Efficient high-throughput sequencing for quantitative immune profiling using unique molecular identifiers

Qiang Li*, Ryuji Tazawa*, Wenwen Xiang, Karthikeyan Swaminathan, Magnolia Bostick, Nidhanjali Bansal, and Andrew Farmer†

Takara Bio USA, Inc., Mountain View, CA, USA

*These authors contributed equally to this project †Corresponding author: andrew_farmer@takarabio.com

Abstract

Next-generation sequencing (NGS) for immune repertoire profiling has become a powerful tool for understanding the role of the adaptive immune system in health and disease. Additionally, unique molecular identifiers (UMIs) have become a vital aspect of this approach and are used for preserving quantitative information (i.e., accurate clonotype counts) of the repertoire by removing PCR/sequencing errors and duplicates. Here, we integrated UMIs with our SMART® technology to detect genuine low-frequency events in full-length variable regions of B-cell receptor (BCR) genes and T-cell receptor (TCR) genes. For each library, >90% of reads were on-target, and the most highly represented clonotypes remained consistent among the technical duplicates in the range of 10 ng–1 µg of input RNA or 50–10,000 cells. A sensitivity assay demonstrated RNA transcripts corresponding to multiple UMIs could be detected when spiked into input RNA at a relative concentration of 0.001%. We also developed software to analyze multiple sequence data with UMIs to generate detailed stats for reads, clonotypes, UMIs, and mapping rate as output. The updated human TCR profiling kit, SMARTer® Human TCR a/b Profiling Kit v2 (TCRv2), is designed to be compatible with any Illumina® platform using 2 x 150 bp reads. Moreover, unique dual indexes (UDIs) were incorporated to avoid crossover contamination caused by index hopping. Thus, our immune profiling technology can be used to observe clonal selection and hypermutation events in rare clonotypes found in blood and tumor tissues. These methods could also serve as a basis for the discovery of antibody-based therapeutics.

1 Improved chemistry, easy workflow

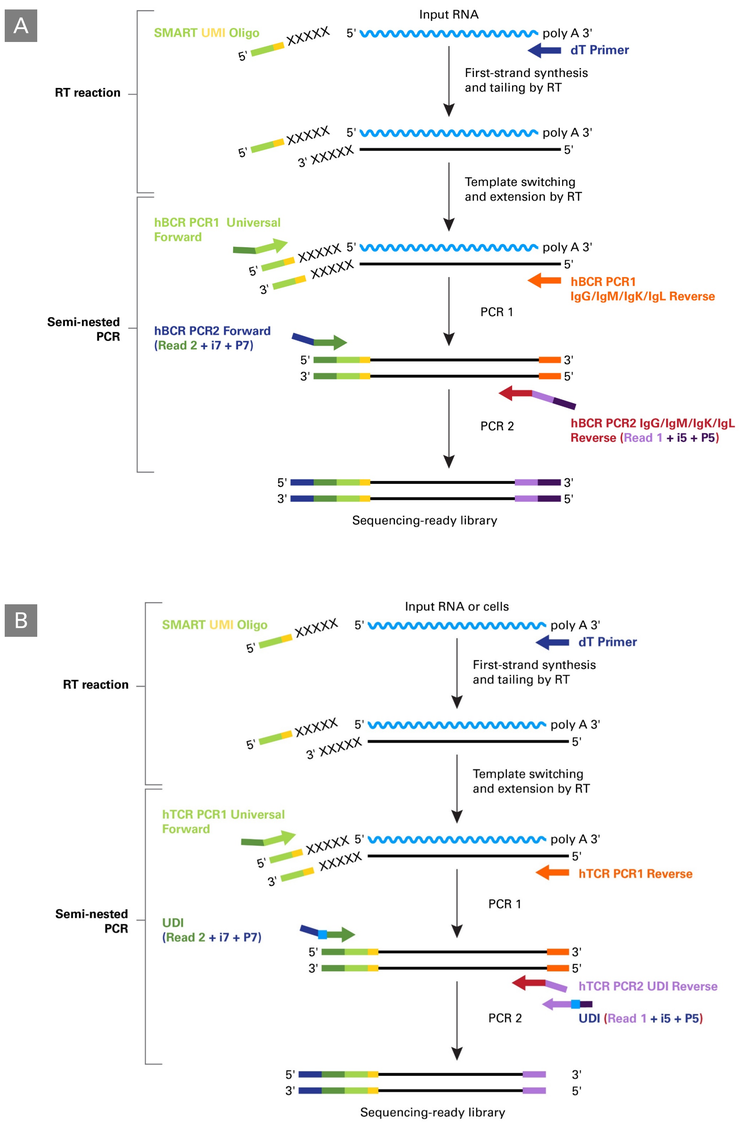


Figure 1. SMARTer immune profiling workflows. Panel A. SMARTer Human BCR IgG IgM H/K/L Profiling Kit workflow. First-strand cDNA synthesis is dT-primed and performed by the MMLV-derived SMARTScribe™ Reverse Transcriptase (RT), which adds nontemplated nucleotides upon reaching the 5' end of each mRNA template. The SMART UMI Oligo anneals to these nontemplated nucleotides and serves as a template for the incorporation of an additional sequence of nucleotides into the first-strand cDNA by the RT (this is the template-switching step). The first-strand cDNA is then subjected to two rounds of gene-specific PCR amplification. The nested PCR in the second round ensures that the vast majority of the reads map to B-cell receptor transcripts **Panel B.** SMARTer Human TCR a/b Profiling Kit v2 workflow. Similar workflow as Panel A, but the primers are specific for TCR genes.

Methods

CD19+ B-cell RNA was purchased from Miltenyi Biotech (Cat. # 130-093-169). PBMC RNA from single donors and RNA from a Jurkat cell line were purchased from Biochain (Cat. # R1234148-10), in addition to in-house RNA extracted from PBMC cells and CD3+ T cells (purity >97%) acquired from AllCells. RNA was extracted using the Macherey-Nagel NucleoSpin RNA Kit (Cat. #s 740955.50, 740955.240C, or 740955.250). B-cell carcinoma cell lines Raji, TIB190, TIB196, and CRL2974 were purchased from ATCC and cultured as suggested. RNA from each cell line was then extracted using the same Macherey-Nagel NucleoSpin RNA Kit as above. 1 ng, 100 pg, 10 pg, and 1 pg each of RNA from the B-cell carcinoma and Jurkat cell lines were spiked into 100 ng of a single donor's PBMC RNA.

All libraries containing TCR-a and TCR-b sequences were generated using the SMARTer Human TCR a/b Profiling Kit v2, as per the user manual. All libraries containing IgG and IgM heavy- and light- chain sequences were generated using the SMARTer Human BCR IgG IgM H/K/L Profiling Kit (Cat #s 634466, 634467), as per the user manual (without the multiplexing option in Appendix A). Libraries were produced using first-strand cDNA as a template in three different PCRs for heavy chain and kappa and lambda light chains. The products of these PCRs were used as templates in a set of nested PCRs, one for each chain. Following purification and size selection, libraries were validated using the Agilent 2100 Bioanalyzer. BCR and TCR libraries were spiked in with 10–30% PhiX to increase sequence diversity and sequenced on an Illumina MiSeq® platform with 600-cycle V3 cartridges (Cat. # MS-102-3003). TCR sequencing were performed with 2 x 150 bp or 2 x 300 bp paired-end sequencing with Illumina MiSeq, MiniSeq™, and NextSeq®.

Sequencing data analysis was completed by the Takara Bio Immune Profiler Software, which features MIGEC (Shugay, 2014) and MiXCR (Bolotin, 2015) software

References

Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).

Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

2 UMIs: what they are and their benefits

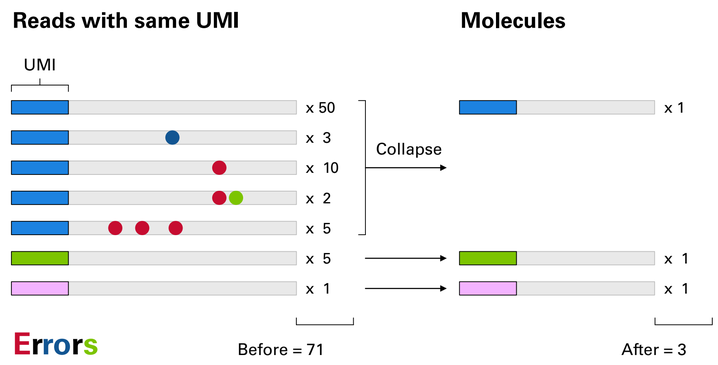


Figure 2. Diagram describing UMI-based error correction. UMIs are sequence tags used to remove PCR duplicates and errors derived from PCR/sequencing. As shown here, without UMI-based correction, the reads are not an accurate representation of the transcripts. If the diversity of UMIs are enough (typically >6 x 10^4), each gene transcript has a unique UMI tag. After the collapse with UMI, duplication and the minor errors are removed, leading to the original transcript sequence.

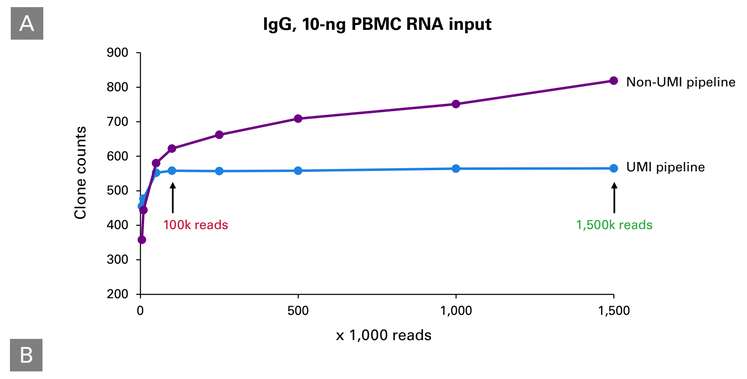


Figure 3. Evaluation of sequencing saturation by utilizing UMIs. BCR profiling libraries from 10 ng of PBMC RNA from a single donor were prepared using the SMARTer Human BCR IgG IgM H/K/L Kit. **Panel A**. IgG clonotype counts at different sequencing depths are shown with (blue line) and without (purple line) UMI-based error correction. IgG and other chains (IgM, IgK, and IgL not shown) were sequenced at 1.5 million reads per library, then downsampled to 1 million, 500k, 200k, 100k, 50k, 10k, and 5k reads. All analyses at different sequencing depths were generated with the Takara Bio Immune Profiler Software. **Panel B**. Venn diagrams showing overlapped clonotypes between libraries with low (100k; red) and high (1,500k; green) sequencing depths.

3 B-cell performance: successful detection of low-abundance clones from various donors

A

100 ng PBMC RNA + spike-in RNA	% spike-in RNA	Detected reads				Total reads
		Raji	TIB190	TIB196	CRL2974	
1 ng spike-in	1%	626	1,650	44,821	27,835	500,000
100 pg spike-in	0.1%	74	125	3,978	2,582	500,000
10 pg spike-in	0.01%	10	13	474	271	500,000
1 pg spike-in	0.001%	1	1	53	28	500,000

B

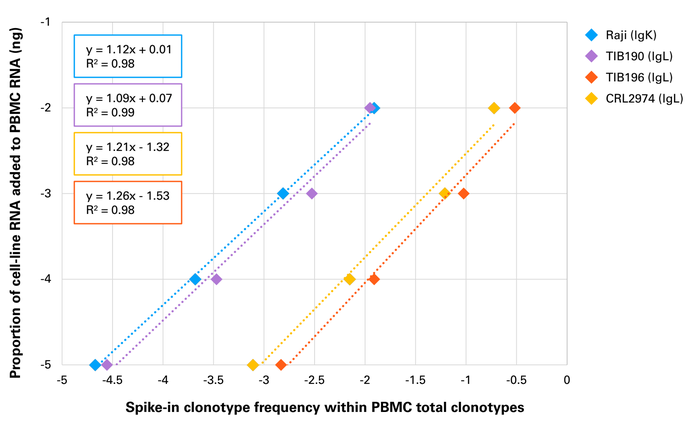


Figure 4. Successful identification of low-abundance clones. 1 ng, 100 pg, 10 pg, and 1 pg RNA of each of the four cell lines (Raji, TIB190, TIB196, and CRL2974) were spiked into 100 ng of a single donor's PBMC RNA. **Panel A.** Clone counts of each cell line at different spike-in levels are listed in the table. Libraries were normalized to 500,000 reads, and all counts were measured after UMI-based consensus collapse. **Panel B**. Calculated correlation between spike-in RNA proportions (log_{10}) and detected clonotype frequencies (log_{10}) for each cell line.

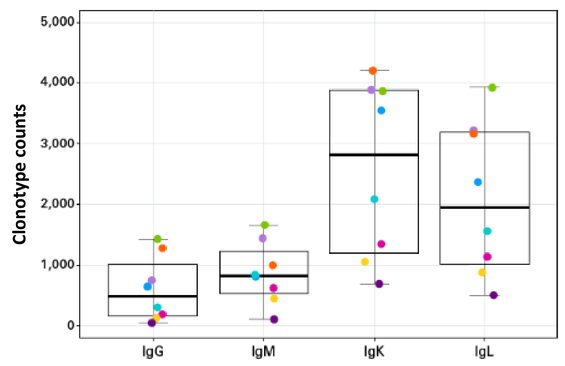


Figure 5. Clonotype counts from PBMC RNA obtained from different donors. IgG, IgM, IgK, and IgL clonotypes from 10 ng each of PBMC RNA from eight donors (represented by different colors) were determined using the SMARTer Human BCR IgG IgM H/K/L Profiling Kit. Libraries were normalized to 100,000 reads for analysis.

4 T-cell performance: increased sensitivity

% Jurkat RNA spiked in 100 ng PBMC RNA	Total reads count	Without UMI collapse			With UMI collapse		
		# of raw reads	# of reads for TRBV12-3-TRBJ1-2	Detected frequency of Jurkat reads	# of detected UMI	# of UMIs for TRBV12-3-TRBJ1-2	Detected frequency of Jurkat UMIs
10.0000%	2,500,000	1,565,005	397,179	2.5E-01	281,280	62,629	2.2E-01
1.0000%	2,500,000	1,422,102	47,160	3.3E-02	219,776	6,426	2.9E-02
0.1000%	2,500,000	1,366,127	5,412	4.0E-03	189,580	631	3.3E-03
0.0100%	2,500,000	1,218,025	521	4.3E-04	196,615	74	3.8E-04
0.0010%	2,500,000	1,331,465	909	6.8E-04	197,870	6	3.0E-05
0.0001%	2,500,000	1,409,199	-	0%	124,149	-	0%
0.0000%	2,500,000	1,222,245	-	0%	197,933	-	0%

Table 1. Assessing the sensitivity and reproducibility of the SMARTer approach. Spike-in analysis was performed in replicate on PBMC RNA samples spiked at varying concentrations (10%, 1%, 0.1%, 0.01%, 0.001%, 0.0001%) with RNA obtained from a homogenous population of Leukemic Jurkat T cells (TRBV12-3-TRBJ1-2 clonotypes). TCRβ CDR3 regions were amplified using the SMARTer Human TCR a/b Profiling Kit v2 and sequenced from 100 ng of total RNA . Reads of 2 x 150 bp were obtained using the Illumina NextSeq System. The sequencing reads were downsampled to 2.5M reads. Read results for spike-in concentrations identified as the reliable concentration limit for each criteria (without and with UMI collapse) have data highlighted in gray. With the UMI collapse, you can confidently get down to a level of 0.001% Jurkat RNA spike-in.

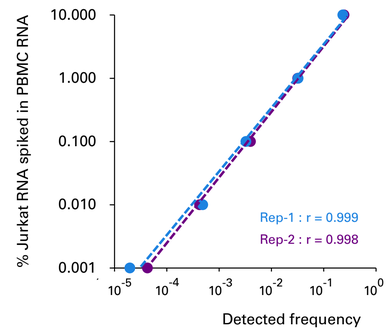


Figure 6. Linear regression results. Analysis of sequence reads of TRB library generated by 100 ng of PBMC revealed a linear correlation between the amount of spiked-in Jurkat RNA and the UMIs for Jurkat transcripts (r = 0.99, TRBV12-3-TRBJ1-2). This result demonstrates that differences in the relative abundance of transcripts for a particular TCR clonotype are faithfully and reproducibly represented in sequencing libraries generated using the SMARTer and UMI approach. Comparison of the number of TRBV12-3-TRBJ1-2-specific sequence reads obtained for the control vs. spike-in samples suggests that added Jurkat RNA at a concentration of 0.001% is detectable with UMI analysis at a depth of ~2,500,000 reads, evidence of the sensitivity afforded by the SMARTer approach. The TCRv2 kit can accommodate detection of rare TCR clones.

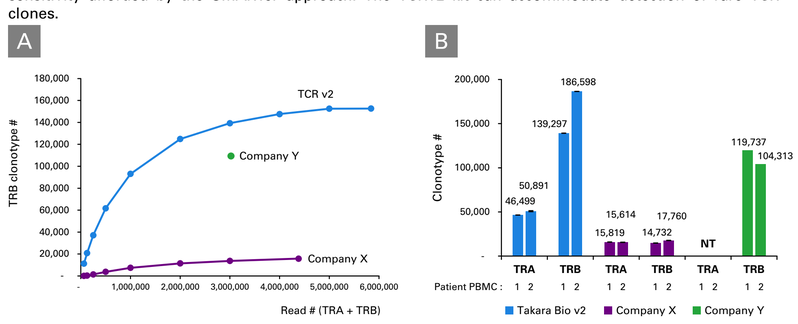
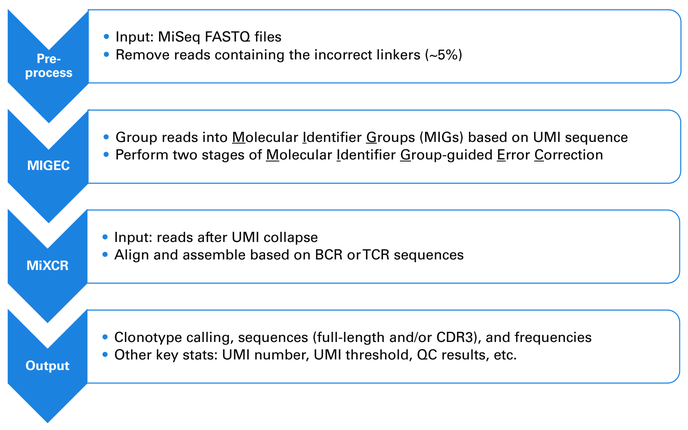


Figure 7. Superior sensitivity and reproducibility with TCRv2. We were able to demonstrate that Takara Bio TCRv2 generates data with superior sensitivity and reproducibility than competitors. We split 5M CD3+ T cells from 2 different healthy donors for RNA and gDNA extraction. 1.6 µg of gDNA was used for library preparation according to manufacture's instructions (15% of total amount of extracted gDNA). 100 ng of RNA were used for library preparation (2% of total amount of extracted RNA). **Panel A.** We observed a dramatically higher clonotype number for TRB clonotypes after downsampling with the TCRv2 kit (TRA results were similar, but not shown). **Panel B.** Clonotype numbers for TCR a/b were shown from each company. (**NT**: not tested). In the comparison, TCRv2 generated 48.7k and 163k clonotypes for TCR a/b respectively, representing a 290% increase against Company X and 145% increase against a gDNA-based approach (Company Y). Importantly, RNA methods used only 2% of total RNA from 5M PBMCs. These results demonstrate Takara Bio RNA methods have greater sensitivity and reproducibility than gDNA methods

5 Easy-to-use analysis tool: Takara Bio Immune Profiler Software



Takara Bio Immune Profiler Software is a bioinformatic tool designed to analyze sequence data stored in FASTQ files generated from SMARTer Human BCR IgG IgM H/K/L Profiling Kit. The Profiler software incorporates two third-party software packages: MIGEC and MiXCR. Output includes UMI number, UMI threshold counts, QC results, and sequences.

Conclusions

- Takara Bio immune profiling kits include unique molecular indexes (UMIs) to correct for PCR duplicates and errors which allow for greater confidence in the results
- Both immune kits provide high reproducibility and are highly sensitive, detecting low-abundance clones
- The TCR kit has the flexibility to be sequenced on all Illumina instruments
- The compatible software provides a user-friendly way to analyze UMI and clonotypes

Takara Bio USA, Inc.
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999

800.662.2566
Visit us at takarabio.com

Clontech TaKaRa cellartis