The Thesis Committee for Chenfeng He
Certifies that this is the approved version of the following Thesis:

DEVELOPMENT OF COMPUTATIONAL METHODS FOR IMMUNE REPERTOIRE ANALYSIS: FROM SEQUENCE TO SPECIFICITY

COMMITTEE:

Ning Jiang, Supervisor

Mia Markey

Pengyu Ren

Keke Chen

DEVELOPMENT OF COMPUTATIONAL METHODS FOR IMMUNE REPERTOIRE ANALYSIS: FROM SEQUENCE TO SPECIFICITY

By

Chenfeng He

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

In Partial Fulfillment

Of the Requirements

For the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2018

Dedication

To my family, thank you for all your support and understanding during my graduate study! I would easily give up without your support!

Acknowledgements

I would like to thank my fellow Jiang lab and Ren lab members for all your help throughout the years. Ben Wendel and Keyue Ma, for all your help on immunology and generating valuable high quality experimental data. Di Wu, for guiding me into the computational immunology field. Chengwen Liu and David Bell, for always helping me when I have some problem on the HPC clusters. Changsheng Zhang, Qiantao Wang, Zhifeng Jing, Sara Cheng and Xiaojia Mu, for your advice on molecular modeling and hints on understanding TCR antigen recognition from the molecular structure point of view. Chad Williams, Michael Malone and Alex Schonnesen, for all the valuable suggestions on biology. Brittain Sobey, Michale Don, Elynna Day and Lacy White, for all your help on the administrative standpoint. Mingjuan Qu, Evan Cohan, Qian Shi, Tian Li, Wei Wang, Jie Lan and everyone else in the labs, for all your help and company along the way.

I would like to acknowledge Pengyu Ren, Keke Chen, Jun Xiao and Mia Markey, for your help and advice on computational methods development and validation. Luara Su and Daniel Alcazar, for gathering and providing HIV patient samples. James Pollard, for the help to maintain my internet access security. My wife, Ting Guo, for accompany me during my whole PhD procedure.

Finally, I would like to thank my advisor, Ning Jiang, for all your advice and direction during my PhD study; for your patience and support to help me slowly catch up on immunology. None of this would happen without you!

DEVELOPMENT OF COMPUTATIONAL METHODS FOR IMMUNE REPERTOIRE ANALYSIS: FROM SEQUENCE TO SPECIFICITY

Chenfeng He, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Ning Jiang

The immune system plays a key role in maintaining human health. Accurately characterizing the immune receptors with immune repertoire sequencing (IRseq) provides an essential way for understanding the adaptive immune system. Towards this goal, we developed a bioinformatics tool, Molecular Identifier Clustering-based IR-Seq (MIDICRS), to quantitatively measure immune repertoire. We have demonstrated MIDCIRS' accuracy, high coverage and wide dynamic range, which allow us to analyze various types of immune repertoires.

Immune repertoire is continuously shaped by encountered antigens; thus, its components reflect an individual's historical disease status. We applied MIDCIRS to measure the antibody repertoire from malaria-experienced individuals and found unexpected mutable capability of infants adaptive immune system. We also used MIDCIRS to measure Follicular helper T cells (Tfhs) directly obtained from untreated HIV patients' lymph nodes and found (1) evidence for intact antigen-driven clonal expansion of Tfh cells and (2) selective utilization of specific complementarity-determining region 3 (CDR3) motifs during chronic HIV infection. Both studies demonstrated MIDCIRS functionality and versatility for studying antigen driven immune response.

Bridging the gap between immune receptor sequences and their biological function (i.e. antigen specificity) is attractive and useful for directly measuring immune repertoire changes with respect to pathogen infection. Using experimentally validated CD8+ TCR sequences with their antigen specificity, we developed a computational tool, Linear programming based Motif Pick and Enrichment analysis for Tcrs (LiMPETs), to find significant motifs within the TCR CDR3 region for determining antigen specificity. We demonstrated LiMPETs' advantages by comparing with existing tools on both public and in-house data.

TABLE OF CONTENT

List of Tables

List of Figures

Chapter 1: Background

1.1 IMMUNE SYSTEM

Our immune system serves as a complex, multifaceted 'army' that continuously protects our health. It fights against threats from foreign invaders (e.g., bacteria, virus) by killing pathogens and infected cells[1]; it also fights against internal threats (e.g., cancerous cells, misfolded proteins) by monitoring and destroying damaged cells[2]. Two main immunity strategies comprise the human immune system, one is the so-called 'Innate immunity', which is responds quickly and provides non-specific defense to pathogens; the other is called 'Adaptive immunity', which is 'learned' and provides specific defense to certain pathogens. In this thesis, the work summarized focuses on Adaptive immunity.

Innate immunity contains non-specific defense to pathogens, this system includes the skin barrier around the body, mucus to trap pathogens, hairs to move mucus trapped pathogens out and neutrophils to kill invading bacteria, etc. This system provides immediate protection against pathogen infection, and the protection is non-specific, which means it does not distinguish among pathogens[3].

Compared with innate immunity, the other immune strategy is 'Adaptive immunity', which has a slower response but is much more complicated and provides long-lasting protection against specific invading pathogens.

1.2 ADAPTIVE IMMUNITY

Adaptive immunity is also known as 'acquired immunity'[3], and as its name implies, this type of immune response is acquired and 'learned' after the immune system encounters antigens. After the immune system first encounters a certain type of pathogen, it will generate immune cells specific to this pathogen and 'memorize' it. In this way immune system can respond and subdue this pathogen very quickly when it comes across the pathogen for a second time. People have been making use of this property of the immune system for a long time, i.e., various types of vaccines have been developed since 1796[4] to artificially 'pre-train' the adaptive immune system to respond to an actual late-stage infection, including influenza[5], HIV[6], several types of cancers[7,8], etc.

The adaptive immune system contains two types of lymphocytes, which are the T-lymphocyte and B-lymphocyte. Both of them are initially differentiated from stem cells in bone marrow[9]. After their initial differentiation, B cells will continue to mature in bone marrow, while T cells will migrate to thymus and mature there[10]. Both of them express cell surface receptors which determine their antigen specificity, B cells express B cell receptor (BCR) while T cells express T cell receptor (TCR). B cells can secrete their BCR, which will become a so-called 'antibody' acting to neutralize infected pathogens. In contrast, the TCR remains on T cell surface where it is used to contact and recognize other cells, e.g. antigen-presenting cells (APC), and become activated. All the BCRs within an individual are termed as the 'antibody repertoire', while all the TCRs are termed the 'TCR repertoire'. The recognition of TCR/BCR to antigen can induce clonal expansion of the lymphocytes, which results in further composition change of the immune repertoire.

Generally speaking, T lymphocytes can be divided into CD8+ T cells and CD4+ T cells, depending on which glycoprotein (CD4 or CD8) is expressed on their surface[3]. These glycoproteins determine the lymphocytes' function: CD8+ T cells can recognize peptides presented by MHC class I molecules (MHC-I), which are expressed by all nucleated cells, and can destroy virus infected cells and tumor cells. CD4+ T cells recognize antigen peptides presented by MHC class II molecules (MHC-II), which are expressed by antigen-presenting cells, and upon activation they moderate and recruit other immune cells to the antigen source.

All nucleated cells express MHC-I molecules on their surface and this molecule constantly samples peptides from inside the cell[11]. If there are some abnormal changes within the cell, e.g., virus infected cells express virus proteins, tumor cells express cancer neoantigens[12], peptides from these proteins are going to be randomly sampled by MHC-I and presented to TCR on CD8+ T cells. CD8+ T cells that recognize these peptides as non-self-antigens will be activated and release granzymes to trigger abnormal cells apoptosis. This is why CD8+ T cells are also called cytotoxic T cells because they can trigger cell apoptosis.

Compared to CD8+ T cells, CD4+ T cells are much more heterogeneous. CD4+ T cells perform supporting or regulation roles in immune response, which are termed 'cell-mediated immunity'. CD4+ T cells that function as supporting roles are generally called helper T cells (Th cells). According to their cytokine/transcriptome files/function differences, helper T cells can be distinguished into several types, e.g., Th1 cells, Th2 cells and Th17 cells. These cells can be activated through TCR recognizing antigen peptides or by cytokine secreted from other immune cells[13]. Activated Th1 cells produce cytokine IFN-γ and IL-2, which activate macrophages and CD8+ T cells to induce monocytic inflammation, which kills intracellular bacteria, protozoans and viruses. Activated Th2 cells produce IL-13, IL-4,IL-5,IL-25, etc., which induces

eosinophilic/basophilic/mast cell inflammation to kill helminths. Activated Th17 cells produce IL-17, IL-22 etc, which induce neutrophilic inflammation to kill extracellular bacteria and fungi[14]. Another specialized type of helper T cells is follicular helper T cells (Tfh cells), which mainly provide help to B cells within secondary lymphoid organs (e.g., lymph node, spleen and tonsil), as well as induce and stabilize the formation of germinal center within B cell follicles located in secondary lymphoid organs[15]. Finally, regulatory T cells (Tregs[16]) produce IL-10, TGF-β, etc., which suppresses immune response and are important for autoimmune diseases.

1.3 FOUR SOURCES OF T/B RECEPTOR DIVERSITY

Because of their antigen specificity, one T(B)CR can only recognize a certain antigen (or a few antigens, because of cross reactivity). One natural question to ask is how our immune system can recognize such a huge amount of different antigens within our living environment. The answer is our body maintains a great diversity of T(B)CRs pool, and the generation of their diversity lies at the DNA sequence level.

Lymphocytes can induce irreversible changes to their genomic DNA sequences, which results in the great diversity of T(B)CRs pool. The procedure of these changes differs slightly between TCR and BCR, but the overall methods are similar with the key part being a step called 'V(D)J recombination'[3] (Figure 1.1). Take the BCR molecule as an example, which is composed of two separate chains: light chain and heavy chain. For the heavy chain, the precursor cells of B cells have 38-46 Variable (V) genes, 23 Diversity (D) genes and 6 Joining (J) genes in their genomic DNA. For each B cell to mature from stem cells into naive B cells, one V gene, one D gene and one J gene will be randomly picked out and joined together to become a new DNA sequence for translating into BCR amino acid sequence. During the recombination process, random insertion or deletion of nucleotides can be added into the junction region between V-D

and D-J. This process can induce a great amount of changes on its protein product, including BCR heavy chain length, reading frame or even stop codon which results into an 'unproductive' VDJ recombination. This VDJ recombination step is shared between TCR and BCR, except TCR alpha chain and BCR light chain only have V genes and J genes, without D genes. For BCR specifically, after the recombination step, random mutations can be introduced into its DNA sequence. Termed 'somatic hyper mutation' (SHM), this induces another level of diversity for BCRs. The junction region between V gene and J gene (including D gene in TCR-beta/BCR-heavy chain) is called Complementary Determining Region 3 (CDR3), which (as its names implies) plays an important role in determining T/BCR's antigen specificity, because it contains the most variability and has the most contact probability with antigen peptides[17].

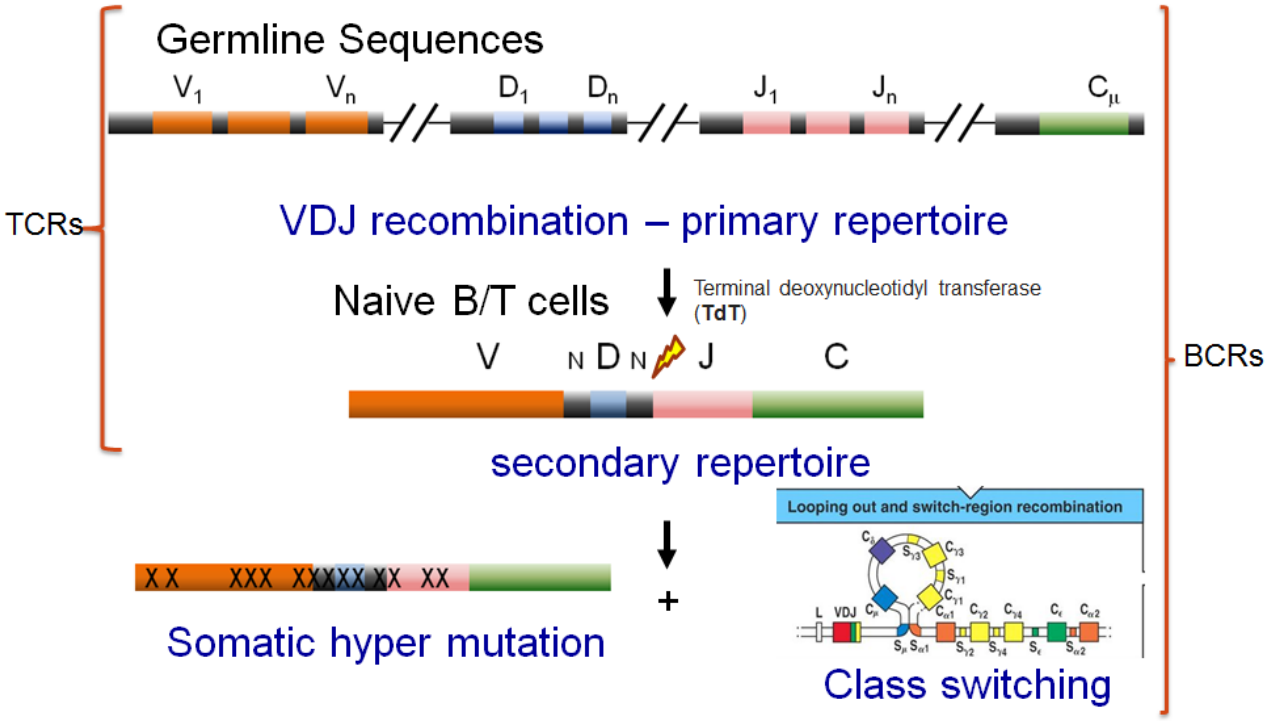


Figure 1.1: Immune repertoire diversity generation

To sum up, there are in total four sources of T(B)CR repertoire diversity:

1). V(D)J combination diversity, which is all the possible combinations between V genes, D genes, J genes;

2). Junction region diversity, which can induce DNA reading frame changes due to insertion/deletion of nucleotides;

3). Alpha-beta chain (TCR) or heavy-light chain (BCR) combinatorial diversity, which is due to the random combination between two chains of the duplex;

4). Somatic hyper mutation diversity, which is introduced by random point mutations within BCR sequences.

It was estimated, this process can generate 10^{13} different BCRs[18] and 10^{18} different TCRs in theory[19], although the repertoire of one individual is much less diversified and the diversity partially depends on the previous antigen experience history of the immune repertoire.

1.4 MID BASED IMMUNE REPERTOIRE SEQUENCING

T/B cells recognize antigens and the recognition will induce changes in the immune repertoire (e.g., T cell clonal expansion will change the composition of TCRs). In other words, immune cells recognize antigens and clear pathogens while pathogens in return shape repertoire, resulting in the change of repertoire composition. Thus, the repertoire becomes a hallmark of an individual's disease states and antigen experience. Traditionally it's difficult to quantify immune repertoire given its huge diversity; however, due to technological advances, researchers nowadays can measure immune repertoire by high-throughput sequencing the T/B receptor sequences in a method called Immune Repertoire Sequencing (IR-Seq)[20].

With the development of high-throughput sequencing, genomic sequencing is becoming widely available and the price has dropped significantly since 2001: human whole genome sequencing has been dropped from $10 million to $1,000 during the past decades[21] and the price will continue to drop, which allows the wide application of IR-Seq in both research and clinical area. However, all current sequencing platforms are suffering from sequencing error, e.g., Illumina MiSeq's reporting error rate is 0.5% errors per bp [22]. This error rate may not be a big problem in some research areas, for example, whole transcriptome sequencing to analyze gene expression value where the sequencing errors can easily corrected by a reference template[23]. However in IR-Seq, due to the random insertion/deletion/mutation, genome template reference cannot be used to distinguish between real mutations and sequencing error. This is extremely important in immune research, because in some cases one single amino acid change can result in totally different antigen specificity[24]. An additional source of error, PCR error, will also happen during the PCR process[25].

In order to perform an accurate measurement of immune repertoire, techniques which can eliminate sequencing/PCR errors are essential. Researchers have developed Molecular identifiers (MID) to reduce the sequencing error rate within IR-Seq [26,27]. MIDs are short, randomly synthesized DNA sequences which can be tagged to cDNAs during reverse transcription. The tagged MIDs go through PCR amplification and high-throughput sequencing in conjunction with their labeled cDNAs. Because MID labelling happens before PCR amplification, all the sequencing reads originating from the same cDNA will be tagged with the same MID. Thus, sequencing reads can be grouped based on their associated MIDs and nucleotides at each position within the original cDNA can be determined by finding a consensus. This process is essentially involves sequencing the same position multiple times, and because the chance of the

same error to happen at the same position is low, the consensus will average out the sequencing errors and improve the sequencing accuracy. MIDs are important to achieve a high sequence quality for IR-Seq, but the accuracy can still be improved as we will discuss in the next Chapter.

Chapter 2: MIDCIRS-Molecular IDentifier Clustering-based Immune Repertoire Sequencing*

2.1 INTRODUCTION

V(D)J recombination can create hundreds of billions of antibodies and T cell receptors that collectively serve as the immune repertoire to protect the host from pathogens. Somatic hypermutation (SHM) further diversifies the antibody repertoire, which makes it impossible to quantify this diversity with nucleotide resolution until the development of high-throughput sequencing-based immune repertoire sequencing (IR-seq)[20,28–30]. Although we and others have developed methods to control for artifacts from high amplification bias and sequencing error rates through data analysis[29,31–35], obtaining accurate sequencing information has now been made possible by the use of molecular identifiers (MID)[26,27,36,37]. MIDs serve as barcodes to track genes of interest through amplification and sequencing. They are short stretches of nucleotide sequence tags composed of randomized nucleotides that are usually tagged to cDNA during reverse transcription to identify sequencing reads that originated from the same mRNA transcript. After PCR amplification, sequencing reads labelled with the same MID are amplified from the same mRNA transcript.

---

*Ma KY†, **He C.**†, et al. Immune Repertoire Sequencing Using Molecular Identifiers Enables Accurate Clonality Discovery and Clone Size Quantification. Front. Immunol. (2018). doi:10.3389/fimmu.2018.00033. K-YM performed all library preparation, data analysis, and wrote the manuscript; CH developed MIDCIRS-TCR analysis pipeline and RNA copy number simulation model; BW helped with naive T cell sorting and manuscript editing; CW helped with CMV-specific T cell sorting and CMV-specific T cell line culture; JX helped to optimize MIDCIRS pipeline. HY helped with sequencing. NJ conceived the idea, designed the study, directed data analysis, and revised the manuscript with contributions from all coauthors.

†: These authors contributed equally.

Despite these advancements, there are still several challenges for correctly applying MID technique in IR-seq: 1). The large amount of input RNA required and low diversity coverage make it challenging to analyze small numbers of cells, such as memory B cells from dissected tissues or blood draws from young children, using IR-seq because these samples require many PCR cycles to generate enough material to make sequencing libraries, thus exacerbating PCR bias and errors; 2). Same MID may label more than one RNA molecules, which result in chimera sequences (Figure 2.1); 3). Erroneous MIDs resulting from PCR or sequencing errors make accurate MID counting difficult.

Here we report the development of MID clustering-based IR-seq (MIDCIRS) that further separates different RNA molecules tagged with the same MID. Using naive B cells, we demonstrate that MIDCIRS has a high coverage of the diversity estimate, or different types of antibody sequences, that is consistent with the input cell number and a large dynamic range of three orders of magnitude compared to other MID-based immune repertoire-sequencing methods[26,27]. We applied MIDCIRS on CD8+ TCR repertoire with various RNA input amount, based on which we demonstrate the necessity of performing MID sub-clustering to eliminate erroneous sequences, the method for eliminating erroneous MIDs and how to estimate T cell clone size from RNA molecule counting. Given the wide use of IR-seq in basic research[32] as well as clinical settings[38], we believe the method outlined here will serve as an important guideline for future IR-seq experimental designs.

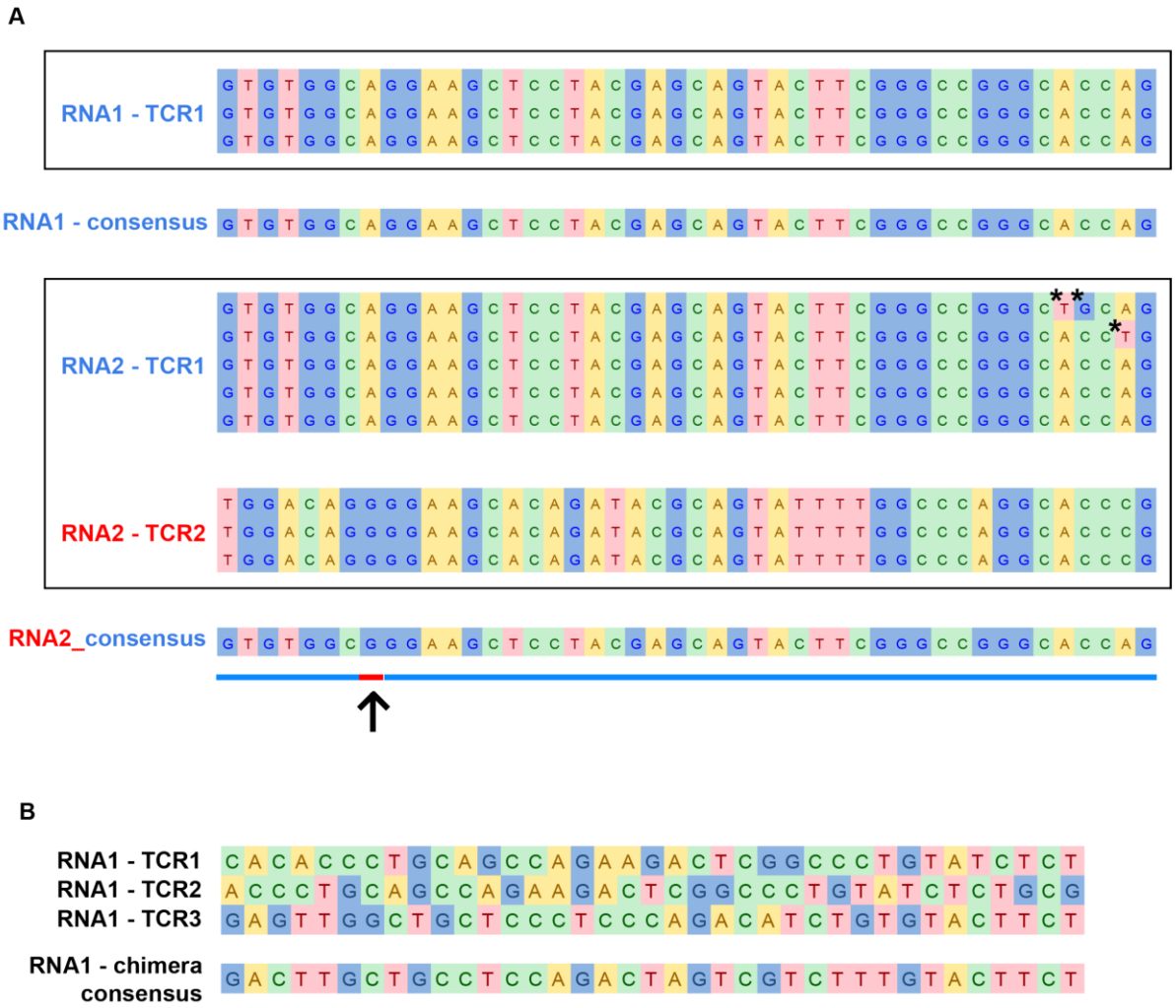


Figure 2.1: Representative demonstration of chimera consensus sequences generated without sub-clustering. (A). Two different TCR RNAs (RNA2-TCR1 and RNA2-TCR2) were tagged with the same MID (RNA2), while one of the TCRs (TCR1) has a sister RNA tagged by another MID (RNA1). After building consensus sequence weighted by quality score and number of reads at each nucleotide position, a chimera consensus sequence was generated from RNA2-tagged TCR sequences (Top box, TCR1 tagged with RNA1; bottom box, two TCR sequences tagged with same MID; *, sequencing or PCR errors that are removed in the consensus building; sequence outside the top box, true TCR1 consensus sequence; sequence outside the bottom box, chimera consensus sequence; arrow, chimera nucleotide base that differs from the rest of consensus sequence was generated by weighing read number and quality score at each nucleotide). (B) Multiple singleton TCR RNAs were tagged with the same MID (RNA1) that were generated by either sequencing or PCR errors. without sub-clustering, these singletons failed to be removed and a chimera consensus sequence was generated.

2.2 RESULTS

2.2.1 MIDCIRS Sub-Clustering Improves Repertoire Diversity Estimation Accuracy

Molecular identifiers have been adopted in IR-seq and DNA/RNA sequencing to reduce error rate. However, during reverse transcription, multiple transcripts could stochastically be tagged with same MID. Previous strategies relied on increasing the length of MID to reduce the probability of non-unique MID tagging when the total RNA molecule copy number was either unknown or very large[39]. However, longer MID length could reduce the efficiency of reverse transcription[40,41]. Thus, we developed a more generalized approach (MIDCIRS) with reduced MID length.

Figure 2.2 shows the overview of MIDCIRS method. Briefly, we fixed the MID length at 12 random nucleotides and developed a generalized approach to identify each individual transcript using a sequence similarity-based clustering method to separate a group of sequencing reads with the same MID into sub-groups. Consensus sequences are then built by taking the average nucleotide at each position within a sub-group, weighted by their quality score.

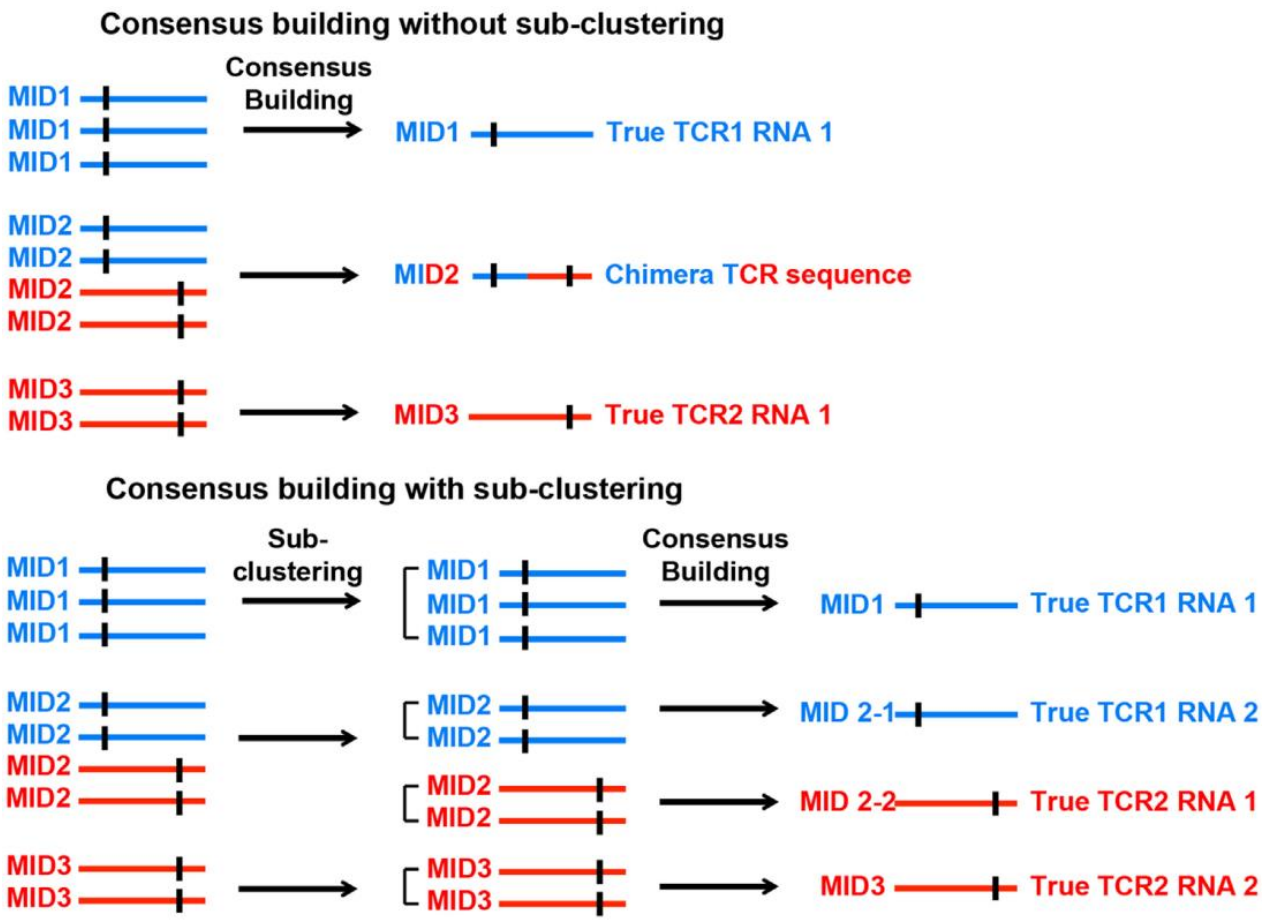


Figure 2.2: Overview of MIDCIRS method. Illustration of consensus TCR sequence building without (top) and with (bottom) sub-clustering. Top: without sub-clustering, chimera sequences are generated when different TCR RNA molecules are tagged with the same MID; bottom: TCR RNA molecules that are tagged with same MID are sub-clustered to reveal truly represented TCR sequences. Short vertical black lines indicate nucleotide differences between two TCR sequences.

Sub-clustering threshold determination

An appropriate clustering threshold must be large enough to cover sequencing reads originated from the same RNA, while be stringent enough to distinguish reads from different RNAs. In order to determine a suitable clustering threshold, we used two template RNAs with known sequences and adopted Levenshtein distance[42] between sequencing reads with the templates to quantify the errors accumulated on reads. As in shown in Figure 2.3, we used 150nt sequencing length, >99% of reads from both template sequences can be covered at threshold of 23

Levenshtein distance, so we set 15% of sequencing length as a threshold for sub-cluster sequence similarity.

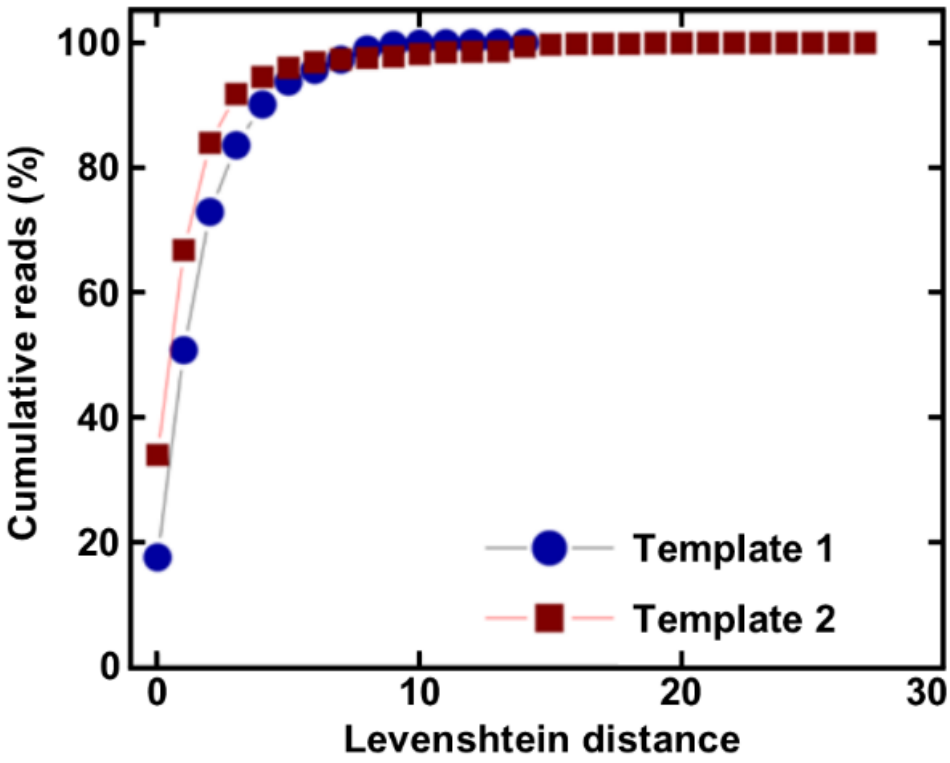


Figure 2.3: Cumulative distribution of reads as a function of Levenshtein distance between RNA control templates and sequencing reads. We performed two replicated experiment to test the robustness of our determined threshold.

Theoretical and experimental verified MIDs need sub-clustering

We reason that in order to comprehensively quantify the overall diversity, a large portion of its RNA must be sampled. However, this will inevitably increase the number of TCR transcripts that need to be tagged with MIDs, which increases the portion of MIDs tagging multiple TCR transcripts. We sought to closely examine the relationship between RNA input and multiple TCR RNAs tagging by the same MID.

The process of MID labeling can be modeled as a Poisson distribution, given the total number of MIDs being M and the number of target molecules being N, the probability that a unique MID will occur k time(s) is:

$$P_k = \frac{(\frac{N}{M})^k}{k!} \times e^{-\frac{N}{M}} \quad (1)$$

Thus, P0 and P1 are the probability that a MID will be tagged 0 and 1 time respectively and the percentage of MIDs that need sub-clustering, F(k>1), is given by:

$$F(k > 1) = \frac{(1-P0-P1)}{(1-P0)} = \frac{[1-e^{-\frac{N}{M}}-\frac{N}{M}\times e^{-\frac{N}{M}}]}{1-e^{-\frac{N}{M}}} \quad (2)$$

The percentage of MIDs with sub-clusters follows an approximate linear trend when the copies of target RNA molecules are less than 5,000,000 (Figure 2.4B). To experimentally validate this, we applied MIDCIRS TCR-seq on a range of sorted naive CD8+ T cells (from 20,000 to 1 million) with three different RNA inputs (10, 30, and 50%). As expected, we found that the observed percentage of MIDs that need sub-clustering is approximately linear with respect to copies of target RNA molecules used in this study (Figure 2.4A). With the highest amount of RNA molecules used in this study, approximately 8.5% of MIDs require further clustering, while previous method treated these sequences as ambiguous [26]. Thus, MIDCIRS sub-clustering significantly improves repertoire diversity coverage.

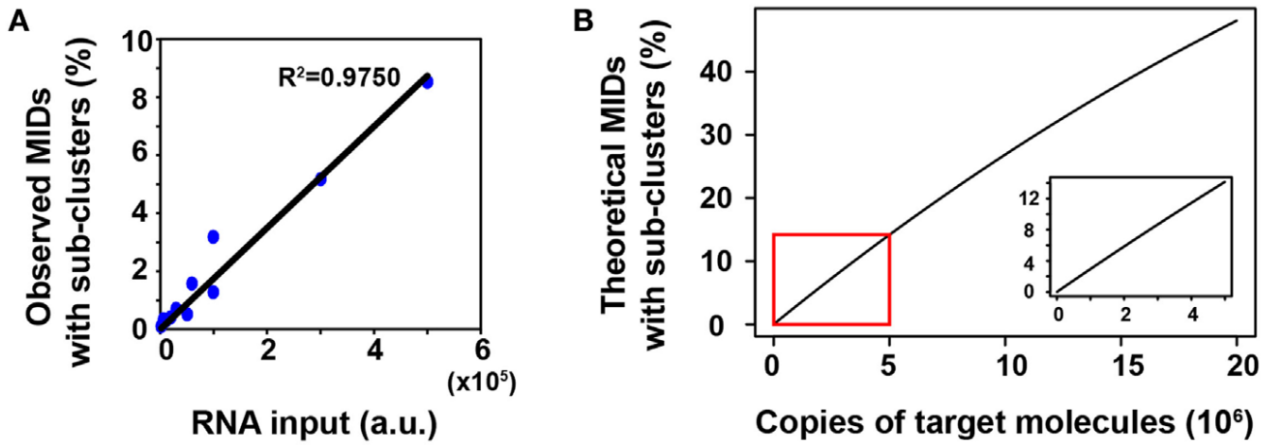


Figure 2.4: Theoretical and experimental verified MIDs need sub-clustering. (A) The percentage of observed molecular identifiers (MIDs) containing sub-clusters is linearly dependent on RNA input, which is defined as cell number multiplied by percentage of RNA (e.g., 20,000 cells with 10%RNA is equivalent to 2,000 RNA input). Line represents linear regression fit, F-test on the slope, $p < 10^{-9}$. (B) The theoretical percentage of MIDs with sub-clusters is approximately linearly dependent on copies of target molecules when copies of target molecules are less than 5,000,000 (bottom right insert).

To evaluate the accuracy of the sub-clustering step by an alternative means, we examined the TCR sequence lengths within MIDs that contain sub-clusters. We reason that if indeed each TCR RNA molecule was tagged with a unique MID, then the lengths of CDR3 for all reads would be identical under each MID. However, we showed that of the 8.5% of MIDs that contain sub-clusters, about 87% of MIDs contain TCR sequencing reads of different CDR3 lengths while only 13% have the same length for one million naive CD8+ T cells (50% RNA input). After performing sub-clustering, over 97% of sub-clusters have a uniform length (Figure 2.5), demonstrating the accuracy of sub-clustering step in MIDCIRS from another point of view.

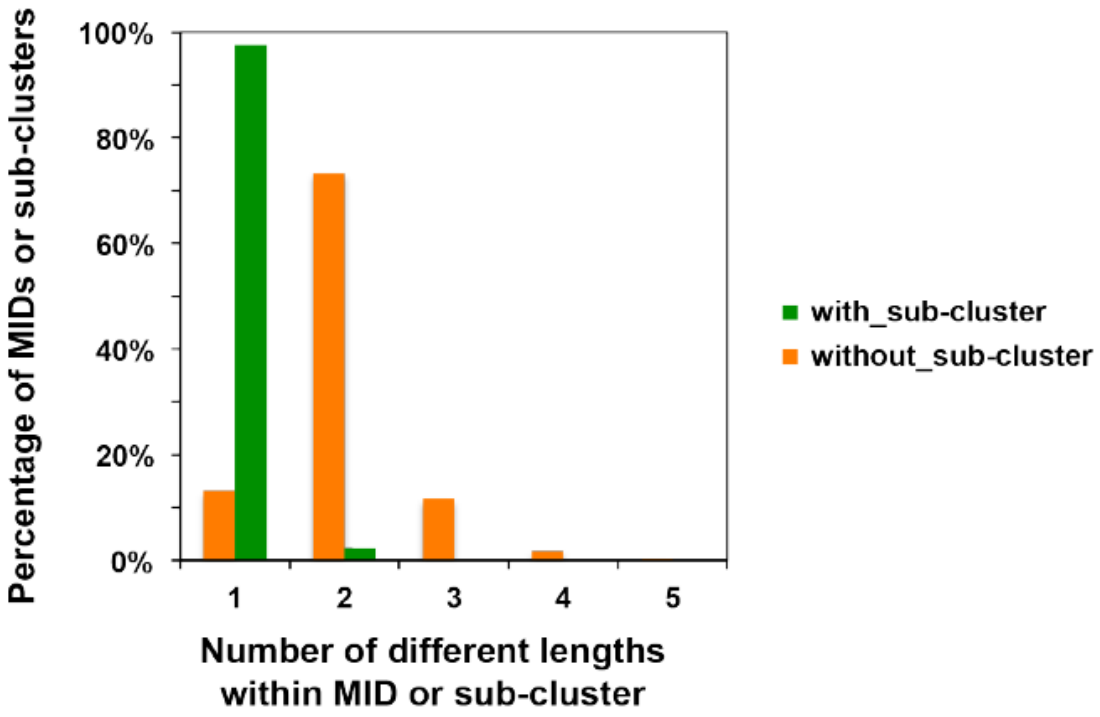


Figure 2.5: CDR3 length differences within multi-RNA containing MIDs before and after sub-clustering. The number of different CDR3 lengths within multi-RNA containing MIDs from one million naive CD8+ T cells (50% RNA input) was plotted before sub-clustering (orange) and within the sub-clusters (green).

Sub-clustering corrects chimera sequences

More importantly, to our surprise, we found that, without performing sub-clustering, the number of unique consensus sequences (unique CDR3 sequences) was overestimated, especially in samples with one million cells (Figure 2.6). This is because chimera sequences were generated in the consensus building step for two scenarios. In one scenario, multiple true TCR sequences could be tagged with the same MID and quality score weighted consensus building will generate chimera sequences (Figure 2.1). In the second scenario, PCR or sequencing errors on MIDs group multiple singletons (MIDs that contain only one read) under the new MID. If sub-clustering is applied, then these singletons will be separated and discarded under the singleton category. However, without sub-clustering, these singletons will be forced to generate a chimera

sequence. Taking together, these chimera sequences cause overestimation of the total TCR diversity. The percentage of chimera sequences can be as high as 47% (Table S1).

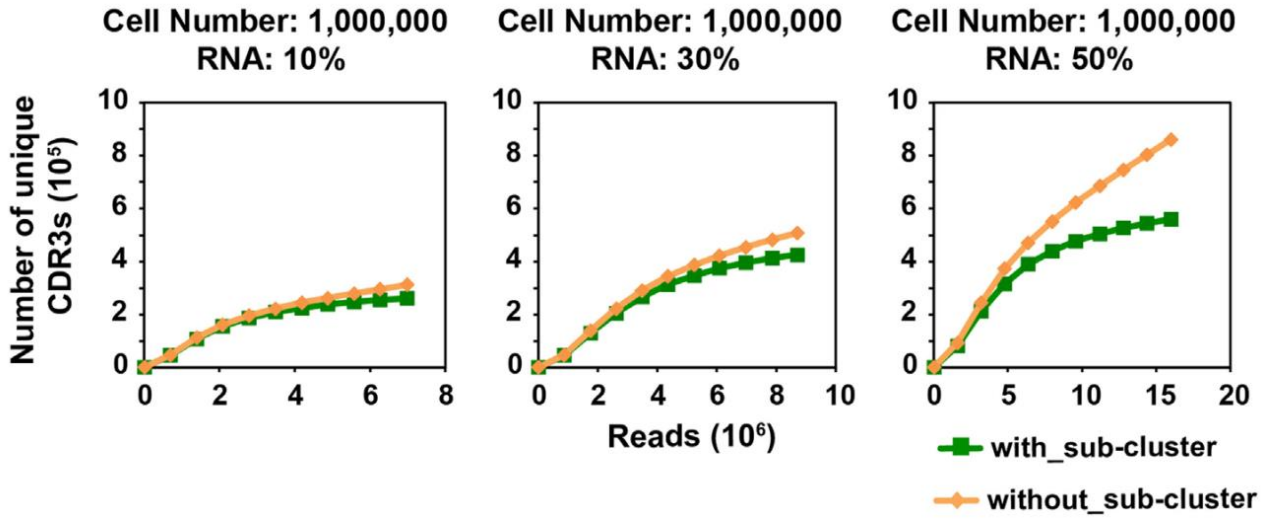


Figure 2.6: Rarefaction curve of unique complementarity-determining regions 3 (CDR3s) with or without sub-clustering. Number of unique CDR3s in three libraries made with three different RNA inputs from sorted one million naive CD8+ T cells.

MIDCIRS significantly reduces error rate

We examined the error rate with or without using MIDCIRS on antibody reperotire[43]. Because the diversity among hundreds of millions of antigen receptors lies in a short stretch of DNA about 60 nucleotides, often two distinct sequences are different by only a few nucleotides. In addition, somatic hypermuation, a process that further diversifies the antibody gene sequences, has a mutation rate that is comparable to the error rate of the next-generation sequencers. This makes estimating the total antigen receptor diversity and tracing the mutational evolution of antibody gene sequences difficult. Using MIDs can reduce the error rate by several orders magnitude and enable an accurate sequencing and diversity comparison. By comparing individual reads within a sub-group to the consensus read, we reached the similar error rate as previously reported for Illumina, which is about 0.5%[22,27].

To calculate the improved error rate using the MIDCIRS, we split the total reads into two groups, performed clustering separately, and compared the consensus of overlapping sub-groups from these two sub-samples. The resulted error rate was 130 fold smaller than the current error rate, which reached a quality score of Q45. In addition, while the raw error rate fluctuated between runs as demonstrated by the error rate from three runs (Figure 2.7, top panel), the improved error rate after using MIDs for these three runs almost did not change (Figure 2.6, bottom panel). This comparison can also be used to guide the cluster generation on the sequencer to maximize the sequence yield without comprising the sequence quality. Without MIDs, the diversity estimate is massively inflated with errors due to PCR and sequencing as demonstrated in one experiment where we obtained 1.3 million of reads for one library made from 10,000 cells. It generated 258,320 unique raw reads and, even after we took out unique sequences represented by only one read, there are still 148,680 unique sequences, which is impossible for a total of 10,000 cells. This also demonstrates the necessity of using MIDCIRS in immune repertoire sequencing.

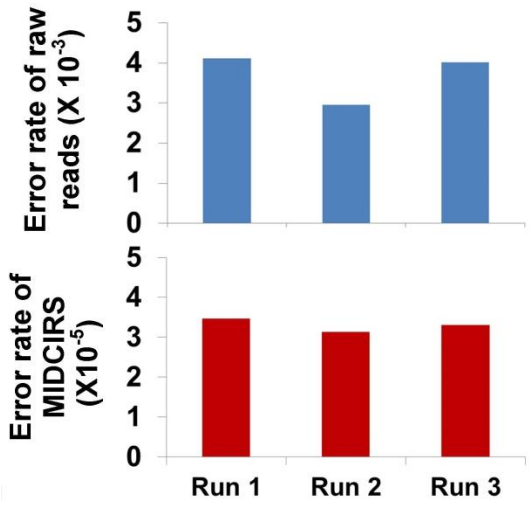


Figure 2.7: Comparison between raw error rate and improved error rate after using MIDCIRS.

2.2.2 MID Read-Distribution-Based Barcode Correction Improves Accuracy and Sensitivity of Counting TCR Transcripts

Besides correcting PCR and sequencing errors, MIDs have also been used for absolute quantification of RNA molecule copy number in single-cell studies to improve precision[44–47]. Here, we demonstrated how to use MIDCIRS TCR-seq to digitally count TCR transcripts. The absolute quantification of TCR transcripts is fundamental for accurate clonal size estimation. We noticed that PCR and sequencing errors also affected MIDs, as seen in single-cell RNA sequencing studies[41,48], leading to an inflated number of RNA molecules when libraries were sequenced exhaustively with respective to the total TCR transcripts in the sample (Figure 2.8). To correct MID errors, we first removed singleton reads, which cannot be confidently used in generating MID groups due to sequencing errors. Then, we adopted a similar approach applied in single-cell RNA-seq by fitting the distribution of reads under each MID subgroup into two negative binomial distributions (Figure 2.9)[48]. Erroneous MIDs generated due to PCR errors generally have distinctively lower read counts compared with true MIDs. These two negative binomial distributions distinctly separated true MIDs from erroneous MIDs. MIDs with low read counts were removed accordingly (see Materials and Methods). After MID correction, number of RNA molecules saturated across libraries (Figure 2.8).

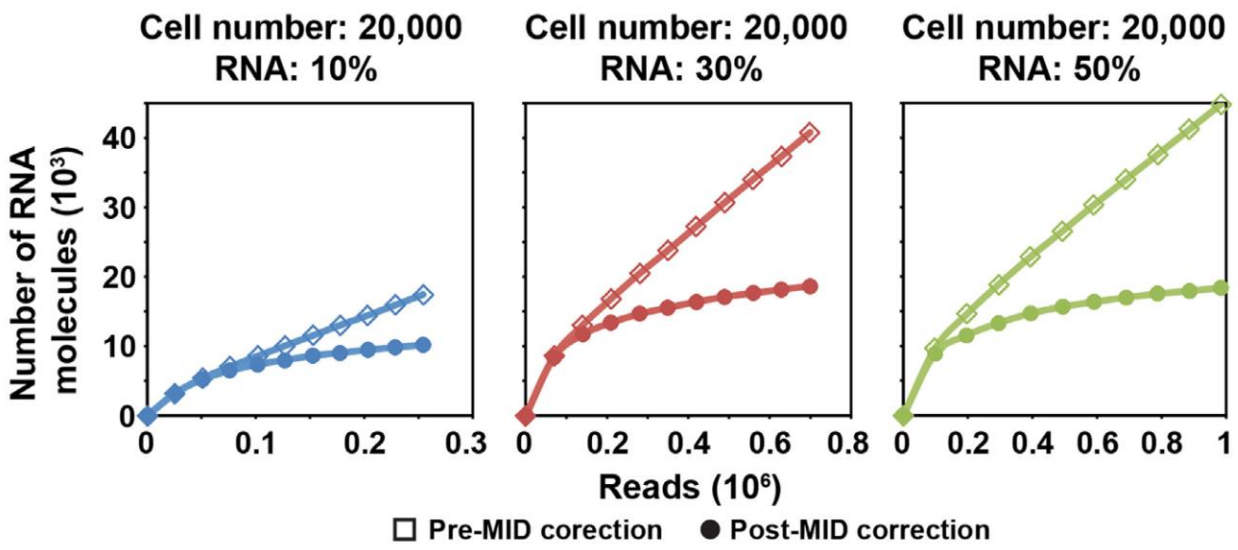


Figure 2.8: Rarefaction curve of detected TCR RNA molecules before and after error correction on molecular identifiers (MIDs) in 20,000 naive CD8+ T cells for three RNA input amounts

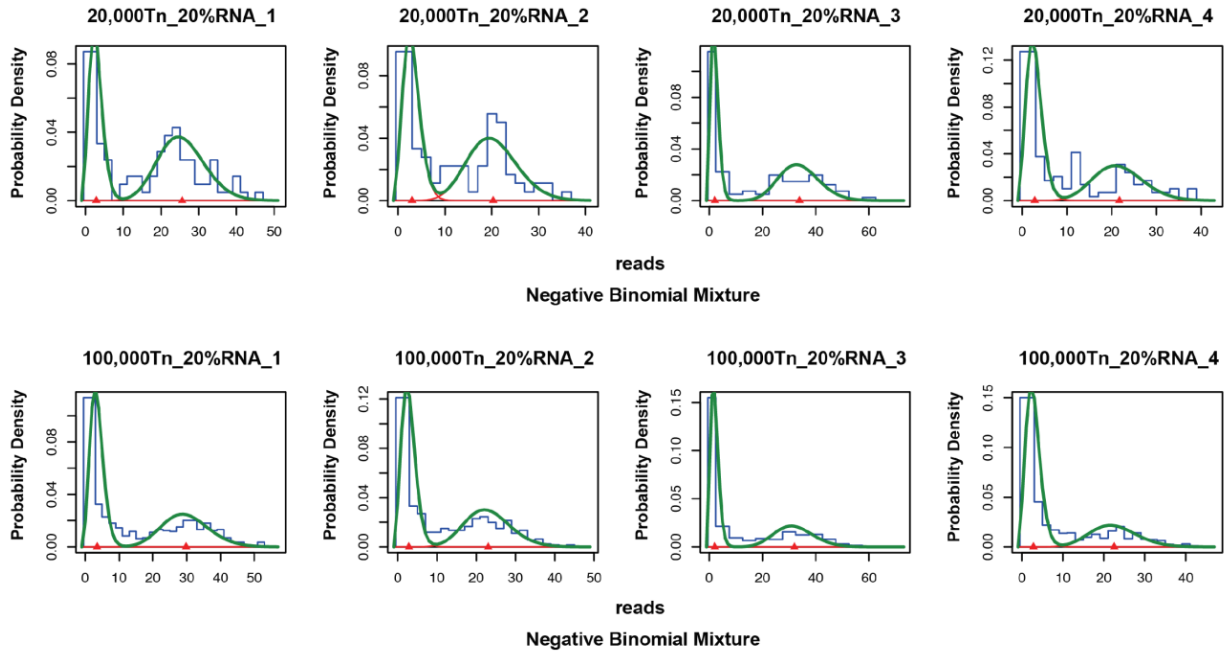


Figure 2.9: Distribution of reads under each MID sub-group within example TCR clones. The distribution of reads is modeled with a mixed negative binomial distribution with two components, reads from the 2nd distribution need to be removed in order to correct MID errors.

2.2.3 MIDCIRS yields high accuracy and coverage down to 1000 cells

We used sorted naive B cells with varying numbers (10^3–10^6) to test the dynamic range of MIDCIRS. The resulting diversity estimates, or different types of antibody sequences, display a strong correlation with cell numbers at 83% coverage (Fig. 2.10, slope). Previous studies have shown that about 80% of naive B cells express distinct heavy chain genes[49,50], thus our method achieves a comprehensive diversity coverage that is much higher than other MID-based antibody repertoire-sequencing techniques[26,27,36,37]. Thus, compared with previous IR-seq with MID method[26], MIDCIRS not only can improve the accuracy of diversity estimation but also can increase diversity coverage of CDR3.

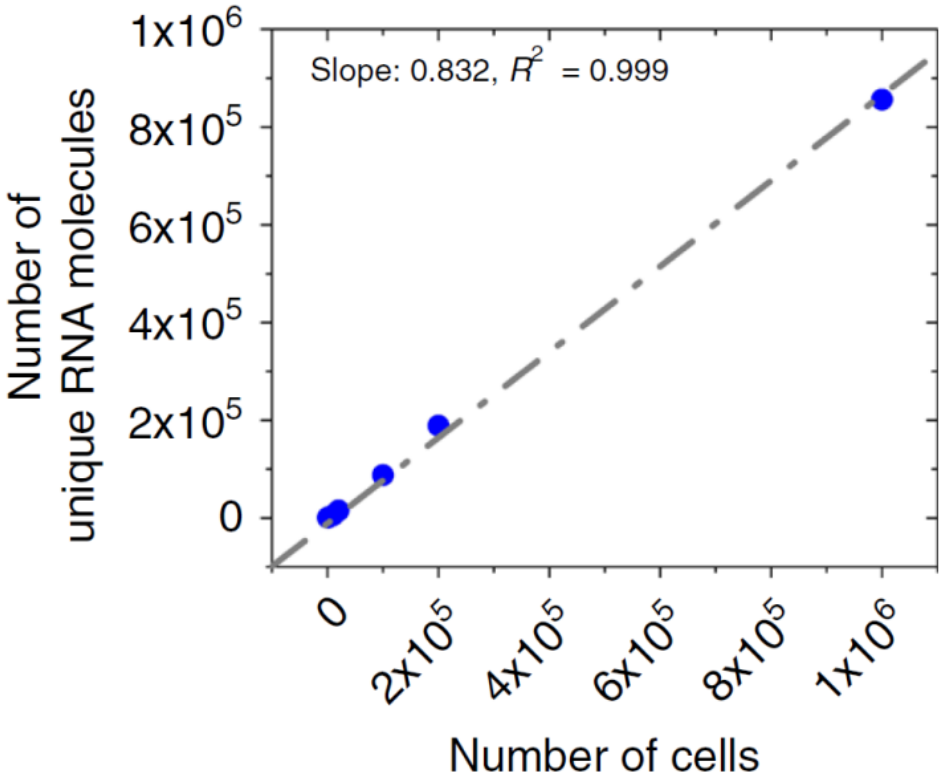


Figure 2.10: Correlation between number of cells and number of unique RNA molecules after using MIDCIRS. RNA from as few as 1000 to as many as 1,000,000 naive B cells was used as input material in generating the amplicon libraries. Slope indicates the estimated diversity coverage.

2.2.4 Theoretically and experimentally measured TCR RNA molecule copy number per cell

After MID sub-clustering and MID correction, theoretically MIDCIRS can capture all the RNA molecules within a sample. However, due to the experiment efficiency and removing erroneous MIDs, we are never going to cover all the RNA molecules in a bio-sample. One question is researchers may be interested is: what's the efficiency of MIDCIRS, i.e. the percentage of TCR RNAs covered within the total TCR RNA molecules. In order to estimate the efficiency, we adopted two ways to calculate the average TCR RNA copy number per CD8+ naive T cell (m), which is an unknown parameter yet: one is to directly fit the number of RNA molecules we observed versus the RNA we inputted and the slope will be 'm'; the other way is to develop a statistical model which predict 'm' based on the number of unique TCR molecules we observed.

Figure 2.11 shows the result of the first way of modeling number of observed RNA versus number of inputted RNA ('number of cells' times 'percentage of RNA input'), we fitted a line across all the data points we measured. The slope shows the average RNA copy number per cell is ~3.

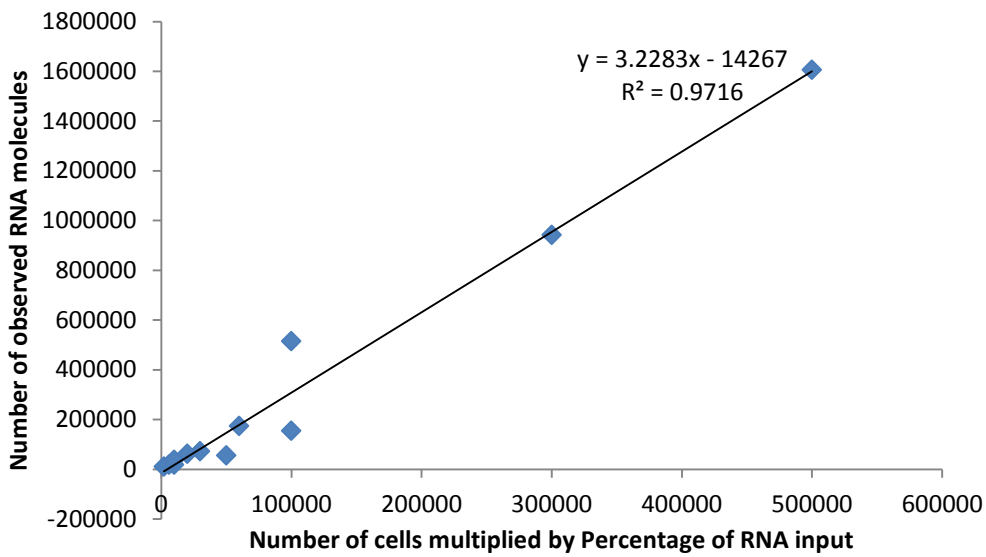


Figure 2.11: Observed RNA versus inputted RNA. X axis shows the experimental cell numbers and RNA input percentage, Y-axis shows the number of MIDCIRS' observed RNA molecules. The slope shows the estimated RNA copy number per cell.

The second way is to develop a statistical model based on number of unique RNA observed versus the inputted RNA while treat the number of observed RNA molecues as an unknown parameter. For N observed RNA molecules, there are K different RNA clones. The RNA molecule copy number of each clone is m_i ($i \in (1, K)$), whose sum equals N. After fitting the data, m_i follows a power law distribution[29,51] (Figure 2.12):

$$m_i = m \times x_i \tag{3}$$

$$f(x_i) = (\alpha - 1)x_i^{-\alpha}, (\alpha > 1) \tag{4}$$

(m is the RNA molecule copy number per cell, which is a constant across all T cells. x_i represents the cell numbers of each clone, which follows a power law distribution, and the parameter α was fitted with an algorithm combining maximum-likelihood fitting and goodness-of-fit test based on Kolmogorov-Smirnov statistic[52]. 'fit_power_law' function in R package igraph was applied[53].

Specifically, we fitted the RNA molecule distribution (Figure 2.10) with equation (5):

$$f(m_i) = \left(\frac{\alpha - 1}{m_{min}}\right)\left(\frac{m_i}{m_{min}}\right)^{-\alpha}, (\alpha > 1) \tag{5}$$

Since 'm' is a constant, the alpha in equation (4) and (5) should be equal. We fitted across all libraries on log-log scale, and the average slope was taken as α in the above model.

When we sample n RNA molecules from this population, the expected detected diversity, E(D), can be calculated as the following:

$$E(D|m, x_i) = K - \frac{\sum_{i=1}^{K}\binom{N - m \times x_i}{n}}{\binom{N}{n}}, x_i = (x_1, x_2, \dots, x_K) \tag{6}$$

And x_i can be sampled from the fitted power law distribution.

Then, the percentage of the RNA diversity coverage, P(D), can be estimated as:

$$P(D|m, x_i) = \frac{E(D|m, x_i)}{K} \tag{7}$$

We then used equation (8) to get estimated m:

$$\min_m \sum_i (P(D_i|m, x_i) - D_{obs})^2, m \in \{1, 2, \dots\} \tag{8}$$

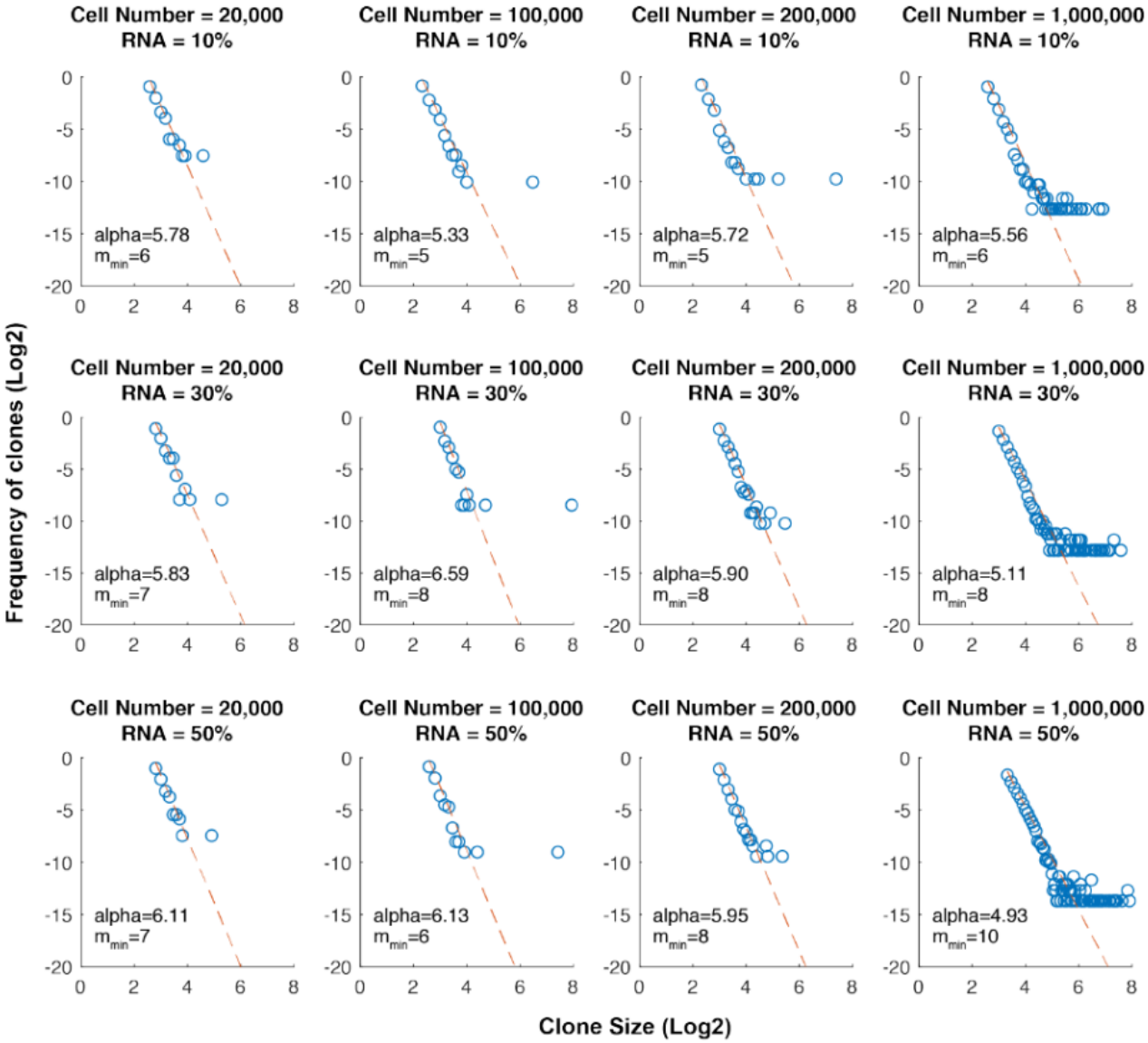


Figure 2.12: TCR clone size distribution of naive CD8+ T cells. Red dashed line is the fitted power law distribution.

Using the model above, we predict 'm', which is the average TCR RNA per cell, to be equal to 3, and the results are shown in Figure 2.13A. The RNA copy numbers estimated by both ways agree with each other, which demonstrate the accuracy of digit counting RNAs with MIDCIRS. The 2nd way of modelling treated MID numbers as an unknown parameter and we used the 1st way to validated out prediction after sequencing much more reads from the library.

This result shows MIDCIRS is able to capture almost all the MID labeled RNA molecules. We experimentally measured the average RNA copy per cell digital PCR, digital PCR measured TCR RNA copy per cell is 8~12(Figure 2.13B), this differ is mainly during the Reverse Transcription step (RT) and PCR steps. In summary, the results all together represent MIDCIRS efficiency is ~30%.

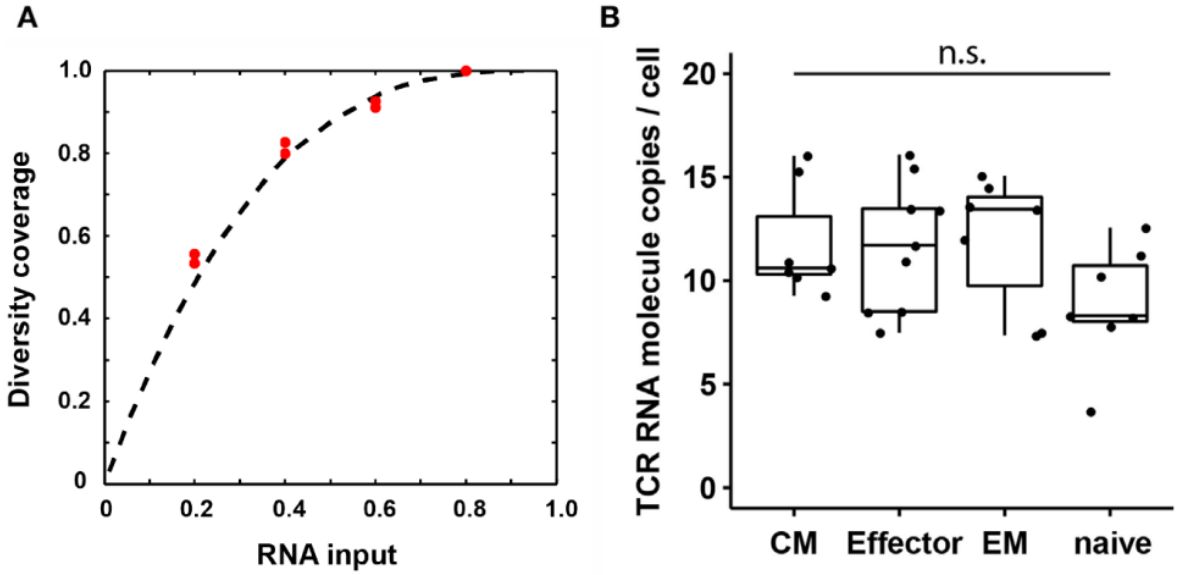


Figure 2.13: Modelling and digital measuring TCR RNA copy per cell. A). Curve fitting of diversity coverage as a function of different RNA input with 3 as predicted TCR RNA molecule copy number per cell. B). Validate TCR RNA molecule copy number with digital PCR.

2.2.5 Naive TCRs serve as guideline for sequencing depth

While designing a sequencing experiment, it's hard to determine a suitable sequencing depth that is able to cover all the diversity (unique RNA) or RNA molecules. One way researchers normally performing is to sequence multiple times, with each time adding more sequencing reads to check whether all the unique RNAs have been covered and pool all the reads together as the final sequencing pool, which is very inefficient. Our experiment on naive TCRs provides a good bottom line for researchers, because naive cells are generally more diverse than other cell populations since naive T cells have little clonal expansion. Certain sequencing depth can cover all the diversity within naive TCR repertoire will definitely be able to cover diversity in other cell population. We found that a shallower sequencing depth is required to saturate unique CDR3s than RNA molecules. In addition, the amount of diversity covered increased with increasing RNA input. Thus, to exhaustively measure the TCR repertoire diversity, with 30–50% of RNA input, a sequencing depth equivalent to 10 times the cell number covers most of the CDR3 diversity, while a sequencing depth equivalent to about 100 times the relative RNA input (defined as cell number multiplied by percentage of RNA input) is required to saturate the RNA molecules. For example, 30% RNA of 20,000 cells is equivalent to 6,000 RNA input. Then, it takes about 600,000 reads to saturate the RNA molecules but only 200,000 reads to saturate the unique CDR3s.

2.3 MATERIALS AND METHODS

Naive CD8+ T Cell Sorting

Human leukocyte reduction system chambers were obtained from de-identified donors at We Are Blood (Austin, TX, USA) with strict adherence to guidelines from the Institutional Review Board of the University of Texas at Austin. CD8+ T cell enrichment was done following the protocol described previously[54] using RosetteSep CD8+ T Cell Enrichment Cocktail (STEMCELL) together with Ficoll-Paque (GE Healthcare). Then, RBCs were lysed using ACK Lysing Buffer (Lonza). After washing in phosphate-buffered saline with fetal bovine serum, the cell mixture was passed through a cell strainer (Corning) and ready for use. Naive CD8+ T cells were FACS-sorted into RLT Plus buffer (Qiagen) supplemented with 1% β-mercaptoethanol (Sigma) based on the phenotype of CD8+CD4-CCR7+CD45RA+ using BD FACSAria II cell sorter.

Bulk TCR Library Generation and Sequencing

Total RNA was purified using All Prep DNA/RNA kit (Qiagen) following the manufacturer's protocol. Library preparation and QC were similar to protocols described previously[43] using TCR primers[55]. Reads of the same library from all runs were combined and analyzed.

dPCR of TCR

Total RNA purified from sorted CD8+ T cells and cultured CMV-specific CD8+ T cell lines were reverse transcribed with polyT primers[55] using Superscript III in 20 µl reaction following the manufacturer's protocol. 2 µl of cDNA was subsequently used on QuantStudio 3D dPCR system following manufacturer's protocol.

Preliminary Read Processing

First, only reads that have exact TCR constant sequences were kept for further analysis. These reads were then cut to 150 nt starting from constant region to eliminate high error-prone region at the end of reads. These preprocessed reads were split into MID groups according to 12-nt barcodes.

MID Sub-Cluster Generating and Filtering

For each MID group, a quality threshold clustering was used to group reads derived from a common ancestor RNA molecule and separate reads derived from distinct RNAs. Briefly, a Levenshtein distance of 15% of the read length was used as the threshold. For each subgroup, a consensus sequence was built based on the average nucleotide at each position, weighted by the quality score. In the case that there were only two reads in an MID subgroup, we only considered them useful reads if both were identical. Each MID subgroup is equivalent to an RNA molecule. Next, we merged all of the identical consensus to form unique consensus sequences. Further, we applied filtering of unique consensus sequences after sub-cluster generation by (a) removing non-functional TCR sequences and (b) removing sequences with lower MID counts that are one Levenshtein distance away from the other. Then, for each unique consensus sequence, we removed MID sub-clusters if their reads are less than 20% of maximum read count based on the fitting of two negative binomial distribution (Figure 2.8). Scripts for this section can be downloaded at https://github.com/utjianglab/MIDCIRS.

Statistical Analysis

Mann–Whitney U test was used to calculate the significance of copy number difference between pairs in naive, effector, effector memory, and central memory CD8+ T cells and p values was adjusted with Benjamini–Hochberg procedure. Adjusted p-value that was less than 0.05 was considered significant.

2.4 DISCUSSION

We applied the MIDCIRS in T cells to demonstrate: 1) the necessity of MID sub-clustering to improve accuracy of repertoire diversity estimation; 2) erroneous MID correction eliminates overestimate of TCR RNAs; 3) the accuracy of counting TCR RNA molecules via MID read-distribution based barcode correction and 4) wide dynamic range of MIDCIRS.

Previous MID-based IR-seq methods, such as MIGEC, build TCR consensus sequences by grouping MIDs[26,56]. However, the number of target molecules could vary significantly with different sample inputs, which could be challenging for choosing the appropriate MID length to ensure that each target RNA molecule is uniquely tagged by MID. Longer MIDs are likely to decrease the reverse transcription efficiency[40,41]. Thus, the MIDCIRS method offers a flexible strategy for MID-barcoded IR-seq. In addition, MIGEC triages MIDs with high diversity as ambiguous. We compared TCR diversity discovered using MIDCIRS with that of MIGEC, using MID with at least two reads as the threshold for both approaches and found that MIGEC led to an underestimated TCR diversity (Figure S1, $p < 0.001$, effect size $r = 0.62$). We demonstrated that using MID-based sub-clustering approach, MIDCIRS could identify new diversities, prevent chimera sequences from being built, and digitally count RNA molecules. This corrected diversity is highly consistent with cell input numbers.

While MIDs are useful to correct for sequencing errors and PCR errors that occur on TCR sequences, such errors are also likely to show up on MID sequences. Although these errors do not affect TCR diversity estimation, they lead to an overestimation of transcript copies, thus misestimating TCR clone size. We corrected MID errors based on the distribution of MID read counts under MID subgroups. With MID correction, we were able to accurately count TCR RNA

molecule copy number, estimate MIDCIRS detection limit as well as detect T cell clonal expansion.

Based on the TCR diversity estimation and its dependency on RNA input, we built a probability model to estimate TCR RNA molecule copies, which resulted in three copies per cell. We would like to point out that this does not mean that on average there are three copies of TCR RNA in a T cell. Because of the efficiency of RNA purification and reverse transcription, we expect our observed RNA molecule per cell to be lower than the true value. In Fact, dPCR results showed an average of 10 copies of TCR RNA molecule per cell, suggesting the efficiency of MIDCIRS in TCR RNA molecule digital counting is about 30%, which is consistent with previous finding that nanoliter reaction volume significantly improved PCR efficiency. Thus, quantifying TCR RNA molecule per cell enables us to estimate the extent of T cell clonal expansion that was not possible until now.

Chapter 3: MIDCIRS application in biomedical study helps understanding antigen driven immune response

Having demonstrated the high coverage, accuracy and dynamic range of MIDCIRS in immune repertoire sequencing in Chapter 2, we applied MIDCIRS on real patient samples to study antigen driven immune responses of immune repertoire.

Chapter 3 contains two separated studies focusing on MIDCIRS applications in TCR repertoire and antibody repertoire:

We applied MIDCIRS on HIV patients, analyzed their TCR repertoire and found evidence for 1) intact antigen-driven clonal expansion of Tfh cells and 2) selective utilization of specific complementarity-determining region 3 (CDR3) motifs during chronic HIV infection[57].

We used MIDCIRS to study malaria infected children samples, analyzed their antibody repertoire and found unexpected mutation capability of immune system in infants[43].

Both of the results demonstrated MIDCIRS potential usage in research and clinical applications.

3.1 Application of MIDCIRS on TCR repertoire: The receptor repertoire and functional profile of follicular T cells in HIV-infected lymph nodes*

3.1.1 Introduction

Follicular helper T cells (Tfh) provide key signals necessary for B cell recruitment and selection to generate protective antibody responses[15,58]. During untreated chronic HIV infection, Tfh cells become highly expanded in the lymph nodes (LNs)[59,60]. Despite this, HIV+ patients generate diminished protective antibody responses against immune challenges. For example, HIV-infected individuals produce lower titers of antibodies and less durable responses to seasonal influenza vaccines[61]. The prevailing model suggests that Tfh cells from HIV patients are ineffective at providing B cell help based on in vitro assays that showed less robust antibody production by B cells co-cultured with Tfh cells from HIV+ patients[62–64]. A proposed mechanism for this involves up-regulation of programmed cell death–ligand 1 (PD-L1) by B cells, which interacts with programmed cell death–1 (PD-1) on Tfh cells to inhibit T cell receptor (TCR)–dependent activation of Tfh cells[62]. However, the extent to which Tfh cells express impaired antigen responsiveness in vivo remains unclear. Because Tfh cells need to appropriately sense antigen signals to discriminate between B cells, defective response to antigen not only impairs provision of T cell help to individual B cells but also imperils the process of B cell selection on a global level.

---

* Wendel, B. S.†, Del Alcazar D†, **He C**†. *et al.* The receptor repertoire and functional profile of follicular T cells in HIV-infected lymph nodes. *Sci. Immunol.* (2018). doi:10.1126/sciimmunol.aan8884. B.S.W. performed TCR sequencing and data analysis, D.A.A. performed CyTOF staining and analysis. , C.H. analyzed TCR sequencing data, B.A. performed in vitro peptide stimulation, P.D.R. and G.R.T established the infrastructure for HIV+ patient recruitment and provided HIV+ LN samples and the associated clinical information. S.M.H. assisted with TCR sequencing library preparation. K.-Y.M. adapted the TCR repertoire sequencing protocol. M.B. contributed to the lymph nodes from healthy individuals. L.F.S. and N.J. designed the study. L.F.S., N.J., and B.S.W. wrote and edited the paper.

†: These authors contributed equally.

Here, we analyzed the TCR repertoire composition of primary Tfh cells isolated directly from LNs from untreated HIV+ individuals. We used the presence or absence of antigen-dependent TCR signatures to address the responsiveness of Tfh cells to antigen engagement. Our TCR repertoire data revealed clonal expansion and convergent selection for Gag-reactive TCRs in Tfh cells in the germinal centers (GCs) of HIV-infected LNs, indicating that Tfh cells remain capable of responding to HIV antigens during chronic HIV infection.

3.1.2 Results

HIV-infected LNs contain clonally expanded GC Tfh cells

LNs from untreated HIV+ patients contain a high frequency of Tfh cells, but the mechanism that drives expansion of Tfh cells remains unclear. The enrichment of HIV antigens[65,66] and the highly pro-inflammatory milieu[67,68] in the LNs could lead to antigen-driven and/or bystander T cell expansion. To address whether proliferation of Tfh cells is antigen-dependent, we tested whether HIV induces selective proliferation of certain T cell clones. We focused on GC Tfh cells because the frequency of these cells becomes greatly increased during chronic HIV infection[59,60]. To identify GC Tfh cells, we selected memory CD4+ T cells that express Tfh cell markers CXCR5 and PD-1. CD57 is a glycan carbohydrate epitope expressed by Tfh cells in the GC, and we used this marker to further demarcate the GC subset [69–72]. Naive CD4+ T cells were identified by CD45RO−CXCR5−CD57−CCR7+ expression, and memory CD4+ T cells were identified by CD45RO+CXCR5− staining and excluded for PD-1 and inducible T cell co-stimulator (ICOS)–positive cells (Figure 3.1A). We sorted 1464 to 15,000 naive, memory, and GC Tfh cells from freshly thawed LN samples and analyzed the TCR sequences of these subsets using MIDCIRS to increase the accuracy of repertoire sequencing [43,55]. Because the variability of TCR sequences is encoded in the complementarity-determining region 3 (CDR3) region, we used the number of transcripts detected for a particular CDR3 sequence to define TCR clone size. On average, 11,839 TCR transcripts were detected for each sample (table 3.1). Unique TCR frequencies range from 1 in 37,129 (0.003%) for the rarest clones to 250 in 2498 (~10%) for the most expanded clone. To compare the degree of relative clonal expansion, we categorized TCR frequency into six groups, ranging from rare (<0.1%) to >2%, according to the clone size relative to the total TCR transcripts detected in that sample. As expected, the TCR repertoire of naive CD4+ T cells was composed mostly of rare clones. In contrast, the TCR repertoire of GC Tfh

cells had a much higher fraction of TCRs occupied by abundant clones (>0.1%) compared with naive and memory CD4+ T cells (Figure 3.1B and figure 3.2). The degree of TCR clonal expansion was quantified by normalized Shannon entropy (NSE)[73,74]. Consistent with the hypothesis that the increase in GC Tfh cell frequency is due to selective proliferation of certain T cell clones, GC Tfh cells had a lower NSE score compared with naive and memory cells (Figure 3.1C). Together, our data demonstrated a notable expansion of clone size in GC Tfh cell populations.

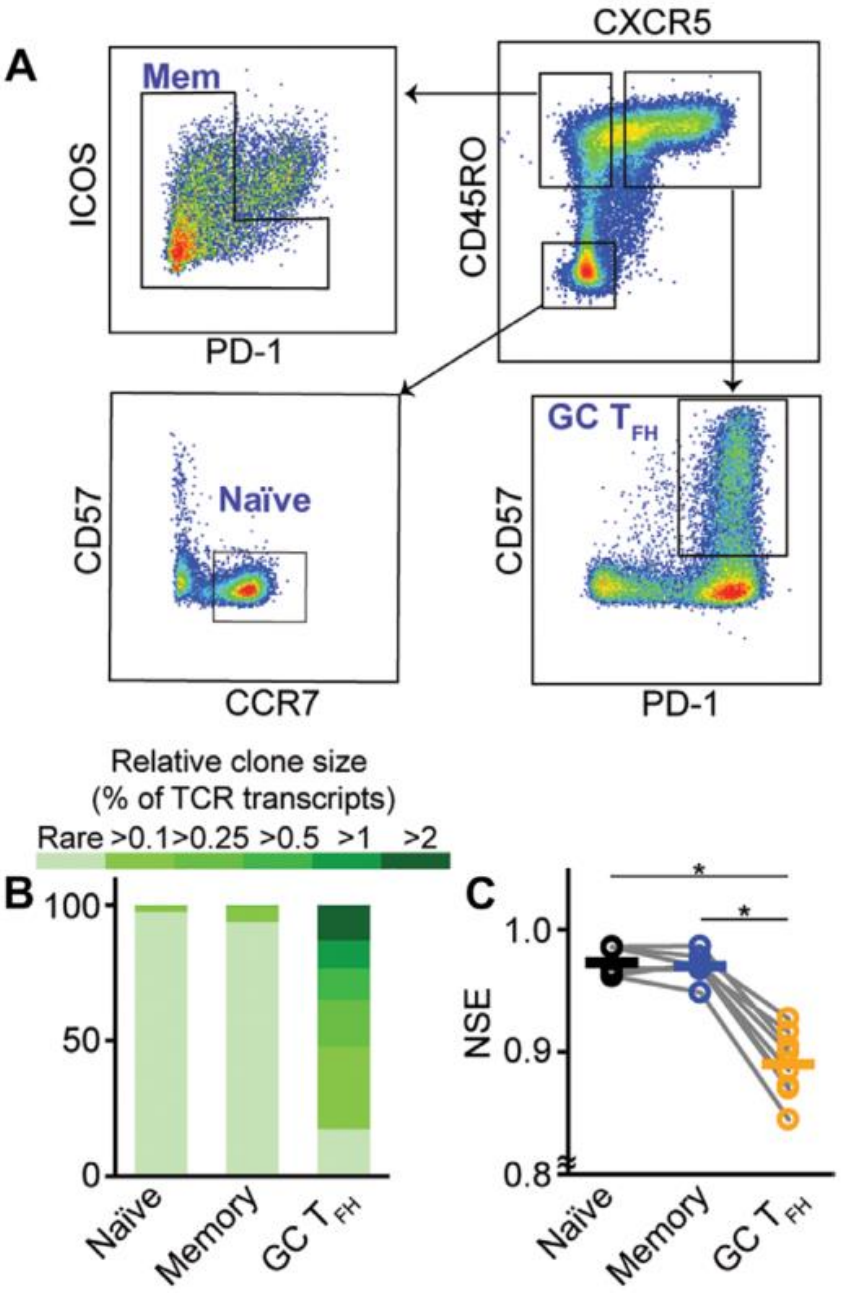


Figure 3.1: GC Tfh cells become clonally expanded. (A) Representative plots showing sorting strategy to identify naive, memory, and GC Tfh cells. (B) Breakdown of the proportion of the TCR repertoire represented by clones of different sizes for sorted naive, memory, and GC Tfh cells from HIV+ LNs. TCR clone size was normalized by the total number of TCR transcripts on nucleotide sequences. (C) NSE of the TCR repertoire of sorted naive, memory, and GC Tfh cells. Gray lines link the same patient. Bars indicate means. *P < 0.05 by two-tailed Wilcoxon signed-rank test (n = 8 HIV-infected LNs).

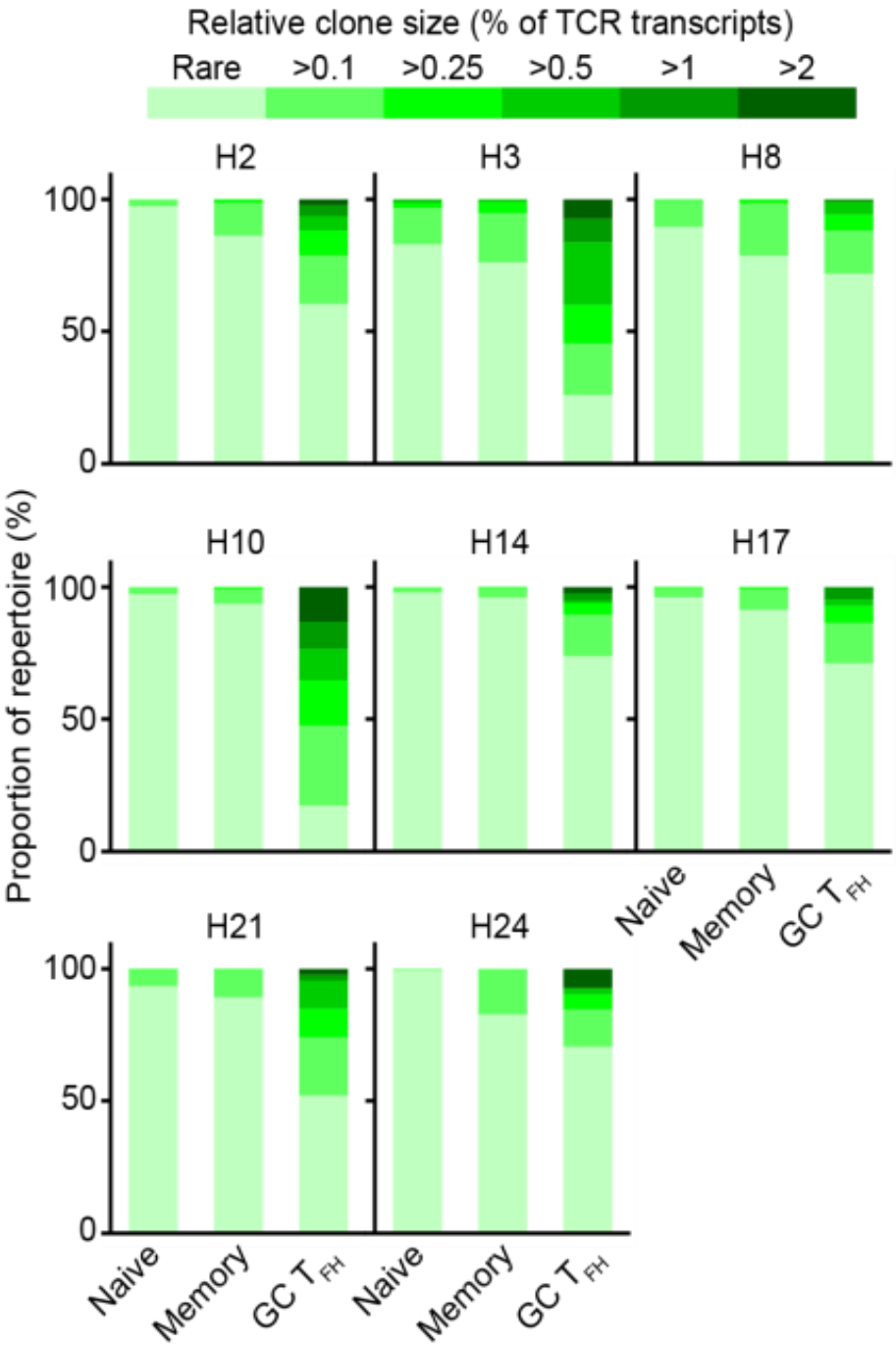


Figure 3.2: GC Tfh cells are clonally expanded. Breakdown of the proportion of the TCR repertoire represented by clones of different sizes for sorted naive, memory, and GC Tfh cells from HIV+ LNs for each individual. TCR clone size was normalized by the total number of TCR transcripts on nucleotide (nt) sequences.

TCRs from GC Tfh cells exhibit signatures of antigen-driven clonal convergence

Next, to test whether clonal expansion in GC Tfh cells from HIV-infected LNs was antigen-driven, we analyzed the TCR sequences for evidence of convergence to the same amino acid sequence from distinct nucleotide sequences. Unlike B cells, which can undergo somatic hypermutation, the TCR sequence of a naive T cell is determined during maturation in the thymus and remains fixed throughout the life spans of the T cell and its progeny. Thus, with the exception of clones that express two TCRa or TCRb sequences, distinct TCR nucleotide sequences necessarily arise from distinct naive T cells. However, multiple nucleotide sequences of different TCRs may encode the same amino acid sequence. These degenerate TCR sequences are typically rare, and the presence of these sequences suggests antigen selection pressure that favors certain TCR motifs that recognize particular antigen(s). Thus, having highly abundant CDR3 amino acid sequences that are encoded by multiple distinct nucleotide sequences indicates preferential expansion of T cells with that specificity[73]. On the other hand, we would not expect multiple nucleotide sequences to converge on the amino acid level in the absence of strong antigen-driven selection. Following this logic, we translated the TCR nucleotide sequences into amino acid sequences and tallied the number of different nucleotide sequences that encode each CDR3 amino acid sequence. These CDR3 amino acid sequences can be broken into four quadrants—Q1, Q2, Q3, and Q4—based on the level of degeneracy and frequency in the repertoire (Figure 3.3A and Figure 3.4). Q1 contains highly expanded amino acid CDR3 sequences that are encoded by two or more nucleotide sequences (Figure 3.3B). These degenerate, abundant clones likely arose from strong antigen-driven selection and proliferation. Q2 contains low-frequency amino acid CDR3 sequences that are also encoded by two or more nucleotide sequences. Degenerate clones can stochastically arise in the repertoire, but these are typically rare as reflected by the low frequency of non-clonally expanded sequences in Q2. Q3

contains amino acid CDR3 sequences that show neither clonal expansion nor amino acid convergence and make up most of the repertoire. Q4 contains expanded amino acid CDR3 sequences derived from a single nucleotide sequence and are therefore nondegenerate. This TCR degeneracy analysis revealed a significantly higher degree of antigen-driven clonal convergence in GC Tfh cells compared with naive and memory T cells (Figure 3.3C). Together with the NSE decrease in GC Tfh cells, these data provide further evidence that antigen-driven clonal expansion is preserved in GC Tfh cells.

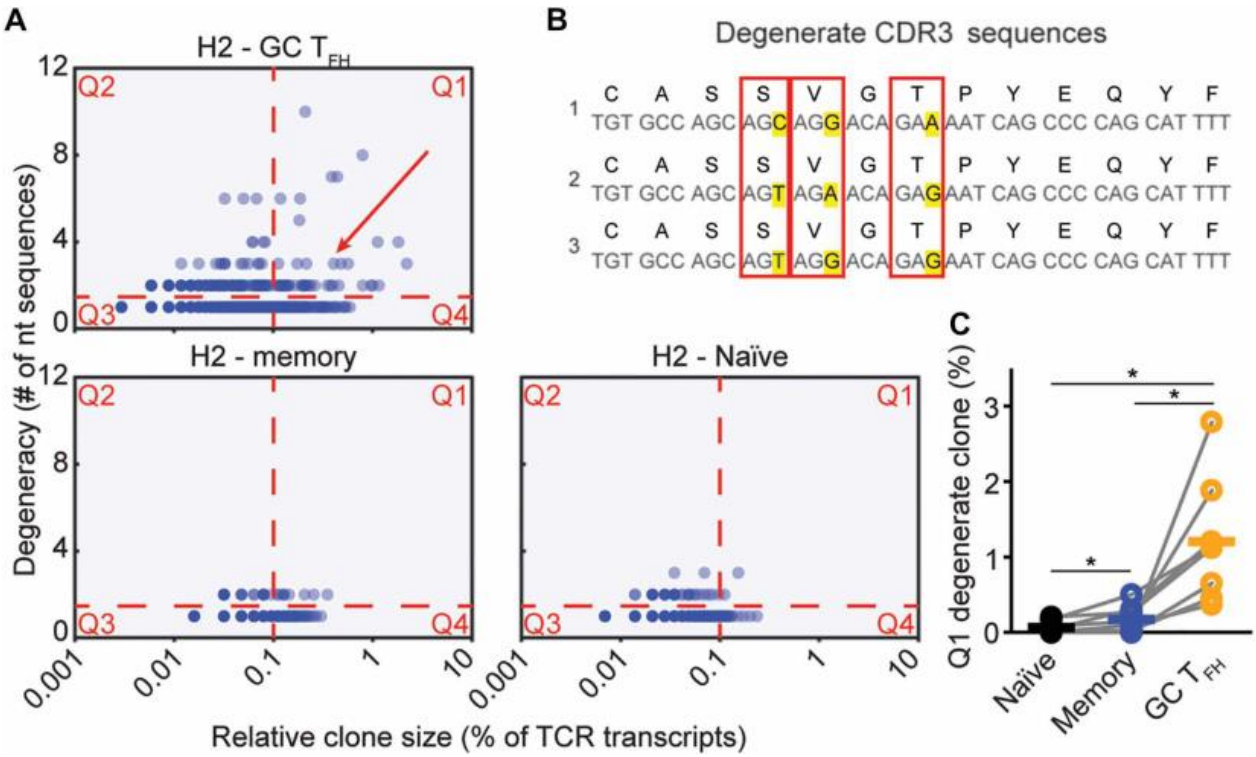


Figure 3.3: Antigen-driven clonal selection signature in GC Tfh cells of HIV-infected LNs. (A) Representative degeneracy plot from sample H2. Coding degeneracy level [number of unique TCR nucleotide (nt) sequences encoding a common CDR3 amino acid sequence] of each CDR3 amino acid sequence is plotted against their frequency (measured as percentage of total TCR transcripts) in naive, memory, and GC Tfh cells. Each dot is a unique CDR3 amino acid sequence. Red dashed lines indicate cutoffs for degenerate (two or more nucleotide sequences coding for the same amino acid sequence; horizontal) and expanded (0.1% or more of TCR transcripts; vertical) clones. Red arrow points to example degenerate clone in (B). (B) Example of CDR3 amino acid degeneracy. Amino acid (top row) and nucleotide (bottom row) sequences for three distinct nucleotide sequences (0.41% of total TCR transcripts) that code for the same amino acid sequence as indicated by arrow in (A): Y = 3 and X = 0.41%. Red boxes and highlights indicate redundant codons. (C) Comparison of Q1 degenerate-abundant clone percentage in naive, memory, and GC Tfh cells. Gray lines link the same patient. Bars indicate means. *P < 0.05 by two-tailed Wilcoxon signed-rank test (n = 8 HIV-infected LNs).

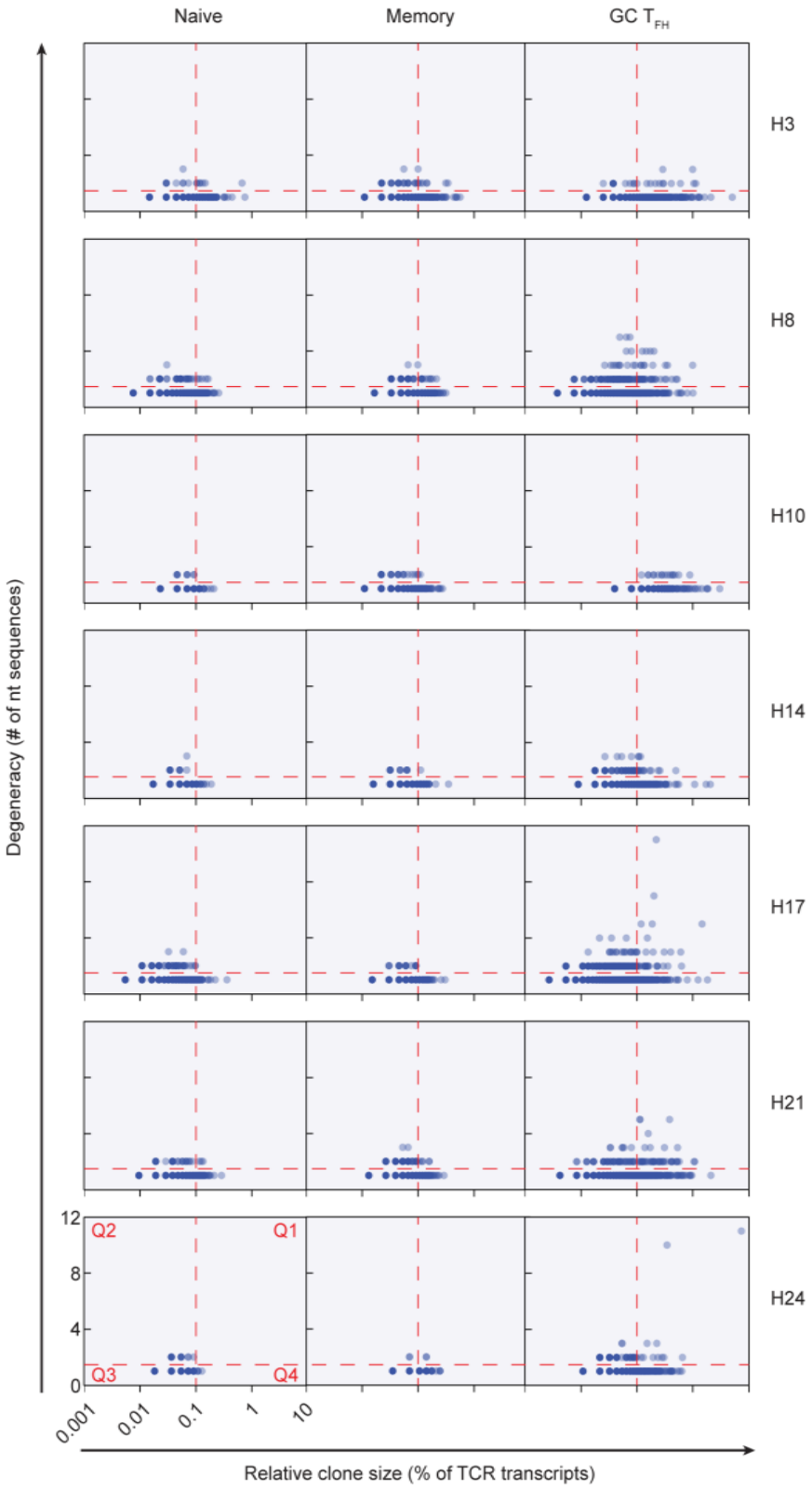


Figure 3.4: Antigen-driven clonal selection signature in GC Tfh cells of HIV-infected LNs. Coding degeneracy level (number of unique TCR nucleotide (nt) sequences encoding a common

Figure 3.4 cont.
CDR3 amino acid (aa) sequence) of each CDR3 aa sequence is plotted against their frequency (measured as % of total TCR transcript) in naive, memory, and GC Tfh cells. Each dot is a unique CDR3 aa sequence. Red dashed lines indicate cutoffs for degenerate (2 or more nt sequences coding for the same aa sequence, horizontal) and expanded (0.1% or more of TCR transcripts, vertical) clones. Each panel is broken into 4 quadrants: Q1: degenerate-abundant clones; Q2: degenerate-rare clones; Q3: nondegenerate-rare clones; Q4: nondegenerate-abundant clones.

HIV promotes selective expansion of HIV-reactive Tfh cells

To determine whether clonally expanded and/or convergently selected TCRs include HIV-specific sequences, about 2 to 3 million thawed LN cells were cultured with an HIV-1 consensus B Gag peptide pool for 3 to 4 weeks and then re-stimulated with the same peptide pool for 4 hours to identify antigen-specific T cells by CD40L and CD69 up-regulation (Figure 3.5). LN cells were also stimulated with an overlapping set of hemagglutinin (HA) peptides from influenza virus (A/California/7/2009) as a non-HIV control. TCRs from CD40L+CD69+ Gag- or HA-reactive T cells were used to generate a reference TCR panel (Table S2). These antigen-specific TCR sequences (Table S4) were mapped onto our bulk T cell sequencing data from freshly thawed LN cells to determine which sequences were Gag- or HA-specific. Common sequences shared between naive, memory, or GC Tfh cells were shown as connecting lines on circos plots (Figure 3.6A).

We found several Gag-specific TCR sequences in the GC Tfh (zero to seven clones) population. Although we did not have enough data points to reach significance, the overlap between Gag-specific TCR sequences was minimal in memory T cells (zero or one clone), and no Gag-specific sequences were found in the naive T cell population (Figure 3.6B). A similar trend of enrichment of antigen-specific clones in the GC Tfh cells was also observed for HA-specific TCR sequences (Figure 3.7). This is unsurprising because these individuals have likely been exposed to influenza infection and/or vaccinated against HA in the past. However, analysis

of combined TCR sequencing data from all individuals showed that these Gag-specific GC Tfh cells, but not the HA-specific clones, were highly expanded compared with the bulk GC Tfh cells of unknown specificity (Figure 3.6C). Translating these antigen-specific TCR sequences into amino acid sequences showed that the Gag-specific TCR sequences within the GC Tfh population, but not the HA-specific sequences, have a significantly higher degree of coding degeneracy (Figure 3.6D). Thus, the Gag-specific GC Tfh cells were preferentially expanded and degenerate. Collectively, these data indicate that Gag-specific Tfh cells respond to antigen stimulation and become selectively expanded in the LNs.

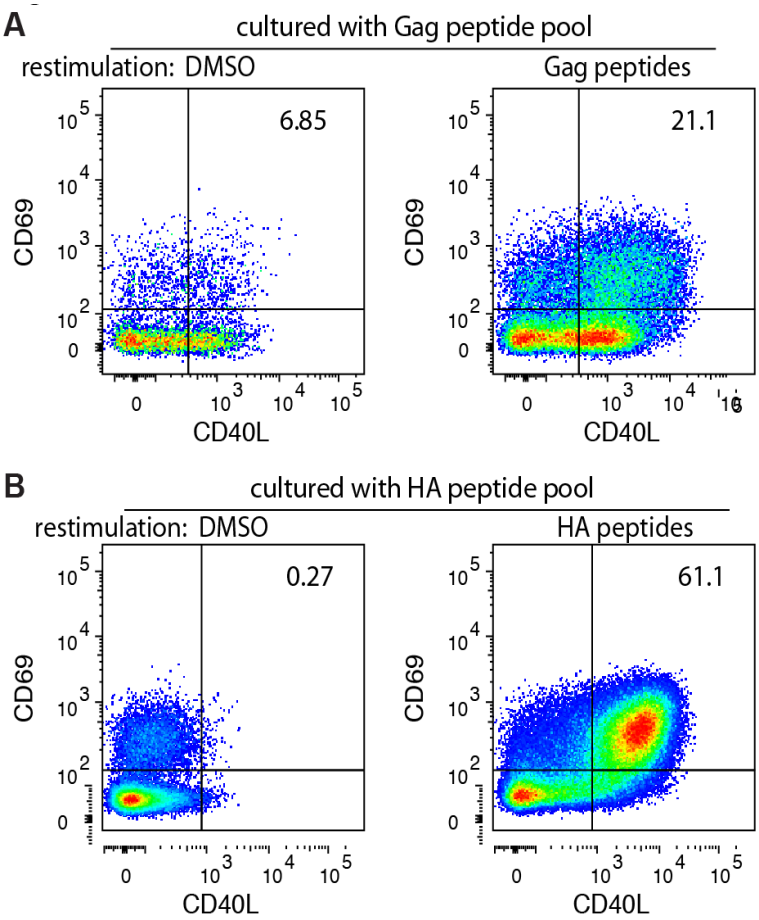


Figure 3.5: Identification of Gag- or HA-reactive T cells in cultured cells. LN cells were cultured with Gag or HA peptide pools for 3-4 weeks, then re-stimulated with peptides for 4

Figure 3.5 cont.
hours. (A-B) T cells specific to Gag or HA were identified by positive CD40L and CD69 staining. Representative plots showing antigen-specific T cells that responded to Gag peptides (A) or HA peptides (B).

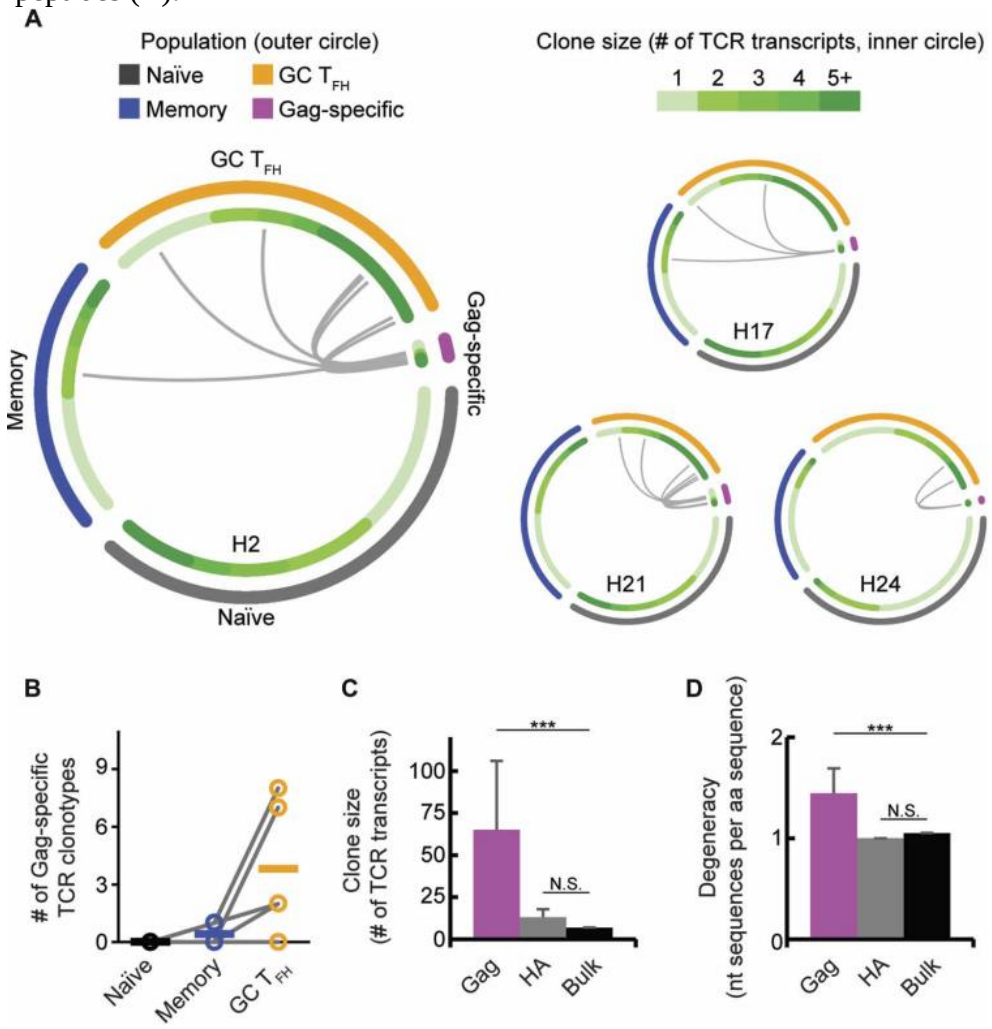


Figure 3.6: GC Tfh cells exhibit HIV antigen–driven clonal expansion and selection. (A) Gag-specific TCR clones overlap with HIV+ LN CD4+ T cell populations. Each thin slice of the arc represents a unique TCR sequence, ordered by the clone size (darker green for larger clones, inner circle). Gray curves indicate Gag-specific TCR nucleotide sequences found in naive (gray, outer circle), memory (blue, outer circle), and GC Tfh (orange, outer circle) populations. No Gag overlapping clones were detected for one individual, H8 (not shown). (B) Number of Gag-specific TCR clones observed in naive, memory, and GC Tfh populations. Gray lines link the same patient. Bars indicate means (P values by two-tailed paired t test). (C) Mean clone size of Gag-specific T cells, HA-specific T cells, and bulk clones of unknown specificity from the GC Tfh population. (D) Number of distinct nucleotide (nt) sequences per CDR3 amino acid (aa) sequence for Gag-specific T cells, HA-specific T cells, or bulk GC Tfh cells. Data from all four individuals were aggregated for (C) and (D). Error bars indicate SEM. N.S., not significant. ***P < 0.001 by two-tailed t test.

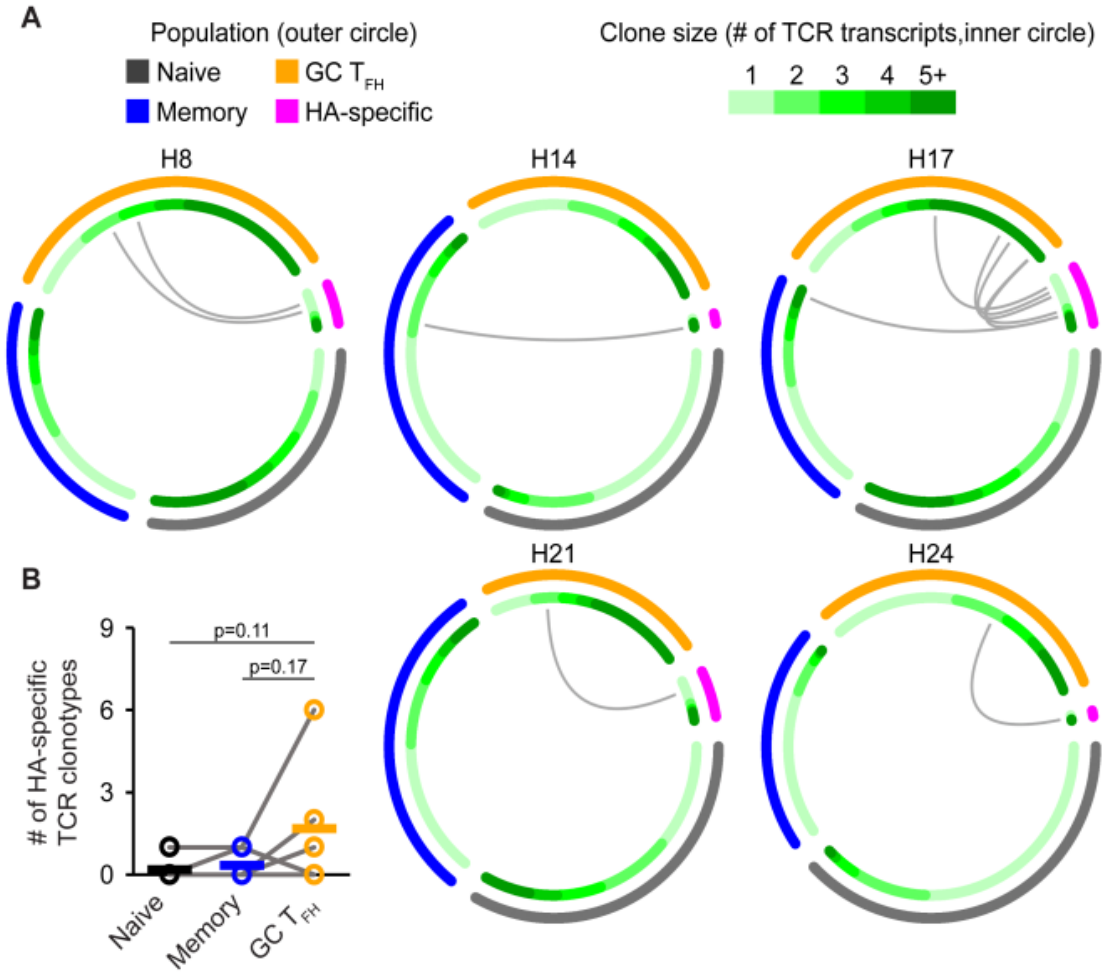


Figure 3.7: HA-specific CD4 T cell clones detected in HIV-infected LNs. (A) HA-specific TCR clones overlap with HIV+ LN CD4+ T cell populations. Each thin slice of the arc represents a unique TCR sequence, ordered by the clone size (darker green for larger clones, inner circle). Grey curves indicate HA-specific TCR nucleotide sequences found in naive (black, outer circle), memory (blue, outer circle), and GC Tfh (orange, outer circle) populations. No HA-overlapping clones were detected for one subject, H2 (not shown). (B) Number of HA-specific TCR clones observed in naive, memory, and GC Tfh populations. Grey lines connect samples from the same patient. Bars indicate means. Indicated P-value by two-tailed paired t test.

3.1.3 Materials and Methods

Study design

The goal of the study was to define Tfh cell diversity in primary human LNs. All samples were de-identified and obtained with Institutional Review Board approval from the University of Pennsylvania.

CyTOF staining

Metal conjugation of CyTOF antibodies was performed according to the manufacturer protocol using the X8 Maxpar kit (Fluidigm). Cells were stained with antibody panel, washed three times, then resuspended in 2% paraformaldehyde (Electron Microscopy Sciences) with 125nM iridium intercalator (Fluidigm) for an overnight incubation at 4°C. The next day, cells were washed three times, including a final wash in distilled water, and resuspended in water containing normalization beads before acquisition on CyTOF 2.

Flow-cytometry based antibody staining

Identifying bulk T cell subsets for direct sequencing: Cryopreserved were stained with CD4-AF700 (OKT4, eBioscience), CD11b-PE/cy5 (ICRF44, Biolegend), CD19-PE/cy5 (HIB19, Biolegend), and CD8-PE/cy5 (HIT8a, Biolegend) in the dump channel, fixable aqua dye (ThermoFisher) for live/dead discrimination, CD57-FITC (HCD57, Biolegend), CD45RO-BV605 (UCHL1, Biolegend), CXC5-PE/TexasRed (J252D4, Biolegend), ICOS-APC (C398.4, Biolegend), PD-1-BV785 (EH12.2H7, Biolegend), CCR7-PE/cy7 (G043H7, Biolegend). LN cells were sorted by naïve (CD45RO-CXCR5-CD57-CCR7+), memory (CD45RO+CXCR5-PD-1-ICOS-), or GC TFH cell phenotypes (CD45RO+CXCR5+PD-1+CD57+) using BD FACSARIA.

Identifying antigen-specific T cells: Peptide-reactive T cells were identified using the following antibodies depending on the experimental condition: CD40L (24-31, Biolegend), CD69 (FN50, BD biosciences), anti-IL-21 APC antibody (3A3-N2, Biolegend), Ox40 (Ber-ACT35, Biolegend), CD25 (BC96, Biolegend).

TCRb sequencing and analyses

TCR sequences from single cells were obtained by a series of three nested polymerase chain reactions, as previously described [75,76]. TCR junctional region analysis was performed using IMGT/V-Quest. For bulk cell analyses, TCR library generation and raw sequence processing were performed using MIDs with primers listed in Table S3[43,55].

Clone size distribution and normalized Shannon entropy: The size of each TCR clone was determined by the number of TCR transcripts of that sequence detected. The sizes were then normalized by the total number of TCR transcripts detected in that sample to yield the relative clone size in percent. Normalized Shannon entropy was calculated as previously described[73].

Amino acid translation and degeneracy: The CDR3 blast module of MIGEC[26] was used to translate the CDR3 nucleotide sequences into amino acid sequences. For each amino acid CDR3 sequence, the number of distinct nucleotide sequences (TCR clones) encoding that amino acid CDR3 sequence was tallied as the degree of degeneracy[73]. Amino acid CDR3 sequences encoded by 2 or more TCR clones were labeled as degenerate, and degeneracy versus relative clone size was analyzed to identify expanded, degenerate clones.

Antigen-specific TCR identification: TCR clones from the peptide-stimulated cells were used to establish donor-specific Gag- and HA-specific TCR sequences. TCR clones found in both the Gag- and HA-stimulated cultures were eliminated, as they likely originated from basally

activated T cells within the initial LN sample. These antigen-specific sequences were then queried in the bulk naïve, memory, and GC TFH cells to identify Gag- and HA-specific clones within the respective populations. Circlize R package[77] was used to visualize circus plots.

Statistical methods

Assessment of normality was performed using D'Agostino-Pearson test. Pearson or Spearman's rank correlation was used depending on the normality of the data to measure the degree of association. The best-fitting line was calculated using least-squares fit regression. Statistical comparisons were performed using two-tailed Student's t test or Wilcoxon signed-rank test, with a P value of <0.05 as a cutoff to determine statistical significance. Multiple comparisons were corrected using Holm-Sidak method. Statistical analyses were performed using GraphPad Prism.

3.1.4 Discussion

How HIV affects lymphoid Tfh cells has been studied under limited settings. In part, the challenge has been the inaccessibility of human lymphoid tissues and the tools available to interrogate a small number of cells. Here, we overcame these challenges using LNs obtained for clinical diagnostics from a mostly untreated HIV+ cohort. The data described here represent a comprehensive phenotype and TCR analysis of Tfh cells in the LNs, including that of HIV-reactive T cells. We also analyzed LN samples from HIV− HCs, but because of ethical and practical limitations, HC-derived LNs were obtained from different body sites and should be interpreted with this potential caveat. Our data based on TCR repertoire sequencing analyses provided evidence for antigen-driven expansion of Tfh cells and selection for certain preferred CDR3 sequences during chronic HIV infection. We further demonstrated that these GC Tfh cells acquire a distinct functional phenotype and become dominated by an IL-21+ functional subset.

We used HIV infection to ask how prolonged antigen stimulation alters the composition of Tfh cells in the LN. In vitro studies have suggested that Tfh cells could become inhibited in the context of chronic inflammation and fail to activate appropriately to TCR stimulation via induction of PD-1–mediated inhibitory signals[62]. The increase in Tfh cells could then be explained by an overabundance of cytokine signals in HIV-infected LNs that activated T cells in an antigen-independent fashion[67,68]. Although our data do not rule out a contribution from bystander T cell expansion, our data are most consistent with the model where Tfh cell pathology, manifested as clonal expansion and reduced poly functionality, is primarily an antigen-driven process. By TCR repertoire sequencing, we showed that certain TCR clonotypes become expanded within GC Tfh cells. A small portion of HIV-specific clones harbor distinct nucleotide sequences that converge to the same amino acid sequence—a signature of antigen-driven

selection. Convergent selection of TCR sequences is expected only when there is external pressure to select for certain CDR3 binding motifs. These data provide strong evidence for an antigen-driven process and additionally suggest that B cells, in their capacity as antigen-presenting cells, also shape the composition of Tfh cells. We measured the extent of clonal restriction by single-cell TCR sequencing. We found different clonal frequencies in Gag-reactive IL-21+ T cells between different HIV+ patients, with expanded clones occupying a significant proportion of Gag-reactive response in some individuals. Although we did not evaluate Env-reactivity directly, reduced TCR diversity will likely also affect T cells that recognize other HIV antigens. How Tfh cell repertoire relates to the selection of protective and/or neutralizing antibody responses remains poorly understood. The density of peptide–major histocompatibility complex (MHC) presented by competing B cells has been shown to be a major factor that determines the magnitude of T cell help [78–80], but the diversification of antigenic variants by viral mutation provides an additional layer of complexity in the selection of relevant B cells during chronic HIV infection[81]. Previous studies have shown that early HIV Env gene diversity predicts development of antibody breadth[82]. This raised the possibility that a diverse repertoire of HIV-specific Tfh cells may be necessary to capture the breadth of viral variants, and an oligoclonal Tfh population that focuses T cell reactivity to nonproductive but common antigenic viral sequences may neglect rare B cells that have neutralizing potential. Future studies to determine whether individuals with more diverse HIV-specific Tfh cell repertoire are more successful at generating broadly neutralizing antibodies will provide additional insights.

3.2 Application of MIDCIRS on antibody repertoire: Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children*

3.2.1 Introduction

MIDCIRS' high coverage and dynamic range allow us to sequence samples with very few cells, for example, infant blood samples. We use MIDCIRS to examine the antibody repertoire diversification in infants (<12 months old) and toddlers (12–47 months old) from a malaria endemic region in Mali before and during acute Plasmodium falciparum infection. Although the antibody repertoire in fetuses[83], cord blood[84], young adults[85], and the elderly[85,86] has been studied, infants and toddlers are among the most vulnerable age groups to many pathogenic challenges, yet their immune repertoires are not well understood. Infants are widely thought to have weaker responses than toddlers to vaccines because of their developing immune systems[87]. Thus, understanding how the antibody repertoire develops and diversifies during a natural infection, such as malaria, not only provides valuable insight into B cell ontology in humans, but also provides critical information for vaccine development for these two vulnerable age groups.

---

* Wendel, B.S.†, **He C.**†, Qu M.† et. al. Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children. Nat. Commun. 8, 531 (2017). B.S.W. performed all malaria related experiment and part of data analysis; C.H. performed antibody sequencing, mutation and selection pressure analysis; M.Q. developed the sequencing protocol using sorted naive B cells; D.W. helped with sequence analysis; S.M.H. helped with library construction; K.Y.M. helped with sequencing; J.X. helped with lineage visualization, E.W.L, P.D.C., and S.K.P. selected malaria patients, provided samples and helped with experimental design; P.R. provided computation resources and helped with analysis; K.C. helped with lineage structure algorithm optimization and lineage visualization; N.J. conceived the idea, designed the study, and directed data analysis; B.S.W. and N.J. wrote the paper with contributions from all co-authors.

†: These authors contributed equally.

Using peripheral blood mononuclear cells (PBMC) from 13 children aged 3–47 months old before and during acute malaria, with two of the children followed for a second year and nine additional pre-malaria individuals we show that infants and toddlers use the same V, D, and J combination frequencies and have similar complementarity determining region 3 (CDR3) length distributions.

Although infants have a lower level of average SHM than toddlers, the number of SHMs in reads that mutated in infants is unexpectedly high. Infants have a similar, if not higher, degree of antigen selection strength, assessed by the likelihood of amino acid-changing SHMs, compared with toddlers. Remarkably, during acute malaria, antibody lineages expand in both infants and toddlers, and this expansion is coupled with extensive diversification to the same degree as in young adults in response to acute malaria[88,89]. In summary, using an accurate and high-coverage IR-Seq method, we discover features of the antibody repertoire that were previously unknown in infants and toddlers, shedding light on the development of the immune system and its interactions with pathogens.

3.2.2 Results

Infants and toddlers have similar VDJ usage and CDR3 lengths.

Equipped with this ultra-accurate and high-coverage antibody repertoire-sequencing tool, we applied it to study the antibody repertoire of infants and toddlers residing in a malaria endemic region of Mali. From an ongoing malaria cohort study[90], we obtained paired PBMC samples collected before and during acute febrile malaria from 13 children aged 3–47 months old (Figure 3.8 and Table S5). Two of the children were followed for an additional year, giving 15 total paired PBMC samples. An average of 3.8 million PBMCs per sample was directly lysed for RNA purification. All PBMCs were subjected to MIDCIRS analysis. An average of 3.75 million sequencing reads was obtained for each PBMC sample (Table S6).

For all PBMC samples, sequencing approximately the same number of reads as the cell numbers saturates the rarefaction curve (Figure 3.9). IgM accounts for more than 50% of the repertoire in infants but reduces to less than 50% in toddlers (Figure 3.10). VDJ gene usage is highly correlated for IgM between infants and toddlers (Figure 3.11), demonstrating that the same mechanism of VDJ recombination is used to generate the primary antibody repertoire in infants and toddlers(Figure 3.11). The diagonal lines in each panel indicate same sample self-correlation, and the two shorter off-diagonal lines indicate correlations from two time points of the same individual. These data recapitulate previous observations in zebrafish that clonal expansion-induced differences on the number of reads in each VDJ class can confound the highly similar VDJ usage during B cell ontology[31]. In addition, infants and toddlers have similar CDR3 length distributions across the three isotypes and both time points (Figure 3.12), consistent with recent studies of PBMCs from 9-month-olds infants[83,84] and adults[91,92], and

confirming the previous results that an adult-like distribution of CDR3 length is achieved around 2 months of age[93].

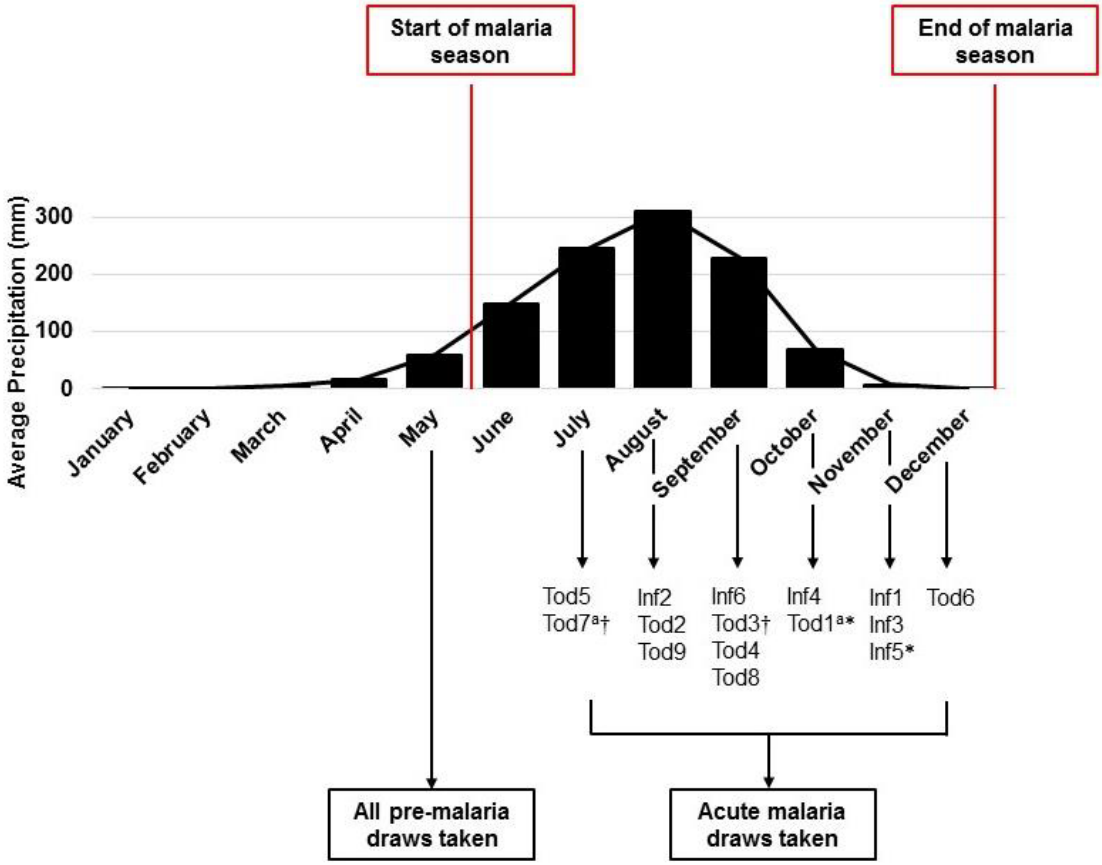


Figure 3.8: Sample collection timeline. All pre-malaria blood draws were taken in May, just before the start of the rainy season. Acute malaria blood draws were taken 7 days after the onset of acute febrile malaria. Unless otherwise indicated (a), all samples were collected during 2011. Average precipitation was estimated from the neighboring city of Bamako, Mali (climatemps.com).

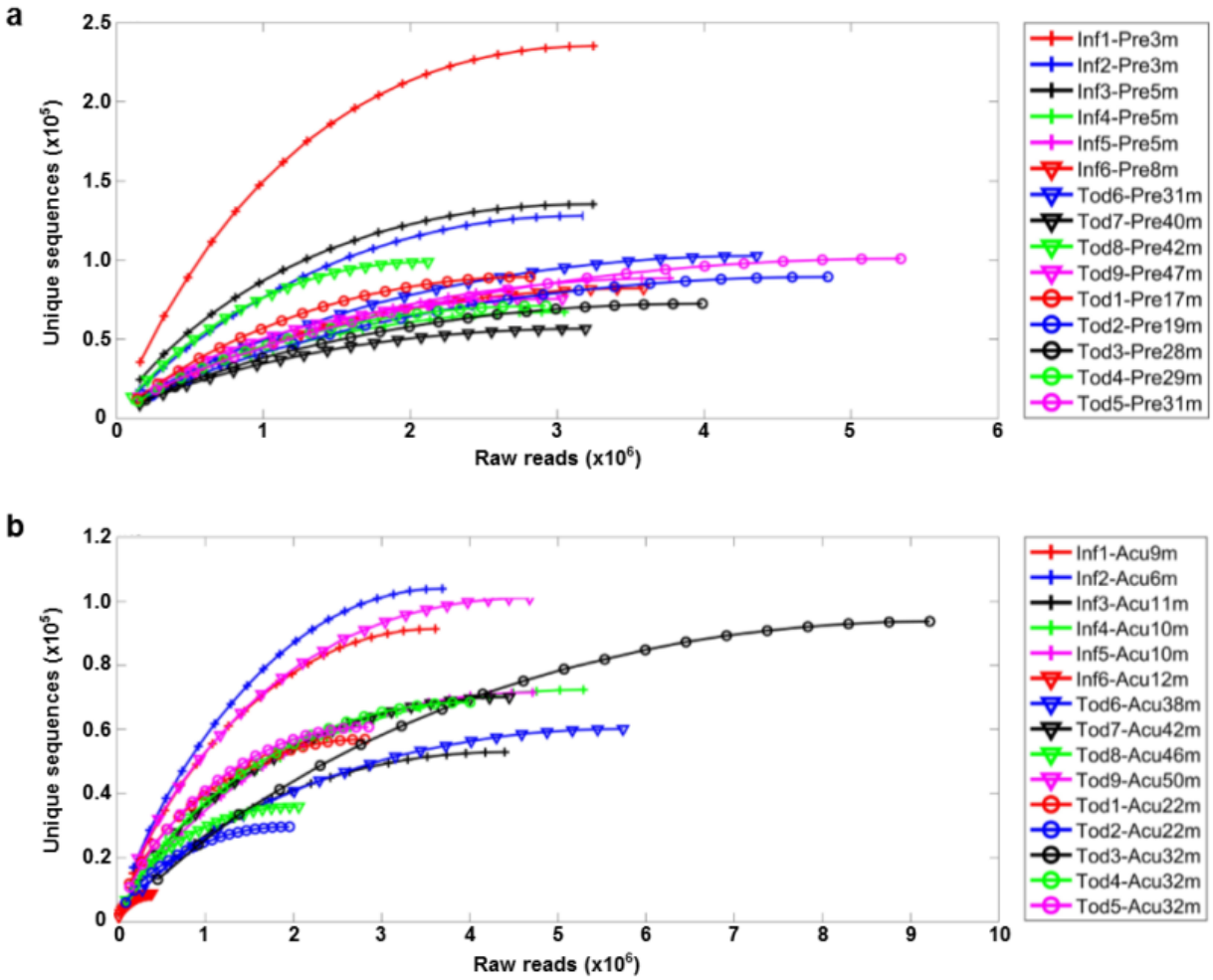
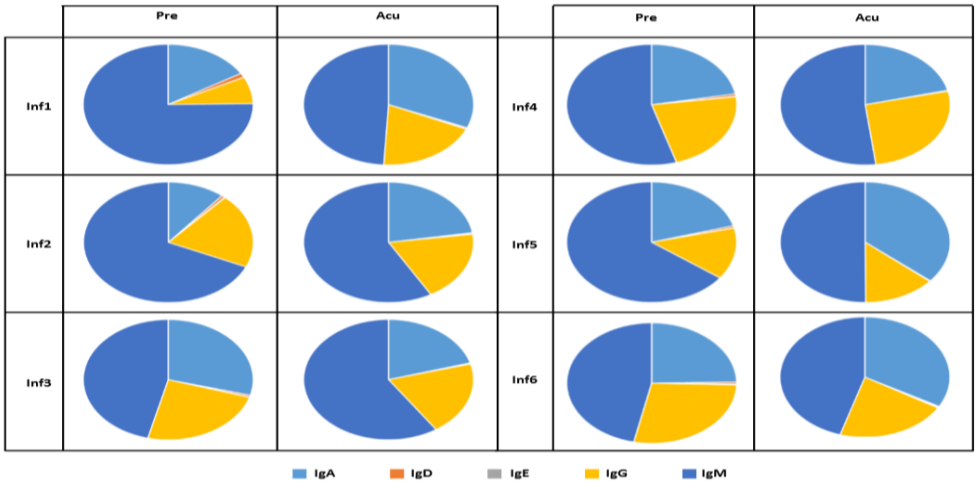


Figure 3.9: Rarefaction analysis of paired PBMC malaria cohort sequencing libraries. (a) Pre-malaria PBMC rarefaction curves (N=15). (b) Acute malaria PBMC rarefaction curves (N=15). Raw reads were subsampled to varying depths, and MIDCIRS was used to determine the number of unique RNA molecules. All single-read sequences that occurred before subsampling were discarded. Single-read sequences that occurred as a results of subsampling were included as unique RNA molecules. The number of unique RNA molecules discovered saturated for all samples, indicating adequate sequencing depth.



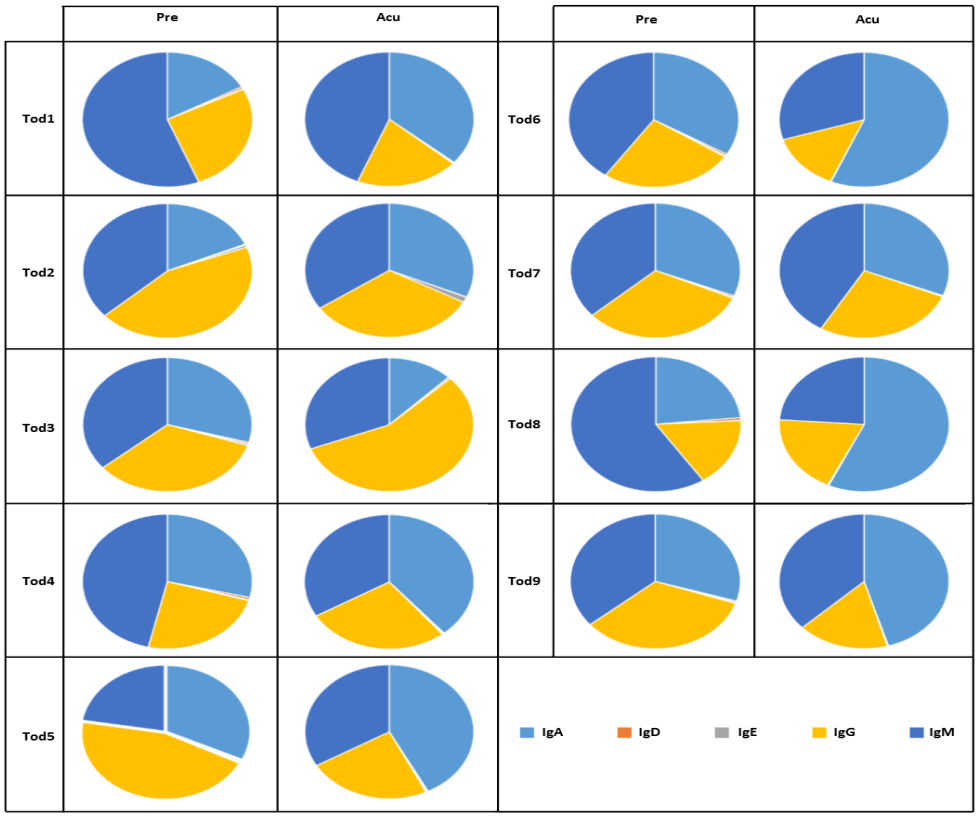


Figure 3.10: Antibody isotype distribution for infants and toddlers. Antibody isotypes were assigned based on the portion of the constant region sequenced for infants (A) and toddlers (B). Isotype distribution was weighted on the number of RNA molecules.

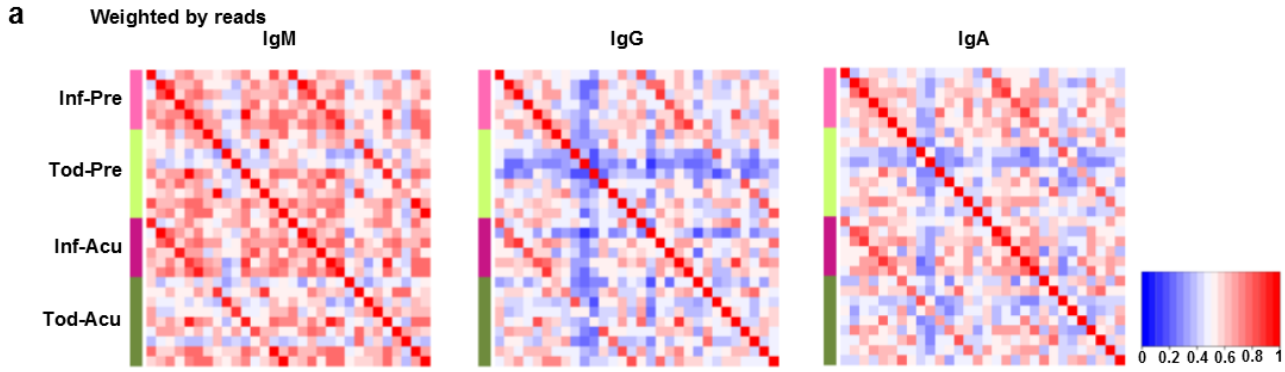


Figure 3.11: Correlation between VDJ usage in paired PBMCs samples (N=15 pairs of pre-malaria and acute malaria). Correlations weighted by reads. The color bar left of each panel as well as in figure legend indicates the sample group: infant pre-malaria (pink), toddler pre-malaria (light green), infant acute malaria (maroon), and toddler acute malaria (dark green). Color indicates strength of Pearson correlation. The diagonal lines in each panel indicate same sample self-correlation; two shorter off-diagonal lines indicate correlations from two time points of the same individual.

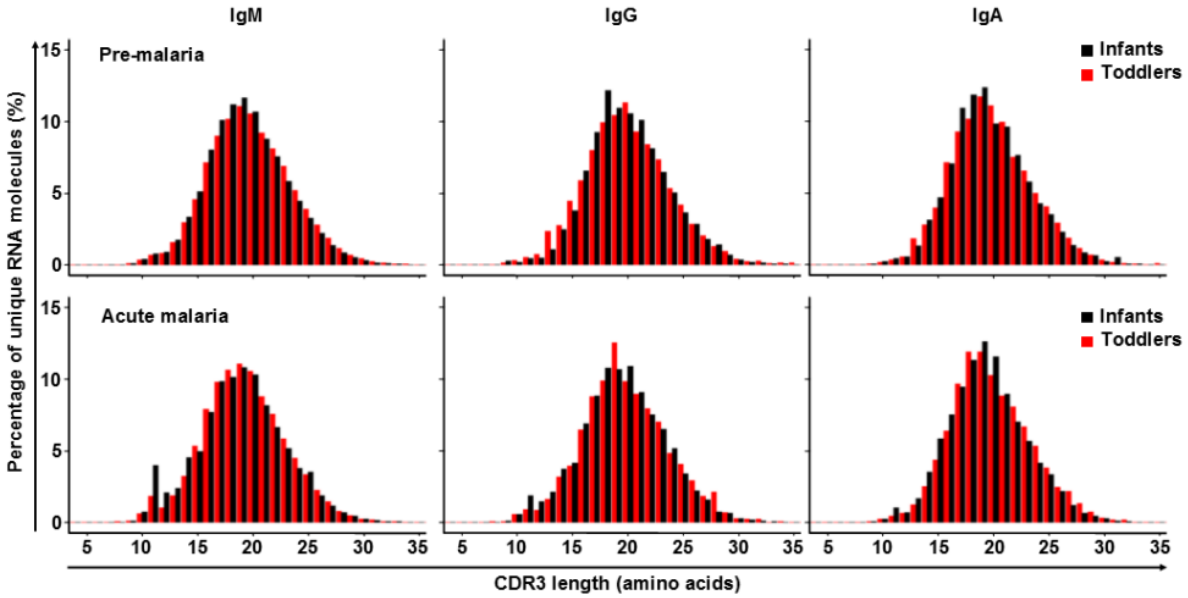


Figure 3.12: CDR3 amino acid lengths of infants (black, N=6) and toddlers (red, N=9) at pre-malaria (top) and acute malaria (bottom) time points, separated by isotype.

Both infants and toddlers have unexpectedly high SHM

SHM is an important characteristic of antibody repertoire secondary diversification due to antigen stimulation[94]. Although it has been demonstrated before that infants have fewer mutations in their antibody sequences than toddlers and adults, the limited number of sequences for only a few V genes does not provide convincing evidence of the levels of SHM in infants[95]. A recent study using the first generation of IR-seq showed that two 9-month-old infants averaged at least six SHMs in IgM of an average length of 500 nucleotides[83]. These numbers are equivalent to, if not higher than, reported SHM rates in IgM sequences from healthy adults day 7 post influenza vaccination[27] and are much higher than a low-throughput infant study using a few V genes and limited antibody sequences[96]. Owing to the inherent errors associated with the first generation of IR-seq as discussed above, it is possible that PCR and sequencing errors had a role[83]. In addition, it remains unclear whether infants (< 12 months old) are able to generate a significant number of mutations in response to infection, which would demonstrate their capacity to diversify the antibody repertoire[97].

Here, we show that infants (< 12 months old) and toddlers (12–47 months old) reach an unexpectedly high level of SHMs in all three major isotypes, particularly IgG and IgA[98] (Figure 3.12a). Although the mutation distributions remain in the low end of the spectrum for IgM, the number of mutations is significantly higher in IgG and IgA for both age groups. The threshold for the 10% most highly mutated unique RNA molecules is around 10 in infant IgG and IgA sequences (Figure 3.12a, infants, right of the blue long vertical lines) and around 20 in toddler IgG and IgA sequences (Figure 3.12a, toddlers, right of the blue long vertical lines). To minimize any possible inflation of SHMs, we excluded all sequences that were mapped to novel alleles, which were identified by both TIgGER[99] and inspecting IgM sequences (Methods).

These putative novel alleles account for 8% of all unique sequences on average. Naive B cells from these same patients, sorted as a control, harbor only 0.55 mutations on average, as expected. Upon acute malaria infection, the SHM histogram shifts rightward for almost all isotypes in almost all individuals (Figure 3.12a, the right shift of pink long vertical line compared to blue long vertical line), including infants. These results demonstrate high levels of SHM that exceed what have been documented previously[9596,98].

SHM load is distinct between infants and toddlers.

The differences in the shapes of SHM distributions of infants and toddlers, steadily decreasing from unmutated for infants in all three isotypes while peaking around 10 for toddlers in IgG and IgA (Figure 3.13a), suggest that the total SHM load might reflect the history of interactions between the antibody repertoire and the environment, including malaria exposure. As the malaria season is synchronized with the 6-month rainy season (Figure 3.8), and > 90% of the individuals in this cohort are infected with P. falciparum during the annual malaria season[90], we hypothesized that the SHM load would increase with age. However, we found that the SHM load rapidly increases with age in infancy and then appears to plateau around 12 months of age (Figure 3.13b). The two-staged trend of SHM load remains for all three isotypes (Figure 3.13b), with samples around the transition having the largest variation. Detailed comparisons show that, consistent with the two-stage trend, toddlers have a higher SHM load compared with infants for all three isotypes at both pre-malaria and acute malaria time points (Figure 3.13c). Although there is a significant increase on SHM load upon acute malaria infection in IgM for both infants and toddlers, bulk PBMC analysis does not show a significant increase in IgG or IgA, possibly because of the already elevated SHM base level. This, along with the two-stage trend (Figure 3.13b), suggests that 12 months is an important developmental threshold for secondary antibody

repertoire diversification: before this threshold, the global repertoire is quite naive but can quickly diversify upon a natural infection.

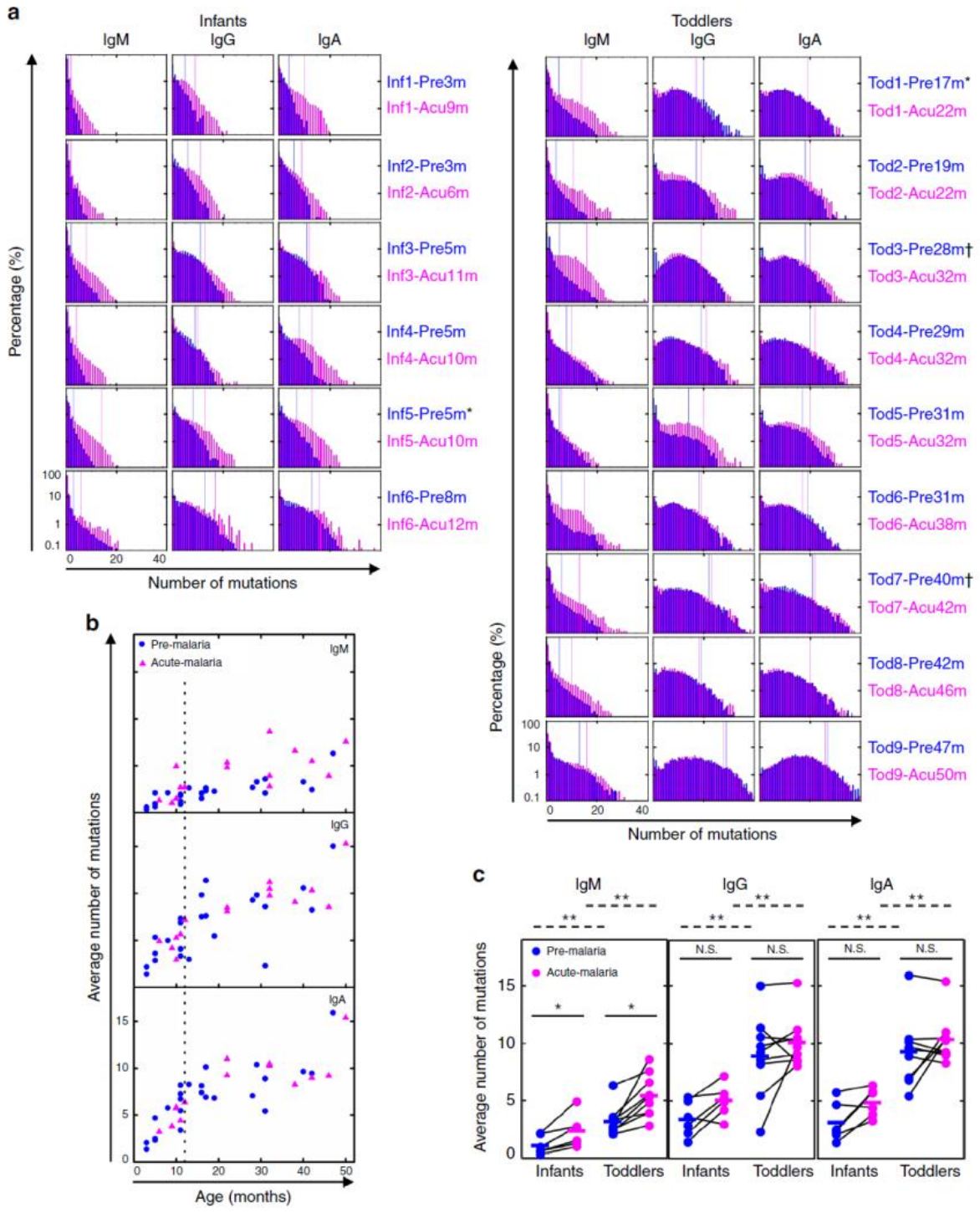


Figure 3.13: Infants and toddlers are separated into two stages based on SHM load. a). Distribution of SHM number for infants (N = 6) and toddlers (N = 9), from whom we had paired pre-malaria (blue) and acute (pink) malaria samples, weighted by unique RNA molecules. Blue and pink long vertical lines represent the number of mutations above which 10% of sequences fall for the respective samples. * and † demarcate samples derived from the same individuals

Figure 3.13 cont.
followed for two malaria seasons. b) Age-related average number of mutations in pre-malaria (blue circle, N = 24, NInfant = 11, NToddler = 13) and acute malaria (pink triangle, N = 15, N-Infant = 6, N-Toddler = 9) samples, weighted by RNA molecules. Dashed line indicates the age boundary for infants (< 12 months old) and toddlers (12–47 months old). c) Comparison of average number of mutations for paired infants and toddlers. Pre- (blue) and acute (pink) malaria samples separated by isotype; lines connect paired samples (N-Infant,paired = 6, N-Toddler paired = 9). Bars indicate means. *P < 0.05, **P < 0.01, N.S. indicates no significant difference by two-tailed Mann–Whitney U test (between age groups, dashed lines) or two-tailed Wilcoxon signed-rank test (between paired time points, solid lines). Differences in variance were not significant by squared ranks test.

SHMs are similarly selected in infants and toddlers.

One of the key features of antibody affinity maturation is antigen selection pressure imposed on an antibody, which is reflected in the enrichment of replacement mutations[100] in the CDRs, the parts of the antibody that interact with antigens, and the depletion of replacement mutations in the framework regions (FWRs), the parts of the antibody responsible for proper folding. The unexpectedly high level of SHMs observed in infants prompted us to ask whether those SHMs have characteristics of antigen selection, as seen in older children and adults. As previous studies have shown that infants have limited CD4 T cell responses and neonatal mice exhibit poor germinal center formation[87], we hypothesized that infant antibody sequences would display weaker signs of antigen selection.

Here, we use a recently published tool, BASELINe[101], to compare the selection strength. BASELINe quantifies the likelihood that the observed frequency of replacement mutations differs from the expected frequency under no selection; a higher frequency implies positive selection and a lower frequency implies negative selection, and the degree of divergence from no selection relates to the selection strength. Surprisingly, despite infants harboring fewer overall mutations, these mutations are positively selected in the CDRs and negatively selected in the FWRs in both IgG and IgA (Figure 3.14b, c, e, f). Contrary to the hypothesis that infants would

have a lower selection strength than toddlers, for both IgG and IgA, infants actually have a higher selection strength at both pre-malaria and acute malaria time points (Figure 3.14). The lower selection strength in infant IgM sequences at the pre-malaria time point is significantly higher during acute malaria infection (Figure 3.14a, d, CDR black curves between two time points, $P < 0.0001$ (numerical integration, as previously described)), suggesting that the significant increase in SHM is antigen-driven and selected upon. To compare with a large amount of historical adult data, we calculated replacement to silent mutation ratios (R/S ratios), which are about 2-3:1 in FWRs and 5:1 in CDRs for both infants and toddlers (Table S6). These results are similar to adults[100,102–104], and much higher than what has been reported for children previously using a very limited number of sequences[105]. We also noticed that R/S ratio in the FWRs of IgM was much higher in infants, contrary to the BASELINe results, which highlights the importance of incorporating the expected replacement frequency when considering selection pressure. These results suggest that as an end result of interactions between antigen selection and SHM, the degree of antibody amino acid changes is comparable in infants, toddlers, and adults. It also suggests that cellular and molecular machineries for antigen selection are already in place in infants.

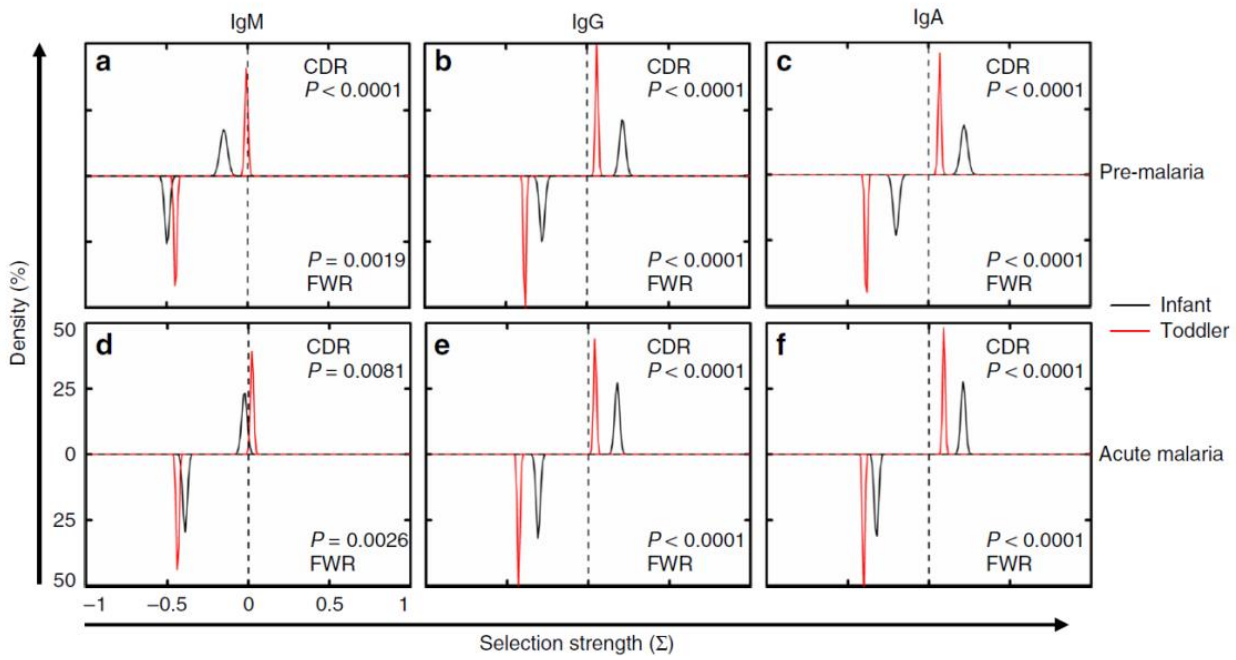


Figure 3.14: Antigen selection strength comparisons between infants and toddlers. Selection strength distributions, as determined by BASELINe[101], were compared between infants (black) and toddlers (red) for PBMCs from pre-malaria (a–c) (N-infant = 6, N-toddler = 9) and acute malaria (d–f) (N-infant = 6, N-toddler = 9) time points, separated by isotype: a, d IgM; b, e IgG; and c, f IgA. Selection strength on CDR (CDR1 and 2, top half of each panel) and FWR (FWR2 and 3, bottom half of each panel) for unique RNA molecules was calculated. CDR3 and FWR4 were omitted due to the difficulty in determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. P value calculated as previously described[101].

Antibody Lineage diversify upon acute malaria

One key feature distinct antibody with TCR is ability of somatic hypermutation, which are random mutations on the receptor sequences. The exhaustive sequencing data obtained by MIDCIRS offers the possibility to reconstruct clonal lineages that trace B cell development. Clonal lineages contain different species of unique antibody sequences that could be progenies derived from the same ancestral B cell, which are normally organized as a 'tree' format[32]. B cell clonal lineage analysis has been used to track affinity maturation and sequence evolution of HIV broadly neutralizing antibodies[106,107]. The structure of antibody clonal lineage provides another way to quantify antibody repertoire and its structure serves as a molecule clock to understand

how one antibody lineage is evolving. We constructed all the lineages of the antibody repertoires with a pre-determined threshold[32] (Figure 3.15a) and compared the clonal expanded lineages before and after malaria season (Figure 3.15b). Results show upon acute malaria infection, the fraction of non-singleton lineages increases in both infants and toddlers, which hints antibody lineages get expanded due to malaria infection.

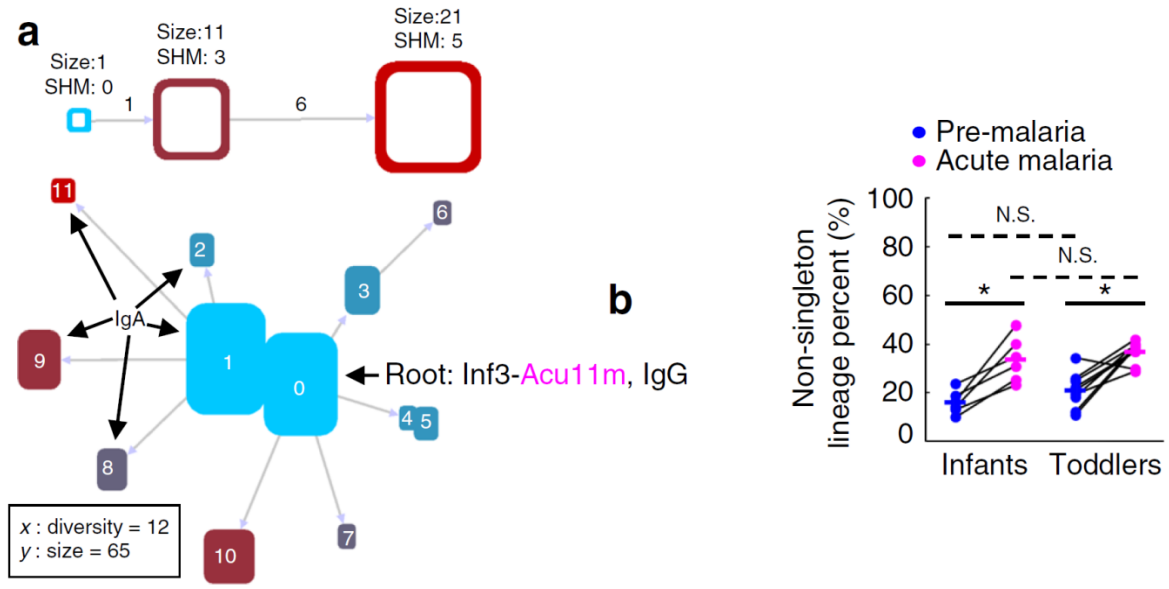


Figure 3.15: Lineages from malaria samples. a). Structure of one example lineage. Each node is a unique RNA molecule species. The height of the node corresponds to the number of RNA molecules of the same species, the color corresponds to number of nucleotide mutations, and the distance between nodes is proportional to the Levenshtein distance between the node sequences, as indicated in the legend above each lineage. All unlabeled nodes share the isotype with the root. b). The non-singleton lineage percent (lineages comprises at least two RNA molecules) between infants and toddlers at pre-malaria (blue) and acute (pink) malaria. *P < 0.05 by two-tailed Wilcoxon signed-rank test (between time points, solid lines); N.S. indicates no significant difference by two-tailed Mann–Whitney U test (between age groups, dashed lines).

3.2.3 Materials and Methods

Preliminary read processing

Raw reads from Illumina MiSeq PE250 were first cleaned up. Only reads that exactly matched the corresponding library indices were included for further processing. The end of each raw read was trimmed such that all bases had a quality score of 25 or higher. Reads 1 and 2 were merged using the SeqPrep tool (https://github.comjstjohn/SeqPrep). The merged reads were filtered with specific V-gene and constant region primers to determine immunoglobulin (Ig) sequencing reads. The primers were then truncated from the reads. The retained reads were further truncated to 320bp for the Naive B cells in method verification experiments and 330bp for samples from malaria cohort.

MID sub-group generating

Raw reads were split into MID groups according to their 12 nucleotide barcodes. For each MID group, quality threshold clustering was used to cluster similar reads. This process groups reads derived from a common template RNA molecule together while separating reads derived from distinct RNA molecules. A Levenshtein distance of 15% of the read length was used as the threshold. This was calibrated using RNA controls with known sequences. For each sub-group, a consensus sequence was built based on the average nucleotide at each position, weighted by the quality score. In the case that there were only two reads in an MID sub-group, we only considered them useful reads if both were identical. Each MID sub-group is equivalent to an RNA molecule. Next, we merged the entire identical consensus to form unique consensus sequences, or unique RNA molecules, which were used to estimate the diversity and assess the sequencing depth in rarefaction analysis

Error rate calculation

The difference between the consensus sequence for an RNA molecule and the raw reads associated with it represent the errors generated in either PCR or sequencing. The error rate can be calculated using the following formula:

$$\mathrm{ErrorRate(Raw)} = \frac{\sum_{i=1}^{N_I} \mathrm{Diff(i, I)}}{N_I \times L}$$

where Diff(i,I) is the Levenshtein distance between read i and the consensus sequence in MID sub-group I, N_I is the number of reads in MID sub-group I, and L is the read length.

In order to estimate the improved error rate using MIDCIRS, we equally divided the raw reads from one library into two datasets. The same MID sub-group generating process was performed on both datasets. By comparing the differences between the consensus sequences with identical MID between these two datasets, we can calculate the improved error rate for using MID sub-groups as:

$$\mathrm{ErrorRate(MID)} = \frac{\sum_{I,J} \mathrm{Diff(I, J)} \times N_I}{\sum_I N_I \times L}$$

,where Diff(I,J) is the Levenshtein distance between the consensus I and consensus J which have the identical MID, $N_I$is the number of reads in MID sub-group I, and L is the read length.

VDJ definition and mutation counts

As described in previous work, similar methods were used to define the V, D, and J gene segments for all sequences[31]. From the International ImMunoGeneTics information system database (IMGT, http://www.imgt.org/textes/vquest/refseqh.html)[108], human heavy chain variable gene segment sequences (249 V-exon, 37 D-exon and 13 J-exon) were downloaded.

Each unique sequence was first aligned to all 249 V gene alleles. The specific V-allele with a maximum Smith-Waterman score was then assigned. In some cases, newly identified germline alleles, defined either by TIgGER or our method (below), were added to the template sequences. J-segments and D-segments were then similarly assigned. The number of mutations from germline sequence was counted as the number of substitutions from the best aligned V and J templates as previously described[29]. The CDR3 was omitted due to the difficulty in determining the germline sequence. The germline sequences of V, D, and J gene segments were grouped by combining similar alleles into families using IMGT designation in VDJ correlation plots. In total, 58 V, 27 D, and 6 J families were used.

Novel allele detection

To address the possibility of novel germline alleles inflating the observed number of mutations, new germline alleles were assembled. In short, IgM sequences for each subject were aligned and assigned to the traditional V-gene alleles in the IMGT database. If novel alleles exist in subjects, parts of unique RNA sequences will be assigned as mutations when they are actually derived from differences between novel and traditional alleles. The ratios of unmutated unique RNA molecules to those with one, two, three and four mutations compared to the IMGT germline were determined, and if any were found to be less than 2 to 1, the alleles were flagged for further inspection. Unique RNA molecules were used to minimize the contributions of clonal expansion, and IgM sequences were used to minimize the contributions of somatic hypermutation. Sequences within flagged alleles were then aligned to the closest IMGT germline to determine if the mutations are truly polymorphisms. When identical mutation patterns were observed in a minimum of 80% of all sequences in a flagged allele family, it was deemed a novel germline

allele. For subjects with sorted NaiBs, novel alleles were generated from the NaiB BCR sequences to complement those found in the bulk IgM sequences.

TIgGER was used as previously reported as another method to discover novel alleles [109]. TIgGER compares the mutation rate at a specific position to the overall number of mutations for sequences within the same assigned V-gene allele. Outliers within the low mutation region suggests the existence of a novel allele, and the shape of the curve can effectively distinguish between individuals homozygous and heterozygous for the novel allele.

Our new method and TIgGER have a 90% percent overlap in newly identified alleles. Discrepancies between the two methods were treated with a conservative estimation on the number of SHM, meaning we liberally included novel alleles as part of the germline gene segments. Non-overlapping novel alleles were manually inspected, and the union of novel alleles detected by TIgGER and our method was included as part of the germline gene segments. Sequences mapped to these novel alleles were excluded from our analysis, which accounts for an average 8% of all sequences.

Translation from nucleotide to amino acid sequences

Nucleotide sequences were translated into amino acid sequences based on codon translation. The unique RNA sequences were inputted to IMGT High V quest to translate into amino acid sequences. The boundary of the CDR3 is defined by IMGT numbering for Ig and two conserved sequence markers of 'Tyr-(Tyr/Phe)-Cys' to 'Trp-Gly.' CDR3 length was determined according to these anchor residues.

Selection pressure

The selection pressure was evaluated via BASELINe[99]. The unique RNA molecules of PBMC, populations were inputted to BASELINe and compared with the closest IMGT germline alleles. The observed number of replacement and silent mutations were compared with the expected number of mutations for the assigned germline sequence. A selection strength as measured by the probability density function (PDF) was generated using BASELINe[99] to indicate the direction and degree for CDR (CDR1 and 2) and FWR (FWR2, and 3) regions for each unique RNA molecule first, then combined for each individual, and then further combined for each of the two groups, infants and toddlers. The PDFs were plotted and compared between infants and toddlers. The associated P values comparing two group PDFs were calculated using BASELINe. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

Replacement/Silent mutation

According to the amino acid sequence translation results and V/D/J gene templates alignment results, we counted the number of nucleotide mutations resulting in amino acid substitutions (replacement, R) or no amino acid substitutions (silent, S) in FWR region (FWR2 and 3) and CDR region (CDR1 and 2). The number of silent and replacement mutations was averaged in each age-group (Infant and Toddler) and the ratio for silent vs. replacement mutation was calculated. The CDR3 and FWR4 were omitted due to the difficulty in determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules.

Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

VDJ usage correlation

The correlation of VDJ usage between infants and toddlers were calculated with Pearson Correlation Coefficient as the following formula:

$$\mathrm{corr} = \frac{\sum_{v=\{V\},d=\{D\},j=\{J\}} (X_{vdj} - < X >)(Y_{vdj} - < Y >)}{\sqrt{\sum_{v=\{V\},d=\{D\},j=\{J\}} (X_{vdj} - < X >)^2 * \sum_{v=\{V\},d=\{D\},j=\{J\}} (Y_{vdj} - < Y >)^2}}$$

vdj refers to the combination of one v allele family from 58 V gene allele families ({V}), one d allele family from 27 D gene allele families ({D}), and one j allele family from 6 J gene allele families ({J}). For the reads weighted correlation, X_{vdj} and Y_{vdj} refer to the fraction of reads assigned to the respective vdj combination for subjects X and Y, respectively. $< X >$ and $< Y >$ are the average reads across all vdj combinations, i.e. 1/9396, where 9396 is the total possible number of vdj allele family combinations. For the lineage weighted correlation, these parameters refer to the fraction of lineages for each vdj allele family combination. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

Lineage structure construction and visualization

Representative lineages were selected to visualize the lineage structures and the evolution of antibody sequences. Lineage structures were generated using COLT (software can be downloaded here: http://www.cs.wright.edu/~keke.chen/software/colt.zip) and validated manually. We implemented a lineage visualization tool, COLT-Viz. In short, COLT considers constraints (e.g., isotype and timepoint) along with mutational patterns to build lineage trees. The

height of each node is proportional to the number of RNA molecules associated with the unique sequence (size), the color of each node relates to the number of SHMs, and the distance between nodes is proportional to the Levenshtein distance between the node sequences.

Code availability

References or links for all software tools used are listed in the relevant "Methods" sections. All other relevant data are available from the authors.

Data availability

Sequencing data that support the findings of this study have been deposited in dbGaP with the accession code phs001209.v1.p1.

3.2.4 Discussion

About 13,000 children under 1 year old die every day worldwide[110], and most of these deaths are caused by infection[87]. It has long been recognized that children's immune systems are immature at birth and require time to develop to provide protection against pathogens or respond to vaccines. However, few studies have focused on children's antibody repertoire development, diversification, and response to infection. Knowledge in this area holds great interest to vaccine development and vaccination strategy design. This is especially urgent for malaria, as it still kills about half a million children each year[111], and the most advanced malaria vaccine confers only partial, short-lived protection in African children[112].

Previous studies showed that there was evidence of SHM and antigen selection in infants 8 months of age or older by examining a few V-gene alleles[95]. However, it is not clear how widespread SHMs are in infant antibody repertoires and to what degree SHMs can be introduced in response to an infection. By using a comprehensive and unbiased analysis, here we show that infants as young as 3 months old can have 10% of sequences with five or more mutations, and they can further introduce mutations upon an acute febrile malaria infection to well over 20 SHM per 270 nt heavy chain V region. Compared with toddlers, there is a separation on SHM load around 12 months: this number gradually increases before 12 months and stays at a plateau after that regardless of repeated malaria incidents. Consistent with this trend is the similar pattern observed in the increase in the percentage of memory B cells and corresponding decrease in percentage of naive B cells with age: both plateaued after 12 months of age. Accordantly, SHM load in IgM, IgG, and IgA correlates with the percentages of naive and memory B cells. Surprisingly, regardless of the lower mutation load in infants, their mutations are similarly, if not more strongly, selected as those of toddlers, suggesting that the molecular machineries and other

cellular components involved in antibody selection are already developed in infants. In future analyses, it will be of interest to tease out the mechanistic contributions to a two-stage increase of average mutation number, in particular, the role of T cell help and germinal center formation. Regardless of these detailed mechanisms, it is clear infants can perform antibody selection as well as toddlers and adults, which provides some assurance of the effectiveness of vaccination in young children.

In summary, we systemically studied the antibody repertoire in malaria-exposed infants and toddlers and discovered several aspects of repertoire development, diversification, and capacity to respond to an infection that were not known before, which provides not only new parameters and approaches in quantifying vaccine efficacy beyond traditional serological titer but also venues for future studies of detailed molecular and cellular mechanisms that drive antibody repertoire differences between infants and toddlers.

Chapter 4: LiMPETs-Linear programming based Motif Pick and Enrichment analyze for Tcrs

4.1 INTRODUCTION

One main challenge of engineering T-cell receptors (TCRs) for immunotherapy is to correctly identify TCR sequences that can recognize certain pathogen associated epitope. Because of its importance, growing efforts have been made in this area. Several experiment techniques have been developed for TCR antigen-binding specificity measurement[113](e.g., Fluorescent pMHC tetramer, CyTOF with isotope-labeled pMHC tetramers, etc), but all the techniques suffer from low throughput.

Computational methods for predicting TCR specificity is attractive, since it's cheap and easy to be high-throughput. TCR recognizing its antigen peptide through protein-protein interaction, so computational methods for protein structure modeling have been applied to predict TCR-peptide binding affinity [114]. A routine way of modeling protein structure is to first model an initial structure (either by homology modeling or ab initio modeling), then use Molecule Dynamics (MD) simulation to refine the structure. However, there are several challenges to apply this for TCR-peptide complex modeling: 1). Homology modeling requires template structure of a very similar sequence, while TCR is highly diverse, current available crystal structures covers only a very small proportion of the whole repertoire; 2). The most important region to determine TCR antigen specificity are the Complementary Determining Regions (CDR1, CDR2, CDR3), since these regions contact directly to peptides. However, these regions normally adopt a linear loop structure, which means they are quite flexible and hard to model. The accuracy now is still unsatisfied and the prediction accuracy of the loops drops when it gets longer (the most accurate prediction has an average RMSD of 1.6~6A depending on the

loop length[115]); 3). Correctly model TCR itself is already hard, while to model the TCR-peptide complex is even harder, since conformation of those contact loops will change before and after binding to peptides[116]. 4). Even though the computation efficiency is growing dramatically (e.g., High-performance computing, GPU, etc.), MD simulation is still computationally expensive and to be used for screening such the diversified TCR repertoire is challenging.

It will be ideal if TCR's binding specificity can be predicted directly from its sequence. Machine learning seems to be a promising way, given its great power and capability to model even very complicated distributions. However, to train machine learning model requires significant amount of data, especially when the problem itself is complicated. A curated database has been designed to record published experimentally validated TCR antigen specificity[117], which provides valuable resource for TCR-antigen binding specificity research. However, these accumulated TCRs can still only cover a very small proportion of the enormous diversity of TCR repertoire. Also data collected in this curated database does not consider cross reactivity between TCRs and peptides.

Although TCR sequences are quite diverse, the specificity of each TCR is determined by the interaction between TCR with peptide-major histocompatibility complex (pMHC), TCRs recognize the same antigen must share some similarity. This similarity should be able to detected by computational methods and possibly been used for predicting TCR specificity. Researchers have shown conserved motifs are shared by TCRs of the same specificity at positions of high-antigen-contact probability [118]. Glanville et al. developed GLIPH [17], which can identify conserved motifs from TCRs CDR3 sequences of the same specificity. GLIPH determines 2-4mers continuous amino acid sequences (motifs) that significantly enrich in certain antigen

specific TCR group by comparing to naive TCRs group. Candidate motifs are extracted from the CDR3 region of TCR beta chain.

Here, we are reporting another tool, named LiMPETs, which has similar function of selecting antigen specific motifs from a group of antigen specific TCRs. Same as GLIPH, we focus only on TCR beta chain CDR3, since this is the most important specificity determine region (most antigen-contact), the most variable region[17] and so with the most data available[117]. LiMPETs forms the process of selecting antigen specific motifs as a least absolute deviation (LAD) problem in multiple linear regression, and solves the optimization problem with Mixed-Integer Linear Programming (see methods). LAD is well known for its robustness to outliers and easy to interpret[119], and find its applications in bioinformatics for variable (feature) selection[120,121]. Here, we implement LAD to select antigen specific motifs and shows LiMPETs is more accurate and is more reliable than GLIPH. Also, compared to GLIPH, LiMPETs can easily be generalized to multiple specificity groups and can take TCR cross-reactivity (i.e., same TCR recognize multiple antigens or same antigen recognized by multiple TCRs) into consideration. Also, LiMPETs is developed under the popular R circumstance as a package, so it’s more easily to be used.

4.2 MATERIALS AND METHODS

4.2.1 Problem formulation

Given a set of TCRs with known antigen specificity, LiMPETs tries to identify motifs enriched in the TCR set by comparing with another set of naive TCRs. This process of selecting motifs is formed as solving the following LAD problem with linear programming method.

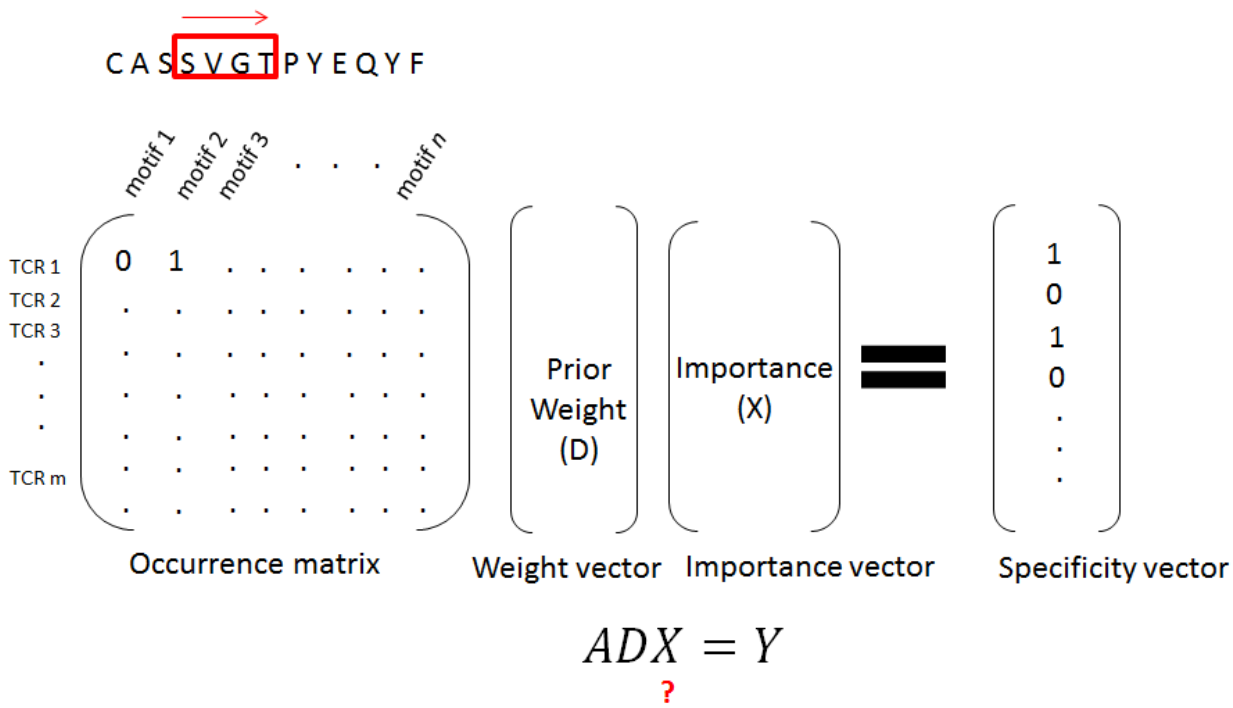


Figure 4.1: Schematic of LiMPETs model

As shown in Figure 4.1, our linear model contains 4 matrices/vectors:

a). **Occurrence matrix(A):** A matrix records whether one motif exists in the CDR3 sequence of a TCR or not. 1 means the motif can be found in the TCR, while 0 means not;

b). **Weight vector(W)/diagonal weight matrix(D):** A vector (W) given prior weights for all motifs, 'W' is user defined, it can either be all-ones vector(i.e., no prior bias on the motifs), or it can be other properties of the motif sequences (e.g., contact probability of the amino acids[122],

etc.). In this work, we defined a weight vector based on the frequency of motifs within the naive TCR dataset, which will be described later. The diagonal weight matrix(D) is a diagonal matrix with all the elements on the diagonal equal to W, i.e., D=diag(X).

c). **Specificity vector(Y):** A vector records whether one TCR is from antigen specificity TCR group or from naive TCR group. 1 means the TCR is from antigen specificity TCR group, while 0 means from naive TCR group;

d). **Importance vector(X):** A vector needs to be solved by linear programming algorithm. The values represent how significant one motif is enriched in the antigen specificity TCR group;

Given 'm' sequences, and 'n' motifs, the linear model is:

$$ADX = Y \tag{1}$$

,where A is a mxn binary matrix, D is a nxn diagonal matrix, X and Y are nx1 vectors.

Given the CDR3 sequence of a TCR, a window of 3 or 4 amino acids (AAs) slides across the CDR3 sequence, and the occurrence of motifs within the TCR are recorded in matrix A, while its antigen specificity is recorded in the vector Y.

After encoded the motif occurrence and antigen specificity, LiMPETs solves for 'X' by minimizing:

$$minimize\ |ADX - Y|_1 = minimize \sum_{i=1}^{m} \left| \left(\sum_{j=1}^{n} (AD)_{ij} X_j \right) - Y_i \right|, \qquad s.t. \quad X_j \geq 0$$

(2)

The advantage of minimizing on the L1-norm is its robustness to outliers [123] and sparsity of residuals after optimization [124]. We need the robustness because our control set is from naive TCRs instead of real true negative TCRs. The sparsity of residuals will prompt lots of insignificant residuals (i.e residuals close to 0), through which the values in 'X' will be either 0 or larger than 1.

To solve the optimization problem in Equation (2) is equivalent to solve:

$$minimize \sum_{i=1}^{m} \varepsilon_i , \qquad s.t.\ X_j \geq 0, \varepsilon_i \geq 0, \left(\sum_{j=1}^{n} (AD)_{ij} X_j\right) - Y_i \leq \varepsilon_i, \left(\sum_{j=1}^{n} (AD)_{ij} X_j\right) - Y_i \geq -\varepsilon_i$$

(3)

4.2.2 Weight vector generation

The weight vector can be any prior information for each motif, in this work we designed a weight vector by analyzing the naive CD8+ TCR dataset and adopted 'markov chain' [125] to calculate the prior weight for each motif:

1). The 'initial state probability' (π) of the markov chain was calculated by counting the frequency of each amino acid within the CDR3 region, which is a vector of 20 elements. The 'transition matrix'(T) was also calculated by counting the frequency of each possible transition between two adjacent amino acids in naive CD8+ CDR3 region, which is a 20x20 matrix.

2). For each motif S, a probability associated can be calculated using the 'initial state probability' (π) and 'transition matrix'(T):

$$P_S = \pi_{S_1} * \prod_{k=2}^{n} T_{(S_k, S_{k+1})} \qquad (4)$$

,where S_1, S_k and S_{k+1} are respectively the 1st, k^{th} and $(k+1)^{th}$ amino acid of motif S, n is the length of S. P_s represents the probability of generate given motif 'S' by random.

3). The probability for all motifs of different lengths can be calculated using equation (4). In order to normalize and compare motifs of different lengths, we further normalize the probability with the motif length:

$$P_N = 1 - P_s^{\frac{1}{n}} \tag{5}$$

,where P_N is the normalized weight for given motif 'S', while n is the length of S.

4.2.3 Implementation

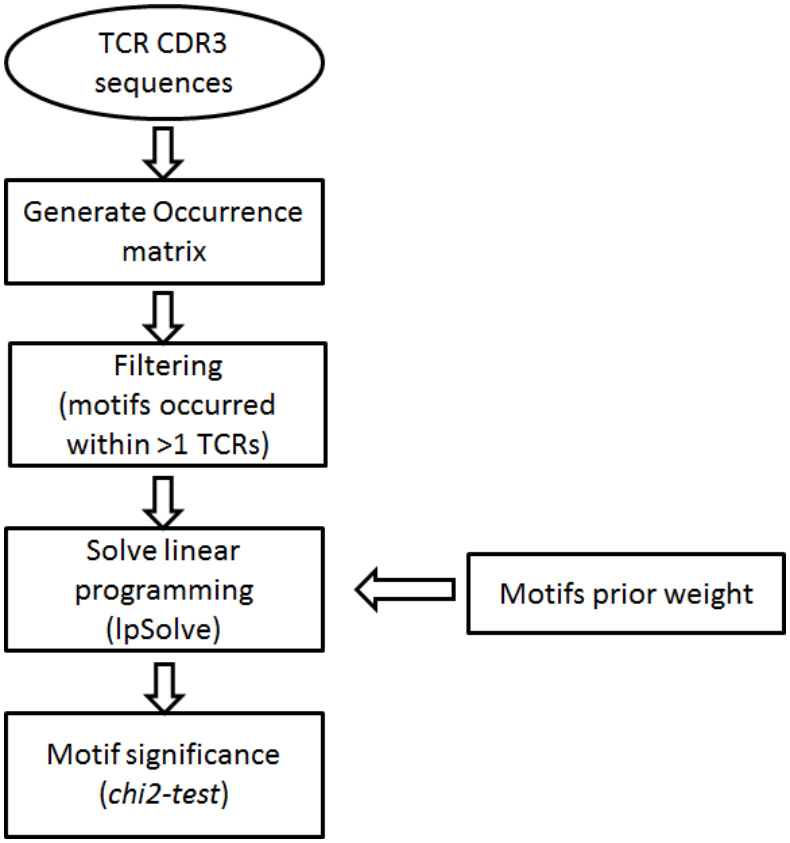


Figure 4.2: Workflow of LiMPETs

Figure 4.2 shows the workflow of LiMPETs, which contains four steps:

a). Given a set of TCR CDR3 sequences, generate the 'Occurrence Matrix' by record whether one motif exist within a TCR; generate 'Specificity Vector' by record whether this TCR is antigen specific or from the naive control group;

b). To accelerate the following steps, LiMPETs filters out motifs exist within only one TCR;

c). Solve the linear programming problem in Equation (3) with the Open Source Mixed-Integer Linear Programming System lp_solve [126];

d). To evaluate the associated significance (p-value) of each motif, LiMPETs adopts chi-square test, Benjamini-Hochberg adjustment (BH) [127] was used to adjust p-value for multiple testing. In this work, motifs with p-values smaller than 0.01 were selected as significant motifs.

The resulting motifs and associated p-value can be utilized in different ways, e.g., it can be used to identify antigen specificity groups from T cell receptor repertoire as GLIPH does; it may also be used to predict the specificity of TCRs by evaluating whether these TCR contain any conserved motifs with known specificity.

4.3 RESULTS

4.3.1 Determine an appropriate sample size for naive TCR dataset

In order to tune and test our model, we designed a dataset (Dataset 1) contains TCRs specific to one of the 3 antigen shown in Table 4.1 by gathering data from Glanville *et al.*[17], Dash *et al.*[128]. Since there are siginficant fewer TCRs for pp65, we also included a random sample of pp65-specific TCRs within VDJdb(downloaded on 02/02/2018)[117] to even the data size. Only TCR-beta sequences have been used for the testing. We chose these 3 antigens because they were associated with the most TCRs across our data sources. We also included the naive CD8+ TCRs in Glanville *et al.*[17] as negative control dataset.

Table 4.1: TCRs contained in Dataset 1

Antigen	Peptide	Species	HLA	Number of unique CDR3s
BMLF1	GLCTLVAML	Epstein-Barr virus(EBV)	HLA-A*02	748
M1	GILGFVFTL	Influenza(FLU)	HLA-A*02	697
pp65	NLVPMVATV	Cytomegalovirus(CMV)	HLA-A*02	682
Naive	N/A	N/A	N/A	27,845

Feature(variable) selection on imbalanced dataset is a challenge problem in data mining research[129], but very common in real medical applications[130], where normally negative controls overwhelmed the whole dataset. One epidemiological case-control study propose a ratio of 4:1 between control versus cases[131]. Here, we used 10-fold cross-validation to determine an appropriate control dataset size. The results are shown in Figure 4.3. We chose the negative dataset to be 50 folds of the positive dataset, if the number of desired controls is more than the naive TCR set, then all the naive TCRs were used as negative dataset.

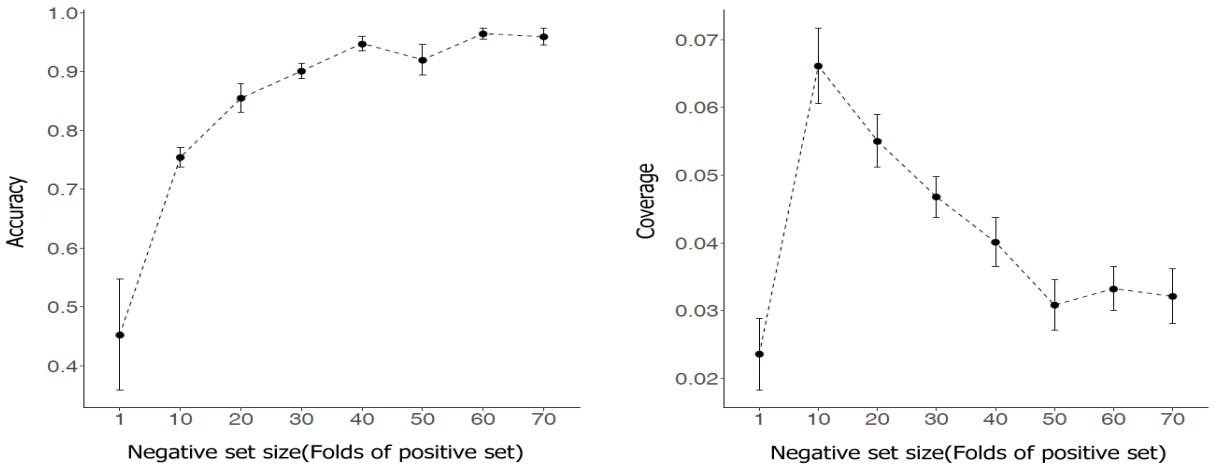


Figure 4.3: Coverage and Accuracy of LiMPETs with different negative control dataset size. With the increase of control (naive) dataset size, the accuracy of LiMPETs prediction reached to plateau.

4.3.2 LiMPETs is more robust than GLIPH

Although more and more measurement have been performed and TCRs with known antigen specificity have been gathered, due to the low throughput of current experiment and the huge diversity of TCR repertoire, TCRs with known specificity are still very limited and imbalanced (https://vdjdb.cdr3.net/). In order to make use of current limited dataset, a tool with robust performance on small dataset is essential. To test the robustness of LiMPETs and GLIPH, we gradually down sampled Dataset 1 from 90% to 10% as training dataset, while use the remaining TCRs as test dataset. Motifs and their specificity identified from training dataset can be used to predict the specificity in the test dataset (i.e., if one sequence in test set contains certain motif, then its specificity will be predicted to be the same as the motif). Coverage (percentage of TCRs predicted) and accuracy (the percentage of correct predictions) of the specificity prediction can be compared between the two methods.

Figure 4.4 shows the testing results of LiMPETs versus GLIPH on different sub-sample size. Overall, LiMPETs performs more accurate than GLIPH, but GLIPH's accuracy drops

significantly when the dataset size becomes small (~200 total TCRs and ~70 for each antigen specificity group). GLIPH identifies significant motifs by repeat sampling from naive dataset of the same number of TCRs in the 'antigen specific group', and a p-value will be calculated by counting the frequency of samples with certain motif more enriched/frequent than the 'antigen specific group'. This numerical way is essentially trying to estimate the frequency of given motif in the naive TCR dataset and calculated the p-value based on the sampling distribution. However, when the sample size becomes smaller, the variance of the sampling distribution will become larger, which means the numerical p-value will become less accurate in small samples. We also compared LiMPETs' performance with/without motif weight, adding the weight did not significantly increase accuracy, but increased coverage of LiMPETs on small dataset.

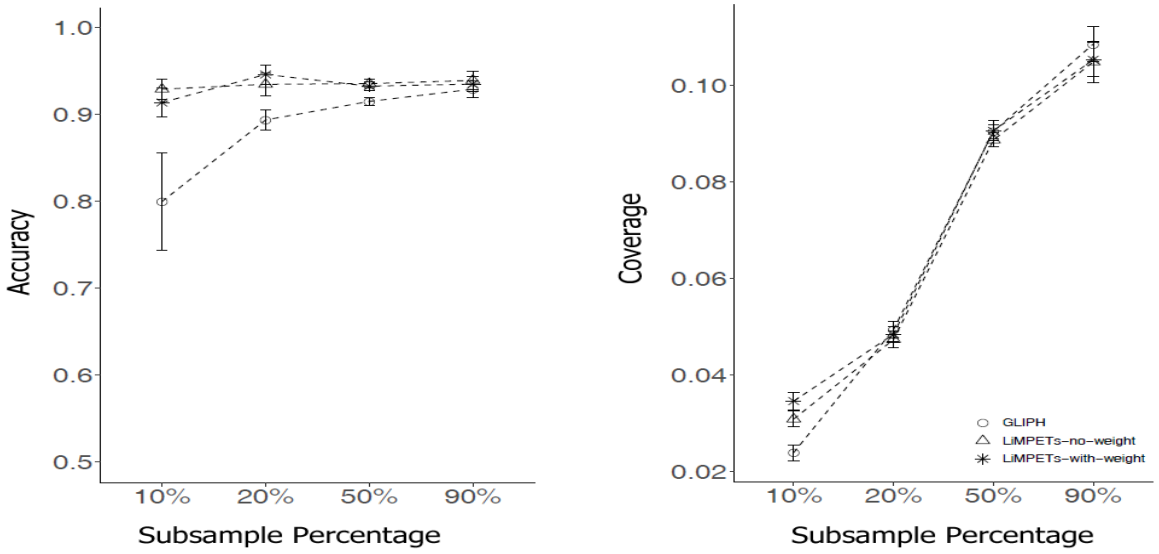


Figure 4.4: Comparison of LiMPETs versus GLIPH on different sample size. LiMPETs is more robust to sample size compared to GLIPH.

4.3.3 LiMPETs is more generalizable than GLIPH

GLIPH identify significant motifs by comparing the frequency of one motif with 1,000 random sampled sets from naive TCR dataset. The way of doing this has an intrinsic drawback when given one set of similar sequences (Figure 4.5), multiple significant motifs can be identified from the same sequence. However, the crystal structure analysis show the 'contact residues with antigen peptide are typically three to four amino acids in length and usually contiguous' [17], so the identified significant motifs cannot all be true (false positive). The design of LiMPETs enforced it to identify only one significant motif for each sequence, which is more stringent than GLIPH and will output fewer motifs, but we argue the identified significant motifs should be more generalizable to novel TCRs. GLIPH will output more motifs but contain more false positive ones, and it's highly possible those false positive motifs will cover TCRs of other antigen specificity.

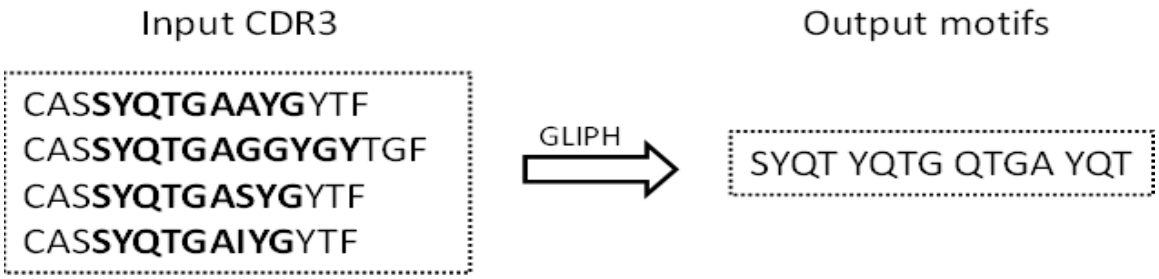


Figure 4.5: Given one set of CDR3 AA sequences, GLIPH may identify multiple significant motifs from the same sequence.

To test whether our expectation is true, we applied both methods on the TCRs collected within VDJdb. We only focus on human CD8+ TCR beta-chain CDR3 and we require each set of certain antigen specific group to contain at least 100 unique TCRs. Based on this criteria, we selected 14 TCR groups with various antigen specificity and MHC restriction (Dataset 2, Table

S8). Each 'antigen-specific group' can further be divided into sub-groups based on their source (i.e., original publications).

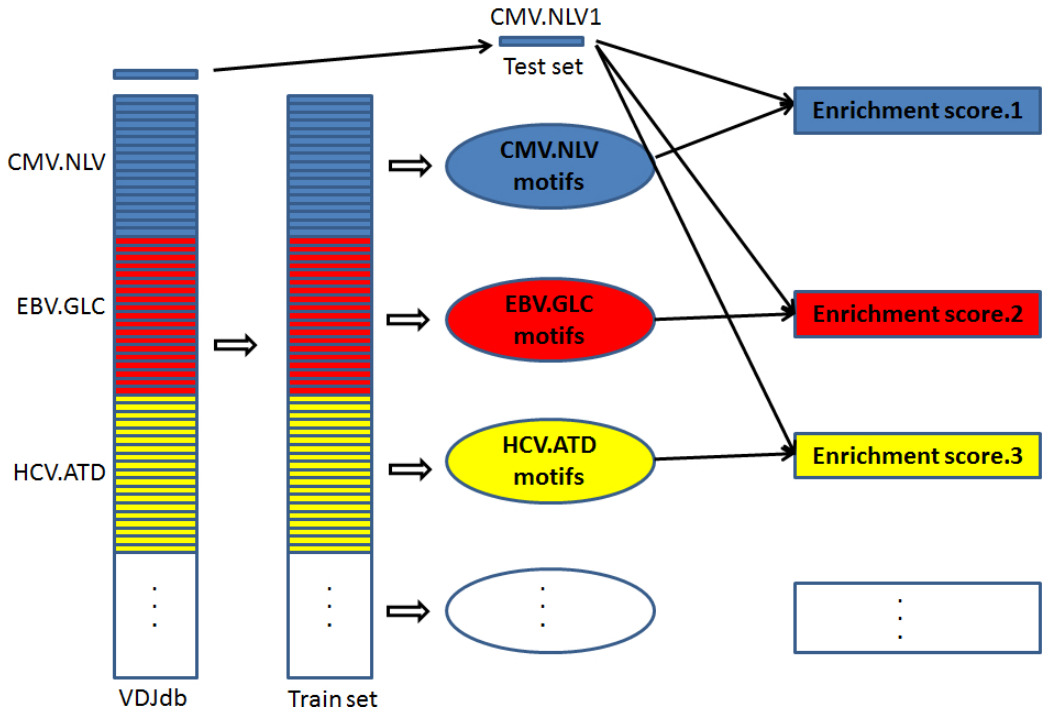


Figure 4.6: Workflow of testing on VDJdb database. Each time a subset of TCRs with the same specificity from one publication is retained as test dataset and the other TCRs were used as training set (i.e., 'leave one publication out' test).

To test both methods, we repeated the following steps (Figure 4.6):

1). For each sub-group (TCRs originated from the same publication), if it contains >100 TCR sequences, then it will be removed from Dataset 2, the remaining dataset will be used as 'training dataset', while the removed set will be retained as 'test dataset';

2). If the training dataset still has >100 TCR sequences for each antigen-specific group, then we apply both methods (LiMPETs and GLIPH) on each antigen-specific group to identify significant motifs, and use the motifs identified to test on the 'test set'.

In total, we identified 9 test groups, 8 of which are HLA-A*02 restricted (Table S9). Since for the same group within the training dataset, LiMPETs and GLIPH will identify different number of significant motifs. Also the number of sequences within each test dataset varies a lot. In order to compare the enrichment of the two sets of motifs within eac test dataset, we defined an 'enrichment score' calculated as a numerical p-value: repeatedly sample the same number of motifs from the naive TCRs (1,000 times), and count the frequency (p-value) that the sampled motif sets covered more sequences than the identified 'significant motif sets'. Then the enrichment score is calculated as:

$$S = -log_{10}P \tag{6}$$

, where P is the numerical p-value, and S is the designed Enrichment score.

Figure 4.7 shows the visualization of enrichment score for each test dataset across all the antigen specificity groups within the training datasets. Clearly, motif set identified by GLIPH for certain antigen-specific group are much more cross enriched to other antigen specific group.

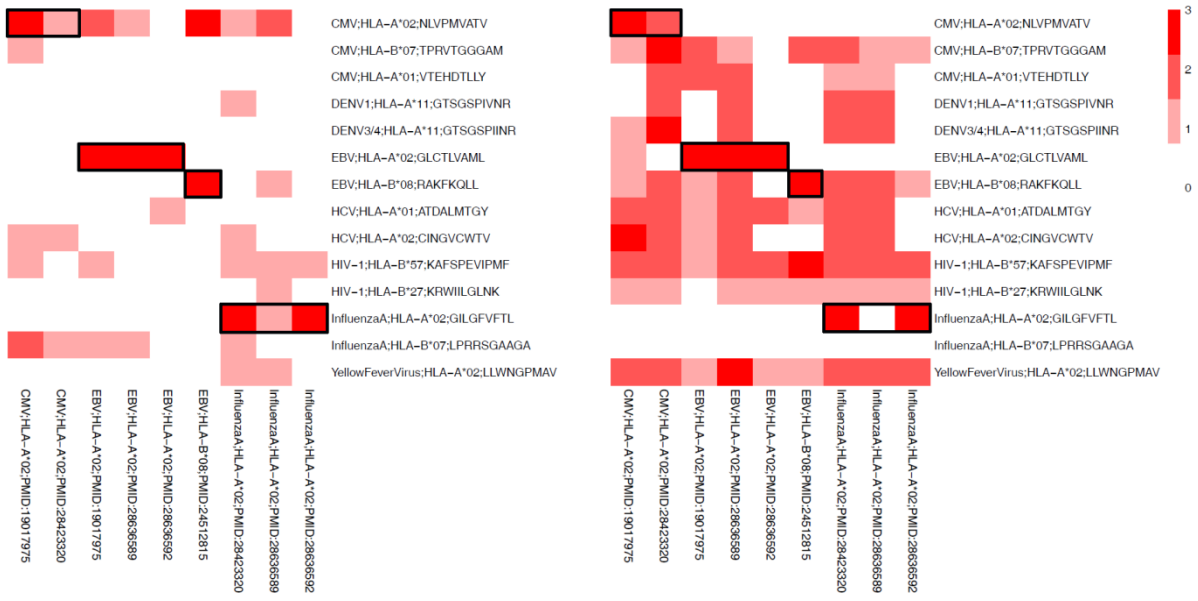


Figure 4.7: Heatmap visualization of enrichment score. The black boxes represent true antigen specificity, which means the row and column are from the same antigen specificity group and restricted by the same MHC.

If we treat this as a multi-class classification problem, then the ROC curves for the cross-validation test described above are shown in Figure 4.8. Strikingly, both methods perform quite well, with Area Under Curve (AUC) higher than 0.9, LiMPETs shows even higher classification power than GLIPH.

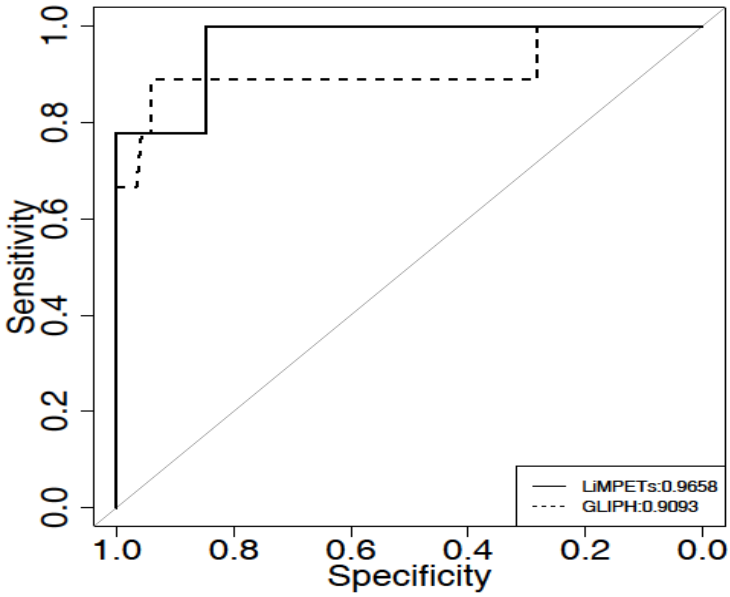


Figure 4.8: ROC curve for LiMPETs and GLIPH on Dataset 2. Legend shows the Area under curve (AUC) of both methods.

4.3.4 LiMPETs eliminates more false positive than GLIPH

The test results across VDJdb shows that LiMPETs has less false positive compared to GLIPH, however, our ROC curve analysis assume there is no cross-reactivity between different antigen specific TCR groups. It's hard to control TCR cross reactivity in Dataset 2, because VDJdb is a curated database gathering data from previous publications, which means we can only know certain TCR will bind to certain peptide, but we do not know whether this TCR will or will not bind to other peptides.

To have a better controlled result, we experimentally validated a set of TCRs (Dataset 3) which are HLA-A2 restricted and specific to M1 peptide (GILGFVFTL) but not specific to HCV peptide(CINGVCWTV) (data not published). This set of TCRs contains 438 unique CDR3 sequences, within which we counted how many TCRs contain significant motifs identified from Dataset 2 for M1 and HCV. Table 4.2 summarizes the performance of LiMPETs versus GLIPH,

both methods cover similar number of M1 specific TCRs within Dataset 3, but 7 TCRs also contain HCV specific motifs identified by GLIPH while zero TCRs contained LiMPETs HCV motifs. Although the number does not vary a lot, considering our sampling space is whole human TCR repertoire, this difference means a lot. The associated numerical p-value has also been calculated as described above (Equation 6), HCV specific motifs identified by GLIPH is almost significant enriched in our experiment validated M1 specific TCRs. The numerically calculated p value is 0.056, which means only 56 out of 1000 times that a random set of motifs will cover more than 7 sequences within Dataset 3 M1 specific TCRs. Another point worth to mention is, LiMPETs identified less than one half of significant motifs compared to GLIPH (Table 4.3) but the coverage are similar, which proves LiMPETs is more stringent but the significant motifs identified are more generalizable.

Table 4.2 Number of TCRs in Dataset 3 contain significant motifs identified from Dataset 2 (with numerical p-value).

Number of TCRs in Dataset 3	M1	HCV
LiMPETs	214(0.005)	0(1)
GLIPH	213(0.009)	7(0.056)

Table 4.3 Number of significant motifs identified from Dataset 2 by both methods

Number of Motifs	M1	HCV
LiMPETs	55	14
GLIPH	128	11

4.4 DISCUSSION

Computationally predicting TCR antigen specificity is attractive given its wide application in basic research and clinical usage (e.g., cancer immunotherapy). Traditional way of computational modeling for protein-protein interaction is not accurate enough, due to the flexibility of the loop contact region [115], the lacking of homology modeling templates and the conformation change before/after TCR bind to pMHC [116]. Also, the high computational cost limits its usage in high-throughput applications. Machine learning (ML) algorithms are powerful and potentially can be used for TCR specificity prediction. However, the fundamental step of applying ML methods is feature engineering, which is to find an appropriate feature set for the algorithms. In this work, we adopted a linear model which can automatically select features during optimization. Tests of LiMPETs on both public and in-house data show its advantages compared to current existing method.

We designed the penalty function as L1-norm format instead of L2-norm. The L2-norm format is well known as least square regression, which is well studied and has explicit solution. Within an overdetermined system (i.e., number of observations > number of features, as in our model), methods (e.g., Singular Value Decomposition based[132]) have been developed to approximating the least square solution. However, we decided on choosing L1-norm because its robustness to outliers (erroneous measurement) while L2-norm's solutions are sensitive to outliers. Figure 4.9 shows the comparison of accuracy between using L1-norm versus L2-norm. Also, the sparsity of L1-norm penalty function will prompt large number of residuals close to 0, which can be used for automatic motif selection. The least square regression was solved by Singular Value Decomposition (SVD), and the most weighted motifs (same number of motifs as LiMPETs identified) were selected as significant motifs.

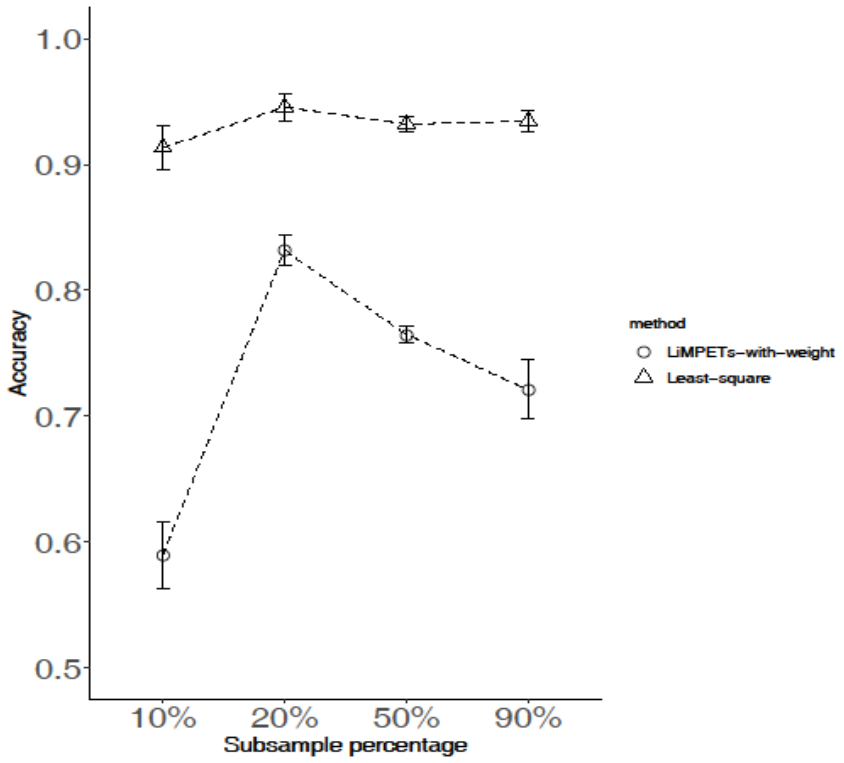


Figure 4.9: LiMPETs accuracy compared with Least square regression. X-axis represents different input sample size from Dataset 1.

Besides the accuracy and robustness of LiMPETs, the framework we proposed here can be easily generalized: 1). By introducing dummy variable, the 'Specificity vector' term can be a binary matrix instead of a vector, through which TCR cross-reactivity (i.e., same TCR recognize multiple antigens) information can be included for motif selection; 2). Current model encode the 'Occurrence Matrix' as binary matrix, although prior information can be incorporated by the 'Weight vector', continuous variables may perform better; 3). Current model does not consider interactions between motifs, cross terms can be added in the linear model in order to generalize it.

LiMPETs adopted linear programming solver to solve for the 'Importance vector', which requires significant amount of computational memory to manipulating matrices, however since our system is sparse, solvers/algorithms[133] more efficient for sparse matrices should perform faster. In this work, we used chi-square test at the final step to calculate an associated p-value,

however chi-square test is not reliable for low-frequency terms [134], possible optimization can be performed.

Chapter 5. Conclusions and Future studies

We developed MIDCIRS computational pipeline for correctly measuring immune repertoire, together with wet lab experimental design, we showed MIDCIRS' high accuracy, high coverage and wide dynamic range compared with other tool/pipeline. We detailed analyzed MIDCIRS performance measured with various metrics to prove MIDCIRS advantages. However, as we demonstrated in Chapter 2, the efficiency of current MIDCIRS is only ~30% and this is mainly due to RT/PCR efficiency drops, design of more efficient primers and experimental procedures will help to improve MIDCIRS' efficiency.

Having demonstrated the benefits of MIDCIRS, we applied MIDCIRS to real world problems to study antigen driven immune response. We used MIDCIRS to measure the antibody repertoire from malaria-experienced individuals and found unexpected mutable capability of baby adaptive immune system by incorporated analysis of antibody somatic hypermutation and antibody clonal lineage structures. We also used MIDCIRS to measure Follicular helper T cells (Tfhs) directly obtained from untreated HIV patients' lymph nodes and found evidence for intact antigen-driven clonal expansion of Tfh cells and selective utilization of specific complementarity-determining region 3 (CDR3) motifs during chronic HIV infection. Both studies demonstrated MIDCIRS is useful for studying antigen driven immune response. MIDCIRS is potentially to be used in other repertoire analysis to help immunologists understand various types of immune response.

We developed another tool, named LiMPETs, to find significant motifs within TCR CDR3 region for different antigen specificity. The goal of LiMPETs is to find set of motifs significant enriched within the sequence of certain antigen specific TCR group and we demonstrated these motifs are generalizable to novel TCRs. We also demonstrated LiMPETs'

advantage by comparing with existing tool on both public and in-house data. Due to the available data right now, we only focused on CD8+ TCR beta chain CDR3 region. Although previous research demonstrated this region is the most important, determining of TCR antigen specificity includes all the structures at the TCR-pMHC interface (i.e., CDR1a, CDR2a, CDR3a, CDR1b, CDR2b, CDR3b, peptide and helix regions of MHC), by considering only a small fraction at the interface limits the accuracy and coverage. One reason that we only focus on beta chain is because current accumulated data are most beta chain, with no information about its paired alpha chains. However, people are developing methods which can measure TCR alpha and beta chain simultaneously with its antigen specificity. With more and more paired-data available, we are hoping to extend the framework and increase LiMPETs' coverage. Also, current version of LiMPETs only considers binary variables, by relaxing them to continuous variables (e.g., adding Amino Acids properties) will potentially increase LiMPETs' performance.

Appendix

Table S1. Metrics of sequencing results with different RNA input

Sample	Raw reads	Mappable reads	Map percentage (%)	Total RNA molecules	Unique productive CDR3	Percentage of MIDs with sub-clusters (%)	Percentage of chimera sequences (%)	Top CDR3 molecules *	Top CDR3 molecule fraction (%)
20,000Tn _10%RNA	402975	254228	63.09	10171	4579	0.11	0.32	24	0.24
20,000Tn _30%RNA	877556	698961	79.65	18670	7253	0.34	0.42	39	0.21
20,000Tn _50%RNA	1188083	984951	82.90	18367	7495	0.32	0.70	30	0.16
100,000Tn _10%RNA	922615	766441	83.07	36949	17632	0.28	0.33	89	0.24
100,000Tn _30%RNA	2409732	2173270	90.19	72257	30428	0.70	1.58	245	0.34
100,000Tn _50%RNA	1744861	1566048	89.75	55058	27280	0.52	0.99	171	0.31
200,000Tn _10%RNA	1000937	788947	78.82	61525	34097	0.41	0.86	166	0.27
200,000Tn _30%RNA	4224183	3902130	92.38	173224	66990	1.57	5.44	498	0.29
200,000Tn _50%RNA	3147293	2889513	91.81	154666	67607	1.28	2.64	628	0.41
1,000,000Tn _10%RNA	7695858	6975703	90.64	514916	237331	3.19	16.14	1430	0.28
1,000,000Tn _30%RNA	9439612	8719649	92.37	942010	382743	5.18	17.02	2387	0.25
1,000,000Tn _50%RNA	17021339	15979187	93.88	1606258	487295	8.52	47.45	4468	0.28

- Top CDR3: CDR3 with highest MID.

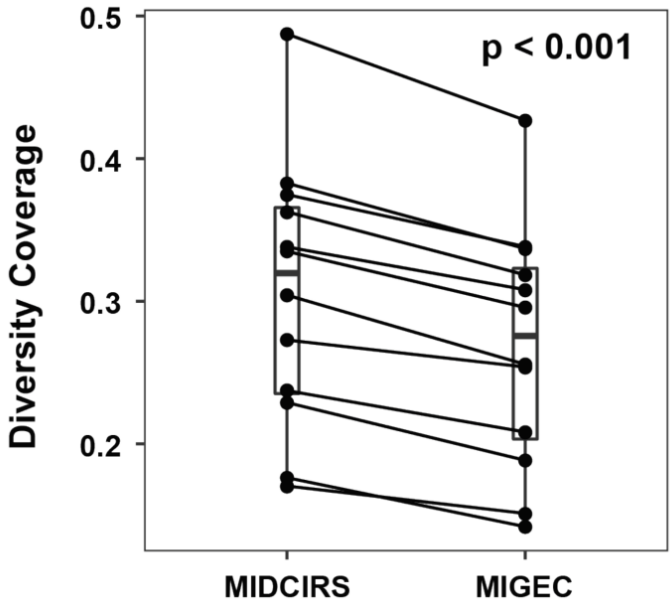


Figure S1: Comparison of diversity coverage between MIDCIRS and MIGEC pipelines on the same set of data presented in this study. P-value was determined by paired Wilcoxon test.

Table S2: TCR repertoire sequencing cell and transcript counts.

	Naive		Memory		GC Tfh	
Subject	Cells	TCR transcripts	Cells	TCR transcripts	Cells	TCR transcripts
H2	10,000	14,420	10,000	6,315	15,000	33,904
H3	10,000	6,750	10,000	8,945	10,000	7,954
H8	10,000	13,225	10,000	5,992	10,000	26,374
H10	10,000	4,314	10,000	9,033	1,464	2,498
H14	10,000	5,822	10,000	6,254	10,000	11,197
H17	10,000	18,376	10,000	6,533	10,995	37,129
H21	10,000	10,404	10,000	7,503	10,227	23,673
H24	10,000	5,488	10,000	2,821	10,000	9,220

Table S3: TCRβ Sequencing Primers. Red Ns indicate 12N random molecular identified (MID). Blue Ns indicate fixed Illumina i7 indexes used for pooling multiple libraries for a single run.

Primer Target	Sequence	PCR Step
TCRb Constant	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN NNN NNN GAC CTC GGG TGG GAA CAC	Reverse Transcription
TRBV1	GAC GTG TGC TCT TCC GAT CTC TGA CAG CTC TCG CTT ATA CCT TCA	1st PCR, forward
TRBV2	GAC GTG TGC TCT TCC GAT CTG CCT GAT GGA TCA AAT TTC ACT CTG	
TRBV3	GAC GTG TGC TCT TCC GAT CTA ATG AAA CAG TTC CAA ATC GMT TCT	
TRBV4	GAC GTG TGC TCT TCC GAT CTC CAA GTC GCT TCT CAC CTG AAT	
TRBV5-1	GAC GTG TGC TCT TCC GAT CTC GCC AGT TCT CTA ACT CTC GCT CT	
TRBV5-2	GAC GTG TGC TCT TCC GAT CTT TAC TGA GTC AAA CAC GGA GCT AGG	
TRBV5-3	GAC GTG TGC TCT TCC GAT CTC TCT GAG ATG AAT GTG AGT GCC TTG	
TRBV5-4/5/6/7/8	GAC GTG TGC TCT TCC GAT CTC TGA GCT GAA TGT GAA CGC CTT G	
TRBV6-1	GAC GTG TGC TCT TCC GAT CTT CTC CAG ATT AAA CAA ACG GGA GTT	
TRBV6-2/3	GAC GTG TGC TCT TCC GAT CTC TGA TGG CTA CAA TGT CTC CAG ATT	
TRBV6-4	GAC GTG TGC TCT TCC GAT CTA GTG TCT CCA GAG CAA ACA CAG ATG	
TRBV6-5/6/7	GAC GTG TGC TCT TCC GAT CTG TCT CCA GAT CAA MCA CAG AGG ATT	
TRBV6-8/9	GAC GTG TGC TCT TCC GAT CTA AAC ACA GAG GAT TTC CCR CTC AG	
TRBV7-1	GAC GTG TGC TCT TCC GAT CTG TCT GAG GGA TCC ATC TCC ACT C	
TRBV7-2	GAC GTG TGC TCT TCC GAT CTT CGC TTC TCT GCA GAG AGG ACT GG	
TRBV7-3	GAC GTG TGC TCT TCC GAT CTC TGA GGG ATC CGT CTC TAC TCT GAA	
TRBV7-4/8	GAC GTG TGC TCT TCC GAT CTC TGA GRG ATC CGT CTC CAC TCT G	
TRBV7-5	GAC GTG TGC TCT TCC GAT CTG GTC TGA GGA TCT TTC TCC ACC T	
TRBV7-6/7	GAC GTG TGC TCT TCC GAT CTG AGG GAT CCA TCT CCA CTC TGA C	
TRBV7-9	GAC GTG TGC TCT TCC GAT CTC TGC AGA GAG GCC TAA GGG ATC T	
TRBV8-1	GAC GTG TGC TCT TCC GAT CTA AGC TCA AGC ATT TTC CCT CAA C	
TRBV8-2	GAC GTG TGC TCT TCC GAT CTA TGT CAC AGA GGG GTA CTG TGT TTC	
TRBV9	GAC GTG TGC TCT TCC GAT CTA CAG TTC CCT GAC TTG CAC TCT G	
TRBV10-1/3	GAC GTG TGC TCT TCC GAT CTA CAA AGG AGA AGT CTC AGA TGG CTA	
TRBV10-2	GAC GTG TGC TCT TCC GAT CTT GTC TCC AGA TCC AAG ACA GAG AA	
TRBV11	GAC GTG TGC TCT TCC GAT CTC TGC AGA GAG GCT CAA AGG AGT AG	
TRBV12-1/2	GAC GTG TGC TCT TCC GAT CTA TCA TTC TCY ACT CTG AGG ATC CAR	
TRVB12-3/4/5	GAC GTG TGC TCT TCC GAT CTA CTC TGA RGA TCC AGC CCT CAG AAC	
TRBV13	GAC GTG TGC TCT TCC GAT CTC AGC TCA ACA GTT CAG TGA CTA TCA T	
TRBV14	GAC GTG TGC TCT TCC GAT CTG AAA GGA CTG GAG GGA CGT ATT CTA	
TRBV15	GAC GTG TGC TCT TCC GAT CTG CCG AAC ACT TCT TTC TGC TTT CT	
TRBV16	GAC GTG TGC TCT TCC GAT CTA TTT TCA GCT AAG TGC CTC CCA AAT	
TRBV17	GAC GTG TGC TCT TCC GAT CTC ACA GCT GAA AGA CCT AAC GGA AC	
TRBV18	GAC GTG TGC TCT TCC GAT CTA TTT TCT GCT GAA TTT CCC AAA GAG	
TRBV19	GAC GTG TGC TCT TCC GAT CTG TCT CTC GGG AGA AGA AGG AAT C	
TRBV20-1	GAC GTG TGC TCT TCC GAT CTG ACA AGT TTC TCA TCA ACC ATG CAA	
TRBV21-1	GAC GTG TGC TCT TCC GAT CTC AAT GCT CCA AAA ACT CAT CCT GT	
TRBV22-1	GAC GTG TGC TCT TCC GAT CTA GGA GAA GGG GCT ATT TCT TCT CAG	
TRBV23-1	GAC GTG TGC TCT TCC GAT CTA TTC TCA TCT CAA TGC CCC AAG AAC	
TRBV24-1	GAC GTG TGC TCT TCC GAT CTG ACA GGC ACA GGC TAA ATT CTC C	
TRBV25-1	GAC GTG TGC TCT TCC GAT CTA GTC TCC AGA ATA AGG ACG GAG CAT	
TRBV26	GAC GTG TGC TCT TCC GAT CTC TCT GAG GGG TAT CAT GTT TCT TGA	
TRBV27	GAC GTG TGC TCT TCC GAT CTC AAA GTC TCT CGA AAA GAG AAG AGG A	
TRBV28	GAC GTG TGC TCT TCC GAT CTA AGA AGG AGC GCT TCT CCC TGA TT	
TRBV29-1	GAC GTG TGC TCT TCC GAT CTC GCC CAA ACC TAA CAT TCT CAA	
TRBV30	GAC GTG TGC TCT TCC GAT CTC CAG AAT CTC TCA GCC TCC AGA C	
ILLUPE1adaptor_short	ACA CTC TTT CCC TAC ACG AC	1st PCR, reverse
ILLUPE2adaptor_full	CAA GCA GAA GAC GGC ATA CGA GAT AA NNN NNN GTG ACT GGA GTT CAG ACG TG	2nd PCR, forward
ILLUPE1adaptor_full	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC	2nd PCR, reverse

Table S4: Gag and HA TCR sequence reference panel.

Subject	Specificity	V allele	J allele	CDR3	Phenotype
H2	Gag	TRBV4-3*01	TRBJ1-4*01	CASSQDTRVTNEKLFF	GC T_{FH}
H2	Gag	TRBV19*01	TRBJ2-3*01	CASSLAGGLDTQYF	GC T_{FH}
H2	Gag	TRBV20-1*01	TRBJ2-4*01	CSARDYARGGGGNIQYF	GC T_{FH}
H2	Gag	TRBV7-2*01	TRBJ1-1*01	CASSLAQYTEAFF	GC T_{FH}
H2	Gag	TRBV12-3*01	TRBJ1-2*01	CASSLGQGAAGYTF	GC T_{FH}
H2	Gag	TRBV4-3*01	TRBJ2-3*01	CASSQDQRAETDTQYF	GC T_{FH}
H2	Gag	TRBV4-3*01	TRBJ1-4*01	CASSQAMGDNEKLFF	GC T_{FH}
H2	Gag	TRBV27*01	TRBJ1-5*01	CASSLRRNQPQHF	Memory
H8	HA	TRBV20-1*01	TRBJ1-5*01	CSARRGADQPQHF	GC T_{FH}
H8	HA	TRBV4-3*01	TRBJ2-5*01	CASSQAGVPGTQYF	GC T_{FH}
H14	HA	TRBV18*01	TRBJ2-2*01	CASSPDRTATGELFF	Memory
H17	Gag	TRBV28*01	TRBJ1-4*01	CASSRIGQGGHEKLFF	GC T_{FH}
H17	HA	TRBV12-3*01	TRBJ2-2*01	CASSLNGVTGELFF	GC T_{FH}
H17	HA	TRBV5-4*01	TRBJ2-3*01	CASSLWTGGADTQYF	GC T_{FH}
H17	HA	TRBV7-3*01	TRBJ1-5*01	CAILEGGPGQPQHF	GC T_{FH}
H17	HA	TRBV30*01	TRBJ2-1*01	CAGRGPSGSNEQFF	GC T_{FH}
H17	HA	TRBV12-3*01	TRBJ1-5*01	CASSLGQGVGQPQHF	GC T_{FH}
H17	HA	TRBV20-1*01	TRBJ2-1*01	CSARGELAEQESYNEQFF	GC T_{FH}
H17	Gag	TRBV11-3*01	TRBJ2-1*01	CASSRPLANEQFF	GC T_{FH}
H17	Gag	TRBV28*01	TRBJ1-4*01	CASSRIGQGGHEKLFF	Memory
H17	HA	TRBV7-3*01	TRBJ1-5*01	CAILEGGPGQPQHF	Memory
H17	HA	TRBV7-9*01	TRBJ2-5*01	CASSLAGEETQYF	Naive
H21	Gag	TRBV5-4*01	TRBJ1-4*01	CASSPGEGLATNEKLFF	GC T_{FH}
H21	Gag	TRBV28*01	TRBJ1-5*01	CASSLSRDQPQHF	GC T_{FH}
H21	Gag	TRBV20-1*01	TRBJ2-7*01	CSARDGGVHEQYF	GC T_{FH}
H21	Gag	TRBV6-2*01	TRBJ1-5*01	CASTKEGQPQHL	GC T_{FH}
H21	Gag	TRBV28*01	TRBJ2-7*01	CASRPGQEAYEQYF	GC T_{FH}
H21	Gag	TRBV20-1*01	TRBJ2-2*01	CSARGLGRGIESGELFF	GC T_{FH}

H21	Gag	TRBV7-2*01	TRBJ2-3*01	CASSPEARGPRTDTQYF	GC T_{FH}
H21	HA	TRBV29-1*01	TRBJ1-2*01	CSVDRAGTNYGYTF	GC T_{FH}
H24	Gag	TRBV6-1*01	TRBJ1-5*01	CASSEARNRGLGQPQHF	GC T_{FH}
H24	Gag	TRBV20-1*01	TRBJ1-1*01	CSARDRGRATEAFF	GC T_{FH}
H24	HA	TRBV6-5*01	TRBJ1-2*01	CASSYSTGTGGGYTL	GC T_{FH}

Table S5: Cohort and cell type availability

Patient	Pre-malaria				Acute malaria		
	Pre-Index	Pre- Age	PBMC	Memory B	Acute-Index	Acute Age	PBMC
Inf1	Inf1-Pre3m	3m	Yes	I.S.	Inf1-Acu9m	9m	Yes
Inf2	Inf2-Pre3m	3m	Yes	J.F.	Inf2-Acu6m	6m	Yes
Inf3	Inf3-Pre5m	5m	Yes	I.S.	Inf3-Acu11m	11m	Yes
Inf4	Inf4-Pre5m	5m	Yes	J.F.	Inf4-Acu10m	10m	Yes
Inf5*	Inf5-Pre5m	5m	Yes	J.F.	Inf5-Acu10m	10m	Yes
Inf6	Inf6-Pre8m	8m	Yes	J.F.	Inf6-Acu12m	12m	Yes
Inf7	Inf7-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf8	Inf8-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf9	Inf9-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf10	Inf10-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf11	Inf11-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Tod1*	Tod1-Pre17m	17m	Yes	Yes	Tod1-Acu22m	22m	Yes
Tod2	Tod2-Pre19m	19m	Yes	Yes	Tod2-Acu22m	22m	Yes
Tod3†	Tod3-Pre28m	28m	Yes	Yes	Tod3-Acu32m	32m	Yes
Tod4	Tod4-Pre29m	29m	Yes	Yes	Tod4-Acu32m	32m	Yes
Tod5	Tod5-Pre31m	31m	Yes	J.F.	Tod5-Acu32m	32m	Yes
Tod6	Tod6-Pre31m	31m	Yes	Yes	Tod6-Acu38m	38m	Yes
Tod7†	Tod7-Pre40m	40m	Yes	Yes	Tod7-Acu42m	42m	Yes
Tod8	Tod8-Pre42m	42m	Yes	Yes	Tod8-Acu46m	46m	Yes
Tod9	Tod9-Pre47m	47m	Yes	Yes	Tod9-Acu50m	50m	Yes
Tod10	Tod10-Pre13m	13m	Yes	Yes	N.A.	N.A.	N.A.
Tod11	Tod11-Pre16m	16m	Yes	Yes	N.A.	N.A.	N.A.
Tod12	Tod12-Pre17m	17m	Yes	Yes	N.A.	N.A.	N.A.
Tod13	Tod13-Pre17m	17m	Yes	Yes	N.A.	N.A.	N.A.

I.S. indicates insufficient PBMC for FACS sorting or analysis.
J.F. indicates just flow cytometry analysis.
N.A indicates samples were not available.
* Same individual
† Same individual

Table S6: Sequencing read statistics of paired PBMCs from the malaria cohort

Sample	PBMCs[a]	Raw reads	Mapped reads	Percent Mapped	Unique RNA molecules
Inf1-Pre3m	3,000,000	3,246,180	2,989,252	92.1%	41,842
Inf1-Acu9m	3,000,000	3,608,436	3,348,589	92.8%	32,800
Inf2-Pre3m	3,000,000	3,176,623	2,987,587	94.0%	35,379
Inf2-Acu6m	3,000,000	3,689,115	3,481,675	94.4%	29,523
Inf3-Pre5m	4,150,000	3,242,619	3,070,458	94.7%	37,234
Inf3-Acu11m	5,000,000	4,396,739	4,153,830	94.5%	42,634
Inf4-Pre5m	5,000,000	3,048,762	2,810,018	92.2%	45,445
Inf4-Acu10m	3,700,000	5,287,767	4,864,629	92.0%	29,694
Inf5-Pre5m*	5,000,000	3,764,663	3,425,015	91.0%	54,516
Inf5-Acu10m*	50,00,000	4,712,120	4,374,600	92.8%	41,774
Inf6-Pre8m	5,000,000	3,588,177	3,456,165	96.3%	47,254
Inf6-Acu12m	400,000	395,765	378,182	95.6%	03,447
Tod1-Pre17m*	5,000,000	2,816,309	2,576,372	91.5%	53,551
Tod1-Acu22m*	1,380,000	2,811,617	2,593,849	92.3%	12,514
Tod2-Pre19m	5,000,000	4,842,338	4,673,875	96.5%	40,600
Tod2-Acu22m	1,920,000	1,956,906	1,886,521	96.4%	15,285
Tod3-Pre28m†	5,000,000	3,988,677	3,687,883	92.5%	35,567
Tod3-Acu32m†	5,000,000	9,218,255	8,565,149	92.9%	47,144
Tod4-Pre29m	5,000,000	2,924,629	2,851,964	97.5%	48,950
Tod4-Acu32m	5,000,000	4,004,416	3,846,197	96.0%	40,628
Tod5-Pre31m	5,000,000	5,338,867	5,126,888	96.0%	31,531
Tod5-Acu32m	3,000,000	2,853,984	2,736,902	95.9%	26,955
Tod6-Pre31m	5,000,000	4,356,975	4,198,929	96.4%	44,665
Tod6-Acu38m	2,170,000	5,738,001	5,460,964	95.2%	22,270
Tod7-Pre40m†	5,000,000	3,192,503	2,893,482	90.6%	34,901
Tod7-Acu42m†	4,740,000	4,448,008	4,079,432	91.7%	34,185
Tod8-Pre42m	5,000,000	2,120,127	2,058,164	97.1%	48,939
Tod8-Acu46m	2,100,000	2,060,234	1,986,239	96.4%	17,039
Tod9-Pre47m	3,000,000	3,035,618	2,682,991	88.4%	20,094
Tod9-Acu50m	3,000,000	4,678,879	3,912,981	83.6%	18,447

[a]Number of PBMCs differs because of the age dependent blood draw volume and cell recovery.
* Same individual
† Same individual

Table S7: Replacement and silent mutations and their ratio for PBMCs in infants and toddlers

			FWR			CDR			Average R/S Ratio	
			R	S	R/S Ratio	R	S	R/S Ratio	FWR	CDR
Infant	Pre	IgM	0.54	0.11	4.98	0.18	0.04	5.15	3.00 ± 1.12	5.54 ± 0.25
		IgG	1.54	0.70	2.21	1.36	0.24	5.67		
		IgA	1.48	0.65	2.28	1.29	0.22	5.75		
	Acute	IgM	1.36	0.34	4.05	0.58	0.11	5.52		
		IgG	1.88	0.85	2.22	1.62	0.30	5.35		
		IgA	2.03	0.90	2.25	1.75	0.30	5.79		
Toddler	Pre	IgM	1.12	0.35	3.20	0.58	0.11	5.54	2.41 ± 0.45	5.34 ± 0.25
		IgG	3.42	1.57	2.17	2.73	0.54	5.05		
		IgA	3.88	1.82	2.14	3.15	0.58	5.41		
	Acute	IgM	2.16	0.79	2.73	1.33	0.24	5.44		
		IgG	4.28	2.02	2.11	3.39	0.68	5.02		
		IgA	4.33	2.04	2.12	3.55	0.64	5.59		

Nucleotide mutations resulting in amino acid substitutions (Replacement, R) or no amino acid substitutions (silent, S) in the framework region (FWR2 and 3) and complementary determining regions (CDR1 and 2) of infants (N=6) and toddlers (N=9), weighted by unique RNA molecules. CDR3 and FWR4 were not included in this analysis due to the difficulty determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. Average displayed as mean ± standard deviation.

Table S8. Data selected for training in VDJdb

ID	host_species	antigen_species	TCRab	MHC	MHC.A	MHC.B	Epitope	number_of_unique_CDR3s
2	HomoSapiens	CMV	TRB	MHCI	HLA-A*02	B2M	NLVPMVATV	4223
3	HomoSapiens	CMV	TRB	MHCI	HLA-B*07	B2M	TPRVTGGGAM	156
4	HomoSapiens	CMV	TRB	MHCI	HLA-A*01	B2M	VTEHDTLLY	199
5	HomoSapiens	DENV1	TRB	MHCI	HLA-A*11	B2M	GTSGSPIVNR	165
6	HomoSapiens	DENV3/4	TRB	MHCI	HLA-A*11	B2M	GTSGSPIINR	158
8	HomoSapiens	EBV	TRB	MHCI	HLA-A*02	B2M	GLCTLVAML	846
9	HomoSapiens	EBV	TRB	MHCI	HLA-B*08	B2M	RAKFKQLL	172
12	HomoSapiens	HCV	TRB	MHCI	HLA-A*01	B2M	ATDALMTGY	165
13	HomoSapiens	HCV	TRB	MHCI	HLA-A*02	B2M	CINGVCWTV	124
15	HomoSapiens	HIV-1	TRB	MHCI	HLA-B*57	B2M	KAFSPEVIPMF	171
16	HomoSapiens	HIV-1	TRB	MHCI	HLA-B*27	B2M	KRWIILGLNK	280
25	HomoSapiens	InfluenzaA	TRB	MHCI	HLA-A*02	B2M	GILGFVFTL	2541
26	HomoSapiens	InfluenzaA	TRB	MHCI	HLA-B*07	B2M	LPRRSGAAGA	159
30	HomoSapiens	YellowFeverVirus	TRB	MHCI	HLA-A*02	B2M	LLWNGPMAV	284

Notes: data downloaded onz02/02/2018

Table S9. Data selected for testing within VDJdb

testID	host_species	antigen_species	TCRab	MHC	MHC.A	MHC.B	Epitope	Reference
1	HomoSapiens	CMV	TRB	MHCI	HLA\-A*02	B2M	NLVPMVATV	PMID:19017975
2	HomoSapiens	CMV	TRB	MHCI	HLA\-A*02	B2M	NLVPMVATV	PMID:28423320
4	HomoSapiens	EBV	TRB	MHCI	HLA\-A*02	B2M	GLCTLVAML	PMID:19017975
5	HomoSapiens	EBV	TRB	MHCI	HLA\-A*02	B2M	GLCTLVAML	PMID:28636589
6	HomoSapiens	EBV	TRB	MHCI	HLA\-A*02	B2M	GLCTLVAML	PMID:28636592
9	HomoSapiens	EBV	TRB	MHCI	HLA\-B*08	B2M	RAKFKQLL	PMID:24512815
12	HomoSapiens	InfluenzaA	TRB	MHCI	HLA\-A*02	B2M	GILGFVFTL	PMID:28423320
13	HomoSapiens	InfluenzaA	TRB	MHCI	HLA\-A*02	B2M	GILGFVFTL	PMID:28636589
14	HomoSapiens	InfluenzaA	TRB	MHCI	HLA\-A*02	B2M	GILGFVFTL	PMID:28636592

References

1. Yatim, K. M. & Lakkis, F. G. A brief journey through the immune system. *Clin. J. Am. Soc. Nephrol.* (2015). doi:10.2215/CJN.10031014

2. Cavallo, F., De Giovanni, C., Nanni, P., Forni, G. & Lollini, P. L. 2011: The immune hallmarks of cancer. in *Cancer Immunology, Immunotherapy* (2011). doi:10.1007/s00262-010-0968-0

3. Murphy, K. & Weaver, C. *Janeway's Immunbiology*. *Janeway's Immunbiology* (2017). doi:10.1007/s13398-014-0173-7.2

4. Riedel, S. Edward Jenner and the History of Smallpox and Vaccination. *Baylor Univ. Med. Cent. Proc.* (2005). doi:10.1080/08998280.2005.11928028

5. Carrat, F. & Flahault, A. Influenza vaccine: The challenge of antigenic drift. *Vaccine* (2007). doi:10.1016/j.vaccine.2007.07.027

6. Barouch, D. H. *et al.* Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* (2013). doi:10.1038/nature12744

7. Topalian, S. L., Drake, C. G. & Pardoll, D. M. Immune checkpoint blockade: A common denominator approach to cancer therapy. *Cancer Cell* (2015). doi:10.1016/j.ccell.2015.03.001

8. Weber, J. S. *et al.* Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): A randomised, controlled, open-label, phase 3 trial. *Lancet Oncol.* (2015). doi:10.1016/S1470-2045(15)70076-8

9. Abbas Abul K., Lichtman;, A. H. & Pillai, S. *Cellular and Molecular Immunology*. *Elsevier* (2014).

10. Alberts, B. Molecular Biology of the Cell. *Yale J. Biol. Med.* (2008). doi:10.1024/0301-1526.32.1.54

11. Janeway, C. J., Travers, P. & Walport, M. Immunobiology: The Immune System in Health and Disease. 5th edition. *Garl. Sci.* (2001). doi:10.1016/S0944-7113(96)80081-X

12. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* (2015). doi:10.1126/science.aaa4971

13. Golubovskaya, V. & Wu, L. Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy. *Cancers* (2016). doi:10.3390/cancers8030036

14. Weaver, C. T., Elson, C. O., Fouser, L. A. & Kolls, J. K. The Th17 Pathway and Inflammatory Diseases of the Intestines, Lungs, and Skin. *Annu. Rev. Pathol. Mech. Dis.* (2013). doi:10.1146/annurev-pathol-011110-130318

15. Crotty, S. T Follicular Helper Cell Differentiation, Function, and Roles in Disease. *Immunity* (2014). doi:10.1016/j.immuni.2014.10.004

16. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. *Nat. Immunol.* (2003). doi:10.1038/ni904

17. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017). doi:10.1038/nature22976

18. Chaudhary, N. & Wesemann, D. R. Analyzing immunoglobulin repertoires. *Frontiers in Immunology* (2018). doi:10.3389/fimmu.2018.00462

19. Calis, J. J. A. & Rosenberg, B. R. Characterizing immune repertoires by high throughput sequencing: Strategies and applications. *Trends in Immunology* (2014). doi:10.1016/j.it.2014.09.004

20. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* (2014). doi:10.1038/nbt.2782

21. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. *www.genome.gov/sequencingcostsdata* (2016).

22. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* (2012). doi:10.1038/nbt.2198

23. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (2012). doi:10.1101/gr.135350.111

24. Iba, Y., Hayashi, N., Sawada, J., Titani, K. & Kurosawa, Y. Changes in the specificity of antibodies against steroid antigens by introduction of mutations into complementarity-determining regions of the V(H) domain. *Protein Eng* **11,** 361–370 (1998).

25. McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol. Biol. Int.* (2014). doi:10.1155/2014/287430

26. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* (2014). doi:10.1038/nmeth.2960

27. Vollmers, C., Sit, R. V., Weinstein, J. A., Dekker, C. L. & Quake, S. R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1312146110

28. Robins, H. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology* (2013). doi:10.1016/j.coi.2013.09.017

29. Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science (80-.).* (2009). doi:10.1126/science.1170020

30. Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell

population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* (2015). doi:10.1038/ni.3175

31. Jiang, N. *et al.* Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci.* (2011). doi:10.1073/pnas.1014277108

32. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* (2013). doi:10.1126/scitranslmed.3004794

33. Bolotin, D. A. *et al.* Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* (2012). doi:10.1002/eji.201242517

34. Michaeli, M., Noga, H., Tabibian-Keissar, H., Barshack, I. & Mehr, R. it. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front. Immunol.* (2012). doi:10.3389/fimmu.2012.00386

35. Zhu, J. *et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1219320110

36. Vander Heiden, J. A. *et al.* PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu138

37. Khan, T. A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* (2016). doi:10.1126/sciadv.1501371

38. Ye, B., Smerin, D., Gao, Q., Kang, C. & Xiong, X. High-throughput sequencing of the immune repertoire in oncology: Applications for clinical diagnosis, monitoring, and immunotherapies. *Cancer Letters* (2018). doi:10.1016/j.canlet.2017.12.017

39. Briney, B., Le, K., Zhu, J. & Burton, D. R. Clonify: Unseeded antibody lineage assignment from next-generation sequencing data. *Sci. Rep.* (2016). doi:10.1038/srep23901

40. Shiao, Y.-H. A new reverse transcription-polymerase chain reaction method for accurate quantification. *BMC Biotechnol.* (2003). doi:10.1186/1472-6750-3-22

41. Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P. & Linnarsson, S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* (2013). doi:10.1371/journal.pone.0085270

42. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* (1966). doi:citeulike-article-id:311174

43. Wendel, B. S. *et al.* Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children. *Nat. Commun.* **8,** 531 (2017).

44. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science (80-.).* (2015). doi:10.1126/science.1258367

45. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci.* (2011). doi:10.1073/pnas.1017621108

46. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* (2014). doi:10.1038/nmeth.2772

47. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci.* (2012). doi:10.1073/pnas.1118018109

48. Fu, G. K., Wilhelmy, J., Stern, D., Fan, H. C. & Fodor, S. P. A. Digital encoding of cellular

mRNAs enabling precise and absolute gene expression measurement by single-molecule counting. *Anal. Chem.* (2014). doi:10.1021/ac500459p

49. Wu, Y. C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010). doi:10.1182/blood-2010-03-275859

50. Dekosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2492

51. Desponds, J., Mora, T. & Walczak, A. M. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1512977112

52. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* (2009). doi:10.1137/070710111

53. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* (2006). doi:10.3724/SP.J.1087.2009.02191

54. Yu, W. *et al.* Clonal Deletion Prunes but Does Not Eliminate Self-Specific αβ $CD8^+$ T Lymphocytes. *Immunity* (2015). doi:10.1016/j.immuni.2015.05.001

55. Ma, K. Y. *et al.* Immune repertoire sequencing using molecular identifiers enables accurate clonality discovery and clone size quantification. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.00033

56. Egorov, E. S. *et al.* Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *J. Immunol.* (2015). doi:10.4049/jimmunol.1500215

57. Wendel, B. S. *et al.* The receptor repertoire and functional profile of follicular T cells in HIV-infected lymph nodes. *Sci. Immunol.* (2018). doi:10.1126/sciimmunol.aan8884

58. Vinuesa, C. G., Linterman, M. A., Yu, D. & MacLennan, I. C. M. Follicular Helper T Cells. *Annu. Rev. Immunol.* (2016). doi:10.1146/annurev-immunol-041015-055605

59. Perreau, M. *et al.* Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. *J. Exp. Med.* (2013). doi:10.1084/jem.20121932

60. Lindqvist, M. *et al.* Expansion of HIV-specific T follicular helper cells in chronic HIV infection. *J. Clin. Invest.* (2012). doi:10.1172/JCI64314

61. Crum-Cianflone, N. F. *et al.* Durability of antibody responses after receipt of the monovalent 2009 pandemic influenza A (H1N1) vaccine among HIV-infected and HIV-uninfected adults. *Vaccine* (2011). doi:10.1016/j.vaccine.2011.02.040

62. Cubas, R. A. *et al.* Inadequate T follicular cell help impairs B cell immunity during HIV infection. *Nat. Med.* (2013). doi:10.1038/nm.3109

63. Boswell, K. L. *et al.* Loss of Circulating CD4 T Cells with B Cell Helper Function during Chronic HIV Infection. *PLoS Pathog.* (2014). doi:10.1371/journal.ppat.1003853

64. Cubas, R. *et al.* Reversible Reprogramming of Circulating Memory T Follicular Helper Cell Function during Chronic HIV Infection. *J. Immunol.* (2015). doi:10.4049/jimmunol.1501524

65. Kohler, S. L. *et al.* Germinal Center T Follicular Helper Cells Are Highly Permissive to HIV-1 and Alter Their Phenotype during Virus Replication. *J. Immunol.* (2016). doi:10.4049/jimmunol.1502174

66. Hufert, F. T. *et al.* Germinal centre CD4+ T cells are an important site of HIV replication in vivo. *AIDS* (1997). doi:10.1097/00002030-199707000-00003

67. Biancotto, A. *et al.* Abnormal activation and cytokine spectra in lymph nodes of people chronically infected with HIV-1. *Blood* (2007). doi:10.1182/blood-2006-11-055764

68. Younes, S.-A. *et al.* IL-15 promotes activation and expansion of CD8+ T cells in HIV-1 infection. *J. Clin. Invest.* (2016). doi:10.1172/JCI85996

69. Pizzolo, G. *et al.* Distribution and heterogeneity of cells detected by HNK-1 monoclonal antibody in blood and tissues in normal, reactive and neoplastic conditions. *Clin. Exp. Immunol.* (1984).

70. Bowen, M. B., Butch, A. W., Parvin, C. A., Levine, A. & Nahm, M. H. Germinal center T cells are distinct helper-inducer T cells. *Hum. Immunol.* (1991).

71. Kim, C. H. *et al.* Subspecialization of CXCR5+ T cells: B helper activity is focused in a germinal center-localized subset of CXCR5+ T cells. *J. Exp. Med.* (2001). doi:10.1084/jem.193.12.1373

72. Kim, J. R., Lim, H. W., Kang, S. G., Hillsamer, P. & Kim, C. H. Human CD57+ germinal center-T cells are the major helpers for GC-B cells and induce class switch recombination. *BMC Immunol.* (2005). doi:10.1186/1471-2172-6-3

73. Jia, Q. *et al.* Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer. *Oncoimmunology* (2015). doi:10.1080/2162402X.2014.1001230

74. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* (1991). doi:10.1109/18.61115

75. Su, L. F., Kidd, B. A., Han, A., Kotzin, J. J. & Davis, M. M. Virus-Specific CD4+Memory-Phenotype T Cells Are Abundant in Unexposed Adults. *Immunity* (2013). doi:10.1016/j.immuni.2012.10.021

76. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2938

77. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu393

78. Victora, G. D. *et al.* Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* (2010). doi:10.1016/j.cell.2010.10.032

79. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu Rev Immunol* (2012). doi:10.1146/annurev-immunol-020711-075032

80. Depoil, D. *et al.* Immunological synapses are versatile structures enabling selective T cell polarization. *Immunity* (2005). doi:10.1016/j.immuni.2004.12.010

81. Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J. & Sanjuán, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol.* (2015). doi:10.1371/journal.pbio.1002251

82. Piantadosi, A. *et al.* Breadth of neutralizing antibody response to human immunodeficiency virus type 1 is affected by factors early in infection but does not influence disease progression. *J. Virol.* (2009). doi:10.1128/JVI.01149-09

83. Rechavi, E. *et al.* Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. *Sci. Transl. Med.* (2015). doi:10.1126/scitranslmed.aaa0072

84. Prabakaran, P. *et al.* Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* (2012). doi:10.1007/s00251-011-0595-8

85. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5,** 171ra19 (2013).

86. Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front. Immunol.* (2012). doi:10.3389/fimmu.2012.00193

87. Prabhudas, M. *et al.* Challenges in infant immunity: Implications for responses to infection and

vaccines. *Nature Immunology* (2011). doi:10.1038/ni0311-189

88. Portugal, S. *et al.* Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function. *Elife* (2015). doi:10.7554/eLife.07218

89. Zinöcker, S. *et al.* The V Gene Repertoires of Classical and Atypical Memory B Cells in Malaria-Susceptible West African Children. *J. Immunol.* (2015). doi:10.4049/jimmunol.1402168

90. Tran, T. M. *et al.* An intensive longitudinal cohort study of malian children and adults reveals no evidence of acquired immunity to plasmodium falciparum infection. *Clin. Infect. Dis.* (2013). doi:10.1093/cid/cit174

91. Boyd, S. D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* (2009). doi:10.1126/scitranslmed.3000540

92. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci.* (2009). doi:10.1073/pnas.0909775106

93. Schroeder, H. W., Zhang, L. & Philips, J. B. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* (2001). doi:10.1182/blood.V98.9.2745

94. Jacob, J., Kelsoe, G., Rajewsky, K. & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature* (1991). doi:10.1038/354389a0

95. Ridings, J., Dinan, L., Williams, R., Roberton, D. & Zola, H. Somatic mutation of immunoglobulin V(H)6 genes in human infants. *Clin. Exp. Immunol.* (1998). doi:10.1046/j.1365-2249.1998.00694.x

96. Ridings, J. *et al.* Somatic hypermutation of immunoglobulin genes in human neonates. *Clin. Exp. Immunol.* (1997). doi:10.1046/j.1365-2249.1997.3631264.x

97. IJspeert, H. *et al.* Evaluation of the antigen-experienced B-cell receptor repertoire in healthy children and adults. *Front. Immunol.* (2016). doi:10.3389/fimmu.2016.00410

98. Biswas, S., Saxena, Q. B., Roy, A. & Kabilan, L. Naturally occurring plasmodium-specific IgA antibody in humans from a malaria endemic area. *J. Biosci.* (1995). doi:10.1007/BF02703849

99. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci.* (2015). doi:10.1073/pnas.1417683112

100. Chang, B. & Casali, P. The CDR1 sequences of a major proportion of human germline Ig VHgenes are inherently susceptible to amino acid replacement. *Immunol. Today* (1994). doi:10.1016/0167-5699(94)90175-9

101. Yaari, G., Uduman, M. & Kleinstein, S. H. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks457

102. Schroder, A. E., Greiner, A., Seyfert, C. & Berek, C. Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis. *Proc. Natl. Acad. Sci.* (1996). doi:10.1073/pnas.93.1.221

103. O'Brien, P. M. *et al.* Immunoglobulin genes expressed by B-lymphocytes infiltrating cervical carcinomas show evidence of antigen-driven selection. *Cancer Immunol. Immunother.* (2001). doi:10.1007/s00262-001-0234-6

104. Machida, K. *et al.* Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. *Proc. Natl. Acad. Sci. U. S. A.* (2004).

doi:10.1073/pnas.0303971101

105. Weitkamp, J. H., LaFleur, B. J., Greenberg, H. B. & Crowe, J. E. Natural evolution of a human virus-specific antibody gene repertoire by somatic hypermutation requires both hotspot-directed and randomly-directed processes. *Hum. Immunol.* (2005). doi:10.1016/j.humimm.2005.02.008

106. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014). doi:10.1038/nature13036

107. Haynes, B. F., Kelsoe, G., Harrison, S. C. & Kepler, T. B. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature Biotechnology* (2012). doi:10.1038/nbt.2197

108. Weisel, F. J., Zuccarino-Catania, G. V., Chikina, M. & Shlomchik, M. J. A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity* (2016). doi:10.1016/j.immuni.2015.12.004

109. Krishnamurty, A. T. *et al.* Somatically Hypermutated Plasmodium-Specific IgM+Memory B Cells Are Rapid, Plastic, Early Responders upon Malaria Rechallenge. *Immunity* (2016). doi:10.1016/j.immuni.2016.06.014

110. UNICEF. *The State of the World's Children 2014 in Numbers: Every Child Counts Revealing disparities, advancing children's rights*. *United Nations Children's Fund* (2014).

111. World Health Organization. *World Malaria Report 2015*. *World Health* (2015). doi:ISBN 978 92 4 1564403

112. White, M. T. *et al.* A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC Med.* (2014). doi:10.1186/s12916-014-0117-2

113. Newell, E. W. & Davis, M. M. Beyond model antigens: High-dimensional methods for the analysis of antigen-specific T cells. *Nature Biotechnology* (2014). doi:10.1038/nbt.2783

114. Knapp, B., Demharter, S., Esmaielbeiki, R. & Deane, C. M. Current status and future challenges in T-cell receptor/peptide/MHC molecular dynamics simulations. *Brief. Bioinform.* (2015). doi:10.1093/bib/bbv005

115. Marks, C. & Deane, C. M. Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal* (2017). doi:10.1016/j.csbj.2017.01.010

116. Armstrong, K. M., Piepenbrink, K. H. & Baker, B. M. Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes. *Biochem. J.* (2008). doi:10.1042/BJ20080850

117. Shugay, M. *et al.* VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx760

118. Holt, R. A. Interpreting the T-cell receptor repertoire. *Nat. Biotechnol.* **35,** 829–830 (2017).

119. Wang, L. The L1 penalized LAD estimator for high dimensional linear regression. *J. Multivar. Anal.* (2013). doi:10.1016/j.jmva.2013.04.001

120. Slawski, M. *et al.* Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics* (2012). doi:10.1186/1471-2105-13-291

121. Wu, C. & Ma, S. A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinform.* (2014). doi:10.1093/bib/bbu046

122. Miyazawa, S. & Jernigan, R. L. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins Struct. Funct. Genet.* (1999).

doi:10.1002/(SICI)1097-0134(19990101)34:1<49::AID-PROT5>3.0.CO;2-L

123. Rosen, J., Park, H., Glick, J. & Zhang, L. Accurate solution to overdetermined linear equations with errors using L1 norm minimization. *Comput. Optim. ...* 329–341 (2000). doi:10.1023/A:1026562601717

124. Boyd, S. & Vandenberghe, L. *Convex Optimization*. *Optimization Methods and Software* (2004). doi:10.1017/CBO9780511804441

125. Gagniuc, P. A. *Markov Chains: From Theory to Implementation and Experimentation*. *Markov Chains: From Theory to Implementation and Experimentation* (2017). doi:10.1002/9781119387596

126. Berkelaar M, Eikland K, N. P. Package 'lpsolve'. 1–11 (2004).

127. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* (1995). doi:10.2307/2346101

128. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017). doi:10.1038/nature22383

129. Yin, L., Ge, Y., Xiao, K., Wang, X. & Quan, X. Feature selection for high-dimensional imbalanced data. *Neurocomputing* (2013). doi:10.1016/j.neucom.2012.04.039

130. Ahmed, I., Pariente, A. & Tubert-Bitter, P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Stat. Methods Med. Res.* (2018). doi:10.1177/0962280216643116

131. Grimes, D. A. & Schulz, K. F. Compared to what? Finding controls for case-control studies. *Lancet* (2005). doi:10.1016/S0140-6736(05)66379-9

132. Trefethen, L. N. & Bau III, D. Numerical linear algebra. *Numer. Linear Algebr. with Appl.* (1997).

doi:10.1137/1.9780898719574

133. Yen, I. E.-H. H. *et al.* Sparse Linear Programming via Primal and Dual Augmented Coordinate Descent. *Adv. Neural Inf. Process. Syst. 28* (2015).

134. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.* (1993).