BIOINFORMATICS INSTITUTE

Application of deep sequencing methods for inferring quasispecies of influenza virus and revealing vaccine-resistant mutations

Ekaterina Sukhinina[1, 2, †] and Nikita Vyatkin[1, 3, †]

[1]Bioinformatics Institute, Saint-Petersburg, Russia
[2]Siberian State Medical University, Tomsk, Russia
[3]Saint-Petersburg State University, Russia

[†]Contributed equally.

Abstract

Next-generation sequencing (NGS) has been successfully applied to analyze viral quasispecies of high diversity. A number of low-frequency drug- or vaccine-resistant mutations with therapeutic importance have been discovered. Despite the influenza virus population diversity is well covered, not much information exists on the influenza virus vaccine resistance due to the presence of a resistant quasispecies in the viral population. Here we identified low-frequency mutations in the vaccine-resistant viral variant A/Hong Kong/4801/2014 strain, separated real mutation from Illumina sequencing errors and examined its effect on the structure of the viral hemagglutinin protein. The detected low-frequency single nucleotide polymorphism (SNP) leads to a Pro/Ser substitution in an epitope region of hemagglutinin. This mutation the most likely connected with vaccine resistance of the examined viral variant.

Key words: Next Generation Sequencing (NGS); Deep Sequencing; Influenza Virus; Quasispecies; Rare variants.

Introduction

Influenza vaccine conveys immunity against the influenza virus by stimulating the production of antibodies toward two types of antigens: hemagglutinin (HA) and neuraminidase (NA). Influenza A virus has 18 HA and 11 NA subtypes, and these antigens are critical for the virulence of the organism [1]. Antigens are combined differently within the virus structure. And further, viral proteins can be altered through random mutations or by reassortment [2].

RNA viruses, including influenza viruses, are characterised by an error-prone replication mechanism. Antigenic drifts are genetic changes occurring in the virus due to various actions of polymerases. It leads to gradual antigenic changes and new variant strains production[3]. The term "viral quasispecies" refers to the fact that RNA viruses exist as heterogeneous populations of closely related genetic variants [4]. Different closely related variants of a highly mutant Influenza A virus can combine within the same host. Deep sequencing systems, such as Illumina, produce thousands of reads for a given location in the genome. This method makes it possible to detect the viral genome diversity within the single host [5].

Methods

Data for analysis

We analyzed data obtained by targeted deep sequencing of HA genes in the influenza virus sample. For this, an Illumina single-end sequencing run was organized. The results were published in the NCBI Sequence Read Archive (SRA) and labeled SRR1705851 [6]. The reference sequence for the influenza A hemagglutinin gene got from NCBI too (strain A/USA/RVD1_H3/2011(H3N2) used) [7].

Aligning reads to reference

Alignment of reads with the reference gene was performed using BWA tools v. 0.7.17-r1188 [8]. First of all, the reference sequence was indexed by the Burrows-Wheeler transform via `bwa index` tool. Then reads were aligned (`bwa mem`), and obtained results are compressed and sorted for further quicker analysis by samtools v. 1.6 (`samtools view`, `samtools sort`, `samtools index`) [9]. All of the utilities listed above were used with standard parameter settings.

Compiled on: November 11, 2022.
Draft manuscript prepared by the author.

Variant calling

To identify common variants (single nucleotide polymorphisms (SNPs)) in our data we used *samtools mpileup* and *varscan mpileup2snp* (VarSacn v2.4.4 [10]). The minimum variant allele frequency threshold was equal to 0.95. As shown in the results, all discovered SNPs don't lead to any missense mutation in produced protein. So we looked up rare variants using a smaller threshold for variant frequency (0.001). Samtools' default behavior is to stop piling up base calls at each location after it reaches 8000 calls to save computational power. Given the potential rarity of our variants, we set the depth limit (using the -d flag) to $MaxDepth = 45000$, which is greater than the nucleotide coverage of every reference gene nucleotide.

Since *MaxDepth* directly affects memory consumption, it should not be too large. There is exists two ways to compute the appropriate *MaxDepth* value. The first way is to estimate average coverage by reads in reference gene. Since we know gene length $L = 1665$ bp, total number of aligned reads $N = 361116$ (can be obtained with the *samtools flagstat*) and maximal lengths of reads $l = 151$ (can be obtained with the FastQC [11], see supplementary materials 5), one can estimate *AverageCoverage*:

$$AverageCoverage \leq \frac{N \cdot l}{L}$$

Considering this, one can use $MaxDepth = w \cdot AverageCoverage$ (where $w > 1$, for example, $w = 1.5$ or 2) hoping that there are no positions in the reference covered by more than $w \cdot AverageCoverage$ reads.

Another approach allows us to get the exact value of *MaxDepth*. We can use *samtools depth* to ask the computer to calculate coverage for each position in the gene. After that, it remains to find the base in the reference genome covered by the largest number of reads and take the corresponding value as an *MaxDepth*. In our case, it equals 44522, but the value 45000 was chosen for ease of perception.

Control data processing

By sequencing an isogenic viral sample derived from a virus clone that matches the reference sequence, we were able to distinguish low-frequency modifications from background noise and sequencing errors. We can identify the real variants by looking at the errors in this control data and comparing them to the mutations in our sample.

An isogenic sample of the standard (reference) H3N2 influenza virus was sequenced three times on an Illumina machine. One can download fastq data for the three controls groups of reads from SRA FTP: SRR1705858, SRR1705859 and SRR1705860.

The alignment and variant calling (with a minimal variant frequency of 0.001) of the control groups were completely similar to the processing of the original data.

Determination of low-frequency mutations

Any "mutations" found in the control samples must be errors since they don't have any real genetic variants. As a result, we can determine what in our data is an error and what is a true variant using the frequency of the errors from the control. We computed the average and standard deviations for the frequencies in each control group to achieve this. After that, variants in studying alignment with frequencies higher than 3 standard deviation away from the averages in the control group, were accepted as real low-frequency mutations.

Results

The initial data consist of 358,265 reads, most of which were 150-151 bp long (see the FastQC report 5) with the reference HA gene of 1,665 bp. Information about control data sets is presented in Table 1. We can see, that number of aligned reads is a bit greater than number of all reads in each case. This is because, due to the very short reference sequence, some of the reads align with equal success in several places.

Table 1. Number of reads

	Studied sequences	Control data sets			
	SRR1705851	SRR1705858	SRR1705859	SRR1705860	Total
Raw data	358 265	256 586	233 327	249 964	739 877
Aligned	361 116	256 658	233 375	250 108	740 141

A total of 5 high-frequency mutations were found (see Table 2), but all of these are synonymous.

Table 2. Detected common variants (allele frequency is greater than 0.95)

Location, bp	Base changes	Codon changes	Amino acid
72	A → G	ACA → ACG	Thr (T)
117	C → T	GCC → GCT	Ala (A)
774	T → C	TTT → TTC	Phe (F)
999	C → T	GGC → GGT	Gly (G)
1 260	A → C	CTA → CTC	Leu (L)

The average and standard deviation of the frequencies from each control data sets are presented in Table 3.

Table 3. Statistics, computed for rare "mutations" in control data sets

	SRR1705858	SRR1705859	SRR1705860
Number of rare variants	57	52	61
Average frequency, %	0.256	0.237	0.250
Standard deviation, %	0.072	0.052	0.078

In the end, we identified two variants as true mutations, but not noise or sequencing errors (see Table 4).

Discussion

Genetic diversity within the infectious pathogen population may be associated with disease progression, vaccine ineffectiveness, and treatment failure. The low fidelity of the influenza virus polymerase, which lacks classic proofreading mechanisms, is responsible for the continuous production of mutated viral copies. The mutant spectrum known as viral quasispecies allows rapid adaptation of virus towards changing environment [12].

Next-generation sequencing (NGS) can detect viral quasispecies by direct sequencing mixed sample with high coverage. Every read obtained in this way represents a contiguous fragment of DNA from a single molecule in the DNA library of the sample. Therefore, the set of reads provides a statistical sample of the DNA library and can be used to make inference about the genetic structure of the population.

Using NGS for low-frequency variants identification is still challenging because of sequencing errors. A consequence of the high error rate is the risk of considering a technical error as a low-frequency variant. Thus, the raw data alone leads to the overestimation of the genetic diversity of the sample [13].

There are different ways to estimate systematic and random sequencing errors at various steps of preparation during NGS workflow. These methods make it possible to improve the detection of

Table 4. Low-frequency mutations

Location, bp	Base changes	Codon changes	Amino acid changes	Mutation	Frequency, %
307	C → T	CCG → TCG	Pro → Ser	missense	0.94
1 458	T → C	TAT → TAC	Tyr → Tyr	synonymous variant	0.84

low-level variants by deep sequencing.

In the present work, we used a case-control design with comparison of the sample of interest with the control sample [14]. Sequencing and alignment of control samples to the reference provided information about the average error rate in our particular experiment. The SNP leading to an amino acid substitution occurred with a greater frequency than three standard deviations from the mean. This method enables the identification of a viral hemagglutinin low-frequency variant (Pro/Ser substitution at position 102) that allowed the virus to avoid specific antibodies.

Another way is a cohort design in which multiple samples are sequenced simultaneously, and each sample is compared individually versus the remaining samples ("aggregate control") [14]. Also, omputational error correction techniques promise to eliminate sequencing errors and improve the results of downstream analyses [15]. Some of these approaches aim to get rid of sequencing errors without losing sensitivity during the detection of low-frequency variants [16]. The great demand for accurate sequencing data within the biomedical community specialists leads to further development in error correction methods.

References

1. Kirkpatrick E, Qiu X, Wilson PC, Bahl J, Krammer F. The influenza virus hemagglutinin head evolves faster than the stalk domain. Sci Rep 2018;8(1):10432.
2. Jones JE, Le Sage V, Padovani GH. Parallel evolution between genomic segments of seasonal human influenza viruses reveals RNA-RNA relationships. Elife 2021;10:e66525.
3. Franco-Paredes C, Carrasco P, Preciado JI. The first influenza pandemic in the new millennium: lessons learned hitherto for current control efforts and overall pandemic preparedness. J Immune Based Ther Vaccines 2009;7:2.
4. Domingo E, Perales C. Species Concepts: Viral Quasispecies. Oxford: Academic Press; 2016.
5. Singer JB, Thomson EC, Hughes J, Aranday-Cortes E, McLauchlan J. Interpreting Viral Deep Sequencing Data with GLUE. Viruses 2019;11(4):323.
6. SRR1705851. NCBI Sequence Read Archive 2015 8; ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/.
7. Influenza A virus (A/USA/RVD1_H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene, partial cds. NCBI Nucleotide database; GenBank: KF848938.1.
8. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 2010;26(5):589–595.
9. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021 02;10(2). https://doi.org/10.1093/gigascience/giab008.
10. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 2012;22(3):568–576.
11. Andrews S, FASTQC. A quality control tool for high throughput sequence data; 2010.
12. Vignuzzi M, Stone J, Arnold J, Cameron C, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 2006;439:344–348.
13. Zagordi O, Klein R, Däumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res 2010;38(21):7400–9.
14. Ma X, Shao Y, Tian L. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 2019;20:50.
15. Mitchell K, Brito JJ, Mandric I. Benchmarking of computational error-correction methods for next-generation sequencing data. Genome Biol 2020;21:71.
16. Davis EM, Sun Y, Liu Y. SequencErr: measuring and suppressing sequencer errors in next-generation sequencing data. Genome Biol 2021;22:37.

Supplementary materials

Table 5. FastQC summary about sequencing reads

Measure	Value
Filename	SRR1705851.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	358265
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	42

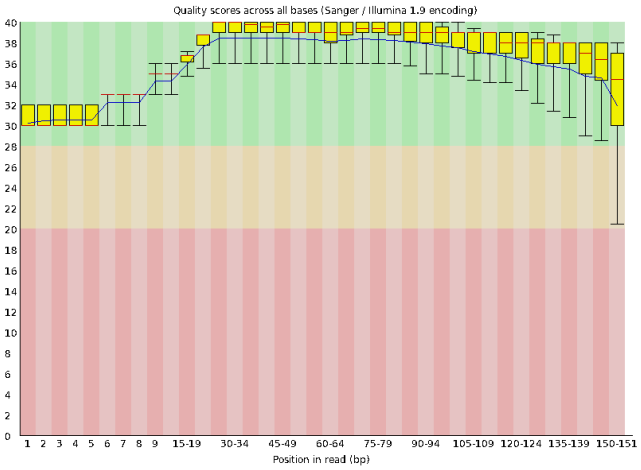

Figure 1. Per base reads quality by the Phred-33 score (FastQC report)