BIOINFORMATICS INSTITUTE

Viral quasispecies occurrences cause vaccines ineffectiveness

Stanin Vladislav[1, 2 *] and Nikita Vaulin[1, 3, * †]

[1]Bioinformatics Institute, Saint-Petersburg, Russia
[2]Peter the Great St.Petersburg Polytechnic University, Saint-Petersburg, Russia
[3]Skolkovo Institute of Science and Technology, Moscow, Russia

*Contributed equally.
†vaulin@ro.ru

Abstract

Seasonal vaccinations are important to prevent the spread of viruses. However, vaccines can protect against only several the most relevant viral strains. Therefore, occurrence of new viral quasispecies can make vaccines less effective. This report presents the results of the analysis of mutations in hemagglutinin of Influenza A/Hong Kong/4801/2014 (H3N2) strain quasispecies obtained from one person. The analysis detected one valid mutation led to the P103S amino acid substitution in the epitope D of glycoprotein. Therefore, we consider this missense mutation caused antigenic drift and, as a result, less efficiency of vaccine.

Key words: NGS, viral evolution, influenza, H3N2

Introduction

Influenza viruses are members of the family *Orthomyxoviridae*. This family represents enveloped viruses the genome of which consists of segmented negative-sense single-stranded RNA segments. A clinical characteristic of human influenza is a sudden rise in body temperature to >38.5 °C, headache, limb ache, tiredness, general faintness and dry cough [1]. The most serious outcomes are primary influenza pneumonia, encephalitis or myocarditis [1].

There are three types of influenza virus (A, B and C). Annual seasonal influenza epidemics and even pandemics are usually caused by subtypes of influenza A virus, including the H3N2 strains studied in this work. While infection immune system recognise the virus by surface glycoproteins that enable attachment to cells, in particular - hemagglutinin (HA). Therefore subdivision into different subtypes based on antigenic differences in them [2].

Influenza infection can be prevented by vaccination. Vaccine allows to increase the level of affine to virus glycoproteins antibodies by showing the pathogen-like particles or avirulent virus to human immune system. Increased levels of antibodies prevent severe form of disease and reduce risk of a pandemic. Two main types of influenza vaccine are widely available: inactivated influenza vaccines (IIV) which include inactivated viruses or their particles and live attenuated influenza vaccines (LAIV) which are made from attenuated, or weakened, viruses [2, 3].

Vaccination have seasonal character because of frequent mutations in viral genome. The most important mutations are related to receptor-binding glycoproteins sites (epitopes). Such mutations are responsible for the antigenic variation that generates drift variant of virus strains. Viral mutations occur so frequently that new quasispecies with rare mutations in epitope regions within the strain appear.

Modern vaccines cover several the most relevant and common strains. In 2022 - 2023 northern hemisphere influenza season WHO recommended usage the following influenza vaccine components in the forms of trivalent or quadrivalent vaccines [4]:

- A/Victoria/2570/2019 (H1N1)pdm09-like virus
- A/Darwin/9/2021 (H3N2)-like virus
- B/Austria/1359417/2021 (B/Victoria lineage)-like virus
- B/Phuket/3073/2013 (B/Yamagata lineage)-like virus

However, new mutations in quasispecies can also cause antigenic drift and even displace the parential strain [5]. Detection of rare mutations in epitope regions of such quasispecies usually can be done by targeted deep sequencing. This kind of sequencing feature a depth of several hundred to several thousand reads at any given position and utilized for calling variants with extremely low population frequencies (under 1%). The problem is next-generation sequencing errors that may be mistaken for rare mutations. There are several sources of errors: inaccuracy of DNA polymerases and sub-optimal handling/storage conditions during the study [6].

In this study we investigate mutations of quasispecies of Influenza A/Hong Kong/4801/2014 (H3N2) strain obtained from the person that could cause antigenic drift. We used the results of deep sequencing of segment 4 hemagglutinin gene to obtain rare variants. Herewith, the problem of sequencing errors was solved by

Compiled on: November 11, 2022.
Draft manuscript prepared by the author.

comparison with variant calling results of trice sequenced isogenic viral sample derived from a virus clone.

Figure 1. The crystal structure of hemagglutinin of A/Port Chalmers/1/1973 influenza virus and location of amino acid substitution (red); PDB: 4WE5 [7]

Materials and methods

Dataset

In this work we used data from Illumina 1.9 deep sequencing of a virus sample from one person (personal data are hidden in order to preserve medical and military secrecy)

NGS data processing

Raw reads were analysed with FastQC v.0.11.9 [8]. All FastQC reports are presented in project repository at GitHub (see Supplemental materials).

Reads alignment to the reference genome was conducted with BWA MEM tools v. 0.7.17-r1188 [9]. Reference genome index was made with *bwa index*.

For alignment data processing *samtools* v. 1.16.1 package was used [10]. Alignment data was sorted with *samtools sort* and converted to lightweight binary format with *samtools view -Sb*. Alignment statistics was collected with *samtools flagstat*.

Variant calling

In order to perform variant calling alignment index was built with *samtools index*. Variant calling was done using *samtools mpileup* with a depth limit specified with *-d* option. Depth limit was calculated for each sample based on number of reads, average read length (151 bp) and reference length (1690 bp):

$$Depth\ Limit = \frac{N_{reads} \cdot L_{reads}}{L_{reference}}$$

This number was confirmed by calculating it with usage of bedtools [11]. We obtained depth of sequencing using *bedtools bedtobed* and *bedtools genomecov* for each position of genome and calculated the average.

Single nucleotide polymorphisms from mpileup file were obtained using *mpileup2snp* function from VarScan tool v. 2.4.4 with the following parameters: "*–min-var-freq 0.001 –variants –output-vcf 1*" (minimum SNP frequency 1%) [12].

The results of variat calling were reviewed with *awk* tool and visualised with IGV v. 2.14.0 [13].

For rare variants calling the the background noise level was computed as $mean + 3 \cdot sd$ and confidence intervals of mutation frequency of 3 technical replicates using R v.4.2.1 in RStudio IDE v. 2022.07.2+576 [14, 15].

Visualisation of substitution

To visualize location of amino acid substitution in hemagglutinin we utilized PyMOL molecular graphics system v. 2.5.0 [16]. We used the crystal structure of hemagglutinin of A/Port Chalmers/1/1973 influenza virus obtained from Protein data Bank . (PDB: 4WE5, Figure 1)[7, 17]

Results

Reads alignment

The initially received data contained 358265 reads most of which was 150–152–nucleotides long (see Supplemental materials for GitHub repository with FastQC_reports folder). Most of the reads had average per read phred 33 quality over 36 and no adapter sequences was found by FastQC so the raw reads were directly used for analysis.

Reads was aligned using BWA-MEM method from the BWA package. 358032 reads were mapped in total (99.9%).

The real coverage calculated with *bedtools* gave two of results differed by less then 3 *standard deviations* (31212.8 ± 1908).

Variant calling

SNP detection was carried out with VarScan as described in Materials and methods. We performed analysis for only those nucleotide mutations with 95% or more frequency considered as SNPs and this resulted only synonymous amino acid substitutions (Table 1).

Position	Mutation	AA substitution	Effect
72	A/G	T24T	Synonymous
117	C/T	A39A	Synonymous
774	T/C	F258F	Synonymous
999	C/T	G333G	Synonymous
1260	A/C	L420L	Synonymous

Table 1. Single nucleotide polymorphisms found in sequencing data with *–min-var-freq* parameter equal to 0.95

Hence, we performed rare variants calling with 1% frequency threshold.

In order to distinguish rare variants from the PCR amplification and sequencing errors we conducted the same analysis for control isogenic viral samples. The control samples sequencing was performed in triplicate, the mean values and confidence intevals for control samples are shown in the Table 2.

After this procedure, we received only 2 rare mutations, the frequency of which was statistically significantly different from the background noise. Detected SNPs are presented in the Table 3. Since one of the rare mutations found also results in a synonymous substitution, we have one candidate for enhanced avoidance

of vaccine immunity.

Sample	Mean	Interval
1	0.26%	[0.04, 0.47]%
2	0.24%	[0.08, 0.39]%
3	0.25%	[0.02, 0.49]%

Table 2. Mean values and confidence intervals for noise level in control isogenic viral samples (in percents)

Position	Mutation	AA substitution	Effect	Frequency
307	C/T	P103S	Missense	0,97%
1458	T/C	Y486Y	Synonymous	0,82%

Table 3. Rare single nucleotide polymorphisms found in sequencing data with *–min-var-freq* parameter equal to 0.001

Discussion

Our results show that the person was infected with a specific strain of the influenza virus H3N2 that carries a C/T mutation at position 307 that results in missense amino acid substitution P103S. According to Muñoz and Deem [18] residue P103 is located within the hemagglutinin epitope D of influenza A H3N2 strain. Proline substitution with serine leads to a significant disruption of the local microenvironment in the epitope, since the hydrophobic amino acid residue is replaced by a polar one. In this epitope, proline performs its popular function being at the base of the α-helix secondary structure, so its replacement can greatly affect the entire local conformation We assume obtained mutation caused antigenic drift and, therefore, antibodies formed after vaccination worse recognized modified site of hemagglutinin. Thus it caused a clinical picture of disease. However, we also found that the influenza strain the person had was not included in this season's standard vaccination kit, so the flu transmission to others may be even more likely.

References

1. Seitz R, Heiden M, Offergeld R, Burger R. Influenza Virus. Transfusion Medicine and Hemotherapy 2009;36(1):32–39. https://www.karger.com/Article/FullText/197314, publisher: Karger Publishers.
2. Sridhar S, Brokstad KA, Cox RJ. Influenza Vaccination Strategies: Comparing Inactivated and Live Attenuated Influenza Vaccines. Vaccines 2015;3(2):373–389. https://www.mdpi.com/2076-393X/3/2/373.
3. Types of seasonal influenza vaccine;. https://www.who.int/europe/news-room/fact-sheets/item/types-of-seasonal-influenza-vaccine.
4. Recommended composition of influenza virus vaccines for use in the 2022-2023 northern hemisphere influenza season;. [Accessed 08-Nov-2022]. WHO recommendations.
5. Boni MF. Vaccination and antigenic drift in influenza. Vaccine 2008 Jul;26:C8–C14. https://www.sciencedirect.com/science/article/pii/S0264410X08004489.
6. Domingo E, Perales C. Viral quasispecies. PLOS Genetics 2019;15(10):e1008271. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008271, publisher: Public Library of Science.
7. Yang H, Carney PJ, Chang JC, Guo Z, Villanueva JM, Stevens J. Structure and receptor binding preferences of recombinant human A(H3N2) virus hemagglutinins. Virology 2015;477:18–31. https://www.sciencedirect.com/science/article/pii/S004268221400573X.
8. FastQC tool;. Accessed: 2022-10-27. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
9. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754 – 1760.
10. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021 02;10(2). https://doi.org/10.1093/gigascience/giab008, giab008.
11. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 Jan;26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
12. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 2012;22 3:568–76.
13. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology 2011 Jan;29(1):24–26. https://doi.org/10.1038/nbt.1754.
14. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria;, https://www.R-project.org.
15. RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA; 2020, http://www.rstudio.com/.
16. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8; 2015.
17. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research 2018 10;47(D1):D520–D528. https://doi.org/10.1093/nar/gky949.
18. Muñoz ET, Deem MW. Epitope analysis for influenza vaccine design. Vaccine 2005;23(9):1144–1148. https://www.sciencedirect.com/science/article/pii/S0264410X0400636X.

Supplemental materials

GitHub repository of the project: Flu_project_BI_2022