production strains for industrially valuable compounds. We hope that MiL1515 will provide researchers with the best possible tools for their experiments.

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.

ACKNOWLEDGMENTS

We would like to thank D. Zielinski, J. Lerman, J. Orth, S. Bowen, and J. Reed for helpful discussions. We’d like to thank M. Abrams for copyediting help. The work was funded by the Novo Nordisk Foundation (Grant NN10CC1016517) and by grant 1R01GM057089 from the NIH/NIGMS and 1-U01AI24316-01 from NIH/NIAID. B.O.P., C.J.L., and A.M.F. received support from the Novo Nordisk Foundation Center for Biosustainability (NNF10CC1016517).

AUTHOR CONTRIBUTIONS


COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jonathan M Monk1,4, Colton J Lloyd1,4, Elizabeth Brunk1,4, Nathan Mih1, Anand Sasytri1, Zachary King1, Rikika Takeuchi2, Wataru Nomura2, Zhen Zhang2, Hirotada Mori2, Adam M Feist1,3 & Bernhard O Palsson1,3

1Department of Bioengineering, University of California, San Diego, La Jolla, California, USA.
2Graduate School of Biological Sciences Nara Institute of Science and Technology Ikoma, Nara, Japan.
3The NNF Center for Biosustainability, The Technical University of Denmark, Lyngby, Denmark.
4These authors contributed equally to this work.

email: palsson@ucsd.edu

To the Editor: Somatic recombination and accumulation of mutations in V-D-J segments result in vast heterogeneity of T-cell receptor (TCR) and immunoglobulin repertoires. High-throughput profiling of immune receptors has become an important tool for studies of adaptive immunity and for the development of diagnostics, vaccines, and immunotherapies. There are efficient molecular and software tools for the targeted sequencing of TCR and immunoglobulin repertoires, including MiXCR, developed by our team. However, sufficient amount and quality of tissue or extracted RNA or DNA are not always available for analysis.

An alternative way of immune profiling is to use TCR and immunoglobulin transcripts that are present in bulk RNA-seq data. Because transcriptome sequencing has become routine in both basic and clinical studies, it could serve as a source of functionally relevant information on immune receptor hypervariable regions.

The major advances in our approach allow unified processing and comparison of immune repertoires obtained from different types of raw sequencing data. From the Cancer Genome Atlas (TGCA, with >10,000 tumor samples: https://gdc-portal.nci.nih.gov/) and other databases could be employed for immune repertoire profiling. Such analysis is of particular interest in cancer immunotherapy studies. Available tumor tissue is often limited, which precludes splitting the samples for separate transcriptome, TCR, and immunoglobulin profiling. Separate immune repertoire profiling also adds complexity and increases the costs for massive clinical studies.

Furthermore, transcriptomic analysis is often employed in comparative studies of functional T- and B-cell subsets, and it could additionally yield the immune receptor repertoires at no cost.

Several groups have reported tools for TCR or immunoglobulin repertoire extraction from bulk or single-cell RNA-seq data. However, a broadly applicable software tool that enables the accurate and efficient extraction of immune repertoires from RNA-seq was not available.

Here we develop a tool on the basis of MiXCR3, implementing a set of algorithms (Supplementary Methods, Supplementary Note 1, Supplementary Figs. 1 and 2, and Supplementary Tables 1 and 2) to extract as many true CDR3 sequences as possible, with nearly zero CDR3-like false positives. The major advances in our approach include the following: 1) a sensitive and highly selective aligner that employs a fast algorithm but switches to a more sensitive modified Smith–Waterman/Needleman–Wunsch algorithm in ambiguous cases; 2) a partial alignments assembler that builds contigs from several initial alignments in a manner protected from artificial diversity generation (http://dx.doi.org/10.5281/zenodo.804326); and 3) CDR3 extension (for TCRs but not immunoglobulins, because of the possible presence of hypermutations) that fills in the edges of the CDR3 based on known information on the relevant germline gene segments. The resulting RNA-seq analysis pipeline employs the same MiXCR modules, the same error-correction algorithms, and has the same output format as for targeted TCR or immunoglobulin profiling. This allows unified processing and comparison of immune repertoires obtained from different types of raw sequencing data.

Software testing with in silico-generated data demonstrated the high extraction efficiency of MiXCR, with zero false-positive clones observed. TRUST software13,16 efficiency was an order of magnitude lower, and the software generated a substantial number of false clonotypes (Supplementary Methods and Supplementary Fig. 3).

Correspondence

Antigen receptor repertoire profiling from RNA-seq data

To further verify the efficiency and specificity of extracting TCR CDR3 repertoires from RNA-seq data, we performed deep, targeted profiling of TCR alpha-(TRA) and beta-(TRB) chain repertoires (TCR-seq), as described\(^\text{18}\) and 100+100-bp paired-end RNA-seq for the same RNA samples, obtained from surgically resected melanoma specimens from two patients, SPX6730 (ileocecal lymph node metastasis) and SPX8151 (small intestine resection).

We next assessed, estimated on the basis of TCR-seq data, the dependence of the number of TCR-seq-confirmed clonotypes on their abundance. MiXCR was able to extract all relatively large TRB-CDR3 clonotypes (frequency in repertoire > 0.15%) from the lymph node metastasis RNA-seq, even with the short paired-end reads (trimmed in silico to 50+50 bp). In contrast, TRUST failed to extract a considerable proportion of high-frequency clonotypes (Fig. 1a and Supplementary Fig. 2). Most MiXCR-reported clonotypes were confirmed by TCR-seq data, while only a minor fraction of CDR3s reported by TRUST could be confirmed (Fig. 1b). The clonotype frequencies extracted by MiXCR from RNA-seq data correlated with TCR-seq (Fig. 1c), demonstrating that RNA-seq-based profiling can be quantitative for large clonotypes in samples that contain a substantial number of T cells.

We also compared MiXCR with V'DJer software\(^\text{15}\) for the extraction of immunoglobulin repertoires. For some samples (including large SPX6730 and SPX8151 RNA-seq data sets) V'DJer failed to extract immunoglobulin heavy-chain (IGH) and light-chain (IGL, IGK) repertoires within 4 d using eight threads on a Xeon E5-2683 CPU with 50 GB of RAM. MiXCR successfully extracted repertoires for IGH, IGK, IGL, TRA, and TRB chains from both melanoma samples, and outperformed both V'DJer and TRUST, yielding many-fold more canonical and minimal numbers of non-canonical clonotypes (Fig. 1d). MiXCR demonstrated superior sensitivity, extracting 10- to 1,000-fold more clonotypes compared to V'DJer (Fig. 1e). Both V'DJer and TRUST require substantially more hands-on time and application of third-party alignment tools with particular versions of the human genome and particular analysis settings, which are not clearly defined in the documentation and required optimization. The output from both tools irretrievably loses useful biological information (Supplementary Table 1).
CORRESPONDENCE

Figure 2 Immune repertoire analysis in human melanoma and sorted mice T cells. (a–d) Kaplan–Meier plots depicting the survival probability over time relative to repertoire-related metrics for TCGA SKCM patient samples. Samples were divided into high (≥cutoff) and low (<cutoff) value cohorts based on the median for the relevant metric. (a) IGH expression (reads per million). (b) IGH clonality. Samples with at least 500 IGH CDR3 sequencing reads were included to minimize the influence of repertoire depth on clonality index. (c) IGH clonality and expression. (d) Kaplan–Meier plots showing survival probability relative to IgG1/IGH and IgA/IGH proportions. P values reflect log-rank test for survival difference between low and high metrics groups. n, number of patients. (e–i) Comparative analysis of CD4 effector and regulatory T-cell repertoires. (e) Correlation of observed TRA versus TRB CDR3 diversity, each per 500 unique CDR3-covering reads. (f) TRB diversity observed per 500 randomly sampled unique CDR3-covering reads (left), and estimated using the Shannon–Wiener index (right) for the spleen (SP) and central nervous system (CNS) samples from Foxp3<sup>+</sup> mice with induced autoimmune encephalomyelitis. (g) Average TRB CDR3 lengths for Teff and Treg subsets in spleen, weighted for clonotype size. (h) Functional characteristics of the middle portion of the CDR3 amino acid sequence for Teff and Treg TRB repertoires in the spleen. Plots show hydrophathy, polarity, and strength metrics as derived from VDJtools software, weighted for clonotype size. (i) Multidimensional space (mds) analysis of TRB CDR3 amino acid repertoire overlaps, as derived from VDJtools, metric D. *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001 (paired, two-tailed Student’s t-test).

single-cell T-cell transcriptome analysis, MiXCR outperformed TraCeR<sup>17</sup> in efficiency of TRA and TRB chain detection (Supplementary Table 3).

Next, we extracted repertoires from TCGA 48+48-bp paired-end RNA-seq for 458 patients with cutaneous melanoma (SKCM) (see http://dx.doi.org/10.6084/m9.figshare.4620739 for clone sets and Supplementary Note 2). Notably, the obtained immunoglobulin repertoires were an order of magnitude larger than the TCRs, indicating the presence of intratumoral immunoglobulin-producing plasma cells.

High intratumoral IGH expression levels were associated with longer survival (Fig. 2a), in agreement with recent work showing a positive correlation of activated B-cell gene signatures with survival in SKCM<sup>19</sup>. High levels of IGH clonality (analyzed as in ref. 20) were also associated with longer survival (Fig. 2b and Supplementary Fig. 4a), and the two parameters had strong cumulative value. Patients with both high IGH expression and clonality had the best prognosis, whereas high IGH expression with low clonality was associated with poor prognosis (Fig. 2c).

In many patients, a single dominant intratumoral immunoglobulin clonotype occupied 30–80% of all immunoglobulin CDR3 sequences in both heavy- and light-chain repertoires. Hypermutating IGH CDR3 variants could be observed even in primary tumor samples (Supplementary Table 4 and Supplementary Fig. 4b), which could reflect the presence of intratumoral germinal centers. Observation of extra-large immunoglobulin expansions and hypermutation processes in tumors raise the question of the antigenic specificity of dominant immunoglobulin variants, for which functional heavy- and light-chain pairs can be identified based on frequency<sup>21</sup>. If such dominant intratumorally produced immunoglobulins are tumor-specific<sup>22</sup>, exploration of their usefulness in precision immunotherapy—for example, in the context of chimeric antigen receptors (CARs<sup>23</sup>)—would be of considerable interest.

We found that a high proportion of IgG1 of all IGH transcripts was associated with longer survival (Fig. 2d and Supplementary Fig. 4), suggesting that intratumorally produced clonal...
IgG1 antibodies could exert opsonizing antitumor effects. A high IgA/IGH proportion was associated with a negative prognosis, consistent with recent work showing that IgA-producing plasma cells function as potent immunosuppressors through the secretion of IL-10 and PD-L1 (ref. 24). Collectively, these results indicate that intratumorally produced immunoglobulins may represent an important component of the anti-tumor response, and extraction of immunoglobulin clonal repertoires from RNA-seq data could be a potential source of clinically useful biomarkers for cancer immunotherapy.

To test the feasibility of extracting TCR repertoire sequences from sorted T cells, we performed 50-50 bp paired-end RNA-seq for the effector (T_{eff}) and regulatory (T_{reg}) CD4 T cells from the spleen and central nervous system (CNS) of Foxp3^{gfcre} mice\(^{25}\) with induced autoimmune encephalomyelitis (Supplementary Note 3). The near-100% abundance of T cells in these samples allowed MiXCR to extract high-quality TCR repertoires comprising, on average, 1,330 TRA and 1,489 TRB unique sequences from sorted T cells, allowing MiXCR to extract high-quality TCR repertoires comprising, on average, 1,330 TRA and 1,489 TRB unique sequences from sorted T cells, and 3,295 TRA and 3,105 TRB sequences from sorted T-cell subsets. RNA-seq analysis is rapidly becoming routine, both in clinical practice and basic research, with implementation gaining steam as the cost of high-throughput sequencing decreases and read-length increases. The MiXCR pipeline now enables straightforward mining of these data sets for immune repertoire profiling. We anticipate that in the near future RNA-seq-based profiling may replace targeted profiling of TCR and immunoglobulin repertoires in many applications.

The software, comprehensive user documentation, and source code can be found at https://github.com/milaboratory/mixcr. Our software is continuously upgraded, and extraction of immunoglobulin clonal repertoires from RNA-seq data could be a component of the anti-tumor response, and extraction of immunoglobulin clonal repertoires from RNA-seq data could be a potentially useful biomarker for cancer immunotherapy.

To test the feasibility of extracting TCR repertoire sequences from sorted T cells, we performed 50–50 bp paired-end RNA-seq for the effector (T_{eff}) and regulatory (T_{reg}) CD4 T cells from the spleen and central nervous system (CNS) of Foxp3^{gfcre} mice\(^{25}\) with induced autoimmune encephalomyelitis (Supplementary Note 3). The near-100% abundance of T cells in these samples allowed MiXCR to extract high-quality TCR repertoires comprising, on average, 1,330 TRA and 1,489 TRB unique sequences from sorted T cells, allowing MiXCR to extract high-quality TCR repertoires comprising, on average, 1,330 TRA and 1,489 TRB unique sequences from sorted T cells, and 3,295 TRA and 3,105 TRB sequences from sorted T-cell subsets. RNA-seq analysis is rapidly becoming routine, both in clinical practice and basic research, with implementation gaining steam as the cost of high-throughput sequencing decreases and read-length increases. The MiXCR pipeline now enables straightforward mining of these data sets for immune repertoire profiling. We anticipate that in the near future RNA-seq-based profiling may replace targeted profiling of TCR and immunoglobulin repertoires in many applications.

The software, comprehensive user documentation, and source code can be found at https://github.com/milaboratory/mixcr. Our software is continuously upgraded, and extraction of immunoglobulin clonal repertoires from RNA-seq data could be a potentially useful biomarker for cancer immunotherapy.
学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具