Title: Ibex: Variational autoencoder for single-cell BCR sequencing.

Authors: Nicholas Borcherding[1,*], Bo Sun[2], David DeNardo[3], Jonathan R. Brestoff[1]

1. Department of Pathology and Immunology, Barnes-Jewish Hospital, Washington University in St Louis, MO 63110, USA

2. Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

3. Department of Medicine, Barnes-Jewish Hospital, Washington University in St Louis, MO 63110, USA

Nicholas Borcherding, MD, PhD

Department of Pathology and Immunology

Division of Laboratory and Genomic Medicine

Washington University School of Medicine

St Louis, MO, 63110

United States

Phone: 314-273-1280

Email: borcherding.n@wustl.edu

Twitter: @theHumanBorch

Abstract

Summary

B cells are critical for adaptive immunity and are governed by the recognition of an antigen by the B cell receptor (BCR), a process that drives a coordinated series of signaling events and modulation of various transcriptional programs. Single-cell RNA sequencing with paired BCR profiling could offer insights into numerous physiological and pathological processes. However, unlike the plethora of single-cell RNA analysis pipelines, computational tools that utilize single-cell BCR sequences for further analyses are not yet well developed. Here we report Ibex, which vectorizes the amino acid sequence of the complementarity-determining region 3 (cdr3) of the immunoglobulin heavy and light chains, allowing for unbiased dimensional reduction of B cells using their BCR repertoire. Ibex is implemented as an R package with integration into both the Seurat and Single-Cell Experiment framework, enabling the incorporation of this new analytic tool into many single-cell sequencing analytic workflows and multimodal experiments.

Availability and Implementation

Ibex is available as an R package at https://github.com/ncborcherding/Ibex. Reproducible code and data for the figure appearing in the manuscript are available at https://github.com/ncborcherding/Ibex.manuscript.

Introduction

Single-cell sequencing has become a mainstay in immunology. In addition to enabling the characterization of various immune populations and their transcriptional profiles, single-cell sequencing technologies allow for pairing of transcriptomic and clonotypic data and can therefore provide insight into the complexities of the adaptive immune response. In the context of lymphocytes, the immune response is regulated by antigen recognition at the site of the adaptive immune receptor. In B cells, the BCR sequence is determined by the somatic recombination genes at heavy and light chain loci (Schatz and Ji, 2011) and in the germinal center via somatic hypermutation.(Odegard and Schatz, 2006) This process collectively produces groups of sequences of “clonal lineages” that are generally used as categorical data in single-cell analysis.

Several tools have been put forth to integrate single-cell transcriptomic and T cell receptor sequences, such as CoNGA (Schattgen *et al.*, 2022), TESSA (Zhang *et al.*, 2021), and mvTCR.(Drost *et al.*, 2022) The latter two tools leverage the use of neural networks to generate a latent representation of the cdr3 sequence. The power of this approach has been shown in cancer detection (Beshnova *et al.*, 2020) and in predicting responses to immunotherapy.(Sidhom *et al.*, 2022) Recent deep-learning approaches have shown promise in the optimization of antibody selection.(Mason *et al.*, 2021) Other groups proposing the potential of using natural language processing to better understand the linguistics of immunoglobulins (Ig) and improve in silico development.(Vu *et al.*, 2022) However, tools for the integrating multimodal single-cell sequencing data with BCR sequences are lacking.

We developed an intuitive tool to encode single-cell BCR light and heavy chain sequences using variational autoencoder models. For each chain, the sequences can be encoded using the amino acid properties of the Atchley factors (Atchley *et al.*, 2005), Kidera factors (Kidera *et al.*,

1985), or directly encoded with the one-hot autoencoding approach. With the Ibex package, the encoded numerical vectors can be incorporated into single-cell objects and used directly for dimensional reduction or multimodal integration. This provides an uncomplicated and adaptable approach for the incorporation of BCR sequences to the analysis of single-cell B cell data. In addition, Ibex is compatible with our previous scRepertoire R package (Borcherding *et al.*, 2020), enabling users to easily attach adaptive immune receptor sequences to single-cell objects. To demonstrate the application of Ibex, we investigated single-cell peripheral blood B cells in COVID19-associated multisystem inflammatory syndrome.(Ramaswamy *et al.*, 2021)

Results

Ibex is an R package that translates the BCR light and heavy chain cdr3 sequences using variational autoencoders. Each available autoencoder was trained on 600,000 human cdr3 sequences of the respective Ig chain, with a 128-64-30-64-128 neuron structure in the keras R package (v2.4.0). Each autoencoding training involved zero padding up to 70 residues to allow for variable cdr3 lengths. In addition, each model was trained using 80:20 data split with a batch size of 128 and an epoch of 30. The output of these autoencoder models is a 30-dimensional latent space representing the amino acid cdr3 sequence (Figure 1A).

To benchmark the autoencoding approach for both heavy and light chains, a bootstrapping approach using randomized novel cdr3 sequences that no model had seen (Figure 1B). For each of the models, 10, 100, 1000, 1e5, and 1e6 novel sequences were tested a total of 10 times. Comparing the number of cdr3 sequences assessed to the unique latent dimensional output for the models demonstrated concordance (Figure 1B, upper panel). The exception was for light chain models assessing 1e6 sequences had two sequences with the same latent dimensional output. The time of encoding between the approaches available in Ibex was

relatively consistent with the encoding of 1e6 sequences taking under 30 seconds of computation (Figure 1B, lower panel).

In order to demonstrate the application of Ibex, we selected a publicly-available data set of pediatric multisystem inflammatory syndrome patients.(Ramaswamy *et al.*, 2021) This data set has the advantage of providing peripheral blood samples with single-cell RNA and protein quantifications with adaptive immune receptor sequencing. After combining the data for patient 1 of the cohort, B cells were isolated using canonical lineage markers and BCR sequences. After which, Ibex encoding for both the light and heavy chains of the B cells was performed by converting the amino acid sequence into Kidera factors. The latent dimensions from the encoding can be used to directly explore sequence similarities, such as the clustering of IgK and Igλ (Figure 1C). Ibex also enables the reduction of cell-level quantifications to clonotype-level quantifications using minimal Euclidean distance across principal component dimensions (Figure 1D), as previously described (Schattgen *et al.*, 2022), or as an average across all cells defined as clones. The clonotype-level single-cell data can be used to generate UMAPs for the RNA, antibody-derived tag (ADT) protein level quantification, or the Ibex-based chain encoded values (Figure 1E). However, each of these assays can be rescaled across modalities to generate normalized values, using packages such as mumosa R package (Lun, Aaron, 2022) (Figure 1F) or weighted-nearest neighbors in the Seurat R package.(Hao *et al.*, 2021) The resulting weighted or normalized values can then be used to create a unified representation of RNA, protein, Ig heavy chain, such as with dimensional reduction or clustering (Figure 1G).

Conclusion

Ibex is an R package designed to combine deep learning with immune repertoire profiling and incorporate them into common single-cell sequencing analytic workflows. The package offers customizable encoding of BCR light and heavy chains to produce latent dimensional

representations of amino acid sequences. Alone or in combination with other single-cell modalities, these latent vectors may assist in the characterization of an immune response, understanding Ig maturation, or possible epitope identification. In future work, we will attempt to add clonal lineage analysis and examine the fidelity of the autoencoding process across the immense diversity of BCR/TCR sequences.

Acknowledgments

Financial Support: This work was supported by the National Institutes of Health (NIH) Common Fund via the NIH Director's Early Independence Award (DP5 OD028125) to J.R.B and Departmental Funding from the Department of Pathology and Immunology at Washington University.

Conflicts of Interest: NB is a consultant for Santa Ana Bio, Inc. The other authors have no conflicts of interest relating to the work described in this manuscript.

References

Atchley,W.R. *et al.* (2005) Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, **102**, 6395–6400.
Beshnova,D. *et al.* (2020) De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci Transl Med*, **12**, eaaz3738.
Borcherding,N. *et al.* (2020) scRepertoire: An R-based toolkit for single-cell immune receptor analysis. *F1000Res*, **9**, 47.
Drost,F. *et al.* (2022) Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis. 2021.06.24.449733.
Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573-3587.e29.
Kidera,A. *et al.* (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem*, **4**, 23–55.
Lun, Aaron (2022) mumosa: Multi-Modal Single-Cell Analysis Methods.
Mason,D.M. *et al.* (2021) Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng*, **5**, 600–612.
Odegard,V.H. and Schatz,D.G. (2006) Targeting of somatic hypermutation. *Nat Rev Immunol*, **6**, 573–583.

Ramaswamy,A. *et al.* (2021) Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children. *Immunity*, **54**, 1083-1095.e7.
Schattgen,S.A. *et al.* (2022) Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat Biotechnol*, **40**, 54–63.
Schatz,D.G. and Ji,Y. (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol*, **11**, 251–263.
Sidhom,J.-W. *et al.* (2022) Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. *Sci Adv*, **8**, eabq5089.
Vu,M.H. *et al.* (2022) ImmunoLingo: Linguistics-based formalization of the antibody language.
Zhang,Z. *et al.* (2021) Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat Methods*, **18**, 92–99.

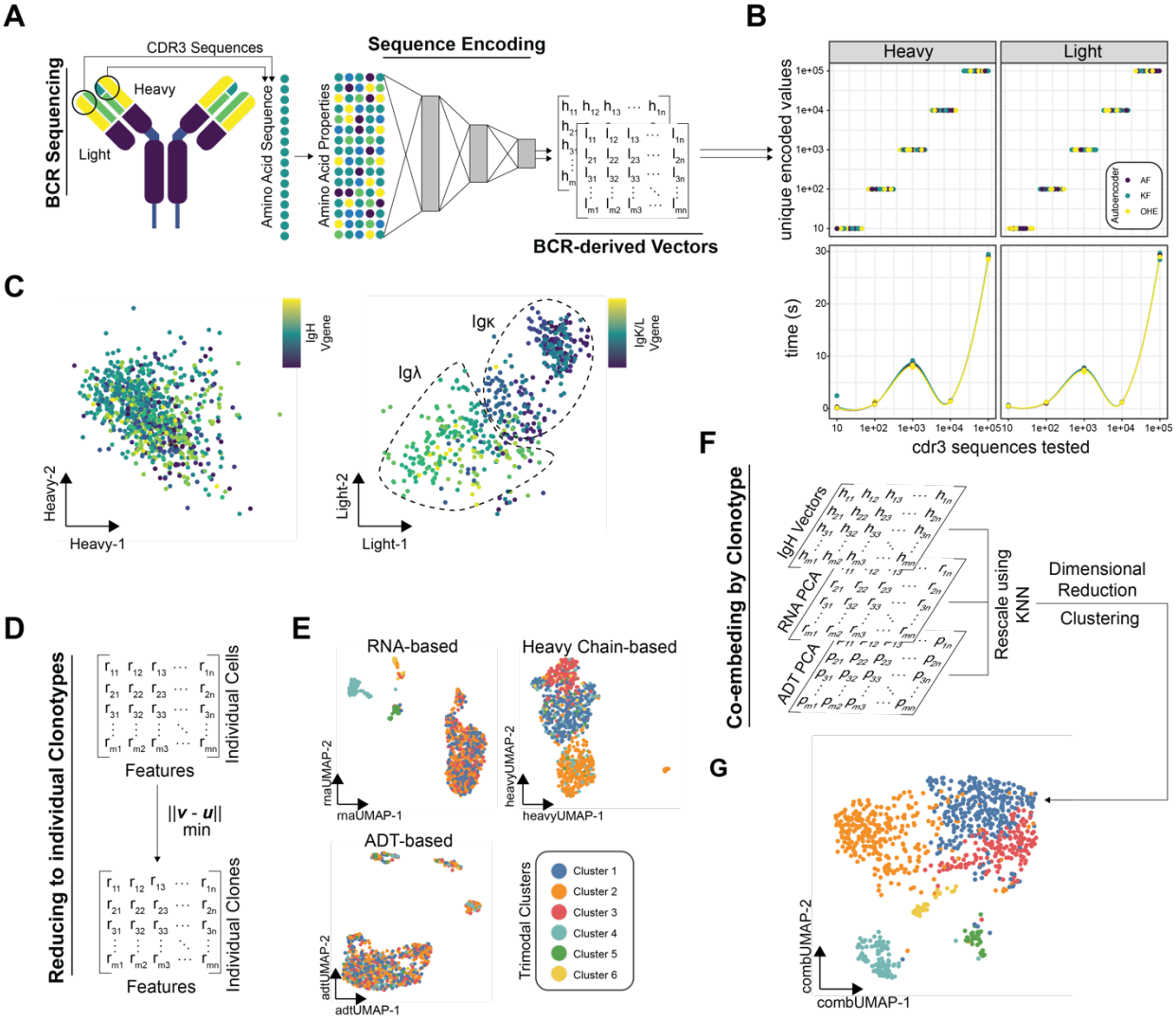

Figure 1: (**A**) Conceptual schematic for the autoencoding approach for the BCR sequences. (B) Benchmarking of the autoencoders in Ibex using novel cdr3 sequences for heavy and light chain in terms of unique latent dimensions returned (upper panels) and time for encoding (bottom panels). Colors are based on the encoding approach, Atchley factors (AF), Kidera factors (KF), and one-hot encoding (OHE). (**C**) Visualization of the latent space dimensions for the heavy (left) and light (right) chains. Colors indicate differing V genes associated with the respective chain. (**D**) Reduction of individual cells to individual clonotypes by selecting cells with the minimal Euclidean distance in principal component vectors across clonotypes. (**E**) Uniformed manifold approximation and projection (UMAP) using RNA, ADT-level protein quantification, and

Ibex-derived heavy chain sequences. Clustering based on trimodal integrated values. (**F**) Approach for the multimodal integration of all three data types through knn rescaling. (**G**) Resulting UMAP from the normalized trimodal integrated values.