frontiers | Frontiers in Immunology

REVIEW
published: 15 July 2022
doi: 10.3389/fimmu.2022.858057

Machine Learning Approaches to TCR Repertoire Analysis

Yotaro Katayama[1*], *Ryo Yokota*[2], *Taishin Akiyama*[3,4] *and Tetsuya J. Kobayashi*[5,1]

[1] Graduate School of Engineering, The University of Tokyo, Tokyo, Japan, [2] National Research Institute of Police Science, Kashiwa, Chiba, Japan, [3] Laboratory for Immune Homeostasis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [4] Graduate School of Medical Life Science, Yokohama City University, Yokohama, Japan, [5] Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

OPEN ACCESS

Edited by:
Loretta Tuosto,
Sapienza University of Rome, Italy

Reviewed by:
Gur Yaari,
Bar-Ilan University, Israel
Pirooz Zareie,
Monash University, Australia
Steven M. Abel,
The University of Tennessee, Knoxville, United States
David Olivieri,
University of Vigo, Spain
Victor Greiff,
University of Oslo, Norway

***Correspondence:**
Yotaro Katayama
yotaro.katayama@gmail.com

Specialty section:
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology

Received: 19 January 2022
Accepted: 07 June 2022
Published: 15 July 2022

Citation:
Katayama Y, Yokota R, Akiyama T and Kobayashi TJ (2022) Machine Learning Approaches to TCR Repertoire Analysis. Front. Immunol. 13:858057. doi: 10.3389/fimmu.2022.858057

Sparked by the development of genome sequencing technology, the quantity and quality of data handled in immunological research have been changing dramatically. Various data and database platforms are now driving the rapid progress of machine learning for immunological data analysis. Of various topics in immunology, T cell receptor repertoire analysis is one of the most important targets of machine learning for assessing the state and abnormalities of immune systems. In this paper, we review recent repertoire analysis methods based on machine learning and deep learning and discuss their prospects.

Keywords: machine learning, deep learning, T cell, T cell receptor, immunoinformatics

INTRODUCTION

Our bodies are constantly exposed to threats from various pathogenic bacteria, viruses, and cancer cells. The immune system is central to maintaining our body in a healthy state by detecting and evicting those pathogens. Among the different types of immune cells, T cells play various roles in the recognition, memory, and eviction of such threats (1). The peptides derived from those pathogens provide information to T cells when they are presented on the major histocompatibility complex (MHC) as antigens. T cells recognize an antigen if their T cell receptors (TCRs) can bind to the antigen-MHC complex. As antigens are diverse and MHC genes are highly polymorphic, TCRs also must be diverse to recognize a wide range of antigens. TCR diversity is generated by V(D)J recombination (2), one of the somatic recombination processes in our body. This process can potentially yield more than 10^{13} patterns of TCR (3). This diversity of TCRs ensures that, even if unknown antigens enter the body, there will be T cells with TCRs that can recognize them with a high probability. Furthermore, the recognition of such antigens by T cells, i.e., the binding of antigens to their TCRs, activates the T cells, inducing their proliferation and/or phenotypic changes (1). These dynamics alter the diversity of TCRs (TCR repertoire) in a T cell population and modulate its collective recognition of antigens. Therefore, quantitative evaluation of the TCR repertoire in individuals enables us to capture the individual's past and present immunological status. It may also be possible to predict its future. Specifically, quantitative measurement of TCR repertoires may contribute to the quantification of abnormalities in the immune status of patients with specific diseases, the identification of the causes, and prediction of the risk of developing immune-related diseases in the future. For example, a diagnosis for a kind of leukemia is already approved by FDA (U.S. Food and Drug Administration) and clinically used. Quantitative measurement of TCR repertoire is performed by sequencing the recombinant genes encoding the TCRs of T cells in blood or other specimens. Since the mainstream of DNA sequencing technology has shifted from the low-throughput Sanger method to high-throughput next-generation sequencers (NGSs), the cost and time required for sequencing TCR repertoires have been dramatically reduced, which makes it practical to exploit TCR repertoires for practical applications. The recent advent of new techniques such as single-cell sequencing further provides ways to characterize different aspects of T cell repertoires (4).

In parallel with the development of TCR repertoire sequencing technology, bioinformatic and machine learning (ML) based data analysis, including deep learning (DL), is pervading the field of immunology. As we will see in more detail later, this is because a typical sample of repertoire data from a single person consists of a set of several hundred thousand sequences, and ML is an effective tool for extracting information from such a large amount of data. ML is already indispensable to repertoire sequencing analysis. It has also allowed new applications based on the repertoire sequencing such as the personal cancer vaccine (neoantigen vaccine) design (5) and the new testing methods for infections such as COVID-19 (6). Not only clinical applications, but also basic researches are assisted by ML based analysis methods. The impact of ML in repertoire sequencing is rapidly growing.

In this paper, we will outline the rapidly developing TCR repertoire analysis methods based on ML with useful tools and databases. We also discuss possible directions for further development of TCR repertoire analysis.

Diversification of TCR

T cell progenitors are generated from hematopoietic stem cells in the bone marrow and undergo differentiation for maturation in the thymus before being exported to the periphery (1). A TCR is a heterodimer of α- and β-chains (certain TCRs consist of the γ- and δ-chain, but these are omitted here for simplicity). In the V(D)J recombination, one gene is selected from each of the V, D, and J gene groups of pre-recombinant genes of each chain (in α-chain, the D gene group does not exist), and the selected genes are combined with random insertions and deletions. Because of the randomness in the gene selection, insertions, and deletions, a variety of TCRs are generated. For example, in the case of the human β-chain, there are 64-67 V genes, 2 J genes, and 14 D genes according to IMGT database[1]. It should be noted that there are two loci for each TCR chain as humans are diploid. 30% of T cells (dual TCR) have two different productive TCR α-chain mRNAs despite the allelic exclusion mechanisms (7).

A TCR recognizes antigens present on the MHC. Antigens are digested into short peptides and presented on the MHC to form peptide-loaded MHC (pMHC) complexes (8). The affinity of a TCR to an pMHC complex is mainly determined by the recombination-dependent highly variable regions called the complementarity determining regions (CDRs) (9). In the sequence of recombinant TCR genes, three CDRs exist, from CDR1 to CDR3. CDR1 and CDR2 engage in binding to the MHC complex presenting an antigen, whereas CDR3 contributes to the binding affinity of the TCR to the antigen itself. Thus, the sequence of CDR3 plays a particularly important role in analyzing repertoires. Many studies to be introduced here also work on CDR3. After recombination, T cells in the thymus undergo positive and negative selection based on their interactions with self-antigens presented by other cells such as thymic epithelial cells (10). In positive selection, T cells with TCRs that have a moderate affinity to some self-antigen-MHC complexes are selected to survive. This process also selects T cells such that they recognize the antigen only if it is presented on the MHC (11). This phenomenon is called MHC restriction (12). Note that TCRs are "personalized" in this process as the MHC genes are highly polymorphic. The impact of genetic background, including MHC polymorphisms, on repertoire dynamics will be revisited in the next section. Cross-reactivity ensures that the selected T cells may recognize some non-self-antigen-MHC complexes too (13). In contrast, T cells with TCRs that have a high affinity to any self-antigen-MHC complex are eliminated in negative selection. This process decreases the number of self-reactive T cells by 60-70% (14). The remaining self-reactive T cells are suppressed by peripheral tolerance (14). By combining these mechanisms, TCRs that can recognize non-self-antigens but do not recognize self-antigens are selected and exported to the periphery. Then, T cells are induced to differentiate and proliferate depending on the antigens encountered in the periphery. From such peripheral T cell population dynamics, an appropriate repertoire is shaped and maintained so that it attains the ability to remember and rapidly respond to experienced antigens while retaining the diversity to respond to unknown ones (15).

Influencing Factors of TCR Repertoire

Various factors affect the formation of a TCR repertoire. As we described earlier, peripheral antigen exposure changes the TCR repertoire. We review other potential factors in this section.

First, the genetic background can affect diverse aspects of repertoire dynamics. As we mentioned in the positive selection in the thymus, TCRs are selected to have MHC restriction. Therefore, the MHC type can influence the formation of repertoire. For example, associations between specific HLA (human MHC) types and specific sequences are observed (16). Furthermore, gene usage in V(D)J recombination might be affected by MHC (17). In addition, some HLA variants are associated with onset of autoimmune diseases (18). These results contrast with those of immunoglobulin for which V(D)J recombination process before selection is found to be highly different even between monozygotic (MZ) twins (19). Moreover, as HLA genes are highly variant (12), TCRs that bind to the same peptide can differ between people. Therefore, we cannot easily assume that T cells with the same TCR recognize the same antigens in different individuals when genetic information such as HLA is not the same.

Not only MHC, but also V(D)J genes themselves have polymorphisms (20, 21). Some of those variants are shown to affect the affinity of TCR-pMHC complex (22), which may result in different repertoire dynamics. We do not fully understand the effect of these variants on repertoire, as many variants remain to be discovered (23). Furthermore, genetic background is not the

[1] http://www.imgt.org/.

only dominant factor in the final peripheral repertoire dynamics. For example, a study in MZ twins revealed that peripheral repertoires of MZ twins are almost as different as those of unrelated individuals in terms of shared TCRs (24). On the other hand, those of the same person are very similar even after years (24). This might be caused by the fact that the probability of generation of the same TCR in different individuals is very low even if the MHC alleles are the same.

Second, aging also affect repertoire dynamics greatly. Age-related changes in the immune system are collectively called "immunosenescence" (25, 26). In the context of TCR repertoire analysis, immunosenescence often refers to the decrease in the proportion of naïve T cells and the increase in that of memory T cells undergoing persistent selection, for example, memory T cells recognizing antigens behind chronic viral infections such as cytomegalovirus (CMV) (27). This phenomenon impairs the diversity of the TCR repertoire. One of the main causes of this change is the decrease in the thymic output of naïve T cells due to the age-related thymic involution (28).

Repertoire Sequencing and Batch Effects

We can quantify TCR repertoires through repertoire sequencing (AIRR-seq) using NGSs. A typical repertoire sequencing procedure is summarized in **Figure 1A**. Samples such as peripheral blood mononuclear cells (PBMC) are collected, and their CDR regions are amplified by polymerase chain reaction (PCR). Then NGSs are used to read the amplified sequences. As CDR3 is the most diversified region in the TCR gene, many protocols are developed for CDR3 sequencing (29, 30).

Repertoire sequencing is one of the most actively developed sequencing technologies. In addition to the conventional procedure described above, single-cell repertoire sequencing has also been developed in recent years (4, 31). Using such protocols, for example, the pairing of the TCRα and TCRβ chains can be measured (4, 31). Furthermore, dual TCRs can be investigated (32–36). As this review is primarily dedicated to repertoire analysis methods, we focus mainly on the potential biases in the conventional sequencing methods, which may skew the results of the ML methods.

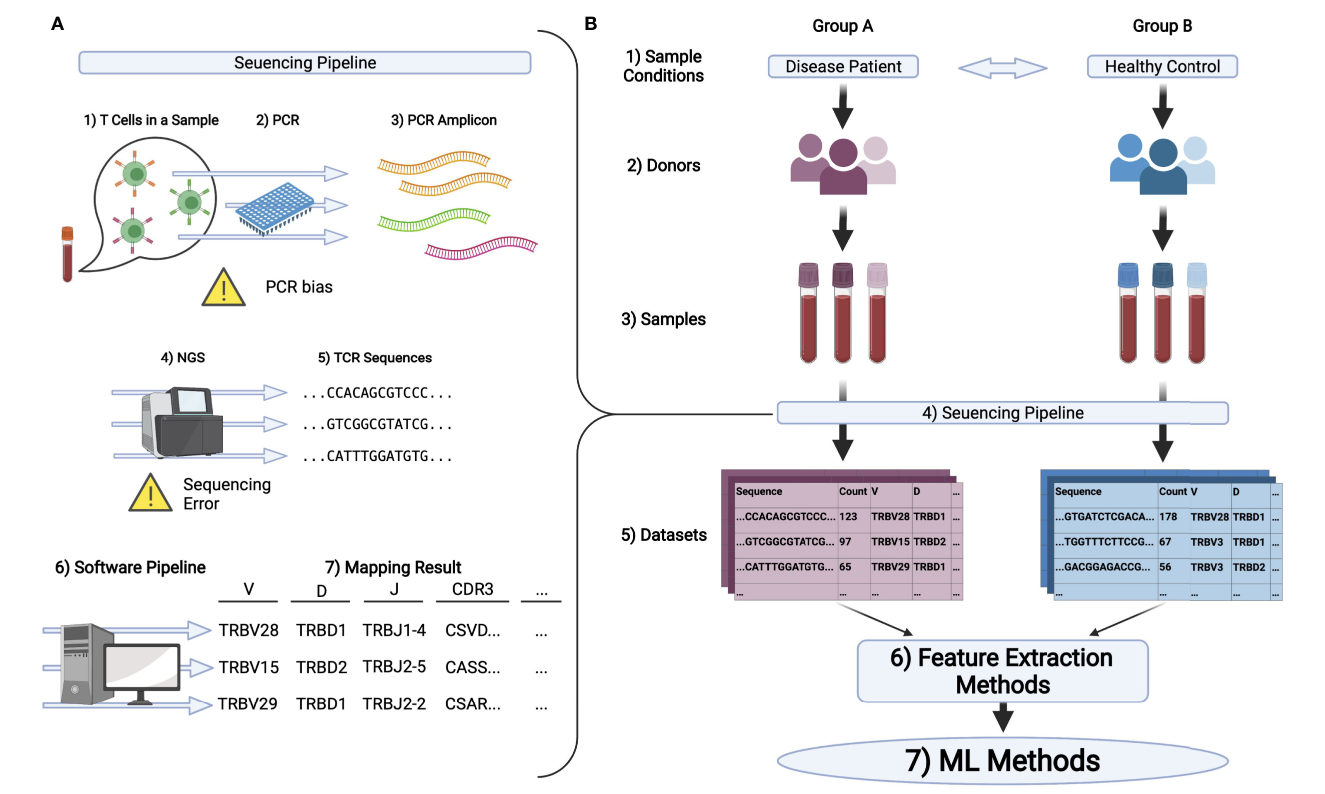

FIGURE 1 | **(A)** Schematic illustration of the pipeline in TCR repertoire analysis. 1) First, T cells in samples, typically being collected from peripheral blood, are processed to extract its DNA or RNA of TCRs. 2,3) PCR is conducted to amplify the signal. 4,5) Then, the amplified DNA or cDNA is sequenced by NGS to obtain TCR sequences. 6,7) Finally, these sequences are mapped to the reference genes by the software pipeline introduced in the main text and analyzed further. **(B)** A typical experimental flow for applying ML methods on repertoire datasets. 1-3) Samples are collected from multiple groups of donors who have different immunological and physiological conditions. 4,5) By the pipeline illustrated in **(A)**, the dataset is obtained for each sample typically in the format of a table or matrix. 6) Datasets are encoded to ML friendly formats (feature vectors) using feature extraction methods. In bioinformatics, it is common to analyze gene expression matrices, which summarize the expression level of each gene for each sample. In repertoire analysis, for each sample, we have a matrix, each raw of which represents the sequence of one TCR, its observation count, its gene usage, and other properties of the TCR. Note that typically 10^4 to 10^5 different sequences are observed per sample and that only a limited number of overlapping sequences are usually detected among samples. Therefore, a relatively large sparse matrix must be handled for repertoire analysis. 7) ML algorithms are performed on the encoded datasets.

First, PCR introduces various biases originating from the amplification. The sequence composition influences the amplification ratio of PCR. Multiple primers are also a source of biases. Multiple primers are commonly used in repertoire sequencing (37) because the edges of the CDR3 region are diverse depending on the choice of V (and J) genes. These primers are designed for known V (and J) genes. As a result, CDR3 sequences composed of unknown V (and J) alleles may not be amplified (30). In addition, multiplex PCR is also influenced by the amplification bias (30). Such quantitative bias affects a variety of ML methods introduced later. For example, diversity-based methods in observation frequency-based methods can be directly skewed. Various proliferated clonotype discovery methods are also affected.

Second, PCR and NGS introduce errors in the TCR sequences (29, 38, 39). It is estimated that about 2% of the PCR amplicons contain some sequencing errors in TCR sequencing (40), and 1-6% of sequences yielded by NGSs (Illumina) are erroneous. Erroneous sequences lead to false-positive clusters and skew the diversity in observation frequency-based methods. In contrast, dissimilarity-based methods aggregate similar sequences into a cluster. Therefore, they can be less affected by such errors.

Third, the starting material matters. We can employ either DNA or RNA of TCRs. In general, DNA-based methods are supposed to be more quantitative than RNA-based ones, as the number of RNA copies fluctuates among cells (30). However, a recent systematic review (41) suggests that the starting material may not always be the determinant of correctness or sensitivity. Moreover, RNA-based methods have some qualitative advantages. For example, some RNA-based protocols can capture the full-length TCR sequence, which contains CDR1,2 and 3 (30).

Many protocols have already been proposed to reduce such biases and errors. Certainly, their magnitudes can differ by protocol (29, 38). However, each method have both advantages and disadvantages. To apply ML methods to any data, we need to mind the protocol used to derive the data and be aware of the introduced bias beforehand. The following reviews are referred for details of each protocol (30, 31, 37).

Moreover, repertoire sequencing is affected by various batch effects. As we reviewed in this section, the choice of experimental protocol affects the result. Even if the protocol is the same, various conditions, such as different batches and different facilities where samples are collected, can be distinguished by ML algorithms (42, 43). These batch effects can be problematic in applying machine learning because of shortcut learning (44). We here adopt a famous example from medical image processing to intuitively explain the concept of shortcut learning. In the pneumonia detection task from an X-ray image, the performance of ML models is known to be dropped if tested by the datasets from other hospitals. It was revealed that ML models seemed to distinguish from which hospital an image was taken (45). As every hospital has different pneumonia prevalence rates, the model outputs positive if the sample seemed to be taken in a hospital with a high prevalence rate and can achieve a decent performance score. However, of course, if an image is not taken at the known hospitals, the model cannot answer correctly. In this situation, the hospital classification task was easier and was thereby used as a "shortcut" for the pneumonia detection task. As ML can distinguish various experimental conditions because of the batch effects, shortcut learning can also happen in ML-based repertoire analysis. There are attempts to remove the batch effects in repertoire sequencing. Some of errors and biases can be corrected by bioinformatic post-processing (29, 38, 40, 46). Such algorithms are implemented in popular software such as MiXCR (47). They are successful in reducing errors and biases (40). However, we have to be aware that the batch effects may not always be corrected. Thus, we must be careful when applying machine learning methods to repertoire datasets. For detailed comparisons of software, refer to (40, 46).

Current Pipeline and Datasets for TCR Repertoire Analysis

Currently, a variety of TCR repertoire datasets are available to the public. There are two main types of platform hosting repertoire datasets. The first one is a public database, Sequence Read Archive (SRA)[2], to which we can register raw sequences (e.g., FASTQ files). To download data, users need to find the accession number of International Nucleotide Sequence Databases (INSD)[3] and use software such as sra-toolkit[4]. Each read sequence in a FASTQ file generated by NGSs is mapped to the reference sequences to annotate CDRs and selected V(D)J genes. Several pipeline tools for the analysis of FASTQ files have been proposed and developed, among which IMGT/HighV-QUEST (48), igBLAST (49), and MiXCR (47) are popularly used in previous studies. For performance comparisons of the major tools, we refer the reader to these review articles (50, 51). This workflow is summarized in **Figure 1A**.

The other is the platforms dedicated to immunosequencing datasets. For example, VDJServer[5] (52) and immuneAccess[6] have been widely used in recent years. Once FASTQ files are uploaded to these services, they will automatically process the files, and various analyses can be performed on the web. Such services seem to be highly appreciated by emancipating users from setting up a local environment or being bothered by complex software options. Still, there is no de facto standard for such repositories, and this has led to the development of curated databases such as iReceptor[7] (53) and TCRdb[8] (54) for scattered datasets.

To efficiently collect information on TCR repertoire analysis, it is also recommended to use other major repositories and communities as follows; VDJdb (55), a database that combines information on TCRs, antigens, and MHCs; Immune Epitope Database (IEDB) (56), a database of immune epitopes; McPAS-TCR (57), a database that organizes and collects TCR sequences related to various pathogens; and Adaptive Immune Receptor Repertoire (AIRR) community (58), a community for sharing antigen and repertoire datasets.

[2] https://www.ncbi.nlm.nih.gov/sra/.

[3] https://www.insdc.org/.

[4] https://github.com/ncbi/sra-tools.

[5] https://vdjserver.org/.

[6] https://clients.adaptivebiotech.com/immuneaccess.

[7] https://gateway.ireceptor.org/.

[8] http://bioinfo.life.hust.edu.cn/TCRdb/#/.

Challenges in TCR Repertoire Analysis

In general, the basic approach for extracting useful information by comparing samples with others of different conditions is to contrast the information shared or not shared between samples in the same condition or those across different conditions like **Figure 1B**. In TCR repertoire analysis, the TCR sequences commonly observed among different individuals (called public TCR sequences) are considered important (59). However, due to the diversity of TCR repertoires, the number of public TCR sequences is very small compared to the total number of sequences observed in each sample. Therefore, they may not be sufficient to characterize the differences among the sample groups.

In addition, for sequences observed uniquely in each sample, it is not easy to distinguish whether they are attributed to the differences of individuals or to those of sample conditions such as abnormalities or diseases. The difficulty of associating observed sequences with sample conditions is one of the major problems in repertoire analysis due to the diversity of TCRs. Moreover, the TCR repertoire may change over time due to the donor's health history such as injury, infection, and aging (1). Thus, the individual difference in TCR repertoire is large. Therefore, we should perform the analysis by taking into account not only the current but also the past health condition of the donors.

Furthermore, T cells isolated from peripheral blood samples are commonly used to measure human TCR repertoire. From each sample, we typically measure TCRs of about 10^4 to 10^6 T cells, which is a tiny fraction of the donor's approximately 10^{11} Tcells (60, 61). Thus, quantitative analysis of TCR repertoire has its own difficulties due to the diversity and chronological variation of TCR repertoire and also to the high population size of T cells compared with the measurable size.

Moreover, we still cannot directly know what antigens a specific TCR sequence recognizes. Therefore, only from sequence information, we cannot compare or measure the similarity of TCRs by their antigen recognition profile. Experimental analyses of antigen-specific TCRs are widely performed (62–64), but they cannot be exhaustive, and we cannot conduct such analysis on every TCR. Although we have TCR-pMHC binding prediction methods, some of which we review later, the performance is still limited. In addition, the diversity of TCRs, MHCs, and antigens makes it impractical to calculate the complete recognition profile. This is problematic because we may need to utilize similar but different sequences to compare or characterize repertoires, as the number of shared identical TCRs is very small because of the high variety of TCRs and the limited sample size we mentioned earlier. Although a lot of experimental evidence such as (64, 65) implies that similar TCR sequences may recognize similar antigens, there is no *a priori* similarity measure. We need to devise a new way to calculate the similarity of TCRs.

These challenges are not the only obstacles in TCR repertoire analysis for understanding the dynamics of TCR repertoire. As we saw earlier, experimental procedures for repertoire sequencing using PCR or NGS inevitably introduce batch effects and errors. Some of the software tools introduced in the previous section correct and debias the sequencing data to some extent. However, not all the errors and batch effects can be removed.

These problems summarized in **Figure 2** are related to the development of bioinformatic and ML methods to be introduced in the next section. Each method approaches to these challenges in a unique way, which can be categorized as in **Figure 3**. In the following sections, we review each category one by one.

OBSERVATION FREQUENCY BASED METHODS

In the conventional analysis of TCR repertoires, statistical indices from ecology have been employed. In ecology, the complexity of ecosystems has been measured by diversity (66). Typically, diversity is calculated by the rarity weighted count of the species. If a species has a dominant population, the diversity of the ecosystem is small. In contrast, if there are many rare species, the diversity increases.

By treating each TCR clonotype as a species, diversity can be measured for a TCR repertoire in a sample. In immunology, the diversity of TCR repertoire is closely related to the clonal expansion (an increase in the proportion of T cells with the same TCR clonotype caused by a proliferation of T cells, which decreases the diversity of TCR repertoire) against specific antigens (1). By applying the species diversity analysis methods in ecology, the degree of clonal expansion has been associated with various sample conditions. A typical example is an approach to quantify the diversity of amino acid sequences in CDR3 using indices such as Hill's number (67). Around 2010, the quantification of the diversity of TCR repertoires using probability models were proposed (68), which enabled us to characterize differences between samples. Both Guindani et al. (69) and Rempala et al. (70) employed the Poisson abundance model (68), to not only fitting the abundance distribution shapes, but also to classify the samples using the estimated parameters of each sample. This approach is still being investigated: PowerTCR (71) proposed a probabilistic model not based on Poisson abundance model in 2018.

However, these approaches employ only frequency information and do not directly utilize the sequence information. As a result, important information can be obscured or lost. For example, even if the samples have very similar frequency distributions, the sequences observed at high frequencies might be completely different. In particular, it is difficult to examine or identify particular sequences that caused the differences between samples only from the frequency information, which is important for practical applications. Moreover, utilizing ML methods enabled the processing of sequence information without compressing it down to the frequency information. Therefore, the recent advances in TCR repertoire analysis have occurred primarily in sequence-information-based methods using ML methods. We categorized them by their approaches, as summarized in **Figure 3**, and will review each of them in the following sections. Nevertheless, frequency-based methods are still

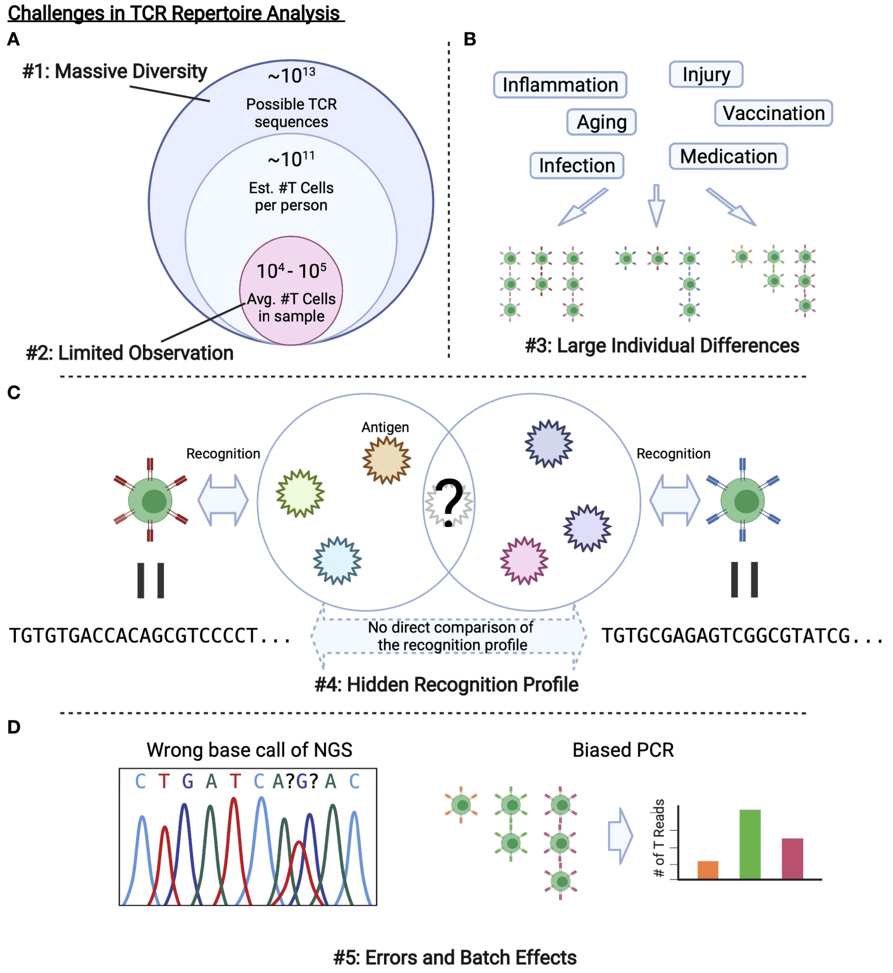

FIGURE 2 | Challenges in TCR repertoire analysis. **(A)** Only limited observations are possible compared with the massive diversity of TCRs in a body or that of possible TCR sequences. **(B)** Various factors alter the repertoire, which results in large individual differences. **(C)** As we cannot observe all the antigens that a TCR recognizes, we cannot directly evaluate the similarity between TCRs with different sequences. **(D)** Experimental procedures including PCR and NGS inevitably introduce errors and batch effects.

practically effective for small datasets as the frequency distribution can be obtained relatively stably even from a sample containing a small number of TCRs.

UTILIZATION OF SEQUENCE INFORMATION

As we saw in the previous section, frequency distribution based methods can only provide the degree of difference between different samples. In particular, specific sequences characterizing the sample differences are of particular importance. Since we are interested in specific sequences characterizing the sample differences, we need another approach that can directly utilize sequence information to identify those specific sequences. One of the most illustrative and important applications of sequence information is monitoring minimum residual disease (MRD), a kind of T(B)-cell leukemia (72). As dominant T(B) cell clones themselves are the direct cause of MRD, unusually proliferated TCRs (BCRs) can be utilized as biomarkers to monitor the progression of the disease. In monoclonal leukemia, the identification of such dominant sequences is fairly easy because the dominant clone sometimes occupies more than 75% of the T cells (73).

However, we may not be able to find such obvious sequences for other diseases or conditions. Unlike leukemia, the most abundant clone in a sample may not be related to diseases or conditions. We must find a portion of the sequences shared between the samples of the same condition, but this is not straightforward. Due to the diversity and individual differences of TCR repertoires, the number of shared sequences is typically very small. Even if we find shared TCRs, we must statistically discriminate whether such shared TCRs are yielded by a

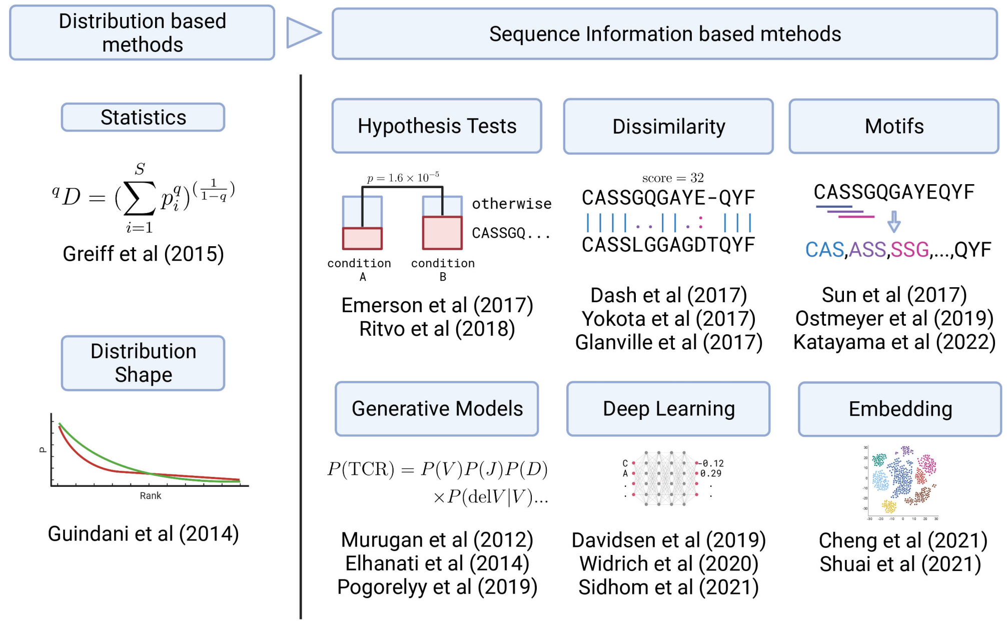

FIGURE 3 | Graphical summary of the development in TCR analysis methods. Early analysis was based on the TCR clonotype abundance (frequency) distribution (left panel). Recently, sequence information has started to be utilized in various ways (right panel) by employing statistical, ML, and DL methods.

condition or by chance. It should be noted that we need to devise a way to evaluate "similar" sequences because we cannot directly observe the similarity of TCRs. To overcome these problems, ML has been utilized. In this section, we review three major categories of methods, which are based on hypothesis tests, dissimilarity, and motifs, respectively.

Hypothesis Test Based Methods

One of the straightforward ways to extract relationships of specific TCRs and sample conditions is to use hypothesis tests to judge whether the number of observed TCRs is significantly large or small for a specific sample condition. For example, Emerson et al. (16) collected peripheral blood samples from a total of 641 donors, 289 affected and 352 unaffected by CMV, and identified the TCRβ sequences specific to the CMV-affected donors using Fisher's exact test. In addition, by utilizing the identified 164 CMV-specific TCRβ sequences as features of a repertoire, they designed a discriminative model of beta-binomial distribution for predicting CMV infection. De Neuter et al. (74) replicated the results on another dataset and showed that a random forest classifier using the observation counts of these TCRs in a sample also works well to predict the infection. Emerson's method, however, ignores sequence similarity completely and only utilizes the information of "Public TCRs."

In contrast, Ritvo et al. (75) proposed a method called TCRNET, which utilizes sequence similarity to estimate clusters of similar TCRs that are significantly proliferated in specific samples. Here, similar TCRs are defined as those derived from the same V and J genes and differ at most by one amino acid sequence. Then, the number of TCRs in the target cluster is contrasted with the number of TCRs with the same V, J genes and CDR3 sequence length as the target cluster. If the proportion is found to be significantly larger in a specific sample by the binomial test, the target cluster is judged as a proliferated cluster.

These methods require counting the same or similar TCRs. This process is very slow because, in a naive implementation, every possible pair of sequences must be compared. CompAIRR[9] is developed for a faster exact or approximate search for shared TCRs.

Dissimilarity Based Methods

The methods introduced above compare only a specific TCR or a cluster of TCRs with the others. Therefore, they abandon the sequence information of the others, even though they constitute most of the sequences in samples. Some methods have been proposed to exploit such information. In particular, we review the methods based on the dissimilarity between TCRs. Network analysis based on sequence similarity has been used for a long time. For example, classification of healthy and leukemic samples is performed on the BCR sequence network of each sample in which all sequences differ at most one residue are connected (76). In TCR, a similar network analysis revealed the public TCRs conserved between mice and humans (77).

More complex dissimilarity indices tailored for TCR analysis have been proposed. Dash et al. (78) quantified the differences between two TCR sequences by weighted Hamming distances and visualized epitope-specific TCR clusters by dimensionality reduction and clustering of their dissimilarity matrices. Their method, called TCRdist (78), has become a popular method to search for epitope-specific TCR sequences. TCRdist focuses

[9] https://github.com/uio-bmi/compairr.

mostly on evaluating the differences of TCRs. By contrast, RECOLD (79), which was proposed by our group, is designed to measure the differences between samples. RECOLD calculates the distance between all the observed sequences in all samples to create a dissimilarity matrix. Then, dimensionality reduction is performed on the matrix, and every observed sequence is embedded in a shared low-dimensional space as a point. In this space, each sample is represented as a probability distribution, and the difference between samples is quantified as the difference of distributions by Jensen-Shannon Divergence. In addition, RECOLD can identify the sequences specifically contributing to the differences of samples using the bootstrap method.

New methods based on TCR-level dissimilarity are still actively and continuously explored. A method called GLIPH (63) integrates sequence information and observed frequency information with CDR3 length and HLA to estimate epitope-specific sequences. iSMART (80) and GLIPH2 (81) have been released recently to improve the performance and the applicable data size. In TCRdist3 (82), TCRdist-based distance can be combined with motifs (introduced in the next section) to characterize TCR clusters.

On the other hand, some methods are devised for calculating the distances between repertoires directly. Repertoire Dissimilarity Index (RDI) (83) compares the usage of V(D)J gene segment. ImmuneREF[10] utilizes various interpretable indices such as diversity indices and positional amino acid frequencies.

As in the case of the hypothesis-based methods, the computation cost is important for dissimilarity-based methods, which also perform a lot of sequence comparison. ClusTCR (84) achieves faster clustering by focusing CDR3 and compromising flexibility in the sequence alignment. GIANA (85) used a different approach. In GIANA, a lightweight linear transformation equivalent to sequence alignment on BLOSUM62 is constructed. Then, every sequence can be encoded into a coordinate in the euclidian space, where clustering is fast.

Motif Based Methods

The dissimilarity based methods characterize TCRs (or samples) by the relative distances between them. Alternatively, we can directly encode TCRs (or samples) into feature vectors and apply ML methods to the vectors. A conventional but effective method to create such feature vectors is the k-mer method. It characterizes a TCR (or a sample) by the observed frequency of all possible k consecutive substrings (motifs) in the sequence (or sequences in the sample). Therefore, in a typical 3-mer method, its feature vector has approximately 21^3 dimensions (21 = 20 human amino acids + a symbol representing the edges of the amino-acid sequences). The k-mer features have been combined with various ML methods: LP-boost (86); Bayesian discriminators (87); and SVMs (87). They were applied to TCR β-chain CDR3 datasets to discriminate whether a sample had been treated with ovalbumin or not, which was used to stimulate immune responses. We also proposed MotifBoost (43), which merges the k-mer encoding and Gradient Boosting Decision Tree (GBDT) (88) for repertoire classification. Along with proposing a new method, we also investigated the nature of the k-mer encoding and revealed that CMV infected and healthy samples are well separated in the k-mer feature space derived by a PCA-like unsupervised learning method called Gaussian Process Latent Variable Model (GPLVM) (89). This result indicates that the k-mer encoding can naturally capture the intrinsic characteristics of repertoires. Moreover, k-mer based methods work effectively even on smaller samples compared to the other methods.

As we mentioned earlier, we still do not fully understand what kind of factors determine the similarity between different TCRs. However, the success of the dissimilarity-based methods, which is based on the hypothesis that similar TCRs work similarly in the body, implies that the hypothesis is true to some extent. Moreover, the success of k-mer encoding support and strengthen the view that some important motifs play a central role in determining the similarity of TCRs. This is also supported by the fact that shared motifs of antigen-specific TCRs are found in various conditions (62–64).

While being conventional, k-mer encoding and combined ML methods still have room for further improvement and development. For example, Ostmeyer et al. (90) combined the 4-mer method with logistic regression to discriminate between cancerous and healthy tissues. In this work, feature vectors are created differently from the conventional way. Each 4-mer motif is represented as a 20-dimensional vector consisting of four 5-dimensional biophysicochemical feature vectors of each amino acid. Therefore, a TCR is converted into a bag of 4-mer feature vectors. To deal with this setup, they employed the multiple instance learning framework. Specifically, they trained a logistic regression model to assign a score, which is the probability that the motif is related to cancer, for each motif. A sample's score, which is used for sample-level classification, is defined as the maximum score of the motifs found in the sample.

As a good representation of data is decisive in ML, we expect that more applications appear, which are built around k-mer methods or other data representation methods.

APPLICATION OF GENERATIVE MODELS

Most of the methods mentioned above are used for characterizing the differences between samples. Thus, they usually compare samples obtained from different conditions by assuming that the dataset to be analyzed is from a cross-sectional or longitudinal study. However, careful effort is required for obtaining datasets from multiple experiments. Recruiting a sufficient number of donors for every sample condition is difficult, especially if they are rare.

To solve this problem, methods based on generative models have recently been explored. These methods employ mathematical models for the generation of TCRs, which have been intensively developed since 2012. For TCR generation in

[10] https://github.com/GreiffLab/immuneREF.

the thymus, a probabilistic model implementing the biological mechanism of V(D)J recombination was proposed (91). Various extensions to this model, especially for inference methods, have been proposed based on Monte Carlo simulation (92), improved expectation maximization (EM) algorithm (93), and dynamic programming (94). For TCR selection in the periphery, Elhanati et al. (95) devised another probabilistic model. Their model employs the actual peripheral repertoire dataset to estimate the probability distribution of post-selected TCRs, and utilizes the TCR generation model of (91) to infer that of unobserved pre-selected TCRs. This model is trained to predict the difference between the two distributions.

Based on the same idea of substituting the unobserved datasets with a generative model, Pogorelyy et al. (96) developed a method called Antigen-specific Lymphocyte Identification by Clustering of Expanded sequences (ALICE), which can characterize samples obtained from only one condition by contrasting them with the sequences generated by a generative model as reference repertoires. This strategy is also applied to characterizing TCRs (92).

The generative model can pave the way to quantify the abnormality of a sample and to infer its responsible sequences only from a snapshot sampling of the patient's repertoire, without expensive effort to conduct cohort studies. However, challenges remain for its practical and reliable employment. For example, because the TCR generation model utilized in ALICE does not take into account the individual difference that affects the TCR repertoire [e.g., genetic background (97) and age (26)], the parameters of the generative model may need to be adjusted to the conditions of individual samples to further enhance its reliability.

Simulation of Repertoire

The advance of generative models leads to the emergence of some simulation software, which create pseudo repertoire datasets. Simulated datasets have been used to assess the performance of repertoire analysis methods. For example, a simulated dataset was used to assess the performance of the V (D)J genes identification for B cells (98).

IgSimulator (99) is one of the earliest repertoire dataset simulators. AbSim (100) simulates the temporal development of mutations in B cells. However, these simulators were made for antibody sequences, not TCR sequences. ImmuneSIM (101) is capable of simulating TCR repertoires. In addition, its remarkable feature is the simulation of repertoires for classification. It can implant k-mer like sequences into the repertoire dataset. Classification methods can be tested whether they can find the implanted TCR or repertoire or the implanted motif itself. As motifs play an important role in characterizing repertoires (see motif-based methods), k-mer like signal implanting is recently adopted in some studies (102, 103).

Using simulation, further evaluation of analysis methods can be performed. For example, the classification performance was evaluated in various conditions with different density of signal, sample sizes and so on as done in (103). Evaluations like this cannot be conducted using only real datasets.

APPLICATION OF DEEP LEARNING

Deep learning (DL) is a class of ML algorithm, which achieves good performance in various fields. DL has been pervading various areas of biology such as genomics (104) and systems biology (105), and it has also recently been applied to repertoire analysis. Again, DL itself is just another ML algorithm. However, representation learning, which is one of the notable features of deep learning, allow DL models to achieve high performance by learning appropriate representations from data without explicitly providing the mechanism behind it (106). On the other hand, most of the models we introduced earlier used hand-crafted features or were based on the human knowledge. We call such models "hand-crafted model" hereafter. While the generative model of TCRs introduced above is a hand-crafted model that explicitly implements biological mechanisms such as V(D)J recombination, Davidsen et al. (107) proposed a Variational Auto Encoder (VAE) (108) based generative model that treats the TCR generation like a string generation task. Another feature of DL is that representations learned in one task can be easily transferred to other tasks [called transfer learning (109)]. DeepTCR (110) solves classification problems using features obtained from a VAE-based generative model.

Not only generative models like VAE but also discriminative models are utilized for repertoire analysis. For example, DeepRC (102) utilized a popular class of DL model architecture called attention mechanism for the repertoire classification problem. Simply put, the attention mechanism is a kind of learnable weighted average (111). DeepRC encodes each amino acid sequence in the repertoire to a vector and analyzes its importance through the attention mechanism. Classification is made on the weighted average of the encoded vectors.

DL is also being intensively applied to the prediction of affinity between pairs of T cells and antigens (112, 113), as well as triplets including MHCs (114). TCR-pMHC binding prediction task is one of the most actively studied topics in immunoinformatics (115). The task is to predict whether or not the target antigen will be recognized by a TCR using the sequence information of the TCR and the antigen protein. As Alphafold2 (116) has made an innovation in predicting the structure of proteins from their amino acid sequences, DL is expected to make a breakthrough in this area.

At this stage, DL-based methods have not yet demonstrated the performance to dominate hand-crafted models, in which human crafts the feature or the model structure, in this field. For example, a comparison between a hand-crafted generative model (95) and Davidsen's VAE-based generative model (107) was conducted (117). This paper concluded that the hand-crafted model outperforms DL-based models with lower computational cost and higher interpretability. For peptide-MHC binding prediction, according to a systematic performance comparison review conducted in 2020, ML-based models still scored better than DL-based models on average (118). In addition, our group compared a DL model and ML models by changing the available data size for learning on a repertoire classification task and found that the performance of the DL-based model deteriorates on the small datasets (43).

According to the current trend, the application of DL in this field will be investigated even more intensively in the future. For example, some more recent DL-based peptide-MHC methods reviewed in the next section are showing better performance than the traditional methods on some specific datasets. However, DL may not wipe out the need for traditional biological hand-crafted models because of its expensive computation cost, lack of interpretability, and data-intensive nature. Instead, the integration of hand-crafted and DL-based models is being explored. In a recently proposed model for T cell selection called soNNia (119), a hand-crafted generative model for TCR generation probability (95), which was used for comparison in (117), is combined with a DL model of the TCR selection. For TCR-pMHC interaction prediction, a combination of DL and traditional ML methods is also being pursued (120).

Embedding Methods Based on Representation Learning

In the recent advances in Natural Language Processing (NLP), self-supervised representation learning draws attention, which utilizes the nature of data as a target signal to learn good representations. This is realized by the ability of DL to acquire good representations mentioned in the previous section. One of the earliest successful approaches is Word2Vec (121), which encodes a word to a numeric vector (Word Embedding). In a Word2Vec training method called CBOW (continuous bag of words), a neural network (NN) that converts a word to a vector is trained to predict a masked word in a sentence using encoded vectors of its surrounding words (122). Word2Vec is utilized widely to convert textual data to numerical representation in NLP and also is applied to repertoire analysis. Immune2Vec (123) is inspired by Word2Vec and treats a TCR/BCR as a sentence and a k-mer as a word, respectively. Representation of a TCR/BCR, which is composed of many k-mers, is derived by averaging all k-mer vectors, which is a similar procedure to FastText (124) in NLP.

After the success of Word2Vec, various NN architectures for self-supervised representation learning in NLP are developed. One of the noticeable approaches is neural language models. A language model is a generative model to predict words from the context. CBOW is a representative example which predicts a word from context words. Thanks to the invention of a new NN building block called Transformer (125), which utilized the attention mechanism we mentioned in the previous section. NNs can handle more distant dependencies in a text. New neural language models like BERT (126) exploited the Transformer's ability and broke the former models' records in various tasks. These models are trained to predict a masked word similarly to CBOW. However, in contrast to CBOW, they can predict one or more meaningful sentences, not a word. One such language model called GPT-3 can write natural texts, e.g., news articles (127). We can also utilize a neural language model to embed a sentence using the output of the hidden layer (Sentence Embedding). Such sentence embedding is revealed to be a very good representation and can be applied to multiple downstream tasks in NLP, from question answering to translation, with little additional training for each task (called fine-tuning) (128). Training of the language model itself (called pretraining) requires a large corpus and enormous computation resources. However, once the training is done, the same model can be applied to various problems with fine-tuning using little data.

Language models have also been employed in repertoire analysis. Before that, language models have been intensively applied to general protein sequences (129–132). BERTMHC (133) showed utilizing the pre-trained model of (129) actually increases the performance in the peptide-MHC (Class II) binding prediction task. ImmunoBERT (134) used the same pre-trained model for the peptide-MHC (Class I) binding prediction task. Hashemi et al. (135) employed the pre-trained model of (131) and fine-tuned them for peptide-MHC (Class I) binding prediction and achieved higher performance compared to a previous software. Some papers perform pre-training on their own on the repertoire sequencing dataset. In Leem et al. (136), each amino acid in a TCR is treated as a word, and a TCR is treated as a sentence to pre-train a BERT language model (AntiBERTa). AntiBERTa achieved a higher ROC-AUC in a paratope prediction task than other tools.

The utilization of language models is not limited to embedding. In Shuai et al. (137), another language model called GPT-2 (128) is utilized for pretraing on an antibody generation model (IgLM). Because GPT-2 is designed for full sentence generation, unlike BERT, IgLM can generate new antibodies (CDRs). A new antibody design workflow is proposed in the paper and outlined as follows: First, many antibodies are created using IgLM. Then the3D structure for each antibody is calculated. Finally, the properties of the generated structures are computed to select better antibody candidates.

MACHINE LEARNING FOR REPERTOIRE ANALYSIS IN PRACTICE

In this review, we focused mainly on the technical aspects of ML and DL methods and categorized them by their approach. As a result, we cannot cover all topics, especially those being relevant to practical applications. This may be compensated by a thorough review of the repertoire analysis methods before 2019 in (138), and another review that introduce many methods categorized by task (139). In addition, more ML applications can be found on the pMHC-epitope analysis in (140–143), and on longitudinal analysis in (144, 145).

To practice ML methods, we can refer to the author's implementation in most cases. We can find a comprehensive list of such implementations and other software in (146). In addition, there exist some libraries that implement multiple popular methods to be used for general analysis. In particular, VDJTools (42) and tcR (147) (Immunearch[11] is its successor) are equipped with a broad range of basic analysis methods and are widely used in practice. Moreover, new libraries are being

[11] https://immunarch.com/.

developed such as ImmuneML (148), which focuses more on ML methods.

As for the topics that those sources cannot fully cover, we discuss the following two topics in relation to the practice of ML methods in TCR repertoire analysis: One is prospective practical applications of repertoire analysis, such as blood testing and cancer vaccination. The other is repertoire analysis of COVID-19

Applications of Repertoire Analysis

Recently, applications of repertoire analysis have been developed rapidly. One of the most prominent applications is blood testing (149). In this field, the diagnosis of MRD (see UTILIZATION OF SEQUENCE INFORMATION) and the COVID-19 testing (see the next section) are already approved by FDA. There are potentially more diseases that can be diagnosed by repertoire sequencing. For example, autoimmune diseases such as lupus erythematosus (150), rheumatoid arthritis (150), and lupus nephritis (151) have been successfully classified with the V-J gene usage distribution feature and a random forest classifier. In the BCR repertoire, IGHV gene selection was analyzed for multiple autoimmune diseases (152).

In relation to autoimmunity, repertoire analysis revealed the features common to self-reactive T cells. Hydrophobic residues (153, 154) or Cysteine (154) on CDR3 are related to their self-reactivity. Hydrophobic CDRs enrichment in regulatory T cells is replicated by a logistic regression model with 606 T cell features to predict whether a cell becomes a regulatory T cell or not (155). Prediction of self-reactive T cells may play an important role in the diagnosis of autoimmune diseases in the future.

Another prominent application is neoantigen vaccines to treat cancer. Neoantigen is a tumor-specific antigen that can be used to target tumor cells. Thus, neoantigen vaccines stimulate T cells to attack tumor cells. Neoantigen vaccines should be personalized because tumors of different individuals tend to acquire different mutations and express different neoantigens (156, 157). Repertoire analysis is expected to reduce the labor required for finding individual neoantigen (158). The finding of neoantigens *in silico* is typically performed as follows: First, tumor-specific mutations and their transcripted proteins are identified by sequencing. Second, from those proteins, all antigenic peptides that mark cancer cells are listed. Third, the peptides that can bind to the patient's MHC well are screened. Finally, the obtained peptides are tested to determine whether the pMHC complex can be recognized by T cells or not. Repertoire analysis is used in the third step to predict the affinity of peptide and personal MHC. A couple of software was published for this task (118). On the other hand, immunopeptidome is studied as a different approach to find neoantigens (159). This approach is also interesting in relation to repertoire analysis. In this approach, TCR-pMHC complexes in tumor tissues are collected and analyzed to retrieve their peptide sequences. As the peptides are already assured to bind to MHC, we can skip some of the described screening process. Immunopeptidome can be seen as a peptide repertoire, and its analysis might provide insight into TCR repertoire in the future.

We reviewed some potential applications of repertoire analysis in this section. To realize such applications, we need reproducible and robust results. For clinical applications, standardized protocols must be established. For example, a standard experimental protocol is proposed for MRD diagnosis (160). Also, bioinformatic pipelines are not yet standardized. We will expect more standardized workflows to appear in the future. An example is a new standard format for repertoire dataset proposed by AIRR Community (161).

Repertoire Analysis for COVID-19

Understanding COVID-19 has been one of the most important research topics in recent years, and repertoire analysis has revealed various characteristics of COVID-19 so far. In this section, we will see how the ML-based repertoire analysis introduced in this review is used in the COVID-19 study.

Repertoire analysis has been employed to investigate the nature of COVID-19 infection. Most basic observation is the change in diveristy. Many studies reported the low TCR repertoire diversity in active COVID-19 patients (162–165). Some studies further reported that the severity of the symptom is related to the lower diversity (163, 166). However, it should be noted that decrease in TCR diversity is not necessarily specific to COVID-19 infection but common to various virus infections (164). Cheng et al. (167) investigated V(D)J gene usage and found that some Vβ genes, which are estimated to have a high affinity to SARS-Cov2 spike protein antigen, were enriched in severe COVID-19 patients.

Further insights are also provided by using sequence information based ML methods. In Simnica et al. (168), COVID-19 public TCRs are investigated. GLIPH2 (81), one of the dissimilarity-based methods we reviewed, was used to cluster TCRs and select COVID-19 related TCRs by Student's T-test (similar to Emerson et al. (16) introduced as one of the hypothesis test based methods). GLIPH2 was also employed in Chang et al. (166) to characterize the TCRs related to the severity of the symptoms. Minervina et al. (169) also examined the dynamics of COVID-19 patients' repertoires over time using the hypothesis test previously proposed by the same group (170) to distinguish proliferating clones. Quiros-Fernandez et al. (171) revealed the cross-reactivity of CD8+ T cells in unexposed donors to the COVID-19 epitope, which is derived using NetCTLPan (172), an NN-based peptide-MHC binding prediction software.

We cannot cover all the COVID-19 related literature here. For further reading, see (173) for early researches and (174) for recent updates. For repertoire diversity and COVID-19, see (175, 176). Note that, as COVID-19 is still not fully understood, these results should be further validated in the future.

As a practical application, the repertoire analysis is utilized to diagnose COVID-19. Adaptive Biotechnologies, a US-listed company, applied the ML algorithm that they developed for CMV [in Emerson et al. (16), introduced in Section 3.1] to the COVID-19 dataset. It was demonstrated that the algorithm successfully distinguished the sample's COVID-19 infection status (6). Adaptive Biotechnologies received EUA (Emergency Use Authorization) for the COVID-19 test from the FDA. Nevertheless, repertoire-based test may not be the first choice for COVID-19 diagnosis. First, T(B)CR repertoire can not provide direct evidence of SARS Cov-2 virus existence. Second, repertoire-based test requires sequencing, which costs substantially more than PCR or antibody tests. However, repertoire analysis can potentially reveal far more information than such tests (149), and the sequencing cost is decreasing. Therefore, in the future, repertoire-based blood testing can be utilized further (149).

Small Sample Problem of Repertoire Datasets

The size of datasets is the major determinant of the performance of methods and the reliability of their results (43, 103). Therefore, the establishment and development of sufficiently large datasets are important equally to or even more than the development of analysis methods.

TCRdb (54), one of the major databases of TCR repertoire, contains 131 projects with a total of 8,341 samples of public datasets aggregated from various repositories as of November 2021. Since one project is usually associated with one paper, a rough estimate indicates that one paper contains 64 samples on average. In general, this number is considered small for applying ML algorithms, and actually, the classification methods mentioned above do not always work satisfactorily in some different classification tasks, especially when the sample size is less than 100 (43). A simulation also indicates that the number of samples affects the classification performance (103).

This situation is gradually changing with the appearance of large datasets containing several hundred samples, such as the CMV dataset in Emerson et al. (16). In addition, Adaptive Biotechnologies and Microsoft released a new COVID-19 dataset with 1,486 samples, one of the largest released ever as a single dataset[12], which was used in (6). However, such a large dataset is exceptional, especially as that of the human repertoire, in light of the difficulty to collect a large number of patients with the same condition, e.g., infection records. Even though the number of publicly available datasets have been grown steadily (177), and will continue to grow, the small data size problem may not be readily resolved. Note that we might employ other animals' datasets for some basic research (77). In VDJdb (177), datasets of mice and macaques are recorded. However, the number of the dataset is much fewer than that of humans.

Simulations are not only a powerful tool for repertoire analysis, as we saw earlier, but also can contribute to overcoming the situation, as generative models can create an unlimited amount of pseudo datasets. However, the employment of simulations in repertoire analysis may not always be assured, depending on the tasks and situations. For example, simulated datasets for repertoire classification tasks are created by embedding specific k-mer like signals only in repertoires belonging to specific classes (101–103). Though we know such motifs are important to characterize the binding property of TCR (63), other signals may be still missing. Also, each disease may affect repertoire uniquely [e.g., the difference between CMV and varicella zoster virus (VZV) (178)]. Therefore, until we have a plenty of real datasets, we can not know how we can characterize the changes in repertoire caused by a given condition. Therefore, we will still need real datasets, especially to enable new practical applications.

To alleviate the problem, we have to select appropriate methods for a given size of datasets, understand more about the limit of information that can be derived from a given data, and develop new methods that can integrate multiple datasets or work effectively even with small sizes of datasets.

DISCUSSION

In this paper, we have surveyed ML applications to TCR repertoire analysis by following its development from simple statistical indices to DL, as being summarized in **Figure 3**. For reader's convenience, we summarized a detailed comparison between the methods in **Table 1**.

Finally, we discuss the remaining technological challenges and outline the future directions in the development of TCR repertoire analysis. In particular, we focus on two topics, the small sample problem and the multimodal data integration.

The small sample problem of repertoire datasets we reviewed in the previous section is one of the most important problems that should be resolved in repertoire analysis. As mentioned earlier, the cost of large datasets will likely remain high. Thus, we need to address the problem by devising new analysis methods that can work on smaller but practical datasets. We have at least three representative approaches to achieve this goal. First, as we reviewed in the previous section, simulations can be used to create datasets. We expect more simulation software releases in the future. In this section, we discuss the other two approaches further.

Another possible direction is to utilize multiple datasets to solve a task. Two DL-based techniques which we mentioned earlier will play an important role to this end. Transfer learning can be employed to implement such a method. In transfer learning, we prepare a DL model that has already learned a good feature representation after training on a large unsupervised corpus, and then utilize it for feature extraction in the target task (179). This technique improves the performance of the target task, especially when the dataset for the target task is small. Similarly, representation learning is important. Good representations of repertoires may be learned from large amounts of unlabeled repertoire datasets. If such good representations are learned, classification of individual diseases, for example, may become possible with high performance even if only small amounts of data are available for the target diseases. As we saw earlier, this direction was already investigated using VAE (110). However, the size of the model is far smaller than those used in NLP, and the universality of the representation has not yet been discussed. Moreover, there is no standard task in repertoire analysis in contrast to NLP. Therefore, the models are not evaluated in terms of which downstream tasks can be applied *via* transfer learning. Recently, attempts appear, which utilize large language models in repertoire analysis (133–137, 180). In AntiBERTa (137), fine-tuning for a downstream task is also investigated. Currently, these methods are in development. To be

[12] https://clients.adaptivebiotech.com/pub/covid-2020.

TABLE 1 | Qualitative comparison of the methods reviewed in this article. In practice, both feature encoding methods and ML algorithms for specific tasks such as classification or regression are combined. As the choice of ML algorithms is usually arbitrary, this table is organized by the viewpoint of feature extraction.

Methods		Core Idea	TCR-level encoding	Repertoire-level encoding	ML methods combined with	Strength	Weakness	Notable Examples	Relationship with other methods
Distribution based models	Statistics (Diversity)	TCR diversity is related to healthiness and abnormality of immunological states. Diversity indices such as a rarity weighted count of TCR clonotypes can be used as basic parameters of the immunological state.	NA	A diversity index (a scalar value)	NA	Applicable to data with small sample size and/ or small number of sequences.	Too simple and ignoring sequence information.	Grieff et al (67) used multiple diversity indices to create a repertoire-level feature vector.	NA
	Distribution Shape	The distribution of the clonotype frequency is used to analyze the structure of clonotype diversity. By fitting the sample distribution by probabilistic models, characteristic parameters of the distribution are estimated.	NA	Model parameters of distributions	Probablistic Model	Applicable to data with small sample size and/ or small number of sequences. Flexibility of modeling.	Arbitrariness of modeling and ignoring sequence information.	Guidani et al (69) used a bayesian model to infer the number of clonotypes in a sample from the distribution.	NA
Sequence Information based methods	Hypothesis Test	The TCRs shared among the samples in a condition compared to others might be correlated with the condition. Such TCRs can be identified by hypothesis tests.	Significance of presence or absence of specific TCRs in a condition	A bool vector of the existence of the specific condition-related TCRs found by the hypothesis tests	Various Classifiers	Each TCR can be characterized by the relatedness to the conditions.	Ignoring most of the sequences.	Emerson et al (16) used a hypothesis-based method to find CMV-related TCRs and classify CMV infection based on the existence of such TCRs.s Ritvo et al (75) proposed a method to find proliferated clusters using a hypothesis test.	To include similarity, hypothesis tests are combined with dissimilarity-based methods (ex. Glanville et al., 63).
	Dissimilarity	Similar TCRs may play a similar role in the body. Distance between TCRs can be used to detect and cluster the similar TCRs.	Relative distance from other sequences. Manifold learning is sometimes used to calculate absolute position of the TCR in the latent space	Density distribution on the latent space	Clustering Algorithms and Manifold Learning	Utilizing all sequences to characterize samples. Each TCR is characterized by the relative distance from the other TCRs.	Computational cost of pairwise alignment.	Dash et al (78) used a dissimilarity matrix and visualized the epitope-specific clusters by manifold learning. Yokota et al (79) quantified the distance of repertoires by creating the inter-sample dissimilarity matrix. Glanville et al (63) integrate various information into the dissimilarity calculation (ex. length of CDR3)	NA
	Motif	Local patterns such as (k-mer) motifs in a TCR may be related to its function. Encoding TCRs by a vector of local features may be a	Bag of k-mer. Atchley vector is also used to encode the TCR to more dense vector.	Bag of k-mer or aggreation of TCR-level encoding	Various Classifiers / Regressors	Utilize all sequences to characterize samples. Applicable to data with small sample size and/ or small number of sequences. Each TCR	Low flexibility in modeling.	Sun et al (86) used a 3-mer feature vector of each CDR3 and SVM for a repertoire classification task. Ostmeyer et al (90) used a 4-mer vector further encoded by the Atchley vector, which represents the physicochemical nature of amino acids.	Motifs are sometimes used for calculating dissimilarity (ex. Mayer-

(Continued)

TABLE 1 | Continued

Methods		Core Idea	TCR-level encoding	Repertoire-level encoding	ML methods combined with	Strength	Weakness	Notable Examples	Relationship with other methods
		good representation of TCRs.				is directly characterized as a feature vector.		Katayama et al (43) applied a 3-mer feature vector to repertoire classification tasks on small datasets.	Blackwell et al., 82).
	Generative Models	The mechanisms of generation and selection of TCRs are the determinants of TCR repertoire. Their modelling provides additional information to the observed and not-observed repertoires.	NA	Model parameters of the generative models	Probablistic Model and Manifold Learning	Utilizing all sequences to characterize samples. Applicable to data with small sample size and/ or small number of sequences. Generation of pseudo data (for simulatiion, data augmentation etc.)	Validity of assumptions in models.	Murugan et al (91) modelled the biological V(D)J recombination process and used unselected TCRs to fit the model. Elhanati et al (95) modelled the thymic selection of TCRs and combined Murugan's model to estimate the parameter of the selection process. Pogorelyy et al (96) proposed a method to quantify the abnormality of repertoire using the generation probability from a generative model.	NA
	Deep Learning (DL)	Good representations of repertoires may be obtained by Deep learning and may improve the performance of various repertoire analysis	Various encoding based on VAE or language models (See embedding methods)	Inferred parameters of DL-based models	Generative Models and Embedding Methods	High flexibility in modeling. High performance if sufficient amount of data is provided.	Model is not explainable and data expensive.	Davidsen et al (107) proposed a VAE-based model to embed TCR sequences into the latent space. Widrich et al (102) proposed a Transformer-like model for a repertoire classification problem. Sidhom et al (110) used another VAE-based model to solve various regression/classification tasks.	Embedding Methods are closely related with DL.
	Embedding Methods	Because TCR sequences are a collection of strings, encoding TCRs to fixed-length dense vectors using NLP may lead to efficient algorithms.	Sentence embedding	NA	Various Algorithms incl. DL	High flexibility in modeling. Applicable to data with small sample size and/or small number of sequences (after pre-training).	Model is not explainable.	Cheng et al (166) employed a pre-trained general protein language model for the peptide-MHC binding prediction task. Shuai et al (137) performed pre-training using the repertoire sequence dataset (BCR) and measured the performance on a single downstream task.	NA

NA stands for not applicable.

more widely used, we need to further investigate the transferability of the learned models and representations further. In particular, we believe that studies on language models can be explored. Language models are still improved in NLP, with larger models being pursued. The application of these language models in repertoire analysis is also to be investigated.

The other approach is to combine multiple models to exploit more information in repertoire datasets. The hypothesis test-based methods tend to make predictions based on a tiny subset of specific TCRs, especially public TCRs, and ignore most of the other TCRs in the dataset. In other words, these methods are based on the exact match. This is contrary to a certain class of motif-based or deep-learning-based methods that exploit all the sequences in a sample by encoding them with a fixed-length feature vector. In other words, these methods are based on fuzzy matches. Actually, our group compared these two types of methods and revealed that they provide different prediction profiles (43). Two fuzzy-match-based methods yielded similar predictions. This is intriguing because the two methods are based on completely different methods (k-mer encoding on repertoire level + GBDT vs. deep learning-based feature encoding + attention mechanism). On the other hand, a hypothesis-based method yielded very different predictions. This result suggests that these methods may utilize different information and that ensembling these approaches may result in a better performance on smaller datasets.

While the repertoire data may possess the remaining information that can be further exploited, a T cell population cannot be characterized solely by the sequence information of the TCR repertoire. We cannot predict all the nature of TCRs only from the sequence information. Moreover, important information is missing. For example, T-cell subpopulations cannot be determined by sequence data itself. Therefore, the integration of multimodal information is a promising direction for further repertoire analysis. Most of the methods we reviewed in this paper do not employ information other than TCR sequences except one that integrates the physicochemical properties of amino acids to repertoire datasets (90). We may accommodate a lot more sources to analyze the repertoire dataset. Actually, multi-omics analysis is recently explored (181, 182). The multi-omics approach is usually used with single-cell sequencing to connect multiple data at the single-cell level. Currently, such multi-omics data is not yet popularly employed. However, some interesting findings have been reported. For example, single-cell analysis of RNA-seq and CDR3 revealed the correlation between the gene expression and the frequent CDR3 sequences (182). Another source may come from the 3D structure estimation methods, as the nature of a TCR sequence is determined by the binding affinity to antigens. A recent paper (183) encodes a BCR sequence to a feature vector using the estimated 3D structure of the B cell receptor. Another paper (184) utilizes 3D structure information to predict peptides that bind well with a pair of TCR and MHC. In the paper, a binding score matrix between peptide residues and TCR residues is learned from the existing TCR-pMHC structures. The matrix is then used to calculate the possible alternative peptide of the TCR and MHC.

Toward this direction, hand-crafted models, which exploit specific information based on human understanding, can be effectively utilized to complement the data-driven models by DL. By considering the fact that Alphafold2 (116) was realized by the combination of a feature extraction method and loss function based on chemical insights, it would be promising to unite hand-crafted models with data-driven ones and to integrate multimodal data in repertoire analysis.

AUTHOR CONTRIBUTIONS

YK, RY: writing of the manuscript, TA, TK: writing of the manuscript and supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This research is supported by the JSPS KAKENHI Grant Numbers 19H05799, 19K20408, 20H03441, and by the JST CREST Grant Number JPMJCR2011.

ACKNOWLEDGMENTS

We thank the members of our lab for the fruitful discussion. **Figures 1–3** were created with Biorender.com.

REFERENCES

1. Kumar BV, Connors TJ, Farber DL. Human T Cell Development, Localization, and Function Throughout Life. *Immunity* (2018) 48:202–13. doi: 10.1016/j.immuni.2018.01.007
2. Nikolich-Žugich J, Slifka MK, Messaoudi I. The Many Important Facets of T-Cell Repertoire Diversity. *Nat Rev Immunol* (2004) 4:123–32. doi: 10.1038/nri1292
3. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front Immunol* (2018) 9:224. doi: 10.3389/fimmu.2018.00224
4. De Simone M, Rossetti G, Pagani M. Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. *Front Immunol* (2018) 9:1638. doi: 10.3389/fimmu.2018.01638
5. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An Immunogenic Personal Neoantigen Vaccine for Patients With Melanoma. *Nature* (2017) 547:217–21. doi: 10.1038/nature22991
6. Gittelman RM, Lavezzo E, Snyder TM, Zahid HJ, Elyanow R, Dalai S, et al. Diagnosis and Tracking of Past SARS-CoV-2 Infection in a Large Study of Vo', Italy Through T-Cell Receptor Sequencing [Preprint]. *medRxiv* (2020). doi: 10.1101/2020.11.09.20228023
7. Schuldt NJ, Binstadt BA. Dual TCR T Cells: Identity Crisis or Multitaskers? *J Immunol* (2019) 202:637–44. doi: 10.4049/jimmunol.1800904

8. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol* (2016) 37:724–37. doi: 10.1016/j.it.2016.08.010
9. Garcia KC, Adams EJ. How the T Cell Receptor Sees Antigen—A Structural View. *Cell* (2005) 122:333–6. doi: 10.1016/j.cell.2005.07.015
10. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and Negative Selection of the T Cell Repertoire: What Thymocytes See (and Don't See). *Nat Rev Immunol* (2014) 14:377–91. doi: 10.1038/nri3667
11. Van Laethem F, Tikhonova AN, Singer A. MHC Restriction is Imposed on a Diverse T Cell Receptor Repertoire by CD4 and CD8 Co-Receptors During Thymic Selection. *Trends Immunol* (2012) 33:437–41. doi: 10.1016/j.it.2012.05.006
12. La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. Understanding the Drivers of MHC Restriction of T Cell Receptors. *Nat Rev Immunol* (2018) 18:467–78. doi: 10.1038/s41577-018-0007-5
13. Sewell AK. Why Must T Cells be Cross-Reactive? *Nat Rev Immunol* (2012) 12:669–77. doi: 10.1038/nri3279
14. ElTanbouly MA, Noelle RJ. Rethinking Peripheral T Cell Tolerance: Checkpoints Across a T Cell's Journey. *Nat Rev Immunol* (2021) 21:257–67. doi: 10.1038/s41577-020-00454-2
15. Farber DL, Yudanin NA, Restifo NP. Human Memory T Cells: Generation, Compartmentalization and Homeostasis. *Nat Rev Immunol* (2014) 14:24–35. doi: 10.1038/nri3567
16. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing Identifies Signatures of Cytomegalovirus Exposure History and HLA-Mediated Effects on the T Cell Repertoire. *Nat Genet* (2017) 49:659–65. doi: 10.1038/ng.3822
17. Zvyagin IV, Pogorelyy MV, Ivanova ME, Komech EA, Shugay M, Bolotin DA, et al. Distinctive Properties of Identical Twins' TCR Repertoires Revealed by High-Throughput Sequencing. *Proc Natl Acad Sci* (2014) 111:5980–5. doi: 10.1073/pnas.1319389111
18. Zanelli E, Breedveld FC, de Vries RRP. HLA Association With Autoimmune Disease: A Failure to Protect? *Rheumatology* (2000) 39:1060–6. doi: 10.1093/rheumatology/39.10.1060
19. Slabodkin A, Chernigovskaya M, Mikocziova I, Akbar R, Scheffer L, Pavlović M, et al. Individualized VDJ Recombination Predisposes the Available Ig Sequence Space. *Genome Res* (2021) 31:2209–24. doi: 10.1101/gr.275373.121
20. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, et al. Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. *Front Immunol* (2019) 10:435. doi: 10.3389/fimmu.2019.00435
21. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, et al. VDJbase: An Adaptive Immune Receptor Genotype and Haplotype Database. *Nucleic Acids Res* (2019) 48:D1051–6. doi: 10.1093/nar/gkz872
22. Gras S, Chen Z, Miles JJ, Liu YC, Bell MJ, Sullivan LC, et al. Allelic Polymorphism in the T Cell Receptor and Its Impact on Immune Responses. *J Exp Med* (2010) 207:1555–67. doi: 10.1084/jem.20100603
23. Omer A, Peres A, Rodriguez OL, Watson CT, Lees W, Polak P, et al. T Cell Receptor Beta Germline Variability Is Revealed by Inference From Repertoire Data. *Genome Med* (2022) 14:2. doi: 10.1186/s13073-021-01008-4
24. Dupic T, Bensouda Koraichi M, Minervina AA, Pogorelyy MV, Mora T, Walczak AM. Immune Fingerprinting Through Repertoire Similarity. *PloS Genet* (2021) 17:1–16. doi: 10.1371/journal.pgen.1009301
25. Nikolich-Žugich J. The Twilight of Immunity: Emerging Concepts in Aging of the Immune System. *Nat Immunol* (2018) 19:10–9. doi: 10.1038/s41590-017-0006-x
26. Aiello A, Farzaneh F, Candore G, Caruso C, Davinelli S, Gambino CM, et al. Immunosenescence and Its Hallmarks: How to Oppose Aging Strategically? A Review of Potential Options for Therapeutic Intervention. *Front Immunol* (2019) 10:2247. doi: 10.3389/fimmu.2019.02247
27. Pawelec G. Hallmarks of Human "Immunosenescence": Adaptation or Dysregulation? *Immun Ageing* (2012) 9:15. doi: 10.1186/1742-4933-9-15
28. Palmer D. The Effect of Age on Thymic Function. *Front Immunol* (2013) 4:316. doi: 10.3389/fimmu.2013.00316
29. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next Generation Sequencing for TCR Repertoire Profiling: Platform-Specific Features and Correction Algorithms. *Eur J Immunol* (2012) 42:3073–83. doi: 10.1002/eji.201242517
30. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of Methodologies for T-Cell Receptor Repertoire Analysis. *BMC Biotechnol* (2017) 17:61. doi: 10.1186/s12896-017-0379-9
31. Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent Advances in T-Cell Receptor Repertoire Analysis: Bridging the Gap With Multimodal Single-Cell RNA Sequencing. *ImmunoInformatics* (2022) 5:100009. doi: 10.1016/j.immuno.2022.100009
32. Lee ES, Thomas PG, Mold JE, Yates AJ. Identifying T Cell Receptors From High-Throughput Sequencing: Dealing With Promiscuity in TCRα and TCRβ Pairing. *PloS Comput Biol* (2017) 13:1–25. doi: 10.1371/journal.pcbi.1005313
33. Balakrishnan A, Gloude N, Sasik R, Ball ED, Morris GP. Proinflammatory Dual Receptor T Cells in Chronic Graft-Versus-Host Disease. *Biol Blood Marrow Transplant* (2017) 23:1852–60. doi: 10.1016/j.bbmt.2017.07.016
34. Hosoya T, Li H, Ku CJ, Wu Q, Guan Y, Engel JD. High-Throughput Single-Cell Sequencing of Both TCR-β Alleles. *J Immunol* (2018) 201:3465–70. doi: 10.4049/jimmunol.1800774
35. Carter JA, Preall JB, Atwal GS. Bayesian Inference of Allelic Inclusion Rates in the Human T Cell Receptor Repertoire. *Cell Syst* (2019) 9:475–482.e4. doi: 10.1016/j.cels.2019.09.006
36. Yang L, Jama B, Wang H, Labarta-Bajo L, Zúñiga EI, Morris GP. TCRα Reporter Mice Reveal Contribution of Dual TCRα Expression to T Cell Repertoire and Function. *Proc Natl Acad Sci* (2020) 117:32574–83. doi: 10.1073/pnas.2013188117
37. Trück J, Eugster A, Barennes P, Tipton CM, Luning Prak ET, Bagnara D, et al. Biological Controls for Standardization and Interpretation of Adaptive Immune Receptor Repertoire Profiling. *eLife* (2021) 10:e66274. doi: 10.7554/eLife.66274
38. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of Errors Introduced During High Throughput Sequencing of the T Cell Receptor Repertoire. *BMC Genomics* (2011) 12:106. doi: 10.1186/1471-2164-12-106
39. Rouet R, Jackson KJL, Langley DB, Christ D. Next-Generation Sequencing of Antibody Display Repertoires. *Front Immunol* (2018) 9:118. doi: 10.3389/fimmu.2018.00118
40. Gerritsen B, Pandit A, Andeweg AC, de Boer RJ. RTCR: A Pipeline for Complete and Accurate Recovery of T Cell Repertoires From High Throughput Sequencing Data. *Bioinformatics* (2016) 32:3098–106. doi: 10.1093/bioinformatics/btw339
41. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, et al. Benchmarking of T Cell Receptor Repertoire Profiling Methods Reveals Large Systematic Biases. *Nat Biotechnol* (2021) 39:236–45. doi: 10.1038/s41587-020-0656-3
42. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: Unifying Post-Analysis of T Cell Receptor Repertoires. *PloS Comput Biol* (2015) 11:1–16. doi: 10.1371/journal.pcbi.1004503
43. Katayama Y, Kobayashi TJ. Comparative Study of Repertoire Classification Methods Reveals Data Efficiency of K-Mer Feature Extraction. *Front Immunol* (2022). doi: 10.3389/fimmu.2022.797640
44. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut Learning in Deep Neural Networks. *Nat Mach Intell* (2020) 2:665–73. doi: 10.1038/S42256-020-00257-Z
45. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PloS Med* (2018) 15: e1002683. doi: 10.1371/JOURNAL.PMED.1002683
46. Afzal S, Gil-Farina I, Gabriel R, Ahmad S, von Kalle C, Schmidt M, et al. Systematic Comparative Study of Computational Methods for T-Cell Receptor Sequencing Data Analysis. *Briefings Bioinf* (2017) 20:222–34. doi: 10.1093/bib/bbx111
47. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: Software for Comprehensive Adaptive Immunity Profiling. *Nat Methods* (2015) 12:380–1. doi: 10.1038/nmeth.3364
48. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and

IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi: 10.1007/978-1-61779-842-9_32

49. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: An Immunoglobulin Variable Domain Sequence Analysis Tool. *Nucleic Acids Res* (2013) 41:W34–40. doi: 10.1093/nar/gkt382
50. Zhang Y, Yang X, Zhang Y, Zhang Y, Wang M, Ou JX, et al. Tools for Fundamental Analysis Functions of TCR Repertoires: A Systematic Comparison. *Briefings Bioinf* (2019) 21:1706–16. doi: 10.1093/bib/bbz092
51. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D, et al. Benchmarking Immunoinformatic Tools for the Analysis of Antibody Repertoire Sequences. *Bioinformatics* (2019) 36:1731–9. doi: 10.1093/bioinformatics/btz845
52. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, et al. VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements. *Front Immunol* (2018) 9:976. doi: 10.3389/fimmu.2018.00976
53. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. Ireceptor: A Platform for Querying and Analyzing Antibody/B-Cell and T-Cell Receptor Repertoire Data Across Federated Repositories. *Immunol Rev* (2018) 284:24–41. doi: 10.1111/imr.12666
54. Chen SY, Yue T, Lei Q, Guo AY. TCRdb: A Comprehensive Database for T-Cell Receptor Sequences With Powerful Search Function. *Nucleic Acids Res* (2021) 49:D468–74. doi: 10.1093/NAR/GKAA796
55. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: A Curated Database of T-Cell Receptor Sequences With Known Antigen Specificity. *Nucleic Acids Res* (2017) 46:D419–27. doi: 10.1093/nar/gkx760
56. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 Update. *Nucleic Acids Res* (2018) 47:D339–43. doi: 10.1093/nar/gky1006
57. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: A Manually Curated Catalogue of Pathology-Associated T Cell Receptor Sequences. *Bioinformatics* (2017) 33:2924–9. doi: 10.1093/bioinformatics/btx286
58. Rubelt F, Busse CE, Bukhari SAC, Bürckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community Recommendations for Sharing Immune-Repertoire Sequencing Data. *Nat Immunol* (2017) 18:1274–8. doi: 10.1038/ni.3873
59. Attaf M, Huseby E, Sewell AK. $\alpha\beta$ T Cell Receptors as Predictors of Health and Disease. *Cell Mol Immunol* (2015) 12:391–9. doi: 10.1038/cmi.2014.134
60. Lythe G, Callard RE, Hoare RL, Molina-París C. How Many TCR Clonotypes Does a Body Maintain? *J Theor Biol* (2016) 389:214–24. doi: 10.1016/j.jtbi.2015.10.016
61. Mora T, Walczak AM. How Many Different Clonotypes do Immune Repertoires Contain? *Curr Opin Syst Biol* (2019) 18:104–10. doi: 10.1016/j.coisb.2019.10.001
62. McHeyzer-Williams LJ, Panus JF, Mikszta JA, McHeyzer-Williams MG. Evolution of Antigen-Specific T Cell Receptors *In Vivo*: Preimmune and Antigen-Driven Selection of Preferred Complementarity-Determining Region 3 (CDR3) Motifs. *J Exp Med* (1999) 189:1823–38. doi: 10.1084/jem.189.11.1823
63. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying Specificity Groups in the T Cell Receptor Repertoire. *Nature* (2017) 547:94–8. doi: 10.1038/nature22976
64. Chen G, Yang X, Ko A, Sun X, Gao M, Zhang Y, et al. Sequence and Structural Analyses Reveal Distinct and Highly Diverse Human CD8+ TCR Repertoires to Immunodominant Viral Antigens. *Cell Rep* (2017) 19:569–83. doi: 10.1016/j.celrep.2017.03.072
65. Serana F, Sottini A, Caimi L, Palermo B, Natali PG, Nisticò P, et al. Identification of a Public CDR3 Motif and a Biased Utilization of T-Cell Receptor V Beta and J Beta Chains in HLA-A2/Melan-A-Specific T-Cell Clonotypes of Melanoma Patients. *J Trans Med* (2009) 7:1–14. doi: 10.1186/1479-5876-7-21
66. Chao A, Chiu CH, Jost L. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annu Rev Ecology Evolution Systematics* (2014) 45:297–324. doi: 10.1146/annurev-ecolsys-120213-091540
67. 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) 7:49. doi: 10.1186/s13073-015-0169-8
68. Laydon DJ, Bangham CRM, Asquith B. Estimating T-Cell Repertoire Diversity: Limitations of Classical Estimators and a New Approach. *Philos Trans R Soc B: Biol Sci* (2015) 370:20140291. doi: 10.1098/rstb.2014.0291
69. Guindani M, Sepúlveda N, Paulino CD. Müller P. A Bayesian Semiparametric Approach for the Differential Analysis of Sequence Counts Data. *J R Stat Society: Ser C (Applied Statistics)* (2014) 63:385–404. doi: 10.1111/rssc.12041
70. Rempala GA, Seweryn M, Ignatowicz L. Model for Comparative Analysis of Antigen Receptor Repertoires. *J Theor Biol* (2011) 269:1–15. doi: 10.1016/j.jtbi.2010.10.001
71. Koch H, Starenki D, Cooper SJ, Myers RM, Li Q. powerTCR: A Model-Based Approach to Comparative Analysis of the Clone Size Distribution of the T Cell Receptor Repertoire. *PloS Comput Biol* (2018) 14:1–18. doi: 10.1371/journal.pcbi.1006571
72. Rawstron AC, Fazi C, Agathangelidis A, Villamor N, Letestu R, Nomdedeu J, et al. A Complementary Role of Multiparameter Flow Cytometry and High-Throughput Sequencing for Minimal Residual Disease Detection in Chronic Lymphocytic Leukemia: An European Research Initiative on CLL Study. *Leukemia* (2016) 30:929–36. doi: 10.1038/leu.2015.313
73. Gong Q, Wang C, Zhang W, Iqbal J, Hu Y, Greiner TC, et al. Assessment of T-Cell Receptor Repertoire and Clonal Expansion in Peripheral T-Cell Lymphoma Using RNA-Seq Data. *Sci Rep* (2017) 7:11301. doi: 10.1038/s41598-017-11310-0
74. De Neuter N, Bartholomeus E, Elias G, Keersmaekers N, Suls A, Jansens H, et al. Memory CD4+ T Cell Receptor Repertoire Data Mining as a Tool for Identifying Cytomegalovirus Serostatus. *Genes Immun* (2019) 20:255–60. doi: 10.1038/s41435-018-0035-y
75. Ritvo PG, Saadawi A, Barennes P, Quiniou V, Chaara W, El Soufi K, et al. High-Resolution Repertoire Analysis Reveals a Major Bystander Activation of Tfh and Tfr Cells. *Proc Natl Acad Sci* (2018) 115:9604–9. doi: 10.1073/pnas.1808594115
76. Bashford-Rogers RJ, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network Properties Derived From Deep Sequencing of Human B-Cell Receptor Repertoires Delineate B-Cell Populations. *Genome Res* (2013) 23:1874–84. doi: 10.1101/gr.154815.113
77. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, et al. T Cell Receptor Repertoires of Mice and Humans Are Clustered in Similarity Networks Around Conserved Public CDR3 Sequences. *eLife* (2017) 6: e22057. doi: 10.7554/eLife.22057
78. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable Predictive Features Define Epitope-Specific T Cell Receptor Repertoires. *Nature* (2017) 547:89–93. doi: 10.1038/nature22383
79. Yokota R, Kaminaga Y, Kobayashi TJ. Quantification of Inter-Sample Differences in T-Cell Receptor Repertoires Using Sequence-Based Information. *Front Immunol* (2017) 8:1500. doi: 10.3389/fimmu.2017.01500
80. Zhang H, Liu L, Zhang J, Chen J, Ye J, Shukla S, et al. Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers. *Clin Cancer Res* (2020) 26:1359–71. doi: 10.1158/1078-0432.CCR-19-3249
81. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium Tuberculosis Immune Response by T-Cell Receptor Clustering With GLIPH2 and Genome-Wide Antigen Screening. *Nat Biotechnol* (2020) 38:1194–202. doi: 10.1038/s41587-020-0505-4
82. Mayer-Blackwell K, Schattgen S, Cohen-Lavi L, Crawford JC, Souquette A, Gaevert JA, et al. TCR Meta-Clonotypes for Biomarker Discovery With Tcrdist3 Enabled Identification of Public, HLA-Restricted Clusters of SARS-CoV-2 TCRs. *eLife* (2021) 10:e68605. doi: 10.7554/eLife.68605
83. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The Repertoire Dissimilarity Index as a Method to Compare Lymphocyte Receptor Repertoires. *BMC Bioinf* (2017) 18:155. doi: 10.1186/s12859-017-1556-5
84. Valkiers S, Van Houcke M, Laukens K, Meysman P. ClusTCR: A Python Interface for Rapid Clustering of Large Sets of CDR3 Sequences With Unknown Antigen Specificity. *Bioinformatics* (2021) 37:4865–7. doi: 10.1093/bioinformatics/btab446

85. Zhang H, Zhan X, Li B. GIANA Allows Computationally-Efficient TCR Clustering and Multi-Disease Repertoire Classification by Isometric Transformation. *Nat Commun* (2021) 12:4699. doi: 10.1038/s41467-021-25006-7

86. Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, et al. Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. *Front Immunol* (2017) 0:430. doi: 10.3389/FIMMU.2017.00430

87. Cinelli M, Sun Y, Best K, Heather JM, Reich-Zeliger S, Shifrut E, et al. Feature Selection Using a One Dimensional Naïve Bayes' Classifier Increases the Accuracy of Support Vector Machine Classification of CDR3 Repertoires. *Bioinformatics* (2017) 33:951–5. doi: 10.1093/bioinformatics/btw771

88. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat* (2001) 29:1189 – 1232. doi: 10.1214/aos/1013203451

89. Lawrence N. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *J Mach Learn Res* (2005) 6:1783–1816.

90. Ostmeyer J, Christley S, Toby IT, Cowell LG. Biophysicochemical Motifs in T-Cell Receptor Sequences Distinguish Repertoires From Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue. *Cancer Res* (2019) 79:1671–80. doi: 10.1158/0008-5472.CAN-18-2292

91. Murugan A, Mora T, Walczak AM, Callan CG. Statistical Inference of the Generation Probability of T-Cell Receptors From Sequence Repertoires. *Proc Natl Acad Sci* (2012) 109:16161–6. doi: 10.1073/pnas.1212755109

92. Pogorelyy MV, Minervina AA, Chudakov DM, Mamedov IZ, Lebedev YB, Mora T, et al. Method for Identification of Condition-Associated Public Antigen Receptor Sequences. *eLife* (2018) 7:e33050. doi: 10.7554/eLife.33050

93. Marcou Q, Mora T, Walczak AM. High-Throughput Immune Repertoire Analysis With IGoR. *Nat Commun* (2018) 9:561. doi: 10.1038/s41467-018-02832-w

94. Sethna Z, Elhanati Y, Callan J Curtis G, Walczak AM, Mora T. OLGA: Fast Computation of Generation Probabilities of B- and T-Cell Receptor Amino Acid Sequences and Motifs. *Bioinformatics* (2019) 35:2974–81. doi: 10.1093/bioinformatics/btz035

95. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying Selection in Immune Receptor Repertoires. *Proc Natl Acad Sci* (2014) 111:9875–80. doi: 10.1073/pnas.1409572111

96. Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T, et al. Detecting T Cell Receptors Involved in Immune Responses From Single Repertoire Snapshots. *PloS Biol* (2019) 17:1–13. doi: 10.1371/journal.pbio.3000314

97. DeWitt I William S, Smith A, Schoch G, Hansen JA, Matsen I Frederick A, Bradley P. Human T Cell Receptor Occurrence Patterns Encode Immune History, Genetic Background, and Receptor Specificity. *eLife* (2018) 7:e38358. doi: 10.7554/eLife.38358

98. Bonissone SR, Pevzner PA. Immunoglobulin Classification Using the Colored Antibody Graph. *J Comput Biol* (2016) 23:483–94. doi: 10.1089/cmb.2016.0010

99. Safonova Y, Lapidus A, Lill J. IgSimulator: A Versatile Immunosequencing Simulator. *Bioinformatics* (2015) 31:3213–5. doi: 10.1093/bioinformatics/btv326

100. Yermanos A, Greiff V, Krautler NJ, Menzel U, Dounas A, Miho E, et al. Comparison of Methods for Phylogenetic B-Cell Lineage Inference Using Time-Resolved Antibody Repertoire Simulations (AbSim). *Bioinformatics* (2017) 33:3938–46. doi: 10.1093/bioinformatics/btx533

101. Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, et al. immuneSIM: Tunable Multi-Feature Simulation of B- and T-Cell Receptor Repertoires for Immunoinformatics Benchmarking. *Bioinformatics* (2020) 36:3594–6. doi: 10.1093/bioinformatics/btaa158

102. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, Holzleitner M, et al. Modern Hopfield Networks and Attention for Immune Repertoire Classification. *Adv Neural Inf Process Syst* (2020) 33:18832–45. doi: 10.1101/2020.04.12.038158

103. Kanduri C, Pavlović M, Scheffer L, Motwani K, Chernigovskaya M, Greiff V, et al. Profiling the Baseline Performance and Limits of Machine Learning Models for Adaptive Immune Receptor Repertoire Classification [Preprint]. *bioRxiv* (2021). doi: 10.1101/2021.05.23.445346

104. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep Learning: New Computational Modelling Techniques for Genomics. *Nat Rev Genet* (2019) 20:389–403. doi: 10.1038/s41576-019-0122-6

105. Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and Deep Learning Meet Genome-Scale Metabolic Modeling. *PloS Comput Biol* (2019) 15:1–24. doi: 10.1371/journal.pcbi.1007084

106. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. *IEEE Trans Pattern Anal Mach Intell* (2013) 35:1798–828. doi: 10.1109/TPAMI.2013.50

107. Davidsen K, Olson BJ, DeWitt I William S, Feng J, Harkins E, Bradley P, et al. Deep Generative Models for T Cell Receptor Protein Sequences. *eLife* (2019) 8:e46935. doi: 10.7554/eLife.46935

108. Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: Y Bengio and Y LeCun *International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* Banff. ICLR (2014).

109. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A Comprehensive Survey on Transfer Learning. *Proc IEEE* (2021) 109:43–76. doi: 10.1109/JPROC.2020.3004555

110. Sidhom JW, Larman HB, Pardoll DM, Baras AS. DeepTCR is a Deep Learning Framework for Revealing Sequence Concepts Within T-Cell Repertoires. *Nat Commun* (2021) 12:1–12. doi: 10.1038/s41467-021-21879-w

111. Chaudhari S, Mithal V, Polatkan G, Ramanath R. An Attentive Survey of Attention Models. *ACM Trans Intell Syst Technol* (2021) 12:1–32. doi: 10.1145/3465055

112. Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Front Immunol* (2020) 11:1803. doi: 10.3389/fimmu.2020.01803

113. Fischer DS, Wu Y, Schubert B, Theis FJ. Predicting Antigen Specificity of Single T Cells Based on TCR CDR3 Regions. *Mol Syst Biol* (2020) 16:e9416. doi: 10.15252/msb.20199416

114. Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep Learning-Based Prediction of the T Cell Receptor–Antigen Binding Specificity. *Nat Mach Intell* (2021) 3:864–75. doi: 10.1038/s42256-021-00383-2

115. Nielsen M, Andreatta M, Peters B, Buus S. Immunoinformatics: Predicting Peptide–MHC Binding. *Annu Rev Biomed Data Sci* (2020) 3:191–215. doi: 10.1146/annurev-biodatasci-021920-100259

116. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly Accurate Protein Structure Prediction With AlphaFold. *Nature* (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

117. Isacchini G, Sethna Z, Elhanati Y, Nourmohammad A, Walczak AM, Mora T. Generative Models of T-Cell Receptor Sequences. *Phys Rev E* (2020) 101:62414. doi: 10.1103/PhysRevE.101.062414

118. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A Comprehensive Review and Performance Evaluation of Bioinformatics Tools for HLA Class I Peptide-Binding Prediction. *Briefings Bioinf* (2020) 21:1119–35. doi: 10.1093/bib/bbz051

119. Isacchini G, Walczak AM, Mora T, Nourmohammad A. Deep Generative Selection Models of T and B Cell Receptor Repertoires With Sonnia. *Proc Natl Acad Sci* (2021) 118:e2023141118. doi: 10.1073/pnas.2023141118

120. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, et al. A Compact Vocabulary of Paratope-Epitope Interactions Enables Predictability of Antibody-Antigen Binding. *Cell Rep* (2021) 34:108856. doi: 10.1016/j.celrep.2021.108856

121. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: C Burges, L Bottou, M Welling, Z Ghahramani and K Weinberger, editors. *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc (2013).

122. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Y Bengio and Y LeCun, editors. *International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings* Scottsdale. ICLR (2013).

123. Ostrovsky-Berman M, Frankel B, Polak P, Yaari G. Immune2vec: Embedding B/T Cell Receptor Sequences in ^{N} Using Natural Language

Processing. *Front Immunol* (2021) 12:680687. doi: 10.3389/fimmu.2021.680687

124. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors With Subword Information. *Trans Assoc Comput Linguistics* (2017) 5:135–46. doi: 10.1162/tacl_a_00051
125. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: I Guyon, UV Luxburg, S Bengio, H Wallach and R Fergus, editors. *Advances in Neural Information Processing Systems* Vol. 30. New York: Curran Associates, Inc (2017).
126. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* Minneapolis Minnesota: Assoc Comput Linguistics (2019), 4171–86. doi: 10.18653/v1/N19-1423
127. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: H Larochelle, M Ranzato, R Hadsell, MF Balcan and H Lin, editors. *Advances in Neural Information Processing Systems*, vol. 33. New York: Curran Associates, Inc (2020). 1877–901.
128. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. *Language Models are Unsupervised Multitask Learners* (2019). Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf (Accessed May 17, 2022).
129. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating Protein Transfer Learning With TAPE. *Adv Neural Inf Process Syst* (2019) 32:9689–701. doi: 10.1101/676825
130. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans Pattern Anal Mach Intell* (2021) 2021:1–1. doi: 10.1109/TPAMI.2021.3095381
131. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological Structure and Function Emerge From Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc Natl Acad Sci* (2021) 118:e2016239118. doi: 10.1073/pnas.2016239118
132. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics* (2022) 38:2102–10. doi: 10.1093/bioinformatics/btac020
133. Cheng J, Bendjama K, Rittner K, Malone B. BERTMHC: Improved MHC–peptide Class II Interaction Prediction With Transformer and Multiple Instance Learning. *Bioinformatics* (2021) 22:4172–9. doi: 10.1093/bioinformatics/btab422
134. Gasser HC, Bedran G, Ren B, Goodlett D, Alfaro J, Rajan A. Interpreting BERT Architecture Predictions for Peptide Presentation by MHC Class I Proteins [Preprint]. *arXiv* (2021). doi: 10.48550/ARXIV.2111.07137
135. Hashemi N, Hao B, Ignatov M, Paschalidis I, Vakili P, Vajda S, et al. Improved Predictions of MHC-Peptide Binding Using Protein Language Models [Preprint]. *bioRxiv* (2022). doi: 10.1101/2022.02.11.479844
136. Leem J, Mitchell LS, Farmery JH, Barton J, Galson JD. Deciphering the Language of Antibodies Using Self-Supervised Learning [Preprint]. *bioRxiv* (2021). doi: 10.1101/2021.11.10.468064
137. Shuai RW, Ruffolo JA, Gray JJ. Generative Language Modeling for Antibody Design (2021), Paper presented at: *Machine Learning for Structural Biology Workshop at the 35th Conference on Neural Information Processing Systems*, 2021 Dec 13. doi: 10.1101/2021.12.13.472419
138. Bradley P, Thomas PG. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annu Rev Immunol* (2019) 37:547–70. doi: 10.1146/annurev-immunol-042718-041757
139. Greiff V, Yaari G, Cowell LG. Mining Adaptive Immune Receptor Repertoires for Biological and Clinical Information Using Machine Learning. *Curr Opin Syst Biol* (2020) 24:109–19. doi: 10.1016/j.coisb.2020.10.010
140. Zvyagin IV, Tsvetkov VO, Chudakov DM, Shugay M. An Overview of Immunoinformatics Approaches and Databases Linking T Cell Receptor Repertoires to Their Antigen Specificity. *Immunogenetics* (2020) 72:77–84. doi: 10.1007/s00251-019-01139-4
141. Mösch A, Raffegerst S, Weis M, Schendel DJ, Frishman D. Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Front Genet* (2019) 10:1141. doi: 10.3389/fgene.2019.01141
142. Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, Laukens K, et al. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Front Immunol* (2019) 10:2820. doi: 10.3389/fimmu.2019.02820
143. Ogishi M, Yotsuyanagi H. Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front Immunol* (2019) 10:827. doi: 10.3389/fimmu.2019.00827
144. Heather JM, Best K, Oakes T, Gray ER, Roe JK, Thomas N, et al. Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and Following Antiretroviral Therapy. *Front Immunol* (2016) 6:644. doi: 10.3389/fimmu.2015.00644
145. Qi Q, Cavanagh MM, Saux SL, NamKoong H, Kim C, Turgano E, et al. Diversification of the Antigen-Specific T Cell Receptor Repertoire After Varicella Zoster Vaccination. *Sci Trans Med* (2016) 8:332ra46–332ra46. doi: 10.1126/scitranslmed.aaf1725
146. Teraguchi S, Saputri DS, Llamas-Covarrubias MA, Davila A, Diez D, Nazlica SA, et al. Methods for Sequence and Structural Analysis of B and T Cell Receptor Repertoires. *Comput Struct Biotechnol J* (2020) 18:2000–11. doi: 10.1016/j.csbj.2020.07.008
147. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. Tcr: An R Package for T Cell Receptor Repertoire Advanced Data Analysis. *BMC Bioinf* (2015) 16:1–5. doi: 10.1186/s12859-015-0613-1
148. Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneML Ecosystem for Machine Learning Analysis of Adaptive Immune Receptor Repertoires. *Nat Mach Intell* (2021) 3:936–44. doi: 10.1038/s42256-021-00413-z
149. Arnaout RA, Prak ETL, Schwab N, Rubelt F, Arnaout RA, et al. Adaptive Immune Receptor Repertoire Community. The Future of Blood Testing Is the Immunome. *Front Immunol* (2021) 12:626793. doi: 10.3389/fimmu.2021.626793
150. Liu X, Zhang W, Zhao M, Fu L, Liu L, Wu J, et al. T Cell Receptor β Repertoires as Novel Diagnostic Markers for Systemic Lupus Erythematosus and Rheumatoid Arthritis. *Ann Rheumatic Dis* (2019) 78:1070–8. doi: 10.1136/annrheumdis-2019-215442
151. Ye X, Wang Z, Ye Q, Zhang J, Huang P, Song J, et al. High-Throughput Sequencing-Based Analysis of T Cell Repertoire in Lupus Nephritis. *Front Immunol* (2020) 11:1618. doi: 10.3389/fimmu.2020.01618
152. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC, et al. Analysis of the B Cell Receptor Repertoire in Six Immune-Mediated Diseases. *Nature* (2019) 574:122–6. doi: 10.1038/s41586-019-1595-3
153. Stadinski BD, Shekhar K, Gómez-Touriño I, Jung J, Sasaki K, Sewell AK, et al. Hydrophobic CDR3 Residues Promote the Development of Self-Reactive T Cells. *Nat Immunol* (2016) 17:946–55. doi: 10.1038/ni.3491
154. Daley SR, Koay HF, Dobbs K, Bosticardo M, Wirasinha RC, Pala F, et al. Cysteine and Hydrophobic Residues in CDR3 Serve as Distinct T-Cell Self-Reactivity Indices. *J Allergy Clin Immunol* (2019) 144:333–6. doi: 10.1016/j.jaci.2019.03.022
155. Lagattuta KA, Kang JB, Nathan A, Pauken KE, Jonsson AH, Rao DA, et al. Repertoire Analyses Reveal T Cell Antigen Receptor Sequence Features That Influence T Cell Fate. *Nat Immunol* (2022) 23:446–57. doi: 10.1038/s41590-022-01129-x
156. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. A Dendritic Cell Vaccine Increases the Breadth and Diversity of Melanoma Neoantigen-Specific T Cells. *Science* (2015) 348:803–8. doi: 10.1126/science.aaa3828
157. Blass E, Ott PA. Advances in the Development of Personalized Neoantigen-Based Therapeutic Cancer Vaccines. *Nat Rev Clin Oncol* (2021) 18:215–29. doi: 10.1038/s41571-020-00460-2
158. Garcia-Garijo A, Fajardo CA, Gros A. Determinants for Neoantigen Identification. *Front Immunol* (2019) 10:1392. doi: 10.3389/fimmu.2019.01392
159. Vizcaíno JA, Kubiniok P, Kovalchik KA, Ma Q, Duquette JD, Mongrain I, et al. The Human Immunopeptidome Project: A Roadmap to Predict and

Treat Immune Diseases. *Mol Cell Proteomics* (2020) 19:31–49. doi: 10.1074/mcp.R119.001743

160. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V, et al. Standardized Next-Generation Sequencing of Immunoglobulin and T-Cell Receptor Gene Recombinations for MRD Marker Identification in Acute Lymphoblastic Leukaemia; a EuroClonality-NGS Validation Study. *Leukemia* (2019) 33:2241–53. doi: 10.1038/s41375-019-0496-7
161. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, et al. Community Standardized Representations for Annotated Immune Repertoires. *Front Immunol* (2018) 9:2206. doi: 10.3389/fimmu.2018.02206
162. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, 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. doi: 10.1016/j.immuni.2020.06.024
163. Zhang JY, Wang XM, Xing X, Xu Z, Zhang C, Song JW, et al. Single-Cell Landscape of Immunological Responses in Patients With COVID-19. *Nat Immunol* (2020) 21:1107–18. doi: 10.1038/s41590-020-0762-x
164. Wang P, Jin X, Zhou W, Luo M, Xu Z, Xu C, et al. Comprehensive Analysis of TCR Repertoire in COVID-19 Using Single Cell Sequencing. *Genomics* (2021) 113:456–62. doi: 10.1016/j.ygeno.2020.12.036
165. Hou X, Wang G, Fan W, Chen X, Mo C, Wang Y, et al. T-Cell Receptor Repertoires as Potential Diagnostic Markers for Patients With COVID-19. *Int J Infect Dis* (2021) 113:308–17. doi: 10.1016/j.ijid.2021.10.033
166. Chang CM, Feng P, Wu TH, Alachkar H, Lee KY, Chang WC. Profiling of T Cell Repertoire in SARS-CoV-2-Infected COVID-19 Patients Between Mild Disease and Pneumonia. *J Clin Immunol* (2021) 41:1131–45. doi: 10.1007/s10875-021-01045-z
167. Cheng MH, Zhang S, Porritt RA, Noval Rivas M, Paschold L, Willscher E, et al. Superantigenic Character of an Insert Unique to SARS-CoV-2 Spike Supported by Skewed TCR Repertoire in Patients With Hyperinflammation. *Proc Natl Acad Sci* (2020) 117:25254–62. doi: 10.1073/pnas.2010722117
168. Simnica D, Schultheiß C, Mohme M, Paschold L, Willscher E, Fitzek A, et al. Landscape of T-Cell Repertoires With Public COVID-19-Associated T-Cell Receptors in Pre-Pandemic Risk Cohorts. *Clin Trans Immunol* (2021) 10:e1340. doi: 10.1002/cti2.1340
169. Minervina AA, Komech EA, Titov A, Bensouda Koraichi M, Rosati E, Mamedov IZ, et al. Longitudinal High-Throughput TCR Repertoire Profiling Reveals the Dynamics of T-Cell Memory Formation After Mild COVID-19 Infection. *eLife* (2021) 10:e63502. doi: 10.7554/eLife.63502
170. Pogorelyy MV, Minervina AA, Touzel MP, Sycheva AL, Komech EA, Kovalenko EI, et al. Precise Tracking of Vaccine-Responding T Cell Clones Reveals Convergent and Personalized Response in Identical Twins. *Proc Natl Acad Sci* (2018) 115:12704–9. doi: 10.1073/pnas.1809642115
171. Quiros-Fernandez I, Poorebrahim M, Fakhr E, Cid-Arregui A. Immunogenic T Cell Epitopes of SARS-CoV-2 are Recognized by Circulating Memory and NaïVe CD8 T Cells of Unexposed Individuals. *EBioMedicine* (2021) 72:103610. doi: 10.1016/j.ebiom.2021.103610
172. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: Pan-Specific MHC Class I Pathway Epitope Predictions. *Immunogenetics* (2010) 62:357–68. doi: 10.1007/s00251-010-0441-4
173. Gutierrez L, Beckford J, Alachkar H. Deciphering the TCR Repertoire to Solve the COVID-19 Mystery. *Trends Pharmacol Sci* (2020) 41:518–30. doi: 10.1016/j.tips.2020.06.001
174. Maecker HT. Immune Profiling of COVID-19: Preliminary Findings and Implications for the Pandemic. *J ImmunoTherapy Cancer* (2021) 9:e002550. doi: 10.1136/jitc-2021-002550
175. Gallo Marin B, Aghagoli G, Lavine K, Yang L, Siff EJ, Chiang SS, et al. Predictors of COVID-19 Severity: A Literature Review. *Rev Med Virol* (2021) 31:e2146. doi: 10.1002/rmv.2146
176. Bartleson JM, Radenkovic D, Covarrubias AJ, Furman D, Winer DA, Verdin E. SARS-CoV-2, COVID-19 and the Aging Immune System. *Nat Aging* (2021) 1:769–82. doi: 10.1038/s43587-021-00114-7
177. Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: Database Extension, New Analysis Infrastructure and a T-Cell Receptor Motif Compendium. *Nucleic Acids Res* (2019) 48:D1057–62. doi: 10.1093/nar/gkz874
178. Goronzy JJ, Weyand CM. Understanding Immunosenescence to Improve Responses to Vaccines. *Nat Immunol 2013 14:5* (2013) 14:428–36. doi: 10.1038/NI.2588
179. Ruder S, Peters ME, Swayamdipta S, Wolf T. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 15–8. doi: 10.18653/v1/N19-5004
180. Ruffolo JA, Gray JJ, Sulam J. (2021). Deciphering Antibody Affinity Maturation With Language Models and Weakly Supervised Learning. Paper presented at *Machine Learning for Structural Biology Workshop at the 35th Conference on Neural Information Processing Systems*, 2021 Dec 13. doi: 10.48550/ARXIV.2112.07782
181. Samir J, Rizzetto S, Gupta M, Luciani F. Exploring and Analysing Single Cell Multi-Omics Data With VDJView. *BMC Med Genomics* (2020) 13:29. doi: 10.1186/s12920-020-0696-z
182. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-Cell Multi-Omics Analysis of the Immune Response in COVID-19. *Nat Med* (2021) 27:904–16. doi: 10.1038/s41591-021-01329-2
183. Ripoll DR, Chaudhury S, Wallqvist A. Using the Antibody-Antigen Binding Interface to Train Image-Based Deep Neural Networks for Antibody-Epitope Classification. *PloS Comput Biol* (2021) 17:1–42. doi: 10.1371/journal.pcbi.1008864
184. Karnaukhov VK, Shcherbinin DS, Chugunov AO, Chudakov DM, Efremov RG, Zvyagin IV, et al. Predicting TCR-Peptide Recognition Based on Residue-Level Pairwise Statistical Potential [Preprint]. *bioRxiv* (2022). doi: 10.1101/2022.02.15.480516

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.