Fall BIOINF

Detection of virulence-increasing mutations in Influenza A hemagglutinin protein (HA)

Abstract

The influenza virus is a highly contagious virus with the ability to mutate and evolve rapidly. In the course of our work, we studied a sample of influenza virus that possibly consists of quasispecies. One of the quasispecies is an H3N2 variant, and another is a vaccine-resistant variant. As a result of our search, we discovered one missense mutation in epitope D of HA protein, that could contribute into increased virulence of the variant, as it is an epitope that interacts with antibodies.

Introduction

The influenza virus is a virus responsible for causing a widespread infectious disease commonly called the flu. The prevalence of this virus is caused by its ability to evolve rapidly and transmit through respiratory droplets and aerosols. The most effective way for an organism to protect itself from the Influenza virus is to form antibodies to a viral surface protein, called hemagglutinin (HA). HA-specific antibodies interact with the virus and do not allow it to attach to the surface of the target cell [1]. But, the Influenza virus has a segmented negative-sense RNA genome that replicates via an RNA-dependent RNA polymerase (RdRp) complex that lacks proofreading mechanisms that allow the virus to mutate quickly and accumulate mutations [2]. The process of slow accumulation of mutations in the virus genome that code virus-surface proteins is called antigenic drift. This leads to notion that RNA viruses exist as heterogeneous populations of closely related genetic variants called quasispecies [3].

Deep sequencing study of such populations allows detection of all old and newly emerged quasispecies. This approach is a widely used and accepted method of detection of quasispecies, but it also can be a source of false-positive results, which occurs when sequencing error is not distinguished from a new mutation [4]. In this article, we tried to differentiate between them by analyzing and comparing sequencing errors and mutation frequencies in isogenic and mixed populations.

Materials and methods

Materials Raw whole-genome sequencing read for infected person's viral sample *Influenza A* were obtained from [5]. Clean read of *Influenza A* was downloaded from NCBI [6]. Infected person's reads quality were checked with *FastQC*. Analysis showed that sequence quality per base is good for the raw data from infected person so there is no need to trim reads.
To distinguish mutations from sequencing errors were used 3 isogenic reads [7], [8], [9].

Alignment and coverage All operations were conducted using *SnakeMake*. All reads (from both infected person and isogenic variants) were aligned to reference. We deducted estimated and actual average coverage for the aligned reads in Table 1. It is considerably good and from it can be estimated minimal possible option for the *VarScan*.

We used deep sequencing mode in process of creating pileup of the sorted file.

It is also possible to estimate average coverage for the read with such formula [10]: $C_{est} = \frac{LN}{G}$, where C_{est} stands for coverage, N for number of reads, L - average length of read and G - gaploid genome length. For the infected person sample file: $C_{est} = 31662.4$

Variants call We used 3 different levels for the option minimum variant frequency of *VarScan*. According Table 1 we can deduce that it is senseless to use less than 0.0001 minimal frequency for scanning of results.
We used VarScan with 0.95, 0.01, 0.001 values of option *–min-var-freq*. Results of which is that there are only 5 variant for 0.95 and 0.01; and 21 variant for 0.001.

Results

The number of primary reads, primary mapped reads and average coverage per position for mapped reads are presented in Table 1.

Table 1 Variants alignment information

Dataset	Primary reads	Primary mapped reads	Average coverage per read
Infected person's sample	358265	358032	29989.9
Isogenic 1	256586	256500	21778.2
Isogenic 2	233327	233251	19790.5
Isogenic 3	249964	249888	21186.7

Averages and standard deviations for 3 reference dasasets are presented in Table 2.

Table 2 Averages and standard deviations of the frequencies from isogenic datasets

Dataset	Average	Standard deviation	Average + 3 SD
Isogenic 1	0.2565%	0.0717%	0.4717%
Isogenic 2	0.2369%	0.0766%	0.3941%
Isogenic 3	0.0766%	0.0520%	0.4844%

Mutations detected in infected person's sample that have frequencies more than $Average + 3(Standart\ deviations)$ of frequencies from isogenic populations were thresholded. These mutations along with their frequencies and changed nucleotides are presented in Table 3. Mutation on position 307 especially stands out as it is only mutation that causes an amino-acid change in epitope D of HA protein [11]

Table 3 Infected person's common and rare mutations

Position	Nucleotide change	Amino-acid change	Percent
72	A $\rightarrow$ *G*	Thr $\rightarrow$ *Thr*	99.96%
774	T $\rightarrow$ *C*	Phe $\rightarrow$ *Phe*	99.96%
1260	A $\rightarrow$ *C*	Leu $\rightarrow$ *Leu*	99.94%
999	C $\rightarrow$ *T*	Gly $\rightarrow$ *Gly*	99.86%
117	C $\rightarrow$ *T*	Ala $\rightarrow$ *Ala*	99.82%
307	C $\rightarrow$ *T*	Pro $\rightarrow$ *Ser*	0,94%
1458	C $\rightarrow$ *T*	Tyr $\rightarrow$ *Tyr*	0,84%

Discussion

Used error-correction method

We consider all variants of bases in isogenic samples as sequencing errors. As far as coverage for the infected person's sample is higher than for the isogenic samples we expect the maximal frequency for error to be lower than analogous for the isogenic. So we use *3-σ rule* to deduce the upper bound for the frequency of error in isogenic variants. It is considered reasonable to use the maximal deduced bound of isogenic variants 0.4844% Table 2

One of the two newly found mutations is a missense mutation located on the epitope D of the HA protein, which possibly causes conformational changes that affect protein's affinity to antibodies and increase the virus's ability to infect cells [11].

Proposed error-correction method

In this article we used isogenic viral populations to distinguish errors. Another way for error control and increased read depth coverage might be the use of 3rd generation sequencing. Although 3rd generation sequencing methods known for relatively high produce of sequencing errors, use of these methods could help to overcome errors introduced by PCR [4]. In addition before analysis reads can be preprocessed by trimming 5bp at both ends of each read to remove potentially low-quality bases and adapter contaminations. Reads with low-mapping quality also can be removed from further analysis [12].

References

[1] Hyunsuh Kim, Robert G. Webster, and Richard J. Webby. "Influenza Virus: Dealing with a Drifting and Shifting Pathogen". In: *Viral Immunology* 31.2 (2018). PMID: 29373086, pp. 174–183. DOI: 10.1089/vim.2017.0141. eprint: https://doi.org/10.1089/vim.2017.0141. URL: https://doi.org/10.1089/vim.2017.0141.

[2] Lily Chan et al. "Review of Influenza Virus Vaccines: The Qualitative Nature of Immune Responses to Infection and Vaccination Is a Critical Consideration". In: *Vaccines* 9.9 (2021). ISSN: 2076-393X. DOI: 10.3390/vaccines9090979. URL: https://www.mdpi.com/2076-393X/9/9/979.

[3] Cyril Barbezange et al. "Seasonal Genetic Drift of Human Influenza A Virus Quasispecies Revealed by Deep Sequencing". In: *Frontiers in Microbiology* 9 (Oct. 2018), p. 2596. DOI: 10.3389/fmicb.2018.02596.

[4] Kerensa Mcelroy, Torsten Thomas, and Fabio Luciani. "Deep sequencing of evolving pathogen populations: Applications errors and bioinformatic solutions". In: *Microbial informatics and experimentation* 4 (Jan. 2014), p. 1. DOI: 10.1186/2042-5783-4-1.

[5] ENA-EMBL-EBI. URL: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz.

[6] NLM-NCBI. *Influenza A virus 2011(H3N2) segment 4 hemagglutinin (HA) gene*. URL: https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta.

[7] ENA-EMBL-EBI. URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz.

[8] ENA-EMBL-EBI. URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz.

[9] ENA-EMBL-EBI. URL: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz.

[10] Illumina. URL: https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf.

[11] Enrique T. Muñoz and Michael W. Deem. "Epitope analysis for influenza vaccine design". In: *Vaccine* 23.9 (2005), pp. 1144–1148. ISSN: 0264-410X. DOI: https://doi.org/10.1016/j.vaccine.2004.08.028. URL: https://www.sciencedirect.com/science/article/pii/S0264410X0400636X.

[12] Xiaotu Ma et al. "Analysis of error profiles in deep next-generation sequencing data". In: *Genome Biology* 20 (Mar. 2019). DOI: 10.1186/s13059-019-1659-6.