Term Project Report

Northern Illinois University

Fall 2022 CSCI 652 - MSTR Algorithmic Bioinformatics I

SARS-Cov-2 Wild Type vs Representative Omicron Variant

Professor Minmei Hou

Team Members:

Mahima Devi Allam (Z1924638)

Dinesh Kolla (Z1935563)

Ganesh Kanadam (Z1939317)

Introduction

Sars-Cov2 is one of the viruses that causes the Covid19 pandemic. It belongs to the coronavirus family, and it also contains other genomes like Omicron variant, which is more contagious than prior variants, it leads to large cause in community cases. The virus that causes a respiratory disease called coronavirus disease 19 (COVID-19). SARS-CoV-2 is a member of a large family of viruses called coronaviruses. The incubation period of SARS CoV-2 wild-type is 5 days. December 15, 2020, to February 28, 2021. Basically, wild type gene is found in its natural, non-mutated which is the unchanged virus form.

These viruses can infect people and some animals. SARS-CoV-2 was first known to infect people in 2019. The virus is thought to spread from person to person through droplets released when an infected person coughs, sneezes, or talks. It may also be spread by touching a surface with the virus on it and then touching one’s mouth, nose, or eyes, but this is less common. Research is being done to treat COVID-19 and to prevent infection with SARS-CoV-2. Also called severe acute respiratory syndrome coronavirus 2. The SARS-CoV-2 Omicron variant is more contagious than prior variants, leading to large increases in community cases, as well as Omicron’s greater contagiousness than wild type. The incubation period of SARSCoV-2 Omicron is 3 days and from December 15, 2021, to February 28, 2022, is when the Omicron variant predominated. Omicron’s variants are especially efficient spreaders of the disease. One explanation was that more than 30 of Omicron’s mutations are on the virus’s spike protein, the part that attaches to human cells, and several of those are believed to increase the probability of infection.

In this project, we will use pairwise alignment for Omicron like genomes, we are calculating the substitution rate and the gap rates for the genome, and we also plotted some graphs in order to study the relation between Sars-Cov2 and Omicron Variant.

Methodology

We used pairwise alignment for Omicron like genomes. Pairwise alignment is defined as an alignment procedure comparing two biological sequences of either protein, DNA or RNA. Pair-wise alignments are also an essential element in genome assembly pipelines and alignment of entire genome sequences can identify genes duplications and deletions.

Pairwise Alignment is used to find areas of similarity between two biological sequences that may point to functional, structural, or evolutionary links (protein or nucleic acid). When aligning two sequences, the algorithm will identify the optimal relationship between them. This is done by comparing every letter in one sequence with every letter in the other. The algorithm will account for matches and mismatches and compute the best mathematical path through these matches and mismatches. Comparing Pairwise to other Protein-DNA alignment methods, it stands out mainly for its ability for frame shifting during alignment. For typical Protein-DNA alignment methods, they first select one of three frames to convert the DNA into a protein sequence, and then they compare it with the specified protein. This alignment is based on the idea that the DNA translation frame is not broken along the entire DNA strand.

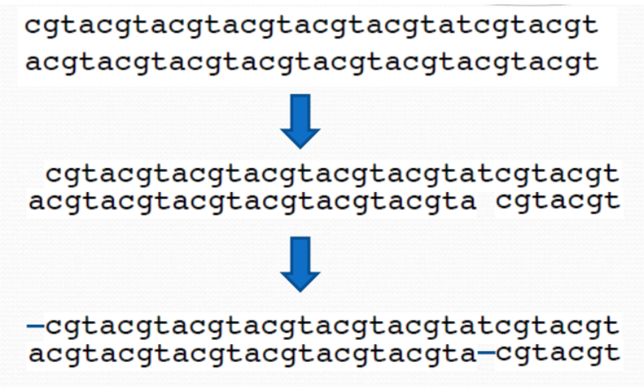
The substitution rate is calculated when the number of mutations in each generation is multiplied by the probability of the new mutation that reaches fixation. In other words, it can be described as the frequency of DNA sequence mutations caused by nucleotide replacement. Based on the substitution of nucleotides in a DNA sequence, substitution rate is the most often measured kind of mutation. With traditional DNA analysis, it is simpler to quantify them. Compared to other groups of mutations, which typically occur at high rates, substitution rates of mutations have distinct rates of mutation per generation.

$$\text{Substitution Rate} = \frac{\text{Mismatching Count}}{\text{Matching Count} + \text{Mismatching Count}}$$

Insertions and deletions are additional elements to take into account when studying sequences. It is expected that while comparing the sequences of different protein family members, we will detect that some of the sequences have one or more extra residues (insertions) or some residues that are missing (deletion). For instance, there will frequently be some extensive segments of insertions and deletions when comparing a group of bacterial sequences to a group of eukaryotic sequences. A whole domain may occasionally be added to or removed from a protein. Depending on how we handle these insertions and deletions, different sequence alignments can be produced.

In bioinformatics, gaps are used to account for genetic mutations occurring from insertions or deletions in the sequence.

$$\text{Gap Rate} = \frac{\text{Gap Count}}{\text{Match Count} + \text{Mismatch Count} + \text{Gap Count}}$$



```
cgtacgtacgtacgtacgtacgtatcgtacgt
acgtacgtacgtacgtacgtacgtacgtacgt

 cgtacgtacgtacgtacgtacgtatcgtacgt
acgtacgtacgtacgtacgtacgta cgtacgt

-cgtacgtacgtacgtacgtacgtatcgtacgt
acgtacgtacgtacgtacgtacgta-cgtacgt
```

In the example alignment above, we introduced a gap (marked by a dash in the above sequence) to maximize the number of matches. A gap in one of the sequences means that one or more amino acid residues have been deleted from the sequence, or we could also say that there is an insertion in the second sequence.

Work Environment

Coding Platform: Visual studio

Programming Language used: Java

Packages used: Java.io.*, Java.util.*, Java.long.character.*

Graphs: Microsoft Excel

Implementation

Input file: sars2.omicron.sing.maf

The input file provided to us is as stated above. In the input file, there are two datasets with various variants: one is the SARS Cov-2 wild type dataset, and the other is the omicron variant dataset. Both of these datasets are equal in size.

Structured file: sarsCov2structure.txt

Additionally, a structured file was provided to us. There are six genomes in the file, which are the Orf1a, Orf1b, S, E, M, and N genomes, respectively. There are distinct beginning and ending points for each genome.

The **Orf1a** and **Orf1b** genes are conserved throughout the genomes of nidoviruses, a class of viruses that also includes coronaviruses. ORF1a and ORF1b, found in genomes of coronaviruses. Orf1a is the first open reading frame at the 5' end of the genome. They produce the polypeptides 1a and b, which are involved in RNA production and interfere the immune system. The two Orfs are collectively referred to as the replicase gene.

S gene: A homotrimeric glycoprotein complex called spike protein, which is encoded by the S-gene, is necessary for infectiousness. The complex consists of two subdomains, the first of which, S1, has a receptor binding domain (RBD) with a high affinity for mammalian ACE2, also known as angiotensin converting enzyme 2, or ACE2.

E (envelope) **genes** are unique to the COVID-19-causing SARS-CoV-2 virus. You have COVID if these viral targets are found. The Cycle Threshold (Ct), which describes how many times the E gene targets must be amplified before they can be detected, is represented by the values mentioned alongside the targets. These numbers serve as the foundation for the test's interpretation of "detected" or "not detected."

Simply put, the coronavirus's **N gene**, also known as a nucleocapsid, is a kind of structural protein. Several viruses' nucleocapsid proteins have been shown to play a variety of regulatory roles throughout viral pathogenesis. They have distinctive structural motifs and/or signature sequences that enable them to interact with other viral and host components and skew the host cellular machinery in a way that makes it more conducive to the virus' survival.

The membrane (**M gene**) glycoprotein, which spans the membrane bilayer three times, is the most prevalent structural protein. It leaves a short NH2-terminal domain outside the virus and a lengthy COOH terminus (cytoplasmic domain) inside the virion.

Process

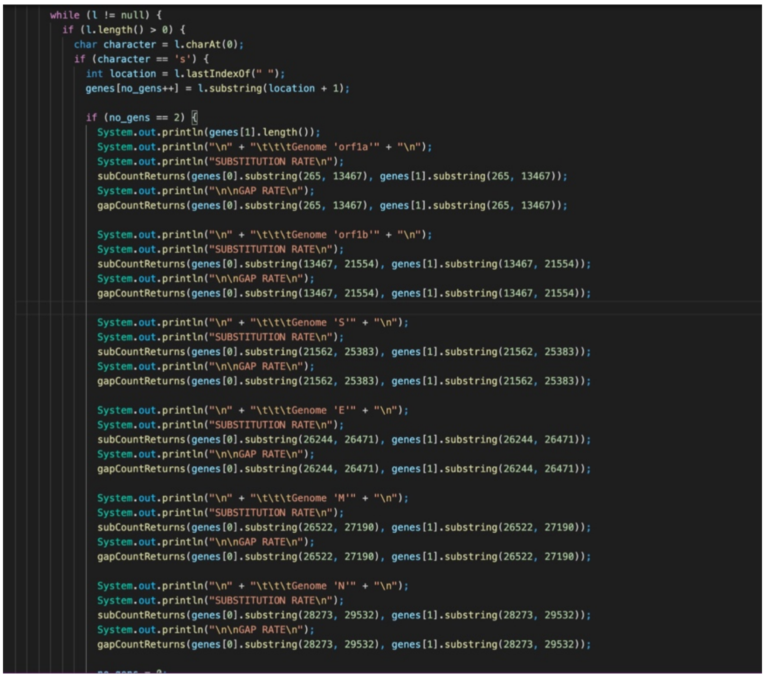
Each of the six genomes' beginning and ending positions is used, and we compare each letter in one sequence (the wild type data set) with each letter in the other sequence (i.e. omicron variant dataset). When comparing the sequences, we skip the gaps and do not count them. Instead, we diagonally match the next letter with the top sequence.

We start by taking the beginning and ending positions of the Orf1a genome, after which we count and break the sequence. Now, this sequence is considered as the Orf1a sub dataset and we perform substitution and gap rate analysis. In a similar manner, we count and separate the sequences at the start and end points of the remaining 5 genomes. After that, we use this sequence as a subset of the sample and run substitution and gap rate analysis on it.

In addition to gap, substitution rate analysis, we also find match count, mismatch count, transitions count and transversion count. The Ti (transitions)/ Tv (transversion) ratio is calculated for each of the six sub data sets.

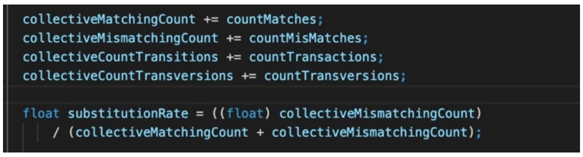
- Match Count means the number of matching residues in the alignment
- Mismatch Count means the number of mismatched residues in the alignment.
- Transitions involve bases with a similar pattern because they are interchanges of one- or two-ring pyrimidine's (A G or C T). Another kind of base substitution is a transversion, in which one base from one class transforms into another base from a different class. So, the purines become pyrimidine's and the pyrimidine's become purines.
- Ti/Tv means the ratio of the number of transitions to the number of transversions. If the distribution of transition and transversion mutations were random (i.e. without any biological influence) we would expect a ratio of 0.5.

CODE



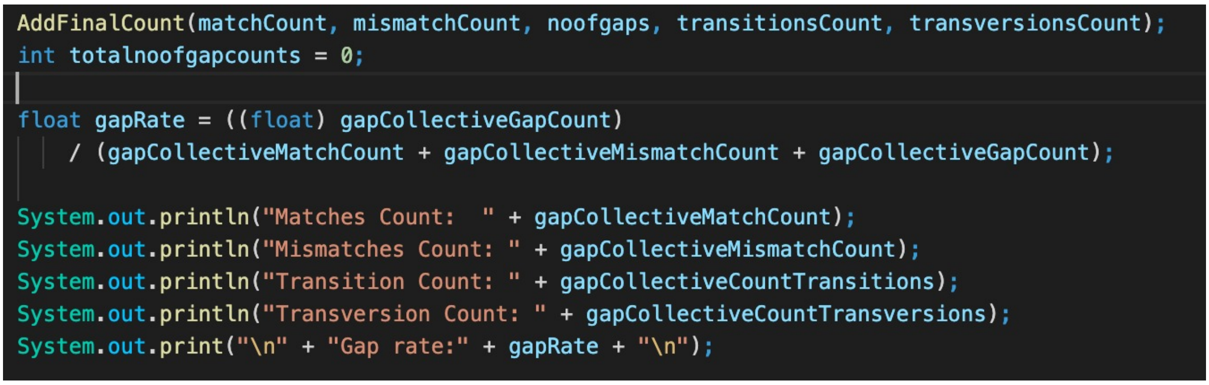
```
while (l != null) {
  if (l.length() > 0) {
    char character = l.charAt(0);
    if (character == 's') {
      int location = l.lastIndexOf(" ");
      genes[no_gens++] = l.substring(location + 1);

      if (no_gens == 2) {
        System.out.println(genes[1].length());
        System.out.println("\n" + "\t\t\tGenome 'orf1a'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(265, 13467), genes[1].substring(265, 13467));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(265, 13467), genes[1].substring(265, 13467));

        System.out.println("\n" + "\t\t\tGenome 'orf1b'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(13467, 21554), genes[1].substring(13467, 21554));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(13467, 21554), genes[1].substring(13467, 21554));

        System.out.println("\n" + "\t\t\tGenome 'S'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(21562, 25383), genes[1].substring(21562, 25383));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(21562, 25383), genes[1].substring(21562, 25383));

        System.out.println("\n" + "\t\t\tGenome 'E'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(26244, 26471), genes[1].substring(26244, 26471));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(26244, 26471), genes[1].substring(26244, 26471));

        System.out.println("\n" + "\t\t\tGenome 'M'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(26522, 27190), genes[1].substring(26522, 27190));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(26522, 27190), genes[1].substring(26522, 27190));

        System.out.println("\n" + "\t\t\tGenome 'N'" + "\n");
        System.out.println("SUBSTITUTION RATE\n");
        subCountReturns(genes[0].substring(28273, 29532), genes[1].substring(28273, 29532));
        System.out.println("\n\nGAP RATE\n");
        gapCountReturns(genes[0].substring(28273, 29532), genes[1].substring(28273, 29532));

        no_gens = 0;
```

- Breaking down the input dataset into six data sets based on sarsCov2structure.txt file



```
collectiveMatchingCount += countMatches;
collectiveMismatchingCount += countMisMatches;
collectiveCountTransitions += countTransactions;
collectiveCountTransversions += countTransversions;

float substitutionRate = ((float) collectiveMismatchingCount)
    / (collectiveMatchingCount + collectiveMismatchingCount);
```

- Code for Substitution Rate calculation



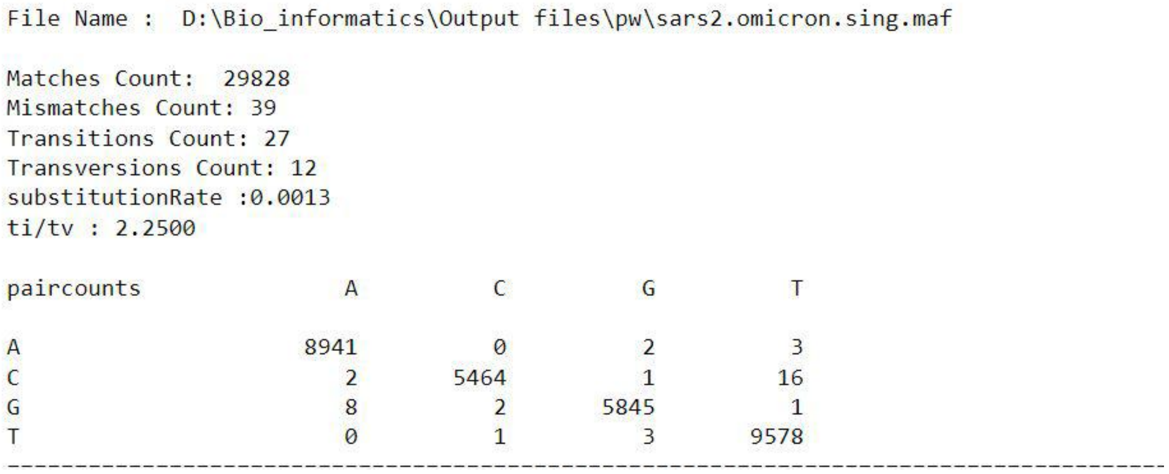
```
AddFinalCount(matchCount, mismatchCount, noofgaps, transitionsCount, transversionsCount);
int totalnoofgapcounts = 0;

float gapRate = ((float) gapCollectiveGapCount)
    / (gapCollectiveMatchCount + gapCollectiveMismatchCount + gapCollectiveGapCount);

System.out.println("Matches Count:  " + gapCollectiveMatchCount);
System.out.println("Mismatches Count: " + gapCollectiveMismatchCount);
System.out.println("Transition Count: " + gapCollectiveCountTransitions);
System.out.println("Transversion Count: " + gapCollectiveCountTransversions);
System.out.print("\n" + "Gap rate:" + gapRate + "\n");
```

- Code for Gap Rate calculation

Results

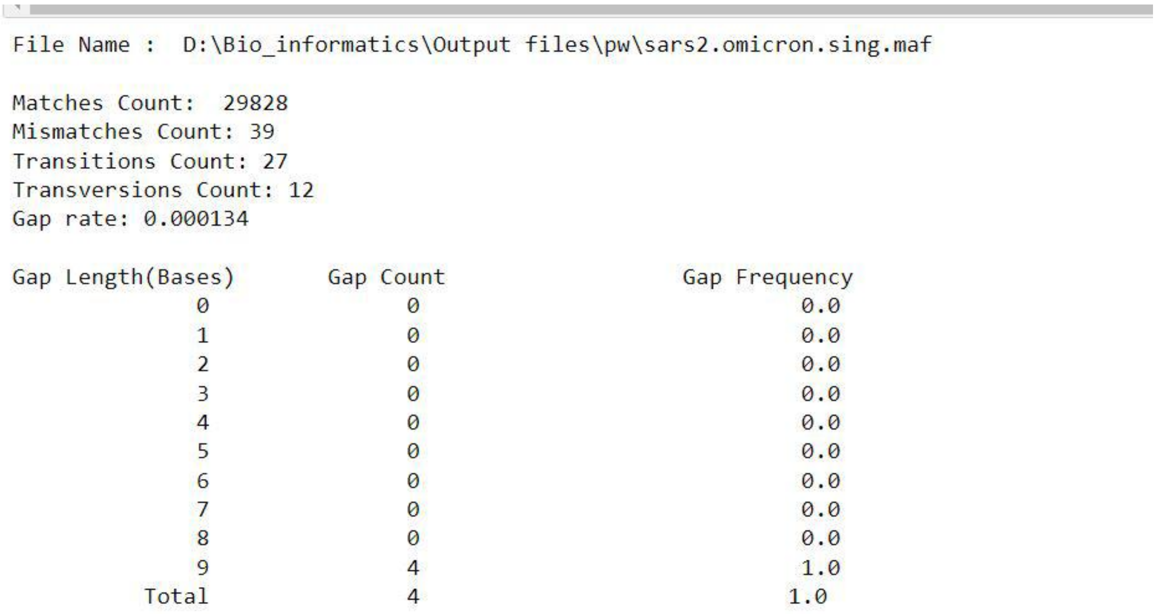
- **Substitution rate and Gap rate results for sar2 and Omicron dataset**

Substitution Rate

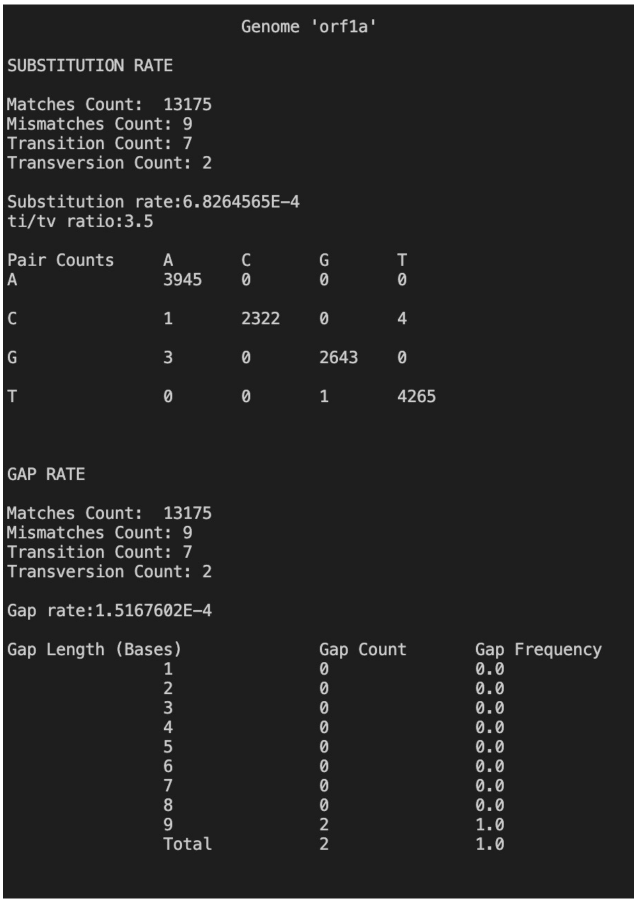


```
File Name :  D:\Bio_informatics\Output files\pw\sars2.omicron.sing.maf

Matches Count:  29828
Mismatches Count: 39
Transitions Count: 27
Transversions Count: 12
substitutionRate :0.0013
ti/tv : 2.2500

paircounts               A          C          G          T

A                     8941          0          2          3
C                        2       5464          1         16
G                        8          2       5845          1
T                        0          1          3       9578
--------------------------------------------------------------------------------------
```

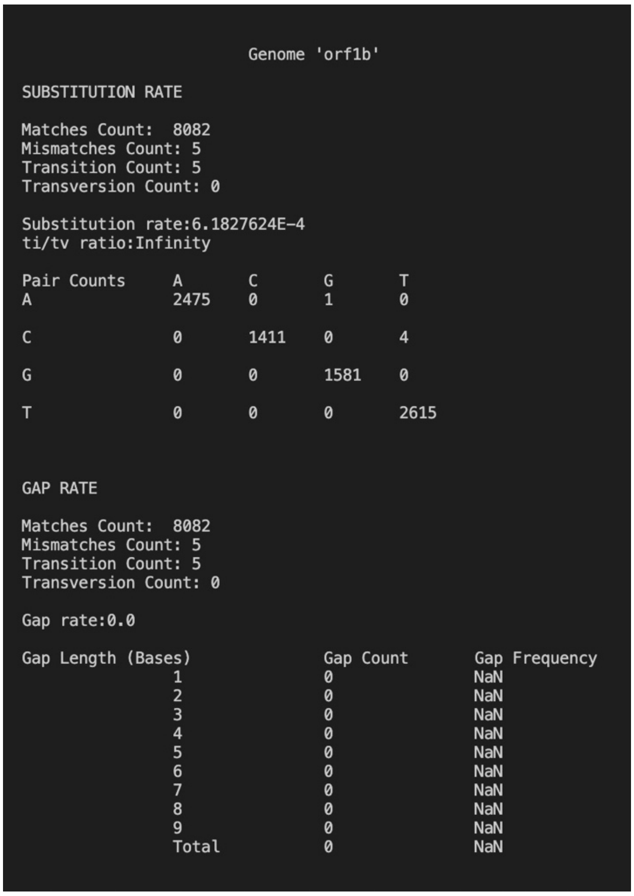
Gap Rate



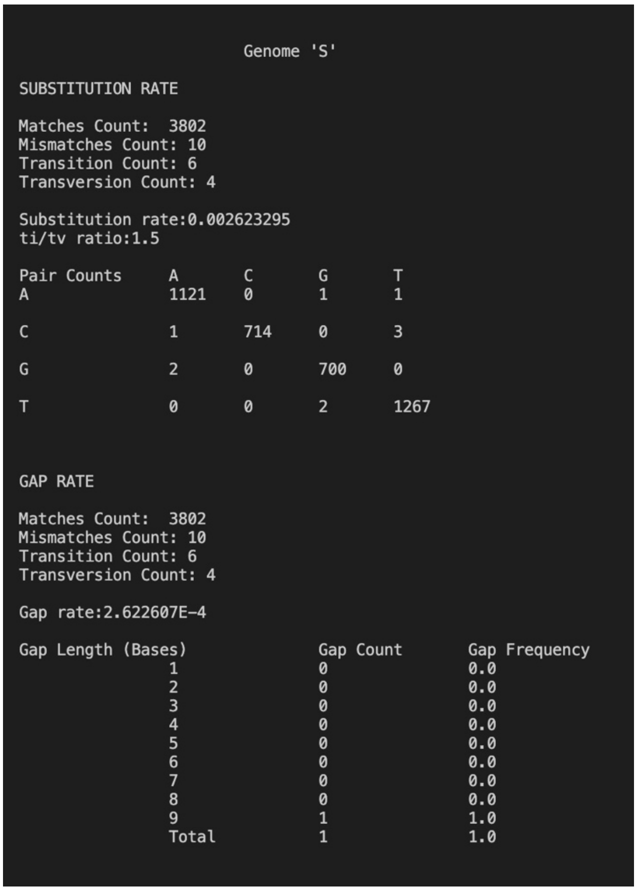
```
File Name :  D:\Bio_informatics\Output files\pw\sars2.omicron.sing.maf

Matches Count:  29828
Mismatches Count: 39
Transitions Count: 27
Transversions Count: 12
Gap rate: 0.000134

Gap Length(Bases)       Gap Count                Gap Frequency
              0              0                           0.0
              1              0                           0.0
              2              0                           0.0
              3              0                           0.0
              4              0                           0.0
              5              0                           0.0
              6              0                           0.0
              7              0                           0.0
              8              0                           0.0
              9              4                           1.0
          Total              4                          1.0
```



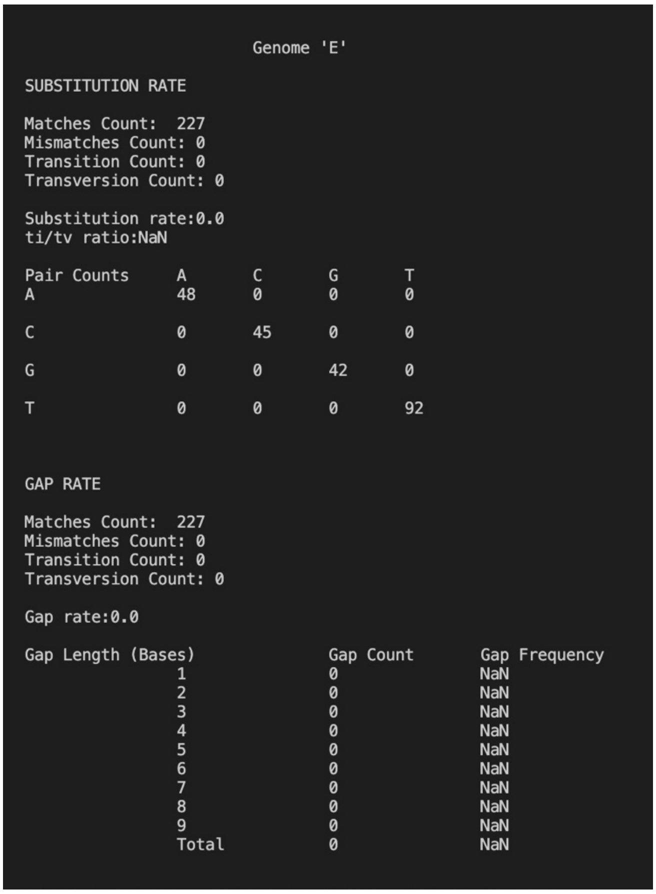
```
                         Genome 'orf1a'

SUBSTITUTION RATE

Matches Count:  13175
Mismatches Count: 9
Transition Count: 7
Transversion Count: 2

Substitution rate:6.8264565E-4
ti/tv ratio:3.5

Pair Counts     A       C       G       T
A               3945    0       0       0

C               1       2322    0       4

G               3       0       2643    0

T               0       0       1       4265



GAP RATE

Matches Count:  13175
Mismatches Count: 9
Transition Count: 7
Transversion Count: 2

Gap rate:1.5167602E-4

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               0.0
                2               0               0.0
                3               0               0.0
                4               0               0.0
                5               0               0.0
                6               0               0.0
                7               0               0.0
                8               0               0.0
                9               2               1.0
                Total           2               1.0
```



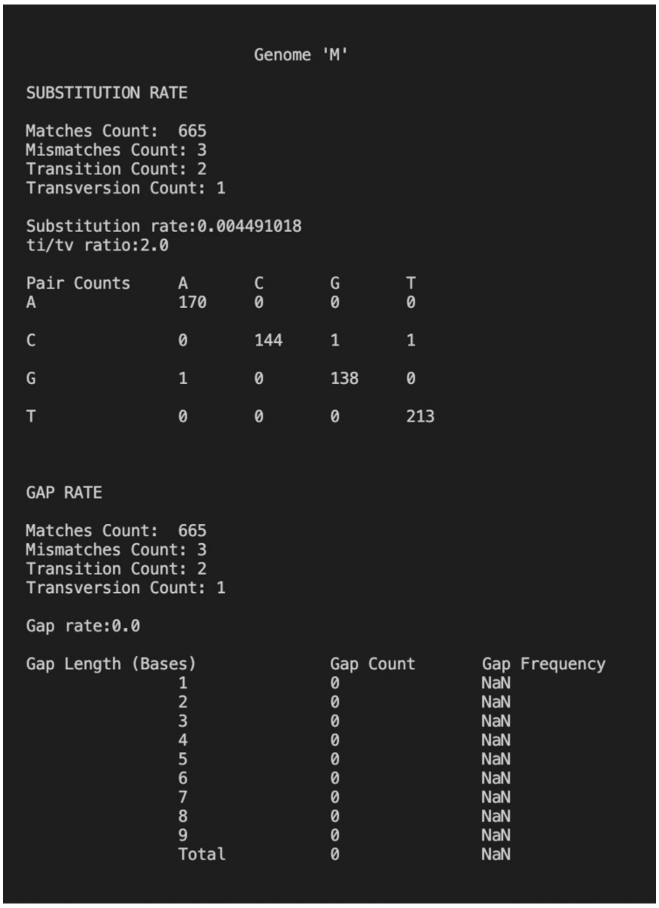
```
                         Genome 'orf1b'

SUBSTITUTION RATE

Matches Count:  8082
Mismatches Count: 5
Transition Count: 5
Transversion Count: 0

Substitution rate:6.1827624E-4
ti/tv ratio:Infinity

Pair Counts     A       C       G       T
A               2475    0       1       0

C               0       1411    0       4

G               0       0       1581    0

T               0       0       0       2615



GAP RATE

Matches Count:  8082
Mismatches Count: 5
Transition Count: 5
Transversion Count: 0

Gap rate:0.0

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               NaN
                2               0               NaN
                3               0               NaN
                4               0               NaN
                5               0               NaN
                6               0               NaN
                7               0               NaN
                8               0               NaN
                9               0               NaN
                Total           0               NaN
```



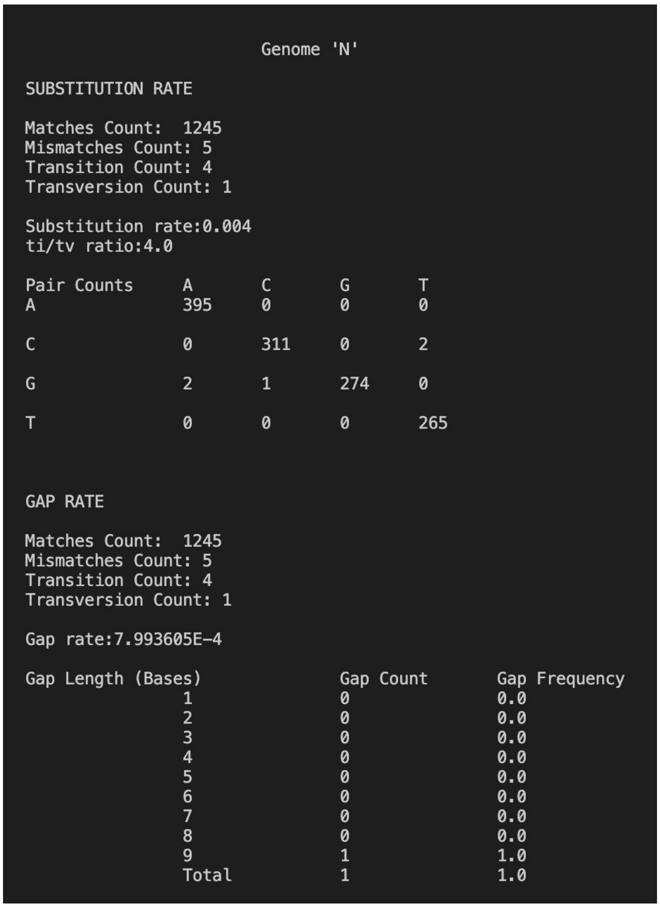
```
                         Genome 'S'

SUBSTITUTION RATE

Matches Count:  3802
Mismatches Count: 10
Transition Count: 6
Transversion Count: 4

Substitution rate:0.002623295
ti/tv ratio:1.5

Pair Counts     A       C       G       T
A               1121    0       1       1

C               1       714     0       3

G               2       0       700     0

T               0       0       2       1267



GAP RATE

Matches Count:  3802
Mismatches Count: 10
Transition Count: 6
Transversion Count: 4

Gap rate:2.622607E-4

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               0.0
                2               0               0.0
                3               0               0.0
                4               0               0.0
                5               0               0.0
                6               0               0.0
                7               0               0.0
                8               0               0.0
                9               1               1.0
                Total           1               1.0
```



```
                         Genome 'E'

SUBSTITUTION RATE

Matches Count:  227
Mismatches Count: 0
Transition Count: 0
Transversion Count: 0

Substitution rate:0.0
ti/tv ratio:NaN

Pair Counts     A       C       G       T
A               48      0       0       0

C               0       45      0       0

G               0       0       42      0

T               0       0       0       92



GAP RATE

Matches Count:  227
Mismatches Count: 0
Transition Count: 0
Transversion Count: 0

Gap rate:0.0

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               NaN
                2               0               NaN
                3               0               NaN
                4               0               NaN
                5               0               NaN
                6               0               NaN
                7               0               NaN
                8               0               NaN
                9               0               NaN
                Total           0               NaN
```

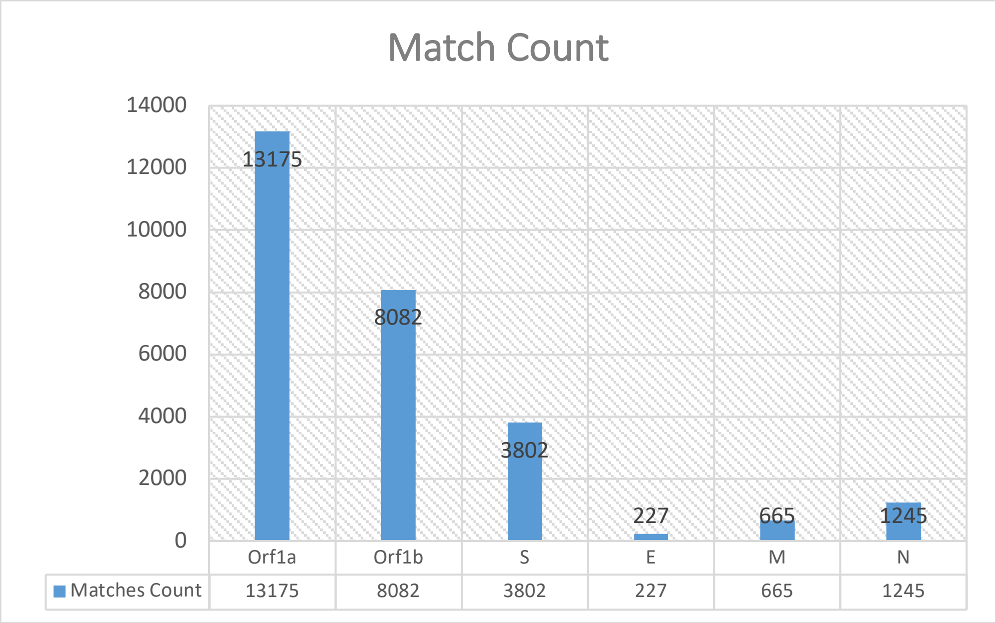


```
                         Genome 'M'

SUBSTITUTION RATE

Matches Count:  665
Mismatches Count: 3
Transition Count: 2
Transversion Count: 1

Substitution rate:0.004491018
ti/tv ratio:2.0

Pair Counts     A       C       G       T
A               170     0       0       0

C               0       144     1       1

G               1       0       138     0

T               0       0       0       213


GAP RATE

Matches Count:  665
Mismatches Count: 3
Transition Count: 2
Transversion Count: 1

Gap rate:0.0

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               NaN
                2               0               NaN
                3               0               NaN
                4               0               NaN
                5               0               NaN
                6               0               NaN
                7               0               NaN
                8               0               NaN
                9               0               NaN
                Total           0               NaN
```



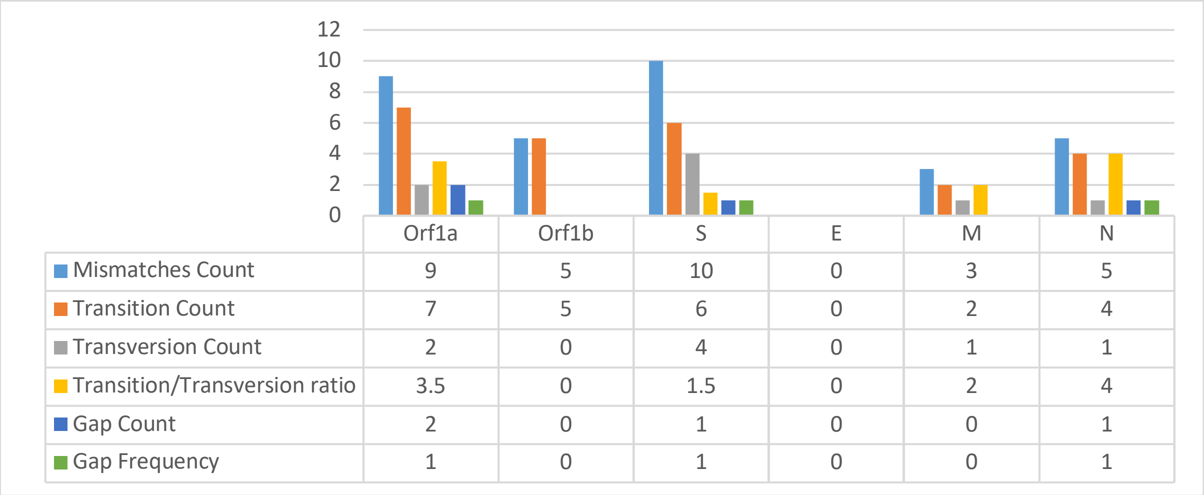
```
                        Genome 'N'

SUBSTITUTION RATE

Matches Count:  1245
Mismatches Count: 5
Transition Count: 4
Transversion Count: 1

Substitution rate:0.004
ti/tv ratio:4.0

Pair Counts     A       C       G       T
A               395     0       0       0

C               0       311     0       2

G               2       1       274     0

T               0       0       0       265


GAP RATE

Matches Count:  1245
Mismatches Count: 5
Transition Count: 4
Transversion Count: 1

Gap rate:7.993605E-4

Gap Length (Bases)              Gap Count       Gap Frequency
                1               0               0.0
                2               0               0.0
                3               0               0.0
                4               0               0.0
                5               0               0.0
                6               0               0.0
                7               0               0.0
                8               0               0.0
                9               1               1.0
                Total           1               1.0
```

Graph Analysis



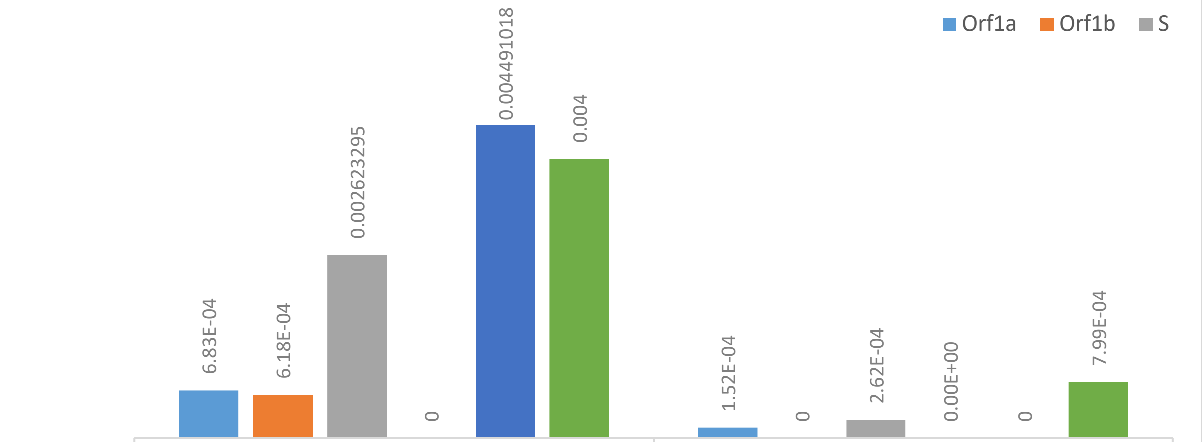
	Orf1a	Orf1b	S	E	M	N
🟦 Matches Count	13175	8082	3802	227	665	1245

The above graph represents the match count for each of the six genomes respectively. As you can see, Gene E has the lowest count when compared to other genes, whereas Orf1a has a higher match count than the other genes.



	Orf1a	Orf1b	S	E	M	N
🟦 Mismatches Count	9	5	10	0	3	5
🟧 Transition Count	7	5	6	0	2	4
⬜ Transversion Count	2	0	4	0	1	1
🟨 Transition/Transversion ratio	3.5	0	1.5	0	2	4
🟦 Gap Count	2	0	1	0	0	1
🟩 Gap Frequency	1	0	1	0	0	1

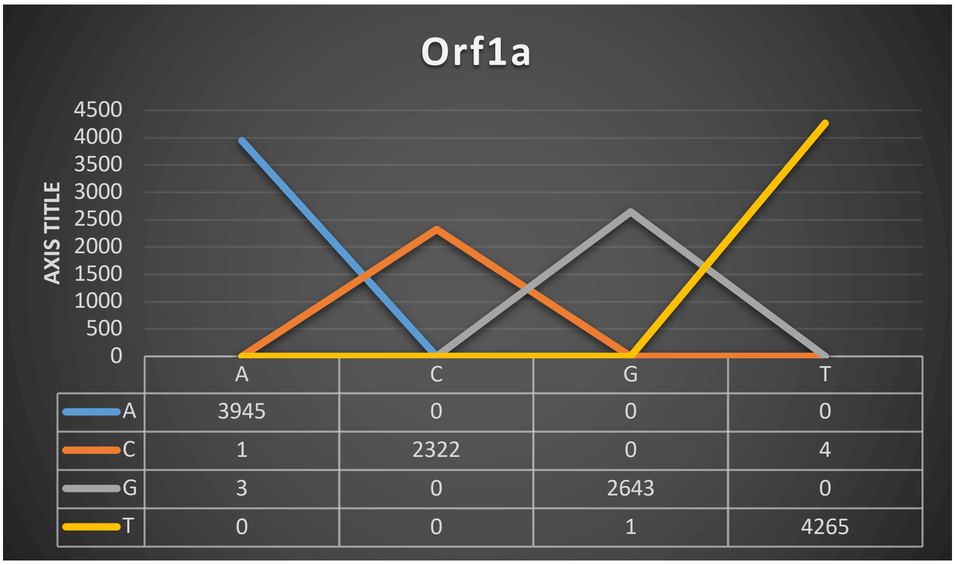
The above graph displays the number of matches, the number of transitions, the number of transversions, the Ti/Tv ratio, the number of gaps, and the Gap frequency for each of the six genomes. We can see that the E gene has zero value for mismatch, transition, and the other attributes.



	Substitution rate	Gap rate
🟦 Orf1a	6.83E-04	1.52E-04
🟧 Orf1b	6.18E-04	0
■ S	0.002623295	2.62E-04
🟨 E	0	0.00E+00
🟦 M	0.004491018	0
🟩 N	0.004	7.99E-04

The substitution and gap rates for the six genomes are represented in the above graph, which is a representation of the whole genome analysis. We can conclude that the E gene has zero value in terms of substitution and gap rate.

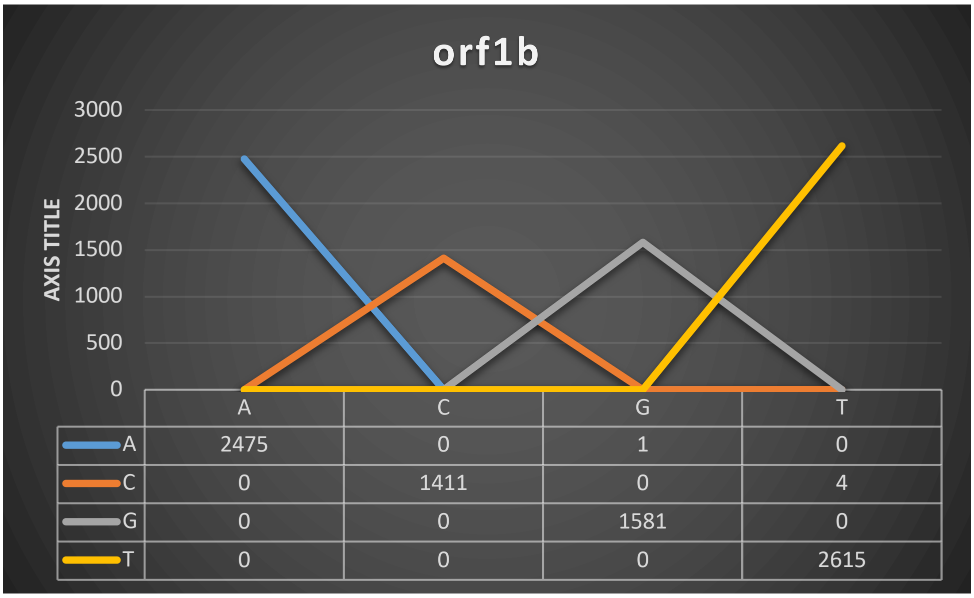
Substitution Rate Graphs



	A	C	G	T
A	3945	0	0	0
C	1	2322	0	4
G	3	0	2643	0
T	0	0	1	4265

Substitution rate: 6.83E-04

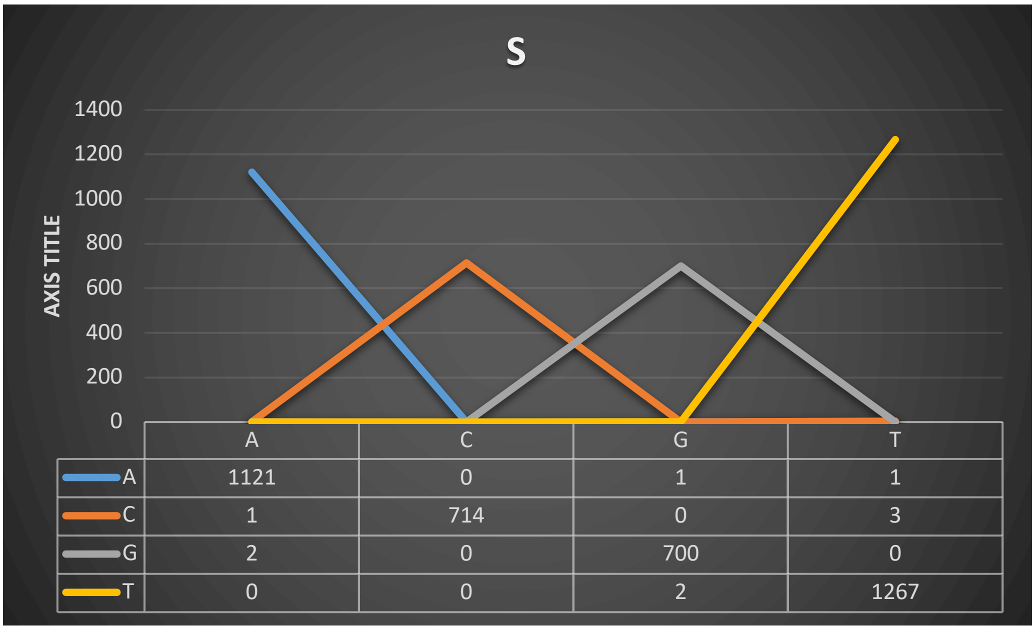
ti/tv ratio: 3.5



	A	C	G	T
A	2475	0	1	0
C	0	1411	0	4
G	0	0	1581	0
T	0	0	0	2615

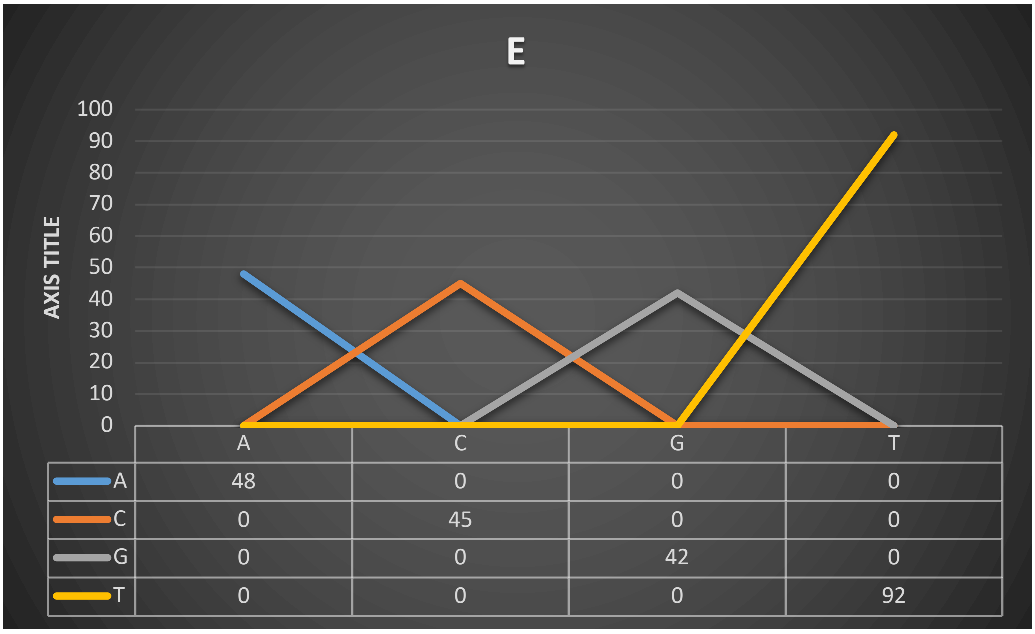
Substitution rate: 6.18E-04

ti/tv ratio: Infinity



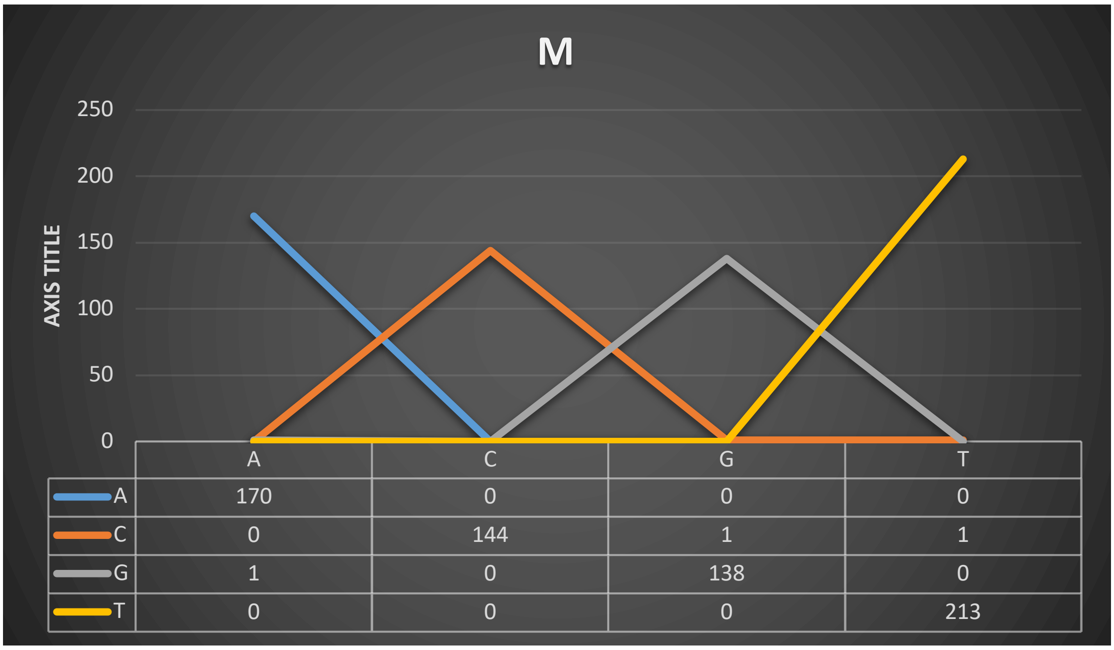
	A	C	G	T
A	1121	0	1	1
C	1	714	0	3
G	2	0	700	0
T	0	0	2	1267

Substitution rate:	0.0026233
ti/tv ratio:	1.5



	A	C	G	T
A	48	0	0	0
C	0	45	0	0
G	0	0	42	0
T	0	0	0	92

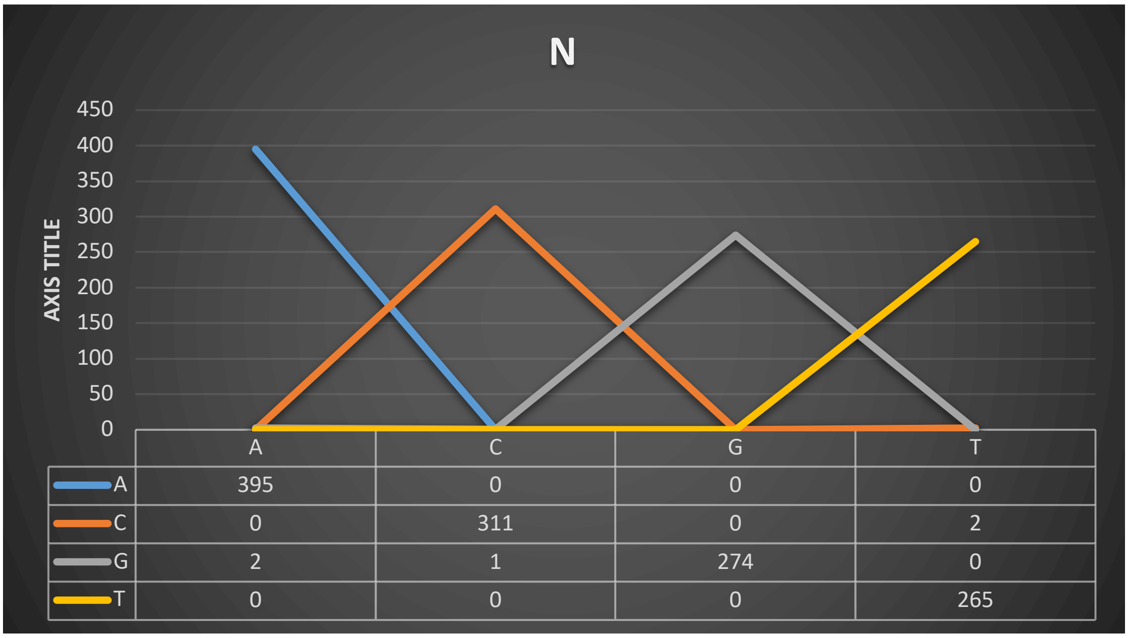
Substitution rate:	0.0
ti/tv ratio:	NaN



	A	C	G	T
A	170	0	0	0
C	0	144	1	1
G	1	0	138	0
T	0	0	0	213

Substitution rate: 0.00449102

ti/tv ratio: 2



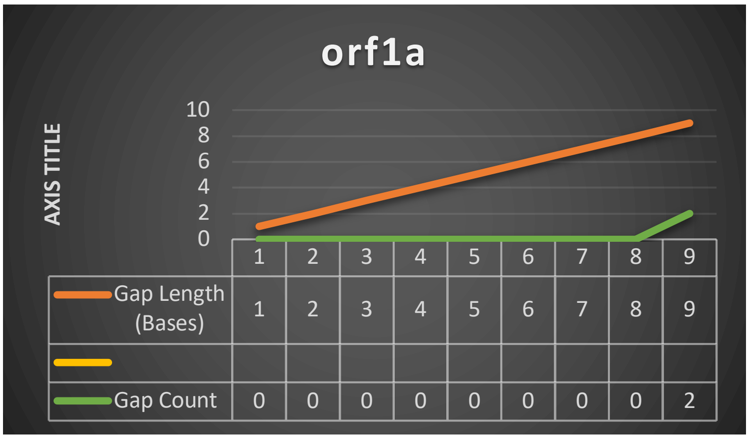
	A	C	G	T
A	395	0	0	0
C	0	311	0	2
G	2	1	274	0
T	0	0	0	265

Substitution rate: 0.004

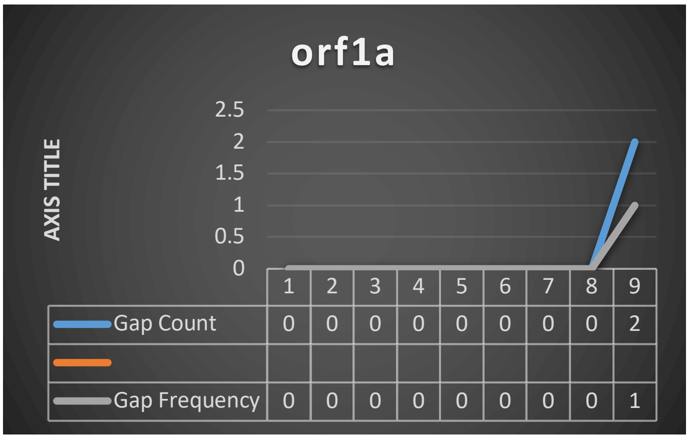
ti/tv ratio: 4

Gap Rate Graph

Graph between Gap length and Gap count

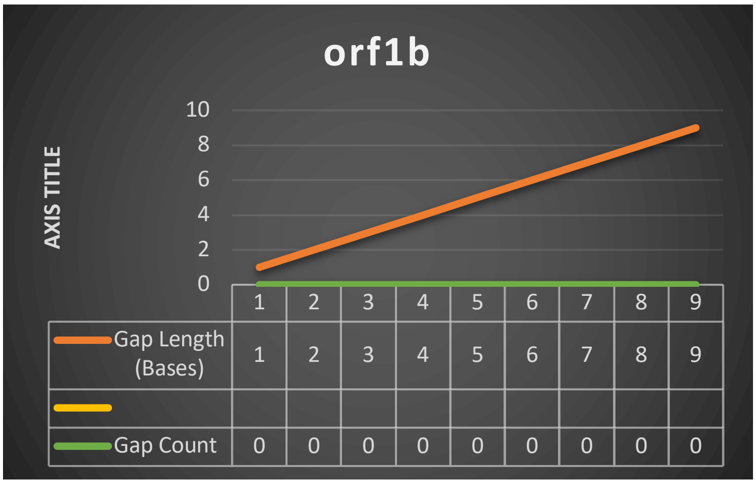


Graph between Gap count & Gap Frequency

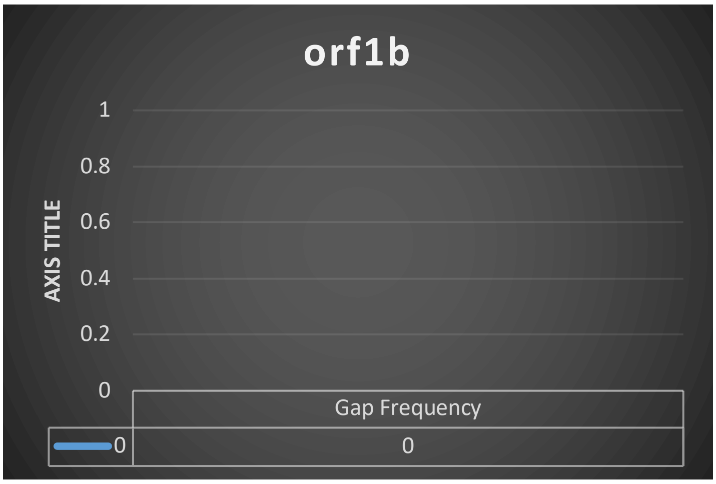


Gap rate: 1.52E-04

Graph between Gap length and Gap count

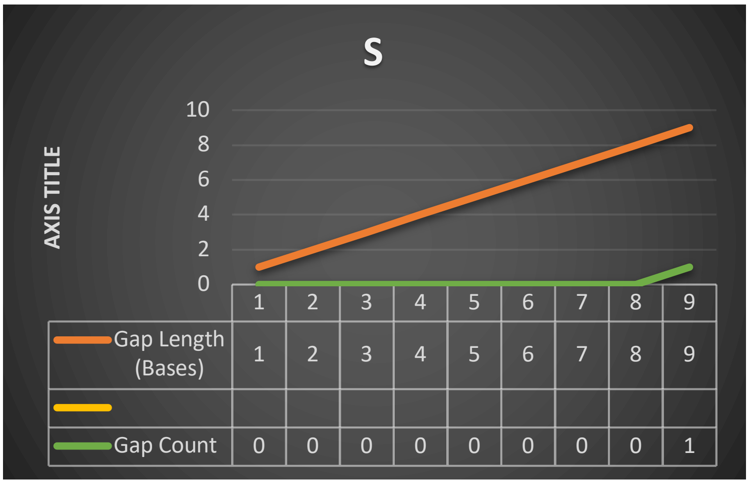


Graph between Gap count & Gap Frequency

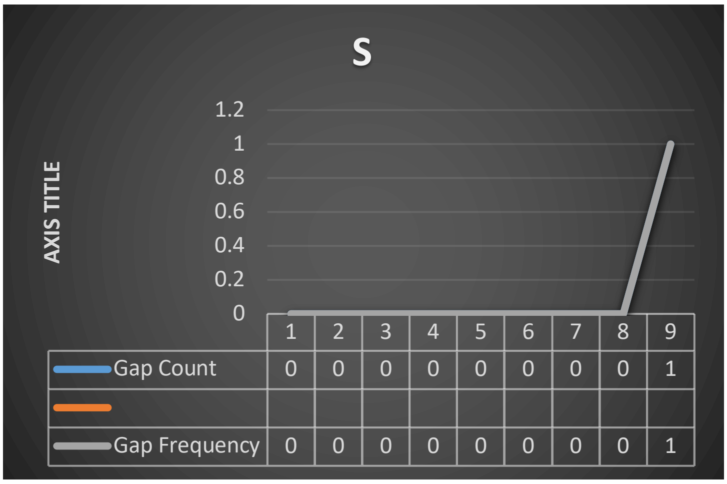


Gap rate: 0

Graph between Gap length and Gap count

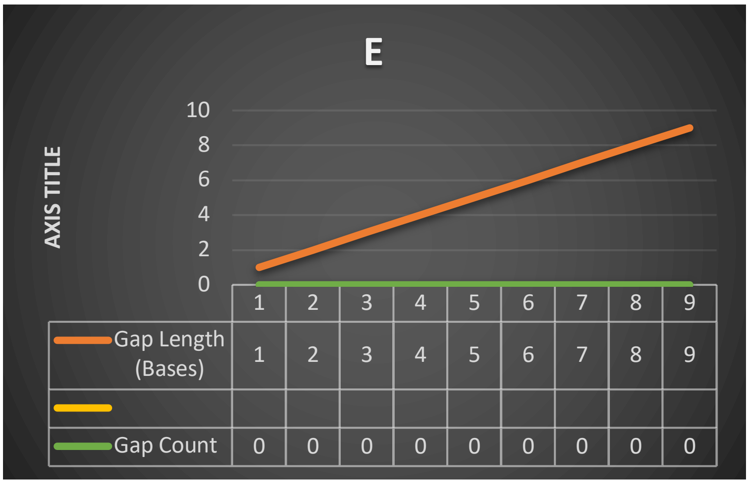


Graph between Gap count & Gap Frequency

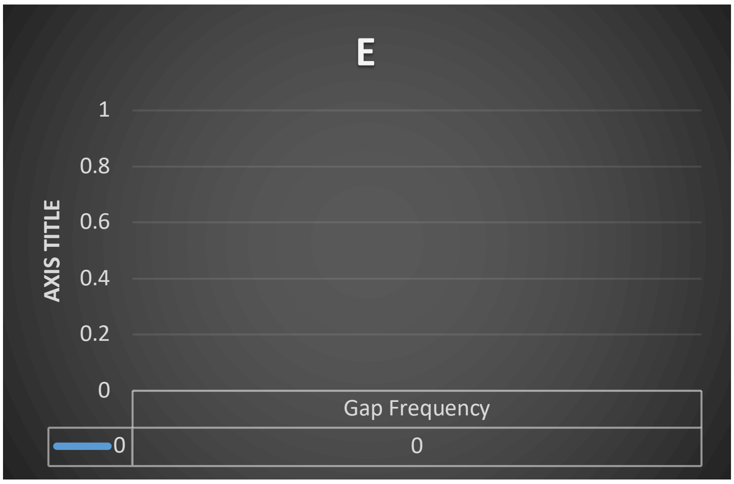


Gap rate: 2.62E-04

Graph between Gap length and Gap count

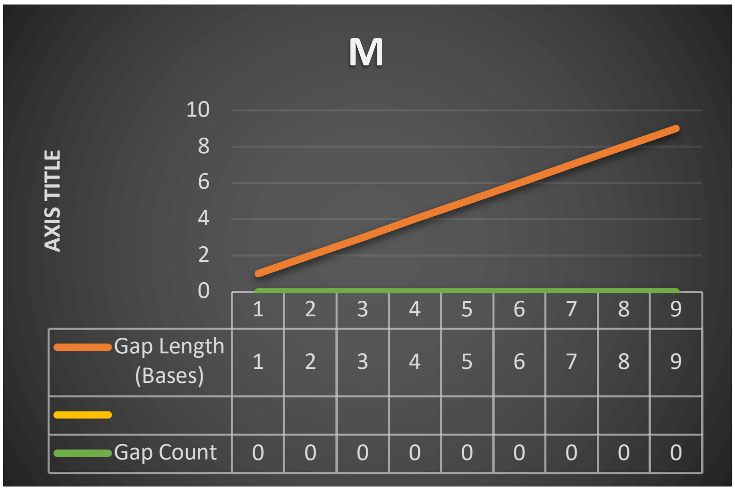


Graph between Gap count & Gap Frequency

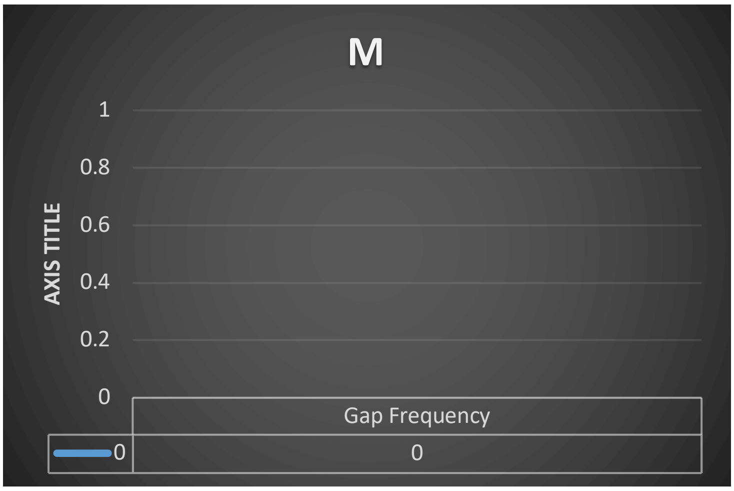


Gap rate: 0.00E+00

Graph between Gap length and Gap count

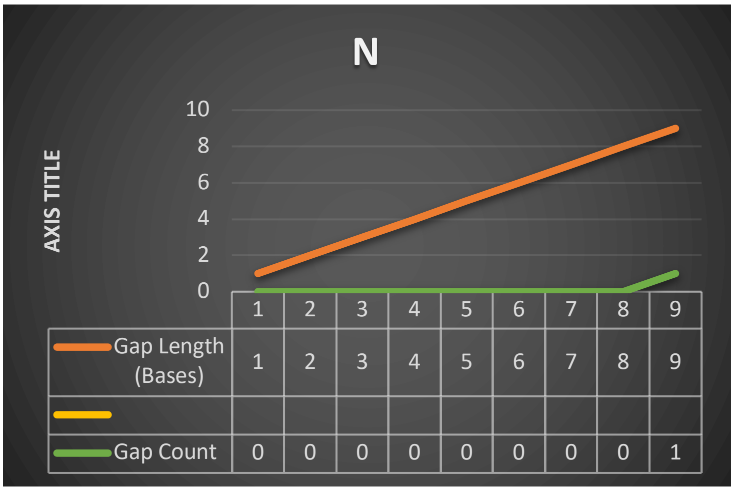


Graph between Gap count & Gap Frequency

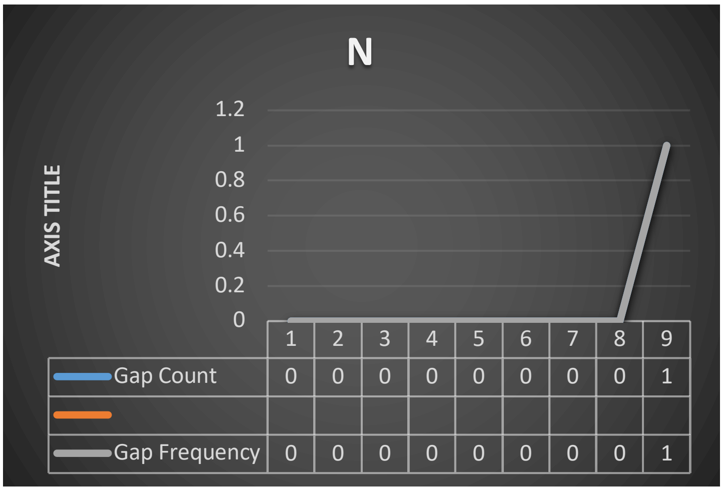


Gap rate: 0.00E+00

Graph between Gap length and Gap count



Graph between Gap count & Gap Frequency



Gap rate: 7.99E-04

Team member's contribution:

Project management:	Mahima (30%)	Dinesh (35%)	Ganesh (35%)
Implementation:	Ganesh (35%)	Dinesh (35%)	Mahima (30%)
Design & Analyses:	Dinesh (35%)	Mahima (35%)	Ganesh (30%)
Report writing:	Dinesh (35%)	Mahima (35%)	Ganesh (30%)