“Why did I get the flu?”. Deep sequencing, error control, p-value, viral evolution…

Author: Kristina Zheltova

Abstract

Flu, or influenza, is a contagious respiratory illness that spreads from person to person through the air. In this work I studied a sample of influenza virus possibly consisted of quasispecies. As a result I got a missense mutation in epitope D of HA protein, that could increase the virus infectivity

Introduction

Influenza, commonly known as "the flu", is an infectious disease caused by influenza viruses. Symptoms range from mild to severe and often include fever, runny nose, sore throat, muscle pain, headache, coughing, and fatigue [1].

Influenza viruses are constantly changing. They can change in two different ways. One way flu viruses change is called “antigenic drift.” Drift consists of small mutations in the genes of influenza viruses that can lead to changes in the surface proteins of the virus, HA (hemagglutinin) and NA (neuraminidase). The HA and NA surface proteins of influenza viruses are “antigens,” which means they are recognized by the immune system and are capable of triggering an immune response, including production of antibodies that can fight infection. Flu vaccines are designed to target one or more of the surface proteins/antigens of flu viruses [2].

Influenza viruses exist as a large group of closely related viral genomes, also called quasispecies Deep sequencing study of such populations allows detection of all old and newly emerged quasispecies [3].

In this work, I tried to differentiate quasispecies by comparing sequencing errors and mutations in isogenic and mixed populations.

Methods

Data

Raw whole-genome sequencing read was taken from here. Reference sequence for the influenza hemagglutinin gene was taken from here. Data for the three controls (from sequencing of isogenic reference samples) were obtained from SRR1705858, SRR1705859, SRR1705860. Epitope locations were taken from here.

Tools

The data were analyzed with FastQC 0.3.2-1 [4] tool with default parameters. Reference file was indexed and reads were aligned with BWA 0.7.17-9 package [5]. BAM files were sorted and indexed with samtools 1.16.1-1 [6]. In variant calling I used VarScan 2.3.9 to call actual variants.

Coverage estimation

We can estimate average coverage for the read like $cov_{est} = \frac{L_{reads} * N_{reads}}{L_{ref}} = \frac{358265 * 151}{1690} = 32011$ where L_{reads}- length of the read, N_{reads}- number of reads, and L_{ref} - length of the reference.

Variants calling
I used VarScan to look for common and rare variants with param *--min-var-freq* setted to 0.95 and 0.001 respectively.

Results

The data about number of reads and mapped reads are presented in Table 1.

Dataset	Number of reads	Mapped reads	Avg coverage
Raw genome sequence	358265	358032	29989
Control 1	256586	256500	21778
Control 2	233327	233251	19790
Control 3	249964	249888	21186

Table 1 – Mapped reads info

Average and standard deviation of the frequencies from each control sample are presented in Table 2.

Dataset	Average	STD
Control 1	0.256%	0.072%
Control 2	0.237%	0.052%
Control 3	0.25%	0.078%

Table 2 - Average and std for reference samples.

Detected mutations, which have more than *avg +/- 3 std* from controls sequences are presented in Table 3.

Position	Change	Amino-acid change	Frequency (%)
72	A-G	-	99.96
117	C-T	-	99.82
307	C-T	Pro-Ser	0.94
774	T-C	-	99.96
999	C-T	-	99.86
1260	A-C	-	99.94
1458	T-C	-	0.84

Discussion

The mutation on 307 position is a missense mutation on the epitope D of the HA protein. Probably, it causes conformational changes that affect proteins affinity to antibodies and increase the virus infectivity.

Also, in this work I used control samples sequencing data to distinguish errors. But it is possible to trim 5bp at both ends of each read to remove potentially low-quality bases.

Another way to control for error in deep sequencing experiments like this is using of the 3rd generation sequencing, that could help to overcome errors from PCR [7].

References

1. Wikipedia - Influenza
2. How Flu Viruses Can Change: “Drift” and “Shift”
3. Van den Hoecke, Silvie et al. “Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing.” BMC genomics vol. 16,1 79. 14 Feb. 2015, doi:10.1186/s12864-015-1284-z
4. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15), 2114–2120 (04 2014)
5. Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., Wingett, S.: FastQC.Babraham Institute (Jan 2019)
6. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H.: Twelve years of SAMtools and BCFtools. GigaScience 10(2) (02 2021)
7. McElroy, Kerensa et al. “Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions.” Microbial informatics and experimentation vol. 4,1 1. 15 Jan. 2014, doi:10.1186/2042-5783-4-1