Investigation of rare mutations that lead to epitope change in influenza quasispecies population

Pavlova N., Dudkovskaya A.

Abstract

Every autumn a seasonal campaign to prevent mass influenza virus infection starts. Each year people are vaccinated from several virus strains due to the high influenza mutation speed. However even vaccinated people can be infected with the flu if its strain is not covered by the vaccine. Another possible reason for that could be mutations in one or several epitopes that make virions "invisible" for the vaccine antibodies. In this research we analyzed the results of targeted single-end deep sequencing of viral probes from person that was not vaccinated yet infected their vaccinated friend with influenza. The main task of this research is to find common and rare SNPs in Influenza A virus (A/USA/RVD1_H3/2011(H3N2)) segment 4 hemagglutinin (HA) of the experimental probe and to define whether these mutations are in the epitope regions of HA. For this we set up a targeted deep sequencing experiment to analyze the HA genes from the second person viral sample. We found 5 common mutations and 2 rare mutations. All common SNPs and one of rare SNPs were synonymous. Another rare SNP (307 position, C -> T, frequency 0.94%) was a missense variant (Pro -> Ser) that belongs to the Epitope D segment of viral hemagglutinin. We conclude that due to the high mutation speed part of the viral population acquired the SNP that changed the Epitope D of viral hemagglutinin protein and thus became contageous even for the vaccinated person. The moral of the story is notorious: every non-vaccinated person gives a perfect opportunity for viruses to evolve and adapt to recent vaccines.

1 Introduction

Every year a global influenza vaccination campaign starts, because the main way to control and restrain influenza infection is vaccination [1]. However, the flu vaccine formula has to be renewed nearly every year to account for the ever-changing virus. To prevent most infections, each dose of seasonal influenza vaccine contains viruses or their HA proteins representing the influenza A(H3N2), A(H1N1) and influenza B strains thought to be most likely to circulate in the upcoming influenza season.

Influenza viruses contain eight single-stranded RNA segments encoding 11 proteins [2]. Hemagglutinin (HA) glycoprotein is the main determinant of virulence. Neuraminidase (HA) is a glycoprotein that prevents viral aggregation by binding the HA protein to other proteins. Due to the errors of polymerases, the genetics of influenza viruses changes and it adducts to gradual antigenic changes in both HA and NA. Because of these changes in influenza populations the virus become more virulent. The first problem is the antigenic drift [3], mutations that occur in the virus and change antigen proteins, so antibodies against the previously circulating subtype do not cross-react. The second one is viral quasispecies [4], a population structure with a set of several strains that interact genetically, live side by side and support each other.

The hemagglutination inhibition (HI) assay is used to titrate the antibody response to a viral infection. The HI assay takes advantage of some viruses' ability to hemagglutinate

(bind) red blood cells, therefore forming a "lattice" and preventing the red blood cells from clumping. If antibodies are present in the sera sample that cross-react with the virus, the antibodies will bind to the virus and prevent the virus from hemagglutinating the red blood cells [5].

We wanted to investigate the case, in which one person, who got vaccinated, was infected by the second person. First we use hemagglutination inhibition assay to find the virus profile. We know that the virus population closely matches the HI profile for an H3N2 strain called A/Hong Kong/4801/2014 (H3N2) and it is covered by first person vaccination. To figure out what is the cause of this infection, we performed a targeted deep sequencing experiment to analyze the HA genes in the second person viral sample with an Illumina single-end sequencing run. As a reference we use a HA gene of A/USA/RVD1_H3/2011(H3N2) strain. The gene contains 1653 nucleotides and the protein contains 550 acids. [6]

The ability of deep sequencing to detect low frequency variants is important in cases like this one, where we have viruses with antigenic drift and viral quasispecies [7], [8]. However it can be difficult to distinguish the rare variants and errors in next generation sequencing due to sequencing errors. To avoid false-positive detections, we explored the frequency of errors. We took the isogenic sample of the reference H3N2 influenza virus, used PCR to amplify the genetic material and put it into a plasmid of bacteria to subclone. Then sequenced the sample three times on the same Illumina machine to calculate the frequency of sequencing errors.

The unquestionable importance of research like this one is in understanding the current antigenic drift of influenza that could contribute to next year vaccine development.

2 Methods

Data acquisition

Reference gene of Influenza A virus (A/USA/RVD_H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene was downloaded fron NCBI database [9]. During sequencing library preparation cDNA was amplified using he Expand High Fidelity PCR System (Roche Applied Sciences, Indianapolis, IN), 40 cycles. Sequencing libraries were prepared with the Nextera XT Sample Prep kit and Nextera XT 24 index kit (Illumina, San Diego, CA), 12 cycles of amplification [10].
The target viral probe single-end Illumina reads were obtained from database [11].
Three control isogenic samples single-end Illumina reads of the reference H3N2 influenza virus were downloaded from database [12], [13], [14].

Readings quality check and alignment to reference

All raw data quality was checked with fastqc tool [15]. Since trimming of raw data with trimmomatic [16] with quality threshold 10 did not change the number of readings, we decided not to increase the quality threshold for the readings due to risks of missing rare SNPs.
Indexation of reference genome and readings to reference genome alignment were performed using bwa tool [17]. The .sam files were compressed to .bam format and then indexed with samtools utilite [18].

Variant calling

Mpileup file was made with samtools utilite [18]. For all readings be counted in mpileup we set the depth limit (-d flag) to the number of reads in the file. Number of reads in target and three control files are in the Table 1.
Variant calling of the target probe data was performed with VarScan.v2.3.4 utilite [19]. For common variants we set the frequency rate to 95%. The results are in the Table 3).
To determine the sequencing mistakes threshold we made variant calling of the control probes data with frequency rate 0.1% and analyzed the means and standard deviation of all variants. The results are in the Table 4).
To find rare SNPs in target probe we set the frequency rate for variant calling to the highest of three control probes data mean frequency + 3 standard deviations (0.507%) so that all variants found would not be the sequencing mistakes. The results are in the Table 5.

Variant analysis

Variants found were visualized in IGV browser [20].
3D-structure of reference and sequenced H3N2 hemagglutinin was reconstructed with usage of SwissModel resource [21] and PDB database rcsb.org [22].

3 Results

Readings general information

The length of the reference gene is 1653 bp. Number of reads, average read length and the percentage of mapped reads to the reference for target sample and three control samples reads are presented in the Table 1.

Table 1: Number of reads and average read lengths in target sample and control samples.

File type	Number of reads	Average read length, bp	% mapped readings
Target sample SRR1705851	358265	147.148	99.94
Control sample SRR1705858	256586	148.561	99.97
Control sample SRR1705859	233327	148.446	99.97
Control sample SRR1705860	249964	148.703	99.97

The quality of raw target and control samples are presented in the Table 2.

Table 2: Redings quality characteristics of target sample (TS) and control samples (CS) readings

Quality characteristic	TS 51	CS 58	CS 59	CS 60
Basic Statistics	normal	normal	normal	normal
Per base seq quality	normal	normal	normal	normal
Per seq quality scores	normal	normal	normal	normal
Per base seq content	very unusual	very unusual	very unusual	very unusual
Per seq GC content	abnormal	very unusual	very unusual	very unusual
Per base N content	normal	normal	normal	normal
Seq Length Distrib	abnormal	abnormal	abnormal	abnormal
Seq Duplic Levels	very unusual	very unusual	very unusual	very unusual
Overrepresented seq	abnormal	very unusual	very unusual	very unusual
Adapter Content	normal	normal	normal	normal

Variant information

Common variants of target probe are presented in Table 3. All of the frequent variants are synonymous.

Table 3: Identified common SNPs in target probe

Position	SNP	Codone	Aminoacid	Type	Frequency
72	A -> G	ACA -> ACG	Thr -> Thr	synonymous	99.96%
117	C -> T	GCC -> GCT	Ala -> Ala	synonymous	99.82%
774	T -> C	TTT -> TTC	Phe -> Phe	synonymous	99.96%
999	C -> T	GGC -> GGT	Gly -> Gly	synonymous	99.86%
1260	A -> C	CTA -> CTC	Leu -> Leu	synonymous	99.94%

Means, standard deviations and means + 3 sd of frequencies of variants caused by sequencing mistakes in control probes are presented in Table 4. All three control probes had 35 shared variants in positions 44, 165, 183, 216, 218, 222, 235, 254, 276, 340, 356, 370, 409, 414, 421, 463, 660, 670, 691, 722, 744, 859, 915, 987, 1031, 1056, 1086, 1213, 1264, 1280, 1358, 1398, 1421, 1460, 1482.

Table 4: Means, standard deviations of SNPs frequencies of control probes

Name of probe	Num of variants	Mean freq	Standard deviation	Mean + 3 sd
Control sample SRR1705858	58	0.2619	0.0814	0.5061
Control sample SRR1705859	54	0.2383	0.0529	0.397
Control sample SRR1705860	61	0.2503	0.0774	0.4825

Rare variants with frequency above the threshold of sequencing mistakes (0.507%) are presented in Table 5.

Only one rare SNP that was found leads to aminoacid change (103Pro -> 103Ser)

Table 5: Identified rare SNPs in target probe

Position	SNP	Codone	Aminoacid	Type	Frequency
307	C -> T	CCG -> TCG	Pro -> Ser	missence	0.94%
1458	T -> C	TAT -> TAC	Tyr -> Tyr	synonymous	0.84%

Influenza H3N2 hemagglutinin protein structure visualization

Structure of the whole H3N2 hemagglutinin protein with 103Pro (reference H3N2 hemagglutinin) is presented in the Figure 1.

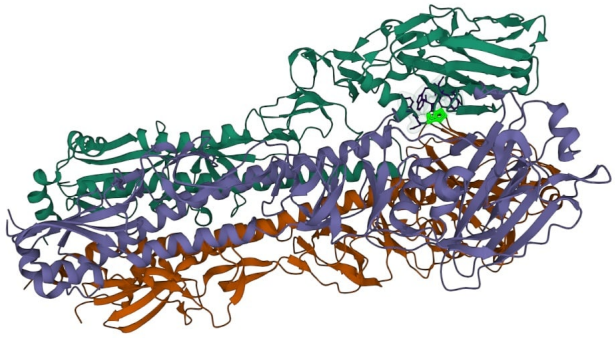

Figure 1. Whole H3N2 hemagglutinin 103Pro structure.

Structure of the whole H3N2 hemagglutinin protein with 103Ser (target probe H3N2 hemagglutinin) is presented in the Figure 2.

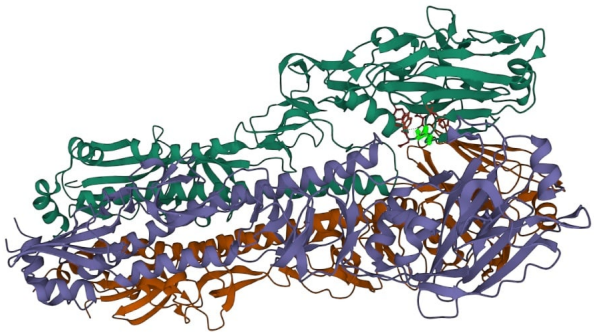

Figure 2. Whole H3N2 hemagglutinin 103Ser structure.

Structure of the partial Epitope D H3N2 hemagglutinin protein with 103Pro (target probe H3N2 hemagglutinin) is presented in the Figure 3.

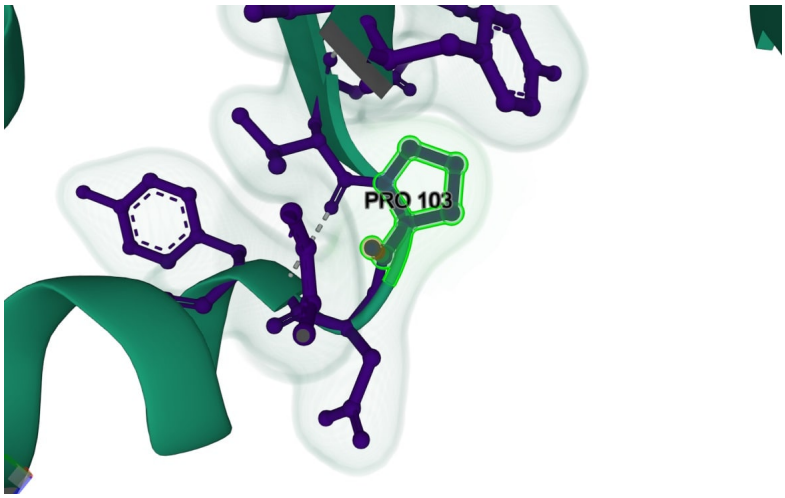

Figure 3. Partial H3N2 hemagglutinin Epitope D 103 Pro structure.

Structure of the partial Epitope D H3N2 hemagglutinin protein with 103Ser (target probe H3N2 hemagglutinin) is presented in the Figure 4.

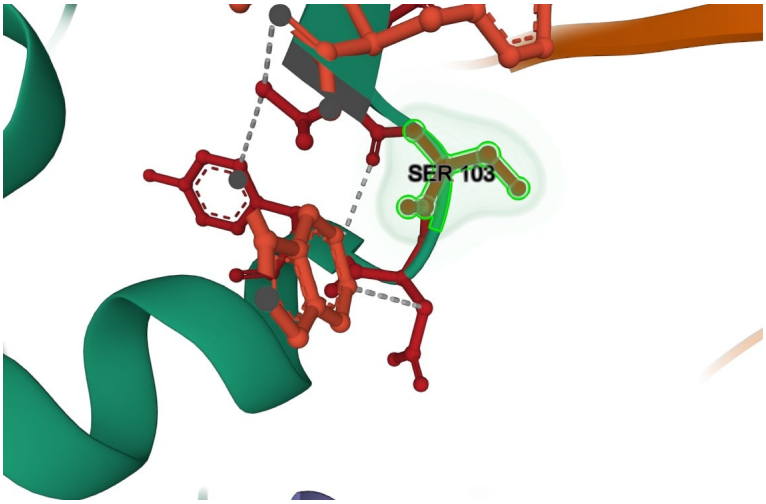

Figure 4. Partial H3N2 hemagglutinin Epitope D 103 Ser structure.

4 Discussion

Most vaccines nowadays contain only antigens of Influenza. Majority of them contains hemagglutinin proteins of different Influenza strains, because hemagglutinin is one of two proteins in the lipid bilayer that surrounds the viral capsid and it is enough for antibodies to recognize and bind to the virion. Two-component vaccines also contain neuraminidase protein which is also exposed to the virion surface. Antibodies recognize the protein with

high affinity and specificity. Protein sites that antibodies bind to are called epitopes and even a small change in their primary, secondary or tertiary structure affects their recongition by antibodies. HA protein in H3N2 type of influenza has 5 epitopes (A-E) [23]. However, if a mutation occurs in these regions, antibodies no longer recognize the protein, because its epitopes perhaps change their structure.

Influenza RNA-dependent RNA polymerase is characterized by low fidelity and lacks the 3' to 5' exonuclease proofreading capability which leads to rapid mutation rate. If the mutation is non-deleterious and changes the viral epitope it would be conserved and as far as virions containing this mutation would be not recognized by antibodies thus advantageous, positively selected and amplified in the viral population [24].

This is one of the main purposes of changing the compound of the vaccine every year. Because of the quite fast mutation process viral population of the infected person is genetically heterogenous and is called quasispecies. Therefore, vaccines that consist of only one Influenza strain hemagglitinin may not work and lead to infection even if a person got vaccinated. So, contemporary vaccines contain antigens of several most represented strains which change from year to year. For understanding which are the next year Influenza strain that needs to be added to future vaccine it is crucial to detect which new epitopic variants it has obtained. Because of the effect of quasispecies and genetic drift, the percentage of these mutations can be very small. Deep sequencing is usually used for detection of such rare variants.

NGS methods present challenge for detecting rare variants due to high error rate of NGS (0.1-1%). NGS errors may occur during library preparation (in that case they would be called as rare variants in all control isogenic probes) or during the sequencing process (SPNs would be unique in each isogenic control) [25]. There are several methods to decrease or calculate NGS errors. First of them would be using different barcodes for different DNA molecules during the sequencing process and requires the depth of coverage at least 10x for enough replication clusters number which leads to increase of sequencing cost. Another approach is increasing threshold of quality score of readings that helps to discard readings where sequencing error occured in the middle or in the end of sequencing process. Third method is counting the sequencing errors frequency in isogenic probes and calculating the threshold value for rare variants detection. In current research we used and described that approach. To lower the error rate during library preparation process high-fidelity polymerases could be used. If during NGS some specific regions are found to have a lot of rare variants it is reasonable to perform further Sanger sequencing of these regions in several repetitions [26].

In our case we have found one missence SNP in the 307 nucleotide that changes the 103 aminoacid from Proline to Serine. The 103 position of the protein lies in the Epitope D [23]. It is easy to select viral mutants that escape most antibodies [27] and also due to the observation that antigenicity-altering mutations frequently occur during the evolution of influenza worldwide [28] and in people with long-term infections [29]. Thus we can conclude that the first person got the flu because of the mutation in the Epitope D.

If this new variant would be detected in enough vaccinated yet infected people it would be reasonable to consider this protein sequence as one of candidates for including to the

vaccine of next year.

5 Optional Extra-Credit Challenge Question

Actual average coverage.
We can count the actual average coverage per position for data set of reads using this formula: $\frac{\text{sum of lengths of reads}}{\text{length of reference}}$. The length of the reference genome is 1653. Here we present the table of average coverage of each data set (see Table 6).

Table 6: Actual average coverage

Name of probe	Sum of lengths	Average coverage
Target sample SRR1705851	52717864	31892
Control sample SRR1705858	38118673	23060
Control sample SRR1705859	34636460	20953
Control sample SRR1705860	37170397	22486

The code:
awk NR%4==2sum+=length($0)ENDprint sum' SRR1705851.fastq -> 52717864
awk NR%4==2sum+=length($0)ENDprint sum' SRR1705858.fastq -> 38118673
awk 'NR%4==2sum+=length($0)ENDprint sum' SRR1705859.fastq -> 34636460
awk 'NR%4==2sum+=length($0)ENDprint sum' SRR1705860.fastq -> 37170397

Strains of Influenza if current vaccine
We can find information in Rospotrebnadzor cite [30] - "A/H3N2 (A/ Darwin/9/2021), Victoria B/Austria/1359417/2021, A(H1N1) pdm09 (A/Victoria/2570/2019)".

A/Hong Kong/4801/2014 vaccine strain was in the 2016.

PCR ("upstream") and sequencing ("during") errors.
In three control isogenic probes we have found 58, 54 and 61 error variants respectively. 35 variants amongst them were supported by all three isogenic probes, thus we consider them as "upstream" errors that occured during library preparation (see Table 7).

Table 7: Means and standard deviations of PCR and sequencing errors frequencies in isogenic probes

Name of probe	Mean PCR error	std PCR error	Mean seq error	std seq error
SRR1705858	0.271	0.099	0.255	0.066
SRR1705859	0.231	0.051	0.243	0.054
SRR1705860	0.239	0.097	0.259	0.058

As we see, in our case frequency of error SNPs "upstream" and "during" are comparable and we can not assume they differ significantly (for each isogenic probe confidence interval of PCR and sequencing errors intersect). This fits into the frame of common sense because both PCR and sequencing errors affect only one cluster.

Labjournal.
Our labjournal labjournal.md and two pbd files (pro.pdb for reference HA; ser.pdb for HA with detected rare SNP) could be found in repository https://github.com/pavlovanadia/BI_2022_project_2. stats.ipynb provides code we used to count statistics on our data.

References

[1] **Nichol KL, Treanor JJ.** Vaccines for seasonal and pandemic influenza. J Infect Dis. 2006 Nov 1;194 Suppl 2:S111-8. doi: 10.1086/507544. PMID: 17163383.

[2] **Bouvier NM, Palese P.** The biology of influenza viruses. Vaccine. 2008 Sep 12;26 Suppl 4(Suppl 4):D49-53. doi: 10.1016/j.vaccine.2008.07.039. PMID: 19230160; PMCID: PMC3074182.

[3] **Doherty, P., Turner, S., Webby, R. et al.** Influenza and the challenge for immunology. Nat Immunol 7, 449–455 (2006). https://doi.org/10.1038/ni1343

[4] **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.

[5] **Bourgeois, M. A., Oaks, J. L.** (2014). Laboratory diagnosis of viral infections. In Equine Infectious Diseases (pp. 132-140). WB Saunders.

[6] https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=genbank

[7] **McElroy K, Thomas T, Luciani F.** Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb Inform Exp. 2014 Jan 15;4(1):1. doi: 10.1186/2042-5783-4-1. PMID: 24428920; PMCID: PMC3902414.

[8] **Sobel Leonard A, McClain MT, Smith GJ, Wentworth DE, Halpin RA, Lin X, Ransier A, Stockwell TB, Das SR, Gilbert AS, Lambkin-Williams R, Ginsburg GS, Woods CW, Koelle K.** Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic Diversification. J Virol. 2016 Nov 28;90(24):11247-11258. doi: 10.1128/JVI.01657-16. PMID: 27707932; PMCID: PMC5126380.

[9] https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta

[10] **Cushing, A., Kamali, A., Winters, M., Hopmans, E. S., Bell, J. M., Grimes, S. M., Xia, L. C., Zhang, N. R., Moss, R. B., Holodniy, M., Ji, H. P.**, (2015) Emergence of Hemagglutinin Mutations During the Course of Influenza Infection. *Scientific reports 5*, 16178.

[11] http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/

[12] ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz

[13] ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz

[14] ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

[15] **Wingett SW, Andrews S.** (2018) FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res. 24*, 1338.

[16] **Bolger AM, Lohse M, Usadel B.** (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics 1*, 2114–2120.

[17] **Li, Heng and Durbin, Richard** (2009) Fast and accurate short read alignment with Burrows–Wheeler transform *bioinformatics 25*(14), 1754–1760.

[18] **Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li** (2021) Twelve years of SAMtools and BCFtools *GigaScience 10*(2), 2047–217X.

[19] **Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L.** (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics 25*(17), 2283–2285.

[20] **Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP**, (2011) Integrative genomics viewer. *Nat Biotechnol. 29*(1), 24–6.

[21] **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.**, (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res. 2*(46), 296–303.

[22] **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.

[23] **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.

[24] **Wang, Y., Tang, C. Y., Wan, X. F.**, (2022) Antigenic characterization of influenza and SARS-CoV-2 viruses. *Analytical and bioanalytical chemistry 414*(9), 2841–2881.

[25] **Yeom, H., Lee, Y., Ryu, T., Noh, J., Lee, A. C., Lee, H. B., Kang, E., Song, S. W., Kwon, S.**, (2019) Barcode-free next-generation sequencing error validation for ultra-rare variant detection. *Nature communications 10*(1), 977.

[26] **Savage SA, Jones K, Teshome K, Lori A, McReynolds LJ, Niewisch MR.**, (2022) Next-generation sequencing errors due to genetic variation in WRAP53 encoding TCAB1 on chromosome 17. *Hum Mutat. online publication.*

[27] **R.G. Webster, W.G. Laver**| Determination of the number of nonoverlapping antigenic areas on Hong Kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance, Virology, Volume 104, Issue 1, 1980, Pages 139-148, ISSN 0042-6822, https://doi.org/10.1016/0042-6822(80)90372-4.

[28] **Benjamin S. Chambers, Kaela Parkhouse, Ted M. Ross, Kevin Alby, Scott E. Hensley**, Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014–2015 Influenza Season, Cell Reports, Volume 12, Issue 1, 2015, Pages 1-6, ISSN 2211-1247, https://doi.org/10.1016/j.celrep.2015.06.005.

[29] **Katherine S Xue, Terry Stevens-Ayers, Angela P Campbell, Janet A Englund, Steven A Pergam, Michael Boeckh, Jesse D Bloom**, (2017) Parallel evolution of influenza across multiple spatiotemporal scales eLife 6:e26875.

[30] https://www.rospotrebnadzor.ru/about/info/news/news_details.php?ELEMENT_ID=22630#:~:text=12.09.2022%20%D0%B3.,%D1%87%D1%82%D0%BE%D0%B1%D1%8B%20%D1%83%D1%81%D0%BF%D0%B5%D0%BB%20%D1%81%D1%84%D0%BE%D1%80%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D1%82%D1%8C%D1%81%D1%8F%20%D0%BD%D0%B5%D0%BE%D0%B1%D1%85%D0%BE%D0%B4%D0%B8%D0%BC%D1%8B%D0%B9%20%D0%B8%D0%BC%D0%BC%D1%83%D0%BD%D0%B8%D1%82%D0%B5%D1%82.