Finding causes of influenza virus infection after vaccination

by deep Illumina sequencing

Abstract

An airborne respiratory virus known as seasonal influenza, or flu, strikes every year from late fall to early spring.The straightforward and accurate method of identifying subtype-specific antigens on the influenza A virus particle is the hemagglutination inhibition test. However, while using this method to analyze non-isogenic samples, we run into issues. Therefore, when the virus population transformed and evolved while replicating inside the organism, this method may not be able to detect the disease. This practical work investigates rare variants within the influenza A H3N2 virus using deep sequencing. Analyzing data from a roommate's infection and three technical replicates, the work aims to distinguish real mutations from sequencing errors. Results identify common and rare mutations in the roommate's sample, with two rare mutations potentially impacting vaccine-induced immunity. However, only one of the determined mutations is located within HA epitope regions. One of the possible mechanisms of that enabled the fluvirus to evade vaccination is proposed.

Introduction

Influenza viruses are negative stranded RNA viruses of the *Orthomyxoviridae* family and influenza A and B are the most common circulating types which are capable of infecting humans. The hemagglutinin (HA) and neuraminidase (NA) proteins, major targets of the immune system and components of many influenza vaccines, are located on the surface of the viral particle. At the same time influenza virus is characterized by extreme genome variability. One of the possible mechanisms of influenza virus variability is **an antigenic drift**, which presents the gradual accumulation of mutations due to errors made by the viral polymerase during genome copying [1]. Due to gradual point changes in HA and NA, virus strains emerge that are so different from previous variants that our immune system recognizes them as completely new. Extremely high mutation rates in viruses can lead to the viral quasispecies formation, which are defined as a complex distributions of closely related variant genomes subjected to genetic variation, competition and selection. Standard clonal analyses and deep sequencing methodologies have confirmed the presence of myriads of mutant genomes in viral populations, and their participation in adaptive processes [2].

As different viral quasispecies may represent extremely rare clonal types (at the level of 0,1%) against some major subtype, very sensitive methods (like deep sequencing) should be used to detect these rare variants. Deep sequencing refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times. This NGS allows detection of rare clonal types, cells, or microbes comprising as little as 1% of the original sample [3]. The NGS measurement error arises in sample preparation (including reverse transcription of RNA genomes to cDNA and amplification of viral genomes), library preparation, sequencing and base calling. And the problem is that at such low frequencies of viral mutation occurrence it is very difficult to distinguish between real rare mutants and sequencing errors [3-4].

The goal of this practical work was to check out the hypothesis that a small portion of the known virus population (strain A/Hong Kong/4801/2014(H3N2) according to HI profile) mutated and evolved while replicating inside the roommate's cells using deep sequencing data.

Methods

The reference sequence for the influenza HA gene was taken from NCBI GenBank.

The roommate's sequencing data was taken from NCBI Sequence Read Archive (SRR1705851). The analyzed sample presents the amplicons from individual patients infected with influenza A H3N2, which were further sequenced on an Illumina MiSeq (2 runs in total) in a single-end manner.

To distinguish between real rare mutations and sequencing errors, three controls (the sample is the same, 3 technical replicas) were used: SRR1705858, SRR1705859 and SRR1705860. The isogenic (100% pure) sample of the standard (reference) H3N2 influenza virus was PCR amplified and subcloned into a plasmid. The control gene amplicon was generated from a single clonally derived plasmid with the HA gene and sequenced 3 times.
To perform further data processing and analysis further tools were used:

- FastQC v0.12.1 to evaluate the quality of reads [5]
- bwa-0.7.17 to align reads to a reference genome [6]
- samtools 1.7 for SAM file compression, BAM file sorting and indexing of [7]
- VarScan.v2.3.9 for variant calling [8]

To call the actual variants in roommate's sample VarScan was used with the minimum variant allele frequency set to 0.95 (or 95%) to look for common variants and after the minimum variant allele frequency set to 0.001 (0,1%) to call for rare variants. To call the variants in control samples VarScan was used with the minimum variant allele frequency set to 0.001 (0,1%).

Results

The total number of reads obtained for each sample as well as number of mapped reads are presented in Table 1.

Sample	Number_of_reads_total	Number_of_reads_mapped
SRR170585 (roommate)	361349	361116
SRR1705858 (control_1)	256744	256658
SRR1705859 (control_2)	233451	233375
SRR1705858 (control_3)	250184	250108

Table 1. Total number of reads and mapped reads for each of 4 datasets.

Based on VarScan results the following averages and standard deviations of the frequencies from each reference sample were obtained (Table 2):

Sample	Mean Frequency	Standard deviation
SRR1705858 (control_1)	0.0025	0.00046
SRR1705859 (control_2)	0.0024	0.00039
SRR1705858 (control_3)	0.0025	0.00053

Table 2. The averages and standard deviations of the frequencies from each reference sample.

Based on VarScan results, the following common mutations in roommate's sample were found:
72 (A>G) AC**A** (Thr) -> AC**G** (Thr) **synonymous (first frame)**
117(C>T) GC**C** (Ala) **-** GC**T** (Ala) -> **synonymous (first frame)**
774(T>C) TT**T**(Phe) -> TT**C**(Phe) -> **synonymous (first frame)**
999(C>T) GG**C**(Gly) -> GG**T**(Gly) -> **synonymous (first frame)**
1260(A>C) CT**A**(Leu) -> CT**C**(Leu) -> **synonymous (first frame)**

Based on VarScan results, after adjusting for control samples the following rare mutations in roommate's sample were found:
307 (C>T) **C**CG (P) -> **T**CG (S) -> non-synonymous (first frame); frequency = 0.0097
1458 (T>C) TA**T**(Tyr) -> TA**C**(Tyr) -> synonymous (first frame); frequency 0.0083

Discussion

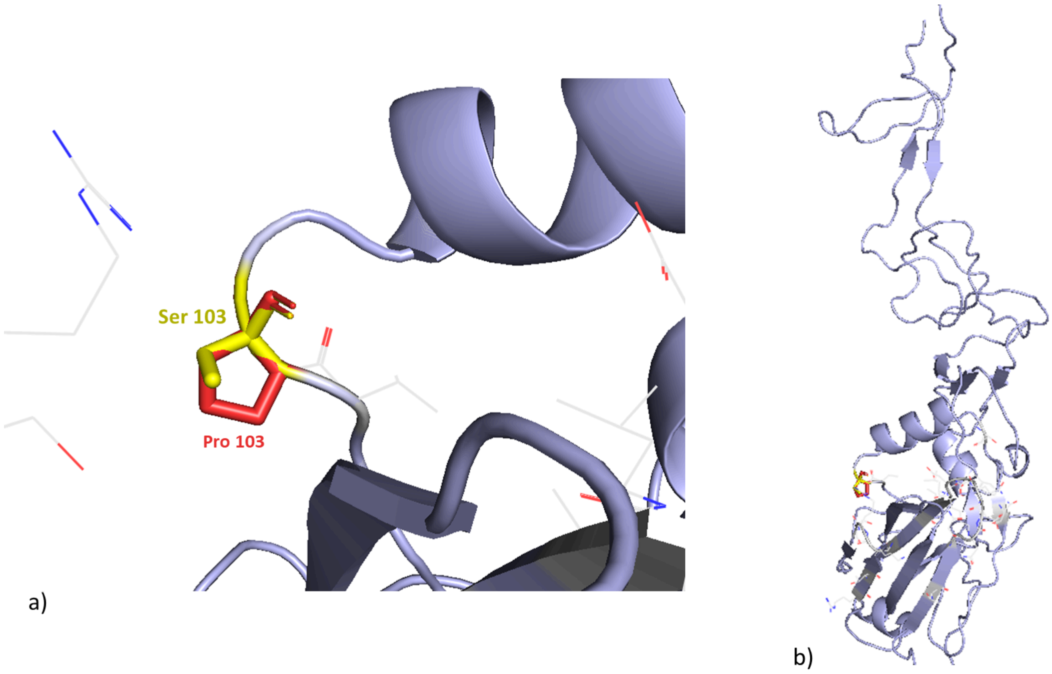
The frequency of mutations was determined and the mean frequency was calculated as a mean of three means and is equal 0.0025. The mean standard deviation was also calculated as a mean of three means and is equal to 0.00046. All the mutations in roommate's sample with the frequencies higher than mean value plus three standard deviations were considered to be real mutations and not mistakes, because in this case the probability that they are not true rare mutations does not exceed 0.3%.

307 (C>T) **C**CG (P) -> **T**CG (S) -> non-synonymous (first frame); frequency = 0.0097
1458 (T>C) TA**T**(Tyr) -> TA**C**(Tyr) -> synonymous (first frame); frequency 0.0083

According to *Munoz et al.* only one of these mutations, 307(C>T) or P103S affects the epitope region, namely region Epitope D (residues 96,102,**103**,117,121,167,170–177,179, 182, 201, 203, 207–209, 212–219, 226–230, 238, 240, 242, 244, 246–248) [9].

One of the possible mechanisms how the vaccinated person could be infected with influenza A (H3N2) virus even those this person was vaccinated could be the following one: the person's roommate got a flu and during the infection process the initial strain got this 307(C>T) mutation in epitope region which possible led to such an extreme change in virus antigen HA part which is normally recognized by the antibodies developed because of vaccination, that this antibodies and immune system can not recognize this modified virus anymore.

For rare variants, accurate quantification and error control is crucial, otherwise there is always a big risk of confusing a real signal with an error. In this work technical replicates with known sequences were used to determine the error rate. There also exist techniques which allow to reduce the error rate at the step of sample preparation and amplification. For example, Primer ID protocol is designed in such a way that it allows control for the artifacts during sample preparation and tracking of individual viral genomes through the PCR and sequencing protocol and direct error correction [10].



*Optional. PDB ID 4O5N. Crystal structure of A/Victoria/361/2011 (H3N2) influenza virus hemagglutinin A chain (the sequence of shown structure fragment is almost the same as of (strain A/Hong Kong/4801/2014(H3N2)). a) Close view of two aligned structures with Pro 103 and mutant with Ser 103. Other aminoacids which comprise the epitope D region are presented with wire lines. b) Complete view of influenza virus hemagglutinin A chain.

Citations

1. Peter C. Soema, Ronald Kompier, Jean-Pierre Amorij, Gideon F.A. Kersten. (2015). Current and next generation influenza vaccines: Formulation and production strategies. European Journal of Pharmaceutics and Biopharmaceutics. 94, 251-263

2. Domingo E, Perales C. Viral quasispecies. PLoS Genet. 2019 Oct 17;15(10):e1008271. doi: 10.1371/journal.pgen.1008271. PMID: 31622336; PMCID: PMC6797082.

3. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426. doi:10.1016/j.tig.2014.07.001

4. Zagordi O, Klein R, Däumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 2010 Nov;38(21):7400-9. doi: 10.1093/nar/gkq655.

5. Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

6. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

8. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568-76. doi: 10.1101/gr.129684.111.

9. 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.

10. Zhou S, Hill CS, Clark MU, Sheahan TP, Baric R, Swanstrom R. Primer ID Next-Generation Sequencing for the Analysis of a Broad Spectrum Antiviral Induced Transition Mutations and Errors Rates in a Coronavirus Genome. Bio Protoc. 2021 Mar 5;11(5):e3938.