Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Automatic reconstruction of persistent human virus sequences

MARIA J. P. SOUSA[1, 2] and DIOGO PRATAS[1, 2, 3, 4]
[1]Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Portugal
[2]Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
[3]Intelligent Systems Associate Laboratory, University of Aveiro, Portugal
[4]Department of Virology, University of Helsinki, Finland

Corresponding author: Maria J. P. Sousa (e-mail: mj.sousa@ua.pt) and Diogo Pratas (e-mail: pratas@ua.pt).

This work was partially funded by National Funds through the FCT (Foundation for Science and Technology), in the context of the project UIDB/00127/2020. M.S. has received funding from the FCT - reference BI/UI62/10835/023. D.P. is funded by national funds through FCT, I.P., under the Scientific Employment Stimulus - Institutional Call - reference CEECINST/00026/2018.

ABSTRACT The increasing availability of human viral sequenced samples, namely from clinical and forensic contexts, has led to the emergence of many optimized genome reconstruction tools. Although the number of new tools is steadily increasing, it is complex to identify functional and accurate tools for improved human viral genome reconstruction. In this work, we survey open-source programs used for human viral genome reconstruction, identifying specific characteristics, features, similarities, and dissimilarities between these tools. For the quantitative comparison, we create an open-source reconstruction benchmark. The benchmark consists in the use of different human viruses with different mutation rates, contamination, mitochondrial DNA inclusion, while the sequencing process is simulated with different coverage depths. The evaluation measures include the identity, a Normalized Compression Semi-Distance (NCSD), and the Normalized Relative Compression (NRC) between the genomes before and after reconstruction, as well as the computational time and memory spent by each tool. Additionally, because our results show that there is not a best tool for all cases since it depends on the data characteristics, we provide an enhanced pipeline for human DNA viral reconstruction that uses the cooperation between the reconstruction tools used in this study. We show that this pipeline is able to better adapt to the sample while reconstructing the viral genomes with increasing accuracy. The benchmark and the cooperation pipeline are fully reproducible and freely available online.

INDEX TERMS human viral genomes, sequence assembly, cooperation, reproducibility, survey.

I. INTRODUCTION

VIRAL particles are submicroscopic infectious agents that replicate inside living cells and infect all known life forms. Viruses are associated with human evolution [1] but also responsible for a considerable number of health issues, namely diseases such as cancer [2] or pandemics such as Covid-19. Therefore, the sensitive sequencing followed by the fast and accurate reconstruction (or assembly) of sequenced human viral genomes is of critical importance [3].

The primary goal of this work was to explore and improve the automatic reconstruction of human persistent viral genomes. To study this topic, we started with a systematic review. Then, we provided a fully reproducible benchmark tool for reconstructing the viral genomes for assessing (qualitatively and quantitatively) the performance of each methodology (HVRS). Finally, because our results show that there is not a best tool for all the cases, we created a novel cooperative pipeline (CoopPipe) that is able to improve upon the results obtained using pre-existent reconstruction programs.

II. SYSTEMATIC REVIEW

For selecting the most suitable tools for the projects developed in our study, there was the necessity to make a systematic review of the programs that perform viral genome reconstruction.

The search strategy targeted studies that investigated the role of viral genome reconstruction tools. The search included only the English language, and the databases of articles searched were Pubmed, IEEE Xplore Digital Library (XDL), and Google Scholar. Boolean logic with MeSH terminology was used, including terms of reconstruction and general computational terms. The term "assembly or reconstruction" was used to search IEEE XDL. Also, top results from Google

Scholar, including related references and studies/reviews, were screened.

All results obtained were filtered using the criteria previously stipulated. This review exclusively included studies that provided a computational tool for the reconstruction of human viral genomes, specifically open-source tools. This review excluded any tools that were not able to be installed locally, tools that were unable to be installed without a registration and tools that required aligned reads, contigs or the result of other tools as input. Only articles with full texts were included. The publication date was specified as 01-2000 to 03-2023.

Using this methodology, 27 reconstruction programs were found, from which, 15 were able to be installed and reconstruct genomes. These programs show different characteristics, from the methods used, to the genomes they reconstruct and the programming languages they were written in. These 15 tools were considered for both the benchmark and the cooperative reconstruction pipeline.

III. BENCHMARK - HVRS

HVRS is an open-source and fully reproducible benchmark tool that allows its users to simulate metagenomic datasets based on real viral sequences, install all of the tools necessary to execute HVRS, reconstruct the viral datasets, evaluate the results obtained and generate plots. The benchmark tool is publicly available at https://github.com/viromelab/HVRS.

HVRS uses GTO to mutate the reads, AlcoR to generate contamination DNA, ART to simulate the sequencing process, 15 reconstruction tools from different reconstruction methodologies to reconstruct the reads, and dnadiff and GeCo3 to evaluate the results obtained.

The benchmark contains 65 different datasets, each constituted by several *real* viral sequences, which are mutated and to which contamination and mitochondrial DNA is added. The datasets were evaluated in terms of their identity, NCSD, and NRC, as well as the computational resources they utilise.

According to the analysis done, there is no tool that can be considered to be the best, as different tools have different characteristics that affect the performance among datasets. However, the performance of all of the tools considered tends to decrease with low fold coverage and with the increase of the percentage of SNPs present in the dataset reconstructed.

IV. RECONSTRUCTION PIPELINE - COOPPIPE

The CoopPipe is a reconstruction pipeline capable of reconstructing viral genomes from human viral samples, using the output of a myriad of tools publicly available (from HVRS) and improving the results obtained using each one of them. CoopPipe includes the installation of all of the tools used in its execution and the viral databases used in the classification process. Additionally, the tool is also capable of classifying both input and output data in terms of their viral content, which can be useful to detect which viruses are present in the sample. CoopPipe is able to evaluate the reconstruction process and generate plots with the results. The CoopPipe is publicly available at https://github.com/viromelab/CoopPipe.

CoopPipe classifies the input reads and uses that information to retrieve suitable references from a viral database. The genomes are then reconstructed, using the references when required and the output is classified to assess what viruses the reconstruction tool was able to reconstruct and the viruses are extracted from the viral database. After, the reconstructed sequences are aligned against the references, which provides us with information about what parts of which viruses were reconstructed by each reconstruction tool. The sequences corresponding to each virus are then joined in a multi-FASTA file, which is aligned against itself. From each multi-FASTA file, a consensus sequence is generated using either EMBOSS or Adaptive Weighted-K (AWK). AWK was also developed in the scope of this work. AWK is a method to generate consensus sequences that takes into consideration the performance of each reconstruction tool in the last k positions while picking which bases to consider for the consensus sequence. Lastly, the sequences for each virus are evaluated in terms of the average identity, NCSD, NRC, and computational resources used with dnadiff and GeCo3.

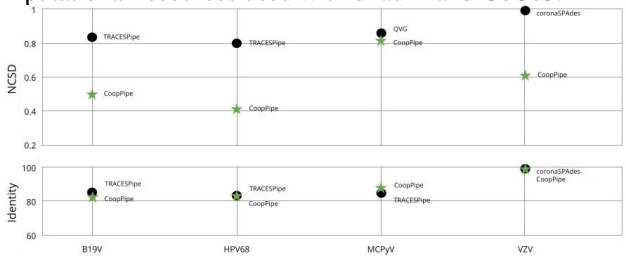

Fig. 1 - Performance of CoopPipe in DS24.

As depicted in Fig.1, CoopPipe tends to have a better performance than the other tools considered, especially in terms of NCSD and NRC.

V. CONCLUSIONS

Choosing a reconstruction tool that is able to perform well under every scenario is an impossibility, as different tools have different characteristics, which make them more or less able to reconstruct data in certain datasets. However, with the cooperation between the different reconstruction tools, it is possible to improve the reconstruction process and obtain higher accuracy. Additionally, it was found that the use of compression based metrics is suitable for the evaluation of reconstructed genomes.

REFERENCES

[1] D. Zhang *et al.*, HBV Integration Induces Complex Interactions between Host and Viral Genomic Functions at the Insertion Site," *Journal of Clinical and Translational Hepatology*, vol. 9, no. 3, pp. 399, 2021.
[2] M. Plummer *et al.*, Global burden of cancers attributable to infections in 2012: a synthetic analysis," *The Lancet Global Health*, vol. 4, no. 9, pp. e609–e616, 2016.
[3] L. Pyöriä *et al.*, Unmasking the tissue-resident eukaryotic DNA virome in humans," *Nucleic Acids Research*, vol. 51, no. 7, pp. 3223–3239, 2023.