Development of a bioinformatics pipeline for the routine analysis of viral whole genome sequencing data

WETENSCHAPPELIJK INSTITUUT VOLKSGEZONDHEID
INSTITUT SCIENTIFIQUE DE SANTÉ PUBLIQUE

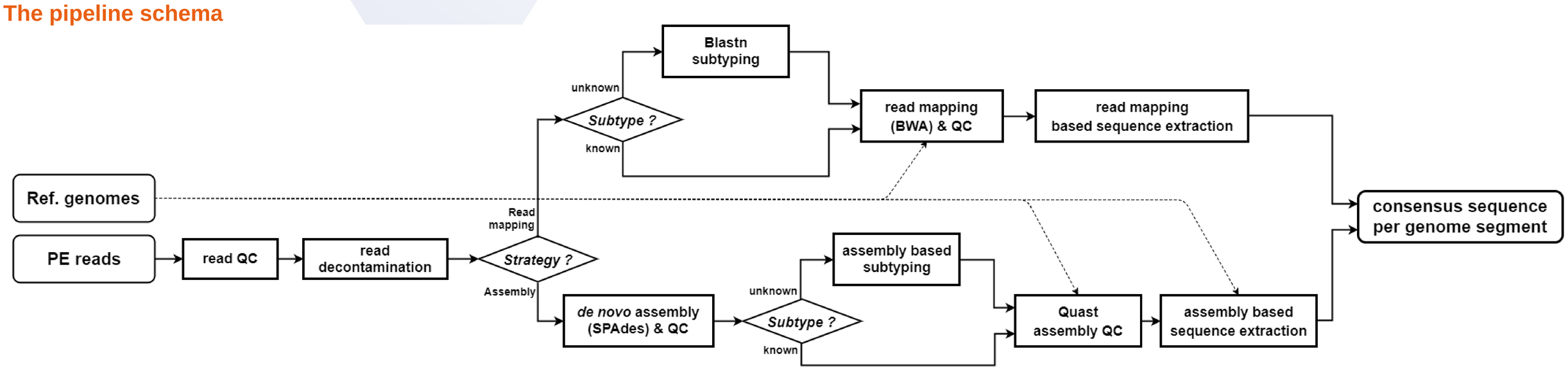
Qiang Fu[1], Bert Bogaerts[1], Raf Winand[1], Julien Van Braekel[1], Cyril Barbezange[2], Veronik Hutse [2], Steven Van Gucht [2], Sigrid De Keersmaecker[1], Nancy Roosens[1], and Kevin Vanneste[1]

1. *Platform Biotechnology and Molecular Biology; Scientific Institute of Public Health, Brussels (1050), Belgium*
2. *Department of Viral Diseases; Scientific Institute of Public Health, Brussels (1050), Belgium*

ABSTRACT

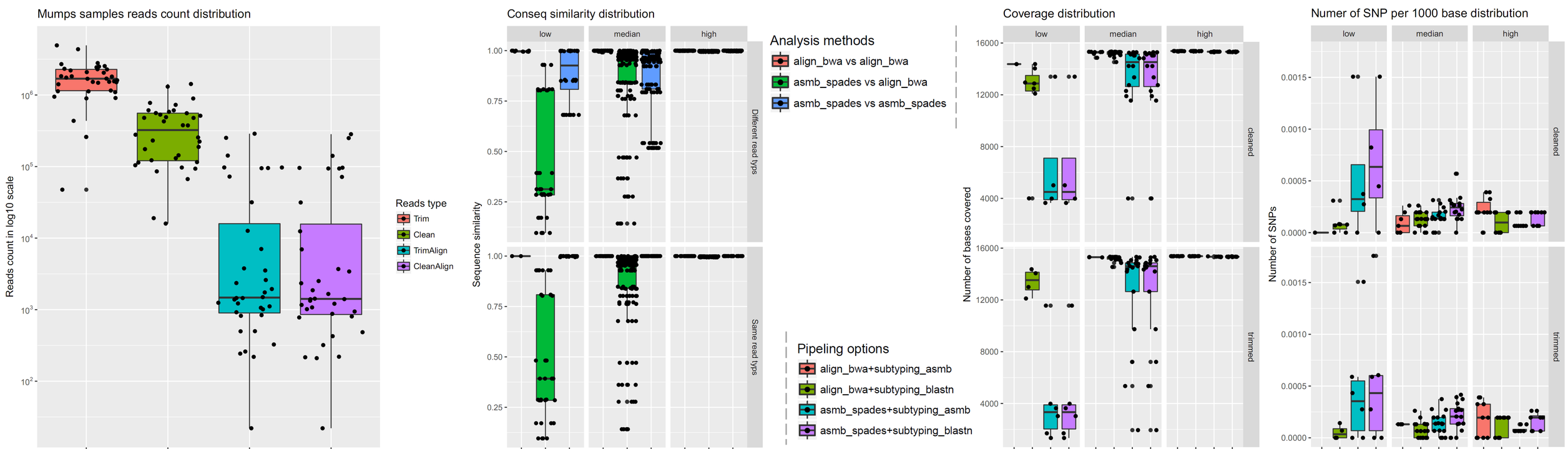
Viral infections can be a major public health threat, sometimes causing massive epidemics. To obtain the viral genome sequence is important for both characterization and prevention. As the viral genotype can diverge rapidly because of high mutation rates, traditional sequencing approaches such as genome walking are laborious and time consuming, while often only generating fragmented sequences. Alternatively, utilizing whole genome sequencing (WGS), the complete viral genome can be obtained providing unprecedented resolution for the study of viral genomics. The lack of the required bioinformatics expertise for analyzing WGS data is however a hurdle preventing its broad adaptation in many national reference centers. We developed a pipeline specifically designed to bridge this gap. Our pipeline is flexible and moreover species-agnostic. It performs automated quality control based on Illumina data (including optional removal of host DNA and/or DNA of other contaminants), and then generates the viral consensus sequence contained within a sample based on either the reference genome of choice or the closest one identified by *de novo* subtyping from a user-provided set of reference genomes. Additionally, a detailed output report containing intermediary results and important quality parameters is created. We deployed a user-friendly interface in an in-house Galaxy instance to facilitate access to a broad audience of scientists. Preliminary validation on mumps data demonstrates that our pipeline is capable of obtaining high-quality viral consensus sequences, providing a solid basis for downstream analyses such as viral genotyping, *in silico* serotyping, and virulence and/or resistance characterization. Our pipeline can easily be adapted to other viral species and will be made publicly available upon its publication.

The pipeline schema



Preliminary results on in-house Mumps data

Test data set and procedure: An in-house data set containing 36 samples of human saliva spiked with Mumps virus was available for testing. Data of each sample was first analyzed with both assembly and Blastn based subtyping procedure to identify the closest reference. Afterwards, the consensus sequences were extracted utilizing both the read mapping strategy (BWA) and the *de novo* assembly strategy (SPAdes).



Number of reads: The number of reads in each of the sample is very heterogeneous. The total number of valid reads (Trim and Clean) per sample ranges from 50,000 up to 6 million, while the number of Mumps reads (TrimAlign and CleanAlign, aligned by BWA) ranges from 22 to around 250,000. In downstream analysis presented here, we categorize the samples based on the number of TrimAlign Mumps reads they have. *Low* are samples with less than 500 Mumps reads, *median* are samples with more than 500 but less than 10,000 Mumps reads, and *high* are samples with more than 10,000 Mumps reads.

Consensus sequence similarity: with high coverage data, sequences extracted are very similar to each other no matter which method is used. With median and low coverage data, sequences extracted using BWA still maintain high similarity, whereas sequences extracted using SPAdes show more variations when different types of reads are used (trimmed vs cleaned). The similarity between sequences extracted using different methods is much lower with low coverage data, which prove to be problematic for the *de novo* assembly based method.

Sequence statistics: with high coverage data, all extracted sequences cover almost the complete Mumps genome (~15,300 bases). With median coverage data and using the SPAdes based method, there is a large variation on the percentage of genome covered, while the BWA based method maintained the same performance as on high coverage data. On low coverage data, the BWA based method clearly outperforms the SPAdes one (left figure). The number of SNPs per 1,000 extracted sequence bases is used to judge the quality of the reference identified by the subtyping method. The assembly based subtyping method shows a clear improvement compared to the Blastn based method on low and median coverage data. On high coverage data, both methods performance equally well (right figure).

CONCLUSION AND FUTURE PERSPECTIVES

The preliminary evaluation of our pipeline based on Mumps dataset shows that the pipeline can extract consensus sequences covering almost the complete genome with good quality data. For samples with low virus concentration, consensus sequences covering large parts of the genome could be extracted utilizing the read mapping strategy. The pipeline will be further evaluated using other viral datasets. It will be publically available through a Galaxy instance hosted at the institute. The pipeline will be actively maintained and regularly updated, incorporating new tools and methods with proven performance.

WIV-ISP | Rue Juliette Wytsmanstraat 14 | 1050 Brussels | Belgium
T +32 2 642 52 58 | F +32 2 642 52 92 | e-mail: nancy.roosens@wiv-isp.be | web: www.wiv-isp.be