PhD Thesis

Xiao Liu

Immune Repertoire Characteristics and Dynamics in Cancer

Supervisor: Karsten Kristiansen

Submitted on: 7th January, 2017
In this thesis, I describe the experimental and analytical methodologies we developed for immune repertoire analysis. We then use the tools to investigate the characteristics and dynamic of immune repertoire on B-Acute lymphoblastic lymphoma and breast cancer.
# Table of contents

ABSTRACT ........................................................................................................................................5

INTRODUCTION .................................................................................................................................6

LIST OF ARTICLES RELATED TO THE TOPIC OF THIS THESIS ..........31

LIST OF ARTICLES NOT RELATED TO THE TOPIC OF THIS THESIS AND COMPLETED DURING THE PHD CAREER .........................32

RESULTS ....................................................................................................................................34

Part One. Optimization, evaluation and comparison of experimental approaches ....34

Result summary of Article 1: ........................................................................................................34

Part Two. Development of the analytical toolset “IMonitor” .........................................................35

Result summary of Article 2: ........................................................................................................35

Part Three. BCR repertoire analysis in blood cancer: Minimal residual disease detection and evolved IGH clones analysis in acute B lymphoblastic leukemia using IGH deep sequencing ..................................................................................................................35

Result summary of Article 3: ........................................................................................................35

Part Four. TCR repertoire analysis in solid cancer: investigation on infiltrated T lymphocytes in breast tumors and adjacent tissues ..................................................................................................................37

Result summary of the Article 4: ..................................................................................................37

DISCUSSIONS AND PERSPECTIVE ......................................................................................38

CONCLUSIONS ...............................................................................................................................43
Abstract

The diversity of T and B cells in terms of their receptors is huge in the invertebrate’s immune system, to provide broad protection against the vast diversity of pathogens. Immune repertoire is defined as the sum of total subtypes that makes the organism’s immune system, either T cell receptor or immunoglobulin. Before the emergence of high-throughput sequencing, the study on immune repertoire is limited by methodology, as it is impossible to capture the whole picture by the low-throughput tools. The massive paralleled sequencing suits perfectly the study on immune repertoire. In this thesis, I describe the experimental and analytical methodologies we developed for immune repertoire analysis. We have devoted extensive efforts on the optimization and evaluation of these pipelines. We then use the tools to investigate the characteristics and dynamic of immune repertoire on both blood and solid cancer. Specifically, we provide analysis on leukemic IGH sequences and their evolution in B-cell acute lymphoblastic leukemia, and monitor the dynamics of IGH repertoire during chemotherapy. Further, the TCR repertoire of infiltrated T lymphocytes in breast tumor and adjacent tissues was analyzed. We describe numerous interesting findings that might provide hints on immunotherapy of breast cancer.
Introduction

1. The functional role of T and B lymphocytes in the adaptive immunity

The immune system of vertebrates consists mainly of two subsystems or units, which are innate immune system and adaptive immune system (or acquired immune system), though both are not mutual exclusive. While the innate immune system composes cells and elements that protects living body from the invasion of pathogens through a non-specific and non-memory way, the adaptive immunity, which consists of specialized and systemic cells and components, can prevent and eliminate the foreign and environmental virus, pathogens and other immunogens in a specific and targeted way. Acquired immunity is “acquired” through its initial contact and immunological response to the foreign antigens, and creates a specific immunological memory, which will lead to enhanced immune response with the secondary stimulation by the same foreign antigens. Several key steps are consisted in the “acquired” process, including recognition of the previously encountered antigens, the activation of the effector cells, and their clonal expansion. On the other hand, the immune system is “adaptive” because it can adapted to the diversity and dynamic changes of the foreign pathogen attack, and construct and strengthen its immune response to protect the body through the processes such as cell clonal selection and affinity maturation.

The adaptive immunity consists two major components of cellular immunity and humoral immunity. In the adaptive immune system, numerous cell population and subpopulation are involved to maintain the immune balance and the normal function of life. The dysfunction of the adaptive immunity will elicit serious abnormalities such as infection, autoimmune diseases and cancer. The white blood cells which generate or are involved in the generation of adaptive immune response are known as lymphocytes. T and B lymphocytes are two prolific and functionally important cells that play pivotal roles in the cellular and humoral immune response. One of the two major T lymphocyte subpopulation, CD8+ T
cells, can serve as cytotoxic cells that directly attack and destroy the pathogens. B lymphocytes and the other subpopulation of T lymphocytes-CD4+ T cells, are both involved in the humoral immunity. CD4+ T cells, also known as helper T cells, can help to activate B lymphocytes, and the activated B cells is able to produce antibody, which is the circulated format of the membrane-bounding protein - B cell receptors (BCR). Antibody is highly specific to the foreign or the innate antigens, and is able to directly bind and neutralize them, which prevents the antigen to bind the host targets and elicit downstream immunological cascade to harm the body. The immunological pathways and networks involving in T and B lymphocytes are quite complex and are still under continuous research, and there are also other numerous cell types being involved in the signaling. Some details in the process have already been uncovered. After the invasion of the foreign pathogens or microbe, a subset of B lymphocytes can bind and recognize the antigens by their trans-membrane BCR. The T lymphocytes can also express their trans-membrane receptor proteins on the cell surface, which is known as T cell receptors (TCR). The binding and recognition of TCR to the invading antigens is reliant upon the antigen-presenting cells (APC). Abundant cell types can serve as APC, such as B lymphocytes, macrophages, however, the most important APCs are dendritic cells (DCs). DCs express the major histocompatibility complex (MHC) type I and type II molecules on their surface membrane. After fragmentation and internalizing of the foreign antigens through the process of phagocytosis or receptor-mediated endocytosis, DCs can recognize and interact with the antigens with the expressed MHC II molecules, and present the antigen to the TCRs on the membrane of T cells to form a TCR-antigen-MHC protein complex. Additionally, DCs can express the co-stimulation factor B7 protein molecules to interact with the CD28 molecules expressed on the surface of T cells, to secret the cytokines and signals to activate the naïve T cells. Taken together, with the help of the other immunological cells such as APCs, and those which produce necessary cytokines to prime the activating pathways, the mature and naïve CD4+ T lymphocytes can recognize and bind the antigens
presented by the MHC molecules of APCs with the TCRs, and be activated to turn to effector helper T (Th) cells. The Th cells consequently produce cytokines that will help to activate the mature and naïve B cells when they encounter the antigens and recognize the “non-self” antigens in the presence of “self” . The B lymphocytes will further experience clonal expansion, and their trans-membrane BCRs will continuously undergo somatic hypermutation to enhance the ability to bind to the epitopes of the antigen, which is known as affinity maturation. Thereafter, the activated B cells will be differentiated into memory cells and the effector cells. The memory B cells will stay in the immune system for long-lasting protection, and easily triggered and expanded when the same antigens are encountered later on. Upon stimulation, the memory B cells will differentiate to plasmablasts and secret abundant neutralizing antibody which are specific to the antigen to eliminate it. This process is significantly faster than the first infection of the antigens. On top of the humoral immunity, the activated CD8+ T cells can serve as cytotoxic T lymphocytes (CTL) that directly recognize and destroy the infected cells expressing antigens. In summary, the T and B lymphocytes play the central roles in the adaptive immune system.

2. **The development of TCR and BCR diversity**

Bearing the significant functions in the immune system, T and B lymphocytes are derived, differentiated and mature in the different organs and tissues. Both T and B cells stem from the multi-potent stem cells in the bone marrow. The T progenitor cells will migrate through peripheral blood to thymus, which is the place for the differentiation and maturation of T lymphocyte, while the bone marrow is the home for all the mature B lymphocytes. The proteins TCR and BCR which mostly expressed on the cell surface of T and B lymphocytes respectively, are encoded by a very unique protein-coding system and germline genes. The protein structure of BCR also known as immunoglobulin (Ig) is composed of both heavy and light chains, in which light chain can be lambda (\(\lambda\)) or kappa (\(\kappa\)) chain. During the process of B cell maturation, the BCR coding genes undergo a somatic
recombination procedure called V(D)J recombination (rearrangement) in which basically Variable (V) genes, Diversity (D) genes, Joining (J) genes and Constant (C) genes are involved. All these genes are big gene families usually contained numerous gene copies bearing high sequence homology. For instance, the human immunoglobulin heavy chain gene locus contain 44 V gene segments, 27 D genes, 6 J genes plus 2 C genes. The immunoglobulin light chain gene region also scatters numerous V, J and C gene locus, but without D genes as it is not involved in the light chain gene rearrangement. The human heavy, lambda and kappa chain gene locus of immunoglobulin locate on the chromosome 14, 22 and 2 respectively. TCRs also have variable peptide chains such as alpha ($\alpha$), beta ($\beta$), gamma ($\gamma$) and delta ($\delta$), while most T cells express $\alpha/\beta$ TCRs (approximately 95% human T cells express $\alpha/\beta$ TCRs and 5% are $\gamma/\delta$ T cells). The TCR genes are similar to the described BCR genes, that TCR beta genes rearrange with the multiple V, D, J and C segments and alpha genes recombine with the V, J and C gene locus.

During the maturation of B cells, the D and J genes rearrange first, then V genes recombine with the rearranged D/J genes. Only one V/D/J gene can take part in each recombination which involves the RAG enzyme. The variety of immunoglobulin genes that could be used to recombine provides the basis for the diversity of BCR, and the additional modification on the recombination products, such as the random nontemplated nucleotide addition and deletion in the V-D and D-J joining regions, somatic hypermutation, heavy/light chain pairing and transcription(Kirkham and Schroeder, 1994; Yancopoulos and Alt, 1986), finally yield a astronomical number of possible BCR in the human immunoglobulin repertoire ($10^{13}$) (Schroeder, 2006) which exceeds the total number of B lymphocyte in the human body ($10^{11}$). This vast diversity of potent BCRs enable their abilities to recognize a variety of exogenous and endogenous antigens, and continuously increase the affinities by consistent exposure, which fundamentally maintain the balance the humoral immunity(Muramatsu et al., 2000). In the development of thymocytes, the TCR undergoes essentially the same sequence of
ordered recombination process that ensembles BCR. The TCR locus rearrangement begins from the double negative pro-T cell when the beta chain D genes and J genes finish recombination first, and V genes involve later on to form the V/D/J products and express the TCR beta chain. During the turning process of the pro-T cell to pre-T cell, CD4 and CD8 express simultaneously on the T cell surface to form the double positive cells, and the alpha chain V and J genes are rearranged and paired with the beta chain to form the whole TCR heterodimer. Likewise, the vast number of TCR gene segments that serve as potent candidates for somatic recombination, and the possibilities of alpha/beta chain pairing, constitute the diversified TCR repertoire to recognize the MHC/antigen complex, which help forming the cellular immunity of CD8+ TCR repertoire, as well as the humoral immunity of helper T cells expressing CD4 molecules.

A complete DNA molecule or RNA transcript is divided into the leader sequence, the framework region (FR) and the complementarity-determining region (CDR). Typically four FRs and three CDRs are juxtaposed in the sequence of FR1, CDR1, FR2, CDR2, FR3, CDR3 and FR4. The antigen binding sites of TCR and BCR are commonly thought to be in the CDR regions, which are far more hypervariable than the FR regions. CDR3 is usually the most variable region to act as the central part of the antigen-binding site (Xu and Davis, 2000), where V-D and D-J joining sections are included and both nontemplated addition/deletion and somatic hypermutation contribute to its diversification.

3. **Researches prior to the NGS technology**

3.1 **prior technology**

Considering the importance of diversification of TCR and BCR repertoire to reflect the adaptive immune status, researchers for long have been keen to study its relationship with infection and other immunological disease, and uncover specific TCRs and BCRs. Sanger sequencing was the only technique to read the nucleotide sequences of the rearranged DNA and the transcribed RNA at that time, however, due its low capacity, it’s extremely difficult for Sanger sequencing to
grasp the enormous sequence diversity of TCRs and BCRs. The largest throughput in the literature was to sequence hundreds of TCRs/BCRs by Sanger sequencing (Hindley et al., 2011; Pacholczyk et al., 2006). Before the emergence of sequencing, several low-resolution techniques had been exploited to visualize the diversity of TCR/BCR repertoire, such as various forms of gel electrophoresis. Specifically, a molecular technique known as spectratyping or immunoscope which employed gel electrophoresis to dissect TCR/BCR repertoires after V-gene specific polymerase chain reaction based on the CDR3 length were adopted in numerous studies (Luo et al., 2008). This method is quick and cheap compared to Sanger sequencing, however, it can’t give any information on the nucleotides or amino acids of CDR3 loops, which is crucial to infer the structure and uncover how TCRs/BCRs recognize the MHC and antigens. On the protein level, flow cytometry had been applied from the early days to quantify the T cells that expressed the α/β TCRs from a particular V gene family. The T cell population can be sorted and quantified by a panel of stained or fluorescence-tagged TCR Vβ and Vα-specific monoclonal antibodies, either individually or in a multiplex format. At least 24 Vβ-specific and three Vα-specific monoclonal antibodies have been available for use. Even today, due to the infeasibility of protein quantification technology such as Mass-spectrum and nuclear magnetic resonance (NMR) which involving fragmentation to apply to the TCR proteins, flow cytometry still seems to be the only available technique to quantify TCRs from the protein level. The limitations of this method is also obvious, which is quite rough and of low resolution. It can only access whether the TCR repertoire is bias towards a particular V gene family, without the exploration of the prolific diversity of the TCR clonatypes within it. The limitation of the availability of the TCR monoclonal antibodies directly determines the resolution of the technique. Additionally, to simplify the TCR model, several groups had developed TCR mini transgenic mice model to reduce the enormous diversity of the fully polyclonal TCR repertoire. Genetically engineered mice are knocked out part of V/D/J germline genes to recombined a limited size of rearranged TCR repertoire while
maintain the normal development and function of the T cells and the adaptive immune response (Lathrop et al., 2011; Pacholczyk et al., 2006). Overall, all these efforts and technologies had paved the way and made it possible to study the great diversity of TCR and BCR repertoire before the emergence of the high throughput sequencing.

3.2 research fields

The above technologies have been employed for more than two decades to investigate the immune repertoire across the full spectrum of cellular and humoral immune response in animal models and human beings. Generally speaking, researches in this field have been focusing on two directions. Firstly, the VDJ pattern of the TCR/BCR being characteristic of T/B cell identifier in the cell population, can be utilized to compare and track the origin of the cells in the particular environment or of certain subtypes/groups. Combined with the traditional technique, such as TCR mini mice model, PCR and gel electrophoresis, spectratyping and Sanger sequencing, researchers have studied the origin of regulatory T cells (represented as Foxp3+CD4+CD25+ T cells) and compared their TCR diversity with other T cell types. They found the majority of regulatory T cells originated from the immature cells in the thymus rather than converting and recruiting the mature T cells from the periphery. Additionally, the diversity of the regulatory T cells exceeded the naïve T cells in the periphery (Pacholczyk et al., 2006). Another study investigated the origin of regulatory T cells infiltrated in the carcinoma and found the regulatory T cells expressed distinct TCRs from other conventional T cells and the antigen-experienced effector/memory cells, which implied tumor-infiltrating lymphocytes (TIL) mostly contained natural regulatory T cells rather than induced or converted cells (Hindley et al., 2011).

Secondly, TCR/antibody repertoire and specific TCR/antibody in association with various diseases has been extensively studied. In the traditional immunological theory, VDJ recombination is mostly randomized which makes the probability of two individuals sharing the same TCR/BCR regarding to a specific antigen or
pMHC antigen vanishingly small. However, after the first reports of TCR bias in the 1990s from the studies of human beings exposed to influenza A virus and Epstein-Barr virus (EBV) (Argeat et al., 1994; Moss et al., 1991), growing reports have emerged to describe the predictable distortions of TCR/BCR repertoire specific for individual antigens or pMHC antigens. These striking phenomena which shook the foundation of the classic immunology, has been referred as “VDJ bias”, “skewing”, “restriction” or “public sequences”. People have discovered various forms and extents of VDJ bias with regard to all kinds of exposures/diseases. Among them, the most obvious examples were found in the field of infections and vaccinations, as mentioned above. Some pathogens or vaccines can produce superantigens which are proteins that bind to certain antibody V domains, thus will result in skewed or even shared antibodies among individuals. Notably, the public TCRs are often restricted by MHC molecules. Abundant pathogens have been found to induce public TCRs in the background of pMHC, including influenza A virus, EBV, cytomegalovirus (Khan et al., 2002), hepatitis B (Sing et al., 2001) and C virus (Umemura et al., 2000) and Human immunodeficiency virus (Yu et al., 2007).

Skewed TCR/BCR repertoire in malignancy has also been found. Unlike pathogen infection and vaccination, the tumor-associate antigens (TAA) are much more complicated and involve various proteins and small molecules, which cause them much difficult to detect and determine. In addition to that, the affinity between the TAA and TCR/BCR is always weak which lead to the detection and isolation of TAA-specific TCR/BCR by ex vivo/in vivo techniques unreliably difficult. Nevertheless, the blood carcinoma and lymphoma which induced by malignant T cells or B cells, have been found to carry stereotyped repertoire among individuals (Agathangelidis et al., 2012). Though these clones are just identifiers for the malignant T/B cells rather than the effect of training or education of the TAAs in the tumor microenvironment or the periphery, this phenomenon is still against our recognition to the blood malignancies, that the malignant T/B cells are originated from somatic mutated clones and the probability of different
individuals carrying the same TCR/BCR clone is relatively small. It inferred the malignant T/B cells are not just driven by somatic mutations and other commonly known factors, but also involving selection of the VDJ recombinations. The lymphocytes infiltrated in the solid tumors have been recurrently reported to have diagnostic and prognostic values, and play pivotal roles in the entire process of tumorigenesis. The regulatory T cells infiltrated in the solid tumor have been found to skewed towards public sequences, which strongly imply that the interaction between the lymphocytes and TAAs has led to the selection of these public cells (Sainz-Perez et al., 2012). Another interested phenomenon have implied these bias has not been driven solely by TAAs, but also can be triggered by certain pathogens in some viral-associated cancer.

Immune-related diseases especially the autoimmune disorders have also been extensively studied for public TCRs. The self-antigens of autoimmunity have not been clarified thus far. As a result, the targets of the self antibodies and TCRs are always difficult to define, both in terms of peptide specificity and MHC restriction. Regardless of the unclear mechanism, people have found the TCR alpha and beta chains are conserved interindividually in systemic lupus erythematosus (Luo et al., 2008). Besides, several studies have reported the TCR bias and MHC restriction in multiple sclerosis (Oksenberg et al., 1993; Wucherpfennig et al., 1990).

4. the application of NGS in the TCR and BCR repertoire study

4.1 the technology

If the entire T/B cell repertoire and the antigen specific T/B cells are to be treated as a haystack and a needle, the emergence of next generation sequencing (NGS) or high-throughput sequencing (HTS) provides a powerful weapon to identify/discover a needle in a haystack. The characteristics of NGS, such as relatively short read-length (compared to Sanger sequencing), enormous reads number and cost-effectiveness per read, fits ideally the investigation on TCR and BCR repertoire. The earliest NGS platform used on this area is the Roche 454 system, which provide on average 500bp
read-length and modest read number per run. It is able to capture the entire VDJ regions of TCR and BCR, and can sequence millions of molecules per repertoire, which significantly surpass the original Sanger sequencing technique. In 2009, the first study utilizing 454 to sequence the zebrafish antibody repertoire was published in *Science* (Weinstein et al., 2009), and ignited the passion on this research area. Most of the several first papers adopted 454 sequencer to cover the entire VDJ recombinated region of the receptor genes(Boyd et al., 2009; Wang et al., 2010). The first study utilizing Illumina sequencer to cover the most variable CDR3 regions was published in *Genome Research* to investigate the TCR beta chain characteristics in the human periphery(Freeman et al., 2009). Compared to 454, the Illumina Hiseq platform provides shorter read-length, but significant higher throughput of reads, and lower cost per read, which is brilliantly suitable to deep sequence a quite complex repertoire. In the current applications, read-length and throughput are still two factors that investigators should compromise and do trade-off. For those applications which require reading the entire VDJ region such as antibody discovery, long reads from 454 or Illumina Miseq are the best fit; for TCR-seq, in which CDR3 regions are mainly focused, Illumina Hiseq platform is quick and cost-effective.

Stepping into the NGS era to recover the complexity of TCR/BCR repertoire, the most critical issues before we use this technology are dealing with the bias and errors occurred in the various parts of the process. Both experimental and bioinformatical methodology should take into account of these problems. No matter starting from gDNA or mRNA samples, the VDJ diversity of TCR/BCR should be enriched with high fidelity before processing to NGS sequencing. However, due to the rareness of the specific VDJ molecules especially the low-frequent copies, the amplification is usually evitable and bias comes along with it. Multiplex PCR is the most convenient and straightforward amplification approach and various isoforms of it have been developed(Klarenbeek et al., 2010; Wang et al., 2010). Multiplex PCR always
incorporates V gene and J gene (or C gene) specific multiplex primers to amplify the full VDJ combined or CDR3 region of the gDNA or cDNA, depending on the region of the V genes the primers target. This one-tube reaction is convenient, but the different efficiencies and cross-reaction of the various primers will definitely bring into certain extent of bias to the amplified products. In order to standardize the procedure to detect the clonally rearranged Immunoglobulin and TCR in a diagnostic setting, such as minimal residual disease detection, the BIOMED-2 collaborative effort has developed and validated standard multiplex PCR primers for different regions of V-J/C genes early before the emergence of NGS (van Dongen et al., 2003). Several groups are still using the BIOMED-2 primers or the optimizations based on them. Other groups, however, are trying to implement various approaches to minimize the gene-specific bias. Isoforms such as multiplex linear PCR as mentioned before is one way. Other ways include, meticulous analysis of bias, changing/selecting the best primers, and adjusting the primer input in the PCR reaction (Carlson et al., 2013), or incorporating random barcodes in the primers to adjust the bias and errors (Vollmers et al., 2013). Through all these efforts, the PCR bias can only be minimized but never be diminished. The other experimental enrichment of 5’RACE on the mRNA, on the other hand, can theoretically avoid V gene bias as it only incorporate C-gene primers. 5’RACE is short for rapid amplification of cDNA 5’ ends, and can only work on mRNA but not genomic DNA. Primers are designed against a known region in the 3’mRNA, which is the constant gene region in the repertoire analysis. The mRNA is reverse transcribed, and a homopolymeric tail is synthetically added to the 3’ends of the cDNA using dNTP (e.g., dCTP) by terminal transferase through the template-switch process (Peters et al., 1999). Therefore, the products can be amplified by the constant gene primer and the homopolymeric primer. Some alterations have been put into this process in order to shorten the product length to get it sequenced by NGS platforms with short-read length, such as Illumina Hiseq system. For instance, biotins are
labeled to the constant gene primers, and the cDNA products are randomly fragmented to appropriate length to be sequenced by single reads or paired-end reads of Hiseq machine. The biotin-labeled fragments are specifically captured by the magnetic beads to enrich our targeted fragments which contain CDR3 and the constant gene region. The RACE approach has been extensively adopted in the current TCR/BCR studies (Li et al., 2013; Warren et al., 2011).

Both multiplex-PCR and RACE approaches for investigation on TCR/BCR have their pros and cons, but the comparison between the two hasn’t been achieved. In this study, I will present a systemic evaluation and comparison of the both methodologies and discuss in details of their applications.

In terms of the enrichment of VDJ recombinations, there is also other approaches such as using probes/baits that are designed against the V/J germline gene sequences to capture the rearranged VDJ fragments (He et al., 2011). This approach, however, requires random shearing of gDNA/cDNA before the capturing step, and will induce the infrequent complete rearrangements being fragmented and lost in capturing and sequencing. Therefore, it is applicable only in recovering the abundant rearrangements in cases such as lymphoma or leukemia.

The chain-pairing information of V_H:V_L for BCR and V_α: V_β for TCR is critical to elucidate the function of T/B cells. Previous approaches amplify each chain separately, or enrich two chains simultaneously in the same reaction whereas in a population level, which couldn’t tell every V_H:V_L or V_α: V_β pair. Until recently, the study on native chain pairs has been achieved through single-cell cloning by limiting dilution and sanger sequencing on individual cells. As discussed, this strategy is low-throughput, expansive and limited in paired-chain sequences. After incorporating with NGS, new strategies have been developed to enhance their throughput and feasibility to uncover native chain-pairing information. A cell-based emulsion RT-PCR approach was developed to allow the selective fusion of the native pairs of
TCR $\alpha$: $\beta$ genes, and more than 700 pairs could be recovered in a single emulsion experiment (Turchaninova et al., 2013). DeKosky et al. developed a $V_H$: $V_L$ pairing technology that relies on sequestering single B cells into subnanoliter volume wells, lysing the cells, capturing RNA on poly-dT beads and generating amplicons encoding linked $V_H$: $V_L$ segments by emulsion overlap extension PCR (DeKosky et al., 2013). Thousands of unique endogenous $V_H$: $V_L$ pairs with 97% validated pairing accuracy in a one-day experiment was achieved. Recently, Bryan Howie et al. reported a pairSEQ approach in which a fixed number of T cells are randomly allocated to each well on a 96-well plate, and mRNA was reverse transcribed, attached with well-specific barcodes, amplified, pooled together and sequenced. Any pairs that share a unique set of wells (barcodes) are considered to come from the same clone, since the probability of two clones sharing the same set of wells is minimal (Howie et al., 2015). Approaches with higher throughput pairing combining emulsion PCR and NGS are under development, and we can anticipate the routine adoption of TCR/BCR pairing sequencing in this field in the near future. Besides the pairing information, single-cell approach can also link the T/B cell phenotype such as gene expression signatures with the receptor pairing sequences, which could help to elucidate the T/B cell evolution and heterogeneity in function (Han et al., 2014).

Feasible bioinformatic tools had been developed to align and analyze TCR and BCR sequences before the emergence of NGS sequencing, including iHMMune-align (Gaeta et al., 2007), IMGT/V-QUEST (Brochet et al., 2008), and NCBI's IgBLAST. These tools are very useful to determine the V/D/J genes recombined, the CDR region annotation and the characteristics of the TCR/BCR Sanger sequences. NGS provides us with unprecedented depth and diversity of immune repertoire ever, and more powerful tools/software are warranted to accurately and fastly handle these enormous data. A couple of new or updated computational tools have been published to analyze NGS repertoire data. New editions for NGS data analysis have been developed by
IMGT and NCBI, such as IMGT/HighV-QUEST (Li et al., 2013) and new IgBLAST (Ye et al., 2013). Other tools including MiTCR (Bolotin et al., 2013) and Decombinator (Thomas et al., 2013) have been published to analyze TCR repertoire data, while MiTCR has been further updated to MiXCR to mine both TCR and BCR data (Bolotin et al., 2015). These established tools are available for VDJ gene assignment, CDR and FR annotation, CDR3 length identification, insertion and deletions analysis, and mutation spectrum analysis (BCR). Several technique aspects should be considered in the improvement of the analytical tools, including identification of PCR and sequencing errors from the true biological variations, and adjusting the PCR bias. While experimental optimization are helpful in these aspects, in silico methods could also be beneficial. Taking them into account, we have established our own bioinformatic tools entitled IMonitor, which is short for Immune Monitor, to analyze both TCR and BCR repertoire data (Zhang et al., 2015). We have validated IMonitor on both simulated data and biological data, and comparison with existed tools reveals the accuracy and superiority of IMonitor. Currently, IMonitor has been utilized to analyze all kinds of repertoire data, including minimal residue disease detection in leukemia, analysis on tumor infiltration lymphocytes, and study on TCR and BCR response upon various vaccination.

On top of the basic analytical tools, additional computational methods have developed to dig into the data in greater details. In terms of scoring the clonality and diversity of the repertoire. Shannon entropy index and simpson’s diversity index have been introduced from ecology to estimate the repertoire diversity but bear obvious limitations including low sensitivity and dependence upon sequencing depth. New bioinformatic frameworks has been established to link the immunological status with the diversity indices (Greiff et al., 2015), and to estimate the species richness and distribution (Laydon et al., 2014). Furthermore, various computational pipelines have been developed to analyze the B cell clone lineages and evolution, and identify the
broadly neutralizing antibody.

4.2 research fields

4.2.1 basic research

Since NGS employment has forwarded the TCR/BCR repertoire research into a unprecedented level, the population characteristics on various species has been investigated in a new trend, including zebrafish (Weinstein et al., 2009), mouse (Madi et al., 2014) and rhesus monkey (Sundling et al., 2012). We have also recently submitted an analysis on camel antibody repertoire. The most crucial problem for these analysis is the lack of V/D/J gene reference sequences for alignment and annotation. While these model animals are widely used in immunological research, such as rhesus used in vaccine design and pre-clinical test, and camel adopted for the unique VHH antibody discovery, we can anticipated that in combination with the traditional animal model study, immune repertoire data collection and analysis will complement and extend our understanding in their immunological status. In this paper, to elucidate the T and B cell response on the novel Dengue vaccine, we have also used rhesus monkey as the animal model.

Bulk and subsets of periphery T/B cells in human have also been investigated for their repertoire heterogeneity by the NGS deep sequencing. T cells in periphery have been studied for their repertoire size (Warren et al., 2011) and the different subsets have been compared for the identical TCRs to seek for the cell fate determination (Wang et al., 2010). Human memory T cells have been found to consist mainly of unexpanded T cells with a relatively small expanded proportion (Klarenbeek et al., 2010). With regards to T cell fate driven by pathogens, a paper published in Science has revealed the functional heterogeneity of memory T cells primed by pathogens or vaccine, and naïve T cells could be induced to multiple fates by the pathogens in vitro (Becattini et al., 2015). Studies on monozygotic twins have revealed that the naïve antibody repertoire is highly heritable (Glanville et al., 2011), but the situation for the
TCR repertoire is more complex (Zvyagin et al., 2014). Interestingly, researchers have identified a antigen-independent selection of maturing B cells (Kaplinsky et al., 2014).

4.2.2 translational research

TCR/BCR spectrum profiling by NGS has been applied to enormous translational research in these years and has a wide range of clinical and healthcare applications. Their application area includes but not limits to antibody discovery, infection and allergy, cancer diagnosis and prognosis, transplantation monitoring, autoimmune disease, vaccine development and response monitoring, and hematology diagnosis and prognosis.

Traditional technique for antibody discovery replies on hybridoma preparation and large-scale screening on cells with high affinity. This method is reliable, but time-consuming and labor-intensive. With direct NGS sequencing on antibody repertoire of immunized animals, we are able to capture the entire antibody response in the periphery or the splenic B cells. By analytical algorithm, the potential antibodies with high affinity to the immunogen can be predicted by their abundance, heavy/light chain pairing and somatic evolution (Reddy et al., 2010). Combining with protein profiling of Mass spectrum in the serum, the antibody proteins can be more accurately identified and quantified, and potentially provide better predictions and outcomes (Cheung et al., 2012). The antibodies with known sequences can then be directly cloned, expressed, and test affinities. TILs are good prognostic biomarkers and has been extensively studied on melanoma, breast cancer and other solid tumors. However, due to limitation of the technique, tumor-infiltrating T cells have only been taken a glimpse of and the heterogeneity of T cells in response to the immunotherapy such as PD-1 or CTLA-4 blockade has not been able to investigated. Recently, studies on CTLA4 blockade treatment on the periphery T cells revealed that the treatment could broaden the TCR repertoire and increase its diversity, and the responders
maintained a more stable repertoire with less expanded or contracted TCR clones compared with the non-responders (Cha et al., 2014; Robert et al., 2014). The B cell isotypes in periphery blood of allergy were also investigated and the IgE was found to be more frequently switched from IgG1 rather than IgM (Looney et al., 2015). As for the autoimmune disease, multiple sclerosis (MS) and systemic lupus erythematosus (SLE) were studied and B cells populating in the MS brains were found to be mature in the draining lymph node (Stern et al., 2014; Thapa et al., 2015; von Budingen et al., 2012). T cell repertoire was investigated following the autologous stem cell transplantation for MS (Muraro et al., 2014).

There are abundant unanswered questions regarding the magnitude and persistence of protection for the new vaccines designed to market. Recently, significant efforts have been put into the understanding of innate immunity and T/B cell response induced by the vaccines, and the advances in this area have benefited a lot to the rational design of vaccines. The hallmarks of vaccine-induced immune response are essentially antibodies secreted by B lymphocytes and various involving of different T lymphocytes. Both antibody responses including primary low-affinity antibody secreting in the extra-follicular reaction and affinity maturation in the germinal centers together with the secondary responses of the memory recall, and the antigen specific effector and memory T cells coordinated with each other and other cell types, provide effective protection of the vaccines to the host.

Reasonably, the dynamic antibody and T cell responses induced by vaccines can benefit substantially from the NGS TCR/BCR repertoire studies. The induction and proliferation of antigen-specific T/B cells, the class-switching of antibody isotypes, and the somatic hyper-mutations acquired in the selection process of affinity maturation, can be investigated in details by analyzing the TCR/antibody repertoire sequences. Together, the magnitude and breadth of responses can be assessed by the change of the repertoire and compared among different dosing time-points and vaccine groups. In addition, the
heritability and environmental contribution to vaccine responses remain an unanswered question.

Bearing all these interesting topics, NGS has facilitated researches on various aspects of vaccine responses recently to aid the design of novel vaccines. Non-human primates of rhesus as the preclinical animal model for vaccine design has been utilized to investigate the presence and development of antigen-specific antibodies post immunization of a HIV vaccine, with the prior knowledge of identified HIV neutralizing antibodies (Dai et al., 2015). Human antibody responses post influenza vaccination has been traced for antigen-specific antibody clones and a certain extent of their convergent rearrangement has been identified (Jackson et al., 2014). Longitudinal study on dynamics of antibody response following years post vaccination in three volunteers have demonstrated significant antibody proliferation and accumulation of somatic hypermutations on isotype-switched antibodies (Laserson et al., 2014). Magnitude of isotype switching from IgM to IgG or IgA has been compared between two different vaccines (Jiang et al., 2013) and lineage analysis and comparison between repeated yearly vaccinated samples reveal signatures of memory B cell activation (Vollmers et al., 2013). Most studies on antibody responses include analysis on breadth and diversity of antibodies induced by vaccination and TLR adjuvants have been found to expand the B cell repertoire following malaria vaccination (Wiley et al., 2011). The heritability of antibody responses post-dose of vaccines has been investigated from identical twin pairs and IgM instead of IgG B cells are more similar between twins and show higher level of heritability (Wang et al., 2015). The aging effect to vaccine responses has also been studied (Jiang et al., 2013). Currently, most studies on vaccine effects have focused on antibody responses, while Becattini et al. have revealed that CD4+ memory T cells induced by distinct vaccines and pathogens have functional heterogeneity and multiple fates post priming (Becattini et al., 2015).
Taken together, the application of NGS to the TCR/BCR discovery and identification has facilitate its broader use as biomarkers in the disease prognosis and monitoring. Furthermore, people include us are trying to construct large database for disease associated repertoire and keen to applied it to the disease prediction, subtyping and prognosis. As for the antibody discovery and the vaccine development, each of them are tens-of-billion-dollar industry and NGS will definitely demonstrate its bright future in these fields.

4.3 challenges and problems to be solved

Though NGS has demonstrated its power to probe the tremendous diversity of TCR and BCR repertoire, challenges still remain ahead for its usage in wider research and clinical applications. As stated earlier, various experimental approaches have been adopted to construct the VDJ libraries before the NGS sequencing. Both genomics DNA (gDNA) and mRNA has been reasonably selected as the research material in various studies. Whether to choose gDNA or mRNA to analyze depends on the aim of the investigation. gDNA of TCR and BCR exists in single copy in the T and B cells, thus represents the cellular proportion and heterogeneity of T/B cell population. Therefore, gDNA is suitable for calculating the proportion of antigen-specific or specific clone of T and B cells, such as vaccine-specific clones or leukemia/lymphoma clones, and studying the functional/phenotypic evolution of specific TCR/BCR clonotypes. On the other hand, as the expression of TCR/BCR varied significantly among different types/status of T/B cells, such as naïve cells and plasmablasts, the quantification of TCR/BCR mRNA would be more correlated with the cell function/activation. Technically, gDNA extracted from a T/B cell population usually contains the pre- and post-recombined VDJ fragments, and the non-productive VDJ rearrangements, which would serve as the background to interfere with the proper amplification of the TCR/BCR repertoire. Nevertheless, gDNA would always be easier to access than the mRNA, including storage and shipment. For gDNA, multiplex PCR on V/J
genes has mostly been adopted, while RACE method and multiplex PCR on V/C genes have often been selected as experimental approaches in different studies for mRNA. Different experimental approaches harbor their own characteristics and bias, and the current divergent selection of methods in different labs makes the published results difficult to compare and do meta-analysis. Consequently, it’s critically required to get a deeper understanding of the characteristics of experimental approaches, and we are able to provide a comprehensive evaluation and comparison of the two commonly used approaches in this paper.

Another problem in this field is to distinguish the biological variations from the errors and bias introduced in the different steps. The improved sensitivity of NGS sequencing to read the individual sequences would also amplify the errors. These errors could be derived from the reverse transcription, the PCR amplification and the NGS sequencing. The latter two are the major sources to introduce various types of chimeras and nucleotide errors. The PCR process is carried along the whole pipeline, including amplification of VDJ rearrangement (and even in 5’ RACE), and the sequencing libraries construction. the mis-annealing of PCR primers to the templates could introduce unspecific amplification and the mis-incorporation of PCR templates would introduce chimeric artifacts. In addition, The DNA polymerase will introduce certain extent of amplification errors in every PCR cycles. All these chimeras and nucleotide errors will accumulate with the augment of PCR cycles. Hence some PCR errors incorporated in the early cycles could be retained in relative higher frequencies in the final NGS data.

NGS sequencing errors have been an issue since the very beginning. Currently, several platforms have been utilized to sequence the immune repertoire, including the Roche 454, Illumina Hiseq and Miseq, and the Life PGM and Proton. These techniques constitute their own characteristics in error rate and pattern, and deserve separate discussion. Pyrosequencing-based technologies including Roche 454, Life PGM and Proton, are usually enriched with indel
errors. These indels arise at frequencies around $5 \times 10^{-3}$ and are often enriched in the nucleotide positions with polymers. Algorithms have been developed according to this pattern to correct these polymer indels with certain success. However, the high frequency of indel errors sometimes is difficult to be distinguished from biological variations thus significantly reduces the rate and accuracy of alignment to the germline V/D/J genes. Alternatively, dye-labeled reversible terminator technology (Illumina HiSeq and Miseq) is dominated with base substitutions (Nguyen et al., 2011), and the overall error rate is lower than the pyrosequencing technology. Collectively, the Illumina platform is more suitable to sequence the TCR/BCR repertoire with the combination of lower base-calling error rate and cost.

Several strategies have been applied to minimize and correct these errors, including using high-fidelity enzymes, introducing random barcodes in the fragments to correct the errors, and apply various algorithms to do the in silico correction (Benichou et al., 2012; Zhang et al., 2015). Incorporating random barcodes to the DNA fragments was originally developed to detect the extremely low-frequency mutations, especially in the tumor tissue or circulating tumor DNA. By adding the highly diversified barcodes (the unique number of barcodes is designed to surpass the number of unique TCRs and BCRs in the studied repertoire), the set of VDJ fragments with the same barcodes could be grouped together and treated as the same template or molecule in the original repertoire before amplification and sequencing. Thereafter, the amplification bias and errors in the group could be corrected simultaneously.

Abundant analytical tools have been developed in this field to uncover the biological meaning of the diversified TCR and antibody omics data, as stated earlier. We have also developed our own in-house software – IMonitor. The challenge is that, with the prosperous development in this field, none of the existing analytical tools have satisfied all the needs and obviously surpassed the others, and the latter published tools have become superior in speed and
accuracy. Perhaps this is one of the reasons that researchers working in this area generally prefer to use custom-designed bioinformatics pipelines. Therefore, the standardization and open-resource computational tools is the only solution to facilitate the reproducing and meta-analysis of results generated from different laboratories, and this is in urgent need with the area mature. As stated, interchangeable and standardized data formats deposited in a central database open to public, validated open-resource algorithms for data analysis and standards for experimental description analogous to the microarray experiment could be eventually beneficial to the application in this area (Georgiou et al., 2014).

5 BCR repertoire analysis in B- acute lymphoblastic leukemia (ALL) and TCR investigation of TILs in breast cancer

The main aim of this manuscript is to discuss the T and B lymphocytes in cancer. T cell or B cell derived lymphoblastic leukemia is an important and prevalent blood cancer, in which these lymphoid cells are direct disease-causing malignant cells. Sequencing the immune repertoire including the malignant cells before and after the diagnosis can help identify the leukemic cell clones, track the minimal residual disease dynamics and monitor the evolved clones. Besides the blood cancer, solid tumor such as breast tumor cells also interacts intimately with the TILs in the tumor microenvironment. Various sections of T lymphocytes, including CD4+, CD8+ and regulatory T cells, play pivotal roles in elimination, equilibrium, and escape of tumor cells by the immune surveillance. Therefore, investigating and understanding the TCR repertoire of these TILs in breast tumor, and their relationship with the adjacent tissues, such as draining lymph nodes and distant normal breast tissues, become fascinating.

5.1 Minimal Residual Disease and clonal evolution analysis of leukemic BCR in ALL

As stated above, one important field for translational research and medicine for
TCR/BCR NGS sequencing is the minimal residual disease (MRD) detection for hematology including T/B cell derived leukemia and lymphoma. The sensitivity and prognosis use for NGS in comparison with the traditional techniques such as flow cytometry and Allelic-specific PCR have been recurrently investigated ever since the beginning of this field (Boyd et al., 2009; Logan et al., 2011; Pui et al., 2015; Pulsipher et al., 2015; Wu et al., 2012). NGS MRD detection has been demonstrated to correlated well with previous techniques but with much higher sensitivity, and this improvement showed better prediction of remission and relapse. Recently, various subtypes of lymphoma has been applied for MRD detection by NGS, and the plasma has exhibited superior detection power than the cellular repertoire (Kurtz et al., 2015; Weng et al., 2013).

However, though the technological sensitivity and reproducibility of NGS based MRD detection have been validated in a handful of studies, there are several unanswered questions regarding the superiorities of the new technique, before the translation can take a step further. 1. Whether the higher sensitive detection of the MRD positive patients correlate with a higher chance of relapse and shorter survival? 2. Will the malignant lymphocyte clones undertake evolution in their cell receptor sequences (TCR/BCR), and whether the evolution occurs under the pressure of therapy, and correlates with relapse? 3. How is the immune reconstruction benefits the patients?

Regarding to the first question, studies have been undertaken to correlate the clinical outcomes of benefit to higher sensitivity, and the data seem in supportive of the positive correlation and the benefit (Logan et al., 2014; Pui et al., 2015). However, prospective clinical trials with larger sample size are still required for clinical validation. As for question three, the repertoire sequencing facilitate the clinicians to evaluation the immune reconstruction after transplantation or other therapies, but limited studies have pursued to quantify and qualify the immune reconstruction with the NGS data. My investigation of child B-ALL not only tried to demonstrate the higher sensitivity of NGS MRD, and compare it with flow cytometry, but also evaluated the emergence and dynamics of evolved IGH clones
during treatment, aiming in providing contributions to question two.

5.2 TCR repertoire in tumor infiltrating T lymphocyte of breast tumor and adjacent tissues

Besides B cells in lymphoblastic cancer, my interest extended to infiltrated T cells in breast cancer. Breast cancer has become one of the most deadly disease worldwide, and has brought significant personal and social burden to Chinese women. Though surgeries and targeted therapies have profoundly improved the clinical outcomes, more precision diagnosis and treatment are required to further increase the survival, especially for the late-stage breast cancer. The TILs in breast cancer has demonstrated promising biomarkers for the prognosis, but the whole profile of their TCR repertoire is not yet investigated.

6 The significance of current study

Four main parts have been included in the content of this paper. Firstly, to set up a effective and unbiased technological platform to investigate the TCR/BCR repertoire with the NGS data, we have designed our own multiplex PCR system with repetitively optimized gene-segment specific primers, and have thoroughly evaluated and compared it with another widely-used approach of 5′RACE, to demonstrate that both methods have their own characteristics and high fidelity to amplify and reflect the clonal diversity of the repertoire. This study stands out as the first attempt of its kind and should be of interest to the researchers in this field. secondly, a high efficient analytical pipeline we have developed is introduced to deal with both TCR and BCR data for numerous applications. Thirdly, we have adopted the above validated experimental and bioinformatic pipelines to investigate the BCR repertoire of child B-ALL, to compare the detection of MRD with the flow cytometry result, and to evaluate the clonal evolutions of the malignant B cells. Lastly, we used our pipelines to characterize the TCR repertoire of the infiltrated T cells in breast tumor, as an example of solid cancer, and to investigate the heterogeneity of TILs, the relationship with adjacent tissues, and
uncover the associations with clinical features. This is one of the first studies of its kind.
This is a comprehensive work including method development and biological investigation. We focused on cancer immunology, but the pipelines we developed can be applied to answer broad scientific questions in T/B cell immunology.
List of articles related to the topic of this thesis


4. Ting Wang, Changxi Wang, Jinghua Wu, Chenyang He, Wei Zhang, Jiayun Liu, Ruifang Zhang, Yonggang Lv, Yongping Li, Xiaoqing Zeng, Hongzhi Cao, Xiuqing Zhang, Xun Xu, Chen Huang, Ling Wang and Xiao Liu. (2016) The different T-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues. Cancer Immunology Research. DOI: 10.1158/2326-6066.CIR-16-0107 (corresponding author)
List of articles not related to the topic of this thesis and completed during the PhD career


4. Xinyang Li, Xiaobo Duan, Kai Yang,…Xiao Liu*, Wen Tan*. Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. PLOS ONE. 2016 Sep 2;11(9):e0161801. doi: 10.1371/journal.pone.0161801. eCollection 2016. (Joint Correspondence)

5. Xiao Liu, Zesong Li, Zheng Su, et al., Novel Y-chromosomal microdeletions associated with non-obstructive azoospermia uncovered by high throughput sequencing of sequence-tagged sites (STSs). Scientific Reports. 6:21831 · February 2016. DOI: 10.1038/srep21831


integration in hepatocellular carcinoma. Nature Genetics, doi:10.1038/ng.2295, 2012 (co-first author)


Results

Part One. Optimization, evaluation and comparison of experimental approaches

Result summary of Article 1:

1. A comprehensive comparison between multiplex PCR and 5’RACE methods of NGS-based immune repertoire study, where TCR beta chain repertoire sequencing was adopted as the example, was conducted in different aspects and by various materials. In specific, two types of “gold standard” samples and samples from healthy individuals were used;

2. High reproducibility of both methods were achieved, and comparison on V/J gene usage revealed that the gene usage were similar for more than 60% of the genes. Together, relatively high proportion of TCR clones were recovered by both methodologies, considering their unique clonotypes and frequencies;

3. There were some discrepancies between the two approaches in terms of the V/J gene usage and multiplex PCR recovered less V-J pairings than 5’RACE, however, the variability was smaller than the biological variability. In contract, The effective data rate for 5’RACE was smaller than multiplex PCR.

4. In summary, both experimental approaches showed high reproducibility, limited technical bias, and are adapted to different applications of immune repertoire research. For instance, 5’RACE is suited for RNA material and species without comprehensive annotations of the TCR/BCR germline gene sequences. Multiplex PCR, on the other hand, is appropriate to study human gDNA and RNA samples with different set of multiplex primers.
Part Two. Development of the analytical toolset “IMonitor”.

Result summary of Article 2:

1. We have introduced the systemic design and development of IMonitor. The whole pipeline includes four steps: raw data processing, VDJ annotation, structural determination and visualization. In the workflow, the alignment-realignment procedure ensures the accuracy of VDJ and structural annotation and serves as a critical step;

2. We evaluated and validated the tool using the simulated in silico data and real public data downloaded from the database. Compared with other existed tools such as HighV-QUEST, Igblast and Decombinator, IMonitor outperforms them in accuracy in almost all chains in both TCR and BCR. For the public data, our toolset also provide superiority in accuracy.

3. IMonitor also provide a solution to minimize PCR bias and correct PCR and sequencing errors. With all these aspects, we utilized IMonitor to visualize sequencing data in cases, such as MRD detection in ALL patients.

Part Three. BCR repertoire analysis in blood cancer: Minimal residual disease detection and evolved IGH clones analysis in acute B lymphoblastic leukemia using IGH deep sequencing

Result summary of Article 3:

1. We recruited 51 pediatric B-ALL patients in this study. Bone marrow (BM) samples were collected upon diagnosis, and in different time slots after chemotherapy. We also collected the peripheral blood (PB) from 15 patients upon diagnosis for comparison. BCR repertoire sequencing from extracted genomics DNA was performed and the data were analyzed. For all the samples, the flow cytometry (FCM) was conducted as the traditional method to evaluate the MRD,
and was compared with NGS.

2. We identified in 92.2% of patients (47/51) the leukemic IGH CDR3 according to the criteria. In total, 77 leukemic clones were identified, with some samples being poly-clonal. Interestingly, we found the number of leukemic clones seem to associate with the disease risk. More high risk (HR) patients had two and more disease clones, and more standard risk (SR) patients tended to be mono-clonal. We also found only about one third of leukemic clones were functional, in accordance with the fact that most pediatric B-ALLs were originated from immature precursor B cells.

3. Replication result showed that the MRD detection from our method had very high reproducibility. Comparison between the result from FCM and NGS showed that, all the MRD positive results from FCM were detected by NGS with similar MRD levels. Twenty-eight samples were MRD negative in FCM, but MRD positive with NGS, which demonstrated its higher sensitivity.

4. We also compared the detection from PB and BM, to investigate whether PB could serve as non-invasive biopsy for MRD detection. We found the correlation of MRD levels between PB and BM was as high as 0.786, however, the MRD levels in the PB were relatively lower than in BM in the diagnostic samples, which could result in false negative in some cases.

5. An important part of this investigation is the analysis and monitor of the evolved IGH clones during the chemotherapy. We identified a significant number of evolved IGH clones arising from V gene replacement during therapy. The sequence analysis of the evolved IGHs revealed that, these clones were mainly produced from the ancestral leukemic clones and underwent continuous V gene replacement. We didn’t find the correlation of the frequencies of the evolved clones with the clinical features. However, we did find the dynamic change of the evolved IGH clones during chemotherapy, which may reflect the selection pressure.
Part Four. TCR repertoire analysis in solid cancer: investigation on infiltrated T lymphocytes in breast tumors and adjacent tissues

Result summary of the Article 4:

1. We studied the TCR repertoire of infiltrated T cells in 16 sets of breast tumors and matched draining lymph nodes and distant normal tissues, and utilized another sample set of 49 patients for validation. We found the TCR diversity was lower in tumor and normal tissues, compared to draining lymph nodes. The infiltrated lymphocytes in breast tumors had preferential usage of V and J genes for their TCR sequences.

2. Comparison of TCR repertoire between tumor and matched lymph nodes revealed that, a large proportion of tumor infiltrating T lymphocytes could be detected in draining lymph nodes, and experienced significant clonal expansion in tumors. Additionally, more TCRs in tumors were expanded (with frequency of more than 0.1%) than they were in lymph nodes, implying their activation state.

3. Interestingly, we found the lymph node ratio (LNR) is positively correlated with the number of infiltrated T cells in tumor, and significantly more expanded clones were present in the LNs with high LNR, than they were in the node-negative patients. These findings were reported for the first time.

4. When correlating the TCR repertoire with the clinical features of the patients, we found luminal-like breast tumors had higher similarity of TCR profile between tumor and normal tissues, than the basal-like subtype. Furthermore, we identified a set of public intra-tumor clones observed in multiple patients, which could arise from both mechanisms of combinatory bias and convergent recombination.
Discussions and Perspective

The diversity of immune repertoire lays the foundation for the T and B lymphocyte to recognize a variety of antigens and protect the invertebrate’s body through adaptive immunity. The traditional tools to study it are quite limited. The emergence of massive paralleled sequencing provides us a unprecedented opportunity to profile the heterogeneity of immune repertoire. However, abundant issues remain regarding technology. Firstly, the material used. The cells/tissues we collected for immune repertoire study usually represent a minor subset the entire repertoire. If spatial heterogeneity exists, such as in tumor microenvironment, sampling could be a serious problem to interpret the biology. In my studies, I include both the peripheral and tissue-resident immune repertoire as research targets. In article 1, I added biological replications and found the reproducibility of TCRb repertoire reached as high as 0.98 to 0.99. This demonstrate at least for peripheral TCRb repertoire, the sampling bias is minimal. However, our other data show that the biological reproducibility of BCR repertoire, which is usually much more diverse than TCR repertoire due to somatic hypermutations, is not that high. The spatial heterogeneity of tissue-resident immune repertoire, such as the TILs, is an ongoing investigation and debate for different tumors (Emerson et al., 2013; Gerlinger et al., 2013; Mani et al., 2016). Though breast tumor reveals a property of spatial homogeneity for the TCR repertoire in its microenvironment (Mani et al., 2016), we still collected spatially distinct tissues and pooled them in article 4, to make a better representation for the sampling. Using genomic DNA or mRNA is both a biological and methodological choice. To quantify the MRD level, which is equivalent to the proportion of leukemic B cells in total B cells (or in total nucleated cells), genomic DNA is utilized in article 3 and in most other relative articles. In article 4, a major aim of the study is to compare the composition of infiltrated T cells in adjacent tissues, therefore genomic DNA is selected to prevent the effect of transcription abundances. RACE experiment can only start from mRNA material, in article 1, to compare the performance of RNA and
MPCR, we have to choose mRNA. Amplification bias and amplification/sequencing errors have always been technological obstacles in immune repertoire research field. To understand the true biological variation/heterogeneity, a optimized and reliable pipeline to produce data which faithfully represent repertoire composition of the input biopsies, is required. In this regard, we devote to design and optimize a MPCR method, in which multiplex primers are carefully designed, tested and optimized. We aim to equalize the primer efficiencies and reduce the non-specific amplification. Afterwards, we test it in pooled synthetic templates, and spiked-in samples. The V gene amplification bias is acceptable, and faithful representation of the abundance is revealed for the spiked-in controls. These results strengthen our confidence and compel us to compare this MPCR method to the 5’RACE pipeline, which is commonly known as V-gene bias free. Our detailed head-to-head comparison provide real data for some long-lasting opinion and debate, such as the data utilizing efficiency. Overall, our data demonstrate pros and cons in both methodologies, in terms of bias, data efficiency and convenience. Article 1 also provide evidences for the reliability of our MPCR method, which is utilized in article 4.

Regarding technology in the immune repertoire research, three directions are under extensive exploration. Firstly, how to retain the chain pairing information. The translation of therapeutic antibodies and T cells must contain pairing information for both chains in the immunotherapy, no matter as drugs or cell therapy. Linkage RT-PCR has been developed to retain the chain pairing information (DeKosky et al., 2013), however, more accessible and convenient methods are still required to be applied in routine practice. Secondly, single-cell analysis combining genetic and transcription information of both the receptor gene and other genes, has recently demonstrated its powerfulness in studying the heterogeneity and pathogenesis of immune cells in various immune microenvironment, such as the immune cell atlas of TILs. TCRs/BCRs will serve as proxies to track the dynamics, expansion and lineage of the cells, while other information, such as whole transcriptome, will reflect the cells’ phenotypes and eventually help explain the biology. Currently, I have developed
methods to extract/assemble TCR/BCR sequences from the single-cell RNA-seq data (scRNA-seq) of the T/B cells, and I am working on deciphering the heterogeneity of immune cells in tumor microenvironment by the high-throughput scRNA-seq approach. This is a yet elusive area underscoring fascinating biological findings. Lastly, identifying the un-identifiable, which is rare clone in the complex background of the repertoire, requires more accurate and error-prone approaches. I have utilized a UID (unique molecular identifier) experimental approach in another study of mucosa immunology. The UID approach can cluster the reads with errors under the same random barcode, and help correct the errors in the cluster. Besides, appropriate algorithm can be quite useful to reduce the errors.

In IMonitor, we also incorporate the error correction step. Firstly, reads containing low-quality bases are mapped to the high-quality reads, and the low-quality bases are corrected; Secondly, since errorous bases are enriched in the low abundant sequences, these sequences are also mapped to the relatively higher abundance sequences, and the mismatches are corrected according to the abundant sequences. These steps ensure the easiest errors to be corrected. However, the pattern of errors should be analyzed before errors can be further minimized. For instance, different pattern of errors are produced in the different sequencing platform. Sequence reads from Roche 454 and Proton are always enriched with homologous insertions and deletions, rather than mismatches. Therefore, the error correction algorithm should be designed accordingly.

We provide examples for immune repertoire research in Cancer discovery. Our method and analysis on B-ALL can be easily extended to study T cell leukemia, or chronic lymphoblastic leukemia (CLL). For T cell derived disease, alpha/beta and gamma/delta T cells should both be targeted, as the malignant cells could be derived from both. Similar methodology could be followed in T/B cell derived lymphoma. The clinical translation of MRD detection by NGS is still on the way. Currently, large scale multi-center trials are being conducted and have been reported in a number of studies (Rawstron et al., 2016). The advantages of NGS approach, such as tracking the clonal drift (sequence change of leukemic clones), identifying the emergence of newly evolved clones, and monitoring the reconstruction of the immune diversity, are only
validated in the initial proof-of-concept stage, and warrant further large scale validation. Moreover, the clinical benefits of these advantages, such as lower incidence of relapse, improved survival, should be validated in multi-center prospective clinical trials.

The IgH clonal evolution in B-lineage leukemia is driven by both the V-gene replacement (also known as secondary recombination) and somatic hypermutations. The later contributes more significantly in CLL. In ALL, however, its contribution is difficult to assess due to the interference of sequencing errors. In the data of our ALL patients, we also identified the evolved IgH clones in the same lineage with the leukemic clone with plenty of mutations. However, the lineage sizes were relatively small, and their distribution patterns were not significantly discrepant from the pattern of sequencing errors. Therefore, we only discussed the evolved IgH clones derived from V-gene replacement. Interestingly, we found four leukemic clones were derived apparently due to V replacement from other leukemic clones, which demonstrated the clonal evolution was an important path for the generation of leukemic clones. I anticipate a lot more interesting discoveries being made after intensive investigations on this research field. The number and frequencies of the evolved clones in the process of treatment, may rather be regulated under the balance of tumorigenesis and therapy than be stochastic. Involving more longitudinal samples, combining with other technologies such as single-cell analysis, and extending analytical targets to other genes than IgH sequences, will benefit a lot to this research purpose.

While the blood cancer involves the peripheral circulating system, primary breast tumor cells have a intimate contact with their surrounding immune cells within the microenvironment. In our study, we validated the hypothesis that the surrounding T lymphocytes show a expanding feature and skewness of V,J segments usage, implying the restriction of their TCR repertoire. Surprisingly, we found a certain extent of resemblance of TCR repertoire between tumor and adjacent normal tissues. This might be either due to direct migration of T cells from tumor to normal, or sporadic migration of tumor antigens to the normal tissue to further educate the tissue resident T cells. The TCR resemblance of tumor and normal in luminal-like tumor is higher
than it in basal-like tumor, provide a direct evidence that tumor subtypes have different impacts on their immune environments. The limitation of the current study is the lack of molecular profile of the tumors, thus we are unable to bridge the gap between the tumor and immune cells. In future studies, by large scale of population sequencing of both tumors and immune repertoire, we aim to construct a specificity database between tumor neo-antigens and TCRs/BCRs. This will be fulfilled by data statistics and functional validation. Plenty of applications, including tumor prognosis monitoring, and early prediction, will become realistic with this weapon. The hypothesis of tumor early prediction by tracking the content and extent of circulating tumor DNA (ctDNA), is very attractive these days. However, in my view, this approach should be combined with evaluating the tumor specific T/B cells in the peripheral repertoire, as the early malignant cells could be eradicated by the host immune cells and not necessarily lead to cancer. Other factors such as the expression of PD-1/PD-L1 genes also contribute to the tumorigenesis process. Taken together, in the future, the fascinating idea of predicting cancer in a very early stage will only be possible when the involvement of monitoring immune system is considered.
Conclusions

In this thesis, I provide a full solution for immune repertoire analysis, including validated experimental pipelines and a in-house developed analytical toolset. I devoted to comprehensively evaluate and optimize the pipelines. With the help of validated methodology, I provide analysis on BCR repertoire of B-ALL, and TCR repertoire of TILs in breast cancer. Interesting discoveries were provided with aim at aiding understanding cancer immunology and immunotherapy, including evolutionary analysis on BCR sequences of leukemic B cells, and the profile and relationship among breast tumor and adjacent tissues. For B-ALL, the evolution on IGH sequences of malignant B cells continuously arise due to V gene replacement. This evolution is very likely accumulated under the pressure of therapy, and could give hints to clinical outcomes. For breast cancer, the infiltrated T cells in tumor have a unique profile than they are in the adjacent compartments. The migration of these T lymphocytes between tumor and normal tissues is associated with the subtypes of breast tumor, and node positivity is correlated with the extent of T cell infiltration in tumor and expansion in lymph nodes. These findings definitely provide new insight to the understanding of TILs in breast cancer.
References


lymphocytes in breast cancer. Breast cancer research : BCR 18, 78.


antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome research 21, 790-797.


Appendix

Paper 1

Paper 2

Paper 3

Paper 4
Ting Wang, Changxi Wang, Jinghua Wu, Chenyang He, Wei Zhang, Jiayun Liu, Ruifang Zhang, Yonggang Lv, Yongping Li, Xiaojing Zeng, Hongzhi Cao, Xiuqing Zhang, Xun Xu, Chen Huang, Ling Wang and Xiao Liu. (2016) The different T-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues. Cancer Immunology Research. DOI: 10.1158/2326-6066.CIR-16-0107
Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire

Xiao Liu1,3☯, Wei Zhang1,2☯, Xiaojing Zeng1,2☯, Ruifang Zhang1, Yuanping Du1, Xueyu Hong1, Hongzhi Cao1, Zheng Su1, Changxi Wang1, Jinghua Wu1, Chao Nie1, Xun Xu1, Karsten Kristiansen3*

1 BGI-Shenzhen, Shenzhen, 518083, China, 2 Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen, 518083, China, 3 Department of Biology, University of Copenhagen, Copenhagen, Denmark

☯ These authors contributed equally to this work.

* kk@bio.ku.dk

Abstract

High-throughput sequencing has recently been applied to profile the high diversity of antibody/B cell receptors (BCRs) and T cell receptors (TCRs) among immune cells. To date, Multiplex PCR (MPCR) and 5' RACE are predominately used to enrich rearranged BCRs and TCRs. Both approaches have advantages and disadvantages; however, a systematic evaluation and direct comparison of them would benefit researchers in the selection of the most suitable method. In this study, we used both pooled control plasmids and spiked-in cells to benchmark the MPCR bias. RNA from three healthy donors was subsequently processed with the two methods to perform a comparative evaluation of the TCRβ chain sequences. Both approaches demonstrated high reproducibility (R² = 0.9958 and 0.9878, respectively). No differences in gene usage were identified for most V/J genes (>60%), and an average of 52.03% of the CDR3 amino acid sequences overlapped. MPCR exhibited a certain degree of bias, in which the usage of several genes deviated from 5' RACE, and some V-J pairings were lost. In contrast, there was a smaller rate of effective data from 5' RACE (11.25% less compared with MPCR). Nevertheless, the methodological variability was smaller compared with the biological variability. Through direct comparison, these findings provide novel insights into the two experimental methods, which will prove to be valuable in immune repertoire research and its interpretation.

Introduction

T cells and B cells are the most important components of the adaptive immune system. These cells enable the host to resist a vast array of potential pathogens via the production of diverse T cell receptor (TCR) or B cell receptor (BCR)/immunoglobulin (Ig) repertoire. The nucleotide sequences that encode the TCR or BCR are produced via somatic rearrangement of Variable, Diverse (D) and Joining (J) segments, as well as a set of non-template nucleotide insertions and
deletions at the V-(D)-J junction region. Together, these processes contribute to the generation of diversity in TCRs or BCRs. Most of the variation in each chain is located in complementary-determining region 3 (CDR3), which is encoded by the V(D)J junction. The CDR3 of the TCR β-chain (TRB) region, defined by the IMGT collaboration [1], begins with the second conserved cysteine encoded by the 3' portion of the V gene segment and ends with the conserved phenylalanine encoded by the 5' portion of the Jβ gene segment. Besides, Ig introduces point mutations during the process of affinity maturation of B-cells in the germinal center, which further increases the diversity of the Ig repertoire. Theoretically, the potential diversity of TCRs-αβ (TCR α chain and β chain) may attain $10^{18}$ for humans according to the combinatorial mechanism, whereas the potential diversity of the B-cell repertoire is even higher considering the somatic hyper-mutation (SHM)[2, 3]. Wang et al. estimated a sample with $0.47 \times 10^6$ TCR-α (TRA) unique nucleotide sequences and $0.35 \times 10^6$ TCR-β unique CDR3 nucleotide sequences [4]. Jacob Glanville et al. estimated that a human donor has at least $3.5 \times 10^{10}$ unique IgM sequences using a capture-recapture method [5].

As a result of the high diversity of the immune repertoire, the limited output of Sanger sequencing and the low resolution of GeneScan provide only limited visualization of this process. The next generation sequencing (NGS) platforms, such as Roche 454 and Illumina HiSeq, are ideally suited to derive a holistic snapshot of the process and extensively characterize and visualize the complexity and plasticity of the TCR and BCR repertoires [4, 6–9].

Irrespective of the sequencing platform, it is imperative to enrich the V(D)J rearranged genomic or transcript sequences prior to sequencing, using methods such as multiplex PCR (MPCR), Rapid Amplification of cDNA Ends (5'RACE), linear PCR or sequence capture. He et al. have used a sequence capture method denoted IgCap to enrich the somatically rearranged DNA sequences from lymphomas [10]. However, IgCap captures all fragments that contain relevant Ig heavy chain (IgH) gene sequences, whereas it does not simply capture the rearranged fragments; thus, it provides limited sensitivity and resolution. The linear amplification method begins with the V gene primer set, followed by the C (constant)-gene primers in the second direction extension [11]. It requires more starting material and has less amplification bias, in theory. However, the strong bias of linear amplification was also observed in our test (data not shown).

Of all of these methods, MPCR and 5'RACE are the two most commonly used approaches because of their obvious advantages compared with the other approaches. MPCR utilizes a pool of primers that targets all V and J germline genes (or C genes) to amplify the whole V(D)J rearrangements or specifically the CDR3 regions [8, 12–14]. Although MPCR may cause PCR bias because of the inconsistent reaction efficiencies of multiple primers [2], it requires only one PCR step prior to library construction and is suitable for both genomic DNA and RNA materials, which makes it a convenient and cost effective approach.

The 5'RACE is another commonly used method [7, 15, 16] for investigating the immune repertoire. It only utilizes one primer set that targets the known C-gene region of the mRNA transcripts. The second general primer is synthetically added to the 3' end of the cDNA using terminal deoxynucleotidyl transferase (TdT) and cCTP. This method introduces negligible PCR bias compared with other methods [2] although it may only be used for RNA samples.

Despite their wide-ranging applications demonstrated by various studies, a direct and systematic comparative evaluation between these two methods is lacking. Benichou et al. have provided comments regarding published papers [2]; however, limited head to head comparisons of the two methods have been published. In this study, we designed experiments to comparatively evaluate these two methods. First, pooled-plasmid controls and cell populations spiked with known TCRβ rearrangements were designed to benchmark the MPCR bias (Fig 1A); limited bias was identified. Second, we simultaneously performed MPCR and 5'RACE
using identical RNA samples extracted from the peripheral blood of 3 individuals to profile the TCRβ repertoire. We also prepared a replicate for MPCR and 5’RACE (Fig 1B). We compared the effective data, V and J gene usage, sequence diversity, and V-J pairing, in addition to the level of clonal sequences. Finally, saturation analyses were conducted for different samples (Fig 1B). Overall, the respective features of these two methods provide deeper insights in the investigation of the immune repertoire.

Materials and Methods

Blood Samples

Samples of peripheral blood mononuclear cells (PBMC) from 2 healthy men and 1 healthy woman (S01, S02, and S03) aged 22 to 34 years were obtained with written informed consents. PBMCs were immediately isolated from each sample using Ficoll-Paque (GE Healthcare)
gradient centrifugation. RNA was isolated using Trizol (Invitrogen) according to the manufacturer’s specifications. The RNA concentration and sample integrity were determined on an Agilent Bioanalyzer (Agilent).

**Plasmid Pools**

Thirty-three plasmid clones were selected from the TA-cloning library of one TCR β-chain repertoire. Each plasmid was cloned with a specific TCRβ CDR3 region sequence. All TCRβ V and J subfamilies were included in the plasmid pools. One plasmid pool was prepared with equal moles of each plasmid, and the other two plasmid pools were prepared with different pooling ratios (ratios of 1:10, 1:100, 1:1000, and 1:10000). All three pools were replicated. The details regarding the different pooling ratios are shown in S3 Table.

**Spiked-In Samples**

Spiked-in samples were generously donated by professor Karen Cerosaletti [17]. In brief, five CD4+ T cell clones, previously reported as specific for GAD65, which had unique TCRβ CDR3 sequences, were spiked into CD4+CD45RA+ naïve T cells sorted from the fresh PBMC of a control donor. Three CD4+T cell mixes, which contained different amounts of the five CD4 + T cell clones as measured via flow cytometry, were prepared. In mix 1, 10⁵, 10⁴, 1000, 100 and 10 T cells of the five clones were doped in a background of 1 million naïve T cells. In mix 2, the same 1000 cells for each clone were doped, and 10, 100, 1000, 10⁴ and 10⁵ were doped in mix 3. gDNA and total RNA were extracted from the cells using commercially available kits. The details of the five clones are shown in S4 Table.

**MPCR Method**

We used MPCR primers from a published paper [18] to amplify the rearranged CDR3 regions of TCRβ, including 30 forward V primers and 13 reverse J primers. One to 3 μg of total PBMC RNA was subjected to DNase I digestion (NEB) for 10 min at 37°C to remove genomic DNA. The digestion reaction was terminated via the addition of 2 μL 1 mM EDTA for 10 min at 75°C. The product was subsequently purified via ethanol precipitation. First strand cDNA was synthesized using SuperScript I reverse transcriptase (Invitrogen) with a primer specific to the TCRβ constant region (5’>CACGTGGTCGGGGWAGAAGC<3’). First-strand cDNA was used as a template for multiplex PCR with an equimolar pool of the 30 TCR Vβ forward primers (the VβF pool) and an equimolar pool of the 13 TCR Jβ reverse primers (the JβR pool). The primer set is available upon request. The reaction conditions were as follows: 1×QIAGEN Multiplex PCR Master Mix, 0.5×Q solution, 0.2 μM VβF pool and JβR pool in a 50-μL volume. The cycling conditions were as follows: 95°C for 15 min, 25–30 cycles at 94°C for 30 sec, 60°C for 90 sec, and 72°C for 30 sec, plus a final extension at 72°C for 5 min. The PCR product was run on a 2% agarose gel; the band size between 110–180 bp was excised from the gel and purified.

**5’-RACE Method**

5’-RACE was performed using the 5’ RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (Invitrogen). First-strand cDNA was synthesized using a published TRBC (TRB C region) primer (5’>CACGTGGTCGGGGWAGAAGC<3’). The reaction conditions were as follows: 3–5 μg RNA, TRBC primer 2.5 pmole, 1×PCR buffer, 2.5 mM MgCl₂, 0.4 mM each dNTP, 10 mM DTT, and 200 units of Superscript II in a 25 μL volume. Extension occurred at 42°C for 50 min, followed by inactivation at 70°C for 15 min; 1 μl of RNase mix was subsequently added, followed by incubation at 37°C for 30 min to remove RNA. The cDNA was
then purified with a S.N.A.P. Column. The purified cDNA was subsequently used in the TdT-tailing reaction to add an oligo-dC tail in the 3’ end of cDNA using TdT and dCTP. The dC-tailed cDNA was directly amplified via PCR. PCR was performed using Taq DNA polymerase (TAKARA) in a 50 μL volume with a deoxynosine-containing abridged anchor primer (AAP) provided with the 5’ RACE system (5’-GGCCACCGTGGCATACTACGGGGIIGGGI<3’) and an equimolar combination of two biotinylated GSPs, (5’-ACACTTAATTACACGGGTGGGAACACCTTGTTCAGGT<3’) and (5’-ACACCTAATACGGGGTAACACGTTTTTCAGGT<3’), which contain 5’ PacI sites and were specific for TRBC1 and TRBC2, respectively[15]. The reaction conditions were as follows: 1×PCR buffer, 1.5 mM MgCl2, 0.2 mM each dNTP, 0.4 μM AAP, 0.4 μM biotinylated GSP, 5 μL dC-tailed cDNA, and 2.5 units Taq DNA polymerase. A 2 min denaturation at 94°C was followed by 30 cycles of 30 sec at 94°C, 30 sec at 55°C and 45 sec at 72°C, as well as a final extension at 72°C for 5 min. The PCR product was run on a 2% agarose gel; the band centered at 520 bp was excised and purified. Approximately 3 micrograms of purified PCR product were subsequently sheared using a Covaris S2 (Applied Biosystems). The parameters of the reaction conditions were as follows: 100 μL reaction volume, 12 cycles with a duty cycle of 10%, intensity of 5, and cycles per burst at 200 for 30 sec. Biotinylated fragments were then purified using 50 μL of Dynabeads M-270 Streptavidin (Invitrogen) prepared according to the manufacturer’s specifications. Washed and bound biotinylated fragments were subsequently cleaved with PacI [reaction conditions; 1× NEB buffer 1, 1× BSA, and 10 U PacI (NEB) in a 50 μL volume for 1 h at 37°C, followed by 20 min at 65°C]. The sample was run on a 2% agarose gel; the fraction from 150–190 bp was excised and purified.

Library Preparation and High Throughput Sequencing
The MPCR and 5’-RACE products were prepared for Hiseq 2000 sequencing according to the manufacturer’s protocol with modifications. The end repair was performed using T4 DNA polymerase, Klenow fragment of E.coli DNA polymerase I and T4 PNK in T4 PNK buffer (all Enzymatics) at 20°C for 30 min. Following a Qiaquick PCR purification (Qiagen), a deoxynucleotides were added to polished double strands using the Klenow fragment (3’-5’-exo-) in the presence of 200 μM dATP in 1×blue buffer (all Enzymatics) at 37°C for 30 min. Following MinElute PCR purification (Qiagen), Illumina PE adapters were ligated. The reaction conditions were as follows: 1×Rapid ligation buffer, 3000 U T4 DNA ligase (Rapid) (all Enzymatics), and 0.8 μM PE adapters in a 50 μL volume at 20°C for 15 min. The product was purified using 1.2 times the volume of Ampure XP (AGENCOURT) and eluted in a volume of 23 μL. The ligation product was subsequently amplified via PCR using Phusion HF Master mix (Finnzymes) with 0.2 μM of Illumina primers 1.0 and 2.0 in a 50 μL volume. The cycling conditions were as follows: a 2-min denaturation at 98°C was followed by 12 cycles of 20 sec at 98°C, 30 sec at 65°C, and 30 sec at 72°C, as well as a final extension for 5 min at 72°C. The PCR product was purified using 1.2 times the volume of Ampure XP (AGENCOURT) and eluted in a volume of 15 μL. Libraries with insert sizes of 100–200 bp were analyzed via Bioanalyzer analysis and QPCR and were stored at -20°C until use. Paired-end 100 bp sequencing was performed using an Illumina Hiseq 2000 system following the manufacturer’s instructions. Fluorescent images were processed into sequences using the Illumina data processing pipeline.

Data Analysis
Sequencing data were analyzed using IMonitor [18], which is briefly introduced as follows. First, it performs basic QC and filters the low quality reads of NGS data; it subsequently merges the cleaned pair end reads to be connected. Second, the merged data are used to perform a
BLAST [19–21] alignment to the V, D, J germline genes and alleles (IMGT, http://www.imgt.org/), respectively. After acquiring the BLAST alignment results, it performs a re-alignment for each result and selects the best V/D/J alignment for each merged read. Third, it filters the sequences in which the abundance (support reads) of the corresponding CDR3 nucleotide sequence is less than 5 and translates these sequences to protein sequences. Finally, it performs multiple statistics of the TCR data, such as V-J pairing, V/J usage, CDR3 sequence frequency statistics, and CDR3 length distribution.

Statistical Analysis

We used Pearson correlation coefficients[22] to measure the linear correlation between two variables. The analysis included MPCR and 5’RACE reproducibility evaluation (observed frequencies and expected frequencies), as well as the evaluation of the similarity of two samples using CDR3 frequencies. The statistical significance of the difference in the V/J frequency between the MPCR and 5’RACE samples was determined using paired t tests. The Chao1 has been used to estimate the target richness for individual-based data for ecology. Here, we used this method to estimate the maximum number of unique CDR3 amino acid sequences of the samples. Estimated values using the Chao1 algorithm and observed values were used to create the saturation curves [23, 24]. The Chao1 bias corrected method was quoted, and the formula was as follows:

\[
\hat{S}_{\text{choa1}} = S_{\text{obs}} + \frac{F_1(F_1 - 1)}{2(F_2 + 1)}
\]

\(S_{\text{obs}}\) represents the total number of observed clono-types in a sample; \(F\) denotes the number of clono-types (\(F_1, \) the number of clono-types detected one time. \(F_2, \) the number of clono-types detected two times).

This research was reviewed and approved by a duly constituted ethics committee (the Institutional Review Board on Bioethics and Biosafety of BGI).

Results

MPCR Bias Evaluation

**Plasmid pools.** One major concern for MPCR is its potential bias towards specific clones because of the inconsistent amplification efficiencies among the multiplex primers. To evaluate this issue, three-plasmid pools, one pool with an equal mole of each plasmid and the other two pools with different pooling ratios (S3 Table), were used to compare the expected and observed frequencies of the inserted TCR\(\beta\) CDR3 sequences. Each plasmid pooling experiment was replicated (Fig 2). The plasmid clones C-16 and C-17 were removed in the subsequent analysis because nearly none of corresponding reads were identified in all six pools. An average Pearson R\(^2\) (correlation coefficient) of 0.9247 was identified between the replications, which demonstrated an excellent stability for MPCR. Although some observed frequencies did not exactly match the expected pattern, the differences were not significant in all six pools. The average high correlation coefficient (R\(^2\) (average) = 0.9130) between the observed and expected frequencies demonstrated an overall low bias for MPCR in these plasmid-mix samples.

**Spiked-in samples.** Three spiked-in T cell populations with defined ratios of T cells of known TCR sequences were also adopted to evaluate MPCR. gDNA was extracted, amplified with MPCR, and sequenced for an average of 10 million TCR \(\beta\) sequences (S4 Table). Overall, the five spiked-in clones were identified in each of the three samples at levels comparable to the expected frequencies (Fig 3). Specifically, clones A and C exhibited high concordance between
the expected and observed frequencies in all three mixtures; clone B was slightly over amplified; clones D and G were nearly identical with the expected frequencies of 1:10 and 1:103 ratios; however, they were over-represented in the mixture of the 1:105 ratio. Furthermore, in the spiked-in experiment, 10 clone cells of one million were identified, which illustrated the high sensitivity of the MPCR method to track low frequency or rare clones.

Effective Data

RNA was extracted from the PBMC of three healthy donors and was subsequently divided into two parts to amplify the TCR β region. Together, 4 MPCR libraries and 4 5′RACE libraries, including the replicates, were sequenced with paired-end 100 bp using a HiSeq2000 (Illumina Inc.) The data statistics are shown in Table 1. Each sample was sequenced for 8–46 M paired-end reads. The insert sizes of some 5′RACE fragments were either too short (less than 100bp) or too long (more than 200bp) because of random shearing. The fragments less than 100bp were easily polluted by sequencing adapter, so the filtered data of 5′RACE samples (22.05% on average) were more than that of MPCR samples (2.69% on average). The fragments longer than 200bp could not be merged, so less sequences were merged for 5′RACE samples.

![Fig 2. MPCR bias evaluated by the mixed plasmid samples. Thirty-three plasmids were mixed in the sample with three different pooling gradients (a, b, and c represent plasmid mixes 1, 2, and 3, respectively, details in S3 Table). Their replications are showed in d, e, and f. The observed frequencies of the thirty-three templates were calculated and compared with the expected pooling frequencies. The numbers in the top left corner (b, c, e, and f) represent the Pearson correlation coefficients between the observed and expected frequencies.](https://doi.org/10.1371/journal.pone.0152464.g002)
Furthermore, the VJ alignment rate of each MPCR sample was greater than 98%, whereas it was less than 79% for each of the 5'RACE samples. It was because the sequences of 5'RACE samples included an approximately 30 bp C region, which left some sequences with very short or even no V gene nucleotides for identification. Most sequences were identified the correct ORF (open reading frame) for all samples. To decrease the sequencing errors, the sequences detected for less than 5 times (5 reads supported) were filtered. Overall, the average effective data rate for MPCR was 80.98%, whereas it was 47.48% for 5'RACE. With the exception of S02-R-2, the number of unique CDR3 nucleotide and amino acid sequences were all greater than $10^5$, and its number of S03-M and S03-R were several times more than the other samples because they were started with more cells and RNA amounts. Simpson's diversity index \cite{25} of the CDR3 AA (amino acid) sequences was calculated to quantify the distribution of diversity in each sample (Table 1).

![Graphs showing expected versus observed frequencies for 5 spiked-in clones in three mixtures.](image)

**Fig 3. Expected versus observed frequencies for 5 spiked-in clones in three mixtures.** Five CD4+ T cell clones (A, B, C, D, and G) were spiked into sorted CD4+ T cell populations at concentrations that spanned five orders of magnitude. TRB CDR3 was amplified via MPCR, and the observed clone frequency was calculated. The results demonstrated a high concordance between the expected and observed clone frequencies in three samples (Mix1, Mix2, and Mix3).
Reproducibility of MPCR and 5'RACE

To investigate the reproducibility of MPCR and 5'RACE, two separate aliquots (S01-M-1 and S01-M-2) of RNA from individual S01 were used to prepare MPCR sequencing libraries, whereas S02-R-1 and S02-R-2 from S02 were used for 5'RACE. All subsequent analyses were based on the same amount of effective reads for the replicated samples. Specifically, effective data were randomly sub-sampled to 6,880,594 sequences per library for the MPCR method and 11,023,329 sequences per library for the 5'RACE. The basic statistics are shown in S1 Table.

For MPCR, the data analysis indicated that the number of unique CDR3 was similar between the two replicates. In contrast, the two replicates of 5'RACE exhibited a greater difference (S1 Table), which may be attributed to the variable fragmentation in the 5'RACE experiment. The replicated samples for MPCR shared 69,634 unique CDR3 amino acid (AA) sequences, which corresponded to 78.63% and 74.93%, respectively, of their total sequences. For 5'RACE, 25,773 unique CDR3 AA sequences were shared, which represents 50.01% and 59.68%, respectively, of the total reads. Moreover, both MPCR and 5'RACE exhibited high reproducibility for CDR3 AA abundance, with Pearson correlation coefficients of 0.9907 for MPCR and 0.9878 for 5'RACE (Fig 4A and 4C).

To further evaluate the reproducibility, we calculated the consistency between the replicated samples. CDR3 AA sequences were ranked by frequencies in each replicate; we subsequently calculated the shared CDR3 number (accumulative shared number) and percentage (accumulative percentage) between the two replicates beginning with the top ranking clones, e.g., top 10 and top 100. For example, ranked CDR3 1000 (at x axis in Fig 4B) indicates 78.5% CDR3s were shared among the top 1000 CDR3s. The shared percentage of replicates slowly decreased along with the decrease in the CDR3 abundance for both MPCR and 5'RACE. In MPCR, the top 19 CDR3 sequences of the two libraries were identical with the exception that some sequences ranked differently between the two replicates. For the top 1000 sequences, up to 78.5% of the sequences were shared; While overall, approximately 43.48% of the unique sequences were shared (Fig 4B). In 5'RACE, there were 96% identical sequences in the top 30 CDR3 sequences.

Table 1. Statistics of the multiplex PCR and 5'RACE libraries.

<table>
<thead>
<tr>
<th>Sample</th>
<th>S01-M-1</th>
<th>S01-M-2</th>
<th>S02-M-1</th>
<th>S03-M-1</th>
<th>S01-R-1</th>
<th>S02-R-1</th>
<th>S02-R-2</th>
<th>S03-R-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total reads</td>
<td>8,126,815</td>
<td>14,588,545</td>
<td>12,612,786</td>
<td>39,754,077</td>
<td>24,555,544</td>
<td>24,450,175</td>
<td>24,204,014</td>
<td>46,409,300</td>
</tr>
<tr>
<td>Clean reads&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7,955,514</td>
<td>14,293,710</td>
<td>12,396,695</td>
<td>37,797,350</td>
<td>18,590,382</td>
<td>17,713,913</td>
<td>17,739,268</td>
<td>40,325,274</td>
</tr>
<tr>
<td>PE reads merged&lt;sup&gt;b&lt;/sup&gt;</td>
<td>7,934,499</td>
<td>14,251,076</td>
<td>12,365,441</td>
<td>37,447,896</td>
<td>16,755,290</td>
<td>17,713,913</td>
<td>15,704,110</td>
<td>37,861,325</td>
</tr>
<tr>
<td>VJ alignment&lt;sup&gt;c&lt;/sup&gt;</td>
<td>7,821,564</td>
<td>14,054,378</td>
<td>12,150,255</td>
<td>36,496,389</td>
<td>10,763,131</td>
<td>13,726,534</td>
<td>12,427,498</td>
<td>29,851,422</td>
</tr>
<tr>
<td>Correct ORF reads&lt;sup&gt;d&lt;/sup&gt;</td>
<td>7,439,792</td>
<td>13,448,043</td>
<td>11,122,904</td>
<td>34,838,414</td>
<td>9,585,671</td>
<td>12,638,442</td>
<td>11,460,936</td>
<td>28,265,147</td>
</tr>
<tr>
<td>Effective reads&lt;sup&gt;e&lt;/sup&gt;</td>
<td>6,880,594</td>
<td>12,664,905</td>
<td>10,398,513</td>
<td>27,823,441</td>
<td>9,585,671</td>
<td>12,019,778</td>
<td>11,023,329</td>
<td>26,067,794</td>
</tr>
<tr>
<td>Effective reads Rate</td>
<td>84.67%</td>
<td>86.81%</td>
<td>82.44%</td>
<td>69.99%</td>
<td>39.04%</td>
<td>49.16%</td>
<td>45.54%</td>
<td>56.17%</td>
</tr>
<tr>
<td>Unique CDR3 nucleotides&lt;sup&gt;f&lt;/sup&gt;</td>
<td>148,262</td>
<td>204,041</td>
<td>170,733</td>
<td>1,061,044</td>
<td>265,458</td>
<td>156,655</td>
<td>81,008</td>
<td>541,446</td>
</tr>
<tr>
<td>Unique CDR3 peptides&lt;sup&gt;g&lt;/sup&gt;</td>
<td>138,831</td>
<td>188,164</td>
<td>158,991</td>
<td>858,389</td>
<td>247,403</td>
<td>147,035</td>
<td>76,636</td>
<td>483,083</td>
</tr>
<tr>
<td>Simpson’s diversity index (CDR3)</td>
<td>0.9961</td>
<td>0.9949</td>
<td>0.9288</td>
<td>0.9878</td>
<td>0.9881</td>
<td>0.9885</td>
<td>0.9885</td>
<td>0.9828</td>
</tr>
</tbody>
</table>

<sup>a</sup>Clean reads refer to reads after filtering the low-quality sequences
<sup>b</sup>PE merged reads refer to paired-end reads merged filtered by the sequences with a stop codon, incorrect CDR3 length, and mapped to pseudo genes
<sup>c</sup>VJ alignment refers to reads mapped to V and J genes
<sup>d</sup>Correct ORF reads were determined to have the CDR3 region
<sup>e</sup>Effective reads were filtered by the low-frequency sequences
<sup>f</sup>Unique CDR3 nucleotide represents a non-redundant fragment (CDR3) of nucleotide
<sup>g</sup>Unique CDR3 peptide represents a non-redundant fragment (CDR3) of amino acids.

doi:10.1371/journal.pone.0152464.t001
of the two replicates, 68% in the top 1000, and approximately 22.96% overall were shared (Fig 4D). Overall, the reproducibility of MPCR and 5'RACE was very high, especially for MPCR.

**V and J Gene Usage Comparison between MPCR and 5'RACE**

There are 48 functional V genes and 13 J genes for the TRB, which contribute to the high diversity in the CDR3 region. In this study, the V and J usage of six samples were analyzed. The results are shown in Fig 5. Furthermore, the average frequency difference (AFD) between the 5'RACE and MPCR samples was calculated. Paired T tests were used to determine the difference between the two methods.

The usage of most V and J gene exhibited similar patterns without significant differences, including 32 V genes (66.67%) and 7 J genes (53.85%), indicating no overall significant bias in
the MPCR method. However, several genes, such as TRBV29-1 and TRBJ2-1, exhibited a usage variance between the two methods (p < 0.05, paired t test). Among these genes, 3 J had a higher frequency in MPCR compared with 5'RACE (AFD > 0), whereas the majority of V genes (14/16 or 87.5%) exhibited a lower frequency in the MPCR method (AFD < 0). This finding may be a result of the lower kinetics efficiency during the amplification process for primers that target these genes. Furthermore, we determined that the overall 72.92% (35/48) V gene frequency of MPCR was less than 5'RACE (AFD < 0, Fig 5), which was accounted for the bias of several abundant genes, such as TRBV29-1 and TRBV20-1.

CDR3 Sequence Comparison between MPCR and 5'RACE
The variations associated with specific V and J genes were identified in this study; however, it was unclear how this bias may impact CDR3 sequences. To investigate this issue, the CDR3
AA sequences were sorted by frequencies in descending order for the MPCR and 5’RACE data sets (Fig 6A, 6B and 6C). The consistencies of the two sorted data sets produced by the two different experimental methods were calculated. For the 3 individuals, the sequences in the top 100 unique CDR3 amino acid sequences of the two libraries had concordances of 24.0%, 30.0% and 40.0%, respectively. The overlapping rates of the unique CDR3 sequences were 21.07%, 17.35% and 17.07% for individual S01, S02 and S03, respectively, based on the entire dataset. S01 had a lower rate in the top 100; however, its overall rate was 4% higher compared with the other individuals, implying the sharing of identical sequences in lower frequency for the two methods.

The overlap rate in unique CDR3 AA sequences was approximately 20%; however, the overlap rate increased to an average of 55.42% when their abundances were considered (S2 Table). In particular, the overlap rates were 59.40% and 54.10%, respectively, for S01, 66.32% and 43.03%, respectively, for S02, and 55.98% and 53.69%, respectively, for S03. Thus, abundant clonotypes had a higher consistency in frequency between the two methods. The correlation was also calculated from 3 individuals (Fig 6D, 6E and 6F). The coefficient of the correlation in the abundance of CDR3 AA sequences between MPCR and 5’RACE sequencing for each individual ranged from 0.2817 to 0.6034 (0.4230 on average).
When we examined the V-J gene pairing covered by each sample, we determined that the 5′RACE samples contained nearly all 624 potential V-J pairs (48 TRBV * 13 TRBJ), whereas the MPCR samples, on average, lost 63 pairs (Fig 7). On average, the overlapping V-J pairing number for each sample was 560 (89.80%). The more diversified pattern of V-J pairing identified by the 5′RACE method indicated that it introduced less bias compared with MPCR. Overall, the correlation coefficient of V-J pairing abundance between the two methods ranged from 0.2777 to 0.7202.

Method Variation in Comparison to Biological Variation

The direct comparison of the two methods in the aspects of the CDR3 AA sequence and V-J gene pairing, especially the latter, clearly illustrated the existence of a method-specific pattern. To determine whether the method variation was more significant than the biological variation among individuals, we compared the method variation with the biological variation between two donors using the same methodology. Using 5′RACE or MPCR, the overlapping rate of the CDR3 AA sequences between two individuals was very low, i.e., less than 10% (S1C and S1F Fig), which was significantly less than the rate of the same sample with different methods (Fig 6A, 6B and 6C). Moreover, the correlation coefficient (less than 0.017) of the CDR3 sequences between two different donors (S1A and S1D Fig) was also far less than that of the same individual (Fig 6D, 6E and 6F). However, for the VJ pairing, the correlation coefficient between two different individuals, which ranged from 0.3431 to 0.5450 (S1B and S1E Fig), was slightly increased (Fig 7). This finding clearly indicated that although it had an influence, method specificity was not as significant as individual specificity in the TCR sequence level.

Saturation Analysis and Error Comparison

How the number of clonotypes tended to saturate with the increase in the sequencing amount in a TCR repertoire remained unclear, and the answer would be of interest to researchers.
Saturation curves of the samples using MPCR or 5’ RACE are shown (S2 Fig). Sequences were randomly subsampled at an interval of 0.5 million sequences or 1 million sequences. All samples reached saturation with the increase of the effective data, which indicated that the sequencing data size of each sample was sufficient and the analyses were credible. The saturating values were different for the 5’RACE and MPCR samples. More specifically, the number of unique CDR3 that reached the saturation was higher in MPCR compared with 5’RACE when starting with the same amount of RNA. More RNA input was required by 5’RACE to reach the same level of saturation with MPCR, which may be because of the random shearing step in which some clonotypes were lost in 5’RACE.

TCRs do not undergo somatic hyper-mutation in the recombination process; thus, the aligning mismatches with their germline genes were induced by the PCR and sequencing error, together with a certain degree of germline gene polymorphism. We subsequently compared the mismatches of both methods to estimate their errors. As expected, MPCR composed more errors compared with 5’RACE when calculated both by bases and sequences (Fig 8), which was probably contributed to its multiplex amplification.

Discussion

MPCR and 5’RACE are the most common experimental methods used to investigate the immune repertoire. 5’RACE is less biased [2], whereas the effect of MPCR appears to be
influenced by the usage of multiple primers. Therefore, we first evaluated the bias of MPCR using two different control samples. Few templates exhibited bias; moreover, the correlation coefficient between the observed and expected frequencies for the plasmid pools reached 0.9130. The frequencies of five spiked-in clones in three samples were close to the expected frequencies, which was similar with the previous reported[17]. These evaluations demonstrated that although some MPCR primers exhibited certain bias, it was within the acceptable range for it to be used in the subsequent analysis.

Next, the experimental reproducibility of MPCR and 5’RACE was evaluated. It must be emphasized that the two RNA aliquots used as replicates only represent parts of the entire repertoire, which may yield deviation for some TCR clones, especially the low frequency clones. The results demonstrated a strong similarity between the MPCR and 5’RACE samples (correlation coefficients of 0.9907 for MPCR and 0.9878 for 5’RACE) based on their CDR3 amino acid sequence abundance analysis, in which more than 68% of the sequences were identical in the top 1000 CDR3 sequences of the two sorted replicates. The reproducibility of 5’RACE was slightly lower than MPCR, which may be a result of the random shearing step in which only the biotinylated sequences were selected for 5’RACE. This finding also explains a potential reason for the greater difference in the number of unique CDR3 nucleotides and amino acids for 5’RACE. Overall, both methods were highly reproducible with excellent stability.

We identified considerable diversity in the TCRs, as well as biological differences between the divided blood samples or RNA from the same individual. Most V and J genes (~60%) exhibited comparable frequencies between the two methods for the three individuals investigated (Fig 5), with the exception of several genes that exhibited a significant difference. Sixteen V genes and 6 J genes had a significant difference with p-values < 0.05, which creates a strong bias for several genes in the MPCR samples. Notably, Freeman et al. described a healthy individual V and J usage with the 5’RACE method, in which the top 3 V genes were TRBV20-1, TRBV5-1, and TRBV29-1, respectively, whereas the top 5 J genes were TRBJ2-1, TRBJ1-1, TRBJ2-7, TRBJ2-3 and TRBJ2-5, respectively. Most of these genes also appeared in high frequency in our samples, which implies a similarity in the V/J gene usage distribution among healthy individuals and the reliability of our method [7].

We demonstrated that there were 17.07–21.07% overlapping unique CDR3 sequences between the MPCR and 5’RACE samples from the same individual (Fig 6). This finding was lower than expected and may be attributed to the biotinylated sequence selection of 5’RACE; the bias of MPCR; the amplification and sequencing errors and the diversity of RNA replicates. Considering the sequence abundance, the overlapping rate increased to an average of 55.42% (S2 Table), which indicated that the corresponding high-frequency sequences had been accurately profiled. The consistency and correlation of the VJ pairing were also calculated (Fig 7). Moreover, the sequences of the 5’RACE samples included nearly all of the V-J pairings, whereas the MPCR samples lost 63 V-J pairings on average (Fig 7), which indicates that 5’RACE had a higher efficiency to capture the different clones. As a result of the different RNA amounts used in the two methods for S01/S02, the correlation, consistency and V-J pairing of S01/S02 were worse compared with S03. Furthermore, the consistency and correlation of the CDR3 AA sequence for different samples utilizing the same experimental methods were analyzed, and the methodological differences were compared; however, the latter was demonstrated to be trivial as expected.

In summary, MPCR is a convenient method to enrich the CDR3 region. Notably, the primers designed and adjusted for the MPCR approach influence its performance and the subsequent results of the comparison. The primers adopted in this study were individually optimized for efficient amplification to ensure a satisfactory performance in our pooling plasmids of known clones. However, when tested in biological samples, the performance of several
specific genes may remain problematic, such as the lower usage of several J genes and the loss of several V-J parings in MPCR compared with 5’RACE. Our experiments have demonstrated the potential to achieve comparable results between the two methods; however, there are ways to optimize MPCR, such as adjusting the primer efficiency and fine-tuning the ratio of the primers, which is typically difficult. Furthermore, reducing the cycle number of MPCR was an effective way to avoid bias; however, an excessive reduction caused failure to enrich the CDR3 region. To overcome this problem, two-step PCR, in which a universal sequence is added to the 5’ end of the gene specific MPCR primers to enable a reduction in the cycles of the gene specific amplification and subsequent enrichment by the universal primers, was used[26]. In addition, in some research, random barcodes are added to the gene specific MPCR primers to correct the mismatches induced by PCR or sequencing and the amplification bias[26, 27].

5’RACE appeared to be a less biased method to enrich the CDR3 region. In the current setting, the 5’RACE products needed to be subjected to fragmentation and affinity purification, and these procedures result in the loss of low copy transcripts to a certain degree, which causes an inaccurate profiling of the immune repertoire. Therefore, for the 5’RACE immune repertoire libraries, direct sequencing without fragmentation would prove ideal, which could be achieved in other sequencing platforms with longer reads, such as Miseq or 454 sequencer.

Immune repertoire analysis has been widely applied to various fields, such as the monitoring of minimal residue disease for lymphocytic leukemia, vaccine performance evaluation and isolation of monoclonal antibodies. The selection of the proper method and the evaluation of its properties are crucial to scientific discoveries and a better understanding/interpretation of the captured repertoire diversity. For example, extraction of the genuine frequencies of cancer clones in the relapse of leukemia is the key for the application.

Our study provides a direct head-to-head comparison of the two major experimental methodologies commonly used to investigate the immune repertoire. We believe that the systematic evaluation of the platform-specific characteristics and bias would provide useful suggestions for researchers to select the suitable method and diminish the platform-induced bias/errors. One limitation was the use of Hiseq 2000, which was constraining because of the short read lengths that resulted from this platform. A similar constraint is associated with standard 5’RACE, which cannot be investigated without the fragmentation of cDNA and has been selected by other studies performed in combination with 454 sequencer. Overall, considering that we used the most widely used sequencing platform, the current findings will be of value to immunology research and will benefit the development of related applications in the field.

**Supporting Information**

**S1 Fig. Comparison of different individuals with the same method.** (a). The correlation of the CDR3 AA sequences between samples S01 and S02 using 5’RACE. Each dot represents a unique CDR3 AA sequence. (b). The correlation of VJ pairing between samples S01 and S02 using 5’RACE. Each dot represents a unique type of VJ pairing. (c). The consistency of the CDR3 AA sequences between samples S01 and S02 using 5’RACE. Black circles represent the overlapping rate of the two sorted data sets; grey circles represent the number of overlapping sequences of the two sorted data sets. In (a) and (b), the number in the top left corner indicates the Pearson correlation coefficient of the two data sets. (d). The correlation of the CDR3 AA sequences between samples S01 and S02 using MPCR. (e). The correlation of the VJ pairing between samples S01 and S02 using MPCR. (f). The consistency of the CDR3 AA sequences between samples S01 and S02 using MPCR. Black, overlapping rate; Grey, number of overlapped sequences.

(EOF)
S2 Fig. Saturation analysis for the three individuals. The saturation curve was plotted for each individual using the observed and predicted (Chao1) numbers of unique CDR3 AA for both MPCR and 5’RACE. (a). Sample S01. (b). Sample S02. (c). Sample S03. (d). Saturation value distribution for all samples.

S1 Table. Basic statistics of two repetitive MPCR and 5’RACE libraries.

S2 Table. Comparison of two methods for 3 individuals.

S3 Table. Plasmid mix pattern.

S4 Table. Experimental design for five CD4+ T cell clones.

Acknowledgments

We thank Nitin Saksena (University of Sydney, Australia) for editing this manuscript.

Author Contributions

Conceived and designed the experiments: XL KK. Performed the experiments: XZ XH RZ. Analyzed the data: WZ YD. Contributed reagents/materials/analysis tools: WZ XZ XH. Wrote the paper: WZ XL. Assisted with sequencing: CW HC ZS JW CN XX.

References


IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis

Wei Zhang,*†1 Yuanping Du,*†1 Zheng Su,*†1 Changxi Wang,* Xiaojing Zeng,* Ruifang Zhang,* Xueyu Hong,* Chao Nie,* Jinghua Wu,* Hongzhi Cao,* Xun Xu,* and Xiao Liu*,‡
*BGI-Shenzhen and †Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen 518083, China, and ‡Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark

ABSTRACT The advance of next generation sequencing (NGS) techniques provides an unprecedented opportunity to probe the enormous diversity of the immune repertoire by deep sequencing T-cell receptors (TCRs) and B-cell receptors (BCRs). However, an efficient and accurate analytical tool is still on demand to process the huge amount of data. We have developed a high-resolution analytical pipeline, Immune Monitor ("IMonitor") to tackle this task. This method utilizes realignment to identify V(D)J genes and alleles after common local alignment. We compare IMonitor with other published tools by simulated and public rearranged sequences, and it demonstrates its superior performance in most aspects. Together with this, a methodology is developed to correct the PCR and sequencing errors and to minimize the PCR bias among various rearranged sequences with different V and J gene families. IMonitor provides general adaptation for sequences from all receptor chains of different species and outputs useful statistics and visualizations. In the final part of this article, we demonstrate its application on minimal residual disease detection in patients with B-cell acute lymphoblastic leukemia. In summary, this package would be of widespread usage for immune repertoire analysis.

KEYWORDS next generation sequencing; bioinformatics; immune repertoire; TCR/BCR

The diversity of T-cell receptors (TCRs), B-cell receptors (BCRs), and secreting form antibodies makes up the core of the complicated immune system and serves as pivotal defensive components to protect the body against invading virus, bacteria, and other pathogens. The TCR consists of a heterodimeric αβ chain (~95%, TRA, TRB) or γδ chain (~5%), while the BCR is assembled with two heavy chains (IGH) and two light chains (IGK or IGL). Structurally, each chain can be divided into the variable domain and the constant domain (Lefranc and Lefranc 2001a,b). The diversity of the TCR and BCR repertoire is enormous, owing to the process of V(D)J gene rearrangement, random deletion of germ-line nucleotides, and insertion of uncertain length of nontemplate nucleotides between V-D and D-J junctions (TRB, IGH) or V-J junctions (TRA, IGK, IGL). In humans, it has been estimated theoretically that the diversity of TCR-αβ receptors exceeds $10^{18}$ in the thymus, and the diversity of the B-cell repertoire is even larger, considering the somatic hypermutation (Janeway 2005; Benichou et al. 2012). The T- and B-cell repertoire could undergo dynamic changes under different phenotypic status. Recently, deep sequencing enabled by different platforms including Roche 454 and Illumina Hiseq (Freeman et al. 2009; Robins et al. 2009; Wang et al. 2010; Fischer 2011; Venturi et al. 2011) has been applied to unravel the dynamics of the TCR and BCR repertoire and extended to various translational applications such as vaccination, cancer, and autoimmune diseases.

Several tools and software have been developed for TCR and BCR sequence analysis, including iHMMune-align (Gaeta et al. 2007), HighV-QEUST (Li et al. 2013), IgBLAST (Ye et al. 2013), Decombinator (Thomas et al. 2013), and MiTCR (Bolotin et al. 2013). These tools are equipped with useful functions, including V(D)J gene alignment, CDR3 sequence identification, and more, yet with obvious limitations. For instance, HighV-QEUST can be adopted to analyze both TCRs and BCRs, but its online version limits maximum sequence input to 150,000 at a time for regular users. Decombinator and MiTCR can only be used to analyze the TCR sequences.
Besides, most tools lack specific solutions to some common problems like systematic statistics and visualizations, PCR and sequencing errors, and amplification bias correction.

Here, we introduce a novel pipeline, Immune Monitor (IMonitor) for both TCR and BCR deep sequencing analysis. It includes four steps in its core component: basic data processing, V(D)J assignment, structural analysis, and statistics/visualization. One feature that makes IMonitor stand out is its realignment process to identify V(D)J genes and alleles with significantly enhanced precision. We simulated 15 data sets for five chains (TRA, TRB, IGH, IGK, IGL) of different sequencing error rates and hypermutation rates, together with actual rearranged sequences, to test performance of various tools. IMonitor performs quite well in accuracy and clonotype rearranged sequences, to test performance of various tools. IMonitor incorporates a process to correct PCR and sequencing errors, utilizing the data from six plasmid mixed samples, and an in silico model was modulated to reduce the PCR bias. Finally, we validate IMonitor in detection of minimal residual disease (MRD) of B-cell acute lymphoblastic leukemia (B-ALL) to show its wide utility potential.

Materials and Methods

The core component of IMonitor consists of four steps: basic data processing, V(D)J assignment, structural analysis, and statistics/visualization, as shown in Figure 1. IMonitor can utilize data generated by a variety of next generation sequencing (NGS) platforms, such as Illumina, Roche 454, and Life Ion Proton, in both FASTQ and FASTA format. The final results of IMonitor include a complete map of sequences and data analysis in depth, and the latter is visualized and presented with viewer-friendly graphs and figures.

IMonitor for basic data processing

In the first step, the reads were checked for inclusion of adaptor sequences. If any adaptor sequence was detected and located within 50 bp of the 3’ end of the read, it was deleted from the read. Reads bearing adaptor sequence at the 5’ end or >5% “N” bases were discarded. The average base quality of each read was calculated after removing the low-quality bases (base quality <10) at the 3’ end. Further filtration left out reads with average quality <15. For Illumina paired-end (PE) sequencing, the PE reads were merged at their overlapping region. For PE reads with insertion length longer than the length of a single read, the COPE (Liu et al. 2012) tool was used; otherwise reads were assembled by an in-house program. The main parameters for both tools included the maximum overlapping length (read length), minimum overlapping length (10 bp), mismatch rate (10%) at the overlapping region, and ratio (best overlap length/second-best overlap length, 0.7).

IMonitor for V(D)J assignment

The V/D/J reference sequences were downloaded from the IMGT database, the international ImMunoGeneTics information system (http://www.imgt.org/). Processed sequences were aligned to the V, (D), J references, respectively, by BLAST (Altschul et al. 1990; Zhang et al. 2000; Ye et al. 2006) and specific parameters were applied to accommodate the differences in lengths of V, (D), J segments (BLAST parameters: V, -W 15 -K 3 -v 1 -b 3; D, -W 4 -K 3 -v 3 -b 5; and J, -W 10 -K 3 -v 1 -b 3).

The high similarity among the genes and alleles of the germline sequences, along with the diversity of V/D/J gene rearrangement, gave rise to difficulties for accurate alignment. This might eventually lead to an incorrect structural analysis (CDR3 identification, deletion, or insertion). To improve the accuracy, a second alignment procedure was developed to identify exactly the V/D/J genes (Figure 2). First, a global alignment strategy, which attempted to align every base in every sequence, was used for the non-CDR3 region of the sequence. The mapped region generated from BLAST became a new seed and served as starting points for bootstrapping (base-by-base) extension to both directions, until the entire non-CDR3 region in the query was mapped to the target (reference) sequence. The mapping score was calculated according to these rules: reward for a nucleotide match was 5 and penalty for a nucleotide mismatch was −4. Second, the M-mismatch extension model of local alignment strategy was applied to locate the exact end positions of V and J genes during CDR3 region realignment. The procedure began at the CDR3 start position in the V gene or the CDR3 end position in the J gene and continuously extended in one direction until the preset mismatch limit was reached, generating the longest possible interval with the highest score. The mismatch numbers allowed for V/D/J genes were determined based on the analysis result of publicly available rearrangement sequences (http://www.imgt.org/ligmdb/) (Supporting Information, Figure S2A) and adjusted accordingly for different TCR and BCR chains (mismatches allowed: TRBV/J, TRAV/J, 0;IGHV/J, 2; IGKV/J, IGLV/J, 7). As shown in Figure S2A, these mismatch limits took mutations into consideration and covered >99.5% of all defined rearrangement sequences. Because the entire D gene was located within the CDR3 region, only the M-mismatch extension model was used for its realignment (mismatches allowed: TRBD, 0; IGHD, 4). Finally, all data including alignment score, identity, mismatch number, and alignment length were processed, and the alignment with highest score and identity larger than the threshold (>80%) was selected as the best hit. However, there might be several best hits with the same score due to the homology among the germline genes and alleles. In this case, the reference with the fewest deletions was selected, as shorter deletions are more likely to happen according to previous reported results (Warren et al. 2009) and our analysis from actual public rearrangement data (Figure S2B).

IMonitor for structural analysis

The IMGT collaboration (Yousfi Monod et al. 2004) outlined the CDR3 region of all chains, starting from the second conserved cysteine encoded by the V segment and ending with
the conserved phenylalanine or tryptophane encoded by the J segment. Combining this information with our selected reference from the previous step, the CDR3 region of target sequence could be readily identified. For unmapped sequences, the CDR3 region was determined by searching through for a conservative amino acids module within both ends of the CDR3 region (“YXC” for start and “[FW]GXG” for end, where “X” stands for any amino acid). The rearrangement frame was tagged as “in-frame” if the length of CDR3 was a multiple of three and no stop codon was found in whole sequence; otherwise it was tagged “out-of-frame.” The structure of the sequence was clearly described, including V, (D), J segments used, the CDR3 region, and the deletions and insertions at rearrangement sites. Then the nucleotide sequences were translated into peptides. However, some sequences must be filtered out to ensure the accuracy of the immune repertoire, which include (1) sequences without CDR3 region and (2) sequences with V and J alignment orientation conflict. The sequences that were aligned to pseudogenes, were out-of-frame, and included a stop codon were marked.

**IMonitor for statistics and visualization**

The basic statistics of IMonitor include CDR3 frequency distribution, V-J paring, V/J usage, 5’V/5’D/3’D/5’J deletion length distribution, V-D/D-J insertion length distribution, V/J base composition, CDR3 length distribution, CDR3 segmental frequency statistics, Top10 CDR3 frequency, hypermutation of BCRs, etc. Figures were plotted to visually demonstrate each result. For V-J paring, a three-dimensional...
figure was generated. R script was used to draw most of the figures while the V/J base composition was plotted with weblogo 2.8 (Crooks et al. 2004). To completely eliminate sequencing error effect, a sequence detected $<10$ times was excluded to calculate hypermutation. The mutation rate consisted of base mutation and sequence mutation. The former was the content of mutational bases in total bases, and the latter was rate of the sequence containing mutation in total sequences.

The Shannon–Weiner index (Shannon 1997), as shown below, used as an immunological diversity value in several previous works (Sherwood et al. 2013), provided a good estimate of diversity in a large-scale study and was suitable for immune repertoire. We used it to calculate the diversity of CDR3, the V gene, the J gene, and V-J pairing,

$$H' = -\sum_{i=1}^{S} p(i) \ln p(i),$$

where CDR3 is an example, $S$ denotes the total number of unique CDR3, and $p(i)$ denotes the frequency of CDR3.

IMonitor also provided the saturation analysis with the Chao1 algorithm, which was used to estimate the target richness for individual-based data in previous ecological studies. Estimated values generated by the Chao1 bias-corrected algorithm were used to predict the maximum number of clones in the sample, while observed values were drawn separately with the rarefaction curves (Chao 1984, 1987):

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{F_1(F_1 - 1)}{2F_2 + 1},$$

In Equation 2, $S_{\text{obs}}$ stands for total number of observed clonotypes in a sample; $F$ denotes the number of clonotypes whose base qualities all were $>Q20$ (Q20 were the best cutoff according to Figure S5, A and B); (2) sequences with more than five (three for only CDR3 region correction) low-quality bases were unwanted and discarded; and (3) the rest were defined as low-quality sequences. Second, the low-quality sequences were mapped to the high-quality ones. When the mismatches were no more than five (three for only CDR3 region correction) and all located at low-quality positions, the mismatches were corrected; otherwise the sequence was discarded. Finally, to eliminate PCR errors, sequences with low abundance were compared to ones with high abundance (at least fivefold difference). If fewer than three mismatches were found in the low-abundance sequences, they were corrected to the corresponding high-abundance sequences. To test the effectiveness of this method, samples made from mixtures of six plasmids were used for error characteristics analysis and further evaluation.

**Figure 2** The workflow of realignment. The program takes the BLAST alignment results as input, realigns the sequence to reference for both the non-CDR3 region and the CDR3 region, calculating the score and identity, and then selects the maximal score as the best hit. The reference with shortest length of deletion is preferred if it finds multiple references with the same maximal scores. It outputs the optimal alignment result.

**PCR and sequencing error correction**

PCR and sequencing error of NGS is one of the toughest problems in immune repertoire analysis. The method we developed to correct this error could be utilized on either the whole sequence or just the CDR3 region. The procedure consisted of three steps. First, sequences were divided into three groups: (1) high-quality sequences whose base qualities all were $>Q20$ (Q20 were the best cutoff according to Figure S5, A and B; (2) sequences with more than five (three for only CDR3 region correction) low-quality bases were unwanted and discarded; and (3) the rest were defined as low-quality sequences. Second, the low-quality sequences were mapped to the high-quality ones. When the mismatches were no more than five (three for only CDR3 region correction) and all located at low-quality positions, the mismatches were corrected; otherwise the sequence was discarded. Finally, to eliminate PCR errors, sequences with low abundance were compared to ones with high abundance (at least fivefold difference). If fewer than three mismatches were found in the low-abundance sequences, they were corrected to the corresponding high-abundance sequences. To test the effectiveness of this method, samples made from mixtures of six plasmids were used for error characteristics analysis and further evaluation.
**Multiplex PCR bias minimization**

We established a new bioinformatic approach to minimize PCR amplification bias. The approach is built on the hypothesis that there are two factors affecting a clone’s frequency during multiplex PCR (MPCR): the template’s concentration and the multiple primers’ efficiency. Using six plasmid mixture samples (Table S4), we could compare the observed with the expected frequency and simulate an effective formula.

The streamlined procedure of this method is demonstrated by the flowchart in Figure 6. Two factors, templates’ concentration and primers’ efficiency, were considered to be affecting the bias and examined here. The samples were mixed properly, and clones were grouped to explore PCR bias rules. First, we analyzed the bias’ correlation with templates’ concentration (concentration analysis without primer effect). Each clone had three different concentrations in all samples. To eliminate potential effects caused by the multiple primers efficiencies, the clones that had the same concentration ratio among all samples were grouped together, generating five groups in total [groups were named 10_2E4_1E5, 1000_2E4, 100_1E3_2E4, 100_1E4_2E4, and 10_1E4_2E4; for instance, 10_2E4_1E5 denoted the three concentrations (10, 2E4, 1E5) in respective samples]. Then, within each group, the clone frequencies were normalized by multiplying the same coefficient \( k \), which could generate a minimal sum of absolute deviation \( D_{\text{min}}(i) \). The 10_2E4_1E5 group sets an example as follows:

\[
\mu(j) = \frac{1}{n} \sum_{i=1}^{n} f(i, j), \quad j = \{10, 2E4, 1E5\} \tag{3}
\]

\[
D_{\text{min}}(i) = \min \{ |f(i, 10) k - \mu(10)| + |f(i, 2E4) k - \mu(2E4)| + |f(i, 1E5) k - \mu(1E5)| \}, \quad k \in (0, +\infty) \tag{4}
\]

\[
f_{\text{norm}}(i, j) = f(i, j) k, \quad j = \{10, 2E4, 1E5\}, \tag{5}
\]

where \( n \) is clone number in a group, \( i \) is a clone, and \( f \) is clone frequency. \( k \) was set consecutively from 0 and a series of \( D(i) \) was calculated, after which the \( k \) that generated the smallest \( D(i) \) was selected.

After normalization, five groups were combined on the basis of 2E4 copies, which existed in all groups. We used a regression module to fit a curve (Equation 6) that reflected the relationship between concentration and PCR bias,

\[
y = 0.60636 \log_{1.8}^{x+1} \tag{6}
\]

where \( y \) is the observed sequence’s frequency and \( x \) is the expected sequence’s frequency.

Second, we analyzed the bias caused only by primer efficiencies (primer analysis without concentration effect). To remove the effect of clone concentration, clones with the same concentration in all samples were collected into one group; thus six groups were generated (10, 100, 1E3, 1E4, 2E4, and 1E5, where group 10 contained all clones that had the concentration 10 in any of the samples) (Figure 6). After calculation, each group was normalized by multiplying the same coefficient \( l \), with the following details,

\[
r(i, j) = \begin{cases} 
\frac{f(i, j) l}{f(i, 2E4)}, & \text{if } f(i, j) l > f(i, 2E4) \\
\frac{f(i, 2E4)}{f(i, j) l}, & \text{if } f(i, 2E4) > f(i, j) l,
\end{cases} \quad l \in (0, +\infty), \quad j = \{10, 100, 1E3, 1E4, 2E4, 1E5\} \tag{7}
\]

\[
R_{\text{min}}(i, j) = \min \left\{ \sum_{l=1}^{n} r(i, j) \right\} \tag{8}
\]

\[
f_{\text{norm}}(i, j) = f(i, j) l, \tag{9}
\]

where, \( n \) is clone number in a group, \( i \) is a clone, and \( f \) is clone frequency. \( l \) was set consecutively from 0 and a series of \( R(i) \) was calculated, after which the \( l \) that generated the minimal \( R(i) \) was selected.

Each primer’s efficiency was calculated after normalization. Then, analysis of the two factors was integrated into a formula that minimized the PCR bias (Equation 10),

\[
f_{\text{correct}} = 1.8^{S} / (S 0.60636p) - 1, \quad p = 0.5 p(v) + 0.5 p(j) + 0.05, \tag{10}
\]

where \( f_{\text{correct}} \) is the corrected frequency, \( S \) is the clone’s observed abundance, \( S \) is the sum abundance of the sample, \( p(v) \) is the primer efficiency value for the V gene, and \( p(j) \) is the primer efficiency value for the J gene.

**Multiplex PCR amplification**

To amplify rearranged CDR3 regions, multiple forward primers in the V region and reverse primers in the J region were designed. For the RNA sample, the first-strand cDNAs were synthesized using SuperScript II Enzyme according to the manufacturer’s instructions. Then two individual equimolar pools of the forward primers and the reverse primers were used for a multiplex PCR (QIAGEN, Valencia, CA) of 30 cycles according to the provided protocol. The fractions between 110 and 180 bp of the PCR products were excised and purified.

**Simulation of in silico sequences**

A total of \( 10^6 \) sequences were generated in silico for each data set with a length of 200–300 bp by simulating the relevant biological processes that occur during B-cell and T-cell development. The sequences of V(D)J genes of TRA/TRB/IGH/IGK/IGL were downloaded from IMGT (http://www.imgt.org/). First, to simulate recombination, a V allele and a J allele (an extra D allele for TRB and IGH) were selected at random to generate a V-D-J (V-J for TRA and light chain) combination. Second, to simulate deletion and
insertion, some bases were deleted at the 3’ end of V and the 5’ end of J (both ends of D for TRB and IGH) according to the deletion length distribution from previous research (Figure S1); meanwhile some random bases were inserted at V-D and D-J junctions (V-J junction for TRA and light chain), using the established insertion length distribution (Figure S1). Third, both somatic hypermutation during BCR maturation and sequencing error were also taken into consideration. Although the typical error rate of Illumina HTS technology is ~1%, it can be reduced after merging paired-end reads (Figure 5A). Therefore, for each chain, the error rate was set to 0.1%, 0.5%, and 2% evenly at every position of the sequence. The hypermutation rate was set to 1% for IGH and 4% for IGK/IGL.

Public rearrangement data sets
The data sets were obtained from the IMGT/LIGM-DB database (http://www.imgt.org/ligmdb/); searched by “Homo sapiens,” “rearranged,” “TRB,” or “IGH”; and then the selected sequences were annotated manually (Annot. level = “manual”) and annotated by V, D, J genes. Finally, 24 TRB and 1763 IGH sequences met these requirements. The sequence was in FASTA format, without sequencing quality. The length of sequences downstream of the J gene was limited to 50 bp.

Samples

Plasmid mix samples: Thirty-three different functional TCR β-chain sequences, which included all the TCR β, V, and J genes, were integrated into plasmid vectors. Three mixing pattern pools were used: one with an equal mole number of each plasmid and the other two pools with different pooling gradients. The mix patterns are listed in Table S4, and each pattern was replicated to produce six plasmid mix samples.

Spiked-in DNA samples: Spiked-in samples were generously donated by Karen Ceresaletti (Benaroya Research Institute, Seattle, WA) (Robins et al. 2012). Five CD4+ T-cell clones were spiked in a background of sorted CD4+CD45RA+ naive T cells and each unit had 1 million cells. The five clones in three different units had different numbers of unique TCRβ CDR3 sequences and are shown in Table S7. The background cells for these doped samples were sorted from a fresh peripheral blood mononuclear cell (PBMC) sample obtained from a control donor with informed consent. DNA was extracted from the cells with a commercial kit.

Healthy donor samples: Samples of peripheral blood from two healthy human donors (H-H-1 and H-B-1) were obtained by venipuncture with informed consent. PBMCs were isolated immediately and RNA was extracted using Trizol reagent (Invitrogen, Carlsbad, CA). DNA was extracted with a QIAamp DNA Blood Mini Kit and stored at −20°C. The CDR3 region was amplified by multiplex PCR (Table S9) and sequenced by an Illumina platform (Table S6).
Data sets of TRB (24 sequences)

<table>
<thead>
<tr>
<th>Datasets/Tools</th>
<th>V_gene</th>
<th>V_allele</th>
<th>D_gene</th>
<th>D_allele</th>
<th>J_gene</th>
<th>J_allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMonitor</td>
<td>24 (100%)</td>
<td>23 (96%)</td>
<td>17 (71%)</td>
<td>12 (50%)</td>
<td>24 (100%)</td>
<td>24 (100%)</td>
</tr>
<tr>
<td>IgBLAST</td>
<td>23 (96%)</td>
<td>22 (92%)</td>
<td>13 (54%)</td>
<td>11 (46%)</td>
<td>24 (100%)</td>
<td>24 (100%)</td>
</tr>
<tr>
<td>Deconvinator*</td>
<td>21 (87%)</td>
<td>—</td>
<td>—</td>
<td>23 (96%)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Data set of IGH (1763 sequences)

<table>
<thead>
<tr>
<th>Datasets/Tools</th>
<th>V_gene</th>
<th>V_allele</th>
<th>D_gene</th>
<th>D_allele</th>
<th>J_gene</th>
<th>J_allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMonitor</td>
<td>1735 (98%)</td>
<td>1509 (86%)</td>
<td>1037 (59%)</td>
<td>952 (54%)</td>
<td>1619 (92%)</td>
<td>1533 (87%)</td>
</tr>
<tr>
<td>IgBLAST</td>
<td>1716 (97%)</td>
<td>1518 (86%)</td>
<td>986 (56%)</td>
<td>956 (54%)</td>
<td>1563 (89%)</td>
<td>1498 (85%)</td>
</tr>
</tbody>
</table>

The data sets were obtained from the IMGT/LIGM-DB database (http://www.imgt.org/ligmdb); searched by “Homo sapiens,” “rearranged,” “TRB,” or “IGH”; and then the selected sequences were annotated manually (Annot. level==“manual”) and annotated by V, D, J genes. So these sequences have a fairly high level of annotation confidence. The data sets and HighV-QEUST came from the same website, so HighV-QEUST was not used here. The references used for tools were the same and were from the IMGT database (http://www.imgt.org).

*Deconvinator analyzed just the gene level of V and J.

**MRD samples**: Bone marrow samples from two patients (M001 and M002) with B-ALL were provided with informed consent. The samples of pretreatment, day 15, and day 33 post-treatment were assayed. A total of 1.2 μg DNA of each sample was used for multiplex PCR amplification (Table S9). The library of ~150–270 bp insert-size length was extracted and sequenced using the 2 × 100 PE Illumina platform (Table S6).

The research was prospectively reviewed and approved by a duly constituted ethics committee.

**Data availability**

The source code of IMonitor is freely available for download at https://github.com/zhangwei2015.

**Table S1**: Simulated TRB with 0.1%, 0.5% and 2% sequencing error.

**Table S2**: Simulated IGH with 0.1%, 0.5% and 2% sequencing error and 0.1% hyper-mutation.

**Table S3**: Simulated Data with 0.5% sequencing error (TRA/IGK/IGL) and 4% hyper-mutation.

**Table S4**: Plasmid mixing pattern.

**Table S5**: Data process for PCR and sequencing error statistics.

**Table S6**: Samples information.

**Table S7**: Experimental design for five CD4+ T cell clones in the three spiked in mix.

**Table S8**: Performance of IMonitor and other tools on the simulated dataset.

**Table S9**: TRB and IGH V/J primers.

**Figure S1**: Insertion and deletion length distribution for simulated data.

**Figure S2**: IGH-VDJ Mutation and deletion/insertion analysis on the public sequences.

**Figure S3**: Outputs of IMonitor, H-B-01 as an example.

**Figure S4**: H-B-01 sample output figure of IMonitor.

**Figure S5**: Error characteristics of 6 plasmid mix samples.

**Figure S6**: V-J pairing dynamics for M002.

**Figure S7**: MiTCR and IMonitor performance in 3 spiked-in samples.

**Figure S8**: Nucleotide composition of V/J genes.

**Results**

**System design of IMonitor**

Four steps are described in Figure 1:

1. **Basic data process**: Sequence containing adapter sequence was processed and low-quality bases at the 3’ end of the sequence were removed. PE reads were merged to one sequence by an in-house program and COPE (Liu et al. 2012).

2. **V(D)J assignment**: The reference germline sequences were downloaded directly from IMGT (http://www.imgt.org/). Processed data were aligned to the references by BLAST (Altschul et al. 1990; Zhang et al. 2000; Ye et al. 2006) and realigned to improve the map accuracy, after which the optimal alignment was selected for every sequence (Figure 2).

3. **Structural analysis**: A novel method was established to correct PCR and sequencing errors. The CDR3 region was identified with the help of both V/J references and conservative amino acids and then translated into amino acids.

4. **Statistics and visualization**: Characteristic data of the immune repertoire of the samples, such as repertoire diversity, clonotype frequency, CDR3 length distribution, V/J usage, V-J pairing, hypermutation, deletion, and insertion, were collected (Figure S3) and presented with corresponding graphs. More specifically, the V-J pairing was visualized by a three-dimensional graph (Figure 4, Figure S4, and Figure S8).

**IMonitor outperforms other analytical tools in various aspects**

To evaluate the performance of IMonitor, we designed a head-to-head comparison between IMonitor and other publicly available tools with both simulated data and public rearrangement sequences. The TRA/TRB data were analyzed by HighV-QEUST (Li et al. 2013), Decombinator (Thomas et al. 2013), IgBLAST (Ye et al. 2013), and IMonitor, while the IGH/IGK/IGL data were analyzed by HighV-QEUST, IgBLAST (Ye et al. 2013), and IMonitor. Thomas et al. (2013) reported that...
iHMMune-align (Gaeta et al. 2007) generated a similar result to IgBLAST so it was excluded from the comparison. MiTCR (Bolotin et al. 2013) performance is strongly related to sequencing quality, so neither simulated data nor public sequences were suitable for it. Three spiked-in samples sequenced by Illumina were used to compare MiTCR and IMonitor, the result of which is shown in Figure S7. Although MiTCR finished the run much faster than IMonitor, clone G
was missed because of erroneous J-gene assignment and an incorrect CDR3 region. All tools employed their default parameters.

By processing the in silico data with different tools, we calculated the accuracy of V/D/J genes and V/D/J alleles of each tool analyzed (Figure 3 and Table S1, Table S2, and Table S3). For all 24 V/D/J genes and alleles in all TCR/BCR chains, IMonitor showed superior performance in 14 of them (58.33%), with 7 of 24 slightly lower in performance than the best tool (no more than 1% difference) and the remaining 3 from 1.6 to 3.25% difference (TRAV gene, TRBD gene, and TRBD allele). For the highly homologous V-gene family (at least 40 for each chain), the accuracy of IMonitor was >95% for almost all chains and exceeded 99% for all J genes. For D genes, which were short and embedded with deletions and insertions at both ends, IMonitor performed significantly better for IGH (>=80% accuracy), while slightly worse for TRB. Decombinator was not designed to identify alleles and D genes.

In addition to the simulating data, public rearranged sequences that are annotated manually with clear V(D)J genes and extracted from the IMGT/LIGM-DB database (http://www.imgt.org/ligmdb/) were utilized to test IMonitor. Twenty-four TRB sequences and 1763 IGH sequences were analyzed by different tools (Table 1). For TRB, IMonitor and IgBLAST performed better than Decombinator in general, whereas IMonitor outperformed IgBLAST in D genes. For IGH, IMonitor performed similarly to IgBLAST in V and D alleles, but was superior in all other genes and alleles. The good performance in D genes by IMonitor demonstrated the effectiveness of the M-mismatch extension model during D gene realignment. The accuracies of IGH-J genes and alleles were both slightly lower for these two tools, because some public IGH sequences have only a partial J segment (<30 bp) and they are difficult to distinguish from other homologous genes and alleles.

To assess running time and memory needed, 10^5 simulated TRB and IGH data sets were analyzed by IMonitor, IgBLAST, and Decombinator separately; the results are shown in Table S8. When only one CPU was used, IMonitor took 12 min 52 sec and 21 min 96 sec to analyze the two data sets separately, with peak memory of 226 ~ 325 Mb. Of all tools tested, Decombinator was the fastest and IMonitor ranked second.

Overall, IMonitor produced satisfactory results for both simulated and published sequences. It generated similar results to IgBLAST in some genes, while it outperformed other tools in most occasions. It is also direct proof that the realignment strategy for V(D)J identification is useful.

The output of IMonitor

One of the features that distinguish IMonitor from others is its ability to export comprehensive statistics for characteristics of TCR/BCR repertoire and accessible graphs. The statistics include not only basic statistics but also in-depth statistics (Figure S3). The former elucidates the process from raw data to effective sequences, such as clean data rate and V(D)J gene mapped rate, which all provide sequence number, rate of input, and rate of raw data. The latter consists of multiple statistics based on effective sequence, such as functional classification, V/J/V-J gene usage rate, clone number, diversity calculated by Shannon index, and hypermutation. IMonitor is also able to translate obscure data into self-explanatory graphs. Important statistics like V/J usage, top 10 clone frequencies, CDR3 segmental frequency statistics (split into four segments after frequency sorted: top100, 100-1E3, 1E3-1E4, >1E4), insertion and deletion length distribution, V/J nucleotide composition, and V-J pairing

![Figure 5](error_rate.png)
Figure 6 Minimize MPCR bias flowchart. Six samples are mixed together and search the bias rules under two independent pathways, concentration analysis and primer analysis. For concentration analysis, six groups are created to eliminate primer effect, and each group is normalized; then five groups.
can all be presented as figures (Figure 4, Figure S4, and Figure S8). V-J pairing diversity is visualized (Figure 4M, Figure S4M), which can be applied to track the changes of the immune system over time and reveal immunological conditions.

**PCR and sequencing errors correction**

IMonitor also integrates a process to correct PCR and sequencing errors. The six plasmid mixture samples with three different pooling gradients (Materials and Methods and Table S4) were used to analyze the error characteristics and evaluate the effectiveness of error correction. The processed sequences were mapped to V/D/J references, and the final effective data were used to summarize the error characteristics shown in Figure S5. The known template sequences were used as references to calculate the error rate, and the results are shown in Figure 5 and Table S5. The average percentage of high-quality sequences was 74.88%. A total of 12.86% of sequences with low quality were corrected according to high-quality sequences, while the remaining 12.26% were discarded. Thereafter, an average 6.33% of low-abundance sequences were also corrected (Table S5). In consideration of the influence posed on error statistics by impurity of the plasmids during the cell culture before we mixed the plasmids, if the “erroneous” sequence was found in on less than four samples and was detected >100 times in each sample, it was excluded from error rate calculation. After the correction process, the mean error rate of all sequences was decreased from 0.082 to 0.013%, and the percentage of error-bearing sequences was decreased from 6.313 to 0.912% (Figure 5).

**Minimization of multiplex PCR bias**

We successfully developed a novel method to minimize the MPCR bias under a given set of multiplex primers. Details of this method are shown in Figure 6 and Materials and Methods. Cross-validation was used to evaluate this method, as shown in Figure 6. Residual sum of squares (RSS) (Draper and Smith 1998) was calculated for each test, where $$RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$$, $y_i$ is the observed frequency, and $\hat{y}$ is the expected frequency. Except for the mix 1-2 sample, the other five samples reduced the RSS value, which demonstrated that this method is evidently effective.

Moreover, the method was tested using three spiked-in samples, in which five known clones were spiked in $10^6$ cells. Compared to the expected frequency, clones B, C, D, and G had obvious bias. After modifying the frequency with this method, the bias was relieved to a certain extent (Figure 7, A–C). The bias ratio was defined as the observed frequency bias divided by the corrected frequency bias. If the ratio is $>1$, it means the frequency is corrected positively. A total of 86.7% of the clones (except clone A in Index-R/T) generated a ratio $>1$, particularly clone D in Index-R. These results conclude that the method is indeed capable of minimizing MPCR bias (Figure 7, D and E).

**IMonitor to monitor MRD**

To test the feasibility of IMonitor in translational research, we applied it to analyze the data from two patients with B-ALL.
Samples were collected upon diagnosis and on day 15 and day 33 post-treatment (details in Materials and Methods). We first tried to identify the cancer clones in the B-cell receptor repertoire. Patients with B-ALL showed a clear pattern of deficient clonal diversity (Figure 8A, Figure S6). By the cutoff of 10% for the cancer clone frequency, the cancer clone was identified in patient M001 with clonal frequency of 87.02% upon diagnosis, whereas it substantially decreased to 14.65% on day 15 and further to 0.28% on day 33, suggesting the effect of treatment (Figure 8B). However, with the increased sensitivity of our method, MRD was detected on day 33. In contrast, flow cytometry was not able to detect any MRD in this time slot. Interestingly, two cancer clones (48.39% and 42.13%) were identified in patient M002, and the data of day 33 post-treatment showed an MRD level of 2.54% (1.46% and 1.08% corresponding to each clone), compared with the negative result in detection with flow cytometry.

![Figure 8](image.png)

**Figure 8** Detection of MRD in B-ALL using IMonitor. (A) Repertoire (V-J pairing) of IGH is shown for pretreatment (day 0) and post-treatment (day 15 and day 33). (B) Cancer clone frequency is shown for each clone in the two patients (M001 and M002) before treatment (day 0) and post-treatment (day 15 and day 33).

Discussion

We have developed a comprehensive methodology for analysis of the T-cell receptor repertoire and B-cell receptor repertoire made available by next generation sequencing technology. IMonitor provides an arsenal of solutions for four steps: basic data processing, V(D)J assignment, structural analysis, and statistics visualization. IMonitor distinguishes itself from other analysis tools with several features. The first important feature is its realignment process. The high homology among genes and alleles together with random base deletion and insertion at gene junctions have been affecting the accuracy of alignment. Therefore, global or local alignment by itself is not sufficient to complete the whole picture. During the realignment process, CDR3 regions are scrutinized with the M-mismatch extension model of local alignment while non-CDR3 regions are covered by global alignment. The test using simulated data and published rearrangement sequences demonstrated IMonitor’s unquestionably better performance than other tools. The second feature of IMonitor is its ability to correct PCR and sequencing error and minimize MPCR bias, whose usage can be extended to other fields of research. IMonitor can be used to analyze any chain of T- and B-cell receptors and multiple species such as humans, monkeys, and rabbits. Furthermore, IMonitor results are presented with intuitive graphs. For example, the overall diversity of the immune system can be interpreted easily from a three-dimensional V-J pairing graph.

PCR and sequencing error of NGS is one of the problems that remain untackled for immune repertoire analysis. Preliminary results from previous studies show that a significant number of errors accumulate, and these errors can potentially lead to overestimating the actual TCR clonotypes. Besides,
sequencing error also results in artificially increased diversity of the TCR repertoire (Nguyen et al. 2011; Warren et al. 2011). Simply filtering out all low-quality sequences not only removes the sequencing error bases, but also leaves out a lot of genuine sequences. Our method, however, manages to decrease the error rate while rescuing most of the sequences. Actually, the efficiency of the approach would be improved if more factors are considered when correcting the errors. For example, erroneous bases have some bias for certain sequencing platforms, and the occurrence probability of sequencing error at each position of sequence is different. Using the six plasmid mixture samples, the characteristics of PCR and sequencing errors can be analyzed, as shown in Figure S5. The base error rate declines as the base quality in the sequence improves, whereas the percentage of discarded sequences increases sharply (Figure S5, A and B). The quality indexes of incorrect bases mainly fall into two categories: Q ≤ 10 and Q ≥ 35. Apparently, the former mostly results from sequencing error, while the cause of the latter is mostly PCR error (Figure S5D). More than 85% of sequences have only one error base, and the rate rises after removing the sequences with minimal base quality of Q20 (Figure S5C).

Here we have introduced a new bioinformatics methodology to reduce the PCR bias of MPCR samples. We found that the bias originated from two factors: template concentration and inconsistent primer efficiencies. Using six plasmid mixture samples, we designed a formula to reduce the bias. By applying it to the spiked-in samples, we validated its effectiveness. However, due to the limited size of training data, some bias persisted in spiked-in samples. We believe when the training data contain ≥100 templates, the effect of the approach would be more significant. Besides, different primer sets should be trained to generate a suitable formula to reduce bias, so this article mainly introduces a bioinformatics approach showing how to create a suitable formula to adjust the bias. Previous literature reports that it reduce the bias mainly through an experimental method to optimize the primers and primer concentration (Carlson et al. 2013). It is a scientific and systemic experimental method to adjust primers. It would be ideal to use this method for optimizing primers in the first step and then to use our bioinformatics method for further reducing PCR bias. Stephen R. Quake and colleagues developed a consensus read sequencing approach that incorporated unique barcode labels (UIDs) on each starting RNA molecule (Vollmers et al. 2013). It could eliminate PCR bias completely in theory if the synthetic UIDs were random enough.

IMonitor for analyzing the TCRs and BCRs repertoire in human and other animal models has the widest applications among the available tools in basic and translational research. We have demonstrated its utility in identifying the cancer clonotypes and monitoring MRD in B-ALL, while at the same time evaluating the clonal diversity for immune remodeling following treatment by its graphic visualization. We believe that IMonitor can also be applied in many other areas, such as tracing emerging clonotypes upon vaccination and following their frequencies during the process, selecting monoclonal antibodies based on sequencing the immune repertoire. With the importance of immune repertoire research becoming more recognized, we believe IMonitor will play a role in advancing our understanding of the immune system.

Acknowledgments

We thank the Science and Technology Planning Project of Guangdong Province (no. 2012A031100010) for support and the Shenzhen Municipal Government of China (no. GJHZ20130417140835564) for support. This study is supported by the International Science and Technology Cooperation Program of Shenzhen, China (GJHZ20130417140835564).

Literature Cited


Nguyen, P., J. Ma, D. Pei, C. Obert, C. Cheng et al., 2011 Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics 12: 106.
Warren, R. L., J. D. Freeman, T. Zeng, G. Choe, S. Munro et al., 2011 Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 21: 790–797.

Communicating editor: G. D. Stormo

W. Zhang et al.
Supporting Information inventory

Table S1 Simulated TRB with 0.1%, 0.5% and 2% sequencing error.
Table S2 Simulated IGH with 0.1%, 0.5% and 2% sequencing error and 0.1% hyper-mutation
Table S3 Simulated Data with 0.5% sequencing error (TRA/IGK/IGL) and 4% hyper-mutation
Table S4 Plasmid mixing pattern
Table S5 Data process for PCR and sequencing error statistics
Table S6 Samples information
Table S7 Experimental design for five CD4+ T cell clones in the three spiked in mix
Table S8 Performance of IMonitor and other tools on the simulated dataset
Table S9 TRB and IGH V/J primers

Figure S1 Insertion and deletion length distribution for simulated data
Figure S2 IGH-VDJ Mutation and deletion/insertion analysis on the public sequences
Figure S3 Outputs of IMonitor, H-B-01 as an example
Figure S4 H-B-01 sample output figure of IMonitor
Figure S5 Error characteristics of 6 plasmid mix samples
Figure S6 V-J pairing dynamics for M002
Figure S7 MiTCR and IMonitor performance in 3 spiked-in samples
Figure S8 Nucleotide composition of V/J genes
Table S1. Simulated TRB with 0.1%, 0.5% and 2% sequencing error.

<table>
<thead>
<tr>
<th>Mismatch Rate</th>
<th>Hyper-Mutation Rate</th>
<th>Software</th>
<th>TRBV-Genome(%)</th>
<th>TRBD-Genome(%)</th>
<th>TRBJ-Genome(%)</th>
<th>TRBV-allele(%)</th>
<th>TRBD-allele(%)</th>
<th>TRBJ-allele(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10%</td>
<td>0.00%</td>
<td>IMMonitor</td>
<td>98.25</td>
<td>77.04</td>
<td>100.00</td>
<td>82.95</td>
<td>72.13</td>
<td>99.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>91.46</td>
<td>86.31</td>
<td>99.24</td>
<td>65.94</td>
<td>81.69</td>
<td>99.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>97.87</td>
<td>80.55</td>
<td>99.20</td>
<td>72.46</td>
<td>76.33</td>
<td>99.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decombinator</td>
<td>70.28</td>
<td>-</td>
<td>80.76</td>
<td>29.53</td>
<td>-</td>
<td>70.17</td>
</tr>
<tr>
<td>0.50%</td>
<td>0.00%</td>
<td>IMMonitor</td>
<td>98.18</td>
<td>76.90</td>
<td>100.00</td>
<td>82.34</td>
<td>71.60</td>
<td>99.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>91.44</td>
<td>85.77</td>
<td>99.24</td>
<td>66.04</td>
<td>80.75</td>
<td>99.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>97.80</td>
<td>79.32</td>
<td>99.02</td>
<td>72.58</td>
<td>74.95</td>
<td>98.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decombinator</td>
<td>68.33</td>
<td>-</td>
<td>78.64</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2%</td>
<td>0.00%</td>
<td>IMMonitor</td>
<td>98.13</td>
<td>77.82</td>
<td>100.00</td>
<td>80.01</td>
<td>70.77</td>
<td>99.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>91.59</td>
<td>83.89</td>
<td>99.30</td>
<td>65.39</td>
<td>77.27</td>
<td>98.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>97.83</td>
<td>75.13</td>
<td>98.61</td>
<td>71.82</td>
<td>70.07</td>
<td>98.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decombinator</td>
<td>59.58</td>
<td>-</td>
<td>68.45</td>
<td>24.85</td>
<td>-</td>
<td>60.18</td>
</tr>
<tr>
<td>Mismatch Rate</td>
<td>Hyper-Mutation Rate</td>
<td>Software</td>
<td>IGHV-G Gene(%)</td>
<td>IGHD-G Gene(%)</td>
<td>IGHJ-G Gene(%)</td>
<td>IGHV-allele(%)</td>
<td>IGHD-allele(%)</td>
<td>IGHJ-allele(%)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------</td>
<td>----------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>0.10%</td>
<td>0.10%</td>
<td>IMonitor</td>
<td>98.30</td>
<td>83.97</td>
<td>99.98</td>
<td>81.48</td>
<td>81.41</td>
<td>99.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>87.97</td>
<td>72.90</td>
<td>98.60</td>
<td>69.38</td>
<td>70.81</td>
<td>97.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>98.76</td>
<td>75.68</td>
<td>99.83</td>
<td>80.74</td>
<td>73.7</td>
<td>99.73</td>
</tr>
<tr>
<td>0.50%</td>
<td>0.10%</td>
<td>IMonitor</td>
<td>98.19</td>
<td>83.86</td>
<td>99.98</td>
<td>81.3</td>
<td>81.28</td>
<td>99.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>87.51</td>
<td>72.73</td>
<td>98.61</td>
<td>68.96</td>
<td>70.66</td>
<td>97.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>98.63</td>
<td>75.47</td>
<td>99.83</td>
<td>80.63</td>
<td>73.52</td>
<td>99.5</td>
</tr>
<tr>
<td>2%</td>
<td>0.10%</td>
<td>IMonitor</td>
<td>98.02</td>
<td>83.79</td>
<td>99.96</td>
<td>80.46</td>
<td>81.01</td>
<td>98.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>87.72</td>
<td>72.33</td>
<td>98.59</td>
<td>68.43</td>
<td>70.04</td>
<td>96.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>98.52</td>
<td>74.61</td>
<td>99.79</td>
<td>79.63</td>
<td>72.22</td>
<td>98.55</td>
</tr>
</tbody>
</table>
Table S3. Simulated Data with 0.5% sequencing error (TRA/IGK/IGL) and 4% hyper-mutation for IGK/IGL

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mismatch Rate</th>
<th>Hyper-Mutation Rate</th>
<th>Software</th>
<th>V-Gene (%)</th>
<th>J-Gene (%)</th>
<th>V-allele (%)</th>
<th>J-allele (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGK</td>
<td>0.50%</td>
<td>4.00%</td>
<td>IMMonitor</td>
<td>93.57</td>
<td>99.99</td>
<td>86.91</td>
<td>92.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>73.13</td>
<td>99.97</td>
<td>64.94</td>
<td>93.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>91.53</td>
<td>100.00</td>
<td>85.46</td>
<td>92.94</td>
</tr>
<tr>
<td>IGL</td>
<td>0.50%</td>
<td>4.00%</td>
<td>IMMonitor</td>
<td>100.00</td>
<td>99.57</td>
<td>77.24</td>
<td>98.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>98.82</td>
<td>97.54</td>
<td>76.85</td>
<td>97.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>100.00</td>
<td>99.67</td>
<td>78.25</td>
<td>99.53</td>
</tr>
<tr>
<td>TRA</td>
<td>0.50%</td>
<td>0.00%</td>
<td>IMMonitor</td>
<td>98.35</td>
<td>100.00</td>
<td>86.75</td>
<td>99.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HighV-QUST</td>
<td>100.00</td>
<td>94.12</td>
<td>83.31</td>
<td>93.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IgBLAST</td>
<td>100.00</td>
<td>100.00</td>
<td>85.32</td>
<td>99.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Decombinator</td>
<td>68.36</td>
<td>72.83</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table S4. Plasmid mixing pattern

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C-01</td>
<td>TRBV10-1</td>
<td>TRBJ2-7</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>C-02</td>
<td>TRBV10-2/3</td>
<td>TRBJ2-7</td>
<td>2000</td>
<td>2000</td>
<td>1000</td>
<td>1000</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>C-03</td>
<td>TRBV11-1/2/3</td>
<td>TRBJ1-3</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>C-04</td>
<td>TRBV11-1/2/3</td>
<td>TRBJ1-5</td>
<td>2000</td>
<td>2000</td>
<td>100</td>
<td>100</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>C-05</td>
<td>TRBV12-3/4</td>
<td>TRBJ2-1</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-06</td>
<td>TRBV12-5</td>
<td>TRBJ2-1</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>C-07</td>
<td>TRBV13</td>
<td>TRBJ1-1</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-08</td>
<td>TRBV14</td>
<td>TRBJ2-7</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-09</td>
<td>TRBV15</td>
<td>TRBJ1-6</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-10</td>
<td>TRBV15</td>
<td>TRBJ2-4</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-11</td>
<td>TRBV16</td>
<td>TRBJ1-1</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-12</td>
<td>TRBV19</td>
<td>TRBJ1-6</td>
<td>2000</td>
<td>2000</td>
<td>100</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-13</td>
<td>TRBV20-1</td>
<td>TRBJ1-5</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-14</td>
<td>TRBV20-1</td>
<td>TRBJ1-5</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-15</td>
<td>TRBV20-1</td>
<td>TRBJ2-2</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-16</td>
<td>TRBV24-1</td>
<td>TRBJ1-2</td>
<td>2000</td>
<td>2000</td>
<td>100</td>
<td>100</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-17</td>
<td>TRBV25</td>
<td>TRBJ2-5</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-18</td>
<td>TRBV27/28</td>
<td>TRBJ2-4</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-19</td>
<td>TRBV29-1</td>
<td>TRBJ2-3</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-20</td>
<td>TRBV2</td>
<td>TRBJ2-6</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-21</td>
<td>TRBV30</td>
<td>TRBJ1-1</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-22</td>
<td>TRBV3-1</td>
<td>TRBJ1-2</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-23</td>
<td>TRBV4-1/2/3</td>
<td>TRBJ2-7</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-24</td>
<td>TRBV5-1</td>
<td>TRBJ2-1</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C-25</td>
<td>TRBV5-4/5/6/8</td>
<td>TRBJ2-3</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-26</td>
<td>TRBV6-1/2/3/5/8</td>
<td>TRBJ2-1</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>C-27</td>
<td>TRBV6-4</td>
<td>TRBJ2-5</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-28</td>
<td>TRBV6-6</td>
<td>TRBJ1-6</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-29</td>
<td>TRBV6-9</td>
<td>TRBJ1-3</td>
<td>2000</td>
<td>2000</td>
<td>100000</td>
<td>100000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-30</td>
<td>TRBV7-2/4/6/7/8</td>
<td>TRBJ2-6</td>
<td>2000</td>
<td>2000</td>
<td>10</td>
<td>10</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-31</td>
<td>TRBV7-3</td>
<td>TRBJ2-7</td>
<td>2000</td>
<td>2000</td>
<td>100</td>
<td>100</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>C-32</td>
<td>TRBV7-9</td>
<td>TRBJ1-4</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C-33</td>
<td>TRBV9</td>
<td>TRBJ1-2</td>
<td>2000</td>
<td>2000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Note: * the clone ratio in the sample.
Table S5. Data process for PCR and sequencing error statistics.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sum Sequence</th>
<th>High Quality Sequence(%)</th>
<th>Filter Sequence(%)</th>
<th>Low Quality Corrected(%)</th>
<th>PCR Error Corrected(%)</th>
<th>Effective Data</th>
<th>Before Correction</th>
<th>After Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Base Error(%)</td>
<td>Sequence Error(%)</td>
</tr>
<tr>
<td>index-1</td>
<td>4,273,571</td>
<td>78.50</td>
<td>9.28</td>
<td>12.21</td>
<td>8.36</td>
<td>3,876,775</td>
<td>0.098</td>
<td>7.304</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.016</td>
<td>1.152</td>
</tr>
<tr>
<td>index-2</td>
<td>4,217,557</td>
<td>78.59</td>
<td>9.77</td>
<td>11.64</td>
<td>9.10</td>
<td>3,805,551</td>
<td>0.111</td>
<td>8.628</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.022</td>
<td>1.564</td>
</tr>
<tr>
<td>index-3</td>
<td>3,603,556</td>
<td>70.16</td>
<td>16.25</td>
<td>13.59</td>
<td>4.44</td>
<td>3,018,100</td>
<td>0.070</td>
<td>5.218</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
<td>0.590</td>
</tr>
<tr>
<td>index-10</td>
<td>5,078,119</td>
<td>81.28</td>
<td>8.80</td>
<td>9.92</td>
<td>4.70</td>
<td>4,631,376</td>
<td>0.058</td>
<td>4.359</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
<td>0.569</td>
</tr>
<tr>
<td>index-11</td>
<td>2,785,059</td>
<td>64.59</td>
<td>18.61</td>
<td>16.80</td>
<td>5.32</td>
<td>2,266,836</td>
<td>0.081</td>
<td>6.484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.010</td>
<td>0.779</td>
</tr>
<tr>
<td>index-12</td>
<td>3,335,881</td>
<td>76.15</td>
<td>10.87</td>
<td>12.98</td>
<td>6.04</td>
<td>2,973,382</td>
<td>0.072</td>
<td>5.885</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.010</td>
<td>0.816</td>
</tr>
<tr>
<td>Sample</td>
<td>species</td>
<td>Gene</td>
<td>Library</td>
<td>Experimental method</td>
<td>Sequencer</td>
<td>Type</td>
<td>Amount</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
<td>---------------------</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>H-B-01</td>
<td>Human</td>
<td>TRB</td>
<td>cDNA</td>
<td>MPCR</td>
<td>Hisseq2500</td>
<td>PE100</td>
<td>1ug</td>
<td></td>
</tr>
<tr>
<td>H-H-01</td>
<td>Human</td>
<td>IGH</td>
<td>DNA</td>
<td>MPCR</td>
<td>Hisseq2000</td>
<td>PE100</td>
<td>3ug</td>
<td></td>
</tr>
<tr>
<td>M001</td>
<td>Human</td>
<td>IGH</td>
<td>DNA</td>
<td>MPCR</td>
<td>Hisseq2500</td>
<td>PE150</td>
<td>1.2ug</td>
<td></td>
</tr>
<tr>
<td>M002</td>
<td>Human</td>
<td>IGH</td>
<td>DNA</td>
<td>MPCR</td>
<td>Hisseq2500</td>
<td>PE150</td>
<td>1.2ug</td>
<td></td>
</tr>
</tbody>
</table>
Table S7. Experimental design for five CD4+ T cell clones in the three spiked in mix.

<table>
<thead>
<tr>
<th>Clone</th>
<th>TCRB V</th>
<th>TCRB J</th>
<th>CDR3</th>
<th>Mix 1</th>
<th>Mix 2</th>
<th>Mix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>VB8</td>
<td>TRBJ1-1</td>
<td>CASSLGGQGVG</td>
<td>100,000</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>A</td>
<td>VB5.1</td>
<td>TRBJ2-5</td>
<td>CASSPGIAELKETQY</td>
<td>10,000</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>VB6.7</td>
<td>TRBJ2-7</td>
<td>CASHTGFVSYEQQY</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>C</td>
<td>VB4</td>
<td>TRBJ1-4</td>
<td>CSVGTGDNEKLF</td>
<td>100</td>
<td>1000</td>
<td>10,000</td>
</tr>
<tr>
<td>D</td>
<td>VB4</td>
<td>TRBJ1-4</td>
<td>CSVGQGDNEKLF</td>
<td>10</td>
<td>1000</td>
<td>100,000</td>
</tr>
</tbody>
</table>
Table S8. Performance of IMonitor and other tools on the simulated dataset.

<table>
<thead>
<tr>
<th>Data set of TRB ($10^5$ sequences)</th>
<th>Peak Memory (MB)</th>
<th>Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMonitor $^a$</td>
<td>325.88M</td>
<td>12m52s</td>
</tr>
<tr>
<td>IgBLAST $^b$</td>
<td>327.95M</td>
<td>27m46s</td>
</tr>
<tr>
<td>Decombinator $^c$</td>
<td>209.00M</td>
<td>1m23s</td>
</tr>
<tr>
<td>HighV-QUST $^d$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data set of IGH (10$^5$ sequences)</th>
<th>Peak Memory (MB)</th>
<th>Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMonitor $^a$</td>
<td>226.22M</td>
<td>21m96s</td>
</tr>
<tr>
<td>IgBLAST $^b$</td>
<td>196.22M</td>
<td>92m30s</td>
</tr>
<tr>
<td>HighV-QUST $^d$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: $^a$, run with 1cpu and blast (-a 1); $^b$, run with 1cpu and igblast (-num_threads 1);
$c$, run with command prompt; $^d$, run online, send the results to user after 1-2weeks
### Table S9. TRB and IGH V/J primers

<table>
<thead>
<tr>
<th>IGH V/J Primers</th>
<th>TRB V Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHV1-18</td>
<td>TRBV2</td>
</tr>
<tr>
<td>IGHV1-2/1-46</td>
<td>TRBV3-1</td>
</tr>
<tr>
<td>IGHV1-24</td>
<td>TRBV4-1/2/3</td>
</tr>
<tr>
<td>IGHV1-3/1-45</td>
<td>TRBV5-1</td>
</tr>
<tr>
<td>IGHV1-69/1-f</td>
<td>TRBV5-4/5/6/8</td>
</tr>
<tr>
<td>IGHV1-8</td>
<td>TRBV6-4.1</td>
</tr>
<tr>
<td>IGH2-70/26/5</td>
<td>TRBV6-8/5/1.2</td>
</tr>
<tr>
<td>IGHV3</td>
<td>TRBV6-9/7/1.1/6</td>
</tr>
<tr>
<td>IGH4</td>
<td>TRBV6-4.2</td>
</tr>
<tr>
<td>IGHV5-51</td>
<td>TRBV6-2/3</td>
</tr>
<tr>
<td>IGHV6-1</td>
<td>TRBV7-2/4/6/7/8</td>
</tr>
<tr>
<td>IGHV7</td>
<td>TRBV7-3</td>
</tr>
<tr>
<td>IGHJ</td>
<td>TRBV7-9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRB J Primers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TRBJ1.1</td>
<td>TRBV9</td>
</tr>
<tr>
<td>TRBJ1.2</td>
<td>TRBV10-1</td>
</tr>
<tr>
<td>TRBJ1.3</td>
<td>TRBV10-2/3</td>
</tr>
<tr>
<td>TRBJ1.4</td>
<td>TRBV11-1/2/3</td>
</tr>
<tr>
<td>TRBJ1.5</td>
<td>TRBV12-3.2/5.2</td>
</tr>
<tr>
<td>TRBJ1.6</td>
<td>TRBV12-3.1/4/5.1</td>
</tr>
<tr>
<td>TRBJ2.1</td>
<td>TRBV13</td>
</tr>
<tr>
<td>TRBJ2.2</td>
<td>TRBV14</td>
</tr>
<tr>
<td>TRBJ2.3</td>
<td>TRBV15</td>
</tr>
<tr>
<td>TRBJ2.4</td>
<td>TRBV16</td>
</tr>
<tr>
<td>TRBJ2.5</td>
<td>TRBV18</td>
</tr>
<tr>
<td>TRBJ2.6</td>
<td>TRBV19</td>
</tr>
<tr>
<td>TRBJ2.7</td>
<td>TRBV20-1</td>
</tr>
<tr>
<td>TRBJ2.8</td>
<td>TRBV21</td>
</tr>
<tr>
<td>TRBJ2.9</td>
<td>TRBV22-1</td>
</tr>
<tr>
<td>TRBJ2.10</td>
<td>TRBV23</td>
</tr>
<tr>
<td>TRBJ2.11</td>
<td>TRBV24-1</td>
</tr>
<tr>
<td>TRBJ2.12</td>
<td>TRBV25-1</td>
</tr>
<tr>
<td>TRBJ2.13</td>
<td>TRBV26</td>
</tr>
<tr>
<td>TRBJ2.14</td>
<td>TRBV27-1</td>
</tr>
<tr>
<td>TRBJ2.15</td>
<td>TRBV28</td>
</tr>
<tr>
<td>TRBJ2.16</td>
<td>TRBV29-1</td>
</tr>
<tr>
<td>TRBJ2.17</td>
<td>TRBV30-F5</td>
</tr>
</tbody>
</table>

Wei Zhang et al. 11 SI
Figure S1. Insertion and deletion length distribution for simulated data.
Figure S2. IGH-VDJ Mutation and deletion/insertion analysis on the public sequences. (A) VDJ mutation number statistics. (B) VDJ deletion/insertion length statistics. The data sets were obtained from IMGT/LIGM-DB database([http://www.imgt.org/ligmdb/](http://www.imgt.org/ligmdb/)), searched by “Homo sapiens”, “rearranged”, ”TRB” or “IGH”, and then selected the sequences annotated by manual(Annot. level==”manual”) and annotated by V,D,J genes. So these sequences have fairly high level of annotation confidence.
1. Sample Basical Statistics

<table>
<thead>
<tr>
<th></th>
<th>Rate_of_input(%)</th>
<th>Rate_of_rawdata(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw_seq_number(PE=1)</td>
<td>8126815</td>
<td>-</td>
</tr>
<tr>
<td>Clean_data</td>
<td>7955514 97.89</td>
<td>97.89</td>
</tr>
<tr>
<td>PE_read_merged</td>
<td>7934499 99.74</td>
<td>97.63</td>
</tr>
<tr>
<td>Merged_with_highquality</td>
<td>7806823 98.13</td>
<td>96.06</td>
</tr>
<tr>
<td>V_alignment</td>
<td>7692062 98.53</td>
<td>94.65</td>
</tr>
<tr>
<td>D_alignment</td>
<td>3944006 50.52</td>
<td>48.53</td>
</tr>
<tr>
<td>J_alignment</td>
<td>7509383 96.19</td>
<td>92.40</td>
</tr>
<tr>
<td>VJ_alignment</td>
<td>7419604 95.04</td>
<td>91.29</td>
</tr>
<tr>
<td>CDR3_found_VJ</td>
<td>7313503 98.57</td>
<td>89.99</td>
</tr>
<tr>
<td>CDR3_found_byconserve</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PCR_Sequencing_correct</td>
<td>6569719 89.83</td>
<td>80.84</td>
</tr>
<tr>
<td>Effective_data</td>
<td>6188283 94.19</td>
<td>76.14</td>
</tr>
</tbody>
</table>

----------Note:-----------------------------------------------
Clean_data: filter the Adapter pollution, low quality sequence
Effective_data: filter the sequence: 1. cannot find CDR3;
2. V and J strand conflict; 3. CDR3 less than 0bp;
4. sequence abundance filter.

2. Sample Further Statistics

in-frame: 5986614 96.74
out-of-frame(stop_codon): 33909 0.55
out-of-frame(CDR3_length): 105860 1.71
non-function: 61899 1.00
V_gene_used: 48 100.00
J_gene_used: 13 100.00
V-J_pairing: 558 89.42
Uniq_number(seq_nt,seq_aa): 1152945 926184
Uniq_number(cdr3_nt,cdr3_aa): 204878 182609
Shannon_index(seq,seq_aa): 16.23 15.74
Shannon_index(cdr3_nt,cdr3_aa): 14.47 14.25
Shanono_index(V,J,V-J): 3.84 2.54 6.22
Hyper-mutation(base_rate,seq_rate): 0.00 0.00

Figure S3. Outputs of IMonitor, H-B-01 as an example. Sample basic statistics show the data procedure, from raw data to effective data, such as paired-end reads merged, V(D)J alignment rate. Sample further statistics, show the multiple statistics based on effective data.

14 SI Wei Zhang et al.
Figure S4. H-B-01 sample output figure of IMonitor. (A) Sequence length distribution. (B) Saturation curve, rarefaction studies of sequences. Sub-sequences are randomly selected and observed unique CDR3 number and predicted CDR3 number (Chao1 corrected algorithm) are calculated. (C) CDR3 nucleotide length distribution. (D) CDR3 abundance distribution. (E) CDR3 amino acid frequencies sectional content. (F) Top ten frequency of CDR3 amino acid. (G) Length distribution of V/D/J gene in CDR3 region. (H) Deletion length distribution of V/D/J gene. (I) Insertion length distribution of between V and D gene, D and J gene. (J) Hyper-mutation, Only for Ig. (K), J gene usage. (L) V gene usage. (M) Three-dimensional graph of V-J pairing.
Figure S5. Error characteristics of 6 plasmid mix samples. (A) Error base rate after sequence filtering by different minimal quality value. For example, Q20 means filter the sequence with at least one base quality less than Q20. (B) Removed data rate after sequence filtering by different minimal quality. (C) Mismatch number distribution, raw sequences (Q0, no filtration) and sequences after filtering by minimal quality 20(Q20). (D) Error base distribution with base quality. Only unique sequences are considered.
Figure S6. V-J pairing dynamics for M002. Day 0 for pre-treatment, Day 15 and Day 33 for post-treatment.
Figure S7. MiTCR and IMonitor performance in 3 spiked-in samples.
Figure S8. Nucleotide composition of V/J genes. (A) H-H-01 sample. (B) H-B-01 sample.
Minimal Residual Disease Detection and Evolved IGH Clones Analysis in Acute B Lymphoblastic Leukemia Using IGH Deep Sequencing

Jinghua Wu1,2, Shan Jia3†, Changxi Wang1,2, Wei Zhang1,2, Sixin Liu3, Xiaojing Zeng1,2, Huirong Ma3, Xiuli Yuan4, Yuanping Du2, Xiaodong Wang3, Xueyu Hong1,2, Xuemei Li1,2, Feiqiu Wen3, Xun Xu1,2, Jianhua Pan4, Changgang Li3* and Xiao Liu1,2,5*

1BGI-Shenzhen, Shenzhen, China; 2China National Genebank-Shenzhen, BGI-Shenzhen, Shenzhen, China; 3Hematology and Oncology Department, Shenzhen Children’s Hospital, Shenzhen, China; 4KingMed Diagnostics, Guangzhou, China; 5Department of Biology, University of Copenhagen, Copenhagen, Denmark

Acute B lymphoblastic leukemia (B-ALL) is one of the most common types of childhood cancer worldwide and chemotherapy is the main treatment approach. Despite good response rates to chemotherapy regimens, many patients eventually relapse and minimal residual disease (MRD) is the leading risk factor for relapse. The evolution of leukemic clones during disease development and treatment may have clinical significance. In this study, we performed immunoglobulin heavy chain (IGH) repertoire high throughput sequencing (HTS) on the diagnostic and post-treatment samples of 51 pediatric B-ALL patients. We identified leukemic IGH clones in 92.2% of the diagnostic samples and nearly half of the patients were polyclonal. About one-third of the leukemic clones have correct open reading frame in the complementarity determining region 3 (CDR3) of IGH, which demonstrates that the leukemic B cells were in the early developmental stage. We also demonstrated the higher sensitivity of HTS in MRD detection and investigated the clinical value of using peripheral blood in MRD detection and monitoring the clonal IGH evolution. In addition, we found leukemic clones were extensively undergoing continuous clonal IGH evolution by variable gene replacement. Dynamic frequency change and newly emerged evolved IGH clones were identified upon the pressure of chemotherapy. In summary, we confirmed the high sensitivity and universal applicability of HTS in MRD detection. We also demonstrated the higher sensitivity of HTS in MRD detection and investigated the clinical value of using peripheral blood in MRD detection and monitoring the clonal IGH evolution. In addition, we confirmed the high sensitivity and universal applicability of HTS in MRD detection. We also reported the ubiquitous evolved IGH clones in B-ALL samples and their response to chemotherapy during treatment.

Keywords: acute B lymphoblast leukemia, minimal residual disease, high throughput sequencing, IGH, clonal evolution

Abbreviations: ALL, acute lymphoblastic leukemia; B-ALL, B cell lineage ALL; BM, bone marrow; CDR3, complementarity determining region 3; FCM, flow cytometry; FR3, framework region 3; HTS, high throughput sequencing; IGH, immunoglobulin heavy chain; MRD, minimal residual disease; ORF, open reading frame; PB, peripheral blood; RQ-PCR, real-time quantitative polymerase chain reaction; TCR, T cell receptor.
INTRODUCTION

Acute lymphoblastic leukemia (ALL) is one of the most common malignant disease in children worldwide and represents approximately a quarter of cancer diagnoses among children younger than 15 years. During the past 25 years, there has been a gradual increase in ALL incidence according to the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program, as well as published reports (1). About 85% of pediatric ALL is classified as B cell lineage ALL (B-ALL). With the improvement of treatment strategies, the 5-year survival rate of ALL has greatly increased from 57.2% during 1975–1977 to 91.8% during 2004–2010 according to the cancer statistics by National Institutes of Health (NIH). In spite of increased cure rates in pediatric ALL, relapse still occurs in approximately 15–20% of patients (2), and minimal residual disease (MRD) is the main risk factor for ALL relapse. Therefore, MRD level has become an important clinical index for doctors to assess treatment response, to adjust the treatment strategy during ALL therapy, and to predict relapse after treatment (3–8). On the other side, some patients are likely to be over-treated and experience serious side effects of chemotherapy. Thus, it is important to consider increasing cure rates together with decreasing toxicity during treatment, and monitoring MRD can provide great help.

Flow cytometry (FCM), fusion genes analysis, and molecular analysis of immunoglobulin (Ig) or T cell receptor (TCR) gene rearrangements are three principal methods for MRD detection in childhood ALL. FCM is a method based on the immunophenotype, which distinguishes the ALL cells from the normal leukocytes by the cell markers they expressed (9). FCM has become a widely used method for clinical investigations due to the rapidity of method, but the immunophenotypic shifts induced by chemotherapy could affect the confidence and accuracy of MRD detection (10, 11). Detection of MRD based on the gene fusions caused by chromosomal translocation can achieve very high sensitivity, but this method only applicable to ALL patients with chromosomal translocations that are identified only in a minority of patients (12). Another ALL MRD detection method is based on the antigen receptor gene rearrangement by real-time quantitative polymerase chain reaction (RQ-PCR) (13). This method is also very sensitive (14), but is laborious and time-consuming due to the need to design patients-specific primers (15). Besides, false-negative result due to clonal evolution is a major problem using unique Ig or TCR rearrangement as RQ-PCR target for MRD detection (7). Several studies have reported the application of high throughput sequencing (HTS) of TCR repertoire or Ig repertoire in MRD detection (16–18). Bone marrow (BM) is the most used specimen for MRD detection in leukemia; however, it has been proposed that peripheral blood (PB) might represent a more convenient specimen for monitoring MRD. For acute myeloid leukemia (AML), Maurillo et al. showed that PB could be used as the cell source for MRD detection using FCM, and MRD levels in PB were positively correlated with MRD measured in BM (19). Coustan-Smith et al. (20) demonstrated PB may be used to monitor MRD in T-lineage ALL patients. By contrast, in B-lineage ALL patients, PB did not work well for MRD detection, but still provided valuable prognostic information (20). However, these studies did not analyze and describe the feasibility and clinical significance of MRD detection using PB by HTS.

Tumors, including hematological tumors, are composed of heterogeneous subpopulations, and tumors with higher heterogeneity had higher rates of resistance to chemotherapy compared with tumors with lower phenotypic variability (21). Some of the subpopulations were produced due to on-going cancer clonal evolution in disease development and disease treatment. The massive evolution of the leukemic cell plays a pivotal role in disease progression and relapse (22), and a prior study reported the mechanisms of clonal evolution in B-ALL (23). The evolution related to V gene replacement in pre-treatment B-ALL samples and the clinical significance has been reported by several studies (17, 24). However, to our knowledge, the extent and dynamics of the clonal evolution of leukemic cells during treatment, as well as its clinical significance, have never been reported.

In this study, we investigated the immunoglobulin heavy chain (IGH) repertoires of 51 B-ALL patients before and during the chemotherapy using HTS technology. We identified disease-bearing IGH rearrangements in 92.2% of the patients and half of the patients harbored two or more leukemic IGH clones. Those clonal B cells were in the early developmental stages. We also analyzed the IGH repertoires of nine diagnostic samples in the RNA level and the result showed that all the disease-causing clones, including those without function, can transcribe mRNA. Next, we demonstrated the high sensitivity of HTS technology in MRD detection and the prognostic value of MRD detection using PB samples. Lastly, we analyzed the evolved IGH clones induced by V gene replacement, and at the same time evaluated the changes of the numbers and the frequencies of those evolved IGH clones during treatment.

MATERIALS AND METHODS

Patients and Samples

We studied 51 childhood B-ALL patients diagnosed in Shenzhen Children's Hospital. BM specimens and (or) PB were obtained at diagnosis and during treatment, and in total 169 specimen were collected (Table S1 in Supplementary Material). The study was carried out in accordance with the recommendations of Declaration of Helsinki and was approved by BGI-IRB. Written informed consent was obtained from the parent(s) or guardian(s) of each child. All the BM and PB specimens were collected in heparin and stored at −80°C until analysis was conducted. High-quality gDNA were extracted from the frozen BM and PB samples using DNA Blood mini kit (QIAGEN, Cat. no.51106).

Disease Risk Stratification

B-ALL patients were stratified into three risk groups according to the following criteria:

Standard risk (SR): (1) age at diagnosis between 1 and 6 years; (2) WBC < 20 × 10^9/L; (3) good prednisone respond (GPR) at 7 days treatment, peripheral blasts < 1.0 × 10^9/L at day 8; (4) BM aspiration results M1 (blasts < 5%) or M2 (blasts 5% ~25%) at day 15 post induction; and (5) BM aspiration results M1 at day 33 post induction.
Intermediate risk (IR); (1) age at diagnosis <1 year or ≥6 years; (2) WBC ≥ 20 × 10^9/L; (3) GPR; (4) BM aspiration results M1 or M2 at day 15 post induction; (5) BM aspiration results M1 at day 33 post induction; (6) reach to SR but BM aspiration results M3 (blasts > 25%) on day 15 of induction therapy, and BM aspiration results M1 on day 33 of induction therapy.

High risk (HR); (1) IR but BM aspiration results M3 on day 15 of induction therapy; (2) poor prednisone respond (PPR), peripheral blasts > 1.0 × 10^9/L at day 8; (3) BM aspiration results M2 or M3 at day 33 of induction; (4) t(9:22) (BCR/ABL) positive or t(4:11) (MLL/AF4); (5) testicle leukemia at diagnosis and did not clinically resolve by day 33 of chemotherapy; (6) large mediastinum mass at diagnosis, and did not completely resolve by day 33; (7) central nervous system leukemia (CNSL) at diagnosis.

### Immunophenotyping and MRD Detection by FCM

Immunophenotyping and MRD detection by FCM was performed at KINGMED CENTER FOR CLINICAL LABORATORY. The red blood cells of BM samples were lysed to get nucleated cells. For MRD detection, BM nucleated cells were stained with the following monoclonal antibody combinations: (1) CD58 FITC, CD34 PE, 7AAD PerCP-CY5.5, CD10 PE-Cy7, CD19 APC, CD38 V450, CD45 V500; and (2) CD66c FITC, CD13 PE, CD34 PE, 7AAD PerCP-CY5.5, CD10 PE-Cy7, CD19 APC, CD15 V450, CD45 V500. The labeled cells were analyzed using a BD CANTO II flow cytometer (Becton Dickinson, San Jose, CA, USA) with at least 100,000 events were acquired. Data were analyzed using BD FACSDiva (Becton Dickinson) and FCS Express (De Novo Software, Los Angeles, CA, USA) software. The compensation matrix was set up using BD CompBeads (Becton Dickinson) for fluorochrome-conjugated antibodies. Quality control was performed using BD Cytometer Setup and Tracking Beads (Becton Dickinson).

### High Throughput Sequencing of IGH Repertoire

The complementarity determining region 3 (CDR3) of the variable regions of IGH was amplified by multiplex PCR. Concretely, the complete VDJ rearrangements of IGH were amplified from 1200 ng gDNA with 12 degenerate forward primers annealed to the 55 functional variable genes (V) and 4 reverse primers annealed to the 6 functional joining genes (J) listed in the IMGT database. The primers have been carefully evaluated to minimize PCR bias, and were listed in Table S4 in Supplementary Material. PCRs (50 μL) were set up at 25 μL of 2x QIAGEN Multiplex PCR master mix, 5 μL of QIAGEN Q solution, 1 μL of 10 μM forward primer pool, 1 μL of 10 μM reverse primer pool, and 18 μL of 67 ng/μL gDNA. The reaction cycling conditions were: 95°C for 15 min, 30 cycles at 94°C for 30 s, 60°C for 90 s, and 72°C for 30 s, followed by a final extension at 72°C for 5 min. The target amplified products (120–200 bp) was purified by electrophoresis on 2% agarose gel, and the Illumina Hiseq sequence adaptors were ligated. Then the sequencing libraries were sequenced with standard 2 × 150 paired end reads on Illumina Hiseq2500 platform.

### Analyses of the Sequencing Data

Sequencing data were analyzed by the TCR and BCR repertoire analyzing pipeline IMonitor (25). About 2–5 million mapped reads (include correct V and J genes) remained after the above pipeline for each sample, and we used a random subset of 2 million mapped reads for continued analysis. In order to remove the false IGH clones derived from sequencing error or contamination from other samples in the same Illumina sequencing lane, we filtered the clonotypes with supported reads lower than the inferred average sequencing depth for every B cells in each sample. The average sequencing depth was inferred as following: first, the total input cell number for amplification was 2 × 10^6 cells, equating to 1200 ng input gDNA. Second, we calculated the B cell number by multiplying the total input cell number by the B cell percentage (detected by FCM according the CD19 marker). Third, the average sequencing depth was inferred by dividing the total used reads (2 million) by the B cell number used in the sample. The CDR3s were classified as non-functional if frame-shift mutations or stop codons existing in CDR3 region.

### MRD Detection by HTS

Similar to previous studies, we defined the CDR3s in diagnostic samples with frequency higher than 10% as leukemic clones (17, 18). In the post-treatment samples, we determined the MRD by calculating the frequencies of identical CDR3 sequences with the leukemic clones in the level of total nucleated cells. The MRD is calculated by multiplying the frequencies of leukemic clones by B cell percentage in the total nucleated cells, which is determined by FCM.

### Identifying the Evolved IGH Clones of Leukemic Clones

The method used by Gawad et al. (24) was used to identify the evolved IGH clones from the leukemic clones with minor modification. Concretely, the evolved IGH clones were defined as: (1) identical J sequence with the leukemic clone; (2) share at least eight bases of identical NDN sequence with the leukemic clone; (3) different V gene from the leukemic clone; (4) more than three mismatches in CDR3 sequence if the evolved IGH clone had the same length as the leukemic clone.

### RESULTS

#### Clinical Characteristics of the B-ALL Patients

We randomly, without any prior knowledge of the clonal IGH rearrangement existence, collected 51 pediatric B-ALL patients in total. According to the disease risk stratification criteria, 19 of them were classified as SR, 18 as IR, and 14 as HR (Table S1 in Supplementary Material, details in the section “Materials and Methods”). The characteristics and clinical information of the patients, including diagnostic age, gender, and cytogenetics were collected and listed in Table S1 in Supplementary Material. Ninety-six percent of patients (49/51) were aged between 1 and 10, and 2 patients were more than 10 years old. Among the patients, 35.3% (18/51) were females, and 23.5% (12/51) were diagnosed
with TEL-AML1 gene fusion. All the 51 patients showed good treatment result after 2–3 months of chemotherapy.

**Clonal Leukemic IGH Identification in Pre-Treatment BM Samples**

According to the definition of leukemic clones (see Materials and Methods), we identified leukemic IGH CDR3s with complete V-D-J rearrangement in a frequency of above 10% in the diagnostic BM samples in 47 of 51 patients (92.2%). Of the 47 patients with leukemic IGH clones, 24 had two or more leukemic clones (Table 1; Table S2 in Supplementary Material). Therefore, we identified 77 leukemic clones from the 47 patients in total (Table S2 in Supplementary Material). The four patients without identification of leukemic clones did not show any significantly distinct immunophenotypes. They may harbor incomplete VDJ rearrangement, or use the pseudo genes to recombine VDJ which cannot be amplified by our multiplex PCR primer set. We then investigated if the number of leukemic clones related with the disease risk. The result showed that more patients in SR group have just one disease clone; however, more patients in HR group have two or more disease clones (Figure 1).

Because B-ALL is a kind of cancer originating from the immature stage of B cells, we investigated if the leukemic IGH rearrangements underwent positive selection. Of the 77 leukemic clones, 28 (about one-third) were functional with correct open reading frame (ORF) in the CDR3, and the other 49 were classified as non-functional due to the occurrence of frame-shift or stop-codon mutations in the CDR3 (Table S2 in Supplementary Material). The ratio of functional to non-functional clones was proportional with that expected by chance. The data above demonstrated that these B cells were immature precursor B cells without experiencing positive selection. We next compared the mRNA transcription of the functional and non-functional leukemic clones by sequencing the IGH repertoire of RNA extracted from the same BM specimen of nine patients. All the leukemic clones, including the non-functional clones, were detected in RNA with high frequency. Surprisingly, the transcription of non-functional leukemic clones was not eliminated or even decreased compared with the functional clones (Figure S1 in Supplementary Material), implying the resistance to the nonsense mRNA decay pathway (26, 27).

However, compared with the frequencies in DNA samples, the frequencies of most leukemic clones in RNA decreased largely, both for functional and non-functional ones (Table 2), which could be due to decreased transcription of the pre-B cells (28), including B-ALL leukemic cells, and it is consistent with other report (29).

**TABLE 1 | Leukemic clones identification in 52 B-ALL patients.**

<table>
<thead>
<tr>
<th>Disease IGH clone</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>7.8</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>45.1</td>
</tr>
<tr>
<td>&gt;1</td>
<td>24</td>
<td>47.1</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>100</td>
</tr>
</tbody>
</table>

**FIGURE 1 | Correlation of B-ALL disease risk with the number of leukemic clones in diagnostic sample.** HR, high risk; IR, intermediate risk; SR, standard risk; “mono_clone,” just one leukemic clone in the diagnostic sample; “poly_clone,” two or more leukemic clones in the diagnostic sample.
Comparison of MRD Results Detected by HTS with FCM in Follow-up BM Samples

With the precise IGH CDR3 sequence information of the leukemic clones, we determined the frequencies of the leukemic clones in the follow-up BM samples. MRD based on the total nucleated cells was calculated by multiplying the frequency of the leukemic clone by the percentage of the B cells in total nucleated cells, which was determined by FCM for those samples. MRD levels and leukemic clone frequencies decreased as treatment progressed (Table S2 in Supplementary Material and Figure S2 in Supplementary Material). With the purpose of evaluating the reproducibility of MRD detection by HTS, we performed two replicated MRD detections from equal BM specimen for three samples of two patients (P049-15, P049-33, and P051-15) and the result showed good reproducibility (Table S3 in Supplementary Material).

In order to assess the accuracy of the MRD detection by IGH repertoire HTS, we compared the detection results of HTS with the results detected by FCM, which were performed as the golden standard for MRD detection. All 34 MRD positive samples by FCM were detected as positive by HTS with similar MRD levels. Twenty-eight samples were detected as MRD positive by HTS, but MRD negative by FCM. The other 19 samples were MRD negative by both HTS and FCM. No MRD positive samples by FCM were detected negative by HTS, which demonstrated the higher sensitivity of HTS (Figure 2).

Identification of Leukemic Clones and MRD Detection in the PB Samples

Because BM aspiration can be painful, using PB instead of BM was evaluated for MRD detection. For 15 patients, we collected

---

**TABLE 2** The frequency of leukemic clones in DNA and RNA of the same bone marrow samples.

<table>
<thead>
<tr>
<th>Patients</th>
<th>Leukemic clone ID</th>
<th>Leukemic clones frequency in DNA (%)</th>
<th>Leukemic clones frequency in RNA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P001</td>
<td>P001B</td>
<td>35.79</td>
<td>10.94</td>
</tr>
<tr>
<td></td>
<td>P001A</td>
<td>15.42</td>
<td>4.30</td>
</tr>
<tr>
<td></td>
<td>P001D</td>
<td>14.82</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>P001C</td>
<td>14.80</td>
<td>4.59</td>
</tr>
<tr>
<td>P005</td>
<td>P005</td>
<td>84.56</td>
<td>34.53</td>
</tr>
<tr>
<td>P006</td>
<td>P006</td>
<td>78.62</td>
<td>30.36</td>
</tr>
<tr>
<td>P007</td>
<td>P007A</td>
<td>45.52</td>
<td>27.39</td>
</tr>
<tr>
<td></td>
<td>P007B</td>
<td>40.96</td>
<td>1.16</td>
</tr>
<tr>
<td>P008</td>
<td>P008</td>
<td>90.87</td>
<td>20.19</td>
</tr>
<tr>
<td>P013</td>
<td>P013</td>
<td>25.12</td>
<td>11.79</td>
</tr>
<tr>
<td>P011</td>
<td>P011A</td>
<td>62.80</td>
<td>32.14</td>
</tr>
<tr>
<td></td>
<td>P011B</td>
<td>18.44</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>P011C</td>
<td>11.60</td>
<td>17.85</td>
</tr>
<tr>
<td>P015</td>
<td>P015A</td>
<td>27.64</td>
<td>10.02</td>
</tr>
<tr>
<td></td>
<td>P015B</td>
<td>10.59</td>
<td>4.30</td>
</tr>
<tr>
<td>P009</td>
<td>P009</td>
<td>87.59</td>
<td>19.61</td>
</tr>
</tbody>
</table>

*Underline in the clone frequency means those clones with correct ORF.*

---

**FIGURE 2** Comparison of MRD results detected by high throughput sequencing (HTS) with flow cytometry (FCM) assays. The samples were classified into three subsets: (1) MRD was detected positive by both HTS and FCM (left); (2) MRD were detected positive by HTS, but negative by FCM (middle); and (3) MRD was detected negative by both methods (right).
both diagnostic BM and PB samples (Table S1 in Supplementary Material). We investigated if the leukemic clones identified in BM also existed in PB with high frequencies. The results revealed that all leukemic clones could be detected in PB with frequency of above 5%, except one clone, which was 18.405% in BM, but was just 0.810% in PB (Figure 3A). The Pearson correlation coefficient between the frequencies of leukemic clones in diagnostic PB and BM was 0.7860. Therefore, PB can be used instead of BM in disease diagnosis and in leukemic clone identification for most patients.

To test MRD detection in PB specimens and to compare it with BM, post-treatment BM and PB samples were collected simultaneously in 15 pairs from 11 patients (Table S1 in Supplementary Material). We determined the frequencies of leukemic clones using the same method. In total, 22 leukemic clones were identified in pre-treatment BM samples for the 11 patients, and 18 clones were positive in post-treatment BM. When detecting the 18 clones in the corresponding PB samples, 12 clones were detected and other 6 were PB negative. When evaluating their frequencies in BM, 11/12 PB positive clones had relatively higher frequency than the 6 PB negative clones. Three leukemic clones were MRD negative in both PB and BM; one was positive in PB with clonal frequency of less than 0.01%, but was negative in BM. We then assessed if the leukemic clone frequency in PB correlated positively with that in BM, and a weak correlation was found with a Pearson correlation coefficient of 0.322 (Figure 3B). Overall, although PB could be used to detect MRD to some extent, the sensitivity of MRD detection using PB was still lower than using BM and it resulted in false negative for some samples.

**Monitoring the Evolved IGH Clones of the Leukemic Clones Before and during Therapy**

On-going change at the IGH gene is one of the most important ways leading to the diversity of cancer cells in B-ALL. After finishing IGH variable (V) gene, diversity (D) gene, and joining (J) gene rearrangement during B cell development (30), the V segments upstream of the used V gene are still reserved, which make the V gene replacement possible by rearranging an upstream V gene with the complete V–D–J rearrangement. Because the loss of 5′ recombination signal sequences (RSS) in D segment of the rearranged V–D–J exon, cryptic RSS within the used V gene is involved in V gene replacement, which effectively increases the length of the V–D–J junction of the new rearrangement (31). The identified evolved IGH rearrangements of leukemic clones in this paper could be generated by three different mechanisms. The first and second mechanisms are based on V replacement in IGH gene. The first mechanism is that the leukemic clone is the ancestral clonotype and the evolved IGH clones derived from continuous V replacement. In this situation, the V segment of the evolved IGH clones will be located at the upstream of the leukemic clone, and the CDR3 length of evolved IGH clones would be longer than that of leukemic clone (31). The second mechanism is that the leukemic clone is produced by V replacement from a more ancestral pre-leukemic clonotype. In this situation, the V segment of leukemic clone will located at the downstream of the ancestral clone, and the CDR3 length of the leukemic clone would be longer than the ancestral clone. The third mechanism is that the ancestral (pre-leukemic) clone finished an incomplete D–J rearrangement of IGH locus, and the leukemic clone and other
evolved IGH clones underwent independent IGH V rearrangement. Therefore, the CDR3 length of the leukemic clone and the evolved IGH clones would be comparable.

To explore the roles of these mechanisms, we assessed and compared the V segments used and the CDR3 length between the leukemic clones and the evolved IGH clones. In total, 51.62% of evolved IGH clones were mainly replaced by upstream V genes compared to 19.67% by downstream V segments (Figures 4A,B). For the evolved IGH clones using upstream V segments, most of the CDR3s were longer than the corresponding leukemic clones; however, the CDR3 lengths of the evolved IGH clones using downstream V segments were comparable with the leukemic clones (Figure S3 in Supplementary Material). The above result demonstrated that the evolved IGH clones we identified were produced mainly by the first mechanism; in addition, the third mechanism also contributed to the evolved IGH clones. Because the pre-leukemic clones were also important in leukemic development and could lead to relapse of disease, we included all those evolved IGH clones in subsequent analysis.

In the diagnostic BM samples, we identified evolved IGH clones in most of the patients, and only two patients did not have evolved IGH clones. We also found that four leukemic clones were produced due to V gene replacement of the other four leukemic clones in three patients (Table S2 in Supplementary Material). The number of identified evolved IGH clones varied from 0 to 4558 for each leukemic clone (Figure 4A), with 14 (18.2%) of the leukemic clones did not experience IGH evolution. When investigating the ORF of the evolved IGH clones, only 11.5% of those clones were functional with correct ORF, which is much lower than the proportion of the leukemic clones. The total frequency of these clones varied from 0 to 38.55% for each leukemic clone (Figure 4C). There was a positive correlation between the

FIGURE 4 | The evolved clones in pre-treatment and post-treatment bone marrow (BM) samples. (A,B) The number of evolved clones in pre-treatment (A) and follow-up (B) BM samples. Different colors indicate different relative position (green for upstream, orange for downstream, and blue for uncertain) for the evolved clonotypes compared to the leukemic clones. (C,D) The correlation between the leukemic clonal frequencies and the evolved clones in pre-treatment (C) and post-treatment (D) samples. The size of the point indicates the unique number of evolved clones from each leukemic clone, and colors in (D) indicated the time of BM obtained after the beginning of chemotherapy.
a significant number of evolved chemotherapy in the begin of treatment. Interestingly, we found IGH positive selection of the evolved samples (Figure S5 in Supplementary Material), which may imply the day 15 samples showed higher ratio than the diagnostic treatment for most of clones; however, for several clones, IGH clones to chemotherapy. The ratios of total frequencies of evolved IGH clones showed very weak correlation with the frequency of leukemic clones (Pearson’s r = 0.0490) and total frequencies (Pearson’s r = −0.0985) of evolved IGH clones was not associated with the frequency of the cancer clone (Figure 4C). The association of evolved IGH clones with clinical features is another important question; however, we did not identify their correlation with disease characteristics in our dataset (Figures S4A,B in Supplementary Material).

We next investigated the dynamic change of the evolutionary IGH clones following chemotherapy treatment in BM. Forty-six percentage of the MRD positive samples contained evolved IGH clones, and we did not detect evolved IGH clones in MRD negative samples (Figure 4D). We also investigated the ORF of these evolved IGH clones, and the results showed only 12.7% of those clones were functional with correct ORF, which was similar to the proportion of evolved IGH clones in diagnostic samples. The total frequency of these evolved IGH clones showed very weak correlation with the frequency of leukemic clones (Pearson’s r = 0.2203, Figure 4D). Tracking the fluctuation of frequency ratios of the evolved IGH clones to the leukemic clones revealed the imbalanced response of leukemic clones and evolved IGH clones to chemotherapy. The ratios of total frequencies of evolved IGH clones to the frequencies of leukemic clones decreased during treatment for most of clones; however, for several clones, the day 15 samples showed higher ratio than the diagnostic samples (Figure S5 in Supplementary Material), which may imply positive selection of the evolved IGH clones upon the pressure of chemotherapy in the begin of treatment. Interestingly, we found a significant number of evolved IGH clones newly emerged in the post-treatment samples (Figure S6 in Supplementary Material), even though these clones eventually disappeared with the negative detection of MRD after chemotherapy.

We compared the number and total frequency of evolved IGH clones in PB and BM of the same patients. The result showed that the number and total frequency of evolved IGH clones in PB were correlated positively with that in BM (Pearson’s r = 0.9941 for number; Pearson’s r = 0.7154 for total frequency). However, for most samples, the total frequency of evolved IGH clones in PB was lower than that in BM (Figure 5). This result revealed that PB has the potential to serve as a non-invasive specimen to monitor clonal IGH evolution during therapy, but it may underestimate the frequency of evolved IGH clones.

**DISCUSSION**

Leukemia is a clonal disease, and pediatric B-ALL is derived from an early stage of B cell development. During early B cell development, the germline variable (V), diversity (D) and joining (J) gene segments of Ig recombine and, therefore, each B cell obtains a particular combination of V–D–J segments (30). The random deletion at the end of germline segments and the random insertion of nucleosides between the V–D and D–J gene segments during rearrangement creates huge diversity of CDR3 of the IGH gene. According to the above production mechanism of the antibody diversity, the CDR3 sequence of the IGH can be used as a unique proxy to distinguish different B cells. IGH sequence has been used as a tumor-specific biomarker to detect MRD for decades by conventional polymerase chain reaction (PCR) or RQ-PCR (7, 32, 33). Since the development of immune repertoire HTS technology in 2009 (34–36), several studies have tried to detect MRD of pediatric ALL, including T-ALL (36) and B-ALL (17, 18) using this technology. In order to evaluate the wider applicability of this technology, we detected the MRD
of B-ALL patients with different treatment days using immune repertoire HTS. We also investigated the evolved IGH clones in pre-treatment and post-treatment samples.

We identified clonal IGH rearrangements in 92.2% of the unselected cohort of childhood B-ALL patients (Table 1). The four patients without clonal IGH rearrangement detected did not show any difference in clinical phenotype and FCM result. It is possible that these leukemic cells were in an earlier developmental stage in those patients, just finished the D–J rearrangement, but not the complete V–D–J rearrangement, which has been observed in previous studies (17, 18, 37). It is also possible those leukemic cells were in the earliest recognizable B-lineage cells, which did not begin the rearrangement. Previous study had also reported that most of the childhood B-ALL (>83%), but not all the patients, had IGH gene rearrangement (7, 33). Therefore, even though the high sensitivity of MRD detection using HTS, the limitation in failure to diagnose the pro-B-ALL, which consists of about 10% of B-ALL patients, indicated that FCM was still needed to accurately diagnose B-ALL and detect MRD for those patients.

Importantly, we found that disease risk of B-ALL is related to the number of leukemic clones in diagnostic samples for the first time. Concretely, sample with more leukemic clones indicates higher disease risk, which means slower reaction to chemotherapy reagent (Figure 1). More leukemic clones may imply higher RAG1 and RAG2 activity, which could not only target the Ig gene but also the non-Ig gene (such as tumor suppressor gene IKZF1, CRLF2, BTG1, etc.), and lead to the genomic instability according to Swaminathan et al. (23). They reported that high RAG1 mRNA expression predicted poor ALL patient outcome, which could explain our result.

During the V–D–J rearrangement of IGH in B cell development, the random deletion and addition of nucleotides between V–D and D–J segments at the time of joining could result in variability at the junctions and different ORF in CDR3. The possibility that the length of the CDR3 is exact multiple of three nucleotides is one-third expected from random rearrangement. Therefore, in theory, only one in three IGH rearranged B cells could make an in-frame and functional rearrangement before selection, which has been demonstrated previously (38). In our study, about one-third of the disease clones had in-framed CDR3 ORF; and two-thirds of the disease clones had frame-shift mutation in the CDR3 (Table S2 in Supplementary Material). Therefore, our data are consistent with random recombination and suggested that leukemic cells were in the very early development stage without experiencing positive selection.

In the normal B cell development, precursor-BCR (pre-BCR) checkpoint controls critical B-cell developmental processes and is essential to make sure that the B cells develop into mature B cell (39–41). The pre-BCR has to be expressed on the cell surface to induce the B cell to pass this checkpoint. In order to pass the checkpoint, the B cell should complete in-frame IGH VDJ recombination to express a functional Pre-BCR. If a precursor B cell possesses an out-of-frame IGH rearrangement, there is no pre-BCR expressed on the surface of the B cell, therefore, the cell does not receive necessary survival signals, and it will undergo a programmed cell death. In our study, we found that two-thirds of the leukemic cells do not have functional IGH V–D–J rearrangement (Table S2 in Supplementary Material), but they survived the selection and proliferated to clonal cells. A similar result was also observed in another study in BCR-ABL1 positive B-ALL cells (42), which reported only 3 out of 12 BCR-ABL1 positive B-ALL cases harbored potentially functional IGH gene rearrangements. Several studies have reviewed that leukemic B cells could manage to evade the pre-BCR checkpoint and avoid clonal extinction by mimicking or by-passing the Pre-BCR signaling pathway (43–45).

As to MRD detection, our data demonstrated the robustness of HTS similar to previous studies (16–18). Compared to FCM, HTS did not show false-negative results, and the HTS method can detect MRD at levels lower than 0.01%, which cannot be detected by FCM (Figure 2). In general, our study confirmed the wide applicability of HTS in MRD detection and higher sensitivity than FCM. There were eight post-treatment samples, whose MRD were higher than 0.1% by HTS, but were reported negative in the FCM detection. Wu et al. discovered a similar phenomenon in their study, and verified and explained that this is possible due to immunophenotypic normalization induced by therapy (17). Because most the patients in our cohort were followed up for less than 2 years, and no patient relapsed, we cannot assess the clinical significance of the higher MRD detection sensitivity using HTS in this panel. Wu et al. discussed the prognostic significance by reviewing previous studies, and concluded that very low-level MRD detected by HTS is likely to be meaningful and, therefore, they suggests that these patients should be closely monitored (17).

To date, BM is the most used specimen for MRD detection in leukemia; however, it has been proposed that PB might represent a more convenient specimen for monitoring MRD. In this study, we found that for samples with relatively high leukemic clone frequency in BM (approximately above 1%), PB samples could detect MRD but with relatively lower MRD levels than BM samples (Figure 3), and this result agreed with the Coustan-Smith et al’s study in B-ALL (20). Overall, there is a false-negative rate and the MRD level could be underestimated in PB, which limited the application of PB in clinic for MRD detection today. Larger cohort with more comprehensive design is needed to determine how PB sample should be used directly in clinic. For example, the MRD level in PB 7 day post-treatment may correlate well with the MRD level in BM 15 day post-treatment; therefore, it is possible we could use the MRD level in PB 7 day post-treatment to predict the prognosis of leukemia. Notably, in our study, one negative leukemic clone in BM was found to be positive in PB, which implied PB might provide compensated prognostic value to BM.

Clonal evolution is an important question in leukemia development and treatment. In this study, we identified evolved IGH clones induced by V replacement, and also the correlation of those clones with the frequencies of leukemic clones in diagnostic BM samples (Figure 4C). More importantly, we found that four leukemic clones were produced due to V gene replacement from other leukemic clones (Table S2 in Supplementary Material), which suggested that evolved IGH clones could expand greatly and lead to disease relapse if not eliminated during treatment. We also investigated the reaction of those evolved IGH clones during chemotherapy. Most of the day 15 and part of the day 33 post-treatment samples contained evolved IGH clones.
(Figure 4D) and abundant new evolved IGH clones not present in the diagnostic samples emerged upon the pressure of therapy (Figure S6 in Supplementary Material), which demonstrated that chemotherapy can induce continuous evolution of the leukemic cells and increase the heterogeneity of cancer. When investigating the response of evolved IGH clones to chemotherapy, both the clones identified in pre-treatment samples and those newly generated during treatment decreased during the course of therapy and disappeared when the MRD was negative (Figure 4D). Therefore, those evolved IGH clones were also therapy sensitive in primary cancer, which indicate good prognosis of those patients. However, we cannot make sure if a clone harboring an evolved IGH gene actually is to be equaled to the generation of a new clone with new genomic variants. In order to validate this correlation, single-cell sequencing used by Gawad et al. (46) is required to investigate the genome-wide somatic variants and the IGH DNA sequence of each single cell.

In this paper, we confirmed the wide applicability and higher sensitivity of HTS in MRD detection of B-ALL patients. Considering the invasiveness of BM extraction, we assessed the value of using PB as MRD detection specimen, and suggested that although MRD in PB could not represent the exact MRD level of the patients, PB could be used to evaluate the response of cancer cells to chemotherapy in the beginning of the treatment. During the disease development and disease treatment, cancer cells could evolve and generate high genetic heterogeneity of genome, including IGH gene. We identified many evolved IGH clones in the diagnostic samples. However, nearly all those evolved IGH clones decreased with on-going treatment and disappeared when the MRD became negative in these samples. It is possible some of those evolved IGH clones could lead to relapse of the disease.

DATA ACCESSION

The raw data has been deposited to Sequence Read Archive (SRA) under accession SRA456729.

AUTHOR CONTRIBUTIONS

CL and XL designed the study and supervised research. JW and CW performed research, analyzed data, and wrote the paper. SJ collected the specimens and clinical information. WZ, YD, XL, and XX analyzed the data. XZ and XH performed the experiments. SL, HM, XY, XW, and FW contributed samples. JP checked the FCM MRD detection result. All authors contributed to the preparation of the manuscript and approved the submission in its current form.

FUNDING

This work was supported in part by the Science, Technology and Innovation Committee of Shenzhen Municipality JCYJ201404163004841, JCYJ2014041614331552 and JCYJ20140416141331490.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fimmu.2016.00403

REFERENCES


Conflict of Interest Statement: JP has employment with KingMed Diagnostics. All the other authors declare no competing financial interests.

Copyright © 2016 Wu, Jia, Wang, Zhang, Liu, Zeng, Mai, Yuan, Du, Wang, Hong, Li, Wen, Xu, Pan, Li and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
The different T-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues

Ting Wang1†, Changxi Wang2†, Jinghua Wu2†, Chenyang He1†, Wei Zhang2, Jiayun Liu3, Ruifang Zhang2, Yonggang Lv1, Yongping Li1, Xiaojing Zeng2, Hongzhi Cao2, Xiuqing Zhang2, Xun Xu2, Chen Huang4, Ling Wang1* and Xiao Liu2,5*

Authors Affiliations:
1Department of Vascular and Endocrine Surgery, Xijing Hospital, Fourth Military Medical University, Xi'an, 710032 China.
2BGI-Shenzhen, Shenzhen, 518083 China.
3Institute of Clinical Laboratory Medicine, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, Shaanxi, China.
4Department of Nephrology, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, Shaanxi, China.
5Department of biology, University of Copenhagen, Copenhagen 2200, Denmark

†These authors contributed equally to this work.

*Correspondences: Xiao Liu (liuxiao@genomics.cn) or Ling Wang (vascular@fmmu.edu.cn)

Running title: TCR repertoire of breast tumor and adjacent tissues

Conflict of interest: The authors declare no conflicts of interest.
Abstract

T lymphocytes infiltrate the microenvironment of breast cancer (BC) tumors and play a pivotal role in tumor immune surveillance. Relationships between the T-cell receptors (TCRs) borne by T cells within tumors, in the surrounding tissues, and in draining lymph nodes are largely unexplored in human breast cancer. Consequently, information about the relative extent of possible T-cell exchange between these tissues is also lacking. Here we have analyzed the TCR repertoire of T cells using multiplex PCR and high throughput sequencing of the TCRβ chain in the tissues of tumor, adjacent nontumor, and axillary lymph-nodes of breast cancer patients. T-cell repertoire diversity in tumors was lower than in lymph nodes, but higher than in nontumor tissue, with a preferential use of variable and joining genes. These data are consistent with the hypothesis that most of the T cells in tumors derive from the lymph node, followed by their expansion in tumor tissue. Positive nodes appeared to enhance T-cell infiltration into tumors and T cell clonal expansion in lymph nodes. Additionally, the similarity in TCR repertoire between tumor and nontumor tissue was significantly higher in luminal-like, rather than basal-like, BC. Our study elucidated the high heterogeneity of the TCR repertoire, and provides potential for future improvements in immune-related diagnosis, therapy, and prognosis for BC patients.

Introduction

Tumorigenesis is regulated by immune responses to neoantigens in the tumor microenvironment, including the plasticity of macrophage (1) and suppression of antitumor immune response (2,3). A decade ago, Schreiber proposed the three “E roles” (elimination, equilibrium, and escape) of immune surveillance, which interprets the interaction between tumor and host immune system in tumor progression (4). Lymphocytes are the major antitumor cells in tumor microenvironments and many studies have demonstrated that tumor-infiltrating lymphocytes (TILs), especially CD8+ T cells, are related to improved clinical prognosis in most of the tumor types (5), such as ovarian cancer (6), cervical cancer(7), colorectal cancer (8-10), and breast cancer (11-14). The presence and antitumor function of TILs makes cures of tumor by immunotherapy possible (15,16).

To counteract tumor-induced immunosuppressive microenvironments, several immunotherapy strategies have been developed to enhance the antitumor immune response. Adoptive cell transfer (ACT) is one such strategy, which depends directly on the presence of antitumor T cells, by treating the patients with T cells isolated from the tumor mass and expanded ex vivo (15). Immunotherapies based on the adoptive transfer of TILs have had surprisingly positive results for some patients with metastatic melanoma (16). Another immunotherapeutic strategy directly relying on the existence of TILs is immune checkpoint blockade, which inhibits the interaction of cytotoxic T lymphocyte antigen 4 (CTLA-4) or programmed cell death 1 receptor (PD-1) with their ligands using blocking antibodies. Immune checkpoint blockade can activate the T cells in the microenvironment and has led to durable antitumor effects in patients with metastatic melanoma (17,18), renal cell carcinoma, and non-small-cell lung cancer (18). Checkpoint blockade had such therapeutic efficacy for various types of cancer, that cancer immunotherapy was chosen by the editors of Science as the biggest scientific breakthrough of 2013. However, current immunotherapies do not work for every patient, emphasizing the need to further
understand and monitor the infiltrating T cells in cancer.

Breast cancer (BC) is the second leading cause of cancer death among women in the United States\textsuperscript{19}, and has become one of the most lethal diseases worldwide. Infiltrating lymphocytes are an independent prognostic factor associated with better BC patient survival. Recently, next generation sequencing has been utilized to investigate the T-cell receptor (TCR) repertoire of infiltrating T lymphocyte in renal cell carcinoma(\textsuperscript{19}) and ovarian cancer(\textsuperscript{20}), but not in BC. The clonal diversity of infiltrating lymphocytes in breast tumor tissues, and its comparison with that of adjacent tissues, remains largely unexplored.

The tumor-draining lymph nodes (LNs) are the LNs that lie immediately downstream of tumors, and for breast cancer, axillary LNs are the most important draining LNs that provide significant diagnostic and prognostic value. The degree of overlap of TCR repertoire between the tumor tissue and draining LNs can be used to assess the percentage of infiltrating T cells derived from draining LNs (\textsuperscript{21}). Therefore, we have also investigated the derivation of tumor infiltrating T cells, by comparing the TCR repertoire of tumor-infiltrating T cells with that of breast-draining LN T cells from the same BC patient.

It is reported that the association of TILs with clinical outcome is observed just in basal-like or triple-negative BC, but not in luminal BC (\textsuperscript{13}). One previous study has shown that basal type BC cell lines express more PD-L1 compared to luminal BC cell lines (\textsuperscript{22}). Therefore, it is possible that the TILs in luminal BC are different from those in basal-like BC. Comparing the TCR repertoire of infiltrating T cells of the luminal BC with that of the basal BC can help verify this hypothesis. The shared (public) TCRs observed in multiple individuals, and their association with disease antigens, such as tumor neoantigens, have been investigated for decades (\textsuperscript{23,24}). Recombinational biases and convergent recombination are proposed to contribute to the occurrence of public TCRs (\textsuperscript{23}). In this study, we present the public intratumoral TCRs for BC and incorporate this information in our explanation of the contributions of different mechanisms.

Overall, in this study, we attempt to (1) assess the amount of T-cell infiltration and TCR diversity of adjacent tissues; (2) investigate the percentage of tumor-infiltrating T cells that were derived from draining LN; (3) uncover associations between the TCR repertoire and other clinical features, including node positivity and tumor subtypes; and (4) characterize public TCR response among breast tumors. The results we present could help us to better understand the infiltrating T lymphocytes in BC, be beneficial in the effective design and better monitoring of BC immunotherapy, and allow the improved stratification of the patients for treatment and outcome prediction.

**Materials and Methods**

*Clinical samples*

This study is based mainly on analyses of sixteen female breast cancer patients from Xijing hospital, the Fourth Military Medical University, China. All the tissues were treatment naïve, and were removed during surgery of the patients who hadn’t experienced any chemotherapy. The study
was approved by the Institutional Review Board in Xijing hospital and BGI-Shenzhen. Written informed consents were obtained from every participant. The clinical information of these patients was shown in Table 1. The sixteen patients can be divided into two subgroups (eight in luminal type and the other eight in non-luminal type) according to the expressions of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (Her2) and Ki-67 protein (see below).

Cancer tissue, normal breast tissue, and the draining LN (axillary LN) were collected during the surgery. Spatially different tumor specimens were collected and combined to be more representative of the tumor tissue. All of the solid specimens were cut into small pieces (1 cubic millimeter) immediately after surgery and transferred into a 2 ml freezing tube with preservation solution (RPMI 1640 with 15% fetal calf serum and 10% DMSO). After immersion in liquid nitrogen for 30 min, the samples were placed in -80 °C for long term storage. High quality genomic DNA (gDNA) was extracted from frozen tissues by the common salting-out method. Additionally, 20 breast tumor tissues were selected from the tissue repository in the Xijing hospital, in which 14 were node-positive tumors with LN ratio (the number of positive LNs (pLN) divided by the number of LNs removed during surgery) of more than 50%, and six were node-negative tumors. These tissues were combined with the 16 tumors for analysis, to validate the association between tumor infiltration rate and node-positivity. TCR repertoire data from another independent cohort of 49 patients, who had been consecutively recruited in year 2013 in Xijing hospital, were used for validation (data not published), to compensate for the limited sample size and statistical power in our discovery cohort, for some of the analyses. These tissues were also treatment naïve, and were representative of the hospitalized patients in that period. Samples from this cohort were collected following the same protocol described above.

**Immunohistochemistry and molecular subtype of patients**

Molecular subtypes of BC were classified according to the expression of ER, PR, HER2, and Ki-67, detected by immunohistochemistry stained methods. Luminal A was defined as ER+ and/or PR+, HER2−, Ki-67 less than 14%; luminal B was defined as ER+ and/or PR+, HER2+ or HER2−, and Ki-67 more than 14%; HER2 was defined as ER−, PR−, and HER2+; and triple-negative (TNP) or basal-like subgroup was defined as ER−, PR−, and HER2− (25).

**Infiltrating T lymphocytes quantification**

The T lymphocytes infiltrating in tumor and adjacent nontumor tissues were analyzed by immunohistochemistry using antibody against human CD3, CD4, and CD8 based on formalin-fixed paraffin-embedded slides. Stained slides scoring and quantified by digital scanning. Positive T lymphocytes were visually scored by a pathologist who was blind to the clinical characteristics of the patients. Tumor-infiltrating T lymphocytes were defined as CD3+ cells in tumor tissues including those located within tumor cell nests and in the adjacent peritumoral stroma. The T lymphocytes infiltrating percentage was calculated by dividing the number of total nucleated cells by the CD3+ cells. To assess the reproducibility and reliability of the scoring, all cases were scored again by the same pathologist after a period of time (4 weeks), and were re-scored by a second pathologist. Both the scorings of the same pathologist, and from the two pathologists, showed very high reproducibility (Supplementary Table S1). We used Pearson correlation analysis to represent the reproducibility of the same pathologist, and the coefficient
was 0.996. The intraclass correlation (ICC) is a better indicator to show the reproducibility between two pathologists, and the ICC consistency of our data was 0.986.

**High throughput sequencing and analysis of TCR repertoire**

The third complementary determining region (CDR3) of TCRs were amplified by multiplex PCR and sequenced using methods described previously (26-28). Briefly, gDNA (1500 ng) for each sample were amplified using QIAGEN Multiplex PCR Kit (QIAGEN) with 32 forward primers annealed to the FR3 region and 13 reverse primers annealed to the junction (J) region of TCR published in our previous study (29). The reaction cycling conditions were: 95°C 15 min, 30 cycles of 30 s at 94°C, 90 s at 60°C, and 30 s at 72°C, plus a final extension of 5 min at 72°C. The primers and multiplex PCR reaction have been optimized to minimize the multiplex PCR bias before the beginning of the study, and the bias has been assessed using the synthetic templates (Supplementary Fig. S1). The target amplified product (100–200 bp) was purified by electrophoresis on 2% agarose gel and then were sequenced with standard 2 × 150 paired end reads on Illumina Hiseq2000 platform.

Sequencing data were analyzed by an in-house developed pipeline **IMonitor** (29). A brief summary of the pipeline was as follows: (1) The low quality reads and badly adapter contaminated reads were filtered to get clean reads. (2) The cleaned pair-end reads were merged, and the reads that could not be merged were discarded. (3) The merged reads were aligned to their respective V, D, J germline sequences (IMGT, http://www.imgt.org/) using BLAST. (4) Re-alignment of the correctly mapped reads was performed to select the best V/D/J alignment.

**Statistics**

To study the distribution of TRBV and TRBJ gene of these three tissues of breast cancer patients, pairwise V-segment and J-segment profile comparisons were performed between tumor and other tissues. Because V gene frequencies are not all independent, we employed a permutation test based on the Mann-Whitney U statistics (Shuffled-U Test)(30). The resulting P values were corrected for multiple testing using false discovery rates. Statistical significance for the difference between two groups in other analyses were all determined using Mann-Whitney U test. Paired test was used in case samples from the same patient were compared. For calculation of Shannon diversity index, we included the top 1000 unique clones to minimize the effect of sequencing errors. A P value lower than 0.05 was considered statistically significant. All analyses were performed with R version 3.0.2.

**Data availability**

The raw sequence data of the TCR repertoire from the sixteen patients have been deposited at the NCBI Sequence Read Archive (SRA) under the accession number of SRA455606. Besides, we have uploaded all processed CDR3 sequences analyzed in this manuscript to a public database (Pan Immune Repertoire Database, PIRD, [http://db.cngb.org/pird/](http://db.cngb.org/pird/)), which facilitates sequences query, comparison and further analysis.
**Results**

**Comparison of TCR diversity and gene usage between tumors and adjacent tissues**

Immunohistochemical (IHC) staining and careful scoring of the T lymphocytes in the tissues revealed that the proportion of infiltrated T lymphocytes into the tumor microenvironment varied considerably among tumors and adjacent nontumor tissues (Table 1). Overall, breast tumors exhibited a significantly higher extent of T-lymphocyte infiltration than their paired nontumor tissues ($P = 0.0023$, Fig. 1A). The CD4+ and CD8+ T cells in tumors and nontumors were also analyzed (Supplementary Table S2).

Genomic DNA from the three types of tissues was used to profile the TCR $\beta$ chain by high throughput sequencing. The abundantly diversified TRB CDR3 sequences, which were unlikely to be shared by two unrelated T cells, served as unique barcodes for the T lymphocytes. We first evaluated the diversity of T cells in the neighboring tissues. The Shannon diversity index was used to calculate T-cell diversity (29). The T cell population that had infiltrated tumors was significantly less diverse than that in LN ($P = 6.104 \times 10^{-5}$, Fig. 1B, LN stated here and in the following refers to draining LN), but were more diversified than the matched nontumor tissue ($P = 1.825 \times 10^{-2}$). The diversity of T lymphocytes in nontumor tissue were significantly lower than LN ($P = 1.526 \times 10^{-4}$), and varied extensively among individuals (Fig. 1B). We also calculated the curve of cumulative frequency of total clones with cumulative unique clones (Supplementary Fig. S2A), and the proportion of the abundant TCR clones (> 0.01%) (Supplementary Fig. S2B). The above two indexes demonstrated the same trends, with TCR repertoire diversity in nontumor < in tumor < in LN.

To further explore the properties of infiltrated T cells in tumors, we compared the TRBV and TRBJ gene usage of TCR repertoire in tumor tissue with matched LNs and nontumors. We tested all the V and J gene usage frequencies and identified the biased usage of J1-1, J1-6, J2-7, and J2-4 of J genes, between tumor and LN. Only the V24-1 gene usage discrepancy was observed between tumor and nontumor (Supplementary Fig. S3). Those differences between tumor and other tissue implied the unique V or J usage pattern for TCR in tumors.

**Most tumor-infiltrated T lymphocytes detected in draining LNs and expanded in tumors**

Draining LNs serve as a T-lymphocyte reservoir that includes both naïve and activated lymphocytes, where transferred tumor antigens are presented by antigen-presenting cells to naïve T lymphocytes that become activated if they recognize the antigen. We had a unique opportunity to characterize the T cells that underwent this process by comparing the TRB CDR3s among the tumor and related tissues. Specifically, we found that about 60% of CDR3 rearrangements in tumor samples could be tracked in matched LNs, whereas only 10%–20% of CDR3 rearrangements in the LNs were found in tumor samples. The similar ratio could also be found between adjacent nontumor samples and LNs (Fig. 2). The high proportion of CDR3s in tumor and normal samples that corresponds to LNs, suggests that draining LNs provide a major source for the T-cell infiltration in tumor and normal tissues.

Furthermore, when assessing the clonality of T-cell repertoire in tumors and LNs, we found that most (61.91%–96.42%) T cells in tumors were expanded (> 0.1%) or medium-sized (0.01%–0.1%) clones, whereas the small clones (< 0.01%) were more prevalent in the LNs (Fig. 3A and B). Large and expanded T-cell clones have activated phenotypes (31), so we infer from our results that
the infiltrated T cells in the tumor microenvironment were highly activated and expanded. In contrast, T-cell clonal expansion in the LNs was much lower, where most of the T cells are naïve. Additionally, for the CDR3 sequences identified in both tumors and LNs, which we defined as “migrated clones”, we found that their frequencies were significantly higher in tumors than in LNs for all patients except BC0009A, who was node-positive with high LN ratio (LNR) (Supplementary Fig. S4). These data are consistent with the hypothesis that T lymphocytes represented by migrated clones are less expanded in the LN and then undergo remarkable expansion after they traffic to the tumor microenvironment, where they encounter and recognize tumor antigens. In some node-positive patients with high LN ratios (such as BC0009A), the T lymphocytes may also be activated and expanded in the LNs, perhaps due to more thorough encounter with tumor antigens after tumor metastasis. We also investigated the frequency distribution of migrated T cells versus total T cells, and found that significantly more migrated clones were expanded and high-frequency (expanded and large), in both tumors (Fig. 3B and D, and Supplementary Fig. S5A) and LNs (Fig. 3 A and C, and Supplementary Fig. S5B). This implied that activated, but not naïve T cells, were more likely to be recruited from the LN, and migrated T cells could be further expanded in the tumor environment, where they could modulate immunological functions during tumorigenesis.

**LN positivity correlated with more infiltration of T cells in tumors and expansion in LNs**

The LN ratio (LNR), defined as the number of positive LNs (pLN) divided by the number of LNs removed during surgery, provides an independent score for prediction of prognosis and survival (32,33). We therefore investigated the association among the LNR, number of pLN, degree of infiltration of T cells, and their clonality and expansion.

We found that node-positive tumors had a higher degree of T-cell infiltration than node-negative tumors, though the difference was not significant, perhaps due to the limited number of node-positive samples, and in 3/5 only one pLN was detected ($P = 0.1416$, Fig. 4A). The difference was more apparent in tumors with higher LNR. The two tumors with LNR above 70% had infiltration rate of 40% and 30% (defined as the percentage of CD3+ T cells in the total nucleated cells), respectively (Table 1). To validate this observation, we collected 14 additional tumor samples with high LNR (> 50%) and 6 additional tumor samples from node-negative patients. The two groups shared matched tumor subtypes and grades. Indeed, we found the infiltration ratios of high LNR tumors were significantly higher than the node-negative tumors ($P = 0.00022$, Fig. 4B).

We also noticed that more expanded T cells were present in the LN with high LNR, compared with node-negative patients (Fig. 3A). To validate this observation, data from another cohort of 49 BC patients were analyzed (Supplementary Table S3). We found node-positive patients had significantly higher T-cell expansion in their LNs, and the increment was even more obvious only taking into account of LNs with high LNR (30% were used here as to include more samples in the group for statistical test, Fig. 4C). Other clinical features including tumor subtypes and grades were matched between the node-positive and -negative groups. It implies that the metastatic tumor cells in the LN might enhance activation and expansion of the tumor-associated T cells. We also calculated the TRB Pearson coefficient between tumor and LN, indicative of similarity of T-cell repertoire between the two compartments, and found that the T-cell repertoires between tumor and
LN were more similar in the high LNR patients (Supplementary Fig. S6A).

**Similarity of T-cell repertoire between tumors and nontumors affected by tumor subtype**

To explore the relationship of T lymphocytes in different tissues, we calculated the Pearson coefficient between each pair of tissues and found that T lymphocytes between tumors and nontumors, as well as between LNs and tumors, were significantly higher than between LNs and nontumors (Supplementary Fig. S6B). In addition, the similarities of T-cell repertoire between tumors and nontumors, as well as LNs and tumors, varied extensively in different patients. To investigate if the diversified intertissue T-cell repertoire similarity among patients was related to breast cancer phenotype, the intertissue correlations among the breast cancer subtypes were compared, and the Pearson coefficients between tumors and nontumors were significantly higher in luminal-like subtype (luminal A and B) than in basal-like subtype \((P = 0.03596, \text{Fig. } 5A)\). Data of the independent validation cohort were also analyzed, but only five basal-like tumors were identified out of 49 patients, which provided insufficient power for statistical test. Therefore, we combined the two cohorts together, and demonstrated a significantly higher correlation between luminal-like tumors and paired nontumor tissues \((P = 0.02017, \text{Fig. } 5B)\).

**Public intratumoral TCR response**

The TCR response between individuals for various diseases including cancer can be “public” (found in more than one individual) (23,24). To examine whether any intratumoral TCR clones that responded to breast cancer were public, we compared the large and expanded TRB amino-acid sequences (>0.01%) from the tumors of multiple patients, and filtered the public clones (existing in at least two healthy individuals), with a TCR database collected from peripheral blood of 661 healthy individuals (not published). Thus, we identified 24 TCR amino-acid sequences that were shared by at least two patients (Table 2), and the top clone in the list was found in three out of 16 patients. The frequencies of these public sequences mostly accounted for 0.01% to 1% of the total sequences in an individual. We found that 29.2% (7/24) of public TCRs use TRBJ2-3 for recombination, which greatly surpassed the frequency of its usage for all the intratumor TCRs in various patients (10.46%–15.65%), implying the contribution of combinatorial bias to public TCR usage. In addition, some of the public TCRs were recombined by different germline V genes, which further formed variable nucleotide sequences, and eventually encoded the same amino-acid sequences. This clearly illustrates the role of “convergent recombination” in the development of public TCR response (34). Taken together, our data imply that multiple mechanisms are involved in this process.

**Discussion**

We have presented here a comprehensive investigation on TCR repertoire for breast cancer. Using IHC, and TRB CDR3 deep sequencing, we unraveled an unpredicted degree of heterogeneity of T lymphocyte in tumor and other tissues in breast cancer patients. The presence of tumor-infiltrating lymphocytes (TIL), especially CD8+ T cells, correlates with improved clinical prognosis in most of the tumor types (5). With the help of our complete breast cancer sample set, we not only could assess and compare the extent of TILs in different tissues, but also systematically investigated
TCR diversity and clonality, and correlated it with tumor molecular subtypes. The characteristics and relationship of TCR repertoire between tumor and adjacent tissues, especially LNs, helped us to assess the process of TIL activation, recruitment, expansion, and selection.

Many studies show that the draining LN is where tumor antigens are first presented to T cells by antigen-presenting cells, which will induce the activation and proliferation of tumor-specific T cells (35-37). Then the proliferating T cells infiltrate the tumor microenvironment by complex mechanisms (38). The tumor mass can also serve as a potential site for T-cell activation (37,39). Here we found that T cells in draining LN were more diversified, with fewer large and expanded clones (Fig. 1B and Supplementary Fig. S2), which implied the presence of more naïve cells or unexpanded activated cells. In contrast, TILs had less diversity, with a greater proportion of large and expanded clones (Fig. 1B and Supplementary Fig. S2), which illustrated the universal T-cell activation and proliferation in the tumor microenvironment, where plenty of tumor mass and neoantigens could stimulate the process. Nontumor tissues harbor few T cells, and those present could be tissue-resident memory T cells (TRM) (40) migrating to the nontumor breast tissue due to historical inflammation, or because of minimal migration and the presence of tumor antigens in the nontumor tissues to recruit the T cells.

The relationship between the amount of T cell infiltration into tumors and pLN/LNR has not been reported to date. In our study, we found that node-positive patients tended to have more TILs, and patients with high LNR (> 50%) had significantly more TILs than node-negative patients (Fig. 4A and B). More studies are needed to uncover the mechanism underlying this phenomenon. A possible explanation is that enhanced activation and expansion of T lymphocytes by tumor antigens in pLN increases the probability of T cells crossing vasculature and stromal barriers, thereby reaching the tumor site and encountering tumor cells. In addition, the TRB repertoire was more similar between tumors and LNs in patients with high LNRs, compared with node-negative patients (Supplementary Fig. S6A). This strongly illustrated the effect of tumor-cell training on the surrounding T lymphocytes, and indicated that many T-cell clones in tumor tissues were tumor-reactive.

The correlation between tumor and adjacent nontumors is higher in luminal A/B type BC, whereas is significantly lower in basal-like type BC (Fig. 5A,B). The heterogeneity of the infiltrated TCR repertoire is shaped and modulated by the variety of neoantigens in the tumor microenvironment. Aberrantly expressed tumor genes derived by their somatic mutations contribute significantly to these tumor neoantigens. The overall mutation rate in luminal A/B BC is the lowest among all the BC subtypes, whereas the mutation rate is higher in the basal-like subtype (34). It is possible that the antigenic microenvironment is more similar between tumor and normal tissues in luminal A/B BC, due to its lower mutation rate, and in turn, it shaped the higher similarity of the TCR repertoire between tumor and nontumor tissue, compared with basal-like BC. Further studies and evidences are required to support this explanation.

The data on the TCR repertoire used by T cells that infiltrate tumors and neighboring tissues in breast cancer has implications for future diagnoses, monitoring, and predicting prognoses. The identification of tumor-reactive lymphocytes will benefit from the analysis of immune repertoires in tumor microenvironments. Tumor-reactive lymphocytes circulating in the peripheral blood
could also serve as potential noninvasive biomarkers to monitor the response of the solid tumor to treatment, track minimal residual disease, and predict relapse. We detected numerous tumor infiltrated TCR clones in the peripheral blood of the patients before and after surgery (data not shown). However, whether the dynamic change of the frequency of these TCR clones after surgery and further treatment have clinical implications still warrants prospective studies and long-term follow-up of patients. On the other hand, tumor-reactive lymphocytes as direct antitumor weapons in immunotherapy, can be tracked and monitored in the ACT process, by deep-sequencing and immune repertoire analysis. In summary, the comprehensive analysis of the characteristic TCR repertoire in breast cancer has depicted a landscape of immunosurveillance and interactions with tumor cells, as well as the relationship of TCR repertoires to various clinical features.

Acknowledgements
This work was supported by the National Natural Science Fund 81272899, 81672593, 30973463 & 81470120; Shaanxi Fund 2013K12-03-03, 2014JM4087 and Xi’an Found SF1323(3) and Discipline booster plan of Xi Jing Hospital XJZT12Z07, 2014JM4087.

References
9. Ohtani H. Focus on TILs: prognostic significance of tumor infiltrating lymphocytes in human


25. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. Clinical medicine &


Tables

Table 1. The pathological data of the discovery cohort in the study.
Patient ID | ER | HER | PR | Ki-6 | Lymph Node Ratio | Tumor subtype | Tumor infiltration (%) | Nontumor infiltration (%) |
--- | --- | --- | --- | --- | --- | --- | --- | --- |
BC0001A | 3+ | - | 2+ | 20% | 0/16 | Luminal B | NA | NA |
BC0002A | 1+ | + | - | 35% | 0/6 | Luminal B | 10 | 2 |
BC0003A | - | - | - | 30% | 1/14 | Basal-like | 20 | 0.1 |
BC0004A | 3+ | - | 3+ | 10% | 0/16 | Luminal A | 8 | 1 |
BC0005A | 3+ | + | 2+ | 5% | 1/18 | Luminal B | 0.3 | 0.1 |
BC0006A | - | - | - | 95% | 0/5 | Basal-like | 25 | 0.5 |
BC0007A | - | - | - | 20% | 0/23 | Basal-like | 2 | 1 |
BC0008A | - | + | - | 40% | 0/29 | HER2 enriched | NA | NA |
BC0009A | 1+ | - | - | 30% | 14/18 | Luminal B | 30 | 0.1 |
BC0010A | - | - | - | 50% | 15/15 | Basal-like | 40 | 0.2 |
BC0011A | 3+ | - | 2+ | 5% | 0/20 | Luminal A | 8 | 0.1 |
BC0012A | 2+ | + | 1+ | 28% | 0/16 | Luminal B | 3 | 1 |
BC0013A | - | - | - | 50% | 0/18 | Basal-like | 2 | 0.2 |
BC0014A | 3+ | - | 3+ | 28% | 0/23 | Luminal B | 0.2 | 2 |
BC0015A | - | - | - | 60% | 0/5 | Basal-like | 15 | 3 |
BC0016A | 3+ | - | 2+ | 10% | 1/10 | Luminal A | 12 | 0.1 |

NA means this sample was not analyzed.

Table 2. Public TRB CDR3 clones among Tumor tissues of all 16 patients. All these clones were filtered with a TRB database of 661 healthy individuals.
<table>
<thead>
<tr>
<th>Sequence</th>
<th>TRBV</th>
<th>TRBJ</th>
<th>BC0004A</th>
<th>BC0005A</th>
<th>BC0006A</th>
<th>BC0007A</th>
<th>BC0008A</th>
<th>BC0009A</th>
<th>BC0010A</th>
<th>BC0013A</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSASPGLAGRETQYF</td>
<td>TRB20-1</td>
<td>TRB2-5</td>
<td>0.0131</td>
<td>0.0102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSVEEAYEQYF</td>
<td>TRB29-1</td>
<td>TRB2-7</td>
<td>0.0453</td>
<td>0.0135</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSANFRERNSPLHF</td>
<td>TRB20-1</td>
<td>TRB1-6</td>
<td>0.8711</td>
<td>0.0107</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSFQGPEYQYF</td>
<td>TRB7-8</td>
<td>TRB2-7</td>
<td>0.0188</td>
<td>0.0837</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSASQTDTQYF</td>
<td>TRB20-1</td>
<td>TRB2-3</td>
<td>0.0278</td>
<td>0.0358</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSFAASKHEQYF</td>
<td>TRB12-4 TRBV12-3</td>
<td>TRB2-7</td>
<td>0.0152</td>
<td>0.0352</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSASRRDGTDTQYF</td>
<td>TRB20-1</td>
<td>TRB2-3</td>
<td>0.0171</td>
<td>0.0149</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATSLAGGETQYF</td>
<td>TRBV24/OR9-2 TRBV24-1</td>
<td>TRB2-5</td>
<td>0.0100</td>
<td>0.1671</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATSDINEQFF</td>
<td>TRBV24/OR9-2 TRBV24-1</td>
<td>TRB2-1</td>
<td>0.0123</td>
<td>0.0211</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAWSPSGLPQYF</td>
<td>TRBV30</td>
<td>TRB2-3</td>
<td>0.4246</td>
<td>0.0113</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATSRENNEQFF</td>
<td>TRBV15</td>
<td>TRB2-1</td>
<td>0.0113</td>
<td>0.0115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSAPGQGSGNTIYF</td>
<td>TRB20-1</td>
<td>TRB1-3</td>
<td>0.0108</td>
<td>0.0109</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSLVRGRDYGYTF</td>
<td>TRBV7-2</td>
<td>TRB1-2</td>
<td>0.0141</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSATGTGANSPLHF</td>
<td>TRBV5-1</td>
<td>TRB1-2</td>
<td>0.0177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSATGTGANSPLHF</td>
<td>TRBV5-1</td>
<td>TRB1-6</td>
<td>0.0147</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASSATGTGANSPLHF</td>
<td>TRBV2</td>
<td>TRB1-6</td>
<td>0.1580</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASTTNTDTQYF</td>
<td>TRBV12-4 TRBV12-3</td>
<td>TRB2-3</td>
<td>0.0182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASTTNTDTQYF</td>
<td>TRBV19</td>
<td>TRB2-3</td>
<td>0.0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure legends**

Fig. 1. Proportion and diversity of infiltrated T lymphocytes in breast tumor and other tissues. (A) Comparison of infiltration proportion of T lymphocytes in tumor and nontumor breast tissue of 14 patient samples analyzed by IHC (P-value=0.0023). (B) Shannon diversity index of LN, nontumor and tumor tissue.

Fig. 2. Shared clones between different tissues. The percentage was calculated by dividing the total clones of each tissue by the shared clones. (A) The proportion of shared clones between LN and nontumor tissue. (B) The proportion of shared clones between LN and tumor tissue.

Fig. 3. Percentage of T-cell clones of tumor and LN tissue in different degree of expansion for each patient. Colored bars represent the percentage of T-cell clones in each proportion. (Expanded clone: > 0.1%; Medium clone: 0.01%–0.1%; Small clone: < 0.01%) (A) Total T-cell clones of LN tissue. (B) Total T-cell clones of tumor tissue. (C) T-cell clones of LN shared with tumor tissue. (D) T-cell clones of tumor tissue shared with LN tissue.

# Node-positive patients with high LNR (> 70%). The patients were ordered in sequence of their percentage of expanded clones for the total T cells in the LNs, from the smallest to the largest.

Fig. 4. Correlation of node positivity with T lymphocyte ratio in tumor and expansion in LN. (A) Comparison of infiltration proportion of T lymphocytes in tumor tissue of LN-positive patients and LN-negative patients (P = 0.1416). (B) Comparison of infiltration proportion of T lymphocytes in tumor tissue of patients with high LNR and LN-negative patients (P = 0.0022). (C) Comparison of T-cell expansion in LN for node-negative, node-positive and high LNR patients.
(*P < 0.05, ** P < 0.01, *** P < 0.001 according to Mann-Whitney U test)

LNR: lymph node ratio (LNR)

Fig. 5. The Pearson correlation coefficients between tumor and nontumor in luminal-like and basal-like BC subtypes (A) The comparison in the main cohort. (B) The comparison combining the main cohort with the validation cohort. (*P < 0.05 according to Mann-Whitney U test).
Figure 1

(A) T lymphocyte percentage (%)

(B) Shannon’s diversity index

Legend:
- **: p < 0.01
- ***: p < 0.001
- ****: p < 0.0001
Figure 3

A

% of Total clones

B

% of Migrated clones

C

% of Total clones

D

% of Migrated clones

###

Expanded

Medium

Small

Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.
The different T-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues

Ting Wang, Changxi Wang, Jinghua Wu, et al.

Cancer Immunol Res  Published OnlineFirst December 30, 2016.

Updated version

Access the most recent version of this article at:
doi:10.1158/2326-6066.CIR-16-0107

Supplementary Material

Access the most recent supplemental material at:
http://cancerimmunolres.aacrjournals.org/content/suppl/2016/12/30/2326-6066.CIR-16-0107.DC1.html

Author Manuscript

Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts

Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions

To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions

To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.