“Why did I get the flu?”. Deep sequencing and influenza

Evdokimova Anastasiia, Bioinformatics Institute

Abstract

This research is more like a detective investigation. Two people live together. One of them did not get the flu shot and became ill. Fortunately, the second turned out to be more responsible and got the vaccine. But after a while he also fell ill. How did this happen? Let's find out!

Introduction

Humanity has been controlling the influenza virus for about 70 years thanks to annual mass vaccinations. The vaccine allows organisms to produce antibodies targeting the hemagglutinin - viral surface protein, playing a key role in the penetration of the virus into the cell. But the flu virus constantly mutates, and sometimes this leads to the appearance of particularly aggressive mutations, which is why we still face deadly flu epidemics, for example pandemic H1N1/09 virus (swine flu).

Influenza mutates at a rate of one mutation per genome per replication, and can exist as quasispecies, even within a single host. If more than one strain of flu infects the same host, they can recombine.

The main role in the evolution of influenza viruses is played by antigenic drift and antigenic shift. As a result of antigenic drift, a gradual change in the properties of the virus occurs due to mutations in the viral genome and natural selection of the most “successful” variants that can easily infect animals and humans. Antigenic shift is associated with reassortment and represents the "birth" of a new viral particle from two different strains.

To find out rare mutations deep sequencing is used. Deep sequencing refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times. This NGS approach allows to detect rare clonal types, cells, or microbes comprising as little as 1% of the original sample. But in this case, it is problematic to find low-frequency mutations against the background of errors arising during sample handling, library preparation, PCR enrichment, and sequencing. In this work we try to overcome this problem and find out why the person who got the flu vaccine was infected and sick.

Methods

For the analysis, deep sequencing data were taken from NCBI Sequence Read Archive (an Illumina single-end sequencing run) [1]. The quality of raw reads was assessed using the program FastQC (version 0.10.1) [2].

For reference we used the influenza A virus (A/USA/RVD1_H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene [3] (length = 1735 bp).

To align reads to the reference sequences was used aligner BWA-MEM (version 0.7.5a-r405) [4]. To compress, index and simply statistic of SAM-file with alignments was used Samtools *view*, *index*, *flagstat*, *sort* (version 0.1.19-44428cd) [5] .

Samtools mpileup was used to search for mutations. Since our variants may be quite rare, we set that depth limit to something we know is higher than our coverage with the -d flag. The value of this flag was calculated manually, considering the length of the reads, the length of the reference sequence and the number of reads (d = number of reads / (length of reference / length of reads)).
To annotate SNP in our data was used prograng VarScan (version 2.3.9.), command *mpileup2snp* [6]. To look for common variants we set the minimum variant frequency to 0.95 (95 %), to look for rare variants - 0.001 (0.1 %).
To visualize results was used IGV Browser (version 2.8.11) [7].
To see the difference between real, rare mutants in the viral population, and errors introduced in the sequencing and amplification process, we used the control sample - the isogenic (100% pure) sample of the standard H3N2 influenza virus which was PCR amplified, subcloned into a plasmid and sequenced three times on an Illumina machine. We used these three sequencing data sets to estimation of the frequency of the errors to help figure out what's an error and what's a true variant in the data from your roommate [8, data SRR1705858, SRR1705859, SRR1705860]. We did the same procedure with this data as with the test sample (quality control, aligning to the same reference, converting to bam file, indexing, search mutations using samtools mpileup and running VarScan with the minimum variant frequency to 0.001).
After that we had three files with variants. We calculate average and standard deviation of frequencies and compare it with frequencies of variants in the test sample file.For the next analize we took mutations with frequencies that are more than three standard deviations away from the averages in the reference files.
To visualize results was used IGV Browser.

Results

The result of quality control of raw reads from test sample data and control data you can see in Supplementary Information (SI). We have a good quality of all reads and some problems with GC-content, but GC distribution is not bimodal, so it is not a contamination of the sample. Besides that, we have a high level of duplicates and overrepresented sequences, which is normal in a deep sequencing experiment.
After quality control we aligned reads on the reference. Table 1 shows information about the total number of reads in each fastq file, the number of unmapped reads and the percentage of mapped reads.

		control data		
	test sample	SRR1705858	SRR1705859	SRR1705860
total	358265	1026344	933308	999856
unmapped	233	86	78	77
% of mapped reads	99.94%	99.97%	99.97%	99.97%

Table 1. The total number of reads in test sample and control files, the number of unmapped reads and the percentage of mapped reads

Searching for common variants didn't give us any interesting results: we found five mutations, but all of that doesn't change amino acids (Table 2) (to see commands go to SI).

№	Position	Ref.	Ref. amino acid	Alt.	Alt. amino acid
1	72	ACA	Threonine (T)	ACG	Threonine (T)
2	117	GCC	Alanine (A)	GCT	Alanine (A)
3	774	TTT	Phenylalanine (F)	TTC	Phenylalanine (F)
4	999	GGC	Glycine (G)	GGT	Glycine (G)
5	1260	CTA	Leucine (L)	CTC	Leucine (L)

Table2. . Common variants in the test sample data.

Next we look at rare variants. We found 14 mutations. To understand which is real, and which is just errors introduced in the sequencing and amplification process, we used the isogenic sample of the standard H3N2 influenza virus. We aligned three fastq files on the reference (see Table 1) and looked for rare mutations (see SI). All mutations in the control sample are just errors, so we calculate average and standard deviation of errors in each file (Table 2B). In SI you can find the table with all rare variants in control data. Among the rare mutations in the test sample, we are only interested with frequencies that are more than 3 standard deviations away from the averages in the reference files (Table 2A, C). So, we have two significant mutations, and only one of them, in position 307, leads to replacement of the amino acid.

A

pos	Ref	Alt	Freq, %
95	A	G	0.18
254	A	G	0.19
307	C	T	0.95
340	T	C	0.18
389	T	C	0.23
722	A	G	0.23
744	A	G	0.18
802	A	G	0.24
915	T	C	0.2
1043	A	G	0.19
1086	A	G	0.21
1213	A	G	0.22
1280	T	C	0.18
1458	T	C	0.83

B

	average	sd	average + 3* sd
SRR1705858	0,2467241	0,04601251	0,38476163
SRR1705859	0,232963	0,04601251	0,37100053
SRR1705860	0,232963	0,04601251	0,37100053
mean	0,2375500333	0,04601251	0,3755875633

C

pos	Ref	Ref. amino acid	Alt	Alt. amino acid
307	CCG	Proline (P)	TCG	Serine (S)
1458	TAT	Tyrosine (Y)	TAC	Tyrosine (Y)

Table 3. A. Position, reference base, alternative base and frequency of rare mutations in test sample data. B. Average and standard deviation of errors in control data. C. Significant mutations in the test sample. Only SNP in 307 position leads to the replacement of the amino acid.

Discussion

We found out that a hemagglutinin mutation P103S exists in the test sample. We made this conclusion based on the fact that the frequency of this mutation is more than three standard deviations higher than the standard sequencing error rate calculated from the analysis of control samples - the isogenic sample of the standard H3N2 influenza virus. The hemagglutinin H3 protein has five epitope regions (A–E), mutations P103S located in epitope region D [9]. We know that the flu vaccine in season 2017/2018 contains strain A/Hong Kong/4801/2014 (H3N2) [10]. So, we can suggest that mutation in epitope regonin D leads to preventing antibodies, targeting the hemagglutinin, from binding, which causes the disease.

In this research, we encountered the problem of sequencing errors, which is a key confounding factor for detecting low-frequency genetic variants, because mutation rates are often in the error rate range. How to reduce sequencing errors? First of all, we can increase the threshold for the quality of reads, and trimm the bases if their quality is less than 30 Phred score, for example (in our experiment the lowest quality was 20 Phred score). It will reduce the chance of error from 0.01% to 0.001%. But it can lead to very short reads and difficulty aligning. Besides that, we can set the threshold of reads mapping quality. In [11] was developed an algorithm to identify rare variants in experiments with paired sequencing. It was designed to account for the concordance between forward and reverse readouts so that discordant readouts were not counted and concordant readouts were counted only once. Also, in this article it was found that error rates have substitution-type and sequence context dependencies, which reflect fidelity of DNA polymerases. They also found that C>A/G>T errors are enriched in a subset of samples, which indicates sub-optimal handling/storage conditions. So, it makes sense to think about studying the optimal handling/storage conditions required to minimize errors.

Citations

1. Sequencing data: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz
2. FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. Reference: https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta
4. BWA: http://bio-bwa.sourceforge.net/bwa.shtml
5. Samtools: https://www.htslib.org/
6. VarScan: http://dkoboldt.github.io/varscan/
7. IGV Browser: https://igv.org/
8. Control data: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/
9. Enrique T. Munoz, Michael W. Deem. Epitope analysis for influenza vaccine design. Vaccine 2005 Jan 19;23(9):1144-8. doi:10.1016/j.vaccine.2004.08.028.
10. WHO: https://www.who.int/
11. Ma, X., Shao, Y., Tian, L. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* 20, 50 (2019). https://doi.org/10.1186/s13059-019-1659-6
12. https://en.wikipedia.org/wiki/DNA_codon_table
13. https://docs.google.com/presentation/d/1LWmmcP2HJMDevZbjykfVhJig0MFaY-pMBjGqQeHa8t8/edit#slide=id.g1d69fc4019_0_91

Supplementary Information

FastQC reports:
https://drive.google.com/file/d/1YODGL5McKcD7_6Be47h1BJ_tvZbq_0Q7/view?usp=sharing

Looking for common mutations in the test sample
Searching variants: d = number of reads / (length of reference / length of reads) = 358265 / (1735 / 151) = 32000 (round up)
samtools mpileup -d32000 -f reference.fa alignment.bam > my.mpileup
Running VarScan: *java -jar ../Project_1/VarScan.v2.3.9.jar mpileup2snp my.mpileup --min-var-freq* 0.95 *--variants --output-vcf* 1 > *VarScan_results.vcf*

Looking for rare mutations
Test sample:
java -jar ../Project_1/VarScan.v2.3.9.jar mpileup2snp my.mpileup --min-var-freq 0.001 *--variants --output-vcf* 1 > *VarScan_results.vcf*

Control sample: d = number of reads / (length of reference / length of reads) = 1026344 / (1735 / 151) = 90000 (round up)

samtools mpileup -d90000 -f ../reference.fa alignment58.bam > my58.mpileup
samtools mpileup -d90000 -f ../reference.fa alignment59.bam > my59.mpileup
samtools mpileup -d90000 -f ../reference.fa alignment60.bam > my60.mpileup

java -jar ../Project_1/VarScan.v2.3.9.jar mpileup2snp my.mpileup --min-var-freq 0.001 *--variants --output-vcf* 1 > *VarScan_58.vcf*
java -jar ../Project_1/VarScan.v2.3.9.jar mpileup2snp my.mpileup --min-var-freq 0.001 *--variants --output-vcf* 1 > *VarScan_59.vcf*
java -jar ../Project_1/VarScan.v2.3.9.jar mpileup2snp my.mpileup --min-var-freq 0.001 *--variants --output-vcf* 1 > *VarScan_60.vcf*

Table with rare variants:
https://docs.google.com/spreadsheets/d/1zGeANFBL4L0TCxi5JUoKpq1DTJZUiqm1Dryy0xNaMyk/edit?usp=sharing

Lab Notebook:
https://docs.google.com/document/d/1RCuia9kTY7thXSFkOu1DO_dxi9I7jKCFzti1USnczNU/edit?usp=sharing