**Bolotin *et al.* reply:** Clinical and biological studies are often aimed at discovering a small fraction of T cell receptor (TCR) or immunoglobulin variants shared between samples or groups of samples—for example, when searching for the common clonotypes involved in response to tumor antigens, infectious agent antigens or self antigens[1]. In this context, even a single false intersection between immune repertoires extracted from different samples may compromise the utility of the results. Among the scientific community, complementarity-determining region 3 (CDR3) with designated V and J gene segments is the accepted identifier that unambiguously defines TCR α- or a β-chain sequence.

MiXCR adheres to the conventional approach of identifying TCRs by their complete CDR3 sequence. Both V and J segments are mapped using conventional, dynamic programming-based alignment algorithms optimized for nearly absolute specificity. Each CDR3 in MiXCR output is extracted on the basis of this alignment information, as a sequence located between the fixed positions inside V and J regions. Thus, CDR3 boundaries are always exactly positioned in the gene context[2].

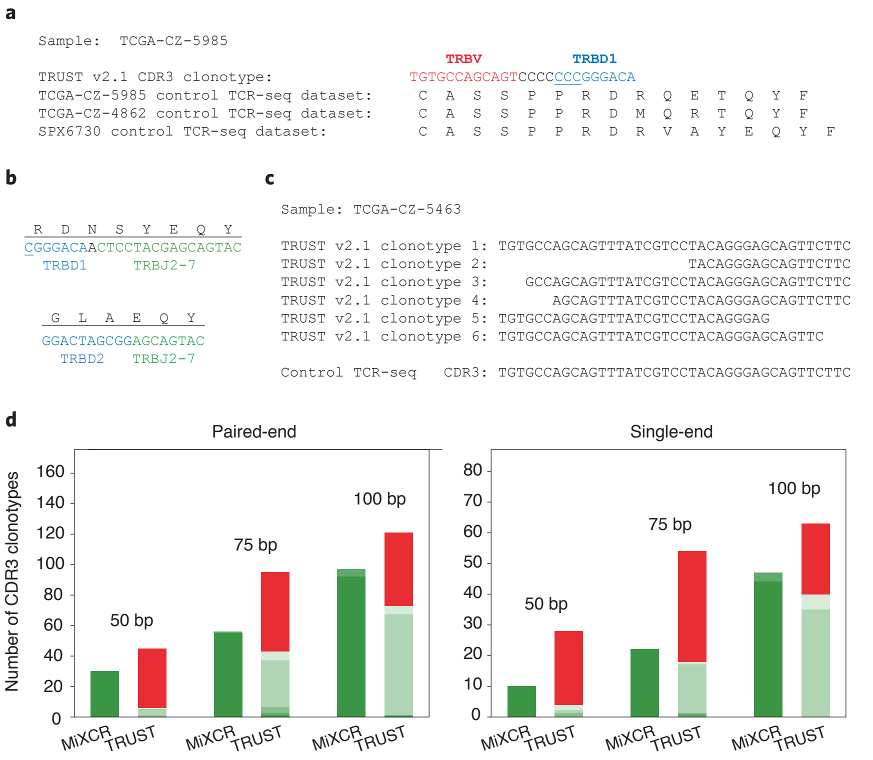
In contrast, TRUST v2.1 (ref. 3) uses fragments of TCR sequence with no strictly defined start and end positions to define unique clonotypes. Because of the combinatorial nature of TCR and immunoglobulin generation, a fragment that partially covers the V(D)J junction can be present in multiple different clonotypes. Therefore, TRUST v2.1 generates a substantial portion of fragments that cannot be unambiguously assigned to particular CDR3 clonotypes. For example, a sequence extracted by TRUST v2.1 from The Cancer Genome Atlas (TCGA) RNA sequencing (RNA-seq) sample TCGA-CZ-5985 matches different CDR3 clonotypes in unrelated control TCR sequencing (TCR-seq) datasets (Fig. 1a). We found many such examples in TRUST v2.1-extracted repertoires (https://bitbucket.org/liulab/trust/src/nbt-response), with up to 18% of reported sequences mapping to dozens and even hundreds of CDR3 clonotypes in unrelated control TCR-seq datasets. Therefore, on the basis of the results of TRUST v2.1, one could reach incorrect conclusions about the existence of TCR variants shared among T cell repertoires, even though there might be no actual connection between them in terms of common clonotypes.

Hu *et al.* claim that partial CDR3 fragments produced by TRUST software may be informative for modeling TCR binding specificity. We believe this statement is misleading for two reasons: first, rational sequence-based prediction of TCR specificity requires positioning of a motif in the context of the full TCR chain[4]; and second, most of the TRUST v2.1-derived fragments that are shared between samples largely represent germline-segment sequences. For example, in ref. 5, the authors identified three CDR3 fragments, the presence of which in TCGA tumor RNA-seq data correlates with three immunogenic somatic mutations. However, the identified fragments (RDNSYEQY, GESEQY and GLAEQY) are well represented in most of the TCR repertoires of memory and naive, CD4 and CD8 peripheral blood T-cell subsets of healthy donors reported in ref. 6. Such sequences are often formed at the D-J junction, with no or few randomly added nucleotides in between (Fig. 1b). Thus, they represent frequent recombination events[7] and belong to a pool of highly public clonotypes[8].

In addition, multiple fragments generated by TRUST v2.1 and designated as independent CDR3 clonotypes match to the same single clonotype in a control dataset (Fig. 1c). This leads to substantial overestimation of the number of confirmed clonotypes in a TRUST v2.1 performance evaluation[3]. In general, the output of TRUST v2.1 does not reflect the number of identified unique CDR3s, but instead reflects the count of extracted sequence fragments related to TCR regions.

Recent work on TRUST that incorporates V/J germline sequence-based CDR3 extension could be beneficial for TRUST users, potentially producing full CDR3 sequences for conventional comparison techniques. However, the community should be aware that one must be vigilant to avoid producing false clonotypes when incorporating CDR3 extension.

As was indicated in the Supplementary Methods of our original publication[2], we used TopHat aligner with hg19 to produce


a

```
Sample:  TCGA-CZ-5985
                                           TRBV          TRBD1
TRUST v2.1 CDR3 clonotype:                 TGTGCCAGCAGTCCCCCCCGGGACA
TCGA-CZ-5985 control TCR-seq dataset:       C  A  S  S  P  P  R  D  R  Q  E  T  Q  Y  F
TCGA-CZ-4862 control TCR-seq dataset:       C  A  S  S  P  P  R  D  M  Q  R  T  Q  Y  F
SPX6730 control TCR-seq dataset:            C  A  S  S  P  P  R  D  R  V  A  Y  E  Q  Y  F
```

b

```
 R  D  N  S  Y  E  Q  Y
CGGGACAACTCCTACGAGCAGTAC
  TRBD1      TRBJ2-7

   G  L  A  E  Q  Y
  GGACTAGCGGAGCAGTAC
    TRBD2    TRBJ2-7
```

c

```
Sample: TCGA-CZ-5463

TRUST v2.1 clonotype 1: TGTGCCAGCAGTTTATCGTCCTACAGGGAGCAGTTCTTC
TRUST v2.1 clonotype 2:                      TACAGGGAGCAGTTCTTC
TRUST v2.1 clonotype 3:    GCCAGCAGTTTATCGTCCTACAGGGAGCAGTTCTTC
TRUST v2.1 clonotype 4:       AGCAGTTTATCGTCCTACAGGGAGCAGTTCTTC
TRUST v2.1 clonotype 5: TGTGCCAGCAGTTTATCGTCCTACAGGGAG
TRUST v2.1 clonotype 6: TGTGCCAGCAGTTTATCGTCCTACAGGGAGCAGTTC

Control TCR-seq   CDR3: TGTGCCAGCAGTTTATCGTCCTACAGGGAGCAGTTCTTC
```

**Figure 1** A comparison of the outputs of TRUST v2.1 and MiXCR. TRUST v2.1 (ref. 3) uses fragments of TCR sequence with no strictly defined start and end positions to define unique clonotypes. (**a**) Example of a sequence fragment designated by TRUST v2.1 as a CDR3 clonotype mapping to three different CDR3 clonotypes in both control and unrelated TCR-seq datasets. P-segment underlined. (**b**) Example of sequence fragments designated by TRUST v2.1 as CDR3 clonotypes that were reported to correlate with an immunogenic somatic mutation[5]. The fragments map to the D–J junction. (**c**) Example of six distinct CDR3 clonotypes reported by TRUST v2.1 that map to a single control TCR-seq clonotype. (**d**) MiXCR v2.1.3 and TRUST v2.1 performance comparison on *in silico*–generated data as in Supplementary Figure 3 of ref. 2. BAM files were rebuilt using STAR with disabled local alignment, hg19. Dark green corresponds to fully matched CDR3 sequences (without any mismatches or indels), lighter shades of green to CDR3 sequences matched with mismatches or indels (up to three mutations), red to reported CDR3 sequences that did not match any of the original synthetic clonotypes.

BAM files in all tests except tests with *in silico* data. Thus, all results obtained for TRUST v2.1 presented in Figure 1 of our original paper[2] were performed using the aligner options recommended by TRUST developers at the time and are fully valid.

TRUST documentation available from the official code repository (https://bitbucket.org/liulab/trust/src/2b9d8af155938f1b8d4f56326963d1610b2638f5/README.txt?at=master&fileviewer=file-view-default) does not indicate recommended aligner parameters for STAR. We apologize that we did not find this information in the Supplementary Material of the paper[3]. We have now reproduced the *in silico* tests shown in Supplementary Figure 3 of our original publication[2] using STAR with disabled local alignment, as Hu *et al.* and their original publication[3] recommend (see **Supplementary Note 1** for commands used). Using these parameters, the previously observed difference between results obtained for paired-end and single-end data using TRUST v2.1 disappeared; however, the general trends remained the same (**Fig. 1d**). TRUST v2.1 output contained fewer confirmed (coinciding with the full CDR3 sequence) or partially confirmed (with up to three mismatches or indels allowed) CDR3 sequences than MiXCR output. Many TRUST v2.1-generated sequences failed to match the original set of synthetic clonotypes, whereas none of the MiXCR-generated sequences failed to match.

With regards to our method of negative control random transcripts generation, Hu *et al.* suggest using the transcriptome assembly with only protein-coding sequences (gencode.pc_v19), as TRUST v2.1 relies on this assembly. We disagree with this approach and maintain that real data contain a more diverse set of transcripts, which is better reflected by the most comprehensive set of transcripts available (for example, gencode.v26.transcripts.fa). Using the latter set of transcripts, TRUST v2.1 produces false positives with either enabled or disabled local alignments.

We regret that Supplementary Table 1 in our original paper[2] indicated that TRUST v2.1. is not an open source software. This was a mistake on our part.

In closing, we want to emphasize that the absence of ambiguously mapped calls is critical for the comparative analysis aimed at identifying the TCR or immunoglobulin variants shared between samples (for example, associated with a certain disease condition, response to vaccination, or functional subsets of T or B cells). Conversely, the presence of multiple distinct calls derived from the same clonotypes (**Fig. 1c**) precludes summary statistical analysis, such as estimation of repertoire diversity or clonality. The absence of anchor points for positioning of CDR3 impairs both rational searches for the characteristic TCR or immunoglobulin motifs[4] and comparisons of the physicochemical properties of immune repertoires[9].

Further evolution of software tools for extraction of immune repertoires from RNA-seq data remains desirable, but will require substantial time and efforts to perform scrupulous verification of the logic of the algorithms used and the techniques used for validation, as well as quality checks on the program code.

**Code availability.** MiXCR is available at https://github.com/milaboratory/mixcr. Data analysis code is available at https://github.com/milaboratory/mixcr-rna-seq-paper.

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

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.4296).*

ACKNOWLEDGEMENTS
Research work supported by grant of the Ministry of Education and Science of the Russian Federation Number 14.W03.31.0005.

AUTHOR CONTRIBUTIONS
D.A.B. and S.P. worked on the code. All authors analyzed the data. All authors contributed to the writing of the manuscript.

COMPETING INTERESTS
MiLaboratory LLC develops MiXCR software and has exclusive rights for its commercial distribution.

*Dmitriy A Bolotin[1–3,7], Stanislav Poslavsky[1,3,7], Alexey N Davydov[4] & Dmitriy M Chudakov[2–6]*

*[1]MiLaboratory LLC, Skolkovo Innovation Center, Moscow, Russia. [2]Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia. [3]Pirogov Russian National Research Medical University, Moscow, Russia. [4]Central European Institute of Technology, Brno, Czech Republic. [5]Privolzhsky Research Medical University, Nizhny Novgorod, Russia. [6]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia. [7]These authors contributed equally to this work.*
*e-mail: ChudakovDM@gmail.com*

1. Pogorelyy, M.V. *et al. Elife* **7**, e33050 (2018).
2. Bolotin, D.A. *et al. Nat. Biotechnol.* **35**, 908–911 (2017).
3. Li, B. *et al. Nat. Genet.* **49**, 482–483 (2017).
4. Dash, P. *et al. Nature* **547**, 89–93 (2017).
5. Li, B. *et al. Nat. Genet.* **48**, 725–732 (2016).
6. Qi, Q. *et al. Proc. Natl. Acad. Sci. USA* **111**, 13139–13144 (2014).
7. Murugan, A., Mora, T., Walczak, A.M. & Callan, C.G. Jr. *Proc. Natl. Acad. Sci. USA* **109**, 16161–16166 (2012).
8. Venturi, V. *et al. J. Immunol.* **186**, 4285–4294 (2011).
9. Izraelson, M. *et al. Immunology* **153**, 133–144 (2018).