HW 1, basic analysis of antibody sequences

The purpose of this assignment is to perform basic analysis of antibody sequences with different properties. First, you need to download three samples: NAIVE, MEMORY, and PLASMA from Google Drive. Each sample represents a fragment of a real Rep-seq library taken from the corresponding type of B cells. For each sample, you need to compute alignment against known V and J genes (or *germline genes*) of all sequences. Alignment can be done using one of the following tools:

- IgBlast: https://www.ncbi.nlm.nih.gov/igblast/igblast.cgi (has both command line and online versions).
- DiversityAnalyzer: https://immunotools.github.io/immunotools/ (available as a command line version only; works on Linux and MacOS).

Note that these tools can produce slightly different results because of the difference in alignment algorithms. We will not take these differences into account during the grading. However, we ask you to specify the chosen tool in the report.

For each sample,

1. Analyze the joint usage of V and J genes: for each sequence find the closest germline V and J genes, list of all VJ pairs occurring in the sample, and create a plot (e.g., heatmap) showing the number of sequences for each VJ pair.

NOTE: the same gene can be presented by several allelic variations. E.g., IGHV1-18*01 and IGHV1-18*02 are variations of the same V gene IGHV1-18.

2. Find 10 most used V genes in the sample and analyze their *mutability*. For each gene, analyze sequences aligned to it and compute the number of differences in each alignment. Mutability is the distribution of the number of differences. Visualize mutability of 10 most used V genes in any convenient form (e.g., using boxplot).

3. Visualize distributions of CDR3 lengths.

4. Compute the fraction of non-productive sequences in the sample. Both IgBlast and DiversityAnalyzer report productiveness of input sequences as a part of the output.

For three samples:

1. Compare VJ usages in three samples and find samples with the smallest and highest number of VJ combinations.

2. Compare distributions of CDR3 lengths and explain why CDR3 lengths in the PLASMA sample differ from the NAIVE and MEMORY samples.

3. Compare mutability of V genes, find samples with the smallest and highest mutability, and provide a biological explanation of differences between samples.

4. Compare fractions of productive sequences in all samples.

Please combine all results (including plots) into a single PDF report and upload it to Google Disk

Deadline: October 6th (Friday), 11:59 pm PST.