Why did I get the flu?

Authors: Daria Nikanorova, Anna Rybina

Abstract

Influenza virus evolves extremely fast, which poses a threat even to vaccinated people. One of promising approaches to investigate influenza's subspecies is deep sequencing. High level of coverage allows to detect rare mutations. However SNPs must be accurately distinguished from technical errors during library preparation and sequencing. In this study we applied deep sequencing analysis to find rare mutations, presumably responsible for avoidance of immune response after vaccination. We found a mutation (C/T) in the 307 position of hemagglutinin (HA) influenza's gene that causes non-synonymous amino acid substitution in epitope D of HA protein. We propose that this mutation may alter immunogenic epitope, which in turn impairs antibody recognition and binding to its target.

Introduction

The high rate of evolution of the influenza virus challenges vaccinating against it. Rapid *antigenic drift* is forced by influenza's RNA polymerase that lacks proof-reading activity and makes mistakes every replication cycle with high rate. *Antigenic shift* occurs when genetic segments from different strains or species are rearranged and mixed up inside host cells (Kim, Webster, and Webby 2018). According to quasispecies theory, viral populations are heterogeneous and consist of different subpopulations inside one host (Domingo 2006). These small subpopulations evolve independently and obtain different mutations while replicating in unvaccinated people. Some of these mutations may occur in immunogenic epitopes of proteins that serve as targets for antibodies, produced by immune cells after vaccination. As a result, this influenza subspecies can not be detected by the immune system, thus posing a threat even to vaccinated people. Low number of viral subunits in a whole population within a host results in low frequencies of nucleotide variants in sequencing samples. One possible way to identify SNPs is deep sequencing. This approach provides a very high level of coverage and accurate detection of minor alleles and mutations from heterogeneous samples (Flaherty et al. 2012). However, distinguishing real rare mutations from errors occurred during library preparation and sequencing challenges deep sequencing analysis (Orton et al. 2015). Most of approaches aim to eliminate errors by using technical replicates or employing different computational models (most of them are based on binomial distribution) (Robasky, Lewis, and Church 2014). This study is aimed to identify the mutations in hemagglutinin (HA) influenza's gene that may reflect occurrence of a new subspecies, which may stay unnoticed by antibodies produced after vaccination.

Methods

One amplicon of hemagglutinin (HA) was obtained from an individual patient infected with influenza A H3N2. Three control amplicon replicates were generated in the following way: sample of the isogenic reference influenza A H3N2 strain was PCR amplified, cloned into plasmid. Single purified plasmid containing HA gene, was RT-PCR amplified. Both experimental and three control libraries were indexed, pooled together in equal

concentrations and sequenced on Illumina MiSeq in a single-end mode with 151 cycles. Raw reads were trimmed to remove adapter sequencing and binned by indexes to separate fastq files. Processed sequencing data of experimental and control samples in fastq format were downloaded from SRA database (run accession numbers: SRR1705851 and SRR1705858 - SRR1705860, respectively). RVD_H3/2011_H3N2 reference sequence of HA gene was downloaded from NCBI (GenBank accession number KF848938.1).
Quality of reads were checked using FastQC v0.11.9 (Andrews, S). To align reads to the reference genome, we used BWA tool v0.7.17-r1188 (Li and Durbin 2009)) (Supplementary materials, section 2 and 5). Reference genome was indexed with default parameters. Reads were mapped implementing the BWA-MEM algorithm with default parameters. Resulting alignment in SAM format was compressed by converting it to BAM format, then sorted and indexed using SAMtools v1.9 (Li et al. 2009). Basic statistics, a fraction of mapped/unmapped reads and the coverage of sequencing was estimated manually and using SAMtools (commands view -f4, flagastat, stats, idxstats, coverage commands, see Supplementary materials, sections 2, 3 and 5). Pileup files were created using SAMtools (mpileup command). Maximum depth (-d) option was set to the value higher than estimated average coverage of sequencing to catch rare SNPs (Supplementary materials, section 3 and 6).
SNP calling was performed using VarScan v2.4.1 (mpileup2snp command) (Koboldt, Larson, and Wilson 2013) (Supplementary materials, section 4 and 6) and different thresholds for minimum variant frequency (--min-var-freq option) were specified depending on which kind of variant were searched. To find common variants in the clinical sample, a minimum variant frequency of 95 % was used. For detecting rare variants in all samples, the parameter was lowered to 0.001 (0.1%). To distinguish rare mutations from an error, we determined the variance introduced by PCR amplification, the library preparation and Illumina sequencing based on control samples. VarScan was run on each control sample with --min-var-freq 0.001 and then mean frequency and standard deviation were calculated using either R or python (Supplementary materials, section 7). Average values of mean frequency (av_freq) and standard deviation (av_std) throughout three control samples were calculated. Assuming that rare genetic variants detected in reference control samples are errors, we set the interval of possible frequencies associated with true genetic mutations as following: < (av_freq - 3*av_std) or > (av_freq + 3*av_std). If the frequency of variants in the clinical sample fell into a specified interval (are more than 3 standard deviations away from averages of reference data) then the variant was considered as true genetic mutation. Selected variants from the clinical sample were analyzed in the IGV browser (Robinson et al. 2011). Epitope mapping was performed referring to the work (Muñoz and Deem 2005).

Results

Alignment

In this study, Illumina deep sequencing data on amplicons of HA from influenza A H3N2 were analyzed: one clinical and three control isogenic samples. The reference sequence of the HA gene was chosen correctly as only about 0.06 % of reads from the clinical sample were not mapped (**Table 1**). Estimated coverage of sequencing varied from about 30 000 up about 45 000 across all samples.

Table 1. Number of reads in analysis

Sample	# of reads before mapping	# of mapped reads	Estimated coverage
SRR1705851 (clinical)	358265	358032	45176
SRR1705858 (control 1)	256586	256500	32585
SRR1705859 (control 2)	233327	233251	29631
SRR1705860 (control 3)	249964	249888	31745

Variant calling

While searching for rare variants in the clinical sample, we should set a criteria to distinguish them from errors introduced during PCR amplification, library preparation and Illumina sequencing. To take it into account, we analyzed control samples, assuming that rare SNPs in the isogenic reference data might be errors (**Table 2**). We determined the interval of frequencies associated with false-positive mutations as following:
[av_freq - 3*av_std; av_freq + 3*av_std] = [0.249 - 3*0.067; 0.249 - 3*0.067]
where av_freq and av_std are average frequency and average standard deviation of rare variant frequencies across all control samples. In other words, if a variant frequency lies from the control sample stands more than 3 standard deviations away from averages of reference data, then the variant might be a true genetic mutation.

Table 2. Average and standard deviations of variant frequencies in control samples.

Sample	Mean	Std
SRR1705858	0.257	0.071
SRR1705859	0.237	0.052
SRR1705860	0.251	0.078
Average	0.249	0.067

Based on this criteria, we identified 5 common SNPs with minimal frequency of 95 % in the clinical sample (**Table 3**). These genetic alterations resulted in synonymous mutations and therefore probably could not impair the antigen recognition by antibodies derived from the vaccine (see also **Figure S3 - S7**). Moreover, we revealed 2 rare SNPs in the experimental sample which might be true mutations, among them only one substitution resulted in non - synonymous mutation in protein (Proline was replaced by Serine at position 103: P103S) (**Table 3**).

Table 3. Common and rare genetic mutations, identified in the clinical sample

Position	Reference	Alternative	Frequency,%	Mutation	Type
72	A	G	99.96	T24T	synonymous

117	C	T	99.82	A34A	synonymous
307	C	T	0.94	P103S	non - synonymous
774	T	C	99.96	F258F	synonymous
999	C	T	99.86	G333G	synonymous
1260	A	C	99.94	L420L	synonymous
1458	T	C	0.84	T485T	synonymous

According to the result of epitope mapping, identified rare mutation P103S is located within the epitope D region (Muñoz and Deem 2005; Cushing et al. 2015)).

We inspected the structure of HA protein of influenza A (strain A/Victoria/361/2011(H3N2)). According to our analysis, the mutation P103S occurred within the loop, preceding α-helix. Proline is considered to be very rigid amino acid residue and might create a fixed kink in a protein chain. Its substitution by Serine could cause critical conformational changes in protein resulting in reduced recognition by antibodies produced by immune cells after vaccination.

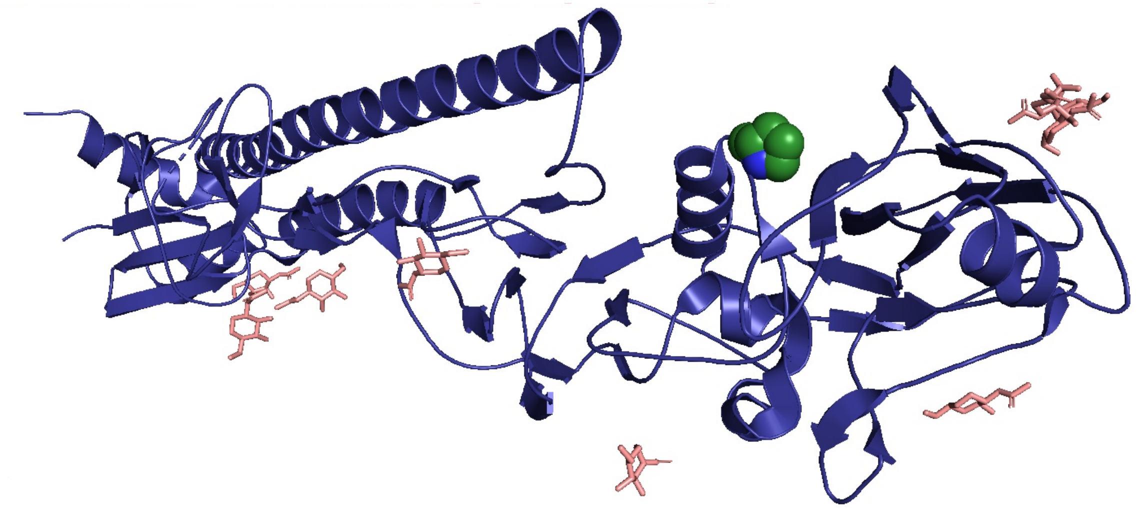

Figure 1. The crystal structure of hemagglutinin of influenza virus A/Victoria/361/2011(H3N2) with 2-acetamido-2-deoxy-beta-D-glucopyranose as a ligand (PDB ID: 4WE8 (Yang et al. 2015)). Missense mutation revealed in the experimental sample in our study occurred at the position 103 where Proline was substituted with Serine. Proline at the position 103 is highlighted and its atoms are depicted as spheres. Image was created using PyMol.

Discussion

A single mutation makes vaccine helpless

In this study we performed an analysis of deep sequencing data of the influenza hemagglutinin gene (strain A/Hong Kong/4801/2014, H3N2). Although this strain was covered by the vaccine, it remained infectious and caused flu in vaccinated people. After alignment of reads we found 5 SNPs with frequency of 95% and high. We found that all of them lead to synonymous substitutions in a HA protein and can not alter epitope recognition by antibodies.

We focused next on SNPs with lower frequency (> 0.01 %). The idea was that rare mutations may reflect heterogeneity of the viral population, which consists of different subpopulations, so called quasispecies (Domingo 2006). We suspected that a small subpopulation evolves while replicating in unvaccinated people, thus obtaining some mutations in immunogenic epitopes. These mutations are thought to be rare, as far as only a small part of a viral population hold them. But if a vaccinated person is infected by viruses of this small subpopulation, it will lead to a disease, as far as the mutated epitopes of quasispecies can not be recognized by antibodies. Deep sequencing analysis with coverage ~ 35000 reads per base allows us to highlight 21 rare variants. In order to differentiate real mutations from PCR and sequencing errors, we used three technical replicates of a control sample with known sequence (SRR1705858, SRR1705859, SRR1705860). We found that the mean frequency of SNVs in all three replicates is equal to 0.249 with average standard deviation of 0.067. We applied the statistical idea that real mutations in our sample must not belong to the same distribution. Instead they must lie more than 3 standard deviations away from the averages. Two rare SNPs satisfy this condition: non-synonymous substitution in 307 position (C/T) and synonymous substitution in 1458 position (T/C). The first one causes an amino acid substitution P103S in the epitope D of the HA protein in influenza virus A (Muñoz and Deem 2005). Interestingly, no mutations in the epitope D were found in A/Fujian/411/2002 strain, while the same substitution P103S of the epitope D was found in recent study (Muñoz and Deem 2005; Cushing et al. 2015). We assume that this substitution P103S of the epitope D may spoil antibodies recognition and binding thus averting immune defence. Our results explain how a vaccinated person is able to get the flu from an unvaccinated one, who serves as a reservoir for antigenic drift of influenza virus.

Fatal errors

Errors during sample preparation (RNA/DNA degradation, contamination or low input of nucleic acid), library preparation (PCR amplification errors, primer biases, adaptor errors) and sequencing (dephasing during sequencing homopolymeric regions, damaged fluorophores) may spoil further detection of low-frequency genetic variants (Robasky, Lewis, and Church 2014). One possible way of reducing technical errors is proposed by Patrick Flaherty and colleagues. This method detects rare SNVs by comparing the baseline error rate from multiple reference replicates to the sample error rate at each position (Flaherty et al. 2012). Firstly, reference and sample DNA are independently prepared and tagged with adapters. Moreover, the amplification of the replicates is carried out by polymerases with a high fidelity (e.g Phusion) and a small number of amplification cycles that reduce the PCR-induced errors. After preparation both libraries are sequenced on the same lane. A baseline error rate for each position is computed by obtaining error rate distribution from replicates by a Beta-Binomial model. Finally, the error rate of the sample SNPs is compared to the baseline error rate to call rare variants. According to authors this Beta-Binomial algorithm exceeds other popular approaches such as Genome Analysis Toolkit (GATK) and SAMtools.

Apart from technical biases sequence context and nucleotide substitution types affect error rate. According to a recent study, A>G/T>C changes occur with the highest rate among all substitution types (10-4) (Ma et al. 2019). Our results confirm this observation, as far as all errors found in replicates were of A>G/T>C type.

Optional: calculation of actual average coverage per position

Depth of feature coverage for each base for the given sorted bam file might be computed using bedtools genomecov (v2.27.0) with the specified -d option. Output of this command consists of 3 columns: reference sequence, reference sequence position and number of features (reads) overlapping this reference sequence position. Output of genomecov command is pipelined to awk command which summarizes all values from the third column and then print the average coverage (sum of mapped reads number which divided by the length of reference sequence):

```
genomeCoverageBed -ibam SRR1705851.sorted.bam -d | awk -F '\t' '{sum +=
$3} END {print sum/NR}'
31212.7

bedtools genomecov -ibam SRR1705851.sorted.bam -d | awk -F '\t' '{sum +=
$3} END {print sum/NR}'
31212.7
```

References

Cushing, Anna, Amanda Kamali, Mark Winters, Erik S. Hopmans, John M. Bell, Susan M. Grimes, Li C. Xia, et al. 2015. “Emergence of Hemagglutinin Mutations During the Course of Influenza Infection.” *Scientific Reports* 5 (November): 16178.

Domingo, Esteban. 2006. *Quasispecies: Concept and Implications for Virology*. Springer Science & Business Media.

Flaherty, Patrick, Georges Natsoulis, Omkar Muralidharan, Mark Winters, Jason Buenrostro, John Bell, Sheldon Brown, Mark Holodniy, Nancy Zhang, and Hanlee P. Ji. 2012. “Ultrasensitive Detection of Rare Mutations Using next-Generation Targeted Resequencing.” *Nucleic Acids Research* 40 (1): e2.

Kim, Hyunsuh, Robert G. Webster, and Richard J. Webby. 2018. “Influenza Virus: Dealing with a Drifting and Shifting Pathogen.” *Viral Immunology* 31 (2): 174–83.

Koboldt, Daniel C., David E. Larson, and Richard K. Wilson. 2013. “Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection.” *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi1504s44.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Ma, Xiaotu, Ying Shao, Liqing Tian, Diane A. Flasch, Heather L. Mulder, Michael N. Edmonson, Yu Liu, et al. 2019. “Analysis of Error Profiles in Deep next-Generation Sequencing Data.” *Genome Biology* 20 (1): 50.

Muñoz, Enrique T., and Michael W. Deem. 2005. “Epitope Analysis for Influenza Vaccine Design.” *Vaccine* 23 (9): 1144–48.

Orton, Richard J., Caroline F. Wright, Marco J. Morelli, David J. King, David J. Paton, Donald P. King, and Daniel T. Haydon. 2015. “Distinguishing Low Frequency Mutations from RT-PCR and Sequence Errors in Viral Deep Sequencing Data.” *BMC Genomics* 16 (March): 229.

Robasky, Kimberly, Nathan E. Lewis, and George M. Church. 2014. “The Role of Replicates for Error Mitigation in next-Generation Sequencing.” *Nature Reviews. Genetics* 15 (1): 56–62.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Yang, H., P. J. Carney, J. C. Chang, Z. Guo, J. M. Villanueva, and J. Stevens. 2015. “The Crystal Structure of Hemagglutinin of Influenza Virus A/Victoria/361/2011.” https://doi.org/10.2210/pdb4we8/pdb.