



Comparison of different tools for Influenza genome sequencing data analysis

Artem Fadeev,
Smorodintsev Research Institute of Influenza

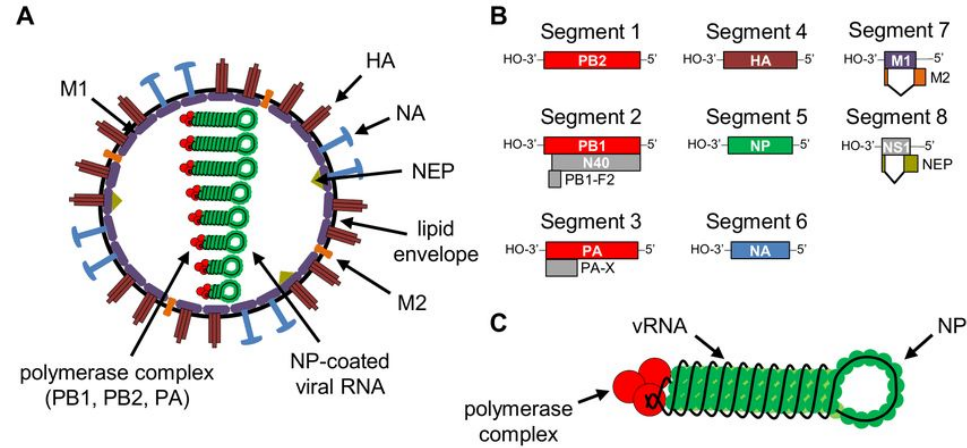


Mikhail Ushakov, Saint-Petersburg State Institute of Technology
Daria Tsyba, Pavlov University

Introduction

Despite the emergence of COVID-2019 epidemic, the high incidence of **Influenza** remains an urgent problem.

Unlike SARS-CoV-2, the vaccine against various strains of the influenza virus is developed but needs constant adjustment based on analysis of genome structure of circulating viruses. The introduction of NGS technologies allows to do this quite accurately and efficiently. But a huge amount of data requires process optimization and the selection of the most accurate, fast and effective analysis tools.



Influenza virus genome consists of 8 RNA segments: PB2, PB1, PA, HA, NP, NA, M and NS. For clinical practice and vaccine production the most significant are segments coding hemagglutinin (HA) and neuraminidase (NA).



The project purpose is:
to compare different tools designed for
Influenza virus whole-genome NGS data
analysis (consensus sequence assembly +
● SNP calling).



Tools to Compare

1. BWACycle – Smorodintsev Research Institute of Influenza (<https://github.com/Molecular-virology-lab/bwacycle>)
2. IRMA – CDC Atlanta (<https://wonder.cdc.gov/amd/flu/irma/>)
3. INSaFLU – Instituto Nacional de Saude (INSa) Doutor Ricardo Jorge (<https://github.com/INSaFLU/INSaFLU>)
4. FluLINE – WHO CC Melbourne (<https://github.com/UmaSangumathi/FluLINE>)
5. FluSeq – Singapore (<https://github.com/hkailee/FluSeq>)

Project Objectives & Workplan

February 2020

Available tools analysis

- Looking for available tools
- Installation

Processing of
Illumina & Nanopore Data
with different tools

- Running tools with test data
- Running tools with real data
(downloaded from open sources)


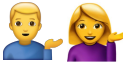
Output comparison

- Output investigation
- Coding script for analysis
- Visualization & Comparison

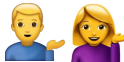
Project Objectives & Workplan

30 May 2020

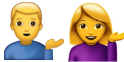



Available tools analysis


- Looking for available tools  ✓
- Installation  ✓

Processing of Illumina & Nanopore Data with different tools

- Running tools with test data  ✓
- Running tools with real data

Output comparison

- Output investigation  ✓
- Coding  script for analysis  ✓
- Visualization & Comparison  ✓



Exploratory ~~data~~ tools analysis

Before running the selected tools on big datasets, we conducted a “surface” analysis of their characteristics: capabilities and limitations.

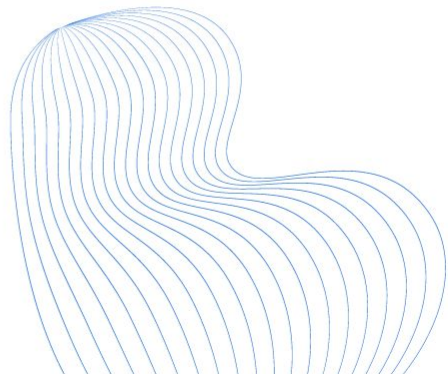
* Assessment criteria are subjective enough, but before this project, such an analysis was not carried out, so we can afford some liberties :)

Tools Comparison

	BWAcycle	IRMA	INSaFLU	FluLINE	FluSeq
Web or local	local	local	web/local	local	local
User friendly	+/- help, dependencies	+	+/- output, local installation NEW Docker is available	-	-
Illumina support	+	+	+	+	+
Nanopore support	+	+	- (web)	?	?
Speed	20-26 min per sample	10-17 min per sample	hours - days (web)	?	?

At this stage, we decided to abandon the explicit "outsiders" (FluLINE and FluSeq) and continue comparing 3 other tools (3 for Illumina and only 2 for Nanopore).

Coverage Comparison



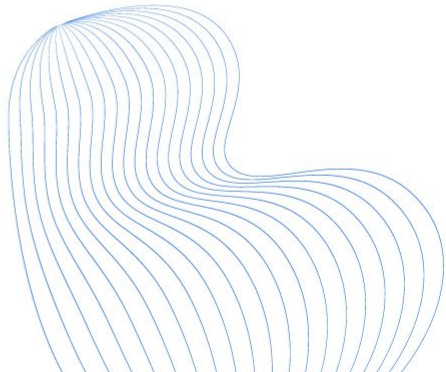
Evaluation of the resulting **coverage** is critical at the initial stage, since the high coverage can overcome errors in further analysis: base calling and assembly. We coded a simple visualization of coverage for the faster comparison.

Coverage Comparison



Each graph reflects the coverage obtained for each segment of Influenza virus genome. BWAcycle and IRMA create statistics per position, INSaFLU shows only mean coverage for whole segment.

SNP Calling Comparison



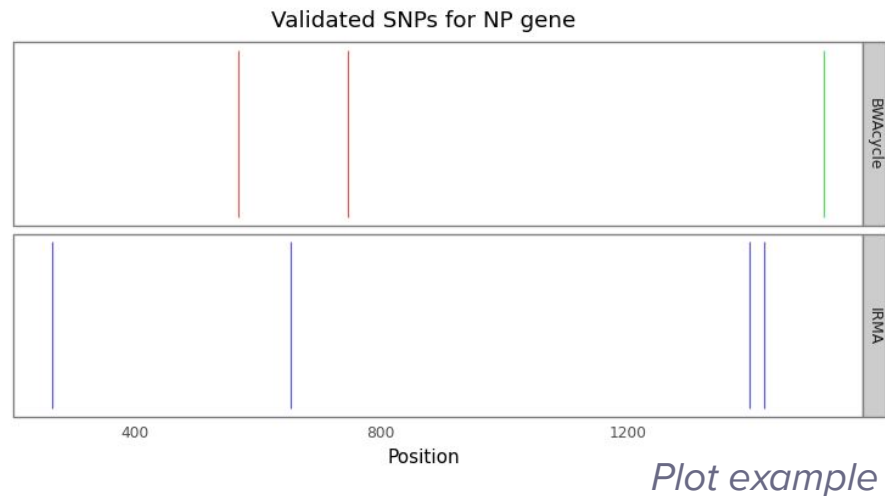
The most important point in the analysis of whole-genome sequencing data is SNP calling. The new variants in the Influenza virus genome determine the characteristics of the circulating strains, and therefore the specifics of seasonal vaccines. The accuracy and reliability of the obtained SNPs are based on two components:

- 1) correct assembly and identification of the consensus sequence
- 2) SNP calling actually

SNP Comparison - Principles

To visualize SNPs found by different tools, we also wrote a small script that allows to display the position of the variant and its meaning - the type of replacement determines its color on the plot.

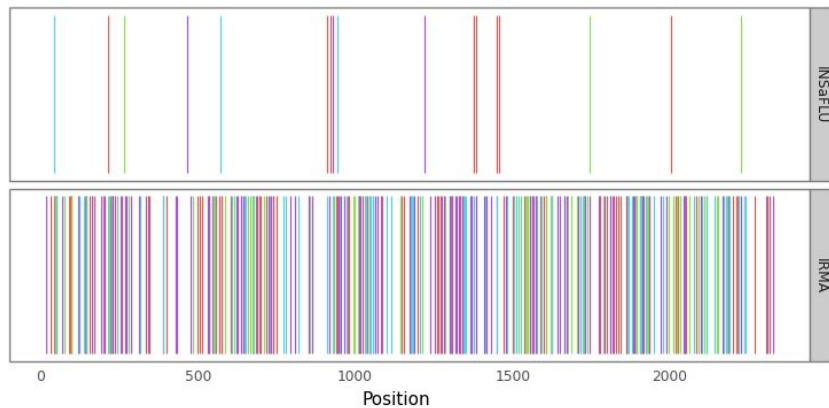
Next 2 slides demonstrate such plots created for the test sample.



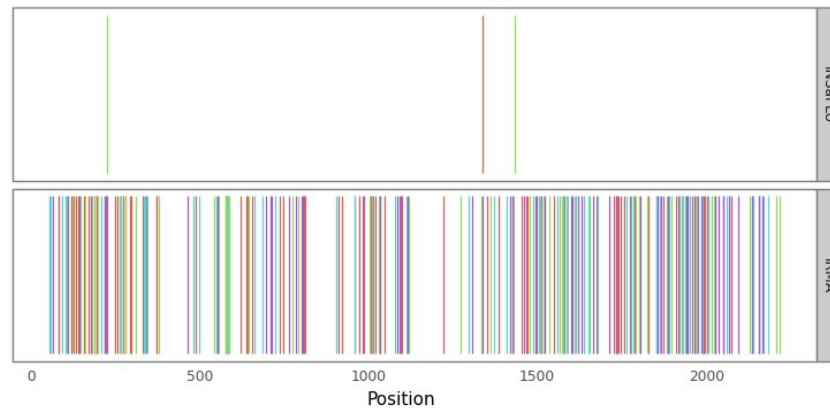
This approach allows us to identify gross errors in the assembly of the consensus sequence and / or its identification, without resorting to an eyeball analysis of dozens of output files. For example, in this way we found out IRMA cannot adequately analyze our test sample while BWAcycle didn't find SNPs at all :(The most likely reason is the reference that was incorrectly chosen by the tool for alignment.

SNP Comparison - Test Sample (1)

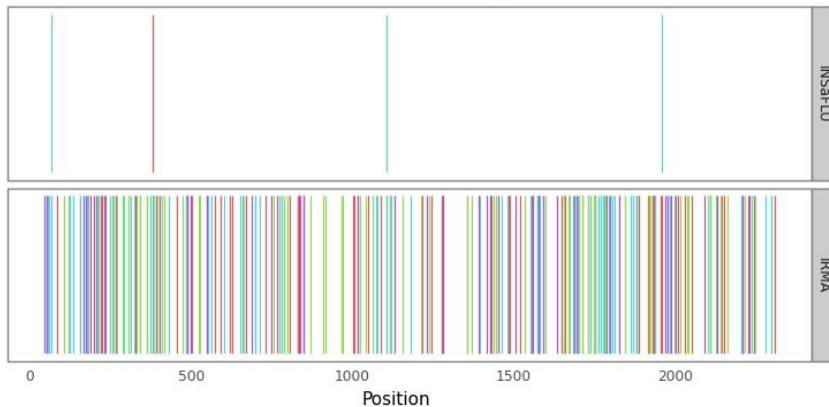
Validated SNPs for PB2 gene



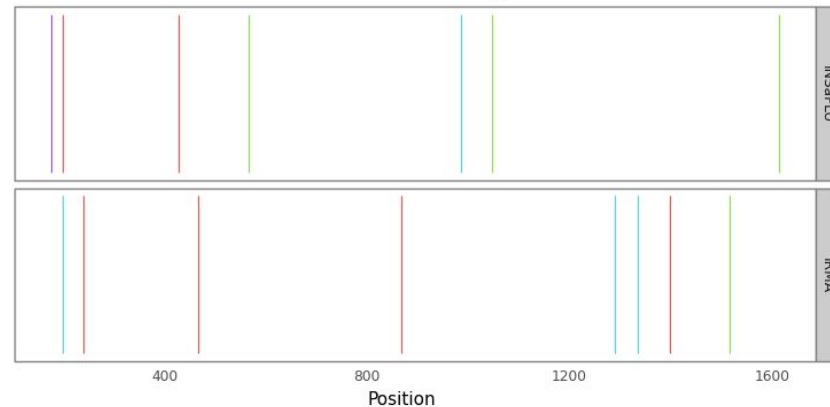
Validated SNPs for PA gene



Validated SNPs for PB1 gene

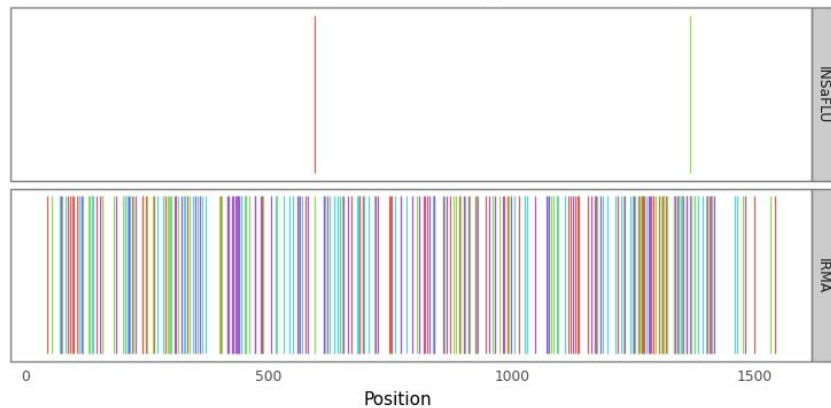


Validated SNPs for HA gene

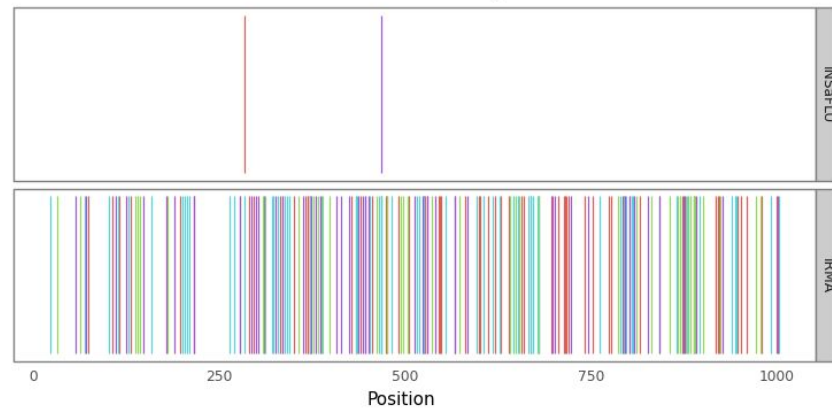


SNP Comparison - Test Sample (2)

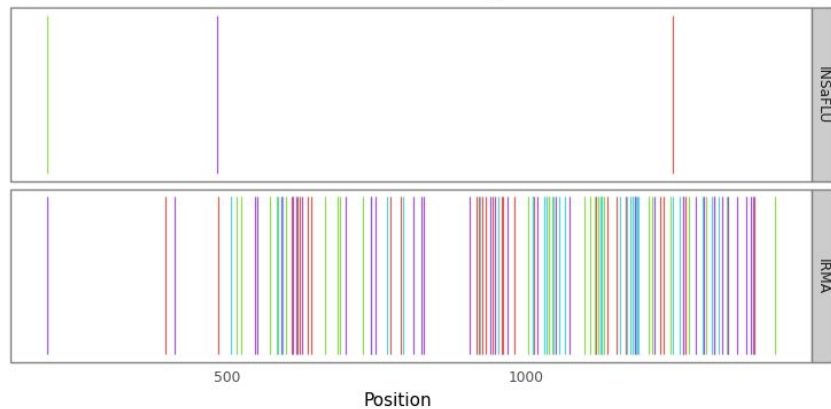
Validated SNPs for NP gene



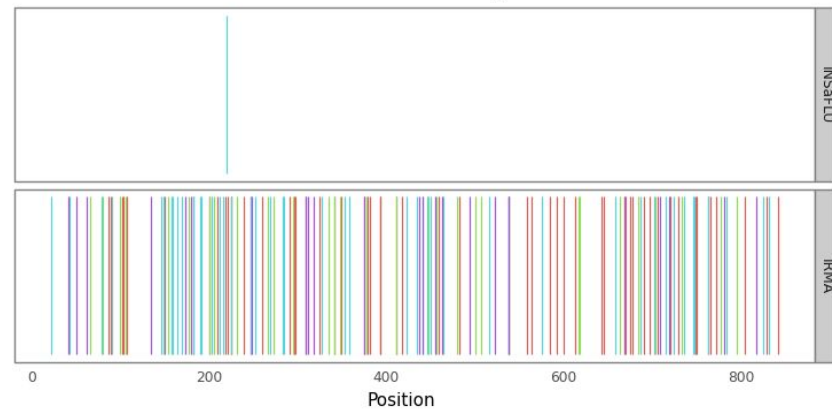
Validated SNPs for M gene



Validated SNPs for NA gene



Validated SNPs for NS gene



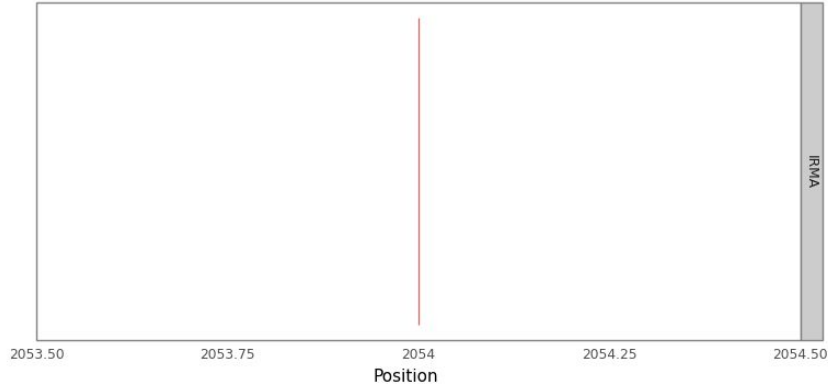
SNP Comparison - Nanopore

After an unsuccessful run on the Illumina data, we analyzed the Nanopore test data using BWAcycle and IRMA (web-version of INSaFLU doesn't support Nanopore data analysis, Docker was not available at that moment).

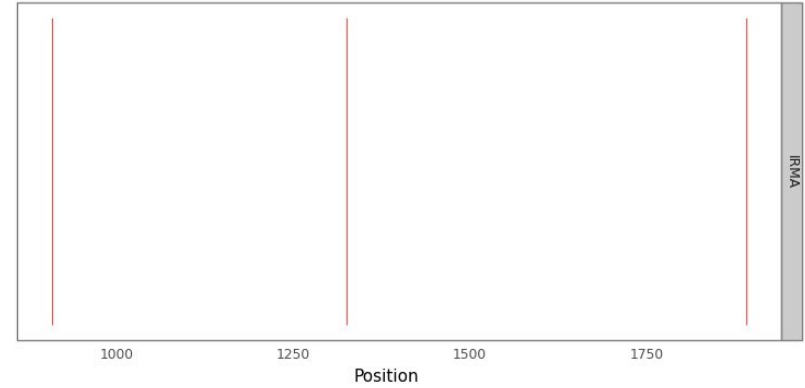
From the next 2 slides, we can verify that we got more adequate results. However, the reasons for the mismatch between the positions and the type of SNPs remain unclear and undetectable in the absence of metadata for test sample.

SNP Comparison - Nanopore (1)

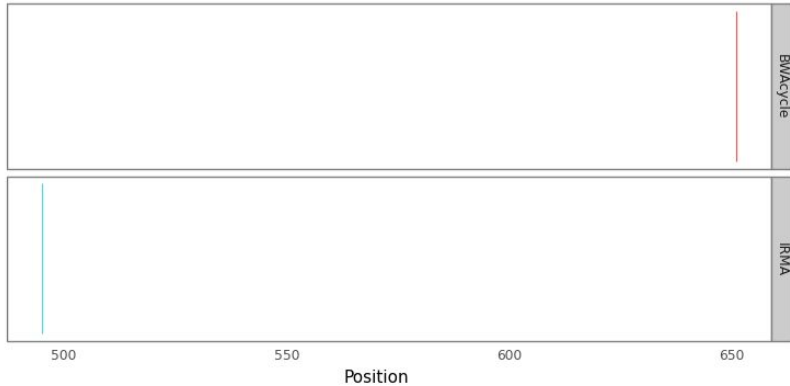
Validated SNPs for PB2 gene



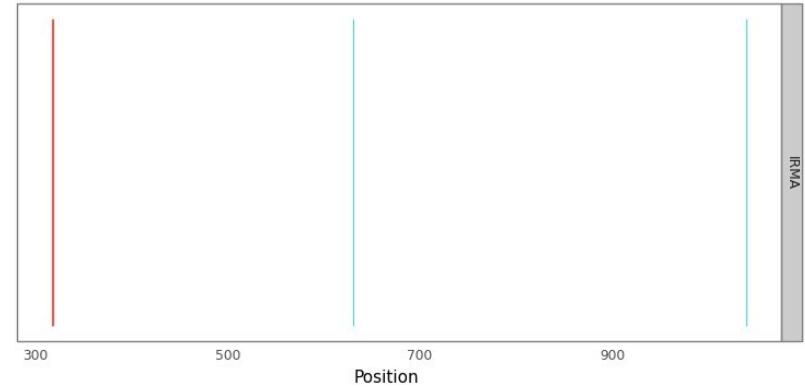
Validated SNPs for PA gene



Validated SNPs for PB1 gene

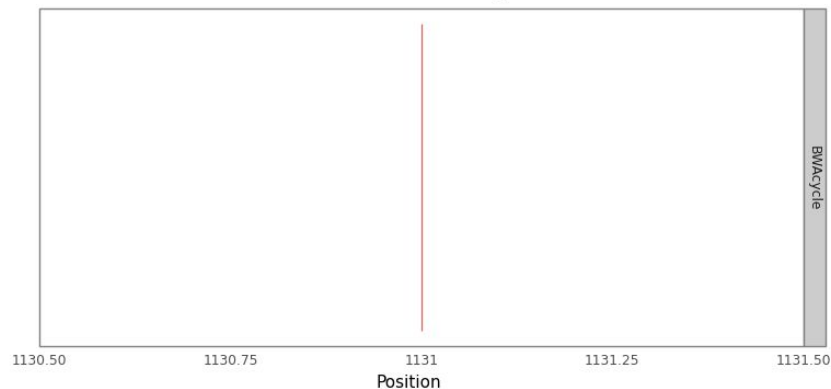


Validated SNPs for NA gene

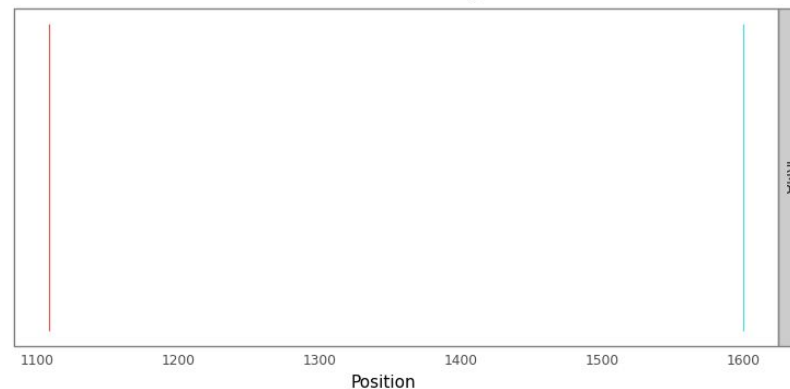


SNP Comparison - Nanopore (2)

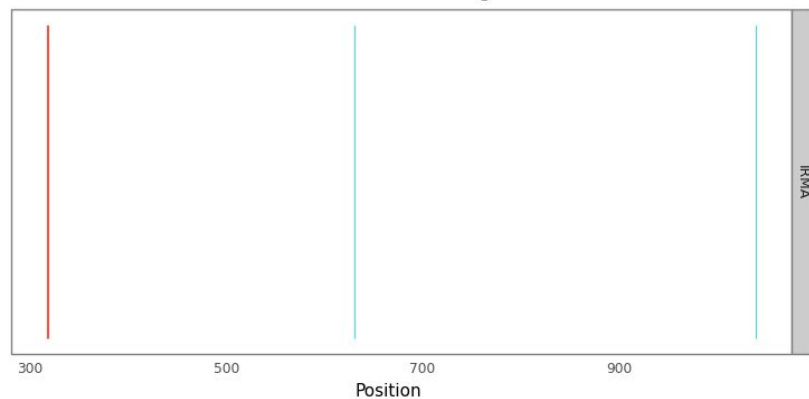
Validated SNPs for NP gene



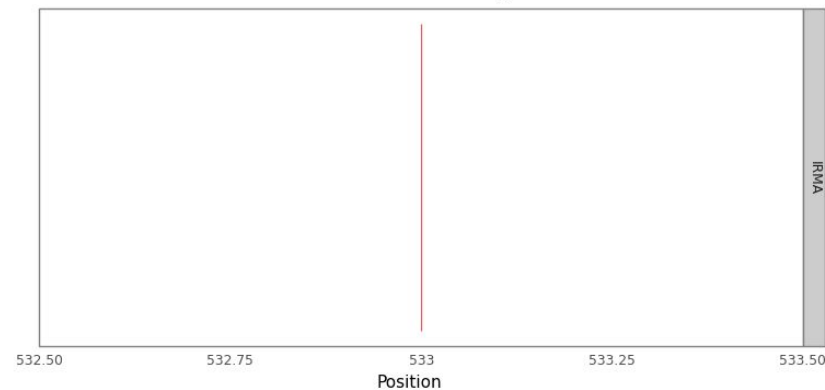
Validated SNPs for M gene

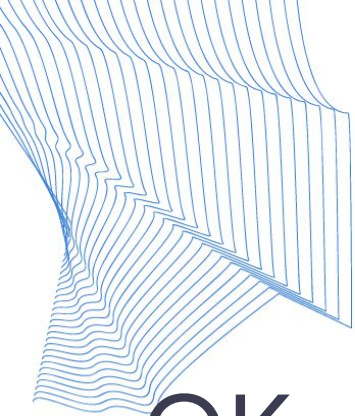


Validated SNPs for NA gene



Validated SNPs for NS gene





OK, but which
one is the best
tool?

To answer this question it is **not enough** to look at a couple (of dozens) of graphs. We need to compare the data obtained in the analysis of a large number of “reference” samples, i.e. samples for which we **know** the **strain** and other **metadata exactly**. In addition to the data themselves, such analysis requires significant **computing power**.

Troubles & Conclusions

Since all human and computational resources of **Smorodintsev Research Institute of Influenza** are thrown into the study and fight against the COVID-19 epidemic, we were forced to suspend the analysis and draw **intermediate conclusions**.



Based on the capabilities and limitations of the tools being compared, as well as the results of the analysis of the test sample, we came to the conclusion that the most suitable tools are **BWAcycle** and **IRMA**.

We hope to confirm this conclusion statistically soon, as well as develop a strategy for improving BWAcycle based on the difficulties and limitations discovered during the analysis.

Further Plans

Build
BWAcycle,
IRMA &
INSaFLU on
server

Perform
analysis
using
bigger set
of samples

Selection of the
best of the best
tools for Influenza
virus genome
sequencing
analysis in SRII

[Project on GitHub](#)

