CORRESPONDENCE

Evaluation of immune repertoire inference methods from RNA-seq data

To the Editor: Characterizing tumor-infiltrating T cell receptor (TCR) repertoire is a critical step toward identifying cancer antigens and developing new immunotherapies[1]. We previously developed a computational algorithm named TRUST[2,3] to extract TCR hypervariable complementarity-determining region 3 (CDR3) sequences from unselected bulk tumor RNA sequencing (RNA-seq) data. When applied to large cancer cohorts, TRUST found associations between tumor mutation load and TCR repertoires[2]. A recent study by Bolotin et al.[4] reported a new version of their MiXCR tool that enables assembly of TCR clonotypes from RNA-seq data. In comparing MiXCR to TRUST, Bolotin et al.[4] concluded that MiXCR was superior and that the output of TRUST includes unconfirmed and potentially false positive results. Here, we point out important differences between TRUST and MiXCR, as well as differences in interpretation of results, which may complicate direct comparison of the tools.

TRUST uses TCR variable (V) and joining (J) gene motifs to search and annotate CDR3-containing reads and performs de novo assembly on the CDR3-overlapping reads. The output of TRUST v2.1 contains reads with V or J gene motifs and partial CDR3 sequences (**Supplementary Fig. 1**), which were considered "non-canonical unconfirmed" by Bolotin et al.[4]. We recognize that partial CDR3 sequences or reads that extend beyond the accepted limits of CDR3 cannot be unambiguously counted as unique clonotypes. However, we would note that single-chain CDR3 from bulk RNA-seq may also not be ideal to identify a unique clonotype because, for a strict definition of clonotype, generally both chains would be required. We also point out that partial CDR3 sequences of reasonable length (6–30 amino acids) and perfect match to a subregion of the respective CDR3 molecule are informative for modeling TCR binding specificity. This is because structural studies indicate that only a small region in the complete CDR3 makes contact with the antigen peptide[5,6], and the recent GLIPH (grouping of lymphocyte interactions by paratope hotspots)[5] method can cluster TCRs with

likely shared specificity from enriched local motifs within many distinct CDR3 molecules. Therefore, partial CDR3 sequences, such as those contained in the output of TRUST v2.1, may be valuable when seeking to gain insights into the frequency of shared specificities.

TRUST[3] v2.1 is open source and compatible with TopHat[7], MapSplice[8] and STAR[6] mapping to Human Genome Reference 37 (GRCh37/hg19), with some conditions. We noted these conditions, such as disabling local alignment, in the Supplementary Notes and software ReadMe associated with our paper[3], although we did not specify the full STAR command line. Bolotin et al.[4] incorrectly stated that TRUST v2.1 was not open source, that it requires raw reads to be aligned using TopHat, and that TRUST v2.1 on STAR alignments did not produce results. We are able to run TRUST v2.1 on STAR-aligned RNA-seq data and obtain results using the parameters in **Supplementary Note 1**, and suspect that Bolotin et al.[4] either used The Cancer Genome Atlas (TCGA) data mapping to Human Genome Reference 38 (GHCh38/hg38) or failed to disable local alignment.

We note that, when using simulated data, Bolotin et al.[4] identified "false CDR3 sequences" in TRUST v2.1 outputs by examining CDR3 calls on negative control sequences generated from random hg38 transcripts. Theoretically, TRUST and MiXCR should not make CDR3 calls from negative control sequences, so any calls were treated as false positives. However, because some V/J genes were annotated in hg38 but were not annotated to hg19 on which TRUST v2.1 relies, TRUST v2.1 would naturally assemble some CDR3 sequences from these unmappable reads. The correct approach would have been to generate negative control random transcripts using hg19 genome annotation, which indeed did not yield any CDR3 calls from TRUST v2.1 (**Supplementary Note 2**).

Since publication of TRUST v2.1 in April 2017, we have continued to develop and maintain TRUST. Recent updates include B cell receptor CDR3 calling functions, added compatibility with hg38 reference genome, and a postprocessing module for easier downstream analyses. We

also improved the computational efficiency of TRUST through multithread processing and simple instruction multiple data (SIMD) accelerations. We respect the MiXCR developers' continued development and maintenance of their valuable tool, and hope fair and collegial competitions between the algorithms will ultimately benefit the scientific user community.

Code availability. TRUST is available at https://bitbucket.org/liulab/trust. Code to run simulations and process TRUST outputs is available at https://bitbucket.org/liulab/trust/src/nbt-response.

Editor's note: This article has been peer-reviewed.

AUTHOR CONTRIBUTIONS
B.L. designed the TRUST algorithm. X.H. organized the codebase. J.Z. performed in silico simulations and generated figures and tables. X.H., B.L., J.S.L. and X.S.L. wrote the manuscript.

COMPETING INTERESTS
The authors declare no competing interests.

Xihao Hu[1,5], Jian Zhang[2,5], Jun S Liu[3], Bo Li[4,6] & Xiaole Shirley Liu[1,3,6]

[1]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.* [2]*Beijing Institute of Basic Medical Sciences, Beijing, China.* [3]*Statistics Department, Harvard University, Boston, Massachusetts, USA.* [4]*Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas, USA.* [5]*These authors contributed equally to this work.* [6]*These authors jointly supervised this work.*
e-mail: xsliu.dfci@gmail.com

1. Pardoll, D.M. *Nat. Rev. Cancer* **12**, 252–264 (2012).
2. Li, B. *et al. Nat. Genet.* **48**, 725–732 (2016).
3. Li, B. *et al. Nat. Genet.* **49**, 482–483 (2017).
4. Bolotin, D.A. *et al. Nat. Biotechnol.* **35**, 908–911 (2017).
5. Glanville, J. *et al. Nature* **547**, 94–98 (2017).
6. Dash, P. *et al. Nature* **547**, 89–93 (2017).
7. Kim, D. *et al. Genome Biol.* **14**, R36 (2013).
8. Wang, K. *et al. Nucleic Acids Res.* **38**, e178 (2010).