Breakthrough flu flew through my door

Zhandos Ayupov, Polina Guseva

November 12, 2022

Abstract

Hitherto flu significantly affects the human population. Yet general vaccination lowers noticeably the possibility of negative effects. Breakthrough influenza caught from our roommate has been recognised as the A/Hong Kong/4801/2014 (H3N2) substrain. In this paper, deep-sequenced reads are aligned to the hemagglutinin gene of the A/USA/RVD1 H3/2011 (H3N2) reference together with three controls from isogenic viruses. The only missense mutation is the P103S substitution in the epitope D causing slight conformation changes and possible affinity alteration. Probably, it appears as a mix of viral quasi-species leading to antigenic shift and less vaccine efficiency.

Introduction

Annually seasonal flu causes around half a million deaths worldwide [1], leads to a loss of 3.7-5.9 working days per episode [2], and a fear of a pandemic [3]. While vaccination decreases complications in general up to 60% [4], [5].

Influenza virus has several types with dozen of substrains each [6]. By recombining their genetic material, antigenic shift [7] rarely produces new variants via viral quasispecies [8] as the virus picks itself up in pieces. Whereas, point mutations cause antigenic drift [9] as influenza's RNA polymerase errs 1 time/genome/replication [10].

One of the main targets for host antibodies is hemagglutinin (H or HA) [11] as this protein covers a viral envelope [12]. With evolution pressure on it, many modifications occur in its structure with a rate of ~ 5.7 nucleotide substitutions/year (HA1 in A(H3N2) [13] with a general rate of ~ 2 nucleotide substitutions/year (for influenza A) [14]. Therefore, it is crucial to keep vaccine composition up to date [15]. For season 2022-2023 WHO recommends vaccines including: an A/Victoria/2570/2019 or an A/Wisconsin/588/2019 (H1N1)pdm09-like virus; an A/Darwin/6 or 9/2021 (H3N2)-like virus; a B/Austria/1359417/2021 (B/Victoria lineage)-like virus; and a B/Phuket/3073/2013 (B/Yamagata lineage)-like virus [16].

To accurately detect rare mutations leading to mixed viral populations, these modifications should be redundantly supported by a lot of reads. This assumption is called deep sequencing (RNASeq for RNA) [17]. Moreover, a frequency of rare (0.1% of bases sequenced) substitutions could be the same as the rate of NGS errors [18], [19]. This happens due to signal misinterpretation during NGS or integrating of a wrong nucleotide during PCR (or previously as library preparation or storing) [18].

Our hypothesis is that HA of our roommate's infectant mutated because of antigenic shift and/or antigenic shift and led to epitope transformations so that antibodies formed after vaccinations did not recognise this virus anymore. In this paper, we examine: 1) aligning of reads from hemagglutinin of our roommate's virus to the reference HA gene from the H3N2 strain, 2) locations of obtained mutations with sufficient frequency, 3) and their changes to the structure of epitopes.

Methods

During this project, we worked with our roommate's flu [20] (hereinafter referred to as the test) which was previously identified as similar to Influenza A virus the A/Hong Kong/4801/2014 (H3N2) strain. As for the reference sequence, the hemagglutinin gene of the A/USA/RVD1_H3/2011(H3N2) strain is chosen (GenBank # KF848938.1). Also, to eliminate PCR and sequence errors, three control sequencing of an isogenic referral virus were executed [21], [22], [23].

Table 1: Description of the filtered single nucleotide polymorphisms (SNPs) in the roommate's sequence

Location, bp	72	117	307	774	999	1260	1458
Frequency, %	99.96	99.82	0.94	99.96	99.86	99.94	0.84
Codon change	AC**A** → AC**G**	GC**C** → GC**T**	**C**CG → **T**CG	TT**T** → TT**C**	GG**C** → GG**T**	CT**A** → CT**C**	TA**T** → TA**C**
Amino acid change	Thr → Thr	Ala → Ala	Pro → Ser	Phe → Phe	Gly → Gly	Leu → Leu	Tyr → Tyr
Mutation	synonymous	synonymous	missense	synonymous	synonymous	synonymous	synonymous

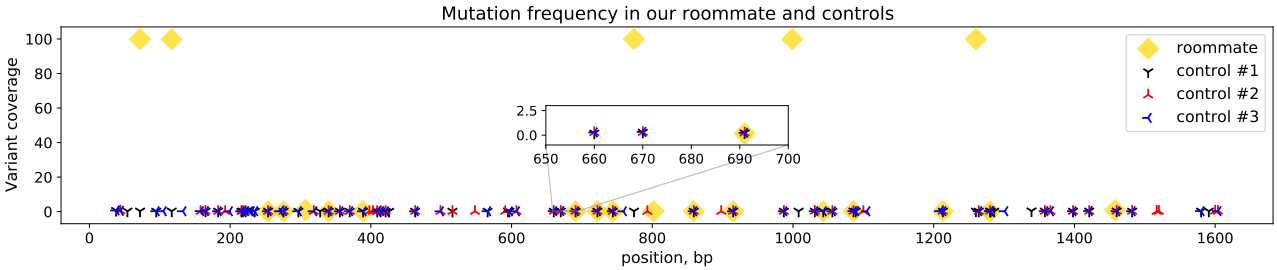

Figure 1: The frequency (%) of SNPs in the our roommate's sequence in comparison to controls

Both the test and all controls were aligned to the reference gene and filtered; proven variants are linked to protein changes. Firstly, we performed quality control of our test influenza with Fast Quality Control (FastQC, `fastqc`) [24]. After that, we aligned our initial reads with the indexed (`bwa index`) reference sequence of hemagglutinin with Burrows-Wheeler Aligner (BWA, particularly used algorithm `bwa mem`) [25]. Then we transformed the obtained *.sam format file of alignment to *.bam format (`samtools view -b`), sorted (`samtools sort`) and indexed (`samtools index`) that file with samtools [26]. For the following steps, variants within the whole depth were pilled up (`samtools mpileup -f`). Later we used VarScan [27] to look for common (> 95% frequency) and rare (> 0.1% frequency) single nucleotide polymorphisms (SNPs) in the aligned sequences (`varscan mpileup2snp`). We used this command without a strand filter. Then we repeated the same processes for three control examples. Based on the broadest 99,7% confidence interval, SNPs from the test were filtered in comparison to the controls. If the opposite is not stated, the default settings are used.

Then we manually analyzed discovered mutations from previous steps to identify any changes in amino acids. Finally, we visualized the protein sequence for the reference sequence and the test by SWISS-MODEL [28] and PDB MolStar Viewer [29] to compare the difference.

Results

The total sequence length of reads is 52 717 864 bp. Therefore, the coverage for the 1 665 bp reference hemagglutinin gene is $\sim 3.2 \times 10^4$ reads/nucleotide. Generally, the quality of all reads is above 30 on the Phred-33 scale. Despite the warning about duplication levels by the FastQC report [24], it could be ignored due to the very deep coverage by reads (fig.**??**). Per base sequence content is abnormal due to unequal random primer hybridisation throughout RNA library preparation. All three controls have coverage of 28-31 reads/nucleotide. Both test and control reads were mapped without a significant loss (less than 0.1%).

There are 7 piled-up SNPs (tab.1) filtered based on the noise distribution from controls (fig.1), see the rest of the results (tab.2). The most radical 99.7% noise confidence interval (tab.3) suggests neglecting all substitutions with a frequency below 0,48%. Only one of the found mutations is missense affecting the 103rd amino acid (from proline to serine) in the hemagglutinin protein, whereas others are synonymous. Two SNPs have low occurrence rates, while ones of others are above 99%.

Based on the visualisation (fig.2A), there are no radical changes in the protein conformation. However, the substitution by serine relaxes a linker between two beta-sheets (fig.2B-C).

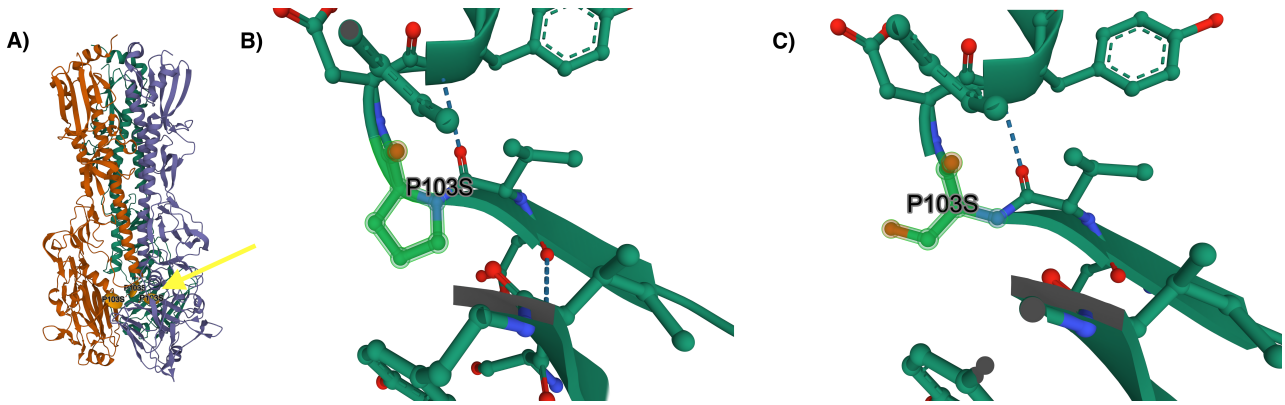

Figure 2: The 3d visualization of hemagglutinin: A) general view with P103S pointed, B) zoomed of the 103rd amino acid as in the reference (proline), C) proximity site of the same area for the test (as serine).

Discussion

While most of the found SNPs are identified as noise or synonymous, one missense mutation is detected. It affects the epitope D in hemagglutinin 3 of influenza A [30] as P103S which is already known previously in the 3C.2a2 group [31], [32], [33]. As proline ensures a rigid protein secondary structure, its substitution by serine could relax the epitope conformation (fig.2B-C). Furthermore, serine is more polar which could affect antibody affinity.

Antigenicly shifted variants could drop the effectiveness of immunizations up to 1% [34]. Since it happens due to the mixing of different variants [7], this viral quasispecies [8] occurs rarely as the event is unlikely [35]. Therefore, the found SNPs with low frequency (P103S, Y1458Y) represent that mechanism. As others almost totally cover this substrain, they substituted too long ago and probably already do not cause new antigenic changes.

Highly likely, the real mutations are above the noise signal. As sequencing is just a reading of a read, it produces random noise [18]. Yet, the sequence of the same nucleotide could lead to a signal merge. Whereas the mutations at the same site are caused by PCR (not deep enough coverage) or library preparations [36], [19]. To decrease this noise, the elaborated software could be used to implement k-mer counting or the Full-text index in Minute space [37]. Also, more careful handling will improve the quality as well as cutting-edge sequencing technologies.

According to the correlation for flu provided [30], the seasonal severity should increase in $\sim$1.6-1.7$\times$ because of this mutation. However, since a comparison with the hemagglutinins from the recommended vaccines is not provided, this estimation could only be approximate. Among current variants [16], vaccination has the lowest effectiveness against H3N2: it decreases the likelihood to get flu only by 27% with the current season shot [5].

Moreover, vaccination is not a silver bullet against the disease but a way to prevent severe course. From circumstantial evidence, a person with a breakthrough infection was able to "start analyzing data **right away**" despite some under-the-weather symptoms. On these grounds, it is likely to conduct that the flu course was less harsh than it could be without vaccination.

References

[1] J. Paget, P. Spreeuwenberg, V. Charu, R. J. Taylor, A. D. Iuliano, J. Bresee, L. Simonsen, C. Viboud, *et al.*, "Global mortality associated with seasonal influenza epidemics: New burden estimates and predictors from the GLaMOR Project," *Journal of Global Health*, vol. 9, no. 2, 2019.

[2] M. Keech and P. Beardsworth, "The impact of influenza on working days lost," *Pharmacoeconomics*, vol. 26, no. 11, pp. 911–924, 2008.

[3] Y. Guan, D. Vijaykrishna, J. Bahl, H. Zhu, J. Wang, and G. J. Smith, "The emergence of pandemic influenza viruses," *Protein & Cell*, vol. 1, no. 1, pp. 9–13, 2010.

[4] E. A. Belongia, M. D. Simpson, J. P. King, M. E. Sundaram, N. S. Kelley, M. T. Osterholm, and H. Q. McLean, "Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies," *The Lancet Infectious Diseases*, vol. 16, no. 8, pp. 942–951, 2016.

[5] S. S. Kim, B. Flannery, I. M. Foppa, J. R. Chung, M. P. Nowalk, R. K. Zimmerman, M. Gaglani, A. S. Monto, E. T. Martin, E. A. Belongia, *et al.*, "Effects of prior season vaccination on current season vaccine effectiveness in the United States Flu Vaccine Effectiveness Network, 2012–2013 through 2017–2018," *Clinical Infectious Diseases*, vol. 73, no. 3, pp. 497–505, 2021.

[6] F. M. Burnet, "A genetic approach to variation in influenza viruses; the characters of three substrains of influenza virus A (WS).," *Journal of General Microbiology*, vol. 5, no. 1, pp. 46–53, 1951.

[7] D. J. Earn, J. Dushoff, and S. A. Levin, "Ecology and evolution of the flu," *Trends in Ecology & Evolution*, vol. 17, no. 7, pp. 334–340, 2002.

[8] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS Pathogens*, vol. 6, no. 7, p. e1001005, 2010.

[9] W. Shao, X. Li, M. U. Goraya, S. Wang, and J.-L. Chen, "Evolution of influenza a virus by mutation and re-assortment," *International Journal of Molecular Sciences*, vol. 18, no. 8, p. 1650, 2017.

[10] J. W. Drake, "Rates of spontaneous mutation among RNA viruses.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 9, pp. 4171–4175, 1993.

[11] N. Green, H. Alexander, A. Olson, S. Alexander, T. M. Shinnick, J. G. Sutcliffe, and R. A. Lerner, "Immunogenic structure of the influenza virus hemagglutinin," *Cell*, vol. 28, no. 3, pp. 477–487, 1982.

[12] I. T. Schulze, "Structure of the influenza virion,"

[13] W. M. Fitch, R. M. Bush, C. A. Bender, and N. J. Cox, "Long term trends in the evolution of H (3) HA1 human influenza type A," *Proceedings of the National Academy of Sciences*, vol. 94, no. 15, pp. 7712–7718, 1997.

[14] E. Nobusawa and K. Sato, "Comparison of the mutation rates of human influenza A and B viruses," *Journal of Virology*, vol. 80, no. 7, pp. 3675–3678, 2006.

[15] F. Carrat and A. Flahault, "Influenza vaccine: the challenge of antigenic drift," *Vaccine*, vol. 25, no. 39-40, pp. 6852–6862, 2007.

[16] World Health Organization = Organisation mondiale de la Santé, "Recommendations announced for influenza vaccine composition for the 2022-2023 northern hemisphere influenza season," *Weekly Epidemiological Record (WER) = Relevé épidémiologique hebdomadaire*, vol. 97, no. 12, pp. 109–132, 2022.

[17] D. Goldman and K. Domschke, "Making sense of deep sequencing," *International Journal of Neuropsychopharmacology*, vol. 17, pp. 1717–1725, 06 2014.

[18] X. Ma, Y. Shao, L. Tian, D. A. Flasch, H. L. Mulder, M. N. Edmonson, Y. Liu, X. Chen, S. Newman, J. Nakitandwe, *et al.*, "Analysis of error profiles in deep next-generation sequencing data," *Genome biology*, vol. 20, no. 1, pp. 1–15, 2019.

[19] K. Mitchell, J. J. Brito, I. Mandric, Q. Wu, S. Knyazev, S. Chang, L. S. Martin, A. Karlsberg, E. Gerasimov, R. Littman, *et al.*, "Benchmarking of computational error-correction methods for next-generation sequencing data," *Genome biology*, vol. 21, no. 1, pp. 1–13, 2020.

[20] NCBI Sequence Read Archive (SRA), "SRR1705851," 8 2015.

[21] NCBI Sequence Read Archive (SRA), "SRR1705858," 8 2015.

[22] NCBI Sequence Read Archive (SRA), "SRR1705859," 8 2015.

[23] NCBI Sequence Read Archive (SRA), "SRR1705860," 8 2015.

[24] S. Andrews, "FASTQC. A quality control tool for high throughput sequence data," 2010.

[25] H. Li and R. Durbin., "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics (Oxford, England)*, vol. 25,14, 2009.

[26] R. H. e. a. Ramirez-Gonzalez, "Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments.," *Source code for biology and medicine*, vol. 7, 2012.

[27] D. C. e. a. Koboldt, "VarScan: variant detection in massively parallel sequencing of individual and pooled samples.," *Bioinformatics (Oxford, England)*, vol. 25, 2009.

[28] A. e. a. Waterhouse, "SWISS-MODEL: homology modelling of protein structures and complexes.," *Nucleic acids research*, vol. 46, 2018.

[29] S. e. a. Bittrich, "RCSB Protein Data Bank: Improved Annotation, Search, and Visualization of Membrane Protein Structures Archived in the PDB.," *Bioinformatics (Oxford, England)*, vol. 38, 2021.

[30] E. T. Munoz and M. W. Deem, "Epitope analysis for influenza vaccine design," *Vaccine*, vol. 23, no. 9, pp. 1144–1148, 2005.

[31] A. Cushing, A. Kamali, M. Winters, E. S. Hopmans, J. M. Bell, S. M. Grimes, L. C. Xia, N. R. Zhang, R. B. Moss, M. Holodniy, *et al.*, "Emergence of hemagglutinin mutations during the course of influenza infection," *Scientific reports*, vol. 5, no. 1, pp. 1–12, 2015.

[32] "DoD Global, Laboratory-Based, Influenza Surveillance Program, Respiratory Highlights, Surveillance Weeks 36 - 39," tech. rep., USAF School of Aerospace Medicine Defense Health Agency, 2017.

[33] S. Buda, K. Prahm, R. Dürrwald, B. Biere, J. Schilling, U. Buchholz, W. Haas, *et al.*, "Bericht zur Epidemiologie der Influenza in Deutschland Saison 2017/18," 2018.

[34] B. Flannery, R. K. Zimmerman, L. V. Gubareva, R. J. Garten, J. R. Chung, M. P. Nowalk, M. L. Jackson, L. A. Jackson, A. S. Monto, S. E. Ohmit, *et al.*, "Enhanced genetic characterization of influenza A (H3N2) viruses and vaccine effectiveness by genetic group, 2014–2015," *The Journal of Infectious Diseases*, vol. 214, no. 7, pp. 1010–1019, 2016.

[35] Y. Wang, C. Y. Tang, and X.-F. Wan, "Antigenic characterization of influenza and SARS-CoV-2 viruses," *Analytical and Bioanalytical Chemistry*, pp. 1–41, 2021.

[36] W. L. Matochko and R. Derda, "Error analysis of deep sequencing of phage libraries: peptides censored in sequencing," *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013.

[37] D. Laehnemann, A. Borkhardt, and A. C. McHardy, "Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 154–179, 2016.

Supplementary materials

Table 2: Description of the noise-level SNPs filtered-out from the roommate's sequence

Location, bp	254	276	340	389	691	722	744	802	859	915	1043	1086	1213	1280
Frequency, %	0.17	0.17	0.17	0.22	0.17	0.2	0.17	0.23	0.18	0.19	0.18	0.21	0.22	0.18
Nucleotide change	A→G	A→G	T→C	T→C	A→G	A→G	A→G	A→G	A→G	T→C	A→G	A→G	A→G	T→C

Table 3: Confidence intervals for controls with different acceptance levels, %

Confidence interval	**Control #1**	**Control #2**	**Control #3**	**Combined**
68%	(0.185, 0.328)	(0.185, 0.289)	(0.172, 0.328)	(0.180, 0.317)
95%	(0.113, 0.400)	(0.132, 0.342)	(0.094, 0.406)	(0.111, 0.386)
99.7%	(0.041, 0.472)	(0.080, 0.394)	(0.016, 0.484)	(0.042, 0.455)