A hybrid pipeline for reconstruction and analysis of viral genomes at multi-organ level

Diogo Pratas[1,2], Mari Toppinen[2], Lari Pyöriä[2], Klaus Hedman[2,3], Antti Sajantila, Maria F Perdomo[2]

Background: Advances in sequencing technologies have enabled the characterization of multiple microbial and host genomes, opening new frontiers of knowledge while kindling novel applications and research perspectives. Among these is the investigation of the viral communities residing in the human body and their impact on health and disease. To this end, the study of samples from multiple tissues is critical, yet, the complexity of such analysis calls for a dedicated pipeline. We provide an automatic and efficient pipeline for identification, assembly, and analysis of viral genomes that combines the DNA sequence data from multiple organs. TRACESPipe relies on cooperation among 3 modalities: compression-based prediction, sequence alignment, and *de-novo* assembly. The pipeline is ultra-fast and provides, additionally, secure transmission and storage of sensitive data.

Findings: TRACESPipe performed outstandingly when tested on synthetic and *ex vivo* datasets, identifying and reconstructing all the viral genomes, including those with high levels of single-nucleotide polymorphisms (Fig.1). It also detected minimal levels of genomic variation between different organs.

Conclusions: TRACESPipe's unique ability to simultaneously process and analyze samples from different sources enables the evaluation of within-host variability. This opens up the possibility to investigate viral tissue tropism, evolution, fitness, and disease associations. Moreover, additional features such as DNA damage estimation and mitochondrial DNA reconstruction and analysis, as well as exogenous-source controls, expand the utility of this pipeline to other fields such as forensics and ancient DNA studies. TRACESPipe is released under GPLv3 and is available for free download at https://github.com/viromelab/tracespipe.

1 – Department of Electronics, Telecommunications and Informatics & IEETA, University of Aveiro

2 – Department of Virology, University of Helsinki

3 – HUSLAB, Helsinki University Hospital

4 – Department of Forensic Medicine, University of Helsinki

5 – Forensic Medicine Unit, Finnish Institute of Health and Welfare

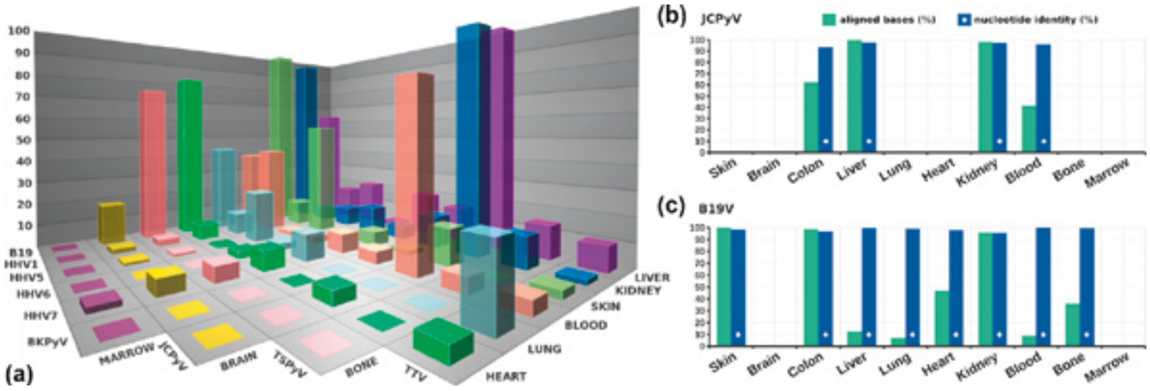

FIGURE 1

a) Breadth coverage percentage (z-axis) of the (real) mapped reads against the best reference virus for each organ sample. The plot is restricted to viral types with a minimum similarity of 10% in ≥1 of the organs. The bottom corner had shallow values, which due to space constraints were not included. (b,c) Percentage of aligned bases (green) and nucleotide identity (blue) between the best reference and reconstructed genomes of JCPyV and B19V, respectively, calculated using dnadiff. Low breadth coverages may not have corresponding aligned-data values as they may have fallen under the minimal quality or similarity thresholds. The latter was set before the run to exclude noise.