Bioinformatics, 31(20), 2015, 3356–3358
doi: 10.1093/bioinformatics/btv359
Advance Access Publication Date: 10 June 2015
Applications Note

Sequence analysis

Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data

Namita T. Gupta[1,†], Jason A. Vander Heiden[1,†], Mohamed Uduman[2], Daniel Gadala-Maria[1], Gur Yaari[3] and Steven H. Kleinstein[1,2,*]

[1]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA, [2]Department of Pathology, Yale University School of Medicine, New Haven, CT 06511, USA and [3]Bioengineering Program, Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on October 17, 2014; revised on April 30, 2015; accepted on June 5, 2015

Abstract

Summary: Advances in high-throughput sequencing technologies now allow for large-scale characterization of B cell immunoglobulin (Ig) repertoires. The high germline and somatic diversity of the Ig repertoire presents challenges for biologically meaningful analysis, which requires specialized computational methods. We have developed a suite of utilities, Change-O, which provides tools for advanced analyses of large-scale Ig repertoire sequencing data. Change-O includes tools for determining the complete set of Ig variable region gene segment alleles carried by an individual (including novel alleles), partitioning of Ig sequences into clonal populations, creating lineage trees, inferring somatic hypermutation targeting models, measuring repertoire diversity, quantifying selection pressure, and calculating sequence chemical properties. All Change-O tools utilize a common data format, which enables the seamless integration of multiple analyses into a single workflow.
Availability and implementation: Change-O is freely available for non-commercial use and may be downloaded from http://clip.med.yale.edu/changeo.
Contact: steven.kleinstein@yale.edu

1 Introduction

Large-scale characterization of immunoglobulin (Ig) repertoires is now feasible due to dramatic improvements in high-throughput sequencing technology. Repertoire sequencing is a rapidly growing area, with applications including detection of minimum residual disease, prognosis following transplant, monitoring vaccination responses, identification of neutralizing antibodies and inferring B cell trafficking patterns (Robins, 2013; Stern *et al.*, 2014). We previously developed the repertoire sequencing toolkit (pRESTO) for producing assembled and error-corrected reads from high-throughput lymphocyte receptor sequencing experiments (Vander Heiden *et al.*, 2014), which may then be fed into existing methods for alignment against V(D)J germline databases [e.g. IMGT/HighV-QUEST (Alamyar *et al.*, 2012), IgBLAST (Ye *et al.*, 2013), iHMMune-align (Gaëta *et al.*, 2007)]. However, extracting measures of biological and clinical interest from the resulting germline-annotated repertoire remains a time-consuming and error-prone process that is often dependent upon custom analysis scripts. Here, we introduce Change-O, a suite of utilities that cover a range of complex analysis tasks for Ig repertoire sequencing data.

2 Features

The Change-O suite is composed of four software packages: a collection of Python commandline tools (changeo-ctl) and three separate R (R Core Team, 2015) packages (alakazam, shm, and tigger)

Table 1. Summary of Change-O features

Package	Analysis tasks
changeo-clt	Parsing of V(D)J assignment output
	Basic database manipulation
	Multiple alignment of sequence records
	Assignment of sequences into clonal groups
	Calculation of CDR3 physiochemical properties
alakazam	Clonal diversity analysis
	Lineage reconstruction
shm	SHM hot/cold-spot modeling
	Quantification of selection pressure
tigger	Inference of novel germline alleles
	Construction of personalized germline genotype

(Table 1). Data are passed to Change-O utilities in the form of a tab-delimited text file. Each utility identifies the relevant input data based on standardized column names and adds new columns to the file with the output information to be carried through to the next analysis step. Change-O provides tools to import data from the frequently used IMGT/HighV-QUEST (Alamyar *et al.*, 2012) tool as well as a set of utilities to perform basic database operations, such as sorting, filtering and modifying annotations.

The more computationally expensive components have built-in multiprocessing support. Each utility includes detailed help documentation and optional logging to track errors. Example workflow scripts are provided on the website, which can easily be modified by adding, removing or reordering analysis steps to meet different analysis goals. As detailed later, several repertoire analyses may be carried out, depending on the nature of the study.

2.1 Inference of novel alleles and individual genotype

Germline segment assignment tools, such as IMGT/HighV-QUEST, work by aligning each sequence against a database of known alleles. However, this process is inaccurate for sequences that utilize previously undetected alleles. In this case, the sequence will be assigned to the closest known allele and any polymorphisms will be incorrectly identified as somatic mutations. To address this problem, the Tool for Immunoglobulin Genotype Elucidation (TIgGER) (Gadala-Maria *et al.*, 2015) has been implemented as an R package for inclusion in Change-O. TIgGER determines the complete set of variable region gene segments carried by an individual and identifies novel alleles, yielding a set of germline alleles personalized to an individual. The germline variable region allele assignments are then adjusted based on this individual Ig genotype. This process significantly improves the quality of germline assignments, thus increasing the confidence of downstream analysis dependent upon mutation profiles.

2.2 Partitioning sequences into clonally related groups

Identifying sequences that are descended from the same B cell (clonal groups) is important to virtually all Ig repertoire analyses. Clonal group sizes and lineage structures provide information on the underlying response, and clonally related sequences cannot be treated independently in statistical analyses and models. Change-O provides several methods for partitioning sequences into clones. Along with published methods based on hierarchical clustering (Ademokun *et al.*, 2011; Chen *et al.*, 2010; Glanville *et al.*, 2009), users also have the option to employ several published somatic hypermutation (SHM) hot/cold-spot targeting models as distance metrics in the clustering methods (Smith *et al.*, 1996; Yaari *et al.*, 2013; Stern *et al.*, 2014). Users may alter the clustering thresholds, and Change-O also includes tools to tune the thresholds based on distance patterns in the repertoire (Glanville *et al.*, 2009).

2.3 Quantification of repertoire diversity

To assess repertoire diversity, Change-O provides an implementation of the general diversity index (${}^qD$) proposed by Hill (1973), which encompasses a range of diversity measures as a smooth curve over a single varying parameter q. Special cases of this general index of diversity correspond to the most popular diversity measures: species richness ($q=0$), the exponential Shannon-Weiner index (as $q \to 1$), the inverse of the Simpson index ($q=2$), and the reciprocal abundance of the largest clone (as $q \to \infty$). Resampling strategies are also provided to perform significance tests and allow comparison across samples with varying sequencing depth (Wu *et al.*, 2014; Stern *et al.*, 2014).

2.4 Generation of B cell lineage trees

Lineage trees provide a means to trace the ancestral relationships of cells within a clone. This information has been used to estimate mutation rates (Kleinstein *et al.*, 2003), infer B cell trafficking patterns (Stern *et al.*, 2014) and trace the accumulation of mutations that drive affinity maturation (Uduman *et al.*, 2014; Wu *et al.*, 2012). Change-O provides a tool for generating lineage trees using `PHYLIP`'s maximum parsimony algorithm (Felsenstein, 1989), with modifications to meet the requirements of an Ig lineage tree (Barak *et al.*, 2008; Stern *et al.*, 2014). Trees may be viewed and exported into different file formats using the igraph (Csardi and Nepusz, 2006) R package.

2.5 Somatic hypermutation hot/cold-spot motifs

SHM is a process that operates in activated B cells and introduces point mutations into the DNA coding for the Ig receptor at a very high rate ($\approx 10^{-3}$ per base-pair per division) (Kleinstein *et al.*, 2003; McKean *et al.*, 1984). Accurate background models of SHM are critical, since SHM displays intrinsic hot/cold-spot biases (Yaari *et al.*, 2013). Change-O provides utilities for estimating the mutability and substitution rates of DNA motifs from large-scale Ig sequencing data to construct hot/cold-spot motif models. Furthermore, models may be generated based solely on silent mutations, thereby avoiding the confounding influence of selection pressures (Yaari *et al.*, 2013). These tools can be used to build models of SHM targeting and gain insight into the relative contributions of different error-prone repair pathways in SHM.

2.6 Analysis of selection pressure

For quantifying selection pressure in Ig sequences, Change-O includes the BASELINe (Yaari *et al.*, 2012) method, which has been implemented as an R package for inclusion in the suite. BASELINe quantifies deviations in the frequency of replacement mutations compared with a background model of SHM. Users may choose between published background models (Smith *et al.*, 1996; Yaari *et al.*, 2013) or infer the background from their own data using the SHM model building tools described above.

3 Conclusion

Change-O is a suite of utilities implementing a wide range of B cell repertoire analysis methods. Together these tools allow researchers to quickly implement advanced analysis pipelines for large datasets generated by repertoire sequencing experiments. A simple tab-delimited file with standardized column names allows for

communication between the utilities and can easily be viewed using any spreadsheet application. This format also allows research groups the flexibility to incorporate other analysis tools into their in-house analysis pipelines by simply adding additional columns of information to the central file. Change-O, along with pRESTO (Vander Heiden *et al.*, 2014), provides key components of an analytical ecosystem that enables sophisticated analysis of high-throughput Ig repertoire sequencing datasets.

Acknowledgements

The authors thank the Yale University Biomedical High Performance Computing Center [funded by National Institutes of Health grants RR19895 and RR029676-01] for use of their computing resources. The authors also thank Chris Bolen, Moriah Cohen, Jingli Shan and Sonia Timberlake for testing Change-O and providing helpful feedback.

Funding

This work was supported by the National Institutes of Health [R01AI104739 to S.H.K.; T15LM07056 to N.T.G., T15LM07056 to J.A.V.H. from National Library of Medicine (NLM)] and by the United States-Israel Binational Science Foundation [2013395 to G.Y. and S.H.K.].

Conflict of Interest: none declared.

References

Ademokun,A. *et al.* (2011) Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging cell*, **10**, 922–930.

Alamyar,E. *et al.* (2012) IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol. Biol.*, **882**, 569–604.

Barak,M. *et al.* (2008) IgTree: creating immunoglobulin variable region gene lineage trees. *J. Immunol. Methods*, **338**, 67–74.

Chen,Z. *et al.* (2010) Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res.*, **6**(Suppl. 1), S4.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Felsenstein,J. (1989) PHYLIP - Phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

Gadala-Maria,D. *et al.* (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. USA*, **112**, 201417683.

Gaëta,B. a. *et al.* (2007) iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, **23**, 1580–1587.

Glanville,J. *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA*, **106**, 20216–20221.

Hill,M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427.

Kleinstein,S.H. *et al.* (2003) Estimating hypermutation rates from clonal tree data. *J. Immunol.*, **171**, 4639–4649.

McKean,D. *et al.* (1984) Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl. Acad. Sci. USA*, **81**, 3180–3184.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins,H. (2013) Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.*, **25**, 646–652

Smith,D. *et al.* (1996) Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J. Immunol.*, **156**, 2642–2652.

Stern,J.N.H. *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.*, **6**, 248ra107.

Uduman,M. *et al.* (2014) Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J. Immunol.*, **192**, 867–874.

Vander Heiden,J.A. *et al.* (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, **30**, 1930–1932

Wu,Y.-C.B. *et al.* (2012) Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front. Immunol.*, **3**, 193.

Wu,Y.-C.B. *et al.* (2014) Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *J. Allergy Clin. Immunol.*, **134**, 604–612.

Yaari,G. *et al.* (2012) Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.*, **40**, e134.

Yaari,G. *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.*, **4**, 358.

Ye,J. *et al.* (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, **41**(Web Server Issue), W34–W40.