IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis

Wei Zhang,*,†,1 Yuanping Du,*,†,1 Zheng Su,*,†,1 Changxi Wang,* Xiaojing Zeng,* Ruifang Zhang,* Xueyu Hong,* Chao Nie,* Jinghua Wu,* Hongzhi Cao,* Xun Xu,* and Xiao Liu*,†,2

*BGI-Shenzhen and †Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen 518083, China, and ‡Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark

ABSTRACT The advance of next generation sequencing (NGS) techniques provides an unprecedented opportunity to probe the enormous diversity of the immune repertoire by deep sequencing T-cell receptors (TCRs) and B-cell receptors (BCRs). However, an efficient and accurate analytical tool is still on demand to process the huge amount of data. We have developed a high-resolution analytical pipeline, Immune Monitor ("IMonitor") to tackle this task. This method utilizes realignment to identify V(D)J genes and alleles after common local alignment. We compare IMonitor with other published tools by simulated and public rearranged sequences, and it demonstrates its superior performance in most aspects. Together with this, a methodology is developed to correct the PCR and sequencing errors and to minimize the PCR bias among various rearranged sequences with different V and J gene families. IMonitor provides general adaptation for sequences from all receptor chains of different species and outputs useful statistics and visualizations. In the final part of this article, we demonstrate its application on minimal residual disease detection in patients with B-cell acute lymphoblastic leukemia. In summary, this package would be of widespread usage for immune repertoire analysis.

KEYWORDS next generation sequencing; bioinformatics; immune repertoire; TCR/BCR

THE diversity of T-cell receptors (TCRs), B-cell receptors (BCRs), and secreting form antibodies makes up the core of the complicated immune system and serves as pivotal defensive components to protect the body against invading virus, bacteria, and other pathogens. The TCR consists of a heterodimeric αβ chain (~95%, TRA, TRB) or γδ chain (~5%), while the BCR is assembled with two heavy chains (IGH) and two light chains (IGK or IGL). Structurally, each chain can be divided into the variable domain and the constant domain (Lefranc and Lefranc 2001a,b). The diversity of the TCR and BCR repertoire is enormous, owing to the process of V(D)J gene rearrangement, random deletion of germline nucleotides, and insertion of uncertain length of nontemplate nucleotides between V-D and D-J junctions (TRB, IGH) or V-J junctions (TRA, IGK, IGL). In humans, it has been estimated theoretically that the diversity of TCR-αβ receptors exceeds 10^{18} in the thymus, and the diversity of the B-cell repertoire is even larger, considering the somatic hypermutation (Janeway 2005; Benichou *et al.* 2012). The T- and B-cell repertoire could undergo dynamic changes under different phenotypic status. Recently, deep sequencing enabled by different platforms including Roche 454 and Illumina Hiseq (Freeman *et al.* 2009; Robins *et al.* 2009; Wang *et al.* 2010; Fischer 2011; Venturi *et al.* 2011) has been applied to unravel the dynamics of the TCR and BCR repertoire and extended to various translational applications such as vaccination, cancer, and autoimmune diseases.

Several tools and software have been developed for TCR and BCR sequence analysis, including iHMMune-align (Gaeta *et al.* 2007), HighV-QEUST (Li *et al.* 2013), IgBLAST (Ye *et al.* 2013), Decombinator (Thomas *et al.* 2013), and MiTCR (Bolotin *et al.* 2013). These tools are equipped with useful functions, including V(D)J gene alignment, CDR3 sequence identification, and more, yet with obvious limitations. For instance, HighV-QEUST can be adopted to analyze both TCRs and BCRs, but its online version limits maximum sequence input to 150,000 at a time for regular users. Decombinator and MiTCR can only be used to analyze the TCR sequences.

doi: 10.1534/genetics.115.176735
Manuscript received March 25, 2015; accepted for publication August 16, 2015; published Early Online August 21, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176735/-/DC1.
[1]These authors contributed equally to this work.
[2]Corresponding author: Main Bldg., Beishan Industrial Zone, Yantian, Shenzhen 518083, China. E-mail: liuxiao@genomics.cn

Besides, most tools lack specific solutions to some common problems like systemic statistics and visualizations, PCR and sequencing errors, and amplification bias correction.

Here, we introduce a novel pipeline, Immune Monitor (IMonitor) for both TCR and BCR deep sequencing analysis. It includes four steps in its core component: basic data processing, V(D)J assignment, structural analysis, and statistics/visualization. One feature that makes IMonitor stand out is its realignment process to identify V(D)J genes and alleles with significantly enhanced precision. We simulated 15 data sets for five chains (TRA, TRB, IGH, IGK, IGL) of different sequencing error rates and hypermutation rates, together with actual rearranged sequences, to test performance of various tools. IMonitor performs quite well in accuracy and clonotype recovery. Furthermore, IMonitor incorporates a process to correct PCR and sequencing errors, utilizing the data from six plasmid mixed samples, and an *in silico* model was modulated to reduce the PCR bias. Finally, we validate IMonitor in detection of minimal residual disease (MRD) of B-cell acute lymphoblastic leukemia (B-ALL) to show its wide utility potential.

Materials and Methods

The core component of IMonitor consists of four steps: basic data processing, V(D)J assignment, structural analysis, and statistics/visualization, as shown in Figure 1. IMonitor can utilize data generated by a variety of next generation sequencing (NGS) platforms, such as Illumina, Roche 454, and Life Ion Proton, in both FASTQ and FASTA format. The final results of IMonitor include a complete map of sequences and data analysis in depth, and the latter is visualized and presented with viewer-friendly graphs and figures.

IMonitor for basic data processing

In the first step, the reads were checked for inclusion of adaptor sequences. If any adaptor sequence was detected and located within 50 bp of the 3′end of the read, it was deleted from the read. Reads bearing adaptor sequence at the 5′ end or >5% "N" bases were discarded. The average base quality of each read was calculated after removing the low-quality bases (base quality <10) at the 3′ end. Further filtration left out reads with average quality <15. For Illumina paired-end (PE) sequencing, the PE reads were merged at their overlapping region. For PE reads with insertion length longer than the length of a single read, the COPE (Liu *et al.* 2012) tool was used; otherwise reads were assembled by an in-house program. The main parameters for both tools included the maximum overlapping length (read length), minimum overlapping length (10 bp), mismatch rate (10%) at the overlapping region, and ratio (best overlap length/second-best overlap length, 0.7).

IMonitor for V(D)J assignment

The V/D/J reference sequences were downloaded from the IMGT database, the international ImMunoGeneTics information system (http://www.imgt.org/). Processed sequences were aligned to the V, (D), J references, respectively, by BLAST (Altschul *et al.* 1990; Zhang *et al.* 2000; Ye *et al.* 2006) and specific parameters were applied to accommodate the differences in lengths of V, (D), J segments (BLAST parameters: V, -W 15 -K 3 -v 1 -b 3; D, -W 4 -K 3 -v 3 -b 5; and J, -W 10 -K 3 -v 1 -b 3).

The high similarity among the genes and alleles of the germline sequences, along with the diversity of V/D/J gene rearrangement, gave rise to difficulties for accurate alignment. This might eventually lead to an incorrect structural analysis (CDR3 identification, deletion, or insertion). To improve the accuracy, a second alignment procedure was developed to identify exactly the V/D/J genes (Figure 2). First, a global alignment strategy, which attempted to align every base in every sequence, was used for the non-CDR3 region of the sequence. The mapped region generated from BLAST became a new seed and served as starting points for bootstrapping (base-by-base) extension to both directions, until the entire non-CDR3 region in the query was mapped to the target (reference) sequence. The mapping score was calculated according to these rules: reward for a nucleotide match was 5 and penalty for a nucleotide mismatch was −4. Second, the M-mismatch extension model of local alignment strategy was applied to locate the exact end positions of V and J genes during CDR3 region realignment. The procedure began at the CDR3 start position in the V gene or the CDR3 end position in the J gene and continuously extended in one direction until the preset mismatch limit was reached, generating the longest possible interval with the highest score. The mismatch numbers allowed for V/D/J genes were determined based on the analysis result of publicly available rearrangement sequences (http://www.imgt.org/ligmdb/) (Supporting Information, Figure S2A) and adjusted accordingly for different TCR and BCR chains (mismatches allowed: TRBV/J, TRAV/J, 0; IGHV/J, 2; IGKV/J, IGLV/J, 7). As shown in Figure S2A, these mismatch limits took mutations into consideration and covered >99.5% of all defined rearrangement sequences. Because the entire D gene was located within the CDR3 region, only the M-mismatch extension model was used for its realignment (mismatches allowed: TRBD, 0; IGHD, 4). Finally, all data including alignment score, identity, mismatch number, and alignment length were processed, and the alignment with highest score and identity larger than the threshold (>80%) was selected as the best hit. However, there might be several best hits with the same score due to the homology among the germline genes and alleles. In this case, the reference with the fewest deletions was selected, as shorter deletions are more likely to happen according to previous reported results (Warren *et al.* 2009) and our analysis from actual public rearrangement data (Figure S2B).

IMonitor for structural analysis

The IMGT collaboration (Yousfi Monod *et al.* 2004) outlined the CDR3 region of all chains, starting from the second conserved cysteine encoded by the V segment and ending with

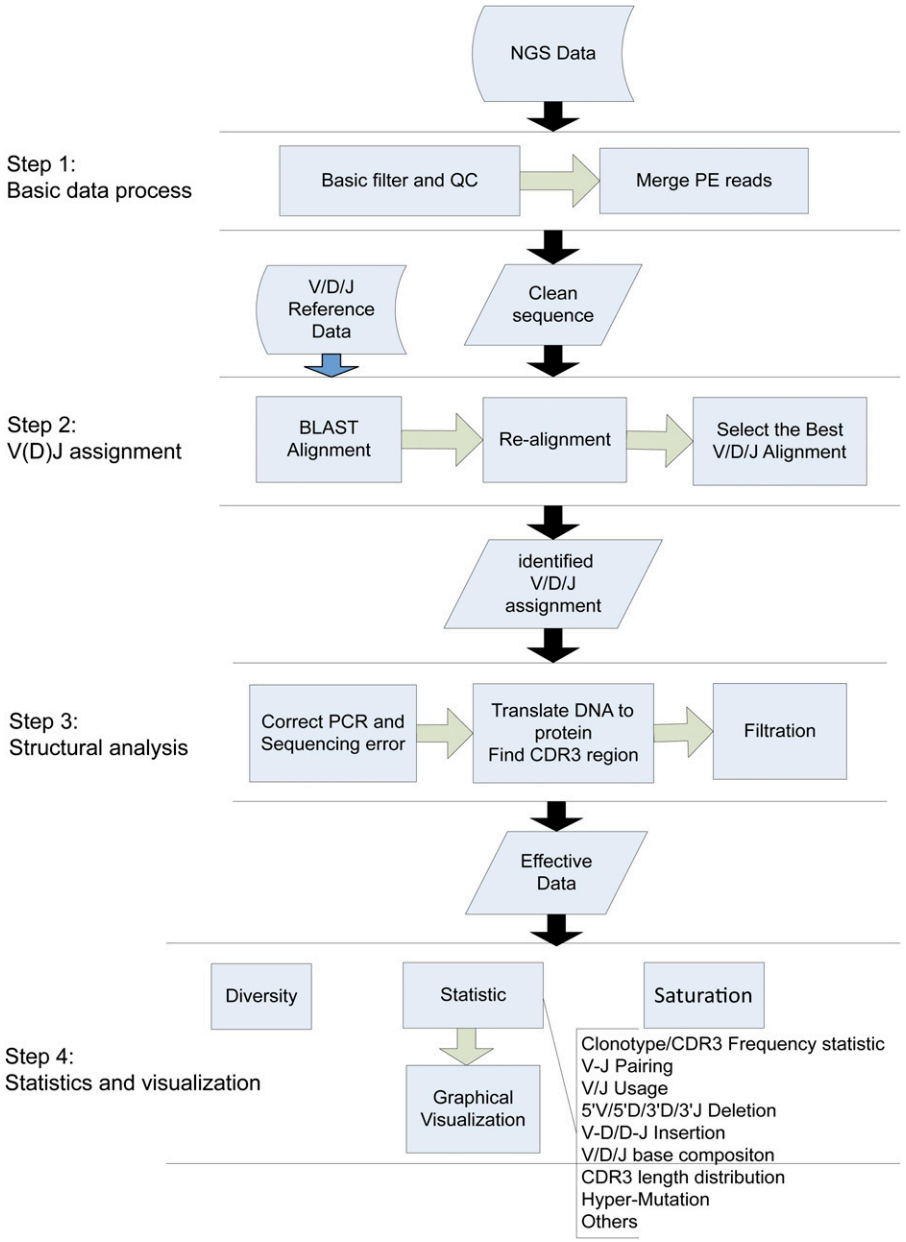

Figure 1 Overview of workflow of IMonitor. Although the program includes four steps, we have several parameters to control whether the module runs or not. The program takes raw NGS (FASTA or FASTQ) as input and outputs the V(D)J assignment of the sequence, some statistics, and corresponding figures.

the conserved phenylalanine or tryptophane encoded by the J segment. Combining this information with our selected reference from the previous step, the CDR3 region of target sequence could be readily identified. For unmapped sequences, the CDR3 region was determined by searching through for a conservative amino acids module within both ends of the CDR3 region ("YXC" for start and "[FW]GXG" for end, where "X" stands for any amino acid). The rearrangement frame was tagged as "in-frame" if the length of CDR3 was a multiple of three and no stop codon was found in whole sequence; otherwise it was tagged "out-of-frame." The structure of the sequence was clearly described, including V, (D), J segments used, the CDR3 region, and the deletions and insertions at rearrangement sites. Then the nucleotide sequences were translated into peptides. However, some sequences must be filtered out to ensure the accuracy of the immune repertoire, which include (1) sequences without CDR3 region and (2) sequences with V and J alignment orientation conflict. The sequences that were aligned to pseudogenes, were out-of-frame, and included a stop codon were marked.

IMonitor for statistics and visualization

The basic statistics of IMonitor include CDR3 frequency distribution, V-J paring, V/J usage, 5′V/5′D/3′D/5′J deletion length distribution, V-D/D-J insertion length distribution, V/J base composition, CDR3 length distribution, CDR3 segmental frequency statistics, Top10 CDR3 frequency, hypermutation of BCRs, etc. Figures were plotted to visually demonstrate each result. For V-J pairing, a three-dimensional

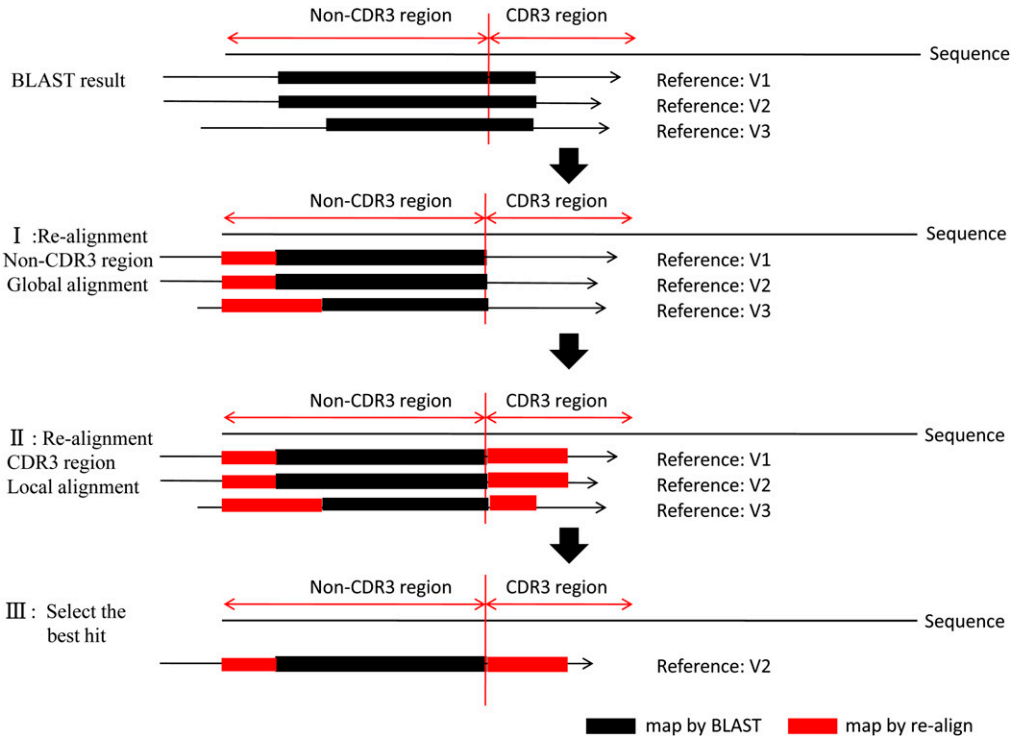

Figure 2 The workflow of realignment. The program takes the BLAST alignment results as input, realigns the sequence to reference for both the non-CDR3 region and the CDR3 region, calculating the score and identity, and then selects the maximal score as the best hit. The reference with shortest length of deletion is preferred if it finds multiple references with the same maximal scores. It outputs the optimal alignment result.

figure was generated. R script was used to draw most of the figures while the V/J base composition was plotted with weblogo 2.8 (Crooks *et al.* 2004). To completely eliminate sequencing error effect, a sequence detected <10 times was excluded to calculate hypermutation. The mutation rate consisted of base mutation and sequence mutation. The former was the content of mutational bases in total bases, and the latter was rate of the sequence containing mutation in total sequences.

The Shannon–Weiner index (Shannon 1997), as shown below, used as an immunological diversity value in several previous works (Sherwood *et al.* 2013), provided a good estimate of diversity in a large-scale study and was suitable for immune repertoire. We used it to calculate the diversity of CDR3, the V gene, the J gene, and V-J pairing,

$$H' = -\sum_{i=1}^{S} p(i)\ln p(i), \tag{1}$$

where CDR3 is an example, S denotes the total number of unique CDR3, and $p(i)$ denotes the frequency of CDR3.

IMonitor also provided the saturation analysis with the Chao1 algorithm, which was used to estimate the target richness for individual-based data in previous ecological studies. Estimated values generated by the Chao1 bias-corrected algorithm were used to predict the maximum number of clones in the sample, while observed values were drawn separately with the rarefaction curves (Chao 1984, 1987):

$$\hat{S}_{\text{Chao1}} = S_{\text{obs}} + \frac{F_1(F_1-1)}{2(F_2+1)}. \tag{2}$$

In Equation 2, S_{obs} stands for total number of observed clonotypes in a sample; F denotes the number of clonotypes (F_1, the number of clonotypes detected one time; F_2, the number of clonotypes detected two times).

PCR and sequencing error correction

PCR and sequencing error of NGS is one of the toughest problems in immune repertoire analysis. The method we developed to correct this error could be utilized on either the whole sequence or just the CDR3 region. The procedure consisted of three steps. First, sequences were divided into three groups: (1) high-quality sequences whose base qualities all were >Q20 (Q20 were the best cutoff according to Figure S5, A and B; (2) sequences with more than five (three for only CDR3 region correction) low-quality bases were unwanted and discarded; and (3) the rest were defined as low-quality sequences. Second, the low-quality sequences were mapped to the high-quality ones. When the mismatches were no more than five (three for only CDR3 region correction) and all located at low-quality positions, the mismatches were corrected; otherwise the sequence was discarded. Finally, to eliminate PCR errors, sequences with low abundance were compared to ones with high abundance (at least fivefold difference). If fewer than three mismatches were found in the low-abundance sequences, they were corrected to the corresponding high-abundance sequences. To test the effectiveness of this method, samples made from mixtures of six plasmids were used for error characteristics analysis and further evaluation.

Multiplex PCR bias minimization

We established a new bioinformatic approach to minimize PCR amplification bias. The approach is built on the hypothesis that there are two factors affecting a clone's frequency during multiplex PCR (MPCR): the template's concentration and the multiple primers' efficiency. Using six plasmid mixture samples (Table S4), we could compare the observed with the expected frequency and simulate an effective formula.

The streamlined procedure of this method is demonstrated by the flowchart in Figure 6. Two factors, templates' concentration and primers' efficiency, were considered to be affecting the bias and examined here. The samples were mixed properly, and clones were grouped to explore PCR bias rules. First, we analyzed the bias' correlation with templates' concentration (concentration analysis without primer effect). Each clone had three different concentrations in all samples. To eliminate potential effects caused by the multiple primers efficiencies, the clones that had the same concentration ratio among all samples were grouped together, generating five groups in total [groups were named 10_2E4_1E5, 1000_2E4, 100_1000_2E4, 100_1E4_2E4, and 10_1E4_2E4; for instance, 10_2E4_1E5 denoted the three concentrations (10, 2E4, 1E5) in respective samples]. Then, within each group, the clone frequencies were normalized by multiplying the same coefficient k, which could generate a minimal sum of absolute deviation $D_{\min}(i)$. The 10_2E4_1E5 group sets an example as follows:

$$\mu(j) = \frac{1}{n}\sum_{i=1}^{n} f(i,j), \quad j = \{10, 2E4, 1E5\} \tag{3}$$

$$\begin{aligned} D_{\min}(i) = \min\{&|f(i,10)\,k - \mu(10)| \\ &+ |f(i,2E4)\,k - \mu(2E4)| \\ &+ |f(i,1E5)\,k - \mu(1E5)|\}, \quad k \in (0,+\infty) \end{aligned} \tag{4}$$

$$f_{\text{norm}}(i,j) = f(i,j)\,k, \quad j = \{10, 2E4, 1E5\}, \tag{5}$$

where n is clone number in a group, i is a clone, and f is clone frequency. k was set consecutively from 0 and a series of $D(i)$ was calculated, after which the k that generated the smallest $D(i)$ was selected.

After normalization, five groups were combined on the basis of 2E4 copies, which existed in all groups. We used a regression module to fit a curve (Equation 6) that reflected the relationship between concentration and PCR bias,

$$y = 0.60636\,\mathbf{log}_{1.8}^{x+1}, \tag{6}$$

where y is the observed sequence's frequency and x is the expected sequence's frequency.

Second, we analyzed the bias caused only by primer efficiencies (primer analysis without concentration effect). To remove the effect of clone concentration, clones with the same concentration in all samples were collected into one group; thus six groups were generated (10, 100, 1E3, 1E4, 2E4, and 1E5, where group 10 contained all clones that had the concentration 10 in any of the samples) (Figure 6). After calculation, each group was normalized by multiplying the same coefficient l, with the following details,

$$\begin{aligned} r(i,j) = &\begin{cases} \dfrac{f(i,j)\,l}{f(i,2E4)}, & \text{if } \{f(i,j)\,l > f(i,2E4)\} \\ \dfrac{f(i,2E4)}{f(i,j)\,l}, & \text{if } \{f(i,2E4) > f(i,j)\,l\}, \end{cases} \\ &l \in (0,+\infty), \quad j = \{10, 100, 1E3, 1E4, 2E4, 1E5\} \end{aligned} \tag{7}$$

$$R_{\min}(i,j) = \min\left\{\sum_{i=1}^{n} r(i,j)\right\} \tag{8}$$

$$f_{\text{norm}}(i,j) = f(i,j)\,l, \tag{9}$$

where, n is clone number in a group, i is a clone, and f is clone frequency. l was set consecutively from 0 and a series of $R(i)$ was calculated, after which the l that generated the minimal $R(i)$ was selected.

Each primer's efficiency was calculated after normalization. Then, analysis of the two factors was integrated into a formula that minimized the PCR bias (Equation 10),

$$\begin{aligned} f_{\text{correct}} = 1.8^{S_i/(S^*0.60636^*p)} - 1, \quad p = 0.5\,p(v) \\ + 0.5\,p(j) + 0.05, \end{aligned} \tag{10}$$

where f_{correct} is the corrected frequency, S_i is the clone's observed abundance, S is the sum abundance of the sample, $p(v)$ is the primer efficiency value for the V gene, and $p(j)$ is the primer efficiency value for the J gene.

Multiplex PCR amplification

To amplify rearranged CDR3 regions, multiple forward primers in the V region and reverse primers in the J region were designed. For the RNA sample, the first-strand cDNAs were synthesized using SuperScript II Enzyme according to the manufacturer's instructions. Then two individual equimolar pools of the forward primers and the reverse primers were used for a multiplex PCR (QIAGEN, Valencia, CA) of 30 cycles according to the provided protocol. The fractions between 110 and 180 bp of the PCR products were excised and purified.

Simulation of in silico sequences

A total of 10^5 sequences were generated *in silico* for each data set with a length of 200–300 bp by simulating the relevant biological processes that occur during B-cell and T-cell development. The sequences of V(D)J genes of TRA/TRB/IGH/IGK/IGL were downloaded from IMGT (http://www.imgt.org/). First, to simulate recombination, a V allele and a J allele (an extra D allele for TRB and IGH) were selected at random to generate a V-D-J (V-J for TRA and light chain) combination. Second, to simulate deletion and

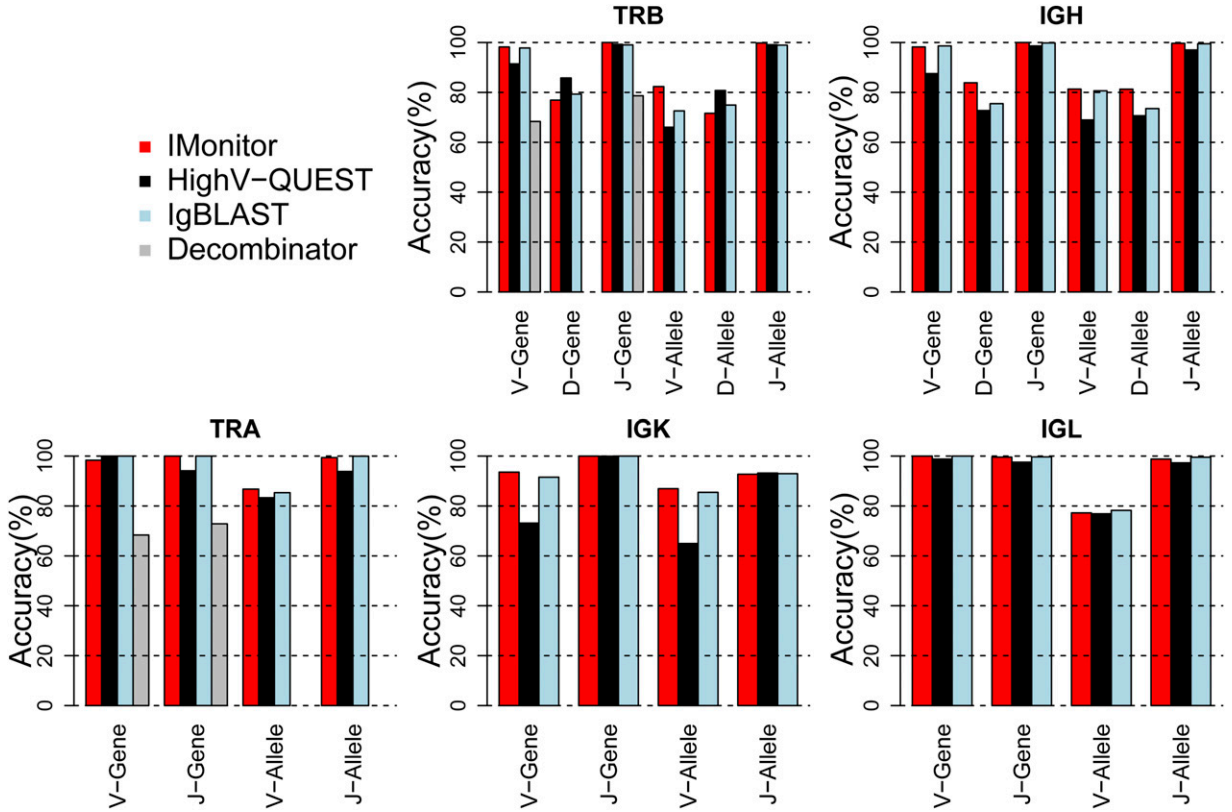

Figure 3 Performance of different types of software on 10^5 *in silico* sequences. Shown is accuracy of TRA, TRB, IGH, IGK, and IGL sequences with 0.5% sequencing error. The accuracy is calculated for both gene and allele, but Decombinator has only V and J genes identified.

insertion, some bases were deleted at the 3′ end of V and the 5′ end of J (both ends of D for TRB and IGH) according to the deletion length distribution from previous research (Figure S1); meanwhile some random bases were inserted at V-D and D-J junctions (V-J junction for TRA and light chain), using the established insertion length distribution (Figure S1). Third, both somatic hypermutation during BCR maturation and sequencing error were also taken into consideration. Although the typical error rate of Illumina HTS technology is ∼1%, it can be reduced after merging paired-end reads (Figure 5A). Therefore, for each chain, the error rate was set to 0.1%, 0.5%, and 2% evenly at every position of the sequence. The hypermutation rate was set to 1% for IGH and 4% for IGK/IGL.

Public rearrangement data sets

The data sets were obtained from the IMGT/LIGM-DB database(http://www.imgt.org/ligmdb/); searched by "Homo sapiens," "rearranged," "TRB," or "IGH"; and then the selected sequences were annotated manually (Annot. level==="manual") and annotated by V, D, J genes. Finally, 24 TRB and 1763 IGH sequences met these requirements. The sequence was in FASTA format, without sequencing quality. The length of sequences downstream of the J gene was limited to 50 bp.

Samples

Plasmid mix samples: Thirty-three different functional TCR β-chain sequences, which included all the TCR β, V, and J genes, were integrated into plasmid vectors. Three mixing pattern pools were used: one with an equal mole number of each plasmid and the other two pools with different pooling gradients. The mix patterns are listed in Table S4, and each pattern was replicated to produce six plasmid mix samples.

Spiked-in DNA samples: Spiked-in samples were generously donated by Karen Cerosaletti (Benaroya Research Institute, Seattle, WA) (Robins *et al.* 2012). Five $CD4^+$ T-cell clones were spiked in a background of sorted $CD4^+CD45RA^+$ naive T cells and each unit had 1 million cells. The five clones in three different units had different numbers of unique TCRβ CDR3 sequences and are shown in Table S7. The background cells for these doped samples were sorted from a fresh peripheral blood mononuclear cell (PBMC) sample obtained from a control donor with informed consent. DNA was extracted from the cells with a commercial kit.

Healthy donor samples: Samples of peripheral blood from two healthy human donors (H-H-1 and H-B-1) were obtained by venipuncture with informed consent. PBMCs were isolated immediately and RNA was extracted using Trizol reagent (Invitrogen, Carlsbad, CA). DNA was extracted with a QIAamp DNA Blood Mini Kit and stored at −20°. The CDR3 region was amplified by multiplex PCR (Table S9) and sequenced by an Illumina platform (Table S6).

Table 1 Number of sequences with correctly identified V, D, and J genes or alleles in public sequence data sets

Datasets/Tools	V_gene	V_allele	D_gene	D_allele	J_gene	J_allele
Data set of TRB (24 sequences)						
IMonitor	24 (100%)	23 (96%)	17 (71%)	12 (50%)	24 (100%)	24 (100%)
IgBLAST	23 (96%)	22 (92%)	13 (54%)	11 (46%)	24 (100%)	24 (100%)
Decombinator[a]	21 (87%)	—	—	—	23 (96%)	—
Data set of IGH (1763 sequences)						
IMonitor	1735 (98%)	1509 (86%)	1037 (59%)	952 (54%)	1619 (92%)	1533 (87%)
IgBLAST	1716 (97%)	1518 (86%)	986 (56%)	956 (54%)	1563 (89%)	1498 (85%)

The data sets were obtained from the IMGT/LIGM-DB database (http://www.imgt.org/ligmdb/); searched by "Homo sapiens," "rearranged," "TRB," or "IGH"; and then the selected sequences were annotated manually (Annot. level=="manual") and annotated by V, D, J genes. So these sequences have a fairly high level of annotation confidence. The data sets and HighV-QEUST came from the same website, so HighV-QEUST was not used here. The references used for tools were the same and were from the IMGT database (http://www.imgt.org).

[a] Decombinator analyzed just the gene level of V and J.

MRD samples: Bone marrow samples from two patients (M001 and M002) with B-ALL were provided with informed consent. The samples of pretreatment, day 15, and day 33 post-treatment were assayed. A total of 1.2 μg DNA of each sample was used for multiplex PCR amplification (Table S9). The library of ~150–270 bp insert-size length was extracted and sequenced using the 2×100 PE Illumina platform (Table S6).

The research was prospectively reviewed and approved by a duly constituted ethics committee.

Data availability

The source code of IMonitor is freely available for download at https://github.com/zhangwei2015.

Table S1: Simulated TRB with 0.1%, 0.5% and 2% sequencing error.
Table S2: Simulated IGH with 0.1%, 0.5% and 2% sequencing error and 0.1% hayper-mutation.
Table S3: Simulated Data with 0.5% sequencing error (TRA/IGK/IGL) and 4% hyper-mutation.
Table S4: Plasmid mixing pattern.
Table S5: Data process for PCR and sequencing error statistics.
Table S6: Samples information.
Table S7: Experimental design for five CD4+ T cell clones in the three spiked in mix.
Table S8: Performance of IMonitor and other tools on the simulated dataset.
Table S9: TRB and IGH V/J primers.

Figure S1: Insertion and deletion length distribution for simulated data.
Figure S2: IGH-VDJ Mutation and deletion/insertion analysis on the public sequences.
Figure S3: Outputs of IMonitor, H-B-01 as an example.
Figure S4: H-B-01 sample output figure of IMonitor.
Figure S5: Error characteristics of 6 plasmid mix samples.
Figure S6: V-J pairing dynamics for M002.
Figure S7: MiTCR and IMonitor performance in 3 spiked-in samples.
Figure S8: Nucleotide composition of V/J genes.

Results

System design of IMonitor

Four steps are described in Figure 1:

1. Basic data process: Sequence containing adapter sequence was processed and low-quality bases at the 3′ end of the sequence were removed. PE reads were merged to one sequence by an in-house program and COPE (Liu *et al.* 2012).
2. V(D)J assignment: The reference germline sequences were downloaded directly from IMGT (http://www.imgt.org/). Processed data were aligned to the references by BLAST (Altschul *et al.* 1990; Zhang *et al.* 2000; Ye *et al.* 2006) and realigned to improve the map accuracy, after which the optimal alignment was selected for every sequence (Figure 2).
3. Structural analysis: A novel method was established to correct PCR and sequencing errors. The CDR3 region was identified with the help of both V/J references and conservative amino acids and then translated into amino acids.
4. Statistics and visualization. Characteristic data of the immune repertoire of the samples, such as repertoire diversity, clonotype frequency, CDR3 length distribution, V/J usage, V-J pairing, hypermutation, deletion, and insertion, were collected (Figure S3) and presented with corresponding graphs. More specifically, the V-J pairing was visualized by a three-dimensional graph (Figure 4, Figure S4, and Figure S8).

IMonitor outperforms other analytical tools in various aspects

To evaluate the performance of IMonitor, we designed a head-to-head comparison between IMonitor and other publicly available tools with both simulated data and public rearrangement sequences. The TRA/TRB data were analyzed by HighV-QEUST (Li *et al.* 2013), Decombinator (Thomas *et al.* 2013), IgBLAST (Ye *et al.* 2013), and IMonitor, while the IGH/IGK/IGL data were analyzed by HighV-QEUST, IgBLAST (Ye *et al.* 2013), and IMonitor. Thomas *et al.* (2013) reported that

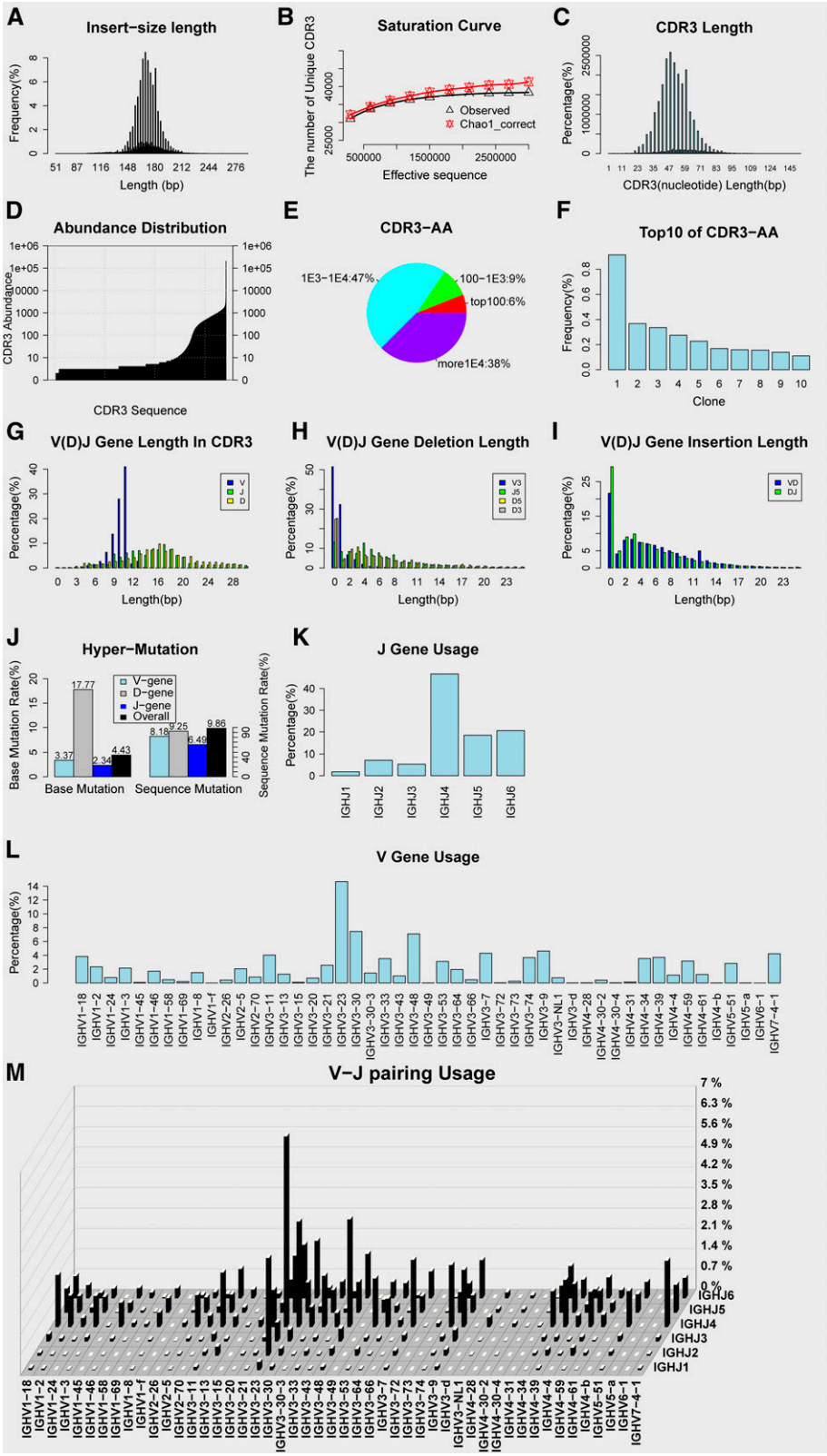

Figure 4 Output figure of IMonitor. H-H-01 sample is shown as an example. (A) Sequence length distribution. (B) Saturation curve, rarefaction studies of sequences. Subsequences are randomly selected and the observed unique CDR3 number and predicted CDR3 number (Chao1-corrected algorithm) are calculated. (C) CDR3 nucleotide length distribution. (D) CDR3 abundance distribution. (E) CDR3 amino acid frequencies sectional content. (F) Top 10 frequency of CDR3 amino acid. (G) The length distribution of the V/D/J gene in the CDR3 region. (H) Deletion length distribution of the V/D/J gene. (I) Insertion length distribution between V and D genes, D and J genes. (J) Hypermutation, only for Ig. (K) J-gene usage. (L) V-gene usage. (M) Three-dimensional graph of V-J pairing.

iHMMune-align (Gaeta *et al.* 2007) generated a similar result to IgBLAST so it was excluded from the comparison. MiTCR (Bolotin *et al.* 2013) performance is strongly related to sequencing quality, so neither simulated data nor public sequences were suitable for it. Three spiked-in samples sequenced by Illumina were used to compare MiTCR and IMonitor, the result of which is shown in Figure S7. Although MiTCR finished the run much faster than IMonitor, clone G

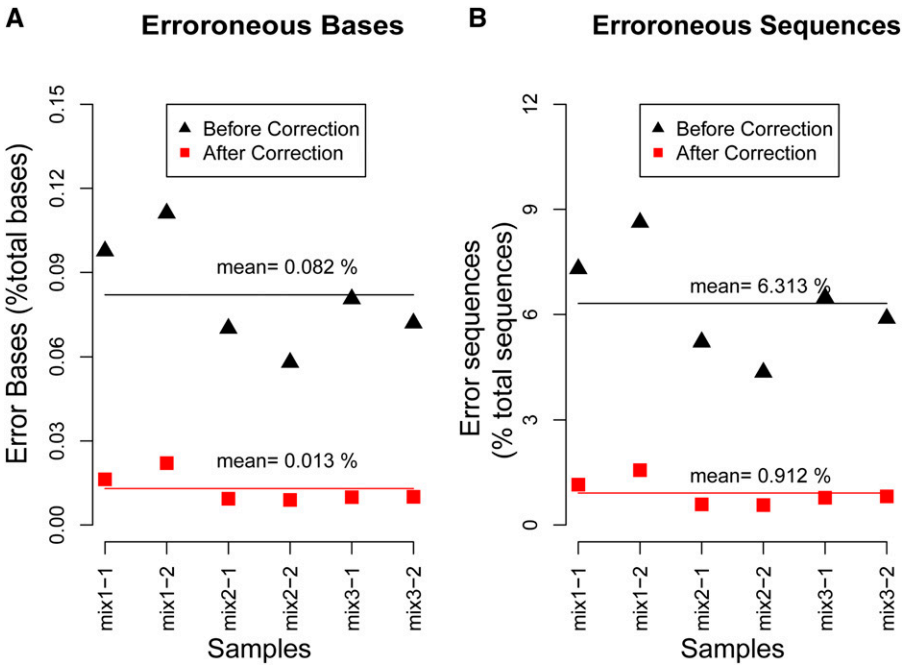

Figure 5 Error rate before and after correction. (A) The percentage of erroneous nucleotide bases divided by the total bases sequenced. (B) The percentage of sequences containing erroneous bases divided by total sequences. The primers at each sequence are excluded for the calculation of error rate.

was missed because of erroneous J-gene assignment and an incorrect CDR3 region. All tools employed their default parameters.

By processing the *in silico* data with different tools, we calculated the accuracy of V/D/J genes and V/D/J alleles of each tool analyzed (Figure 3 and Table S1, Table S2, and Table S3). For all 24 V/D/J genes and alleles in all TCR/BCR chains, IMonitor showed superior performance in 14 of them (58.33%), with 7 of 24 slightly lower in performance than the best tool (no more than 1% difference) and the remaining 3 from 1.6 to 3.25% difference (TRAV gene, TRBD gene, and TRBD allele). For the highly homologous V-gene family (at least 40 for each chain), the accuracy of IMonitor was >95% for almost all chains and exceeded 99% for all J genes. For D genes, which were short and embedded with deletions and insertions at both ends, IMonitor performed significantly better for IGH (>80% accuracy), while slightly worse for TRB. Decombinator was not designed to identify alleles and D genes.

In addition to the simulating data, public rearranged sequences that are annotated manually with clear V(D)J genes and extracted from the IMGT/LIGM-DB database (http://www.imgt.org/ligmdb/) were utilized to test IMonitor. Twenty-four TRB sequences and 1763 IGH sequences were analyzed by different tools (Table 1). For TRB, IMonitor and IgBLAST performed better than Decombinator in general, whereas IMonitor outperformed IgBLAST in D genes. For IGH, IMonitor performed similarly to IgBLAST in V and D alleles, but was superior in all other genes and alleles. The good performance in D genes by IMonitor demonstrated the effectiveness of the M-mismatch extension model during D gene realignment. The accuracies of IGH-J genes and alleles were both slightly lower for these two tools, because some public IGH sequences have only a partial J segment (<30 bp) and they are difficult to distinguish from other homologous genes and alleles.

To assess running time and memory needed, 10^5 simulated TRB and IGH data sets were analyzed by IMonitor, IgBLAST, and Decombinator separately; the results are shown in Table S8. When only one CPU was used, IMonitor took 12 min 52 sec and 21 min 96 sec to analyze the two data sets separately, with peak memory of 226 ∼ 325 Mb. Of all tools tested, Decombinator was the fastest and IMonitor ranked second.

Overall, IMonitor produced satisfactory results for both simulated and published sequences. It generated similar results to IgBLAST in some genes, while it outperformed other tools in most occasions. It is also direct proof that the realignment strategy for V(D)J identification is useful.

The output of IMonitor

One of the features that distinguish IMonitor from others is its ability to export comprehensive statistics for characteristics of TCR/BCR repertoire and accessible graphs. The statistics include not only basic statistics but also in-depth statistics (Figure S3). The former elucidates the process from raw data to effective sequences, such as clean data rate and V(D)J gene mapped rate, which all provide sequence number, rate of input, and rate of raw data. The latter consists of multiple statistics based on effective sequence, such as functional classification, V/J/V-J gene usage rate, clone number, diversity calculated by Shannon index, and hypermutation. IMonitor is also able to translate obscure data into self-explanatory graphs. Important statistics like V/J usage, top 10 clone frequencies, CDR3 segmental frequency statistics (split into four segments after frequency sorted: top100, 100-1E3, 1E3-1E4, >1E4), insertion and deletion length distribution, V/J nucleotide composition, and V-J pairing

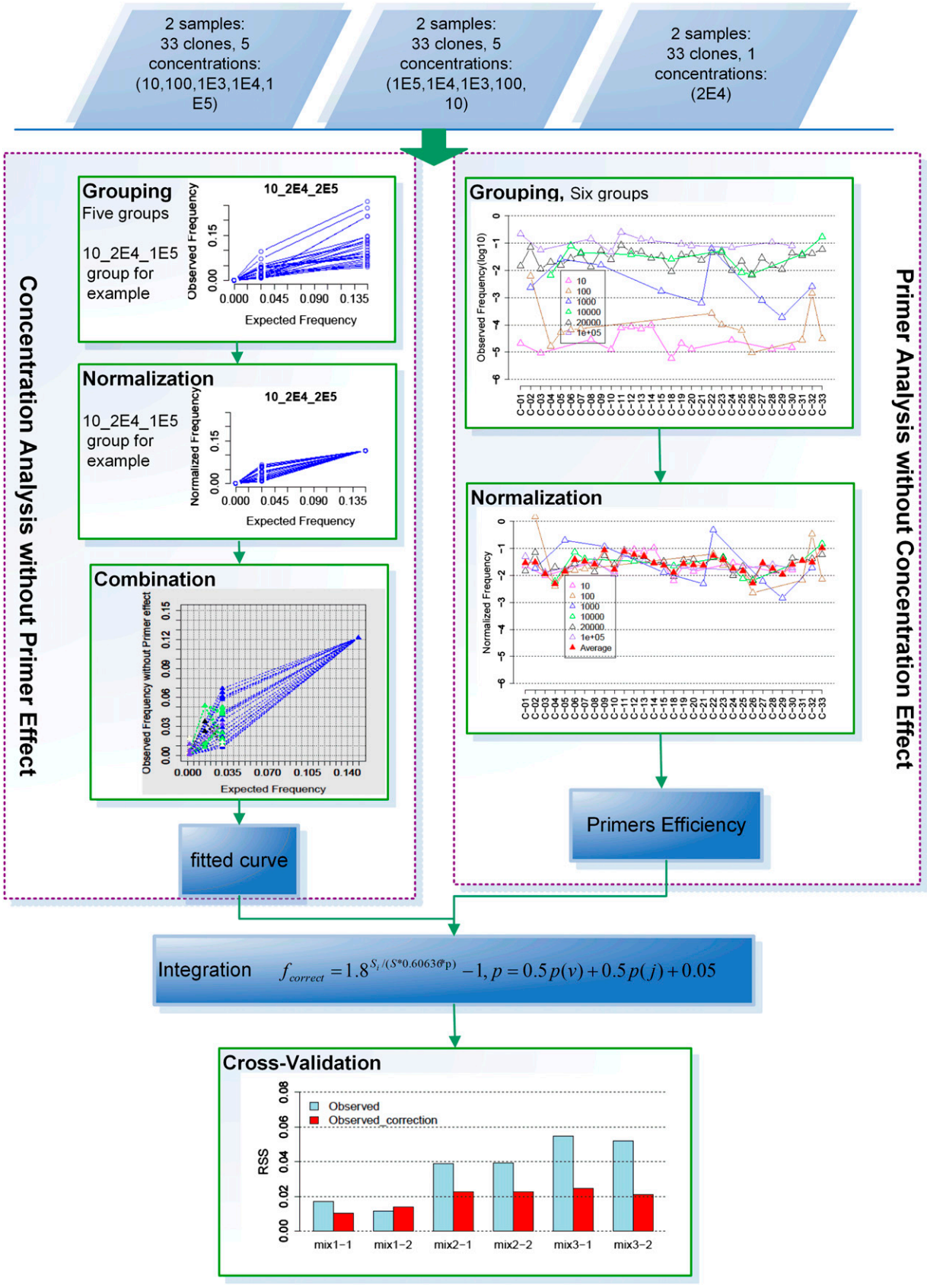

Figure 6 Minimize MPCR bias flowchart. Six samples are mixed together and search the bias rules under two independent pathways, concentration analysis and primer analysis. For concentration analysis, six groups are created to eliminate primer effect, and each group is normalized; then five groups

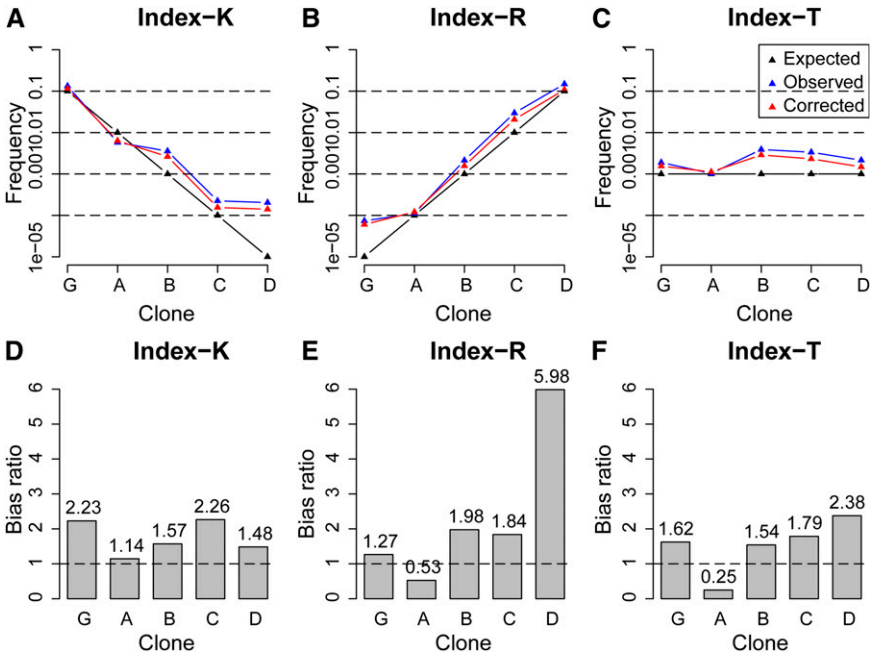

Figure 7 Evaluation of bias reduction in three spiked-in samples. Both observed frequencies and PCR bias-corrected frequencies were calculated for the five clones in three samples, and each clone provides a bias ratio to show bias change after correction. The bias ratio is defined as the observed frequency bias (observed frequency minus expected frequency) divided by corrected frequency bias (corrected frequency minus expected frequency). (A) Sample: Index-K, five clone frequencies. (B) Sample: Index-R, five clone frequencies. (C) Sample: Index-T, five clone frequencies. (D) Sample: Index-K, bias ratio. (E) Sample: Index-R, bias ratio. (F) Sample: Index-T, bias ratio.

can all be presented as figures (Figure 4, Figure S4, and Figure S8) . V-J pairing diversity is visualized (Figure 4M, Figure S4M), which can be applied to track the changes of the immune system over time and reveal immunological conditions.

PCR and sequencing errors correction

IMonitor also integrates a process to correct PCR and sequencing errors. The six plasmid mixture samples with three different pooling gradients (*Materials and Methods* and Table S4) were used to analyze the error characteristics and evaluate the effectiveness of error correction. The processed sequences were mapped to V/D/J references, and the final effective data were used to summarize the error characteristics shown in Figure S5. The known template sequences were used as references to calculate the error rate, and the results are shown in Figure 5 and Table S5. The average percentage of high-quality sequences was 74.88%. A total of 12.86% of sequences with low quality were corrected according to high-quality sequences, while the remaining 12.26% were discarded. Thereafter, an average 6.33% of low-abundance sequences were also corrected (Table S5). In consideration of the influence posed on error statistics by impurity of the plasmids during the cell culture before we mixed the plasmids, if the "erroneous" sequence was found in on less than four samples and was detected >100 times in each sample, it was excluded from error rate calculation. After the correction process, the mean error rate of all sequences was decreased from 0.082 to 0.013%, and the percentage of error-bearing sequences was decreased from 6.313 to 0.912% (Figure 5).

Minimization of multiplex PCR bias

We successfully developed a novel method to minimize the MPCR bias under a given set of multiplex primers. Details of this method are shown in Figure 6 and *Materials and Methods*. Cross-validation was used to evaluate this method, as shown in Figure 6. Residual sum of squares (RSS) (Draper and Smith 1998) was calculated for each test, where $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y})^2$, y_i is the observed frequency, and $\hat{y}$ is the expected frequency. Except for the mix 1-2 sample, the other five samples reduced the RSS value, which demonstrated that this method is evidently effective.

Moreover, the method was tested using three spiked-in samples, in which five known clones were spiked in 10^6 cells. Compared to the expected frequency, clones B, C, D, and G had obvious bias. After modifying the frequency with this method, the bias was relieved to a certain extent (Figure 7, A–C). The bias ratio was defined as the observed frequency bias divided by the corrected frequency bias. If the ratio is >1, it means the frequency is corrected positively. A total of 86.7% of the clones (except clone A in Index-R/T) generated a ratio >1, particularly clone D in Index-R. These results conclude that the method is indeed capable of minimizing MPCR bias (Figure 7, D and E).

IMonitor to monitor MRD

To test the feasibility of IMonitor in translational research, we applied it to analyze the data from two patients with B-ALL.

are combined together and a curve is fitted. For primer analysis, six groups (10, 100, 1000, 1E4, 2E4, 1E5) are created to eliminate the concentration effect. Each group is normalized by multiplying a constant term and combined together, and then the primer efficiencies are obtained. A formula for reducing PCR bias is generated by integrating these two factors. Finally, cross-validation is used to evaluate this method's robustness. Each time, five samples are used as training data, and the remaining one is used as testing data. The residual sum of squares, $\mathrm{RSS} = \sum_i (y_i - \hat{y})^2$, where y_i is the observed value, and $\hat{y}$ is the expected value.

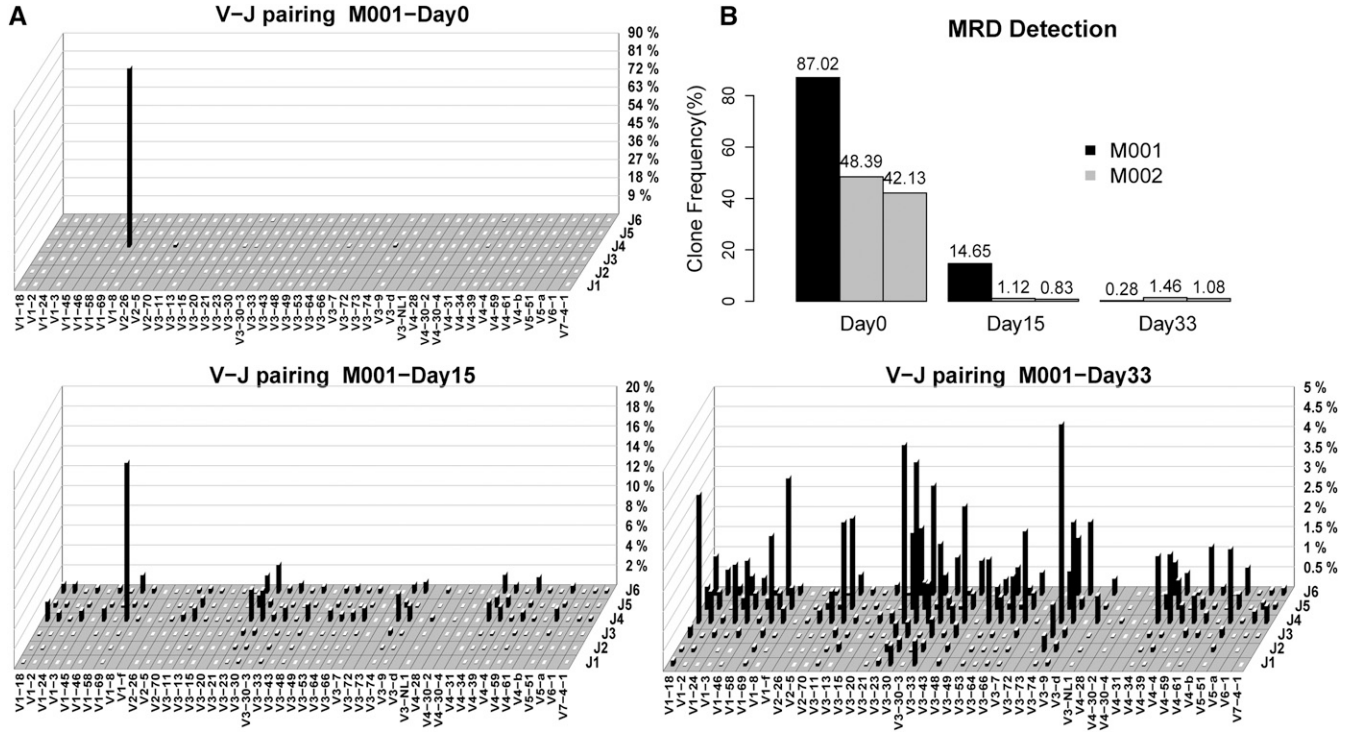

Figure 8 Detection of MRD in B-ALL using IMonitor. (A) Repertoire (V-J pairing) of IGH is shown for pretreatment (day 0) and post-treatment (day 15 and day 33). (B) Cancer clone frequency is shown for each clone in the two patients (M001 and M002) before treatment (day 0) and post-treatment (day 15 and day 33).

Samples were collected upon diagnosis and on day 15 and day 33 post-treatment (details in *Materials and Methods*). We first tried to identify the cancer clones in the B-cell receptor repertoire. Patients with B-ALL showed a clear pattern of deficient clonal diversity (Figure 8A, Figure S6). By the cutoff of 10% for the cancer clone frequency, the cancer clone was identified in patient M001 with clonal frequency of 87.02% upon diagnosis, whereas it substantially decreased to 14.65% on day 15 and further to 0.28% on day 33, suggesting the effect of treatment (Figure 8B). However, with the increased sensitivity of our method, MRD was detected on day 33. In contrast, flow cytometry was not able to detect any MRD in this time slot. Interestingly, two cancer clones (48.39% and 42.13%) were identified in patient M002, and the data of day 33 post-treatment showed an MRD level of 2.54% (1.46% and 1.08% corresponding to each clone), compared with the negative result in detection with flow cytometry. Although follow-up of the patients' condition was required to demonstrate its prognostic value, BCR sequencing plus IMonitor analysis showed its superiority in convenience and sensitivity for MRD detection. More importantly, the three-dimensional V-J pairing figure outputted by IMonitor revealed the remodeling of clonal diversity in the BCR repertoire following treatment, demonstrating its application in monitoring immune reconstruction (Figure 8A and Figure S6).

Discussion

We have developed a comprehensive methodology for analysis of the T-cell receptor repertoire and B-cell receptor repertoire made available by next generation sequencing technology. IMonitor provides an arsenal of solutions for four steps: basic data processing, V(D)J assignment, structural analysis, and statistics visualization. IMonitor distinguishes itself from other analysis tools with several features. The first important feature is its realignment process. The high homology among genes and alleles together with random base deletion and insertion at gene junctions have been affecting the accuracy of alignment. Therefore, global or local alignment by itself is not sufficient to complete the whole picture. During the realignment process, CDR3 regions are scrutinized with the M-mismatch extension model of local alignment while non-CDR3 regions are covered by global alignment. The test using simulated data and published rearrangement sequences demonstrated IMonitor's unquestionably better performance than other tools. The second feature of IMonitor is its ability to correct PCR and sequencing error and minimize MPCR bias, whose usage can be extended to other fields of research. IMonitor can be used to analyze any chain of T- and B-cell receptors and multiple species such as humans, monkeys, and rabbits. Furthermore, IMonitor results are presented with intuitive graphs. For example, the overall diversity of the immune system can be interpreted easily from a three-dimensional V-J pairing graph.

PCR and sequencing error of NGS is one of the problems that remain untackled for immune repertoire analysis. Preliminary results from previous studies show that a significant number of errors accumulate, and these errors can potentially lead to overestimating the actual TCR clonotypes. Besides,

sequencing error also results in artificially increased diversity of the TCR repertoire (Nguyen *et al.* 2011; Warren *et al.* 2011). Simply filtering out all low-quality sequences not only removes the sequencing error bases, but also leaves out a lot of genuine sequences. Our method, however, manages to decrease the error rate while rescuing most of the sequences. Actually, the efficiency of the approach would be improved if more factors are considered when correcting the errors. For example, erroneous bases have some bias for certain sequencing platforms, and the occurrence probability of sequencing error at each position of sequence is different. Using the six plasmid mixture samples, the characteristics of PCR and sequencing errors can be analyzed, as shown in Figure S5. The base error rate declines as the base quality in the sequence improves, whereas the percentage of discarded sequences increases sharply (Figure S5, A and B). The quality indexes of incorrect bases mainly fall into two categories: $Q \leq 10$ and $Q \geq 35$. Apparently, the former mostly results from sequencing error, while the cause of the latter is mostly PCR error (Figure S5D) . More than 85% of sequences have only one error base, and the rate rises after removing the sequences with minimal base quality of Q20 (Figure S5C).

Here we have introduced a new bioinformatics methodology to reduce the PCR bias of MPCR samples. We found that the bias originated from two factors: template concentration and inconsistent primer efficiencies. Using six plasmid mixture samples, we designed a formula to reduce the bias. By applying it to the spiked-in samples, we validated its effectiveness. However, due to the limited size of training data, some bias persisted in spiked-in samples. We believe when the training data contain $\geq$100 templates, the effect of the approach would be more significant. Besides, different primer sets should be trained to generate a suitable formula to reduce bias, so this article mainly introduces a bioinformatics approach showing how to create a suitable formula to adjust the bias. Previous literature reports that it reduce the bias mainly through an experimental method to optimize the primers and primer concentration (Carlson *et al.* 2013). It is a scientific and systemic experimental method to adjust primers. It would be ideal to use this method for optimizing primers in the first step and then to use our bioinformatics method for further reducing PCR bias. Stephen R. Quake and colleagues developed a consensus read sequencing approach that incorporated unique barcode labels (UIDs) on each starting RNA molecule (Vollmers *et al.* 2013). It could eliminate PCR bias completely in theory if the synthetic UIDs were random enough.

IMonitor for analyzing the TCRs and BCRs repertoire in human and other animal models has the widest applications among the available tools in basic and translational research. We have demonstrated its utility in identifying the cancer clonotypes and monitoring MRD in B-ALL, while at the same time evaluating the clonal diversity for immune remodeling following treatment by its graphic visualization. We believe that IMonitor can also be applied in many other areas, such as tracing emerging clonotypes upon vaccination and following their frequencies during the process, selecting monoclonal antibodies based on sequencing the immune repertoire. With the importance of immune repertoire research becoming more recognized, we believe IMonitor will play a role in advancing our understanding of the immune system.

Acknowledgments

We thank the Science and Technology Planning Project of Guangdong Province (no. 2012A031100010) for support and the Shenzhen Municipal Government of China (no. GJHZ20130417140835564) for support. This study is supported by the International Science and Technology Cooperation Program of Shenzhen, China (GJHZ20130417140835564).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni, 2012 Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135: 183–191.

Bolotin, D. A., M. Shugay, I. Z. Mamedov, E. V. Putintseva, M. A. Turchaninova *et al.*, 2013 MiTCR: software for T-cell receptor sequencing data analysis. Nat. Methods 10: 813–814.

Carlson, C. S., R. O. Emerson, A. M. Sherwood, C. Desmarais, M. W. Chung *et al.*, 2013 Using synthetic templates to design an unbiased multiplex PCR assay. Nat. Commun. 4: 2680.

Chao, A., 1984 Non-parametric estimation of the number of classes in a population. Scand. J. Stat. 11: 265–270.

Chao, A., 1987 Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43: 783–791.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner, 2004 WebLogo: a sequence logo generator. Genome Res. 14: 1188–1190.

Draper, N. R., and H. Smith, 1998 *Applied Regression Analysis*. Wiley, New York.

Fischer, N., 2011 Sequencing antibody repertoires: the next generation. MAbs 3: 17–20.

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.

Gaeta, B. A., H. R. Malming, K. J. Jackson, M. E. Bain, P. Wilson *et al.*, 2007 iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 23: 1580–1587.

Janeway, C., 2005 *Immunobiology: The Immune System in Health and Disease*. Garland Science, New York.

Lefranc, M.-P., and G. Lefranc, 2001a *The Immunoglobulin Factsbook*. Academic Press, San Diego.

Lefranc, M.-P., and G. Lefranc, 2001b *The T Cell Receptor Factsbook*. Academic Press, San Diego.

Li, S., M. P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli *et al.*, 2013 IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat. Commun. 4: 2333.

Liu, B., J. Yuan, S. M. Yiu, Z. Li, Y. Xie *et al.*, 2012 COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics 28: 2870–2874.

Nguyen, P., J. Ma, D. Pei, C. Obert, C. Cheng *et al.*, 2011 Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics 12: 106.

Robins, H., C. Desmarais, J. Matthis, R. Livingston, J. Andriesen *et al.*, 2012 Ultra-sensitive detection of rare T cell clones. J. Immunol. Methods 375: 14–19.

Robins, H. S., P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle *et al.*, 2009 Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114: 4099–4107.

Shannon, C. E., 1997 The mathematical theory of communication. 1963. MD Comput. 14: 306–317.

Sherwood, A. M., R. O. Emerson, D. Scherer, N. Habermann, K. Buck *et al.*, 2013 Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. Cancer Immunol. Immunother. 62: 1453–1461.

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.

Venturi, V., M. F. Quigley, H. Y. Greenaway, P. C. Ng, Z. S. Ende *et al.*, 2011 A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J. Immunol. 186: 4285–4294.

Vollmers, C., R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake, 2013 Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. USA 110: 13463–13468.

Wang, C., C. M. Sanders, Q. Yang, H. W. Schroeder, Jr., E. Wang *et al.*, 2010 High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc. Natl. Acad. Sci. USA 107: 1518–1523.

Warren, R. L., B. H. Nelson, and R. A. Holt, 2009 Profiling model T-cell metagenomes with short reads. Bioinformatics 25: 458–464.

Warren, R. L., J. D. Freeman, T. Zeng, G. Choe, S. Munro *et al.*, 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.

Ye, J., S. McGinnis, and T. L. Madden, 2006 BLAST: improvements for better sequence analysis. Nucleic Acids Res. 34: W6–W9.

Ye, J., N. Ma, T. L. Madden, and J. M. Ostell, 2013 IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 41: W34–W40.

Yousfi Monod, M., V. Giudicelli, D. Chaume, and M. P. Lefranc, 2004 IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20(Suppl 1): i379–i385.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller, 2000 A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7: 203–214.

Communicating editor: G. D. Stormo

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176735/-/DC1

IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis

Wei Zhang, Yuanping Du, Zheng Su, Changxi Wang, Xiaojing Zeng, Ruifang Zhang, Xueyu Hong, Chao Nie, Jinghua Wu, Hongzhi Cao, Xun Xu, and Xiao Liu

DOI: 10.1534/genetics.115.176735

Supporting Information inventory

Table S1. Simulated TRB with 0.1%, 0.5% and 2% sequencing error.

Mismatch Rate	Hyper-Mutation Rate	Software	TRBV-Gene(%)	TRBD-Gene(%)	TRBJ-Gene(%)	TRBV-allele(%)	TRBD-allele(%)	TRBJ-allele(%)
0.10%	0.00%	**IMMonitor**	98.25	77.04	100.00	82.95	72.13	99.98
		HighV-QUST	91.46	86.31	99.24	65.94	81.69	99.23
		IgBLAST	97.87	80.55	99.20	72.46	76.33	99.19
		Decombinator	70.28	-	80.76	29.53	-	70.17
0.50%	0.00%	**IMMonitor**	98.18	76.90	100.00	82.34	71.60	99.83
		HighV-QUST	91.44	85.77	99.24	66.04	80.75	99.14
		IgBLAST	97.80	79.32	99.02	72.58	74.95	98.92
		Decombinator	68.33	-	78.64	-	-	-
2%	0.00%	**IMMonitor**	98.13	77.82	100.00	80.01	70.77	99.42
		HighV-QUST	91.59	83.89	99.30	65.39	77.27	98.96
		IgBLAST	97.83	75.13	98.61	71.82	70.07	98.27
		Decombinator	59.58	-	68.45	24.85	-	60.18

Table S2. Simulated IGH with 0.1%, 0.5% and 2% sequencing error and 0.1% hyper-mutation

Mismatch Rate	Hyper-Mutation Rate	Software	IGHV-Gene(%)	IGHD-Gene(%)	IGHJ-Gene(%)	IGHV-allele(%)	IGHD-allele(%)	IGHJ-allele(%)
0.10%	0.10%	**IMonitor**	98.30	83.97	99.98	81.48	81.41	99.88
		HighV-QUST	87.97	72.90	98.60	69.38	70.81	97.21
		IgBLAST	98.76	75.68	99.83	80.74	73.7	99.73
0.50%	0.10%	**IMonitor**	98.19	83.86	99.98	81.3	81.28	99.66
		HighV-QUST	87.51	72.73	98.61	68.96	70.66	97.01
		IgBLAST	98.63	75.47	99.83	80.63	73.52	99.5
2%	0.10%	**IMonitor**	98.02	83.79	99.96	80.46	81.01	98.68
		HighV-QUST	87.72	72.33	98.59	68.43	70.04	96.17
		IgBLAST	98.52	74.61	99.79	79.63	72.22	98.55

Table S3. Simulated Data with 0.5% sequencing error (TRA/IGK/IGL) and 4% hyper-mutation for IGK/IGL

Gene	Mismatch Rate	Hyper-Mutation Rate	Software	V-Gene(%)	J-Gene(%)	V-allele(%)	J-allele(%)
IGK	0.50%	4.00%	**IMMonitor**	93.57	99.99	86.91	92.70
			HighV-QUST	73.13	99.97	64.94	93.19
			IgBLAST	91.53	100.00	85.46	92.94
IGL	0.50%	4.00%	**IMMonitor**	100.00	99.57	77.24	98.79
			HighV-QUST	98.82	97.54	76.85	97.26
			IgBLAST	100.00	99.67	78.25	99.53
TRA	0.50%	0.00%	**IMMonitor**	98.35	100.00	86.75	99.39
			HighV-QUST	100.00	94.12	83.31	93.85
			IgBLAST	100.00	100.00	85.32	99.93
			Decombinator	68.36	72.83	-	-

Table S4. Plasmid mixing pattern

Plasmid No.	V gene	J gene	Plasmid mix 1-1*	Plasmid mix 1-2*	Plasmid mix 2-1*	Plasmid mix 2-2*	Plasmid mix 3-1*	Plasmid mix 3-2*
C-01	TRBV10-1	TRBJ2-7	2000	2000	10	10	100000	100000
C-02	TRBV10-2/3	TRBJ2-7	2000	2000	1000	1000	1000	1000
C-03	TRBV11-1/2/3	TRBJ1-3	2000	2000	10	10	100000	100000
C-04	TRBV11-1/2/3	TRBJ1-5	2000	2000	100	100	10000	10000
C-05	TRBV12-3/4	TRBJ2-1	2000	2000	10000	10000	100	100
C-06	TRBV12-5	TRBJ2-1	2000	2000	100	100	10000	10000
C-07	TRBV13	TRBJ1-1	2000	2000	10000	10000	100	100
C-08	TRBV14	TRBJ2-7	2000	2000	100000	100000	10	10
C-09	TRBV15	TRBJ1-6	2000	2000	1000	1000	1000	1000
C-10	TRBV15	TRBJ2-4	2000	2000	100000	100000	10	10
C-11	TRBV16	TRBJ1-1	2000	2000	100000	100000	10	10
C-12	TRBV19	TRBJ1-6	2000	2000	100	100	10000	10000
C-13	TRBV20-1	TRBJ1-4	2000	2000	100000	100000	10	10
C-14	TRBV20-1	TRBJ1-5	2000	2000	10	10	100000	100000
C-15	TRBV20-1	TRBJ2-2	2000	2000	1000	1000	1000	1000
C-16	TRBV24-1	TRBJ1-2	2000	2000	100	100	10000	10000
C-17	TRBV25	TRBJ2-5	2000	2000	10000	10000	100	100
C-18	TRBV27/28	TRBJ2-4	2000	2000	100000	100000	10	10
C-19	TRBV29-1	TRBJ2-3	2000	2000	100000	100000	10	10
C-20	TRBV2	TRBJ2-6	2000	2000	10	10	100000	100000
C-21	TRBV30	TRBJ1-1	2000	2000	1000	1000	1000	1000
C-22	TRBV3-1	TRBJ1-2	2000	2000	10000	10000	100	100
C-23	TRBV4-1/2/3	TRBJ2-7	2000	2000	10000	10000	100	100
C-24	TRBV5-1	TRBJ2-1	2000	2000	100000	100000	10	10
C-25	TRBV5-4/5/6/8	TRBJ2-3	2000	2000	10000	10000	100	100
C-26	TRBV6-1/2/3/5/8	TRBJ2-1	2000	2000	10000	10000	100	100
C-27	TRBV6-4	TRBJ2-5	2000	2000	1000	1000	1000	1000
C-28	TRBV6-6	TRBJ1-6	2000	2000	10	10	100000	100000
C-29	TRBV6-9	TRBJ1-3	2000	2000	1000	1000	1000	1000
C-30	TRBV7-2/4/6/7/8	TRBJ2-6	2000	2000	10	10	100000	100000
C-31	TRBV7-3	TRBJ2-7	2000	2000	100	100	10000	10000
C-32	TRBV7-9	TRBJ1-4	2000	2000	1000	1000	1000	1000
C-33	TRBV9	TRBJ1-2	2000	2000	100	100	10000	10000

Note: * the clone ratio in the sample.

Table S5. Data process for PCR and sequencing error statistics.

Sample	Sum Sequence	High Quality Sequence(%)	Filter Sequence(%)	Low Quality Corrected(%)	PCR Error Corrected(%)	Effective Data	Before Correction		After Correction	
							Base Error(%)	Sequence Error (%)	Base Error(%)	Sequence Error(%)
index-1	4,273,571	78.50	9.28	12.21	8.36	3,876,775	0.098	7.304	0.016	1.152
index-2	4,217,557	78.59	9.77	11.64	9.10	3,805,551	0.111	8.628	0.022	1.564
index-3	3,603,556	70.16	16.25	13.59	4.44	3,018,100	0.070	5.218	0.009	0.590
index-10	5,078,119	81.28	8.80	9.92	4.70	4,631,376	0.058	4.359	0.009	0.569
index-11	2,785,059	64.59	18.61	16.80	5.32	2,266,836	0.081	6.484	0.010	0.779
index-12	3,335,881	76.15	10.87	12.98	6.04	2,973,382	0.072	5.885	0.010	0.816

Table S6. Samples information

Sample	species	Gene	Library	Experimental method	Sequencer	Type	Amount
H-B-01	Human	TRB	cDNA	MPCR	Hiseq2500	PE100	1ug
H-H-01	Human	IGH	DNA	MPCR	Hiseq2000	PE100	3ug
M001	Human	IGH	DNA	MPCR	Hiseq2500	PE150	1.2ug
M002	Human	IGH	DNA	MPCR	Hiseq2500	PE150	1.2ug

Table S7. Experimental design for five CD4+ T cell clones in the three spiked in mix.

Clone	TCRB V	TCRB J	CDR3	Mix 1	Mix 2	Mix 3
G	VB8	TRBJ1-1	CASSLGGQGVG	100,000	1000	10
A	VB5.1	TRBJ2-5	CASSPGIAELKETQY	10,000	1000	100
B	VB6.7	TRBJ2-7	CASHTGFVSYEQY	1000	1000	1000
C	VB4	TRBJ1-4	CSVGTGDNEKLF	100	1000	10,000
D	VB4	TRBJ1-4	CSVGQGDNEKLF	10	1000	100,000

Table S8. Performance of IMonitor and other tools on the simulated dataset.

	Peak Memory(MB)	Run Time
Data set of TRB(10^5 sequences)		
IMonitor [a]	325.88M	12m52s
IgBLAST [b]	327.95M	27m46s
Decombinator [c]	209.00M	1m23s
HighV-QUST[d]	-	-
Data set of IGH (10^5 sequences)		
IMonitor [a]	226.22M	21m96s
IgBLAST[b]	196.22M	92m30s
HighV-QUST[d]	-	-

Note: [a], run with 1cpu and blast (-a 1); [b], run with 1cpu and iglast (-num_threads 1);

[c], run with command prompt; [d], run online, send the results to user after 1-2weeks

Table S9. TRB and IGH V/J primers

IGH V/J Primers		**TRB V Primers**	
IGHV1-18	AGAGTCACCATGACCACAGAC	TRBV2	ATTTCACTCTGAAGATCCGGTCCAC
IGHV1-2/1-46	AGAGTCACCAKKACCAGGGAC	TRBV3-1	AAACAGTTCCAAATCGMTTCTCAC
IGHV1-24	AGAGTCACCATGACCGAGGAC	TRBV4-1/2/3	CAAGTCGCTTCTCACCTGAATG
IGHV1-3/1-45	AGAGTCACCATTACYAGGGAC	TRBV5-1	GCCAGTTCTCTAACTCTCGCTCT
IGHV1-69/1-f	AGAGTCACGATWACCRCGGAC	TRBV5-4/5/6/8	TCAGGTCGCCAGTTCCCTAAYTAT
IGHV1-8	AGAGTCACCATGACCAGGAAC	TRBV6-4.1	CACGTTGGCGTCTGCTGTACCCT
IGH2-70/26/5	ACCAGGCTCACCATYWCCAAGG	TRBV6-8/5/1.2	CAGGCTGGTGTCGGCTGCTCCCT
IGHV3	GGCCGATTCACCATCTCMAG	TRBV6-9/7/1.1/6	CAGGCTGGAGTCAGCTGCTCCCT
IGH4	CGAGTCACCATRTCMGTAGAC	TRBV6-4.2	AGTCGCTTGCTGTACCCTCTCAG
IGHV5-51	CAGCCGACAAGTCCATCAGC	TRRBV6-2/3	GGGGTTGGAGTCGGCTGCTCCCT
IGHV6-1	AGTCGAATAACCATCAACCCAG	TRBV7-2/4/6/7/8	GGGATCCGTCTCCACTCTGAMGAT
IGHV7	GACGGTTTGTCTTCTCCTTG	TRBV7-3	GGGATCCGTCTCTACTCTGAAGAT
IGHJ	CTGAGGAGACGGTGACCRKKGT	TRBV7-9	GGGATCTTTCTCCACCTTGGAGAT
		TRBV9	CCTGACTTGCACTCTGAACTAAACCT
		TRBV10-1	CCTCACTCTGGAGTCTGCTGCC
		TRBV10-2/3	CCTCACTCTGGAGTCMGCTACC
		TRBV11-1/2/3	GCAGAGAGGCTCAAAGGAGTAGACT
		TRBV12-3.2/5.2	GAAGGTGCAGCCTGCAGAACCCAG
TRB J Primers		TRBV12-3.1/4/5.1	GAAGATCCAGCCCTCAGAACCCAG
TRBJ1.1	CTTACCTACAACTGTGAGTCTGGTG	TRBV13	TCGATTCTCAGCTCAACAGTTC
TRBJ1.2	CTTACCTACAACGGTTAACCTGGTC	TRBV14	GGAGGGACGTATTCTACTCTGAAGG
TRBJ1.3	CTTACCTACAACAGTGAGCCAACTT	TRBV15	TTCTTGACATCCGCTCACCAGG
TRBJ1.4	AAGACAGAGAGCTGGGTTCCACT	TRBV16	CTGTAGCCTTGAGATCCAGGCTACGA
TRBJ1.5	CTTACCTAGGATGGAGAGTCGAGTC	TRBV18	TAGATGAGTCAGGAATGCCAAAG
TRBJ1.6	CATACCTGTCACAGTGAGCCTG	TRBV19	TCCTTTCCTCTCACTGTGACATCGG
TRBJ2.1	CCTTCTTACCTAGCACGGTGA	TRBV20-1	AACCATGCAAGCCTGACCTT
TRBJ2.2	CTTACCCAGTACGGTCAGCCT	TRBV24-1	CTCCCTGTCCCTAGAGTCTGCCAT
TRBJ2.3	CCGCTTACCGAGCACTGTCAG	TRBV25-1	GCCCTCACATACCTCTCAGTACCTC
TRBJ2.4	AGCACTGAGAGCCGGGTCC	TRBV27-1	GATCCTGGAGTCGCCCAGC
TRBJ2.5	CGAGCACCAGGAGCCGCGT	TRBV28	ATTCTGGAGTCCGCCAGC
TRBJ2.6	CTCGCCCAGCACGGTCAGCCT	TRBV29-1	AACTCTGACTGTGAGCAACATGAG
TRBJ2.7	CTTACCTGTGACCGTGAGCCTG	TRBV30-F5	CAGATCAGCTCTGAGGTGCCCCA

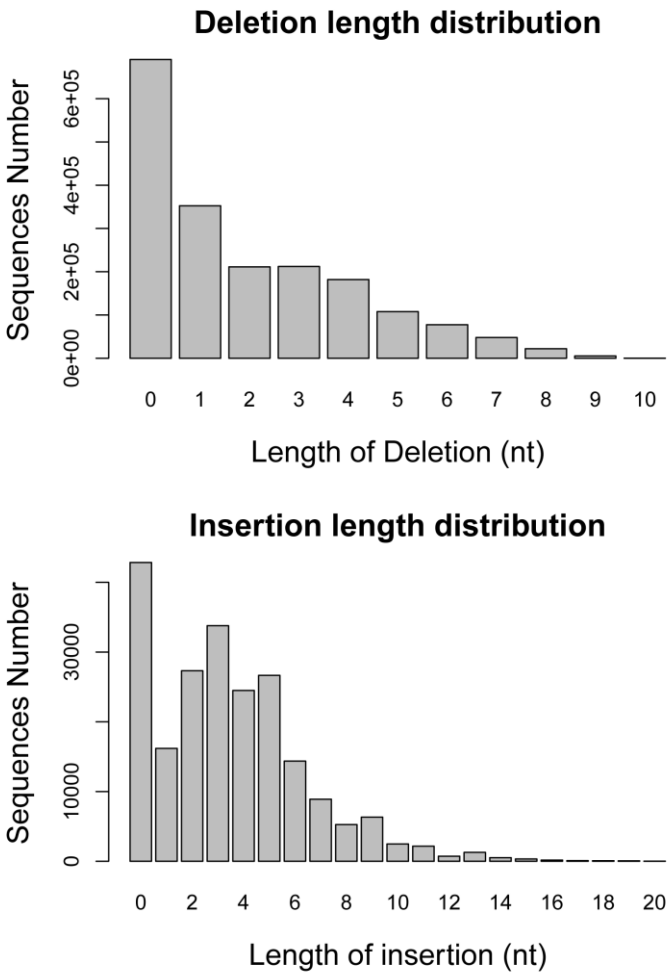

Figure S1. Insertion and deletion length distribution for simulated data.

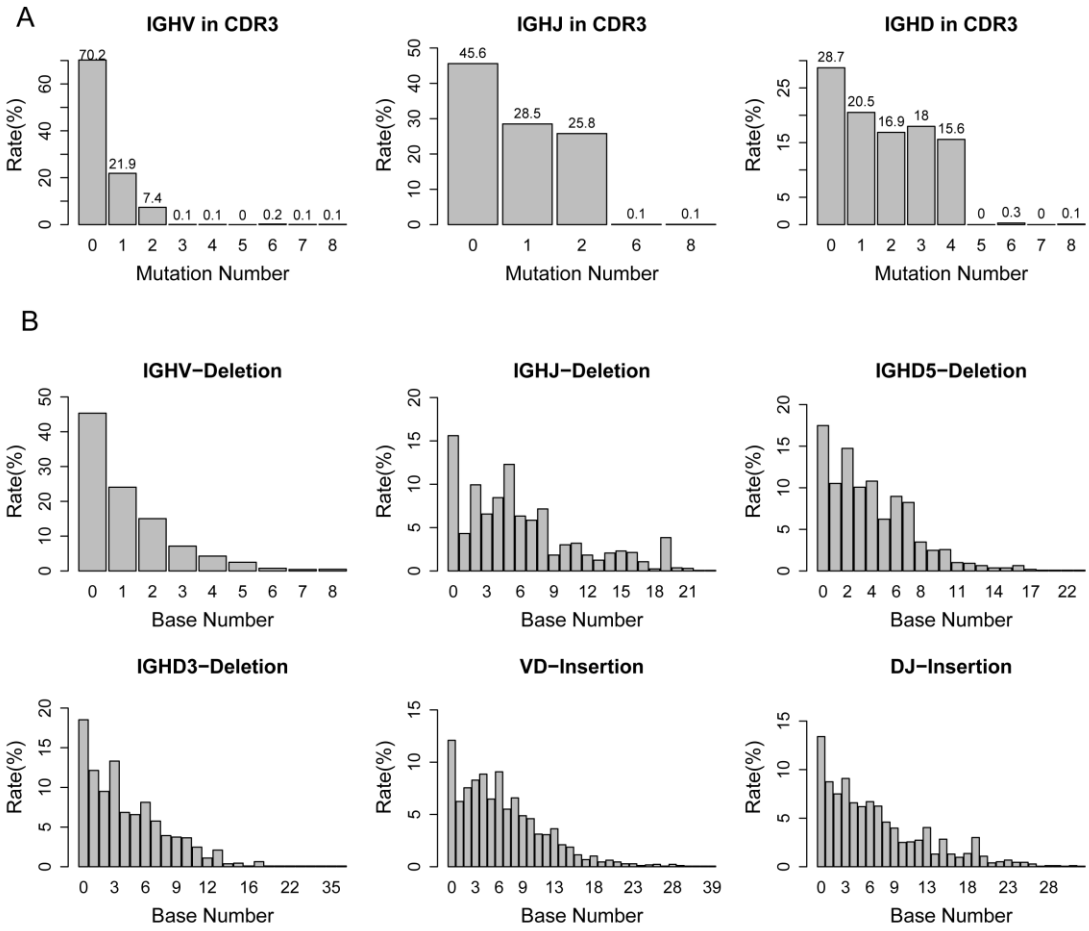

Figure S2. IGH-VDJ Mutation and deletion/insertion analysis on the public sequences. (A) VDJ mutation number statistics. (B) VDJ deletion/insertion length statistics. The data sets were obtained from IMGT/LIGM-DB database(http://www.imgt.org/ligmdb/), searched by "Homo sapiens", "rearranged", "TRB" or "IGH", and then selected the sequences annotated by manual(Annot. level=="manual") and annotated by V,D,J genes. So these sequences have fairly high level of annotation confidence.

1. Sample Basical Statitstics

#Title Seq_num Rate_of_input(%) Rate_of_rawdata(%)
Raw_seq_number(PE=1) 8126815 - -
Clean_data 7955514 97.89 97.89
PE_read_merged 7934499 99.74 97.63
Merged_with_highquality 7806823 98.13 96.06
V_alignment 7692062 98.53 94.65
D_alignment 3944006 50.52 48.53
J_alignment 7509383 96.19 92.40
VJ_alignment 7419604 95.04 91.29
CDR3_found_VJ 7313503 98.57 89.99
CDR3_found_byconserve - - -
PCR_Sequencing_correct 6569719 89.83 80.84
Effective_data 6188283 94.19 76.14

----------Note:----------
Clean_data: filter the Adapter pollution,low quality sequence
Effective_data: filter the sequence: 1. cannot find CDR3;
2. V and J strand conflict; 3. CDR3 less than 0bp;
4. sequence abundance filter.

2. Sample Further Statistics

in-frame: 5986614 96.74
out-of-frame(stop_codon): 33909 0.55
out-of-frame(CDR3_length): 105860 1.71
non-function: 61899 1.00
V_gene_used: 48 100.00
J_gene_used: 13 100.00
V-J_pairing: 558 89.42
Uniq_number(seq_nt,seq_aa): 1152945 926184
Uniq_number(cdr3_nt,cdr3_aa): 204878 182609
Shannon_index(seq,seq_aa): 16.23 15.74
Shannon_index(cdr3_nt,cdr3_aa): 14.47 14.25
Shanono_index(V,J,V-J): 3.84 2.54 6.22
Hyper-mutation(base_rate,seq_rate): 0.00 0.00

Figure S3. Outputs of IMonitor, H-B-01 as an example. Sample basic statistics show the data procedure, from raw data to effecive data, such as paired-end reads merged, V(D)J alignment rate. Sample further statistics, show the multiple statistics based on effective data.

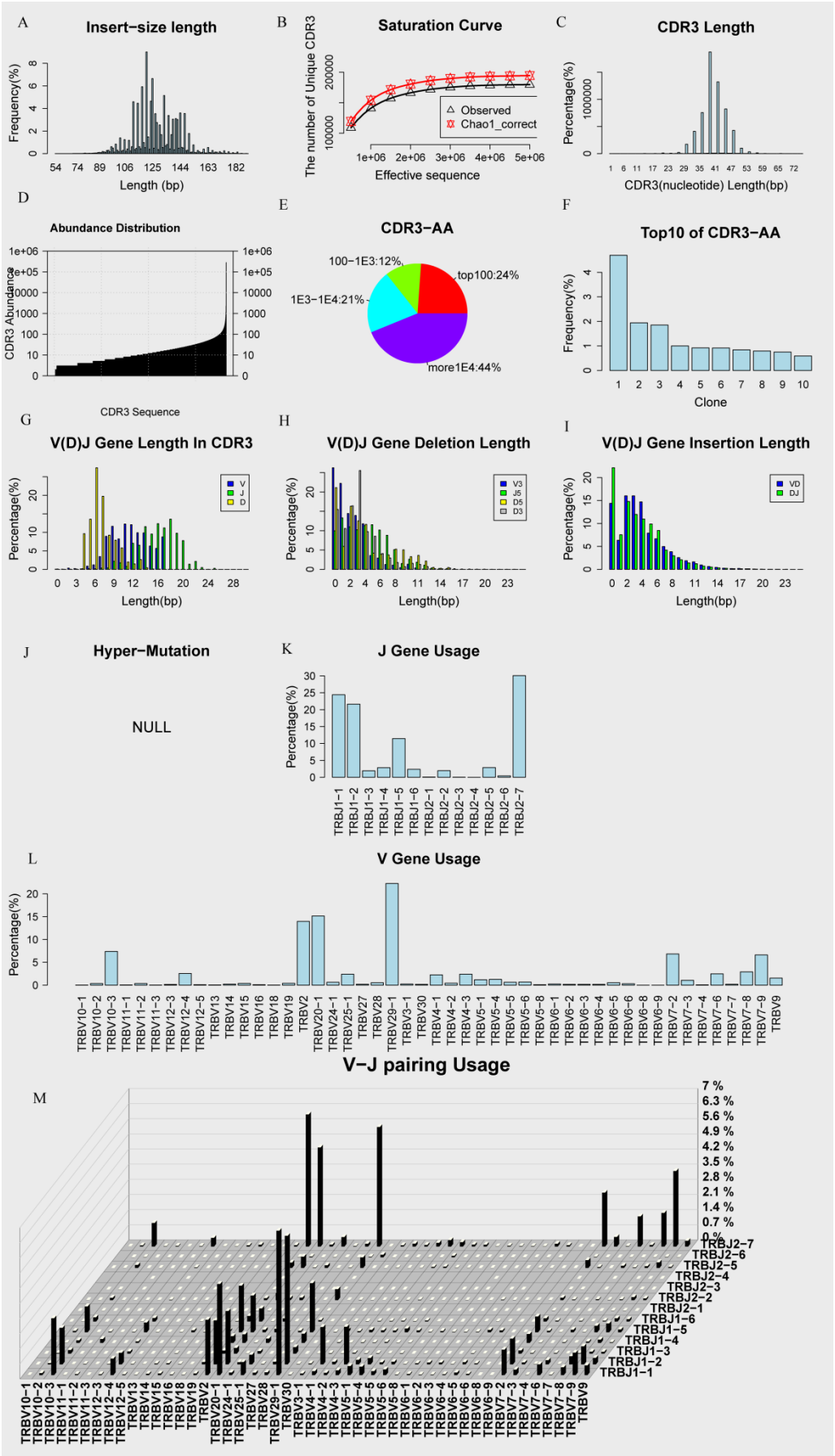
A
Insert−size length
Frequency(%)
Length (bp)
B
Saturation Curve
The number of Unique CDR3
Observed
Chao1_correct
Effective sequence
C
CDR3 Length
Percentage(%)
CDR3(nucleotide) Length(bp)
D
Abundance Distribution
CDR3 Abundance
CDR3 Sequence
E
CDR3−AA
100−1E3:12%
top100:24%
1E3−1E4:21%
more1E4:44%
F
Top10 of CDR3−AA
Frequency(%)
Clone
G
V(D)J Gene Length In CDR3
Percentage(%)
Length(bp)
H
V(D)J Gene Deletion Length
Percentage(%)
Length(bp)
I
V(D)J Gene Insertion Length
Percentage(%)
Length(bp)
J
Hyper−Mutation
NULL
K
J Gene Usage
Percentage(%)
L
V Gene Usage
Percentage(%)
V−J pairing Usage
M

Figure S4. H-B-01 sample output figure of IMonitor. (A) Sequence length distribution. (B) Saturation curve, rarefaction studies of sequences. Sub-sequences are randomly selected and observed unique CDR3 number and predicted CDR3 number (Chao1 corrected algorithm) are calculated. (C) CDR3 nucleotide length distribution. (D) CDR3 abundance distribution. (E) CDR3 amino acid frequencies sectional content. (F) Top ten frequency of CDR3 amino acid. (G) Length distribution of V/D/J gene in CDR3 region. (H) Deletion length distribution of V/D/J gene. (I) Insertion length distribution of between V and D gene, D and J gene. (J) Hyper-mutation, Only for Ig. (K), J gene usage. (L) V gene usage. (M) Three-dimensional graph of V-J pairing.

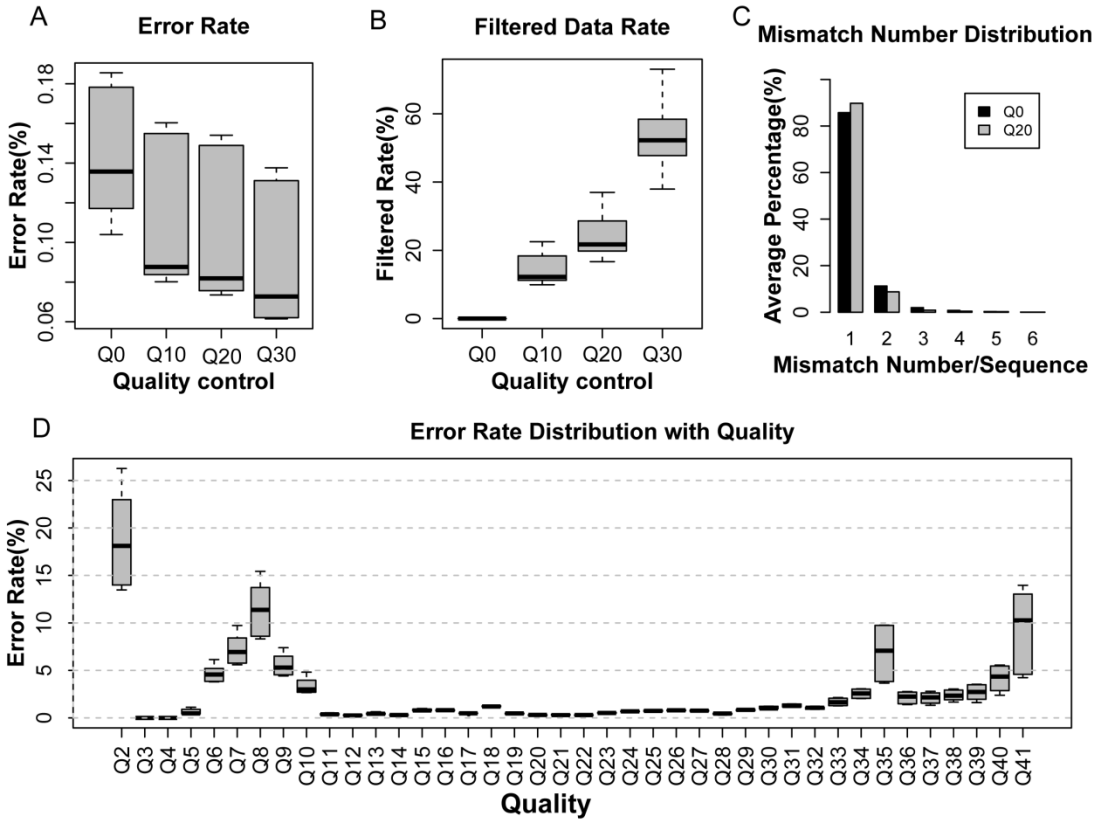

Figure S5. Error characteristics of 6 plasmid mix samples. (A) Error base rate after sequence filtering by different minimal quality value. For example, Q20 means filter the sequence with at least one base quality less than Q20. (B) Removed data rate after sequence filtering by different minimal quality. (C) Mismatch number distribution, raw sequences (Q0, no filtration) and sequences after filtering by minimal quality 20(Q20). (D) Error base distribution with base quality. Only unique sequences are considered.

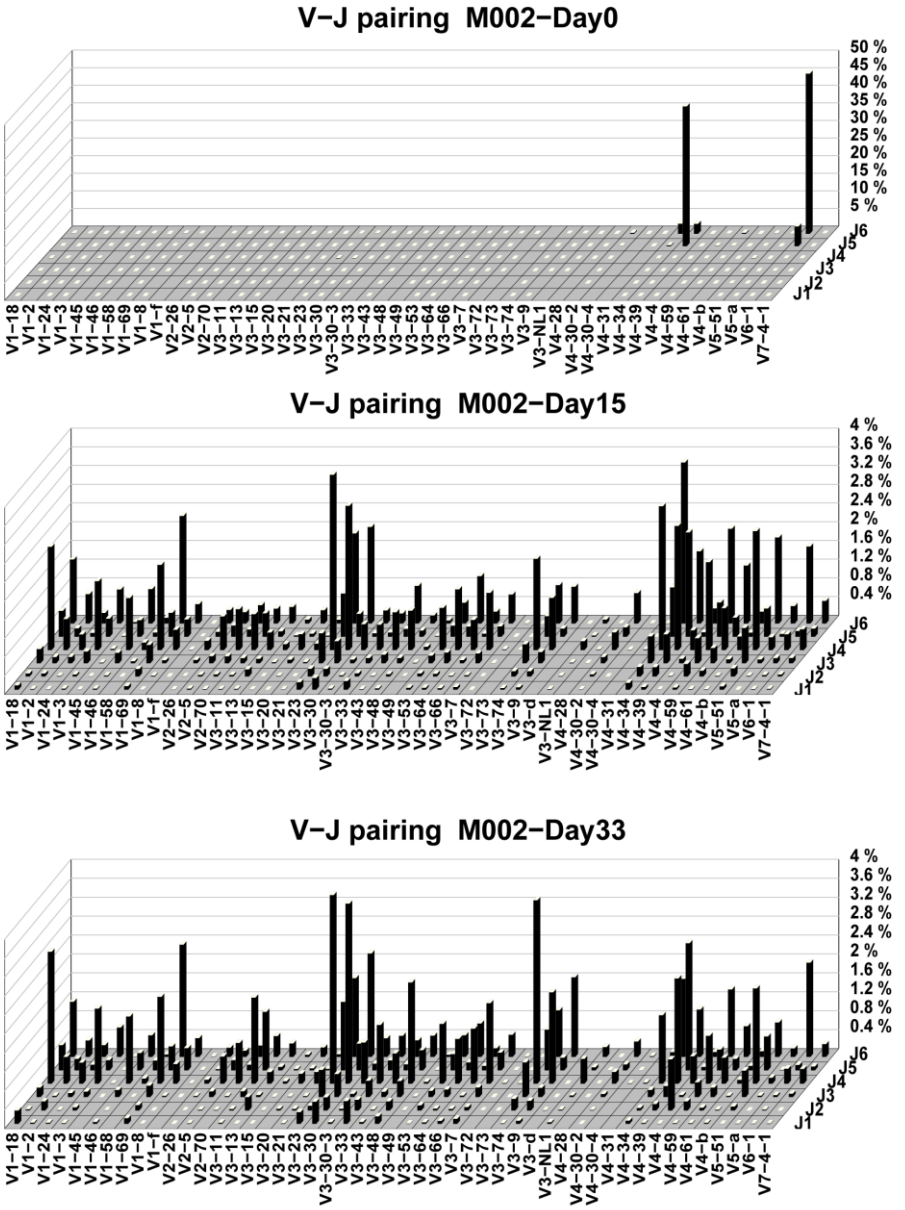

Figure S6. V-J pairing dynamics for M002. Day 0 for pre-treatment, Day 15 and Day 33 for post-treatment.

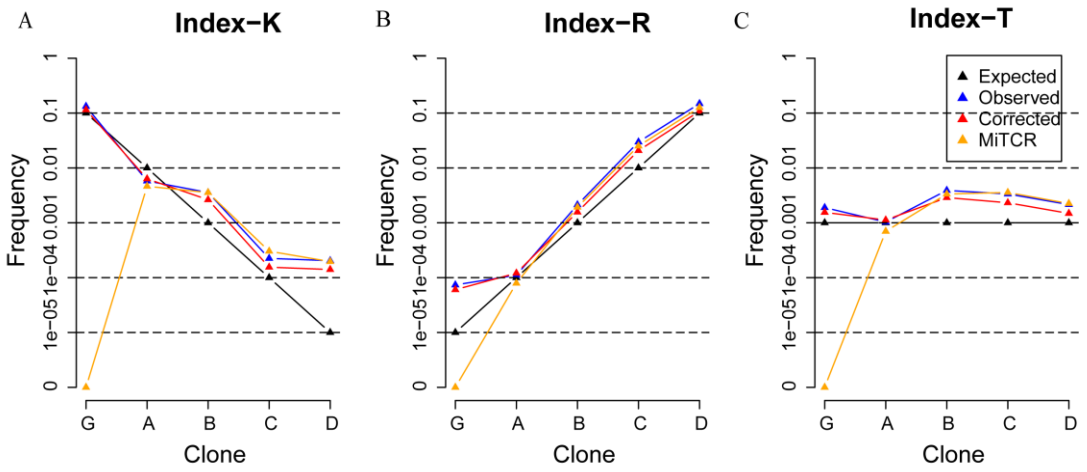

Figure S7. MiTCR and IMonitor performance in 3 spiked-in samples.

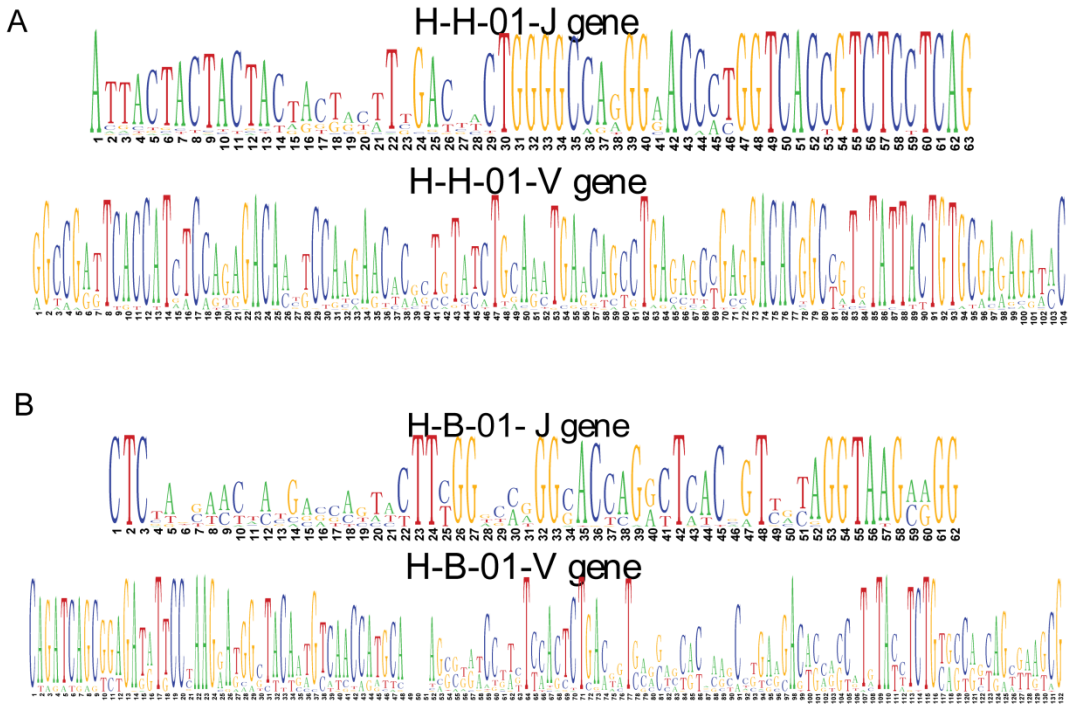

Figure S8. Nucleotide composition of V/J genes. (A) H-H-01 sample. (B) H-B-01 sample.