13, 14]. This manifests as a ridge in parameter space that confounds likelihood maximization. A transformation that minimizes off-diagonal components of the Fisher information matrix is therefore introduced, resulting in the more orthogonal parameterization

$$d\mu_\theta(\lambda) = \frac{\xi^{-\gamma}}{2K_\gamma(\omega)} \lambda^{\gamma-1} e^{-\frac{\omega}{2}\left(\frac{\xi}{\lambda} + \frac{\lambda}{\xi}\right)} d\lambda, \qquad (4)$$

with parameter vector $\theta = (\gamma, \omega, \xi)$. $K_\gamma(\omega)$ denotes the modified Bessel function of the second kind, arising by imposing normalization. Excellent fits to naive and memory occupancy data were obtained with mixtures of two such distributions (see section IV). Lognormal and Pareto distributions were also considered, but produced substantially worse results.

Under the parameterization (4), the distribution (2) becomes

$$r_\theta(i) = \binom{L}{i} \frac{\xi^{-\gamma}}{2K_\gamma(\omega)} \int_0^\infty d\lambda \, \frac{\lambda^{\gamma-1} \left(e^{\frac{\lambda}{L}} - 1\right)^i}{e^{\lambda + \frac{\omega}{2}\left(\frac{\xi}{\lambda} + \frac{\lambda}{\xi}\right)}}. \qquad (5)$$

Although not available in closed-form, these $L + 1$ integrals can be approximated by quadrature to evaluate the

likelihood (3). Modeling as a mixture of two distributions of the form (4) adds a mixing parameter, $0 \leq \alpha \leq 1$, with $r_\theta(i) = (1-\alpha) r_{\theta_1}(i) + \alpha\, r_{\theta_2}(i)$.

C. Maximum likelihood diversity estimation

Direct maximization of the likelihood (3) is computationally formidable, as it constitutes a mixed integer nonlinear programming problem. However, it may be factorized in the suggestive form

$$\mathcal{L}(\theta, S|o) = \mathcal{L}_b(\theta, S|o)\, \mathcal{L}_m(\theta|o),$$

where we define the binomial

$$\mathcal{L}_b(\theta, S|o) = \binom{S}{s} r_\theta(0)^{S-s} \left(1 - r_\theta(0)\right)^s,$$

and zero-truncated multinomial

$$\mathcal{L}_m(\theta|o) = s! \prod_{i=1}^{L} \frac{1}{o_i!} \left(\frac{r_\theta(i)}{1 - r_\theta(0)}\right)^{o_i}.$$

An approach to approximate maximization of $\mathcal{L}(\theta, S|o)$, proposed by Sanathanan as conditional maximum likelihood estimation [15], is to first compute

$$\hat{\theta} = \arg\max_\theta \mathcal{L}_m(\theta|o),$$

which is independent of S and can be obtained by nonlinear numerical maximization of the log-likelihood $\ell_m(\theta|o) = \log \mathcal{L}_m(\theta|o)$. A constrained gradient ascent algorithm [16] was used in the present work. Differentiation gives gradient components of the form

$$\frac{\partial \ell_m(\theta|o)}{\partial \theta_j} = \frac{s}{1 - r_\theta(0)} \frac{\partial r_\theta(0)}{\partial \theta_j} + \sum_{i=1}^{L} \frac{o_i}{r_\theta(i)} \frac{\partial r_\theta(i)}{\partial \theta_j},$$

with

$$\frac{\partial r_\theta(i)}{\partial \theta_j} = \binom{L}{i} \int_0^\infty d\left(\frac{\partial \mu_\theta(\lambda)}{\partial \theta_j}\right) e^{-\lambda} \left(e^{\frac{\lambda}{L}} - 1\right)^i,$$

which may be evaluated by quadrature for the parameterization (4).

Having computed $\hat{\theta}$, it remains to maximize the richness piece of the likelihood. A lemma due to Chapman [17] can be invoked to give

$$\begin{aligned} \hat{S} &= \arg\max_{S \in \mathbb{N}} \mathcal{L}\left(\hat{\theta}, S \middle| o\right) \\ &= \arg\max_{S \in \mathbb{N}} \mathcal{L}_b\left(\hat{\theta}, S \middle| o\right) \\ &= \left\lfloor \frac{s}{1 - r_{\hat{\theta}}(0)} \right\rfloor. \end{aligned}$$

Sanathanan's articulation of an asymptotic theory for the estimator $\hat{S}$ showed it to be equivalent to direct maximization of $\mathcal{L}(\theta, S|o)$ for large S. A corresponding approach was taken by Rodrigues [18] in an empirical Bayes treatment to approximate a posterior distribution for S. The density $\mu'_\theta(\lambda)$ is viewed as a prior which is realized in the repertoire for large S. Due to the large diversity of the BCR repertoire, we employ the diversity estimator $\hat{S}$, first investigating its accuracy via simulation.

D. Shannon diversity in a Poisson abundance model

To quantify the degree of uniformity in repertoire clonal abundance we derive a standard entropy-based index of diversity applied to a Poisson abundance model. For a repertoire with richness S and clone-wise population fractions given by $\pi_1, \pi_2, \ldots \pi_S$, the Shannon index [19] is defined as the information entropy of the clone-wise abundance distribution.

$$H = -\sum_{i=1}^{S} \pi_i \log \pi_i.$$

The maximum entropy, $H_o = \log S$, occurs when $\pi_i = 1/S$ for all clones. We define clonality, C, as the complement of the normalized Shannon entropy $C = 1 - H/H_o = 1 - H/\log S$. C ranges on the unit interval, with zero denoting maximally uniform abundance across clones and unity denoting the most disparity (dominated by a single clone).

In a Poisson abundance model, each clone, i, is assigned a Poisson frequency, λ_i, which is related to its population fraction, π_i, by $\lambda_i = \langle n \rangle \pi_i$, where $\langle n \rangle$ denotes the expected sample size.

$$\langle n \rangle = \sum_{i=1}^{S} \lambda_i.$$

With the measure parameterized by θ this becomes

$$\begin{aligned} \langle n \rangle_{S,\theta} &= S \int_0^\infty d\mu_\theta(\lambda) \lambda \\ &= S I_1(\theta), \end{aligned}$$

where we've defined the integral

$$I_1(\theta) = \int_0^\infty d\mu_\theta(\lambda) \lambda.$$

The Shannon index for a Poisson abundance model is then

$$\begin{aligned} H(S, \theta) &= -S \int_0^\infty d\mu_\theta(\lambda) \frac{\lambda}{\langle n \rangle} \log \frac{\lambda}{\langle n \rangle} \\ &= \log \langle n \rangle - \frac{S}{\langle n \rangle} \int_0^\infty d\mu_\theta(\lambda) \lambda \log \lambda \\ &= \log S + \log I_1(\theta) - \frac{I_2(\theta)}{I_1(\theta)}, \end{aligned}$$

with

$$I_2(\theta) = \int_0^\infty d\mu_\theta(\lambda) \lambda \log \lambda.$$

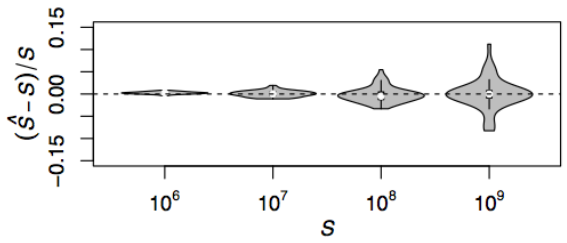


FIG. S1: **Performance of diversity estimation on simulated data.** One hundred simulations were performed for each of four diversity values, and Diversity estimates were computed for each. The resulting fractional errors are summarized as violin plots.

The clonality is then evaluated at the MLE as

$$C(\hat{S},\hat{\theta}) = 1 - \frac{H(\hat{S},\hat{\theta})}{\log \hat{S}} = \frac{1}{\log \hat{S}}\left(\log I_1(\hat{\theta}) - \frac{I_2(\hat{\theta})}{I_1(\hat{\theta})}\right).$$

For parameterization (4) the necessary integrals may be evaluated in terms of modified Bessel functions as

$$I_1(\hat{\theta}) = \xi \frac{K_{-\gamma-1}(\omega)}{K_\gamma(\omega)}$$

and

$$I_2(\hat{\theta}) = \frac{\xi}{K_\gamma(\omega)}\left(\log \xi\ K_{\gamma+1}(\omega) - \left[\frac{d}{dx}K_x(\omega)\right]_{x=-\gamma-1}\right)$$

It is trivial to extended this to a model with two mixed generalized inverse Gaussians.

IV. RESULTS

A. Simulation validation

To validate our methodology for inferring richness, simulations were performed by generating random draws from the likelihood (2). Fig. S1 shows violin plots for fractional error in diversity estimation for four sets of 100 simulations. Each violin is for a set of 100 simulations with identical diversity (S) and shows the distribution of the fractional error of the MLE, $(\hat{S}-S)/S$.

The values of S for the four sets are 10^6, 10^7, 10^8, and 10^9. For all four sets, the expected sample size is fixed at about 4.7 million cells. This is achieved by tuning the scale parameter ξ inversely as S. This is necessary to address a property of the sampling model: the expected sample size is proportional to both S and ξ, but we want expected sample size to be the same in all simulations as we tune S. Remaining parameters were fixed at $\gamma = -1$ and $\omega = 0.01$ across all simulations (values similar to those arising in analyzing real data). Even at the high end of diversity, the expected error is only a few percent, demonstrating the efficacy of conditional maximum likelihood estimation in estimating an unknown population parameter.

B. B cell diversity estimation

Details of diversity estimation applied to experimental data are presented in Fig. S2 and Table. S1, and diversity metric inferences are summarized in main text Fig. 2. Occupancy data with visualized fits of the MLE are shown in Fig. S2. Excellent fits are obtained for all data sets, as assessed by comparison to expectation values $\langle o_i \rangle = \hat{S}\ r_{\hat{\theta}}(i)$, $i = 1, 2, \ldots, L$, and with variation characterized by the quantile functions of the binomial marginal of likelihood (3) at $\hat{S}$ and $\hat{\theta}$. Estimated richness, clonality, and parameterization values are very consistent between the two samples of subject 1's naive BCR repertoire, and take on characteristic values according to cell population.

V. DISCUSSION

By synthesizing flow cytometry and replicate immunosequencing, approximate digital cell counting of memory and naive B cell repertoires of three adults was enabled, providing the deepest and most quantitatively robust characterization of the repertoire yet available. Diversity of the repertoire was inferred using a novel likelihood model devised for replicate-based presence-absence data. Estimates of both clonal richness and evenness of abundance distributions were attained, showing consistency across individuals, but distinct clustering by cell population. Across naive samples, the estimated richness is similar to the expected number of total naive B cells in circulation, suggesting that the typical naive B cell undergoes no proliferation prior to antigen stimulation. Memory richness is consistent with several divisions on average, but higher disparity in abundance (indicated by lower clonality), likely corresponding to clonal expansions in response to antigen stimulation.

[1] J. Bunge and M. Fitzpatrick, J. Am. Statist. Assoc. **88**, 364 (1993).

[2] C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, and

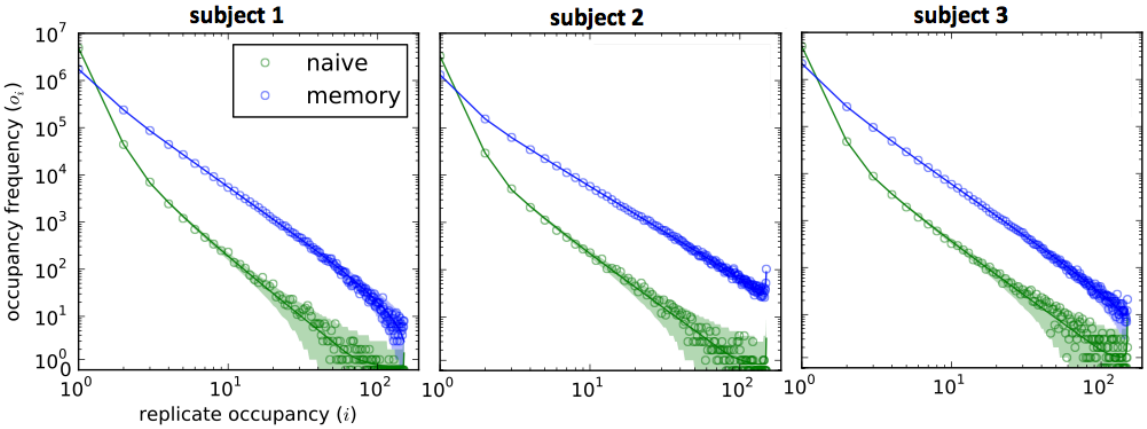


FIG. S2: **Diversity estimation results.** Replicate occupancy data (circles) for naive (green) and memory (blue) samples, with expected occupancies at the MLE (solid lines) and 99% marginal intervals (colored bands), indicating goodness of fit. Data for the second naive sample for patient 1 is omitted because results were not visually distinct. See Table S1 for numerical details of MLE results.

TABLE S1: **MLE details.** Diversity, ($\hat{S}$), parameterization, $\hat{\theta} = \theta_1 + \alpha\theta_2$, and clonality, $C(\hat{S}, \hat{\theta})$, for all samples.

subject	population	$\hat{S}$ (10^9)	θ_1	θ_2	α	$C(\hat{S}, \hat{\theta})$
1	**naive 1**	2.39	(-1.46, 9.04 10^{-4}, 8.69)	(-.097, .379 , 9.83 10^{-4})	.944	.033
	naive 2	2.46	(-1.47, 9.44 10^{-4}, 8.69)	(-.102, .410, 8.74 10^{-4})	.945	.031
	memory	.0527	(-1.13, 7.98 10^{-2}, 9.07)	(-.0722, .293, .0132)	.960	.043
2	**naive**	1.11	(-1.23, 7.28 10^{-3}, 8.69)	(-.0944, .416, 1.68 10^{-3})	.998	.029
	memory	.0765	(-.860, 3.67 10^{-2}, 9.08)	(-.0706, .285, 6.38 10^{-3})	.973	.042
3	**naive**	1.97	(-1.25, 1.80 10^{-3}, 8.69)	(-.0962, .393, 1.39 10^{-3})	.989	.030
	memory	.116	(-1.13, 5.03 10^{-2}, 9.07)	(-.0778, .287, 6.67 10^{-3})	.968	.042

S. R. Quake, Proc. Natl. Acad. Sci. U.S.A. **110**, 13463 (2013).

[3] H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, Blood **114**, 4099 (2009).

[4] I. J. Good and G. H. Toulmin, Biometrika **43**, 45 (1956).

[5] B. Efron and R. Thisted, Biometrika **63**, 435 (1976).

[6] T. Daley and A. D. Smith, Nat. Meth. (2013).

[7] G. A. Rempala, M. Seweryn, and L. Ignatowicz, J. Theor. Biol. **269**, 1 (2011).

[8] J. Greene, M. R. Birtwistle, L. Ignatowicz, and G. A. Rempala, J. Theor. Biol. **326**, 1 (2013).

[9] C. S. Carlson, R. O. Emerson, A. M. Sherwood, C. Desmarais, M.-W. Chung, J. M. Parsons, M. S. Steen, M. A. LaMadrid-Herrmannsfeldt, D. W. Williamson, R. J. Livingston, et al., Nat. Comms. **4** (2013).

[10] H. S. Wilf, *Generatingfunctionology* (Academic Press, London, 2000).

[11] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Jr, Proc. Natl. Acad. Sci. U.S.A. **107**, 5405 (2010).

[12] B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, vol. 9 of *Lecture Notes in Statistics* (Springer, New York, NY, 1982).

[13] G. Z. Stein, W. Zucchini, and J. M. Juritz, J. Am. Statist. Assoc. **82**, 938 (1987).

[14] G. E. Willmot, J. Am. Statist. Assoc. **83**, 517 (1988).

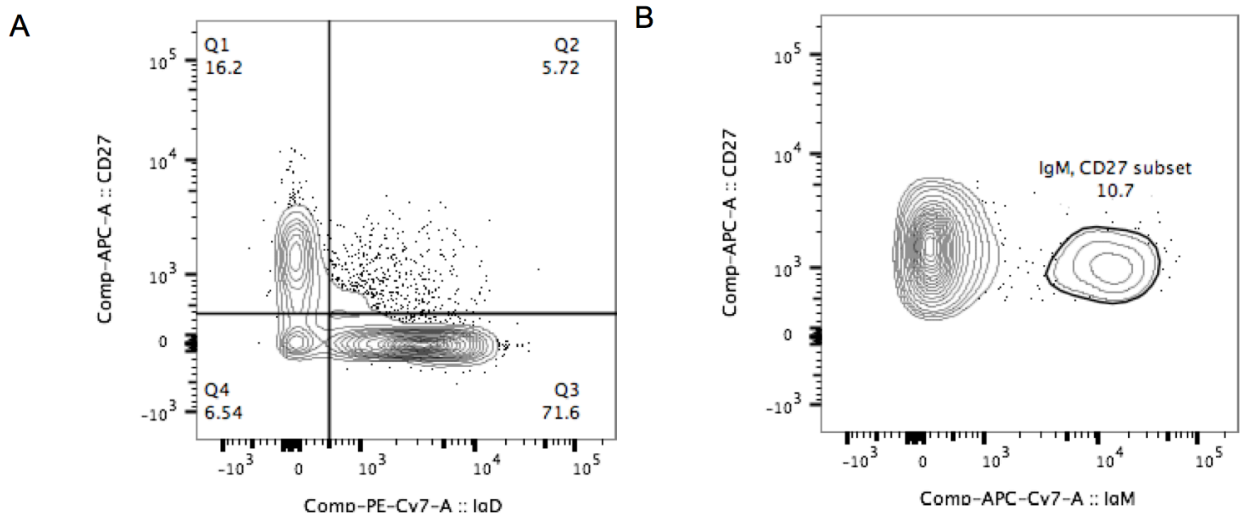
[15] L. Sanathanan, J. Am. Statist. Assoc. **72**, 669 (1977).

[16] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, ACM Trans. Math. Softw. **23**, 550 (1997).

[17] D. Chapman, *Some properties of the hypergeometric distribution with applications to zoological sample censuses*, University of California publications in statistics (University of California Press, 1951).

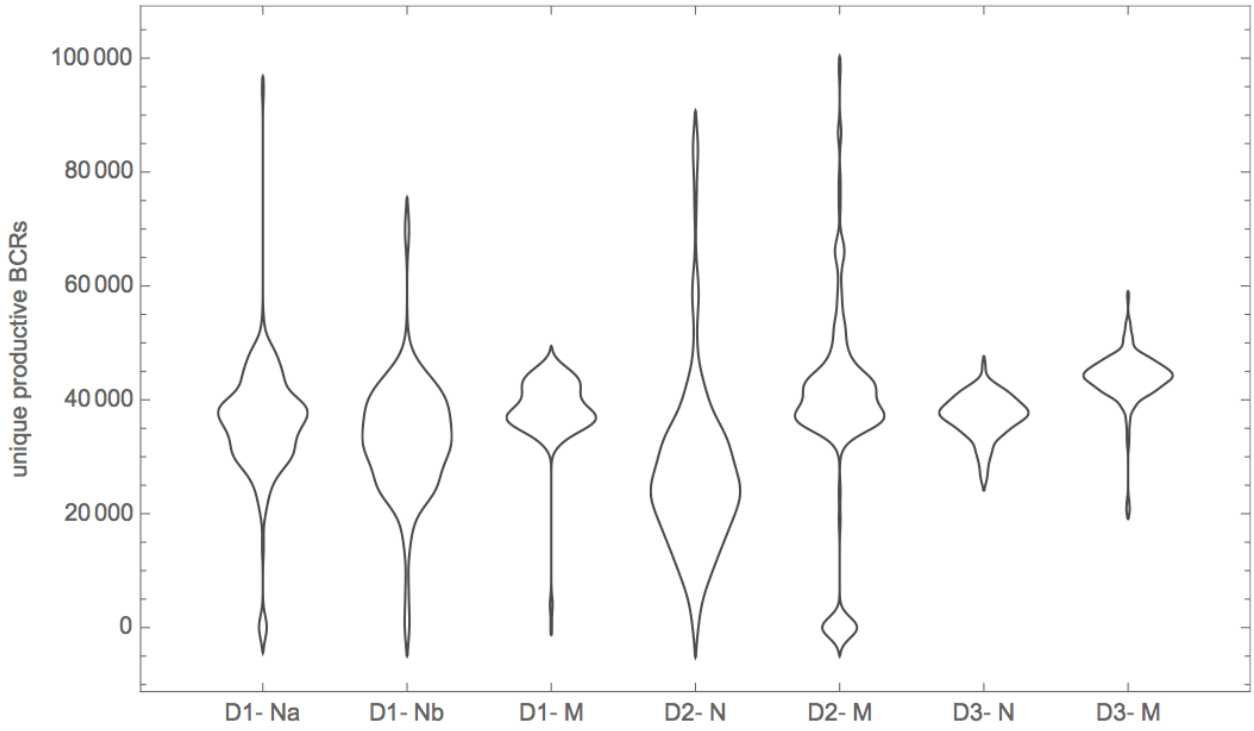
[18] J. Rodrigues, L. A. Milan, and J. G. Leite, Biom. J. **43**, 737 (2001).

[19] C. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

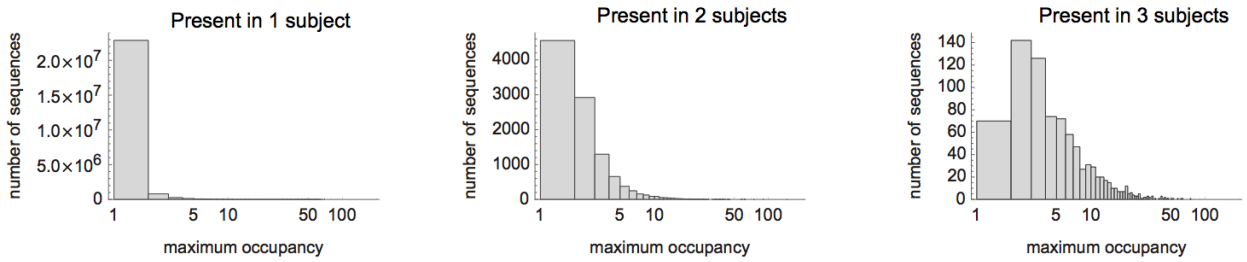


Supplementary Figures

S1 Fig: **Representative contour plots of peripheral blood B cell subsets.** (**a**) CD27 (y-axis) and IgD (x-axis) expression on gated $CD19^{+}$ B cells. (**b**) CD27 (y-axis) and IgM (x-axis) expression on gated $CD19^{+}CD27^{+}$ B cells.



S2 Fig: **Distribution of the number of unique sequences across 188 wells for each sample used in this study.**



S3 Fig: **Distribution of maximum occupancy among sequences found in only one subject, in any two subjects, and in all three subjects.**

3.9 Notes

Acknowledgments

We are grateful to Dr. Christopher Tipton for helpful suggestions, and to Katie Moran for help with editing. W.S.D. thanks Kameron Decker Harris for helpful mathematical discussions and Erick Matsen for comments and corrections to the S1 Method included in the Supporting Information.

Author Contributions

Conceived and designed the experiments: WSD PL TMS AMS CSC PDG ND ROE HSR. Performed the experiments: PL AMS. Analyzed the data: WSD PL TMS MV ROE. Contributed reagents/materials/analysis tools: HSR. Wrote the paper: WSD PL TMS MV HSR.

This work was published in *PLOS One* as

DeWitt WS*, Lindau P*, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS. A public databse of memory and naïve B cell receptor sequences. PLoS One. 2016 Aug 11:11(8)

4 Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection

4.1 Abstract

A detailed characterization of the dynamics and breadth of the immune response to an acute viral infection, as well as the determinants of recruitment to immunological memory, can greatly contribute to our basic understanding of the mechanics of the human immune system and can ultimately guide the design of effective vaccines. In addition to neutralizing antibodies, T cells have been shown to be critical for the effective resolution of acute viral infections. We report the first in-depth analysis of the dynamics of the $CD8^+$ T cell repertoire at the level of individual T cell clonal lineages upon vaccination of human volunteers with a single dose of YF-17D. This live attenuated yellow fever virus vaccine yields sterile, long-term immunity and has been previously used as a model to understand the immune response to a controlled acute viral infection. We identified and enumerated unique $CD8^+$ T cell clones specifically induced by this vaccine through a combined experimental and statistical approach that included high-throughput sequencing of the CDR3 variable region of the T cell receptor β-chain and an algorithm that detected significantly expanded T cell clones. This allowed us to establish that (i) on average, ~2,000 $CD8^+$ T cell clones were induced by YF-17D, (ii) 5 to 6% of the responding clones were recruited to long-term memory 3 months postvaccination, (iii) the most highly expanded effector clones were preferentially recruited to the memory compartment, and (iv) a fraction of the YF-17D-induced clones could be identified from peripheral blood lymphocytes solely by measuring clonal expansion.

4.2 Introduction

During the acute response to a viral infection, viral antigen (Ag)-specific effector CD8 T cell clones (also known as cytotoxic T lymphocytes, or CTLs) become activated and expand as they recognize and eliminate infected host cells (1, 2). The Ag specificity of a T cell clone is determined by the T cell receptor (TCR), which is encoded by random, RAG-mediated V(D)J re- combination. Thus, each T cell clone may be identified by its unique TCRβ CDR3 region, formed from the joining of the V, D, and J gene segments along with deletions and nontemplated insertions at the junctions, with CDR3 being the primary determinant of Ag specificity (3, 4). The identification and tracking of virus-specific CTLs has resulted in the extensive characterization of their phenotype and function (5–8). The identification of virus- specific T cells during the course of an infection has allowed the measurement of the number of unique clones responding to a particular viral epitope (9–11). These studies suggested that the magnitude of the T cell clonal response to different viral Ags is not uniform; for example, in the case of the yellow fever vaccine (YFV), peptide NS4b induces a more robust T cell response than peptide NS5 (9, 12). Moreover, there is extensive variability in the number of unique clones activated by a particular viral epitope (13, 14), which depends both on the quantity of peptide presented (15) and on the microenvironment of the lymph node where the T cell encounters the Ag (7). In addition, responses to chronic and acute viruses seem to be characterized by different patterns of activation and waning of effector cells, as well as different memory cell phenotypes, which might be related to the different patterns of exposures to viral Ags in these two different types of infection (reviewed in reference 16). Finally, major histocompatibility complex polymorphisms lead to variable epitope presentation in

different individuals (17, 18), complicating the characterization of dominant and nondominant clonal CTL responses.

The formation of virus-specific $CD8^+$ memory T cells is also believed to be dependent on the magnitude of the clonal response to Ag (19, 20). After an acute infection is resolved, the virus-specific effector $CD8^+$ T cell pool contracts (21), and a much smaller number of long-lived memory T cells that are capable of responding to subsequent infections is maintained (22). It is thought that effector T cell clones present in high abundance are recruited to the memory repertoire with higher frequency than less abundant clones (11, 23, 24), but it is not clear whether this simply reflects the limitations of currently available techniques. Therefore, highly sensitive techniques are necessary to establish the contribution of less abundant clones to the memory pool (12). Furthermore, to date it has not been possible to relate the magnitude and diversity of the effector T cell response to the subsequent abundance of individual clones in the memory T cell repertoire. Thus, the detailed characterization of the dynamics of the T cell repertoire in response to an acute viral infection can increase our understanding of the breadth of the immune response, the formation of immunological memory, and how the human immune system responds to acute viral infections and immunization with viral vaccines.

We used vaccination with the yellow fever (YF) virus vaccine YF-Vax, which is based on the YF-17D204 attenuated strain, as a model of acute viral infection. YF-17D harbors only 20 amino acid changes compared to the wild-type strain, most of which are found in the E protein and are thought to result in changes in viral tissue tropism (25). In addition, this attenuated virus is replication competent, so that administration of the YFV results in a mild viral infection that is predicted to elicit an immune response that is almost

identical in quality to that induced by wild-type infection (26). Since exposure to YF virus is geographically limited, and YFV is a very effective vaccine that elicits an optimal, long-term protective immune response upon administration of a single dose, this model has been used extensively to explore the human immune response to a controlled, self-resolving acute viral infection (reviewed in references 16 and 27). These seminal studies have shown that (i) the ability of YF-17D to infect dendritic cells and signal through multiple Toll-like receptors may be related to the effectiveness of this vaccine (28); (ii) neutralizing antibodies (nAbs) are the best surrogate marker for protection against YF virus and remain detectable for many years (29, 30); and (iii) $CD8^+$ T cells expand massively before nAbs can be detected (and are thus likely involved in the control of viremia) and persist in the memory compartment for decades (6, 30).

Our understanding of the $CD4^+$ response to YFV is limited. Although helper T cells are clearly required for the production of YFV-specific Abs (including nAbs), different studies have reported variable levels of induction of $CD4^+$ T cells upon vaccination with YFV (30, 31). Some analyses have revealed that cytokine-producing YFV-specific $CD4^+$ T cells can be detected as early as day 2 postvaccination and that they return to baseline by day 28, suggesting that the kinetics of $CD4^+$ T cells precede those of $CD8^+$ T cells (12, 32). Recently, James et al. used class II HLA-DR restricted, YFV-specific tetramers to characterize the $CD4^+$ response to YFV in more depth, showing that all 10 proteins in the YF virus genome contain antigenic epitopes recognized by $CD4^+$ T cells (33). This study also revealed a wide range of frequencies of $CD4^+$ T cells specific for a limited number of YFV epitopes in peripheral blood (from 0 to 100 cells per million $CD4^+$ T cells) and established that YFV-specific T cells, which display a predominant Th1-like memory

phenotype, occur at ~10- to 100-fold- higher frequencies in vaccinated versus unvaccinated individuals, depending on the time point considered (33).

In contrast, there have been several detailed analyses of the kinetics and phenotype of $CD8^+$ T cells induced by vaccination with YFV. For example, Miller et al. (6) showed that activated effector $CD8^+$ T cells (T_{AE}) peak 2 weeks after administration of the YFV and defined the YFV-specific subpopulation of $CD8^+$ CTL cells as $CD38^+$ HLA-DR^+ Ki-67^+ Bcl-2^{lo}. In addition, this study established a strong correlation between the levels of $CD38^+$ HLA-DR^+ $CD8^+$ T cells and the expression of gamma interferon (IFN-γ) by total $CD8^+$ T cells in response to YF virus-infected cells, and it demonstrated that stimulation of $CD8^+$ T cells from YFV-vaccinated volunteers with a comprehensive pool of peptides that span the YF virus polyprotein also induced IFN-γ. Since un- related memory $CD8^+$ T cells (such as those specific for chronic viruses like Epstein-Barr virus [EBV] and cytomegalovirus [CMV] and therefore presumed to preexist at the time of vaccination with YFV) were not found among the expanded $CD8^+$ T cell population, these observations suggest that, at least in the case of YFV, the bystander effect is minimal, and they also imply that the vast majority of T_{AE} clones observed after administration of YF-17D are YF virus specific. Finally, those authors showed that Ag-specific cells could be identified more than 30 days postvaccination, indicating that the YFV-specific effector $CD8^+$ T cells had waned and also that a certain proportion of them had entered the memory compartment (6). Subsequent work from the same group employed an array of overlapping peptides that spanned the entire YF virus polyprotein to demonstrate that vaccination with YFV induces a broad $CD8^+$ T cell response that targets several epitopes in each of the 10 viral proteins (9). The use of tetramers carrying an immunodominant

epitope from the nonstructural NS4b protein helped define the phenotypes of YFV-specific $CD8^+$ T cells through the expansion, contraction, and memory phases of the immune response, further confirming that $CD38^+$ $HLA\text{-}DR^+$ $CD8^+$ T cells dramatically expand after YFV-17D administration and produce cytotoxic effector molecules (9). Similar results were observed by Co et al., who identified YFV-specific proliferation and cytolytic responses on day 14 postvaccination and isolated $CD8^+$ T cell lines that were specific for epitopes from structural and nonstructural YF virus proteins, some of which persisted for up to 19 months postvaccination (10). Again, follow-up data from a tetramer-based approach showed that YFV-specific $CD8^+$ T cells could be identified as early as 7 to 9 days postvaccination, before IFN-γ production was detectable, that memory cells corresponded mostly to a differentiated effector phenotype ($CD45RA^-$ $CCR7^-$ $CD62L^-$), and that these peptide-specific responses lasted for at least 54 months (34). A more recent study using a limited set of YF virus HLA-tetramer epitopes suggested that the $CD8^+$ response to YFV is broad and complex and that responses to different epitopes vary in magnitude and duration (12). Those authors also found that YFV-specific effector $CD8^+$ T cells were $CD45RA^{hi}$ $CCR7^-$ $PD1^+$ $CD27^{hi}$ and that only some of these cells transition to the T cell memory compartment, at which point they became $CD45RA^+$ $CCR7^-$ $PD1^-$ $CD27^{lo}$(12).

In this study, we developed a complementary approach to study the dynamics of the human effector and memory $CD8^+$ T cell repertoire upon acute viral infection. First, we isolated total peripheral blood mononuclear cells (PBMCs) from volunteers who received the YF-17D vaccine prevaccination (on day 0) and on days 14 and 90 postvaccination, and we used flow cytometry to sort a fraction of these samples into $CD8^+$

$CD38^+$ $HLA\text{-}DR^+$ activated, effector T cells on day 14 (i.e., at the peak of their abundance) and into memory $CD8^+$ T cells on days 0 and 90. High- throughput sequencing of the rearranged TCRβ locus in each sample, combined with a computational method to identify expanded T cell clones, allowed the characterization of individual, YFV-induced $CD8^+$ T cell clones during the acute phase, to estimate the abundance of each of these clones and to track them into the memory phase of the antiviral response. This synthesis of flow cytometry sorting protocols and high-throughput sequencing enabled the measurement of the T cell response to viral infection at an unprecedented resolution. Finally, we show that our approach allows the identification of many YFV-induced $CD8^+$ T cell clones by assessing clonal expansion directly from peripheral blood samples (i.e., without previous sorting of activated, effector, or memory $CD8^+$ T cells) and that a large proportion of these clones overlap those identified through immunosequencing of the flow cytometry-sorted activated, effector $CD8^+$ T cell population.

4.3 Materials and Methods

Vaccination and sample collection

Nine volunteers between the ages of 18 and 45 years consented under Fred Hutchinson Cancer Research Center (FHCRC) and the University of Washington Vaccine Research Clinic (UWVRC) IRB protocols to receive the yellow fever single-dose vaccine YF-VAX (based on the YF-17D204 strain of the yellow fever virus [26]) and to have 200 ml of blood drawn at three different time points: immediately before vaccination (day 0), 2 weeks postvaccination (day 14), and 3 months postvaccination (day 90) (Table 1). Written informed consent to use the blood samples in this study was obtained from each

subject. The administration of the YF vaccine and all blood draws and were performed at the UWVRC.

Cell sorting

All cell sorting was performed at FHCRC. Whole-blood samples (200 ml) were collected and PBMCs were isolated by using Histopaque (Sigma-Aldrich, St. Louis, MO) density gradient centrifugation. $CD8^+$ T cells were isolated from total PBMCs by magnetic separation using CD8 MicroBeads and the autoMACs Pro separator (both from Miltenyl Biotec, Auburn, CA), followed by staining with anti-CD3–Alexa Fluor 700, anti-CD8–allophycocyanin (APC)-H7, anti-CD38–phycoerythrin (PE), HLA-DR–fluorescein isothiocyanate, anti-CD14–Pacific Blue, anti-CD19–V450, anti-CD45RO–PE Cy7, anti-CD45RA–APC, anti-CD62L–peridinin chlorophyll protein-Cy5.5, and 4=,6-diamidino-2-phenylindole (DAPI) (all obtained from BD BioSciences, San Jose, CA). T cell subpopulations were sorted using the BD FACSAria II system and FACSDiva v6.1.3 software (BD Biosciences). First, we gated on propidium iodide-negative (PI^-) $CD14^-$ $CD19^-$ to remove dead cells, monocytes, and B cells and then on $CD3^+$ $CD8^+$ to exclude non-T cell lymphocytes and $CD4^+$ T cells. Finally, we isolated four different $CD8^+$ T cell subsets: from the day 0 prevaccination samples we isolated $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ memory T cells (T_{M-0}); from the day 14 postvaccination samples we isolated $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD38^+$ $HLA\text{-}DR^+$ Ag-experienced, activated effector T cells (T_{AE-14}); and from the day 90 postvaccination samples we isolated $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ $CD62L^{lo}$ effector memory T cells (T_{EM-90}) and $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ $CD62L^{hi}$ central memory T cells (T_{CM-90}). To

avoid contamination, $CD38^+$ $HLA\text{-}DR^+$ cells were excluded from the effector memory and central memory T cell populations. Day 90 samples from three of the volunteers had to be discarded due to contamination.

DNA extraction and immunosequencing

Genomic DNA was purified from total PBMCs and each sorted T cell population sample by using the QIAmp DNA blood minikit (Qiagen). For each sample, DNA was extracted from ~1 million T cells, and the TCRβ CDR3 regions were amplified and sequenced using ImmunoSEQ (Adaptive Biotechnologies, Seattle, WA) as previously described (35). In brief, bias-controlled V and J gene primers were used to amplify rearranged V(D)J segments for high- throughput sequencing at ~20x coverage. After correcting sequencing errors via a clustering algorithm, CDR3 segments were annotated according to the International ImMunoGeneTics Collaboration (36, 37) to identify the V, D, and J genes that contributed to each rearrangement. Sequences were classified as nonproductive if it was determined that nontemplated insertions or deletions produced frameshifts or premature stop codons. We used a mixture of synthetic TCR analogs in each PCR to estimate the absolute template abundance (i.e., the number of cells bearing each unique TCR sequence) from sequencing data, as previously described (38).

Identification of expanded and enriched effector T cell clones

Given two samples from the same subject (perhaps drawn at different time points or under differing cell sorting conditions) we wish to identify which T cell clones have significantly increased in relative abundance in the repertoire. Data from each sample consists of abundance for each TCRβ clone in the sample.

Let the repertoire contain *S* distinct clones, and that their proportional abundances at time points 1 and 2 be given by the multinomial vectors $\boldsymbol{\pi}^{(1)} = \{ \pi^{(1)}_1, \pi^{(1)}_2, \ldots, \pi^{(1)}_S\}$ and $\boldsymbol{\pi}^{(2)} = \{ \pi^{(2)}_1, \pi^{(2)}_2, \ldots, \pi^{(2)}_S\}$, with $\sum_{i=1}^{S} \pi_i^{(j)} = 1$. Suppose that *n* clones have changed in abundance between the two time points. Identify these clones with the *n*-element index vector $\boldsymbol{\Delta}$.

We next assume that the aggregated proportional change of all truly changed clone abundances is small, that is: $\sum_{i\in\boldsymbol{\Delta}}(\pi_i^{(2)} - \pi_i^{(1)}) \ll 1$. In this regime, each observed clone can be independently tested for significance using a 2x2 contingency table. We employ the Fisher exact test to compute a p-value for each clone across the two samples. Specifically, suppose clone *i* is observed with abundance $k_i^{(1)}$ at time point 1 and $k_i^{(2)}$ at time point 2. We compute a p-value for the 2x2 contingency table with these abundances on one row, and the remaining abundances (for clones other than *i*) on the other. By summing over hypergeometric probabilities for all more extreme contingency tables, the Fisher exact test gives the p-value for the null hypothesis that the proportion of clone *i* in the repertoire is the same at both time points, to wit: $\pi_i^{(1)} = \pi_i^{(2)}$.

Let *s* represent the number of distinct clones observed across the two samples, where in general $s \leq S$. Without loss of generality, indices 1 through *s* of the repertoire clones are the observed clones. After performing the above analysis on each of the *s* observed clones, we have a vector of p-values $\boldsymbol{p} = \{p_1, p_2, \ldots, p_s\}$.

To choose a rejection region (thereby identifying a set of significantly changed clones) we use the positive false discovery rate (pFDR) method of Storey(39) The pFDR is defined as the expected proportion of true null hypotheses among all rejected hypothesis.

$$\begin{aligned}\mathrm{pFDR}(\gamma) &= \Pr\left(\pi_i^{(1)} = \pi_i^{(2)} \;\middle|\; p_i \le \gamma\right) \\ &= \frac{\pi_0 \Pr\left(p_i \le \gamma \;\middle|\; \pi_i^{(1)} = \pi_i^{(2)}\right)}{\Pr\left(p_i \le \gamma\right)} \\ &= \frac{\pi_0 \gamma}{\Pr\left(p_i \le \gamma\right)}\end{aligned}$$

The second equality follows from Bayes' theorem with π_0 the prior probability that a hypothesis is null. The last equality follows from the definition of a p-value, if the p-values themselves are regarded as independently and identically distributed random variables.

For each *P* value, P_i, the associated *Q* value, Q_i, may be estimated; this is the minimum pFDR that can occur when rejecting *P* values less than or equal to P_i. By examining the number of significant tests at various *Q* value thresholds, an appropriate threshold can be selected (see Fig. 1, below). Control of pFDR is preferred for control of the familywise error rate (FWER), i.e., the probability of one or more false alternative hypotheses. The latter (typically controlled by the Bonferroni method) is overly conservative, as it fails to reject many false null hypotheses in order to attain any nontrivial FWER. The pFDR, on the other hand, rejects these hypotheses at the cost of a specifiably small proportion of rejected true null hypotheses.

The resulting set of significance tests allows the identification of T cell clones whose frequencies are different in the two samples (i.e.,dynamic T cell clones). For example, applying this algorithm to the comparison of total PBMCs isolated on day 14 postvaccination to activated $CD8^+$ T cells purified from the same sample identifies a set of enriched, activated $CD8^+$ T cells that are expected to be YFV specific. In contrast, the comparison of total PBMCs obtained from the same volunteer on day 0 (prevaccination)

and on day 14 postvaccination identifies a set of putative YFV-reactive clones based on clonal expansion.

4.4 Results

It is well established that effector $CD8^+$ T cells expand in response to an acute viral infection (39). Expanded clones can either bind specifically to a pathogen-derived epitope presented by a type I HLA molecule, or they can be induced to expand nonspecifically by cytokines released by other cells, in a process known as the bystander effect (40). In the case of the YFV model, which results in a self-limited, acute viral infection (16, 27) and has thus been extensively used to characterize the human antiviral immune response, activated effector $CD8^+$ T cells peak 2 weeks postvaccination (6, 10) and express a particular set of phenotypic markers, including CD38, HLA-DR, Ki-67, and Bcl-2 (6). The massive expansion of these activated, effector $CD8^+$ cells in response to vaccination with YFV is specific, since these cells have been shown to produce cytokines in response to stimulation with peptides from YF-17D proteins (9, 34), and existing memory $CD8^+$ T cells specific for other viruses, such as CMV or EBV, do not contribute to the activated, proliferating pool of $CD8^+$ T cells (6).

To further explore the dynamics of the T cell repertoire in response to an acute viral infection, we administered a single dose of the live attenuated YFV YF-VAX, based on the YF-17D204 strain of the YF virus (26), to nine healthy volunteers, none of whom reported being previously exposed to the YF virus or having received a YFV. We drew 200 ml of peripheral blood from each subject on day 0 (immediately prior to vaccination) and on days 14 and 90 postvaccination (Table 1). To identify $CD8^+$ T cells present in the

memory compartment prior to vaccination, we sorted a fraction of the total PBMCs obtained from all 9 subjects on day 0 into $CD8^+$ memory T cells (T_{M-0}, defined as $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ cells [41]). Similarly, to characterize the activated effector $CD8^+$ T cells induced by vaccination with YFV, we also sorted a fraction of the total PBMCs obtained from all 9 subjects on day 14 postvaccination by selecting $CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD38^+$ HLA-DR^+ activated effector $CD8^+$ T cells (T_{AE-14}) (6). Finally, to determine which of these clones enter the memory compartment, we sorted PBMCs obtained on day 90 from 6 of the subjects into effector memory (T_{EM-90}) and central memory (T_{CM-90}) $CD8^+$ T cells (respectively, $CD8^+$ $CD45RO^+$ $CD62L^{lo}$ and $CD8^+$ $CD45RO^+$ $CD62L^{hi}$) (42). We were unable to characterize the T_{EM-90} and T_{CM-90} cell populations from the 3 other subjects because these samples had to be discarded due to contamination.

To identify and quantify YFV-induced T cell clones, we extracted genomic DNA from ~1 million T cells for either total PBMCs or sorted T cell populations (Table 1). Next, we used PCR amplification and high-throughput sequencing to characterize the CDR3 regions of rearranged TCRβ loci as previously described (35). TCRβ sequences are nearly unique for each clone, so that the data can be used to assess the dynamics of the cellular adaptive immune response both over time and between T cell subpopulations. Additionally, we determined the number of original templates corresponding to each PCR-amplified clonal sequence by assessing the amplification of a set of synthetic tem- plates, thus providing an estimate of the cellular abundance for each clone in each sample (38).

Identification of vaccine-induced clones

To assess the dynamics of the YFV-induced $CD8^+$ T cell repertoire, we determined whether each unique clone (defined by sequencing the CDR3 region of the TCRβ chain) was enriched in the day 14 postvaccination, YFV-induced effector $CD8^+$ T cell compartment (T_{AE-14}), defined by the expression of CD38 and HLA-DR (6), in comparison to the corresponding total PBMC sample from that time point from the same subject. To do this, we developed a statistical method to identify clones that had significant proportional abundance differences between two samples (see Materials and Methods) (Fig. 1A). Our approach controls for the false-positive rate and takes into account experimental errors that result in the presence of false positives in the YFV-induced T_{AE-14}compartment (i.e., cells that do not have the indicated surface markers). This avoids overstating the number of YFV-induced clones, which would result from a simple enumeration of clones present in the T_{AE-14}compartment, since it includes low levels of many clones that are no more frequent than in the corresponding total PBMC sample. Instead, we considered a clone to be YFV induced if (i) it was significantly enriched in the T_{AE-14} compartment with respect to the corresponding total PBMC sample, and (ii) it carried a productive TCRβ rearrangement. Since the subjects who participated in this study had not been previously exposed to either the YF virus or a YFV, we also took into consideration whether each unique $CD8^+$ T cell clone identified was present in the day 0 prevaccination memory cell compartment (T_{M-0}). Based on these criteria, we classified T cell clones into four categories, as follows: YFV-induced clones (i.e., enriched in the T_{AE-14} compartment versus the day 14 postvaccination total PBMC sample from that individual but absent in the corresponding T_{M-0} compartment); cross-reacting or bystander clones

(i.e., enriched in the T_{AE-14} compartment versus the corresponding total PBMC sample but present in T_{M-0}), and those not enriched in the T_{AE-14} compartment that were either present or absent in T_{M-0} (Fig. 2).

For the nine subjects in the study, we detected on average 2,000 clones that were enriched in the T_{AE-14} compartment compared to the corresponding day 14 postvaccination total PBMC sample from the same individual (2,135 $\pm$ 770) (Table 2). This number constitutes a direct estimate of the number of activated, effector $CD8^+$ T cell clones that expand upon binding to HLA:YFV-derived epitope complexes in response to vaccination with YF-17D. In addition, the vast majority of these clones (on average, 91.5% [Table 2]) were absent in the T_{M-0} population and were thus clearly induced by vaccination with YFV-17D.

Characterization of the recruitment of individual clones to immunological memory

Next, we determined which of the YFV- induced clones entered the long-term central and effector memory compartments by analyzing samples obtained from six of the subjects 90 days postvaccination (Table 1). Preliminary studies demonstrated that YFV-induced $CD8^+$ T_{AE} cells return to baseline levels 30 days postvaccination and suggest that YFV Ag-specific cells that are detected beyond this time point correspond to memory cells (6). Therefore, we tracked in the day 90 postvaccination samples YFV-induced clones that were identified as enriched for the T_{AE-14} compartment but that were absent from the T_{M-0} compartment (i.e., the putative YFV-specific clones), to determine which were contained in the effector memory compartment (T_{EM-90}, defined as $CD3^+$ $CD8^+$

CD14$^-$ CD19$^-$ CD45RA$^-$ CD45RO$^+$ CD62L^{lo}), the central memory compartment (T_{CM-90}, defined as CD3$^+$ CD8$^+$ CD14$^-$ CD19$^-$ CD45RA$^-$ CD45RO$^+$ CD62L^{hi}), or both. Figure 3A and Table 3 show that 3.1% and 2.5% of YFV-induced clones absent in T_{M-0} were identified exclusively in the T_{EM-90} or the T_{CM-90} compartments, respectively, while 6.7% were identified in both. Moreover, we saw that the degree of expansion of a clone (defined as the abundance of a clone in the day 14 postvaccination total PBMC sample for clones absent in the day 0 prevaccination total PBMC sample) correlated with the efficiency of its recruitment to the memory T cell compartment (Fig. 3B).

The YFV-induced clones that were newly recruited to the T_{EM-90} or T_{CM-90}compartments represent 0.43% and 0.45% (as measured by unique clone counts), or 0.41% and 0.28% (as measured by template abundance) of the corresponding memory compartment aggregated over all samples (Fig. 4A). While the number of templates per unique CD8$^+$ T cell clone in the T_{EM-90} compartment averaged 8.3, those in the T_{CM-90}compartment averaged 2.8, indicating that YFV-induced clones recruited to the effector memory compartment are more significantly expanded than those recruited to central memory (Fig. 4B).

Finally, we were interested in determining whether expanded CD8$^+$ clones that are recruited to the memory compartment possess any particular characteristics that distinguish them from those that become activated upon vaccination but then wane. To address this, we analyzed whether several indicators of specificity (such as CDR3 length or V-J gene usage) correlated with the probability that a given CD8$^+$ T cell clone would be recruited to memory. Although no simple indicator showed an association with

recruitment to memory, we found that both the degree of expansion of a clone and the specificity determined by effector sorting (i.e., the fold enrichment in the T_{AE-14} compartment versus that in the corresponding total PBMC sample from day 14 postvaccination) were positively associated with recruitment.

Concordance between expansion in total PBMCs and enrichment in the activated effector $CD8^+$ T cell compartment

In addition to the data presented above, our approach also allowed the identification of activated, effector $CD8^+$ T cells that expanded massively in response to YFV through the direct comparison of the unsorted, total PBMC samples obtained on days 0 and 14 postvaccination. The statistical method described in detail in Materials and Methods can be applied to the identification of T cell clones that have significantly expanded in a day 14 postvaccination total PBMC sample, compared to the corresponding total PBMC sample from the same individual collected prevaccination (Fig. 1B and 5). Among all the cells present in the day 14 postvaccination sample, we identified a set that was highly expanded but that was not captured by the antiviral-specific T_{AE-14} flow cytometry sort (i.e., $CD38^+$ HLA-DR^+ $CD8^+$ T cells from the day 14 postvaccination). These clones likely corresponded to non-$CD8^+$ T cells that expressed the TCRβ receptor (such as $CD4^+$ T cells), but they could also belong to YF-induced $CD8^+$ T cells that possess different surface markers than those previously reported by Miller et al. (6), or to clonal expansions not induced by the vaccine.

Finally, to assess how well the expanded $CD8^+$ T cell clones detected in the total PBMC population based only on immunosequencing (i.e., not sorting particular cell

populations by flow cytometry) were in concordance with the previously identified T_{AE-14} clones (i.e., those identified statistically after flow cytometric sorting of $CD38^+$ $HLA\text{-}DR^+$ $CD8^+$ T cells), we counted how many expanded $CD8^+$ T cells carrying productive rearrangements identified in the total PBMC sample analysis were classified as YFV-induced through the statistical analysis of the flow cytometry-sorted $CD38^+$ $HLA\text{-}DR^+$ $CD8^+$ cell clones described above. Table 4 shows that a significant proportion of these "putatively reactive" clones— between 25% and 95.2%, depending on the subject—were present in the T_{AE-14} compartment, suggesting they were induced by YF-17D. In aggregate, 62% of the putatively reactive clones identified as having expanded in the day 14 post- vaccination total PBMC sample (compared to the corresponding prevaccination sample) could be classified as YFV induced. These data suggest that our approach has the potential of identifying vaccine-specific responding clones through the characterization of clones expanded in the total PBMC population using exclusively immunosequencing.

4.5 Discussion

Using the power of high-throughput immunosequencing, we identified and tracked YFV-induced activated, effector $CD8^+$ T cells as they clonally expanded and underwent phenotypic modification in response to vaccination of human volunteers with the YF-17D vaccine. Previous work using the YFV as a model for an acute, self-resolving viral infection illustrated the general kinetics of the human antibody, $CD4^+$-, and $CD8^+$-based antiviral immune response (6, 9, 10, 12, 16, 30–34). Although some of these studies used either peptide pools or tetramer-based approaches to address the response to a limited

number of viral epitopes (6, 9, 33), or calculated the percentage of different immune cellular compartments that represented clones induced by vaccination with YFV (reviewed in reference 16), the clonal breadth and complexity of the response to a viral infection has been beyond the reach of available methods.

In this study, we determined that an average of approximately 2,000 different $CD8^+$ T cell clonal lineages are activated by vaccination with YFV during the acute phase of the immune response and that about 12% of them can be detected in the long-term memory compartment (including both central and effector memory $CD8^+$ T cells). It would be interesting to determine if a similar number of $CD8^+$ T cell clonal lineages are induced by other viral vaccines or by naturally occurring acute viral infections. We also observed that clones that were most expanded on the total PBMC sample from day 14 postvaccination were also more likely to enter the memory compartment 3 months postvaccination, in agreement with previous data (12). Although we were unable to identify other defining characteristics that differentiate $CD8^+$ T cell clones that expand in response to YFV vaccination and are present in the memory compartment on day 90 postvaccination from those that wane during that period, future studies will attempt to characterize these two populations further, including their epitope specificity, since this would yield valuable information that could guide the design of vaccines against other pathogens. Interestingly, almost all of the clones that were markedly expanded in the total PBMC sample from day 14 postvaccination (compared to the corresponding day 0 prevaccination total PBMC sample from the same individual) were classified as YFV-induced $CD8^+$ T cells by the combination of flow cytometry and statistical analysis. In fact, we observed very few clonally expanded T cells in the periphery that were not identified as YFV-induced clones,

in agreement with previous reports showing that while $CD8^+$ T cells greatly expand in response to vaccination with YFV, the $CD4^+$ expansion is much less dramatic (6, 30, 32, 33). It is important to consider that the sampling depth used in this study limits the detection of bystander $CD8^+$ cells or of $CD4^+$ T cells that are only modestly expanded. Thus, our current level of detection is likely not sufficient to distinguish $CD4^+$ T cell expansion above the intrinsic system noise.

We did not observe a particular pattern of V(D)J gene usage among the expanded $CD8^+$ T clone repertoire. This result partially agrees with those of a preliminary study of V gene usage performed by Co et al. (34), which used a limited set of anti-human Vβ antibodies. Those authors did not observe a dominant Vβ family that predominated among the tetramer-specific $CD8^+$ T cells in two individuals vaccinated with YFV, but they reported that although gene usage changed over time from the acute to the memory phase, no particular V genes persisted between the acute and memory phases of the antiviral response (34).

Finally, it is noteworthy that many of the $CD8^+$ T cell clones identified as expanded through the comparison of the day 14 post- vaccination and the day 0 prevaccination total PBMC samples were classified as likely YFV specific in our initial characterization of clones enriched in the activated effector $CD8^+$ T cells versus the total PBMC sample on day 14 postvaccination. Thus, our approach is capable of identifying a fraction of the highly expanded $CD8^+$ T cells by immunosequencing of total PBMCs prior to infection or vaccination and during the acute response (i.e., 10 to 14 days postvaccination), and our method could be used to ascertain the establishment of long-term memory by sorting

memory T cells a few months after infection (or later) and tracking the $CD8^+$ T cells previously identified as being virus induced. Future experiments will address the epitope specificity of the YFV-induced $CD8^+$ clones, by using, for example, tetramer technology to purify clones that bind to previously identified immunodominant YFV epitopes.

A similar strategy could also be applicable to the evaluation of the B cell response to vaccines and viral infections. In conclusion, immunosequencing can be used to characterize the strength and breadth of the B and T cell responses induced by vaccines and viral infections, and it has the potential to be utilized to evaluate novel vaccines in terms of their potential ability to induce effective long- term protective immune responses.

4.6 Figures and Tables

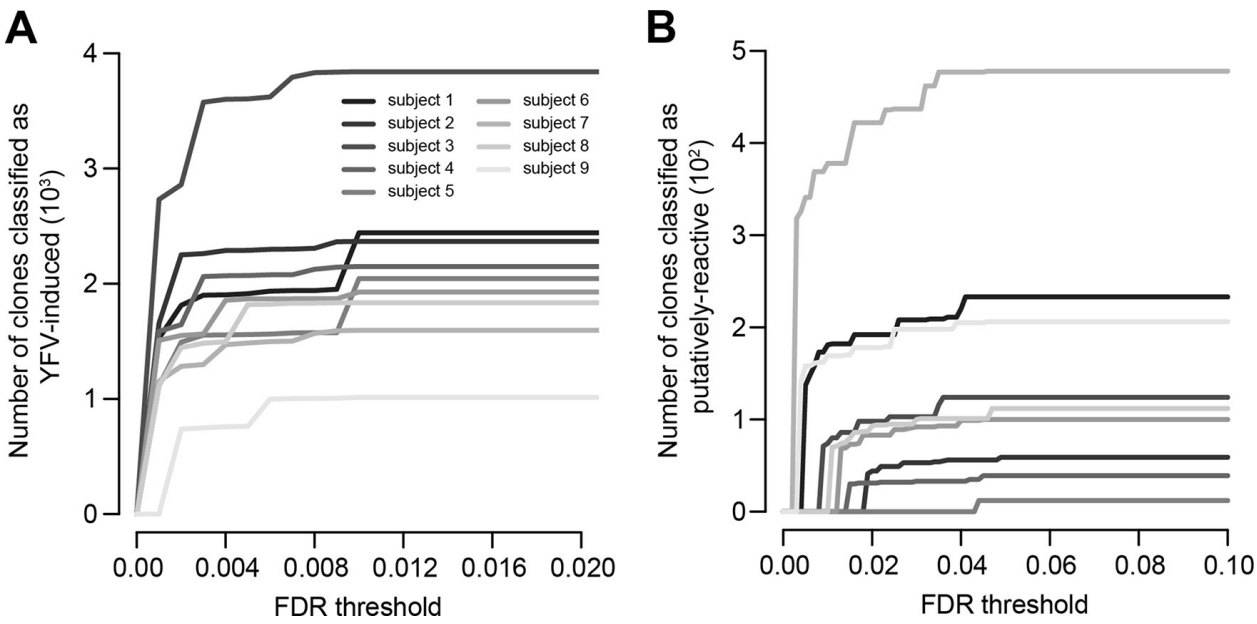


FIG 1 Selection of FDR thresholds. (A) Number of clones classified as YFV induced for various FDR significance thresholds for all subjects. A threshold of 0.01 was selected. (B) Number of clones classified as putatively reactive clones for various FDR significance thresholds for all

subjects. A threshold of 0.05 was selected. Each subject is represented by a different tone of gray, as indicated in the legend.

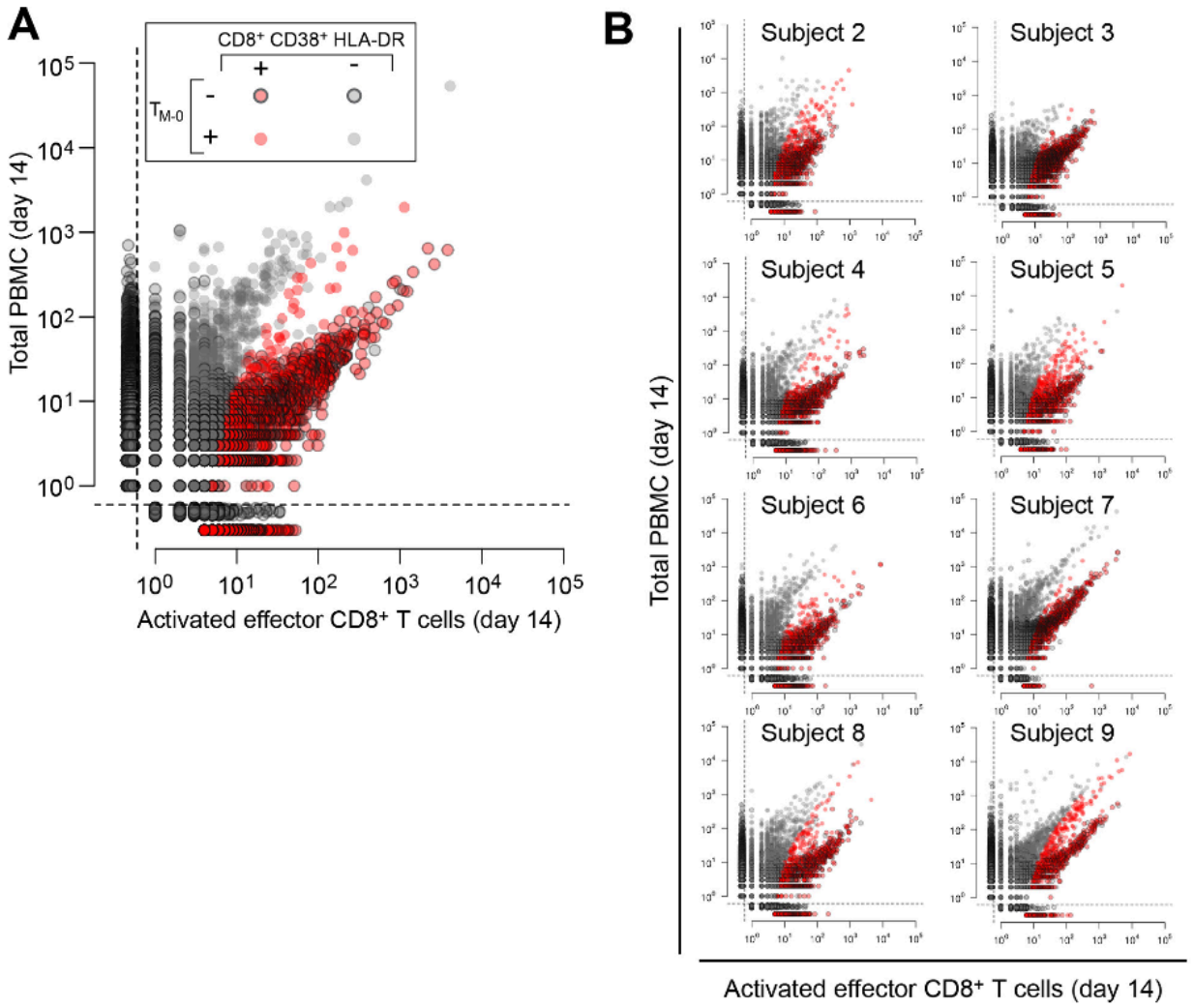


FIG 2 Identification of YFV-induced clones. The graphs show the abundance of unique clones identified by statistical enrichment on the activated effector $CD38^+$ $HLA\text{-}DR^+$ $CD8^+$ T cell compartment on day 14 postvaccination ($T_{AE\text{-}14}$) versus those present in the corresponding total PBMC sample from the same time point for subject 1 (A) and for subjects 2 to 9 (B). Clones were classified into four categories based both on their presence in the $T_{AE\text{-}14}$ and the $T_{M\text{-}0}$ compartments, as indicated in the legend. Red clones are present in the $T_{AE\text{-}14}$ compartment,

whereas gray clones are not; while clones absent in the T_{M-0} compartment have a black edge and those present in the T_{M-0} compartment do not. Darker colors indicate that multiple data points have been superimposed in that particular position. Regions bound by dashed lines indicate clones present in only one sample. YFV-induced clones were significantly enriched in the $CD38^+$ $HLA\text{-}DR^+$ $CD8^+$ T cell-sorted population compared to the corresponding total PBMC sample.

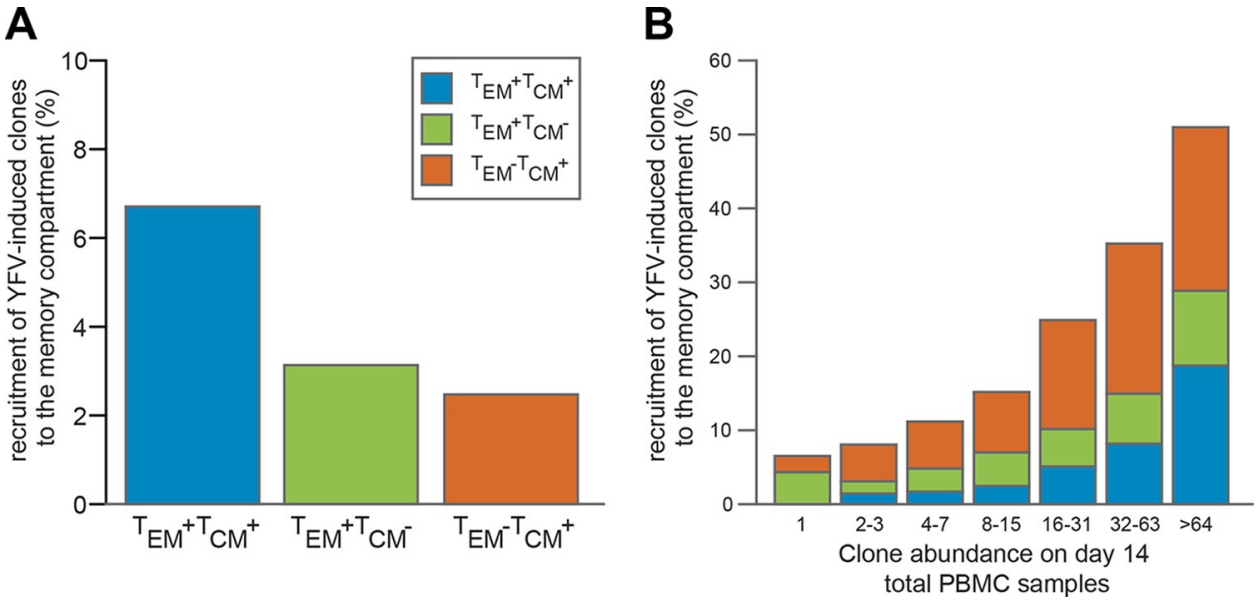


FIG 3 Recruitment of YFV-induced clones to immunological memory compartments. (A) Efficiency of recruitment of YFV-induced clones to the effector (T_{EM}^+ T_{CM}^-) and central (T_{EM}^- T_{CM}^+) memory compartments, or both (T_{EM}^+ T_{CM}^+), as a percentage of all clones classified as YFV induced. (B) Efficiency of recruitment to the effector and central memory compartments (or both) for YFV-induced clones absent from the day 0 prevaccination total PBMC samples, classified into categories based on their abundance in the day 14 postvaccination total PBMC samples. Clones with a higher degree of expansion are more efficiently recruited to the memory compartment. The aggregated data for all subjects are shown; subject-wise source data can be found in Table SI in the supplemental material.

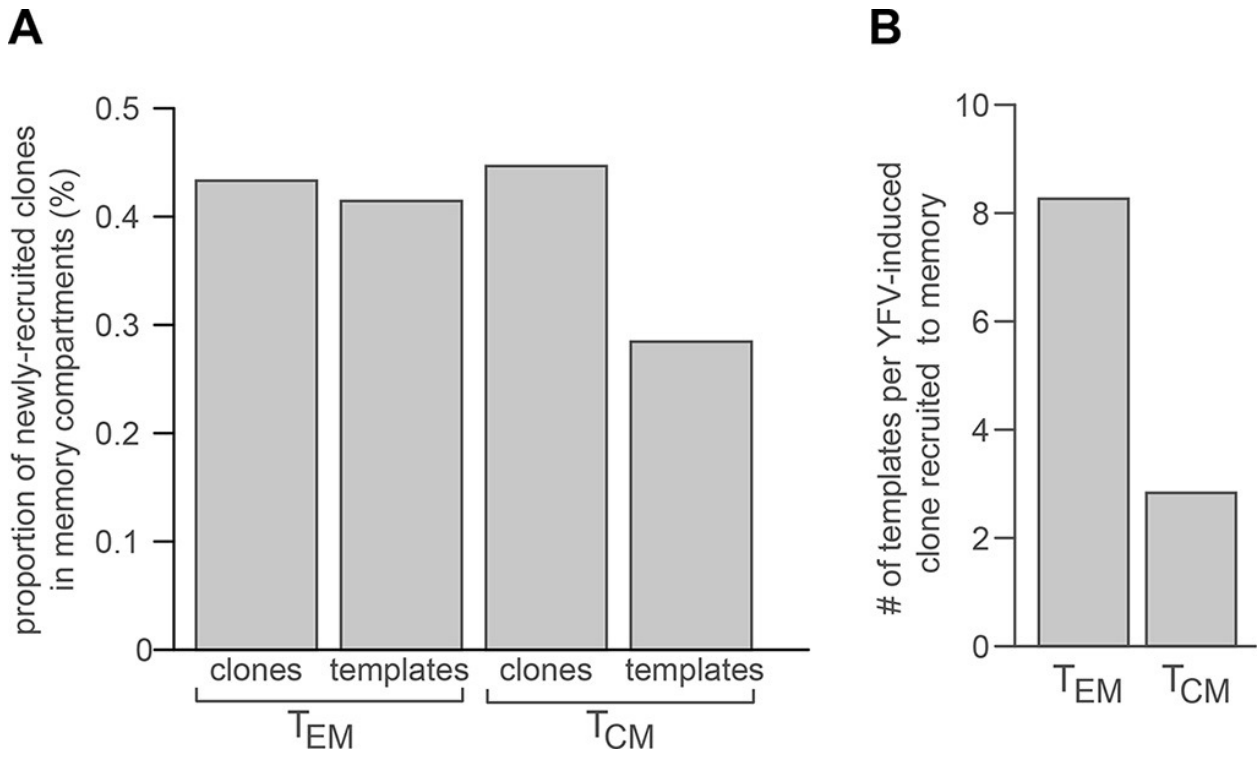


FIG 4 Composition of the effector and central memory compartments on day 90 postvaccination. (A) Proportion of YFV-induced clones newly recruited to the effector (T_{EM-90}) and central (T_{CM-90}) memory compartments on day 90 postvaccination, computed both by clone and template counts. (B) Number of templates per YFV-induced clone identified in the T_{EM-90} and T_{CM-90} memory compartments. More templates per clone were observed in the T_{EM-90} compartment, indicating that these clones were more highly expanded. The aggregated data for all subjects are shown; subject-wise source data can be found in Table SII in the supplemental material.

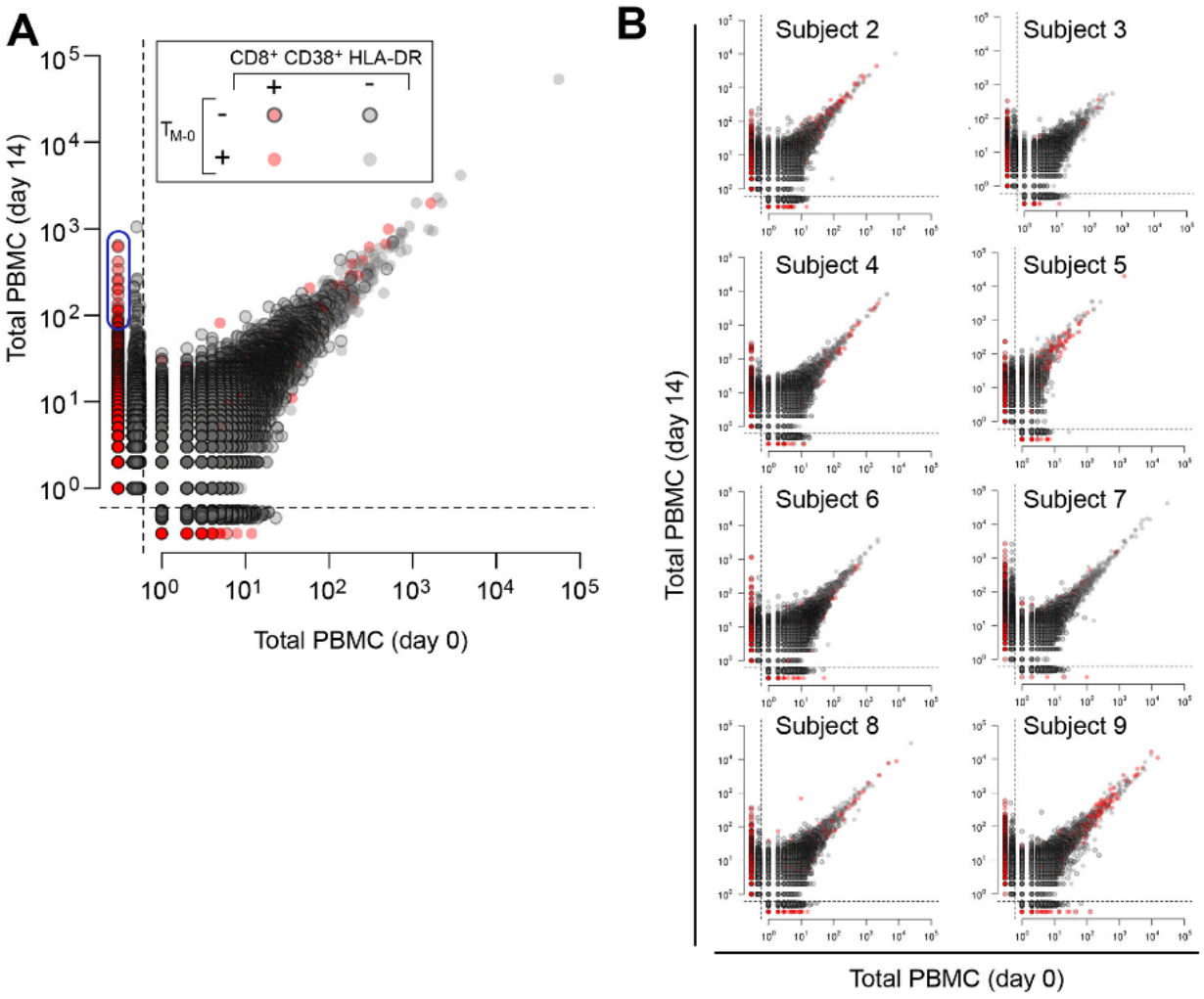


FIG 5 Identification of YFV putatively reactive clones. The graphs show the abundance of unique clones identified by statistical enrichment in the day 14 postvaccination total PBMC sample compared to the prevaccination day 0 total PBMC sample from subject 1 (A) and for subjects 2 to 9 (B). Putatively reactive clones are enclosed by a blue box. Significant enrichment (or expansion) was defined based on a *q* value threshold, with 1% and 5% expected false-positive rates for YFV-induced and putatively reactive clones, respectively (see Materials and Methods). Clones were classified into four categories based both on their presence in the T_{AE-14} and theT_{M-0} compartments, as indicated in the legend. Darker colors indicate that multiple data points are

superimposed in that particular position. Regions bound by dashed lines indicate clones present in only one sample.

TABLE 1 Experimental design[a]

Cell population	Surface markers used for sorting	No. of subjects analyzed on: Day 0	Day 14	Day 90[b]
Total PBMCs	NA	9	9	
YFV-induced effector $CD8^+$ T cells	$CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD38^+$ $HLA\text{-}DR^+$		9	
$CD8^+$ memory T cells ($T_{M\text{-}0}$)	$CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$	9		
$CD8^+$ effector memory T cells ($T_{EM\text{-}90}$)	$CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ $CD62L^{lo}$			6
$CD8^+$ central memory T cells ($T_{CM\text{-}90}$)	$CD3^+$ $CD8^+$ $CD14^-$ $CD19^-$ $CD45RA^-$ $CD45RO^+$ $CD62L^{hi}$			6

[a] Included are the cell populations studied, the surface markers used for sorting by flow cytometry, the days the samples were collected (day 0 prevaccination and days 14 and 90 postvaccination), and the number of subjects analyzed in each group. NA, not applicable.

[b] Day 90 samples from 3 subjects had to be discarded due to contamination.

TABLE 2 Number of YFV-induced clones[a]

Presence or absence in T_{M-0}	No. of YFV-induced clones in subject no.:										
	1	2	3	4	5	6	7	8	9	Avg	Total (%)
+	139	241	36	139	426	163	57	181	256	182	1,638 (8.5)
−	2,303	2,126	3,804	2,010	1,618	1,764	1,538	1,653	757	1,953	17,573 (91.5)
Total (% absent)	2,442 (94.3)	2,367 (89.8)	3,840 (99.1)	2,149 (93.5)	2,044 (79.2)	1,927 (91.5)	1,595 (96.4)	1,834 (90.1)	1,013 (74.7)	2,135 (91.5)	21,346 (82.3)

[a] For each subject, the table shows the number of YFV-induced clones present (+) or absent (−) in the memory compartment on day 0 before vaccination (T_{M-0}), as well as the total number of YFV-induced clones identified and the percentage of those that were absent from T_{M-0}. The last two columns correspond to the aggregated data (average, total, and percentage) from all 9 subjects.

TABLE 3 Number of YFV-induced clones newly recruited to the T_{CM-90} and T_{EM-90} memory compartments[a]

Memory compartment	No. of YFV-induced clones in compartment in subject no.: 2	4	5	7	8	9	Total no. (%) of clones in compartment
CM^+ EM^+	37	59	63	33	19	29	240 (2.5)
CM^+ EM^-	59	65	76	33	32	39	304 (3.1)
CM^- EM^+	107	92	162	151	76	63	651 (6.7)
CM^- EM^-	1,923	1,794	1,317	1,321	1,526	626	8,507 (87.7)
Total	2,126	2,010	1,618	1,321	1,653	757	9,702

[a] For each subject, the table shows the number of YFV-induced clones newly recruited to the T_{CM-90} and T_{EM-90} memory compartments (CM^+ EM^+), T_{CM-90} only (CM^+ EM^-), T_{EM-90} only (CM^- EM^+), or neither (CM^- EM^-), as well as the total number of clones. The last column corresponds to the aggregated data (total and percentage) from the 6 subjects for whom the memory populations were studied.

TABLE 4 Concordance between clones identified as "putatively reactive" in the total PBMC sample and YFV-induced clones identified by their presence in the activated, effector $CD8^+$ T cell compartment[a]

Presence or absence in T_{AE-14} compartment	No. of "putatively reactive" clonse in subject no.: 1	2	3	4	5	6	7	8	9	Avg	Total (%)
+	127	39	118	36	3	44	288	74	119	94.2	848 (62.2)
−	106	20	6	3	9	56	190	38	87	57.2	515 (37.8)
Total (% present in T_{AE-14})	233 (54.5)	59 (66.1)	124 (95.2)	39 (92.3)	12 (25.0)	100 (44.0)	478 (60.3)	112 (66.1)	206 (57.8)	151.4 (62.4)	1,363 (62.2)

[a] For each subject, the table shows the number of "putatively reactive" clones identified in the total PBMC sample that were present (+) or absent (−) in the corresponding T_{AE-14} compartment, as well as the total number of putatively reactive clones for each subject and the percentage of them present in T_{AE-14}. The last two columns correspond to the aggregated data (average, total, and percentage) from all 9 subjects.

4.7 Supplementary Material

		Abundance						
		1	**2-3**	**4-7**	**8-15**	**16-31**	**32-63**	**≥64**
Subject 2	**CM^-EM^-**	14	346	385	231	67	17	6
	CM^+EM^+	0	4	8	5	6	4	2
	CM^-EM^+	0	10	13	10	7	2	0
	CM^+EM^-	0	16	18	20	15	4	1
Subject 4	**CM^-EM^-**	7	328	304	190	53	11	2
	CM^+EM^+	0	6	8	6	10	8	8
	CM^-EM^+	0	4	16	14	5	3	0
	CM^+EM^-	0	16	24	18	3	2	1
Subject 5	**CM^-EM^-**	9	172	186	114	46	10	0
	CM^+EM^+	0	8	6	8	6	2	3
	CM^-EM^+	2	5	11	15	9	0	1
	CM^+EM^-	1	21	28	21	16	5	2
Subject 7	**CM^-EM^-**	1	131	168	274	277	99	76
	CM^+EM^+	0	0	0	3	8	1	14
	CM^-EM^+	0	0	2	4	9	6	8
	CM^+EM^-	0	3	10	18	51	31	27
Subject 8	**CM^-EM^-**	11	227	219	133	61	17	7
	CM^+EM^+	0	1	2	4	5	0	1
	CM^-EM^+	0	3	3	6	2	3	4
	CM^+EM^-	0	9	8	13	8	6	5
Subject 9	**CM^-EM^-**	1	74	98	88	42	28	11
	CM^+EM^+	0	1	2	4	2	8	11
	CM^-EM^+	0	1	3	6	5	5	8
	CM^+EM^-	0	4	9	9	14	9	10

Supplemental Table SI: Number of YFV-induced clones absent on the T_{M-0} compartment, classified based on their recruitment to the T_{CM-90} and T_{EM-90} compartments as well as the level of expansion, measured by their abundance on the day 14 post-vaccination total PBMC samples.

		clone counts		template counts	
		new YFV-induced	not YFV - induced	new YFV-induced	not YFV - induced
Subject 2	EM	96	34,584	668	173,024
	CM	144	39,683	245	119,443
Subject 4	EM	124	20,309	431	159,933
	CM	151	37,911	321	148,867
Subject 5	EM	139	30,260	327	132,558
	CM	225	51,247	769	136,699
Subject 7	EM	66	27,295	117	171,953
	CM	184	33,914	375	102,844
Subject 8	EM	51	11,485	135	104,340
	CM	95	18,658	563	131,258
Subject 9	EM	68	14,163	2818	119,083
	CM	92	27,776	251	93,823

Supplemental Table SII. Composition of the day 90 memory compartment. Shown are the number of new, YFV-induced clones contributing to the T_{EM-90} and T_{CM-90} memory compartmentss as compared to the non-YFV-induced clones, counted both by number of clones and by number of templates.

4.8 Notes

Acknowledgements

We thank Rafi Ahmed and Rama Akondy for sharing their flow cytometry sorting protocols. H.R. thanks Melissa Alvendia for coordinating the study, and W.D. thanks Bryan Howie for helpful discussions.

This work was funded in part by grant R56 AI0181860 from the NIH to H.R.

H.R. owns stock and receives consulting fees from Adaptive Biotechnologies. W.S.D., R.O.E., M.V., T.M.S., C.D., and C.S. are employees of Adaptive Biotechnologies with salary and stock options.

This work was published in the *Journal of Virology* as

DeWitt WS*, Emerson RO*, Lindau P, Vignali M, Snyder TM, Desmarais C, Sanders C, Utsugi H, Warren EH, McElrath J, Makar KW, Wald A, Robins HS. Dynamics of the cytotoxic T cell response to a model of acute viral infection. J Virol. 2015 Apr 15;89(8):4517-26.

5 CMV Infection in the Elderly Does Not Reduce Cytotoxic T cell Repertoire Diversity

5.1 Abstract

With age, the immune system becomes less effective, causing increased susceptibility to infection. Chronic cytomegalovirus (CMV) infection further impairs immune function and is associated with increased mortality in the elderly. CMV exposure elicits massive $CD8^+$ T cell clonal expansions and diminishes the cytotoxic T cell response to subsequent infections, leading to the hypothesis that, in order to maintain homeostasis, clones are expelled from the repertoire, reducing T cell repertoire diversity and diminishing the ability to combat new infections in the CMV-seropositive elderly. However, the impact of CMV infection on the structure and diversity of the underlying T cell repertoire remains uncharacterized. Here we show that the T cell repertoire in the elderly grows to accommodate CMV-driven clonal expansions while preserving its underlying diversity and structure. Using T cell receptor β chain immunosequencing, we observed that the proportion of the peripheral blood T cell repertoire occupied by the most frequent 0.1% of clones is larger in the CMV seropositive across a wide range of ages. Furthermore, we found that in elderly CMV+ individuals, the most frequent clonotypes are CMV specific and comprised the majority of the $CD8^+$ memory T cell repertoire. We also discovered that the structure and diversity of the naïve T cell repertoire was similar in subjects with and without CMV. Our observations suggest that a lifetime of CMV-driven T cell clonal expansions does not compromise the underlying repertoire. Alternatively, we propose that

the diminished immunity in elderly individuals with CMV is due to a decline in cellular function rather than a reduction in cytotoxic T cell repertoire diversity.

5.2 Introduction

As we age, immune function declines, a phenomenon known as immunosenescence. Large-scale changes in both the innate and adaptive immune system enhance susceptibility to infections and diminish responsiveness to vaccines, leading to increased morbidity and mortality (1-4). Many of these changes are exacerbated by pathogens that lead to chronic or persistent infections like cytomegalovirus (CMV) (4-6). CMV is a widely prevalent Herpesvirus that causes latent infections with phases of subclinical reactivation. In the elderly, CMV seropositivity has been classified as an immune risk phenotype (7), and directly linked with increased mortality (8).

Over time, massive CMV-driven $CD8^+$ T cell clonal expansions are thought to compound a decline in immune function (9, 10). CMV-specific memory T cells terminally differentiate into T Effector Memory cells expressing CD45RA (T_{EMRA}), which have limited proliferative potential and resistance to apoptosis (5, 11). These cells possess a late-differentiated antigen-experienced phenotype that does not undergo replicative senescence due to repeated stimulation (5, 12). The accumulation of apoptosis-resistant T_{EMRA} clones in the CMV seropositive elderly is believed to compromise T cell repertoire diversity (13-15).

T cell repertoire diversity is defined as the number, frequency and distribution of clones within the T cell repertoire, and its reduction has been shown to decrease the breadth of the immune response against a wide spectrum of epitopes (16, 17). In CMV, massive $CD8^+$ T cell clonal expansions are thought to result in the expulsion of T cell

clones from the repertoire, thus reducing overall diversity (4, 16). This loss of T cell clones, combined with an age-related decline in polyfunctional T cell responses suggest a mechanism for the increased mortality observed amongst the CMV seropositive elderly (2, 15, 18). However, it is important to note that previous methods -including V-J tracking and spectratyping- lacked the sensitivity and specificity to interrogate the underlying T cell repertoire in CMV (9, 16, 19-21). To gain insights into the nature of the entire cytotoxic $CD8^+$ T cell repertoire in the natural setting of immune aging and chronic stimulation by CMV, we combine flow cytometry and immunosequencing of the TCRβ chain of the TCR receptor as a measure of the diversity of the T cell repertoire.

To characterize the effects of aging and CMV on the T cell repertoire, we surveyed millions of T cell clones across a broad age range and observed that a handful of clones dominate the repertoire in CMV seropositive aging individuals. When we specifically examined the $CD8^+$ T cell repertoires of CMV seropositive elderly, we found that the most frequent 0.1% of peripheral blood clones comprise the majority of classical antigen-experienced $CD45RO^+$ memory T cells and CD45RA-revertant T_{EMRA} compartments. We were able to examine in detail the impact of CMV on the structure of the underlying repertoire of these elderly individuals and failed to find evidence of compromised repertoire diversity in the presence of CMV-induced clonal expansions. Overall, our data suggests that the space occupied by CMV-specific clones grows considerably in the CMV seropositive elderly without affecting the rest of the repertoire, and that the repertoire broadens to accommodate these large clonal expansions.

5.3 Material and Methods

Experimental cohort and study approval

For the survey cohort of 553 donors, peripheral blood samples were obtained from the Fred Hutchinson Cancer Research Center Research Cell Bank biorepository of healthy bone marrow donors. These samples were HLA-typed and tested for CMV serostatus. For the aged cohort of 8 donors, fresh peripheral blood samples were obtained from the Fred Hutchinson Cancer Research Center Prevention Center. The complete blood count (CBC), HLA type and CMV and EBV serostatus were obtained for each sample. For both cohorts, donor protocols were approved and supervised by the Fred Hutchinson Cancer Research Center Institutional Review Board.

Cell sorting

Cells were kept at 4˚C throughout the entire process of enrichment, labeling and sorting. $CD3^+$ T cells were enriched from peripheral blood mononuclear cells by immunomagnetic selection using CD3 MicroBeads (Miltenyl Biotec, Auburn, CA). Cells were stained in the dark for 15 minutes with the following anti-human antibodies: CD45RO PE-Cy7 (BD Biosciences, San Jose, CA), CD3 AlexaFluor700 (BD Biosciences, San Jose, CA), CD62L-PE (BD Biosciences, San Jose, CA), CD45RA-APC (BD Biosciences, San Jose, CA), CD8-Pacific Blue (BD Biosciences, San Jose, CA), CD4 APC-Cy7 (BD Biosciences, San Jose, CA) and LIVE/DEAD Aqua fluorescent reactive dye (Invitrogen, Grand Island, NY). $CD8^+$ T cell subsets were isolated using the BD FACSAria cell-sorting system (BD Biosciences), including $CD8^+$ $CD45RA^-$ $CD45RO^+$ (for $CD8^+$ Memory), $CD8^+$ $CD45RA^+$ $CD45RO^-$ $CD62L^{hi}$ ($CD8^+$ naïve), and $CD8^+$ $CD45RA^+$ $CD45RO^-$ $CD62L^{lo}$ ($CD8^+$ T_{EMRA}).

FlowJo (TreeStar Inc, Ashland, OR) analysis was used to determine the proportions of the different subsets as a fraction of total $CD8^+$ T cells.

Immunosequencing

For the survey cohort, genomic DNA was extracted from peripheral blood samples using the Qiagen DNeasy Blood Extraction kit (Qiagen). An average of 2.5 µg of input DNA was used for each sample. For the aged cohort, total genomic DNA was extracted from sorted T cells using the QIAamp DNA Blood Mini Kit (Qiagen). At least 3.2 µg of input DNA was used for sequencing each population. For all samples, the CDR3 region of the rearranged TCRβ locus, defined according to IMGT, was amplified and sequenced using previously described protocols (22). Raw sequence data were preprocessed to remove errors in the primary sequence of each read, and to compress the data. A nearest-neighbor algorithm was used to collapse the data into unique sequences in order to remove PCR and sequencing errors.

CMV stimulation

RV798 CMV-infected fibroblasts (23) from subjects 4 and 5 were used to stimulate autologous sort-purified $CD8^+CD45RA^-CD45RO^+$ memory and $CD8^+CD45RA^+CD45RO^-CD62L^{lo}$ T_{EMRA} cells. CMV-reactive T cells from each stimulated $CD8^+$ T cell subset were sorted as $CD8^+CD137^+$ events (24). Memory and T_{EMRA} $CD8^+$ T cells stimulated with uninfected fibroblasts were used as a negative control for CD137 expression.

Repertoire diversity and clonality metrics

The Shannon entropy or diversity (*H*) is an index that combines measurements of species richness and abundance. For a sample with richness *S* and clone-wise population fractions given by $\pi_1, \pi_2, \ldots, \pi_S$, the Shannon diversity is defined the entropy of the clone-

$$H = -\sum_{i=1}^{S} \pi_i \log \pi_i.$$

wise abundance distribution. The Shannon entropy favors neither rare nor dominant clones disproportionately because each clone is weighted by its frequency in the sample. Clonality describes the degree to which expanded clones dominate the repertoire. The Shannon equitability (E_H) is defined as $E_H = H / H_O$, where H_O is the maximum entropy, $H_O = \log N$. Clonality is defined as $(1 - E_H)$ with larger values indicating more oligoclonal repertoires.

5.4 Results

CMV Exacerbates Large Clonal Expansions with Age

To survey the impact of age and chronic CMV infection on the accumulation of high-frequency T cell clones, we examined productively rearranged TCRβ DNA sequences from a previous study of 553 healthy CMV seropositive (CMV^+) and CMV seronegative (CMV^-) subjects (**see Methods, Table S1**). Consistent with previous studies showing that repeated CMV stimulation induces large T cell clonal expansions (25, 26), we found that the most frequent 0.1% of clonotypes (~200) comprised a much greater proportion of the T cell repertoire in CMV^+ individuals across age groups (**Fig. 1**). In addition, we observed a gradual increase in the proportion of the repertoire dedicated to these high-frequency clonotypes with age. Ultimately, we found that, on average, the most frequent

0.1% of clonotypes in the oldest age group constituted nearly 30% of the peripheral blood cell repertoire in CMV^+ subjects.

Large Clones Dominate the Memory Repertoires of the Elderly

To further investigate the effect of large clonal expansions on the T cell repertoire of the elderly, we recruited 5 CMV^+ and 3 CMV^- subjects between the ages of 70 and 74. We then isolated PBMCs, as well $CD4^+$, $CD8^+$ naïve, $CD8^+$ memory and $CD8^+$ T_{EMRA} T cell subsets and examined rearranged TCRβ DNA sequences in these samples (**see Methods, Table S2**). As with the larger cohort, we observed that the most frequent 0.1% of clonotypes occupied a larger fraction of the peripheral blood T cell repertoire in CMV^+ subjects (**fig. S1**). We then searched the different T cell subsets to determine the distribution of the most frequent 0.1% of PBMC clonotypes in the repertoire. In order to minimize the effect of contamination during sorting, we bioinformatically removed from the naïve repertoire high-frequency PBMC clonotypes that were also present in memory and T_{EMRA} samples from the same subject. Between 63 and 115 clonotypes were removed from the naïve repertoire of the different subjects. Unexpectedly, the fraction of the naïve repertoire occupied by these clonotypes did not positively correlate with the purity of the sort (**fig. S2A, B**). This observation could be accounted for by our use of CD62L as opposed to CCR7 to divide the T cell subsets; however, CD62L was chosen because CMV-reactive clones are mostly $CD62L^-$. We found that the vast majority of the most frequent 0.1% of PBMC clonotypes resided in the $CD4^+$, $CD8^+$ memory and $CD8^+$ T_{EMRA} repertoires (**Fig. 2A**). Furthermore, 36.3% to 85.1% of these large PBMC clonotypes resided in both $CD4^+$ and $CD8^+$ T cell populations. When we compared CMV^+

and CMV$^-$ subjects, we observed that the most frequent PBMC clonotypes were similarly distributed amongst the different T cell subsets. In addition, as a proportion, these high-frequency PBMC clonotypes constituted the majority of the memory and T_{EMRA} repertoires in CMV$^+$ individuals (**Fig. 2B**). These results are consistent with previous studies demonstrating that CMV-fueled clonal expansions occur within the memory subsets, and especially in the T_{EMRA} subset (27, 28).

High-frequency Clonotypes are CMV-Reactive

To determine the proportion of the memory and T_{EMRA} repertoires devoted to CMV, we isolated skin fibroblasts derived from CMV$^+$ subjects 4 and 5 and infected them with CMV (**see Methods**). Aliquots of memory and T_{EMRA} cells were then stimulated with these fibroblasts and activated T cells were sorted and sequenced. When we examined the frequency of CMV reactive clonotypes in each subject, we observed that, in both memory and T_{EMRA} subsets, the highest frequency clonotypes were CMV-reactive (**Fig. 3A, B**). The majority of these high-frequency CMV-reactive clonotypes were present in both the memory and T_{EMRA} subsets. Some clonotypes identified as CMV-reactive in the memory subset were found in the CD137$^-$ T_{EMRA} population confirming that T_{EMRA} cells contain CMV-reactive clones with limited proliferative potential and resistance to apoptosis (29) (**Fig. 3B**). Therefore, we combined CD137$^+$ memory and CD137$^+$ T_{EMRA} clonotypes to determine the proportion of PBMC, memory and T_{EMRA} subsets allocated to CMV (**Fig. 3C**). We found that, in the two subjects studied in this manner, the ten most frequent CMV-reactive memory clonotypes comprised 34.4% and 45.9% of the memory repertoire, while the ten largest CMV-reactive T_{EMRA} clonotypes made up 82.8% and 62.3% of the

T_{EMRA} repertoire. We also observed that the proportion of the peripheral blood repertoire dedicated to CMV is very similar to the proportion of the repertoire occupied by the most frequent 0.1% of T cell clonotypes. Notably, 37.4% and 42.0% of the most frequent 0.1% of peripheral blood clonotypes were found to be CMV-reactive in each of the subjects. However, as a proportion, CMV-reactive clonotypes accounted for 75.4% and 78.0% of this subpopulation of high-frequency peripheral blood clonotypes. Together, these results demonstrate that the highest-frequency clonotypes in the peripheral blood T cell repertoire are CMV-reactive, and that they reside in the memory and T_{EMRA} subsets.

CMV Does Not Reduce Naïve Repertoire Diversity

Considering that a significant proportion of the repertoire is dedicated to CMV, we next sought to determine the effect of these high-frequency clonotypes on the underlying T cell repertoire. We calculated the Shannon diversity and clonality metrics for each T cell subset under study (**see Methods, Fig. 4A, B**). In order to remove the possible confounding effect of the purity of the cell sort, we removed all memory and T_{EMRA} TCRβ sequences found in the most frequent 0.1% of PBMC clonotypes from the naïve repertoires of each participant. As expected, we observed that the $CD4^+$, $CD8^+$ naïve and PBMC repertoires of all subjects are significantly more diverse as compared to the memory and T_{EMRA} subsets from the same individuals. In addition, we found that the PBMC, memory and T_{EMRA} repertoires of CMV^+ individuals appear to be less diverse and more clonal than those of CMV^- subjects. Significantly, we could not detect a difference in the diversity of the naïve repertoires based on CMV serostatus (**Fig. 4A**). To determine whether CMV altered the structural characteristics of the naïve repertoire, we examined

the distribution of low-frequency clonotypes in the naïve repertoires of each subject. We found nearly identical clone frequency distributions between CMV^+ and CMV^- individuals (**Fig. 4C**). Although naïve T cells make up a smaller fraction of the total $CD8^+$ T cell population in CMV^+ subjects (**Table S2**), the overall structure of their repertoire appears to remain unmodified.

The $CD8^+$ Repertoire Expands to Accommodate Large Clones

To assess whether the differences in PBMC, memory and T_{EMRA} repertoire diversities were the result of large CMV driven clonal expansions, we removed the most frequent 0.1% of PBMC clonotypes from the PBMC, memory and T_{EMRA} repertoires of each subject. When we recalculated the Shannon diversity index, we found that the PBMC and memory repertoire diversities were indistinguishable based on CMV serostatus (**Fig. 5A**). In contrast, the T_{EMRA} repertoire in 3/5 CMV^+ subjects remained less diverse compared to CMV seronegative individuals. Nevertheless, removing the largest clonotypes increased repertoire diversity in all subjects suggesting that large clonal expansions do not dramatically affect the composition of the rest of the repertoire.

In order to reconcile the observation that very high frequency CMV-reactive clonotypes dominate the repertoire of elderly, CMV^+ subjects without altering its underlying diversity and structure, we obtained diagnostic-quality T cell counts from 6 of the subjects in this study. We then used relative frequencies of $CD8^+$ T cell subsets obtained using flow cytometry to calculate the number of cells present in each subset (**Fig. 5B**). Interestingly, we observed that the $CD8^+$ T cell repertoire broadens to accommodate large, CMV-fueled clonal expansions. This observation contrasts with

conclusions from previous studies suggesting that large clonal expansions expel smaller clones from the repertoire in order to maintain homeostasis. Instead, our data suggests that the number of cells in the T cell repertoire increases over the lifespan of a subject in order to accommodate new antigen exposures as well as the continued expansion of high-frequency clonotypes that react to chronic viral infections.

5.5 Discussion

In this study, we define the effect of CMV on the diversity and clonal structure of the aging immune system. Over time, repeated stimulation with latent CMV antigens leads to the accumulation of large CMV-specific clones in the circulating T cell repertoire (4, 16). Consistent with previous studies, we found that the proportion of the repertoire occupied by the most frequent 0.1% of clones in PBMCs dramatically increases with age in CMV seropositive individuals. This demonstrates the substantial burden that CMV infection places on the aging adaptive immune.

To explore the influence of CMV on the structure of the T cell repertoire in greater detail, we examined rearranged TCRβ DNA sequences derived from the $CD8^+$ T cell subsets implicated in suppressing CMV reactivation in a group of elderly CMV^+ and CMV^- subjects. We found that the most frequent 0.1% of peripheral blood clones occupy a significant proportion of the memory and T_{EMRA} subsets in the CMV^+ elderly, which is consistent with a previous study by Paul Moss' group (28). Although we observed the greatest clonal expansions in the memory and T_{EMRA} subsets, many clonotypes were shared between the $CD4^+$ and $CD8^+$ lineages. Nevertheless, these clonotypes were found at much lower frequencies in the $CD4^+$ T cell population, further demonstrating that

a significant portion of the entire T cell repertoire is dedicated to CMV in elderly CMV^+ subjects.

Given that a significant proportion of the memory and T_{EMRA} repertoires were composed of high-frequency clones in the CMV^+ elderly, we next sought to determine whether these clones were CMV reactive. Thus, we determined the proportion of the memory and T_{EMRA} repertoire that is dedicated to CMV in two elderly subjects (25, 26). We found that, in both subjects, the highest frequency CMV-reactive clonotypes were shared among antigen-experienced T cell subsets. Significantly, we observed that some of the highest frequency shared clones in the T_{EMRA} subset were unresponsive to CMV. Our results confirm previous reports that T_{EMRA} cells are not clonally deleted upon replicative senescence (30). Importantly, we found that the highest-frequency clones in the peripheral blood repertoire are CMV-reactive. We suspect that some of the remaining high-frequency clonotypes that are not CMV-specific could recognize Epstein Barr virus (EBV), given that all participants were EBV seropositive. Nevertheless, our results demonstrate that CMV-reactive T cell clones expand to dominate the overall T cell repertoire.

Previous studies have suggested that naïve T cell clones are eradicated from the repertoire in order to accommodate the large clonal expansions observed in the CMV seropositive elderly. However, we found no difference in the diversity of the naïve T cell repertoire or in the distribution of low-frequency clonotypes in the naïve repertoire based on CMV serostatus. Our observations suggest that, from the standpoint of the $CD8^+$ repertoire, CMV-driven clones expand without altering the rest of the repertoire (**Fig. 5 C, D, E**). Given our depth of sampling, the fact that we could not detect a compensatory

shrinking of the repertoire in the presence of CMV-induced expansions implies that this phenomenon occurs either rarely or not at all. Nevertheless, when we bioinformatically removed the most frequent 0.1% of peripheral blood clones from the PBMC, memory and T_{EMRA} repertoires, we observed an increase in diversity, and in fact the diversity of these modified repertoires was highly similar to that observed in the repertoires of CMV^- subjects. This suggests that the expansion of high-frequency, shared memory and T_{EMRA} clones may be a general phenomenon of aging in the context of a latent viral infection. In support of this claim, we observed that the total number of T cells in each $CD8^+$ T cell subset is increased in CMV^+ individuals. Altogether, these results demonstrate that the T cell repertoire grows to accommodate the increased clonal expansions due to CMV.

Although CMV infection alters the T cell repertoire across a variety of ages, CMV is not associated with increased mortality until the late stages of life. We note that while our inferences are based on a few participants, the differences in repertoire structure based on CMV status in the elderly are quite unremarkable. Thus, our observations merit further research into the cellular mechanisms of chronic immune responses that perpetuate immunosenescence.

5.6 Figures and Tables

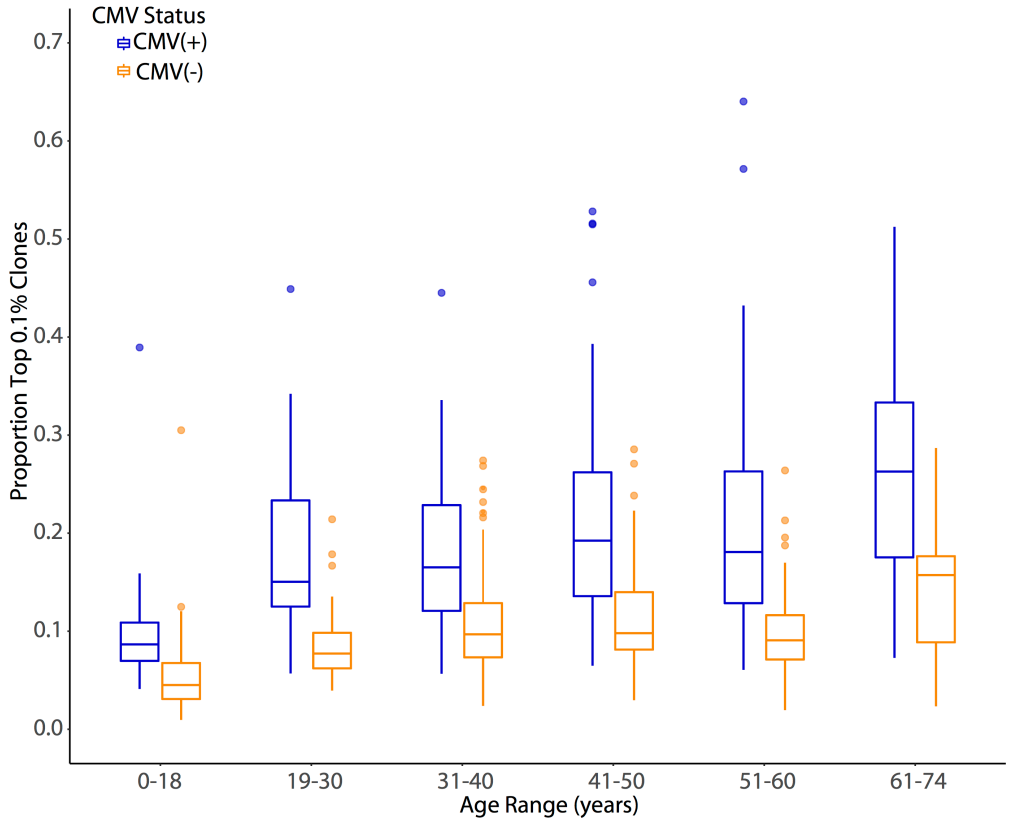


Fig.1. **Impact of CMV on the proportion of high-frequency clonotypes with age**. Boxplot comparing the proportion of the most frequent 0.1% of clonotypes in the peripheral blood T cell repertoire of CMV^+ (blue) and CMV^- (orange) subjects. TCRβ chain sequencing was performed on the PBMCs of 553 subjects across a wide range of ages. Clonotypes with productive TCRβ rearrangements were ranked based on frequency. The cumulative abundance of the most frequent 0.1% of clonotypes was divided by the total sample abundance to yield a proportion. The band inside each box represents the median and the whiskers extend to values that are within 1.5 * interquartile range. Outliers are represented by dots.

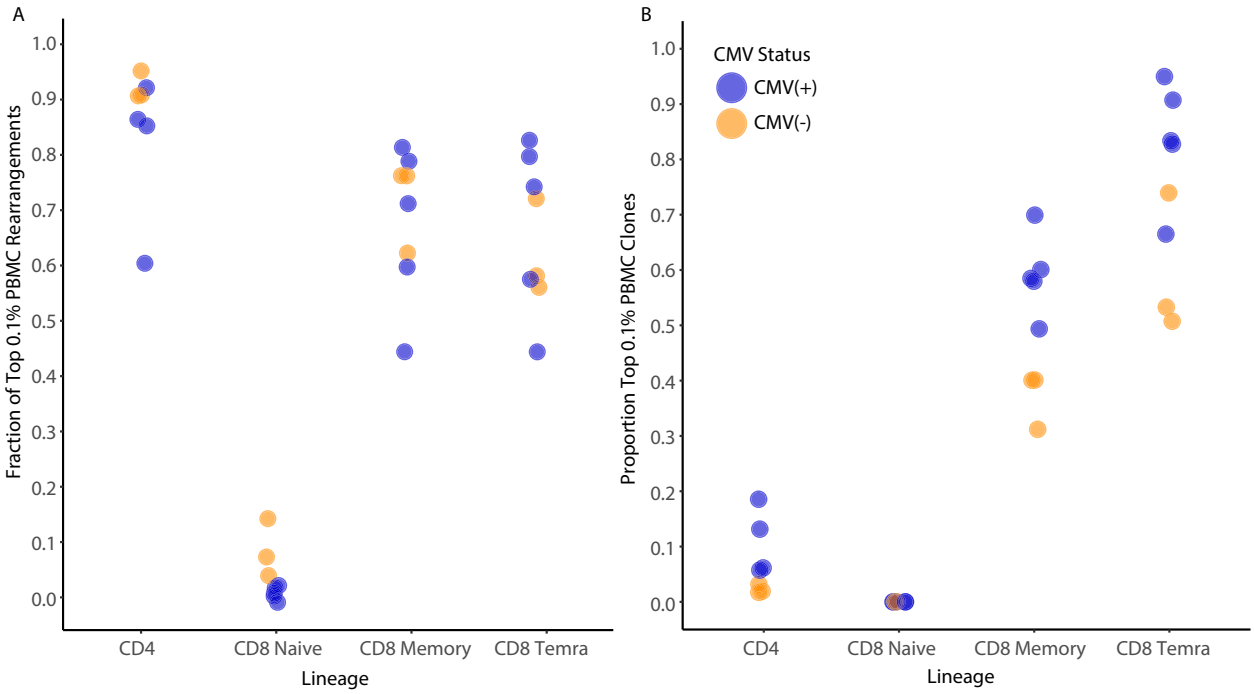


Fig. 2. **Distribution of high-frequency PBMC clonotypes in the elderly**. (**A**) Scatterplot comparing the fraction of the most frequent 0.1% of PBMC clonotypes found in each sorted T cell subset in CMV$^+$ (blue) and CMV$^-$ (orange) subjects. TCRβ sequencing was performed on the PBMCs and T cell subsets of 8 subjects greater than 70 years old. Clonotypes in the PBMC sample with productive TCRβ rearrangements were ranked based on frequency. The fraction of the most frequent 0.1% of PMBC clonotypes present in each T cell subset is depicted. (**B**) Comparison of the proportion of each sorted T cell subset composed of the most frequent 0.1% of PBMC clonotypes in CMV$^+$ and CMV$^-$ subjects. The cumulative abundance of the most frequent 0.1% of PBMC clonotypes present in each T cell subset was divided by the total abundance of each subset to yield a proportion. High-frequency PBMC clonotypes found in both naïve and memory samples were bioinformatically removed from the naïve repertoire. CD4, Bulk CD4$^+$ T cells; CD8 Naïve, CD8$^+$ naïve T cells; CD8 Memory, CD8$^+$ central and effector Memory T cells; CD8 Temra, CD8$^+$ memory T cells expressing CD45RA.

Figure 3

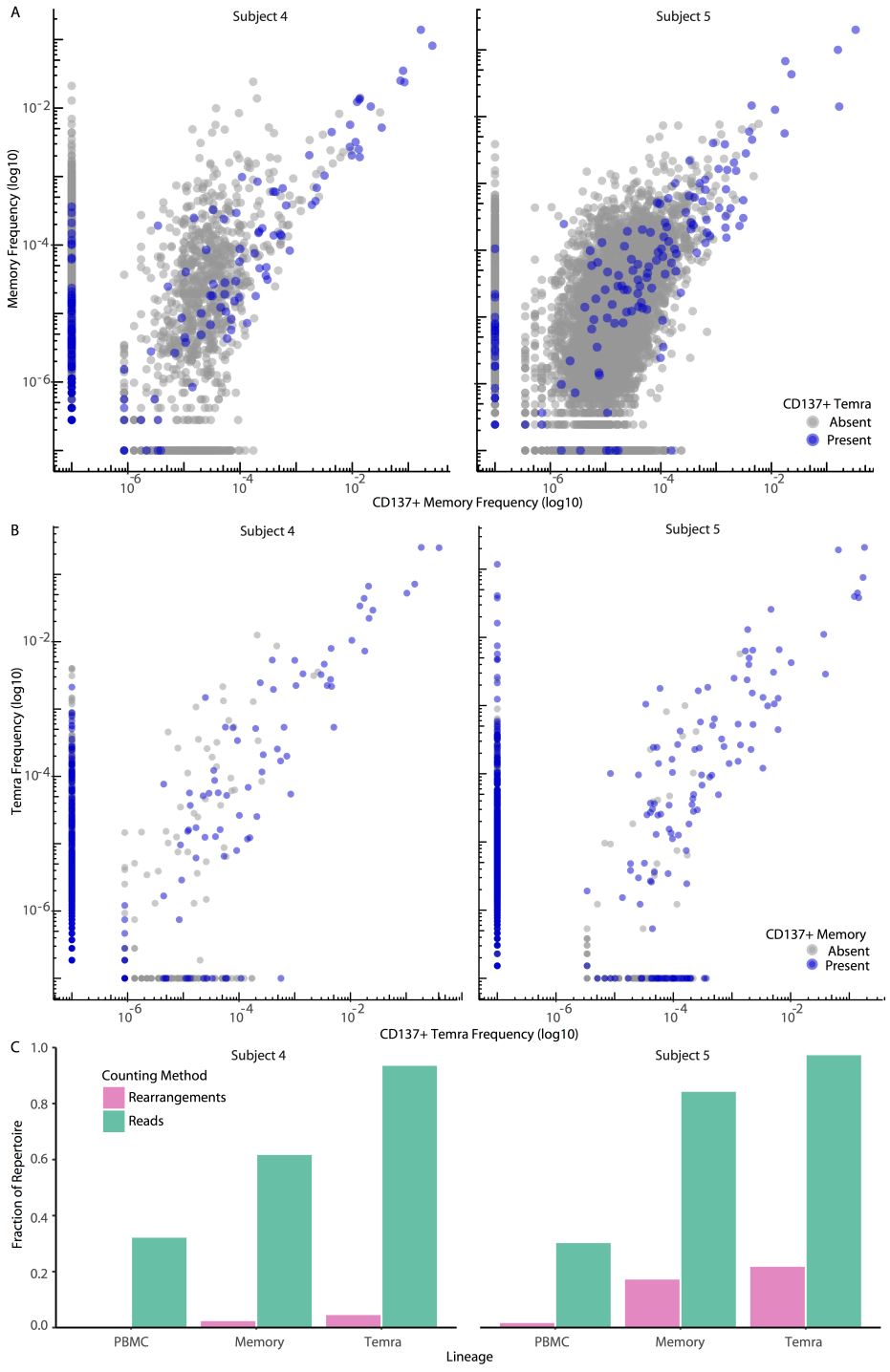


Figure. 3. **Identification of CMV-reactive T cell clonotypes in the elderly**. (**A, B**) Scatterplot comparing clonotype frequencies in CMV-stimulated CD137^{+} and resting memory (**A**) or T_{EMRA}

(**B**) subsets from CMV^+ subjects 4 and 5. $CD45RO^+$ memory and $CD45RA^+$ T_{EMRA} T cells were sorted and stimulated with autologous CMV infected fibroblasts for 24 hours. $CD137^+$ T cells were then sorted and TCRβ was performed. The frequency of productive TCRβ sequences from unstimulated memory and T_{EMRA} samples are plotted against the frequency productive TCRβ sequences from the corresponding $CD137^+$ sample. Each point represents clonotype. Points along the axis represent clonotypes present in one sample. Points colored blue in **(A)** represent clonotypes also present in the $CD137^+$ T_{EMRA} sample and in **(B)** represent clonotypes also present in the $CD137^+$ memory T cell sample. Logarithmic scale, base-10. (**C**) Comparison of the fraction of each T cell subset as unique rearrangements (pink) or reads (green) composed of $CD137^+$ CMV-reactive clonotypes. Memory and T_{EMRA} $CD137^+$ TCRβ sequences were combined to capture all CMV-reactive clonotypes. Unstimulated PBMC, memory and T_{EMRA} samples were then searched for these sequences. Pink represents the fraction of unique CMV-reactive clonotypes found in each unstimulated sample. Green represents the cumulative abundance CMV-reactive clonotypes found in each unstimulated sample divided by the total abundance of the sample. Memory, $CD8^+$ central and effector memory; Temra, $CD8^+$ memory T cells expressing CD45RA.

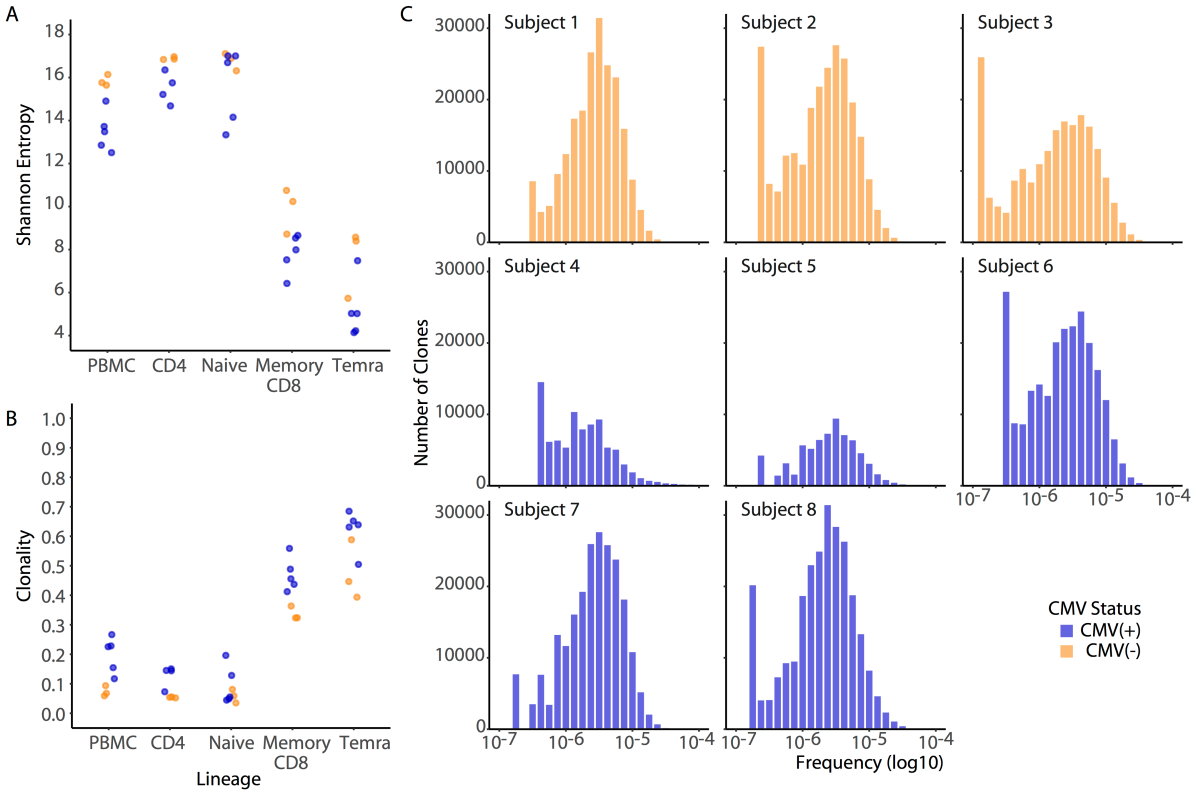


Fig. 4. **Effect of CMV on the underlying T cell repertoire**. (**A, B**) Scatterplot comparing the Shannon entropy (**A**) or clonality (**B**) of each T cell subset in CMV^+ (blue) and CMV^- (orange) subjects. The most frequent 0.1% of PBMC clonotypes found in both naïve and memory samples were bioinformatically removed from naïve T cell entropy and clonality calculations. (**C**) Histogram comparing the frequency distribution of naïve T cell clonotypes in CMV^+ and CMV^- subjects. Each bar represents the total number of unique clonotypes present at a particular frequency. Naïve T cell clonotypes with frequencies greater than 10^{-4} were removed. Logarithmic scale, base-10.

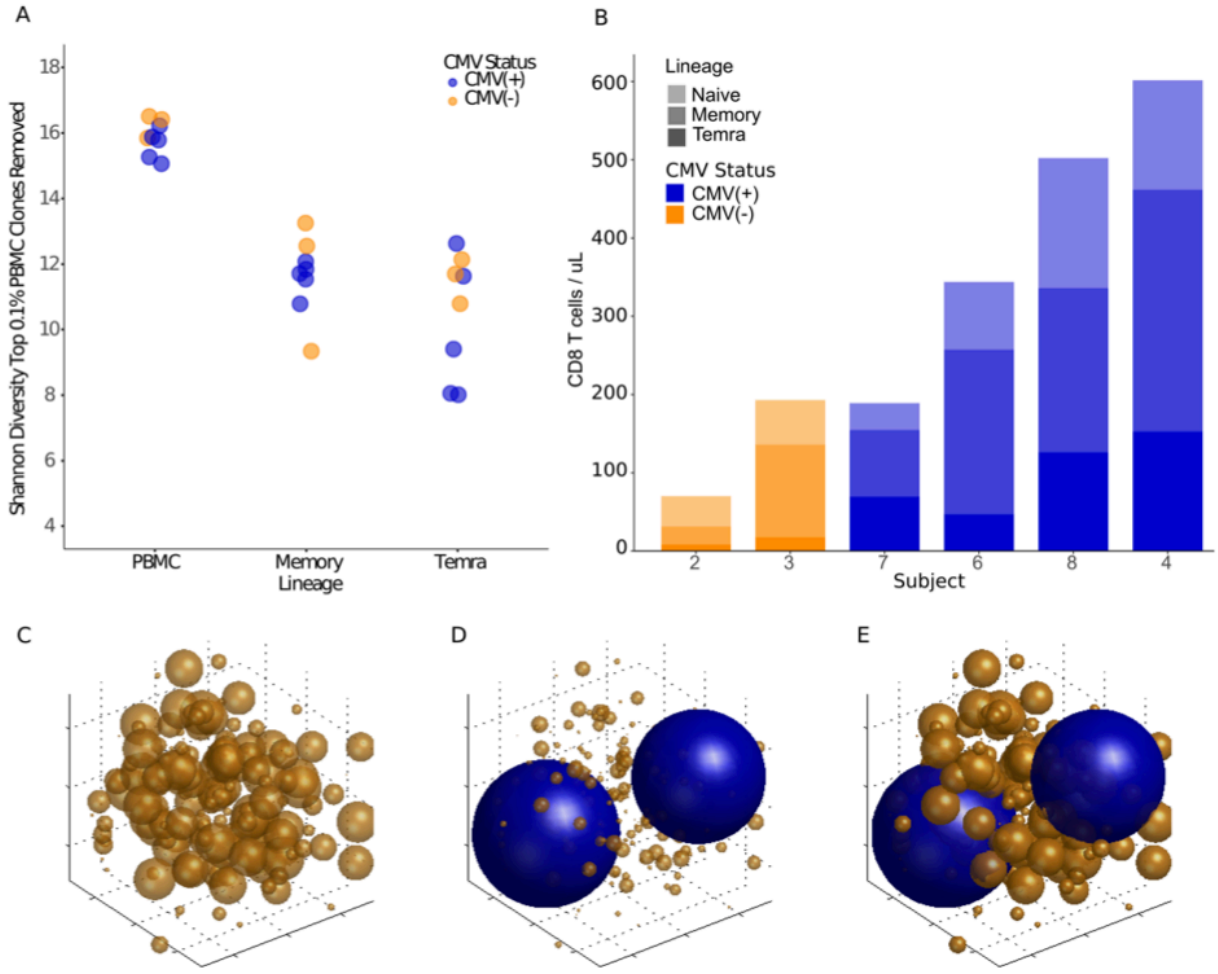


Figure. 5. **Accommodation of high-frequency clonotypes in the repertoire**. (**A**) Scatterplot comparing the Shannon entropy in CMV^+ (blue) and CMV^- (orange) subjects after bioinformatic removal of the most frequent 0.1% of PBMC clonotypes from each T cell subset. PBMC clonotypes were ranked based on frequency and the most frequent 0.1% were selected. These clonotypes were then removed from each sample and Shannon entropy values were recalculated. Memory, $CD8^+$ central and effector memory; Temra, $CD8^+$ memory T cells expressing CD45RA. (**B**) Stacked bar chart comparing the total number of T cells in each subset in 4 CMV^+ and 2 CMV^- subjects. Clinical $CD8^+$ T cell counts were performed on blood samples from 6/8 study subjects. Naïve T cells, light grey; Memory T cells, medium grey; T_{EMRA} cells, dark grey. (**C, D, E**) A model of the T cell repertoire with subordinate clonotypes (orange) and CMV-reactive clonotypes (blue)

depicting the potential impact of CMV infection in (**C**) T cell repertoire at baseline prior to CMV exposure. All clonotypes share similar frequency distributions. (**D, E**) T cell repertoire after CMV exposure with massive clonal expansions that supplant low frequency naïve T cell clonotypes from the repertoire (**D**) or that does not modify the underlying naïve T cell repertoire (**E**). Sphere size corresponds to clone frequency.

5.7 Supplementary Information

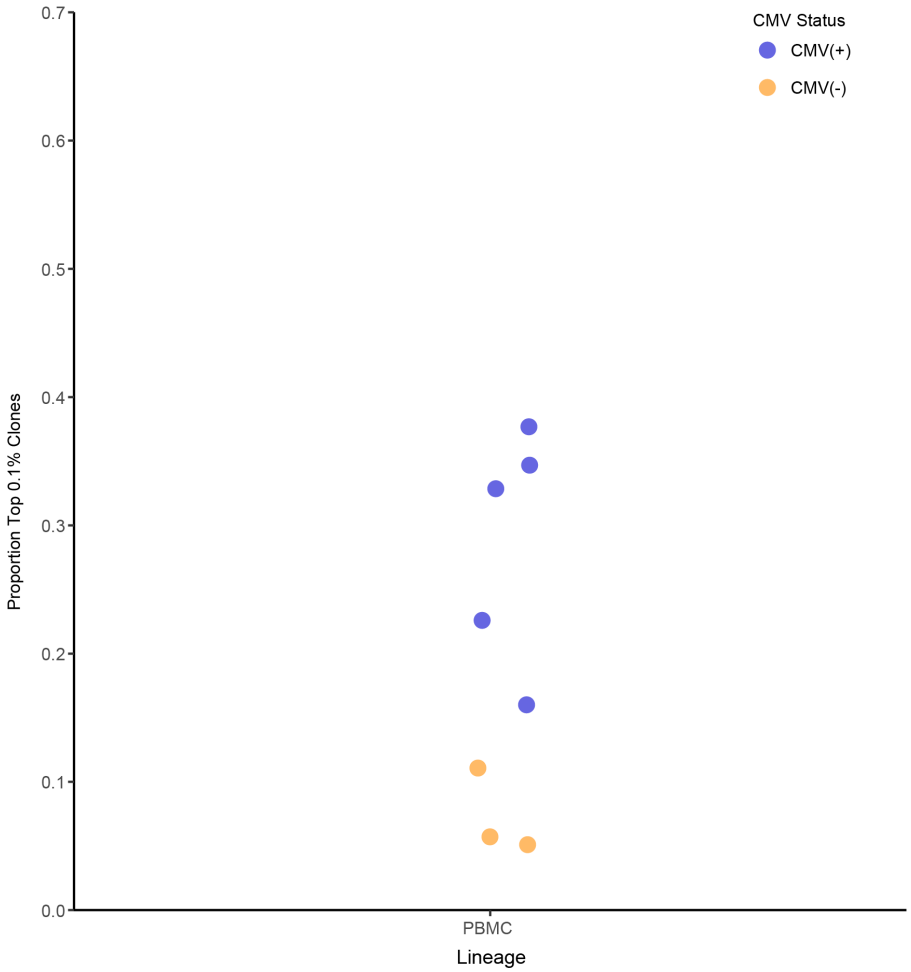
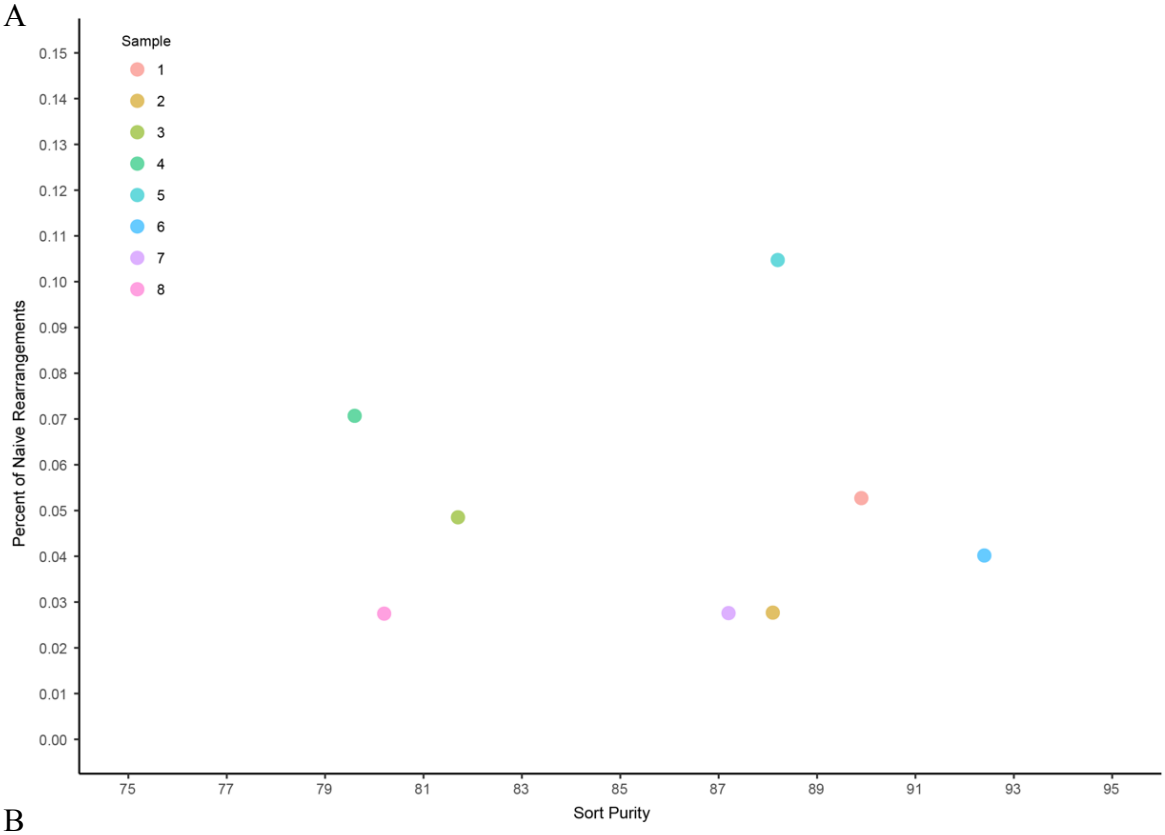


Fig. S1. **High-frequency clonotypes in 8 elderly subjects**. Scatterplot comparing the proportion of the most frequent 0.1% of clonotypes in the peripheral blood repertoires of 5 CMV^{+} (blue) and 3 CMV^{-} (orange) elderly subjects. PBMC clonotypes were ranked based on frequency.

The cumulative abundance of the most frequent 0.1% of clonotypes was divided by the total sample abundance to yield a proportion.

A


Sample
1
2
3
4
5
6
7
8
Percent of Naive Rearrangements
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
75
77
79
81
83
85
87
89
91
93
95
Sort Purity

B

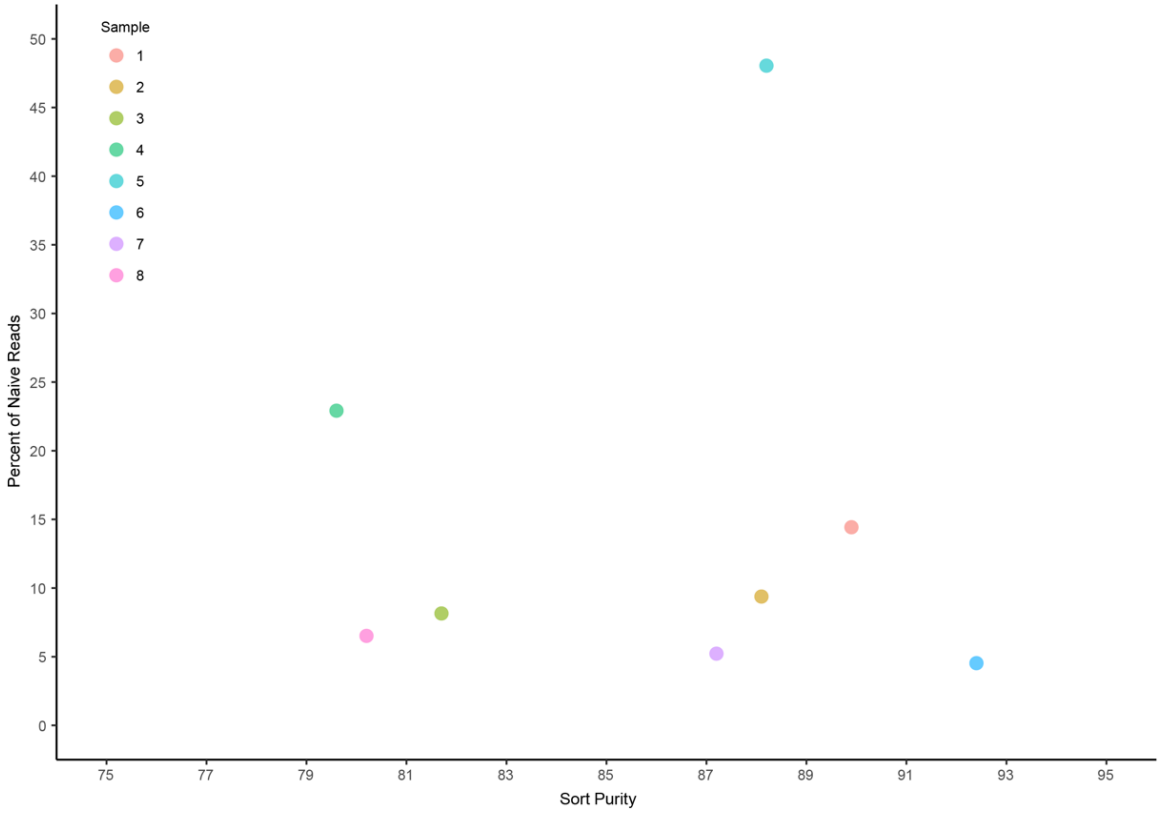

Sample
1
2
3
4
5
6
7
8
Percent of Naive Reads
50
45
40
35
30
25
20
15
10
5
0
75
77
79
81
83
85
87
89
91
93
95
Sort Purity

Fig. S2. **High-frequency clonotype contamination in the naïve repertoire**. Comparison of the most frequent 0.1% of PBMC clonotypes found in naïve T cell samples versus naïve T cell sort purity. For each subject, the most frequent 0.1% of PBMC rearrangements found in the naïve repertoire are expressed **(A)** as a fraction of the total naïve repertoire rearrangements or **(B)** as a proportion of the total frequency of all clonotypes in the naïve repertoire. PBMC clonotypes were ranked based on frequency. The corresponding naïve T cell sample was then searched for TCRβ sequences matching the most frequent 0.1% of PMBC clonotypes. In **(A)** the fraction of unique naïve clonotypes matching the most frequent 0.1% of PMBC clonotypes is divided by the total number of unique clonotypes in the naïve T cell sample. In **(B)** the cumulative abundance of naïve clonotypes matching the most frequent 0.1% of PMBC clonotypes is divided by the total abundance of the naïve T cell sample.

Age Group	CMV(+)	CMV(-)	Total
0-18	13	30	43
19-30	36	54	90
31-40	56	84	140
41-50	78	77	155
51-60	52	41	93
61-74	23	9	32

Table S1. **Number of subjects in each age group used in Fig. 1**. Breakdown of the number of CMV^+ and CMV^- donors in each age group. In total 553 subjects were examined.

			CD8 Fraction (%)			Total Reads (10^6)				Total Rearrangements (10^4)			
Sample	CMV	Age	Naive	Memory	T_{EMRA}	PBMC	Naive	Memory	T_{EMRA}	PBMC	Naive	Memory	T_{EMRA}
1	(-)	71	40.78	42.12	17.00	10.42	5.93	9.74	3.21	17.18	21.73	6.11	1.56
2	(-)	70	55.46	32.42	12.10	13.27	9.47	12.90	0.57	11.28	25.28	1.33	1.47
3	(-)	72	29.40	61.40	9.17	14.02	16.09	13.04	19.46	14.59	22.05	3.57	4.60
4	(+)	71	23.31	51.32	25.40	6.12	3.01	7.17	10.75	13.61	9.68	3.68	0.90
5	(+)	73	19.60	57.65	22.80	9.45	3.67	8.19	13.10	18.16	7.63	2.62	0.33
6	(+)	73	25.21	61.12	13.70	7.21	7.03	15.08	17.12	12.05	23.89	2.42	2.22
7	(+)	70	18.11	45.12	36.80	7.36	10.41	15.16	16.70	7.71	22.84	2.72	3.52
8	(+)	74	33.01	41.80	25.20	3.21	12.54	9.52	5.04	9.93	26.21	2.68	1.26

Table S2. **Characteristics of 8 elderly study subjects**. Age, CMV status, sort purity and a summary of the immunosequencing data for each of the 8 elderly study subjects. The fraction of each $CD8^+$ subset was determined using flow cytometry. Only productively rearranged nucleotide sequences are counted in total rearrangements and sequencing reads for each subject.

5.8 Notes

Acknowledgements: The authors wish to thank Jeanne DaGloria and Heidi Utsugi for technical assistance. **Author contributions:** PL preformed all analysis and wrote the manuscript. MVG and MV provided critical feedback and helped write the manuscript. CJT designed and conducted the CMV stimulation assay and provided critical feedback on the results. EHW and SRR designed the study and provided feedback on the results. KWM designed and oversaw the T cell sorting and DNA extraction procedures. RM designed the experiments, preformed a preliminary analysis on the results and helped write the manuscript. HSR designed and directed the study, provided critical feedback on the results and helped write the manuscript. **Competing interests:** HSR and MV have employment and equity ownership with Adaptive Biotechnologies **Data and materials availability:** All samples are available on the Adaptive Biotechnologies immuneAccess website.

6 Conclusion

6.1 Summary

Understanding the components of an effective immune response is central to the development of vaccines as well as treatments for cancer and autoimmunity. The ability to sequence millions of B and T cells has allowed us to probe the relationship between adaptive immune function and repertoire diversity. In this thesis, I addressed the diversity of the B cell receptor repertoire in healthy adults and the effector T cell repertoire in response to an acute viral infection. I also assessed the impact of chronic CMV infection on the diversity of the underlying T cell repertoire in the elderly. Overall, the work presented in this thesis highlights the extent to which a diverse immune repertoire effects immune function.

My thesis first addresses the diversity of the naïve and memory B cell repertoire in healthy adults. By developing a multi-replicate sequencing method to approximate digital cell counts, we were able to estimate that the B cell repertoire is composed of $1x10^9$ unique naïve B cell clones and $1x10^8$ unique memory B cell clones. This diversity estimate suggests that the nearly each naïve B cell in circulation is a unique clonotype. Furthermore, we created a public database containing $3.7x10^7$ unique BCR sequences with a suite of analytical tools to characterize V gene usage patterns and SHM motifs.

My work goes on to examine the diversity of the effector $CD8^+$ T cell repertoire in response to an acute viral infection using YFV vaccination. In this study, we integrated experimental and computational techniques to identify the number of effector T cell clones responding to the virus at the peak of the immune response. We observed that approximately 2000 unique T cell clones respond to vaccination but only 12% form stable

memory T cells after the immune response has concluded. We went on to show that effector clonotype frequency at the peak of the antiviral immune response correlated with recruitment to the memory population.

My work concludes by examining the diversity of the underlying $CD8^+$ T cell repertoire in elderly individuals with chronic CMV infection. In this study, we tested the hypothesis that CMV driven clonal expansions result in a contraction of the naïve T cell repertoire. We observed that diversity and structure of the underlying naïve T cell repertoire is similar between the CMV seropositive and seronegative elderly. We also found that the T cell repertoire expanded to accommodate large clones in individuals with CMV. Together these results suggest that compromised immune function in the CMV seropositive elderly is not the result of a reduction in naïve T cell repertoire diversity.

This thesis furthers our understanding of the link between B and T cell repertoire diversity and adaptive immunity.

6.2 Discussion and Future Directions

Repertoire diversity and pathogen exposure

Both B and T cell repertoires are highly diverse each containing at least $1x10^8$ unique clones. Given this upper-bound on diversity and our results that ~2,000 clonotypes respond to YFV vaccination, the adaptive immune system has the capacity to recognize approximately 50,000 different pathogens at any particular time. However, there are at most a few thousand pathogens capable of infecting humans. Some insights into this discrepancy come from studies examining immunodominant antigens, which demonstrated that high avidity TCRs clear infections more efficiently. A more diverse

immune repertoire could increase the probability that a high avidity/affinity antigen receptor is incorporated into the adaptive immune response. In this case, measuring the global diversity of the immune repertoire is unlikely to correlate with immunological fitness. New techniques will be needed to determine the diversity of high avidity/affinity antigen receptors in the repertoire. I believe that the combination of structural biology and machine learning will dramatically enhance our ability to identify these receptors.

Identifying antigen-specific T cells

Presently, three distinct techniques are used to identify antigen-specific T cells. In the YFV study, we sorted and sequenced effector $CD8^+$ T cells previously identified to react to vaccination and developed a statistical method to detect clonal expansions. We reasoned that most clonally expanded effector T cells were likely to be specific for YFV antigens. In the CMV study, we stimulated T cells *in vitro* with fibroblasts infected with CMV. By sorting and sequencing activated T cells after stimulation, we were able to identify CMV-reactive clones. A third method not utilized in this work is peptide-MHC tetramer sorting followed by sequencing. This technique allows the isolation of T cells clones that recognize a particular cognate peptide antigen. However, T cell clones not specific to this antigen are also isolated by all of these techniques. More accurate methods to isolate and identify antigen-specific T cell clones need to be developed. Given the biophysical constraints of the TCR:peptide:MHC interaction, molecular modeling techniques could be combine with quantum computing to better identify the determinants of antigen recognition. Alternatively, reagents that more accurately mimic the 3D TCR:peptide:MHC:co-receptor interaction could be created using synthetic cellular

membranes. The development of new methods to identify antigen-specific T cells will allow us to predict an individual's susceptibility to infection based on the composition of their immune repertoire.

7 References

7.1 Chapter 2 References

1. I. Sela-Culang, V. Kunik, Y. Ofran, The structural basis of antibody-antigen recognition. *Frontiers in immunology* **4**, 302 (2013).

2. J. J. Miles, J. McCluskey, J. Rossjohn, S. Gras, Understanding the complexity and malleability of T cell recognition. *Immunology and cell biology* **93**, 433-441 (2015).

3. J. Rossjohn *et al.*, T cell antigen receptor recognition of antigen-presenting molecules. *Annual review of immunology* **33**, 169-200 (2015).

4. M. E. Birnbaum *et al.*, Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073-1087 (2014).

5. G. P. Morris, P. M. Allen, How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature immunology* **13**, 121-128 (2012).

6. C. Milstein, A. J. Munro, The genetic basis of antibody specificity. *Annu Rev Microbiol* **24**, 335-358 (1970).

7. M. D. Cooper, M. N. Alder, The evolution of adaptive immune systems. *Cell* **124**, 815-822 (2006).

8. T. Boehm, J. B. Swann, Origin and evolution of adaptive immunity. *Annu Rev Anim Biosci* **2**, 259-283 (2014).

9. D. L. Farber, N. A. Yudanin, N. P. Restifo, Human memory T cells: generation, compartmentalization and homeostasis. *Nature reviews. Immunology* **14**, 24-35 (2014).

10. D. Tarlinton, K. Good-Jacobson, Diversity among memory B cells: origin, consequences, and utility. *Science* **341**, 1205-1211 (2013).

11. E. W. Newell, N. Sigal, S. C. Bendall, G. P. Nolan, M. M. Davis, Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity* **36**, 142-152 (2012).

12. S. Becattini *et al.*, T cell immunity. Functional heterogeneity of human memory CD4(+) T cell clones primed by pathogens or vaccines. *Science* **347**, 400-406 (2015).

13. L. F. Su, M. M. Davis, Antiviral memory phenotype T cells in unexposed adults. *Immunological reviews* **255**, 95-109 (2013).

14. A. Han, J. Glanville, L. Hansmann, M. M. Davis, Linking T cell receptor sequence to functional phenotype at the single-cell level. *Nature biotechnology* **32**, 684-692 (2014).

15. T. W. LeBien, T. F. Tedder, B lymphocytes: how they develop and function. *Blood* **112**, 1570-1580 (2008).

16. T. L. Rothstein, D. O. Griffin, N. E. Holodick, T. D. Quach, H. Kaku, Human B 1 cells take the stage. *Ann N Y Acad Sci* **1285**, 97-114 (2013).

17. D. Allman, S. Pillai, Peripheral B cell subsets. *Curr Opin Immunol* **20**, 149-157 (2008).

18. F. Weisel, M. Shlomchik, Memory B Cells of Mice and Humans. *Annual review of immunology* **35**, 255-284 (2017).

19. C. T. Watson *et al.*, Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *American journal of human genetics* **92**, 530-546 (2013).

20. F. Melchers, Checkpoints that control B cell development. *J Clin Invest* **125**, 2203-2210 (2015).

21. D. G. Schatz, M. A. Oettinger, M. S. Schlissel, V(D)J recombination: molecular biology and regulation. *Annual review of immunology* **10**, 359-383 (1992).

22. C. Vettermann, M. S. Schlissel, Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunological reviews* **237**, 22-42 (2010).

23. G. D. Victora, M. C. Nussenzweig, Germinal centers. *Annual review of immunology* **30**, 429-457 (2012).

24. G. Teng, F. N. Papavasiliou, Immunoglobulin somatic hypermutation. *Annu Rev Genet* **41**, 107-120 (2007).

25. A. H. Ellebedy *et al.*, Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nature immunology*, (2016).

26. H. W. Schroeder, Jr., L. Cavacini, Structure and function of immunoglobulins. *J Allergy Clin Immunol* **125**, S41-52 (2010).

27. J. K. Hwang, F. W. Alt, L. S. Yeap, Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination. *Microbiol Spectr* **3**, MDNA3-0037-2014 (2015).

28. R. N. Germain, T cell development and the CD4-CD8 lineage decision. *Nature reviews. Immunology* **2**, 309-322 (2002).

29. J. Zhu, H. Yamane, W. E. Paul, Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* **28**, 445-489 (2010).

30. P. Wong, E. G. Pamer, CD8 T cell responses to infectious pathogens. *Annual review of immunology* **21**, 29-70 (2003).

31. F. Sallusto, J. Geginat, A. Lanzavecchia, Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annual review of immunology* **22**, 745-763 (2004).

32. M. P. Lefranc, IMGT, The International ImMunoGeneTics Information System, http://imgt.cines.fr. *Methods in molecular biology* **248**, 27-49 (2004).

33. D. J. Pennington, B. Silva-Santos, A. C. Hayday, Gammadelta T cell development--having the strength to get there. *Curr Opin Immunol* **17**, 108-115 (2005).

34. H. von Boehmer *et al.*, Pleiotropic changes controlled by the pre-T cell receptor. *Curr Opin Immunol* **11**, 135-142 (1999).

35. N. R. Gascoigne, V. Rybakin, O. Acuto, J. Brzostek, TCR Signal Strength and T Cell Development. *Annu Rev Cell Dev Biol* **32**, 327-348 (2016).

36. J. J. Calis, B. R. Rosenberg, Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in immunology* **35**, 581-590 (2014).

37. H. Robins, Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* **25**, 646-652 (2013).

38. C. D. Surh, J. Sprent, Homeostasis of naive and memory T cells. *Immunity* **29**, 848-862 (2008).

39. M. C. van Zelm, T. Szczepanski, M. van der Burg, J. J. van Dongen, Replication history of B lymphocytes reveals homeostatic proliferation and extensive antigen-induced B cell expansion. *The Journal of experimental medicine* **204**, 645-655 (2007).

40. G. Lythe, R. E. Callard, R. L. Hoare, C. Molina-Paris, How many TCR clonotypes does a body maintain? *Journal of theoretical biology* **389**, 214-224 (2016).

41. A. M. Wertheimer *et al.*, Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans. *J Immunol* **192**, 2143-2155 (2014).

42. K. Rubtsova, A. V. Rubtsov, M. P. Cancro, P. Marrack, Age-Associated B Cells: A T-bet-Dependent Effector with Roles in Protective and Pathogenic Immunity. *J Immunol* **195**, 1933-1937 (2015).

43. J. L. Xu, M. M. Davis, Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37-45 (2000).

44. H. S. Robins *et al.*, Comprehensive assessment of T cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099-4107 (2009).

45. H. S. Robins *et al.*, Overlap and Effective Size of the Human CD8(+) T Cell Receptor Repertoire. *Sci Transl Med* **2**, (2010).

46. K. Larimore, M. W. McCormick, H. S. Robins, P. D. Greenberg, Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing. *J Immunol* **189**, 3221-3230 (2012).

47. O. V. Britanova *et al.*, Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J Immunol* **192**, 2689-2698 (2014).

48. M. A. Turchaninova *et al.*, High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* **11**, 1599-1616 (2016).

49. K. Best, T. Oakes, J. M. Heather, J. Shawe-Taylor, B. Chain, Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci Rep* **5**, 14629 (2015).

50. C. S. Carlson *et al.*, Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* **4**, (2013).

51. B. Howie *et al.*, High-throughput pairing of T cell receptor alpha and beta sequences. *Sci Transl Med* **7**, 301ra131 (2015).

52. B. J. DeKosky *et al.*, High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology* **31**, 166-169 (2013).

53. B. J. DeKosky *et al.*, In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* **21**, 86-91 (2015).

54. J. R. McDaniel, B. J. DeKosky, H. Tanno, A. D. Ellington, G. Georgiou, Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat Protoc* **11**, 429-442 (2016).

55. N. L. La Gruta, P. G. Thomas, Interrogating the relationship between naive and immune antiviral T cell repertoires. *Curr Opin Virol* **3**, 447-451 (2013).

56. G. C. Wang, P. Dash, J. A. McCullers, P. C. Doherty, P. G. Thomas, T cell receptor alphabeta diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Sci Transl Med* **4**, 128ra142 (2012).

57. J. Nikolich-Zugich, M. K. Slifka, I. Messaoudi, The many important facets of T cell repertoire diversity. *Nature reviews. Immunology* **4**, 123-132 (2004).

58. K. M. Roskin *et al.*, IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Sci Transl Med* **7**, 302ra135 (2015).

59. Y. N. Lee *et al.*, Characterization of T and B cell repertoire diversity in patients with RAG deficiency. *Sci Immunol* **1**, (2016).

60. A. Fischer, Severe combined immunodeficiencies (SCID). *Clinical and experimental immunology* **122**, 143-149 (2000).

61. H. Morbach, E. M. Eichhorn, J. G. Liese, H. J. Girschick, Reference values for B cell subpopulations from infancy to adulthood. *Clinical and experimental immunology* **162**, 271-279 (2010).

62. V. V. Ganusov, R. J. De Boer, Do most lymphocytes in humans really reside in the gut? *Trends in immunology* **28**, 514-518 (2007).

63. D. A. Bolotin *et al.*, Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* **42**, 3073-3083 (2012).

64. A. Martin-Fontecha, A. Lanzavecchia, F. Sallusto, Dendritic cell migration to peripheral lymph nodes. *Handb Exp Pharmacol*, 31-49 (2009).

65. G. Freer, D. Matteucci, Influence of dendritic cells on viral pathogenicity. *PLoS pathogens* **5**, e1000384 (2009).

66. C. J. Henry, D. A. Ornelles, L. M. Mitchell, K. L. Brzoza-Lewis, E. M. Hiltbold, IL-12 produced by dendritic cells augments CD8+ T cell activation through the production of the chemokines CCL1 and CCL17. *J Immunol* **181**, 8576-8584 (2008).

67. A. J. Sant, A. McMichael, Revealing the role of CD4(+) T cells in viral immunity. *The Journal of experimental medicine* **209**, 1391-1395 (2012).

68. S. M. Kaech, E. J. Wherry, Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection. *Immunity* **27**, 393-405 (2007).

69. S. Sridhar *et al.*, Cellular immune correlates of protection against symptomatic pandemic influenza. *Nat Med* **19**, 1305-1312 (2013).

70. D. A. Price *et al.*, T cell receptor recognition motifs govern immune escape patterns in acute SIV infection. *Immunity* **21**, 793-803 (2004).

71. I. Messaoudi, J. A. Guevara Patino, R. Dyall, J. LeMaoult, J. Nikolich-Zugich, Direct link between mhc polymorphism, T cell avidity, and diversity in immune defense. *Science* **298**, 1797-1800 (2002).

72. M. Hashimoto *et al.*, CD8 T Cell Exhaustion in Chronic Infection and Cancer: Opportunities for Interventions. *Annu Rev Med* **69**, 301-318 (2018).

73. L. A. Vella, R. S. Herati, E. J. Wherry, CD4(+) T Cell Differentiation in Chronic Viral Infections: The Tfh Perspective. *Trends Mol Med* **23**, 1072-1087 (2017).

74. K. E. Pauken, E. J. Wherry, Overcoming T cell exhaustion in infection and cancer. *Trends in immunology* **36**, 265-276 (2015).

75. E. J. Wherry, R. Ahmed, Memory CD8 T cell differentiation during viral infection. *J Virol* **78**, 5535-5545 (2004).

76. L. Cicin-Sain *et al.*, Cytomegalovirus infection impairs immune responses and accentuates T cell pool changes observed in mice with aging. *PLoS pathogens* **8**, e1002849 (2012).

77. B. Reinhardt, R. Jaspert, M. Niedrig, C. Kostner, J. L'Age-Stehr, Development of viremia and humoral and cellular parameters of immune activation after vaccination with yellow fever virus strain 17D: a model of human flavivirus infection. *Journal of medical virology* **56**, 159-167 (1998).

78. T. Querec *et al.*, Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. *The Journal of experimental medicine* **203**, 413-424 (2006).

79. T. D. Querec *et al.*, Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature immunology* **10**, 116-125 (2009).

7.2 Chapter 3 References

1. Davis MM, Calame K, Early PW, Livant DL, Joho R, Weissman IL, et al. An immunoglobulin heavy- chain gene is formed by at least two recombinational events. Nature. 1980; 283(5749):733–9.

2. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet. 2013; 92(4):530–46.

3. Murphy K, Travers P, Walport M. Janeway's Immunobiology. 7th ed. New York, NY: Garland Science; 2008.

4. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000; 102(5):553–63.

5. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. Nature. 1991; 354(6352):389–92.

6. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature. 2003; 424(6944):103–7.

7. Calis JJ, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol. 2014; 35(12):581–90.

8. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The past, present, and future of immune repertoire biology—the rise of next-generation repertoire analysis. Front Immunol. 2013; 4:413.

9. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 2006; 34(Database issue):D781–4.

10. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. PLoS One. 2011; 6(8):e22365.

11. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol. 2010; 184 (12):6986–92.

12. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009; 1(12):12ra23.

13. Briney BS, Willis JR, McKinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. Genes Immun. 2012; 13(6):469–73.

14. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol. 2013; 31(2):166–9.

15. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med. 2015; 21(1):86–91.

16. Galson JD, Truck J, Clutterbuck EA, Fowler A, Cerundolo V, Pollard AJ, et al. B cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B cell activation. Genome Med. 2016; 8(1):68.

17. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014; 16(1):105–14.

18. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol. 2012; 189(6):3221–30.

19. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci USA. 2014; 111 (13):4928–33.

20. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, et al. Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. Immunogenetics. 2012; 64(5):337–50.

21. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B cell recall using antibody repertoire sequencing. Proc Natl Acad Sci USA. 2013; 110(33):13463–8.

22. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med. 2016; 8(1):60.

23. Wu D, Emerson RO, Sherwood A, Loh ML, Angiolillo A, Howie B, et al. Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. Clin Cancer Res. 2014; 20 (17):4540–8.

24. Cortina-Ceballos B, Godoy-Lozano EE, Tellez-Sosa J, Ovilla-Munoz M, Samano-Sanchez H, Aguilar- Salgado A, et al. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. Genome Med. 2015; 7:124.

25. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol. 2013; 25(5):646–52.

26. Elhanati Y, Murugan A, Callan CG Jr., Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci USA. 2014; 111(27):9875–80.

27. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009; 114(19):4099– 107.

28. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010; 2(47):47ra64.

29. Sherwood AM, Desmarais C, Livingston RJ, Andriesen J, Haussler M, Carlson CS, et al. Deep sequencing of the human TCRgamma and TCRbeta repertoires suggests that TCRbeta rearranges after alphabeta and gammadelta T cell commitment. Sci Transl Med. 2011; 3(90):90ra61.

30. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med. 2012; 4 (134):134ra63.

31. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature. 2014; 509(7498):55–62.

32. Muraro PA, Robins H, Malhotra S, Howell M, Phippard D, Desmarais C, et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. J Clin Invest. 2014; 124(3):1168–72.

33. Schneider-Hohendorf T, Mohan H, Bien CG, Breuer J, Becker A, Gorlich D, et al. CD8(+) T cell pathogenicity in Rasmussen encephalitis elucidated by large-scale T cell receptor sequencing. Nat Commun. 2016; 7:11153.

34. DeWitt WS, Emerson RO, Lindau P, Vignali M, Snyder TM, Desmarais C, et al. Dynamics of the cytotoxic T cell response to a model of acute viral infection. J Virol. 2015.

35. Morris H, DeWolf S, Robins H, Sprangers B, LoCascio SA, Shonts BA, et al. Tracking donor-reactive T cells: Evidence for clonal deletion in tolerant kidney transplant patients. Sci Transl Med. 2015; 7 (272):272ra10.

36. Emerson RO, Mathew JM, Konieczna IM, Robins HS, Leventhal JR. Defining the alloreactive T cell repertoire using high-throughput sequencing of mixed lymphocyte reaction culture. PLoS One. 2014; 9 (11):e111943.

37. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, et al. High- throughput sequencing of T cell receptors reveals a homogeneous repertoire of tumor-infiltrating lymphocytes in ovarian cancer. J Pathol. 2013; 231(4):433–40.

38. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014; 515(7528):568–71.

39. Hsu MS, Sedighim S, Wang T, Antonios JP, Everson RG, Tucker AM, et al. TCR Sequencing Can Identify and Track Glioma-Infiltrating T Cells after DC Vaccination. Cancer Immunol Res. 2016; 4(5):412–8.

40. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun. 2013; 4:2680.

41. Chothia C, Gelfand I, Kister A. Structural determinants in the sequences of immunoglobulin variable domain. J Mol Biol. 1998; 278(2):457–79.

42. Fisher RA, Corbet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol. 1943; 12:42–58.

43. Sanathanan L. Estimating the size of a truncated sample. J Am Statist Assoc. 1977; 72(356):669–72.

44. Rodrigues J, Milan LA, Leite JG. Hierarchical bayesian estimation for the number of species. Biom J. 2001; 43(6):737–46.

45. Morbach H, Eichhorn EM, Liese JG, Girschick HJ. Reference values for B cell subpopulations from infancy to adulthood. Clin Exp Immunol. 2010; 162(2):271–9.

46. Agematsu K, Nagumo H, Yang FC, Nakazawa T, Fukushima K, Ito S, et al. B cell subpopulations separated by CD27 and crucial collaboration of CD27+ B cells and helper T cells in immunoglobulin production. Eur J Immunol. 1997; 27(8):2073–9.

47. van Zelm MC, Szczepanski T, van der Burg M, van Dongen JJ. Replication history of B lymphocytes reveals homeostatic proliferation and extensive antigen-induced B cell expansion. J Exp Med. 2007; 204(3):645–55.

48. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 1998; 188 (11):2151–62.

49. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B cell populations. Blood. 2010; 116(7):1070–8.

50. Wu YC, Kipling D, Dunn-Walters DK. The relationship between CD27 negative and positive B cell populations in human peripheral blood. Front Immunol. 2011; 2:81.

51. Glanville J, Kuo TC, von Budingen HC, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A. 2011; 108(50):20066–71.

52. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity. 2000; 13(1):37–45.

53. Rock EP, Sibbald PR, Davis MM, Chien YH. CDR3 length in antigen-specific immune receptors. J Exp Med. 1994; 179(1):323–8.

54. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Front Immunol. 2014; 5:96.

55. Herzenberg LA, Black SJ, Tokuhisa T, Herzenberg LA. Memory B cells at successive stages of differentiation. Affinity maturation and the role of IgD receptors. J Exp Med. 1980; 151(5):1071–87.

56. Weiss U, Rajewsky K. The repertoire of somatic antibody mutants accumulating in the memory compartment after primary immunization is restricted through affinity maturation and mirrors that expressed in the secondary response. J Exp Med. 1990; 172(6):1681–9.

57. Kocks C, Rajewsky K. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. Proc Natl Acad Sci USA. 1988; 85(21):8206–10.

58. Zhang Y, Meyer-Hermann M, George LA, Figge MT, Khan M, Goodall M, et al. Germinal center B cells govern their own fate via antibody feedback. J Exp Med. 2013; 210(3):457–64.

59. Peron S, Laffleur B, Denis-Lagache N, Cook-Moreau J, Tinguely A, Delpy L, et al. AID-driven deletion causes immunoglobulin heavy chain locus suicide recombination in B cells. Science. 2012; 336 (6083):931–4.

60. Victora GD, Nussenzweig MC. Germinal centers. Annu Rev Immunol. 2012; 30:429–57.

61. Bransteitter R, Pham P, Calabrese P, Goodman MF. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. J Biol Chem. 2004; 279(49):51612–21.

62. Greene J, Birtwistle MR, Ignatowicz L, Rempala GA. Bayesian multivariate Poisson abundance models for T cell receptor data. J Theor Biol. 2013; 326:1–10.

63. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. J Theor Biol. 2011; 269(1):1–15.

7.3 Chapter 4 References

1. Kaech SM, Wherry EJ. 2007. Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection. Immunity 27:393– 405.

2. Wherry EJ, Ha SJ, Kaech SM, Haining WN, Sarkar S, Kalia V, Subramaniam S, Blattman JN, Barber DL, Ahmed R. 2007. Molecular signature of CD8+ T cell exhaustion during chronic viral infection. Immunity 27:670 – 684.

3. Engel I, Hedrick SM. 1988. Site-directed mutations in the VDJ junctional region of a T cell receptor beta chain cause changes in antigenic peptide recognition. Cell 54:473– 484.

4. Jorgensen JL, Esser U, Fazekas de St Groth B, Reay PA, Davis MM. 1992. Mapping T cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics. Nature 355:224–230.

5. McHeyzer-Williams MG, Davis MM. 1995. Antigen-specific development of primary and memory T cells *in vivo*. Science 268:106 –111.

6. Miller JD, van der Most RG, Akondy RS, Glidewell JT, Albott S, Masopust D, Murali-Krishna K, Mahar PL, Edupuganti S, Lalor S, Germon S, Del Rio C, Mulligan MJ, Staprans SI, Altman JD, Feinberg MB, Ahmed R. 2008. Human effector and memory CD8+ T cell responses to smallpox and yellow fever vaccines. Immunity 28:710–722.

7. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. 2012. Cytometry by time-of-flight shows combinatorial cytokine expression and virus- specific cell niches within a continuum of CD8+ T cell phenotypes. Immunity 36:142–152.

8. Precopio ML, Betts MR, Parrino J, Price DA, Gostick E, Ambrozak DR, Asher TE, Douek DC, Harari A, Pantaleo G, Bailer R, Graham BS, Roederer M, Koup RA. 2007. Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. J Exp Med 204:1405–1416.

9. Akondy RS, Monson ND, Miller JD, Edupuganti S, Teuwen D, Wu H, Quyyumi F, Garg S, Altman JD, Del Rio C, Keyserling HL, Ploss A, Rice CM, Orenstein WA, Mulligan MJ, Ahmed R. 2009. The yellow fever virus vaccine induces a broad and polyfunctional human memory CD8+ T cell response. J Immunol 183:7919–7930.

10. Co MD, Terajima M, Cruz J, Ennis FA, Rothman AL. 2002. Human cytotoxic T lymphocyte responses to live attenuated 17D yellow fever vac- cine: identification

of HLA-B35-restricted CTL epitopes on nonstructural proteins NS1, NS2b, NS3, and the structural protein E. Virology 293:151– 163.

11. Turner SJ, Diaz G, Cross R, Doherty PC. 2003. Analysis of clonotype distribution and persistence for an influenza virus-specific CD8+ T cell response. Immunity 18:549 –559.

12. Blom K, Braun M, Ivarsson MA, Gonzalez VD, Falconer K, Moll M, Ljunggren HG, Michaelsson J, Sandberg JK. 2013. Temporal dynamics of the primary human T cell response to yellow fever virus 17D as it matures from an effector- to a memory-type response. J Immunol 190: 2150 –2158.

13. Manuel ER, Charini WA, Sen P, Peyerl FW, Kuroda MJ, Schmitz JE, Autissier P, Sheeter DA, Torbett BE, Letvin NL. 2006. Contribution of T cell receptor repertoire breadth to the dominance of epitope-specific CD8+ T-lymphocyte responses. J Virol 80:12032–12040.

14. Miconnet I, Marrau A, Farina A, Taffe P, Vigano S, Harari A, Pantaleo G. 2011. Large TCR diversity of virus-specific CD8 T cells provides the mechanistic basis for massive TCR renewal after antigen exposure. J Im- munol 186:7039 –7049.

15. Henrickson SE, Perro M, Loughhead SM, Senman B, Stutte S, Quigley M, Alexe G, Iannacone M, Flynn MP, Omid S, Jesneck JL, Imam S, Mempel TR, Mazo IB, Haining WN, von Andrian UH. 2013. Antigen availability determines CD8(+) T cell-dendritic cell interaction kinetics and memory fate decisions. Immunity 39:496–507.

16. Ahmed R, Akondy RS. 2011. Insights into human CD8(+) T cell memory using the yellow fever and smallpox vaccines. Immunol Cell Biol 89: 340 –345.

17. Achour A, Michaelsson J, Harris RA, Odeberg J, Grufman P, Sandberg JK, Levitsky V, Karre K, Sandalova T, Schneider G. 2002. A structural basis for LCMV immune evasion: subversion of H-2D^b and H-2K^b presentation of gp33 revealed by comparative crystal structure analyses. Immunity 17:757–768.

18. Eckle SB, Turner SJ, Rossjohn J, McCluskey J. 2013. Predisposed $\alpha\beta$ T cell antigen receptor recognition of MHC and MHC-I like molecules? Curr Opin Immunol 25:653– 659.

19. Hou S, Hyland L, Ryan KW, Portner A, Doherty PC. 1994. Virus- specific CD8+ T cell memory determined by clonal burst size. Nature 369:652– 654.

20. Vezys V, Yates A, Casey KA, Lanier G, Ahmed R, Antia R, Masopust D. 2009. Memory CD8 T cell compartment grows in size with immunological experience. Nature 457:196 –199.

21. Badovinac VP, Porter BB, Harty JT. 2002. Programmed contraction of CD8(+) T cells after infection. Nat Immunol 3:619 – 626.

22. Sung JH, Zhang H, Moseman EA, Alvarez D, Iannacone M, Henrickson SE, de la Torre JC, Groom JR, Luster AD, von Andrian UH. 2012. Chemokine guidance of central memory T cells is critical for antiviral recall responses in lymph nodes. Cell 150:1249 –1263.

23. Flynn KJ, Belz GT, Altman JD, Ahmed R, Woodland DL, Doherty PC. 1998. Virus-specific CD8+ T cells in primary and secondary influenza pneumonia. Immunity 8:683– 691.

24. Sourdive DJ, Murali-Krishna K, Altman JD, Zajac AJ, Whitmire JK, Pannetier C, Kourilsky P, Evavold B, Sette A, Ahmed R. 1998. Conserved T cell receptor repertoire in primary and memory CD8 T cell responses to an acute viral infection. J Exp Med 188:71– 82.

25. Lee E, Lobigs M. 2008. E protein domain III determinants of yellow fever virus 17D vaccine strain enhance binding to glycosaminoglycans, impede virus spread, and attenuate virulence. J Virol 82:6024 – 6033.

26. SanofiPasteur. YF-VAX prospectus. Document LE6445-LE6446. SanofiPasteur, Rockville, MD.

27. Pulendran B. 2009. Learning immunology from the yellow fever vaccine: innate immunity to systems vaccinology. Nat Rev Immunol 9:741–747.

28. Querec T, Bennouna S, Alkan S, Laouar Y, Gorden K, Flavell R, Akira S, Ahmed R, Pulendran B. 2006. Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. J Exp Med 203:413–424.

29. Jonker EF, Visser LG, Roukens AH. 2013. Advances and controversies in yellow fever vaccination. Ther Adv Vaccines 1:144 –152.

30. Reinhardt B, Jaspert R, Niedrig M, Kostner C, L'Age-Stehr J. 1998. Development of viremia and humoral and cellular parameters of immune activation after vaccination with yellow fever virus strain 17D: a model of human flavivirus infection. J Med Virol 56:159 –167.

31. Santos AP, Bertho AL, Dias DC, Santos JR, Marcovistz R. 2005. Lymphocyte subset analyses in healthy adults vaccinated with yellow fever 17DD virus. Mem Inst Oswaldo Cruz 100:331–337.

32. Kohler S, Bethke N, Bothe M, Sommerick S, Frentsch M, Romagnani C, Niedrig M, Thiel A. 2012. The early cellular signatures of protective immunity induced by live viral vaccination. Eur J Immunol 42:2363–2373.

33. James EA, LaFond RE, Gates TJ, Mai DT, Malhotra U, Kwok WW. 2013. Yellow fever vaccination elicits broad functional CD4 T cell re- sponses that recognize structural and nonstructural proteins. J Virol 87: 12794 –12804.

34. Co MD, Kilpatrick ED, Rothman AL. 2009. Dynamics of the CD8 T cell response following yellow fever virus 17D immunization. Immunology 128:e718 – e727.

35. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. 2009. Comprehensive assessment of T cell receptor beta-chain diversity in $\alpha\beta$ T cells. Blood 114:4099 – 4107.

36. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi- Saljoqi S, Sasorith S, Lefranc G, Kossida S. 2015. IMGT, the international ImMunoGeneTics information system 25 years on. Nucleic Acids Res 43:D413–D422.

37. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. 2004. IMGT/ JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J junctions. Bioinformatics (Oxford) 20(Suppl 1):i379 –i385.

38. Wu D, Emerson RO, Sherwood A, Loh ML, Angiolillo A, Howie B, Vogt J, Rieder M, Kirsch I, Carlson C, Williamson D, Wood BL, Robins H. 2014. Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. Clin Cancer Res 20:4540 – 4548.

39. Pulendran B, Ahmed R. 2011. Immunological mechanisms of vaccina- tion. Nat Immunol 12:509 –517.

40. Murali-Krishna K, Altman JD, Suresh M, Sourdive DJ, Zajac AJ, Miller JD, Slansky J, Ahmed R. 1998. Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. Immunity 8:177–187.

41. Hamann D, Baars PA, Rep MH, Hooibrink B, Kerkhof-Garde SR, Klein MR, van Lier RA. 1997. Phenotypic and functional separation of memory and effector human CD8+ T cells. J Exp Med 186:1407–1418.

42. Sallusto F, Lenig D, Forster R, Lipp M, Lanzavecchia A. 1999. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401:708 –712.

7.4 Chapter 5 References

1. Shaw, A. C., S. Joshi, H. Greenwood, A. Panda, and J. M. Lord. 2010. Aging of the innate immune system. *Current opinion in immunology* 22: 507-513.

2. Goronzy, J. J., and C. M. Weyand. 2013. Understanding immunosenescence to improve responses to vaccines. *Nature immunology* 14: 428-436.

3. Reber, A. J., T. Chirkova, J. H. Kim, W. Cao, R. Biber, D. K. Shay, and S. Sambhara. 2012. Immunosenescence and Challenges of Vaccination against Influenza in the Aging Population. *Aging and disease* 3: 68-90.

4. Nikolich-Zugich, J., G. Li, J. L. Uhrlaub, K. R. Renkema, and M. J. Smithey. 2012. Age-related changes in CD8 T cell homeostasis and immunity to infection. *Seminars in immunology* 24: 356-364.

5. Nikolich-Zugich, J. 2008. Ageing and life-long maintenance of T cell subsets in the face of latent persistent infections. *Nature reviews. Immunology* 8: 512-522.

6. Almanzar, G., S. Schwaiger, B. Jenewein, M. Keller, D. Herndler-Brandstetter, R. Wurzner, D. Schonitzer, and B. Grubeck-Loebenstein. 2005. Long-term cytomegalovirus infection leads to significant changes in the composition of the CD8+ T cell repertoire, which may be the basis for an imbalance in the cytokine production profile in elderly persons. *Journal of virology* 79: 3675-3683.

7. Wikby, A., F. Ferguson, R. Forsey, J. Thompson, J. Strindhall, S. Lofgren, B. O. Nilsson, J. Ernerudh, G. Pawelec, and B. Johansson. 2005. An immune risk phenotype, cognitive impairment, and survival in very late life: impact of allostatic load in Swedish octogenarian and nonagenarian humans. *The journals of gerontology. Series A, Biological sciences and medical sciences* 60: 556-565.

8. Savva, G. M., A. Pachnio, B. Kaul, K. Morgan, F. A. Huppert, C. Brayne, and P. A. Moss. 2013. Cytomegalovirus infection is associated with increased mortality in the older population. *Aging cell* 12: 381-387.

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

10. Hadrup, S. R., J. Strindhall, T. Kollgaard, T. Seremet, B. Johansson, G. Pawelec, P. thor Straten, and A. Wikby. 2006. Longitudinal studies of clonally expanded CD8 T cells reveal a repertoire shrinkage predicting mortality and an increased number of dysfunctional cytomegalovirus-specific T cells in the very elderly. *Journal of immunology (Baltimore, Md. : 1950)* 176: 2645-2653.

11. Khan, N., D. Best, R. Bruton, L. Nayak, A. B. Rickinson, and P. A. Moss. 2007. T cell recognition patterns of immunodominant cytomegalovirus antigens in primary and persistent infection. *Journal of immunology (Baltimore, Md. : 1950)* 178: 4455-4465.

12. Michie, C. A., A. McLean, C. Alcock, and P. C. Beverley. 1992. Lifespan of human lymphocyte subsets defined by CD45 isoforms. *Nature* 360: 264-265.

13. Derhovanessian, E., A. B. Maier, K. Hahnel, R. Beck, A. J. de Craen, E. P. Slagboom, R. G. Westendorp, and G. Pawelec. 2011. Infection with cytomegalovirus but not herpes simplex virus induces the accumulation of late-differentiated CD4+ and CD8+ T cells in humans. *The Journal of general virology* 92: 2746-2756.

14. Griffiths, S. J., N. E. Riddell, J. Masters, V. Libri, S. M. Henson, A. Wertheimer, D. Wallace, S. Sims, L. Rivino, A. Larbi, D. M. Kemeny, J. Nikolich-Zugich, F. Kern, P. Klenerman, V. C. Emery, and A. N. Akbar. 2013. Age-associated increase of low-avidity cytomegalovirus-specific CD8+ T cells that re-express CD45RA. *J Immunol* 190: 5363-5372.

15. Smithey, M. J., G. Li, V. Venturi, M. P. Davenport, and J. Nikolich-Zugich. 2012. Lifelong persistent viral infection alters the naive T cell pool, impairing CD8 T cell immunity in late life. *Journal of immunology (Baltimore, Md. : 1950)* 189: 5356-5366.

16. Messaoudi, I., J. Lemaoult, J. A. Guevara-Patino, B. M. Metzner, and J. Nikolich-Zugich. 2004. Age-related CD8 T cell clonal expansions constrict CD8 T cell repertoire and have the potential to impair immune defense. *The Journal of experimental medicine* 200: 1347-1358.

17. Yager, E. J., M. Ahmed, K. Lanzer, T. D. Randall, D. L. Woodland, and M. A. Blackman. 2008. Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *The Journal of experimental medicine* 205: 711-723.

18. Jankovic, V., I. Messaoudi, and J. Nikolich-Zugich. 2003. Phenotypic and functional T cell aging in rhesus macaques (Macaca mulatta): differential behavior of CD4 and CD8 subsets. *Blood* 102: 3244-3251.

19. Naylor, K., G. Li, A. N. Vallejo, W. W. Lee, K. Koetz, E. Bryl, J. Witkowski, J. Fulbright, C. M. Weyand, and J. J. Goronzy. 2005. The influence of age on T cell generation and TCR diversity. *Journal of immunology (Baltimore, Md. : 1950)* 174: 7446-7452.

20. Cicin-Sain, L., S. Smyk-Pearson, N. Currier, L. Byrd, C. Koudelka, T. Robinson, G. Swarbrick, S. Tackitt, A. Legasse, M. Fischer, D. Nikolich-Zugich, B. Park, T. Hobbs, C. J. Doane, M. Mori, M. K. Axthelm, D. A. Lewinsohn, and J. Nikolich-Zugich. 2010. Loss of naive T cells and repertoire constriction predict poor response to vaccination in old primates. *J Immunol* 184: 6739-6745.

21. Cicin-Sain, L., J. D. Brien, J. L. Uhrlaub, A. Drabig, T. F. Marandu, and J. Nikolich-Zugich. 2012. Cytomegalovirus infection impairs immune responses and accentuates T cell pool changes observed in mice with aging. *PLoS Pathog* 8: e1002849.

22. Robins, H. S., S. K. Srivastava, P. V. Campregher, C. J. Turtle, J. Andriesen, S. R. Riddell, C. S. Carlson, and E. H. Warren. 2010. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* 2: 47ra64.

23. Manley, T. J., L. Luy, T. Jones, M. Boeckh, H. Mutimer, and S. R. Riddell. 2004. Immune evasion proteins of human cytomegalovirus do not prevent a diverse CD8+ cytotoxic T cell response in natural infection. *Blood* 104: 1075-1082.

24. Wolfl, M., J. Kuball, W. Y. Ho, H. Nguyen, T. J. Manley, M. Bleakley, and P. D. Greenberg. 2007. Activation-induced expression of CD137 permits detection, isolation, and expansion of the full repertoire of CD8+ T cells responding to antigen without requiring knowledge of epitope specificities. *Blood* 110: 201-210.

25. Khan, N., A. Hislop, N. Gudgeon, M. Cobbold, R. Khanna, L. Nayak, A. B. Rickinson, and P. A. Moss. 2004. Herpesvirus-specific CD8 T cell immunity in old age: cytomegalovirus impairs the response to a coresident EBV infection. *Journal of immunology (Baltimore, Md. : 1950)* 173: 7481-7489.

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

27. Chidrawar, S., N. Khan, W. Wei, A. McLarnon, N. Smith, L. Nayak, and P. Moss. 2009. Cytomegalovirus-seropositivity has a profound influence on the magnitude of major lymphoid subsets within healthy individuals. *Clinical and experimental immunology* 155: 423-432.

28. Wills, M. R., A. J. Carmichael, M. P. Weekes, K. Mynard, G. Okecha, R. Hicks, and J. G. Sissons. 1999. Human virus-specific CD8+ CTL clones revert from CD45ROhigh to CD45RAhigh in vivo: CD45RAhighCD8+ T cells comprise both naive and memory cells. *Journal of immunology (Baltimore, Md. : 1950)* 162: 7080-7087.

29. Ouyang, Q., W. M. Wagner, W. Zheng, A. Wikby, E. J. Remarque, and G. Pawelec. 2004. Dysfunctional CMV-specific CD8(+) T cells accumulate in the elderly. *Exp Gerontol* 39: 607-613.

30. Akbar, A. N., and J. M. Fletcher. 2005. Memory T cell homeostasis and senescence during aging. *Current opinion in immunology* 17: 480-485.