Bioinformatics
doi.10.1093/bioinformatics/xxxxxx

OXFORD

Genome analysis

Analysis of H3N2 influenza mutations affecting antibiotic resistance

Rubinova Valeriya[1], Vlasenko Daniil[2]

[1]Bioinformatics Institute, Saint-Petersburg, Russia, E-mail: valeriyakudr98@gmail.com
[2]St. Petersburg State University, School of Mathematics and Mechanics, St. Petersburg, Russia, E-mail: vlasnko.daniil.vl@gmail.com.

Abstract

This work is aimed at studying the cause of the disease despite the flu vaccination. We analyzed data from targeted deep sequencing of a patient with a new strain of the H3N2 influenza virus and tried to find the mutation that is most likely to cause the disease. 3 controls were also analyzed to correctly interpret sequencing errors and real mutations. Average and standard deviations of noise frequencies were calculated. Based on these values, the obtained SNPs in the test sample were filtered and only SNPs which frequencies differ from the average noise frequency by three standard deviations of the noise frequency were analyzed. It was clarified that mutation 307 C→T (0.94) is the only non-synonymous among all found (Pro → Ser).

Supplementary information: https://github.com/Daniil-Vlasenko/IBBioinformaticsWorkshop/tree/main/Project%202

1 Introduction

The flu vaccine contains a weakened or killed virus. After its injection into the body, it contributes to the production of antibodies and special immune cells that protect against the effects of infection [1]. This takes an average of 14 days. The positive effect of vaccines is that they form immunity and thus prevent the risks of a severe course of the disease. This reduces the number of deaths and serious complications.

There is such a phenomenon as antigen drift - a variant of antigenic variability through mutations, which is slow, prolonged and random changes in the immunoforming surface proteins of viruses (antigens) [2]. As a result, various populations of viruses are formed - quasi-species - a set of several lines that interact genetically and support each other. A distinctive feature of quasispecies of viruses is their significant variability with the formation of many simultaneously existing, immunologically different antigenic variants [3].

The antibodies produced after vaccination bind to the antigen of the virus (with its active center - the epitope). If a mutation occurs in an epitope of a virus (for example, through genetic drift), antibodies cannot bind to it, and thus do not protect the body from disease by this strain [2].

In this case, genetically different strains of the virus are present in the body. Whole genome sequencing (WGS) is used to assess the genetic variability of virus strains. WGS on NGS platforms can become one of the main methods of molecular epidemiology, since it has a higher resolution for detecting any changes in the genome that can lead to specific manifestations of the pathogenicity of infectious agents [4].

To correctly interpret the results, it is important to remember the main sources of error in sequencing:

1. at the stage of sample preparation - are caused by the presence of non-target molecules in the sample, by enzyme errors (polymerase); they cannot be detected at the sequencing stage.
2. at the stage of sample sequencing - due to the peculiarities instrument measurement system and sequencing technologies; such errors can be detected or predicted. The mechanism of error prediction is based on the assignment of the quality of reading the base by the device [5].

In this work, we tried to distinguish between real mutations and sequencing errors, and thus find the true cause of the disease.

2 Methods

We used influenza A virus (H3N2) segment 4 hemagglutinin (HA) gene as a reference sequence [6] and raw Illumina sequencing reads of the patient's influenza [7] that is resistant to the flu vaccine. Also, we used an isogenic samples of the reference H3N2 influenza virus PCR amplified and subcloned into a plasmid [8, 9, 10] which was sequenced with Illumina to separate sequencing errors from influenza variations.

We aligned all reads to the reference with BWA [11]. We used SAMtools [12] and VarScan [13] mpileup2snp option with a minimum allele frequency threshold of variants equal to 0.001 to select SNPs. We then calculated mean and standard deviation of SNPs frequencies for sequencing errors using the sequencing results of isogenic reference H3N2

Table 1. Average and standard deviation of the frequencies from each reference sample.

Sample	Average	Standard derivation
SRR1705858	0.2564	0.07172
SRR1705859	0.2369	0.05237
SRR1705860	0.2503	0.07803

Accordingly, the mean of the average noise frequencies is 0.2479. The mean standard deviation of the noise frequencies is 0.06738. We used the last two numbers as the mean and standard deviation of the noise.

Table 2. Remaining patient influenza SNPs after filtering.

Position	Reference	Alternate	Allele frequency
72	A	G	99.96
117	C	T	99.82
307	C	T	0.94
774	T	C	99.96
999	C	T	99.86
1260	A	C	99.94
1458	T	C	0.84

Position is a 1-based position of the variation in the reference sequence; Reference is a reference base at the given position in the reference sequence; Alternate is an alternative allele at this position; Coverage is a depth coverage of the position; Allele frequency is a number of the allele encounters.

samples, and left those patient influenza SNPs which frequencies differed from the mean sequencing error frequencies by three standard deviations.

We then used IGV browser [14] to select positions where SNPs could lead to a change in the protein structure.

3 Results

There were 358265 patient influenza reads and 256586, 233327, 249964 isogenic reference H3N2 reeds, 99.93%, 99.97%, 99.97% and 99.97% of which were mapped by BWA respectively.

The results of calculations of average and standard deviations of noise frequencies are presented in Table 1. Table 2 shows all the SNPs that were found in the patient's influenza, which frequencies differ from the average noise frequency by three standard deviations of the noise frequency.

Of these, only the third SNP is non-synonymous and change proline to serine in the protein.

4 Discussion

After receiving patient influenza SNPs after filtering it was clarified that mutation 307 C $\rightarrow$ T (0.94) is the only non-synonymous among all found (Pro $\rightarrow$ Ser). According to research [15] this amino acid is located in an epitope C region of hemagglutinin. A non-synonymous mutation that occurred in one of the epitopes of hemagglutinin changed its amino acid sequence. As a result of vaccination, antibodies have been developed in the body that bind to the unchanged epitope of the influenza virus. Thus, the antibodies did not bind to the epitope in which the amino acid change occurred, and the person fell ill with this strain of the influenza virus.

In some cases, rare gene sequence subvariants are often indistinguishable from subvariants resulting from sequencing errors and prior PCR amplification. So we want to suggest an interesting method for error control in deep sequencing experiments like this.

The study [16] describes a variant of the analysis of molecular-barcoded data, called Molecular Identifier Groups-based Error Correction (MIGEC). It is based on a two-stage bioinformatics analysis. First stage. Read sequences carrying the same unique molecular identifier (barcode) are combined into one group (cluster) - Molecular Identifier Group (MIG). The presence of an identical molecular identifier indicates that these read sequences were generated from the same starting DNA or cDNA molecule. Clustering the sequencing data by a unique identifier makes it possible to establish the original nucleotide sequence by the dominant sequence in each group. The second stage of the analysis is based on the fact that high-frequency (in each specific DNA context) PCR errors are repetitive, which makes it possible to distinguish them from real diversity. Such errors give themselves away by appearing as a minor subvariant in a large number of MIGs and, accordingly, can be identified based on the relative frequency of occurrence of a variant sequence in the form of "major" or "minor" in the MIG.The Two-stage algorithm allows filtering out erroneous variants of sequences with high accuracy, while maintaining the natural diversity of the library, and provides the ability to conduct deep error-free analysis of complex libraries.

References

[1]Carrat F., Flahault A. Influenza vaccine: the challenge of antigenic drift //Vaccine. – 2007. – T. 25. – №. 39-40. – C. 6852-6862.

[2]Boni M. F. Vaccination and antigenic drift in influenza //Vaccine. – 2008. – T. 26. – C. C8-C14.

[3]Andino R., Domingo E. Viral quasispecies //Virology. – 2015. – T. 479. – C. 46-51.

[4]Sanjuán R, Domingo-Calap P. Genetic Diversity and Evolution of Viral Populations. Encyclopedia of Virology. 2021:53–61. doi: 10.1016/B978-0-12-809633-8.20958-8. Epub 2021 Mar 1. PMCID: PMC7157443.

[5]Ma X. et al. Analysis of error profiles in deep next-generation sequencing data //Genome biology. – 2019. – T. 20. – №. 1. – C. 1-15.

[6]Gene. National Center for Biotechnology Information; Influenza A virus (A/USA/RVD1_H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene, partial cds; Available from: https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta.

[7]Gene. Available from: http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851.

[8]Gene. Available from: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858.

[9]Gene. Available from: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705859.

[10]Gene. Available from: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705860.

[11]Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

[12]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, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9 [19505943]

[13]Daniel C. Koboldt, Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Li Ding, VarScan: variant detection in massively parallel sequencing of individual and pooled samples, Bioinformatics, Volume 25, Issue 17, 1 September 2009, Pages 2283–2285,

https://doi.org/10.1093/bioinformatics/btp373.

[14]James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).

[15]Munoz E. T., Deem M. W. Epitope analysis for influenza vaccine design //Vaccine. – 2005. – Т. 23. – №. 9. – С. 1144-1148.

[16]Egorov ES, Israelson MA, Kasatskaya SA, Chudakov DM, Lukyanov SA. Qualitative error-free analysis of mass sequencing data using molecular barcoding. Bulletin of RSMU. 2015; (4): 4–9. DOI: