Rare Mutation In Influenza A Virus Hemagglutinin Epitope D Leads To Antigenic Variations

M. Uzun[a], A. Fedorenko[b]

[a]*Research Center Biotechnology RAS, , Moscow, Russia*
[b]*Skolkovo Institute of Science and Technology, , Moscow, Russia*

Abstract

Seasonal influenza, caused by respiratory tract-infecting influenza viruses, prompts the annual design of the influenza vaccine, primarily based on comparisons of viral coat proteins like hemagglutinin. Despite these efforts, vaccines may not consistently protect individuals due to the viral escape phenomenon. Deep sequencing technologies are used to examine the frequency of Single Nucleotide Polymorphisms (SNPs) within epitope regions. However, the significant rate of errors in the process poses challenges in accurately detecting SNPs and low-abundance mutations. This study focuses on an exceptional case, highlighting a rare SNP within the epitope D of the human Influenza A virus, demonstrating its potential impact on the effectiveness of the influenza vaccine.

Keywords: Single nucleotide polymorphisms, Influenza A virus, Deep sequencing, VarScan

1. Introduction

Influenza virus infections pose a major public health threat [1]. The flu vaccine functions by priming the immune system to recognize and combat specific strains of the influenza virus [2], which is achieved through the introduction of inactivated or attenuated virus components. This prompts the production of antibodies targeting immunodominant epitopes, particularly those found in the viral hemagglutinin (HA) protein [3]. However, the influenza virus is known for its dynamic nature, characterized by antigenic drift - a gradual genetic evolution leading to changes in surface proteins [2]. Additionally, within an infected host, the influenza virus exhibits viral quasispecies, representing a diverse population of closely related genetic variants that coexist, influencing the virus's adaptability and evolution [4].

Targeted deep sequencing is a specialized method employed in molecular biology to investigate mixed populations of genetic material by selectively analyzing specific genomic regions [5]. Typical next-generation sequencing (NGS) workflows involve sample processing, DNA extraction, and PCR amplification followed by sequencing, and errors may occur at each of these steps [5]. In this study we analyzed targeted deep sequencing data to investigate genetic variations in the HA gene, seeking insights into the observed ineffectiveness of the flu vaccine.

2. Materials and Methods

Samples of the sequencing data belong to the HA genes of the Influenza A virus A/Hong Kong/4801/2014 (H3N2) strain (roommate's sample), obtained through Illumina single-end sequencing run (NCBI accession number - SRR1705851). Additional samples of the sequencing data were obtained from the isogenic (100% pure) sample of the standard (reference) H3N2 influenza virus. These reference samples were PCR amplified and subcloned into a plasmid, sequenced three times on an Illumina machine, with NCBI accession numbers - SRR1705858, SRR1705859, SRR1705860.

Basic statistics of raw reads were acquired using seqkit v.2.5.1 with the "stats" flag [6]. Subsequently, raw reads underwent quality checks with FastQC v.0.12.1 [7]. The alignment of raw sequence reads to the reference genome of Influenza A virus A/Hong Kong/4801/2014 (H3N2) strain was carried out using bwa v.0.7.17-r1188 with the "-mem" flag [8]. The resulting sam-format file was compressed to a bam-formatted file, sorted, and indexed using samtools v.1.13 [9]. Variant calling was performed using VarScan.v2.4.0, with a minimum variant frequency of 0.95 for the roommate's sample and 0.001 for the reference samples [10]. The alignment and variant calling data were visualized using IGV v.2.16.1 [11].

3. Results

The downloaded sequencing data comprised single-end Illumina reads (Table1). These reads were of good quality and were aligned to the reference genome (Genbank accession: KF848938.1).

Sequence	Number of reads	Mapped reads (%)
SRR1705851 (roommate)	358265	99.94
SRR1705858	256586	99.97
SRR1705859	233327	99.97
SRR1705860	249964	99.97

Table 1: The number of reads in the studied samples and the number of reads that mapped for roommate's and reference samples.

For the studied roommate's and reference samples, SNPs and their frequencies of occurrence were detected. For frequencies of reference samples, the average and standard deviation were calculated (Table 2).

Sequence	mean	SD	mean + 3SD	mean - 3SD
SRR1705858	0,256491	0,071726	0,471669	0,041313
SRR1705859	0,236923	0,052376	0,394051	0,079795
SRR1705860	0,250328	0,078038	0,016215	0,484441

Table 2: Average and standard deviation of the frequencies from each reference sample.

Afterward, mean values of 'mean + 3SD' (0.450054) and 'mean - 3SD' (0.045774) were utilized to identify rare SNPs in the roommate's sample (Supplementary Table 1). In general, 5 common mutations with frequencies different from 95% were identified. Additionally, 2 rare mutations, with frequencies 0.94 and 0.84, were detected. Among all mutations, one was non-synonymous (position 307). It was also observed that this mutation affects the epitope region D in the hemagglutinin protein [12].

4. Discussion

In this work, we detected SNPs for the 'roommate' sample, separating them from errors. To do this, we calculated the average frequency of single nucleotide substitutions in the reference strain sequenced three times, as well as the standard deviation of these frequencies. The 'mean[+3*std*]' values allowed us to obtain the limits within which sequencing errors occur and not to consider such substitutions as SNPs. This approach is based on the fact that errors during NGS workflow steps are less common than real SNPs. Therefore, the more often we encounter this polymorphism, the more likely it is that it is a SNP.

The change in the epitope sequence as we observed in the 'roommate' sample may lead to reduced antibody binding [14] or to its complete absence. Thus, the antibodies contained in the vaccine may not interact with the specific viral epitope causing infection and re-infection to the human [13].

Deep sequencing allows to find such mutations, however, the constraints on practical detection capabilities arise from inaccuracies introduced throughout the sample preparation (PCR amplification) and sequencing processes [15]. Firstly, due to the high error rate of NGS technologies rare genetic changes get overlooked as mistakes and they are lost. Secondly, errors occuring from the previous rounds of PCR amplification have the potential to be amplified in subsequent PCR processes causing false mutations [16]. To improve the accuracy of sequencing the most common and the easiest way is to conduct two independent PCR reactions before sequencing, ensuring that mutations persistently observed in the sequencing results of both reactions are authentic. Nevertheless, tagging technique is a more precise way to overcome errors. Unique tag allows to identify particular DNA template where the error occurs.

Another way to minimize errors is to use bioinformatics tools. For example, GATK uses logistic regression to model base errors, hidden Markov models to compute read likelihoods, and naive Bayes classification to identify variants, which are then filtered to remove likely false positives [17]. Another tools is DeepVariant - a variant caller that utilizes deep learning, specifically Convolutional Neural Networks (CNNs), to accurately identify genetic variants, including single nucleotide polymorphisms and small insertions/deletions (indels), from DNA sequencing data [18].

References

[1] Iuliano, A. D., et al. (2018). Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*, 391, 1285–1300.

[2] Yamayoshi, S., & Kawaoka, Y. (2019). Current and future influenza vaccines. *Nat. Med.*, 25, 212–220.

[3] Herold, S., & Sander, L.-E. (2020). Toward a universal flu vaccine. *Science*, 6480, 852–853.

[4] Domingo, E., & Perales, C. (2019). Viral quasispecies. *PLoS Genet.*, 15, 1–20.

[5] Ma, X., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.*, 20, 1–15.

[6] Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, 11, 1–10.

[7] FastQC. (n.d.). *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*

[8] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv.org*, 00, 1–3.

[9] Danecek, P., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10, 1–4.

[10] Koboldt, D. C., et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22, 568–576.

[11] Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, 39, 23–24.

[12] Muñoz, E. T., & Deem, M. W. (2005). Epitope analysis for influenza vaccine design. *Vaccine*, 23, 1144–1148.

[13] Chakraborty, C., Sharma, A. R., Bhattacharya, M., & Lee, S. S. (2022). A Detailed Overview of Immune Escape, Antibody Escape, Partial Vaccine Escape of SARS-CoV-2 and Their Emerging Variants With Escape Mutations. *Front. Immunol.*, 13, 1–26.

[14] Kugelman, J. R., et al. (2015). Emergence of Ebola Virus Escape Variants in Infected Nonhuman Primates Treated with the MB-003 Antibody Cocktail. *Cell Rep.*, 12, 2111–2120.

[15] Gundry, M., & Vijg, J. (2012). Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, 729, 1–15.

[16] Cheng, C., Fei, Z., & Xiao, P. (2023). Methods to improve the accuracy of next-generation sequencing. *Front. Bioeng. Biotechnol.*, 11, 1–13.

[17] Depristo, M. A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43, 491–501.

[18] Poplin, R., et al. (2018). A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, 36, 983.

Supplementary

Roommate	KF848938.1	KF848938.1	KF848938.1	KF848938.1	KF848938.1	KF848938.1	KF848938.1
Nucleic position	72	117	307	774	999	1260	1458
Reference base	A	C	C	T	C	A	T
Alternative base	G	T	T	C	T	C	C
Amino acid change	Thr (ACA → ACG)	Ala (GCC → GCT)	Pro → Ser (CCG → TCG)	Phe (TTT → TTC)	Gly (GGC → GGT)	Leu (CTA → CTC)	Tyr (TAT → TAC)
Synonym or not	syn	syn	non/missence	syn	syn	syn	syn
Variant frequency (%)	99.96	99.82	0.94	99.96	99.86	99.94	0.84
Amino acid position	24	39	102	258	333	420	486
Affect the epitope regions	-	-	+/ D epitope	-	-	-	-

Supplementary Table 1: Common and rare mutations represented in the viral population and their affection to the epitope regions.