

Genetics and population analysis

immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking

Cédric R. Weber ¹, Rahmad Akbar², Alexander Yermanos¹, Milena Pavlović³, Igor Snapkov², Geir K. Sandve³, Sai T. Reddy^{1,*} and Victor Greiff ^{2,*}

¹Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland, ²Department of Immunology, University of Oslo, 0372 Oslo, Norway and ³Department of Informatics, University of Oslo, 0373 Oslo, Norway

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 9, 2019; revised on February 3, 2020; editorial decision on March 2, 2020; accepted on March 4, 2020

Abstract

Summary: B- and T-cell receptor repertoires of the adaptive immune system have become a key target for diagnostics and therapeutics research. Consequently, there is a rapidly growing number of bioinformatics tools for immune repertoire analysis. Benchmarking of such tools is crucial for ensuring reproducible and generalizable computational analyses. Currently, however, it remains challenging to create standardized ground truth immune receptor repertoires for immunoinformatics tool benchmarking. Therefore, we developed immuneSIM, an R package that allows the simulation of native-like and aberrant synthetic full-length variable region immune receptor sequences by tuning the following immune receptor features: (i) species and chain type (BCR, TCR, single and paired), (ii) germline gene usage, (iii) occurrence of insertions and deletions, (iv) clonal abundance, (v) somatic hypermutation and (vi) sequence motifs. Each simulated sequence is annotated by the complete set of simulation events that contributed to its *in silico* generation. immuneSIM permits the benchmarking of key computational tools for immune receptor analysis, such as germline gene annotation, diversity and overlap estimation, sequence similarity, network architecture, clustering analysis and machine learning methods for motif detection.

Availability and implementation: The package is available via <https://github.com/GreiffLab/immuneSIM> and on CRAN at <https://cran.r-project.org/web/packages/immuneSIM>. The documentation is hosted at <https://immuneSIM.readthedocs.io>.

Contact: sai.reddy@ethz.ch or victor.greiff@medisin.uio.no

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Targeted deep sequencing of adaptive immune receptor repertoires (AIRR-seq data, [Breden *et al.*, 2017](#)) has become a key resource for immunodiagnostics and immunotherapeutics research. Consequently, there exists a rapidly growing number of immune receptor informatics tools for germline gene annotation, diversity and overlap estimation, network architecture (sequence similarity) and machine learning analysis ([Brown *et al.*, 2019](#); [Yaari and Kleinstein, 2015](#)). To benchmark and assess the performance of these tools, synthetic ground truth immune receptor datasets with complete information on all repertoire feature dimensions investigated or used in these tools (e.g. germline gene usage, insertion and deletions, and clonal abundance; [Fig. 1](#)) are required ([Brown *et al.*, 2019](#); [Yaari and Kleinstein, 2015](#)). Therefore, there is a need for a computational framework that enables the simulation of native-like immune

receptor repertoires as well as repertoires that differ in single- or multiple feature dimensions while simultaneously allowing for the tracing of all immunologically relevant simulation parameters. To address this gap in the landscape of immune receptor simulation tools that are focused predominantly on generating native-like repertoires ([Marcou *et al.*, 2018](#); [Safonova *et al.*, 2015](#); [Yermanos *et al.*, 2017](#)), we here present the immuneSIM R package, which allows the tunable multi-feature simulation of human and mouse BCR and TCR repertoires (single-chain and paired full-length variable regions) with traceable simulation event-level annotation for each of the simulated sequences.

The user has full control over the following immunological features: V-, D-, J-germline gene set and usage, occurrence of insertions and deletions, clonal sequence abundance and somatic hypermutation. Post-sequence simulation, the generated immune receptor sequences may be further altered by the addition of custom sequence

- Brown,A.J. *et al.* (2019) Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Mol. Syst. Des. Eng.*, **4**, 701–736.
- Dash,P. *et al.* (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, **547**, 89–93.
- Emerson,R.O. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, **49**, 659–665.
- Giudicelli,V. and Lefranc, MP. (2011) IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb. Protoc.*, **2011**, 716–725.
- Glanville,J. *et al.* (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature*, **547**, 94–98.
- Greiff,V. *et al.* (2017) Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.*, **199**, 2985–2997.
- Marcou,Q. *et al.* (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, **9**, 561.
- Safonova,Y. *et al.* (2015) IgSimulator: a versatile immunosequencing simulator. *Bioinformatics*, **31**, 3213–3215.
- Yaari,G. and Kleinstein,SH. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.*, **7**, 121.
- Yermanos,A. *et al.* (2017) Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics*, **33**, 3938–3946.