LETTERS
https://doi.org/10.1038/s41591-020-1000-7

nature medicine

Check for updates

Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling

Rebecca J. Rockett[1,2,10], Alicia Arnott[1,2,3,10], Connie Lam[1,2], Rosemarie Sadsad[1,2,4], Verlaine Timms[1,2], Karen-Ann Gray[1,2], John-Sebastian Eden[1,5], Sheryl Chang[6], Mailie Gall[3], Jenny Draper[3], Eby M. Sim[2,3], Nathan L. Bachmann[2,3], Ian Carter[3], Kerri Basile[3], Roy Byun[7], Matthew V. O'Sullivan[1,2,3], Sharon C-A Chen[1,2,3], Susan Maddocks[3], Tania C. Sorrell[1,2,8], Dominic E. Dwyer[1,2,3], Edward C. Holmes[1,9], Jen Kok[1,2,3], Mikhail Prokopenko[1,6] and Vitali Sintchenko[1,2,3,8] ✉

In January 2020, a novel betacoronavirus (family *Coronaviridae*), named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was identified as the etiological agent of a cluster of pneumonia cases occurring in Wuhan City, Hubei Province, China[1,2]. The disease arising from SARS-CoV-2 infection, coronavirus disease 2019 (COVID-19), subsequently spread rapidly causing a worldwide pandemic. Here we examine the added value of near real-time genome sequencing of SARS-CoV-2 in a subpopulation of infected patients during the first 10 weeks of COVID-19 containment in Australia and compare findings from genomic surveillance with predictions of a computational agent-based model (ABM). Using the Australian census data, the ABM generates over 24 million software agents representing the population of Australia, each with demographic attributes of an anonymous individual. It then simulates transmission of the disease over time, spreading from specific infection sources, using contact rates of individuals within different social contexts. We report that the prospective sequencing of SARS-CoV-2 clarified the probable source of infection in cases where epidemiological links could not be determined, significantly decreased the proportion of COVID-19 cases with contentious links, documented genomically similar cases associated with concurrent transmission in several institutions and identified previously unsuspected links. Only a quarter of sequenced cases appeared to be locally acquired and were concordant with predictions from the ABM. These high-resolution genomic data are crucial to track cases with locally acquired COVID-19 and for timely recognition of independent importations once border restrictions are lifted and trade and travel resume.

The World Health Organization (WHO) declared COVID-19 a pandemic on 11 March 2020, when 118,000 cases had been reported from 110 countries. At the time of writing (18 April 2020), the number of global cases had surpassed 2,000,000 after multiple independent importations of infection from visitors and returned travelers, making the control of this disease of prime global public health importance[3,4]. Major outbreaks have been documented in South Korea, Iran, the United States and Europe[2,3] and person-to-person transmission has been documented primarily through household contacts[5], with up to 85% of human-to-human transmission occurring in family or household clusters[6,7]. To date, our understanding of the mechanisms of disease spread is limited.

These events, combined with estimations from epidemic models, have led to unprecedented measures of disease control being implemented by national governments with profound costs to citizens and economies. Epidemic models of COVID-19 have suggested that virus transmission can be substantially disrupted by rapid detection and quarantine of infectious cases and their contacts[8]. However, validation of COVID-19 modeling predictions has become increasingly important since many models were built using incomplete data and thus produced divergent outcomes[9,10], affecting confidence in public health policy directives. In this study, we used the combination of near real-time SARS-CoV-2 genomic and public health surveillance data to verify inferences from computational models. Genomic epidemiology is a high-resolution tool for public health surveillance and disease control[11–13]. Genomics can augment conventional epidemiological methods by consolidating contentious links and minimizing the impact of case recall bias, which is particularly important in the context of mild or asymptomatic infection.

[1]Marie Bashir Institute for Infectious Diseases and Biosecurity, University of Sydney, Sydney, New South Wales, Australia. [2]Centre for Infectious Diseases and Microbiology–Public Health, Westmead Hospital, Westmead, New South Wales, Australia. [3]Centre for Infectious Diseases and Microbiology Laboratory Services, NSW Health Pathology–Institute of Clinical Pathology and Medical Research, Westmead, New South Wales, Australia. [4]Sydney Informatics Hub, Core Research Facilities, University of Sydney, Sydney, New South Wales, Australia. [5]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, New South Wales, Australia. [6]Centre for Complex Systems, Faculty of Engineering, University of Sydney, Sydney, New South Wales, Australia. [7]Health Protection NSW, NSW Ministry of Health, Sydney, New South Wales, Australia. [8]Centre for Infectious Diseases and Microbiology, Westmead Institute for Medical Research, Westmead, New South Wales, Australia. [9]School of Life and Environmental Sciences and School of Medical Sciences, University of Sydney, Sydney, New South Wales, Australia. [10]These authors contributed equally: Rebecca J. Rockett, Alicia Arnott. ✉e-mail: vitali.sintchenko@sydney.edu.au

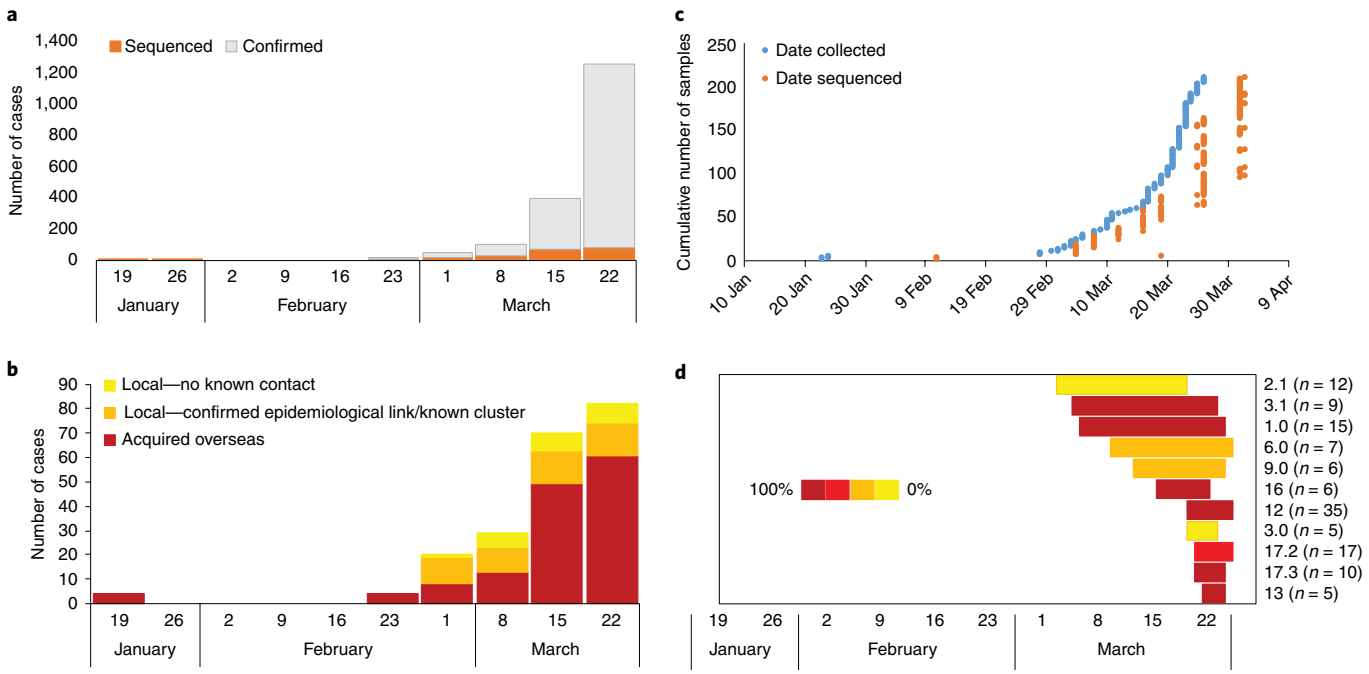


Fig. 1 | Timeline of COVID-19 cases and SARS-CoV-2 genome sequencing in NSW. a, Weekly proportions of confirmed COVID-19 cases (gray) that underwent genome sequencing (orange) during the study period. **b**, Total counts and proportions of imported (red) and locally acquired cases, both epidemiologically linked (orange) and of unknown source (yellow) during the study period by epidemiological week. **c**, Turnaround time of genome sequencing during the first phase of the epidemic. The date of collection (blue) and date of whole-genome sequencing (orange) for each of the 209 samples included in this study are shown. **d**, Timelines of genomically defined clusters. Cluster number and the number of cases (in brackets) are indicated next to the bars. (Only clusters containing five or more cases are shown.) The color gradient reflects the proportion of overseas-acquired cases within the cluster, with the darkest red representing 100% overseas-acquired cases and yellow representing zero overseas-acquired cases in the cluster.

The COVID-19 pandemic has also triggered unrivaled efforts in real-time genome sequencing of SARS-CoV-2. Indeed, thousands of SARS-CoV-2 genomes have already been sequenced internationally and made publicly available from the Global Initiative on Sharing All Influenza Data (GISAID)[14]. Importantly, the ongoing analysis of this global dataset suggests no notable links between SARS-CoV-2 genome sequence variation and virus transmissibility or disease severity[15]. However, even during these early stages of the pandemic, genomic surveillance has been applied to differentiate currently circulating strains into distinct, geographically based lineages and reveal multiple SARS-CoV-2 importations into geographical regions of China and the United States[16,17].

Australia, as an island country between the Pacific and Indian oceans with strong movement of people to and from COVID-19 hotspots in Asia, Europe and North America, has experienced unique challenges and opportunities in responding to the pandemic. The first laboratory-confirmed COVID-19 patients were diagnosed in Melbourne and Sydney on 25 and 26 January 2020, respectively. Since then, and as of the end of this study period on 28 March 2020, 3,635 cases had been confirmed in Australia with 1,617 cases (44.5%) occurring in New South Wales (NSW), the most populous state of Australia (24.5 per 100,000 population)[18]. The Australian Government introduced progressive epidemic mitigation measures from 23 March 2020 to limit social interactions, reduce virus spread and prevent community-based transmissions. This strategy has been supported by widely available testing for SARS-CoV-2 in NSW, with 1,541 tests performed per 100,000 residents during that time[19].

In this study, we examine the value of near real-time genome sequencing of SARS-CoV-2 in understanding local transmission pathways during the containment stage of the COVID-19 epidemic and compare findings from the genomic surveillance of SARS-CoV-2 with predictions of a computational ABM. This comparison was performed to assess the impact of potential sampling bias in genomic surveillance and validate model-based inferences using experimental data. The synergistic use of high-resolution genomic surveillance and computational agent-based modeling not only improves our understanding of SARS-CoV-2 transmission chains in the community and the evolution of this new virus but is also important to help mitigate community-based transmissions.

Between 21 January and 28 March 2020, 1,617 cases of COVID-19 were diagnosed and reported to the NSW Ministry of Health. All patients resided in metropolitan Sydney. Before 29 February 2020, only 4 cases of COVID-19 were detected in NSW, all of which were imported. The first locally acquired case in NSW was reported on 3 March 2020, after which a sharp spike in both imported and locally acquired cases occurred during the week commencing 15 March 2020 (Fig. 1a). Between 1 and 21 March 2020, the weekly proportion of imported cases was between 5 and 20%. During the same period, cases epidemiologically defined as 'unknown origin/under investigation' increased from none during the week beginning 1 March 2020 to between 31 and 35% from 8 to 21 March 2020 (Fig. 1b). Of the 1,617 COVID-19 cases reported during the study period, complete viral genomes were obtained from 209 (13%) (Fig. 1a). Following an initial delay of 21 d between the date of sample collection and sequencing for the first 3 samples received, the median number of days between clinical sample collection and sequencing was 5 d (range: 1–21 d; Fig. 1c). The proportion of COVID-19 cases sequenced weekly peaked during the week commencing 1 March 2020, at 25%. Consensus genome sequences were obtained from all specimens with reverse transcription PCR (RT–PCR) cycle threshold values ≤30. The amplicon sequencing method used in this study provided consistently complete, high-quality consensus sequences with a median coverage depth between 227 and 3,409

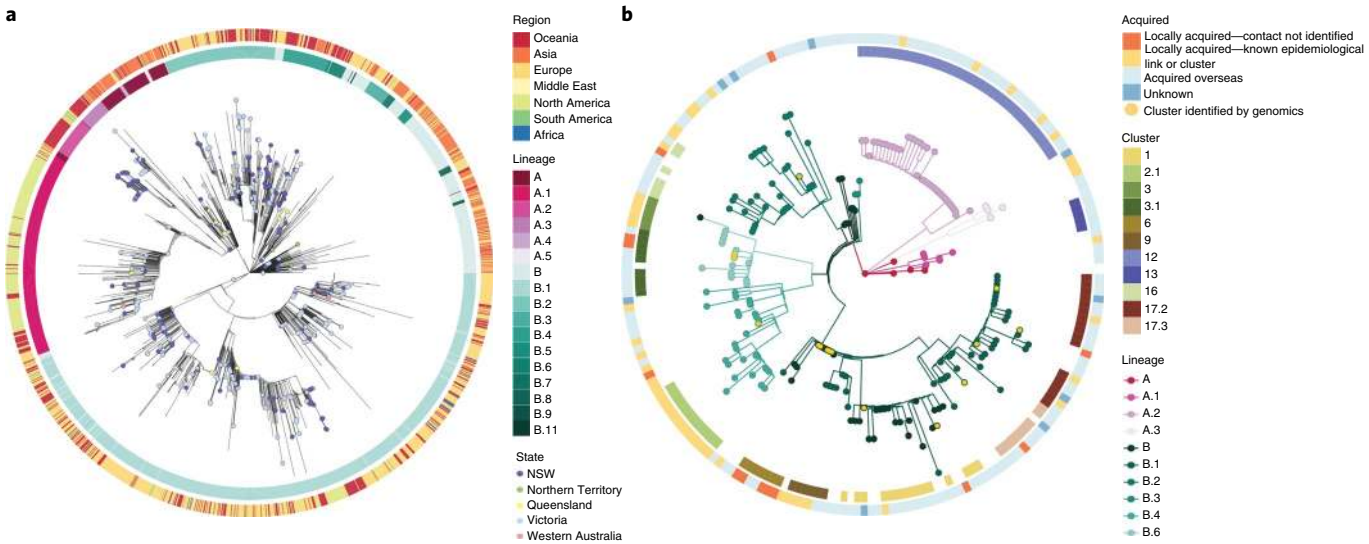


Fig. 2 | Phylogenetic analysis of SARS-CoV-2 genomes. a, SARS-CoV-2 genomes from Australia in the global context provided by GISAID. Australian SARS-CoV-2 genomes are color-coded by the state of the initial diagnosis. The inner ring represents global phylogenetic lineages inferred from the final alignment using pangolin (https://github.com/hCoV-2019/pangolin). The outer ring shows the country of origin of the GISAID genomes. **b**, Phylogenetic relationships between SARS-CoV-2 genomes recovered from patients in NSW. The inner ring represents the allocation of clusters in NSW (only clusters equal or larger than five genomes are presented). The outer ring shows the classification of cases as locally or overseas-acquired based on genomic and epidemiological data. Bootstrap data for lineage inference is presented in Supplementary Table 1.

(Supplementary Table 1). No important genomic changes were identified in the amplicon primer sites of the genomes produced in this study. However, changes in primer binding sites must be monitored to ensure the ongoing ability to produce complete SARS-CoV-2 genomes from clinical samples.

The 209 NSW SARS-CoV-2 genomes were dispersed across the global SARS-CoV-2 phylogeny (Fig. 2a). All major genomic lineages of SARS-CoV-2 were detected in NSW (Fig. 2a and Supplementary Table 1), including the introduction and then local transmission of lineage B.4.2, which is dominated by genomes from Australia that were likely imported from Iran (Fig. 2b). Phylogenetic analysis identified multiple independent introductions of SARS-CoV-2 into NSW over time. Most of the genomes from imported NSW cases grouped with viral sequences from the country of origin. SARS-CoV-2 genomes from ten patients (that is, five household contact pairs) were indistinguishable (three of five pairs) or differed by one single single-nucleotide polymorphism (SNP) (two of five pairs). However, institutional outbreaks, where cases could have only contracted the virus within the institution, demonstrated more genomic diversity. Three institutional outbreaks representing 35, 17 and 12 cases demonstrated up to 2 SNP differences (a single linked case had three SNP differences) between genomes during the outbreak. The duration of these three institutional outbreaks was between 6 and 17 d (detailed in Supplementary Table 1). Therefore, we chose to define clusters as sequences that differed by no more than 2 SNPs from the index case in each outbreak (Fig. 3a), with the index case being the earliest observation of a given sequence. In this study, 27 clusters were identified, of which 8 (29.6%) consisted of 2 cases and 11 (40.7%) contained 5 or more cases. With a single exception, all clusters remained active during the study period (that is, they were associated with continuing onward transmission). The 11 clusters consisting of 5 or more cases were associated with different institutions with no overlapping epidemiological connections (Fig. 1d). The largest cluster contained 35 cases linked by COVID-19 exposure in a single institution (Fig. 3b). Due to the low genetic diversity of SARS-CoV-2, both genomic and epidemiological data were needed to define SARS-CoV-2 clusters. Indeed, when comparing the NSW genomes to the international GISAID dataset, many international genomes clustered closely and would fall within 2 SNPs of our index cases, thus supporting the requirement for epidemiological evidence of local acquisition and contact tracing to confirm genomic clustering. The phylodynamics of the epidemic in NSW were also investigated (Extended Data Fig. 1). However, because genomic clusters were sampled for a limited period (maximum 19 d) they displayed only weak temporal structure ($R^2 = 0.171$ in a root-to-tip regression). The tracking of cluster evolution over time will become important to identify active clusters over a longer sampling period.

Genomic surveillance identified previously unsuspected links to other cases and decreased the proportion of COVID-19 cases with contentious links (Fig. 3b,c). Figure 3 quantifies the added clarity offered by genomics by illuminating the potential sources of infections that could not be identified using conventional epidemiological methods. Genomic evidence was used to cluster 38.7% (81 out of 209) of cases for which the available epidemiological data could not identify direct links (Fig. 3c). This included clustering 12.4% (26 out of 209) of cases with a history of recent arrival from overseas with other cases without a travel history and 5.3% (11 out of 209) of locally acquired cases with unknown epidemiological links. Twenty-two (10.5%) of the 209 cases included in this study were epidemiologically classified as 'locally acquired—contact not identified'. Genomic evidence reduced the proportion of such cases from 10.5% (22 out of 209) to 3.3% (7 out of 209); 15 out of 22 (68.2%) were identified as belonging to 9 genomic clusters containing cases with known epidemiological links (Fig. 3b). The remaining seven cases were genomic singletons, not clustering with other genomes. These findings highlight the use of genomic information in the context of defining disease clusters to inform public health action.

In the ABM, the COVID-19 pandemic spread in Australia originated from overseas, with some infections probabilistically generated in proportion to the average daily number of incoming passengers at airports, and binomially distributed within a 50-km radius of each airport[20]. Figure 4a presents a network formed by community transmission chains, or clusters of linked cases, produced by an ABM run simulating the period corresponding to the time interval between weeks 6 and 10 of the study, that is, the period

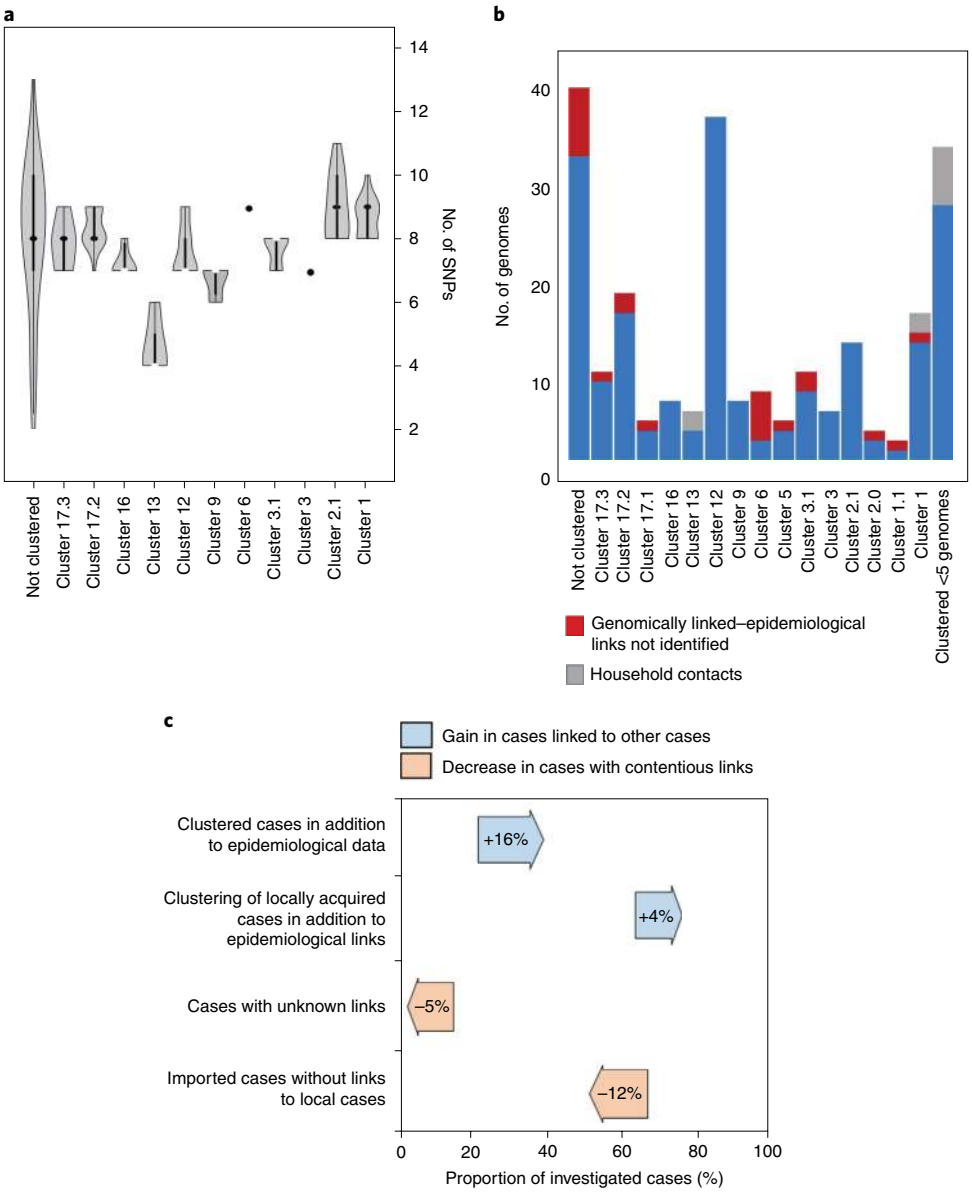


Fig. 3 | Added value of genomic surveillance to the SARS-CoV-2 public health response in NSW. a, Box plot contrasting SNP differences between genomic clusters (containing more than five genomes) and nonclustered genomes of SARS-CoV-2 in NSW. The black circles represent the median; the box limits indicate the 25th and 75th percentiles; the whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; the polygons represent the density estimates of the data and extend to extreme values. **b**, Proportion of cases assigned to SARS-CoV-2 clusters with unknown epidemiological links (red bars), clustered using both genomic and epidemiological data (blue bars) and cases that form part of household clusters (gray bars). **c**, Change bar graph illustrating the impact of SARS-CoV-2 genomic surveillance on epidemiological investigations. The arrow heads indicate the direction of change. The base and tip of the arrow indicate the prior value without genomic evidence and subsequent value after genomic evidence has been included, respectively. The color reflects two major types of added value, increasing recognition of connections between cases with minimal epidemiological data and decreasing in cases with contentious links.

preceding the introduction of major lockdown strategies. A distribution of chain lengths, or cluster sizes, is shown in Fig. 4b and displays similar trends when compared to the distribution of genomic cluster sizes. The average local transmissions within household and household clusters (comprising four adjacent houses) obtained over ten simulation runs, amounting to 18.6% in total (range 15.0–23.9%, with an s.d. of 2.9%) of all infections simulated over the 35-day period, increasing over the weeks to 11.9% per week (range 8.0–14.7%, with an s.d. of 2.1%) during the last week of the study (week 10) (Fig.4c and Supplementary Table 2). When community transmissions are considered within local government areas, which sets an upper bound for all local transmissions, the average local transmissions (across household, household cluster and statistical local area (SLA)), mapped to statistical area level 2 (described in the 2016 national census contexts), reached an upper bound of 34.9% in total (s.d. = 8.2%) of all infections, peaking during the last week at 24.4% per week (s.d. = 5.7%). These fractions are also in strong concordance with their counterparts defined through genomic surveillance (25.8% for all local transmissions, with 17.1–18.3% during the final week; Fig. 4c). As expected, only a proportion of inferred transmission chains were detected by genomic surveillance based on identified COVID-19 cases (Fig.4b).

This is the first report of the convergent application of SARS-CoV-2 genomic surveillance and agent-based modeling to

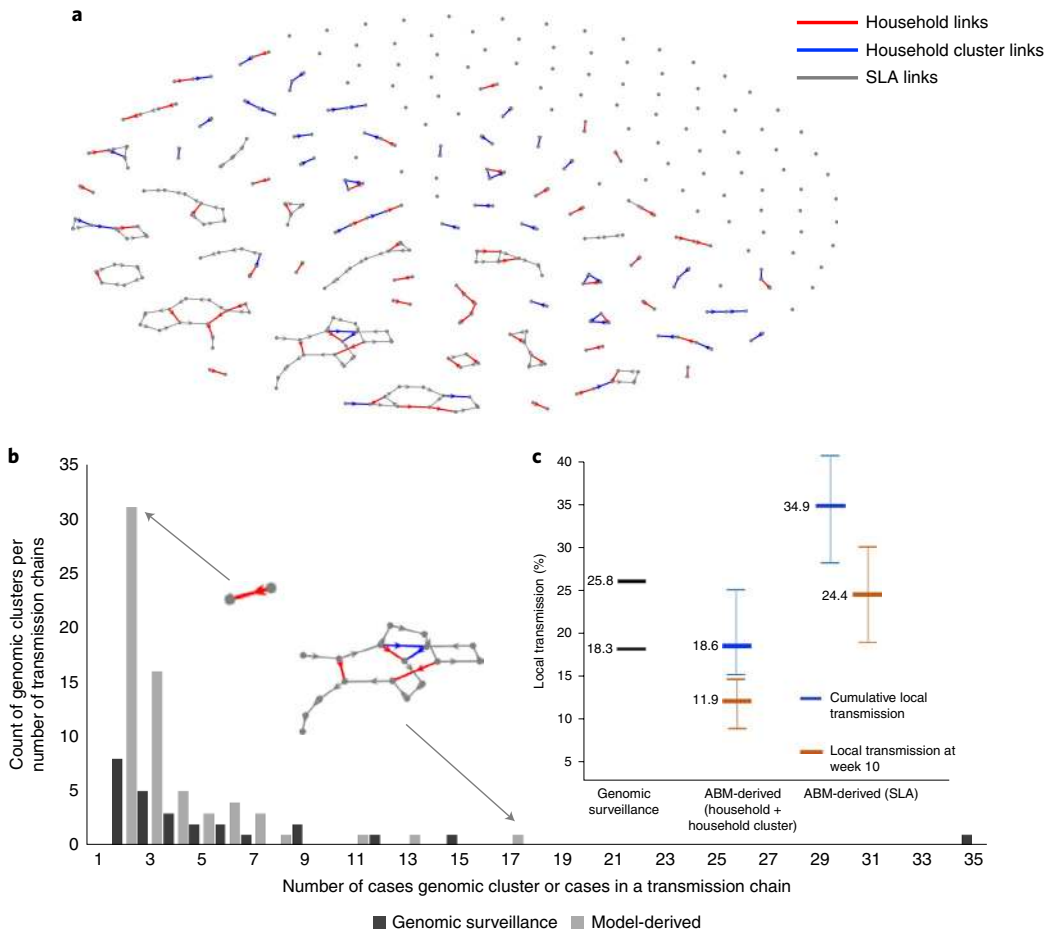


Fig. 4 | Network of SARS-CoV transmission cases generated by the ABM. a, Network of transmission chains originated by initial infections simulated by the ABM and detected within a given social (mixing) context, such as households, household clusters and SLAs. A directed household (household cluster or SLA) link connects two infected individuals in the same household (household cluster or local area), with the direction inferred from the corresponding simulation time steps. A local transmission is defined as any chain formed by links detected within these mixing contexts and may include links across several contexts. **b**, Distribution of chain lengths contrasted with the sizes of genome sequencing-defined clusters. Examples of a transmission chain with 2 cases linked in a household and a transmission chain of 17 cases where cases are linked within household, household cluster and SLA contexts are shown. **c**, Comparison of local transmission rates in NSW defined by SARS-CoV-2 genome sequencing and simulated in Australia by the ABM model. The average over the ten simulation runs is depicted and the error bars represent the minimum and maximum percentages of estimated local transmission events. Genomic surveillance indicates that 25.8% of SARS-CoV-2 cases in NSW are classified as community-based transmission events over the study period, with 18.3% of cases collected in the last week of the study classified as community-based transmission.

investigate local transmission of COVID-19. Strengths of this study include integration of high-resolution genomic data with local epidemiological data and inferences made by agent-based modeling, providing context and confirmation for the genomic results and clustering. The fine-scale resolution provided by the genomic analyses presented in this study is increasingly important for the containment of local outbreaks through identification of secondary cases and providing context for cases of community transmission without apparent epidemiological links[21].

Our findings extend the value of genomic surveillance in depicting transmission pathways and evolution of emerging pathogens[13,22,23]. They also improve our understanding and support the implementation of genomically enhanced surveillance for more efficient COVID-19 control strategies and the investigation of cases with unclear infection sources, with a short turnaround-time[16,17,22,23]. The power of such surveillance is augmented by international sharing of SARS-CoV-2 genomic data collected by researchers and public health microbiology service providers[24,25]. Our clinical specimens originated from a mixture of private and public pathology providers serving hospitals and primary care settings; thus, our sample probably represents the general population with a spectrum of clinical disease. We relied on an average sequence depth greater than 200 to ensure the quality of short read mapping and genome comparisons in line with previous reports[21,26], although further optimization may be required for sustainable and cost-effective SARS-CoV-2 surveillance.

The multiple lineages of SARS-CoV-2 in Australia can be explained by numerous recent, concurrent and independent introductions of COVID-19 from regions with a high frequency of local transmission of COVID-19, such as Asia, Western Europe and North America (Fig. 2a). These observations confirm the dominance of overseas-acquired infections in these early stages of the epidemic in Australia before the implementation of border control measures on 20 March 2020. Our phylogenetic analysis of unique sequences suggested or confirmed the country of origin in specific cases and correlated genomically clustered cases with epidemiologically linked cases (Fig. 1b). This analysis also documented genomically similar cases associated with concurrent transmission in several institutions. This observation was based on the first 10 weeks of the SARS-CoV-2 epidemic in Australia and does not

capture the impact of border control and social/physical distancing measures introduced during this study. The duration of this study was too short to delineate active and inactive clusters. It was also not possible to infer a directionality of virus spread during these early stages of the epidemic since the rate of virus transmission was greater than the rate of evolutionary change in the virus genome[16]. Although the genomic surveillance results are based on a relatively small number of COVID-19 cases in Australia's most populous state, they offer a high-resolution picture of virus spread in the community; the concordance of results from genomic surveillance and agent-based modeling reinforces our conclusions.

This study leveraged the high resolution of inferences offered by agent-based infectious disease models in contrast to the simulation of population-level dynamics models that have been typically employed for public health policy decisions. ABMs capture the interactions of individuals in time and space using government census data and detailed transmission pathways, providing insight into settings where transmission is occurring and how clusters emerge. In our case, the ABM revealed the predominance of early COVID-19 transmission at the household and local community level. The convergent estimates of local transmission from the computational model and genome sequencing not only validated COVID-19 modeling predictions but also suggested that our sampling of COVID-19 cases has been adequate for discovery and monitoring of local transmission events in the time of co-circulation of genomically similar strains of SARS-CoV-2.

While genomic surveillance and the ABM produced congruent results, they still show some difference, for example, recent transmission estimates from ABM were slightly higher than the findings from genomic surveillance, reflecting differences in the scope and detection capacity of these approaches. In addition, the genomic study did not describe transmissions from asymptomatic carriage, while the ABM did not explicitly simulate transmissions in hospitals, residential age care facilities, or introduced by maritime traffic, for example, cruise ships. At the same time, the two approaches reinforced one another largely by addressing these important limitations. Genomic analysis identified transmission occurring in households, the local community and in institutions within the state of NSW, and the ABM captured the complexity of different mixing contexts across the entire nation, including schools and workplaces, with different age-dependent transmission rates and heterogeneous contact patterns, accounting for both symptomatic and asymptomatic infections. The observations of community transmissions obtained using genomic surveillance (25.8% cumulative, with 18.3% during the last week) lay within the uncertainty intervals estimated by the ABM: household/neighborhood-related transmissions (lower bound: 18.6% cumulative, with 11.9% during the last week); and local government area-related estimates (upper bound: 34.9% cumulative, with 24.4% during the last week; Fig.4c).

The synergistic use of genomic surveillance and ABM highlighted the role of transmission within households and local communities, even within a large metropolis such as Sydney, and that genomics data in our context may be less affected by assumptions required for phylogenetic inference and other biases such as sampling density, timing of sample collection and quality of sequencing. Thus, the multimethod 'socio-genomic' approach presented in this study for the examination of emerging epidemics makes our findings more robust and aligned with recent calls for caution in applying virus genomic data to public health investigations[27,28]. Fusion of these two approaches to infer local transmissions reduced overall uncertainty in the scale of the community transmission, presenting actionable inputs for public health policymakers. It must be noted that the actual transmission paths remain uncertain, presenting an open challenge to disease surveillance and warranting further research.

Our team has implemented an ongoing real-time SARS-CoV-2 surveillance system to inform public health response. This genomic survey of SARS-CoV-2 in a subpopulation of infected patients in the first 10 weeks of COVID-19 activity in Australia enabled proactive monitoring of local transmission events in the critical time of implementation of national COVID-19 control measures. Only a quarter of cases appeared to be locally acquired and genomics-derived local transmission rates were concordant with predictions from the computational ABM. This convergent assessment improves our understanding of the transition of COVID-19 from outbreak to pandemic. Integrated analysis of outputs from SARS-CoV-2 genomic surveillance and computational models can refine our understanding of the evolution of the COVID-19 pandemic and are equally relevant for assessment and informing policy for other emerging pathogens. Application of this high-resolution genomic analysis will be crucial to not only track, trace and place locally acquired COVID-19 cases to ensure targeted and informed public health action, but also for the identification of independent importations once current international border restrictions are lifted and trade and travel resume.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-1000-7.

Received: 21 April 2020; Accepted: 25 June 2020;
Published online: 9 July 2020

References

1. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
2. *Coronavirus Disease 2019 (COVID-19). Situation Report 48* https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200308-sitrep-48-covid-19.pdf?sfvrsn=16f7ccef_4 (World Health Organization, 2020).
3. *Coronavirus Disease 2019 (COVID-19). Situation Report 67* https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200327-sitrep-67-covid-19.pdf?sfvrsn=b65f68eb_4 (World Health Organization, 2020).
4. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* **368**, 489–493 (2020).
5. Ghinai, I. et al. First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. *Lancet* **395**, 1137–1144 (2020).
6. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)* https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19) (World Health Organization, 2020).
7. Chan, J. F.-W. et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514–523 (2020).
8. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395**, 931–934 (2020).
9. Verity, R. et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020).
10. Lourenco, J. et al. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. Preprint at *medRxiv* https://doi.org/10.1101/2020.03.24.20042291(2020).
11. Grubaugh, N. D. et al. Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019).
12. Sintchenko, V., Gallego, B., Chung, G. & Coiera, E. Towards bioinformatics assisted infectious disease control. *BMC Bioinformatics* **10**, S10 (2008).
13. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).
14. Shu, Y. & McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data: from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

15. Bedford, J. et al. COVID-19: towards controlling of a pandemic. *Lancet* **395**, 1015–1018 (2020).
16. Deng, X. et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* eabb9263 (2020).
17. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003.e9 (2020).
18. Australian Government Department of Health. Coronavirus (COVID-19) current situation and case numbers https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers (2020).
19. *NSW COVID-19 Case Statistics by Local Health District* (NSW Government, 2020); https://www.health.nsw.gov.au/Infectious/covid-19/Pages/recent-case-updates.aspx
20. Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M. & Prokopenko, M. Modelling transmission and control of the COVID-19 pandemic in Australia. Preprint at *arXiv* https://arxiv.org/abs/2003.10218 (2020).
21. Gudbjartsson, D. F. et al. Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.* **382**, 2302–2315 (2020).
22. Sintchenko, V. & Holmes, E. C. The role of pathogen genomics in assessing disease transmission. *BMJ* **350**, h1314 (2015).
23. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
24. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
25. Dos S Ribeiro, C., Koopmans, M. P. & Haringhuizen, G. B. Threats to timely sharing of pathogen sequencing data. *Science* **362**, 404–406 (2018).
26. Shen, Z. et al. Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. *Clin. Infect. Dis.* ciaa203 (2020).
27. Sintchenko, V., Iredell, J. R. & Gilbert, G. L. Genomic profiling of pathogens for disease management and surveillance. *Nat. Rev. Microbiol.* **5**, 464–470 (2007).
28. Villabona-Arenas, C. J., Hanage, W. P. & Tully, D. C. Phylogenetic interpretation during outbreaks requires caution. *Nat. Microbiol.* **5**, 876–877 (2020).

Methods

Samples for genomic surveillance. All clinical respiratory samples collected between 21 January and 28 March 2020, which tested positive by RT–PCR for SARS-CoV-2 at the Institute of Clinical Pathology and Medical Research (ICPMR) were included in the study. The RT–PCR assay used WHO-recommended primers and probes targeting regions in the E gene and RdRp domain.[29] Samples were submitted directly to the ICPMR for RT–PCR testing and also referred from other NSW Health Pathology and private pathology laboratories for genome sequencing. Genome sequencing was attempted on all PCR-positive samples with a RT–PCR cycle threshold value ≤30.

Epidemiological data and case definitions. Public health follow-up was conducted in parallel for each COVID-19 case. Travel histories were collected and used to determine whether infections were acquired locally or overseas. The origin and number of confirmed COVID-19 cases detected in NSW during the study period were publicly available and obtained from the NSW Health website[30].

An imported case was defined as a person who tested SARS-CoV-2 RT–PCR-positive and reported international travel in the 14 d before illness onset. A locally acquired case was a person who tested SARS-CoV-2 RT–PCR-positive and had not traveled outside Australia in the 14 d before illness onset. Locally acquired cases were further classified on epidemiological grounds as confirmed contacts of a case, as part of a known cluster or of unknown origin (that is, no known contact with a case). Close contacts were defined as individuals who had been in the same closed space for at least 2 h with a laboratory-confirmed case (that is, someone who tested positive for COVID-19) when that person was infectious[31].

The decision to close the Australian border to all except Australian citizens, residents and immediate family members was announced on 20 March 2020. Stage 1 and 2 restrictions were implemented on 23 and 26 March 2020, respectively. Stage 1 restrictions included closure of all restaurants, gymnasiums, cinemas and places of worship, as well as avoiding nonessential travel, outdoor gatherings of >500 and indoor gatherings of >100 people. Stage 2 restrictions were an extension of stage 1 and included advising the public to stay at home unless going to work or education, shopping for essential supplies, undertaking personal exercise or attending medical appointments or compassionate visits[31].

SARS-CoV-2 genome sequencing. We undertook SARS-CoV-2 whole-genome sequencing using an existing amplicon-based Illumina sequencing approach[32,33]. Briefly, RT–PCR-positive samples were reverse-transcribed using SuperScript IV VILO Master Mix (Thermo Fisher Scientific). The viral cDNA was used as input for multiple overlapping PCR reactions (approximately 2.5 kilobases each) that spanned the viral genome using Platinum SuperFi PCR Master Mix (Thermo Fisher Scientific). Amplicons were then pooled equally, purified and quantified before Nextera XT library preparation and multiplex sequencing on an Illumina iSeq or MiniSeq (150 cycles)[34]. All consensus SARS-CoV-2 genomes identified in the study have been uploaded to GISAID (Supplementary Table 1).

SARS-CoV-2 genome analysis. The raw sequence data were subjected to an in-house quality control procedure before further analysis. Demultiplexed reads were quality-trimmed using Trimmomatic (version 0.36; sliding window of 4, minimum read quality score of 20, leading/trailing quality of 5)[35]. The taxonomic identification of the sample was verified using Centrifuge v.1.0.4, based on a database compiled from human, prokaryote and viral sequences (including SARs-CoV-2 RefSeq sequences available before March 2020)[36]. All samples had >99% assignment to the genus 'betacoronavirus'. Reference mapping and consensus calling was performed using iVar v.1.2 (ref. [32]). Briefly, reads were mapped to the reference SARS-CoV-2 genome (National Center for Biotechnology Information (NCBI) GenBank accession MN908947) using Burrows–Wheeler Aligner-MEM v.0.7.17, with unmapped reads discarded. iVar trim was used to soft-clip reads containing primer sequences and discard reads <20 length discarded after trimming. A consensus sequence was called for positions where depth >10 and quality >20 with a minimum frequency threshold of 0.1. The 5′ and 3′ untranslated regions were masked from the consensus due to the poor quality of these regions. QUAST v.5.0.2 was used to evaluate consensus sequence quality in addition to manual inspection in Geneious Prime v.2020.0.5 (ref. [37]). The SARS-CoV-2 genomes from NSW were compared with one another as well as with complete, or near-complete, global genomes available at GISAID (www.gisaid.org, accessed 28 March 2020; see Supplementary Table 2 for the complete list of international genomes used in this study). The quality of GISAID genomes was evaluated using QUAST, with sequences retained only if they were >28,000-base pairs in length and contained <0.05% missing bases (n = 1,985 reference genomes). The GISAID and NSW genomes were aligned with MAFFT v.7.402 (FFT-NS-2, progressive method)[38]. Genomes were trimmed to remove the 5′ and 3′ untranslated regions. Phylogenetic analysis was performed using the maximum likelihood approach (IQ-TREE v.1.6.7 (substitution model: GTR + F + R2, tree root: GISAID accession no. EPI-ISL-413856)) with 1,000 bootstrap replicates[39]; SARS-CoV-2 genomic lineages were inferred using phylogenetic assignment of named global outbreak lineages (pangolin) (https://github.com/hCoV-2019/pangolin). Total SNP numbers between the index SARS-CoV-2 genome from NSW (GISAID accession no. NSW01/EPI-ISL-407893) and each genome in the study was calculated using SNP-dists (excluding ambiguities). SNP differences between the 11 clusters containing 5 or more cases were represented by violin plots using BoxPlotR[40] (http://shiny.chemgrid.org/boxplotr/). Temporal structure and distribution of genomic clusters in NSW was visualized using TreeTime (version 0.7.3)[41]. Phylogenetic trees were constructed using the R package ggtree (version 1.99.1)[42].

Synthesis of genomic and epidemiological information. 'Genomic transmission' in the context of this study refers to the identification of pairs or clusters of cases between which the virus appears to have been transmitted based on high genomic similarity of no more than 2 SNPs. Epidemiological information was used to validate and confirm these genomic links. In instances where a discrepancy between the probable source of acquisition based on genomic and epidemiological data occurred, the source identified on the basis of genomic data was accepted. This was implemented to overcome any limitations of epidemiological information, such as incomplete or incorrect case recall regarding whereabouts and close contacts during their infectious period.

All data described in this study were reported and used to inform public health action in NSW. As such, on the basis of combined genomic and epidemiological information, all cases were categorized using classifications applied by NSW Health, that is, overseas-acquired, local transmission with a confirmed epidemiological link/part of a known cluster or local transmission with no known epidemiological links. Local, or local genomic, transmission refers to cases who did not report overseas travel in the 14 d before illness onset.

Cases with no known genomic or epidemiological links were classified as follows. Genomically, these genomes were classified as 'singletons' and reported as not clustering with other genomes. If the case had traveled overseas during the 14 d before illness onset, such a case would have been classified as 'not clustered, overseas-acquired'. If the case had not traveled during the 14 d before illness onset, the case was classified as 'not clustered, local transmission with no known epidemiological links'. To identify links between all sequenced cases, each new genome generated was analyzed in the context of all others sequenced to date from NSW and indeed, internationally. Continued and prospective analysis of all sequences against our expanding local dataset often subsequently revealed links between new cases and those that did not cluster previously. In the event that cases with no known links subsequently clustered with those newly sequenced, the classification for the case in question was updated in our analysis and reported to NSW Health to inform public health action.

ABM. We used an ABM developed to trace the ongoing COVID-19 epidemic in Australia[20]. This model has been specifically calibrated to reproduce several key characteristics of COVID-19 transmission in Australia, including its reproductive number (R_0 = 2.27), the length of incubation and generation periods (5.0 and 6.4 d, respectively), age-dependent fractions of the symptomatic cases (0.669 and 0.134 for adults and children, respectively) and the probability of transmission from asymptomatic/presymptomatic agents (0.3 of symptomatic individuals). The ABM comprises approximately 24 million software agents, each with attributes of an anonymous individual, matching key demographic statistics from the 2016 Australian census data as well as other data from the Australian Bureau of Statistics and the Australian Curriculum, Assessment and Reporting Authority. Contact and transmission rates were set to differ across distinct social contexts such as households, household clusters, local neighborhoods, schools, classrooms and workplaces. The ABM simulation processes and updates agents' states over time, starting from initial infections, seeded in international airports around Australia, using demographics and mobility layers previously developed and applied to trace influenza pandemics[43,44].

The ABM simulated several social (mixing) contexts, including households, household clusters comprising 4 adjacent houses and SLAs, mapped to statistical area level 2 in the 2016 census (SLAs are local government areas or part thereof). To infer a directed transmission link from the simulation results, we connected any two infected individuals with the same household identifier, the same neighborhood (household cluster) label or the same wider community (SLA) index. A direction of each link represented the temporal lag, as determined by the simulation steps when their (symptomatic) infections were detected by the ABM. A local transmission was defined as any chain formed by links detected within these mixing contexts and could include links across several contexts. For example, individuals A_1 and A_2, infected in the same household A, at different simulation steps were considered to form a transmission link $A_1 \rightarrow A_2$. Similarly, when two infections were detected in the same household cluster, for example, across two households A and B, a directed link was formed, for example, $A_1 \rightarrow B_1$, where household cluster(A_1) = household cluster(B_1). Finally, a link $A_1 \rightarrow C_1$ was constructed when two infections were identified across the individuals sharing the same SLA index, that is, SLA(A_1) = SLA(C_1). We accounted for all links resulting from the wider community transmission within an SLA, potentially capturing the directed triplet network motifs $A_1 \rightarrow A_2$, $A_1 \rightarrow C_1$ and $C_1 \rightarrow A_2$, where A_1 is deemed to infect an individual C_1 in the same SLA at step t_1, before C_1 infects another member A_2 of household A at a later step t_2. The link formation algorithm still registers the potential transmission $A_1 \rightarrow A_2$ within the household context, as it appears in the expected temporal order. In general, triplet motifs are present in various real-world friendship and small-world networks[45,46]. The existence of directed triplet motifs

reflects the uncertainty in identifying the transmission source ($C_1 \rightarrow A_2$ or $A_1 \rightarrow A_2$ in our example), being informed only by temporal dependencies and the context labels. Thus, the inclusion of SLA-wide transmissions provides an upper bound on the extent of community transmission. In other words, transmissions occurring across the wider community are detected to investigate the overall connectivity of emerging local 'clusters'. The COVID-19 model has been validated with the most recent disease surveillance data, being able to predict both the incidence and prevalence peaks in early to mid-April 2020 (ref. [20]).

Statistical methods. Descriptive statistics were employed and measurements were taken from distinct samples. For genome comparison, continuous variables are presented as the median (range: min–max); discrete variables are presented using violin plots to display the median, interquartile range and distribution. Categorical variables are presented as counts and percentages. For the ABM output analysis, averages and s.d. values were calculated from ten simulation runs.

Ethics statement. Clinical specimens were routinely processed at the ICPMR and deemed not research. A nonresearch determination for this project was granted by Health Protection NSW since it was a designated communicable disease control activity.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The original/raw SARS-CoV-2 genome sequencing data are available at the NCBI GenBank (BioProject no. PRJNA633948). The consensus genome sequences for national and international genomes are available from GISAID (www.gisaid.org; see Supplementary Tables 1 and 3). The results of the ABM simulations are presented in Supplementary Table 2 and the ABM data sources have been detailed elsewhere[42,43].

Code availability

No unique pipelines or source code were developed for this project.

References

29. *Laboratory Testing for 2019 Novel Coronavirus (2019-nCoV) in Suspected Human Cases. Interim Guidance* https://www.who.int/publications-detail/laboratory-testing-for-2019-novel-coronavirus-in-suspected-human-cases-20200117 (World Health Organization, 2020).
30. *Coronavirus Disease 2019 (COVID-19). Control Guideline for Public Health Units* https://www.health.nsw.gov.au/Infectious/controlguideline/Pages/novel-coronavirus.aspx (NSW Government, 2020).
31. *Limits on Public Gatherings for Coronavirus (COVID-19)* https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/how-to-protect-yourself-and-others-from-coronavirus-covid-19/limits-on-public-gatherings-for-coronavirus-covid-19 (Australian Goverment Department of Health, 2020).
32. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
33. Di Giallonardo, F. et al. Evolution of human respiratory syncytial virus (RSV) over multiple seasons in New South Wales, Australia. *Viruses* **10**, 476 (2018).
34. Eden, J.-S. et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, veaa027 (2020).
35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
36. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
37. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
38. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
39. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
40. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat. Methods* **11**, 121–122 (2014).
41. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
42. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
43. Cliff, O. M. et al. Investigating spatiotemporal dynamics and synchrony of influenza epidemics in Australia: an agent-based modelling approach. *Simul. Model. Pract. Theory* **87**, 412–431 (2018).
44. Zachreson, C. et al. Urbanization affects peak timing, prevalence, and bimodality of influenza pandemics in Australia: results of a census-calibrated model. *Sci. Adv.* **4**, eaau5294 (2018).
45. Fournet, J. & Barrat, A. Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks. *Sci. Rep.* **6**, 24593 (2016).
46. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

Acknowledgements

Members of the ICPMR SARS-CoV-2 Study Group include L. Donovan, S. Kumar, T. Tran, H. Rahman, D. Ko, T. Sivaruban, E. Martinez, A. Ginn, Q. Wang and K. Pey. Members of the ABM team include N. Harding, C. Zachreson and O. Cliff. We acknowledge the Sydney Informatics Hub and use of the University of Sydney's high-performance computing cluster, Artemis. We thank the NSW Pathology partner laboratories, ACT Pathology, Douglass Hanly Moir, Australian Clinical Laboratories and Laverty Pathology for referring samples for genomic surveillance. Expert advice and epidemiological information provided by the NSW public health surveillance units at the NSW Health Protection are also gratefully acknowledged. The authors are indebted to all researchers and their organizations who have shared SARS-CoV-2 genome data on GISAID. This study was supported by the Prevention Research Support Program funded by the NSW Ministry of Health and the National Health and Medical Research Council Centre for Research Excellence in Emerging Infectious Diseases (no. GNT1102962). V.S. and M.P. are supported by an Australian Research Council grant (no. DP200103005). The funders of this study had no role in study design, data collection, data analysis and interpretation, or writing of the article.

Author contributions

V.S., R.J.R., A.A., D.E.D., J.K., J.-S.E., E.C.H. and M.P. conceived and designed the study. R.J.R., C.L., V.T., K.-A.G., E.M.S., N.L.B. and I.C. processed and tested the samples. R.J.R., C.L., R.S., V.T., K.-A.G., J.-S.E., M.G., J.D., K.B., R.B., V.S. and E.C.H. carried out the sequencing and analysis. S.C. and M.P. carried out the agent-based modeling. V.S., S.C.-A.C., M.V.O., S.M., T.C.S., D.E.D. and J.K. coordinated the study. V.S., R.J.R. and A.A. wrote the first manuscript draft with editing from E.C.H., R.S., T.C.S, D.E.D., S.C.-A.C. and M.P. The final manuscript was approved by all authors.

Competing interests

The authors declare no competing interests.

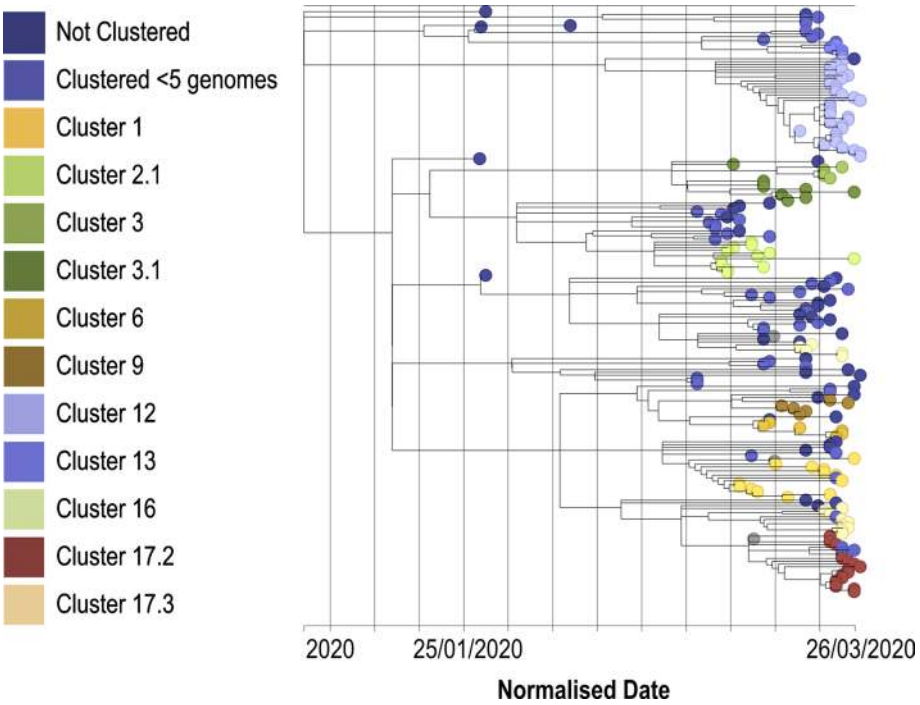
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-020-1000-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-1000-7.

Correspondence and requests for materials should be addressed to V.S.

Peer review information Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



Extended Data Fig. 1 | Phylodynamics of SARS-CoV-2 clusters in NSW. Timestamped tree with selected genomic clusters (five or more cases); first Australian case was used as reference point. The genomic clusters were sampled for a limited period (maximum 19 days) and displayed weak temporal signal ($R^2 = 0.171$).

nature research

Corresponding author(s): Sintchenko NMED-L103844C

Last updated by author(s): Jun 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for collection, 2008 near complete genomes were sourced from GISAID (www.gisaid.org: accessed 28th March 2020).
Data analysis	Demultiplexed reads were trimmed using Trimmomatic (version 0.36). The taxonomic identification of the sample was verified using Centrifuge version 1.0.4. iVar version 1.2 and BWA-mem version 0.7.17 were used for reference mapping. QUAST version 5.0.2 was used to evaluate the consensus sequence quality in addition to manual inspection in Geneious Prime (2020.0.5). The GISAID and NSW genomes were aligned with MAFFT v7.402. Phylogenetic analysis was performed using the maximum likelihood approach (IQTree v1.6.7). SARS-CoV-2 genomic lineages were inferred using Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) (https://github.com/hCoV-2019/pangolin). Total SNP numbers between the index SARS-CoV-2 genome from NSW (GISAID Accession: NSW01/EPI-ISL-407893) and each genome in the study was calculated using SNP-dists (https://github.com/tseemann/snp-dists version 0.6). SNP differences by violin plots using BoxPlotR. Temporal structure and distribution of genomic clusters in NSW was visualised using Treetime (version 0.7.3). Phylogenetic trees were constructed using R package ggtree (version 1.99.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The original/raw SARS-CoV-2 genome sequencing data is available in the National Center for Biotechnology Information (NCBI) GenBank (BioProject PRJNA633948). Consensus genome sequences for national and international genomes are available from the GISAID, (the Global Initiative on Sharing All Influenza Data; www.gisaid.org) (see Supplementary Tables S1 and S3 for data file details). The ABM simulation data is provided in Supplementary Table S2. NSW Health COVID-19 reports can be accessed at https://www.health.nsw.gov.au/Infectious/controlguideline/Pages/novel-coronavirus.aspx

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All clinical samples initially positive by RT-PCR for SARS-CoV-2 in the study period in our laboratory and the network of laboratories referring samples to our laboratory were subjected to culture-independent genome sequencing. Our sample represents 13% of all cases diagnosed in NSW in the study period.
Data exclusions	Duplicate samples and sequential positive samples from same patients were excluded from virus population analysis to minimise sampling bias
Replication	Bootstrap analysis was conducted on phylogenetic tree approximation. 1000 bootstraps were achieved successfully. The agent-based model simulation was repeated ten times to estimate reproducibility of model estimates.
Randomization	Randomization was not applicable as we did not have intervention and control arms in the study. It would be ethically challenging to decline genome sequencing for some of COVID-19 cases and to separate public health investigations between cases or clusters within the same geographic area or jurisdiction
Blinding	Scientists involved in genome data analysis were not blinded to the epidemiological data in order to establish and verify initial criteria for virul genome clustering

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging