# **HW 2. Constructing clonal lineages**

The purpose of this assignment is to reconstruct clonal lineages for an antibody repertoire taken after a seasonal flu vaccination and analyze their properties. To complete the assignment, follow the instructions:

1. Download the [repertoire file](https://drive.google.com/file/d/1pMYx0-3pD2UpnCTrXeFtW6999Zhvgr-n/view?usp=sharing) in FASTA format. It consists of 10,000 sequences.
2. (5 points + 5 points) Implement reconstruction of clonal lineages using the Hamming graph (HG) on CDR3s through the following steps:
   1. Compute CDR3s of sequences (e.g., using IgBlast or DiversityAnalyzer, see HW #1 for the details).
   2. Compute Hamming distances (HD) for pairs of the computed CDR3s.
   3. Connect two sequences *s1* and *s2* by an edge if (i) their CDR3s have the same lengths *L* and (ii) the *HD*(CDR3\_*s1*, CDR3\_*s2*) / *L* ≤ 0.2 (i.e., similarity is at least 80%). Since processing all pairs of sequences takes too much time, optimizations are welcome. Implementation and its brief description of any optimization add 5 points to this part.
   4. Report connected components of the HG as clonal lineages.
   5. Attach the implemented code to your report.

**I didn’t do any optimisation except for divide reads by length – and to compare pairs by the same length.**

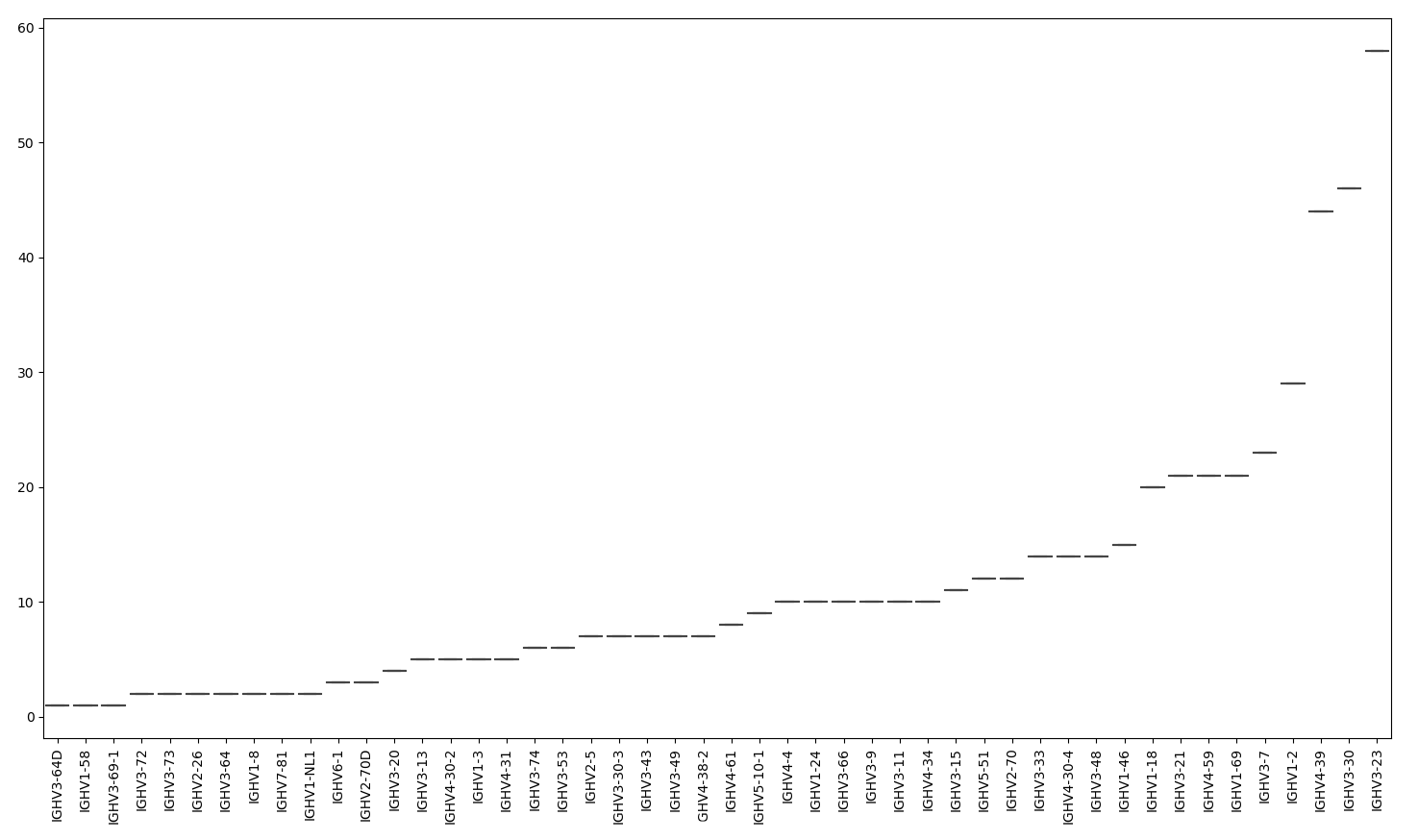
**There was an idea to divide reads by length and their ACGT content. My estimate show that it will save me 10% of comparisons. But, if the treshold will be 0.1, it will save me 60% of comparisons. Anyway, I did not realise that algorithm.**

1. (5 points) Analyze the computed clonal lineages and fill blank cells in Table 1:

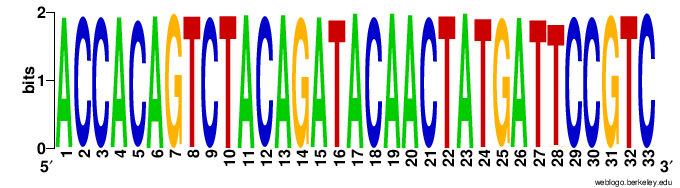
|  |  |
| --- | --- |
| The number of clonal lineages | **546** |
| The number of sequences in the largest lineage | **584** |
| The number of clonal lineages presented by at least 10 sequences | **266** |

Table 1.

1. (5 points) For each lineage, compute the closest V gene (e.g., using IgBlast or DiversityAnalyzer, see HW #1 for the details), create a usage plot of the computed V genes (x axis shows V genes, y axis shows the number of clonal lineages formed by each of V genes), and insert it below:



1. (5 points) Create a [web logo](https://weblogo.berkeley.edu/logo.cgi) plot of CDR3s from the largest clonal lineage and insert it below:



1. (5 points) Extract VDJ sequences from the largest lineage and compute their phylogenetic tree (e.g., using Clustal Omega <https://www.ebi.ac.uk/tools/msa/clustalo>). Visualize the resulting tree (e.g., via Iroki tool: <http://www.iroki.net/viewer>) and add it below:

[Tree figure]

Total: 30 points.

Deadline: November 22nd (Sunday), 11:59 pm PST.