Investigation of T cell receptor and Immunoglobulin repertoire through next generation sequencing data

Yaxuan Yu

A thesis submitted to

The School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Science

Supervised by

Prof. Cathal Seoighe

&

Prof. Rhodri Ceredig

March 2017

Abstract

The diversity of the immunological repertoire has long been a subject of research focus, providing important insights into the adaptive immune system. Rapid developments in next generation sequencing technologies have revolutionized the way immunological repertoires are analyzed, providing unprecedented high-resolution data. Nonetheless, these high-throughput approaches also present unique computational challenges that must be addressed through the development of accurate and efficient bioinformatics pipelines. In this thesis, we demonstrated a complete bioinformatics workflow for processing and analysis of high-throughput sequences from immune receptors, and applied these tools to explore research questions relating to the diversity of immune receptor genes in human populations.

An aspect of the immunological repertoire that is frequently of immediate interest to immunologists is the distribution of different immune receptor clonotypes among individuals, as knowledge of this could lead to a better understanding of the dynamics of the immune system in different conditions. We first implemented a bioinformatics pipeline to analyze next generation sequencing data from T cell receptors and immunoglobulins. This pipeline featured an ultra-fast and accurate fast-tag-searching algorithm for VDJ alignments, which outperformed all the other similar pipelines on benchmarking. In addition to that, this pipeline included two novel functional components. The first function was polymorphism analysis, which reports putative novel SNPs found in the input sequences. The second novel function was the ability to construct lineage mutation trees to describe the affinity maturation process of immunoglobulins.

No matter how sophisticated the alignment algorithms are, accurate gene alignment always requires the right reference database. Unfortunately, the IMGT database, which is the most widely used reference database in immunological repertoire analysis pipelines, has been shown to be incomplete and to contain numerous errors. Thus, the second task undertaken in this PhD

thesis was to create a more comprehensive reference database for T cell receptors and immunoglobulin genes by exploiting the large volume of publicly available human genome resequencing data generated in recent years. Based on the variant calling information retrieved from the 1000 Genomes Project and the current human reference genomes, we were able to infer a set of putative alleles of immune receptor genes. Lym1k, our database of these inferred alleles, provided a more comprehensive collection of immune receptor alleles found in global human populations, as evidenced by a significantly improved alignment performance on real datasets compared to IMGT.

The immune receptor loci are among the most dynamic regions of the human genome, with a high rate of structural variation, as well as high allelic diversity. Previous analyses of the allelic diversity of immune receptor genes in global human populations were constrained by the limited size of human genome resequencing data. We focused on addressing three research questions relating to the allelic diversity of immune receptor genes in our last research chapter. Firstly, it has been shown by many studies that African populations have greater overall allelic richness than other human populations, we thus compared the allelic diversity between African and Non-African populations for immune receptor genes. Not surprisingly, the immune receptor alleles in African populations were more diversified compared to Non-African populations. As the immune receptor genes with the same gene type are located adjacent to each other on the chromosome, we secondly investigated if genomic location was associated with allelic diversity, potentially reflecting differences in the frequency of receptor gene use between genes located towards the proximal or distal ends of the arrays of genes of a given type. However, we did not find an effect of position on allelic diversity. Lastly, we hypothesized that immune receptor genes that are more frequently selected during rearrangement are under higher diversifying selection pressure, and this would lead to a higher allelic diversity. Surprisingly, the correlation was absent from most of the gene types except for weak positive correlations in TCRA genes.

In conclusion, this thesis demonstrated several novel high-throughput approaches and strategies for immunological repertoire analysis. It also addressed some important biological questions relating to the allelic diversity of immune receptor genes by exploiting public biological resources, which could potentially inform subsequent studies.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisors Prof. Cathal Seoighe and Prof. Rhodri Ceredig for guiding me through one of my most unforgettable adventures in science. They have been tremendous mentors for me by providing their insightful suggestions and enormous patience during my PhD. I have been absolutely privileged and lucky to have had the chance to work with these two fantastic scientists.

Second, I would like to thank the members from my graduate research committee, Dr. John Newell, Dr. Haixuan Yang and Dr. Tim Downing for their support and advices during my PhD.

I would like to acknowledge the fellow bioinformatics PhD students in the School of Mathematics in NUI Galway, and other fellows from the QUANTI network for exchanging opinions and experiences in research. I appreciate that each time when I got stuck at some research problems, they have always been able to provide their insights from different angles. They also made my time in Galway very special.

Last but not the least, my special thanks goes to my family. I am deeply grateful for their encouragements during my PhD. Each time I have felt lost or confused about life, I have known they are always there watching my back, giving me strength to carry on. It would have been impossible to finish this thesis without their support, and I am proud to share the honor with them.

Declaration

I hereby declare that this thesis which I now submit for assessment as partial fulfillment of the requirements for the award of Doctor of Philosophy in Science is entirely my own work and had not been taken from the work of others; save and to the extent that such work has been cited and acknowledged in the text. I have not obtained a degree in this university, or elsewhere, on the basis of this work.

Signed: 

Date: 23/07/2017

Contents

List of figures

List of tables

Chapter 1 – Introduction

A range of highly diversified immune receptors trigger the responses of the adaptive immune system to pathogens. In this chapter, we review the basic cellular mechanisms involved in generating diverse immunological repertoires as well as the high-throughput techniques that have recently been developed for immunological repertoire analysis. In brief, in chapter 1.1, we discuss the development of T and B cells, which are characterized by highly diversified cell-surface receptors, generated through a unique somatic gene rearrangement mechanism. In addition, we review the public biological resources generated in recent years that can be used to generate insights into immune repertoires, as well as the unique computational challenges these data present.

1.1 Origins and importance of T and B cell receptors

1.1.1 An overview of the evolution of the adaptive immune system

The vertebrate adaptive immune system, together with the innate immune system, provides protection against potential pathogens. The capacity of the adaptive immune system to recognize the broad range of pathogens is largely due to the extreme diversity of lymphocyte receptors (T cell receptors and immunoglobulins). Fundamentally, what distinguishes the adaptive immune system from the innate immune system is its high specificity towards certain pathogens during the immune response. Moreover, after successfully recognizing a specific pathogen, the adaptive immune system forms an immunological memory towards that pathogen, and it will initialize an enhanced response to this pathogen in a subsequent encounter. For instance, vaccination, exploits this property of the adaptive immune system [1].

Adaptive immunity in vertebrates arose rather abruptly in evolution [2]. The most evolutionarily ancient adaptive immune systems, dating to 500 million years before the present, are found in jawed fish (cartilaginous fish), which have organized lymphoid tissues, T cell receptors and immunoglobulins [3, 4]. Although adaptive immunity is often regarded as a unique mechanism in jawed vertebrates, in which specificity is achieved through somatic gene rearrangement strategies (VDJ recombination), studies also suggest that jawless vertebrates evolved alternative forms of "adaptive immunity" that makes use of different genetic strategies [5-7]. Furthermore, scientists have found that TCRs from mice continue to function well even after being substituted by TCRs from sharks, frogs or trout [8]. This shows that many vertebrate species even though they share common ancestors that can date back to 500 million years ago, still have similar TCRs, suggesting conservation of the TCR-MHC interaction in the adaptive immune system [9, 10]. In addition to that, Immunoglobulin M (IgM), as one of most ancient antibodies, is first found in cartilaginous fish and shared by all the jawed vertebrates with the same immunological functions [11].

1.1.2 The development of T and B cells

The major cellular components of the adaptive immune system are T and B cells. B cells were discovered and characterized in the mid 1960s to 1970s using animal models and clinical information from patients with immune deficiency conditions [12-14]. B cells (B lymphocytes) are involved in the humoral immune response (the immunity mediated by substances from humors or body fluids, such as secreted antibodies and complement proteins) of adaptive immunity [15]. The common ancestors of all blood cells are the hematopoietic stem cells [16], which are multipotent and self-renewing stem cells that reside in the bone marrow. These cells first differentiate into several multipotent progenitor cells, which are stem cells that can produce lymphoid and myeloid cells but are not self-renewable. Subsequently, multipotent progenitor cells differentiate into common lymphoid progenitors [17].

1.1.2.1 B cell maturation

Starting from the lymphoid progenitors, there are several stages involved in the maturation of B cells or the definitive B-cell fate (figure 1.2.1). First, common progenitor cells give rise to the progenitor B cells (pro-B cells), which can be specified by the induction of B-lineage-specific transcription factor E2A [18]. This stage, also known as early pro-B cell maturation is characterized by the rearrangement of D-J genes on the immunoglobulin heavy chain locus. Subsequently, The rearrangements of V-DJ genes take place and this stage is known as the pro-B cell stage. The sign of successful completion of the pro-B cell stage and entering the pre-B cell stage is the expression of a functional heavy chain. Then, during the pre-B cell maturation stage, the cell begins to rearrange V-J genes on the immunoglobulin light chain. The cellular mechanisms involved in the VDJ rearrangements will be discussed in greater detail in the following chapter.

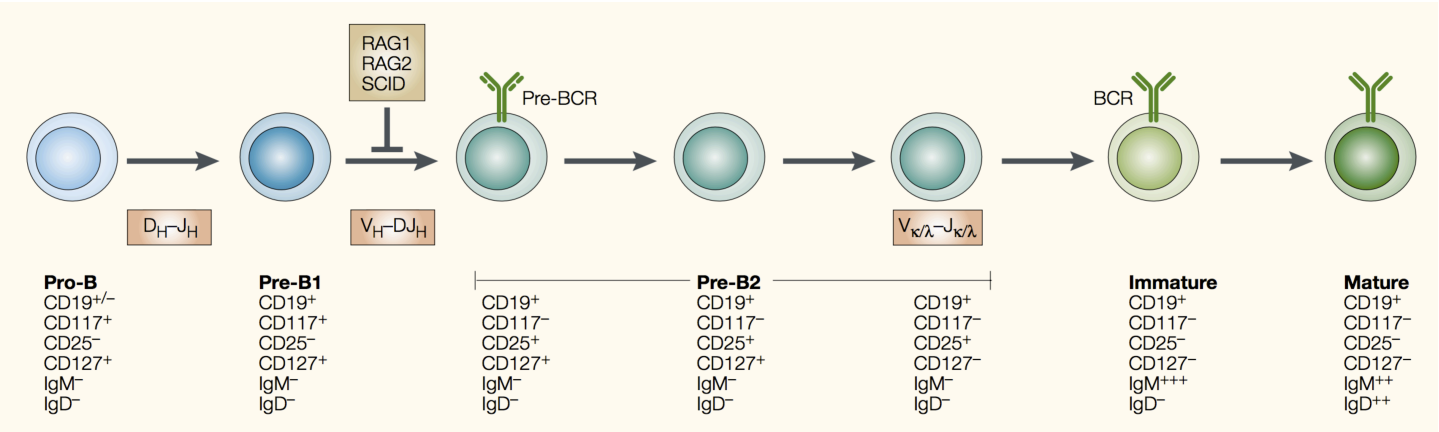


Figure 1.1.1 step-wise B cell development. B cells mature from left to right in this schema. D_H-J_H represents the recombination of heavy chain D gene with heavy chain J gene. V_H-DJ_H refers to the recombination of V gene with the D-J recombination on heavy chain. $V_{\kappa/\lambda}$ - $J_{\kappa/\lambda}$ represents the gene rearrangements of the light chain. Reproduced from [24] with permission.

Notably, The chance of successful gene rearrangements on heavy and light chains is about 1/3 [19, 20], thus many pre and pro B cells may undergo

apoptosis if the rearrangements are nonproductive. After successfully assembling the immunoglobulin light chain, the pre-B cell turns into an immature B cell with the expression of complete immunoglobulin M (IgM) molecule (an immunoglobulin isotype) on the cell surface. After that, it becomes mature B cells with the additional occurrence of IgD molecules through alternative splicing [21-23].

There is a very important mechanism of B cell tolerance called clonal deletion when the immature B cells are formed [25, 26]. The immature B cells undergo negative selection when the IgM molecules attached on their cell surface recognize self-antigens (these B cells are known as auto-reactive B cells) [27]. There may be two types of self-antigens involved in the negative selection: The first type involves the ubiquitous self-cell-surface (multivalent) antigens such as major histocompatibility complex (MHC). It is believed that auto-reactive B cells will subsequently undergo apoptosis. However, studies have suggested that not all B cells with strong auto-reactive receptors are deleted, instead, there is an interval before cell death happens [28-32]. The auto-reactive B cells may be rescued by going through further gene rearrangements, termed as receptor editing [33-37]. The second type of self-antigen is the soluble self-antigen (low valence). Under this circumstance, the premature B cells do not die. Nonetheless, there is a down-regulation of receptor synthesis, which results in the loss of the ability to express IgM on their cell surfaces. These B cells therefore only express IgD and lose their ability to recognize antigens; they are, therefore, called anergic B cells [38, 39]. There are also possibilities that some auto-reactive B cells simply “escaped” the negative selection and they are in a state of immunological ignorance [40]. These B cells retain a certain level of affinity towards self-antigens, but do not respond to them for some reason that is still largely unknown [30].

Overall, B cell negative selection or central tolerance is not perfect. This is a double-edged sword. The relatively loose negative selection process results in a more diversified receptor repertoire, which can recognize a wider range of pathogens. This arises, however, at the price of escaped auto-reactive B cells that can be reactivated under certain conditions, such as an increase in the

concentration of certain self-antigens, which results in autoimmune responses [41-43].

1.1.2.2 The development of T cells

T cells represent another important component in the adaptive immune system with their crucial role in cell-mediated immunity. Like the B cells, the journey of T cell development also starts from the major hematopoietic stem cells. In addition to that, the development of T cells shares a lot of similarities with the B cells, such as step-wise rearrangement of the VDJ genes, negative selections to eliminate self-reactive antigens and the assembly of heterodimeric antigen receptors. Nevertheless, T cell maturation involves their own characteristics and some unique mechanisms. First, although both progenitor T and B cells are initially derived from hematopoietic stem cells in bone marrow, progenitor T cells migrate to the thymus, where all the important events of their development occur [44-46]; this is the origin of the term T cell (Thymus dependent lymphocyte). When progenitor T cells first enter into the thymus, no gene arrangements have happened and they lack cell-surface molecules (CD4, CD8) [47] that are the main characteristics of mature T cells. These T cells are classified as “double negative” thymocytes (CD4-, CD8-). One fundamental difference comparing to B cell development is that there are two distinct lineages of T cell maturations starting from here [44]: the majority of the double negative thymocytes give rise to the α:β T cells with both CD4 and CD8 expressed on their cell surfaces[48, 49]; the minority population form the γ:δ lineage, comprising T cells that lack CD4 and CD8 expression [50, 51].

The double negative stage of T cell development can be subdivided into four stages (DN1, DN2, DN3, DN4) based on the expression of three types of cell-surface molecules: CD25, CD44 and Kit [52-54]. Initially DN1 cells have the germline configuration of genes encoding for both chains of the T cell receptors and only express CD44 on their cell surfaces. The gene rearrangements of D to J genes first happen on the β chain of the thymocytes during DN2. The sign of a thymocyte that successfully finishes DN2 is the

expression of both CD44 and CD25 on their cell surface [55-57]. After that, they enter into DN3 and the expression of CD44 and Kit are reduced. Meanwhile, D to J rearrangements continues in some thymocytes and other thymocytes that have already finished the D-J arrangements begin to undergo the V-DJ rearrangements. DN3 thymocytes that fail to form functional β chain will not progress into DN4 and soon die in the thymus. On the contrary, the thymocytes that successfully form a productive β chain and lose the expression of CD25 enter into DN4, where they begin to proliferate.

When the thymocytes express both CD4 and CD8 on their cell surfaces, they are also known as the double positive T cells (CD4+, CD8+). The majority of T cells are double positive thymocytes. In the double positive stage, the genes encoding the T cell receptor α chain begin to rearrange. There are two main regions in the thymus: the peripheral cortex and the central medulla [58]. The peripheral cortex can be further subdivided into outer cortex and inner cortex. The development of the thymocytes until the double positive stage, as we described above all happens in the outer cortex. The double positive thymocytes now migrate into the inner cortex and undergo the process of positive and negative selections (figure 1.1.2) [59, 60].

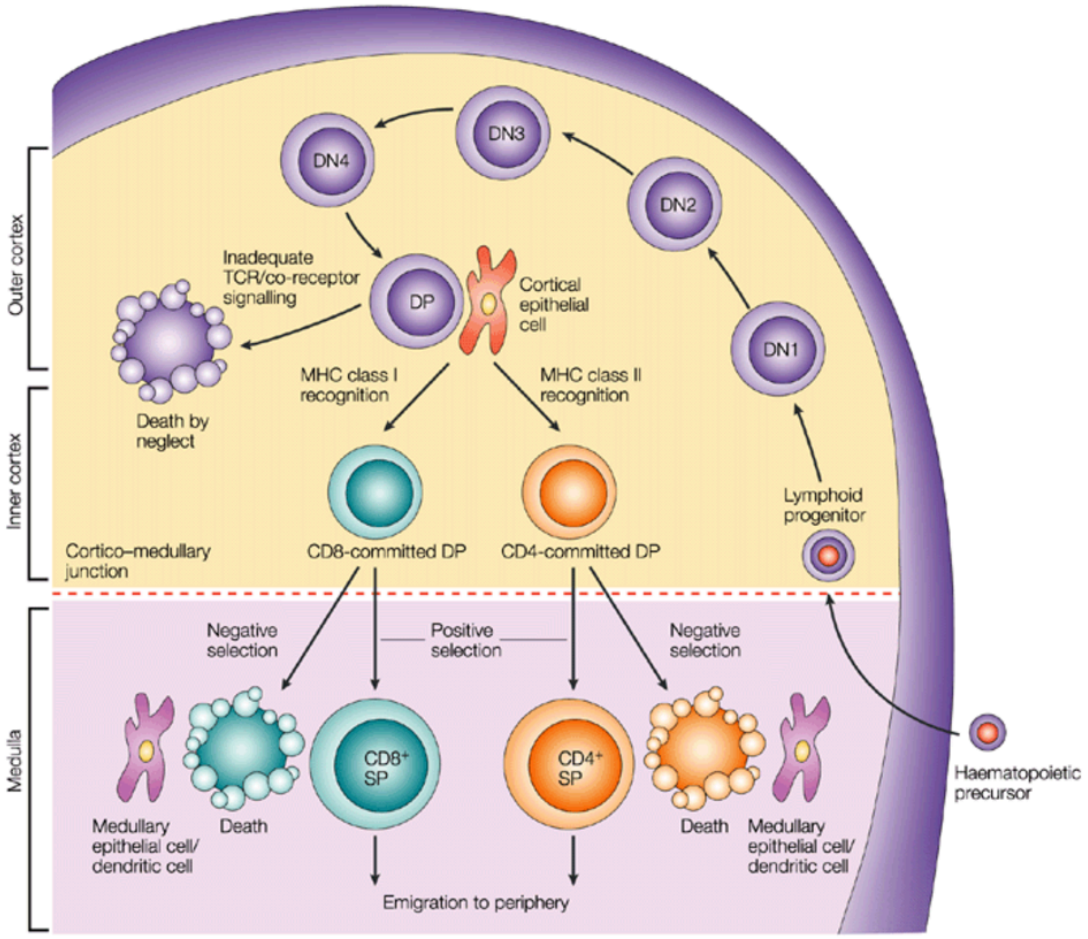


Figure 1.1.2 Maturation and migration of thymocytes in the thymus. The thymus is subdivided into cortex (outer and inner cortex) and medulla from top to bottom in the figure. Different stages of T cell maturation happen in different areas of the thymus, illustrated by arrows with time course. Sourced from [44] with permission.

The ultimate objective of positive selection is to ensure T cells capable of interacting with MHCs for serving useful immune functions. During positive selection, the double positive thymocytes are presented with self-antigens expressed by the cortical epithelial cells with their MHC molecules attached on their surface. Only those thymocytes that can interact with the MHC-1 or MHC-2 (two subtypes of MHC) molecules with moderate affinity will pass through positive selection and receive a "survival signal"; the remainder fail the selection and undergo apoptosis [61]. Note that the fate of T cells is determined early at this point: the thymocytes that can interact well with MHC-1 molecules will become CD4+ T cells, while other thymocytes that can

interact with MHC-2 molecule further progress to CD8+ T cells [62]. Finally, thymocytes that survive positive selection migrate to the boundary between medulla and inner cortex to undergo negative selection. They are presented with self-antigens again on medullary thymic epithelial cells with their MHCs. Similar to the negative selection in B cell development the thymocytes that strongly interact with the self-antigens fail the negative selection and are eliminated. The remaining thymocytes that pass negative selection become naïve T cells and exit the thymus to circulate to the lymphoid organs. It is estimated that 98% of thymocytes die during their maturation in the thymus, as a result of either failing negative or positive selection [63,64].

In summary, both T and B cells share the ability to recognize non-self-antigens of invading pathogens but differ on their specialized immunological mechanisms and functions. They share a very similar step-wise gene rearrangement process forming the double chain structures of cell-surface receptors (T cell receptor and immunoglobulin) during their maturation. They both have central tolerance mechanisms involved in their maturation by which the T and B cells undergo negative selections to eliminate the cells that are strongly reactive to self-peptides. Nevertheless, compared to B cells, T cells undergo more rigorous negative selection processes, as well as an additional positive selection process to ensure that T cells are capable of interacting with self-antigens restricted by MHC molecules during their development.

1.1.3 The generation of the T and B cell receptor diversity

Both T and B cells have antigen receptors attached on their cell surfaces, namely, T cell receptors (TCR) for T cells and immunoglobulins for B cells. Each T and B cell contains numerous copies of a single lymphocyte receptor, which can recognize antigens from pathogens in their environments with their unique antigen binding sites [65]. Because the immune system consists of billions of T and B cells, collectively with a large diversity of the lymphocyte receptor repertoire, these T and B cells enable extensive adaptive immune responses to a broad range of foreign antigens [66]. The ability of lymphocyte

receptors to target a wide range of antigens results from the variability of the amino acid sequences of antigen-binding sites. Not surprisingly, given such large diversity of TCR/IG repertoire (estimated to be 10^{12} different TCRs and immunoglobulins in human) [67], complex and elegant genetic mechanisms during their maturation are involved. In this chapter we will review the genetic mechanisms involved in the generation of the diversity of TCR and immunoglobulin repertoire.

1.1.3.1 Diversification of immunoglobulins

Immunoglobulin, also known as antibody or B cell receptor, is a large "Y" shaped protein attached on the surface of the B cell (figure 1.1.3) [68]. Each arm of the "Y" shaped protein represents a fragment antigen-binding (Fab) component, which consists of a constant domain (C region) and a variable domain (V region). Although the general structure of different immunoglobulins is similar, each of the unique immunoglobulins is characterized by a hyper-variable region within the variable domain. The variable domain is a double chain structure, which includes a heavy chain (IGH) and a light chain (IGL). The hyper-variable region, also known as the antigen binding sites or paratopes is where the antigen to receptor binding happens. Antigen binding sites mainly include three regions named as Complementary Determining Region one (CDR1), two (CDR2), and three (CDR3) [69, 70].

Different immunoglobulin clonotypes present distinctive immune responses to different antigens, depending on their binding affinities. Virtually all microbes can be the targets of immunoglobulins, based on the extensive diversity of the immunoglobulin repertoire. One early hypothesis for the diversification of immunoglobulins is that there are separate germline immunoglobulin genes located on the chromosome(s), which are translated into different immunoglobulin proteins [71]. However, although there are several immunoglobulin gene groups consisting of multiple gene segments located on the chromosome [72], the number of genes included in these gene groups is insufficient to make up such large diversity of the repertoire [66, 67].

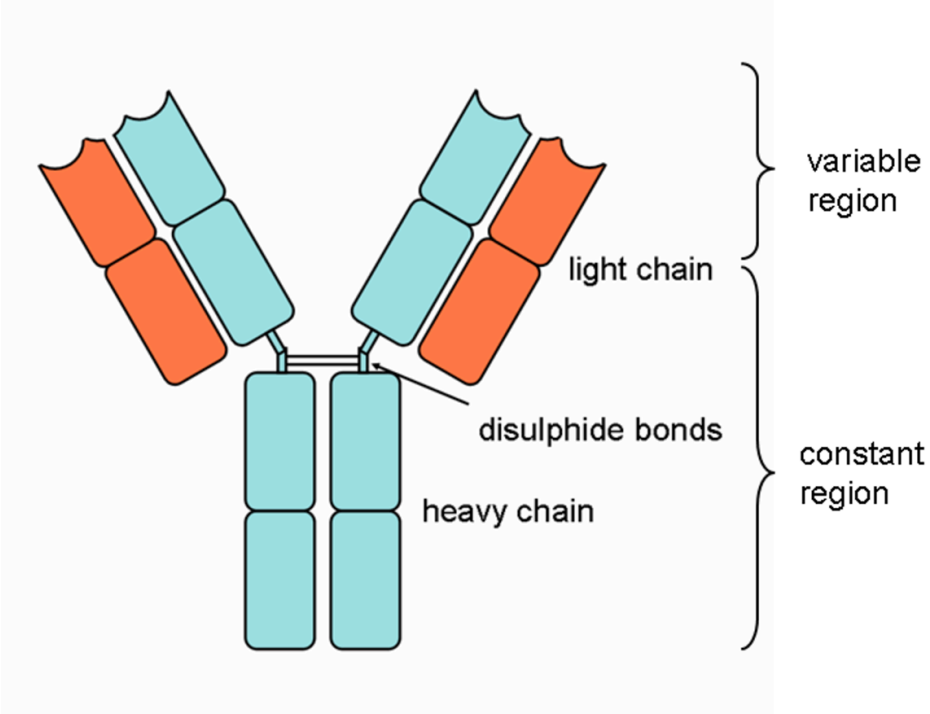


Figure 1.1.3 An illustration of immunoglobulin structure. Sourced from [73] with permission.

Further studies of cloned immunoglobulin genes have shown that there are two genetic mechanisms involved in the diversification of the immunoglobulin repertoire [74, 75]. The first mechanism, called V (D) J recombination, also known as somatic recombination or rearrangement, involves a largely random process of gene rearrangements, generating unique immunoglobulin variable regions. The variable region of the immunoglobulins is encoded by several gene segments (subgenes). Specifically, IGH is encoded by three types of gene segments, namely variable (IGHV), diversity (IGHD) and joint (IGHJ) gene segments [76], while IGL only contains V (IGLV) and J (IGLJ) gene segments [77]. Gene segments within the same chain type are located adjacent to each other on the chromosome. The number of different gene segments varies between different gene types. For instance, there are 54 functional IGHV gene segments and only thirteen functional IGHJ gene segments in human reported by IMGT database [78].

Here we use IGH as an example to illustrate this recombination process in greater detail since it includes one more gene type (IGHD) than IGL and,

therefore, one more joining event (figure 1.1.4) [79]. In brief, VDJ recombination in IGH can be divided into two steps with similar genetic mechanisms. Firstly, D-J recombination occurs between one D and one J gene segment with deletion of all the DNA between the selected D and J gene segment on the heavy chain locus, forming a recombined D-J segment. Secondly, one V gene segment is selected and joined with the D-J recombination from the first step by removing the entire DNA between them. The recombination is guided by the recombination signal sequences (RSSs), which are conserved non-coding DNA sequences near the point where the recombination takes place [80, 81]. During the recombination of D-J, the recombinase encoded by lymphocyte-specific recombination activating genes (RAG-1 and RAG-2) target one RSS that is adjacent to one particular D gene segment and one RSS that is next to one particular J gene segment forming a "hairpin" structure of RAG complexes, and subsequently initiating DNA cleavages on both targets [82, 83]. The cleavages result in a piece of DNA sequence that contains all the D and J gene segments between the two cutoff-targeted genomic position, flowing away from the genome, and this DNA sequence will be further joined by DNA ligase forming a circular signal joint. On the other side, the remaining two broken DNA ends are firstly repaired by several proteins such as KU70, KU80 and Artemis [84, 85]. Furthermore, the cut ends are modified by terminal deoxynucleotidyl transferase (TdT), with non-templated additions of nucleotides to the junction between the D and J gene segment [86]. The additions are mostly stochastic but TdT does show a G/C preference [87]. In addition to that, random deletions are introduced on the new joint with the help of exonucleases. All of these processes result in a hyper-variable antigen-binding site, even for a small group of germline gene segments.

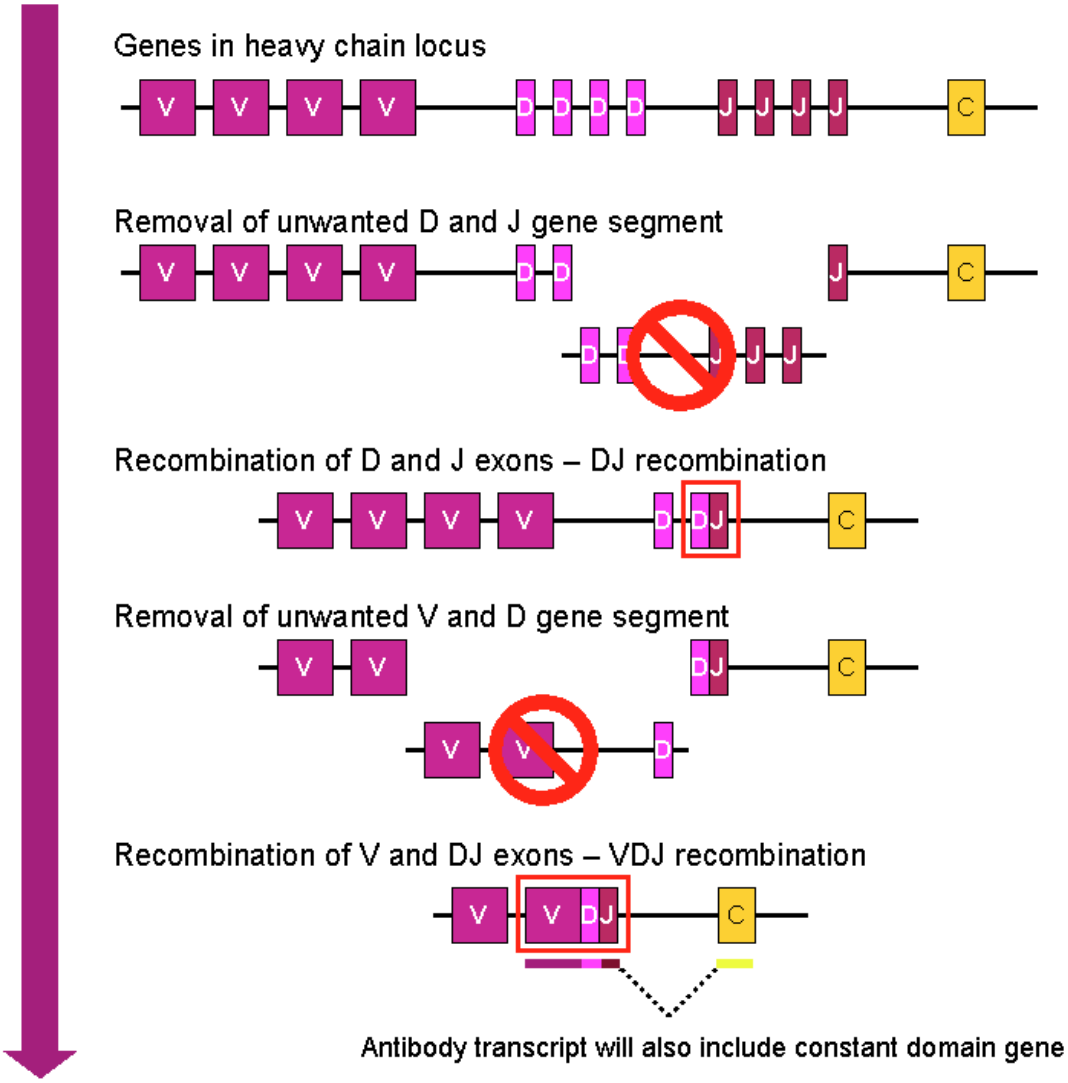


Figure 1.1.4 Illustration of VDJ recombination in immunoglobulins. Different gene segments are labeled in different colors. Sequential events in the somatic recombination are illustrated by the arrow from top to bottom. Sourced from [88] with permission.

Besides gene rearrangements, somatic hypermutation is another important cellular mechanism that further expands the diversity of the immunoglobulin repertoire. Somatic hypermutation is a major component of affinity maturation, which is the ability of B cells to produce immunoglobulins with enhanced affinity towards foreign antigens during an immune response [89]. Somatic hypermutation occurs in activated B cells (i.e. cells that recognize foreign antigens and start to proliferate) in which V (D) J recombination has finished and the immunoglobulin genes have been transcribed. Somatic

hypermutation leads to at least 10^5-10^6 fold greater mutation rates on the immunoglobulin locus than the normal mutation rates across the genome [90]. The mutations mainly consist of single base substitutions, with few insertions and deletions. Notably, the mutations mainly take place on the three complementary determining regions [75].

1.1.3.2 Diversification of T cell receptors

The T cell receptor is a molecule found on the surface of T cells, responsible for recognizing antigens presented by major histocompatibility complex molecules (MHC) on antigen-presenting cells (APCs) [91]. There are many similarities between the structure of immunoglobulins and TCRs. First, each of the T cell receptors (TCRs) also consists of two chains. The majority of the T cells carrying TCRs consist of an α chain (TCRA) and a β chain (TCRB) and these T cells are thus referred as α:β T cells [8]. The remaining 5% of T cells are γ:δ T cells, which contain TCRs comprised of γ (TCRG) and δ (TCRD) chains [50]. The TCRA and TCRG loci include V and J gene segments, similar to IGL locus, while TCRB and TCRD include V, D and J gene segments, like that for IGH. The diversity of the TCR repertoire is mainly generated by the process of V (D) J gene rearrangements during the development of T cells.

As mentioned above, the maturation of T cells occur in the thymus where V (D) J recombination takes place [44-46]. The recombination process of TCR genes is essentially the same as that of immunoglobulin genes. Briefly, for TCRB, D to J recombination first occurs by using RAG1 and RAG2 to target the RSSs, remove the unselected gene segments and form a recombined D-J segment, followed by V to D-J recombination with the same cellular process. There are also random insertions and deletions added to the V-(D)-J junctions during the rearrangements, further diversifying the TCR repertoire.

Although TCR and immunoglobulin genes share the same somatic recombination process, there are a few differences between the mechanisms involved for diversifying TCR and immunoglobulin repertoires [84]. The major difference is that TCR does not undergo somatic hypermutation for

further diversification of the sequences after the V (D) J recombination. In addition to that, TCR diversity mainly concentrates in the CDR3 region. Similar to the immunoglobulin, each of the TCR chains also includes three complementary determining (CDR) regions where antigen bindings take place. Within the variable domain of each TCR chain, the CDR1 and CDR2 are found in the V region, whereas CDR3 includes part of the V, all the D (only for TCRB) and part of the J gene segment. Since there are no somatic hypermutations in TCRs, which can add diversities to all the three complimentary determining regions as in immunoglobulins, the most variable region in TCR is the CDR3 region that includes the most gene segments as well as the junctional diversities [92, 93].

In summary, the adaptive immune system, which is characterized by its remarkably diversified lymphocyte repertoires, initiates specific immune responses to a broad range of pathogens. The development of both T and B lymphocytes starts from the uncommitted hematopoietic stem cells. Mammalian B lymphocyte development mainly takes place in the bone marrow; although T lymphocytes also originate from the bone marrow, most of their development activities are in the thymus. During the maturation of B and T lymphocytes, they undergo a similar gene rearrangement process for their cell-surface receptors - T cell receptor for T cells and immunoglobulins for B cells, involving the use of RAG1/RAG2 and several DNA-repair enzymes to achieve the large diversity of their receptor repertoire. Once their rearrangements have finished, T and B cells need to survive their corresponding selection processes tested in two ways. T cells undergo positive selections to ensure moderate binding affinity towards antigen-MHC complex. In addition to that, negative selection occurs throughout the whole development stages of T and B cells, eliminating the auto-reactive lymphocytes that can bind with self-antigens with strong affinity. The complex and precise genetic mechanisms involved in the diversifications of immune repertoires ensured the striking ability of adaptive immune system to response to a wide range of the foreign invaders.

1.2 Next generation sequencing of immunological repertoires

1.2.1 An overview of the evolution of DNA sequencing technologies

DNA sequencing technology refers to the methods of determining the order of the nucleotides in DNA molecules. In the early 1970s, Ray Wu and colleagues at Cornell University carried out the first approach of determining DNA sequences by using the location-specific primer extension strategy [94, 95]. Based on this strategy, Frederick Sanger and colleagues developed much faster DNA sequencing methods, later called Sanger sequencing (also known as chain-termination sequencing) [96], and it is one of the most widely used DNA sequencing technologies in the last century.

1.2.1.1 Sanger sequencing technology

Sanger sequencing can be divided into four steps (figure 1.2.1). First, the double stranded DNA sequence is denatured into single strand using heat. Second, the location-specific primer, which is designed to be complementary to the region of the DNA sequence that is next to the region of interest, is annealed to the template DNA sequence. Next the solution is divided into four separate reactions, each includes one type of the dideoxynucleotides (ddATP,ddGTP,ddCTP, ddTTP), all four of the deoxynucleotides (dATP, dGTP, dCTP, dTTP) and DNA polymerase. The DNA sequence is then synthetized with the nucleotides added to the growing chain by the DNA polymerase. Last, the DNA is again denatured and separated by size using gel electrophoresis. The gel is then exposed to UV light or X-ray and the DNA sequence can be directly read from the gel image or the X-ray film. However, the major shortcomings in Sanger sequencing is the limit of the sequence length: if the length of the DNA being sequenced is more than 1000 base pairs, the result

will be vastly inaccurate. Therefore it is very ineffective in studies involving sequencing of large genomes such as human genome, which is almost 3 billion base pair long [97].

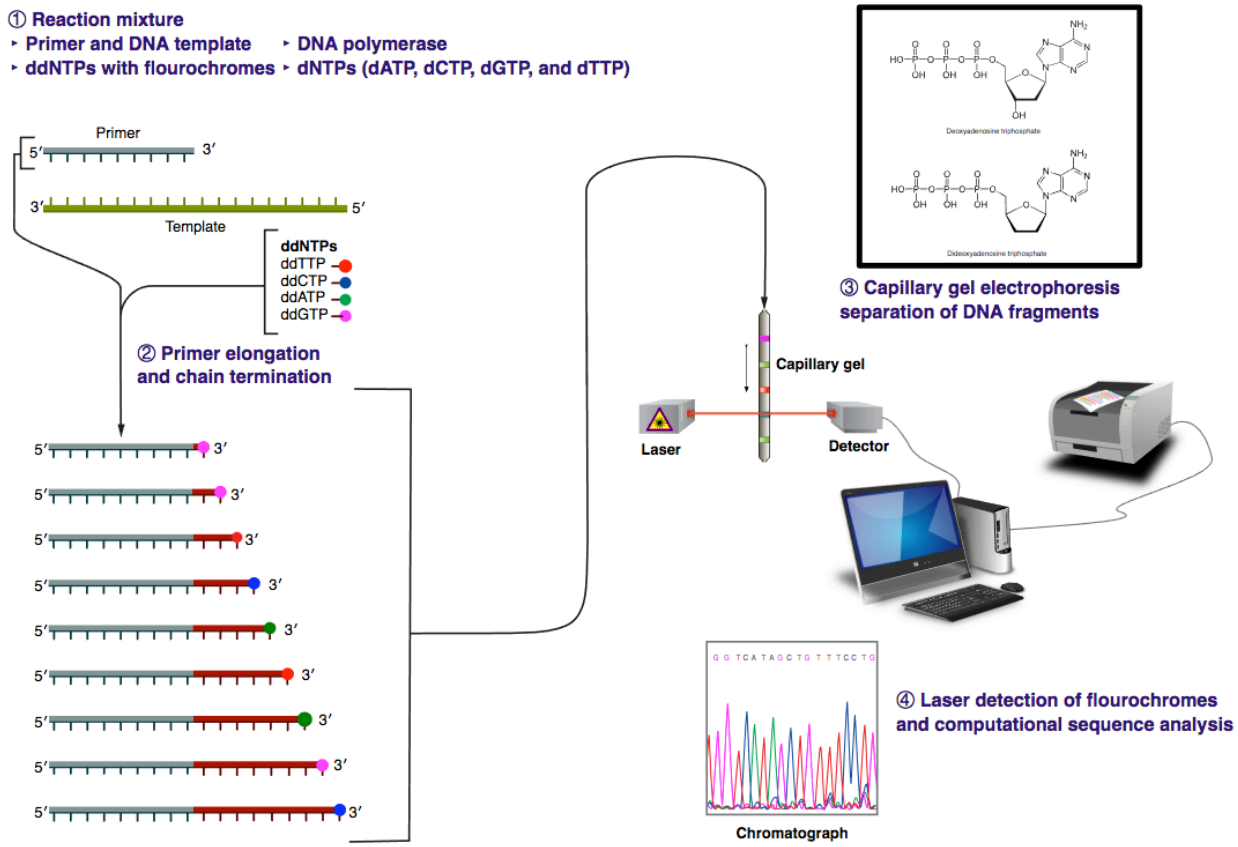


Figure 1.2.1 A stepwise illustration of Sanger sequencing experiments. Sourced from [98] with permission.

1.2.1.2 High-throughput sequencing technologies

In the mid to late 1990s, several new DNA sequencing technologies were developed aiming to improve the accuracy and effectiveness of DNA sequencing. Together they were so called Next Generation Sequencing or high-throughput sequencing technologies. In 1996, Pål Nyrén and colleagues published their method of pyrosequencing [99], which differs from Sanger sequencing by relying on "sequencing by synthesis" principle. In 1997, Pascal Mayer and Laurent Farinelli submitted a patent about their new sequencing method, which mainly exploited random surface-PCR arraying methods [100].

The first commercialized NGS method is from Lynx Therapeutics, which applied Massively paralleled signature sequencing in 2000 [101]. A major improvement of the efficiency of DNA sequencing happened in 2004, where 454 Life Sciences marketed their parallelized version of pyrosequencing [102]; this approach reduced 6-fold the sequencing costs comparing to Sanger sequencing. Since then, with the fast growing needs of scientists in DNA sequencing, several companies have developed NGS methods with further improvements in sequencing efficiency, resulting in significant reduction of sequencing costs (figure 1.2.2) [103-107].

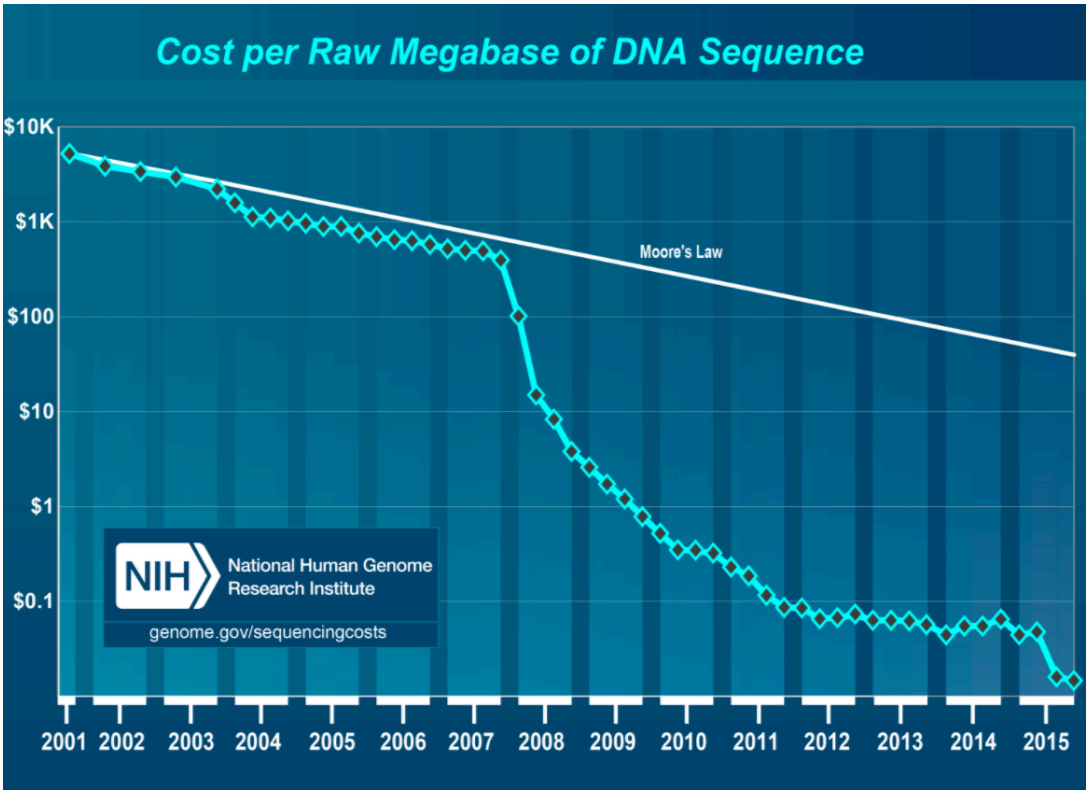


Figure 1.2.2 The change of the cost of DNA sequencing from 2001 to 2015 by using Sanger and Next Generation sequencing. The data from 2001 to 2007 represents the cost of generating DNA sequence per Megabase by using Sanger sequencing. From 2008 to 2015, the data is about DNA sequencing cost by using next generation sequencing. Sourced from [108] with permission.

Currently, the most widely used NGS platforms/technologies are Illumina (Solexa) sequencing [109], Roche 454 sequencing [110], Ion torrent [103] and

SoliD sequencing [111]. All of these technologies share the ability to sequence a large number of DNA molecules in parallel, providing advantages over classical Sanger sequencing in terms of speed, cost, sample size and accuracy. We take Illumina DNA sequencing (Illumina currently has the most sequencing machines in operation, globally) as an illustration of basic strategies involved in NGS sequencing (figure 1.2.3).

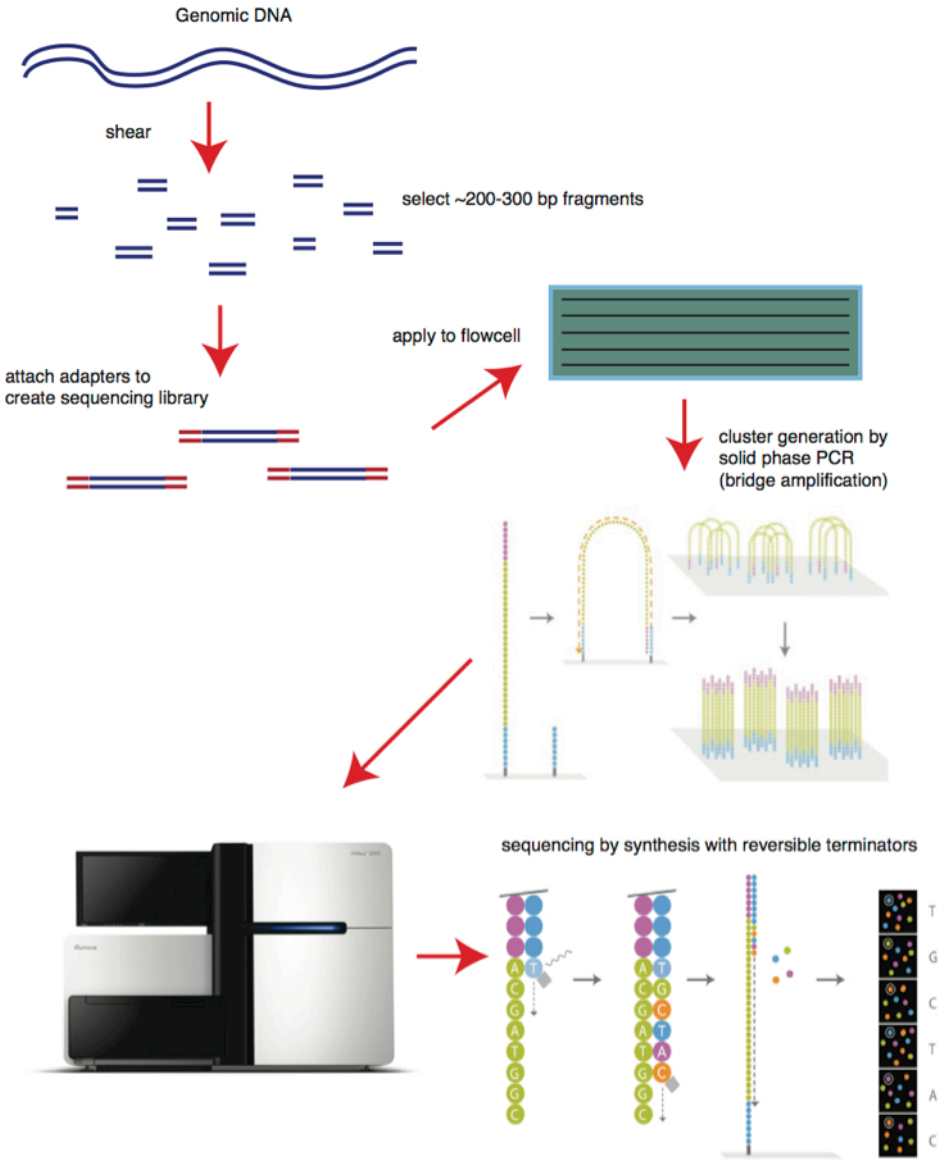


Figure 1.2.3 The workflow of DNA sequencing in Illumina platform. Sourced from [112] with permission.

First, the genomic DNA templates are sheared into more manageable short fragments usually with lengths between 200 and 300 base pairs. Second, the "adaptors" (short sequences of DNA) are attached to the DNA fragments and these combinations are then applied onto the flowcell with primers attached on the surface. The complementary DNA fragments bind to the primers and become attached on the flowcell. Once prepared, these DNA fragments are used as templates to generate many replicates of the DNA sequences using Polymerase Chain Reaction (PCR) amplification forming bridges [113]. Then the double-stranded bridged DNA sequences in the flowcell are denatured into single strands using heat, generating millions of dense clusters of identical DNA sequences. Last, several cycles of nucleotide base reading process are introduced. In each cycle, fluorescently labeled terminators (dideoxynucleotides), primers and DNA polymerase are added to the flowcell, and lasers are applied on the flowcell to activate the fluorescent labels of the terminator nucleotide bases, which can be detected by a camera and read by computers automatically.

A typical run of this sequencing process usually takes only a few hours to generate millions of DNA sequences with much lower sequencing cost comparing to Sanger sequencing. However, the fast accumulating sequencing data from NGS platforms also challenges researchers in terms of data storage and data interpretation. Furthermore, because current NGS platforms generate sequences with length usually ranging from 50 to 250 bp long, which are much shorter than the classical Sanger sequencing, this also poses challenges to the efficiency and accuracy of bioinformatics algorithms in genome assembly and sequence alignments. NGS technologies nowadays are fast innovating towards lower cost, higher accuracy and longer sequence lengths, which suggests that the analysis of NGS data will remain an important focus of bioinformatics research in the foreseeable future.

1.2.2 The construction of the human reference genome

The human reference genome is a digital nucleic acid sequence database assembled from DNA sequences obtained from a number of individuals as a representative example of a human genome sequence [114]. It is important because it provides a guide when new genomes are sequenced and it is also used as a template in genetic variation studies in individuals.

In April 2003, the first whole human genome was successfully completed under the Human Genome Project (HGP)[115]. The HGP was a 13-year-long publicly funded international project that aimed to determine the order of nucleotides in the human genome, performed by The International Human Genome Sequence Consortium, which included twenty research institutes and universities across the globe. Since then, with the emergence of new sequencing technologies [103-111], the human reference genome is maintained and frequently updated by the Genome Reference Consortium (GRC) [116]. The current reference genome build at the time this thesis is being written is GRCh38.p9, which was released on September 26, 2016.

A major computational challenge in determining the full human genome is genome assembly. This is because the total length of the human genome is about three billion base pairs, making it impossible for current sequencing technology to directly sequence from one end to the other. As described above, current sequencing methods normally generate short reads, usually less than 1000 base pairs. These reads are assembled like pieces of a puzzle to progressively form contiguous pieces, with larger pieces subsequently mapped, where possible, onto full chromosome sequences. There are two approaches to assemble a genome. When a complete and accurate reference genome is already available for the same species (or sometimes a closely related species), short reads can be mapped against the existing reference genome sequence, and assembled sequences, which are similar but not necessarily identical to the reference sequence, can be constructed. This approach requires less

coverage and it is very efficient for finding single nucleotide polymorphisms (SNP) in individuals. The major shortcoming of this method is that it may miss novel sequences or rearrangements that are significantly different from the reference sequence. The second method is called *de novo* assembly, which consists of assembling short reads from scratch to form full-length sequence without the help of reference template [117]. This approach requires more coverage and is more computationally intensive. The major advantage of this method is that the assembly process is not biased towards the reference sequence and it may identify novel sequences. Because of the different shortcomings of these two approaches, hybrid genome assembly approaches are widely used in recent years [118-121].

1.2.3 Databases of human genome variation

The construction of a human reference genome and development of sequencing technologies provide the possibility to systematically investigate human genome variations. Human genome variation refers to the variation in the DNA sequence of the genome of different individuals. It results in different forms of genes, known as alleles, which may directly influence the biological traits of individuals such as eye color, skin tone and nose shape [122]. Genetic variations can be divided into several forms based on the size of the mutations on the genome and the types of variation. Single nucleotide polymorphisms are the most common form of genetic variation in human (accounting for approximately 90% of the genetic variation) [123]. Single nucleotide polymorphisms are differences at a single nucleotide (insertion, deletion or substitution) across individuals. Structural variations include all the variations in the structures of the chromosome. For instance, copy number variation (CNV) is one important category of structural variations. CNV refers to the phenomenon that some sections of the genome are duplicated and the numbers of copies may vary across different individuals. Approximately 50% of the entire human genome is composed of repetitive sequence [124] and around 4.8-9.5% of the human genome can be classified as CNVs [125]. Genetic variation is particularly important in both evolutionary studies and medical applications. It can help scientists understand ancient human

population genetics, population migrations and allele frequency differences between global populations. For medicine, genetic variation can help explain some differences in susceptibility to certain diseases and different reactions to drugs.

1.2.3.1 dbSNP and dbVar

A number of archives of common and rare human genome variants are available in the public domain [126-129]. One of the most widely used public short-nucleotide-variation (variation < 50 base pairs) databases is the Single Nucleotide Polymorphism Database (dbSNP) hosted by the National Center for Biotechnology Information (NCBI) [130]. Although by name, dbSNP implies that it includes only SNPs, it actually includes different types of genetic variations besides SNPs, such as short deletions and insertions (indels), short tandem repeats (STRs) etc. Another widely used database hosted by NCBI named dbVar [131], contains longer variants (structural variations >= 50 base pairs) such as copy number variations, inversions and so on. Both dbSNP and dbVar accept submission of variants for any organisms from a wide range of sources such as individual research laboratories, genome sequencing centers, private businesses etc. In the current build of dbSNP (build 149), there are over 557 million submissions including more than 89 million human SNPs. However, the large number of submissions also makes it unrealistic for dbSNP to validate the quality of each of the submissions thoroughly. Several research groups have questioned the quality of the SNPs included in dbSNP, suspecting a high false positive rate [132-137]. These false SNPs could be easily submitted to dbSNP from experiments with flawed sequencing or bioinformatics analysis pipelines. Musemeci and colleagues reported that up to 8.32% of biallellic coding SNPs in dbSNP are noise resulting from highly similar DNA sequences [138].

1.2.3.2 Human genome resequencing projects

To resolve this chaotic situation, several genome resequencing projects aiming to provide an account of genomic variation found in global human populations were launched [139-143]. Among them, the most comprehensive public projects are the International HapMap project (HapMap) and 1000 genomes project, and both of these are used as validation sources in dbSNP. HapMap was launched in 2002 and since then, datasets from three phases of this project have been published. The most recent phase (phase three), of the project included DNA samples from 1,397 individuals in nine populations and reported 1.4 million SNPs.

Aiming to create the largest public catalogue of human genetic variation data, the 1000 Genomes Project (G1K) was launched in 2008. G1K project was completed in 2015 and by that point it had reconstructed the genomes of 2504 individuals from 26 populations (figure 1.2.4) using a combination of genome sequencing methods (whole-genome sequencing, deep exome sequencing, dense microarray genotyping) and several NGS platforms (Illumina, 454, SOLiD) [143]. In total G1K identified 88 million variants, consisting of 84.7 million SNPs, 3.6 million indels (insertions and deletions) and 60,000 structural variants. The G1K project provided by far the most detailed description of human variation, which facilitates studies of human evolution, genetic variant distribution and disease association. Furthermore, the G1K project broadens the knowledge of scientists about germline variations of genes within highly variable regions such as the T cell receptor and immunoglobulin genes.

The G1K project uses Variant Calling Format (VCF) to store sequence variations. Compared to the General Feature Format (GFF), VCF greatly reduces the size of the variant data by removing the redundant genetic information, which is shared across all the genomes [144]. Notably, VCF files provided by the latest phase of the G1K project (phase 3) are phased, which means that, in principle, complete sequences of gene alleles found on

individual haplotypes can be recovered, when the corresponding genome region has been sequenced to sufficient depth.

Figure 1.2.4 Populations of the samples that are included in the 1000 Genomes Project. Each circle in the figure represent a population and the circles labeled with the same color are from the same super populations (African, Ad Mixed African, East Asian, European, South Asian). Sourced from [145] with permission.

1.2.4 Transcriptome sequencing and analysis

Transcription is one of the fundamental processes leading to the generation of functional gene products (proteins and functional RNAs), which gives rise to phenotypes (observable traits) in individuals. The first step of gene expression is transcription, in which a particular section of the DNA sequence is copied into a new molecule of messenger RNA (mRNA). The transcriptome is defined as a set of all the mRNA molecules that are involved in the gene expression. The analysis of transcriptomes helps us reveal the functionalities of many genes, as well as illustrating the molecular basis of many diseases and indicating possible medical treatments.

1.2.4.1 Gene expression Microarrays

DNA microarray technology has been one of the most widely used methods for large-scale studies of gene expression since its invention in the 1990s [146]. DNA Microarrays revolutionized gene expression analysis by enabling the expression of tens of millions of genes to be investigated in parallel, along with a very affordable cost, which makes DNA microarray very popular in both academic and industrial research. Nonetheless, there are several major shortcomings of DNA microarray technology. Firstly, microarrays are limited to measuring the expression of genes for which the probes are designed (known transcripts), which makes it unsuitable for discovering novel transcripts. Another issue is that during the hybridization process, cross-hybridization can occur, adding background noise to the analysis. Last and most importantly, the results of microarray are represented as images with different fluorescently labeled spots; it still remains as an open question in bioinformatics to precisely quantify the expression level of the genes based on these images [147, 148].

1.2.4.2 Bulk RNA sequencing technology

RNA sequencing (RNAseq or whole transcriptome shotgun sequencing), which was first introduced by Mortazavi et al in 2008 [149], has gained in popularity as a means to analyze the transcriptome because it overcomes several of the problems that microarray technology has, as we mentioned above. RNAseq exploits high-throughput sequencing technologies to obtain high depth of coverage of RNA transcripts [150-152].

The first step of a typical run of RNAseq experiment is library preparation (figure 1.2.5): a population of targeted RNA sequences is first converted into a library of cDNA fragments using DNA or RNA fragmentation and reverse transcription. Second, adaptors are attached to the cDNA fragments on one end or both ends and NGS methods (Illumina, 454 etc) are applied to the cDNA sequences to amplify and obtain millions of short sequences. The final

and also the most challenging step is interpreting the RNAseq data. These short reads generated in a high-through manner are usually mapped to the reference genome using specialist bioinformatics tools. Along with the increasing use of RNAseq, the number of bioinformatics software tools for RNAseq analysis is also increasing rapidly. We will discuss the bioinformatics algorithms involved in RNAseq analysis in details in the following sections.

The RNAseq protocol described above is based on RNAseq data retrieved from a bulk population of cells, therefore it is also called bulk RNAseq. Current gene expression analysis methods involved in bulk RNAseq and microarray analysis estimate the mean value of a bulk population of cells by averaging the expression signal of individual cells. However, different cell types within a sample (for instance T cells, B cells, Natural killer cells etc) and cells at different life stages may present distinctive gene expression patterns and this information is not available from bulk RNAseq. For instance, the difference in the expression level of specific transcripts can vary as much as 1000-fold between presumably identical cells [153]. These problems can be resolved by methods in which gene expression analysis is applied at the single cell level. Although transcriptome analysis at the single cell level has been available for some years, until recently the data generated was generated in a low-throughput manner [154-156].

1.2.4.3 Single cell RNA sequencing technology

In more recent years, with the advent of more sophisticated sequencing and isolation technologies, various methods of single cell RNAseq (scRNAseq) have been reported [157-162]. In general, these scRNAseq methods include two independent techniques. The first technique consists of single-cell isolation: individual cells are first harvested by micromanipulation such as patch pipette or nanotube, but most commonly, by using fluorescence activated cell sorting (FACS) [163]. The second technique concerns preparation of the cDNAs libraries and amplifying the cDNA by using NGS methods, which is similar to the bulk RNAseq as we described above. More recently, spike-ins and Unique Molecular Identifiers (UMIs) are widely used

in scRNAseq experiments to quantify the technical biases between different scRNAseq libraries [158,159].

There are still some major shortcomings of current scRNAseq methods. First, unlike the sequencing libraries in bulk RNAseq, in scRNAseq, each sequencing library represents one cell, which may introduce a discrepancy between different libraries as a result of artificial noises such as different amplification levels between cells. Specialized normalization or spike-in methods have to be carefully introduced into scRNAseq analysis to resolve this problem. Furthermore, scRNAseq data may suffer "gene dropout" problem, some of the genes are moderately expressed and may not be detected in a given cell (164). Finally, current datasets of scRNAseq range from 10^2 to 10^5 cells because of the relatively high cost, which may be unsatisfactory to investigate the expression of different cells in complex systems such as the immune system. However, many of these challenges may be resolved with the anticipated fast pace of innovation in sequencing technologies and scRNAseq is likely to become a major focus of research in the foreseeable future.

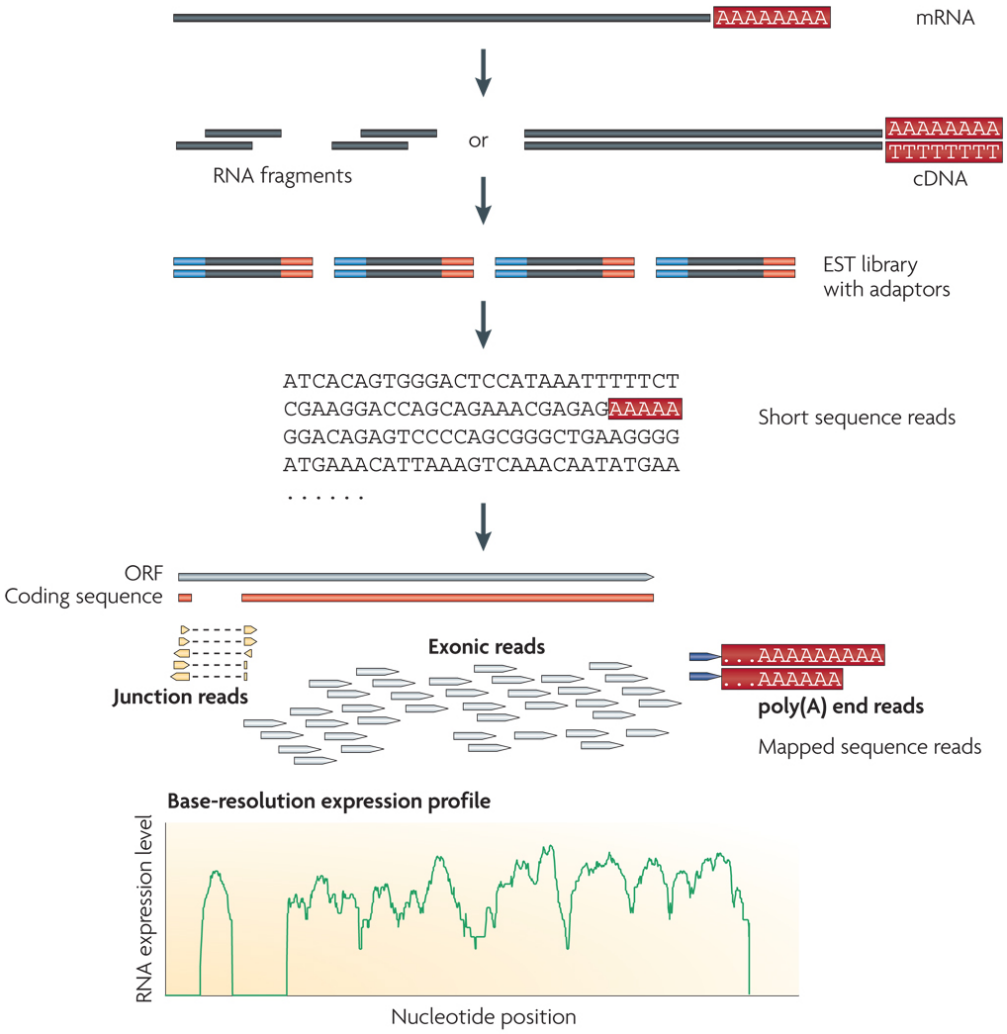


Figure 1.2.5. A typical workflow of RNAseq experiments and analysis. Sourced from [165] with permission.

1.2.4.4 Computational analysis of RNAseq data

A typical run of RNAseq analysis includes quality control (QC), sequence mapping, expression quantification and differential expression analysis (figure 1.2.6).

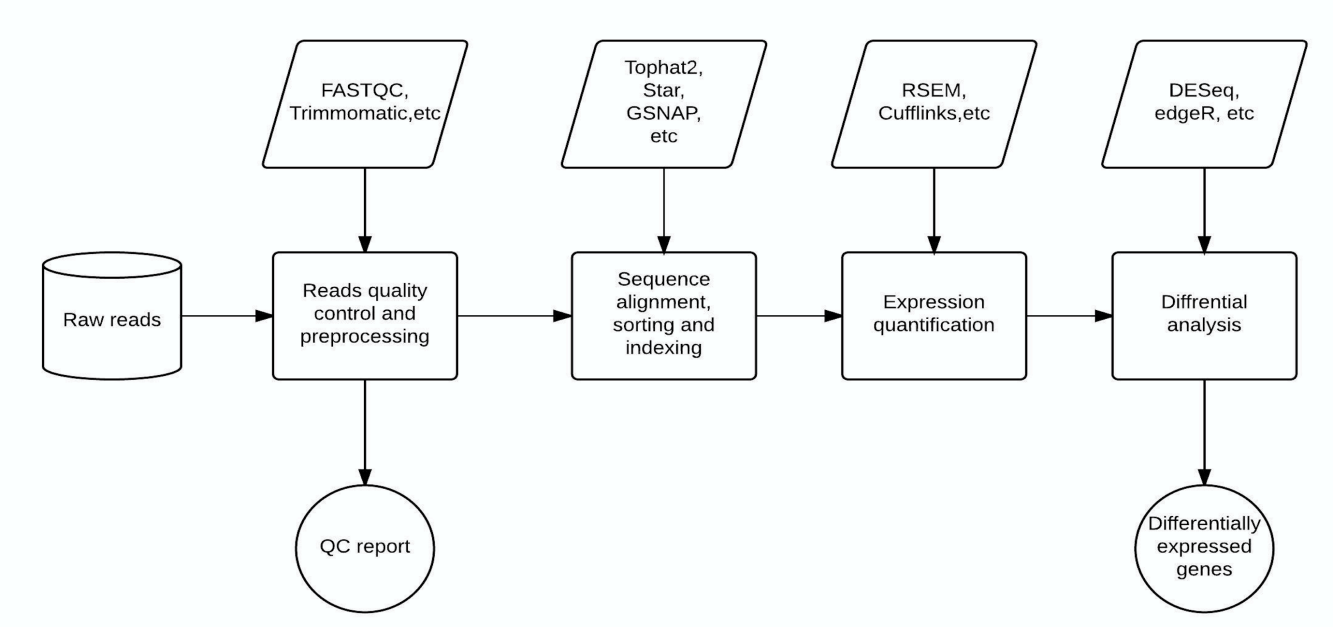


Figure 1.2.6 a step-wise workflow of typical steps involved in RNAseq data analysis.

Quality control is a critical process in all kinds of NGS data analysis. As we have described in previous chapter, a typical run of deep sequencing yields millions of nucleotide sequences, and the quality of each sequence may vary due to different platform-specific artifacts. Poorly sequenced reads substantially influence the accuracy of the subsequent analysis. The per-base quality score was introduced to reflect the confidence that the base is correctly called. The Phred quality score is one of the most widely used quality scores and it is defined as follows [166, 167]:

$$Q = -10 * log_{10} p$$

Where p is the probability that the base call is incorrect. Typically, the sequences generated from NGS platforms are reported with both the corresponding per-base quality scores, and sequence data in this form are named as FASTQ format. In order to optimize the data storage of FASTQ files, the quality score is encoded as ASCII codes by converting the numerical value to the its corresponding ASCII character. One of the major approaches of quality control is analyzing the sequence quality and trimming the error-

prone sequences with poor quality scores. In addition to that, there are other quality issues that need to be resolved with special care, such as sequence duplications and the presence of adaptors etc. There are many tools designed for addressing the above issues, such as FastQC [168], NGSQC [169], FASTX-toolkit (170) and Trimmomatic [171].

After quality control, the analysis-ready reads are mapped to a set of reference sequences to determine their genomic locations and corresponding transcripts. The basic objective of sequence mapping or sequence alignment is to find the most similar region in the reference sequence for each input read. The strategies involved in sequence mapping can vary across different RNAseq datasets depending on the availability of reference data from the targeted organism, thus this step can be further divided into two scenarios: mapping to the reference genome or mapping to the reference transcriptomes. Mapping to the reference genome is mainly used in overall gene expression studies to investigate the abundance of all the included transcripts; bioinformatics tools such as TopHat [172] and STAR [173] are often used in this scenario. In studies that are focusing on a small sets of transcriptomes (for instance the transcriptomes from T cell receptor or immunoglobulin genes) and the corresponding transcriptome reference sequences already exist and are well annotated, the reads are mapped to a small set of reference transcriptomes, which saves large amounts of computational resources. Tools such as Bowtie [174] and HTSeq [175] are widely used in this scenario.

De novo transcriptome assembly is often applied on RNAseq data from organisms without a reference model (figure 1.2.7). Transcriptome assembly is different from genome assembly as we described earlier, by presenting some unique challenges, thus a range of customized transcriptome assemblers have been developed [177-180]. Briefly, the basic idea of these assemblers is that they first assemble the short reads into longer contigs using different algorithms such as overlap graphs or de-Bruijn graphs, subsequently, they treat these long contigs as the reference transcriptomes, and the raw reads are mapped back to them. Either with reference-based mapping or *de novo* assembly, transcriptome mapping based on short reads remains as a

challenging problem. There are growing trends of long-read technologies such as Smart-Seq [162] and SMRT (Single Molecule Real-time Sequencing Technology) [181] in recent years, which are believed to be promising approaches that can improve the power of the transcriptome analysis.

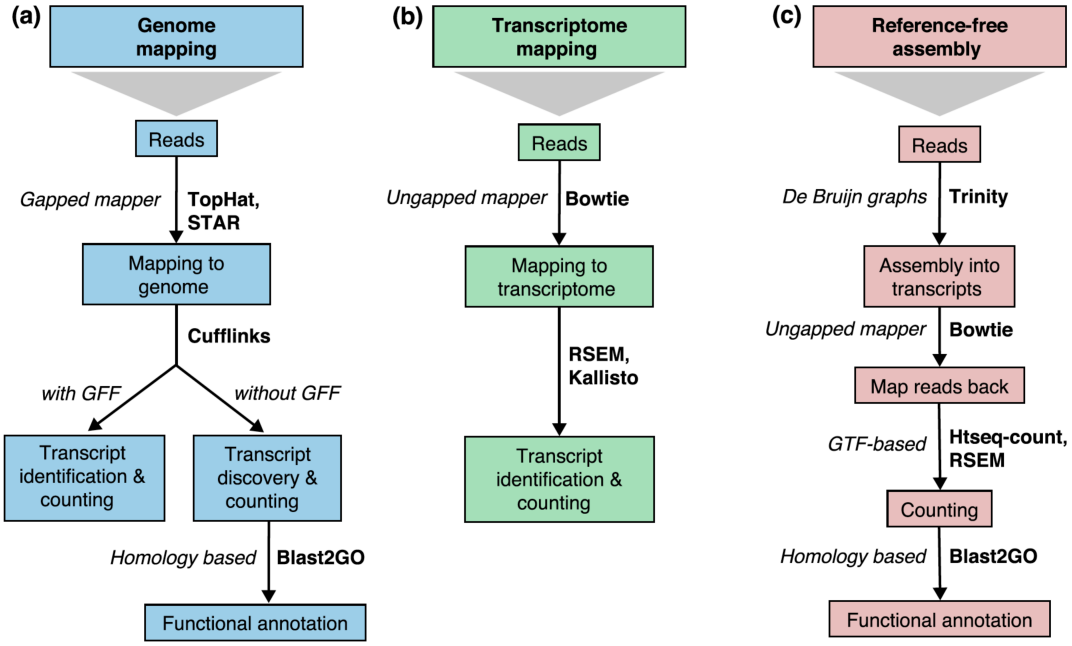


Figure 1.2.7 Three different approaches used in transcriptome mapping. a. reference genome based approach. b. transcriptome reference based approach. c. non-reference based (de novo) approach. Sourced from [176] with permission

One critical challenge in transcriptome analysis is to accurately determine or quantify the expression level of different genes. The intuitive approach for quantifying the expression of one particular gene is to simply count the number of sequences that are mapped to it (this approach is used in tools such as HTSeq-count and feature counts) [182]. However, counting the raw reads alone is problematic due to several factors, including different length of the transcripts, different total number of the reads (library size) across samples and sequencing biases in different batches of experiments. RPKM (read per kilobase per million mapped reads) is often used in transcript quantification by calculating the mapped reads normalized by the library size and the transcript length [183]. Another similar metric, FPKM (fragment per base per

million) differs from RPKM by counting the mapped fragments instead of the mapped reads, which is preferred in paired-ended RNAseq experiments. Because in the paired-ended RNAseq data, one cDNA fragment corresponds to a pair of sequence reads, and this does not mean that both reads can be successfully mapped, thus, counting reads may lead to bias towards some of the cDNA fragments that contain both reads successfully mapped. TPM (transcript per million) is a relatively new metric reported in RSEM [184], which provides a more comparable expression value between different samples, a detailed description of TPM calculation is described in the methods section in Chapter Four.

Transcriptome quantification is the essential prerequisite for differential gene expression analysis, which is the most fundamental research problem that most of the RNAseq experiments are designed for. The main objective of differential gene expression analysis is to find the genes that are differently expressed between conditions such as healthy or diseased, which help us understand the molecular basis of phenotypic variations in these conditions. There are several bioinformatics software applications developed for detecting differentially expressed genes by using a range of statistical models such as edgeR [185], DESeq [186], baySeq [187], NOISeq etc [188]. However, special attention has to be paid while choosing different bioinformatics analysis pipelines as many benchmarking studies have suggested that there was no "gold standard" pipeline that can significantly outperform all the other tools on different datasets [189-191].

As we have described in previous chapters, single cell RNAseq differs from bulk RNAseq in that each cell in single cell RNAseq experiments represents one library, instead of a population of cells as in bulk RNAseq experiments. Additional cell quality control is introduced to filter out cells with poor sequence quality and unsatisfactory amplification [192]. In addition to that, the amount of RNA sequenced from different cells may vary considerably; thus, between-cell normalization is an essential step in single cell RNAseq analysis. There are also some additional controls introduced in single cell RNAseq experiments. The most common approaches are spike-ins and

Unique Molecular Identifiers (UMI) [193, 194]. The spike-ins are artificial RNA molecules, such as those from the External RNA Control Consortium. Spike-ins with known concentration are added to lysate of the cell before the transcription reaction, which can help reveal the quality of the library. For instance, if the sequences are mainly mapping to the spike-ins instead of the endogenous RNAs, it indicates that the quality of the library is poor. Furthermore, because the concentrations of the spike-ins are known, these can be used to predict the extent of the amplification, which can be used in the cell-to-cell normalization process. Single cell RNAseq did not gain wide popularity until 2014 when the sequencing cost became sufficiently low, and currently there are only a handful of tools that are specialized for single cell RNAseq analysis [195-197]. Although several computational analysis methods in bulk RNAseq can be applied on single cell RNAseq data, special adaptions of these methods are needed. With growing interest from scientists in this technology, more powerful single cell RNAseq analysis tools will be developed in the near future.

1.2.5 Lymphocyte receptor repertoire profiling

As discussed earlier V (D) J gene rearrangements and somatic hypermutations (for immunoglobulins only) lead to a very high level of diversity of immunoglobulin receptors. In addition to that, many TCR/IG sequences share very high similarities within their repertoire. The large diversity and high similarity of this genomic region provides challenges both for the sequencing technologies and computational analysis. Therefore, specialized immunological repertoire sequencing protocols and bioinformatics analysis pipelines have been developed.

1.2.5.1 Lymphocyte receptor sequencing protocols

There were many studies using low-throughput methods to investigate the lymphocyte receptor repertoire since the 1990s [198-200]. Most of these studies investigated the VDJ recombinants of T cell receptors and

immunoglobulins to determine the CDR3 region in up to hundreds of cells per experiment. These studies yielded many important immunological insights concerning the recognition mechanisms of TCRs and immunoglobulins. However, although most of these discoveries were successful, the number of TCR and immunoglobulin sequences involved in these studies were usually up to hundreds, which is far below the theoretical diversity (up to 10^{12} clonotypes), thus providing only a very limited view of the full lymphocyte receptor repertoire.

The application of NGS technologies in recent years has shed some light onto immunological repertoires by generating much higher coverage of TCR/IG sequences, providing the possibility to gain a much broader picture of the immune repertoire. There are several NGS platforms that can be applied to lymphocyte receptor sequencing [201]. A typical run of lymphocyte receptor sequencing (figure 1.2.8) starts with careful selection of the subpopulation of T and B cells of interest, due to the well-acknowledged heterogeneity of T cell receptors and immunoglobulin between individuals and cell types (CD8+ T cells, CD4+ T cells, naïve and memory T/B cells etc.) [202, 203]. Most commonly, human peripheral blood is the main source for harvesting T and B cells in these studies because of its convenience. Notably, using the T and B cells retrieved from peripheral blood only revealed a limited view of the human lymphocyte repertoire, because only a small fraction of T and B cells are present in peripheral blood in human. For instance, only approximately 2% of the B cells are found in peripheral blood [204].

Similar to other NGS protocols described above, the second step in lymphocyte receptor sequencing is library preparation and amplification. The primers here need to be carefully redesigned because different V and J gene segments are included in different TCRs and immunoglobulins, ruling out the possibility of creating a common primer for the CDR3. Thus multiplex PCR is usually introduced in lymphocyte receptor sequencing experiments by using unique primers for all the possible combinations of the V and J segments [205]. However, there are many shortcomings of the multiplex PCR approach, such as non-specific amplification, primer-dimer formation and uneven

reactions. Most importantly, multiplex PCR methods may introduce biases: The amplification levels of different primers may vary among different combinations of V and J segments, resulting in biases towards certain TCR or immunoglobulins, which contain V and J combinations with primers that are more amplified [206]. Further normalization steps based on barcoding approaches could be useful to resolve this issue [207].

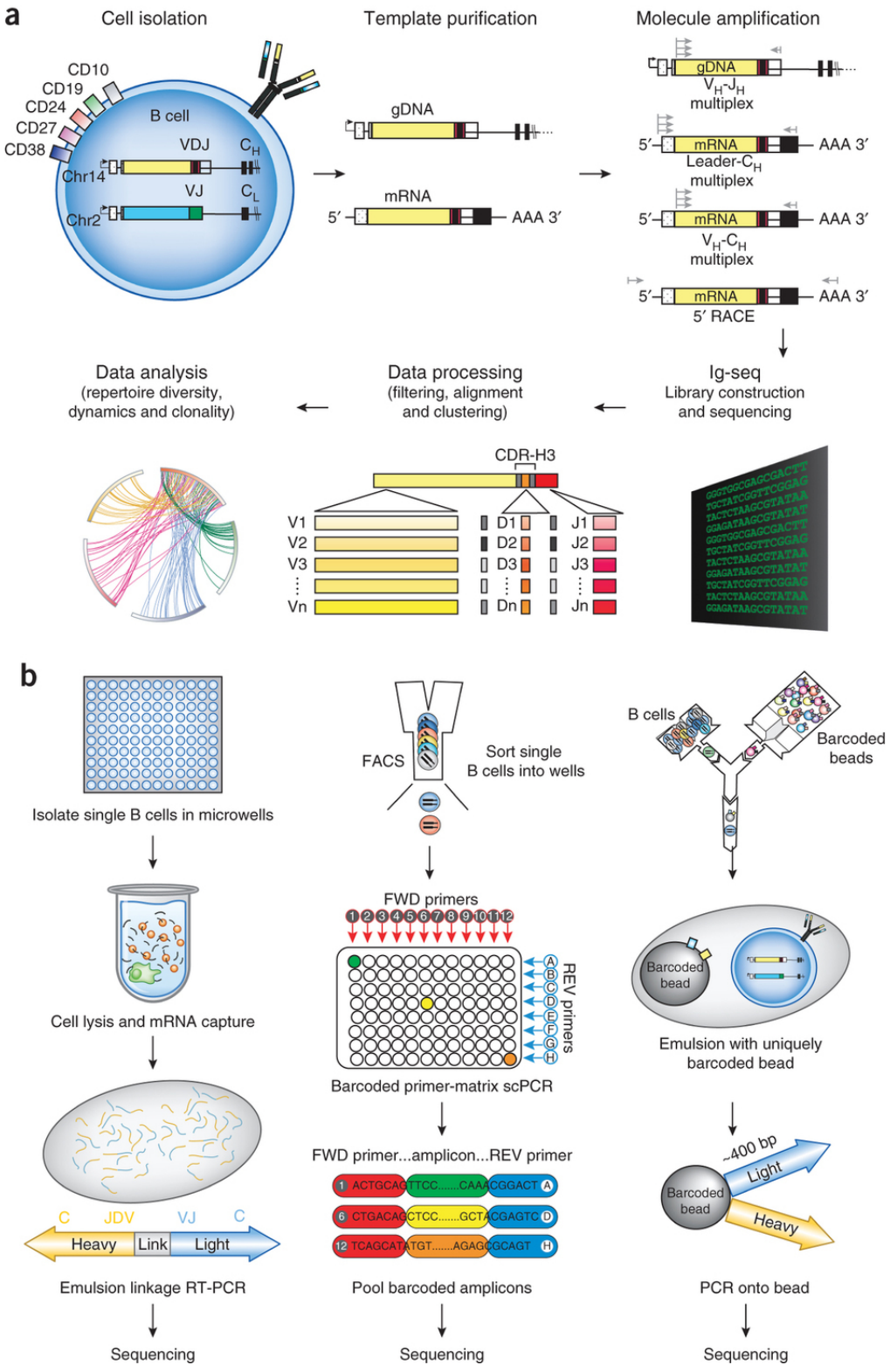


Figure 1.2.8 Workflows of immunoglobulin sequencing by using NGS approaches. (a). Immunoglobulin sequencing from a bulk population of B cells. (b). Three different methods in Single B cell RNA sequencing. Sourced from [204] with permission.

1.2.5.2 Bioinformatics pipelines for immune repertoire analysis

The TCR and immunoglobulin repertoire provides a screenshot of the immune status of individuals in different conditions such as disease, aging or autoimmunity, and they can be further regarded as important biomarkers for personalized medicine. The accurate understanding of immune repertoire relies on sophisticated computational analysis methods. However, computational analysis of TCR and Immunoglobulin NGS data present some unique challenges due to the extraordinary heterogeneity of the immune repertoire [204]. In recent years, a range of bioinformatics tools were developed, specialized for immune repertoire analysis [208-219]. The fundamental objectives included in all the tools are to correctly map the raw sequences to their corresponding genomic V, (D) and J gene segments, determine the CDR3 region, and then to calculate the abundances of different clonotypes. Using these results, further tasks include determining the frequency distribution across clonotypes, calculating different V (D) J usage and measuring the diversity of the repertoire (figure 1.2.9).

Due to the unique genetic recombination process involved in TCRs and immunoglobulins, gene-mapping algorithms used for lymphocyte receptor sequences are different from most of the alignment algorithms used in other bioinformatics alignment tasks. A raw TCR or immunoglobulin sequence generated from a NGS platform usually consists of part of the V, all the D (for IGH and TCRB only) and part of the J gene segment. The ultimate purpose of sequence mapping is to correctly determine which V, D (immunoglobulin heavy chain and TCR beta chain only) and J gene is used in one particular lymphocyte receptor sequence. This process includes two important factors that can directly influence the reliability of the analysis. The first factor concerns the comprehensiveness and accuracy of the reference sequences of lymphocyte receptor genes. Currently, IMGT is the most widely used reference database in all the TCR and immunoglobulin sequence specialized bioinformatics tools. However, as we outline in Chapter three, several studies

have suggested that IMGT might be incomplete and contain errors [220-222]. Second, immune receptor gene mapping requires robust and fast alignment algorithms, which can effectively determine the correct reference gene segments based on truncated and heavily mutated sequences. In chapter two, we described in detail the VDJ alignment algorithm we developed in our bioinformatics pipeline for immune receptor repertoire analysis [217].

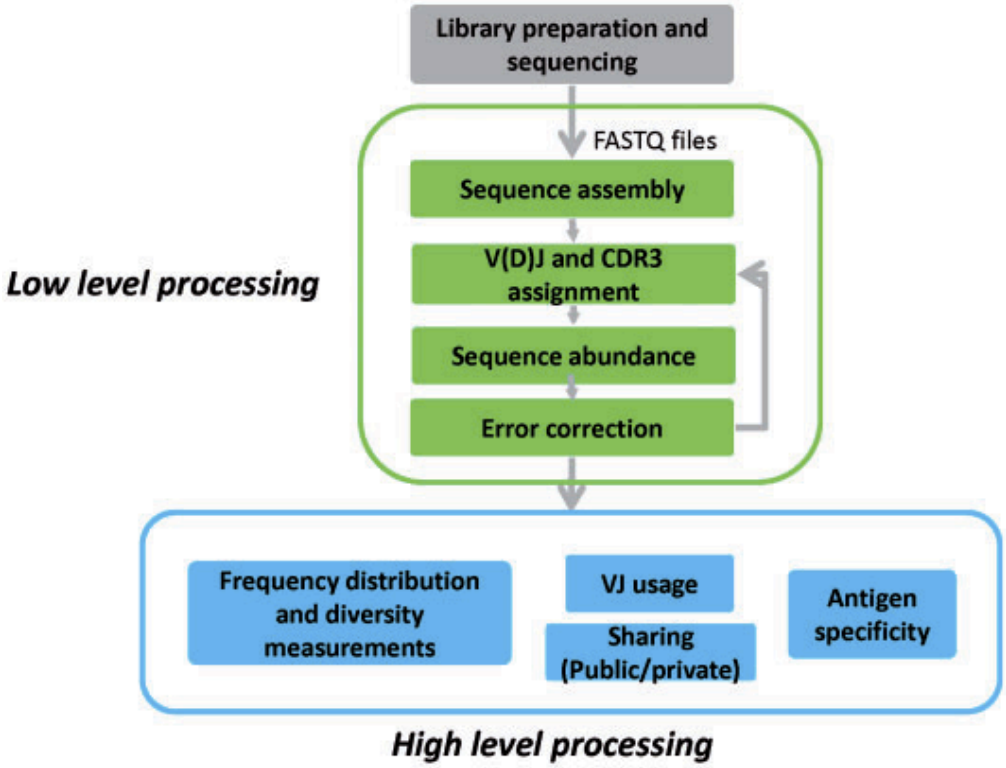


Figure 1.2.9 Common steps involved in the computational analysis of T cell receptors and Immunoglobulins. The analysis can be further divided into low-level processing and high-level processing. In low-level processing (green box), the raw sequences are assigned to their corresponding reference V (D) J gene segments with different alignment and error correction algorithms. The high level processing (blue box) mainly involves interpreting the results to reveal the clinical and biological meaning of the given TCR/IG repertoire. Sourced from [208] with permission.

In addition to the V (D) J gene alignment, immune repertoire analysis usually includes the determination of hyper-variable CDR3 region, which plays a crucial role in antigen binding. The CDR3 region is defined as the sequence

starts from the last cysteine of the V gene segment to the conserved [FW] GXG motif (X represents any amino acids) on the J gene segment. The determination of the CDR3 region is relatively straightforward by targeting the conserved motifs, but it needs special care while determining the last cysteine of the V gene segment. This is because for some of the TCR and immunoglobulin sequences, there may be two or more cysteines located before the [FW] GXG motifs of the J gene, thus it is hard to determine which one is the last cysteine on the V gene. For this reason most of the bioinformatics tools firstly determine the reference V (D) J sequence of the raw sequences, and use the reference location of the V genes to find the last cysteine.

After assigning the V, (D) and J genes and extracting the CDR3 region, one particular TCR or immunoglobulin clonotype can be represented by its CDR3 nucleotide sequence, reference V, D and J gene, which is used as the standard output format in most of the immune receptor specialized bioinformatics tools mentioned above. By counting the abundance of each clonotype, we can construct the frequency distribution of different clonotypes and this can be further used to determine the direction of the clonal expansion of the TCR/IG repertoire in different conditions. In addition to that, diversity measurements of the TCR and immunoglobulin may be of direct interest, such as for the analysis of the repertoire diversity in clinical samples [223, 224]. Notably, it is important to use suitable diversity measures to reveal the true diversity of the TCR/IG repertoires. There are several diversity measures that can be applied to immune repertoire analysis, such as Shannon entropy, Simpson Index and Rényi entropy [225]. All of these take into account two factors: the number of different clonotypes (richness) and the evenness of the distribution across clonotypes. Measures of immune receptor repertoire diversity are discussed in greater detail in Chapter four.

In summary, immunological sequencing technologies are evolving continuously in recent years, providing a more precise estimate of immunological repertoires from high-resolution sequencing data. However, considering the size and complexity of the immune repertoire, further

improvements are still needed in constructing standard immunological sequencing experiments with less background noise. In addition to that, normalization steps based one spike-in approaches for different primers of VJ combinations are necessary in sequencing protocols. Furthermore, during the TCR/IG sequences analysis, besides carefully choosing the alignment algorithms (described in detail in Chapter Two), the selection of the reference sequences also directly influences the accuracy of the results (further described in Chapter three).

1.3 Aims and objectives

In this thesis, we set out to create a sophisticated high-throughput solution for immunological repertoire analysis, and to apply it to investigate biological questions. Specifically, we first aimed to develop a robust and effective bioinformatics algorithm to process next generation sequencing data from lymphocyte receptors. Subsequently, we asked if it is possible to infer a more comprehensive collection of TCR and immunoglobulin reference sequences based on the abundant human resequencing data accumulated in recent years. Lastly, we aimed to analyze the characteristics of the allelic diversities of the TCR and immunoglobulins, using high throughput genomic and transcriptome sequencing data available in the public domain.

In the first research chapter, we set up the objective to develop a software application that integrates sophisticated alignment algorithms and additional features for statistical analysis of TCR and immunoglobulin repertoires. Immunologists have shown growing interest in understanding the frequency distribution of different clonotypes of TCR and immunoglobulins in different conditions. More efficient and accurate bioinformatics pipelines specifically tailored for TCR and immunoglobulin sequencing analysis are needed because of the complexity of the lymphocyte receptor repertoire.

In addition to the alignment algorithm, the completeness and accuracy of reference sequences for TCR and immunoglobulin genes used during the alignments has a large effect on analysis pipelines. Thus in the second

research chapter, we aimed to create a more comprehensive collection of reference sequences of TCR and immunoglobulin gene based on the human resequencing data, given the fact that currently the most widely used reference database IMGT appeared to be incomplete [220-222].

In the last research chapter, we aimed to investigate some research questions relating to the germline allelic diversity of human immune receptor genes by utilizing the high-throughput approaches we developed in the previous projects. We first asked if there is any difference in the allelic diversities of different TCR and immunoglobulin genes among the global populations, given the fact that there were more genetic variations in African populations than the Non-African populations [227]. Furthermore, we asked if allelic diversities of the immune receptor genes were associated with their genomic locations. Last, we hypothesized that more frequently used genes are under higher diversifying selection pressure, resulting in higher allelic diversity. To investigate this hypothesis we examined the correlations between the allelic diversities of the TCR and immunoglobulin genes with their RNA expression.

Chapter 2 – LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins

The content of this chapter was published as:

Yu, Y., Ceredig, R. and Seoighe, C., 2015. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. Nucleic acids research, p.gkv1016.

2.1 ABSTRACT

The adaptive immune system includes populations of B and T cells capable of binding foreign epitopes via antigen specific receptors, called immunoglobulin (IG) for B cells and the T cell receptor (TCR) for T cells. In order to provide protection from a wide range of pathogens, these cells display highly diverse repertoires of IGs and TCRs. This is achieved through combinatorial rearrangement of multiple gene segments in addition, for B cells, to somatic hypermutation. Deep sequencing technologies have revolutionized analysis of the diversity of these repertoires; however, accurate TCR/IG diversity profiling requires specialist bioinformatics tools. Here we present LymAnalzyer, a software package that significantly improves the completeness and accuracy of TCR/IG profiling from deep sequence data and includes procedures to identify novel alleles of gene segments. On real and simulated datasets LymAnalyzer produces highly accurate and complete results. Although, to date we have applied it to TCR/IG data from human and mouse, it can be applied to data from any species for which an appropriate database of reference genes is available. Implemented in Java, it includes both a command line version and a graphical user interface and is freely available at https://sourceforge.net/projects/lymanalyzer/.

2.2 INTRODUCTION

T cell receptors and Immunoglobulins recognize diverse arrays of foreign antigens and play important roles in the adaptive immune response. The diversity of TCRs and IGs is achieved by V(D)J recombination (for both TCRs and IGs) and somatic hypermutation (for IGs). V(D)J recombination is a stochastic process of rearrangement of variable (V), joining (J) and diversity (D, for the TCR beta chain and IG heavy chain only) gene segments during the early stages of T and B cell maturation. Somatic hypermutation is the T cell-dependent process through which IGs undergo extremely high rates of somatic mutation during the proliferation of B cells in germinal centres. As a consequence of this hypermutation process, B cells are selected for their expression of higher affinity IG's, a process called affinity maturation. The complementarity determining region 3 (CDR3) which includes part of the V, all of the D and some of the J gene segments is the most variable region of TCR/IG sequences and plays the major role in binding specificity. In man, theoretical estimates of the number of distinct TCR and IG generated by this mechanism are around 10^{10} [228]. The analysis of CDR3 diversity within individuals reveals insights into the mechanisms of adaptive immunity as well as clinically relevant information about the state of the immune system in individual patients [229]. Therefore, robust bioinformatics pipelines for comprehensive analysis of TCR/IG diversity are required.

Compared to the Sanger sequencing technology, next generation sequencing (NGS) technology provides information at much higher resolution about the DNA sequences of TCR and IG, allowing more complete analysis of lymphocyte repertoires. This gives us an opportunity to gain a better understanding of adaptive immunity. Typically, the main objectives are to identify the VDJ genes, extract the CDR3 region and estimate the diversity of the lymphocyte repertoire. Existing software packages are available for VDJ identification and CDR3 extraction. IgBlast [210] and IMGT/High-V-Quest [209] are both web-based tools for TCR/IG sequence analysis that make use of dynamic programming sequence alignment algorithms. These tools include user-friendly graphical user interfaces (GUIs), and they are fast and robust

enough for the analysis of small numbers of TCR/IG sequences. iHMMune-align [230] uses a hidden Markov model to align IG sequences. However, for high throughput sequencing datasets, these three tools are no longer suitable due to the limited numbers of sequences they can process (no more than 150,000 reads), as all of these tools were developed for sequence data generated by traditional sequencing technologies. More recently, Decombinator [211] and MiTCR [214] were developed specifically for the analysis of NGS data from TCRs (neither tool currently allows the analysis of IG sequences). MiXCR [214] is the most recently developed tool for the analysis of TCR/IG data. However, we demonstrate here that techniques used to achieve the speed required for the analysis of NGS data by these tools result in reduced accuracy in VDJ gene assignment and an incomplete profile of TCR diversity. Here we present LymAnalyzer, a software package for the comprehensive and accurate analysis of TCR/IG NGS data. The alignment step in LymAnalyzer, which is based on a fast-tag-searching algorithm, results in rapid identification of VDJ gene segments, with significantly improved accuracy and completeness compared to existing tools applied to TCR data. In addition, LymAnalyzer can be applied to IG sequences, includes an integrated single nucleotide polymorphism (SNP) calling algorithm that identifies novel alleles of the VDJ gene segments and produces lineage mutation trees to represent the affinity maturation process of the IGs.

2.3 MATERIALS AND METHODS

2.3.1 The workflow of LymAnalyzer

LymAnalyzer consists of four functional components: VDJ gene alignment, CDR3 extraction, polymorphism analysis and lineage mutation tree construction (figure 2.1).

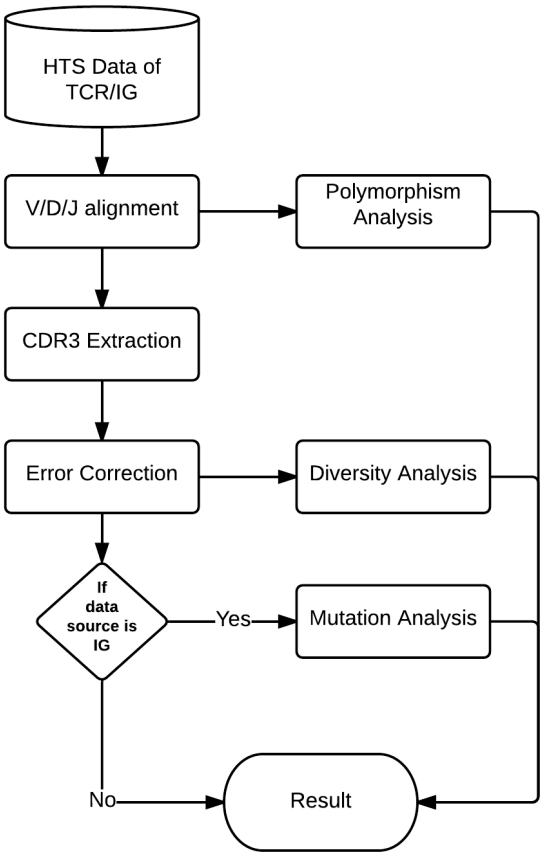


Figure 2.1. The stepwise workflow of LymAnalyzer.

TCR/IG Diversity analysis is the first process in the pipeline. This process includes three steps, the first of which is V/D/J alignment. For each input sequence, we use a fast-tag-searching algorithm, described in detail below, to determine the reference V, D and J genes from which this input sequence is derived. Each input sequence is aligned against all sequences in the International Immunogenetics Database (IMGT) [78] and the best matching V, D and J sequences are selected. In the second step we extract the CDR3 region from the sequence. The CDR3 region of TCR/IG begins with the last cysteine of the V segment and ends with the conserved motif [FW] GXG (X represents any amino acid) in the J segment. The conserved motif in the J segment is straightforward to identify in the input sequence because 12 nucleotides are sufficient to ensure unique occurrence within the TCR/IG sequences. But for the cysteine motif in the V segment, there may be two

cysteines both located towards the end of the sequence. In our pipeline, for each of the input sequences, we first find the location of the last cysteine of the corresponding reference V gene we obtained from the alignment step. We then map this location back to the input sequence to determine the position of the last cysteine on the input sequence. This allows us to determine the sequence of the CDR3 region. The third and final step involves classification of the CDR3 sequences. After we obtain the CDR3s from previous step, we classify the input sequences. CDR3 sequences are clustered into clonotypes and the number of clonotypes and the number of sequences per clonotype are calculated. This process is discussed in detail below.

Users can also choose to perform polymorphism analysis to identify novel SNPs that do not correspond to alleles that are not contained in the IMGT database. Each potential SNP, as well as the observed frequencies of the alternative alleles, is included in the result report. For immunoglobulins, by default LymAnalyzer will also create lineage trees that describe the stepwise somatic hypermutation of immunoglobulin sequences in the germinal center [231].

2.3.2 NGS data for TCRs/IGs

We obtained TCR/IG sequence data from the NCBI Sequence Read Archive (SRA) to test the performance of LymAnalyzer. The experimental data consisted of two datasets. The TCR sequence data (SRA index: PRJNA229070) used here is from Putintseva et al [232]. It consists of nine samples; the number of reads in each sample ranges from 4,202,419 to 13,872,805. The reads are all from the beta chain of TCR covering part of V, all the D and part of the J region (100bp long). A second dataset, consisting of immunoglobulin sequences, (SRA index: SRP017087) is from Doria-Rose et al [233]. This dataset contains seven samples, with read counts varying from 271,382 to 23,191,224. Each sequence is 250bp long and comes from the heavy or light chain of the IG. It contains part of V (all the D for heavy chain) and part of the J region. Putintseva et al used MiTCR to analyse the TCR data, Doria-Rose et al exploited their own bioinformatics pipelines, which included BLAST to process the IG data.

2.3.3 Simulated dataset

We used simulation to compare the accuracy of LymAnalyzer and existing tools. Firstly we created a reference gene database: the reference V, D and J gene database used in our simulation pipeline is obtained from the latest version of IG/TCR repertoire of IMGT database. For each of the simulated sequences, we selected the V, D and J gene segments assuming uniform gene usage from the reference gene database. Mismatches were introduced to simulate the combined effects of PCR errors, sequencing errors and mutations/polymorphisms in the input sequences, each of which can lead to differences between the input sequences and the corresponding gene segments in the database. Three mismatch levels were used: no mismatches, a low mismatch level and a high mismatch level. For the low mismatch level, there were 0-7 mismatch(es) on the V gene, 0-1 mismatch on the D gene and 0-3 mismatch(es) on the J gene. For the high mismatch level, there were 0-15 mismatch(es) on the V gene, 0-2 mismatch(es) on the D gene and 0-5 mismatch(es) on the J gene. The number and the position of the introduced mismatch(es) in the corresponding gene were both uniformly distributed. After obtaining the "mutated" V, D and J segments, we added 0-6 randomly generated nucleotides to the V-D and D-J junction to simulate nucleotide insertions during VDJ recombination. We generated three datasets, with varying mismatch rates, each consisting of 20 samples. Each of the samples contained 200,000 TCR/IG sequences.

2.3.4 Fast-tag-search based alignment algorithm

Due to the large size of NGS datasets fast algorithms are required for sequence alignment. LymAnalyzer uses an alignment algorithm based on fast-tag-searching to map the input sequence to reference V and J segments. We first define a detection tag set that consists of multiple short detection tags from the input string. Iteratively we use detection tags from the tag set to search for perfect matches in the second string and store the indexes that obtain perfect matches. Subsequently we calculate the Hamming distance (the number of positions at which the corresponding symbols are different) of these two strings by extending from each perfect match index (figure 2.2). By default, the reference VDJ genes used by LymAnalyzer are derived from the most

recent update of the IMGT database; however, users can also choose to import their own reference gene database. For each of the reference genes, we select the last five nucleotides from the 3' end of the sequence as our first detection tag T1. Then we select another five nucleotides, which are located adjacent to the previous detection tag by extending towards the 5' end. The same operations are repeated until we obtain five detection tags (The number of the tags is an adjustable parameter that can be defined by users; it is five in the default setting) and we get the detection set V as

$$V = \{T_1, T_2, T_3, T_4, T_5\} \quad (1)$$

For any reference J genes, instead of choosing the last five nucleotides of the 3' end, the algorithm starts from the 5' end. The same operations as we described for the reference V gene are repeated three times (This is an adjustable parameter that can be defined by users; it is three in the default setting) to get the J set where

$$J = \{T_1, T_2, T_3\} \quad (2)$$

Furthermore, we locate the indexes that have perfect matches with the first tag, T_1, for each of the input sequences. However this may not be successful due to mutations and sequencing errors in the matching region; hence, the five detection tags used for the V genes and three detection tags used for the J genes. If the preceding tag fails, we used the subsequent tag to repeat the matching. Once we find perfect matching index(es), we extend in both directions from this index and calculate the matching score, which consists of the number and percentage of matches between the input and reference sequence. If the number of matches passes the minimum threshold (i.e. 90% match and 30bp matches), we keep the corresponding reference gene in the candidate set. After this process is applied to all of the reference genes, we choose the sequence with the highest percentage of matches from the candidate set. As J genes are shorter than V genes, we use only three detection tags in the J genes and the minimum threshold requires only 20bp matches. D gene alignment is different from V and J gene alignment because D genes are short (12-16bp) relative to V and J genes, D genes are quite similar to each other and the D gene is inside the CDR3 region, which is hyper variable. Once we have successfully aligned the V and J genes, we remove the region that is aligned with them. The remaining sequence contains only the D gene. We

choose the last three nucleotides from the 3' end and 5' end of each reference D gene as our detection tags. Again we use the detection tags to find a perfect match and extend to get a matching score. Subsequently we select the D gene, which has the highest matching score and passes the minimum matching threshold (90% match and 10bp matches), as our aligned D gene segment.

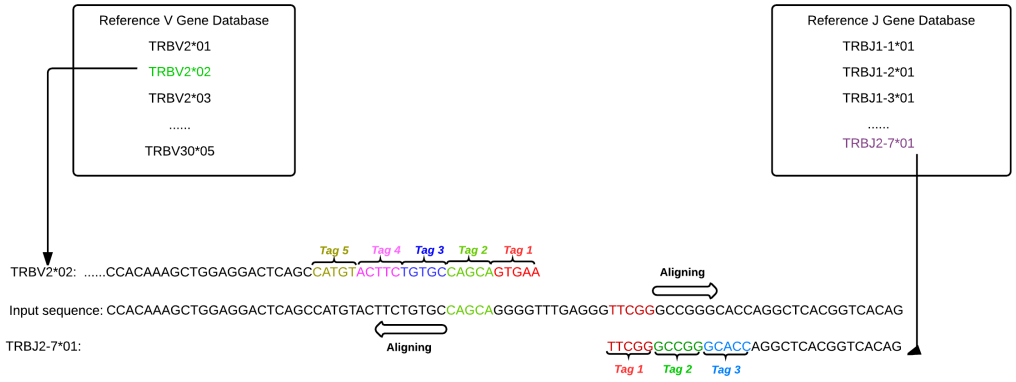


*Figure 2.2. Alignment algorithm: Reference V and J genes are shown in the boxes on the right and left. The 5 detection tags of one of the reference V genes (TRBV2*02) are labelled with different colours within the V gene sequence (left). Tag 1 (red) fails to achieve an exact match with the input sequence but an exact match with tag 2 (green) can be used to extend an alignment in both directions. Similarly exact matching of tag 1 (red) seeds the alignment of J sequence TRBJ2-7*01 (right) to the input.*

2.3.5 CDR3 extraction and classification

Once the reference VDJ genes of each input sequence have been determined, we extract the CDR3 sequence and classify the input sequences by their CDR3. Input sequences are in the same CDR3 class if they are mapped to the same V(D)J genes and have identical CDR3 region nucleotide sequences. CDR3 classification takes place in two stages: We firstly perform preliminary classification based on exact matching of the extracted CDR3 sequence and count the number of each clonotype (CDR3 classes). This results in large numbers of singleton clonetypes and clonetypes that have small numbers of copies. CDR3 sequences with counts below an adjustable portion (default = 0.001%) of the sequences are labelled as “minimum sequences”, with the rest of the sequences labelled as “core sequences”. For each of the minimum

sequences, we calculate its Hamming distance to the core sequences of the same length. If the Hamming distance is less than M steps (M is an adjustable parameter, default = 2), we merge the minimum sequence with the corresponding core sequence. This process is repeated by iterating over the minimum sequences.

2.3.6 SNP calling

After the input sequence has been aligned with the corresponding reference genes, it is straightforward to locate nucleotides that do not match the reference sequence. These are considered as potential SNPs in the input sequences. In order to avoid treating PCR errors as potential polymorphisms we use two criteria described by Schott et al [221]. The first is that the same gene variant should occur in multiple V(D)J combinations. For instance, when we are searching for V gene SNPs we require the potential non-reference allele on the V gene to associate with more than three different J genes. As we have more V genes than J genes, the minimum number of different V genes required to define a potential SNP on the J gene is five. The second criterion to identify a candidate SNP is that the non-reference allele should occur at a frequency of at least 10% among the sequences of the corresponding gene. This is informed by the assumption that somatic point mutations should occur in fewer than 5% of all the sequences, unless they are within the G/C mutation hot spot, in which case they can reach a frequency of 10% [221].

In order to efficiently store and index the potential SNPs, we use a nested Hashmap structure (figure 2.3). For each of the mutations found in a given reference gene, we calculate the percentage of mutations and the number of different gene types associated with them. If the two criteria mentioned above are met we store it in the potential SNP dataset for downstream analysis.

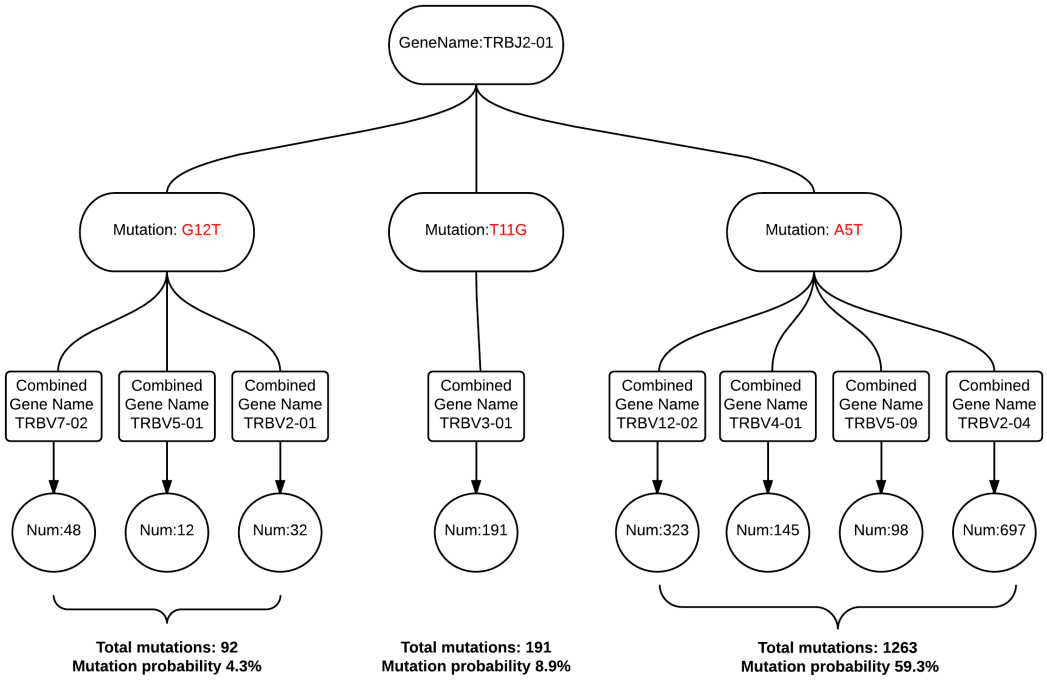


Figure 2.3. Data structure used in SNP analysis. For each of the reference genes, we have a Hashmap storing the mutation information. For reference J gene TRBJ2-01, three mutations are detected (shown in red). For example, G12T means Glycine at position 12, mutated to Tyrosine. This mutation is found in combination with three different V segments with a combined frequency of 4.3%.

2.3.7 Mutation tree construction

LymAnalyzer creates lineage mutation trees for immunoglobulins. The lineage mutation tree construction algorithm used in LymAnalyzer is based on the modification of the distance method concept exploited by Barak et al [234]. As noted by the authors, this method does not aim to simulate the particular mutation process that occurred. Instead it aims to reveal the minimal steps that could have led to the observed sequences. We firstly define the root sequences, which are those sequences with the original germline configuration. For each of the root sequences, we find the sequences that are within ten Levenshtein steps (the minimum steps required to change one string to another only using insertions, deletions or substitutions). Each layer of the tree is created according to the distance to the root node.

2.3.8 Statistical test

We used two approaches to test the statistical significance of differences in proportions of mapped reads. Treating individual sequence reads as the statistical unit, the equality of the proportion of mapped reads (and, in the case of the simulated data, the proportion of correctly mapped reads) between two methods was tested using the chi-square test. In the case of the real data it may be more appropriate to treat samples as the statistical unit because there may be differences between samples that affect the performance of different methods (e.g. different levels of mismatch with the reference genes). Therefore, we also used the Wilcoxon signed-rank test to perform a paired comparison of the median proportions of reads from the biological samples mapped by each method. The significance threshold for both tests was set at 0.01.

2.3.9 Implementation and software resources

The algorithms, command line console and graphic user interface of LymAnalyzer were implemented in Java 1.8 by using Eclipse (4.5.0). IMGT database was used as the reference database and test sample datasets were acquired from NCBI database.

2.3.10 Authorship contribution statement

YY, RhC and CS carried out the conception and the design of the study. YY developed the algorithm, implemented the software packages, performed the analysis and drafted the manuscript.

2.4 RESULTS

2.4.1 Accurate CDR3 extraction and VDJ identification

LymAnalyzer was first applied to a dataset in the public domain (SRA: PRJNA229070), consisting of short read TCR sequences from nine samples. LymAnalyzer consistently mapped a significantly higher proportion of the reads (figure 2.4A), compared to MiTCR, MiXCR and Decombinator ($p < 0.01$ in all cases; See Materials and Methods for details of statistical tests). The decline in the proportion of extracted reads from all three tools from sample

SRR103674 to sample SRR1033679 was due to differences in sequencing quality. We also compared the performance of LymAnalyzer with MiXCR on IG sequences using a publicly available dataset (SRA: SRP017087). Because many of the reads in this dataset do not cover the CDR3 region, a large proportion of the reads remained unmapped by both tools; however, LymAnalyzer mapped a larger portion of reads compared to MiXCR in all cases (figure 2.4B)($p < 0.01$).

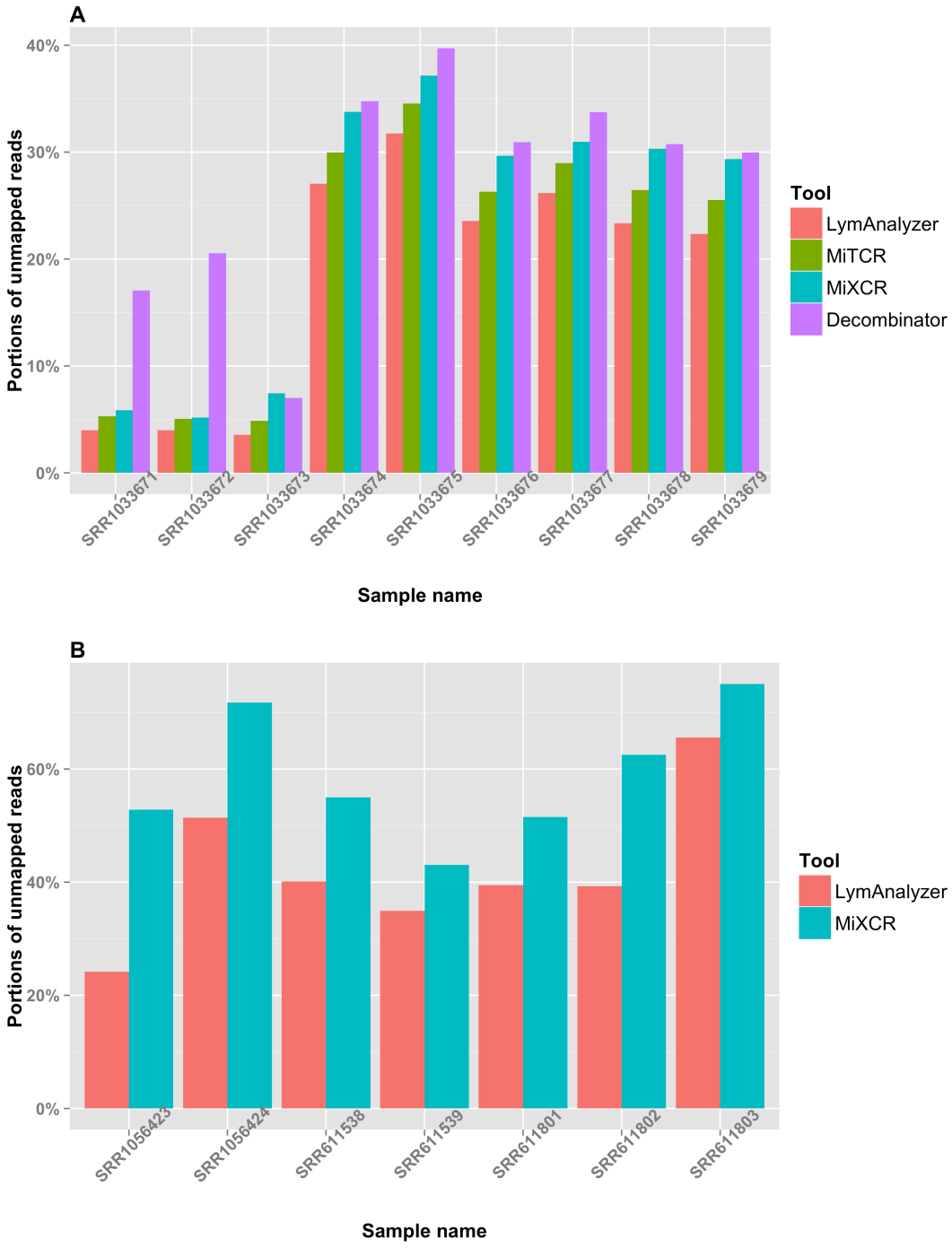


Figure 2.4. Results based on real dataset. (A) The comparison of mapping completeness of TCR data among LymAnalyzer, MiTCR, MiXCR and Decombinator. LymAnalyzer outperformed the other tools in all nine sample based on the completeness of the alignment. (B) The comparison of mapping completeness of IG data between LymAnalyzer and MiXCR. LymAnalyzer consistently mapped larger proportions of reads compared to MiXCR in all the samples.

We used simulated datasets to investigate the accuracy and completeness of the results generated by LymAnalyzer. Each simulation consisted of VDJ recombination together with different mismatch levels. For simulated TCR data, in the absence of mismatches, LymAnalyzer can map all of the sequences, whereas MiTCR only mapped 91% of the sequences (figure 2.5A). As the mismatch level increased, the number of reads that MiTCR and LymAnalyzer mapped declined gradually, as expected; however, LymAnalyzer still mapped a greater proportion of the reads than MiTCR. In addition to mapping a greater proportion of the reads, LymAnalyzer was also significantly more accurate than other methods, with 99.25% of the sequences mapped correctly, compared to 91.45% correctly mapped sequences with MiTCR in the absence of mismatches (figure 2.5B).

We also compared the results from Decombinator and MiXCR based on simulated data (Supplementary Figure A1). However, Decombinator and MiXCR can only give us the particular gene name of each sequence, instead of allele name, which is the standard output of MiTCR and LymAnalyzer. Therefore we only compared the results under gene name level. Under the high mutation level, Decombinator missed more than one third of the sequences. Therefore we compared the performance of LymAnalyzer, MiTCR and MiXCR separately (figure 2.6). At the gene name level, both LymAnalyzer and MiTCR had increased accuracy as expected. MiTCR had higher accuracy (98.1%) comparing to MiXCR (96.4%) in the no mismatch dataset. However, as the mismatch level increased, MiXCR achieved higher accuracy than MiTCR. LymAnalyzer still achieved the highest accuracy and completeness among the three tools at all mismatch levels. For simulated IG data, we

compared LymAnalyzer with MiXCR since they are, to date, the only tools that can process IG NGS data (figure 2.7). LymAnalyzer showed both improved accuracy and completeness relative to MiXCR. In terms of accuracy, in the absence of mismatches, MiXCR mapped more than 20% of the sequences incorrectly, compared to 0.7% mapping errors in LymAnalzyer. Under the high mismatch level, LymAnalyzer retained accuracy above 95%, while the accuracy of MiXCR declined to 75.9%.

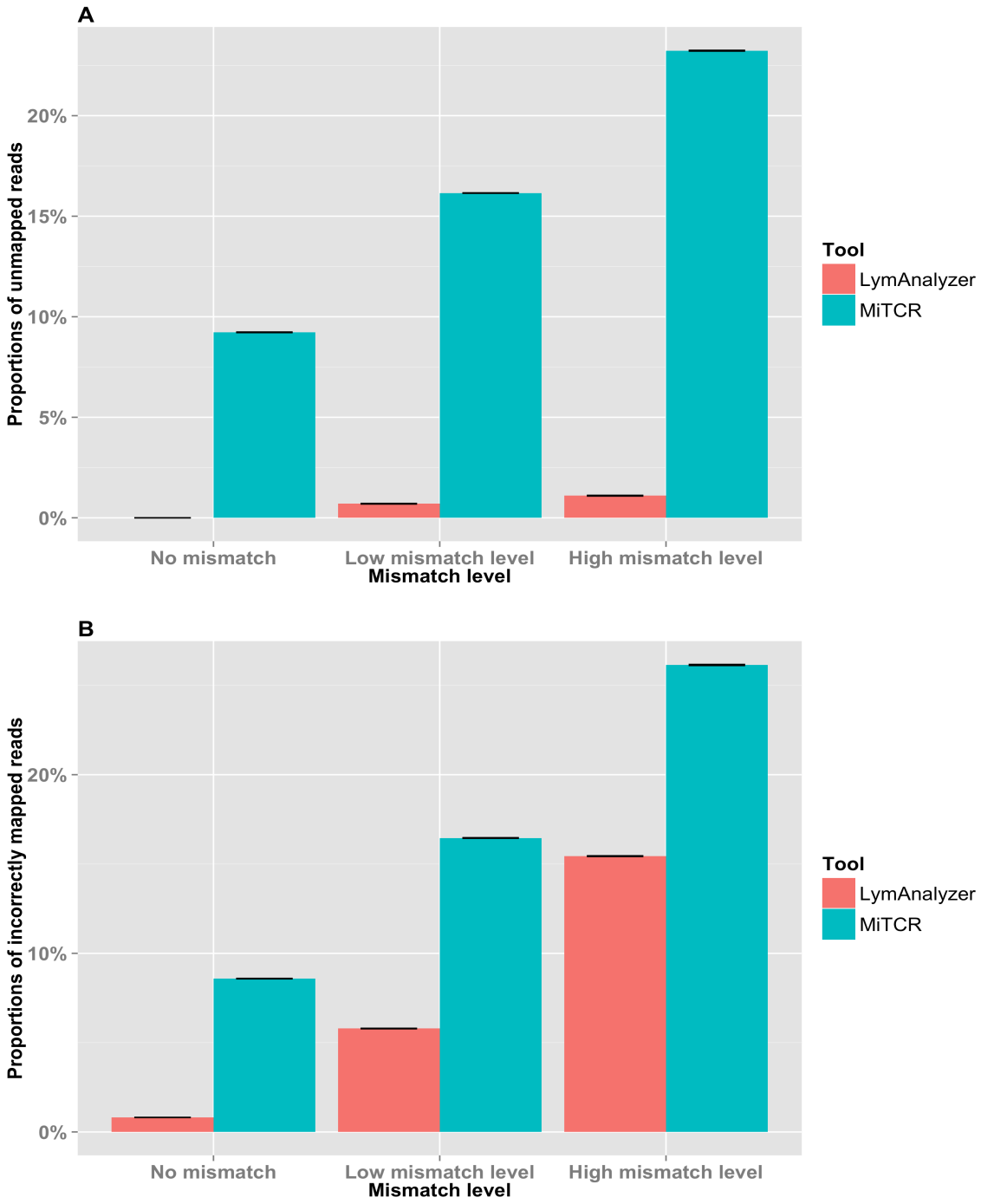


Figure 2.5. Results based on simulated TCR data on the allele name level. (A) Comparison of completeness of the results from LymAnalyzer and MiTCR. (B) Comparison of the accuracy of LymAnalyzer and MiTCR. The completeness and accuracy values shown are the means of the results from twenty simulated samples. The error bars shown at the top of each bar indicate the standard error of the mean of the simulated datasets.

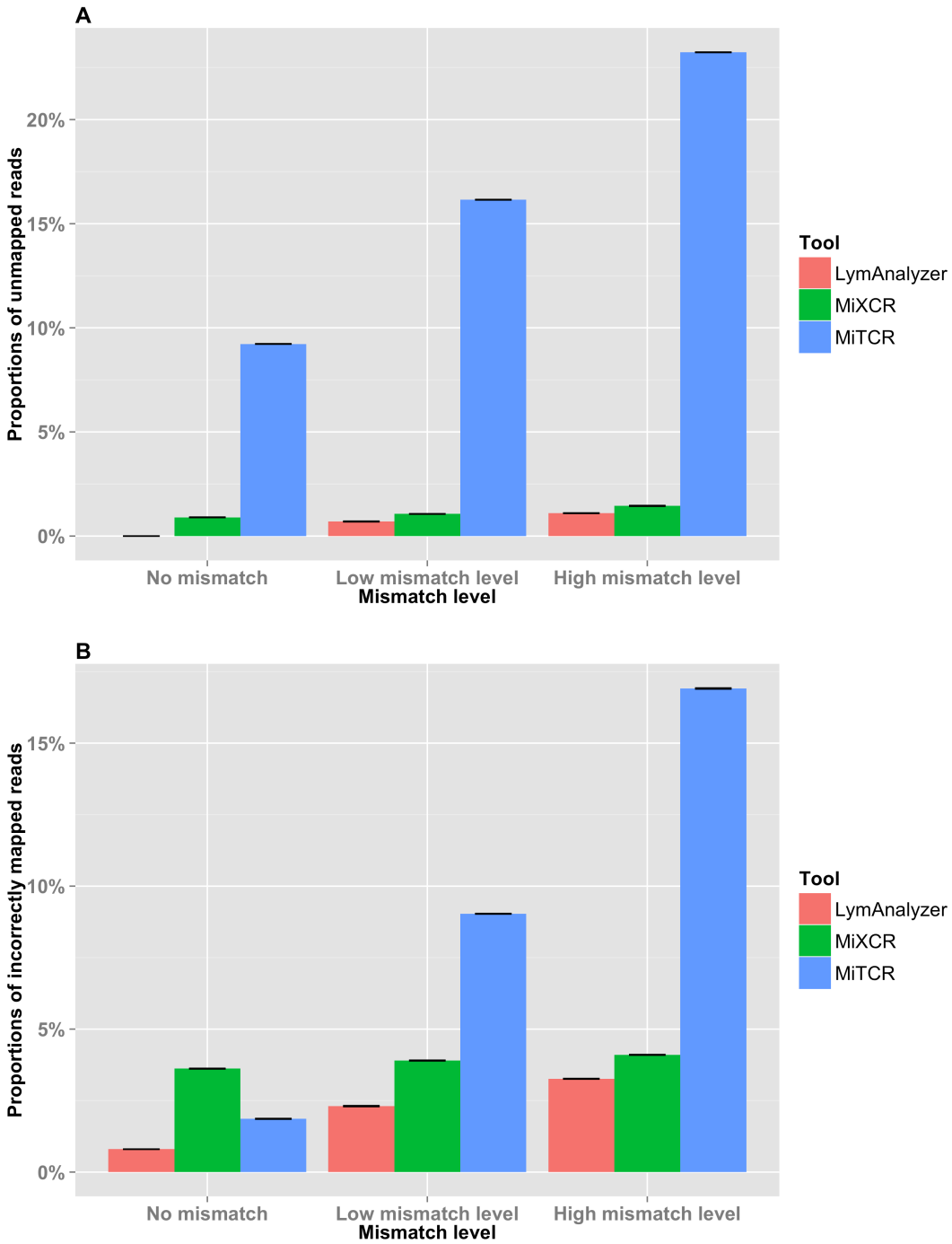


Figure 2.6. Results based on simulated TCR data at the gene name level. (A) Comparison of the completeness of the results from LymAnalyzer, MiTCR

and MiXCR. LymAnalyzer had the highest completeness among the three tools at all mismatch levels. The completeness decreased with increasing mismatch level; however LymAnalyzer retained above 98% completeness even at the highest mismatch level. (B) Comparison of the accuracy of LymAnalyzer, MiTCR and MiXCR at different mismatch levels; LymAnalyzer outperformed MiTCR and MiXCR in terms of accuracy.

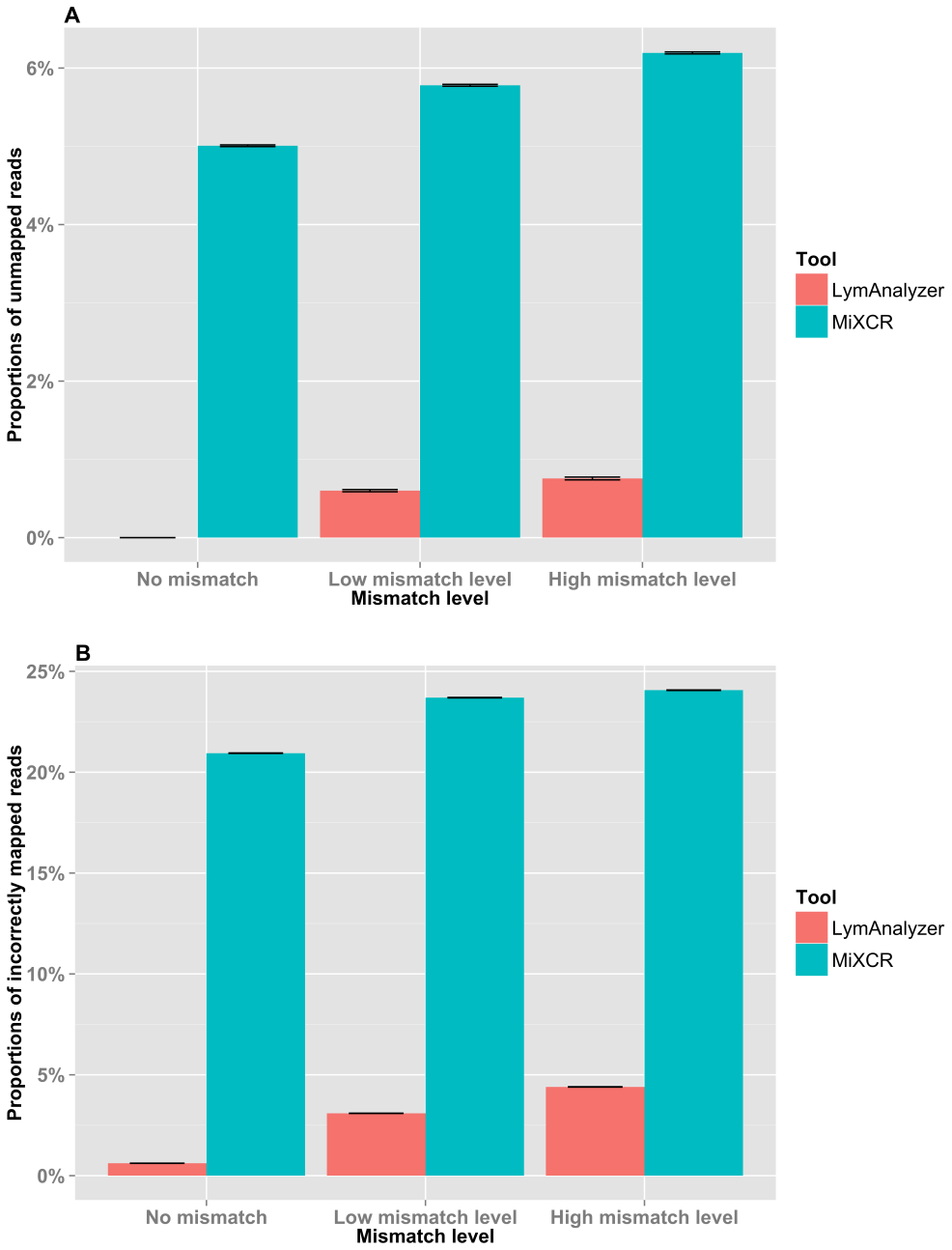


Figure 2.7. Results based on simulated IG data at the gene name level. (A) The comparison of completeness of the results from LymAnalyzer and MiXCR. (B) The comparison of the accuracy of LymAnalyzer and MiXCR. LymAnalyzer had both significantly improved accuracy and completeness compared to MiXCR.

2.4.2 Running Time

LymAnalyzer runs on Windows, Linux and Mac OS X. We tested the running performance of LymAnalyzer on both a Linux cluster and a personal computer (MacBook). On a MacBook, for 125,000 to 200,000 sequences, the full analysis can be finished in 9-12 seconds. For large datasets, with 10-15 million sequences, the full analysis can be accomplished in 25-40 minutes on our cluster server (Hardware configuration: 6-core 2.2Ghz AMD Opteron Processor 2427 with 32GB memory). As can be seen from the plot in the supplementary figure A2, the running time scales linearly with the number of reads. The estimated processing speed of LymAnalyzer for sequences of 100bp long is 8461 reads per second. (Table 2.1)

Table 2.1. Feature comparisons of different TCR/IG sequencing analysis tools.

	IgBlast	IMGT/High-V-Quest	iHMMune-align	Decombinator	MiTCR	MiXCR	LymAnalyzer
CDR3 Extraction	√	√	×	√	√	√	√
V, D and J gene alignment	√	√	√	√	√	√	√
Large NGS data (>1 million sequences)	×	×	×	√ (5.7K reads/second)	√ (14.6K reads/second)	√ (13.9K reads/second)	√ (8.5K reads/second)
TCR	√	√	×	√	√	√	√
Immunoglobulins	√	√	√	×	×	√	√
Polymorphism Analysis	×	×	×	×	×	×	√
IG mutation Anlysis	×	×	×	×	×	×	√

2.4.3 Additional features of LymAnalyzer

In addition to features that are common to existing tools, LymAnalyzer can perform polymorphism analysis and generate hypermutation trees for IG sequences. LymAnalyzer provides both command line and GUI version and is

implemented in JAVA for cross-platform application. A comparison of features available in different tools is provided in Table 2.1.

In order to test if our SNP calling algorithm is capable of recognizing potential unreported alleles, we manually modified the reference gene database, retaining just one allele of each distinct V gene. We then used LymAnalyzer to map a subset of 125,000 reads from sample SRR1033674 to this modified reference gene database. From the result file, we found seven putative SNPs (Table 2.2). By mapping these suspected SNPs back to the original reference gene database, we found that all but one of them can be accounted for by the known alleles that were removed from our reference gene database at the beginning of the test (Table 2.2). There was one variant of TRBV29-1*03, corresponding to a substitution from A to C, that could not be mapped to an existing allele in the IMGT database. Given the high frequency (45.44%) and read count (4,488) for this mutation, it could correspond to an allele that is not found in the database. In support of this hypothesis we found that there is a known A/C SNP (rs17214) at the genomic position corresponding to this mutation in dbSNP [133]. This variant leads to an amino acid change (Methionine to Leucine) on TRVB29-1.

Table 2.2. Suspected SNPs and their true allele on the V genes.

Mapped allele	Suspected SNP	Allele frequency	Allele counts	IMGT allele
TRBV7-9*07	G167C	98.68%	4872	TRBV7-9*03
TRBV15*01	A257G	97.77%	1931	TRBV15*02
TRBV11-2*03	A240G	38.64%	807	TRBV11-2*01
TRBV9-2*01	A256G	97.33%	474	TRBV9-2*02
TRBV9-2*01	G278T	96.51%	470	TRBV9-2*02
TRBV6-6*05	G247C	97.87%	2123	TRBV6-6*02
TRBV29-1*03	A191C	45.44%	4488	Not found

Mutation trees are generated in Newick format and can be visualized using several existing software packages (figure 2.8) [235]. The tree does not necessarily represent the real mutation process that took place; it shows the minimal steps that can explain the observed sequences. Adjacent layers are

separated by a Levenshtein distance of one, which represents one nucleotide change. Each of the nodes in a given layer is one step away from the nodes to which it is linked in the previous and subsequent layers. However it is not guaranteed that there is always a parent node that is one Levenshtein distance away from the current node. Therefore we create a hypothetical node in each layer of the tree. The hypothetical node is not a real sequence that exists in the dataset, but instead represents the collection of unobserved intermediate sequences between two nodes that are separated by a Levenshtein distance greater than one.

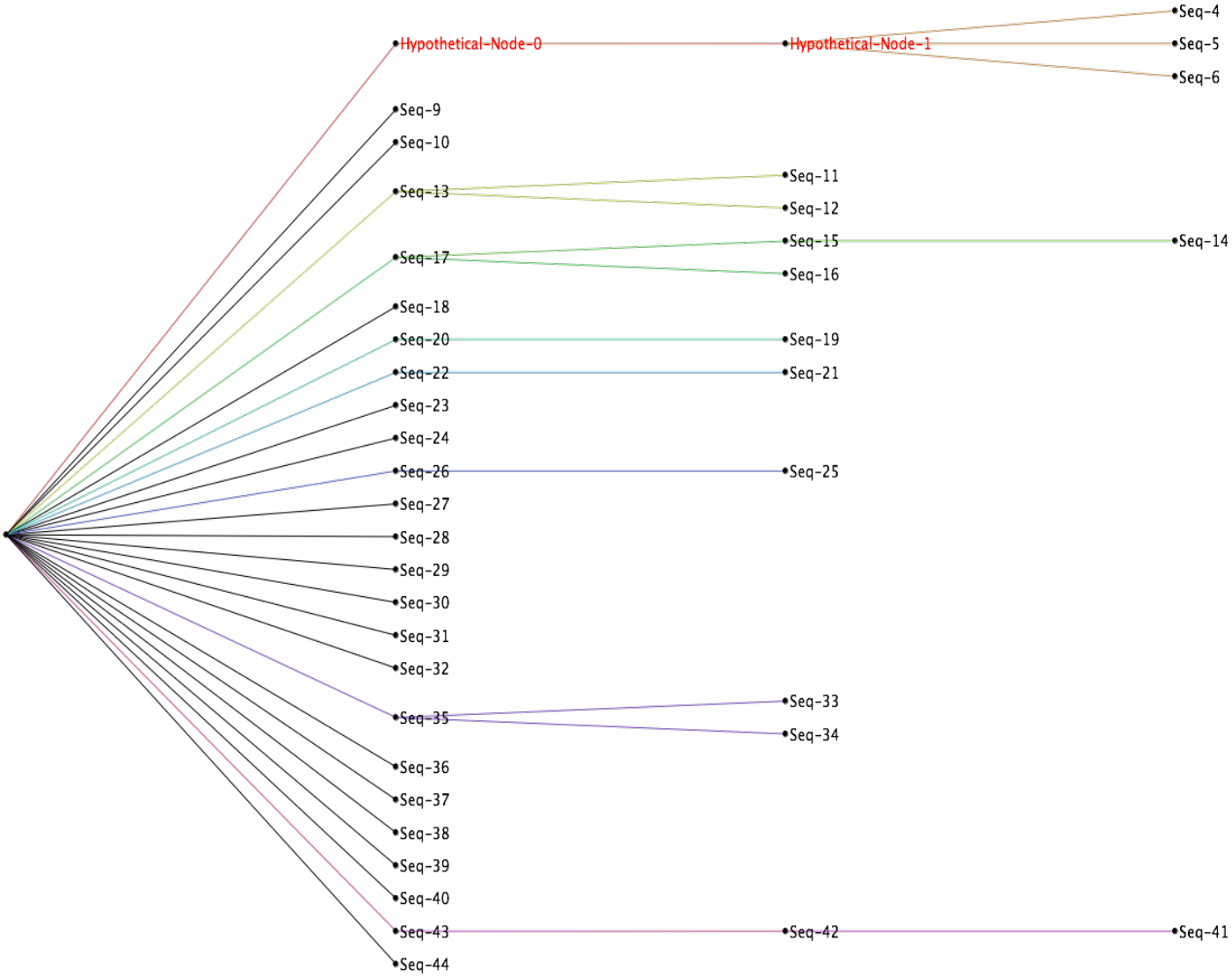


Figure 2.8. Example of a mutation tree generated by LymAnalyzer. (visualized using FigTree) Each node represents an individual clone. The nodes on the same level are one Levenshtein step (nucleotide change) away from their corresponding nodes on the previous and subsequent layers. The

hypothetical nodes are shown in red. These are required to connect the nodes that are more than one step away from the closest observed sequence.

2.5 DISCUSSION

Next generation sequencing technology gives researchers an opportunity to study lymphocyte repertoire diversity at high resolution. However current bioinformatics pipelines for identification and annotation of large TCR/IG sequence datasets are unsatisfactory due to their suboptimal accuracy and completeness. Here we present LymAnalyzer, a software package for comprehensive analysis of TCR/IG sequence data.

LymAnalyzer consists of four functional components: VDJ gene identification followed by CDR3 extraction, SNP calling and lineage mutation tree generation. We performed multiple tests of accuracy using publicly available and simulated datasets and compared the performance of LymAnalyzer to existing tools (MiTCR, MiXCR and Decombinator). In our evaluation using real data, LymAnalyzer mapped more reads than the other tools. In terms of accuracy, we have shown using simulated data that LymAnalyzer provides significantly improved mapping accuracy compared to MiTCR, MiXCR and Decombinator. MiTCR had the fastest running performance among the tools; however, it trades accuracy and completeness for speed. Given that TCR/IG sequencing datasets are tractable on a personal computer for typical datasets or on computer clusters for large projects this trade off is unnecessary. Despite significant improvement in accuracy and completeness, the running time of LymAnalyzer is better than Decombinator and remains comparable to MiTCR and MiXCR.

The majority of lymphocyte sequence analysis tools can only process TCRs. LymAnalyzer makes the analysis of IGs also available. Furthermore, LymAnalyzer is to date the only tool that can generate mutation trees for IGs. Another novel feature of LymAnalyzer is the ability to detect the SNPs. We tested the reliability of this function by running LymAnalyzer on the same dataset again with the reference gene database that only kept one representative allele for each gene and compared the results from both runs.

Indeed, LymAnalyzer revealed the SNPs, which can also be found in the removed alleles. However, the accuracy of the SNP detection can be hampered by allelic imbalance. We only report the suspected SNPs that exceed particular mutation rate threshold (10%), and we may miss some imbalanced alleles that are lower than this threshold. Therefore, we set this threshold as an adjustable parameter. Users can change this threshold value based on their requirements. Previous studies have shown that the IMGT database appears to be incomplete, as many reported IG heavy chain variable alleles are not found in the database [236-239]. Moreover, many IG heavy chain variable alleles polymorphisms may have been reported in error [220]. The unreported SNP on TRBV29-1 found in our study shows that there are also TCR beta chain variable alleles missing from the IMGT database. A more updated and robust reference gene database for TCR/IG sequences is required. By taking advantage of the increased availability of TCR/IG sequence datasets, the SNP detection function implemented in LymAnalyzer could help to discover novel alleles and improve the coverage of the TCR/IG reference gene database.

Chapter 3 – A database of human immune receptor alleles recovered from population sequencing data

The content of this chapter was published as:

Yu, Y., Ceredig, R. and Seoighe, C., 2017. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. The Journal of Immunology, p.1601710.

3.1 ABSTRACT

High throughput sequencing data from T cell receptors (TCRs) and immunoglobulins (IGs) can provide valuable insights into the adaptive immune response, but bioinformatics pipelines for analysis of these data are constrained by the availability of accurate and comprehensive repositories of TCR and IG alleles. We have created an analytical pipeline to recover immune receptor alleles from genome sequencing data. Applying this pipeline to data from the 1000 Genomes Project we have created Lym1K, a collection of immune receptor alleles that combines known, well-supported alleles with novel alleles found in the 1000 Genomes Project data. We show that Lym1K leads to a significant improvement in the alignment of short read sequences from immune receptors and that the addition of novel alleles discovered from genome sequence data is likely to be particularly significant for comprehensive analysis of populations that are not currently well represented in existing repositories of immune alleles.

3.2 INTRODUCTION

The adaptive immune system provides protection from disease by initiating specific immune responses to specific foreign antigens following their recognition by the clonally-distributed surface bound receptors of T and B lymphocytes, namely T cell receptors (TCRs) for T cells and immunoglobulins

(IG) for B cells. In its secreted form, the B cell's IG molecule is called an antibody. The variability in molecular shapes of foreign antigens represents a significant challenge for the immune system, requiring development of an extensive receptor repertoire for antigen recognition. In response to this challenge, TCRs and IGs achieve vast diversity by a unique mechanism called VDJ recombination, which is a largely stochastic process of gene rearrangement encoding the variable parts of TCRs and IGs respectively. In addition, genes encoding IGs undergo somatic hypermutation during T cell-dependent B cell responses, resulting in further expansion in antibody diversity and subsequent selection of B cell clones generating IG's with higher affinity for antigen. Quantitative analysis of the diversity of TCR and IG repertoires can shed light on receptor-antigen interactions and reveal insights into the state of immune system under diverse conditions, such as aging [240-242], chronic viral infections [243-245] and autoimmune diseases [246-249].

Next generation sequencing (NGS) technology revolutionized the analysis of TCR/IG diversity by providing vast amounts of data at much higher resolution, leading also to a need for effective and robust bioinformatics pipelines. Generally, these pipelines align the input sequences with the reference V, D (for IG heavy chain and TCR beta chain) and J genes, determine the CDR3 region and estimate the repertoire diversity. Two main factors that directly influence the quality of these analyses are the alignment algorithm and the completeness and accuracy of the reference database. In the past three years, several software packages have been developed with different alignment algorithms for analyzing NGS sequence data of TCR and (or) IG [208-219]. These studies all focused on improving the accuracy, effectiveness and completeness of their alignment algorithms. Most of the software packages use the reference V(D)J sequences provided by International ImMunoGeneTics Database (IMGT) [78], which is currently the most comprehensive collection of TCR and IG germline gene sequences available.

However, previous studies have reported numerous novel V alleles that are not found in the IMGT database [236-239], suggesting that this database is incomplete. In addition, a study of 226 IGHV alleles in IMGT reported a high

rate of false positives (of 226 alleles 104 were classified as very likely to contain sequencing errors) [220]. This is because most of the IMGT TRBV and IGHV alleles were included from early studies, conducted between 1984 and 1996 [250, 251]. After that, there were only two major updates on IGHV alleles in 2002 and 2009 introducing in total 36 alleles, and no updates on TRBV alleles submitted in the past twenty years. Furthermore, many alleles were reported in only one study, and multiple allelic variants of the same gene were derived from the same individual. For instance, an individual can have at most two alleles of any gene; however, seven IGHV3-15 alleles were derived from one individual. This was pointed out by Wang et al. [220] but the issue remains unresolved in the current version of IMGT. In addition, many TRBV and TRAV alleles in IMGT are annotated as having come from rearranged sequences, which may contain somatic mutations and partially truncated 3' ends resulting from VDJ recombination.

Since the development of IMGT, a large and ever-increasing number of full human genomes have been sequenced. The chromosomal regions containing the immune receptor alleles have both a high level of intra-chromosomal duplication and a high level of diversity and thus present challenges for genome assembly. However, improvement of the reference genome over time has the potential to enable recovery of a comprehensive set of immune gene alleles found in human populations from population resequencing studies. We have developed a bioinformatics pipeline that can be used to generate a TCR/IG reference database using data from population resequencing studies. We applied this pipeline to data from the 1000 Genomes Project (G1K) [143], which contains the most comprehensive collection of global human variation currently available in the public domain, to create the Lym1K database. We aligned short read sequence data from human immune receptors to Lym1K and obtained significantly improved alignments compared to what could be achieved using the IMGT database as reference. We also report evidence that the diversity of immune receptor alleles observed in non-African populations is particularly under-represented in the existing database.

3.3 MATERIALS AND METHODS

3.3.1. Immune receptor sequencing dataset

We obtained two datasets from the NCBI Sequence read archive (SRA). The first dataset was for the TCR beta chain (SRA index: PRJNA310731) and consisted of 34 samples. The number of reads ranged from 4,571,596 to 13,742,272. Each read was 200bp long and included part of the V, all the D and part of the J region. The second dataset contained both IG heavy chain and IG light chain sequences (SRA index: PRJNA179099). We used four samples from this dataset (all samples that included more than one million reads). The number of reads in the samples ranged from 1,046,575 to 23,191,224. Each read was 250bp long and included part of V (all the D for the heavy chain) and part of the J region.

3.3.2 Simulated dataset

Each simulated sequence was created based on the process of VDJ recombination of TCR and IG and included mismatches to stimulate the combined effects of sequencing errors and mutations/polymorphisms. First, we randomly chose a V, D (only for the heavy chain of IGs and the beta chain of TCRs) and J gene assuming uniform gene usage. We then decided the particular allele of the chosen V gene based on the frequency of its occurrence in the chosen population. Secondly, we "mutated" the chosen V and J segments by introducing 0-10 mismatch(es) on the V region and 0-5 mismatch(es) on the J region. To simulate junction diversity, we inserted 0-6 randomly generated nucleotides into the V-D and D-J junction (V-J junction only for IG light chain and TCR alpha chain).

We applied our simulation pipeline for TCRB, TCRA, IGH and IGL genes separately. In all we produced 27 datasets, with separate datasets corresponding to each of the 26 populations included in G1K and one dataset derived from the pool of all populations. Each dataset consisted of twenty simulated samples, and each sample contained 200,000 simulated sequences.

3.3.3 1000 Genomes Project data

We retrieved VCF files corresponding to the TCRB, TCRA, IGH and IGL regions from phase 3 of the 1000 Genomes Project. We used the R package SNPRelate [252] to perform principal component analysis separately on variant data from each of the four regions.

3.3.4 AlleleMiner pipeline

The goal of AlleleMiner is to infer all the possible alleles of the target genes from the input VCF files. For the Lym1K database, we first retrieved the location information of TCR and IG genes from BioMart (GRCH38.p2) and phased VCF files from G1K project (GRCH38; Phase 3). Note that the current phase 3 VCF files from G1K project were generated based on the GRCH37 human genome assembly; however, VCF files in the coordinates of the GRCH38 assembly are also provided.

For each gene, AlleleMiner extracts the corresponding reference genome sequence from the UCSC database and SNPs from the input VCF file. Subsequently, AlleleMiner retrieves all alleles of each gene from the haplotypes spanning the gene of interest. Identical haplotypes are merged, and the number of times a particular haplotype appears is counted. Users can define a threshold for the haplotype occurrence times to eliminate rare haplotypes, some of which may be the result of sequencing error; only the haplotypes that occur more than the threshold number of times are stored as potential alleles in the new TCR/IG database.

3.3.5 Implementation and software resource

The pipelines were implemented in Java 1.8 by using Eclipse (4.5.0) and are freely available at https://sourceforge.net/projects/alleleminer/. The Lym1K database of immune receptor alleles can be obtained from http://maths.nuigalway.ie/biocluster/database/.

3.3.6 Authorship contribution statement

YY, RhC and CS carried out the conception and the design of the study. YY developed the algorithm, implemented the software packages, performed the analysis and drafted the manuscript.

3.4 RESULTS

3.4.1 Construction of TCR/IG reference database and performance comparison

The sequencing of increasing numbers of whole human genomes and improvements in the reference human genome assembly give rise to opportunities to expand substantially the reference database of human immune receptor alleles and to profile the diversity of immune receptor alleles across global populations. Currently, the 1000 Genomes Project represents the most extensive collection of whole genomes sampled from global human populations in the public domain. Focusing specifically on TCRs and IGs, the majority of human immune receptor genes have been sequenced to sufficiently high coverage to recover novel alleles, the median coverage of IGHV, IGLV, TRBV, TRAV are accordingly 121, 120.5, 76 and 53 (Supplementary Figure B1). We retrieved known and novel TCR/IG alleles from 2504 human individuals in the phase three cohort of the 1000 Genomes Project. We then tested the performance of the reference database we derived from population resequencing data and compared it to the existing IMGT reference database of human immune receptors using simulated data as well as short read receptor sequencing datasets (figure 3.1).

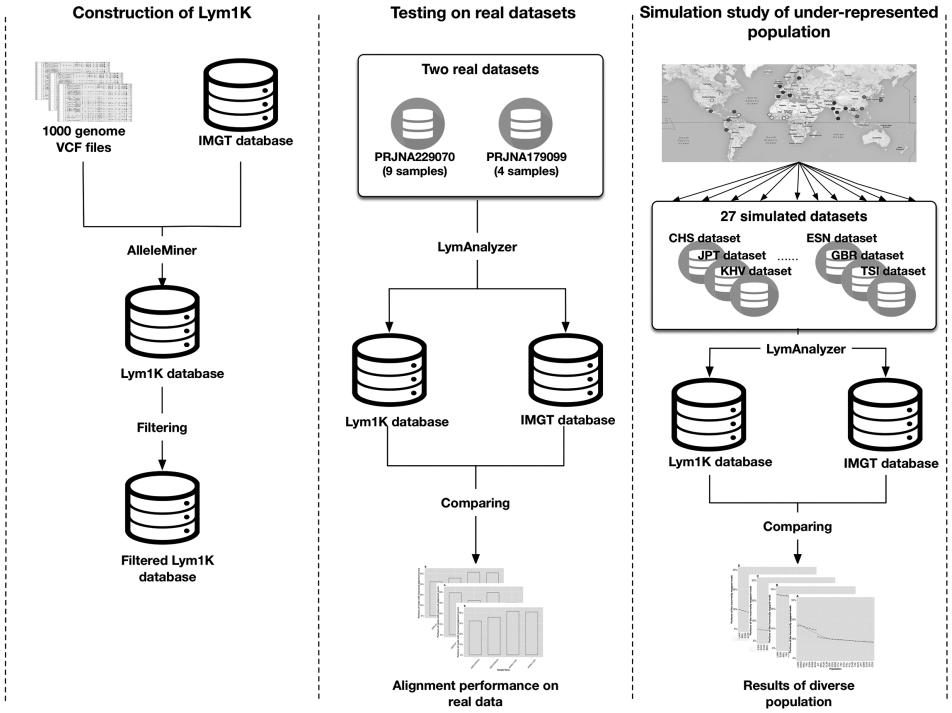


Figure 3.1. Study outline. AlleleMiner retrieves all immune receptor haplotypes in data from the 1000 Genomes Project, merging it with alleles from IMGT to produce the Lym1K reference database. We compared the performance of the Lym1K and IMGT reference datasets using real IG and TCR short read sequence data. Finally, we carried out a simulation study to investigate the extent of under-representation of alleles found in diverse human populations in IMGT.

3.4.2 AlleleMiner workflow

AlleleMiner is the core component of the pipeline and is used to infer immune receptor alleles from genomic data. The algorithm used in AlleleMiner is described in detail in the Experimental Procedures. In brief, AlleleMiner first enumerates all possible haplotypes of each TCR/IG gene, using as input genome-wide sequence variants in variant call format (VCF). Subsequently, the inferred haplotypes that occur in more than a user-defined minimum number of individuals are merged with known alleles (here we merged discovered variants with alleles in the IMGT database to yield the Lym1K

database of known and inferred immune receptor alleles). A filtering step is then applied to assess the support for each non-reference SNP allele found among the observed haplotypes (figure 3.2).

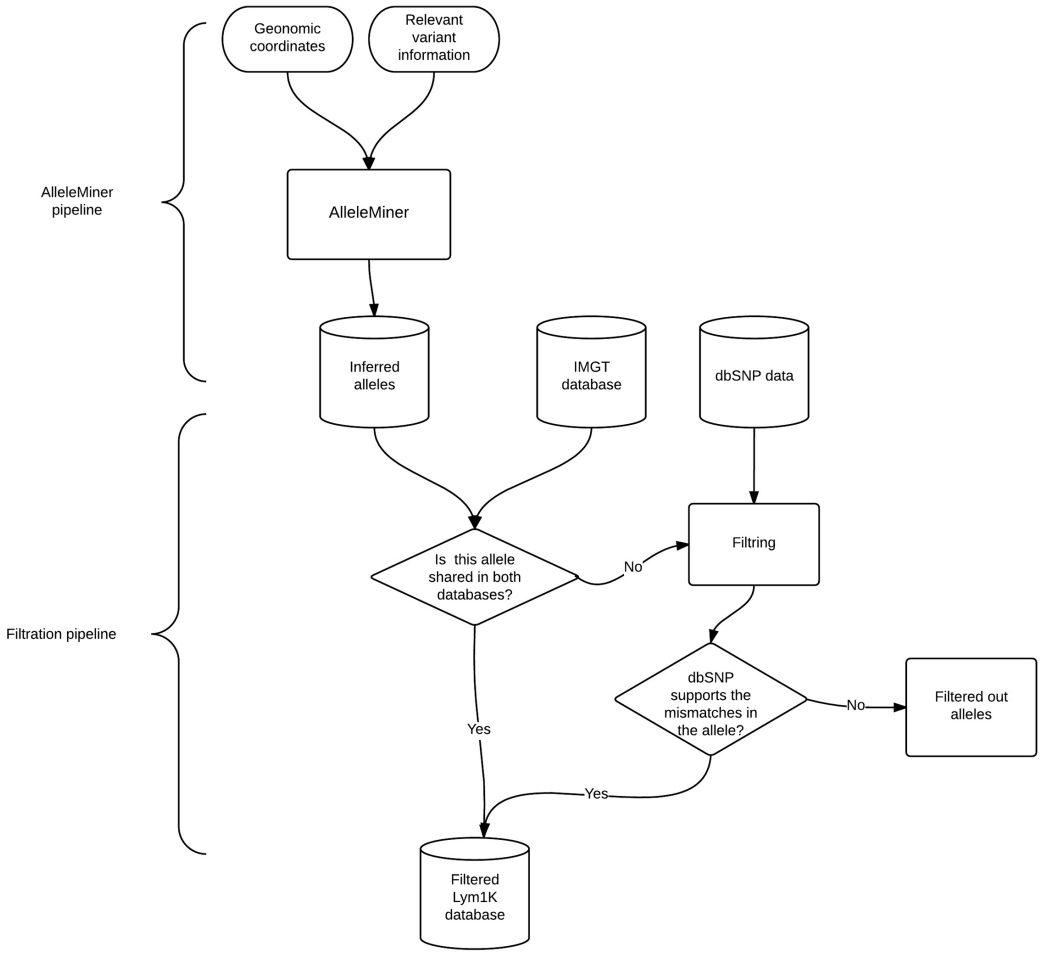


Figure 3.2. Workflow for database construction. AlleleMiner is used to infer all the possible alleles from VCF data. The Lym1K database is constructed by merging the inferred alleles with the original IMGT alleles. For alleles not found in IMGT we apply a filtering step that takes account of the validation status of all non-reference SNP alleles contained on the haplotypes corresponding to the putative novel allele.

3.4.3 Performance assessment

To assess the extent to which known immune receptor alleles could be recovered from population resequencing data we set the initial inclusion threshold to one (i.e. we considered all alleles found at least once in the G1K data). For IGL, TRA and TRB the majority of the alleles in IMGT could be

recovered (Table 3.1 and Supplementary Table B1). We then examined the alleles that were not recovered. In the case of TRAV and TRBV, many of these alleles were annotated as originating from rearranged genomic DNA or cDNA, rather than originating from unrearranged germline DNA. Rearranged sequences may have additional somatic variants that we do not expect to find among the alleles we recovered from the genomic sequence. Many of the remaining alleles are from the 22 TRBV genes that are not included in either the GRCh37 or GRCh38 human genome assemblies (GRCh37 was used in the G1K project and the current human genome assembly, GRCh38, is used as a reference genome by AlleleMiner). Excluding these genes and the alleles corresponding to rearranged sequences, we were able to retrieve all of the remaining TRAV and all but one allele (TRBV19*02) of the remaining TRBV alleles in IMGT (Table 3.1).

By contrast, although a majority of IGLV alleles were recovered, fewer than half of the IGHV and IGHJ alleles from IMGT were found among the inferred alleles (Table 3.1). This could reflect the greater diversity of IG genes (Supplementary Figure B2) or the presence of false positive alleles in IMGT, particularly in IGHV genes. Indeed, a high rate of false positive IGHV alleles in IMGT has previously been reported [220]. We found evidence for both of these effects. The greater diversity of IG genes is apparent from the relationship between the number of inferred alleles and the inclusion threshold used in AlleleMiner (Supplementary Figure B3A). The number of shared IGHV alleles dropped significantly more rapidly than the other three gene groups, suggesting that IMGT includes IGHV alleles that are found at relatively low frequencies in the G1K populations (Supplementary Figure B3B). To explore the impact of false positive alleles in IMGT on the extent to which known alleles could be recovered from genomic data, we used the validation classes from Wang et al [220]. We found that most of the IMGT alleles that were classified as high confidence (levels 1 and 2) were successfully recovered. By contrast, there were only a few IMGT alleles labeled as low confidence (levels 3, 4 and 5) found among the inferred alleles (Supplementary Figure B3D).

Table 3.1. The shared allele counts between IMGT database and inferred alleles with minimum repeat threshold of one. The compared alleles were restricting to the genes that are GRCH37 and GRCH38. The IMGT alleles retrieved from non-germline sequences were not included in the comparison.

	IGHV	IGLV	TRBV	TRAV	IGHJ	IGLJ	TRBJ	TRAJ	IGHD	TRBD
Number of alleles in IMGT	208	133	88	56	13	12	14	50	34	3
Number of alleles inferred from G1K	3746	3917	370	568	100	160	22	162	249	2
Number of putative novel alleles	3609	3740	318	511	90	141	8	114	219	0
Number of alleles shared	83	99	54	46	6	12	13	48	30	2
Proportion of IMGT alleles found in G1K	39.90%	74.40%	98.10%	100.00%	46.15%	100.00%	92.86%	96.00%	88.24%	66.67%
Proportion of IMGT alleles not found in G1K	60.10%	25.60%	1.90%	0.00%	53.85%	0.00%	7.14%	4.00%	11.76%	33.33%

In addition to the known alleles we recovered large numbers of novel putative immune receptor alleles from the population sequence data not found in IMGT. These numbered in the hundreds for TRVA and TRBV and in the thousands for IGLV and IGHV, when we applied the minimal threshold of a single occurrence (Table 3.1). The number of putative novel alleles decreased rapidly, as this threshold was increased in the case of IG, but far more slowly in the TCR case, suggesting saturation with respect to the number of samples in G1K of the TCR but not of the IG allelic diversity (Supplementary Figure B1).

3.4.4 Comparison of IMGT and Lym1K on real datasets

We compared the performance of the IMGT and Lym1K databases on real TCR and IG short-read sequence datasets. Because there are no high throughput sequencing datasets for TCRA currently in the public domain (to our knowledge) this comparison was restricted to IGH, IGL and TCRB. For each input sequence, if the alignment score of the best match of the sequence is higher using Lym1K rather than IMGT alone as the reference database, it demonstrates that the Lym1K database contains additional allele(s) that are more similar to the input read than the most similar allele from IMGT. Consequently, this corresponds to a sequence for which an improved match can be found when alleles recovered from genomic sequence data are included in the reference database. The median length of the IGHD genes is only nineteen nucleotides, and the entire D gene is within the CDR3 region with high mutation rates. During alignment, the median of the perfect matching nucleotides of the D gene is only six, which hampered the accuracy of the assessments on IGHD genes. Therefore we focused our comparison on V and J gene alleles.

We calculated the portions of the input sequences with improved alignment performances (figure 3.3). To assess the differences in alignment performance in greater detail, we classified the aligned reads based on the number of mismatches with the best-matching reference allele. Furthermore, In order to rule out the possibilities that Lym1K achieved improved alignment performance due to the larger number of included alleles, in each individual, we only kept the top two best aligned alleles (corresponding to the number of haplotypes per diploid individual) for each gene in the following comparison. We then compared the proportion of reads within each class between the two reference databases (figure 3.4). The proportions of unaligned reads were consistently lower using

Lym1K while the proportions of reads with zero or 1-2 mismatches were consistently higher. This demonstrates that the Lym1K database contains alleles that are more similar to the repertoire of the input sample. For instance, IGHV1-2*75p is a novel allele which we inferred from the G1K data and included in Lym1K. For sample SRR611358 there were 67 input sequences aligned with IGHV1-2*75p with an improved alignment score compared to the best-matching alleles (IGHV1-2*02 or IGHV1-2*03) in IMGT (figure 3.5).

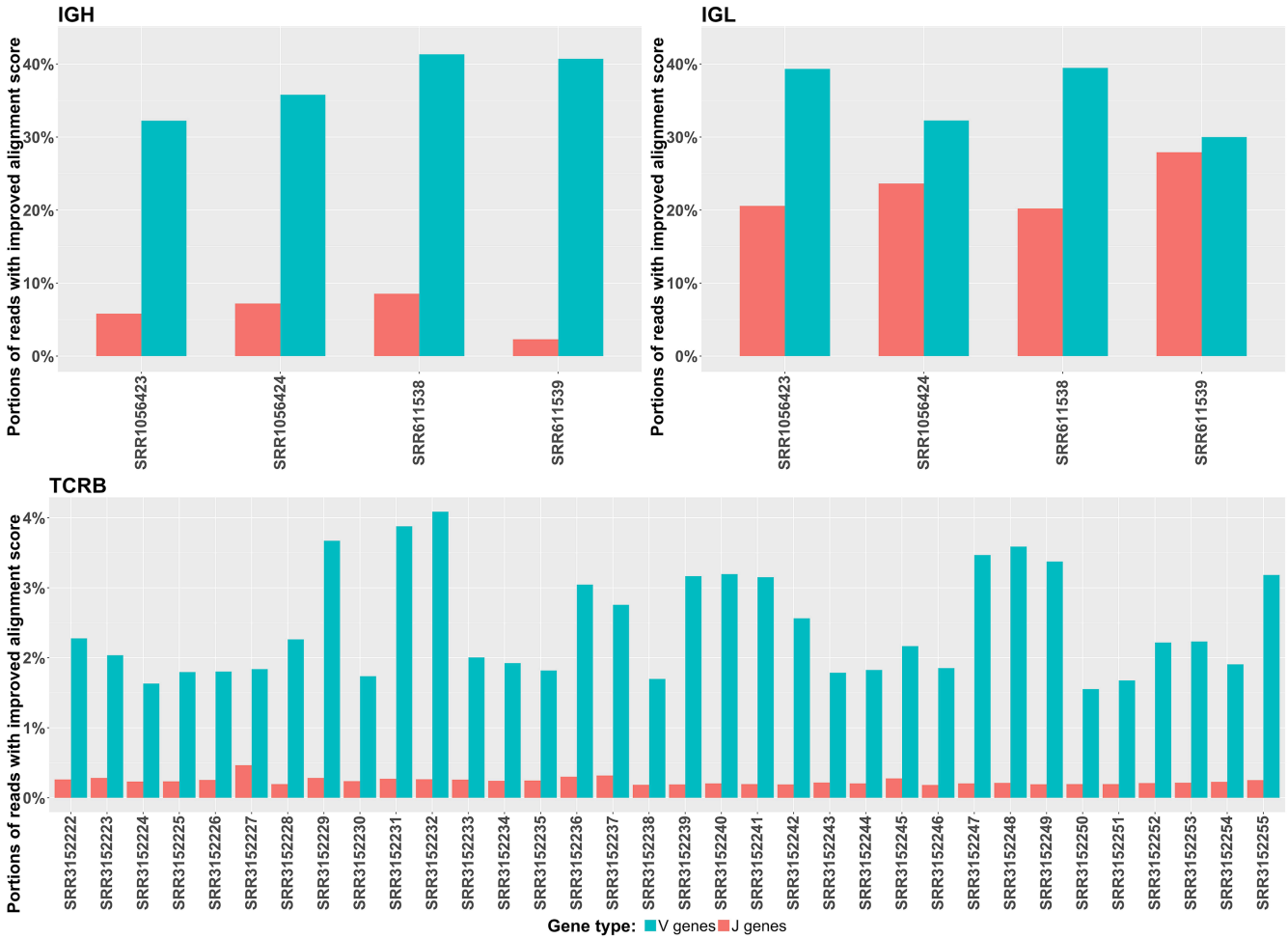


Figure 3.3. Improvements in alignment performance using the Lym1K database.

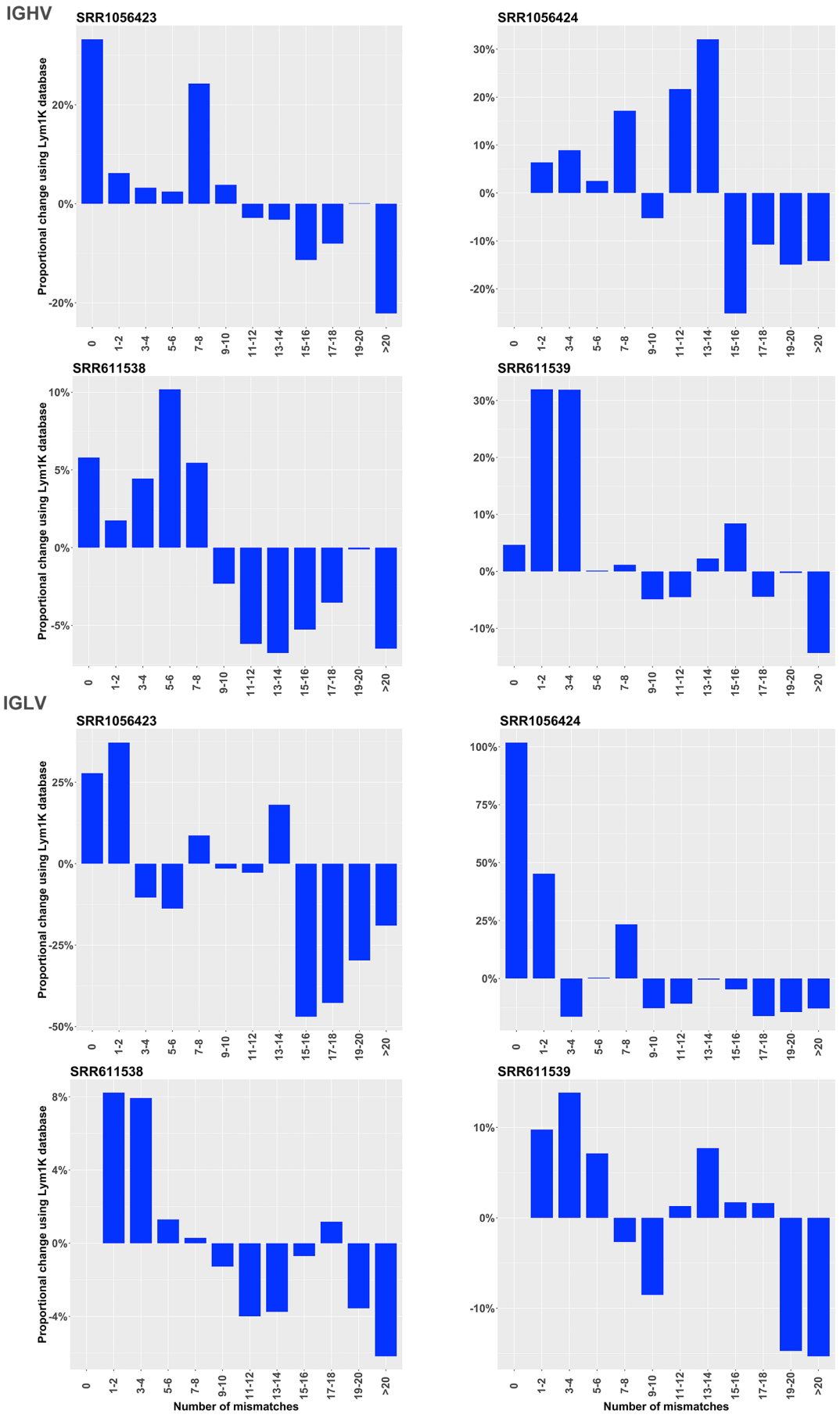

IGHV
SRR1056423
SRR1056424
SRR611538
SRR611539
IGLV
SRR1056423
SRR1056424
SRR611538
SRR611539
Proportional change using Lym1K database
Number of mismatches
0
1-2
3-4
5-6
7-8
9-10
11-12
13-14
15-16
17-18
19-20
>20

Figure 3.4. Alignment performance difference between IMGT and Lym1K. The aligned reads were classified into thirteen categories starting from zero mismatches with increments of two mismatches, to twenty mismatches and more than twenty mismatches. For example, in Figure A the number of alignments with zero mismatches increased by 33.2% using Lym1K for sample SRR1056423.

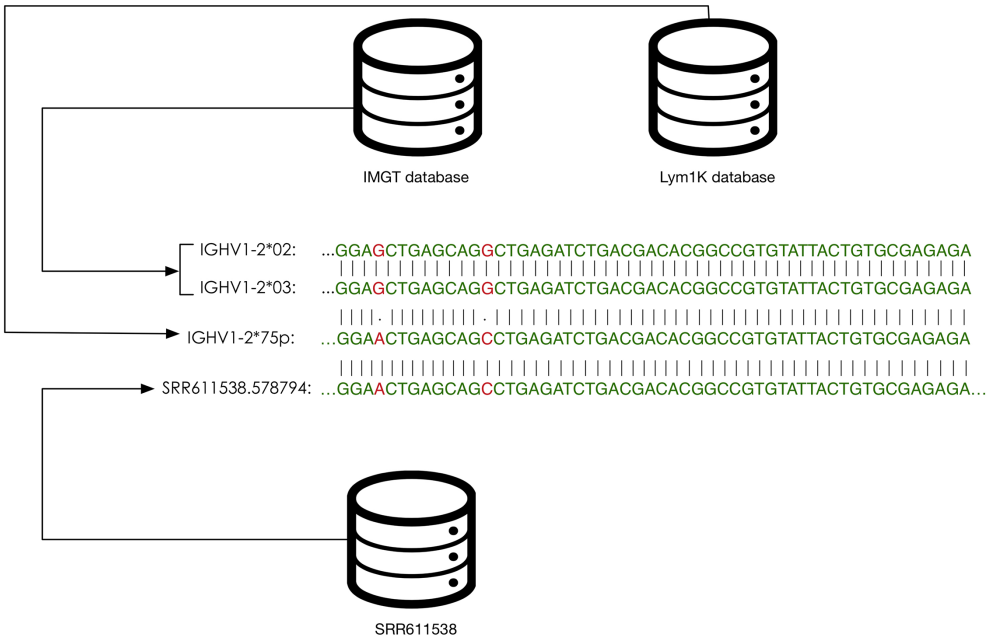


*Figure 3.5. Example of a novel allele (IGHV1-2*75p) in Lym1K that matches a short-read sequence from a real IGHV sequencing dataset (SRR611538). IGHV1-2*02 and IGHV1-2*03 are both found in the IMGT database and the sequence (SRR611538.578794) was aligned to these alleles in the original study. These two alleles both have two mismatches with the observed sequence. The inferred allele differed from the two most similar alleles found in IMGT, but showed a perfect match with the input sequence read.*

In the case of TCRB we introduced 318 putative novel TRBV alleles in Lym1K (Table 3.1). However, the improvements of the alignment performance of the TCRB sequences were small (figure 3.3). Among 34 samples, the highest proportion of reads with improved alignment was 4.1% (in sample SRR1033671) and the average proportion was only 2.4%. Therefore, it appears likely that IMGT contains adequate representation of the TCR diversity in the individuals from whom these samples were derived.

3.4.5 Improved population coverage of Lym1K

The IG dataset consisted only of samples collected from the South African population [233]. The TCRB samples were from Caucasians and African Americans. This may have affected the results of the performance comparisons due to the under or over representation of certain populations in the IMGT database. For instance, if the IMGT database mainly consists of alleles from the Caucasian population, this could result in better alignment results for datasets that originate from Caucasian populations compared to datasets from African populations. In general, African populations display greater genetic diversity than non-African populations [227] and this greater genetic diversity can be also be seen in the immune receptor regions (figure 3.6). This suggests that a comprehensive collection of human immune alleles requires sampling from diverse global populations, and particularly those from the African continent. We carried out simulations to assess the potential for differential performance of the existing IMGT database on datasets derived from different global populations. The simulation pipeline is described in detail in the Experimental Procedures section. For each chain, we simulated separate datasets for each of the populations in G1K as well as one pooled dataset corresponding to all populations.

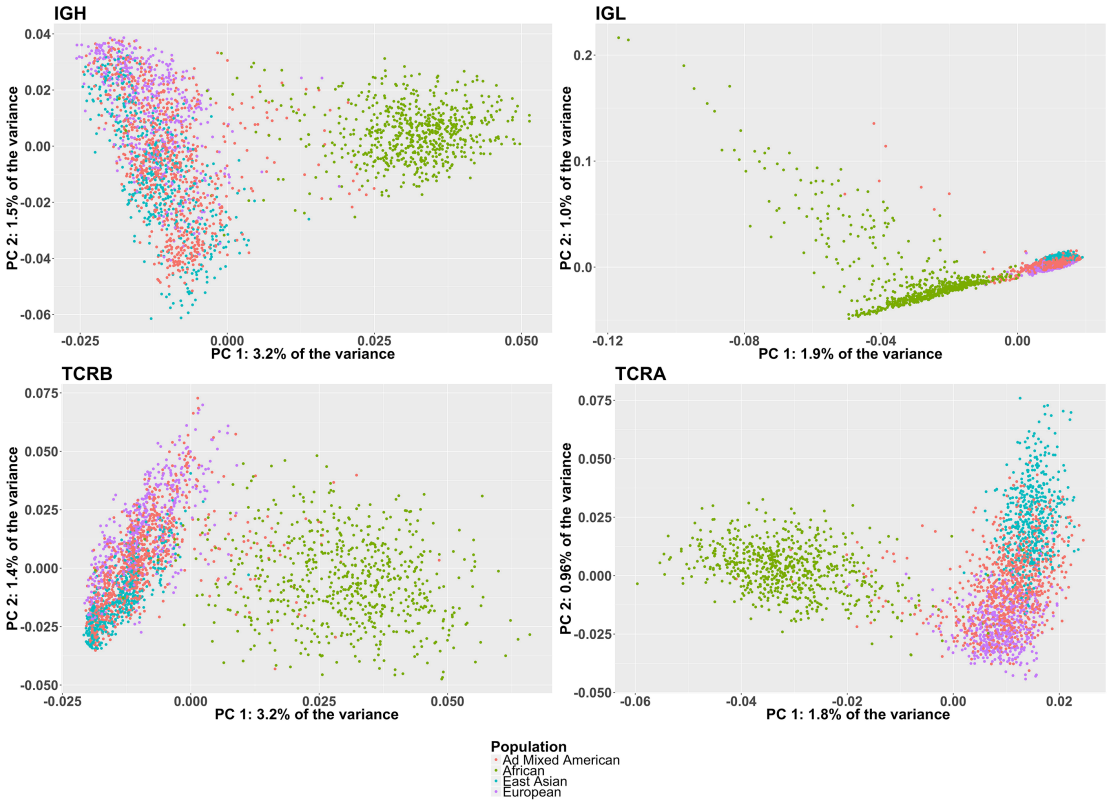


Figure 3.6. Principal component analysis of TCR/IG gene variation in human populations.

We used LymAnalyzer [217] to map simulated reads to reference alleles contained in IMGT and determined the proportion of reads that were mapped to the correct allele. Note that, in this case, our objective was not to compare IMGT to Lym1K as the simulation is based on the same haplotypes used to construct Lym1K. Instead we wished to compare the accuracy of the alignments obtained when we used the existing IMGT database to map simulated reads from different human populations.

For all but one of the genes (TCRA), a significantly lower mapping accuracy was achieved for African than for non-African populations, indicating that the additional variation in immune genes found in African populations has not been captured adequately in IMGT (figure 3.7). For the IGH datasets, The African populations (LWK, GWD, MSL, YRI, ESN, ASW and ACB) contained significantly more incorrectly mapped reads than in Non-African populations ($p < 2.2e-16$). The IGL datasets showed the largest portions of reads that were mapped inaccurately using IMGT. Again, using the IMGT database, there were noticeably larger portions of reads incorrectly mapped in the African populations than in Non-African populations ($p < 2.2e-16$). The portions of incorrectly mapped reads in TCRB datasets ranged from 4.9% (KHV) to 9.9% (LWK), and there were also more incorrectly mapped reads in African populations compared to Non-African populations ($p < 2.2e-16$). The TCRA datasets had the highest accuracy among the four chains using the IMGT database as the reference, with the smallest variation across populations. In contrast to other chains, the top three most incorrectly mapped populations in TCRA datasets are three East Asian populations (CDX, KHV and CHS), which were among the least incorrectly mapped populations in other chains.

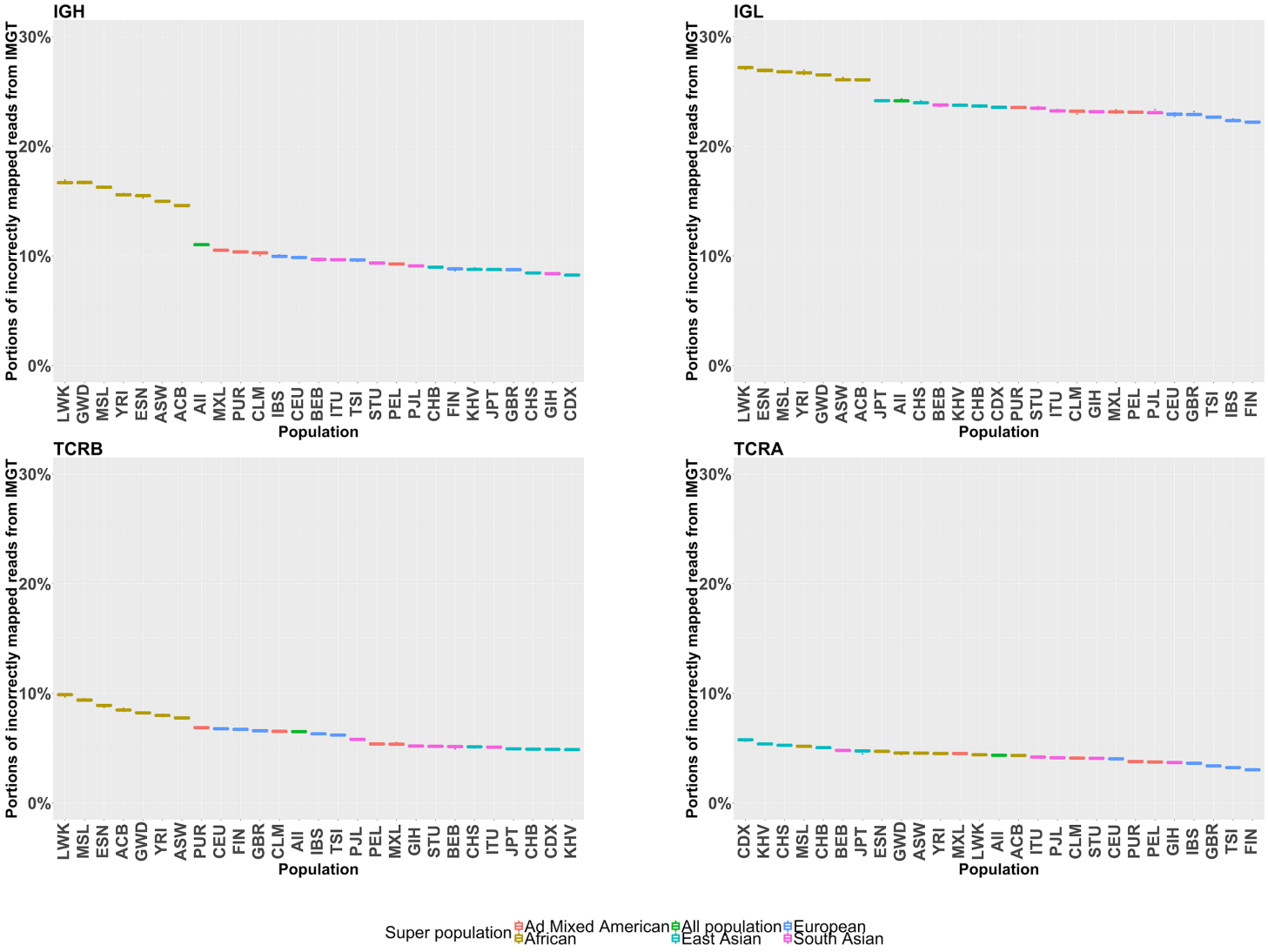


Figure 3.7. Comparison of the proportions of incorrectly mapped simulated reads among different populations using the IMGT database as the reference. Error bars, corresponding to two standard errors are shown in each plot. Population code: CHB - Han Chinese in Bejing, JPT - Japanese in Tokyo, CHS - Southern Han Chinese, CDX - Chinese Dai in Xishuangbanna, KHV - Kinh in Ho Chi Minh City, CEU - Utah Residents (CEPH) with Northern and Western Ancestry, TSI - Toscani in Italia, FIN - Finnish in Finland, GBR - British in England and Scotland, IBS - Iberian Population in Spain, YRI - Yoruba in Ibadan, Nigeria, LWK - Luhya in Webuye, Kenya, GWD - Gambian in Western Divisions in the Gambia, MSL - Mende in Sierra Leone, ESN - Esan in Nigeria, ASW - Americans of African Ancestry in SW USA, ACB - African Caribbeans in Barbados, MXL - Mexican Ancestry from Los Angeles USA, PUR - Puerto Ricans from Puerto Rico, CLM - Colombians from Medellin, PEL - Peruvians from Lima, GIH - Gujarati Indian from Houston, PJL - Punjabi from Lahore, BEB - Bengali from Bangladesh, STU - Sri Lankan Tamil from the UK, ITU - Indian Telugu from the UK.

3.5 DISCUSSION

The advent of high throughput sequencing technologies has enabled detailed studies of immune receptor repertoire diversity and changes in repertoire diversity over time, with many applications, including assessing the efficacy of leukemia treatment, understanding immune responses to disease and determining the causes and consequences of changes in receptor diversity with age. Recently, a number of dedicated algorithms and bioinformatics tools have been developed to support such studies [208-219]; however, in addition to accurate alignment algorithms, the results of bioinformatics pipelines for the analysis of receptor diversity depend on the accuracy and completeness of the reference database of immune receptors and their alleles.

The IMGT database is currently the most widely used global reference in TCR and IG diversity analysis. However, multiple previous studies have shown that IMGT is incomplete [236-239] and one study suggested that some reported IGHV alleles might contain sequencing errors [220]. This is mainly due to quality control issues arising from the inclusion of alleles from early studies and inadequate updating since the first release of IMGT database. Here we provide evidence that alignment results obtained using IMGT as the reference source for TRBV, IGH and IGL datasets derived from African populations are likely to be worse than those obtained for non-African populations, suggesting that IMGT does not adequately capture the global diversity of human immune receptors and that certain global populations are particularly under represented. This supports the value of incorporating immune alleles inferred from population sequencing studies that were specifically designed to profile the genetic diversity of global human populations.

In order to create a more comprehensive TCR/IG reference database, we developed a novel bioinformatics pipeline that can be used to infer alleles from variant calling information, obtained from population sequencing projects. We applied this pipeline to data from G1K to create the Lym1K database and have made both Lym1K as well as the bioinformatics pipeline used to create it freely available (See Software Resource). Using real TCRB, IGH and IGL datasets, we have shown that the incorporation of alleles inferred from population sequencing studies leads to improvements in alignment performance. These improvements are substantial in the case of IG and result from novel

alleles that provide a better match to the receptor repertoire of the input dataset than can be obtained using the alleles in the existing reference database.

There were in total 22 genes missing either from the current genome build (GRCh38) or GRCh37 and the depth of sequencing was unsatisfactory on several TRBV genes from G1K project (Supplementary Figure B3). This represents a limit to the comprehensiveness of receptor alleles inferred from the currently available population sequencing data; however, further improvements in the human reference genome and future population sequencing studies, will enable alleles of these genes to be discovered.

Our study suggests that the diversity of IG genes may not be profiled completely with existing sample numbers (Supplementary Figure B1). The availability of genomic variants from larger numbers of individuals sampled from global populations could lead to further improvements in the comprehensiveness of immune receptor alleles recovered from genomic data. Our bioinformatics pipeline is freely available and can be applied to recover additional immune alleles from further public or restricted access VCF files and updated genome builds, as they become available. Furthermore, we have designed our short-read mapping package, LymAnalyzer [217], so that Lym1K or any other collections of immune alleles produced using our pipeline or by other means can be substituted seamlessly as the reference allele database.

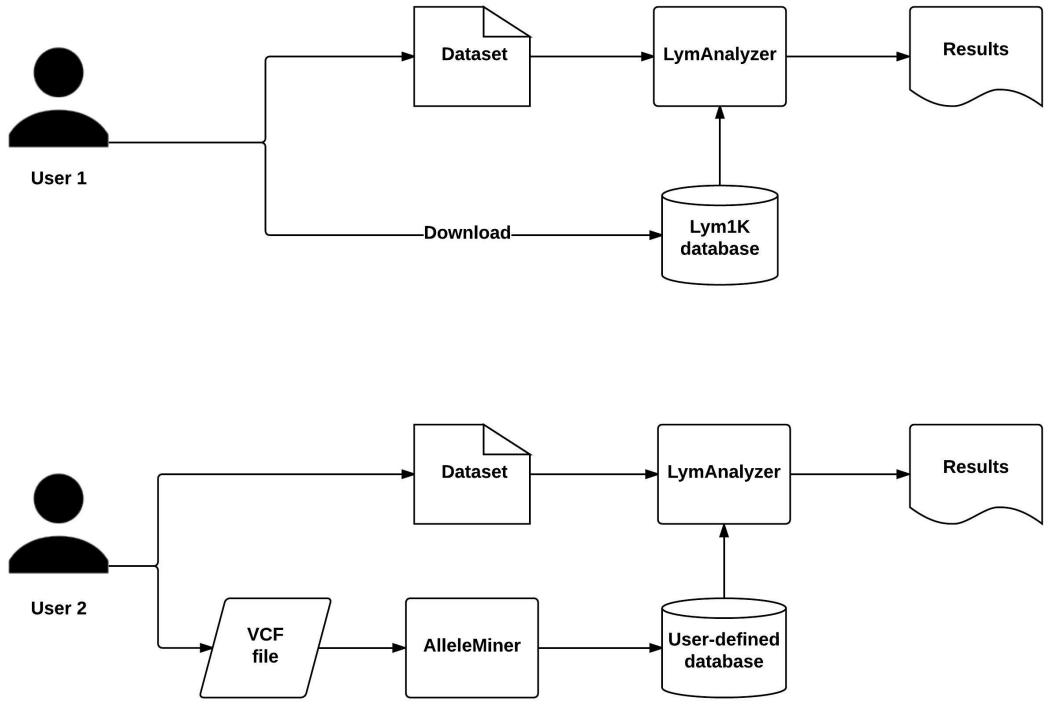


Figure 3.8. Two use cases. Lym1K enables User 1 to perform a more comprehensive analysis of immune receptor diversity by combining alleles inferred from G1K with known alleles in IMGT. User 2 applies AlleleMiner to recover additional receptor alleles from VCF data to generate a user-defined database, which is subsequently used as the reference database for receptor repertoire analysis.

Two use cases are envisaged for the resources described here (figure 3.8). First, users can apply Lym1K directly as their TCR/IG reference database for their analysis. Currently, G1K represents the largest public catalogue of human genome variation. However, as new public as well as restricted-access population sequencing datasets become available, and with further improvements in the human genome assembly, users may wish to replace or extend the reference dataset with their own database of reference alleles inferred from genomic variant data. Such users can apply AlleleMiner to their own VCF files to generate a customized TCR/IG reference database, which can then be selected as the reference database for LymAnalyzer. Indeed, depending on the specific data access restrictions that apply, such users may be able to contribute alleles discovered by applying AlleleMiner to restricted-access data into public repositories (such as IMGT), without violating data protection rules.

Chapter 4 - Diversity of T cell Receptor and Immunoglobulin Genes

4.1 Abstract

Among the most dynamic regions of human genome are the immune receptor (T cell receptor and immunoglobulin) loci, which contain a high degree of allelic variation. This feature, along with somatic rearrangements contributes to the extensive diversity of immunological repertoires. However, current understandings about the diversity of T cell receptor and immunoglobulin genes are constrained by the unsatisfactory coverage of immunological allelic information from the global population. To tackle this problem, we conducted a range of analysis about the diversity of T cell receptor and immunoglobulin genes based on allelic information retrieved from one of the largest human genome-resequencing catalogue, the 1000 Genomes Project. First, we compared the allelic diversity of immune receptor genes between African and Non-African populations and found that African populations had greater allelic diversity than Non-African populations. Subsequently, we hypothesized that the allelic diversity of immune receptor genes might associate with their genomic location and gene expression. However, based on comprehensive statistical tests, we did not find significant correlation between genomic location and allelic diversity in immune receptor genes. We hypothesized that RNA sequencing data for immune receptor genes derived from human whole blood could be used as a proxy for the frequency of gene segment use, and that more frequent gene segments would be under greater diversifying selective pressure. However, there was a lack of correlation between immune receptor gene expression and gene diversity, except for TCRA genes, which showed weak positive correlation. Overall, we present an analysis of the diversity of immune receptor genes, based on data currently available in the public domain. Further investigations with improved population coverage of human resequencing data, as well as, ideally, single cell RNA sequencing will be required to provide definitive answers to some of the questions posed.

4.2 Introduction

The immune system in vertebrates is complex and provides protection from foreign entities such as viruses, bacteria and parasites. The pathogenic world challenges the immune system by presenting enormous variability of antigens, which requires

specific recognition. In response, the adaptive immune system, as one subsystem of the overall immune system, has the amazing ability to initiate specific responses to a very large variety of antigens. This is based on it's extensively diversified antigen recognition receptors attached on the surfaces of T and B cells, namely, T cell receptors (TCR) for T cells and Immunoglobulins (IG, B cell receptors or antibodies) for B cells. During the early stages of maturation of T and B cells, genes for these receptors undergo a unique somatic gene rearrangement mechanism called VDJ recombination, which involves a largely random process of rearrangements of V (variable), J (joining) and D gene segments (diversity, only in TCR beta chain and IG heavy chain). Immunoglobulins are furthermore diversified by additional somatic hypermutations during the immune responses.

The mechanism of diversification of T cell receptors and Immunoglobulins has been investigated in great detail in previous studies [20, 74], mainly focusing on VDJ recombination and somatic hypermutation. However, although the diversity produced by VDJ recombination ultimately depends on the diversity of the TCR and IG V, D and J segments themselves, there have been fewer studies of the genomic diversity of these genes and of how this diversity both shapes and is determined by TCR/IG repertoire diversity and immunological functions. TCR and IG V gene loci are among the most segmentally duplicated regions in human genome and they are known to exhibit very high frequency of single nucleoside polymorphisms and copy number variations [221, 253]. Furthermore, germline variations in IGHV genes have been shown to associate with antibody function [254]. A better understanding of allelic diversity of TCR and IG V genes can help us address the structural characteristics of T cell receptors and Immunoglobulins and may reveal the connections between susceptibility to specific diseases and specific immune gene polymorphisms.

Previous studies investigating the allelic diversity of TCRs and IGs were mainly based on the alleles from IMGT database and some reported putative alleles from low-throughput sequencing technologies [255-258]. However, several studies have shown that this database is incomplete and may contain errors. More importantly, many more novel V gene alleles were discovered in recent studies [220-222], especially from African populations, suggesting that we may have very limited understanding of the real variability of TCR/IG V genes and our current reference database may contain

unsatisfactory coverage of the alleles from some populations, particularly populations from the African continent [236].

Next Generation Sequencing (NGS) technologies shed some light into this important question with its much lower sequencing cost and efficiency comparing to the traditional Sanger sequencing technology [108, 110]. Thus, there were many improvements of human reference genome over time with frequent updates in recent years. In addition to that, the most recent phase (phase three) of 1000 genomes project (G1K project) provided a comprehensive description of genetic variants of the human genomes of 2504 individuals from 26 populations across the world [143]. Furthermore, many recent studies were focused on unraveling the genetics of human gene expression, which led to the accumulation of large amount of whole blood RNA sequencing (RNAseq) data. These high throughput datasets provide opportunities to address important questions about the diversity of human immune responses. For example, using RNAseq data from whole blood, a recent study investigated the association between major histocompatibility complex (MHC) alleles and TCR gene expression [259], providing a better understanding of the long-standing question about TCR-MHC genetic relationship.

In this chapter, we aimed to investigate the global human diversity of TCR and Immunoglobulin V genes and to relate allelic diversities of TCR/IG V genes both to their physical locations on chromosomes and to their frequency with which they are found in RNA expression data. In our previous study, we reconstructed a database of alleles of TCR/IG V genes based on the variants information of G1K project [143, 222]. Using this the new database, we compared the diversities of the TCR/IG V gene alleles across the 26 populations and found that, as is the case for other loci, African populations have the largest allelic diversities of TCR/IG V genes. Our results suggested that there was no clear evidence showing that allelic diversities of TRBV and IGHV genes are associated with their expressions or genomic locations.

4.3 Results

4.3.1 Diversity analysis of TCR/IG alleles inferred from G1K project

There has been a growing interest from immunologists in understanding the diversity of T cell receptor and Immunoglobulin repertoires, but to-date there have been insufficient studies of germline variations of TCR/IG genes due to the limitation of available human resequencing data. In our previous work, we developed AlleleMiner, a bioinformatics pipeline to infer all putative alleles from variant calling data. Subsequently, by using this tool along with the G1K data, we created the Lym1k database, currently the most comprehensive collection of putative TCR/IG alleles [222]. Here, we exploited our Lym1k database to investigate in depth the diversity of germline TCR/IG V genes.

4.3.2 Distinctive allelic diversities in different genes and subpopulations

Comparative studies of genetic diversities among different human populations help us understand the genetic basis of phenotypic adaption and disease [253]. The African population is particularly important in these comparisons because previous studies have shown that, due to the African origins of modern humans, there is greater genetic diversity in African populations than in all non-African populations combined [227]. Therefore, we first hypothesized that the allelic diversity in TCR/IG genes from African population is also larger than that found in non-African populations. To test this, we calculated Shannon entropy (also in this context referred to as the diversity index) as a measure of the allelic diversity of each gene (see Materials and Methods for details). In short, the Shannon entropy of a given gene describes the richness (the number of the alleles) and evenness of the frequency distribution (how close in frequency the alleles are) of the alleles of the gene.

From the three types of TCR/IG gene segments (V, D, J), we excluded D gene segments in our analysis due to their short lengths and high mutation rates (they are fully located within the CDR3 region), which limited their accuracy in Lym1K database, as described in detail in our previous work [222]. We also excluded TRBJ and IGHJ genes because of their insufficient coverage in G1K project (half of the TRBJ and IGHJ genes were with median coverage less than 10). Therefore, we restricted our analysis to the IGHV, IGLV, TRBV, TRAV, IGLJ and TRAJ genes. We first calculated the overall Shannon entropies of each gene using data from all the populations combined. We discovered that the Shannon entropies of the alleles of

different TCR/IG genes vary significantly (figure 4.1, Supplementary figure C1). For instance, the most diversified IGHV gene, IGHV4-4, has a Shannon entropy 65 fold larger than IGHV3-72, which is the least diversified IGHV gene. Additionally, IGHV genes showed far greater Shannon Entropy (median Shannon Entropy 1.04) than the remaining genes (median Shannon Entropy 0.42, 0.23, 0.3 and 0.22 for IGLV, TRBV, TRAV, IGLJ and TRAJ, respectively). The entropy of the two J segments included in the study was lower than for any of the V segments.

We also calculated the Shannon entropies of the TCR/IG genes within each of the populations separately (figure 4.2, supplementary figure C2, supplementary table C1). Not surprisingly, we found that the allelic diversity in African populations (YRI, LWK, GWD, MSL, ESN ASW and ACB; population code can be found in supplementary table 2) is significantly larger than in the non-African populations in IGHV ($p = 6.57e-09$), IGLV ($p = 3.73e-11$), TRBV ($p = 1.4e-04$), TRAV ($p = 1.28e-06$), IGLJ ($p = 4.5e-03$) and TRAJ ($p = 1.8e-02$) genes.

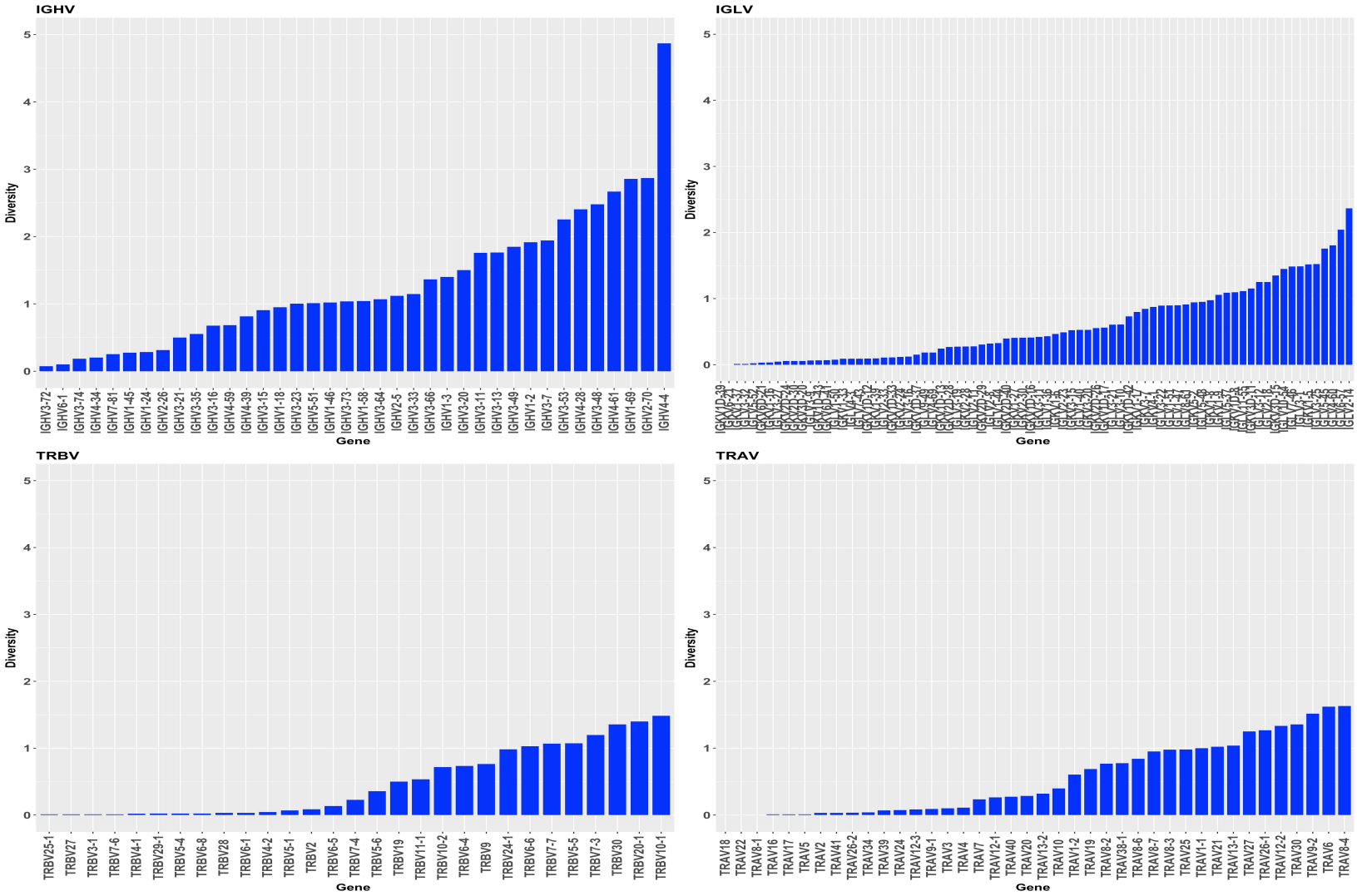


Figure 4.1. Shannon entropies of IGHV, IGLV, TRBV and TRAV genes.

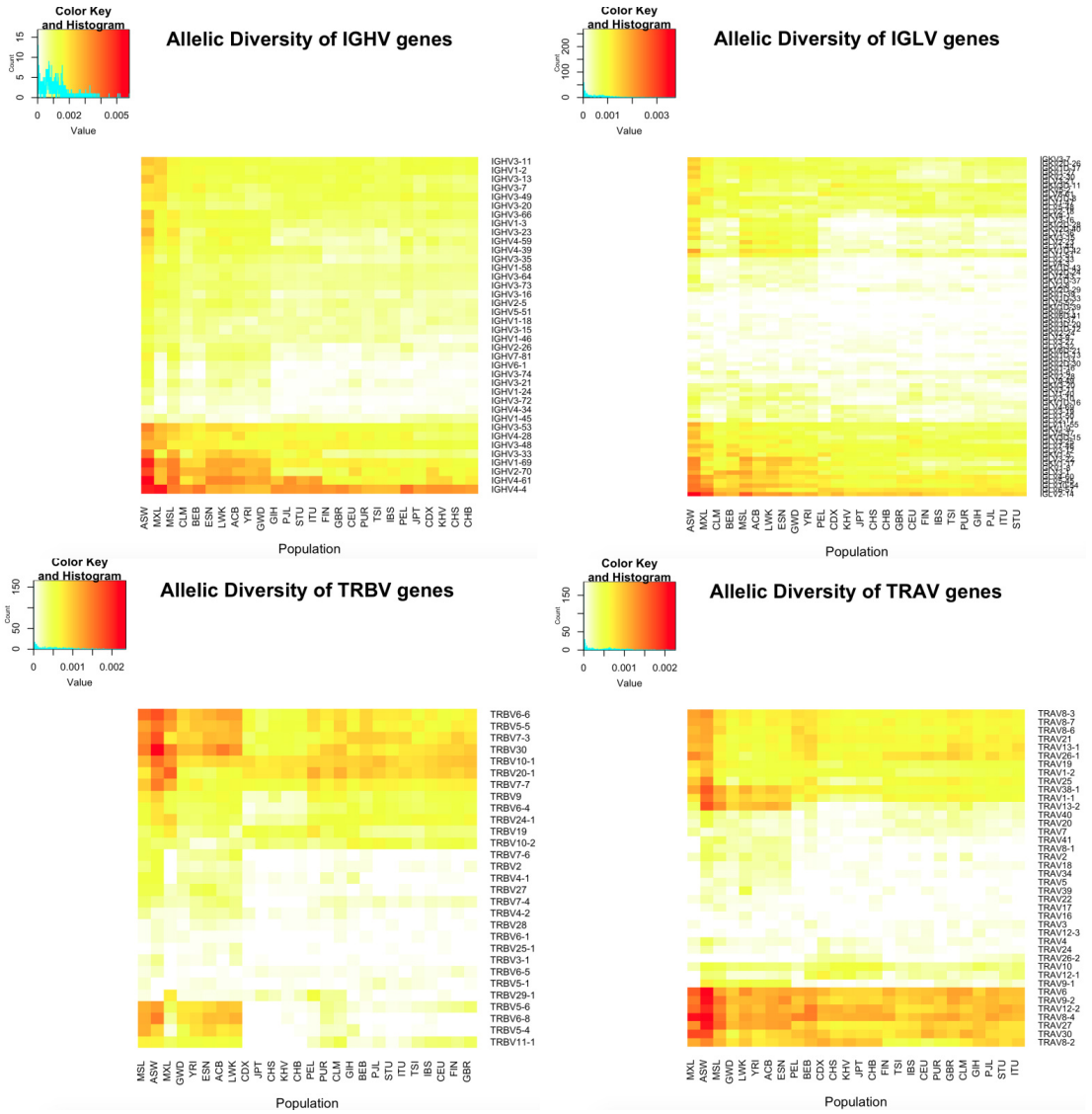


Figure 4.2. Allelic diversity of TCR/IG V genes represented in heatmap from different populations.

4.3.3 Immune gene expression in MDD

There have been several reports of possible links between depression and human immune system [260-263]. Since half of the samples of our whole blood RNAseq data were derived from individuals with major depressive disorder (MDD), it is important to determine whether the expression patterns in TCR/IG V genes differed between MDD and healthy samples, as this could introduce a bias in our expression analysis. Therefore, we compared the expression of IGHV, IGLV, TRBV and TRAV genes respectively from MDD and healthy samples (figure 4.3). No significant difference between these two groups in their immune gene expression

levels was found, suggesting that we can apply our expression analysis on all the samples despite healthy or MDD.

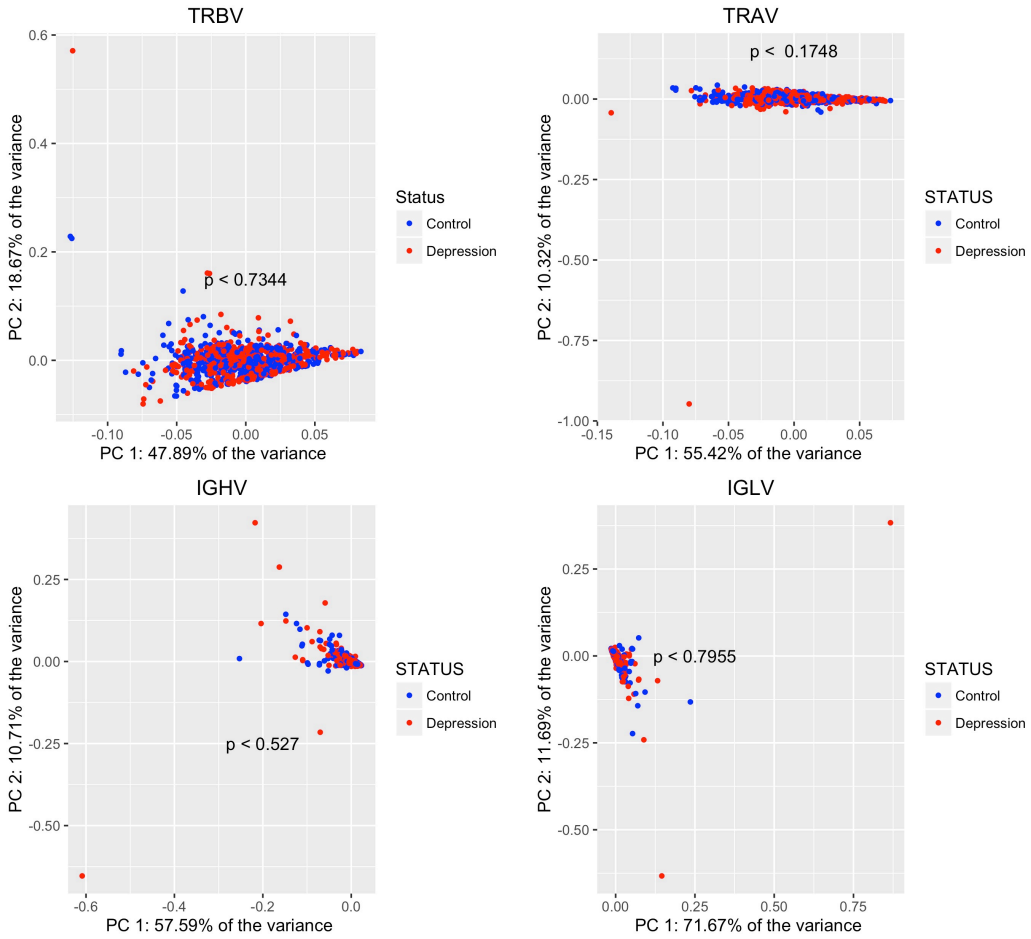


Figure 4.3. Principal component analysis of TCR/IG V gene expression between samples from MDD and healthy.

4.3.4 Relationship between allelic diversity and genomic location

The TCR/IG V genes from the same chain types are all located adjacent to each other forming gene groups on the chromosomes. One long-standing question is whether the physical locations of the TCR/IG genes play an important role during the selection of genes in V (D) J recombination and how this would influence the allelic diversity of the genes. Studies on mice have shown that there are more rearrangements using the TRAV genes on the 3' end than the 5' end [264], we therefore hypothesized that the usage of certain gene segment during VDJ recombination may associate with their genomic location: The more distally located (relative to the 3' end) gene segments might be less used during the VDJ

recombination, resulting in lower allelic diversity (as a consequence of weaker diversifying selective pressure). Assuming that this hypothesis is correct, we expected to find a correlation between genomic position of TCR/IG genes and their allelic diversity. However, no such correlation was evident in the data (figure 4.4; Spearman rank-order correlation tests non-significant for all genes). The lack of correlation between genomic position and allelic diversity in the case of TRAV genes may indicate either that the tendency for more rearrangements involving genes towards the 3' end that has been observed for mouse is not found in humans or that allelic diversity is independent of the frequency of V gene use. In support of the former, we found no relationship between the genomic locations of all types of TCR/IG genes with their RNA expression level (figure 4.5).

4.3.5 Relationship between allelic diversity and gene expression

Although VDJ rearrangement is regarded as a largely stochastic process, based on numerous studies about TCR/IG repertoires, the expressions of different TCR/IG genes are not uniformly distributed. Instead, some of the TCR/IG genes are regarded as more "public" than other TCR/IG genes: these TCR/IG genes are more frequently selected across many individuals during the VDJ recombination even prior to positive and negative selection. In addition to that, the usages of TCR/IG genes differ significantly within different conditions such as disease and autoimmune, as a result of antigen-related selection. For instance, in Michael et al.'s study [226], they found that TRBV19 dominated the TRBV gene use in leukemia patients' TCR repertoire, which accounted for 99.8% of the clones. One intuitive question is whether the more frequent use and higher expression of certain genes would result in greater diversifying selective pressure acting on these genes. We therefore hypothesized that the more frequently expressed genes might associate with greater germline allelic diversities, resulting in a positive correlation between expression level and allelic diversity. We discovered relatively weak positive correlations that were statistically significant (significance level < 0.1) between gene expression and allelic diversity in TRAV and IGLV genes (table 4.1). However, we did not find any correlations that were statistically significant in IGHV, TRBV, IGLJ and TRAJ genes (table 4.1, supplementary figure C5),

suggesting that the expression of these genes may not be associated with their allelic diversities.

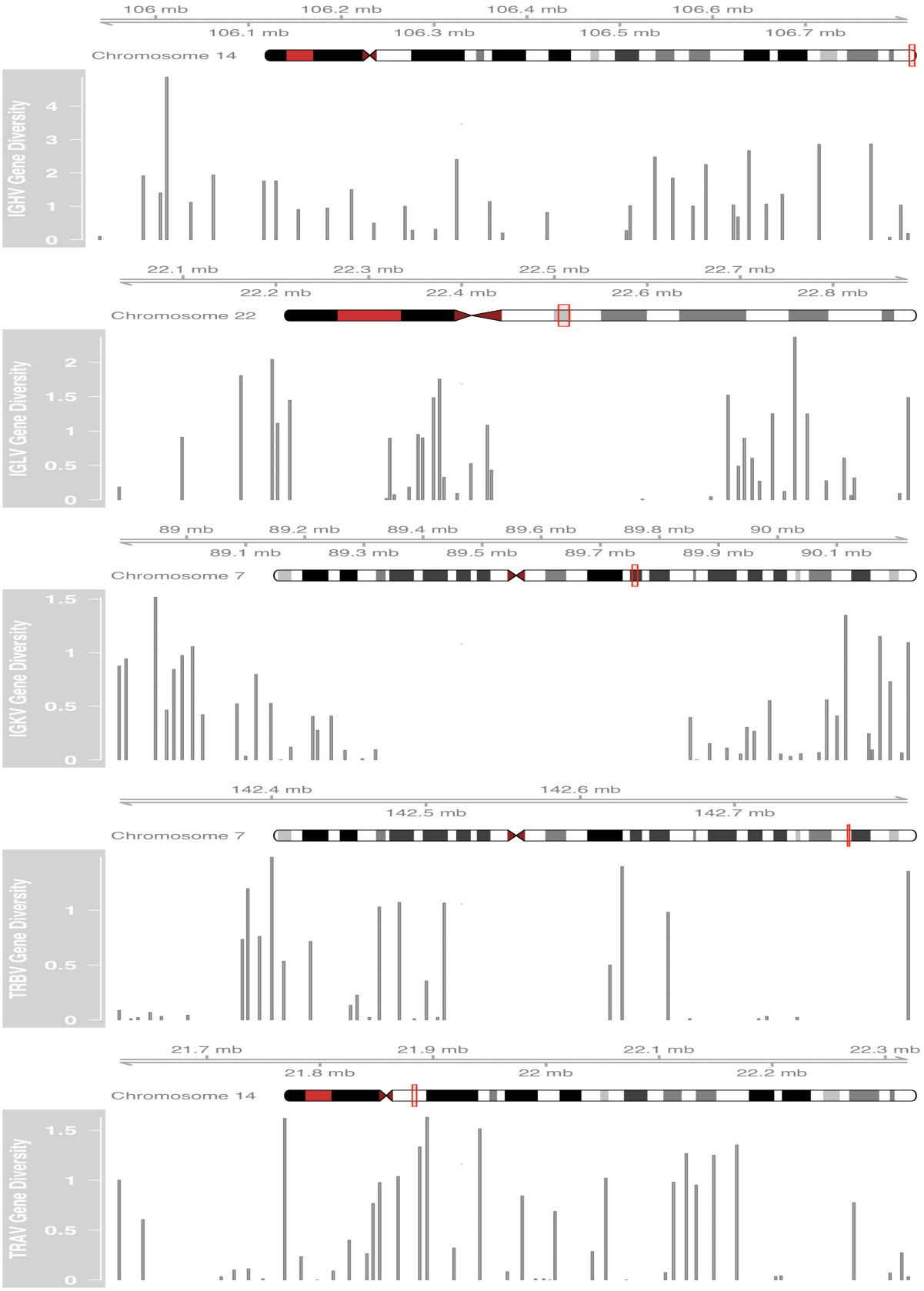


Figure 4.4. The diversities of immune genes along the chromosome with physical coordinates mapped. Each bar in the plot represents the Shannon entropy of one particular gene and the distance between genes were relative to their physical locations on chromosome. The plot was generated using Gviz [267].

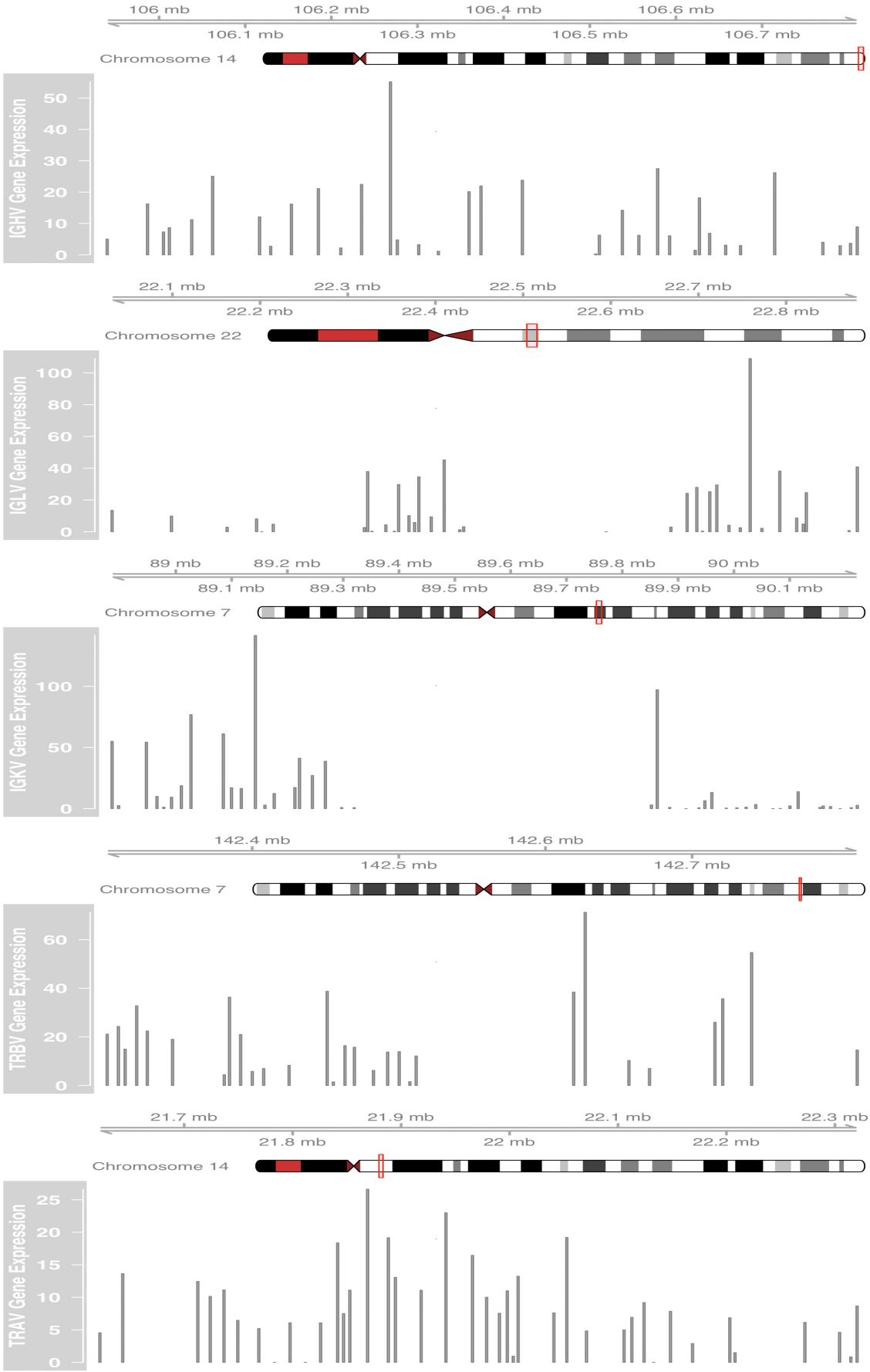


Figure 4.5. The expression of immune genes along the chromosome. Each bar in this plot represents the RNA expression level of a particular gene.

Table 4.1. Correlation between TCR/IG gene expression and allelic diversity. The p-values were corrected by using Benjamini-Hochberg correction.

	IGHV	IGLV	TRBV	TRAV	TRAJ	IGLJ
Spearman's rank correlation coefficient	rho =0.16	rho=0.25	rho = -0.15	rho = 0.29	rho=0.19	rho = 0.17
	p-value=0.55	p-value=0.04	p-value=0.71	p-value=0.07	p-value=0.35	p-value=0.60
Pearson product-moment correlation coefficient	cor=0.05	cor= 0.17	cor= -0.003	cor=0.32	cor=-0.04	cor=0.20
	p-value=0.78	p-value=0.13	p-value=0.99	p-value=0.07	p-value=0.74	p-value=0.60
Kendall rank	tau = 0.1	tau = 0.17	tau = -0.09	tau = 0.20	tau=0.12	tau = 0.13
Correlation coefficient	p-value=0.55	p-value=0.04	p-value=0.71	p-value= 0.07	p-value=0.35	p-value= 0.60

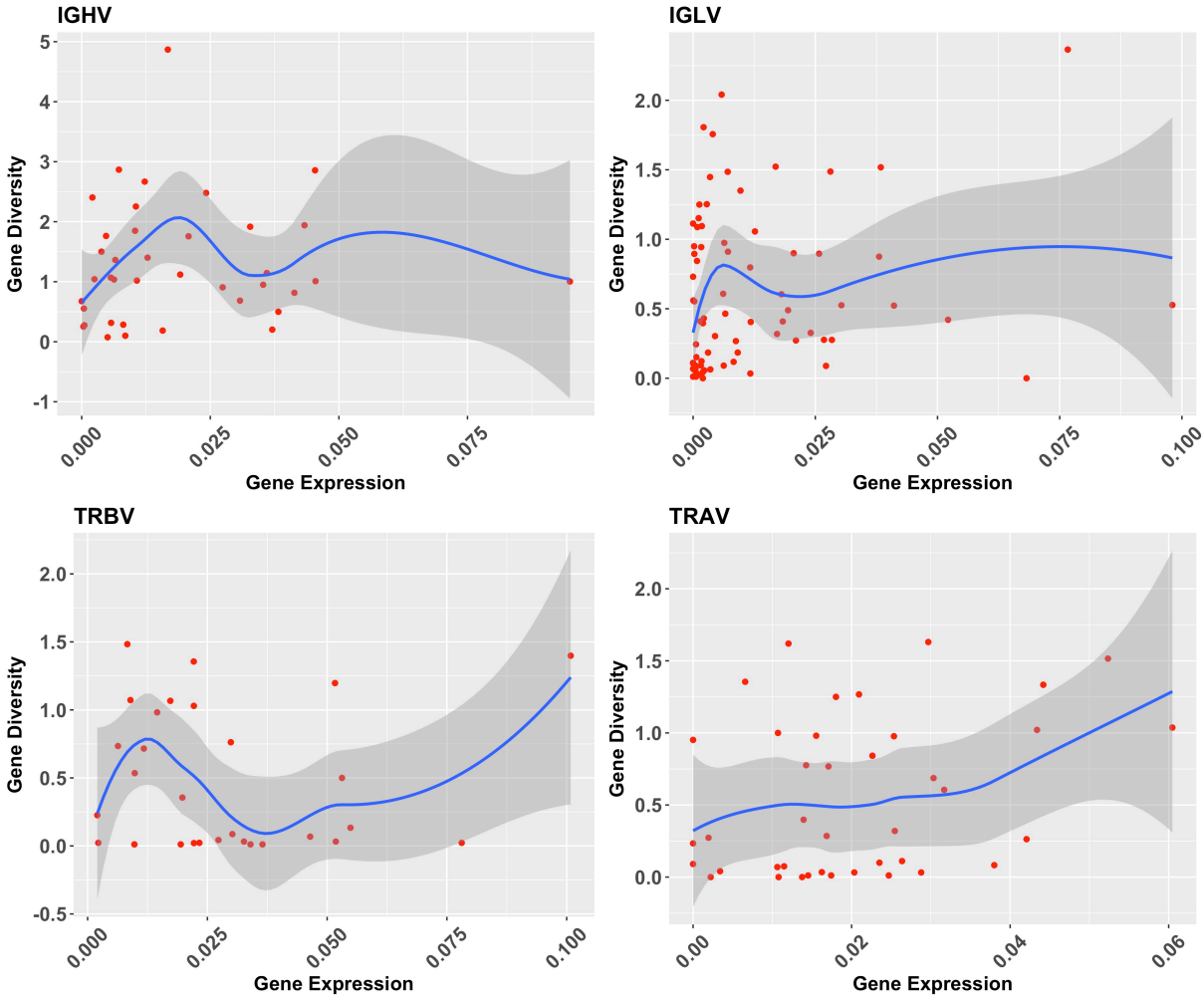


Figure 4.6. The relationship between gene expression and allelic diversity in IGHV, IGLV, TRBV and TRAV genes. The data is fitted by using Local Polynomial Regression (LOESS). The shaded areas are the standard errors.

4.4 Materials and Methods

4.4.1 TCR/IG alleles from Lym1K

We used the collection of TCR/IG alleles included in Lym1K database, which has been introduced in our previous work [222]. For completeness, we used the lowest inclusion threshold of Lym1K database. More specifically, we regarded each of the inferred unique haplotype as an allele.

4.4.2 Normalization of RNAseq data

We obtained the whole blood RNAseq data from the Depression Genes and

Networks Project created by Battle et al [268] (National Institute of Mental Health grant 5RC2MH089916). The dataset consists of 922 individuals, with 463 individuals with major depressive disorder and 459 healthy individuals.

The expressions of TCR/IG genes were calculated by using RSEM [184]. We normalized the expression level of TCR/IG genes based on Transcripts Per Kilobase million (TPM) by introducing the total number of reads of the specific gene type that the given gene belongs to. Our customized TPM for each gene is calculated as follows:

$$TPM_i = \frac{R_i * M * 10^6}{L_i * N * T} \quad (1)$$

Where R_i is the read count of the i_{th} gene, M is the mean of the read length, T is the total number of the reads of the given gene type (TRBV, TRAV, IGHV etc.), L_i is the length of the i_{th} gene and N is the total number of reads.

4.4.3 Shannon entropy

We used Shannon entropy as a measure of gene diversity. Shannon entropy is affected by both the allelic richness (how many different alleles belong to a given gene) as well as the evenness of the frequency distribution across alleles (for a given number of alleles, the Shannon entropy is maximized when all alleles have equal frequencies). The Shannon Entropy of each gene is calculated as follows:

$$H = -\sum_{i=0}^{n} p_i \, log_2 p_i \quad (2)$$

where p_i is the proportion of haplotypes carrying the i_{th} allele of the gene and n is the total number of the alleles. When calculating the diversity index within different populations, equation 2 is not suitable anymore because different populations consist of different numbers of alleles. We therefore normalized the sample size in the Shannon Entropy calculation as follows:

$$H = -\sum_{i=0}^{n} \frac{p_i log_2 p_i}{log_2 n} \quad (3)$$

4.5 Discussion

T cell receptors and Immunoglobulins play crucial roles in the human adaptive immune system by providing specific recognitions of foreign antigens and preventing growth of the corresponding pathogens. The sets of TCR and IG genes within the same chain type are located adjacent to each other in genomic regions that are among the most dynamic genomic region in human genome with high gene duplications and point mutations. Studies of the variability of the TCR/IG V genes at germline level may provide a better understanding of the evolutionary diversification of TCRs and IGs. However, such studies were lacking to date because of the unsatisfactory coverage of genome sequencing and genotype data for these regions in global populations.

In recent years, with improvements of the sequencing technologies, it has become possible to investigate the allelic diversity of immune genes based on human resequencing data with much higher resolution and larger sample sizes. In this study, we first retrieved the collection of alleles of TCR/IG genes included in the Lym1k database [222]. Based on these alleles, we investigated the allelic diversity within different genes and different populations. We found that among the four chain types (IGH, IGL, TRB, TRA), the alleles of IGHV genes contain the largest diversities. Furthermore, we discovered that, as is the case for most genomic regions, African populations demonstrate the highest allelic diversities of TCR/IG V genes.

We further investigated the effect of genomic locations of TCR/IG genes and their allelic diversities. We did not find any significant correlations between the genomic locations of different immune genes and their allelic diversities, suggesting that genomic locations of TCR/IG genes may not associate with their allelic diversities. Lastly, we investigated the relationship between expression of TCR/IG V genes and allelic diversity. Our expectation was that the frequent use and higher expression of certain V genes would result in greater diversifying selective pressure affecting these genes, resulting in greater germline diversity. However, correlations between gene expressions were either weakly positive (for IGLV and TRAV genes) or absent altogether (for IGHV and TRBV genes).

In this work, the data mainly came from two sources: the TCR/IG allele information retrieved from Lym1k database, which was created based on human resequencing data (G1K project) with the current human reference genome (GRCH38) and the TCR/IG RNAseq data generated from human blood samples. Limitations include the fact that coordinates of a some TCR/IG genes are not available in the current human reference genome: there were 22 TCR/IG V genes missing from the main build of the reference genome, which were required by AlleleMiner for the allele database construction. The second issue that may limit our findings is that many TCR/IG genes had insufficient sequencing coverage in the G1K project (genes with a cross-sample median coverage less than 10 were excluded). Nearly half of the IGHJ, IGHD, and TRBJ genes and eleven out of 45 TRBV had median depth less than ten in the current (phase 3) release of the G1K project. In addition, the real allelic diversity or the variability of TCR/IG V genes may still not be fully reflected by only 2504 individuals in G1K project. For instance, IGHV2-70 contains 591 distinct haplotypes out of 5008 haplotypes in total. Further improvements in both the human reference genome and in the coverage of population re-sequencing data are needed for a better understanding of the allelic diversity of immune genes.

Relating to the gene expression data, there might be three reasons explaining the lack of correlations between the expression and allelic diversity of TCR/IG genes. One reasons is that, since we used gene expression as an indicator or proxy of the frequency of the gene usage during the selection process, the relationship between gene usage and gene expression might not be strong. For example, one study illustrated that a single mutation on IGKV2-29 gene yielded the IGKV2D-29*02 allele, and this drastically decreased the expression of this gene [269]. For some of the genes, even though they were selected during the selection process, their expression level might be lower than other selected genes. Thus RNA expression level of one particular gene might not be an accurate reflection of how often this gene was selected. Second, the RNAseq data were retrieved from B and T Cells harvested from peripheral blood, which only revealed a limited view of the real diversity of the immune receptor repertoire. For instance, approximately only two percent of the B cells are included in the peripheral blood [204]. The usage of the different immune receptor genes might not be satisfactorily reflected by the limited immune receptor repertoire of peripheral blood. Lastly, the frequency of human immune receptor gene usage might not be stable over time, even if the frequency

was correctly reflected from the gene expression data, the frequency of gene usage inferred from this dataset may not reflect the mean of the frequency of the gene usage among the human populations, over the time during which the human allelic diversity in these regions has accumulated.

The bulk RNAseq dataset used in our analysis is very extensive with a large sample size, and it captured a snapshot of net results of average immune gene expression. However, there are still some improvements that can be made to precisely reveal the TCR/IG gene expression in higher resolution. Because the accuracy might be limited in bulk RNAseq experiments by cell heterogeneity, especially for T and B cells with such high diversity. The composition of different T and B cells (such as CD4, CD8 T cells) with different TCRs and IGs in the pooled samples can vary in RNAseq experiments, and this might result in different gene expression results. Single Cell RNAseq (scRNA-seq) technology provides a solution for this problem by allowing in-depth transcriptome analysis in single cells to ensure the same subtype of the T and B cells being compared. However, current single cell datasets commonly range from 10^2 to 10^5 cells from a very limited number of individuals, which may be insufficient for investigating the diversity of TCR/IG repertoire. But with fast growing popularity and lower sequencing costs in recent years, more comprehensive scRNAseq dataset will be available in the near future and we can gain much more insights into the relationship of immune gene expression and diversity.

Chapter 5 – Discussion

The adaptive immune system in mammals is characterized by its ability to recognize millions of different kinds of foreign antigens and to mount highly precise and enhanced immune responses to the invading pathogens. This is achieved through it's extensively diversified repertoire of antigen recognition receptors attached on T (T cell receptor) and B cells (immunoglobulin). Previous understandings about the complexity and functional characteristics of the immune receptor repertoire are limited by the lack of high-resolution data. In this thesis, we described three research projects that make use of cutting-edge high throughput approaches to interpret the diversity of T cell receptor and immunoglobulin genes. We first developed a software tool for systematically analyzing next generation sequencing data of TCRs and immunoglobulins, which provides an in-depth view of the frequency distribution of the TCR and immunoglobulin clonotypes. Second, we inferred a collection of TCR and immunoglobulin gene reference sequences based on large quantities of human resequencing data in the public domain. Last, we investigated the characteristics of allelic diversities of immune receptor genes.

Our first research question focused on tackling the computational challenges involved in the analysis of large-volume next generation sequencing data of TCR and immunoglobulin repertoires. In chapter two, we implemented LymAnalyzer, a high-throughput application that contains three analysis components for interpreting TCR and immunoglobulin sequence data in an accurate and efficient manner. The first analysis component of LymAnalyzer was to calculate the frequency distribution of the clonotypes of TCR and immunoglobulin repertoire, which was the major functional component of other similar tools [209-216]. This process included correct VDJ alignment and CDR3 extraction. We developed a fast-tag-searching algorithm for mapping the input sequences of the TCRs and immunoglobulins to their corresponding reference sequences in IMGT database. We compared the alignment results of LymAnalyzer with Decombinator, MiTCR and MiXCR on both real and *in silico* datasets. LymAnalyzer outperformed all the other tools with significant improved mapping completeness and accuracy. In addition to that, LymAnalyzer integrated two novel analysis components. To illustrate the mutation process occurring during affinity maturation in immunoglobulins, LymAnalyzer provides a lineage mutation tree construction function that can reveal the minimal steps required to

obtain the observed immunoglobulin sequences. Last, LymAnalyzer provides polymorphism analysis, which can detect putative novel SNPs in the input TCR or immunoglobulin sequences that have not been reported in the reference databases such as IMGT.

There are still future improvements that can be made on LymAnalyzer. Immunological repertoire sequencing technologies are fast evolving towards Multiplex-PCR with spike-in approaches to resolve the primer-bias issues. Additional normalization steps are needed for sequence data generated from such platforms. Furthermore, one potential factor that might hamper the accuracy of the polymorphism analysis is allelic imbalance. Because the SNP inclusion threshold used in LymAnalyzer is 10%, imbalanced alleles that were present in less than this threshold were ignored from the analysis. Therefore, we made the inclusion threshold user-adjustable, but special caution is needed when choosing the appropriate threshold. One important question that arose while applying polymorphism analysis to public immunoglobulin sequence data is whether our reference database is complete. This was supported by the fact that there was a substantial number of SNPs identified by LymAnalyzer that were not included in the IMGT database.

Indeed, several previous studies have suggested that the IMGT database is incomplete and might contain errors [220-222]. Thus in the second research chapter, we aimed to redefine a more comprehensive collection of TCR and immunoglobulin reference sequences. In Chapter 3, we described AlleleMiner, a novel bioinformatics pipeline that can be used to infer alleles based on variant calling data and the reference genome. Furthermore, we created the Lym1K database by using AlleleMiner with human variant information retrieved from the 1000 Genomes Project. Compared to IMGT, Lym1k presented a much more comprehensive set of TCR and immunoglobulin alleles, and the accuracy of Lym1k was further validated by its significant improvements of sequence alignment performance on real TCR and immunoglobulin datasets. In addition to that, we revealed that IMGT did not adequately cover the allelic diversity of immune receptor genes in global human populations, evident by worse alignment results in African populations than the Non-African populations. This further supported the need of expanding current TCR/IG reference database by integrating the immune receptor alleles inferred from human resequencing data in global populations.

Nonetheless, further improvements of the completeness and accuracy of Lym1K could be made based on further updates on the reference genome and larger variation calling datasets. For instance, there were 22 genes not included in the reference genome builds (CRCh 37, GRCh38). In addition to that, the sequencing coverage of several TRBV genes was unsatisfactory in the 1000 genomes project. Therefore, we made AlleleMiner user-adjustable and freely available online, so that it can be easily adapted to infer new alleles when more updated reference genome builds and variant calling data are available. Last, our studies suggested that the real size of the allelic variations of immune receptor genes, especially for immunoglobulins, might not be adequately represented from the variant calling information included in 1000 genomes project with "only" 2504 samples.

In the last research chapter, based on the construction of the more comprehensive collection of alleles of immune receptor genes from our previous work, some long-standing questions about the characteristics of germline allelic diversity of immune receptor genes could be properly addressed. We first asked if there was more allelic diversity in the immune receptor genes in African populations than in Non-African populations, as many studies have shown that the overall genetic diversity in African population is larger than the Non-African population combined [227]. Indeed, we found that the diversity index of immune receptor genes from IGH, IGL and TCRB is significantly larger in African population than the rest. We further hypothesized that there might be an association between the allelic diversity of immune receptor genes and their genomic locations, based on the distribution of immune receptor genes on the chromosome. However, we did not find evident correlations between genomic location and allelic diversity of immune receptor genes. Last, we asked if the more frequent selection of particular genes during the rearrangements would result in greater selection pressure, yielding higher allelic diversity. To test this, we used RNA expression level of different genes as the "proxy" to represent the frequency of their usage during somatic rearrangements, and conducted correlation analysis between the gene expression and allelic diversity. Surprisingly, we only found weak positive correlations in TRAV genes, and no significant correlations in all the other gene types.

However, we need to pay special attention to several confounding factors that might cause the absence of correlations. The first factor concerns the "proxy" we chose to represent the frequency of gene usage. Because gene expression might not adequately represent how often the gene is used during the selection process, given the fact that

the expression of some alleles of the genes is much lower than others [269]. Furthermore, although the size of the RNAseq dataset used in our analysis is very large, the source of the RNAseq data is from peripheral blood, which only includes approximately 2% of the overall immunoglobulin repertoire [204]. Thus further RNAseq experiments applied to lymphoid organs such as bone marrow and thymus are needed for comprehensive diversity analysis of immune receptor genes. Another reason that might explain the absence of correlation between gene expression and allelic diversity is that, over time, the frequency of the usage of different immune receptor genes may not be stable among populations. At the end, these problems could be eventually resolved by analyzing single cell RNAseq data from extensive coverage of immune cells from different lymphoid organs, from global populations. Although current sample sizes of most of the single cell RNAseq datasets range from 10^2 to 10^5, which is still unsatisfactory for describing the diversity of immune repertoire, it is increasing steadily every year with the decrease of sequencing costs.

Next generation sequencing has revolutionized immune repertoire analysis by allowing quantification of immunoglobulin and T cell receptor diversity. The interpretation of immune repertoires using LymAnalyzer and Lym1K could have medical applications in three major areas: tracking changes of the immune system in different diseases, therapeutic antibody design and the development of vaccines.

There are many immune-related diseases such as Type 1 diabetes, Leukemia, Rheumatoid arthritis etc. Deep sequencing captures a high-resolution snapshot of the immune repertoire, which includes important metrics such as diversity index for describing the status of the immune system. For instance, there is an overall decrease in diversity of the TCR repertoire during the course of HIV infection [270-272] and the immune repertoire is highly skewed in Leukemia [226]. The Lym1K database could provide a more comprehensive understanding of the genetic variations of TCR/IG genes and the utility of LymAnalyzer could enable more complete analysis of TCR/IG repertoires in different conditions, for example in ageing where repertoires become increasingly oligoclonal. The output file of clonotypes generated by LymAnalyzer could be easily annotated for downstream analysis such as repertoire diversity analysis, dominant clonotype analysis etc.

In addition to that, recombinant monoclonal antibody selection in vitro for specific antigens has become an important tool for generating effective therapeutic antibodies

for many diseases [273]. However, this method is relatively ineffective because it mainly focuses on analyzing a small number of randomly picked clonotypes and then applying low-throughput methods such as enzyme-linked immunosorbent assays [274]. By utilizing deep sequencing methods together with LymAnalyzer and the Lym1K database, we can obtain a view of the selected antibody population in unprecedented depth and in a more effective manner, facilitating the identification of antigen-specific antibodies.

Vaccination is one of the most effective and feasible approaches to prevent many infectious diseases. Nonetheless, there are still many infectious diseases for which no effective vaccine exists (such as AIDS, Malaria, Leishmaniasis etc) due to our constrained understandings about the dynamics of the immune response during these infections. Recent vaccination studies using high throughput methods often generate large amounts of sequencing data of the immune repertoire from blood samples at some well defined time-points [275]. LymAnalyzer could be applied for the analysis of post-vaccination immune responses at different time points using immune receptor sequencing data. In addition, recent studies have suggested the importance of determining the ontogeny of broadly-neutralizing antibodies for HIV[233, 236] The more comprehensive and accurate collection of genetic variants in immunoglobulin genes included in the Lym1K database could help these analyses and further facilitate antibody-guided vaccine development.

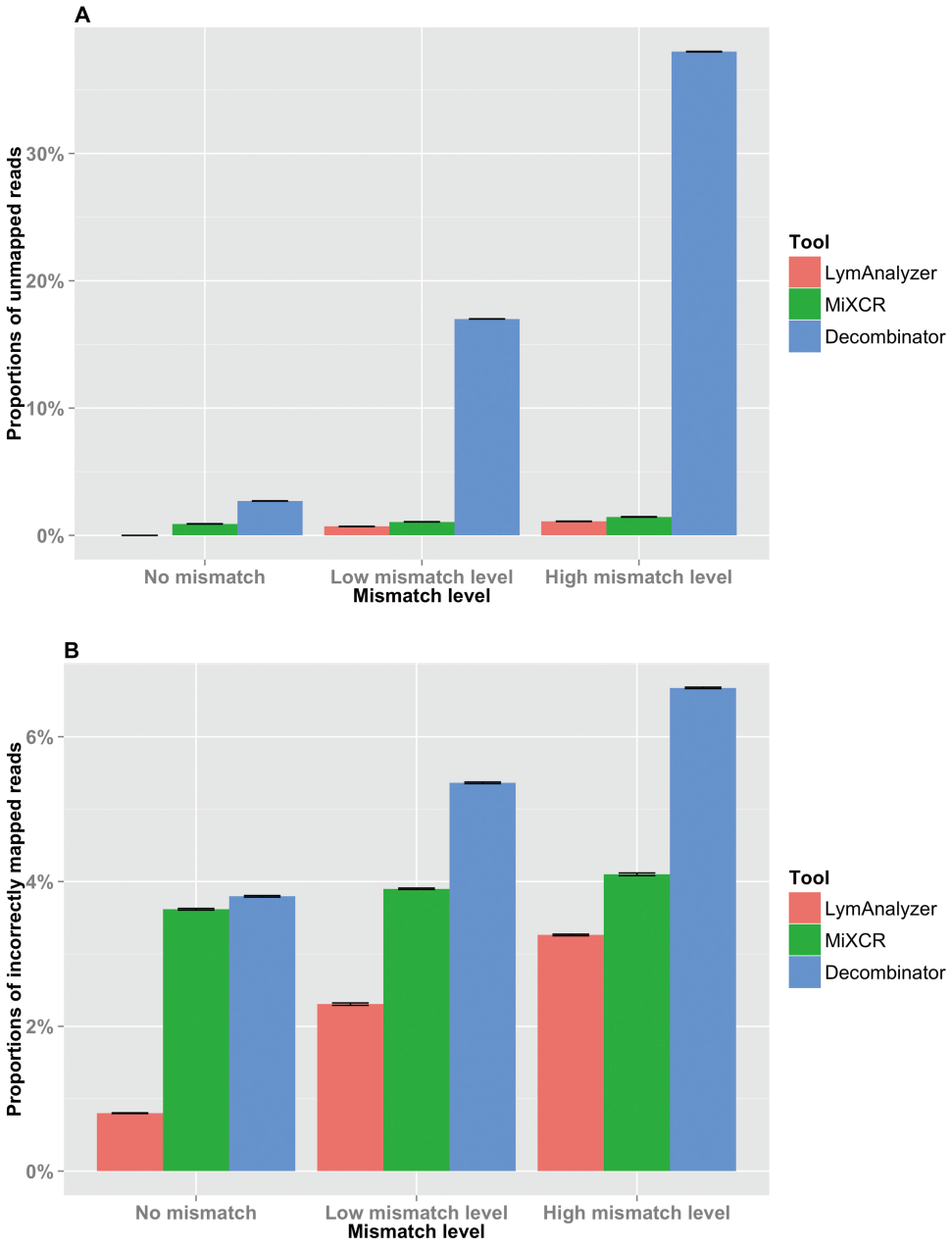
There are some biological implications and research questions that could be addressed by analyzing variant information of the immune genes included in the Lym1K database. We observed that allelic diversity of IGHV genes was much larger than the TRBV genes. This might be explained by the difference of the nature of epitopes that T and B cells recognize. B cell epitopes are generally conformational and immunoglobulins bind to them in free solution whereas T cell epitopes are linear and bound within the groove of the MHC. The variability of 3D conformational epitopes for antibodies might be much larger than the linear epitopes for TCR and over time, this may have lead to higher germline diversity in immunoglobulin genes. Moreover, another interesting observation we have described in the last research chapter is that there is no clear evidence showing that allelic diversity of the immune genes is related to the frequency of their usage during VDJ recombination. Based on this finding, one research question that needs further investigation is whether the selection of immune genes during recombination is stable over time. The distribution of the immune repertoire is

thought to be shaped by the selection pressure of distinctive pathogens in different environments [276, 277]. The different conditions such as hygiene between now and the past might present different pathogenic environments to the immune system and further alter the selection pressure.

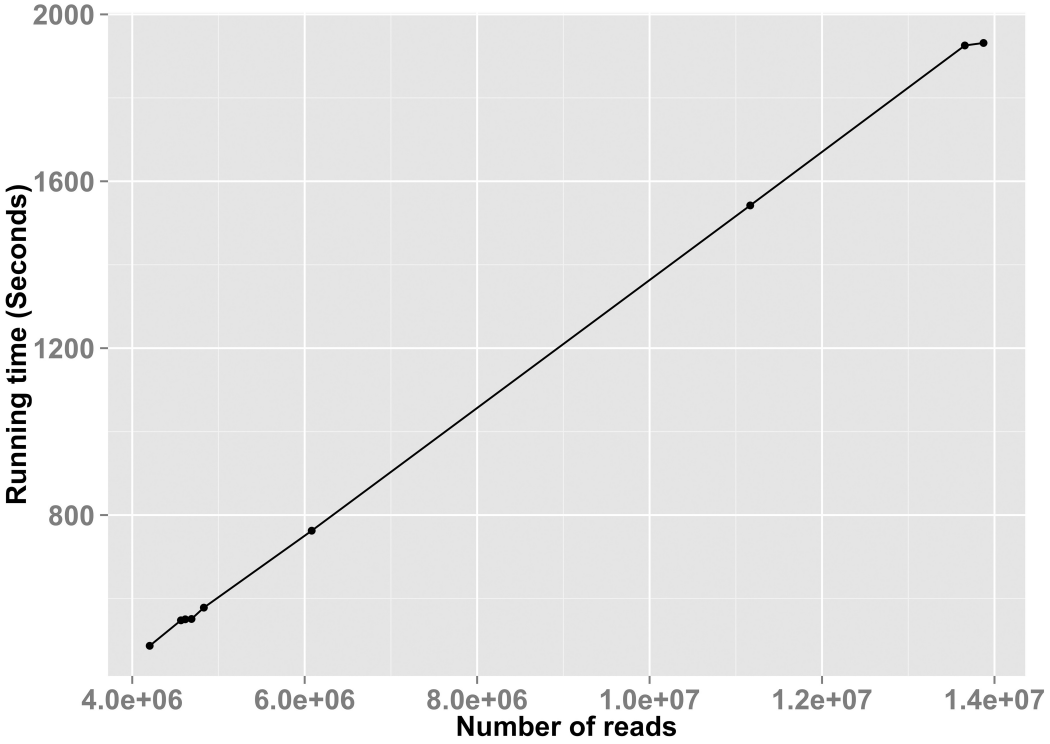
Both LymAnalyzer and Lym1K contain many flexible interfaces and adjustable parameters. Thus users can either directly use or tune the workflow when there are more variation data available (figure 3.8). Furthermore, since both tools are open-source software under the Apache license, expert users can continue tuning the algorithms or adding new features to the tools thereby providing more contributions to the immunology research community.

In summary, this thesis has demonstrated the utility of high-throughput approaches in analyzing immune receptor repertoires. More importantly, this thesis has illustrated a complete workflow of creating useful biological resources for the scientific community from large volumes of publicly available biological data, and beyond that, deriving interesting biological insights. This thesis sheds some light on the allelic diversity of human immune receptor genes, which could be used as a stepping stone for further investigations as more comprehensive datasets become available.

Appendix A

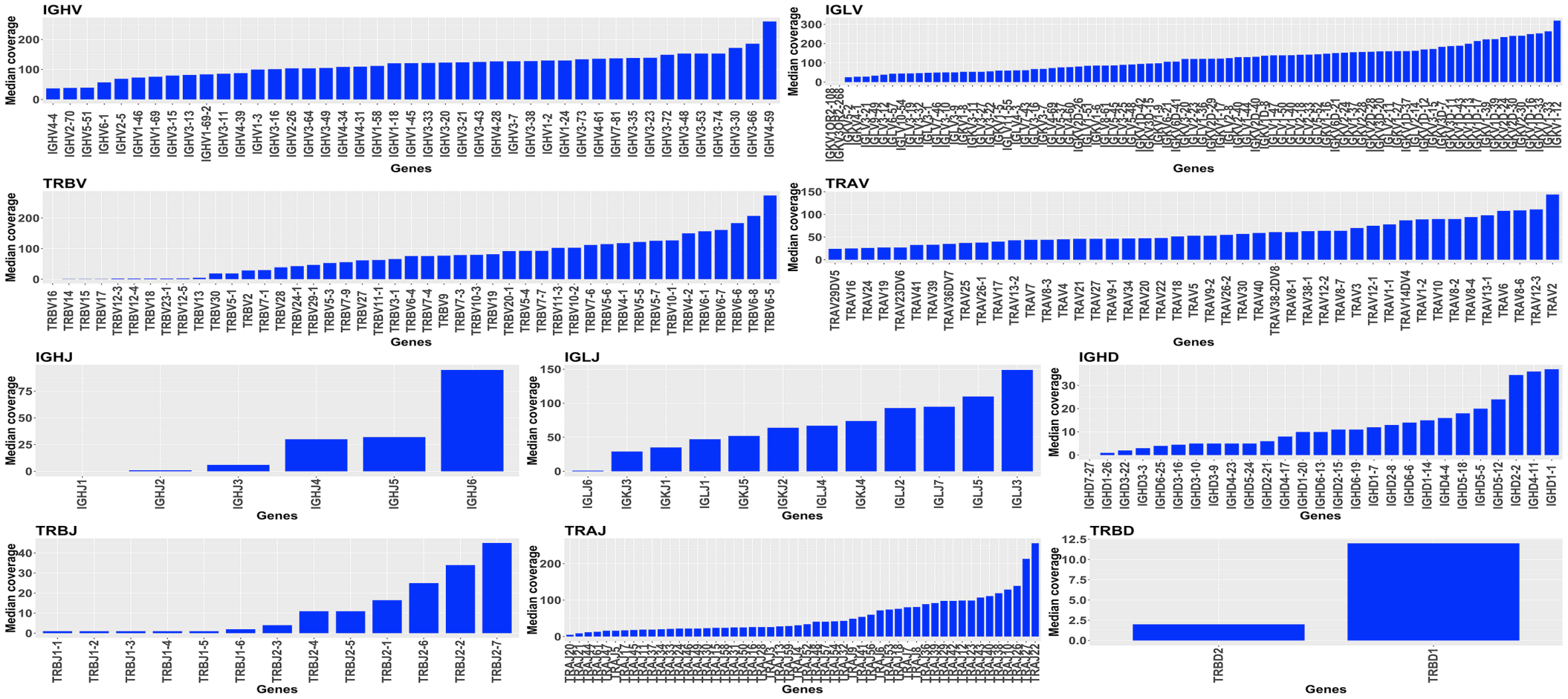


Supplementary figure A1. Comparisons of completeness and accuracy among LymAnalyzer, MiXCR and Decombinator based on simulated TCR data on the gene name level.

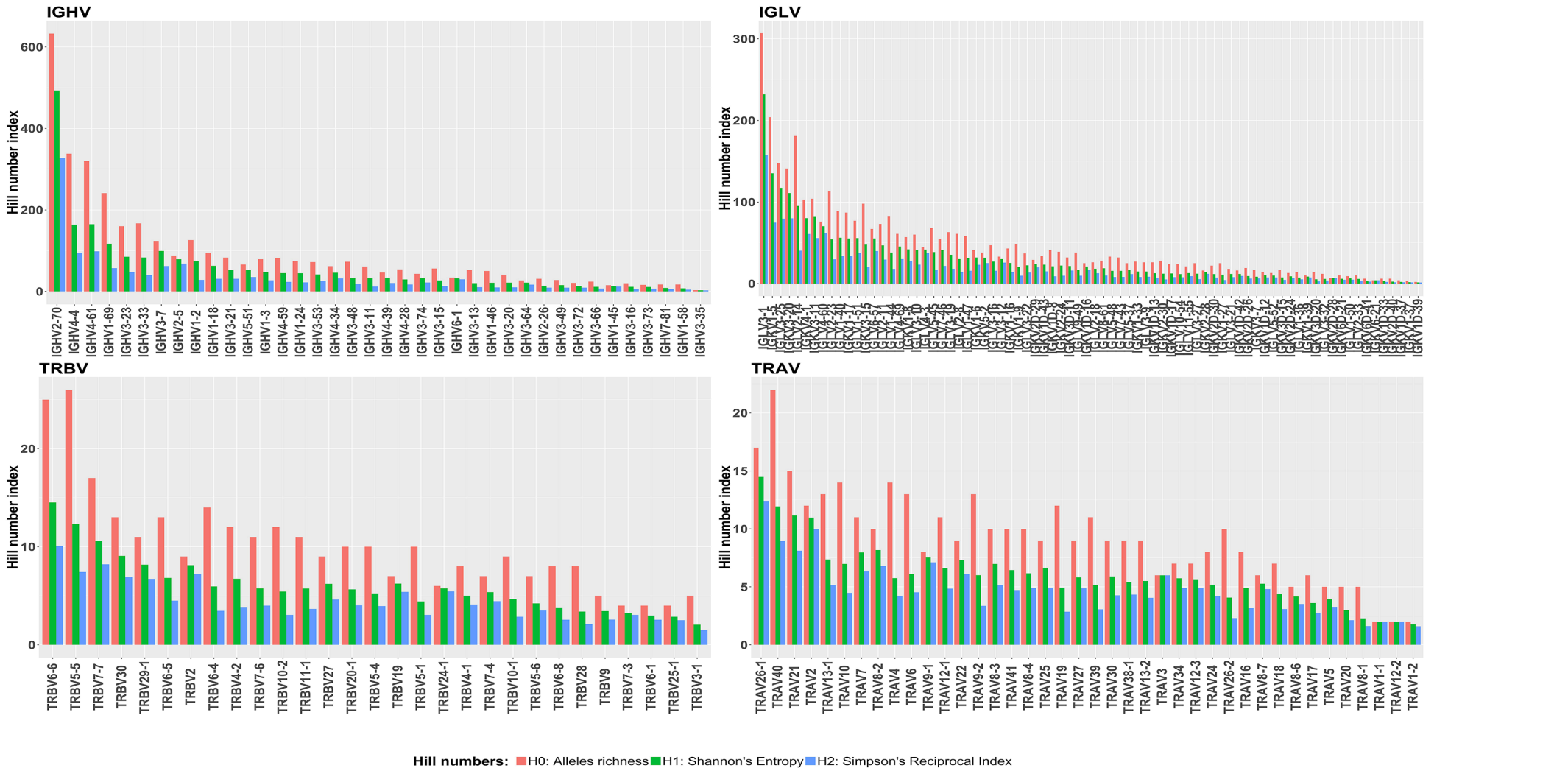


Supplementary figure A2. Running performance of LymAnalyzer.

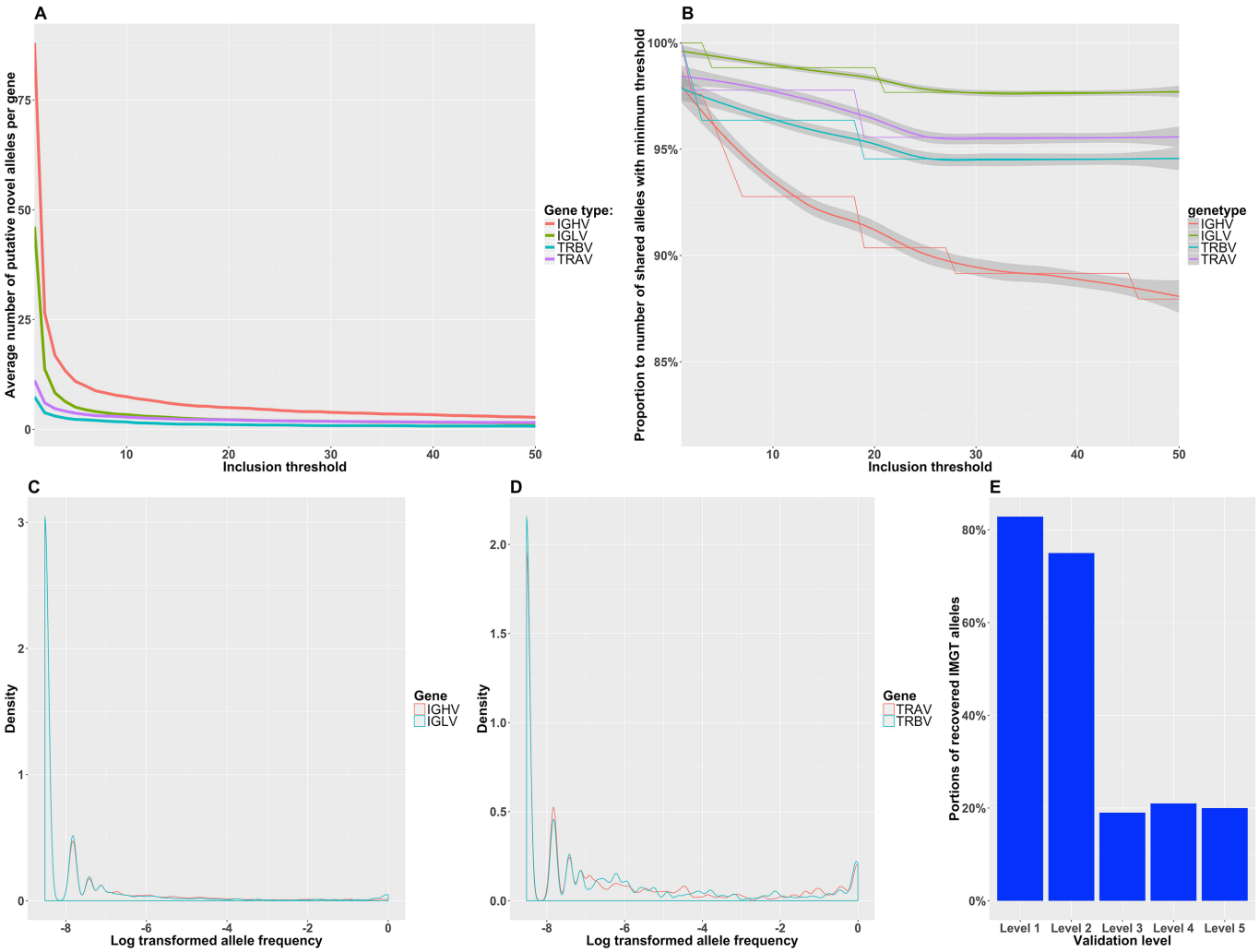
Appendix B



Supplementary figure B1. Median depth of TCR and IG genes in the resequencing data of 1000 genomes project.



Supplementary figure B2. Allelic diversity of IGHV, IGLV, TRBV and TRAV genes. For each gene, each type of allele was regarded as a species, and then the hill numbers were calculated by using Recon. We used Species richness, Shannon's Entropy and Simpson's Reciprocal index to illustrate the diversity of each gene.



Supplementary figure B3. The difference of inferred putative alleles comparing to IMGT alleles. A. The change of the average numbers of the putative novel alleles per gene with the adjustments of the inclusion threshold. B. The change of the numbers of the shared alleles with the adjustments of the inclusion threshold. The value on Y-axis is the number of the shared alleles proportion to the maximal shared alleles between Lym1K and IMGT (Inclusion threshold equal to one). The data is fitted by local polynomial regression, and the shaded area is the 95% confidence interval. C. Frequency distribution of the number of occurrences of IGHV and IGLV allele in 1000 genomes project. D. Frequency distribution of the number of occurrences of IGHV and IGLV allele in 1000 genomes project. The frequency in both C and D are naturally log transformed. E. Recovery rate of IMGT alleles with different validation levels.

Supplementary table B1. The shared alleles between IMGT and inferred alleles. The coordinates of some of the IMGT genes may vary from their coordinates in the genome assembly and a few IMGT alleles are only partial in the 5' or 3' end. One IMGT allele may correspond to multiple recovered alleles form G1K data. For IGHV, we used the validation classes introduced by Wang et al., the alleles were classified into five confidence levels from one (highest) to five (lowest) [220].

IGHV2-5*07	IGHV2-5*47p	Recovered
IGHV5-51*03	IGHV5-51*15p	Recovered
IGHV5-51*01	IGHV5-51*51p,IGHV5-51*21p,IGHV5-51*56p,IGHV5-51*62p	Recovered
IGHV3-48*03	IGHV3-48*20p,IGHV3-48*75p,IGHV3-48*21p,IGHV3-48*53p	Recovered
IGHV4-59*01	IGHV4-59*30p,IGHV4-59*27p,IGHV4-59*87p	Recovered
IGHV5-51*05	IGHV5-51*46p	Recovered
IGHV4-59*08	IGHV4-59*4p	Recovered
IGHV3-53*02	IGHV3-53*66p	Recovered
IGHV3-53*01	IGHV3-53*6p,IGHV3-53*13p	Recovered
IGHV3-73*02	IGHV3-73*11p	Recovered
IGHV3-74*02	IGHV3-74*8p	Recovered
IGHV3-74*01	IGHV3-74*15p	Recovered
IGHV3-73*01	IGHV3-73*1p	Recovered
IGHV4-34*02	IGHV4-34*25p	Recovered
IGHV4-34*03	IGHV4-34*42p	Recovered
IGHV3-30*18	IGHV3-30*1p	Recovered
IGHV4-34*01	IGHV4-34*1p	Recovered
IGHV1-3*01	IGHV1-3*3p,IGHV1-3*71p,IGHV1-3*4p	Recovered
IGHV1-3*02	IGHV1-3*49p,IGHV1-3*37p	Recovered
IGHV1-58*02	IGHV1-58*7p	Recovered
IGHV1-58*01	IGHV1-58*13p,IGHV1-58*2p	Recovered
IGHV3-49*03	IGHV3-49*6p	Recovered
IGHV3-49*04	IGHV3-49*10p	Recovered
IGHV3-49*05	IGHV3-49*32p	Recovered

IGHV3-33*01	IGHV3-33*134p,IGHV3-33*169p	Recovered
IGHV3-33*03	IGHV3-33*121p	Recovered
IGHV7-81*01	IGHV7-81*8p,IGHV7-81*19p	Recovered
IGHV1-24*01	IGHV1-24*73p,IGHV1-24*63p,IGHV1-24*48p,IGHV1-24*27p	Recovered
IGHV2-5*01	IGHV2-5*56p	Recovered
IGHV3-20*01	IGHV3-20*31p,IGHV3-20*41p	Recovered
IGHV2-5*05	IGHV2-5*74p	Recovered
IGHV2-5*02	IGHV2-5*55p,IGHV2-5*42p	Recovered
IGHV3-72*01	IGHV3-72*21p	Recovered
IGHV3-72*02	IGHV3-72*5p,IGHV3-72*1p	Recovered
IGHV2-70*11	IGHV2-70*66p,IGHV2-70*165p,IGHV2-70*290p,IGHV2-70*613p,IGHV2-70*491p	Recovered
IGHV1-2*04	IGHV1-2*66p,IGHV1-2*50p	Recovered
IGHV1-2*05	IGHV1-2*47p	Recovered
IGHV1-2*02	IGHV1-2*28p	Recovered
IGHV6-1*01	IGHV6-1*26p	Recovered
IGHV4-61*01	IGHV4-61*286p	Recovered
IGHV4-61*07	IGHV4-61*104p,IGHV4-61*286p,IGHV4-61*117p,IGHV4-61*210p,IGHV4-61*12p,IGHV4-61*49p,IGHV4-61*322p,IGHV4-61*55p	Recovered
IGHV3-15*07	IGHV3-15*47p	Recovered
IGHV3-15*01	IGHV3-15*14p,IGHV3-15*45p	Recovered
IGHV3-13*01	IGHV3-13*5p	Recovered
IGHV3-13*03	IGHV3-13*14p	Recovered
IGHV3-13*04	IGHV3-13*58p	Recovered
IGHV2-26*01	IGHV2-26*6p	Recovered
IGHV3-30*03	IGHV3-30*1p	Recovered
IGHV3-64*02	IGHV3-64*15p	Recovered
IGHV3-64*01	IGHV3-64*17p	Recovered
IGHV4-4*02	IGHV4-4*235p	Recovered
IGHV2-70*01	IGHV2-70*366p,IGHV2-70*378p,IGHV2-70*619p,IGHV2-70*263p,IGHV2-70*315p	Recovered
IGHV3-7*03	IGHV3-7*34p,IGHV3-7*21p	Recovered

IGHV3-7*02	IGHV3-7*77p,IGHV3-7*8p	Recovered
IGHV3-7*01	IGHV3-7*57p,IGHV3-7*129p,IGHV3-7*113p	Recovered
IGHV3-35*01	IGHV3-35*3p,IGHV3-35*1p	Recovered
IGHV3-66*03	IGHV3-66*4p	Recovered
IGHV4-39*01	IGHV4-39*20p	Recovered
IGHV4-39*04	IGHV4-39*29p,IGHV4-39*31p	Recovered
IGHV3-21*01	IGHV3-21*25p,IGHV3-21*23p	Recovered
IGHV3-21*04	IGHV3-21*43p	Recovered
IGHV3-21*03	IGHV3-21*38p	Recovered
IGHV1-46*01	IGHV1-46*44p,IGHV1-46*35p,IGHV1-46*11p,IGHV1-46*24p,IGHV1-46*5p	Recovered
IGHV1-46*03	IGHV1-46*31p	Recovered
IGHV1-45*02	IGHV1-45*14p,IGHV1-45*1p,IGHV1-45*3p	Recovered
IGHV1-45*03	IGHV1-45*15p,IGHV1-45*8p	Recovered
IGHV3-11*05	IGHV3-11*57p,IGHV3-11*66p,IGHV3-11*51p	Recovered
IGHV1-69*07	IGHV1-69*138p,IGHV1-69*234p,IGHV1-69*44p,IGHV1-69*43p,IGHV1-69*245p	Recovered
IGHV1-69*05	IGHV1-69*217p	Recovered
IGHV1-69*06	IGHV1-69*150p	Recovered
IGHV1-18*03	IGHV1-18*48p	Recovered
IGHV3-43*01	IGHV3-43*1p	Recovered
IGHV1-69*01	IGHV1-69*251p	Recovered
IGHV4-28*02	IGHV4-28*15p,IGHV4-28*13p	Recovered
IGHV1-18*01	IGHV1-18*1p,IGHV1-18*44p	Recovered
IGHV4-28*01	IGHV4-28*50p,IGHV4-28*51p,IGHV4-28*24p,IGHV4-28*18p	Recovered
IGHV4-28*05	IGHV4-28*30p,IGHV4-28*19p,IGHV4-28*28p	Recovered
IGHV3-23*01	IGHV3-23*101p,IGHV3-23*16p,IGHV3-23*32p	Recovered
IGHV1-69*12	IGHV1-69*1p,IGHV1-69*80p	Recovered
IGHV3-23*04	IGHV3-23*63p	Recovered
IGHV1-69*13	IGHV1-69*177p,IGHV1-69*255p	Recovered
IGHV3-23*03	IGHV3-23*125p	Recovered
IGHV1-69*11	IGHV1-69*15p	Recovered
IGHV2-5*06	NA	Validation level 5

IGHV2-5*08	NA	Validation level 5
IGHV2-5*09	NA	Validation level 5
IGHV3-48*01	NA	Validation level 2
IGHV5-51*02	NA	Validation level 3
IGHV3-48*04	NA	Not recovered
IGHV3-48*02	NA	Validation level 2
IGHV4-59*04	NA	Validation level 3
IGHV4-59*03	NA	Validation level 5
IGHV1-8*02	NA	Not in GRCh38
IGHV4-59*02	NA	Validation level 4
IGHV1-8*01	NA	Not in GRCh38
IGHV5-51*04	NA	Validation level 4
IGHV4-59*09	NA	Validation level 5
IGHV2-5*10	NA	Not recovered
IGHV4-59*07	NA	Validation level 5
IGHV4-59*06	NA	Validation level 5
IGHV4-59*05	NA	Validation level 5
IGHV1-69-2*01	NA	Not in GRCh37
IGHV1-69-2*02	NA	Not in GRCh37
IGHV4-59*10	NA	Validation level 3
IGHV4-30-4*06	NA	Not in GRCh38
IGHV3-53*03	NA	Validation level 3
IGHV4-30-4*05	NA	Not in GRCh38
IGHV3-53*04	NA	Not recovered
IGHV4-30-4*02	NA	Not in GRCh38
IGHV4-30-4*01	NA	Not in GRCh38
IGHV4-30-4*04	NA	Not in GRCh38
IGHV4-30-4*03	NA	Not in GRCh38
IGHV3-74*03	NA	Validation level 1

IGHV4-34*04	NA	Validation level 5
IGHV4-34*05	NA	Validation level 5
IGHV4-34*08	NA	Validation level 5
IGHV4-34*09	NA	Validation level 3
IGHV4-34*06	NA	Validation level 5
IGHV4-34*07	NA	Validation level 5
IGHV3-30*19	NA	Not recovered
IGHV3-30*15	NA	Not recovered
IGHV3-30*14	NA	Not recovered
IGHV3-30*17	NA	Not recovered
IGHV3-30*16	NA	Not recovered
IGHV3-30*11	NA	Not recovered
IGHV3-30*10	NA	Not recovered
IGHV3-30*13	NA	Not recovered
IGHV3-30*12	NA	Not recovered
IGHV3-49*01	NA	Validation level 3
IGHV3-49*02	NA	Validation level 5
IGHV4-34*13	NA	Validation level 5
IGHV4-31*10	NA	Not recovered
IGHV4-34*10	NA	Validation level 3
IGHV4-34*11	NA	Validation level 3
IGHV4-34*12	NA	Validation level 4
IGHV4-31*06	NA	Not recovered
IGHV4-31*05	NA	Not recovered
IGHV4-31*04	NA	Not recovered
IGHV4-31*03	NA	Not recovered
IGHV4-31*02	NA	Not recovered
IGHV4-30-2*04	NA	Not in GRCh37
IGHV4-31*01	NA	Not recovered
IGHV4-30-2*03	NA	Not in GRCh37

IGHV4-30-2*05	NA	Not in GRCh37
IGHV3-33*05	NA	Validation level 3
IGHV3-33*04	NA	Validation level 5
IGHV4-30-2*02	NA	Not in GRCh37
IGHV3-33*02	NA	Validation level 3
IGHV4-30-2*01	NA	Not in GRCh37
IGHV3-33*06	NA	Not recovered
IGHV2-5*04	NA	Validation level 5
IGHV2-5*03	NA	Validation level 5
IGHV4-31*09	NA	Not recovered
IGHV4-31*07	NA	Not recovered
IGHV4-31*08	NA	Not recovered
IGHV2-70*10	NA	Validation level 5
IGHV2-70*12	NA	Validation level 3
IGHV2-70*13	NA	Validation level 4
IGHV1-2*03	NA	Validation level 5
IGHV6-1*02	NA	Validation level 5
IGHV4-61*02	NA	Validation level 1
IGHV4-61*03	NA	Validation level 5
IGHV4-61*05	NA	Validation level 5
IGHV4-61*04	NA	Validation level 5
IGHV3-15*06	NA	Validation level 5
IGHV3-15*08	NA	Validation level 5
IGHV4-61*08	NA	Validation level 4
IGHV3-15*02	NA	Validation level 4
IGHV3-15*03	NA	Validation level 5
IGHV3-NL1*01	NA	Not in GRCh38
IGHV3-15*04	NA	Validation level 5

IGHV3-15*05	NA	Validation level 5
IGHV3-13*02	NA	Validation level 3
IGHV3-30*01	NA	Not recovered
IGHV3-64*04	NA	Validation level 5
IGHV3-30*02	NA	Not recovered
IGHV3-64*05	NA	Validation level 5
IGHV3-30*04	NA	Not recovered
IGHV3-64*03	NA	Validation level 5
IGHV3-30*05	NA	Not recovered
IGHV4-b*02	NA	Not in GRCh38
IGHV3-30*06	NA	Not recovered
IGHV4-b*01	NA	Not in GRCh38
IGHV3-30*08	NA	Not recovered
IGHV3-30*07	NA	Not recovered
IGHV3-30*09	NA	Not recovered
IGHV4-4*05	NA	Validation level 5
IGHV4-4*06	NA	Validation level 5
IGHV4-4*03	NA	Validation level 5
IGHV4-4*04	NA	Validation level 5
IGHV4-4*01	NA	Validation level 2
IGHV2-70*02	NA	Validation level 5
IGHV2-70*03	NA	Validation level 3
IGHV4-4*07	NA	Validation level 1
IGHV3-66*01	NA	Validation level 1
IGHV3-9*02	NA	Not in GRCh38
IGHV3-9*01	NA	Not in GRCh38
IGHV2-70*08	NA	Validation level 5
IGHV2-70*07	NA	Validation level 5
IGHV3-66*04	NA	Validation level

		4
IGHV2-70*06	NA	Validation level 5
IGHV2-70*05	NA	Validation level 5
IGHV1-2*01	NA	Validation level 2
IGHV3-66*02	NA	Validation level 3
IGHV2-70*04	NA	Validation level 5
IGHV4-39*02	NA	Validation level 4
IGHV4-39*06	NA	Validation level 5
IGHV4-39*05	NA	Validation level 5
IGHV4-39*03	NA	Validation level 5
IGHV4-39*07	NA	Validation level 1
IGHV5-a*01	NA	Not in GRCh38
IGHV5-a*04	NA	Not in GRCh38
IGHV5-a*03	NA	Not in GRCh38
IGHV3-21*02	NA	Validation level 4
IGHV1-46*02	NA	Validation level 4
IGHV3-30-3*01	NA	Not in GRCh38
IGHV3-30-3*02	NA	Not in GRCh38
IGHV1-45*01	NA	Validation level 2
IGHV3-11*01	NA	Validation level 1
IGHV7-4-1*03	NA	Not in GRCh38
IGHV7-4-1*02	NA	Not in GRCh38
IGHV7-4-1*05	NA	Not in GRCh38
IGHV7-4-1*04	NA	Not in GRCh38
IGHV3-11*04	NA	Not recovered
IGHV7-4-1*01	NA	Not in GRCh38
IGHV3-11*03	NA	Validation level 4
IGHV1-69*08	NA	Validation level 4
IGHV1-69*09	NA	Validation level

		2
IGHV1-69*03	NA	Validation level 5
IGHV1-69*04	NA	Validation level 2
IGHV3-43*02	NA	Not recovered
IGHV1-69*02	NA	Validation level 2
IGHV1-18*02	NA	Validation level 5
IGHV4-28*06	NA	Not recovered
IGHV4-28*04	NA	Validation level 5
IGHV4-28*03	NA	Validation level 5
IGHV3-23*02	NA	Validation level 3
IGHV1-69*10	NA	Validation level 3
IGHV3-23*05	NA	Validation level 3
IGLV4-60*03	IGLV4-60*71p,IGLV4-60*65p,IGLV4-60*57p,IGLV4-60*83p	Recovered
IGLV4-60*02	IGLV4-60*77p	Recovered
IGLV4-69*01	IGLV4-69*18p,IGLV4-69*53p,IGLV4-69*33p	Recovered
IGLV3-10*01	IGLV3-10*54p,IGLV3-10*27p	Recovered
IGLV3-32*01	IGLV3-32*13p,IGLV3-32*2p	Recovered
IGLV11-55*01	IGLV11-55*8p,IGLV11-55*26p	Recovered
IGLV1-50*01	IGLV1-50*8p	Recovered
IGLV2-18*02	IGLV2-18*30p,IGLV2-18*28p	Recovered
IGLV2-18*01	IGLV2-18*21p,IGLV2-18*29p	Recovered
IGLV6-57*01	IGLV6-57*26p	Recovered
IGLV1-47*02	IGLV1-47*14p	Recovered
IGLV1-47*01	IGLV1-47*40p,IGLV1-47*50p	Recovered
IGLV5-45*03	IGLV5-45*39p,IGLV5-45*20p	Recovered
IGLV5-45*01	IGLV5-45*42p	Recovered
IGLV5-45*02	IGLV5-45*5p,IGLV5-45*64p	Recovered
IGLV3-12*01	IGLV3-12*9p,IGLV3-12*14p	Recovered
IGLV3-12*02	IGLV3-12*11p,IGLV3-12*12p	Recovered
IGLV8-61*01	IGLV8-61*5p	Recovered
IGLV8-61*02	IGLV8-61*15p	Recovered

IGLV2-14*01	IGLV2-14*84p,IGLV2-14*28p	Recovered
IGLV2-11*01	IGLV2-11*15p,IGLV2-11*67p,IGLV2-11*5p,IGLV2-11*71p,IGLV2-11*74p	Recovered
IGLV7-46*02	IGLV7-46*8p	Recovered
IGLV7-46*01	IGLV7-46*1p	Recovered
IGLV3-22*01	IGLV3-22*22p	Recovered
IGLV5-37*01	IGLV5-37*9p,IGLV5-37*11p,IGLV5-37*23p	Recovered
IGLV2-33*01	IGLV2-33*11p,IGLV2-33*9p	Recovered
IGLV1-36*01	IGLV1-36*13p	Recovered
IGLV3-1*01	IGLV3-1*214p,IGLV3-1*196p,IGLV3-1*12p,IGLV3-1*266p,IGLV3-1*59p,IGLV3-1*164p,IGLV3-1*97p,IGLV3-1*199p,IGLV3-1*285p,IGLV3-1*146p,IGLV3-1*20p	Recovered
IGLV1-51*02	IGLV1-51*11p	Recovered
IGLV1-44*01	IGLV1-44*12p	Recovered
IGLV1-51*01	IGLV1-51*34p	Recovered
IGLV2-8*01	IGLV2-8*43p,IGLV2-8*8p	Recovered
IGLV1-40*01	IGLV1-40*23p,IGLV1-40*9p	Recovered
IGLV9-49*01	IGLV9-49*16p,IGLV9-49*22p	Recovered
IGLV9-49*02	IGLV9-49*32p	Recovered
IGLV3-9*01	IGLV3-9*15p,IGLV3-9*10p	Recovered
IGLV7-43*01	IGLV7-43*15p,IGLV7-43*9p	Recovered
IGLV10-54*02	IGLV10-54*26p	Recovered
IGLV10-54*01	IGLV10-54*8p,IGLV10-54*19p	Recovered
IGLV5-52*01	IGLV5-52*3p,IGLV5-52*7p	Recovered
IGLV3-27*01	IGLV3-27*14p	Recovered
IGLV4-3*01	IGLV4-3*19p,IGLV4-3*46p,IGLV4-3*37p	Recovered
IGLV4-60*01	NA	Not recovered
IGLV4-69*02	NA	Not recovered
IGLV3-10*02	NA	Not recovered
IGLV3-16*01	NA	Not recovered
IGLV2-18*04	NA	Not recovered
IGLV2-18*03	NA	Not recovered
IGLV3-19*01	NA	Not recovered
IGLV8-61*03	NA	Not recovered

IGLV2-14*03	NA	Not recovered
IGLV2-23*01	NA	Not recovered
IGLV2-14*02	NA	Not recovered
IGLV2-23*02	NA	Not recovered
IGLV2-11*03	NA	Not recovered
IGLV2-23*03	NA	Not recovered
IGLV2-14*04	NA	Not recovered
IGLV2-11*02	NA	Not recovered
IGLV2-33*03	NA	Not recovered
IGLV5-48*01	NA	Not recovered
IGLV2-8*02	NA	Not recovered
IGLV2-8*03	NA	Not recovered
IGLV1-40*03	NA	Not recovered
IGLV9-49*03	NA	Not recovered
IGLV1-40*02	NA	Not recovered
IGLV3-9*02	NA	Not recovered
IGLV3-21*01	NA	Not recovered
IGLV3-21*02	NA	Not recovered
IGLV3-21*03	NA	Not recovered
IGLV10-54*03	NA	Not recovered
IGLV5-39*02	NA	Not in GRCh38
IGLV5-39*01	NA	Not in GRCh38
IGKV3D-20*01	IGKV3D-20*11p,IGKV3D-20*14p	Recovered
IGKV1-8*01	IGKV1-8*51p,IGKV1-8*47p	Recovered
IGKV2D-26*01	IGKV2D-26*16p	Recovered
IGKV1-9*01	IGKV1-9*24p,IGKV1-9*4p	Recovered
IGKV1-5*03	IGKV1-5*32p,IGKV1-5*28p,IGKV1-5*64p,IGKV1-5*147p	Recovered
IGKV2D-24*01	IGKV2D-24*13p	Recovered
IGKV1-5*01	IGKV1-5*197p	Recovered
IGKV2D-30*01	IGKV2D-30*13p	Recovered
IGKV1D-42*01	IGKV1D-42*12p,IGKV1D-42*8p,IGKV1D-42*9p	Recovered
IGKV4-1*01	IGKV4-1*90p,IGKV4-1*9p,IGKV4-1*82p,IGKV4-1*17p,IGKV4-1*84p,IGKV4-1*20p,IGKV4-1*38p	Recovered

IGKV2-30*01	IGKV2-30*25p,IGKV2-30*18p	Recovered
IGKV2-30*02	IGKV2-30*3p,IGKV2-30*29p,IGKV2-30*17p	Recovered
IGKV1-17*01	IGKV1-17*38p	Recovered
IGKV1D-16*02	IGKV1D-16*2p	Recovered
IGKV1D-16*01	IGKV1D-16*9p,IGKV1D-16*13p,IGKV1D-16*27p	Recovered
IGKV1-12*01	IGKV1-12*8p,IGKV1-12*34p	Recovered
IGKV2-24*01	IGKV2-24*18p,IGKV2-24*29p,IGKV2-24*8p	Recovered
IGKV5-2*01	IGKV5-2*27p,IGKV5-2*6p,IGKV5-2*16p	Recovered
IGKV1-33*01	IGKV1-33*15p,IGKV1-33*17p,IGKV1-33*25p	Recovered
IGKV3-20*01	IGKV3-20*56p,IGKV3-20*62p,IGKV3-20*92p,IGKV3-20*76p,IGKV3-20*113p,IGKV3-20*98p,IGKV3-20*55p,IGKV3-20*41p,IGKV3-20*107p,IGKV3-20*85p,IGKV3-20*129p,IGKV3-20*83p,IGKV3-20*12p,IGKV3-20*45p,IGKV3-20*79p,IGKV3-20*114p,IGKV3-20*19p,IGKV3-20*2p	Recovered
IGKV3-11*01	IGKV3-11*47p,IGKV3-11*97p,IGKV3-11*105p,IGKV3-11*39p,IGKV3-11*106p,IGKV3-11*17p,IGKV3-11*22p,IGKV3-11*3p,IGKV3-11*34p,IGKV3-11*101p,IGKV3-11*66p	Recovered
IGKV1D-12*02	IGKV1D-12*10p,IGKV1D-12*9p,IGKV1D-12*15p,IGKV1D-12*12p	Recovered
IGKV2-40*01	IGKV2-40*1p	Recovered
IGKV1-39*01	IGKV1-39*4p	Recovered
IGKV1D-17*01	IGKV1D-17*1p	Recovered
IGKV2D-40*01	IGKV2D-40*2p,IGKV2D-40*8p	Recovered
IGKV1D-33*01	IGKV1D-33*4p	Recovered
IGKV1D-37*01	IGKV1D-37*1p	Recovered
IGKV1D-39*01	IGKV1D-39*1p	Recovered
IGKV2D-28*01	IGKV2D-28*3p	Recovered
IGKV1D-43*01	IGKV1D-43*18p,IGKV1D-43*36p,IGKV1D-43*6p	Recovered
IGKV3-7*04	IGKV3-7*15p,IGKV3-7*14p,IGKV3-7*9p	Recovered
IGKV1D-8*01	IGKV1D-8*40p,IGKV1D-8*4p,IGKV1D-8*43p,IGKV1D-8*16p	Recovered
IGKV2D-29*01	IGKV2D-29*28p,IGKV2D-29*14p,IGKV2D-29*30p,IGKV2D-29*12p	Recovered
IGKV1-16*02	IGKV1-16*8p,IGKV1-16*42p	Recovered

IGKV1-37*01	IGKV1-37*2p	Recovered
IGKV1-27*01	IGKV1-27*12p,IGKV1-27*9p	Recovered
IGKV2-28*01	IGKV2-28*2p,IGKV2-28*13p	Recovered
IGKV3D-11*01	IGKV3D-11*27p,IGKV3D-11*18p	Recovered
IGKV3-15*01	IGKV3-15*14p,IGKV3-15*66p,IGKV3-15*36p,IGKV3-15*100p,IGKV3-15*62p,IGKV3-15*23p	Recovered
IGKV3D-15*01	IGKV3D-15*20p	Recovered
IGKV6-21*01	IGKV6-21*5p,IGKV6-21*3p	Recovered
IGKV3D-7*01	IGKV3D-7*1p	Recovered
IGKV2D-26*02	NA	Not recovered
IGKV1-5*02	NA	Not recovered
IGKV6D-41*01	NA	Not recovered
IGKV1-17*02	NA	Not recovered
IGKV1-12*02	NA	Not recovered
IGKV3/OR2-268*01	NA	Not in GRCh38
IGKV3/OR2-268*02	NA	Not in GRCh38
IGKV3-20*02	NA	Not recovered
IGKV1-NL1*01	NA	Not in GRCh38
IGKV3-11*02	NA	Not recovered
IGKV1D-12*01	NA	Not recovered
IGKV2-40*02	NA	Not recovered
IGKV1D-17*02	NA	Not recovered
IGKV6D-21*01	NA	Not recovered
IGKV1-6*01	NA	Not recovered
IGKV1D-13*01	NA	Not recovered
IGKV3-7*01	NA	Not recovered
IGKV3-7*03	NA	Not recovered
IGKV3-7*02	NA	Not recovered
IGKV1-13*02	NA	Not recovered
IGKV1-16*01	NA	Not recovered
IGKV2D-29*02	NA	Not recovered
TRBV5-6*01	TRBV5-6*12p	Recovered
TRBV5-4*01	TRBV5-4*1p,TRBV5-4*13p	Recovered
TRBV13*01	TRBV13*1p	Recovered

TRBV25-1*01	TRBV25-1*4p	Recovered
TRBV18*01	TRBV18*1p	Recovered
TRBV5-5*02	TRBV5-5*12p	Recovered
TRBV5-5*01	TRBV5-5*19p,TRBV5-5*30p,TRBV5-5*16p,TRBV5-5*11p	Recovered
TRBV14*01	TRBV14*1p	Recovered
TRBV7-6*01	TRBV7-6*5p,TRBV7-6*7p	Recovered
TRBV6-8*01	TRBV6-8*4p,TRBV6-8*9p	Recovered
TRBV10-2*02	TRBV10-2*9p,TRBV10-2*10p	Recovered
TRBV9*01	TRBV9*7p	Recovered
TRBV4-2*01	TRBV4-2*9p	Recovered
TRBV9*02	TRBV9*3p	Recovered
TRBV6-2*01	TRBV6-2*1p	Recovered
TRBV10-2*01	TRBV10-2*3p,TRBV10-2*7p	Recovered
TRBV6-4*01	TRBV6-4*15p	Recovered
TRBV6-4*02	TRBV6-4*14p	Recovered
TRBV11-1*01	TRBV11-1*10p	Recovered
TRBV3-1*01	TRBV3-1*2p	Recovered
TRBV15*02	TRBV15*1p	Recovered
TRBV15*01	TRBV15*1p	Recovered
TRBV24-1*01	TRBV24-1*3p	Recovered
TRBV5-1*01	TRBV5-1*9p,TRBV5-1*8p	Recovered
TRBV4-1*01	TRBV4-1*8p	Recovered
TRBV6-5*01	TRBV6-5*13p,TRBV6-5*6p,TRBV6-5*3p	Recovered
TRBV28*01	TRBV28*6p,TRBV28*1p	Recovered
TRBV6-1*01	TRBV6-1*3p	Recovered
TRBV16*01	TRBV16*1p	Recovered
TRBV27*01	TRBV27*9p	Recovered
TRBV7-3*01	TRBV7-3*4p	Recovered
TRBV30*04	TRBV30*7p	Recovered
TRBV20-1*05	TRBV20-1*4p	Recovered
TRBV20-1*02	TRBV20-1*11p	Recovered
TRBV20-1*01	TRBV20-1*8p	Recovered
TRBV19*03	TRBV19*10p,TRBV19*2p	Recovered

TRBV12-3*01	TRBV12-3*1p	Recovered
TRBV19*01	TRBV19*9p	Recovered
TRBV7-7*01	TRBV7-7*3p,TRBV7-7*16p	Recovered
TRBV30*02	TRBV30*17p	Recovered
TRBV30*01	TRBV30*4p	Recovered
TRBV10-1*02	TRBV10-1*6p,TRBV10-1*10p	Recovered
TRBV10-1*01	TRBV10-1*2p	Recovered
TRBV6-6*02	TRBV6-6*29p,TRBV6-6*8p	Recovered
TRBV7-4*01	TRBV7-4*10p,TRBV7-4*1p	Recovered
TRBV11-2*03	TRBV11-2*6p	Recovered
TRBV6-6*01	TRBV6-6*25p	Recovered
TRBV6-6*03	TRBV6-6*16p	Recovered
TRBV2*02	TRBV2*7p	Recovered
TRBV2*01	TRBV2*6p	Recovered
TRBV11-2*01	TRBV11-2*5p	Recovered
TRBV12-4*01	TRBV12-4*1p	Recovered
TRBV29-1*01	TRBV29-1*2p	Recovered
TRBV12-5*01	TRBV12-5*1p	Recovered
TRBV13*02	NA	Non-germline sequence
TRBV14*02	NA	Non-germline sequence
TRBV6-6*05	NA	Non-germline sequence
TRBV7-6*02	NA	Non-germline sequence
TRBV9*03	NA	Non-germline sequence
TRBV4-2*02	NA	Non-germline sequence
TRBV5-4*04	NA	Non-germline sequence
TRBV5-4*02	NA	Non-germline sequence
TRBV5-4*03	NA	Non-germline sequence
TRBV3-1*02	NA	Non-germline sequence
TRBV15*03	NA	Non-germline sequence
TRBV4-1*02	NA	Non-germline sequence

TRBV5-1*02	NA	Non-germline sequence
TRBV6-9*01	NA	Not in GRCh38
TRBV6-3*01	NA	Not in GRCh38
TRBV7-3*05	NA	Non-germline sequence
TRBV7-3*04	NA	Non-germline sequence
TRBV20-1*06	NA	Non-germline sequence
TRBV20-1*07	NA	Non-germline sequence
TRBV30*05	NA	Non-germline sequence
TRBV20-1*04	NA	Non-germline sequence
TRBV20-1*03	NA	Non-germline sequence
TRBV16*03	NA	Non-germline sequence
TRBV7-2*01	NA	Not in GRCh38
TRBV19*02	NA	Not recovered
TRBV7-7*02	NA	Non-germline sequence
TRBV10-3*04	NA	Not in GRCh37
TRBV10-3*03	NA	Not in GRCh37
TRBV10-3*02	NA	Not in GRCh37
TRBV10-3*01	NA	Not in GRCh37
TRBV7-2*04	NA	Not in GRCh37
TRBV7-2*02	NA	Not in GRCh37
TRBV7-2*03	NA	Not in GRCh37
TRBV11-3*03	NA	Not in GRCh37
TRBV11-3*02	NA	Not in GRCh37
TRBV11-3*01	NA	Not in GRCh37
TRBV4-3*02	NA	Not in GRCh38
TRBV4-3*03	NA	Not in GRCh38
TRBV4-3*01	NA	Not in GRCh38
TRBV4-3*04	NA	Not in GRCh38
TRBV6-6*04	NA	Non-germline sequence
TRBV5-8*02	NA	Not in GRCh38

TRBV5-8*01	NA	Not in GRCh38
TRBV2*03	NA	Non-germline sequence
TRBV11-2*02	NA	Non-germline sequence
TRBV7-8*01	NA	Not in GRCh38
TRBV7-8*02	NA	Not in GRCh38
TRBV7-8*03	NA	Not in GRCh38
TRBV7-9*05	NA	Not in GRCh37
TRBV7-9*06	NA	Not in GRCh37
TRBV12-4*02	NA	Non-germline sequence
TRBV7-9*07	NA	Not in GRCh37
TRBV7-9*01	NA	Not in GRCh37
TRBV7-9*02	NA	Not in GRCh37
TRBV5-5*03	NA	Non-germline sequence
TRBV7-9*03	NA	Not in GRCh37
TRBV7-9*04	NA	Not in GRCh37
TRBV29-1*02	NA	Non-germline sequence
TRBV29-1*03	NA	Non-germline sequence
TRAV20*01	TRAV20*2p	Recovered
TRAV13-2*01	TRAV13-2*4p	Recovered
TRAV13-1*01	TRAV13-1*2p	Recovered
TRAV3*01	TRAV3*5p,TRAV3*8p	Recovered
TRAV25*01	TRAV25*7p	Recovered
TRAV35*01	TRAV35*13p	Recovered
TRAV4*01	TRAV4*10p	Recovered
TRAV10*01	TRAV10*3p	Recovered
TRAV27*01	TRAV27*9p,TRAV27*7p	Recovered
TRAV22*01	TRAV22*10p,TRAV22*6p	Recovered
TRAV21*01	TRAV21*13p	Recovered
TRAV36/DV7*01	TRAV36/DV7*11p	Recovered
TRAV38-2/DV8*01	TRAV38-2/DV8*1p,TRAV38-2/DV8*3p,TRAV38-2/DV8*2p	Recovered
TRAV16*01	TRAV16*2p	Recovered

TRAV9-2*01	TRAV9-2*15p	Recovered
TRAV8-3*01	TRAV8-3*4p	Recovered
TRAV5*01	TRAV5*4p	Recovered
TRAV1-1*01	TRAV1-1*6p	Recovered
TRAV19*01	TRAV19*3p,TRAV19*16p,TRAV19*17p	Recovered
TRAV8-1*01	TRAV8-1*1p	Recovered
TRAV39*01	TRAV39*11p,TRAV39*3p,TRAV39*8p	Recovered
TRAV7*01	TRAV7*14p	Recovered
TRAV2*01	TRAV2*1p	Recovered
TRAV24*01	TRAV24*1p	Recovered
TRAV8-4*01	TRAV8-4*11p	Recovered
TRAV18*01	TRAV18*7p,TRAV18*3p	Recovered
TRAV6*01	TRAV6*5p	Recovered
TRAV29/DV5*01	TRAV29/DV5*6p,TRAV29/DV5*8p	Recovered
TRAV38-1*01	TRAV38-1*12p	Recovered
TRAV8-2*01	TRAV8-2*6p	Recovered
TRAV8-6*01	TRAV8-6*1p	Recovered
TRAV8-6*02	TRAV8-6*2p	Recovered
TRAV41*01	TRAV41*8p	Recovered
TRAV12-3*01	TRAV12-3*7p,TRAV12-3*6p	Recovered
TRAV26-1*01	TRAV26-1*13p	Recovered
TRAV26-2*01	TRAV26-2*12p,TRAV26-2*4p	Recovered
TRAV40*01	TRAV40*13p	Recovered
TRAV30*01	TRAV30*14p	Recovered
TRAV34*01	TRAV34*8p,TRAV34*6p	Recovered
TRAV17*01	TRAV17*7p,TRAV17*3p,TRAV17*4p,TRAV17*2p	Recovered
TRAV14/DV4*01	TRAV14/DV4*04*1p	Recovered
TRAV12-2*01	TRAV12-2*2p	Recovered
TRAV14/DV4*02	TRAV14/DV4*04*9p	Recovered
TRAV9-1*01	TRAV9-1*6p	Recovered
TRAV12-1*01	TRAV12-1*5p	Recovered
TRAV1-2*01	TRAV1-2*2p	Recovered
TRAV13-1*03	NA	Non-germline sequence

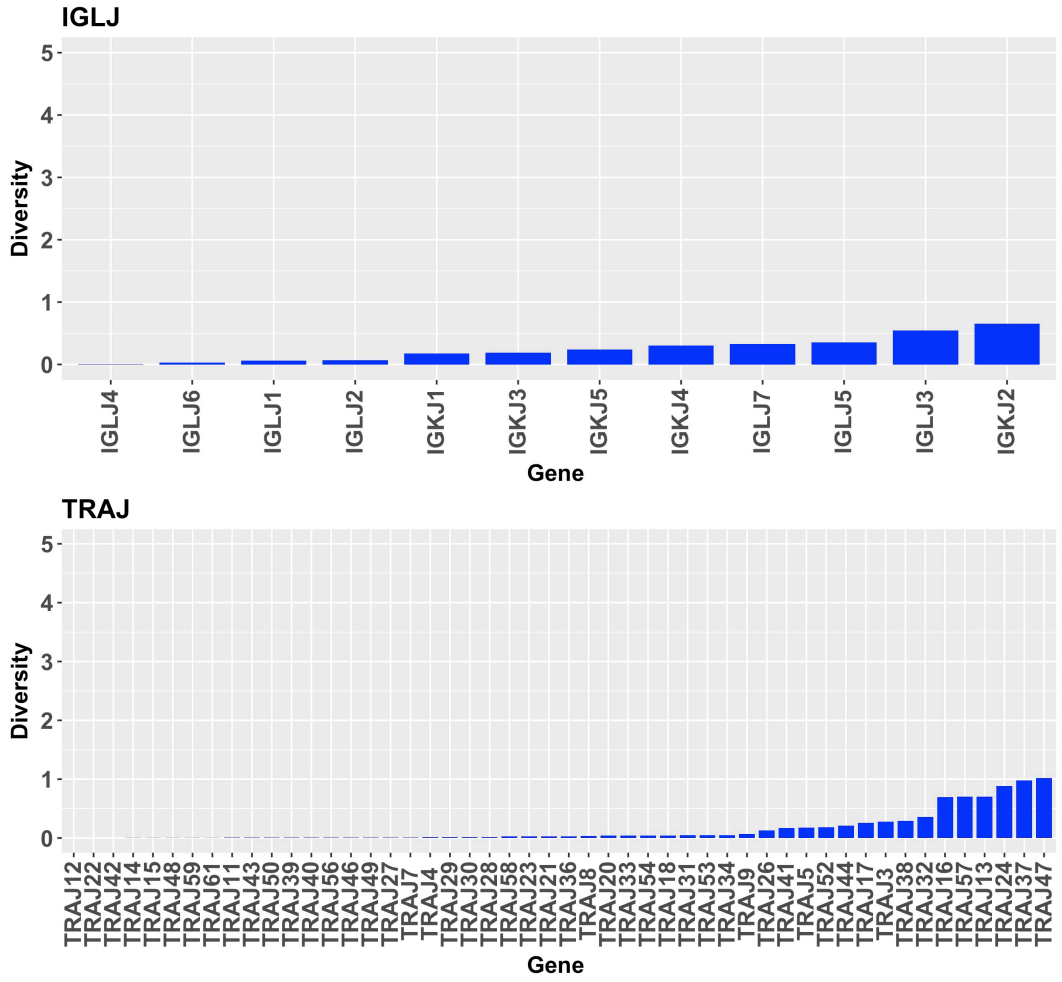
TRAV8-7*01	NA	Non-germline sequence
TRAV8-1*02	NA	Non-germline sequence
TRAV30*04	NA	Non-germline sequence
TRAV6*06	NA	Non-germline sequence
TRAV29/DV5*03	NA	Non-germline sequence
TRAV26-2*02	NA	Non-germline sequence
TRAV14/DV4*04	NA	Non-germline sequence
TRAV23/DV6*01	NA	Non-germline sequence
TRAV1-2*02	NA	Non-germline sequence
IGKJ2*02	IGKJ2*2p	Recovered
IGKJ2*01	IGKJ2*4p	Recovered
IGKJ1*01	IGKJ1*8p	Recovered
IGKJ2*04	IGKJ2*11p	Recovered
IGKJ2*03	IGKJ2*6p	Recovered
IGLJ4*01	IGLJ4*3p	Recovered
IGKJ5*01	IGKJ5*13p	Recovered
IGLJ6*01	IGLJ6*2p	Recovered
IGLJ2*01	IGLJ2*2p	Recovered
IGKJ4*02	IGKJ4*15p	Recovered
IGKJ4*01	IGKJ4*20p	Recovered
IGLJ3*01	IGLJ3*2p	Recovered
IGLJ3*02	IGLJ3*8p	Recovered
IGLJ5*02	IGLJ5*2p	Recovered
IGLJ5*01	IGLJ5*3p	Recovered
IGLJ7*01	IGLJ7*1p	Recovered
IGKJ3*01	IGKJ3*18p	Recovered
IGLJ7*02	IGLJ7*3p	Recovered
IGLJ1*01	IGLJ1*8p	Recovered
TRBJ1-1*01	TRBJ1-1*1p	Recovered
TRBJ2-5*01	TRBJ2-5*1p	Recovered
TRBJ1-5*01	TRBJ1-5*1p	Recovered

TRBJ2-4*01	TRBJ2-4*2p	Recovered
TRBJ2-1*01	TRBJ2-1*1p	Recovered
TRBJ1-2*01	TRBJ1-2*1p	Recovered
TRBJ2-6*01	TRBJ2-6*1p	Recovered
TRBJ1-4*01	TRBJ1-4*1p	Recovered
TRBJ2-2*01	TRBJ2-2*2p	Recovered
TRBJ2-3*01	TRBJ2-3*1p	Recovered
TRBJ2-7*01	TRBJ2-7*1p	Recovered
TRBJ1-3*01	TRBJ1-3*1p	Recovered
TRBJ1-6*01	TRBJ1-6*1p	Recovered
TRBJ1-6*02	NA	Not recovered
TRAJ41*01	TRAJ41*6p	Recovered
TRAJ37*01	TRAJ37*2p	Recovered
TRAJ3*01	TRAJ3*2p	Recovered
TRAJ28*01	TRAJ28*1p	Recovered
TRAJ22*01	TRAJ22*1p	Recovered
TRAJ30*01	TRAJ30*2p	Recovered
TRAJ56*01	TRAJ56*3p	Recovered
TRAJ44*01	TRAJ44*2p	Recovered
TRAJ7*01	TRAJ7*2p	Recovered
TRAJ43*01	TRAJ43*1p	Recovered
TRAJ52*01	TRAJ52*5p	Recovered
TRAJ54*01	TRAJ54*2p	Recovered
TRAJ5*01	TRAJ5*1p	Recovered
TRAJ27*01	TRAJ27*3p	Recovered
TRAJ39*01	TRAJ39*3p	Recovered
TRAJ57*01	TRAJ57*1p	Recovered
TRAJ10*01	TRAJ10*1p	Recovered
TRAJ8*01	TRAJ8*1p	Recovered
TRAJ9*01	TRAJ9*1p	Recovered
TRAJ38*01	TRAJ38*3p	Recovered
TRAJ48*01	TRAJ48*2p	Recovered
TRAJ11*01	TRAJ11*1p	Recovered
TRAJ17*01	TRAJ17*1p	Recovered

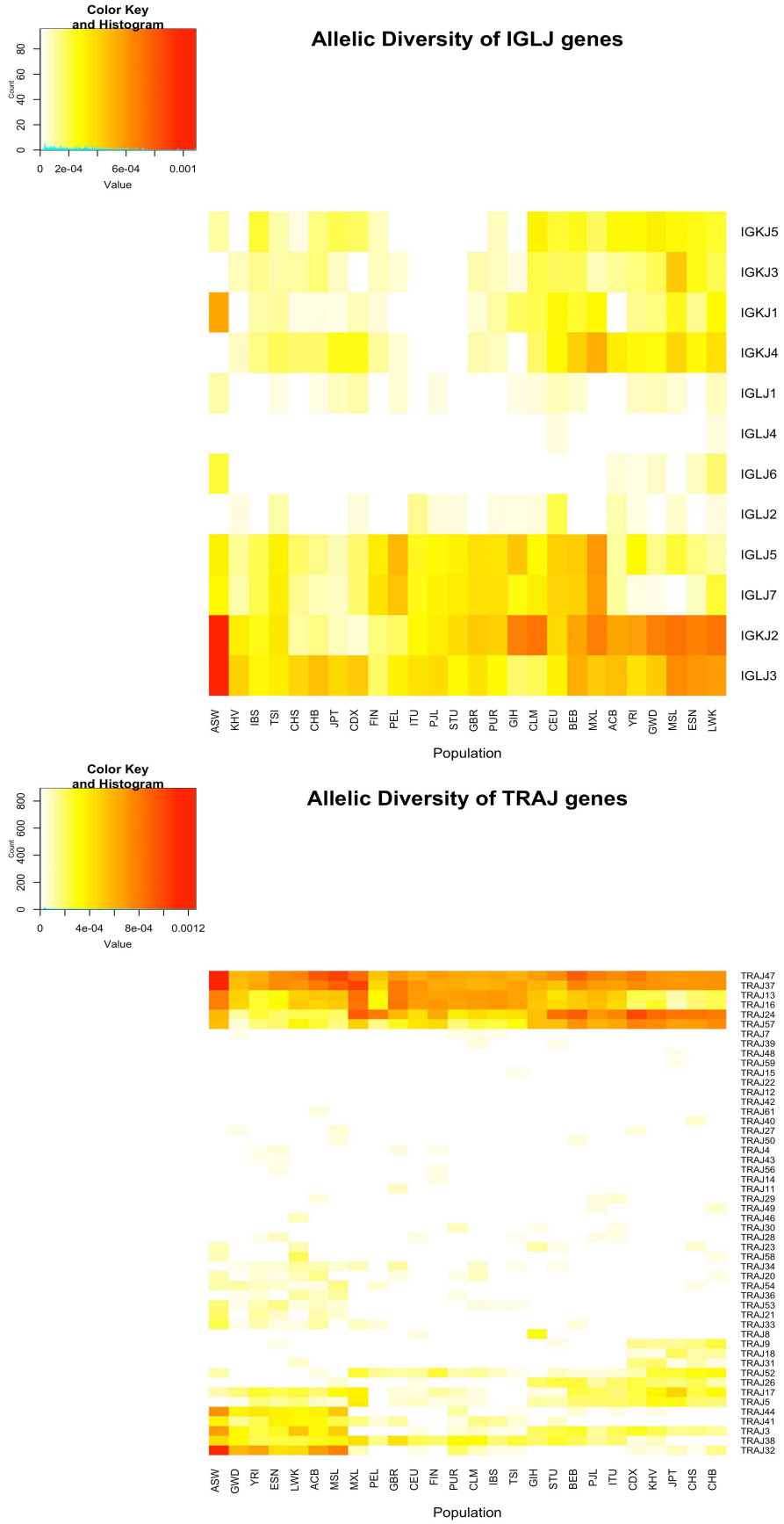
TRAJ20*01	TRAJ20*1p	Recovered
TRAJ21*01	TRAJ21*2p	Recovered
TRAJ16*01	TRAJ16*3p	Recovered
TRAJ13*01	TRAJ13*3p	Recovered
TRAJ42*01	TRAJ42*1p	Recovered
TRAJ4*01	TRAJ4*3p	Recovered
TRAJ26*01	TRAJ26*2p	Recovered
TRAJ46*01	TRAJ46*2p	Recovered
TRAJ40*01	TRAJ40*3p	Recovered
TRAJ33*01	TRAJ33*5p	Recovered
TRAJ53*01	TRAJ53*4p	Recovered
TRAJ6*01	TRAJ6*1p	Recovered
TRAJ49*01	TRAJ49*1p	Recovered
TRAJ12*01	TRAJ12*2p	Recovered
TRAJ50*01	TRAJ50*1p	Recovered
TRAJ45*01	TRAJ45*1p	Recovered
TRAJ15*01	TRAJ15*1p	Recovered
TRAJ47*01	TRAJ47*4p	Recovered
TRAJ34*01	TRAJ34*3p	Recovered
TRAJ18*01	TRAJ18*2p	Recovered
TRAJ32*01	TRAJ32*4p	Recovered
TRAJ29*01	TRAJ29*5p	Recovered
TRAJ23*01	TRAJ23*2p	Recovered
TRAJ31*01	TRAJ31*1p	Recovered
TRAJ14*01	TRAJ14*1p	Recovered
TRAJ24*01	NA	Not recovered
TRAJ36*01	NA	Not recovered
IGHD4-4*01	IGHD4-4*1p	Recovered
IGHD1-14*01	IGHD1-14*1p	Recovered
IGHD4-17*01	IGHD4-17*4p	Recovered
IGHD3-9*01	IGHD3-9*6p	Recovered
IGHD2-15*01	IGHD2-15*3p	Recovered
IGHD6-19*01	IGHD6-19*4p	Recovered
IGHD1-20*01	IGHD1-20*2p	Recovered

IGHD1-1*01	IGHD1-1*1p	Recovered
IGHD5-24*01	IGHD5-24*7p	Recovered
IGHD3-22*01	IGHD3-22*11p	Recovered
IGHD1-7*01	IGHD1-7*5p	Recovered
IGHD4-23*01	IGHD4-23*7p	Recovered
IGHD2-8*02	IGHD2-8*10p	Recovered
IGHD2-21*02	IGHD2-21*2p	Recovered
IGHD2-8*01	IGHD2-8*6p	Recovered
IGHD2-21*01	IGHD2-21*3p	Recovered
IGHD6-25*01	IGHD6-25*12p	Recovered
IGHD6-6*01	IGHD6-6*2p	Recovered
IGHD2-2*01	IGHD2-2*27p	Recovered
IGHD3-3*01	IGHD3-3*14p	Recovered
IGHD2-2*02	IGHD2-2*14p	Recovered
IGHD6-13*01	IGHD6-13*9p	Recovered
IGHD5-18*01	IGHD5-18*2p	Recovered
IGHD7-27*01	IGHD7-27*1p	Recovered
IGHD5-5*01	IGHD5-5*5p	Recovered
IGHD5-12*01	IGHD5-12*1p	Recovered
IGHD3-10*01	IGHD3-10*5p	Recovered
IGHD3-16*02	IGHD3-16*15p	Recovered
IGHD4-11*01	IGHD4-11*2p	Recovered
IGHD1-26*01	IGHD1-26*1p	Recovered
IGHD3-3*02	NA	Not recovered
IGHD2-2*03	NA	Not recovered
IGHD3-10*02	NA	Not recovered
IGHD3-16*01	NA	Not recovered
TRBD1*01	TRBD1*1p	Recovered
TRBD2*02	TRBD2*1p	Recovered
TRBD2*01	NA	Not recovered

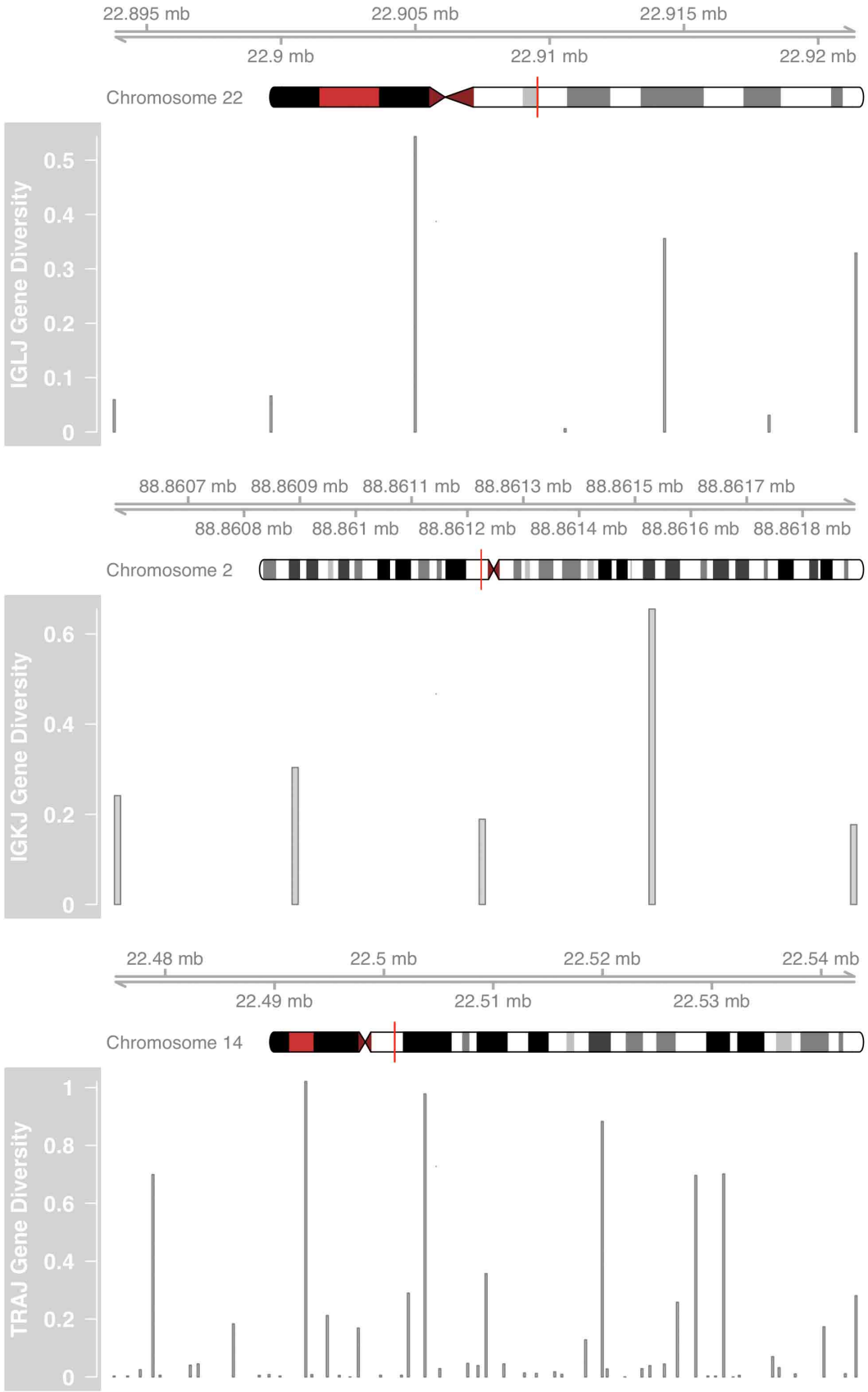
Appendix C



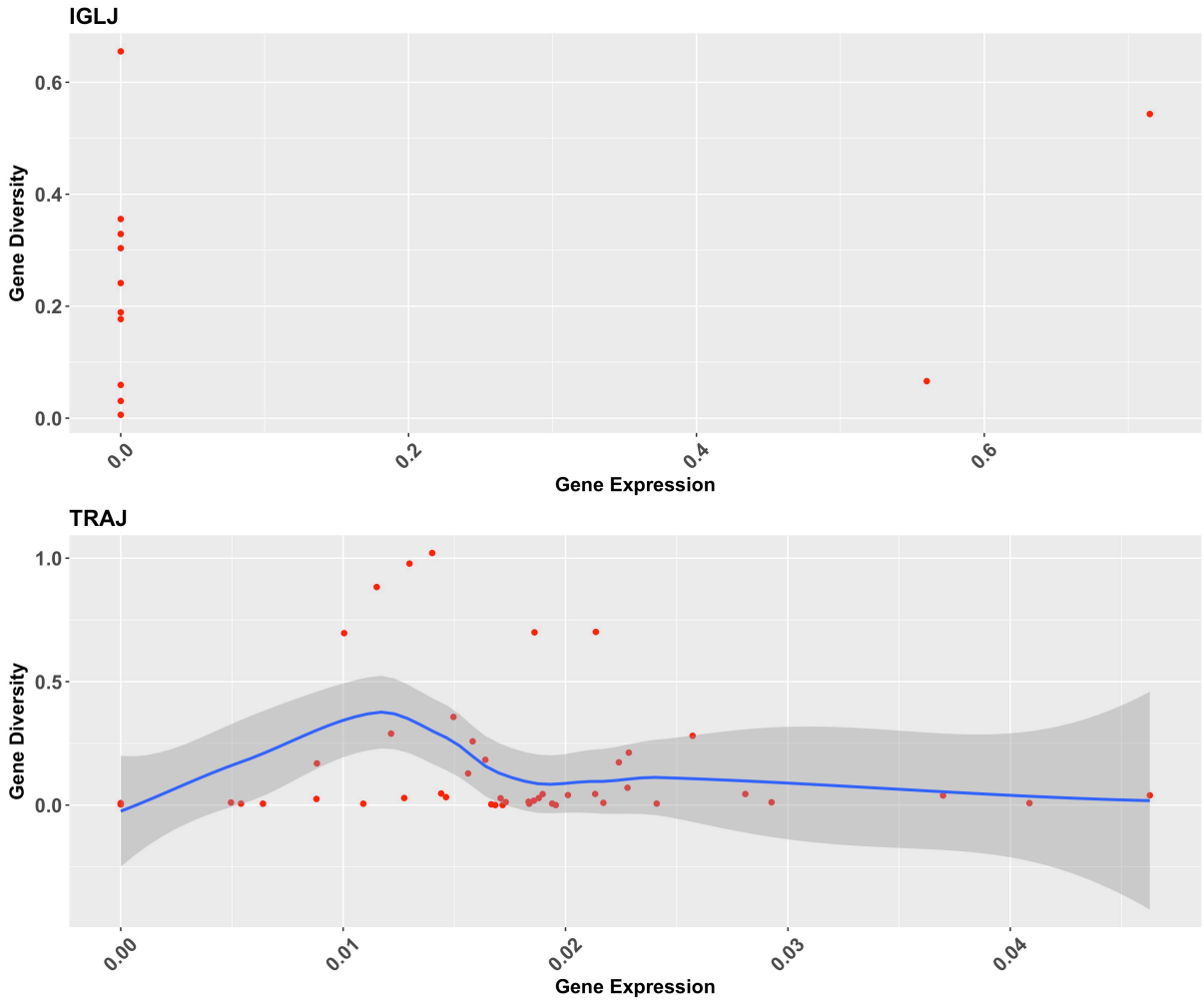
Supplementary figure C1. Shannon Entropy of TRAJ and IGLJ genes.



Supplementary figure C2. Allelic diversity of TRAJ and IGLJ genes within different populations.



Supplementary figure C3. The diversities of IGLJ, IGKJ and TRAJ genes along the chromosome.



Supplementary figure C4. The relationship between gene expression and diversity in TRAJ and IGLJ genes.

Supplementary table C1. Shannon Entropy of TCR/IG V genes in different populations.

	PEL	MXL	BEB	GBR	CDX	CLM	ACB	MSL	FIN	IBS	JPT	CHS	LWK	CHB	STU	KHV	CEU	GIH	ITU	TSI	ASW	YRI	PJL	GWD	PUR	ESN
IGHV1-69	1.7E-03	3.0E-03	2.1E-03	1.8E-03	1.8E-03	2.2E-03	3.0E-03	3.6E-03	1.7E-03	1.6E-03	1.4E-03	1.4E-03	2.9E-03	1.5E-03	1.8E-03	1.6E-03	1.9E-03	1.6E-03	1.6E-03	1.7E-03	5.1E-03	2.6E-03	1.8E-03	2.6E-03	1.7E-03	3.0E-03
IGHV2-26	7.7E-04	8.5E-04	3.2E-04	1.3E-04	2.5E-04	3.5E-04	9.1E-04	1.2E-03	3.0E-05	3.0E-04	2.9E-04	3.3E-04	8.7E-04	5.1E-04	3.9E-04	3.8E-04	6.1E-05	2.5E-04	3.8E-04	3.0E-04	1.3E-03	7.7E-04	2.8E-04	8.5E-04	3.5E-04	8.8E-04
IGHV1-2	1.5E-03	2.3E-03	1.6E-03	1.2E-03	1.3E-03	1.5E-03	1.4E-03	1.6E-03	1.2E-03	1.1E-03	1.4E-03	1.1E-03	1.3E-03	1.3E-03	1.3E-03	1.3E-03	1.4E-03	1.4E-03	1.3E-03	1.1E-03	2.1E-03	1.1E-03	1.3E-03	1.2E-03	1.1E-03	1.3E-03
IGHV5-51	2.8E-04	7.3E-04	8.7E-04	5.4E-04	5.4E-04	5.9E-04	7.1E-04	9.5E-04	6.8E-04	5.4E-04	4.1E-04	5.3E-04	8.8E-04	6.5E-04	3.5E-04	4.6E-04	6.0E-04	4.2E-04	4.6E-04	7.1E-04	1.3E-03	7.8E-04	6.3E-04	6.2E-04	4.6E-04	8.4E-04
IGHV1-3	9.0E-04	1.4E-03	1.1E-03	7.7E-04	8.4E-04	1.1E-03	1.2E-03	1.7E-03	8.4E-04	7.8E-04	7.8E-04	7.4E-04	1.2E-03	8.3E-04	7.6E-04	7.8E-04	8.7E-04	7.2E-04	6.6E-04	1.0E-03	2.2E-03	1.1E-03	9.1E-04	1.2E-03	7.6E-04	1.1E-03
IGHV7-81	3.8E-04	2.2E-04	4.5E-04	8.9E-05	8.5E-05	5.8E-04	1.1E-03	1.4E-03	1.3E-04	2.1E-04	6.8E-05	1.0E-04	9.5E-04	8.7E-05	2.7E-04	2.4E-04	1.1E-04	3.0E-04	3.2E-04	2.0E-04	1.7E-03	8.5E-04	1.4E-04	8.3E-04	3.9E-04	9.3E-04
IGHV3-49	1.6E-03	2.1E-03	1.3E-03	1.3E-03	1.2E-03	1.3E-03	1.1E-03	1.0E-03	1.2E-03	1.1E-03	1.1E-03	1.1E-03	1.1E-03	1.1E-03	1.1E-03	1.2E-03	1.1E-03	9.3E-04	1.0E-03	1.1E-03	2.0E-03	7.2E-04	1.2E-03	7.2E-04	1.1E-03	9.0E-04
IGHV3-48	1.6E-03	2.7E-03	1.8E-03	1.8E-03	1.4E-03	1.7E-03	1.5E-03	1.7E-03	1.7E-03	1.4E-03	1.3E-03	1.3E-03	1.6E-03	1.3E-03	1.7E-03	1.4E-03	1.6E-03	1.6E-03	1.6E-03	1.4E-03	2.7E-03	1.5E-03	1.8E-03	1.6E-03	1.5E-03	1.8E-03
IGHV1-24	1.2E-04	2.9E-04	3.2E-04	1.3E-04	1.0E-04	3.0E-04	2.9E-04	6.1E-04	2.0E-04	1.3E-04	2.7E-05	1.0E-04	4.7E-04	5.6E-05	2.9E-05	1.1E-04	2.2E-04	1.1E-04	1.2E-04	1.2E-04	6.8E-04	3.8E-04	3.2E-05	3.6E-04	1.1E-04	3.9E-04
IGHV2-5	5.8E-04	1.0E-03	8.5E-04	8.5E-04	5.4E-04	7.4E-04	8.7E-04	1.1E-03	7.2E-04	6.7E-04	4.3E-04	4.8E-04	1.2E-03	6.0E-04	4.5E-04	5.6E-04	5.7E-04	4.9E-04	5.4E-04	6.7E-04	1.6E-03	9.4E-04	5.7E-04	8.4E-04	5.6E-04	1.1E-03
IGHV1-46	2.2E-04	7.5E-04	6.2E-04	6.9E-04	2.4E-04	4.2E-04	6.3E-04	7.7E-04	5.9E-04	6.3E-04	1.9E-04	1.6E-04	6.9E-04	1.5E-04	4.5E-04	3.0E-05	5.8E-04	5.3E-04	5.9E-04	7.0E-04	1.0E-03	6.5E-04	6.0E-04	6.6E-04	4.6E-04	6.0E-04
IGHV3-56	1.0E-03	1.4E-03	1.3E-03	8.7E-04	1.2E-03	1.1E-03	1.5E-03	1.8E-03	8.4E-04	8.0E-04	8.9E-04	9.7E-04	1.4E-03	9.6E-04	1.2E-03	1.1E-03	9.4E-04	9.9E-04	1.0E-03	8.8E-04	2.6E-03	1.3E-03	1.0E-03	1.2E-03	8.9E-04	1.5E-03
IGHV1-45	5.4E-04	7.2E-04	5.4E-04	2.1E-04	7.5E-04	3.9E-04	1.8E-04	3.5E-04	1.9E-04	1.4E-04	5.9E-04	5.7E-04	2.4E-04	6.1E-04	4.8E-04	6.9E-04	1.6E-04	3.0E-04	2.8E-04	1.9E-04	4.2E-04	2.5E-05	4.3E-04	1.2E-04	2.3E-04	1.4E-04
IGHV3-54	8.3E-04	1.2E-03	7.5E-04	6.8E-04	7.2E-04	8.2E-04	8.1E-04	1.1E-03	6.4E-04	7.4E-04	7.1E-04	6.4E-04	1.1E-03	6.8E-04	5.7E-04	6.5E-04	6.3E-04	6.0E-04	6.6E-04	6.8E-04	1.6E-03	7.1E-04	6.5E-04	7.6E-04	7.1E-04	7.9E-04
IGHV4-28	1.7E-03	2.5E-03	1.6E-03	1.8E-03	9.1E-04	1.9E-03	2.1E-03	2.2E-03	1.4E-03	1.5E-03	8.2E-04	9.3E-04	1.6E-03	1.0E-03	1.4E-03	1.0E-03	1.5E-03	1.4E-03	1.4E-03	1.3E-03	3.1E-03	1.6E-03	1.6E-03	1.4E-03	1.7E-03	1.8E-03
IGHV3-20	9.5E-04	1.7E-03	1.1E-03	1.1E-03	8.3E-04	1.0E-03	8.6E-04	8.9E-04	1.0E-03	8.8E-04	7.7E-04	8.7E-04	7.7E-04	8.7E-04	1.2E-03	1.0E-03	1.1E-03	9.9E-04	1.1E-03	8.0E-04	1.6E-03	6.6E-04	1.2E-03	7.9E-04	9.0E-04	7.4E-04
IGHV1-18	2.9E-04	8.6E-04	8.2E-04	6.8E-04	7.6E-04	4.5E-04	4.8E-04	8.9E-04	4.2E-04	5.4E-04	3.5E-04	4.9E-04	5.4E-04	5.8E-04	6.1E-04	5.9E-04	6.4E-04	6.2E-04	6.3E-04	6.4E-04	1.1E-03	2.7E-04	5.7E-04	6.5E-04	4.4E-04	6.8E-04
IGHV3-21	2.7E-04	2.6E-04	2.6E-04	2.1E-04	6.9E-05	4.8E-04	5.1E-04	9.2E-04	6.1E-05	2.1E-04	1.0E-04	8.1E-05	9.0E-04	1.7E-04	2.9E-05	5.4E-05	5.3E-04	5.6E-05	5.1E-05	3.0E-04	1.1E-03	3.5E-04	3.2E-05	7.5E-04	8.2E-05	6.3E-04
IGHV4-61	1.5E-03	2.6E-03	2.4E-03	1.5E-03	2.1E-03	1.9E-03	3.4E-03	3.9E-03	1.6E-03	1.6E-03	1.3E-03	1.6E-03	3.5E-03	1.9E-03	2.8E-03	2.2E-03	1.5E-03	2.5E-03	3.1E-03	1.7E-03	5.8E-03	3.0E-03	3.3E-03	3.0E-03	1.8E-03	3.5E-03
IGHV3-23	5.0E-04	1.5E-03	7.9E-04	6.6E-04	6.8E-04	7.2E-04	1.4E-03	2.0E-03	5.5E-04	7.1E-04	9.0E-04	5.7E-04	1.5E-03	7.0E-04	4.2E-04	6.1E-04	6.9E-04	5.8E-04	7.1E-04	3.4E-04	2.8E-03	1.3E-03	4.2E-04	1.4E-03	5.6E-04	1.6E-03
IGHV4-4	3.8E-03	5.2E-03	3.7E-03	3.4E-03	3.3E-03	3.3E-03	2.7E-03	3.7E-03	3.2E-03	2.6E-03	3.2E-03	3.2E-03	2.9E-03	3.1E-03	3.2E-03	3.3E-03	3.0E-03	3.3E-03	3.3E-03	2.6E-03	5.1E-03	2.5E-03	3.2E-03	2.5E-03	2.6E-03	2.8E-03
IGHV3-72	0.0E+00	0.0E+00	4.0E-05	0.0E+00	6.9E-05	1.4E-04	3.3E-04	4.7E-04	3.0E-05	5.2E-05	4.9E-05	1.3E-04	3.5E-04	2.8E-05	7.1E-05	8.4E-05	9.1E-05	5.0E-05	2.9E-05	2.6E-05	8.3E-04	2.8E-04	3.2E-05	3.7E-04	2.7E-05	4.1E-04
IGHV3-53	1.5E-03	2.6E-03	1.8E-03	1.5E-03	1.0E-03	2.0E-03	1.9E-03	2.4E-03	1.4E-03	1.4E-03	1.1E-03	1.0E-03	2.1E-03	1.1E-03	1.6E-03	1.3E-03	1.4E-03	1.6E-03	1.6E-03	1.4E-03	3.5E-03	1.8E-03	1.8E-03	1.7E-03	1.5E-03	2.1E-03
IGHV3-74	0.0E+00	7.4E-05	0.0E+00	3.6E-05	1.4E-04	6.7E-05	6.1E-04	6.3E-04	6.1E-05	5.2E-05	1.4E-04	0.0E+00	5.9E-04	5.0E-05	5.1E-05	3.0E-05	2.7E-04	0.0E+00	2.9E-05	7.8E-05	1.3E-03	4.6E-04	1.2E-04	4.3E-04	1.4E-04	5.4E-04
IGHV3-73	7.8E-04	1.1E-03	8.6E-04	6.9E-04	7.1E-04	7.3E-04	9.9E-04	1.1E-03	7.3E-04	6.4E-04	6.6E-04	6.2E-04	9.5E-04	6.3E-04	5.5E-04	6.5E-04	6.6E-04	6.5E-04	6.0E-04	6.0E-04	1.7E-03	9.4E-04	6.0E-04	8.5E-04	6.6E-04	9.6E-04

IGHV1-58	8.6E-04	1.3E-03	7.5E-04	6.9E-04	8.1E-04	8.9E-04	8.9E-04	1.3E-03	6.4E-04	6.4E-04	6.8E-04	7.2E-04	9.4E-04	7.7E-04	5.5E-04	7.5E-04	6.7E-04	6.0E-04	5.3E-04	6.9E-04	1.7E-03	7.2E-04	6.1E-04	7.1E-04	7.4E-04	8.6E-04
IGHV3-11	1.5E-03	2.3E-03	1.4E-03	1.3E-03	1.5E-03	1.5E-03	1.5E-03	1.7E-03	1.1E-03	1.1E-03	1.3E-03	1.2E-03	1.5E-03	1.3E-03	1.1E-03	1.3E-03	1.3E-03	1.1E-03	1.1E-03	9.3E-04	2.5E-03	1.3E-03	1.1E-03	1.3E-03	1.3E-03	1.5E-03
IGHV3-7	1.2E-03	2.0E-03	1.8E-03	1.3E-03	1.1E-03	1.4E-03	9.1E-04	1.6E-03	1.2E-03	1.1E-03	6.8E-04	8.6E-04	1.3E-03	8.3E-04	1.1E-03	1.0E-03	1.4E-03	1.2E-03	1.1E-03	1.1E-03	2.0E-03	7.9E-04	1.2E-03	9.4E-04	9.6E-04	1.2E-03
IGHV4-39	9.6E-04	1.1E-03	1.1E-03	4.7E-04	7.3E-04	7.1E-04	1.3E-03	1.5E-03	4.5E-04	4.3E-04	4.6E-04	6.2E-04	1.3E-03	7.6E-04	9.4E-04	6.0E-04	5.8E-04	1.0E-03	1.0E-03	5.4E-04	2.0E-03	1.2E-03	8.9E-04	1.1E-03	6.7E-04	1.2E-03
IGHV3-33	7.2E-04	1.3E-03	1.9E-03	5.6E-04	7.6E-04	9.6E-04	2.1E-03	2.8E-03	6.1E-04	6.5E-04	6.7E-04	8.1E-04	2.5E-03	6.4E-04	1.6E-03	9.3E-04	7.8E-04	1.2E-03	1.8E-03	7.1E-04	3.8E-03	2.2E-03	1.1E-03	1.9E-03	8.3E-04	2.3E-03
IGHV6-1	0.0E+00	7.4E-05	0.0E+00	1.1E-04	1.7E-04	1.9E-04	6.5E-04	8.6E-04	0.0E+00	1.0E-04	0.0E+00	8.1E-05	5.8E-04	1.1E-04	2.9E-05	9.1E-05	0.0E+00	0.0E+00	7.9E-05	5.2E-05	1.3E-03	5.8E-04	0.0E+00	5.8E-04	8.6E-05	6.5E-04
IGHV3-35	6.1E-04	7.4E-04	9.9E-04	3.9E-04	5.8E-04	6.3E-04	1.1E-03	1.2E-03	2.2E-04	3.9E-04	5.5E-04	6.2E-04	8.1E-04	6.3E-04	7.6E-04	5.6E-04	3.4E-04	7.1E-04	9.0E-04	3.9E-04	1.7E-03	9.3E-04	8.0E-04	8.5E-04	4.8E-04	1.1E-03
IGHV3-15	3.4E-04	9.8E-04	8.2E-04	5.2E-04	3.8E-04	5.7E-04	7.1E-04	8.2E-04	5.4E-04	5.9E-04	5.2E-04	3.7E-04	3.2E-04	3.6E-04	6.4E-04	3.3E-04	6.1E-04	5.1E-04	4.2E-04	5.2E-04	1.5E-03	5.2E-04	5.9E-04	4.4E-04	4.5E-04	6.3E-04
IGHV3-16	7.5E-04	1.4E-03	4.6E-04	4.0E-04	4.8E-04	6.9E-04	9.0E-04	1.2E-03	5.3E-04	3.1E-04	3.7E-04	3.5E-04	9.9E-04	3.1E-04	3.7E-04	4.8E-04	5.1E-04	1.8E-04	3.7E-04	3.4E-04	1.6E-03	7.0E-04	3.7E-04	7.5E-04	6.1E-04	8.2E-04
IGHV4-59	1.2E-03	1.4E-03	1.0E-03	3.3E-04	1.1E-03	8.2E-04	1.1E-03	1.7E-03	3.8E-04	5.0E-04	6.4E-04	8.8E-04	1.7E-03	7.8E-04	6.2E-04	7.3E-04	4.7E-04	2.8E-04	5.2E-04	3.3E-04	2.2E-03	1.2E-03	5.3E-04	1.3E-03	5.4E-04	1.3E-03
IGHV3-13	1.0E-03	2.0E-03	1.4E-03	1.3E-03	9.5E-04	1.4E-03	1.3E-03	1.6E-03	1.0E-03	1.1E-03	7.2E-04	7.2E-04	1.1E-03	8.8E-04	1.3E-03	8.9E-04	1.2E-03	1.1E-03	1.3E-03	9.8E-04	2.6E-03	1.2E-03	1.2E-03	1.4E-03	1.3E-03	1.1E-03
IGHV2-70	1.4E-03	2.6E-03	2.7E-03	1.0E-03	2.2E-03	2.7E-03	2.6E-03	3.6E-03	1.3E-03	1.8E-03	1.4E-03	1.5E-03	3.0E-03	1.7E-03	8.9E-04	1.5E-03	2.1E-03	1.2E-03	8.5E-04	1.6E-03	4.5E-03	2.6E-03	8.1E-04	2.8E-03	1.5E-03	2.9E-03
IGHV4-34	8.3E-05	7.4E-05	1.2E-04	7.2E-05	1.4E-04	1.4E-04	1.6E-04	5.7E-04	6.1E-05	9.8E-05	8.2E-05	8.1E-05	1.5E-04	2.2E-04	1.0E-04	3.0E-05	2.4E-04	2.8E-05	1.9E-04	5.2E-05	2.4E-04	1.3E-04	0.0E+00	2.3E-04	0.0E+00	2.7E-04
IGLV1-40	4.1E-05	2.9E-04	3.2E-04	3.3E-04	1.4E-04	1.6E-04	3.1E-04	3.1E-04	1.1E-04	2.3E-04	8.2E-05	9.4E-05	5.4E-04	1.3E-04	1.1E-04	2.0E-04	6.1E-04	2.0E-04	5.7E-05	1.8E-04	6.1E-04	1.6E-04	2.2E-04	3.9E-04	2.8E-04	1.8E-04
IGLV1-44	8.3E-05	2.9E-04	5.3E-04	1.1E-04	2.6E-04	1.6E-04	6.9E-04	9.6E-04	2.0E-04	2.1E-04	1.9E-04	3.4E-04	8.2E-04	1.7E-04	2.9E-05	1.9E-04	1.5E-04	1.1E-04	7.9E-05	2.0E-04	6.7E-04	6.3E-04	3.2E-05	5.0E-04	5.5E-05	8.5E-04
IGLV4-60	1.3E-03	2.0E-03	1.4E-03	1.2E-03	1.2E-03	1.5E-03	1.4E-03	1.7E-03	1.1E-03	1.2E-03	9.0E-04	1.1E-03	1.5E-03	1.1E-03	1.1E-03	1.0E-03	1.1E-03	1.2E-03	1.0E-03	1.1E-03	2.4E-03	1.3E-03	1.2E-03	1.1E-03	1.2E-03	1.3E-03
IGKV1-17	9.4E-04	8.3E-04	1.1E-03	7.3E-04	1.1E-03	8.2E-04	1.2E-03	1.7E-03	2.5E-04	3.8E-04	1.0E-03	9.9E-04	1.6E-03	1.0E-03	5.8E-04	8.1E-04	5.2E-04	4.2E-04	3.5E-04	5.2E-04	2.3E-03	1.3E-03	2.5E-04	1.2E-03	4.3E-04	1.6E-03
IGKV1-16	0.0E+00	0.0E+00	0.0E+00	3.6E-05	1.2E-04	1.5E-04	2.6E-04	8.3E-05	0.0E+00	5.2E-05	2.1E-04	6.7E-05	2.6E-04	1.5E-04	5.7E-05	8.4E-05	0.0E+00	2.8E-05	5.1E-05	0.0E+00	4.7E-04	1.9E-04	3.2E-05	2.7E-04	2.7E-05	3.0E-05
IGLV1-47	5.5E-04	7.5E-04	7.6E-04	5.0E-04	5.5E-04	4.3E-04	7.4E-04	1.0E-03	7.3E-04	4.0E-04	5.7E-04	5.4E-04	7.0E-04	4.8E-04	4.5E-04	5.1E-04	6.2E-04	5.5E-04	4.3E-04	5.1E-04	1.5E-03	8.4E-04	5.3E-04	6.8E-04	5.3E-04	9.5E-04
IGLV5-37	4.7E-04	1.1E-03	1.1E-03	7.0E-04	7.2E-04	6.8E-04	8.2E-04	8.6E-04	7.5E-04	6.5E-04	6.3E-04	6.4E-04	9.8E-04	6.4E-04	8.7E-04	7.6E-04	7.1E-04	8.8E-04	6.9E-04	6.6E-04	1.7E-03	5.9E-04	8.5E-04	8.6E-04	7.2E-04	7.0E-04
IGKV2D-26	7.7E-04	1.1E-03	8.8E-04	4.7E-04	7.5E-04	6.7E-04	8.9E-04	9.8E-04	3.0E-04	3.1E-04	6.2E-04	6.1E-04	8.1E-04	6.1E-04	6.8E-04	5.8E-04	4.5E-04	5.7E-04	7.0E-04	2.7E-04	1.4E-03	7.0E-04	6.7E-04	6.9E-04	7.2E-04	7.5E-04
IGLV2-23	2.4E-04	2.8E-04	4.3E-04	2.6E-04	2.3E-04	4.4E-04	1.1E-03	1.5E-03	1.2E-04	3.4E-04	2.7E-04	3.0E-04	1.2E-03	0.0E+00	2.0E-04	2.2E-04	4.6E-04	1.8E-04	3.5E-04	1.7E-04	7.0E-04	8.2E-04	1.7E-04	7.2E-04	4.1E-04	8.0E-04
IGKV2D-24	0.0E+00	0.0E+00	1.4E-04	0.0E+00	3.4E-05	1.2E-04	1.9E-04	1.8E-04	6.1E-05	5.2E-05	0.0E+00	2.7E-05	3.0E-04	0.0E+00	1.4E-04	3.0E-05	3.0E-05	2.8E-05	0.0E+00	0.0E+00	4.8E-04	2.1E-04	6.5E-05	3.0E-04	5.5E-05	2.0E-04
IGLV6-57	1.5E-03	2.4E-03	1.6E-03	1.4E-03	1.4E-03	1.6E-03	1.4E-03	1.6E-03	1.3E-03	1.3E-03	1.2E-03	1.3E-03	1.2E-03	1.3E-03	1.2E-03	1.2E-03	1.3E-03	1.2E-03	1.2E-03	1.2E-03	2.4E-03	1.0E-03	1.3E-03	9.6E-04	1.3E-03	1.0E-03
IGKV2-30	7.7E-04	9.9E-04	6.4E-04	2.8E-04	6.9E-04	5.7E-04	7.9E-04	1.0E-03	2.0E-04	2.0E-04	8.0E-04	6.2E-04	8.4E-04	6.2E-04	4.8E-04	6.1E-04	2.4E-04	3.9E-04	4.6E-04	3.2E-04	1.5E-03	6.2E-04	5.1E-04	6.0E-04	5.3E-04	7.2E-04
IGKV1D-39	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	1.0E-04	0.0E+00	0.0E+00	2.7E-05	0.0E+00	3.0E-05	0.0E+00	2.9E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.3E-05	0.0E+00	0.0E+00
IGLV8-61	1.1E-03	1.2E-03	8.9E-04	6.6E-04	8.7E-04	9.2E-04	4.1E-04	3.8E-04	7.7E-04	3.9E-04	1.0E-03	7.4E-04	4.0E-04	7.2E-04	5.3E-04	7.6E-04	5.6E-04	5.9E-04	7.0E-04	4.8E-04	1.3E-03	3.1E-04	6.0E-04	2.7E-04	5.5E-04	4.5E-04
IGKV2D-28	8.3E-05	1.8E-04	7.2E-05	1.9E-04	0.0E+00	3.0E-04	7.3E-04	7.6E-04	5.4E-05	1.8E-04	0.0E+00	2.7E-05	7.6E-04	2.8E-05	8.9E-05	0.0E+00	1.9E-04	2.8E-05	5.1E-05	2.1E-04	1.2E-03	5.9E-04	5.7E-05	5.8E-04	4.1E-04	6.8E-04
IGKV2D-29	4.1E-05	4.5E-04	0.0E+00	2.5E-04	0.0E+00	1.4E-04	6.5E-05	1.7E-04	2.1E-04	7.8E-05	3.7E-04	1.6E-04	2.4E-04	1.6E-04	2.9E-05	0.0E+00	1.9E-04	0.0E+00	2.9E-05	1.9E-04	5.7E-04	2.6E-04	1.4E-04	1.3E-04	3.8E-04	9.1E-05

IGKV1D-37	4.1E-05	1.3E-04	4.0E-05	1.5E-04	0.0E+00	1.7E-04	3.2E-04	3.6E-04	3.0E-05	6.4E-05	0.0E+00	0.0E+00	3.0E-04	0.0E+00	0.0E+00	0.0E+00	1.5E-04	0.0E+00	0.0E+00	9.7E-05	7.4E-04	2.9E-04	3.2E-05	3.6E-04	0.0E+00	2.1E-04
IGLV3-32	4.1E-05	1.5E-04	4.0E-05	0.0E+00	1.8E-04	0.0E+00	0.0E+00	4.1E-05	0.0E+00	2.6E-05	1.8E-04	1.5E-04	3.0E-05	1.3E-04	5.7E-05	1.1E-04	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.2E-05	0.0E+00	2.7E-05	3.0E-05
IGLV1-50	1.2E-04	2.2E-04	4.9E-04	0.0E+00	3.4E-04	1.3E-04	2.3E-04	4.5E-04	0.0E+00	7.2E-05	4.0E-04	3.1E-04	6.3E-04	3.2E-04	2.4E-04	4.8E-04	5.4E-05	2.3E-04	2.5E-04	5.2E-05	8.2E-04	4.2E-04	2.6E-04	3.4E-04	1.5E-04	2.9E-04
IGKV1D-8	4.2E-04	7.3E-04	9.2E-04	9.3E-04	4.9E-04	6.9E-04	9.3E-04	1.0E-03	8.1E-04	3.6E-04	5.7E-04	4.0E-04	8.2E-04	5.2E-04	6.6E-04	6.2E-04	7.6E-04	5.6E-04	5.7E-04	5.8E-04	1.4E-03	6.6E-04	6.0E-04	7.2E-04	7.9E-04	7.3E-04
IGKV1D-33	4.1E-05	0.0E+00	0.0E+00	3.6E-05	0.0E+00	6.0E-05	3.2E-05	4.1E-05	3.0E-05	9.7E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	1.1E-04	5.0E-05	0.0E+00	6.4E-05	8.1E-05	4.5E-05	0.0E+00	0.0E+00	0.0E+00	3.0E-05
IGLV1-51	2.6E-04	1.0E-03	5.7E-04	4.8E-04	2.1E-04	6.4E-04	8.9E-04	1.2E-03	4.3E-04	5.3E-04	1.3E-04	2.0E-04	1.1E-03	1.9E-04	6.7E-04	2.4E-04	7.1E-04	5.4E-04	5.7E-04	5.4E-04	1.3E-03	1.0E-03	4.5E-04	8.3E-04	3.8E-04	8.7E-04
IGKV1D-42	4.5E-04	5.3E-04	6.3E-04	6.0E-04	2.3E-04	5.3E-04	1.3E-03	1.4E-03	4.0E-04	2.7E-04	4.6E-04	1.8E-04	1.2E-03	2.4E-04	4.2E-04	4.4E-04	4.7E-04	4.6E-04	4.1E-04	5.5E-04	2.0E-03	1.1E-03	3.6E-04	1.0E-03	6.9E-04	1.2E-03
IGKV1-27	7.8E-04	9.3E-04	7.3E-04	3.0E-04	9.2E-04	6.7E-04	7.7E-04	1.1E-03	9.4E-05	1.9E-04	9.7E-04	9.0E-04	8.7E-04	8.4E-04	4.7E-04	8.1E-04	3.0E-04	3.8E-04	4.5E-04	3.3E-04	1.4E-03	7.8E-04	6.3E-04	6.7E-04	5.4E-04	7.5E-04
IGLV11-55	8.6E-04	1.2E-03	8.4E-04	7.9E-04	7.7E-04	7.6E-04	1.1E-03	1.2E-03	7.8E-04	6.7E-04	7.1E-04	6.9E-04	8.9E-04	7.0E-04	6.3E-04	7.2E-04	6.8E-04	6.2E-04	6.5E-04	6.3E-04	1.9E-03	9.6E-04	7.7E-04	1.0E-03	7.8E-04	9.8E-04
IGKV3-7	8.1E-04	1.1E-03	7.8E-04	6.9E-04	6.6E-04	7.3E-04	9.5E-04	9.5E-04	4.3E-04	5.5E-04	6.2E-04	6.2E-04	8.6E-04	6.5E-04	6.3E-04	6.5E-04	5.5E-04	5.6E-04	6.8E-04	4.5E-04	1.5E-03	8.4E-04	7.1E-04	4.6E-04	5.1E-04	6.2E-04
IGLV2-11	4.1E-05	2.2E-04	7.5E-04	0.0E+00	5.4E-04	3.4E-05	1.1E-04	6.6E-04	1.5E-04	5.2E-05	4.5E-04	3.5E-04	2.4E-04	3.9E-04	3.0E-04	2.7E-04	3.6E-04	3.1E-04	3.1E-04	1.6E-04	3.0E-04	3.0E-04	1.8E-04	2.8E-04	1.8E-04	3.0E-04
IGKV2-28	1.7E-04	2.0E-04	3.4E-04	1.7E-04	1.1E-04	1.6E-04	3.2E-05	4.1E-05	6.1E-05	2.2E-04	2.7E-05	0.0E+00	1.1E-04	2.8E-05	2.3E-04	0.0E+00	1.3E-04	1.7E-04	2.5E-04	2.2E-04	0.0E+00	2.5E-05	4.0E-04	2.3E-05	1.2E-04	3.0E-05
IGKV1-8	1.1E-03	1.3E-03	1.1E-03	8.4E-04	6.9E-04	9.3E-04	1.2E-03	1.6E-03	4.3E-04	5.7E-04	6.9E-04	7.1E-04	1.5E-03	6.7E-04	8.2E-04	7.4E-04	6.7E-04	7.7E-04	8.4E-04	5.5E-04	2.2E-03	1.1E-03	8.6E-04	1.2E-03	7.2E-04	1.2E-03
IGKV1-5	1.3E-03	1.4E-03	1.6E-03	1.2E-03	8.4E-04	1.0E-03	6.1E-04	1.3E-03	7.3E-04	8.7E-04	9.0E-04	7.3E-04	1.2E-03	9.3E-04	1.0E-03	9.3E-04	8.2E-04	8.6E-04	1.0E-03	9.2E-04	1.3E-03	8.3E-04	1.1E-03	1.1E-03	7.1E-04	1.0E-03
IGKV1-6	3.4E-04	4.5E-04	5.4E-04	4.5E-04	2.6E-04	2.3E-04	1.3E-04	1.7E-04	2.5E-04	2.6E-04	1.5E-04	5.4E-05	1.5E-04	1.1E-04	3.7E-04	1.1E-04	3.0E-04	3.5E-04	3.8E-04	2.2E-04	2.4E-04	7.6E-05	3.6E-04	1.3E-04	2.0E-04	1.2E-04
IGLV9-49	4.1E-05	7.4E-05	1.9E-04	3.6E-05	1.0E-04	1.4E-04	3.2E-05	2.5E-04	1.9E-04	2.6E-05	2.5E-04	1.2E-04	3.0E-05	1.7E-04	3.2E-04	1.1E-04	9.1E-05	2.1E-04	1.6E-04	1.5E-04	8.1E-05	4.5E-05	9.0E-05	6.5E-05	0.0E+00	1.1E-04
IGLV2-14	1.4E-03	3.0E-03	1.8E-03	1.5E-03	1.8E-03	2.0E-03	2.0E-03	2.7E-03	1.5E-03	1.4E-03	1.6E-03	1.5E-03	2.0E-03	1.6E-03	1.4E-03	1.4E-03	1.7E-03	1.4E-03	1.5E-03	1.2E-03	3.8E-03	1.8E-03	1.3E-03	1.7E-03	1.3E-03	2.1E-03
IGKV1-39	0.0E+00	1.5E-04	4.0E-05	3.6E-05	0.0E+00	0.0E+00	3.2E-05	1.2E-04	1.1E-04	5.2E-05	5.5E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.0E-05	5.6E-05	0.0E+00	2.6E-05	8.1E-05	5.1E-05	0.0E+00	2.3E-05	2.7E-05	6.1E-05
IGKV1-9	8.7E-04	1.2E-03	8.2E-04	8.8E-04	7.4E-04	7.8E-04	7.5E-04	1.1E-03	4.0E-04	6.3E-04	7.0E-04	6.9E-04	7.9E-04	7.2E-04	6.0E-04	6.9E-04	7.9E-04	5.4E-04	6.0E-04	6.0E-04	1.5E-03	8.4E-04	6.8E-04	6.7E-04	6.0E-04	9.2E-04
IGKV1-37	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.0E-05	5.7E-05	1.5E-04	0.0E+00	2.6E-05	0.0E+00	0.0E+00	5.4E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.2E-04	2.5E-05	0.0E+00	4.1E-05	2.7E-05	1.5E-04
IGLV7-43	8.3E-05	7.4E-05	1.9E-04	3.6E-05	0.0E+00	1.0E-04	3.4E-04	5.2E-04	3.0E-05	7.8E-05	5.5E-05	8.1E-05	2.5E-04	0.0E+00	1.2E-04	0.0E+00	3.0E-05	2.3E-04	7.1E-05	5.2E-05	8.5E-04	2.5E-04	1.0E-04	2.5E-04	5.5E-05	3.0E-04
IGLV3-27	4.1E-05	7.4E-05	1.1E-04	0.0E+00	3.4E-05	0.0E+00	3.2E-05	0.0E+00	0.0E+00	0.0E+00	1.5E-04	9.4E-05	6.1E-05	1.3E-04	1.1E-04	1.4E-04	1.2E-04	1.3E-04	7.1E-05	0.0E+00	0.0E+00	1.2E-04	5.7E-05	2.3E-05	0.0E+00	0.0E+00
IGLV2-18	4.9E-04	9.5E-04	6.3E-04	9.4E-04	4.3E-04	7.7E-04	4.6E-04	3.7E-04	7.8E-04	6.7E-04	1.6E-04	2.6E-04	5.3E-04	4.3E-04	6.7E-04	3.9E-04	9.6E-04	7.6E-04	6.1E-04	6.5E-04	9.2E-04	4.8E-04	6.8E-04	3.7E-04	6.9E-04	5.2E-04
IGKV4-1	7.1E-04	8.1E-04	1.0E-03	5.7E-04	3.4E-04	6.2E-04	5.0E-04	6.7E-04	4.7E-04	4.5E-04	2.7E-05	2.2E-04	4.0E-04	2.3E-04	4.2E-04	2.0E-04	5.8E-04	6.2E-04	4.1E-04	5.1E-04	8.9E-04	6.6E-04	4.5E-04	5.6E-04	2.8E-04	8.1E-04
IGLV4-69	1.1E-04	7.4E-05	3.8E-04	1.1E-04	6.0E-04	1.6E-04	4.0E-04	5.8E-04	3.0E-05	5.2E-05	3.6E-04	3.9E-04	4.4E-04	3.2E-04	3.6E-04	5.0E-04	1.5E-04	2.3E-04	2.4E-04	1.4E-04	5.4E-04	2.4E-04	1.3E-04	4.8E-04	4.9E-05	2.2E-04
IGLV3-25	6.1E-04	1.2E-03	1.0E-03	8.2E-04	7.3E-04	9.0E-04	1.0E-03	9.3E-04	7.6E-04	9.3E-04	7.1E-04	7.0E-04	8.7E-04	6.3E-04	4.1E-04	7.8E-04	1.2E-03	4.9E-04	5.0E-04	8.3E-04	1.4E-03	9.1E-04	6.0E-04	7.8E-04	5.7E-04	1.3E-03
IGKV1-33	1.4E-04	2.5E-04	0.0E+00	0.0E+00	3.4E-05	1.8E-04	1.8E-04	0.0E+00	3.0E-05	2.6E-05	8.2E-05	4.8E-05	9.1E-05	2.8E-05	0.0E+00	0.0E+00	9.1E-05	0.0E+00	5.7E-05	7.8E-05	8.1E-05	1.1E-04	0.0E+00	1.6E-04	9.5E-05	0.0E+00
IGLV3-22	4.7E-04	1.3E-03	8.9E-04	5.7E-04	9.7E-04	7.9E-04	1.6E-03	2.0E-03	5.5E-04	5.4E-04	8.1E-04	9.4E-04	1.7E-03	8.0E-04	6.9E-04	9.0E-04	5.8E-04	5.6E-04	7.3E-04	5.6E-04	2.3E-03	1.4E-03	5.3E-04	1.4E-03	7.4E-04	1.7E-03
IGKV1D-43	4.1E-05	7.4E-05	1.4E-04	3.6E-05	1.0E-04	1.0E-04	1.8E-04	3.0E-04	0.0E+00	2.6E-05	0.0E+00	2.7E-05	1.9E-04	2.8E-05	8.6E-05	6.1E-05	8.4E-05	0.0E+00	0.0E+00	2.6E-05	4.9E-04	2.1E-04	3.2E-05	3.0E-04	2.8E-04	2.8E-04

IGKV6D-21	0.0E+00	0.0E+00	0.0E+00	6.4E-05	6.9E-05	1.1E-04	8.0E-05	4.1E-05	0.0E+00	2.6E-05	1.4E-04	1.6E-04	1.8E-04	1.9E-04	0.0E+00	1.7E-04	0.0E+00	0.0E+00	2.9E-05	0.0E+00	8.1E-05	1.3E-04	0.0E+00	7.0E-05	0.0E+00	0.0E+00
IGKV2-24	8.3E-05	7.4E-05	1.2E-04	0.0E+00	0.0E+00	0.0E+00	3.2E-05	1.5E-04	0.0E+00	1.3E-04	8.2E-05	2.7E-05	9.1E-05	8.4E-05	1.2E-04	3.0E-05	1.5E-04	2.1E-04	7.9E-05	2.6E-05	1.6E-04	5.1E-05	3.2E-05	7.0E-05	7.6E-05	1.2E-04
IGKV6-21	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.4E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.8E-05	0.0E+00	0.0E+00	8.1E-05	0.0E+00	0.0E+00	2.3E-05	0.0E+00	0.0E+00
IGLV3-21	3.7E-04	9.6E-04	2.7E-04	4.5E-04	6.7E-04	5.8E-04	6.3E-04	8.8E-04	4.1E-04	3.6E-04	6.3E-04	5.5E-04	5.3E-04	6.0E-04	2.8E-04	6.6E-04	3.8E-04	1.5E-04	3.7E-04	3.6E-04	1.2E-03	5.5E-04	2.6E-04	4.8E-04	4.4E-04	6.9E-04
IGKV3D-20	1.5E-04	2.0E-04	0.0E+00	0.0E+00	3.4E-05	0.0E+00	1.1E-04	1.7E-04	7.5E-05	2.6E-05	0.0E+00	0.0E+00	0.0E+00	2.8E-05	2.9E-05	3.0E-05	0.0E+00	5.6E-05	0.0E+00	5.2E-05	8.1E-05	6.3E-05	3.2E-05	9.6E-05	2.7E-05	1.2E-04
IGKV5-2	8.8E-04	1.3E-03	9.1E-04	7.2E-04	6.6E-04	8.1E-04	5.6E-04	8.6E-04	4.3E-04	5.8E-04	6.8E-04	6.3E-04	5.5E-04	6.1E-04	6.5E-04	6.7E-04	6.3E-04	7.1E-04	6.7E-04	5.8E-04	9.8E-04	4.9E-04	7.9E-04	6.3E-04	6.0E-04	7.0E-04
IGKV2D-40	3.4E-04	1.3E-04	3.6E-04	3.5E-04	6.1E-05	3.7E-04	7.5E-04	8.8E-04	2.7E-04	1.8E-04	6.8E-05	7.5E-05	8.4E-04	2.8E-05	1.8E-04	0.0E+00	3.2E-04	2.1E-04	2.1E-04	1.4E-04	1.4E-03	7.3E-04	2.0E-04	6.6E-04	3.3E-04	7.7E-04
IGLV2-8	8.3E-05	5.0E-04	2.0E-04	2.8E-04	1.2E-04	2.2E-04	3.5E-04	2.0E-04	7.5E-05	9.8E-05	6.8E-05	1.3E-04	5.4E-05	1.5E-04	5.7E-05	7.5E-05	2.1E-04	2.8E-05	8.6E-05	2.3E-04	3.9E-04	1.1E-04	9.0E-05	2.3E-04	1.0E-04	2.9E-04
IGLV3-19	0.0E+00	0.0E+00	4.4E-04	1.4E-04	5.2E-04	6.7E-05	2.4E-04	3.9E-04	6.1E-05	1.6E-04	2.4E-04	2.9E-04	5.0E-04	1.9E-04	2.1E-04	3.0E-04	2.4E-04	2.7E-04	5.1E-05	1.0E-04	4.0E-04	3.6E-04	1.2E-04	2.8E-04	8.2E-05	4.2E-04
IGKV1D-12	0.0E+00	0.0E+00	4.0E-05	6.4E-05	0.0E+00	6.7E-05	6.5E-05	2.6E-04	0.0E+00	7.8E-05	2.7E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.1E-05	3.0E-05	0.0E+00	0.0E+00	7.2E-05	0.0E+00	7.9E-05	9.0E-05	4.6E-05	0.0E+00	1.3E-04
IGKV1D-13	0.0E+00	1.5E-04	4.0E-05	2.0E-04	0.0E+00	1.8E-04	9.7E-05	4.1E-05	0.0E+00	1.6E-04	0.0E+00	0.0E+00	6.1E-05	0.0E+00	2.9E-05	3.0E-05	1.9E-04	5.0E-05	5.1E-05	1.6E-04	0.0E+00	7.1E-05	1.5E-04	7.0E-05	2.2E-04	9.1E-05
IGLV3-10	3.4E-04	3.9E-04	1.4E-04	3.4E-04	1.4E-04	4.1E-04	5.6E-04	7.6E-04	2.9E-04	4.4E-04	0.0E+00	2.7E-05	6.7E-04	2.8E-05	7.1E-05	9.1E-05	5.1E-04	2.3E-04	1.2E-04	2.9E-04	8.6E-04	4.8E-04	1.8E-04	5.6E-04	3.8E-04	5.4E-04
IGLV7-46	1.0E-03	1.7E-03	1.4E-03	9.3E-04	1.1E-03	8.9E-04	1.0E-03	1.1E-03	9.5E-04	8.6E-04	9.6E-04	9.6E-04	1.2E-03	8.0E-04	8.6E-04	9.6E-04	1.1E-03	9.8E-04	8.9E-04	9.1E-04	1.8E-03	8.2E-04	1.0E-03	7.2E-04	9.3E-04	8.8E-04
IGKV3-11	7.3E-05	5.9E-04	2.0E-04	1.4E-04	2.1E-04	4.1E-04	4.1E-04	6.0E-04	1.2E-04	2.3E-04	1.1E-04	1.3E-04	4.9E-04	1.1E-04	0.0E+00	1.5E-04	3.5E-04	1.1E-04	2.9E-05	2.5E-04	5.0E-04	2.6E-04	5.7E-05	3.3E-04	1.1E-04	5.2E-04
IGLV3-12	8.3E-04	1.2E-03	1.3E-03	9.1E-04	1.1E-03	1.1E-03	1.1E-03	1.3E-03	6.7E-04	8.1E-04	8.1E-04	8.5E-04	8.9E-04	8.6E-04	9.9E-04	9.8E-04	8.1E-04	9.7E-04	9.4E-04	7.7E-04	1.5E-03	8.2E-04	9.9E-04	9.7E-04	8.6E-04	1.0E-03
IGKV1D-16	3.1E-04	2.5E-04	3.6E-04	4.5E-04	9.6E-05	2.3E-04	5.1E-04	5.7E-04	3.0E-04	1.5E-04	4.9E-05	6.7E-05	2.2E-04	5.0E-05	3.1E-04	6.1E-05	3.3E-04	2.9E-04	3.0E-04	8.1E-05	7.5E-04	3.9E-04	2.6E-04	1.7E-04	2.2E-04	4.4E-04
IGKV1D-17	7.1E-04	9.0E-04	6.8E-04	5.7E-04	6.2E-04	6.3E-04	9.6E-04	1.1E-03	2.5E-04	3.4E-04	7.9E-04	6.1E-04	9.4E-04	6.6E-04	5.0E-04	5.4E-04	3.2E-04	4.7E-04	5.4E-04	3.1E-04	1.6E-03	8.9E-04	5.8E-04	8.0E-04	4.8E-04	8.9E-04
IGKV3-15	1.7E-04	6.4E-04	6.2E-04	1.8E-04	2.4E-04	4.4E-04	8.9E-04	8.3E-04	3.2E-04	2.9E-04	1.9E-04	2.7E-05	8.6E-04	5.6E-05	1.3E-04	9.1E-05	1.7E-04	2.3E-04	2.2E-04	4.7E-04	1.5E-03	7.7E-04	1.7E-04	7.0E-04	3.7E-04	9.3E-04
IGLV3-16	8.3E-05	3.5E-04	1.9E-04	0.0E+00	3.4E-05	2.7E-04	8.5E-04	1.1E-03	3.0E-05	5.2E-05	5.5E-05	5.4E-05	7.8E-04	0.0E+00	1.4E-04	6.1E-05	1.1E-04	2.8E-05	1.6E-04	1.0E-04	1.4E-03	7.4E-04	9.0E-05	6.3E-04	2.7E-04	8.6E-04
IGLV10-54	1.2E-03	2.1E-03	1.4E-03	1.1E-03	9.5E-04	1.2E-03	1.7E-03	1.9E-03	8.1E-04	8.8E-04	8.6E-04	7.7E-04	1.7E-03	8.2E-04	1.1E-03	1.0E-03	9.5E-04	1.2E-03	1.1E-03	8.6E-04	2.5E-03	1.4E-03	1.2E-03	1.4E-03	9.4E-04	1.6E-03
IGLV1-36	2.3E-04	4.4E-04	1.1E-04	2.4E-04	1.2E-04	3.5E-04	8.9E-04	1.2E-03	2.9E-04	2.3E-04	1.6E-04	1.0E-04	1.1E-03	1.6E-04	1.1E-04	1.1E-04	2.8E-04	5.0E-05	1.2E-04	2.9E-04	1.2E-03	7.0E-04	2.9E-04	8.4E-04	3.8E-04	8.1E-04
IGKV3D-11	6.6E-04	8.1E-04	9.1E-04	9.2E-04	1.2E-03	7.3E-04	6.4E-04	8.0E-04	5.6E-04	6.2E-04	7.6E-04	7.9E-04	5.8E-04	7.9E-04	8.0E-04	9.9E-04	8.0E-04	6.6E-04	7.3E-04	7.5E-04	1.2E-03	5.1E-04	8.2E-04	5.9E-04	7.2E-04	6.7E-04
IGKV3D-15	4.9E-04	1.0E-03	1.0E-03	1.2E-03	4.7E-04	8.7E-04	1.1E-03	1.2E-03	6.8E-04	5.9E-04	4.5E-04	3.9E-04	1.1E-03	4.1E-04	7.6E-04	7.3E-04	1.0E-03	7.5E-04	7.4E-04	8.5E-04	1.8E-03	1.0E-03	8.5E-04	9.1E-04	8.3E-04	1.0E-03
IGKV2D-30	0.0E+00	0.0E+00	1.5E-04	6.4E-05	8.5E-05	1.7E-04	0.0E+00	2.4E-04	0.0E+00	2.6E-05	1.8E-04	2.4E-04	1.7E-04	2.0E-04	2.6E-04	9.1E-05	6.1E-05	0.0E+00	1.1E-04	0.0E+00	3.6E-04	7.1E-05	9.0E-05	1.9E-04	2.7E-05	0.0E+00
IGLV5-48	6.1E-04	7.9E-04	6.9E-04	4.6E-04	7.4E-04	4.7E-04	6.7E-04	8.9E-04	7.6E-04	4.9E-04	5.6E-04	5.4E-04	7.6E-04	5.9E-04	5.2E-04	5.5E-04	7.5E-04	7.6E-04	5.5E-04	5.0E-04	1.0E-03	4.6E-04	6.9E-04	6.1E-04	5.0E-04	3.3E-04
IGLV2-33	4.1E-05	1.5E-04	0.0E+00	8.9E-05	3.4E-05	1.5E-04	3.5E-04	3.1E-04	3.0E-05	2.6E-05	0.0E+00	0.0E+00	3.1E-04	0.0E+00	5.7E-05	3.0E-05	1.1E-04	0.0E+00	2.9E-05	7.2E-05	5.2E-04	1.9E-04	3.2E-05	3.1E-04	7.6E-05	4.3E-04
IGLV3-1	6.7E-04	1.1E-03	2.0E-03	5.3E-04	9.2E-04	1.7E-03	1.4E-03	2.0E-03	6.1E-04	1.1E-03	8.3E-04	8.1E-04	1.7E-03	7.5E-04	4.9E-04	7.4E-04	9.4E-04	6.1E-04	4.2E-04	7.7E-04	2.0E-03	1.5E-03	3.9E-04	1.4E-03	6.7E-04	1.4E-03
IGLV5-45	1.2E-03	1.8E-03	1.3E-03	1.2E-03	7.7E-04	1.1E-03	1.4E-03	1.5E-03	9.8E-04	1.0E-03	8.2E-04	7.7E-04	1.4E-03	7.6E-04	1.0E-03	8.3E-04	1.2E-03	1.2E-03	1.0E-03	1.1E-03	2.4E-03	1.1E-03	1.1E-03	1.2E-03	1.1E-03	1.2E-03
IGLV5-52	0.0E+00	7.4E-05	0.0E+00	0.0E+00	6.9E-05	0.0E+00	3.2E-05	7.3E-05	0.0E+00	0.0E+00	6.8E-05	0.0E+00	5.4E-05	2.8E-05	5.7E-05	3.0E-05	0.0E+00	5.0E-05	0.0E+00	4.6E-05	8.1E-05	0.0E+00	5.7E-05	0.0E+00	2.7E-05	5.4E-05

IGLV3-9	0.0E+00	1.5E-04	4.0E-05	7.2E-05	6.1E-05	0.0E+00	1.5E-04	1.4E-04	0.0E+00	4.6E-05	1.1E-04	8.1E-05	1.7E-04	5.6E-05	0.0E+00	5.4E-05	6.1E-05	2.8E-05	0.0E+00	0.0E+00	3.2E-04	0.0E+00	0.0E+00	7.3E-05	2.7E-05	1.5E-04
IGKV3-20	4.1E-05	3.7E-04	5.8E-04	6.4E-05	1.7E-04	4.0E-04	2.8E-04	4.4E-04	1.8E-04	3.6E-04	3.2E-04	1.6E-04	4.8E-04	2.0E-04	7.9E-05	1.2E-04	5.0E-04	8.4E-05	8.9E-05	2.3E-04	4.7E-04	4.5E-04	0.0E+00	3.9E-04	8.2E-05	5.7E-04
IGKV6D-41	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	8.3E-05	1.6E-04	1.5E-04	1.4E-04	0.0E+00	4.9E-05	0.0E+00	9.4E-05	2.8E-05	0.0E+00	0.0E+00	3.0E-05	2.8E-05	5.7E-05	0.0E+00	2.0E-04	1.5E-04	0.0E+00	1.1E-04	0.0E+00	0.0E+00
IGLV4-3	8.3E-05	7.4E-05	0.0E+00	7.2E-05	1.4E-04	1.0E-04	3.3E-04	3.9E-04	3.0E-05	2.6E-05	0.0E+00	0.0E+00	3.7E-04	2.8E-05	2.9E-05	6.1E-05	1.2E-04	0.0E+00	0.0E+00	0.0E+00	4.2E-04	2.9E-04	0.0E+00	2.1E-04	4.9E-05	3.8E-04
TRBV10-2	2.4E-04	3.8E-04	6.1E-04	4.5E-04	2.2E-04	3.5E-04	2.1E-04	4.5E-04	5.6E-04	4.5E-04	4.2E-04	3.3E-04	2.9E-04	3.7E-04	5.3E-04	3.0E-04	5.1E-04	6.7E-04	5.3E-04	3.1E-04	3.6E-04	7.1E-05	6.6E-04	3.0E-04	3.5E-04	6.1E-05
TRBV10-1	9.1E-04	1.3E-03	9.7E-04	1.1E-03	8.5E-04	9.4E-04	9.2E-04	1.1E-03	1.0E-03	9.0E-04	7.2E-04	7.7E-04	8.4E-04	8.1E-04	9.0E-04	8.0E-04	9.8E-04	9.2E-04	8.0E-04	8.3E-04	1.6E-03	7.7E-04	1.1E-03	8.1E-04	8.8E-04	8.7E-04
TRBV25-1	4.1E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	5.7E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.2E-04	0.0E+00	2.9E-05	0.0E+00	3.0E-05	0.0E+00	0.0E+00	0.0E+00	8.1E-05	2.5E-05	0.0E+00	2.3E-05	0.0E+00	1.1E-04
TRBV3-1	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	1.3E-04	0.0E+00	0.0E+00	2.7E-05	0.0E+00	1.8E-04	0.0E+00	0.0E+00	0.0E+00	0.0E+00	5.0E-05	2.9E-05	2.6E-05	2.9E-04	4.5E-05	0.0E+00	5.8E-05	0.0E+00	1.1E-04
TRBV11-1	1.5E-04	4.1E-04	7.2E-05	4.2E-04	0.0E+00	4.9E-04	6.2E-04	7.0E-04	3.0E-04	3.6E-04	0.0E+00	0.0E+00	5.9E-04	0.0E+00	2.9E-05	0.0E+00	3.7E-04	7.8E-05	1.2E-04	2.7E-04	6.9E-04	5.5E-04	1.7E-04	5.6E-04	3.3E-04	6.1E-04
TRBV20-1	1.2E-03	1.7E-03	1.0E-03	9.8E-04	7.7E-04	1.1E-03	6.3E-04	7.4E-04	8.6E-04	9.0E-04	7.6E-04	7.5E-04	5.9E-04	7.8E-04	8.2E-04	8.4E-04	9.4E-04	8.8E-04	9.0E-04	8.1E-04	1.3E-03	4.1E-04	9.4E-04	5.0E-04	9.2E-04	3.9E-04
TRBV4-1	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.1E-05	6.0E-05	2.4E-04	5.7E-04	3.0E-05	0.0E+00	2.7E-05	7.5E-05	2.8E-04	6.9E-05	0.0E+00	3.0E-05	3.0E-05	2.7E-04	0.0E+00	0.0E+00	6.0E-04	2.6E-04	0.0E+00	2.7E-04	1.9E-04	2.7E-04
TRBV4-2	1.1E-04	7.4E-05	0.0E+00	0.0E+00	1.4E-04	1.8E-04	2.1E-04	2.6E-04	5.4E-05	5.2E-05	0.0E+00	8.4E-05	2.9E-04	2.8E-05	0.0E+00	3.0E-05	0.0E+00	5.0E-05	0.0E+00	0.0E+00	8.1E-05	2.4E-04	5.7E-05	5.8E-05	5.5E-05	2.4E-04
TRBV30	7.6E-04	1.5E-03	8.5E-04	9.8E-04	5.2E-04	1.0E-03	1.5E-03	1.5E-03	8.9E-04	8.2E-04	4.0E-04	5.5E-04	1.3E-03	4.7E-04	6.7E-04	5.4E-04	9.0E-04	7.1E-04	7.8E-04	7.8E-04	2.4E-03	1.1E-03	7.5E-04	1.1E-03	9.5E-04	1.3E-03
TRBV24-1	5.0E-04	9.2E-04	6.5E-04	6.6E-04	1.1E-04	5.8E-04	4.8E-04	5.8E-04	5.5E-04	6.3E-04	2.3E-04	2.4E-04	4.1E-04	2.4E-04	4.9E-04	1.9E-04	6.3E-04	5.8E-04	5.5E-04	6.4E-04	8.1E-04	3.8E-04	6.1E-04	4.0E-04	5.9E-04	4.4E-04
TRBV5-1	0.0E+00	0.0E+00	9.9E-05	3.6E-05	0.0E+00	0.0E+00	3.2E-05	1.3E-04	0.0E+00	5.2E-05	0.0E+00	0.0E+00	3.0E-05	2.8E-05	5.1E-05	5.4E-05	0.0E+00	2.8E-05	1.2E-04	7.8E-05	0.0E+00	7.6E-05	2.0E-04	4.1E-05	2.7E-05	9.4E-05
TRBV5-6	8.3E-05	4.3E-04	4.0E-05	3.7E-04	0.0E+00	2.4E-04	1.0E-03	1.1E-03	2.7E-04	1.5E-04	0.0E+00	0.0E+00	9.7E-04	2.8E-05	5.7E-05	3.0E-05	2.3E-04	1.1E-04	7.1E-05	1.3E-04	1.3E-03	7.8E-04	1.1E-04	7.7E-04	3.0E-04	8.6E-04
TRBV19	7.8E-04	7.6E-04	2.9E-04	3.3E-04	6.1E-04	5.9E-04	4.0E-04	3.8E-04	3.3E-04	3.8E-04	5.2E-04	4.5E-04	1.5E-04	4.8E-04	2.5E-04	5.2E-04	3.4E-04	5.0E-04	2.3E-04	2.1E-04	5.7E-04	2.9E-04	2.5E-04	2.6E-04	4.0E-04	3.8E-04
TRBV5-4	4.1E-05	7.4E-05	0.0E+00	0.0E+00	0.0E+00	1.4E-04	7.6E-04	1.2E-03	0.0E+00	2.6E-05	0.0E+00	0.0E+00	8.1E-04	2.8E-05	0.0E+00	0.0E+00	0.0E+00	2.7E-04	0.0E+00	2.6E-05	9.0E-04	8.1E-04	3.2E-05	4.7E-04	1.5E-04	9.2E-04
TRBV5-5	8.4E-04	1.2E-03	9.0E-04	7.1E-04	6.4E-04	9.3E-04	1.0E-03	1.2E-03	7.5E-04	6.0E-04	3.7E-04	6.0E-04	1.2E-03	5.7E-04	7.0E-04	5.9E-04	5.9E-04	6.5E-04	6.5E-04	7.4E-04	1.5E-03	1.0E-03	6.6E-04	8.3E-04	6.5E-04	9.9E-04
TRBV6-8	4.1E-05	2.5E-04	0.0E+00	0.0E+00	0.0E+00	1.7E-04	1.0E-03	1.2E-03	0.0E+00	5.2E-05	0.0E+00	0.0E+00	1.1E-03	0.0E+00	0.0E+00	6.1E-05	0.0E+00	2.8E-05	0.0E+00	0.0E+00	1.5E-03	8.6E-04	0.0E+00	8.0E-04	2.8E-04	1.1E-03
TRBV9	6.9E-04	5.0E-04	3.5E-04	5.2E-04	1.3E-04	6.6E-04	6.4E-04	7.4E-04	5.5E-04	5.0E-04	1.3E-04	3.5E-04	5.5E-04	1.4E-04	4.5E-04	1.1E-04	5.0E-04	4.3E-04	4.2E-04	3.8E-04	9.5E-04	4.8E-04	5.5E-04	4.6E-04	6.0E-04	5.4E-04
TRBV6-6	8.8E-04	1.4E-03	9.0E-04	7.3E-04	6.3E-04	8.5E-04	1.3E-03	1.6E-03	6.3E-04	7.3E-04	3.8E-04	4.8E-04	1.3E-03	5.4E-04	7.6E-04	5.5E-04	5.9E-04	6.9E-04	7.8E-04	5.7E-04	1.8E-03	1.0E-03	7.6E-04	9.2E-04	7.6E-04	1.1E-03
TRBV6-5	7.3E-05	0.0E+00	7.2E-05	6.4E-05	3.4E-05	6.0E-05	3.2E-05	7.3E-05	1.6E-04	4.6E-05	1.0E-04	5.4E-05	0.0E+00	6.9E-05	5.7E-05	6.1E-05	6.1E-05	2.8E-05	0.0E+00	7.2E-05	1.6E-04	0.0E+00	3.2E-05	2.3E-05	2.7E-05	0.0E+00
TRBV6-4	3.5E-04	4.8E-04	3.9E-04	4.7E-04	3.4E-05	3.7E-04	5.1E-04	5.9E-04	5.5E-04	4.5E-04	1.3E-04	1.1E-04	5.7E-04	1.4E-04	4.8E-04	1.1E-04	4.8E-04	5.8E-04	4.3E-04	4.0E-04	8.0E-04	3.9E-04	5.8E-04	3.9E-04	4.0E-04	4.0E-04
TRBV28	4.1E-05	7.4E-05	0.0E+00	3.6E-05	3.4E-05	6.0E-05	1.2E-04	4.1E-05	0.0E+00	4.6E-05	0.0E+00	0.0E+00	7.5E-05	2.8E-05	5.1E-05	3.0E-05	0.0E+00	2.8E-05	0.0E+00	0.0E+00	1.4E-04	1.2E-04	0.0E+00	2.3E-05	2.7E-05	1.6E-04
TRBV2	0.0E+00	0.0E+00	8.1E-05	1.1E-04	6.1E-05	1.1E-04	1.6E-04	3.3E-04	0.0E+00	2.6E-05	0.0E+00	0.0E+00	2.3E-04	2.8E-05	0.0E+00	0.0E+00	5.4E-05	7.8E-05	5.7E-05	8.1E-05	5.7E-04	1.2E-04	0.0E+00	1.6E-04	2.7E-05	2.1E-04
TRBV6-1	0.0E+00	7.4E-05	4.0E-05	0.0E+00	0.0E+00	6.7E-05	3.2E-05	4.1E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.0E-05	0.0E+00	0.0E+00	0.0E+00	3.0E-05	0.0E+00	0.0E+00	4.6E-05	1.4E-04	2.5E-05	0.0E+00	8.7E-05	0.0E+00	3.0E-05
TRBV7-3	4.9E-04	7.5E-04	9.6E-04	8.1E-04	3.6E-04	7.7E-04	1.1E-03	1.4E-03	9.5E-04	6.9E-04	5.7E-04	5.5E-04	1.1E-03	4.8E-04	8.6E-04	5.6E-04	7.8E-04	9.4E-04	8.9E-04	6.0E-04	1.8E-03	8.8E-04	9.4E-04	9.2E-04	6.8E-04	1.1E-03
TRBV27	4.1E-05	1.5E-04	8.1E-05	0.0E+00	3.4E-05	6.0E-05	2.8E-04	4.9E-04	0.0E+00	2.6E-05	0.0E+00	2.7E-05	2.4E-04	2.8E-05	0.0E+00	3.0E-05	0.0E+00	0.0E+00	5.7E-05	0.0E+00	4.9E-04	4.5E-04	0.0E+00	2.1E-04	4.9E-05	5.0E-04

TRBV7-7	8.3E-04	1.1E-03	7.5E-04	7.3E-04	6.3E-04	7.9E-04	5.9E-04	9.8E-04	7.1E-04	6.6E-04	5.8E-04	5.2E-04	6.4E-04	5.5E-04	7.0E-04	5.8E-04	7.1E-04	6.2E-04	6.6E-04	6.2E-04	1.5E-03	7.4E-04	6.9E-04	5.9E-04	6.7E-04	4.8E-04
TRBV7-6	0.0E+00	0.0E+00	7.2E-05	0.0E+00	0.0E+00	6.0E-05	3.3E-04	4.5E-04	0.0E+00	0.0E+00	0.0E+00	0.0E+00	5.1E-04	2.8E-05	0.0E+00	0.0E+00	3.0E-05	0.0E+00	0.0E+00	0.0E+00	5.8E-04	2.6E-04	0.0E+00	2.1E-04	2.7E-05	2.5E-04
TRBV29-1	6.3E-04	7.4E-04	8.1E-05	0.0E+00	6.9E-05	2.7E-04	6.5E-05	7.3E-05	0.0E+00	0.0E+00	2.5E-04	1.7E-04	8.4E-05	1.8E-04	2.9E-05	5.4E-05	6.1E-05	0.0E+00	2.9E-05	0.0E+00	0.0E+00	7.1E-05	0.0E+00	6.5E-05	2.7E-04	5.4E-05
TRBV7-4	7.3E-05	1.8E-04	1.1E-04	1.7E-04	0.0E+00	2.0E-04	3.8E-04	4.0E-04	1.9E-04	1.5E-04	0.0E+00	0.0E+00	3.1E-04	0.0E+00	7.9E-05	0.0E+00	1.3E-04	1.1E-04	1.6E-04	8.1E-05	2.9E-04	2.4E-04	2.0E-04	1.6E-04	1.7E-04	3.2E-04
TRAV1-1	4.9E-04	7.4E-04	7.6E-04	7.0E-04	3.4E-04	6.6E-04	8.3E-04	1.1E-03	6.5E-04	6.3E-04	2.8E-04	2.3E-04	8.9E-04	3.1E-04	5.7E-04	3.0E-04	6.5E-04	6.5E-04	6.0E-04	6.0E-04	1.5E-03	7.9E-04	5.7E-04	7.6E-04	6.1E-04	8.6E-04
TRAV1-2	5.5E-04	7.0E-04	5.0E-04	4.0E-04	5.4E-04	4.3E-04	5.7E-04	6.0E-04	5.3E-04	3.3E-04	5.6E-04	5.3E-04	4.6E-04	5.8E-04	4.5E-04	5.8E-04	3.3E-04	4.5E-04	5.3E-04	3.5E-04	1.0E-03	5.1E-04	4.8E-04	4.6E-04	3.9E-04	5.4E-04
TRAV26-2	0.0E+00	0.0E+00	1.4E-04	0.0E+00	2.6E-04	3.4E-05	3.2E-05	0.0E+00	0.0E+00	2.6E-05	1.9E-04	2.6E-04	0.0E+00	1.9E-04	1.4E-04	2.7E-04	3.0E-05	8.4E-05	2.0E-04	2.6E-05	0.0E+00	0.0E+00	1.1E-04	0.0E+00	0.0E+00	0.0E+00
TRAV9-2	9.4E-04	1.4E-03	1.1E-03	1.2E-03	8.4E-04	1.0E-03	1.2E-03	1.5E-03	9.4E-04	9.2E-04	8.5E-04	7.4E-04	1.1E-03	8.0E-04	9.4E-04	8.3E-04	1.0E-03	9.7E-04	9.6E-04	8.1E-04	2.0E-03	1.1E-03	1.1E-03	1.0E-03	8.8E-04	1.3E-03
TRAV9-1	4.1E-05	0.0E+00	7.2E-05	1.6E-04	0.0E+00	9.4E-05	5.0E-04	5.7E-04	3.0E-05	2.6E-05	5.5E-05	2.7E-05	5.9E-04	0.0E+00	1.9E-04	0.0E+00	0.0E+00	2.8E-05	9.9E-05	5.2E-05	6.9E-04	4.8E-04	5.7E-05	4.3E-04	1.5E-04	5.0E-04
TRAV30	9.9E-04	1.7E-03	1.1E-03	8.5E-04	9.5E-04	1.1E-03	8.4E-04	9.8E-04	6.8E-04	8.3E-04	7.5E-04	7.4E-04	8.6E-04	7.9E-04	7.9E-04	7.9E-04	1.1E-03	1.0E-03	7.9E-04	8.2E-04	1.7E-03	6.8E-04	8.1E-04	4.7E-04	9.5E-04	8.7E-04
TRAV10	2.0E-04	1.8E-04	3.6E-04	2.4E-04	3.9E-04	2.9E-04	3.3E-04	4.7E-04	3.3E-04	2.7E-04	5.9E-04	5.9E-04	3.8E-04	6.3E-04	2.5E-04	5.9E-04	2.1E-04	4.3E-04	3.5E-04	2.1E-04	5.2E-04	3.5E-04	3.3E-04	3.3E-04	2.1E-04	3.5E-04
TRAV2	4.1E-05	1.5E-04	4.0E-05	0.0E+00	3.4E-05	2.2E-04	2.7E-04	3.6E-04	0.0E+00	7.2E-05	0.0E+00	0.0E+00	3.5E-04	5.6E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	4.1E-04	2.1E-04	3.2E-05	1.6E-04	1.7E-04	2.1E-04
TRAV34	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.0E-05	1.9E-04	7.3E-05	0.0E+00	4.6E-05	8.6E-05	2.7E-05	2.2E-04	0.0E+00	2.9E-05	3.0E-05	0.0E+00	0.0E+00	0.0E+00	5.2E-05	2.5E-04	1.6E-04	3.2E-05	7.3E-05	8.6E-05	2.1E-04
TRAV26-1	6.9E-04	1.4E-03	8.2E-04	9.2E-04	6.3E-04	8.2E-04	5.5E-04	6.4E-04	7.6E-04	7.7E-04	6.5E-04	5.5E-04	3.5E-04	6.0E-04	6.9E-04	6.5E-04	8.6E-04	6.7E-04	7.4E-04	7.7E-04	1.2E-03	4.5E-04	8.3E-04	5.3E-04	8.1E-04	6.4E-04
TRAV8-7	7.5E-04	1.1E-03	7.8E-04	6.7E-04	6.4E-04	6.4E-04	7.2E-04	7.6E-04	6.9E-04	4.9E-04	4.8E-04	5.7E-04	7.1E-04	5.6E-04	6.3E-04	5.8E-04	6.6E-04	6.1E-04	6.3E-04	5.5E-04	1.2E-03	6.0E-04	6.4E-04	5.5E-04	6.1E-04	6.6E-04
TRAV8-6	8.8E-04	1.0E-03	7.4E-04	6.0E-04	6.5E-04	7.1E-04	6.3E-04	8.6E-04	6.1E-04	4.9E-04	6.1E-04	6.0E-04	6.2E-04	6.5E-04	6.7E-04	6.4E-04	5.1E-04	6.1E-04	6.1E-04	5.1E-04	1.2E-03	5.5E-04	6.2E-04	6.0E-04	5.2E-04	6.2E-04
TRAV39	4.1E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.5E-05	1.3E-04	9.4E-05	0.0E+00	0.0E+00	4.8E-05	5.0E-04	0.0E+00	0.0E+00	0.0E+00	8.4E-05	0.0E+00	0.0E+00	0.0E+00	1.6E-04	6.3E-05	6.5E-05	2.3E-05	2.7E-05	1.1E-04
TRAV8-4	1.4E-03	2.0E-03	1.4E-03	1.2E-03	1.0E-03	1.0E-03	1.2E-03	1.5E-03	1.2E-03	9.5E-04	1.1E-03	1.1E-03	1.2E-03	1.1E-03	1.1E-03	1.1E-03	1.1E-03	9.7E-04	1.0E-03	8.3E-04	2.3E-03	1.1E-03	1.2E-03	9.6E-04	1.1E-03	1.2E-03
TRAV8-3	8.0E-04	1.2E-03	7.7E-04	7.4E-04	6.9E-04	7.0E-04	6.9E-04	8.1E-04	5.3E-04	5.9E-04	6.2E-04	6.2E-04	7.3E-04	6.2E-04	7.1E-04	6.0E-04	6.4E-04	6.2E-04	6.4E-04	6.2E-04	1.3E-03	6.2E-04	6.9E-04	6.5E-04	6.4E-04	6.9E-04
TRAV19	5.9E-04	8.0E-04	6.3E-04	4.4E-04	3.6E-04	5.0E-04	5.4E-04	7.2E-04	4.5E-04	4.0E-04	5.4E-04	3.3E-04	5.5E-04	6.1E-04	5.2E-04	6.0E-04	4.4E-04	5.2E-04	4.9E-04	4.7E-04	9.5E-04	4.8E-04	5.2E-04	6.3E-04	4.1E-04	5.4E-04
TRAV8-2	8.7E-04	1.1E-03	1.0E-03	5.5E-04	1.1E-03	7.1E-04	7.7E-04	1.1E-03	4.9E-04	5.5E-04	9.5E-04	9.5E-04	9.5E-04	9.6E-04	7.9E-04	1.1E-03	3.9E-04	8.7E-04	8.9E-04	4.7E-04	1.4E-03	7.2E-04	9.3E-04	7.4E-04	5.2E-04	8.6E-04
TRAV18	0.0E+00	0.0E+00	4.0E-05	0.0E+00	3.4E-05	3.4E-05	1.9E-04	1.0E-04	0.0E+00	0.0E+00	6.8E-05	0.0E+00	2.2E-04	2.8E-05	5.1E-05	5.4E-05	0.0E+00	7.8E-05	7.1E-05	0.0E+00	2.9E-04	2.2E-04	5.7E-05	8.7E-05	2.7E-05	2.7E-04
TRAV8-1	4.1E-05	0.0E+00	1.1E-04	0.0E+00	0.0E+00	3.4E-05	2.3E-04	2.2E-04	0.0E+00	0.0E+00	0.0E+00	2.7E-05	1.7E-04	2.8E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.9E-04	2.3E-04	0.0E+00	2.1E-04	2.7E-05	2.2E-04
TRAV17	4.1E-05	0.0E+00	0.0E+00	3.6E-05	0.0E+00	0.0E+00	3.2E-05	1.9E-04	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.0E-05	5.0E-05	0.0E+00	0.0E+00	0.0E+00	2.8E-05	0.0E+00	0.0E+00	8.1E-05	7.1E-05	3.2E-05	4.1E-05	6.8E-05	3.0E-05
TRAV16	7.3E-05	0.0E+00	0.0E+00	3.6E-05	3.4E-05	3.4E-05	3.2E-05	4.1E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	1.5E-04	0.0E+00	2.9E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.2E-05	4.1E-05	0.0E+00	0.0E+00
TRAV12-3	0.0E+00	0.0E+00	4.0E-05	3.6E-05	0.0E+00	0.0E+00	8.0E-05	4.1E-05	3.0E-05	2.6E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.9E-05	3.0E-05	8.4E-05	1.4E-04	5.1E-05	7.2E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.0E-05
TRAV12-1	4.1E-05	0.0E+00	4.5E-04	1.1E-04	6.7E-04	2.0E-04	1.4E-04	3.3E-04	3.0E-05	2.5E-04	3.0E-04	4.1E-04	9.4E-05	3.4E-04	2.8E-04	5.3E-04	2.2E-04	1.4E-04	3.3E-04	1.5E-04	3.4E-04	9.5E-05	1.2E-04	9.9E-05	2.5E-04	1.9E-04
TRAV12-2	1.0E-03	1.6E-03	1.4E-03	1.0E-03	1.1E-03	8.7E-04	1.2E-03	1.4E-03	9.8E-04	7.6E-04	1.1E-03	1.1E-03	1.1E-03	1.1E-03	1.2E-03	1.0E-03	7.7E-04	9.9E-04	1.1E-03	7.3E-04	1.9E-03	9.9E-04	1.1E-03	9.2E-04	8.9E-04	1.1E-03
TRAV5	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	1.2E-04	1.9E-04	0.0E+00	2.6E-05	6.8E-05	0.0E+00	2.2E-04	2.8E-05	2.9E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	8.1E-05	1.5E-04	0.0E+00	1.3E-04	0.0E+00	1.4E-04

TRAV6	1.1E-03	1.6E-03	1.2E-03	1.2E-03	1.1E-03	1.2E-03	1.1E-03	1.2E-03	1.1E-03	9.5E-04	8.5E-04	8.7E-04	1.3E-03	8.5E-04	9.9E-04	8.6E-04	9.8E-04	8.5E-04	9.9E-04	9.1E-04	2.1E-03	1.1E-03	1.1E-03	9.3E-04	1.0E-03	1.2E-03
TRAV3	0.0E+00	7.4E-05	1.7E-04	1.1E-04	3.4E-05	0.0E+00	0.0E+00	0.0E+00	3.0E-05	4.6E-05	2.7E-05	0.0E+00	6.1E-05	0.0E+00	7.1E-05	0.0E+00	7.5E-05	8.7E-05	1.1E-04	6.4E-05	0.0E+00	0.0E+00	8.0E-05	0.0E+00	0.0E+00	3.0E-05
TRAV4	4.1E-05	1.3E-04	0.0E+00	8.9E-05	2.3E-04	1.1E-04	1.0E-04	2.2E-04	0.0E+00	8.1E-05	1.2E-04	1.0E-04	6.1E-05	5.6E-05	2.9E-05	1.6E-04	7.5E-05	2.8E-05	0.0E+00	9.0E-05	4.1E-04	7.9E-05	6.5E-05	8.8E-05	0.0E+00	1.1E-04
TRAV7	4.1E-05	2.5E-04	1.5E-04	1.5E-04	0.0E+00	1.8E-04	9.0E-05	1.6E-04	1.5E-04	1.4E-04	0.0E+00	5.4E-05	2.4E-04	0.0E+00	1.2E-04	0.0E+00	2.1E-04	1.3E-04	5.1E-05	1.1E-04	8.1E-05	1.4E-04	9.7E-05	2.0E-04	1.2E-04	1.3E-04
TRAV40	0.0E+00	2.2E-04	2.1E-04	2.6E-04	3.4E-05	1.8E-04	1.3E-04	2.9E-04	2.0E-04	7.8E-05	5.5E-05	2.7E-05	3.2E-04	5.6E-05	1.1E-04	0.0E+00	1.8E-04	1.7E-04	1.9E-04	1.5E-04	5.3E-04	1.9E-04	1.2E-04	1.5E-04	7.6E-05	2.0E-04
TRAV24	4.1E-05	7.4E-05	0.0E+00	0.0E+00	2.0E-04	3.4E-05	3.2E-05	4.1E-05	0.0E+00	9.2E-05	4.9E-05	2.7E-05	3.0E-05	7.8E-05	2.9E-05	1.1E-04	8.4E-05	2.8E-05	2.9E-05	2.6E-05	2.2E-04	7.9E-05	0.0E+00	4.1E-05	2.7E-05	0.0E+00
TRAV13-1	7.7E-04	1.0E-03	8.4E-04	8.5E-04	5.6E-04	7.9E-04	6.3E-04	6.6E-04	6.2E-04	6.2E-04	4.8E-04	5.8E-04	6.4E-04	5.5E-04	6.6E-04	5.6E-04	6.4E-04	6.8E-04	8.1E-04	6.2E-04	1.1E-03	5.0E-04	6.7E-04	5.1E-04	6.1E-04	5.9E-04
TRAV41	0.0E+00	0.0E+00	8.1E-05	0.0E+00	3.4E-05	3.4E-05	2.7E-04	4.1E-04	0.0E+00	5.2E-05	2.7E-05	0.0E+00	1.4E-04	1.3E-04	0.0E+00	3.0E-05	0.0E+00	6.9E-05	2.9E-05	2.6E-05	3.3E-04	2.5E-04	3.2E-05	1.8E-04	9.8E-05	2.3E-04
TRAV25	6.4E-04	9.3E-04	7.7E-04	7.2E-04	3.0E-04	6.3E-04	5.1E-04	5.5E-04	6.1E-04	6.3E-04	3.2E-04	1.7E-04	5.7E-04	2.5E-04	6.5E-04	1.9E-04	6.4E-04	5.7E-04	6.1E-04	5.8E-04	1.4E-03	4.2E-04	6.0E-04	2.1E-04	6.4E-04	4.6E-04
TRAV13-2	2.3E-04	4.1E-04	2.4E-04	2.4E-04	0.0E+00	4.7E-04	1.1E-03	1.3E-03	1.3E-04	2.5E-04	0.0E+00	2.7E-05	9.8E-04	2.8E-05	1.2E-04	5.4E-05	2.2E-04	1.9E-04	1.7E-04	1.7E-04	1.7E-03	1.0E-03	2.1E-04	8.5E-04	3.1E-04	1.0E-03
TRAV20	4.1E-05	1.3E-04	1.1E-04	2.4E-04	3.4E-05	2.1E-04	8.0E-05	2.2E-04	7.5E-05	2.1E-04	8.6E-05	5.4E-05	1.6E-04	6.9E-05	5.1E-05	3.0E-05	1.5E-04	1.4E-04	1.2E-04	2.2E-04	4.8E-04	7.9E-05	8.0E-05	2.4E-04	1.6E-04	1.5E-04
TRAV22	0.0E+00	7.4E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.2E-05	8.3E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	8.4E-05	7.8E-05	7.1E-05	3.0E-05	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.2E-04	5.1E-05	3.2E-05	6.5E-05	2.7E-05	0.0E+00
TRAV21	7.9E-04	1.1E-03	8.9E-04	7.3E-04	7.1E-04	7.0E-04	6.7E-04	6.5E-04	6.6E-04	6.0E-04	6.5E-04	6.4E-04	5.7E-04	6.5E-04	6.6E-04	6.6E-04	6.9E-04	6.3E-04	7.0E-04	6.3E-04	1.2E-03	5.7E-04	7.3E-04	4.8E-04	6.2E-04	5.7E-04
TRAV27	8.1E-04	1.3E-03	9.5E-04	8.2E-04	7.0E-04	8.4E-04	1.2E-03	1.6E-03	7.0E-04	7.6E-04	6.7E-04	6.2E-04	1.2E-03	6.6E-04	7.6E-04	6.7E-04	9.6E-04	9.6E-04	7.7E-04	7.3E-04	2.0E-03	1.1E-03	8.8E-04	1.0E-03	8.9E-04	1.2E-03
TRAV38-1	7.5E-04	1.2E-03	6.4E-04	6.0E-04	6.9E-04	5.9E-04	9.8E-04	1.2E-03	5.1E-04	4.7E-04	6.4E-04	6.2E-04	1.1E-03	6.5E-04	5.4E-04	6.5E-04	4.8E-04	4.8E-04	5.2E-04	4.7E-04	1.7E-03	8.6E-04	5.7E-04	7.8E-04	5.0E-04	9.0E-04
TRGV8	5.1E-05	0.0E+00	0.0E+00	0.0E+00	7.8E-05	1.3E-04	4.5E-04	6.7E-04	0.0E+00	0.0E+00	0.0E+00	6.9E-05	3.8E-04	0.0E+00	0.0E+00	7.5E-05	3.7E-05	0.0E+00	0.0E+00	0.0E+00	8.3E-04	3.6E-04	3.9E-05	4.2E-04	1.2E-04	6.0E-04
TRGV5	2.7E-04	3.6E-04	2.5E-04	1.8E-04	0.0E+00	2.3E-04	5.4E-04	9.3E-04	1.8E-04	1.3E-04	1.0E-04	0.0E+00	6.3E-04	1.2E-04	1.3E-04	6.6E-05	1.9E-04	6.7E-05	9.8E-05	9.2E-05	1.2E-03	3.7E-04	2.0E-04	6.4E-04	3.1E-04	6.0E-04
TRGV9	8.0E-04	1.2E-03	1.2E-03	8.4E-04	7.2E-04	8.8E-04	4.3E-04	7.6E-04	4.4E-04	7.2E-04	6.5E-04	6.2E-04	6.9E-04	7.0E-04	9.2E-04	7.1E-04	6.9E-04	8.2E-04	7.7E-04	8.1E-04	9.7E-04	3.3E-04	8.9E-04	3.7E-04	5.8E-04	3.8E-04
TRGV10	8.5E-04	8.5E-04	3.3E-04	1.8E-04	3.8E-04	6.3E-04	1.1E-03	1.5E-03	2.3E-04	3.0E-04	4.9E-04	4.7E-04	1.1E-03	6.1E-04	2.6E-04	3.2E-04	2.0E-04	2.9E-04	4.0E-04	1.0E-04	2.0E-03	1.1E-03	3.1E-04	7.9E-04	4.0E-04	1.1E-03
TRGV11	3.1E-04	7.2E-04	6.9E-04	4.4E-04	1.2E-03	4.7E-04	8.8E-04	8.9E-04	5.0E-04	4.2E-04	7.4E-04	8.4E-04	7.7E-04	7.2E-04	6.6E-04	9.3E-04	4.1E-04	5.1E-04	5.1E-04	5.6E-04	1.5E-03	7.4E-04	8.1E-04	6.0E-04	8.2E-04	7.0E-04
TRGV3	4.9E-05	5.0E-04	4.8E-04	2.4E-04	5.0E-04	5.7E-04	1.3E-03	1.8E-03	1.5E-04	2.3E-04	5.7E-04	4.4E-04	1.2E-03	4.5E-04	5.5E-04	4.4E-04	2.5E-04	2.5E-04	3.9E-04	1.9E-04	2.3E-03	1.4E-03	4.5E-04	1.3E-03	2.9E-04	1.5E-03
TRGV4	1.1E-03	1.8E-03	1.1E-03	1.2E-03	1.3E-03	1.4E-03	1.8E-03	2.1E-03	8.6E-04	9.7E-04	9.9E-04	1.0E-03	1.6E-03	1.0E-03	9.4E-04	9.9E-04	9.9E-04	8.8E-04	8.5E-04	1.1E-03	2.9E-03	1.3E-03	1.1E-03	1.6E-03	1.4E-03	1.6E-03
TRGV1	0.0E+00	1.5E-04	0.0E+00	2.1E-04	0.0E+00	4.1E-05	8.3E-05	0.0E+00	0.0E+00	1.5E-04	3.7E-05	3.4E-05	0.0E+00	0.0E+00	0.0E+00	3.9E-05	9.1E-05	0.0E+00	0.0E+00	2.5E-04	0.0E+00	0.0E+00	3.9E-05	3.1E-05	1.3E-04	3.7E-05
TRGV2	1.1E-03	2.0E-03	1.4E-03	1.3E-03	1.2E-03	1.3E-03	1.2E-03	1.3E-03	1.2E-03	1.0E-03	1.2E-03	1.1E-03	1.3E-03	1.2E-03	1.1E-03	1.1E-03	1.2E-03	1.2E-03	1.1E-03	1.0E-03	2.0E-03	9.7E-04	1.2E-03	9.2E-04	1.1E-03	1.0E-03

REFERENCES

1. Plotkin, S.A., 2005. Vaccines: past, present and future. Nature medicine, 11, pp.S5-S11.
2. Market, E. and Papavasiliou, F.N., 2003. V (D) J recombination and the evolution of the adaptive immune system. PLoS Biol, 1(1), p.e16.
3. Cooper, M.D. and Alder, M.N., 2006. The evolution of adaptive immune systems. Cell, 124(4), pp.815-822.
4. Litman, G.W., Rast, J.P. and Fugmann, S.D., 2010. The origins of vertebrate adaptive immunity. Nature Reviews Immunology, 10(8), pp.543-553.
5. Saha, N.R., Smith, J. and Amemiya, C.T., 2010, February. Evolution of adaptive immune recognition in jawless vertebrates. In Seminars in immunology (Vol. 22, No. 1, pp. 25-33). Academic Press.
6. Pancer, Z., Amemiya, C.T., Ehrhardt, G.R., Ceitlin, J., Gartland, G.L. and Cooper, M.D., 2004. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature, 430(6996), pp.174-180.
7. Alder, M.N., Rogozin, I.B., Iyer, L.M., Glazko, G.V., Cooper, M.D. and Pancer, Z., 2005. Diversity and function of adaptive immune receptors in a jawless vertebrate. Science, 310(5756), pp.1970-1973.
8. Scott-Browne, J.P., Crawford, F., Young, M.H., Kappler, J.W., Marrack, P. and Gapin, L., 2011. Evolutionarily conserved features contribute to αβ T cell receptor specificity. Immunity, 35(4), pp.526-535.
9. Litman, G.W., Anderson, M.K. and Rast, J.P., 1999. Evolution of antigen binding receptors. Annual review of immunology, 17(1), pp.109-147.
10. Flajnik, M.F. and Kasahara, M., 2001. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. Immunity, 15(3), pp.351-362.
11. Flajnik, M.F., 2002. Comparative analyses of immunoglobulin genes: surprises and portents. Nature Reviews Immunology, 2(9), pp.688-698.
12. Cooper, M.D., Peterson, R.D. and Good, R.A., 1965. Delineation of the thymic and bursal lymphoid systems in the chicken. Nature, 205(4967), pp.143-146.
13. Burnet, S.F.M., 1959. The clonal selection theory of acquired immunity (Vol. 3). Nashville: Vanderbilt University Press.
14. Cooper, M.D., 2015. The early history of B cells. Nature Reviews Immunology, 15(3), pp.191-197.

15. Murphy, K. and Weaver, C., 2016. Janeway's immunobiology. Garland Science.
16. Quesenberry, P. and Levitt, L., 1979. Hematopoietic stem cells. New England Journal of Medicine, 301(16), pp.868-872.
17. Kondo, M., Weissman, I.L. and Akashi, K., 1997. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell, 91(5), pp.661-672.
18. Kwon, K., Hutter, C., Sun, Q., Bilic, I., Cobaleda, C., Malin, S. and Busslinger, M., 2008. Instructive role of the transcription factor E2A in early B lymphopoiesis and germinal center B cell development. Immunity, 28(6), pp.751-762.
19. Janeway Jr, C.A., Travers, P., Walport, M. and Shlomchik, M.J., 2001. The rearrangement of antigen-receptor gene segments controls lymphocyte development.
20. Schatz, D.G., Oettinger, M.A. and Schlissel, M.S., 1992. V (D) J recombination: molecular biology and regulation. Annual review of immunology, 10(1), pp.359-383.
21. Ohta, Y. and Flajnik, M., 2006. IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates. Proceedings of the National Academy of Sciences, 103(28), pp.10723-10728.
22. Edholm, E.S., Bengten, E. and Wilson, M., 2011. Insights into the function of IgD. Developmental & Comparative Immunology, 35(12), pp.1309-1316.
23. Chen, K., Xu, W., Wilson, M., He, B., Miller, N.W., Bengten, E., Edholm, E.S., Santini, P.A., Rath, P., Chiu, A. and Cattalini, M., 2009. Immunoglobulin D enhances immune surveillance by activating antimicrobial, proinflammatory and B cell–stimulating programs in basophils. Nature immunology, 10(8), pp.889-898.
24. Ceredig, R. and Rolink, T., 2002. A positive look at double-negative thymocytes. Nature Reviews Immunology, 2(11), pp.888-897.
25. Nemazee, D.A., 1989. Clonal deletion of B lymphocytes in a transgenic mouse bearing anti-MHC. Nature, 337, p.562.
26. Nemazee, D. and Buerki, K., 1989. Clonal deletion of autoreactive B lymphocytes in bone marrow chimeras. Proceedings of the National Academy of Sciences, 86(20), pp.8039-8043.
27. Wardemann, H., Yurasov, S., Schaefer, A., Young, J.W., Meffre, E. and Nussenzweig, M.C., 2003. Predominant autoantibody production by early human B cell precursors. Science, 301(5638), pp.1374-1377.
28. Hayakawa, K., Asano, M., Shinton, S.A., Gui, M., Allman, D., Stewart, C.L., Silver, J. and Hardy, R.R., 1999. Positive selection of natural autoreactive B cells. Science, 285(5424), pp.113-116.

29. Shlomchik, M.J., 2008. Sites and stages of autoreactive B cell activation and regulation. Immunity, 28(1), pp.18-28.
30. Janeway, C.A., Travers, P., Walport, M. and Shlomchik, M.J., 1997. Immunobiology: the immune system in health and disease (Vol. 1, p. 11). Singapore: Current Biology.
31. Grimaldi, C.M., Hicks, R. and Diamond, B., 2005. B cell selection and susceptibility to autoimmunity. The Journal of Immunology, 174(4), pp.1775-1781.
32. Hudson, K.E., Hendrickson, J.E., Cadwell, C.M., Iwakoshi, N.N. and Zimring, J.C., 2012. Partial tolerance of autoreactive B and T cells to erythrocyte-specific self-antigens in mice. haematologica, 97(12), pp.1836-1844.
33. Tiegs, S.L., Russell, D.M. and Nemazee, D., 1993. Receptor editing in self-reactive bone marrow B cells. Journal of Experimental Medicine, 177(4), pp.1009-1020.
34. Gay, D., Saunders, T., Camper, S. and Weigert, M., 1993. Receptor editing: an approach by autoreactive B cells to escape tolerance. Journal of Experimental Medicine, 177(4), pp.999-1008.
35. Casellas, R., Shih, T.A.Y., Kleinewietfeld, M., Rakonjac, J., Nemazee, D., Rajewsky, K. and Nussenzweig, M.C., 2001. Contribution of receptor editing to the antibody repertoire. science, 291(5508), pp.1541-1544.
36. Nemazee, D., 2006. Receptor editing in lymphocyte development and central tolerance. Nature Reviews Immunology, 6(10), pp.728-740.
37. Retter, M.W. and Nemazee, D., 1998. Receptor editing occurs frequently during normal B cell development. The Journal of experimental medicine, 188(7), pp.1231-1238.
38. Merrell, K.T., Benschop, R.J., Gauld, S.B., Aviszus, K., Decote-Ricardo, D., Wysocki, L.J. and Cambier, J.C., 2006. Identification of anergic B cells within a wild-type repertoire. Immunity, 25(6), pp.953-962.
39. Cambier, J.C., Gauld, S.B., Merrell, K.T. and Vilen, B.J., 2007. B-cell anergy: from transgenic models to naturally occurring anergic B cells?. Nature Reviews Immunology, 7(8), pp.633-643.
40. Fang, W., Weintraub, B.C., Dunlap, B., Garside, P., Pape, K.A., Jenkins, M.K., Goodnow, C.C., Mueller, D.L. and Behrens, T.W., 1998. Self-reactive B lymphocytes overexpressing Bcl-x L escape negative selection and are tolerized by clonal anergy and receptor editing. Immunity, 9(1), pp.35-45.
41. Barr, T.A., Shen, P., Brown, S., Lampropoulou, V., Roch, T., Lawrie, S., Fan, B., O'Connor, R.A., Anderton, S.M., Bar-Or, A. and Fillatreau, S., 2012. B cell depletion

therapy ameliorates autoimmune disease through ablation of IL-6–producing B cells. Journal of Experimental Medicine, 209(5), pp.1001-1010.

42. Shlomchik, M.J., Madaio, M.P., Ni, D., Trounstein, M. and Huszar, D., 1994. The role of B cells in lpr/lpr-induced autoimmunity. Journal of Experimental Medicine, 180(4), pp.1295-1306.

43. Falcone, M., Lee, J., Patstone, G., Yeung, B. and Sarvetnick, N., 1998. B lymphocytes are crucial antigen-presenting cells in the pathogenic autoimmune response to GAD65 antigen in nonobese diabetic mice. The Journal of Immunology, 161(3), pp.1163-1168.

44. Germain, R.N., 2002. T-cell development and the CD4–CD8 lineage decision. Nature Reviews Immunology, 2(5), pp.309-322.

45. Kappler, J.W., Roehm, N. and Marrack, P., 1987. T cell tolerance by clonal elimination in the thymus. Cell, 49(2), pp.273-280.

46. Bhan, A.K., Reinherz, E.L., Poppema, S.I.B.R.A.N.D., McCluskey, R.T. and Schlossman, S.F., 1980. Location of T cell and major histocompatibility complex antigens in the human thymus. J Exp Med, 152(4), pp.771-782.

47. Murphy, K.M., Heimberger, A.B. and Loh, D.Y., 1990. Induction by Antigen of Intrathymic Apoptosis of CD4 (+) CD8 (+) TCR (lo) Thymocytes in Vivo. Science, 250(4988), p.1720.

48. Saizawa, K., Rojo, J. and Janeway, C.A., 1987. Evidence for a physical association of CD4 and the CD3: α: β T-cell receptor. Nature, 328(6127), pp.260-263.

49. Teh, H.S., Kisielow, P., Scott, B., Kishi, H., Uematsu, Y., Blüthmann, H. and von Boehmer, H., 1988. Thymic major histocompatibility complex antigens and the αβ T-cell receptor determine the CD4/CD8 phenotype of T cells.

50. Mackay, C.R., Beya, M.F. and Matzinger, P., 1989. γ/δ T cells express a unique surface molecule appearing late during thymic development. European journal of immunology, 19(8), pp.1477-1483.

51. Morita, C.T., Verma, S., Aparicio, P., Spits, H. and Brenner, M.B., 1991. Functionally distinct subsets of human γ/δ T cells. European journal of immunology, 21(12), pp.2999-3007.

52. Nishimura, H., Agata, Y., Kawasaki, A., Sato, M., Imamura, S., Minato, N., Yagita, H., Nakano, T. and Honjo, T., 1996. Developmentally regulated expression of the PD-1 protein on the surface of double-negative (CD4–CD8–) thymocytes. International immunology, 8(5), pp.773-780.

53. Shivakumar, S.U.M.A.T.I., Tsokos, G.C. and Datta, S.K., 1989. T cell receptor alpha/beta expressing double-negative (CD4-/CD8-) and CD4+ T helper cells in humans augment the production of pathogenic anti-DNA autoantibodies associated with lupus nephritis. The Journal of Immunology, 143(1), pp.103-112.
54. Robey, E. and Fowlkes, B.J., 1994. Selective events in T cell development. Annual review of immunology, 12(1), pp.675-705.
55. Shimizu, Y., Van Seventer, G.A., Siraganian, R., Wahl, L. and Shaw, S., 1989. Dual role of the CD44 molecule in T cell adhesion and activation. The Journal of Immunology, 143(8), pp.2457-2463.
56. Huet, S.T.E.P.H.A.N.E., Groux, H.E.R.V.E., Caillou, B.E.R.N.A.R.D., Valentin, H.E.L.E.N.E., Prieur, A.M. and Bernard, A.L.A.I.N., 1989. CD44 contributes to T cell activation. The Journal of Immunology, 143(3), pp.798-801.
57. Haynes, B.F., Telen, M.J., Hale, L.P. and Denning, S.M., 1989. CD44—a molecule involved in leukocyte adherence and T-cell activation. Immunology today, 10(12), pp.423-428.
58. Gray, H., 1918. "4c. The Thymus". Anatomy of the Human Body.
59. Starr, T.K., Jameson, S.C. and Hogquist, K.A., 2003. Positive and negative selection of T cells. Annual review of immunology, 21(1), pp.139-176.
60. Blackman, M., Kappler, J. and Marrack, P., 1990. The role of the T cell receptor in positive and negative selection of developing T cells. Science, 248(4961), pp.1335-1341.
61. Fowlkes, B.J. and Schweighoffer, E., 1995. Positive selection of T cells. Current opinion in immunology, 7(2), pp.188-195.
62. M Opata, M. and Stephens, R., 2013. Early decision: effector and effector memory T cell differentiation in chronic infection. Current immunology reviews, 9(3), pp.190-206.
63. Shortman, K., Egerton, M., Spangrude, G.J. and Scollay, R., 1990, January. The generation and fate of thymocytes. In Seminars in immunology (Vol. 2, No. 1, pp. 3-12).
64. Surh, C.D. and Sprent, J., 1994. T-cell apoptosis detected in situ during positive and negative selection in the thymus. Nature, 372(6501), p.100.
65. Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L. and Wiley, D.C., 1987. The foreign antigen binding site and T cell recognition regions. Nature, 329, pp.512-518.
66. Schroeder, H.W., 2006. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Developmental & Comparative Immunology, 30(1), pp.119-135.

67. Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J. and Kourilsky, P., 1999. A direct estimate of the human αβ T cell receptor diversity. Science, 286(5441), pp.958-961.
68. Rhoades, R. and Pflanzer, R.G., 1989. Human physiology (pp. 754-779). Saunders College Pub.
69. Abbas, A.K., Lichtman, A.H. and Pillai, S., 2014. Cellular and molecular immunology. Elsevier Health Sciences.
70. Greenberg, S., Silverstein, S.C. and Paul, W.E., 1993. Fundamental immunology. Fundamental Immunology, 509.
71. Klinman, N.R., Press, J.L., Sigal, N.H. and Gerhart, P.J., 1976. The acquisition of the B cell specificity repertoire: the germ-line theory of predetermined permutation of genetic information. The generation of antibody diversity, pp.127-150.
72. Ichihara, Y., Matsuoka, H. and Kurosawa, Y., 1988. Organization of human immunoglobulin heavy chain diversity gene loci. The EMBO journal, 7(13), p.4141.
73. Je W., 2006, Antibody structure, WikiMedia Commons. https://en.wikivet.net/Immunoglobulins_-_Overview
74. Schatz, D.G., 2004. V (d) j recombination. Immunological reviews, 200(1), pp.5-11.
75. Li, Z., Woo, C.J., Iglesias-Ussel, M.D., Ronai, D. and Scharff, M.D., 2004. The generation of antibody diversity through somatic hypermutation and class switch recombination. Genes & development, 18(1), pp.1-11.
76. Early, P., Huang, H., Davis, M., Calame, K. and Hood, L., 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell, 19(4), pp.981-992.
77. Das, S., Nikolaidis, N., Klein, J. and Nei, M., 2008. Evolutionary redefinition of immunoglobulin light chain isotypes in tetrapods using molecular markers. Proceedings of the National Academy of Sciences, 105(43), pp.16647-16652.
78. Lefranc, M.P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbié, V. and Chaume, D., 1999. IMGT, the international ImMunoGeneTics database. Nucleic acids research, 27(1), pp.209-212.
79. Ruiz, M., Pallarès, N., Contet, V.E.R., Barbié, V. and Lefranc, M.P., 1999. The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Experimental and clinical immunogenetics, 16(3), pp.173-184.
80. Bassing, C.H., Alt, F.W., Hughes, M.M., D'auteuil, M., Wehrly, T.D., Woodman, B.B., Gärtner, F., White, J.M., Davidson, L. and Sleckman, B.P., 2000. Recombination signal

sequences restrict chromosomal V (D) J recombination beyond the 12/23 rule. Nature, 405(6786), pp.583-586.

81. Cuomo, C.A., Mundy, C.L. and Oettinger, M.A., 1996. DNA sequence and structure requirements for cleavage of V (D) J recombination signal sequences. Molecular and Cellular Biology, 16(10), pp.5683-5690.

82. Kapitonov, V.V. and Jurka, J., 2005. RAG1 core and V (D) J recombination signal sequences were derived from Transib transposons. PLoS Biol, 3(6), p.e181.

83. Akamatsu, Y. and Oettinger, M.A., 1998. Distinct roles of RAG1 and RAG2 in binding the V (D) J recombination signal sequences. Molecular and Cellular Biology, 18(8), pp.4670-4678.

84. Krangel, M.S., 2003. Gene segment selection in V (D) J recombination: accessibility and beyond. Nature immunology, 4(7), pp.624-630.

85. Moshous, D., Pannetier, C., de Chasseval, R., le Deist, F., Cavazzana-Calvo, M., Romana, S., Macintyre, E., Canioni, D., Brousse, N., Fischer, A. and Casanova, J.L., 2003. Partial T and B lymphocyte immunodeficiency and predisposition to lymphoma in patients with hypomorphic mutations in Artemis. The Journal of clinical investigation, 111(3), pp.381-387.

86. McCaffrey, R., Harrison, T.A., Parkman, R. and Baltimore, D., 1975. Terminal deoxynucleotidyl transferase activity in human leukemic cells and in normal human thymocytes. New England Journal of Medicine, 292(15), pp.775-780.

87. Motea, E.A. and Berdis, A.J., 2010. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1804(5), pp.1151-1166.

88. Gustavocarra, 2008, Simplistic overview of V(D)J recombination. https://en.wikipedia.org/wiki/V(D)J_recombination#/media/File:VDJ_recombination.png

89. Neuberger, M.S. and Milstein, C., 1995. Somatic hypermutation. Current opinion in immunology, 7(2), pp.248-254.

90. French, D.L., Laskov, R. and Scharff, M.D., 1989. The role of somatic hypermutation in the generation of antibody diversity. Science, 244(4909), pp.1152-1158.

91. Hamilos, D.L., 1989. Antigen presenting cells. Immunologic research, 8(2), pp.98-117.

92. Jorgensen, J.L., Reay, P.A., Ehrich, E.W. and Davis, M.M., 1992. Molecular components of T-cell recognition. Annual review of immunology, 10(1), pp.835-873.

93. Danska, J.S., Livingstone, A.M., Paragas, V., Ishihara, T. and Fathman, C.G., 1990. The presumptive CDR3 regions of both T cell receptor alpha and beta chains determine T cell specificity for myoglobin peptides. Journal of Experimental Medicine, 172(1), pp.27-33.
94. Padmanabhan, R., Jay, E. and Wu, R., 1974. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4. Proceedings of the National Academy of Sciences, 71(6), pp.2510-2514.
95. Wu, R., 1972. Nucleotide sequence analysis of DNA. Nature, 236(68), pp.198-200.
96. Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences, 74(12), pp.5463-5467.
97. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. and Gocayne, J.D., 2001. The sequence of the human genome. science, 291(5507), pp.1304-1351.
98. Estevezj, 2012, The Sanger (chain-termination) method for DNA sequencing, Wikipedia, https://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg
99. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P., 1996. Real-time DNA sequencing using detection of pyrophosphate release. Analytical biochemistry, 242(1), pp.84-89.
100. Erlich, H.A. and Higuchi, R.G., Hoffman-La Roche Inc., 1994. Methods for nucleic acid amplification. U.S. Patent 5,314,809.
101. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. and Roth, R., 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature biotechnology, 18(6), pp.630-634.
102. Elahi, E. and Ronaghi, M., 2004. Pyrosequencing: a tool for DNA sequencing analysis. Bacterial Artificial Chromosomes: Volume 1 Library Construction, Physical Mapping, and Sequencing, pp.211-219.
103. Rusk, N., 2011. Torrents of sequence. Nature Methods, 8(1), pp.44-45.
104. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 309(5741), pp.1728-1732.

105. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology, 4(4), pp.265-270.
106. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. and Dahl, F., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science, 327(5961), pp.78-81.
107. Thompson, J.F. and Steinmann, K.E., 2010. Single molecule sequencing with a HeliScope genetic analysis system. Current protocols in molecular biology, pp.7-10.
108. National Human Genome Research Institute. 2014. The cost of sequencing a human genome. https://www.genome.gov/sequencingcosts/
109. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. and Boutell, J.M., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. nature, 456(7218), pp.53-59.
110. Voelkerding, K.V., Dames, S.A. and Durtschi, J.D., 2009. Next-generation sequencing: from basic research to diagnostics. Clinical chemistry, 55(4), pp.641-658.
111. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K. and Sidow, A., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome research, 18(7), pp.1051-1063.
112. Stuart, M.B., 2012, Sequencing-by-Synthesis: Explaining the Illumina Sequencing Technology
113. Schochetman, G., Ou, C.Y. and Jones, W.K., 1988. Polymerase chain reaction. The Journal of infectious diseases, 158(6), pp.1154-1157.
114. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. and Gocayne, J.D., 2001. The sequence of the human genome. science, 291(5507), pp.1304-1351.
115. Collins, F.S., Morgan, M. and Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. Science, 300(5617), pp.286-290.
116. Schneider, V. and Church, D., 2013. Genome reference consortium.
117. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. and Li, S., 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome research, 20(2), pp.265-272.

118. Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. and Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology, 30(7), pp.693-700.

119. Bashir, A., Klammer, A.A., Robins, W.P., Chin, C.S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P. and Sebra, R., 2012. A hybrid approach for the automated finishing of bacterial genomes. Nature biotechnology, 30(7), pp.701-707.

120. Cerdeira, L.T., Carneiro, A.R., Ramos, R.T.J., de Almeida, S.S., Schneider, M.P.C., Baumbach, J., Tauch, A., McCulloch, J.A., Azevedo, V.A.C. and Silva, A., 2011. Rapid hybrid de novo assembly of a microbial genome using only short reads: Corynebacterium pseudotuberculosis I19 as a case study. Journal of microbiological methods, 86(2), pp.218-223.

121. Deshpande, V., Fung, E.D., Pham, S. and Bafna, V., 2013, September. Cerulean: A hybrid assembly using high throughput short and long reads. In International Workshop on Algorithms in Bioinformatics (pp. 349-363).

122. Wood, E.J. 1995. The encyclopedia of molecular biology. Biochemical Education. 23 (2): 1165.

123. Chakravarti, A., 2001. Single nucleotide polymorphisms: to a future of genetic medicine. Nature, 409(6822), pp.822-823.

124. Treangen, T.J. and Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics, 13(1), pp.36-46.

125. Zarrei, M., MacDonald, J.R., Merico, D. and Scherer, S.W., 2015. A copy number variation map of the human genome. Nature Reviews Genetics, 16(3), pp.172-183.

126. Cruts, M., Theuns, J. and Van Broeckhoven, C., 2012. Locus-specific mutation databases for neurodegenerative brain diseases. Human mutation, 33(9), pp.1340-1344.

127. Touitou, I., Lesage, S., McDermott, M., Cuisset, L., Hoffman, H., Dode, C., Shoham, N., Aganna, E., Hugot, J.P., Wise, C. and Waterham, H., 2004. Infevers: an evolving mutation database for auto-inflammatory syndromes. Human mutation, 24(3), pp.194-198.

128. Krawczak, M. and Cooper, D.N., 1997. The human gene mutation database. Trends in Genetics, 13(3), pp.121-122.

129. Higasa, K., Miyake, N., Yoshimura, J., Okamura, K., Niihori, T., Saitsu, H., Doi, K., Shimizu, M., Nakabayashi, K., Aoki, Y. and Tsurusaki, Y., 2016. Human genetic variation database, a reference database of genetic variations in the Japanese population. Journal of human genetics.

130. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), pp.308-311.

131. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G. and Paschall, J., 2013. DbVar and DGVa: public archives for genomic structural variation. Nucleic acids research, 41(D1), pp.D936-D941.

132. Musumeci, L., Arthur, J.W., Cheung, F.S., Hoque, A., Lippman, S. and Reichardt, J.K., 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Human mutation, 31(1), p.67.

133. Mitchell, A.A., Zwick, M.E., Chakravarti, A. and Cutler, D.J., 2004. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics, 20(7), pp.1022-1032.

134. Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L. and Nickerson, D.A., 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nature genetics, 33(4), pp.518-521.

135. Cutler, D.J., Zwick, M.E., Carrasquillo, M.M., Yohn, C.T., Tobin, K.P., Kashuk, C., Mathews, D.J., Shah, N.A., Eichler, E.E., Warrington, J.A. and Chakravarti, A., 2001. High-throughput variation detection and genotyping using microarrays. Genome research, 11(11), pp.1913-1925.

136. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. and Liu-Cordero, S.N., 2002. The structure of haplotype blocks in the human genome. Science, 296(5576), pp.2225-2229.

137. Reich, D.E., Gabriel, S.B. and Altshuler, D., 2003. Quality and completeness of SNP databases. Nature genetics, 33(4), pp.457-458.

138. Musumeci, L., Arthur, J.W., Cheung, F.S., Hoque, A., Lippman, S. and Reichardt, J.K., 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Human mutation, 31(1), p.67.

139. Gudbjartsson, D.F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S.A., Zink, F., Oddson, A., Magnusson, G., Halldorsson, B.V., Hjartarson, E. and Sigurdsson, G.T.,

2015. Sequence variants from whole genome sequencing a large group of Icelanders. Scientific data, 2.

140. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. and Cancer Genome Atlas Research Network, 2013. The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), pp.1113-1120.

141. Genomics England, 2015, 100,000 Genomes Project. Genomicsengland.co.uk.

142. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y. and Tam, P.K.H., 2003. The international HapMap project. Nature, 426(6968), pp.789-796.

143. 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature, 467(7319), pp.1061-1073.

144. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. and McVean, G., 2011. The variant call format and VCFtools. Bioinformatics, 27(15), pp.2156-2158.

145. 1000 Genomes Project Consortium, A map of 1000 genome project in globe, http://www.internationalgenome.org/.

146. Heller, M.J., 2002. DNA microarray technology: devices, systems, and applications. Annual review of biomedical engineering, 4(1), pp.129-153.

147. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1), pp.262-267.

148. Allison, D.B., Cui, X., Page, G.P. and Sabripour, M., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature reviews genetics, 7(1), pp.55-65.

149. Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J. and Marra, M.A., 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques, 45(1), p.81.

150. Chu, Y. and Corey, D.R., 2012. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic acid therapeutics, 22(4), pp.271-274.

151. Wang, Z., Gerstein, M. and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), pp.57-63.

152. Metzker, M.L., 2010. Sequencing technologies—the next generation. Nature reviews genetics, 11(1), pp.31-46.
153. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S., 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol, 4(10), p.e309.
154. Levsky, J.M., Shenoy, S.M., Pezo, R.C. and Singer, R.H., 2002. Single-cell gene expression profiling. Science, 297(5582), pp.836-840.
155. Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A. and Xie, X.S., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. science, 329(5991), pp.533-538.
156. Raj, A., Van Den Bogaard, P., Rifkin, S.A., Van Oudenaarden, A. and Tyagi, S., 2008. Imaging individual mRNA molecules using multiple singly labeled probes. Nature methods, 5(10), p.877.
157. Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E. and Lin, X., 2013. RNA-sequencing from single nuclei. Proceedings of the National Academy of Sciences, 110(49), pp.19802-19807.
158. Hashimshony, T., Wagner, F., Sher, N. and Yanai, I., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell reports, 2(3), pp.666-673.
159. Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O. and Dor, Y., 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome biology, 17(1), p.77.
160. Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Sandberg, R. 2012. Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. Nature Biotechnology, 30(8), 777–782.
161. Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B.B., Siddiqui A., Lao K., Surani M.A., mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82.
162. Picelli S., Björklund Å.K., Faridani O.R., Sagasser S., Winberg G., Sandberg R., Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
163. Fulwyler, M.J., 1965. Electronic separation of biological cells by volume. Science, 150(3698), pp.910-911.
164. Stegle, O., Teichmann, S.A. and Marioni, J.C., 2015. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics, 16(3), pp.133-145.

165. Xuehua, Z., Novel next generation sequencing tool - RNA Sequencing, advantages, challenges and opportunities. University of Manitoba. http://home.cc.umanitoba.ca/~zhangx39/PLNT7690/presentation/presentation.html
166. Ewing, B., Hillier, L., Wendl, M.C. and Green, P., 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research, 8(3), pp.175-185.
167. Ewing, B., Hillier, L., Wendl, M.C. and Green, P., 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research, 8(3), pp.175-185.
168. Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.
169. Patel, R.K. and Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one, 7(2), p.e30619.
170. Gordon, A. and Hannon, G.J., 2010. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished), http://hannonlab. cshl. edu/fastx_toolkit.
171. Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, p.btu170.
172. Trapnell, C., Pachter, L. and Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), pp.1105-1111.
173. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), pp.15-21.
174. Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), pp.357-359.
175. Anders, S., Pyl, P.T. and Huber, W., 2014. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics, p.btu638.
176. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), p.13.
177. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. and Chen, Z., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29(7), pp.644-652.

178. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S. and Zhou, X., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics, 30(12), pp.1660-1666.

179. Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics, 28(8), pp.1086-1092.

180. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. and MacManes, M.D., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), pp.1494-1512.

181. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E. and Turner, S.W., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods, 10(6), pp.563-569.

182. Liao, Y., Smyth, G.K. and Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), pp.923-930.

183. Wagner, G.P., Kin, K. and Lynch, V.J., 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences, 131(4), pp.281-285.

184. Li, B. and Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 12(1), p.323.

185. Robinson, M.D., McCarthy, D.J. and Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), pp.139-140.

186. Anders, S. and Huber, W., 2012. Differential expression of RNA-Seq data at the gene level–the DESeq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL).

187. Hardcastle, T.J. and Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC bioinformatics, 11(1), p.422.

188. Tarazona, S., García, F., Ferrer, A., Dopazo, J. and Conesa, A., 2012. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnet. journal, 17(B), pp.pp-18.

189. Seyednasrollah, F., Laiho, A. and Elo, L.L., 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics, 16(1), pp.59-70.

190. Zhang, Z.H., Jhaveri, D.J., Marshall, V.M., Bauer, D.C., Edson, J., Narayanan, R.K., Robinson, G.J., Lundberg, A.E., Bartlett, P.F., Wray, N.R. and Zhao, Q.Y., 2014. A comparative study of techniques for differential expression analysis on RNA-Seq data. PloS one, 9(8), p.e103207.

191. Soneson, C. and Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC bioinformatics, 14(1), p.91.

192. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C. and Heisler, M.G., 2013. Accounting for technical noise in single-cell RNA-seq experiments. Nature methods, 10(11), pp.1093-1095.

193. Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R. and Oliver, B., 2011. Synthetic spike-in standards for RNA-seq experiments. Genome research, 21(9), pp.1543-1551.

194. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature methods, 11(2), pp.163-166.

195. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. and Xu, Y., 2015. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol, 11(11), p.e1004575.

196. McCarthy, D.J., Campbell, K.R., Lun, A.T. and Wills, Q.F., 2016. scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. bioRxiv, p.069633.

197. Ntranos, V., Kamath, G.M., Zhang, J.M., Pachter, L. and David, N.T., 2016. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome biology, 17(1), p.112.

198. Klein, U., Rajewsky, K. and Küppers, R., 1998. Human immunoglobulin (Ig) M+ IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. Journal of Experimental Medicine, 188(9), pp.1679-1689.

199. Küppers, R., Zhao, M., Hansmann, M.L. and Rajewsky, K., 1993. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. The EMBO Journal, 12(13), p.4955.

200. Ehlich, A., Martin, V., Müller, W. and Rajewsky, K., 1994. Analysis of the B-cell progenitor compartment at the level of single cells. Current Biology, 4(7), pp.573-583.
201. Benichou, J., Ben-Hamo, R., Louzoun, Y. and Efroni, S., 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology, 135(3), pp.183-191.
202. Becattini, S., Latorre, D., Mele, F., Foglierini, M., De Gregorio, C., Cassotta, A., Fernandez, B., Kelderman, S., Schumacher, T.N., Corti, D. and Lanzavecchia, A., 2015. Functional heterogeneity of human memory CD4+ T cell clones primed by pathogens or vaccines. Science, 347(6220), pp.400-406.
203. Ronowicz, E. and Coutinho, A., 1975. Functional analysis of B cell heterogeneity. Immunological Reviews, 24(1), pp.3-40.
204. Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H. and Quake, S.R., 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nature biotechnology, 32(2), pp.158-168.
205. Robins, H., 2013. Immunosequencing: applications of immune repertoire deep sequencing. Current opinion in immunology, 25(5), pp.646-652.
206. Henegariu, O., Heerema, N.A., Dlouhy, S.R., Vance, G.H. and Vogt, P.H., 1997. Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques, 23(3), pp.504-511.
207. Britanova, O.V., Putintseva, E.V., Shugay, M., Merzlyak, E.M., Turchaninova, M.A., Staroverov, D.B., Bolotin, D.A., Lukyanov, S., Bogdanova, E.A., Mamedov, I.Z. and Lebedev, Y.B., 2014. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. The Journal of Immunology, 192(6), pp.2689-2698.
208. Heather, J.M., Ismail, M., Oakes, T. and Chain, B., 2017. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. Briefings in Bioinformatics, p.bbw138.
209. Brochet, X., Lefranc, M.P. and Giudicelli, V., 2008. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized VJ and VDJ sequence analysis. Nucleic acids research, 36(suppl 2), pp.W503-W508.
210. Ye, J., Ma, N., Madden, T.L. and Ostell, J.M., 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic acids research, p.gkt382.

211. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. and Chain, B., 2013. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics, p.btt004.

212. Duez, M., Giraud, M., Herbert, R., Rocher, T., Salson, M. and Thonier, F., 2016. Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PloS one, 11(11), p.e0166126.

213. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V. and Chudakov, D.M., 2015. MiXCR: software for comprehensive adaptive immunity profiling. Nature methods, 12(5), pp.380-381.

214. Bolotin, D.A., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Turchaninova, M.A., Zvyagin, I.V., Britanova, O.V. and Chudakov, D.M., 2013. MiTCR: software for T-cell receptor sequencing data analysis. Nature methods, 10(9), pp.813-814.

215. Zhang, W., Du, Y., Su, Z., Wang, C., Zeng, X., Zhang, R., Hong, X., Nie, C., Wu, J., Cao, H. and Xu, X., 2015. IMonitor: a robust pipeline for TCR and BCR repertoire analysis. Genetics, 201(2), pp.459-472.

216. Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A.U., Babel, N., Reinert, K. and Robinson, P.N., 2015. IMSEQ-a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics, p.btv309.

217. Yu, Y., Ceredig, R. and Seoighe, C., 2015. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. Nucleic acids research, p.gkv1016.

218. Imkeller, K., Arndt, P.F., Wardemann, H. and Busse, C.E., 2016. sciReptor: analysis of single-cell level immunoglobulin repertoires. BMC bioinformatics, 17(1), p.67.

219. Yang, X., Liu, D., Lv, N., Zhao, F., Liu, F., Zou, J., Chen, Y., Xiao, X., Wu, J., Liu, P. and Gao, J., 2015. TCRklass: a new k-string–based algorithm for human and mouse TCR repertoire characterization. The Journal of Immunology, 194(1), pp.446-454.

220. Wang, Y., Jackson, K.J., Sewell, W.A. and Collins, A.M., 2008. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunology and cell biology, 86(2), pp.111-115.

221. Boyd, S.D., Gaëta, B.A., Jackson, K.J., Fire, A.Z., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D. and Simen, B.B., 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. The Journal of Immunology, 184(12), pp.6986-6992.

222. Yu, Y., Ceredig, R. and Seoighe, C., 2017. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. The Journal of Immunology, p.1601710.

223. Pannetier, C., Even, J. and Kourilsky, P., 1995. T-cell repertoire diversity and clonal expansions in normal and clinical samples. Immunology today, 16(4), pp.176-181.

224. Tumeh, P.C., Harview, C.L., Yearley, J.H., Shintaku, I.P., Taylor, E.J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V. and West, A.N., 2014. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature, 515(7528), pp.568-571.

225. Hill, M.O., 1973, Diversity and evenness: a unifying notation and its consequences. Ecology. 54: 427–432.

226. Clemente, M.J., Przychodzen, B., Jerez, A., Dienes, B.E., Afable, M.G., Husseinzadeh, H., Rajala, H.L., Wlodarski, M.W., Mustjoki, S. and Maciejewski, J.P., 2013. Deep sequencing of the T cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes. Blood, pp.blood-2013.

227. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O. and Ibrahim, M., 2009. The genetic structure and history of Africans and African Americans. science, 324(5930), pp.1035-1044.

228. Wang, C., Sanders, C.M., Yang, Q., Schroeder, H.W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R.M., Hudson, J.R., Davis, R.W. and Han, J., 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proceedings of the National Academy of Sciences, 107(4), pp.1518-1523.

229. Clemente, M.J., Przychodzen, B., Jerez, A., Dienes, B.E., Afable, M.G., Husseinzadeh, H., Rajala, H.L., Wlodarski, M.W., Mustjoki, S. and Maciejewski, J.P., 2013. Deep sequencing of the T cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes. Blood, pp.blood-2013.

230. Gaëta, B.A., Malming, H.R., Jackson, K.J., Bain, M.E., Wilson, P. and Collins, A.M., 2007. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics, 23(13), pp.1580-1587.

231. Lossos, I.S., Alizadeh, A.A., Eisen, M.B., Chan, W.C., Brown, P.O., Botstein, D., Staudt, L.M. and Levy, R., 2000. Ongoing immunoglobulin somatic mutation in

germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. Proceedings of the National Academy of Sciences, 97(18), pp.10209-10213.

232. Putintseva,E., Britanova,O., Staroverov,D., Merzlyak,E., Turchaninova,M., Shugay,M., Bolotin,D., Pogorelyy,M., Mamedov,I., Bobrynina,V., et al, 2013, Mother and Child T Cell Receptor Repertoires: Deep Profiling Study. Front Immunol, 4.

233. Doria-Rose, N.A., Schramm, C.A., Gorman, J., Moore, P.L., Bhiman, J.N., DeKosky, B.J., Ernandes, M.J., Georgiev, I.S., Kim, H.J., Pancera, M. and Staupe, R.P., 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature, 509(7498), pp.55-62.

234. Barak, M., Zuckerman, N.S., Edelman, H., Unger, R. and Mehr, R., 2008. IgTree©: Creating Immunoglobulin variable region gene lineage trees. Journal of immunological methods, 338(1), pp.67-74.

235. Andrew, R., et al. 2007, FigTree [Computer software]. Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

236. Scheepers, C., Shrestha, R.K., Lambson, B.E., Jackson, K.J., Wright, I.A., Naicker, D., Goosen, M., Berrie, L., Ismail, A., Garrett, N. and Karim, Q.A., 2015. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. The Journal of Immunology, 194(9), pp.4371-4378.

237. Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D. and Nadeau, K.C., 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Science translational medicine, 1(12), pp.12ra23-12ra23.

238. Parameswaran, P., Liu, Y., Roskin, K.M., Jackson, K.K., Dixit, V.P., Lee, J.Y., Artiles, K.L., Zompi, S., Vargas, M.J., Simen, B.B. and Hanczaruk, B., 2013. Convergent antibody signatures in human dengue. Cell host & microbe, 13(6), pp.691-700.

239. Jackson, K.J., Liu, Y., Roskin, K.M., Glanville, J., Hoh, R.A., Seo, K., Marshall, E.L., Gurley, T.C., Moody, M.A., Haynes, B.F. and Walter, E.B., 2014. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell host & microbe, 16(1), pp.105-114.

240. Gibson, K.L., Wu, Y.C., Barnett, Y., Duggan, O., Vaughan, R., Kondeatis, E., Nilsson, B.O., Wikby, A., Kipling, D. and Dunn-Walters, D.K., 2009. B-cell diversity decreases in old age and is correlated with poor health status. Aging cell, 8(1), pp.18-25.

241. Boyd, S.D., Liu, Y., Wang, C., Martin, V. and Dunn-Walters, D.K., 2013. Human lymphocyte repertoires in ageing. Current opinion in immunology, 25(4), pp.511-515.

242. Dunn-Walters, D.K. and Ademokun, A.A., 2010. B cell repertoire and ageing. Current opinion in immunology, 22(4), pp.514-520.

243. Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N.S., Louder, M., McKee, K. and O'Dell, S., 2011. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science, 333(6049), pp.1593-1602.

244. Wang, C., Liu, Y., Xu, L.T., Jackson, K.J., Roskin, K.M., Pham, T.D., Laserson, J., Marshall, E.L., Seo, K., Lee, J.Y. and Furman, D., 2014. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. The Journal of Immunology, 192(2), pp.603-611.

245. Wu, X., Zhang, Z., Schramm, C.A., Joyce, M.G., Do Kwon, Y., Zhou, T., Sheng, Z., Zhang, B., O'Dell, S., McKee, K. and Georgiev, I.S., 2015. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. Cell, 161(3), pp.470-485.

246. Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I.R. and Friedman, N., 2014. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome research, 24(10), pp.1603-1612.

247. Tipton, C.M., Fucile, C.F., Darce, J., Chida, A., Ichikawa, T., Gregoretti, I., Schieferl, S., Hom, J., Jenks, S., Feldman, R.J. and Mehr, R., 2015. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. Nature immunology, 16(7), pp.755-765.

248. Hehle, V., Fraser, L.D., Tahir, R., Kipling, D., Wu, Y.C., Lutalo, P.M., Cason, J., Choong, L., D'Cruz, D.P., Cope, A.P. and Dunn-Walters, D.K., 2015. Immunoglobulin kappa variable region gene selection during early human B cell development in health and systemic lupus erythematosus. Molecular immunology, 65(2), pp.215-223.

249. Muraro, P.A., Robins, H., Malhotra, S., Howell, M., Phippard, D., Desmarais, C., Sousa, A.D.P.A., Griffith, L.M., Lim, N., Nash, R.A. and Turka, L.A., 2014. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. The Journal of clinical investigation, 124(3), pp.1168-1172.

250. Pallarès, N., S. Lefebvre, V. E. R. Contet, F. Matsuda, and M. P. Lefranc. 1999. The human immunoglobulin heavy variable genes. Experimental and clinical immunogenetics, 16(1): 36-60.

251. Folch, G. E. R. and M. P. Lefranc. 2000. The human T cell receptor beta variable (TRBV) genes. Experimental and clinical immunogenetics, 17(1): 42-54.
252. Zheng, X., 2012. SNPRelate: parrallel computing toolset for genome-wide association studies. R package version, 95.
253. Liu, L. and Lucas, A.H., 2003. IGH V3-23* 01 and its allele V3-23* 03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b. Immunogenetics, 55(5), pp.336-338.
254. Watson, C.T. and Breden, F., 2012. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes and immunity, 13(5), pp.363-373.
255. Watson, C.T., Steinberg, K.M., Huddleston, J., Warren, R.L., Malig, M., Schein, J., Willsey, A.J., Joy, J.B., Scott, J.K., Graves, T.A. and Wilson, R.K., 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. The American Journal of Human Genetics, 92(4), pp.530-546.
256. Romo-González, T. and Vargas-Madrazo, E., 2005. Structural analysis of substitution patterns in alleles of human immunoglobulin VH genes. Molecular immunology, 42(9), pp.1085-1097.
257. Romo-González, T. and Vargas-Madrazo, E., 2006. Substitution patterns in alleles of immunoglobulin V genes in humans and mice. Molecular immunology, 43(6), pp.731-744.
258. Romo-González, T., Morales-Montor, J., Rodríguez-Dorantes, M. and Vargas-Madrazo, E., 2005. Novel substitution polymorphisms of human immunoglobulin VH genes in Mexicans. Human immunology, 66(6), pp.731-739.
259. Sharon, E., Sibener, L.V., Battle, A., Fraser, H.B., Garcia, K.C. and Pritchard, J.K., 2016. Genetic variation in MHC proteins is associated with T cell receptor expression biases. Nature genetics.
260. Pace, T.W., Mletzko, T.C., Alagbe, O., Musselman, D.L., Nemeroff, C.B., Miller, A.H. and Heim, C.M., 2006. Increased stress-induced inflammatory responses in male patients with major depression and increased early life stress. American Journal of Psychiatry, 163(9), pp.1630-1633.
261. Bierhaus, A., Wolf, J., Andrassy, M., Rohleder, N., Humpert, P.M., Petrov, D., Ferstl, R., von Eynatten, M., Wendt, T., Rudofsky, G. and Joswig, M., 2003. A mechanism

converting psychosocial stress into mononuclear cell activation. Proceedings of the National Academy of Sciences, 100(4), pp.1920-1925.

262. Maes, M., 1995. Evidence for an immune response in major depression: a review and hypothesis. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 19(1), pp.11-38.

263. Miller, A.H. and Raison, C.L., 2016. The role of inflammation in depression: from evolutionary imperative to modern treatment target. Nature Reviews Immunology, 16(1), pp.22-34.

264. Pasqual, N., Gallagher, M., Aude-Garcia, C., Loiodice, M., Thuderoz, F., Demongeot, J., Ceredig, R., Marche, P.N. and Jouvin-Marche, E., 2002. Quantitative and Qualitative Changes in VJ α Rearrangements During Mouse Thymocytes Differentiation Implication For a Limited T Cell Receptor α Chain Repertoire. The Journal of experimental medicine, 196(9), pp.1163-1174.

265. Lim, A., Trautmann, L., Peyrat, M.A., Couedel, C., Davodeau, F., Romagné, F., Kourilsky, P. and Bonneville, M., 2000. Frequent contribution of T cell clonotypes with public TCR features to the chronic response against a dominant EBV-derived epitope: application to direct detection of their molecular imprint on the human peripheral T cell repertoire. The Journal of Immunology, 165(4), pp.2001-2011.

266. Bowerman, N.A., Falta, M.T., Mack, D.G., Wehrmann, F., Crawford, F., Mroz, M.M., Maier, L.A., Kappler, J.W. and Fontenot, A.P., 2014. Identification of multiple public TCR repertoires in chronic beryllium disease. The Journal of Immunology, 192(10), pp.4571-4580.

267. Brodlie, K., Wood, J., Duce, D. and Sagar, M., 2004, September. gViz: visualization and computational steering on the Grid. In Proceedings of the UK e-Science All Hands Meeting (pp. 54-60).

268. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R. and Urban, A.E., 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome research, 24(1), pp.14-24.

269. Scott, M.G., Crimmins, D.L., McCourt, D.W., Zocher, I., Thiebe, R., Zachau, H.G. and Nahm, M.H., 1989. Clonal characterization of the human IgG antibody repertoire to Haemophilus influenzae type b polysaccharide. III. A single VKII gene and one of several JK genes are joined by an invariant arginine to form the most common L chain V region. The Journal of Immunology, 143(12), pp.4110-4116.

270. Wiehe K., Easterhoff D., Luo K., Nicely N.I., Bradley T., Jaeger F.H., et al. 2014. Antibody light-chain-restricted recognition of the site of immune pressure in the RV144 HIV-1 vaccine trial is phylogenetically conserved. Immunity. 41(6): 909–18.

271. Costa A.I., Koning D., Ladell K., McLaren J.E., Grady B.P., Schellens I.M., et al. 2015. Complex T-cell receptor repertoire dynamics underlie the CD8+ T-cell response to HIV-1. J Virol. 89(1):110–9.

272. Zhu J., Peng T., Johnston C., Phasouk K., Kask A.S., Klock A., et al. 2013. Immune surveillance by CD8 alpha alpha+ skin-resident T cells in human herpes virus infection. Nature 497(7450):494–7.

273. Hoogenboom H.R., 2005. Selecting and screening recombinant antibody libraries. Nat Biotechnol. 23(9):1105–16.

274. Hornbeck, P.V., 1991. Enzyme-linked immunosorbent assays. Current protocols in immunology, pp.2-1.

275. Galson J.D., Pollard A.J., Truck J., Kelly D.F., 2014. Studying the antibody repertoire after vaccination: practical applications. Trends in Immunology. 35(7):319–31.

276. Barreiro, L.B., and Quintana-Murci, L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. 11, 17–30.

277. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. 2009. Evolutionary dynamicsof human Toll-like receptors and their different contributions to hostdefense. PLoS Genet. 5, e1000562.