Why did I get the flu

Ginzburg G., Kupaeva D.

Abstract

Seasonal flu H3N2 is a seasonal flu virus, causing significant sickness rate and morbidity all over the world. Although seasonal vaccine is developed to contain H3N2 strains components, due to short evolution time of it's antigens and the emergence of it's quasistrains it is possible even for vaccinated patients to get sick. In our work we analyze reasons for ineffectiveness of the seasonal vaccine against current flu strain. Also we examine some ways of detecting sequencing errors and preventing them from influencing our results.

Introduction

Influenza is a respiratory illness caused by viruses belonging to the Orthomyxoviridae family. This family consists of four genera of influenza viruses (A, B, C, and D), which are classified based on differences in their internal glycoproteins, nucleoproteins, and matrix [1]. Influenza is a single-stranded RNA virus that encodes 13 genes, including neuraminidase and hemagglutinin proteins, which are expressed on the virus itself and the surface of infected cells[2] . Infection can be prevented by vaccination, which induces neutralizing antibodies that usually target the glycoprotein hemagglutinin (HA). HA is incorporated into the viral membrane as a homotrimer, in which each monomer consists of two disulfide-linked polypeptides, HA1 and HA2.

HA1 forms a globular head domain containing a receptor-binding site that targets sialic acid residues on host cells, while HA2 contains a transmembrane anchor domain and a fusion domain. The vast majority of neutralizing antibodies induced by infection or vaccination are directed against the open and highly variable loops surrounding the receptor-binding site in HA1 and prevent attaching the virus to cell receptors [3] .

Continuous genetic variation in influenza viruses, due to the low fidelity of viral polymerase, can transform into antigenic variation as mutations in HA and/or NA, a phenomenon also known as antigenic drift. Due to the positive selection of such variations in influenza populations, several quasispecies of the same strain circulates simultaneously, which makes neutralization of the virus strains particularly difficult. To assess the diversity of quasispecies of the virus, methods of targeted deep sequencing are currently used, in which DNA is enriched by a specific fragment of the viral genome [4].

In this paper, we focused on the analysis of mutations present in quasispecies of a patient who infected a person with an actual seasonal flu vaccine. A distinct difficulty in data analysis was the detection of sequencing errors, the frequency of which is close to the frequency of rare viral quasi-populations; however, statistical methods make it possible to distinguish real mutations with sufficient accuracy[5] .

Methods

Our work was focused on mutations of the A/USA/RVD1_H3/2011 strain of the H3N2 virus. For our data we had preprocessed Illumina single-end reads with 151 rounds of sequencing: our roommate's sequencing data (SRR1705851), and the control sample sequencing data(SRR1705858, SRR1705859, SRR1705860). We used **fastqc**[6] for reads quality examination. After it we used **bwa** [7] (v.0.7.17-r1188) package **(**standard option of *mem* usage**), samtools** [8](v.1.7) for view (*view* option), sort (*sort* option), inmapped reads counting (*view -f4 flag*), coverage estimate (*depth -d 50000*) snp detection (*mpileup -d max_coverage*). First we used **VarScan**[9] (v.2.4.0) **--variants** with --min-var-freq parameter set to 0.95 to identify possible mutations in the patient's data and also we ran it with --min-var-freq parameter set to 0.001 to identify rare variants.**IGV Browser**[10] (v.2.8.10) was used to visualize alingment results. All commands, which are used in this study, are listed in supplementary materials 1. All substitution are listed in supplementary materials 2.

3d models of protein were constructed by Swiss-Modell[11] with standard settings. For control sequence modeller predict that our protein has 98.19% of identity with 4we8.1.A Hemagglutinin from influenza virus A/Victoria/361/2011, and QMEAN consist of -0.27, and it is a high rate for predicted models. For a sequence with SNP found equal similar models and QMEAN consist -0.31, which are high rate too. Models and swiss-modeller's reports attached in supplementary materials 3. The model was visualized in the pymol v.2.4.1 software.

Results

For getting SNP data we prepare row sequences data using methods from the previous chapter. Statistical data about processing our data is presented in table 1.

	Count of row sequences	Count of mapped sequences	Median frequency of SNP, %	Standard deviation of SNP, %	Count of SNP, frequency > 0.1%
Patient data	358265	358032	24.0	43.5	21
Control 1	256586	256500	0.256	0.0717	57
Control 2	233327	233251	0.237	0.0524	52
Control 3	249964	249888	0.250	0.0780	61

Table 1. Basic statistics of processing of reads.

Analysis of common mutation from patient samples showed that more than 95% of the viral population have 5 synonymous substitutions. No one from this common substitution changes amino acids, and,

consequently, they can not protect the virus from antibodies. Information about position and changes in codons provided in table 2.

Position	Reference nucleotide	Nucleotide after change	Codon change	Amino Acid change	Frequency
72	A	G	aca - acg	T - T	99,96
117	C	T	gcc - gct	A - A	99,82
774	T	C	ttt - ttc	F - F	99,97
999	C	T	ggc - ggt	G - G	99,86
1260	A	C	cta - ctc	L - L	99,94

Table 2. Common mutation in gene hemagglutination inhibition (HI)

Search of rare substitution (frequency 0.1% < N < 95%) occurs 16 SNP, including non-synonymous. For excluding SNP which are the result of sequence error, we search for rare SNP in control sequences, the genotype of each veraciously matches with reference. Count of founded SNP provided in table 1. For these SNPs in control samples, we calculated mean and standard deviation. Their value occurs 0.2482941 mean and 0.06898269 standard deviations. We excluded all values which frequency occurs smaller than control mean and 3 standard deviations. This value consists of 0.4552422%. This manipulation detected 2 rare SNPs. Information about this SNPs is demonstrated in table 3.

Position	Reference nucleotide	Nucleotide after change	Codon change	Amino Acid	Frequency
307	C	T	ccg - tcg	**P - S**	0,95
1458	T	C	tat - tac	Y - Y	0,84

Table 3. Rare SNPs in patient’s sequences.

One of the 2 presented mutations (at position 1458) is synonymous and cannot be the cause of the change in the antigen of the influenza quasi-population. Another mutation, located at position 103, results in a nonsynonymous substitution of proline for serine. Visualisation of location of this SNP and spatial changes presented in figure 1.

This mutation is of interest for 2 reasons: First, these amino acids have different chemical properties. Proline is a non-polar amino acid, while serine is polar. Second, position 307 refers to epitope D of the virus antigens [12].

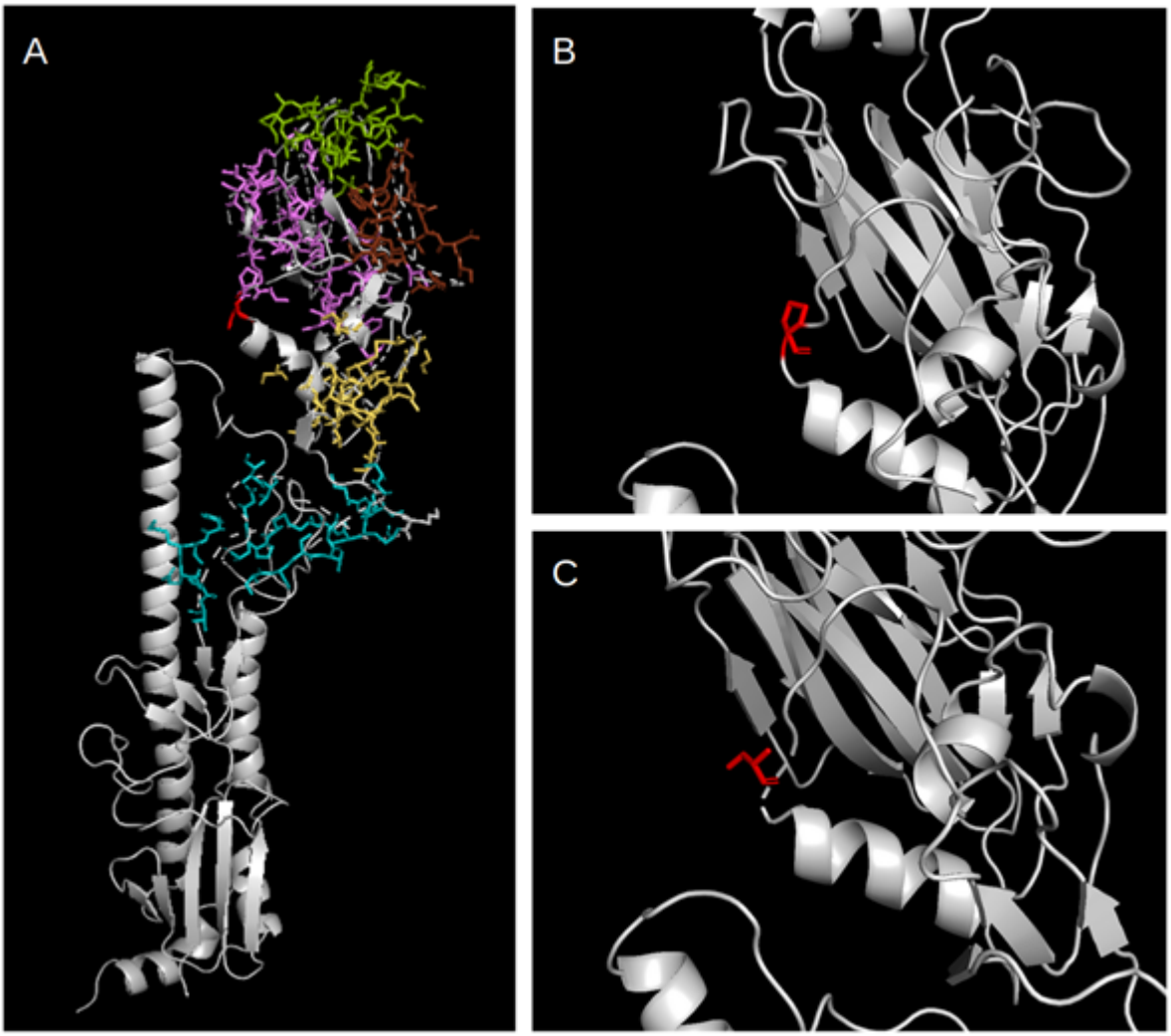

Figure 1. Visualisation of SNP P103S in HA A/Hong Kong/4801/2014 (H3N2) . A - visualisation of 1 subunits of hemagglutinin (brown - epitope A, green - epitope B, blue - epitope C, pink - epitope D, yellow - epitope E, red - P103), B - 103P (like in reference), C - 103S (SNP from quasispecies influenza from patient)

Discussion

So, one of the virus subpopulations of this patient was able to infect a person with the current seasonal vaccination. Of all the SNPs we found, only one - changing the 103 amino acids - could affect the ability of the virus of this strain to ignore antibodies. P103S changes the aliphatic amino acid to the polar one, thereby significantly changing the biochemical properties of the antigen. P103C is part of the D epitope, and, probably, its replacement leads to the impossibility of binding the antigen by the antibody. Indeed, the 3D model of the protein (Fig.1B, C) shows that these amino acids have significantly different spatial characteristics. Moreover, early this SNP was identified in quasispecies in a sick person [13].

In our research, we encountered methodological problems that require complex solutions to prevent incorrect results. When analyzing rare mutations, we obtained a large number of SNPs both in the

experimental and in the control sample, despite the fact that it had to exactly match the reference. We filtered the found SNPs by frequency of occurrence, setting the significance threshold to the mean SNP frequency in the control + 3 standard deviations.

Our method allows you to quickly and inexpensively filter out most of the sequencing errors; however, it does not allow detecting ultra-rare SNPs (frequency <0.45), and also does not guarantee protection against rare outliers.

In order to distinguish between a sample preparation error, a sequencing error, and real SNPs, we need to carefully study the conditions in which the error occurs. In our case, a careful study of the errors showed that the distribution of errors in 4 repetitions was very unlikely to happen if it was to happen randomly. Thus, in 11 positions, a rare substitution is found in all 4 replicates, and another 22 substitutions are present in 3 out of 3 samples (count of errors per position presented in the table from supplementary materials 4).

Such a distribution of substitutions leads to the idea that these substitutions are site-specific and, therefore, may be associated with sample preparation: conditions of reverse transcription, amplification, and fragmentation. It has also been shown that when using DNA as a template or a small number of amplification cycles, it reduces the frequency of errors[5] . A detailed analysis of the occurrence of errors in the process of sample preparation comes to the conclusion that most of the errors are a consequence of the shortening of the sequences, however, the elimination of the shortened sequences does not affect the frequency of errors in the "hot spots" that seem to arise during the amplification process [14]. To prevent such errors, it is necessary to use not only biological replicates (such as 3 control sequencing samples), but also technical ones: using different enzymes in the process of sample preparation and with a different number of amplification cycles. Also, it may be suggested to use error-correcting software (for example, **Fiona**[15] software, as it shows one of the best results among read-error-correcting tools[16]).The authors strongly advise to exclude truncated samples from analysis as they are a significant source of error [14]. It should also be noted that sequence alignment may also present a source of error as sequence alingment packages may not take quality of reads into account (**bwa**, for example)[17].

Citations

1. Vemula, S. V. *et al.* Current approaches for diagnosis of influenza virus infections in humans. *Viruses* **8**, (2016).

2. Rossman, J. S. & Lamb, R. A. Influenza virus assembly and budding. *Virology* **411**, 229–236 (2011).

3. Beer, K. *et al.* Characterization of neutralizing epitopes in antigenic site B of recently circulating influenza A(H3N2) viruses. *J. Gen. Virol.* **99**, 1001–1011 (2018).

4. Barbezange, C. *et al.* Seasonal Genetic Drift of Human Influenza A Virus Quasispecies Revealed by Deep Sequencing. *Front. Microbiol.* **9**, (2018).

5. King, D. J. *et al.* A systematic evaluation of high-throughput sequencing approaches to identify low-frequency single nucleotide variants in viral populations. *Viruses* **12**, (2020).

6. Andrews, S. FastQC A Quality Control Tool for High Throughput Sequence Data. *Babraham Bioinformatics* (2010). Available at: www.bioinformatics.babraham.ac.uk/projects/fastqc/.

7. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

8. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

9. Koboldt, D. C. *et al.* VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

10. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

11. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

12. Muñoz, E. T. & Deem, M. W. Epitope analysis for influenza vaccine design. *Vaccine* **23**, 1144–1148 (2005).

13. Martin, P. On the origin of the Hirudinea and the demise of the Oligochaeta. *Proc. R. Soc. B Biol. Sci.* **268**, 1089–1098 (2001).

14. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 1–14 (2018).

15. Schulz, M. H. *et al.* Fiona: A parallel and automatic strategy for read error correction. in *Bioinformatics* **30**, i356 (Oxford University Press, 2014).

16. Mitchell, K. *et al.* Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* **21**, 71 (2020).

17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Supplemental resources

Materials 1. Lab Notebook
https://docs.google.com/document/d/1ChheguT3wOg-2WiviWsop9AvC3Q5d-6iXfE8zBrbRNc/edit#heading=h.rojdcj7h7mfq

Material 2. Table with all found substitutions:
https://docs.google.com/spreadsheets/d/1hYM5J_LjFZaCSB7Yk9qsxFN8pfnu3390QOSYXvZyQXM/edit?usp=sharing

Material 3. Built models: reference HA, HA with SNP S103S, and SNP reports.
https://drive.google.com/drive/folders/1ih574VOCnSy9jIQL1Mr97NEXyCmIlges?usp=sharing

Material 4. Count of substitution by nucleotide position in 4 replies:
https://docs.google.com/spreadsheets/d/178OFe9AaPjp60j3rejBdpKfYPibJrjrX-1GvPRSI4I8/edit?usp=sharing