Learning signatures of functional responses from immune repertoires

Zachary Montague

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Armita Nourmohammad, Chair
Paul Wiggins
Frederick A. Matsen

Program Authorized to Offer Degree:
Physics

University of Washington

Abstract

Learning signatures of functional responses from immune repertoires

Zachary Montague

Chair of the Supervisory Committee:
Armita Nourmohammad
Department of Physics

The adaptive immune system is a diverse, complex system that detects, eradicates, and stores memories of the multitude of evolving pathogens encountered over a lifetime. It is composed of T cells and B cells whose receptors are generated stochastically from a large sample space via genomic recombination and selection. Upon activation, B cells can diversify their receptors further through rapid, Darwinian evolution. Though high-throughput immune repertoire sequencing technologies have enabled the unparalleled sampling of millions of adaptive immune receptor sequences, the resulting datasets are still highly sparse with regard to an individual's immune repertoire as well as the universe of receptors that can be generated. Leveraging principled, multiscale statistical analyses robust to undersampled systems, I investigate the dynamics and differential features of adaptive immune receptors likely to be associated with functional responses to COVID-19 and post-acute sequelae of COVID-19. Moreover, I propose a method to elucidate the role of memory B cells in secondary infections by introducing a stochastic telegraph process to model the lifecycle of B cell lineages. In doing so, I provide a means to reconstruct time-resolved evolutionary histories of B cells and enable a widescale characterization of their effective rates of phenotypic switching.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The adaptive immune system is an organ system present in all jawed vertebrates that confers protection against invaders and malfunctioning or unhealthy cells. Its composition is complex and extraordinarily dynamic relative to a majority of other organ systems in the human body. The surface receptors of constituent cells are generated in a stochastic manner from a large sample space, and a compartment of this system has the ability to evolve and mutate its genetic code as quickly as rapidly evolving pathogens [135]. Through this breadth and plasticity, the adaptive immune system has the potential to recognize and eliminate a multitude of threats. Understanding the principles which shape the adaptive immune system and characterizing the actors of the adaptive immune system in the context of a threat are crucial for paving the way for ameliorating human health outcomes and treatments.

The immune repertoire is the collection of realized immune cells in an organism. Its most basal units are a subset of lymphocytes called T cells and B cells, which process and encode information in a local manner. Each T and B cell possesses its own specialized immune receptors, coined T cell receptors and B cell receptors, by which it binds and interacts with pathogenic molecules [135]. Lineages of cells which were activated and participated in mediating an immune response may persist as memory in the organism for months and up to an entire lifetime, providing efficient protection against past encounters and similar future invaders [54]. Theoretical studies have shown that this decentralized encoding of information and memory is an optimal strategy for identifying evolving patterns, providing evidence as to why the adaptive immune system takes this form [263, 264]. When mounting an immune response, the microscale units of the adaptive immune repertoire, B cell and T cells, along with the innate, i.e., static and fast-acting, immune system work in concert, giving rise to

multiscale, collective behavior.

Roughly 10^{12} cells constitute the human immune repertoire, with 5×10^{11} in the T cell repertoire and 3×10^{11} in the B cell repertoire [268]. High-throughput sequencing technologies within the last two decades has ushered in the collection of large datasets of partially sequenced immune receptor repertoires [29, 91, 277, 251, 77, 31, 140]. And yet, the coverage of a typical sample represents a meager 0.0001% of an individual's repertoire at a single snapshot in time. More recent advances have demonstrated the possibility of even higher sequencing depths with fully sequenced receptors [232]. In principle, the immune repertoire contains information about ongoing ailments in addition to the history of previous encounters an individual has had over the course of their lifetime. Extracting meaningful information from these datasets, however, remains an ongoing challenge.

Though the number of cohorts sampled and the depth of datasets has grown, immune repertoire datasets are highly undersampled with regard to an individual's entire repertoire and with regard to the entire space of realizable immune receptors. Despite many active efforts to utilize sequence and protein structure information to ascertain immune receptor specificity, there is no complete mapping from immune receptor sequences to their target antigens [180, 38, 336, 207, 228, 316, 347, 138, 88, 164, 318, 192, 126, 261, 68, 306, 114, 208]. Databases [275, 98, 297, 216] have been constructed which contain experimental results from dextramer sorting [49], tetramer sorting [71], IFN-γ secretion assay [171], multiplex identification of antigen-specific T cell receptors assay (MIRA) [152], etc., which clarify and validate the function of observed immune receptors. However, it is infeasible to apply such experiments to every receptor observed, and these experiments require knowing the full receptor sequence to then synthesize and test them, which most repertoire datasets currently lack. The complexity of this task is further underscored by the mapping being highly degenerate. For example, a single T cell receptor has been reported to bind to one million different antigens [327], i.e., is highly cross-reactive, and many different B cell receptors have been found to be specific and bind to the same SARS-CoV-2 epitope [50]. Specificity alone, too, does not determine the role an adaptive lymphocyte will play when a response is mounted. T cells

with different receptor sequences but which can bind to the same antigen attain different fates depending on their avidity, i.e., the strength of their specificity [165, 166]. The relationship between how the repertoire is organized with respect to specificity and cross-reactivity is not fully understood, though strides have been made phenomenologically with regard to B cells [262, 45]. Moreover, T cells and B cells in their respective lineages follow many possible differentiation pathways that ultimately lead to starkly different fates [135, 123, 2]; the factors dictating the casting of these roles at this level, too, is not well understood.

Like many other developments in history wherein theory has followed novel technology or observations in experiments [134, 86], a revolution has taken place in quantitative immunology to describe immune repertoire data collected from these next-generation high-throughput sequencing experiments [10]. Both biophysical mechanistic models and deep learning models have been used to reveal insight from immune repertoires. This thesis is a continuation of this work. I employ methods from physics and statistical inference to understand signatures of functional responses to acute and chronic diseases. I examine the relationship between the severity of COVID-19 and the B cell response in chapter 2. In chapter 3, I investigate differences in the T cell repertoires among those who had only COVID-19 and those who went on to develop long COVID. Furthermore, in chapter 4, I derive a mathematical model inspired by statistical physics and queueing theory and apply it to phylogenetic inference with a goal of characterizing the presence and timescales of B cell phenotypic switching in individuals with untreated HIV infections. Finally, I conclude my observations and detail future directions in chapter 5. For the rest of this chapter, I proceed by describing the key components of the adaptive immune system.

1.1 An overview of T cells and B cells

The role of T cells in the adaptive immune system is to survey and patrol cells native to an organism. Fragments of proteins, called peptides, are produced by the degradation of a cell's internal proteins and give a window into how the cell is functioning. Cells present peptides on their surface via proteins called major histocompatibility complexes (known as human

leukocyte antigens (HLAs) in humans), forming peptide-MHC complexes. T cells interact with cells' peptide-MHC complexes with their T cell receptors. Cytotoxic CD8+ T cells kill cells that present peptides unexpected to originate from healthy cells, e.g., from cancerous or virus-infected cells. Helper CD4+ T cells, on the other hand, act mainly to support various immune functions. They release signaling molecules called cytokines to help orchestrate immune responses, prevent autoimmune responses, and mediate B cell maturation [135].

B cells through their B cell receptors and secretion of antibodies, the free, soluble form of their B cell receptors, enable the immune system to detect any molecules, such as those on foreign sources like bacteria, fungi, and parasites. B cell receptors and antibodies bind to a threat's surface receptors, neutralizing their ability to harm cells. Moreover, bound antibodies flag threats in a process called opsonization, enabling innate immune system cells to dispose of invaders more efficiently [178] (Fig 1.1). Antibodies also regulate inflammation and modulate the recruitment and priming of innate immune system cells to attack [11].

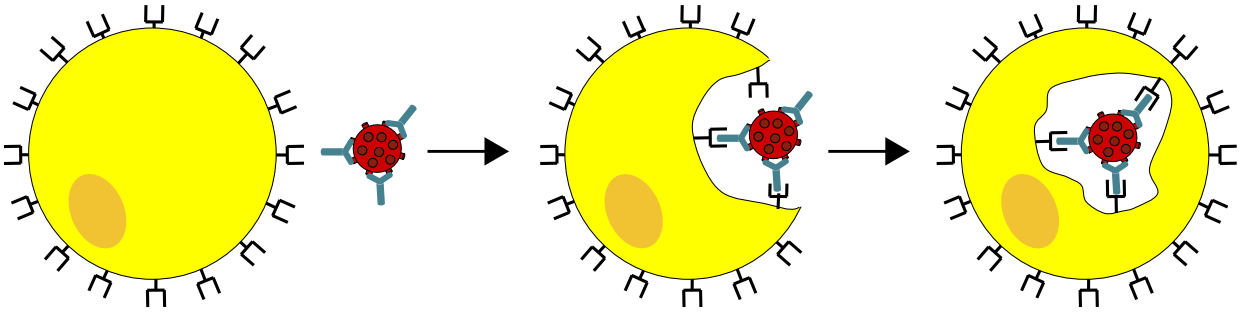


Figure 1.1: **Antibody opsonization.** Cell membranes are typically negatively charged [190] while antibodies appear to prefer a net positive charge [317]. By binding to an invader's surface receptors, antibodies alert the more "blind" innate immune system cells, such as a phagocyte, that a threat is present and provide a bridge to what otherwise would have been an unfavorable interaction. The phagocyte can then engulf and dispose of the threat [178]. Image adapted from wikimedia user Maher33 under CC-BY-SA 4.0 license.

1.2 Repertoire formation

T cells and B cells begin their existence in the bone marrow as pluripotent hematopoietic stem cells from which all blood cells are derived [79, 246]. After undergoing the necessary differentiation steps which tune transcription factor expression in a process known as hematopoiesis, hematopoietic stem cells may become T cell or B cell progenitors. T cell progenitors migrate from the bone marrow to the thymus while B cell progenitors remain in the bone marrow.

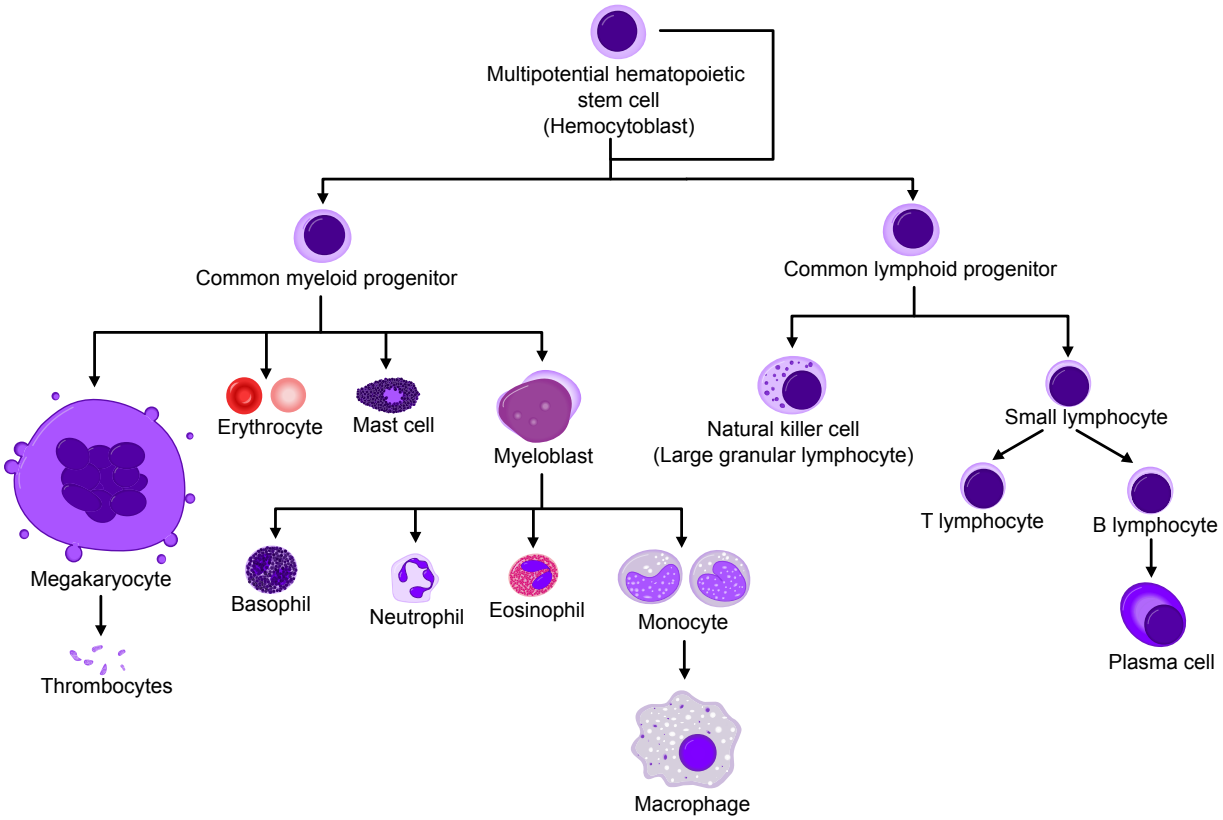


Figure 1.2: **Hematopoiesis.** Hematopoiesis is the process by which blood cells are made. This flow chart depicts the various differentiation pathways that can be taken and includes the creation of T and B cells, the cells forming the adaptive immune system. Image reproduced from wikimedia users A. Rad and M. Häggström under CC-BY-SA 3.0 license.

1.2.1 V(D)J recombination

Lymphocyte progenitors undergo a process called V(D)J recombination to create their surface receptors that impart their vital function. T and B cell receptors are constructed from two chains: T cell receptors have an α and a β chain whereas B cell receptors have a κ or λ and a heavy chain. The potentially more diverse chains, the β chain for T cells and heavy chain for B cells, are believed to more strongly constrain what molecules the receptor can sense; therefore, these chains are used to identify and study receptors, cells, and the repertoires at-large. Along the cell's genome are key regions containing genomic templates called the variable (V), diversity (D), and joining (J) genes for the β and heavy chains and solely V and J genes for the others. The number of genomic templates depends on the species of the organism, lymphocyte type, and chain.

V(D)J recombination begins by selecting one of each of these segments in a nonuniform manner. The chosen V and D segments and the D and J segments are then spliced together by deleting nucleotides and inserting palindromic and non-templated nucleotides at the junctions where they will be joined. The number of deletions and insertions can span anywhere from 0 to more than 10 nucleotides for deletions and 20 nucleotides for insertions [206, 184] (Fig. 1.3). The resulting potential diversity of T cell receptors in humans from recombination alone is estimated to be anywhere from 10^{15} to 10^{61} [60, 201].

Due to the random nature of V(D)J recombination, about 60% to 90% of recombinations result in a nonfunctional (unproductive) rearrangement in which the genome contains a stop codon; is out-of-frame, i.e., the number of nucleotides which encode the receptor is not divisible by three; or uses a pseudogene segment, which are nonfunctional V, D, or J genes. (Estimates were obtained from using IGoR [184].) In this instance, the sister chromosome is then V(D)J recombined, and, if a successful rearrangement is produced, the progenitor survives and proceeds to mature through selection. Allelic exclusion guarantees that only one productive V(D)J product will be expressed [90, 113]. Consequently, the unproductive rearrangement can hitchhike. While the productive rearrangement will later be subject

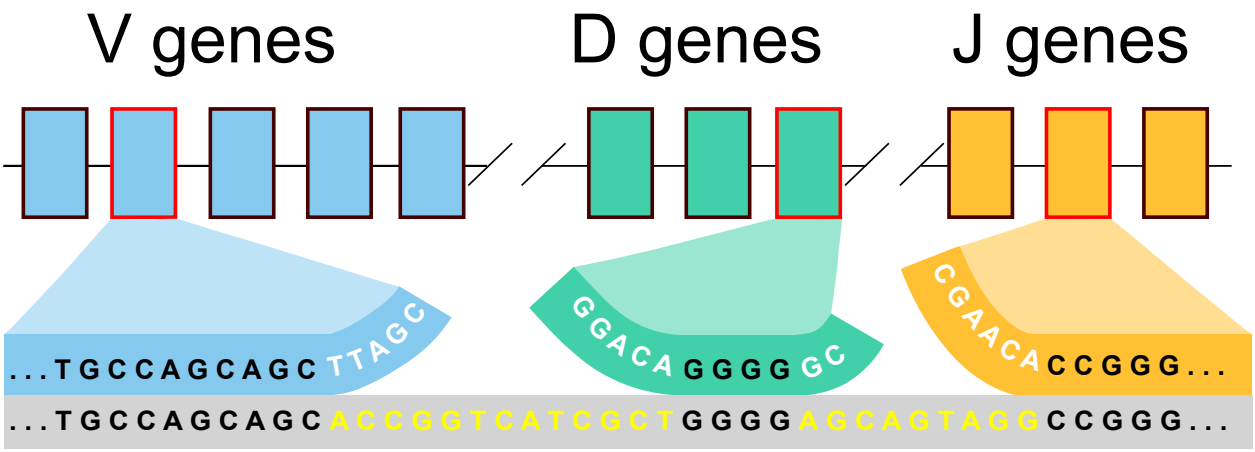


Figure 1.3: **VDJ recombination.** Each of the variable (V), diversity (D), and joining (J) genes are chosen at random from templates along the genome (red boxes). VDJ recombination then trims a stochastic amount of nucleotides from each gene at the two junctions (white characters) and pastes them together by inserting non-templated nucleotides at random (yellow characters). Adapted from [206, 64].

to tolerance and antigenic selection, the unproductive arrangement will not. Therefore, unproductive repertoires of B and T cells can be investigated to characterize the statistics and mechanistic principles governing their respective V(D)J recombination [206, 184].

The receptor chains produced in V(D)J recombination can be divided into seven segments which alternate between two types along the genome. Three complementarity determining regions (CDRs) are responsible for binding to antigens. The first two CDRs, CDR1 and CDR2, are along the V gene. The third CDR, CDR3, is the portion of the receptor at which V(D)J combination took place. Flanking these CDRs on both sides are four framework regions that provide structural support for the CDRs. All together, this is referred to as the variable region of the receptor (Fig. 1.4).

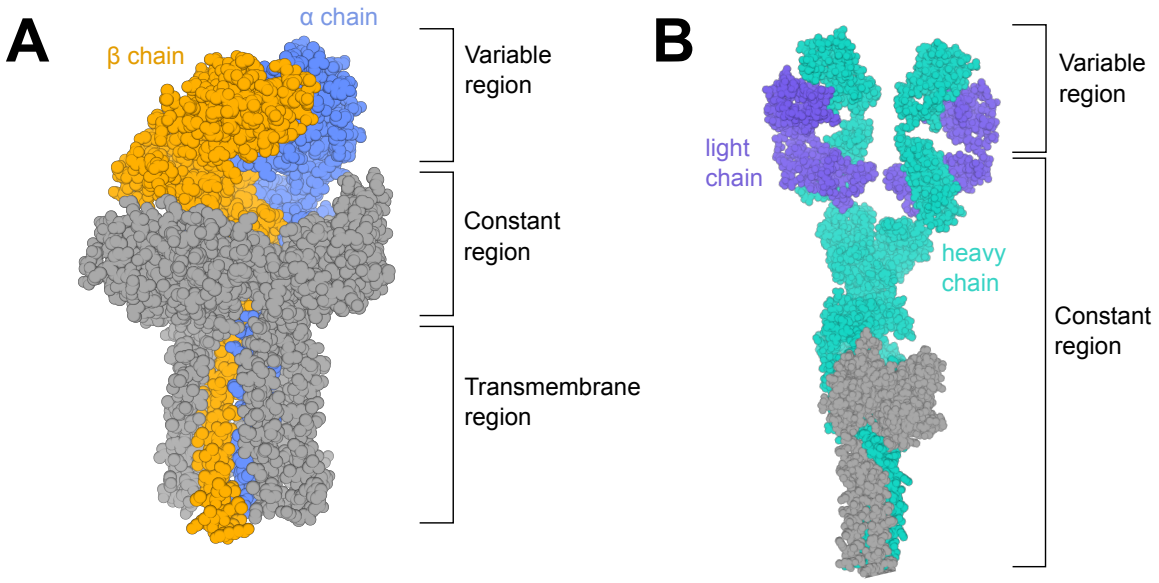


Figure 1.4: **T cell receptor and antibody structures. (A)** A T cell receptor consists of an α (blue) and β chain (orange). The variable region contains the outcome of V(D)J recombination—the four framework regions and three CDR regions—and interacts directly with antigens. The constant region links the α and β chain together, and the transmembrane region anchors the receptor to the cell. **(B)** The structure of an antibody is shown with its heavy chain colored in teal and light chain colored in purple. Observe its characteristic Y-shape. The structures illustrated here are not exactly to scale, though they are highly similar in size. A T cell receptor is about 11.5 nm long, with 7.5 nm of the receptor being extracellular [70]. An antibody is approximately 10 nm long [249].

1.3 Repertoire refinement and maturation

1.3.1 T cell tolerance and maturation

After V(D)J recombination, T cells undergo thymic selection. Thymic selection begins by subjecting T cells to positive selection for their binding to self-antigens as well as to MHC molecules of the host, leaving approximately 5% of functional recombinations [16]. MHCs are extremely polymorphic within species [135] while the parameters of V(D)J recombination appear highly consistent across individuals [269], so it is not guaranteed that receptors will bind to an individual's pMHCs. At this stage of selection, T cells also differentiate into CD4+ or CD8+ cells, with a cell's fate being highly determined by its receptor sequence [131, 165, 46]. After differentiating, T cells undergo negative selection in which they are tested for binding against self-antigen. T cells which bind too strongly undergo apoptosis, mitigating downstream autoimmune responses. In principle, T cells which have low affinity for self-antigens, but not too low, enter in the periphery. The efficacy of this negatively selection, however, is believed to be subpar in eliminating all self-reactive T cells [202]. For example, T cells spend only a few days in the thymus and, therefore, can probe just a small fraction of the self-antigens [36]. T cells that have successfully passed thymic selection enter the periphery.

T cells which have yet to trigger an immune response, i.e., naive T cells, may persist in the periphery for years [314]. Upon recognizing a cognate antigen, activated T cells proliferate, increasing their population by many folds, and can differentiate into effector cells to participate actively in an immune response or into quiescent memory that can be easily reactivated upon similar encounters. After the infection has cleared, effector T cell populations contract, leaving memory T cell lineages that can be maintained for entire lifetimes [167, 181, 160].

1.3.2 B cell tolerance and maturation

Central tolerance prevents B cells with ineffective and, to a degree, self-reactive B cell receptors from leaving the bone marrow and entering the periphery [210]. If a B cell binds poorly to self-antigens, it may undergo receptor editing by which it may rearrange its light chain and be tested again, or it may more immediately undergo apoptosis. Like T cells, B cells are subject to central tolerance for only a few days over which it is not possible to scan and be tested against the entire library of self-antigens. Inevitably, many autoreactive B cells enter into the periphery [92]. However, as in the case for T cells, many tolerance mechanisms exist in the periphery and secondary lymphoid organs to prevent autoimmune responses.

B cells which successfully pass central tolerance migrate to secondary lymphoid organs such as lymph nodes and the spleen [135]. There, with a half-life of two months [141], they await activation to their cognate antigens [6]. Activated B cells which bind their cognate antigen validate the pathogenic nature of the antigen with CD4+ T cells and proliferate if the antigen does, indeed, have harmful origins. From there, they have three avenues forward (Fig. 1.5). They can differentiate into plasma B cells—cells which can secrete thousands of antibodies per second [278]—that typically exist only over the course of the infection, so-called short-lived plasma cells [28]. At this stage, plasma B cells' receptors and secreted antibodies have low affinity for their antigens and are not the most effective mechanism for clearing for the pathogens; however, mounting a quick response is imperative to buying the adaptive immune system time to mature and for collecting antigens to mediate B cell evolution. Activated B cells can also differentiate into low-affinity memory B cells for rapid responses for future or variant encounters. The third pathway activated B cells can take is to form microanatomical structures called germinal centers and improve their binding to their cognate antigens via a process called affinity maturation [311].

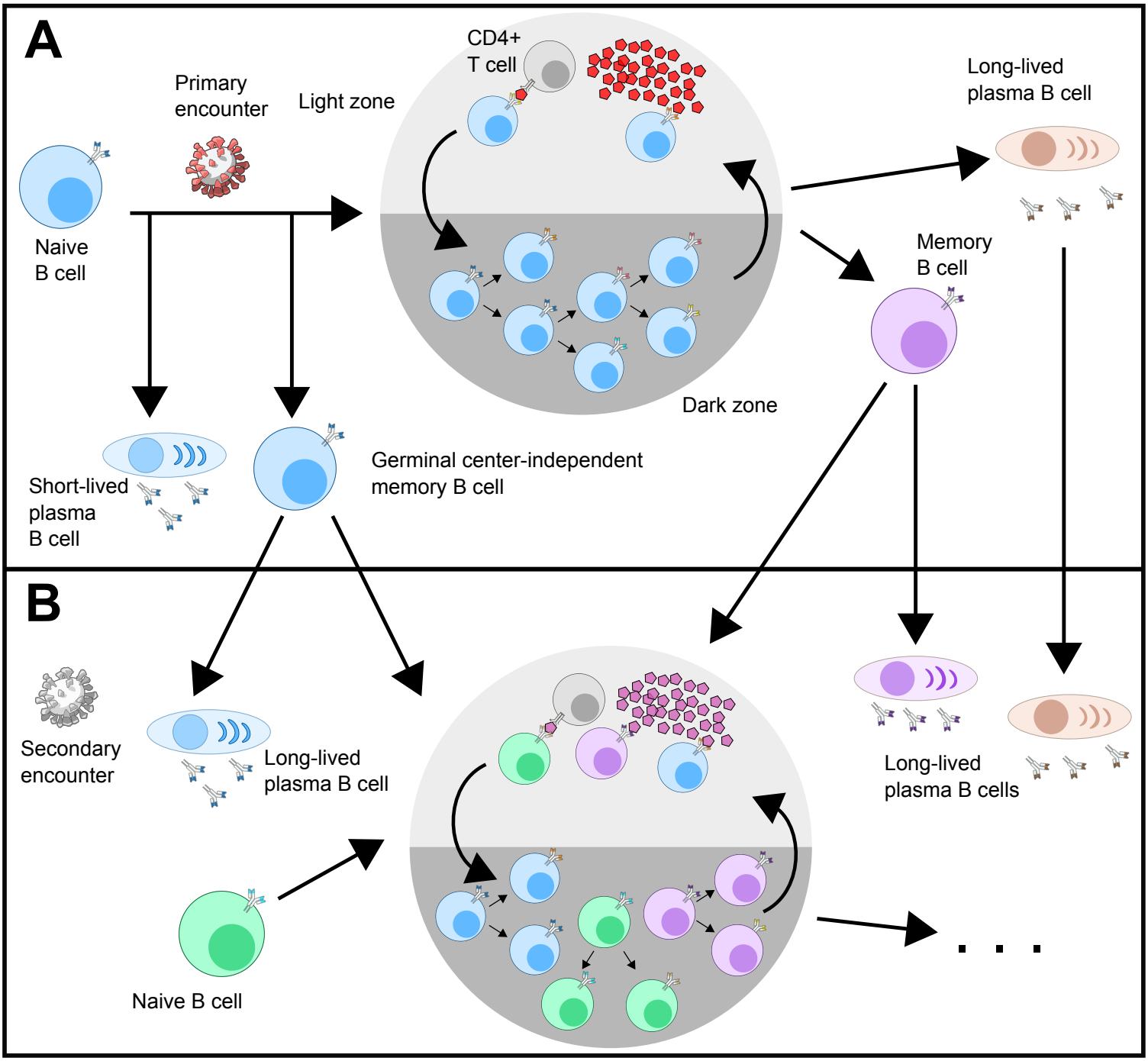


Figure 1.5: **B cell fates in primary and secondary responses. (A)** When a primary infection occurs, naive B cells can differentiate into short-lived plasma B cells, memory B cells with low affinity, or germinal center B cells.

Figure 1.5: (Previous page.) Germinal center B cells undergo affinity maturation where they proliferate and somatically hypermutate their receptors in the dark zone and are selected for increased binding to cognate antigens in the light zone. B cells which survive affinity maturation may differentiate into long-lived plasma B cells or memory B cells. **(B)** When a secondary infection occurs, long-lived plasma B cells increase antibody production, and memory B cells may differentiate into plasma cells to secrete antibodies too. Newly generated naive B cells may become activated and have germinal center-independent fates (not shown, as in (A)). Both germinal center-independent and dependent memory cells may reseed germinal centers, competing with each other as well as naive B cells. Memory B cells may also not be recalled, and only naive B cells may undergo affinity maturation (not shown, as in (A)). Illustration, in part, from NIAID NIH BIOART Sources (bioart.niaid.nih.gov/bioart/173, bioart.niaid.nih.gov/bioart/464, bioart.niaid.nih.gov/bioart/510, bioart.niaid.nih.gov/bioart/17).

1.3.3 B cell evolution

Affinity maturation is a Darwinian evolutionary process in which B cells mutate their B cell receptors, proliferate rapidly, are selected for enhanced binding to the presented antigens, and compete with other B cells [311]. Remarkably, B cells mutate approximately 10^6 times faster than the baseline rate of mutation in the body, and their cell cycles are as quick as six hours during affinity maturation, about four times faster than the typical human cell. During affinity maturation, B cells can potentially increase their affinity 100- to more than 10^5-fold [4]. However, the affinity maturation process exhibits highly stochastic behavior [15, 293], and, though recent experiments have shown that the median affinity of B cell lineages increases on average [15], it is possible that improvements in affinity may not even transpire [161].

The germinal center is organized into a "light" zone and a "dark" zone, reflecting the population density of B cells in those regions (Fig. 1.5). B cells clonally expand in the dark

zone and mutate their receptors in a sequence context- and position-dependent manner [283]. In the light zone, the affinity of their B cell receptors is evaluated by CD4+ T cells on antigens collected over the course of the immune challenge. Thus, B cells are competing against other B cells for limited CD4+ T cell and antigen resources. Positively selected B cells can undergo further cycling between the light and dark zones.

Germinal centers can last from weeks to several months [302, 186, 169, 149]. While B cell lineages that established long-lived germinal centers may accrue mutations and increase their affinity over the entire existence of a germinal center, experimental observations have shown that lineages unrelated to the immune challenge enter and compete with these founder lineages [61]. Characterizing the timescales of evolution in germinal centers and how they are influenced by immune challenges and population variability remains an open problem.

1.3.4 B cell differentiation

B cells which survive affinity maturation exit germinal centers and differentiate into memory or long-lived plasma cells, the latter taking lodge in the bone marrow. Experimental observations suggest B cells more likely to differentiate into plasma cells have higher affinity, and memory B cells have lower affinity [311]. This strategy permits the immune system to clear the ongoing immune challenge efficiently and prime it for similar, but not identical, future encounters [262]. Unlike their naive counterparts, both of these cells are non-proliferative and can persist for timescales that span months up to a lifetime [148, 141]. When a previously encountered immune challenge arises again, memory B cells can be reactivated. They can differentiate into plasma cells for fast clearance of invaders and facilitate more enhanced antigen collection for germinal center responses. Additionally, memory B cells can seed or enter germinal centers, competing with naive B cells, and evolve their lineage further (Fig. 1.5). The degree to which this occurs across the distribution of immune challenges and what controls memory B cell recall into germinal centers is poorly understood [69, 224, 188, 326, 191, 309, 356, 76, 14, 13, 303]. Contributing to shedding light on memory B cell recall and secondary germinal center formation, I showcase in chapter 4

a mathematical inference framework to characterize B cell lineage engagement in and the effective timescales of secondary germinal center responses.

Chapter 2

DYNAMICS OF B CELL REPERTOIRES AND EMERGENCE OF CROSS-REACTIVE RESPONSES IN COVID-19 PATIENTS WITH DIFFERENT DISEASE SEVERITY

The content of this chapter was been published in:

Montague Z, et al. "Dynamics of B cell repertoires and emergence of cross-reactive responses in patients with different severities of COVID-19." Cell Reports 35.8 (2021).

Individuals with the 2019 coronavirus disease (COVID-19) show varying severity of the disease, ranging from asymptomatic to requiring intensive care. Although monoclonal antibodies specific to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have been identified, we still lack an understanding of the overall landscape of B cell receptor (BCR) repertoires in individuals with COVID-19. We use high-throughput sequencing of bulk and plasma B cells collected at multiple time points during infection to characterize signatures of the B cell response to SARS-CoV-2 in 19 individuals. Using principled statistical approaches, we associate differential features of BCRs with different disease severity. We identify 38 significantly expanded clonal lineages shared among individuals as candidates for responses specific to SARS-CoV-2. Using single-cell sequencing, we verify the reactivity of BCRs shared among individuals to SARS-CoV-2 epitopes. Moreover, we identify the natural emergence of a BCR with cross-reactivity to SARS-CoV-1 and SARS-CoV-2 in some individuals. Our results provide insights important for development of rational therapies and vaccines against COVID-19.

2.1 Introduction

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the 2019 coronavirus disease (COVID-19), has now spread to 223 countries and caused more than 143 million infections with a mortality rate around 2.2% (World Health Organization, 2021). Individuals with COVID-19 show varying disease severity, ranging from asymptomatic to requiring intensive care. Although epidemiological and clinical data report that many factors, such as age, gender, genetic background, and preexisting conditions, are associated with disease severity, host immunity against virus infection is the crucial component of controlling disease progression [101, 102, 189, 304, 330]. Shedding light on signatures of a protective immune response against SARS-CoV-2 infection can help elucidate the nature of COVID-19 and guide therapeutic agent development as well as vaccine design and assessment.

Adaptive immunity is considered one of the core protective mechanisms of humans against infectious diseases. A vast diversity of surface receptors on B and T cells enables us to recognize and counter new or repeated invasion from a multitude of pathogens [136, 213]. In particular, antibodies produced by B cells can provide long-lasting protection against specific pathogens through neutralization or other antibody-mediated immune mechanisms [136]. During the early phase of an infection, antigens of a pathogen are recognized by a group of naive B cells, which then undergo affinity maturation in a germinal center through somatic hypermutation and selection. The B cell receptors (BCRs) of mature B cells can react strongly to infecting antigens, resulting in B cell stimulation, clonal expansion, and, ultimately, secretion of high-affinity antibodies in the blood [33, 34, 55]. The specificity of a BCR is determined by a number of features, such as V, D, or J gene usage and length and sequence composition of the HCDR3 region. SARS-CoV-2-specific immunoglobulin G (IgG) antibodies can be detected in plasma samples of individuals with COVID-19 starting from the first week after symptom onset [227]. These antibodies bind to different antigens, including the spike protein and nucleoprotein as well as other structural or non-structural proteins [105]. In addition, multiple studies have isolated SARS-CoV-2-specific B cells from

individuals with COVID-19 and determined their germline origin [21, 32, 41, 47, 107, 108, 128, 142, 156, 157, 176, 218, 250, 252, 270, 271, 273, 332, 343, 351]. However, we still lack a comprehensive view of individuals' entire BCR repertoires during SARS-CoV-2 infection.

Antibody repertoire sequencing has advanced our understanding of the diversity of adaptive immune repertoires and their response to pathogens [29, 91, 155, 251]. A few studies have performed BCR repertoire bulk sequencing to characterize the statistical signatures of the immune response to SARS-CoV-2 [87, 214, 215, 266]. However, these studies have limited data regarding the dynamics of BCR repertoires, which could provide significant insight into responses specific to the infection. Moreover, they do not probe the composition of plasma B cells during infection, which is the direct indicator of antibody production in an individual.

In this study, we established a principled statistical approach to study the statistics and dynamics of bulk and plasma B cell repertoires and to characterize the immune responses in 19 individuals with different severities of COVID-19. By combining information from the statistics of sequence features in BCR repertoires, the expanding dynamics of clonal lineages during infection, and sharing of BCRs among individuals with COVID-19, we identified 38 clonal lineages that are potential candidates for a response to SARS-CoV-2. Importantly, eight of these lineages contain BCRs from the plasma B cell repertoire and, hence, are likely to have been secreting antibodies during infection. Moreover, using single-cell sequencing, we verified the reactivity of BCRs shared among individuals to the epitopes of the receptor-binding domain (RBD) and N-terminal domain (NTD) of SARS-CoV-2. Last, we identified cross-reactive responses to SARS-CoV-1 in some individuals with COVID-19 and a natural emergence of a previously isolated SARS-reactive antibody [231] in three individuals.

2.2 Strong correlation between composition of bulk and plasma B cell repertoires

We obtained total RNA from peripheral blood mononuclear cells (PBMCs) isolated from 19 individuals infected with SARS-CoV-2 and three healthy individuals (Appendix A; Data S1; Table A.1). To broaden our healthy control pool, we also incorporated into our analyses

IgG B cells from 10 individuals in the Great Repertoire Project (GRP) [31]. Sequence statistics for the first three biological replicates pooled together for each individual from the GRP are shown in Data S1 (Appendix A). The individuals with COVID-19 showed different severities of symptoms, forming three categories of infected cohorts: 2 individuals with mild symptoms, 12 with moderate symptoms, and 5 with severe symptoms. Specimens from all but one individual were collected at two or more time points during the course of the infection (Data S1). In addition to the bulk repertoire, we also isolated CD38+ plasma B cells from PBMC samples at at least two time points from seven individuals in this cohort (six moderate and one severe) and from seven additional individuals (two asymptomatic, three mild, and two moderate) and three healthy individuals (Figure 2.1; Data S1). The sampled time points for all individuals in this study are indicated in Figure 2.2 and Data S1. IgG heavy chains of B cell repertoires were sequenced by next-generation sequencing, and the statistics of the collected BCR read data from each sample are shown in Data S1. Statistical models were applied to analyze the length of the HCDR3 region, IGHV or IGHJ gene usage, and expansion and sharing of clonal lineages (Figure 2.2).

The bulk repertoire is a collection of all BCRs circulating in the blood, including receptors from naive, memory, and plasma B cells. Plasma B cells are actively producing antibodies, so their receptors are more likely engaged in responding to an ongoing infection. Interestingly, the abundance of B cell clonal lineages in the bulk and plasma repertoires are strongly correlated (Figure 2.3A), with Pearson correlations ranging from $0.55-0.88$ across individuals and significance $p < 5 \times 10^{-8}$ across individuals; correlations and p values for each individual are shown in Figure 2.3. The significant correspondence between the bulk and plasma B cell repertoires in Figure 2.3 indicates that samples from the bulk, which cover a larger depth, are representative of functional immune responses, at least over the course of the infection.

2.3 B cell repertoires differ in receptor compositions across cohorts

We aimed to investigate whether cohorts with different disease severities can be distinguished by molecular features of their B cell repertoires. Because sequence features of immune

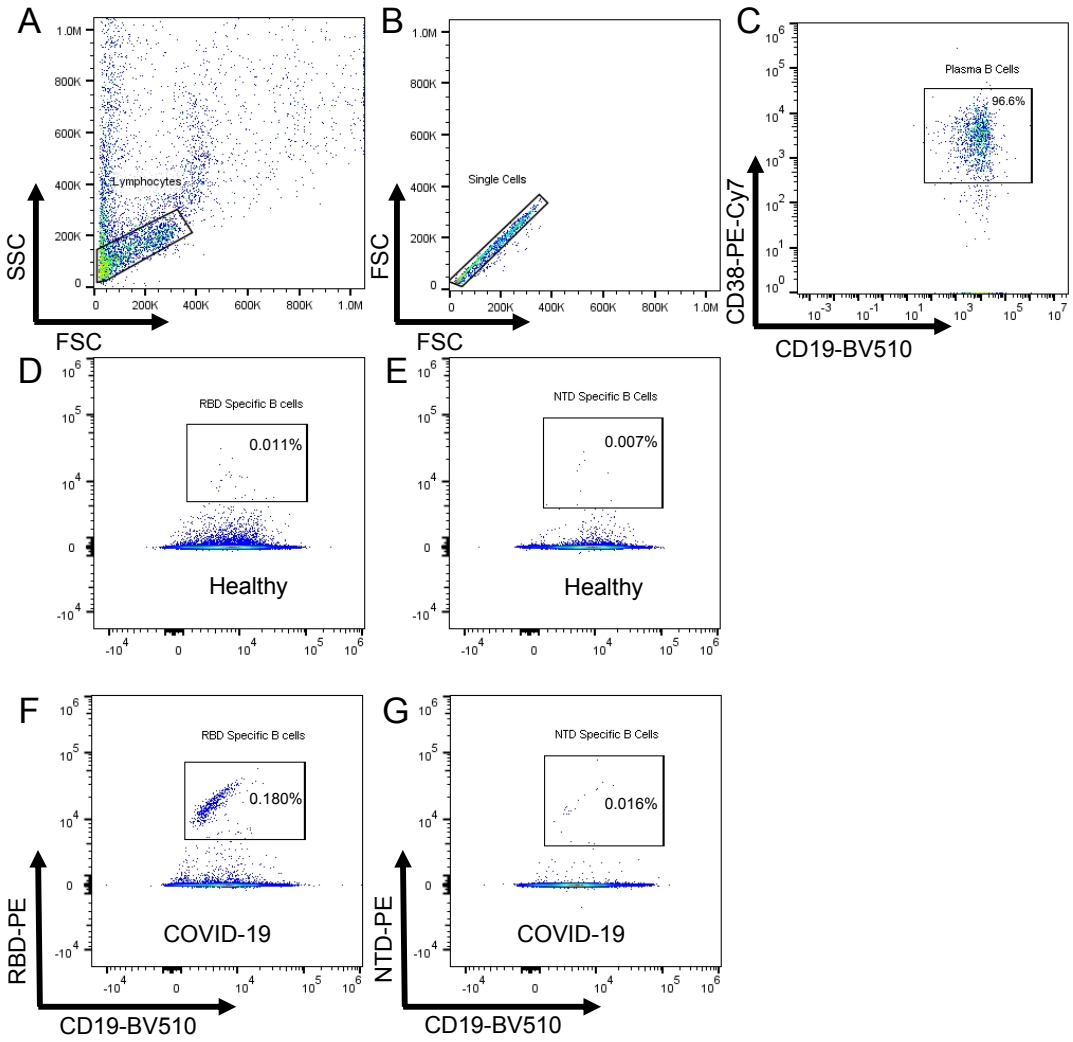


Figure 2.1: **Gating strategy for CD38+ plasma B cells and for RBD- or NTD-specific B cells. (A - C)** Flow cytometry stainings of CD38+ plasma B cells from patients infected with COVID- 19 using two fluorescent markers, anti-human CD19-BV510 (BioLegend) and CD38-PE-Cy7 (BioLegend) in the same tube. Percentages indicate the proportions of CD19+CD38+ plasma B cells within total B cells. **(D-G)** Flow cytometry stainings of RBD- or NTD-specific B cells from healthy donors and individuals infected with COVID-19 using anti-human CD19-BV510 (BioLegend) and PE fluorescent dye for RBD protein (D, F) or NTD protein (E, G) in the same staining tube. Percentages indicate the proportions of CD19+ and RBD- or NTD-protein double positive specific B cells.

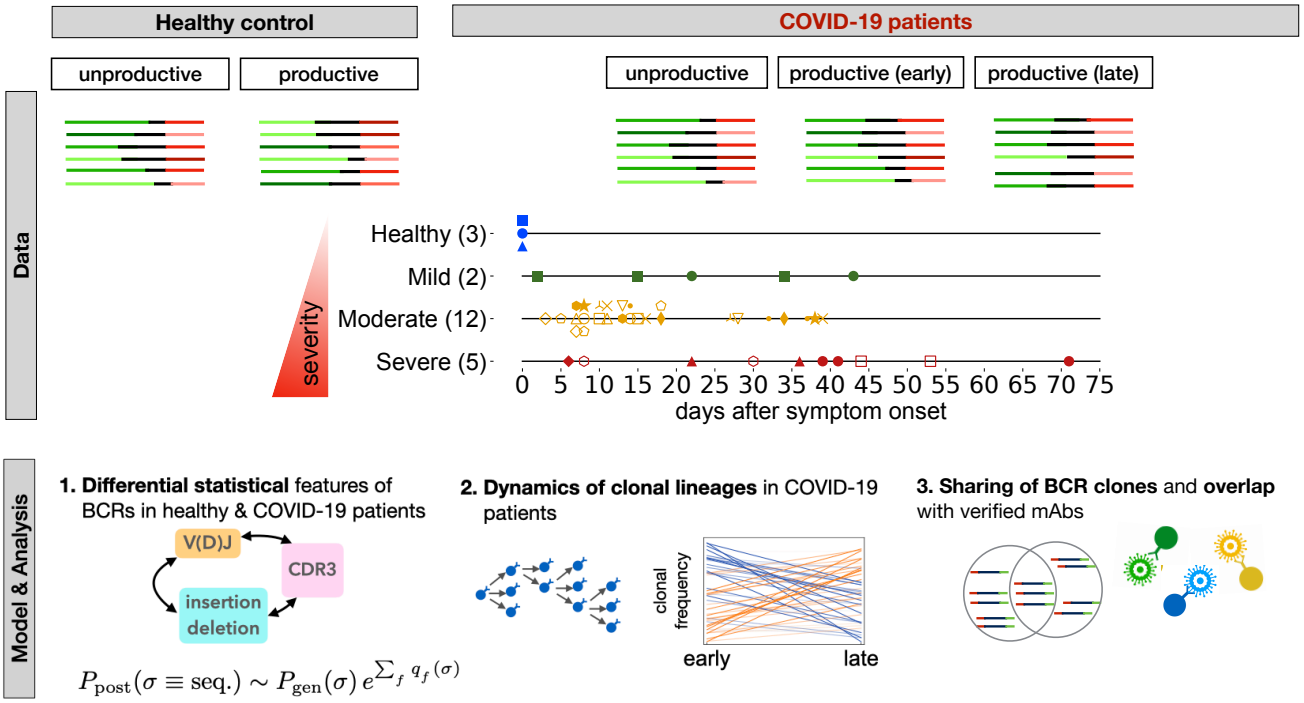


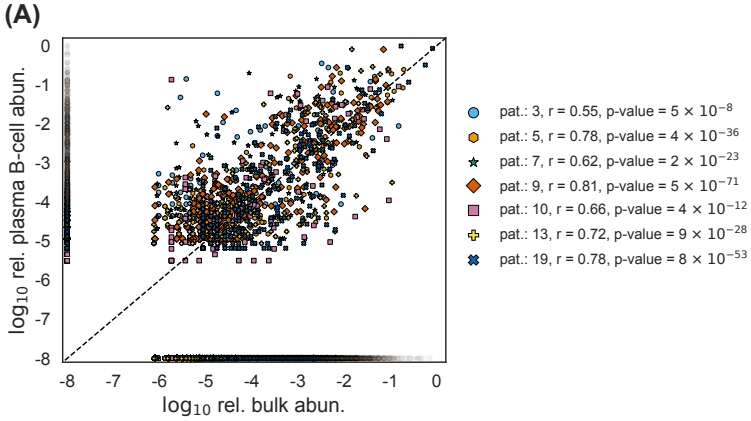
Figure 2.2: **Roadmap for analysis of BCR repertoires. Top:** we collected bulk blood IgG BCR samples from 3 healthy individuals and 2 individuals with mild, 12 with moderate, and 5 with severe symptoms of COVID-19 (different markers and colors). We also collected CD38+ plasma B cells from PBMC samples of 7 individuals in this cohort (6 moderate and 1 severe) and from 7 additional individuals (2 asymptomatic, 3 mild, and 2 moderate) and 3 healthy individuals (Data S1). Samples were collected at different time points during infection (shown in the center for bulk repertoires). We distinguished between productive and unproductive receptors that had frameshifts because of V(D)J recombination. Line segments of varying lengths represent full V(D)J rearrangements (colors). For each individual, we constructed clonal lineages for productive and unproductive BCRs and inferred the naive progenitors of lineages (Appendix A). **Bottom:** (1) Using the set of unproductive inferred naive BCRs, we inferred a model to characterize the null probability for generation of receptors $P_{\text{gen}}(\sigma)$ [184].

Figure 2.2: (Previous page.) We inferred a selection model [269] to characterize the deviation from the null among inferred naive productive BCRs, with the probability of entry to the periphery $P_{\text{post}}(\sigma)$ and selection factors $q_f(\sigma)$ dependent on receptor sequence features. (2) Based on temporal information of sampled BCRs, we identified clonal lineages that expanded significantly during infection. (3) We identified progenitors of clonal lineages shared among individuals and assessed the significance of these sharing statistics based on the probabilities to find each receptor in the periphery. The shared, expanding clonal lineages that contain plasma B cells are likely candidates for secreting responsive antibodies during infection. We verified the reactivity of receptors to SARS-CoV-2 antigenic epitopes using sorted single-cell data. We also identified previously characterized monoclonal antibodies (mAbs) specific to SARS-CoV-2 and SARS-CoV-1.

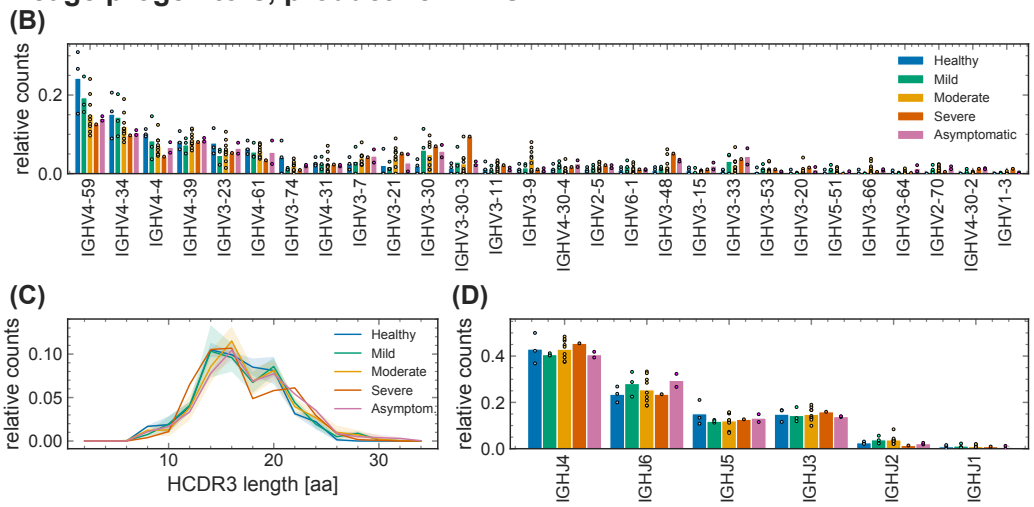
receptors are often associated with their binding specificity, we used statistical Appendix A to compare these features at the level of clonal lineages, including the inferred receptor sequence of lineage progenitors in the bulk (Figures 2.4 and 2.5) and in the plasma B cell repertoires (Figure 2.3) and also the unique sequences in the bulk (Figure 2.5) and in the plasma B cell repertoires (Figure 2.3); see Data S1 for details.

Lineage progenitors of IgG repertoires are closest to the ensemble of naive receptors in the periphery. Features of lineage progenitors reflect receptor characteristics that are necessary for activating and forming a clonal lineage in response to an infection. In particular, the subset of lineages that contain plasma BCRs can signal specific responses for antibody production against the infecting pathogen. Statistics of unique sequences in the bulk and the plasma B cell repertoires, on the other hand, contain information about the size of the circulating lineages. Importantly, these statistical ensembles are relatively robust to PCR amplification biases that directly affect read abundances (Appendix A).

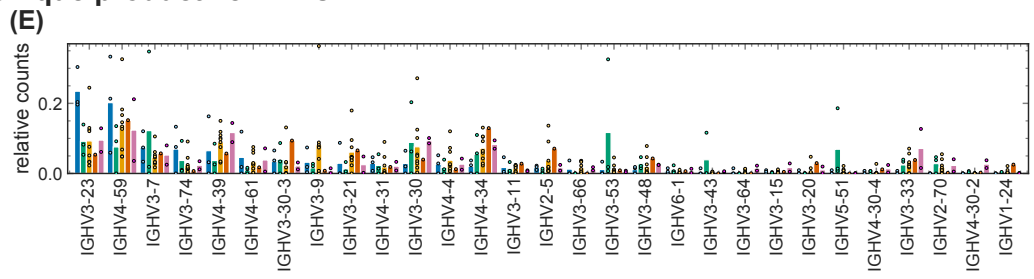
IGHV genes cover a large part of pathogen-engaging regions of BCRs, including the

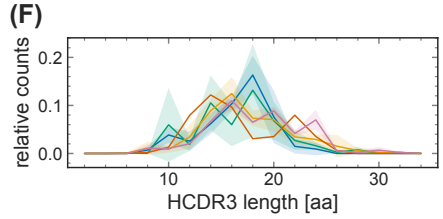


Lineage progenitors, productive BCRs



Unique productive BCRs





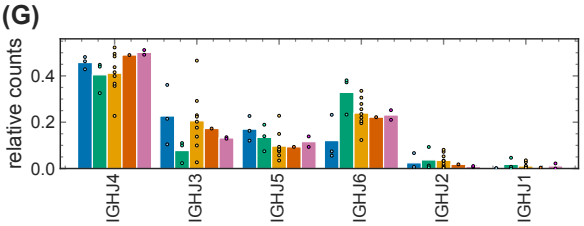


Figure 2.3: **Sequence features of immune receptors in the plasma B cell repertoire across cohorts. (A)** Scatter plot shows $\log_{10}$ relative abundance of clonal lineages constructed from the plasma B cell and bulk repertoire data from all time points and technical replicates in each individual (colors). To avoid primer-specific amplification biases, the relative abundance is estimated as the total read count of a clonal lineage relative to the total reads in the data associated with a specific primer amplification. Lineages with only bulk reads or only plasma reads are displayed as having $\log_{10}$ relative abundance $= 10^{-8}$. Pearson correlations (r) between abundances of lineages which were present in both the bulk and the plasma B cell repertoires and the corresponding p values are indicated in the legend for each individual. **(B - D)** Similar statistics are shown as in Fig. 2 (A,C,D), but for progenitors of clonal lineages with minimum size of three, in which at least one BCR is found in the plasma B cell repertoire data; statistics of these lineages are reported in Data S1. Smaller read counts in the plasma B cell data compared to the bulk do not allow for comparative analysis of receptor statistics across co horts. Markers shown in the histograms indicate statistics for each individual (biological replicate) in a given cohort. **(E - G)** Similar statistics are shown as in Fig. S2 (A - C), but for unique receptors harvested from the plasma B cell repertoires. Statistics of these receptors for each individual is described in Data S1. Smaller read counts in the plasma B cell data compared to the bulk don't allow for comparative analysis of receptor statistics across cohorts. Colors are consistent across panels.

three complementarity-determining regions HCDR1, HCDR2, and a portion of HCDR3. Therefore, we investigated whether there are any differences in V gene usage across cohorts, which may indicate preferences relevant for response to a particular pathogen. We found that the variation in V gene usage among individuals within each cohort was far larger than differences among cohorts in the bulk (Figure 2.4A) and plasma B cell repertoires (Figure 2.3B). Data from unique sequences also indicated large background amplitudes because of

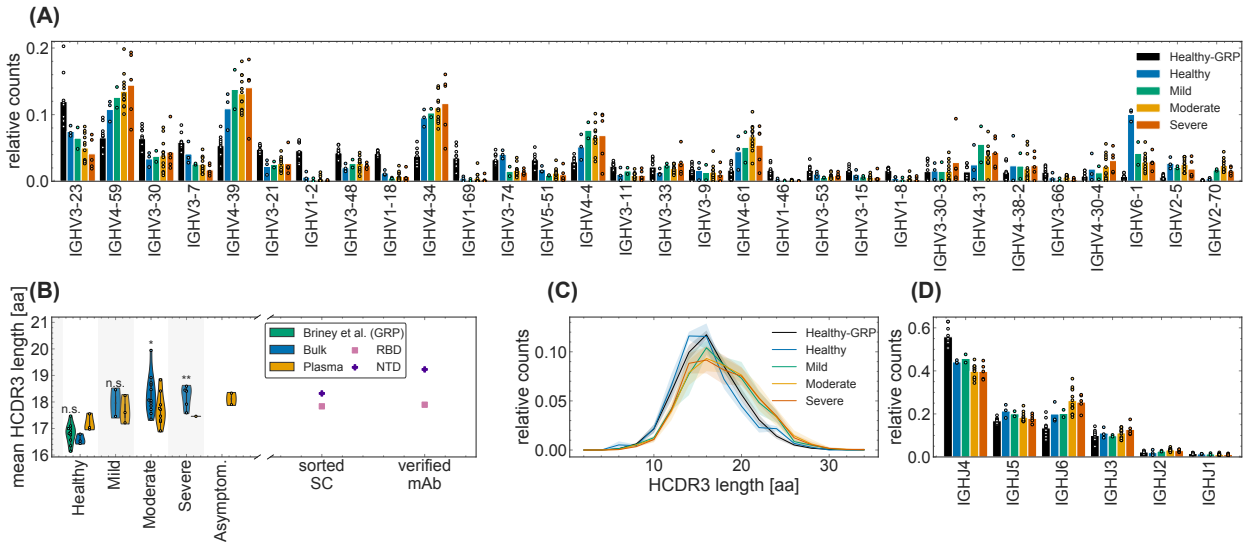
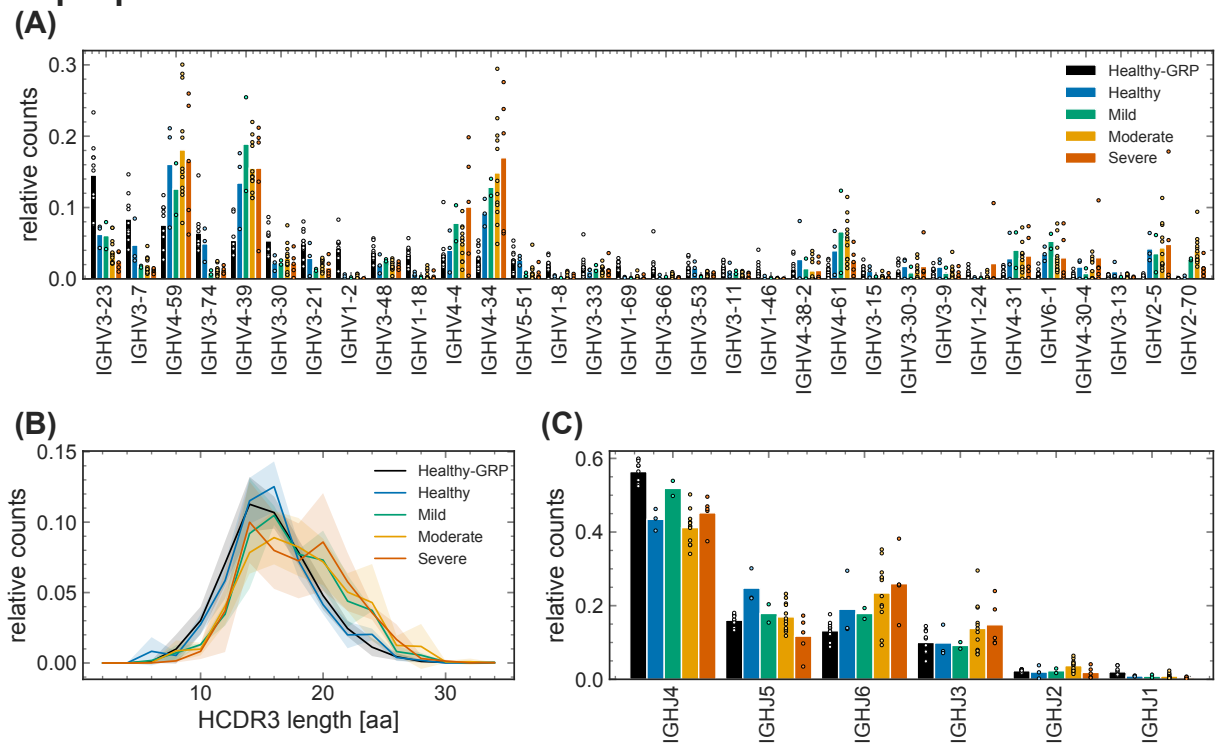


Figure 2.4: **Sequence features of immune receptors in the bulk repertoire across cohorts.** **(A)** The relative counts for IGHV gene usage for inferred naive progenitors of clonal lineages in healthy individuals and those with mild, moderate, and severe COVID-19 symptoms. The bars indicate the usage frequency averaged over individuals in each cohort, and dots indicate the variation in V gene frequencies across individuals (biological replicates) within each cohort. **(B - C)** Statistics of length of HCDR3 amino acid sequence for different individuals (biological replicates) in each cohort. The violin plots in (B) show the mean HCDR3 length of each individual (dots) in a given cohort (color), with the violin plot cut parameter set to 0.1. The mean HCDR3 lengths of the sorted single cells are verified mAbs (axis) for RBD-reactive (pink squares) and NTD-reactive (purple plus symbols) receptors are shown on the right. Full lines in (C) show distributions averaged over individuals (biological replicates) in each cohort (color), and shading indicates regions containing on standard deviation of variation across individuals within a cohort. One-way ANOVA statistical tests were performed, comparing the mean HCDR3 of all COVID-19 cohorts and the healthy repertoires from the Great Repertoire Project (GRP) dataset [31] with the healthy control from this study. Healthy-mild: $F_{1,3} = 12.0$, $p = 0.04$; healthy-moderate: $F_{1,13} = 15.7$, $p = 0.0016$; healthy-severe: $F_{1,6} = 37.5$, $p = 0.00087$; healthy-GRP: $F_{1,11} = 0.9$, $p = 0.359$. Significance cutoffs: n.s. > 0.01, *p ≤ 0.01, **p < 0.001.

vast differences in the sizes of lineages within a repertoire (Figures 2.5A and 2.3E). Similarly, IGHJ gene usage was also comparable across different cohorts for bulk and plasma B cell repertoires (Figures 2.4D, 2.5C, 2.3D, and 2.3G). We do not see a significant distinction in statistics of gene usage between the bulk and plasma B cell repertoires (see Figures 2.4 and 2.5 for bulk and 2.3 for plasma B cells). Our results suggest that the SARS-CoV-2 V gene-specific responses are highly individualized at the repertoire level.

HCDR3 is part of the variable chain of BCRs and is often a crucial region in determining specificity Importantly, HCDR3 is highly variable in its sequence content and length because of insertion and deletion of sequence fragments at the VD and DJ junctions of the germline receptor. Therefore, differential characteristics of the HCDR3 sequence in BCR repertoires of different cohorts can signal preferences for sequence features specific to a class of antigens. We found that HCDR3s of lineages in individuals with COVID-19 with moderate and severe symptoms are significantly longer than in healthy individuals from this study and the GRP ([31]; Figures 2.4B and 2.4C; one-way ANOVA statistics for differences in mean HCDR3 length: healthy-moderate: $F_{1,13} = 15.7$, $p = 1.6 \times 10^{-3}$; healthy-severe: $F_{1,6} = 37.5$, $p = 8.7 \times 10^{-4}$; GRP-moderate: $F_{1,20} = 34.0$, $p = 1.1 \times 10^{-5}$; GRP-severe: $F_{1,13} = 41.5$, $p = 2.2 \times 10^{-5}$). The difference between HCDR3 length in healthy individuals and individuals with mild symptoms were less significant. These differences are also observed at the level of unique productive BCRs (Figure 2.5B). These findings are consistent with previous reports of longer HCDR3 lengths in individuals with COVID-19 [87, 214, 266]. Despite differences in experimental protocols, the HCDR3 lengths of the healthy cohort from this study and from the GRP [31] are comparable with each other (Figures 2.4B, 2.4C, and 2.5B). In addition, we found no significant difference between the HCDR3 length of the unproductive BCR repertoires of healthy individuals and individuals with COVID-19 (Figure 2.5E), which should reflect biases in generation of receptors prior to functional selection. These findings indicate that BCRs with longer HCDR3s tend to be elicited preferentially in repertoires of individuals responding to SARS-CoV-2. This preference seems to have functional significance because longer HCDR3s are also observed among monoclonal antibodies (mAbs) specific to

Unique productive BCRs



Lineage progenitors, unproductive BCRs

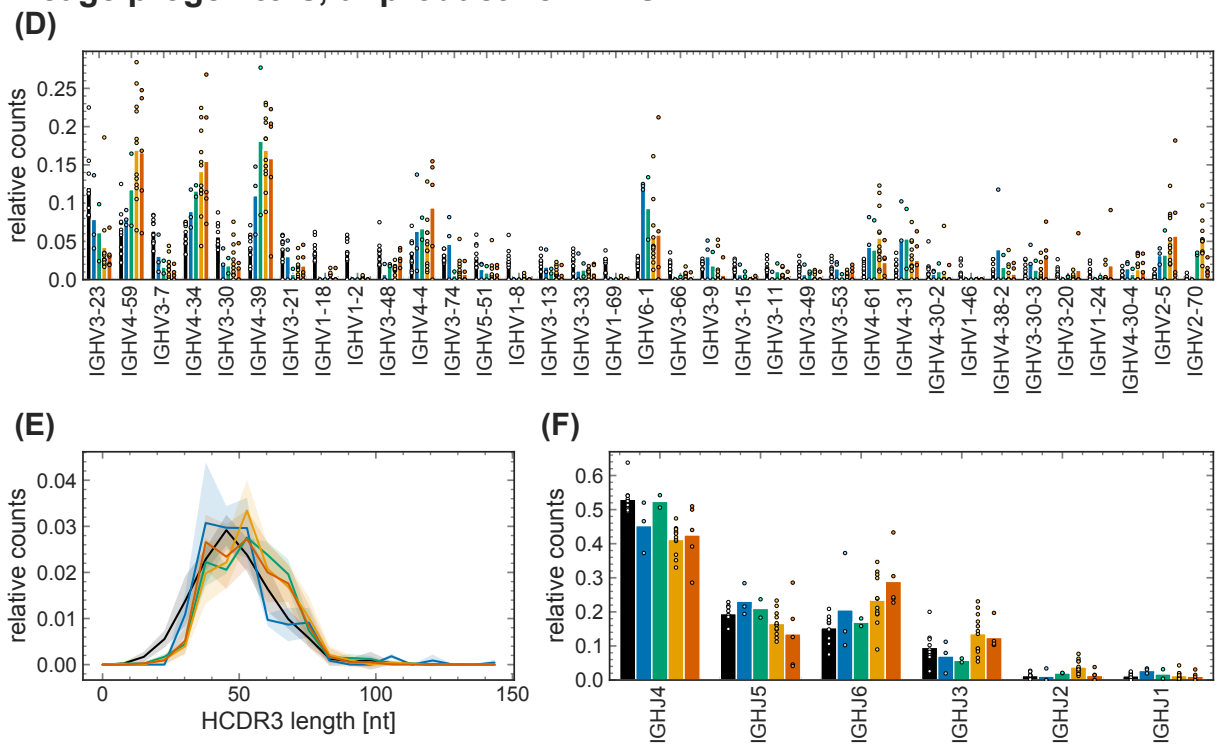


Figure 2.5: **Bulk repertoire sequence statistics. (A - C)** Similar statistics are shown as in Fig 2.4 (A, C - D) but for unique receptors excluding singletons (Appendix A). Unique BCRs in healthy individuals (our control and the GRP [31]) show significantly shorter HCDR3s compared to moderate and severe cohorts. ANOVA statistics for mean HCDR3 length between cohorts are as follows. Healthy-mild: $F_{1,3} = 8.7$, $p = 0.06$; healthy-moderate: $F_{1,13} = 17.2$, $p = 0.001$, healthy-severe: $F_{1,6} = 10.0$, $p = 0.020$; GRP-mild: $F_{1,10} = 11.3$, $p = 0.0073$; healthy-GRP: $F_{1,11} = 0.074$, $p = 0.791$; GRP-moderate: $F_{1,20} = 34.0$, $p = 0.000011$; GRP-severe: $F_{1,13} = 41.5$, $p = 0.000022$. **(D - F)** Similar statistics are shown as in Fig. 2.4 (A, C - D) but for unproductive lineage progenitors. The differences in the statistics of HCDR3 length between the unproductive repertoires of healthy individuals and the COVID-19 cohorts are insignificant (ANOVA $p > 0.01$). Colors are consistent across panels. Detailed statistics on the biological and technical replicates can be found in Data S1.

the RBD and NTD of SARS-CoV-2 (Figure 2.4B), which were identified in previous studies [32, 107, 128, 157, 231, 250, 332, 351].

2.4 Differential selection on B cell repertoires in response to SARS-CoV-2

Longer HCDR3 sequences can introduce more sequence diversity at the repertoire level. Quantifying sequence diversity of a B cell repertoire can be sensitive to the sampling depth in each individual. Despite progress in high-throughput repertoire sequencing techniques, sequenced BCRs still present a highly under-sampled view of the entire repertoire. To characterize the diversity of repertoires and the statistics of sequence features, we inferred models of repertoire generation and selection for entry of receptors into the periphery (Appendix A; [74, 184, 269]). We first used data from unproductive lineage progenitors of BCRs in the bulk repertoire to infer the highly non-uniform baseline model that characterizes the probability $P_{\text{gen}}(\sigma)$ to generate a given receptor sequence, dependent on its sequence features, including the V, D, and J gene choices and also the inserted and deleted sequences at the VD and DJ junctions ([74, 184, 269]; Figure 2.2; Appendix A). The resulting model reflects the biased preferences in generating BCRs in the bone marrow by V(D)J recombination.

The functional but pathogen-naive BCRs that enter the periphery experience selection through processes known as central tolerance [136]. In addition, the inferred progenitors of clonal lineages in the IgG repertoire have undergone antigen-dependent selection that led to expansion of their clonal lineages in response to an infection. These two levels of selection make sequence features of functional lineage progenitors distinct from the pool of unproductive BCRs. In addition, differential selection on receptor features can be used to quantify a distance between repertoires of different cohorts that reflect their functional differences in responses to immune challenges [133].

To identify these distinguishing sequence features, we inferred a selection model for lineage progenitors (Appendix A). We characterized the probability to observe a clonal lineage ancestor in the periphery as $P_{\text{post}}(\sigma) \sim P_{\text{gen}}(\sigma) e^{\sum_{f:\,\text{features}} q_f(\sigma)}$, which deviates from the inferred generation probability of the receptor $P_{\text{gen}}(\sigma)$ by selection factors $q_f(\sigma)$ [132, 130, 133, 269].

These selection factors $q_f(\sigma)$ depend on sequence features, including IGHV and IGHJ genes, HCDR3 length, and amino acid preferences at different positions in HCDR3 (Appendix A; [74, 132, 130, 133, 184, 269]. The inferred selection models are robust to differences in the sample size of the repertoires when enough data are available to train the models (Appendix A; Figures 2.6C-2.6F). As a result, selection models offer a robust approach to compare functional differences even between repertoires with widely different sample sizes, as is the case for our cohorts (Appendix A; Figures 2.6C-2.6F).

The distribution of the log probability $\log_{10} P_{\text{post}}(\sigma)$ for the inferred progenitors of clonal lineages observed in individuals from different cohorts is shown in Figure 2.7A. We find an overabundance of BCR lineages with progenitors that have a low probability of entering the periphery (i.e., a lower $P_{\text{post}}(\sigma)$) in individuals with COVID-19 compared with healthy individuals (Figure 2.7A). A similar pattern is observed at the level of generation probability $P_{\text{gen}}(\sigma)$ for functional receptors in healthy individuals versus individuals infected with COVID-19 (Figure 2.6A). Notably, the inferred selection models from the GRP healthy repertoires are comparable with the healthy cohort in this study (Figure 2.6B). Thus, the overabundance of rare receptors in individuals with COVID-19 is likely to be linked to functional responses associated with stimulation of the repertoires against SARS-CoV-2.

We estimated the diversity of the repertoires in each cohort by evaluating the entropy of receptor sequences generated by the respective repertoire models (Appendix A). In particular, diverse repertoires that contain B cell lineages with rare receptors (i.e., those with a lower $P_{\text{post}}(\sigma)$), should have larger entropies. We found that immune repertoires are more diverse in individuals with COVID-19 compared with healthy individuals (Figure 2.7A; Appendix A). The entropy of BCR bulk repertoires grows with disease severity, from 39.18 bits in the healthy cohort to 40.81 ± 0.03 bits in the mild cohort, 41.03 ± 0.25 bits in the moderate cohort, and 41.32 ± 0.11 bits in the severe cohort (Appendix A). The error bars show the standard error of the mean obtained by averaging over entropy estimates from different models inferred in each of the COVID-19 cohorts, from repertoires subsampled to the same size as the healthy control (Appendix A). As indicated in Figure 2.6, the models inferred

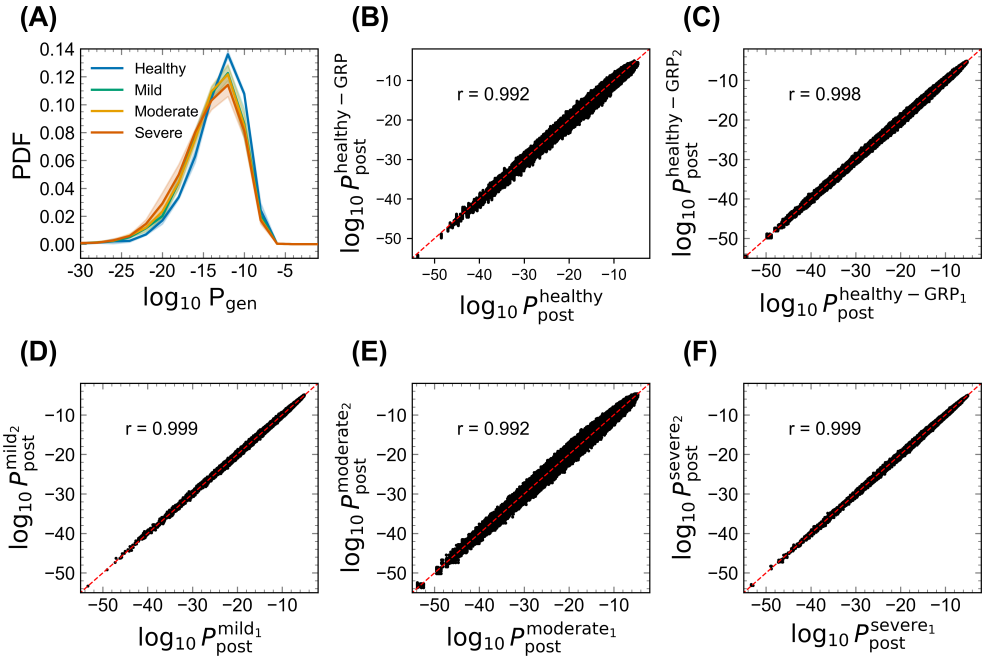


Figure 2.6: **Robustness of SONIA selection models. (A)** The distribution of the log-generation probability of a sequence σ $\log_{10} P_{\text{gen}}(\sigma)$, evaluated using the inferred generation models by the IGoR software [184], is shown as a normalized probability density function (PDF) for inferred naive progenitors of productive clonal lineages in cohorts of healthy individuals and the mild, moderate, and severe cohorts of individual with COVID-19 (colors). Full lines show distributions averaged over individuals in each cohort, and shadings indicate regions containing one standard deviation of variation among individuals (biological replicates) within a cohort. **(B)** The scatterplot shows $\log P_{\text{post}}$ obtained by evaluating 500,000 generated sequences using the inferred selection S(ONIA) models [269] trained on the healthy cohort (x-axis) and 30 SONIA models trained on independent samples of the GRP dataset [31] downsampled to the size of the healthy cohort in this study (7,161 receptors) (y-axis). The scatters show all unique pairwise comparisons between SONIA models trained on independent subsets with each cohort for **(C)** GRP (30 models), and individuals with COVID-19 with **(D)** mild (two models), **E** moderate (13 models), and **F** severe (three models) symptoms (Appendix A). The pearson correlation for all pairwise model comparisons is shown in each panel. Detailed statistics on the biological replicates can be found in Data S1.

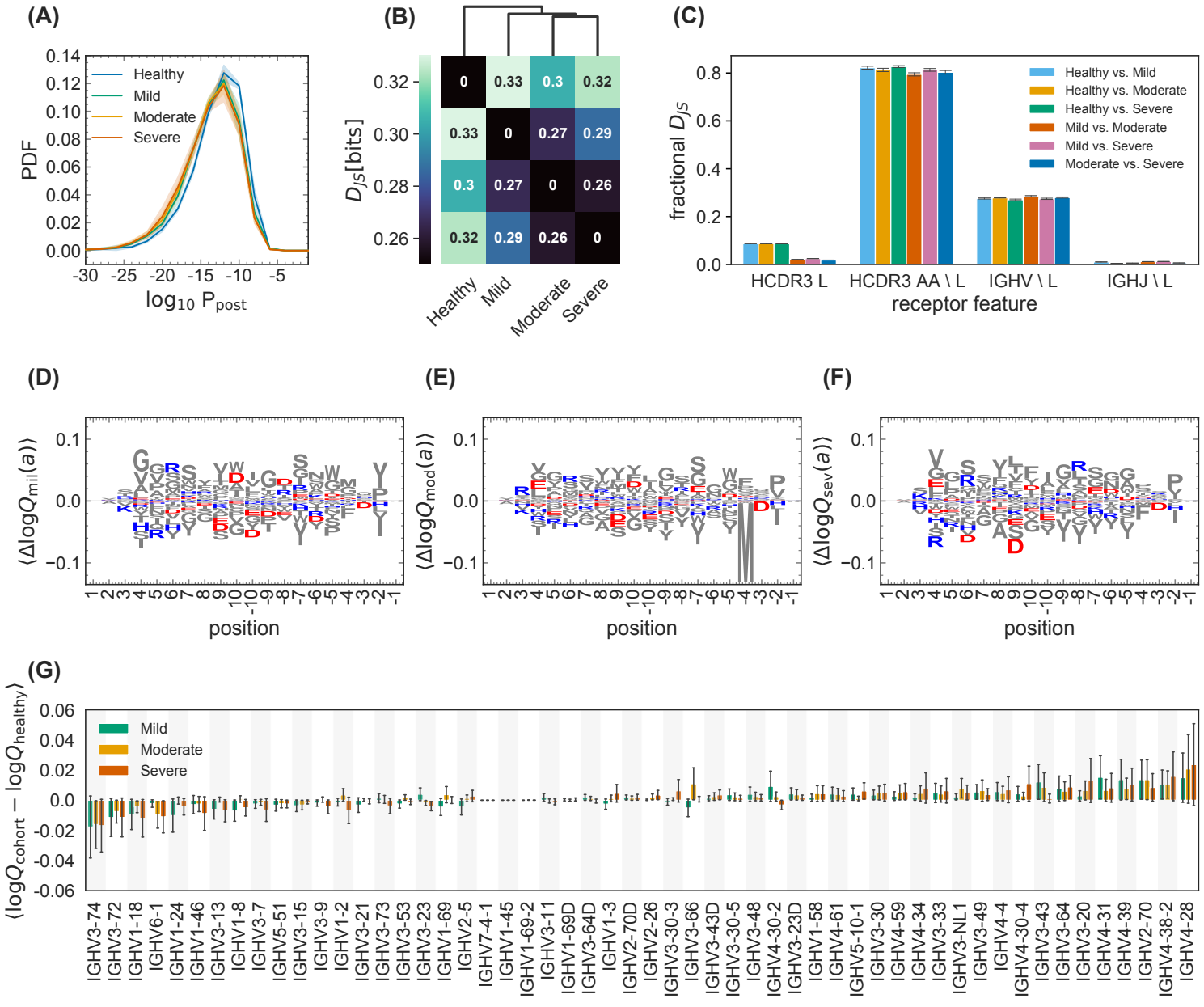

(A)
PDF
Healthy
Mild
Moderate
Severe
log10 Ppost
(B)
DJS[bits]
Healthy
Mild
Moderate
Severe
(C)
fractional DJS
Healthy vs. Mild
Healthy vs. Moderate
Healthy vs. Severe
Mild vs. Moderate
Mild vs. Severe
Moderate vs. Severe
HCDR3 L
HCDR3 AA \ L
IGHV \ L
IGHJ \ L
receptor feature
(D)
⟨ΔlogQmil(a)⟩
position
(E)
⟨ΔlogQmod(a)⟩
position
(F)
⟨ΔlogQsev(a)⟩
position
(G)
⟨logQcohort − logQhealthy⟩
Mild
Moderate
Severe

Figure 2.7: **Differential statistics of immune repertoires across cohorts. (A)** The distribution of the log probability to observe a sequence σ in the periphery $\log_{10} P_{\text{post}}(\sigma)$ is shown as a normalized probability density function (PDF) for inferred naive progenitors of clonal lineages in cohorts of healthy individuals and the mild, moderate, and severe cohorts of individuals with COVID-19. Full lines show distributions averaged over individuals (biological replicates; Data S1) in each cohort, and shading indicates regions containing one standard deviation of variation among individuals within a cohort. **(B)** Clustering of cohorts based on their pairwise Jensen-Shannon divergences (D_{JS}) as a measure of differential selection on cohorts (Appendix A). **(C)** The bar graph shows how incorporating different features into a SONIA selection model contributes to the fractional D_{JS} between models trained on different cohorts. The error bars show the standard deviation of these estimates, using five independent sets of 100,000 generated BCRs for each selection model (Appendix A). **(D - F)** Logo plots show the expected differences in the log-selection factors for amino acid usage. $\langle \Delta \log Q_{\text{cohort}}(a) \rangle = \langle \log Q_{\text{cohort}}(a) - \log Q_{\text{healthy}}(a) \rangle$, for the (D) mild, (E) moderate, and (F) severe COVID-19 cohorts. The expectation values $\langle \cdot \rangle$ are evaluated on the mixture distribution $\frac{1}{2}\left(P_{\text{post}}^{\text{cohort}} + P_{\text{post}}^{\text{healthy}}\right)$. Positively charged amino acids (lysine, K; arginine, R; and histidine, H) are shown in blue, and negatively charged amino acids (aspartate, D, and glutamate, E) are shown in red. All other amino acids are shown in gray. Positions along the HCDR3 are shown up to 10 residues starting from the 3' (position values) and 5' ends (negative values). **(G)** The bar graph shows the average mean difference between the log-selection factors for IGHV gene usage for the mild (green), moderate (yellow), and severe (red) COVID-19 cohorts, with the mean differences computed using the mixture distribution $\frac{1}{2}\left(P_{\text{post}}^{\text{cohort}} + P_{\text{post}}^{\text{healthy}}\right)$, and the average is taken over 30 independently trained SONIA mdoels for each cohort. Error bars show the standard deviation of these estimates across the inferred SONIA models (Appendix A).

from subsampled repertoires are highly consistent within each cohort.

Selection factors $q_f(\sigma)$ determine the deviation in preferences for different sequence features of BCRs in each cohort. A comparison of selection factors among cohorts can characterize their distinctive sequence features. To quantify the selection differences across cohorts, we evaluated the Jensen-Shannon divergence (D_{JS}) between repertoires of different cohorts, which measures the distance between the features of their receptor distributions ([133]; Appendix A). Clustering of the cohorts based on their pairwise D_{JS} indicates that repertoires diverge with growing disease severity and that COVID-19 cohorts are more similar to each other than to the healthy cohort (Figure 2.7B; Appendix A).

The inferred selection models enabled us to quantify how different receptor features affect the pairwise D_{JS} of BCR repertoires (Appendix A). We found that HCDR3 length contributes the most to differences in receptor distributions between healthy and COVID-19 cohorts (Figure 2.7C), consistent with the significant differences in HCDR3 length distributions shown in Figure 2.4C. We also found that the amino acid composition of HCDR3 is the second most distinguishing factor between repertoires (Figure 2.7C), indicating that negatively charged amino acids are slightly suppressed at the center of HCDR3s in COVID-19 cohorts compared with healthy repertoires (Figures 2.7D - F). The selection differences of IGHV and IGHJ gene usage between healthy individuals and those with COVID-19 are insignificant (Figures 2.7C and 2.7G), consistent with our previous analysis of lineage characteristics in Figures 2.4A and 2.4D. HCDR3 length and composition are the molecular features that are most distinguishable at the repertoire level across different cohorts. Nonetheless, further work is necessary to understand the molecular underpinnings that may make these receptor features apt in response to SARS-CoV-2.

2.5 Expansion of BCR clonal lineages over time indicates responses to SARS-CoV-2

We examined the dynamics of BCR repertoires in individuals with COVID-19. The binding level (measured by optical density 450 [OD_{450}] in ELISAs) of IgM and IgG antibodies against

the RBD or NTD of SARS-CoV-2 increased in most individuals with COVID-19 in our study over the course of their infection (Figures 2.8A and 2.9). We expect that the increase in OD_{450} binding level is associated with activation of specific B cells, increasing mRNA production of the corresponding BCRs. Detecting expansion of specific clonal lineages is challenging because of subsampling of the repertoires. Only a limited overlap of BCR lineages was found when we compared data between different time points or between technical replicates of a repertoire sampled at the same time point (Figures 2.10A and 2.10B). To identify expanding clonal lineages, we examined lineages only in individuals whose plasma showed an increase in binding level (OD_{450}) to the RBD of SARS-CoV-2 and compared the sequence abundance of those lineages in the bulk repertoire that appeared at two or more time points (Figures 2.8A and 2.9; Appendix A). Using a hypothesis test with a false discovery rate of 7.5%, as determined by analyzing technical replicate data (Appendix A; Figure 2.10), we detected significant expansion of clonal lineages of receptors harvested from the bulk repertoire within all investigated individuals. The results reflect a dynamic repertoire in all individuals, ranging from 5%-15% of lineages with significantly large changes in sequence abundances over time (Figures 2.8 and 2.10). The expanding lineages have an HCDR3 length comparable with the rest of the repertoire in individuals with COVID-19 (Figure 2.10). Moreover, the expanding lineages show V gene preferences comparable with previously identified antibodies against SARS-CoV-2 (RBD). This includes the abundance of IGHV4-59, IGHV4-39, IGHV3-23, IGHV3-53, IGH3-66, IGHV2-5, and IGHV2-70 [32, 142, 231, 252]. However, these preferences in V gene usage among expanding lineages are also comparable with the overall biases in V gene usage within individuals, and expanded lineages roughly make up 25% of lineages with a given V gene (Figure 2.8C). Therefore, our results suggest that the overall response to SARS-CoV-2 is not driven by only specific classes of IGHV genes.

We expect clonal expansions to reflect responses to SARS-CoV-2 during infection. Indeed, we observe that expanding lineages (based on the bulk data) show an overrepresentation of receptors harvested from plasma B cells, which are likely to be associated with antibody-secreting B cells (Figure 2.8D; Appendix A); specific p values for each individual are reported

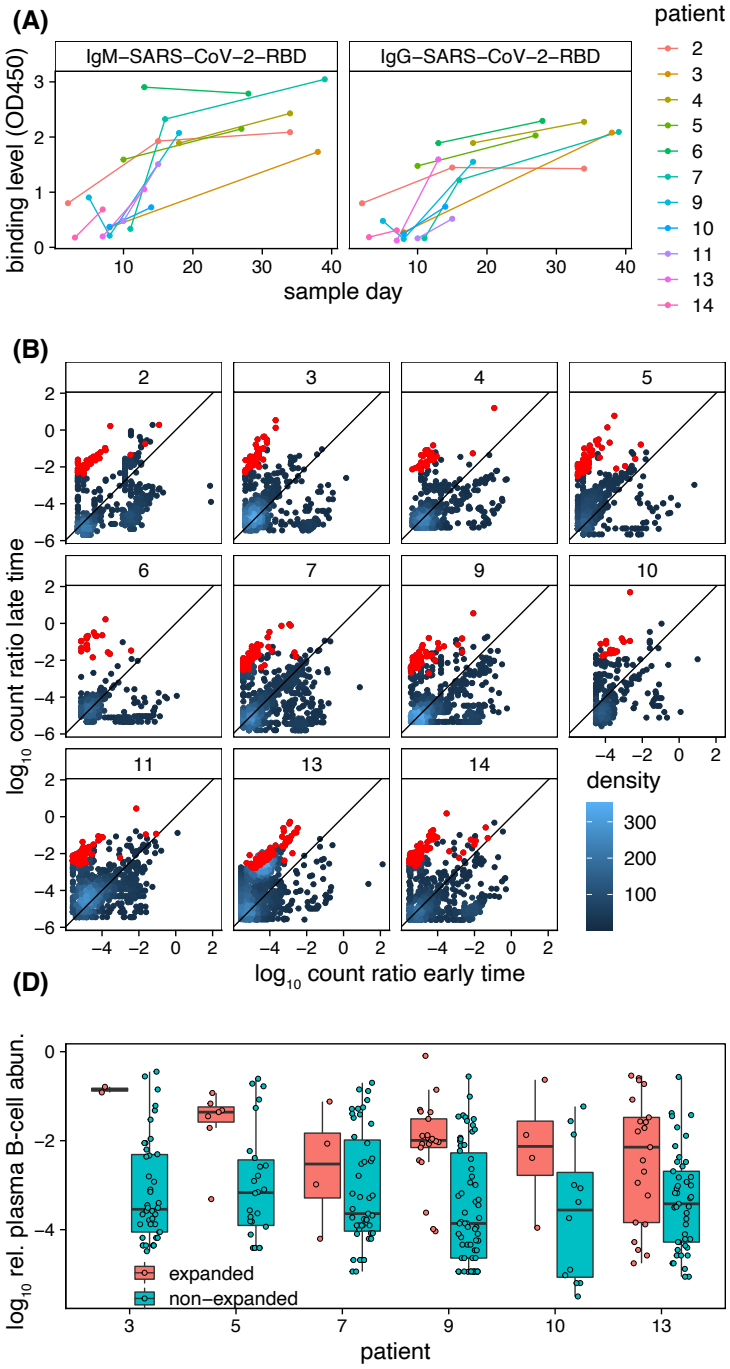

(A)
IgM−SARS−CoV−2−RBD
IgG−SARS−CoV−2−RBD
binding level (OD450)
sample day
patient
2
3
4
5
6
7
9
10
11
13
14
(B)
log10 count ratio late time
log10 count ratio early time
density
300
200
100
(C)
IGHV4−59
IGHV4−39
IGHV4−34
IGHV6−1
IGHV4−61
IGHV4−4
IGHV4−31
IGHV3−23
IGHV2−5
IGHV2−70
IGHV3−30
IGHV3−21
IGHV3−48
IGHV3−7
IGHV4−30−4
IGHV4−38−2
IGHV3−33
IGHV3−74
IGHV3−30−3
IGHV5−51
IGHV3−11
IGHV4−30−2
IGHV3−53
IGHV3−13
IGHV3−9
IGHV3−66
IGHV3−43
IGHV5−10−1
IGHV3−20
IGHV1−18
IGHV1−8
IGHV3−15
IGHV1−24
IGHV4−28
IGHV3−64
IGHV1−2
IGHV3−49
IGHV3−73
IGHV3−72
IGHV1−69
IGHV1−46
IGHV1−58
non−expanded
expanded
fraction expanded
patient
lineages
(D)
log10 rel. plasma B−cell abun.
expanded
non−expanded
patient

Figure 2.8: **Dynamics of BCR repertoires during infection. (A)** The binding level (measured by OD_{450} in ELISA) of the IgM (left) and IgG (right) repertoires to SARS-CoV-2 (RBD) epitopes increases over time in most individuals. **(B)** The log ratio of BCR (mRNA) abundance at late time versus early time is shown for all clonal lineages that are present in at least two time points (Appendix A). Each panel shows dynamics of lineages for a given individual, as indicated in the label. The analysis is shown for individuals for whom the binding level (OD_{450} of the IgG repertoire increases over time (shown in A). The count density indicates the number of lineages at each point. Lineages that expanded significantly over time are indicated in red (Appendix A). **(C)** IGHV gene use of lineages is shown for non-expanded (left) and expanded (center) lineages in all individuals (colors). The right panel shows, for each individual (colors), the fraction of expanded lineages with a given IGHV gene as the number of expanded lineages divided by the total number of lineages with that given IGHV gene. The size of the circles indicates the total number of lineages in each category. **(D)** Boxplot of the $\log_{10}$ relative read abundance in the plasma B cell repertoire (Appendix A) for expanding (red) and nonexpanding (cyan) lineages that contain reads from plasma B cells in different individuals. Receptors from the plasma B cell repertoire are significantly more abundant in expanding lineages in four individuals based on ANOVA test statistics. Individual 3: $F_{1,42} = 5.4$, $p = 0.02$; individual 5: $F_{1,31} = 0.5$, $p = 0.5$; individual 7: $F_{1,49} = 0.01$, $p = 0.91$; individual 9: $F_{1,42} = 4.1$, $p = 0.04$; individual 10: $F_{1,42} = 2.9$, $p = 0.1$; individual 13: $F_{1,64} = 7.7$, $p = 0.007$.

in the caption of Figure 2.8D.

2.6 Sharing of BCR among individuals

Despite the vast diversity of BCRs, we observe a substantial number of identical progenitors of BCR clonal lineages among individuals with COVID-19 (Figure 2.11) and among healthy individuals from our dataset and from the GRP (Figure 2.12). Previous work has also identified sharing of BCRs among individuals with COVID-19, which was interpreted by the authors as evidence of large-scale convergence of immune responses [87, 214, 266]. Although BCR sharing can be due to convergent responses to common antigens, it can also arise from convergent recombination leading to the same receptor sequence [75, 233] or from experimental biases. Therefore, it is imperative to formulate a statistical model to quantify the significance of BCR sharing. Convergent recombination defines a null expectation for the amount of sharing within a cohort based on only the underlying biases for receptor generation within a repertoire ([75, 233]; Appendix A). Intuitively, sharing is more likely among commonly generated receptors (i.e., with a high) and within cohorts with larger sampling (Appendix A). Consequently, rare receptors (i.e., with a low) that are shared among individuals in a common disease group can signal commonality in function and a response to a common antigen, as observed previously for T cell receptors (TCRs) in response to a yellow fever vaccine [235], cytomegalovirus (CMV), and diabetes [233].

We used the receptors' probabilities $P_{\text{post}}(\sigma)$ to assess the significance of sharing by identifying a probabilistic threshold to limit the shared outliers among individuals with COVID-19 (dashed line in Figure 2.11) and healthy individuals (dashed lines in Figure 2.12). Of a total of 40,128 (unique) progenitors of clonal lineages reconstructed from the pooled bulk+plasma B cell repertoires (Figure 2.11A; Data S1), we found 10,146 progenitors to be shared among at least two individuals, and 761 of these lineages contained receptors found in plasma B cells of at least one individual. 167 of the 10,146 lineage progenitors were classified as rare, having a probability of occurrence below the indicated threshold (dashed line in Figure 2.11B). 30 of these contain receptors harvested from plasma B cells,

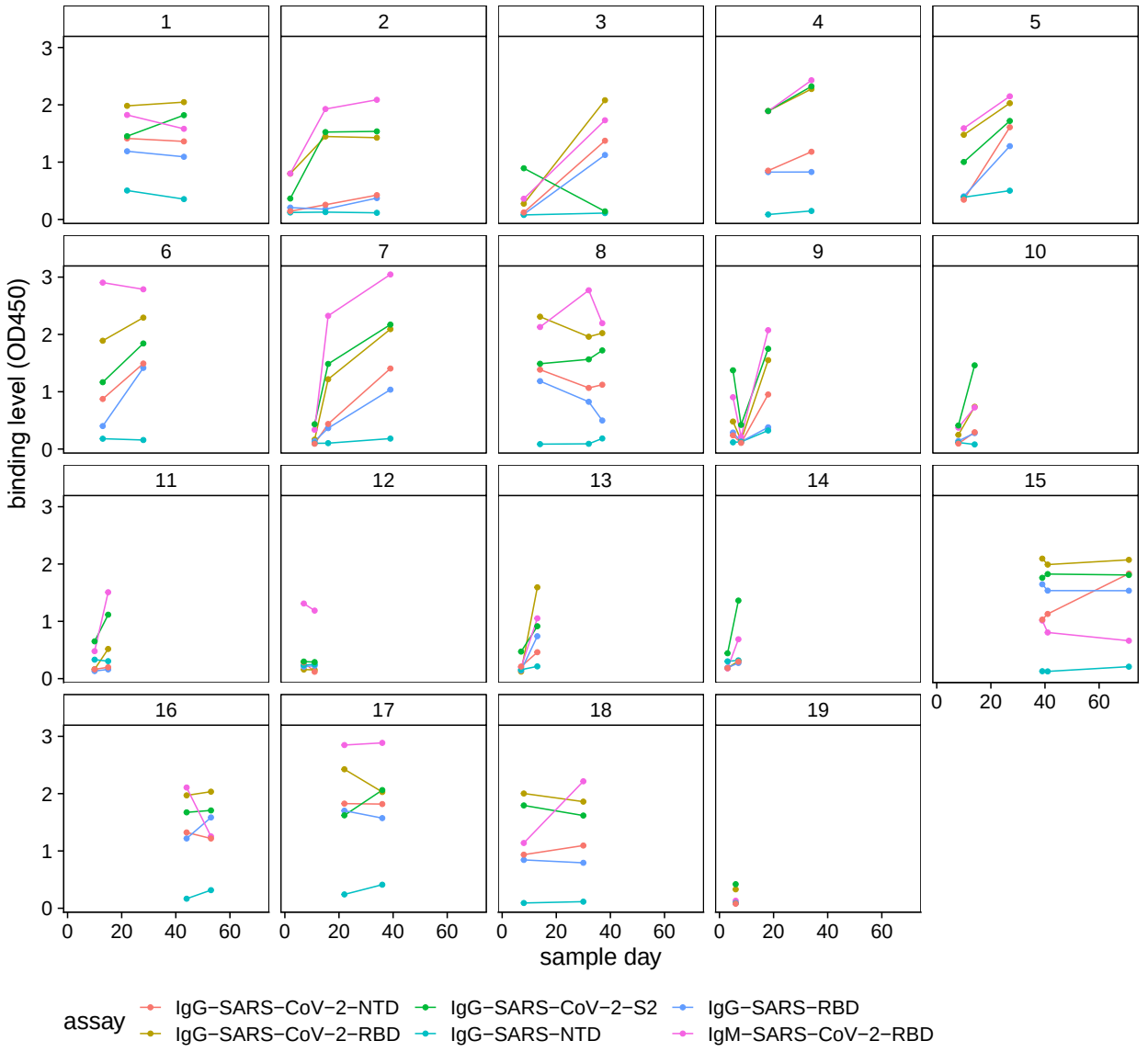


Figure 2.9: **ELISA binding assays for IgG and IgM repertoires against SARS-CoV-2 and SARS-CoV.** Plasma binding levels (measured by OD_{450} against RBD, NTD, and S2 subdomain of SARS-CoV-2 and against RBD and NTD epitopes of SARS-CoV are shown. As seen in the binding assays, many individuals developed a cross-reactive response to SARS-CoV-2 and SARS-CoV. Some individuals showed no increase in IgG binding to SARS-CoV-2 RBD due to already high levels at sampling time or natural variation. For the expansion analysis (Fig. 2.8), we only analyzed individuals whose IgG repertoires showed an increase in binding to SARS-CoV-2 (RBD): 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, and 14.

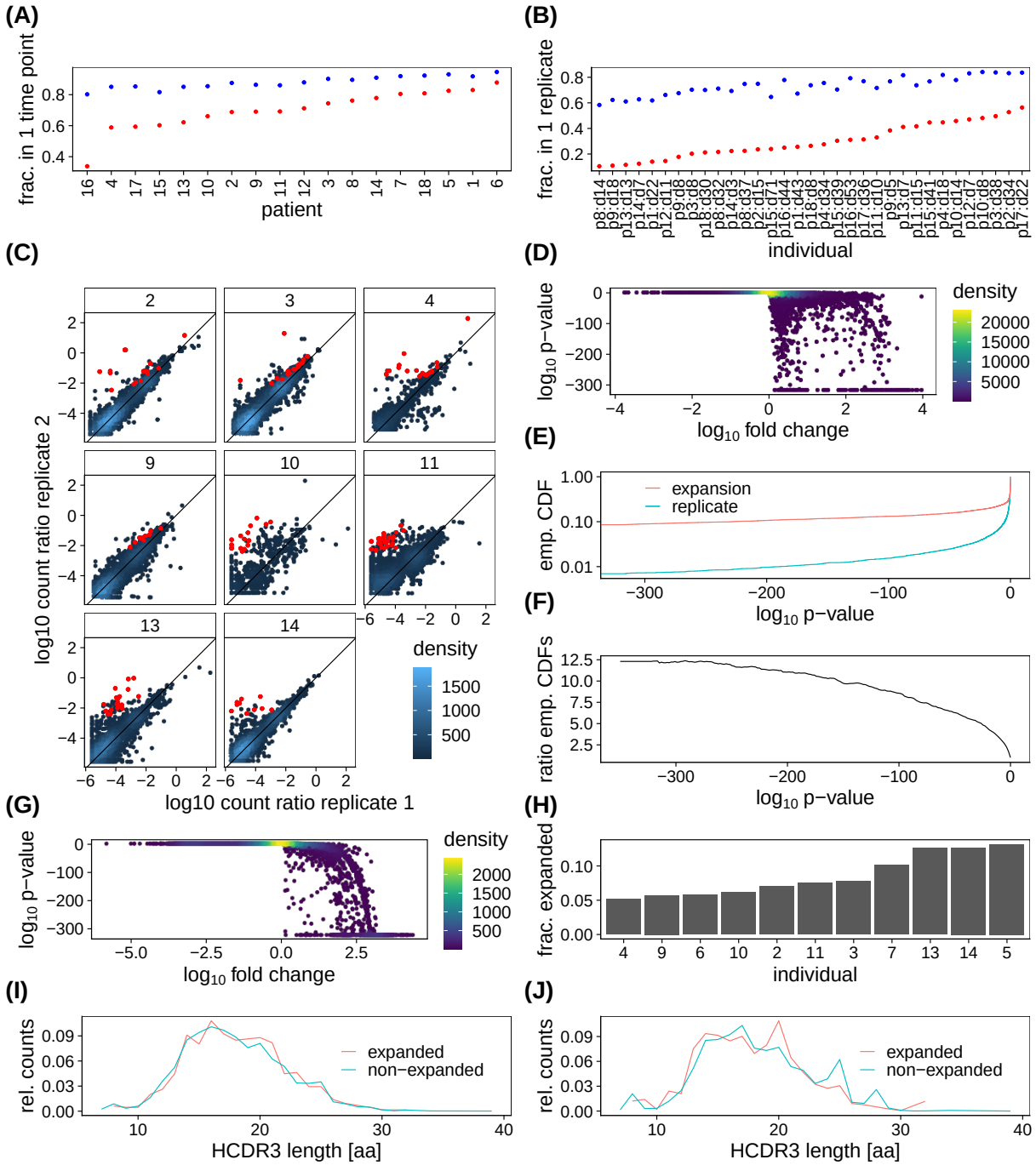

(A)
frac. in 1 time point
patient
(B)
frac. in 1 replicate
individual
(C)
log10 count ratio replicate 2
log10 count ratio replicate 1
density
(D)
log10 p−value
log10 fold change
density
(E)
emp. CDF
expansion
replicate
log10 p−value
(F)
ratio emp. CDFs
log10 p−value
(G)
log10 p−value
log10 fold change
density
(H)
frac. expanded
individual
(I)
rel. counts
HCDR3 length [aa]
expanded
non−expanded
(J)
rel. counts
HCDR3 length [aa]
expanded
non−expanded

Figure 2.10: **Statistics of clonal expansion.** BCR repertoires are highly undersampled, and relatively few BCR lineages appear in multiple time points and technical replicates. **(A)** Fraction of lineages present in only one time point before (blue) and after (red) filtering out small lineages (i.e., those with less than three unique sequences per time point) are shown. **(B)** Fraction of lineages present in only one technical replicate before (blue) and after (red) filtering out small lineages (i.e., those with less than three unique sequences per time point) are shown. **(C)** The log-ratio of abundance of receptors for all clonal lineages present in two technical replicates is shown. Each panel shows the test result for a given patient, as indicated in the label. The count density indicates the number of lineages at each point. Lineages that show a significant expansion (false positives) are indicated in red. **(D)** $\log_{10}$ p values of the expansion test versus $\log_{10}$ fold change (or odds ratio) for the technical replicate data are shown. Color indicates density of points, and p values of zero are displayed at the minimum nonzero value. See Appendix A for normalization, data processing, and hypothesis testing. **(E)** Empirical cumulative density functions (ECDF) of expansion data from multiple time points (red) and the technical replicate data (blue) show that many more tests performed on multiple time point data result in low p values compared to the technical replicate data. **(F)** Ratio of ECDFs indicates that at a significant threshold of 10^{-300} there are roughly 12.3 times more true positives than false positives. **(G)** $\log_{10}$ p values of the expansion test versus $\log_{10}$ fold change for the data corresponding to Fig. 2.8B is shown. Color indcates density of points, and p values of zero are displayed at the minimum nonzero value. **(H)** Fraction of lineages expanded for different individuals is shown. HCDR3 length distributions of expanded and nonexpanded lineages **(I)** with each lineage having equal weight and **(J)** with each lineage weighted by the number of unique sequences per time point (excluding singletons) are shown.

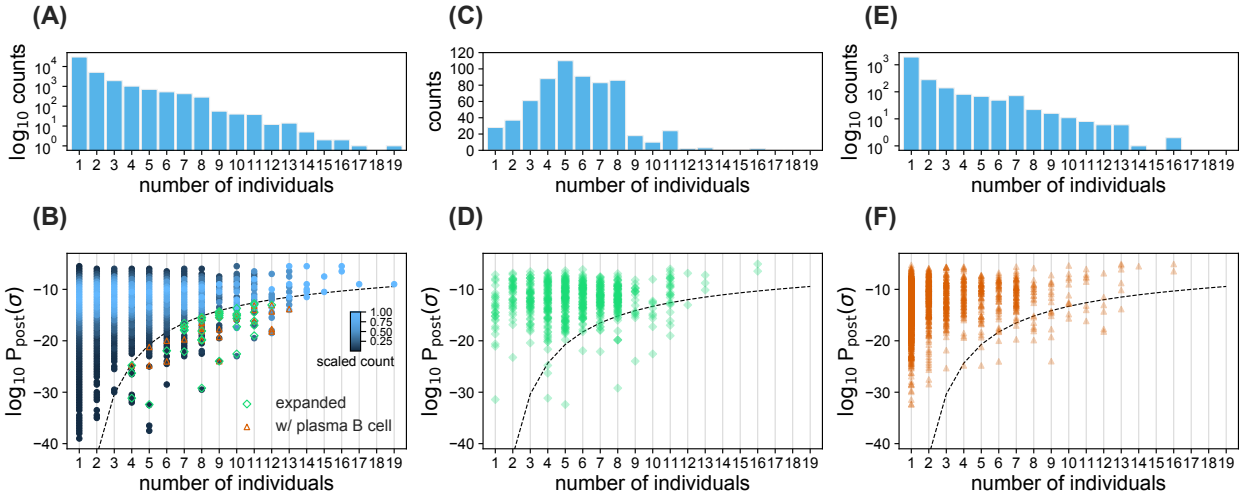


Figure 2.11: **Sharing of BCRs among individuals. (A)** The histogram shows the number of clonal lineages that share a common progenitor in a given number of individuals, indicated on the horizontal axis. **(B)** The density plot shows the distribution of $\log_{10} P_{\text{post}}$ for progenitors of clonal lineages shared in a given number of individuals, indicated on the horizontal axis. The histogram bin size is 0.5. The scaling of sequence counts sets the maximum of the density in each column to one. Sharing of rare lineages with $\log_{10} P_{\text{post}}$ below the dashed line is statistically significant (Appendix A). Green diamonds indicate clonal lineages below the dashed line with significant expansion in at least one of the individuals. Orange triangles indicate clonal lineages below the dashed line that contain reads from the plasma B cell repertoire in at least one of the individuals. **(C, E)** Histograms showing the numbers of clonal lineages that share a common progenitor in a given number of individuals that have expanded significantly during infection in at least one of the individuals (C) or contained reads from the plasma B cell repertoire in at least one of the individuals (E). **(D, F)** Scatterplots with transparent, overlapping markers show $\log_{10} P_{\text{post}}$ for progenitors of clonal lineages shared in a given number of individuals that have expanded (D) or contain reads from the plasma B cell repertoire (F) in at least one individual. The dashed line is similar to (B).

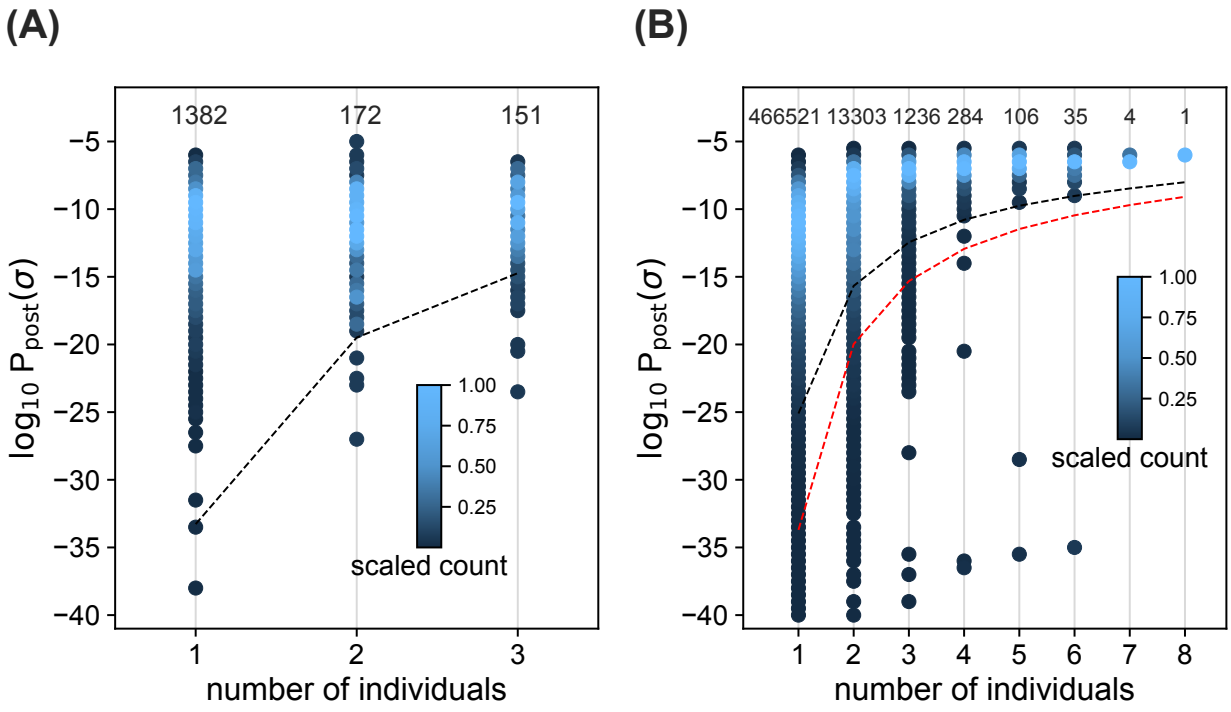


Figure 2.12: **Sharing of BCRs among healthy individuals. (A)** The density plot shows the distribution of $\log_{10} P_{\text{post}}$ for progenitors of clonal lineages shared in a given number of healthy individuals, indicated on the horizontal axis. (Histogram bin size is 0.5.) The clonal lineages are constructed from the bulk data. The counts in each bin are scaled such that the maximum is equal to one for each column. The numbers above each column indicate the total number of sequences in the respective column. Sharing of rare lineages with $\log_{10} P_{\text{post}}$ below the dashed line is statistically significant (see Appendix A). **(B)** Similar statistics are in (A) are shown but for healthy individuals in the Great Repertoire Project [31]. Sharing of rare lineages with $\log_{10} P_{\text{post}}$ below the black, dashed line is statistically significant (see Appendix A). For comparison, the dashed line in (A) is shown as a red dashed line in (B) and extended to eight individuals.

indicating a significant over-abundance of plasma B cells among the rare, shared receptors ($p = 7.2 \times 10^{-6}$) (Figures 2.11E and 2.11F). Moreover, we found that 615 lineages shared a common sequence ancestor in at least two individuals and had expanded in at least one of the individuals (Figures 2.11C and 2.11D). 38 of these shared, expanding lineages stemmed from rare naive progenitors (below the dashed line in Figures 2.11B and 2.11D), eight of which contain receptors found in plasma B cells of at least one individual. The over-abundance of plasma BCRs in the rare shared, expanding lineages is significant ($p = 0.04$). The sharing of these rare, expanding BCRs among individuals with COVID-19, with an overabundance of receptors associated with antibody production in the plasma B cell data, indicates a potentially convergent response to SARS-CoV-2; these receptors are listed in Data S2.

We found that 24% of receptors in the 38 rare shared, expanding lineages contain multiple cysteines in their HCDR3s, in contrast to only 10% of the receptors in the whole repertoire. Sequence patterns with cysteine pairs in the HCDR3 have been associated with stabilization of the HCDR3 loop by forming disulfide bonds with particular patterns and spacings of the cysteines [170, 239]. Disulfide bonds in the HCDR3 can decrease the conformational flexibility of the loop, thus decreasing the entropic cost of binding to improve the affinity of the receptor [9]. The significantly larger fraction of multi-cysteine HCDR3s among the candidate SARS-CoV-2-responsive receptors ($p = 0.013$ based on binomial sampling) indicates an underlying molecular mechanism for developing a potent response to SARS-CoV-2.

2.7 Presence of SARS-CoV-2 and SARS-CoV-1 specific neutralizing antibodies within repertoires

To further investigate the functional response in the repertoire of individuals with COVID-19, we performed single-cell sequencing on pooled samples from all individuals, sorted for reactivity to RBD or NTD epitopes of SARS-CoV-2 (Appendix A). This analysis suggests that about 0.2% of these single cells are RBD reactive as opposed to only 0.02% that are NTD reactive (Figure 2.1). This inferred fraction of reactive antibodies is consistent with previous estimates [156].

We characterized the sequence features of RBD- and NTD-sorted antibodies. The IGHV gene usage of these reactive receptors is shown in Figure 2.13 and is compared with gene usage in mAbs identified in previous studies [32, 107, 128, 157, 157, 231, 250, 332, 351]. Despite the broad range of IGHV gene usage associated with epitope reactivity, sorted single cells show IGHV gene preferences common to the previously identified mAbs against SARS-CoV-2 epitopes. This includes an abundance of IGHV1-69, IGHV4-59, IGHV3-30-3, IGHV3-33, IGHV1-18, IGHV5-51, and IGHV1-46 against the RBD and IGHV3-23, IGHV4-59, IGHV4-39, IGHV3-21, and IGHV3-48 against the NTD (Figure 2.13A). Similarly, we observe consistent biases in V and J gene usage of the and light chains for the sorted single cells and verified mAbs (Figure 2.14). Moreover, the HCDR3 length distributions of the sorted single cells are comparable to those of the verified mAbs (Figure 2.14). The average lengths of the HCDR3 for the verified mAbs and sorted single-cell receptors are comparable to those of bulk repertoires from individuals with COVID-19, which are significantly longer than those of healthy individuals (Figure 2.4B).

To characterize how SARS-CoV-2-reactive receptors make up individuals' repertoires, we mapped the heavy chain receptors from the sorted single cells onto BCR lineages constructed from the bulk+plasma B cell data in individuals with COVID-19 (Appendix A; Data S2). We found that 13 (of 237) RBD-sorted and 13 (of 330) NTD-sorted antibodies from the single cells matched receptor lineages in at least one individual (Figure 2.13B). Interestingly, we found broad sharing of these antibodies with 10 RBD- and 6 NTD-sorted single cells present in at least two individuals (Figure 2.13B).

In repertoires of individuals with COVID-19, we found that several HCDR3s matched with SARS-CoV-2-specific mAbs that have been isolated previously in other studies [32, 107, 128, 157, 157, 231, 250, 332, 351]. Specifically, a total of 20 mAb families specific to SARS-CoV-2 epitopes were found to be close in sequence to HCDR3s in our data (with up to one amino acid difference), among which are 14 RBD-specific, one NTD-specific, and five S1-specific (reactive to the RBD or NTD) mAbs (Figure 2.13B; Data S3). Interestingly, nine of these mAbs are shared among at least two individuals, and the NTD-specific antibody is

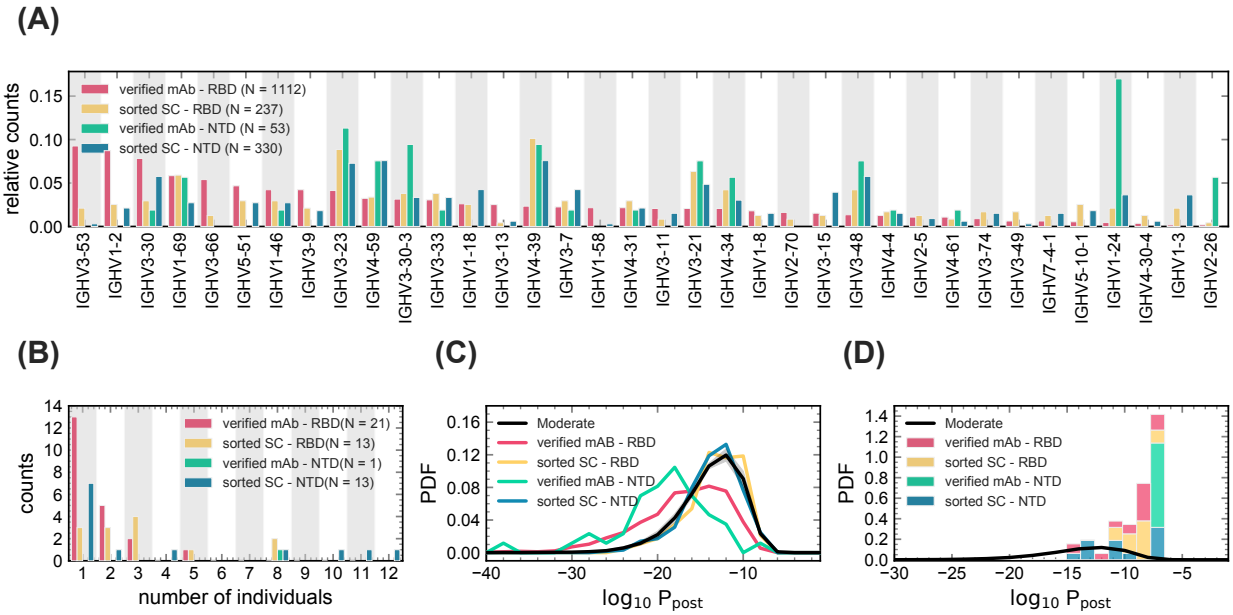


Figure 2.13: **Statistics of BCRs reactive to RBD and NTD epitopes. (A)** Relative counts for IGHV gene usage for known mAbs (Data S3) reactive to RBD (pink) and NTD (green) epitopes of SARS-CoV-2 and for receptors obtained from single-cell sequencing of the pooled sample from all individuals (Appendix A), sorted for RBD (yellow) and NTD (blue) epitopes. **(B)**The histogram shows the number of NTD-sorted receptors from single cell sequencing (Data S2) and RBD- and NTD-specific verified mAbs (Data S3) found in the bulk+plasma B cell repertoires of a given number of individuals (Appendix A), indicated on the horizontal axis. **(C)** The distribution of the log probability to observe a sequence σ in the periphery $\log_{10} P_{\rm post}(]\sigma)$ is shown as a normalized PDF for inferred naive progenitors of known RBD- and NTD-specific mAbs and for RBD- and NTD-sorted receptors from single-cell sequencing. $P_{\rm post}(\sigma)$ values were evaluated based on the repertoire model created from individuals with moderate symptoms. The corresponding $\log_{10} P_{\rm post}$ distribution for bulk repertoires of the moderate cohort (similar to Fig. 2.7A is shown in black as a reference. **(D)** Similar to (C) but restricted to receptors that are found in the bulk+plasma repertoire of at least one individual in the cohort (Data S2 and S3). Colors are consistent between panels, and the number of samples used in each panel is indicated in the legend.

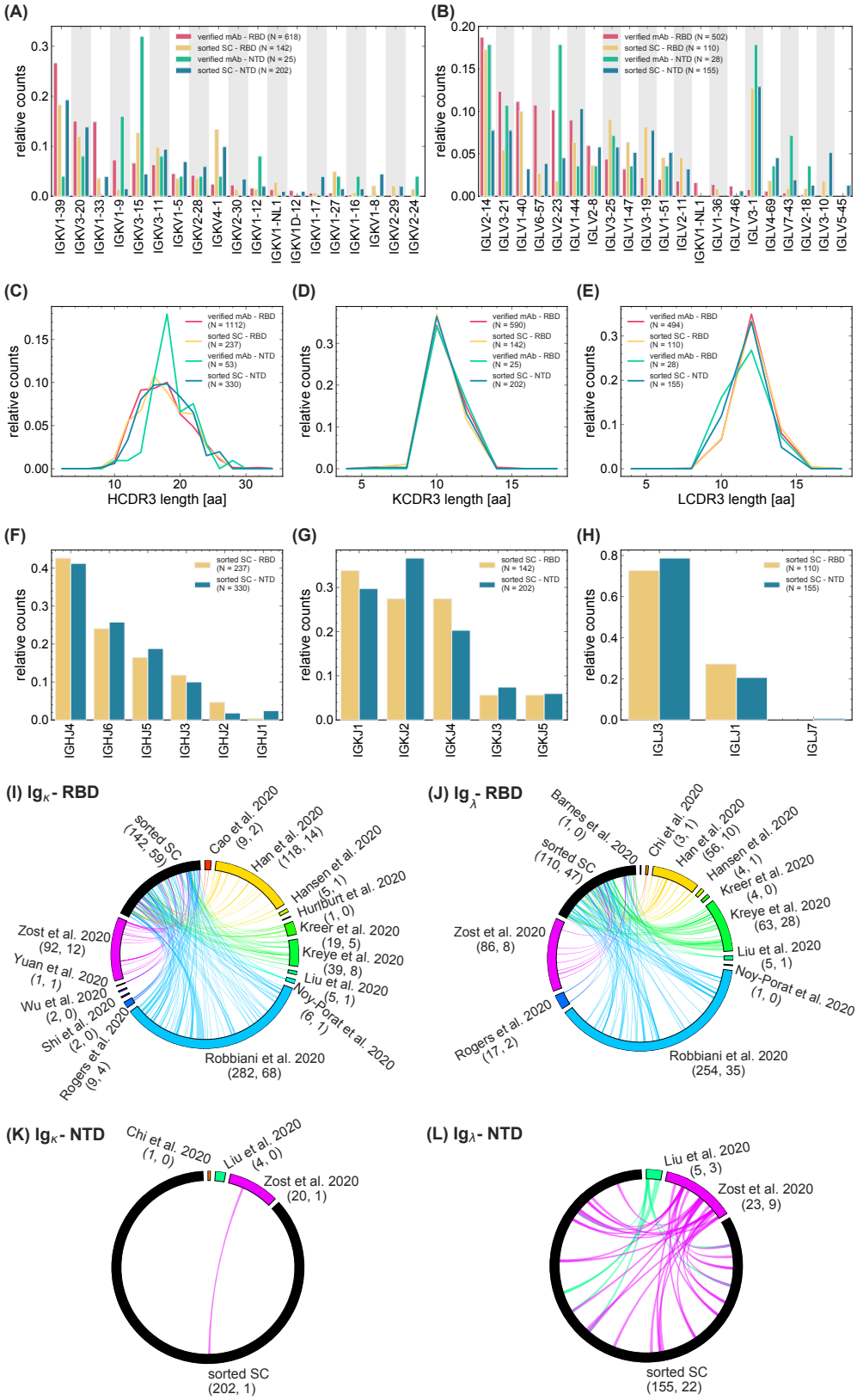

(A)
relative counts
verified mAb - RBD (N = 618)
sorted SC - RBD (N = 142)
verified mAb - NTD (N = 25)
sorted SC - NTD (N = 202)
IGKV1-39
IGKV3-20
IGKV1-33
IGKV1-9
IGKV3-15
IGKV3-11
IGKV1-5
IGKV2-28
IGKV4-1
IGKV2-30
IGKV1-12
IGKV1-NL1
IGKV1D-12
IGKV1-17
IGKV1-27
IGKV1-16
IGKV1-8
IGKV2-29
IGKV2-24
(B)
relative counts
verified mAb - RBD (N = 502)
sorted SC - RBD (N = 110)
verified mAb - NTD (N = 28)
sorted SC - NTD (N = 155)
IGLV2-14
IGLV3-21
IGLV1-40
IGLV6-57
IGLV2-23
IGLV1-44
IGLV2-9
IGLV3-25
IGLV1-47
IGLV3-19
IGLV1-51
IGLV2-11
IGKV1-NL1
IGLV1-36
IGLV7-46
IGLV3-1
IGLV4-69
IGLV7-43
IGLV2-18
IGLV3-10
IGLV5-45
(C)
relative counts
verified mAb - RBD (N = 1112)
sorted SC - RBD (N = 237)
verified mAb - NTD (N = 53)
sorted SC - NTD (N = 330)
HCDR3 length [aa]
(D)
relative counts
verified mAb - RBD (N = 590)
sorted SC - RBD (N = 142)
verified mAb - RBD (N = 25)
sorted SC - NTD (N = 202)
KCDR3 length [aa]
(E)
relative counts
verified mAb - RBD (N = 494)
sorted SC - RBD (N = 110)
verified mAb - RBD (N = 28)
sorted SC - NTD (N = 155)
LCDR3 length [aa]
(F)
relative counts
sorted SC - RBD (N = 237)
sorted SC - NTD (N = 330)
IGHJ4
IGHJ6
IGHJ5
IGHJ3
IGHJ2
IGHJ1
(G)
relative counts
sorted SC - RBD (N = 142)
sorted SC - NTD (N = 202)
IGKJ1
IGKJ2
IGKJ4
IGKJ3
IGKJ5
(H)
relative counts
sorted SC - RBD (N = 110)
sorted SC - NTD (N = 155)
IGLJ3
IGLJ1
IGLJ7
(I) Ig$_\kappa$- RBD
sorted SC (142, 59)
Cao et al. 2020 (9, 2)
Han et al. 2020 (118, 14)
Hansen et al. 2020 (5, 1)
Hunburt et al. 2020 (1, 0)
Kreer et al. 2020 (19, 5)
Kreye et al. 2020 (39, 8)
Liu et al. 2020 (5, 1)
Noy-Porat et al. 2020 (6, 1)
Robbiani et al. 2020 (282, 68)
Rogers et al. 2020 (9, 4)
Shi et al. 2020 (2, 0)
Wu et al. 2020 (2, 0)
Yuan et al. 2020 (1, 1)
Zost et al. 2020 (92, 12)
(J) Ig$_\lambda$- RBD
Barnes et al. 2020 (1, 0)
sorted SC (110, 47)
Chi et al. 2020 (3, 1)
Han et al. 2020 (56, 10)
Hansen et al. 2020 (4, 1)
Kreer et al. 2020 (4, 0)
Kreye et al. 2020 (63, 28)
Liu et al. 2020 (5, 1)
Noy-Porat et al. 2020 (1, 0)
Robbiani et al. 2020 (254, 35)
Rogers et al. 2020 (17, 2)
Zost et al. 2020 (86, 8)
(K) Ig$_\kappa$- NTD
Chi et al. 2020 (1, 0)
Liu et al. 2020 (4, 0)
Zost et al. 2020 (20, 1)
sorted SC (202, 1)
(L) Ig$_\lambda$- NTD
Liu et al. 2020 (5, 3)
Zost et al. 2020 (23, 9)
sorted SC (155, 22)

Figure 2.14: **Sequence features of heavy and light chain receptors in sorted single-cells and monoclonal antibodies.** The bar graphs show the relative counts for **(A)** the κ-chain IGKV-gene usage and **(B)** the λ-chain IGLV gene usage for the verified mAbs reactive to RBD (pink) and NTD (green) epitopes of SARS-CoV-2 (Data S3) and the light chain receptors obtained from the RBD- (yellow) and NTD- (blue) sorted single cell data (Appendix A). Distributions of the lengths of **(C)** HCDR3 (heavy chain), **(D)** KCDR3 (κ-chain), and **(E)** LCDR3 (λ-chain) amino acid sequences are shown. **(F)** IGHJ-gene usage, **(G)** IGKJ-gene usage, and **(H)** IGLJ-gene usage of the sorted single-cells is shown in relative counts in bar graphs. Colors are consistent between panels and the number of samples used to evaluate the statistics in each panel is indicated in the legend. **(I - L)** Circos plots show matches between the light chain CDR3 sequences of progenitors in the sorted single-cell dataset (black) and light chain CDR3 sequences in the verified mAbs (colors) for RBD-reactive (I) IG_κ and (J) IG_λ sequences and for NTD-reactive (K) IG_κ and (L) IG_λ sequences. Different colors indicate different studies from which mAbs were pooled. The reference to each study, the total number of mAbs in the study, and the number of mAbs with matching light chain CDR3 to the sorted single-cell data are reported in each panel.

found in eight individuals (Figure 2.13B).

We also found that two individuals with COVID-19 had exact HCDR3 matches to a previously identified antibody, S304, that has cross-reactivity to SARS-CoV-1 and SARS-CoV-2 [231]. We observed in one individual an HCDR3 with only one amino acid difference to this antibody (Data S3). Importantly, the plasma in these individuals showed a substantial binding level (OD_{450}) to SARS-CoV-1 (Figure 2.9), which indicates possible cross-reactive antibody responses to SARS-CoV-1 and SARS-CoV-2.

We investigated the matches between the RBD- and NTD-sorted single-cell receptors with verified mAbs from previous studies [32, 107, 128, 157, 231, 250, 332, 351]. Although we found no matches between the heavy chain CDR3 of sorted single-cell receptors and the verified mAbs, we found a large number of matches between the κ and λ light chain CDR3s of the sets (Figure 2.14). Notably, 59 of 142 IG_κ and 47 of 110 IG_λ from the RBD-reactive single cells and 1 of 202 IG_κ and 22 of 155 IG_λ from the NTD-reactive single cells matched to light chain CDR3s of mAbs in those respective subsets (Figure 2.14). Given the low sequence diversity of light chain receptors, it remains to be seen whether these matches between the light chain mAbs and sorted single cells are statistically significant–a question that requires modeling the generation and selection of the light chain receptor repertoires.

Last, we observed that previously verified mAbs have a lower probability $P_{\text{post}}(\sigma)$ of generation and entry to the periphery compared with the overall repertoire (Figure 2.13C). This is partly expected because the selection models used to evaluate these probabilities were trained on different repertoires than those from which these mAbs were originally harvested. Consistently, the evaluated probabilities for the sorted single-cell receptors are within the range for the bulk repertoire (Figure 2.13C) because the two datasets were derived from the same cohort. Notably, all the verified mAbs and the sorted single-cell receptors that we can match to the individuals' repertoires have a relatively high probability $P_{\text{post}}(\sigma)$ (Figure 2.13D). This is not surprising as it is unlikely to observe rare BCRs to be shared across different cohorts. Overall, our results are encouraging for vaccine development because they indicate that even common antibodies can confer responses specific to SARS-CoV-2.

2.8 Discussion

COVID-19 will remain an ongoing threat to public health until an effective SARS-CoV-2 vaccine is available globally. Understanding the human B cell immune response to SARS-CoV-2 is critical for vaccine development and assessment [319]. A repertoire of immune receptor sequences represents a unique snapshot of the history of immune responses in an individual [29, 91, 155, 251], and changes in a repertoire during an infection can signal responses specific to pathogens [124, 217]. Identifying signatures of a functional response to a given pathogen from a pool of mostly unspecific BCRs collected from the blood is challenging—it is like finding a needle in a haystack. Therefore, principled statistical inference approaches are necessary to extract functional signals from such data. Here we systematically characterize the B cell repertoire response to SARS-CoV-2 in individuals with different severities of COVID-19 by combining evidence from the overall statistics of repertoires with dynamics of clonal lineages during infection and sharing of immune receptors among individuals.

At the repertoire level, we showed that the HCDR3 of BCRs in individuals with COVID-19 are significantly longer than the HCDR3 in healthy individuals and that the amino acid composition of this receptor region varies among cohorts of individuals with mild, moderate, and severe symptoms. We observed large-scale sharing of BCRs among individuals with COVID-19, consistent with previous findings in those with COVID-19 [87, 214, 266]. Sharing of BCRs among individuals can signal common immune responses to a pathogen. However, BCR sharing can also be due to convergent recombination leading to the same receptor sequence or other experimental biases that influence statistics of shared sequences. These statistical nuances can substantially sway conclusions drawn from the sharing analysis and should be carefully accounted for. Here we established a null expectation of BCR sharing due to convergent recombination by inferring a model of receptor generation and selection. Our analysis identified a subset of rare BCRs shared among individuals with COVID-19, which appears to signal convergent responses to SARS-CoV-2.

Bulk B cell repertoires predominantly contain a mixture of naive, memory, and plasma

B cells. At the early stages of viral infection, antigen-specific plasma B cells may develop that act as antibody factories and confer neutralization against the infecting pathogen [328]. Almost all prior work on immune repertoires has focused on bulk repertoires, which are often easier to sample and analyze. Moreover, functional studies, using single-cell sequencing of antigen-sorted BCRs, have often been disconnected from large-scale analysis of receptor repertoires. Our study synergizes data from bulk and plasma B cell sequencing with antigen-sorted single-cell BCRs to draw a more complete picture of the human immune response to SARS-CoV-2. Importantly, our joint longitudinal analysis of the bulk and plasma B cell repertoires in individuals with COVID-19 provides insight into the dynamics of antigen-specific B cells as well as the statistics of receptor sequence features associated with responses to SARS-CoV-2.

In addition to the statistics of repertoires, we observed that the activity of many B cell lineages (i.e., mRNA production) in individuals with COVID-19 increases during infection, accompanied by an increase in the binding level (OD_{450}) of the individuals' plasma to the RBD and NTD of SARS-CoV-2. The dynamics of clonal lineages during an infection provide significant insights into the characteristics of responsive antibodies [124, 217]. By taking advantage of data collected at multiple time points in most individuals, we identified expanded lineages shared among individuals and found 38 clonal lineages that are candidates for a response specific to SARS-CoV-2 antigens (Figure 2.11; Data S2). Importantly, the over-representation of plasma B cells among these shared expanding lineages signifies their potential role in mounting protective antibody responses against SARS-CoV-2. It should be noted that none of these 38 clonal lineages matched with the verified mAbs. This is in part expected because the verified mAbs that matched the bulk repertoires have relatively high probabilities P_{post} (Figure 2.13D), whereas these 38 lineages are chosen explicitly to be rare.

Our analysis of repertoire dynamics has identified a large-scale expansion of B cell clonal lineages (5%-15% of lineages) over the course of COVID-19 infection. However, it is hard to imagine that all of these expanding clones that account for a sizeable portion of the repertoire are engaged in responding specifically to SARS-CoV-2. In contrast, our single-cell analysis

identified about 0.2% of receptors as reactive to RBD and 0.02% as reactive to NTD epitopes (Figure 2.1)—an estimate that is consistent with previous findings [156]. This disparity raises an open question: why do we observe such a large-scale expansion of clonal lineages during an acute immune response?

Identifying antibodies with cross-reactive neutralization abilities against viruses in the SARS family is of significant interest. Although cross-neutralization antibodies have been isolated from individuals with COVID-19 [32, 175, 349], it remains unclear how prevalent they are. In nine individuals, we see a substantial increase in the binding level (OD_{450}) of their plasma to SARS-CoV-1 epitopes over the course of COVID-19 infection. In three individuals, we identify a BCR identical to the heavy chain of antibody S304 [231], which has been isolated previously from an individual who recovered from a SARS-CoV-1 infection. This antibody has been shown to be moderately cross-reactive to SARS-CoV-1 and SARS-CoV-2, and our results indicate a possibility for such cross-reactive antibodies to emerge naturally in response to SARS-CoV-2 [32, 179, 252]. Our findings provide substantial insight into and strong implications for devising vaccines and therapies with a broad applicability against SARS-CoV-2.

Chapter 3

THE T CELL REPERTOIRE RESPONSE IN INDIVIDUALS WITH POST-ACUTE SEQUELAE OF COVID-19

The post-acute sequelae of SARS-CoV-2 infection presents a broad spectrum of symptoms encompassing a plethora of organ systems. The immunological mechanisms underlying the emergence and persistence of these chronic ailments remain poorly understand despite numerous research fronts due to the diversity of the disease and numerous etiologies. Here, we investigate the role of T cell receptor (TCR) repertoires in PASC by longitudinally profiling peripheral blood TCRs in individuals with PASC and comparing these repertoires with repertoires from individuals who recovered from COVID-19. While the repertoires look similarly globally, we find differences in the presence of TCRs and motifs by honing in on abundant clones, dynamic clones, or highly public clones. We identify 1,091 potentially differentiating receptors which were highly abundant, dynamic, or shared in their respective cohort. Our results provide the basis for detecting T cell receptors potentially specific to individuals with PASC.

3.1 Introduction

Five years since the inception of the COVID-19 pandemic, millions of individuals afflicted previously with SARS-CoV-2 infections have persisting maladies, a condition called post-acute sequelae of SARS-CoV-2 infection (PASC) and, colloquially, Long Covid [59, 99, 193, 225, 139, 63]. PASC is characterized by symptoms spanning a diversity of organ systems and include incessant fatigue, weakened neurocognitive function, persistent hyperglycemia, cardiovascular disease, autonomic dysfunction, endometriosis, and gastrointestinal disorders [182, 146, 222, 300, 334, 146, 7, 237, 335]. The etiologies underlying some of these

symptoms are believed to be uncleared viral reservoirs [291], autoimmune responses [290], chronic inflammation [59], and the reactivation of quiescent viruses [352, 325, 150, 226] among others [59, 198, 341]. PASC presents in individuals regardless of age, sex, or COVID-19 severity, though the likelihood of its presentation is associated positively with the severity an individual's COVID-19 [37]. While methods for preventing PASC mirror those for SARS-CoV-2, significant breakthroughs in diagnosing and treating PASC are hampered by the wide heterogeneity of the disease and unclear pathogeneses [8].

T cells are crucial for clearing infections and providing protection against previously encountered pathogens for years and up to an individual's lifetime [135, 167, 181, 160]. Using their T cell receptors (TCRs), they detect pathogenic peptides presented extracellularly by cells on major histocompatibility complexes. Their TCRs are generated in a stochastic process called V(D)J recombination that ultimately creates an ensemble of T cells with enough functional diversity to recognize a wide spectrum of pathogen-derived peptides. Upon recognition of such peptides, killer T cells will destroy unhealthy or infected cells whereas helper T cells will secrete cytokines to facilitate the killer T cell, B cell, and innate immune system responses [135]. The power and importance of T cells in the adaptive immune system is underscored by recent studies suggesting that T cells alone can control Orthoflavivirus challenges as well as SARS-CoV-2 without B cell responses [144, 243]. T cell dysregulation has been proposed as one of the etiologies of PASC [341], and the resolution of dysregulation in the adaptive immune system has been associated with an improvement in the quality of life of individuals with PASC [230]. Machine learning has revealed different compartments of T cells with different cell surface markers as being associated with the severity of PASC lung diseases and breathlessness [40]. Given that etiologies, such as uncleared SARS-CoV-2 reservoirs and inflammation, might be mediated by T cells directly or indirectly, studying the T cell response in the context of SARS-CoV-2 and PASC is critical for understanding the determinants of PASC and crafting appropriate therapies.

Investigating T cell repertoires has elucidated how the adaptive immune system responds to various pathogens and challenges [77, 232, 197, 196, 247, 233, 234, 315, 94]. Though studies

have been performed by characterizing the T cell response to PASC [259, 290, 150, 174, 324, 223, 341, 230], none have investigated the bulk T cell repertoires of individuals with PASC longitudinally. Examining bulk T cell repertoires collected longitudinally provides a view into the T cell response with incredible depth and permits the tracking of clones across time. Understanding the dynamics of T cells in the context of PASC and identifying commonalities in TCRs observed in individuals with PASC could provide further insight for characterizing PASC pathogenesis.

In this study, we characterize the T cell repertoires collected longitudinally from a cohort of individuals with heterogenous symptoms of PASC and individuals who contracted SARS-CoV-2 but did not develop PASC using principled global and local statistical analyses. We examine different subsets of the repertoire, such as expanding and contracting clones, large clones, and widely shared clones, to probe for T cell clones that might be relevant to the presentation of PASC, scrutinizing the differences in molecular composition and features among TCRs between the two cohorts. Ultimately, we identify 1,091 TCRβ receptors that are candidates for potentially being specific to PASC+ or PASC− responses.

3.2 Overview of the cohort

The cohort of individuals studied here is a subset of the INCOV cohort. The INCOV cohort contained 209 individuals who presented a broad range of COVID-19 symptoms and was assembled by a collaboration between the Institute for Systems Biology and the Swedish Medical Center to study the effects of SARS-CoV-2 originally [289] and later on PASC [290]. Peripheral blood mononuclear cells were isolated from 120 of the 209 individuals in the INCOV cohort and sent to Adaptive Biotechnologies to sequence the T cell receptor β (TCRβ) locus from bulk gDNA (Appendix B). The times at which the cohort's repertoires were sampled display a large amount of heterogeneity, with two to three samples taken two days to half a year (169 days) after the onset of symptoms from their initial encounter with SARS-CoV-2 (Fig. 3.1B-D). The ages of individuals in our cohort range from 19 to 86, with ages well-represented across subsets of the cohort stratified by PASC status and sex (Fig. 3.1E). The

individuals in this cohort had bouts with SARS-CoV-2 during the first year of the COVID-19 pandemic and displayed a diversity of severities (asymptomatic to severe) over the course of their infection and recovery as measured by the WHO ordinal severity scale (WOSS) [185]. Fig. 3.1F shows the distribution of WOSS measurements for observations within nearly three weeks, between three weeks and two months, and after two months since symptom onset. Only the distributions of WOSS values from the PASC−, female individuals and PASC+, male individuals in the first window of time differed significantly (Mann-Whitney $U = 469$, $n_{-,F} = n_{+,M} = 25$, Bonferroni-corrected p-value $= 0.036$). Males presented slightly higher severities on average within three weeks since symptom onset compared to females, consistent with previous observations [147, 276]. Overall, median severities were mild across groups stratified by PASC status and sex regardless of when severity was measured after three weeks since symptom onset.

There is no clear consensus on what exactly defines PASC [56, 248, 109, 294, 8]. Here, we adopt the definition of PASC as having at least one of the following symptoms present at least two months after one's initial SARS-CoV-2 infection: shortness of breath, inability to exercise, a persistent cough, excretion of sputum, fatigue, diarrhea, nausea, abdominal pain, loss of taste, or loss of smell. Though these criteria span many organ systems—cardiovascular, neuromuscular, gastrointestinal, respiratory, etc.—and the underlying causes may differ [59], we partition the cohort using the presence of at least one symptom due to a high overlap of symptoms among individuals and statistical limitations imposed by the small size of the cohort (Fig. 3.1A). The PASC+ and PASC− cohorts consisted of 71 (58% female, median age 53) and 49 individuals (57% female, median age 50), respectively (Fig. 3.1E). We quantified differences in the display of symptoms using the Jaccard distance and performed Ward clustering. We found that symptoms grouped as expected and observed five clusters of symptoms—diarrhea and nausea (gastrointestinal); loss of taste and smell (anosmia/dysgeusia); abdominal pain; cough and excretion of sputum (respiratory); and fatigue, shortness of breath, and inability to exercise (neurological). We sought to decipher relationships between sex and symptoms and age and symptoms using mutual information.

The mutual information quantifies the statistical dependence between two random variables, i.e., how much uncertainty about the outcome of an event is reduced given another event has been observed [51].

$$I(x,y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3.1}$$

The significance of the observed mutual information can be computed using a randomization test by permuting the features of the dataset and then computing the mutual information [285]. Even though correlations are destroyed by shuffling, the null dataset will have nonzero mutual information due to finite size effects. The observed mutual information on the unshuffled data can be compared to the distribution of mutual information obtained from an ensemble of permuted datasets. A p-value can be computed as $1 - q_{I(x,y)}$, where $q_{I(x,y)}$ is the quantile of the mutual information of the unshuffled dataset relative to the null distribution. Using ensembles of 1,000 permuted datasets for each sex-symptom and age-symptom comparison and after applying Bonferroni-corrections [212], we did not find any meaningful associations with symptom presentation and sex or age (Fig. 3.1A).

3.3 Sequences features and repertoire composition do not differ among cohorts

We sought to determine how PASC shapes the repertoires at the level of coarse-grained features, such as TRβV gene usages, TRβJ gene usages, and TCRβ CDR3 amino acid sequence lengths. These sequence features, while also associated with an individual's HLAs, can contain information pointing to a receptor's binding specificity. We compared these features at the level of unique TCRβ recombinations (Fig. 3.2) and TCRβ recombinations weighted by cell counts (Fig. B.1), aggregating feature statistics within a group of individuals with the same PASC status and sex. While we observed variability in gene usages among cohorts stratified by PASC status and sex, no significant differences were observed among the cohorts (Mann-Whitney U test with FDR-BH corrections [116]). TCRβ CDR3 length distributions from PASC+ individuals were statistically indistinguishable from the TCRβ CDR3 length distributions derived from repertoires of PASC− individuals. As a control, we compared the

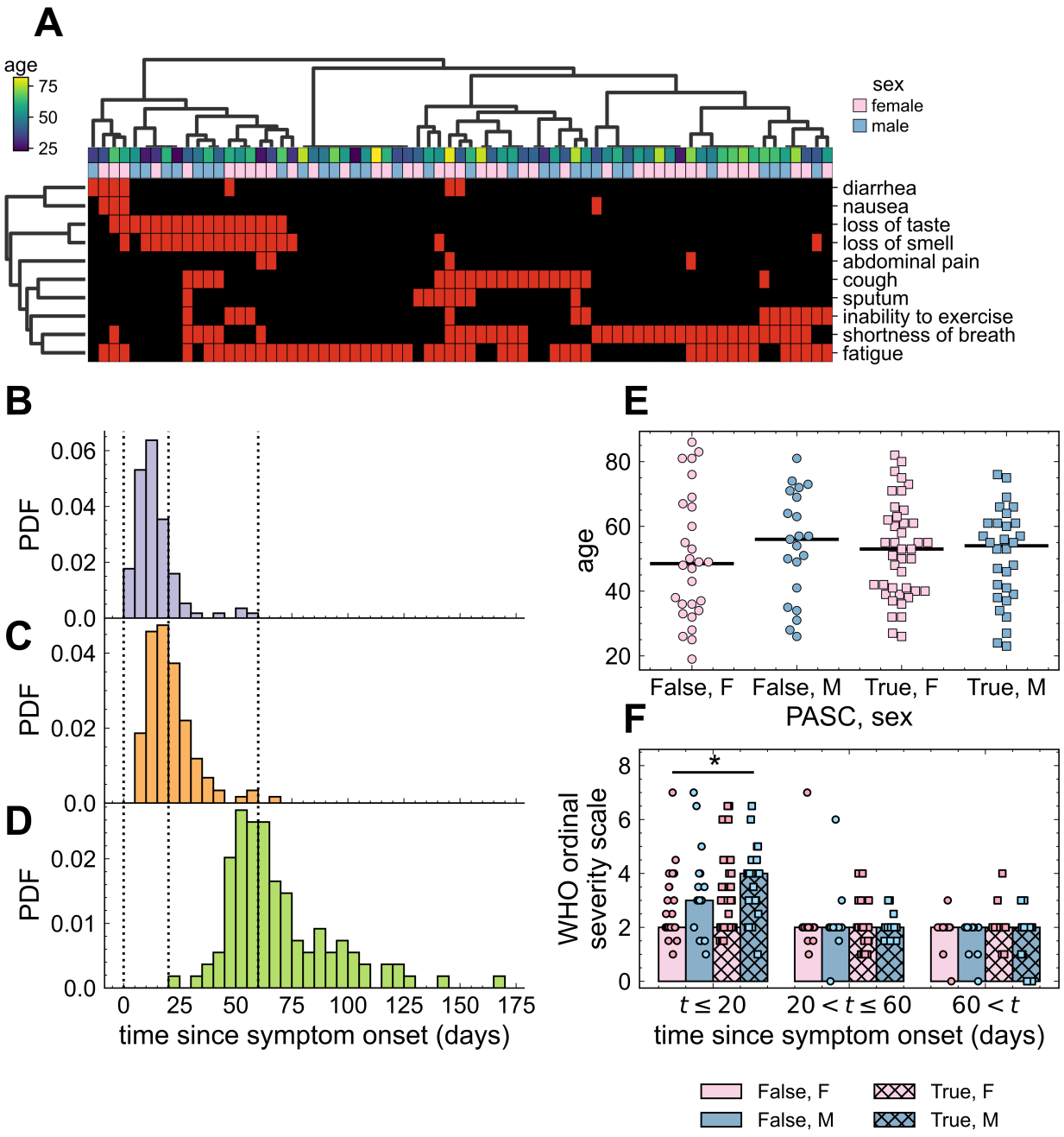

A
age
75
50
25
sex
female
male
diarrhea
nausea
loss of taste
loss of smell
abdominal pain
cough
sputum
inability to exercise
shortness of breath
fatigue
B
C
D
PDF
0.06
0.04
0.02
0.0
0.04
0.02
0.0
0.02
0.01
0.0
0
25
50
75
100
125
150
175
time since symptom onset (days)
E
age
80
60
40
20
False, F
False, M
True, F
True, M
PASC, sex
F
WHO ordinal severity scale
8
6
4
2
0
*
t ≤ 20
20 < t ≤ 60
60 < t
time since symptom onset (days)
False, F
False, M
True, F
True, M

Figure 3.1: **PASC symptoms, sampling time distributions, and cohort demographics. (A)** PASC symptoms (y-axis) presented by individuals in the PASC+ cohort (x-axis) are shown as a clustermap created using Ward clustering by Jaccard distances between sets of symptoms. Red rectangles display which symptoms were observed. The left dendrogram shows the relationships among symptom presentation whereas the top dendrogram shows clusters of individuals presenting similar symptoms. The colored rectangles directly above the heatmap show the sex (blue for male, pink for female) of each individual in the PASC+ cohort, and the viridis-colored rectangles show the age of each individual in the PASC+ cohort, with the associated colorbar displayed on the on the left. **(B - D)** The histograms show the distributions of the times at which repertoires were sampled relative to an individual's onset of symptoms. (B) shows the distribution for the earliest sampled repertoires, (C) shows the distribution for the second set of sampled repertoires, and (D) shows the distribution for the third sample. The dotted vertical lines are drawn at x values of 0, 20, and 60 to indicate how samples are binned. **(E)** The swarmplot shows the distribution of ages of the PASC+ and PASC− cohorts stratified by sex (x-axis, colors, markers). Horizontal lines show the median age of each swarm. **(F)** The WHO ordinal severity scale (WOSS) measures the disease severity of an individual via their respiratory status [185]. Bars show the median WOSS value (y-axis) for each group partitioned by PASC status (hatches) and sex (colors as in (E)) for each binning of observation times (x-axis). The markers (shapes as in (E)) depict the full distribution of WOSS values. *: Bonferroni-corrected p-value < 0.05.

repertoires' TCRβ CDR3 length distributions to those from repertoires of healthy individuals (and without CMV) from the Emerson cohort [77]. We observed that the means of the TCRβ CDR3 distributions in individuals who had SARS-CoV-2 were about half an amino acid longer and statistically significant (two-sided t-test with Bonferroni corrections) compared to the healthy cohort (t statistics and p-values in Table B.1). Therefore, PASC does

not appear to influence the peripheral blood T cell repertoire in a global manner distinct from acute COVID-19.

We next investigated the spread of clone sizes, here, interpreted as the number of cells, in repertoires of PASC+ and PASC− individuals using a statistic called clonality. Clonality is a measure used to quantify how imbalanced the distribution of clones are in a sample based on clonal abundances. The frequency of the i-th clone is given by its abundance, a_i, normalized by the total abundance of the sample.

$$f_i = a_i \Big/ \sum_i a_i \tag{3.2}$$

The clonality is defined here using the normalized Shannon entropy:

$$\mathcal{C} = 1 + \frac{\sum_{i=1}^{N} f_i \log f_i}{\log N}, \tag{3.3}$$

where N is the number of clones in a sample. By normalizing the observed entropy, i.e., $-\sum_{i=1}^{N} f_i \log f_i$, with the maximal entropy, $\log N$, clonality accounts for the heterogenous sizes that sampled repertoires may have, placing all compared clonalities on the same scale. Conversely, Shannon entropy alone is an extensive quantity, so it is highly contingent on the depth of a sample. $\mathcal{C} = 0$ indicates each clone has the same abundance whereas $\mathcal{C} \approx 1$ indicates one or a few clones are overwhelmingly abundant compared to all the other clones. Other definitions of clonality have been used to understand lymphomas [329], cytomegalovirus (CMV) and graph-versus-host disease [39], and how HLA shapes the TCR repertoire [158]. The distribution of clonalities binned by observation time relative to symptom onset (as in Fig. 3.1F) is shown in Fig. 3.3A. No distributions of clonalities were statistically different from one another (Mann-Whitney U test with Bonferroni corrections, $p > 0.05$). We also studied how clonality changes over time. Fig. 3.3B shows the log-fold change in clonality among the three coarse-grained bins of days since symptom onset. While it appears that the repertoires of PASC+ individuals have become less uneven than PASC− individuals, none of these distributions of log-fold changes are significantly different from one another (Mann-Whitney U test with Bonferroni corrections, $p > 0.05$). This suggests

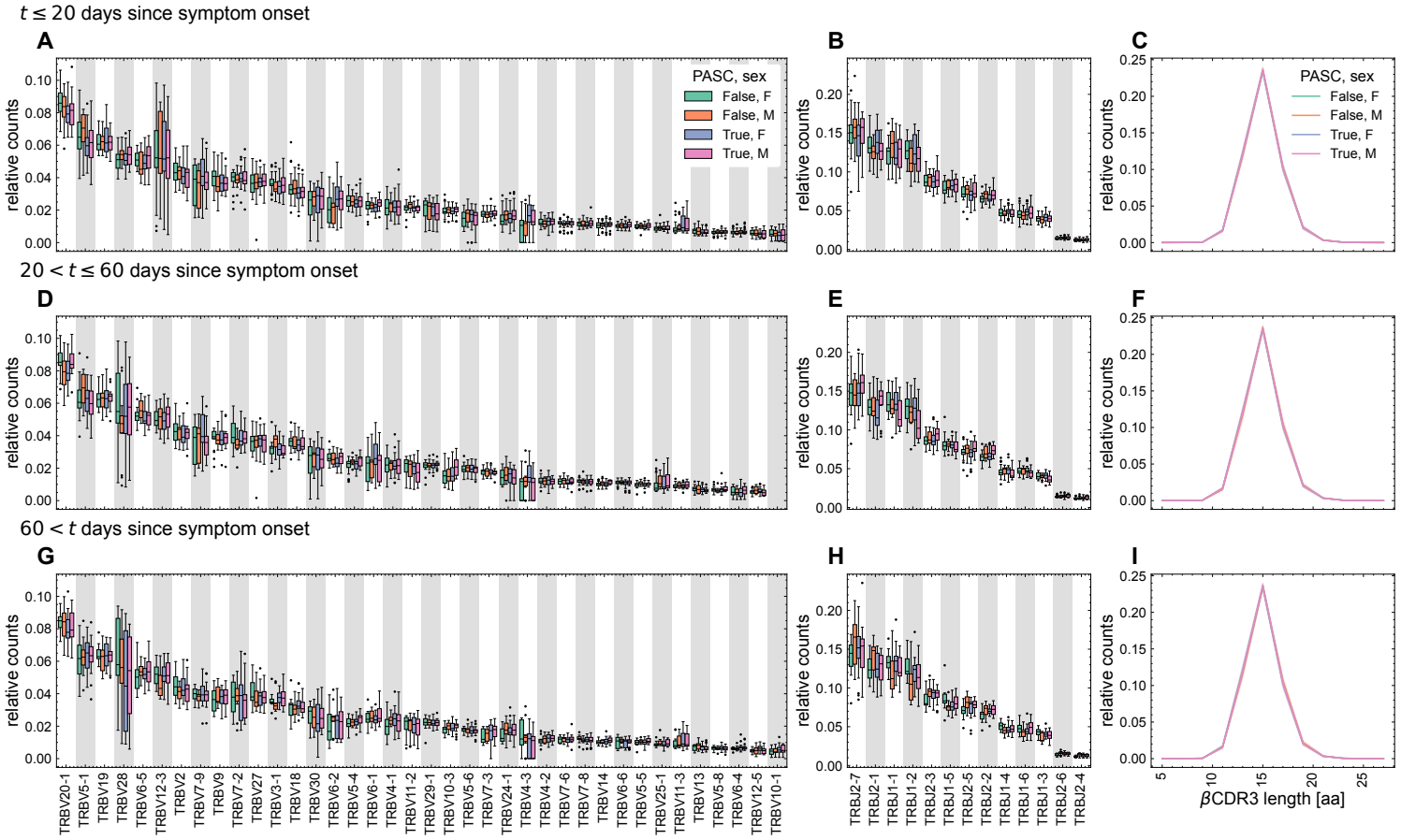


Figure 3.2: **Receptor composition at the level of unique recombinations across cohorts. (A, D, G)** The distribution of TRβV gene usages is shown as boxplots, stratified by PASC status and sex (colors) and sampling time relative to symptom onset as indicated above each triple of plots. Interquartiles are depicted in the box, with the median shown as a black, horizontal line. The rest of the distribution is shown as whiskers, and outliers are plotted as black dots. TRβV genes are shown only if their usages were above 1% of the entire repertoire in any one of the cohorts. **(B, E, H)** The distribution of TRβJ gene usages is shown as boxplots as in (A, D, G). TRβJ genes are shown only if their usages were above 1% of the entire repertoire in any one of the cohorts. **(C, F, I)** The distribution of amino acid TCRβ CDR3 lengths is shown. Lines indicate average relative counts for each cohort, and shading indicates regions containing one standard deviation of variation across individuals within a cohort.

that PASC does not modulate the distribution of clone sizes in a manner distinct from acute COVID-19.

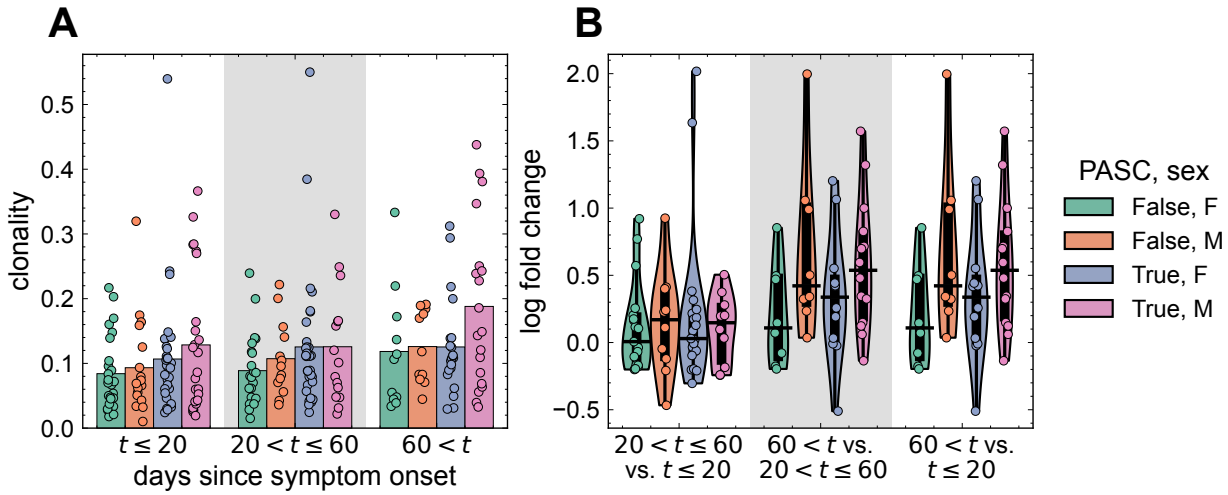


Figure 3.3: **Clonalities across cohorts and sexes. (A)** The distribution of clonalities of the TCRβ repertoires is shown stratified by PASC status and sex (colors) as well as by sampling time relative to symptom onset (x-axis). The bars indicate the clonality averaged over individuals in each PASC-sex cohort, and dots indicate the variation in clonalities across individuals within each PASC-sex cohort. **(B)** The distribution of log-fold changes in clonality is shown as a combination of violinplots and swarmplots, with the medians demarcated by a horizontal black line. Colors are as in (A). The x-axis denotes the numerator of the fold change and the denominator of the fold change as "numerator vs. denominator."

3.4 Motif enrichment among the most populated clones

Given the global repertoire statistics showed no signatures of any significant discrepancies between PASC+ and PASC− individuals, we proceeded by studying the TCRβ repertoires in a much more local fashion by examining specific subsets of the repertoire and identifying

sequences specific to each cohort.

3.4.1 TCR-OT: comparing repertoires using motifs

Due to the high-dimensional nature of TCRβ amino acid sequences and the smaller sizes of these subsets, we used a nonparameteric, interpretable statistical method based on optimal transport (TCR-OT) [220] that balances interpretability with biologically meaningful predictions. TCR-OT takes as input two repertoire datasets—$\mathcal{R}_1$, $\mathcal{R}_2$—each containing deduplicated, independent V(D)J recombination events, i.e., TCR sequences defined by their TRβV gene, TRβJ gene, TCRβ CDR3 nucleotide sequence, and individual from which they were sampled. Both datasets undergo preprocessing which deduplicates the TCRs at the level of amino acid sequences (TRβV gene and βCDR3 amino acid sequence) and retains their multiplicities. Probability mass distributions for each dataset, $\mathcal{P}_1, \mathcal{P}_2$, are formed by normalizing the multiplicities of the amino acid sequences in their respective datasets. A distance matrix D of TCRdists [57, 187] (Appendix B) is computed between all pairs of TCRs in $\mathcal{R}_1$ and $\mathcal{R}_2$. Using the probability mass distributions and the distance matrix D, TCR-OT finds the optimal transport plan $\Pi^\star$. A transport plan Π for datasets $\mathcal{R}_1$, $\mathcal{R}_2$ is a joint probability distribution whose marginals are $\mathcal{P}_1$ and $\mathcal{P}_2$, i.e.,

$$\mathbf{1}_j \Pi_{ij} = \mathcal{P}_1, \quad \mathbf{1}_i \Pi_{ij} = \mathcal{P}_2, \tag{3.4}$$

where $\mathbf{1}$ is a vector of ones and summation is performed as in Einstein notation. The optimal transport plan $\Pi^\star$ is the joint probability distribution whose marginals are $\mathcal{P}_1$ and $\mathcal{P}_2$ and which minimizes the Frobenius inner product between itself and the distance matrix.

$$\Pi^\star = \text{argmin}_{\Pi \in U(\mathcal{P}_1, \mathcal{P}_2)} \langle \Pi, D \rangle_{\text{F}}, \tag{3.5}$$

where $U(\mathcal{P}_1, \mathcal{P}_2)$ is the space of all transport plans with row- and column-wise marginals $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. In practice, TCR-OT finds the optimal transport plan by regularizing the Frobenius norm by the entropy of the transport plan for speedier convergence of the

algorithm and for obtaining a more parsimonious solution.

$$\Pi^\star = \text{argmin}_{\Pi \in U(\mathcal{P}_1, \mathcal{P}_2)} \langle \Pi, D \rangle_{\text{F}} - \lambda H(\Pi), \tag{3.6}$$

where $H(\Pi) = -\sum_{i,j} \Pi_{ij} \log \Pi_{ij}$ is the entropy of plan Π, and λ is the strength of the regularization. The Wasserstein distance matrix is the Hadamard product of $\Pi^\star$ and D.

$$W = \Pi^\star \odot D \tag{3.7}$$

It gives the minimum "effort" that must be exerted to move mass from one probability distribution to the other.

For a given TCR sequence $\sigma \in \mathcal{R}_1$, its set of neighbors is given by

$$\mathcal{N}(\sigma) = \{\sigma' \mid \text{TCRdist}(\sigma, \sigma') \leq t\}. \tag{3.8}$$

We mirror [220] and let $t = 48$, which amounts to a difference of about 4 amino acids in the TCRβ CDR3 amino acid sequences or 12 amino acid differences between the TRβV genes of σ and σ'. The TCR-OT enrichment statistic for σ is computed as the overall Wasserstein distance between a neighborhood of TCRs around σ, $\mathcal{N}(\sigma)$, and all the TCRs in $\mathcal{R}_2$

$$s(\sigma) = N_1 \sum_{\sigma' \in \mathcal{N}(\sigma)} \delta(\sigma')_i \mathbf{1}_j W_{ij}, \tag{3.9}$$

where $\delta(\sigma')$ is a vector of 0s except at the position indexed by σ' where it is 1, and scaling the statistic by the size of the repertoire N_1 ensures the statistic is comparable across datasets of different sizes. Enrichment statistics for sequences in $\mathcal{R}_2$ are computed similarly.

To quantify the significance of a TCR's enrichment, TCR-OT performs a randomization test. Each realization of the randomization test shuffles the labels that specify whether an independent recombination comes from $\mathcal{R}_1$ or $\mathcal{R}_2$. The enrichment scores are computed on this shuffled dataset with the procedure above. After generating many realizations, a distribution of enrichment scores is obtained for each input TCR, characterizing the enrichment scores we would expect if $\mathcal{R}_1$ and $\mathcal{R}_2$ were samples from an identical population of TCRs.

A p-value is calculated for each TCR sequence using the quantile of its observed enrichment score—the score computed using the unshuffled labels—relative to this distribution. Randomizations were performed until each TCR was sampled in their respective dataset 500 times. BH-corrections [116] were applied with a BH-corrected $p < 0.05$ indicating a sequence is significantly enriched.

TCR-OT also constructs clusters of TCRs that appear to be enriched in their respective dataset. Its algorithm is as follows. Consider $\mathcal{R}_1$ which can be partitioned into a set which contains TCRs which have been clustered already $\mathcal{R}_{1,c}$ and those which haven't $\mathcal{R}_{1,\tilde{c}}$. Initially $\mathcal{R}_{1,c} = \emptyset$ and $\mathcal{R}_{1,\tilde{c}} = \mathcal{R}_1$. Consider the most enriched TCR in $\mathcal{R}_{1,\tilde{c}}$: $\sigma^\star$. Moving away from $\sigma^\star$ in steps of Δr TCRdists until $r_{\max}$, we identify the subset of TCRs within the r_i and $r_i + \Delta r$ annulus and compute the mean enrichment m_i

$$\mathcal{A}_i = \{\sigma | r_i \leq \text{TCRdist}(\sigma^\star, \sigma) < r_i + \Delta r\}, \tag{3.10}$$

$$m_i(\sigma^\star) = \langle s(\sigma) \rangle_{\mathcal{A}_i(\sigma^\star)}, \tag{3.11}$$

where $i \in [0, 1, ..., i_{\max}]$ indexes the annulus we are studying, and $i_{\max} r = r_{\max}$. Using segmented linear regression [204], a breakpoint in the mean enrichment vs. TCRdist annulus radii relationship can be detected which demarcates when the mean enrichment has saturated (Fig. B.2). All TCRs with TCRdist less than this breakpoint are clustered with the focal TCR $\sigma^\star$, and $\mathcal{R}_{1,c}$ and $\mathcal{R}_{1,\tilde{c}}$ are updated accordingly. TCRs clustered together in this manner are believed to signal similar function since focal sequences were characterized as atypical with respect to the reference repertoire via the TCR-OT enrichment statistic. TCR-OT's clustering proceeds by finding a pre-specified number of clusters or until the segmented linear regression fails to identify a breakpoint. In this analysis, $\Delta r = 5$ and $r_{\max} = 200$ were chosen. We retained only clusters with cluster radius $r_{\text{cluster}} \leq 120$, corresponding to a difference of at most 10 amino acids in the TCRβ CDR3 sequence assuming the same TRβV gene, for downstream analyses. This choice balances sequence diversity within a cluster with intolerance toward constituent cluster sequences being too functionally dissimilar to the focal sequence. Moreover, segmented linear regressions producing $r_{\text{cluster}} > 120$ ensued from

performing inference on sparse subsets of the dataset, so this conservative choice ensures we examined only robust clusters downstream. (Most cluster radii inferred in this study were between 50 and 100 TCRdist. This corresponds to roughly at most 4 and 8 amino acid differences, respectively, in the TCRβ CDR3 sequences relative to the cluster focal sequence. See. Figs. B.2, B.7, B.8.)

3.4.2 Differential statistics of the most abundant clones between PASC+ and PASC− individuals

We collected the top rank-50 most abundant functional clones, defined using TCRβ CDR3 amino acid sequences and TRβV genes, in each individual who was observed at least 60 days post symptom onset (see Appendix B for how top rank-50 clones were collected). (We ignore the TRβJ gene since it has been shown that it is redundant when the TCRβ CDR3 sequence is studied [110].) We filtered the dataset to include only individuals who were observed at least 60 days post symptom onset because we did not want the dataset to be dominated by clones that were not observed well after the onset of PASC in PASC+ individuals or convalescence in PASC− individuals. Having an imbalanced dataset with regard to sampling time might, for instance, lead to washing out signals of motifs associated with TCRs that were abundant at later times and participated in immune dysregulation in PASC+ individuals or motifs associated with populations of TCRs whose presence staved off PASC in PASC− individuals. Being the most populated, these clones were likely to have expanded relative to when individuals' repertoires were sampled and thus may be associated with SARS-CoV-2 responses or the development of PASC. This resulted in 2,628 functional clones (3,357 independent βCDR3 recombinations) from the PASC+ cohort and 1,535 functional clones (2,019 independent βCDR3 recombinations) from the PASC− cohort. We first inspected the coarse-grained receptor composition statistics in the PASC+ and PASC− cohorts at this level of the repertoire (Fig. 3.4A-C). No significant differences were observed among the distributions of gene usages or TCRβ CDR3 lengths. We detected 159 functional clones in the PASC+ and 105 functional clones in the PASC− cohorts that

were significantly atypical in amino acid composition relative to the other respective cohort (FDR BH-corrected $p < 0.05$) (Fig. 3.4D). We applied the TCR-OT algorithm for clustering around functional clones with the highest TCR-OT statistics and obtained 20 clusters for each cohort (Fig. 3.4F,G, B.2). The sizes of the clusters varied from 2 to 147 functional clones (2 to 196 independent recombinations) in the PASC+ most abundant clones and 1 to 57 functional clones in the PASC− subset (7 to 128 independent recombinations). Because the significances of the sequences varied within clusters (Fig. 3.4E,F,G) and to ensure clusters reflected a degree of collective neighborhood amplification, we refined which clusters we examined downstream by restricting the clusters to have at least 20% of their constituent functional clones be significant. We chose this fraction based on the median fraction of sequences which were significant in clusters observed in the PASC− cohort (Fig. 3.4E). Since we were interested in more general trends to distinguish molecular features of TCRs between PASC+ and PASC− repertoires, we further refined clusters examined downstream by requiring that recombinations from at least two individuals be present in each TCR-OT cluster (Fig. 3.5C,D).

Sequence logos constructed from neighborhoods with a radius of 48 TCRdist (at most 4 amino acid differences between TCRβ CDR3 sequence or at most 12 amino acids differences between TRβV genes) from the focal sequences of significant TCR-OT clusters are shown in Fig. 3.5A,B. (Logos of complete TCR-OT clusters are shown in Fig. B.3A,B.) Similar to the TRβV genes associated with SARS-CoV-2-reactive TCRs discovered by ref. [236], we saw the usage of TRβV5-5 and TRβV20-1, in the PASC+ clusters and TRβV5-1, TRβV7-9, TRβV20-1 and TRβV29-1 in the PASC− clusters (Fig. 3.5A,B). The frequencies of physicochemical properties, such as charge or polarity, among TCRβ CDR3 sequences in the TCR-OT clusters were not significantly different between PASC+ and PASC− (Mann-Whitney U test, Bonferroni-corrected p-values) (Fig. B.3). To associate function with these clusters, we matched sequences in the TCR-OT clusters to sequences previously reported as specific to SARS-CoV-2, Epstein-Barr Virus (EBV), Cytomegalovirus (CMV), or InfluenzaA in the VDJdb [275, 98]. We defined a match as a TCRβ sequence being within 24 TCRdist (roughly

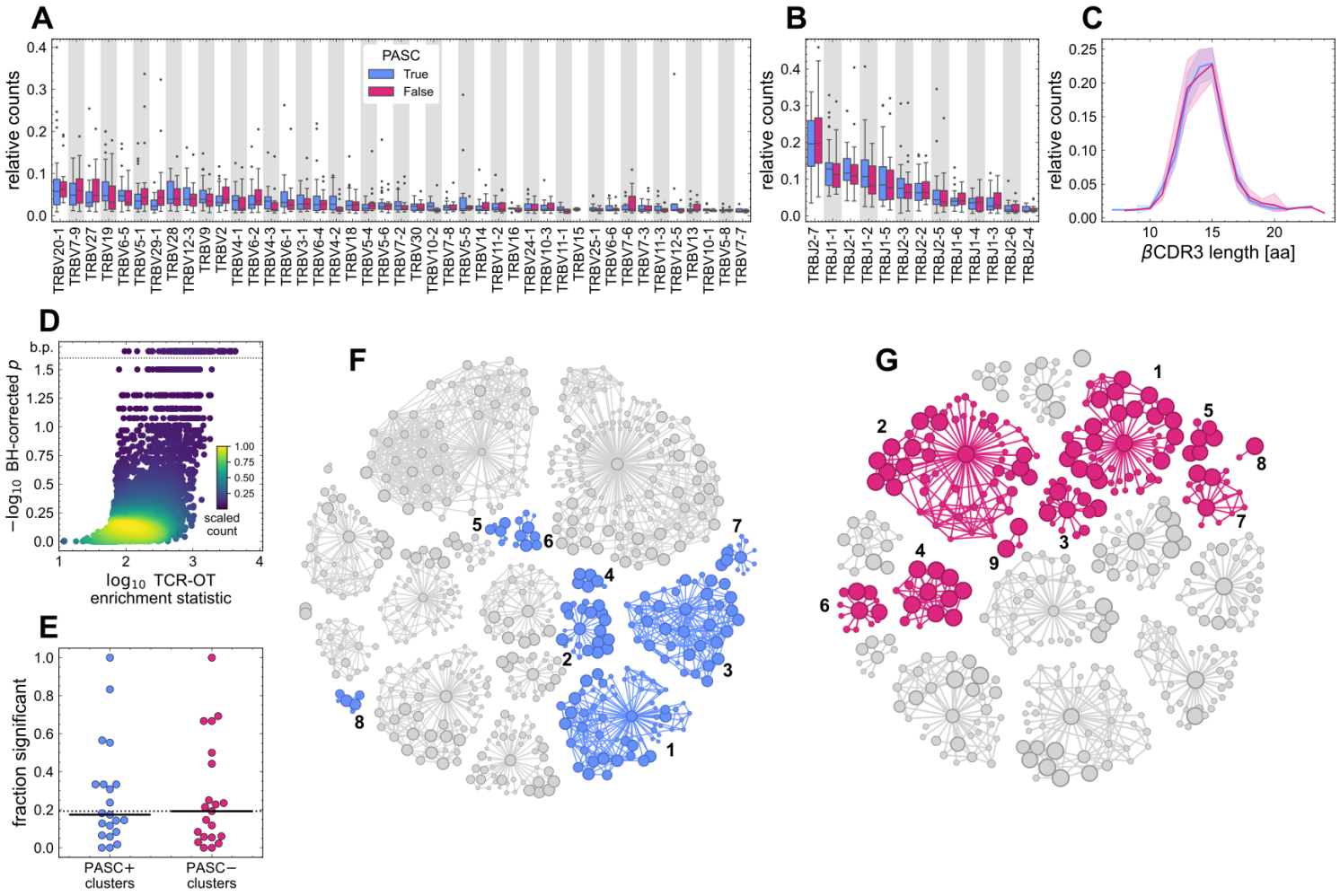

A
relative counts
PASC
True
False
TRBV20-1
TRBV7-9
TRBV27
TRBV19
TRBV6-5
TRBV5-1
TRBV29-1
TRBV28
TRBV12-3
TRBV9
TRBV2
TRBV4-1
TRBV6-2
TRBV4-3
TRBV6-1
TRBV3-1
TRBV6-4
TRBV4-2
TRBV18
TRBV5-4
TRBV5-6
TRBV7-2
TRBV30
TRBV10-2
TRBV7-8
TRBV5-5
TRBV14
TRBV11-2
TRBV16
TRBV24-1
TRBV10-3
TRBV11-1
TRBV15
TRBV25-1
TRBV6-6
TRBV7-6
TRBV7-3
TRBV11-3
TRBV12-5
TRBV13
TRBV10-1
TRBV5-8
TRBV7-7
B
relative counts
TRBJ2-7
TRBJ1-1
TRBJ2-1
TRBJ1-2
TRBJ1-5
TRBJ2-3
TRBJ2-2
TRBJ2-5
TRBJ1-6
TRBJ1-4
TRBJ1-3
TRBJ2-6
TRBJ2-4
C
relative counts
βCDR3 length [aa]
D
b.p.
−log10 BH-corrected p
scaled count
log10 TCR-OT enrichment statistic
E
fraction significant
PASC+ clusters
PASC− clusters
F
G

Figure 3.4: **Top rank-50 receptor composition and TCR-OT clusters. (A)** The distribution of TRβV gene usages is shown as boxplots, stratified by PASC status (colors). Interquartiles are depicted in the box, with the median shown as a black, horizontal line. The rest of the distribution is shown as whiskers, and outliers are plotted as black dots. **(B)** The distribution of TRβJ gene usages is shown as boxplots as in (A). **(C)** The distribution of amino acid TCRβ CDR3 lengths is shown. Lines indicate average relative counts for each cohort, and shading indicates regions containing one standard deviation of variation across individuals within a cohort. **(D)** The volcano plot shows the $\log_{10}$ FDR BH-corrected p-value vs. the $\log_{10}$ TCR-OT enrichment statistic for each functional clone tested. Points above the dotted line, annotated along the y-axis as having a value of b.p. were assigned p-values beyond the precision of the randomization test, i.e., a value of 0. **(E)** The swarm plot shows the distribution of the fraction of functional clones which were deemed significant within a TCR-OT cluster for each cohort (colors). The black horizontal lines show the median for each distribution. Markers above the horizontal, dotted line indicate clusters which have more than 20% of their constituent clones called as significant. **(F, G)** The network plots show the topology of the TCR-OT clusters for the PASC+ (F) and PASC− (G) cohorts. Sizes of the nodes show the strength of the $-\log_{10}$ FDR-BH corrected p-values associated with a functional clone, with larger sizes indicating more significantly atypical functional clones. Edges between nodes and the cluster's focal sequence indicate sequence similarity within the TCR-OT inferred cluster radius, and all other edges indicate sequence similarity within 48 TCRdist. Network topologies with colored nodes had at least 20% of their nodes deemed significant and had recombinations from at least two individuals. The numbers next to each colored cluster indicate the ranking of the clusters with respect to the cluster sizes, with 1 indicating the cluster with the largest amount of clones.

2 amino acid differences in the TCRβ CDR3 amino acid sequence if the TRβV genes are identical) from a VDJdb TCRβ sequence and also required that the individual from which the sequence originated had at least one HLA allele group in common with the reported donor's HLA allele groups (Fig. 3.5E). We analyzed TCRβ sequences in our repertoires for these pathogens since EBV [96, 150, 253, 290] and CMV [44] viral or immune reserviors have been hypothesized to be reactivated in PASC+, though studies are not completely concordant in these observations. In particular, ref. [290] observed no associations with CMV viremia and hypothesized bystander action, ref. [117] did not observe elevated IgM or IgG antibodies specific for EBV in PASC+ individuals with less severe cases of COVID-19, and ref. [44] observed elevated anti-CMV and anti-EBV IgG antibody levels, though serum positivity was consistent between PASC− and PASC+ individuals. Of the sequences in the PASC+ clusters that matched to the VDJdb, we observed 1–7 matches to CMV, 1–4 matches to InfluenzaA, 1–6 matches to SARS-CoV-2, and 2–4 matches to EBV; of the sequences in the PASC− clusters that matched to the VDJdb, we observed 1–14 matches to CMV, 1–8 matches to InfluenzaA, 1–11 matches to SARS-CoV-2, and 1–7 matches to EBV. While TCRβ sequences with known InfluenzaA epitopes were slightly enriched in PASC+ clusters and those with CMV epitopes were slightly enriched in PASC− clusters, these enrichments were not statistically different (Mann-Whitney U test, Bonferroni-corrected $p > 0.5$) (Fig. 3.5).

3.5 Characterizing responses longitudinally by inspecting expansion and contraction

Characteristic behaviors of T cell dynamics are well-documented and show that T cell populations engaging with an immune challenge reach capacity nearly two weeks after the challenge's onset [195, 233, 295]. T cells specific to SARS-CoV-2 have been detected roughly a week after symptom onset for individuals with mild COVID-19 and over two weeks for individuals with more severe cases [23]. Refs. [322, 266, 196] also showed that following SARS-CoV-2 infections, T cell populations in some individuals may continue to expand well over two weeks, reaching their peak over a month after the onset of the infection. In this

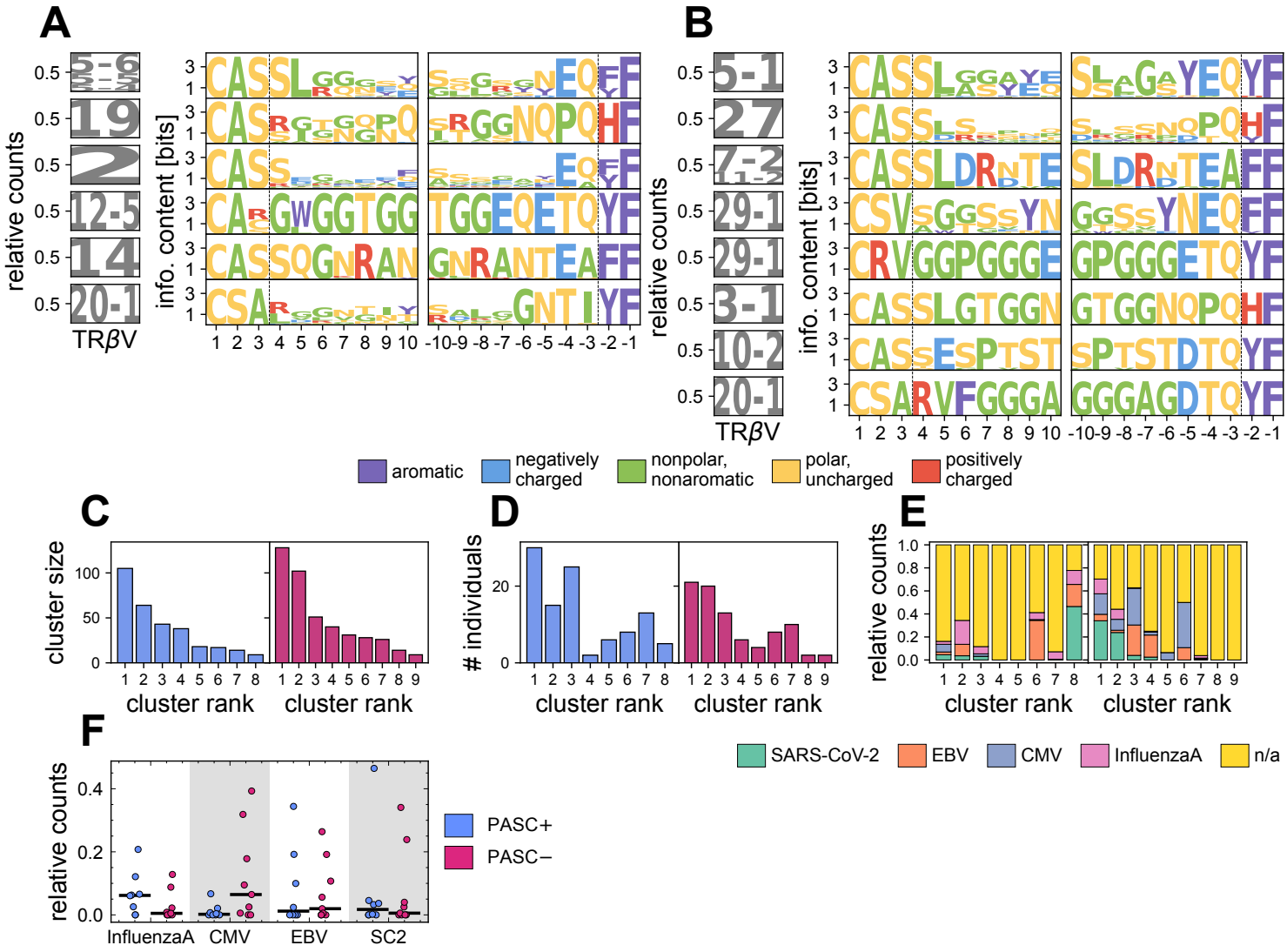

A
B
relative counts
info. content [bits]
TRβV
5-6
19
2
12-5
14
20-1
5-1
27
7-2
29-1
29-1
3-1
10-2
20-1
1 2 3 4 5 6 7 8 9 10 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1
aromatic
negatively charged
nonpolar, nonaromatic
polar, uncharged
positively charged
C
cluster size
cluster rank
D
individuals
cluster rank
E
relative counts
cluster rank
SARS-CoV-2
EBV
CMV
InfluenzaA
n/a
F
relative counts
InfluenzaA
CMV
EBV
SC2
PASC+
PASC−

Figure 3.5: **Sequence motifs and VDJdb matching for significant top rank-50 TCR-OT clusters. (A)** Sequence logos describing the composition of amino acid sequences are shown for the significant TCR-OT clusters obtained from the PASC+ cohort. Logos were constructed from focal sequence(s) and members in the cluster that were within 48 TCRdist of the focal sequence(s). Logos are ordered top to bottom, with the top logo describing the cluster with the largest size and the bottom logo describing the cluster with the smallest size. Only clusters with at least 10 clones within a neighborhood of 48 TCRdist of the focal sequence(s) are plotted. (Left) TRβV gene usage frequencies within a cluster are shown as a logo. (Middle, right) The logo plots shown the information content at each position in the TCRβ CDR3 sequence from left to right (middle) and right to left (right). Amino acids are colored according to their physicochemical properties indicated in the legend. The first three positions and last two positions of the amino acid sequence do not enter into TCRdist calculations (Appendix B), and a vertical, dashed line demarcates where this occurs in each respective logo. **(B)** Logos are shown as in (A) but for clusters obtained from the PASC− cohort. **(C)** The amount of independent recombinations in the clusters for each cohort (colors) is shown as a bar graph. Cluster rank corresponds to the size of the clustering. **(D)** The number of individuals represented in a cluster for each cohort (colors) is shown as a bar graph. **(E)** The stacked bar graphs show the distribution of epitopes of TCRβ sequences in VDJdb that were within 24 TCRdist of a constituent TCRβ in a cluster and whose donor HLA group matched to at least one HLA group of the individual from which the similar TCRβ sequence originated. The left bar graph shows the distribution for the PASC+ clusters while the right bar graph shows the distribution for the PASC− clusters. **(F)** The epitope distributions of (E) are shown as a strip plot for the PASC+ and PASC− cohorts (colors), with the horizontal black lines indicating the medians of the distributions.

study, sample times were taken with respect to symptom onset, which for the original Wuhan variant of SARS-CoV-2, occurs approximately 6.5 days after infection onset [333, 183]. The median times at which the first, second, and third repertoire samples were collected were 12 days, 19 days, and 63 days since symptom onset (Fig. 3.1B-D), translating to roughly 18.5, 25.5, and 69.5 days after the onset of infection. Hence, the initial repertoire samples in this study may not contain the hallmark first peak of expansion in T cells. Nevertheless, long-term dynamics of clones may shed light on the features of TCR repertoires which are associated with the onset of PASC. We examined expansion and contraction between repertoires sampled before 20 days after symptom onset and repertoires sampled after 60 days since symptom onset.

We first explored the different modes of dynamics among those individuals who were observed within the aforementioned windows of time. Following the procedure by refs. [197, 196], we grouped functional clones by applying hierarchical clustering to the principal components computed from the normalized clonal frequency trajectories of the top rank-1000 clones (Appendix B). While ref. [196] observed three dynamical modes largely consistent among the two individuals in their study, we report a wide range of dynamical modes, which we hypothesize is due to the heterogeneity of sampling and the diversity of responses among our much larger cohort (Figs. 3.7A,B; B.5; B.6). For instance, we do not consistently observe in each individual a subpopulation of the top rank-1000 clones that is large and roughly stable over time. Moreover, while there are some trends within each cohort, such as a dynamical mode of a large expansion and then a large contraction, there are several exceptions, and the consistency of these behaviors is clouded by the heterogeneity of the sampling times.

3.5.1 NoisET: statistical model for clonal dynamics

We employed NoisET, a Bayesian method for detecting differential expansion and contraction, to identify significantly dynamic clones from the entire functional repertoires of the aforementioned sets of individuals [153]. NoisET first characterizes two null models—one for the earlier time point and one for the later—using biological replicates of repertoires. It

detects significantly contracted or expanded clones as those whose fold changes in abundance fluctuated beyond that expected by the null models. The frequencies of TCRβ clones are distributed according to a power law [62, 85],

$$\rho(f) = Cf^{\alpha}, \tag{3.12}$$

with α the parameter of the distribution and C the normalization constant. NoisET models the observed abundance using a Poisson distribution, negative binomial distribution, or a Poisson-negative binomial mixture in which the number of T cells is treated as a latent variable. Because our dataset did not feature biological replicates, we characterized TCRβ abundances n using a Poisson distribution,

$$P(n|f; N_r) = \frac{(fN_r)^n}{n!}e^{-fN_r}, \tag{3.13}$$

where f is the frequency of the TCRβ sequences (relative to an individual's entire repertoire, i.e., observed and unobserved), and N_r is the total abundance in the sample. Of the three models offered by NoisET, the Poisson distribution is the most parsimonious for modeling noise, and it affords us the ability to create *in silico* replicates with few assumptions by using Poisson bootstrapping in lieu of biological replicates.

For a given TCRβ clone, the probability of observing an abundance of n in one replicate and n' in another replicate given α, the exponent of the clone frequency distribution (Eq. 3.12), and $f_{\min}$, the minimum frequency of a TCRβ clone relative to an individual's entire repertoire, is

$$P(n, n'|\alpha, f_{\min}) = \int_{f_{\min}}^{1} df\, \rho(f|\alpha)P(n|f)P(n'|f). \tag{3.14}$$

$\rho(f|\alpha)$ is the clone frequency distribution given in Eq. 3.12, and f, the frequency of a clone relative to an individual's entire repertoire, is unobserved and therefore marginalized out. α and $f_{\min}$ are inferred using maximum likelihood estimation with the likelihood given by

$$\mathcal{L}(\alpha, f_{\min}) = \prod_{i=1}^{N_{\text{obs}}} P(n_i, n_i'|n_i + n_i' > 0, \alpha, f_{\min}), \tag{3.15}$$

where N_{obs} is the number of TCRβ clones observed and i indexes the set of clones. The observation of clone i's abundances in the two replicates n, n' is treated as independent from the observation of any other clone's abundances in the two replicates. Therefore, the likelihood is factorized as the product of the individual likelihoods for each clone's n_i, n_i'. Crucially, a clone must have been observed at least once in either replicate for it to enter into our dataset, so the probability of n_i, n_i' must be conditioned on the clone being seen at least once $P(n_i, n_i'|n_i + n_i' > 0, \alpha, f_{\min})$. This conditional probability can be rewritten using the law of total probability as

$$P(n_i, n_i'|n_i + n_i' > 0, \alpha, f_{\min}) = \frac{P(n_i, n_i', n_i + n_i' > 0|\alpha, f_{\min}))}{P(n_i + n_i' > 0|\alpha, f_{\min})}. \tag{3.16}$$

Instead of computing $P(n_i + n_i' > 0)$ by summing over the probabilities of every possible combination of n_i, n_i' subject to $n_i + n_i' > 0$, we calculate the probability of never observing clone i and subtract it from one.

$$P(n_i + n_i' > 0|\alpha, f_{\min}) = 1 - P(0, 0|\alpha, f_{\min}) \tag{3.17}$$

Substituting Eq. 3.17 into Eq. 3.16 and then Eq. 3.16 into Eq. 3.15, the likelihood now appears as

$$\mathcal{L}(\alpha, f_{\min}) = \prod_{i=1}^{N_{\text{obs}}} \frac{P(n_i, n_i', n_i + n_i' > 0|\alpha, f_{\min})}{1 - P(0, 0|\alpha, f_{\min})}. \tag{3.18}$$

While the abundances of the clones are treated as independent, the frequency f_i of clone i is not independent from all the other frequencies. The optimization must be constrained such that the frequencies sum to 1. Letting N be the (estimated) total amount of clones in an individual's repertoire, the inference should yield $N\langle f\rangle = 1$ [240]. NoiseET requires

$$N_{\text{obs}} \frac{P(0, 0|\alpha, f_{\min})}{1 - P(0, 0|\alpha, f_{\min})} \langle f\rangle_{\rho(f|n+n'=0)} + \sum_{i=1}^{N_{\text{obs}}} \langle f\rangle_{\rho(f|n_i, n_i')} = 1. \tag{3.19}$$

N_{obs} multiplied by the odds ratio gives the estimated number of clones that were not observed, so the first term calculates the frequency of the entire repertoire that was unobserved. The second term is the sum of the average posterior frequencies over the observed clones, i.e., the

model's estimates for the clones' frequencies with respect to an individual's entire repertoire. An α and $f_{\min}$ are obtained for each replicate, $\theta_1 = (\alpha(t_1), f_{\min}(t_1))$ for the earlier time point and $\theta_2 = (\alpha(t_2), f_{\min}(t_2))$ for the later time point.

Next, NoisET learns a posterior distribution of the log-fold change η of a clone's frequency over time and assumes a prior that is a mixture of a Laplace distribution and a delta function

$$P(\eta|\gamma, \bar{\eta}) = \gamma \frac{1}{2\bar{\eta}} e^{-|\eta|/\bar{\eta}} + (1-\gamma)\delta(\eta), \tag{3.20}$$

where $\gamma \in [0, 1]$ is the mixture parameter and represents the fraction of clones that are responding. The abundance of a responding clone changes over time more widely than can be captured by the noise model alone. η is the log-fold change of this deviation from the noise model and is characterized by a Laplace distribution. The Laplace distribution is parameterized by $\bar{\eta}$, which is proportional to the distribution's standard deviation and represents the strength of the bulk of responses. The Laplace distribution is similar to the Gaussian distribution ($\propto e^{-\eta^2}$) but features longer tails since $e^{-|\eta|}$ decays slower than $e^{-\eta^2}$. Therefore, a larger $|\eta|$ has a higher probability in a Laplace distribution compared to a Gaussian distribution. Consequently, the Laplace distribution gives some more flexibility for the posterior distribution to characterize $\eta \neq 0$ and was found to be more consistent with dynamical repertoire data [240]. The delta function captures $\eta = 0$ behavior, meaning a clone's variation across time was completely consistent with the noise model. If the delta function were not present, the Laplace distribution alone would have to capture the probability, here $(1-\gamma)$, of a clone being unresponsive. Since a nontrival fraction of the repertoire is expected to be unresponsive [240], a small $\bar{\eta}$ would be required to inflate the probability of $\eta = 0$; however, $\bar{\eta}$ being too small would preclude modeling large log-fold changes that are observed in the data since they would be given low probabilities. The prior constructed as this mixture gives the necessary flexibility for modeling responsive and unresponsive clones.

The probability of observing a clone with abundance $n(t_1)$ and $n(t_2)$ at times t_1 and t_2

respectively and null model parameters θ_1 and θ_2 is

$$P(n(t_1), n(t_2)|\gamma, \bar{\eta}, \theta_1, \theta_2) = \int\int df\, \rho(f|\theta_1)\, d\eta\, P(\eta|\gamma, \bar{\eta})\, P(n(t_1)|f, \theta_1) P(n(t_2)|fe^{\eta}, \theta_2). \quad (3.21)$$

The parameters of the prior are learned by maximizing the likelihood of the pairs of observed abundances over time

$$(\gamma^*, \bar{\eta}^*) = \underset{(\gamma,\bar{\eta})}{\text{argmax}} \prod_{i=1}^{N_{\text{obs}}} \frac{P(n_i(t_1), n_i(t_2), n_i(t_1) + n_i(t_2) > 0|\gamma, \bar{\eta}, \theta(t_1), \theta(t_2))}{1 - P(0, 0|\gamma, \bar{\eta}, \theta(t_1), \theta(t_2))}. \quad (3.22)$$

We note that the likelihood here takes a similar form as the likelihood used to learn the parameters of the noise model by similar arguments (Eq. 3.18). The denominator ensues from the fact that each clone in our data was observed at least once longitudinally:

$$P(n_i(t_1), n_i(t_2)|n_i(t_1) + n_i(t_2) > 0) = \frac{P(n_i(t_1), n_i(t_2), n_i(t_1) + n_i(t_2) > 0)}{P(n_i(t_1) + n_i(t_2) > 0)}, \quad (3.23)$$

where we have supressed the model parameters for the moment. We identify the numerator and denominator of Eq. 3.23 with that in Eq. 3.22 since

$$P(n_i(t_1) + n_i(t_2) > 0|\gamma, \bar{\eta}, \theta(t_1), \theta(t_2)) = 1 - P(0, 0|\gamma, \bar{\eta}, \theta(t_1), \theta(t_2)). \quad (3.24)$$

Once γ and $\bar{\eta}$ are inferred, posteriors for the log-fold change of clones can be computed readily

$$P(\eta|n(t_1), n(t_2), \gamma, \bar{\eta}, \theta(t_1), \theta(t_2)) = \frac{P(n(t_1), n(t_2)|\gamma, \bar{\eta}, \theta(t_1), \theta(t_2)) P(s|\gamma, \bar{\eta})}{P(n(t_1), n(t_2))}. \quad (3.25)$$

NoisET tests for expansion and contraction by calculating a one-sided p-value using the posterior distribution

$$p = P(\eta \leq 0|n(t_1), n(t_2), \gamma, \bar{\eta}, \theta(t_1), \theta(t_2)). \quad (3.26)$$

Consistent with this definition, contraction is tested by switching the counts and null model parameters by sending $t_1 \to t_2$ and $t_2 \to t_1$.

3.5.2 Differential statistics of dynamical clones between PASC+ and PASC− individuals

We studied clonal dynamics at the level of TCRs with the same TRβV gene and TCRβ CDR3 amino acid sequence to hone in on detecting population changes more likely to be associated with the role these TCRs played in repertoire reponses. In the absence of replicates in this study, we opted to generate *in silico* replicates via Poisson bootstrapping. Specifically, given a repertoire sampled at a single time point, we produced two replicates by Poisson resampling using the observed TCR's abundance as the Poisson rate. For each sampled repertoire, we generated 100 sets of *in silico* paired replicates, learned 100 NoisET Poisson noise models, and took the mean of the resulting parameters for that sample as input for NoisET to infer expansion and contraction across time. TCRs were considered significantly dynamic if their BH-FDR adjusted $p < 0.01$, with BH-FDR corrections applied separately for each individual.

We identified 2,194 clones that significantly expanded in the PASC+ cohort, 1,015 clones that expanded in the PASC− cohort, 1,355 clones that contracted in the PASC+ cohort, and 1,045 clones that contracted in the PASC− cohort (Figs. 3.7C-D, 3.6A,B). We quantified how much of these repertoires might be enriched for specificity to SARS-CoV-2 by comparing them to a database containing clonotypes known to be specific to SARS-CoV-2 peptides from a Multiplex Identification of Antigen-Specific (MIRA) assay [216] (Appendix B). We opted to use the MIRA database since it is comprised of nearly 138,000 TCRβ sequences which tested positive for SARS-CoV-2 reactivity whereas the VDJdb database is more sparse, containing roughly 5,400 SARS-CoV-2-specific TCRβ sequences. We considered a sequence as specific to SARS-CoV-2 and matched to the MIRA database if it was within 12 TCRdist (1 amino acid difference in TCRβ CDR3 sequence, or 3 amino acid differences in the TRβV genes) of at least two TCRβ sequences in the MIRA database. We imposed this more stringent requirement on matching, in comparison to our criteria for matching into VDJdb that requires a TCR to be within 24 TCRdist of only one VDJdb sequence, since the MIRA database was produced by surveying TCRs for their specificity toward exogenously loaded, untested SARS-CoV-2 peptides. Thus, it's possible that the MIRA database contains a

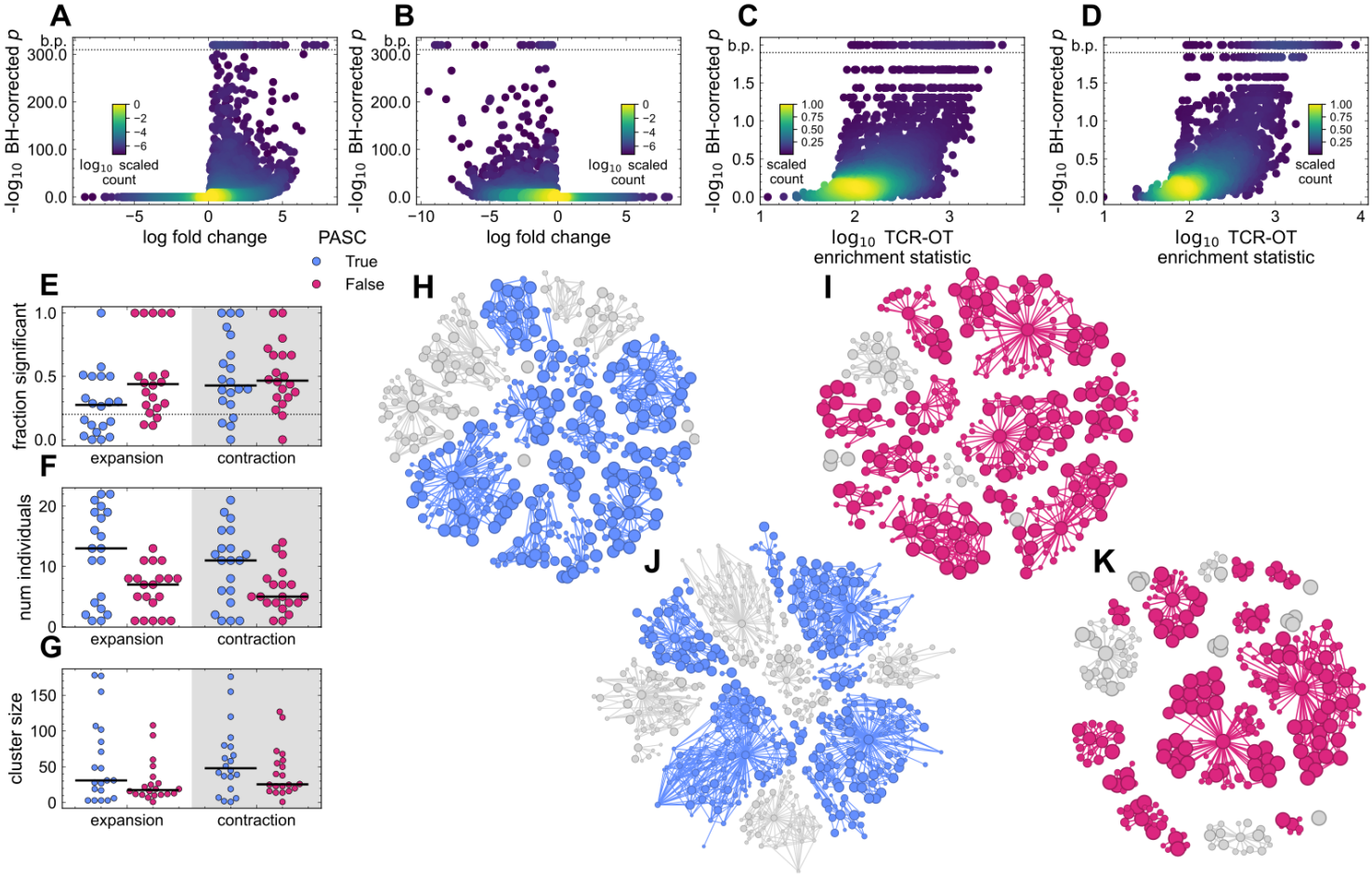

A
B
C
D
b.p.
-log10 BH-corrected p
log fold change
log10 scaled count
scaled count
log10 TCR-OT enrichment statistic
PASC
True
False
E
F
G
fraction significant
num individuals
cluster size
expansion
contraction
H
I
J
K

Figure 3.6: **NoisET volcano plots and TCR-OT summary statistics and clusters.** **(A)** The volcano plot shows the $-\log_{10}$ BH-corrected p-value vs. the NoisET-inferred mean log-fold change for each functional clone tested for expansion. p-values annotated as b.p. are beyond machine precision. **(B)** As in (A) but for each functional clone tested for contraction. **(C)** The volcano plot shows the $-\log_{10}$ BH-corrected p-value vs. the $\log_{10}$ TCR-OT enrichment statistic for each functional clone in the expanded PASC+ and PASC− repertoires. p-values annotated as b.p. are beyond machine precision. **(D)** As in (C) but for each functional clone in the contracted PASC+ and PASC− repertoires. **(E)** The swarm plot shows the distribution of the fraction of functional clones which were deemed significant within a TCR-OT cluster stratified by dynamic (x-axis) and PASC status (colors). The black horizontal lines show the median for each distribution. Markers above the horizontal, dotted line indicate clusters which have more than 20% of their constituent clones called as significant. **(F)** The swarm plot shows the distribution of individuals in each cluster stratified by dynamic (x-axis) and PASC status (colors). **(G)** The swarm plot shows how many independent recombinations were in each TCR-OT cluster stratified by dynamic (x-axis) and PASC status (colors). The black horizontal lines show the median for each distribution. **(H - K)** The network plots show the topology of the TCR-OT clusters for the PASC+ contracted subset (H), PASC− contracted subset (I), PASC+ expanded subset (J), and PASC− expanded subset. Sizes of the nodes show the strength of the $-\log_{10}$ FDR-BH corrected p values associated with a functional clone, with larger sizes indicating more significantly atypical functional clones. Edges between nodes and the cluster's focal sequence indicate sequence similarity within TCR-OT inferred cluster radius, and all other edges indicate sequence similarity within 48 TCRdist. Network topologies with colored nodes had at least 20% of their nodes deemed significant and had recombinations from at least two individuals.

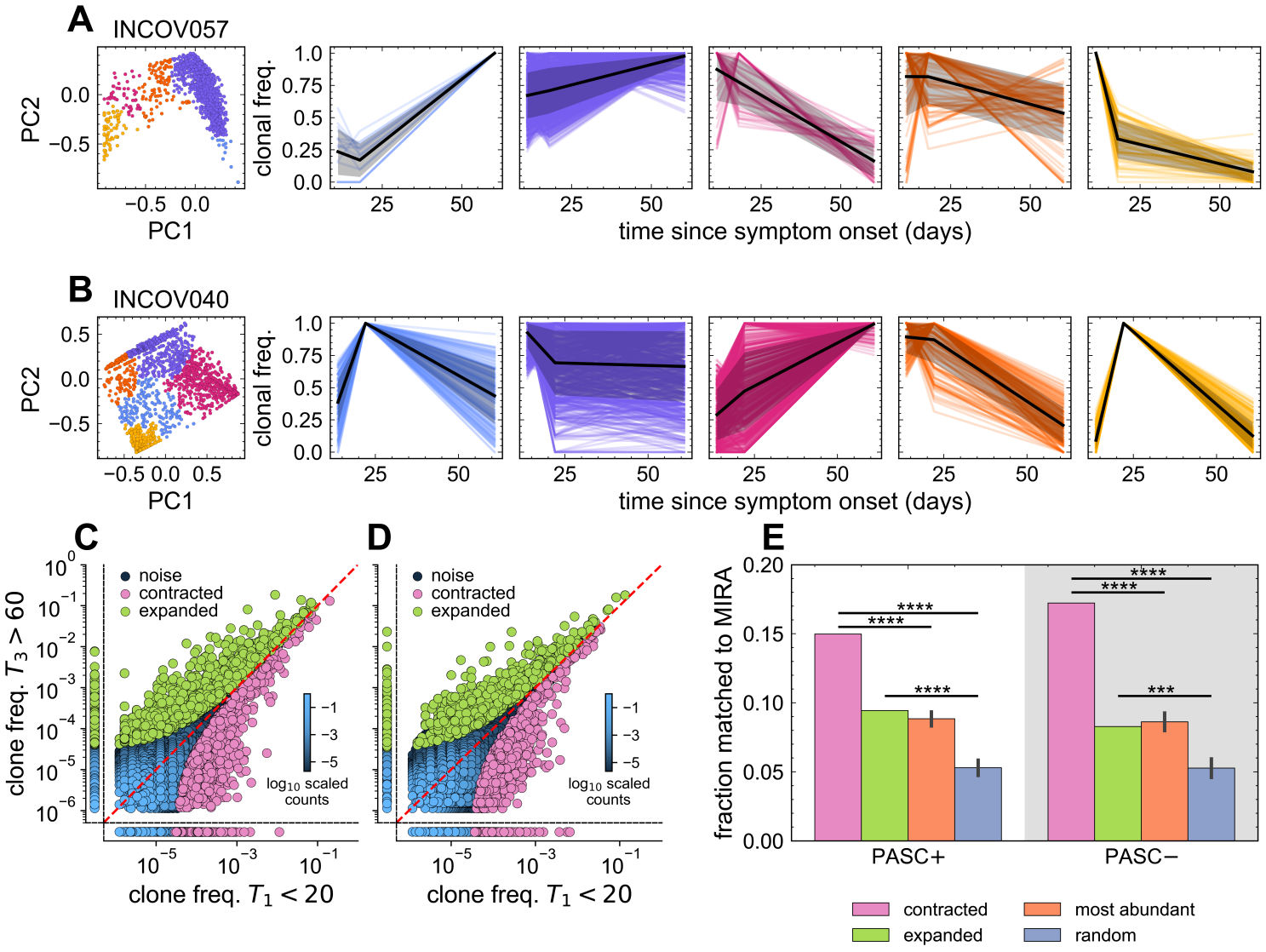

A
INCOV057
PC2
PC1
clonal freq.
time since symptom onset (days)
B
INCOV040
PC2
PC1
clonal freq.
time since symptom onset (days)
C
noise
contracted
expanded
clone freq. $T_3 > 60$
clone freq. $T_1 < 20$
$\log_{10}$ scaled counts
D
noise
contracted
expanded
clone freq. $T_1 < 20$
$\log_{10}$ scaled counts
E
fraction matched to MIRA
PASC+
PASC−
contracted
most abundant
expanded
random

Figure 3.7: **Dynamics of TCR repertoires. (A)** Dynamical modes of the TCR repertoire of individual INCOV057. (Left) PCA of normalized top rank-1000 clonal frequencies. Points are colored by the output of hierarchical clustering. (Middle) The normalized clonal trajectories of clones within a cluster (colors as in Left) are plotted as lines. The black line shows the average clonal trajectory, and shading indicates regions containing one standard deviation of variation across trajectories in that mode. **(B)** As in (A) but for the top rank-1000 in individual INCOV040. **(C, D)** The normalized clonal frequencies observed in the earlier sampled repertoire (x-axis) and later sampled repertoire (y-axis) are shown as a scatter plot for clones pooled in the PASC+ cohort (C) and clones pooled in the PASC− cohort(D). Regions bounded by black, dashed lines show clones present in only the early sample (bottom) or later sample (left). Lighter colors indicate higher densities of points. Clones detected as significantly contracted are colored pink, and clones detected as significantly expanded are colored green. **(E)** Bars show the average fraction of clones which had a match to MIRA (i.e., was a neighbor to at least two MIRA sequences within 12 TCRdist) for the dynamically significant clones (pink, green), 250 subsamples of top rank-200 clones (orange) from the respective dataset (x-axis), and 250 subsamples of randomly chosen clones (purple) from the respective dataset (x-axis). ****: p-value $< 10^{-4}$; ***: p-value $< 10^{-3}$.

fair amount of false positives. [216]. We found 9.4% of the PASC+ expanded clones, 8.3% of the PASC− expanded clones, 15.0% of the PASC+ contracted clones, and 17.2% of the PASC− contracted clones to overlap with the MIRA assay (Fig. 3.7E). The fractions of MIRA overlaps between the PASC+ and PASC− expanded repertoires and the fractions of MIRA overlaps between the PASC+ and PASC− contracted repertoires were not significantly different (one-tailed Fisher exact test with Bonferroni corrections, $p > 0.05$). As a control, we subsampled the top rank-200 clones from the repertoires of individuals tested for expansion and contraction to the sizes of the dynamical subsets and computed the fraction of matches to

the MIRA database. We observed $8.85\% \pm 0.3\%$ and $8.81\% \pm 0.5\%$ of the top rank-200 clones from PASC+ repertoires subsampled to the sizes of the PASC+ expanded and contracted subsets, respectively, to match to the MIRA database; similarly, we found $8.65\% \pm 0.6\%$ and $8.56\% \pm 0.6\%$ of the top rank-200 clones from PASC− repertoires subsampled to the sizes of the PASC− expanded and contracted subsets, respectively, to match to the MIRA database. As another control, we subsampled from the pooled repertoires of the PASC+ cohort tested for dynamics and found $5.3\% \pm 0.4\%$ and $5.3\% \pm 0.6\%$ matched to the MIRA database for the sizes of the PASC+ expanded and contracted subsets, respectively. Additionally, we subsampled from the pooled repertoires of the PASC− cohort tested for dynamics and observed $5.3\% \pm 0.7\%$ and $5.2\% \pm 0.6\%$ matched to the MIRA database for the sizes of the PASC− expanded and contracted subsets, respectively. The contracted repertoires were significantly enriched for specificity to SARS-CoV-2 compared to the most abundant TCRs and random TCRs whereas the expanded repertoires were significantly enriched only when compared to the random subset of TCRs (Z-test statistics and Bonferroni-corrected p-values shown in Table B.2) (Fig. 3.7E).

To characterize the sequences which were substantially atypical between the subsets of PASC+ and PASC− expanded and contracted repertoires, respectively, we applied the TCR-OT algorithm. TCR-OT detected 255 significantly enriched TCRs in the PASC+ expanded subset, 137 in the PASC− expanded subset, 210 in the PASC+ contracted subset, and 205 in the PASC− contracted subset. Moreover, we identified 13 clusters in the contracted PASC+ repertoire, 16 clusters in the contracted PASC− repertoire, 10 clusters in the expanded PASC+ repertoire, and 12 clusters in the expanded PASC− repertoire. As before, we required that each cluster had at least 20% of their constiuents be significantly enriched to ensure neighborhoods rather than singletons were amplified, and we restricted the clusters to contain TCRs which were present in at least two individuals (Figs. 3.6; B.7; B.8; 3.8E,F). We report the sequence motifs of neighbors within 48 TCRdist of the focal sequences in these clusters in Fig. 3.8A-D and the sequence motifs of the entire clusters in Fig. B.9. While motifs seemed to be slightly enriched for polar, uncharged amino acids in the PASC+ contracted

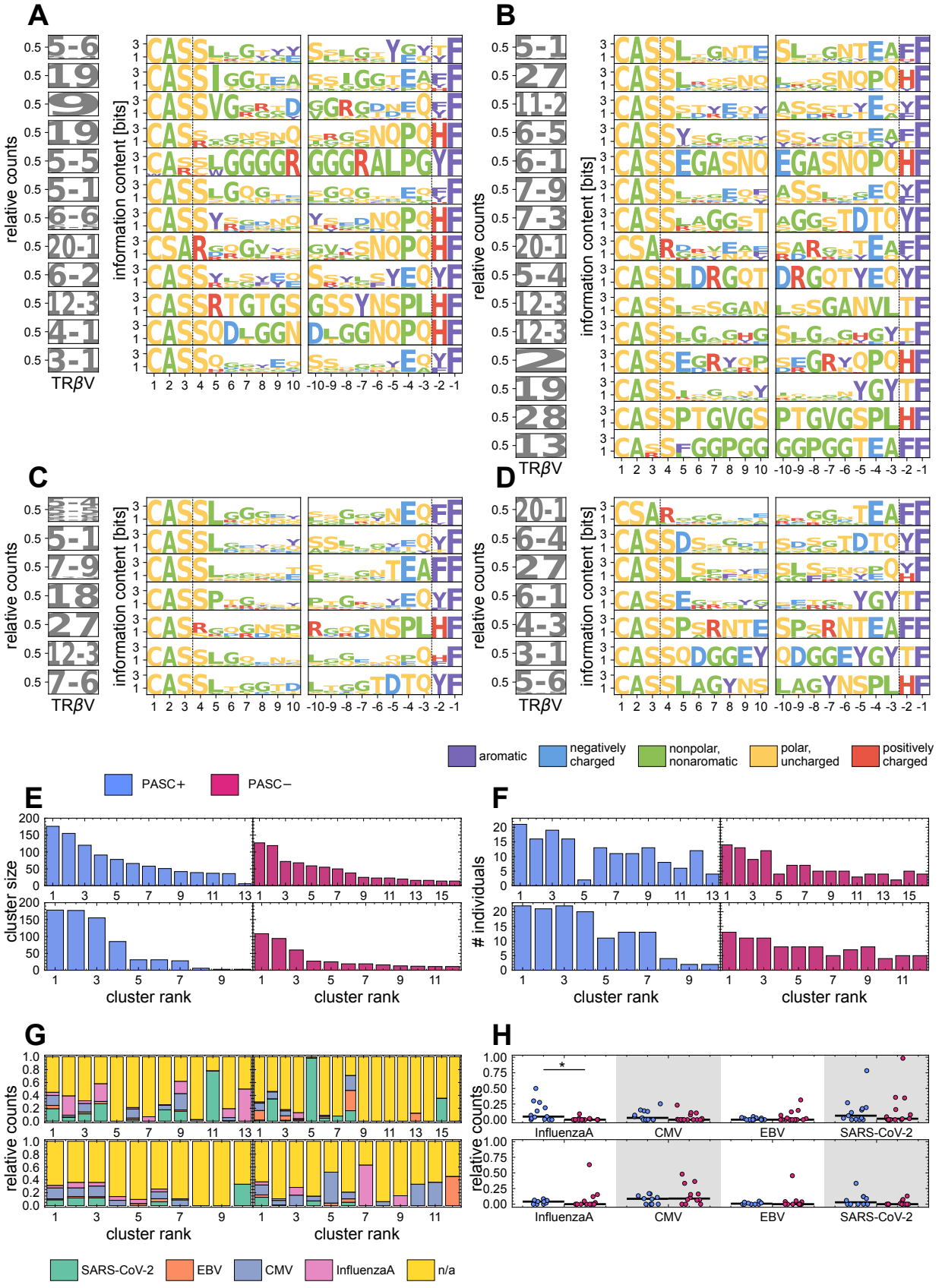

A
B
C
D
relative counts
information content [bits]
TRβV
aromatic
negatively charged
nonpolar, nonaromatic
polar, uncharged
positively charged
PASC+
PASC−
E
F
cluster size
individuals
cluster rank
G
H
relative counts
InfluenzaA
CMV
EBV
SARS-CoV-2
n/a

Figure 3.8: **Motifs and VDJdb matching among dynamical TCR-OT clusters. (A, B, C, D)** Sequence logos describing the composition of amino acid sequences are shown for the significant TCR-OT clusters obtained from the contracted PASC+ subset (A), contracted PASC− subset (B), expanded PASC+ subset (C), and expanded PASC− subset (D). Logos were constructed from focal sequence(s) and members in the cluster that were within 48 TCRdist of the focal sequence(s). Logos are shown only if 10 sequences were present in this 48 TCRdist neighborhood. Logos are ordered top to bottom by cluster size, with the largest cluster size at the top. (Left) TRβV gene usage frequencies within a cluster are shown as a logo. (Middle, right) The logo plots shown the information content at each position in the amino acid TCRβ CDR3 sequence from left to right (middle) and right to left (right). Amino acids are colored according to their physicochemical properties indicated in the legend. **(E)** The amounts of independent recombinations in the clusters from the contracted (top) and expanded (bottom) repertoires of PASC+ and PASC− individuals (colors) are shown as bar graphs. Cluster rank corresponds to the size of the clustering. **(F)** The number of individuals represented in a cluster for the contracted (top) and expanded (bottom) repertoires of PASC+ and PASC− individuals (colors) are shown as bar graphs. **(G)** The stacked bar graphs show the distribution of epitopes of TCRβ sequences in VDJdb that were within 24 TCRdist of a constituent TCRβ in a cluster and whose donor HLA group matched to at least one HLA group of the individual from which the similar TCRβ sequence originated. The left bar graphs shows the distribution for the PASC+ clusters while the right bar graphs shows that for the PASC− clusters. The top and bottom bar graphs correspond to the contracted and expanded repertoires, respectively. **(H)** The epitope distributions (G) are shown as a strip plot for the contracted (top) and expanded (bottom) repertoires of PASC+ and PASC− (colors). The horizontal black lines show the medians. *: Bonferroni-corrected p-value 0.05.

repertoire and for polar, uncharged and negative amino acids in the PASC− repertoire, the differences were not significant (Mann-Whitney U test with Bonferroni corrections, $p > 0.05$) (Figs. 3.8A-D; B.9; B.10A,C).

We matched sequences in these clusters to sequences with known specificity given by VDJdb [275, 98] and found that significant clusters created from the contracted repertoire of the PASC+ cohort were enriched for specificity to InfluenzaA epitopes (Mann-Whitney $U = 170$, $n_+ = 13$, $n_- = 16$, Bonferroni-corrected $p = 0.016$) (Fig. 3.8G,H). Given the individuals in our cohorts were examined in early 2020, we hypothesize that some PASC+ individuals may have had encounters with flu around the time they contracted SARS-CoV-2 since there is little evidence supporting bystander or viremia activation associated with influenza. Thus, while this cohort provides a unique window into understanding the development of PASC early in the pandemic, some repertoire differences between the PASC+ and PASC− cohorts might be due to influenza infections.

3.6 Functional responses in the public repertoire

Though an individual's TCR repertoire is generated stochastically and highly private [73], identical TCRs can be sampled from individuals. They may signal similar pathogenic exposures and thus the publicness of a TCR can be used to detect if it was mounted functionally against an immune challenge. Public TCRs have been identified as responsive to or predictive of receiving a yellow fever vaccine [235] or having CMV [77, 233], diabetes [233], or SARS-CoV-2 [254, 236]. However, TCRs may be public due to biases in V(D)J recombination and commonality in HLA (i.e., similar thymic selection) [308, 75]. Deciphering which TCRs are highly incident in a sample due to shared antigen history requires careful statistical analysis.

We began by testing the enrichment of shared TCRβ clones using the method from ref. [77], which was also applied to SARS-CoV-2 cohorts in ref. [236]. Ref. [77] performed one-sided Fisher exact tests on the sharing of clones among CMV+ and CMV− individuals. They defined a TCRβ clone by its TCRβ CDR3 amino acid sequence, TRβV gene, and TRβJ gene and examined clones that were shared in at least two individuals only. They

applied multiple testing corrections by quantifying their false discovery rate for a given p-value by comparing the number of rejected null hypotheses in the observed dataset to datasets with permuted CMV statuses. Specifically, given m permutations of the disease status of individuals and n_0 rejected null hypotheses in the observed dataset, the false discovery rate (FDR) can be approximated as

$$\text{FDR} = \frac{1}{m} \frac{\sum_{i=1}^{m} n_{\text{permuted},i}}{n_0}. \tag{3.27}$$

FDR < 1, and FDR $\ll 1$ for a strong signal, results from the number of clones determined as enriched in the observed dataset being larger on average than the number of clones called enriched in the shuffled datasets. This would indicate that the cohort labels and sharing numbers of clones are informative for discriminating overshared clones, as in ref. [77]. We performed one-sided Fisher exact tests with the same procedure to identify TCRβs overshared in our PASC+ cohort and found $22,381$ clones that had $p < 0.05$. To quantify our false discovery rate, we permuted the labels specifying which cohort a clone was observed in, and then we computed the significance of the enrichment of clones in the PASC+ using these shuffled labels. We repeated the randomization of labels and significance computation 100 times to obtain a distribution of n_{permuted} and an estimate of the false discovery rate. We observed that the amount of clones that was called enriched in the observed dataset, $22,381$, was not statistically different from the number of clones called enriched in the shuffled datasets (randomization test, $p = 0.63$), resulting in FDR $= 1.27$ (Fig. B.11). Therefore, counting statistics alone cannot discriminate overshared TCRβ receptor sequences in PASC+ individuals.

We continued to study sharing by employing models that learn how the statistics of repertoires have been shaped by V(D)J recombination and selection and ultimately allow us to estimate the probability of observing sequences in repertoires [74, 184, 269, 133]. Sharing of sequences can be due to biases in V(D)J recombination, similar HLAs and convergent thymic selection, convergent pathogenic selection, or experimental biases, and using these models allows us to quantify our expectations of a sequence shared due to V(D)J recombination and

the average effects of selection [75, 233, 200, 254]. We first learned a model characterizing the V(D)J recombination process of each individual by pooling their unproductive TCRβ sequences and using IGoR (version 1.4) [184] to learn the relationships between TRβV, TRβD, and TRβJ choices as well as the insertion and deletion profiles at the VD and DJ junctions. IGoR yields the probability of generating a receptor $P_{\text{gen}}(\sigma)$. Next, we used soNNia (version 0.3.2) [133] to model the effects of selection by studying the statistics of productive TCRβs observed in data compared to the productive TCRβ repertoire generated from V(D)J recombination alone. soNNia infers selection factors $Q^{\theta}(\sigma)$ to estimate the fold change difference between the probability $P_{\text{post}}(\sigma)$ to observe a productive sequence in the periphery and the probability $P_{\text{gen}}(\sigma)$ of generating the sequence,

$$P_{\text{post}}(\sigma) = \frac{1}{Z} P_{\text{gen}}(\sigma) Q^{\theta}(\sigma). \tag{3.28}$$

Z is a normalization factor, and $Q^{\theta}(\sigma)$ is a neural network which takes as input TRβV gene and TRβJ gene usages, TCRβ CDR3 sequence length, and TCRβ CDR3 amino acid composition [133]. We trained a soNNia model for each individual, pooling their repertoires sampled across time, to characterize selection due to individual HLA restrictions as well as different immune responses mounted over the course of their infections (Appendix B).

We studied the sharing of TCRβs among individuals by aggregating repertoires over time for each individual and deduplicating the data at the level of TRβV gene, TRβJ gene, TCRβ CDR3 amino acid sequence, and individual. Let σ be a TCRβ clone defined here as as the TRβV gene, TRβJ gene, and TCRβ CDR3 amino acid sequence. The observation of σ in each individual is a Bernoulli trial. Because each σ has a different probability of being observed in individual, the number of times a sequence is shared among individuals is modeled as a Poisson binomial process. We approximate the Poisson binomial distribution characterizing these phenomena using a binomial distribution parameterized by the maximum amount of possible incidences and the average probability of observing a sequence [48]. This approximation will yield conservative estimates when quantifying outliers because the binomial distribution is more dispersed than the Poisson binomial distribution. The probability of

observing σ at least once in individual i's repertoire is

$$P_{\text{obs}}(\sigma; i) = 1 - (1 - P_{\text{post}}(\sigma; i))^{N_{\text{nt}}(i)}, \tag{3.29}$$

where $P_{\text{post}}(\sigma; i)$ is the probability of observing σ in the periphery estimated using the soNNia model associated with individual i, and $N_{\text{nt}}(i)$ is the number of unique TCRβ nucleotide sequences, i.e., sequences defined by the TRβV gene, TRβJ gene, and TCRβ CDR3 nucleotide sequence, in individual i's pooled repertoire. We next computed the probability of sharing σ in m out of M individuals:

$$P_{\text{share}}(m; M, \sigma) = \binom{M}{m} \rho(\sigma)^m (1 - \rho(\sigma))^{M-m}. \tag{3.30}$$

The probability $\rho(\sigma)$ is the average probability of observing σ in each individual at least once,

$$\rho(\sigma) = \frac{1}{M} \sum_{i=1}^{M} P_{\text{obs}}(\sigma; i). \tag{3.31}$$

To identify outliers in the P_{share} distribution, we fix a threshold c such that

$$c = \binom{M}{m} q(m)^m (1 - q(m))^{M-m}, \tag{3.32}$$

with $q(m) \in [0,1]$, a probability tuned for each m.[1] This gives a parametric curve of probabilities $q(m)$ versus sharing number $2 \leq m \leq M$, and sequences below this curve are said to be outliers and rare. We chose c such that 0.5% of the repertoire lie under the parametric curve $q(m)$ (see Appendix B for a comparison to binomal p-values).

We observed large scale sharing of TCRs within the PASC+ and PASC− cohorts (Fig. 3.9A,B). From 3,969,443 shared TCRs in the PASC+ cohort, we detected 19,847 that were rare; from 2,683,467 TCRs shared among PASC− individuals, we identified 13,417 that were rare (Fig. 3.9C,D). We examined the datasets for specificity to SARS-CoV-2 by matching to sequences in the MIRA database [216] (Appendix B). We observed that 23.7% of the rare TCRs from the PASC+ and 24.3% of the rare TCRs from the PASC− subsets were specific for SARS-CoV-2. The subset of rare TCRs from the PASC− cohort was not significantly more enriched for SARS-CoV-2 specificity compared to the PASC+ cohort's rare subset (one-sided Fisher exact test, Bonferroni-corrected $p = 0.20$). Notably, 5,762 rare, shared TCRs were identical between the two cohorts, translating to 29% of the PASC+ subset and

[1] Fixing M, the binomial probability $b(q,m) = \binom{M}{m} q^m (1-q)^{M-m}$ will have two q which satisfy $b(q,m) = c$ when $0 < m < M$ and $c \in (0, \max_q b(q,m))$. To illustrate this, we inspect the derivatives of $\log b(q,m)$ with respect to q since log preserves the functional behavior of $b(q,m)$ but is easier to manipulate algebraically. $d_q \log b(q,m) = m/q - (M-m)/(1-q)$, indicating the function is increasing when $m/q > (M-m)/(1-q)$ and decreasing when $m/q < (M-m)/(1-q)$. $b(q,m)$ is only decreasing when $m = 0$ since the first derivative is negative, and $b(q,m)$ is only increasing when $m = M$. Thus, $b(q,m) = c$ at only one q for these conditions as claimed. The second derivative is $d_q^2 \log b(q,m) = -m/q^2 - (M-m)/(1-q)^2 < 0$, so $\log b(q,m)$ and $b(q,m)$ are concave-down. Because $b(q,m)$ is concave down for $0 < m < M$, $b(q,m) = c$ at two distinct values of q when c is chosen such that $c \in (0, \max_q b(q,m))$. We require the parametric curve which is used to call a shared sequence an outlier $q(m)$ to be monotonically increasing. If, on the other hand, $q(m)$ is not monotically increasing, then $n > m$ could imply $q(n) < q(m)$. In other words, we would be a placing a looser requirement on identifying outlier sequences which are shared in fewer individuals than those shared in more individuals; however, a sequence is, in general, expected to be shared in fewer individuals, and so $m < n$ must imply $q(m) < q(n)$ in order for $q(m)$ to detect outliers in a statistically and biologically meaningful way. To enforce the monotonicity of $q(m)$, we choose the smaller q in constructing $q(m)$ whenever two q satisfy $b(q,m) = c$.

42% of the PASC− subset. Excluding this overlap to characterize the rare, shared TCRs local to each cohort, we found 18.0% of the rare TCRs from the PASC+ subset and 14.4% of the rare TCRs from the PASC− subset were specific for SARS-CoV-2 as detected by comparing them to the MIRA database. In this case, the rare TCRs local to the PASC+ cohort were significantly enriched for SARS-CoV-2 (one-sided Fisher exact test, Bonferroni-corrected $p = 4.18 \times 10^{-12}$). To investigate how SARS-CoV-2 specificity changes with the publicness of a TCR, we investigated the fraction of MIRA matches as a function of the number of individuals in which a TCR was shared. We studied this relationship for TCRs in the rare subsets and, as a control, the pooled shared repertoire for each cohort (Fig. 3.9E,F). Only 8.1% and 9.7% of the entire shared repertoires were specific to SARS-CoV-2 for the PASC+ and PASC− cohorts, respectively, with these percentages being significantly lower than that for the rare subsets (one-sided Fisher exact tests, Bonferroni-corrected $p = 0$). The fraction of MIRA matches in the rare subsets for low sharing numbers exceeds that of the pooled shared datasets, which was expected by performing outlier detection. As the number of individuals in which a TCR is shared is increased, the two functions intersect, and then the rare subsets contain slighltly fewer MIRA matches. We attribute this disagreement in the fraction of MIRA matches to our model understimating P_{obs} by using the binomial approximation and the inherent loss in sensitivity when studying only outliers in the shared dataset (Fig. 3.9E,F). Finally, we compared the fraction of MIRA matches in our rare subsets to MIRA matches from subsets of the bulk repertoires constructed from matching the statistics of $P_{\text{obs}}(\sigma)$ in the rare subsets, as outlined by ref. [254] (Appendix B) (Fig. 3.9G). We performed 200 randomizations and observed $1.55\% \pm 0.09\%$ and $1.81\% \pm 0.11\%$ to match to the MIRA database from the controls. Indeed, the rare, shared subsets exhibit significantly increased specificity to SARS-CoV-2 compared to these controls (Z-test, Bonferroni-corrected $p = 0$).

Having identified TCRs which were shared more than expected within each cohort, we next sought to distinguish which rare, shared TCRs were unique to each cohort and applied the TCR-OT algorithm to these subsets. Applying the TCR-OT algorithm to the datasets

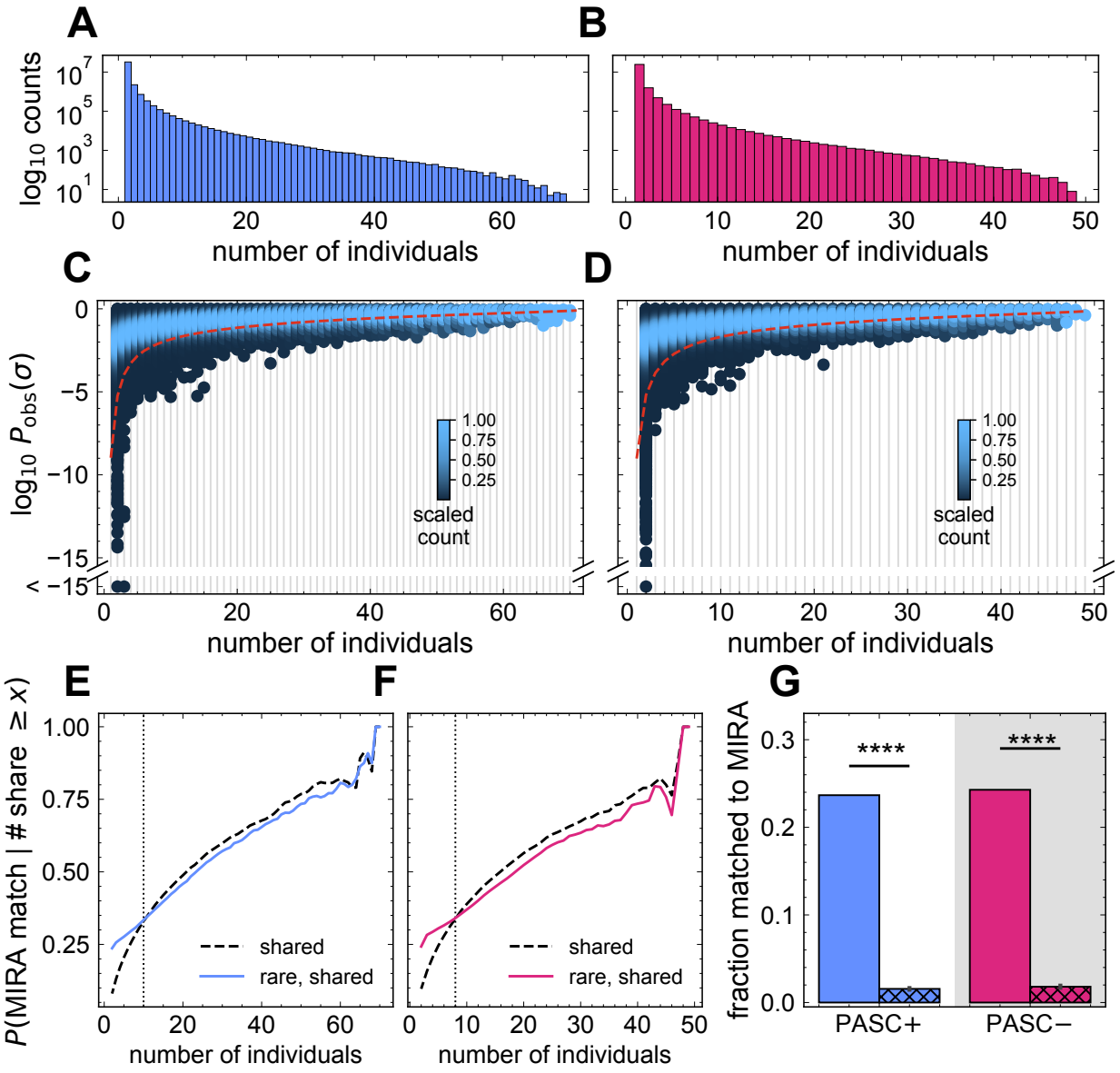

A
B
log10 counts
10^7
10^5
10^3
10^1
0
20
40
60
number of individuals
0
10
20
30
40
50
number of individuals
C
D
log10 Pobs(σ)
0
−5
−10
−15
< −15
1.00
0.75
0.50
0.25
scaled count
number of individuals
E
F
G
P(MIRA match | # share ≥ x)
1.00
0.75
0.50
0.25
shared
rare, shared
number of individuals
fraction matched to MIRA
0.3
0.2
0.1
0.0
****
PASC+
PASC−

Figure 3.9: **Sharing of TCR repertoires. (A, B)** The sharing spectrums of the (A) PASC+ cohort and (B) PASC− cohort are shown as histograms and illustrate the number of TCRs (y-axis) which are incident in a given number of individuals (x-axis). **(C, D)** The density plot shows the distribution of $\log_{10} P_{\text{obs}}$ (y-axis) for TCRs shared in a given number of individuals (x-axis) for the (C) PASC+ cohort and (D) PASC− cohort. The scaling of the counts is set for each number of individuals (column) separately, with the maximum of the density given a value of one. Sharing of rare lineages with $\log_{10} P_{\text{obs}}$ below the parametric curve defined by P_{share} (the red, dashed line) is statistically significant. **(E, F)** The line plot shows how the fraction of MIRA matches changes as a function of the minimum number of individuals among which TCRs are shared for the (E) PASC+ cohort and (F) PASC− cohort. The black, dashed line shows the expectation from the entire shared repertoire whereas the solid, colored line displays the relationship constructed from the subset of rare TCRs. The vertical, black, dotted line indicates where the two curves first approximately intersect. **(G)** The bar plots show the fraction of MIRA matches in the rare TCR subsets (colored bar) and the average fraction of MIRA matches from 200 realizations of subsets drawn randomly from the bulk repertoires consistent with the P_{obs} distribution of the rare subsets (hatched, colored bar). The error bar on the hatched bars indicates one standard deviation of variation. ****: Bonferroni-corrected p-value $< 10^{-4}$.

at the level of independent recombinations would necessarily lead to a majority of sequences deemed significantly enriched in their own cohorts because the sharing analysis detected sequences enriched in public neighborhoods. Therefore, we applied the TCR-OT algorithm at the level of TCRβ CDR3 amino acid sequence, TRβV gene, and TRβJ gene, as in the sharing analysis, merely to detect differing motifs among the rare, shared subsets. We found 51 (0.26%) and 20 (0.15%) sequences in the PASC+ and PASC− rare, shared subsets to be significantly different (TCR-OT randomization test, BH-FDR adjusted $p < 0.05$), and

only 9 PASC+ and 0 PASC− TCR-OT clusters had at least 20% of their constituents called significant (Fig. 3.10A-D). We observed slight specificity toward CMV and InfluenzaA in a couple clusters; however, most sequences in any of the TCR-OT did not match to the VDJdb (Fig. 3.10E,F).

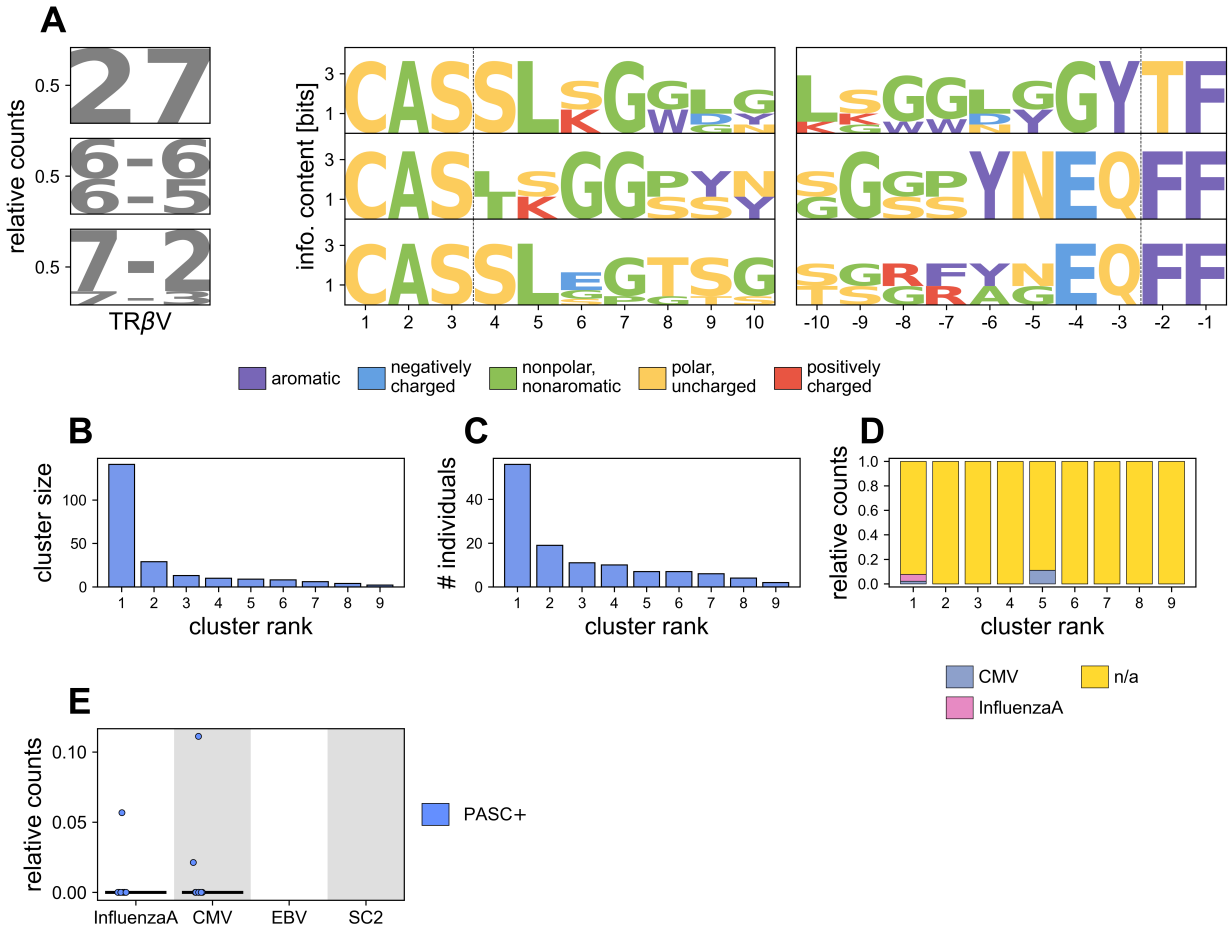


Figure 3.10: **Atypical rare, shared sequences. (A)** Sequence logos describing the composition of amino acid sequences are shown for the significant TCR-OT clusters obtained from the rare, shared PASC+ repertoire. Logos were constructed from focal sequence(s) and members in the cluster that were within 48 TCRdist of the focal sequence(s), and logos are shown only if 10 sequences were present in the constrained neighborhood for a cluster. Logos are ordered top to bottom, with the top logo describing the cluster with the largest size and the bottom logo describing the cluster with the smallest size. (Left) TRβV gene usage frequencies within a cluster are shown as a logo. (Middle, right) The logo plots shown the information content at each position in the amino acid TCRβ CDR3 sequence from left to right (middle) and right to left (right). Amino acids are colored according to their physicochemical properties indicated in the legend. **(B)** The amounts of amino acid sequences in the clusters is shown as a bar graphs. Cluster rank corresponds to the size of the clustering. **(C)** The number of individuals represented in a cluster is shown as a bar graphs. **(D)** The stacked bar graph show the distribution of epitopes of TCRβ sequences in VDJdb that were within 24 TCRdist of a constituent TCRβ in a cluster and whose donor HLA group matched to at least one HLA group of the individual from which the similar TCRβ sequence originated. **E** The epitope distributions of (D) are shown as a strip plot, with the horizontal black lines indicating the medians of the distributions.

3.7 Discussion

COVID-19 has transitioned to endemicity [221], and, coupled with vaccination exhaustion, the lack of well-defined pathogeneses, and the slow pace at which clinical studies are being conducted, PASC will likely continue to afflict millions of individuals and arise in many more [8]. Understanding what components of the T cell response lead to PASC emerging and how the T cell repertoire is organizing during PASC is crucial for elucidating etiologies of PASC and guiding the development of efficacious therapies. Here, we investigated T

cell repertoires collected longitudinally from individuals who had COVID-19 early on in the pandemic, with some subsequently proceeding to present symptoms of PASC.

Studying the TCR repertoire to uncover functional responses is a nontrivial task. An overwhelming majority of observed TCRs are expected to be unspecific with regard to an ongoing immune challenge. To sieve through the repertoire for functional responses, we began by examining global features of the repertoire. We observed that the TCR repertoires of the INCOV cohort were indistinguishable regardless of the PASC status when viewed at the level of receptor features alone. Thus, we continued by looking for local sequence usages which differed between the two cohorts of PASC+ and PASC− individuals by examining three subsets of the repertoire: the most abundant clones, significantly dynamic clones, and public TCRs predicted to be rarely observed and shared.

The most abundant clones in a repertoire ensue from recent expansions or persistently large clones. Large pre-existing clones control early responses to SARS-CoV-2 infections [194] that may affect the likelihood of developing PASC. Additionally, clones specific to SARS-CoV-2 have been observed to be large and stable over eight months post infection in COVID-19 convalescent individuals [350], underscoring the importance of abundant clones in the resolution of SARS-CoV-2 infections. We distinguished atypical sequences between the most abundant TCRs in the cohorts using an optimal transport method adapted for TCRs [220]. From thousands of TCRs, we detected a few hundred which were unique to each cohort. We validated their specificities by comparing their sequences and associated HLAs with databases of TCRs with known reactivity and found no differences.

We next investigated the dynamics of the TCRs in our cohort. We observed many dynamical modes of TCRs likely due to the heterogeneity of sampling of our cohort's repertoires as well as the diversity of responses of individuals recovering and developing PASC. We identified expanding and contracting clones using statistical inference that characterizes the undersampling of the repertoire from biological and technical noise. In lieu of replicate data, we performed *in silico* bootstrapping to estimate the noise. From roughly 20 million receptors, we identified a few thousand TCRs which were significantly dynamic. We tested for the

specificity of the TCRs to SARS-CoV-2 by calculating their overlap with sequences in the COVID-19 MIRA database. Interestingly, only the contracted TCRs were highly enriched for specificity to SARS-CoV-2, with at least 15% of them matching to the MIRA database, when compared to the most abundant TCRs in the repertoires. We refined these dynamic subsets by applying the optimal transport algorithm and found 465 dynamic TCRs that were unique to the PASC+ cohort and 342 dynamic TCRs that were unique to the PACS− cohort. Additionally, we constructed clusters of the dynamic clones and observed that contracted clusters in the PASC+ cohort were enriched for flu. Because our cohort was observed in the beginning of the pandemic, it's likely that some individuals may have had bouts with the flu prior to contracting COVID-19. It's also possible that bystander T cells may have been activated during their SARS-CoV-2 infections, though bystanders in this context have not been reported to being specific to influenza.

The public sharing of TCRs can be indicative of convergent responses to similar pathogenic encounters; however, TCRs can also be incident in many individuals due to similar V(D)J recombination biases, HLA restrictions, and thymic and peripheral selective processes. To characterize our expectations of a TCR's publicness, we learned generation models for each individual and selection models for each sampled repertoire in order to characterize the differences between individuals and longitudinal selection pressures as accurately as possible. We used the models' predictions for the probabilities of observing TCRs in the periphery to estimate the likelihoods of sharing the observed TCRs and identify outliers. Ultimately, we found 19,847 and 13,417 clones from 4 million and 2.6 million shared clones in the PASC+ and PASC− cohorts, respectively, which were unexpected to be shared as widely as they were. These rare subsets were substantially enhanced for specificity to SARS-CoV-2: about 24% of each subset matched to sequences in the MIRA database compared to entire shared repertoire in which roughly 9% matched. We compared the sequence compositions of these two rare, shared subsets and identified 310 sequences specific to the PASC+ cohort and 301 sequences specific to the PASC− cohort. We observed no differences in epitope specificity when comparing them to the VDJdb.

The roles played by the TCRs that we have identified as unique to each cohort in the development or supression of PASC remains an open question. Those detected from the PASC+ cohort may be responding to the reactivation of quiescent viruses, driving autoimmune responses, or contributing to dysregulation. Our cohort was observed very early on in the pandemic and, while on the larger end for immune repertoire studies, is quite small considering the breadth of PASC definitions. Moreover, some observations associated with the earlier responses in the PASC+ cohort may be contaminated by individuals having or recovering from the flu. Additionally, our study probed only the peripheral blood TCR repertoire whereas characterizing the TCR repertoire local to other organ systems, such as the respiratory or gastrointenstinal systems, might yield more insight into possible pathogeneses of PASC. Nevertheless, we present the first study characterizing repertoires longitudinally sampled from individuals who presented symptoms of PASC. From about 91 million TCRs, we've identified 1,091 TCRs which we believe are candidates for discriminating what may lead to PASC or which play an active role in the persistence of PASC.

Chapter 4

THE LIFECYCLE OF B CELL LINEAGES: INFERENCE OF EFFECTIVE MEMORY RECALL TIMESCALES

B cells are the central actors in the adaptive immune system and encode highly diverse and mutable pathogen-engaging receptors. They can counter a multitude of pathogens by directly neutralizing invaders or by storing memory to respond to reinfections efficiently. Upon infection, activated B cells seed germinal centers (GCs) where they hypermutate and are selected for enhanced affinity to pathogens. On longer timescales, memory B cells from previous GC reactions can seed new GCs during reinfection and mutate further; however, the extent of their role in response to reinfections is unclear. Because B cells evolve only in GCs, standard dynamical models with continuous accumulation of mutations fail to describe this interrupted evolution. We introduce a stochastic telegraph process to model the B cell lifecycle by capturing the entry and exit of B cells from GCs and constraining the accumulation of mutations to the GC residents only. We use this model to reconstruct time-resolved evolutionary histories of B cells with an outlook to apply this to a longitudinal dataset of immune repertoires from individuals with HIV and obtain distributions of the rates across repertoires and individuals.

4.1 Introduction

The B cell repertoire is the principal component of the adaptive immune system. B cells express B cell receptors (BCRs) that permit B cells to bind to antigens, ensuing in the neutralization and opsonization of invaders [205, 55]. BCRs are created through a process called V(D)J recombination wherein gene segments coding for BCRs are chosen at random and then combined by inserting and deleting nucleotides randomly at the gene junctions.

When they encounter their cognate antigens, naive B cells activate and can differentiate into into short-lived plasma B cells to secrete antibodies or memory B cells (MBCs) for efficient responses to similar future encounters [24, 219]. Activated B cells can also diversify their BCRs further by migrating to germinal centers (GCs) and undergoing affinity maturation. In GCs, B cells somatically hypermutate their BCRs, clonally expand, and are selected for better binding affinity to the target epitopes displayed in the GC microenvironment [310]. B cells with higher affinity after affinity maturation typically differentiate into long-lived plasma B cells whereas those with lower affinity differentiate into MBCs [280, 229, 274].

Upon reinfection or encountering an antigen from a variant of a previous infection, quiescent MBCs can be reactivated. They can clonally expand and quickly differentiate into short- or long-lived plasma B cells, form and enter GCs to improve their affinity and then differentiate, or act as bystanders [69, 292, 84, 163, 203, 303, 100, 112]. The nature and conditions of whether and how B cell clonal families experience continued affinity maturation through the recall of memory B cells is not well understood. Isotype [69, 224, 188], expression of cell surface markers [12, 298, 353, 321], affinity [309, 305], longevity of pre-existing GCs and time between recurring immune challenges [305], and location of the MBC [17, 162] seem to contribute to the recruitment and fate of an activated MBC. Immune history, such as the order in which pathogenic epitopes are encountered [76, 14, 13, 191, 303]; age [35]; and the immune challenge itself, for instance, flavivirus compared to SARS-CoV-2 [326, 356], also appear to impact the role of MBCs in secondary infections. Uncovering the ontogeny of recalled MBCs and their continued evolution in secondary GCs is obstructed further by dominating naive responses in secondary infections [305, 303]. Moreover, granular properties, such as the sequence features of BCRs which describe the homo- or heterogeneity of clonal families that are recalled within and across individuals, remain to be explored.

Since B cells evolve during affinity maturation, the field of phylogenetics is of great utility for characterizing and understanding B cell biology. Previous work has been devoted to reconstructing phylogenetic topologies specifically for B cells [19, 281, 27, 104, 265, 119, 80, 122, 66, 83, 339, 257, 5, 137], investigating the mutability of the BCR genome [331, 119,

313, 80, 122, 283], clarifying the role of selection [337, 313, 122, 124, 217, 244, 211], improving clonal family inference [282], understanding phenotype information [123, 303, 121, 302, 120, 149], and annotating time on B cell trees [331, 265, 272, 313]. However, no such phylogenetic studies have been conducted yet which annotate times on B cell trees accounting for the mutability of different cell types (quiescent MBCs vs. GC B cells) and the intermittent evolution of B cells in response to successive immune challenges.

The basis of inferring time on a phylogeny is the molecular clock hypothesis: rates at which genomes evolve are constant over time [354]. Since the inception of this hypothesis, different clock models of varying complexities have been created to model processes more consistent with their underlying biology. A strict clock assumes a constant evolutionary rate over the tree [355]. Other models relax this assumption to varying degrees, permitting lineages to have their own evolutionary rate in uncorrelated and correlated fashions [296, 258, 72], to share evolutionary rates among clades [342], and to evolve sites along the genome [301, 18] or evolve branches [127, 18] independently. Though these methods are able to characterize an impressive range of biological phenomena, they are misspecified for modeling B cell evolution and annotating time on phylogenies of B cell lineages due to the quiescent state of MBCs.

Recent developments in the field of phylodynamics have led to multitype birth-death models inferred using Bayesian methods that have been applied to tumor and virus systems in which different subpopulations evolve with different evolutionary parameters [284, 159, 20, 267, 173]. In these models, each branch effectively receives its own local clock but uses much fewer parameters due to the multitype structured nature of the populations. While phenomenologically similar to that of B cell evolution since they study heterogeneous, yet structured populations, these models assume that the parameters, such as the mutation rate and birth rate, are nonzero for all of their population types. Because their inference of the birth rate is inextricably coupled to the mutation rate, the underlying parameters characterizing the deactivated periods of B cell lineages may be at the boundary of the model manifold, leading to algorithmic instability and poor mixing of chains. Additionally, B cell repertoires are sampled from an individual's blood typically, where B cells are not undergoing

affinity maturation (though there are exceptional cases of extrafollicular maturation [67]), making some of the methods inappropriate that assume a mixture of cell types are present in the observed data [173]. Bayesian phylogenetics is also prohibitive computationally in the setting of B cell repertoires in which we desire to study thousands of moderately sized ($\sim 10^2$) lineages.

The seedbank model in population genetics is used to investigate a population that contains dormant constituents [172]. This dormant compartment is constrained not to reproduce, but it is allowed to mutate at a rate slower than the active subset of the population. Extensions of coalescent theory to seedbank populations illustrate that they differ fundamentally from structured multitype populations in which all population compartments can reproduce [26]. Statistical inference on genomic data using site frequency spectra demonstrated that seedbank and multitype populations are indeed identifiable from one another [25]. Recently, a Bayesian inference framework was created to characterize seedbank genomic data phylogenetically [42]. Ref. [42] could infer the tree structure, transition rates between the active and dormant states, and evolutionary parameters when many switches occurred over a lineage and the respective mutation rates of the active and dormant populations were comparable. However, when the timescale of dormancy was increased and the dormant mutation rate was substantially smaller than the active mutation rate, the uncertainty of the branch lengths increased leading to the inference failing to converge unless the tree topology was fixed. Notably, we expect the underlying parameters of B cell genealogies to be concordant with the regime in which their inference was unstable due to larger genealogical uncertainty. Finally, ref. [42] did not demonstrate the performance of their inference when the input data was sampled solely from the dormant compartment of the population; therefore, how its performance will translate in principle to B cell genomic data remains to be understood.

A flexible, maximum-likelihood framework that has been invaluable for simultaneously inferring evolutionary parameters and producing time trees for large phylogenies is TreeTime [256]. TreeTime infers the evolutionary parameters and annotates the time on tree topologies by factorizing the likelihood's dependence on the ancestral sequences, rates, and

times and by using heuristics. This iterative optimization scheme means that its runtime scales linearly with the size of the tree, so it can analyze large phylogenies ($> 10^3$ samples) of rapidly evolving viruses, such as Ebola, SARS-CoV-2, or Influenza [106]. However, while TreeTime has shown convergence to an approximate global optimum for strict molecular clocks, its algorithm is not guaranteed to converge in general.

In this chapter, we derive a mathematical theory for B cell evolution by coarse-graining their activity into two states: activated GC B cells which can accrue mutations and dormant, immutable MBCs which can persist for long periods of time. Our theory is a subset of the seedbank coalescent. We assume that the timescale of branches coalescing along the phylogeny is fast enough such that the slower dynamics of phenotypic switching dictate the time of a lineage. In this reduced model structure which focuses solely on phenotypic switching– the telegraph model–we hope to increase sensitivity for inferring the switching rates. We derive a likelihood which marginalizes over the unseen periods of activity and dormancy in a lineage, and we identify the inter-arrival time distribution for these processes for effective simulation. We demonstrate the model's statistical consistency on star phylogenies. We adapt TreeTime's maximum likelihood approach to infer the evolutionary parameters and time-annotated genealogy, with an outlook for its application to the large amount of phylogenies obtained from B cell repertoires. Ultimately, we show that TreeTime's factorized inference is unable to resolve the parameters fully.

4.2 Deep sequencing of the IgH repertoire in HIV-infected individuals shows measurably evolved clonal lineages

We use the dataset from ref. [288] which contains 10 to 20 longitudinal samples taken from 10 HIV-infected male individuals, with a majority of the samples collected before the men received antiretroviral therapy (Fig. 4.1A). Ref. [288] performed deep sequencing of the variable region of the immunoglobulin heavy chain locus (IgH), i.e., the BCR repertoire. We created a pipeline to download the BCR repertoires, process the data, and identify clonal lineages (Appendix C). We reconstructed phylogenies using the GY94 substitution model [97]

as implemented in IgPhyML [122] for only those lineages in which BCRs from at least two timepoints were observed and in which the sum of unique BCRs observed at each time point was at least 100, amounting to 0.2% of the entire dataset (Fig. 4.1B; see also Fig. C.1). We required the latter condition so inference of the phylogeny and its associated substitution rate would be more robust. To resolve a phylogeny in time, it is necessary to ensure the tree has sufficient temporal signal, i.e., is measurably evolving. A tree possesses this property if there is sufficient change in the genetic sequences between tips sampled at different times, which yields a positive correlation between sampling times of the nodes and their root-to-tip divergences [245]. We applied a permutation test, which takes into account idiosyncrasies of B cell lineages, to our reconstructed phylogenies to determine if lineages were measurably evolving [121]. Lineages with $p < 0.05$ were said to be measurably evolving and retained for downstream analyses. Of those trees tested, we detected 12%-20% of lineages to be measurably evolving (Fig. 4.1C). The distribution of IGHV gene usages was similar between lineages detected as measurably evolving and lineages without temporal signal whereas the HCDR3 length distribution was shifted to longer lengths for the measurably evolved lineages (Manny-Whitney U test, $n_- = 10,984$, $n_+ = 1,938$, $U = 10952973$, $p = 0.0081$).

Fig. 4.1D shows the distribution of the inferred substitution rate obtained from regressing root-to-tip divergence onto sample time for the measurably evolved phylogenies. Notably, this regression assumes a molecular clock hypothesis, i.e., a likelihood given by the Poisson distribution with a fixed substitution rate. The distribution spans roughly two orders of magnitude, with the median of the distribution about 1.6 orders of magnitude less than the observed mutation rate of 4×10^{-3} day^{-1} site^{-1} [151, 312]. The discrepancy can be due to the following reasons. (i) The substitution rate is the rate at which sites in the sequence mutate and then reach fixation, a function of selection pressures, and will necessarily be less than the mutation rate since most mutations are deleterious and therefore not observed. Even BCRs collected from healthy individuals exhibit a large range of selection coefficients, as quantified by the ratio of nonsynonymous to synonymous mutations in a lineage d_N/d_S [217, 282]. (ii) Additionally, phylogenies of B cells are inherently undersampled views of the entire

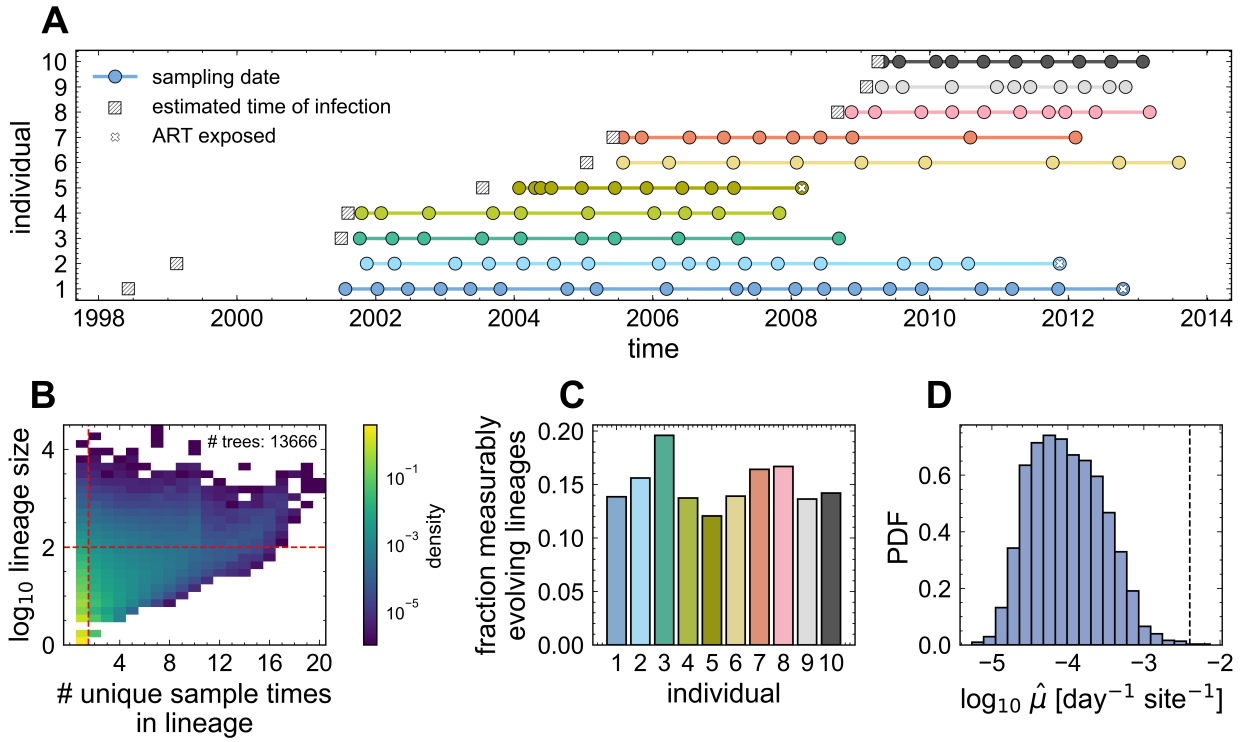


Figure 4.1: **Cohort of individuals infected with HIV has measurably evolving lineages.** **(A)** The scatter and line plot shows the estimated time of infection and times at which repertoire samples were taken for each individual (y-axis, colors). **(B)** The two-dimensional histogram shows how many lineages were observed at a given number of time points (x-axis) and their size, defined here as the sum of the number of unique receptors observed at each time point (y-axis). The horizontal red, dashed line is drawn at a lineage size of 100, and the vertical, red dashed line indicates that lineages binned to the right of it were seen at least two time points. The amount of lineages in the upper right quadrant demarcated by red, dashed lines is annotated in the upper right of the plot. **(C)** The bar plot shows the fraction of lineages which were detected as measurably evolving for each individual (x-axis, colors). **(D)** The histogram shows the distribution of $\log_{10}$ substitution rates inferred from Poissonian root-to-tip regression performed on the trees detected as measurably evolving. The vertical black, dashed line shows the expected mutation rate for a B cell [151].

mutational history of lineages due to limited sampling longitudinally as well as at a single time point. In effect, intermediate mutations, mutations which affect the same site, and reversions may be missed. However, some argue that using a continuous-time Markov model of evolution, as is done in root-to-tip regression, captures the full mutational history and mitigates this undersampling [81]. (iii) Lastly, substitution rates may be underestimated due to phenotypic switching along the lineages and the intermittent evolution of B cells when not correctly accounting for the time a B cell lineage is active. The substitution rate inferred under a molecular clock hypothesis has been computed assuming that mutations can occur along any point in the lineage; however, B cell lineages sampled across months and years are most likely not undergoing continuous evolution over those timescales.

Fig. C.2C-D shows how the Poissonian substitution rate changes as a function of IGHV gene usage (C) and HCDR3 length (D). While the inferred substitution rate does not appear to be modulated in a meaningful way by the IGHV gene, it increases with increasing HCDR3 length (Pearson $r = 0.082$, $p = 0.00029$). Curiously, this trend holds even though the substitution rate should be invariant to the sequence length since it is measured of units of $\text{site}^{-1}\,\text{day}^{-1}$.

4.3 Modeling the lifecycle of B cell lineages

The dynamics of B cell lineages concomitant with the timescales at which repertoires are sampled can be coarse-grained into two states. Lineages exist either in an active state (on), meaning some cells in the lineage are in a GC and accrue mutations, or they are quiescent (off), meaning all cells in the lineage are circulating in the periphery and not mutating. The effective rate of substitution $\mu(t)$ at time t along a lineage is given by

$$\mu(t) = \begin{cases} \mu, & \psi(t) = \text{on}, \\ 0, & \psi(t) = \text{off} \end{cases} \tag{4.1}$$

where $\psi(t) \in \{\text{on}, \text{off}\}$ is the state of the lineage at time t and is a stochastic quantity. The process of accumulating mutations can then be described by a Cox process, a Poisson process

with randomly varying intensity [52].

We model the process governing state transitions as a *telegraph process* [89], otherwise known as a dichotomous Markov process [209]. We assume that the duration of time a lineage spends in the GC and the duration of time a lineage spends in circulation are exponentially distributed with characteristic timescales α_-^{-1} and α_+^{-1}, respectively. In other words, α_- is the rate at which active lineages deactivate, and α_+ is the rate at which inactive lineages reactivate. All together, this sort of model has been referred to as a renewal Cox process with Markovian intensity [279].

The probability of observing m substitutions on a B cell lineage with lifetime $\tau \geq 0$ is Poissonian with a time-dependent substitution rate:

$$P(m|\tau;\mu) = \frac{\left(\int_0^\tau dt\,\mu(t)\right)^m}{m!} \exp\left(-\int_0^\tau dt\,\mu(t)\right). \tag{4.2}$$

The integral evaluates to

$$\int_0^\tau dt\,\mu(t) = \mu \int_0^\tau dt\,\mathbb{1}_{\text{on}}(\psi(t)) = \mu\tau x, \tag{4.3}$$

where $\mathbb{1}_{\text{on}}$ is the indicator function yielding 1 if $\psi(t) = \text{on}$, and x is the fraction of a lineage's lifetime spent in the activated state

$$x = \frac{1}{\tau}\int_0^\tau dt\,\mathbb{1}_{\text{on}}(\psi(t)). \tag{4.4}$$

The model that characterizes m substitutions on a branch is

$$P(m|x,\tau;\mu) = \frac{(\mu x\tau)^m}{m!}\exp(-\mu x\tau). \tag{4.5}$$

Crucially, x is a stochastic variable which is not observable. We seek to marginalize x out of our model. Thus, we must construct the fractional occupation time distribution which characterizes the probability of the fraction of time x that a lineage spends in the active state.

4.3.1 The propagator of the telegraph process

We first derive the solutions to the Kolmogorov equations, known in physics as the propagator, to understand how the states of the system evolve in time. These solutions will be instrumental in marginalizing out the fractional occupation time. With $\psi \in \{\text{on}, \text{off}\}$ the state of a B cell lineage and t the time on a lineage, let $|\psi_i, t_i\rangle$, called ket, be a state vector that contains the probabilities of being in state ψ_i at time t_i. (In this context, it is shorthand for a row vector.) Conversely, $\langle\psi_f, t_i + t|$, called bra, is dual to the ket and denotes the probability associated with the lineage being in state ψ_f at time $t_i + \tau$ as a column vector. The propagator is defined as

$$P(\psi_f, t_i + \tau|\psi_i, t_i) = \langle\psi_f, t_i + \tau|\psi_i, t_i\rangle = \langle\psi_f| \boldsymbol{T}(t_i, t_i + \tau) |\psi_i\rangle , \tag{4.6}$$

where $\boldsymbol{T}$ is the time evolution operator that transforms ψ_i into ψ_f.

We determine $\boldsymbol{T}$ by defining the infinitesimal transition rate matrix $\boldsymbol{Q}$ which generates the probability that an event occurs within an infinitesimally small interval of time δt. Letting $\boldsymbol{I}$ denote the identity matrix,

$$\langle\psi_f|\boldsymbol{T}(t_i, t_i + \delta t)|\psi_i\rangle = \langle\psi_f|\boldsymbol{I}|\psi_i\rangle + \langle\psi_f|\boldsymbol{Q}|\psi_i\rangle \, \delta t + o(\delta t) \tag{4.7}$$

For finite times,

$$\boldsymbol{T}(t_i, t_i + \tau) = e^{\boldsymbol{Q}\tau} = \begin{pmatrix} P(\psi_f = \text{on}, t_i + \tau|\psi_i = \text{on}, t_i) & P(\psi_f = \text{on}, t_i + \tau|\psi_i = \text{off}, t_i) \\ \\ P(\psi_f = \text{off}, t_i + \tau|\psi_i = \text{on}, t_i) & P(\psi_f = \text{off}, t_i + \tau|\psi_i = \text{off}, t_i) \end{pmatrix} . \tag{4.8}$$

We define the basis for our two-state system as

$$|\text{on}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \quad |\text{off}\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} . \tag{4.9}$$

Constructing the transition rate matrix as a left stochastic matrix, in agreement with the ordering of Eq. 4.8, gives

$$\boldsymbol{Q} = \begin{pmatrix} -\alpha_- & \alpha_+ \\ \alpha_- & -\alpha_+ \end{pmatrix} . \tag{4.10}$$

The off-diagonal entries specify the probabilities of transitioning during infinitesimally small intervals of time while the diagonal entries ensure a null net probability flow. $\boldsymbol{Q}$ has eigenvalues $\lambda_0 = 0$, and $\lambda_1 = -(\alpha_- + \alpha_+)$, with corresponding eigenvectors

$$\vec{v}_0 = \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix}, \quad \vec{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \tag{4.11}$$

We diagonalize $\boldsymbol{Q} = \boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{V}$ where

$$\boldsymbol{V} = \begin{pmatrix} -1 & \alpha_+ \\ 1 & \alpha_- \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} -(\alpha_- + \alpha_+) & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.12}$$

Then the entries of the propagator for the telegraph process are

$$\begin{aligned} \boldsymbol{T}(t_i, t_i + \tau) = e^{\boldsymbol{Q}\tau} &= \boldsymbol{V}^{-1} e^{\boldsymbol{D}\tau} \boldsymbol{V} \\ &= \frac{1}{\alpha_- + \alpha_+} \begin{pmatrix} \alpha_+ + \alpha_- e^{-(\alpha_- + \alpha_+)\tau} & \alpha_+ \left(1 - e^{-(\alpha_- + \alpha_+)\tau}\right) \\ \alpha_- \left(1 - e^{-(\alpha_- + \alpha_+)\tau}\right) & \alpha_- + \alpha_+ e^{-(\alpha_- + \alpha_+)\tau} \end{pmatrix}, \end{aligned} \tag{4.13}$$

where the boundary conditions for each entry are notated in Eq. 4.8.

The fractional occupation time distribution

We next derive the fractional occupation time distribution which characterizes the distribution of x. This will allow us to marginalize x from our likelihood since, as remarked, it is unobservable for our process. Additionally, the fractional occupation time distribution can be used to compute posteriors of x on the branches of trees.

The occupation time distribution describing any iid telegraphing process was described by refs. [344, 345]. We summarize and quote their work here generally. Let $\{T_{\text{on},1}, T_{\text{on},2}, T_{\text{on},3}, ...\}$ and $\{T_{\text{off},1}, T_{\text{off},2}, T_{\text{off},3}, ...\}$ be sequences of independent, positive random variables which represent the periods of time spent in the on state and off state, respectively. Let $F(t; \theta_{\text{on}})$ and $G(t; \theta_{\text{off}})$ be the cumulative distribution functions of $T_{\text{on},i}$ and $T_{\text{off},i}$, respectively, and $f(t; \theta_{\text{on}})$ and $g(t; \theta_{\text{off}})$ denote the respective probability densities. The sum of n on-periods

$\sum_{i=1}^{n} T_{\text{on},i}$ is characterized by the n-fold convolution of the cumulative distribution function $F(t;\theta_{\text{on}})$ and density $f(t;\theta_{\text{off}})$, denoted $F(t|n;\theta_{\text{on}})$ and $f(t|n;\theta_{\text{on}})$. A similar construction is used for the sum of n off-periods $\sum_{i=1}^{n} T_{\text{off},i}$. The joint probability $P(x, n, \psi_f|\psi_i, \tau; \theta_{\text{on}}, \theta_{\text{off}})$ of the fractional occupation time x, the final state $\psi_f = \psi(\tau)$, and the number of switches n given the initial state $\psi_i = \psi(0)$, lifetime τ, and parameters is derived as follows.

Consider $\psi_i = \psi_f = \text{on}$. When $n = 0$, no switching occurs and $x = 1$. The probability that $x = 1$ is the probability of no switching over the interval τ: $e^{-\theta_{\text{on}}\tau}$. Now, fix $n \geq 1$. The sequence of times appears as $T_{\text{on},1} + T_{\text{off},1} + ... + T_{\text{on},n} + T_{\text{off},n} + T_{\text{on},n+1}$. This corresponds to n transitions from on $\rightarrow$ off, n transitions from off $\rightarrow$ on, $n + 1$ periods of T_{on}, and n periods of T_{off} occurring over the interval τ. Crucially, it does not matter when the $n + 1^{\text{th}}$ T_{on} ends, only that it began within the time interval τ, i.e., n switches from on $\rightarrow$ off took place. On the other hand, only n periods of T_{off} could have occurred in the time interval τ, and the total time spent in the off state is constrained $\sum_{i=1}^{n} T_{\text{off},i} = \tau(1-x)$. Thus, the joint probability of x, n, and $\psi_f = \text{on}$ given $\psi_i = \text{on}$ and τ is

$$P(x, n, \text{on}|\text{on}, \tau; \theta_{\text{on}}, \theta_{\text{off}})\, dx =$$
$$dx \left(\delta_{n,0}\delta(x-1)e^{-\theta_{\text{on}}\tau} + \mathbb{1}_{n>0}\Big[F(\tau x|n;\theta_{\text{on}}) - F(\tau x|n+1;\theta_{\text{on}})\Big] g(\tau(1-x)|n;\theta_{\text{off}})\,\tau\right). \quad (4.14)$$

$\delta_{n,0}$ is a Kronecker delta function that is 1 when $n = 0$ and 0 otherwise, $\delta(x)$ is a delta function, and $\mathbb{1}_{n>0}$ is an indicator function which gives 1 when $n > 0$ and 0 otherwise. The term in square brackets gives the difference between the probability that $\sum_{i=1}^{n} T_{\text{on},i} = \tau x$ and the probability that $\sum_{i=1}^{n+1} T_{\text{on},i} = \tau x$. This difference translates to the probability of observing n events of switching from the on state to the off state over the interval τx [145] and, in other words, constrains the number of switches from on $\rightarrow$ off. The unbracketed density term, $g(\tau(1-x)|n;\theta_{\text{off}})\,\tau\, dx$, is the probability of spending a total time of $\tau(1-x)$ in the off state over n periods of T_{off}, so it constrains the time spent in the off state. The extra factor τ ensues from us investigating the probability of the fractional occupation time x rather than the occupation time τx:

$$P(\tau x, n, \text{on}|\text{on}, \tau, \theta_{\text{on}}, \theta_{\text{off}})d(\tau x) = P(x, n, \text{on}|\text{on}, \tau, \theta_{\text{on}}, \theta_{\text{off}})\tau dx. \quad (4.15)$$

The joint probability for starting and ending in the off state follows similarly,

$$P(x, n, \text{off}|\text{off}, \tau; \theta_{\text{on}}, \theta_{\text{off}})\, dx = \\ dx \left(\delta_{n,0} \delta(x)\, e^{-\theta_{\text{off}} \tau} + \mathbb{1}_{n>0} \Big[G(\tau(1-x)|n; \theta_{\text{off}}) - G(\tau(1-x)|n+1; \theta_{\text{off}}) \Big] f(\tau x|n; \theta_{\text{on}})\, \tau \right), \tag{4.16}$$

Observe that when $\psi_i = \psi_f = \text{off}$, there is term which is a delta function at $x = 0$. This term results from when no switching occurs (as seen by the Kronecker delta func $\delta_{n,0}$), and its coefficient is $e^{-\theta_{\text{off}}\tau}$, the probability of no switching occurring when $\psi_i = \psi_f = \text{off}$.

Now, consider $\psi_i = \text{on}$ and $\psi_f = \text{off}$ with $n \geq 0$ fixed. The sequence of time periods appears as $T_{\text{on},1} + T_{\text{off},1} + ...T_{\text{on},n} + T_{\text{off},n}$. This corresponds to constraining that there must have been $n+1$ transitions from on $\rightarrow$ off and n transitions from off $\rightarrow$ on. Because $\psi_f = \text{off}$, the magnitude of $\sum_{i=1}^{n+1} T_{\text{off},i}$ is irrelevant since it exceeds the time interval τ; however, $\psi_f = \text{off}$ necessitates that there must have been n transitions from off $\rightarrow$ on. Moreover, $\sum_{i=1}^{n+1} T_{\text{on},i} = \tau x$ must hold within the interval τ because, again, $\psi_f = \text{off}$. Thus, we have

$$P(x, n, \text{off}|\text{on}, \tau; \theta_{\text{on}}, \theta_{\text{off}}) dx = \Big[G(\tau(1-x)|n; \theta_{\text{off}}) - G(\tau(1-x)|n+1; \theta_{\text{off}}) \Big] f(\tau x|n+1; \theta_{\text{on}})\, \tau dx \tag{4.17}$$

where the bracketed term specifies n switches from off $\rightarrow$ on, as described previously, and the density $f(\tau x|n+1, \theta_{\text{on}})\, \tau dx$ constrains that $\sum_{i=1}^{n+1} T_{\text{on},i} = \tau x$. Similarly for $\psi_i = \text{off}$ and $\psi_f = \text{on}$,

$$P(x, n, \text{on}|\text{off}, \tau; \theta_{\text{on}}, \theta_{\text{off}})\, dx = \Big[F(\tau x|n; \theta_{\text{on}}) - F(\tau x|n+1; \theta_{\text{on}}) \Big] g(\tau(1-x)|n+1; \theta_{\text{off}})\, \tau dx. \tag{4.18}$$

Specifying that the periods $T_{\text{on},i}$, $T_{\text{off},i}$ be exponentially distributed with α_-^{-1} giving the average length of $T_{\text{on},i}$ and α_+^{-1} giving the average length of $T_{\text{off},i}$, the distributions that characterize the sum of n on-periods and the sum of n off-periods are Erlang(n, α_-) and Erlang(n, α_+), respectively, i.e., the n-fold convolution of the exponential distribution. We chose to model the periods of time with exponentially distributed random variables because

we expect switching to be rare in short time intervals. Moreover, these choices yield the most parsimonious telegraph model in which the on state and off state have different characteristic timescales. Additionally, modeling $T_{\text{on},i}$, $T_{\text{off},i}$ with exponential distributions with different characteristic timescales is the most parsimonious telegraph model. The exponential and Erlang distributions are, respectively,

$$f(t;\alpha_-) = \alpha_- e^{-\alpha_- t}, \quad f(t|n;\alpha_-) = \frac{\alpha_-^n t^{n-1}}{(n-1)!} e^{-\alpha_- t}. \tag{4.19}$$

The difference in consecutive terms of the CDF of the Erlang distribution is the Poisson distribution:

$$F(t|n;\alpha_-) - F(t|n+1;\alpha_-) = P(n|t;\alpha_-) = \frac{(\alpha_-\tau)^n}{n!} e^{-\alpha_- t}. \tag{4.20}$$

With these assumptions, we have for $\psi_i = \psi_f = \text{on}$

$$\begin{aligned} &P(x, n, \psi_f = \text{on}|\psi_i = \text{on}, \tau; \alpha_-, \alpha_+)dx = \\ &dx\left(\delta_{n,0}\delta(x-1)\, e^{-\alpha_-\tau} + \mathbb{1}_{n>0}\Big[F(\tau(1-x)|n;\alpha_-) - F(\tau(1-x)|n+1;\alpha_-)\Big] g(\tau x|n;\alpha_+)\,\tau\right) \\ &\quad = dx\left(\delta_{n,0}\delta(x-1)e^{-\alpha_-\tau} + \mathbb{1}_{n>0}\,\tau\left[\frac{(\alpha_-\tau x)^n}{n!}e^{-\alpha_- t x}\right]\frac{\alpha_+^n (t(1-x))^{n-1}}{(n-1)!}e^{-\alpha_+ t(1-x)}\right) \\ &\quad = dx\left(\delta_{n,0}\delta(x-1)e^{-\alpha_-\tau} + \mathbb{1}_{n>0}\,\frac{(\alpha_-\alpha_+\tau^2 x)^n (1-x)^{n-1}}{n!\,(n-1)!}e^{-(\alpha_-\tau x + \alpha_+\tau(1-x))}\right). \end{aligned} \tag{4.21}$$

Substituting the appropriate Poisson and Erlang distributions into the other joint probabilities with $\psi_i = \psi_f = \text{off}$; $\psi_i = \text{on}, \psi_f = \text{off}$; and $\psi_i = \text{off}, \psi_f = \text{on}$ (Eqs. 4.16 - 4.18) yields

$$\begin{aligned} &P(x, n, \psi_f = \text{off}|\psi_i = \text{off}, \tau; \alpha_-, \alpha_+)dx = \\ &\qquad dx\left(\delta_{n,0}\delta(x)\, e^{-\alpha_+\tau} + \mathbb{1}_{n>0}\frac{(\alpha_-\alpha_+\tau^2(1-x))^n x^{n-1}}{n!\,(n-1)!}e^{-(\alpha_-\tau x + \alpha_+\tau(1-x))}\right), \end{aligned} \tag{4.22}$$

$$P(x, n, \psi_f = \text{off}|\psi_i = \text{on}, \tau; \alpha_-, \alpha_+)dx = dx\,\frac{\alpha_-\tau(\alpha_-\alpha_+\tau^2 x(1-x))^n}{(n!)^2}e^{-(\alpha_-\tau x + \alpha_+\tau(1-x))}, \tag{4.23}$$

$$P(x, n, \psi_f = \text{on}|\psi_i = \text{off}, \tau; \alpha_-, \alpha_+)dx = dx\,\frac{\alpha_+\tau(\alpha_-\alpha_+\tau^2 x(1-x))^n}{(n!)^2}e^{-(\alpha_-\tau x + \alpha_+\tau(1-x))}. \tag{4.24}$$

The number of switching events is unobservable; therefore, we marginalize over n by summing from $n = 0$ to $n = \infty$. This gives the joint probability distributions of the final state ψ_f and the fractional occupation time x,

$$P(x, \psi_f = \text{on}|\psi_i = \text{on}, \tau; \alpha_-, \alpha_+)dx = dx\Big[\delta(x-1)\, e^{-\alpha_- \tau} + \tau\sqrt{\alpha_-\alpha_+}e^{-\alpha_+\tau}e^{-(\alpha_- - \alpha_+)\tau x}\sqrt{\frac{x}{1-x}}I_1\Big(2\tau\sqrt{\alpha_-\alpha_+}\sqrt{x(1-x)}\Big)\Big], \quad (4.25)$$

$$P(x, \psi_f = \text{off}, |\psi_i = \text{off}, \tau; \alpha_-, \alpha_+)dx = dx\Big[\delta(x)e^{-\alpha_+\tau} + \tau\sqrt{\alpha_-\alpha_+}e^{-\alpha_+\tau}e^{-(\alpha_- - \alpha_+)\tau x}\sqrt{\frac{1-x}{x}}I_1\Big(2\tau\sqrt{\alpha_-\alpha_+}\sqrt{x(1-x)}\Big)\Big], \quad (4.26)$$

$$P(x, \psi_f = \text{off}|\psi_i = \text{on}, \tau; \alpha_-, \alpha_+)dx = dx\,\alpha_-\tau e^{-\alpha_+\tau}e^{-(\alpha_- - \alpha_+)\tau x}I_0\Big(2\tau\sqrt{\alpha_-\alpha_+}\sqrt{x(1-x)}\Big), \quad (4.27)$$

$$P(x, \psi_f = \text{on}|\psi_i = \text{off}, \tau; \alpha_-, \alpha_+)dx = dx\,\alpha_+\tau e^{-\alpha_+\tau}e^{-(\alpha_- - \alpha_+)\tau x}I_0\Big(2\tau\sqrt{\alpha_-\alpha_+}\sqrt{x(1-x)}\Big). \quad (4.28)$$

I_n is the modified Bessel function of the first kind. Observe that Eq. 4.25 and Eq. 4.26 are similar except for their delta function term and the square-root term ($\sqrt{x/(1-x)}$ versus $\sqrt{(1-x)/x}$, where the former term gives much larger weight to $x \in (0.5, 1)$ and the latter gives larger weight to $x \in (0, 0.5)$). On the other hand, Eq. 4.27 and Eq. 4.28 mirror each other exactly up to a normalization constant due to time-reversal symmetry.

We expect that the nodes in B cell phylogenies will occupy only one state ψ_f. Then the probability we are interested in is $P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+)$ instead of $P(x, \psi_f|\psi_i, \tau; \alpha_-, \alpha_+)$. The probability of x given the final state ψ_f is obtained by normalizing the joint probability by the telegraph propagator (Eq. 4.13).

$$P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+) = \frac{P(x, \psi_f|\psi_i, \tau; \alpha_-, \alpha_+)}{P(\psi_f|\psi_i, \tau; \alpha_-, \alpha_+)} \quad (4.29)$$

4.3.2 The likelihood of observing m mutations given a branch with time τ

Now that we have the telegraph propagator and the fractional occupation time distribution, we can derive the likelihood of observing m mutations given the branch has time τ. This

likelihood will be used to characterize μ, α_-, and α_+ since we will treat the tree topology and branches annotated with mutations as our observables. Using Eq. 4.5 and Eqs. 4.25 - 4.28, we want to construct the likelihood $P(m|\psi_i, \psi_f, \tau; \mu, \alpha_-, \alpha_+)$ by marginalizing over x:

$$P(m|\psi_i, \psi_f, \tau; \mu, \alpha_-, \alpha_+) = \int_0^1 dx \, \frac{(\mu x \tau)^m}{m!} e^{-\mu x \tau} P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+). \tag{4.30}$$

However, we can also obtain a differential-form of the likelihood by rearranging Eq. 4.30.

$$\begin{aligned}\int_0^1 dx \, \frac{(\mu x \tau)^m}{m!} e^{-\mu x \tau} P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+) &= \frac{\mu^m}{m!} \int_0^1 dx \, (x\tau)^m e^{-\mu x \tau} P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+) \\ &= \frac{\mu^m}{m!} \frac{d^m}{d(-\mu)^m} \int_0^1 dx \, e^{-\mu x \tau} P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+) \\ &= \frac{\mu^m}{m!} \frac{d^m}{d(-\mu)^m} \langle e^{-\mu x \tau} \rangle_{P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+)}\end{aligned} \tag{4.31}$$

where $\langle y \rangle_{P(x)}$ denotes the expectation of y with respect to $P(x)$. Thus, the likelihood can be computed by finding the moment-generating function of the fractional occupation time distribution:

$$P(m|\psi_i, \psi_f; \mu, \alpha_-, \alpha_+, \tau) = \frac{\mu^m}{m!} \frac{d^m}{d(-\mu)^m} \langle e^{-\mu x \tau} \rangle_{P(x|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+)}. \tag{4.32}$$

Similar methods have been employed for deriving likelihoods in coalescent theory [177] and approaching population genetics models using phase-type distributions [115].

To compute the moment-generating function, the transition rate matrix $\boldsymbol{Q}$ defined in Eq. 4.10 can be altered to include a source term which, when exponentiated, effectively keeps track of when time is spent in the on-state. Choosing $-\mu$ as the source term to match Eq. 4.32, the matrix which allows the states to evolve over time and which keeps track of when the on-state is occupied is

$$\boldsymbol{M} = \begin{pmatrix} -\alpha_- - \mu & \alpha_+ \\ \alpha_- & -\alpha_+ \end{pmatrix} = \boldsymbol{Q} + \begin{pmatrix} -\mu & 0 \\ 0 & 0 \end{pmatrix} = \boldsymbol{Q} + \boldsymbol{J}. \tag{4.33}$$

In the following, we use path integration to show $\boldsymbol{M}$ is the desired matrix and illustrate that $\boldsymbol{J}$ keeps track of time spent in the on-state as claimed. (For a review on path integration,

see refs. [82, 129]) We calculate

$$\langle\psi_f|e^{(\boldsymbol{Q}+\boldsymbol{J})\tau}|\psi_i\rangle\,. \tag{4.34}$$

To compute the path integral, we divide time τ into N segments with each segment having time period $\epsilon = \tau/N$.

$$\langle\psi_f|\left(e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}\right)^N|\psi_i\rangle \tag{4.35}$$

Because $|\psi\rangle$ forms a complete set of states

$$\sum_{\psi}|\psi\rangle\langle\psi| = \boldsymbol{I}, \tag{4.36}$$

inserting $\boldsymbol{I}$ between all factors of $e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}$ yields

$$\langle\psi_f|\left(e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}\right)^N|\psi_i\rangle = \left(\prod_{j=1}^{N-1}\sum_{\psi_j}\right)\langle\psi_f|e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}|\psi_{N-1}\rangle\ldots\langle\psi_1|e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}|\psi_i\rangle\,. \tag{4.37}$$

For short-time propagators, we need only $\mathcal{O}(\epsilon)$ accuracy as $N\to\infty$, and the Trotter decomposition rule [299] gives

$$e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon} = e^{\boldsymbol{Q}\epsilon}e^{\boldsymbol{J}\epsilon}. \tag{4.38}$$

Using this approximation,

$$\langle\psi_k|e^{\boldsymbol{Q}\epsilon}e^{\boldsymbol{J}\epsilon}|\psi_i\rangle = \begin{cases} e^{-\mu\epsilon}\langle\psi_k|e^{\boldsymbol{Q}\epsilon}|\psi_l\rangle\,, & \psi_l = \text{on} \\ \langle\psi_k|e^{\boldsymbol{Q}\epsilon}|\psi_l\rangle\,, & \psi_l = \text{off}. \end{cases}$$

The path integral now appears as

$$\lim_{N\to\infty}\langle\psi_f|\left(e^{(\boldsymbol{Q}+\boldsymbol{J})\epsilon}\right)^N|\psi_i\rangle = \lim_{N\to\infty}\left(\prod_{j=1}^{N-1}\sum_{\psi_j}\right)\langle\psi_f|e^{\boldsymbol{Q}\epsilon}|\psi_{N-1}\rangle\ldots\langle\psi_1|e^{\boldsymbol{Q}\epsilon}|\psi_i\rangle\, e^{-\mu\epsilon\left[\delta_{\psi_i,\,\text{on}}+\sum_{j=1}^{N-1}\delta_{\psi_j,\,\text{on}}\right]}, \tag{4.39}$$

where $\langle\psi_f|e^{\boldsymbol{Q}\epsilon}|\psi_{N-1}\rangle\ldots\langle\psi_1|e^{\boldsymbol{Q}\epsilon}|\psi_i\rangle$ is time-ordered and represents the probability of a path being traversed. The exponential term, $e^{-\mu\epsilon\left[\delta_{\psi_i,\,\text{on}}+\sum_{j=1}^{N-1}\delta_{\psi_j,\,\text{on}}\right]}$, counts how many periods of time of length ϵ were spent in the on-state for a given path. Taking n derivatives with

respect to the source term $-\mu$, letting $-\mu \to 0$, and summing over all paths permits us to average n powers of $\left[\epsilon\left(\delta_{\psi_i,\,\mathrm{on}} + \sum_{j=1}^{N-1}\delta_{\psi_j,\,\mathrm{on}}\right)\right]$. Thus, we have found the form of the moment-generating function.

Assuming that $\tau \geq 0$, the (unnormalized) moment-generating functions that characterize the occupation time distributions for each combination of boundary conditions are

$$\langle \psi_f = \mathrm{on}|\, e^{\boldsymbol{M}\tau}\, |\psi_i = \mathrm{on}\rangle = \left(e^{-\gamma_- \tau}\frac{d\gamma_-}{d\mu} + e^{-\gamma_+\tau}\frac{d\gamma_+}{d\mu}\right), \tag{4.40}$$

$$\langle \psi_f = \mathrm{off}|\, e^{\boldsymbol{M}\tau}\, |\psi_i = \mathrm{off}\rangle = \left(e^{-\gamma_- \tau}\frac{d\gamma_+}{d\mu} + e^{-\gamma_+\tau}\frac{d\gamma_-}{d\mu}\right), \tag{4.41}$$

$$\langle \psi_f = \mathrm{off}|\, e^{\boldsymbol{M}\tau}\, |\psi_i = \mathrm{on}\rangle = \frac{\alpha_-}{\eta}\left(e^{-\gamma_- \tau} - e^{-\gamma_+\tau})\right), \tag{4.42}$$

$$\langle \psi_f = \mathrm{on}|\, e^{\boldsymbol{M}\tau}\, |\psi_i = \mathrm{off}\rangle = \frac{\alpha_+}{\eta}\left(e^{-\gamma_- \tau} - e^{-\gamma_+\tau}\right), \tag{4.43}$$

where we have defined

$$\lambda = \mu + \alpha_- + \alpha_+, \quad \rho = \mu + \alpha_- - \alpha_+, \quad \eta = \sqrt{\lambda^2 - 4\mu\alpha_+} = \sqrt{\rho^2 + 4\alpha_-\alpha_+}, \tag{4.44}$$

and

$$\gamma_\pm = \frac{\lambda \pm \eta}{2} = \frac{\lambda\left(1 \pm \sqrt{1 - 4\mu\alpha_+/\lambda^2}\right)}{2}. \tag{4.45}$$

Note that $-\gamma_\pm$ are the eigenvalues of $\boldsymbol{M}$, and $(-(\rho \pm \eta)/2\alpha_-, 1)^\top$ are the eigenvectors of $\boldsymbol{M}$. Importantly, λ, η, $\gamma_\pm > 0$ and $\lambda \geq \eta \geq \rho$. Moreover, the derivatives of $\gamma_\pm$ with respect to μ are

$$\frac{d\gamma_\pm}{d\mu} = \frac{1}{2}\left(1 \pm \frac{\rho}{\eta}\right) \in [0, 1]. \tag{4.46}$$

The telegraph rates in terms of $\gamma_\pm$ are

$$\alpha_- = \gamma_+ + \gamma_- - \left(\frac{\gamma_+\gamma_-}{\mu} + \mu\right), \qquad \alpha_+ = \frac{\gamma_+\gamma_-}{\mu}. \tag{4.47}$$

The unnormalized moment-generating functions given by Eqs. 4.40-4.43 are the joint distributions for observing m mutations and final state ψ_f given an initial state ψ_i and lifetime τ in their respective ensembles:

$$P(m, \psi_f|\psi_i, \tau; \mu, \alpha_-, \alpha_+) = \frac{\mu^m}{m!}\frac{d^m}{d(-\mu)^m}\langle \psi_f|\, e^{\boldsymbol{M}\tau}\, |\psi_i\rangle\,. \tag{4.48}$$

As expected, the normalization of the moment-generating function, given by the zeroth moment, is the telegraph propagator (Eq. 4.13)

$$\lim_{-\mu\to 0} \frac{d^0}{d(-\mu)^0} \langle \psi_f |\, e^{\boldsymbol{M}\tau} \,| \psi_i \rangle = \langle \psi_f |\, e^{\boldsymbol{Q}\tau} \,| \psi_i \rangle = P(\psi_f | \psi_i, \tau; \alpha_-, \alpha_+), \tag{4.49}$$

where we take the derivative with respect to $-\mu$ since it is the source term used to keep track of the moments of the fractional occupation time (see Eqs. 4.32 and 4.39). The normalized moment-generating function gives the likelihood of m mutations conditioned on ψ_i, ψ_f, and τ:

$$P(m|\psi_i, \psi_f, \tau; \mu, \alpha_-, \alpha_+) = \frac{\mu^m}{m!} \frac{d^m}{d(-\mu)^m} \frac{\langle \psi_f |\, e^{\boldsymbol{M}\tau} \,| \psi_i \rangle}{P(\psi_f | \psi_i, \tau; \alpha_-, \alpha_+)}. \tag{4.50}$$

The likelihood of a tree with n external branches and k internal branches is the product of the likelihood of each branch

$$P(\vec{m}|\vec{\tau}; \mu, \alpha_-, \alpha_+) = \prod_{j=1}^{n} P(m_j | \text{on}, \text{off}, \tau_j; \mu, \alpha_-, \alpha_+) \prod_{l=1}^{k} P(m_l | \text{on}, \text{on}, \tau_l; \mu, \alpha_-, \alpha_+). \tag{4.51}$$

We constrain each internal branch to begin and end in the active state since we assume branching and mutations occur only in the active state. On the other hand, external branches start on since they were created from a branching (mutation) event and end off since most BCR repertoire studies sample BCRs from B cells in the peripheral blood. In most cases, it is unlikely for B cells in the peripheral blood to be active and form GCs. We note, however, that GC formation can occur outside of secondary lymphoid organs. B cells activated extrafollicularly have been observed during salmonella infection [67], and GCs can form in tertiary lymphoid structures, which are technically considered peripheral relative to the immune system, during chronic infections and cancer [260]. Of course, if we wanted to characterize a phylogeny from a study in which fine-needle aspiration (FNA) was used to sample B cells from GCs, as in ref. [149], the external branches associated with BCRs from those FNA samples would need to be constrained to end in the active state. Thus, while we assume the form of the likelihood above, our framework is flexible and amenable to BCRs sampled from GCs.

4.3.3 Closed-form solutions for the likelihoods

The previous section showcased that the likelihood of observing m mutations could be computed either by integrating out the fractional occupation time (Eq. 4.30) or by taking derivatives with respect to the mutation rate (Eq. 4.50). In this section, we show how to derive a closed-form for the likelihood of observing m mutations on a branch given the branch has a lifetime τ by deriving the distribution which characterizes the time between mutation events, the so-called distribution of waiting times. Crucially, we use our results from our derivation of the moment-generating function of x in the previous section. These closed-form solutions can be computed using efficient numerical routines and therefore should be preferred for studying the likelihood and its landscape, optimizing the likelihood to find the maximum likelihood estimates of the parameters, and computing the derivatives of the likelihood using autograd packages.

The likelihood for ancestral branches

We find a closed form of Eq. 4.40 by studying the distribution which characterizes the waiting time before the m-th mutation occurs. Notably, a lineage must be in the active state for a mutation to occur. The cumulative distribution function that characterizes the waiting time necessary to observe the first mutation when $\psi_i = \text{on}$ is given by the probability that at least one mutation has occurred within an interval τ:

$$F(\tau|\psi_i = \text{on}, 1) = P(m \geq 1|\psi_i = \text{on}, \tau) = 1 - P(0|\psi_i = \text{on}, \tau) = \tag{4.52}$$

$$1 - \langle \text{on}| \, e^{\boldsymbol{M}\tau} \, |\text{on}\rangle - \langle \text{off}| \, e^{\boldsymbol{M}\tau} \, |\text{on}\rangle \tag{4.53}$$

$$= 1 - \left(e^{-\gamma_- \tau} \left[\frac{d\gamma_-}{d\mu} + \frac{\alpha_-}{\eta} \right] + e^{-\gamma_+ \tau} \left[\frac{d\gamma_+}{d\mu} - \frac{\alpha_-}{\eta} \right] \right). \tag{4.54}$$

The coefficient of $e^{-\gamma_- \tau}$ is simplified algebraically using the relations in Eqs. 4.44 - 4.47:

$$\begin{aligned}
&\frac{d\gamma_-}{d\mu} + \frac{\alpha_-}{\eta} = \frac{d\gamma_-}{d\mu}\left(1 + \frac{\alpha_-}{\eta}\left(\frac{d\gamma_-}{d\mu}\right)^{-1}\right) = \frac{d\gamma_-}{d\mu}\left(1 + \frac{\alpha_-}{\eta}\frac{2\eta}{\eta-\rho}\right) \\
&= \frac{d\gamma_-}{d\mu}\left(1 + \frac{2\alpha_-}{\eta-\rho}\right) = \frac{d\gamma_-}{d\mu}\left(1 + \frac{2\alpha_-(\eta+\rho)}{\eta^2-\rho^2}\right) = \frac{d\gamma_-}{d\mu}\left(1 + \frac{2\alpha_-(\eta+\rho)}{4\alpha_-\alpha_+}\right) \\
&= \frac{d\gamma_-}{d\mu}\left(1 + \frac{(\eta+\rho)}{2\alpha_+}\right) = \frac{d\gamma_-}{d\mu}\left(\frac{2\alpha_+ + \eta + \rho}{2\alpha_+}\right) = \frac{d\gamma_-}{d\mu}\left(\frac{\gamma_+}{\alpha_+}\right) = \frac{d\gamma_-}{d\mu}\frac{\mu}{\gamma_-}
\end{aligned} \tag{4.55}$$

The coefficient of $e^{-\gamma_+ \tau}$ is obtained through similar algebraic relations and leads to

$$\frac{d\gamma_+}{d\mu} - \frac{\alpha_-}{\eta} = \frac{d\gamma_+}{d\mu}\frac{\mu}{\gamma_+}. \tag{4.56}$$

Thus, the coefficients of the exponentials are probabilities since the coefficients add to 1 and each coefficient is nonnegative:

$$\left(\frac{d\gamma_-}{d\mu} + \frac{\alpha_-}{\eta}\right) + \left(\frac{d\gamma_+}{d\mu} - \frac{\alpha_-}{\eta}\right) = 1, \quad \frac{d\gamma_\pm}{d\mu}\frac{\mu}{\gamma_\pm} \geq 0 \implies \frac{d\gamma_\pm}{d\mu}\frac{\mu}{\gamma_\pm} \in [0,1] \tag{4.57}$$

Now the cumulative distribution function appears as

$$F(\tau|\psi_i = \text{on}, 1) = 1 - \left(e^{-\gamma_- \tau}\frac{d\gamma_-}{d\mu}\frac{\mu}{\gamma_-} + e^{-\gamma_+ \tau}\frac{d\gamma_+}{d\mu}\frac{\mu}{\gamma_+}\right). \tag{4.58}$$

Because the term in parentheses of Eq. 4.58 has the form

$$e^{-\gamma_- \tau} q + e^{-\gamma_+ \tau}(1-q), \tag{4.59}$$

with $q \in [0,1]$ a probability, the cumulative distribution function (which uniquely determines the probability distribution of a random variable [43]) given in Eq. 4.58 can be identified as a mixture of exponential distributions, otherwise known as a hyperexponential distribution.

We next characterize the distribution which models how much time elapses before the m-th mutation is observed. Consider a random variable X which is characterized by the cumulative distribution function given by Eq. 4.58. $X \sim \text{Exp}(\gamma_-)$ with probability $q = \mu d_\mu \log(\gamma_-)$, and $X \sim \text{Exp}(\gamma_+)$ with probability $(1-q)$. Now consider $\sum_{i=1}^m X_i$, with X_i iid and distributed identically to the aforementioned X. Each X_i is either distributed as $\text{Exp}(\gamma_-)$ with probability q or distributed according to $\text{Exp}(\gamma_+)$ with probability $(1-q)$. Let

$0 \leq k \leq m$ be the number of times an X_i is distributed as $\text{Exp}(\gamma_-)$. Then k is binomially distributed: $k \sim \text{Binom}(m, q)$. $\sum_{i=1}^{m} X_i$ can be rewritten to appear as

$$\sum_{i=1}^{m} X_i = \sum_{i=1}^{k} X_{i,-} + \sum_{j=k}^{m} X_{j,+}, \tag{4.60}$$

where $X_{i,-} \sim \text{Exp}(\gamma_-)$ and $X_{j,+} \sim \text{Exp}(\gamma_+)$. The k-fold convolution of an exponential distribution with rate γ_- is an Erlang distribution with shape k and rate γ_-

$$f(t|k;\gamma_-) = \frac{\gamma_-^k\, t^{k-1}}{(k-1)!} e^{-\gamma_- t}. \tag{4.61}$$

Similarly, the $(m-k)$-fold convolution of an exponential distribution with rate γ_+ is an Erlang distribution with shape $m-k$ and rate γ_+. Altogether, $\sum_{i=1}^{m} X_i$ is the sum of two Erlang-distributed random variables whose rates are γ_- and γ_+ and whose shapes are k and $m-k$, respectively, where k is binomially distributed with probability $q = \mu d_\mu \log(\gamma_-)$ and number of trials m. The probability density function which models the waiting time that elapses before the m-th mutation occurs is therefore obtained by using a convolution of two Erlang distributions [323] with binomially distributed shape parameters and marginalizing over the shapes [348]:

$$\begin{aligned}
&P(\tau|\psi_i = \text{on}, m) = \\
&\qquad \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d\gamma_-}{d\mu}\frac{\mu}{\gamma_-}\right)^k \left(\frac{d\gamma_+}{d\mu}\frac{\mu}{\gamma_+}\right)^{m-k} \int_0^\tau du\, f(u|k;\gamma_-) f(\tau - u|m-k;\gamma_+) \\
&= \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d\gamma_-}{d\mu}\frac{\mu}{\gamma_-}\right)^k \left(\frac{d\gamma_+}{d\mu}\frac{\mu}{\gamma_+}\right)^{m-k} \int_0^\tau du\, \frac{\gamma_-^k\, u^{k-1}}{(k-1)!} e^{-\gamma_- u} \frac{\gamma_-^{m-k}\,(\tau-u)^{m-k-1}}{(m-k-1)!} e^{-\gamma_+(\tau-u)} \\
&\qquad = \frac{e^{-\gamma_-\tau}\tau^{m-1}\mu^m}{(m-1)!} \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d\gamma_-}{d\mu}\right)^k \left(\frac{d\gamma_+}{d\mu}\right)^{m-k} {}_1F_1(m-k, m, -\eta\tau),
\end{aligned} \tag{4.62}$$

where $m \geq 1$, consistent with the definition of the waiting time distributions, and ${}_1F_1$ is the confluent hypergeometric function. Since mutations can occur at any time when a lineage is in an on state, we naturally have $\psi_f = \text{on}$. The function describing the joint probability of m and $\psi_f = \text{on}$ given $\psi_i = \text{on}$ is determined by taking the difference of consecutive terms in

this waiting time's cumulative distribution function [125].

$$P(m, \text{on}|\text{on}, \tau) = F(\tau|\text{on}, m) - F(\tau|\text{on}, m+1). \tag{4.63}$$

Thus, another form for Eq. 4.40 is

$$P(m, \psi_f = \text{on}|\psi_i = \text{on}, \tau) = \frac{(\tau\mu)^m e^{-\gamma_- \tau}}{m!} \sum_{k=0}^{m+1} \binom{m+1}{k} \left(\frac{d\gamma_-}{d\mu}\right)^k \left(\frac{d\gamma_+}{d\mu}\right)^{m+1-k} {}_1F_1(m+1-k, m+1, -\eta\tau). \tag{4.64}$$

We examine our solutions in the Poisson regime where $\alpha_+ \gg \mu \gg \alpha_-$, i.e., lineages are always active. This gives

$$P(\psi_f = \text{on}|\psi_i = \text{on}, \tau) \approx 1, \quad \gamma_- \approx \mu, \quad \gamma_+ \approx \alpha_+, \quad \frac{d\gamma_-}{d\mu} \approx 1, \quad \frac{d\gamma_+}{d\mu} \approx 0. \tag{4.65}$$

The likelihood of m mutations conditioned on all other random variates becomes (substituting Eq. 4.40 into Eq. 4.50)

$$P(m|\text{on}, \text{on}, \tau) \approx \frac{\mu^m}{m!} \frac{d^m}{d(-\mu)^m}\left(e^{-\mu\tau}\right) = \frac{(\mu\tau)^m}{m!} e^{-\mu\tau}. \tag{4.66}$$

This is also apparent from Eq. 4.64 in which the ${}_1F_1(\cdot)$ terms are negligible except when $k = m+1$ at which ${}_1F_1(0, \cdot, \cdot) = 1$ and $\left(\frac{d\gamma_-}{d\mu}\right)^{m+1} \approx 1$. Additionally, we see that $P(\tau|\psi_i = \text{on}, m+1)$ (Eq. 4.62) converges to an Erlang distribution in the Poisson limit as expected in the same regime of parameter space.

The likelihood for external branches

The cumulative distribution function of the waiting time until the first mutation when $\psi_i = \text{off}$ is computed using the probability of observing at least one mutation over the time interval τ when $\psi_i = \text{off}$:

$$F(\tau|\psi_i = \text{off}, 1) = P(m \geq 1|\psi_i = \text{off}, \tau) = 1 - P(0|\psi_i = \text{off}, \tau) = \tag{4.67}$$

$$1 - \langle \text{on}| \, e^{\boldsymbol{M}\tau} \, |\text{off}\rangle + \langle \text{off}| \, e^{\boldsymbol{M}\tau} \, |\text{off}\rangle \tag{4.68}$$

$$= 1 - \left(e^{-\gamma_- \tau} \left[\frac{d\gamma_+}{d\mu} + \frac{\alpha_+}{\eta} \right] - e^{-\gamma_+ \tau} \left[\frac{d\gamma_-}{d\mu} - \frac{\alpha_+}{\eta} \right] \right). \tag{4.69}$$

The coefficient of $e^{-\gamma_- \tau}$ is

$$\frac{d\gamma_+}{d\mu} + \frac{\alpha_+}{\eta} = \frac{1}{2}\left(1 + \frac{\rho}{\eta}\right) + \frac{\alpha_+}{\eta} \tag{4.70}$$

$$= \frac{\eta + \rho + 2\alpha_+}{2\eta} \tag{4.71}$$

$$= \frac{\gamma_+}{\eta}. \tag{4.72}$$

A similar procedure shows that the coefficient of $e^{-\gamma_+ \tau}$ is γ_-/η. The cumulative distribution function now appears as

$$F(\tau|\psi_i = \text{off}, 1) = 1 - \left(e^{-\gamma_- \tau}\frac{\gamma_+}{\eta} - e^{-\gamma_+ \tau}\frac{\gamma_-}{\eta}\right), \tag{4.73}$$

which we identify as the cumulative distribution function of the hypoexponential distribution. The hypoexponential distribution is a generalized Erlang distribution: if a random variable X is characterized by the cumulative distribution function given by Eq. 4.73, it is generated as $X = Y + Z$, where $Y \sim \text{Exp}(\gamma_-)$ and $Z \sim \text{Exp}(\gamma_+)$. By time-reversal symmetry, we argue that a telegraph process that started off, telegraphs into the on state, and eventually mutates is the reverse of the telegraph process which starts in the on state (perhaps from a mutation event), telegraphs off, and ends off. Thus, the time characterizing both of these processes should be identical.

$$F(\tau|\psi_f = \text{off}, \psi_i = \text{on}, 0) = F(\tau|\psi_i = \text{off}, 1). \tag{4.74}$$

Finally, $F(\tau|\psi_i = \text{off}, 1)$ is the cumulative distribution function which models the time it takes for a telegraph process to start from the on state and end in the off state without any mutations occurring.

The time of an external branch with m mutations is given as the waiting time for m mutations (i.e., the time which characterizes an internal branch that starts and ends on and has m mutations) with the addition of the time necessary for the branch to ultimately end off without any additional mutations. Therefore, for general m, an external branch's time is given by the random variables $X + Y$, where X is distributed as the binomial + convolved

Erlang obtained in the previous section (Eq. 4.62), and Y is hypoexponentially distributed with rates γ_- and γ_+. The probability density function that characterizes the time of an external branch is obtained by convolving the distribution in Eq. 4.62 with a hypoexpnoential distribution with rates γ_-, γ_+.

$$P(\tau|\psi_i = \text{on}, \psi_i = \text{off}, m) = \frac{\alpha_+(\mu t)^{m+1}e^{-\gamma_- \tau}}{(m+1)!} \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d\gamma_-}{d\mu}\right)^{m-k} \left(\frac{d\gamma_+}{d\mu}\right)^{k} {}_1F_1(k+1, m+2, -\eta\tau) \quad (4.75)$$

Notably, $m \geq 0$ here. Taking the difference between consecutive terms of the cumulative distribution function for this probability density function gives $P(m, \psi_f = \text{on}|\psi_i = \text{off}, \tau)$.

$$P(m, \psi_f = \text{off}|\psi_i = \text{on}, \tau) = \frac{\alpha_-\mu^m t^{m+1}e^{-\gamma_- \tau}}{(m+1)!} \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d\gamma_-}{d\mu}\right)^{m-k} \left(\frac{d\gamma_+}{d\mu}\right)^{k} {}_1F_1(k+1, m+2, -\eta\tau) \quad (4.76)$$

4.4 *Telegraph model resolves switching and mutation parameters for star phylogenies*

We validated the mutation-telegraph likelihood of a tree, $P(\vec{m}|\vec{\tau}; \mu, \alpha_-, \alpha_+)$ (Eq. 4.51) by performing maximum likelihood estimation over a range of realistic α_- and α_+ on simulated star trees. A star tree is a phylogeny in which all lineages are terminal and have a common root. To simulate such a phylogeny for a given α_- and α_+, we sampled the branch lengths of the lineages with a uniform distribution, $t \sim U(0, 1/\alpha_- + 1/\alpha_+)$, where $1/\alpha_- + 1/\alpha_+$ is the average time of one full cycle. We used the Gillespie algorithm [93] to simulate the occupation time along each branch and sampled each at their respective time t. We constructed the star trees by sampling 100 branches that started on and ended off and 98 branches that started on and ended on, congruent with what we expect from bifurcating phylogenies. In other words, given n external branches on a bifurcating tree, there will be $n-2$ internal branches. Only phylogenies which had at most 30 mutations on a branch were used as inputs for likelihood inference downstream. We inferred the parameters by performing minimization

on the negative log likelihood of the tree, with the likelihood given by Eq. 4.51, and used the simplicial homology global optimization routine (SHGO) [78]. This method is a deterministic algorithm that efficiently samples the parameter space and has been shown to have excellent convergence properties.

The mutation-telegraph model performs excellently on the simulated star phylogeny data (Fig. 4.2A-C). For each group of $(\mu, \alpha_-, \alpha_+)$, we computed the relative bias and the relative root mean square error (RMSE):

$$\text{rel. bias} = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \frac{\hat{\theta}_i - \theta_{\text{true}}}{\theta_{\text{true}}}, \qquad \text{rel. RMSE} = \sqrt{\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \left(\frac{\hat{\theta}_i - \theta_{\text{true}}}{\theta_{\text{true}}} \right)^2}, \tag{4.77}$$

where $\hat{\theta}$ is the maximum likelihood estimate, θ_{true} is the input parameter used for the simulations, and N_{sim} is the number of simulations. We observe no bias in the inference of the mutation rate and a small positive relative bias for the switching rates (Fig. C.3A). The inference of μ has the smallest relative RMSE and the switching rates have higher relative RMSE, with α_+ having the highest relative RMSE (Fig. C.3B). As expected, we observe positive correlations between the inferred μ and α_- and between the inferred α_- and α_+ (Fig. 4.2D,F). On the other hand, we observe a small positive correlation between μ and α_+ whereas we expected these two values to be correlated negatively (Fig. 4.2E). That being stated, we do not observe here any identifiability issues with the model.

4.5 Inferring the mutation-telegraph model on a phylogeny of an evolving population

Star trees are a very simple example of a phylogeny and are not representative of the phylogenies expected to be consistent with those characterizing B cell lineages. In other words, the sequences of a B cell lineage do not share a common root. In fact, B cell phylogenies present a lot of peculiarities for phylogenetic inference since B cells are rapidly evolving. For example, multifurcating trees were shown to be more consistent with B cell phylogenies using tree topologies generated by a simulation of multi-round BCR evolution. In comparison, most phylogenetics methods assume a bifurcating tree in which every parent node has

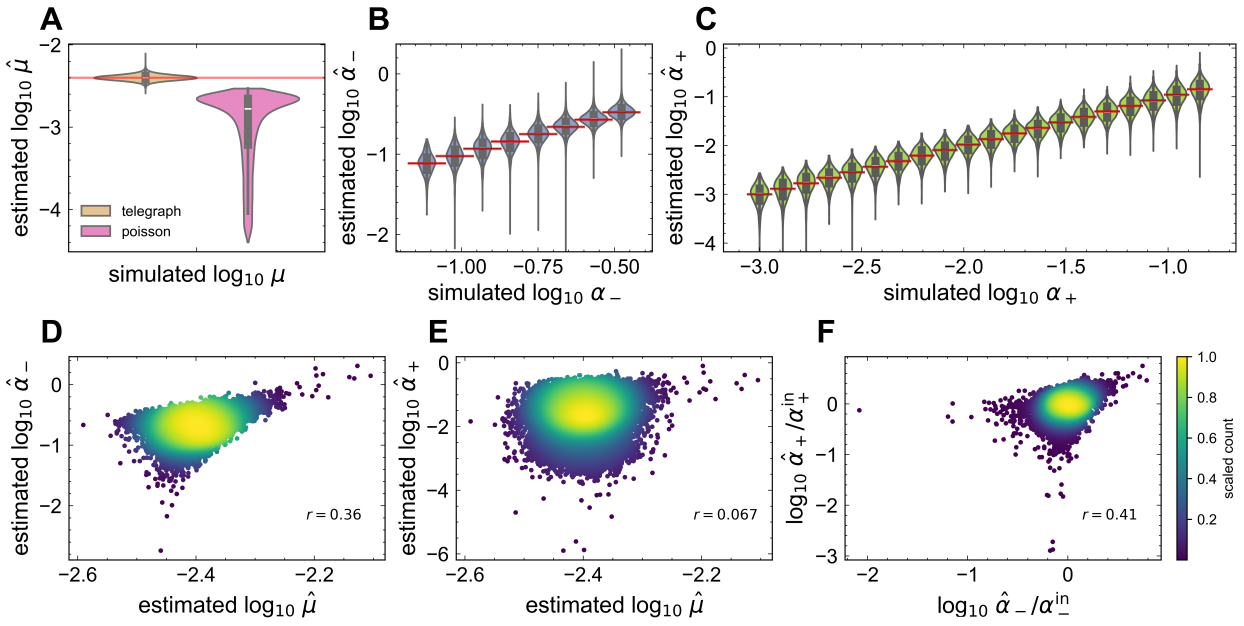


Figure 4.2: **Telegraph-mutation model resolves true rates in simulated data. (A)** The violin plot shows the distribution of maximum likelihood estimates of μ using the mutation-telegraph model (orange) and the Poisson model (pink). The red horizontal line demarcates the true rate. **(B)** The violin plots shows the distribution of maximum likelihood estimates of α_- for each simulated α_-. The red horizontal lines demarcate the true rate to which each violin should be compared. **(C)** The violin plots shows the distribution of maximum likelihood estimates of α_- for each simulated α_-. The red horizontal lines demarcate the true rate to which each violin should be compared. **(D)** The estimates of $\log_{10}\alpha_-$ versus $\log_{10}\mu$ are plotted as a scatter density. The Pearson correlation is shown in the bottom right. **(E)** The estimates of $\log_{10}\alpha_+$ versus $\log_{10}\mu$ are plotted as a scatter density. The Pearson correlation is shown in the bottom right **(F)** The estimates of the normalized $\log_{10}\alpha_-$ versus $\log_{10}\alpha_+$ are plotted as a scatter density. The Pearson correlation is shown in the bottom right.

only two child nodes. (See refs. [118, 340] for reviews of idiosyncrasies associated with B cell phylogenetics.) In practice, phylogenies inferred from data contain only the following information obtain via inference: the topology of the tree and the expected number of mutations on each branch. The sampling date information for each tip is also known, and we assume this is fixed by the experiment.[1] Most phylogenetic inference concerns only reconstructing the tree topology and investigating evolutionary relationships among the observed data of sequences [338], so the dates of the internal nodes are not characterized.

Our mutation-telegraph tree likelihood crucially relies on information about all the node dates, and therefore time on a branch (Eq. 4.51), since we want to characterize the effective time scales of the mutation process and phenotypic switching in B cell lineages. Comparatively, root-to-tip regression circumvents the need for the internal branch times since it assumes mutations are characterized by a Poisson distribution. The Poisson distribution is convolution-closed, i.e., the sum of two Poisson random variables is another Poisson variable, so the distribution of a mutation observed on a single branch is in the same statistical family as the sum of mutations observed on many branches [81]. On the other hand, the mutation-telegraph distribution is not convolution-closed, so we must have each branch's length in units of time. Therefore, we need a method which enables us to situate the internal nodes in time. Moreover, due to the size of trees expected from BCR repertoire data ($\sim 100 - 1000$) and the number of measurably evolved lineages that can be observed (10s to 100s) (Fig. C.1, 4.1), we desired a statistical method whose computational demand would permit us to analyze BCR repertoire data at scale.

[1]The sampling date could be treated as a noisy estimate in principle. The likelihood of the sampling date, for instance, might be modeled by a Gaussian distribution with some pre-specified width and mean given by the observed sampling date [217].

4.5.1 TreeTime: phylogenetic inference of the node dates, evolutionary parameters, and ancestral sequences

We solve the problem of situating the internal nodes in time by using an algorithm called TreeTime. TreeTime is an approximate maximum likelihood optimization algorithm that infers time-annotated trees and the evolutionary rates [256]. TreeTime takes as input the tree topology, the sequences of the tips, and the sample dates of the tips. After preprocessing, such as rerooting the tree, resolving polytomies, and pruning short branches, TreeTime proceeds by inferring the ancestral sequences using a message-passing routine [241] to globally optimize the likelihood of the substitution model. TreeTime then estimates the substitution rate in units of inverse time using the provided sampling times of the tips and root-to-tip regression. Finally, TreeTime infers the times of the ancestral nodes using a message-passing algorithm similarly to how it infers ancestral sequences. While TreeTime can perform joint or marginal inference of the times, we recount here only the algorithm for inferring the times of the nodes by marginalizing over all other messages since it is more stable and what we will ultimately use.

A message-passing routine situates internal nodes in time

The algorithm begins by computing in post-order the likelihood $H_n(t)$ that node n sits at time t

$$H_n(t|\mathcal{C}_n) = E_n(t) \prod_{c \in \mathcal{C}_n} C_c(t), \tag{4.78}$$

where $\mathcal{C}_n$ is the set of node n's children, and $E_n(t)$ is a constraint about n's date information irrespective of its children. $C_n(t)$ is the information propagated from node n to its parent p. It is defined as the convolution of the probability $b_n(\tau)$ of the branch length between n and p and the probability distribution H_n of n's date:

$$C_n(t_p) = \int_0^\infty d\tau \, b_n(\tau) H_n(t_p - \tau|\mathcal{C}_n). \tag{4.79}$$

In TreeTime, $b_n(\tau)$ is the Erlang distribution. When the post-order computation reaches the root, r, its likelihood is calculated by

$$P_r(t) = \frac{E_r(t)}{Z_r} \prod_{c \in \mathcal{C}_r} C_c(t), \tag{4.80}$$

where Z_r is a normalization factor. Traversing the tree now in pre-order, the marginal likelihood node that n is positioned at t is

$$P_n(t) = \frac{1}{Z_n} H_n(t|\mathcal{C}_n) \int_0^\infty d\tau \, b(\tau) \frac{P_p(t+\tau)}{C_n(t+\tau)}, \tag{4.81}$$

where $H_n(t|\mathcal{C}_n)$ contributes information propagated from the leaves of node n, and the integral gives information about n's position from subtrees excluding n and its children.

A substitution model for telegraph processes

TreeTime's algorithm for inferring the substitution rate and internal node dates requires the inference of the ancestral sequence. While the tree topology gives the branch length as the expected number of mutations on that branch, TreeTime annotates the nodes with inferred sequences. This provides an integer-value used for specifying the branch length distribution (see Eqs. 4.61, 4.62, and 4.75) when inferring the dates of the internal nodes. Additionally, inferring the ancestral sequences, as opposed to using the expected number of mutations on a branch, can be used to study how perturbations at the sequence level affect the node date and parameter inference downstream. Substitution models used by TreeTime assume a molecular clock hypoythesis. Therefore, we must construct a new substitution model consistent with the mutation-telegraph model.

Because we are chiefly interested in characterizing the time and the effective rates of phenotype switching rather than the identity of nucleotide substitutions and substitution model parameters, we construct a version of the Jukes-Cantor model [143], the most parsimonious substitution model, amenable to telegraph processes and use it for inferring ancestral sequences. The Jukes-Cantor model describes substitutions made along the genome and assumes that the stationary frequencies of bases are equal and that mutation rates from one

character to another are homogenous. Let $P(m=0|t;\mu)$ be the probability of no mutations occurring over a time t. The Jukes-Cantor model assumes $P(m=0|t;\mu)=e^{-\mu t}$ is Poissonian [143]. Then, given an alphabet $\mathcal{A}$ of size $a=|\mathcal{A}|$ (e.g., the DNA alphabet has size 4), the probability of character i mutating into character j given the time along a branch t and mutation rate μ is

$$P(j|i,t;\mu,a)=\begin{cases}\frac{1}{a}+\frac{a-1}{a}e^{-t\mu}, & i=j\\ \frac{1}{a}-\frac{1}{a}e^{-t\mu}, & i\neq j.\end{cases} \tag{4.82}$$

Assuming a telegraph process, the time over which substitutions can occur is shortened, and the probabilities change accordingly. Given the boundary constraints imposed by phylogenies, the relevant probabilities, i.e., Eqs. 4.40 and 4.42 normalized by their respective propagators, are:

$$P(m=0|\text{on},\text{on},t;\mu,\alpha_-,\alpha_+)=\frac{1}{P(\text{on}|\text{on},t;\alpha_-,\alpha_+)}\left(e^{-\gamma_-\tau}\frac{d\gamma_-}{d\mu}+e^{-\gamma_+\tau}\frac{d\gamma_+}{d\mu}\right) \tag{4.83}$$

and

$$P(m=0|\text{on},\text{on},t;\mu,\alpha_-,\alpha_+)=\frac{1}{P(\text{off}|\text{on},t;\alpha_-,\alpha_+)}\frac{\alpha_+}{\eta}\left(e^{-\gamma_-\tau}(1-e^{-\eta\tau})\right). \tag{4.84}$$

Therefore, the Jukes-Cantor model adopted for a telegraph process is

$$P(j|i,t,\psi_i,\psi_f;\mu,\alpha_-,\alpha_+,a)=\begin{cases}\frac{1}{a}+\frac{a-1}{a}P(m=0|\psi_i,\psi_f,t;\mu,\alpha_-,\alpha_+) & i=j\\ \frac{1}{a}-\frac{1}{a}P(m=0|\psi_i,\psi_f,t;\mu,\alpha_-,\alpha_+), & i\neq j.\end{cases} \tag{4.85}$$

4.5.2 Performance of the mutation-telegraph model on more realistic phylogenies

We adapted TreeTime to use our mutation-telegraph model in the following ways. The inference is initialized by running one round of TreeTime with the Poisson model in order to get estimates of the branch times. The algorithm then uses these times to infer the mutation and switching parameters using SHGO [78], as we did for the star phylogeny analysis. Using these rates, the branch lengths are inferred using the mutation-telegraph inter-arrival time

distributions and the message-passing algorithm, which yields the marginal likelihoods of the nodes' dates. Finally, the ancestral sequences are re-inferred using the telegraph JC model. The algorithm then re-infers the mutation-telegraph rates and continues through this sequence until convergence.

Validating the performance of our adaptation of TreeTime requires an oracle which produces ground truth phylogenies that resembles those we anticipate analyzing, i.e., phylogenies reconstructed using BCR lineages. However, this is a nontrivial problem due to the niche BCRs occupy in the field of phylogenetics [118]. While simulation frameworks have been produced to generate phylogenies from BCR lineages undergoing a single round [58] and multiple rounds of evolution [346], we instead generated ground truth data by simulating times on tree topologies from the HIV cohort which were detected as measurably evolving. The performance on this ground truth would be much closer to that we'd expect on our data since the two are highly similar. We simulated ancestral sequences on the trees using RAxML-NG [154] and the GY94 substitution model [97]. We then placed times on the branches using our waiting time distributions (Eqs. 4.62, 4.75, Appendix C).

B cell topologies given as input to our adapted TreeTime undergo no preprocessing. The topologies produced by IgPhyML [122] are rooted using the germline sequence as the outgroup, so rerooting the tree would be inconsistent with the evolution of the B cell lineage. Further, we do not resolve polytomies because doing so introduces virtual branches with 0 mutations which, unlike root-to-tip regression which is agnostic to virtual branches, would introduce biases into the inference of the rates.

The performance of the mutation-telegraph adaptation of TreeTime is shown in Fig. 4.3. In stark contrast to the results shown in Fig. 4.2, the estimated $\hat{\mu}$ has a large spread and is negatively biased, being roughly one order of magnitude smaller than the true μ across all input parameters and tree topologies (Fig. 4.3A). α_- is estimated better when its true value is smaller, and a negative bias grows as the rate becomes larger (Fig. 4.3B). Remarkably, the slowest timescale given by α_+ appears to be consistently estimated well across different (α_-, α_+) (Fig. 4.3C). The distribution of the estimated mutation rates from the mutation-

telegraph adaptation of TreeTime is rightward shifted compared to the estimates obtained from the Poisson version of TreeTime. We compared the estimated branch lengths in units of time output by the mutation-telegraph TreeTime to the simulated branch lengths in units of time and observed a correlation of 0.78 (Fig C.4).

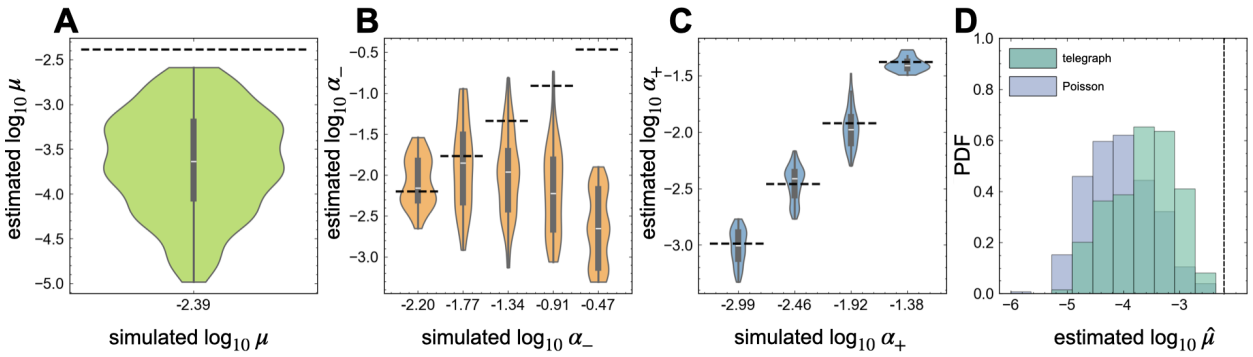


Figure 4.3: **Performance of TreeTime using the mutation-telegraph model. (A)** The violin plot shows the distribution of maximum likelihood estimates of μ using the mutation-telegraph model. The horizontal black, dashed line indicates the value of the true rate. μ is measured in units of days^{-1} site^{-1}. **(B)** The violin plots shows the distribution of maximum likelihood estimates of α_- for each simulated α_-. The horizontal black, dashed line indicates the value of the true rates. α_-^{-1} is measured in units of days. **(C)** The violin plots shows the distribution of maximum likelihood estimates of α_+ for each simulated α_+. The horizontal black, dashed line indicates the value of the true rates. α_+^{-1} is measured in units of days. **(D)** The histograms shows the distribution of estimated $\log_{10}\hat{\mu}$ using the mutation-telegraph adaptation of Treetime (green) and the original, Poissonian Treetime (lavender). The vertical black, dashed line indicates the value of the simulated μ.

TreeTime's approach separates the inference of the node dates and the parameters of the branch length distribution. This approximation infers the substitution rate accurately since root-to-tip regression relies on the sampling times, and each root-to-tip divergence is calculated by summing over the expected amount of mutations on the branches encountered

when traversing from a tip to the root. This summation produces more robust inference (as opposed to using each individual time-annotated branch length) since noisy estimates of the times are effectively averaged out. Moreover, using Erlang distributions, which are consistent with root-to-tip regression using a Poisson model, to infer the lengths of branches in time results in fairly stable estimates since Erlang distributions do not have long tails. All these factors contribute to the factorization and approximations working well. On the other hand, our adapted version uses each branch's length in time for inferring the parameters μ, α_-, and α_+. With the branch lengths themselves already being noisy estimates, these errors are propagated when inferring the parameters, whose likelihood landscape is much more jagged and rough than a likelihood associated with the Poisson distribution, and thus root-to-tip regression. Moreover, the mutation-telegraph waiting time distributions have much longer tails and greater widths than the Erlang distribution (Fig. C.5A). This can be seen in part by the variance statistic of the waiting time distributions (Eq. C.14) (Fig. C.5B). For instance, when a branch starts and ends in the active state, the variance of the inter-arrival time is the variance of an Erlang distribution ($[m+1]/\mu^2$, with $m \geq 0$) scaled by a factor of $[(\alpha_- + \alpha_+)^2 + 2\mu\alpha_-]/\alpha_+^2$. When α_+ is small compared to μ and α_-, this can be approximated as $[\alpha_- \times (\alpha_- + \mu)]/\alpha_+^2$. This expression can be significantly greater than 1 when the denominator is much smaller than the numerator, as expected for the regime of parameters in which B cell phenotypic switching takes place ($\mu \gg \alpha_- \gg \alpha_+$) (Fig. C.5B). We attempted to mitigate uncertainty in branch length estimates using message-passing which marginalizes messages. While that stabilizes the inference some (as opposed to using joint message-passing [256]), it alone cannot resolve the uncertainty inherent to the mutation-telegraph waiting time distributions. Thus, estimating node dates with the mutation-telegraph model is substantially more difficult. Inferring these μ, α_-, α_+ in a factorized, approximate manner ensues in biases and poor inference is incurred.

4.6 Discussion and outlook

B cell lineages evolve intermittently and estimating their phenotypic switching rates is paramount to gaining insight into the role of memory B cells in secondary infections. These rates have yet to measured *in vivo*, and the conditions under which it does occur continues to elude researchers. Being able to detect statistically if lineages were recalled would open a new avenue for longitudinal B cell repertoire analyses to identify functionally responsive lineages against recurring immune challenges like vaccinations, influenza, and SARS-CoV-2 or chronic challenges like HIV. It would also enable researchers to use datasets of incredibly large depth, increasing the possibility of detecting lineages which experienced recall, and couple their findings with other statistics derived from studying repertoires, such as using proabability P_{gen} of generating a receptor [184] to quantify a recall versus generation event.

Endeavoring to create this statistical framework for B cell recall, we derived an exact two-state model—the mutation-telegraph model—that infers the mutation rate, deactivation rate, and activation rate without having any knowledge of the lineages' occupation times spent in the activated and deactivated states. We demonstrated the model's likelihood had excellent performance in recovering the rates when the lineages' branch times were known using star phylogeny simulations. However, in practice, the dates of ancestral phylogenies are unknown and so our likelihood alone cannot recapture the underlying parameters of the system. B cell repertoire analyses can ensue in hundreds or thousands of measurably evolving phylogenies, so we adapted a previously developed approximate maximum likelihood approach, TreeTime, to infer the dates of a phylogeny's ancestral nodes and the parameters of the mutation-telegraph model. We evaluated its performance by simulating node dates consistent with the mutation-telegraph model on B cell topologies. Because the the inference of the nodes' dates and the rates are inextricably coupled, the approximate inference algorithm introduces biases into the inference of the rates and the nodes' dates.

One simple avenue forward would be to incorporate prior information and use a penalized likelihood approach. For instance, to break the degeneracy between the mutation rate and

the deactivation rate, we could place a prior on the mutation rate which has been measured previously [151] using a lognormal distribution. However, inference of the model rates would still be performed disjointly from the inference of the node dates, which is statistically inconsistent.

Though inferring the dates of the ancestral nodes of the phylogeny would yield more information about the lineage's evolution, especially in the context of studying co-evolving systems, we could treat the ancestral node dates as nuisance parameters that are marginalized out to improve estimates of the mutation-telegraph rates. In order to do this, TreeTime's message-passing algorithm could be used along with the mutation-telegraph inter-arrival time distributions to obtain the marginal likelihood of the root node's date. Integrating over this likelihood would give the the marginal likelihood of the mutation-telegraph rates and could be used as the cost for a global optimization routine. The difficulty here is the instability in numerical precision that must be handled with care in order to obtain the marginal likelihood of the root. Because TreeTime uses its message-passing algorithm for situating the ancestral nodes in time, it only requires computing the likelihoods of the nodes' dates up to a constant. This allows the algorithm to negate incorporating normalization by the Poisson clock rate or mutation-telegraph rates as well as renormalize the likelihoods to be on similar scales to avoid numerical instabilities when multiplying distributions or convolving distributions. However, for a global optimization routine to use the aforementioned loss, likelihoods must be normalized appropriately for values of the loss to be compared accurately when traversing the loss landscape. Thus, these numerical instabilities must be handled appropriately.

Bayesian phylogenetic methods could also be used at the expense of time. With the release of the BEAST SeedbankTree package [42], a Bayesian treatment can be given to the mutation-telegraph model in which the model parameters and node dates are inferred jointly. The seedbank coalescent is a generalization of the mutation-telegraph model. Its model structure is therefore consistent with what we would expect from B cell lineages. Ref. [42] validated the SeedbankTree package using simulations where at least some of the tips were still active, the mutation rate of the dormant population was nonzero, and the

deactivation rate was smaller than the activation rate. On the other hand, we assume that the tips of B cell phylogenies occupy only the dormant state, the mutation rate of the dormant state is zero, and the deactivation rate should be greater than the activation rate. Thus, the SeedbankTree package's performance in the regime of B cell phylogenies may be less accurate than what the authors gathered from their simulations, and it is not yet clear how stable the SeedbankTree MCMC when the dormant mutation rate is zero. We would also need to require that BEAST use a fixed tree topology since our input topology would be generated from software such as IgPhyML [122] which roots the tree using the germline sequence as the outgroup. Preventing the exploration of tree space would speed up the inference as well as potentially improve the inference of the switching parameters, as noted by ref. [42].

Inferring a time-annotated tree and the rates associated with the branch length distributions is a difficult problem. Ideally Bayesian phylogenetic methods are used to jointly infer the dates of the nodes as well as the associated model parameters. Though we were unsuccessful in inferring time-annotated trees using an approximate maximum likelihood method, we have laid the groundwork for inference of the phenotypic rates for B cell phylogenies. Nevertheless, with the method developed here, aforementioned forward directions, and seedbank models developed elsewhere, we believe that the statistical detection of recalled lineages as well as the measurement of phenotypic switching rates is within reach.

Chapter 5

CONCLUSIONS AND OUTLOOK

Immune repertoires contain a vast amount of information about one's history of immune challenges and the unique responses mounted toward those challenges. While high-throughput sequencing experiments have enabled the collection of an unprecedented number of receptors from individuals, these samples are yet only a meager window into the actual diversity of an individual's repertoire, which, moreover, is microscopic relative to the potential diversity in the universe of adaptive immune receptors. To make sense of this data in the absence of a complete mapping which characterizes the cognate antigens of adaptive immune receptors, we developed and leveraged quantitative theoretical and computational tools to identify receptor sequences as candidates for functional responses to immune challenges.

In chapter 2, we characterized the B cell response to COVID-19 differentiated by severity of the disease, and, in chapter 3, we characterized the T cell response in individuals who had only COVID-19 and those who went on to present symptoms of PASC. We used the unproductive compartment of the repertoires to understand the features of V(D)J recombination, yielding the probability of generation P_{gen}. We then used the productive compartment to learn how selection modulates the statistics of the generation process, giving the probability P_{post} of observing a receptor sequence in the periphery. We used these models to characterize the importance of features of receptors for distinguishing healthy individuals from individuals with COVID-19 as well as setting null expectations for receptor sequences incident in many individuals. These models can also be used to detect functional responses in single repertoire snapshots by characterizing the expected size of receptor sequence neighborhoods. Though an algorithm using P_{gen} was developed [234] and deep P_{post} models have been shown to model neighborhoods in sequence space more accurately than P_{gen} or linear P_{post} mod-

els [131], a thoroughly studied refinement using P_{post} models has yet to be made. Further, while [206, 184] developed principled models for characterizing the generation of receptor sequences, some mechanistic features are yet to be modeled correctly. For example, current P_{gen} models underestimate ligation scenarios due to microhomology [255] and overestimate the diversity of nucleotides which can be inserted at the gene junctions. In both cases, P_{gen} is underestimated and requires corrections, such as an ad hoc q which calibrates P_{gen} and P_{post} based on fitting discrepancies in observed sequence degeneracy between the data and model [254]. Incorporating these findings into P_{gen} models will ameliorate all P_{gen} and P_{post} downstream analyses such as annotating clonal families [282] or the receptor sequences themselves, characterizing the publicness of receptors among individuals or the expansion of neighborhoods of receptors within an individual, and increase the sensitivity and specificity for detecting functional responses as we have performed in this thesis.

Receptors similar in sequence have been shown to be similar in their specificity and function [57, 95], leading to the development of TCRdist [57, 187] which we have used to a great extent in chapter 3 for characterizing the enrichment of sequence motifs and clustering TCRs. TCRdist is a distance metric which penalizes substitutions using the BLOSUM62 matrix learned from the general protein universe. Distances learned from the niche of TCRs [242, 238] and which are also site-dependent [208] may improve identifying relations between TCRs with similar specificity. As the amount of data on TCR specificities increases, it would be interesting to see the returns gained from context-dependent distance metrics or contrastive learning using a unified structural and language model. Similarly, BCRs are typically related to each other using simpler metrics such as Hamming distances, as we have done in chapter 2, or Levenshtein distances [3]. Developing a sequence distance metric associated with BCR specificity and which takes into account both sequence and isotype information (which shapes BCR affinity and cross-reactivity) would be useful for similar analyses.

We also studied the dynamics of BCR clonal lineages and TCR functional sequences in chapters 2 and 3, the characterization of which is limited by the undersampling of the

repertoires. Here, replicates quantifying the biological and technical noise are crucial for understanding the fluctuations in mRNA expression or cell counts. Though assumed elsewhere [214], the observation of clones in multiple replicates alone (given that clones are "rare") is neither necessary nor sufficient for a clone to have expanded. When studying repertoires with the intent of tracking clones over time, replicates must be incorporated into experimental designs and used for analyses.

We narrowed the search for functional receptors from the repertoires of our cohorts consisting of $10^7 - 10^8$ receptors to $10^1 - 10^3$ in chapters 2 and 3. These receptors warrant experimental verification to understand their role in their respective immune challenges. A caveat in both studies, and most repertoire studies, is that we only had access to the more diverse chain of the receptor repertoires—the heavy chain for B cells and the β chain for T cells. To identify a receptor completely and synthesize it for testing against antigens, both the heavy and light or the β and α chains must be known. Recently, an affordable method for deeply sequencing paired-chain repertoires has been developed [232]. Going forward, paired-chain sequencing will be paramount to understanding the adaptive immune system, solving problems such as deciphering the mapping between immune receptors and their cognate antigens, and enabling repertoire analyses to be efficacious in clinical and diagnostic settings.

In chapter 4, we derived the mathematical theory and inference framework for characterizing intermittent B cell evolution. The inference procedure we adapted and presented was not performant enough on ground truth data in resolving the switching rates of the phenotypes because we used an algorithm which factorizes and infers independently the switching rates and the branch times. Therefore, we did not use it to investigate the timescales of phenotypic switching for *in vivo* repertoire data; however, we believe the model itself has merit and can be adapted in the following ways. One avenue forward is performing maximum likelihood inference of the observed tree topologies and sampling times by not only marginalizing over the occupation time in the activated and deactivated states as we have done, but also marginalizing over the times of the branches. Although branch times of BCR phylogenies

would be interesting in the context of studying coevolution, they are nuisance parameters if we are keenly interested in studying the phenotypic switching rates. Another avenue forward, however computationally restrictive, would be to develop a Bayesian phylodynamic method that characterizes seedbanks [42], which are generalizations of the mutation-telegraph model we developed here. As coevolutionary models are produced and models of germinal center dynamics are improved, simulation-based inference may also be of utility when characterizing B cell phylogenies and the role of recall; however, these methods crucially rely on well-designed statistics of the phylogenies which may be nontrivial to develop or may not be identifiable. The role of memory B cells in secondary infections is not well understood, and characterizing these switching rates is pivotal for enhancing our knowledge of a fundamental adaptive immune process in addition to vaccine design against rapidly evolving pathogens, such as influenza and SARS-CoV-2.

Appendix A

SUPPLEMENT FOR CHAPTER 2

Data and code availability

The accession numbers for the BCR repertoire raw fastq data and single-cell data reported in this paper are BioProject: PRJNA645245 and PRJNA679920. All code for data processing and statistical analysis can be found at: https://github.com/StatPhysBio/covid-BCR

Cell lines

Sf9 cells (Spodoptera frugiperda ovarian cells, female, ATCC catalog no. CRL-1711) and High Five cells (Trichoplusia ni ovarian cells, female; Thermo Fisher Scientific, Waltham, United States (US), catalog number: B85502) were maintained in HyClone (GE Health Care, Chicago, US) insect cell culture medium.

Patients and samples

Specimens of heparinized blood were collected from the RT-PCR-confirmed patients with COVID-19 at the Infectious Disease Centre of the Princess Margaret Hospital, Hong Kong. The study was approved by the institutional review board of the Hong Kong West Cluster of the Hospital Authority of Hong Kong (approval number: UW20-169). All study procedures were performed after informed consent was obtained. Day 1 of clinical onset was defined as the first day of the appearance of clinical symptoms. The severity of the COVID-19 cases was classified based on the adaptation of the Sixth Revised Trial Version of the Novel Coronavirus Pneumonia Diagnosis and Treatment Guidance. The severity of the patients was categorized as follows: Mild - no sign of pneumonia on imaging, mild clinical symptoms; Moderate - fever, respiratory symptoms and radiological evidence of pneumonia; Severe -

dyspnea, respiratory frequency > 30/min, blood oxygen saturation 93%, partial pressure of arterial oxygen to fraction of inspired oxygen ratio < 300, and/or lung infiltrates > 50% within 24 to 48 hours; Critical - respiratory failure, septic shock, and/or multiple organ dysfunction or failure or death. For details on age, sex, and severity of each individual, see Data S1.

PBMC isolation

The blood samples were first centrifuged at 3000 xg for 10 minutes at room temperature for plasma collection. The remaining blood was diluted with equal volume of PBS buffer, transferred onto the Ficoll-Paque Plus medium (GE Healthcare), and centrifuged at 400 xg for 20 minutes. Peripheral Blood Mononuclear Cells (PBMC) samples were then collected and washed with cold RPMI-1640 medium for three times. The isolated PBMC samples were finally stored at cell freezing solution (10% DMSO + 90% FBS) and kept in −80°C until used.

RNA extraction and reverse transcription

Total RNA was extracted from 5×10^5 PBMC using the RNeasy Mini isolation kit (QIAGEN) according to the manufacturer's protocol. Reverse transcription of the RNA samples was performed using the Proto- Script® II Reverse Transcriptase kit (New England Biolabs, NEB) with random hexamer primers according to the manufacturer's protocol. The thermal cycling conditions were designed as follows: 25°C for 5 minutes, 42°C for 60 minutes, and 80°C for 5 minutes. The resulting cDNA samples were stored in 80°C freezer before PCR was performed.

Amplification of B cell repertoire from the samples by PCR

The cDNA samples were used as a template to amplify the antibody IgG heavy chain gene with six FR1-specific forward primers and one constant region-specific reversed primer using

the Phusion® High-Fidelity DNA Polymerase. The primer sequences were the same as previously described [331]; primer sequences are listed in Table A.1. The thermal cycling conditions were set as follows: 98°C for 30 s; 30 cycles of 98°C for 10 s, 58°C for 15 s, and 72°C for 30 s; and 72°C for 10 minutes. Then 10 ng of the PCR product was used as a template for the next round of gene amplification with sample-specific barcode primers. The thermal cycling conditions were set as follow: 98°C for 3 min; 30 cycles of 98°C for 10 s, 58°C for 15 s, and 72°C for 15 s; and a final extension at 72°C for 10 min using Phusion® High-Fidelity DNA Polymerase. The PCR product was purified by QIAquick Gel Extraction Kit (QIAGEN), and quantified by NanoDrop Spectrophotometers (Thermofisher).

5'-end primer	Sequence (5' - 3')
IGHV1	CCTCAGTGAAGGTCTCCTGCAAGG
IGHV2	TCCTGCGCTGGTGAAACCCACACA
IGHV3	GGTCCCTGAGACTCTCCTGTGCA
IGHV4	TCGGAGACCCTGTCCCTCACCTGC
IGHV5	CAGTCTGGAGCAGAGGTGAAA
IGHV6	CCTGTGCCATCTCCGGGGACAGTG
3'-end primer	**Sequence (5' - 3')**
CHG-R	GCGCCTGAGTTCCACGACAC

Table A.1: List of primers used for PCR amplification of B cell repertoire samples.

Protein expression and purification

The receptor-binding domain (RBD, residues 319–541) and N-terminal domain (NTD, residues 14 to 305) of the SARS-CoV-2 spike protein (GenBank: QHD43416.1) as well as the RBD (residues 306-527) and NTD (residues 14-292) of SARS-CoV-1 spike protein (GenBank: ABF65836.1) were cloned into a customized pFastBac vector [179, 320]. The RBD and NTD

constructs were fused with an N-terminal gp67 signal peptide and a C-terminal His6 tag. Recombinant bacmid DNA was generated using the Bac-to-Bac system (Life Technologies, Thermo Fisher Scientific). Baculovirus was generated by transfecting purified bacmid DNA into Sf9 cells using FuGENE HD (Promega, Madison, US) and subsequently used to infect suspension cultures of High Five cells (Life Technologies) at a multiplicity of infection (moi) of 5 to 10. Infected High Five cells were incubated at 28°C with shaking at 110 rpm for 72 h for protein expression. The supernatant was then concentrated using a Centramate cassette (10 kDa molecular weight cutoff for RBD, Pall Corporation, New York, USA). RBD and NTD proteins were purified by Ni-NTA Superflow (QIAGEN, Hilden, Germany), followed by size exclusion chromatography and buffer exchange to phosphate-buffered saline (PBS).

CD38+ plasma B cell enrichment

CD38+ plasma B cells were isolated from the PBMC samples by performing two subsequent magnetic separation steps according to the manufacturer's protocol (Plasma Cell Isolation Kit II, human, Miltenyi Biotec). Briefly, non-plasma B cells are labeled with magnetic beads combined with cocktail antibodies and separated using the MACS column. Then, CD38+ plasma B cells are directly labeled with CD38 MicroBeads and isolated from the pre-enriched B cell pool. Purified CD38+ plasma B cells were eluted and washed in PBS containing 2% (v/v) fetal bovine serum (FBS) and kept for the following RNA isolation step. In order to test the purity of the CD38+ plasma B cells, we also added staining antibodies and 10 µL of Anti-human CD19-BV510 (BioLegend) and CD38-PE-Cy7 (BioLegend) and incubated them for 15 minutes in the dark in the refrigerator (2-8°C). Cells were finally fixed with 4% PFA for 20 minutes on ice. The stained samples were acquired by flow cytometry on a FACS Attune (Invitrogen) and analyzed with FlowJo software (Figure 2.1).

RBD and NTD protein specific binding B cell enrichment

B cells were enriched from the PBMC samples according to the manufacture's protocol (B Cell Isolation Kit II, human, Miltenyi Biotec). Briefly, non-B cells are labeled with a

cocktail of biotin-conjugated antibodies and separated by the MACS column. Purified B cells were eluted and kept in the PBS buffer with 2% (v/v) FBS. The enriched B cells were then incubated with 2 µg Biotin-RBD or NTD protein for 30 min at 4°C. After incubation, Anti-Biotin MicroBeads were added and incubated for 30 min. RBD and NTD specific bead binding B cells were washed and eluted in PBS and stored on ice until use. In order to test the purity of the RBD- or NTD-specific B cells, we also added staining antibodies, 10 µL of Anti-human CD19-BV510 (BioLegend), and 2 µg of SARS-CoV-2 RBD-PE or NTD-PE and incubated them for one hour in the dark in the refrigerator (2-8°C). Cells were finally fixed with 4% PFA for 20 minutes on ice. The stained samples were acquired by flow cytometry on a FACS Attune (Invitrogen) and analyzed with FlowJo software (Figure 2.1).

Single B cell 5' mRNA and VDJ sequencing

After RBD or NTD specific B cells enrichment, cells were counted by using 0.4% (w/v) trypan blue stain solution in the microscope and directly loaded on the 10X Chromium™ Single Cell A Chip. Then single B cell lysis and RNA first-strand synthesis were carried out following the 10X Chromium™ Single Cell 5' Library & Gel Bead Kit protocol. The RNA sample were used for the next step B cell VDJ library construction following the Chromium™ Single Cell V(D)J Enrichment Kits protocol. VDJ library sequencing was performed on a NovaSeq PE150 and the sequencing data were processed by Cell Ranger.

ELISA

A 96-well enzyme-linked immunosorbent assay (ELISA) plate (Nunc MaxiSorp, Thermo Fisher Scientific) was first coated overnight with 100 ng per well of purified recombinant protein in PBS buffer. The plates were then blocked with 100 µL of Chonblock blocking/sample dilution ELISA buffer (Chondrex Inc, Redmon, US) and incubated at room temperature for 1 h. Each human plasma sample was diluted to 1:100 in Chonblock blocking/sample dilution ELISA buffer. Each sample was then added into the ELISA plates for a two-hour incubation at 37°C. After extensive washing with PBS containing 0.1% Tween 20, each well in the plate

was further incubated with the anti-human IgG secondary antibody (1:5000, Thermo Fisher Scientific) for 1 hour at 37°C. The ELISA plates were then washed five times with PBS containing 0.1% Tween 20. Subsequently, 100 µL of HRP substrate (Ncm TMB One; New Cell and Molecular Biotech Co. Ltd, Suzhou, China) was added into each well. After 15 min of incubation, the reaction was stopped by adding 50 µL of 2 M H2SO4 solution and analyzed on a Sunrise (Tecan, Männedorf, Switzerland) absorbance microplate reader at 450 nm wavelength.

BCR preprocessing

We used a similar procedure for processing of the bulk and the plasma B cell receptor repertoires. For initial processing of the raw reads, we used pRESTO (version 0.5.13) [307] to assemble paired-end reads, remove sequences with a mean quality score less than 30, mask primer subsequences, and collapse duplicate sequences into unique sequences. The small fraction of paired-end reads that overlapped were assumed to be anomalous and were discarded from the analysis. Additionally, after preprocessing with pRESTO, we discarded unique reads that contained ambiguous calls (N's) in their receptor sequence.

BCR error correction

We performed two rounds of error correction on sequences that passed the quality control check. In the first round, we clustered singletons and other low-frequency sequences into larger sequences if they were similar in sequence. The intent of this round was to correct for sequencing errors (e.g., from reverse transcription of mRNA to cDNA) that caused large abundance clones to be split into many similar sequences. We used two parameters: Δ_r, the marginal Hamming distance tolerance per decade in log-ratio abundance (each $\log_{10}$ unit allowing Δ_r additional sequence differences), and Δ_a, the marginal abundance tolerance of clusterable sequences per decade in log-ratio abundance (each log10 unit allowing abundance Δ_a higher as clusterable). For example, a sequence with abundance a_1 and a Hamming distance d away from a higher abundance sequence with abundance a_2 was absorbed into the

latter only if $d \leq \Delta_r \log_{10} \frac{a_2}{a_1}$ and $a_1 \leq \Delta_a \log_{10} \frac{a_2}{a_1}$. We used the output of this first round as input for the second round of error correction, in which we more aggressively target correction of reverse transcriptase errors. In the second round, we used two different parameters to assess sequence similarity: d_{thresh}, the Hamming distance between sequences, and a_{thresh}, the ratio of sequence abundances. A sequence with abundance a_1 and a Hamming distance d away from a sequence of larger abundance a_2 was absorbed into the latter only if $d \leq d_{\text{thresh}}$ and the ratio of the sequence abundances was greater than a_{thresh}, i.e., $\frac{a_2}{a_1} \geq a_{\text{thresh}}$. This round of error correction allows much larger abundance sequences to potentially be clustered than is possible in the first round. For both of the above steps, we performed clustering greedily and approximately by operating on sequences sorted by descending abundance, assigning the counts of the lower abundance sequence to the higher abundance one iteratively.

After error correction, the sequences still contained a large number of singletons, i.e., sequences with no duplicates (Data S1) We discarded these singletons from all analyses that relied on statistics of unique sequences (i.e., the results presented in Figures 2.5A–S2.5C and 2.3E–2.3G).

BCR annotation

For each individual, error-corrected sequences from all time points and technical replicates were pooled and annotated by abstar (version 0.3.5) [30]. We processed the output of abstar, which included the estimated IGHV gene/allele, IGHJ gene/allele, location of the HCDR3 region, and an inferred naive sequence (germline before hypermutation). Sequences which had indels outside of the HCDR3 were discarded. We partitioned the sequences into two sets: productive BCRs, which were in-frame and had no stop codons, and unproductive BCRs, which were out-of-frame.

Unproductive BCRs

Due to a larger sequencing depth in healthy individuals, we were able to reconstruct relatively large unproductive BCR lineages. Unproductive sequences are BCRs that are generated but,

due to a frameshift or insertion of stop codons, are never expressed. These BCRs reside with productive (functional) BCRs in a nucleus and undergo hypermutation during B cell replication and, therefore, provide a suitable null expectation for generation of BCRs in immune repertoires.

Clonal lineage reconstruction

To identify BCR clonal lineages, we first grouped sequences by their assigned IGHV gene, IGHJ gene, and HCDR3 length and then used single-linkage clustering with a threshold of 85% Hamming distance. A similar threshold has been suggested previously by [103] to identify BCR lineages. Defining size as the sum of the number of unique sequences per time point within a lineage, clusters of size smaller than three were discarded from most analyses. They were retained only for training receptor generation and selection models and were not discarded in the sharing analysis only if the progenitor of that small cluster was also a progenitor of a cluster of size at least three in another patient. For each cluster, there may have been multiple inferred naive sequences, as this was an uncertain estimate. Therefore, the most common naive sequence was chosen to be the naive progenitor of the lineage. When the most common naive sequence of a productive lineage contained a stop codon, the progenitor of the lineage was chosen iteratively by examining the next most common naive sequence until it did not contain any stop codons. If all inferred naive sequences in a productive lineage had a stop codon, that lineage was discarded from the analysis. Data S1 shows the statistics of constructed clonal lineages in each individual for the bulk repertoire and combined bulk+plasma B cell repertoire, respectively.

Mapping of single-cell data onto reconstructed clonal lineages

Like the repertoire datasets, the single-cell sequences were annotated by abstar [30]. For each receptor acquired by single-cell sequencing, we identified a subset of reconstructed clonal lineages from the bulk repertoire which had identical HCDR3 length as the sequence and which also had an IGHV gene which was 90% similar to that of the single-cell receptor. This

flexibility in V-gene choice would identify functionally homologous receptors and associate a receptor to a lineage with a sequence divergence in the V-segment, compatible with the expectation under somatic hypermutations [168]. A single-cell sequence was matched to a reconstructed clonal lineage from this subset if its HCDR3 could be clustered with other members of the lineages, using single-linkage clustering with a similarity threshold of 85% Hamming distance (similar to the criteria for lineage reconstruction for bulk repertoires).

Inference of generation probability and selection for BCRs

We used IGoR (version 1.4) [184] to obtain a model of receptor generation. This model characterized the probability of generation $P_{\text{gen}}(\sigma)$ of a receptor dependent on the features of the receptor, including the IGHV, IGHD, and IGHJ genes and the deletion and insertion profiles at the VD and DJ junctions. To characterize the parameters of this model, we trained IGoR on the progenitors of unproductive lineages, regardless of size, pooled from the bulk repertoire of all individuals, restricted to progenitors whose HCDR3 began with a cysteine and ended with a tryptophan. For consistency with our receptor annotations based on abstar, we used abstar's genomic templates and the HCDR3 anchors of abstar's reference genome as inputs for IGoR's genomic templates and HCDR3 anchors. distributions of the healthy and COVID-19 cohorts in this study are shown in Figure 2.6A.

We used SONIA (version 0.45) [269] to infer a selection model for progenitors of productive clonal lineages. The SONIA model evaluated selection factors q to characterize the deviation in the probability $P_{\text{post}}(\sigma)$ to observe a functional sequence in the periphery from the null expectation based on the generation probability $P_{\text{gen}}(\sigma)$: $P_{\text{post}}(\sigma) = \frac{1}{Z} P_{\text{gen}}(\sigma) e^{\sum_{f:\,\text{features}} q_f(\sigma)}$, where Z is the normalization factor and $q_f(\sigma)$ are selection factors dependent on the sequence features f. These sequence features include IGHV-gene and IGHJ-gene usages and HCDR3 length and amino acid composition [269].

In our analysis, we used the SONIA left-right model with independent IGHV- and IGHJ-gene usages [269]. We used the output from IGoR [184] as the receptor generation model for SONIA. We trained four cohort-specific selection models on progenitors of productive

lineages, regardless of size, pooled from the bulk repertoire of all individuals within a cohort, restricted to progenitors whose HCDR3 began with a cysteine and ended with a tryptophan. 150 epochs, L2 regularization with strength 0.001, and 500,000 generated sequences were used to train each SONIA model. Figure 2.7 shows the distributions for the probabilities of observing productive receptors sampled from each cohort $P_{\text{post}}(\sigma)$. A selection model was also trained on all the productive lineage progenitors in the GRP dataset [31] and used 5,000,000 generated sequences, keeping the other parameters unchanged. We refrain from comparing directly associated with GRP BCRs to BCRs in this study due to experimental differences.

It should be noted that the (pre-selection) generation model inferred by IGoR [184] is robust to sequence errors due to experimental errors or hypermutations in the IgG repertoires. However, hypermutations in BCRs could introduce errors in inference of selection models and estimation of receptor probabilities by SONIA [269]. Therefore, we have restricted our selection analyses to only the inferred progenitors of clonal lineages. Although the inferred progenitors of lineages can still deviate from the true (likely IgM naive) progenitors, the selection models inferred from ensembles of inferred progenitors in IgG repertoires seem to be comparable to the models inferred from the IgM repertoires (M. Ruiz Ortega, personal communication). The resulting selection models, trained on either true or inferred progenitors, reflect preferences for sequence features of unmutated receptors, including IGHV- and IGHJ-genes and HCDR3 length and composition, but they do not account for the hypermutation preferences that may distinguish one cohort from another.

Characterizing the robustness of selection inference

To test the sensitivity of the inferred selection models on the size of the training sets, we downsampled the receptor data of each COVID-19 cohort to a size comparable to the smallest cohort, i.e., the healthy repertoire sequenced in this study. This downsampling resulted in two independent training datasets for the mild COVID-19 cohort, 13 independent training datasets for the moderate COVID-19 cohort, and three independent training datasets for the

severe COVID-19 cohort. Though this downsampling resulted in over 400 independent training datasets for the GRP, we elected to use only 15. We then inferred a separate selection model with SONIA for each of these training datasets and used each model to evaluate the receptor log-probabilities $\log_{10} P_{\text{post}}(\sigma)$ for a set of 500,000 generated receptors. The evaluated probabilities are strongly correlated between models inferred from the downsampled data in each cohort, with a Pearson correlation of r $> 0.99$and p value $= 0$ (p value is smaller than machine precision) (see Figures 2.6C–2.6F). We used a similar approach to compare the selection model inferred from the healthy repertoires sequenced in this study and the GRP study [31]. Figure 2.6B shows that, using the model inferred with our healthy repertoire and 30 downsampled independently inferred selection models using the GRP dataset, the evaluated log-probabilities $\log_{10} P_{\text{post}}(\sigma)$ based on these two datasets are strongly correlated, with a Pearson correlation of r > 0.99 and p value $= 0$ (p value is smaller than machine precision). See Figure 2.6B.

Characterizing repertoire diversity

We quantified the diversity of each cohort by evaluating the entropy of receptor sequences in each cohort. Entropy can be influenced by the size of the training dataset for the selection models. To produce reliable estimates of repertoires' diversities (and entropies), we used the procedure described above to learn independent selection models for subsampled repertoires in each cohort. We then used the inferred IGoR and SONIA models to generate 500,000 synthetic receptors based on each of the subsampled, cohort-specific models. We evaluated cohort entropies H as the expected log-probabilities to observe a functional sequence in the respective cohort:

$$H = -\sum_{\sigma} P_{\text{post}}(\sigma) \log P_{\text{post}}(\sigma) = \langle \log P_{\text{post}}(\sigma) \rangle_{P_{\text{post}}(\sigma)}. \tag{A.1}$$

The estimates based on the generated receptors are reported in the Chapter 2. The error bars reported for these entropy estimates are due to variations across the inferred models in each cohort.

For comparison, we also evaluated the entropy estimated on the repertoire data in each cohort, which showed a similar pattern to the estimates from the generated cohorts. Specifically, the entropy of BCR repertoires estimated from the data follows: 39.8 ± 0.3 bits in healthy individuals, 41.9 ± 0.7 bits for patients in the mild cohort, 42.7 ± 0.3 bits for patients in the moderate cohort, and 42.9 ± 0.5 for patients in the severe cohort. The error bars indicate the standard error due to differences among individuals within a cohort.

Comparing selection between repertoires of cohorts

Selection models enable us to characterize the sequence features of immune repertoires that differ between cohorts. We evaluated the Jensen-Shannon divergence $D_{\mathrm{JS}}(r, r')$ between the distribution of repertoires r and r', P^r_{post} and $P^{r'}_{\mathrm{post}}$, defined as

$$D_{\mathrm{JS}}(r,r') = \frac{1}{2} \sum_{\sigma:\,\text{sequences}} P^r_{\mathrm{post}}(\sigma) \log \frac{P^r_{\mathrm{post}}(\sigma)}{\left(P^r_{\mathrm{post}}(\sigma) + P^{r'}_{\mathrm{post}}(\sigma)\right)/2} + P^{r'}_{\mathrm{post}}(\sigma) \log \frac{P^{r'}_{\mathrm{post}}(\sigma)}{\left(P^r_{\mathrm{post}}(\sigma) + P^{r'}_{\mathrm{post}}(\sigma)\right)/2} \tag{A.2}$$

$$= \frac{1}{2} \sum_{\sigma:\,\text{sequences}} P^r_{\mathrm{post}}(\sigma) \log \frac{2Q^r(\sigma)}{Q^r(\sigma) + Q^{r'}(\sigma)} + P^{r'}_{\mathrm{post}}(\sigma) \log \frac{2Q^{r'}(\sigma)}{Q^r(\sigma) + Q^{r'}(\sigma)}, \tag{A.3}$$

where we used the relationship between a receptor's generation probability $P_{\mathrm{gen}}(\sigma)$ and its probability after selection $P^r_{\mathrm{post}}(\sigma)$, using the inferred selection factor $Q^r = \frac{1}{Z} e^{\sum_{f:\,\text{features}} q^r_f(\sigma)}$ in repertoire r: $P^r_{\mathrm{post}}(\sigma) = P_{\mathrm{gen}}(\sigma) Q^r(\sigma)$. The Jensen-Shannon divergence $D_{\mathrm{JS}}(r, r')$ is a symmetric measure of distance between two repertoires, which we can calculate using their relative selection factors [133]. Figure 2.7 shows the expected partial Jensen-Shannon divergences evaluated over five independent realizations of 100,000 generated sequences for each partial selection model. The error bars show the variations of these estimates (i.e., standard deviation) over the five independent realizations in this procedure.

Clonal lineage expansion

We studied clonal lineage expansion of BCR repertoires in individuals that showed an increase in the binding level ($OD_4 50$) of their plasma to SARS-CoV-2 (RBD) during infection (Figures

2.8A and 2.9): patients 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14. Other individuals showed no increase in IgG binding to SARS-CoV-2 (RBD), either due to already high levels of binding at early time points or to natural variation and noise (Figure 2.9). Our expansion test compared two time points. Therefore, for individuals with three time points, we combined data from different time points such that the separated times coincided with larger changes in binding levels (OD_{450}). Specifically, we combined the last two time points for patients 2 and 7 and the first two time points for patient 9. In addition, we combined the technical replicates at the same time point and filtered out small lineages with size less than three, where size was defined as the sum of the amount of unique sequences per time-point within a lineage.

To test for expansion, we compared lineage abundances (i.e., total number of reads in a lineage) between early and late time points. Many lineages appeared only in one time point due to the sparse sampling of clonal lineages and the cells that generate them (Figure 2.10). Therefore, we tested for expansion only for lineages that had nonzero abundances at both time points.

Our expansion test relied on comparing the relative abundance of a given lineage with other lineages. However, due to primer-specific amplification biases, abundances were not comparable between reads amplified with different primers. Therefore, in our analysis we only compare a lineage with all other lineages that were amplified with the same primer.

We applied a hypergeometric test (Fisher's exact test) to characterize significance of abundance fold change for a focal lineage. A similar method was used to study clonal expansion in TCRs [65]. For each focal clonal lineage (in a given individual), we defined a 2×2 contingency matrix C,

$$C = \begin{pmatrix} n_i^{\text{early}} & N_{/i}^{\text{early}} \\ n_i^{\text{late}} & N_{/i}^{\text{late}} \end{pmatrix}, \tag{A.4}$$

where n_i^{early} and n_i^{late} are the abundances of the focal lineage at the early and late time, and $N_{/i}^{\text{early}}$ and $N_{/i}^{\text{late}}$ are the total abundances of all the reads (with the same primer) minus those from lineage i at the early and late times. The ratio $\frac{n_i^{\text{early}}}{n_i^{\text{late}}} / \frac{N_{/i}^{\text{late}}}{N_{/i}^{\text{early}}}$ describes the fold change, or odds ratio, of lineage i relative to the rest of the reads in the same primer group. Based

on the contingency matrix C, one-sided p values for Fisher's exact test were calculated using the "fisher.test" function in R version 4.0. Fold change and p values are shown in Fig. 2.10G.

To determine a significance threshold for the Fisher's exact test, we examined the technical replicate data from samples collected from the same time point in each individual because we did not expect any significant expansion among technical replicates. We performed the expansion test on pairs of technical replicates (Figure 2.10C) and compared the empirical cumulative distributions of the time point and replicate expansion data (Figures 2.10E and 2.10F) [286, 287]. We chose a p value threshold of 10^{-300}, where there were 12.3 as many significant expansions as in the replicate data, and therefore the false discovery rate was approximately $1/(1+12.3) = 0.075$.

Significance of BCR sharing among individuals

The probability that receptor σ is shared among a given number of individuals due to convergent recombination can be evaluated based on the probability to observed a receptor in the periphery $P_{\text{post}}(\sigma)$, the size of the cohort M, and the size of the repertoire (sequence sample size) N. First, we evaluated the probability $\rho(\sigma; N)$ that receptor σ with probability $P_{\text{post}}(\sigma)$ appears at least once in a sample of size N,

$$\rho(\sigma; N) = 1 - (1 - P_{\text{post}}(\sigma))^N \simeq 1 - e^{-NP_{\text{post}}(\sigma)}. \tag{A.5}$$

The probability that receptor σ is shared among m individuals out of a cohort of M individuals, each with a (comparable) sample size N, follows a the binomial distribution,

$$P_{\text{share}}(\sigma; m, M, N) = \binom{M}{m} [\rho(\sigma; N)]^m [1 - \rho(\sigma; N)]^{M-m}. \tag{A.6}$$

We aimed to identify shared receptors that were outliers such that their probability of sharing is too small to be explained by convergent recombination or other biases in the data. To do so, we identify the receptors with the smallest sharing probabilities P_{share} and found a threshold of P_{post} (dashed lines in Figures 2.11 and 2.12) at the 2% quantile of P_{share} in the data. Specifically, since P_{share} is a function of P_{post} and m (number of individuals sharing a

receptor), for each m we solved for P_{post} such that $P_{\text{share}} = c$. We tuned c such that only 2% of the data lie below P_{share}. This was a conservative choice to identify the rare shared outliers in the data.

Quantification and statistical analysis

Differences in the mean HCDR3 lengths and $\log_{10}$ relative read abundance in the plasma B cell repertoire for expanded and non-expanded lineages were studied by ordinary one-way ANOVA tests using SciPy 1.5 and python 3.8.5. Results can be found in the captions of Figures 2.4 and 2.5 and in the caption of Figure 2.8, respectively. The Pearson correlation coefficients and p values associated with testing for non-correlation for the correspondences between the bulk and plasma repertoires of patients, results shown in the legend of Figure 2.3A, and between the $\log_{10} P_{\text{post}}$ of independently trained SONIA models, results shown in plots and the caption of Figure 2.6, were found using SciPy 1.5. Fisher exact tests for the expansion analyses were calculated using the "fisher.test" function in R version 4.0 and in Python using the fisher module found at https://github.com/brentp/fishers˙exact˙test. The details of how the Fisher exact tests were constructed can be found above, and the results are shown in Figures 4, 5, and S6. IGoR 1.4 was used to infer a baseline generation model, and SONIA 0.45 was used to infer a selection model. Details on how both were used can be found in Methods. Binomial sampling p values were obtained using SciPy 1.5. Jensen-Shannon divergences, entropy estimates, and the statistical analysis for identifying rare receptors were detailed in Methods and were developed in-house and can be found in the GitHub repository for this paper at https://github.com/StatPhysBio/covid-BCR.

Appendix B

SUPPLEMENT FOR CHAPTER 3

Data and code availability

All code for data processing and statistical analysis can be found at: https://github.com/StatPhysBio/pasc

Donors and samples

The cohort studied here was composed of 120 individuals from the INCOV cohort [290]. Donors were recruited at five hospitals of Swedish Medical Center and affiliated clinics located in the Puget Sound region near Seattle, with all providing written in-person consent. The WHO Ordinal Scale for Clinical Improvement score (WOS) [185] was used to quantify disease severity at each blood draw.

TCR repertoire sequencing

Immunosequencing of the TRβ chains was performed by Adaptive Biotechnologies.

TCR repertoire preprocessing

Unproductive repertoires were obtained by selecting from a sample only those clones which were annotated as out-of-frame. Productive repertoires were obtained by selecting only those clones which were in-frame, had no stop codons, and had no pseudogenes. Gene names were conformed to follow IMGT conventions. TCRs were deduplicated at the level of TRβV gene, TRβJ gene, and TCRβ CDR3 nucleotide sequence in each sampled repertoire and abundances of identical sequences were summed. We group TCRs by their TRβV gene and TCRβ CDR3

amino acid sequence when applying the TCR-OT algorithm or studying dynamics since the TRβJ gene has been shown to be redundant when information on the tcrb CDR3 sequence is used [110]. The term functional clone, used throughout chapter 3 and this appendix, is defined as a TCR's amino acid sequence, i.e., its TRβV gene identity and TCRβ CDR3 amino acid sequence.

TCRdist

TCRdist is a distance metric between TCRs that compares all regions of TCR sequences which are believed to be important for binding peptides, as opposed to the CDR3 region only, in a physicochemical manner [57, 187]. TCRdist is computed using the algorithm described in [187].

$$\text{TCRdist}(\sigma, \sigma') = \sum_{c \in \mathcal{C}} w(c) \sum_{i \in \mathcal{A}} \tilde{d}(\sigma_{c,i}, \sigma'_{c,i}) \tag{B.1}$$

$\mathcal{C}$ is the collection of regions in the TCR sequences that are compared pairwise,

$$\mathcal{C} = \{\text{CDR1}, \text{CDR2}, \text{CDR2.5}, \text{CDR3}\}, \tag{B.2}$$

where CDR2.5 is the pMHC-facing loop between CDR2 and CDR3. $w(c)$ is a function weighing the contributions of each sequence portion.

$$w(c) = \begin{cases} 3, & c = \text{CDR3}\beta \\ 1, & \text{else} \end{cases} \tag{B.3}$$

The CDR3 sequence is given a weight of 3 since it plays a larger role in binding to epitopes, and this puts it on equal footing with the contributions from the 3 other non-CDR3 loops combined. $\mathcal{A}$ is the best gapped alignment between σ and σ', and i indexes the amino acid in the sequences. Notably, the gapped alignment for the CDR3 is obtained by ignoring the first three and last two residues of the CDR3 sequences since these residues play a smaller role in binding epitopes because they form a conserved beta sheet. The mismatch distance

between two aligned CDRs uses the BLOSUM62 matrix [111]:

$$\tilde{d}(\sigma_i, \sigma_i') = \begin{cases} 0, & \sigma_i = \sigma_i' \\ 4, & \sigma_i = \text{`}-\text{'} \text{ or } \sigma_i' = \text{`}-\text{'} \\ \min(4, 4 - \text{BLOSUM62}(\sigma_i, \sigma_i')), & \text{else} \end{cases} \tag{B.4}$$

The best gapped alignment $\mathcal{A}$ is the alignment which yields the minimum summed mismatch distance between the shorter sequence with inserted gaps and the longer sequence. Fast, parallelizable TCRdist calculations are performed using the `tcrdist-rs` package [199].

Collecting the top rank-k clones

TCRs were grouped by their TCRβ CDR3 amino acid sequence and TRβV gene and their abundances totaled within each group. The top rank-k TCRs were identified as being among the k-most abundant in at least one of the samples at a given time point. We handled ties by assigning them the minimum rank.

Molecular composition of sequence motifs

Let x be an amino acid characteristic having a value of aromatic, polar and uncharged, nonpolar and nonaromatic, negative, or positive. For a TCRβ CDR3 amino acid sequence σ, we compute the amino acid characteristic weight as

$$w_x(\sigma) = \sum_{i=4}^{L-2} \frac{\mathbb{1}_x(X(\sigma_i))}{L-5}, \tag{B.5}$$

where L is the length of the TCRβ CDR3 amino acid sequence, $X(\sigma_i)$ gives the amino acid characteristic at site i, $\mathbb{1}(y)_x$ is the indicator function which returns 1 only if $y \in x$, and we ignore the first three and last two positions along the TCRβ CDR3 sequence, consistent with TCRdist. This ensures that sequences with different lengths are treated similarly. The within-cluster amino acid characteristic weight is then computed as

$$w_{x,\text{cluster}} = \frac{1}{N} \sum_{\sigma \in \text{cluster}} n(\sigma) w_x(\sigma), \tag{B.6}$$

where N is the number of independent recombinations in a cluster, and $n(\sigma)$ is the number of independent recombinations which translate to σ.

Graphical analysis

We used Gephi (v.10.1) [22] to create undirected network graphs for the TCR-OT clusters (Fig. 3.4, 3.6). For each given TCR-OT cluster, edges connected nodes if nodes were within 48 TCRdist of each other, and all nodes in a cluster had edges to the focal sequences of the cluster (those with the highest enrichment), save for self-edges. To arrange the clusters, we used the ForceAtlas 2 layout with stronger gravity enabled, gravity set to 1, and a scaling of 10. Nodes sizes range from 2 to 6, depending on the strength of their BH-corrected $-\log_{10}$ significance of their TCR-OT enrichment statistic.

Matching into VDJdb

The VDJdb (dated 2024.06.13) [275, 98] was downloaded from the vdjdb-db GitHub repository. We restricted the database to TCRs from humans and which were detected as reactive to SARS-CoV-2, EBV, CMV, and InfluenzaA. Additionally, we removed 10x Genomics data known to be contaminated [53], i.e., TCR sequences from [1] which were annotated as being paired with HLA-A*03:01 or HLA-A*11:01. We deduplicated the remaining data by their paired-chain TCR sequence; the annotated antigen species, epitope, and gene; and the annotated HLAs. A sequence from the INCOV TCRβ repertoires was matched to a sequence in the VDJdb if it was within 24 TCRdist of a VDJdb sequence and there was at least one identical HLA allele group between the INCOV individual from which the TCR originated and the VDJdb annotation. After a sequence σ was compared to the VDJdb, its weight for matching to each antigen species was computed as

$$w_a(\sigma) = c_a(\sigma) \Big/ \sum_a c_a(\sigma), \tag{B.7}$$

where $c_a(\sigma)$ is the number of matches σ had with VDJdb TCRs reactive to antigen a.

Matching into the MIRA database

The COVID-19 MIRA database (accessed November 14, 2024) was downloaded from Adaptive Biotechnologies COVID-2020 repository [216]. TCRβ sequences in the database underwent preprocessing in which gene names were conformed to IMGT conventions, and sequences were removed if they were unproductive or missing their TRβV or TRβJ genes. Moreover, the database was deduplicated at the level of TCRβ CDR3 nucleotide sequence, TRβV gene, TRβJ gene, and individual. A sequence from the INCOV TCRβ repertoires was said to be matched to the MIRA database, and thus specific for SARS-CoV-2, if it was within 12 TCRdist of at least two sequences in the MIRA database.

Training generation and selection models

We used IGoR (version 1.4) [184] to learn a model of VDJ recombination for each individual by pooling all their observed unproductive sequences. Since the provided annotations did not contain a TCRβ CDR3 nucleotide sequence for unproductive sequences, IGoR was used to align the full TCRβ nucleotide sequence against genomic templates to identify the nucleotide TCRβ CDR3 sequence. After this alignment, the unproductive sequences were deduplicated at the level of TCRβ CDR3 nucleotide sequence, TRβV allele, and TRβJ allele to retain only those sequences which most likely came from independent recombination events. IGoR models were initialized from the default TCRβ IGoR model and trained using 10 epochs of expectation-maximization.

soNNia (version 0.3.2) [133] TCRβ models were trained for each individual's repertoire using the deep architecture with 150 epochs, 500,000 P_{gen} sequences, L_2 regularization of 3×10^{-4}, joint V-J features, and a batch size of $5,000$. Training datasets were constructed from productive TCRs whose TCRβ CDR3 amino acid sequences began with a cysteine and whose TCRβ CDR3 amino acid sequences lengths did not exceed 30 residues. To limit biases due to expansion, the training datasets were deduplicated at the level of TCRβ CDR3 nucleotide sequences, TRβV gene, and TRβJ gene. When learning a soNNia model for each

sampled repertoire, P_{gen} sequences were generated from the associated individual's bespoke IGoR model.

Parametric equi-P_{share} vs. binomial p-values

p-values computed using the binomial distribution can also be used to quantify whether sequences are rare. Choosing a p-value threshold such that 0.5% of the shared repertoire is rare, as in the main text, results in almost the identical set (roughly 99.8% overlap) of sequences selected as that obtained using the equi-P_{share} parametric curve. Notably, sequences detected by the equi-P_{share} parametric curve and not binomial significance tended to be incident in more individuals.

Creating a control for the subset of rare clones

We mirror ref. [254] for creating a control cohort and summarize their method here. We recall that constituents of the repertoire studied for sharing are defined by their TRβV gene, TRβJ gene, and TCRβ CDR3 amino acid sequence. To create a control relative to the subset of clones identified as rare, we want to sample from the bulk TRβ repertoire such that we reproduce the $P_{\text{obs}}(\sigma)$ distribution of the rare clones. We histogram the $\log_{10} P_{\text{obs}}(\sigma)$ distribution of the rare clones evenly with n bins, yielding a height h_i of counts in each bin; here, we use 500 bins spaced evenly between -40 and 0. We bin the $\log_{10} P_{\text{post}}(\sigma)$ of the sequences in the deduplicated bulk TRβ repertoire and sample h_i TRβ sequences without replacement in each respective bin. When then check the overlap of those sequences with the MIRA database (Fig. 3.9).

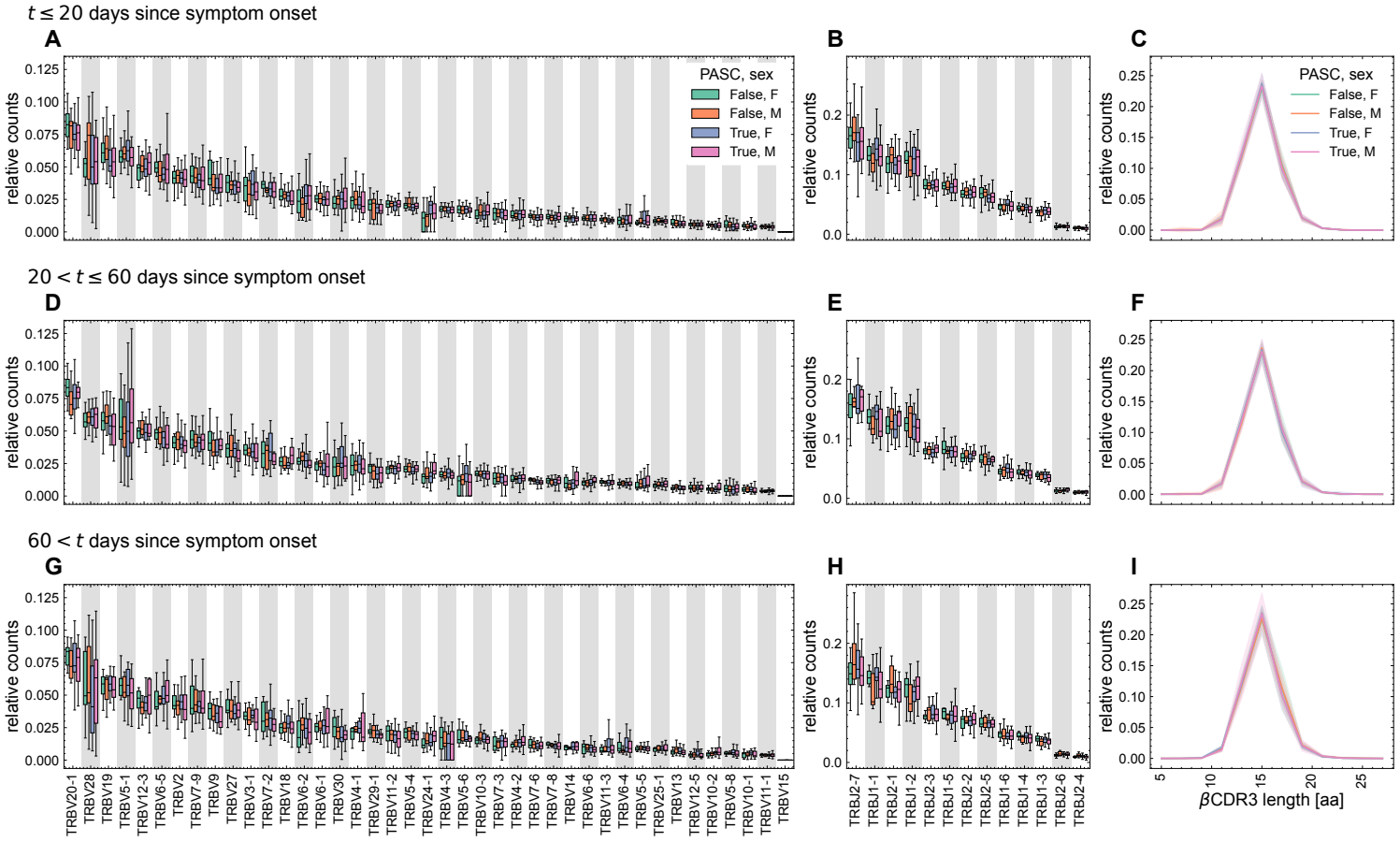


Figure B.1: **Receptor composition at the level of recombinations weighted by abundance across cohorts. (A, D, G)** The distribution of V gene usages is shown as boxplots, stratified by PASC status and sex (colors) and sampling time relative to symptom onset as indicated above each triple of plots. Interquartiles are depicted in the box, with the median shown as a black, horizontal line. The rest of the distribution is shown as whiskers. Outliers are suppressed due to the high variability so as not to impair visualization of the bulk behavior. V genes are shown only if their usages were above 1% of the entire repertoire in any one of the cohorts. **(B, E, H)** The distribution of J gene usages is shown as boxplots as in (A, D, G). J genes are shown only if their usages were above 1% of the entire repertoire in any one of the cohorts. **(C, F, I)** The distribution of βCDR3 lengths is shown. Lines indicate average relative counts for each cohort, and shading indicates regions containing one standard deviation of variation across individuals within a cohort.

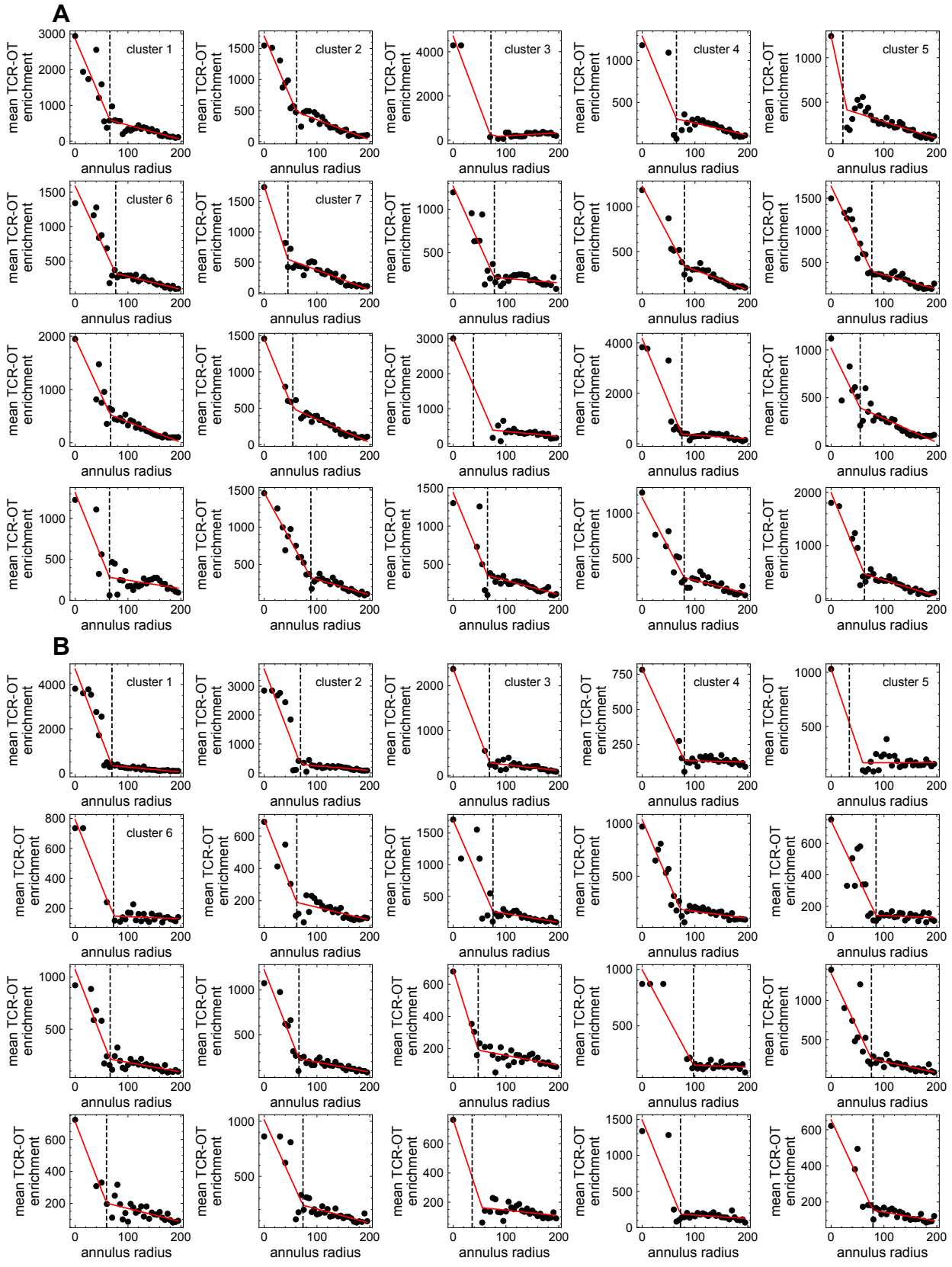

A
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
mean TCR-OT enrichment
annulus radius
B
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
mean TCR-OT enrichment
annulus radius

Figure B.2: **TCR-OT cluster breakpoint analyses on top rank-50 repertoires** The mean TCR-OT enrichment vs annulus radius (given in units TCRdist) for cluster detection in the PASC+ top 50-rank repertoire **(A)** and PASC− top 50-rank repertoire **(B)**. The black dots show the observed statistic at within the annulus, and the red line shows the segmented linear fit (Methods). The breakpoint, shown as a vertical black, dashed line, in each segmented linear regression determines the cluster radius. Regressions associated with clusters used in downstream analyses are annotated with their cluster size rank, as in Fig. 3.4F,G.

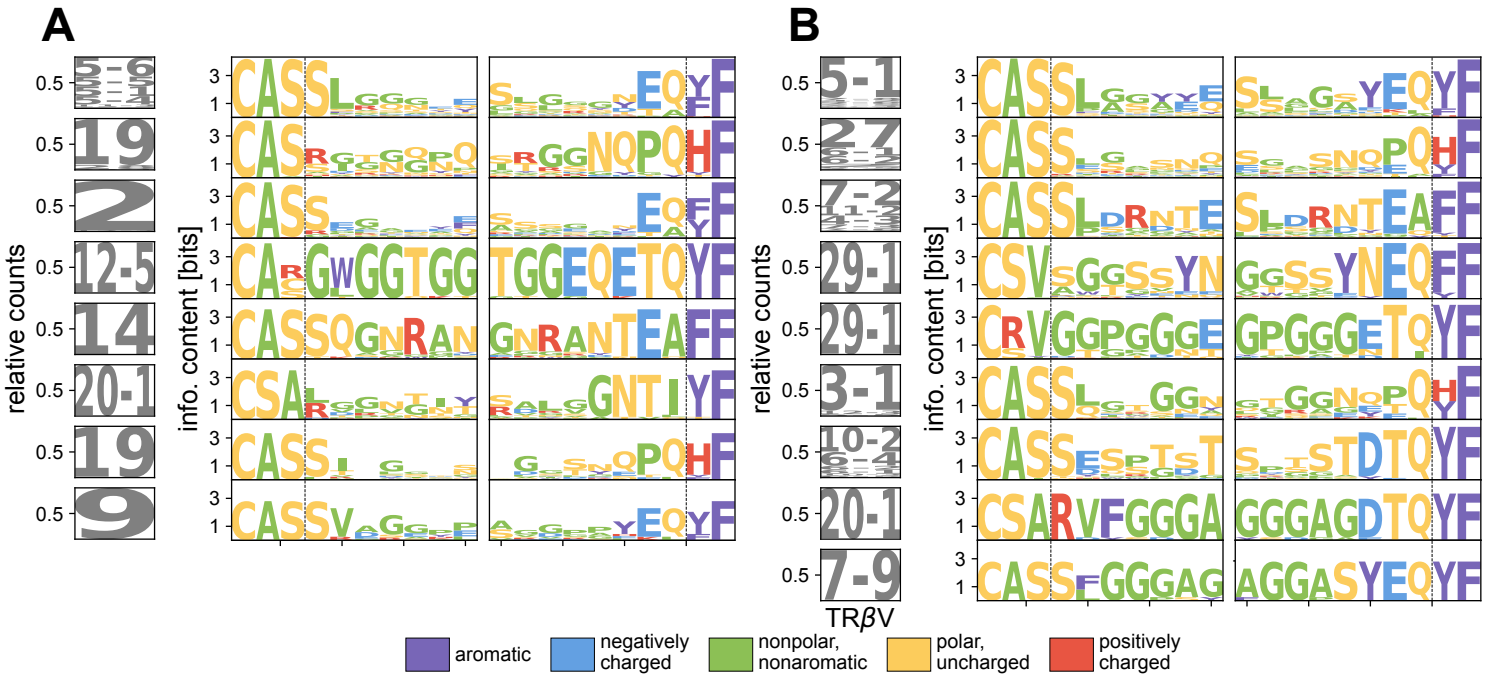


Figure B.3: **Full sequence motifs for significant top rank-50 TCR-OT clusters.** The plot resembles Fig. 3.5 except that sequence logos are constructed using all members within the TCR-OT cluster.

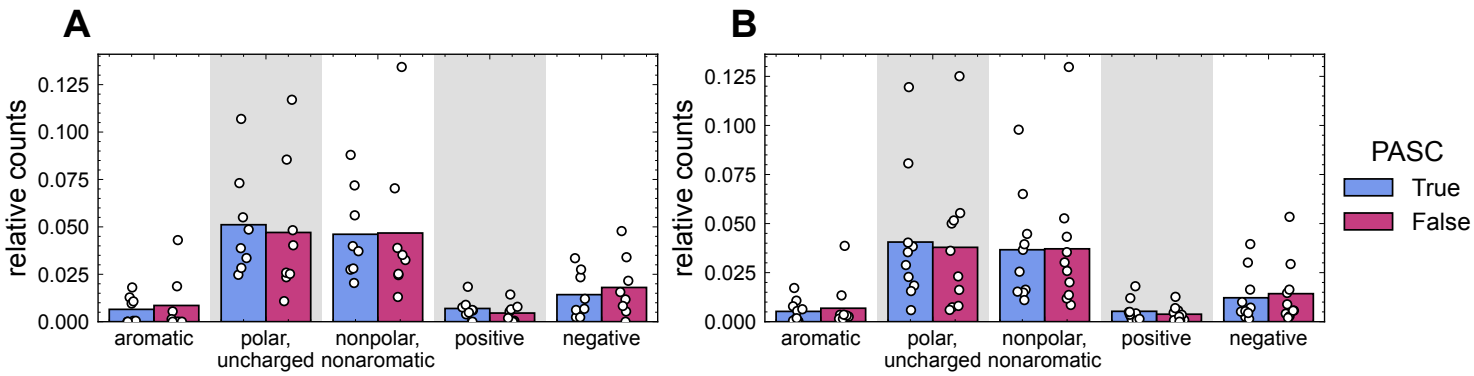


Figure B.4: **Physicochemical properties of rank-50 TCR-OT clusters. (A)** The relative counts of each amino acid characteristic for the logos shown in Fig. 3.5 for the PASC+ and PASC− cohorts (colors). **(B)** As in (A) but for the full sequence logos shown in Fig. B.3.

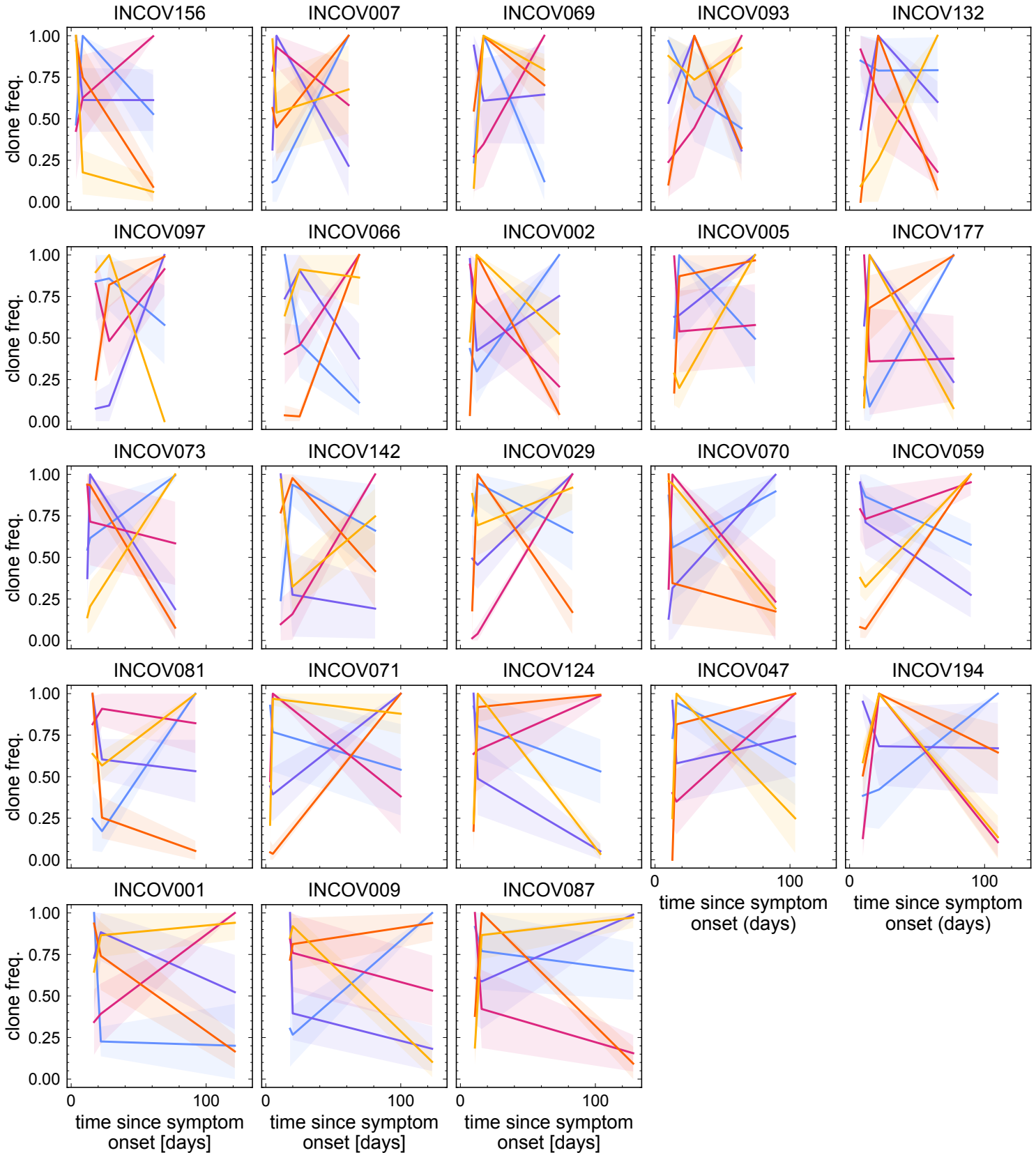

INCOV156
INCOV007
INCOV069
INCOV093
INCOV132
INCOV097
INCOV066
INCOV002
INCOV005
INCOV177
INCOV073
INCOV142
INCOV029
INCOV070
INCOV059
INCOV081
INCOV071
INCOV124
INCOV047
INCOV194
INCOV001
INCOV009
INCOV087
clone freq.
0.00
0.25
0.50
0.75
1.00
0
100
time since symptom onset [days]
time since symptom onset (days)

Figure B.5: **Dynamical modes of clonal frequency trajectories for PASC+ cohort.** The modes of clonal trajectories from the top rank-1000 clones within an individual identified using PCA and hierarchical clustering. Lines indicate average clonal frequency within each mode (colors), and shading indicates regions containing one standard deviation of variation across the mode. Plots are ordered from left-to-right and top-to-bottom using the time at which the individual's repertoire was last sampled.

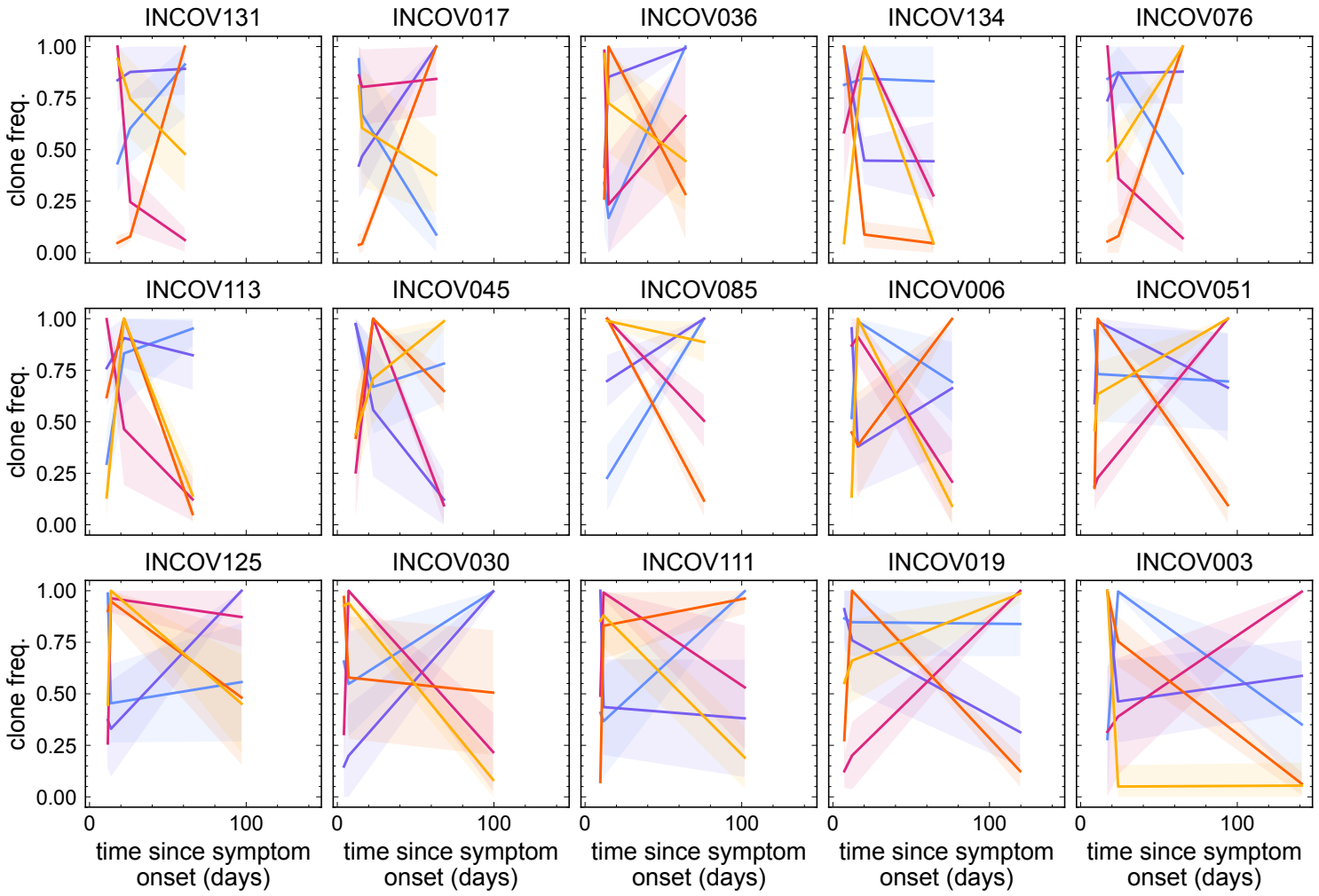


Figure B.6: **Dynamical modes of clonal frequency trajectories for PASC− cohort.** The modes of clonal trajectories from the top rank-1000 clones within an individual identified using PCA and hierarchical clustering. Lines indicate average clonal frequency within each mode (colors), and shading indicates regions containing one standard deviation of variation across the mode. Plots are ordered from left-to-right and top-to-bottom using the time at which the individual's repertoire was last sampled.

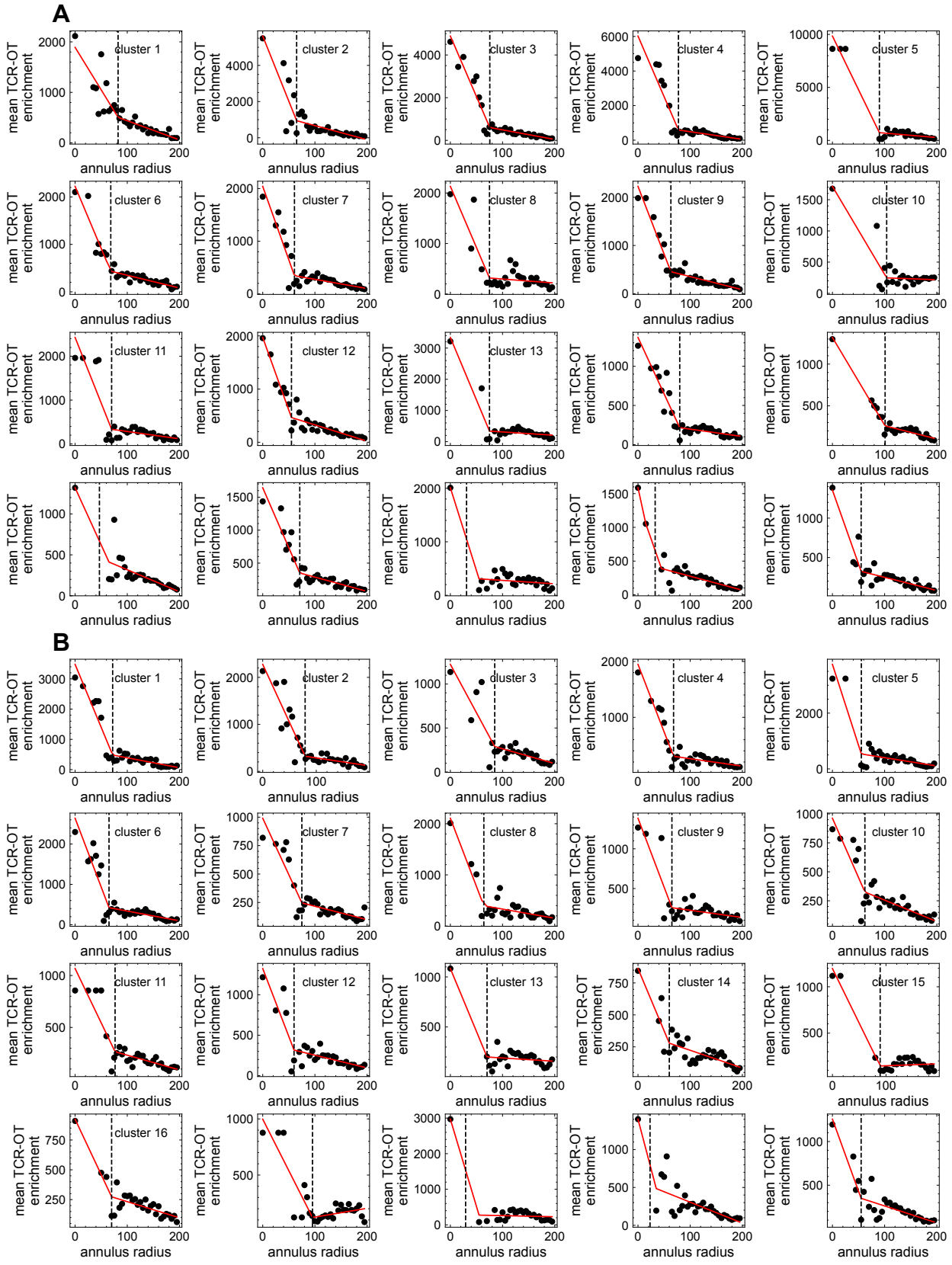

A
B
mean TCR-OT enrichment
annulus radius
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
cluster 9
cluster 10
cluster 11
cluster 12
cluster 13
cluster 14
cluster 15
cluster 16

Figure B.7: **TCR-OT cluster breakpoint analyses on contracting repertoires** The mean TCR-OT enrichment vs annulus radius (given in units TCRdist) for cluster detection in the contracting PASC+ repertoire **(A)** and contracting PASC− **(B)**. The black dots show the observed statistic at within the annulus, and the red line shows the segmented linear fit (Methods). The breakpoint, shown as a vertical black, dashed line, in each segmented linear regression determines the cluster radius. Regressions associated with clusters used in downstream analyses are annotated with their cluster size rank, as in Fig. 3.4H,I.

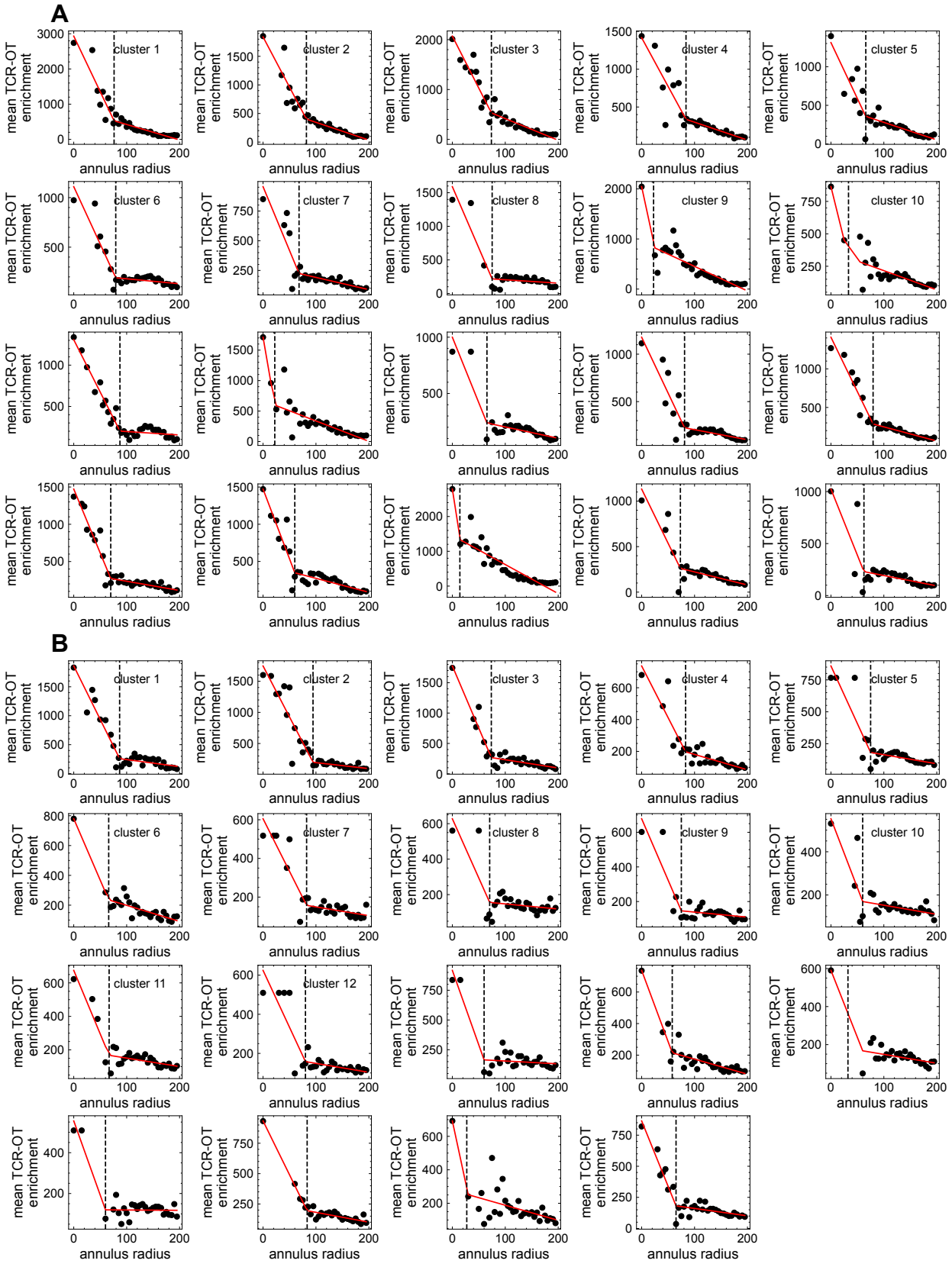

A
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
cluster 9
cluster 10
B
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
cluster 9
cluster 10
cluster 11
cluster 12
mean TCR-OT enrichment
annulus radius

Figure B.8: **TCR-OT cluster breakpoint analyses on expanding repertoires** The mean TCR-OT enrichment vs annulus radius (given in units TCRdist) for cluster detection in the expanding PASC+ repertoire **(A)** and expanding PASC– **(B)**. The black dots show the observed statistic at within the annulus, and the red line shows the segmented linear fit (Methods). The breakpoint, shown as a vertical black, dashed line, in each segmented linear regression determines the cluster radius. Regressions associated with clusters used in downstream analyses are annotated with their cluster size rank, as in Fig. 3.4J,K.

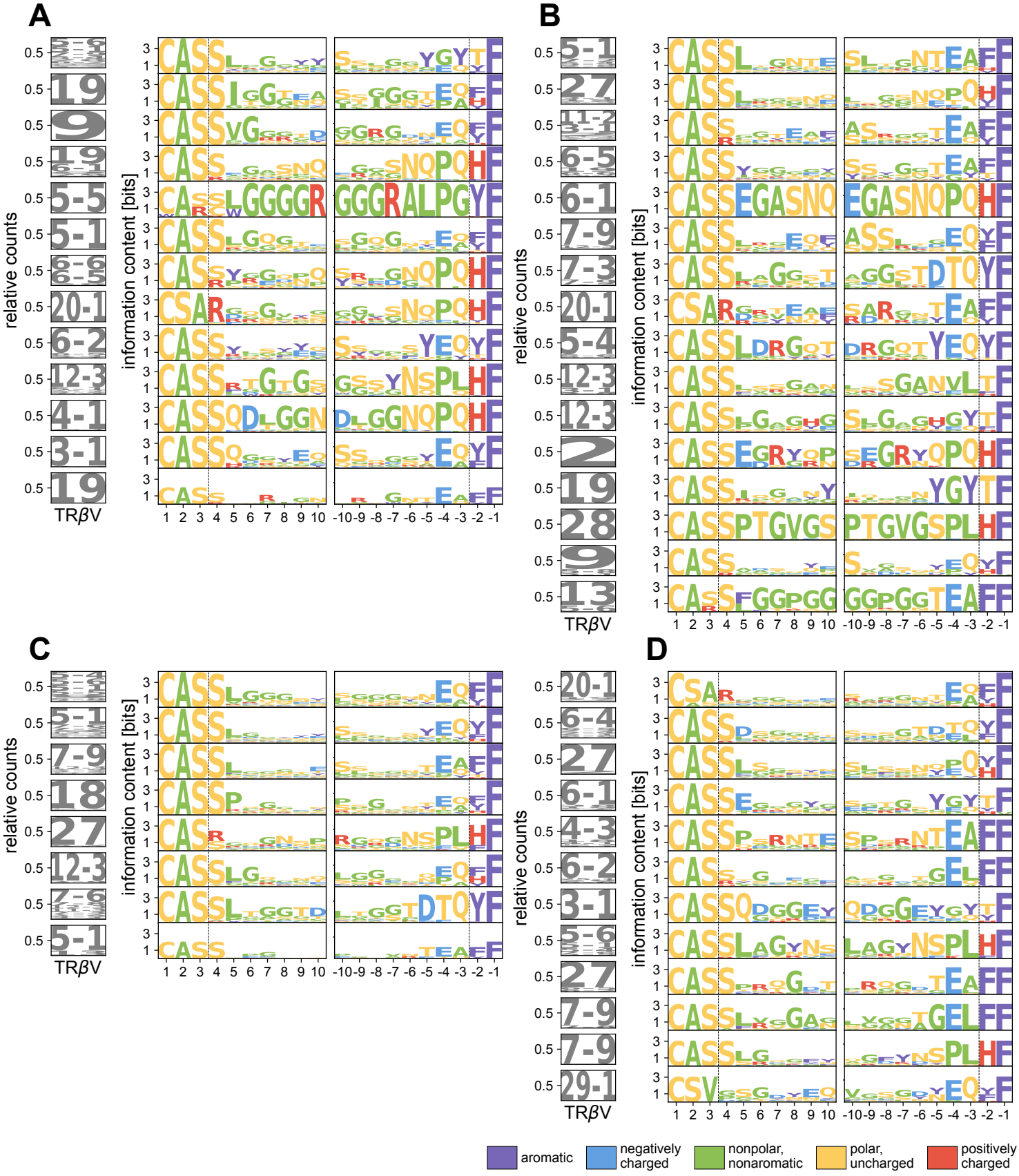

A
B
C
D
relative counts
information content [bits]
TRβV
1 2 3 4 5 6 7 8 9 10
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1
aromatic
negatively charged
nonpolar, nonaromatic
polar, uncharged
positively charged

Figure B.9: **Full sequence motifs for significant dynamical TCR-OT clusters.** The plot resembles Fig. 3.8 except that sequence logos are constructed using all members within the TCR-OT clusters from the **(A)** PASC+ contracted repertoire, **(B)** PASC− contracted repertoire, **(C)** PASC+ contracted repertoire, and **(D)** PASC− contracted repertoire. Sequence logos are shown only for clusters with at least 6 sequences.

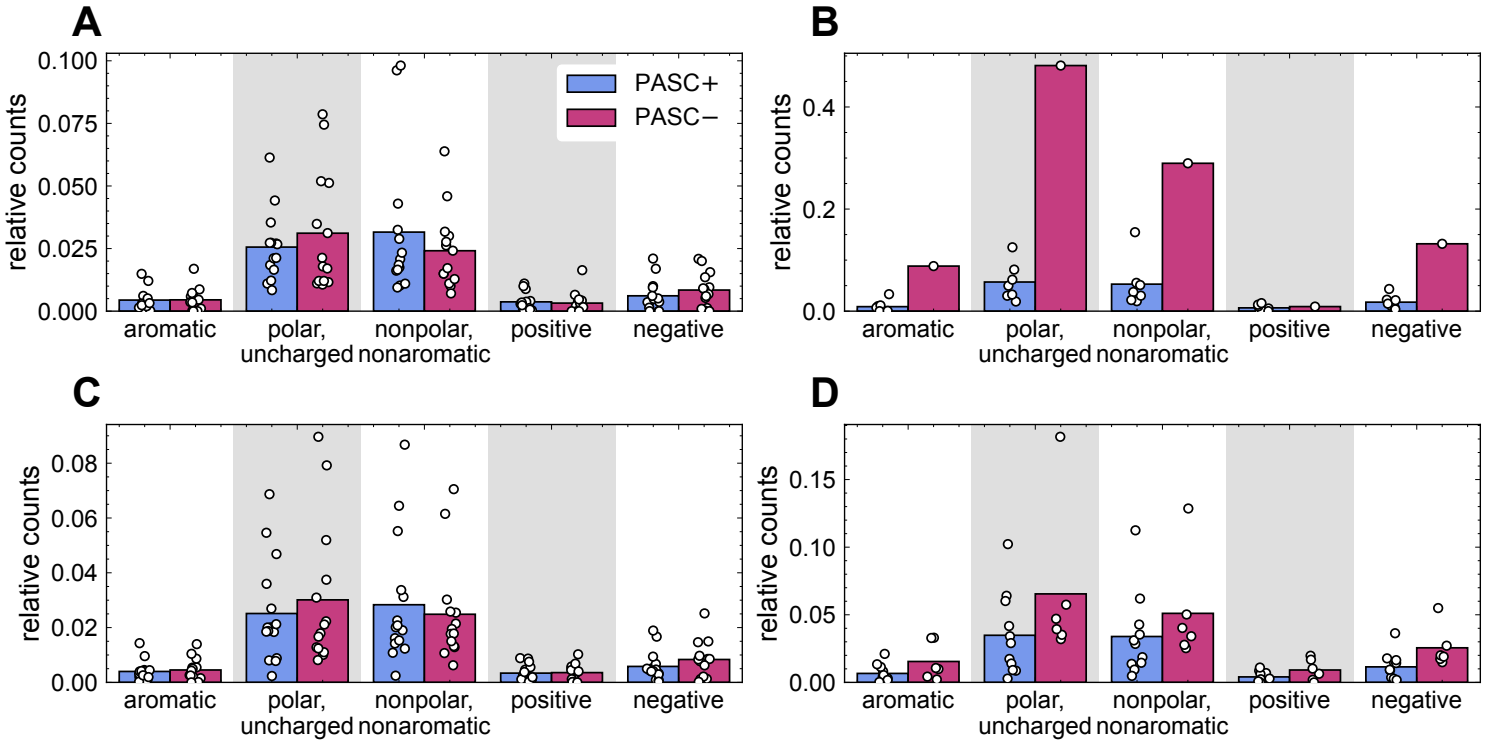


Figure B.10: **Physicochemical properties of dynamical TCR-OT clusters. (A, B)** The relative counts of each amino acid characteristic for the logos shown in Fig. 3.8 for the TCR-OT clusters constructed from the contracted (A) and expanded (B) PASC+ and PASC− cohorts (colors). **(C, D)** As in (A,B) but for the full sequence logos shown in Fig. B.9.

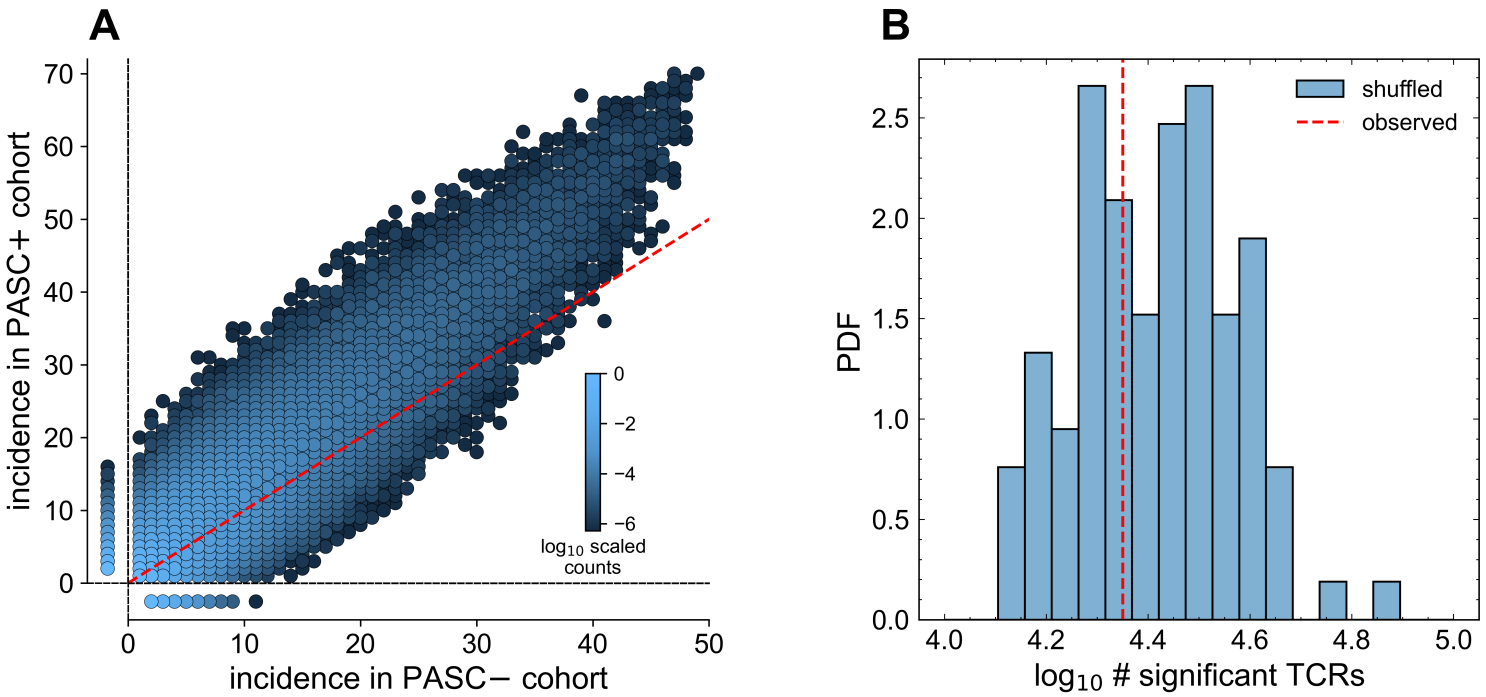


Figure B.11: **Fisher exact test detects no meaningful PASC-associated TCRβs. (A)** The scatter plot shows the number of TCRβ amino acid sequences shared in the PASC− cohort (x-axis) and PASC+ cohort (y-axis), with the colors of the dots indicating the logarithm of the frequency normalized by the maximum frequency. Regions bounded by dashed lines show clones present in only the PASC− cohort (bottom) or PASC+ cohort (left). **(B)** The number of significant clones found from applying one-tailed Fisher exact tests on the observed dataset is shown as a red dashed line, and the distribution shows the number of significant clones found from applying one-tailed Fisher exact tests on datasets with shuffled PASC status.

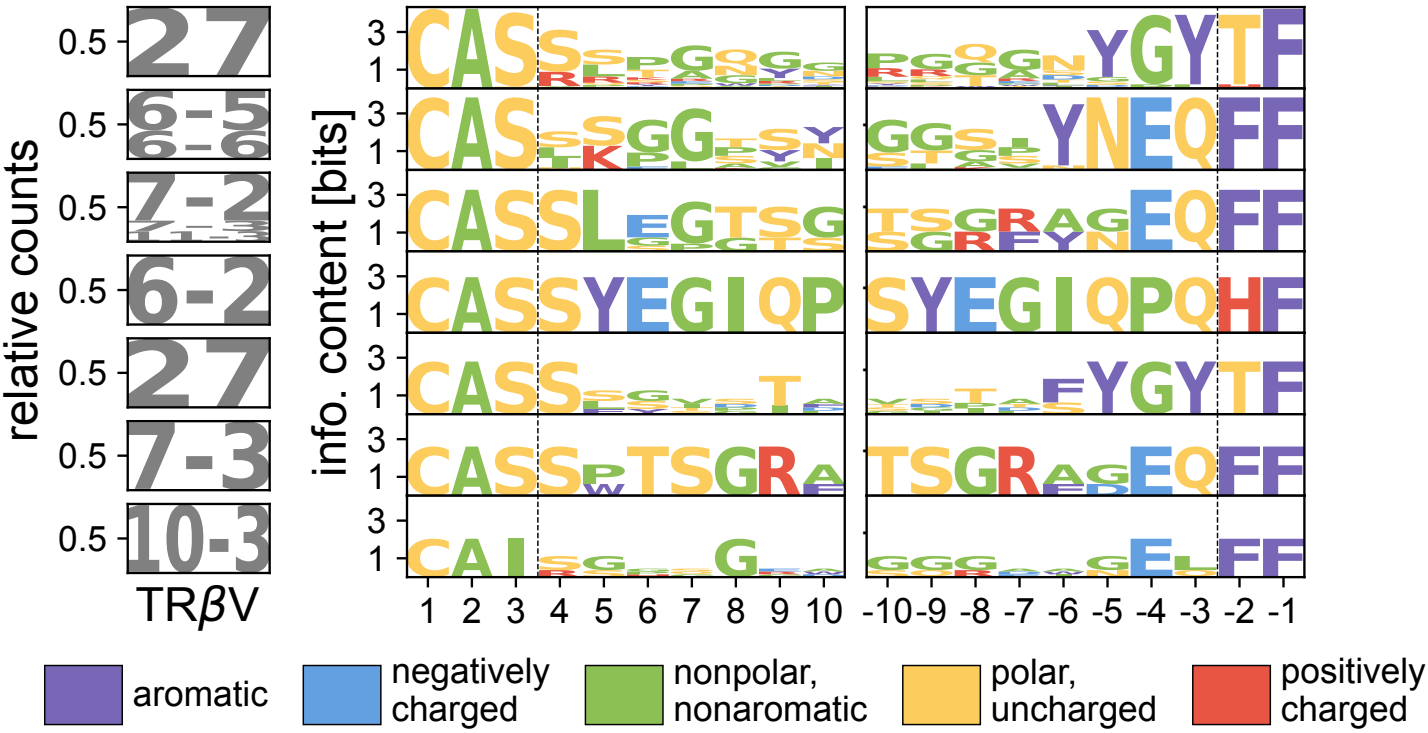


Figure B.12: **Sequence logos of atypical rare, shared clusters.** TCR-OT clusters of rare, shared sequences in the PASC+ cohort atypical with respect to rare, shared sequences in the PASC−. Clusters are plotted only if they have at least 6 sequences. No clusters from the PASC− met the criterion of having at least 20% of their constituents be significantly enriched and therefore are not shown.

PASC status	sex	time bin	t statistic	Bonferroni-corrected p-value
PASC−	female	$t \leq 20$	$t(377) = 21.5$	$p = 2.71 \times 10^{-66}$
PASC−	female	$20 < t \leq 60$	$t(374) = 20.63$	$p = 1.58 \times 10^{-62}$
PASC−	female	$60 < t$	$t(363) = 15.13$	$p = 2.62 \times 10^{-39}$
PASC−	male	$t \leq 20$	$t(370) = 18.57$	$p = 1.06 \times 10^{-53}$
PASC−	male	$20 < t \leq 60$	$t(365) = 16.71$	$p = 8.36 \times 10^{-46}$
PASC−	male	$60 < t$	$t(363) = 15.84$	$p = 3.33 \times 10^{-42}$
PASC+	female	$t \leq 20$	$t(384) = 23.65$	$p = 9.42 \times 10^{-76}$
PASC+	female	$20 < t \leq 60$	$t(383) = 23.32$	$p = 2.72 \times 10^{-74}$
PASC+	female	$60 < t$	$t(373) = 19.34$	$p = 4.91 \times 10^{-57}$
PASC+	male	$t \leq 20$	$t(377) = 21.39$	$p = 7.59 \times 10^{-66}$
PASC+	male	$20 < t \leq 60$	$t(367) = 16.97$	$p = 6.22 \times 10^{-47}$
PASC+	male	$60 < t$	$t(371) = 19.23$	$p = 1.67 \times 10^{-56}$

Table B.1: INCOV TCRβ CDR3 length statistics compared to CMV− Emerson cohort.

PASC status	dynamic	null repertoire	Z statistic	Bonferroni-corrected p-value
PASC+	contracted	most abundant	10.02	$p = 4.99 \times 10^{-23}$
PASC+	contracted	random	15.47	$p = 2.31 \times 10^{-53}$
PASC+	expanded	most abundant	1.84	$p = 0.27$
PASC+	expanded	random	8.51	$p = 6.68 \times 10^{-17}$
PASC–	contracted	most abundant	13.43	$p = 1.69 \times 10^{-40}$
PASC–	contracted	random	16.88	$p = 2.33 \times 10^{-63}$
PASC–	expanded	most abundant	-0.74	$p = 1.0$
PASC–	expanded	random	4.53	$p = 2.38 \times 10^{-5}$

Table B.2: Dynamic repertoires and matching into the MIRA database.

Appendix C

SUPPLEMENT FOR CHAPTER 4

C.0.1 Data and code availability

BCR repertoire raw FASTQ data were obtained from: PRJNA543982. All code for data processing and statistical analysis can be found at: https://github.com/zacmon/anarep and https://github.com/zacmon/telegraph˙treetime.

C.0.2 Data preprocessing

All FASTQ data were downloaded from the SRA using the SRA toolkit. From there, preprocessing was performed as in ref. [200] (see Appendix A); however, the primers used for sequencing the repertoires of this cohort were proprietary and thus we exclude any preprocessing steps pertaining to primers. We used pRESTO (version 0.7.0) [307] to assemble paired-end reads, remove sequences with a mean quality score less than 30, and collapse duplicate sequences into unique sequences. The small fraction of paired-end reads that overlapped were assumed to be anomalous and were discarded from the analysis. Additionally, after preprocessing with pRESTO, we discarded unique reads that contained ambiguous calls (N's) in their receptor sequence. We then performed two rounds of error correction on sequences that passed the quality control check, with the algorithm given in ref. [200] (see Appendix A).

For each individual, error-corrected sequences from all time points and technical replicates were pooled and annotated by abstar (version 0.4.4) [30]. We processed the output of abstar, which included the estimated IGHV gene/allele, IGHJ gene/allele, location of the HCDR3 region, and an inferred naive sequence (germline before hypermutation). Sequences which had indels outside of the HCDR3 were discarded.

To identify BCR clonal lineages, we first grouped sequences by their assigned IGHV gene, IGHJ gene, and HCDR3 length and then used single-linkage clustering with a threshold of 85% Hamming distance. A similar threshold has been suggested previously by [103] to identify BCR lineages. The inferred progentior sequence of the lineage was taken as the most common inferred naive sequence of the receptors in the lineage.

C.0.3 Tree reconstruction

For each lineage, multiple sequence alignments (MSAs) were constructed using an in-house algorithm. Observed BCRs and the inferred progenitor were aligned by their CDR3 sequences. Ambiguous nucleotides (N's) were added to pad the beginning and ending of sequences such that all sequences had the same length and the integrity of the amino acid sequences was preserved. We inferred phylogenies using the GY94 substitution model [97] as implemented in IgPhyML [122].

C.0.4 Probabilities of the number of switches

For completeness, we include the joint probability distributions for the number of periods n and the final state ψ_f by marginalizing out x in Eqs.4.25-4.28. Let

$$C = \frac{\sqrt{\pi} t^{n+1/2} (|\alpha_- - \alpha_+|)^{1/2-n} (\alpha_- \alpha_+)^n e^{-t(\alpha_- + \alpha_+)/2}}{n!} \tag{C.1}$$

for the following equations. If $\alpha_- \neq \alpha_+$,

$$P(n, \text{on}|\text{on}, \tau; \alpha_-, \alpha_+) = C \left[\frac{I_{n-1/2}(t|\alpha_- - \alpha_+|/2) + (-1)^{\mathbb{1}_{\alpha_- > \alpha_+}} I_{n+1/2}(t|\alpha_- - \alpha_+|/2)}{2} \right], \tag{C.2}$$

$$P(n, \text{off}|\text{off}, \tau; \alpha_-, \alpha_+) = C \left[\frac{I_{n-1/2}(t|\alpha_- - \alpha_+|/2) + (-1)^{\mathbb{1}_{\alpha_- < \alpha_+}} I_{n+1/2}(t|\alpha_- - \alpha_+)/2|}{2} \right], \tag{C.3}$$

$$P(n, \text{off}|\text{on}, \tau; \alpha_-, \alpha_+) = \frac{\alpha_- C}{|\alpha_- - \alpha_+|} I_{n+1/2}(t|\alpha_- - \alpha_+|/2), \tag{C.4}$$

$$P(n, \text{on}|\text{off}, \tau; \alpha_-, \alpha_+) = \frac{\alpha_+ C}{|\alpha_- - \alpha_+|} I_{n+1/2}(t|\alpha_- - \alpha_+|/2), \tag{C.5}$$

If $\alpha_- = \alpha_+$, then one switching process is indistinguishable from the other,

$$P(n, \psi_f | \psi_i, \tau; \alpha_-, \alpha_+) = \text{Pois}(2n + (1 - \delta_{\psi_i,\psi_j}) | \alpha_-, \tau). \tag{C.6}$$

Normalizing by the appropriate entry of the telegraph propagator (Eq. **??**) gives $P(n|\psi_i, \psi_f, \tau; \alpha_-, \alpha_+)$.

C.0.5 Statistics of distributions characterizing the inter-arrival time τ

The expected time of a branch $\langle\tau\rangle$ (with uniform prior given ψ_i) given m mutations is

$$\langle\tau\rangle_{\psi_i=\psi_f=\text{on}} = \frac{m+1}{\mu}\frac{\alpha_- + \alpha_+}{\alpha_+} \tag{C.7}$$

$$\langle\tau\rangle_{\psi_i\neq\psi_f} = \frac{(\alpha_- + \alpha_+)(m+1) + \mu}{\mu + \alpha_-} \tag{C.8}$$

$$\langle\tau\rangle_{\psi_i=\psi_f=\text{off},\, m=0} = \frac{\mu + \alpha_-}{\mu\alpha_+} + \frac{\alpha_-}{\alpha_-\mu + \mu^2} \tag{C.9}$$

$$\langle\tau\rangle_{\psi_i=\psi_f=\text{off},\, m>0} = \frac{(\alpha_- + \alpha_+)(m+1) + 2\mu}{\mu\alpha_+} \tag{C.10}$$

The variances of the distribution of inter-arrival times are

$$\text{Var}(\tau)_{\psi_i=\psi_f=\text{on}} = \frac{m+1}{\mu^2}\frac{(\alpha_- + \alpha_+)^2 + 2\mu\alpha_-}{\alpha_+^2} \tag{C.11}$$

$$\text{Var}(\tau)_{\psi_i\neq\psi_f} = \frac{(m+1)\left((\alpha_- + \alpha_+)^2 + 2\mu\alpha_-\right) + \mu^2}{\mu^2\alpha_+^2} \tag{C.12}$$

$$\text{Var}(\tau)_{\psi_i=\psi_f=\text{off},\, m=0} = \frac{\mu^2 + \alpha_-^2 + 2\alpha_-(\mu + \alpha_+)}{\mu^2\alpha_+^2} + \frac{1}{\mu^2} - \frac{1}{(\mu + \alpha_-)^2} \tag{C.13}$$

$$\text{Var}(\tau)_{\psi_i=\psi_f=\text{off},\, m>0} = \frac{(m+1)(\alpha_- + \alpha_+)^2 + 2(m+1)\mu\alpha_- + 2\mu^2}{\mu^2\alpha_+^2} \tag{C.14}$$

In the limit of $\alpha_+ \gg \alpha_-$, we observed that the moments for the $\psi_i = \psi_f = \text{on}$ distributions approach those of the Erlang distribution as expected.

C.0.6 Simulations

To simulate the inter-arrival times on a tree, we used the insight we gained when deriving the inter-arrival time distributions Eqs. 4.62-4.75, i.e., the inter-arrival time distribution for

a branch starting and ending on is a convolution of gamma distributions with binomially distributed shape parameters. The inter-arrival time for a branch that starts on and ends off is obtained by adding a hypoexponentially distributed random variable to the time sampled from the aforementioned distribution for starting and ending on. Concretely, suppose a branch has $m \geq 0$ mutations and begins and ends in the active state. Let $\tilde{m} = m + 1$. To simulate the time on the branch, we first sample the shapes of the gamma distributions. The probability associated with the slower exponential rate γ_- is

$$p_- = \frac{d\gamma_-}{d\mu}\frac{\mu}{\gamma_-}. \tag{C.15}$$

Then

$$n_- \sim \text{Binom}(\tilde{m}, p_-), \quad n_+ = \tilde{m} - n_-. \tag{C.16}$$

Finally,

$$t_{\text{on,on}}(m) \sim \text{Gamma}(n_-, \gamma_-) + \text{Gamma}(n_+, \gamma_+). \tag{C.17}$$

To simulate the time on a branch that starts on and ends off, we sample

$$t_{\text{on,off}}(m) \sim \text{Exp}(\gamma_-) + \text{Exp}(\gamma_+) + t_{\text{on,on}}(m) \tag{C.18}$$

Given a tree annotated with mutations, the times can be simulated in a preorder traversal.

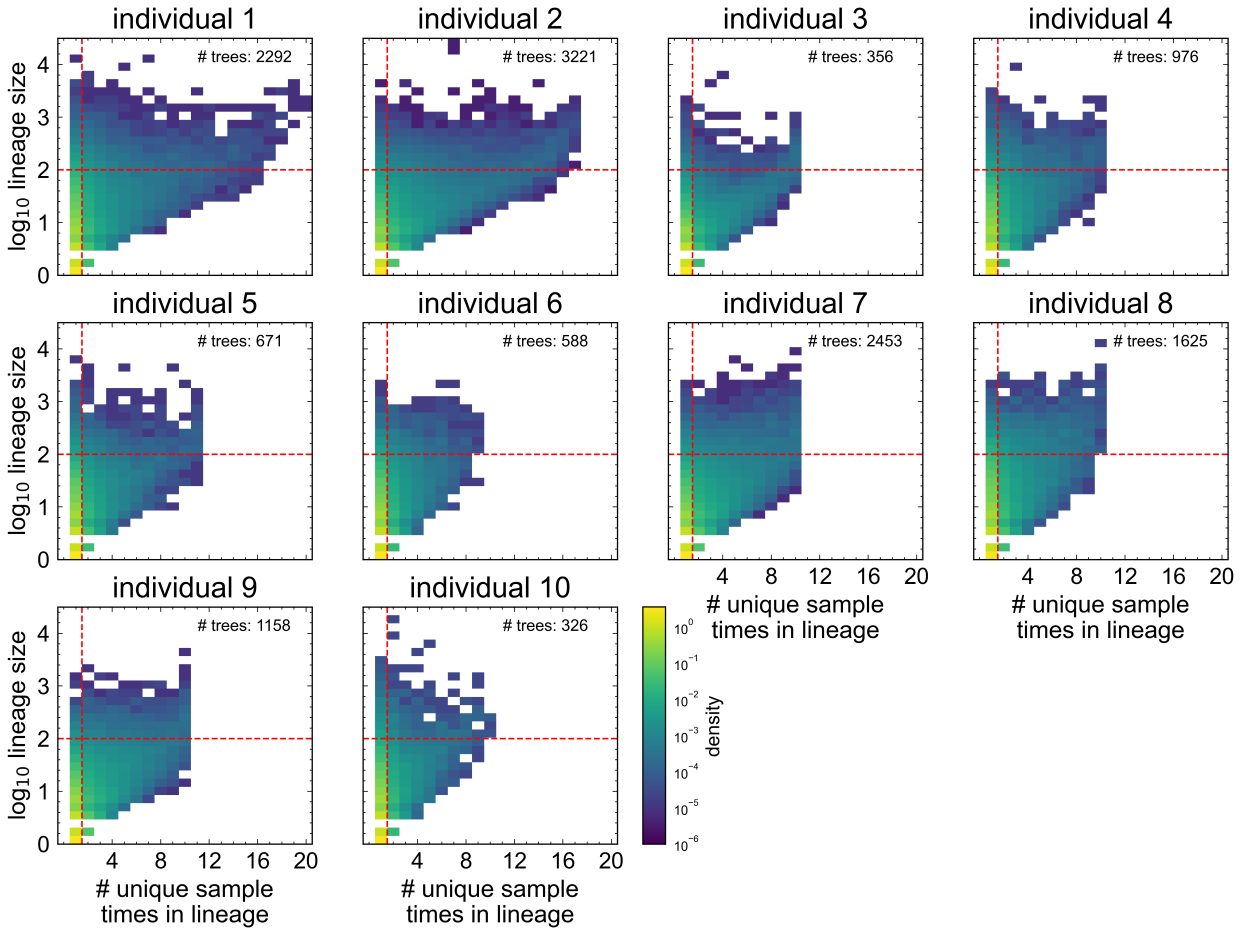


Figure C.1: **Lineage distributions in each individual.** The plot is as in Fig. 4.1B but for each individual in the cohort.

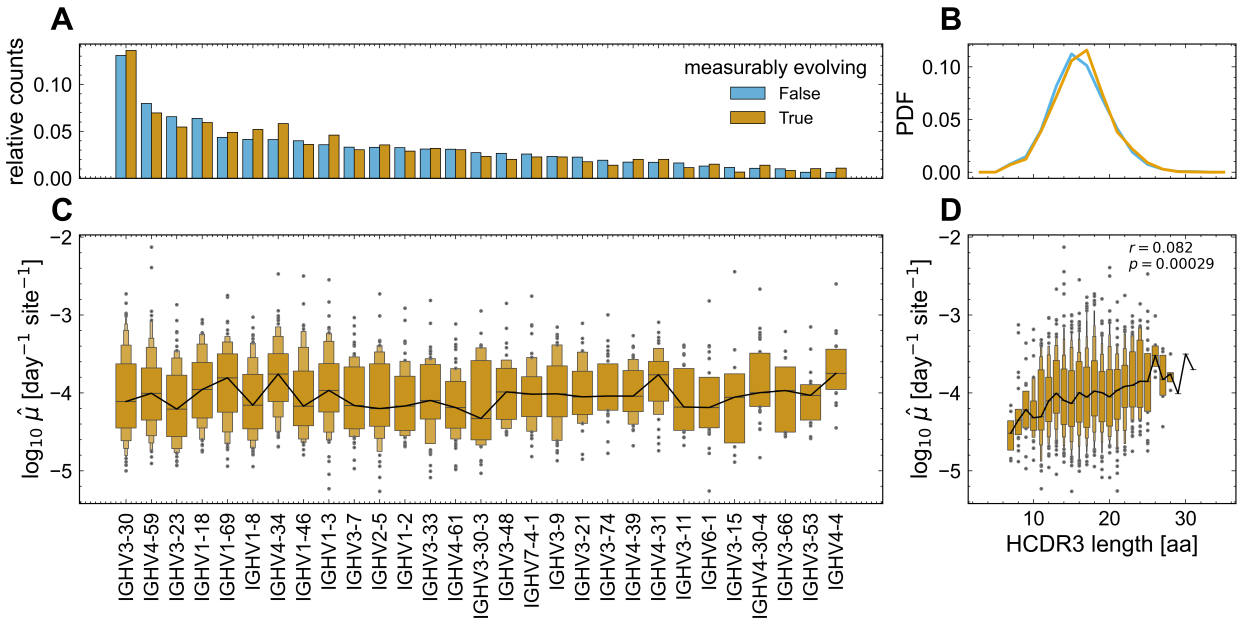


Figure C.2: **Receptor features of measurably evolved lineages (A)** The bar plot shows the distribution of IGHV gene usages in lineages detected as measurably evolved and those that were not (colors). **(B)** The line plots show the distribution of HCDR3 lengths in the lineages not detected as measurably evolved and those that were (colors). The shift to longer HCDR3 lengths in the measurably evolving lineages is significant (see main text). **(C)** The distribution of root-to-tip inferred substitution rates for each IGHV gene of the measurably evolving lineages is shown as a boxenplot. The line shows how the median substitution rate of each IGHV gene changes by descending in IGHV gene usage (from left to right). **(D)** The distribution of the root-to-tip inferred substitution rates for each HCDR3 amino acid length observed in the set of measurably evolving lineages is shown as a boxenplot. The line shows how the median substitution rate changes as the HCDR3 length is increased. The Pearson r correlation and its associated p-value between the substitution rate and HCDR3 length is annotated in the upper right of the plot.

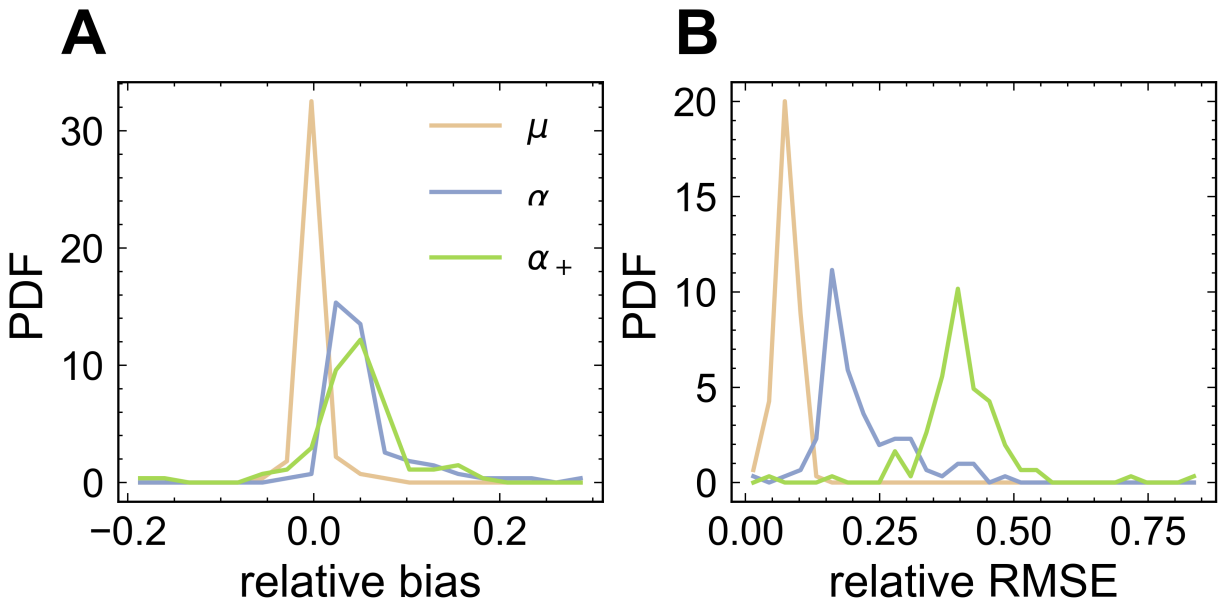


Figure C.3: **Relative bias and RMSE for star phylogeny analysis. (A)** The line plots show the histograms of relative bias for the inferred rates (colors). **(B)** The line plots show the histograms of relative RMSE for the inferred rates (colors).

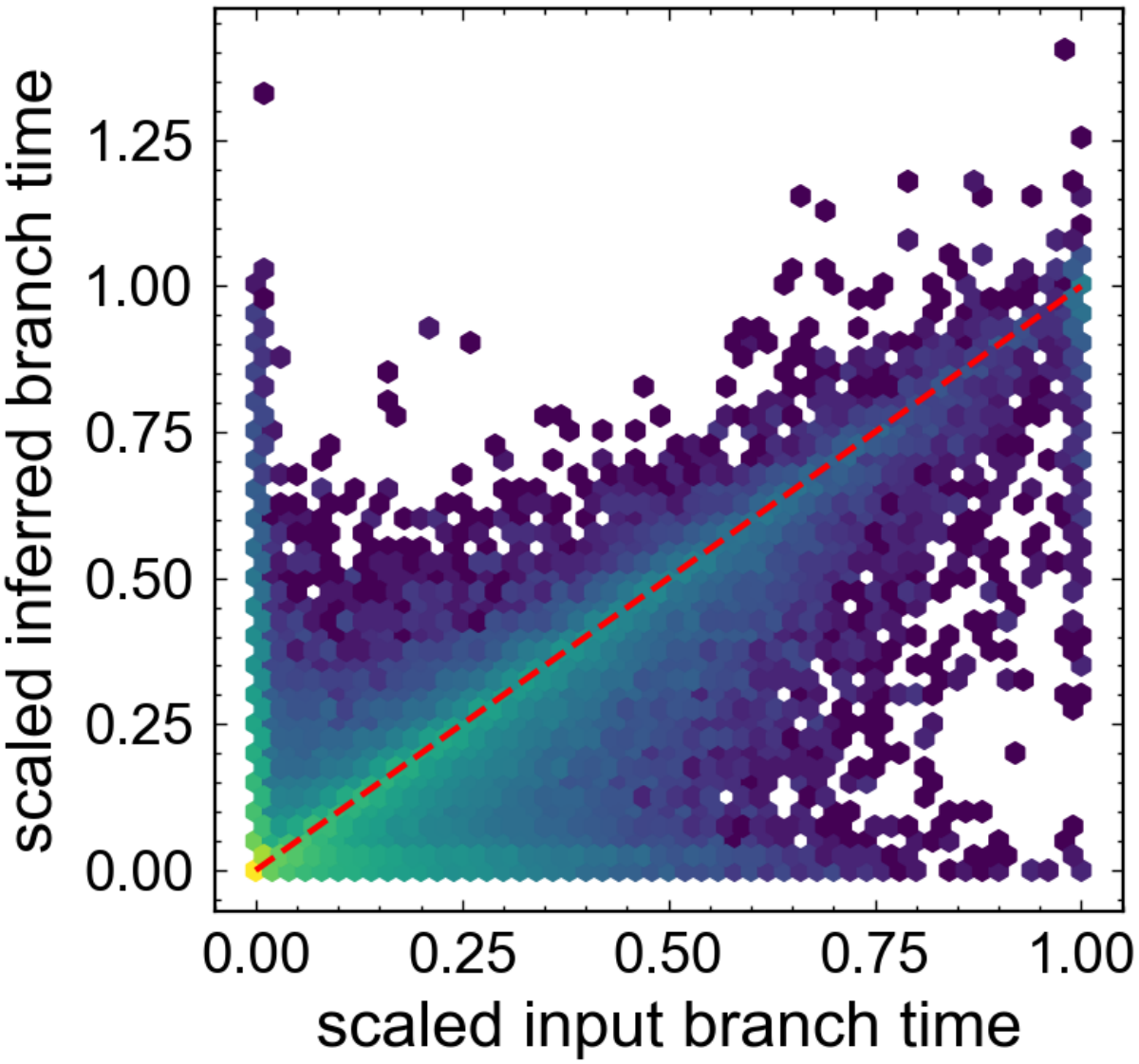


Figure C.4: **Comparison of estimated branch lengths in units of time.** The hexbin plot shows the scaled inferred branch time versus the scaled input branch time for each simulation. Given a simulation, both the inferred and true branch times are scaled by the maximum true branch time to ensure branch times are comparable across simulations. The density of the hexbins was performed using a logarithmic scale. The diagonal red, dashed line has unit slope. The Pearson correlation between the estimated and true branch times is 0.78.

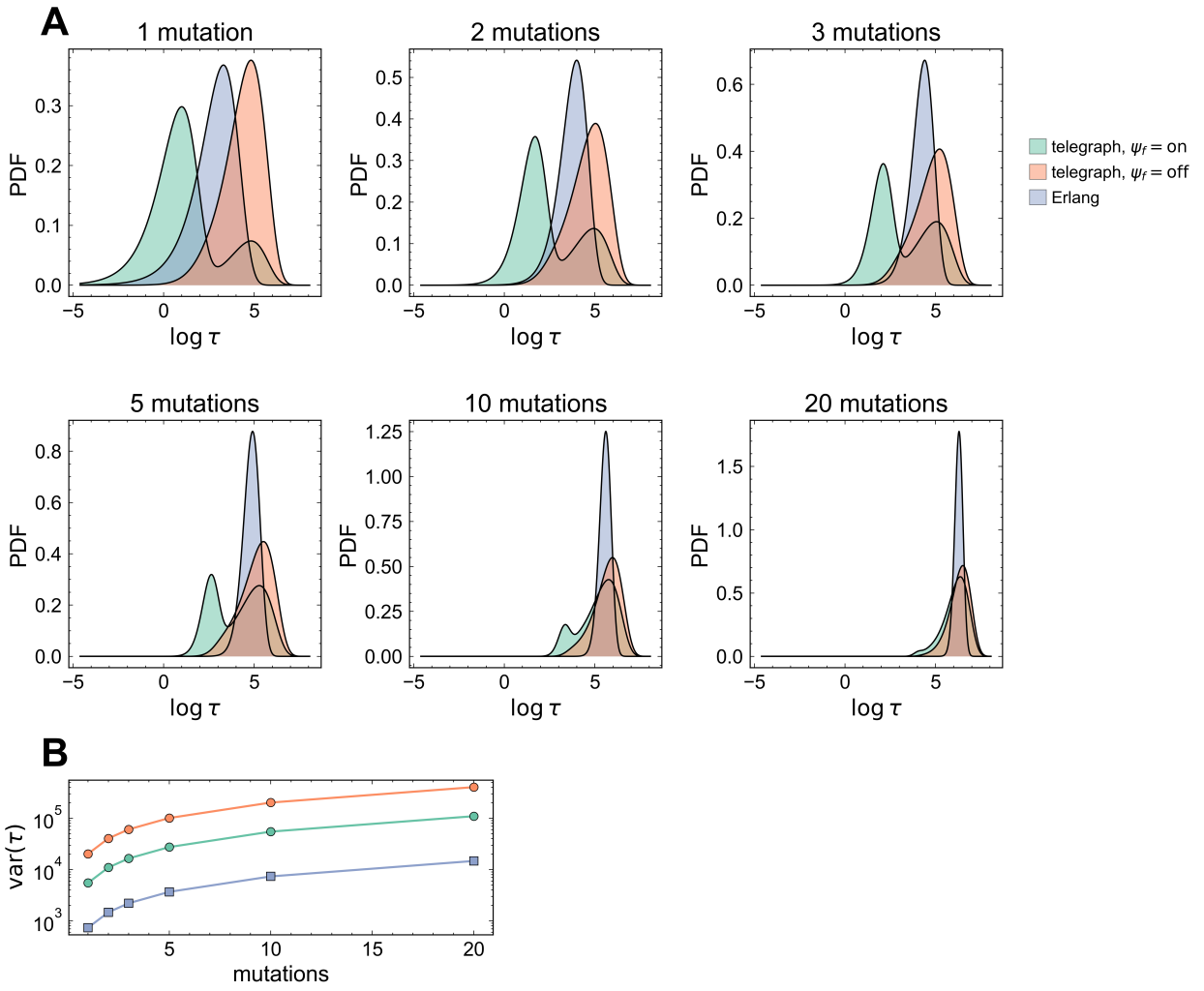


Figure C.5: **Comparison of waiting time distributions. (A)** The distributions of the log of the waiting time τ until m mutations occur are shown as histograms for the mutation-telegraph distributions (Eqs. 4.62, 4.75) and Erlang distribution (Eq. 4.61) (colors). The titles of each subplot indicate m. Because the x-axis is plotted as $\log \tau$, the PDFs computed using the likelihood of τ using the aforementioned equations are multiplied τ due to change-of-variables for proper normalization. We set $\mu = 0.3$, $\alpha_- = 1/14$, and $\alpha_+ = 1/100$. The rate used for the Erlang distribution is $\mu_{\text{pois}} = \mu\alpha_+/(\alpha_- + \alpha_+) = 0.0368$. **(B)** The line plot shows how the variance of τ changes as a function of the number mutations m. We observe that the telegraph distributions have orders of magnitude higher variances than the Erlang distribution, resulting in high uncertainty when inferring the node dates.

BIBLIOGRAPHY

[1] 10x Genomics. A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. 2019.

[2] Kathleen Abadie, Elisa C Clark, Rajesh M Valanparambil, Obinna Ukogu, Wei Yang, Riza M Daza, Kenneth KH Ng, Jumana Fathima, Allan L Wang, Judong Lee, et al. Reversible, tunable epigenetic silencing of tcf1 generates flexibility in the t cell memory decision. *Immunity*, 57(2):271–286, 2024.

[3] Maria Francesca Abbate, Thomas Dupic, Emmanuelle Vigne, Melody A Shahsavarian, Aleksandra M Walczak, and Thierry Mora. Computational detection of antigen-specific b cell receptors following immunization. *Proceedings of the National Academy of Sciences*, 121(35):e2401058121, 2024.

[4] Robert K Abbott and Shane Crotty. Factors in b cell competition and immunodominance. *Immunological reviews*, 296(1):120–131, 2020.

[5] Nika Abdollahi, Lucile Jeusset, Anne de Septenville, Frederic Davi, and Juliana Silva Bernardes. Reconstructing b cell lineage trees with minimum spanning tree and genotype abundances. *BMC bioinformatics*, 24(1):70, 2023.

[6] Munir Akkaya, Kihyuck Kwak, and Susan K Pierce. B cell memory: building two walls of protection against pathogens. *Nature Reviews Immunology*, 20(4):229–238, 2020.

[7] Ziyad Al-Aly and Clifford J Rosen. Long covid and impaired cognition—more evidence and more work to do, 2024.

[8] Ziyad Al-Aly and Eric Topol. Solving the puzzle of long covid. *Science*, 383(6685):830–832, 2024.

[9] Juan Carlos Almagro, Gopalan Raghunathan, Eric Beil, Dariusz J Janecki, Qiang Chen, Thai Dinh, Ann LaCombe, Judy Connor, Mark Ware, Paul H Kim, et al. Characterization of a high-affinity human antibody with a disulfide bridge in the third complementarity-determining region of the heavy chain. *Journal of molecular recognition*, 25(3):125–135, 2012.

[10] Grégoire Altan-Bonnet, Thierry Mora, and Aleksandra M Walczak. Quantitative immunology for physicists. *Physics Reports*, 849:1–83, 2020.

[11] Galit Alter, Tom HM Ottenhoff, and Simone A Joosten. Antibody glycosylation in inflammation, disease and vaccination. In *Seminars in immunology*, volume 39, pages 102–110. Elsevier, 2018.

[12] Shannon M Anderson, Mary M Tomayko, Anupama Ahuja, Ann M Haberman, and Mark J Shlomchik. New markers for murine memory b cells that define mutated and unmutated subsets. *The Journal of experimental medicine*, 204(9):2103, 2007.

[13] Sarah F Andrews, Michael J Chambers, Chaim A Schramm, Jason Plyler, Julie E Raab, Masaru Kanekiyo, Rebecca A Gillespie, Amy Ransier, Sam Darko, Jianfei Hu, et al. Activation dynamics and immunoglobulin evolution of pre-existing and newly generated human memory b cell responses to influenza hemagglutinin. *Immunity*, 51(2):398–410, 2019.

[14] Sarah F Andrews, Yunping Huang, Kaval Kaur, Lyubov I Popova, Irvin Y Ho, Noel T Pauli, Carole J Henry Dunand, William M Taylor, Samuel Lim, Min Huang, et al. Immune history profoundly affects broadly protective b cell responses to influenza. *Science translational medicine*, 7(316):316ra192–316ra192, 2015.

[15] Tatsuya Araki. Replaying life's tape with intraclonal germinal center evolution. 2023.

[16] K Maude Ashby and Kristin A Hogquist. A guide to thymic selection of t cells. *Nature Reviews Immunology*, 24(2):103–117, 2024.

[17] Assia Asrir, Meryem Aloulou, Mylène Gador, Corine Pérals, and Nicolas Fazilleau. Interconnected subsets of memory follicular helper t cells have different effector functions. *Nature communications*, 8(1):847, 2017.

[18] Guy Baele, Mandev S Gill, Paul Bastide, Philippe Lemey, and Marc A Suchard. Markov-modulated continuous-time markov chains to identify site-and branch-specific evolutionary variation in beast. *Systematic biology*, 70(1):181–189, 2021.

[19] Michal Barak, Neta S Zuckerman, Hanna Edelman, Ron Unger, and Ramit Mehr. Igtree©: creating immunoglobulin variable region gene lineage trees. *Journal of immunological methods*, 338(1-2):67–74, 2008.

[20] Joëlle Barido-Sottani, Timothy G Vaughan, and Tanja Stadler. A multitype birth–death model for bayesian inference of lineage-specific birth and death rates. *Systematic biology*, 69(5):973–986, 2020.

[21] Christopher O Barnes, Anthony P West, Kathryn E Huey-Tubman, Magnus AG Hoffmann, Naima G Sharaf, Pauline R Hoffman, Nicholas Koranda, Harry B Gristick, Christian Gaebler, Frauke Muecksch, et al. Structures of human antibodies bound to sars-cov-2 spike reveal common epitopes and recurrent features of antibodies. *Cell*, 182(4):828–842, 2020.

[22] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[23] Antonio Bertoletti, Anthony T Tan, and Nina Le Bert. The t-cell response to sars-cov-2: kinetic and quantitative aspects and the case for their protective role. *Oxford Open Immunology*, 2(1):iqab006, 2021.

[24] Deepta Bhattacharya, Ming T Cheah, Christopher B Franco, Naoki Hosen, Christopher L Pin, William C Sha, and Irving L Weissman. Transcriptional profiling of antigen-dependent murine b cell differentiation and memory formation. *The Journal of Immunology*, 179(10):6808–6819, 2007.

[25] Jochen Blath, Eugenio Buzzoni, Jere Koskela, and Maite Wilke Berenguer. Statistical tools for seed bank detection. *Theoretical Population Biology*, 132:1–15, 2020.

[26] Jochen Blath, Adrián González Casanova, Noemi Kurt, and Maite Wilke-Berenguer. A new coalescent for seed-bank models. 2016.

[27] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. Mixcr: software for comprehensive adaptive immunity profiling. *Nature methods*, 12(5):380–381, 2015.

[28] Alexandra Bortnick and David Allman. What is and what should always have been: Long-lived plasma cells induced by t cell–independent antigens. *The Journal of Immunology*, 190(12):5913–5918, 2013.

[29] Scott D Boyd, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel vdj pyrosequencing. *Science translational medicine*, 1(12):12ra23–12ra23, 2009.

[30] Bryan Briney and Dennis R Burton. Massively scalable genetic analysis of antibody repertoires. *BioRxiv*, page 447813, 2018.

[31] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744):393–397, 2019.

[32] Philip JM Brouwer, Tom G Caniels, Karlijn van der Straten, Jonne L Snitselaar, Yoann Aldon, Sandhya Bangaru, Jonathan L Torres, Nisreen MA Okba, Mathieu Claireaux, Gius Kerster, et al. Potent neutralizing antibodies from covid-19 patients define multiple targets of vulnerability. *Science*, 369(6504):643–650, 2020.

[33] FM Burnet. *The clonal selection theory of acquired immunity*. Vanderbilt University Press, 1959.

[34] FM Burnet. Immunity as an aspect of general biology. *Mechanisms of Antibody Formation. Academia Publishing House of the Czech Academy of Sciences*, pages 15–21, 1960.

[35] Alice R Burton, Stephane M Guillaume, William S Foster, Adam K Wheatley, Danika L Hill, Edward J Carr, and Michelle A Linterman. The memory b cell response to influenza vaccination is impaired in older persons. *Cell reports*, 41(6), 2022.

[36] Thomas Charles Butler, Mehran Kardar, and Arup K Chakraborty. Quorum sensing allows t cells to discriminate between self and nonself. *Proceedings of the National Academy of Sciences*, 110(29):11833–11838, 2013.

[37] Miao Cai, Yan Xie, Eric J Topol, and Ziyad Al-Aly. Three-year outcomes of post-acute sequelae of covid-19. *Nature medicine*, pages 1–10, 2024.

[38] Michael Cai, Seojin Bang, Pengfei Zhang, and Heewook Lee. Atm-tcr: Tcr-epitope binding affinity prediction using a multi-head self-attention model. *Frontiers in immunology*, 13:893247, 2022.

[39] Zuleika Calderin Sollet, Antonia Schäfer, Sylvie Ferrari-Lacraz, Stavroula Masouridi-Levrat, Anne-Claire Mamez, Amandine Pradier, Federico Simonetta, Yves Chalandon, Jean Villard, and Stéphane Buhler. Cmv serostatus and t-cell repertoire diversity 5 years after allogeneic hematopoietic stem cell transplantation. *Leukemia*, 37(4):948–951, 2023.

[40] Glenda Canderan, Lyndsey M Muehling, Alexandra Kadl, Shay Ladd, Catherine Bonham, Claire E Cross, Sierra M Lima, Xihui Yin, Jeffrey M Sturek, Jeffrey M Wilson, et al. Distinct type 1 immune networks underlie the severity of restrictive lung disease after covid-19. *Nature immunology*, 2025.

[41] Yunlong Cao, Bin Su, Xianghua Guo, Wenjie Sun, Yongqiang Deng, Linlin Bao, Qinyu Zhu, Xu Zhang, Yinghui Zheng, Chenyang Geng, et al. Potent neutralizing antibodies against sars-cov-2 identified by high-throughput single-cell sequencing of convalescent patients' b cells. *Cell*, 182(1):73–84, 2020.

[42] Lorenzo Cappello, Wai Tung Jack Lo, Joy Zhichun Zhang, Peiyu Xu, Daniel Barrow, Ishani Chopra, Andrew G Clark, Martin T Wells, and Jaehee Kim. Bayesian phylodynamic inference of population dynamics with dormancy. *bioRxiv*, pages 2025–01, 2025.

[43] George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

[44] Carlo Cervia-Hasler, Sarah C Brüningk, Tobias Hoch, Bowen Fan, Giulia Muzio, Ryan C Thompson, Laura Ceglarek, Roman Meledin, Patrick Westermann, Marc Emmenegger, et al. Persistent complement dysregulation with signs of thromboinflammation in active long covid. *Science*, 383(6680):eadg7942, 2024.

[45] Victor Chardès, Massimo Vergassola, Aleksandra M Walczak, and Thierry Mora. Affinity maturation for an optimal balance between long-term immune coverage and short-term resource constraints. *Proceedings of the National Academy of Sciences*, 119(8):e2113512119, 2022.

[46] Daniel G Chen, Jingyi Xie, Yapeng Su, and James R Heath. T cell receptor sequences are the dominant factor contributing to the phenotype of cd8+ t cells with specificities against immunogenic viral antigens. *Cell reports*, 42(11), 2023.

[47] Xiangyang Chi, Renhong Yan, Jun Zhang, Guanying Zhang, Yuanyuan Zhang, Meng Hao, Zhe Zhang, Pengfei Fan, Yunzhu Dong, Yilong Yang, et al. A neutralizing human antibody binds to the n-terminal domain of the spike protein of sars-cov-2. *Science*, 369(6504):650–655, 2020.

[48] KP Choi and Aihua Xia. Approximating the number of successes in independent trials: Binomial versus poisson. *The Annals of Applied Probability*, 12(4):1139–1148, 2002.

[49] William Chour, Alexander M Xu, Alphonsus HC Ng, Jongchan Choi, Jingyi Xie, Dan Yuan, Diana C DeLucia, Rick A Edmark, Lesley C Jones, Thomas M Schmitt, et al. Shared antigen-specific cd8+ t cell responses against the sars-cov-2 spike protein in hla-a* 02: 01 covid-19 participants. *MedRxiv*, pages 2020–05, 2020.

[50] Mathieu Claireaux, Tom G Caniels, Marlon de Gast, Julianna Han, Denise Guerra, Gius Kerster, Barbera DC van Schaik, Aldo Jongejan, Angela I Schriek, Marloes Grobben, et al. A public antibody class recognizes an s2 epitope exposed on open conformations of sars-cov-2 spike. *Nature communications*, 13(1):4539, 2022.

[51] TM Cover and Joy A Thomas. Elements of information theory, 2006.

[52] David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

[53] Giancarlo Croce, Sara Bobisse, Dana Léa Moreno, Julien Schmidt, Philippe Guillame, Alexandre Harari, and David Gfeller. Deep learning predictions of tcr-epitope interactions reveal epitope-specific chains in dual alpha t cells. *Nature Communications*, 15(1):3211, 2024.

[54] Shane Crotty and Rafi Ahmed. Immunological memory in humans. In *Seminars in immunology*, volume 16, pages 197–203. Elsevier, 2004.

[55] Jason G Cyster and Christopher DC Allen. B cell responses: cell interaction dynamics and decisions. *Cell*, 177(3):524–540, 2019.

[56] Valerie Danesh, Alejandro C Arroliga, James A Bourgeois, Leanne M Boehm, Michael J McNeal, Andrew J Widmer, Tresa M McNeal, and Shelli R Kesler. Symptom clusters seen in adult covid-19 recovery clinic care seekers. *Journal of general internal medicine*, 38(2):442–449, 2023.

[57] Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi HO Nguyen, Katherine Kedzierska, et al. Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature*, 547(7661):89–93, 2017.

[58] Kristian Davidsen and Frederick A Matsen IV. Benchmarking tree and ancestral sequence inference for b cell receptor sequences. *Frontiers in immunology*, 9:2451, 2018.

[59] Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. Long covid: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146, 2023.

[60] Mark M Davis and Pamela J Bjorkman. T-cell antigen receptor genes and t-cell recognition. *Nature*, 334(6181):395–402, 1988.

[61] Renan VH de Carvalho, Jonatan Ersching, Alexandru Barbulescu, Alvaro Hobbs, Tiago BR Castro, Luka Mesin, Johanne T Jacobsen, Brooke K Phillips, Hans-Heinrich Hoffmann, Roham Parsa, et al. Clonal replacement sustains long-lived germinal centers primed by respiratory viruses. *Cell*, 186(1):131–146, 2023.

[62] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences*, 113(2):274–279, 2016.

[63] Andrea DeVries, Sonali Shambhu, Sue Sloop, and J Marc Overhage. One-year adverse outcomes among us adults with post–covid-19 condition vs those without covid-19 in a large commercial insurance database. In *JAMA Health Forum*, volume 4, pages e230010–e230010. American Medical Association, 2023.

[64] William S DeWitt. *Some problems in probabilistic modeling of germline and somatic evolutionary processes*. University of Washington, 2022.

[65] William S DeWitt, Ryan O Emerson, Paul Lindau, Marissa Vignali, Thomas M Snyder, Cindy Desmarais, Catherine Sanders, Heidi Utsugi, Edus H Warren, Juliana McElrath, et al. Dynamics of the cytotoxic t cell response to a model of acute viral infection. *Journal of virology*, 89(8):4517–4526, 2015.

[66] Amrit Dhar, Duncan K Ralph, Vladimir N Minin, and Frederick A Matsen IV. A bayesian phylogenetic hidden markov model for b cell receptor sequence analysis. *PLoS computational biology*, 16(8):e1008030, 2020.

[67] Roberto Di Niro, Seung-Joo Lee, Jason A Vander Heiden, Rebecca A Elsner, Nikita Trivedi, Jason M Bannock, Namita T Gupta, Steven H Kleinstein, Francois Vigneault, Tamara J Gilbert, et al. Salmonella infection drives promiscuous b cell activation followed by extrafollicular affinity maturation. *Immunity*, 43(1):120–131, 2015.

[68] Connor S Dobson, Anna N Reich, Stephanie Gaglione, Blake E Smith, Ellen J Kim, Jiayi Dong, Larance Ronsard, Vintus Okonkwo, Daniel Lingwood, Michael Dougan, et al. Antigen identification and high-throughput interaction mapping by reprogramming viral entry. *Nature methods*, 19(4):449–460, 2022.

[69] Ismail Dogan, Barbara Bertocci, Valérie Vilmont, Frédéric Delbos, Jérome Mégret, Sébastien Storck, Claude-Agnès Reynaud, and Jean-Claude Weill. Multiple layers of b cell memory with different effector functions. *Nature immunology*, 10(12):1292–1299, 2009.

[70] De Dong, Lvqin Zheng, Jianquan Lin, Bailing Zhang, Yuwei Zhu, Ningning Li, Shuangyu Xie, Yuhang Wang, Ning Gao, and Zhiwei Huang. Structural basis of assembly of the human t cell receptor–cd3 complex. *Nature*, 573(7775):546–552, 2019.

[71] Tao Dong, Guillaume Stewart-Jones, Nan Chen, Philippa Easterbrook, Xiaoning Xu, Laura Papagno, Victor Appay, Michael Weekes, Chris Conlon, Celsa Spina, et al. Hiv-specific cytotoxic t cells from long-term survivors select a unique t cell receptor. *The Journal of experimental medicine*, 200(12):1547–1557, 2004.

[72] Alexei J Drummond, Simon Y W Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

[73] Thomas Dupic, Meriem Bensouda Koraichi, Anastasia A Minervina, Mikhail V Pogorelyy, Thierry Mora, and Aleksandra M Walczak. Immune fingerprinting through repertoire similarity. *PLoS Genetics*, 17(1):e1009301, 2021.

[74] Yuval Elhanati, Anand Murugan, Curtis G Callan Jr, Thierry Mora, and Aleksandra M Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, 2014.

[75] Yuval Elhanati, Zachary Sethna, Curtis G Callan Jr, Thierry Mora, and Aleksandra M Walczak. Predicting the spectrum of tcr repertoire sharing with a data-driven model of recombination. *Immunological reviews*, 284(1):167–179, 2018.

[76] Ali H Ellebedy, Florian Krammer, Gui-Mei Li, Matthew S Miller, Christopher Chiu, Jens Wrammert, Cathy Y Chang, Carl W Davis, Megan McCausland, Rivka Elbein, et al. Induction of broadly cross-reactive antibody responses to the influenza ha stem region following h5n1 vaccination in humans. *Proceedings of the National Academy of Sciences*, 111(36):13133–13138, 2014.

[77] Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature genetics*, 49(5):659–665, 2017.

[78] Stefan C Endres, Carl Sandrock, and Walter W Focke. A simplicial homology algorithm for lipschitz optimisation. *Journal of Global Optimization*, 72:181–217, 2018.

[79] Farbod Famili, Anna-Sophia Wiekmeijer, and Frank JT Staal. The development of t cells from stem cells in mice and humans. *Future science OA*, 3(3):FSO186, 2017.

[80] Jean Feng, David A Shaw, Vladimir N Minin, Noah Simon, and Frederick A Matsen IV. Survival analysis of dna mutation motifs with penalized proportional hazards. *The annals of applied statistics*, 13(2):1268, 2019.

[81] Roux-Cil Ferreira, Emmanuel Wong, and Art FY Poon. bayroot: Bayesian sampling of hiv-1 integration dates by root-to-tip regression. *Virus Evolution*, 9(1):veac120, 2023.

[82] Richard Phillips Feynman. Space-time approach to non-relativistic quantum mechanics. *Reviews of modern physics*, 20(2):367, 1948.

[83] Mathilde Foglierini, Leontios Pappas, Antonio Lanzavecchia, Davide Corti, and Laurent Perez. Ancestree: An interactive immunoglobulin lineage tree visualizer. *PLoS computational biology*, 16(7):e1007731, 2020.

[84] Daniela Frölich, Claudia Giesecke, Henrik E Mei, Karin Reiter, Capucine Daridon, Peter E Lipsky, and Thomas Dörner. Secondary immunization generates clonally related antigen-specific plasma cells and memory b cells. *The Journal of Immunology*, 185(5):3103–3110, 2010.

[85] Mario U Gaimann, Maximilian Nguyen, Jonathan Desponds, and Andreas Mayer. Early life imprints the hierarchy of t cell clone sizes. *Elife*, 9:e61639, 2020.

[86] Peter Galison. Einstein's clocks: The place of time. *Critical Inquiry*, 26(2):355–389, 2000.

[87] Jacob D Galson, Sebastian Schaetzle, Rachael JM Bashford-Rogers, Matthew IJ Raybould, Aleksandr Kovaltsuk, Gavin J Kilpatrick, Ralph Minter, Donna K Finch, Jorge Dias, Louisa K James, et al. Deep sequencing of b cell receptor repertoires from covid-19 patients reveals strong convergent immune signatures. *Frontiers in immunology*, 11:605170, 2020.

[88] Yicheng Gao, Yuli Gao, Yuxiao Fan, Chengyu Zhu, Zhiting Wei, Chi Zhou, Guohui Chuai, Qinchang Chen, He Zhang, and Qi Liu. Pan-peptide meta learning for t-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, 5(3):236–249, 2023.

[89] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer Series in Synergetics. Springer-Verlag, Berlin, fourth edition, 2009.

[90] Nicholas RJ Gascoigne and S Munir Alam. Allelic exclusion of the t cell receptor α-chain: developmental regulation of a post-translational event. In *Seminars in immunology*, volume 11, pages 337–347. Elsevier, 1999.

[91] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, 32(2):158–168, 2014.

[92] Andrew Getahun. Role of inhibitory signaling in peripheral b cell tolerance. *Immunological reviews*, 307(1):27–42, 2022.

[93] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[94] Rachel M. Gittelman, Enrico Lavezzo, Thomas M. Snyder, H. Jabran Zahid, Cara L. Carty, Rebecca Elyanow, Sudeb Dalai, Ilan Kirsch, Lance Baldo, Laura Manuto, Elisa Franchin, Claudia Del Vecchio, Monia Pacenti, Caterina Boldrin, Margherita Cattai, Francesca Saluzzo, Andrea Padoan, Mario Plebani, Fabio Simeoni, Jessica Bordini, Nicola I. Lorè, Dejan Lazarević, Daniela M. Cirillo, Paolo Ghia, Stefano Toppo, Jonathan M. Carlson, Harlan S. Robins, Andrea Crisanti, and Giovanni Tonon. Longitudinal analysis of t cell receptor repertoires reveals shared patterns of antigen-specific response to sars-cov-2 infection. *JCI Insight*, 7(10), May 2022.

[95] Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M Krams, Christina Pettus, et al. Identifying specificity groups in the t cell receptor repertoire. *Nature*, 547(7661):94–98, 2017.

[96] Jeffrey E Gold, Ramazan A Okyay, Warren E Licht, and David J Hurley. Investigation of long covid prevalence and its relationship to epstein-barr virus reactivation. *Pathogens*, 10(6):763, 2021.

[97] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.

[98] Mikhail Goncharov, Dmitry Bagaev, Dmitrii Shcherbinin, Ivan Zvyagin, Dmitry Bolotin, Paul G Thomas, Anastasia A Minervina, Mikhail V Pogorelyy, Kristin Ladell, James E McLaren, et al. Vdjdb in the pandemic era: a compendium of t cell receptors specific for sars-cov-2. *Nature methods*, 19(9):1017–1019, 2022.

[99] Trisha Greenhalgh, Manoj Sivan, Alice Perlowski, and Janko Ž Nikolich. Long covid: a clinical update. *The Lancet*, 404(10453):707–724, August 2024.

[100] Claude Gregoire, Lionel Spinelli, Sergio Villazala-Merino, Laurine Gil, María Pía Holgado, Myriam Moussa, Chuang Dong, Ana Zarubica, Mathieu Fallet, Jean-Marc Navarro, et al. Viral infection engenders bona fide and bystander subsets of lung-resident memory b cells through a permissive mechanism. *Immunity*, 55(7):1216–1233, 2022.

[101] Severe Covid-19 GWAS Group. Genomewide association study of severe covid-19 with respiratory failure. *New England Journal of Medicine*, 383(16):1522–1534, 2020.

[102] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David SC Hui, et al. Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18):1708–1720, 2020.

[103] Namita T Gupta, Kristofor D Adams, Adrian W Briggs, Sonia C Timberlake, Francois Vigneault, and Steven H Kleinstein. Hierarchical clustering can identify b cell clones with high confidence in ig repertoire sequencing data. *The Journal of Immunology*, 198(6):2489–2499, 2017.

[104] Namita T Gupta, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H Kleinstein. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356–3358, 2015.

[105] Asmaa Hachim, Niloufar Kavian, Carolyn A Cohen, Alex WH Chin, Daniel KW Chu, Chris KP Mok, Owen TY Tsang, Yiu Cheong Yeung, Ranawaka APM Perera, Leo LM Poon, et al. Beyond the spike: identification of viral targets of the antibody responses to sars-cov-2 in covid-19 patients. *MedRxiv*, pages 2020–04, 2020.

[106] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[107] Xiaojian Han, Yingming Wang, Shenglong Li, Chao Hu, Tingting Li, Chenjian Gu, Kai Wang, Meiying Shen, Jianwei Wang, Jie Hu, et al. A rapid and efficient screening system for neutralizing antibodies and its application for sars-cov-2. *Frontiers in immunology*, 12:653189, 2021.

[108] Johanna Hansen, Alina Baum, Kristen E Pascal, Vincenzo Russo, Stephanie Giordano, Elzbieta Wloga, Benjamin O Fulton, Ying Yan, Katrina Koon, Krunal Patel, et al. Studies in humanized mice and convalescent humans yield a sars-cov-2 antibody cocktail. *Science*, 369(6506):1010–1014, 2020.

[109] Sarah Wulf Hanson, Cristiana Abbafati, Joachim G Aerts, Ziyad Al-Aly, Charlie Ashbaugh, Tala Ballouz, Oleg Blyuss, Polina Bobkova, Gouke Bonsel, Svetlana Borzakova, et al. Estimated global proportions of individuals with persistent fatigue, cognitive, and respiratory symptom clusters following symptomatic covid-19 in 2020 and 2021. *Jama*, 328(16):1604–1615, 2022.

[110] James Henderson, Yuta Nagano, Martina Milighetti, and Andreas Tiffeau-Mayer. Limits on inferring t cell specificity from partial information. *Proceedings of the National Academy of Sciences*, 121(42):e2408696121, 2024.

[111] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[112] Brittany Henry and Brian J Laidlaw. Functional heterogeneity in the memory b-cell response. *Current Opinion in Immunology*, 80:102281, 2023.

[113] Sebastian Herzog, Michael Reth, and Hassan Jumaa. Regulation of b-cell proliferation and differentiation by pre-b-cell receptor signalling. *Nature Reviews Immunology*, 9(3):195–205, 2009.

[114] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, 2024.

[115] Asger Hobolth, Iker Rivas-González, Mogens Bladt, and Andreas Futschik. Phase-type distributions in mathematical population genetics: An emerging framework. *Theoretical Population Biology*, 2024.

[116] Yosef Hochberg and Yoav Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7):811–818, 1990.

[117] Alexandra Domnica Hoeggerl, Verena Nunhofer, Wanda Lauth, Natalie Badstuber, Nina Held, Georg Zimmermann, Christoph Grabmer, Lisa Weidner, Christof Jungbauer, Nadja Lindlbauer, et al. Epstein-barr virus reactivation is not causative for post-covid-19-syndrome in individuals with asymptomatic or mild sars-cov-2 disease course. *BMC Infectious Diseases*, 23(1):800, 2023.

[118] Kenneth B Hoehn and Steven H Kleinstein. B cell phylogenetics in the single cell era. *Trends in Immunology*, 45(1):62–74, 2024.

[119] Kenneth B Hoehn, Gerton Lunter, and Oliver G Pybus. A phylogenetic codon substitution model for antibody lineages. *Genetics*, 206(1):417–427, 2017.

[120] Kenneth B Hoehn, Oliver G Pybus, and Steven H Kleinstein. Phylogenetic analysis of migration, differentiation, and class switching in b cells. *PLoS computational biology*, 18(4):e1009885, 2022.

[121] Kenneth B Hoehn, Jackson S Turner, Frederick I Miller, Ruoyi Jiang, Oliver G Pybus, Ali H Ellebedy, and Steven H Kleinstein. Human b cell lineages associated with germinal centers following influenza vaccination are measurably evolving. *Elife*, 10:e70873, 2021.

[122] Kenneth B Hoehn, Jason A Vander Heiden, Julian Q Zhou, Gerton Lunter, Oliver G Pybus, and Steven H Kleinstein. Repertoire-wide phylogenetic models of b cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences*, 116(45):22664–22672, 2019.

[123] Felix Horns, Christopher Vollmers, Derek Croote, Sally F Mackey, Gary E Swan, Cornelia L Dekker, Mark M Davis, and Stephen R Quake. Lineage tracing of human b cells reveals the in vivo landscape of human antibody class switching. *Elife*, 5:e16578, 2016.

[124] Felix Horns, Christopher Vollmers, Cornelia L Dekker, and Stephen R Quake. Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *Proceedings of the National Academy of Sciences*, 116(4):1261–1266, 2019.

[125] Chaoran Hu, Vladimir Pozdnyakov, and Jun Yan. Density and distribution evaluation for convolution of independent gamma variables. *Computational Statistics*, 35:327–342, 2020.

[126] Dan Hudson, Ricardo A Fernandes, Mark Basham, Graham Ogg, and Hashem Koohy. Can we predict t cell specificity with digital biology and machine learning? *Nature Reviews Immunology*, 23(8):511–521, 2023.

[127] John P Huelsenbeck, Bret Larget, and David Swofford. A compound poisson process for relaxing the molecular clock. *Genetics*, 154(4):1879–1892, 2000.

[128] Nicholas K Hurlburt, Emilie Seydoux, Yu-Hsin Wan, Venkata Viswanadh Edara, Andrew B Stuart, Junli Feng, Mehul S Suthar, Andrew T McGuire, Leonidas Stamatatos, and Marie Pancera. Structural basis for potent neutralization of sars-cov-2 and role of antibody affinity maturation. *Nature communications*, 11(1):5413, 2020.

[129] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

[130] Giulio Isacchini, Carlos Olivares, Armita Nourmohammad, Aleksandra M Walczak, and Thierry Mora. Sos: online probability estimation and generation of t-and b-cell receptors. *Bioinformatics*, 36(16):4510–4512, 2020.

[131] Giulio Isacchini, Valentin Quiniou, Pierre Barennes, Vanessa Mhanna, Hélène Vantomme, Paul Stys, Encarnita Mariotti-Ferrandiz, David Klatzmann, Aleksandra M Walczak, Thierry Mora, et al. Local and global variability in developing human t-cell repertoires. *PRX Life*, 2(1):013011, 2024.

[132] Giulio Isacchini, Zachary Sethna, Yuval Elhanati, Armita Nourmohammad, Aleksandra M Walczak, and Thierry Mora. Generative models of t-cell receptor sequences. *Physical Review E*, 101(6):062414, 2020.

[133] Giulio Isacchini, Aleksandra M Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of t and b cell receptor repertoires with sonnia. *Proceedings of the National Academy of Sciences*, 118(14):e2023141118, 2021.

[134] Shannon P Israelsen. The scientific theories of michael faraday and james clerk maxwell. *The Purdue Historian*, 7(1):1, 2014.

[135] CA Janeway. *Immunobiology: the immune system in health and disease*. Garland Science, 9 edition, 2005.

[136] CA Janeway. *Immunobiology: the immune system in health and disease*. Garland Science, 6 edition, 2005.

[137] Cole G Jensen, Jacob A Sumner, Steven H Kleinstein, and Kenneth B Hoehn. Inferring b cell phylogenies from paired heavy and light chain bcr sequences with dowser. *bioRxiv*, 2023.

[138] Mathias Fynbo Jensen and Morten Nielsen. Nettcr 2.2-improved tcr specificity predictions by combining pan-and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *bioRxiv*, pages 2023–10, 2023.

[139] David H. Jiang, Darius J. Roy, Brett J. Gu, Leslie C. Hassett, and Rozalina G. McCoy. Postacute sequelae of severe acute respiratory syndrome coronavirus 2 infection. *JACC: Basic to Translational Science*, 6(9–10):796–811, September 2021.

[140] Gustav Johansson, Melita Kaltak, Cristiana Rimniceanu, Avadhesh K Singh, Jan Lycke, Clas Malmeström, Michael Hühn, Outi Vaarala, Susanna Cardell, and Anders Ståhlberg. Ultrasensitive dna immune repertoire sequencing using unique molecular identifiers. *Clinical chemistry*, 66(9):1228–1237, 2020.

[141] Derek D Jones, Joel R Wilmore, and David Allman. Cellular dynamics of memory b cell populations: Igm+ and igg+ memory b cells persist indefinitely as quiescent cells. *The Journal of Immunology*, 195(10):4753–4759, 2015.

[142] Bin Ju, Qi Zhang, Jiwan Ge, Ruoke Wang, Jing Sun, Xiangyang Ge, Jiazhen Yu, Sisi Shan, Bing Zhou, Shuo Song, et al. Human neutralizing antibodies elicited by sars-cov-2 infection. *Nature*, 584(7819):115–119, 2020.

[143] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3(24):21–132, 1969.

[144] Shirin Kalimuddin, Christine YL Tham, Yvonne FZ Chan, Shou Kit Hang, Kamini Kunasegaran, Adeline Chia, Candice YY Chan, Dorothy HL Ng, Jean XY Sim, Hwee-Cheng Tan, et al. Vaccine-induced t cell responses control orthoflavivirus challenge infection without neutralizing antibodies in humans. *Nature Microbiology*, pages 1–14, 2025.

[145] Edward PC Kao. *An introduction to stochastic processes*. Courier Dover Publications, 2019.

[146] Claudia Kedor, Helma Freitag, Lil Meyer-Arndt, Kirsten Wittke, Leif G. Hanitsch, Thomas Zoller, Fridolin Steinbeis, Milan Haffke, Gordon Rudolf, Bettina Heidecker, Thomas Bobbert, Joachim Spranger, Hans-Dieter Volk, Carsten Skurk, Frank Konietschke, Friedemann Paul, Uta Behrends, Judith Bellmann-Strobl, and Carmen Scheibenbogen. A prospective observational study of post-covid-19 chronic fatigue syndrome following the first pandemic wave in germany and biomarkers associated with symptom severity. *Nature Communications*, 13(1), August 2022.

[147] Nick Keur, Antine W Flikweert, Isis Ricaño-Ponce, Anneke C Muller Kobold, Izabela A Rodenhuis-Zybert, Kieu TT Le, Matijs van Meurs, Marco J Grootenboers, Peter HJ van der Voort, Peter Heeringa, et al. Delineating sex-specific circulating host response signatures associated with covid-19 severity and mortality. *Iscience*, 27(11), 2024.

[148] Laleh Khodadadi, Qingyu Cheng, Andreas Radbruch, and Falk Hiepe. The maintenance of memory plasma cells. *Frontiers in immunology*, 10:721, 2019.

[149] Wooseob Kim, Julian Q Zhou, Stephen C Horvath, Aaron J Schmitz, Alexandria J Sturtz, Tingting Lei, Zhuoming Liu, Elizaveta Kalaidina, Mahima Thapa, Wafaa B Alsoussi, et al. Germinal centre-driven maturation of b cell response to mrna vaccination. *Nature*, 604(7904):141–145, 2022.

[150] Jon Klein, Jamie Wood, Jillian R. Jaycox, Rahul M. Dhodapkar, Peiwen Lu, Jeff R. Gehlhausen, Alexandra Tabachnikova, Kerrie Greene, Laura Tabacof, Amyn A. Malik, Valter Silva Monteiro, Julio Silva, Kathy Kamath, Minlu Zhang, Abhilash Dhal, Isabel M. Ott, Gabrielee Valle, Mario Peña-Hernández, Tianyang Mao, Bornali Bhattacharjee, Takehiro Takahashi, Carolina Lucas, Eric Song, Dayna McCarthy, Erica Breyman, Jenna Tosto-Mancuso, Yile Dai, Emily Perotti, Koray Akduman, Tiffany J. Tzeng, Lan Xu, Anna C. Geraghty, Michelle Monje, Inci Yildirim, John Shon, Ruslan Medzhitov, Denyse Lutchmansingh, Jennifer D. Possick, Naftali Kaminski, Saad B. Omer, Harlan M. Krumholz, Leying Guan, Charles S. Dela Cruz, David van Dijk,

Aaron M. Ring, David Putrino, and Akiko Iwasaki. Distinguishing features of long covid identified through immune profiling. *Nature*, 623(7985):139–148, September 2023.

[151] Steven H Kleinstein, Yoram Louzoun, and Mark J Shlomchik. Estimating hypermutation rates from clonal tree data. *The Journal of Immunology*, 171(9):4639–4649, 2003.

[152] Mark Klinger, Francois Pepin, Jen Wilkins, Thomas Asbury, Tobias Wittkop, Jianbiao Zheng, Martin Moorhead, and Malek Faham. Multiplex identification of antigen-specific t cell receptors using a combination of immune assays and immune receptor sequencing. *PloS one*, 10(10):e0141561, 2015.

[153] Meriem Bensouda Koraichi, Maximilian Puelma Touzel, Andrea Mazzolini, Thierry Mora, and Aleksandra M Walczak. Noiset: noise learning and expansion detection of t-cell receptors. *The Journal of Physical Chemistry A*, 126(40):7407–7414, 2022.

[154] Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, 2019.

[155] Christoph Kreer, Henning Gruell, Thierry Mora, Aleksandra M Walczak, and Florian Klein. Exploiting b cell receptor analyses to inform on hiv-1 vaccination strategies. *Vaccines*, 8(1):13, 2020.

[156] Christoph Kreer, Matthias Zehner, Timm Weber, Meryem S Ercanoglu, Lutz Gieselmann, Cornelius Rohde, Sandro Halwe, Michael Korenkov, Philipp Schommers, Kanika Vanshylla, et al. Longitudinal isolation of potent near-germline sars-cov-2-neutralizing antibodies from covid-19 patients. *Cell*, 182(4):843–854, 2020.

[157] Jakob Kreye, S Momsen Reincke, Hans-Christian Kornau, Elisa Sánchez-Sendin, Victor Max Corman, Hejun Liu, Meng Yuan, Nicholas C Wu, Xueyong Zhu, Chang-Chun D Lee, et al. A sars-cov-2 neutralizing antibody protects from lung pathology in a covid-19 hamster model. *bioRxiv*, 2020.

[158] Chirag Krishna, Diego Chowell, Mithat Gönen, Yuval Elhanati, and Timothy A Chan. Genetic and environmental determinants of human tcr repertoire diversity. *Immunity & Ageing*, 17:1–7, 2020.

[159] Denise Kühnert, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116, 2016.

[160] Marco Künzli and David Masopust. Cd4+ t cell memory. *Nature immunology*, 24(6):903–914, 2023.

[161] Masayuki Kuraoka, Aaron G Schmidt, Takuya Nojima, Feng Feng, Akiko Watanabe, Daisuke Kitamura, Stephen C Harrison, Thomas B Kepler, and Garnett Kelsoe. Complex antigens drive permissive clonal selection in germinal centers. *Immunity*, 44(3):542–552, 2016.

[162] Masayuki Kuraoka, Chen-Hao Yeh, Goran Bajic, Ryutaro Kotaki, Shengli Song, Ian Windsor, Stephen C Harrison, and Garnett Kelsoe. Recall of b cell memory depends on relative locations of prime and boost immunization. *Science immunology*, 7(71):eabn5311, 2022.

[163] Tomohiro Kurosaki, Kohei Kometani, and Wataru Ise. Memory b cells. *Nature Reviews Immunology*, 15(3):149–159, 2015.

[164] Bjørn PY Kwee, Marius Messemaker, Eric Marcus, Giacomo Oliveira, Wouter Scheper, Catherine J Wu, Jonas Teuwen, and Ton N Schumacher. Stapler: efficient learning of tcr-peptide specificity prediction from full-length tcr-peptide data. *bioRxiv*, pages 2023–04, 2023.

[165] Kaitlyn A Lagattuta, Joyce B Kang, Aparna Nathan, Kristen E Pauken, Anna Helena Jonsson, Deepak A Rao, Arlene H Sharpe, Kazuyoshi Ishigaki, and Soumya Raychaudhuri. Repertoire analyses reveal t cell antigen receptor sequence features that influence t cell fate. *Nature immunology*, 23(3):446–457, 2022.

[166] Kaitlyn A Lagattuta, Ayano C Kohlgruber, Nouran S Abdelfattah, Aparna Nathan, Laurie Rumker, Michael E Birnbaum, Stephen J Elledge, and Soumya Raychaudhuri. The t cell receptor sequence influences the likelihood of t cell memory formation. *Cell Reports*, 44(1), 2025.

[167] Nora Lam, YoonSeung Lee, and Donna L Farber. A guide to adaptive immune memory. *Nature Reviews Immunology*, pages 1–20, 2024.

[168] Donald W Lee, Ilja V Khavrutskii, Anders Wallqvist, Sina Bavari, Christopher L Cooper, and Sidhartha Chaudhury. Brilia: integrated tool for high-throughput annotation and lineage tree assembly of b-cell repertoires. *Frontiers in immunology*, 7:681, 2017.

[169] Jeong Hyun Lee, Henry J Sutton, Christopher A Cottrell, Ivy Phung, Gabriel Ozorowski, Leigh M Sewall, Rebecca Nedellec, Catherine Nakao, Murillo Silva, Sara T Richey, et al. Long-primed germinal centres with enduring affinity maturation and clonal migration. *Nature*, 609(7929):998–1004, 2022.

[170] Peter S Lee, Nobuko Ohshima, Robyn L Stanfield, Wenli Yu, Yoshitaka Iba, Yoshinobu Okuno, Yoshikazu Kurosawa, and Ian A Wilson. Receptor mimicry by antibody f045–092 facilitates universal binding to the h3 subtype of influenza virus. *Nature communications*, 5(1):3614, 2014.

[171] Paul J Lehner, EC Wang, PA Moss, Sheila Williams, Kaye Platt, Steven M Friedman, John I Bell, and Leszek K Borysiewicz. Human hla-a0201-restricted cytotoxic t lymphocyte recognition of influenza a is dominated by t cells bearing the v beta 17 gene segment. *The Journal of experimental medicine*, 181(1):79–91, 1995.

[172] Jay T Lennon, Frank den Hollander, Maite Wilke-Berenguer, and Jochen Blath. Principles of seed banks and the emergence of complexity from dormancy. *Nature Communications*, 12(1):4807, 2021.

[173] Maya A Lewinsohn, Trevor Bedford, Nicola F Müller, and Alison F Feder. State-dependent evolutionary models reveal modes of solid tumour growth. *Nature Ecology & Evolution*, 7(4):581–596, 2023.

[174] Katherine M Littlefield, Renée O Watson, Jennifer M Schneider, Charles P Neff, Eiko Yamada, Min Zhang, Thomas B Campbell, Michael T Falta, Sarah E Jolley, Andrew P Fontenot, et al. Sars-cov-2-specific t cells associate with inflammation and reduced lung function in pulmonary post-acute sequalae of sars-cov-2. *PLoS pathogens*, 18(5):e1010359, 2022.

[175] Hejun Liu, Nicholas C Wu, Meng Yuan, Sandhya Bangaru, Jonathan L Torres, Tom G Caniels, Jelle Van Schooten, Xueyong Zhu, Chang-Chun D Lee, Philip JM Brouwer, et al. Cross-neutralization of a sars-cov-2 antibody to a functionally conserved site is mediated by avidity. *Immunity*, 53(6):1272–1280, 2020.

[176] Lihong Liu, Pengfei Wang, Manoj S Nair, Jian Yu, Micah Rapp, Qian Wang, Yang Luo, Jasper F-W Chan, Vincent Sahi, Amir Figueroa, et al. Potent neutralizing antibodies against multiple epitopes on sars-cov-2 spike. *Nature*, 584(7821):450–456, 2020.

[177] Konrad Lohse, Richard J Harrison, and Nicholas H Barton. A general method for calculating likelihoods under the coalescent process. *Genetics*, 189(3):977–987, 2011.

[178] Lenette L Lu, Todd J Suscovich, Sarah M Fortune, and Galit Alter. Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews Immunology*, 18(1):46–61, 2018.

[179] Huibin Lv, Nicholas C Wu, Owen Tak-Yin Tsang, Meng Yuan, Ranawaka APM Perera, Wai Shing Leung, Ray TY So, Jacky Man Chun Chan, Garrick K Yip, Thomas

Shiu Hong Chik, et al. Cross-reactive antibody response between sars-cov-2 and sars-cov infections. *Cell reports*, 31(9), 2020.

[180] Cedric Malherbe and Talip Uçar. Igblend: Unifying 3d structures and sequences in antibody language models. *bioRxiv*, pages 2024–10, 2024.

[181] Jonathan S Maltzman. Cd8 t cells are forever. *Science Immunology*, 8(80):eadg8279, 2023.

[182] Donna M. Mancini, Danielle L. Brunjes, Anuradha Lala, Maria Giovanna Trivieri, Johanna P. Contreras, and Benjamin H. Natelson. Use of cardiopulmonary stress testing for patients with unexplained dyspnea post–coronavirus disease. *JACC: Heart Failure*, 9(12):927–937, December 2021.

[183] Mattia Manica, Maria Litvinova, Alfredo De Bellis, Giorgio Guzzetta, Pamela Mancuso, Massimo Vicentini, Francesco Venturelli, Eufemia Bisaccia, Ana I Bento, Piero Poletti, et al. Estimation of the incubation period and generation time of sars-cov-2 alpha and delta variants from contact tracing data. *Epidemiology & Infection*, 151:e5, 2023.

[184] Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with igor. *Nature communications*, 9(1):561, 2018.

[185] John C Marshall, Srinivas Murthy, Janet Diaz, N K Adhikari, Derek C Angus, Yaseen M Arabi, Kenneth Baillie, Michael Bauer, Scott Berry, Bronagh Blackwood, Marc Bonten, Fernando Bozza, Frank Brunkhorst, Allen Cheng, Mike Clarke, Vu Quoc Dat, Menno de Jong, Justin Denholm, Lennie Derde, Jake Dunning, Xiaobin Feng, Tom Fletcher, Nadine Foster, Rob Fowler, Nina Gobat, Charles Gomersall, Anthony Gordon, Thomas Glueck, Michael Harhay, Carol Hodgson, Peter Horby, YaeJean Kim, Richard Kojan, Bharath Kumar, John Laffey, Denis Malvey, Ignacio Martin-Loeches, Colin McArthur, Danny McAuley, Stephen McBride, Shay McGuinness, Laura Merson, Susan Morpeth, Dale Needham, Mihai Netea, Myoung-Don Oh, Sabai Phyu, Simone Piva, Ruijin Qiu, Halima Salisu-Kabara, Lei Shi, Naoki Shimizu, Jorge Sinclair, Steven Tong, Alexis Turgeon, Tim Uyeki, Frank van de Veerdonk, Steve Webb, Paula Williamson, Timo Wolf, and Junhua Zhang. A minimal common outcome measure set for covid-19 clinical research. *The Lancet Infectious Diseases*, 20(8):e192–e197, 2020.

[186] Hanover C Matz, Katherine M McIntire, and Ali H Ellebedy. Persistent germinal center responses: slow-growing trees bear the best fruits. *Current opinion in immunology*, 83:102332, 2023.

[187] Koshlan Mayer-Blackwell, Stefan Schattgen, Liel Cohen-Lavi, Jeremy C Crawford, Aisha Souquette, Jessica A Gaevert, Tomer Hertz, Paul G Thomas, Philip Bradley, and Andrew Fiore-Gartland. Tcr meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, hla-restricted clusters of sars-cov-2 tcrs. *Elife*, 10:e68605, 2021.

[188] Louise J McHeyzer-Williams, Pierre J Milpied, Shinji L Okitsu, and Michael G McHeyzer-Williams. Class-switched memory b cells remodel bcrs within secondary germinal centers. *Nature immunology*, 16(3):296–305, 2015.

[189] Julia L McKechnie and Catherine A Blish. The innate immune system: fighting on the front lines or fanning the flames of covid-19? *Cell host & microbe*, 27(6):863–869, 2020.

[190] Stuart McLaughlin. The electrostatic properties of membranes. *Annual review of biophysics and biophysical chemistry*, 18(1):113–136, 1989.

[191] Luka Mesin, Ariën Schiepers, Jonatan Ersching, Alexandru Barbulescu, Cecília B Cavazzoni, Alessandro Angelini, Takaharu Okada, Tomohiro Kurosaki, and Gabriel D Victora. Restricted clonality and limited germinal center reentry characterize memory b cell reactivation by boosting. *Cell*, 180(1):92–106, 2020.

[192] Barthelemy Meynard-Piganeau, Christoph Feinauer, Martin Weigt, Aleksandra M Walczak, and Thierry Mora. Tulip: A transformer-based unsupervised language model for interacting peptides and t cell receptors that generalizes to unseen epitopes. *Proceedings of the National Academy of Sciences*, 121(24):e2316401121, 2024.

[193] Melina Michelen, Lakshmi Manoharan, Natalie Elkheir, Vincent Cheng, Andrew Dagens, Claire Hastie, Margaret O'Hara, Jake Suett, Dania Dahmash, Polina Bugaeva, Ishmeala Rigby, Daniel Munblit, Eli Harriss, Amanda Burls, Carole Foote, Janet Scott, Gail Carson, Piero Olliaro, Louise Sigfrid, and Charitini Stavropoulou. Characterising long covid: a living systematic review. *BMJ Global Health*, 6(9):e005427, September 2021.

[194] Martina Milighetti, Yanchun Peng, Cedric Tan, Michal Mark, Gayathri Nageswaran, Suzanne Byrne, Tahel Ronel, Tom Peacock, Andreas Mayer, Aneesh Chandran, et al. Large clones of pre-existing t cells drive early immunity against sars-cov-2 and lcmv infection. *Iscience*, 26(6), 2023.

[195] Joseph D Miller, Robbert G van der Most, Rama S Akondy, John T Glidewell, Sophia Albott, David Masopust, Kaja Murali-Krishna, Patryce L Mahar, Srilatha Edupuganti,

Susan Lalor, et al. Human effector and memory cd8+ t cell responses to smallpox and yellow fever vaccines. *Immunity*, 28(5):710–722, 2008.

[196] Anastasia A Minervina, Ekaterina A Komech, Aleksei Titov, Meriem Bensouda Koraichi, Elisa Rosati, Ilgar Z Mamedov, Andre Franke, Grigory A Efimov, Dmitriy M Chudakov, Thierry Mora, et al. Longitudinal high-throughput tcr repertoire profiling reveals the dynamics of t-cell memory formation after mild covid-19 infection. *Elife*, 10:e63502, 2021.

[197] Anastasia A Minervina, Mikhail V Pogorelyy, Ekaterina A Komech, Vadim K Karnaukhov, Petra Bacher, Elisa Rosati, Andre Franke, Dmitriy M Chudakov, Ilgar Z Mamedov, Yuri B Lebedev, et al. Primary and secondary anti-viral response captured by the dynamics and phenotype of individual t cell clones. *Elife*, 9:e53704, 2020.

[198] Sindhu Mohandas, Prasanna Jagannathan, Timothy J Henrich, Zaki A Sherif, Christian Bime, Erin Quinlan, Michael A Portman, Marila Gennaro, and Jalees Rehman. Immune mechanisms underlying covid-19 pathology and post-acute sequelae of sars-cov-2 infection (pasc). *Elife*, 12:e86014, 2023.

[199] Zachary Montague. https://github.com/zacmon/tcrdist_rs. 2024.

[200] Zachary Montague, Huibin Lv, Jakub Otwinowski, William S DeWitt, Giulio Isacchini, Garrick K Yip, Wilson W Ng, Owen Tak-Yin Tsang, Meng Yuan, Hejun Liu, et al. Dynamics of b cell repertoires and emergence of cross-reactive responses in patients with different severities of covid-19. *Cell Reports*, 35(8), 2021.

[201] Thierry Mora and Aleksandra M Walczak. Quantifying lymphocyte receptor diversity. In *Systems Immunology*, pages 183–198. CRC Press, 2018.

[202] Thierry Mora and Aleksandra M Walczak. Towards a quantitative theory of tolerance. *Trends in Immunology*, 44(7):512–518, 2023.

[203] Imogen Moran, Akira Nguyen, Weng Hua Khoo, Danyal Butt, Katherine Bourne, Clara Young, Jana R Hermes, Maté Biro, Gary Gracie, Cindy S Ma, et al. Memory b cells are reactivated in subcapsular proliferative foci of lymph nodes. *Nature communications*, 9(1):3372, 2018.

[204] Vito MR Muggeo. Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19):3055–3071, 2003.

[205] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[206] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan Jr. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.

[207] Alexander Myronov, Giovanni Mazzocco, Paulina Król, and Dariusz Plewczynski. Bertrand—peptide: Tcr binding prediction using bidirectional encoder representations from transformers augmented with random tcr pairing. *Bioinformatics*, 39(8):btad468, 2023.

[208] Yuta Nagano, Andrew GT Pyo, Martina Milighetti, James Henderson, John Shawe-Taylor, Benny Chain, and Andreas Tiffeau-Mayer. Contrastive learning of t cell receptor representations. *Cell Systems*, 16(1), 2025.

[209] Randolph Nelson. *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*. Springer Science & Business Media, 2013.

[210] David Nemazee. Mechanisms of central tolerance for b cells. *Nature Reviews Immunology*, 17(5):281–294, 2017.

[211] Hadas Neuman, Jessica Arrouasse, Meirav Kedmi, Andrea Cerutti, Giuliana Magri, and Ramit Mehr. Igtreez, a toolkit for immunoglobulin gene lineage tree-based analysis, reveals cdr3s are crucial for selection analysis. *Frontiers in Immunology*, 13:822834, 2022.

[212] Jerzy Neyman and Egon S Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference part i. *Biometrika*, 20(1-2):175–240, 1928.

[213] Sandra CA Nielsen and Scott D Boyd. Human adaptive immune receptor repertoire analysis—past, present, and future. *Immunological reviews*, 284(1):9–23, 2018.

[214] Sandra CA Nielsen, Fan Yang, Katherine JL Jackson, Ramona A Hoh, Katharina Röltgen, Grace H Jean, Bryan A Stevens, Ji-Yeun Lee, Arjun Rustagi, Angela J Rogers, et al. Human b cell clonal expansion and convergent antibody responses to sars-cov-2. *Cell host & microbe*, 28(4):516–525, 2020.

[215] Xuefeng Niu, Song Li, Pingchao Li, Wenjing Pan, Qian Wang, Ying Feng, Xiaoneng Mo, Qihong Yan, Xianmiao Ye, Jia Luo, et al. Longitudinal analysis of t and b cell receptor repertoire transcripts reveal dynamic immune response in covid-19 patients. *Frontiers in Immunology*, 11:582010, 2020.

[216] Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N Dines, Ian M Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitch Pesesky, Rachel M Gittelman, et al. A large-scale database of t-cell receptor beta (tcrβ) sequences and binding associations from natural and synthetic exposure to sars-cov-2. *Research square*, 2020.

[217] Armita Nourmohammad, Jakub Otwinowski, Marta Łuksza, Thierry Mora, and Aleksandra M Walczak. Fierce selection and interference in b-cell repertoire response to chronic hiv-1. *Molecular biology and evolution*, 36(10):2184–2194, 2019.

[218] Tal Noy-Porat, Efi Makdasi, Ron Alcalay, Adva Mechaly, Yinon Levy, Adi Bercovich-Kinori, Ayelet Zauberman, Hadas Tamir, Yfat Yahalom-Ronen, Ma'ayan Israeli, et al. A panel of human neutralizing mabs targeting sars-cov-2 spike at multiple epitopes. *Nature communications*, 11(1):4303, 2020.

[219] Stephen L Nutt, Philip D Hodgkin, David M Tarlinton, and Lynn M Corcoran. The generation of antibody-secreting plasma cells. *Nature Reviews Immunology*, 15(3):160–171, 2015.

[220] Branden J Olson, Stefan A Schattgen, Paul G Thomas, Philip Bradley, and Frederick A Matsen IV. Comparing t cell receptor repertoires using optimal transport. *PLOS Computational Biology*, 18(12):e1010681, 2022.

[221] Sarah P Otto, Ailene MacPherson, and Caroline Colijn. Endemic does not mean constant as sars-cov-2 continues to evolve. *Evolution*, 78(6):1092–1108, 2024.

[222] Clovis S Palmer, Chrysostomos Perdios, Mohamed Abdel-Mohsen, Joseph Mudd, Prasun K Datta, Nicholas J Maness, Gabrielle Lehmicke, Nadia Golden, Linh Hellmers, Carol Coyne, et al. Non-human primate model of long-covid identifies immune associates of hyperglycemia. *Nature communications*, 15(1):6664, 2024.

[223] Krystallenia Paniskaki, Sarah Goretzki, Moritz Anft, Margarethe J Konik, Toni L Meister, Stephanie Pfaender, Klara Lechtenberg, Melanie Vogl, Burcin Dogan, Sebastian Dolff, et al. Increased sars-cov-2 reactive low avidity t cells producing inflammatory cytokines in pediatric post-acute covid-19 sequelae (pasc). *Pediatric Allergy and Immunology*, 34(12):e14060, 2023.

[224] Kathryn A Pape, Justin J Taylor, Robert W Maul, Patricia J Gearhart, and Marc K Jenkins. Different b cell populations mediate early and late memory during an endogenous immune response. *Science*, 331(6021):1203–1207, 2011.

[225] Matteo Parotto, Mariann Gyöngyösi, Kathryn Howe, Sheila N Myatra, Otavio Ranzani, Manu Shankar-Hari, and Margaret S Herridge. Post-acute sequelae of covid-19:

understanding and addressing the burden of multisystem manifestations. *The Lancet Respiratory Medicine*, 11(8):739–754, August 2023.

[226] Michael J. Peluso, Tyler-Marie Deveau, Sadie E. Munter, Dylan Ryder, Amanda Buck, Gabriele Beck-Engeser, Fay Chan, Scott Lu, Sarah A. Goldberg, Rebecca Hoh, Viva Tai, Leonel Torres, Nikita S. Iyer, Monika Deswal, Lynn H. Ngo, Melissa Buitrago, Antonio Rodriguez, Jessica Y. Chen, Brandon C. Yee, Ahmed Chenna, John W. Winslow, Christos J. Petropoulos, Amelia N. Deitchman, Joanna Hellmuth, Matthew A. Spinelli, Matthew S. Durstenfeld, Priscilla Y. Hsue, J. Daniel Kelly, Jeffrey N. Martin, Steven G. Deeks, Peter W. Hunt, and Timothy J. Henrich. Chronic viral coinfections differentially affect the likelihood of developing long covid. *Journal of Clinical Investigation*, 133(3), February 2023.

[227] Ranawaka APM Perera, Chris KP Mok, Owen TY Tsang, Huibin Lv, Ronald LW Ko, Nicholas C Wu, Meng Yuan, Wai Shing Leung, Jacky MC Chan, Thomas SH Chik, et al. Serological assays for severe acute respiratory syndrome coronavirus 2 (sars-cov-2), march 2020. *Eurosurveillance*, 25(16):2000421, 2020.

[228] My-Diem Nguyen Pham, Thanh-Nhan Nguyen, Le Son Tran, Que-Tran Bui Nguyen, Thien-Phuc Hoang Nguyen, Thi Mong Quynh Pham, Hoai-Nghia Nguyen, Hoa Giang, Minh-Duy Phan, and Vy Nguyen. epitcr: a highly sensitive predictor for tcr–peptide binding. *Bioinformatics*, 39(5):btad284, 2023.

[229] Tri Giang Phan, Didrik Paus, Tyani D Chan, Marian L Turner, Stephen L Nutt, Antony Basten, and Robert Brink. High affinity germinal center b cells are actively selected into the plasma cell compartment. *The Journal of experimental medicine*, 203(11):2419, 2006.

[230] Chansavath Phetsouphanh, Brendan Jacka, Sara Ballouz, Katherine JL Jackson, Daniel B Wilson, Bikash Manandhar, Vera Klemm, Hyon-Xhi Tan, Adam Wheatley, Anupriya Aggarwal, et al. Improvement of immune dysregulation in individuals with long covid at 24-months following sars-cov-2 infection. *Nature Communications*, 15(1):3315, 2024.

[231] Dora Pinto, Young-Jun Park, Martina Beltramello, Alexandra C Walls, M Alejandra Tortorici, Siro Bianchi, Stefano Jaconi, Katja Culap, Fabrizia Zatta, Anna De Marco, et al. Cross-neutralization of sars-cov-2 by a human monoclonal sars-cov antibody. *Nature*, 583(7815):290–295, 2020.

[232] Mikhail V Pogorelyy, Allison M Kirk, Samir Adhikari, Anastasia A Minervina, Balaji Sundararaman, Kasi Vegesana, David C Brice, Zachary B Scott, Paul G Thomas,

SJTRC Study Team, et al. Tirtl-seq: Deep, quantitative, and affordable paired tcr repertoire sequencing. *bioRxiv*, 2024.

[233] Mikhail V Pogorelyy, Anastasia A Minervina, Dmitriy M Chudakov, Ilgar Z Mamedov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Method for identification of condition-associated public antigen receptor sequences. *Elife*, 7:e33050, 2018.

[234] Mikhail V Pogorelyy, Anastasia A Minervina, Mikhail Shugay, Dmitriy M Chudakov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Detecting t cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biology*, 17(6):e3000314, 2019.

[235] Mikhail V Pogorelyy, Anastasia A Minervina, Maximilian Puelma Touzel, Anastasiia L Sycheva, Ekaterina A Komech, Elena I Kovalenko, Galina G Karganova, Evgeniy S Egorov, Alexander Yu Komkov, Dmitriy M Chudakov, et al. Precise tracking of vaccine-responding t cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences*, 115(50):12704–12709, 2018.

[236] Mikhail V Pogorelyy, Elisa Rosati, Anastasia A Minervina, Robert C Mettelman, Alexander Scheffold, Andre Franke, Petra Bacher, and Paul G Thomas. Resolving sars-cov-2 cd4+ t cell specificity via reverse epitope discovery. *Cell Reports Medicine*, 3(8), 2022.

[237] Beth Pollack, Emelia von Saltza, Lisa McCorkell, Lucia Santos, Ashley Hultman, Alison K Cohen, and Letícia Soares. Female reproductive health impacts of long covid and associated illnesses including me/cfs, pots, and connective tissue disorders: a literature review. *Frontiers in Rehabilitation Sciences*, 4:1122673, 2023.

[238] Anna Postovskaya, Koen Vercauteren, Pieter Meysman, and Kris Laukens. tcrblosum: an amino acid substitution matrix for sensitive alignment of distant epitope-specific tcrs. *Briefings in Bioinformatics*, 26(1):bbae602, 2025.

[239] Ponraj Prabakaran and Partha S Chowdhury. Landscape of non-canonical cysteines in human vh repertoire revealed by immunogenetic analysis. *Cell reports*, 31(13), 2020.

[240] Maximilian Puelma Touzel, Aleksandra M Walczak, and Thierry Mora. Inferring the immune response from repertoire sequencing. *PLOS Computational Biology*, 16(4):e1007873, 2020.

[241] Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution*, 17(6):890–896, 2000.

[242] Andrew GT Pyo, Yuta Nagano, Martina Milighetti, James Henderson, Curtis G Callan Jr, Benny Chain, Ned S Wingreen, and Andreas Tiffeau-Mayer. Data-driven discovery of biophysical t cell receptor co-specificity rules. *arXiv preprint arXiv:2412.13722*, 2024.

[243] Isabella Quinti, Vassilios Lougaris, Cinzia Milito, Francesco Cinetto, Antonio Pecoraro, Ivano Mezzaroma, Claudio Maria Mastroianni, Ombretta Turriziani, Maria Pia Bondioni, Matteo Filippini, et al. A possible role for b cells in covid-19? lesson from patients with agammaglobulinemia. *Journal of Allergy and Clinical Immunology*, 146(1):211–213, 2020.

[244] Duncan K Ralph and Frederick A Matsen IV. Using b cell receptor lineage structures to predict affinity. *PLOS Computational Biology*, 16(11):e1008391, 2020.

[245] Andrew Rambaut, Tommy T Lam, Luiz Max Carvalho, and Oliver G Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus evolution*, 2(1):vew007, 2016.

[246] Julita Ramírez, Kara Lukin, and James Hagman. From hematopoietic progenitors to b cells: mechanisms of lineage restriction and commitment. *Current opinion in immunology*, 22(2):177–184, 2010.

[247] Puneet Rawat, Melanie R Shapiro, Leeana D Peters, Michael Widrich, Koshlan Mayer-Blackwell, Keshav Motwani, Milena Pavlovi, Ghadi al Hajj, Amanda L Posgai, Chakravarthi Kanduri, et al. Identification of a type 1 diabetes-associated t cell receptor repertoire signature from the human peripheral blood. *medRxiv*, pages 2024–12, 2024.

[248] Justin T Reese, Hannah Blau, Elena Casiraghi, Timothy Bergquist, Johanna J Loomba, Tiffany J Callahan, Bryan Laraway, Corneliu Antonescu, Ben Coleman, Michael Gargano, et al. Generalisable long covid subtypes: findings from the nih n3c and recover programmes. *EBioMedicine*, 87, 2023.

[249] Michael Reth. Matching cellular dimensions with molecular sizes. *Nature immunology*, 14(8):765–767, 2013.

[250] Davide F Robbiani, Christian Gaebler, Frauke Muecksch, Julio CC Lorenzi, Zijun Wang, Alice Cho, Marianna Agudelo, Christopher O Barnes, Anna Gazumyan, Shlomo Finkin, et al. Convergent antibody responses to sars-cov-2 in convalescent individuals. *Nature*, 584(7821):437–442, 2020.

[251] Harlan Robins. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology*, 25(5):646–652, 2013.

[252] Thomas F Rogers, Fangzhu Zhao, Deli Huang, Nathan Beutler, Alison Burns, Wanting He, Oliver Limbo, Chloe Smith, Ge Song, Jordan Woehl, et al. Isolation of potent sars-cov-2 neutralizing antibodies and protection from disease in a small animal model. *Science*, 369(6506):956–963, 2020.

[253] Johanna Rohrhofer, Marianne Graninger, Lisa Lettenmaier, Johannes Schweighardt, Salvatore Alessio Gentile, Larissa Koidl, Davide Ret, Michael Stingl, Elisabeth Puchhammer-Stöckl, and Eva Untersmayr. Association between epstein-barr-virus reactivation and development of long-covid fatigue. *Allergy*, pages 10–1111, 2022.

[254] María Ruiz Ortega, Natanael Spisak, Thierry Mora, and Aleksandra M Walczak. Modeling and predicting the overlap of b-and t-cell receptor repertoires in healthy and sars-cov-2 infected individuals. *PLoS Genetics*, 19(2):e1010652, 2023.

[255] Magdalena L Russell, Assya Trofimov, Philip Bradley, and Frederick A Matsen IV. Statistical analysis of repertoire data demonstrates the influence of microhomology in v (d) j recombination. *bioRxiv*, 2024.

[256] Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

[257] Matthew Salaciak. *Revealing the Heterogeneity of T Cells in a Relapsed Hodgkin Lymphoma Patient Treated with Immunotherapy using Single-cell RNA Sequencing*. McGill University (Canada), 2022.

[258] Michael J Sanderson. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular biology and evolution*, 19(1):101–109, 2002.

[259] André Santa Cruz, Ana Mendes-Frias, Marne Azarias-da Silva, Sónia André, Ana Isabel Oliveira, Olga Pires, Marta Mendes, Bárbara Oliveira, Marta Braga, Joana Rita Lopes, et al. Post-acute sequelae of covid-19 is characterized by diminished peripheral cd8+ β7 integrin+ t cells and anti-sars-cov-2 iga response. *Nature Communications*, 14(1):1772, 2023.

[260] Yuki Sato, Karina Silina, Maries van den Broek, Kiyoshi Hirahara, and Motoko Yanagita. The roles of tertiary lymphoid structures in chronic diseases. *Nature Reviews Nephrology*, 19(8):525–537, 2023.

[261] Lonneke Scheffer, Eric Emanuel Reber, Brij Bhushan Mehta, Milena Pavlović, Maria Chernigovskaya, Eve Richardson, Rahmad Akbar, Fridtjof Lund-Johansen, Victor Greiff, Ingrid Hobæk Haff, et al. Predictability of antigen binding based on short motifs in the antibody cdrh3. *Briefings in Bioinformatics*, 25(6):bbae537, 2024.

[262] Oskar H Schnaack and Armita Nourmohammad. Optimal evolutionary decision-making to store immune memory. *Elife*, 10:e61346, 2021.

[263] Oskar H Schnaack, Luca Peliti, and Armita Nourmohammad. Risk-utility tradeoff shapes memory strategies for evolving patterns. *arXiv preprint arXiv:2110.15008*, 2021.

[264] Oskar H Schnaack, Luca Peliti, and Armita Nourmohammad. Learning and organization of memory for evolving patterns. *Physical Review X*, 12(2):021063, 2022.

[265] Chaim A Schramm, Zizhang Sheng, Zhenhai Zhang, John R Mascola, Peter D Kwong, and Lawrence Shapiro. Sonar: a high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of b cell transcripts. *Frontiers in immunology*, 7:372, 2016.

[266] Christoph Schultheiß, Lisa Paschold, Donjete Simnica, Malte Mohme, Edith Willscher, Lisa von Wenserski, Rebekka Scholz, Imke Wieters, Christine Dahlke, Eva Tolosa, et al. Next-generation sequencing of t and b cell receptor repertoires from covid-19 patients showed signatures associated with severity of disease. *Immunity*, 53(2):442–455, 2020.

[267] Jérémie Scire, Joëlle Barido-Sottani, Denise Kühnert, Timothy G Vaughan, and Tanja Stadler. Robust phylodynamic analysis of genetic sequencing data from structured populations. *Viruses*, 14(8):1648, 2022.

[268] Ron Sender, Yarden Weiss, Yoav Navon, Idan Milo, Nofar Azulay, Leeat Keren, Shai Fuchs, Danny Ben-Zvi, Elad Noor, and Ron Milo. The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences*, 120(44):e2308511120, 2023.

[269] Zachary Sethna, Giulio Isacchini, Thomas Dupic, Thierry Mora, Aleksandra M Walczak, and Yuval Elhanati. Population variability in the generation and selection of t-cell repertoires. *PLOS Computational Biology*, 16(12):e1008394, 2020.

[270] Emilie Seydoux, Leah J Homad, Anna J MacCamy, K Rachael Parks, Nicholas K Hurlburt, Madeleine F Jennewein, Nicholas R Akins, Andrew B Stuart, Yu-Hsin Wan, Junli Feng, et al. Analysis of a sars-cov-2-infected individual reveals development of potent neutralizing antibodies with limited somatic mutation. *Immunity*, 53(1):98–105, 2020.

[271] Emilie Seydoux, Leah J Homad, Anna J MacCamy, K Rachael Parks, Nicholas K Hurlburt, Madeleine F Jennewein, Nicholas R Akins, Andrew B Stuart, Yu-Hsin Wan, Junli Feng, et al. Characterization of neutralizing antibodies from a sars-cov-2 infected individual. *BioRxiv*, 2020.

[272] Zizhang Sheng, Chaim A Schramm, Mark Connors, Lynn Morris, John R Mascola, Peter D Kwong, and Lawrence Shapiro. Effects of darwinian selection and mutability on rate of broadly neutralizing antibody evolution during hiv-1 infection. *PLoS computational biology*, 12(5):e1004940, 2016.

[273] Rui Shi, Chao Shan, Xiaomin Duan, Zhihai Chen, Peipei Liu, Jinwen Song, Tao Song, Xiaoshan Bi, Chao Han, Lianao Wu, et al. A human neutralizing antibody targets the receptor-binding site of sars-cov-2. *Nature*, 584(7819):120–124, 2020.

[274] Ryo Shinnakasu, Takeshi Inoue, Kohei Kometani, Saya Moriyama, Yu Adachi, Manabu Nakayama, Yoshimasa Takahashi, Hidehiro Fukuyama, Takaharu Okada, and Tomohiro Kurosaki. Regulated selection of germinal-center cells into the memory b cell compartment. *Nature immunology*, 17(7):861–869, 2016.

[275] Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, et al. Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic acids research*, 46(D1):D419–D427, 2018.

[276] Johanna Sieurin, Gunnar Brandén, Cecilia Magnusson, Maria-Pia Hergens, and Kyriaki Kosidou. A population-based cohort study of sex and risk of severe outcomes in covid-19. *European Journal of Epidemiology*, 37(11):1159–1169, 2022.

[277] Adrien Six, Maria Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M Walczak, and Pierre Boudinot. The past, present, and future of immune repertoire biology–the rise of next-generation repertoire analysis. *Frontiers in immunology*, 4:413, 2013.

[278] Mark K Slifka, Rustom Antia, Jason K Whitmire, and Rafi Ahmed. Humoral immunity due to long-lived plasma cells. *Immunity*, 8(3):363–372, 1998.

[279] James A Smith and Alan F Karr. A point process model of summer season rainfall occurrences. *Water Resources Research*, 19(1):95–103, 1983.

[280] Kenneth GC Smith, Amanda Light, GJV Nossal, and David M Tarlinton. The extent of affinity maturation differs between the memory and antibody-forming cell compartments in the primary immune response. *The EMBO journal*, 1997.

[281] Devin Sok, Uri Laserson, Jonathan Laserson, Yi Liu, Francois Vigneault, Jean-Philippe Julien, Bryan Briney, Alejandra Ramos, Karen F Saye, Khoa Le, et al. The effects of somatic hypermutation on neutralization and binding in the pgt121 family of broadly neutralizing hiv antibodies. *PLoS pathogens*, 9(11):e1003754, 2013.

[282] Natanael Spisak, Gabriel Athènes, Thomas Dupic, Thierry Mora, and Aleksandra M Walczak. Combining mutation and recombination statistics to infer clonal families in antibody repertoires. *Elife*, 13:e86181, 2024.

[283] Natanael Spisak, Aleksandra M Walczak, and Thierry Mora. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic acids research*, 48(19):10702–10712, 2020.

[284] Tanja Stadler and Sebastian Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120198, 2013.

[285] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2):S231–S240, 2002.

[286] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

[287] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[288] Nicolas Strauli, Emily Kathleen Fryer, Olivia Pham, Mohamed Abdel-Mohsen, Shelley N Facente, Christopher Pilcher, Pleuni Pennings, Satish Pillai, and Ryan D Hernandez. The genetic interaction between hiv and the antibody repertoire. *BioRxiv*, page 646968, 2019.

[289] Yapeng Su, Daniel Chen, Dan Yuan, Christopher Lausted, Jongchan Choi, Chengzhen L Dai, Valentin Voillet, Venkata R Duvvuri, Kelsey Scherler, Pamela Troisch, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. *Cell*, 183(6):1479–1495, 2020.

[290] Yapeng Su, Dan Yuan, Daniel G Chen, Rachel H Ng, Kai Wang, Jongchan Choi, Sarah Li, Sunga Hong, Rongyu Zhang, Jingyi Xie, et al. Multiple early factors anticipate post-acute covid-19 sequelae. *Cell*, 185(5):881–895, 2022.

[291] Zoe Swank, Yasmeen Senussi, Zachary Manickas-Hill, Xu G Yu, Jonathan Z Li, Galit Alter, and David R Walt. Persistent circulating severe acute respiratory syndrome coronavirus 2 spike is associated with post-acute coronavirus disease 2019 sequelae. *Clinical Infectious Diseases*, 76(3):e487–e490, September 2022.

[292] Stuart G Tangye and David M Tarlinton. Memory b cells: Effectors of long-lived immune responses. *European journal of immunology*, 39(8):2065–2075, 2009.

[293] Jeroen MJ Tas, Luka Mesin, Giulia Pasqual, Sasha Targ, Johanne T Jacobsen, Yasuko M Mano, Casie S Chen, Jean-Claude Weill, Claude-Agnès Reynaud, Edward P Browne, et al. Visualizing antibody affinity maturation in germinal centers. *Science*, 351(6277):1048–1054, 2016.

[294] Tanayott Thaweethai, Sarah E Jolley, Elizabeth W Karlson, Emily B Levitan, Bruce Levy, Grace A McComsey, Lisa McCorkell, Girish N Nadkarni, Sairam Parthasarathy, Upinder Singh, et al. Development of a definition of postacute sequelae of sars-cov-2 infection. *Jama*, 329(22):1934–1946, 2023.

[295] Irani Thevarajan, Thi HO Nguyen, Marios Koutsakos, Julian Druce, Leon Caly, Carolien E van De Sandt, Xiaoxiao Jia, Suellen Nicholson, Mike Catton, Benjamin Cowie, et al. Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe covid-19. *Nature medicine*, 26(4):453–455, 2020.

[296] Jeffrey L Thorne, Hirohisa Kishino, and Ian S Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular biology and evolution*, 15(12):1647–1657, 1998.

[297] Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. Mcpas-tcr: a manually curated catalogue of pathology-associated t cell receptor sequences. *Bioinformatics*, 33(18):2924–2929, 2017.

[298] Mary M Tomayko, Natalie C Steinel, Shannon M Anderson, and Mark J Shlomchik. Cutting edge: Hierarchy of maturity of murine memory b cell subsets. *The Journal of Immunology*, 185(12):7146–7150, 2010.

[299] Hale F Trotter. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society*, 10(4):545–551, 1959.

[300] Vasiliki Tsampasian, Maria Bäck, Marco Bernardi, Elena Cavarretta, Maciej Debski, Sabiha Gati, Dominique Hansen, Nicolle Kränkel, Konstantinos C Koskinas, Josef Niebauer, et al. Cardiovascular disease as part of long covid: a systematic review. *European journal of preventive cardiology*, page zwae070, 2024.

[301] Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical biosciences*, 147(1):63–91, 1998.

[302] Jackson S Turner, Jane A O'Halloran, Elizaveta Kalaidina, Wooseob Kim, Aaron J Schmitz, Julian Q Zhou, Tingting Lei, Mahima Thapa, Rita E Chen, James Brett Case, et al. Sars-cov-2 mrna vaccines induce persistent human germinal centre responses. *Nature*, 596(7870):109–113, 2021.

[303] Jackson S Turner, Julian Q Zhou, Julianna Han, Aaron J Schmitz, Amena A Rizk, Wafaa B Alsoussi, Tingting Lei, Mostafa Amor, Katherine M McIntire, Philip Meade, et al. Human germinal centres engage memory and naive b cells after influenza vaccination. *Nature*, 586(7827):127–132, 2020.

[304] Nicolas Vabret, Graham J Britton, Conor Gruber, Samarth Hegde, Joel Kim, Maria Kuksin, Rachel Levantovsky, Louise Malle, Alvaro Moreira, Matthew D Park, et al. Immunology of covid-19: current state of the science. *Immunity*, 52(6):910–941, 2020.

[305] Viviana Valeri, Akhésa Sochon, Chaoliang Ye, Xinru Mao, Damiana Lecoeuche, Simon Fillatreau, Jean-Claude Weill, Claude-Agnès Reynaud, and Yi Hao. B cell intrinsic and extrinsic factors impacting memory recall responses to srbc challenge. *Frontiers in Immunology*, 13:873886, 2022.

[306] Daphne van Ginneken, Anamay Samant, Karlis Daga-Krumins, Andreas Agrafiotis, Evgenios Kladis, Sai T Reddy, and Alexander Yermanos. Protein language model pseudolikelihoods capture features of in vivo b cell selection and evolution. *bioRxiv*, pages 2024–12, 2024.

[307] Jason A Vander Heiden, Gur Yaari, Mohamed Uduman, Joel NH Stern, Kevin C O'connor, David A Hafler, Francois Vigneault, and Steven H Kleinstein. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13):1930–1932, 2014.

[308] Vanessa Venturi, David A Price, Daniel C Douek, and Miles P Davenport. The molecular basis for public t-cell responses? *Nature Reviews Immunology*, 8(3):231–238, 2008.

[309] Charlotte Viant, Georg HJ Weymar, Amelia Escolano, Spencer Chen, Harald Hartweger, Melissa Cipolla, Anna Gazumyan, and Michel C Nussenzweig. Antibody

affinity shapes the choice between memory and germinal center b cell fates. *Cell*, 183(5):1298–1311, 2020.

[310] Gabriel D Victora and Michel C Nussenzweig. Germinal centers. *Annual review of immunology*, 30(1):429–457, 2012.

[311] Gabriel D Victora and Michel C Nussenzweig. Germinal centers. *Annual review of immunology*, 40(1):413–442, 2022.

[312] Gabriel D Victora, Tanja A Schwickert, David R Fooksman, Alice O Kamphorst, Michael Meyer-Hermann, Michael L Dustin, and Michel C Nussenzweig. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell*, 143(4):592–605, 2010.

[313] Marcos C Vieira, Daniel Zinder, and Sarah Cobey. Selection and neutral mutations drive pervasive mutability losses in long-lived anti-hiv b-cell lineages. *Molecular biology and evolution*, 35(5):1135–1146, 2018.

[314] Nienke Vrisekoop, Ineke den Braber, Anne Bregje de Boer, An FC Ruiter, Mariëtte T Ackermans, Saskia N van der Crabben, Elise HR Schrijver, Gerrit Spierenburg, Hans P Sauerwein, Mette D Hazenberg, et al. Sparse production but preferential incorporation of recently produced naive t cells in the human peripheral pool. *Proceedings of the National Academy of Sciences*, 105(16):6115–6120, 2008.

[315] Alexandra Vujkovic, My Ha, Tessa de Block, Lida van Petersen, Isabel Brosius, Caroline Theunissen, Sabrina H Van Ierssel, Esther Bartholomeus, Wim Adriaensen, Guido Vanham, et al. Diagnosing viral infections through t-cell receptor sequencing of activated cd8+ t cells. *The Journal of Infectious Diseases*, 229(2):507–516, 2024.

[316] Meng Wang, Jonathan Patsenker, Henry Li, Yuval Kluger, and Steven H Kleinstein. Language model-based b cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic acids research*, 52(2):548–557, 2024.

[317] Meryl Wang, David Zhu, Jianwei Zhu, Ruth Nussinov, and Buyong Ma. Local and global anatomy of antibody-protein antigen recognition. *Journal of Molecular Recognition*, 31(5):e2693, 2018.

[318] Anna Weber, Jannis Born, and María Rodriguez Martínez. Titan: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement_1):i237–i244, 2021.

[319] Anna Z Wec, Denise Haslwanter, Yasmina N Abdiche, Laila Shehata, Nuria Pedreño-Lopez, Crystal L Moyer, Zachary A Bornholdt, Asparouh Lilov, Juergen H Nett, Rohit K Jangra, et al. Longitudinal dynamics of the human b cell response to the yellow fever 17d vaccine. *Proceedings of the National Academy of Sciences*, 117(12):6675–6685, 2020.

[320] Anna Z Wec, Daniel Wrapp, Andrew S Herbert, Daniel Maurer, Denise Haslwanter, Mrunal Sakharkar, Rohit K Jangra, M Eugenia Dieterle, Asparouh Lilov, Deli Huang, et al. Broad sarbecovirus neutralizing antibodies define a key site of vulnerability on the sars-cov-2 spike protein. *BioRxiv*, 2020.

[321] Florian J Weisel, Griselda V Zuccarino-Catania, Maria Chikina, and Mark J Shlomchik. A temporal switch in the germinal center determines differential output of memory b and plasma cells. *Immunity*, 44(1):116–130, 2016.

[322] Daniela Weiskopf, Katharina S Schmitz, Matthijs P Raadsen, Alba Grifoni, Nisreen MA Okba, Henrik Endeman, Johannes PC van den Akker, Richard Molenkamp, Marion PG Koopmans, Eric CM van Gorp, et al. Phenotype and kinetics of sars-cov-2–specific t cells in covid-19 patients with acute respiratory distress syndrome. *Science immunology*, 5(48):eabd2071, 2020.

[323] Carl A Wesolowski, Surajith N Wanasundara, Michal J Wesolowski, Belkis Erbas, and Paul S Babyn. A gamma-distribution convolution model of 99m tc-mibi thyroid time-activity curves. *EJNMMI physics*, 3:1–19, 2016.

[324] Gregory P Williams, Esther Dawen Yu, Kendra Shapiro, Eric Wang, Antoine Freuchet, April Frazier, Cecilia S Lindestam Arlehamn, Alessandro Sette, and Ricardo da Silva Antunes. Investigating viral and autoimmune t cell responses associated with post-acute sequelae of covid-19. *Human Immunology*, 85(3):110770, 2024.

[325] Andrea C Wong, Ashwarya S Devason, Iboro C Umana, Timothy O Cox, Lenka Dohnalová, Lev Litichevskiy, Jonathan Perla, Patrick Lundgren, Zienab Etwebi, Luke T Izzo, et al. Serotonin reduction in post-acute sequelae of viral infection. *Cell*, 186(22):4851–4867, 2023.

[326] Rachel Wong, Julia A Belk, Jennifer Govero, Jennifer L Uhrlaub, Dakota Reinartz, Haiyan Zhao, John M Errico, Lucas D'Souza, Tyler J Ripperger, Janko Nikolich-Zugich, et al. Affinity-restricted memory b cells dominate recall responses to heterologous flaviviruses. *Immunity*, 53(5):1078–1094, 2020.

[327] Linda Wooldridge, Julia Ekeruche-Makinde, Hugo a Van Den Berg, Anna Skowera, John J Miles, Mai Ping Tan, Garry Dolton, Mathew Clement, Sian Llewellyn-Lacey,

David A Price, et al. A single autoimmune t cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry*, 287(2):1168–1177, 2012.

[328] Jens Wrammert, Kenneth Smith, Joe Miller, William A Langley, Kenneth Kokko, Christian Larsen, Nai-Ying Zheng, Israel Mays, Lori Garman, Christina Helms, et al. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature*, 453(7195):667–671, 2008.

[329] David Wu, Anna Sherwood, Jonathan R Fromm, Stuart S Winter, Kimberly P Dunsmore, Mignon L Loh, Harvey A Greisman, Daniel E Sabath, Brent L Wood, and Harlan Robins. High-throughput sequencing detects minimal residual disease in acute t lymphoblastic leukemia. *Science translational medicine*, 4(134):134ra63–134ra63, 2012.

[330] Joseph T Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M de Salazar, Benjamin J Cowling, Marc Lipsitch, and Gabriel M Leung. Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. *Nature medicine*, 26(4):506–510, 2020.

[331] Xueling Wu, Zhenhai Zhang, Chaim A Schramm, M Gordon Joyce, Young Do Kwon, Tongqing Zhou, Zizhang Sheng, Baoshan Zhang, Sijy O'Dell, Krisha McKee, et al. Maturation and diversity of the vrc01-antibody lineage over 15 years of chronic hiv-1 infection. *Cell*, 161(3):470–485, 2015.

[332] Yan Wu, Feiran Wang, Chenguang Shen, Weiyu Peng, Delin Li, Cheng Zhao, Zhaohui Li, Shihua Li, Yuhai Bi, Yang Yang, et al. A noncompeting pair of human neutralizing antibodies block covid-19 virus binding to its receptor ace2. *Science*, 368(6496):1274–1278, 2020.

[333] Yu Wu, Liangyu Kang, Zirui Guo, Jue Liu, Min Liu, and Wannian Liang. Incubation period of covid-19 caused by unique sars-cov-2 strains: a systematic review and meta-analysis. *JAMA network open*, 5(8):e2228008–e2228008, 2022.

[334] Yan Xie, Evan Xu, Benjamin Bowe, and Ziyad Al-Aly. Long-term cardiovascular outcomes of covid-19. *Nature Medicine*, 28(3):583–590, February 2022.

[335] Evan Xu, Yan Xie, and Ziyad Al-Aly. Long-term gastrointestinal outcomes of covid-19. *Nature communications*, 14(1):983, 2023.

[336] Ying Xu, Xinyang Qian, Yao Tong, Fan Li, Ke Wang, Xuanping Zhang, Tao Liu, and Jiayin Wang. Attntap: a dual-input framework incorporating the attention mechanism for accurately predicting tcr-peptide binding. *Frontiers in Genetics*, 13:942491, 2022.

[337] Gur Yaari, Jennifer IC Benichou, Jason A Vander Heiden, Steven H Kleinstein, and Yoram Louzoun. The mutation patterns in b-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140242, 2015.

[338] Ziheng Yang. *Molecular evolution: a statistical approach*. Oxford University Press, 2014.

[339] Alexander Yermanos, Andreas Agrafiotis, Raphael Kuhn, Damiano Robbiani, Josephine Yates, Chrysa Papadopoulou, Jiami Han, Ioana Sandu, Cédric Weber, Florian Bieberich, et al. Platypus: an open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR genomics and bioinformatics*, 3(2):lqab023, 2021.

[340] Alexander Dimitri Yermanos, Andreas Kevin Dounas, Tanja Stadler, Annette Oxenius, and Sai T Reddy. Tracing antibody repertoire evolution by systems phylogeny. *Frontiers in immunology*, 9:2149, 2018.

[341] Kailin Yin, Michael J Peluso, Xiaoyu Luo, Reuben Thomas, Min-Gyoung Shin, Jason Neidleman, Alicer Andrew, Kyrlia C Young, Tongcui Ma, Rebecca Hoh, et al. Long covid manifests with t cell dysregulation, inflammation and an uncoordinated adaptive immune response to sars-cov-2. *Nature Immunology*, 25(2):218–225, 2024.

[342] Anne D Yoder and Ziheng Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular biology and evolution*, 17(7):1081–1090, 2000.

[343] Meng Yuan, Nicholas C Wu, Xueyong Zhu, Chang-Chun D Lee, Ray TY So, Huibin Lv, Chris KP Mok, and Ian A Wilson. A highly conserved cryptic epitope in the receptor binding domains of sars-cov-2 and sars-cov. *Science*, 368(6491):630–633, 2020.

[344] S. Zacks. Generalized Integrated Telegraph Processes and the Distribution of Related Stopping Times. *Journal of Applied Probability*, 41(2):497–507, 2004. Publisher: Applied Probability Trust.

[345] Shelemyahu Zacks. Distribution of the Total Time in a Mode of an Alternating Renewal Process with Applications. *Sequential Analysis*, 31(3):397–408, July 2012. Publisher: Taylor & Francis.

[346] Chao Zhang, Andrey V Bzikadze, Yana Safonova, and Siavash Mirarab. A scalable model for simulating multi-round antibody evolution and benchmarking of clonal tree reconstruction methods. *Frontiers in immunology*, 13:1014439, 2022.

[347] Yu Zhang, Xingxing Jian, Linfeng Xu, Jingjing Zhao, Manman Lu, Yong Lin, and Lu Xie. itcep: a deep learning framework for identification of t cell epitopes by harnessing fusion features. *Frontiers in Genetics*, 14:1141535, 2023.

[348] Zhehao Zhang. Renewal sums under mixtures of exponentials. *Applied Mathematics and Computation*, 337:281–301, 2018.

[349] Daming Zhou, Helen ME Duyvesteyn, Cheng-Pin Chen, Chung-Guei Huang, Ting-Hua Chen, Shin-Ru Shih, Yi-Chun Lin, Chien-Yu Cheng, Shu-Hsing Cheng, Yhu-Chering Huang, et al. Structural basis for the neutralization of sars-cov-2 by an antibody from a convalescent patient. *Nature structural & molecular biology*, 27(10):950–958, 2020.

[350] Ksenia V Zornikova, Alexandra Khmelevskaya, Savely A Sheetikov, Dmitry O Kiryukhin, Olga V Shcherbakova, Aleksei Titov, Ivan V Zvyagin, and Grigory A Efimov. Clonal diversity predicts persistence of sars-cov-2 epitope-specific t-cell response. *Communications biology*, 5(1):1351, 2022.

[351] Seth J Zost, Pavlo Gilchuk, Rita E Chen, James Brett Case, Joseph X Reidy, Andrew Trivette, Rachel S Nargi, Rachel E Sutton, Naveenchandra Suryadevara, Elaine C Chen, et al. Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the sars-cov-2 spike protein. *Nature medicine*, 26(9):1422–1427, 2020.

[352] Svitlana Zubchenko, Iryna Kril, Olena Nadizhko, Oksana Matsyura, and Valentyna Chopyak. Herpesvirus infections and post-covid-19 manifestations: a pilot observational study. *Rheumatology International*, 42(9):1523–1530, June 2022.

[353] Griselda V Zuccarino-Catania, Saheli Sadanand, Florian J Weisel, Mary M Tomayko, Hailong Meng, Steven H Kleinstein, Kim L Good-Jacobson, and Mark J Shlomchik. Cd80 and pd-l2 define functionally distinct memory b cell subsets that are independent of antibody isotype. *Nature immunology*, 15(7):631–637, 2014.

[354] Emile Zuckerkandl. Molecular disease, evolution, and genic heterogeneity. *Horizons in biochemistry*, pages 189–225, 1962.

[355] Emile Zuckerkandl and Linus Pauling. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier, 1965.

[356] Yves Zurbuchen, Jan Michler, Patrick Taeschler, Sarah Adamo, Carlo Cervia, Miro E Raeber, Ilhan E Acar, Jakob Nilsson, Klaus Warnatz, Michael B Soyka, et al. Human memory b cells show plasticity and adopt multiple fates upon recall response to sars-cov-2. *Nature immunology*, 24(6):955–965, 2023.