Improving algorithms for amplicon sequencing

Christopher Kent, Dr Joshua Quick

Author Affiliations: **University of Birmingham (IMI), ARTIC Network**

ARTICnetwork
http://artic.network | @NetworkArtic

Abstract

Tilling amplicon sequencing has emerged as a dominant method for viral whole genome sequencing, due to its; low cost, high sensitivity, and "anywhere" approach. Furthermore, due to the widescale adoption, and experience gained from SARs-CoV-2 sequencing, this is an ideal and impactful time to further expand the application of amplicon sequencing.

However, amplicon sequencing has traditionally struggled with; exponentially increasing numbers of primer-primer interactions with higher multiplexing, and covering highly diverse viral clades, due to variation within the primer binding sites reducing primer binding affinity, and hence causing reduced coverage of the amplicon. Hence sites with mutations have been avoided reducing the number of possible primers and hence amplicons available to use. Here we describe improved algorithms to mitigate diversity, utilising novel data structures, T_m normalization and dynamic primer length.

The scale of these improvements can be shown via the generation of a primer scheme for the recent outbreaks of Marburg Virus in West Africa.

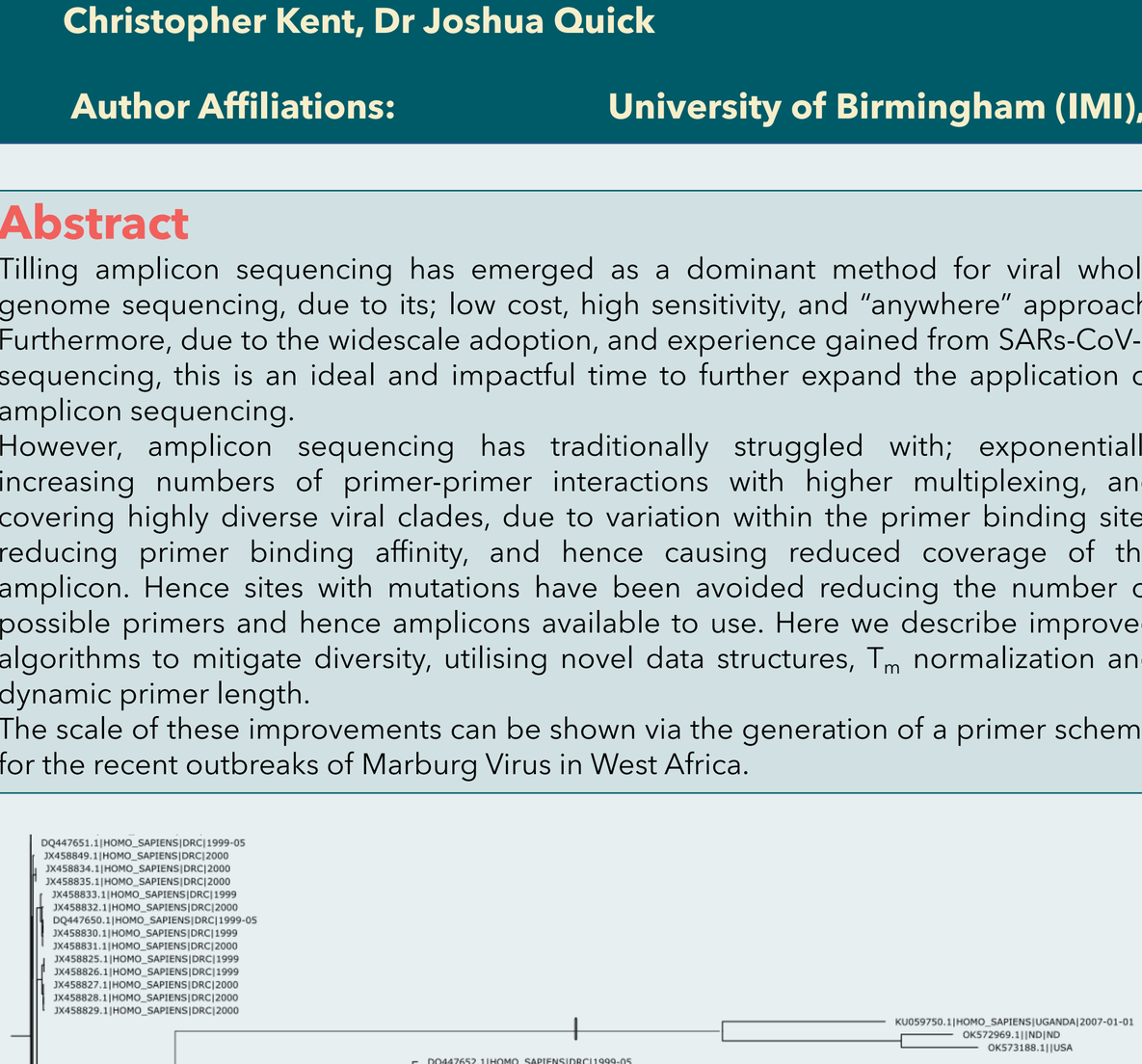

Figure 1.

During 2022, a Marburg Virus outbreak (shown in red) started to appear in West Africa **[1]**, followed by a partial genome from a separate outbreak in East Africa (shown in blue) in 2023.

Hence, any primer scheme made in response to this outbreak should be designed to be as broad as possible, and tolerant of all current variations to ensure it can prevent amplicon dropouts across these current outbreak clades and any flare-ups similarly related to this clade.

Therefore, a primer scheme called MARV-2023 was created utilising 38 similar genomes (shown in box), which encompasses all SNPs and Indels mutations found within any genomes similar to these outbreak clades.

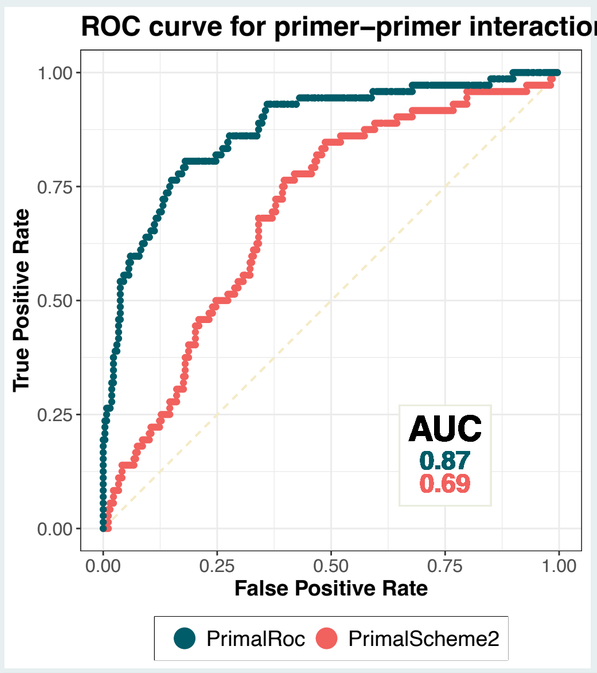

Figure 3.

A Receiver operating characteristic (ROC) curve for PrimalRoc (Blue), a custom implementation of the algorithm described in [2] vs the custom interaction checker utilised in PrimalScheme2 (Red). Each algorithm returns a numerical "dimer score" for how likely two primers are to form a 3' extendable dimer product.

Each point on the graph represents a different dimer score cut-off threshold for dimerization, and the statistical power each threshold produces, using a testing set of experimentally validated primer dimers.

PrimalRoc has higher specificity and sensitivity, allowing both; more primers to be multiplexed without interactions, and decreasing the number of primers incorrectly specified as a dimer-causing.

PrimalRoc

```
SARS-CoV-2_3_LEFT      5'-GTAATAAAGGAGCTGGTGGCCA
                                        ||||||
SARS-CoV-2_3_LEFT                    3'-ACCGGTGGTCGAGGAAATAATG

SARS-CoV-2_17_RIGHT    5'-TTCAACTCTATTTGTTGGAGTGTTAACAA
                                  ||||||||| | ||| |
SARS-CoV-2_71_LEFT           3'-GATAAACAAGCGCACCAAACGG

SARS-CoV-2_64_LEFT     5'-GCCTATTTTGGAATTGCAATGTCGA
                                           |||||||
SARS-CoV-2_60_RIGHT                   3'-TACAGCTCCCCACAGTACGAT
```

Improving interaction checkers enable looking back at past schemes, and interactions contained within them. Shown above are three interactions detected from the v4.1.0 SARs-CoV-2 Scheme.

Dynamic Digestion

```
Alignment                                  fprimer sequences          Tm
..AGTGCGGAAGTTGCAGTTAAAATGTTTGA..      CGGAAGTTGCAGTTAAAATGTTTGA 59.6 ºC
..AGTGCGGAAGTTGCAGCTA---TGTTTGA..        CGGAAGTTGCAGCTATGTTTGA 59.7 ºC
..AGTGCGGAAGTTGCAGTTAAAATGTTTGA..  AGTGCGGAAGTTATAGATAAAATGTATGA 59.8 ºC
..AGTGCGGAAGTTATAGATAAAATGTATGA..      GGAAGTTGCAGTTAAAATGCCTGA 60.6 ºC
..AGTGCGGAAGTTGCAGTTAAAATGTTTGA..      GGAAGATGCAGTTAAGGTGTTTGA 59.8 ºC
..AGTGCGGAAGTTGCAGTTAAAATGCCTGA..
..AGTGCGGAAGATGCAGTTAAGGTGTTTGA..
```

This figure shows how dynamic digestion can handle vastly increased variation located within a primer binding site, using a "cloud" of primers all anchored to the same genomic position on the 3' end. On the left, we have a slice of an MSA, for which we want to generate primers. On the right, we have fprimer sequencing to cover all mutations present in the slice, and the corresponding T_m for each sequence.

The 3' anchor has been shown to reduce the probability of both primer-primer interactions and off-target primer miss-priming, furthermore, the primer length is modulated to provide uniformity in primer annealing temperature. This method can reliably generate primers across regions with; ambiguous bases, insertions/deletions and high sequence diversity, or combinations of all. This approach does increase the number of primers, however, polishing tools that can remove primers with mutations that are non-significant to the primer's binding.

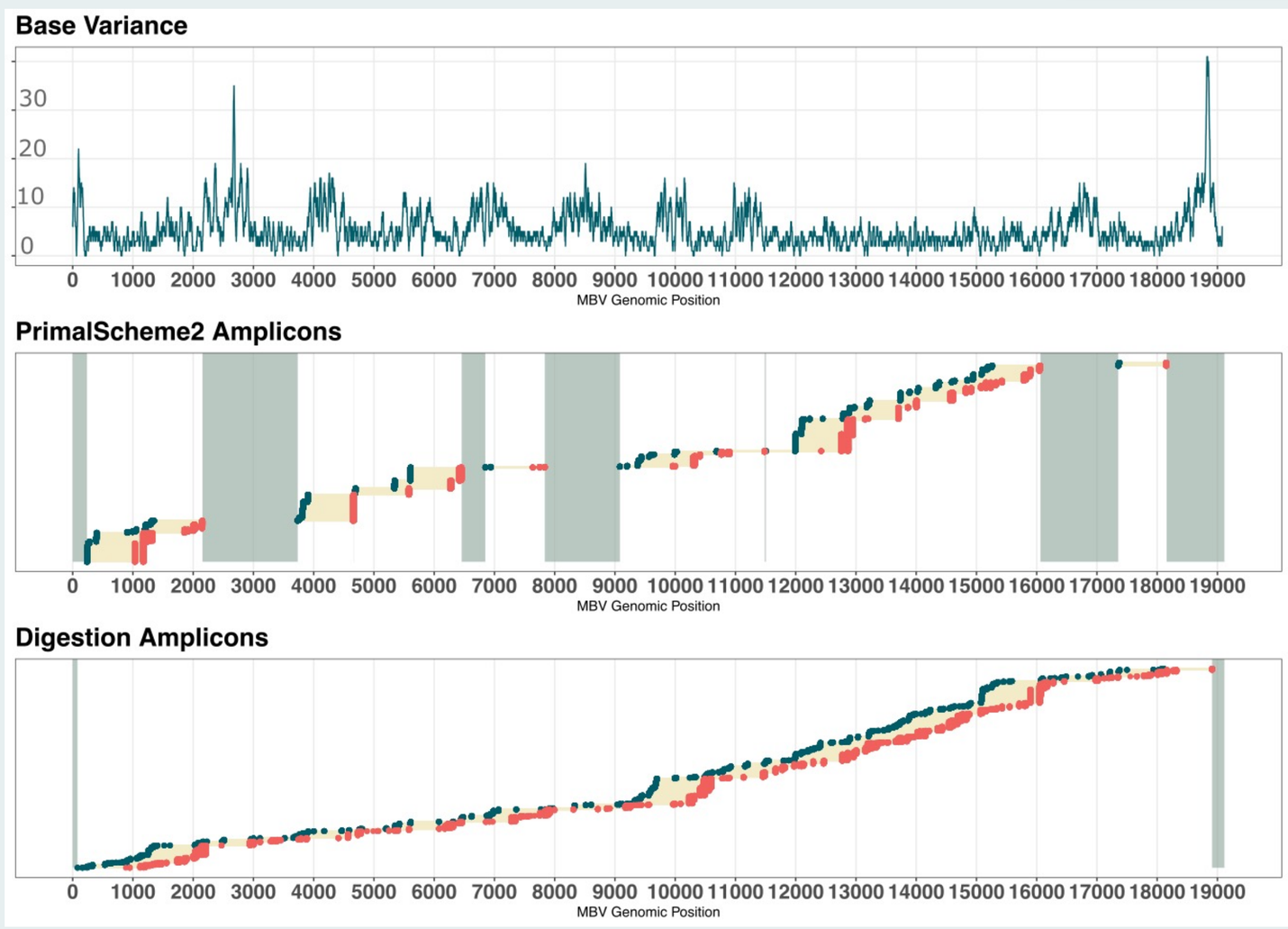

Figure 2.

From a multiple sequence alignment (MSA) of the 38 genomes shown in figure 1, we attempted to design a primer scheme, with 1000-bp amplicons using both PrimalScheme2 and a new research algorithm called dynamic digestion.

Base Variance: The line shows the number of unique mutations contained within a rolling 30mer region of the genome.

PrimalScheme2 Amplicons: This figure shows how using the traditional approach of N-Masking, where primers are prevented from being designed across genomic positions containing variation reduced the number of available amplicons.

Dynamic Digestion Amplicons: This figure shows how utilising both T_m normalisation and dynamic primer length vastly increases the number of available primers, amplicons and regions that can be covered.

Forward and Reverse primers are shown in Blue and Red respectively, with the formed amplicon shown in beige. Grey regions represent areas unable to be sequenced due to no possible amplicon coverage (after primer trimming)

References:

[1] https://virological.org/t/first-ever-outbreak-of-marburg-virus-disease-declared-in-ghana-2022/897

[2] Johnston, A.D., Lu, J., Ru, Kl. *et al.* PrimerROC: accurate condition-independent dimer prediction using ROC analysis. *Sci Rep* **9**, 209 (2019). https://doi.org/10.1038/s41598-018-36612-9

Acknowledgments:

Quick Lab (Natalie Sparks, Andy Smith, Dr Greg McCallum), **Loman Lab** (Dr Jeremy Mirza, Tom Brier, Sam Wilkinson, Dr Nicola Cumley)

UNIVERSITY OF BIRMINGHAM