Fragmenting Genome Sequences by Coding Regions to Improve Performance of the AmpliDiff Algorithm for Large Genomes

Author: Samuel Karskens
s.m.f.karskens@student.tudelft.nl
Responsible Professor: Jasmijn Baaijens
Supervisor: Jasper van Bemmelen

1. Introduction

AmpliDiff algorithm [4] finds parts of DNA that can differentiate between different input genomes and, at the same time, finds their corresponding primers.

- Primers are small parts of DNA that bind to a DNA strand and are used as a starting point for replication in Polymerase Chain Reactions (PCR) to replicate DNA.

AmpliDiff uses Multiple Sequence Alignment (MSA) to align the input genome sequences, which facilitates the identification of DNA sequences that differ between the input genomes.

- MSA takes a long time for larger genomes

AmpliDiff is not able to scale well to larger genomes since it needs to go over all reference genomes in each phase of the algorithm

2. Research question

The main research question: How to modify the existing algorithm to make it more scalable to larger and more complex genomes?

Sub-questions:

- How to fragment the input genomes to minimize the loss of potential primers and discriminatory amplicons?
- How to reduce the runtime of the AmpliDiff algorithm by fragmenting the input genomes?

3. Methods

The new pre-processing strategy:

1. Fragment input sequences by coding regions
2. Caculate the differentiability of every coding region separatly with Mash [3]
3. Use MSA on the most differentiating coding regions
4. Combine aligned coding regions

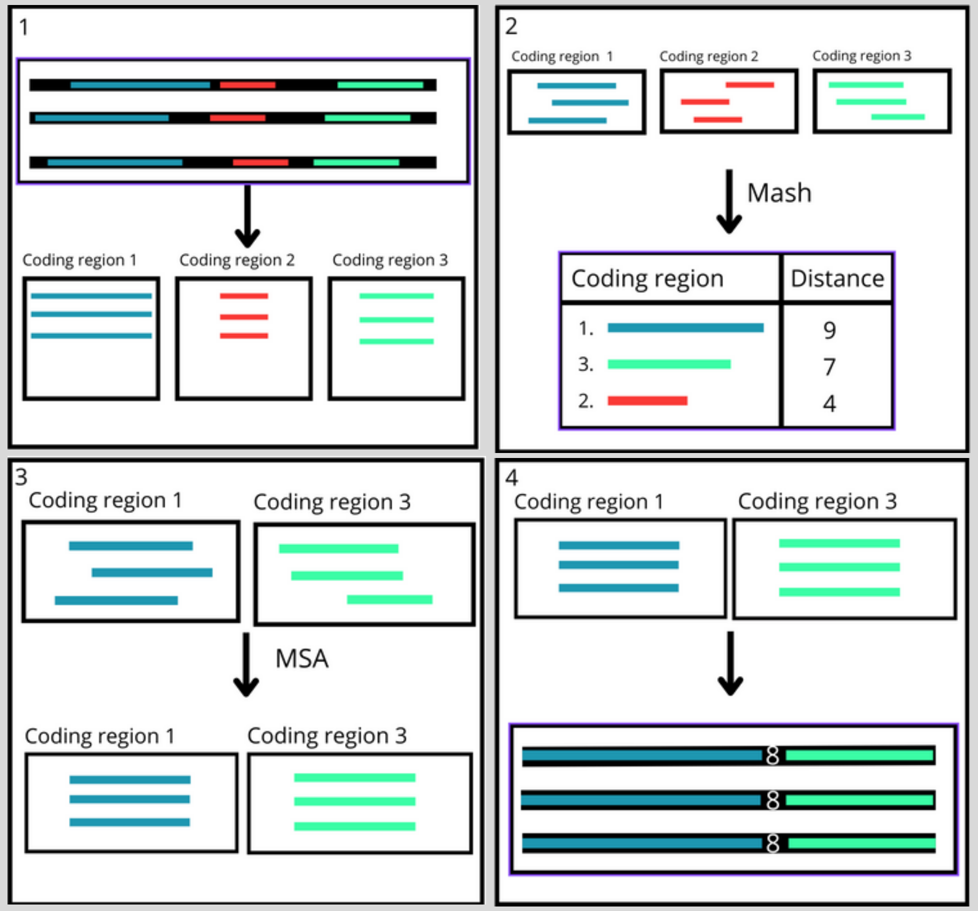

Figure 1: overview of the pre-processing strategy

Modifications of AmpliDiff:

Since the modification of AmpliDiff does not know about the DNA between the coding regions anymore:

- Disallow AmpliDiff to select primers or discriminatory parts of DNA that have overlap between coding regions.
- Only allow AmpliDiff to find corresponding primers in the same coding region as the discriminatory part of DNA. Because it is possible that the space between the primer and the discriminatory part of DNA gets too large and therefore the part will not be copied in PCR.

4. Results

Abundance estimation using the VLQ[1] pipeline and simulated reads from ART[2]

Comparing modified AmpliDiff using the 5, 10 or 15 most differentiating coding regions, to AmpliDiff and whole genome sequencing (WGS)

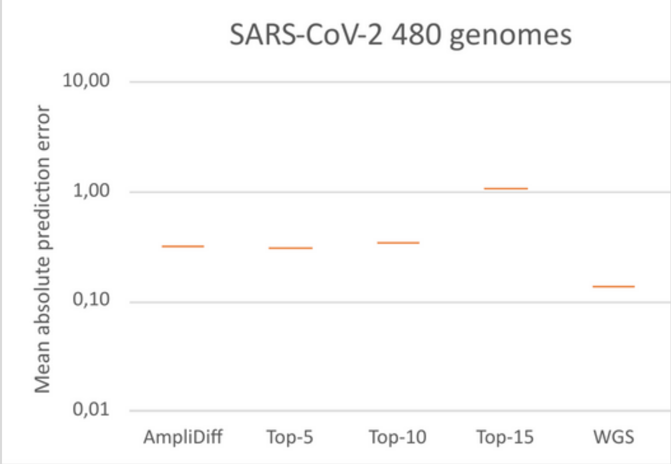

Figure 2: SARS-CoV-2: 480 genomes, 123 lineages. Showing top-5 and top-10 errors are comparable to AmpliDiff. Top-15 is worse than the top-10 and top-5. WGS has the lowest Mean absolute prediction error

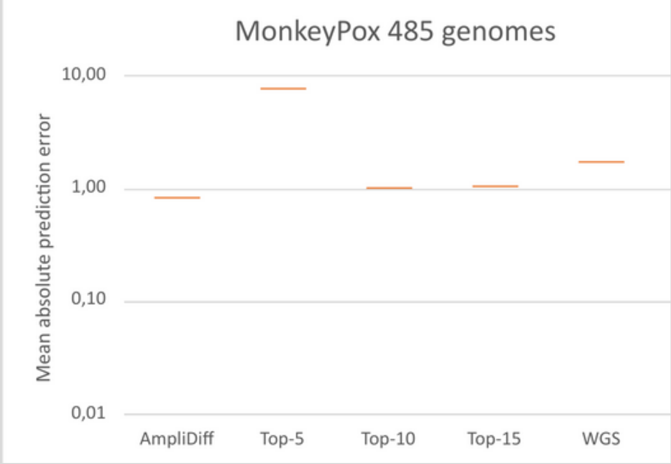

Figure 3: MonkeyPox: 485 genomes, 24 clades. Top-10 and Top-15 run are worse than AmpliDiff but still better than WGS

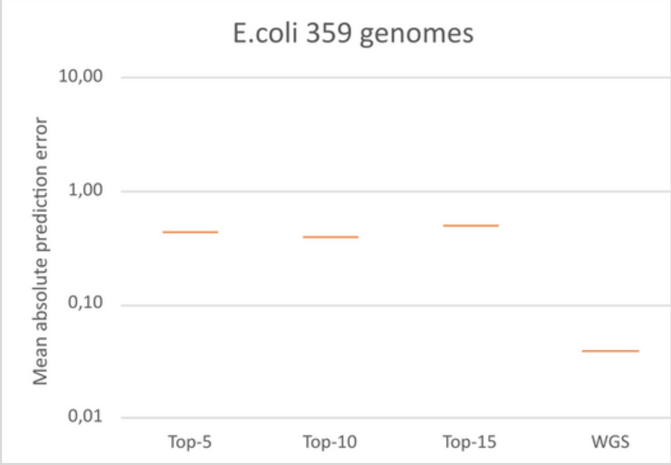

Figure 4: E. coli: 359 genomes, 358 strains. Top-5, Top-10 and Top-15 have all higher mean absolute prediction errors compared to WGS.

Runtime (including pre-processing)

Comparing AmpliDiff to modified AmpliDiff using the 5, 10 or 15 most differentiating coding regions.

- 185GB RAM and 12 CPU cores used for every run

Table 1: SARS-CoV-2

AmpliDiff	Top-5	Top-10	Top-15
6,11h	7,11h	5,74h	7,42h

Table 2: MonkeyPox

AmpliDiff	Top-5	Top-10	Top-15
13,13h	0,44h	0,65h	0,97h

Table 3: E. coli

AmpliDiff	Top-5	Top-10	Top-15
>24h	0,55h	0,83h	0,87h

5. Future work

- Preventing unwanted amplifications, since only coding regions are used, the other parts of DNA are not checked for primers occurring multiple times. This is a potential explanation for the increase in error from Top-10 to Top-15
- Research how many coding regions we actually need

6. Conclusion

- For SARS-CoV-2 the new pre-processing did not make the process faster, due to extra constraints and having to calculate distances between coding regions.
- For MonkeyPox and E.coli, a substantial reduction can be seen in runtime as well as comparable results in abundance estimation to the original AmpliDiff algorithm.

The new method looks promising but needs more research to verify and be applied correctly.

7. References

[1] Jasmijn A Baaijens, Alessandro Zulli, Isabel M Ott, Ioanna Nika, Mart J van der Lugt, Mary E Petrone, Tara Alpert, Joseph R Fauver, Chaney C Kalinich, Chantal BF Vogels, et al. Lineage abundance estimation for sars-cov-2 in wastewater using transcriptome quantification techniques. Genome biology, 23(1):236, 2022.

[2] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. Bioinformatics (Oxford, England), 28(4):593–594, February 2012.

[3] Brian D. Ondov, Todd J. Treangen, Pall Melsted, ´ Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology, 17(1):132, June 2016.

[4] Jasper van Bemmelen, Davida S. Smyth, and Jasmijn A. Baaijens. “AmpliDiff: An Optimized Amplicon Sequencing Approach to Estimating Lineage Abundances in Viral Metagenomes”. In: (2023).