BIOINFORMATICS INSTITUTE

Low-frequency mutation in the hemagglutinin gene of the influenza virus H3N2 alter the epitope D and may account for cases of vaccination ineffectiveness

November 10, 2023

Homework number 2

by Kirill Petrikov and Ilia Popov

Abstract

The high mutation rate of the influenza virus is one of the key reasons, not only requires constant vaccine updates, but the newest vaccines may be ineffective. This report describes a case of influenza virus infection, strain H3N2 covered by the vaccine, in a vaccinated individual. Haemagglutinin gene deep sequencing data obtained from suspected source of infection and from three isogenic controls were examined. Five high-frequency (>98%) SNPs, all silent, and 16 low-frequency (<1%) SNPs were found in the patient sample. A mean sequencing error rate of 0.25 ± 0.07% was calculated for three controls. This allowed us to separate two true mutations with frequencies of 0.84% and 0.94% from the erroneous ones (<0.25%). The first is silent, but the second results in a P103S substitution in epitope D of the haemagglutinin. This may account for the vaccination ineffectiveness.

Introduction

Influenza virus is among the most significant pathogens, leading the way in mortality from infectious diseases and results in an estimated 250,000 to 500,000 deaths every year [1,2].

Vaccination is an effective way of disease control [3,4]. Because the virus is rapidly mutating, vaccines need to be constantly updated to maintain effective protection.

When viral RNA is transcribed, error-prone polymerase provides genetic changes that result in new variant strains, a process known as antigenic drift [5]. The main mutation target is the surface hemagglutinin glycoprotein (HA). It binds to sialic acid on surface glycoproteins and glycolipid, allowing effective contact with cells [6]. HA is the primary target of antibodies that provide protective immunity to influenza viruses. It contains several epitopes that serve as targets for the development of modern vaccines. Thus, mutations in HA allow the virus to avoid the immune response of the infected organism [7].

Such genetic features ensure the formation of so-called subpopulations: diverse subsets of viral particles of the same strain that possess a distinct genotype and phenotype [8]. Because the frequency of occurrence of a genotype in a particular population can be very low, deep sequencing methods are required to detect them, allowing for sequencing errors with conventional coverage [9]. This approach requires distinguishing between low frequency biological variants and sequencing errors. This can be achieved by assessing the occurrence of sequencing errors in isogenic samples.

The aim of this work was to detect rare mutations that could provide to viral subpopulations anti-immune protection based on the analysis of data from deep targeting sequencing of the HA protein.

Materials and methods

Data accession

The Influenza A virus (H3N2) hemagglutinin gene (GenBank No KF848938.1) was utilized as a reference [10].

Raw reads from patient's material [11] and three isogenic controls obtained from The European Nucleotide Archive [12–14]. Control amplicon was generated from a signle clonally derived plasmid with the HA gene.

Data preprocessing

FastQC v0.12.1 was used to control the quality of raw reads [15].

Reads mapping and variant calling

Reads mapping against a reference HA gene was performed by BWA 0.7.17-r1188 with BWA-MEM algorithm [16].

Mpileup file was generated using samtools 1.18 [17], command “mpileup” used with “--max-depth” parameter values as described below.

Subsequent variant calling was performed by VarScan v2.3 [18], command “mpileup2snp” used with “--min-var-freq” parameter values 0.95 or 0.001.

IGV 2.16.2 was used for data visualization [19].

Results

Data assessment

The phred quality of raw reads is quite high. A lot of duplicates are normal because of deep coverage for a short target region. Therefore, no additional reads trimming was carried out.

Quantification of Aligned Reads

Table 1 presents read mapping statistics, showcasing the number of reads, those successfully mapped to the reference, and the corresponding percentage for the experimental and three control samples. The high mapping percentages indicate robust alignment to the reference genome across all samples.

Table 1. Read mapping statistics for experimental and control samples

Data	Number of reads	Mapped to the reference reads	Percentage
Experimental sample	1433060	361116	99.94%
Control_1	1026344	256658	99.97%
Control_2	933308	233375	99.97%
Control_3	999856	250108	99.97%

Sequencing Depth Analysis: Comparative Average Coverage of Experimental and Control Samples

Table 2 presents the average coverage values resulting from read alignment for the experimental sample and three control samples. The experimental sample exhibits a higher average coverage at 31212.7, suggesting robust sequencing depth compared to the control samples.

Table 2. Average coverage values obtained after reads alignment

Data	Value
Experimental sample	31212.7
Control_1	22630.8
Control_2	20655.5
Control_3	22048.1

Optimizing Variant Calling: Impact of Coverage Depth on SNP Detection Using 'samtools mpileup'

Table 3 displays the results of assessing the optimal coverage depth for generating a pileup file in variant calling using the "samtools mpileup" command. The table reveals the number of identified single nucleotide polymorphisms (SNPs) at different "--max-depth" parameter values, demonstrating an incremental trend in SNP detection with increasing coverage depth.

Table 3. Determining the optimal coverage depth for forming a pileup file for variant calling.

"--max-depth" parameter value	Found SNP (VarScan results for 0.1% frequency)
30'000	16
35'000	18
40'000	20
50'000	**21**
60'000	21

Variant Frequencies in Isogenic Controls

Table 4 provides insights into the variant frequencies within isogenic controls, presenting the total number of single nucleotide polymorphisms, their respective frequencies (in percentage), and the associated standard deviations (SD). The data highlights the consistency of SNP occurrence across samples, aiding in the characterization of genetic variability in isogenic backgrounds.

Table 4. Variants frequency in isogenic controls

Sample No	Total SNP number	Frequency, %	SD
1	57	0.26	0.08
2	52	0.24	0.06
3	61	0.25	0.08

Genetic Variants Analysis

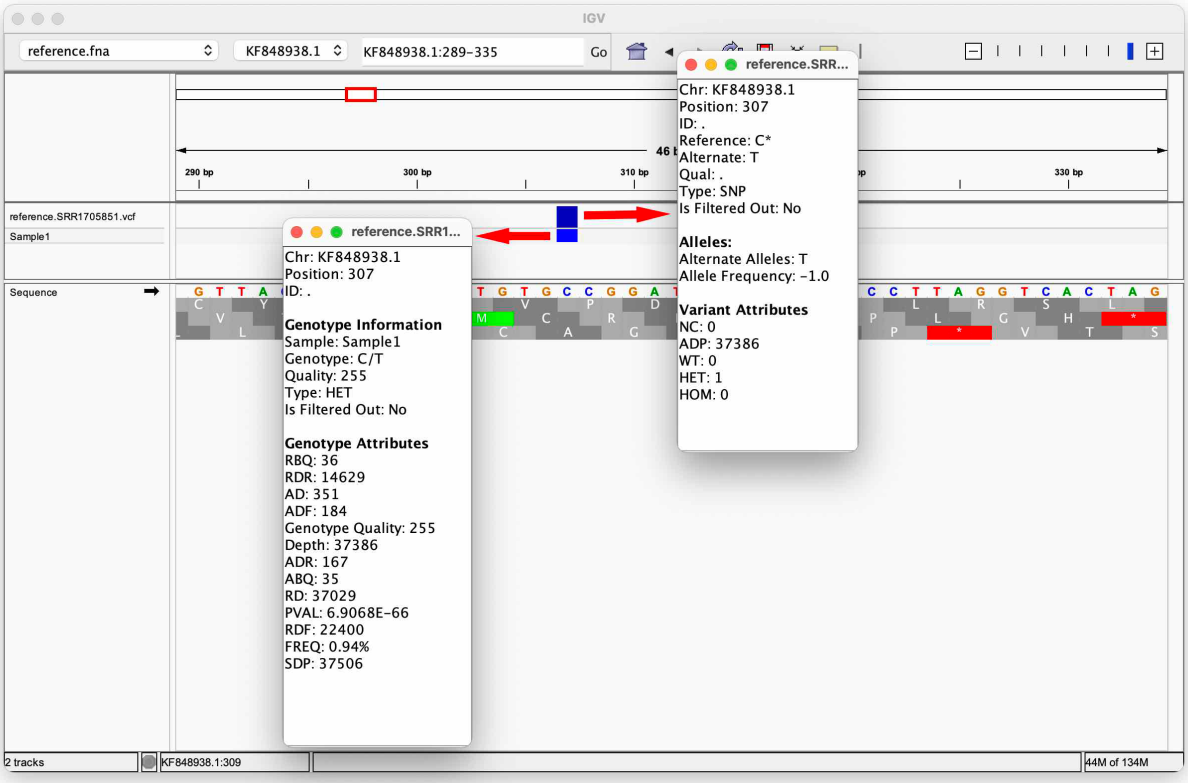
Table 5 details the variants identified in the experimental sample, providing information on the reference base, genomic coordinates, alternative bases, variant frequencies in percentage, and their corresponding status. Notably, the table indicates potential sequencing errors, silent mutations, and missense mutations, shedding light on the genetic alterations present in the analyzed sample. Figure 1 provides detailed visual information on missense mutation.

Table 5. Variants found in experimental sample

Reference	Coordinate	Alternative	Frequency, %	Mutation type
A	254	G	0.17	Sequencing error
A	276	G	0.17	-//-
T	340	C	0.17	-//-
A	691	G	0.17	-//-
A	744	G	0.17	-//-
A	859	G	0.18	-//-
A	1043	G	0.18	-//-
T	1280	C	0.18	-//-
T	915	C	0.19	-//-
A	722	G	0.2	-//-
A	1086	G	0.21	-//-
T	389	C	0.22	-//-
A	1213	G	0.22	-//-
A	802	G	0.23	-//-
T	1458	C	0.84	Silence
C	**307**	**T**	**0.94**	**Missense (Pro103Ser)**
C	117	T	99.82	Silence
C	999	T	99.86	-//-
A	1260	C	99.94	-//-

A	72	G	99.96	-//-
T	774	C	99.96	-//-

Figure 1 Specification of Missense (Pro103Ser) mutation in IGV



Chr: KF848938.1
Position: 307
ID: .
Reference: C*
Alternate: T
Qual: .
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: T
Allele Frequency: -1.0

Variant Attributes
NC: 0
ADP: 37386
WT: 0
HET: 1
HOM: 0

Chr: KF848938.1
Position: 307
ID: .

Genotype Information
Sample: Sample1
Genotype: C/T
Quality: 255
Type: HET
Is Filtered Out: No

Genotype Attributes
RBQ: 36
RDR: 14629
AD: 351
ADF: 184
Genotype Quality: 255
Depth: 37386
ADR: 167
ABQ: 35
RD: 37029
PVAL: 6.9068E-66
RDF: 22400
FREQ: 0.94%
SDP: 37506

Discussion

The emergence of influenza virus variants that evade vaccine-induced immunity poses a significant challenge to public health. In this study, deep sequencing of the hemagglutinin gene from a patient infected with an H3N2 influenza virus, despite prior vaccination, revealed subtle genetic changes that may contribute to vaccine ineffectiveness. Our analysis identified two mutations with frequencies of 0.84% and 0.94%, one silent and the other resulting in a Pro103Ser substitution in epitope D of the HA.

Epitope D is a critical region of the HA protein targeted by the immune system for generating protective antibodies. The Pro103Ser substitution observed in our study is of particular significance, as it occurs within this epitope. This mutation introduces a change in the amino acid sequence, potentially altering the conformation of the epitope and affecting the binding affinity of neutralizing antibodies. Previous studies have emphasized the importance of epitope variability in influenza virus immune evasion [20,21].

To determine the likelihood of these mutations being genuine biological variants and not sequencing errors, we implemented a rigorous approach. By calculating a mean sequencing error rate of 0.25 ± 0.07% from isogenic controls, we established a threshold to differentiate true mutations from background noise. Variants with frequencies exceeding this threshold were considered biologically relevant. This method ensures the specificity of mutation calls and minimizes the inclusion of false positives, a crucial consideration in deep sequencing studies [22,23].

Despite vaccination, our results suggest that the identified mutations allowed the virus to escape immune surveillance, leading to breakthrough infection. The vaccine's inability to confer complete protection against all viral strains is well-documented due to the high mutation rate of influenza viruses, necessitating frequent vaccine updates [24]. In today's world, this is quite a serious problem, but there is already an example of improving vaccine efficacy: finding new target epitopes with high conservativity [25].

To enhance the reliability of deep sequencing experiments and control for potential errors, implementing additional measures will be optimal. One approach involves increasing sequencing depth, as this can improve the accuracy of variant detection, especially for low-frequency variants [26]. Moreover, incorporating unique molecular identifiers during library preparation can help distinguish true variants from PCR or sequencing errors, reducing false positives [27]. Additionally, leveraging error correction algorithms in bioinformatics pipelines, such as those integrated into tools like LoFreq and FreeBayes, can further enhance the precision of variant calling [28,29].

References

1. Paget, J.; Spreeuwenberg, P.; Charu, V.; Taylor, R.J.; Iuliano, A.D.; Bresee, J.; Simonsen, L.; Viboud, C. Global Mortality Associated with Seasonal Influenza Epidemics: New Burden Estimates and Predictors from the GLaMOR Project. *J Glob Health 9*, 020421, doi:10.7189/jogh.09.020421.
2. Cozza, V.; Campbell, H.; Chang, H.H.; Iuliano, A.D.; Paget, J.; Patel, N.N.; Reiner, R.C.; Troeger, C.; Viboud, C.; Bresee, J.S.; et al. Global Seasonal Influenza Mortality Estimates: A Comparison of 3 Different Approaches. *American Journal of Epidemiology* **2021**, *190*, 718–727, doi:10.1093/aje/kwaa196.
3. Blyth, C.C.; Fathima, P.; Pavlos, R.; Jacoby, P.; Pavy, O.; Geelhoed, E.; Richmond, P.C.; Effler, P.V.; Moore, H.C. Influenza Vaccination in Western Australian Children: Exploring the Health Benefits and Cost Savings of Increased Vaccine Coverage in Children. *Vaccine: X* **2023**, *15*, 100399, doi:10.1016/j.jvacx.2023.100399.
4. Moa, A.; Kunasekaran, M.; Akhtar, Z.; Costantino, V.; MacIntyre, C.R. Systematic Review of Influenza Vaccine Effectiveness against Laboratory-Confirmed Influenza among Older Adults Living in Aged Care Facilities. *Human Vaccines & Immunotherapeutics* **2023**, *19*, 2271304, doi:10.1080/21645515.2023.2271304.
5. Doherty, P.C.; Turner, S.J.; Webby, R.G.; Thomas, P.G. Influenza and the Challenge for Immunology. *Nat Immunol* **2006**, *7*, 449–455, doi:10.1038/ni1343.
6. Couceiro, J.N.S.S.; Paulson, J.C.; Baum, L.G. Influenza Virus Strains Selectively Recognize Sialyloligosaccharides on Human Respiratory Epithelium; the Role of the Host Cell in Selection of Hemagglutinin Receptor Specificity. *Virus Research* **1993**, *29*, 155–165, doi:10.1016/0168-1702(93)90056-S.
7. Lorenzo, M.M.G.; Fenton, M.J. Immunobiology of Influenza Vaccines. *CHEST* **2013**, *143*, 502–510, doi:10.1378/chest.12-1711.
8. Ghorbani, A.; Ngunjiri, J.M.; Lee, C.-W. Influenza A Virus Subpopulations and Their Implication in Pathogenesis and Vaccine Development. *Annu. Rev. Anim. Biosci.* **2020**, *8*, 247–267, doi:10.1146/annurev-animal-021419-083756.
9. Flaherty, P.; Natsoulis, G.; Muralidharan, O.; Winters, M.; Buenrostro, J.; Bell, J.; Brown, S.; Holodniy, M.; Zhang, N.; Ji, H.P. Ultrasensitive Detection of Rare Mutations Using Next-Generation Targeted Resequencing. *Nucleic Acids Research* **2012**, *40*, e2–e2, doi:10.1093/nar/gkr861.
10. Influenza A Virus (A/USA/RVD1_H3/2011(H3N2)) Segment 4 Hemagglutinin (HA) Gene, Partial Cds 2015.
11. Index of /Vol1/Fastq/SRR170/001/SRR1705851 Available online: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/ (accessed on 10 November 2023).
12. Index of /Vol1/Fastq/SRR170/008/SRR1705858 Available online: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/ (accessed on 10 November 2023).

13. Index of /Vol1/Fastq/SRR170/009/SRR1705859 Available online: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/ (accessed on 10 November 2023).
14. Index of /Vol1/Fastq/SRR170/000/SRR1705860 Available online: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/ (accessed on 10 November 2023).
15. Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 27 October 2023).
16. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760, doi:10.1093/bioinformatics/btp324.
17. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008, doi:10.1093/gigascience/giab008.
18. Koboldt, D.C.; Chen, K.; Wylie, T.; Larson, D.E.; McLellan, M.D.; Mardis, E.R.; Weinstock, G.M.; Wilson, R.K.; Ding, L. VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples. *Bioinformatics* **2009**, *25*, 2283–2285, doi:10.1093/bioinformatics/btp373.
19. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative Genomics Viewer. *Nat Biotechnol* **2011**, *29*, 24–26, doi:10.1038/nbt.1754.
20. Smith, D.J.; Lapedes, A.S.; de Jong, J.C.; Bestebroer, T.M.; Rimmelzwaan, G.F.; Osterhaus, A.D.M.E.; Fouchier, R.A.M. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **2004**, *305*, 371–376, doi:10.1126/science.1097211.
21. van de Sandt, C.E.; Kreijtz, J.H.C.M.; Rimmelzwaan, G.F. Evasion of Influenza A Viruses from Innate and Adaptive Immune Responses. *Viruses* **2012**, *4*, 1438–1476, doi:10.3390/v4091438.
22. Nielsen, R.; Williamson, S.; Kim, Y.; Hubisz, M.J.; Clark, A.G.; Bustamante, C. Genomic Scans for Selective Sweeps Using SNP Data. *Genome Res* **2005**, *15*, 1566–1575, doi:10.1101/gr.4252305.
23. Schirmer, M.; Ijaz, U.Z.; D'Amore, R.; Hall, N.; Sloan, W.T.; Quince, C. Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform. *Nucleic Acids Res* **2015**, *43*, e37, doi:10.1093/nar/gku1341.
24. Krammer, F. The Human Antibody Response to Influenza A Virus Infection and Vaccination. *Nat Rev Immunol* **2019**, *19*, 383–397, doi:10.1038/s41577-019-0143-6.
25. Raymond, D.D.; Bajic, G.; Ferdman, J.; Suphaphiphat, P.; Settembre, E.C.; Moody, M.A.; Schmidt, A.G.; Harrison, S.C. Conserved Epitope on Influenza-Virus Hemagglutinin Head Defined by a Vaccine-Induced Antibody. *Proceedings of the National Academy of Sciences* **2018**, *115*, 168–173, doi:10.1073/pnas.1715471115.
26. Laehnemann, D.; Borkhardt, A.; McHardy, A.C. Denoising DNA Deep Sequencing Data-High-Throughput Sequencing Errors and Their Correction. *Brief Bioinform* **2016**, *17*, 154–179, doi:10.1093/bib/bbv029.
27. Kühnemund, M.; Wei, Q.; Darai, E.; Wang, Y.; Hernández-Neuta, I.; Yang, Z.; Tseng, D.; Ahlford, A.; Mathot, L.; Sjöblom, T.; et al. Targeted DNA Sequencing and in Situ Mutation Analysis Using Mobile Phone Microscopy. *Nat Commun* **2017**, *8*, 13913, doi:10.1038/ncomms13913.
28. Wilm, A.; Aw, P.P.K.; Bertrand, D.; Yeo, G.H.T.; Ong, S.H.; Wong, C.H.; Khor, C.C.; Petric, R.; Hibberd, M.L.; Nagarajan, N. LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets. *Nucleic Acids Res* **2012**, *40*, 11189–11201, doi:10.1093/nar/gks918.
29. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing 2012.
30. KirPetrikov/BI_2023_HW_Report at HW2 Available online: https://github.com/KirPetrikov/BI_2023_HW_Report (accessed on 10 November 2023).

Supplementary Materials
Available at: https://github.com/KirPetrikov/BI_2023_HW_Report/tree/HW2