Bioinformatics Advance Access published September 11, 2016

Sequence Analysis

ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data

Aaron M. Rosenfeld [1], Wenzhao Meng [2], Eline T. Luning Prak [2], Uri Hershberg [1,3,*]

[1] School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, 19104, USA
[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104
[3] Department of Microbiology and Immunology, College of Medicine, Drexel University, Philadelphia, PA, 19129, USA

*To whom correspondence should be addressed.

Associate Editor: Dr. Inanc Birol

Abstract

Summary: As high-throughput sequencing of B cells becomes more common, the need for tools to analyze the large quantity of data also increases. This paper introduces ImmuneDB, a system for analyzing vast amounts of IGHV sequences and exploring the resulting data. It can take as input raw FASTA/FASTQ data, identify genes, determine clones, construct lineages, as well as provide information such as selection pressure and mutation analysis. It uses an industry leading database, MySQL, to provide fast analysis and avoid the complexities of using error prone flat-files.
Availability:
ImmuneDB is freely available at http://immunedb.com
A demo of the ImmuneDB web interface is available at: http://immunedb.com/demo
Contact: uh25@drexel.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

This paper introduces ImmuneDB, a system for analysing, storing, and exploring immune repertoire data, focused specifically on B-cell heavy chain variable region (IGHV) sequences from high-throughput sequencing. It reduces ad-hoc scripting, avoids flat-files in favor of a database, and provides a web-based analysis interface. The system has been tested on a dataset of 100 million sequences and required 48 hours to process on a 12-core Intel i7 with 64 gigabytes of memory. ImmuneDB is provided both as a Python package and as a set of Docker images interconnected by Docker Compose.

2 Workflow

ImmuneDB is a set of parallelized Python command line programs which interact with an underlying MySQL database. Each step of the analysis workflow builds upon data generated in the previous steps and is initiated with a different command. As shown in Supplementary Figure 1, some stages of the workflow may be optionally replaced with other software. For example, HighV-Quest (Alamyar *et al.*, 2012) may be used for gene identification, while still using ImmuneDB for later stages.

2.1 Variable & Joining Region Gene Identification

Provided by ImmuneDB is an implementation of the V and J identification method proposed in (Zhang *et al.*, 2015), using V and J gene anchoring. It takes as input raw FASTA or FASTQ files with full or partial IGHV sequences, optionally pre-processed by pRESTO (Vander Heiden *et al.*, 2014). Insertions and deletions are detected and flagged, but not corrected in this step. ImmuneDB can alternatively import pre-identified sequences from HighV-Quest (Alamyar *et al.*, 2012) or CSV files with IGMT aligned sequences and adequate metadata. For either method, germlines are specified by the user and are not pre-defined.

Whether identified by ImmuneDB or imported from another system, additional information is extracted from the IGHV sequences including the identity to germline, the third complementarity determining region (CDR3) length, and if they are productive.

ImmuneDB incorporates the notion of gene ties. After all sequences in a sample are identified, their average mutation rate and length is used to determine which genes are statistically indistinguishable from each other.

2.2 Optional Local Alignment

ImmuneDB can utilize the Needleman-Wunsch alignment method (Needleman and Wunsch, 1970) to correct insertions and deletions in sequences and to identify sequences with mutations in conserved regions necessary for anchoring. Each J gene is aligned to the query sequence and, if an alignment is found, the same process is applied to each V gene. The position(s) of all insertions and deletions are stored along with the sequence in the database.

2.3 Sequence Collapsing

After sequences are identified, they are collapsed within each subject. Collapsing occurs when two sequences are identical excluding positions where either has an unknown nucleotide (indicated with an "N"). The sequence with the higher copy number within its sample serves as the representative sequence, and maintains the collapsed copy number.

2.4 Clonal Assignment & Lineages

Clonal assignment aggregates sequences into groups such that all sequences in a given group likely share a common ancestor. ImmuneDB iterates over each sequence, from highest to lowest copy-number. For each, it finds the largest clone with the same V gene, J gene, and CDR3 length in nucleotides such that all sequences already belonging to that clone differ in CDR3 amino-acid sequence by no more than 15% (Zhang *et al.*, 2015). This percentage can be changed by the user along with other restrictions. Alternatively, users may import their own clonal assignment.

Following clonal assignment, a lineage tree may be constructed for each clone. ImmuneDB uses clearcut (Sheneman *et al.*, 2006) as a basis for neighbor-joining and additionally annotates each branching point with its associated mutations, sequences present, and metadata. Users can modify the specifics of tree construction by specifying which mutations and sequences should be included.

2.5 Statistics & Mutation Analysis

Statistics are calculated for every sample and for every clone. For samples, the distributions of gene usage, region length, germline identity, and copy number are determined for all sequences, unique sequences, unique sequences with a copy number greater than one, and clones.

For clone statistics, mutations from the germline are calculated for each gene region. Using this information, selection pressure is calculated with BASELINe (Yaari *et al.*, 2012) for the entire clone as well as each sample in which the clone exists.

2.6 APIs & Web Interface

ImmuneDB includes a REST API, providing a set of HTTP endpoints for common read-only queries that can be accessed in a language-agnostic manner. For more complex interactions with ImmuneDB, the included Python API can be used.

A web interface to ImmuneDB is provided, allowing users to interactively explore data after analysis with the pipeline. Most data including raw sequences, metadata, mutation analysis, and clonal assignments can be easily downloaded in a variety of formats including FASTA, FASTQ, and CSV allowing further processing with external tools.

For analyzing large amounts of data, users can perform queries on the data to analyze sequences and clones across subjects, tissues, and other attributes. Clones and sequences can be further filtered based on if they are functional, and sequences can be filtered based on if they are unique and their copy number.

3 Comparison to IMGT HighV-Quest

ImmuneDB's gene identification process was compared to IMGT HighV-Quest. Overall, ImmuneDB quickly identifies a reasonably high number of sequences compared to HighV-Quest (about 75%) without local alignment. After local alignment, which took 4 hours, ImmuneDB identified all but 58 (out of 530,104) of the sequences of which HighV-Quest did but identified 3,684 sequences that HighV-Quest did not. Further, ImmuneDB requires no remote processing — everything can be run on local hardware. Finally, custom germline files can be used for identification, a feature that is missing with HighV-Quest. A more detailed analysis is included in Supplementary Text 1.

4 Conclusion

This paper has introduced ImmuneDB, a package to assist in the analysis of high-throughput B-cell IGHV sequences. It provides a single integrated platform of common analysis tools including gene identification, clonal assignment, lineage construction, and statistics aggregation. After analysis, resulting data can then be easily visualized, queried, and exported through its web-based interface. The ImmuneDB package is freely available at http://immunedb.com.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number P01AI106697 and by NIH P30-CA016520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and Lefranc1, M.-P. (2012). Imgt/highv-quest: The imgt web portal for immunoglobulin (ig) or antibody and t cell receptor (tr) analysis from ngs high throughput and deep sequencing. *Immunome Research*, **8**(1), –.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443 – 453.

Sheneman, L., Evans, J., and Foster, J. A. (2006). Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, **22**(22), 2823–2824.

Vander Heiden, J., Yaari, G., Uduman, M., Stern, J., O'Connor, K., Halfer, D., Vigneault, F., and Kleinstein, S. (2014). presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires (tech1p. 863). *The Journal of Immunology*, **192**(1 Supplement), 69–31.

Yaari, G., Uduman, M., and Kleinstein, S. H. (2012). Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Research*, **40**(17), e134.

Zhang, B., Meng, W., Prak, E. T. L., and Hershberg, U. (2015). Discrimination of germline v genes at different sequencing lengths and mutational burdens: A new tool for identifying and evaluating the reliability of v gene assignment. *Journal of Immunological Methods*, **427**, 105 – 116.