TCRosetta: a powerful server for analyzing and annotating T-cell receptor repertoire

Tao Yue
Huazhong University of Science and Technology
Si-Yi Chen
Huazhong University of Science and Technology
Wen-Kang Shen
Huazhong University of Science and Technology
Liming Cheng
Huazhong University of Science and Technology
An-Yuan Guo (guoay@hust.edu.cn)
Huazhong University of Science and Technology

Research Article

Keywords: TCR repertoire, network analysis, TCR CDR3 annotation, web server

Posted Date: May 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1621224/v1

Abstract

Background

Analyzing the T-cell receptor (TCR) repertoire is important with the advent of precision medicine and immunotherapy, as TCR repertoire could serve as a biomarker of immune response and disease progression. Though many TCR analysis methods have been developed for analyzing various aspects of the TCR repertoire, the usage of these methods requires experience in bioinformatics and is technically cumbersome. There is an urgent need to develop an easy-used online server for TCR repertoire analysis.

Methods

TCR CDR3 sequence with clinical information were collected from TCRdb. In TCR repertoire general analysis, Renyi entropy and 1-Pielou's index was used to calculate diversity and clonality of the TCR repertoire, respectively. GIANA and iGraph were used to discover possible disease-specific TCR CDR3 sequences and construct a TCR network in network analysis. OLGA was used to calculate the generation probability of a CDR3 sequence in healthy individuals in public analysis. PHATE was used to embed TCR repertoire in embedding analysis.

Results

In this study, we introduce TCRosetta, a powerful platform for analyzing and annotating TCR repertoire. Above 244 million complementary determining region 3 (CDR3) sequences of TCR beta chain (TRB) with disease information are integrated into the webserver, enabling big-data-based CDR3 annotation for the first time. The main functions of TCRosetta are as follows: (i) General feature analysis for TCR repertoire, including diversity, V/J gene usage, CDR3 length distribution, and clonality etc.; (ii) Annotate the disease preference of TCR repertoire and TRB CDR3 sequences; (iii) TCR repertoire network construction and analysis; (iv) Calculate generation probability of TRB CDR3 sequences. TCRosetta is the first comprehensive online server for TCR repertoire analysis and is useful for immunology research.

Background

T-cell receptor (TCR) is the protein complex on the T-cell surface, which recognizes antigenic peptides bound to major histocompatibility complex molecules. A TCR consists of heterodimer of two chains (αβ or γδ), both of which are products of V(D)J recombination. Estimates of the number of different TCRs in the human body range from 10^{12} to 10^{18} [1]. The high diversity of TCR is caused by genetic rearrangement of the variable (V), diversity (D), and joining (J) genes, also known as V(D)J recombination, a mechanism of somatic recombination during the T-cell maturation [2]. Complementary-determining region 3 (CDR3) is the most variable portion in TCR because the CDR3 is encoded by the

junctions between V, (D), and J genes. The CDR3 region contacts with the presented epitope and is considered the main driver of T-cell specificity [3].

All TCRs make up the TCR repertoire, reflecting the immune status of an individual and associating with human health. Analyzing general features of TCR repertoire (e.g., diversity and V/J gene usage) and discovering key TCR CDR3 sequences can contribute to a understanding of immune response [4] and support further clinical development of immunotherapy [5]. For example, the diversity of TCR repertoire is considered a potential biomarker for tracking the response to immunotherapy [6]. Features of TCR beta chain (TRB) can also assist cancer early-stage diagnosis [7] and treatment selection [8]. TRB has been used to detect malignant clones in some blood diseases and demonstrated better sensitivity and accuracy than traditional methods [9]. Additionally, discovering disease-specific TCR sequences could help to make TCR-engineered T-cells specifically recognize tumor antigens [10].

Currently, the post-processing of TCR repertoire data to reveal the function of T cells in the immune microenvironment has gained more attention. Various methods have been developed for analyzing and visualizing general features of TCR repertoire, such as ImmunExplorer [11], tcR[12], VDJtools [13], and VDJviz [14]. Some methods such as TCRdist [15], GLIPH [16], and GIANA [17], focus on identifying antigen-specific TCR by TCR similarity. Other methods, IGoR [18] and OLGA [19] could calculate generation probabilities of TCR amino acid sequences. However, these methods can only be run locally and are difficult to users without programming skills.

Therefore, it is essential to develop an online platform integrating all these TCR analysis methods for comprehensive TCR analysis. In this study, we propose TCRosetta (http://bioinfo.life.hust.edu.cn/TCRosetta/), a powerful server for analyzing and annotating the TCR repertoire based on big data.

Methods

TCRosetta overview

TCRosetta is a user-friendly and powerful server for analyzing and annotating the TCR repertoire. Because the TRB has higher diversity than TCR alpha chain, most TCR-Seq data only focus on the TRB [20]. Thus, all analyses for TCR repertoire in TCRosetta refer to the TRB. TCRosetta supports several input formats, including CDR3 sequence list, AIRR-compatible formatted file, .csv/.tsv formatted file, as well as output file from TCR extraction software like MiXCR [21], CATT [22], IMSEQ [23] and RTCR[24]. The input data will firstly get through pre-processing to ensure data quality (Figure 1A). TCRosetta provides two modes of analyses for TCR repertoire. The general analyses include repertoire diversity, CDR3 length distribution, V/J gene usage, V-J gene utilization, and clonality (Figure 1B). The advanced analyses contain network analysis, public analysis, embedding analysis, and enrichment analysis (Figure 1C).

Data pre-processing

All input TCR sequences are further processed to ensure the reliability and quality of the sequences. The quality control will be performed based on the following rules: (i) Identical CDR3 sequences with different V/J genes will be merged and only keep the V/J gene with the highest frequency; (ii) Different alleles of the same V/J family will be merged and only keep the family information; (iii) Only sequences containing the V gene, J gene, and complete in-frame CDR3 sequences will be retained; (iv) The complete CDR3 sequence should begin with the cysteine (C), end with the phenylalanine (F) or tryptophan (W) and contain no stop codon according to the IMGT (ImMunoGeneTics) rules [25]; (v) Then the low-quality CDR3 sequences with length less than 8 and greater than 24 will be removed.

Measure TCR repertoire diversity by Renyi entropy

The diversity measures the number of distinct clones and their frequencies in the T-cell repertoire. We adopt the widely used Renyi entropy to quantify the TCR repertoire diversity [26]. Renyi entropy can graphically represent the distribution of abundant clones and rare clones within a given repertoire, in addition to assigning a numerical value to repertoire diversity. Here is the frequency of the sequence in TCR repertoire; is the number of unique sequences in TCR repertoire; is the base of the logarithm, which determines the choice of units of entropy measure. The order sets the degree of sensitivity of the diversity index to sequences abundance in TCR repertoire, as shown in formula (1). When , all sequences are weighted equally. The entropy measure becomes a function of the number of unique sequences, but not dependent on their abundance. When , the Renyi entropy is equivalent to the Shannon entropy.

$$\text{Renyi entropy} = \frac{1}{1-\alpha}\log_b\left(\sum_{i=1}^{N} Pi^{\alpha}\right) \qquad (1)$$

The TCR repertoire clonality analysis

Aside from the diversity of TCR repertoire, the description of equivalency in species abundance can also be used to measure the dominance of clones in a repertoire; thus, referred as clonal evenness [27]. The clonal evenness of a repertoire can be calculated using Pielou's index, and the complement of clonal evenness is often used as clonality. Thus, we use 1-Pielou's index to calculate the TCR repertoire clonality [28]. A clonal score of 0 represents a maximally diverse population with even frequencies, and a value close to 1 means a repertoire driven by clonal dominance, as shown in formula (2). Here, is the frequency of the sequence; is the number of unique sequences, and is the base of the logarithm.

$$\text{Pielou's index} = -\frac{\sum_{i=1}^{N} Pi \log_b Pi}{\log_b(N)} \qquad (2)$$

Cluster TCR sequence and construct TCR network

Similarity in TCR CDR3 sequences implies structural resemblance for antigen recognition, which may share antigen specificity [29]. Therefore, clustering similar CDR3 sequences is an important way to identify antigen-specific receptors. Although there are a few methods for TCR clustering, GIANA is the best tool for clustering large scale TCR repertoire (> 10^6 sequences) with higher clustering accuracy, specificity and efficiency[17]. GIANA converts the sequence alignment and clustering problem into a classic nearest neighbor search in high-dimensional Euclidean space. Thus, we integrate the GIANA into network analysis in TCRosetta to cluster TCR sequences and then construct TCR network. The workflow of network analysis contains the following steps: (i) Use GIANA to cluster similar CDR3 sequences; (ii) Use the Muscle [30] to align sequences obtained in the previous step and then convert it to a distance matrix by calculating Hamming distance between these sequences; (iii) Transform the distance matrix into a adjacency matrix, where two sequences are connected if their Hamming distance is less than 3; (iv) Construct network by igraph (https://igraph.org/) based on the distance matrix from the previous step; (v) Calculate the betweenness of each node in the network by betweenness centrality, which is a measure of the centrality in a graph based on the shortest paths and reflects the control of the node in graph theory; (vi) Mix the betweenness and degree as the weight of a node in the network; (vii) Use the random walk algorithm [31] to discover network communities in the TCR network.

Calculate public score of TCR sequence

Each TCR sequence can be generated in a large number of ways, comprising the recombination of V(D)J segments, random insertions and deletions [18], which make the generation process cannot be described exactly. High-throughput sequencing of large TCR repertoires has enabled the development of methods to predict the probability of generation by V(D)J recombination. Among several methods calculating the generation probability of TCR sequence, OLGA [19] is the best one balanced the accuracy and efficiency by dynamic programming methods. Therefore, we apply OLGA to calculate the generation probability in TCRosetta. We define the public score as the generation probability of a TCR sequence generated in healthy individuals to reflect the healthy degree of the sequence. The parameters of OLGA model were trained by a background, which consists of 5,000,000 randomly selected TCR sequences from healthy samples in TCRdb [32]. The Mann–Whitney U test is used to compare the difference in the public score distribution between input data and background to distinguish whether the input data are healthy.

Embedding analysis based on 3-mer TCR motif

Previous studies demonstrated that only partial CDR3 sequences called “motifs” would contact specific peptides, which forms a part of the TCR specificity [33]. Based on this observation, we employ the k-mer (k = 3) abundance distribution of a TCR repertoire to represent the feature of the TCR repertoire. Each TCR repertoire is represented by a distribution of 3-mer abundance over 20^3 = 8000 dimensions. Such high dimension has more specificity information of TCR repertoire but also introduces lots of noise in the data. Among those dimensionality-reduction methods, PHATE [34] generates a low-dimensional embedding

specific for visualization, which provides an accurate, denoised representation of a TCR repertoire, and is highly scalable both in memory and runtime. Therefore, we apply PHATE to embedding TCR repertoire in TCRosetta.

We use the TCR data of eight diseases (Breast cancer, COVID-19, Crohn's disease, Melanoma, Yellow fever vaccine, classical Hodgkin lymphoma, Non-small cell lung cancer and Cytomegalovirus) in TCRdb to train a PHATE model to learn the embedding representation of k-mer abundance distribution of TCR repertoire. We perform the following steps for training the model. (i) For each sample, split all sequences into 3-mer using a sliding window with step length 1; (ii) Count all 3-mer motifs to form a TCR motif count matrix and exclude samples with either an extremely large (top 20%) or small number (bottom 20%) of motifs; (iii) Filter out motifs that do not appear in more than 50% of the disease samples because that a disease-specific motif should be present in most of samples of this disease; (iv) We use the Mann–Whitney U test with P-value > 0.05 to calculate the difference between the motifs distribution of healthy samples and disease samples to filter out non-disease-specific motifs; (v) Remove the batch effect by scprep (https://scprep.readthedocs.io/en/stable/index.html) and normalize using Z-score for integrating the motif count matrix from eight diseases; (vi) Use PHATE to train a distribution model. For input data, we split all sequences into 3-mers using a sliding window and count all 3-mers motifs to form a TCR motif count matrix. We then embed the matrix into two-dimensional space by our pre-trained model (described above) to obtain its position in the two-dimensional distribution and to discover its possible disease information.

Annotation-based disease preference evaluation

The existing amount of TCR sequences with clinical information motivates us to further explore the possibility of annotating clinical information for unknown samples, which is a function missing in all current tools. The annotation function in TCRosetta allows users to search multiple CDR3 sequences in our reference consisting of 244 million high-quality and annotated CDR3 sequences over 55 clinical conditions. We use Elasticsearch (https://www.elastic.co/), a fast and distributed search engine for all types of data, to quickly and exactly search in large scale CDR3 sequence data. Because only a few CDR3 sequences are potential disease-specific sequences and they are usually cloned with relatively high frequencies in a TCR repertoire [35]. TCRosetta only keeps the top 3000 (sorted by frequency) sequences for fuzzy search with 0 or 1 mismatch amino acid for CDR3 sequence annotation. We then perform statistics to access potential disease preferences based on annotation results. Because sequences annotated for different diseases are more likely to be nonspecific sequences, we exclude sequences annotated with more than five different diseases in the above annotation results. Then we calculate the number of unique sequences for each disease in the upload sample. To further investigate the possible disease preference for the upload data, we use the Fisher Exact test to calculate for disease , which represents the significance of difference between input data and reference, as shown in formula (3). Only diseases with will be left. We calculate the for each disease , which represents the number of

sequences annotated with this disease. Here is the number of total sequences in annotation result, is number of total sequences in reference. is the number of sequences for disease in reference.

$$pi = \frac{\binom{N}{Mi}\binom{Z}{Si}}{\binom{N+Z}{Mi+Si}} \quad (3)$$

Results

Input

TCRosetta provides a user-friendly file manager for users to upload and manage data (upload, remove, and clear). TCRosetta supports three different types of input data: (i) Output file of MiXCR [21], IMSEQ [23], CATT [22] or RTCR [24]; (ii) CDR3 sequences list; (iii) AIRR-compatible formatted file and .csv/.tsv formatted file. The resulting file from MiXCR, IMSEQ, CATT and RTCR can be directly uploaded from the user local machine on the “Analysis” page by clicking the “Choose file” button. File or data uploaded by users will be deleted after 24 hours. Users can also download example files by clicking the download button or “Run example” to run the example directly. To meet different needs of users and save time, we developed two analysis modes in TCRosetta. If users would like to annotate their interested sequences instead of analyzing the entire TCR repertoire, they could enter CDR3 sequences in the input box and select Annotation, Enrichment analysis, and Public analysis. If users would like to analyze TCR repertoire and discover possible disease-specific CDR3 sequences, users could upload the TCR repertoire directly and additionally select Network analysis, General analysis of the TCR repertoire and Embedding analysis (Figure 2A).

Download

TCRosetta allows users to download the analysis results. By clicking the “Export to TSV” button below the data table in analysis results, users can download data in “.tsv” format and open it in Excel. The download data contain TCR sequences information. TCRosetta also allows users to download charts by clicking the download icon in the upper right corner of the chart.

General analysis of TCR repertoire

Some changes like diversity and clonality allow for sensitive tracking of dynamic changes in antigen-specific T-cells and help to predict response to immunotherapy [36]. TCRosetta provides general feature analyses for the TCR repertoire: (i) the CDR3 length distribution; (ii) TCR diversity (calculated by Renyi entropy); (iii) the V-J gene utilization; (iv) the sequence logo of the first and last five amino acids of CDR3 sequences; (v) the V/J gene usage; (vi) the clonality and its’ distribution; (vii) the top ten CDR3 clonotypes by frequency (Figure 2B). TCRosetta also provides a background sequence and V/J gene reference, which

are 100000 CDR3 sequences and TRB V/J genes randomly selected from healthy samples in TCRdb database. The reference CDR3 sequences are used to find the difference between the input data and background.

Public analysis for calculating probability generated from healthy individuals

Due to the high diversity of TCR, most TCRs are private to individuals and don't exist in other individuals. However, some TCRs are shared by multiple individuals, called public TCRs. Previous studies presented that public TCR sequences may be associated with self-related immunity [37] and SARS-CoV-2 epitopes [38]. In public analysis, TCRosetta calculates the public score (details in Methods) of a CDR3 sequence and obtains the TCR repertoire public score distribution using OLGA. TCRosetta also calculates the significance of the difference in public score between sample and background to determine whether there is skewing in the public score distribution of input samples (details see Methods). The public analysis results are shown in tables and graphs (Figure 2C). The public score distribution of the sample is displayed in an interactive box chart. The red and black colors are the distribution of sequences in samples and background, respectively. The P-value is shown at the top of the boxplot, with the higher P-value indicating that the input data are less likely to be generated from healthy individuals. The relationship between public score and frequency of sequences in TCR repertoire is displayed in a scatter chart. The public scores of sequences are displayed in a dynamic table with a filtering function.

Embedding analysis for discovering TCR disease information

CDR3 is the main peptide-contacting region of TCR and the distribution of CDR3 sequence motifs may reflect the disease-specific information (43). TCRosetta splits all CDR3 amino acid sequences in upload data into 3-mer to reveal possible disease specificity from CDR3 motif distribution and obtain a high-dimensional distribution of motifs. Then, TCRosetta embeds it into a pre-trained model PHATE (details see Methods). The embedding analysis result is shown in a scatter chart, where different colors represent different diseases. We used a lung cancer data set from immunoACCESS to evaluate the function of the embedding analysis. We embedded the data set into two-dimensional space by our pre-trained model and found that the points of the data set exactly fell in the region of Non-small Cell Lung Cancer (NSCLC) samples (Figure 2D).

Annotation and Enrichment analysis for annotating unknown sample

TCR-Seq has been developed rapidly and a large number of TCR-Seq data with disease information are available in public. An important issue is how to use such a large amount of TCR sequences to annotate the disease condition of unknown samples. TCRosetta annotates sample by batch search similar TRB CDR3 sequences in reference and then annotates sample disease information based on the statistic of search results (details in Methods). Results of annotation are displayed in a dynamic table with filter and export function on the “Annotation by batch search” page. Users can filter the table by selecting disease or inputting interested CDR3 sequence and export the table with TSV format by clicking the “Export to TSV” button. The statistics for search results are displayed on the “Enrichment analysis” page as follows. 1) The enriched disease distribution of upload data is displayed in a treemap (Figure 2E). Different square colors represent different diseases, and the size of the square is the number of unique sequences annotated with this disease. Users can further browse the CDR3 sequence, Vregion, and Jregion usage by clicking the square. 2) The corrected enriched disease distribution is shown in a pie chart. The size of the sector represents the ratio of the disease. 3) The result of the Fisher exact test is represented in the table sorted by the significance from largest to smallest. Each row represents a disease, and each column contains the Condition, P-value, and Effect Size. A large effect size means that a finding has practical significance. Users could discover the possible disease information of the upload data through the P-value and Effect size.

Network analysis for clustering similar CDR3 sequences

In TCRosetta, the upload TCR repertoire will be transformed to a TCR network to visualize similar sequences. The TCR network is displayed in an interactive and zoomable chart. Each node represents a unique CDR3 sequence, and the edge between two nodes represents the Hamming distance of two sequences less than 3. The node size indicates the node importance and different colors of nodes represent different communities (modules) in the network (Figure 3A). The similarity of sequences within a community is greater than that outside the community. The sequences of the network are displayed in a dynamic table with filtering and exporting functions (Figure 3B). To better observe the usage of amino acids in each position on the CDR3 sequence, we make the sequence logo of each cluster in network to show the position weight matrix of the complete CDR3 sequence, reflecting cluster conservation (Figure 3C).

Case study: analyses of TCR repertoire for an immunotherapy dataset by TCRosetta

To illustrate the powerful function of TCRosetta in TCR repertoire analysis, a dataset of anti-PD-1 immunotherapy for oral carcinoma [39] has been used, including 10 pre-treatment samples and 10 post-treatment samples. A total of 136,992 and 174,885 TRB CDR3 sequences with V, J gene usage were obtained for pre-treatment and post-treatment sample, respectively. Two groups were separately uploaded into TCRosetta with all analyses selected. The results were summarized in Figure 4. There is no

difference between the two groups on the CDR3 length distribution. By comparing the two groups, we observed trends of the lower diversity upon anti-PD-1 treatment (Figure 4A). For clonality, the post-treatment's clonality increases during the treatment, which is consistent with the original study, suggesting T cell clonal expansion could be beneficial in oral carcinoma immunotherapy. Next, we compared the clonal expansion of CDR3 sequences between the two groups. The CDR3 sequence with highest clonal frequency in the pre-treatment group (CASSEEAGTIYEQYF) is different with the post-treatment-group (CATSRESPGQGIDEQ) (Figure 4B). Then, we explored V and J gene pairing usage bias for two groups. In the pre-treatment group, the most frequently used V-J gene pairing segment is TRBV5-1/TRBJ1-2, while in the post-treatment group, it is TRBV15-1/TRBJ2-1 (Figure 4C). For every single V and J gene segment, there are some V and J gene segments with increased or decreased frequency in the post-treatment group comparing with the pre-treatment group, such as TRBJ1-1, TRBJ1-2, TRBV7-9 and TRBV21-1 (Figure 4D). These differences in V/J gene usage may help predict immunotherapy response. Finally, we studied the potential disease-specific CDR3 sequences by clustering similar CDR3 sequences in the pre-treatment group and the post-treatment group. We found that the distributions of network nodes are different in the two groups, potentially supporting the unique function of TRB CDR3s in antigen recognition. For each CDR3 cluster in the network, we observed that "GTG" is the conserved residues based on the distribution of amino acid usage frequency in the pre-treatment group, which could be favorable residues for antibody-antigen binding (Figure 4E).

Web implementation

For the website front end, we use Vue.js (https://vuejs.org/), a progressive JavaScript framework, to build the web server and communicate with user clients. The website is designed to be user-friendly with six pages: Home, Analysis, Document, Help, Contact, and Result. The main pages are Analysis, Result, and Help. On the Analysis page, users can upload a file or input a CDR3 sequence list to analyze. The Result page shows the interactive charts using Echarts (https://echarts.apache.org), a free, powerful charting, and visualization library. The manual of the webserver is on the Help page. For the back end, we build the webserver using Flask (https://flask.palletsprojects.com/en/2.0.x/) and Python 3.7 (https://www.python.org/). Flask is a lightweight web application framework.

Discussion And Conclusions

Analysis of the TCR repertoire may help to gain a better understanding of the immune system. However, the huge increase in analysis methods requires extensive skills in bioinformatics and programming. Thus, we developed a comprehensive T-cell repertoire analysis platform: TCRosetta. TCRosetta could not only analyze the features of TCR repertoire and display them in interactive plots but also is the first platform with batch search function and TCR annotation function.

TCRosetta can be applied to many situations. For example, calculating the features (CDR3 sequence length distribution, diversity, V-J utilization, and clonality) of the TCR repertoire can be used as

biomarkers for predicting response to immunotherapy [36]. Besides, Network analysis could help to identify T-cell clones specific to the tumor by clustering similar TCR sequences, which could play an important role in some tumor immunotherapy and TCR-T therapy [29]. Moreover, Public analysis in TCRosetta may assist in predicting the autoimmune disease prognosis by discovering TCR CDR3 sequences shared between individuals associated with self-related immunity [37].

TCR beta chain has the highest diversity and plays an important role in tumor antigen recognition. Some studies have shown that TCR alpha chain also has function in tumor antigen recognition. TCRosetta mainly studies the features on the TCR beta chain and ignores the alpha chain, which may limit the its usage. TCRosetta integrated More than 244 million reliable TRB CDR3 sequences with clinical condition in annotation reference, which make it possible to annotate TCR repertoire with big data. But due to the high diversity of TCR in human body (range from 10^{12} to 10^{18}), these data are not sufficient to annotate all TCRs. With the development of TCR studies, we will continue to add more analysis functions for the TCR repertoire analysis and support more TCR profiling tools. Future updates will further improve the data volume of the reference to ensure that more TCR sequences could be annotated. We believe that TCRosetta would facilitate TCR-related research and clinical applications.

Abbreviations

TCR
T-cell receptor
CDR3
Complementary determining region 3
TRB
TCR beta chain
V
Variable genes
J
Joining genes
D
Diversity genes
C
Cysteine
F
Phenylalanine
W
Tryptophan

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

TCRosetta is a web server freely accessible without login requirement at (We used a lung cancer data set from immunoACCESS (https://doi.org/10.21417/MC2021NC)). The source code is available at (https://github.com/yttyhhh/TCRosetta_code). Data for anti-PD-1 immunotherapy for oral carcinoma can be obtained at https://doi.org/10.21417/SL2021CRM and the lung cancer data set is from https://doi.org/10.21417/MC2021NC.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Key R&D Program of China (2021YFF0703704) and
Natural Science Foundation for Distinguished Young Scholars of Hubei Province of China [2020CFA070].

Authors’ contributions

AYG design the study. YT and SYC collected TCR-Seq data, constructed website. LC help on the data collection. All the authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Arstila T. Petteri, Casrouge Armanda, Baron Véronique, Even Jos, Kanellopoulos Jean, Kourilsky Philippe. A Direct Estimate of the Human αβ T Cell Receptor Diversity. Science. American Association for the Advancement of Science; 1999;286:958–61.

2. Davis MM, Boyd SD. Recent progress in the analysis of αβT cell and B cell receptor repertoires. Curr Opin Immunol. 2019;59:109–14.
3. Miles JJ, Douek DC, Price DA. Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. Immunol Cell Biol. 2011;89:375–87.
4. Gate D, Saligrama N, Leventhal O, Yang AC, Unger MS, Middeldorp J, et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. Nature. 2020;577:399–404.
5. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell. Elsevier; 2017;169:1342–1356.e16.
6. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat Med. 2019;25:1251–9.
7. Valentina Giudice, Xingmin Feng, Zenghua Lin, Wei Hu, Fanmao Zhang, Wangmin Qiao, et al. Deep sequencing and flow cytometric characterization of expanded effector memory CD8 + CD57 + T cells frequently reveals T-cell receptor Vβ oligoclonality and CDR3 homology in acquired aplastic anemia. Haematologica. 2018;103:759–69.
8. Levine AG, Hemmers S, Baptista AP, Schizas M, Faire MB, Moltedo B, et al. Suppression of lethal autoimmunity by regulatory T cells with a single TCR specificity. J Exp Med. 2017;214:609–22.
9. Nazarov VI, Minervina AA, Komkov AY, Pogorelyy MV, Maschan MA, Olshanskaya YV, et al. Reliability of immune receptor rearrangements as genetic markers for minimal residual disease monitoring. Bone Marrow Transplant. 2016;51:1408–10.
10. Roth TL, Puig-Saus C, Yu R, Shifrut E, Carnevale J, Li PJ, et al. Reprogramming human T cell function and specificity with non-viral genome targeting. Nature. 2018;559:405–9.
11. Schaller S, Weinberger J, Jimenez-Heredia R, Danzer M, Oberbauer R, Gabriel C, et al. ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. BMC Bioinform. 2015;16:252.
12. 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 Bioinform. 2015;16:175.
13. 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. Public Library of Science; 2015;11:e1004503.
14. Bagaev DV, Zvyagin IV, Putintseva EV, Izraelson M, Britanova OV, Chudakov DM, et al. VDJviz: a versatile browser for immunogenomics data. BMC Genom. 2016;17:453.
15. 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.
16. 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.
17. 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.

18. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nat Commun. 2018;9:561.
19. Sethna Z, Elhanati Y, Callan CG Jr, 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.
20. Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. Genome Med. 2013;5:98.
21. 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.
22. Chen S-Y, Liu C-J, Zhang Q, Guo A-Y. An ultra-sensitive T-cell receptor detection method for TCR-Seq and RNA-Seq data. Bioinformatics. 2020;36:4255–62.
23. Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K, et al. IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics. 2015;31:2963–71.
24. 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.
25. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res. 2015;43:D413–22.
26. 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;9:00224.
27. Reuben A, Zhang J, Chiou S-H, Gittelman RM, Li J, Lee W-C, et al. Comprehensive T cell repertoire characterization of non-small cell lung cancer. Nat Commun. 2020;11:603.
28. 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. Chakraborty AK, editor. eLife. eLife Sciences Publications, Ltd; 2017;6:e22057.
29. Zhao L, Cao YJ. Engineered T Cell Therapy for Cancer in the Clinic. Front Immunol. 2019;10:02250.
30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
31. F. Berardo de Sousa, L. Zhao. Evaluating and Comparing the IGraph Community Detection Algorithms. 2014 Brazilian Conference on Intelligent Systems. 2014. p. 408–13.
32. Chen S-Y, Yue T, Lei Q, Guo A-Y. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. Nucleic Acids Res. 2021;49:D468–74.
33. Rudolph MG, Stanfield RL, Wilson IA. HOW TCRS BIND MHCS, PEPTIDES, AND CORECEPTORS. Annu Rev Immunol. Annual Reviews; 2006;24:419–66.
34. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37:1482–92.

35. Fairfax BP, Taylor CA, Watson RA, Nassiri I, Danielli S, Fang H, et al. Peripheral CD8 + T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma. Nature Medicine. 2020;26:193–9.
36. Kidman J, Principe N, Watson M, Lassmann T, Holt RA, Nowak AK, et al. Characteristics of TCR Repertoire Associated With Successful Immune Checkpoint Therapy Responses. Front Immunol [Internet]. Frontiers; 2020 [cited 2021 Jan 4];11. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2020.587014/full
37. Goda S, Hayakawa S, Karakawa S, Okada S, Kawaguchi H, Kobayashi M. Possible involvement of regulatory T cell abnormalities and variational usage of TCR repertoire in children with autoimmune neutropenia. Clin Exp Immunol. 2021;204:1–13.
38. Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova KV, Kiryukhin D, Titov A, et al. SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity. Elsevier; 2020;53:1245–1257.e5.
39. Liu S, Knochelmann HM, Lomeli SH, Hong A, Richardson M, Yang Z, et al. Response and recurrence correlates in individuals treated with neoadjuvant anti-PD-1 therapy for resectable oral cavity squamous cell carcinoma. Cell Rep. Elsevier; 2021;2:100411.

Figures

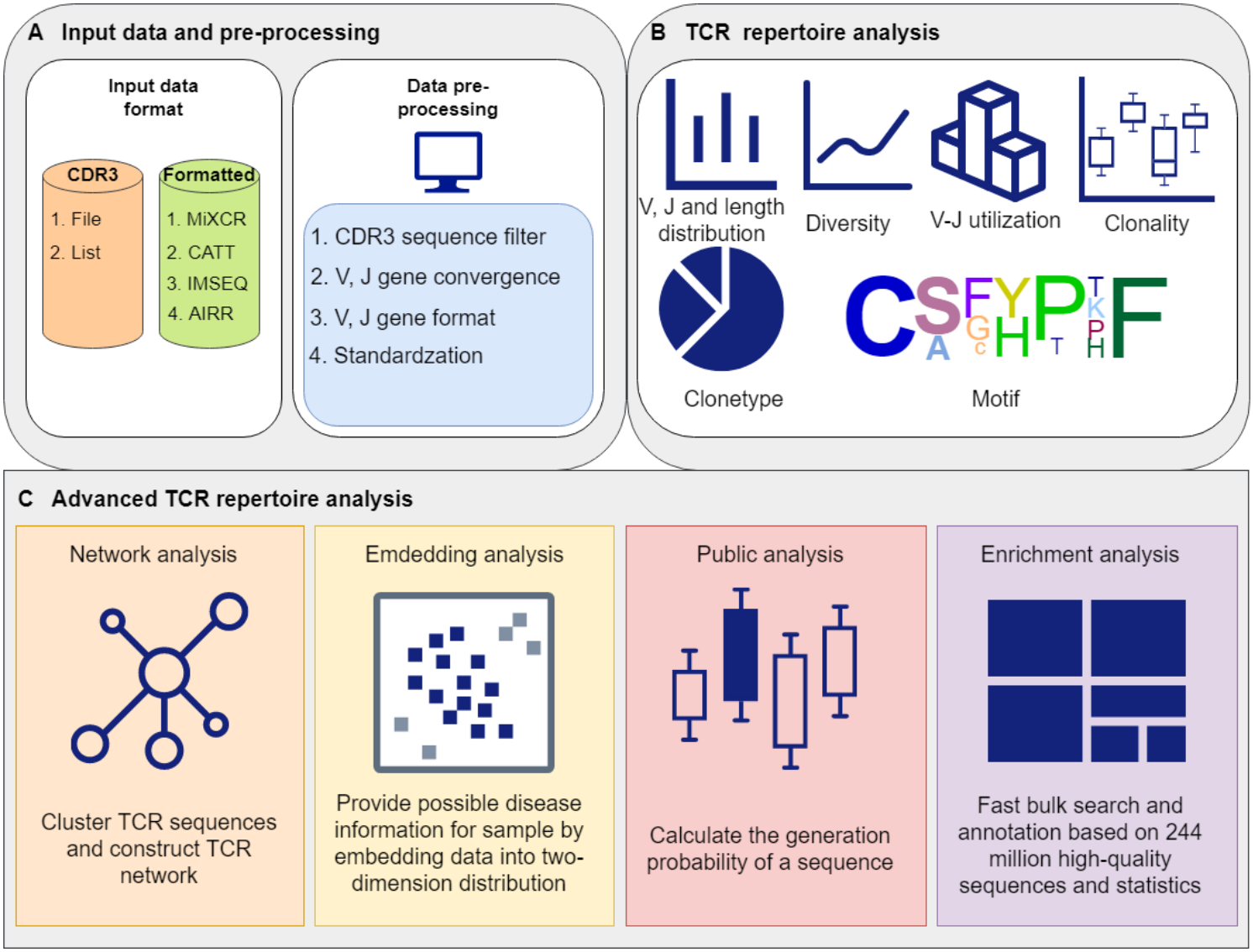


Figure 1

Overview of TCRosetta. (A) Input data format and pre-processing in TCRosetta. TCRosetta supports several input formats, including TCR sequence list and output files from TCR extraction software like MiXCR, CATT, and IMSEQ. Data pre-processing is used to obtain reliable and high-quality TCR sequences. (B) The general analysis for TCR repertoire in TCRosetta includes diversity, clonality, clonetype, V/J usage, V-J gene utilization and CDR3 length distribution. (**C**) The advanced analysis in TCRosetta contains network analysis, embedding analysis, public analysis and enrichment analysis.

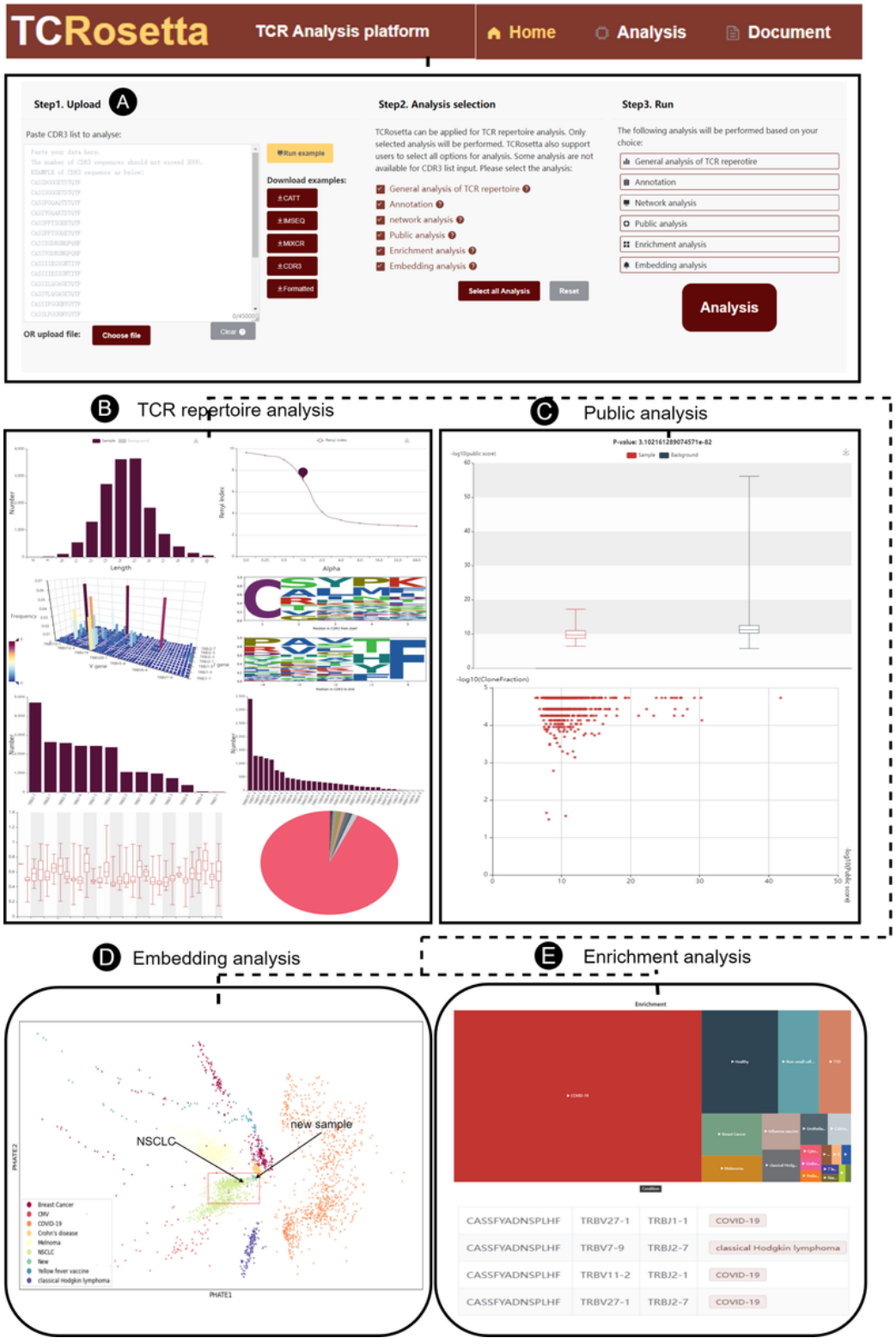
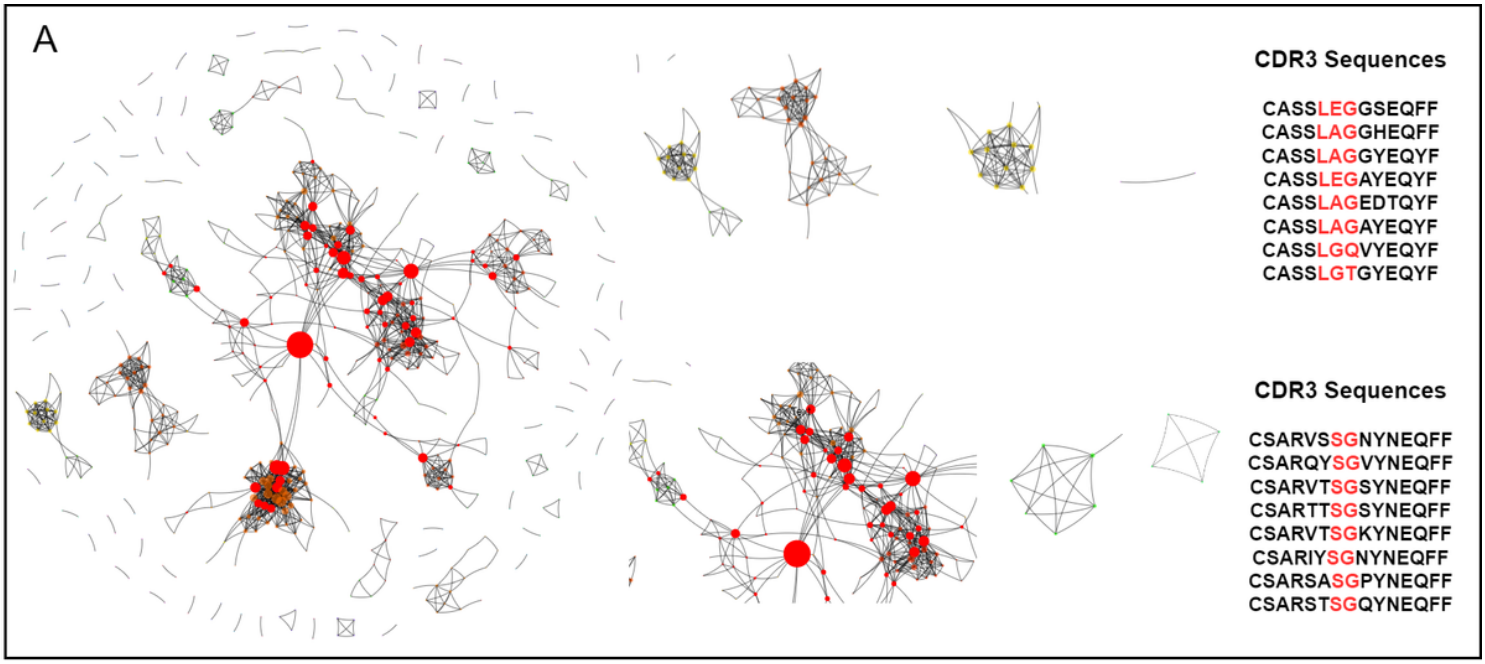


Figure 2

Functional web pages of TCRosetta. (A) Analysis page. Users can input their interested TCR sequences or upload file for analysis. (B) TCR repertoire general analysis. The TCR repertoire features of the upload file include V/J gene usage, diversity, clonality and so on. (C) Public analysis. Calculate the public score for TCR sequences. (D) The result of embedding analysis on an independent test set. The legend is at the bottom left of the picture and “New” in the legend represents the samples in the test set. The green points

in the red box are the result of embedding analysis for samples in the test set. (E) Annotation analysis and enrichment analysis, Users can annotate their interested TCR sequences and obtain enriched disease information for TCR repertoire.



B

CDR3 sequence	Vregion	Jregion	clonaFraction	Ratio
CASGRTKNTEAFF	TRBV3-1	TRBJ12-1	1.70369e-05	0.0035678
CASRLGQLKEQYF	TRBV2-1	TRBJ12-2	1.257e-05	0.010896
CASSLGDYEF	TRBV12	TRBJ3-5	3.7096e-05	0.0077309
CASSLLGGDROYF	TRBV4-1	TRBJ10	1.4796e-05	0.0217497
CSAARGLDQYEYF	TRBV3-7	TRBJ7-7	1.2576e-05	0.0035906

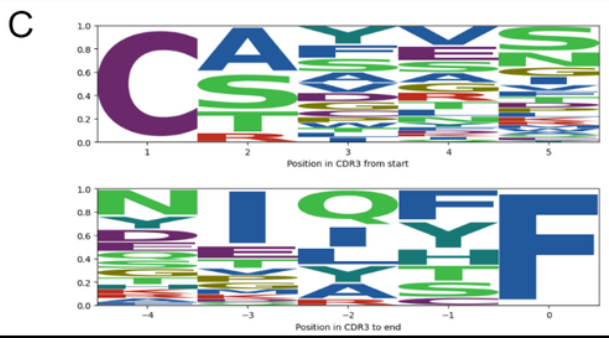


Figure 3

Results of network analysis in TCRosetta. (A) Network of TCRsequences. Red nodes are hub sequences. (B) Table of hub sequences. Each column contains CDR3 sequence, Vregion, Jregion, cloneFraction, and Ratio. (C) Sequence logo of hub sequences in the network.

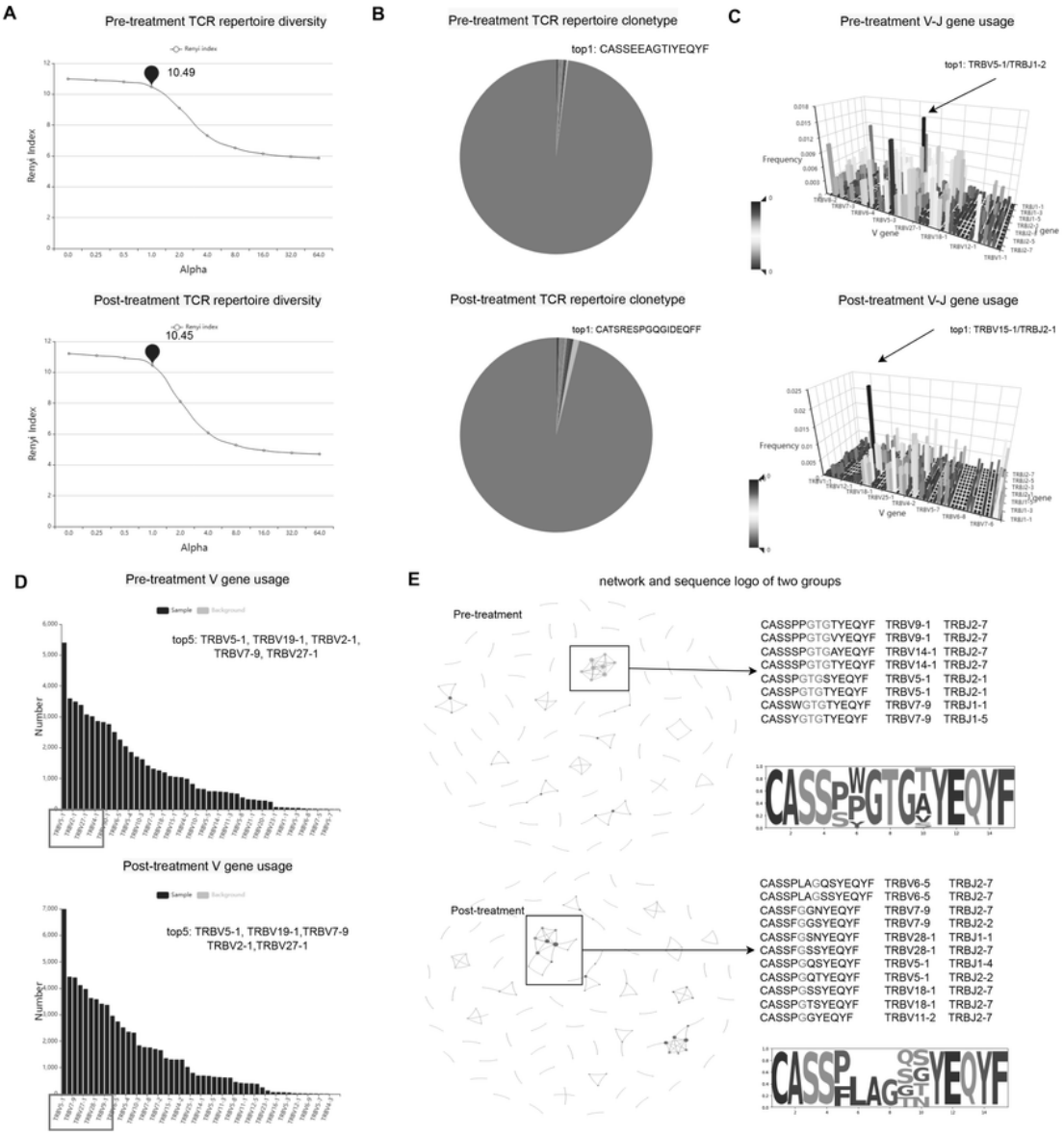


Figure 4

A case study analyzing T-cell receptor repertoire by TCRosetta. (A) The Renyi diversity plots of the pre-treatment group and the post-treatment group. The gradient of the slope increases as the distribution of the repertoire becomes more monoclonal. (B) The TRB CDR3 clone type distribution in the pre-treatment group and the post-treatment group. The top1 clone has been shown in the figure. (C) The TRB V-J gene pairing usage of two groups. The color represents the frequency, where darker color represents higher V-J

gene pairing usage. The top1 V-J gene pairing usage is shown in the figure. (D) The TRB V gene usage in two groups. (E) The network and sequence logo of the pre-treatment group and the post-treatment group. TRB CDR3 sequences and sequence logo of cluster in network (black box) are shown at the right of the network. Conserved CDR3 motif residues in the middle are highlighted in red in CDR3 alignments.