CSE 597 –006 Introduction to computational immunogenomics

Homework 1: Computing basic characteristics of antibody sequences

Haonan Wu PSU ID: hvw5426

February 20, 2025

1 Analysis for each sample

All the intermediate data and the code to parse and analyze the data can be found in the GitHub repertoire https://github.com/Wu-Haonan/Comp_Immuno_HW1.

1.1 VJ recombination

I run DiversityAnalyzer to align the sequence from plasma, naive, and memory sample data to the human immunoglobulin database and find the closest germline V and J genes. I list each VJ pair and its frequency in Tables 4, 5-7, and 8-10 and generate the heatmap in Figure 7, 8.

1.2 SHM rates of Top 10 V genes

In the plasma sample, the top 10 used V genes are listed in Tables 1, 2 and 3. For each V gene, I compute the SHM rates of each corresponding alignment and draw boxplots (see Figures 1, 2, 3).

V Gene	**Frequency**
IGHV3-7	760
IGHV2-26	85
IGHV3-30	32
IGHV3-21	31
IGHV1-2	19
IGHV4-59	17
IGHV5-10-1	14
IGHV1-69	11
IGHV3-33	5
IGHV4-34	5

Table 1: Top 10 Most Used V Genes in Plasma Sample

V Gene	Frequency
IGHV4-59	175
IGHV4-34	61
IGHV4-39	55
IGHV3-33	54
IGHV1-18	53
IGHV5-51	51
IGHV3-23	49
IGHV1-46	47
IGHV1-69	45
IGHV3-21	37

Table 2: Top 10 Most Used V Genes in Naive Sample

V Gene	Frequency
IGHV3-23	112
IGHV3-30	91
IGHV3-7	68
IGHV1-2	60
IGHV1-18	55
IGHV5-51	55
IGHV1-69	38
IGHV1-8	37
IGHV3-53	30
IGHV3-66	29

Table 3: Top 10 Most Used V Genes in Memory Sample

1.3 CDR3 length distribution and non-productive rate

The distribution of CDR3 lengths of the plasma sample is shown in Figure 4, 5 and 6.

The fraction of non-productive sequences in the plasma sample is 0.032.
The fraction of non-productive sequences in the naive sample is 0.031.
The fraction of non-productive sequences in the naive sample is 0.078.

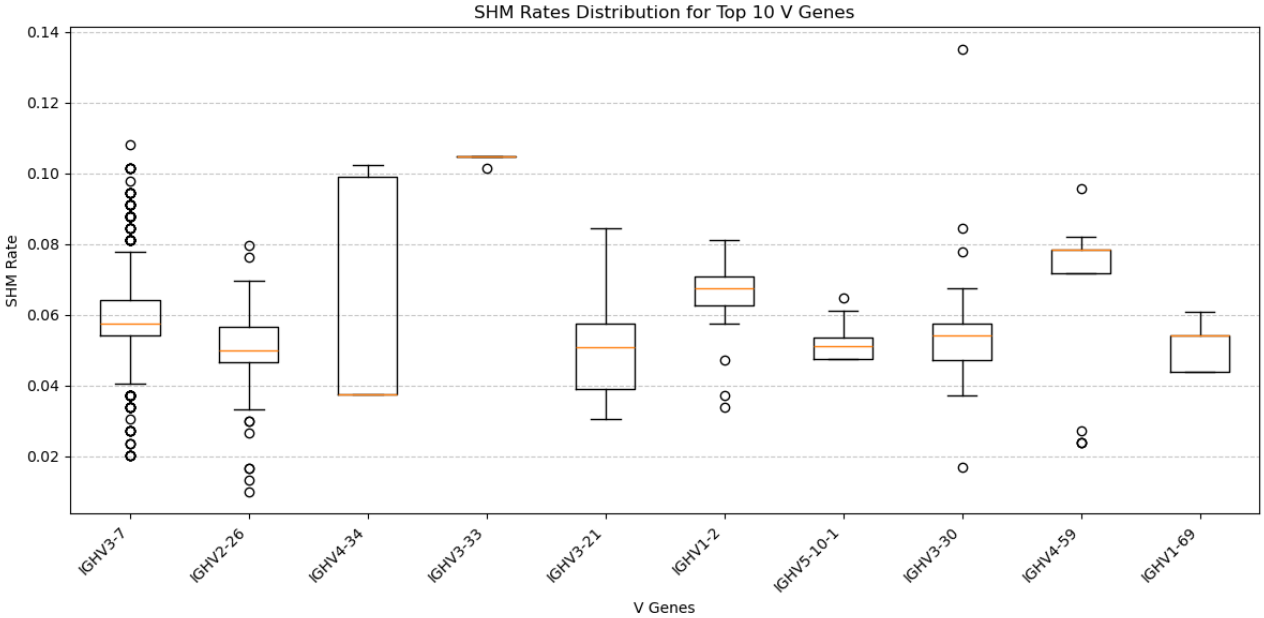

Figure 1: Number of sequences for each VJ pair in Plasma B cell Sample.

2 Integration analysis

1. The plasma sample has 32 VJ combinations, the naive sample has 70 VJ combinations, and 60. So, the plasma sample has the smallest VJ combinations and the naive sample has highest combinations.

2. The CDR3 length distribution of the plasma sample mainly lies at a fixed length, but naive and memory samples have a wide CDR3 length distribution. Because the plasma B cell has undergone somatic hypermutation (SHM) and clonal selection to optimize antigen-binding affinity and CDR3 is an important domain of immunoglobulin to neutralize antigens, it has a relatively stable length of CDR3 length. The naive B cells have not undergone selection, so their CDR3 length is primarily determined by random V(D)J recombination. Memory B cells retain a relatively high level of diversity to prepare for future antigen variants. Hence, naive and memory B cells lead to broader distributions.

3. The V mutability boxplots of each sample are shown in Figure 10, 11 and 10. We can observe that

   - Naive sample has smaller SHM rates and many sequences do not have any mutation, because it does not undergo somatic hypermutation.
   - Plasma sample has relatively high SHM rates and is centralized, because the plasma B cell has undergone somatic hypermutation (SHM) and clonal selection to optimize antigen-binding.
   - Memory sample has high SHM rates and a high variance, because it needs high diversity to prepare for future antigen variants.

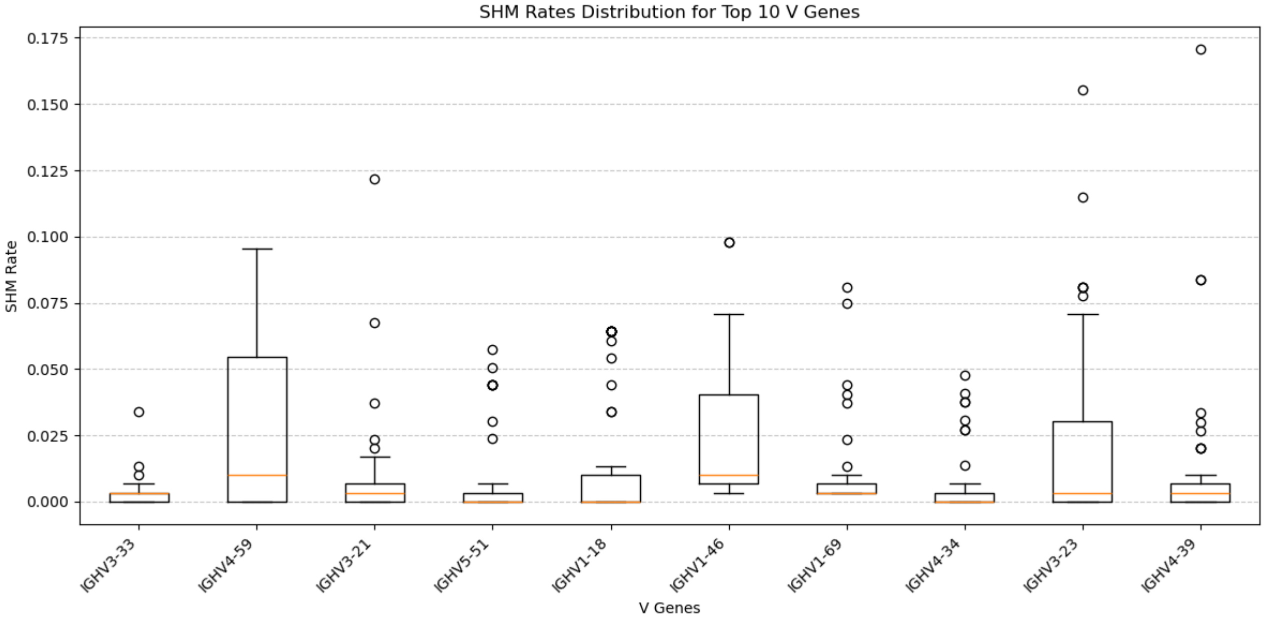

Figure 2: Number of sequences for each VJ pair in Naive B cell Sample.

4. The intersection of three top used v gene sets is **IGHV1-69** gene. The results (see Figure 13) also match what we observed (see point 3) i.e. the naive sample has the smallest SHM rate and the memory sample has the highest.

5. The naive B cells have not undergone SHM, so the non-productive rate is quite low. The memory B cells are selected by their affinity, so non-productive sequence will be dropped after selection. The memory B cells have the highest non-productive rate, because they need high diversity but without selection pressure.

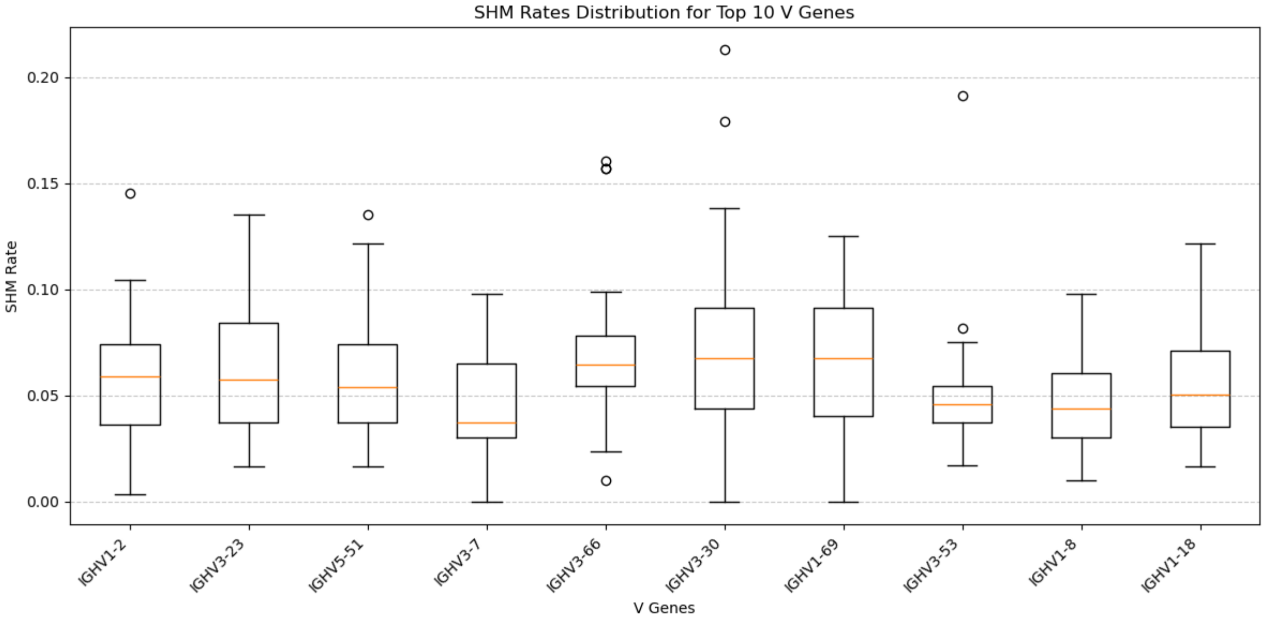

Figure 3: Number of sequences for each VJ pair in Memory B cell Sample.

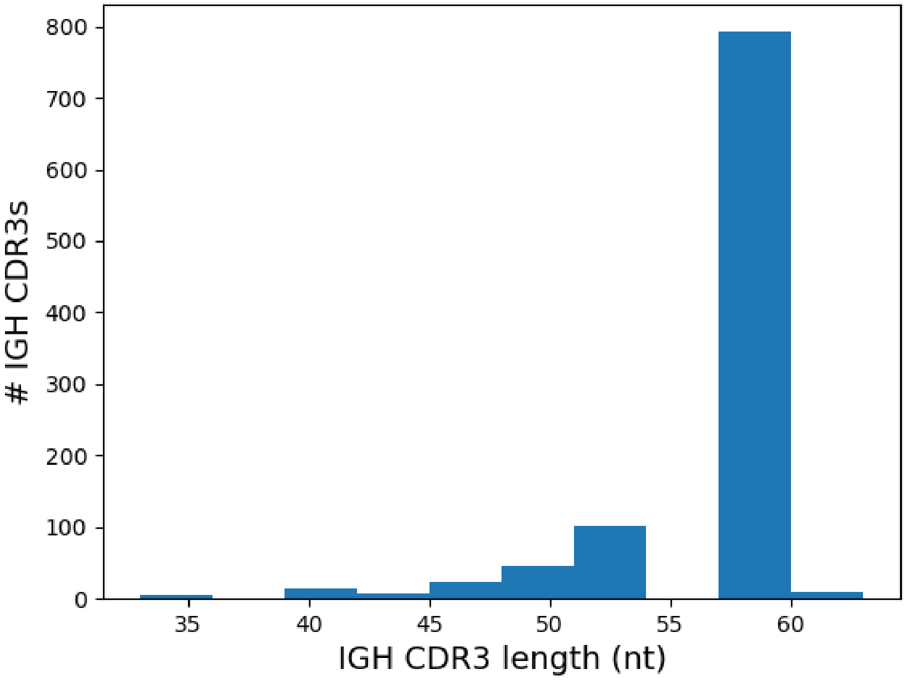

Figure 4: CDR3 length distribution of plasma sample

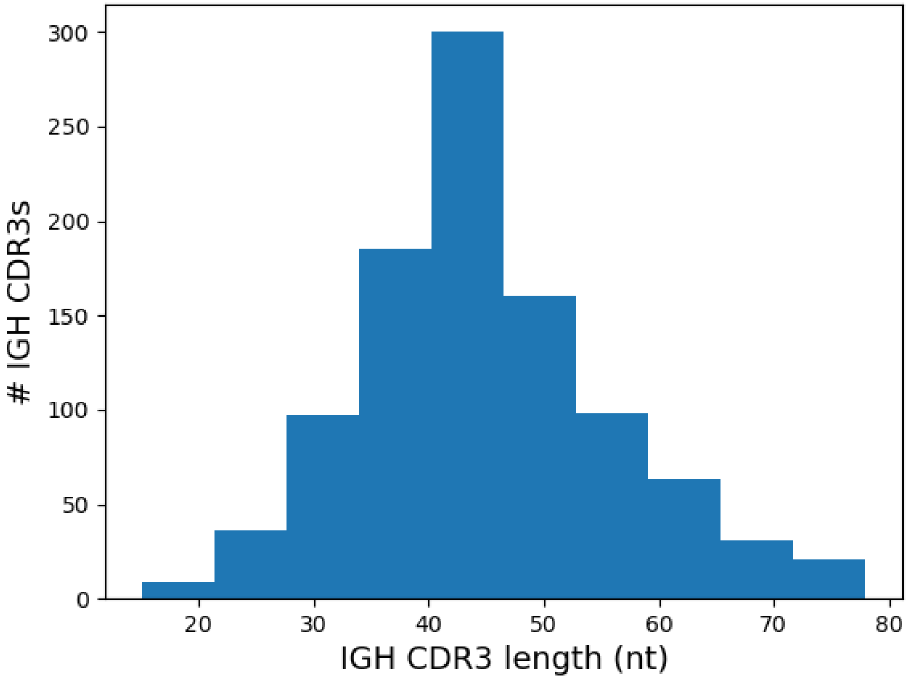

Figure 5: CDR3 length distribution of naive sample

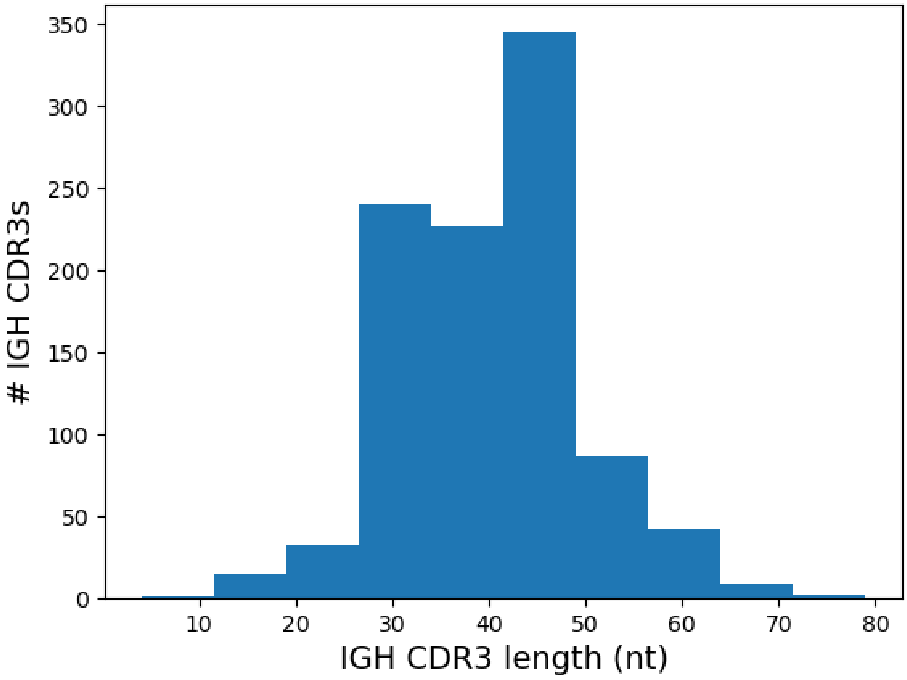

Figure 6: CDR3 length distribution of memory sample

V Hit	**J Hit**	**Frequency**
IGHV3-7	IGHJ5	721
IGHV2-26	IGHJ4	85
IGHV4-34	IGHJ1	5
IGHV3-7	IGHJ4	29
IGHV3-33	IGHJ6	5
IGHV3-21	IGHJ4	24
IGHV1-2	IGHJ6	12
IGHV5-10-1	IGHJ5	11
IGHV3-9	IGHJ5	3
IGHV1-2	IGHJ3	2
IGHV3-30	IGHJ6	14
IGHV3-7	IGHJ6	10
IGHV3-48	IGHJ4	1
IGHV5-51	IGHJ4	1
IGHV3-30	IGHJ5	17
IGHV4-59	IGHJ3	16
IGHV1-2	IGHJ4	5
IGHV1-69	IGHJ4	11
IGHV5-51	IGHJ6	1
IGHV1-24	IGHJ5	3
IGHV3-21	IGHJ5	6
IGHV5-10-1	IGHJ6	3
IGHV3-30	IGHJ4	1
IGHV4-39	IGHJ5	4
IGHV3-23	IGHJ1	2
IGHV3-11	IGHJ6	1
IGHV3-15	IGHJ4	1
IGHV3-48	IGHJ5	2
IGHV3-21	IGHJ6	1
IGHV3-74	IGHJ5	1
IGHV4-39	IGHJ1	1
IGHV4-59	IGHJ6	1

Table 4: V-J Combination Frequency Table of Plasma Sample

V Hit	J Hit	Frequency
IGHV3-49	IGHJ4	7
IGHV3-33	IGHJ6	16
IGHV4-59	IGHJ6	22
IGHV4-4	IGHJ2	1
IGHV2-26	IGHJ2	1
IGHV3-21	IGHJ4	23
IGHV4-4	IGHJ5	6
IGHV1-8	IGHJ6	5
IGHV3-66	IGHJ4	10
IGHV1-58	IGHJ3	1
IGHV5-51	IGHJ5	6
IGHV1-18	IGHJ4	22
IGHV5-51	IGHJ3	12
IGHV1-46	IGHJ4	30
IGHV1-18	IGHJ6	11
IGHV4-59	IGHJ4	126
IGHV3-9	IGHJ4	11
IGHV1-46	IGHJ2	2
IGHV1-69	IGHJ6	11
IGHV4-59	IGHJ3	11

Table 5: V-J Combination Frequency Table of Naive Sample (Part 1)

V Hit	J Hit	Frequency
IGHV4-34	IGHJ4	36
IGHV2-5	IGHJ1	1
IGHV3-33	IGHJ5	4
IGHV1-69	IGHJ3	9
IGHV1-2	IGHJ4	12
IGHV1-69	IGHJ4	19
IGHV3-23	IGHJ4	22
IGHV3-13	IGHJ4	6
IGHV2-5	IGHJ3	5
IGHV4-4	IGHJ3	2
IGHV3-7	IGHJ6	6
IGHV3-15	IGHJ5	7
IGHV4-4	IGHJ4	8
IGHV1-18	IGHJ3	8
IGHV3-30-3	IGHJ6	5
IGHV2-5	IGHJ4	6
IGHV1-8	IGHJ5	1
IGHV4-59	IGHJ5	14
IGHV4-30-2	IGHJ3	2
IGHV3-33	IGHJ2	3

Table 6: V-J Combination Frequency Table of Naive Sample (Part 2)

V Hit	J Hit	Frequency
IGHV4-30-4	IGHJ4	5
IGHV4-39	IGHJ5	9
IGHV3-33	IGHJ4	20
IGHV5-51	IGHJ6	15
IGHV1-46	IGHJ3	4
IGHV3-23	IGHJ6	9
IGHV1-8	IGHJ4	11
IGHV5-51	IGHJ4	17
IGHV4-39	IGHJ4	30
IGHV3-48	IGHJ4	14
IGHV4-34	IGHJ6	6
IGHV1-24	IGHJ4	8
IGHV3-53	IGHJ3	2
IGHV3-7	IGHJ3	3
IGHV4-34	IGHJ5	9
IGHV4-31	IGHJ4	6
IGHV1-46	IGHJ6	2
IGHV3-30	IGHJ4	17
IGHV3-64D	IGHJ4	2
IGHV4-30-4	IGHJ5	1
IGHV3-23	IGHJ5	5
IGHV3-48	IGHJ5	3
IGHV6-1	IGHJ4	2
IGHV1-58	IGHJ2	1
IGHV2-70	IGHJ4	6
IGHV3-33	IGHJ3	10
IGHV3-13	IGHJ6	1
IGHV3-74	IGHJ6	4
IGHV3-7	IGHJ4	12
IGHV3-66	IGHJ3	4

Table 7: V-J Combination Frequency Table of Naive Sample (Part 3)

V Hit	**J Hit**	**Frequency**
IGHV1-2	IGHJ4	23
IGHV3-23	IGHJ1	2
IGHV4-39	IGHJ4	13
IGHV5-51	IGHJ4	33
IGHV3-48	IGHJ4	15
IGHV3-7	IGHJ3	7
IGHV3-66	IGHJ2	4
IGHV3-30	IGHJ3	12
IGHV4-59	IGHJ5	2
IGHV1-2	IGHJ6	14
IGHV5-10-1	IGHJ4	6
IGHV5-51	IGHJ2	5
IGHV3-30-3	IGHJ4	14
IGHV1-69	IGHJ4	16
IGHV4-34	IGHJ4	15
IGHV1-3	IGHJ4	4
IGHV3-23	IGHJ5	22
IGHV1-46	IGHJ4	17
IGHV3-30-3	IGHJ6	4
IGHV1-69	IGHJ6	7

Table 8: V-J Combination Frequency Table of Memory Sample (Part 1)

V Hit	**J Hit**	**Frequency**
IGHV3-30	IGHJ4	65
IGHV3-15	IGHJ6	1
IGHV3-23	IGHJ3	21
IGHV4-38-2	IGHJ4	8
IGHV3-53	IGHJ4	24
IGHV5-51	IGHJ6	7
IGHV3-66	IGHJ4	19
IGHV3-7	IGHJ4	54
IGHV3-64D	IGHJ4	12
IGHV4-34	IGHJ2	2
IGHV3-9	IGHJ3	6
IGHV3-64	IGHJ4	7
IGHV4-39	IGHJ3	5
IGHV1-8	IGHJ6	7
IGHV2-70	IGHJ4	2
IGHV1-18	IGHJ4	34
IGHV4-4	IGHJ4	17
IGHV6-1	IGHJ4	4
IGHV4-30-2	IGHJ3	2
IGHV1-58	IGHJ5	2

Table 9: V-J Combination Frequency Table of Memory Sample (Part 2)

V Hit	**J Hit**	**Frequency**
IGHV3-13	IGHJ4	2
IGHV1-2	IGHJ1	9
IGHV3-48	IGHJ5	2
IGHV3-11	IGHJ5	3
IGHV3-33	IGHJ6	7
IGHV1-8	IGHJ1	1
IGHV4-61	IGHJ5	2
IGHV3-21	IGHJ4	14
IGHV3-23	IGHJ4	56
IGHV3-33	IGHJ4	11
IGHV3-30	IGHJ1	1
IGHV3-49	IGHJ4	3
IGHV3-23	IGHJ6	9
IGHV3-7	IGHJ5	3
IGHV1-8	IGHJ4	23
IGHV3-11	IGHJ4	6
IGHV3-43	IGHJ4	6
IGHV1-3	IGHJ1	1
IGHV1-18	IGHJ3	4
IGHV4-4	IGHJ2	3

Table 10: V-J Combination Frequency Table of Memory Sample (Part 3)

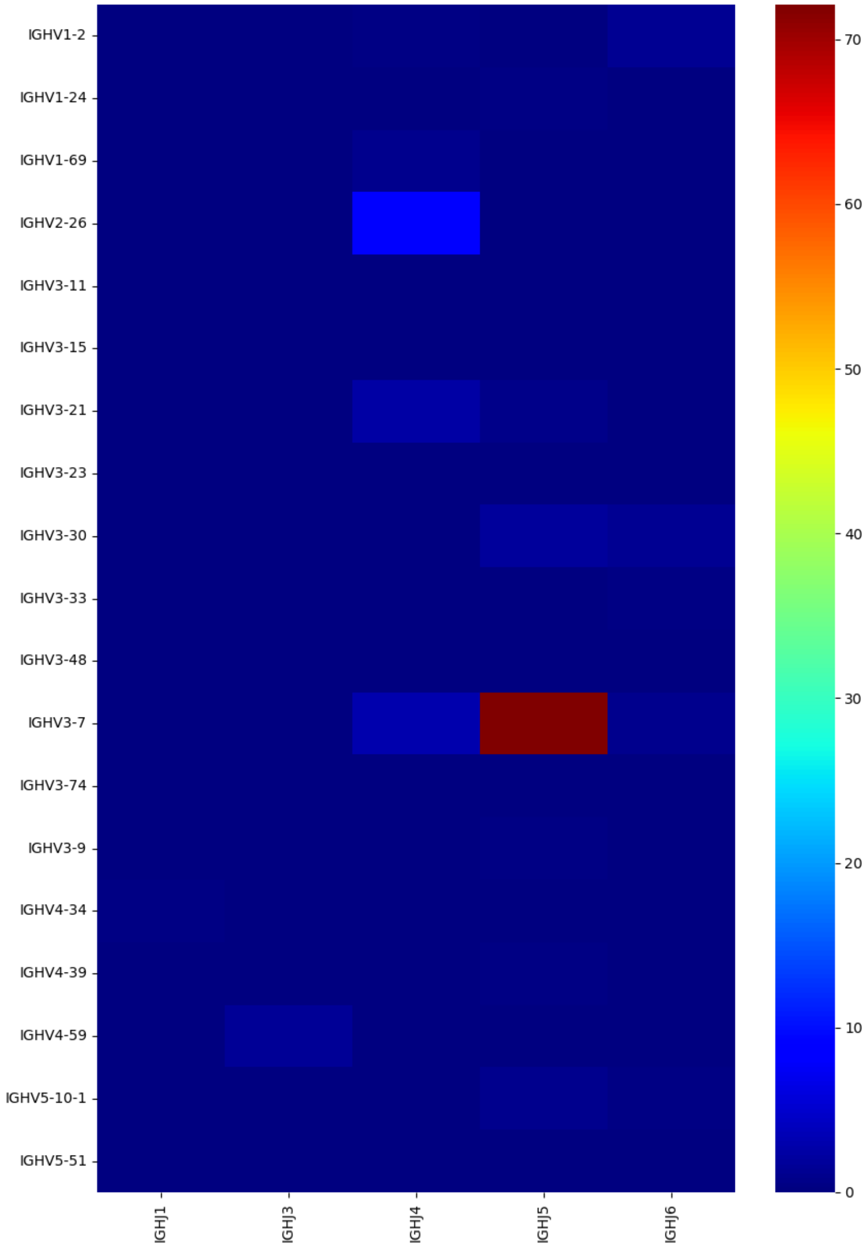

Figure 7: Number of sequences for each VJ pair in Plasma B cell Sample.

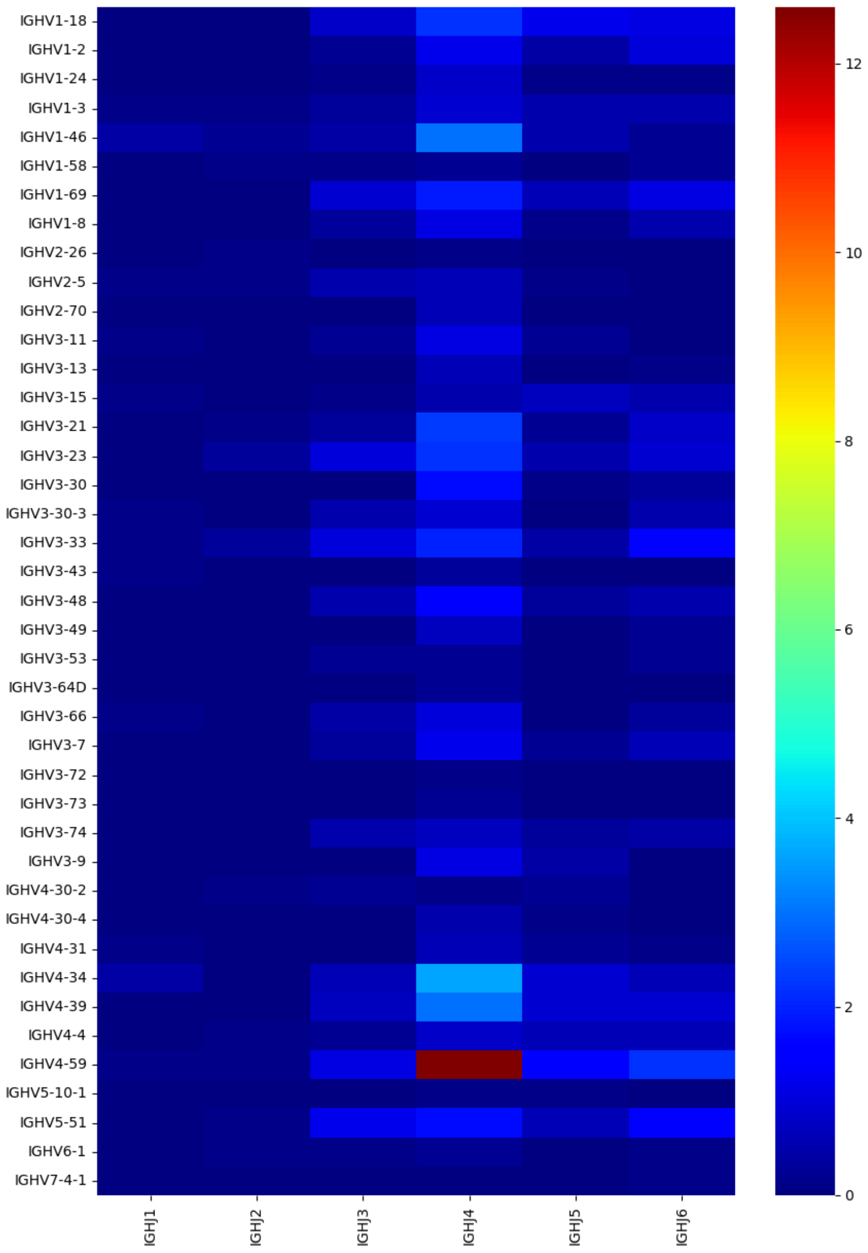

Figure 8: Number of sequences for each VJ pair in Naive B cell Sample.

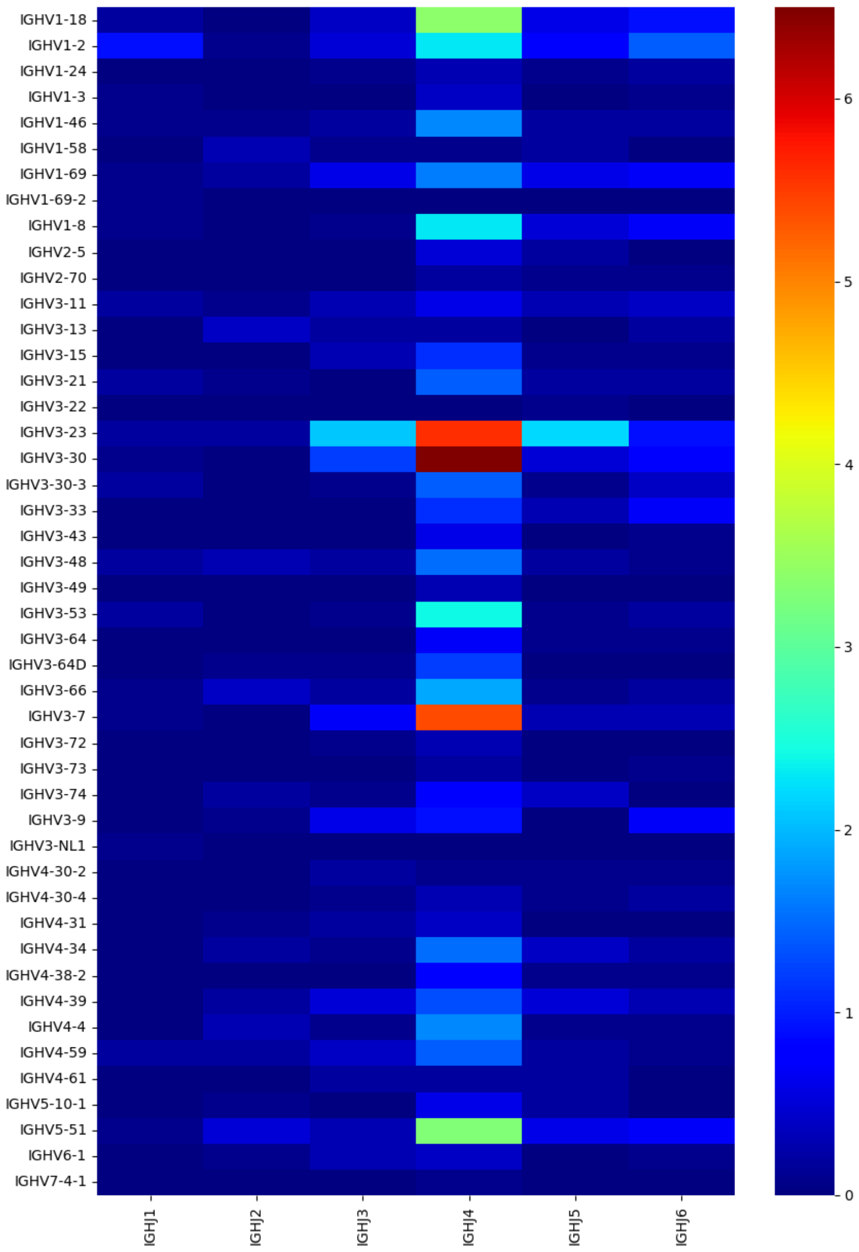

Figure 9: Number of sequences for each VJ pair in Memory B cell Sample.

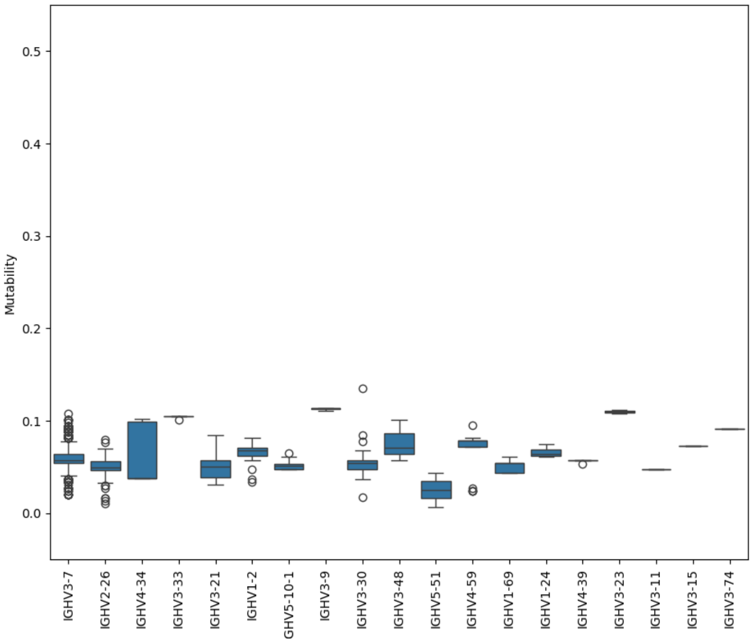

Figure 10: V mutability of the plasma sample.

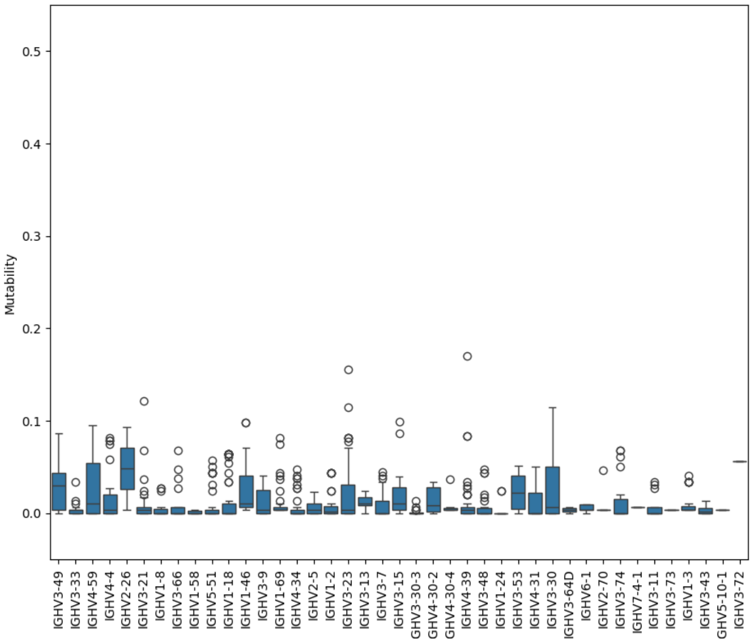

Figure 11: V mutability of the naive sample.

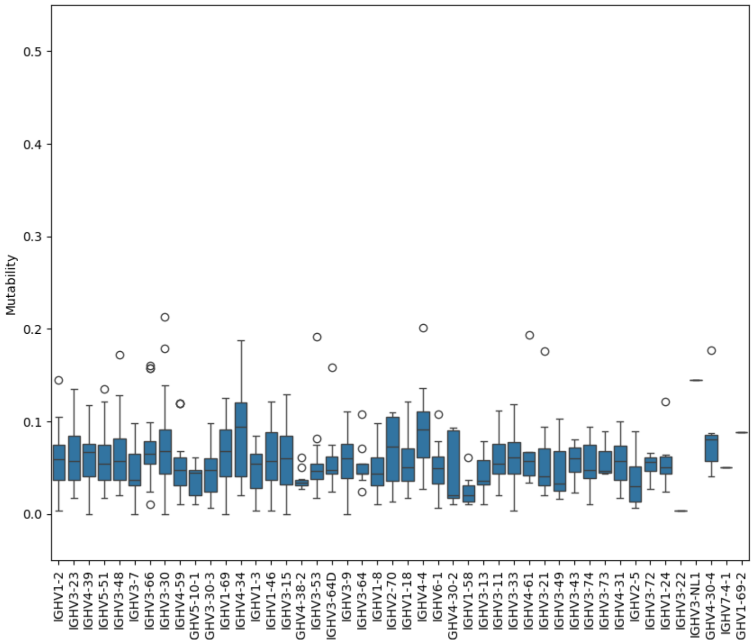

Figure 12: V mutability of the memory sample.

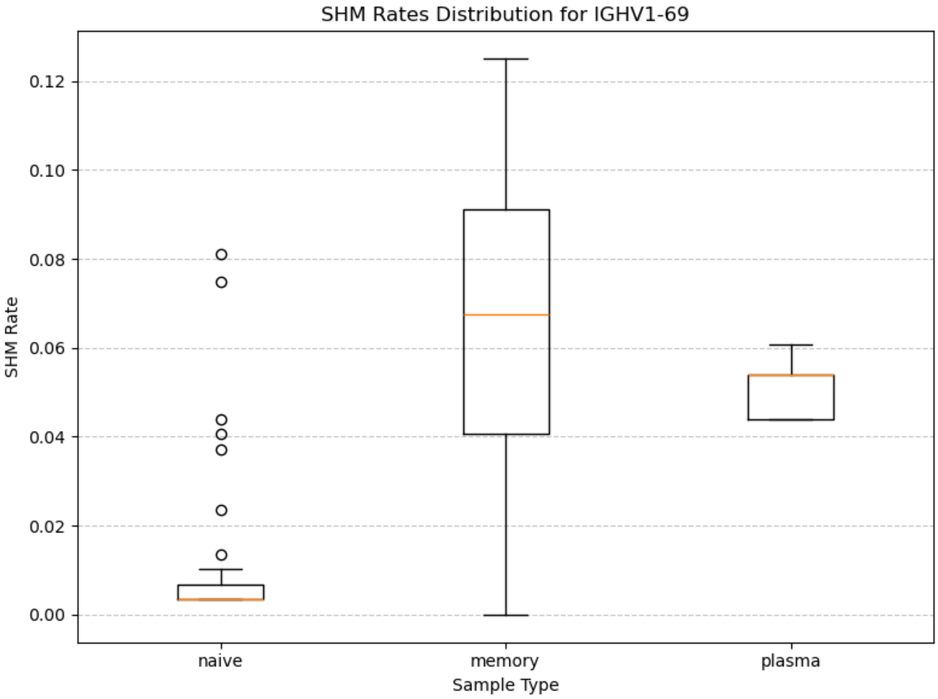

Figure 13: SHM rates across three samples on IGHV1-69 gene