ImmunoInformatics 6 (2022) 100012

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.elsevier.com/locate/immuno

A Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data

Teresa Rubio[a], Maria Chernigovskaya[b], Susanna Marquez[c], Cristina Marti[d], Paula Izquierdo-Altarejos[a], Amparo Urios[a,e], Carmina Montoliu[e,f], Vicente Felipo[a], Ana Conesa[g], Victor Greiff[b,*], Sonia Tarazona[h]

[a] Laboratory of Neurobiology, Centro Investigación Príncipe Felipe, Valencia, Spain
[b] Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway
[c] Doctoral Program in Design, Manufacture, and Management of Industrial Projects, Universitat Politècnica de València, Valencia, Spain
[d] Laboratory of Biological Noise and Cell Plasticity, Centro Investigación Príncipe Felipe, Valencia, Spain
[e] Neurological Impairment Laboratory. Fundación Investigación Hospital Clínico Universitario de Valencia, Instituto de Investigación Sanitaria-INCLIVA, Valencia, Spain
[f] Department of Pathology, Facultad de Medicina, Universidad de Valencia, Spain
[g] Genomics of Gene Expression Laboratory, Institute for Integrative Systems Biology, Spanish National Research Council, Valencia, Spain
[h] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Keywords:
T-cell receptor
CD4 isolation
RNA sequencing
Immune repertoire analysis
Minimal Hepatic Encephalopathy

ABSTRACT

T-cell receptor (TCR) analysis is relevant for the study of immune system diseases. The expression of TCRs is usually measured with targeted sequencing approaches where TCR genes are selectively amplified. However, many non-targeted RNA-seq experiments also contain reads of TCR genes, which could be leveraged for TCR expression analysis while reducing sample requirements and costs. Moreover, a step-by-step pipeline for the processing of transcriptome RNA-seq reads to deliver immune repertoire data is missing, and these types of analyses are usually not included in RNA-seq studies of immunological conditions. This represents a missed opportunity for complementing them with the analysis of the immune repertoire.

We present a Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data. We used a case study where TCR repertoire profiles were recovered from bulk RNA-seq of isolated CD4 T cells from control patients, cirrhotic patients without and with Minimal Hepatic Encephalopathy (MHE). MHE is a neuropsychiatric syndrome, mediated by peripheral inflammation, that may affect cirrhotic patients. After the recovery of 498-1,114 distinct TCR beta chains per patient, repertoire analysis of patients resulted in few public clones, high diversity and elevated within-repertoire sequence similarity, independently of immune status. Additionally, TCRs associated with celiac disease and inflammatory bowel disease were significantly over-represented in MHE patient repertoires. The provided computational pipeline functions as a resource to facilitate TCR profiling from RNA-seq data boosting immunophenotype analyses of immunological diseases.

1. Introduction

T-cell receptors (TCR) are able to recognize an immense variety of processed antigens. The approximate diversity of unique TCRs in a human individual is ~10^8–10^{10} [1–3]. A T-cell clonotype is a set of cells that share the same TCR, and the set of unique T-cell clonotypes in an individual is called a TCR repertoire.

The TCR is a two-chain protein and most human T cells consist of α/β chains (TRA and TRB) with a small proportion being γ/δ (TRC and TRD). TCR genes are formed by a process called V(D)J recombination, which consists of the rearrangement of the variable (V), diversity (D), joining (J), and constant (C) gene segments. Three complementarity-determining regions (CDRs) are important for recognizing antigens, with CDR3β (CDR3 region of the TRB chain) as the preferential target of many TCR repertoire studies due to its high diversity and primary importance for antigen binding [4].

High-throughput sequencing (HTS) is a powerful tool for the analysis of these highly diverse immune repertoires, contributing to lymphocyte biology research, antibody engineering, and vaccination [5,6]. For capturing TCRs of α/β T cells, most of the HTS immune repertoire studies (also called Adaptive Immune Receptor Repertoire sequencing, AIRR-seq [7]) apply specific library preparation methods targeting receptor

* Corresponding author.
E-mail address: victor.greiff@medisin.uio.no (V. Greiff).

https://doi.org/10.1016/j.immuno.2022.100012
Received 21 November 2021; Received in revised form 1 March 2022; Accepted 2 March 2022

transcripts [8,9]. Available HTS protocols for immune repertoire profiling include multiplex PCR, with guided primers for the amplification of TCR transcripts at the CDR3 region; target enrichment to capture the sequences of interest using complementary RNA baits; and the more standard 5'RACE (rapid amplification of 5′complementary DNA ends), which is able to retrieve the complete 5′end of the mRNA [6]. However, immune repertoires can also be efficiently extracted from transcriptome RNA-seq data [10–12], as TCR transcripts are part of the bulk-sequenced transcriptome. Using RNA-seq for immune receptor analysis reduces costs and sample amount since both gene expression and immune receptor transcripts are measured in the same experiment, at the expense of less sensitivity as only the most abundant and expanded clones can be recovered. Different computational methods are able to reconstruct TCR from RNA-seq data such as MiXCR [10], CATT [13], or TRUST4 [12]. To a certain extent, these repertoire reconstruction algorithms have been evaluated and compared against each other in terms of the sensitivity and specificity of TCR extraction. For instance, for the MiXCR software it was suggested that it is able to extract high-frequency clonotypes better than the first version of TRUST at any tested read sequencing length and most of the MiXCR-reported clonotypes were confirmed by control TCR-seq data [10]. Additional methods for immune receptor reconstruction have been emerging, such as BASIC [14], BRACER [15], and BALDR [16], but they are able to reconstruct only B-cell receptors (BCR) from single-cell RNA-seq data, which was not applicable in our study. Despite the available collection of TCR repertoire reconstruction tools, a detailed step-by-step pipeline for the processing of immune repertoire data from standard bulk RNA-seq data is not readily available, representing a missed opportunity for complementing gene expression studies of immunological conditions with the analysis of the immune repertoire.

Here, we present an end-to-end pipeline for the analysis of TCR repertoire profiles from bulk RNA-seq, implemented in Nextflow [17] for easy distribution and robust utilization. Nextflow is a workflow management system that provides native support to run pipelines in multiple compute environments and with containerization systems. AIRR-seq and immune-related Nextflow pipelines have become very popular [18–20], mainly due to the simplicity to run an analysis while transparently managing common issues of shell scripting (e.g., required dependencies, computational resources, code failure tracking, cumbersome transfer between collaborators). We apply our Nextflow pipeline to a dataset of CD4 T cells isolated from control patients, cirrhotic patients without, and with Minimal Hepatic Encephalopathy (MHE). MHE is a neuropsychiatric syndrome affecting about 40% of cirrhotic patients who show attention deficits, mild cognitive impairment, psychomotor slowing, and impaired visuomotor coordination [21]. The main hypothesis of MHE etiology is that peripheral inflammation together with hyperammonemia leads to neuroinflammation, which alters neurotransmission and produces cognitive/motor impairment [22]. Specific alterations in the immunophenotype of cirrhotic patients with MHE pointed to CD4 T cells as key factors in the immune shift that triggers the appearance of MHE [23]. The study of CD4 T-cell repertoires might therefore help understand the immune status of MHE patients.

2. Material and Methods

2.1. Patient recruitment

Three groups of patients (healthy control, cirrhotic without, and with MHE) were recruited from the outpatient clinics of Hospital Clinico and Hospital Arnau de Vilanova (Valencia, Spain). The diagnosis of cirrhosis was based on clinical, biochemical, and ultrasonographic data. Cognitive function was evaluated by the Psychometric Hepatic Encephalopathy Score (PHES), a set of five psychometric tests used as the gold standard for MHE diagnosis. Patients were classified as MHE when the score was ≤ -4 points [24]. All participants were enrolled after signing a written informed consent form. Study protocols were approved by the Scientific and Research Ethics Committees of Hospital Clinico Universitario and Arnau de Vilanova Hospital of Valencia, Spain, and were in accordance with the ethical guidelines of the Helsinki Declaration.

Blood samples were collected in BD Vacutainer® (Becton, Dickinson and Company, Franklin Lakes, NJ, USA) tubes with EDTA. Peripheral blood mononuclear cells were centrifuged over a density gradient medium (Lymphoprep™, Palex Medical, SA), according to the manufacturer's instructions and CD4 T cells were purified from 5×10^6 PBMCs by immunomagnetic negative selection using the EasySep™ Human CD4 T Cell Isolation Kit (STEMCELL Technologies Inc.).

2.2. RNA sequencing experimental design

Whole RNA from CD4 T cells was isolated using the miRNeasy Micro Kit (QIAGEN) following the instructions of the manufacturer and sequenced on an Illumina HiSeq2500 machine using HiSeq Sequencing v4 Chemistry. Ultra-low input RNA library preparation with polyA selection and strand-specificity was used for RNA-seq. Paired-end of 125 bp and 50 million reads of sequencing depth was selected for short-read sequencing.

2.3. Read trimming and filtering

Reads were trimmed using Trimmomatic v0.38 [25] when the average Phred quality score was below 20 in a sliding window of 20 bp and removed if the resulting read length was less than 80 bp. These parameters were selected as optimal after a comparative analysis of different sliding window values (from 4-20 bp).

2.4. Repertoire reconstruction

MiXCR v3.0.13 [26] was used to align and assemble TCR repertoire from RNA-seq data using the "analyze shotgun" command and default parameters. Clones (i.e., CDR3 amino acid sequences of the TRB chain) were included in the analysis if they had a minimal abundance read of 2 read counts and their CDR3βs were of 4 amino acids minimum length as described previously [27].

2.5. Hill-based evenness profiles

Common diversity indices are Hill numbers defined as:

$$D_\alpha = \left(\sum_{i=1}^{n} f_i^\alpha \right)^{\frac{1}{1-\alpha}}, \tag{1}$$

where n is the number of unique clones in a repertoire, f_i is the frequency distribution (proportional abundance of clones) and $\boldsymbol{\alpha}$ is a scale parameter in (0,1) and (1, $+\infty$).

An α-Diversity profile (D_α) was previously defined [28] as:

$$D_\alpha = SR \times E_\alpha, \tag{2}$$

where E_α is the evenness and SR is the species richness or the number of unique clones in a repertoire dataset.

Here, we calculated Evenness profiles (E_α), defined as D_α / SR according to the above Equation 2. We used different values of $\boldsymbol{\alpha}$, ranging from 0 to 10 with a step size of 0.2, to obtain the Diversity profiles (D_α) [28]. Diversity is not defined for the case $\boldsymbol{\alpha}$=1 but L'Hospital's rule defines that as $\boldsymbol{\alpha}$ tends to 1, diversity tends to the exponential of the Shannon entropy:

$$D_{\alpha=1} = \exp\left(- \sum_{i=1}^{n} f_i \ \ln\left(f_i\right) \right). \tag{3}$$

All pairwise Pearson correlation coefficients of the Evenness profiles were calculated between samples. Hierarchical clustering was performed based on Euclidean distance for correlations and heatmaps were generated for visualization.

2.6. Shannon Evenness

Shannon Evenness (S-E) is defined as Shannon entropy divided by the Species Richness (SR). S-E is 1 if all clones in a repertoire have the same frequency (an "even" repertoire), or it nears 0 if very few clones dominate in the repertoire ("polarized" repertoire).

2.7. Jaccard similarity

Pairwise clonal convergence between two repertoires A and B was quantified using the Jaccard similarity coefficient, defined as the size of the intersection of A and B divided by the size of the union of A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

Values range between 0 and 1, where 1 means complete overlap of repertoire A and B, and 0 indicates no overlapping receptor sequences between repertoires A and B.

2.8. K-mer-based TCR analysis

Overlapping k-mers of length 3 (k = 3) were extracted from the amino acid CDR3β sequences [27] in each TCR repertoire and condensed into a k-mer frequency distribution matrix using the **immunarch** R package [29]. Hierarchical clustering and heatmap visualization were performed as described above in Section 2.5.

2.9. TCR sequence similarity architecture

TRB repertoire similarity networks were generated as previously described [27,30,31], where nodes represent amino acid CDR3β sequences and edges were drawn between sequences differing by 1 amino acid (Levenshtein distance = 1). The degree (number of links per node) distributions of each repertoire were calculated using the degree function from the R package **igraph** [32].

2.10. Graphics generation

Statistical analysis and graphics were performed using the programming environment R v4.0.5 [33]. The matrix of public clones generated in the repertoire overlap analysis was generated using the **immunarch** R package [29] with the *repOverlap()* and *vis()* functions. All heatmaps were created using the *aheatmap()* function of the **NMF** R package [34]. Mean quality sequencing plots for paired-end reads were obtained from FastQC [35] report and barplots summarizing MiXCR output were drawn using the **ggplot2** R package [36]. The **ggpubr** R package [37] functions *ggboxplot()* and *ggscatter()* were used for clone statistics and correlations, respectively.

2.11. Antigen/Disease-specific TCR databases

McPAS-TCR is a curated database of TCR sequences linked to the associated antigen target or pathology based on published literature [38]. The database was downloaded and filtered by pathological category, maintaining human TRB sequences associated with autoimmunity and pathogens. VDJdb is a curated repository of antigen-specific TCR sequences utilizing experimental information from recently published TCR specificity assays [39]. At the moment of download, McPAS-TCR and VDJdb databases had been last updated on 6 March 2021 and 2 February 2021, respectively.

2.12. Fisher's exact test analysis

The overrepresentation of clones associated with diseases or antigens (McPAS-TCR and VDJdb) in our CDR3β sequences was evaluated using a one-tailed Fisher's exact test applied to each group of patients (control, with MHE, without MHE), using the disease categories included in the McPAS-TCR and VDJdb databases. TCRs present in the samples but absent in the McPAS-TCR and VDJdb databases were excluded from the analysis as described previously [30]. Fisher's exact test was used to test the overrepresentation of specific disease-associated receptors in the database within the measured receptors of the sample. The obtained p-values were adjusted for multiple testing using Benjamini-Hochberg FDR correction considering both the number of diseases and the number of sample groups tested.

2.13. Nextflow pipeline

Nextflow v21.10.6 was used to implement the pipeline. In addition, the DSL2 syntax extension was enabled at the beginning of the workflow script to allow the definition of module libraries and simplify the writing of the data analysis pipeline.

2.14. Data and code availability

The transcriptomic dataset used in this study is available in the GEO database repository, GSE184200, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184200. The code and complete documentation of the Nextflow pipeline are publicly available from the Github repository:https://github.com/ConesaLab/TCR_nextflow.

3. Results and Discussion

3.1. Step-by-step analysis overview

The step-by-step pipeline for the processing of immune repertoire data from whole transcriptome RNA-seq reads is summarized in Fig. 1.

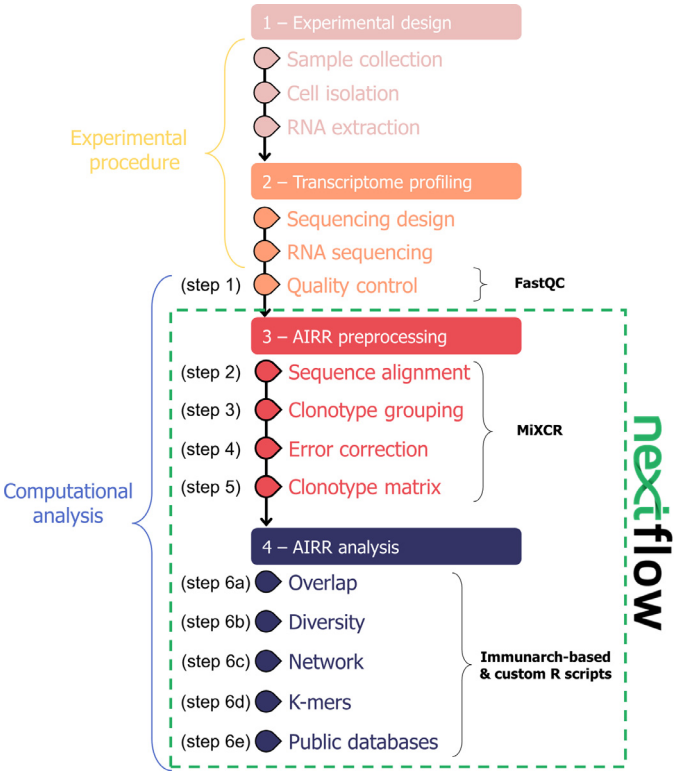

Fig. 1. Pipeline overview. Experimental steps comprise sample collection, cell isolation, RNA extraction, and sequencing. The computational pipeline analysis implemented in Nextflow included repertoire quantification, and repertoire properties screening. AIRR: Adaptive Immune Receptor Repertoire.

The pipeline consists of four main steps representing both the experimental procedure and the computational analysis steps: (1) Experimental design, (2) Transcriptome profiling, (3) AIRR (Adaptive Immune Receptor Repertoire) pre-processing, and (4) AIRR analysis. The experimental procedure includes sample collection, immune cell isolation, RNA extraction, sequencing design (e.g., library strand specificity, paired- or single-end reads), and RNA sequencing (RNA-seq). The computational analysis starts with 'fastq' files in which the sequencing quality (**step 1**) needs to be verified. The MiXCR software [26] was used to assemble TCR repertoires from sequencing reads after quality control. We chose this well-established repertoire reconstruction tool because it is able to extract high-frequency clonotypes from RNA-seq data with a comparable yield to other similar tools (TRUST3, TRUST4) in case of sufficient sequencing read length [12]. MiXCR assembly algorithm avoids the introduction of false-positive clones, which might appear either by alignment of reads to non-target molecules or by overlapping between two sequences from different clones in the reconstruction of partially covered CDR3 regions. The Nextflow pipeline starts with the MiXCR repertoire extraction step (Fig. 1) and was performed using the MiXCR function "analyze shotgun" command, which consists of the following workflow: sequence alignment against reference V, D, J, and C genes (**step 2**), followed by clustering of identical sequences into clonotypes (default, clustering by CDR3β) (**step 3**) and correction of PCR/sequencing errors (**step 4**) to output a tab-delimited file containing the quantification as a clonotype matrix (**step 5**).

Additional AIRR analysis steps are needed to study different immune receptor features: overlap for clonal convergence (**step 6a**), diversity indexes for clonal expansion (**step 6b**), network analysis for clonal sequence architecture (**step 6c**), k-mer distribution for repertoire sequence similarity (**step 6d**) and public databases screening for antigen specificity (**step 6e**). Steps 1–6 were implemented in the Nextflow pipeline as parallel processes that receive MiXCR files as input and provide ready-to-publish plots and tables as well as a final report summarising all results for better user interpretability.

3.2. T-cell receptor sequences can be recovered from RNA-seq data (steps 1–5)

CD4 T cells were isolated and sequenced by bulk paired-end RNA-seq from a total of 20 patients (8 control, 6 cirrhotic without MHE, and 6 cirrhotic with MHE). Sequencing read pre-processing included trimming and filtering (see Methods), which resulted in good quality scores (mean q > 30) for all samples (Supplementary Fig. 1A).

Read alignment against the reference VDJ genes (IMGT database [40]) showed a range of successfully aligned reads between 0.05–0.1%. The majority of reads matched TCR regions (24.1–67.6% of successfully aligned reads), although some reads aligned with immunoglobulin (IG) chains (27.1–75.0% of successfully aligned reads) indicating slight contamination during T-cell isolation (Supplementary Fig. 1B). A high proportion (52.5–87.7%) of the recovered clones, i.e., CDR3 amino acid sequences, matched TRA and TRB chains. Bulk RNA-seq data cannot determine the pairing of specific α/β receptor chains within the population of T cells, something that can only be achieved by sequencing single T cells. Therefore, we decided to focus on the TRB chain for all subsequent analyses. TRB is more appropriate than TRA for identifying T-cell clones because around 7–30% of T cells may have two different alpha chains expressed on the same clone [41] while only 1% of T cells may have two different beta chains on the same clone [42].

We have compared three groups of patients using the Kruskal-Wallis test for various repertoire statistics – number of isolated cells, RNA quantity, number of reads obtained, the number of recovered clones, i.e., CDR3β amino acid sequence, and their Shannon evenness (Supplementary Fig. 2A). The clone recovery yield ranged from 498 to 1,114 distinct CDR3 amino acid sequences per individual in TRB. None of the mentioned measurements showed any significant difference between groups of patients (Kruskal-Wallis test, p-value > 0.05) except for the number of clones, which was significantly increased in cirrhotic patients without MHE versus control (post hoc Wilcoxon test, p-value = 0.024).

To determine whether sequencing depth sufficiently covered the clonal repertoire of the samples, we calculated pairwise correlations as previously described [27] between the cell number, the number of clones, and the Shannon evenness across all samples (Supplementary Fig. 2B). When sequencing depth is saturating with respect to clone detection, the number of clones solely depends on the sample type and not on the number of reads. We found a positive correlation (Pearson coefficient = 0.77, p-value = 6.5×10^{-5}), that may indicate insufficient sequencing depth. Nevertheless, the number of distinct CDR3 sequences assembled was of similar magnitude as reported in other studies of TCR reconstruction from bulk RNA-seq data: 367–936 TRB extracted clonotypes from the central nervous system and 1,684–2,977 extracted TRB clonotypes from the spleen using paired-end data from isolated CD4 T cells [10].

3.3. TCR sequences profiling in MHE

3.3.1. Low clonal convergence among patient samples (step 6a)

Repertoire overlap analysis is the most common approach to uncover clonotypes shared between given individuals, which are also denominated as "public" clones [43–46]. Using the Jaccard similarity (see Methods), we found a low clonal convergence (0.00027±0.00041 Jaccard average measure) between healthy and cirrhotic patients with or without MHE (Fig. 2A).

3.3.2. High clonal expansion in all samples independently of immune status (step 6b)

The expansion of individual T-cell clones that bind their matching antigen can be analyzed using Hill-based evenness profiles, a diversity measurement derived from ecology (see Methods). Unlike single diversity indices, which can produce different clonal expansion results, diversity profiles capture the entire immune repertoire and reflect immunological statuses more sensitively [28]. CD4 T cells in this work showed a positive correlation in diversity profiles of T-cell clones (Pearson coefficient from 0.59 to 1), regardless of the cognitive impairment or cirrhosis condition of the studied patients (Fig. 2B).

3.3.3. Increased within-repertoire similarity based on repertoire architecture is unconstrained by immune status (steps 6c-d)

The adaptive immune response is determined by immune receptor sequences: the higher their dissimilarity, the wider the range of antigens they are able to recognize. The all-to-all sequence similarity within a repertoire represents the repertoire architecture, which was measured in our patients using both k-mers and network analysis.

First, the number of overlapped 3-mers (k-mers length = 3 amino acids) were calculated per patient, as previously described [27]. A large positive Pearson correlation, ranging from 0.96 to 1 was obtained between k-mers vectors (Fig. 2C). This result might indicate that patients share similar sequence architecture patterns independently of their immune status.

To complete the repertoire architecture analysis, a sequence similarity network was constructed using CDR3β amino acid sequence as nodes and adding an edge when sequences differed in 1 amino acid (Levenshtein distance = 1). Then, the number of similar clones (network degree) was calculated and represented as a heatmap (Fig. 2D). 96.8% of the clones had degree = 0 (single nodes) in all the samples and the maximum degree obtained was 5 in one cirrhotic patient without MHE (PC149). This substantial proportion of clones with degree zero was also found in CD4 cells from patients with Multiple Sclerosis while CD8 cells presented a more homogeneous degree distribution [27]. Moreover, the majority of cirrhotic patients without MHE showed a significantly higher (Wilcoxon test, p-value = 0.013) number of single nodes (Fig. 2E), which may be related to underlying differences in

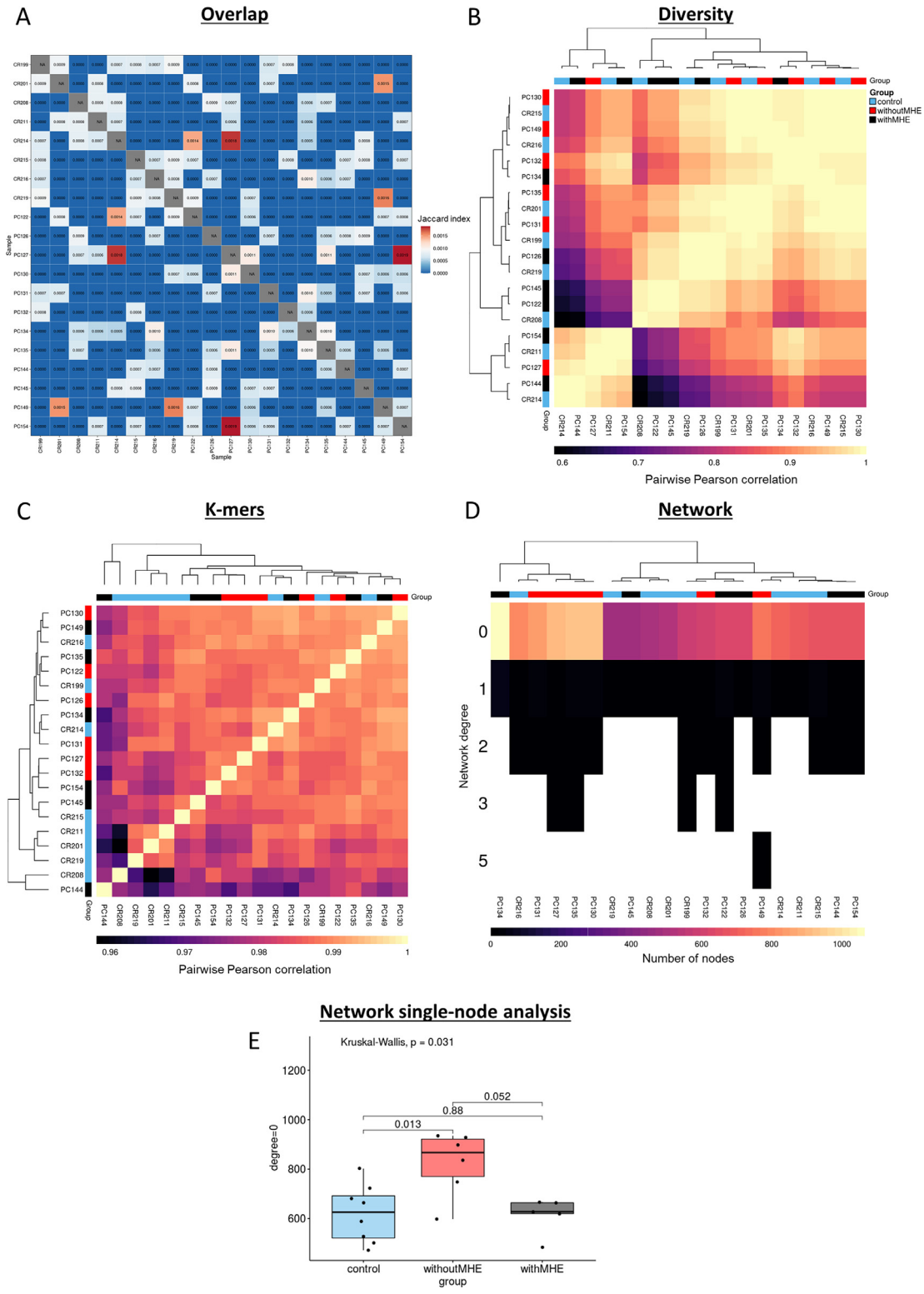

Fig. 2. Different characteristics of TCR (beta chain) repertoire analysis showed specific within-sample profiles independent of immune status. A) Overlap analysis of CDR3β sequences to uncover clonal convergence or shared clonotypes between individuals. Jaccard similarity coefficient matrix indicates a high (red) or low (blue) match between two sample sets. B) Diversity profile analysis to measure clonal expansion. The heatmap contains high (light yellow) or low (dark purple) pairwise Pearson correlation coefficients of the Evenness profiles calculated for control (blue), cirrhotic without MHE (red) or cirrhotic with MHE (black) patients. C) K-mer analysis to examine repertoire similarity. The heatmap represents the highest (light yellow) to the lowest (dark purple) pairwise Pearson correlation coefficients of the 3-mers (amino acid k-mers of length 3) frequency distribution matrix. D) Network analysis to study clonal architecture. Colors from light yellow to dark purple represent the number of nodes with degree 0–5 of each repertoire. Most of the nodes are single (i.e. degree=0). E) Network single node (degree=0) distribution between patient groups to test if the number of degree=0 nodes in cirrhotic patients without MHE (red) were significantly higher than control (blue) or cirrhotic with MHE (black) patients. The test was performed after removing PC134 (outlier in the withMHE group).

Table 1
Top p-values of the overrepresentation CDR3 sequence analysis for the diseases collected in the McPAS-TCR database [38].

McPAS-TCR Disease association	control	withoutMHE	withMHE
Celiac disease	0.011*	0.214	<0.001**
Cytomegalovirus (CMV)	1	1	0.723
Diabetes Type 1	1	1	0.510
Epstein Barr virus (EBV)	1	0.510	1
HTLV-1	1	0.877	0.823
Inflammatory bowel disease (IBD)	1	<0.001**	0.001**
Influenza	0.510	0.510	0.066
Narcolepsy	0.723	1	1
Psoriatic Arthritis	0.064	1	1
Ulcerative Colitis	1	0.066	1
Yellow fever virus	1	0.723	1

P-values were adjusted for multiple testing using Benjamini-Hochberg FDR correction considering both numbers of diseases and the number of sample groups. HTLV1: Human T cell Leukaemia virus 1; *p-value<0.05; **p-value<0.01

the number of TRB clones between the three different groups (Supplementary Fig. 2A). Noteworthy, sample PC134 (cirrhotic patient without MHE) presented 1064 single nodes, the highest number compared with the rest of patients (472–935 single nodes), which could be explained by the highest number of recovered clones (1,114 clones) in this sample, and we considered it as an outlier in this analysis (Fig. 2E).

3.4. TCRs with known disease association are overrepresented in MHE CDR3 repertoires (step 6e)

To evaluate if the clones in our patients were significantly associated with previously described diseases or antigens, we assessed the overlap (see Methods) between our CDR3β sequences and two different TCR databases: VDJdb and McPAS-TCR.

VDJdb contained a total of 41,169 human TRB sequences with *Cytomegalovirus* being the species epitope with the highest number of sequences, compressing nearly a half of them. McPAS-TCR contained a total of 30,052 TRB sequences within the autoimmune and pathology categories, with over half of them belonging to *Mycobacterium tuberculosis*. The sequence intersection between the two databases was low, ranging from 2 to 1,032 in the common pathologies/pathogens: InfluenzaA, Cytomegalovirus, Epstein-Barr Virus, Human Immunodeficiency Virus, Yellow fever virus, Human T cell Lymphotropic Virus, Hepatitis C virus, Mycobacterium tuberculosis, Herpes Simplex Virus 2 and Covid-19, sorted by decreasing order of shared sequences (Supplementary Table 1).

CDR3β sequences were grouped by type of patient (control, cirrhotic without MHE and cirrhotic with MHE) to test overlap with McPAS-TCR database (Table 1 and Supplementary Table 2). We found that Inflammatory bowel disease (IBD) has a significant overrepresentation in cirrhotic without MHE (p-value = 3.082×10^{-5}) and with MHE patients (p-value = 5.470×10^{-4}). Traditionally, IBD has been associated with a Th1-mediated inflammation [47,48], but more recent discoveries have shown the involvement of Th17 cells contributing to inflammation by the secretion of proinflammatory cytokines such as IL-17 and IL-21 [49–51]. Previous studies have also shown alterations in Th1, Th17, IL-17 and IL-21 in patients with MHE [23,52], which may constitute a plausible link between these two disorders. Celiac disease was also significant in control (p-value = 1.127×10^{-2}) and cirrhotic patients without MHE group (p-value = 4.836×10^{-4}). There was no significant overlap with a p-value level below 0.05 between our dataset and VDJdb associated CDR3βs (Supplementary Table 3). While the overlap studies of bulk sequencing data with antigen-specific data are interesting, results remain challenging to interpret, as it is unclear why specific antigens are enriched. The polyreactivity of the TCR repertoire might be a reason. Specifically, each TCR maps to several antigen specificities, or in other words, there is no one-to-one TCR-antigen map, only a many-to-many. We have observed a similar association of bulk sequencing data with seemingly unrelated antigens in two recent publications of ours [27,30].

4. Conclusion

We have presented here a step-by-step computational pipeline for the processing of immune repertoire data from whole transcriptome RNA-seq reads that leverages the presence of immunological receptor sequences (TCR) extracted from RNA-seq transcriptomics datasets. As far as we know, this is the first pipeline that includes both TCR extraction from RNA-seq data as well as a complete immune repertoire data analysis. Different repertoire features can be calculated to interpret the immune repertoire variation. Repertoire overlap analysis is the most common approach to uncover shared clonotypes between individuals also known as “public” clones [45]. Diversity measurement helps understand the expansion of individual T-cell clones [28]. Repertoire architecture is represented by receptor sequence likeness, which determines adaptive immune response and can be quantified both by k-mers and network analysis [27]. Finally, evaluating the overrepresentation of immune receptors with a known pathological association in patient immune repertoires guides the assessment of cross-reactivity with other immunological conditions [30].

We present a case study where TCR repertoire profiles were recovered from bulk RNA-seq of isolated CD4 T cells from control patients, cirrhotic patients without and with MHE. A total of 498–1,114 distinct clones (i.e. CDR3β amino acid sequences) per individual were reconstructed using MiXCR. The authors of this tool have shown similar yields on 100 bp paired-end RNA-seq data using ileocecal lymph node metastasis samples (around 3,000 recovered TRB), small intestine resection samples (around 150 TRB), or isolated CD4 T-cells from spleen (1,684–2,977 TRB) and central nervous system (367–936 TRB) [10]. Our resulting range of clones across patients from isolated CD4 T-cells from human blood samples is halfway between those from spleen and central nervous system, showing a good extraction of clonotypes from the RNA-seq dataset. From clonotype data analysis, we found that the immune repertoires of our three groups of patients are highly similar. The high similarity could have either a biological or technological origin. Three main different reasons can be suggested. (1) The fact that repertoire changes to immune perturbations are more subtle than previously thought. This is in line with recent results on larger cohorts. Specifically, our findings suggest that for autoimmune diseases, the immune signal is very weak if not isolated by cell type for example [53]. (2) Another reason could be the low sample number. However, we have previously shown that even large sample numbers may not lead to separation between patient classes unless specific machine learning algorithms and feature encoding are used [54,55]. (3) It may also be that the proposed features (repertoire overlap, repertoire diversity, network analysis, k-mer, and comparison with existing databases) do not capture the full biological heterogeneity of TCR repertoires. However, we have recently shown that these features cover a large part of immune repertoire diversity [53]. That said, crucial features that have been taken into account are HLA-associated or antigen-specific sequences [54,56,57]. Our pipeline can be applied to any bulk RNA-seq dataset obtained from a sample containing T cells, thanks to the Nextflow implementation. We believe that is a useful resource to study the immune repertoire similarity landscape across different biological scenarios (e.g., health, disease, autoimmunity, infection, vaccination).

Taking advantage of the generation of a vast amount of RNA-seq datasets for different immune cell populations in the last few decades, our Nextflow pipeline can be applied for the study of TCR repertoires to understand patient immune status in multiple diseases. The current version of this pipeline is useful for the study of T-cell subtypes (CD8 and CD4 subpopulations) but it can be easily adapted to the study of B cells. Additionally, it only supports two input species for the moment

(*Homo sapiens* and *Mus musculus*), but they can be expanded as soon as the information on antigen/disease-specific TCR databases increases. However, only the most abundant and expanded clones can be recovered using bulk RNA-seq data for immune repertoire quantification and pairing of specific α/β receptor chains cannot be determined rendering simulation [58,59] and benchmarking [8,60] studies necessary, which will need to determine to what extent the present workflow can be used to identify immune-state-related immune signals. Since bulk RNA-seq datasets combine dual biological information measured in one single experiment, the Nextflow pipeline will allow for parallel analysis of immune repertoires and gene expression. In previous work, we performed the gene expression analysis of the RNA-seq dataset used here, in which the integration of both RNA-seq and miRNA-seq datasets from CD4 T-cells of our MHE patient cohort were analyzed (data not shown). Formal integration of TCR and gene expression analysis results will require the development of adequate mathematical methods that are able to deal with the different structure of both datasets [53,61].

Declaration of Competing Interest

VG declares advisory board positions in aiNET GmbH, Enpicom B.V, Specifica Inc, Adaptyv Biosystems, and EVQLV. VG is a consultant for Roche/Genentech.

Author contributions

AC, VF, ST, and CMontoliu designed the study. AU and PI collected samples and performed T-cell isolation. RNA extraction was made by CMarti. VG designed and supervised data analysis. MC designed Fisher's exact test analysis. VF, AC, and CMontoliu obtained funding. TR analyzed the data and wrote the manuscript. SM and TR assembled the Nextflow workflow. AC, VG, and ST supervised manuscript write-up. All authors reviewed, revised, and approved the final manuscript.

Funding

We acknowledge generous support by The Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to VG), UiO World-Leading Research Community (to VG), UiO:LifeScience Convergence Environment Immunolingo (to VG), EU Horizon 2020 iReceptorplus (#825821) (to VG), a Research Council of Norway FRIPRO project (#300740, to VG), a Research Council of Norway IKTPLUSS project (#311341, to VG), a Norwegian Cancer Society Grant (#215817, to VG). This work was also supported in part by Fundación Ramón Areces (to CM), the Ministerio de Ciencia e Innovación Spain (SAF2017-82917-R and PID2020-113388RB-I00 to VF; FIS PI18/00150 to CM), Consellería Educación Generalitat Valenciana (PROMETEOII/2018/051 to VF), Ministerio de Economía y Competitividad (BIO2015-71658-R to AC), Centro de Investigación Príncipe Felipe (Ayudas para proyectos de investigación intergrupos to TR) and co-funded with European Regional Development Funds (ERDF to VF, CM, AC).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2022.100012.

References

[1] Y. Elhanati, A. Murugan, C.G. Callan, T. Mora, A.M. Walczak, Quantifying selection in immune receptor repertoires, Proc. Natl. Acad. Sci. 111 (2014) 9875–9880. https://doi.org/10.1073/pnas.1409572111.
[2] A. Murugan, T. Mora, A.M. Walczak, C.G. Callan, Statistical inference of the generation probability of T-cell receptors from sequence repertoires, Proc. Natl. Acad. Sci. 109 (2012) 16161–16166. https://doi.org/10.1073/pnas.1212755109.
[3] Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. Blood 2009;114:4099–107. doi:10.1182/blood-2009-04-217604.
[4] Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature 1988;334:395–402. doi:10.1038/334395a0.
[5] Brown AJ, Snapkov I, Akbar R, Pavlović M, Miho E, Sandve GK, Greiff V. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. Mol. Syst. Des. Eng. 2019;4:701–36. doi:10.1039/C9ME00071B.
[6] 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.
[7] Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG, Watson CT, Marthandan N, Faison WJ, Hershberg U, Laserson U, Corrie BD, Davis MM, Peters B, Lefranc M-P, Scott JK, Breden F, Luning Prak ET, Kleinstein SH. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat. Immunol. 2017;18:1274–8. doi:10.1038/ni.3873.
[8] Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, Uddin I, Ismail M, Oakes T, Chain B, Eugster A, Kashofer K, Rainer PP, Darko S, Ransier A, Douek DC, Klatzmann D, Mariotti-Ferrandiz E. 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.
[9] Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends Immunol 2015;36:738–49. doi:10.1016/j.it.2015.09.006.
[10] Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva EV, Obraztsova AS, Shugay M, Ataullakhanov RI, Rudensky AY, Schumacher TN, Chudakov DM. Antigen receptor repertoire profiling from RNA-seq data. Nat. Biotechnol. 2017;35:908–11. doi:10.1038/nbt.3979.
[11] Farmanbar A, Kneller R, Firouzi S. RNA sequencing identifies clonal structure of T-cell repertoires in patients with adult T-cell leukemia/lymphoma. Npj Genomic Med 2019;4:1–9. doi:10.1038/s41525-019-0084-9.
[12] Song L, Cohen D, Ouyang Z, Cao Y, Hu X, Liu XS. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat. Methods. 2021. doi:10.1038/s41592-021-01142-2.
[13] Chen SY, Liu CJ, Zhang Q, Guo AY. An ultra-sensitive T-cell receptor detection method for TCR-Seq and RNA-Seq data. Bioinforma. Oxf. Engl. 2020;36:4255–62. doi:10.1093/bioinformatics/btaa432.
[14] Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinforma. Oxf. Engl. 2017;33:425–7. doi:10.1093/bioinformatics/btw631.
[15] Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao S-W, Sollid LM, Teichmann SA, Stubbington MJT. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat. Methods. 2018;15:563–5. doi:10.1038/s41592-018-0082-3.
[16] Upadhyay AA, Kauffman RC, Wolabaugh AN, Cho A, Patel NB, Reiss SM, Havenar-Daughton C, Dawoud RA, Tharp GK, Sanz I, Pulendran B, Crotty S, Lee FE-H, Wrammert J, Bosinger SE. BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. Genome Med 2018;10:20. doi:10.1186/s13073-018-0528-3.
[17] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 2017;35:316–19. doi:10.1038/nbt.3820.
[18] Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. 2020;38:276–8. doi:10.1038/s41587-020-0439-x.
[19] Garrett ME, Galloway JG, Wolf C, Logue JK, Franko N, Chu HY, Matsen FA, Overbaugh J. Comprehensive characterization of the antibody responses to SARS-CoV-2 Spike protein after infection and/or vaccination. BioRxiv 2021. doi:10.1101/2021.10.05.463210.
[20] J.G. Galloway, E. Matsen, phip-flow, Matsen Group, 2022. https://github.com/matsengrp/phip-flow (Accessed 9 February 2022).
[21] Weissenborn K, Giewekemeyer K, Heidenreich S, Bokemeyer M, Berding G, Ahl B. Attention, Memory, and Cognitive Function in Hepatic Encephalopathy. Metab. Brain Dis. 2005;20:359–67. doi:10.1007/s11011-005-7919-z.
[22] Cabrera-Pastor A, Llansola M, Montoliu C, Malaguarnera M, Balzano T, Taoro-Gonzalez L, García-García R, Mangas-Losada A, Izquierdo-Altarejos P, Arenas YM, Leone P, Felipo V. Peripheral inflammation induces neuroinflammation that alters neurotransmission and cognitive and motor function in hepatic encephalopathy: Underlying mechanisms and therapeutic implications. Acta Physiol 2019;226:e13270. doi:10.1111/apha.13270.
[23] Mangas-Losada A, García-García R, Urios A, Escudero-García D, Tosca J, Giner-Durán R, Serra MA, Montoliu C, Felipo V. Minimal hepatic encephalopathy is associated with expansion and activation of CD4+CD28-, Th22 and Tfh and B lymphocytes. Sci. Rep. 2017;7:6683. doi:10.1038/s41598-017-05938-1.
[24] Weissenborn K, Ennen JC, Schomerus H, Rückert N, Hecker H. Neuropsychological characterization of hepatic encephalopathy. J. Hepatol. 2001;34:768–73. doi:10.1016/S0168-8278(01)00026-5.
[25] Bolger AM, Lohse M, Usadel B. a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
[26] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM. MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods. 2015;12:380–1. doi:10.1038/nmeth.3364.
[27] Amoriello R, Greiff V, Aldinucci A, Bonechi E, Carnasciali A, Peruzzi B, Repice AM, Mariottini A, Saccardi R, Mazzanti B, Massacesi L, Ballerini C. The TCR Repertoire Reconstitution in Multiple Sclerosis: Comparing One-Shot and Continuous Immunosuppressive Therapies. Front. Immunol. 2020;11:559. doi:10.3389/fimmu.2020.00559.
[28] 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.

[29] ImmunoMind Team, immunarch: An R Package for Painless Analysis of Large-Scale Immune Repertoire Data, (2019).
[30] Amoriello R, Chernigovskaya M, Greiff V, Carnasciali A, Massacesi L, Barilaro A, Repice AM, Biagioli T, Aldinucci A, Muraro PA, Laplaud DA, Lossius A, Ballerini C. TCR repertoire diversity in Multiple Sclerosis: High-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood. EBioMedicine 2021;68:103429. doi:10.1016/j.ebiom.2021.103429.
[31] Miho E, Roškar R, Greiff V, Reddy ST. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. Nat. Commun. 2019;10:1321. doi:10.1038/s41467-019-09278-8.
[32] Csardi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst 2006;1695:9.
[33] Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
[34] Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010;11:367. doi:10.1186/1471-2105-11-367.
[35] S. Andrews, FASTQC. A quality control tool for high throughput sequence data, 2010.
[36] Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016. https://ggplot2.tidyverse.org.
[37] A. Kassambara, ggpubr: "ggplot2" Based Publication Ready Plots. R package version 0.4.0., (2020). https://CRAN.R-project.org/package=ggpubr.
[38] Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinforma. Oxf. Engl. 2017;33:2924–9. doi:10.1093/bioinformatics/btx286.
[39] Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, Greenshields-Watson A, Attaf M, Egorov ES, Zvyagin IV, Babel N, Cole DK, Godkin AJ, Sewell AK, Kesmir C, Chudakov DM, Luciani F, Shugay M. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res 2020;48:D1057–62. doi:10.1093/nar/gkz874.
[40] Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 2005;33:D256–61. doi:10.1093/nar/gki010.
[41] Rybakin V, Westernberg L, Fu G, Kim HO, Ampudia J, Sauer K, Gascoigne NRJ. Allelic exclusion of TCR α-chains upon severe restriction of Vα repertoire. PloS One 2014;9:e114320. doi:10.1371/journal.pone.0114320.
[42] Steinel NC, Brady BL, Carpenter AC, Yang-Iott KS, Bassing CH. Posttranscriptional silencing of VbetaDJbetaCbeta genes contributes to TCRbeta allelic exclusion in mammalian lymphocytes. J. Immunol. Baltim. Md 2010;185:1055–62. doi:10.4049/jimmunol.0903099.
[43] Elhanati Y, Sethna Z, Callan CG, Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol. Rev. 2018;284:167–79. doi:10.1111/imr.12665.
[44] Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, Valai A, Lopes T, Radbruch A, Winkler TH, Reddy ST. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep 2017;19:1467–78. doi:10.1016/j.celrep.2017.04.054.
[45] Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, Reddy ST. Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. J. Immunol. 2017;199:2985–97. doi:10.4049/jimmunol.1700594.
[46] Putintseva E, Britanova O, Staroverov D, Merzlyak E, Turchaninova M, Shugay M, Bolotin D, Pogorelyy M, Mamedov I, Bobrynina V, Maschan M, Lebedev Y, Chudakov D. Mother and Child T Cell Receptor Repertoires: Deep Profiling Study. Front. Immunol. 2013;4:463. doi:10.3389/fimmu.2013.00463.
[47] Brand S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut 2009;58:1152–67. doi:10.1136/gut.2008.163667.
[48] Molnár T, Tiszlavicz L, Gyulai C, Nagy F, Lonovics J. Clinical significance of granuloma in Crohn's disease. World J. Gastroenterol. WJG. 2005;11:3118–21. doi:10.3748/wjg.v11.i20.3118.
[49] Nemeth ZH, Bogdanovski DA, Barratt-Stopper P, Paglinco SR, Antonioli L, Rolandelli RH. Crohn's Disease and Ulcerative Colitis Show Unique Cytokine Profiles. Cureus 2017;9:e1177. doi:10.7759/cureus.1177.
[50] Imam T, Park S, Kaplan MH, Olson MR. Effector T Helper Cell Subsets in Inflammatory Bowel Diseases. Front. Immunol. 2018;9:1212. doi:10.3389/fimmu.2018.01212.
[51] Bushara O, Escobar DJ, Weinberg SE, Sun L, Liao J, Yang G-Y. The Possible Pathogenic Role of IgG4-Producing Plasmablasts in Stricturing Crohn's Disease. Pathobiol. J. Immunopathol. Mol. Cell. Biol. 2022:1–11. doi:10.1159/000521259.
[52] Rubio T, Felipo V, Tarazona S, Pastorelli R, Escudero-García D, Tosca J, Urios A, Conesa A, Montoliu C. Multi-omic analysis unveils biological pathways in peripheral immune system associated to minimal hepatic encephalopathy appearance in cirrhotic patients. Sci. Rep. 2021;11:1907. doi:10.1038/s41598-020-80941-7.
[53] Weber CR, Rubio T, Wang L, Zhang W, Robert PA, Akbar R, Snapkov I, Wu J, Kuijjer ML, Tarazona S, Conesa A, Sandve GK, Liu X, Reddy ST, Greiff V. Reference-based comparison of adaptive immune receptor repertoires. bioRxiv 2022. doi:10.1101/2022.01.23.476436.
[54] Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, Rieder M, Robins HS. 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.
[55] Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, Waagan K, Bernal FLM, Costa AA, Corrie B, Akbar R, Al Hajj GS, Balaban G, Brusko TM, Chernigovskaya M, Christley S, Cowell LG, Frank R, Grytten I, Gundersen S, Haff IH, Hovig E, Hsieh P-H, Klambauer G, Kuijjer ML, Lund-Andersen C, Martini A, Minotto T, Pensar J, Rand K, Riccardi E, Robert PA, Rocha A, Slabodkin A, Snapkov I, Sollid LM, Titov D, Weber CR, Widrich M, Yaari G, Greiff V, Sandve GK. 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.
[56] Francis JM, Leistritz-Edwards D, Dunn A, Tarr C, Lehman J, Dempsey C, Hamel A, Rayon V, Liu G, Wang Y, Wille M, Durkin M, Hadley K, Sheena A, Roscoe B, Ng M, Rockwell G, Manto M, Gienger E, Nickerson J, Moarefi A, Noble M, Malia T, Bardwell PD, Gordon W, Swain J, Skoberne M, Sauer K, Harris T, Goldrath AW, Shalek AK, Coyle AJ, Benoist C, Pregibon DC, Jilg N, Li J, Rosenthal A, Wong C, Daley G, Golan D, Heller H, Sharpe A, Abayneh BA, Allen P, Antille D, Armstrong K, Boyce S, Braley J, Branch K, Broderick K, Carney J, Chan A, Davidson S, Dougan M, Drew D, Elliman A, Flaherty K, Flannery J, Forde P, Gettings E, Griffin A, Grimmel S, Grinke K, Hall K, Healy M, Henault D, Holland G, Kayitesi C, LaValle V, Lu Y, Luthern S, Schneider JM, Martino B, McNamara R, Nambu C, Nelson S, Noone M, Ommerborn C, Pacheco LC, Phan N, Porto FA, Ryan E, Selleck K, Slaughenhaupt S, Sheppard KS, Suschana E, Wilson V, Carrington M, Martin M, Yuki Y, Alter G, Balazs A, Bals J, Barbash M, Bartsch Y, Boucau J, Carrington M, Chevalier J, Chowdhury F, DeMers E, Einkauf K, Fallon J, Fedirko L, Finn K, Garcia-Broncano P, Ghebremichael MS, Hartana C, Jiang C, Judge K, Kaplonek P, Karpell M, Lai P, Lam EC, Lefteri K, Lian X, Lichterfeld M, Lingwood D, Liu H, Liu J, Ly N, Hill ZM, Michell A, Millstrom I, Miranda N, O'Callaghan C, Osborn M, Pillai S, Rassadkina Y, Reissis A, Ruzicka F, Seiger K, Sessa L, Sharr C, Shin S, Singh N, Sun W, Sun X, Ticheli H, Trocha-Piechocka A, Walker B, Worrall D, Yu XG, Zhu AM.C.-19 C. and P. Team2‡. Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. Sci. Immunol. 2021. https://www.science.org/doi/abs/10.1126/sciimmunol.abk3070 .
[57] DeWitt WS III, Smith A, Schoch G, Hansen JA, Matsen IV FA, 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.
[58] Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, Reddy ST, Greiff V. 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.
[59] C. Kanduri, M. Pavlović, L. Scheffer, K. Motwani, M. Chernigovskaya, V. Greiff, G.K. Sandve, Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification, (2021) 2021.05.23.445346. https://doi.org/10.1101/2021.05.23.445346.
[60] Dahal-Koirala S, Balaban G, Neumann RS, Scheffer L, Lundin KEA, Greiff V, Sollid LM, Qiao S-W, Sandve GK. TCRpower: quantifying the detection power of T-cell receptor sequencing with a novel computational pipeline calibrated by spike-in sequences. Brief. Bioinform. 2022:bbab566. doi:10.1093/bib/bbab566.
[61] Schattgen SA, Guion K, Crawford JC, Souquette A, Barrio AM, Stubbington MJT, Thomas PG, Bradley P. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat. Biotechnol. 2022;40:54–63. doi:10.1038/s41587-021-00989-2.