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

for a recent study that employed public omics and imaging data sets. Yu developed an approach to improve prediction of patient survival[2]. He and his coauthors reanalyzed histopathology images from The Cancer Genome Atlas (TCGA) and extracted features using CellProfiler. They then employed different machine learning approaches using packages from the R programming language. Yu and coauthors provided source code and data under an open license, enabling future parasites to build on this work[2].

The inaugural recipient of the Sustained Parasite Award was Erick Turner from Oregon Health & Science University. Turner's research revealed pervasive selective reporting of clinical studies and showed how this skewed the broader biomedical literature. For example, more than 90% of publications on antidepressant trials reported a positive outcome; however, the US Food and Drug Administration (FDA)-registered studies underlying that body of literature only had positive outcomes 51% of the time[3]. Trials with negative or ambiguous results were spun as having positive results and published at a lower rate than trials with positive findings. Turner continued to identify reporting biases for multiple drug classes[4].

These two awardees, in different ways, captured the essence of research parasitism. Their work enhanced understanding of the world and promoted rigorous biomedical research.

The act of rigorous secondary data analysis is critical for maintaining the accuracy and efficiency of scientific discovery. As scientists, we make predictions, perform experiments and generate data to test those predictions. When we ask rigorous questions, we obtain more accurate findings that can prevent harm. For example, Vioxx was evaluated for use in treating pain associated with rheumatoid arthritis[5]. Questions were raised shortly thereafter about its cardiovascular effects[6]. Independent researchers, using data from multiple studies, identified a drug-associated increase in cardiovascular event risk[7]. These research parasites identified important side-effects of this drug, correcting incomplete information on the drug's safety profile.

Journals, although they conduct peer review, do not validate each experimental result or claim. Research parasites fill this gap. In the case of the initial VIGOR study[5] published in *The New England Journal of Medicine*, Jeffrey Drazen, the editor-in-chief of the journal, stated: "We can't be in the business of policing every bit of data that we put out. We think that that's the role of people who know the field" (ref. 6). Research parasites help to maintain the self-correcting nature of scientific inquiry. Scientists who perform rigorous parasitism put scientific work to the test, and their results may support or challenge what we think we know.

Parasites also improve efficiency: many data sets were originally designed for specific questions, but these data may also answer distinct but related questions. Investigators can refocus data sets via meta-analysis to reveal general patterns that become apparent only with many studies. Data sets can also be individually useful. New researchers can often bring their own creative ideas to existing data, leading to novel breakthroughs and disruptive innovations.

The robust culture of research parasitism is not limited to clinical or biomedical research. A team of eagle-eyed parasites recently identified inconsistencies in a study of food consumption at a buffet restaurant[8]. Others raised issues with the genetic structure of polar bear populations[9] and social welfare spending[10]. This process by which scientists selectively retest assumptions and previous findings forms the foundation of scientific discovery in every discipline.

Under some proposals for data reuse, data would be shared with researchers working in concert with the investigators who initially analyzed the data[1]. We expect that this would counteract the recent focus of the US National Institutes of Health (NIH) on rigor, transparency and reproducibility. Any procedure that includes data generators as gatekeepers has the potential to compromise rigor and robustness. As gatekeepers, researchers could withhold data from those with contrary views or a reputation of challenging the status quo. We must expect data sharing to lead to some conclusions being challenged and, ultimately, refuted. If this fails to occur, it indicates a problem with the process and not the correctness of conclusions.

Good research parasites also have an obligation to credit their research hosts. We do not feel that parasites individually have an obligation to benefit those from whom they gather data (that is, in research symbiosis). However, we do feel that parasites should advocate strongly for recognition and support for those who generate and share valuable data without restricting their use.

We enjoyed the work that research parasites shared with us this year and expect research parasites to continue to separate real findings from erroneous conclusions. To celebrate such contributions, we are announcing an open call for applications for the 2018 Research Parasite Awards. Winners receive partial travel support to PSB, a prize and recognition on the award website. Applications are due by 5 p.m. Hawaii Standard Time on 30 September 2017. For full instructions, visit the website at http://research-parasite.com/. We have a positive and inclusive view of parasitism and particularly encourage individuals from groups that are under-represented in their discipline to apply.

ACKNOWLEDGMENTS

Numerous individuals and organisms contributed time and money toward the 2017 Research Parasite Awards. We would like to thank the PSB organizers, who helped us hone our selection criteria and award committee, and *GigaScience*, the Gordon and Betty Moore Foundation, *Nature Genetics* and *Scientific Data*, which provided financial and travel support for the awards. The authors were supported in part by grants from the Gordon and Betty Moore Foundation (GBMF 4552 to C.S.G.) and the US National Institutes of Health (R01LM012373 and K01ES025434 to L.X.G. and R01LM008111 to L.E.H.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Casey S Greene[1], Lana X Garmire[2], Jack A Gilbert[3], Marylyn D Ritchie[4] & Lawrence E Hunter[5]

[1]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Cancer Epidemiology Program, University of Hawaii Cancer Center, University of Hawaii, Honolulu, Hawaii, USA. [3]Department of Surgery, University of Chicago School of Medicine, Chicago, Illinois, USA. [4]Biomedical and Translational Informatics Program, Geisinger Health System, Danville, Pennsylvania, USA. [5]Department of Pharmacology, University of Colorado School of Medicine, Aurora, Colorado, USA.

e-mail: csgreene@upenn.edu

1. Longo, D.L. & Drazen, J.M. *N. Engl. J. Med.* **374**, 276–277 (2016).
2. Yu, K.-H. *et al. Nat. Commun.* **7**, 12474 (2016).
3. Turner, E.H., Matthews, A.M., Linardatos, E., Tell, R.A. & Rosenthal, R. *N. Engl. J. Med.* **358**, 252–260 (2008).
4. Roest, A.M. *et al. JAMA Psychiatry* **72**, 500 (2015).
5. Bombardier, C. *et al. N. Engl. J. Med.* **343**, 1520–1528 (2000).
6. Armstrong, D. *Wall Street Journal* (15 May 2006).
7. Mukherjee, D. *et al. J. Am. Med. Assoc.* **286**, 954 (2001).
8. van der Zee, T., Anaya, J. & Brown, N.J.L. *PeerJ Preprints* http://dx.doi.org/10.7287/peerj.preprints.2748v1 (2017).
9. Malenfant, R.M., Davis, C.S., Cullingham, C.I. & Coltman, D.W. *PLoS One* **11**, e0148967 (2016).
10. Breznau, N. *Sociol. Sci.* **2**, 440–441 (2015).