Rare mutation analysis in context of viral mutagenesis

Ekaterina Grigorieva[1, 2] and Gleb Khegai[1, 2]

[1]Bioinformatics Institute
[2]Moscow Institute of Physics and Technology

Abstract

Viral quasispecies are subpopulations of nonidentical but closely related mutant and recombinant viral genomes [1]. We suggest quasispecies can cause vaccinated people getting ill. In our study we investigated the sample of an H3N2 flu strain using the advantages of deep sequencing. We present detected missense single-nucleotide polymorphism (SNP) which changes the epitope and makes virus envade the vaccine.

Keywords: Viral quasispecies; deep sequencing; viral influenza

Introduction

Flu, also known as influenza, is a viral infection affecting parts of respiratory system: nose, throat and lungs. While most people manage to cope with the disease on their own, there are cases, in which the flu can lead to irreversible and even deadly complications [2]. One of the most effective ways to prevent this from happening is to receive a vaccine in advance.

Most influenza vaccine types contain damaged or weakened forms of viruses. This leads to immune system activation by stimulating of specific to the flu antibody production. Due to such preparation, the person's immunity is ready for the viral invasion. However, there are instances when the vaccination does not protect from getting the flu, and viral quasispecies is one of them.

The flu, as a part of a larger group of RNA viruses, exists in the host as a population of closely related but genetically distinct variants called "quasispecies". It was shown that characterization of RNA viruses diversity within one populations is more important for pathogenesis understanding than focusing on any specific viral haplotypes [3].

With the development of sequencing methods interest in studying influenza quasispecies has grown rapidly. Newer next generation sequencing technologies have expedited research in virology, due to their high throughput sequence data production [4]. Nowadays, the deep sequencing method is used for the detection of viral variants.

However, it is easy to make mistakes in finding rare mutations because of sequencing errors. In our study we suggest to solve this problem by filtering errors. Three control sequences were analysed in order to model sequence error pattern. Thus rare mutations could be filtered out of the PCR and sequencing errors.

Materials and methods

In this work we used raw data received from the NCBI Sequence Read Archive [5]. The reference sequence was gotten from NCBI Genbank [6]. The quality of reads was calculated using FastQC [7]. In order to improve the quality of reads we used Trimmomatic tool with the following options: cut bases off the start of a read if their quality was below 20, cut bases off the end of a read if their quality was below 20, trimmed reads with the window size 10 and the average quality within the window 20, dropped reads if they were below length 20 [8]. The alignment and calculating the amount of mapped reads was performed with BWA-MEM and samtools respectively[9,10]. To call variants we used Varscan tool with 0.95 minimum variant frequency for common mutations, 0.001 for rare ones and 0.7388 for trusted rare variants. We visualized received mutations in the IGV browser [11].For calculating the trusted frequency we used python3 and pandas library [12]. The PDB structure of the H3N2 hemagglutinin molecule was received from PDB database [13, 14]. For 3D protein structure visualization we used PyMol [15].

Results

Common variant calling

First, we analysed target sample. The amount of raw and mapped reads for all the samples can be found in Table 1. The FastQC was quite good. There is a slight problem of low quality in the end of reads. Nevetheless, we decided not to use Trimmomatic, becuse in futher analysis there will be errors filtration and analysis of rara variants. So it is important to use initial reads in such subtle research. Then the alignment and variant calling was performed. Common variants are presented in Table 2.

Sequencing errors

After primary analysis of target sample, sequence analysis of three reference lines was carried out. FastQC metrics were similar to those in the target sample from out roommate. In order to analyse sequence and PCR mistakes we decided not to use Trimmomatic as well as in studied sample. Then each of three samples was aligned. With VarScan tool we inspected mutations with frequency higher than 0.1%. All obtained mutations are errors as we considered reference cell lines. Error average and its standard deviation can be found in Table 3. Results are consistent between different technical repetitions. As we know

Table 1 Read number information

Sample	Total reads	Mapped reads
SRR1705851	361349	361116
SRR1705858	256744	256658
SRR1705859	233451	233375
SRR1705860	250184	250108

Table 2 Common and possible qiasispesies variants

Position	Reference	Alternative	Type	Amino acid substitution
72	A	G	SILENT	-
117	C	T	SILENT	-
774	T	C	SILENT	-
999	C	T	SILENT	-
1260	A	C	SILENT	-
307	C	T	MISSENSE	Pro -> Ser
1458	T	C	SILENT	-

some information about error statistics, so we can analyse rare mutation in target sample. If frequency of a mutation is more then average + 3*(standard deviation), we can trust them as type I error will be fewer than 1%.

Rare variants

Then the same analysis was performed for rare variants. You can find them in the Table 2 as well as the common variants.

Epitope mapping

Received missense SNP referring to 103 residue (proline) leads to Epitope D according to already known epitopes [17].

Mutation visualisation

It is important to notice that proline is non-polar amino acid and serine is polar one. This can also give structural disadvantages. We looked at mutated residue in Pymol. Figure 1 and Figure 2 present the structure before and after mutating.

PCR and sequencing error

The analysis for different errors is described in supplementary and presented in Figure 3.

Discussion

In this work we explored rare variants with the help of deep sequencing. As mostly reliable variants were the common ones and did not have significant effect on some molecular features, we found low frequency variant. This could explain the fact, that a person got flu, despite of the fact, that he was vaccinated. We propose, that this happened due to the presence of qiasispecies [18].

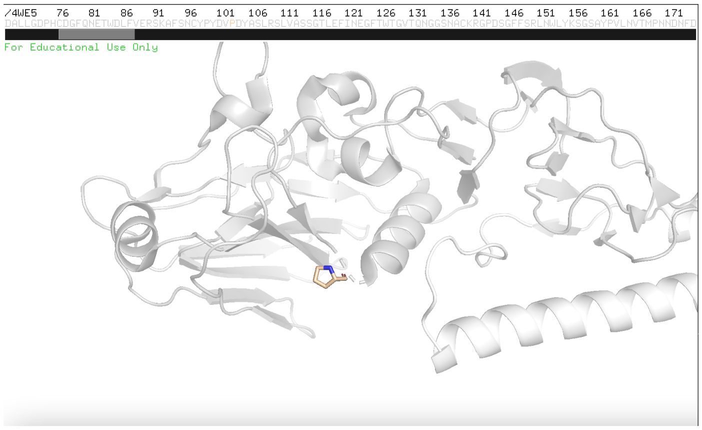
/4WE5 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
DALLGDPHCDGFQNETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEFINEGFTWTGVTQNGGSNACKRGPDSGFFSRLNWLYKSGSAYPVLNVTMPNNDNFD
For Educational Use Only

Figure 1 The original structure of the H3N2 hemagglutinin molecule

The analysis of sequencing data of 3 control cell lines was performed. Several technical replicas allowed us to estimate main statistics for error distributions including average and standard deviation. Error pattern knowledge enabled filtering out sequencing and PCR errors from real rare variants (Table 2). Among them mutation in 307 position. It leads to Proline substitution by Serine. Moreover, this position corresponds to epitope D [17]. So this mutation probably provided the virus with the resistance to vaccine, as immune system antibodies can not recognise the protein [19].

As one of the main issues in variant calling is error filtration we had 3 control lines to model error patterns. Another way to filter out non-existing mutations is to have several technical replicas of the target sample. Comparing between several

Table 3 Average and standard deviation for mutation errors frequency

Sample	Average	Standard deviation
SRR1705858	0.261%	0.071%
SRR1705859	0.239%	0.053%
SRR1705860	0.251%	0.078%
Integrative	0.250%	0.067%

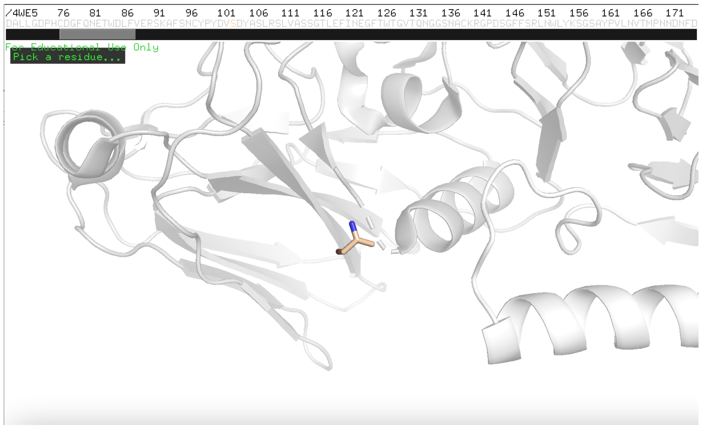
/4WE5 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
DALLGDPHCDGFQNETWDLFVERSKAFSNCYPYDVSDYASLRSLVASSGTLEFINEGFTWTGVTQNGGSNACKRGPDSGFFSRLNWLYKSGSAYPVLNVTMPNNDNFD
For Educational Use Only
Pick a residue...

Figure 2 The original structure of the H3N2 hemagglutinin molecule

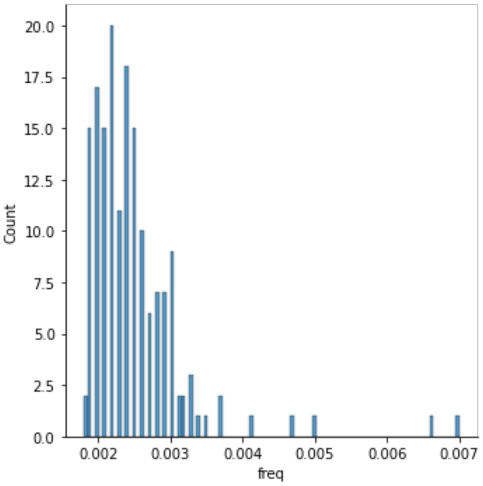

Figure 3 Errors distribution

repetitions allows both to filter out error and to find real existing variants. Apart from this, we can use additional metrics to analyse mutations. Typical signs of sequencing errors are low position coverage, low caller inner metrics, low quality of reads covering the position, presence of several low frequency variant near to each other. Validating on control samples enables dramatically enhance multiparametrical mutation filtration.

In conclusion, we analysed rare mutations of target deep sequenced sample with the help of error filtration based on information regarding sequencing errors from several technical replicas of control samples.

Supplementary data

The code and more details are available at https://github.com/grigorievaekaterina/BI_Project_2.

Acknowledgments

We thank Mike Rayko for supervising and supporting us.

References

1. Domingo E, Perales C. Viral quasispecies. PLoS Genet. 2019 Oct 17;15(10):e1008271. doi: 10.1371/journal.pgen.1008271. PMID: 31622336; PMCID: PMC6797082.

2. Rothberg MB, Haessler SD, Brown RB. Complications of viral influenza. Am J Med. 2008;121(4):258-264. doi:10.1016/j.amjmed.2007.10.040

3. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature. 2006;439(7074):344-348. doi:10.1038/nature04388

4. Isakov O, Bordería AV, Golan D, et al. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. Bioinformatics. 2015;31(13):2141-2150. doi:10.1093/bioinformatics/btv101

5. http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170

6. https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1

7. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

8. Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

9. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

10. Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li. GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

11. Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing Genome Research DOI: 10.1101/gr.129684.111

12. 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). (Free PMC article here).

13. W. McKinney, AQR Capital Management, pandas: a python data analysis library, http://pandas.sourceforge.net

14. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242

15. Yang, H., Carney, P. J., Chang, J. C., Guo, Z., Villanueva, J. M., Stevens, J. (2015). Structure and receptor binding preferences of recombinant human A (H3N2) virus hemagglutinins. Virology, 477, 18-31.

16. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

17. Muñoz, Enrique T, and Michael W Deem. "Epitope analysis for influenza vaccine design." Vaccine vol. 23,9 (2005): 1144-8. doi:10.1016/j.vaccine.2004.08.028

18. Bonomo, M. E.; Kim, R. Y.; Deem, M. W. Modular Epitope Binding Predicts Influenza Quasispecies Dominance and Vaccine Effectiveness: Application to 2018/19 Season. Vaccine 2019, 37 (24), 3154–3158. https://doi.org/10.1016/j.vaccine.2019.03.068.

19. 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. 2019, 20 (1), 50. https://doi.org/10.1186/s13059-019-1659-6.