Detection of the rare variants for analyzing hemagglutinin genes

Ignat Sonets, Kamilla Faizullina

We analyze the amplicon data to detect rare variants. As occuring errors exist, we use three control sequences to understand which variants are real and which of them was detected as mutation due to error.

1 Introduction

Vaccine is a formulation of killed or attenuated pathogens, or antigens derived from them which help prevent, ameliorate, or treat infectious disease by stimulating antibody production or cellular immunity against the pathogen [1]. Epitope is a molecular region, usually an amino acid sequence, on the surface of an antigen that is capable of eliciting a specific immune response. [2]. Any changes of epitopes lead to decreasing efficiency of antibodies.

Influenza viruses are important human respiratory tractpathogens responsible for the seasonal epidemics andsporadic pandemics around the world [4] Influenza exists as quasispecies. Quasispecies are viral variants leading to diversification of the original strain [3]. Deep sequencing enables us to study mixed populations. However, the detection of rare varints might be challenging due to errors which accure prior to or during sequencing.

2 Methods

2.1 Data

In order to analyze the hemaaglutinin genes, we use Amplicon of H3N2 HA infecting Homosapien labeled SRR1705851 from The National Center for Biotechnology Information database [5]. We analyze the quality of the amplicon data via Fastqc [6]. Figure 1 illustrates the quality of this sequencing data. It is quite good and we do not need to filter the reads. We use the reference data [9]. Also, we use three control sequencing data labeled SRR1705858, SRR1705859 and SRR1705860 [10].

The sequence	Reads	Mapped
Homo sapiens(SRR1705851)	358265	361116
SRR1705858	256586	256658
SRR1705859	233327	233375
SRR1705860	249964	250108

Table 1: Data

2.2 Methods

In order to to map the data from the resistant strain to the reference sequence, we use the aligner called BWA-MEM [7]. VarScan is used to find SNP [8].

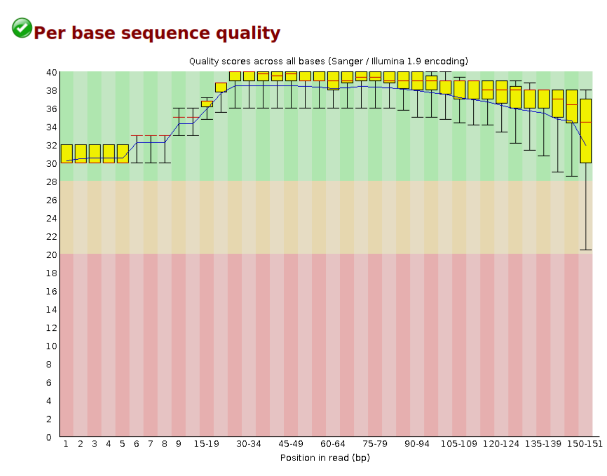

Figure 1: The quality of the sequence labeled SRR1705851

3 Results

3.1 Common variants

The data labeled SRR1705851 is used to detect the variants. The parameter –min-var-frequency equals 0.95 for searching common variants. Five SNPs were reported (Table 2). We used the codon table to check if these mutations can affect the protein, but all these variants do not change the amino acids.

3.2 Rare variants

Next, we run VarScan to detect rare parameters, but –min-var-frequency equals 0.001. We obtained 21 SNPs. To find the difference between real rare variants and possible sequencing error, we use three isogenic viral samples. Again, we align these samples

Position	Ref	Alt	Triplets	Amino acids
72	A	G	ACA → ACG	Threonine → Threonine
117	C	T	GCC → GCT	Alanine → Alanine
774	T	C	TTT → TTC	Phenylalanine → Phenylalanine
999	C	T	GGC → GGT	Glycine → Glycine
1260	A	C	CTA → CTC	Leucine → Leucine

Table 2: The variants, –min-var-frequency=0.95

The control sequence	Mean frequency	Standard deviation
SRR1705858	0,24928571	0,04716921734
SRR1705859	0,2369230769	0,05237640771
SRR1705860	0,25032787	0,07803775183

Table 3: Data

to the reference and get SNPs. Table3 presents the mean and standard deviation for frequencies of detected variants. Table 4 presents the rare SNPs with frequencies that are more than 3 standard deviations away from the averages in the reference results.

4 Discussion

Position	Ref	Alt	Triplets	Amino acids
307	C	T	CCG → TCG	Proline → Serine
1458	T	C	TAT → TAC	Tyrosine → Tyrosine

Table 4: The rare variants,

References

[1] Vaccine. In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer, Berlin, Heidelberg, 2006.

[2] Epitope. In: Vohr HW. (eds) Encyclopedia of Immunotoxicology. Springer, Berlin, Heidelberg, 2016.

[3] Quasispecies. In: Schwab M. (eds) Encyclopedia of Cancer. Springer, Berlin, Heidelberg, 2008.

[4] H. Ghaffari, A. Tavakoli, A. Moradi, A. Tabarraei, F. Bokharaei-Salim, M. Zahmatkeshan, M. Farahmand, D. Javanmard, S. J. Kiani, M. Esghaei, V. Pirhajati-Mahabadi, S. H. Monavari, and A. Ataei-Pirkooh, "Inhibition of h1n1 influenza virus infection by zinc oxide nanoparticles: another emerging application of nanomedicine," Journal of Biomedical Science , vol. 26, p. 70, Sep 2019

[5] NCBI: https://www.ncbi.nlm.nih.gov/sra/?term=SRR1705851

[6] Fastqc : https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[7] Burrows-Wheeler Aligner: http://bio-bwa.sourceforge.net/

[8] VarScan: http://dkoboldt.github.io/varscan/

[9] Reference data: http://public.dobzhanskycenter.ru/mrayko/Week2/KF848938.1.fasta

[10] SRR1705858: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1705858
SRR1705859: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1705859
SRR1705860: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1705860

[11] Munoz E. T., Deem M. W. Epitope analysis for influenza vaccine design,Vaccine,2005.