A step towards the use of selective sequencing approaches for viral genomes

Cameron Robert Ferguson

Supervised by: Guillaume Fournié, Sarah Hill and Jayna Raghwani

Royal Veterinary College University of London

Introduction:

Selective sequencing refers to an Oxford Nanopore Technologies (ONT) sequencer's ability to reject individual DNA molecules while being sequenced and enrich for non-rejected molecules of interest. This is beneficial as the total amount of DNA that can be sequenced per flow cell is finite, so rejecting sequences early allows you to reduce the amount of unwanted molecules "waste" that is sequenced, hence leaving more available capacity for wanted sequencing. Traditional methods for enriching viral DNA or RNA content for sequencing rely on expensive reagents and machines, with every pre-processing step in the lab adding time and expense, making it often difficult and impractical to deploy them in the field. Due to their low cost and portability, the handheld sequencing devices ONT MinION's are an attractive option for in-field sequencing. Unsurprisingly there has been considerable interest in this unique ability of ONT sequencers. As such, many different programs and pipelines with diverse methodologies have been developed to take advantage of it. In this rotation, we focused on one such pipeline called Readfish, focusing on improving the metagenomic classifier section of the pipeline.

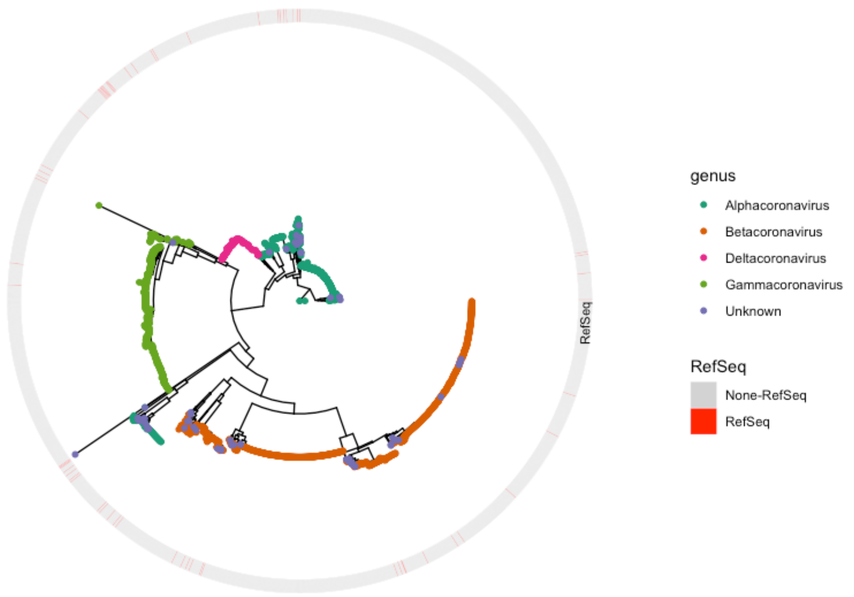

Fig.1 Maximum likelihood phylogenetic tree coloured by genus with sequences in RefSeq database shown in red in the outer circle.

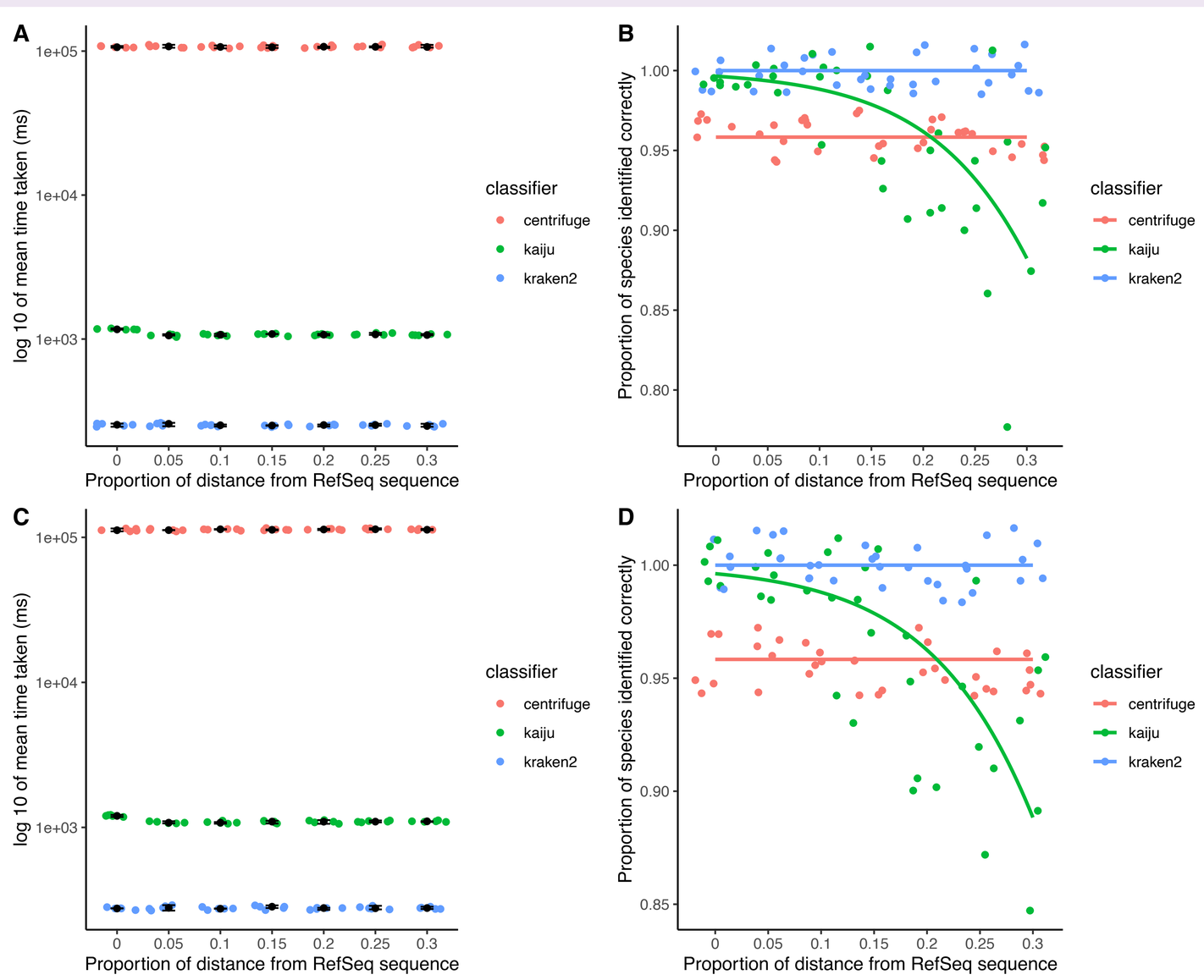

Fig.2 Classification assessment on the simulated data without homopolymer errors. A) A scatter plot depicting the log10 of mean time taken against the proportion of the distance from the closest respective RefSeq sequences for the data without homopolymer errors. For each proportion, the mean time for each simulated dataset is plotted for each classifier (n = 5 replicates of simulated data), and then the median for each classifier is plotted as a black dot with error bars. B) A scatter plot depicting the proportion of species identified correctly by each classifier against the proportion of the distance from the RefSeq sequences for the data without homopolymer errors. For each proportion of the distance from the RefSeq sequences, the proportion of correctly identified species for each simulated dataset is plotted for each classifier (n = 5 replicates of simulated data). C) A scatter plot depicting the log10 of mean time taken against the proportion of the distance from the closest respective RefSeq sequences for the data with homopolymer errors. For each proportion, the mean time for each simulated dataset is plotted for each classifier (n = 5 replicates of simulated data), and then the median for each classifier is plotted as a black dot with error bars. D) A scatter plot depicting the proportion of species identified correctly by each classifier against the proportion of the distance from the RefSeq sequences for the data with homopolymer errors. For each proportion of the distance from the RefSeq sequences, the proportion of correctly identified species for each simulated dataset is plotted for each classifier (n = 5 replicates of simulated data).

Assessment of the classifiers :

Perhaps unsurprisingly, our analysis indicates that the classifier used has a highly significant effect (df = 2, p = 0) on the mean time taken for the classifier to run, where centrifuge performed significantly worse than kaiju and kraken2 ($z = 1056.524$, $p < 0.0001$ and $z = 779.143$, $p < 0.0001$ respectively), and kaiju performed significantly worse then kraken2 ($z = 166.443$, $p < 0.0001$) (Fig.2A). Interestingly, our analysis has found that whether the data contains simulated homopolymer errors or not has a highly significant effect (df = 1, p = 0) on the mean time taken for the classifier to run.

Furthermore, it appears that the effect of the sequence divergence from the RefSeq sequences on the proportion of species correctly identified is dependent on the classifer used ($F = 18.589$, df = −12, $p < 2.2 \times 10{-}16$) with both main effects (the classifier used and the distance from the RefSeq sequences) having a highly significant effect ($F = 158.52$, df = −2, $p < 2.2{\times}10{-}16$ and $F = 13.388$, df = −6, $p = 9.43{\times}10{-}13$ respectively).

Assessing the distance between the sequences and the closest RefSeq sequence:

The maximum likelihood phylogenetic tree constructed from the spike protein gene alignment (Fig.1) revealed that most coronaviruses for which genus classification has not yet been assigned clustered within the *Alphacoronavirus* and *Betacoronavirus* genera, with only one of the unknown-genus viruses clustering quite basally within the *Gamacoronavirus* genus and none within the *Deltacoronavirus* genus. The distribution of the RefSeq sequences within the tree is far from uniform with the *Gammacoronavirus* genus seeming to have relatively few RefSeq sequences when compared to the number of sequences in the genus. Assessment of the pairwise distance between sequences indicates that 98% of the sequences in the alignment are between 0-30% different at the nucleotide level from the closest RefSeq sequence. With only 8% of the sequences being more then 20% different from the closest RefSeq sequence in regions of high homology.

Summary:

In summary our analysis of three classifiers (kraken2, kaiju and centrifuge) revealed that potential substantial gains in classification speed could be achieved by changing the ReadFish metagenomic classifier from centrifuge to kraken2. This could mark a significant first step towards improving the viability of the ReadFish pipeline for viral genomes enabling the use of read-until approaches with shorter-segmented viruses like influenza.

Show me the maths!

LIDo London Interdisciplinary Doctoral Programme