The identification of rare vaccine-resistant Influenza A H3N2 virus variants by deep sequencing

Michael Greenberbg, Polina Vaganova
Bioinformatics Institute, Saint Petersburg, Russia

Abstract. Influenza A viruses represent a significant global health threat. The surface hemagglutinin (HA) virus protein is the primary target for current vaccines. However, the virus's ability to undergo antigenic changes, such as antigenic drift and shift, challenges vaccine efficacy. This study employs deep sequencing to identify both common and rare variants of an influenza A/H3N2 strain not covered by the vaccine. The analysis reveals rare missense mutation, which may affect antibody affinity to HA, potentially impacting vaccine effectiveness. The study highlights the importance of deep sequencing in discerning rare variants and suggests strategies for improving variant identification accuracy.

Keywords: Influenza A virus, A/H3N2, Hemagglutination Inhibition test (HI), Influenza hemagglutinin (HA), vaccine resistance, deep sequencing, angtigenic drift.

1 Introduction

Influenza viruses cause seasonal epidemics as well as pandemics and are a significant concern for human health. Current influenza virus vaccines primary target the surface hemagglutinin (HA) protein of the virus. Vaccines show efficacy when they are antigenically well matched to circulating strains. However, the efficacy of vaccines can be diminished as the virus undergoes antigenic changes.

Influenza virus has two major mechanisms for antigenic change, antigenic drift, and antigenic shift. Antigenic drift is the process by which minor changes are introduced into key viral epitopes through point mutations in the viral genome.[1] Antigenic shift occurs by genetic reassortment between two different influenza viruses which infect the same host cell and can lead to the emergence of a novel influenza strain.[2] Rather than the gradual changes seen with antigenic drift, antigenic shift results in a complete exchange of HA genes between viruses.[3]

The effectiveness of these evasion mechanisms is further increased because influenza virus has quasispecies nature. This refers to the fact that RNA viruses exist as heterogeneous populations of closely related genetic variants which independently mutate and actively shifting their antigens.[4,5] Because of that the Hemagglutination Inhibition (HI) test which is often used to estimate vaccine efficacy[6] can give false positive results in case of a new variant closely related to the strains used for antibody production.

The identification of virus variants through sequencing methods may not always prove effective, as infrequent variants can be mistaken for sequencing and PCR errors. However deep-sequencing technologies address this issue by significantly increasing the depth of coverage,[7] thereby enabling the study of intrinsic heterogeneity within the virus genome and identification of emerging virus variants.[8]

In this study we focusing on identification both common and rare influenza variants from patient X which are not covered by vaccine protection against A/Hong Kong/4801/2014 (H3N2) strain using deep sequencing technology.

2 Materials and methods

For this study raw Illumina sequencing reads from shotgun sequencing of the HA genes from viral sample which is not covered by vaccine protection were used.

Alignment and indexation of reads was performed using BWA (version 0.7.17).[9] Sorting and indexation of alignment file was performed using SAMtools[10] (version 1.3.1). Variants were identified using Varscan[11] (version 2.4.6) with a minimum variant frequency parameter set to 0.1% (–min-var-frequency=0.001) for rare variants and depth parameter set to maximum (-d=0). Results were analyzed using IGV[12] (version 2.16.2).

For the control experiment, Illumina sequencing reads from three shotgun sequencing runs of the HA gene sequences were employed using the isogenic reference sample. The processing of reads and the identification of variants was made as described above. SNPs in experimental sample with frequencies that are more than three standard deviations away from the averages in the control sample were identified using Pandas library.[13]

Impact of identified mutation on HA structure was visualized using AlphaFold2[14] and pictures were made using Pymol.[15] Original HA structure was obtained from rcsb database (pdb code: 4WE8).[16]

3 Results

A total of 358265 reads of hemagglutinin gene were obtained from the patient X sample, with an average read length approximately 150bp. The reads were indexed and aligned to partial reference CDC of hemagglutinin gene of influenza A virus A/USA/RVD1_H3/2011(H3N2) (GeneBank accession KF848938.1). The resulting coverage was 99.94% with all reads being mapped. Aligned reads were examined for variant calling with Varscan with minimum variant frequency parameter set to 0.1% and no depth limit. In result 21 SNPs were identified, 5 of which had high ($> 95\%$) frequencies in reeds.

To verify which low-frequency SNPs corresponded to actual influenza variants, we used three controls containing reads from isogenic influenza virus. For each control, we performed the same procedures as with the patient X sample. Additionally, we computed the mean frequency of variants and their corresponding standard deviations for each control sample. The initial read number, coverage, number of mapped sequences, and associated statistics for the controls are listed in table 1.

Variants from the patient X sample were filtered based on acquired statistics. Only variants with frequencies more than 3 standard deviations away from the averages in the control samples were selected for further examination. As the result there were identified 2 rare variants alongside with 5 high-frequency ones (table 2).

Control	Starting reads	Coverage	Mapped reads	Mean frequency	Std deviation
Control 1	256,586	99.97%	256,586	0.257	0.072
Control 2	233,327	99.97%	233,327	0.237	0.052
Control 3	249,964	99.97%	249,964	0.250	0.078

Table 1: *Summary for controls.*

Position	Base Change	Frequency	Type of SNP
72	A→G	99.96%	syn (Thr → Thr)
117	C→T	99.82%	syn (Ala → Ala)
307	C→ T	0.94%	mis (Pro → Ser)
774	T→C	99.96%	syn (Phe → Phe)
999	C→T	99.86%	syn (Gly → Gly)
1260	A→C	99.94%	syn (Leu → Leu)
1458	T→C	0.84%	syn (Tyr → Tyr)

Table 2: *Summary of SNPs. Syn - synonymous mutation, mis - missense. The single missense mutation is highlighted in yellow.*

Among all identified SNPs only one rare variant corresponds to missense mutation 103Pro → Ser (numeration from PDB: 4W8E[16] was used), while all other including high-frequency ones were identified as synonymous.

Structure of identified variant HA protein (P103S) was obtained using AlphaFold2.[14] 103Pro substitution (fig. 1 A) took place in loop region (fig. 1 C) and didn't crucially affect the global protein topology (fig. 1 B). However, according to Munoz et al,[17] 103Pro incorporated in D-epitope of the HA protein, which is among the primary targets of human antibodies against this antigen.

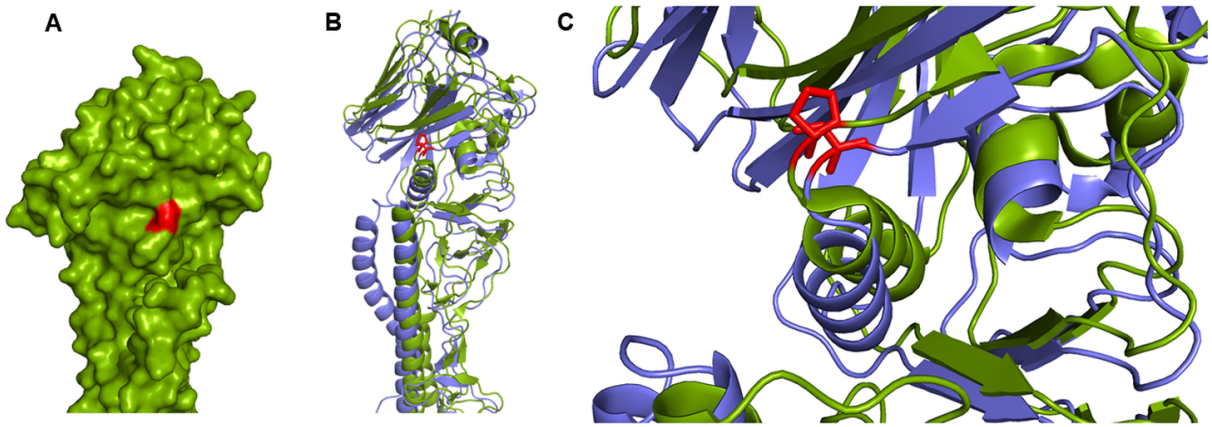

Fig 1: Structure of HA protein. **A**, Structure of HA protein of influenza A H3N2 (PDB: 4WE8) shown in surface representation (green) and specified position of 103Pro residue (red). **B**, Comparison of the native HA protein (green) with AlphaFold2 P103S mutant model (slate) tertiary structures. **C**, Comparison of the native HA protein (green) with AlphaFold2 P103S mutant model (slate) local conformation. Proline at position 103 is located in loop region and was substituted to Serine.

4 Discussion

Our findings suggest that the influenza virus affecting patient X, which showed no HI reaction, may be the result of antigenic drift from the original virus. This is supported by the discovery of only point SNPs compared to the reference, with only one of them influencing the HA protein. The 103Pro → Ser substitution doesn't shift overall HA protein conformation, but due to its location in D-epitope this mutation can significantly affect antibody affinity. This explains vaccine inefficiency, while low frequency of occurrence of this variant may explain the positive results of the HI test.

It is important to note that there is a challenge associated with identifying rare mutations due to the presence of noise. The less common the variant, the more difficult it is to differentiate it from errors originating from sample library preparation or sequencing. Deep sequencing enables a more precise differentiation between errors and genuine variants by achieving higher coverage depth. Still, if variant frequency lies near to the "error" frequency, as observed in our data, distinguishing between them even in such analysis can be challenging.

Several approaches can be used to improve such cases. Errors associated with library preparation can be reduced using high-fidelity polymerases in PCR. Another way is to reduce sequencing error by increasing replicates. It can be achieved by analyzing more samples with same source. Moreover, the same library can be sequenced multiple times. Based on this, approaches such as using overlapping read pairs (ORP)[18] allow for the identification of extremely rare variants. This methodology, as its name suggests, involves pair-end sequencing of the sample to generate pairs of overlapping forward and reverse reads. Combined with a control check to empirically derive error rates in PCR and sequencing, this approach enables the capture of variants with frequencies less than 0.05%.[18]

Finally, if it is financially possible, different platforms can be used to create independent reads of the same library which then can be compared to exclude obvious errors. This can be highly beneficial, hence different sequencing platforms have different prevailing type of errors. For example for Illumina reads, that were used in our work it is mainly nucleotide substitutions, while for PacBio it is insertions or deletions.[19]

References

1 G. W. Both, M. J. Sleigh, N. J. Cox, *et al.*, "Antigenic drift in influenza virus h3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites," *Journal of Virology* **48**, 52–60 (1983).

2 K. Shimizu, "[mechanisms of antigenic variation in influenza virus]," *Nihon rinsho. Japanese journal of clinical medicine* **58**, 2199–205 (2000).

3 R. G. Webster, W. G. Laver, G. M. Air, *et al.*, "Molecular mechanisms of variation in influenza viruses," *Nature* **296**, 115–121 (1982).

4 E. Domingo, V. Martín, C. Perales, *et al.*, "Viruses as quasispecies: Biological implications," in *Current Topics in Microbiology and Immunology*, 51–82, Springer Berlin Heidelberg (2006).

5 D. Steinhauer, J. C. de la Torre, and J. Holland, "High nucleotide substitution error frequencies in clonal pools of vesicular stomatitis virus," *Journal of virology* **63**(5), 2063–2071 (1989).

6 J. C. Pedersen, "Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus," in *Methods in Molecular Biology*, 11–25, Springer New York (2014).

7 M. E. Quiñones-Mateu, S. Avila, G. Reyes-Teran, *et al.*, "Deep sequencing: Becoming a critical tool in clinical virology," *Journal of Clinical Virology* **61**, 9–19 (2014).

8 H. Chen, X. Wen, K. K. W. To, *et al.*, "Quasispecies of the d225g substitution in the hemagglutinin of pandemic influenza a(h1n1) 2009 virus from patients with severe disease in hong kong, china," *The Journal of Infectious Diseases* **201**, 1517–1521 (2010).

9 H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics* **25**, 1754–1760 (2009).

10 P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of SAMtools and BCFtools," *GigaScience* **10** (2021).

11 D. C. Koboldt, Q. Zhang, D. E. Larson, *et al.*, "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research* **22**, 568–576 (2012).

12 J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, *et al.*, "Integrative genomics viewer," *Nature Biotechnology* **29**, 24–26 (2011).

13 T. pandas development team, "pandas-dev/pandas: Pandas," (2020).

14 J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature* **596**, 583–589 (2021).

15 Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8." (2015).

16 H. Yang, P. Carney, J. Chang, *et al.*, "The crystal structure of hemagglutinin of influenza virus a/victoria/361/2011," (2015).

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

18 H. Chen-Harris, M. K. Borucki, C. Torres, *et al.*, "Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs," *BMC Genomics* **14**(1), 96 (2013).

19 X. Yang, S. P. Chockalingam, and S. Aluru, "A survey of error-correction methods for next-generation sequencing," *Briefings in Bioinformatics* **14**, 56–66 (2012).

5 Supplementary materials

Code availability

Scripts for running deep sequencing analysis for rare variants is available at GitHub repository: https://github.com/PolinaVaganova/deep_sequencing_influenza_variants_identifying.

Control Samples

The fastq files for the three control samples are available at the following links:

1. Control 1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz
2. Control 2: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz
3. Control 3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

Analysed Data

Reeds from the patient X can be accessed at the following link: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/.