The power and limitations of genomics to track COVID-19 outbreaks: a case study from New Zealand

Jemma L Geoghegan PhD*[1,2], Jordan Douglas PhD*[3], Xiaoyun Ren PhD[2], Matthew Storey BSc[2], James Hadfield PhD[4], Olin K Silander PhD[5], Nikki E Freed PhD[5], Lauren Jelley MAppSc[2], Sarah Jefferies MD[2], Jillian Sherwood MBChB[2], Shevaun Paine MAE[2], Sue Huang PhD[2], Andrew Sporle MA[6,7], Michael G Baker MBChB[8], David R Murdoch MD[9], Alexei J Drummond PhD[3], David Welch PhD[3], Colin R Simpson PhD[10,11], Nigel French PhD[12], Edward C Holmes PhD[13], Joep de Ligt PhD[2].

[1]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand.

[2]Institute of Environmental Science and Research, Wellington, New Zealand.

[3]Centre for Computational Evolution, School of Computer Science, University of Auckland, Auckland, New Zealand.

[4]Fred Hutchinson Cancer Research Centre, Seattle, Washington, USA.

[5]School of Natural and Computational Sciences, Massey University, Auckland, New Zealand.

[6]Department of Statistics, University of Auckland, New Zealand.

[7]iNZight Analytics Ltd., Auckland, New Zealand.

[8]Department of Public Health, University of Otago, Wellington, New Zealand.

[9]Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand.

[10]School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, New Zealand.

[11]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.

[12]School of Veterinary Science, Massey University, Palmerston North, New Zealand.

[13]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia.

*contributed equally

Author for correspondence: jemma.geoghegan@otago.ac.nz

Keywords: SARS-CoV-2; COVID-19; coronavirus; genomics; infectious disease; New Zealand

Summary

Background. Real-time genomic sequencing has played a major role in tracking the global spread and local transmission of SARS-CoV-2, contributing greatly to disease mitigation strategies. After effectively eliminating the virus, New Zealand experienced a second outbreak of SARS-CoV-2 in August 2020. During this August outbreak, New Zealand utilised genomic sequencing in a primary role to support its track and trace efforts for the first time, leading to a second successful elimination of the virus.

Methods. We generated the genomes of 80% of the laboratory-confirmed samples of SARS-CoV-2 from New Zealand's August 2020 outbreak and compared these genomes to the available global genomic data.

Findings. Genomic sequencing was able to rapidly identify that the new COVID-19 cases in New Zealand belonged to a single cluster and hence resulted from a single introduction. However, successful identification of the origin of this outbreak was impeded by substantial biases and gaps in global sequencing data.

Interpretation. Access to a broader and more heterogenous sample of global genomic data would strengthen efforts to locate the source of any new outbreaks.

Funding. This work was funded by the Ministry of Health of New Zealand, New Zealand Ministry of Business, Innovation and Employment COVID-19 Innovation Acceleration Fund (CIAF-0470), ESR Strategic Innovation Fund and the New Zealand Health Research Council (20/1018 and 20/1041).

Main Text

Only twelve days after the novel coronavirus SARS-CoV-2 was identified, a genome of the virus was first published[1]. This information was pivotal to the subsequent, rapid, development of diagnostic tests and identification of potential treatments[2,3]. Since then, as of 25 September 2020, over 110,000 genomes of SARS-CoV-2 have been shared publicly[4]. The underlying genome sequencing has occurred so quickly that, for the first time during an infectious disease outbreak, it has enabled virological and epidemiological data to be integrated in real time[5]. Analysis of these data has played an important role in informing the COVID-19 response by tracking the global spread and evolution of SARS-CoV-2, including identification of the number, source, and timing of introductions into individual countries. This has led to a greater understanding of COVID-19 outbreaks around the world[6-13].

Of the 185 countries that have reported positive cases of COVID-19 to the World Health Organization[14], 60% (n=112) have sequenced and shared SARS-CoV-2 genomes on the GISAID database[4] (September 2020). This immense global sequencing effort has facilitated ongoing

genomic surveillance of the pandemic, including monitoring viral genetic changes of interest[12], which has informed public health responses[15-18]. Nevertheless, the number and proportion of positive cases sequenced, and genomes published, varies dramatically between countries and over time (Figure 1). The COVID-19 Genomics UK (COG-UK) consortium, for example, has led to the UK being the most represented sampling location, totalling over 42,000 genomes and comprising 39% of the global data set despite recording only 1% of the world's positive cases (n = 412,241). Conversely, SARS-CoV-2 genomes sequenced in India represent just 3% of the global data set but 18% of the world's total reported cases (n = 5,646,010).

We show here that such disparities in sequencing effort can have major implications for data interpretation and must be met with careful consideration. Real-time sequencing of SARS-CoV-2 genomes has, however, had particular utility in tracking the re-emergence of the virus in New Zealand. By June 2020 New Zealand had effectively eliminated COVID-19 in the community and positive cases were limited to those linked to managed quarantine facilities at the border[7,13,19]. Following over 100 days with no detected community transmission of COVID-19, four new cases emerged on August 12, 2020 with no apparent epidemiological link to any known case. We compared the genomes of these cases to sequenced cases from both New Zealand's first wave and those in quarantine facilities and, again, found no link. The vast majority of available sequence data from cases in New Zealand's quarantine facilities were of different virus lineages than that of the August 2020 outbreak. However, this observation was of limited value given that only 42% of cases in those quarantine facilities had adequate viral RNA for successful genomic sequencing. In order to determine the likely origins of this outbreak we compared genomes from the new community outbreak to the global data set.

An initial genomic sequence analysis found that the re-emergence of COVID-19 in New Zealand was due to a SARS-CoV-2 from lineage B.1.1.1[20]. Of the countries that have contributed SARS-CoV-2 data, 40% had genomes of this lineage. Remarkably, 85% of B.1.1.1. genomes were from the UK and were generated between March – September 2020, the vast majority of which were sampled during the first wave of disease in the UK (Figure 2). Phylogenetic analysis of the most recently sampled B.1.1.1. genomes (1,996 of 4,544) identified genomes sampled from Switzerland, South Africa and England in August as the most likely to be contained within the sister clade (Figure 2): these were the closest sampled genomic relatives of the viruses associated with New Zealand's August 2020 outbreak (Supplementary Figure 1). Additional Bayesian analysis estimated that the outbreak originated 11 days before the first transmission event, with a 95% highest posterior density (HPD) of 0 – 28 days. We also estimated that the first transmission event in the outbreak occurred between 26 July – 13 August 2020 (95% HPD; mean date of 5 August). Epidemiological data showed that two confirmed cases linked to the outbreak had a symptom onset date of 31 July, although the most probable sampled genomes within the sister clade were

sampled later, between 6 August – 28 August. Hence, is unlikely that the currently available global genomic data set contains the source of this outbreak.

Genomic epidemiological analysis on the possible origins of New Zealand's re-emergence was found to be inconclusive, likely due to missing genomic data within the quarantine border facilities as well as in the global data set. A glimpse into the genomic diversity likely omitted from the global data set can be seen in the genomes sequenced in New Zealand from positive quarantine cases which comprise citizens and residents returning from across the globe . For example, 12 SARS-CoV-2 genomes from returnees to New Zealand from India who arrived on the same flight fell across at least four genomic lineages and comprised sequence divergence of up to 34 single nucleotide polymorphisms. This represented far more genomic mutations than was observed in New Zealand during the first outbreak in March – May 2020. Such a high level of diversity in just a small sample of positive cases from India suggest that the currently available genomic data fails to encompass the true diversity that existed locally let alone globally.

The re-emergence of COVID-19 into New Zealand in August 2020 exemplified one of the most complete genomic data sets for a specific outbreak compiled to date, comprising 81% of positive cases (145 of 179 total PCR positive cases). Real-time genomic sequencing quickly informed track and trace efforts to control the outbreak, setting New Zealand on track to eliminate the virus from the community for the second time. The rapid genome sequencing of positive samples provided confidence to public health teams regarding links to the outbreak and identified that cases and sub-clusters were linked to a single genomic lineage, resulting from a single introduction event. Indeed, the timing and length of lockdown measures were partly informed based on these data.

Nevertheless, the biased nature of global sampling, including the contribution of very few genome sequences from certain geographic locations, clearly limited the power of genomics to attribute the geographical origin of New Zealand's August 2020 outbreak. We therefore advocate that careful consideration of the potential sampling biases and gaps in available genomic data be made whenever attempting to determine the geographic origins of a specific outbreak of SARS-CoV-2. Analysis should consider all available evidence, including from genomic and epidemiological sources.

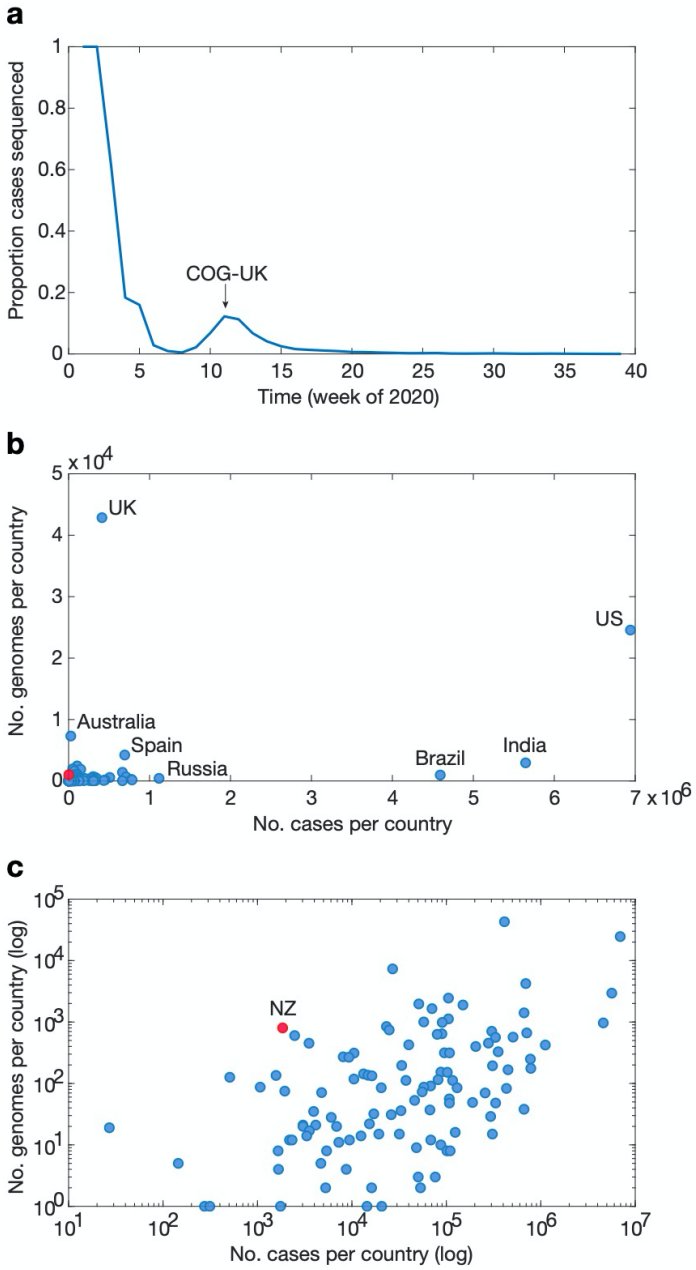

Figure 1. **a** Proportion of global cases sequenced and shared on GISAID between December 2019 to September 2020, where the second mode was largely driven by the COG-UK consortium as illustrated; **b** Number of genomes sequenced and number of reported cases per country (New Zealand is represented in red) on a linear scale; **c** Number of genomes sequenced and number of reported cases per country on a logarithmic scale.

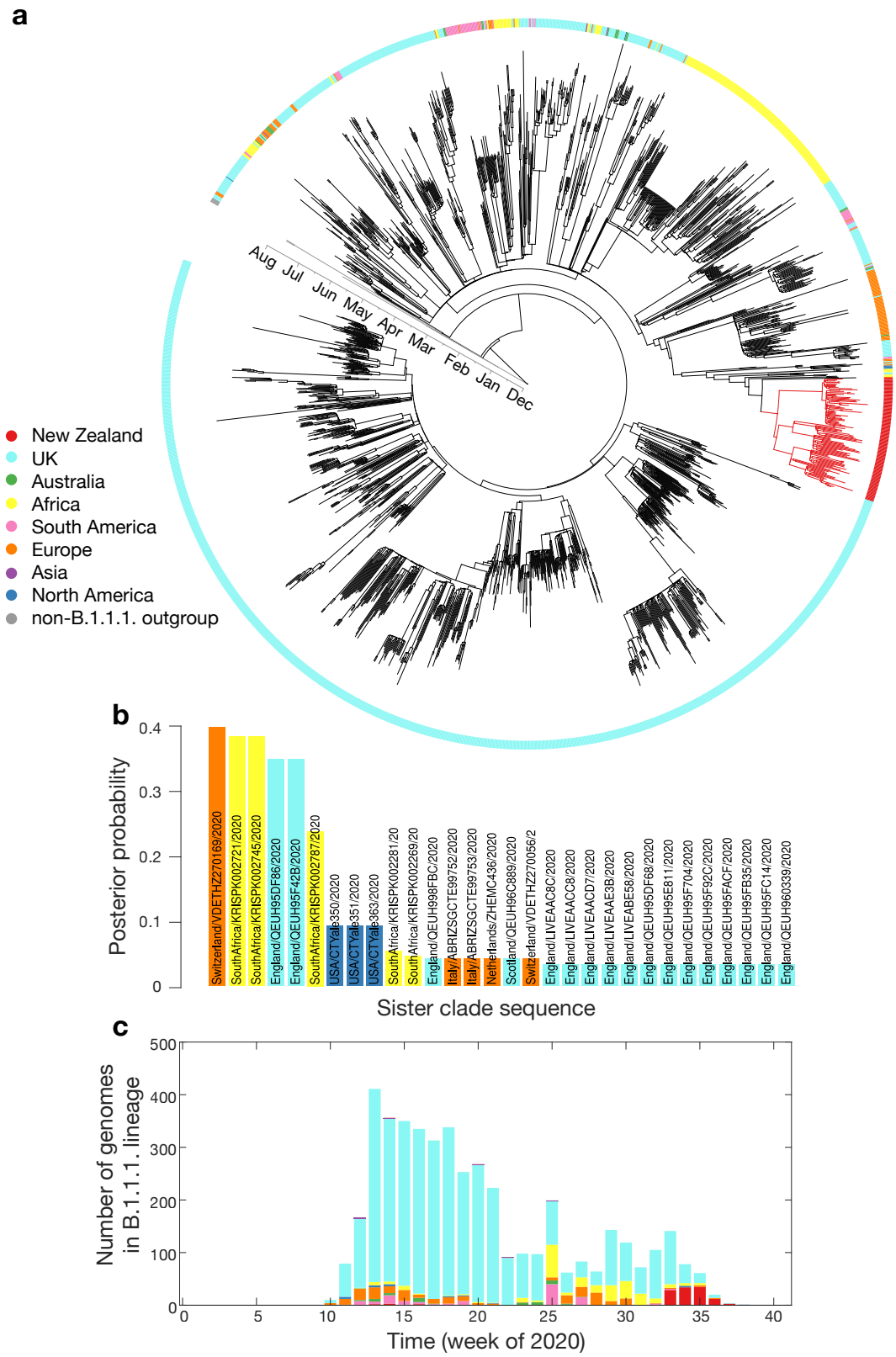

Figure 2. **a** Maximum clade credibility phylogenetic tree of 2,000 subsampled global genomes (1,996 most recently sampled B.1.1.1. plus four non-B.1.1.1. used as an outgroup) with an outer ring coloured by sampling region; **b** Posterior probability of genomes within the sister clade to New Zealand's August outbreak, colour-coded by sampling location; **c** Proportion of genomes within lineage B.1.1.1. in the global data set over time, colour-coded by sampling location.

Methods

Ethics statement. Nasopharyngeal samples testing positive for SARS-CoV-2 by real-time polymerase chain reaction (RT-PCR) were obtained from public health medical diagnostics laboratories located throughout New Zealand. All samples were de-identified before receipt by the researchers. Under contract for the New Zealand Ministry of Health, ESR has approval to conduct genomic sequencing for surveillance of notifiable diseases.

Genomic sequencing of SARS-CoV-2: New Zealand's second wave. A total of 172 (out of 179) laboratory-confirmed samples of SARS-CoV-2 were received by ESR for whole genome sequencing, comprising samples from New Zealand's August 2020 outbreak. Genome sequencing of SARS-CoV-2 samples was performed as before[7]. In short, viral extracts were prepared from respiratory tract samples where SARS-CoV-2 was detected by RT-PCR using WHO recommended primers and probes targeting the E and N gene. Extracted RNA from SARS-CoV-2 positive samples were subject to whole genome sequencing following the ARTIC network protocol (V3) (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye) and the Massey University 1200 bp primer set (https://www.protocols.io/view/ncov-2019-sequencing-protocol-rapid-barcoding-1200-bh7hj9j6)[21].

Briefly, one of the tiling amplicon designs were used to amplify viral cDNA prepared with SuperScript IV. Sequence libraries were then constructed, using the Oxford Nanopore ligation sequencing and native barcoding expansion kit for samples amplified with the ARTIC V3 primer sets, and the Oxford Nanopore rapid barcoding kit for samples amplified with the 1200 bp primer sets. The 1200 bp primers and rapid barcoding was used in cases where the genomes were required urgently. Libraries were sequenced using R9.4.1 MinION flow cells, respectively. Near-complete (>90% recovered) viral genomes were subsequently assembled through reference mapping. Steps included in the pipeline are described in detail online (https://github.com/ESR-NZ/NZ_SARS-CoV-2_genomics). The reads generated with Nanopore sequencing using ARTIC primer sets (V3) were mapped and assembled using the ARTIC bioinformatics medaka pipeline (v 1.1.0). In total, 145 (out of 172) genomes from New Zealand's August 2020 outbreak passed quality control. All data are available on GISAID.

Phylogenetic analysis of SARS-CoV-2. All human SARS-CoV-2 genomes available on GISAID[4] as of 25 September 2020 (n = 109,379) were downloaded and their lineages assigned according to the proposed nomenclature[16] using pangolin (https://github.com/hCoV-2019/pangolin). Genomes assigned to the B.1.1.1. lineage (n = 4,544) were subsampled to include 1,996 most recent-in-time sequences along with four outgroup (non-B.1.1.1.) sequences, and were aligned with those from the recent New Zealand outbreak (n = 140) using MAFFT (v 7)[22] employing the FFT-NS-2 progressive alignment algorithm. Bayesian phylogenetic analyses were performed

using BEAST 2.5[23]. We used a strict clock model with one HKY substitution model (estimated frequencies) for each codon position and one for non-coding positions. We employed the Bayesian skyline model[24] as a tree prior to allow effective population sizes to change over time intervals. These components of the model and their prior distributions are those used by Douglas et al. 2020[13]. Phylogenetic trees were annotated using FigTree (v 1.4)[25] and Tree of Life (v 4)[26].

Acknowledgements

We thank the ARTIC network for making their protocols and tools openly available and specifically Josh Quick for sending the initial V1 and V3 amplification primers. We thank the diagnostic laboratories that performed the initial RT-PCRs and referred samples for sequencing as well as the public health units for providing epidemiological data. We thank Genomics Aotearoa for their support, the NextStrain team for their support and timely global and local analysis, and all those who have contributed SARS-CoV-2 sequences to GenBank and GISAID databases. The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment's Research Infrastructure programme (https://www.nesi.org.nz).

References

1. Holmes EC. Novel 2019 coronavirus genome. 2020. https://virological.org/t/novel-2019-coronavirus-genome/319.
2. Shin MD, Shukla S, Chung YH, et al. COVID-19 vaccine development and a potential nanomaterial path forward. *Nature Nanotechnology* 2020; **15**(8): 646-55.
3. Stebbing J, Phelan A, Griffin I, et al. COVID-19: combining antiviral and anti-inflammatory treatments. *Lancet Infect Dis* 2020; **20**(4): 400-2.
4. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 2017; **1**(1): 33-46.
5. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018; **34**(23): 4121-3.
6. Bedford T, Greninger AL, Roychoudhury P, et al. Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv* 2020: 2020.04.02.20051417.
7. Geoghegan JL, Ren X, Storey M, et al. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *medRxiv* 2020: 2020.08.05.20168930.
8. Candido DS, Claro IM, de Jesus JG, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* 2020: eabd2161.
9. Eden J-S, Rockett R, Carter I, et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evolution* 2020; **6**(1).
10. Filipe ADS, Shepherd J, Williams T, et al. Genomic epidemiology of SARS-CoV-2 spread in Scotland highlights the role of European travel in COVID-19 emergence. *medRxiv* 2020: 2020.06.08.20124834.
11. Seemann T, Lane C, Sherry N, et al. Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv* 2020: 2020.05.12.20099929.
12. Zhang L, Jackson CB, Mou H, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020: 2020.06.12.148726.
13. Douglas J, Mendes FK, Bouckaert R, et al. Phylodynamics reveals the role of human travel and contact tracing in controlling COVID-19 in four island nations. *medRxiv* 2020: 2020.08.04.20168518.
14. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 2020; **20**(5): 533-4.
15. Oude Munnink BB, Nieuwenhuijse DF, Stein M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nature Medicine* 2020; **26**(9): 1405-10.
16. Kalinich CC, Jensen CG, Neugebauer P, et al. Real-time public health communication of local SARS-CoV-2 genomic epidemiology. *PLOS Biology* 2020; **18**(8): e3000869.
17. Rockett RJ, Arnott A, Lam C, et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature Medicine* 2020; **26**(9): 1398-404.
18. Bauer DC, Tay AP, Wilson LOW, et al. Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. *Transboundary and Emerging Diseases* 2020; **67**(4): 1453-62.
19. Jefferies S, French N, Gilkison C, et al. COVID-19 in New Zealand and the impact of the national response: a descriptive epidemiological study. *The Lancet Public Health*.
20. Rambaut A, Holmes EC, Hill V, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020: 2020.04.17.046086.
21. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biology Methods and Protocols* 2020; **5**(1).
22. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; **30**(4): 772-80.

23. Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* 2019; **15**(4): e1006650.
24. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005; **22**(5): 1185-92.
25. Rambaut A, Drummond AJ. FigTree version 1.4.0. 2012.
26. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* 2019; **47**(W1): W256-W9.