Hemagglutinin mutation harms vaccine efficiency against H3N2 flu virus strain

V. Zvezdin, D. Kirpa

November 2022

1 Abstract

In this report we explored the reasons behind the case of developing flu symptoms in a patient protected with vaccine from the strain that caused them (H3N2). The viral samples were taken from their non-vaccinated cohabitant. Using deep sequencing approach, we identified a rare mutation in the viral HA gene leading to an aminoacid change Pro103Ser, in the locus that resides in the epitope D of the protein the vaccine targets. We theorise that the viral variants bearing this mutation are not recognised properly by vaccine antibodies. Thus, they might be dangerous even for the people vaccinated against this strain.

2 Introduction

The flu virus belongs to a large group of RNA viruses, utilizing RNA-dependent RNA polymerase to copy its genetic material. This enzyme is prone to errors, introducing new mutations with each duplication cycle. The flu virus also lacks the proofreading proteins that would allow to fix those errors. [1] This, in turn, results in viruses ability to mutate very fast. In a single host, it creates the broad range of "quasispecies" that bear their own particular mutations and are able to recombine with each other, distributing those mutations across the virome. Being the result of random errors, most mutation are shared only by a small part of viral populations, making them hard to detect. However, they might occur in any part of the genome, including the antigen loci, which can drastically change the strength of antigen-antibody interaction. This constant reconstruction of antigens through mutations is called antigenic drift. Due to it, some viral particles might obtain mutations protecting them from vaccines [2]. It gives them selective advantage and allows to quickly supplant other variants. With flu virus in particular, each year it typically results in new variants arising that are never more respondent to the vaccines in use [3].

With the development of NGS methods, it became possible to find such mutations in viral genetic material. For small antigenic sequences, current NGS techniques easily allow coverage of dozens of thousands of reads per position, the so-called "deep sequencing", detecting most mutations present. However, such high-throughput analyses always bring errors of their own, which creates the problem of properly distinguishing between real mutations and various technical errors in the data. The latter ones might occur during sample RNA (DNA) amplification, when new copies with mismatches or indels are created, or in the sequencing itself during nucleotide attachment, signal detection or basecalling. Therefore, analysing deep sequencing data requires using the control data to refer to, such as data from more samples, or repeated runs on the same or other platforms. In this study, we used isogenic viral samples to determine the level of noise in our experiments and identify variations whose frequency significantly exceeds it.

3 Methods

The segment 4 of the Hemagglutinin gene of Influenza A virus H3N2 strain was used for the analysis. (The reference id in GenBank: KF848938.1). Illumina single-end sequencing runs were performed on samples isolated from the patient and reads were recovered SRA FTP archive:
http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/.

Reference sequence was indexed and the reads were aligned to it with bwa utility [4]. Alignment was then converted to .bam, sorted and indexed with Samtools. Number of total and mapped reads was assessed by Samtools flagstat command on alignment .bam file [5]. We extracted the average length of reads in alignment using this command:

samtools stats "alignment file name".bam | grep "average length"

Given also the reference length of 1665 nucleotides, average coverage was calculated (number of reads * average length / length of reference). We made several runs of mpileup Samtools command, starting with depth

equal that number and increasing it to see when the file size would not increase anymore, including as much reads in the analysis as possible. In the end, we chose to use the depth = 50000.

samtools mpileup -f "reference name".fasta "alignment name".bam -d 50000 > "mpileup name".mpileup

VarScan tool was used for variant calling on the .mpileup file with output format in .vcf. First, the variants occuring more often than in 95% cases were extracted (–min-var-freq 0.95 –variants) [6].

Another run for mutations occuring at least in 0.1% reads (–min-var-freq 0.001) was performed as well. Three isogenic Influenza samples subcloned from the plasmid were used as biological controls for the analysis. They were as well sequenced on Illumina, and the reads were recovered from SRA archives:
SRR1705858: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz
SRR1705859: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz
SRR1705860: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

The reads were put through the same pipeline. Since there were less reads for these samples, the depth we chose for mpileup remained unchanged. For each sample the mean and sd of frequencies of procured variations were calculated. In the experimental sample, the variations were considered real if their frequency was higher than mean + 3sd of all the control sample. All the common and statistically significant rare mutations were assessed in the IGV browser to see if they bring changes to aminoacid sequence.

4 Results

Out of 361349 reads, 361116 were mapped to the reference (99,97%) in the patient samples. In the control samples, there were 256586, 233327, and 249964 reads respectively, and the percentage of mapped reads was the same. The average read length equaled 148 nucleotides per read in sample and 147 in controls. The coverage in sample thus was around 32k reads per position, with around 22k in all the controls. 5 commonly occurring mutations (frequency greater than 99%) were found in the first run on sample. All of them were identified to be synonymous in the IGV browser, however. The second run reported back 23 variations, 2 of which were indels and 21 were mismatches. All 16 new mismatch variations were rare (frequency less than 1%). The calculations for the controls are presented in the Table 1.

Table 1. Statistical values for variations in control samples

	№ of variations	Mean	Sd
Control 1	57	0,26	0,07
Control 2	52	0,24	0,05
Control 3	61	0,25	0,08

Two of the variations in the patient sample alignment were deemed statistically significant:
reference pos ref alt freq
KF848938.1, 307, C, T, 0.94%
KF848938.1, 1458, T, C, 0.84%

The first of these mutations was also found to be non-synonymous. It results in aminoacid change Pro103Ser. This aminoacid resides within the epitope D of Influenza virus, as described in the following article [7]

5 Discussion

In this research, we were able to find several mutations occuring in the viral samples isolated from the patient who hasn't undergone vaccination against the flu. One of these rare variants is non-synonymous, changing the aminoacid in the epitope D of the Influenza hemagglutinin protein. This part of the protein is targeted by antibodies in vaccines, and the mutation, Pro103Ser notably alters its properties, introducing the polar aminoacid in place of the non-polar one. Such a mutation might weaken the antigen-antibody interaction. [2],[8]. In the absence of other notable mutations we can conclude that this one is most likely responsible for virus being able to avoid the vaccine designed against it. Even if only a small fraction of viral particles contains this mutation, it might still be enough for them to infect the vaccinated person and inflict the disease. [9]

Even the variants significantly different from the noise, such as the one we've found, can still be brought in by PCR errors - the earlier in amplification they happen, the higher the percentage in the final reads would be. To prove the obtained result is indeed the acting mutation, it is possible to compare the interaction with antibodies between the wild type strain and the artificially modified strain bearing just this mutation. It can be done *in vitro* in viral titration assays as well as in cell cultures, for example, in HA inhibition assay. Also, these type of errors may be eliminated with additional runs, especially with combination of 2nd and 3rd generation platforms, as the latter ones are able to produce long reads and do not heavily rely on sample amplification.

We provide the Google colab link with the Bash and Python code used in this study:
https://colab.research.google.com/drive/1mAuiasb2XB4MYwxmwfDAecLjR2wFTcBU?usp=sharing

References

[1] Laura D Manzanares-Meza and Oscar Medina-Contreras. "SARS-CoV-2 and influenza: a comparative overview and treatment implications". In: *Boletın médico del Hospital Infantil de México* 77.5 (2020), pp. 262–273.

[2] David A Steinhauer and JJ Holland. "Rapid evolution of RNA viruses". In: *Annual Reviews in Microbiology* 41.1 (1987), pp. 409–431.

[3] Gabriele Neumann, Takeshi Noda, and Yoshihiro Kawaoka. "Emergence and pandemic potential of swine-origin H1N1 influenza virus". In: *Nature* 459.7249 (2009), pp. 931–939.

[4] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[5] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2 (Feb. 2021). giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

[6] Daniel Koboldt et al. "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome research* 22 (Mar. 2012), pp. 568–76. DOI: 10.1101/gr.129684.111.

[7] Enrique T Munoz and Michael W Deem. "Epitope analysis for influenza vaccine design". In: *Vaccine* 23.9 (2005), pp. 1144–1148.

[8] John J Skehel and Don C Wiley. "Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin". In: *Annual review of biochemistry* 69 (2000), p. 531.

[9] Jonathan M Gershoni et al. "Epitope mapping". In: *BioDrugs* 21.3 (2007), pp. 145–156.