A method for accurate reconstruction of persistent human viral sequences

Maria J. P. Sousa[1]
https://github.com/mirakaya

Diogo Pratas[1,2]
https://pratas.github.io

[1]IEETA/DETI/LASI
University of Aveiro, Portugal

[2]Department of Virology,
University of Helsinki, Finland

Abstract

The accurate reconstruction of viral genomes has become increasingly important due to the growing availability of diverse human viral sequenced samples. This necessity is particularly pronounced in clinical and forensic scenarios, where specialized tools capable of correctly reconstructing genomes are required. In this article, we present CoopPipe, a method capable of improving the reconstruction of human DNA viruses. CoopPipe utilizes an adaptive cooperation between publicly available viral reconstruction tools and is capable of improving the reconstruction process by, on average, 19.3% in terms of the NCSD and 17.0% in terms of the NRC, in relation to the best tool and for each virus. The implementation of CoopPipe is entirely reproducible and publicly available at `https://github.com/viromelab/CoopPipe`.

1 Introduction

The advancement of high-throughput sequencing technologies has made it easier to study the genetic makeup of human viral genomes [13]. The study of human viruses is of utmost importance since they are responsible for a myriad of diseases such as cancer, AIDS, or Covid-19 and can affect all of the major tissues and systems [7]. The sequenced genomes are, however, fragmented and in order to be studied in this context, they need to be reconstructed, which is a challenging and complex process both at the laboratory and computational levels.

The primary goal of this method is to improve the reconstruction of human DNA viruses, focusing especially on genomes with similar characteristics to those found in samples obtained from forensic and clinical scenarios. Therefore, we developed a novel cooperative pipeline (CoopPipe) that is able to improve upon the results of the best existent reconstruction programs.

2 Methodology

The CoopPipe is a reconstruction pipeline capable of reconstructing viral genomes from human viral samples, combining the output of other reconstruction tools, such as metaSPAdes [9], PEHaplo [1], QVG [16], TRACESPipe [12], and LAZYPIPE 2 [10]. CoopPipe is available as an open-source tool.

CoopPipe includes the automatic installation of all of the programs 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 the viral sequences that are contained within it, which can be used to detect which viruses are present in the sample given and the viruses each tool is capable of reconstructing. Moreover, CoopPipe can also evaluate the reconstruction process and plotting the results.

Figure 1 illustrates the different phases of the execution of CoopPipe. This methodology requires only the FASTQ reads from a metagenomic human sample as input.

CoopPipe has two different execution paths: the upper one, used by the reconstruction tools that do not require a reference to reconstruct the genomes and the lower path, followed by tools that require a reference.

The initial classification process, present on the lower path, serves to retrieve information about the contents of the input sample, which is used to extract suitable references from a viral database. This process is made using either Kraken2 [17], Centrifuge [4], or FALCON-meta [11].

Afterwards, the input genomes are reconstructed using the reconstruction programs selected and the results are stored as FASTA files. The last part includes the reclassification, where the reconstructed genomes are analysed to determine which genomes were able to be reconstructed.

The identification of which viruses or parts of viruses are contained in a sample is made by CoopPipe by aligning the reconstructed genome to the references retrieved, using either Bowtie2 [6] or BWA [8]. The reconstructed genomes for each virus are joined together in a multi-FASTA file, which is further multi-aligned using either MAFFT [3] or Muscle [2].

Lastly, a consensus sequence for each virus is generated using either EMBOSS [14] or Adaptive Weighted-K (AWK). The AWK is also a contribution of this work and its objective is to produce the most complete approximation of a genome from a multi-FASTA file. The AWK considers that if any of the tools is capable of reconstructing a nucleotide base, that base is considered for the consensus sequence and it employs models that evaluate the performance of each tool in the last k positions and makes decisions based on them. CoopPipe uses AWK with several models, setting the value of k to both low and high values, which makes it more adaptable to different scenarios.

After generating the consensus sequence, the reconstruction is evaluated based on the average identity, Normalized Compression Semi-Distance (NCSD), Normalized Relative Compression (NRC) and the computational resources used with dnadiff from MUMmer 4 [5] and GeCo3 [15].

3 Results

To benchmark CoopPipe, sixty-five nearly synthetic datasets based on *real* human viral sequences were generated. These datasets included different viral compositions, contamination, mitochondrial DNA and varying percentages of SNPs (substitutions) added. Additionally, the sequencing process was simulated with depth coverage ranging from 2x to 40x.

To illustrate the performance obtained by CoopPipe and given space constrains of this paper, we include results for two of the datasets generated are illustrated in the Figures 2 and 3. Both datasets considered had a read length of 150 bp, depth coverage of 2x and contained contamination, mitochondrial DNA and the viruses B19V, HPV, VZV and MCPyV, differing only on the percentage of SNPs each of them contains, with dataset 18 containing 3% of SNPs and dataset 24 containing 15% of SNPs.

Figure 2 shows the results obtained by CoopPipe when reconstructing dataset 18 using all reconstruction tools. The average identity of CoopPipe is worse than the one obtained using the best tool for each virus. The average identity is a metric that relies on alignments and the differences between the alignments and the reference, and it may not be the most reliable metric for this case since unaligned data is not taken into consideration. We used this metric only as a control. In terms of the NCSD, CoopPipe was able to improve the reconstruction made in relation to the remaining tools for the viruses HPV68, MCPyV and VZV, however, it reconstructed B19V less accurately than TRACESPipe. TRACESPipe was designed to accurately reconstruct the B19V virus, and the other tools that were able to reconstruct B19V did it less accurately, which negatively affected CoopPipe. Similar to the results obtained for the NCSD, CoopPipe has obtained better results in terms of the NRC in every virus except B19V, which can once again be attributed to the poor performance of the other reconstruction tools in relation to TRACESPipe.

Figure 3 shows the results obtained by CoopPipe when reconstructing dataset 24 using all reconstruction tools. CoopPipe has a better performance in terms of the average identity than the best tool in the virus MCPyV, with a slightly worse performance in the remaining viruses. In terms of the NCSD, CoopPipe has obtained better results than all other tools considered with significant improvements to the reconstruction of the viruses B19V, HPV68 and VZV. CoopPipe has also significantly improved the reconstruction of the viruses B19V, HPV68 and VZV in terms of NRC in comparison to all other reconstruction programs, however, its performance was slightly worse than the performance obtained by the best tool when reconstructing MCPyV.

In the whole datasets, CoopPipe was able to improve the reconstruction of viral genomes in relation to the best reconstruction tool by, on average, 19.3% in terms of the NCSD and 17.0% in terms of the NRC.

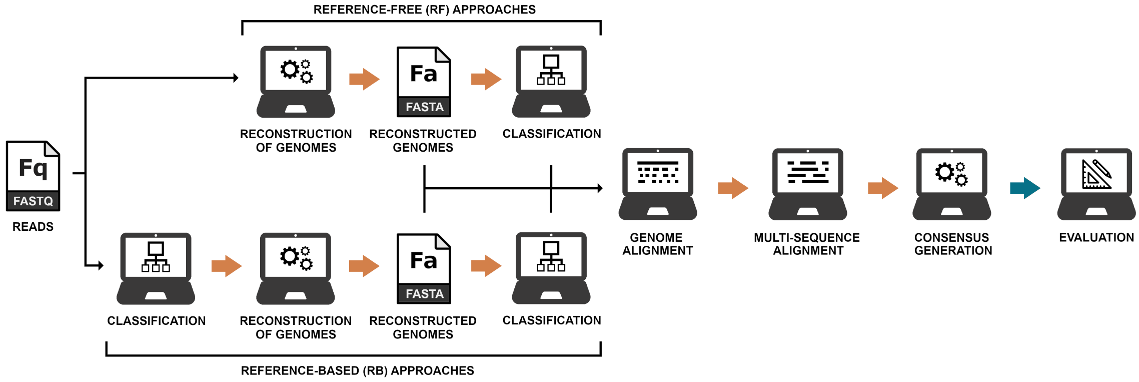

Figure 1: CoopPipe methodology depicting the different phases of execution. Hybrid reconstruction programs that depend on explicitly given references (VirGenA and V-pipe), follow the bottom path, whereas programs that are able to determine which references to utilize (LAZYPIPE and TRACESPipe) follow the top path. The blue arrow indicates that the step is optional in the execution of CoopPipe.

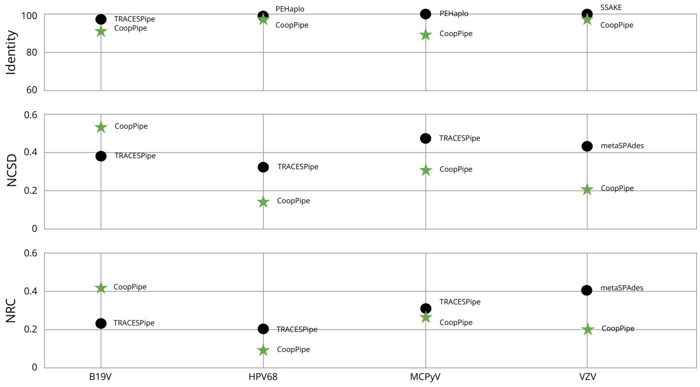

Figure 2: Results obtained using CoopPipe in relation to the tool that obtained the best performance for each virus of the dataset 18.

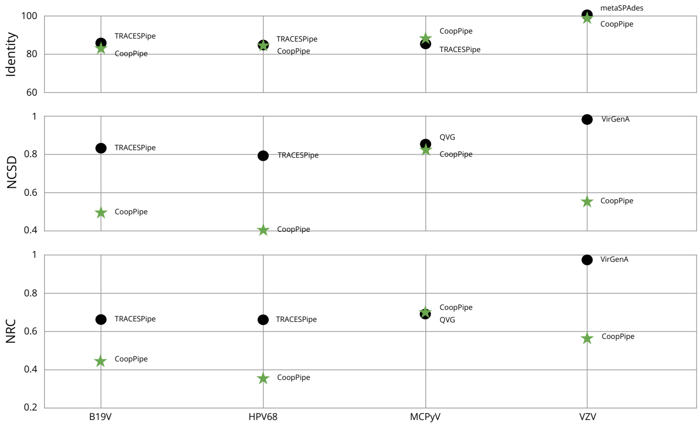

Figure 3: Results obtained using CoopPipe in relation to the tool that obtained the best performance for each virus of the dataset 24.

These values show the substantial improvement that CoopPipe provides in this very important and competitive area.

4 Conclusions

The precise reconstruction of human viral genomes is vital, yet finding a tool that consistently achieves optimal performance for diverse viral genomes and situations is challenging. We show that CoopPipe can leverage the cooperation between different reconstruction tools and significantly improve the reconstruction of different genomes, even in unfavourable scenarios, namely low-depth coverage and a high percentage of SNPs. Furthermore, we show that CoopPipe can improve the reconstruction of viral genomes in relation to the best reconstruction tool by, on average, 19.3% and 17.0% in NCSD and NRC, respectively.

Funding: 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 the FCT, I.P., under the Scientific Employment Stimulus - Institutional Call - reference `CEECINST/00026/2018`.

References

[1] J. Chen et al. De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding. *Bioinformatics*, 34(17): 2927–2935, 2018.

[2] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[3] K. Katoh et al. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[4] D. Kim et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, 2016.

[5] S. Kurtz et al. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):1–9, 2004.

[6] B. Langmead et al. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[7] K. Leppard et al. *Introduction to modern virology*. Oxford: Blackwell Publishing Limited, 2007.

[8] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

[9] S. Nurk et al. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, 2017.

[10] I. Plyusnin et al. Enhanced viral metagenomics with lazypipe 2. *Viruses*, 15(2):431, 2023.

[11] D. Pratas et al. Metagenomic composition analysis of an ancient sequenced polar bear jawbone from Svalbard. *Genes*, 9(9):445, 2018.

[12] D. Pratas et al. A hybrid pipeline for reconstruction and analysis of viral genomes at multi-organ level. *GigaScience*, 9(8):giaa086, 2020.

[13] L. Pyöriä et al. Unmasking the tissue-resident eukaryotic dna virome in humans. *Nucleic Acids Research*, 51(7):3223–3239, 2023.

[14] P. Rice et al. Emboss: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

[15] M. Silva et al. Efficient dna sequence compression with neural networks. *GigaScience*, 9(11):giaa119, 2020.

[16] A. Váradi et al. Rapid genotyping of targeted viral samples using Illumina short-read sequencing data. *Plos one*, 17(9):e0274414, 2022.

[17] D. E. Wood et al. Improved metagenomic analysis with kraken 2. *Genome Biology*, 20:1–13, 2019.