Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand

Jemma L Geoghegan[1,2], Xiaoyun Ren[2], Matthew Storey[2], James Hadfield[3], Lauren Jelley[2], Sarah Jefferies[2], Jill Sherwood[2], Shevaun Paine[2], Sue Huang[2], Jordan Douglas[4], Fábio K Mendes[4], Andrew Sporle[5,6], Michael G Baker[7], David R Murdoch[8], Nigel French[9], Colin R Simpson[10,11], David Welch[4], Alexei J Drummond[4], Edward C Holmes[12], Sebastián Duchêne[13], Joep de Ligt[2]

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

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

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

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

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

[6]McDonaldSporle Ltd., Auckland, New Zealand.

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

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

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

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

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

[12]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.

[13]Department of Microbiology and Immunology, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia.

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

Keywords: SARS-CoV-2; COVID-19; coronavirus; genomics; phylodynamics; phylogenetics; virus evolution; infectious disease; New Zealand

Abstract

New Zealand, a geographically remote Pacific island with easily sealable borders, implemented a nation-wide lockdown of all non-essential services to curb the spread of COVID-19. New Zealand has now effectively eliminated the virus, with low numbers of new cases limited to new arrivals in managed quarantine facilities at the border. Here, we generated 649 SARS-CoV-2 genome sequences from infected patients in New Zealand with samples collected between 26 February and 22 May 2020, representing 56% of all confirmed cases in this time period. Despite its remoteness, the viruses imported into New Zealand represented nearly all of the genomic diversity sequenced from the global virus population. The proportion of D614G variants in the virus spike protein increased over time due to an increase in their importation frequency, rather than selection within New Zealand. These data also helped to quantify the effectiveness of public health interventions. For example, the effective reproductive number, R_e, of New Zealand's largest cluster decreased from 7 to 0.2 within the first week of lockdown. Similarly, only 19% of virus introductions into New Zealand resulted in a transmission lineage of more than one additional case. Most of the cases that resulted in a transmission lineage originated from North America, rather than from Asia where the virus first emerged or from the nearest geographical neighbour, Australia. Genomic data also helped link more infections to a major transmission cluster than through epidemiological data alone, providing probable sources of infections for cases in which the source was unclear. Overall, these results demonstrate the utility of genomic pathogen surveillance to inform public health and disease mitigation.

Main Text

New Zealand is one of a handful of countries that aimed to eliminate coronavirus disease 19 (COVID-19). The disease was declared a global pandemic by the World Health Organisation (WHO) on 11 March 2020. The causative virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1], was first identified and reported in China in late December 2019, and is the seventh coronavirus known to infect humans, likely arising through zoonotic transmission from wildlife[2]. Because of its relatively high case fatality rate[3-5], and virus transmission from asymptomatic or pre-symptomatic individuals[6,7], SARS-CoV-2 presents a significant public health challenge. Due to its high rate of transmission, morbidity and mortality, SARS-CoV-2 has resulted in world-wide lockdowns, economic collapses and led to healthcare systems being overrun.

Since the publication of the first SARS-CoV-2 genome on 10 January 2020[8], there has been a substantial global effort to contribute and share genomic data to inform local and international communities about key aspects of the pandemic[9]. Analyses of genomic data have played an important role in tracking the epidemiology and evolution of the virus, often doing so in real time[10],

and leading to a greater understanding of COVID-19 outbreaks globally[11-15].

New Zealand reported its first case on 26 February 2020 and within a month implemented a stringent, country-wide lockdown of all non-essential services. To investigate the origins, time-scale and duration of virus introductions into New Zealand, the extent and pattern of viral spread across the country, and to quantify the effectiveness of intervention measures, we generated whole genome sequences from 56% of all documented SARS-CoV-2 cases from New Zealand and combined these with detailed epidemiological data.

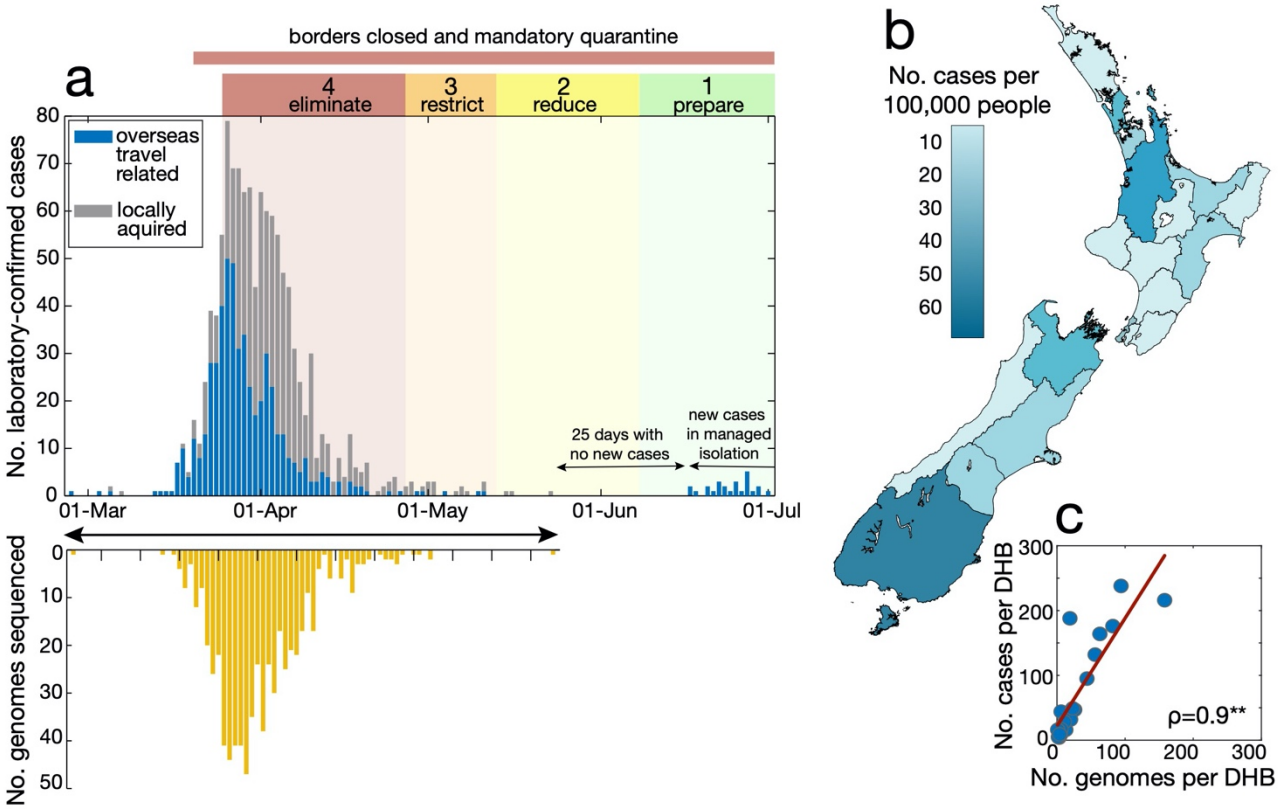

Figure 1. (**a**) Number of laboratory-confirmed cases by reported date, both locally acquired (grey) and linked to overseas travel (blue) in New Zealand, highlighting the timing of public health alert levels 1-4 ('eliminate', 'restrict', 'reduce', 'prepare') and national border closures. The number of genomes sequenced in this study is shown over time. (**b**) Map of New Zealand's District Health Boards shaded by the incidence of laboratory-confirmed cases of COVID-19 per 100,000 people. (**c**) Number of laboratory-confirmed cases per District Health Board (DHB) versus number of genomes sequenced, indicating Spearman's ρ, where asterisks indicate statistical significance (p<0.0001).

82 **Table 1**. Demographic data for confirmed (n=1178) and probable (n=350) cases of SARS-CoV-2
83 in New Zealand between 26 February and 1 July 2020. The percentage of genomes sequenced in
84 each category is shown.

85

Age group	Number of cases	Deceased	Percentage of genomes in data set
0 to 9	37	0	6%
10 to 19	122	0	38%
20 to 29	365	0	45%
30 to 39	238	0	39%
40 to 49	221	0	42%
50 to 59	248	0	44%
60 to 69	180	3	45%
70 to 79	78	7	45%
80 to 89	30	7	50%
90+	9	5	56%
Gender	**Number of cases**	**Percentage of cases**	**Percentage of genomes in data set**
Female	848	55%	42%
Male	680	45%	41%
Ethnicity	**Number of cases**	**Percentage of cases**	**Percentage of genomes in data set**
European or other	1067	70%	46%
Asian	210	14%	27%
Māori	130	9%	42%
Pacific peoples	81	5%	35%
Middle Eastern / Latin American / African	33	2%	42%
Unknown	7	0.50%	86%
Transmission type	**Number of cases**	**Percentage of cases**	**Percentage of genomes in data set**
Imported cases	572	37%	48%
Locally-acquired cases	956	63%	39%

86
87
88
89

Between 26 of February and 1 July 2020 there were a total of 1,178 laboratory-confirmed cases and a further 350 probable cases of SARS-CoV-2 in New Zealand (a probable case is defined as a person who has returned a negative laboratory result or could not be tested, but the medical officer of health has assigned the case classification based on exposure history and clinical symptoms). Of these combined laboratory-confirmed and probable cases, 55% were female and 45% were male, with the highest proportion of cases in the 20-29 age group (Table 1). Many cases were linked to overseas travel (37%). Geographic locations in New Zealand with the highest number of reported cases did not necessarily reflect the human population size or density in that region, with the highest incidence reported in the Southern District Health Board (DHB) region rather than in highly populated cities (Figure 1). The number of laboratory-confirmed cases peaked on 26 March 2020, the day after New Zealand instigated an Alert Level 4 lockdown – the most stringent level, ceasing all non-essential services and stipulating that the entire population self-isolate (Figure 1). From 23 May 2020, New Zealand experienced 25 consecutive days with no new reported cases until 16 June, when new infections, linked to overseas travel, were diagnosed. All subsequent new cases have been from patients in managed quarantine facilities.

We sequenced a total of 649 virus genomes from samples taken between 26 February (first reported case) and 22 May 2020 (to date, the last confirmed case unassociated with managed quarantine facilities). This represented 56% of all New Zealand's confirmed cases sampled during that period. The data generated originated from the 20 DHBs from across New Zealand. DHBs submitted between 0.1% and 81% of their positive samples to the Institute of Environmental Science and Research (ESR), Wellington, for sequencing. Despite this disparity, a strong nationwide spatial representation was achieved (Figure 1).

Notably, the genomic diversity of SARS-CoV-2 sequences sampled in New Zealand represented nearly all of the genomic diversity present in the global viral population, with nine second-level A and B lineages from a recently proposed global SARS-CoV-2 genomic nomenclature[16] identified. This high degree of genomic diversity was observed throughout the country (Figure 2). The SARS-CoV-2 genomes sampled in New Zealand comprised 24% aspartic acid (S^{D614}) and 73% glycine (S^{G614}) at residue 614 in the spike protein (Figure 2). Preliminary studies suggest that the D614G mutation can enhance viral infectivity in cell culture[17]. Nevertheless, it is noteworthy that the increase in glycine in New Zealand samples is due to multiple importation events of this variant rather than selection for this mutation within New Zealand. We also inferred a weak yet significant temporal signal in the data, reflecting the low mutation rate of SARS-CoV-2, which is consistent with findings reported elsewhere (Figure 2).

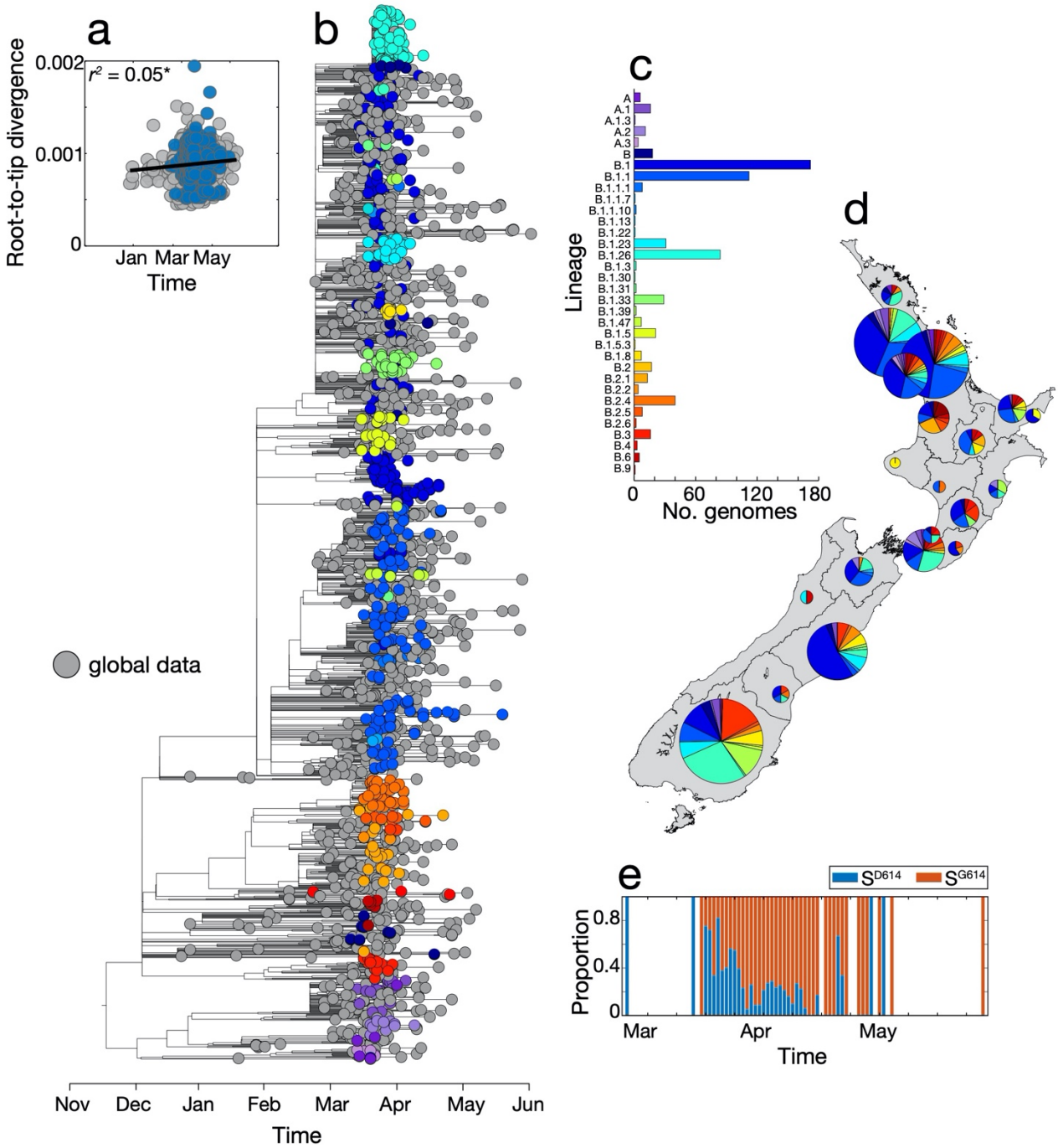

Figure 2. (**a**) Root-to-tip regression analysis of New Zealand (blue) and global (grey) SARS-CoV-2 sequences, with the determination coefficient, r^2 (an asterisk indicates statistical significance; $p<0.05$). (**b**) Maximum-likelihood time-scaled phylogenetic analysis of 649 viruses sampled from New Zealand (coloured circles) on a background of 1000 randomly subsampled viruses from the globally available data (grey circles). Viruses sampled from New Zealand are colour-coded according to their genomic lineage[16]. (**c**) The number of SARS-CoV-2 genomes sampled in New Zealand within each lineage[16]. (**d**) The sampling location and proportion of SARS-CoV-2 genomes sampled from each viral genomic lineage is shown on the map of New Zealand. (**e**) The frequency of D (blue) and G (red) amino acids at residue 614 on the spike protein over time.

Despite the small size of the New Zealand outbreak, there were 277 separate introductions of the virus out of the 649 cases considered. Of these, we estimated that 24% (95% CI: 23-30) led to only one other secondary case (i.e. singleton) while just 19% (95% CI: 15-20) of these introduced cases led to ongoing transmission, forming a transmission lineage (i.e. onward transmission to more than one individual; Figure 3). The remainder (57%) did not lead to a transmission event. New Zealand transmission lineages most often originated in North America, rather than in Asia where the virus first emerged, likely reflecting the high prevalence of the virus in North America during the sampling period. By examining the time of the most recent common ancestor, or TMRCA, of the samples, we found no evidence that the virus was circulating in New Zealand before the first reported case on 26 February. Finally, we found that detection was more efficient (i.e. fewer cases were missed) later in the epidemic in that the detection lag (the duration of time from the first inferred transmission event to the first detected case) declined with the age of transmission lineages (as measured by the time between the present and the TMRCA; Figure 3).

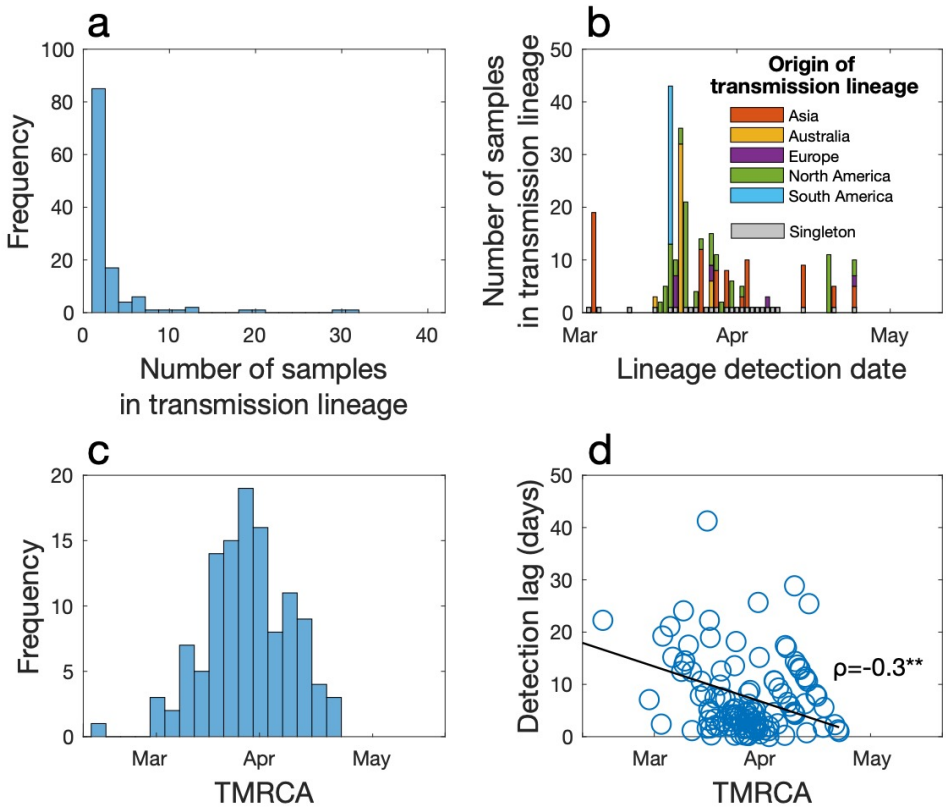

Figure 3. (**a**) Frequency of transmission lineage size. (**b**) The number of samples in each transmission lineage as a function of the date at which the transmission lineage was sampled, coloured by the likely origin of each lineage (inferred from epidemiological data). Importation events that led to only one additional case (singletons) are shown in grey over time. (**c**) Frequency of TMRCA (the time of the most recent common ancestor) of importation events over time. (**d**) The difference between the TMRCA and the date as which a transmission lineage was detected (i.e. detection lag) as a function of TMRCA. Spearman's ρ indicates a significant negative relationship ($p<0.01$).

The largest clusters in New Zealand were often associated with social gatherings such as weddings, hospitality and conferences[18]. The largest cluster, which comprised lineage B.1.26, most likely originated in the USA according to epidemiological data, and significant local transmission in New Zealand was probably initiated by a superspreading event at a wedding in Southern DHB (geographically the most southern DHB) prior to lockdown. Examining the rate of transmission of this cluster enables us to quantify the effectiveness of the lockdown. Its effective reproductive number, R_e, decreased over time from 7 at the beginning of the outbreak (95% credible interval, CI: 3.7-10.7) to 0.2 (95% CI: 0.1-0.4) by the end of March (Figure 4). The sampling proportion of this cluster, a key parameter of the model, had a mean of 0.75 (95% CI: 0.4-1), suggesting sequencing captured the majority of cases in this outbreak. In addition, analysis of genomic data has linked five additional cases to this cluster that were not identified in the initial epidemiological investigation, highlighting the added value of genomic analysis. This cluster, seeded by a single-superspreading event that resulted in New Zealand's largest chain of transmission, illustrates the link between micro-scale transmission to nation-wide spread (Figure 4).

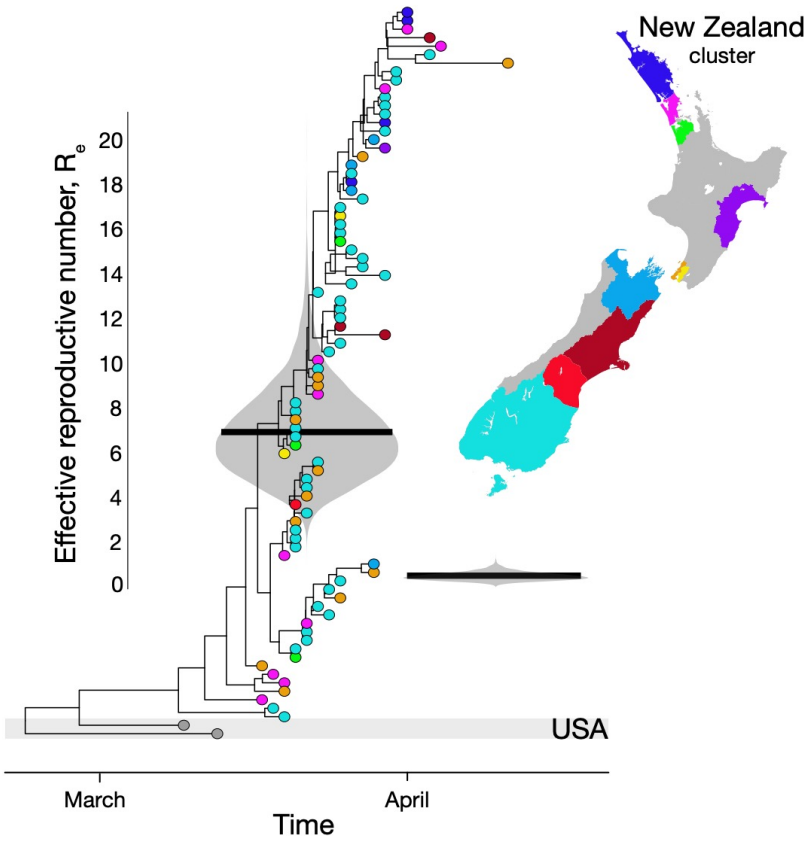

Figure 4. Maximum clade credibility phylogenetic tree of New Zealand's largest cluster with an infection that most likely originated in the USA. Estimates of the effective reproductive number, R_e, are shown in violin plots superimposed onto the tree, grouping the New Zealand samples into two time-intervals as determined by the model. Black horizontal lines indicate the mean R_e. Tips are coloured by the reporting District Health Board and their location shown on the map.

198 The dramatic decrease in R_e of this large cluster coupled with the relatively low number of virus
199 introductions that resulted in a transmission lineage suggests that implementing a strict and early
200 lockdown in New Zealand rapidly reduced multiple chains of virus transmission. As New Zealand
201 continues to remain free of COVID-19 community transmission, but with positive cases still
202 detected amongst individuals quarantined at the border reflecting high virus incidence in other
203 localities, it is imperative that ongoing genomic surveillance is an integral part of the national
204 response to monitor any re-emergence of the virus, particularly when border restrictions might
205 eventually be eased.

206

207 Methods

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

213 **Genomic sequencing of SARS-CoV-2**. A total of 733 laboratory-confirmed samples of SARS-
214 CoV-2 were received by ESR for whole genome sequencing. Viral extracts were prepared from
215 respiratory tract samples where SARS-CoV-2 was detected by RT-PCR using WHO
216 recommended primers and probes targeting the E and N gene. Extracted RNA from SARS-CoV-2
217 positive samples were subject to whole genome sequencing following the ARTIC network
218 protocol (V1 and V3) and the New South Wales (NSW) primer set[15].

219 Briefly, three different tiling amplicon designs were used to amplify viral cDNA prepared with
220 SuperScript IV. Sequence libraries were then constructed using Illumina Nextera XT for the NSW
221 primer set or the Oxford Nanopore ligation sequencing kit for the ARTIC protocol. Libraries were
222 sequenced using Illumina NextSeq chemistry or R9.4.1 MinION flow cells, respectively. Near-
223 complete (>90% recovered) viral genomes were subsequently assembled through reference
224 mapping. Steps included in the pipeline are described in detail online (https://github.com/ESR-
225 NZ/NZ_SARS-CoV-2_genomics).

226 The reads generated with Nanopore sequencing using ARTIC primer sets (V1 and V3) were
227 mapped and assembled using the ARTIC bioinformatics medaka pipeline (v 1.1.0)[19]. For the NSW
228 primer set, raw reads were quality and adapter trimmed using trimmomatic (v 0.36)[20]. Trimmed
229 paired reads were mapped to a reference using the Burrows-Wheeler Alignment tool[21]. Primer
230 sequences were masked using iVar (v 1.2)[22]. Duplicated reads were marked using Picard (v
231 2.10.10)[23] and not used for SNP calling or depth calculation. Single nucleotide polymorphisms

232 (SNPs) were called using bcftools mpileup (v 1.9)[24]. SNPs were quality trimmed using vcflib (v
233 1.0.0)[25] requiring 20x depth and overall quality of 30. Positions that were less than 20x were
234 masked to N in the final consensus genome. Positions with an alternative allele frequency
235 between 20% to 79% were also masked to N. In total, 649 sequences passed our quality control
236 (BioProject: PRJNA648792; a list of genomes and their sequencing methods are provided in
237 Supplementary Table 1).

238 **Phylogenetic analysis of SARS-CoV-2**.

239 SARS-CoV-2 sequences from New Zealand, together with 1,000 genomes uniformly sampled at
240 random from the global population from the ~50,000 available sequences from GISAID[26] (June
241 2020), were aligned using MAFFT(v 7)[27] using the FFT-NS-2 algorithm. A maximum likelihood
242 phylogenetic tree was estimated using IQ-TREE (v 1.6.8)[28], utilising the Hasegawa-Kishino-Yano
243 (HKY+Γ)[29] nucleotide substitution model with a gamma distributed rate variation among sites (the
244 best fit model was determined by ModelFinder[30]), and branch support assessment using the
245 ultrafast bootstrap method[31]. We regressed root-to-tip genetic divergence against sampling dates
246 to investigate the evolutionary tempo of our SARS-CoV-2 samples using TempEst (v 1.5.3)[32].
247 Lineages were assigned according to the proposed nomenclature[16] using pangolin
248 (https://github.com/hCoV-2019/pangolin). To depict virus evolution in time, we used Least
249 Squares Dating[33] to estimate a time-scaled phylogenetic tree using the day of sampling.

250 With the full set of New Zealand sequences, we used a time-aware coalescent Bayesian
251 exponential growth model available in BEAST (v 1.10.4)[34]. The HKY+Γ model of nucleotide
252 substitution was again used along with a strict molecular clock. Because the data did not display
253 a strong temporal signal, we used an informative prior reflecting recent estimates for the
254 substitution rate of SARS-CoV-2[35]. The clock rate had a Γ prior distribution as a prior with a mean
255 of 0.8×10^{-3} subs/site/year and standard deviation of 5×10^{-4} (parameterised using the shape and
256 rate of the Γ distribution). Parameters were estimated using Bayesian Markov Chain Monte Carlo
257 (MCMC) framework, with 2×10^{8} steps-long chains, sampling every 1×10^{5} steps and removing
258 the initial 10% as burn-in. Sufficient sampling was assessed using Tracer (v 1.7.1)[36], by verifying
259 that every parameter had effective sampling sizes above 200. Virus sequences were annotated as
260 'imported' (including country of origin) or 'locally acquired', according to epidemiological data
261 provided by EpiSurv[37]. From a set of 1,000 posterior trees, we estimated a number of statistics
262 using NELSI[38]. We determined the number of introductions of the virus into New Zealand as well
263 as the changing number of local transmission lineages through time, with the latter defined as two
264 or more New Zealand SARS-CoV-2 cases that descend from a shared introduction event of the
265 virus into New Zealand[39]. Importation events that led to only a single case rather than a
266 transmission lineage are referred to as 'singletons'. For each transmission lineage and singleton,
267 we inferred the TMRCA.

To estimate R_e through time we analysed New Zealand sequences from the clade identified to be associated with a wedding. We used a Bayesian birth-death skyline model using BEAST (v 2.5)[40], estimating R_e for two time-intervals, as determined by the model, and with the same parameter settings as above. We assumed an infectious period of 10 days, which is consistent with global epidemiological estimates[41].

Online Supplementary Material

Supplementary Table 1. A list of genomes and which amplification and sequencing method was used in for each case.

Acknowledgements

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). We thank the ATRIC network for making their protocols and tools openly available and specifically Josh Quick for sending the initial V1 and V3 amplification primers. We thank Genomics Aotearoa for their support. 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 the Nextstrain team for their support and timely global and local analysis. We thank all those who have contributed SARS-CoV-2 sequences to GenBank and GISAID databases.

References

1 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

3 Wu, J. T. *et al.* Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506-510, doi:10.1038/s41591-020-0822-7 (2020).

4 Russell, T. W. *et al.* Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill.* **25**, doi:10.2807/1560-7917.Es.2020.25.12.2000256 (2020).

5 Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 669-677, doi:10.1016/s1473-3099(20)30243-7 (2020).

6 Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936, doi:10.1126/science.abb6936 (2020).

7 Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro surveill.* **25**, 2000180, doi:10.2807/1560-7917.ES.2020.25.10.2000180 (2020).

8 Holmes, E. C. Novel 2019 coronavirus genome, https://virological.org/t/novel-2019-coronavirus-genome/319 (2020).

9 Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10-19, doi:10.1038/s41564-018-0296-2 (2019).

10 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinform* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

11 Candido, D. d. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* eabd216, 10.1126/science.abd2161 (2020).

12 Filipe, A. D. S. *et al.* Genomic epidemiology of SARS-CoV-2 spread in Scotland highlights the role of European travel in COVID-19 emergence. *medRxiv*, 2020.2006.2008.20124834, doi:10.1101/2020.06.08.20124834 (2020).

13 Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv*, 2020.2005.2012.20099929, doi:10.1101/2020.05.12.20099929 (2020).

14 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv*, 2020.2004.2002.20051417, doi:10.1101/2020.04.02.20051417 (2020).

15 Eden, J.S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, doi:10.1093/ve/veaa027 (2020).

16 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv*, 2020.2004.2017.046086, doi:10.1101/2020.04.17.046086 (2020).

17 Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*, 2020.2006.2012.148726, doi:10.1101/2020.06.12.148726 (2020).

18 Leclerc, Q. J. *et al.* What settings have been linked to SARS-CoV-2 transmission clusters? *Wellcome Open Res.* **5**, 83, doi:10.12688/wellcomeopenres.15889.2 (2020).

19 Loman, N. R., W; Rambaut, A. nCoV-2019 novel coronavirus bioinformatics protocol, https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html (2020).

20 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinform.* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

21 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinform.* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

22 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8, doi:10.1186/s13059-018-1618-7 (2019).

23 Picard Toolkit. Broad Institute, http://broadinstitute.github.io/picard/ (2019).

24 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinform.* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

25 Garrison, E. Vcflib, a simple C++ library for parsing and manipulating VCF files. https://github.com/vcflib/vcflib (2016).

26 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

27 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

28 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).

29 Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174, doi:10.1007/BF02101694 (1985).

30 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

31 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518-522, doi:10.1093/molbev/msx28 (2017).

32 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007, doi:10.1093/ve/vew007 (2016).

33 To, T. H., Jung, M., Lycett, S. & Gascuel, O. Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82-97, doi:10.1093/sysbio/syv068 (2016).

34 Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214, doi:10.1186/1471-2148-7-214 (2007).

35 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450-452, doi:10.1038/s41591-020-0820-9 (2020).

36 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901-904, doi:10.1093/sysbio/syy032 (2018).

37 EpiSurv: national notifiable disease surveillance database, https://surv.esr.cri.nz/episurv/index.php, (2020).

38 Ho, S. Y., Duchêne, S. & Duchêne, D. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol. Ecol. Resour.* **15**, 688-696, doi:10.1111/1755-0998.12320 (2015).

39 Pybus, O. G. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages, https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507 (2020).

40 Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. USA* **110**, 228-233, doi:10.1073/pnas.1207965110 (2013).

41 He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672-675, doi:10.1038/s41591-020-0869-5 (2020).