Example Analysis Pipeline

This example illustrates the use of the Trigs toolset and third-party tools in a recent analysis.

In this analysis, post-vaccination plasma samples were taken from a single rabbit. PBMCs were prepared for Rep-seq analysis using PCR amplification. Samples were sequenced on an Illumina MiSeq producing 300x300 bp paired-end reads covering the entire variable region (manuscript in preparation). Fig. 1 shows the subsequent analysis pipeline: the numbers on the figure refer to descriptions below.

1 – Quality Control and Junction Analysis

Reads from the two sequencing runs were demultiplexed by Illumina software, following which they were pre-processed using the Repertoire Sequencing Toolkit (pRESTO) [1]. Quality-checked heavy and kappa chain reads (identified from primer matches) were written to separate FASTA files. Lambda chain reads were not processed in this analysis.

To minimise the impact of residual sequencing errors on downstream processing, resulting read sets were clustered to a minimum identity of 97% using uparse [2], following which the sequences were parsed with IgBLAST [3] using the IMGT germline library for the rabbit (Oryctolagus cuniculus) [4].

The output from IgBLAST was converted to tab-separated analysis format (similar to IMGT output) by **`IgBLASTPlus`**, and a Linux sed script was used to merge the files from each sample, creating consolidated heavy and kappa chain analysis files. The sed script modified the Sequence IDs to include a two letter sample code, used in downstream processing to distinguish sequences derived from each sample.

2 – Clonal Analysis

Representative mAbs were isolated from the samples by hybridoma, and characterized in terms of their specificity. The heavy chain junction sequences of these mAbs were merged with heavy chain junction sequences from the NGS analysis, extracted from the analysis file by **`ExtractFromIMGT`**. **`NeighbourDist`** was used to plot the distribution of nearest neighbours. Such plots typically show two peaks: the first peak is taken to reflect the distance distribution of clonally-related sequences, and the trough between the two peaks provides an indicative threshold for the clustering of clonally related sequences [5]. For performance reasons, the analysis was run on a number of random samples of size 75,000 records using the **`-l`** option of **`NeighbourDist`**, and the results were compared to confirm consistency.

To obtain an overview of the heavy chain clonal relationships, **`ClusterSeqs`** was used to determine clusters of the junction sequences, using the cutoff threshold inferred from **`NeighbourDist`**. Resulting clusters were processed by **`ClusterGraph`** in order to produce data that can be rendered by Gephi [6], providing a plot in which each sequence is represented by a point, colour coded by sample ID, and joined to nearest neighbours by a line. Gephi's Yifan Hu layout algorithm was used to lay out data in the plot.

Researchers have not as yet reached consensus on the approach to clustering that best reflects the underlying process of clonal development. While we consider that the approach above provides a good performance tradeoff for the overall analysis of a repertoire consisting of several million reads, the pipeline has been designed to facilitate adoption of different methods. In particular, CD-HIT [7], which provides a more conservative clustering algorithm, can be used as a direct replacement for ClusterSeqs. Another alternative is CHANGE-O [8], which implements a number of published methods.

3 - Phylogenetic Analysis of Selected Families

To gain insights into the development of mAbs of interest, the analysis records of NGS heavy chain junction sequences clustering with their junction sequences were obtained using **`ClusterExtract`**. Full-length nucleotide sequences of those whose inferred V- and J- germlines matched the mAB were extracted using **`ExtractFromIMGT`**. These sequences were codon-aligned using TranslatorX [9] and MUSCLE [10]. Down-sampled data sets were created with **`FastaSample`** and phylogenetic trees were inferred with IQ-Tree and plotted in Python using the ETE Toolkit. Downsampling was used both for performance reasons (the largest sequence set had >30,000 records) and also to provide a more even comparison between timepoints.

4 – Germline Analysis and CDR3 Length Distribution

Plots were obtained directly from **`PlotGermline`** and **`SpectraType`**. These tools are capable of producing plots for multiple samples at once, using the sample coding embedded in the Sample ID.

Bibliography

1. Vander Heiden, J. A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinforma. Oxf. Engl.* **30,** 1930–1932 (2014).
2. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinforma. Oxf. Engl.* **26,** 2460–2461 (2010).
3. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41,** W34-40 (2013).
4. Lefranc, M.-P. & Lefranc, G. *The Immunoglobulin FactsBook*. (Academic Press, 2001).
5. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5,** 171ra19 (2013).
6. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. in *Third International AAAI Conference on Weblogs and Social Media* (2009). at <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>
7. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28,** 3150–3152 (2012).
8. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinforma. Oxf. Engl.* (2015). doi:10.1093/bioinformatics/btv359
9. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38,** W7-13 (2010).
10. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32,** 1792–1797 (2004).