BIOINFORMATICS INSTITUTE

Bio-informatics Practice

November 11, 2023

Exploring Hemagglutinin Mutations Throughout Influenza Infection

by Artem Vasilev and Tatiana Lisitsa

Abstract

Influenza, caused by influenza viruses, poses a significant global health threat. This study focusing on the hemagglutinin (HA) gene and its mutations. The project employs deep high-throughput sequencing and control samples to differentiate low-frequency variants from sequencing errors.

Results indicate 21 SNPs in the sample and varying numbers in control samples. Five SNPs in the sample exhibit a minor allele frequency (MAF) exceeding 99%, suggesting their association with the A/Hong Kong/4801/2014 (H3N2) strain. Control samples show mean frequencies ranging from 0.2369% to 0.2565%, with two SNPs deviating from the average by more than 3 standard deviations. Further analysis identifies a synonymous substitution (Y486=) and a missense mutation (P103S) affecting the epitope D, potentially impacting antibody affinity.

Keywords: Influenza, Infection, Hemagglutinin, Influenza vaccine

Introduction

Influenza is an acute respiratory infection caused by influenza viruses [1]. Influenza viruses belong to the Orthomyxovirus family (*Orthomyxoviridae*). There are 4 types of influenza viruses, types A, B, C and D. The genome is represented as a ribonucleoprotein complex, wherein RNA encodes several proteins. The key proteins in pathogenesis of influenza are hemagglutinins (HA) and neuraminidases (NA). Hemagglutinins proteins bind to sialic acid on the surface of respiratory epithelial cells, inducing endocytosis. Unusually for RNA viruses, the viral genome replicates in the nucleus. New viruses assemble on the cell surface and are released through the action of receptor-splicing neuraminidases [2].

Vaccination is the best way to prevent influenza. The WHO recommends vaccines for use in the 2023-2024 influenza season contain the following [3]:

Egg-based vaccines:

- A/Victoria/4897/2022 (H1N1)pdm09-like virus
- A/Darwin/9/2021 (H3N2)-like virus
- B/Austria/1359417/2021 (B/Victoria lineage)-like virus
- B/Phuket/3073/2013 (B/Yamagata lineage)-like virus

Cell culture- or recombinant-based vaccines:

- A/Wisconsin/67/2022 (H1N1)pdm09-like virus
- A/Darwin/6/2021 (H3N2)-like virus
- B/Austria/1359417/2021 (B/Victoria lineage)-like virus
- B/Phuket/3073/2013 (B/Yamagata lineage)-like virus

In this project, we have analyzed the data derived from the sequencing of a patient with influenza infection. We aimed to identify the presence of HA mutations that present at a low fraction within the overall viral population and allow this subpopulation to avoid postvaccine immunity. Also, our sequencing analysis of the HA gene included a series of three control replicates.

Materials and methods

To achieve our goals, we utilized deep high-throughput sequencing, coupled with the incorporation of a set of control samples. This approach facilitated a comprehensive analysis, enabling us to distinguish variants with low frequency in the population from sequencing errors. The amplicons were generated using custom primers for the HA gene. Sequencing libraries were prepared with the Nextera XT Sample prep kits and Nextera XT 24 index kit. Single-end sequencing was performed on MiSeq (Illumina) platform.

Data processing was carried out using an algorithm that involved quality control of fastq files with FastQC v0.12.1 [4]. In our experimental sample, there were 358,265 total reads, with an average length of 150 bases. Alignment of reads to the reference genome sequence of *Influenza A virus* (A/USA/RVD1 H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene was performed using the BWA (Burrows-Wheeler Alignment) software package (Version: 0.7.17-r1188). Percentage of aligned reads: 99.9%. Expected average coverage $= \frac{1{,}358{,}032 \times 150}{1665} = 32,255$.

Coordinate sorting and generation of an mpileup file were performed using Samtools 1.16.1 [5]. We also calculated the maximum depth of coverage for our sample and controles using Samtools. Nucleotide sequence variants were detected using VarScan v2.4.6 [6]. Annotation of the identified variants was performed using python scripts and manual processing. Visualization was carried out in the IGV desktop application v2.16.2 [7].

The classification of nucleotide sequence variants were based on literature data.

Results

When generating the mpileup file, we used the value -d 0 , to use the full depth of coverage. We analyzed the data with VarScan, setting the minor allele frequency (MAF) threshold to 0.001 (0.1%). In the sample, we identified 21 SNPs, while control 1 had 51 SNPs, control 2 had 58 SNPs, and control 3 had 61 SNPs. Subsequently, we processed the files using awk and a Python script in parallel to examine the frequency of the minor allele (refer to Supplementary Materials for details). Out of the 21 SNPs identified in the sample, 5 exhibited a MAF exceeding 99%. We posit that these variants are determinants of the A/Hong Kong/4801/2014 (H3N2) strain. To obtain information about the MAF in variants identified in control samples, refer to Table 1.

	Control 1	Control 2	Control 3
Mean frequencies, %	0,2565	0,2369	0,2503
SD	0,0717	0,0524	0,0780

Table 1: MAF of variants identified in control samples

We identified 2 SNP's with MAF that deviated from the average MAF of the control samples by more than 3 SD:

- 307 C>T (P103S, Pro103Ser) MAF=0,94%
- 1458 T>C (Y486=, Tyr486Tyr) MAF=0,84%

We attributed the remaining variants to sequencing errors that occurred during library preparation for sequencing (present in all control samples) or directly during the sequencing stage.

Discussion

Of the two low-frequency variants identified in the analysis, Y486= (1458 T>C), was a synonymous substitution with no effect on protein structure. The second variant, a missense mutation, resulted in the substitution of proline for serine at position 103. The P103S (307

C>T) mutation was previously described as affecting the epitope D [8, 9]. This change introduces significant impact on the amino acid sequence and could potentially affect the affinity of antibodies to epitope D. Comparing the identified variants to known influenza strains and their potential impact on protein structure provides information for understanding the dynamics of influenza viruses. This understanding is pivotal for vaccine development.

The application of deep high-throughput sequencing, coupled with the correlation of identified genetic variants with a set of control samples, enabled us to robustly differentiate variants with low minor allele frequency from potential sequencing errors.

Supplementary materials

GitHub

References

[1] *Influenza (Seasonal)*. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: 2023-10-03.

[2] Edward C Hutchinson. "Influenza Virus". en. In: *Trends Microbiol* 26.9 (June 2018), pp. 809–810.

[3] *Recommendations announced for influenza vaccine composition for the 2023-2024 northern hemisphere influenza season*. https://www.who.int/news/item/24-02-2023-recommendations-announced-for-influenza-vaccine-composition-for-the-2023-2024-northern-hemisphere-influenza-season. Accessed: 2023-02-24.

[4] *FastQC*. June 2015. URL: https://qubeshub.org/resources/fastqc.

[5] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2 (Feb. 2021). giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

[6] Daniel C. Koboldt et al. "VarScan: variant detection in massively parallel sequencing of individual and pooled samples". In: *Bioinformatics* 25.17 (June 2009), pp. 2283–2285. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp373. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/17/2283/48993337/bioinformatics_25_17_2283.pdf. URL: https://doi.org/10.1093/bioinformatics/btp373.

[7] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". In: *Briefings in Bioinformatics* 14.2 (Apr. 2012), pp. 178–192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. eprint: https://academic.oup.com/bib/article-pdf/14/2/178/546734/bbs017.pdf. URL: https://doi.org/10.1093/bib/bbs017.

[8] Enrique T Munoz and Michael W Deem. "Epitope analysis for influenza vaccine design". In: *Vaccine* 23.9 (2005), pp. 1144–1148.

[9] Anna Cushing et al. "Emergence of hemagglutinin mutations during the course of influenza infection". In: *Scientific reports* 5.1 (2015), p. 16178.