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

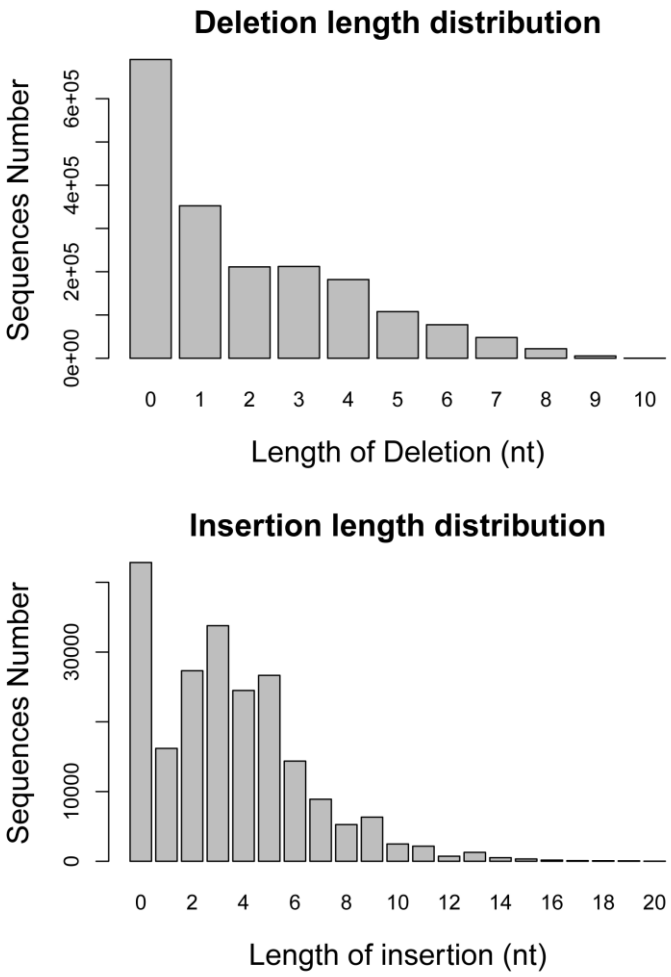

Figure S1. Insertion and deletion length distribution for simulated data.

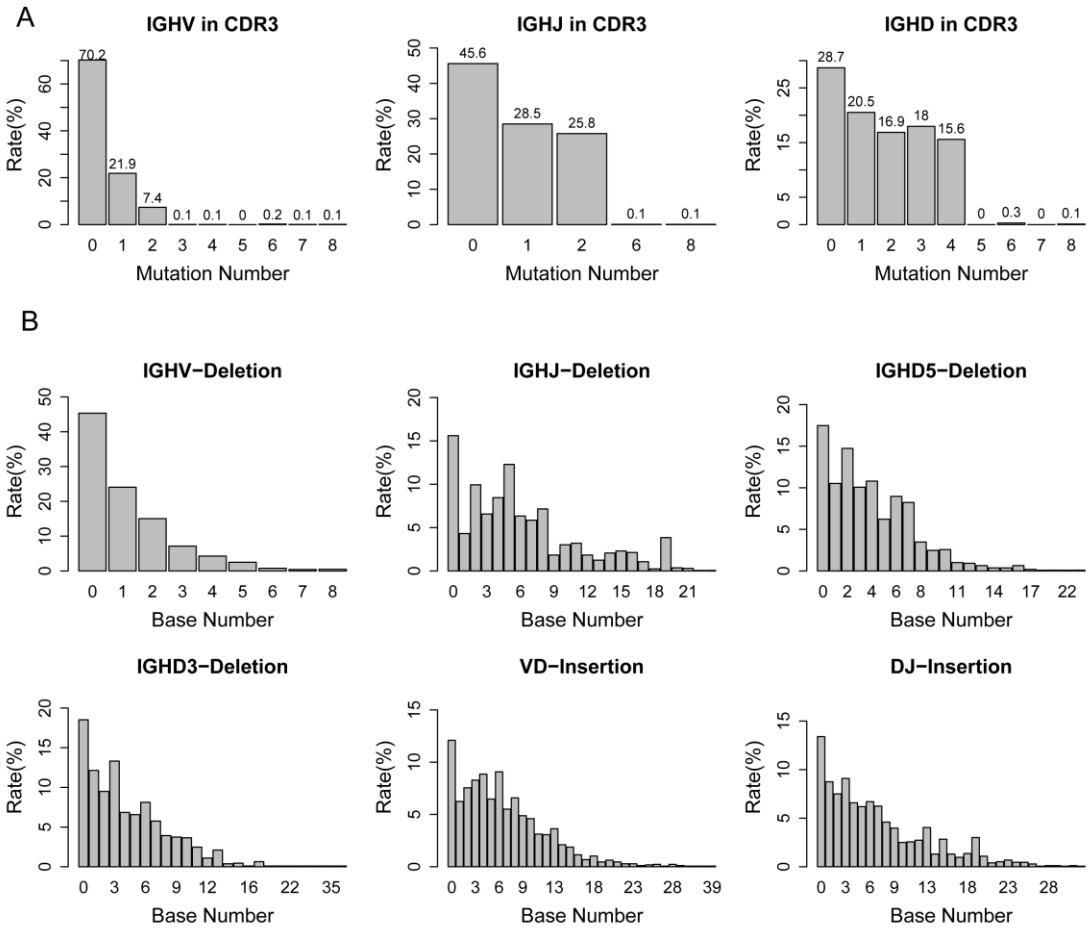

Figure S2. IGH-VDJ Mutation and deletion/insertion analysis on the public sequences. (A) VDJ mutation number statistics. (B) VDJ deletion/insertion length statistics. The data sets were obtained from IMGT/LIGM-DB database(http://www.imgt.org/ligmdb/), searched by "Homo sapiens", "rearranged", "TRB" or "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.

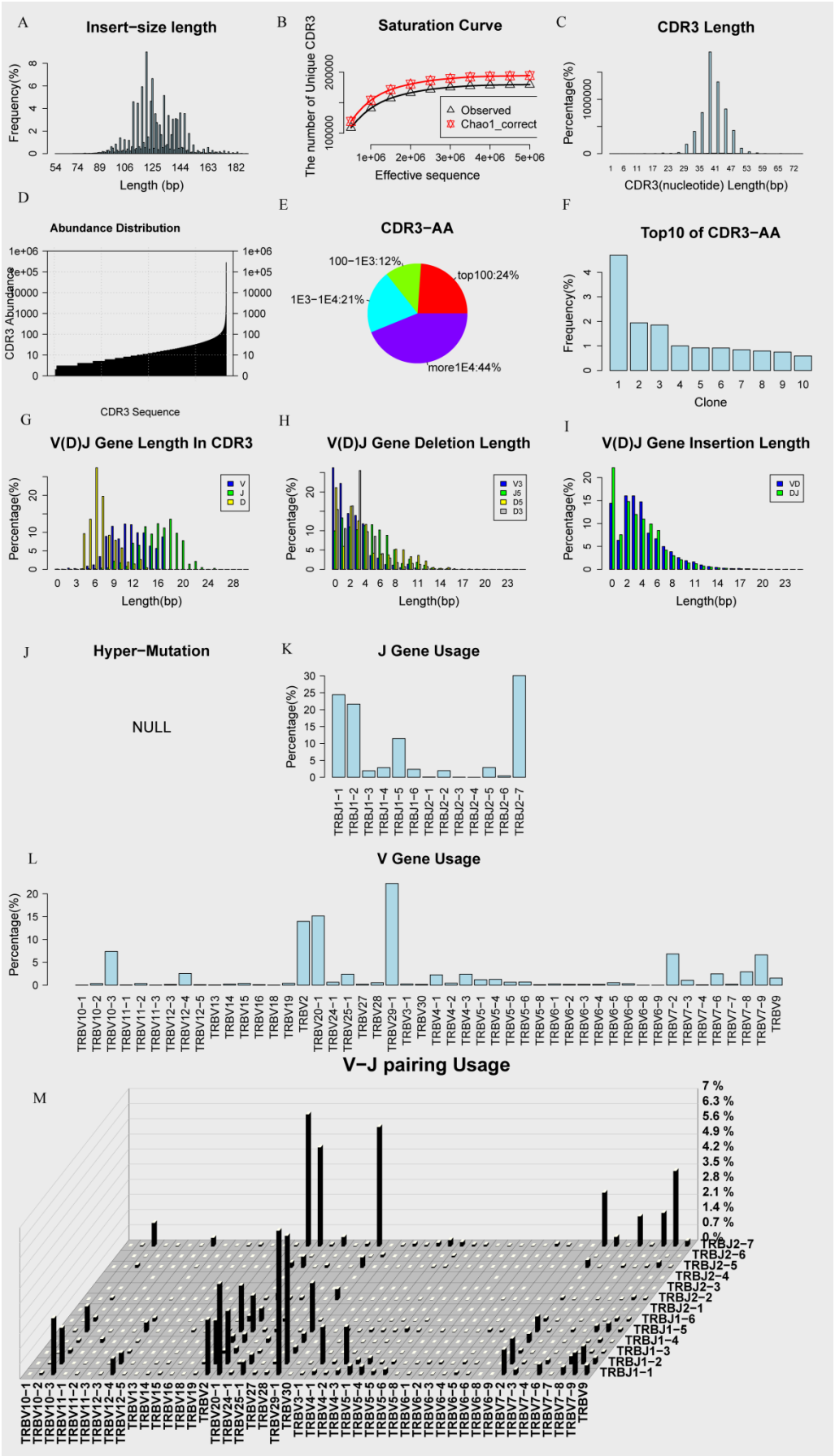
A
Insert−size length
Frequency(%)
Length (bp)
B
Saturation Curve
The number of Unique CDR3
Observed
Chao1_correct
Effective sequence
C
CDR3 Length
Percentage(%)
CDR3(nucleotide) Length(bp)
D
Abundance Distribution
CDR3 Abundance
CDR3 Sequence
E
CDR3−AA
100−1E3:12%
top100:24%
1E3−1E4:21%
more1E4:44%
F
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.

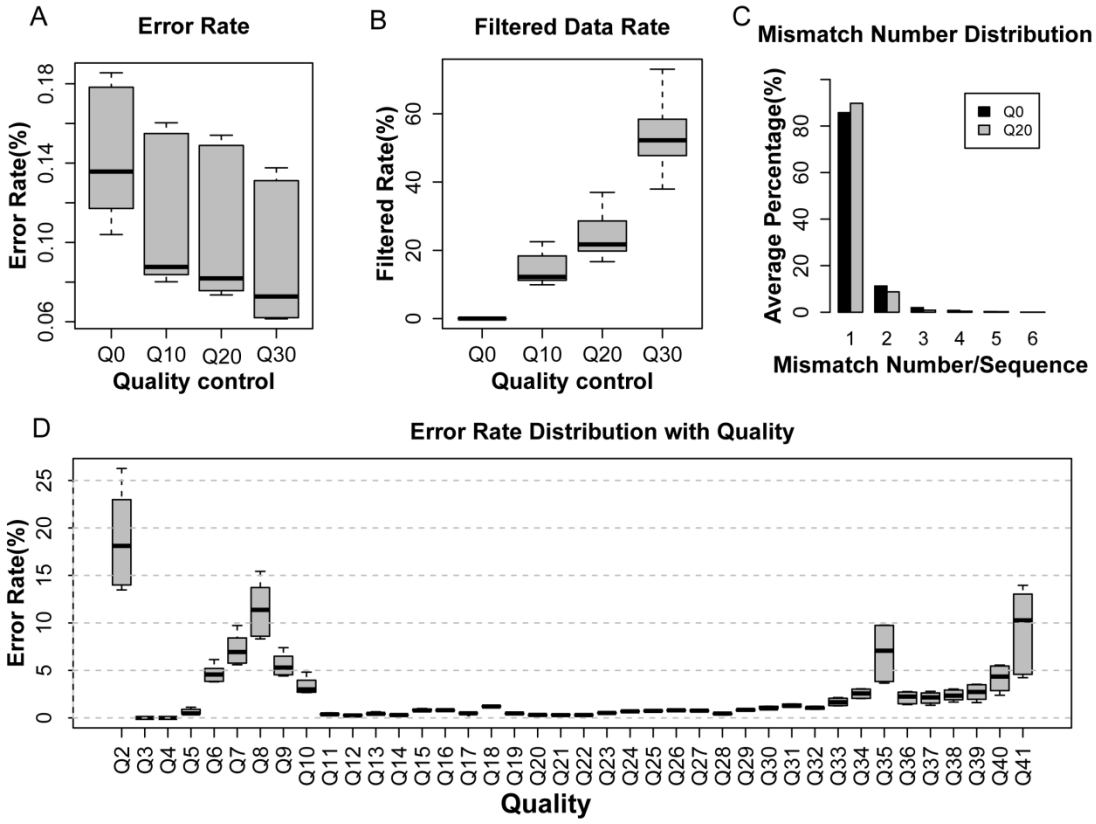

Figure S5. Error characteristics of 6 plasmid mix samples. (A) Error base rate after sequence filtering by different minimal quality value. For example, Q20 means filter the sequence with at least one base quality less than Q20. (B) Removed data rate after sequence filtering by different minimal quality. (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.

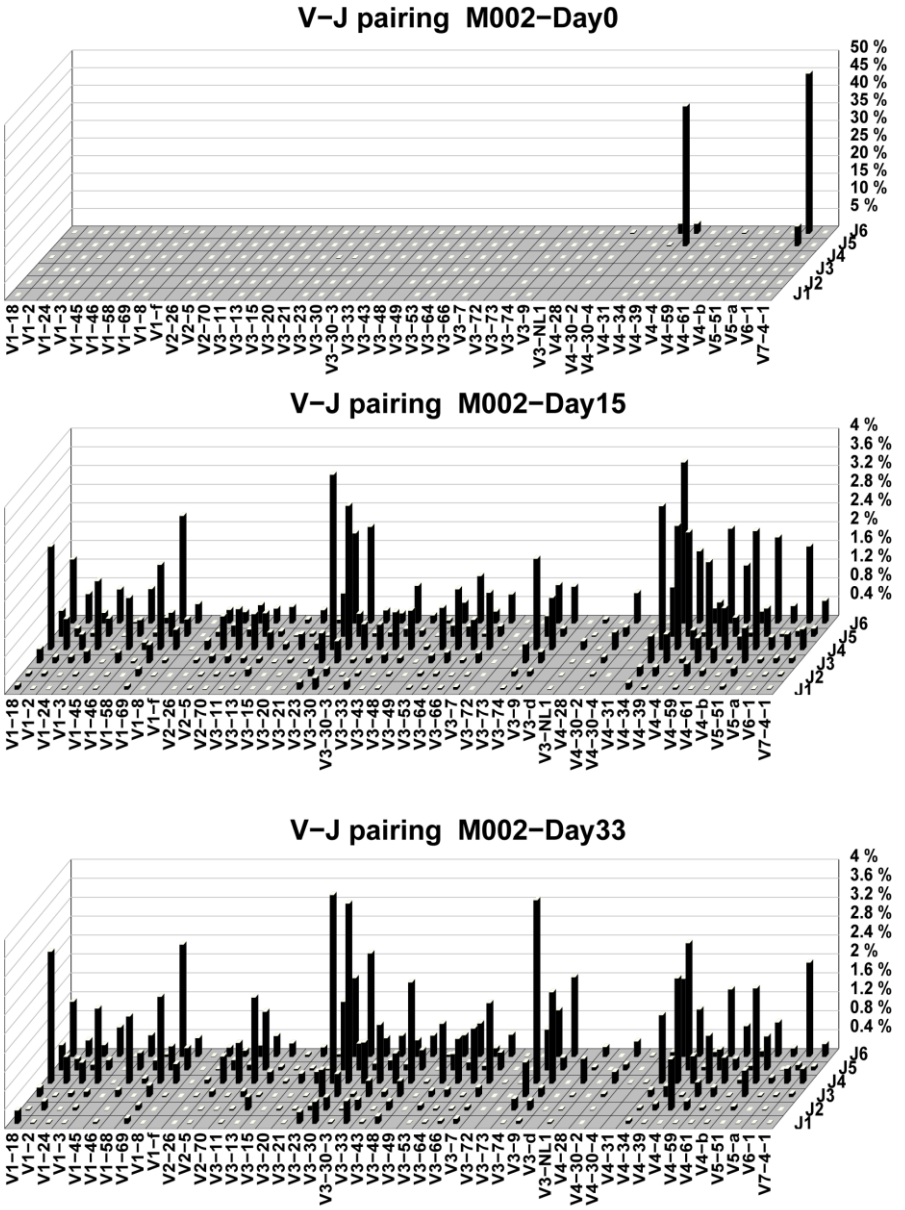

Figure S6. V-J pairing dynamics for M002. Day 0 for pre-treatment, Day 15 and Day 33 for post-treatment.

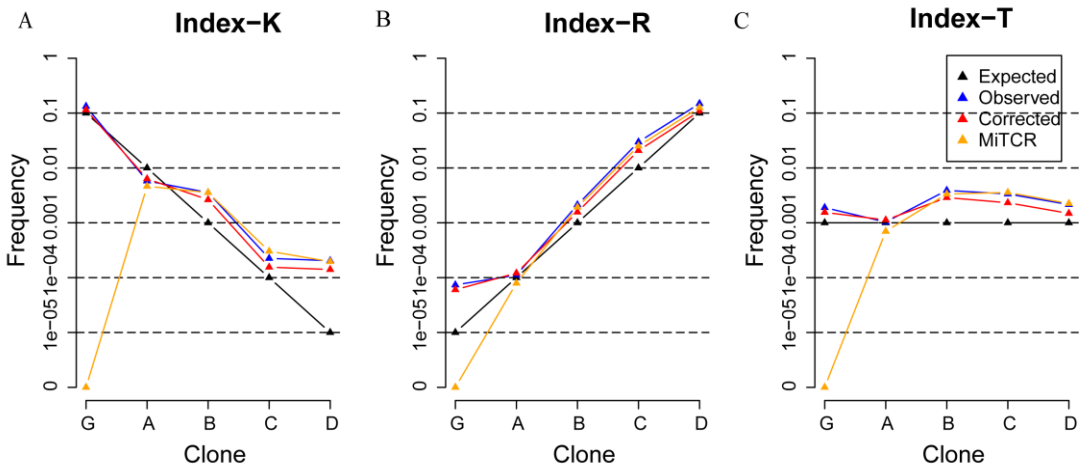

Figure S7. MiTCR and IMonitor performance in 3 spiked-in samples.

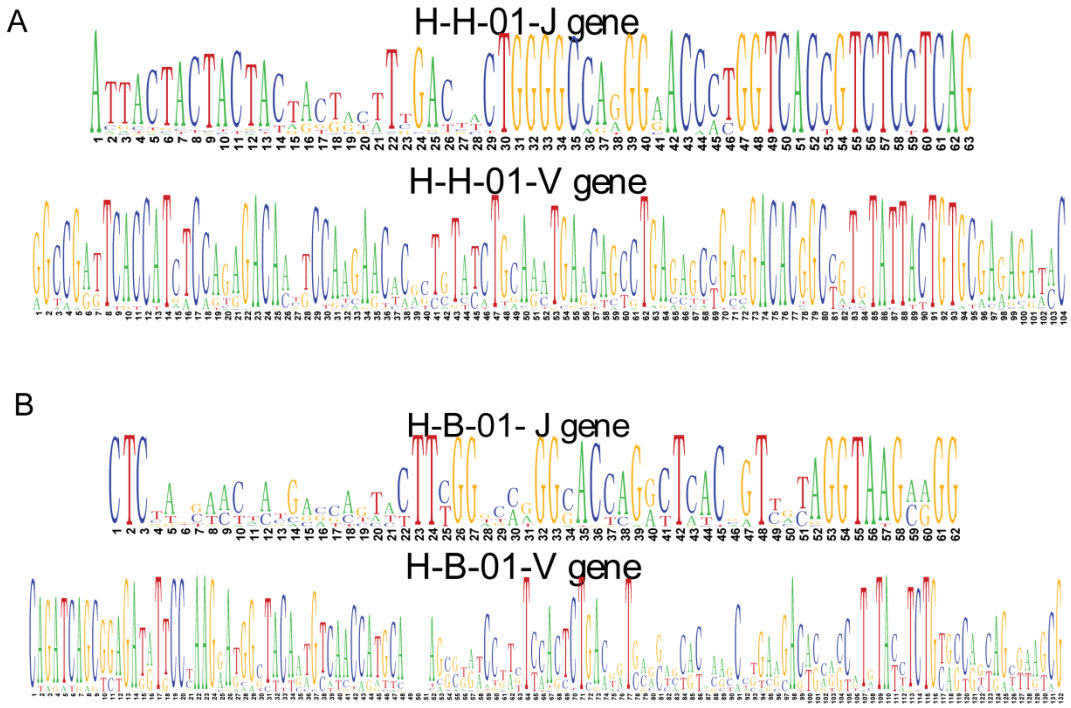

Figure S8. Nucleotide composition of V/J genes. (A) H-H-01 sample. (B) H-B-01 sample.