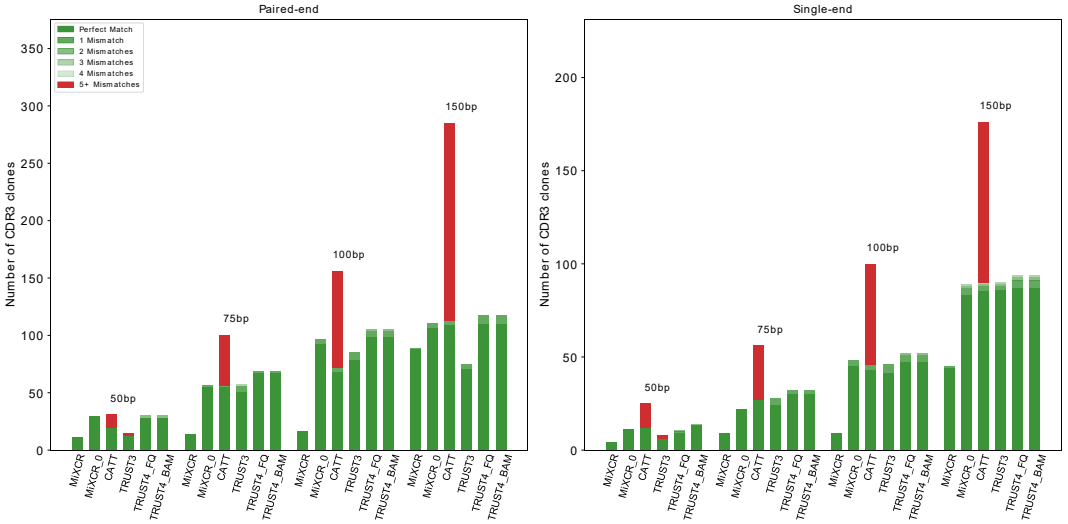
Supplementary Table 1. Running time in minutes of TRUST4, MiXCR, TRUST3 and CATT on tumor RNA-seq samples

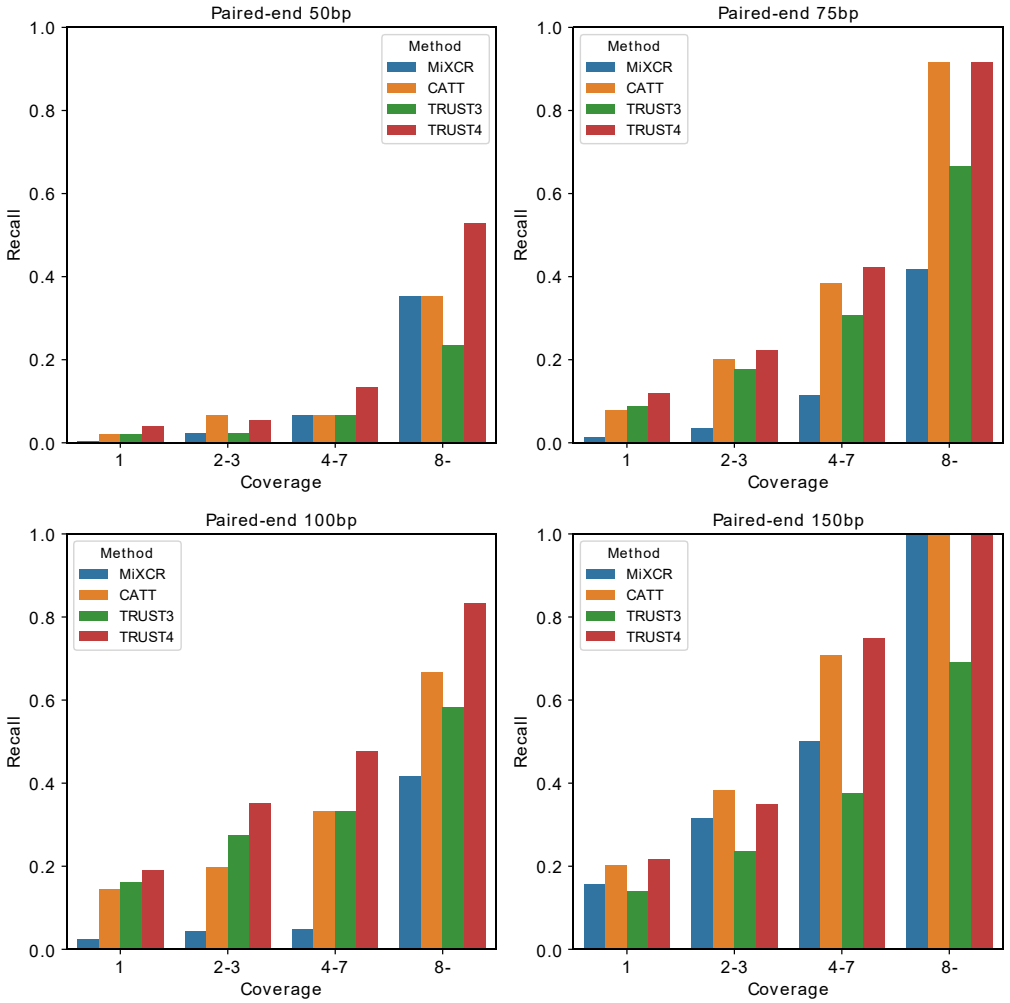
TRUST4's running time has two entries for BAM input and FASTQ input respectively. The numbers in the parenthesis were the running time on processing input FASTQ files. BAM files were generated by STAR and took 2.5 hours on average. The performance was tested with 8 threads, and the running times were displayed for both wall time (in bold) and cpu time. TRUST3 was not optimized for parallelization so wall time equals CPU time; CATT used more than 8 threads even when 8 thread was specified as the parameter. TRUST4 and MiXCR consumed less than 6GB memory, and CATT used 8GB memory on average, whereas TRUST3 required 62GB memory on average.

		FZ-20	FZ-83	FZ-94	FZ-97	FZ-116	FZ-122	Avg
Read pairs		126M	107M	85M	93M	86M	125M	104M
TRUST4 BAM	**Wall**	**44**	**64**	**46**	**35**	**28**	**25**	**40**
	CPU	103	166	110	86	69	57	98
TRUST4 FASTQ	**Wall**	**56 (25)**	**71 (26)**	**49 (20)**	**40 (20)**	**36 (19)**	**40 (27)**	**49 (23)**
	CPU	202	268	181	155	143	169	186
MiXCR	**Wall**	**227 (213)**	**235 (212)**	**175 (131)**	**181 (156)**	**149 (139)**	**240 (188)**	**201 (173)**
	CPU	1726	1810	1353	1404	1188	1893	1562
TRUST3	**Wall**	**790**	**717**	**499**	**507**	**409**	**481**	**567**
CATT	**Wall**	**1057 (1014)**	**834 (798)**	**636 (605)**	**675 (643)**	**624 (598)**	**893 (852)**	**787 (751)**
	CPU	8520	6892	5243	5545	5146	7346	6449

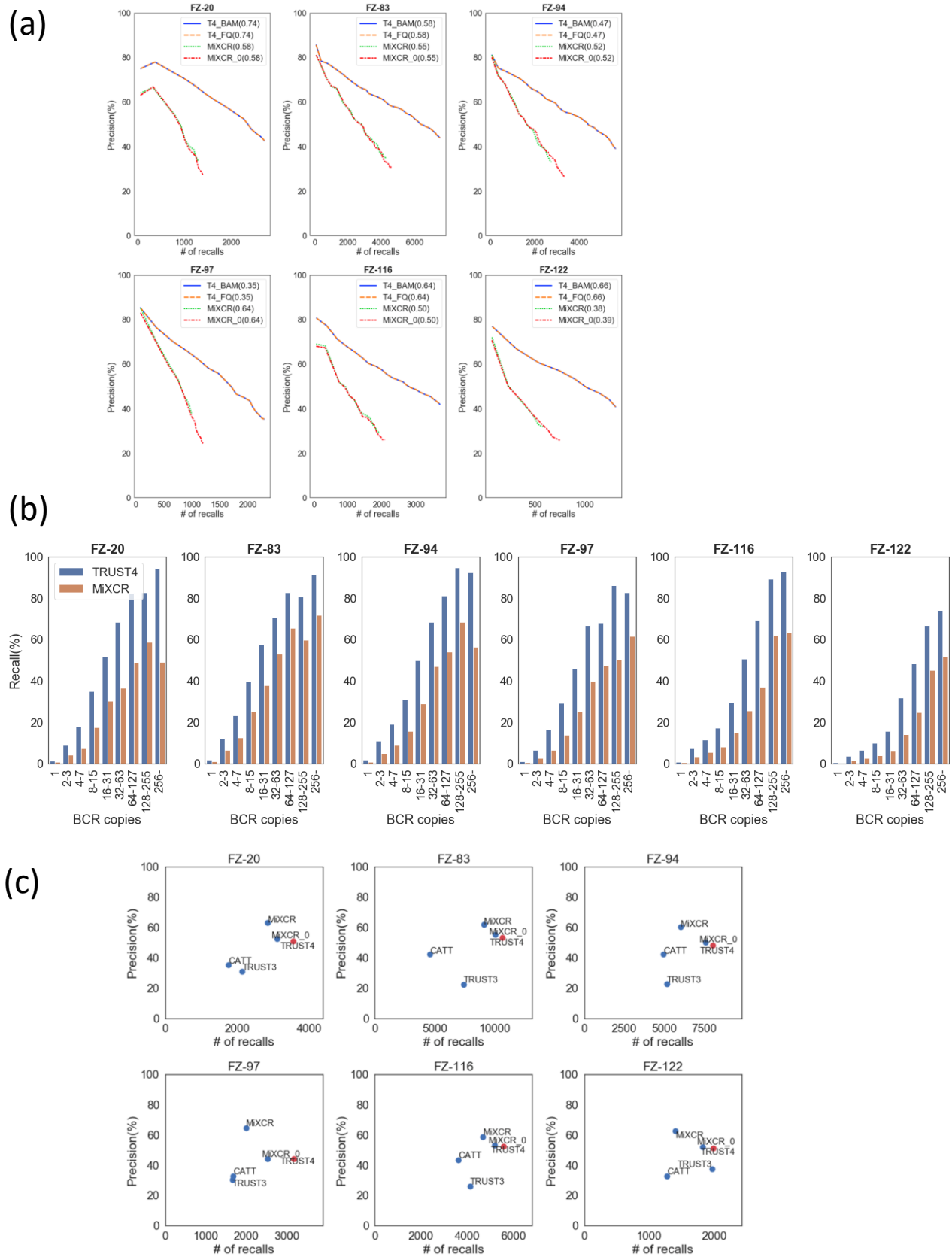
(a)



(b)



Supplementary Figure 1. Evaluation on the *in silico* RNA-seq data for TRB performance
(a) The number of recalled TRB CDR3 from the *in silico* RNA-seq data. with MiXCR, CATT, TRUST3 and TRUST4, including MiXCR with or without filter (MiXCR_0), and TRUST4 with FASTQ or BAM input.
(b) The sensitivity of TRB CDR3 categorized by the number of reads covering the TRB chain.



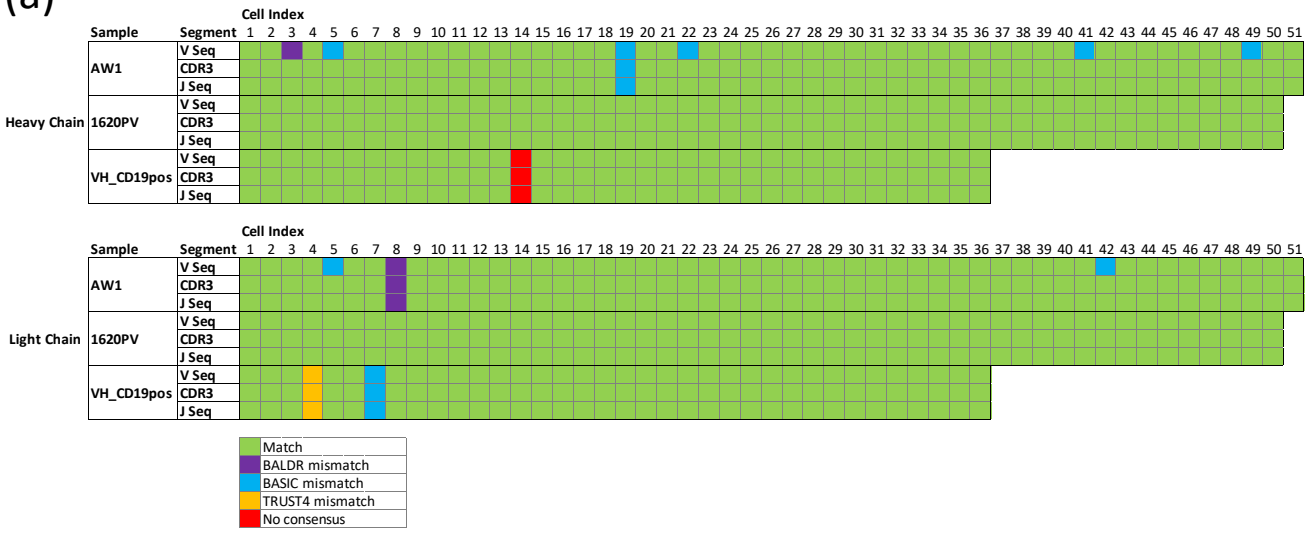
Supplementary Figure 2. Evalutation on the 6 tumor RNA-seq samples using BCR-seq data as the gold standard

(a) Precision-recall curve of IGH assemblies (V, J, C gene assignments and CDR3 sequence). The evaluation included MiXCR, MiXCR_0, TRUST4 with FASTQ and BAM. The curves were drawn by connecting the precision and recall of the top N abundant assemblies (N=100, 500, 1000,…). Numbers in the legend is the Pearson correlation of the IGH abundance between MiXCR/TRUST4 and BCR-seq. T4 is for TRUST4.

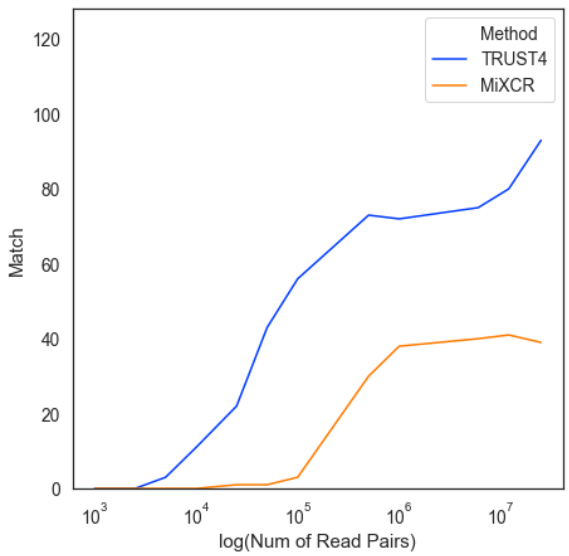
(b) The sensitivity of MiXCR and TRUST4 for IGH assemblies (V, J, C gene assignments and CDR3 sequence) in bulk RNA-seq data on different IGH abundances reported in BCR-seq data

(c) The precision and sensitivity on IGH CDR3 sequence assemblies. Red dots were TRUST4 results.

(a)



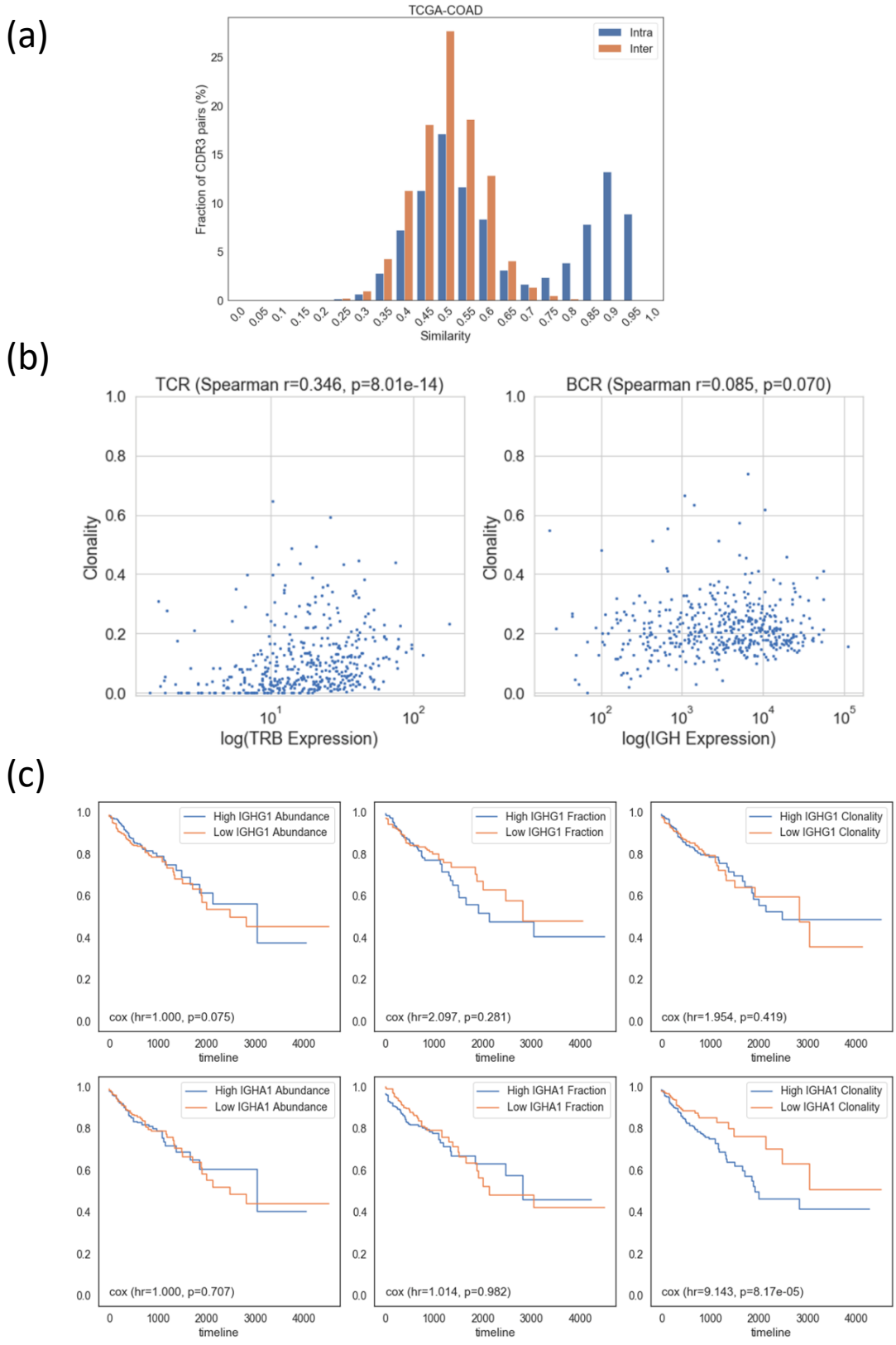
(b)



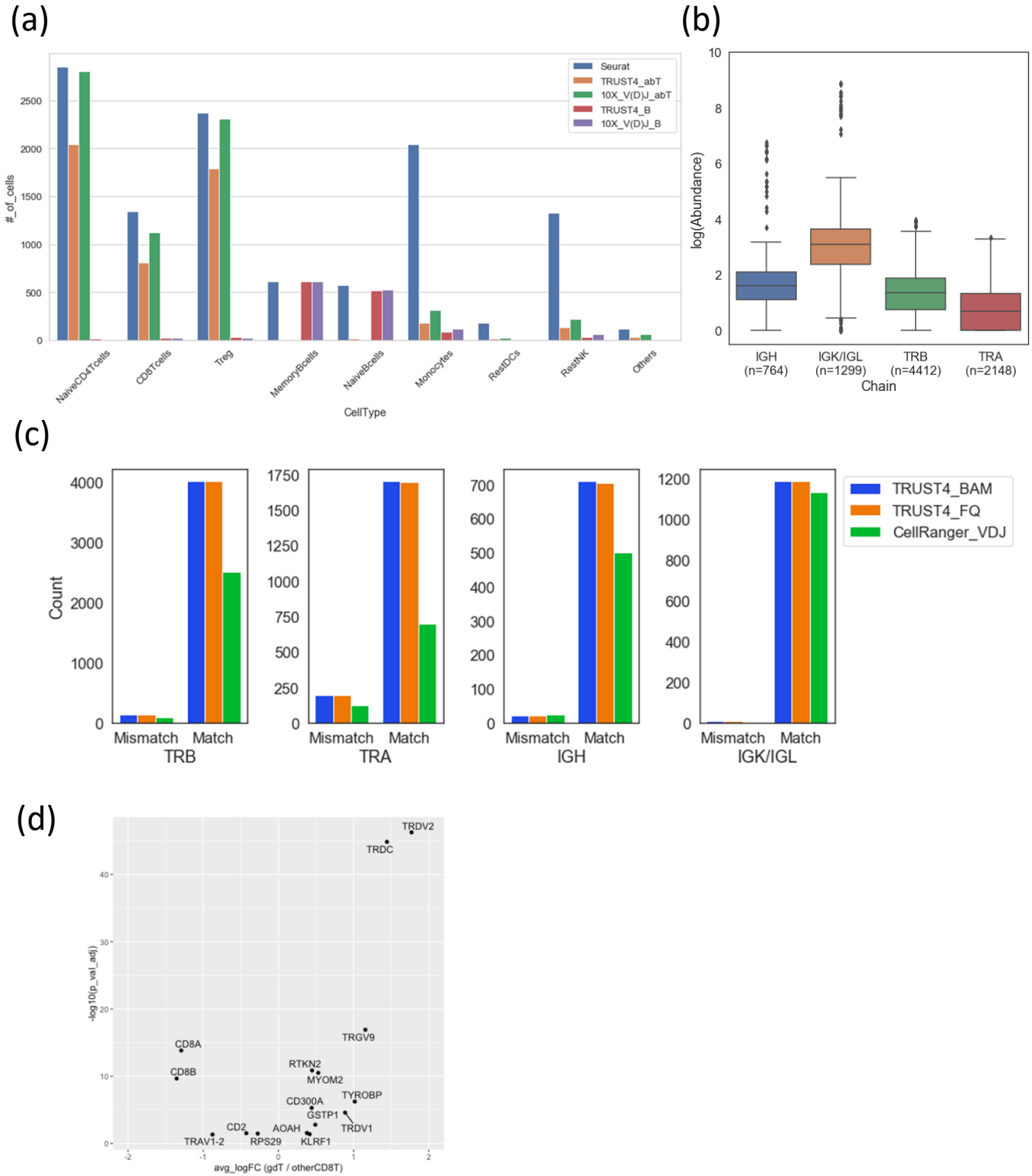
Supplementary Figure 3. Evaluation with SMART-seq data

(a) Comparison of BALDR, BASIC and TRUST4 on three B cell SMART-seq samples. AW1 contains 51 cells, 1620PV contains 50 cells and VH_CD10pos contains 36 cells. A column of three entries represents one cell from a sample. The three entries represent the results on V gene sequence, CDR3 sequence and J gene sequence respectively. The color of the entry means whether the three assemblers have identical sequences as: (1) Match: all three methods generate identical sequences. (2) Method_name mismatch: two other methods assemble identical sequences, but the indicated method does not. (3) No consensus: the three methods report different results.

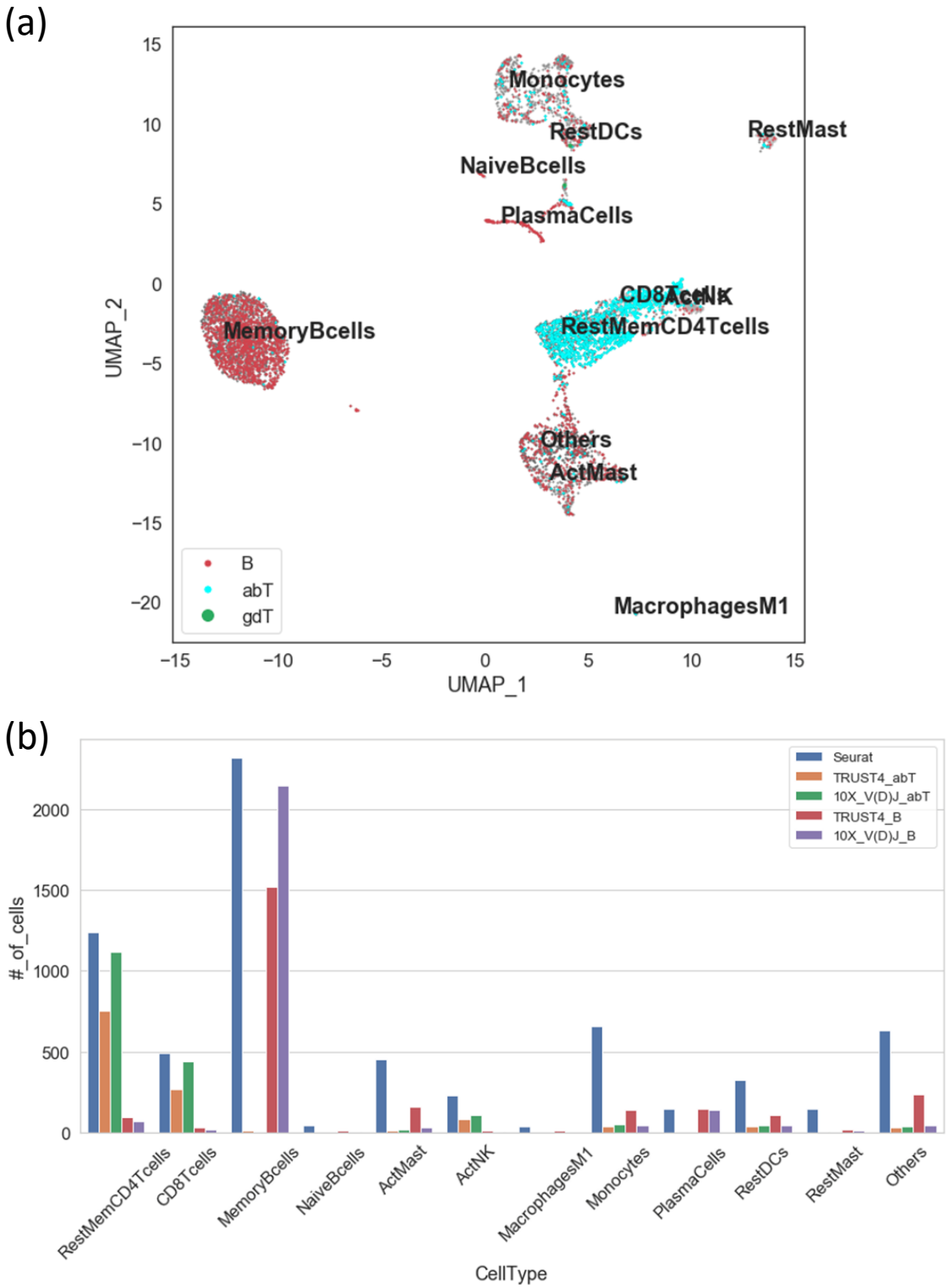
(b) The number of full-length IGH assemblies with different total number of read pairs in the pseudo-bulk data



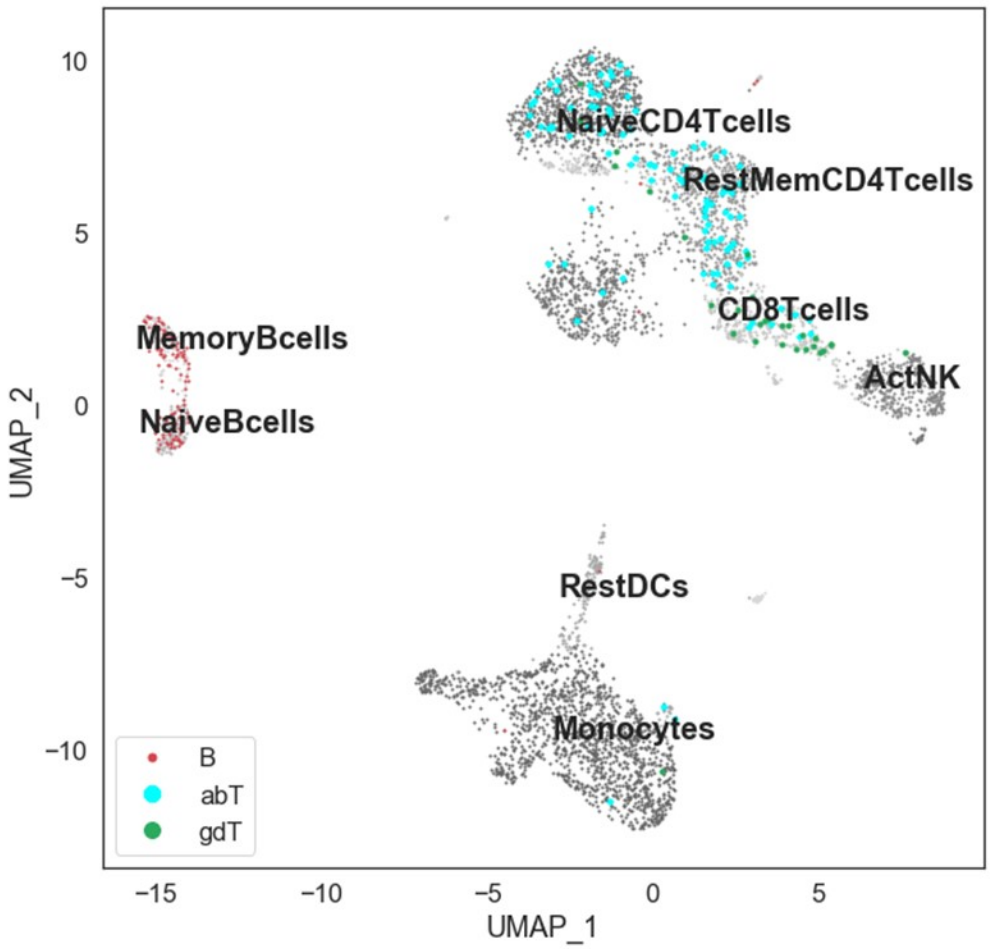
Supplementary Figure 4. Application of TRUST4 on 466 TCGA-COAD RNA-seq samples
(a) The nucleotide similarity distributions of CDR3s pairs with the same length and same V, J gene assignments. The distributions were based on the pairs between samples (inter-patient) and the pairs within the same sample (intra-patient).
(b) Spearman correlation between TRB (left) and IGH (right) gene expression (sum of constant genes' TPMs) and clonality. TRB clonality was based on CDR3 sequences and IGH clonality was based on IGH clusters. Clonality is defined as 1-normalized_Shannon_entropy.
(c) IgA1 clonality was associated with poor prognosis, whereas IgG1 clonality, IgG1 and IgA1 expression level and fraction were not. Hazard ratios (hr) and p-values were computed by Cox proportional hazards regression corrected by patient age.



Supplementary Figure 5. Evaluation on 10X Genomics 5' scRNA-seq of PBMC
(a) TRUST4 and 10X V(D)J immune repertoire calls on Seurat annotated 10X scRNA-seq of PBMC.
(b) Significantly higher read coverage of BCRs than TCRs in 10X scRNA-seq. Abundance is the number of reads found by TRUST4 that supports the CDR3. p-values: IGH vs TRB: 3e-17; IGK/IGL vs TRA: <1e-30 (two-sided Wilcoxon rank-sum test).
(c) Comparison of CellRanger_VDJ, TRUST4 with BAM and FASTQ input using 10X V(D)J data as gold standard. CellRanger_VDJ and TRUST4_FQ both took raw FASTQ data as input. TRUST4 completed in 1.5 hours with 5.5GB memory usage, while CellRanger_VDJ spent 19 hours and 13GB memory. TRUST4 was tested with 8 threads and CellRanger_VDJ was given 8 cores.
(d) Higher expression of TRDV, TRGV and TRDC and lower expression of CD8A and CD8B in the 83 gd T cells compared to all other CD8 T cells. The results were based on Seurat's function "FindMarkers".



Supplementary Figure 6. TRUST4 immune repertoire calls agree with 10X V(D)J and Seurat annotation on NSCLC 10X scRNA-seq data
(a) UMAP of the cells colored by TRUST4 called CDR3 types
(b) Comparison between TRUST4 and 10X V(D)J immune repertoire calls on Seurat annotation



Supplementary Figure 7. TRUST4 calls fewer TCRs and BCRs from 3' 10X scRNA-seq data of PBMC Due to the low recall number, the circle size of abT and gdT called by TRUST4 was enlarged.