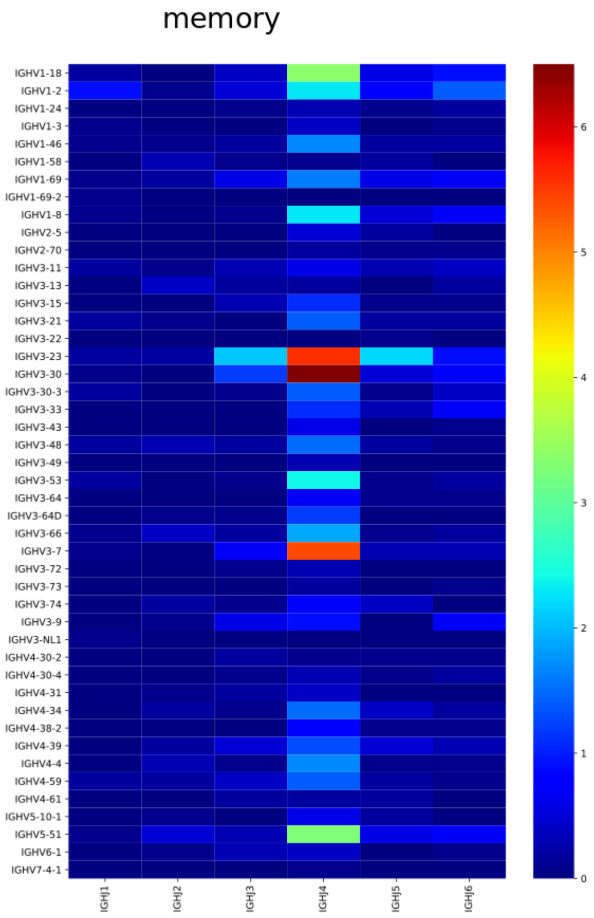
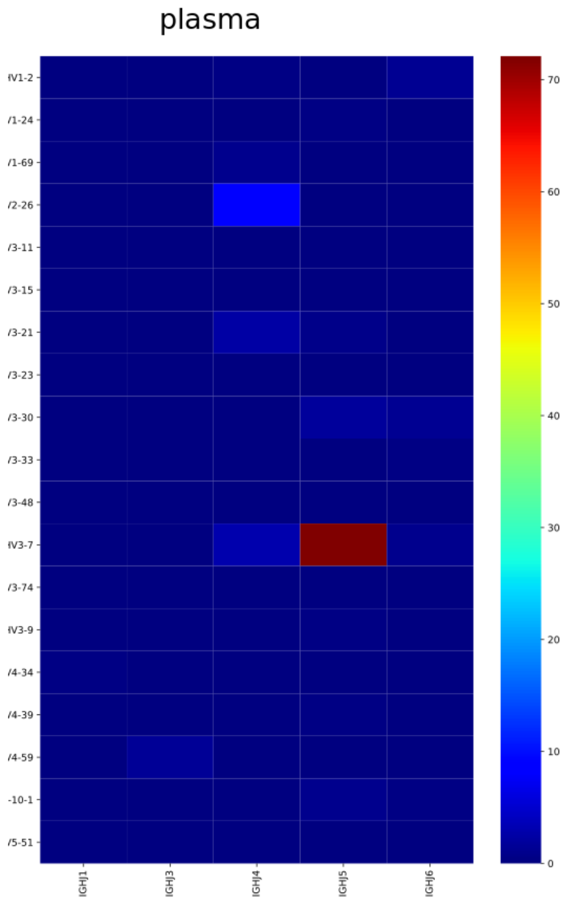
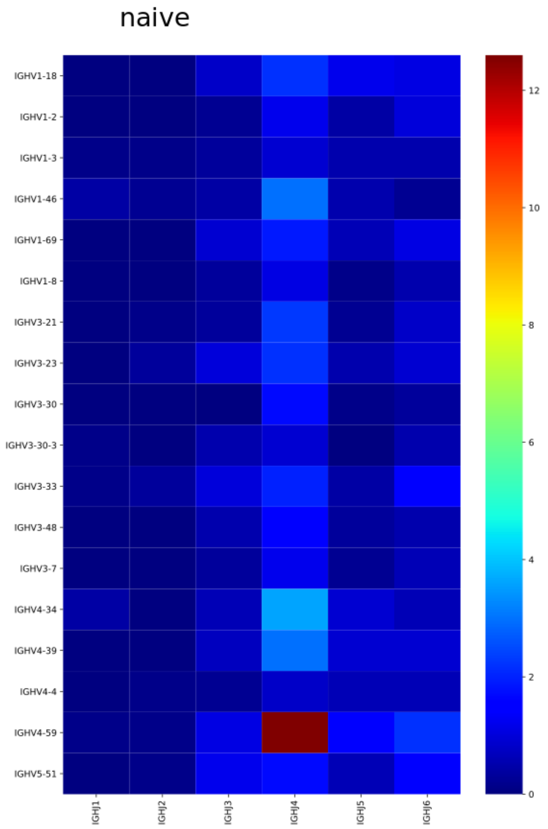
HW#1
Tool: DiversityAnalyzer.

Part 1.

1) Heatmaps, showing the number of sequences for each VJ pair:

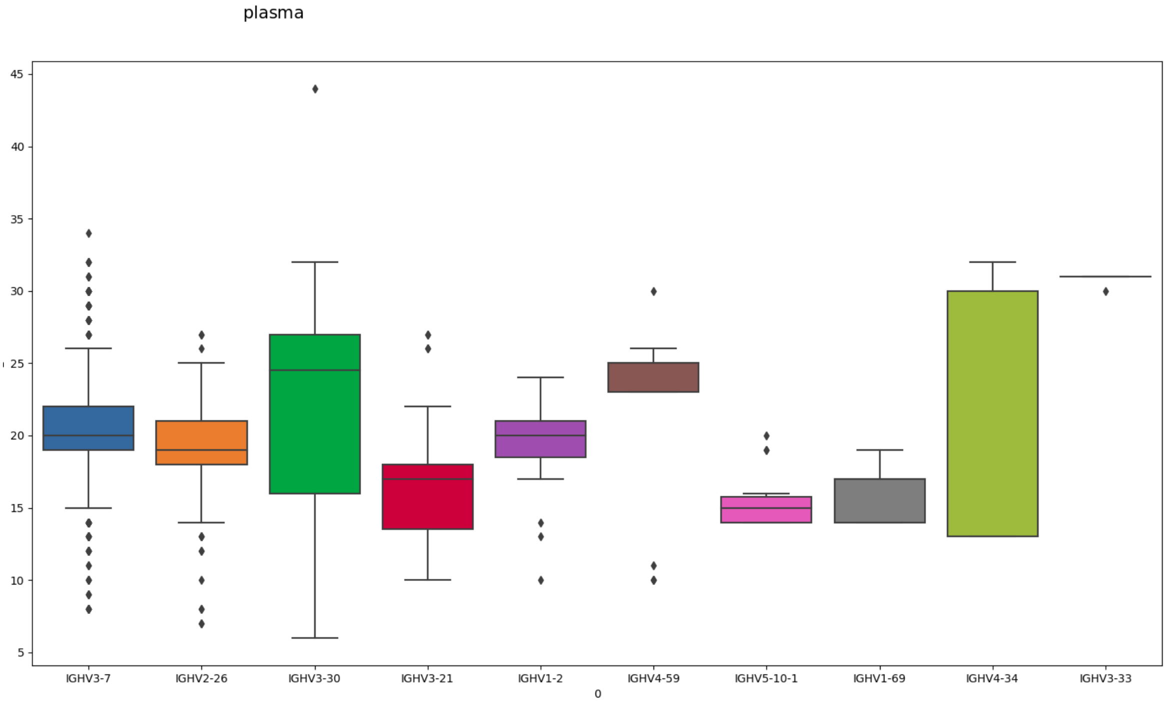


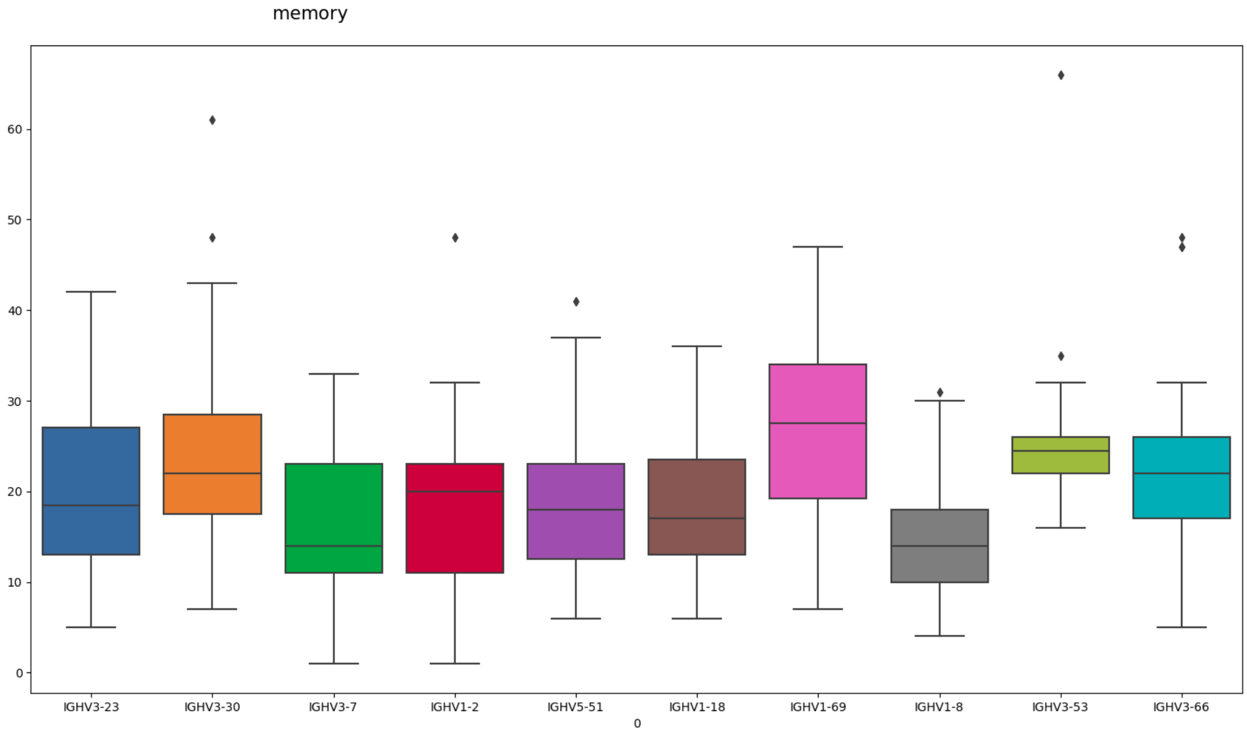


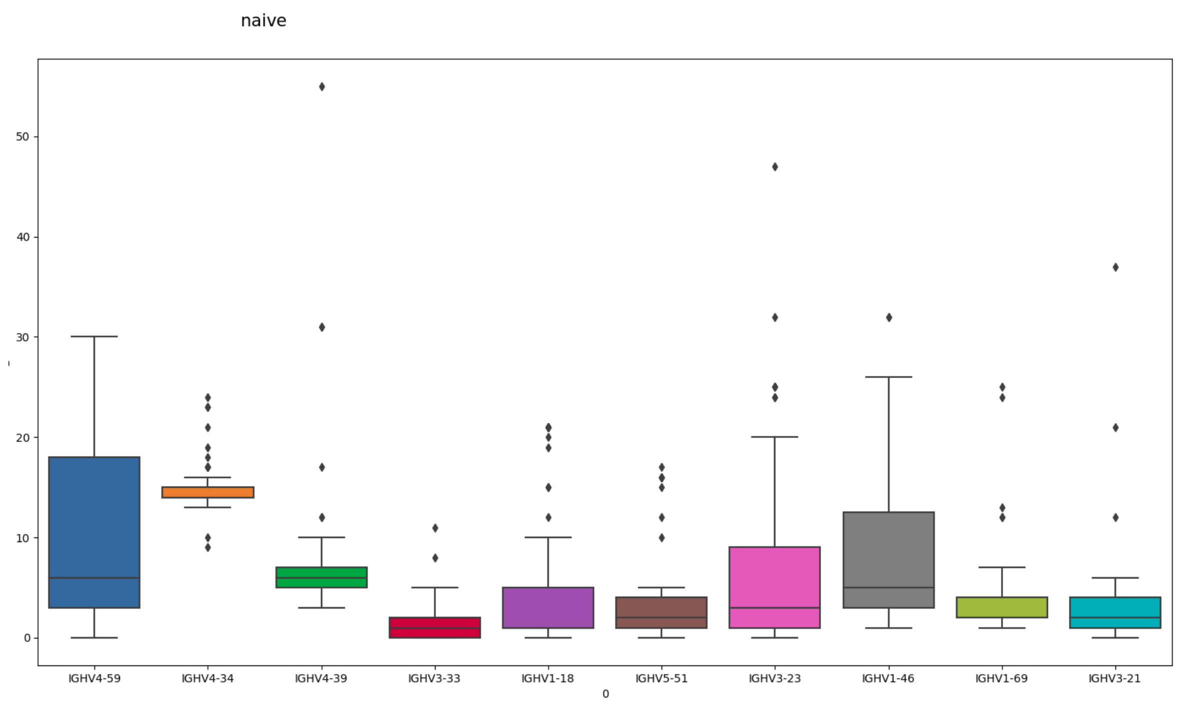


2) Find 10 most used V genes in the sample and analyze their mutability (i.e. distribution of the number of differences).

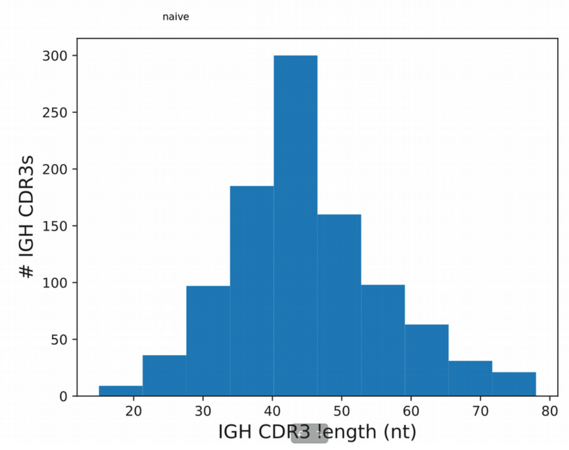
The tool has found most used V genes and reads, aligned with them. Then I`ve used Needleman-Wunsch algorithm to align each gene and it`s reads and count the number of differences. Boxplots for them:

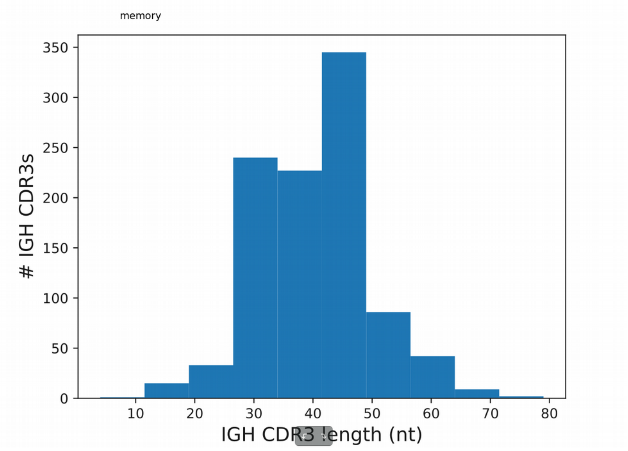


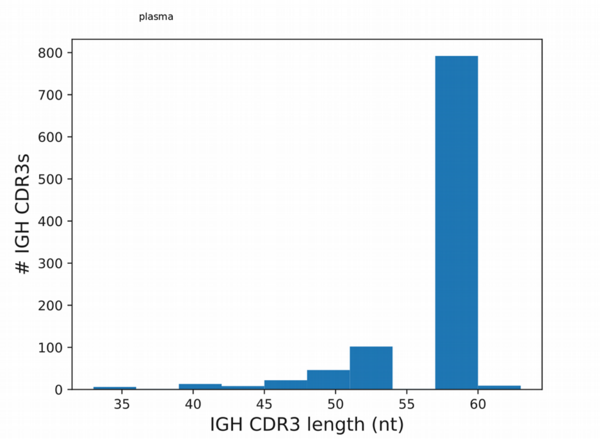




3) Distributions of CDR3 lengths:







4) The fraction of non-productive sequences in the sample:

naive cells:

	IGH	IGK	IGL
# sequences	1000	0	0
# productive sequences	969	0	0
# sequences with stop codon	27	0	0
# out-of-frame sequences	14	0	0

plasma cells:

	IGH	IGK	IGL
# sequences	1000	0	0
# productive sequences	968	0	0
# sequences with stop codon	31	0	0
# out-of-frame sequences	7	0	0

memory cells:

	IGH	IGK	IGL
# sequences	1000	0	0
# productive sequences	922	0	0
# sequences with stop codon	76	0	0
# out-of-frame sequences	14	0	0

Part 2.

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

Memory cells have different usages of VJ combinations — they remember different threats.
Plasma cells have leading IGHV3-7 and IGHJ5 combination, probably, because they are fighting some threat how.
Naive cells have leading IGHJ4 and IGHV4-59 combination. And I don`t understand — why. Probably this is a bug? Or a cancer?

Number of VJ combinations: memory (max ~7) < naive (max ~12) < plasma (max ~70)

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

Plasma cells are trying to get the best affinity with antigens in present time. According to VJ heatmaps, the best affinity in plasma cells is for IGHV3-7 and IGHJ5 combination. Probably, this peak at ~50nt for CDR3 is because it is an average length for this VJ-combination.

For naive cells we see one peak for VJ and CDR3 graphs.
For memory cells we see two peaks for VJ and CDR3 graphs.

Memory and naive cells have some distribution of CDR3 lengths, plasma — nearly not. May be, this is because memory and naive cells are for «different» variants of threats, but plasma — for one present threat.

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

The lowest mutability have naive cells, because they don`t pass throught mutation and selection yet.

Plasma and memory cells have the highest mutability (average rate ~20 differences with reference gene).

Boxplots for memory cells seems a little more symmetric then for plasma cells. Maybe this is because plasma cells have intermediate antibody variants.

4) **Compare fractions of productive sequences in all samples.**

There are MORE non-productive sequences in MEMORY cells then in other ones (twice as more!).

It seems strange — why should we remember cells that don`t produce antibodies?!

Probably, the reason is that memory cells lineages can exists for a very long time. And the replication mistakes rate in them are higher then in naive and plasma cells. Because of that mistakes there are more non-productive sequences.