Bench marking AmpliDiff for Human Monkeypox, Hiv-1 and Influenza-A

Kevin den Boon

Supervisor(s): Jasmijn Baaijens, Jasper van Bemmelen

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2024

Name of the student: Kevin den Boon
Final project course: CSE3000 Research Project
Thesis committee: Jasmijn Baaijens, Jasper van Bemmelen, Chirag Raman

Abstract

AmpliDiff provides a method which takes a list of genomes and their lineages, and finds a set of amplicons and their primers in such a way that these amplicons can be used to differentiate between the lineages of a specific virus. While it has been shown that AmpliDiff find results comparable to whole genome sequencing for SARS-CoV-2 when looking at abundance estimations, it is not know how well it performs for other viruses, or what factors of a virus impacts the performance of the amplicons found by AmpliDiff.

In this paper we will be showing the effectiveness of AmpliDiff on Human monkeypox, HIV-1 and Influenza-A. By running AmpliDiff for the three viruses mentioned above, we obtain sets of amplicons, which are used to do a lineage abundance estimation. By then comparing the estimation to the know abundance we calculate the Mean Average Error (MAE). This MAE will then be used to compare against the MAE obtained from doing a abundance estimation based on whole genome sequencing. By comparing the amplicons against whole genome sequencing (wgs), we show that using viruses with longer genomes positively impacts the performance of the amplicons. We also show that the amount of misalignment characters added by the Multiple Sequence Alignemnt (MSA), impacts the required settings for AmpliDiff to find amplicons, and can negatively impact the MAE. Finally, we show that AmpliDiff can be run, with some minor changes to the code base, on segmented genomes, with performance similar to that of single segment genomes.

1 Introduction

Sequence analysis can be used to predict or find virus outbreaks [4; 2], or check how well treatments for a virus are working [3]. Generally wgs is used for this purpose, however a valuable alternative to wgs is targeted sequencing, also known as amplicon sequencing. By only amplifying these specific targets, amplicon sequencing is not only cost effective, but also provides the possibility for increasing the depth of the sequencing and a smaller resulting dataset, allowing for easier subsequent analysis [7]. AmpliDiff aims to provide an efficient alternative to whole genome sequencing in lineage abundance estimations. AmpliDiff achieves this by finding highly differentiable genome regions, or amplicons, and their corresponding primers. These can then be used in Polymerase Chain Reaction (PCR) [9] to, for example, find what specific lineages of a virus are in the current sample. PCR works by first splitting the strands of the double stranded DNA into single strands, then it attaches the primers to the strands, and finally the DNA polymerase extends the primers, thus multiplying the part of the DNA between the two primers. A visualization can be found in figure 1.

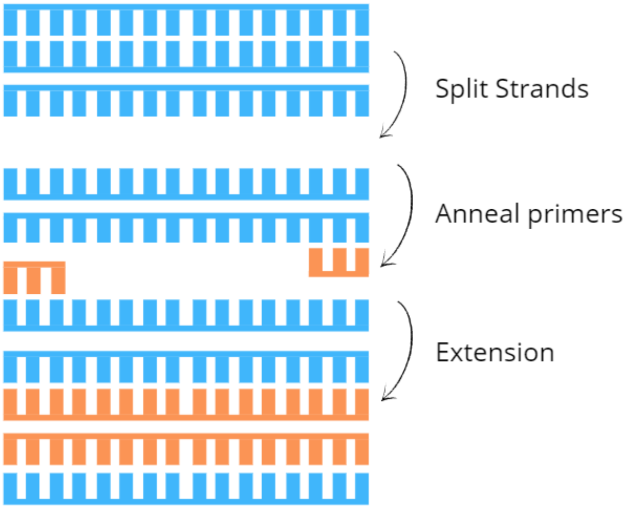

Figure 1: Visualization of the steps of PCR

It has been shown that for SARS-CoV-2, AmpliDiff can find amplicons and their corresponding primers, which can be used to discriminate between the different lineages reasonably well in comparison to wgs [14], however as no other viruses have currently been tested, it is currently not known whether this holds for any virus other than SARS-CoV-2.

In this paper we will be looking at bench marking the AmpliDiff software for HIV-1, human monkeypox and Influenza-A. By showing how the length of a genome can influence how well the resulting amplicons found by AmpliDiff do, when compared to wgs, we try to get a more general idea of how good the amplicons found by AmpliDiff would be for viruses in general. We will also be showing that it is possible to run AmpliDiff for viruses with a different organization of their genome, in this specific case using a segmented virus, which in turn enlarges the amount of possible viruses to run on AmpliDiff.

2 Methodology

AmpliDiff

AmpliDiff is a software finds target amplicons and the corresponding primers, in such a way that the resulting amplicons can be used to differentiate between lineages of a virus. To do this, first AmpliDiff takes the sequences, and uses it to create a database of feasible primers. Next up feasible amplicons are found, and the sequences which they can differentiate are stored. Finally AmpliDiff greedily goes through all the amplicons, selecting the ones that can differentiate the most sequences first, and checks for each amplicon if we have valid primers available. It continues this until either all the sequences are differentiable by the amplicons, or a configurable amount of amplicons has been found.

Picking viruses

The three viruses that were picked for this paper were HIV-1, human monkeypox and Influenza. HIV-1 and human monkeypox have been chosen to see whether the length of the virus genomes have an effect on the quality of the lineage

abundances created with the amplicons found by AmpliDiff. SARS-Cov-2 is generally around 29.500 nucleotides long. Here HIV-1 has been picked as a short genome, being around 9.500 nucleotides long, and human monkeypox as a longer genome, being around 197.000 nucleotides long.
Influenza, however was chosen for a different reason, namely the fact that Influenza is a segmented virus [17]. This means that in the specific case for Inluenza-A, the genome is split into 8 smaller segments, which together make up its genome with a size of around 14.000 nucleotides. To solve this issue for AmpliDiff, we have put the segments together in a single genome, and put a segment break character in between the segments, to make sure that AmpliDiff does not find any amplicons or primers spanning multitple segments.

Finding genome information

To run AmpliDiff with the 3 aforementioned viruses, we need to get both the genome data and their corresponding lineages. This data has been gathered from 3 different databases. Formonkeypox, Gisaid [19] was used, initially taking all genomes available using the complete and high coverage filters for the database. This dataset can be accessed in Gisaid by using Gisaids EPISET identifier system, using .... as the identifier. For HIV-1 the HIV sequence database [1] was used, filtering on HIV-1 and complete genomes. Finally, for Influenza-A, the NCBI influenza [6] database was used, selecting influenza-A, any subtype, full-length only and human host.

Pre-Processing

AmpliDiff requires two files to run, one being an MSA aligned fasta file, the second being a tab separated metadata file. As the databases found contained too many genomes to be able to run AmpliDiff in a feasible time, subsets of the datasets had to be created. In figure 2 the preprocessing pipeline can be found. As all required information needed for the metadata file can be found in the sequence fasta file, the resulting datasets were parsed to create the required metadata file. Stratified sampling was done using a Python script with the Pandas library, using a combination of its groupBy function and its sampling function.

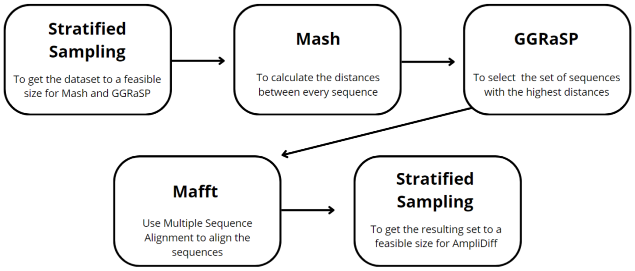
Stratified Sampling
To get the dataset to a feasible size for Mash and GGRaSP
Mash
To calculate the distances between every sequence
GGRaSP
To select the set of sequences with the highest distances
Mafft
Use Multiple Sequence Alignment to align the sequences
Stratified Sampling
To get the resulting set to a feasible size for AmpliDiff

Figure 2: Preprocessing pipeline used to create the datasets for AmpliDiff. The first stratified sampling step is used to get the dataset to a size on which GGRaSP could feasibly be run, and the second stratified sampling step was used to get the set to a set on which AmpliDiff could feasibly be run. Uses Mafft [16], Mash [18] and GGRaSP [10]

Evaluating amplicons and primers obtained by AmpliDiff

To evaluate the amplicons and primers found by AmpliDiff, we will be calculating the Mean Absolute Error (MAE). This will be done by using the VLQ pipeline [5] to calculate the abundances for the found amplicons, and doing the same using the whole genome sequences.
First a reference dataset and a simulation dataset are needed. This reference set is used by Kallisto [8] to create an index. In this case, the set created for the AmpliDiff runs was also used to create the Kallisto index. Next, we need a simulation set. This is created by randomly selecting sequences not included in the reference set. This simulation set is then used by ART [13] to create the reads required for Kallisto to do the abundance estimations. For the amplicon based reads, reads will only be generated from amplicons if the amplicon is amplifiable in a specific sequence. The results from Kallisto are abundances per reference. By summing those we can get the abundances per lineage. Finally, we can calculate MAE by using $MAE = \frac{1}{|L|}\sum_{l \in L} |\phi_l - \hat{\phi_l}|$ where ϕ_l is the actual abundance and $\hat{\phi_l}$ is the estimated abundance per lineage.

3 Your contribution

Improvement of an idea

AmpliDiff was build in such a way that it can only be run with single segment viruses. Influenza, one of the viruses used in this paper however, has eight segments for each genome. A solution could be to split the run into eight runs, and run each segment against each other. While it might be possible to find some differences in specific segments, there is a big chance that one segment does not carry enough information to identify all possible lineages. Another problem would be that the primers found for any amplicon in a specific segment, might also be able to bind to a different segment, and thus make the resulting amplicon unusable. As such AmpliDiff needs to have the information of all segments to create feasible amplicons.
A simple solution could be to just paste all the segments together in the same order for every genome. Thus creating a single segment sequence which could easily be run on AmpliDiff. This however brings a new problem, namely that it is now possible for AmpliDiff to create amplicons or primers that span multiple segments. As these segments are not actually pasted together. If the primers were used for a PCR run, the run would fail as the amplicon would not be able to be multiplied. A small visualization of this problem can be seen in figure 3. To combat this problem, instead of just pasting the segments together, an unused character is added in between segments, for this the ”8” was picked.

Next up, the code for AmpliDiff needed to be slightly changed to make sure it knows what to do with this new ”8” character. To do this, first, the amplicon generation code was changed, such that any amplicon that has the character either within the amplicon, or within the primer search width of the amplicon is rejected. Without checking the primer search window, it would be possible to have an amplicon on the edge

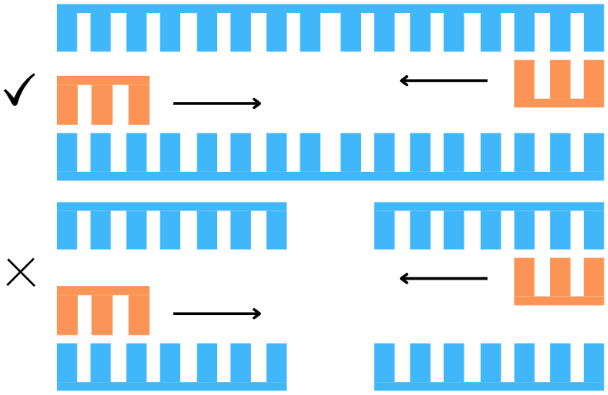

Figure 3: Visualization of a sequence with primers attached. The top sequence will correctly multiply, as the primers can fully duplicate the sequence, the bottom sequence will fail, as it has 2 segments and the primers cannot jump to the other segment to continue the duplication

of segment 1, and a primer on the edge of segment 2. Neither would have the break character in it, and both would thus be valid, however it would not be possible to multiply this amplicon. As AmpliDiff is already looping through every sequence when looking for valid amplicons, the actual implementation was a simple extra check during amplicon generation, which should not cause any significant time loss.

For the primers, there was already a check for degeneracy, which required to check every single character in a possible primer to calculate the degeneracy. As it is already looping through all the characters here, a simple if statement could be used to check for the new break character, and reject the primer if it is found. This should also not cause any significant time loss.

These minimal changes allow for a simple version of running segmented viruses, assuming the segments were aligned separately before pasting them together with the segment break character, to ensure the sequences all have the same length, and the segments do not get shifted in a way such that we are comparing different segments with each other.

4 Experimental Setup and Results

Running AmpliDiff

In table 1 we can see the settings used for the runs done in AmpliDiff. The settings used for monkeypox were the general settings used for all viruses. For HIV-1, as the MSA alignment added a lot of misalignment characters, the maximum allowed misalignment characters in an amplicon (mt), was first upped to half the amplicon width. This resulted in finding feasible amplicons, but no feasible primers, which was also the problem for Influenza-A. To try to solve this issue, first the maximum melting temperature difference for primers (mtd), was increased to 10, as this would increase the size of the primer database. This however was not enough and two more settings were added to separate runs, increasing the area in which we look for primers around the amplicons (sw) to 100, and increasing the self complementary threshold (sc) of the primers to 12.

Table 1: Settings used for the different AmpliDiff runs.

	cov	aw	mna	mt	mtd	sc	sw
monkeypox	[90,95,99.5]	[200,400]	50	x	x	x	x
HIV run 1	90	200	50	100	10	12	x
HIV run 2	90	200	100	100	10	x	100
Influenza run 1	90	200	50	x	10	12	x
Influenza run 2	90	200	100	x	10	x	100

Postprocessing

For all viruses postprocessing was needed to be able to do a lineage abundance estimation and calculate the MAE. This was done by following the VLQ pipeline [5]. To do this, first Art_illumina was used to do reads from the simulation dataset. Five sets of reads for each virus were generated. For the 200bp amplicons this was done using amplicon mode, HiSeqX TruSeq (150bp) paired-end reads, length of 125 bp x 2 and a read depth of 1000. For the 400 bp amplicons, amplicon mode, MiSeq v3 (250bp), length of 225 bp x 2 and a read depth of 1000 was used. For the whole genome sequencing of 200 bp width, HiSeqX TruSeq (150bp) paired-end reads, length of 125 bp x 2, a standard deviation of 10, a mean size of 200bp and a read depth of 100 was used. Lastly, for the 400 bp width, we used MiSeq v3 (250bp), length of 225 bp x 2, a standard deviation of 10, a mean size of 400 bp and a read depth of 100. The runs were done 5 times using seeds [40, 41, 42, 45, 57]. Next up, Kallisto was used to do the actual abundance estimations. First an index had to be created for Kallisto. The index for Kallisto was created using the input set for AmpliDiff. Then the Kallisto quantification algorithm was run with the above created index and the five pair reads generated by ART. Finally, the Kallisto data was parsed and compared to the lineage abundances in the simulation set, and the MAE was calculated.

Results

Pox

For monkeypox, we can see in figure 4 that for the 200bp width, amplicon selection vastly outperforms the wgs. Using a larger width mostly impacts the five amplicon based reads in a negative way, and the wgs in a positive way, bringing it more in line with the amplicon based reads. The amount of amplicons used seem to have a relatively small impact on the abundance estimations, as the MAE keeps hovering around the 0.65 mark. While monkeypox has been run for 3 different amplifiabilities, all 3 resulted in the exact same amplicons being picked, and as such the results for MAE was also the exact same. Thus there are no separate bars or tables for the other amplifiabilities. In figure 5 we can see the amplifiabilities for the actual amplicons based on the dataset used for the AmpliDiff run. Comparing this to the results in figure 4, we can see that, when comparing the 200bp width and 400bp width, a bigger distance between the amplifiabilities, seems to lead to a bigger distance between the widths in the calculated MAE. It also shows us, just like the MAE's that the 200bp runs generally outperform the 400bp runs.

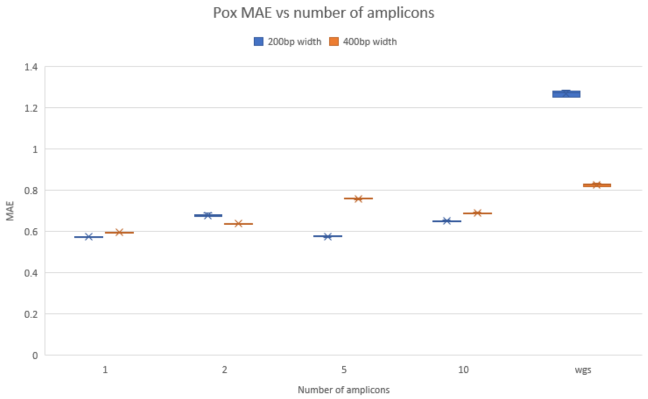

Figure 4: Box plot of the MAE calculated for the five art reads

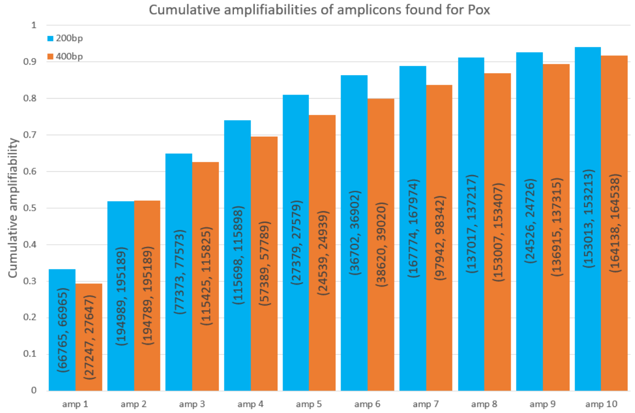

Figure 5: Bar graph showing the cumulative amplifiability of the amplicons for both the 200bp and 400 bp width. Also includes the locations of the found amplicons.

Influenza-A

In figure 6 the results for Influenza-A can be found. From this figure, it can be seen that while the 200bp width wgs has a better result than the rest, both of the Influenza-A runs seem to hold up to the 400 width wgs, and the sc12 run even outperforming the 400bp wgs run by a margin of 0.07. An interesting thing to note here, is that the 200bp wgs seems to do better than the 400bp wgs, which is the exact opposite from what happened for the monkeypox runs. Examining the exact lineage estimations, it can be seen that for all runs the biggest error is found in either the H1N1 or H3N2 lineages, which are also the most prevalent lineages, taking up 41% and 46% respectively of that entire simulation dataset. These have errors of around 6 for all runs except the wgs 200bp run, which only has an error of 3, which is also the reason the 200bp wgs outperforms the rest of the runs.

HIV

Finally in figure 7, we can see that the both runs of the wgs outperform the amplicon runs. For the amplicon runs itself, changing the self complementary threshold for the primers seemed to have a better effect on the outcome then increasing the primer search width, which can also be seen in the Influenza-A results. Also, while relatively the differences in the MAE might be fairly big, as there is about a 4 times difference between the 200bp wgs run and the sw100 run, absolutely the differences are very similar to the results for the other viruses. A final thing to note for HIV-1, is that the MSA alignment resulted in a relatively big increase in size of the virus. While monkeypox went from 197000 to 207000 nucleotides, and Influenza-A 13000 to 13500, an increase of 5% and 3.8% respectively, HIV-1 went from 9700 to 15000, an increase of 54%.

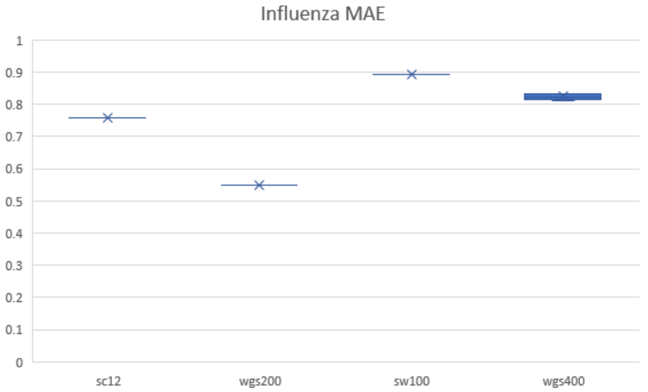

Figure 6: Box plot showing the MAE for the 200bp and 400bp width wgs, the 200bp width with 100bp primer search width run (sw100) and the run using a primer self complementary threshold of 12 (sc12) for Influenza-A

Another thing to note for both the HIV-1 and the Influenza-A results, is that only a single amplicon was found, with which all the sequences in the reference set could be differentiated. However this also means that if this amplicon cannot be amplified with the given primers, we have no information about a specific sequence at all. This means that the dataset used was too small or the sequences were too similar to each other to require multiple amplicons, or that this amplicon is very good at differentiating the lineages of the virus, given the amplicon can be amplified.

General comparison

When comparing the different viruses to each other, the first noticeable thing would be the differences in scale between the MAE's. These however are not very interesting to look at, as they are wholly dependent on the amount of lineages that were used. What is interesting however is their relative relation to the their respective wgs. By taking the average of the amplicon MAE's (μ_a) and the average of the wgs MAE's (μ_w) for a specific virus, we can find the relative error with: $\frac{\mu_a - \mu_w}{\mu_w} * 100$. This results in table 2. One more thing to note, is that while checking whether an amplicon was amplifiable on the set of simulation sequences, it was found that for monkeypox, about 2.5% of the sequences contained an amplicon that was not amplifiable. For HIV-1 and Influenza-A however, about 20% of the sequences contained an amplicon that was not amplifiable.

For both HIV-1 and Influenza-A we can see that the runs with the increase in self complementary threshold performs better

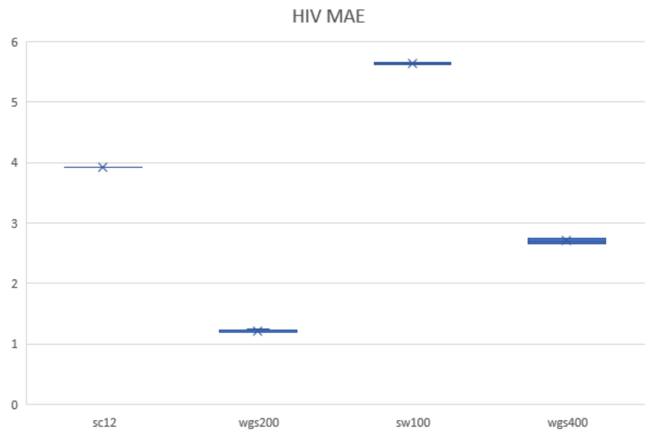

Figure 7: Box plot showing the MAE for the 200bp and 400bp width wgs, the 200bp width with 100bp primer search width run (sw100) and the run using a primer self complementary threshold of 12 (sc12) for HIV-1

than the runs with the increased search width.

Table 2: Table showing the average difference between amplicon based MAE and wgs MAE. negative values indicate that the amplicon based MAE outperformed wgs

monkeypox	Influenza-A	HIV-1
-38.9%	20.2%	144.1%

5 Responsible Research

For any research, reproducibility is a key point. To make sure the experiments done in this paper are fully reproducible, the databases and filters used to acquire the sequences have been detailed in the methodology chapter. For HIV and Influenza, a list of the accessions used for the various datasets can be found on github (https://github.com/Kdenboon/CSE-3000-Research-Project) [15]. For Pox the set can be found on Gisaid, using the following identifier: .... Furthermore, the exact settings used to run AmpliDiff have been provided in the Experimental Setup chapter, and as there is no randomness to AmpliDiff, this should result in the same amplicons being found by anyone trying to recreate this experiment. For ART, specific seeds have been used and added to the experimental setup, to make sure to keep out any randomness and guarantee full reproducibility. All tests were run on the Delft High Performance Computing centre. [11].

All collected genome data has been cited and referenced following the guides for referencing the various databases.

This paper has been written keeping the Netherlands Code of Conduct for Research Integrity [12] in mind.

6 Discussion

The goal of this paper was to get an idea of how well AmpliDiff works for non covid viruses. Here HIV-1 clearly has the worst results, seeing its relatively big gap between the amplicon results and the wgs results. This is probably caused by the MSA alignment, which increased the size of the HIV-1 sequences by 54%, which made it harder for AmpliDiff to find both feasible amplicons and primers. While this was counteracted by changing AmpliDiff settings to allow for more dash characters in both the amplicons and primers, resulting in one feasible amplicon and primer combination, this still seems to negatively impact the performance of AmpliDiff.

Seeing how HIV-1, the virus with the shortest genome length, has the worst results when comparing the AmpliDiff amplicons against the wgs, and monkeypox the best results, an assumption can be made that the length of a virus has significant impact on how well AmpliDiff works in comparison to wgs. Including the results found for Covid in [14], we can see that the results found for Covid are similar to Influenza, where the amplicons perform either worse or similar then wgs. The length of Covid is 29000, around twice the size of Influenza. Doing the same calculation as for the results of table 2 on Covid, taking the results of both the Texas and Netherlands set in account, results in wgs being better by about 36.3% on average. Looking at the Covid results however also clearly indicates that there can be significant differences between datasets, as using only the Texas dataset for example would result in wgs being better by only 10.5%.

Next up, we would like to discuss the size of the reference sets. All the reference sets have a size of around 100 genomes per virus, as it was currently not feasible to run bigger sets of data. This means that the sets the amplicons are based on are fairly limited in the data they can provide, and might thus result in amplicons that are not optimal when looking at the full database for a virus. This could have had a negative influence especially on HIV-1 and Influenza, where only a single amplicon has been found. While running a bigger dataset for Influenza should not run into more problems than an increased runtime, for HIV-1 another problem does show up in the MSA alignment. When aligning bigger sets of sequences, the size and the amount of characters inserted by the alignment increases. In an earlier test with 400 sequences, the alignment resulted in a sequence size of 27000 characters, and aligning the whole HIV-1 dataset resulted in a length of 54000. This means that HIV-1 might be too different between its lineages for AmpliDiff to find any feasible amplicons or primers and that HIV-1 might not be a good candidate to run on AmpliDiff.

The clear best results for the amplicon based MAE in these experiments for AmpliDiff comes from the monkeypox results. It is the only virus that outperforms the wgs, for both 200bp and 400bp widths, and for any amount of amplicons used. Once again this is probably because of the length of the genome. This would make sense, as the amplicon based reads captures the specific spots on which we know we can differentiate the lineages of this virus. While wgs looks at the whole virus and thus has to deal with data that might not help differentiate between the lineages at all, or even influence the abundance estimation in a negative way. For monkeypox, another notable result was that the 200bp runs generally outperform the 400bp runs. This could be explained by the 400bp amplicons not being able to find primers in some highly differentiable areas, while the 200bp amplicons could, resulting in a better MAE for the 200bp runs.

Next, looking at the results for Influenza, the only problem we ran into was that we could not find any feasible primers for the amplicons that were found. While the Influenza genome alignment was not a problem, the size of the segments might be. As we cannot find any amplicons spanning multiple segments, AmpliDiff is really just checking a bunch of relatively small genomes to each other. In the case of Influenza these segments are anywhere between the length of 2500 and 800 basepairs. Looking at the results, seeing that Influenza has a single amplicon found, with 14 forward and 12 reverse primers, it seems that the area in which the differentiable amplicons are found for Influenza also has a relatively different area around it in which the primers must be found. This would explain why using both a smaller set of data, and using looser constraints for the primers helped AmpliDiff find a feasible combination of amplicon and primers.

Finally, for both Influenza and HIV-1, only a single amplicon was found. During the creation of the reads, in about 20% of the sequences this amplicon could not be amplified. As any amplicon not amplifiable was not added to the reads, it follows that for the case of Influenza and HIV-1 the full sequence was not added to the read data. This resulted in less sequences for Kallisto to identify, which might have impacted the results.

7 Conclusions and Future Work

AmpliDiff is a methodology that finds target regions, or amplicons, from a set of viral genomes, in such a way that we can use these amplicons to differentiate between the different lineages of a virus. In this paper we have looked at benchmarking AmpliDiff for the human monkeypox, HIV-1 and Influenza-A viruses, by comparing them to the usually used whole genome sequencing. This resulted in 3 sets of data, from which we can see that for HIV-1 the amplicons generally perform worse then wgs, for Influenza-A they perform about equal to wgs, and for Pox they perform better then wgs. From this it was concluded that we generally expect AmpliDiff to perform better for longer viruses. In the specific case of Influenza-A, we have made small changes to AmpliDiff to allow us to run a segmented genome. Running Influenza-A with these changes, there did not seem to be any other problems stemming from the virus being segmented.

Future recommendations

To see whether the hypothesis of longer viruses resulting in a comparatively better MAE, either more viruses of different lengths could be tested on AmpliDiff, or different datasets of the already run viruses could be tested. The first option to get a wider range of viruses and lengths involved, the second to make sure we did not find any outliers with the currently used sets, as a sizable difference has already been shown between the two Covid sets in [14].
Furthermore in the case of HIV-1 and Influenza-A, as only a single amplicon was found, it would be interesting to see how using a larger reference set would impact the eventual MAE. The expectation would be that in a larger dataset, more sequences need to be differentiated, possibly also against more lineages, and thus this one amplicon might no longer be enough, or a different amplicon might be found depending on the feasibility of the primers. This would also help in making sure the full simulation dataset is actually tested in Kallisto, as when we have multiple amplicons, not being able to amplify a single amplicon for a sequence no longer means we have no reads for that sequence. Finally, running a longer virus which has a bad MSA alignment. As currently HIV-1 is the only virus that we have run that is so different between its sequences, that the alignment increased the size of the genome by over 50% for datasets of around 100 genomes, and up to 450% when using the complete set of HIV-1 data. While HIV-1 has a shorter genome compared to the other viruses run here, this might also have had a negative influence on the results, by running a virus with a comparative length to e.g. Covid, we might be able to get an idea of how much the MSA alignment impacts the final results.

8 acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

References

[1] Hiv sequence database. https://www.hiv.lanl.gov/. Accessed: 2023-12-18.

[2] W. Ahmed, V. J. Harwood, P. Gyawali, J. P. S. Sidhu, and S. Toze. Comparison of concentration methods for quantitative detection of sewage-associated viral markers in environmental waters. 81(6):2042–2049.

[3] Warish Ahmed, Nicola Angel, Janette Edson, Kyle Bibby, Aaron Bivins, Jake W. O'Brien, Phil M. Choi, Masaaki Kitajima, Stuart L. Simpson, Jiaying Li, Ben Tscharke, Rory Verhagen, Wendy J. M. Smith, Julian Zaugg, Leanne Dierens, Philip Hugenholtz, Kevin V. Thomas, and Jochen F. Mueller. First confirmed detection of SARS-CoV-2 in untreated wastewater in australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. 728:138764.

[4] Manuel Ampuero, Constanza Martínez-Valdebenito, Marcela Ferrés, Ricardo Soto-Rifo, and Aldo Gaggero. Monkeypox virus in wastewater samples from santiago metropolitan region, chile. 29(11):2358–2361.

[5] 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 B. F. Vogels, Mallery I. Breban, Claire Duvallet, Kyle A. McElroy, Newsha Ghaeli, Maxim Imakaev, Malaika F. Mckenzie-Bennett, Keith Robison, Alex Plocik, Rebecca Schilling, Martha Pierson, Rebecca Littlefield, Michelle L. Spencer, Birgitte B. Simen, Ahmad Altajar, Anderson F. Brito, Anne E. Watkins, Anthony Muyombwe, Caleb Neal, Chen Liu, Christopher Castaldi, Claire Pearson, David R. Peaper, Eva Laszlo, Irina R. Tikhonova, Jafar Razeq, Jessica E. Rothman, Jianhui Wang, Kaya Bilguvar, Linda Niccolai, Madeline S. Wilson, Margaret L. Anderson, Marie L. Landry, Mark D. Adams, Pei Hui, Randy Downing, Rebecca Earnest, Shrikant Mane, Steven Murphy, William P. Hanage, Nathan D. Grubaugh, Jordan Peccia, Michael Baym, and Yale SARS-CoV-2 Genomic Surveillance Initiative. Lineage abundance estimation for SARS-CoV-2 in wastewater using transcriptome quantification techniques. 23(1):236.

[6] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the national center for biotechnology information. 82(2):596–601.

[7] Findlay Bewicke-Copley, Emil Arjun Kumar, Giuseppe Palladino, Koorosh Korfi, and Jun Wang. Applications and analysis of targeted genomic sequencing in cancer studies. 17:1348–1359.

[8] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. 34(5):525–527.

[9] Kirstie Canene-Adams. General PCR. 529:291–298.

[10] Thomas H Clarke, Lauren M Brinkac, Granger Sutton, and Derrick E Fouts. GGRaSP: a r-package for selecting representative genomes using gaussian mixture models. 34(17):3032–3034.

[11] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1, 2022.

[12] KNAW; NFU; NWO; TO2 federatie; Vereniging Hogescholen; VSNU. Nederlandse gedragscode wetenschappelijke integriteit. 2018.

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

[14] Jasper van Bemmelen, Davida S. Smyth, and Jasmijn A. Baaijens. AmpliDiff: An optimized amplicon sequencing approach to estimating lineage abundances in viral metagenomes. page 2023.07.22.550164. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[15] Jasper van Bemmelen and Kevin den Boon. Modified amplidiff for segmented genomes. https://github.com/Kdenboon/CSE-3000-Research-Project.

[16] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. 30(14):3059–3066.

[17] Florian Krammer, Gavin J. D. Smith, Ron A. M. Fouchier, Malik Peiris, Katherine Kedzierska, Peter C. Doherty, Peter Palese, Megan L. Shaw, John Treanor, Robert G. Webster, and Adolfo García-Sastre. Influenza. 4(1):3.

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

[19] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data - from vision to reality. 22(13):30494.