BraCeR: B-cell receptor reconstruction and clonality inference from single-cell RNA-seq

Ida Lindeman[1,2,3], Guy Emerton[3], Lira Mamanova[3], Omri Snir[1,2], Krzysztof Polanski[3], Shuo-Wang Qiao[1,2,4], Ludvig M. Sollid[1,2,4], Sarah A. Teichmann[3,5] & Michael J.T. Stubbington[3]

[1]Centre for Immune Regulation, University of Oslo and Oslo University Hospital, Oslo, Norway. [2]Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway. [3]Wellcome Sanger Institute, Hinxton, United Kingdom. [4]KG Jebsen Coeliac Disease Research Centre, University of Oslo, Oslo, Norway. [5]Theory of Condensed Matter, Cavendish Laboratory, Cambridge, United Kingdom.

Present address for M.J.T.S is 10x Genomics, Crown Business Center Key Point II, Office Space, Units F-07 & F-08 , Schipholweg 103, 2316 XC Leiden, The Netherlands.

Corresponding author:

M.J.T.S

mjt.stubbington@gmail.com

Reconstruction of antigen receptor sequences from single-cell RNA-sequencing (scRNA-seq) data allows the linking of antigen receptor usage to the full transcriptomic identity of individual B

lymphocytes, without having to perform additional targeted repertoire sequencing (Rep-seq). Here we report BraCeR (freely available at https://github.com/teichlab/bracer/ and as **Supplemental Software**), an extension of TraCeR [1], for reconstruction of paired full-length B-cell receptor sequences and inference of clonality from scRNA-seq data (**Supplementary Note 1**).

BraCeR builds on the well-verified pipeline of TraCeR for assembly of BCR sequences from paired-end or single-end reads, modified to account for somatic hypermutations (SHM) and isotype switching (**Supplementary Notes 2** and **3**). In addition, a 'Build' mode facilitates the creation of resource files for the analysis of species beyond human and mouse.

BraCeR was tested against experimental human and mouse scRNA-seq data with various SHM rates and repertoire diversities [2, 3]. The reconstruction accuracy was similar to that of BASIC, a previously reported tool for BCR reconstruction [2]. Comparatively, BraCeR yielded a reconstruction efficiency that was somewhat superior for the long (125 bp) reads and similar to BASIC for short (50 bp) reads (**Supplementary Note 4** and **Supplementary Tables 1-7**). Importantly however, BraCeR allows for reconstruction of additional heavy and light chains present in a cell, and it can identify non-productively rearranged chains; we validated BCR reconstruction (including additional sequences) by direct comparison with targeted BCR-sequencing in plasma cells (**Supplementary Note 5** and **Supplementary Tables 8-9**). It is also possible to provide BraCeR with FASTA files containing BCR sequences assembled by other means or determined by single-cell Rep-seq for downstream analyses.

BraCeR further improves upon BASIC by using the reconstructed sequences to infer clonal

relationships and perform immunoglobulin lineage reconstruction. BraCeR identifies the isotype of each reconstructed BCR and collapses highly similar reconstructed sequences in a cell. It then quantifies the expression of the BCR sequences and identifies the most highly expressed chains for each locus in each cell. Potential cell multiplets or cross-cell contaminations are reported, and are by default removed from downstream analyses (**Supplementary Note 6**).

Clonally related cells are grouped into clonotypes based on the reconstructed chains in each cell. Productively rearranged reconstructed BCRs for each locus are grouped into clones based on V- and J-gene assignments using custom Python scripts and a previously described method for comparison of CDR3 sequence similarity [4] (**Supplementary Note 7**). Clonal networks are constructed based on shared clonal heavy and light chains and then visualised (**Fig. 1A**). As an optional feature, BraCeR can also use existing tools [4-6] to construct lineage trees for each group of clonal cells (**Fig. 1B**), with the inferred germline sequence as outgroup, facilitating linkage of transcriptomic phenotype to the evolution of immunoglobulin sequences (**Supplementary Note 8**).

With an easy-to-use command-line interface, BraCeR provides a complete pipeline for clonal inference and lineage tracing of B cells; raw scRNA-seq reads can be processed all the way to clonal networks and lineage trees. Additional output includes summary files and graphs summarising reconstructed sequences and isotype usage. BraCeR also creates tab-delimited database files according to the Change-O Data Standard [4], thus facilitating further analysis of the reconstructed BCR sequences using additional downstream tools. BraCeR aims to follow the recommendations of the Adaptive Immune Receptor Repertoire (AIRR) Community (http://airr.irmacs.sfu.ca).

Details on methods and experimental design are described in the **Supplementary Methods**, **Supplementary Notes**, **Life Sciences Reporting Summary** and **Supplementary Table 10**.

Data availability

Raw reads from our verification experiment are available under controlled access through the European Genome-phenome Archive (EGA). Due to the sensitive nature of patient-derived data only the BCR-derived reads were extracted and deposited under accession numbers EGAN00001781663 through EGAN00001781675. The demultiplexed targeted BCR-seq data is deposited under accession numbers EGAN00001719670 through EGAN00001719682. Supplementary Figure SN4.2 has associated raw data.

Acknowledgements

We would like to thank K. E. A. Lundin at Oslo University Hospital for providing clinical material, E. S. Fasouli, A. Jinat and L. M. Eggesbø for assistance with experiments and S. Kleinstein for helpful discussions. This work was funded by grants from the University of Oslo world-leading research program on human immunology (WL-IMMUNOLOGY), the South-Eastern Norway Regional Health Authority (projects 2014045 and 2016113) and Stiftelsen KG Jebsen (SKGHMED-017) to L.M.S., Wellcome Trust Grant 206194 (M.J.T.S., S.A.T., G.E., L.M., and K.P.), and ERC Grant ThDEFINE (646794) (S.A.T.).

Author contributions

I.L., M.J.T.S, G.E. and K.P. developed the software. I.L., L.M., O.S. and S.-W.Q. performed scRNA-seq. I.L. analysed the data. M.J.T.S., L.M.S. and S.A.T. supervised the project. I.L. and M.J.T.S. wrote the manuscript.

Competing financial interests

M.J.T.S. has been employed by 10x Genomics since April 2018; this employment had no bearing on this work. The other authors declare no competing financial interests.

Figure legends

Figure 1. Clonal network (**A**) and lineage trees (**B**) for each clonotype defined by BraCeR for a human plasmablast scRNA-seq dataset [2].

References

1. Stubbington, M.J. *et al. Nat Methods*. **13**, 329-32 (2016).
2. Canzar, S. *et al. Bioinformatics*. **33**, 425-427 (2017).
3. Wu *et al*. *J Exp Med.* **214,** 183 - 196 (2017).
4. Gupta, N.T. *et al. Bioinformatics*. **31**, 3356-8 (2015).
5. Stern, J.N.H. *et al*. *Sci Transl Med*. **6**, 248ra107 (2014).
6. Felsenstein, J. *Cladistics*. **5**, 164-166 (1989).