cells

MDPI

Review

T-Cell Receptor Repertoire Analysis with Computational Tools—An Immunologist's Perspective

Mahima Arunkumar [1,2,3] and Christina E. Zielinski [1,2,*]

[1] Department of Infection Immunology, Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knoell-Institute, 07745 Jena, Germany; M.Arunkumar@campus.lmu.de
[2] Department of Biological Sciences, Friedrich Schiller University, 07743 Jena, Germany
[3] Bioinformatics, Ludwig Maximilians University Munich, 80539 Munich, Germany
* Correspondence: christina.zielinski@leibniz-hki.de

Abstract: Over the last few years, there has been a rapid expansion in the application of information technology to biological data. Particularly the field of immunology has seen great strides in recent years. The development of next-generation sequencing (NGS) and single-cell technologies also brought forth a revolution in the characterization of immune repertoires. T-cell receptor (TCR) repertoires carry comprehensive information on the history of an individual's antigen exposure. They serve as correlates of host protection and tolerance, as well as biomarkers of immunological perturbation by natural infections, vaccines or immunotherapies. Their interrogation yields large amounts of data. This requires a suite of highly sophisticated bioinformatics tools to leverage the meaning and complexity of the large datasets. Many different tools and methods, specifically designed for various aspects of immunological research, have recently emerged. Thus, researchers are now confronted with the issue of having to choose the right kind of approach to analyze, visualize and ultimately solve their task at hand. In order to help immunologists to choose from the vastness of available tools for their data analysis, this review addresses and compares commonly used bioinformatics tools for TCR repertoire analysis and illustrates the advantages and limitations of these tools from an immunologist's perspective.

Keywords: T-cell receptor repertoire; bioinformatic analysis; T cells; systems immunology

Citation: Arunkumar, M.; Zielinski, C.E. T-Cell Receptor Repertoire Analysis with Computational Tools—An Immunologist's Perspective. *Cells* **2021**, *10*, 3582. https://doi.org/cells10123582

Academic Editor: Dieter Kabelitz

Received: 11 November 2021
Accepted: 15 December 2021
Published: 18 December 2021

1. Introduction

Long-lasting T- and B-cell responses are the hallmark of immunological memory. Quantitative shifts in T cells with distinct T-cell receptors occur as a result of proliferation and thus clonal expansion in response to cognate antigens, which can originate from a plethora of microbial pathogens but also autoantigens. Clonally expanded T cells express the same unique TCR. Some of them will persist after a contraction phase to provide long-term immunological memory, i.e., against viral antigens. They can therefore serve as biomarkers of host protection. All individual TCRs of a given individual comprise the TCR repertoire. It represents a reflection of the individual's history of antigen exposures [1]. Correlation of antigen-specific T cells with their respective functional polarization and differentiation state provides further qualitative measures of protection and tolerance [1].

Despite many ingenious methods, technical limitations have made it difficult for many years to create a comprehensive overview of TCR repertoires, until highly specific methods based on next-generation sequencing were developed. They facilitated the parallel analysis of millions of TCR and BCR (B cell receptor) sequences. In particular, single-cell sequencing made it possible to determine the relationship and dynamics of antibody and TCR repertoires with more accuracy on the single-cell level. Thus, high-throughput sequencing has enabled the profiling of TCRs and BCRs in single cells and of transcriptomes at an impeccable resolution, which is greatly adding to our understanding of adaptive immune responses in health and disease [1–3].

The effectiveness of the adaptive immune system strongly depends on the diversity and availability of antigen receptors [4]. How does the adaptive immune system generate antigen receptors with such a huge diversity to cover recognition of a plethora of antigens? The answer is that complete genes that encode a variable region are generated by the somatic recombination of separate gene segments. Multiple contiguous variable gene segments (TCRV) are present at each immune globulin locus. The variable (V), diversity (D) and joining (J) segments are rearranged in a stochastic fashion, which is the cause for the combinatorial diversity in antigen receptors encoding the complementary-determining region three (CDR3) of the TCRb receptor [5]. The hypothetical diversity of the TCR repertoire achieved by combinatorial diversity is by itself huge, with the actual heterogeneity increasing even more by the process of non-template insertion and deletion, as well as pairing of heterogeneous chains. It probably spans the range of 10^{15}–10^{20}, thus exceeding the number of 10^{12} T cells that continuously patrol the body [1,4,6]. The massive hypothetical diversity of the TCR repertoire therefore results in only a limited number of T cells with a unique TCR within an individual. Consequently, rare T-cell clones can easily be overlooked. Novel sequencing-based technologies have strongly improved this bottleneck [2,7].

The information gained from a comprehensive TCR repertoire analysis is immense. The TCR represents a unique identification of T-cell clones due to the low probability of somatic recombination of the exact V(D)J rearrangement in a second T cell within the same individual. An increase in specific TCR frequency can act as a proxy for antigen-specific immune responses, resulting in clonal expansion of antigen-specific T cells with the same TCR (Figure 1). Longitudinal analysis of the TCR repertoire, paired with functional assessment of the T-cell quality, can yield insights into the ensuing immune response to the respective antigen, i.e., a pathogen. This is essential to gain an understanding for the generation of protective immunity, the pathogenesis of immune-mediated diseases and for the design of therapeutic strategies.

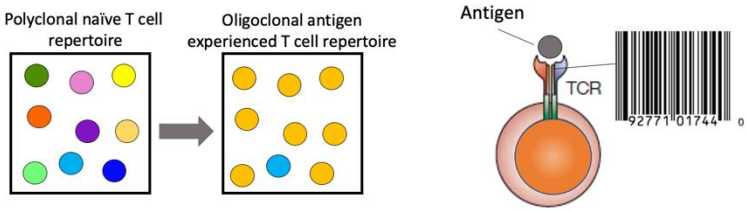

Figure 1. Clonal outgrowth of a T cell with a unique TCR (barcode) due to stimulation with its cognate antigen that generates shifts in the composition of TCR repertoire.

How can we tackle the magnitude and diversity of the human TCR repertoire to map the history of antigen encounter and to retrieve immunological correlates of host protection and tolerance? In the following sections, we provide an overview of the currently available methodologies for repertoire analysis by looking at five specific tools that are frequently used by immunologists. In addition, we describe the different aspects, including the advantages and disadvantages, that the immunologist should consider when choosing the appropriate method for a given research question.

2. Data Analysis for Exploring Immune Repertoires

A variety of computational and statistical methods are currently available for studying large sets of raw data [7–13]. Most of these tools are helpful for identifying features and patterns in the data that have some functional or biochemical significance. The first step in the analysis is the recovery of TCR or BCR sequences from the raw data, and this step is then followed by clustering and annotation. The next step in the workflow is the visualization of the immune repertoire (IR), which commonly includes clonotype abundance, diversity and V(D)J usage. Especially the calculation of gene usage in the different samples is of significance, since a change in the usage of specific genes may be due to alterations in the repertoire caused by the respective underlying disease or immunological pertur-

bation. The last step in the data analysis of the immune repertoire traditionally involves the visualization of repertoire overlap and clustering, as well as specialized analysis of individual clonotypes and generation of publication-ready figures. Particularly changes in TCR overlap and diversity over time are important measurements that could be informative of disease progression. Figure 2 illustrates the data and information flow in standard immunology research.

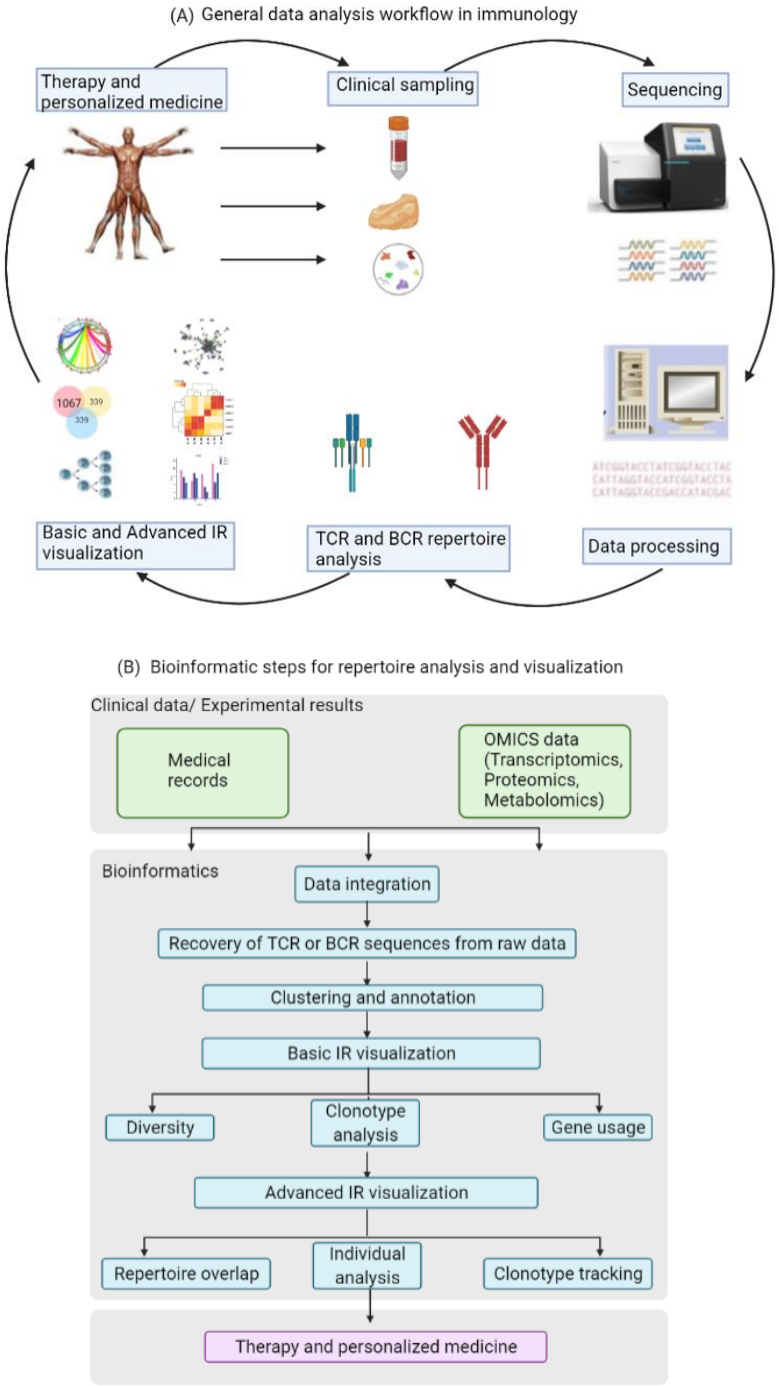

Figure 2. Standardized data analysis workflow for TCR/BCR repertoire analysis. (**A**) Pipeline from the clinic to the sequencing results. (**B**) Bioinformatic steps for basic and advanced immune repertoire analysis and visualization.

To determine TCR diversity, several indices can be employed. In T-cell repertoires, diversity takes the clonal composition into account, specifically the number of unique TCR sequences (richness) and the relative abundance of these sequences (evenness). However, it is important to point out that it is not possible to directly measure the total TCR diversity in a given sample. Since the TCR repertoire is highly diverse and the distributions of individual TCRs can be heavily skewed, the diversity cannot be evaluated from experimental samples alone [4,6,14,15].

Most of the diversity indices stem from the information theory to quantify ecosystems biodiversity. Commonly used indices are the Shannon index, the Inverse Simpson index, Gini coefficient or the DE50 score [4,15–18].

The Shannon index accounts for sample richness and evenness. A large Shannon index value implies that the distribution of the CDR3 sequences is more diverse. When using the Shannon index to compare different samples with each other, researchers must be careful, as this index assumes that the distributions of the clonal frequencies of the samples are similar to each other, and this may not necessarily be true for every sample being compared. Moreover, the Shannon index is sensitive to low-frequency reads. Thus, variations and changes in low-frequency reads could affect the outcome of this diversity index [15,16].

The Inverse Simpson index, in contrast to the Shannon index, emphasizes high-frequency reads, and therefore researchers should only use the Inverse Simpson index when their dataset contains high-frequency reads. High values of this index indicate an even distribution of TCR clones, and low values indicate enrichment of T-cell clones [15,16].

The Gini coefficient measures the inequality among values of a frequency distribution. It was originally proposed to measure the inequality of income or wealth across countries but can also be used in immunology. The scale ranges between zero and one. Zero stands for total equality of clones; thus, all clones will have identical frequencies. On the other hand, one stands for total inequality, thus indicating sample oligoclonality. This index can be used to gauge the inequality of the frequency distribution between different clonotypes in each sample and give a summary of the clonotype abundance distribution [19–23].

The DE50 (diversity evenness score) is considered an indicator for the degree of clonality in a given dataset. Summing up the number of reads in ranked order, from high to low, DE50 presents how many unique reads can be found within 50% of in-frame reads. A low value indicates a high clonality level [14–18]. Since each index only takes a fraction of different diversity aspects into account, they all have certain biases when applied to various underlying populations in terms of size and abundance distribution. If two repertoires are being compared, it is possible that one diversity measure shows a certain trend, while the other shows a very different result (or even the opposite result), considering that they represent different aspects of the underlying abundance distributions.

Regarding overlap measures, i.e., a measure on how similar or different two sets of data are, the Morisita–Horn index and the Jaccard index are commonly used in immunology [14,16]. In contrast to the α-diversity measures (e.g., the Shannon index) discussed above, these overlapping measures are often termed as β-diversity measures [18].

The Morisita–Horn index is a statistical measure of dispersion of individuals in a population. It accounts for both the number and abundance of shared TCRs between two repertoires, and its score ranges between 0, meaning no overlap, and 1, meaning all clones overlap at similar frequencies. The Jaccard index measures similarity between sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. It also ranges between 1 and 0, where 1 indicates complete overlap and 0 represents no overlap. The Yu–Clayton index is one of the few similarity indices that can detect and compare the presence and abundance of same TCRs among samples [22]. It has, however, only rarely been used so far in the field of immunology, as compared to the Morisita–Horn index or the Jaccard index. Most of these indices are very similar (especially the Morisita–Horn index and the Jaccard index), but they differ in the consideration they give to factors such as the species richness or the evenness of the given data. Therefore, it is very important

for scientists to be watchful when comparing results obtained from different tools that use different techniques. Moreover, experimental sampling only partially estimates the overlap and diversity of repertoires [4,6,15]. Therefore, researchers must be careful when dealing with immune repertoire data as uniformity between samples is important. To this end, down-sampling or re-sampling are one of the commonly used strategies to generate more comparable data. Down-sampling refers to the reduction of the dataset to a more manageable size and thus working with a random sample without replacement from our original data. Re-sampling means that random data can be drawn with replacement from our original dataset in such a way that this is comparable to the original data. In this way, re-sampling allows us to make unbiased estimates, as it is drawn from unbiased samples.

3. Scirpy

Scirpy is a Python-toolkit that is used to analyze TCR repertoires from single-cell RNA sequencing (scRNA-seq) data [24,25]. It can easily be integrated with the Scanpy library, which is a toolkit for analyzing single-cell gene expression data. Scanpy is one of the standard tools for preprocessing, visualization, clustering, pseudotime trajectory inference, differential expression analysis and simulation of gene regulatory networks [26,27].

Scirpy is a pipeline and is available for the characterization of T-cell receptors. It can be used for the visualization of immune repertoires from single cells and for integration with transcriptomic data to characterize the TCRs of single T cells. Starting with the process of data loading, Scirpy supports a variety of data formats, including 10× Genomics CellRanger, TraCeR, BraCer or AIRR-compliant data [20,27,28]. Detailed tutorials with examples on data loading and core analysis make it easy for less experienced researchers to work with Scirpy.

Scirpy enables the investigation of the composition and phenotypes of both single and dual TCRs in T cells. With Scirpy, researchers can inspect TCR chain configurations and explore the abundance, diversity, expansion and overlap of clonotype repertoires across samples, patients or cell clusters. It can also integrate transcriptomics data. Finally, it is possible to investigate the intricate relationship between cells and clonotypes, as well as analyze the distribution of CDR3 sequence lengths and V(D)J gene usages. Regarding clonotype analysis, Scirpy implements a network-based approach that enables clustering of cells into clonotypes based on having either identical CDR3 nucleotide sequences, identical CDR3 amino acid sequences or similar CDR3 amino acid sequences based on pairwise sequence alignment. The sequence-alignment-based networks offer the opportunity to identify cells that might recognize the same epitopes.

Let us assume that a basic clonotype analysis on a 10× dataset, which contains blood and skin samples from a patient, should be conducted. After loading, preprocessing and normalizing the data, all T cells, based on the criteria of having a TCR, will be extracted from the dataset. Then the clonality can be assessed and a clonotype network can be constructed by showing all the clones in the dataset based on their assigned clone ID. This allows for the identification of shared clones, i.e., clones that can be found in blood, as well as skin samples. In these shared clones, differential gene expression between blood and skin within each shared clone can be analyzed. This identifies whether T cells with identical precursors differ in gene expression as a result of their differential location (skin versus blood). Based on the top differentially expressed genes, we can formulate a certain hypothesis about our data that can then be accepted or rejected depending on further downstream analysis. An example of the results is visualized in Figure 3.

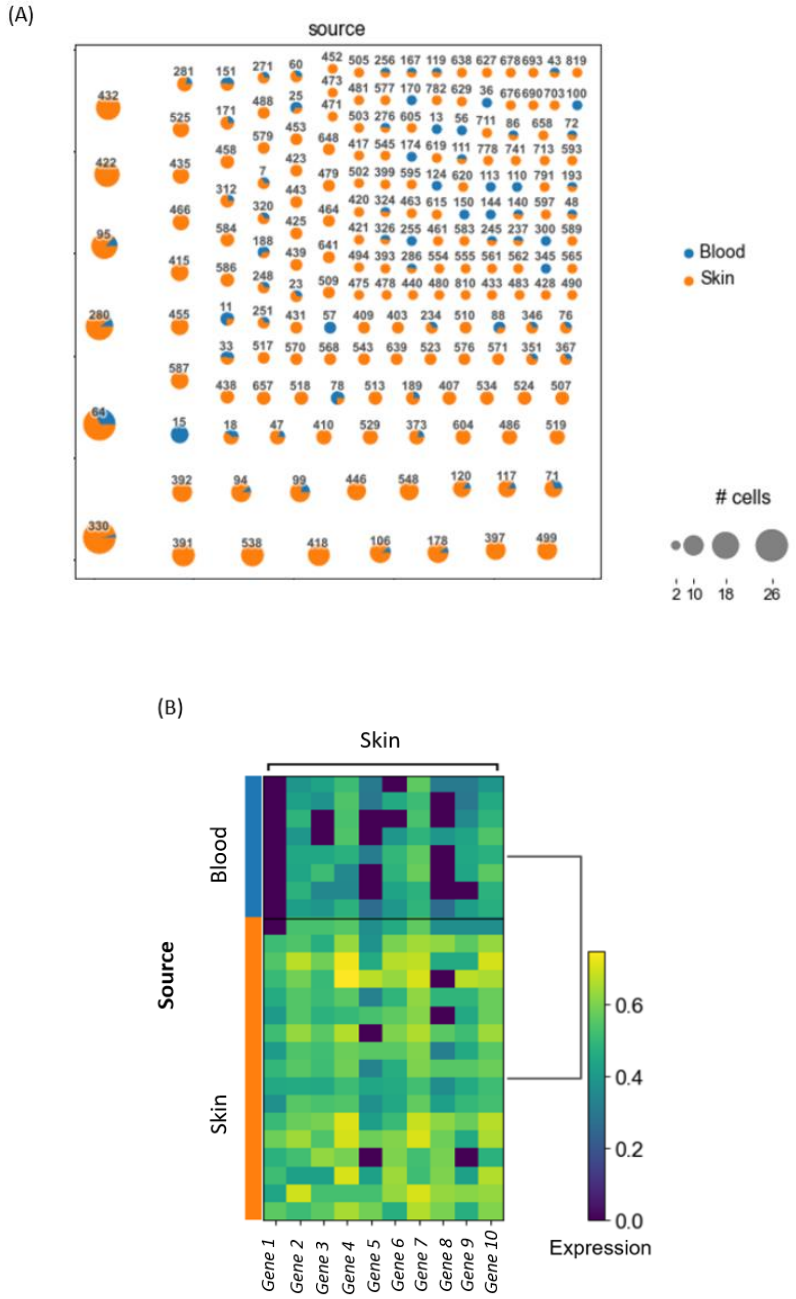

Figure 3. (**A**) Example of a clonotype network constructed by Scirpy, using the default parameters in order to showcase all clones found in the dataset comprising T cells from matched skin and blood of a given patient. (**B**) Heatmap showing the top 10 differentially expressed genes for the shared clone with the clone ID 64 (as depicted in (**A**)), using default parameters in Scirpy.

All results also have the advantage of being generated in a publication-ready style. It can therefore be concluded that Scirpy provides a big range of methods for analyzing T-cell repertoires from single-cell sequencing data. It is open-source and has detailed tutorials to support researchers who conduct the analysis. In this way, prior in-depth knowledge in Python is certainly useful, but not necessary. On the other hand, Scirpy does have two main shortcomings, namely that it does not support bulk data formats or BCR-related analysis. Thus, if this kind of analysis needs to be conducted, researchers need to refer to other tools.

4. Immunarch

Another open-source tool (previously known as tcR) is an R package called Immunarch [29]. It is fully interoperable with Seurat. Seurat, also available as a Bioconductor package, is an R package designed for quality control, analysis and exploration of single-cell RNA-seq data, and thus it contains implementations of commonly applied techniques for exploring single-cell expression data [30]. Immunarch offers data loading, analysis and visualization for all popular TCR and BCR analysis and post-analysis formats, including single-cell data: ImmunoSEQ, IMGT, MiTCR, MiXCR, MiGEC, MigMap, VDJtools, Immunarch, AIRR, 10× Genomics and ArcherDX [27–29,31–35]. Since this tool is constantly being updated, even more useful methods and applications are expected to be available in the future [29]. Additionally, Immunarch works on almost any kind of data, from R data frames and data tables to databases such as MonetDB or Apache Spark data frames via sparklyr. Not only does Immunarch accept all standard immunosequencing formats, but it also automatically detects and parses the format of the uploaded data.

Immunarch implements most of the commonly used analysis methods, such as clonality analysis; estimation of repertoire similarities in the distribution of clonotypes; gene usage and kmer distribution measures; repertoire diversity analysis; clonotype tracking between samples and across time points; and annotation of clonotypes, using external immune receptor databases. This includes the frequently used databases: VDJDB, McPAS-TCR and TBAdb from PIRD [32–34]. All results are generated in a publication style. However, Immunarch also has a built-in tool for additional visualization manipulation, such as changing font sizes, text angles, titles or legends.

Immunarch is beginner-friendly, because a researcher requires little to no knowledge in R in order to conduct the analysis, and most methods are incorporated into a couple of main functions with simple naming. Moreover, Immunarch includes a comprehensive tutorial with examples for an easy start. All of these features make the Immunarch package very popular for research, especially when TCR analysis and clonotype comparison need to be conducted. For example, recently the Immunarch package was used for BCR analysis, which demonstrated that a single human V_H-gene enables a broad antibody response in the blood that targets bacterial lipopolysaccharides [36].

In sum, Immunarch provides a vast range of methods for analyzing both T-cell and B-cell repertoires. The fact that it is an open-source tool and includes detailed explanations and tutorials makes it easy for beginners to conduct advanced analysis in a fast and efficient manner. The only main drawback is the fact that it currently provides paired chain information only for 10× Genomics data.

5. ImmunoSEQ Analyzer 3.0

ImmunoSEQ Analyzer 3.0 is an online web tool for data exploration provided by the company Adaptive Biotechnologies [35]. It enables researchers to understand the contours and dimensions of their data. The immunoSEQ analyzer 3.0 offers all the basic analysis and visualization methods for immunological research. The Sample Overview dashboard includes information regarding the number of productive templates, rearrangements, maximal productive frequencies and clonalities of each sample. The analyzer supports various views, which make it possible to see the details of single samples or even compare two or more samples with each other. The different views allow researchers to identify complete sequence information for all unique TCR or BCR rearrangements in a sample, track how particular clones expand across samples or tissues and compare the gene usage between samples. Additionally, pairwise scatterplots can be generated that reveal the relative abundance of every detected clone. Moreover, the analyzer features tools to conduct additional statistical tests and metrics for immunosequencing data. This includes a tool for plotting a Venn diagram, a down-sampling tool for comparing samples of different sizes, a differential abundance tool for comparing different samples and additional diversity metrics. The generated results are publication-ready and can be exported easily, along with the sample data.

The main advantage of using the analyzer is that it contains a massive database of TCR and BCR sequences. Researchers can incorporate millions of sequences of public data and control samples by using the immunoSEQ Analyzer, compare analysis from multiple investigators, organize all samples into experiment-specific folders and share the created projects with colleagues. The data can even be shared with off-site collaborators by making use of the option of showcasing the data in the company's immuneACCESS database. The biological and translational applications for this methodology are manifold. For example, the immunological correlates of vaccines against SARS-CoV-2 have recently been interrogated with this methodology. Researchers found that polyfunctional spike protein specific Th1 corresponds to a diverse TCR repertoire [33]. In addition, immune cell lineage tracing and cellular developmental pathways can be investigated. It has been shown, for example, that tissue resident memory T cells (T_{RM}) and central memory T cells (T_{CM}) share identical naïve precursor cells due to their overlapping clonal origin [35].

However, this tool also has its drawbacks, such as the fact that the immunoSEQ Analyzer is specifically designed around immunoSEQ data and does not support outside upload. This implies that analysis with non-immunoSEQ data needs to be carried out by using other methods. Additionally, chain pairing information is not supported at present. It is also necessary to mention that it is not open-source software but a commercial tool. However, researchers can make use of the option of creating a free account first in order to explore the full range of the immunoSEQ Analyzer 3.0, including all integrated tools.

6. Immcantation Framework

The Immcantation framework is a massive and powerful analytical system for high-throughput AIRR-seq datasets. Starting with raw reads, this tool contains several Python and R packages that can conduct preprocessing and repertoire analysis and determine population structures. There are eleven core and contributed packages. The seven core packages are pRESTO, Change-O, Alakazam, SHazaM, TIgGER, SCOPer and prestoR. The four contributed packages are [37,38] RDI, RAbHIT, IgPhyML and sumrep [39–41]. All plots are generated in a publication-ready style.

Starting with the tool pRESTO, researchers can perform preprocessing, from raw sequences to paired-end assembly. The main aim of preprocessing is to transform raw reads into sequences where errors have been eliminated. Overall, pRESTO supports multiplexed and RACE samples. It can also perform de-multiplexing, which is usually already performed in sequencing facilities. Moreover, pRESTO supports single-end sequencing, too. Read processing can be performed with or without UMI inclusion, making it compatible for various protocols a researcher might use [3,8]. After this, the pRESTO report package (prestoR) can be used to generate plots.

Regarding clonotyping, the Immcantation framework provides a package called Change-O for standardizing the output of the V(D)J reference alignment software, such as IMGT or IgBLAST [42,43]; clonal clustering; and germline reconstruction. Change-O allows for the processing of reads that contain a premature stop codon. Researchers can also group clonotypes by J or V allele and other parameters, such as the nucleotide Hamming distance, amino acid Hamming distance and many others. When performing hierarchical clustering, researchers are given the choice between single, average or complete linkage [37].

The included R package Alakazam can plot a repertoire lineage tree by using the clone outputs from Change-O. Various diversity measures, such as the species richness, the Shannon index or the inverse Simpson index, are also performed by Alakazam. Additionally, Alakazam calculates V(D)J alleles, determines gene usage, infers clonal abundance and lists the chemical properties of the amino acid sequences [37]. It also provides the option to generate rarefaction curves.

This suite of packages, all integrated into one big framework, makes the Immcantation Portal a mighty asset for almost any kind of repertoire analysis a researcher might want to conduct in the realm of immunology. An example is given by a recent publication that demonstrated stratification of celiac disease patients and controls by naïve B-cell repertoires

by using the TIgGER package to conduct preliminary data processing in order to deduce new alleles and a personalized genotype for all individuals in their data [37,38].

In sum, the Immcantation framework contains various packages which are highly valuable for immunological research. Immcantation includes pRESTO, which handles all stages of sequence processing from raw reads up to V(D)J gene assignment. To facilitate advanced repertoire analysis, Immcantation also contains methods for novel V gene allele detection (TIgGER), subject-specific germline genotype identification, B-cell clone assignment (Change-O and SCOPer), lineage tree construction and analysis (IgPhyML), somatic mutation profiling and selection analysis (BASELINe) [40,41]. Immcantation can start from raw data or read the output of common V(D)J assignment tools, such as IgBLAST. It also supports MiAIRR and the AIRR Community data standard and includes tools to facilitate MiAIRR-compliant submissions to NCBI repositories [28].

However, the Immcantation framework, being a mixture of many sub-packages, may seem confusing and overwhelming at first, even though each package is thoroughly documented. Moreover, Immcantation offers various summary functions for AIRR-seq data, but it does not have a sophisticated method for comparing and visualizing these summaries. Many summaries of interest are implemented in one of the many sub-packages, but there is no single standard data format. This can be a cause for struggle when comparing summaries across packages [37,38,40,41].

7. VDJtools

VDJtools is an open-source software framework for TCR and BCR analysis and is based on Java [44,45]. It can analyze the output of the following VDJ junction mapping and analysis platforms: MiTCR, MiGEC, IgBlast, IMGT, ImmunoSEQ, VDJdb, Vidjil, RTCR, MiXCR and ImSEQ.

The framework also has a built-in tool, called *correct*, that performs frequency-based correction to eliminate erroneous clonotypes. Additionally, VDJtools provides a variety of other filtering options, such as filtering non-functional clonotypes, filtering out all clonotypes found in another sample, filtering by frequency and filtering V(D)J segments that match a specified segment set. VDJtools allows for basic, as well as advanced, IR visualization by applying a diverse set of methods and strategies. Researchers can make use of the command line tool, which is user-friendly and well documented, so that immunologists with little computational background can easily generate publication-ready plots or simply make use of the tabular outputs. For each sample, VDJtools calculates basic statistics of read counts, mean clonotype size and number of non-functional clonotypes. It determines VJ gene usage and the distribution of clonotype abundance by CDR3 sequence lengths [45].

As for diversity indexes, the framework provides many methods, such as Chao 1, Efron–Thisted, Shannon–Wiener index, Normalized Shannon–Wiener index and Inverse Simpson index. The framework performs a comprehensive analysis of clonotype sharing and clonotype tracking, as well. Data can be visualized as scatter plots of overlapping clonotype abundance, abundance plots, sequence clustering dendrograms and (pairwise) overlap plots. VDJtools includes a built-in tool called CalcPairwiseDistances, which performs a pairwise overlap for a set of samples and computes a list of repertoire similarity measures. Additionally, by using databases, such as VDJdb, that contain CDR3 sequences, variable and joining segments can be obtained. VDJtools can also annotate samples based on VDJ junction matching [45].

Considering all features, it becomes evident that VDJtools can be beneficial to researchers for general TCR and BCR data analysis. An example can be found in a recent publication that used the VDJtools software to analyze and visualize their MiXCR bulk TCR output data to demonstrate that a conserved TCR signature dominates a highly polyclonal T-cell expansion during the acute phase of a malaria infection [46,47]. The following publication is another example where VDJtools was used on MiXCR data in order to calculate their average TCR repertoire characteristics weighted by clonotype size. The results indicated that memory $CD4^+$ T cells are also generated in the human fetal intestine, and

this was an unexpected, considering that the fetus is thought to be protected from exposure to foreign antigens [48–50].

In sum, VDJtools provides basic, as well as advanced, methods for analyzing T-cell and B-cell repertoires supporting various file inputs. It is an open-source tool and includes a comprehensive documentation, making it easy to work with. However, VDJtools does have a few major shortcomings, namely that it does not provide support for 10× Genomics or Smart-seq2 data yet. Moreover, paired chain information is not included. Thus, if this kind of analysis needs to be conducted, researchers are advised to use other tools.

8. Discussion

Other sophisticated tools worth mentioning are CoNGA (clonotype neighbor graph analysis) and scRepertoire. CoNGA is a graph-based approach which identifies correlations between gene-expression data and TCR sequences through statistical analysis of gene expression and TCR similarity graphs. It can be useful when studying the complex relationships between TCR sequences and T-cell phenotypes in large heterogeneous single-cell datasets [42,43,51–55]. The R-based tool scRepertoire is used for single-cell immune receptor analysis, and it combines mRNA and immune profiling for data derived from 10× Genomics Chromium Immune Profiling for TCR and BCR analysis [27,54]. At this point, we can conclude that several tools and methods for TCR and BCR repertoire analysis and clonotype identification exist. In this review article, we have illustrated the capabilities, advantages and disadvantages of five commonly used tools in immunology. Table 1 gives an overview of all the five tools that we employed in this review. However, there are many tools that are currently being used in immunology which have not been addressed in this review, and many newer tools are also emerging. This technological progress offers great opportunities for groundbreaking insights, and we can be sure that the methods and tools discussed here will continue to evolve and improve in the future. As of today, a gold-standard method for the field has not yet been identified. Depending on the purpose of the scientific study, some approaches may be more suitable than others. However, it is important to consciously select a method or a tool by keeping all strengths and weaknesses of each approach in mind. Finally, due to the possible method or tool-specific biases, scientists must always be very careful when comparing results obtained from different methods.

Table 1. Overview of commonly used tools in immunology.

Tools	Data Format	Are TCR and BCR Analysis Possible?	Is It Open Source?	Are Costs Involved?	Are Detailed Tutorials Available?	Sharing and Collaborating on Data and Analysis Possible?
Scirpy	Only single cell data supports currently	BCR analysis not supported yet	Yes	No	Yes	No
Immunarch	Compatible with various data formats	TCR and BCR analysis possible	Yes	No	Yes	No
ImmunoSEQ analyzer 3.0	Does not directly support outside data upload	TCR and BCR analysis possible	No	Yes	Yes	Yes
Immcantation portal	Compatible with various data formats	TCR and BCR analysis possible	Yes	No	Yes	No
VDJtools	Compatible with various data formats	TCR and BCR analysis possible	Yes	No	Yes	No

Author Contributions: All authors conceptualized and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the SFB1054 (project B10, project-ID 210592381 to CEZ), SFB1335 (project P18, project-ID 360372040 to CEZ), Collaborative Research Centre (CRC)/Transregio 124 FungiNet (project C7, project number 210879364 to CEZ), the Leibniz Center for Photonics in Infection Research (LPI-BT1, to CEZ) and by Germany's Excellence Strategy—EXC 2051—Project-ID 390713860 to CEZ from the German Research Foundation (DFG) and Carl-Zeiss-Stiftung to CEZ.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zielinski, C.; Corti, D.; Mele, F.; Pinto, D.; Lanzavecchia, A.; Sallusto, F. Dissecting the human immunologic memory for pathogens. *Immunol. Rev.* **2011**, *240*, 40–51. [CrossRef]
2. Six, A.; Mariotti-Ferrandiz, M.E.; Chaara, W.; Magadan, S.; Pham, H.-P.; Lefranc, M.-P.; Mora, T.; Thomas-Vaslin, V.; Walczak, A.M.; Boudinot, P. The past, present, and future of immune repertoire biology—The rise of next-generation repertoire analysis. *Front. Immunol.* **2013**, *4*, 413. [CrossRef] [PubMed]
3. De Simone, M.; Rossetti, G.; Pagani, M. Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. *Front. Immunol.* **2018**, *9*, 1638. [CrossRef] [PubMed]
4. Laydon, D.J.; Bangham, C.; Asquith, B. Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2015**, *19*, 370. [CrossRef]
5. Jung, D.; Alt, F.W. Unraveling V(D)J recombination; insights into gene regulation. *Cell* **2004**, *116*, 299–311. [CrossRef]
6. Arstila, T.P.; Casrouge, A.; Baron, V.; Even, J.; Kanellopoulos, J.; Kourilsky, P. A direct estimate of the human alphabeta T cell receptor diversity. *Science* **1999**, *286*, 958–961. [CrossRef]
7. Pai, J.A.; Satpathy, A.T. High-throughput and single-cell T cell receptor sequencing technologies. *Nat. Methods* **2021**, *18*, 881–892. [CrossRef]
8. Hwang, B.; Lee, J.H.; Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **2018**, *50*, 1–14. [CrossRef] [PubMed]
9. Mouillot, D.; Leprêtre, A. A comparison of species diversity estimators. *Res. Popul. Ecol.* **1999**, *41*, 203–215. [CrossRef]
10. Rosati, E.; Dowds, C.M.; Liaskou, E.; Henriksen, E.K.K.; Karlsen, T.H.; Franke, A. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **2017**, *17*, 61. [CrossRef]
11. Pasetto, A.; Lu, Y.-C. Single-Cell TCR and Transcriptome Analysis: An Indispensable Tool for Studying T-Cell Biology and Cancer Immunotherapy. *Front. Immunol.* **2021**, *12*, 689091. [CrossRef]
12. Tu, A.A.; Gierahn, T.M.; Monian, B.; Morgan, D.M.; Mehta, N.K.; Ruiter, B.; Shreffler, W.G.; Shalek, A.K.; Love, J.C. TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol.* **2019**, *20*, 1692–1699. [CrossRef] [PubMed]
13. Singh, M.; Al-Eryani, G.; Carswell, S.; Ferguson, J.M.; Blackburn, J.; Barton, K.; Roden, D.; Luciani, F.; Phan, T.G.; Junankar, S.; et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.* **2019**, *10*, 3120. [CrossRef] [PubMed]
14. Rempala, G.A.; Seweryn, M. Methods for diversity and overlap analysis in T-cell receptor populations. *J. Math. Biol.* **2013**, *67*, 1339–1368. [CrossRef] [PubMed]
15. Venturi, V.; Kedzierska, K.; Turner, S.J.; Doherty, P.C.; Davenport, M. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J. Immunol. Methods* **2007**, *10*, 321. [CrossRef] [PubMed]
16. Chiffelle, J.; Genolet, R.; Perez, M.A.; Coukos, G.; Zoete, V.; Harari, A. T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.* **2020**, *65*, 284–295. [CrossRef]
17. Aversa, I.; Malanga, D.; Fiume, G.; Palmieri, C. Molecular T-Cell Repertoire Analysis as Source of Prognostic and Predictive Biomarkers for Checkpoint Blockade Immunotherapy. *Int. J. Mol. Sci.* **2020**, *21*, 2378. [CrossRef]
18. Scholz, M. Alpha and Beta Diversity. Available online: https://www.metagenomics.wiki/pdf/definition/alpha-beta-diversity (accessed on 8 November 2021).
19. Hogan, S.A.; Courtier, A.; Cheng, P.F.; Jaberg-Bentele, N.F.; Goldinger, S.M.; Manuel, M.; Perez, S.; Plantier, N.; Mouret, J.-F.; Nguyen-Kim, T.D.L.; et al. Peripheral blood tcr repertoire profiling may facilitate patient stratification for immunotherapy against melanoma. *Cancer Immunol. Res.* **2019**, *7*, 77–85. [CrossRef]

20. Rambaut, A.; Drummond, A.J.; Xie, D.; Baele, G.; Suchard, M.A. Posterior summarisation in Bayesian phylogenetics using Tracer. *Syst. Biol.* **2018**, *67*, 901. [CrossRef]
21. Firebaugh, G. Empirics of World Income Inequality. *Am. J. Sociol.* **1999**, *104*, 1597–1630. [CrossRef]
22. Yue, J.C.; Clayton, M.K. Similarity Measure Based on Species Proportions. *Commun. Stat. Theory Methods* **2005**, *34*, 2123–2131. [CrossRef]
23. Chao, A.; Hsieh, T.C.; Chazdon, R.; Colwell, R.K.; Gotelli, N.J. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology* **2015**, *96*, 1189–1201. [CrossRef]
24. Wang, Y.; Liu, Y.; Chen, L.; Chen, Z.; Wang, X.; Jiang, R.; Zhao, K.; He, X. T Cell Receptor Beta-Chain Profiling of Tumor Tissue, Peripheral Blood and Regional Lymph Nodes From Patients With Papillary Thyroid Carcinoma. *Front. Immunol.* **2021**, *12*, 312. [CrossRef] [PubMed]
25. Sturm, G.; Szabo, T.; Fotakis, G.; Haider, M.; Rieder, D.; Trajanoski, Z.; Finotello, F. Scirpy: A Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* **2020**, *36*, 4817–4818. [CrossRef]
26. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 15. [CrossRef] [PubMed]
27. Chromium Single Cell V(D)J Reagent Kits with Feature Barcoding Technology for Cell Surface Protein. Available online: https://support.10$\times$genomics.com/single-cell-vdj/library-prep/doc/user-guide-chromium-single-cell-vdj-reagent-kits-user-guide-v1-chemistry-with-feature-barcoding-technology-for-cell-surface-protein (accessed on 8 November 2021).
28. Heiden, J.A.V.; Marquez, S.; Marthandan, N.; Bukhari, S.A.C.; Busse, C.; Corrie, B.; Hershberg, U.; Kleinstein, S.H.; Iv, F.A.M.; Ralph, D.K.; et al. AIRR Community Standardized Representations for Annotated Immune Repertoires. *Front. Immunol.* **2018**, *9*, 2206. [CrossRef]
29. ImmunoMind Team. Immunarch: An R Package for Painless Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires. *Zenodo* **2019**, *10*. [CrossRef]
30. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M., III; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587. [CrossRef]
31. Bolotin, D.; Poslavsky, S.; Mitrophanov, I.; Shugay, M.; Mamedov, I.Z.; Putintseva, E.; Chudakov, D.M. MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* **2015**, *12*, 380–381. [CrossRef]
32. Bagaev, D.V.; A Vroomans, R.M.; Samir, J.; Stervbo, U.; Rius, C.; Dolton, G.; Greenshields-Watson, A.; Attaf, M.; Egorov, E.S.; Zvyagin, I.V.; et al. VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **2020**, *48*, 1057–1062. [CrossRef]
33. Tickotsky, N.; Sagiv, T. McPAS-TCR: A manually-curated catalogue of pathology-associated T-cell receptor sequences. *Bioinformatics* **2017**, *33*, 2924–2929. [CrossRef] [PubMed]
34. Zhang, W.; Wang, L.; Liu, K.; Wei, X.; Yang, K.; Du, W.; Wang, S.; Guo, N.; Ma, C.; Luo, L.; et al. PIRD: Pan immune repertoire database. *Bioinformatics* **2020**, *36*, 897–903. [CrossRef]
35. The Power to Propel Your Research. Available online: https://www.adaptivebiotech.com/immunoseq (accessed on 1 September 2021).
36. Sangesland, M.; Yousif, A.; Ronsard, L.; Kazer, S.; Zhu, A.L.; Gatter, G.J.; Hayward, M.R.; Barnes, R.M.; Quirindongo-Crespo, M.; Rohrer, D.; et al. A Single Human V_H-gene Allows for a BroadSpectrum Antibody Response Targeting Bacterial Lipopolysaccharides in the Blood. *Cell Rep.* **2020**, *32*, 108065. [CrossRef] [PubMed]
37. Gupta, N.T.; Heiden, J.A.V.; Uduman, M.; Gadala-Maria, D.; Yaari, G.; Kleinstein, S.H. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **2015**, *31*, 3356–3358. [CrossRef]
38. Heiden, J.A.V.; Yaari, G.; Uduman, M.; Stern, J.N.; O'Connor, K.C.; Hafler, D.A.; Vigneault, F.; Kleinstein, S.H. pRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **2014**, *30*, 1930–1932. [CrossRef]
39. Gadala-Maria, D.; Yaari, G.; Uduman, M.; Kleinstein, S.H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 862–870. [CrossRef]
40. Bolen, C.R.; Rubelt, F.; Heiden, J.A.V.; Davis, M.M. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. *BMC Bioinform.* **2017**, *7*, 155. [CrossRef]
41. Hoehn, K.B.; Heiden, J.A.V.; Zhou, J.Q.; Lunter, G.; Pybus, O.G.; Kleinstein, S.H. Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22664–22672. [CrossRef]
42. Lefranc, M.-P.; Giudicelli, V.; Duroux, P.; Jabado-Michaloud, J.; Folch, G.; Aouinti, S.; Carillon, E.; Duvergey, H.; Houles, A.; Paysan-Lafosse, T.; et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **2015**, *43*, D413–D422. [CrossRef]
43. Ye, J.; Ma, N.; Madden, T.L.; Ostell, J.M. IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **2013**, *41*, W34–W40. [CrossRef]
44. López-Santibáñez-Jácome, L.; Avendaño-Vázquez, S.E.; Flores-Jasso, C.F. The pipeline repertoire for Ig-Seq Analysis. *Front. Immunol.* **2019**, *10*, 899. [CrossRef]
45. Shugay, M.; Bagaev, D.V.; Turchaninova, M.; Bolotin, D.; Britanova, O.V.; Putintseva, E.; Pogorelyy, M.; Nazarov, V.I.; Zvyagin, I.V.; Kirgizova, V.I.; et al. VDJtools: Unifying Post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.* **2015**, *11*, e1004503. [CrossRef]

46. Shemesh, O.; Polak, P.; Lundin, K.E.A.; Sollid, L.M.; Yaari, G. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. *Front. Immunol.* **2021**, *12*, 633. [CrossRef] [PubMed]
47. Smith, N.L.; Nahrendorf, W.; Sutherland, C.; Mooney, J.P.; Thompson, J.; Spence, P.J.; Cowan, G.J.M. A Conserved TCRβ Signature Dominates a Highly Polyclonal T-Cell Expansion During the Acute Phase of a Murine Malaria Infection. *Front. Immunol.* **2020**, *11*, 3055. [CrossRef]
48. Swanson, P.A.; Padilla, M.; Hoyland, W.; McGlinchey, K.; Fields, P.A.; Bibi, S.; Faust, S.N.; McDermott, A.B.; Lambe, T.; Pollard, A.J.; et al. AZD1222/ChAdOx1 nCoV-19 vaccination induces a polyfunctional spike protein-specific Th1 response with a diverse TCR repertoire. *Sci. Transl. Med.* **2021**, *13*, eabj7211. [CrossRef] [PubMed]
49. Gaide, O.; Emerson, R.O.; Jiang, X.; Gulati, N.; Nizza, S.T.; Desmarais, C.; Robins, H.; Krueger, J.G.; Clark, R.A.; Kupper, T.S. Common clonal origin of central and resident memory T cells following skin immunization. *Nat. Med.* **2015**, *21*, 647–653. [CrossRef]
50. Li, N.; Van Unen, V.; Abdelaal, T.; Guo, N.; Kasatskaya, S.; Ladell, K.; McLaren, J.E.; Egorov, E.S.; Izraelson, M.; Lopes, S.M.C.D.S.; et al. Memory $CD4^+$ T cells are generated in the human fetal intestine. *Nat. Immunol.* **2019**, *20*, 301–312. [CrossRef]
51. Bolotin, D.; Shugay, M.; Mamedov, I.Z.; Putintseva, E.; Turchaninova, M.; Zvyagin, I.V.; Britanova, O.V.; Chudakov, D. MiTCR: Software for T-cell receptor sequencing data analysis. *Nat. Methods* **2013**, *10*, 813–814. [CrossRef] [PubMed]
52. Schattgen, S.A.; Guion, K.; Crawford, J.C.; Souquette, A.; Barrio, A.M.; Stubbington, M.J.T.; Thomas, P.G.; Bradley, P. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.* 2021, online ahead of print. [CrossRef] [PubMed]
53. Dash, P.; Fiore-Gartland, A.J.; Hertz, T.; Wang, G.C.; Sharma, S.; Souquette, A.; Crawford, J.C.; Clemens, E.B.; Nguyen, T.H.O.; Kedzierska, K.; et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **2017**, *547*, 89–93. [CrossRef] [PubMed]
54. Borcherding, N.; Bormann, N.L. scRepertoire: An R-based toolkit for single-cell immune receptor analysis. *F1000Research* **2020**, *9*, 47. [CrossRef] [PubMed]
55. Kashima, Y.; Sakamoto, Y.; Kaneko, K.; Seki, M.; Suzuki, Y.; Suzuki, A. Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.* **2020**, *52*, 1419–1427. [CrossRef] [PubMed]