creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학박사 학위논문

신장 이식 환자에서 Next Generation Sequencing 기반 B Cell Repertoire 분석

Next Generation Sequencing Based Analysis of B Cell Repertoire in Kidney Transplantation

울산대학교대학원

의학과

양정석

Doctor of Philosophy

Next Generation Sequencing Based Analysis of B Cell Repertoire in Kidney Transplantation

The Graduate School
of the University of Ulsan
Department of Laboratory Medicine
John Jeongseok Yang

신장 이식 환자에서 Next Generation Sequencing 기반 B Cell Repertoire 분석

지도교수 황상현

이 논문을 의학박사 학위 논문으로 제출함

2022 년 08 월

울산대학교대학원

의학과

양정석

Next Generation Sequencing Based Analysis of B Cell Repertoire in Kidney Transplantation

Supervisor: Sang-Hyun Hwang

A dissertation

Submitted to

the Graduate School of the University of Ulsan

In partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

by

John Jeongseok Yang

Department of Laboratory Medicine

Ulsan, Korea

August 2022

양정석의 의학박사학위 논문을 인준함

심사위원장	신성	인
심사위원	오흥범	인
심사위원	황상현	인
심사위원	고대현	인
심사위원	고선영	인

울산대학교대학원

2022 년 08 월

Next Generation Sequencing Based Analysis of B Cell Repertoire in Kidney Transplantation

This certifies that the dissertation
of John Jeongseok Yang is approved.

Sung Shin

____
Committee Chair Dr.

Heung-Bum Oh

____
Committee Member Dr.

Sang-Hyun Hwang

____
Committee Member Dr.

Dae-Hyun Ko

____
Committee Member Dr.

Sun-Young Ko

____
Committee Member Dr.

Department of Laboratory Medicine

Ulsan, Korea

August 2022

국문요약

배경: 면역체계는 다양한 감염원 및 외부 물질로부터 신체를 보호하는 기능을 담당한다. 면역체계를 구성하는 다양한 종류의 면역세포, 신호전달체계는 체계와 균형을 이루고 있다. 면역 체계는 선천면역계와 적응면역계로 구성되어 있다. 적응면역계는 수많은 종류의 병원체, 항원에 반응하는 능력을 특징으로 하는데 이는 서로 다른 종류의 수용체를 통해 항원을 인식할 수 있고, 수없이 많은 종류의 항체를 통해 작용을 나타낼 수 있기 때문이다. 다양한 종류의 항체를 생성할 수 있는 기전을 repertoire 라고 하며 이를 adaptive immune receptor repertoire (AIRR)이라고 일컫는다. T세포수용체와 면역글로불린 항체의 종류가 수많은 항원을 인식할 수 있는 것은 다양한 클론의 면역세포가 존재하기 때문인데 이들이 생성하는 항체의 다양성을 확보하는 기전으로 V(D)J recombination, somatic hypermutation, class switching과 같은 중요한 기전들이 존재한다. 유전체 내에 넓게 분포하고 있는 V, D, J 유전자가 B cell의 성숙과정에서 재배열됨으로 인해 천문학적으로 많은 수의 다른 항원에 반응할 수 있는 다양한 종류의 항체를 생성할 수 있는 능력을 지니게 된다.

이식 면역은 장기이식 분야에서 면역 체계의 이해와 발전을 통해 이식 성적의 큰 향상을 이루어 낸 바 있다. 특히 신장이식의 성적은 효과적인 면역억제제의 개발로 이식성적이 크게 향상되었으나 장기적인 측면에서 이식 장기의 생존으로 평가하는 이식 성적의 향상은 2000년대 이후 제한적이었다. 현재 장기생존의 가장 큰 저해요인은 항체매개 거부 반응이며 이를 예측하거나 빠르게 진단할 수 있는 진단적 지표의 개발이 절실한 상황이다. AIRR 분석의 신장이식 분야에 대한 적용은 기존 연구에 의해 시도된 바 있지만 그 수가 매우 제한적이며 주로 T세포수용체에 국한되어 있다. 따라서 본 연구는 신장이식 환자의 B cell repertoire 분석을 위해 AIRR에 대한 차세대 염기서열분석을 적용해 보았다. 특히 AIRR 분석에 사용되는 다양한 종류의 도구와 기법 그리고 분석과정의 표준화에 대한 평가를 시행하여 장점과 단점 그리고 제한점 등을 검토하여 보편적 사용에 적합한 pipeline의 구축 가능성을 평가하고자 하였다.

방법: 서울아산병원에서 1996년 12월과 2021년 3월 사이의 기간동안 신장이식을 시행한 환자를 대상으로 하였다. ABO 적합 신장이식으로 한정하여 ABO 항체의 영향을 배제하고자 하였으며 이식 전 후 조직적합성 평가를 위해 시행된 검사 결과가 모두 있는 경우 연구에 포함하였다. 이들 가운데 2018년 1월로부터 2021년 7월 사이에 신장생검을 포함한 이식면역 관련 검사 및 평가가 진행된 환자들의 검체를

사용하였으며 총 15명의 보관 검체가 사용되었다. Banff 2017 Revised 진단기준에 따라 항체매개 거부반응 환자군(ABMR)과 거부반응이 없는 환자군(NR)으로 나누어 분석을 시행하였고 두 군 사이의 AIRR의 특징적인 차이를 살펴보고자 하였다. 또한 AIRR 분석의 재현성과 read depth에 따른 clonotype의 차이를 살펴보고자 NCBI SRA로부터 보다 많은 수의 정상 대조군의 염기서열을 확보해 비교 분석을 진행하였다. 차세대 염기서열 분석은 BIOMED-2 프로토콜을 기반으로 개발 및 상용화 된 LymphoTrack IGH FR1 assay kit - MiSeq을 사용하였으며 이로부터 생성된 염기서열 데이터를 활용하였다. AIRR 분석은 preprocessing 과정의 중요성과 이로부터 나타나는 차이점을 살펴보고자 MiXCR, VDJPipe 그리고 LymphoTrack 세가지 pipeline을 비교하였고 AIRR 분석에는 공통적으로 clonality, diversity, CDR3 analysis 그리고 gene usage 항목에 대한 분석을 시행하였다.

결과: ABMR 군과 NR 군 사이에는 AIRR 분석결과에서 차이가 있음을 확인하였다. AIRR 분석 항목 가운데 clone 수와 clonotype의 수의 경우 개별 검체마다 큰 차이가 존재하였으며 이를 구분하기 위한 clonotype 5,000과 같은 특정 cutoff를 적용할 경우 통계적으로 매우 유의한 차이를 나타냈다. Diversity, clonality 항목에서도 검체 간, 군 간 차이를 보였지만 제한적인 검체 수, 특히 거부반응이 없는 환자군 수의 적음으로 인해 통계적 유의성은 관찰되지 않았다. Diversity의 경우 항체매개 거부반응군에서 분석에 사용된 Chao1, Hill number, D50 등 모든 지표가 낮게 관찰되었고 일치하는 경향성을 살펴보았다. 세가지 종류의 preprocessing pipeline을 비교하였을 때 MiXCR, LymphoTrack의 결과는 유사하였으며 clonotyping의 결과를 제외한 나머지 AIRR 분석의 경우 세가지 모두 일치하는 분석 결과 및 경향을 나타냈다. SRA의 정상 대조군 염기서열 분석을 통해 살펴본 결과 read depth가 높을수록 검출되는 clone의 수와 clonotype의 숫자가 증가하였다. 또한 검사 및 분석의 재현성을 살펴본 결과 read depth의 차이에 따른 clone 수가 다르게 검출되더라도 최종적인 clonotype의 수는 일정하게 검출됨을 확인할 수 있었다.

결론: 신장 이식 이후 환자군의 검체를 사용한 AIRR 분석 결과 항체 매개 거부 반응의 발현 유무에 따른 AIRR의 차이를 확인하였다. 검사의 다양성을 고려한 표준화 및 재현성의 측면에서 상용화된 LymphoTrack IGH assay kit는 사용 목적에 부합하였다. 본 연구에서 AIRR 분석에 사용된 LymphoTrack을 사용한 pipeline은 AIRR 데이터의 preprocessing, repertoire 분석을 효과적으로 시행할 수 있어 이후 보다 많은 수의 검체를 활용한 연구와 다양한 영역에 적용할 수 있을 것으로 기대된다.

중심 단어: B cell, Adaptive immunity, repertoire, sequencing, high-throughput, kidney transplantation, antibody-mediated, T cell-mediated, rejection

목차

표 목차 Table legends

그림 목차 Figure legends

부록 목차 Supplementary Table Legends

서론 Introduction

The human immune system is composed of two fundamental mechanisms, the innate immunity and the adaptive immunity.[1,2] The innate immunity provides defense against invading pathogens within the first few hours after contact until the adaptive immune response develops. The adaptive immunity (or acquired immunity) recognizes and process antigens. The response is mediated by lymphocytes and their products, which are antibodies in secreted form or receptors that recognize the antigens. A cardinal feature of the adaptive immune system is to be able to recognize and respond to specific and diverse antigens, to generate secondary immune responses when exposed to a previously challenged antigen yet maintaining the non-reactivity to self (self-tolerance).

Lymphocytes of the B cell lineage are the principal cells of the adaptive immune system. As the primary source of antibody production, diversity of the B cells is critical for the immune system to recognize and respond to any given number of different antigens. In specific, antibodies secreted from B lymphocytes and consequent activation of the complement system are the essential components of the humoral immune system. Secreted forms of immunoglobulin are mainly secreted by plasma cells and their consequent binding actions include neutralization, opsonization, complement fixation etc. The membrane-bound forms of immunoglobulin act as a receptor for antigens on B cell surfaces (B cell receptor, BCR).

The number of antigens that the diverse repertoire of antibodies can recognize is potentially unlimited. In theory, the immune system is capable of producing an antibody response to any non-self-antigen, with the estimated diversity of at least 10^{12} unique antibodies (Figure 1).[3,4] There are also estimates on the size of human antibody repertoire suggested up to ~10^{15} and as high as 10^{18} using different theoretical calculations.[5–7] However the B lymphocytes, the principal cell of adaptive immunity as the sources of antibody production are outnumbered by this vast diversity. The human antibody repertoire outnumbers the estimated number of total mature ($CD27^-/IgD^+$) B cells in the body (~10^9) and even the number of total cell in the human body.[8,9] The combination of innate immune system and the adequacy of antigen recognition made by somewhat limited diversity is apparently functional, but the paradox of mismatch between the diversity of the immune repertoire and the physical limitation of number of B cells is yet fully understood. The complexity of immune repertoire originates from the recombination of different gene segments of the V, D and J segments and understanding of the resulting antibody gene structure is crucial for further downstream understanding of the immune system.

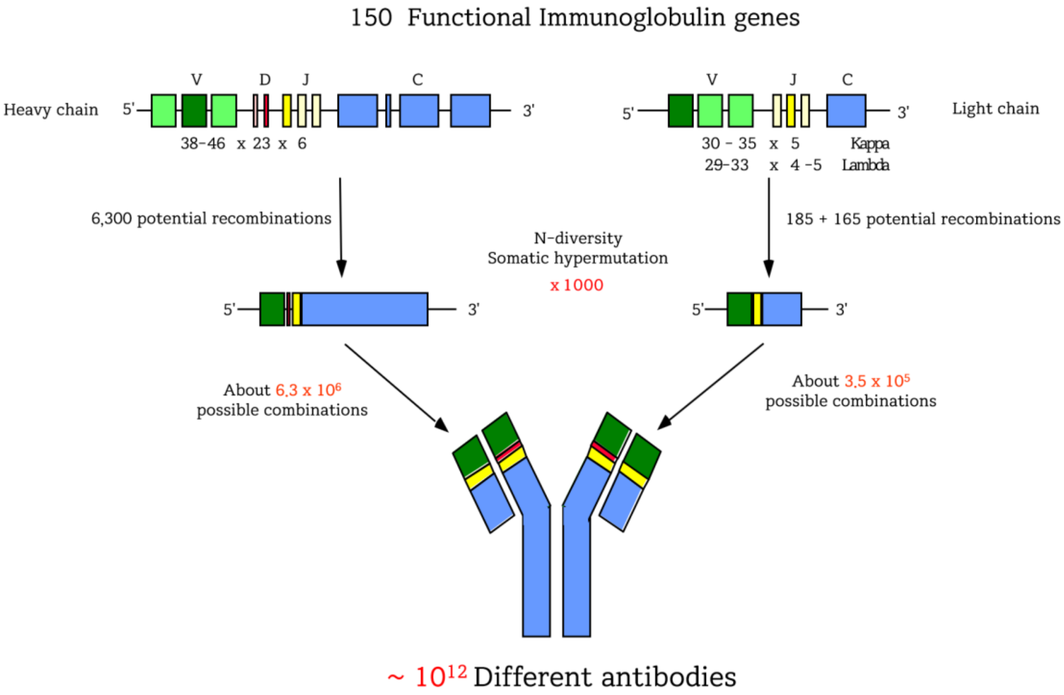


Figure 1. Estimated antibody diversity of the adaptive immune system.

The structure of antibody is important for understanding of their function. Since the discovery of molecule responsible for recognition of pathogenic molecules by Paul Ehrlich in 1891, its structure, function and mechanism of action has been the single most important subject in the field of immunology.[10] Antibodies are generally circulating proteins, composed of two identical heavy chains and two identical light chains. It is a series of homologous chains about 110 amino acids in length, folded in globular shape that constructs the Ig domain. Both chains have amino-terminal variable (V) regions which functions for antigen recognition and a carboxy-terminal constant (C) region that harbor mediative roles for effector functions. Genes encoding a complete immunoglobulin chains do not exist within the DNA of most cells and are widely separated in germlines cells and most somatic cells, requiring assembly and union of separate gene segments.[11] An exception is the B cells in which the genes are rearranged to create a mature immunoglobulin gene. Heavy chain from chromosome loci 14q32, kappa light chain from 2p12 and lambda light chain from 22q12 are rearranged together, from which various number of each gene is combined during rearrangement. The rather imprecise joining of gene segments induces different amino acid sequences, leading to different functional proteins. In brief, the rearrangement process begins with one of the IGHD genes joining with the IGHJ gene, creating a partially rearranged D-J gene. Additional joining of IGHV gene generates a completely rearranged IGHV-D-J gene. Only a completely rearranged immunoglobulin gene can produce a mature immunoglobulin protein, whereas immunoglobulin genes found in other chromosomal loci called orphons are non-functional. Each VH and VL domain encoded by germline V, D and J segments require recombination to form a functional (productive) V gene.[2] Therefore, antibodies are synthesized only by the B lymphocyte lineages and exist in two different forms, membrane-bound on surface of B lymphocytes as antigen receptors and secreted antibodies. In addition to V(D)J rearrangements, random indels of the germline nucleotide sequences and importantly, somatic hypermutation occurs. Through the process of somatic hypermutation, the diversity of B-cell repertoire is further increased to $>10^{14}$. Somatic hypermutation occurs in activated B cells, and the rate of mutation is about 100-folds higher than normal cells owing to the proliferation of B cells.[12]

A key region in the immunoglobulin molecule is the hypervariable region composed of 10 amino acid residues. There are three protruding loops connecting the adjacent β sheets of the V domain. These residues are in complementarity to the three-dimensional shape of the binding antigen, therefore called the 'complementarity-determining regions (CDRs)'. The CDRs are numbered in order, among which the CDR3 is the most variable of CDRs. While all three CDRs are important in antigen recognition, CDR3 is the target of many repertoire studies regarding its highest variability. But is is important to notice the other CDRs (CDR1 and CDR2) of their contribution to antigen specificity.[13] The antigenic specificity of antibody that distinguishes the small differences in chemical structure originates from the structural uniqueness. Structurally related antigen is capable of causing a 'cross-reaction' but the ability of antibodies to specifically bind to a large number of different antigens defines the diversity and the collection of such antibodies defines the adaptive immuune receptor repertoire (AIRR).[14] The definition of AIRR made by the AIRR Community is the collection of BCRs and/or T cell receptors (TCR) in a population of lymphocytes.[15] As each BCRs and/or TCRs have unique combination of gene segments and CDR3 sequences, such sequences are used to define B- and T cell clones. The classic definition of clone is the whole nucleotide of sequences of V, D and J segments.[16] It can also be defined as the receptors with same V and J usage and the amino acid sequences of V(D)J junction sequences.[17] A more recent definition comes down to heavy-light or α/β or γ/δ pairings, but from a sequence-wise perspective, the widely accepted definition of a clone is usually limited to the CDR3 region.[18,19] The reason why many studies focus on the CDR3 has been previously decribed.[20,21] The majority of sequence variation that affects three-dimensional structure occurs in the H3 loop, which is the third CDR of the heavy chain. An antibody clonotype is defined as the collection of sequences using the identical V and J genes to encode the CDR3 amino acid sequences.[22] As such, clonotyping is the process of identifying the unique nucleotide sequences of the CDR3, which has been limited by the bulkiness of the immunoglobulin sequence diversity.[18]

Following the study generating the initial sequences of immunoglobulin gene by Matsuda et al.[23] the human genome sequencing studies have been limited in delineating the immunoglobulin loci, due to its characteristic rearrangements and somatic hypermutations, not to mention the inadequate read-depth to cover the vast diversity of the AIRR diversity.[24] Application of next-generation sequencing (NGS) and the remarkable advancement in high-throughput sequencing (HTS) technology has allowed access to the analysis of immune repertoire and has allowed more in-depth understanding the diversity of the variable regions. Before HTS technology, analysis of the human antibody repertoire has been limited due to its unparalleled size. The AIRR sequencing (AIRR-seq) has become a trending topic in the field of immunoinformatics, which combines both immunology, medicine, bioinformatics, mathematics and computer science as an interdisciplinary science.[25–27] Immunoinformatics has become a rising field of immunology research that comprises T cell therapy, vaccine development, proteomics and computer science. The importance of immunoinformatics was proven invaluable for the vaccine development, especially during the pandemic of COVID-19.[28] Antibody is currently the most prominent class of biotherapeutics with the continuously growing importance.[29–31]

The comprehensive understanding of the adaptive immune system using AIRR-seq is still limited by several factors. There is a variety of workflows for approaching the AIRR, for which efforts to standardize the work flow is continuously made.[15,32–34] Although access to HTS has been facilitated through lowered cost, source of DNA or RNA, process of specimen preparation, library generation, and sequencing runs are variable sources of different outcomes. Even more, the post-sequencing processes which include annotation, clonotyping and repertoire analysis has to be carefully selected according to the purpose of study, scope of interest and its application.[32,35–38] The latter post-sequencing part of the AIRR-seq heavily relies on the bioinformatics technology and computational science, and these fast evolving science has introduced several popular novel tools for AIRR-seq.[29,39] First and of the foremost importance is a proper reference sequences to provide comparison of analytical data. The ImMunoGeneTics (IMGT) information database (http://www.imgt.org) is the largest database of immunoglobulin reference sequences, that also provides useful tools for analysis and visualization. IMGT HIGH/V-Quest has been developed for HTS repertoire data and IMGT clonotype analysis.[40] IgBlast is another popular tool for analysis of immunoglobulin sequences developed from the National Institute for Biotechnology Information (NCBI).[41] Both sequence annotations are optimized for multiple searches for a single immunoglobulin sequence, as does many more tools including iHMMune-aligner, JoinSolver, ImmPort, GWASdb etc.[29,41–44] Prerequisites of background knowledge of medicine and biology and understanding of the clinical need are necessary for the discipline of immunoinformatics. Basis of immunoinformatics also relies on understanding of the computational approaches and its mathematical algorithms. Validation of novel tools and models generated through immunoinformatics is important, and only limited number of studies have been conducted in such scope of analyses.[20,45–48] Selection and establishment of the AIRR-seq workflow requires a considerable efforts to take into consideration, the problems originating from sample source, PCR amplification and sequencing errors and the subsequent repertoire analyses.[49]

For comparison of the clonotypes identified by sequencing, the human immunoglobulin sequences have been defined by a unique numbering, established by the IMGT database. The numbering combines the framework region, CDRs, structural data and the hypervariable loop characteristics, through which the IMGT-ontology classification uses 'locus', 'group', 'subgroup', 'gene' and 'alleles' for naming and classification of the immunoglobulin genes.[50] The definition of a clonotype combines a unique V(D)J rearrangement of the IMGT genes and alleles annotated at the nucleotide level, a conserved anchor sequences and finally a unique CDR3 of the in frame juction. For comparison of repertoires such as the gene usage between samples, annotation of the immunoglobulin genes are commonly performed at the gene levels (e.g. IGHV3-23).

Immune system comprises of the cells, tissues, and various compartments of the body. Lymphocytes develop from the primary lymphoid tissue (bone marrow and thymus), and further circulate or migrate to secondary lymphoid tissues (spleen, lymph nodes, etc.) It is notable that the antigenic stimulus is concentrated in the secondary lymphoid tissues, where antigen is presented to naïve and memory B cells.[1,2] Access to the lymphoid tissues and the lymphatic systems are invasive (i.e. biopsy) and sometimes inaccessible, therefore B cells circulating in peripheral blood is the most favored specimen for analyses of adaptive immune system. The first raw materials obtained from circulating B cells are either genomic DNA (gDNA) or messenger RNA (mRNA), which serves as the templates for library amplification.[51] Choice of the raw material is the first factor to consider when conducting AIRR-seq studies. gDNA is available proportionally to the number of cells while mRNA is associated to the activation status and the function of cells. gDNA has the advantage of easy access and stability, but studies regarding the gene transcription level requires mRNA.[48] After determining the source of genetic material, the next important step is the method of amplification for library preparation. There are several available methods including multiplex PCR, 5'RACE (rapid amplification of cDNA Ends), of which 5'RACE-PCR has been preferred for deep sequencing methods used in functional studies. During the sequencing process errors can occur and understanding the whole process throughout is required for minimizing such errors. There are several sequencing platforms available, among which Illumina sequencing platform has the advantages of shorter read length, reduced cost and importantly higher throughput.[52] With the increasing demand for immunoglobulin repertoire sequencing with clinical purposes of clonality detection and minimal residual disease monitoring, commercially available reagent kits have been developed, validated and approved for clinical use.[53–57] While hematologic malignancies of the B cell lineage such as chronic lymphocytic leukemia, multiple myeloma are first line of disease applications, its use and application has potential for widespread use, with higher sensitivity and specificity.[58] There has been a collaborative effort for the standardization of PCR-based immunoglobulin and TCR clonality testing including both pre- and post-analytical aspects of clonality testing, made by the EuroClonality (BIOMED-2) consortium.[59] Commercially available reagent kits are in accordance with these efforts.

Lastly, the selection of bioinformatics to analyze the massive data generated by HTS requires consideration. There are openly available tools for repertoire analyses, while development of one's own tool is possible, but would require tremendous expertise in the narrow scope of computer science and bioinformatics. In conjunction with the efforts to standardize the AIRR data, known as the AIRR data commons, there are several popular standard tools.[15,34,60] MiXCR/MiTCR is commonly used for profiling of AIRR-seq data, with advantages of faster algorithm and ease of use.[61] Other popular AIRR-seq data processing tools include VDJPipe and Presto.[62,63] These tools allow assignment of V(D)J genes and sequence identification of CDR3 and more. Once sequences are preprocessed and assigned specific V(D)J gene sequences, the following steps in AIRR-seq analysis is called the post-processing or the repertoire analysis step. As mentioned above, a clone originating from the same immunoglobulin requires equal V and J sequences and CDR3 lengths. Grouping such sequences into group, i.e., clonotyping requires additional bioinformatics tools such as Change-O, VDJtools and more.[64–66] Establishing a pipeline for AIRR-seq is up to the purpose and scope of the study, starting from the specimen to post-processing repertoire analysis, to which a gold standard is yet to be set.

Organ transplantation is a clinical application of transplant immunology, in which understanding of the immunology at play has greatly improved the outcome of graft survival. Transplantation is the preferred treatment for chronic kidney disease and development of successful immunosuppressants have been essential in this improvement in clinical outcome, especially for kidney transplantation (KT).[67] Without proper immunosuppression, grafts are at risk of rejections, and the most frequent cause of graft loss remains to be antibody-mediated rejection (ABMR). While the survival and outcome of KT has made a leap through the use of potent immunosuppressants and improving surgical techniques, the long-term graft survival rate has not much improved since 1990's, according to the US renal data system.[68] A recent domestic study on large cohort of KT recipients between 2002 and 2017 describes the graft survival rate to be at 90.3% from living donors and 85.6% from cadaveric donors.[69] Given that graft loss within the first year has been avoided, which also coincides with the report of steady decrease in acute rejection, ABMR remains to be the main cause of long term graft loss.[70,71] There is currently a unmet clinical need for discovery of a biomarker that allows early and specific detection of ABMR and possibly, provide a potential therapeutic target. Complexity of the mechanism through which ABMR occurs have multiple factors involved during its process, which includes but are not limited to, antibody produced by B cells, human leukocyte antigen (HLA) mismatch, degree of sensitization, underlying medical conditions etc. Among these factors the presence of the donor specific antibody (DSA) is probably the most adverse factor for achieving transplantation tolerance.[72] However, antibodies of non-HLA nature are also associated with such alloimmune responses that leads to ABMR, and the role of B cells and its produced antibodies are increasingly highlighted.[73,74] Application of AIRR-seq for the field of organ transplantation is becoming more appropriate as discoveries are being made that antibody repertoire is closely associated with inflammatory responses, alloimmune responses elicited by solid organ transplantation. Such changes in antibody repertoire have been associated with graft dysfunction or loss after KT.[75–79]

A series of previous studies have demonstrated the difference of immune repertoire correlates with the post-transplant rejection risk.[75] Reduction in diversity of the immune repertoire likely resulted from the immune suppression or immunogenic antigens causing expansion of certain persistent clones. Such clonal expansion was not only observed in B cells but also in T cells, in which the clonal T cell expansion correlated with the graft dysfunction demonstrated by elevated serum creatinine levels.[80] The diversity of T cell repertoire was shown to be lower in transplanted groups when compared to non-transplant control groups, and also showed distinct clonotype distributions.[78] A follow-up study demonstrated that such decrease in diversity was also demonstrated in B cell repertoire, and such changes were readily detectable by NGS as soon as day 1 after transplantation.[79] To the best of our knowledge, the decrease in diversity of the immune repertoire after transplantation is well demonstrated, although the contribution of immunosuppression and individual variation needs to be considered.

Many of the previous literature have focused on T cell repertoire using T cells, and only limited number of studies using B cells have been conducted. To understand the sequence related properties of the antibody and measure the diversity of B cell repertoire in correspondence to the clinical outcome, AIRR-seq is applied to a cohort of KT recipients in this study. With the goal of characterizing the antibody repertoire among different clinical outcome and phenotypes, we have evaluated a composite pipeline of AIRR-seq which provides more accessibility and ease of use. The results of AIRR-seq can be analyzed in measures of clonality, diversity and gene usages. The aim of the study is to provide clinical implications of AIRR-seq for detection of ABMR and provide the possibility of AIRR-seq as a universal laboratory test, which can provide diagnosis, response to drug (immunosuppressants) and monitoring of clinically relevant conditions.

This study also aims to compare a variety of available AIRR-compliant tools, in the context that each tool is somewhat limited in certain aspects, e.g., number of sequences able to handle, incompleteness of suite of tools to complete the whole process of AIRR-seq, and metadata handing capabilities, etc. Table 1 provides a summary of commonly used AIRR-seq tools and their functions. Yet only a few software packages are AIRR-compliant but as more are becoming compliant, the AIRR format will be utilized more in terms of standardization and reproducibility. As AIRR studies vary from another in various ways, comparison of popular tools which have been only compared in a broader scope is informative.[81,82]

	Merging paired-end reads	Barcode demultiplexing	PCR primer masking	Read length filter	Read quality filter	Homopolymer filtering	UMI consensus determination	Collapsing duplicate reads	Clonotyping
MiXCR	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
VDJPipe	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No*
pRESTO*	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No*
LymhoTrack	Yes	No	Yes	Yes	No	No	Undisclosed**	No	Yes

Table 1. Comparison of commonly used preprocessing AIRR-seq tools

*pRESTO is provided in VDJServer but was not utilized in this study due to its limitations such as requirement for primer sequences.

**Undisclosed company confidential information.

연구 대상 및 방법 Materials and Methods

1. Study design and population

A cohort of KT recipients who under received transplantation between December 1996 and March 2021 was used for selection of samples used in this study. Kidney biopsy was performed for these patients during the period of January 2018 and July 2021, indicated as either protocol biopsies or in suspicion of rejection. Samples of 15 recipients were carefully selected, based on availability of kidney biopsy data, matching blood samples including serum, buffy coat and/or PBMC. The patients were comprehensively evaluated of their transplant status and the clinical condition of the graft. Kidney biopsy specimens were reviewed by board-certified pathologists applying the semi-quantitative histological scores from the Banff 2019 classification.[83] Additional criteria of C4d positivity to determine the possibility of presence of ABMR (Table 2.) was implemented in this study, because in many clinical situations were mixed-phenotype rejections present. A histologic diagnosis of T cell mediated rejection (TCMR) was used as an exclusion criterion for either of study groups. Also, patient samples without apparent evidence of ABMR were categorized as 'No rejection (NR)'. The possibility of subclinical injury and presence of anti-HLA antibody without the manifestation of C4d positivity is present in NR groups. However, these limited number of NR group patients were strictly selected with the category (Table 1). All specimens were collected with provided informed consent to the participants. The study was conducted in adherence to the Declaration of Helsinki.

	ABMR (*N* = 11)	No rejection (*N* = 4)
Biopsy diagnosis	Antibody-mediated rejection Suspicious for ABMR with additional diagnosis of Borderline T cell-mediated rejection	No evidence of rejection Non-specific changes
Exclusion Criteria	Absence of donor specific antibody Absence of C4d deposition (c = 0) Acute TCMR and/or chronic active TCMR	Presence of any class II HLA antibody >10,000 (MFI) Any of borderline TCMR, chronic active TCMR and ABMR

Table 2. Classification criteria for sample groups

2. Specimen processing and DNA extraction

Purified genomic DNA was extracted from deep freezer (-70℃) stored buffy coat specimens using the QIAamp DNA Mini kit (Qiagen, Valencia, CA, USA) and the QIAcube instrument (Qiagen). With the minimum input quantity of 50 ng high-quality DNA, DNA concentrations were measured by fluorimetry method after extraction and purification using the Qubit dsDNA HS assay kit (Life Technologies, Carlsbad, CA, USA).

3. Library amplification and purification

LymphoTrack IGH FR1 assay kit - Miseq (Invivoscribe, Inc. San Diego, CA, USA) was used according to the manufacturer's instructions. Amplification PCR was performed using 2 ug of gDNA as input, independently barcoded and using EagleTaq DNA polymerase (Roche). PCR purification was done using the Agencourt AMPure XP Bead (Beckman Coulter). Master mix containing gDNA, reagents and DNA polymerase and Illumina linkers to the amplicons were Qiagen Multiplex PCR kit (Qiagen).

The PCR reaction mixture was prepared in 55.3 uL in total with the following: total 8 uL of sample DNA and nuclease free water, 2 uL of LymphoQuant internal control (LQIC) DNA, 45 uL of master mix, 0.3 uL of DNA polymerase. PCR program setting in brief, was as follows; 95℃ for 7 minutes (1 cycle), 95℃ for 45 seconds, 60℃ for 45 seconds and 72℃ for 90 seconds (29 cycles), 72℃ for 10 minutes (1 cycle) and 4℃ cooling cycle to finish. For PCR product purification, the final PCR product was mixed with AMPure XP reagent in 1:1 ratio with nuclease free water as elution solution. Average fragment length of IGH is about 450 bp, which the length (base pairs) and concentration (ng/uL) and the molarity (nmol/L) were measured after elution. The final library is prepared with dilution with 30% PhiX. Total volume of 700 uL consisted of 343 uL of 20pM library, 147 uL of 20pM PhiX, 210 uL of HT1 buffer.

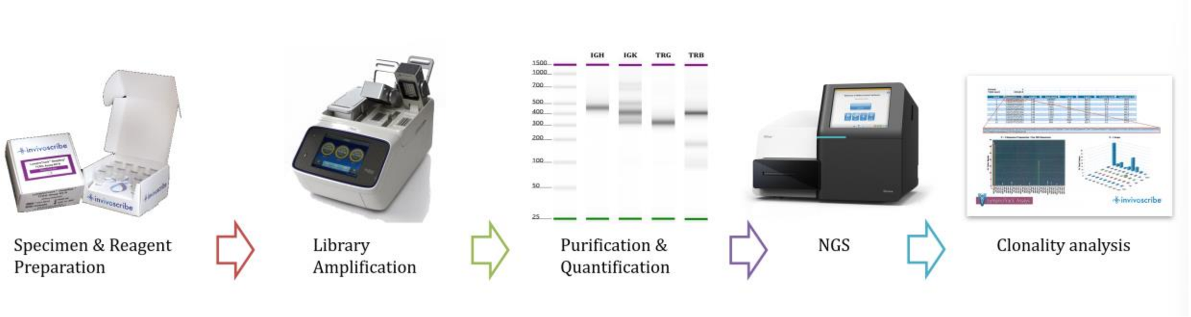


Figure 2. Workflow of AIRR-seq using LymphoTrack IGH FR1 assay kit - MiSeq

4. NGS and data analysis

Libraries were sequenced with with MiSeq Reagent v2 kit (Illumina, San Diego, CA, USA) on a MiSeqDx instrument (Illumina). The generated FASTQ files were stored for further analyses, while LymphoTrack-MiSeq version 2.4.3 (Invivoscribe, Inc.) according to the user manual provided by the manufacturer. NGS library was constructed with the quality control criteria of the following: 1) cluster density: 800-1200 k/mm^2, 2) cluster passing filter: >80% and 3) Q_{30} >30%.

5. AIRR-seq analysis pipeline

The AIRR-seq pipeline following the NGS experiment is described in Figure 2. To overcome the limit of only including small number of samples in NR group, we utilized the NCBI Sequence Reads Archive (SRA) data for recruiting normal healthy control data (https://www.ncbi.nlm.nih.gov/sra). Using the sequence reads and the obtained raw reads in FASTQ format, the downstream analysis was done. Sequences from SRA were publicly available under the BioProject number PRJNA406949.

The workflow of AIRR-seq can be categorized into three major steps: 1) preprocessing, 2) annotation - V(D)J assignment and CDR3 identification, and 3) clonotyping and repertoire analyses (Figure 3). Preprocessing process includes the quality control of the obtained sequences using dedicated tools, such as FAST QC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC.[84] As Illumina platform generates paired-end sequences, both reads R1and R2 were checked for sequence quality scores and sequence length distributions using MultiQC. After passing quality QC, the sequences require merging of R1 and R2, for which MiXCR, LymphoTrack-MiSeq software and VDJPipe were independently used as comparison.[61,62] Figure 4 is a summary flow chart of the AIRR-seq pipelines used in this study.

Merged sequences from paired-end sequencing are passed on for annotation, V(D)J assignment and CDR3 identification, for which the IMGT database was used as a reference sequence.[85] The details of method as to how each pipeline performs such annotation differs from each other, for example the MiXCR tool utilizes an independent algorithm, the kAligner2 algorithm, whereas the VDJPipe tool uses the Hamming distance algorithm. When using the IMGT/HighV-QUEST tool, the Smith-Waterman algorithm is implemented, and the classical BLAST algorithm is used in IgBlast tool. An important distinction between the annotation tools is the ability to perform clonotyping, in which the sequences are grouped and categorized according to their identified CDR3 sequences and V(D)J gene, allele and families. After assignment and CDR3 identification is finished, as clonotyping, the repertoire analysis is performed using the following tools: Immunarch and VDJServer (Figure 5.).[86,87] Cloud- or web-based services such as VDJServer (https://vdjserver.org/) was used with ethernet access and other stand-alone tools were installed and utilized from a local workstation, of which the detailed specification of the workstation used in this study is described in Table 3. The repertoire analysis included clonality, diversity, CDR3 analysis, V(D)J gene usage and somatic hypermutation profiles, which were commonly included in the Immunarch package and VDJServer.

The repertoire analysis consisted of 4 major categories. These were, 1) basic statistics and clonality, 2) diversity, 3) CDR3 analysis, and 4) V(D)J usage. The basic statistics include number of clones, distribution of lengths and counts. The clonality analysis is also included with the basic statistics analysis. The diversity estimation includes number of different approaches, which includes Chao1, Hill numbers, the Gini-Simpson index and the rarefraction analysis. CDR3 analysis mainly compares sample data using distribution of CDR3 lengths, in either nucleotide level or amino acid level. The gene usage analysis compares the gene segment data of each sample, of which the segment data are annotated using the IMGT nomenclature. The gene segments included in this study were all of IGH origin, and the gene usage analysis is limited to the IGHV and IGHJ genes. For uniformity of the analysis, the Immunarch R package was used for visualization of repertoire analyses.

6. Reproducibility and read-depth associated changes in clonotypes

To address the controversial topic of adequate reads required to properly assess the adaptive immune repertoire, we tested limited number samples in duplicates, utilized repository data which includes repeated NGS data of the same individuals for assay reproducibility. For reproducibility, sample data from the SRA accession number PRJNA349143 (influenza vaccination response study). The samples were collected from the same donors at equal time intervals, before 8 days, before 2days, and before 1 hour, before vaccinations were given to subjects. The subject IDs were FV, GMC and IB with serial numbers given in timely order (e.g., before 8 days - FV1, before 2 days – FV2 and before 1hour FV3).

The number of reads were compared to the final clonotype numbers, to assess the relationship between reads and clonotypes. The raw read depth of course does not take into consideration the efficiency of alignment process, but comparison of raw reads and the final clonotype output can provide a simple comparison when identical preprocessing pipeline is applied.

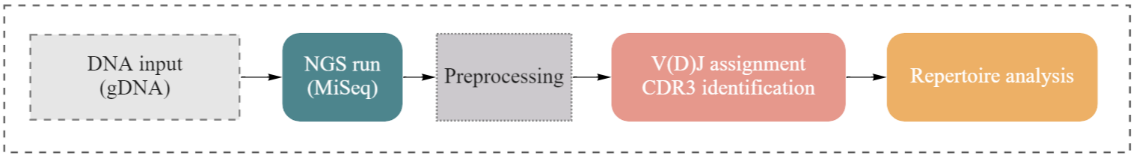


Figure 3. A concise overview of the workflow of AIRR-seq

Processor	Intel i7-7820HQ (2.9 GHz) 8MB Cache
Memory	32GB 2400MHz DDR4
Storage	512GB PCIe M.2 SSD (32Gb/s) and 2TB HDD
Graphic	NVIDIA Quadro M2200, 4GB
Operating system	Windows 10 Pro (64 bit)

Table 3. Specifications of the local workstation used for AIRR-seq.

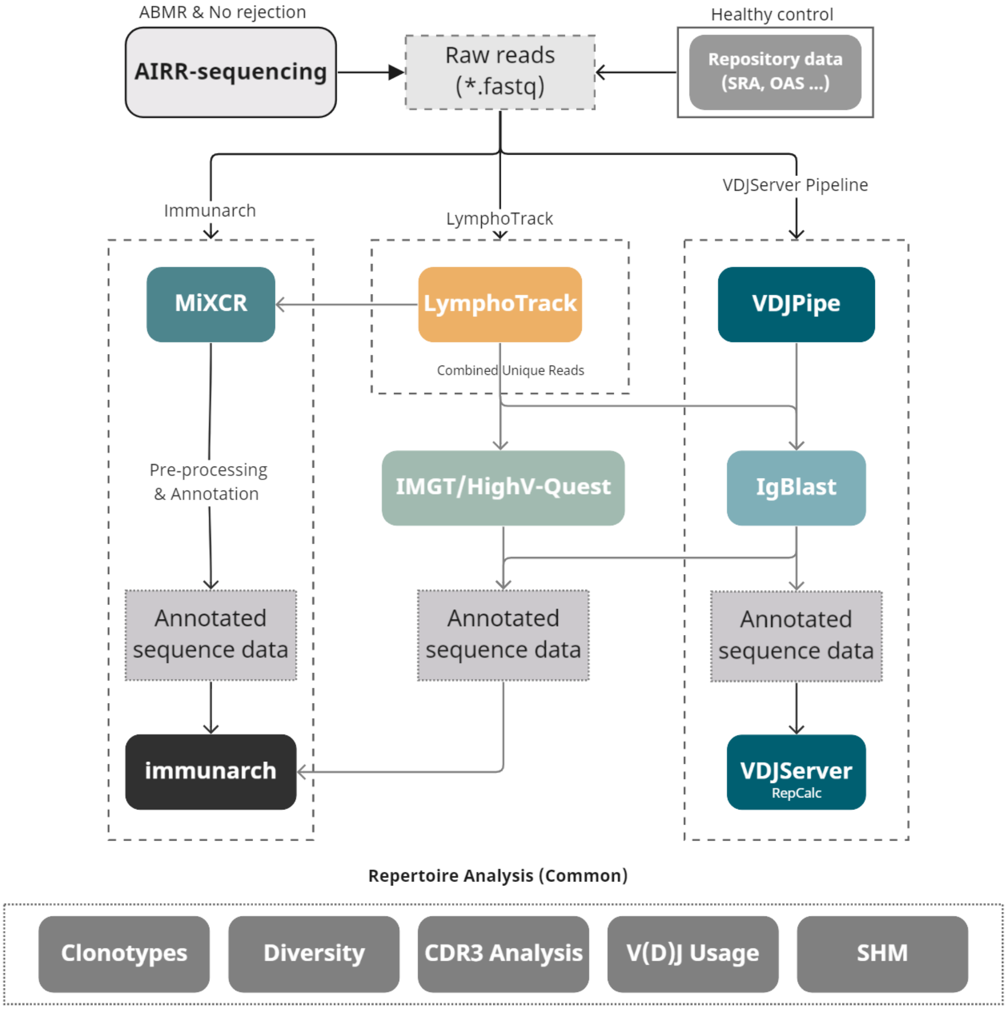
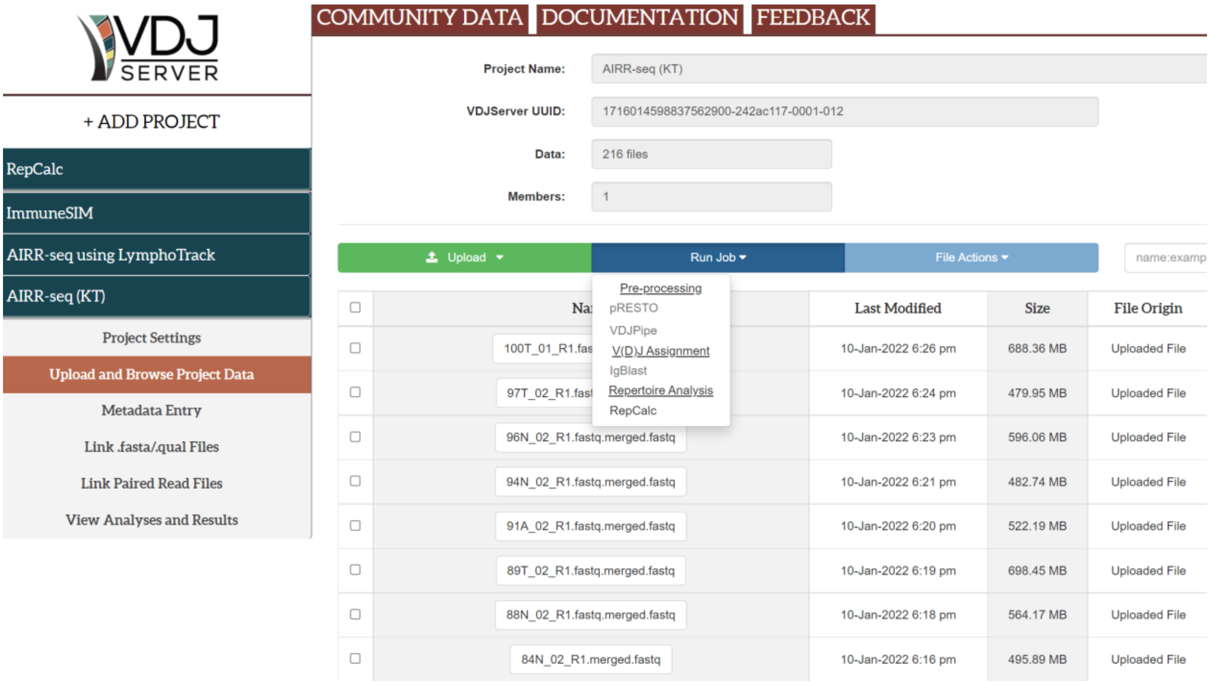


Figure 4. Study Flowchart summarizing the pipeline of AIRR-seq and its comparison



Project Name: AIRR-seq (KT)

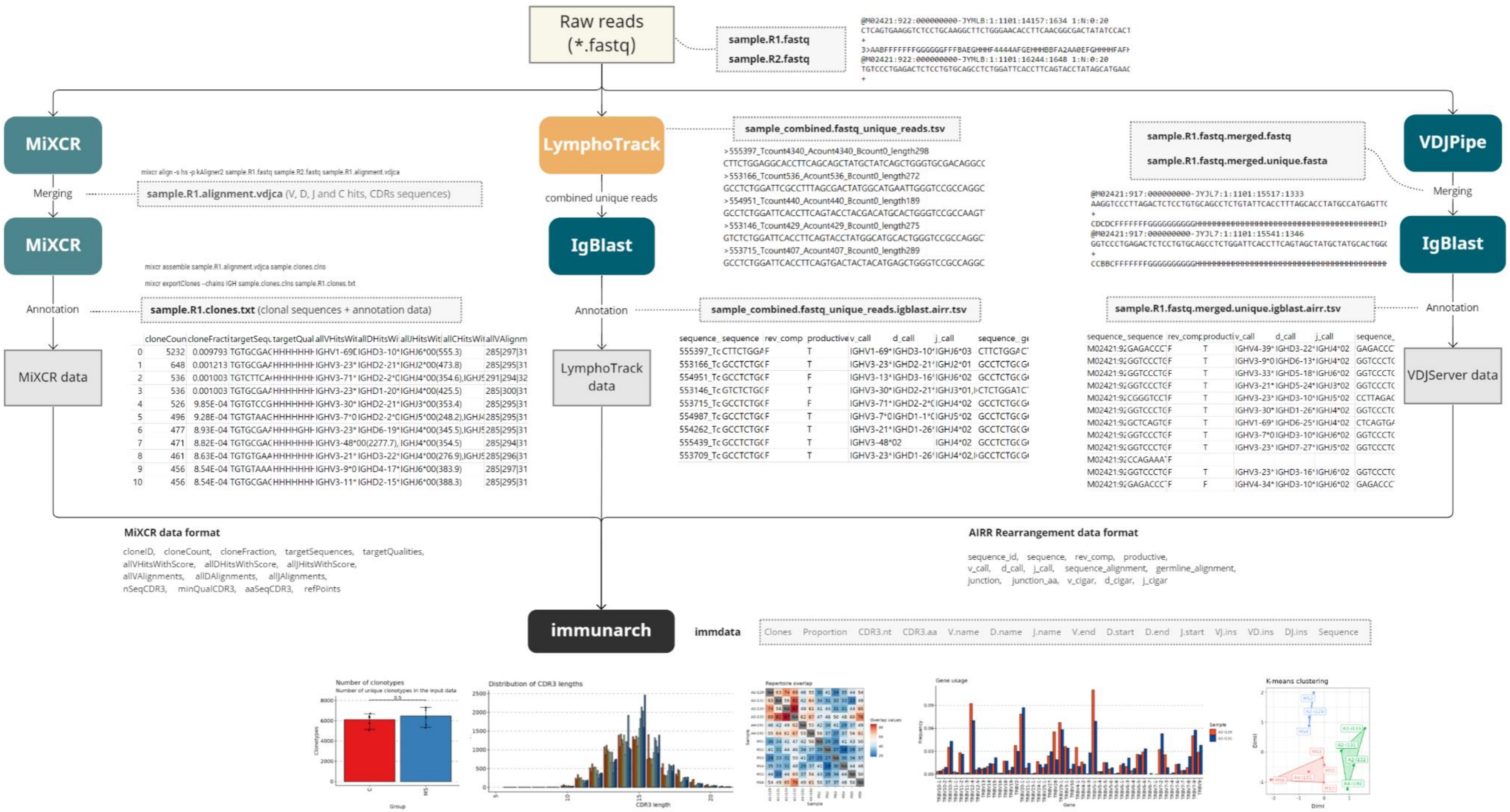
VDJServer UUID: 1716014598837562900-242ac117-0001-012

Data: 216 files

Members: 1

☐	Na	Last Modified	Size	File Origin
☐	100T_01_R1.fas	10-Jan-2022 6:26 pm	688.36 MB	Uploaded File
☐	97T_02_R1.fast	10-Jan-2022 6:24 pm	479.95 MB	Uploaded File
☐	96N_02_R1.fastq.merged.fastq	10-Jan-2022 6:23 pm	596.06 MB	Uploaded File
☐	94N_02_R1.fastq.merged.fastq	10-Jan-2022 6:21 pm	482.74 MB	Uploaded File
☐	91A_02_R1.fastq.merged.fastq	10-Jan-2022 6:20 pm	522.19 MB	Uploaded File
☐	89T_02_R1.fastq.merged.fastq	10-Jan-2022 6:19 pm	698.45 MB	Uploaded File
☐	88N_02_R1.fastq.merged.fastq	10-Jan-2022 6:18 pm	564.17 MB	Uploaded File
☐	84N_02_R1.merged.fastq	10-Jan-2022 6:16 pm	495.89 MB	Uploaded File

Figure 5. VDJServer displaying functions of preprocessing, V(D)J assignment and repertoire analysis



```
@M02421:922:000000000-JYMLB:1:1101:14157:1634 1:N:0:20
CTCAGTGAAGGTCTCCTGCAAGGCTTCTGGGAACACCTTCAACGGCGACTATATCCACT
+
3>AABFFFFFFFGGGGGGFFFBAEGHHHF4444AFGEHHHBBFA2AA0EFGHHHHFAFH
@M02421:922:000000000-JYMLB:1:1101:16244:1648 1:N:0:20
TGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCACCTTCAGTACCTATAGCATGAAC
+
```

```
mixcr align -s hs -p kAligner2 sample.R1.fastq sample.R2.fastq sample.R1.alignment.vdjca
```

sample.R1.alignment.vdjca (V, D, J and C hits, CDRs sequences)

```
mixcr assemble sample.R1.alignment.vdjca sample.clones.clns
mixcr exportClones --chains IGH sample.clones.clns sample.R1.clones.txt
```

sample.R1.clones.txt (clonal sequences + annotation data)

	cloneCoun	cloneFracti	targetSequ	targetQual	allVHitsWit	allDHitsWi	allJHitsWit	allCHitsWit	allVAlignm
0	5232	0.009793	TGTGCGAC	HHHHHHH	IGHV1-69D	IGHD3-10*	IGHJ6*00(555.3)		285|297|31
1	648	0.001213	TGTGCGAA	HHHHHHH	IGHV3-23*	IGHD2-21*	IGHJ2*00(473.8)		285|295|31
2	536	0.001003	TGTCTTCA	HHHHHHH	IGHV3-71*	IGHD2-2*0	IGHJ4*00(354.6),IGHJ5		291|294|32
3	536	0.001003	TGTGCGAA	HHHHHHH	IGHV3-23*	IGHD1-20*	IGHJ4*00(425.5)		285|300|31
4	526	9.85E-04	TGTGTCCG	HHHHHHH	IGHV3-30*	IGHD2-21*	IGHJ3*00(353.4)		285|295|31
5	496	9.28E-04	TGTGTAAG	HHHHHHH	IGHV3-7*0	IGHD2-2*0	IGHJ5*00(248.2),IGHJ4		285|295|31
6	477	8.93E-04	TGTGCGAA	HHHHGHH	IGHV3-23*	IGHD6-19*	IGHJ4*00(345.5),IGHJ5		285|295|31
7	471	8.82E-04	TGTGCGAC	HHHHHHH	IGHV3-48*00(2277.7),		IGHJ4*00(354.5)		285|294|31
8	461	8.63E-04	TGTGTGAA	HHHHHHH	IGHV3-21*	IGHD3-22*	IGHJ4*00(276.9),IGHJ5		285|296|31
9	456	8.54E-04	TGTGTAAA	HHHHHHH	IGHV3-9*0	IGHD4-17*	IGHJ6*00(383.9)		285|297|31
10	456	8.54E-04	TGTGCGAC	HHHHHHH	IGHV3-11*	IGHD2-15*	IGHJ6*00(388.3)		285|295|31

sample_combined.fastq_unique_reads.tsv

```
>555397_Tcount4340_Acount4340_Bcount0_length298
CTTCTGGAGGCACCTTCAGCAGCTATGCTATCAGCTGGGTGCGACAGGCC
>553166_Tcount536_Acount536_Bcount0_length272
GCCTCTGGATTCGCCTTTAGCGACTATGGCATGAATTGGGTCCGCCAGGC
>554951_Tcount440_Acount440_Bcount0_length189
GCCTCTGGATTCACCTTCAGTACCTACGACATGCACTGGGTCCGCCAAGT
>553146_Tcount429_Acount429_Bcount0_length275
GTCTCTGGATTCACCTTCAGTACCTATGGCATGCACTGGGTCCGCCAGGC
>553715_Tcount407_Acount407_Bcount0_length289
GCCTCTGGATTCACCTTCAGTGACTACTACATGAGCTGGGTCCGCCAGGC
```

sample_combined.fastq_unique_reads.igblast.airr.tsv

sequence_	sequence	rev_comp	productive	v_call	d_call	j_call	sequence_	ge
555397_Tc	CTTCTGGA	F	T	IGHV1-69*	IGHD3-10*	IGHJ6*03	CTTCTGGA	CT
553166_Tc	GCCTCTGC	F	T	IGHV3-23*	IGHD2-21*	IGHJ2*01	GCCTCTGC	G
554951_Tc	GCCTCTGC	F	F	IGHV3-13*	IGHD3-16*	IGHJ6*02	GCCTCTGC	G
553146_Tc	GTCTCTGC	F	T	IGHV3-30*	IGHD2-21*	IGHJ3*01,I	CTCTGGA1	CT
553715_Tc	GCCTCTGC	F	F	IGHV3-71*	IGHD2-2*0	IGHJ4*02	GCCTCTGC	G
554987_Tc	GCCTCTGC	F	T	IGHV3-7*0	IGHD1-1*0	IGHJ5*02	GCCTCTGC	G
554262_Tc	GCCTCTGC	F	T	IGHV3-21*	IGHD1-26*	IGHJ4*02	GCCTCTGC	G
555439_Tc	GCCTCTGC	F	T	IGHV3-48*02		IGHJ4*02	GCCTCTGC	G
553709_Tc	GCCTCTGC	F	T	IGHV3-23*	IGHD1-26*	IGHJ4*02,I	GCCTCTGC	G

sample.R1.fastq.merged.fastq

sample.R1.fastq.merged.unique.fasta

```
@M02421:917:000000000-JYJL7:1:1101:15517:1333
AAGGTCCCTTAGACTCTCCTGTGCAGCCTCTGTATTCACCTTTAGCACCTATGCCATGAGTTC
+
CDCDCFFFFFFFGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIH
@M02421:917:000000000-JYJL7:1:1101:15541:1346
GGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCACCTTCAGTAGCTATGCTATGCACTGGC
+
CCBBCFFFFFFFGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
```

sample.R1.fastq.merged.unique.igblast.airr.tsv

sequence_	sequence	rev_comp	producti	v_call	d_call	j_call	sequence_
M02421:92	GAGACCCT	F	T	IGHV4-39*	IGHD3-22*	IGHJ4*02	GAGACCCT
M02421:92	GGTCCCTG	F	T	IGHV3-9*0	IGHD6-13*	IGHJ4*02	GGTCCCTC
M02421:92	GGTCCCTG	F	T	IGHV3-33*	IGHD5-18*	IGHJ6*02	GGTCCCTC
M02421:92	GGTCCCTG	F	T	IGHV3-21*	IGHD5-24*	IGHJ3*02	GGTCCCTC
M02421:92	CGGGTCCT	F	T	IGHV3-23*	IGHD3-10*	IGHJ5*02	CCTTAGAC
M02421:92	GGTCCCTG	F	T	IGHV3-30*	IGHD1-26*	IGHJ4*02	GGTCCCTC
M02421:92	GCTCAGTG	F	T	IGHV1-69*	IGHD6-25*	IGHJ4*02	CTCAGTGA
M02421:92	GGTCCCTG	F	T	IGHV3-7*0	IGHD3-10*	IGHJ6*02	GGTCCCTC
M02421:92	GGTCCCTG	F	T	IGHV3-23*	IGHD7-27*	IGHJ5*02	GGTCCCTC
M02421:92	CCAGAAAT	F					
M02421:92	GGTCCCTG	F	T	IGHV3-23*	IGHD3-16*	IGHJ6*02	GGTCCCTC
M02421:92	GAGACCCT	F	F	IGHV4-34*	IGHD3-10*	IGHJ6*02	GAGACCCT

MiXCR data format

cloneID, cloneCount, cloneFraction, targetSequences, targetQualities, allVHitsWithScore, allDHitsWithScore, allJHitsWithScore, allVAlignments, allDAlignments, allJAlignments, nSeqCDR3, minQualCDR3, aaSeqCDR3, refPoints

AIRR Rearrangement data format

sequence_id, sequence, rev_comp, productive, v_call, d_call, j_call, sequence_alignment, germline_alignment, junction, junction_aa, v_cigar, d_cigar, j_cigar

Clones Proportion CDR3.nt CDR3.aa V.name D.name J.name V.end D.start D.end J.start VJ.ins VD.ins DJ.ins Sequence

Figure 6. Comparison of preprocessing pipelines and example data throughout the pipeline

7. Statistical analysis

All data processing and analyses were done using Excel (Microsoft Corporation, Redmond, WA, USA) and the installed R packages using RStudio v.1.4.1717 (RStudio Inc., Boston, MA, USA). Installed dependencies not specified as an independent tool includes the following: *ggplot2 3.1.0, dplyr 0.8.0, dtplyr 1.0.0, data.table 1.12.6, patachwork, factoextra 1.0.4, fpc, UpSetR 1.4.0, pheatmap 1.0.12, ggrepel 0.8.0, reshape2 1.4.2, circlize, MASS 7.3, Rtsne 0.15, readxl 1.3.1, shiny 1.4.0, shinythemes, airr, ggseqlogo, ggalluvial 0.10.0, Rcpp 1.0, magrittr, methods, scales, ggpubr 0.2, rlang 0.4, plyr, dbplyr 1.4.0, jsonlite, readr, stringr, tibble, tidyselect, purr*, and the versions specified here indicate the minimum requirements. Packages installed with the R commandline 'install.packages()' was the Immunarch R package. While most statistical analysis and visualization of data were done by ggplot package included in R, results visualized in this study are outputs directly from Immunarch and VDJServer as sources. The difference between groups was calculated by the Wilcoxon rank sum test, when only two groups were used for comparison. If there are more than two groups, i.e. ABMR, NR and the normal control data, the Kruskal-Wallis test was performed. The P-value shown above the plots are adjusted by the Holm-Bonferroni correction. All statistical analysis were done using the functions *wilcox.test*, *kruskal.test*, and *p.adjust* implemented within the R package.

결과 Results

1. Demographics

The summary of the comparison of clinical and laboratory data between the groups are shown in Table 4. It is notable that ABMR group included six female individuals, while no females were included in the NR group. The age was shown to be higher in ABMR group, and the ABMR group exhibited longer elapsed time after transplantation (2,421.5 vs 388 days). The presence of anti-HLA antibody was also distinctive between the groups, in which ABMR group showed higher peak antibody MFI compared to NR group (20,353 vs 6,299). Of note, the MFI for ABMR group indicated the DSA MFIs whereas the MFI in NR group indicated any anti-HLA antibody.

2. Sequencing statistics – preprocessing and replicates

The data of individual samples and the sequencing statistics summary of the preprocessing pipelines are shown in Table 5. The summary statistics of the three preprocessing tools used in this study are shown as each steps of preprocessing is done, from left to right. For each preprocessing tools, the same paired-end raw sequence files (R1 and R2) were used as inputs. The data obtained from the SRA were processed by MiXCR tool, and the summary statistics and the results are shown in Table 6. The summary of normal controls used for reproducibility and read-depth associated clonotype analysis are provided in Supplemental Table 1.

From the comparison of preprocessing workflows, the differing preprocessing steps between tools did not allow a direct sequence number to sequence number comparison, as the naming of each step differed between tools. For example, the overlapping and aligning of the MiXCR tool corresponds to the merging process of VDJPipe and combining process of LymphoTrack. However, the final number of reads used for clonotyping provides an idea of the quantitative changes made to the input data. The MiXCR preprocessing provides the number of final clonotypes, whereas the other preprocessing requires immunarch analysis to discover the corresponding data. Due to differences in aligning algorithms and clonotyping methods, the number of reads differed between pipelines. Comparing the ratio of reads used for clonotyping to total input reads, the MiXCR showed that 80.3% of total input reads were used. The ratio was 62.8% in VDJPipe (using IgBLAST) and 82.2% in LymphoTrack software, in which input read count into IgBLAST and total input read into LymphoTrack MiSeq-Software are summarized in Table 5, respectively.

	ABMR	NR
M/F ratio	5:6	4:0
Age	56.3 (46 ~ 66)	41.3 (29~56)
Time since tpl	2421.5 days (72~6985 days)	388 days (99~893 days)
Mean peak antibody (MFI)	20353	6299

Table 4. Demographic summary of the clinical and laboratory data between groups

Sample	Sex/ Age	Total reads	MiXCR				VDJPipe					LymphoTrack	
			Overlapped and aligned	Read used as core	Read used in clonotyping	Final clones	Merged reads	Filtered reads	Unique reads	IgBlast	RepCalc	Total input read count	Combined unique reads
R11	M/50	735,286	577358	541879	535492	5575	735270	694884	574200	400727	400727	555497	102941
R13	M/61	915,486	822270	797536	786932	30099	915482	884018	710218	538904	538904	713689	135331
R25	M/66	932,712	718134	657985	640228	3484	932472	871824	564185	363884	363884	700180	112110
R35	M/57	685,787	543589	526558	518830	19649	685759	654549	594821	429978	429978	541924	127323
R43	F/62	710,829	583419	568244	557996	32251	710789	682788	643421	477824	477824	581303	148919
R63	F/52	750,685	634505	612597	608308	26270	750669	709412	665292	527810	527810	634390	149423
R64	M/46	673,420	534000	521305	518530	5224	673384	624895	518244	393215	393215	551393	108892
R75	F/50	762,975	633478	622773	620710	5344	762936	707445	556033	435379	435379	648872	114425
R80	F/62	682,240	584849	556548	561296	13305	682204	633098	545387	447684	447684	589206	94779
R81	F/52	755,683	665130	647235	644278	21842	755448	713918	662988	552458	552458	678641	138193
R91	F/63	685,787	623356	600118	595551	16768	730928	683243	619921	491795	491795	629072	144718
N30	M/48	766,337	642646	612909	613623	11794	766315	708609	632250	507982	507982	654390	146538
N84	M/32	770,217	671012	610301	604221	37838	770140	743412	611483	455921	455921	533168	131221
N88	M/29	794,116	655729	632228	633306	40554	794054	747419	707260	553303	553303	661619	157349
N96	M/56	834,502	721537	695160	690835	26670	834215	790614	727846	577283	577283	726012	151778

Table 5. Demographic information of the subjects, the sequencing statistics and preprocessing summary

Sample	Total reads	Total number of unique clonotypes		
		MiXCR	VDJPipe	LymphoTrack
R11	735,286	4028	400727	11632
R13	915,486	10521	538904	28768
R25	932,712	2887	363884	8822
R35	685,787	16133	429978	35754
R43	710,829	26264	477824	44920
R63	750,685	22438	527810	45812
R64	673,420	4371	393215	15255
R75	762,975	4343	435379	15160
R80	682,240	11574	447684	29781
R81	755,683	18113	552458	40945
R91	685,787	13631	491795	35774
N30	766,337	9882	507982	35149
N84	770,217	27258	455921	42925
N88	794,116	34206	553303	54922
N96	834,502	21828	577283	44804

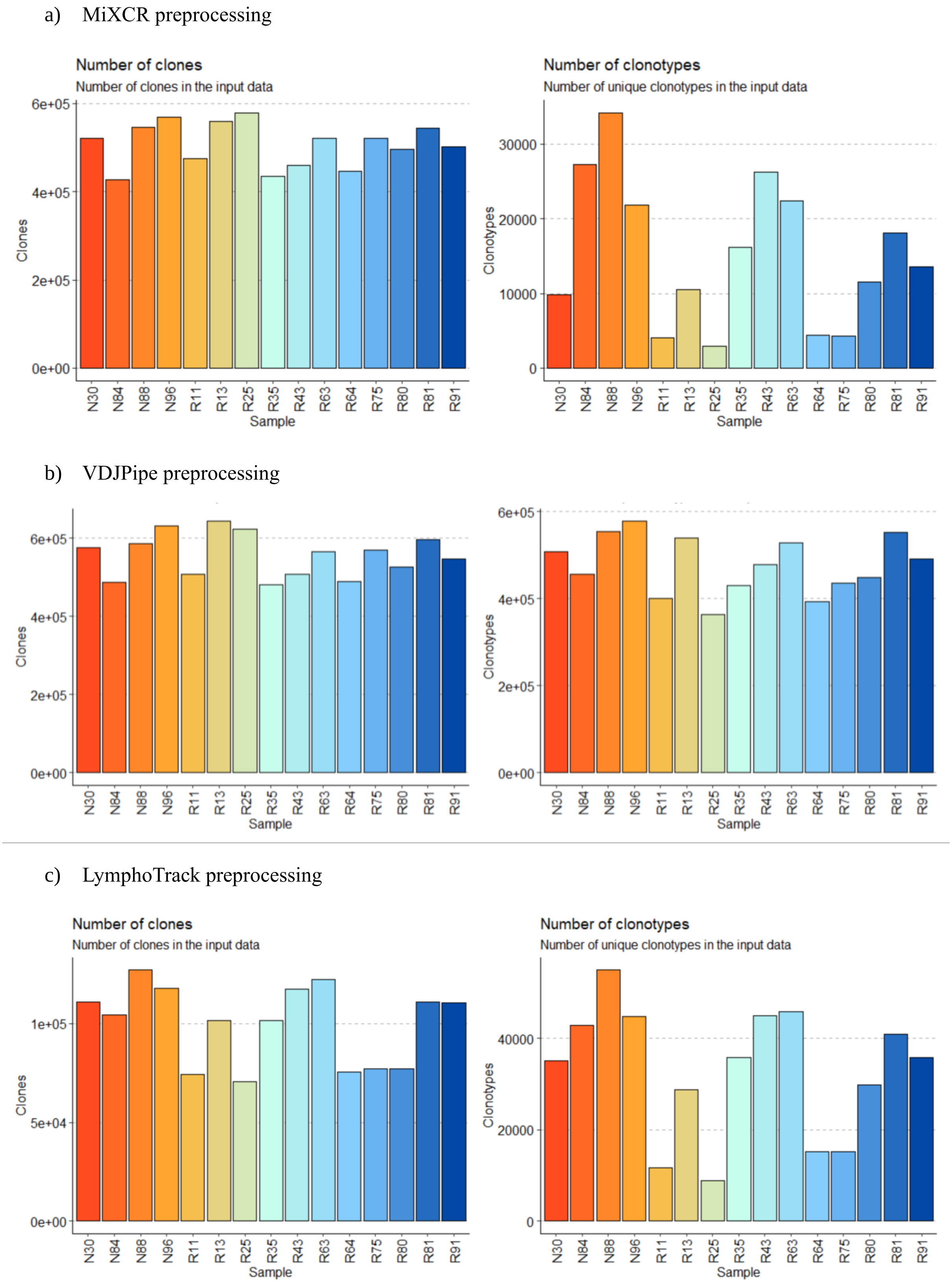
Table 6. The number of unique clonotypes analyzed by Immunarch, using input data from three preprocessing pipelines

3. Basic repertoire statistics and clonality

The AIRR analysis begins with basic repertoire statistics analyses such as number of clones and distribution of lengths and counts. The clonality was analyzed using the repExplore function from the Immunarch R package Table 6 summarizes the total number of unique clonotypes inferred by each preprocessing tools. Figure 7 is the summary of basic repertoire statistics and the comparison of three different preprocessing tools. The results from individual samples are shown, the total number of clones and the number of clonotypes. Along with the process of clonotyping, the number of clones decrease into clonotypes, as sequences of the same progeny are grouped together. This was true for MiXCR and LymphoTrack preprocessed data, while the VDJPipe preprocessed data retained the number of clones (Figure 7, b).

The difference in clonality between ABMR and NR groups were investigated (Figure 8). The data from three preprocessing tools, showed similar distribution of clone numbers and clonotypes, and also similar differences between groups. Although statistical significance was not found in any of the three analyses, a tendency of higher clone number and higher number of clonotypes was consistently observed. It was the initial assumption that there would be a difference of clonality between the groups. However, a statistical significance was absent, and we applied a cutoff of 5,000 clonotypes, re-categorizing the samples into disease category (D) and no rejection category (N). Samples A11, A25, A64 and A75 were classified as D and the other ABMR samples and NR samples were classified as N. The Figure 9 shows that a cutoff of 5,000 clonotypes was statistically significant classifier yet maintaining all NR group samples in the N category (*P* value = 0.002).

The samples showed difference in clonotype results according to different preprocessing, for instance, N30 sample showed particularly lower number of total clones and clonotypes using the MiXCR preprocessing (11,794), compared to the LymphoTrack preprocessing. The difference between preprocessing tools was anticipated, but the results from VDJPipe showed that additional process of clonotyping was required to properly address the repertoire. However, the number of clonotypes shown from MiXCR and LymphoTrack were similar, suggesting that a degree of agreement is present.

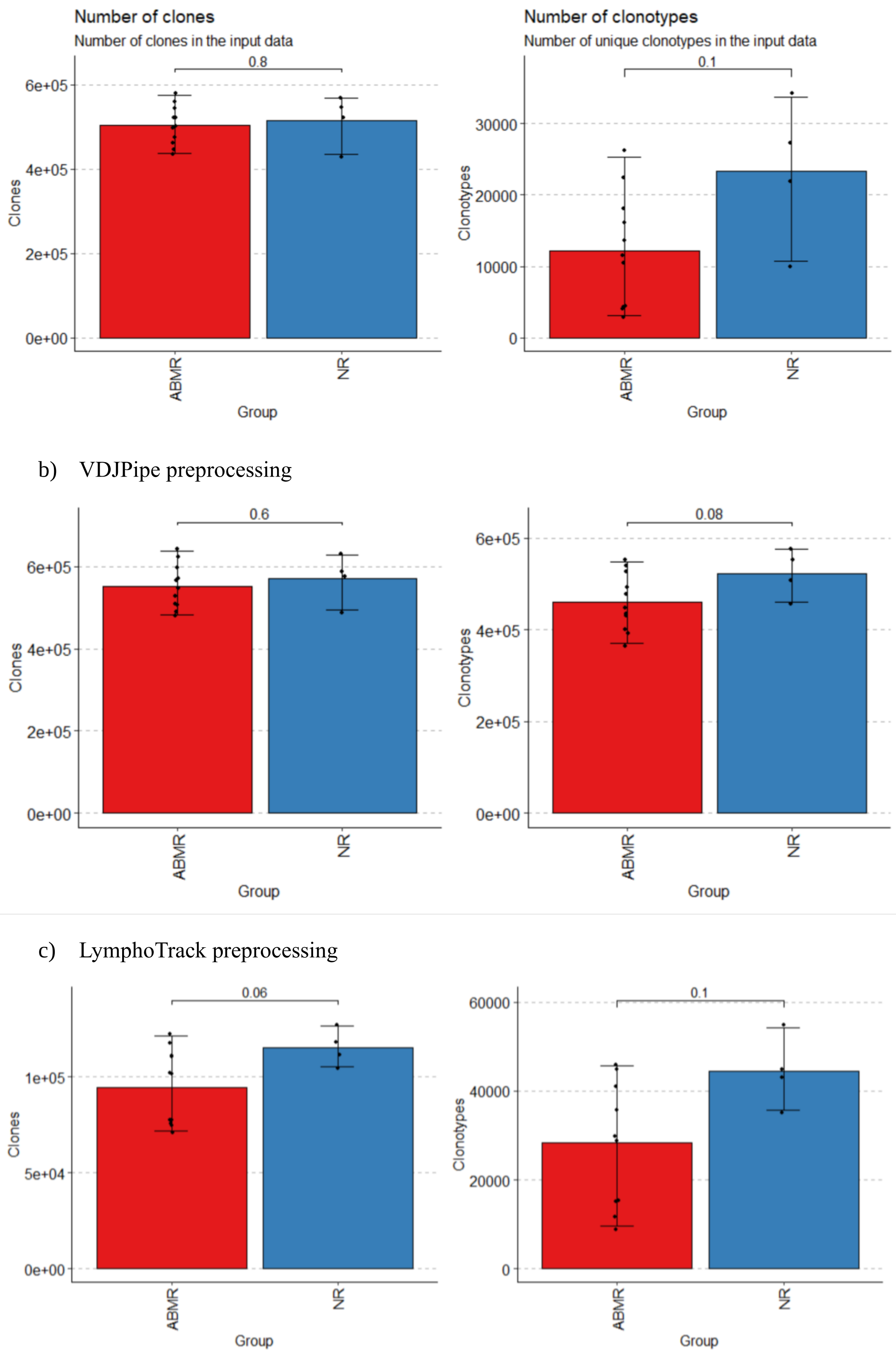

a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 7. The number of clones and clonotypes analyzed by Immunarch using different preprocessing tools.

The number of clones (left) and the number of unique clonotypes (right) are shown.



a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 8. Group comparison of number of clones and number of clonotypes

The number of clones (left) and the number of unique clonotypes (right) are shown.

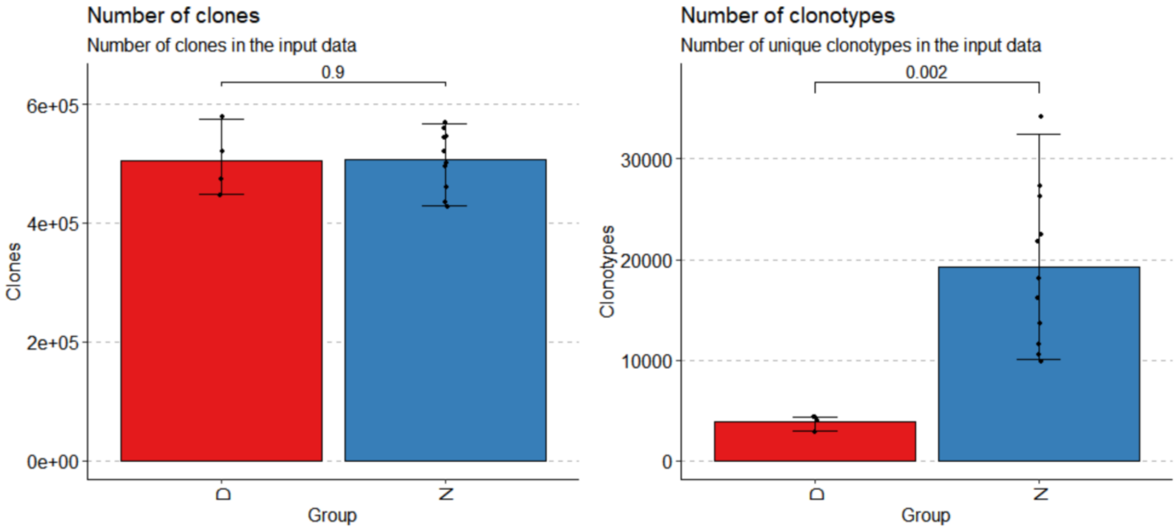
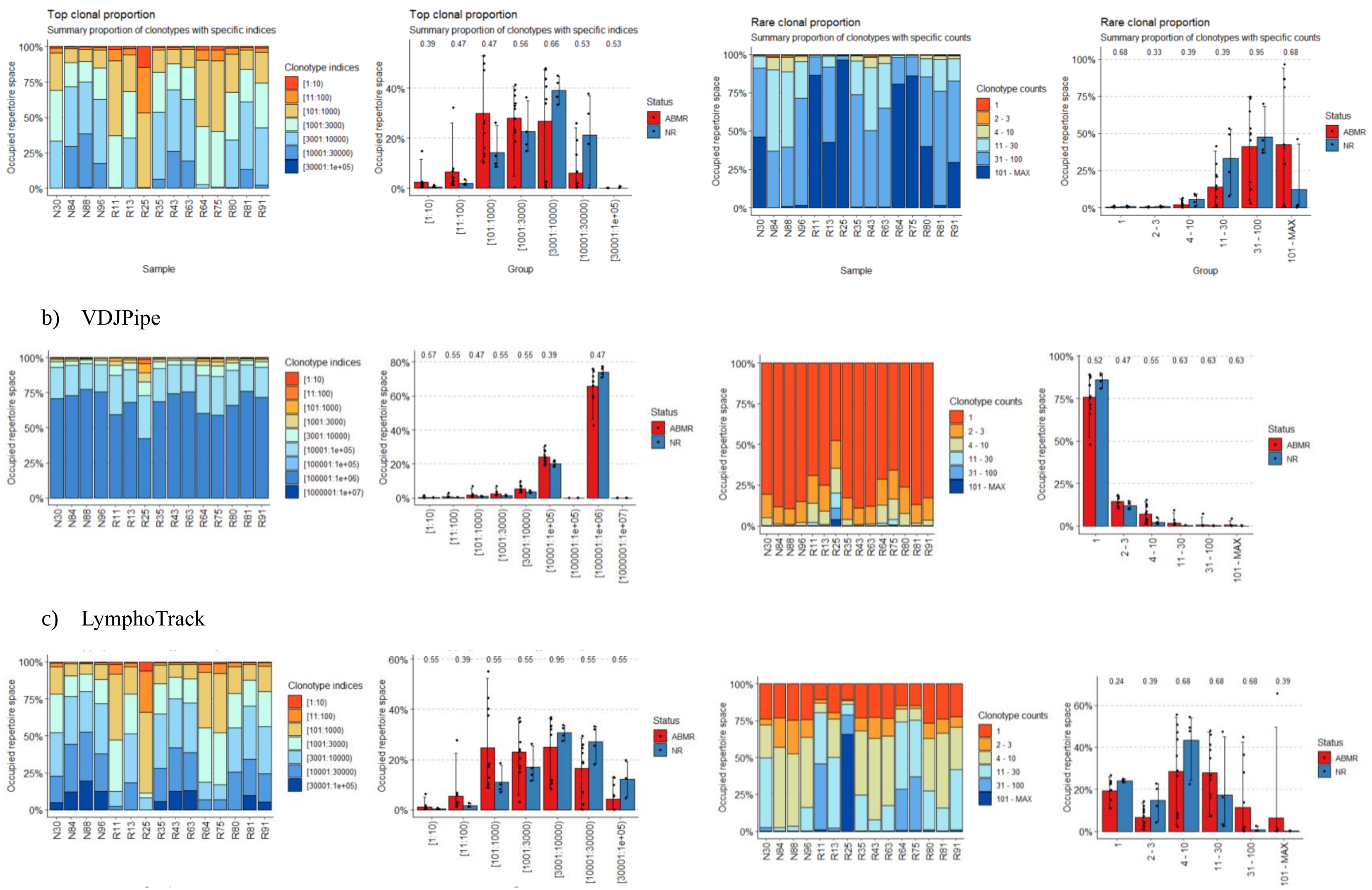


Figure 9. Comparison of clones and clonotypes using the 5,000 clonotype cutoff for reclassification

The clonality analysis also included the estimation and comparison of differences in abundances of clonotypes between samples. This was compared in two methods, 1) proportion of the most abundant clonotypes and 2) proportion of the least abundant clonotypes. Figure 10. shows the clonality analysis results of the three preprocessing tools. Lack of assigning clonotype information in VDJPipe resulted in skewed clonotype analysis results, whereas MiXCR and LymphoTrack were similar in their analyses of top clonal proportions. When assessing the top clonal proportions, ABMR group had higher proportions of upper clonotype indices, suggesting the proliferation of high frequency clonotypes. However, the analyses of rare clonal proportions were dissimilar between all three tools, but the rare fraction (1e-05) in the rare clonal proportion analysis were always higher in the ABMR group.

The repertoire overlap was analyzed to provide the similarities between repertoires of the samples. The 'public' clonotypes were defined as the clonotypes shared between given repertoires, of which their overlap values were visualized into a box plot. We also included the Morisita's overlap index, which measures the dispersion of individuals in the group. There were also other Jaccard, Tversky and cosine indices for calculation of repertoire overlap, but the characteristics of the sample data did not require application of these asymmetric and non-zero vector measures. Figure 11 depicts the overlap of repertoire between samples, of note A25 and N84 showed noticeable overlap of public clonotypes, with overlap value of 1163, 1537 and 1382, respectively. A11 and N88 also showed high overlap values (447, 632 and 528) When Morisita's overlap index of dispersion was applied, the ABMR samples had overlapping results, whereas the NR group samples were mostly independent from all other samples, displaying overlap values of below <0.2, and mostly below <0.1.

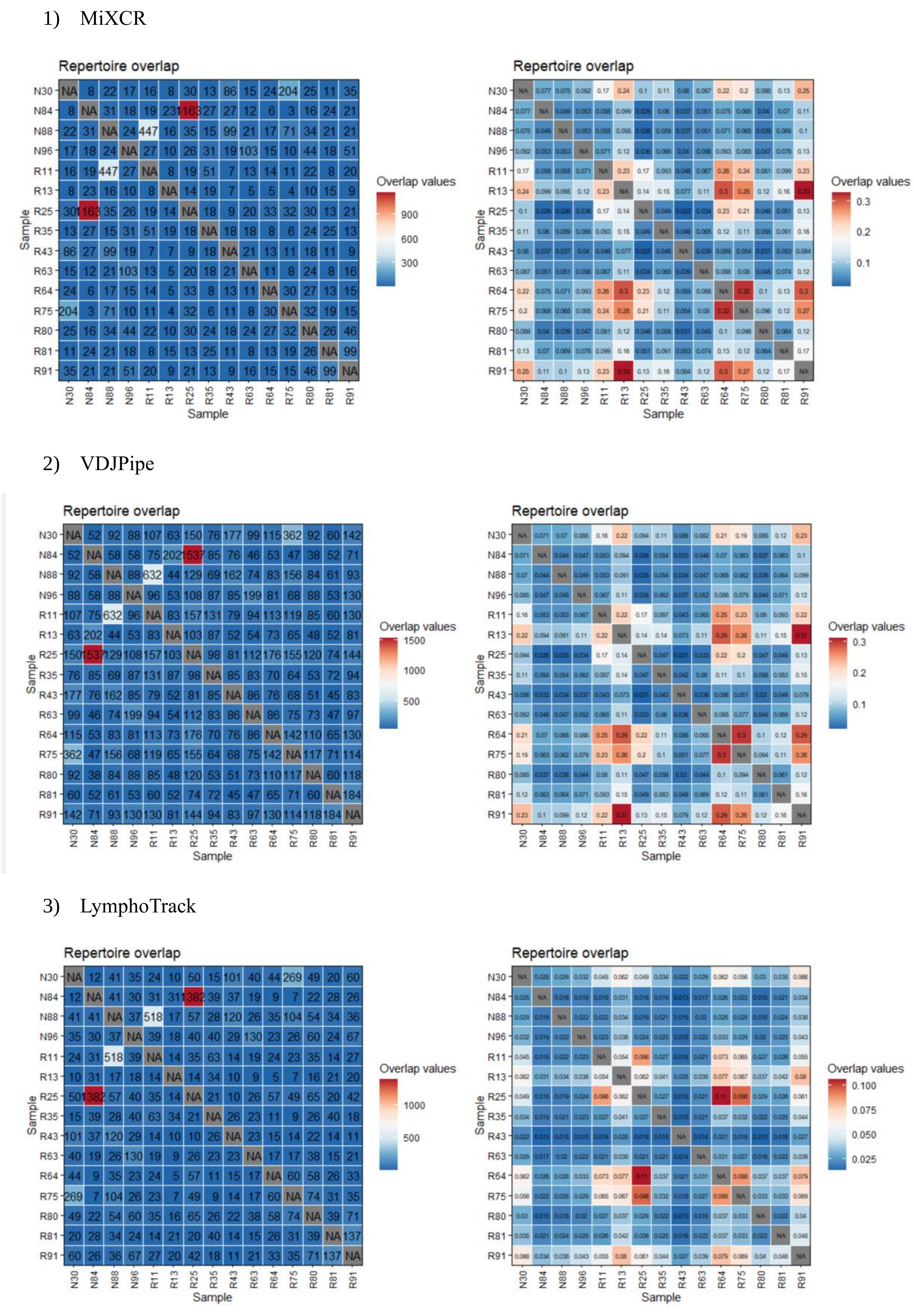


a) MiXCR

b) VDJPipe

c) LymphoTrack

Figure 10. Clonality analyses of the most abundant and the least abundant clonotypes



1) MiXCR

2) VDJPipe

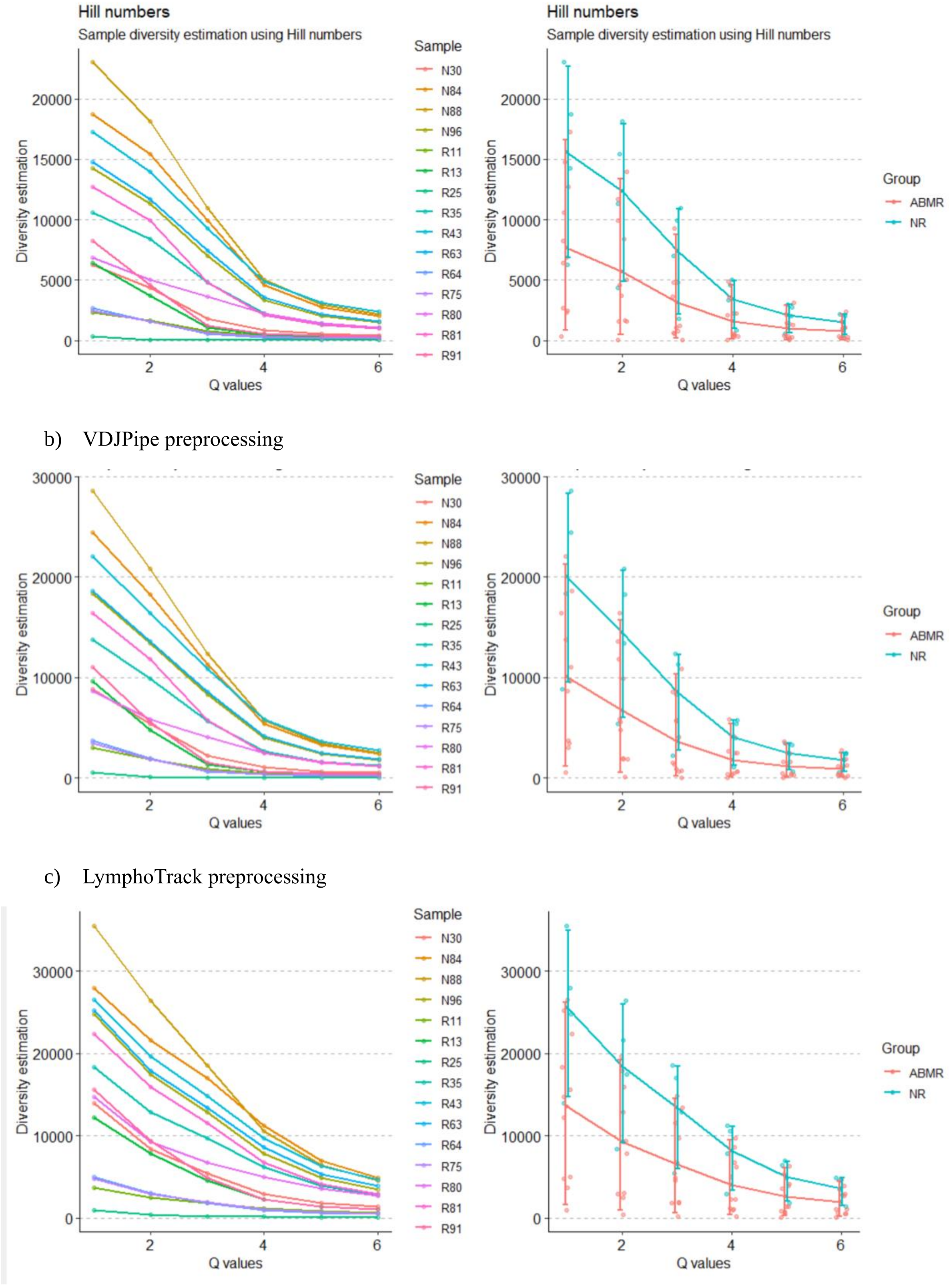
3) LymphoTrack

Figure 11. Repertoire overlap analysis of public clonotypes and the Morisita's overlap index

4. Diversity estimation

The diversity estimation of the repertoire was done using the Chao1, Hill number, True Diversity, the Gini-Simpson and d50 indices. The Hill number, one of the most popularly used diversity index, is shown in Figure 12. The numbers of individual samples are shown in left, and the comparison of sample groups are shown in right. The diversity of NR group was higher than the ABMR groups, in all three preprocessing input data, and this was consistent with the other measures of diversity. The non-parametric estimator of species richness (i.e., the number of clonotypes), Chao1 is shown in Figure 13. The higher Chao1 is too, is indicative of the higher diversity, and the higher Chao1 in NR group compared to ABMR group was observed.

The True diversity, which refers to the number of equally abundant types to reach the average proportional abundance of the types in the dataset, is displayed in Figure 14. The values were consistent across three preprocessing data, and the result of group comparison showed higher diversity in NR groups. The inverse Simpson index which is the effective number of types from weighted arithmetic mean quantifying the proportional abundance of the dataset, is displayed in Figure 15. Lastly, the D50 index, which is calculated by the minimum number of different clonotypes to constitute 50% of the total reads, is shown in Figure 16. In summary of the diversity estimation of the repertoire, the trend observed in all indices used were equal. The D50 index showed the lowest P value for group comparison (0.06), but a statistical significance was not found in any of the indices. Regarding individual samples, the A25 sample showed the lowest diversity using any indices, while N88 appeared to be of the most diverse among the samples. But there were also results from Chao1 and True diversity indicating that A43 was also highly diverse (Figure 13 and 14).

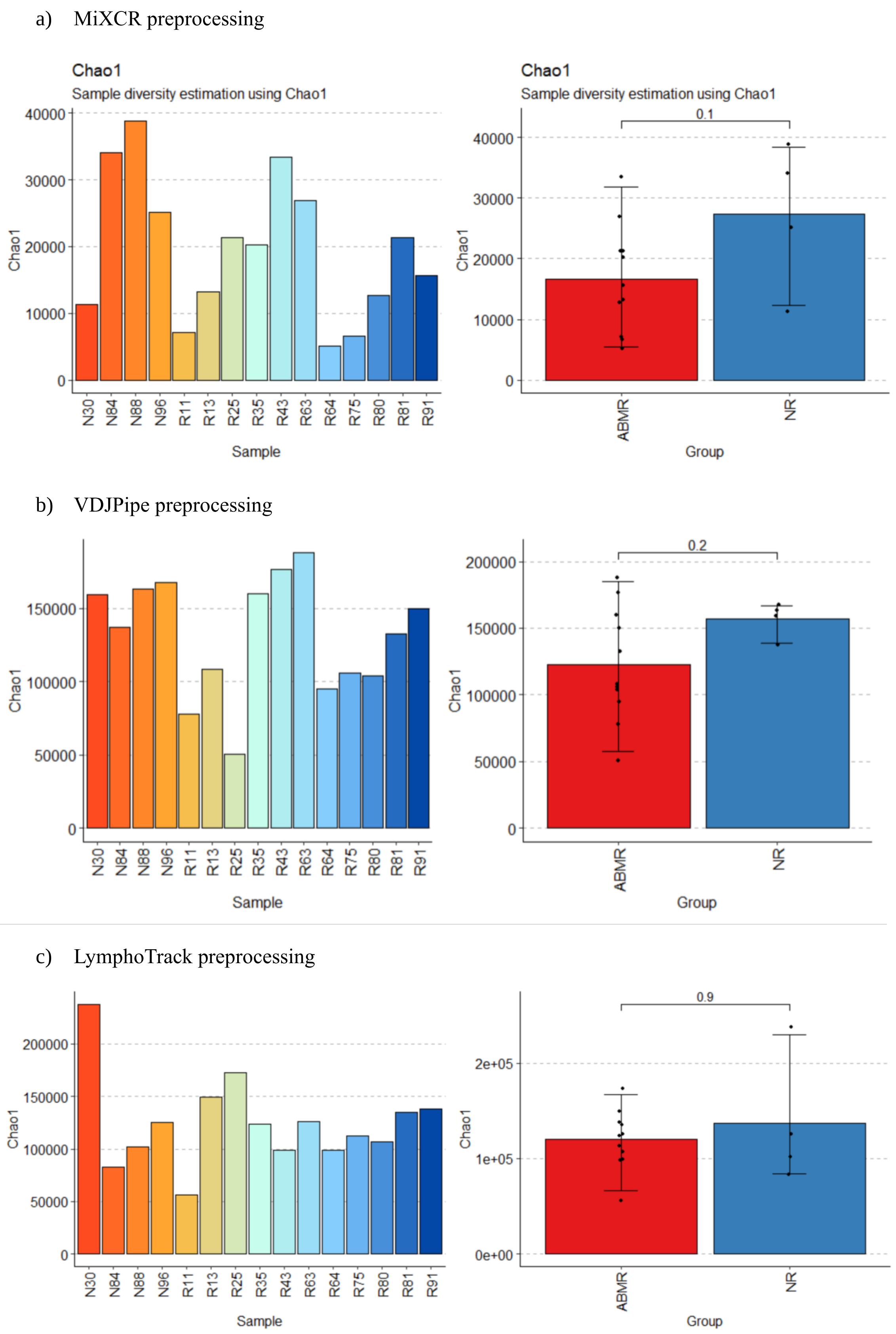


a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

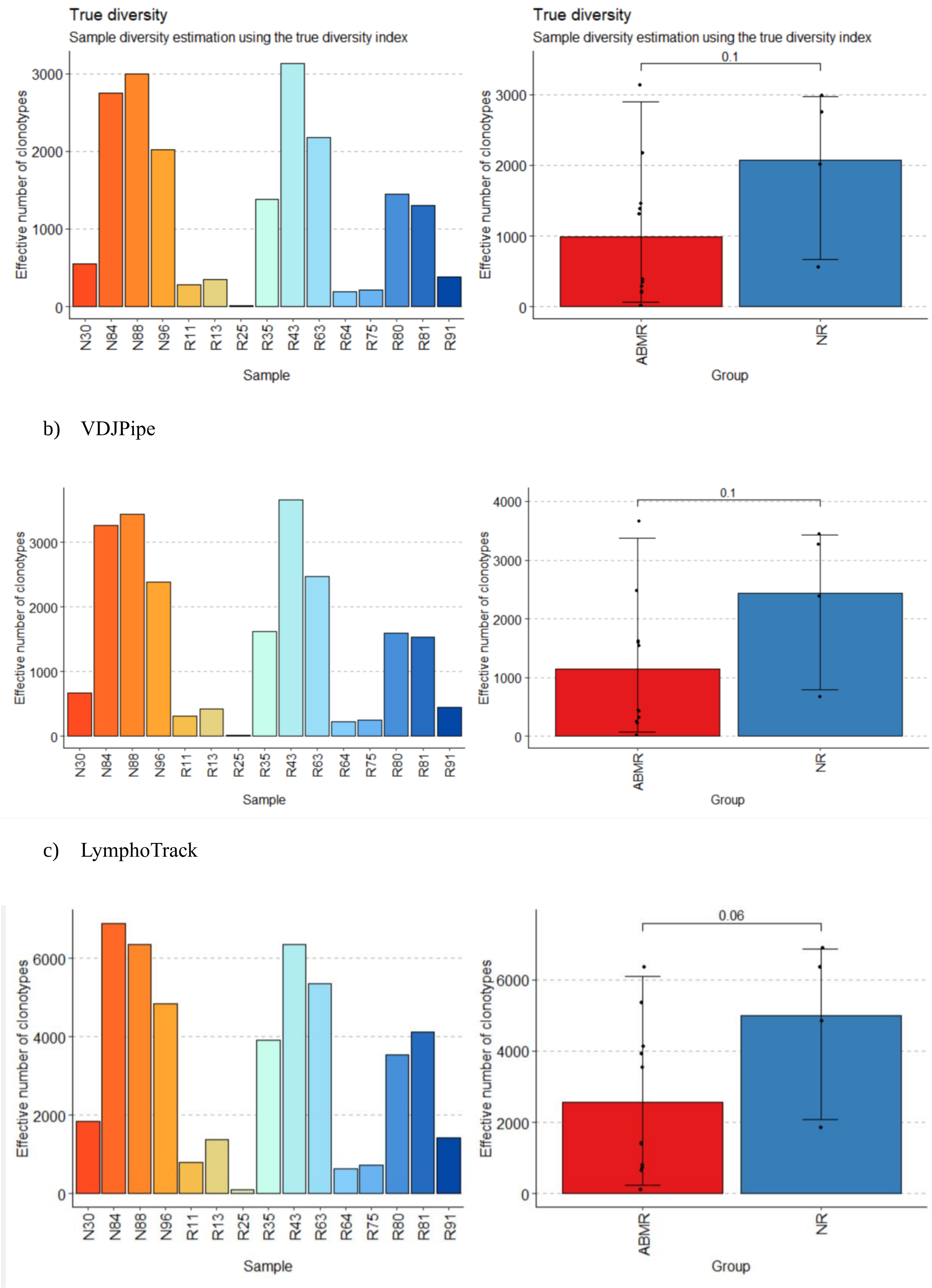
Figure 12. Repertoire diversity estimation - Hill number


a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 13. Repertoire diversity estimation - Chao1

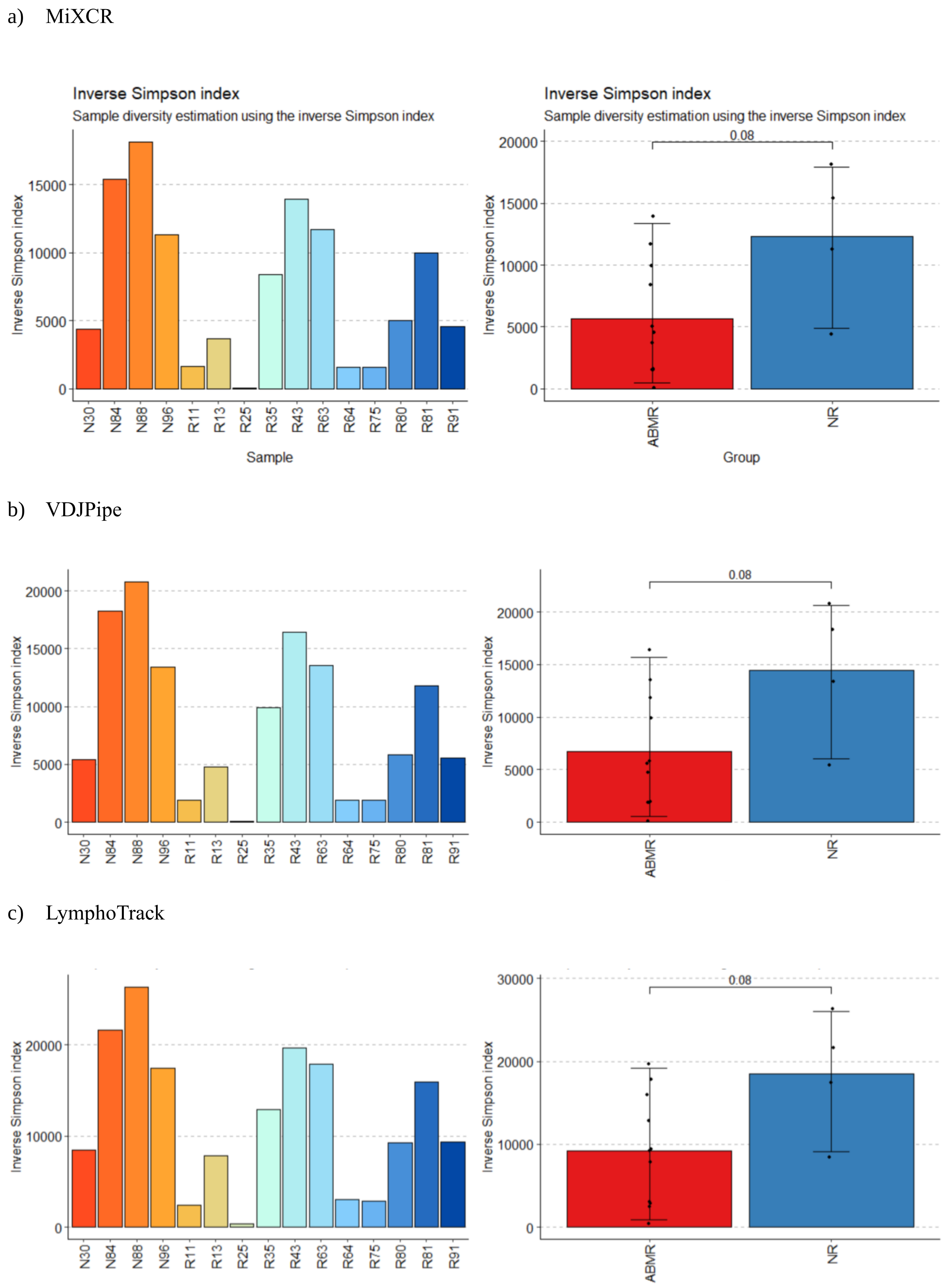


a) MiXCR

b) VDJPipe

c) LymphoTrack

Figure 14. Repertoire diversity estimation - True Diversity (Effective number of types)

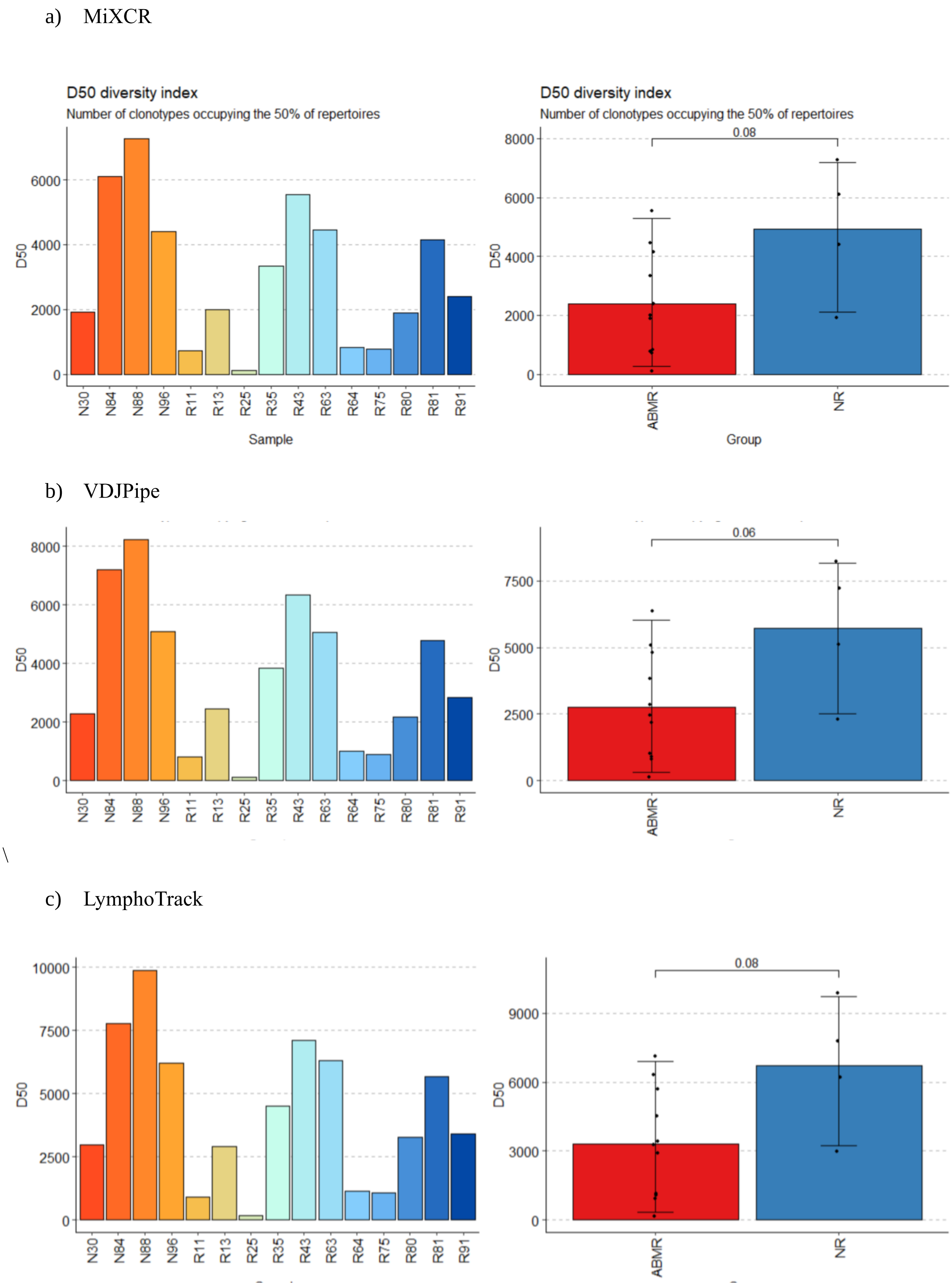


a) MiXCR

b) VDJPipe

c) LymphoTrack

Figure 15. Repertoire diversity estimation - Inverse Simpson index


a) MiXCR

b) VDJPipe

c) LymphoTrack

Figure 16. Repertoire diversity estimation - D50 (minimum number of clonotypes for 50%)

5. CDR3 analysis

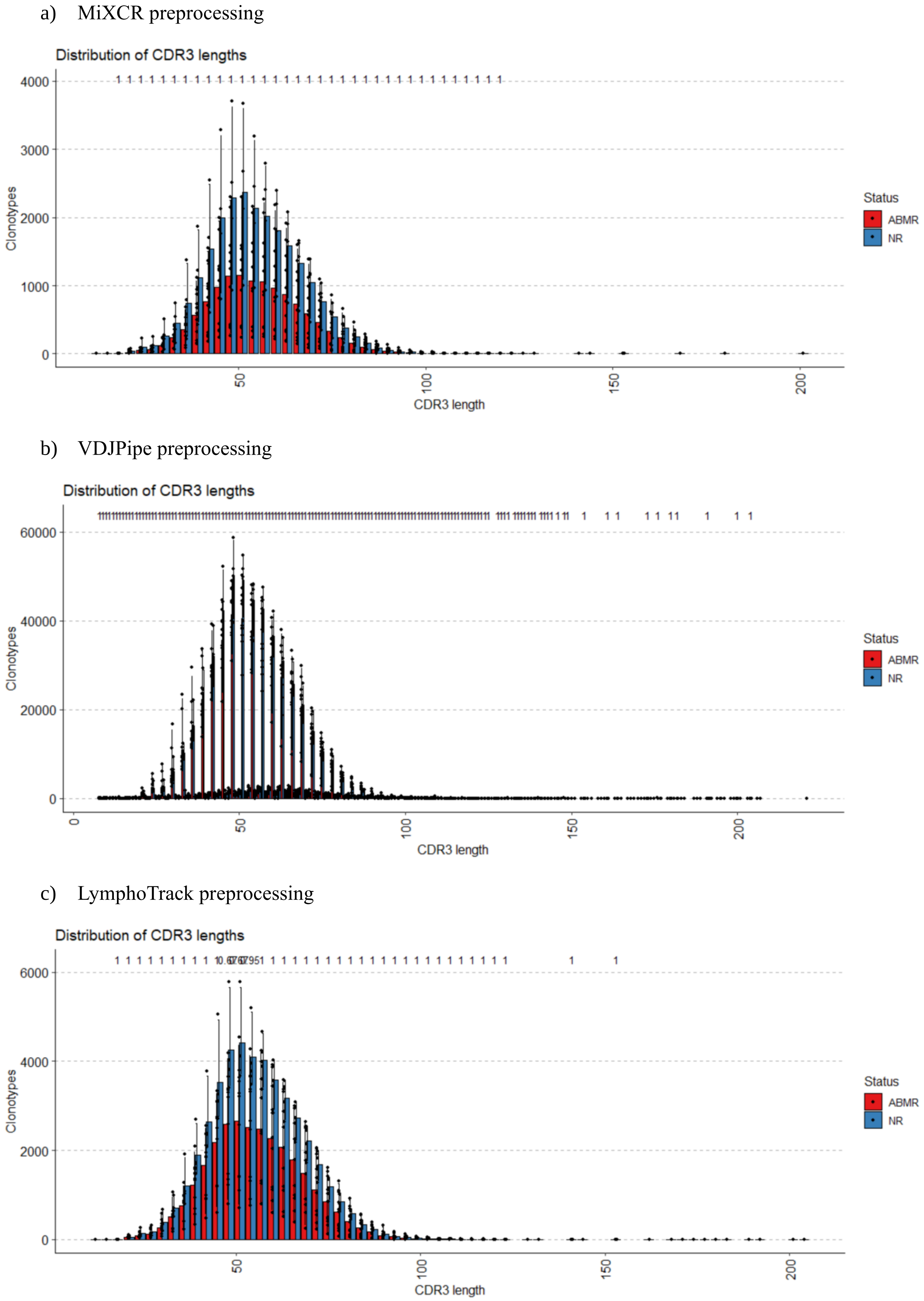
The distribution of CDR3 sequence length was analyzed, using the nucleotide sequences. The plotting of values from each sample results in overplotting within the graph, therefore the group comparison is displayed in Figure 17. Of note, only the coding sequences were analyzed, and the information regarding the coding sequence is imported from the AIRR data format column, 'productive'.

6. V(D)J gene usage analysis

The target gene and the species used for V(D)J gene usage was Homo Sapiens (hs), *IGHV* and *IGHJ* genes. The nomenclature of the genes followed the IMGT nomenclature.[50] The distribution of *IGHV* genes was calculated after being normalized by the individual clonotype counts to avoid sampling bias between samples. By characterizing the samples by usage of specific gene segments and family, a specific gene segment characteristic of sample group was investigated. The results of gene usages for *IGHV* genes and *IGHJ* genes are shown in Figures 18 and 19, respectively.

The *IGHV* gene usage analysis showed interesting results, showing the most common utilization of the *IGHV3-23* gene. Despite the absence of statistical significance, *IGHV3-23* was the only *IGHV* gene with noticeably higher representation in the R group, whereas *IGHV3-11, IGHV3-7* and *IGHV4-39* were slightly higher in NR group (Figure 18). The individual gene usage analysis results are shown in detail in Supplementary Table 3. From the *IGHJ* gene usage analysis, the IGHJ-6 showed higher gene usage in ABMR group of samples, but without statistical significance.

The gene usage was further analyzed by calculating the Jensen-Shannon divergence and the gene usage correlation (Figure 20). These values were calculated as a preprocessing method, and the following analysis of hierarchical clustering displays the structural relationship between the samples (Figure 21). In the case of gene usage, the clustered samples were different between preprocessing data. The MiXCR was the only preprocessing that grouped only ABMR samples by clustering, whereas the other two preprocessing resulted in mixed hierarchical clustering of ABMR and NR group samples (Figure 21).

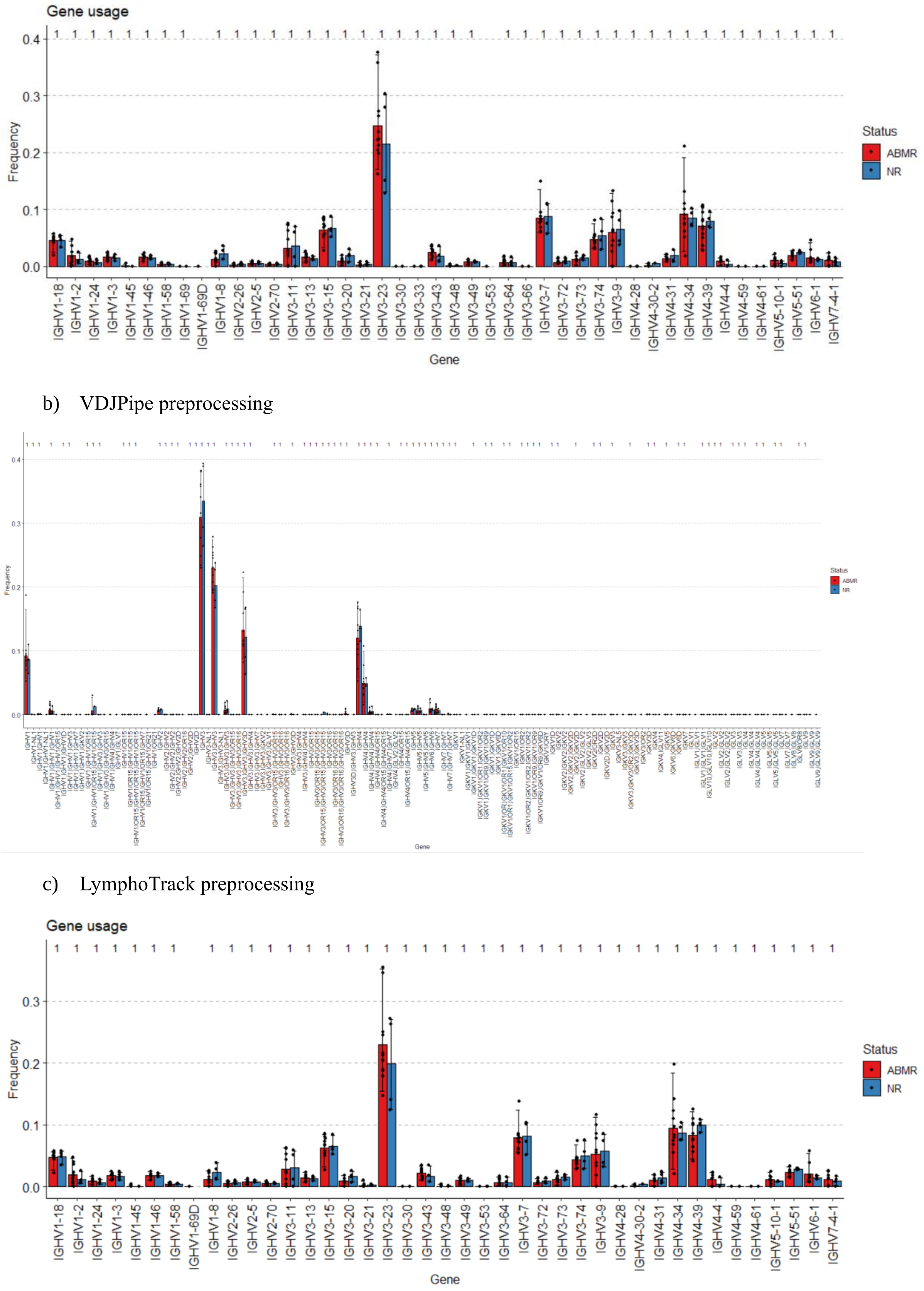

a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 17. Distribution of CDR3 nucleotide lengths

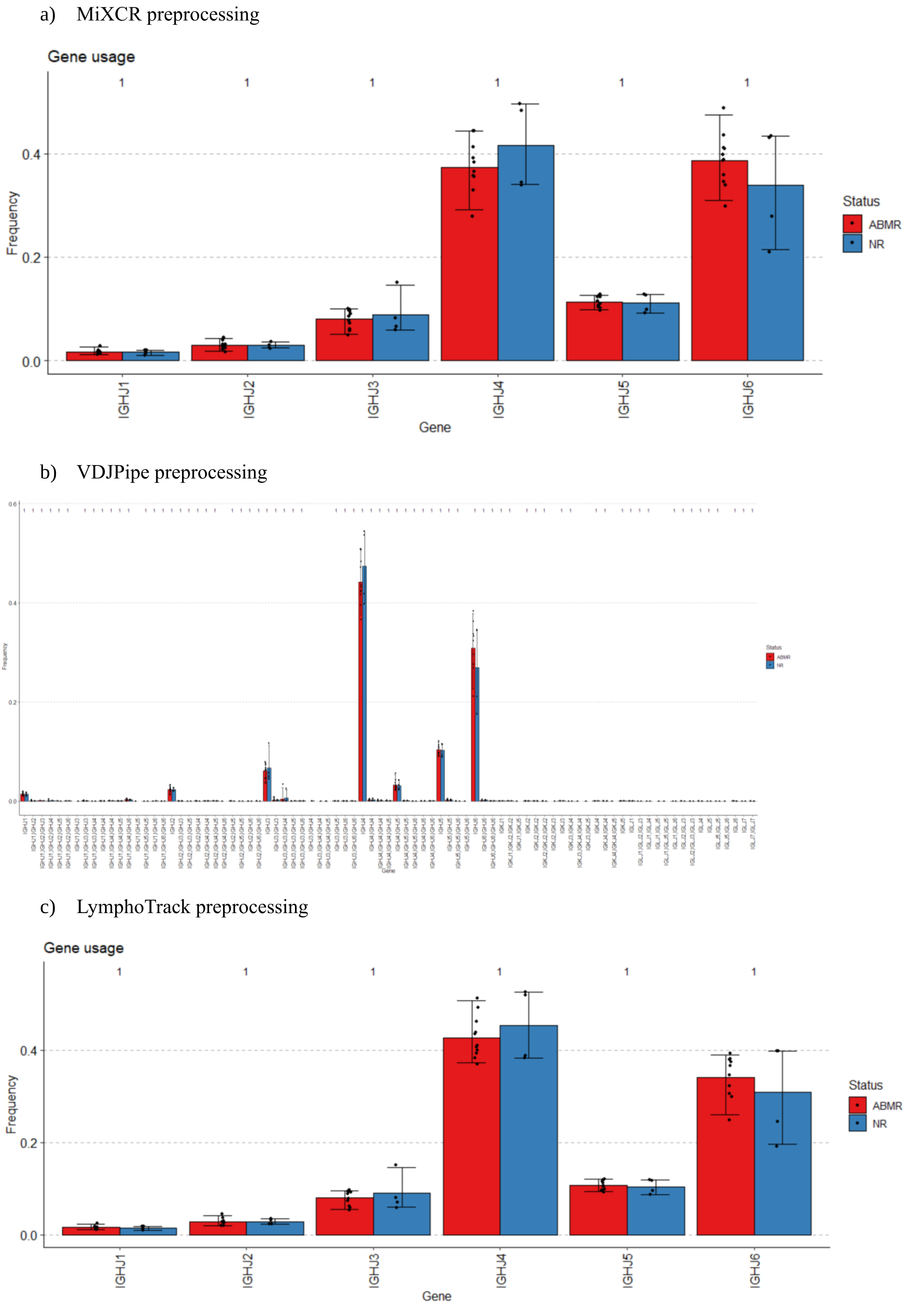
a) MiXCR preprocessing



b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 18. Gene usage analysis - V segment statistics

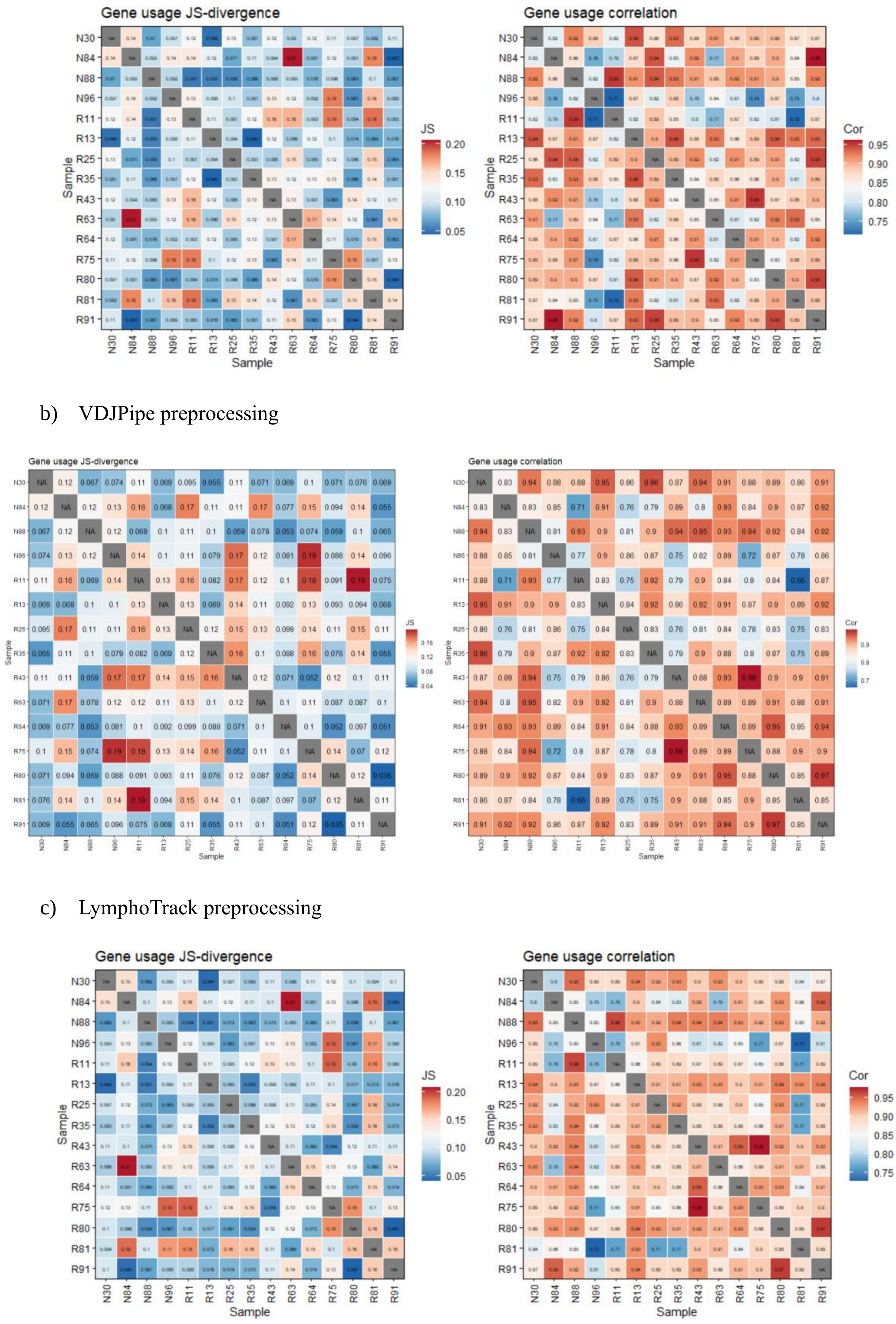


a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 19. Gene usage analysis - J segment statistics

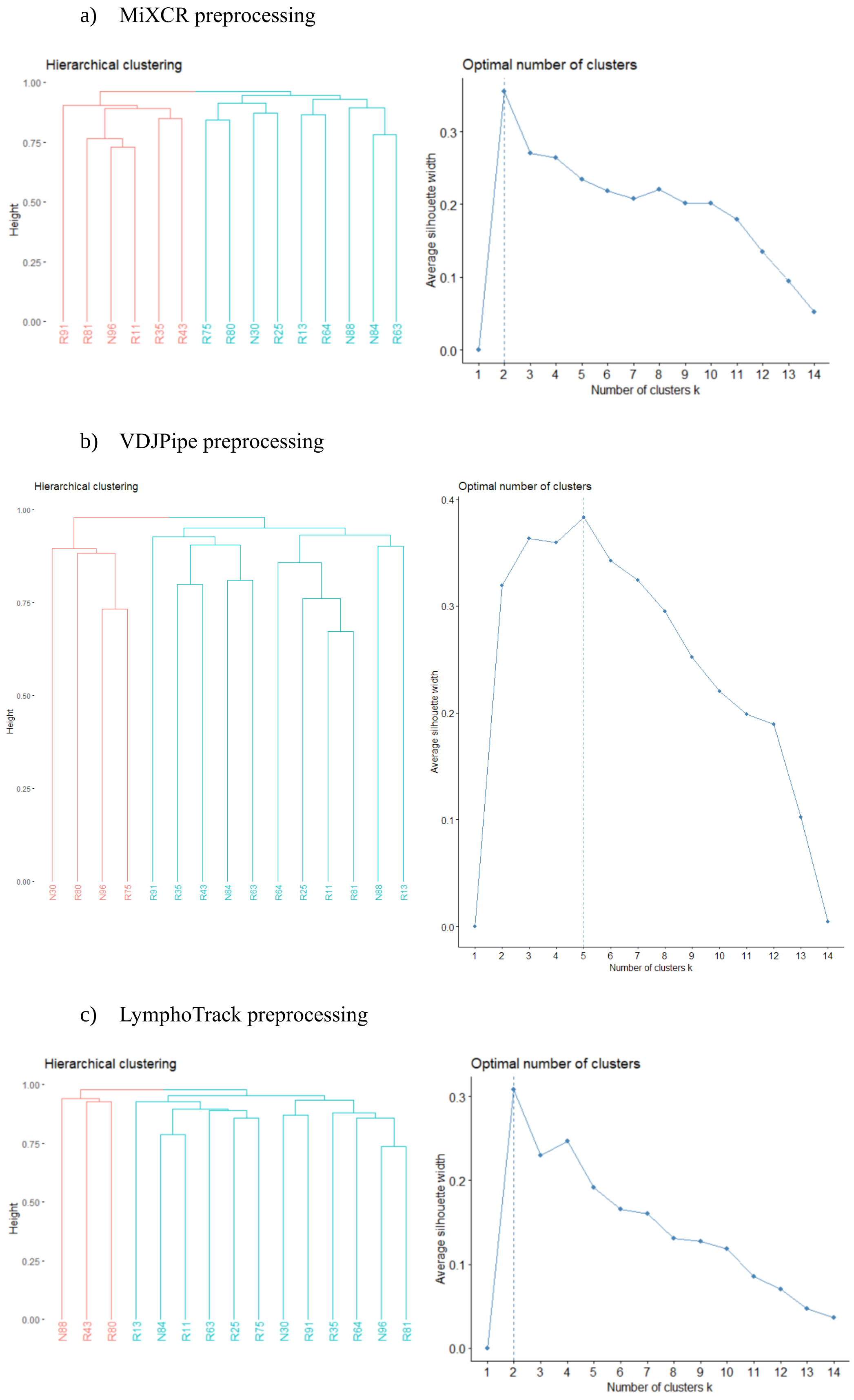


a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 20. Gene usage analysis - JS divergence and usage correlation



a) MiXCR preprocessing

b) VDJPipe preprocessing

c) LymphoTrack preprocessing

Figure 21. Gene usage analysis - hierarchical clustering

7. The reproducibility of repertoire analysis using SRA repository data

The summary of SRA repository data used to assess the reproducibility changes in repertoire according to the read depth, is displayed in Table 7. The MiXCR preprocessing pipeline was used to process SRA repository data, which were paired-end Illumina sequences.

The result of Immunarch analysis is displayed in Figure 22. The total reads and clones from three subjects at their separate samples were variable, while the final number of clonotypes were consistent. And the most representative normal control data, using the largest size of IGH sequence publicly available, showed that the average clonotype number was around 100,000, using total input read of 218,356,368. The highest clonotype number of NR group preprocessed with MiXCR was around 40,554 using total input read of 794,116 (N88). It is anticipated that with higher number of input-read, the higher number of the final clonotypes can be obtained, assuming the normality of the sample.

Sample	description	Total reads	MiXCR			
			Overlapped and aligned	Read used as core	Read used in clonotyping	Final clones
FV1	FV_before_8d	1,248,337	1,085,329	775,570	849,097	46,350
FV2	FV_before_2d	1,057,574	892,113	637,682	695,149	43,737
FV3	FV_before_1h	1,243,523	1058,109	775,787	846,275	50,957
GMC1	GMC_before_8d	1,472,513	1,240,963	891,133	978,631	46,035
GMC2	GMC_before_2d	1,774,057	1,426,720	1,051,114	1,150,558	42,712
GMC3	GMC_before_1h	1,417,170	1,123,734	798,578	878,807	50,554
IB1	IB_before_8d	1,456,496	1,205,880	858,687	940,964	91,915
IB2	IB_before_2d	1,293,820	1,093,866	783,533	857,655	78,258
IB3	IB_before_1h	1,501,664	1,274,653	927,517	1,015,930	72,627

Table 7. Description of normal control samples with replicate results and the sequencing statistics and preprocessing summary

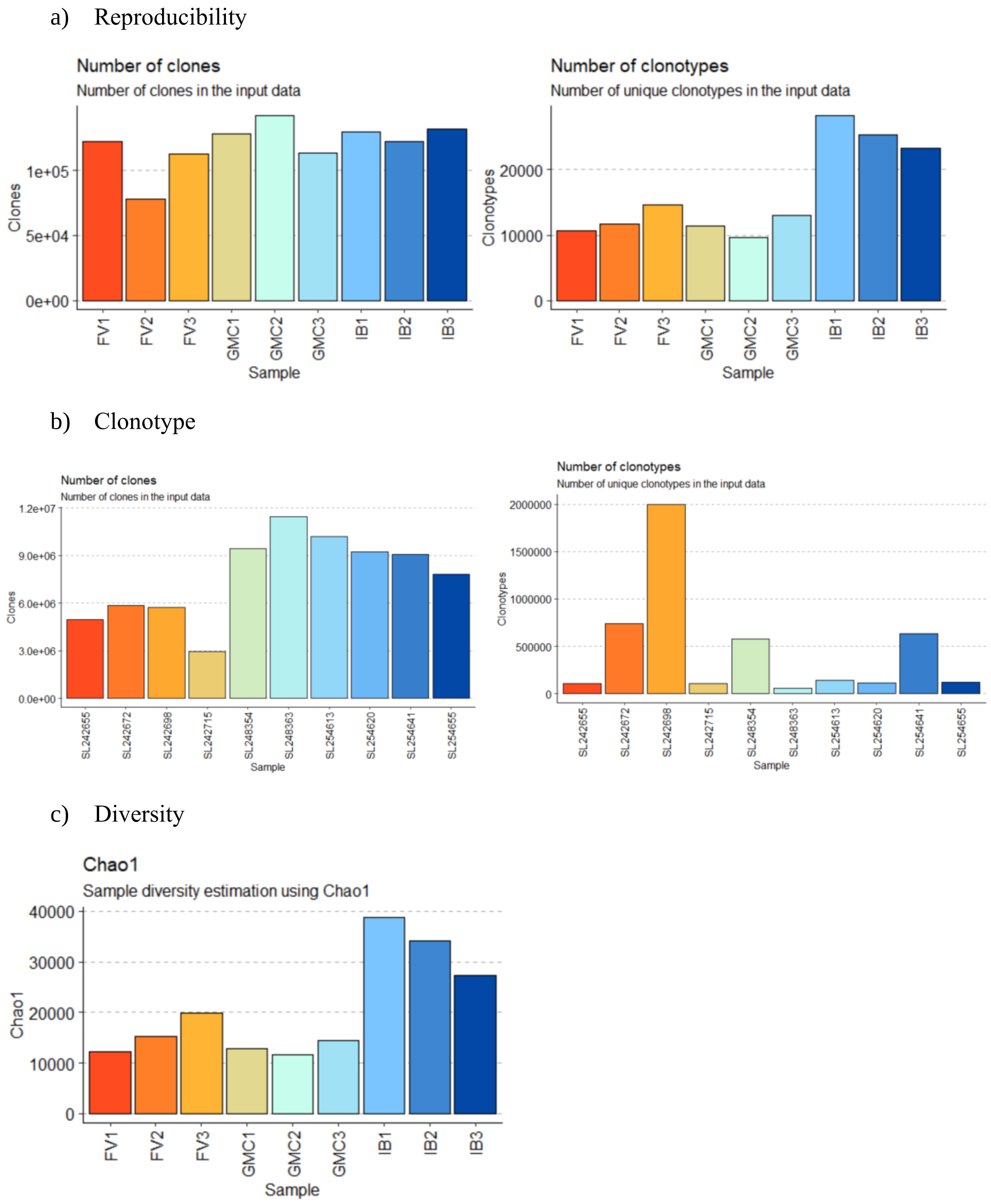


Figure 22. Reproducibility, read-depth associated repertoire and the diversity using SRA sequence data

고찰 Discussion

As the number of researches in the field of AIRR is currently imploding in number and topics, as the number of publication with the query ‘repertoire sequencing’ is continously rising (Figure 7). Molecular genetics has changed the paradigm of how we understand biological phenomenons, and AIRR-seq provides a valuable insight in to clinical conditions where the adaptive immune system plays an important role. The goal of this study was to provide a concise but comprehensive understanding of the currently available AIRR-seq tools, with easier access and completeness, starting from the sequencing raw reads to the repertoire analysis.

Sample selection and the relevant disease categories are important. Here we attempted to explore the key difference of AIRR between those demonstrating ABMR and its non-rejection counterpart. Considering the uncertain clinical significance of borderline TCMR, specimens with the diagnosis of ABMR that accompanied borderline TCMR were categorized as ABMR, in this study. As the clinical course of such cases were more likely to follow the course of classical ABMR, which can be characterized by suboptimal renal function recovery, occurrence of microvascular inflammation, C4d immunopositivity and development of de novo DSA compared to non-borderline TCMR cases.[88] The borderline TCMR category has been continuously assessed by the Banff schema, and in our cases, the borderline TCMR diagnosis implied ambiguous finding due to minor or absent interstitial infiltration (i1 and i0) [89], whereas the diagnosis of ‘suspicious for acute TCMR’ was given otherwise.

Focal segmental glomerulosclerosis (FSGS) is also a common finding from the biopsy specimen of ESRD patients. Despite being a major cause of ESRD, when observed in post-transplant settings, FSGS is considered of lesser clinical significance.[90,91] According to a report, the incidence of FSGS after transplantation reaches up to 30%, either as recurrence of primary FSGS or a post-transplant finding. The current immunosuppression regimen consisting of rituximab, therapeutic plasma exchange and steroids are capable of managing clinical presentations of FSGS, and such findings observed in either of ABMR or NR groups were rather neglected in this study.[92,93] Without a larger, prospective study on management of the post-transplant FSGS, the significance of FSGS in posttransplant graft remains unknown. Despite the sample selection with scrutinizing criteria, the AIRR analyses results were unable to demonstrate statistically significant differences between the groups. The limitation of this study, including the small number of samples (especially the NR group) requires further verification using larger number of samples that distinctively demonstrates disease phenotypes.

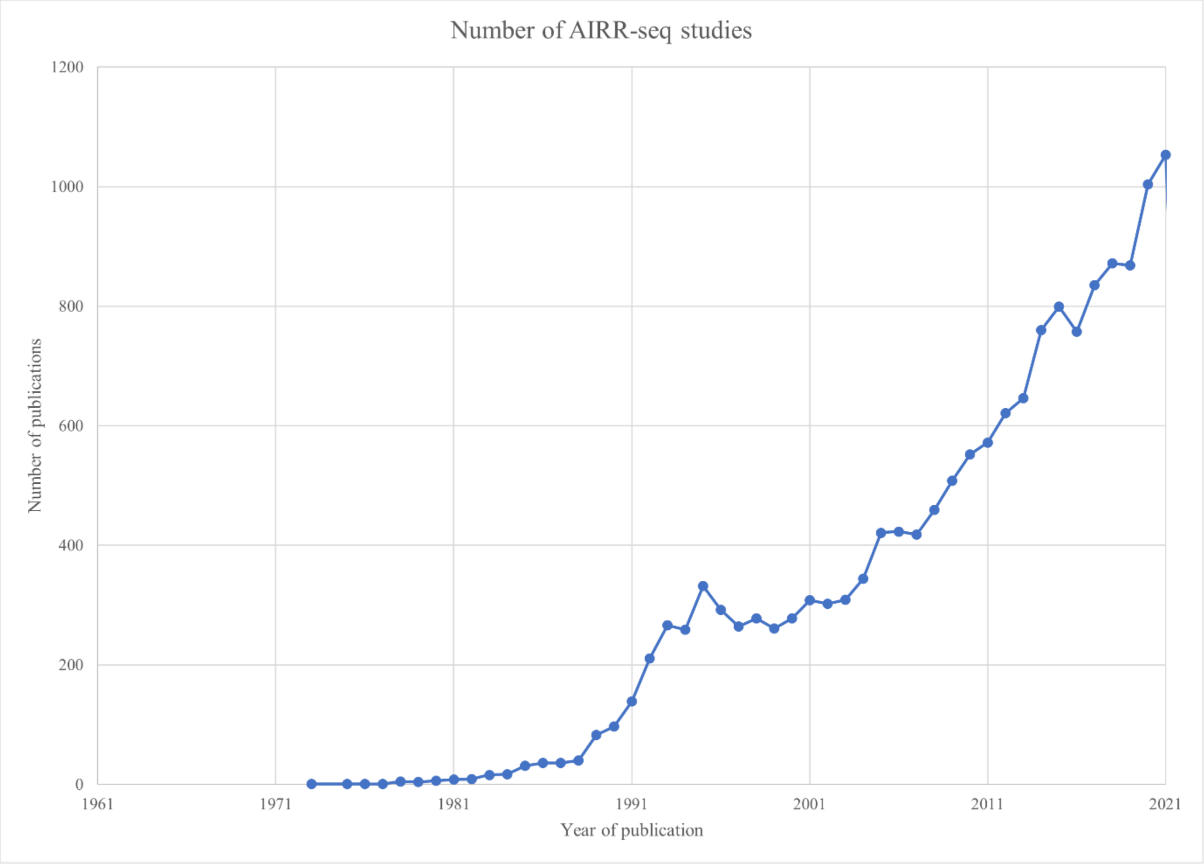


Figure 23. Increasing number of AIRR-seq publications

An important factor to consider in sequencing-based immunoglobulin studies, is the reproducibility of the AIRR testing results. Standardization of such PCR-based immunoglobulin including both TCR and IG (heavy and light chains) have been made by the EuroClonality (BIOMED-2) consortium, in scope of providing the reproducibility and inter-exchangeability of the studies.[59,94] The LymphoTrack IGH assay kit used in this study was developed in parallel with the BIOMED-2 protocol, that ensures reproducible pre- and post-analytical results. Another is the error rate, of which a more stringent criteria is required considering the astronomical diversity of AIRR. The Illumina platform used in this study offers the lowest error rate within the industry of 1.2%.[95,96] The choice of NGS platform was made.

The input DNA volume was determined with several assumptions. The LymphoTrack assay when used for MRD monitoring provides a range of sensitivity between 0.01% to 0.0001% (10^{-4}). The interpretation criteria of LymphoTrack uses a 5% cutoff for MRD to determine presence of clonality for samples with total read between 10,000 and 20,000. For samples with reads over 20,000, 2.5% of total merged sequence indicates presence of clonality. Determination of clonality can vary among studies, but in general the clonality is usually non-evaluable for reads below 10,000 and requires duplicates for reads between 10,000 ~ 20,000.[97] Therefore we calculated with the following assumptions to achieve a more sensitive analysis of clonality. The recommended read depth for clonality is between 20,000~50,000, and >100,000 preferably.

The human diploid cells including lymphocytes contains about 6.5 pg (4-7) per each cells.[98–100] Under the assumption that all DNA is sequenced, an input DNA of 1 ug is equivalent to about 150,000 (153,846) cells, which includes about 4,500 B cells (Lymphocytes account for average of 30% of the white blood cells (WBC) when the differential count is normal). The number of B cells equivalent to 2 ug of input DNA is therefore approximately 10,000 B cells. This number is still only a fraction to the total estimated number of B cells per human which is at about 10^{10}~10^{11} cells.[4] Although previous studies have used much less amount of input DNA ranging from 25 ng to 500 ng, studies differ by the types of input (DNA or RNA) and the depth of reads.[101,102]

Input DNA volume was set with the goal of avoiding NGS mistakes that often produce misleading repertoire results. Oversampling, use of Unique Molecular Identifier (UMI) and computational filtering for error correction are methods that can be used to avoid NGS mistakes, and we approached with the oversampling method in this study. The general rule of 5 – 10 times more reads than the number of input cells is an agreeable measure of calculation (adapted from AIRR-community webinar of April 6, 2021).

The CDR3 length analysis are usually done by either nucleotide length or the amino acid lengths. The initial analysis was done using the amino acid parameter, where no noticeable difference between sample groups was found (data not shown). As some previous literature suggested that use of amino acid based Hamming distance had significantly lower sensitivity, the analysis was conducted using nucleotide lengths.[103,104] Unfortunately this analysis also failed to demonstrate noticeable differences between groups. The definition of clonotype can differ between studies, as 100% identical CDR3 sequence was suggested by Briney et al., whereas lower 80% sequence identity was used to cluster clonotypes in another study.[5,105] The amount of N nucleotides in the V- or J- segment could have affected the otherwise same clonotypes into different clonotypes, leading to a more diverse and indifferent results shown in this study.[106] Investigation by separate clonotyping tools that utilize different clustering algorithms is required, but the use of such tool e.g. ImmuneDB was not implemented in this study. As the comparison study of MiXCR and ImmuneDB described that MiXCR by its more strict clonotyping algorithm, caused separation of clonotypes that may originated from the same clone.[106] The process of MiXCR clustering the clones into clonotype described in detail (https://mixcr.readthedocs.io/en/develop/assemble.html#) shows that fuzzy matches of clonotypes are organized into hierarchical trees, and only their head are considered as final clones. It is possible to adjust the clustering strategy such as the number of cluster layers, maximum number of N nucleotides and probability factor of single nucleotide mutations, but these tweaks within the preprocessing pipeline was outside the scope of this study.

The gene usage analysis is an important component of the repertoire clonality. The *IGHV* gene usage found *IGHV3-23* as the most abundant gene across both groups and all samples. The finding is in consistence with the previous reports that *IGHV3-23* gene was among the most commonly utilized gene.[55,75,107] The findings higher IGHV3-23 usage in most R group samples were possibly associated with the gene ontology of IGHV3-23, such as the B cell receptor signaling pathway, classical complement activation pathway, positive regulation of B cell activation and etc. On the other hand, the previous literatures describing IGHV3-23 as the most used *IGHV* gene suggests that such interpretations may require caution.[108] Gene usage analysis results can be affected by the sources and subsets of B cells and also the locations such as peripheral blood or tissue where it is obtained.[107] Future studies using larger cohort with more distinctive differences in repertoire is likely to demonstrate more drastic differences in gene usage or discover increased expression of selective *IGHV* with statistical significance..

The LymphoTrack preprocessing produced similar results to MiXCR preprocessing, although the results were obviously not identical. The reason for difference originates from the sequence merging process of LymphoTrack software, which uses exact sequence match. According to the manufacture's guideline provided with the software, the exact sequences and the similar sequences up to two mismatched nucleotides are merged, producing the combined unique reads as outputs. The LymphoTrack Software implements a very stringent criteria for the identification of CDR3 regions, which is sometimes considered overly strict. Presence of frameshift mutation or mutations to the anchor amino acid will result in no CDR3 sequences identified using the LymphoTrack Software (Invivoscribe, personal communication, May 12, 2022). Therefore, the combine unique reads generated by LymphoTrack preprocessing, generates different number of clonotypes when compared to MiXCR preprocessing data.

It is important to mention the time factor required for data analysis. The MiXCR tool has been previously mentioned of its rapid capability to handle large sequence data, without specific limitations of input.[61] It is to our experience that MiXCR using the built-in library was able to process the input sequences in expedite, whereas the VDJPipe process performed online, required much more time in addition to the uploading of large sequence data. The VDJServer provides access to the high-performance computing (HPC) at the Texas Advanced Computing Center, which easily outpowers any personal computing system currently available in the consumer market. However, the access and quota to individual users are limited, and can be unpredictable (e.g., maintenance schedule, unexpected errors of server origin and possible override due to overuse). These were the limitations of cloud-based analysis portal that we experienced during the analysis of our data. On the other hand, the LymphoTrack software developed using Java, was also fast for processing input sequences, although this may depend on the systems specifications running the software.

By comparing the preprocessing functions, most of the common and critical preprocessing steps for NGS data was included within each tool. However, the naming of processes varied, and some obvious functions were not specifically mentioned (i.e., quality filtering of LymphoTrack not mentioned). Barcode de-multiplexing and UMI identification is an important aspect of preprocessing, although our study data did not implement UMI for the generation of sequencing data, future studies with different samples will benefit from these features. Merging of paired-end data was done with agreeable percentages in all three preprocessing tools, of which the general sequencing quality was of more importance.

From our evaluation of different pipelines, the computational consideration was an important aspect of the study. As the functions becomes more complex and resource dependent, the burden of computing resource increases almost exponentially when the size of input sequence data becomes larger. The most user and beginner friendly

R language and its RStudio are fundamentally a statistics analysis package, which allocates its memory use directly on the physical memory. The requirements for hardware memory becomes higher in correlation with the size of the sequence data, in which the sequence data of 1 million generates an object size of 800~900 MB (megabytes).[109] Use of parallel processing or virtual memory optimization is perhaps outside the scope of this study, and the sequence data of 15 samples (11,501,246 reads in total) would require at least 8~9 GB of memory. The requirements may vary according to the specific details such as number of clonotypes, sequence analysis level of gene/segment/family, but here we provide a baseline idea of computational requirements for AIRR-seq.

Use of AIRR-seq to explore the repertoire of adaptive immune system begins with the concept that sequence based interpretation can provide antigen-antibody interactions. However, this assumption of similarity in sequences to result in identical or similar antibody response has been proved to be not always accurate.[20,110,111] The high sequence similarity may correspond to structural resemblance, but when it comes down to the the conformation of functional protein interfaces, the similarity of sequences alone needs caution (Figure 24). The three-dimensional structure of two proteins show high similarity of 0.60Å (left), despite the different amino acids sequences (protein database, PDB accession ID: 3PHO, 3QUM). However, the almost matching sequences on the right, generate protein structures with different three-dimensional conformations which measures at 4.15Å (5ILG and 5ILC). The functional aspect of antibodies originates from the sequences, further structured in to eplets, epitopes, CDRs and furthermore, but the interpretation of antibody reactivity (function) would require caution.

Retrospective use and the limited number of samples are some of the limitations of this study. The NR group in this study were included as normal controls, but the indications of kidney biopsies were not always protocol biopsies in which indications of possible rejections were present (Table 8). Due to the stringent criteria for R group, biopsies in R group were indication biopsies where need for evaluation was obvious, such as presence of DSA, urinary symptoms, and imaging findings suggestive of rejection. Also, the small number of NR group was a statistically limiting factor, not to mention the repertoire characteristics of low diversity indices somewhat suggesting presence of rejections. The comparison made between groups were neither statistically significant, although we suspect prospective paired-sample studies using larger number of samples will likely show statistically significant differences in the immune repertoire.

Another limitation of the study is focusing only on the heavy chain data, and the absence of paired-light chain data. The pairing of VH/VL sequencing data supposedly can provide more information on the affinity and antigenic specificity.[112] The definition of clonality of course includes the pairing of VH-VL, however the limited variability of light chain sequences often act as a restriction to evaluation of clonotypes on paired data.[113] The LymphoTrack IGK - MiSeq is available as a separate assay, but was not included for this study.

Further application of the data obtained through AIRR-seq is using as an input to machine-learning models to further investigate the difference between desired study groups. Due to the high diversity of immune repertoires, the comparison of two large datasets is often impossible without the proper tools developed for the specific purpose. For this, a well-structured learning dataset is necessary, and application of the pipeline used in this study is an applicable way of generating such dataset. Although this was not applied in this study, machine-learning model is a promising method of discovering new findings from AIRR-seq data.

Sample	Indication for renal biopsy
R11	Hematuria and imaging findings (US)
R13	DSA: A33(9607)
R25	Graft dysfunction
R35	Imaging findings
R43	DSA: B60(3671), B64(1997), DR9(1346), DQ9(3644)
R63	Severe hydronephrosis
R64	DSA: DQ5(4637)
R75	Graft dysfunction
R80	Thrombotic microangiopathy
R81	Low grade rejection
R91	DSA: DQ5(10023)
N30	Protocol biopsy
N84	Protocol biopsy
N88	Elevated creatinine
N96	Symptomatic fluid collection

Table 8. Indications for the kidney biopsies performed.

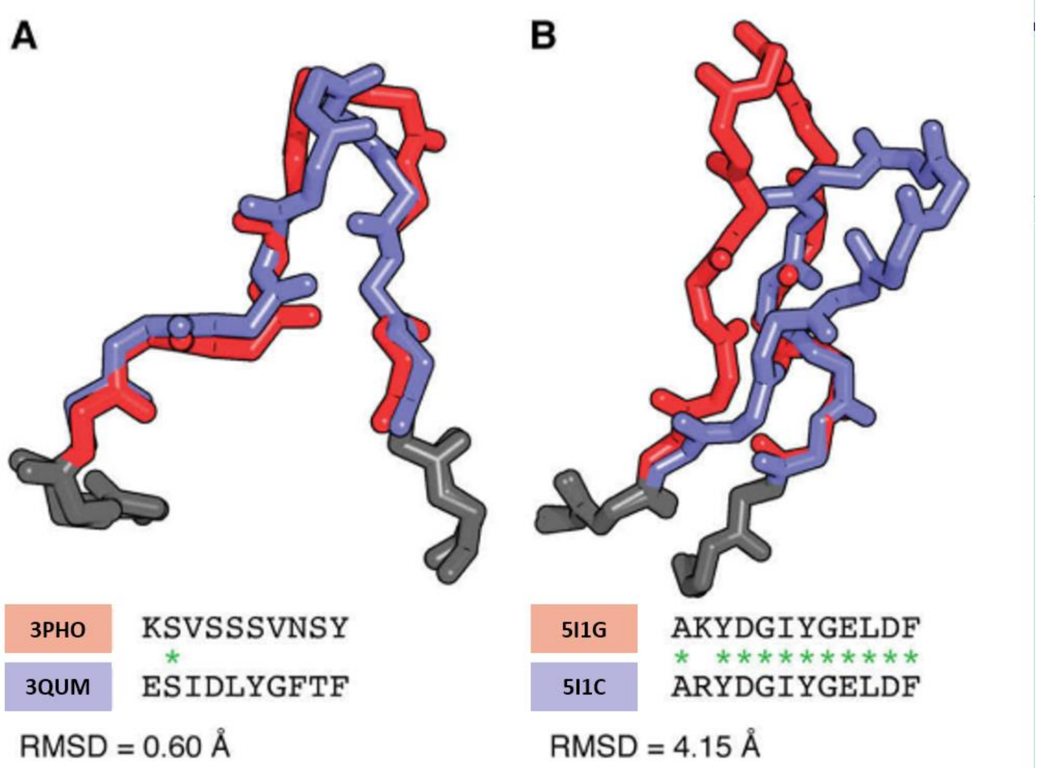


Figure 24. An example of amino acid sequence and conformational structure similarity despite sequence differences (left) and structural difference despite high sequence similarity (right).

결론 Conclusion

In this study, we evaluated the use of the LymphoTrack assay kit and application of three different preprocessing tools to construct an end-to-end repertoire analysis pipeline. The results of the repertoire comparison between ABMR samples and NR samples showed differences in clonality, diversity and the gene usages, despite the limited statistical significance. The application of AIRR-seq for KT proved to be a sensitive and robust method for analysis and comparison of the repertoire between samples and groups. With analytical reproducibility and a measure of correlation between the read depth and clonotypes, AIRR-seq can be used as a reliable indicator of the AIRR.

While the vast variety of the pipelines used for bioinformatics analysis can be problematic for comparison between studies of their results, use of *Immunarch* R package was useful to overcome the hurdles of shell scripting, code writing and standardized method of analysis, using the sequence data generated by three preprocessing pipelines. The VDJServer was also useful tool providing integrated pipeline of tools on cloud-based webserver. Through comparison with most popular preprocessing tools of MiXCR and VDJPipe, LymphoTrack preprocessing was able to provide comparable results with MiXCR for features of the repertoire. Considering the availability of commercialized assay and the widely used platform, LymphoTrack assay in addition to its MRD application can be used for the repertoire analysis of more various purposes. Use of such commercially available assay kit has the strengths of improved reproducibility, accessibility, and standardization.

The difference of results originated from the algorithms and their ability to handle input sequences differently. It is up to the users and the researchers to design their studies accordingly to the characteristics of the AIRR-seq tools that are currently available, and to comprehensively understand their utilities and applications. With the continuing efforts of the AIRR Community and the dedicated researchers, the field of AIRR-seq draws more attention to researchers looking for fundamental mechanism that underlies in the adaptive immune system. The use of AIRR-seq has promising applications in disease diagnosis, monitoring of prognosis, treatment response and possible drug discovery and biomarker developments. The value of this study adds to the very limited number of AIRR-seq studies targeting the BCR in cohort of KT. We conclude that by using commercialized assay kits and AIRR-compliant tools can provide valuable repertoire information in transplant patients, with future applications for novel biomarkers and potential therapeutic target discovery.

참고문헌 References

1. Abbas AK, Lichtman AH, Pillai S. *Cellular and Molecular Immunology*. Elsevier Health Sciences; 2017.

2. Murphy KM, Weaver C. *Janeway's Immunobiology*. Garland Science/Taylor & Francis Group, LLC; 2017.

3. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and PW. *The Generation of Antibody Diversity*. Garland Science/Taylor & Francis Group, LLC; 2002.

4. Ganusov V V, De Boer RJ. Do most lymphocytes in humans really reside in the gut? *Trends Immunol*. 2007;28:514-518.

5. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*. 2019;566:393-397.

6. Schroeder HW. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev Comp Immunol*. 2006;30:119-135.

7. Benichou G, Gonzalez B, Marino J, Ayasoufi K, Valujskikh A. Role of Memory T Cells in Allograft Rejection and Tolerance. *Front Immunol*. 2017;8.

8. Morbach H, Eichhorn EM, Liese JG, Girschick HJ. Reference values for B cell subpopulations from infancy to adulthood. *Clin Exp Immunol*. 2010;162:271-279.

9. Bianconi E, Piovesan A, Facchin F, et al. An estimation of the number of cells in the human body. *Ann Hum Biol*. 2013;40:463-471.

10. Ehrlich P. Experimentelle Untersuchungen über Immunität. I. Ueber Ricin. *Dtsch Medizinische Wochenschrift*. 1891;17:976-979.

11. Owen J, Punt J, Stranford S, Jones P. *Kuby Immunology*. Macmillan Learning; 2018.

12. Odegard VH, Schatz DG. Targeting of somatic hypermutation. *Nat Rev Immunol*. 2006;6:573-583.

13. Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Crowe Jr. JE, Boraschi D, Rappuoli R, eds. *Microbiol Spectr*. 2014;2.

14. Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology*. 2020;160:233-247.

15. Vander Heiden JA, Marquez S, Marthandan N, et al. AIRR Community Standardized Representations for Annotated Immune Repertoires. *Front Immunol*. 2018;9:2206.

16. H Blüthmann, P Kisielow, Y Uematsu, M Malissen, P Krimpenfort, A Berns, H von Boehmer MS. T-cell-specific deletion of T-cell receptor transgenes allows functional rearrangement of endogenous alpha- and beta-genes. *Nature*. 1988;334:156-159.

17. Kent SC, Chen Y, Bregoli L, et al. Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope. *Nature*. 2005;435:224-228.

18. Yassai MB, Naumov YN, Naumova EN, Gorski J. A clonotype nomenclature for T cell receptors. *Immunogenetics*. 2009;61:493-502.

19. Shin S, El-Diwany R, Schaffert S, et al. Antigen Recognition Determinants of γδ T Cell Receptors. *Science (80-)*. 2005;308:252-255.

20. Marks C, Deane CM. How repertoire data are changing antibody science. *J Biol Chem*. 2020;295:9823-9837.

21. Kuroda D, Shirai H, Jacobson MP, Nakamura H. Computer-aided antibody design. *Protein Eng Des Sel*. 2012;25:507-522.

22. Setliff I, Mcdonnell WJ, Raju N, et al. Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection. *Cell Host Microbe*. 2018;23:845-854.e6.

23. Matsuda F, Ishii K, Bourvagnet P, et al. The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. *J Exp Med*. 1998;188:2151-2162.

24. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061-1073.

25. Lefranc M-P, Giudicelli V, Ginestoux C, et al. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol*. 2004;4:17-29.

26. Lefranc M-P. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of

Immunoinformatics. *Front Immunol*. 2014;5.

27. Chatanaka MK, Ulndreaj A, Sohaei D, Prassas I. Immunoinformatics: Pushing the boundaries of immunology research and medicine. *ImmunoInformatics*. March 2022;5:100007.

28. Schultheiß C, Paschold L, Simnica D, et al. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity*. 2020;53:442-455.e4.

29. Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. *Nat Immunol*. 2014;15:118-127.

30. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med*. 2015;7.

31. Robinson WH. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. *Nat Rev Rheumatol*. 2015;11:171-182.

32. Arnaout RA, Prak ETL, Schwab N, et al. The Future of Blood Testing Is the Immunome. *Front Immunol*. 2021;12:228.

33. Trück J, Eugster A, Barennes P, et al. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *Elife*. 2021;10.

34. Breden F, Luning Prak ET, Peters B, et al. Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front Immunol*. 2017;8:1418.

35. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50:1-14.

36. Greiff V, Yaari G, Cowell LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr Opin Syst Biol*. 2020;24:109-119.

37. Teraguchi S, Saputri DS, Llamas-Covarrubias MA, et al. Methods for sequence and structural analysis of B and T cell receptor repertoires. *Comput Struct Biotechnol J*. 2020;18:2000-2011.

38. Liu X, Wu J. History, applications, and challenges of immune repertoire research. *Cell Biol Toxicol*. 2018;34:441-457.

39. Lees WD, Shepherd AJ. Utilities for High-Throughput Analysis of B-Cell Clonal Lineages. *J Immunol Res*. 2015;2015:1-9.

40. Li S, Lefranc M-P, Miles JJ, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun*. 2013;4.

41. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013;41:W34-W40.

42. Gaëta BA, Malming HR, Jackson KJL, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*. 2007;23:1580-1587.

43. Souto-Carneiro MM, Longo NS, Russ DE, Sun H-W, Lipsky PE. Characterization of the Human Ig Heavy Chain Antigen Binding Complementarity Determining Region 3 Using a Newly Developed Software Algorithm, JOINSOLVER. *J Immunol*. 2004;172:6790-6802.

44. Bhattacharya S, Dunn P, Thomas CG, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data*. 2018;5:180015.

45. Ghraichy M, Galson JD, Kelly DF, Trück J. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. *Immunology*. 2018;153:145-160.

46. Smakaj E, Babrak L, Ohlin M, et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. *Bioinformatics*. 2020;36:1731-1739.

47. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med*. 2015;7.

48. Liu H, Pan W, Tang C, et al. The methods and advances of adaptive immune receptors repertoire sequencing. *Theranostics*. 2021;11:8945-8963.

49. Chaudhary N, Wesemann DR. Analyzing Immunoglobulin Repertoires. *Front Immunol*. 2018;9:462.

50. Lefranc M-P. Nomenclature of the Human Immunoglobulin Heavy (IGH) Genes. *Exp Clin Immunogenet*. 2001;18:100-116.

51. Rizzo JM, Buck MJ. Key Principles and Clinical Applications of *"Next-Generation"* DNA Sequencing.

Cancer Prev Res. 2012;5:887-900.

52. Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17:333-351.

53. Syed M, Khedoudja N, Baldi T, et al. MSK-LYMPHOCLONE: Data Analysis Pipeline and Tools for Immune Repertoire Analysis. *J Mol Diagnostics*. 2015;17:804.

54. Gupta SK, Viswanatha DS, Patel KP. Evaluation of Somatic Hypermutation Status in Chronic Lymphocytic Leukemia (CLL) in the Era of Next Generation Sequencing. *Front Cell Dev Biol*. 2020;8:357.

55. Kim M, Jeon K, Hutt K, et al. Immunoglobulin gene rearrangement in Koreans with multiple myeloma: Clonality assessment and repertoire analysis using next-generation sequencing. *PLoS One*. 2021;16:e0253541.

56. Rustad EH, Hultcrantz M, Yellapantula VD, et al. Baseline identification of clonal V(D)J sequences for DNA-based minimal residual disease detection in multiple myeloma. *PLoS One*. 2019;14:e0211600.

57. Looney TJ, Topacio-Hall D, Lowman G, et al. TCR Convergence in Individuals Treated With Immune Checkpoint Inhibition for Cancer. *Front Immunol*. January 2020;10.

58. Sung J-Y, Kang SY, Kim S-H, Kwon JE, Ko Y-H. Analysis of Immunoglobulin Gene Rearrangement: Comparison between BIOMED-2 Multiplex PCR and Conventional Nested PCR. *Lab Med Online*. 2011;1:195.

59. Van Dongen JJM, Langerak AW, Brüggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17:2257-2317.

60. Christley S, Aguiar A, Blanck G, et al. The ADC API: A Web API for the Programmatic Query of the AIRR Data Commons. *Front Big Data*. 2020;3:22.

61. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*. 2015;12:380-381.

62. Christley S, Levin MK, Toby IT, et al. VDJPipe: a pipelined tool for pre-processing immune repertoire sequencing data. *BMC Bioinformatics*. 2017;18.

63. Vander Heiden JA, Yaari G, Uduman M, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*. 2014;30:1930-1932.

64. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data: Table 1. *Bioinformatics*. 2015;31:3356-3358.

65. Shugay M, Bagaev D V, Turchaninova MA, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Comput Biol*. 2015;11:e1004503.

66. Olson BJ, Moghimi P, Schramm CA, et al. sumrep: A Summary Statistic Framework for Immune Receptor Repertoire Comparison and Model Validation. *Front Immunol*. 2019;10:2533.

67. Hakim NS, Danovitch GM. *Transplantation Surgery*.; 2010.

68. Johansen KL, Chertow GM, Foley RN, et al. US Renal Data System 2020 Annual Data Report: Epidemiology of Kidney Disease in the United States. *Am J Kidney Dis*. 2021;77:A7-A8.

69. Lee HS, Kang M, Kim B, Park Y. Outcomes of kidney transplantation over a 16-year period in Korea: An analysis of the National Health Information Database. Stepkowski S, ed. *PLoS One*. 2021;16:e0247449.

70. Schinstock CA, Mannon RB, Budde K, et al. Recommended Treatment for Antibody-mediated Rejection After Kidney Transplantation. *Transplantation*. 2020;104:911-922.

71. Schinstock C, Tambur A, Stegall M. Current Approaches to Desensitization in Solid Organ Transplantation. *Front Immunol*. May 2021;12.

72. Cosimi AB, Ascher NL, Emond JC, et al. The Importance of Bringing Transplantation Tolerance to the Clinic. *Transplantation*. 2021;105:935-940.

73. Zarkhin V, Chalasani G, Sarwal MM. The yin and yang of B cells in graft rejection and tolerance. *Transplant Rev*. 2010;24:67-78.

74. Kwun J, Bulut P, Kim E, et al. The role of B cells in solid organ transplantation. *Semin Immunol*.

2012;24:96-108.

75. Pineda S, Sigdel TK, Liberto JM, Vincenti F, Sirota M, Sarwal MM. Characterizing pre-transplant and post-transplant kidney rejection risk by B cell immune repertoire sequencing. *Nat Commun*. 2019;10.

76. Aschauer C, Jelencsics K, Hu K, et al. Next generation sequencing based assessment of the alloreactive T cell receptor repertoire in kidney transplant patients during rejection: a prospective cohort study. *BMC Nephrol*. 2019;20.

77. Alachkar H, Mutonga M, Kato T, et al. Quantitative characterization of T-cell repertoire and biomarkers in kidney transplant rejection. *BMC Nephrol*. 2016;17.

78. Lai L, Wang L, Chen H, et al. T cell repertoire following kidney transplantation revealed by high-throughput sequencing. *Transpl Immunol*. 2016;39:34-45.

79. Lai L, Zhou X, Chen H, et al. Composition and diversity analysis of the B-cell receptor immunoglobulin heavy chain complementarity-determining region 3 repertoire in patients with acute rejection after kidney transplantation using high-throughput sequencing. *Exp Ther Med*. 2019.

80. Matsutani T, Ohashi Y, Yoshioka T, et al. Skew in T-cell receptor usage and clonal T-cell expansion in patients with chronic rejection of transplanted kidneys. *Transplantation*. 2003;75:398-407.

81. López-Santibáñez-Jácome L, Avendaño-Vázquez SE, Flores-Jasso CF. The Pipeline Repertoire for Ig-Seq Analysis. *Front Immunol*. 2019;10:899.

82. Jackson KJL, Boyd S, Gaëta BA, Collins AM. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics*. 2010;26:3129-3130.

83. Loupy A, Haas M, Roufosse C, et al. The Banff 2019 Kidney Meeting Report (I): Updates on and clarification of criteria for T cell– and antibody-mediated rejection. *Am J Transplant*. 2020;20:2318-2331.

84. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047-3048.

85. Brochet X, Lefranc M-P, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*. 2008;36:W503-W508.

86. ImmunoMind Team. immunarch: An R Package for Painless Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires. *Zenodo*. 2019.

87. Christley S, Scarborough W, Salinas E, et al. VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements. *Front Immunol*. 2018;9:976.

88. Nankivell BJ, Agrawal N, Sharma A, et al. The clinical and pathological significance of borderline T cell–mediated rejection. *Am J Transplant*. 2019;19:1452-1463.

89. Nankivell BJ. The meaning of borderline rejection in kidney transplantation. *Kidney Int*. 2020;98:278-280.

90. Issa N, Cosio FG, Gloor JM, et al. Transplant Glomerulopathy: Risk and Prognosis Related to Anti-Human Leukocyte Antigen Class II Antibody Levels. *Transplantation*. 2008;86:681-685.

91. Cosio FG, Cattran DC. Recent advances in our understanding of recurrent primary glomerulonephritis after kidney transplantation. *Kidney Int*. 2017;91:304-314.

92. Pescovitz MD, Book BK, Sidner RA. Resolution of Recurrent Focal Segmental Glomerulosclerosis Proteinuria after Rituximab Treatment. *N Engl J Med*. 2006;354:1961-1963.

93. Canaud G, Zuber J, Sberro R, et al. Intensive and Prolonged Treatment of Focal and Segmental Glomerulosclerosis Recurrence in Adult Kidney Transplant Recipients: A Pilot Study. *Am J Transplant*. 2009;9:1081-1086.

94. Mcdonald TJ, Kuo L, Kuo FC. Determination of VH Family Usage in B-Cell Malignancies via the BIOMED-2 IGH PCR Clonality Assay. *Am J Clin Pathol*. 2017;147:549-556.

95. Quail M, Smith ME, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*. 2012;13:341.

96. Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 2015;43:e37-e37.

97. Piccaluga PP, Rapezzi D, Gazzola A, Malagola M, Visani G, Gallamini A. Resolving the diagnostic dilemma of T-cell clonal expansion after hematopoietic stem cell transplantation in T-cell lymphoma

patients by TCR-gamma next generation sequencing. *Bone Marrow Transplant*. 2019;54:159-163.

98. Bedi KS, Goldstein DJ. Apparent anomalies in nuclear feulgen-DNA contents. Role of systematic microdensitometric errors. *J Cell Biol*. 1976;71:68-88.

99. Petrakis NL. Microspectrophotometric Estimation of the Desoxyribonucleic Acid (DNA) Content of Individual Normal and Leukemic Human Lymphocytes. *Blood*. 1953;8:905-915.

100. Dorman A, Graham D, Curran B, Henry K, Leader M. Ploidy of smooth muscle tumours: retrospective image analysis study of formalin fixed, paraffin wax embedded tissue. *J Clin Pathol*. 1990;43:465-468.

101. Davis CW, Jackson KJL, Mcelroy AK, et al. Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. *Cell*. 2019;177:1566-1582.e17.

102. Levin M, Levander F, Palmason R, Greiff L, Ohlin M. Antibody-encoding repertoires of bone marrow and peripheral blood—a focus on IgE. *J Allergy Clin Immunol*. 2017;139:1026-1030.

103. Glanville J, Kuo TC, von Büdingen H-C, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci*. 2011;108:20066-20071.

104. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood, J Am Soc Hematol*. 2010;116:1070-1078.

105. Soto C, Bombardi RG, Branchizio A, et al. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*. 2019;566:398-402.

106. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB, a Novel Tool for the Analysis, Storage, and Dissemination of Immune Repertoire Sequencing Data. *Front Immunol*. 2018;9:2107.

107. Louis K, Bailly E, Macedo C, et al. T-bet+CD27+CD21– B cells poised for plasma cell differentiation during antibody-mediated rejection of kidney transplants. *JCI Insight*. 2021;6.

108. Ohm-Laursen L, Nielsen M, Larsen SR, Barington T. No evidence for the use of DIR, D–D fusions, chromosome 15 open reading frames or VHreplacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*. 2006;119:265-277.

109. Bischof J, Ibrahim SM. bcRep: R Package for Comprehensive Analysis of B Cell Receptor Repertoire Data. *PLoS One*. 2016;11:e0161569.

110. Krissinel E. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*. 2007;23:717-723.

111. Pearson WR. An Introduction to Sequence Similarity (“Homology”) Searching. *Curr Protoc Bioinforma*. 2013;42:3.1.1-3.1.8.

112. Adler AS, Bedinger D, Adams MS, et al. A natively paired antibody library yields drug leads with higher sensitivity and specificity than a randomly paired antibody library. *MAbs*. 2018;10:431-443.

113. Jo I, Chung N-G, Lee S, et al. Considerations for monitoring minimal residual disease using immunoglobulin clonality in patients with precursor B-cell lymphoblastic leukemia. *Clin Chim Acta*. January 2019;488:81-89.

부록 Appendices

Sample No. Sex/Age	Ethnicity	Raw reads (bases)	Consensus sequences
SRR8283601 F/18	Caucasian	218,356,368	24,592,893
SRR8283619 F/21	Caucasian	341,880,369	39,963,919
SRR8283655 F/25	African American	228,526,194	90,598,768
SRR8283727 F/29	African American	267,970,240	13,528,917
SRR8283755 M/29	Caucasian	295,183,125	17,991,497
SRR8283773 M/19	African American	298,965,776	86,637,579
SRR8283791 M/29	Caucasian	275,955,787	35,726,036
SRR8283825 F/30	African American	320,844,194	11,767,640
SRR8283843 M/26	Caucasian	332,209,280	30,967,338
D103 M/25	Caucasian	322,781,254	11,746,606

Supplementary Table 1. Demographic information and sequencing statistics of the samples obtained from SRA.

Accession number PRJNA406949

Data loading

```
immdata <- repLoad(“/path to immdata”)
```

Basic statistics and clonality

```
repExplore(immdata$data, .method=”count”) %>% vis()
repExplore(immdata$data, .method=”volume”) %>% vis()
repClonality(immdata$data, "top") %>% vis()
repClonality(immdata$data, "rare") %>% vis()
repClonality(immdata$data, "homeo") %>% vis()
```

Diversity

```
repDiversity(immdata$data, "chao1") %>% vis()
repDiversity(immdata$data, "hill") %>% vis()
repDiversity(immdata$data, "d50") %>% vis()
repDiversity(immdata$data, "div") %>% vis()
```

V(D)J gene usage

```
geneUsage(immdata$data, "hs.ighv", .type="family", .norm = T) %>% vis()
geneUsage(immdata$data, "hs.ighj", .type="family", .norm = T) %>% vis()
geneUsageAnalysis(imm_gu, .method = "js", .verbose = F)
geneUsageAnalysis(imm_gu, .method = "cor", .verbose = F)
imm_gu_js[is.na(imm_gu_js)] <- 0
vis(geneUsageAnalysis(imm_gu, "cosine+hclust", .verbose = F))
```

Supplementary Table 2. *Immunarch* R commandlines used for repertoire analysis

	Names	N30	N84	N88	N96	R11	R13	R25	R35	R43	R63	R64	R75	R80	R81	R91
1	IGHV1-18	0.051152	0.05451	0.033116	0.045567	0.019739	0.050167	0.053717	0.04952	0.0456	0.0446	0.048293	0.044958	0.042339	0.057511	0.042652
2	IGHV1-2	0.005471	0.007989	0.010146	0.0245	0.0004	0.012499	0.01795	0.005844	0.036885	0.016578	0.025248	0.047117	0.000948	0.036337	0.007748
3	IGHV1-24	0.006577	0.012228	0.003634	0.005948	0.002655	0.012724	0.013926	0.007307	0.009606	0.009928	0.011637	0.007138	0.006688	0.018417	0.009421
4	IGHV1-3	0.01391	0.021011	0.016825	0.009646	0.017336	0.022025	0.025434	0.016775	0.008792	0.017774	0.010885	0.011945	0.019009	0.013334	0.018979
5	IGHV1-45	0.000231	NA	NA	8.66E-05	NA	0.00029	0.005577	0.000153	0.000127	0.000165	NA	NA	4.53E-05	3.77E-05	NA
6	IGHV1-46	0.019562	0.017111	0.012527	0.014767	0.008343	0.019185	0.015914	0.015626	0.015103	0.020517	0.014644	0.018235	0.011839	0.023794	0.014564
7	IGHV1-58	0.005279	0.0036	0.003337	0.005492	0.001661	0.007004	0.00227	0.005733	0.003622	0.006775	0.005761	0.001272	0.003496	0.006777	0.004034
8	IGHV1-69	NA	NA	NA	2.79E-06	NA	NA	3.13E-06	NA	NA	NA	NA	NA	NA	NA	NA
9	IGHV1-69D	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.96E-06	NA	NA	NA	NA
10	IGHV1-8	0.012203	0.036247	0.012744	0.027886	0.008766	0.007465	0.010193	0.011284	0.013673	0.010244	0.026942	2.29E-05	0.022104	1.37E-05	0.022037
11	IGHV2-26	0.00624	0.003428	0.001922	0.003858	0.001411	0.003502	0.00041	0.001786	0.004476	0.002572	0.00423	0.003867	0.001565	0.005155	0.004048
12	IGHV2-5	0.005015	0.00824	0.005304	0.004481	0.004762	0.006413	0.009623	0.004908	0.004766	0.003352	0.009686	0.002335	0.005357	0.005172	0.007473
13	IGHV2-70	0.004536	0.003304	0.005265	0.003651	0.002517	0.004413	0.001898	0.004015	0.00609	0.004141	0.005864	0.00326	0.001872	0.005306	0.002576
14	IGHV3-11	0.000442	0.070328	0.016185	0.060574	0.047699	0.017714	0.022133	0.028356	0.026982	0.000429	0.072995	1.53E-05	0.074944	0.001941	0.062712
15	IGHV3-13	0.014293	0.0142	0.010428	0.01777	0.008228	0.021803	0.021663	0.00985	0.016648	0.015609	0.007055	0.01462	0.026889	0.018077	0.019251
16	IGHV3-15	0.088339	0.051034	0.062504	0.065785	0.051499	0.07112	0.056113	0.056738	0.074474	0.068661	0.027496	0.087133	0.056817	0.081257	0.081702
17	IGHV3-20	0.030735	0.020323	0.007509	0.017932	0.000827	0.008369	0.010572	0.006197	0.009785	0.019206	0.005563	0.00551	0.014311	0.019298	0.005086
18	IGHV3-21	0.002242	0.003489	0.007689	0.003218	0.000991	0.001369	0.008295	0.003369	0.004697	0.003905	0.000415	0.000344	0.002882	0.001855	0.002612
19	IGHV3-23	0.303471	0.129024	0.27866	0.15089	0.357296	0.235307	0.203157	0.376055	0.222854	0.264432	0.211697	0.272501	0.197653	0.161266	0.2216
20	IGHV3-30	6.60E-06	4.41E-05	NA	NA	NA	3.41E-06	NA	NA	1.63E-05	NA	NA	NA	NA	NA	7.25E-06
21	IGHV3-33	NA	NA	1.20E-05	7.82E-05	3.28E-06	NA	3.13E-06	4.25E-06	4.09E-06	3.11E-06	NA	3.82E-06	6.97E-06	6.86E-06	3.63E-06
22	IGHV3-43	0.008614	0.035978	0.019988	0.009096	0.01258	0.027575	0.023367	0.016685	0.037159	0.0347	0.019907	0.029363	0.026798	0.011486	0.030206
23	IGHV3-48	0.001007	NA	0.001826	0.00081	0.004188	0.001587	0.00471	NA	4.09E-06	0.002059	3.17E-05	NA	0.001366	0.001152	0.00017
24	IGHV3-49	0.006841	0.009541	0.006662	0.009258	0.003492	0.007454	0.002574	0.011131	0.013154	0.008322	0.0106	0.011494	0.004677	0.006945	0.01122
25	IGHV3-53	NA	NA	NA	NA	NA	NA	NA	NA	NA	6.21E-06	NA	NA	NA	NA	NA
26	IGHV3-64	0.002493	0.016277	0.007836	0.001131	0.001359	0.005413	0.010102	0.00293	0.012491	0.002575	0.007134	0.018377	0.001513	0.009785	0.006671

27	IGHV3-66	NA	NA	9.31E-05	2.79E-06	NA	0.000116	NA	NA	4.09E-06	3.11E-06	0.000241	NA	NA	1.71E-05	NA
28	IGHV3-7	0.057079	0.075997	0.10824	0.110329	0.148968	0.084223	0.07803	0.060536	0.076346	0.087892	0.069078	0.081008	0.095043	0.06092	0.088333
29	IGHV3-72	0.005742	0.007014	0.014831	0.009906	0.015734	0.009905	0.006686	0.004674	0.005637	0.005924	0.00937	0.005411	0.004691	0.007415	0.007523
30	IGHV3-73	0.015607	0.010834	0.017867	0.019589	0.025679	0.017687	0.000849	0.011615	0.010967	0.011285	0.008218	0.013012	0.011554	0.008876	0.011557
31	IGHV3-74	0.045863	0.029736	0.083597	0.0611	0.081209	0.040211	0.056451	0.031095	0.042391	0.050331	0.053951	0.032951	0.05338	0.040034	0.044781
32	IGHV3-9	0.037933	0.081184	0.044219	0.098336	0.064743	0.026435	0.093844	0.061701	0.045191	0.035796	0.132905	2.67E-05	0.114651	2.40E-05	0.08686
33	IGHV4-28	0.000116	0.00019	0.000309	0.000103	NA	0.000102	NA	8.51E-06	6.13E-05	7.14E-05	0.000384	1.15E-05	4.88E-05	0.000195	1.81E-05
34	IGHV4-30-2	NA	0.005801	NA	NA	NA	0.004137	1.25E-05	0.005525	4.09E-06	NA	NA	7.64E-06	NA	NA	0.00234
35	IGHV4-31	0.029041	0.029299	0.008386	0.011663	0.008415	0.021302	0.010544	0.012594	0.015958	NA	0.008337	0.020627	0.009867	0.012954	0.016403
36	IGHV4-34	0.101998	0.093042	0.071358	0.073313	0.018084	0.110119	0.051553	0.061233	0.090816	0.131746	0.074281	0.083052	0.100731	0.21091	0.075473
37	IGHV4-39	0.068262	0.096262	0.070262	0.08363	0.028321	0.080206	0.0818	0.069161	0.0951	0.066179	0.057196	0.107665	0.036512	0.103045	0.050538
38	IGHV4-4	NA	0.000335	6.01E-06	0.010504	NA	NA	0.007678	NA	0.009042	0.008238	0.011775	0.012629	NA	0.016054	0.000943
39	IGHV4-59	NA	4.41E-06	3.90E-05	NA	NA	NA	NA	NA	4.09E-06	NA	NA	3.82E-06	1.05E-05	NA	NA
40	IGHV4-61	NA	2.21E-05	0.000117	2.79E-06	3.28E-06	1.02E-05	3.13E-06	4.25E-06	1.63E-05	3.11E-06	0.000186	0.000218	NA	6.86E-06	0.000236
41	IGHV5-10-1	0.009169	4.41E-06	0.006947	NA	3.28E-06	0.013185	0.004622	0.009349	0.00992	0.010086	NA	0.022129	NA	0.015759	NA
42	IGHV5-51	0.027219	0.025876	0.024832	0.022726	0.01131	0.019868	0.019052	0.020862	0.024705	0.022645	0.016864	0.028109	0.017318	0.023273	0.0155
43	IGHV6-1	0.012243	0.011809	0.014579	0.009071	0.028912	0.014151	0.045416	0.006465	0.006818	0.012922	0.01481	0.012293	0.009194	0.014758	0.007832
44	IGHV7-4-1	0.00107	0.014685	0.0102	0.007291	0.012863	0.014939	0.023856	0.010909	1.23E-05	0.000329	0.016314	0.001334	0.023881	0.000837	0.01689

Supplementary Table 3. IGHV gene usage analysis results of individual samples.

The fraction of which each IGHV genes constitute the repertoire of each specimen are shown.

영문요약 Abstract

Backgrounds: The adaptive immune repertoire is responsible for protection of the body from various pathogens and foreign substances. The immune system is comprised of various immune cells and signaling pathways of which balance and interaction between its components are of importance. The innate immune system and the adaptive immune systems are the two main components of the immune system. The adaptive immune system can be characterized by its ability to respond to almost unlimited number of different pathogens, through antigen recognition using the corresponding receptors. The ability to generate the vast number of antibodies is summarized as the antibody repertoire, so called the adaptive immune receptor repertoire (AIRR). The fundamental ability of T cell receptors and immunoglobulin receptors being able to recognize the different antigens rely on the existence of immune cells of different clones. Such vast diversity of immune cell clones is generated through mechanisms of V(D)J recombination, somatic hypermutation and class switching, etc. The V, D and J segments of the *IGH* gene distributed throughout the genome is rearranged to facilitate immune response to different antigens and the consequent ability to generate antibodies.

The improvements in outcome of solid organ transplantation during the past few decades has benefited from the better understanding of the immune system and histocompatibility. Especially the use of efficient immunosuppressants have improved the outcome of kidney transplantation, although further improvements in long-term outcome has been marginal since the 2000's. The major limiting factor in improving the long-term outcome remains to be antibody-mediated rejection. There remains a great unmet need for development of specific and early biomarkers that could predict development of antibody-mediated rejection or monitor treatment responses. AIRR analysis in kidney transplantation has been applied in limited number of studies, mostly focusing on the T cell receptors. In this study we applied NGS analysis of the AIRR for the assessment of B cell repertoire in kidney transplantation recipients. In specific, here we address the various tools and packages used for AIRR analysis pipelines and compare their characteristic features, limitations and attempt to provide a standardized pipeline that can be used for AIRR-seq.

Method: Kidney transplantation recipients who underwent kidney transplantation during December 1996 and March 2021 were enrolled. Those who received ABO-compatible kidney transplantation with available pre-transplantation immunologic and histocompatibility status results and followed up for post-transplantation monitoring which included kidney biopsy were selected for study. Total of 15 patients were selected and classified by their kidney biopsy results, according to the Banff 2017 revised diagnostic criteria, into two groups: Antibody-mediated rejection (ABMR) group and No rejection (NR) groups. To assess the laboratory analytical features of AIRR-seq, the reproducibility and changes in clonotypes according to target read depth were also analyzed. To compensate the limited number of NR group samples, normal control sequences were obtained from the NCBI SRA database. The NGS analysis used the LymphoTrack assay kit which is a commercialized assay kit based on the BIOMED-2 protocol used for assessment of clonality. AIRR analysis included clonality, diversity, CDR3 analysis and gene usages. The difference of pipelines and the tooled used accordingly were compared by comparing three different pipelines constructed by using MiXCR, VDJPipe and LymphoTrack softwares.

Results: The result of AIRR analysis showed different results between the two groups. The number of clones and clonotypes was different between groups and between samples. Application of a specific cutoff value, such as 5,000 clonotypes generated statistically significant differences between groups. Diversity and clonality was found different between ABMR and NR groups but the limited number of samples were unable to show statistically significant differences. Various indices of diversity, Chao1, Hill number and D50 were all shown to be lower in ABMR group compared to NR group. The results in repertoire analyses were consistent in all three preprocessing pipelines, except for the clonotyping feature, of which consistency between MiXCR and LymphoTrack pipelines was observed. Using the SRA normal control sequence data, the number of clones and clonotypes correlated with the increasing number of read depth, suggesting the need for higher target read depth than conventional targets. The reproducibility was found consistent from the stable clonotype number, despite the varying degree of clones according to different read depths.

Conclusion: The AIRR analysis of kidney transplant recipients have shown different characteristics of AIRR using post-transplantation samples. Use of commercialized assay kit was suitable for AIRR-seq and the repertoire analysis considering the importance of standardization and reproducibility of data. The pipelines constructed and compared in this study using LymphoTrack were capable in terms of preprocessing and repertoire analysis. Its popular application for larger studies and different study subjects is highly anticipated for the future.

Key words: B cell, Adaptive immunity, repertoire, sequencing, high-throughput, kidney transplantation, antibody-mediated, T cell-mediated, rejection