IDENTIFICATION AND STATISTICAL ESTIMATION OF LOW-FREQUENCY VARIANT CALLING FOR INFLUENZA VIRUS A/HONG KONG/4801/2014 (H3N2) STRAIN HEMAGGLUTININ USING TARGETED DEEP SEQUENCING

ANNA MALYKHINA
Bioinformatics Institute
anja12@yandex.ru

OKSANA SIDORENKO
Bioinformatics Institute
oxiksid@gmail.com

MIKHAIL RAYKO
Bioinformatics Institute

November 17, 2022

ABSTRACT

One of the urgent problems of vaccination is the lack of vaccine coverage of new viral strains that actively appear due to their rapid evolution, antigenic drift and accumulation of quasispecies within single host organism. In this work we have analyzed the case of influenza infection after vaccination. The source of infection was the person whose HI test results closely matched the HI profile for A/Hong Kong/4801/2014 H3N2 strain.

Here, a low-frequency missense mutation (Pro103Ser) in the epitope D sequence of viral hemagglutinin (HA) was detected. Due to the different chemical properties of proline and serine this mutation could block epitope specific association with the antibody paratope while high-frequency variants turned out to be synonymous and could not cause a change in the epitope protecting antibodies from binding.

Targeted deep sequencing has proven to be a sensitive, promising, and effective method for detecting rare variants, but the problem of separating errors from the real mutations is still relevant.

***K*eywords** Antigenic drift · Influenza · Vaccination · Variant calling · Rare variants · NGS

1 INTRODUCTION

There are various types of vaccines for the prevention of influenza: living (imitate a natural infection), inactivated (killed, with quite preserved capsid proteins to be recognized as a recipient immune system), subsidiary (do not contact components of the pathogen and differ from inactivated vaccines by containing only the epitopes parts (also known as antigenic determinant) that most readily stimulate the immune system). [1] [2]

In any case, the action of vaccines is based on the provocation of the immune system, which responses by developing antibodies, recognizing epitopes on the surface of antigens and connecting antibodies to antigens. The binding between epitopes and paratopes (the parts of an antibody that binds to the epitope) is exquisitely specific and of high affinity. Comprehensive description of the epitopes/paratopes, ideally to the residue level, is crucial to understand the binding mechanism and to design future therapeutic agents. [3] The slightest change in the epitope can lead to a violation of its specific connection with the paratope which is very dangerous in context of the extremely quick mutation process in viruses: influenza mutates at a rate of one mutation per genome per replication. The rapid evolution of influenza

virus under immune pressure is likely enhanced by the virus's genetic diversity within a host. [4] The result of the accumulation of new mutations that are introduced into currently circulating viral strains genomes over time is called **antigenic drift** - kind of genetic variation in viruses which is attributable to influenza's low fidelity RNA polymerase that lacks a function for error proof-reading. With each replication cycle, polymerase errors create $de\ novo$ mutations, increasing the genetic diversity of the virus. Drift can lead to changes in the hemagglutinin and neuraminidase (NA, another major surface glycoproteins of influenza A virus along with hemagglutinin), explaining why this phenomenon can impacts antibody binding. [5]

Such a high mutation rate and constant antigenic drift are the reason why viruses can exist as **quasispecies**, even within a single host (and be involved in the accumulation of mutations and antigenic drift processes as new reservoirs, increasing the viral pathogenic potential).

Mutant distributions in RNA virus populations (mutant clouds), were first proposed in a theory of molecular evolution termed quasispecies theory. [6]. Viral quasispecies refers to a population structure that consists of extremely large numbers of variant genomes, termed mutant spectra, mutant swarms or mutant clouds. In virology, quasispecies are defined as complex distributions of closely related variant genomes subjected to genetic variation, competition and selection, and that may act as a unit of selection. [7]

Thus, Barbezange et al. demonstrated differences in the intrinsic genetic diversity between subtypes and showed that the composition of the quasispecies evolves season after season by evaluating the virus quasispecies diversity directly in infected human respiratory specimens. Comparison of two seasons for a given subtype strikingly highlighted the evolutionary dynamics and quasispecies plasticity of influenza A viruses, which illustrated the permanent, underlying genetic drift occurring in human influenza viruses. [8]

If more than one influenza strain infects the same host, they can recombine. When genetic segments are mixed with new components, sometimes between species, an **antigenic shift** occurs. Antigenic shift should not be confused with antigenic drift. Antigenic shift an abrupt, major change in a flu A virus, resulting in new HA and/or new HA and NA proteins in flu viruses that infect humans. Antigenic shift can result in a new flu A subtype. Shift can happen if a flu virus from an animal population gains the ability to infect humans. Such animal-origin viruses can contain HA or HA/NA combinations that are different enough from human viruses that most people do not have immunity to the new (e.g., novel) virus. Such a "shift" occurred in the spring of 2009, when an H1N1 virus with genes from North American Swine, Eurasian Swine, humans and birds emerged to infect people and quickly spread, causing a pandemic. When shift happens, most people have little or no immunity against the new virus.

While flu viruses change all the time due to antigenic drift, antigenic shift happens less frequently. Flu pandemics occur rarely; there have been four flu pandemics in the past 100 years. Type A viruses undergo both antigenic drift and shift and are the only flu viruses known to cause pandemics, while flu type B viruses change only by the more gradual process of antigenic drift. [5]

Given all of the above characteristics of viruses (in particular, Influenza A Virus) as highly variable infectious agents, the most accurate identification of all variants in a mixed viral population plays a key role in diagnosis, prevention approaches, and understanding the development and course of viral infections. This is an important and relevant issue regarding vaccination because influenza vaccines must be very frequently reformulated to account for antigenic changes in the viral envelope protein, hemagglutinin. [4] There are several approaches to identifying virus variants, but each of them has its drawbacks: - HI can only detect viruses that match the ones used to make the antibodies - Conventional sequencing can identify new strains, but will miss rare variants - Next generation 'deep sequencing' can study all variants in mixed population, but this approach requires very thorough error control and analysis for accurately identifying and quantitating rare variants. Specific errors may occur at equal or even higher frequencies than true biological mutations, therefore a powerful assessment of low-frequency virus mutations is seriously jeopardized. [14]

Errors acquired during next-generation sequencing (NGS) are key confounding factors of sensitive detection of low-frequency variants by deep sequencing. Errors can occur on various steps of a conventional NGS workflow, such as sample handling, library preparation, PCR enrichment ("upstream" errors) and sequencing ("during" errors: in cluster generation or sequencing by synthesis, e.g. incorrect or missing bases). [9]

In this work we have analyzed the case of influenza infection after vaccination against the viral strain covered by the vaccine. The source of infection was the person whose HI test results closely matched the HI profile for A/Hong Kong/4801/2014 H3N2 strain. We have analyzed results of Illumina single-end targeted deep sequencing of HA genes of viral sample from the infected person. As part of our work based on the identification of rare variants we used a statistical approach to separate errors from the desired rare mutations.

2 METHODS

2.1 SEQUENCE SOURCE

Raw Illumina sequencing reads of the sample (results of Illumina single-end sequencing run on HA genes of viral sample from infected person) were downloaded from the SRA FTP server:

`http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/`

Fastq data for the three controls (from sequencing of isogenic reference samples) were also downloaded from SRA FTP.

`ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz`

`ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz`

`ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz`

2.2 DATA ANALYSIS

For reads quality analysis Fastqc software was used (ver.0.11.9). For aligning sequences to reference bwa (ver.0.7.17-r1188) and samtools (ver.1.16.1) were used with default settings. Variant calling was performed by Varscan (ver.2.4.4) with threshold 90% for frequent variants (there were 5 variants reported back) and 0.1% for rare variants (there were 21 variants reported back: 5 not rare, we had already detected them with threshold 90%, and 16 rare). Automatic SNP annotation was done by SnpEff (v.5.1d). (For detailed description, see section 5.SUPPLEMENTARY). Visualization was done on SWISS-MODEL software. 99% confidence interval was calculated as mean ± 3SD.

3 RESULTS

Quality control analysis on initial file revealed that the reads are of good quality [Figure 1]. As a result of applying bwa and samtools alignment 890569 reads were mapped (99.94%). Varscan with frequency filter of 95% reported 5 variant positions in alignment file in comparison with reference sequence. These variants appeared in sample with high frequency of more than 99%. However, all variants are synonymous, so they unlikely have any impact on protein structure and function.

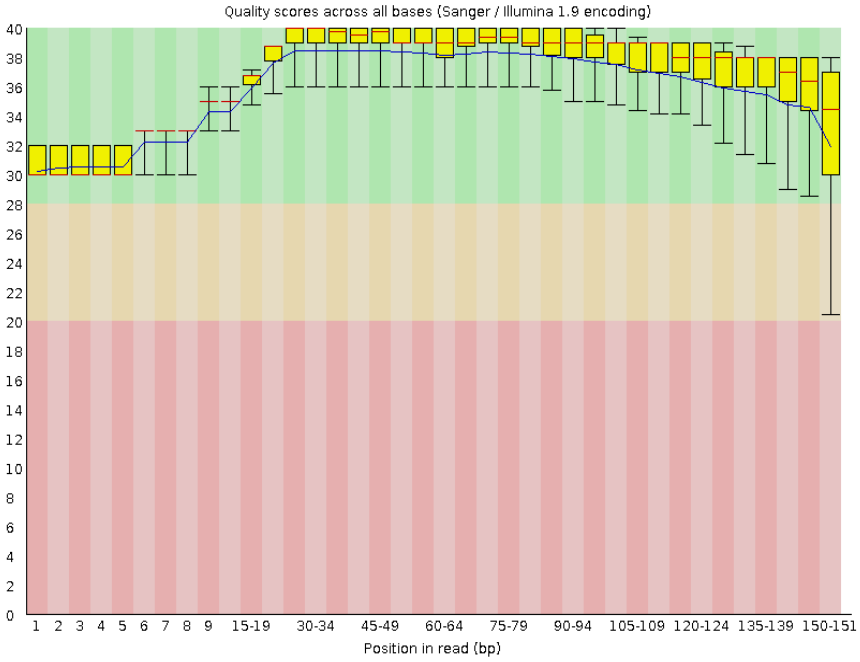

Figure 1: Quality control of the sample reads.

To reveal some rare mutation Varscan filter was set to 0,1%. As a result, additional 18 rare variants were found with occurrence less than 1%. Such low frequency could be not only due to change in sequence itself but also due to errors in sequencing process.

To evaluate noise level in sequencing process, 3 isogenic reference samples were analyzed on the same Illumina machine. Results were processed in the same way as experimental sample. Each repetition showed rare variants with frequency less than 1% [Table 1].

	Number of SNP	Mean frequency value (%)	SD of frequency (%)	99% confidence interval (%)
Control 1	57	0.256	0.072	0.04 - 0.47
Control 2	52	0.237	0.052	0.08 - 0.39
Control 3	61	0.250	0.078	0.02 - 0.48

Table 1: Confidence intervals for all controls are similar. So, we will define noise level as 0.48% and cut off variants below this level. [Figure 2]

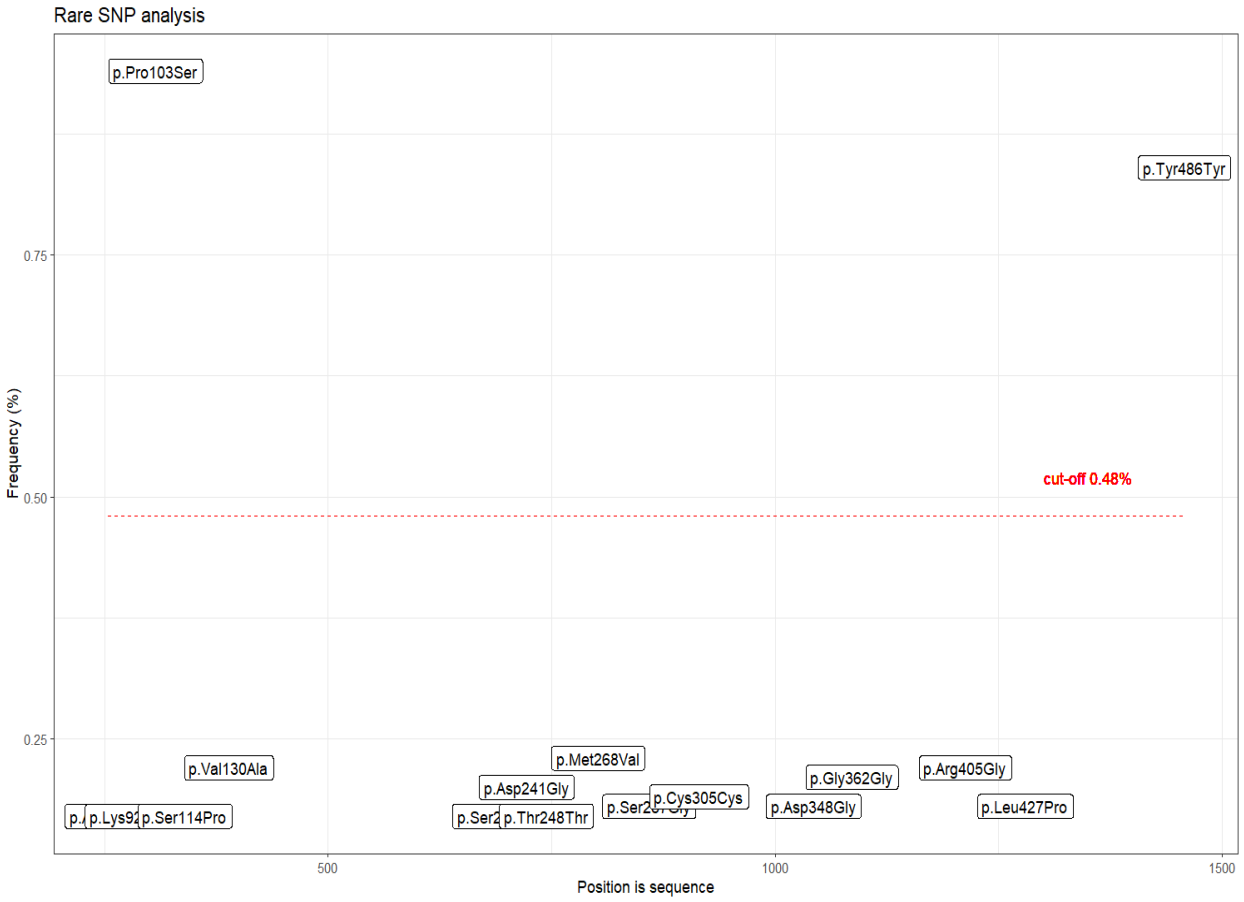

Figure 2: Rare variants in sample with their frequencies. Mutations below cut-off level considered to be a noise.

Position of nucleotide	Frequency	Reference nucleotide	Substitution nucleotide	Position and name of amino acid substitution	Impact of variant
72	99.96%	A	G	Thr24Thr	Synonymous variant
117	99.82%	C	T	Ala39Ala	Synonymous variant
774	99.96%	T	C	Phe258Phe	Synonymous variant
999	99.86%	C	T	Gly333Gly	Synonymous variant
1260	99.94%	A	C	Leu420Leu	Synonymous variant
307	0.94%	C	T	Pro103Ser	Missense variant
1458	0.84%	T	C	Tyr486Tyr	Synonymous variant

Table 2: All mutations found in the sample

All significant mutations found in the sample are sum up in Table 2.

From two rare variants only one is appeared to be missense mutation (Pro103Ser). Serine is classified as a polar amino acid, while proline is an aliphatic amino acid. Such substitution should change tertiary structure significantly. Visual model of this substitution is on Figure 3.

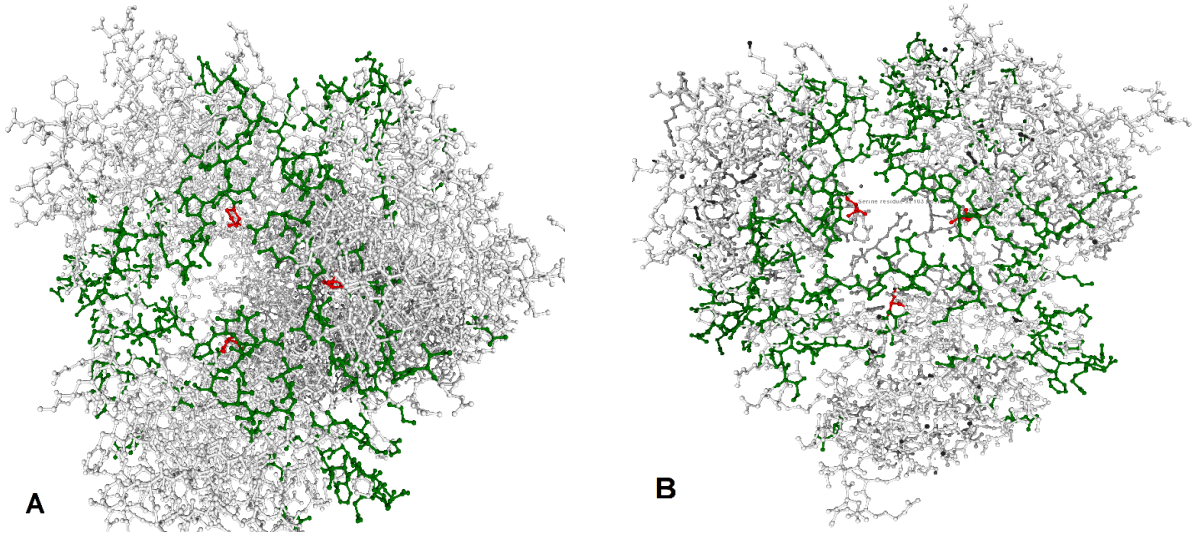

Figure 3: Model of substitution of proline (A) to serine (B) in 103 position in influenza virus hemagglutinin A/Hong Kong/4801/2014 (H3N2). Software: SWISS-MODEL[18]

According Munoz et al [17] amino acid in 103 position is situated in Epitope D. Thus, it plays an important role in antigen-antibody interaction.

Figure 4 shows the visual model of the crystal structure of A/Hong Kong/4801/2014 (H3N2) hemagglutinin epitope D residues and human antibody in complex with H3N2 hemagglutinin.

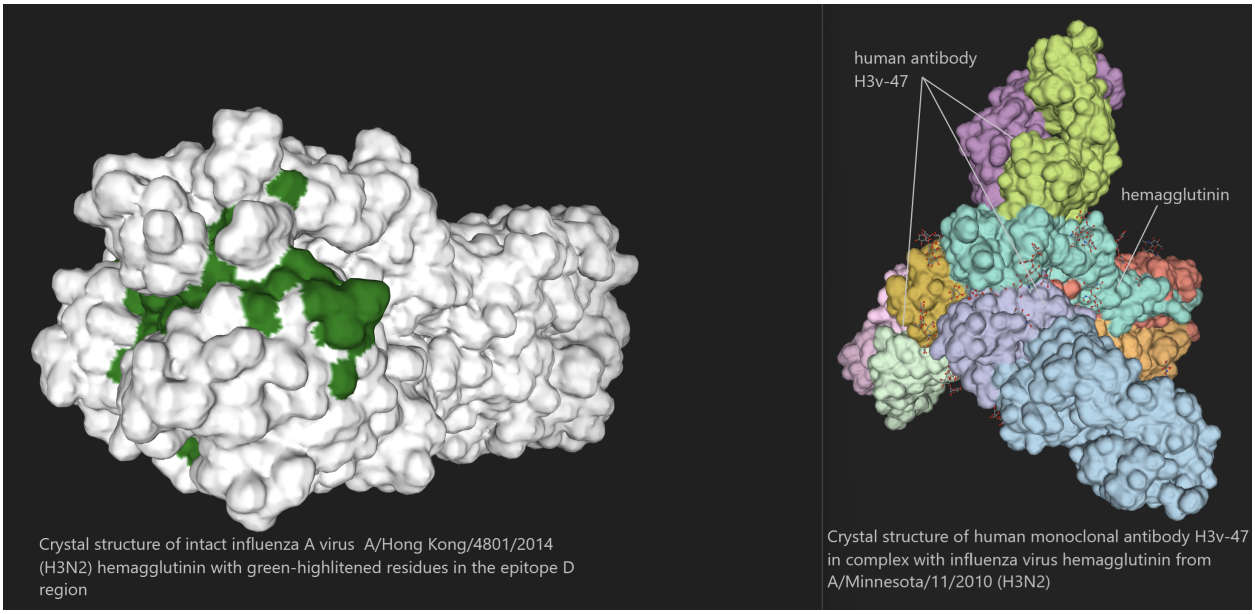

Figure 4: Crystal structures of intact H3N2 hemagglutinin and its complex with human antibody.
Software: SWISS-MODEL[18]

4 DISCUSSION

4.1 DATA

We have established that the common variants could not allow virus to escape the antibody response that developed after influenza vaccination, and the only missense mutation was rare variant Pro103Ser. Such a mutation can only be detected using targeted deep sequencing, but it is necessary to clearly distinguish errors of different levels from a reliable result.

We used the frequency of the errors from the control to figure out what's an error and what's a true variant in the data from the person that was the source of infection by calculating the average and standard deviation of the frequencies reported within each list (3 averages and 3 standard deviations from 3 reference samples). Since the goal of the control analysis was to establish the noise level, we only used low frequent variants ($> 0.1\%$). Our controls don't have high frequences (>90%), but if they did, we wouldn't be using them as they will be very significant outliers that will distort the calculations. Moreover, they would refer to stable mutations supported by an overwhelming number of reads. There can not be so many sequencing errors.

We believe that rare variants falling within the range of less than 3 standard deviations away from the averages in the reference files are errors that can be divided into two groups: "upstream" errors (appeared before the sequencing step and since the library preparation was common for three controls, appear in all three samples) and "during" errors (differed within the three controls as they appeared during sequencing).

4.2 MUTATION EFFECT

As we have already mentioned, proline and serine are chemically different enough amino acids to cause a change in the epitope D sufficient to block its specific association with the antibody paratope. Proline is a hydrophobic cyclic nonpolar imino acid, while serine is a hydrophilic hydroxylic polar amino acid. As shown in the Figure 5, Pro103 forms a weak hydrogen bond with Ile232 and hydrophobic contact with Tyr233, which are absent in the case of Ser103.

The changes we found in the epitope D may explain the considered case of influenza infection after vaccination against the viral strain covered by the vaccine, but a more in-depth analysis is needed to accurately establish the mechanisms, which is beyond the scope of this work.

Moreover, based on the information presented in the section 1.INTRODUCTION, it can be concluded that it is dangerous to completely rely on vaccination. The vaccine may not cover all variants of the virus (a striking example: Recommendations announced for influenza vaccine composition for the 2022-2023 northern hemisphere influenza

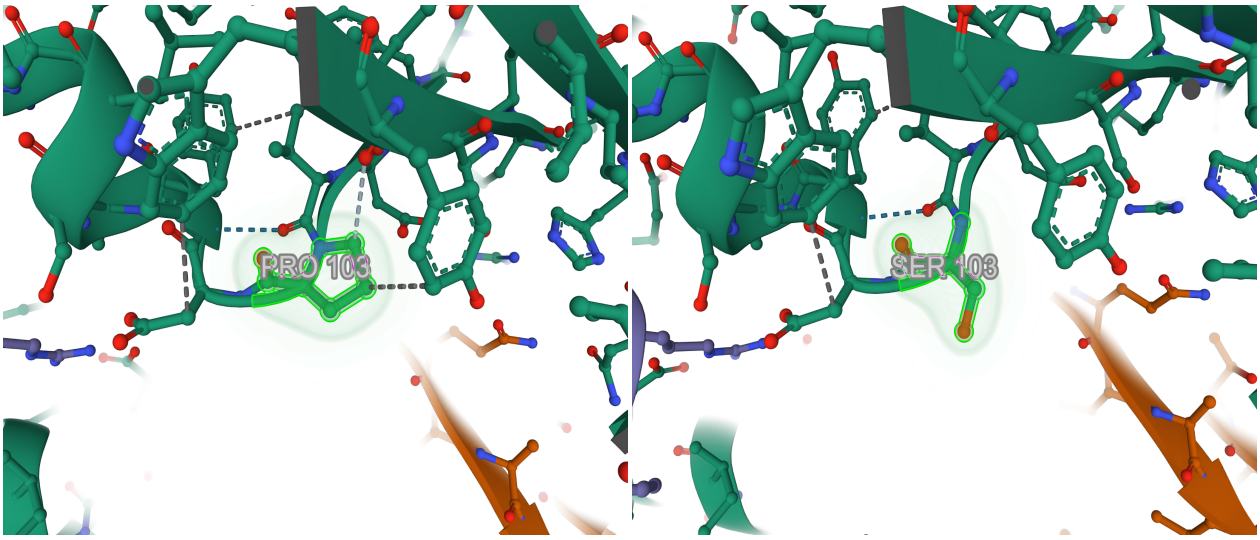

Figure 5: Model of changes in chemical bonds in the case of a proline-to-serine substitution mutation in the 103 position of the hemagglutinin H3N2 strain A/Hong Kong/4801/2014 (H3N2). Software: SWISS-MODEL[18], RCSB PDB[19]

season does not imply A / Hong Kong / 4801/2014 (H3N2) strain, which we are talking about in this paper). Additional preventive measures should be taken and contact with infected people should be avoided.

4.3 ERROR CONTROL PROPOSALS

Even on the example of our fairly simple study, we were convinced of the importance and operose of error control during targeted deep sequencing. With next-generation sequencing technologies, it is now feasible to efficiently sequence patient-derived virus populations at a depth of coverage sufficient to detect rare variants. However, each sequencing platform has characteristic error profiles, and sample collection, target amplification, and library preparation are additional processes whereby errors are introduced and propagated. Many studies account for these errors by using ad hoc quality thresholds and/or previously published statistical algorithms. [12] Most low frequency variant calling algorithms consider multiple sequencing characteristics such as strand bias, base quality, mapping quality, sequence context. [13] But this problem remains relevant and requires an effective solution, since the method itself is very important for public health service.

We can advise using high-fidelity polymerases for PCR to limit the introduction of the “upstream” errors during the viral genome amplification as well as studying and using tools and techniques that are being actively developed to solve the problem of error control (e.g. ViVaMBC [14], CleanDeepSeq [15]). There are also authors’ proposals to the manufacturer Illumina to increase the potential efficiency of sequencers based on the analysis of the miscall pattern that indicated two major sequence patterns that trigger the sequence-specific (SSE) [16] "during error". In general, "the error problem" is a wide field of further research.

5 SUPPLEMENTARY

The lab journal with the detailed pipeline, settings and details can be found at the link `https://docs.google.com/document/d/1aFFATF8HUAuPycAckbowRi5GceaK1PB00iu4R2tZSHM/edit?usp=sharing`.

References

[1] https://web.archive.org/web/20210808211206/https://vaccine-safety-training.org/subunit-vaccines.html

[2] Clem AS. Fundamentals of vaccine immunology. J Glob Infect Dis. 2011 Jan;3(1):73-8. doi: 10.4103/0974-777X.77299. PMID: 21572612; PMCID: PMC3068582.

[3] Zhang MM, Huang RY, Beno BR, Deyanova EG, Li J, Chen G, Gross ML. Epitope and Paratope Mapping of PD-1/Nivolumab by Mass Spectrometry-Based Hydrogen-Deuterium Exchange, Cross-linking, and Molecular Docking. Anal Chem. 2020 Jul 7;92(13):9086-9094. doi: 10.1021/acs.analchem.0c01291. Epub 2020 Jun 10. PMID: 32441507; PMCID: PMC7501946.

[4] Dinis JM, Florek KR, Fatola OO, Moncla LH, Mutschler JP, Charlier OK, Meece JK, Belongia EA, Friedrich TC. Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. J Virol. 2016 Jan 6;90(7):3355-65. doi: 10.1128/JVI.03248-15. Erratum in: J Virol. 2016 Aug 12;90(17):8029. PMID: 26739054; PMCID: PMC4794676.

[5] https://www.cdc.gov/

[6] Domingo E. Quasispecies and the development of new antiviral strategies. Prog Drug Res. 2003;60:133-58. doi: 10.1007/978-3-0348-8012-1_4. PMID: 12790341.

[7] Domingo E, Perales C. Viral quasispecies. PLoS Genet. 2019 Oct 17;15(10):e1008271. doi: 10.1371/journal.pgen.1008271. PMID: 31622336; PMCID: PMC6797082.

[8] Barbezange C, Jones L, Blanc H, Isakov O, Celniker G, Enouf V, Shomron N, Vignuzzi M, van der Werf S. Seasonal Genetic Drift of Human Influenza A Virus Quasispecies Revealed by Deep Sequencing. Front Microbiol. 2018 Oct 31;9:2596. doi: 10.3389/fmicb.2018.02596. PMID: 30429836; PMCID: PMC6220372.

[9] Ma, X., Shao, Y., Tian, L. et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 20, 50 (2019). https://doi.org/10.1186/s13059-019-1659-6

[10] Bertoni, M., Kiefer, F., Biasini, M. et al. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep 7, 10480 (2017). https://doi.org/10.1038/s41598-017-09654-8

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242. https://www.rcsb.org/

[12] McCrone JT, Lauring AS. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. J Virol. 2016 Jul 11;90(15):6884-95. doi: 10.1128/JVI.00667-16. PMID: 27194763; PMCID: PMC4944299.

[13] Van Poelvoorde LAE, Delcourt T, Coucke W, Herman P, De Keersmaecker SCJ, Saelens X, Roosens NHC and Vanneste K (2021) Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing. Front. Microbiol. 12:747458. doi: 10.3389/fmicb.2021.747458

[14] Verbist B, Clement L, Reumers J, Thys K, Vapirev A, Talloen W, Wetzels Y, Meys J, Aerssens J, Bijnens L, Thas O. ViVaMBC: estimating viral sequence variation in complex populations from illumina deep-sequencing data using model-based clustering. BMC Bioinformatics. 2015 Feb 22;16:59. doi: 10.1186/s12859-015-0458-7. PMID: 25887734; PMCID: PMC4369097.

[15] Ma, X., Shao, Y., Tian, L. et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 20, 50 (2019). https://doi.org/10.1186/s13059-019-1659-6

[16] Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011 Jul;39(13):e90. doi: 10.1093/nar/gkr344. Epub 2011 May 16. PMID: 21576222; PMCID: PMC3141275.

[17] Muñoz ET, Deem MW. Epitope analysis for influenza vaccine design. Vaccine. 2005 Jan 19;23(9):1144-8. doi: 10.1016/j.vaccine.2004.08.028. PMID: 15629357; PMCID: PMC4482133.

[18] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296-W303. doi: 10.1093/nar/gky427. PMID: 29788355; PMCID: PMC6030848. https://swissmodel.expasy.org/

[19] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242. https://www.rcsb.org/