Deep sequencing as a way to identify rare mutations in isogenic virus populations

Vinogradova Sofiya, Zherko Nikita
Moscow Institute of Physics and Technology

November 15, 2023

Abstract

Vaccination is one of the most sophisticated ways to prevent viral infections such as the flu. However, even vaccinated people could get infected. In this work, we tried to identify the reasons for such an infection. With the help of deep sequencing data, we managed to identify one Single Nucleotide Polymorphism (SNP) located in the D epitope of the Influenza Hemagglutinin protein that we supposed caused the infection of the vaccinated patient. We also propose a way to deal with the problem of distinguishing SNPs and sequencing errors. Overall, our work suggests a protocol to find out the reasons for the vaccinated patients being still infected.

1 Introduction

The *flu vaccine* works by introducing a small, inactive amount of the flu virus into the body. This triggers the immune system to produce antibodies to fight off the virus. The flu vaccine is typically updated each year to protect against the most common strains of the flu virus that are expected to be circulating [1].

Antigenic drift is a gradual change in the genetic makeup of a virus, such as an influenza virus, over time. As a result of these changes, the virus may slightly differ from the previous strains, making it difficult for the immune system to recognize and respond effectively to it [2].

The presence of several strains of a virus in the same organism can lead to the emergence of *quasispecies*, which result from the new mutations occurring during replication and RNA-polymerase operation with errors [3].

Deep sequencing involves sequencing a specific region of the genome at a very high depth, which allows for the detection of rare variants within the population [4].

The aim of the current work was to analyze the deep sequencing data sets of viral samples derived from a person who infected a vaccinated patient to reveal the reasons for that.

2 Materials and Methods

2.1 Raw data

The reference *reference.fasta* genome was downloaded from NCBI [5]. Roommate's sequencing reads *roommate.fastq.gz* was downloaded from [6] and three control was downloaded from [7], [8], [9].

2.2 Manual data inspection

NGS files were inspected using the following commands. Roommate.fastq contains 1433060 reads.

```
$ wc -l roommate.fastq
```

2.3 FastQC data inspection

NGS files were inspected with FastQC version=0.12.1 [10] using the following command.

```
$ fastqc -o . [filename.fastq.gz]
```

2.4 Alignment to genome

Snakemake version=7.32.4 [11] was used for making sorted BAM files. A Snakefile comprised rules to download, unzip, align, and sort BAM files. The resulting files were visualized using IGV tool version=2.15.4 [12]. The same manipulations were carried out with the three control data sets.

2.5 Variants calling

The mpileup files were made using samtools with a flag -d increased to 60000 since the average sequencing depth was about 30000 and the maximum depth was of 50000 reads. See lab notes for more details.

```
$ samtools mpileup -d 60000 -f reference.fasta
$ [filename].sorted.bam > my.mpileup
```

Possible SNPs were scanned for by VarScan version=2.4.6 with N=95.

```
$ varscan mpileup2snp [filename]
$ --min-var-freq 0.95 --variants --output-vcf 1 >
$ [results1].vcf
```

Then files were checked for rare mutations with N=0.1

```
$ varscan mpileup2snp my.mpileup
$ --min-var-freq 0.001 --variants
$ --output-vcf 1 > [results2].vcf
```

2.6 Vcf files parcing

Parsing was performed by a special utility, called bcftools version=1.18 [13].

```
$ bcftools query -f '%POS %REF %ALT [ %FREQ]'
$ [filename].vcf > [frequencies].txt
```

2.7 Statistical processing files with frequencies

The mean and standard deviation were calculated for each of the 3 vcf files. Using Excel built-in functions AVERAGE and STDIV.

3 Results

We worked with four data sets: the first one, so-called *roommate* corresponds to the deep sequencing results of the HA genes from a vaccinated person who got the flu. Three other data sets, *control1, control2, control3* represent the sequencing of isogenic reference samples. These data sets were analyzed using the FastQC tool (Figure 1). The analysis revealed a high average quality of reads so we did not trim them.

The reads were mapped on the reference HA gene, the results are present in Table 1.

Data set	Initial number of reads	Number of mapped reads
roommate	358265	358032
control1	256586	256500
control2	233327	233251
control3	249964	249888

Table 1: Number of reads on different stages

The roommate's data set was examined first. After alignment to the reference HA gene, we used VarScan tool to reveal the mutations present in the data set. The most interesting mutations are shown in the Table 2. The first 5 rows represent mutations with frequency above 99% and appear to be synonymous mutations. The last two rows represent non-synonymous mutations with a frequency much higher than most of the other mutations revealed by VarScan.

Position in gene	DNA base change	Amino acid change	Type of mutation
72	A ->G	Thr24Thr	synonymous
117	C ->T	Ala39Ala	synonymous
774	T ->C	Phe258Phe	synonymous
999	C ->T	Gly333Gly	synonymous
1260	A ->C	Leu420Leu	synonymous
307	A ->G	Pro103Ser	non-synonymous
1458	T ->C	Tyr486Tyr	synonymous

Table 2: SNPs detected in the roommate sample

We assumed that these mutations enabled the virus to infect the roommate. However, further analysis to distinguish these mutations from sequencing errors was needed.

To do so, we performed easy statistical analysis for *control1, control2, control3* data sets and calculated the range of frequencies of mutations revealed by VarScan in these samples. The results are presented in Table 3. The range of frequencies was calculated as mean(frequencies) $\pm 3 \cdot$ SD(frequencies).

Based on these data we revealed the same two mutations (Pro103Ser and Tyr486Tyr) in the roommate data set, being out of frequency range for sequencing errors. Considering PAPER the Pro103Ser corresponds to the D epitope region of HA.

4 Discussion

In this work, we managed to identify the reason for a person getting the flu from his roommate even after being vaccinated. To do so we used data sets from deep sequencing of the HA gene in the roommate's viral sample. From the first look, we managed to identify 21 SNPs in the HA gene. The were several SNPs with frequencies above 99%, however, these mutations appeared to be synonymous, so they could not contribute to the escape from the vaccine. Most mutations had frequencies of about 0.2 %, however, two other mutations, Pro103Ser and Tyr486Tyr, appeared to be much more common with frequencies of more than 0.8 %. We hy-

Data set	Mean frequency	SD of frequencies	Range
control1	0.26	0.07	0.04 - 0.47
control2	0.24	0.05	0.08 - 0.4
control3	0.25	0.08	0 - 0.5

Table 3: Statistics for mutations frequencies

pothesized that these two mutations could contribute to the infection, however, further analysis to confirm this idea was needed.

To cope with the problem of distinguishing between real but relatively rare SNPs and sequencing errors we performed simple statistics and computed the average and the standard deviation of frequencies for three isogenic viral data sets. Based on the average range of frequencies in Table 3 we further considered SNP to be a real mutation only if its frequency is about 0.5 %. Thus, we confirmed our hypothesis that the Pro103Ser and Tyr486Tyr are the mutations we were looking for. Since the Tyr486Tyr is a synonymous mutation, it was discarded from the further analysis. The Pro103Ser mutation, on the contrary, appeared to be non-synonymous and located in the D epitope of the HA gene. Since the epitope of the HA viral surface protein was altered, the virus could easily escape from the immunity and infect even a vaccinated person.

In this work, we used a basic statistical analysis to cope with the problem of distinguishing between real mutations and sequencing errors. However, there are several more sophisticated approaches to solving this problem. One of the common approaches is the use of Unique Molecular Identifiers (UMIs) [14]. In this method, specific sequences are added during the library preparation and before all PCR steps. During data analysis, the number of reads that contain UMIs is used to calculate the real frequencies of SNPs and distinguish them from sequencing errors.

5 References

[1] Rina Fajri Nuwarda, Abdulsalam Abdullah Alharbi, and Veysel Kayser. *An overview of influenza viruses and vaccines.* Sept. 2021. DOI: 10.3390/vaccines9091032.

[2] Jonathan W. Yewdell. *Antigenic drift: Understanding COVID-19.* Dec. 2021. DOI: 10.1016/j.immuni.2021.11.016.

[3] Andrew Martin et al. *What is Quasispecies?* 1992.

[4] D. Goldman and K. Domschke. *Making sense of deep sequencing.* 2014. DOI: 10.1017/S1461145714000789.

[5] URL: https://www.google.com/url?q=https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report%3Dfasta&sa=D&source=docs&ust=1699907505087147&usg=AOvVaw2GEeZzOih5qkZSp80sdKa-.

[6] URL: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz.

[7] URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz.

[8] URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz.

[9] URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz.

[10] *FastQC*. June 2015. URL: https://qubeshub.org/resources/fastqc.

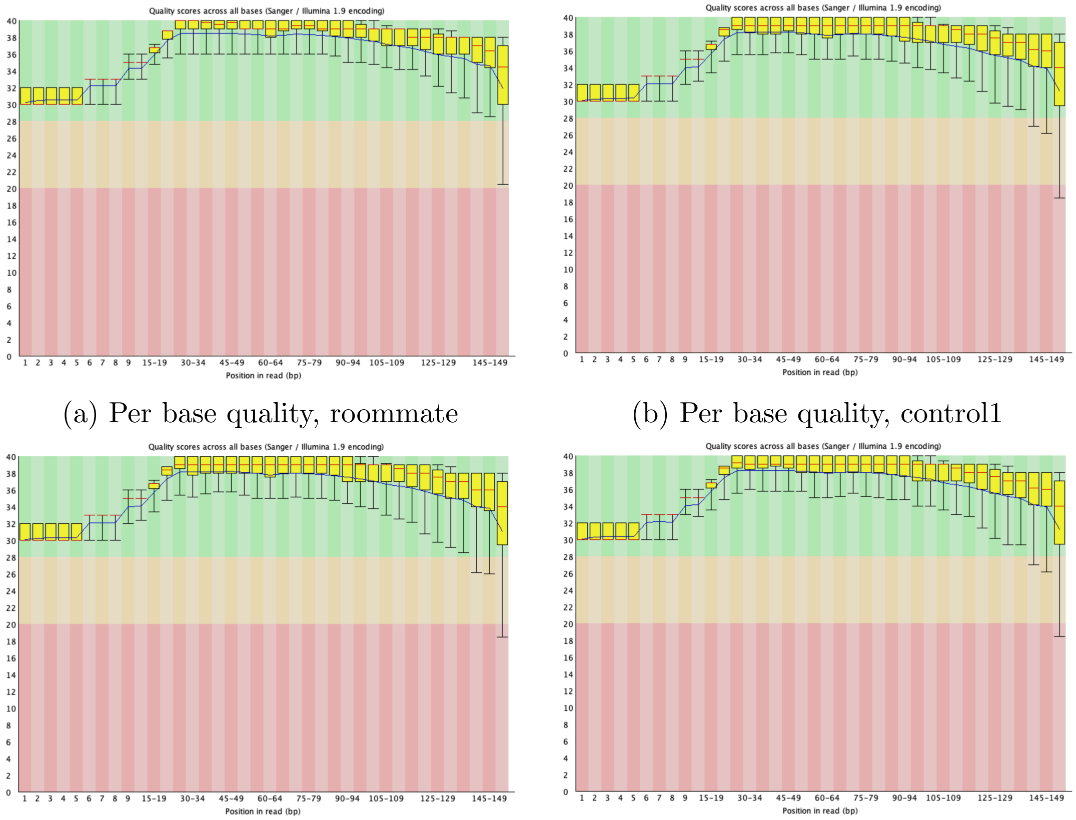
[11] Felix Mölder et al. "Sustainable data analysis with Snakemake". In: *F1000Research* 10 (Jan. 2021), p. 33. DOI: 10.12688/f1000research.29032.1.

[12] James T Robinson et al. *Integrative genomics viewer*. 2011. DOI: 10.1038/nbt0111-24. URL: http://maq.sourceforge.net/.

[13] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27 (21 Nov. 2011), pp. 2987–2993. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr509.

[14] Scott R. Kennedy et al. "Detecting ultralow-frequency mutations by Duplex Sequencing". In: *Nature Protocols* 9 (11 Nov. 2014), pp. 2586–2606. ISSN: 17502799. DOI: 10.1038/nprot.2014.170.

6 Supplementary materials



(a) Per base quality, roommate

(b) Per base quality, control1

(c) Per base quality, control2

(d) Per base quality, control3

Figure 1: FastQC results for four initial data sets