One strain to infect them all: vaccine-resistant mutation in influenza virus's hemagglutinin

Aigul Nugmanova[1*†] and Bogdan Sotnikov[1,1*†]

[1] Bioinformatics institute, Kantemirovskaya street 2A, Saint Petersburg, 197342, Russia.
[2] Medical faculty, Kyrgyz-Russian Slavic university, Kievskaya street 44, Bishkek, 720000, Kyrgyzstan.

*Corresponding author(s). E-mail(s): aya.nugmanova@gmail.com; bogdan.sotnikov.1999@mail.ru;
†These authors contributed equally to this work.

Abstract

Vaccination against influenza is an important task to decrease mortality and morbidity in people population. Cases of vaccine-resistance of flu strains can decrease the confidence of society in vaccines. Unlocking exact mechanisms of each vaccine-resistant morbid event can increase our understanding methods of controlling virus infections in general and influenza particulary. In this paper we investigate hemagglutinin mutations, which let virus to escape vaccine-induced antibodies in patient. The mutation of 307th nucleotide derive to change proline to serine in 103th position of hemagglutinin protein. It causes changes in surface of epitope D and transformed this substrain to vaccine escape mutant.

1 Introduction

Influenza viruses are single-stranded RNA viruses of 3 types: A, B and C[1]. Usually, type A is a cause of epidemic outbreaks. It is classified based on the expression of one of the 17 different subtypes of hemagglutinin (H1-H17) and 9 different subtypes of neuraminidase (N1-N9)[1]. Viruses mutate very fast and some of these variants lead to changes in the surface proteins of the virus and the following spread of these new quasispecies. This process is called antigenic drift.

Most flu vaccines are developed to trigger an immune response to hemagglutinin, thus their effectiveness depends on the spread of strains during the current season[2]. In this work, we made an analysis of hemagglutinin of one new strain that has infected a vaccinated patient. We used the prepared reads received with targeted deep sequencing. Deep sequencing with big coverage for each subsequence gives an opportunity to detect rare mutations related to specific strains when we study mixed populations. In our work, we tried to filter variants that appeared as sequencing mistakes. Such mistakes are produced in library preparation, for example, primer mistakes, as well as technical errors during the sequencing [3, 4].

2 Methods

Our target investigated sample is a mixed population of influenza hemagglutinin genes including A/Hong Kong/4801/2014(H3N2) collected from one person. As a reference sequence, we used an isogenic viral sample of the hemagglutinin gene from A/USA/RVD1_H3/2011(H3N2). Also, we had 3 control samples of sequencing derived from a virus clone that matches the reference sequence. Tabel 1 shows the statistics for each sample.

Table 1 Information about used data

	ID	Number of reads	Mean length of reads
Experimental sample	SRR1705851	358265	148.15
Reference 1	SRR1705858	256586	149.56
Reference 2	SRR1705859	233327	149.45
Reference 3	SRR1705860	249964	149.7

To build the alignment we used the algorithm BWA MEM in BWA v.0.7.17-r1188 [5]. We also used samtools v.1.16.1 to compress, sort and index the alignment. By default, samtools stop piling up the base calls at each position when it gets to 8000 calls, thus we increased this threshold to about 30000 because of deep coverage.

For variant calling used VarScan v.2.4.0[6]. The threshold for the minimum variant allele frequency was selected based on the frequency of sequencing mistakes in control samples. Firstly, for each control sample VarScan returned the mutations with the variant frequency of more than 0.1% (threshold = 0.001). Then, based on the knowledge that these reads haven't any mutations we calculated the mean and standard deviation for sequencing errors on each sample separately. Finally, for the target sample, we used threshold = 0.00493, which is the maximum value of the averages plus 3 standard deviations.

3 Results

We had abnormal results of fastqc analysis sections "Per base sequence content" and "Sequence Duplication Levels" all four sequencing results, and in sections "Per sequence GC content" and "Overrepresented sequences" in standard reads. The fuller list of fastqc output is in table 2.

Table 2 Reads quality characteristics

Quality charasteristic	Experimantal data	Reference 1	Reference 2	Reference 3
Per base sequence quality	Normal	Normal	Normal	Normal
Per sequence quality scores	Normal	Normal	Normal	Normal
Per base sequence content	Very unusual	Very unusual	Very unusual	Very unusual
Per sequence GC content	Slightly abnormal	Very unusual	Very unusual	Very unusual
Per base N content	Normal	Normal	Normal	Normal
Sequence Length Distribution	Slightly abnormal	Slightly abnormal	Slightly abnormal	Slightly abnormal
Sequence Duplication Levels	Very unusual	Very unusual	Very unusual	Very unusual
Overrepresented sequences	Slightly abnormal	Very unusual	Very unusual	Very unusual
Adapter Content	Normal	Normal	Normal	Normal

We found 5 high-frequency SNPs, but all of them had led only to samesense mutations, which didn't change the amino acid sequence. Then we elaborated on less-frequency SNPs. We calculated the mean frequency and standard deviation of SNPs in reference samples to differentiate artefacts from rare variants of SNP. Brief information about a number of mapped reads and statistics of reference alignments showed in table 3.

We also computed mean and standard deviation separately for PCR-associated mistakes (identical in all three sets of reads) and sequencing-associated errors. PCR-associated mistakes mean was equal 0.253 with standard deviation 0.037 and sequencing-assocated mistakes mean and sd were 0.260 and 0.096 respectively.

After filtering statistically non-significant SNPs, we added two supplementary mutations to the five ones. Detailed information about all 7 SNPs showed in table 4.

We identified, that there is only one missense mutation from all seven SNPs (in the 307th nucleotide). It substitutes proline for serine in the 103th amino acid residue. Visualization of hemagglutinin original sequence showed in figure 1. SNP from IGV is in figure 2. This mutation leads to changing epitope D of influenza virus hemagglutinin [7].

Table 3 Reads descriptive statistics

	Number of reads before alignment	Number of mapped reads	Mean frequency of errors	SD
Experimental alignment	358265	358032	-	-
Reference 1	256586	256500	0.262037	0.07135824
Reference 2	233327	233251	0.2420408	0.0515582
Reference 3	249964	249888	0.2554386	0.07928408

Table 4 Brief data about SNPs in vaccine escape virus sample

Nucleotide position	Nucleotide substitution	Codone substitution	Amino acid substitution	Amino acid position	Type of mutation	Frequency of mutation
72	A ->G	ACA ->ACG	Thr ->Thr	24	Samesense	99.96%
117	C ->T	GCC ->GCT	Ala ->Ala	39	Samesense	99.82%
307	C ->T	CCG ->TCG	Pro ->Ser	103	Missense	0.95%
774	T ->C	TTC ->TTT	Phe ->Phe	258	Samesense	99.97%
999	C ->T	GGC ->GGT	Gly ->Gly	333	Samesense	99.86%
1260	A ->C	CTA ->CTC	Leu ->Leu	420	Samesense	99.94%
1458	T ->C	TAT ->TAC	Tyr ->Tyr	486	Samesense	0.83%

4 Discussion

4.1 Basic hypothesis

We wanted to understand how were we infected by the influenza virus after vaccination. The main hypothesis was that our roommate's influenza virus had another combination of hemagglutinin and neuraminidase. It was the miscorrected point: roommates strain was A/Hong Kong/4801/2014 (H3N2) when one component of this year's vaccine was A/Darwin/9/2021 (H3N2)-like virus, which resembled our roommate's strain.

Another hypothesis was roommates strain had some mutations, that lead to changing in its epitopes. It may be the cause of why vaccine-stimulated antibodies didn't stop this virus strain. We investigated sequencing data of roommates influenza strain and conclude, all common SNPs are samesense mutations with no epitope changing (table 3, SNPs with more than 95% frequency of mutations). We disallow this hypothesis too.

Our next idea was too seek the genome for any rare mutations, which can promote the appearance of minority fractions of the influenza virus with resistance to vaccine-induced antibodies. It led to another goal - differentiating between sequencing-dependent artefacts and real rare mutations.

4.2 Computing threshold for rare mutations

For resolving the problem from the previous stage we aligned three sets of reads absolutely identical to the reference. All mutations that would appear in these alignments will be only PCR and sequencing errors. After mathematical calculations, described in the results section we established an upper threshold

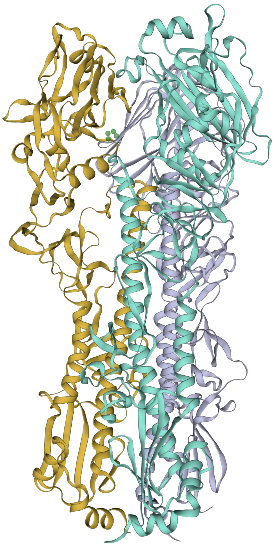

Fig. 1 Non-mutated hemagglutinin vizualisation. Proline in 103th position is highlited.

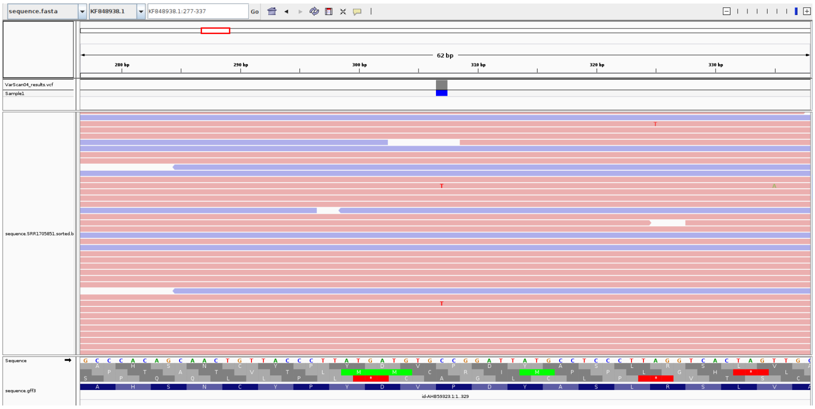
sequence.fasta
KF848938.1
KF848938.1:277-337
Go
62 bp
280 bp
290 bp
300 bp
310 bp
320 bp
330 bp
VarScan04_results.vcf
Sample1
sequence.SRR1705851.sorted.b
T
T
A
T
Sequence
G C C C A C A G C A A C T G T T A C C C T T A T G A T G T G C C G G A T T A T G C C T C C C T T A G G T C A C T A G T T G C
sequence.gff3
A H S N C Y P Y D V P D Y A S L R S L V A
id-AHB59323.1:1..329

Fig. 2 IGV visualization of SNP in the 307th nucleotide

for differentiating rare mutations and sequencing errors as maximal mean plus three standard deviations. This threshold was equal to 0.493%.

The result validated our third hypothesis: we found also two SNPs, one of them was samesense, while another was missense and led to a change in 103th aminoacid position (epitope D). It may derive resistance to vaccine-induced antibodies and cause the illness.

4.3 Another sources of errors in deep sequencing and methods to solve them

The method we used to determine if nucleotide substitution is a rare SNP or an error is only one of the possible ones. There are many potential sources of mistakes in such experiments (contamination from previous reactions[8], himeric reads [9], problems with the sequence context (as example, GC-rich parts of genome and homopolymers) [10], machine failure, user errors and others)[11].

For reducing the level of errors and identifications of deep sequencing defects it may be helpful to use biological, technical and cross-platform replicates.

Another method to decrease the number of deep sequencing mistakes is using some specific software. For an example - CleanDeapSeq. It helps to detect low-quality reads and is functionally equivalent to standard pileup in terms of allele counting [12].

Declarations

- All authors declare that they have no conflicts of interest.

References

[1] Petsch, B., Schnee, M., Vogel, A.B., Lange, E., Hoffmann, B., Voss, D., Schlake, T., Thess, A., Kallen, K.-J., Stitz, L., *et al.*: Protective efficacy of in vitro synthesized, specific mrna vaccines against influenza a virus infection. Nature biotechnology **30**(12), 1210–1216 (2012)

[2] Nachbagauer, R., Krammer, F.: Universal influenza virus vaccines and therapeutic antibodies. Clinical Microbiology and Infection **23**(4), 222–228 (2017)

[3] Behjati, S., Tarpey, P.S.: What is next generation sequencing? Archives of Disease in Childhood-Education and Practice **98**(6), 236–238 (2013)

[4] Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., Loeb, L.A.: Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences **109**(36), 14508–14513 (2012)

[5] Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (2013). https://doi.org/10.48550/ARXIV.1303.3997. https://arxiv.org/abs/1303.3997

[6] Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K.: Varscan 2: somatic

mutation and copy number alteration discovery in cancer by exome sequencing. Genome research **22**(3), 568–576 (2012)

[7] Muñoz, E.T., Deem, M.W.: Epitope analysis for influenza vaccine design. Vaccine **23**(9), 1144–1148 (2005)

[8] Gurr, S.: Pcr protocols-a guide to methods and applications: Edited by m a innis, d h gelfand, j j sninsky and t j white. pp 482. academic press, london 1990. $39.95 isbn 0-12-372181-4. Biochemical Education \mathbf{19}, 45--45 (2010)$

[9] Koboldt, D.C., Ding, L., Mardis, E.R., Wilson, R.K.: Challenges of sequencing human genomes. Brief Bioinform **11**(5), 484–498 (2010)

[10] Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., Kanaya, S.: Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res **39**(13), 90 (2011)

[11] Robasky, K., Lewis, N.E., Church, G.M.: The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet **15**(1), 56–62 (2014)

[12] Ma, X., Shao, Y., Tian, L., Flasch, D.A., Mulder, H.L., Edmonson, M.N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L.L., Levy, S., Easton, J., Zhang, J.: Analysis of error profiles in deep next-generation sequencing data. Genome Biol **20**(1), 50 (2019)