HW 2. Constructing clonal lineages

Весь код доступен в ноутбуке по ссылке:
https://github.com/TohaRhymes/immunogenomics_ib_autumn2020/blob/main/HW2/hw_2_code_and_report.ipynb

1. *Download the repertoire file 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:*

Ответ:

Так как HD будем считать только для строк одинаковой длины, сгруппируем их и создадим для каждой длины свой лист: так мы будем перебирать не $n2$ пар, а в разы меньше - **оптимизация №1**

Напишем функцию для подсчета дистанции Хэмминга (**с оптимизацией №2**):

- если мы при подсчете уже получили HD(CDR3_s1, CDR3_s2) / L ≤ 0.2 - дальше считать не будем, просто вернем False (т.е. строчки связывать ребром не надо).
- для сравнения, добавим функцию, которая не использует эту оптимизацию

Выигрыш: вместо ~25 секунд получили ~20 секунд (~20 процентов времени - не плохо)

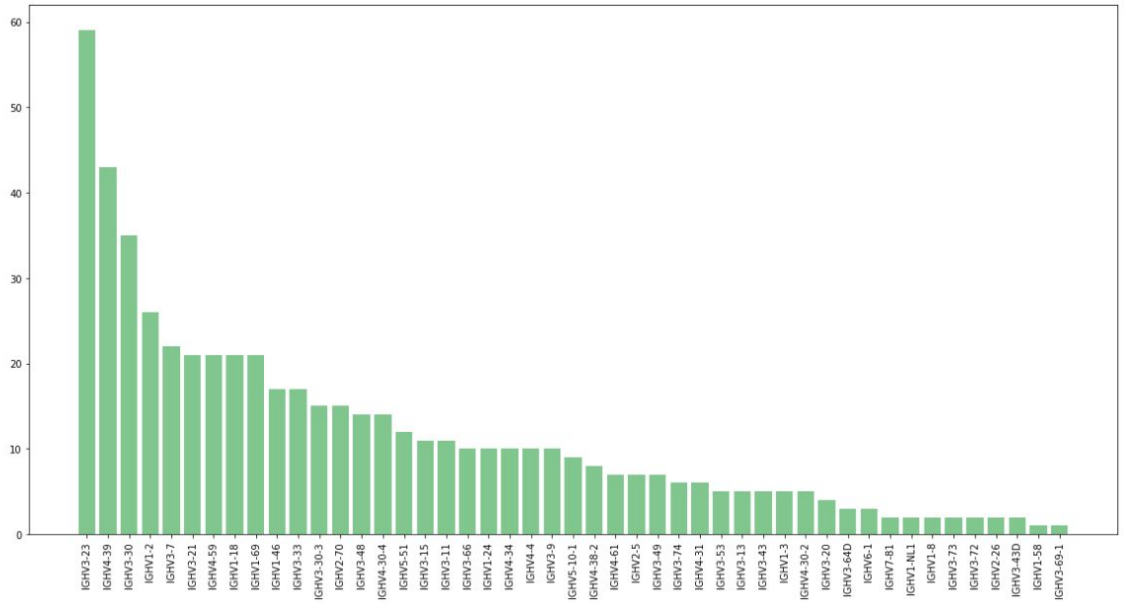
3. *(5 points) Analyze the computed clonal lineages and fill blank cells in Table 1:*

Важный момент: тут вершины - последовательности, поэтому если было 2 одинаковых сиквенса, то они склеились, поэтому число тут - число уникальных сиквенсов.

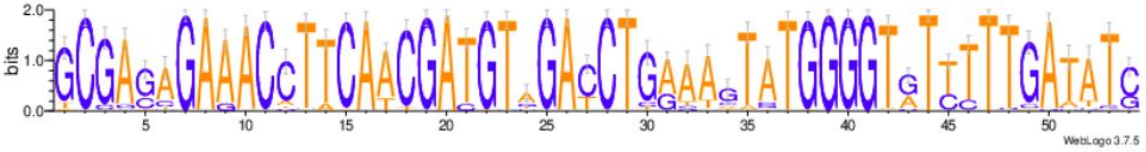
The number of clonal lineages	546
The number of sequences in the largest lineage	122
The number of clonal lineages presented by at least 10 sequences	46

Table 1.

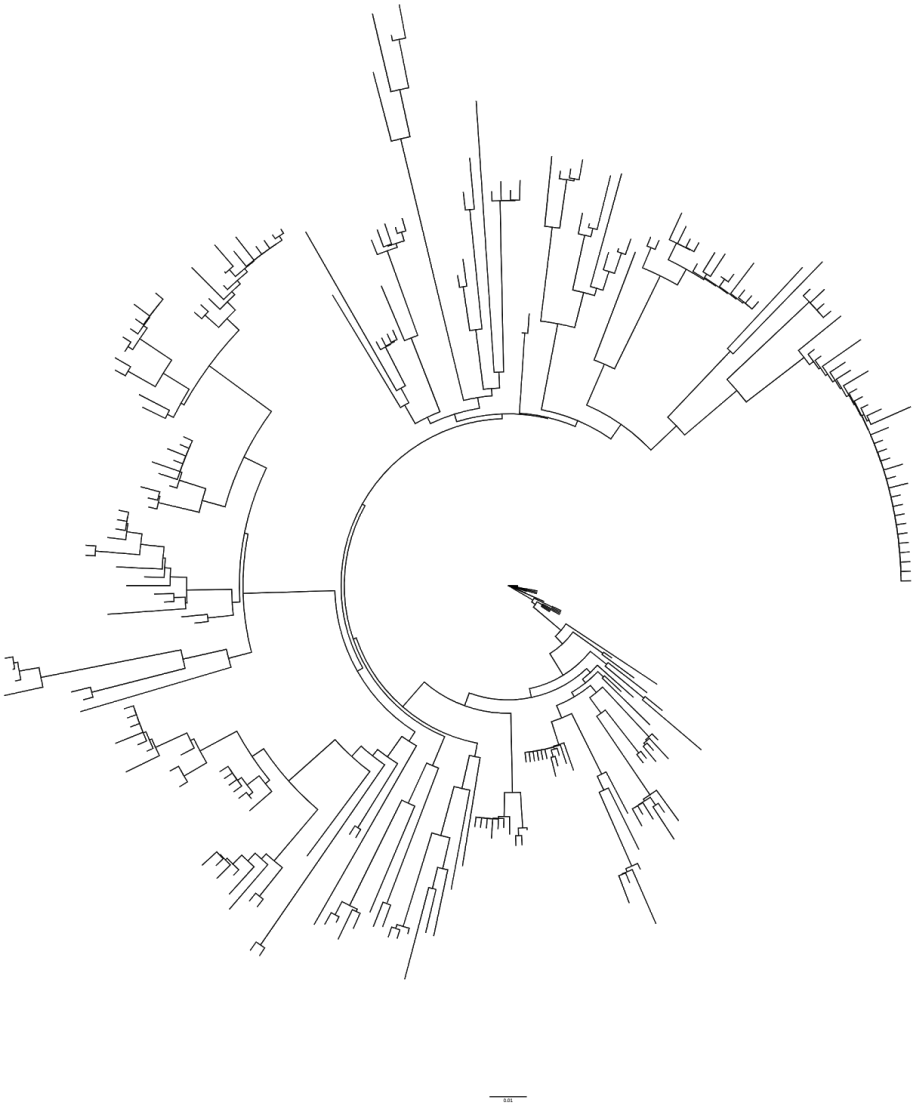
4. *(5 points) For each lineage, compute the closest V gene (e.g., using IgBlast or DiversityAnalyzer, see HW #1 for the details), create a usage plot of the computed V genes (x axis shows V genes, y axis shows the number of clonal lineages formed by each of V genes), and insert it below:*



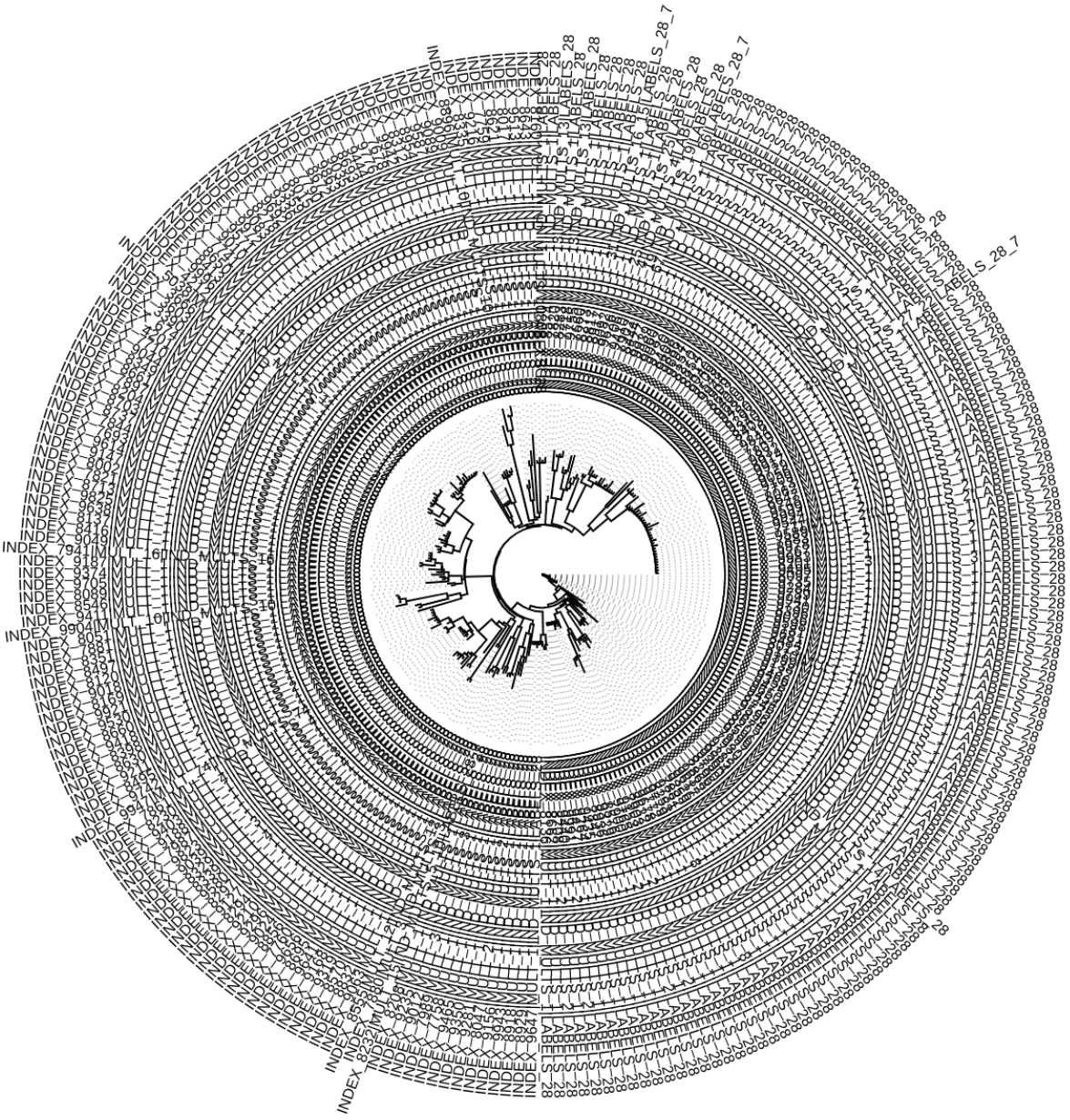
5. *(5 points) Create a [web logo](http://weblogo.berkeley.edu/) plot of CDR3s from the largest clonal lineage and insert it below:*



6. *(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:*



C id:



Total: 30 points.

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