COVID-19 Situation Report - 2020-04-16

Mapping the phylodynamics of COVID-19 cases in New Zealand using Transcendental Information Cascades

Markus Luczak-Roesch[1,2]

1. Victoria University of Wellington, School of Information Management, New Zealand
2. Te Pūnaha Matatini - the Centre for Complex Systems and Networks, New Zealand

Methods

We construct Transcendental Information Cascades over the genomic sequences ordered by their date of collection as described in [1]. The tokenisation captures all unique codons in a particular position in the genomic sequences (truncated flanks to the consensus range 56 to 29,797 [2]) and in the +1, +2 and +3 reading frame. This results in a directed acyclic graph from which we can derive an unrooted phylogenetic tree through community detection.

Results

While we mostly observe the expected pattern of the strongest codon similarity path along the sequence at which virus samples were collected, we also observe a number of links across a longer time span that feature a slightly higher number of similar codons that they do not share with their immediate successors (cf. see Figure 1). Some of these cases and their details are:

- Cases 1 - 20RV0174, 3 - 20RV1109, 5 - 20RV0206 and 7 - 20RV0461 in the network are samples that were taken in three different locations in New Zealand (Auckland, Waikato and Southern) over a time period of 15 days. Cases 1 - 20RV0174 and 5 - 20RV0206 identify as female and male respectively, are both above 60 years of age and have a documented travel link to Iran. The other two individuals from Waikato and Southern DHB are both male and 41 and 33 years of age. Their connection to each other and to the aforementioned travellers is unknown but the codon similarity may be suggestive of an infection pathway from one of the first two individuals to case 3 - 20RV1109.
- Case 9 - 20RV0276 has a known travel history to the USA as per available metadata. The fact that there is a significant link from case 7 - 20RV0635 in the network could be suggestive of an infection pathway in particular because case 7 - 20RV0635 is also located in Wellington even though no common travel link has been reported.
- Case 2 - 20RV0189 and Case 4 - 20RV0195 are documented to have travelled to Italy. They were sampled at a distance of 2 days. Both individuals identify as female and are in their 40s. The link to the intermediate case 3 - 20RV1109 is likely to not be

indicative of a similar origin of the infection because of the aforementioned significant link of case 3 - 20RV1109 to case 7 - 20RV0461.

- Case 8 - 20RV0275 and Case 9 - 20RV0276 are documented to have travelled to the USA and were sampled on the same day. The fact that in the network graph case - RV200275 has not outgoing connections to any other case samples suggests that the genomic sequences of these two cases are almost perfectly identical, which is suggestive of the same source of infection.

Further links will be investigated in more detail as more metadata about travel histories becomes available to us.

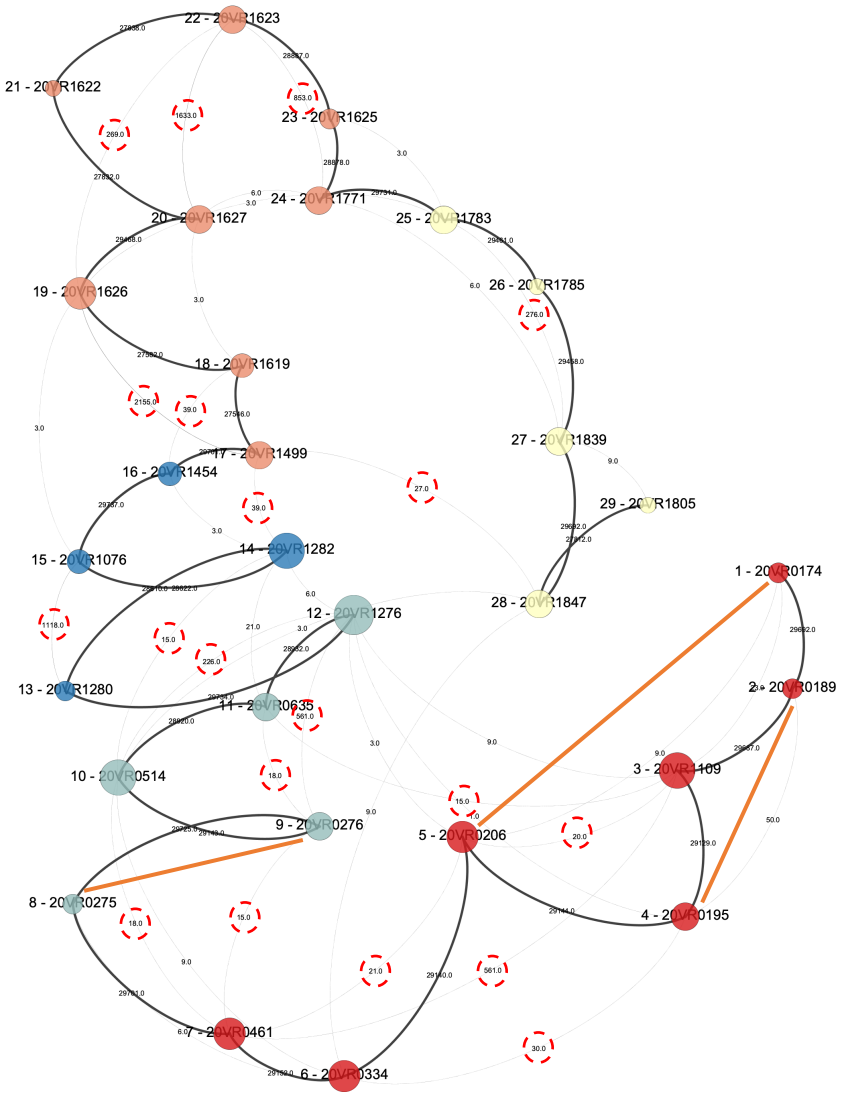

Figure 1: Transcendental Information Cascade network constructed from the genomic sequences using a codon tokeniser. Bold grey edges are the strongest similarities each sequence has with any preceding and succeeding sequences. Circled are weights of what we consider significant edges across longer time spans (sequence steps > 1). These similarities

are unlikely to be a result of general codon bias of the covid-19 genome [3,4]. Orange straight edges indicate known international travel links. Node colours reflect cluster membership.

The suggested infection pathway between cases 1 - 20RV0174, 3 - 20RV1109, 5 - 20RV0206 and 7 - 20RV0461 (or at least the similarity in the samples at the codon level) is also emphasised by their positioning in the same sub-tree within the phylogenetic tree we estimate from the hierarchical network clustering applied to the original Transcendental Information Cascade network (see Figure 2).

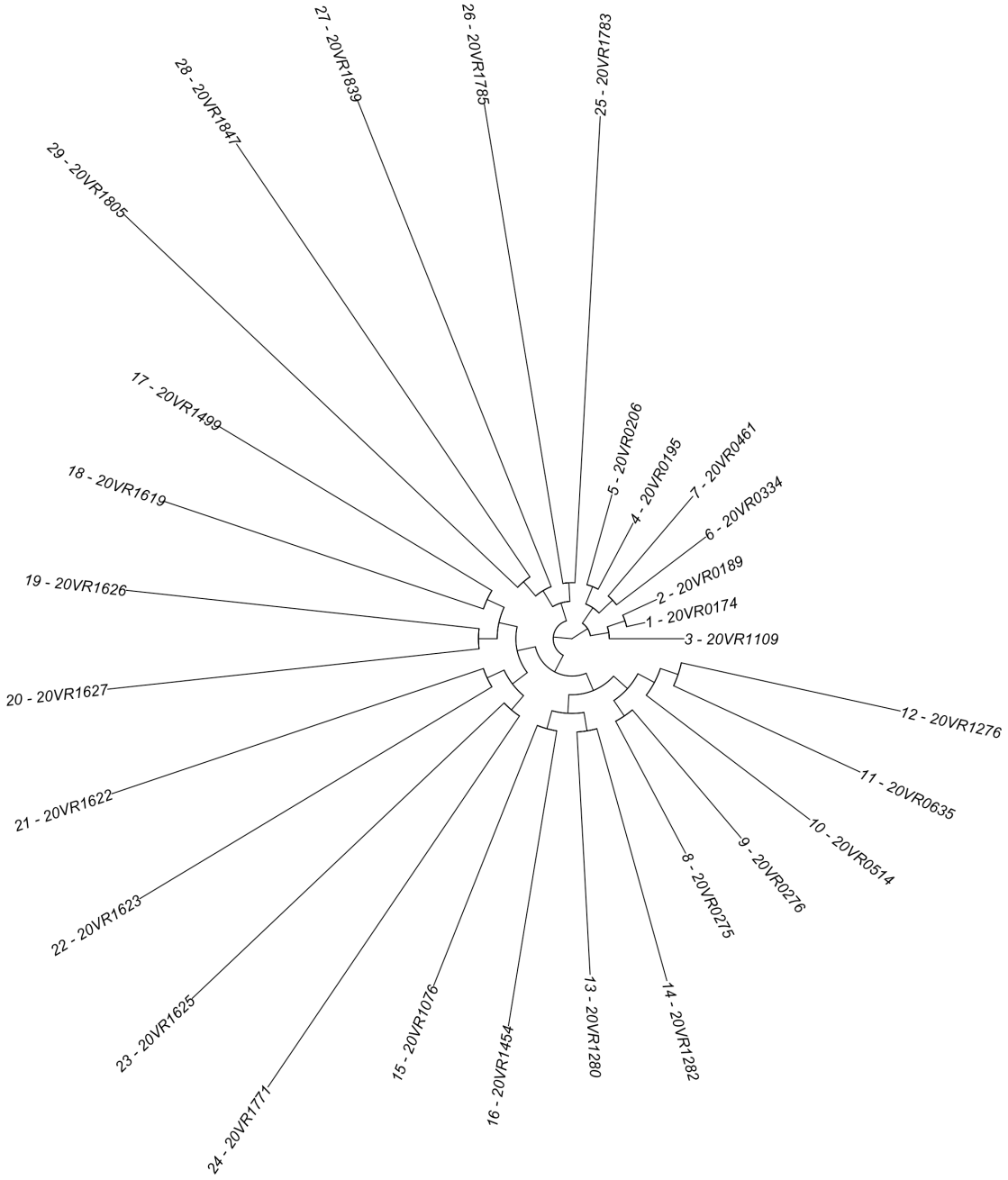

Figure 2: Unrooted phylogenetic tree derived from the clustering of the Transcendental Information Cascade network.

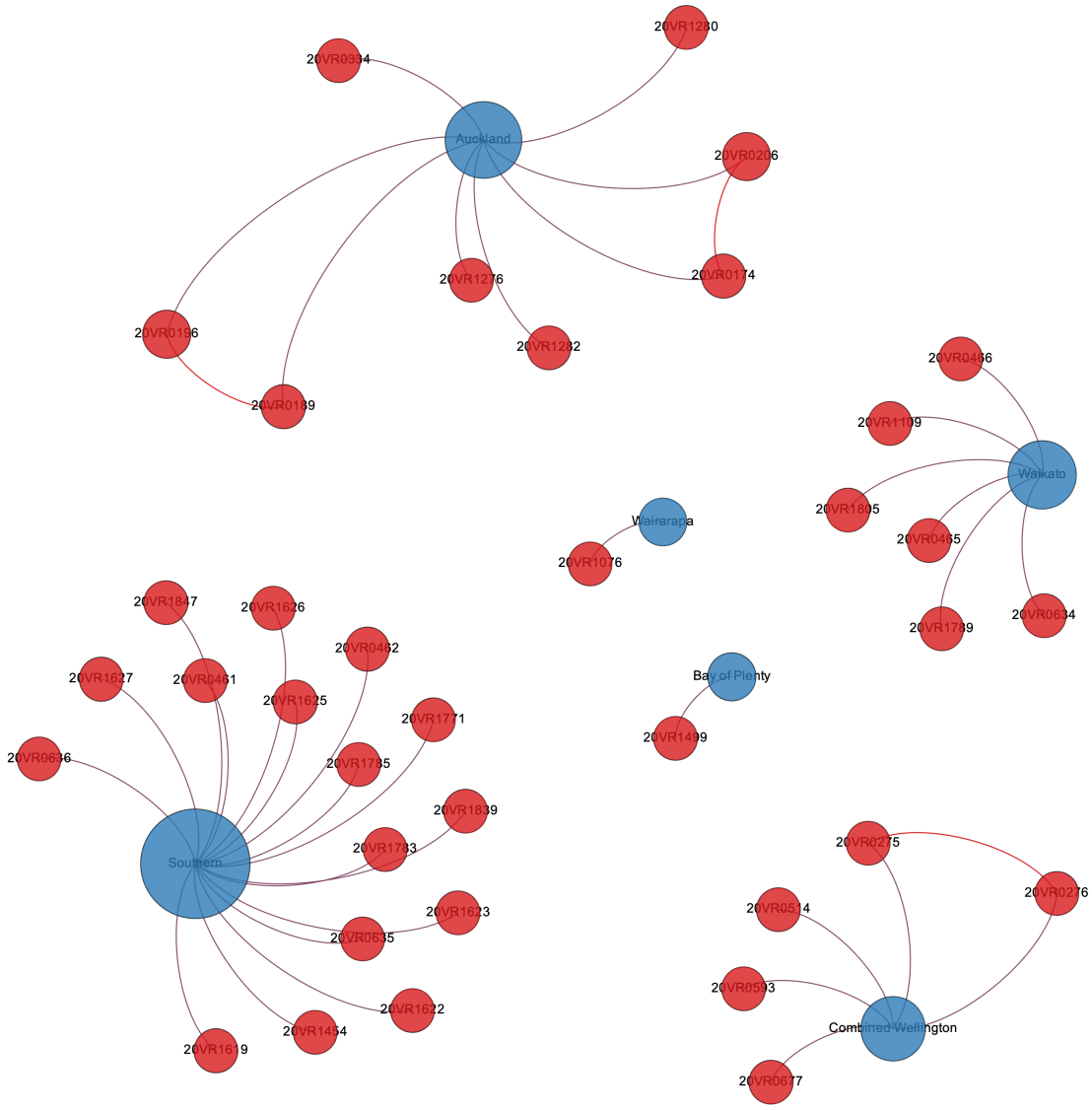

Figure 3: The 22 cases linked to DHBs where they were registered. Red edges between cases indicate common source of overseas travel (same country travelling from).

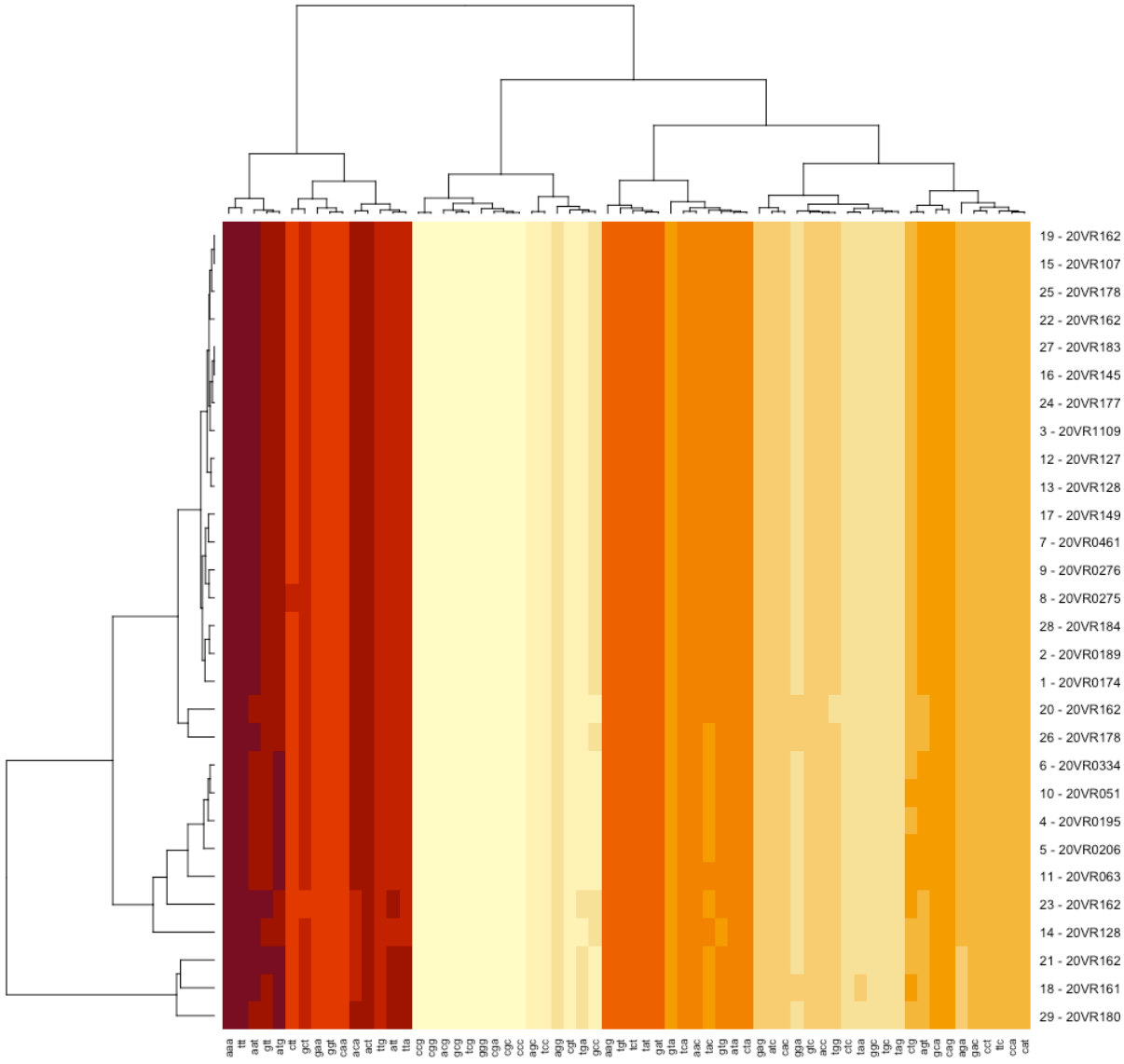

Figure 4: Codon coverage in the +1 reading frame for all 29 genomes.

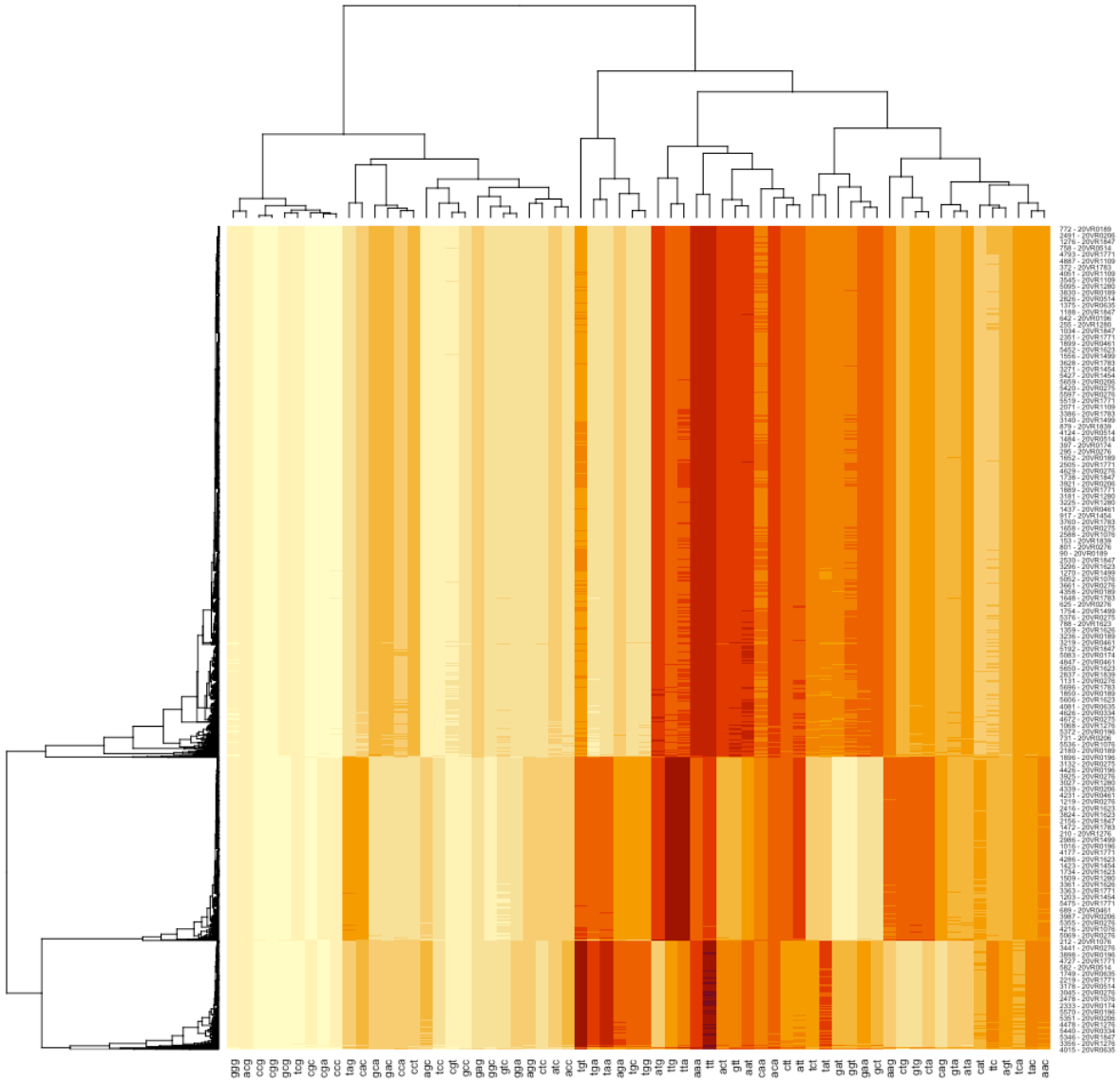

Figure 5: *Codon coverage in the +1 reading frame for approx. 6000 genomes collected globally (data source: GISAID EpiCov database).*

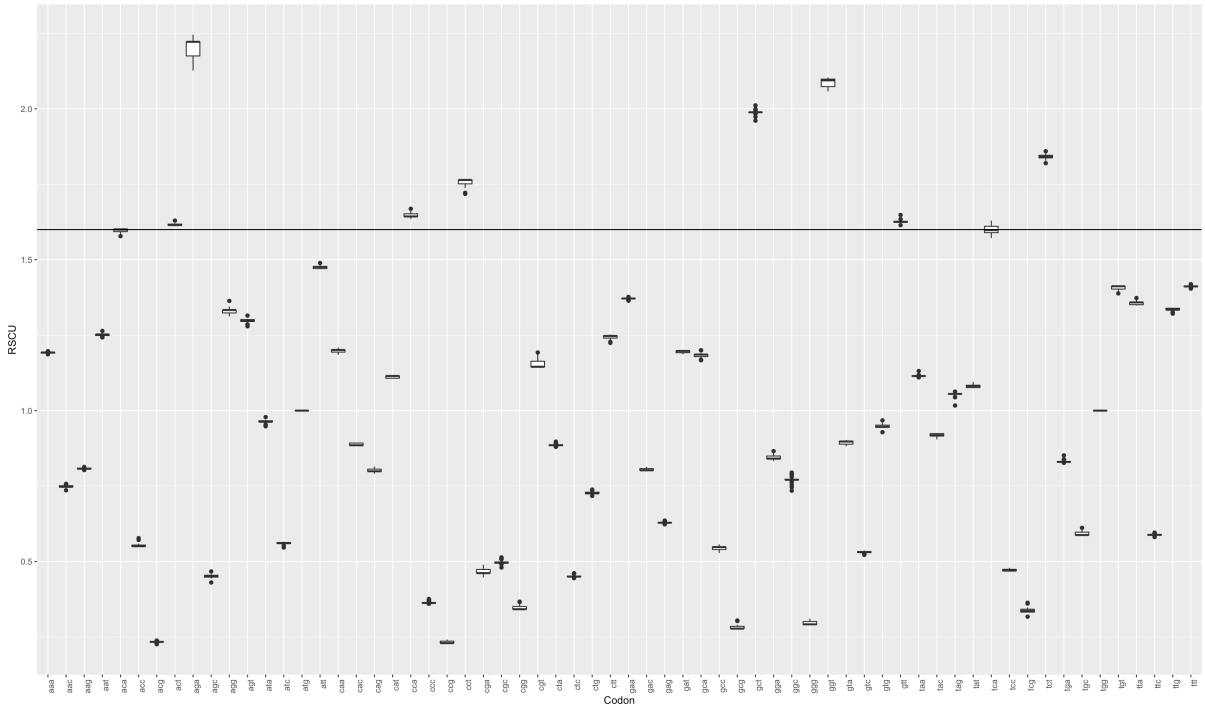

Figure 6: RSCU index for all unique codons computed from the original genomic sequences.

Limitations

This situation report is based on the limited amount of genomic data that is available as of today and a preliminary analysis of the codon recurrence tracing approach using Transcendental Information Cascades only. The network and the inferred phylogenetic tree have not been validated against state-of-the art methods such as phylogenetic trees constructed using a maximum likelihood or Bayesian approach.

Acknowledgements

We want to thank Matthew Tansley for his work on visualising the case data on a geographical map. This work has also been greatly supported by Michael Thingnes' work on the pyhton scripts to construct Transcendental Information Cascades from DNA sequences.

References

[1.] Luczak-Roesch, M. (2020). Networks of information token recurrences derived from genomic sequences may reveal hidden patterns in epidemic outbreaks: A case study of the 2019-nCoV coronavirus. medRxiv.

[2.] Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Yuan, M. L. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265-269.

[3.] Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics, 12(1), 32-42.

[4.] Dilucca, M., & Pavlopoulou, A. (2020). Analysis of codon usage and evolutionary rates of the 2019-nCoV genes. bioRxiv.