Homework2. Yakovleva Yulia

The analysis was done with DiversityAnalyzer: https://immunotools.github.io/immunotools/

2. Код https://github.com/Yulia-Yakovleva/immunogenomics/tree/master/HW2

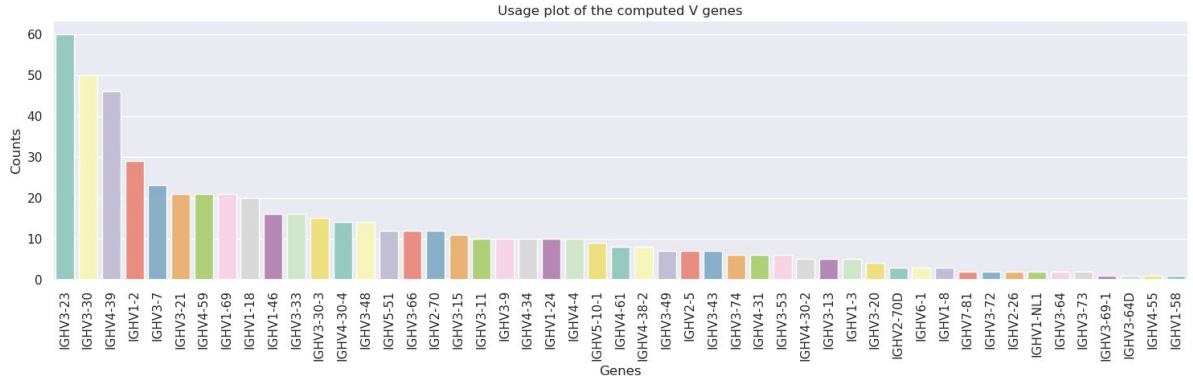
Чтобы не обсчитывать все возможные пары, мы считали Hamming distance только для сиквенсов с одинаковыми длинами. Поскольку CDR3 самый вариабельный среди всех CDRs, мы можем допустить предположение, что все прочтения с одинаковыми CDRs будут одинаковыми, поэтому можем не считать их тоже.

Нереализованная мысль: если две последовательности отличаются на n замен, то найдется такой к-мер, который идентичен в обеих последовательностях, если k <= L / (n + 1), где L - длина последовательности.
То есть можно было разбить последовательности равной длины на множество неперекрывающихся к-меров и для каждой пары делать intersection множеств. Если нет ни одного, то можно смело пропускать данную пару.

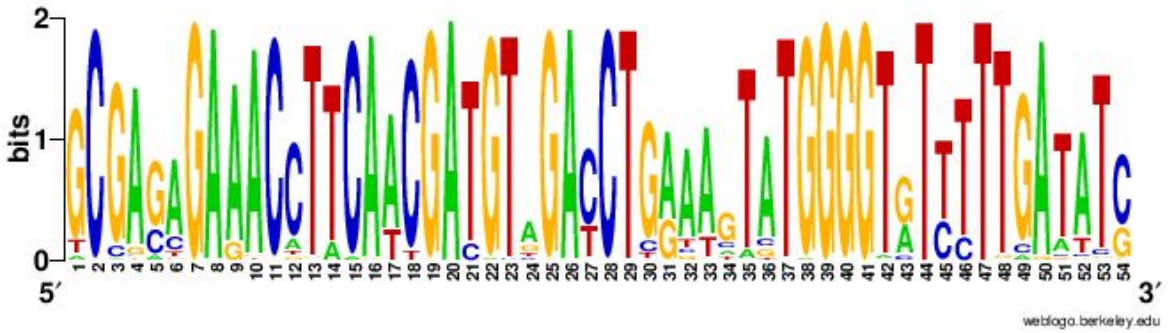
3. Analysis of the computed clonal lineages

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	507

4. V usage plot



5. Web logo plot of CDR3s from the largest clonal lineage



6. Clustal Omega tree