IMMUNE-NETWORK-GENERATION PROGRAM DOCUMENTATION

Program developed by Rachael Bashford-Rogers (2013)
At the Wellcome Trust Sanger Institute under Prof. Paul Kellam
(rb520@cam.ac.uk)

1. Introduction

To date next-generation sequencing (NGS) of BCRs have primarily focused on classifying the IgHV, D and J recombination frequencies to understand the diversity of the BCR repertoire (*1-7*). However, computational assignment of V-D-J sequences to reference databases results in many incompletely assigned IgHV, D and J genes even when the germline alleles are known (*8*). This is most likely due to SHM masking the identity of the germline genes present in the NGS, or the existence of allelic variation relative to the reference IgH genes. Further, investigation of V-D-J gene usage frequencies utilizes only part of the BCR sequence diversity with important information about the V-D-J joining regions and mutational relationships not considered.

Here we propose that analysis of the BCR sequence relationships using the full BCR V-D-J sequence is more informative for human BCR repertoire analysis than V-D-J gene classification. We show that human BCR repertoire diversity can be interpreted through full V-D-J genotype diversity using BCR networks, previously shown to be an intuitive way for understanding B-cell repertoires in zebrafish (*8*). In such networks, the lowest level of organisation in a population of B-cells, namely independent B-cells, is represented by sparse networks whereas highly developed (connected) networks most likely result from clonal expansions of B-cells, arising through antigenic exposure or B-cell malignancies (*8*).

Here, each vertex represents a different sequence, and the number of identical BCR sequences defines the vertex size. Edges are created between vertices that differ by one nucleotide. Clusters are groups of interconnected vertices (*9*). The program described here calculates edges between unique sequences and determines vertex sizes, creating output files in formats that can directly be used in network analysis programs such as networkx (python) or igraph (R or python).

2. User’s guide

a) Installation

Install CD-HIT(*10*):

- Download current CD-HIT at http://bioinformatics.org/cd-hit
- Unpack the file with “tar xvf cd-hit-XXX.tar.gz --gunzip”
- Change dir by “cd cd-hit-2006”
- Compile the programs by “make”

Download Immune-Network-Generation program (python script) and set up to user space:

- Change directory for CD-HIT program to correct location:
  line #36 cd_hit_directory = “cd-hit-v4.5.4-2011-03-07/"
- Install the following python module dependencies:
  sys, collections, os, operator, networkx

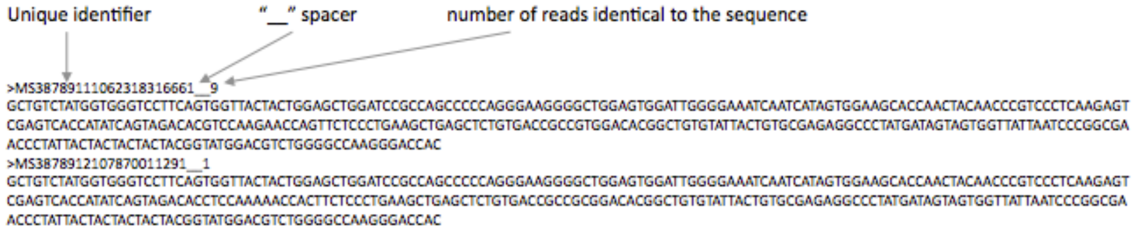
b) Usage

```
python Generate_networks_global.py <fasta file> <sample id> <output directory>
```

Arguments:

<fasta file>	Fasta file of sequences for network analysis.*
<sample id>	Unique sample ID.
<output directory>	The directory in which the output files will be created.

* The sequence IDs should be in multiplicity format:


Unique identifier
“__” spacer
number of reads identical to the sequence

```
>MS38789111062318316661__9
GCTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCCCTCAAGAGT
CGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCCGTGGACACGGCTGTGTATTACTGTGCGAGAGGCCCTATGATAGTAGTGGTTATTAATCCCGGCGA
ACCCTATTACTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCAC
>MS3878912107870011291__1
GCTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCCCTCAAGAGT
CGAGTCACCATATCAGTAGACACCTCCAAAAACCACTTCTCCCTGAAGCTGAGCTCTGTGACCGCCGCGGACACGGCTGTGTATTACTGTGCGAGAGGCCCTATGATAGTAGTGGTTATTAATCCCGGCGA
ACCCTATTACTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCAC
```

c) Example

Run:

```
> python Generate_networks_global_2.0.py TEST_FILES/Sequences_Example2.fasta
Example2 OUTPUT_TEST_FILES/
```

d) Output files

Table 1. Description of output files.

File	Description	Format
Att_SAMPLE.txt	List of unique sequences.	Column 1 = List of unique sequence ids; column 2 = number of reads; column 3 = sequence;
Cluster_identities_SAMPLE.txt	List of sequences within clusters.	Column 1 =sequence number; column 2 = cluster number; column 3 = sequence ID; column 4 = number of reads;
Fully_reduced_SAMPLE.fasta	Fasta file of unique sequences.	Sequence ID multiplicity format;
Plot_ids_SAMPLE.txt	List of sequences for plotting (only sequences that are connected or representing >1 read).	Column 1 =sequence number; column 2 = sequence ID; column 3 = number of reads;

2. References:

If you find Immune-Network-Generation useful, please cite reference #9 (*R. J. Bashford-Rogers et al., 2013*).

1. S. D. Boyd *et al.*, Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *Journal of immunology* **184**, 6986 (Jun 15, 2010).
2. S. D. Boyd *et al.*, Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine* **1**, 12ra23 (Dec 23, 2009).
3. P. J. Campbell *et al.*, Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13081 (Sep 2, 2008).
4. M. L. Sanchez *et al.*, Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. *Blood* **102**, 2994 (Oct 15, 2003).
5. C. Maletzki *et al.*, Ex-vivo clonally expanded B lymphocytes infiltrating colorectal carcinoma are of mature immunophenotype and produce functional IgG. *PloS one* **7**, e32639 (2012).
6. A. Lev *et al.*, The kinetics of early T and B cell immune recovery after bone marrow transplantation in RAG-2-deficient SCID patients. *PloS one* **7**, e30494 (2012).
7. A. R. Wu *et al.*, Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* **11**, 41 (Jan, 2014).
8. J. A. Weinstein, N. Jiang, R. A. White, 3rd, D. S. Fisher, S. R. Quake, High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807 (May 8, 2009).
9. R. J. Bashford-Rogers *et al.*, Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome research* **23**, 1874 (Nov, 2013).
10. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658 (Jul 1, 2006).