OXFORD GENETICS

The problem of detecting low frequency variants in data from deep sequencing: a case report

Dmitrii Traktirov[1, 2] and Egor Shapoval[1, 3]

[1]Institute of BioInformatics
[2]FSBSI "Institute of experimental medicine"
[3]SPbU

Abstract

Influenza, or flu, is an infectious disease caused by airborne respiratory influenza viruses. Virus can spread through coughing, sneezing, talking. It is not 'stable', it continuously evolving, and a new vaccine needs to be created every year to treat a new strain. However, even if current year's flu strain is covered by the vaccine, it is still possible to get infected. The main aim of this article was to determine mutations, in one particular case, that could be responsible for hiding virus from immune system and to distinguish them from sequencing errors (since mutations can be rare).

Keywords: Influenza virus, flu, mutations, viral immune evasion, deep sequencing

Introduction

Influenza is a rapidly changing virus, mostly because of two processes, antigenic drift and antigenic shift. Since the virus do not have any mechanism for correction during it's replication, viral RNA accumulates mutations gradually, mostly in hemagglutinin (HA) and neuraminidase (NA) genes. Slight mutations in this genes lead to changes in antigens (NA and HA proteins), which can lead to a situation when novel strain is able to evade pre-existing antibody-mediated immunity – this situation is called anitgenic drift. Antigenic shift, on the other hand, is a result of reassortment ("shuffle") of the genetic materials of two antigenically different viruses, which leads to hybrid progeny.

A consequence of the high mutation rate is a need for a new vaccine every year. Seasonal influenza vaccines are available as trivalent or quadrivalent intramuscular injection containing hemagglutinin against main circulating types or subtypes (during 2021–2022 Northern Hemisphere influenza season they were H1N1, H3N2, B/Victoria lineage in trivalent vaccines and B/Phuket/3073/2013-like virus in recommended addition in quadrivalent vaccines) (LJ 2019).

However, even vaccinated people can get infected with the influenza virus due to it's high mutation rate and presence of viral quasispecies in host organism. A quasispecies is a well-defined distribution of mutants that is generated by a mutation-selection process. Identification of mutations responsible for antigenic drift in quasispecies can be challenging due to low frequency of occurrence (they present in a small part of all viral particles), and it is important to be able to separate low frequency mutations from method errors.

In this article, we studied sequencing results of HA genes of viral quasispecies and found out which variants might be responsible for the lack of an immune response.

Methods

In this work, the genome of Influenza A virus (H3N2) segment 4 hemagglutinin (HA) gene from NCBI GenBank was used as reference. Sequencing results of HA genes of viral quasispecies are stored in NCBI SRA.

Virus DNA has been aligned to reference genome via bwa (Li 2013) and samtools (Li et al. 2009) packages. The reads have been piled up via `samtool mpileup` with parameter `-d 40000`. Common variants have been founded using VarScan. Firstly VarScan has been started with parameter `--min-var-freq 0.95` and most common mutations were founded. After that via VarScan with parameter `--min-var-freq 0.001` rare variants were founded.

For detecting sequencing errors in virus DNA sequence three control sequences have been downloaded SRR1705858, SRR1705859, SRR1705860 and aligned to reference. Control sequences have been piled up with parameter `-d 0`. Sequencing errors have been detected via VarScan with parameter `--min-var-freq 0.001`.

A full list of used commands can be found in the attached lab journal.

Results

Four files with sequencing results were used in this study, one with influenza hemagglutinin (HA) gene from a patient with A/Hong Kong/4801/2014 (H3N2) strain of influenza virus (hereinafter referred to as 'roommate's data'), and the other three with HA gene from isogenic sample of the reference H3N2 influenza virus ('control 1-3' from now). After standard inspection of reads quality with fastqc, they were mapped to reference Influenza A virus (H3N2) segment 4 HA gene (mapping information is shown in Table 1).

Table 1 Main information about data. Total reads – number of reads present in each file; mapped reads – % of reads that mapped; error rate – average rate of all found variants (mean± sd).

sequence	total reads	mapped reads	error rate
roommate	361349	99.94%	
control 1	256744	99.97%	0.26±0.07
control 2	233451	99.97%	0.24±0.05
control 3	250184	99.97%	0.25±0.08

All variants obtained with VarScan were explored. For three control samples the average and standard deviation of the variants frequencies were calculated (see Table 1) – they are supposed to be due to method errors.

As for roommate's data results, VarScan with parameter `--min-var-freq 0.95` allowed to identify 6 SNPs, however all of them are synonymous (see Table 2, variants #1-6).

Table 2 High confidence variants found with VarScan v.2.3.9 in HA gene

#	position	freq	base change	aa change
1	72	99.96%	A → G	Thr24Thr
2	117	99.82%	C → T	Ala39Ala
3	774	99.96%	T → C	Phe258Phe
4	999	99.86%	C → T	Gly333Gly
5	1260	99.94%	A → C	Leu410Leu
6	307	0.94%	C → T	Pro103Ser
7	1458	0.84%	T → C	Tyr486Tyr

Running VarScan with parameter `--min-var-freq 0.001` allowed to identify many rare variants, however most of them are PCR/sequencing errors. To decide which variants could be real mutations we have used some kind of threshold – if variant frequency in roommate's data was higher than $\mathrm{mean\ variant\ rate}$ (from any of control samples) $+3 \cdot \mathrm{sd}$ (so $\mathrm{p - value} < 0.05$), we labeled it as high confidence variant.

As a result of this step, we got two high confidence mutations in roommate's data (Table 2, #6-7), but only one appeared to be missense (Pro103Ser). All 21 variants that were detected with runninh VarScan with parameter `--min-var-freq 0.001` can be found in lab journal.

On the last step we have used epitope locations listed in (Muñoz and Deem 2005) to determine if any of the high confidence mutations from roommate's data are located in an epitope region of HA protein. It turned out that one of residues forming the epitope D is residue #103, and therefore Pro103Ser change occurs in this epitope.

Discussion

As it was shown, influenza vaccines cannot completely prevent people from getting infected, as mutations accumulate gradually and can easily lead to appearance of new mutants in quasispecies, which are not recognized by the immune system. Sometimes, just one slight mutation is enough to get new subtype of virus, hidden from immune system. It is only necessary that the mutation is located in the special region of HA/NA genes, responsible for virus recognition by the immune system and called epitope.

In present study we made an attempt to recognize appearence of such a mutation in quasispecies in a patient with A/Hong Kong/4801/2014 (H3N2) strain of influenza virus. As mutations appear only in a small part of all viral particles, it is necessary to distinguish such rare mutations from method errors. Here we took for high confidence mutations the variants with observed frequency greater than 3 standard deviations away from reference error rate.

As a result, we found one C307T mutation in HA gene that leads to Pro103Ser change in epitope D of HA protein. This mutation could cause some conformation changes in HA protein, making it unrecognizable by the immune system of a vaccinated person.

It is important to notice that sequencing errors are key confounding factors for detecting low-frequency genetic variants that are important for molecular diagnosis of many diseases using deep next-generation sequencing (NGS), that's why it is important to be able to detect such errors (Ma et al. (2019)). In addition to overall average frequency in control samples, we could also look at read quality and try to define which of variants could appear because of sequencing error. Also, different polymerases are able to lead to different error profiles, so information about polymerase could help us in detecting errors.

Literature cited

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

LJ K. 2019. Seasonal influenza (flu). The Nursing clinics of North America. 54:227–243.

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J et al. 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biology 2019 20:1. 20:1–15.

Muñoz ET, Deem MW. 2005. Epitope analysis for influenza vaccine design. Vaccine. 23:1144–1148.