*Article*

# Decoding Influenza: A Bioinformatics Inquiry into the Mysteries of Flu Infection

**Abstract:** This study investigates the occurrence of flu infection in an individual vaccinated against the virus, focusing on the potential evolution of the virus through antigenic drift. Utilizing deep sequencing technology, the project analyzes the hemagglutinin gene of the influenza virus from a patient who did not receive the vaccine. The approach includes downloading and aligning viral sequence data with a reference sequence and identifying both common and rare genetic variants using VarScan. Additionally, awk one-liners are employed for efficient data parsing. The research aims to elucidate the mechanisms behind the flu virus's capacity to escape vaccine-induced immunity, highlighting the significance of continuous flu vaccine evaluation and modification. This investigation serves as a practical application of bioinformatics in understanding viral evolution and its implications for vaccine efficacy.

## 1. Introduction

Influenza, commonly known as the flu, is a contagious respiratory illness caused by influenza viruses. It poses a significant public health challenge globally due to its capacity for rapid spread and mutation. The flu vaccine, designed to protect against the most prevalent and virulent strains, is a key tool in controlling the spread of the disease. However, the effectiveness of the vaccine can be compromised by the phenomenon of antigenic drift - a process where gradual genetic changes in the virus result in new strains that the immune system does not recognize [1,2].

Antigenic drift is a critical factor in the flu virus's ability to evade the immune response induced by vaccination [3]. This project delves into an intriguing case where an individual, despite being vaccinated, contracts the flu. This raises questions about the potential mismatch between the vaccine strain and the circulating virus, possibly due to antigenic drift.

The study of viral quasispecies, which refers to a population of viruses with genetic variability, is essential in understanding how mutations contribute to the virus's ability to escape vaccine-induced immunity. Quasispecies dynamics are particularly relevant for RNA viruses like the influenza virus, known for their high mutation rates [4].

To investigate these dynamics, this study employs targeted deep sequencing, a powerful tool in modern virology that allows for the detailed analysis of viral genomes. Deep sequencing provides insights into the genetic makeup of virus populations within an individual, including the presence of rare variants that might contribute to vaccine escape. Additionally, this project aims to address the challenges posed by errors

inherent in next-generation sequencing (NGS) technologies. By comparing the viral sequences from the patient with control samples of known sequences, the study seeks to distinguish between real genetic variations and sequencing artifacts.

## 2. Materials and methods

**1. Data Collection and Preparation:**

- **Downloading Sequencing Data:** The sequencing data of the influenza virus from a non-vaccinated individual was retrieved from the NCBI Sequence Read Archive (SRA), specifically the dataset labeled SRR1705851. The data was downloaded from the SRA FTP server and unpacked for analysis.
- **Reference Sequence Acquisition:** The reference sequence for the influenza hemagglutinin gene was obtained from the NCBI GenBank, using the accession number KF848938.1. This sequence was downloaded using the EntrezDirect command 'efetch'.
- **Data availability:** All data analysis steps and files are available by the link.

**2. Quality check:**

- **Quality check** was performed using fastqc tool. Per base sequence quality was good for all samples, so we didn't perform trimming step. There were a high percentage of overrepresented sequences due to PCR amplification of the viral DNA samples.

**3. Sequencing Data Alignment:**

- **Alignment Software:** The alignment of the viral sequencing data to the reference sequence was performed using BWA-MEM alignment algorithm. Prior to alignment BWA index was used to index reference genome..

**4. Variant Analysis with VarScan:**

- **Indexing and mpileup Creation:** SAMtools was used to index the aligned data and create an mpileup file. *Coverage = (Read count * Read length ) / Total genome size* - this formula was used to count average coverage. But `-d` flag was set 0 (*max coverage*) ensure the inclusion of all potential variants.
- **Common and Rare Variant Identification:** VarScan was used to analyze the mpileup file to identify both common and rare variants. The minimum variant frequency threshold was set at 0.95 (95%) for common variants and 0.001 (0.1%) for rare variants.

**5. Pipeline management:**

- **Pipeline managers:** Snakemake was used to organize step 1-3 to one unified pipeline.

**6. Data Parsing and Analysis:**

- **Use of awk for Parsing:** Awk one-liners were utilized for extracting specific data fields from the VarScan output, particularly focusing on the reference, position, and alternate fields in the variant call format (VCF) files.

- **Manual Analysis of Mutations:** The reference and VCF files were examined using the Integrative Genomics Viewer (IGV) and a codon table to assess the potential impact of the identified mutations on the protein structure.

**7. Control Sample Analysis:**

- **Control Data Acquisition and Processing:** Sequencing data from isogenic viral samples were obtained and processed similarly to the main sample. These data were used to differentiate between true viral mutations and sequencing errors.
- **Alignment and Variant Analysis of Control Samples:** Each control sample was aligned to the reference sequence, and VarScan was used to identify rare variants in these samples with 0.001 (0.1%) minimum variant frequency .

**8. Comparative Analysis:**

- **Data Integration and Comparison:** The frequencies of variants in both the patient and control samples were compared to identify significant mutations. *ggplot2*, *dplyr*, *tidyr*, *stringr*, *ComplexUpset* packages from R programming language were implemented to perform analysis for determination of variants in the patient sample that significantly deviated from those observed in the control samples.

**9. Software and Tools:**

- **Bioinformatics Tools:** The analysis utilized various bioinformatics tools, including fastqc, BWA-MEM, SAMtools, VarScan, and awk, Snakemake for pipeline management, as well as the IGV for visual inspection of alignments and variants, Rstudio for comparative data analysis.

## 3. Results

**Sequencing Data Analysis:**

**Data Alignment and Mapping:** The deep sequencing data from the patient's sample (SRR1705851) was successfully aligned to the reference influenza hemagglutinin gene (GenBank accession: KF848938.1). A high percentage of the reads mapped effectively to the reference sequence, indicating good quality of the sequencing data. We counted the percentage of reads that mapped = 361116 (99.94%).

| ID | POS | REF | ALT | FREQ (%) |
|---|---|---|---|---|
| KF848938.1 | 72 | A | G | 99.96 |
| KF848938.1 | 117 | C | T | 99.82 |
| KF848938.1 | 774 | T | C | 99.96 |
| KF848938.1 | 999 | C | T | 99.86 |

| | | | | |
|---|---|---|---|---|
| KF848938.1 | 1260 | A | C | 99.94 |

Table 1. Common variants in patient's sequencing.

**Common Variants Identification:** Analysis using VarScan identified several common variants in the patient's sample when compared to the reference sequence (Table 1). These variants were present in a significant portion of the viral population, suggesting potential areas of mutation contributing to vaccine escape (Supplementary Figure 1).

**Rare Variant Analysis:**

Identification of Rare Variants: Setting the minimum variant frequency threshold to 0.001 in VarScan revealed 16 new variants in the patient's viral sample. These variants represented minor populations within the viral quasispecies and their frequency was lower 1%.

Frequency Analysis: The frequencies of these rare variants were recorded, and a comparative analysis with the control samples was conducted to determine their significance.

**Error Rate Estimation:** Analysis of the control samples (isogenic viral sequences) provides a baseline for distinguishing sequencing errors from true variants. The frequency of mutations detected in the control samples was used as a benchmark to assess the authenticity of the variants found in the patient's sample (All table with mutations are in Folder). Using R, we independently calculated average mutation frequence iner that we cutted-off those variant, which were lower then mean + 3σ (Figure 1). SRR1705858, SRR1705859, SRR1705860, as well as standart deviation (sd). Aft

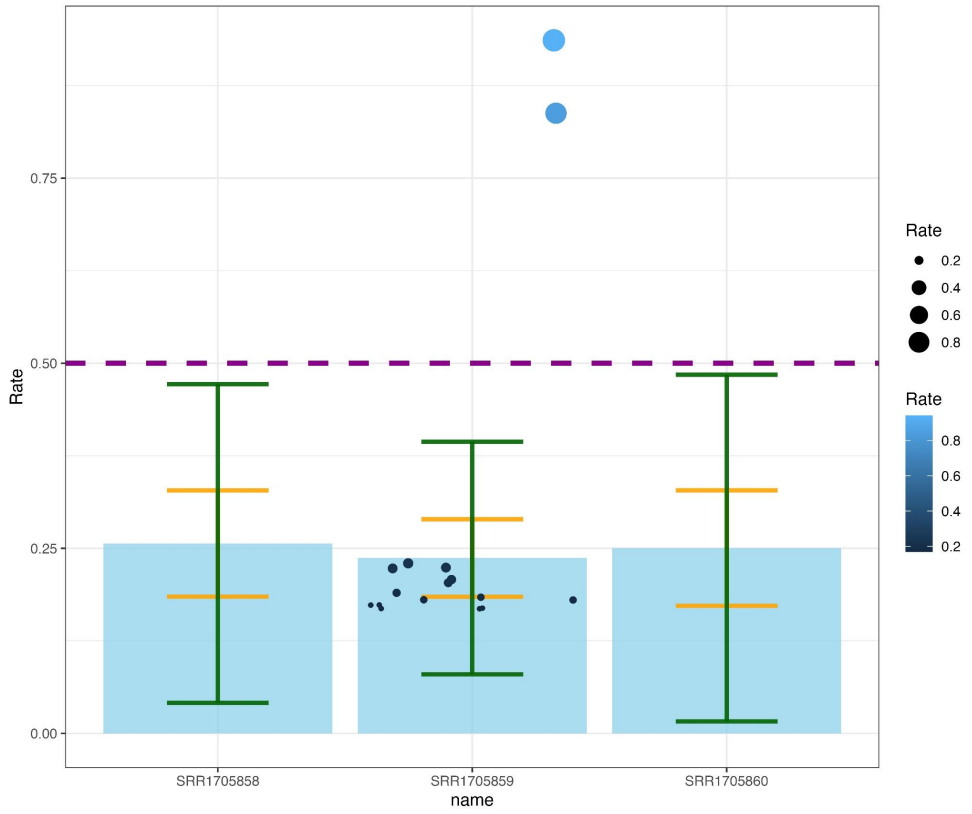



Figure 1. Mutations frequency distribution on control samples. Blue barplot represent mean value of mutations frequency. Yellow error bars represent 1σ, green error bars - 3σ. Purple dashed line represents cut-off which is higher than mean + 3σ. Mutations from our patient a represented with circles, where size and color is established according to mutation FREQ (big and shiny dot have higher FREQ).

After comparative analysis we obtained 7 (5 common + 2 rare) mutations in viral sequencing from patient which are definitely distinguished from sequencing errors.

**Significant Variant Identification:** A few variants in the patient's sample showed frequencies significantly higher than those observed in the control samples, indicating that these were likely true mutations rather than sequencing artifacts. But it is interesting to observe the intersection of this mutations between control samples and patient. We observed 89 unique mutations from which 11 mutations (12%) are common for patient and control samples (Figure 2). 22 mutations (24%) are common for all control samples, which indicates possible PCR errors due to sample preparation. Other non-common mutations in control sampled might occur due to sequencing errors. All positions with each types of intersections are attached as Supplemental Figure 2 (due to high size available by the link).

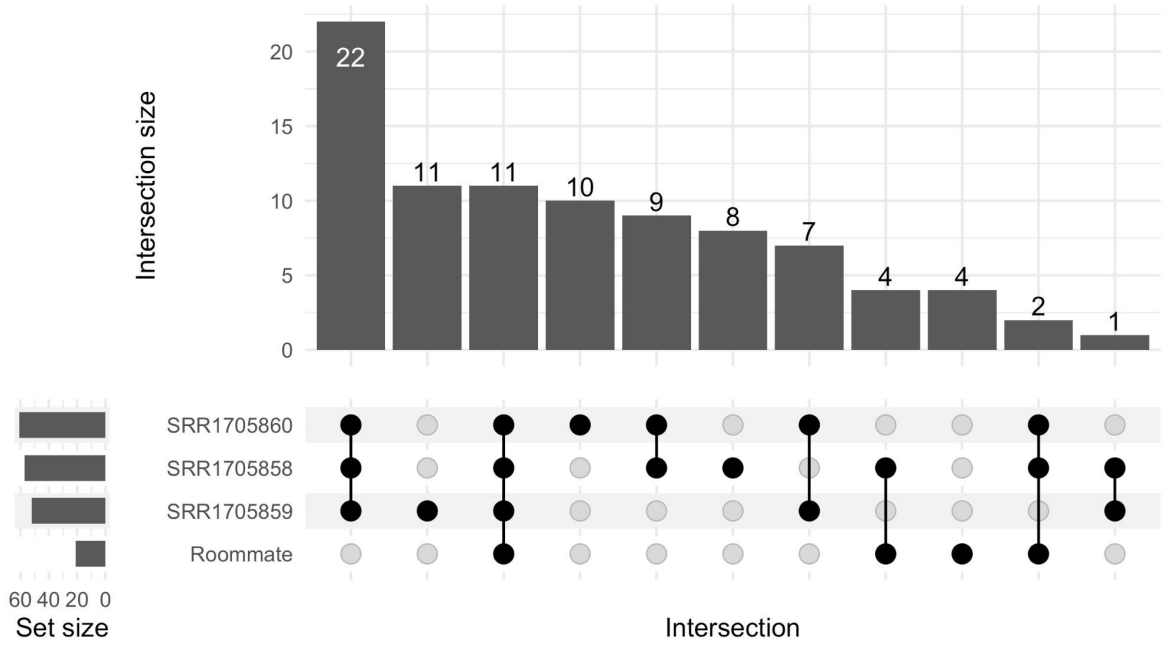
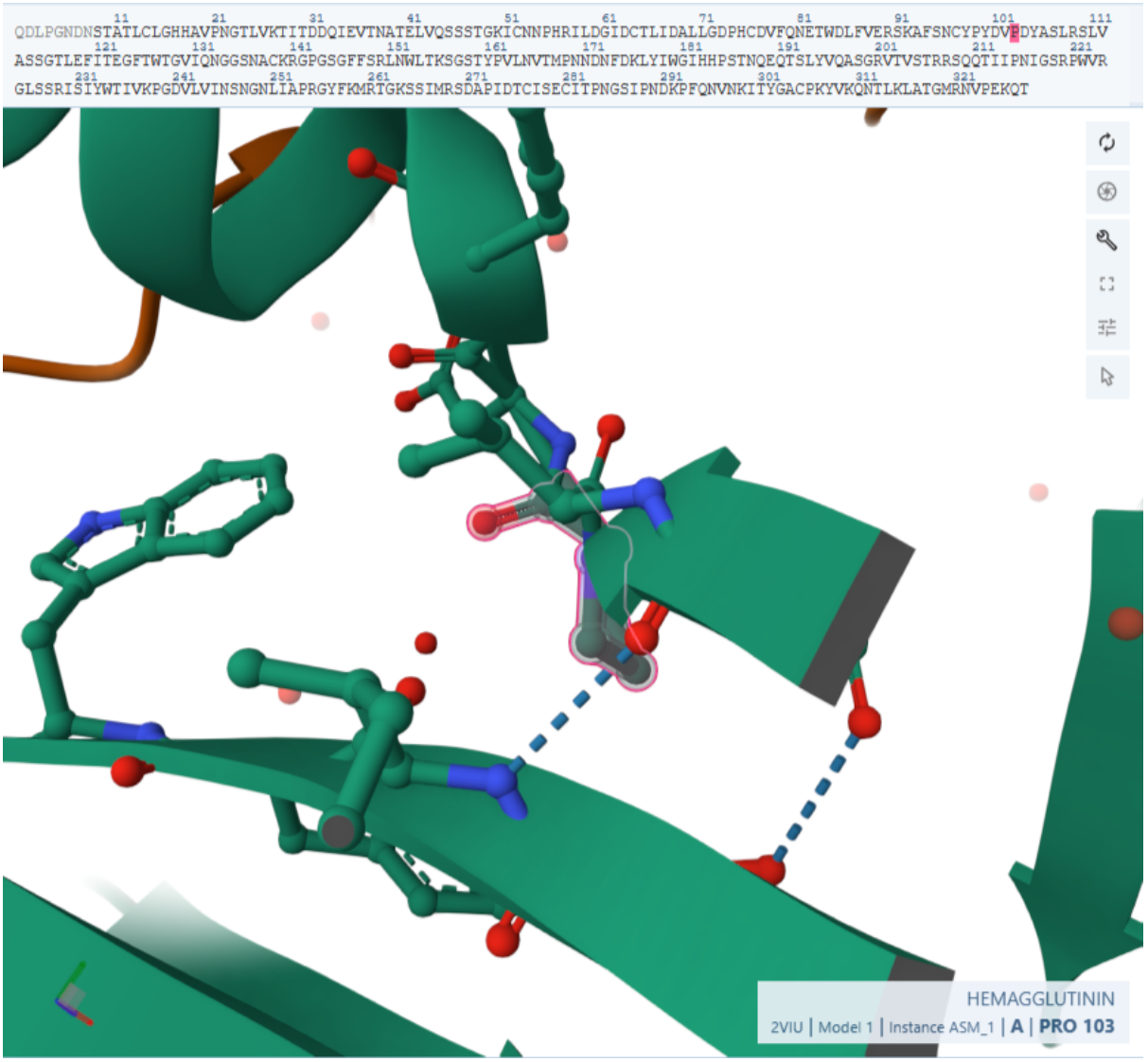



Figure 2. Mutations intersection between patient and control samples. Numbers on grey bars indicate number of mutations in particular intersection that are pictured below each bar.

## Mutational Impact on Protein Function:

**Synonymous vs. Non-synonymous (missense) Mutations:** Examination of the mutations with IGV revealed a mix of synonymous (not affecting the amino acid sequence) and **non-synonymous (missense)** 307 C -> T [Pro -> Ser] (altering the amino acid sequence) mutations (Supplemental Figure 3). The non-synonymous mutation was of particular interest as they could potentially alter the protein structure and function.

**Epitope Mapping:** As synonymous mutations don't alter aminoacids, one missense mutation in 307 nucleotide was in the are of interest in epitope mapping. Epitopes are very important part of the proteins as they represent a part which is recognized by immune system (antibodies). This suggests that these mutations might affect the virus's interaction with antibodies, potentially contributing to the vaccine escape. We performed BLAST alignment of the reference sequence and observed that our reference is lack of first 17 aminoacids comparing with other sequences in NCBI. But we propose that the Munoz et. al.[2] used the same reference, so we performed epitope mapping according to their coordinates. As our missense mutation is located in 307 nucleotide, the 103 aminoacid is substituted [Pro -> Ser]. This substitution may lead to the change of epitope D conformation as proline may affect the protein secondary structure. Moreover, this particular mutation was observed by Cushing A. et. al[5].




## 4. Discussion

Sequencing Data: The patient's sample (SRR1705851) showed a high mapping rate of 99.94% with 361,116 reads mapped.

Mutation Analysis: A total of 7 significant mutations (5 common and 2 rare) were identified in the patient's sample. These mutations were compared to the control samples, revealing 11 common mutations between the patient and control samples. Additionally, 22 mutations common across all control samples were likely due to PCR errors.

Determining Real Mutations: The decision on which mutations were most likely to be real was based on a comparison of their frequencies in the patient's sample against the control samples. Mutations exceeding the mean frequency plus three standard deviations in the control samples were considered significant. This approach helped discern genuine viral mutations from sequencing artifacts.

In line with the methodology detailed in the provided article, we analyzed the high confidence mutations (those greater than 3 standard deviations away from the reference error rate) in our patient's influenza virus sample. Our focus was on the hemagglutinin (HA) protein, specifically its epitope regions, as these are key in the immune system's recognition and response to the virus. Among the mutations identified, the non-synonymous mutation at nucleotide 307 (C -> T) resulted in an amino acid change from Proline to Serine (Pro -> Ser). This mutation was located within an epitope region, as defined by the epitope locations listed in the article [2]. This suggests a potential alteration in epitope conformation, which could impact the virus's interaction with antibodies and explain the vaccine escape in this case. The altered epitope, as a result of the mutation at nucleotide 307, may have changed the protein's structure enough to evade the immune response induced by the vaccine. This suggests that the virus in the patient's roommate, which evolved through antigenic drift, was sufficiently different from the vaccine strain to infect a vaccinated individual.

*Proposals for Error Control in Deep Sequencing:

Laboratory Steps: Implementing more stringent sample preparation protocols can minimize errors. For instance, using high-fidelity enzymes in PCR can reduce the rate of nucleotide misincorporation, thereby lowering error rates.
Bioinformatics Steps: Applying more sophisticated bioinformatics algorithms can enhance the accuracy of mutation detection. Software that incorporates error-correction algorithms can differentiate between true variants and sequencing errors more effectively.
Existing Software Utilization: Utilizing advanced variant calling software, such as GATK or DeepVariant, which are designed to handle the complexities of NGS data, can improve the accuracy of variant identification.
Importance of Error Control:
Error control is crucial in deep sequencing experiments to ensure the reliability of the data, especially when identifying and quantifying rare variants. Accurate identification of these variants is essential for understanding viral evolution, vaccine escape mechanisms, and developing effective vaccines.

In conclusion, the analysis of the influenza virus from the patient's sample, particularly the mutations in the epitope regions of the HA protein, offers valuable insights into the mechanisms of vaccine escape. The findings underscore the importance of continuous monitoring of influenza virus strains and updating vaccine compositions to account for antigenic drift. Moreover, they highlight the need for advanced methodologies in both laboratory practices and bioinformatics analysis to accurately interpret deep sequencing data.

## References

1. Gaitonde DY, Moore FC, Morgan MK. Influenza: Diagnosis and Treatment. Am Fam Physician. 2019 Dec 15;100(12):751-758. PMID: 31845781.
2. Muñoz ET, Deem MW. Epitope analysis for influenza vaccine design. Vaccine. 2005 Jan 19;23(9):1144-8. doi: 10.1016/j.vaccine.2004.08.028. PMID: 15629357; PMCID: PMC4482133.
3. https://www.cdc.gov/flu/about/viruses/change.htm
4. Martínez, M.A., Martrus, G., Capel, E., Parera, M., Franco, S., Nevot, M. (2012). Quasispecies Dynamics of RNA Viruses. In: Witzany, G. (eds) Viruses: Essential Agents of Life. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4899-6_2
5. Cushing, Anna et al. "Emergence of Hemagglutinin Mutations During the Course of Influenza Infection." *Scientific reports* vol. 5 16178. 5 Nov. 2015, doi:10.1038/srep16178

## Data availability

1. http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/
2. https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta
3. SRR1705858:ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz
4. SRR1705859:ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz
5. SRR1705860:ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

## Tools:

NCBI Sequence Read Archive (SRA) and NCBI GenBank: NCBI SRA, NCBI GenBank
EntrezDirect (efetch): NCBI Toolkit
BWA-MEM (Burrows-Wheeler Aligner): BWA GitHub Repository
SAMtools: SAMtools
VarScan: VarScan
Awk
Integrative Genomics Viewer (IGV): IGV
PDB Viewers:

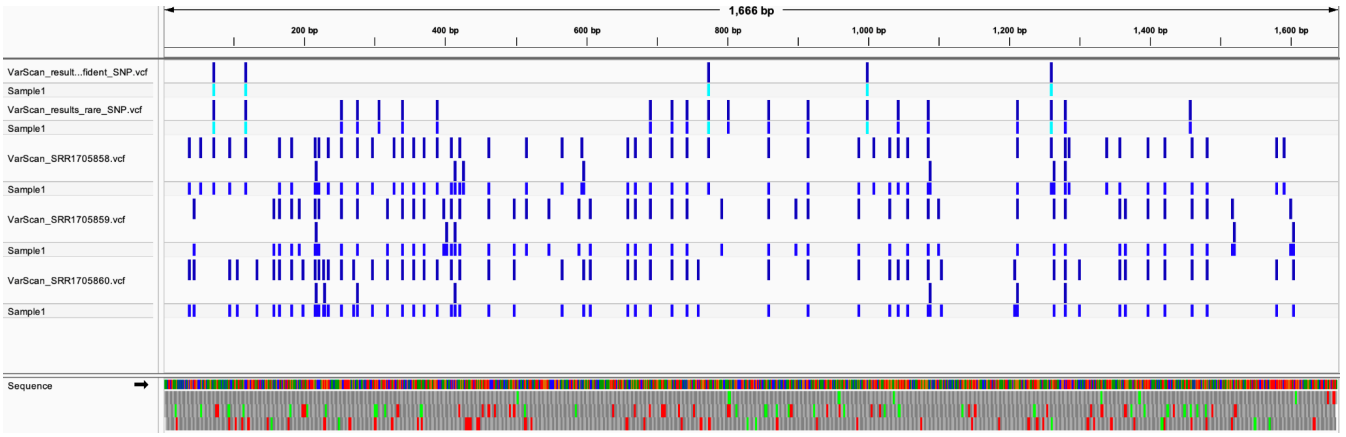
- VMD (Visual Molecular Dynamics): VMD Download
- PyMOL: PyMOL Download
- Jmol: Jmol Download
- RasMol: RasMol Download

R:R Project

## Supplementary

| ID | POS | REF | ALT | FREQ (%) |
|---|---|---|---|---|
| KF848938.1 | 72 | A | G | 99.96 |
| KF848938.1 | 117 | C | T | 99.82 |
| KF848938.1 | 254 | A | G | 0.17 |
| KF848938.1 | 276 | A | G | 0.17 |
| KF848938.1 | 307 | C | T | 0.94 |
| KF848938.1 | 340 | T | C | 0.17 |
| KF848938.1 | 389 | T | C | 0.22 |
| KF848938.1 | 691 | A | G | 0.17 |
| KF848938.1 | 722 | A | G | 0.2 |
| KF848938.1 | 744 | A | G | 0.17 |
| KF848938.1 | 774 | T | C | 99.96 |
| KF848938.1 | 802 | A | G | 0.23 |
| KF848938.1 | 859 | A | G | 0.18 |
| KF848938.1 | 915 | T | C | 0.19 |
| KF848938.1 | 999 | C | T | 99.82 |
| KF848938.1 | 1043 | A | G | 0.18 |
| KF848938.1 | 1086 | A | G | 0.21 |
| KF848938.1 | 1213 | A | G | 0.22 |
| KF848938.1 | 1260 | A | C | 99.96 |
| KF848938.1 | 1280 | T | C | 0.18 |
| KF848938.1 | 1458 | T | C | 0.84 |

Supplemental Table 1. Variants in patient's sequencing.




Supplemental Table 1. Analysis of mutations in IGV.