In collaboration with

Automatization of data analysis in the quality control process

Student(s)
Shakir Suleimanov, Shakir.Suleimanov@skoltech.ru

Program
Life Sciences

Skoltech supervisor
Vera Rybko

Introduction

The immune repertoire sequence refers to the diverse collection of antigen receptor sequences present in an individual's immune system, including T-cell receptors (TCRs) and B-cell receptors (BCRs). MiLaboratory is a biotechnology company specializing in advanced software tools and services for immunogenomics and immunoinformatics. Some of the products in the market developed by MiLaboratory are immunosequencing reagent kits. Every produced kit must pass quality control (QC) as one of the development stages before reaching the customer.

Objectives

The main goal was to create a fully automated quality control procedure for immune repertoire sequencing kits. Objectives:

- Adjust threshold values for QC metrics
- Collect QC results into tables and graphical representation
- Finalize all data into PDF report

Process & Methods

Python and *R* programming languages were chosen to solve this task. The pipeline is a *Jupyter Notebook*, where all commands are running sequentially from raw fastq reads obtained after Illumina sequencing to final reporting tables and images. The main tool that is used for immune-repertoire sequencing analysis - *mixcr* - was implemented in the pipeline to calculate basic statistics. Following and additional statistics were calculated and arranged into table using *Python* based on the output of the *mixcr*. Graphical interpretation of the data was created using *R*.

Finally, all table, metrics and images were arranged into the pdf report using Latex script which was provided by my project supervisor. My goal in this part was to adapt this script for current tables format.

Figure 1 shows main programming tools used for processing of the data and creation of final QC PDF report.

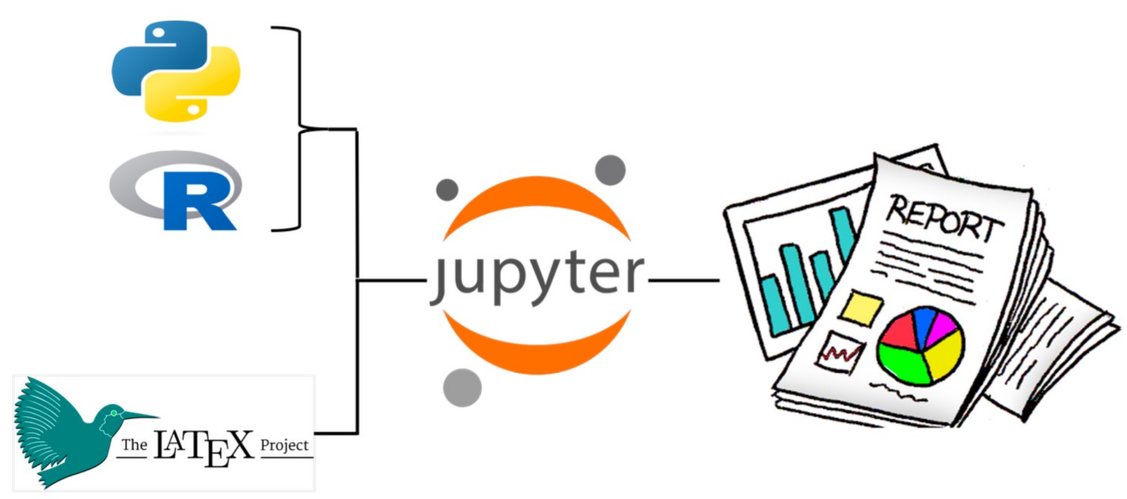

Figure 1. Programming tools and software used for QC report generation

Results

Automated script for quality check of 7 genes-multiplex protocol based immune repertoire sequencing kit was implemented during this project.

Figure 2 demonstrates the alignment statistic for incomplete rearrangements. This part allows to account not only for reads came from completely arranged TRA, TRB, TRG, TRD, IGH, IGK, IGL genes, but also for incomplete rearrangements. This parameter is important for oncological diagnostics.

Figure 3 demonstrates basic V-usage graph. Crosses were added over threshold line to mark those segments that do not pass the threshold.

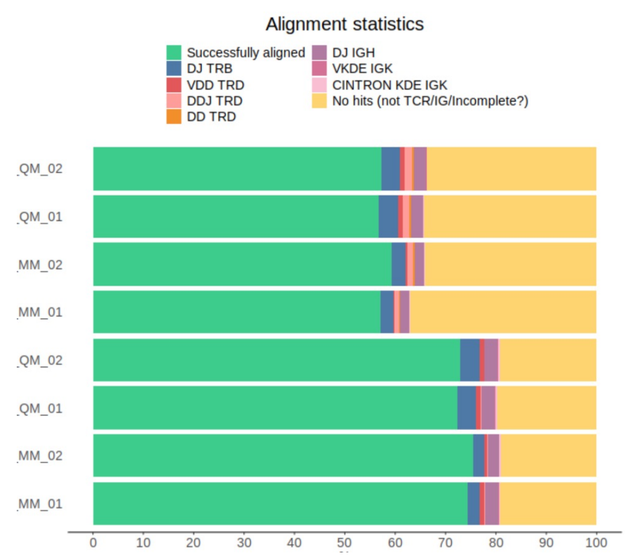

Figure 2. Alignment statistics

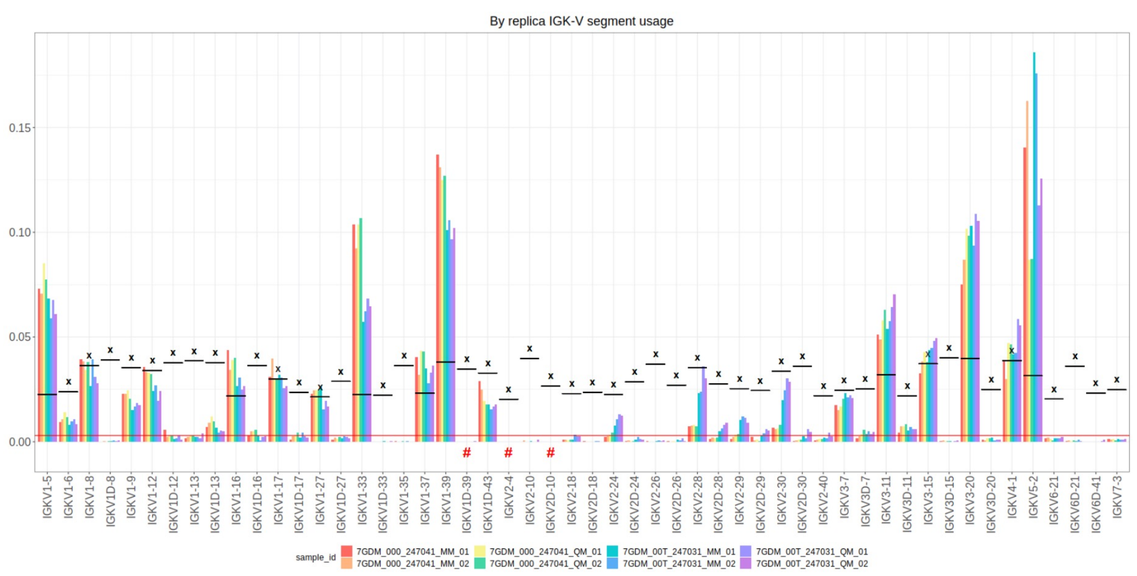

Figure 3. Segment usage

Conclusions

Each quality checking procedure is a time-consuming procedure, because one would need to run different tools and scripts to analyze raw data. Automation of this procedure helps to save a lot of time and to get the results of the QC in a readable format (pdf report). Automating the quality control process ensures that each sequencing kit is evaluated using the same stringent criteria, eliminating human error and variability. This leads to more accurate and consistent results, ensuring that only high-quality kits are distributed to customers.

I would like to thank my project supervisor Mikhail Myshkin for guidance through the project.