Genotyping *A/Hong Kong/4801/2014 (H3N2)* virus hemagglutinin gene. Clinical case.

Ushakov M. Student of the Bioinformatic institute. Saint-Petersburg
Tolmachev M. Student Of the Bioinformatic institute. Saint-Petersburg

Abstract

We genotyped *A/Hong Kong/4801/2014 (H3N2)* virus hemagglutinin gene by new generation sequence method. We found 7 single nucleotide variants and one of them was located in the region which codes amino acid which located in the epitope of hemagglutinin. Therefore we have drawn a conclusion that this variant is responsible for vaccine inefficiency.

Introduction

Vaccination is widely considered as one of the greatest medical achievements of all time. Lots of diseases that were common in the beginning of 19th century are now increasingly rare because of Edward Jenner's development of the vaccine. The importance of organizing vaccination can be traced on the example of the Russian Federation: in 2000, diphteria 771 cases, measles 4.800 cases, mumps 40976 cases, pertussis 29.983 cases, rubella 457.378 cases; in 2018, diphteria 4 cases, measles 2.539, mumps 2.027, pertussis 10.423, rubella 5 [1]. Influenza is an infectious disease caused by an influenza virus. Three of the four types of influenza viruses affect humans: Type A, Type B, and Type C [2]. Despite of regular vaccination against influenza viruses, there is a big risk of getting flu because of Influenza viruses are constantly changing. There are two key mechanism underlying in those processes: antigenic drift and antigenic shift. Antigenic drift means small changes in the genes of influenza virus that can lead to changes in the antigenes structure: hemagglutinin and neuraminidase. Antigen drift is a major change in virus, that leads to new hemagglutinin and neuraminidase and their combinations. Both shift and drift means that immune system wouldn't be prepared for virus even after vaccination against known type.

Clinical case

Life anamnesis

Patient M., white male, 24 years old. Was born in Saint-Petersburg. Early development without features. Got vaccinated according to national vaccitation calendar. In childhood often had colds, don't remember other diseases. Denied chronic and genetic diseases in family. Studied in academic gymnasium № 56. Nowadays studied at First Saint-Petersburg State Medical University n.a. acad. I.P. Pavlov. Patient answered that he smoked one pack of cigarettes per week, drank few bottle of beer per week.

Disease anamnesis

In 14.10.2018 had contact with friend which supposedly had influenza virus. After two days in 16.10.2018 morning felt tired and muscle pain. In 17.10.2018 nose began to release clear fluid, he started coughing without sputum, felt fever and high body temperature – 38.4 degrees of Celsius. Past symptoms still preserved. According to patients words he got vaccinated against influenza virus A/Hong Kong/4801/2014 H3N2 strain .

Materials and methods.

After we have taken patient's informed consent and got new generation sequence results [3] we started our bioinformatic research. The result showed that sample of virus closely matched with A/Hong Kong/4801/2014 H3N2 strain. We supposed that variants in hemagglutinin gene are responsible for avoiding immune system in the first place. We used reference file of that region [4]. We aligned sequences to reference by BWA using mem algorithm and saving results directly in .bam format by samtools[5]. We wrapped up our data in mpileup format by samtools either. To call major variants we used VarScan program (v. 2.4.4) by command java -j --min-var-freq 0.95 –variants. For minor variants we either used VarScan with --min-var-freq 0.0001. For reaction control we used data from isogenic viral sample [6,7,8] providing same protocol. After getting frequency of variants which are 100% known as mistakes we estimated 0.99 percentile and provided 0.99 confidence interval in R by using bootstrap BCA (bias corrected, accelerated) [9] (3.6.1 and packages: dplyr v.0.9; purrr v. 0.3.3; rcompanion v 2.3.7; tidyr 1.0.0). After that, we have drawn a robust conclusion about variants in patient's sequence: if there are random mistakes or actual minor variants. Also we drove a suggestion about source of mistakes in control file: mistakes caused during polymerase chain reaction and actual sequence mistakes. We estimated clusters by Hierarchical Clustering with euclidean distance.

Results

All information about reads and alignments results is contained in table 1.

Table 1 – Reads and alignments information

Sample	Number of reads	Reads mapped	Percentage of reads mapped	Supplementary
Patient	358265	361116	99.94	3084
Control 1	256586	233375	99.97	124
Control 2	233327	256658	99.97	158
Control 3	249964	250108	99.97	220

The results of variant calling showed that 21 snv are present in the patient's sample. Five snv have frequency higher than 95% and the rest have less than 1%.

In the control samples 57, 52 and 61 snv were detected for control 1, control 2 and control 3 respectively. The average and standard deviation for each snv set were calculated (Table 2).

Table 3 – Calculation of confidence intervals

Sample	Average	Sd	Sd * 3	Average + Sd*3
Control 1	0.256	0.072	0.216	0.472
Control 2	0.237	0.052	0.156	0.393
Control 3	0.250	0.078	0.240	0.484

Another way to provide treshold for mistakes is estimate 0.99 quantile for mistakes frequency distribution and 0.99 confidence interval for it: estimation = 0.55, confidence interval = [0.37;0.70]

Then we selected mutations in the patient's sample, the frequency of which is higher than maximum of Average + 3*Sd in the control. All found statistically significant mutations are listed in the Table 3.

Table 3 – High confidence mutations information

Position	Reference	Alternative	Frequency	Codon	Aa	Type

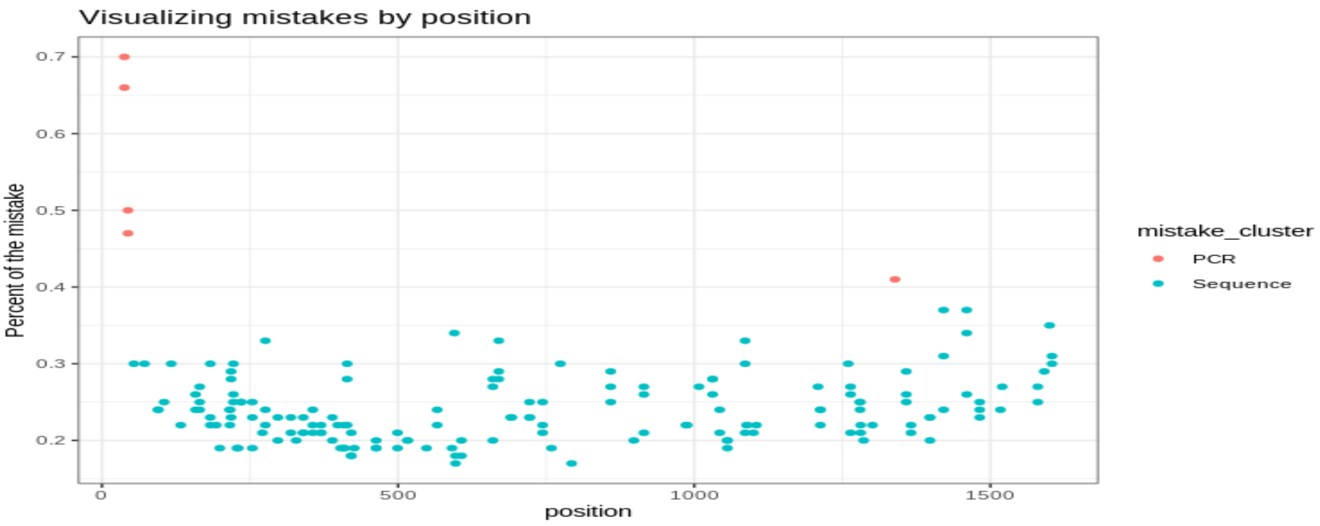
72	A	G	99.96	ACA -> ACG	Thr -> Thr	Synonymous
117	C	T	99.82	GCC -> GCT	Ala -> Ala	Synonymous
307	C	T	0.94	CCG -> TCG	Pro -> Ser	Missense
774	T	C	99.96	TTT -> TTC	Phe -> Phe	Synonymous
999	C	T	99.86	GGC -> GGT	Gly -> Gly	Synonymous
1260	A	C	99.94	CTA -> CTC	Leu -> Leu	Synonymous
1458	T	C	0.84	TAT -> TAC	Tyr -> Tyr	Synonymous

As can be seen from the table, only one of the seven mutations is missense. According to the inspected data this mutation is located in epitope D region of hemagglutinin protein [10].

Discussion

According to the data obtained there is a mutation in the epitope region. This is the main part of antigen and immunogen. Endogenous and intracellularly disposed microbe-derived antigens such as viral antigens are loaded onto the Class I HLA molecules (the HLA I pathway) to be presented to naive CD8+ T cells. The HLA I pathway of the T-cell-mediated response engages naive cytotoxic CD8+T cells, which are activated with the aid of type 1 helper CD4+ T cells. Subsequently, CD8+ T-cell clonal expansion proceeds, and the cells mature until they become effector cytotoxic CD8+ T cells in order to take part in the elimination of such intracellular pathogens like viruses. CD8+ cells, which circulate in bloodstream after vaccination, are unable to recognize immunogen with different epitope. So it will take time for building adaptive immune response, and during that time patient will have disease.

pic.1



Approach that used a rule of 3 sigmas is not robust in this case, because underlying nature of random variable (percent of the mistakes) is beta distribution, not Gauss distribution. There are two common source for the mistakes: polymerase chain reaction and actual sequence mistakes. We expect PCR errors with higher percent of the mistake then sequence errors, because one error would multiply over and over. Cluster analysis proved our suggestions (picture 1).

Conlusion

Vaccination is a great prophylactic tool, but it can’t guarantee 100% defense against highly polymorphic agents. In this particular case, patient was naive for infection, because of mutation in the epitope region. But, patient still resistant for common variant and vaccination is good for your health.

References

1. Available from: http://apps.who.int/immunization_monitoring/globalsummary/countries?countrycriteria%5Bcountry%5D%5B%5D=RUS

2. *Centers for Disease Control and Prevention (CDC)*. 27 September 2017. Retrieved 28 September 2018. Available from: https://www.cdc.gov/flu/about/viruses/types.htm

3. ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz

4. https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta

5.Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

6 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz

7 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz

8 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

9 Carpenter, J. and J. Bithel. 2000. “Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians”. *Statistics in Medicine* 19:1141–1164.

10 Munoz E. et al. Epitope Analysis for Influenza Vaccine Design. Vaccine. 2005; 23(9):1144-1148.