THE LANCET
Psychiatry

Supplementary appendix 1

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Lane CR, Sherry NL, Porter AF, et al. Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study. *Lancet Psychiatry* 2021; published online July 9. http://dx.doi.org/10.1016/S2468-2667(21)00133-X.

Supplementary Information

Genomics-informed responses in the elimination of COVID-19 in Australia

SUPPLEMENTARY METHODS

Setting, data sources and COVID-19 genomics response group

In Australia, all cases of COVID-19 are immediately notified to public health authorities in each State or Territory. The Victorian Infectious Diseases Reference Laboratory (VIDRL) is the public health virology reference laboratory for the State of Victoria in Australia, covering a resident population of approximately 6.24 million. All primary samples testing positive for SARS-CoV-2 by RT-PCR at diagnostic laboratories are forwarded to the Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) for WGS and bioinformatic analysis. We conducted a retrospective, observational study of all patients in Victoria with confirmed COVID-19 with diagnosis on or prior to 31st January 2021. Detailed demographic and risk factor information on each case was obtained from the Victorian Department of Health and Human Services (DHHS) and collected through individual case interviews using standardized case report forms. Data obtained included age, date of symptom onset, and risk factors for infection including whether the case was a healthcare worker, date and location(s) of recent travel, and contact with any other suspected or confirmed cases of COVID-19 prior to illness onset.

To rapidly implement SARS-CoV-2 genomic analysis into local public health responses, a COVID-19 genomics response team was convened. This included representatives from the state health department, virology laboratory, the public health genomics laboratory (genomic epidemiologist, bioinformaticians and medical microbiologists) and academics with expertise in statistical genomics. Laboratory and bioinformatic workflows were developed to enable large-scale rapid genomic processing of samples. The response team held online meetings (weekly plus *ad hoc* as required) to enable interactive reporting of genomic epidemiological analyses and facilitate rapid translation of genomic findings into public health responses.

Genomic sequencing and bioinformatic analysis

RNA was extracted from samples testing positive for SARS-CoV-2 on the QIAsymphony using the DSP Virus/Pathogen Mini Kit (Qiagen) or using the QIAamp 96 Virus QIAcube HT Kit (Qiagen). Tiled amplicons were prepared from RNA extracts using either ARTIC version 1 or version 3 primers (https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019) using published protocols (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w). Sequencing libraries were prepared from amplicons using NexteraXT and sequenced on either the NextSeq500/550 or iSeq100 (Illumina) using 150bp paired-end reads as described by the manufacturer.

Consensus sequence generation

Paired Illumina reads were aligned to the 29903nt Wuhan reference (Genbank MN908947.3) using *minimap2* (v2.17, options "-ax sr")[1]. The output of *samtools*[2] *mpileup* (v1.11, options "-aa -d - -A -B -Q - ") was then used by *ivar consensus*[3] (v1.3, options "-m 10 -t -.9 -n N"). We applied quality control checks on generated sequences, requiring ≥95% genome recovery, ≤42 single nucleotide polymorphisms (SNPs) from the Wuhan reference, and ≤30 ambiguous bases (~1% of the genome) for sequences to 'pass' QC.

Interjurisdictional data-sharing and analysis using AusTrakka

Wherever possible, sequences from cases diagnosed in other Australian states and territories have been included in phylogenomic analyses. Sequencing of inter-state samples was conducted in their originating state genomic laboratories and shared via AusTrakka, a platform developed by the Communicable Diseases Genomic Network for data sharing and collaborative analysis of genomic data among public health genomic laboratories in Australia and New Zealand. Sequencing methodology and protocols vary between states, with some described in the literature[4,5]. AusTrakka uses a combination of consensus sequences publicly available in the GISAID repository or provided directly, for sequences yet to be publicly available. Sequences from states other than Victoria available in AusTrakka that were not publicly available on 11th February 2021 have been excluded from analysis.

Phylogenomic analysis

For phylogenetic analysis, a single sequence was selected to represent each patient based on the best sequencing QC parameters. The Wuhan-1 reference genome was included as an outgroup to add directionality to the tree because it represents the oldest sequenced case. A multiple sequence alignment was generated using MAFFT[6] (v7.453, options "--auto"). Untranslated regions (UTRs) were masked (positions 1-265 and 29675-29903), as were problematic sites as designated at https://github.com/W-L/ProblematicSites_SARS-CoV2 (version 2020-07-15, includes 224 sites[7]). A maximum likelihood tree was then generated with the remaining alignment using IQ-Tree[8] (v2.1.-, options "-mset GTR+G4 -bb 1000"). Results were visualized using R (v3.5.3) using the ggtree package (v.1.14.6)[9].

Cluster discovery

Genomic clusters defined as two or more related sequences using a hierarchical clustering algorithm, determined using a complete-linkage clustering of pairwise genetic distances, derived from the maximum likelihood phylogenetic tree above, with thresholds at 0.001, 0.0005, 0.00025 and 0.000125. Genomic

clusters were subsequently curated using available epidemiological data, with genomic transmission networks used to group multiple genomic clusters where epidemiological and genomic data suggested they arose from a single importation and diversified within Victoria.

Identification of mutations of potential biological significance

To identify mutations, each sample's consensus genome was aligned to the reference genomes using *minimap2* (v2.17, options -r941 -x asm5 --end-bonus=500) and mutations identified using *paftools* (options call -L 50) to generate a VCF file[1]. Each position in the consensus genome having missing data was added to the VCF as having no genotype to distinguish between missing and wildtype sites downstream. VCF files were merged using *bcftools* (options merge -0) and protein consequences of these mutations were called jointly using *bcftools csq* (v1.11)[2].

Estimation of growth rate during the exponential phase using phylodynamic analyses and comparison with reported lineages

A range of phylodynamic analyses were conducted to estimate the growth rate of the exponential phase of TN3 and to compare it with lineage B.1.1.7 in the UK, which has been suggested to be spreading at a higher rate than other lineages circulating[10]. For transmission network G all genome sequences collected from June 4th to July 9th were collected to represent the early exponential phase, while minimizing the impact of population structure.

A birth-death model was used with the molecular evolutionary rate fixed to 1×10^{-3} subs/site/year, as has been estimated previously[11], and the GTR+Γ substitution model. The become uninfectious rate had a Γ prior distribution with shape=25.0 and scale=2.0 (the 95% density corresponds to a duration of infection of between around 5 and 8 days), the reproductive number had a lognormal prior distribution with mean and standard deviation of 1.0. The sampling proportion had a β prior distribution with α=1.0 and β=9.0, with most of the weight concentrated around 0.1 (i.e. 10% sequencing proportion due to our subsampling and to match the expectation of that of the UK). Analyses were conducted in BEAST2.5[12], with a Markov chain Monte Carlo (MCMC) of 1×10^{8} steps, sampling every 10,000 steps. Sufficient sampling was assessed by verifying that the effective sample size for all parameters was at least 200 and we repeated the analyses to ensure convergence of the MCMC to the same stationary distribution.

SUPPLEMENTARY RESULTS

Comparison of the Victorian outbreak with B.1.1.7 emergence in the United Kingdom

We compared our estimate of the growth rate of transmission network G under the birth-death model to independent estimates of the this parameter for B.1.1.7[13]. We found that 0.913 of the posterior density for transmission network G was lower than the previously estimated growth rate of 71.3 from Volz et al. , indicating that transmission network G spread nearly at a lower, but comparable rate to B.1.1.7, in its first month of circulation in Victoria. Assuming our prior on the become uninfectious rate, the reproductive number, R_e, would have had a mean of 1.96 (95% CI: 1.54 – 2.46).

SUPPLEMENTARY TABLES

Supplementary Table 1: Sequences generated at MDU PHL and included in analyses. Please see appendix 2.

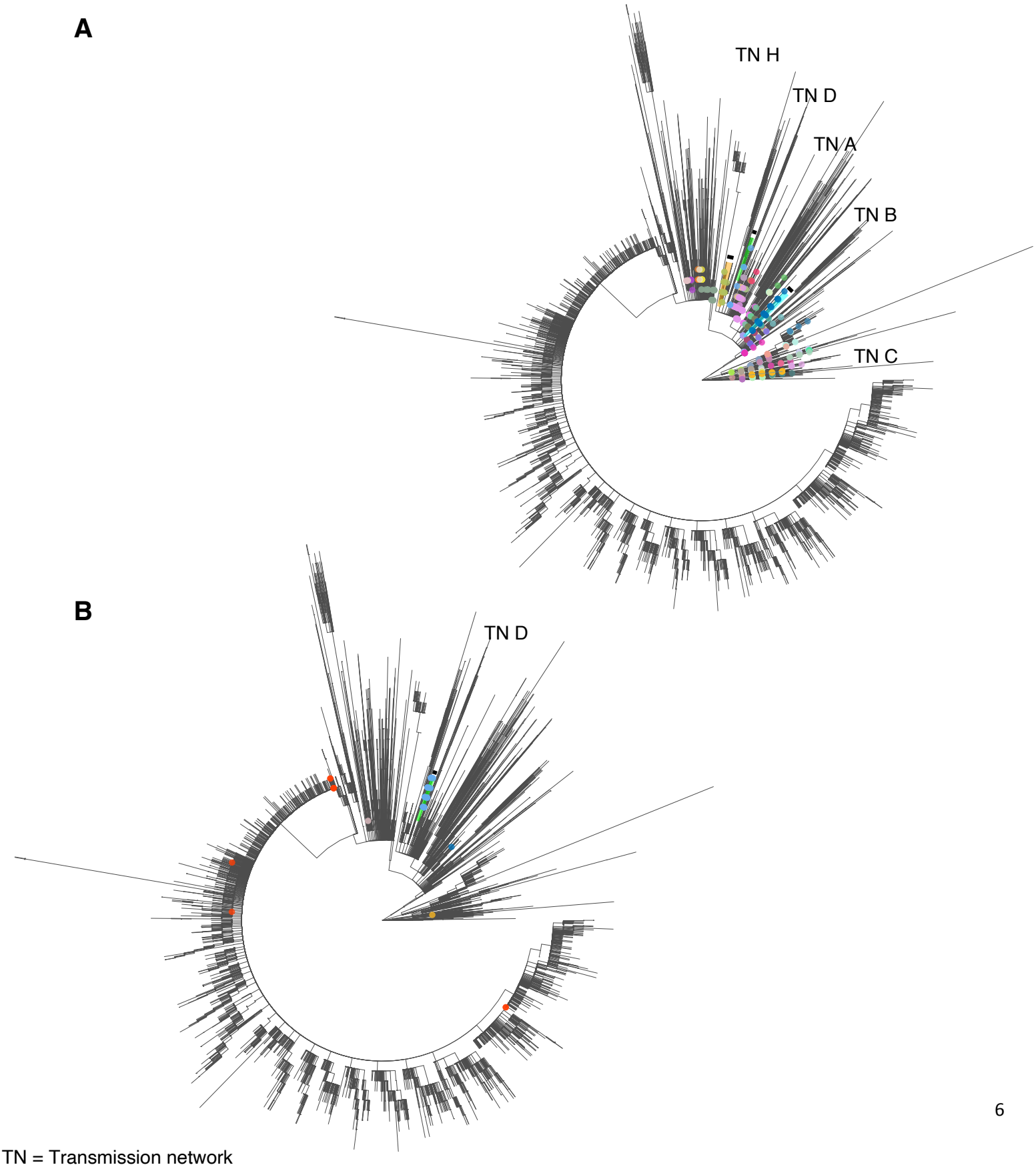
Supplementary Table 2: Publicly available interstate Australian and New Zealand sequences included in anlayses. Please see appendix 3.

Supplementary Table 3: Presence of mutations at location A23063T in the receptor-binding domain of the spike protein hypothesized to contribute to increased transmissibility in included sequences, by genomic cluster or transmission network. Sequences from Victorian cases only. The number of sequences/cases within each genomic cluster with the relevant mutation is presented. Sequences with missing data at the relevant site have been noted as 'missing data'. [a] Other includes genomic clusters with <10 included sequences, clusters where all cases reported overseas travel and unclustered cases.

Genomic cluster / transmission network	Number of cases					
	Wildtype	Variant			Missing data	Total
		N501Y	N501T	N501S		
Transmission network G	9508	-	-	1	917	10426
Transmission network F	136	-	-	-	9	145
Transmission network B	71	-	-	-	6	77
Transmission network D	70	-	-	-	3	73
Transmission network A	65	-	-	-	6	71
1	53	-	-	-	2	55
499	37	-	-	-	1	38
248	28	-	-	-	1	29
30	26	-	-	-	2	28
Transmission network E	-	26	-	-	-	26
57	-	-	23	-	2	25
18	22	-	-	-	2	24
Transmission network C	24	-	-	-	-	24
Transmission network H	17	-	-	-	-	17
107	9	-	-	-	4	13
275	12	-	-	-	1	13
70	12	-	-	-	-	12
489	11	-	-	-	-	11
260	9	-	-	-	2	11
226	8	-	-	-	2	10
Other[a]	510	27	-	-	46	583
Total	**10628**	**53**	**23**	**1**	**1006**	**11711**

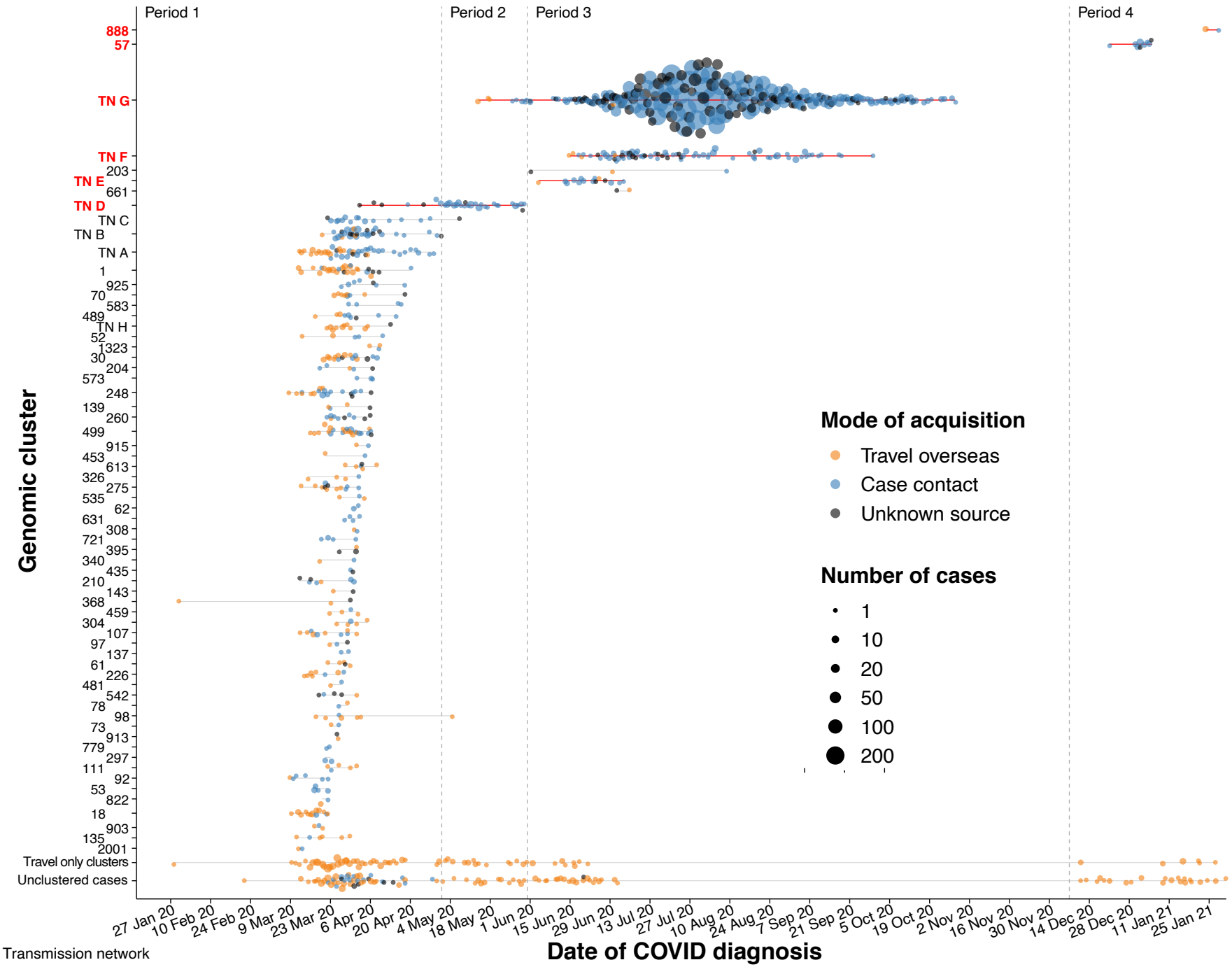
SUPPLEMENTARY FIGURES

Supplementary figure 1: Maximum-likelihood phylogenetic trees of Australian SARS-CoV 2 samples, with cases highlighted by time period, up to 30th January 2021. Includes samples from Victorian cases diagnosed up to and including 31st January 2021, and interstate samples with a collection date prior to 31st January 2021. **A:** Cases diagnosed during period 1 (25th Jan-30th Apr 2020), including those within transmission networks A (shaded area), B (light blue shaded area), C (blue shaded area), D (green shaded area) and H (orange shaded area). **B:** Cases diagnosed during period 2 (1st May-30th May 2020) highlighted, including those within transmission network D (green shaded area).



TN = Transmission network

Supplementary figure 2: Timeline of SARS-CoV-2 genomic clusters, Victoria, Australia, 25th January 2020 – 31st January 2021. Includes Victorian cases diagnosed up to and including 31st January 2021. Cases are plotted by diagnosis date on the x-axis and genomic cluster or transmission network on the y-axis. Mode of acquisition was categorized as (i) travel overseas if the case reported travel in the 14 days prior to symptom onset, (ii) contact with a confirmed case if no overseas travel reported and case contact occurred within the same time period or (iii) unknown source. Bubble size corresponds to the number of cases diagnosed per day, within a genomic cluster or transmission network and with the same mode of acquisition, with larger bubble sizes indicating a greater number of cases.



TN = Transmission network

SUPPLEMENTARY REFERENCES

1. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094-100.
2. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078-9.
3. Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology 2019;20:1-19.
4. 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;26:1398-404.
5. Speake H, Phillips A, Chong T, et al. Flight-associated transmission of severe acute respiratory syndrome coronavirus 2 corroborated by whole-genome sequencing. Emerging infectious diseases 2020;26:2872.
6. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research 2002;30:3059-66.
7. Masking strategies for SARS-CoV-2 alignments. (Accessed 15 July 2020, at https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480/3.)
8. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution 2020;37:1530-4.
9. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017;8:28-36.
10. Leung K, Shum MH, Leung GM, Lam TT, Wu JT. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Eurosurveillance 2021;26:2002106.

11. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus evolution 2020;6:veaa061.
12. 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:e1006650.
13. Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 Lineage B. 1.1. 7 in England: Insights from linking epidemiological and genetic data. medRxiv 2021:2020.12. 30.20249034.