Practice Bioinformatics Institute

# Mutated Influenza A virus (H3N2) hemagglutinin gene causes resistance to vaccine antibodies

Anastasia Kolchina[1] and Anna Chechenina[2]

[1] Faculty of Computer Science, Higher School of Economics, Moscow, 101000, Russia
[2]Bioinformatics department, Escola Superior de Comerc International - Universitat Pompeu Fabra (ESCI-UPF), Barcelona, 08003, Spain

## Abstract

Because of the rapid mutation rate of the influenza virus, it quickly becomes resistant to new vaccines. Nowadays design of the annual influenza vaccine is based on the analysis of the hemagglutinin and neuraminidase mutations. We have recently discovered the new variation of the Influenza A virus strain A/Hong Kong/4801/2014 (H3N2) escaping the vaccine antibodies. We performed the deep sequencing analysis of hemagglutinin H3 protein and located the crucial mutations. Both synonymous and non-synonymous substitutions were detected, seven of them were considered significant in terms of not being sequencing errors.
Particularly, the missense mutation of P103 turned out to be in one of the viral epitopes. We supposed that this mutation could significantly affect the epitope C, and could cause the lack of response to this year vaccine. Besides this investigation, we discussed ways of distinguishing mutations from sequencing errors, and some methodological approaches.
Supplementary materials can be found via link: https://github.com/netwasted/bioinf-institute-project-2.

**Keywords:** Influenza; hemagglutinin; variant calling; mutation; epitope

## Introduction

Seasonal influenza virus infections are a major public health concern and could cause significant mortality. Majority of vaccines nowadays include specific antibodies that block the binding of the hemagglutinin to the host cell receptor (Krammer and Palese 2013). Hemagglutinin (HA) is one of the surface glycoproteins of the influenza virion. When the virus enters the organism, it attaches to polysaccharide chains on the surface of red blood cells containing sialic acid residues. When the antibody attaches to the hemagglutinin, the molecule loses its ability to attach to its receptor and the virus cannot enter the cell.

Hemagglutinin is a homotrimeric membrane glycoprotein, where each domain contains the transmembrane subdomain, the vestigial esterase subdomain, the fusion subdomain, and a receptor-binding globular subdomain (Gamblin *et al.* 2021). However, due to the high variability of this region, the response to the vaccine is strain specific and the antibodies have neutralizing activity against only closely related strains (Margine *et al.* 2013). Hence, the effectiveness of vaccination relies on the accuracy of the prediction of the vaccine strains every season. For example, for the season 2022-2023 among all genetically characterised specimens the H3N2, influenza A, strain was detected in 98% cases (Merced-Morales *et al.* 2022). The recommended A(H3N2) component was changed to an A/Darwin/9/2021 (H3N2)–like virus for egg-based vaccines and an A/Darwin/6/2021 (H3N2)–like virus for cell-based or recombinant vaccine (Merced-Morales *et al.* 2022).

The most common target for the influenza vaccines is a globular head of the protein, but since these subdomains are responsible for the receptor recognition, it is highly variable among different strains. Recently, another class of antibodies directed against the highly conserved stalk domain of the HA has gained increasing attention. Stalk-reactive antibodies tend to exhibit a much broader reactivity and neutralizing activity than antibodies targeting the globular head domain (Krammer and Palese 2013; Margine *et al.* 2013).

The virus evolves by several mechanisms, the most described are antigenic drift, antigenic shift and homologous recombination (De 2018) and exist in diverse, albeit related, viruses, or quasispecies. Antigenic drift is a change in the antigenic region of the protein resulting in the loss of immunity in hosts even in vaccinated population. Antigenic drifts in the HA region of the virus result in polymorphism in the antigenic site, the substitutions are formidable enough to cause an epidemic or a seasonal outbreak (Caton *et al.* 1983). That occurs through replication errors since these RNA viruses do not possess error correction mechanisms (De 2018).

New mutations could be also more dramatic and could be introduced by the genetic reassortments (antigenic shift) in the antigenic regions (Gething *et al.* 1980; Shil *et al.* 2011). Antigenic shift occurs when two or more influenza strains infect the same cell, some viral gene segments may be exchanged and some of the progeny viruses may end up with new combinations of the hemagglutinin and neuraminidase subtypes (De 2018).

Usually, to understand the relation between influenza strains the antigenic distance is calculated which represent the commulative difference between epitopes of these viruses (Shil *et al.* 2011). The changes in the epitope region might affect the interaction with an antibody in a different ways. In order to understand how to much the mutation will affect binding, docking of the antibody onto the 3D structures of HA proteins could be performed (Shil *et al.* 2011). There are several computational methods allowing to identify the antigenic drift of influenza A utilizing the conformation changes on epitopes (Huang and Yang 2011; Tewawong *et al.* 2015), including deep learning approaches (Xia *et al.* 2021).

Homologous recombination occurs in case of presence of different subtypes of the virus in a cell at the same time. Then RNA polymerase could switch the replication template from one viral strand to an equivalent strand of another strain. This type of mutational processes is rare in influenza viruses but still possible with numerous intra-segmental homologous recombination events reported (Hao 2011; He *et al.* 2009).

To describe the strains of the virus currently present in the sample the deep sequencing approach with coverage up to hundreds of thousands per position is widely used. It corresponds to the NGS with a high coverage on the gene of interest, which is usually hemagglutinin and neuraminidase in case of influenza virus (Barbezange *et al.* 2018; Bidzhieva *et al.* 2014). Then using the distribution of the variants frequencies intervals to count mutation as significant are calculated. Mutations with very low frequencies are considered as errors during sample preparation and sequencing. However, heterogeneous population, even with variants at very low frequency, could facilitate or speed up evolution (Barbezange *et al.* 2018).

In our research we focused on the hemagglutinin gene from the strain A/USA/RVD1/H3/2011(H3N2) (Cushing *et al.* 2015), identified in the provided sample. We preformed the analysis of the raw deep sequencing Illumina data using three control samples. Then we aligned the control and experiment reads to the reference gene and infer the frequencies of genetic variations on the different positions. Based on the average and deviations of the control samples we described significant mutations in the experiment sample.

## Methods

The reference sequence, which we consider non-mutated Influenza A virus, is a segment 4 hemagglutinin (HA) gene from a strain called A/USA/RVD1_H3/2011(H3N2), obtained at GenBank database from NCBI. The origin of the strain we study is unknown, but its HI profile is closely matched with A/Hong Kong/4801/2014 (H3N2). In addition to that, three isogenic reference samples from SRA FTP were used as control, their code names are SRR1705858, SRR1705859 and SRR1705860. We aligned the examined sequence to reference using BWA-MEM with default parameters. Samtools was used to sort the resulting file, index it and build a pileup. Common variants were obtained using VarScan, firstly with a high minimum variant frequency cut-off of 0.95 to find common variants, and then with a minimum variant frequency of 0.001 to find some rare variants. We also used IGV browser to visualize the positions of the variants we obtained. As control we aligned the three previously described isogenic reference samples to check which of the found variants (minimum variant frequency = 0.001) are actually mutations and not simply sequencing errors.

The variants with very low frequency variants might be unreliable and might arise due to the polymerase mistakes during PCR or sequencing itself. To identify the reasonable interval for the mutations frequency, when we can count mutation as significant and not a procedure error, we calculated the mean an standard deviations of the mutations frequencies in the control samples. Then if a mutation on the experiment sample will deviate from the average by more than three standard deviations, we consider this mutation significant.

**Table 1** Number of reads

| Name | Total number of reads | Mapped reads to reference |
|---|---|---|
| Examined sequence | 361349 | 361116 (99.94%) |
| SRR1705858 (1st control) | 256744 | 256658 (99.97%) |
| SRR1705859 (2nd control) | 233451 | 233375 (99.97%) |
| SRR1705860 (3rd control) | 250184 | 250108 (99.97%) |

**Table 2** Statistics of variant frequencies in control reference sequences

| Name of control reference sequence | Average of frequencies | Standard deviation of frequencies |
|---|---|---|
| SRR1705858 | 0.2565 | 0.0717 |
| SRR1705859 | 0.2369 | 0.0524 |
| SRR1705860 | 0.2503 | 0.078 |

We calculated average and standard deviation for each of the three samples, selected the result with the biggest standard deviation and constructed a standard confidence interval $(mean - 3\dot{s}td, mean + 3\dot{s}td)$, which accounts for $(0.022, 0.478)$ in our case. Therefore, all the variants with frequency out of the interval we considered significant and actual mutations.

## Results

Raw data contained 230000-37000 reads in total, more than 99% of which were successfully mapped to the reference sequence (see Table 1).

The average frequency in all control samples was around 0.23-0.25, when the standard deviance varied from 0.05 to almost 0.08 (see Table 2). Even though we were mainly focused on low frequency mutations, we identified five mutations with frequency more than 95% in the experiment sample: 72A-G , 117C-T, 774T-C, 999C-T, 1260A-C. All these mutations except 999C-T were present in the control samples.

Then, as stated before, we used the biggest values of standard deviation and average, that corresponds to SRR1705860 sample, to calculate the confidence interval. Using calculated confidence intervals we found seven mutations we consider significant (non-errors), as all of them are out of the confidence interval (see Table 3). Five of them belong to the group of mutations with frequency more than 95%, but two other mutations, 307C-T and 1458T-C, have frequency around 1%. All mutations except 307C-T occurred to be synonymous substitutions, whereas change of this position corresponds to change of amino acid 103 in the protein from proline to serine. We can see that uncharged amino acid was substituted by another polar uncharged amino acid, so chemical properties of the radical did not change dramatically. However, polar charges of serine are more significant because of

**Table 3** Significant variants (proposed mutations)

| Position | Frequency | Base substitution | Amino acid substitution | Type of substitution |
|---|---|---|---|---|
| 1458 | 0.84 | T >C | - | Synonymous |
| 307 | 0.94 | C >T | P >S | Non-synonymous |
| 117 | 99.82 | C >T | - | Synonymous |
| 999 | 99.86 | C >T | - | Synonymous |
| 1260 | 99.94 | A >C | - | Synonymous |
| 72 | 99.96 | A >G | - | Synonymous |
| 774 | 99.96 | T >C | - | Synonymous |

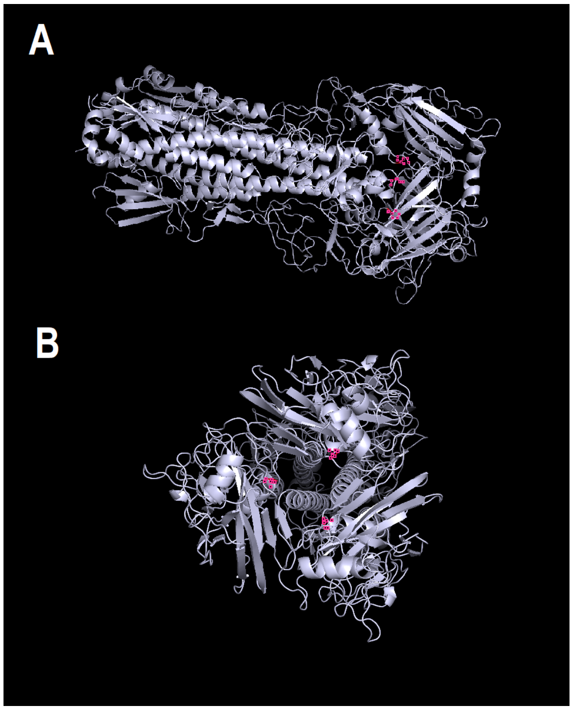

**Figure 1** Model of the Influenza A virus (A/USA/RVD1_H3/2011(H3N2)) hemagglutinin (AHB59323.1). Mutated residue 103 is highlighted in red. A - side view, B - frontal view to the receptor-binding head.

OH-group, so this substitution might affect some coordinating residues around this base. Moreover, this mutation is located in the Epitope C, reported in the (Muñoz and Deem 2005).

## Discussion

We suppose that one of the reasons why the experimental influenza virus strain is not affected by the vaccine antibodies is the mutation at base 307, which corresponds to the position 103 in the protein. This position is located in the Epitope C, so the change might be crucial for the ability of antibody to bind this antigenic region. Even though the significance of Epitope C is cryptic, we suppose its mutation to be the main investment in viral resistance to vaccine. We modelled the structure of the possible hemagglutinin trimer and showed that the mutation (selected in red, see Fig. **??**)

Due to the decreasing sequencing price and rapid progress of these technologies, there is a great need to systematically evaluate sequencing errors at various steps of a NGS analysis workflow, as this knowledge will help improve low-level variant detection by deep sequencing (Ma *et al.* 2019). According to the the mean error rate was around 1% and error-free haplotypes represent 10.09% - 67.57% of the total number of reads, depending on the read length (Skums *et al.* 2012). There are many possible sources of sequencing errors: sample handling, library preparation, PCR enrichment, and sequencing. In our work we differ these errors from possible mutations by calculating the confidence interval for the control samples and selecting only mutations with frequencies out of the interval. However, there are some other more complex methods described in the literature. One of the most popular technologies is based on the reads clustering using, for example, distance measure that models sequencing noise or using Bayesian statistics with the Dirichlet process mixture. Some other ways include construction of multiple sequence alignments, k-mers, or substrings of reads of a fixed length k (Pevzner *et al.* 2001; Zagordi *et al.* 2010; Salmela and Schröder 2011; Skums *et al.* 2012).

The mentioned methods assume that the errors are randomly distributed, but it was shown in some works that the error rate is strongly affected by the presence of homopolymers, position in the sequence, size of the sequence and spatial localization during sequencing (Skums *et al.* 2012). That implies the need of introducing some correction terms to the reads data before analysis. One of the possible steps to avoid errors that does not require any additional advanced software is trimming. During this procedure we can eliminate unpaired reads or cut out the low-quality edges or the adaptor sequences. Speaking about laboratory methods to increase the quality of reads we can suggest to try different polymerases, adjust amplification conditions and maybe perform a DNA repair treatments during library construction (Ma *et al.* 2019).

## Literature cited

Barbezange C, Jones L, Blanc H, Isakov O, Celniker G, Enouf V, Shomron N, Vignuzzi M, van der Werf S. 2018. Seasonal genetic drift of human influenza a virus quasispecies revealed by deep sequencing. Front. Microbiol.. 9:2596.

Bidzhieva B, Zagorodnyaya T, Karagiannis K, Simonyan V, Laassri M, Chumakov K. 2014. Deep sequencing approach for genetic stability evaluation of influenza a viruses. J. Virol. Methods. 199:68–75.

Caton AJ, Raymond FL, Brownlee GG, Yewdell JW, Gerhard W. 1983. Antigenic variation in influenza virus. Biochem. Soc. Trans.. 11:435–441.

Cushing A, Kamali A, Winters M, Hopmans ES, Bell JM, Grimes SM, Xia LC, Zhang NR, Moss RB, Holodniy M *et al.* 2015. Emergence of hemagglutinin mutations during the course of influenza infection. Sci. Rep.. 5:16178.

De A. 2018. Molecular evolution of hemagglutinin gene of influenza a virus. Front. Biosci. (Schol. Ed.). 10:101–118.

Gamblin SJ, Vachieri SG, Xiong X, Zhang J, Martin SR, Skehel JJ. 2021. Hemagglutinin structure and activities. Cold Spring Harb. Perspect. Med.. 11:a038638.

Gething MJ, Bye J, Skehel J, Waterfield M. 1980. Cloning and DNA sequence of double-stranded copies of haemagglutinin genes from H2 and H3 strains elucidates antigenic shift and drift in human influenza virus. Nature. 287:301–306.

Hao W. 2011. Evidence of intra-segmental homologous recombination in influenza a virus. Gene. 481:57–64.

He CQ, Xie ZX, Han GZ, Dong JB, Wang D, Liu JB, Ma LY, Tang XF, Liu XP, Pang YS *et al.* 2009. Homologous recombination as an evolutionary force in the avian influenza a virus. Mol. Biol. Evol.. 26:177–187.

Huang JW, Yang JM. 2011. Changed epitopes drive the antigenic drift for influenza a (H3N2) viruses. BMC Bioinformatics. 12 Suppl 1:S31.

Krammer F, Palese P. 2013. Influenza virus hemagglutinin stalk-based antibodies and vaccines. Curr. Opin. Virol.. 3:521–530.

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J *et al.* 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 20:50.

Margine I, Hai R, Albrecht RA, Obermoser G, Harrod AC, Banchereau J, Palucka K, García-Sastre A, Palese P, Treanor JJ *et al*. 2013. H3N2 influenza virus infection induces broadly reactive hemagglutinin stalk antibodies in humans and mice. J. Virol.. 87:4728–4737.

Merced-Morales A, Daly P, Abd Elal AI, Ajayi N, Annan E, Budd A, Barnes J, Colon A, Cummings CN, Iuliano AD *et al*. 2022. Influenza activity and composition of the 2022-23 influenza vaccine - united states, 2021-22 season. MMWR Morb. Mortal. Wkly. Rep.. 71:913–919.

Muñoz ET, Deem MW. 2005. Epitope analysis for influenza vaccine design. Vaccine. 23:1144–1148.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 98:9748–9753.

Salmela L, Schröder J. 2011. Correcting errors in short reads by multiple alignments. Bioinformatics. 27:1455–1461.

Shil P, Chavan S, Cherian S. 2011. Molecular basis of antigenic drift in influenza A/H3N2 strains (1968-2007) in the light of antigenantibody interactions. Bioinformation. 6:266–270.

Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y. 2012. Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinformatics. 13 Suppl 10:S6.

Tewawong N, Prachayangprecha S, Vichiwattana P, Korkong S, Klinfueng S, Vongpunsawad S, Thongmee T, Theamboonlers A, Poovorawan Y. 2015. Assessing antigenic drift of seasonal influenza A(H3N2) and A(H1N1)pdm09 viruses. PLoS One. 10:e0139958.

Xia YL, Li W, Li Y, Ji XL, Fu YX, Liu SQ. 2021. A deep learning approach for predicting antigenic variation of influenza a H3N2. Comput. Math. Methods Med.. 2021:9997669.

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. 2010. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J Comput Biol. 17:417–428.