Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data

To the Editor:

Neoantigen-specific tumor-infiltrating T lymphocytes are immune effector cells for cancer elimination and are the primary focus of current cancer immunotherapies[1–4]. We previously published a novel method to assemble T cell receptor (TCR) complementarity-determining region 3 (CDR3) sequences using paired-end tumor RNA–seq data[5]. Extending this approach, we have developed 'TCR repertoire utilities for solid tissue', or TRUST, for ultrasensitive detection of tumor-infiltrating T cell CDR3 sequences (**Supplementary Software**). TRUST significantly outperforms our previous method, with a substantial increase in recall (**Supplementary Fig. 1a**), especially for libraries with deeper coverage and longer read length (**Supplementary Fig. 1b**). In addition to exhibiting improved performance, TRUST can also handle single-end RNA–seq data and has demonstrated utility for non-cancerous tissues.

TRUST takes single-end or paired-end library reads mapped to the human reference genome in BAM format as the standard input. It automatically detects input library type, selects informative unmapped reads, assigns reads

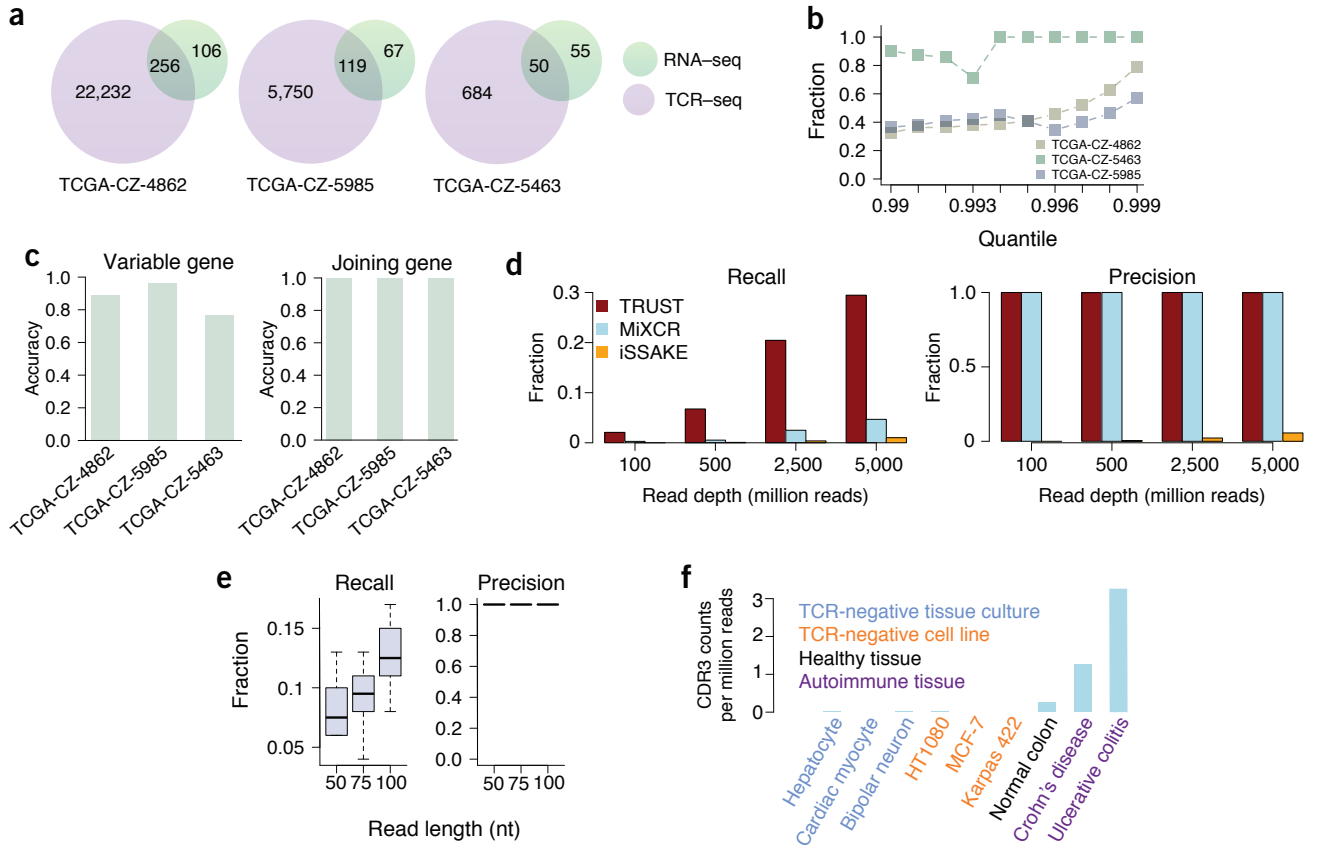

Figure 1 Evaluation of the performance of TRUST in single-end mode. (**a**) Venn diagrams showing the number of CDR3 sequences called using TCR–seq and TRUST, and their overlap. (**b**) TRUST-reported CDR3 sequences are enriched for clonotypes with high abundance. At each quantile, the *y* axis shows the fraction of TRUST-reported CDR3 sequences with a clonal frequency greater than or equal to that for the quantile. (**c**) Accuracy of variable and joining gene estimations by TRUST. (**d**) Recall and precision estimations based on *in silico* simulations at different read depths. (**e**) Recall and precision estimations at different read length settings. Each box includes data between the 25th and 75th percentiles, with the horizontal line representing the median. The upper whisker is min(max(*x*), Q3 + 1.5 × IQR) and the lower whisker is max(min(*x*), Q1 − 1.5 × IQR), where *x* is the data, Q3 is the 75th percentile, Q1 is the 25th percentile and IQR = Q3 − Q1, the interquartile range. (**f**) Application of TRUST to non-cancerous tissue samples.

into TCR genes on the basis of putative motifs, assembles reads into contigs and annotates the assembled CDR3 sequences with International Immunogenetics Information System (IMGT)[6] nomenclatures (**Supplementary Fig. 2** and **Supplementary Note**). To test whether TRUST assembles real CDR3 sequences from single-end libraries, we applied it to three formalin-fixed, paraffin-embedded (FFPE) kidney renal cell carcinoma samples from The Cancer Genome Atlas (TCGA) with both RNA–seq and TCRβ sequencing available[5] (**Supplementary Note**). A median of 64% of the CDR3 calls by TRUST could be confirmed in the TCR–seq data (**Fig. 1a**). We did not expect complete overlap because TCR–seq can only recover 25% to 50% of infiltrating T cells from FFPE samples, owing to DNA fragmentation. TRUST identified a median of 36% of the top 1% most abundant CDR3s from TCR–seq (**Fig. 1b**). Variable (V) and joining (J) segment assignments by TRUST were also highly concordant (median 89% for V and 100% for J segments) with TCR–seq calls (**Fig. 1c**). Similar performance was achieved when TRUST was applied in paired-end mode (**Supplementary Fig. 3a**). Importantly, in comparison to the prototype[5], TRUST recovered a higher percentage of the most abundant CDR3 sequences (**Supplementary Fig. 3b**).

We used *in silico* simulations (**Supplementary Fig. 4** and **Supplementary Note**) with artificially generated TCR transcripts to evaluate TRUST and competing methods[7–9]. With 50-nt single-end reads, at a read depth of 100 million (equivalent to 0.02× coverage[5]), TRUST achieved an average recall of 2.1%, an order of magnitude higher than that for MiXCR (0.12%) or iSSAKE (0%) (**Fig. 1d**). Decombinator failed to assemble any contig, even at a read depth of 5,000 million. Fixing read depth at 500 million, we simulated another set of libraries with read lengths of 50, 75 and 100 nt (**Supplementary Note**). TRUST recall increased with longer reads while high precision was maintained (**Fig. 1e**). We next collected RNA–seq data from six TCR-negative cell lines and three colon tissues from the public domain (**Supplementary Note**) to explore the utility of TRUST on non-cancerous tissues. As expected, T cell content was barely detectable in the cell lines and was higher in tissues from Crohn's disease or ulcerative colitis than in normal colon (**Fig. 1f**).

TRUST is by far the most sensitive method thus far for detecting TCR CDR3 sequences using tumor RNA–seq data. Its improved performance in comparison to our previous algorithm[5] results from optimized CDR3 realignment and use of unmapped reads. The major reason that TRUST outperforms other methods is its application of a thorough pairwise read comparison, which substantially improves the identification of less abundant TCR clones. TRUST is portable and easy to adopt and run. With rapidly accumulating tumor RNA–seq data and continuously decreasing sequencing costs, we anticipate that TRUST will attract broader interest in the immunology and cancer research communities.

Code and data availability. TRUST source code, supporting data and usage are available as **Supplementary Software**, as well as at https://bitbucket.org/liulab/trust/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We acknowledge the following funding sources for supporting our work: NCI grant 1U01 CA180980 and National Natural Science Foundation of China grants 31329003 (to X.S.L.), 31601077 (to R.D.) and 81321002 (to T.L.).

AUTHOR CONTRIBUTIONS
B.L. conceived the project, developed the method and wrote the manuscript. T.L., B.W. and R.D. contributed to data analysis. J.Z. modified TRUST to increase its computational efficiency. J.S.L. and X.S.L. supervised the study and wrote the manuscript with B.L.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

***Bo Li*[1,2]*, Taiwen Li*[3]*, Binbin Wang*[4]*, Ruoxu Dou*[5]*, Jian Zhang*[1]*, Jun S Liu*[2] *& X Shirley Liu*[1,2,4]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA. [3]State Key Laboratory of Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China. [4]School of Life Science and Technology, Tongji University, Shanghai, China. [5]Department of Colorectal Surgery, Sixth Affiliated Hospital, Sun Yat-sen University, Guangdong, China.
e-mail: bli@jimmy.harvard.edu or xsliu@jimmy.harvard.edu

1. Fridman, W.H., Pages, F., Sautes-Fridman, C. & Galon, J. *Nat. Rev. Cancer* **12**, 298–306 (2012).
2. Gajewski, T.F., Schreiber, H. & Fu, Y.X. *Nat. Immunol.* **14**, 1014–1022 (2013).
3. Matsushita, H. *et al. Nature* **482**, 400–404 (2012).
4. Snyder, A. *et al. N. Engl. J. Med.* **371**, 2189–2199 (2014).
5. Li, B. *et al. Nat. Genet.* **48**, 725–732 (2016).
6. Lefranc, M.P. *Cold Spring Harb. Protoc.* **2011**, 595–603 (2011).
7. Warren, R.L., Nelson, B.H. & Holt, R.A. *Bioinformatics* **25**, 458–464 (2009).
8. Bolotin, D.A. *et al. Nat. Methods* **12**, 380–381 (2015).
9. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. & Chain, B. *Bioinformatics* **29**, 542–550 (2013).

Celebrating parasites

To the Editor:

In an editorial published last year[1], Dan Longo and Jeffrey Drazen introduced us to 'research parasites'. These individuals "had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited" (ref. 1). The editorial sparked discussion about the role of secondary data analysis in the scientific process, both in official letters to the editor and informal commentary online. In light of the term's widespread publicity, we chose to use it to honor individuals who practice the craft of data reanalysis for novel ends.

At the Pacific Symposium on Biocomputing (PSB) 2017, we presented the inaugural Research Parasite Awards to researchers selected for their rigorous analysis of publicly accessible data. We specifically sought to honor those whose work extended, replicated or disproved what the original investigators had posited who were not involved in the experimental design or data generation, published independently of the original investigators while appropriately crediting them, and provided their own research products—including source code and intermediate or final results—in a manner that enhanced reproducibility.

We opened a call for nominations and applications in April 2016 and received 41 completed applications. From these, we selected an exemplar of Junior Research Parasitism and a Sustained Parasite. The Junior Parasite Award highlighted work performed as a trainee, while the Sustained Parasite Award required contributions over at least five years of independent research.

The inaugural Junior Parasite Award recipient was Kun-Hsing Yu of Stanford University