TCRklass: A New K-String–Based Algorithm for Human and Mouse TCR Repertoire Characterization

Xi Yang,*,†,[1] Di Liu,‡,[1] Na Lv,* Fangqing Zhao,§ Fei Liu,* Jing Zou,* Yan Chen,* Xue Xiao,* Jun Wu,* Peipei Liu,* Jing Gao,* Yongfei Hu,*,¶ Yi Shi,§ Jun Liu,*,‖ Ruifen Zhang,* Chen Chen,‖ Juncai Ma,‡ George F. Gao,*,§,¶,‖ and Baoli Zhu*,†,¶

The next-generation sequencing technology has promoted the study on human TCR repertoire, which is essential for the adaptive immunity. To decipher the complexity of TCR repertoire, we developed an integrated pipeline, TCRklass, using K-string–based algorithm that has significantly improved the accuracy and performance over existing tools. We tested TCRklass using manually curated short read datasets in comparison with in silico datasets; it showed higher precision and recall rates on CDR3 identification. We applied TCRklass on large datasets of two human and three mouse TCR repertoires; it demonstrated higher reliability on CDR3 identification and much less biased V/J profiling, which are the two components contributing the diversity of the repertoire. Because of the sequencing cost, short paired-end reads generated by next-generation sequencing technology are and will remain the main source of data, and we believe that the TCRklass is a useful and reliable toolkit for TCR repertoire analysis. ***The Journal of Immunology*****, 2015, 194: 446–454.**

The high-level diversity of TCR repertoire is essential for the human cellular immunity, and the size of the TCR repertoire is estimated to be at the magnitude of millions (1). It has been a challenge to determine the size and the diversity of the TCR repertoire in the past with conventional sequencing methods. The first direct estimate was done by Arstila et al. (2) in 1999 based on spectratyping method and Sanger sequencing of a specific band. A decade later, the deep sequencing methods were adopted to directly sequence the TCR repertoire and to determine the real size of the repertoire (3–11). A relatively accurate number of the real size of the TCRβ repertoire was obtained by Warren et al. (4), which is ∼1.3 million for PBMCs. However, the data analysis of the massive amount of sequences generated by deep sequencing is still challenging without proper bioinformatics tools, which are essential for accurate and fast identification of TCR sequences.

The target of TCR sequence analysis is to identify CDR3 sequences and to identify specific V gene segment and J gene segment. Such analysis is based on the sequence characteristics of regions flanking CDR3 (12–15). The CDR3 is the result of recombination of the V and J gene segments and the addition of random nucleotides, which is the major source of TCR diversity. The CDR3 has a conserved cysteine (C) residue on its N terminus, which is part of the V region, and a conserved phenylalanine (F) or tryptophan (W) residue on its C terminus, which is part of the J region. The flanking sequences of these two residues are also conserved. Thus, these two conserved residues can be identified by comparing with reference V and J sequences.

Although the direct sequencing of TCR repertoire using deep sequencing methods has been widely used in recent years, existing tools for the analysis of massive sequence data are not satisfactory. The IMGT is an integrated web toolkit for immunogenomics analysis, which provides tools like V-QUEST and HighV-QUEST, and reference datasets for the annotation of TCR and other important immune-marker genes (16–20). These tools would present illustrated annotations for each TCR gene sequence generated by Sanger sequencing and pyrosequencing. However, when facing short reads generated by Illumina sequencing platform, the sensitivity and specificity were much less satisfactory.

More recently, the iSSAKE-like strategy was developed by Warren et al. (4, 8–10). The algorithm is based on sequencing reads alignment with reference V gene and J gene segments for the identification of TCR sequences, as well as applying a 96% cutoff value on J gene segment diversity. However, the program can only analyze the assembled paired-end reads, and no software was released to the public. In contrast, the *IRmap* is a program that maps the sequencing reads to reference V and J gene segments (5). It is based on *pyromap* that is specifically designed for 454 pyrosequencing platform (21); thus, it does not fit with other sequencing platforms. The Decombinator is a toolkit for analyzing TCR from short reads (22). It is implemented using Python language that is

*Chinese Academy of Sciences Key Laboratory of Pathogenic Microbiology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China; †University of Chinese Academy of Sciences, Beijing 100049, China; ‡Network Information Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China; §Beijing Institutes of Life Sciences, Chinese Academy of Sciences, Beijing 100101, China; ¶Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Zhejiang University, Hangzhou 310003, China; and ‖Chinese Center for Disease Control and Prevention, Beijing 102206, China

[1]X.Y. and D.L. should be regarded as joint first authors.

Received for publication May 9, 2014. Accepted for publication October 16, 2014.

This work was supported by National Basic Research Program of China ("973" project) Grants 2013CB531500 and 2015CB554204.

Address correspondence and reprint requests to Prof. Baoli Zhu, Prof. George F. Gao, or Prof. Juncai Ma, Chinese Academy of Sciences Key Laboratory of Pathogenic Microbiology, Institute of Microbiology, Chinese Academy of Sciences. No. 1 Beichen Road, Chaoyang District, Beijing 100101, China (B.Z., G.F.G.) or Network Information Center, Institute of Microbiology, Chinese Academy of Sciences. No. 1 Beichen Road, Chaoyang District, Beijing 100101, China (J.M.). E-mail addresses: zhubaoli@im.ac.cn (B.Z.), gaof@im.ac.cn (G.F.G.), or ma@im.ac.cn (J.M.)

The online version of this article contains supplemental material.

Abbreviations used in this article: FN, false negative; FP, false positive; HCL, hierarchal clustering analysis; KN, k-string number; NCBI, National Center for Biotechnology Information; Nmk, number of matched K-strings; nNmk, number of matched nucleotide k-strings; Scr, conserved residue support score; SRC, Spearman rank correlation; SRD, Spearman rank distance; TN, true negative; TP, true positive.

much slower than the compiled languages, and its accuracy is only 88%, which is not optimal. The MiTCR is a recently released software that can process massive sequencing reads (23). To find the CDR3 sequence, the algorithm of MiTCR separates high- and low-quality reads in the first step, and then merges low-quality reads to high-quality CDR3 clusters. The whole process misses the V and J sequences without CDR3, which may cause bias for V and J annotations because using only the part of the V and J sequence contained within CDR3 is not adequate for annotate-specific V/J segments.

In this study, we developed a novel algorithm that is based on k-string matching for the identification of CDR3 and V/J gene segments of TCR from millions of short reads. The toolkit, named TCRklass, can work on both assembled and nonassembled paired-end reads, and does not require the CDR3 sequence to identify the V and J gene segments. The TCRklass has shown better sensitivity and specificity than IMGT/HighV-QUEST and MiTCR.

Materials and Methods

Test datasets

The reference human/mouse V/J nucleotide/protein TCRα/β sequences were downloaded from IMGT/GENE-DB. These V and J reference sequences were aligned using MAFFT and indexed using the approach described later in the *Algorithm* section. All alleles of the nucleotide sequences of the V and J gene segments were used for V and J identification, and only the first allele of the amino acid sequences was used for CDR3 identification.

To determine the cutoff value on number of matched K-strings (N_{mk}) for V and J identification, we constructed a test dataset by randomly mutating reference V and J nucleotide sequences. For each reference V and J sequences, 10 sequences were generated and each has 1/20 of bases mutated.

The TCRα/β sequencing reads from three mouse samples and two human samples were used as test dataset (Table I), which were picked from two large-scale studies on human and mouse TCR repertoires. These samples were amplified by 5′-race and 3′-primer in C region, the cDNA were fragmented by sonication, and DNA segments of 150–300bp were gel purified for subsequent sequencing. A total of 40,000 sequences were randomly picked from the datasets mm5a, mm5b, hs1a, and hs1b (10,000 from each) and were manually curated.

By searching the terms of "human TCR CDR3" in National Center for Biotechnology Information (NCBI), a total of 572 CDR3-containing TCR sequences were retrieved. After removing sequences containing frame shift and stop codon in CDR3 region, we used 550 sequences as positive control dataset. In addition, a total of 200,000 sequencing reads were randomly picked from human transcriptome library SRR836529 and were used as negative control dataset.

To evaluate TCRklass on a different dataset, we downloaded the human TCRβ sequencing library SRR060714 (part of sample M1D1) from the NCBI Sequence Read Archive using TCRklass.

Algorithm

In brief, the CDR3 and the V and J identification algorithm in TCRklass are based on substring (k-string) matching. For CDR3 identification, the position of reference V and J sequence k-strings is used to find the CDR3 position on query sequencing reads. For V and J identification, the N_{mk} performs as a measurement of similarity between query sequencing reads and reference V and J sequences, and the most similar V and J gene segment is selected to classify the reads.

Reference database indexing. All 3-strings for each reference sequence of the V and J gene segments were traversed and stored with their position relative to the conserved residue ("C" for V segment and "F(GXG)" for J segment). The database is composed of K-string profiles, and each profile represents to one reference sequence. For V/J identification, 6-strings of nucleotide sequences were used.

Query sequence preprocessing. Before the analysis by TCRklass, the sequencing reads were first assembled in paired ends, producing two subdatasets for each sample: one subdataset of assembled reads and another subdataset of nonassembled read. Both the assembled and the nonassembled subdatasets were used in the following analysis.

V and J gene segment identification. The algorithm for V and J identification uses N_{mk} as the measurement of the distance with reference V and J segments. The query sequences are compared with 6-strings of reference V and J nucleotide sequences, and the V and J gene segment with the highest N_{mk} is selected. If a query sequence is classified to several V and J gene segments that have the same N_{mk}, this query sequence is considered to be ambiguously classified, and those V and J gene segments are all written to output.

CDR3 identification. The algorithm of CDR3 identification has two stages. First, query sequences are translated on all six frames and compared with the 3-string profiles of reference V and J amino acids sequences. The translation frame and the reference profiles with the most N_{mk} were selected for following analysis. Then the position of conserved residue ("C" for V segment and "F(GXG) or W(GXG)" for J segment) on the query sequence is determined by the position of the matched 3-strings. To avoid the disturbance caused by repeated 3-strings, the position of the conserved residue was determined by the number of supported 3-strings. To quantify the concept of "being supported by most 3-strings," we defined a numeric metric named "conserved residue support score (S_{cr})" for each residue on query sequence. Denoting the distance between a matching 3-string to the conserved residue by d_k, the support score is calculated as: $\Sigma(1/d_k)$. The candidate conserved residues with the highest S_{cr} is selected, and the CDR3 region is located between the two conserved residues in V and J gene segments.

Summarizing the raw annotation results. The nucleotide and protein sequences of the CDR3 region for each sequencing read are presented with sequencing quality, and the abundance profile of unique CDR3 clonotypes are summarized in two passes. In the first pass, only the CDR3 sequences whose base qualities are all high are used (quality score 20, by default). The CDR3 clonotypes are generated using these high-quality sequences, and the abundance is simply summed. In the second pass, the low-quality CDR3 sequences are compared with CDR3 clonotypes. If there exists clonotype with identical sequence, it is counted into that clonotype; otherwise, the sequence is deprecated. Unique TCR types are generated by combination of CDR3 sequence and the type of V and J gene segment.

Assessment of errors of sequencing reads. The errors of sequencing reads include: PCR error, sequencing error, and errors in mRNA synthesis. The errors of sequencing reads are assessed using two parallel approaches: 1) a modified CDR3 identification program is run, which allows variations on CDR3 C terminus, and the error rate is calculated by the portion of bases that do not belong to the codon of the conserved Cys residue; and 2) along J sequence, the base composition is summarized for each type of J gene segment. On each position, the most abundant base is considered to be the correct one, and the error rate is calculated by the portion of the other three bases. To handle with possible polymorphism induced by diploid, if the ratio of the most abundant base versus the second abundant base is lower than 3:2, they are both treated as correct bases, and the error rate is calculated by the portion of the other two bases. Both approaches use only those nucleotides whose quality score is >20. The error rate of the whole CDR3 sequence is calculated by powering nucleotide error rate with CDR3 average length: $\mathrm{Error}_{CDR3} = 1 - (1 - \mathrm{Error}_{nucleotide})^{\mathrm{CDR3\ length}}$.

Evaluation of algorithm parameter

K-string size. A series of k-string length (3–9 for nucleotides, 2–6 for amino acids) was tested on the reference V and J sequences and on the sequencing reads of the four test datasets. On each length, the numbers of unique and repeated k-strings on each sequence were recorded. For the sequencing reads, amino acids sequences were translated from all six frames; and for the reference V and J sequences, the amino acid sequences are defined by IMGT. The optimum size of K-string was selected as the possible lowest size that has few amount of repeat.

N_{mk} and S_{cr} cutoff value for CDR3 identification. The randomly picked 4 × 10,000 sequences from four datasets were manually curated one by one to identify the CDR3, to generate a reference dataset of CDR3. The curation was performed by several steps. First, the 8 aa surrounding the conserved residue of the reference V and J sequences were compared with the query sequences, and the CDR3 will be identified if the query sequence can align with both V and J. After that, the remaining unidentified sequences were searched using regular expressions and alignment with reference V and J sequences using NCBI blast.

The 4 × 10,000 sequences were analyzed using TCRklass, and the CDR3 identified by TCRklass were compared with the manually curated ones to determine whether the results are true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The precision and the recall rates were calculated as: precision = TP/(TP + FP), recall = TP/(TP + FN).

The identified CDR3 are filtered by a range of N_{mk} and S_{cr} (2–7 for N_{mk}, 1.0–2.9 for S_{cr}), to determine the best N_{mk} and the S_{cr} cutoff value for CDR3 identification, and a series of precision and recall rates was calculated on these cutoff values. The combination of N_{mk} and S_{cr} was selected to compromise the precision and the recall rate.

N_{mk} cutoff for V/J identification. A series of k-string numbers (KNs; 5–50 for V, 5–30 for J) were tested for whether they can distinguish a V or J

segment from others. On each given KN N, all the combinations of N k-strings in each sequence of the mutated V and J reference dataset were traversed. If a combination of N k-strings exists in any of the sequences of different V or J gene segments, they cannot distinguish this V or J from others. The number of nondistinguishable combinations was recorded, and the probability of nondistinguishable was calculated for each V and J gene segment on each N k-string. The N_{mk} cutoff value was selected by the lowest N that have such probability <0.05.

Comparison with other TCR analysis tools

The IMGT/HighV-QUEST was applied on the four 10,000-sequence manually curated datasets (18, 19). Because the IMGT/HighV-QUEST cannot process nonassembled paired-end reads, these reads were concatenated together by eight N's. The numbers of TP, FN, FP, and TN CDR3 sequences were obtained by comparing with manually curated CDR3, as described in the previous section. The 550 positive control sequences and 200,000 negative control sequences were also analyzed using IMGT/HighV-QUEST and obtained the numbers of TP, FN, FP, and TN CDR3 sequences.

The 10 datasets from three mouse samples and two human samples were analyzed using MiTCR (23). Because the MiTCR also cannot process nonassembled paired-end reads, these reads were concatenated together by eight N's. The abundance profile of CDR3 amino acid sequences (CDR3 profile) and V and J combinations (V/J profile) were compared with those by TCRklass. Specifically, the V/J profiles by MiTCR were compared with the V/J profiles by TCRklass generated using all reads and using CDR3-containing reads. The V/J profiles are hierarchically clustered using Spearman rank distance (SRD) using MeV version 4.8.1. The manually curated 40,000 sequences, 550 positive control sequences, and 200,000 negative control sequences were also analyzed using MiTCR. The numbers of TP, TN, FP, and FN CDR3 sequences were calculated, and the precision and recall rates were compared with TCRklass.

The library SRR060714 was analyzed using TCRklass. The V/J profile and CDR3 profile by TCRklass were compared with those by the original study. The most abundant 10 unique CDR3 sequences whose abundance by TCRklass is 100 times higher than that by the original result (or vice versa) were picked and had their sequencing reads manually curated.

Software implementation

The programs were written in C++ language and used the Standard Template Library and Boost library for data structures and command-line option parsing. The programs that are not computing intensive were written in Perl. The TCRklass software package has four major programs for CDR3 identification and V and J identification for assembled and nonassembled paired-end reads. Besides the four major programs, TCRklass also contains several auxiliary programs for summarizing results, indexing reference sequences, and assessing PCR error. A bundle program, tcrklass.pl, performs all analysis using one command. The toolkit and indexed database can be downloaded at: http://sourceforge.net/projects/tcrklass; the prerequisite HTQC package can be downloaded at: http://sourceforge.net/projects/htqc.

Results

Computational pipeline for TCR sequence analysis

The TCRklass, as shown in Fig. 1, consists of four steps that include: 1) the quality check and reads assembly, 2) the V and J identification, 3) the CDR3 identification, and 4) the results integration. The quality-check process is to discard those low-quality reads, and the high-quality reads are assembled in read-pairs (see the *Materials and Methods* section for details). The original read pairs are classified into two distinct datasets: the assembled reads pairs and the nonassembled ones. Notably, the whole process in this step was performed by HTQC (24).

Then the two datasets are annotated for V and J gene segments and CDR3 sequences in parallel using k-string algorithm. For the V and J identification, reference sequences of V gene and J gene segments are collected from IMGT database and indexed into 6-string matrices. Then the 6-string matrix for each V and J sequence is mapped to each query sequence from both assembled and nonassembled datasets and a score is given. The reads are then classified into a given V or J segment, whichever has the highest score. Similarly, for the CDR3 identification, another set of matrices for V and J genes is generated based on the amino acid sequences and then mapped to the reads translated in all six frames. The in-framed "C" and motif "FGxG" are located and scored, and the segment with the highest score is assigned as the CDR3.

The last step of the pipeline is the integration of results of V/J and CDR3 identification, where the abundance profiling of CDR3 sequences and the V/J gene segments are summarized concerning the sequencing quality. The results are stored in plain-text tables, and a set of colored illustrations was also provided, including the heat maps of V/J combination, CDR3 frequency, and some statistics (Fig. 2).

Datasets used for TCRklass evaluation

To evaluate the performance of TCRklass, we sequenced five TCRα and TCRβ repertoires using Illumina HiSeq2000 from 2 human and 3 mouse T cell samples (10 datasets in total, 5 for α-chain and 5 for β-chain).

For the evaluation of CDR3 identification and installment of the parameters for CDR3 identification, four manually curated datasets were selected from one human and one mouse sample, hs1a, hs1b, mm5a, and mm5b (10,000 sequences from each). In addition to the test using manually curated data, we fetched 550 CDR3-containing TCR sequences as positive control from NCBI, in which 94 sequences are TCRα and 456 are TCRβ. Human RNA-seq data (SRR836529) were used as negative control, where 200,000 of 33 million reads were randomly picked.

To determine the parameters for V/J identification, we generated artificial sequences by inducing mutations on reference V and J sequences, where each sequence generates 10 mutated sequences and each have 1/20 of the residues randomly mutated.

Setting the parameters for the K-string matching algorithm

Both the CDR3 and the V/J identification are based on K-string matching algorithm. Therefore, the size of K-strings has to be defined, where a longer K-string will reduce the number of repeated K-string at the expense of exponential increase of the memory consumption. Thus, an optimum K-string size needs to be defined to find a balance between the number of repeated K-string and the memory consumption. The determination of optimum K-string size for TCRklass was carried out by visualizing the number of repeated K-strings in a range of different sizes. It can be observed that the fraction of nonrepeated K-strings increases rapidly while the K-string size is increased. And when the size of the K-string is increased to 6 for nucleotides and 3 for amino acids, the portion of the repeated K-string became very low (Fig. 2A, 2B). Therefore, we decided to use 6-string nucleotides for V and J identification and 3-string amino acids for CDR3 identification.

In the process of CDR3 identification, the N_{mk} describes the similarity of the query sequences with reference V/J sequences, whereas the S_{cr} is introduced in this article to describe the reliability of the conserved C and F (GXG) residues in the boundary of the CDR3 sequence. To decide the cutoff value of N_{mk} and S_{cr}, we analyzed a manually curated subdataset of 10,000 sequences using TCRklass on a series of different N_{mk} and S_{cr} cutoff values, and the precision and recall rates were calculated by comparing TCRklass analyzed result with the manually curated result. The precision and recall rates were plotted by those N_{mk} and S_{cr} cutoff values (Fig. 2C), and the cutoff values of $N_{mk} = 3$ and $S_{cr} = 1.7$ were selected for an optimal precision rate in expense of a higher recall rate.

In the process of V/J identification, the number of matched nucleotide k-strings (nN_{mk}) describes the similarity of query sequence with reference V and J sequence at nucleotide level, and the sequences are classified to the V and J gene segment with highest nN_{mk}. To define nN_{mk} cutoff value, we tested a series of KNs for their ability of distinguishing specific V/J gene segment type using the artificial dataset generated from the reference V/J sequences.

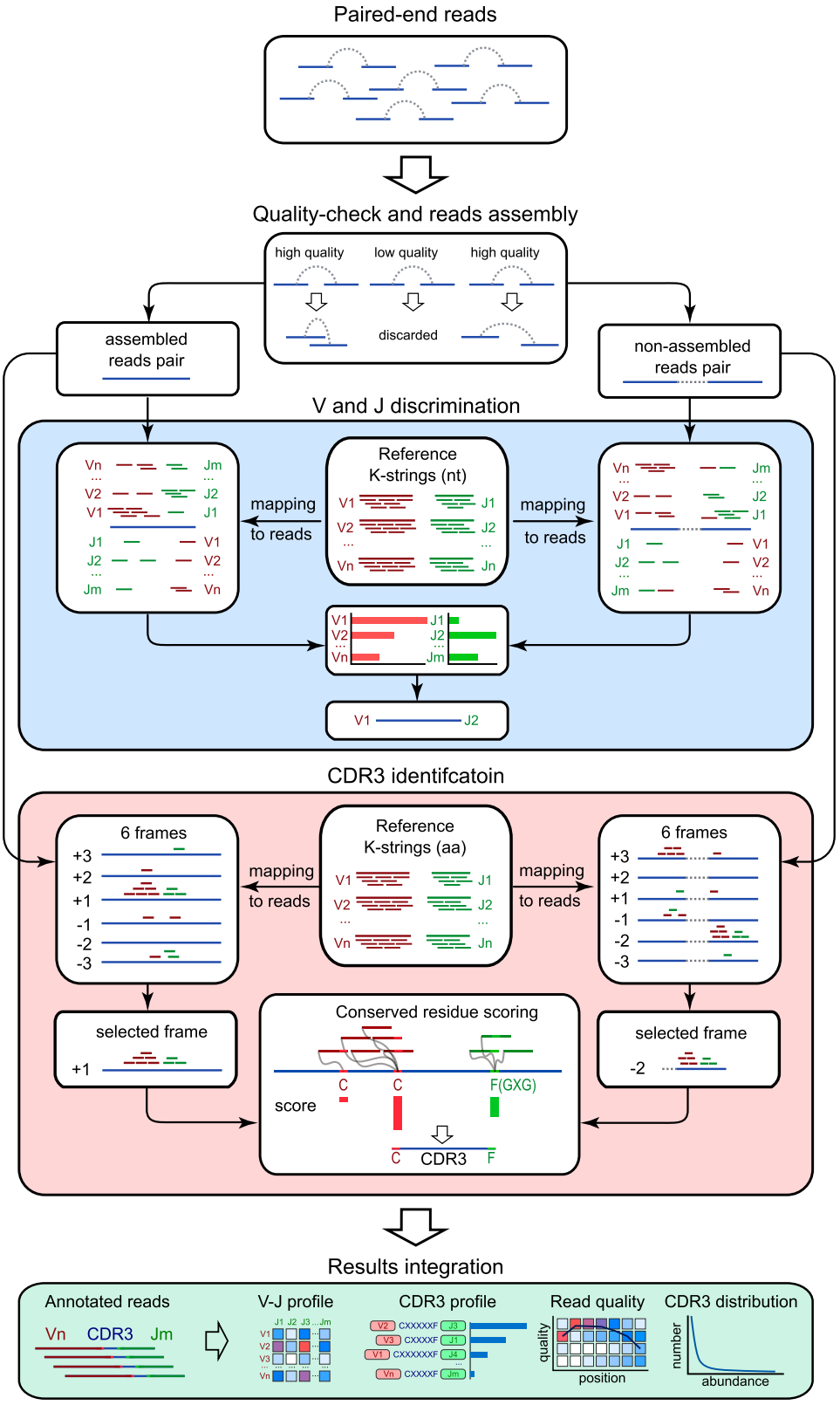

FIGURE 1. TCRklass pipeline.

On each KN, different combinations were tested, and those who cannot identify one type of V/J sequence from other types of V/J sequences were counted (Fig. 2D). The probability (p value) of KN is calculated by dividing the number of undistinguishable combinations by the number of total combinations. The cutoff value for nN_{mk} was defined by the lowest KN whose p value is <0.05. It can be observed that different V/J gene segments have different nN_{mk} cutoff values. Such difference is probably caused by the nature of the sequences that have different identity between different V/J gene families.

Precision and recall rates of TCRklass

The precision and recall rates of CDR3 identification were calculated by comparing the results obtained by TCRklass and by

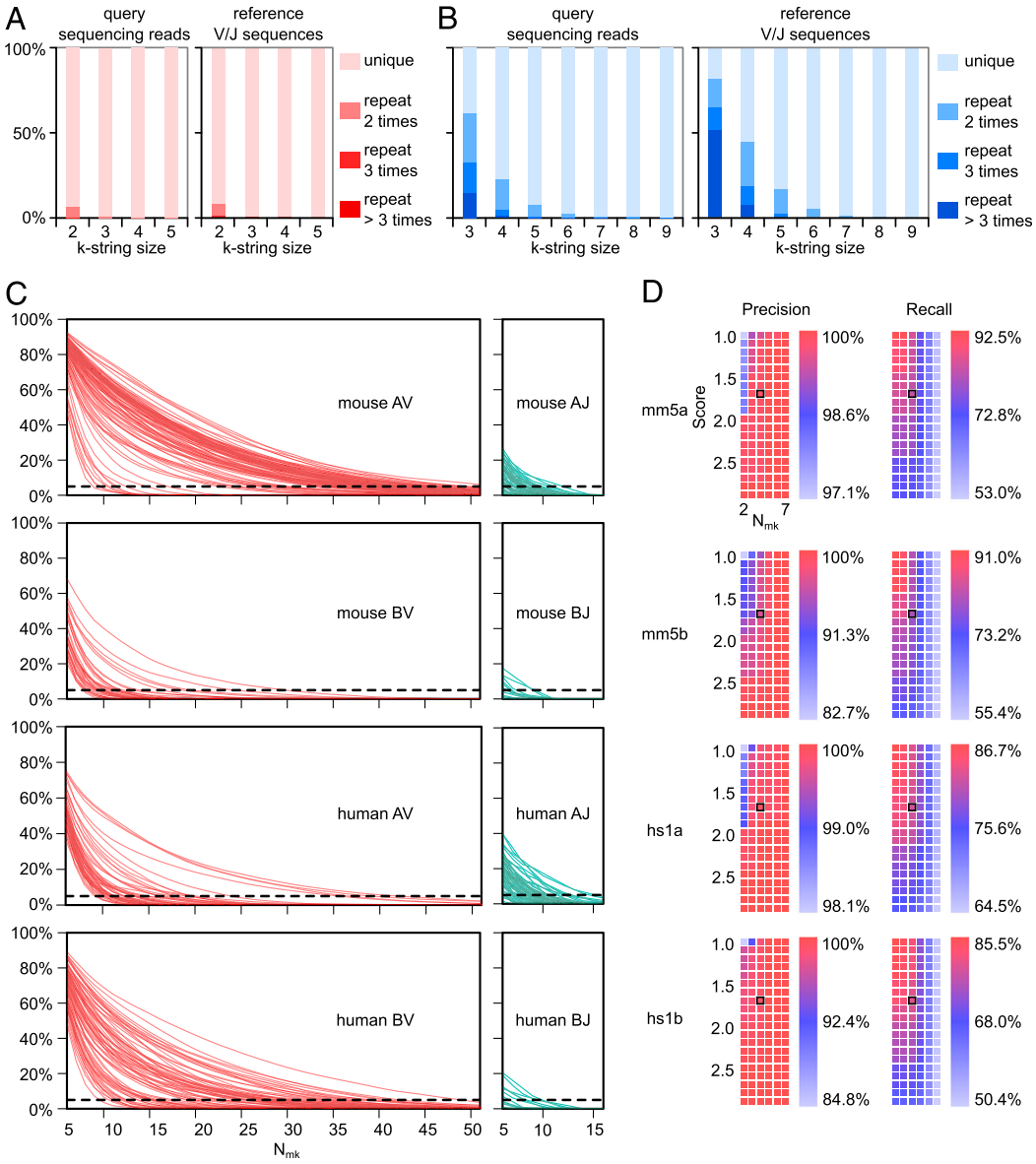

FIGURE 2. Defining the parameters for TCRklass. (**A**) The definition of k-string size for amino acids and (**B**) the definition of k-string size for nucleotide sequences. (**C**) The probability of unrecognizing 6-string combinations, plotted by numbers of 6-string nucleotides. (**D**) The precision-recall rates for CDR3 identification, plotted by different N_{mk} and S_{cr} cutoff values.

manual curation using a randomly picked subset of 40,000 sequences from the datasets mm5a, mm5b, hs1a, and hs1b (Table I). The precision rate for CDR3 identification is 98.5% for α-chain and 99.0% for β-chain in mouse, and 99.8 for α-chain and 98.3% for β-chain in humans. The precision rates have no significant differences between the two individuals. The recall rate for CDR3 identification is 84.6% for α-chain and 87.9% for β-chain in mouse, and 91.4% for α-chain and 93.0% for β-chain in humans. The recall rates for most of the nonassembled reads are lower than those for the assembled reads. As the nonassembled reads are shorter than the assembled ones, and the length distribution of CDR3 is the same (they are from the same sample), the flanking sequences around CDR3 are shorter and harder to be recognized by TCRklass program, but some of them are still recognizable by manual curation. A list of exemplary false-negative sequences is demonstrated in Supplemental Fig. 1.

We also evaluated TCRklass using the 550 positive control sequences and 200,000 negative control sequences downloaded from NCBI. Among the 550 CDR3-containing sequences, 538 were identified by TCRklass. The 12 missed sequences were manually checked, and we found that most of them have very short V region (Supplemental Fig. 2). In contrast, TCRklass identified only one FP CDR3 from negative control sequences. The precision rate of these data is 99.8% and recall rate 97.8%.

TCRklass application on TCRαβ repertoire

All 10 datasets of the repertoires were analyzed using TCRklass to identify CDR3 and V and J gene segments (Table I), and one representative dataset was plotted in Fig. 3A and 3B. For each dataset, a part of the raw reads can be assembled in paired-end reads. The percentage of assembled reads is 63.9 and 29.9% on average for mouse and human samples, respectively. The differences of assembled reads between human and mouse samples might be caused by the discrepancy in DNA library construction. On average, 98.8% of the sequences can be identified to a known reference V or J sequence, indicating that our datasets are mainly composed of TCR sequences. In contrast, the percentages of sequences that have CDR3 identified are 5.6 and 2.6% for mouse

Table I. Test results for TCR repertoire for human and mouse datasets using TCRklass

Dataset	No. of Raw Reads	CDR3 Identified Reads		V/J Identified Reads		CDR3 Precision Rate (%)	CDR3 Recall Rate (%)
		No.	%	No.	%		
mm4a	11,401,155	698,618	6.10	11,286,572	99.00		
mm4b	5,188,006	143,802	2.80	5,175,216	99.80		
mm5a	11,784,033	641,798	5.40	11,663,915	99.00	98.5	84.6
mm5b	6,669,141	139,771	2.10	6,650,024	99.70	99.0	87.9
mm6a	12,023,471	629,978	5.20	11,893,532	98.90		
mm6b	5,578,210	161,512	2.90	5,547,089	99.40		
hs1a	11,257,528	1,232,672	10.90	11,207,381	99.60	99.8	91.4
hs1b	9,533,846	339,744	3.60	9,293,121	97.50	98.3	93.0
hs2a	4,411,224	273,120	6.20	4,389,733	99.50		
hs2b	6,047,973	214,769	3.60	5,802,504	95.90		

α- and β-chain, and 9.6 and 3.6% for human α- and β-chain, respectively. There are more CDR3 sequences identified in α-chain than in β-chain, and the reason for that is unknown. Nevertheless, the assembled subdatasets have higher percentage of CDR3 identified than the nonassembled subdatasets, because these assembled paired-end reads are longer and with higher probability of having a complete CDR3 sequence. The percentage of sequences that can be classified to specific V or J gene segments is much higher than the percentage of CDR3 identified. The reason for such difference could be complicated, and one explanation is that the identification of V and J does not require complete V or J sequence.

From the datasets of α- and β-chains, the abundance for each of the V/J combinations and CDR3 sequences (V/J and CDR3 profiles) was calculated by the number of reads, and the profiles from different datasets were compared. Among the three mouse samples, the average Spearman rank correlations (SRCs) of V/J profiles are 0.968 and 0.933 for α- and β-chains, respectively, whereas those of CDR3 profiles are 0.237 and 0.074. It can be observed that the V/J profiles are more correlated between samples than the CDR3 profiles. For the two human samples, the SRC of V/J profiles is 0.908 for α-chain and 0.922 for β-chain, whereas the SRC of CDR3 profiles is −0.733 for α-chain and −0.806 for β-chain, respectively. The SRCs of CDR3 and V/J profiles for human samples are both lower than those of mouse samples, which indicates that the two human samples have more divergent TCR sequences.

The sequence reads generated have two types of reads, one with CDR3 (CDR3-containing) and the other one without. To verify the effect of the CDR3-containing sequences for the V/J profile, we calculated the SRCs between the V/J profiles generated by all sequences and by CDR3-containing sequences. The hierarchal clustering analysis (HCL) was performed according to SRD between the profiles (Fig. 3E for mouse TCRα and Supplemental Fig. 3D–F for others), and the V/J profiles are clustered together according to whether they have CDR3 sequences. Such a difference indicates that only using CDR3-containing reads for V/J profiling is insufficient, and this can be explained by that the $CDR3^+$ reads have shorter V sequences, which hampered the classification of V gene segments.

The errors of sequencing reads were assessed by base composition on the conserved Cys residue on C terminus of CDR3 sequence, and by base composition of each J gene segment. The average nucleotide error rate assessed by J gene segment polymorphism is 0.46% (Supplemental Table I). The error rate on J sequence 5′-end is lower than 3′-end. The reason is possibly that the paired-end reads overlap on this region, and provides extra nucleotide proofs. The average nucleotide error rate assessed by CDR3-conserved C terminus is 0.41% (Supplemental Table II), which is very similar to that assessed by J sequence. The error rate for the whole CDR3 sequence can be calculated by powering the per-base reliability with average CDR3 length (~40 nt), which is ~15.15–16.84%. The error CDR3 sequences take a notable part of total sequences, but they still cannot be excluded from the CDR3 dataset, because we cannot determine which CDR3 sequences are the false ones.

In addition, the sequencing reads from library SRR060714, published by Warren et al. (4), were analyzed using TCRklass, in which 98.5% of them have complete CDR3 region and all reads have V or J gene segments classified. Comparing with the V/J profile and CDR3 profile generated by the original study (Fig. 3F, 3G), we found that our V/J profiles are highly proportional (SRC = 0.854), whereas our CDR3 profile is not (SRC = −0.294). Specifically, the low-abundance CDR3 sequences are less proportional. The reason for such differences is probably that only one library of sample M1D1 was analyzed in our study, whereas the abundance profile of the original study was generated using all sequences of sample M1D1.

Time and memory consumption of TCRklass

The time consumption of running the CDR3 identification is summarized using scaled amount of input sequences. It can be observed that the time consumption is in proportion to the amount of input sequences and the number of reference V and J sequences (Supplemental Fig. 4A). The average time for processing each input sequence with one reference sequence is calculated to be 3.7–7.4 μs. The nonassembled paired-end reads use more time than the assembled paired-end reads, because the nonassembled read actually has two sequences. For V and J identification, the average time for processing each input sequence with one reference sequence is 2.3–4.7 μs, which is shorter than that of CDR3 identification.

The memory consumption for CDR3 identification is only correlated with the number of reference V and J sequences used, because the input sequences are processed one by one and will not take much memory (Supplemental Fig. 4B). Each K-string profile of the reference sequences needs ~500 kilobytes of memory. This memory size is much larger than the length of the reference sequences, that is, the K-string lookup table for each reference sequence contains all potential K-strings to speed up the search process (26^3 for amino acids and 5^6 for nucleotides). The memory consumption for V and J identification is similar with that of CDR3 determination.

Precision and recall rates in comparison with IMGT/HighV-QUEST using curated datasets

Among all existing tools, the IMGT/HighV-QUEST (HighV-QUEST for short) is the most widely used. To compare the performance of HighV-QUEST and TCRklass, we analyzed the four datasets of manually curated 4 × 10,000 reads using HighV-QUEST. The identified CDR3 sequences were compared with

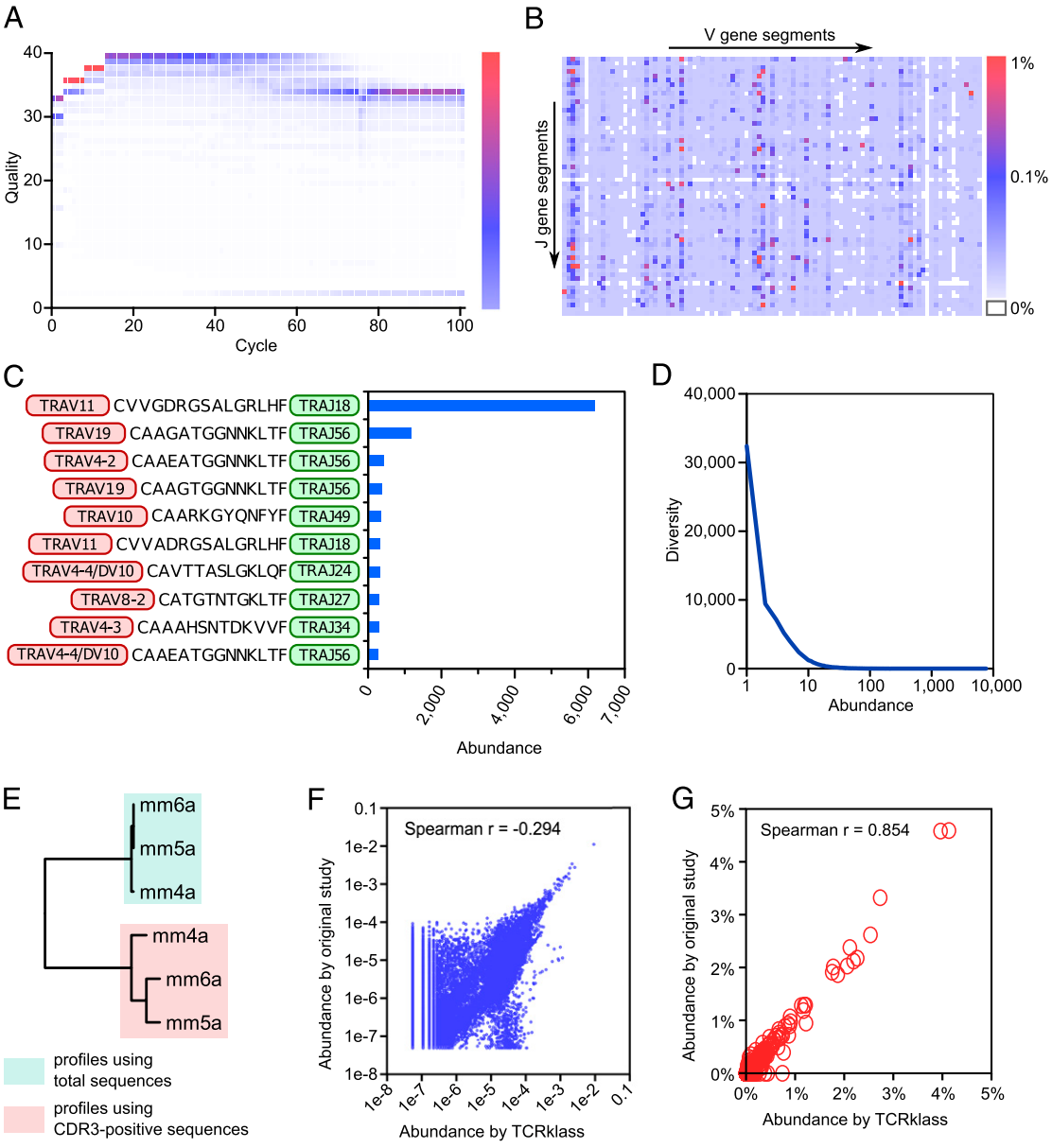

FIGURE 3. Detailed explanation of TCRklass test on dataset mm5a. (**A**) The sequencing reads quality heat map. (**B**) The abundance of V-J combinations. (**C**) The abundance of 10 most frequent TCR clonotypes. (**D**) Number of unique CDR3 sequences in different abundance. (**E**) The HCL tree for V/J pairing profiles in three mouse TCRα samples. The profiles were generated using total sequences (green) and CDR3-containing sequences (red). (**F**) Comparison of the abundance of CDR3 types in amino acid sequence identified by TCRklass (*x*-axis) and by Warren et al. (*y*-axis). (**G**) The abundance of V/J combinations by TCRklass and by the original study (4).

the manually curated ones to determine the number of TP, TN, FP, and FN results, and the precision and recall rates were calculated. On average, the precision rate by HighV-QUEST is 71.4% and the recall rate is 22.5%, which are both much lower than those by TCRklass (99.2 and 89.6%, respectively; Fig. 4A). In all four samples, there are more TP and less FP results by TCRklass than by HighV-QUEST (Fig. 4B).

We further manually examined several HighV-QUEST FP reads (Supplemental Fig. 5) and found the FP CDR3 on these reads are all located near 3′-end and have suspicious J signal, or have frame shifts within the CDR3 sequence that were arbitrarily fixed by HighV-QUEST. Therefore, comparing with HighV-QUEST, the CDR3 identification is improved in TCRklass, especially for short sequences, and TCRklass is more stringent on recognizing V/J sequences. We also applied HighV-QUEST on the 550 positive control sequences and the 200,000 negative control sequences, in which 465 TP and 4 FP CDR3 sequences were identified (Fig. 4C). The precision rate of these sequences was 99.15% and recall rate was 84.55%, which were also less accurate than those by TCRklass (99.81 and 97.82%, respectively).

Improved CDR3 and V/J identification over MiTCR in human and mouse TCR repertoire

For comparative analysis purposes, we ran TCRklass and MiTCR in parallel with 10 human and mouse datasets. To evaluate the precision and recall rates, the manually curated 4 × 10,000 reads were analyzed using MiTCR in parallel with TCRklass (Fig. 4A, 4B). The MiTCR has precision rate of 95.2% and recall rate of 67.1%, which are lower than those of TCRklass (99.2 and 89.6%), showing a better performance of TCRklass on large datasets. However, when we applied MiTCR on the positive control and the negative control datasets (Fig. 4C), the recall rate of MiTCR was

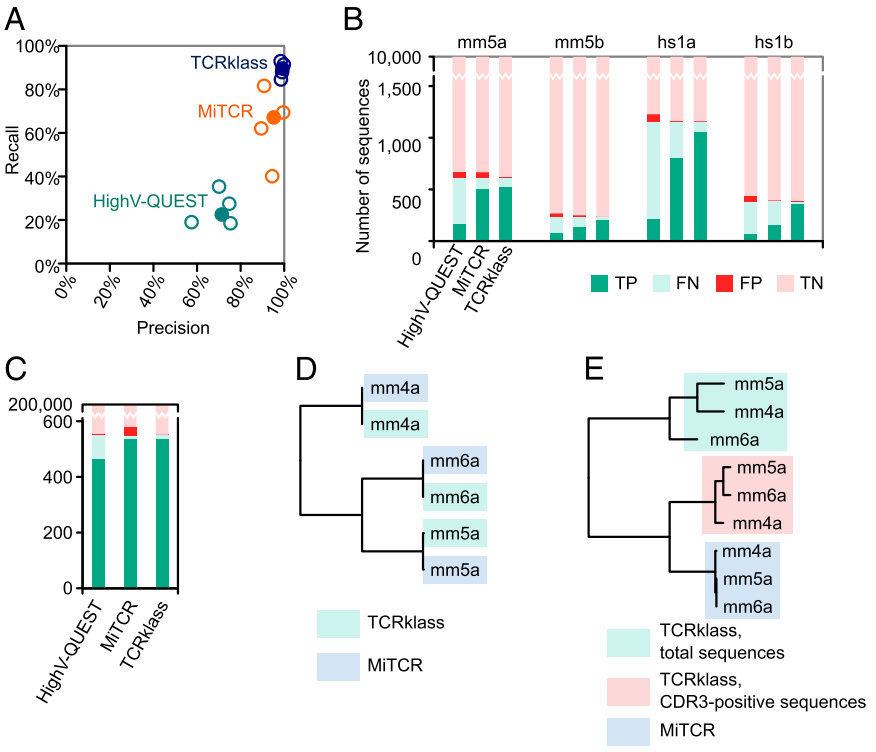

FIGURE 4. Comparison of TCRklass with other tools. (**A**) The precision and recall rates of CDR3 identification by IMGT/HighV-QUEST (green), MiTCR (orange), and TCRklass (blue). (**B**) Number of TP, FN, FP, and TN CDR3 sequences by IMGT/HighV-QUEST, MiTCR, and TCRklass, using test datasets mm5a, mm5b, hs1a, and hs1b. (**C**) Number of TP, FN, FP and TN CDR3 sequences by IMGT/HighV-QUEST, MiTCR, and TCRklass, using the 550 positive control and 200,000 negative control sequences. (**D**) The HCL tree for profiles of CDR3 types in amino acid sequence of three mouse TCRα samples. The profiles were generated using TCRklass (green) and MiTCR (blue). (**E**) The HCL tree for profiles of V/J pairing of three mouse TCRα samples. The profiles were generated using total sequences by TCRklass (green), CDR3-containing sequences by TCRklass (red), and MiTCR (blue).

similar to that of TCRklass; even the precision rate is 5% lower (94.9%). Among all these tests, the precision rate of MiTCR is lower than that in the original study (~1% erroneous CDR3). This is probably because the original evaluation was performed using in silico–generated artificial data, which have no other sequences than $CDR3^+$ ones. This is very different from the actual sequencing reads, where the lack of CDR3 sequences can induce noises into the analysis procedure. Therefore, MiTCR could be prone to noises produced by large datasets of high-throughput sequencing.

Furthermore, the CDR3 abundance profiles were generated at amino acids level. Compared with the profiles by TCRklass, the average SRCs are 0.874 and 0.900 for human α-chain and β-chain, respectively, and 0.865 and 0.820 for mouse α- and β-chain, respectively. We further performed HCL on the profiles by SRD (Fig. 4D, Supplemental Fig. 3A–C) and found that these profiles are closely clustered according to human or mouse samples, which indicates the difference on CDR3 identification by MiTCR and TCRklass does not affect much on the CDR3 abundance profile.

The V/J profiles generated by TCRklass and MiTCR are analyzed by HCL using SRD (Fig. 4E, Supplemental Fig. 3D–F). It can be observed that the profiles are clustered together according to the analysis methods rather than the samples (human or mouse). Specifically, the V/J profiles generated from $CDR3^+$ reads are more closely clustered with the MiTCR V/J profiles. This indicates that the difference between the V/J profiles by TCRklass and MiTCR is significant, and the reason is possibly that MiTCR only use these CDR3-positive reads.

Discussion

The next-generation sequencing technology is a powerful tool for the characterization of highly diverse TCR repertoires that will greatly impact the clinical diagnostics and immunogenomics studies. However, the identification and annotation of TCR clonotypes from the vast amount of sequencing reads is still challenging, especially when short paired-end reads generated by HiSeq2000 are the main source of data due to its low cost and high throughput. The TCRklass, as a new pipeline for TCR repertoire analysis, improved significantly the accuracy for both CDR3 and V/J identification over existing tools.

The k-string matching-based algorithm is widely used in many major software programs such as RDPclassifier (25), NCBI BLAST (26), and others like MetaCV for metagenomic data analysis (27). The matching K-strings can locate conserved sequences, whereas the matching numbers can measure the sequence similarity at the same time. For TCRklass, the CDR3 boundary is defined by multiple matched K-strings of the reference V or J sequences that eliminate the occasional mismatched k-strings, ensuring the uniqueness of CDR3 identified. It shows great advantage over the sequence alignment approach that is very time consuming, and the identified CDR3-containing reads are more accurate than those by MiTCR. Potentially, the TCRklass can also be used for other similar recombined sequences such as immunoglobin (BCR repertoire) analysis.

The IMGT/HighV-QUEST is based on long alignment with V gene segment that limits its use on short reads analysis. The MiTCR is designed to analyze short reads, which is based on read mapping and is focused on CDR3-containing reads. However, its strategy of only processing CDR3-containing reads causes bias on the composition of V and J gene segments. The TCRklass does not require long query sequences, which improved the recoverability of CDR3-containing sequences. And TCRklass identifies CDR3 sequence and V/J gene segments separately, which improved the accuracy of V and J identification.

The abundance profiling of V/J pairing is an important issue in TCR repertoire characterization. The limited length of sequencing reads has generated a large amount of paired-end reads without CDR3 sequence. The TCRklass can take these reads to analyze V/J pairing, and our comparative study showed that the V/J profiles generated using total sequences and using CDR3-containing sequences are different. Actually, the $CDR3^-$ reads do have the CDR3 sequences, but they lay on the gap between two paired-ends that cannot be read by the sequencer. These $CDR3^-$ reads are informative for V/J profiling and should not be discarded. Therefore, the use of both CDR3-containing and non-CDR3–containing reads by TCRklass would be helpful for providing unbiased profile on V/J pairing.

The TCR clonotypes are defined as a combination of CDR3 sequence and V/J gene segments, in which only those $CDR3^+$ reads can be used. The dataset of short sequencing reads of TCR repertoire generated by next-generation sequencing technology can be divided into two parts, one part consists of those reads that contain complete CDR3 sequence ($CDR3^+$), and the other part consists of reads without CDR3 sequence ($CDR3^-$). For MiTCR, it uses only the $CDR3^+$ reads in TCR clonotype profiling. As we have shown that the V/J profiles generated using $CDR3^+$ sequences are biased, the corresponding TCR clonotype profiles by MiTCR are inaccurate because the abundance of those $CDR3^-$ reads is missing. In contrast, TCRklass use both $CDR3^+$ and $CDR3^-$ reads for V/J profiling; thus, it is more accurate. In summary, more TCR clonotypes can be identified using TCRklass; however, the exact number of these additional TCR clonotypes still cannot be defined because of the lack of CDR3 sequence.

The variations of TCR sequences are induced in a few steps that include somatic mutation, mRNA synthesis, PCR amplification, and sequencing. However, it is difficult to identify the error rate in each step. Thus, we assessed a generic per-nucleotide error rate that describes the total variations from genomic DNA sequence to observed sequencing reads, and the error rate assessed by two different ways have similar values that cross-validate each other. The estimated CDR3 error rate is significant ($>15\%$), but we do not think they are able to be excluded from the dataset. An arbitrary and crude method could be excluding CDR3 sequences from the lowest abundant ones. However, because the lowest abundant CDR3 sequences take a significant portion (Supplemental Fig. 6), this approach has the risk for excluding reliable CDR3 sequences or including unreliable ones. For example, the portion of singleton CDR3 sequences for mouse β-chain datasets is $>30\%$, which is much higher than the error rate. Therefore, the study on low abundant CDR3 polymorphism requires better sequencing tools.

In conclusion, TCRklass is a novel toolkit that can identify CDR3 and V/J gene segments from the high-throughput sequencing reads of TCR repertoire. We tested TCRklass using multiple human and mouse TCR sequencing datasets, and demonstrated that TCRklass have higher accuracy than previous tools. The TCRklass is implemented using C++ and Perl; the source code, indexed reference databases, and benchmark data can be freely downloaded at: http://sourceforge.net/projects/tcrklass/. We claim that the TCRklass toolkit will be maintained for at least 10 years.

Disclosures

The authors have no financial conflicts of interest.

References

1. Turner, S. J., P. C. Doherty, J. McCluskey, and J. Rossjohn. 2006. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* 6: 883–894.
2. Arstila, T. P., A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky. 1999. A direct estimate of the human alphabeta T cell receptor diversity. *Science* 286: 958–961.
3. Robins, H. S., P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114: 4099–4107.
4. Warren, R. L., J. D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb, and R. A. Holt. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21: 790–797.
5. Wang, C., C. M. Sanders, Q. Yang, H. W. Schroeder, Jr., E. Wang, F. Babrzadeh, B. Gharizadeh, R. M. Myers, J. R. Hudson, Jr., R. W. Davis, and J. Han. 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl. Acad. Sci. USA* 107: 1518–1523.
6. Genolet, R., B. J. Stevenson, L. Farinelli, M. Osterås, and I. F. Luescher. 2012. Highly diverse TCRα chain repertoire of pre-immune $CD8^+$ T cells reveals new insights in gene recombination. *EMBO J.* 31: 1666–1678.
7. Mueller, C., J. D. Chulay, B. C. Trapnell, M. Humphries, B. Carey, R. A. Sandhaus, N. G. McElvaney, L. Messina, Q. Tang, F. N. Rouhani, et al. 2013. Human Treg responses allow sustained recombinant adeno-associated virus-mediated transgene expression. *J. Clin. Invest.* 123: 5310–5318.
8. Bolotin, D. A., I. Z. Mamedov, O. V. Britanova, I. V. Zvyagin, D. Shagin, S. V. Ustyugova, M. A. Turchaninova, S. Lukyanov, Y. B. Lebedev, and D. M. Chudakov. 2012. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* 42: 3073–3083.
9. Warren, R. L., B. H. Nelson, and R. A. Holt. 2009. Profiling model T-cell metagenomes with short reads. *Bioinformatics* 25: 458–464.
10. Freeman, J. D., R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* 19: 1817–1824.
11. Dziubianau, M., J. Hecht, L. Kuchenbecker, A. Sattler, U. Stervbo, C. Rödelsperger, P. Nickel, A. U. Neumann, P. N. Robinson, S. Mundlos, et al. 2013. TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. *Am. J. Transplant.* 13: 2842–2854.
12. Saito, T., and R. N. Germain. 1988. The generation and selection of the T cell repertoire: insights from studies of the molecular basis of T cell recognition. *Immunol. Rev.* 101: 81–113.
13. Danska, J. S., A. M. Livingstone, V. Paragas, T. Ishihara, and C. G. Fathman. 1990. The presumptive CDR3 regions of both T cell receptor alpha and beta chains determine T cell specificity for myoglobin peptides. *J. Exp. Med.* 172: 27–33.
14. Arden, B., S. P. Clark, D. Kabelitz, and T. W. Mak. 1995. Human T-cell receptor variable gene segment families. *Immunogenetics* 42: 455–500.
15. Arden, B., S. P. Clark, D. Kabelitz, and T. W. Mak. 1995. Mouse T-cell receptor variable gene segment families. *Immunogenetics* 42: 501–530.
16. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol. Biol.* 882: 569–604.
17. Lefranc, M. P., V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 37: D1006–D1012.
18. Li, S., M. P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D. Freeman, V. D. Corbin, J. P. Scheerlinck, M. A. Frohman, et al. 2013. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.* 4: 2333.
19. Alamyar, E., V. Giudicelli, S. Li, P. Duroux, and M. P. Lefranc. 2012. IMGT/HighV-QUEST: the IMGT®web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.* 8: 1–2.
20. Brochet, X., M. P. Lefranc, and V. Giudicelli. 2008. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36: W503–W508.
21. Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R. W. Shafer. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.* 17: 1195–1201.
22. Thomas, N., J. Heather, W. Ndifon, J. Shawe-Taylor, and B. Chain. 2013. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* 29: 542–550.
23. Bolotin, D. A., M. Shugay, I. Z. Mamedov, E. V. Putintseva, M. A. Turchaninova, I. V. Zvyagin, O. V. Britanova, and D. M. Chudakov. 2013. MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* 10: 813–814.
24. Yang, X., D. Liu, F. Liu, J. Wu, J. Zou, X. Xiao, F. Zhao, and B. Zhu. 2013. HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* 14: 33.
25. Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73: 5261–5267.
26. Ye, J., S. McGinnis, and T. L. Madden. 2006. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 34: W6–W9.
27. Liu, J., H. Wang, H. Yang, Y. Zhang, J. Wang, F. Zhao, and J. Qi. 2013. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res.* 41: e3.