**Deep sequencing approach in studying influenza virus quasispecies and vaccine ineffectiveness***Evgenia Khokhlova, Eugenia Ivanova (Strebulaeva)*

**Abstract**

The current study concerns the problem of influenza virus vaccine ineffectiveness. Using a deep sequencing approach and variant frequency analysis we found single nucleotide polymorphisms in hemagglutinin gene that could have been associated with viral quasispecies emergence. Analysis of SNPs localization in protein structure showed that mutational processes took place in antigenic epitope regions of viral protein. According to the results we made a suggestion that presence of the quasispecies with identified mutations could have led to the onset of the disease in the vaccinated patient.

**Introduction**

Vaccination is the primary strategy for prevention and control of influenza [1]. Most seasonal vaccines consist of inactivated influenza virus components, which induce antibody responses against immunodominant epitopes in the viral hemagglutinin (HA) and neuraminidase (NA) proteins. [2] The trimeric HA glycoprotein is responsible for attachment of the virus to specific sialic acid-containing proteins on the host cell surface and fusion between the viral and endosome membranes and release of viral nucleic acids into the cytoplasm. The HA protein contains two structural elements: the head (the primary target of antibodies that confer protective immunity to influenza viruses) and the stalk. Both elements differ in their potential utility as vaccine targets. NA is a glycoprotein that removes sialic acid from viral proteins and prevents aggregation of the virus by the HA protein binding to other proteins. Antibodies to NA aggregate viruses on the cell surface, effectively reducing the amount of virus released from infected cells. Immune response to influenza surface glycoproteins is mainly humoral, but CD4+ and CD8+ T cells also play important roles in immunity to influenza. In contrast to the strain-specific response of antibodies, cell-mediated immunity tends to be more cross-reactive among viral subtypes, recognizing more conserved epitopes on the surface proteins and internal viral proteins. However, influenza viruses as any RNA viruses often undergo genetic changes that lead to gradual antigenic changes in both HA and NA because of their error-prone polymerases. This process is known as antigenic drift that leads to the emergence of new variant strains [1]. The changes associated with antigenic drift happen continually over time as the virus replicates. The small changes that occur from antigenic drift usually produce viruses that are closely related to each other and usually have similar antigenic properties. However, the small changes associated with antigenic drift can accumulate over time and result in quasispecies viruses with a different genome variant that are antigenically different [3]. It is also possible for a single (or small) change in a particularly important location on the HA to result in antigenic drift. When antigenic drift occurs, the immune system may not recognize and prevent sickness caused by the newer influenza viruses. As a result, a person becomes susceptible to flu infection again, as antigenic drift has changed the virus enough that a person’s existing antibodies won’t recognize and neutralize the newer influenza viruses. The deep sequencing approach usually used for investigation of viruses’ genomes because of its small size and fast evolution. It can detect low frequency variants is an important feature of deep sequencing, for example in the context of drug or vaccine resistance [4].However, it’s important to distinguish low frequency variants in the genome from the PCR and sequencing errors. Sequencing errors are key confounding factors for detecting low-frequency genetic variants that are important for viral strain investigation using deep next-generation sequencing (NGS) [5]. So the goal of this project is to figure out how the flu is spreading despite the vaccines action on the example of the H3N2 strain called A/Hong Kong/4801/2014 (H3N2) covered by trivalent vaccines for use in the 2017-2018 northern hemisphere influenza season contain the following an A/Michigan/45/2015 (H1N1)pdm09-like virus, an A/Hong Kong/4801/2014 (H3N2)-like virus and a B/Brisbane/60/2008-like virus.

**Methods**

**Description of the data and quality control**

In this study we operated with sequencing data from the NCBI Sequence Read Archive. We used a hemagglutinin gene sequence from NCBI GenBank as reference (KF848938.1).

Sample:

[ftp://](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz)[ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz](http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz)

Controls:

[ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz](http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz)

[ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz](http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz)

[ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz](http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz)

To assess the quality of the data a fastqc program was used. The overall quality of the reads was good. We observed overrepresented sequences and high level of duplication but it is normal for deep sequencing data. The reads were pre-processed as we saw no adapters and no sufficient decrease of per base sequence quality on the graphs. There were 151 cycles in the sequencing run.

**Alignment of the sequencing data**

Alignment was performed using a bwa-mem tool. We used a samtools program to convert SAM format to BAM format and sort the BAM files.

We calculated approximate coverage by the formula:

Coverage = (mapped\_reads \* read\_length) / reference\_length

**Variant calling**

To find SNPs in our data we run a VarScan program. We set the -d parameter of samtools 33000 while we were creating a mpileup file to find all variants. To observe rare variants with VarScan we set minimum variant frequency 0.1%. For parsing vcf files awk program was used.

**Separation of the rare variants and sequencing errors**

We used statistical approach to find rare variants. Calculations were executed using R version 3.6.3 and dplyr package.

**Visualization**

For visualization of the hemagglutinin molecule we used PyMOL. The crystal structure was taken from RCSB PDB (4WE4).

**Results**

Number of mapped reads after alignment and approximate coverage are shown in table 1.

Table 1 - Number of reads before and after mapping.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Total number of reads | Mapped reads | Coverage |
| Roommate | 358265 | 358032 | 32470 |
| Control\_1 | 256586 | 256500 | 23262 |
| Control\_2 | 233327 | 233251 | 21153 |
| Control\_3 | 249964 | 249888 | 22662 |

Variants in control samples and their frequencies obtained after variant calling is shown in table 3 in Supplementary materials. Some statistics calculated for the frequencies can be found in table 2. Variants with frequencies more than 90% were not used in calculations because those were more likely genuine mutations and not the errors.

Table 2 - Statistics of variant frequencies in control samples.

|  |  |  |
| --- | --- | --- |
| Sample | mean | sd |
| Control\_1 | 0.261090909090909 | 0.0705710235172295 |
| Control\_2 | 0.238846153846154 | 0.0527165504448024 |
| Control\_3 | 0.250983606557377 | 0.077903892030133 |

Using the statistics in table 2 we standardized frequencies of variants for roommate’s sample to distinguish sequencing errors and rare variants. In table 3 we list variants considered as real mutations and information about epitope localization (only variants with Z scores more than 3 are listed, all variants are shown in table 5 in Supplementary materials). Rare variants suggested to belong to quasispecies are shown in red. Visualization of the mutated amino acid and epitope D is displayed in figure 1.

Table 3 - Results of variant calling for roommate’s sample with standardized values of variant frequencies.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Position in gene | Nucleotide change | Amino acid change | Frequency | Frequency Z scores | Epitope |
| 72 | ACA > ACG | Threonine > Threonine, synonymous | 99.96 | 1486.788 | None |
| 117 | GCC > GCT | Alanine > Alanine, synonymous | 99.82 | 1484.701 | None |
| 307 | CCG > TCG | Proline > Serine, non-synonymous | 0.96 | 10.582 | D |
| 774 | TTT > TTC | Phenylalanine > Phenylalanine, synonymous | 99.97 | 1486.937 | None |
| 999 | GGC > GGT | Glycine > Glycine, synonymous | 99.86 | 1485.297 | None |
| 1260 | CTA > CTC | Leucine > Leucine, synonymous | 99.94 | 1486.49 | None |
| 1458 | TAT > TAC | Tyrosine > Tyrosine, synonymous | 0.83 | 8.644 | None |

**

**Figure 1.** Hemagglutinin molecule (HA1 and HA2 chains). Epitope D is shown in space-filling format and Proline103 is shown in magenta color.

**Discussion**

Amino acid mutations in the surface protein, hemagglutinin (HA), of influenza viruses allow these viruses to circumvent neutralization by antibodies induced during infection or vaccination. So, in our project we have found a mutation with low frequency in the 307 genome position in epitope D. This indicates the likelihood of quasispecies emergence. The mutation leads to amino acid change from Proline to Serine in these quasispecies which could be affecting the affinity to antibodies. This change in the structure of the epitope could lead to the onset of the disease even despite the presence of vaccination. Also, we have found several synonymous mutations which indicates a high rate of mutation of the influenza virus. We discovered this rare mutation by comparing standardized frequency values with controls. Control samples had a normal distribution of errors and this allowed us to consider variants with frequencies higher than three standard deviations as true mutations. However, it is very important to be able to correctly distinguish between sample preparation and sequencing errors from true rare mutations. Because it is very easy to confuse them with each other and accept a false positive or false negative result, making a type I error. The best way to control error rate in NGS results is to replicate experiments, however it is not always possible because of different reasons, that’s why scientists usually use a bioinformatic approach of filtering the sequencing data, base call quality approach and short-read alignment quality of reads. All of these post-processing techniques help to reduce uncertainty in the final genotyping variant call [7]. Success on any next-generation sequencing platform begins with optimal sample preparation- from sample isolation and purification to library construction and enrichment. For every of these steps we should use extra quality DNA probes and reagents. If the DNA will be low quality or at low concentration or contaminated, we couldn’t perform any further analysis no matter how good our sequence technology is. Also, correct sample dilution prior to sequencing avoids two different molecules falling into the same spot prevents errors. To reduce error rate in our data we can also try to use error correction tools (EC tools) that identify and correct the sequencing errors. This is achieved by generating a *k*-mer coverage spectrum from the input data and replacing poorly covered (and hence likely erroneous) *k*-mers by similar *k*-mers with a higher coverage. Sometimes, this process is further guided by using the per-base quality scores [8].

**Conclusion**

Deep sequencing approach is a good method for studying influenza virus quasispecies because of its great coverage which allows you to determine whether SNP is an error or real mutation. But you always should make control replicates of your experiment to be sure. Viruses change every replicative cycle so it’s hard to create a vaccine that will be suitable for a long period of time. Our results showed that mutational processes took place in the epitope region of hemagglutinin which can be considered as antigenic drift and lead to vaccine ineffectiveness as a consequence of lower affinity to antibodies generated by the vaccine. That’s why scientists make a new vaccine every flu season, so it helps us to prepare our immune system for different kinds of strain.

**Optional Extra-Credit Challenge Question**

**1) To calculate actual coverage per position samtools depth can be used through the command line (combined with some other programs could be used to calculate the maximum coverage):**

samtools depth -d 0 [file.bam] | cut -f3 | sort -rn | head -1

Here samtools depth calculates real coverage in each position in file.bam;

-d 0 option removing the depth limit (that is 8000 by default)

cut -f3 choose the third field in stdin file

sort -rn sort coverage values from biggest to lowest

head -1 shows the first line - the biggest coverage number

**To calculate average coverage we can use samtools with awk:**

samtools depth -d 0 [file.bam] | awk ‘{sum+=$3} END { print sum/NR}’

Here awk makes a calculation of the mean for values in third field of the stdin file

**2) Using positions reported by VarScan in all 3 of the reference sequences, can you distinguish PCR (“upstream”) and sequencing (“during”) errors? Provide average and standard deviation for both types of error.**

From the literature data we suggest that for the PCR samples, the most common base substitutions are those representing transitions between the purines (A to G and G to A) and pyrimidines (T to C and C to T), which are dominant within our dataset [9].

Also we suggest that Illumina errors more often occurred at G positions and the remaining errors were distributed equally between the A, C and T positions. This suggests that the Illumina machine is more prone to error at G positions and reflects the known issues with GGC and GGX motifs which are prone to signal to noise decline, and are hard to be sequenced [10].

Table 4 - The frequencies of PCR and sequencing errors

|  |  |  |
| --- | --- | --- |
|  | Average | Standard Deviation |
| PCR errors | 0.2530967742 | 0.06927798082 |
| Sequencing errors | 0.3166666667 | 0.01527525232 |

**3) If you are familiar with the** [**PDB database**](http://www.rcsb.org/pdb/home/home.do)**, you can try to explore VMD, PyMOL, Jmol, RasMol, or some other PDB-viewing application to provide an image of the H3N2 hemagglutinin molecule and highlight amino acid changes you’ve found.**

(in results)

**References**

1. Gomez Lorenzo, M.M.; Fenton, M.J. Immunobiology of influenza vaccines. *Chest* 2013, *143*, 502–510, doi:10.1378/chest.12-1711.

2. Kaiser, J. A One-Size-Fits-All Flu Vaccine? *Science* 2006, *312*, 380–382, doi:10.1126/science.312.5772.380.

3. Webster, R.G.; Govorkova, E.A. Continuing challenges in influenza: Continuing challenges in influenza. *Ann. N.Y. Acad. Sci.* 2014, *1323*, 115–139, doi:10.1111/nyas.12462.

4. McElroy, K.; Thomas, T.; Luciani, F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Informatics Exp* 2014, *4*, 1, doi:10.1186/2042-5783-4-1.

5. Ma, X.; Shao, Y.; Tian, L.; Flasch, D.A.; Mulder, H.L.; Edmonson, M.N.; Liu, Y.; Chen, X.; Newman, S.; Nakitandwe, J.; et al. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* 2019, *20*, 50, doi:10.1186/s13059-019-1659-6.

6. Muñoz, E.T.; Deem, M.W. Epitope analysis for influenza vaccine design. *Vaccine* 2005, *23*, 1144–1148, doi:10.1016/j.vaccine.2004.08.028.

7. George M. Church, K.R. The role of replicates for error mitigation in next-generation sequencing. *Nat Rev Genet 15, 56–62* 2014, doi:https://doi.org/10.1038/nrg3655.

8. Heydari, M.; Miclotte, G.; Demeester, P.; Van de Peer, Y.; Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* 2017, *18*, 374, doi:10.1186/s12859-017-1784-8.

9. Orton, R.J.; Wright, C.F.; Morelli, M.J.; King, D.J.; Paton, D.J.; King, D.P.; Haydon, D.T. Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **2015**, *16*, 229, doi:10.1186/s12864-015-1456-x.

10. Abnizova, I.; Boekhorst, R. te; Orlov, Y.L. Computational Errors and Biases in Short Read Next Generation Sequencing. *J Proteomics Bioinform* **2017**, *10*, doi:10.4172/jpb.1000420.