The Open Bioinformatics Journal

Content list available at: https://openbioinformaticsjournal.com

RESEARCH ARTICLE

Ancestral Area Reconstruction of SARS-CoV-2 Indicates Multiple Sources of Entry into Australia

Ngoc Minh Hien Phan[1,2], Helen Faddy[1,2,3], Robert Flower[1,2], Kirsten Spann[1] and Eileen Roulis[1,2,*]

[1]*School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Kelvin Grove, Queensland 4059, Australia*
[2]*Research and Development, Australian Red Cross Lifeblood, Kelvin Grove, Queensland 4059, Australia*
[3]*School of Health and Sport Sciences, University of Sunshine Coast, Petrie, Queensland 4502, Australia*

Abstract:

Background:

The ongoing COVID-19 pandemic is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). International travels to Australia during the early stages of the pandemic prior to border closure provided avenues for this virus to spread into Australia. Studies of SARS-CoV-2 biogeographical distribution can contribute to the understanding of the viral original sources to Australia.

Objective:

This study aimed to investigate the clonality and ancestral sources of Australian SARS-CoV-2 isolates using phylogenetic methods.

Methods:

We retrieved 1,346 complete genomes from Australia along with 153 genomes from other countries from the GISAID and NCBI nucleotide databases as of the 14th May 2020. A representative dataset of 270 Australian and international sequences were resulted from performance of nucleotide redundancy reduction by CD-HIT. We then constructed a median-joining network by Network 10.1.0.0, and phylogenies by IQ-Tree, BEAST and FastTree. The Bayesian statistical dispersal-vicariance analysis (S-DIVA) and Bayesian interference for discrete areas (BayArea) built in RASP were used to reconstruct ancestral ranges over the phylogenetic trees.

Results:

Two major clusters, from Europe and from Asia, were observed on the network of 183 haplotypes with distinct nucleotide variations. Analysis of ancestral area reconstruction over the phylogenies indicated most Australian SARS-CoV-2 sequences were disseminated from Europe and East Asia-Southeast Asia.

Conclusion:

The finding is genetic evidence for the geographic origins of the Australian SARS-CoV-2 sequences. Most Australian sequences were genetically similar to those from Europe and East Asia-Southeast Asia, which were also suggested as two main sources of introduction of SARS-CoV-2 to Australia.

Keyword: SARS-CoV-2, COVID-19, Novel severe acute respiratory syndrome coronavirus 2, Ancestral reconstruction, Source of entry, Dispersal routes, GISAID_EPI_ISL_402131.

Article History	Received: December 9, 2020	Revised: March 18, 2021	Accepted: April 25, 2021

1. INTRODUCTION

Infections caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), were first reported in Wuhan, China in December 2019 [1, 2]. In January 2020, outbreaks of COVID-19 were reported outside China, initially in East and Southeast Asia, and then in the USA and Europe before spreading to Australia [1, 2]. The World Health Organisation (WHO) declared COVID-19 a pandemic on the 11th March 2020. As of the 11th March 2021, over 117 million

* Address correspondence to this author at Research and Development, Australian Red Cross Lifeblood, Kelvin Grove, Queensland 4059, Australia;
Tel: +61-738389048; E-mail: eroulis@redcrossblood.org.au

people have been diagnosed as infected worldwide with over 2.6 million deaths [3].

SARS-CoV-2 belongs to the same family, *Coronaviridae*, and genus, *Betacoronavirus*, as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) coronaviruses [4]. SARS-CoV-2 is an enveloped single-stranded RNA virus with >29kb genome, which share 75-80% genetic similarity to SARS-CoV [5] and 96.2% to the bat coronavirus GISAID_EPI_ISL_402131, which was isolated from Yunnan in China [6]. The virus shares a similar genomic organisation to other coronaviruses, including short untranslated regions at both ends and five open reading frames (ORFs) encoding replicase polyproteins (ORF1ab), spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins [7, 8]. People infected with SARS-CoV-2 can have mild symptoms such as fever, cough, sore throat, muscle pain or fatigue, or more severe symptoms such as acute respiratory distress syndrome and shortness of breath [2, 9, 10]. Other symptoms such as shock, diarrhea, and loss of smell/taste have also been reported [2, 9, 10]. The virus is contagious [11] and can be transmitted from human to human via close contact, small droplets or exposure to infected surfaces [4].

The first four cases in Australia were confirmed on the 25th January 2020, one in the state of Victoria (VIC) and three in the state of New South Wales (NSW) [12]. Australia activated a national COVID-19 emergency plan on the 29th February 2020 [13]. At the height of SARS-CoV-2 outbreak in Australia, an approximate doubling of cases was confirmed every 3 days, rising from under 200 cases on the 13th March 2020 to over 2,000 cases on the 27th March 2020 [12, 13]. Overseas acquired cases were more than two times higher than locally acquired cases prior to the closure of Australian borders to all non-citizens and non-residents on the 20th March 2020 [12]. As from 25th June 2020, Australia recorded 7,558 cases and 104 deaths, including 608 cases from Western Australia (WA), 29 from Northern Territory (NT), 1,066 from Queensland (QLD), 3,162 from NSW, 108 from Capital Territory, 440 from South Australia, 1,917 from VIC and 228 from Tasmania [14].

Studies of SARS-CoV-2 biogeographical distribution can contribute to the understanding of the viral original sources and give an insight to its evolutionary history on a phylogenetic tree. The Bayesian statistical dispersal-vicariance analysis (S-DIVA) and Bayesian interference for discrete areas (BayArea) are two commonly used methods for interference of biogeographical histories. S-DIVA is an event-based method that ignores branch lengths and reconstructs ancestral distribution based on a cost matrix of dispersal, vicariance and extinction events under generalised parsimony approaches [15]. In S-DIVA, a dispersal event incurs a cost of one when an area is included to an ancestral range. A cost of zero is applied when an ancestor shares the same area with its direct descendants or when speciation happens by vicariance. Higher cost may be incurred, and the rule of incurring costs is more complicated when multiple areas are involved in ancestral range [15]. Differing from S-DIVA, BayArea is a model-based method of inferring phylogenetic biogeography through a matrix of instantaneous rate of change within a set of discrete geographic areas [16, 17]. Moreover, BayArea accepts trees with polytomies while at least one binary tree has to be included for S-DIVA analysis. This study used a phylogenetic approach to assess the clonality and ancestry of SARS-CoV-2 isolates in Australia and to determine potential sources of dissemination of this virus, through ancestral reconstruction using S-DIVA and BayArea.

2. MATERIALS AND METHODS

2.1. Sequence Dataset Collection and Preparation

All human SARS-CoV-2 complete genomes from Australia were retrieved from the NCBI nucleotide database and Global Initiative On Sharing All Influenza Data (GISAID) as available on 14th May 2020. In addition, one to fourteen SARS-CoV-2 complete genomes were randomly selected per country or region outside Australia, from either GISAID or NCBI nucleotide database, to represent overseas SARS-CoV-2 cases. The first sequenced SARS-CoV-2 isolate, NC_045512_China/Wuhan (NCBI) was included as a reference sequence and dated 2019-12-20 according to the patient's onset of disease [18]. Duplicate identification or accession numbers were removed from the analysis. We performed manual curation of the dataset and removed 128 sequences having ≥3% of unassigned or ambiguous nucleotides over their entire genomes. CD-HIT EST [19, 20] was used to cluster nucleotide sequences of ≥ 99.5% similarity. As CD-HIT only retained one representative sequence from each cluster, regardless of the country of origin of the sequence, many of the international sequences would have been discarded. Therefore, we re-included SARS-CoV-2 genomes originating outside Australia that were removed by CD-HIT, in order to have at least a representative sequence for each nation even if they could technically be represented by another sequence. This resulted in a final dataset of 117 representative sequences from Australia and 153 representative sequences internationally.

The collected sequences were then aligned using MAFFT 2.1.11 [21, 22]. Misalignments were manually edited and regions with high numbers of Ns and ambiguous nucleotides were trimmed out. The latter regions predominantly occurred within the 3'- and 5'- untranslated regions, corresponding to nucleotide positions start->55 and 29837->end of the reference sequence NC_045512_China/Wuhan|2019-12-20.

2.2. Clonality Analysis by Network Map

For clonality analysis, a variable dataset, in which gaps were not considered and invariable sites were removed, was generated from the DNA alignment of 270 sequences by DNAsp 6 [23]. A Median-Joining (MJ) network was then built using Network 10.1.0.0 [24, 25] (fluxus-engineering.com) with default weights at 10, connection cost for distance calculation, and an epsilon value of 10.

2.3. Ancestral Reconstruction Phylogeny Analysis

Approximated maximum-likelihood (ML) trees were built using FastTree 2.1.11 [26, 27] and IQ-tree 1.6.12 [28]. For FastTree, a default setting was used with 20 rate categories of sites. For IQ-tree, ModelFinder [29] tested 286 DNA models and selected the substitution model GTR+F+R3 as the best-fit

model under Bayesian information criterion (BIC) to compute a consensus ML tree from 1000 ultrafast bootstrap replicates [30]. For additional comparison, we performed Bayesian analysis of molecular sequences under a Markov Chain Monte Carlo (MCMC) method using BEAST 1.10.4 [31] with uncorrelated relaxed clock, GRT substitution model, empirical base frequencies, heterogeneity model of 4 gamma categories and an assumption of constant population size. TreeAnnotator 1.10.4 [32] was used to generate a target Maximum Clade Credibility (MCC) tree summarised from sampled posterior trees produced by BEAST with the first 1,000 tree samples discarded as burn-in.

For all analyses, the trees were rooted to the reference NC_045512_China/Wuhan|2019-12-20. The distribution range of SARS-CoV-2 isolates was assigned into eleven areas as per their country of origin: (A) Australia, (B) Southeast Asia, (C) West Asia, (D) East Asia, (E) South Asia, (F) North Asia, (G) North America, (H) South America, (I) Central Asia, (J) Europe, and (K) Africa. We applied S-DIVA [33] implemented within RASP [34] to the binary trees generated by IQ-tree and BEAST to obtain possible ancestral ranges at each node with averaged costs for dispersals between areas. The number of maximum areas was set to 4 as default. We used the BayArea method [16] within RASP to infer phylogenetic biogeography for the non-binary tree given by FastTree. Posterior probabilities of ancestral states were estimated at nodes on the FastTree phylogeny. The constructed ancestors on the phylogenies were compared in terms of the degree of congruence between the different trees and models. Where similar trees were obtained and to allow for better visualisation, a condensed version of a tree was created with reconstructed ancestral ranges using by using interactive Tree of Life (iTOL) [35].

3. RESULTS

3.1. Sequence Dataset Collection and Preparation

Of the 1,499 human SARS-CoV-2 complete genomes initially collected from GISAID and NCBI nucleotide databases (Table **S1**), we obtained 1,218 genomes isolated within Australia and 153 genomes outside Australia (Fig. **1**) after removing 128 sequences having ≥3% noninformative sites – 124 from the state of VIC, Australia and four from the NT, Australia. By clustering sequences with a nucleotide sequence identity of 99.5% and selecting a representative sequence for each cluster, CD-HIT excluded 1,101 SARS-CoV-2 sequences from Australia from 5 states/territories, and 117 genomes outside Australia, which were then manually re-included, finalising 270 genomes for downstream analysis (Fig. **1**). The sequence MT121215_China|2020-02-02 was representative of the largest cluster of 844 sequences, 702 of which were from Australia – 211 from NSW, 457 from VIC, 19 from QLD, 9 from WA and 6 from NT. The hCoV-19/Australia/VIC683/2020|EPI_ISL_426978|2020-03-29 and hCoV-19/Australia/NT30/2020|EPI_ISL_430633|2020 sequences were representative of two Australian clusters of 158 and 104 SARS-CoV-2 genomes, respectively. The other sequences selected by CD-HIT represented smaller groups of 24 or fewer genomes. The clusters and representative sequences generated by CD-HIT redundancy reduction at a threshold of 99.5% nucleotide similarity was summarised in Table **S2**.

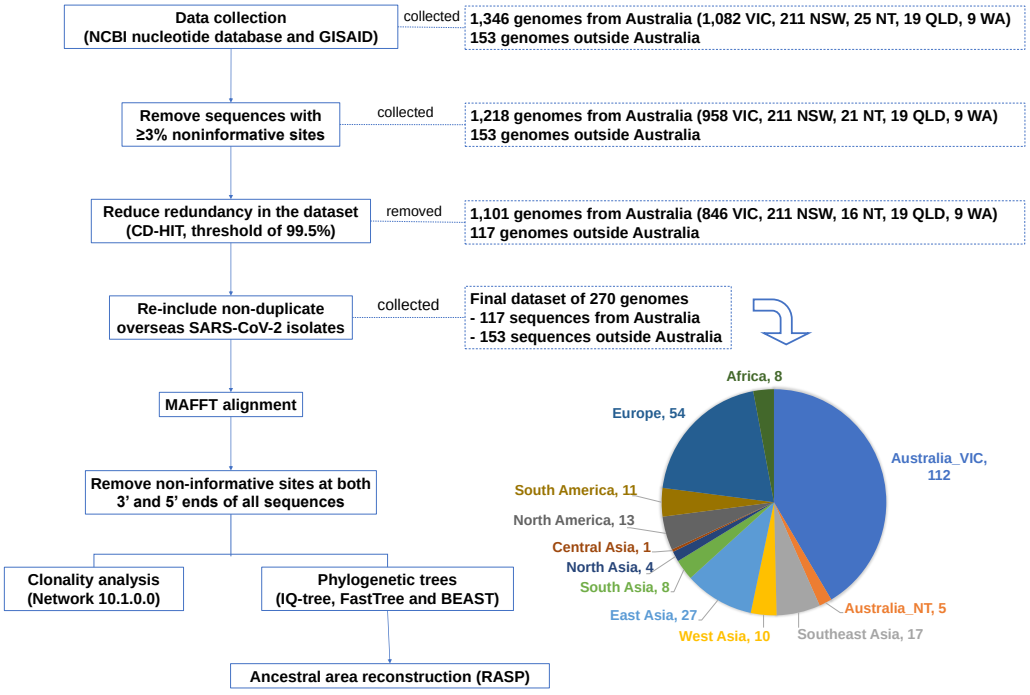

Fig. (1). Number of SARS-CoV-2 complete genomes retrieved or removed at each step of sequence data collection and preparation. State/territory of Australia: VIC - Victoria, NSW - New South Wales, NT - Northern Territory, QLD - Queensland, and WA - Western Australia.

3.2. Clonality Analysis by Network Map

The phylogenetic network of the 270 SARS-CoV-2 sequences from Australia and overseas revealed two obvious clusters (Fig. **2**), implying two main groups of SARS-CoV-2 isolates circulating in Australia. The majority of the Australian isolates were clustered with those from Europe, with the second cluster primarily from Asia, especially East Asia, as well as partly from North America. Smaller groups of Australian SARS-CoV-2 sequences were genetically more similar to isolates from North and South America. These clusters of SARS-Cov-2 sequences with origins from different regions suggest that Australia and Europe / Asia / North America had transmissions from the same sources at the same time. When examining distal nodes, further diversification of Australian isolates, that appears to have occurred subsequent to the main clusters, highlights evidence of local transmission within Australia. Taken all together, these findings suggest that SARS-CoV-2 isolates in Australia are multiclonal.

3.3. Ancestral Reconstruction Phylogeny Analysis

Reconstruction of ancestral area distribution based on ML and MCC trees allowed us to trace the most recent common ancestor of all lineages and to understand the biogeographic relationships among SARS-CoV-2 isolates. Dispersal routes from different (sub)continents to Australia were identified by S-DIVA using estimated costs (for IQ-tree and BEAST), where the cost of dispersal (the movement of a viral species across a geographical barrier to a new environment) is greater than vicariance (the diversification of a viral species within a geographical area) [15, 36]. The greater the cost, the more support for a dispersal route (Table **1**). The data also indicated dissemination of SARS-CoV-2 isolates from Australia to other countries (Table **S1**), particularly to Europe and Africa, the investigation of which is not within the scope of the study. The S-DIVA analysis showed the disseminations of SARS-CoV-2 between Australia (A) and Europe (J) occurred the most frequently, demonstrated by the highest costs for these routes (Tables **1** and **S1**). Like the analysis of SARS-CoV-2 network map, the ancestral area reconstructions showed several clusters of Australian isolates themselves along the MCC and ML trees, implying the occurrence of local transmission events in this country.

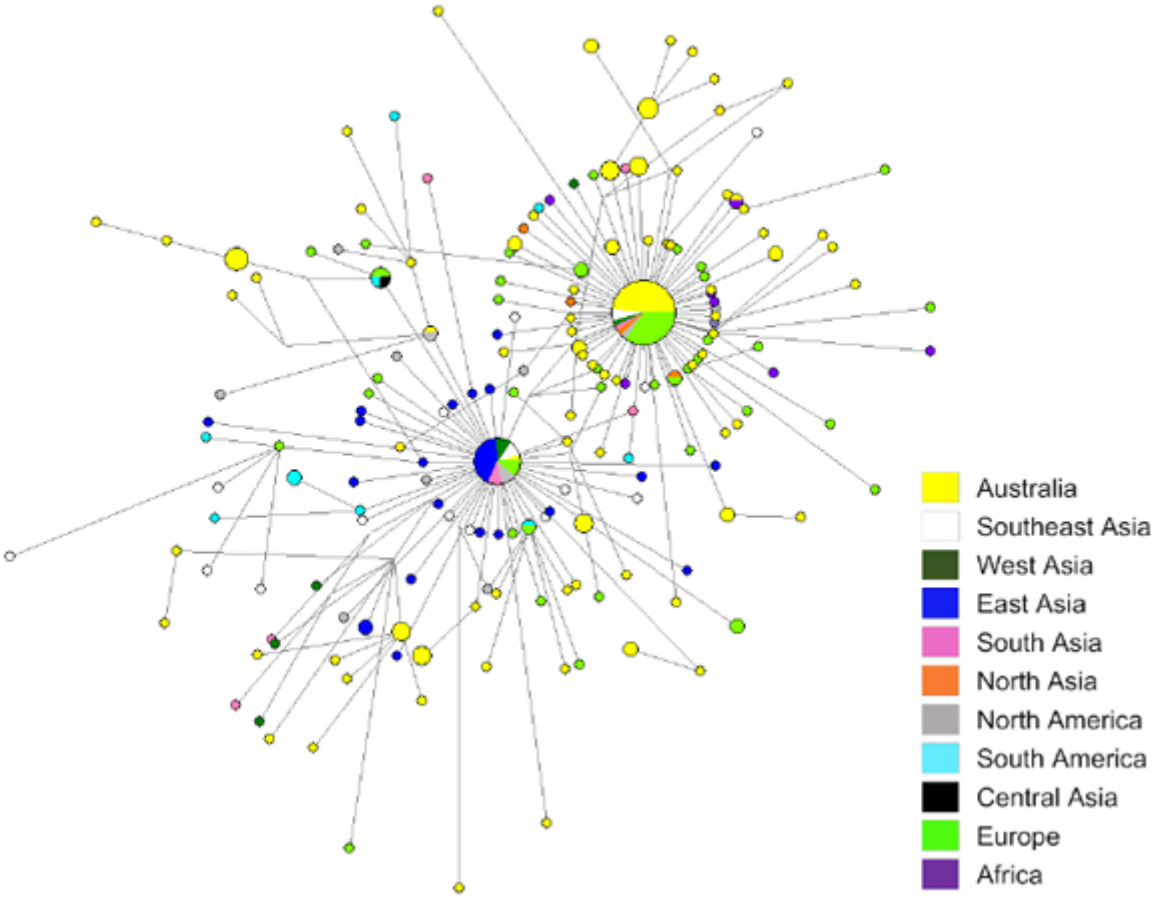

Fig. (2). Median-joining network of 270 sequences of SARS-CoV-2 isolates from Australia and overseas. The size of nodes is proportional to the number of sequences in each cluster and the areas of the nodes are proportional to the number of the sequences from each coloured geographical region. The 270 SARS-CoV-2 isolates yielded 183 haplotypes with distinct nucleotide variations. The branch length is proportional to the number of variable nucleotide positions considered amongst the generated haplotypes.

Table 1. Single area dispersal events to Australia identified for the IQ-tree, BEAST and FastTree phylogenies*.

Phylogeny (method)	On IQ-tree phylogeny (S-DIVA)**	On BEAST phylogeny (S-DIVA)**	On FastTree phylogeny (BayArea)
Dispersal route to Australia	B->A:1.5 C->A:2.5 D->A:1.666667 E->A:1 G->A:1.5 H->A:0.5 J->A:14.33333	B->A:0.5 C->A:1.5 D->A:2 H->A:1 J->A:12	B->A C->A D->A E->A F->A G->A H->A J->A K->A
Cost of dispersals to Australia	23	17	N/A
Cost of all dispersals identified over the tree	122	123	N/A

* (A) Australia, (B) Southeast Asia, (C) West Asia, (D) East Asia, (E) South Asia, (F) North Asia, (G) North America, (H) South America, (I) Central Asia, (J) Europe and (K) Africa. ** Dispersal costs estimated by S-DIVA for corresponding dispersal events.

For the IQ-tree phylogeny, all lineages converged to the node ABDJ, indicating that Southeast Asia, East Asia and Europe were most likely common ancestors of SARS-CoV-2 isolates in Australia (Figs. **3** and **S1**). In addition to ABDJ, Europe (J) or regions including Europe as part of ancestral ranges such as DJ, CJ, or GJ, followed by East Asia (D), had sequences more closely related to the majority of Australian isolates, suggesting those areas as potential sources of SARS-CoV-2 dissemination into the country. West Asia (C), North America (G), South Asia (E) and South America (H) were also observed in internal ancestral nodes for a number of Australian sequences, although to a lesser extent. Similarly, the S-DIVA analysis on BEAST tree (Fig. **S2**) suggests that the region that combined Southeast Asia and East Asia (BD) was the most likely common ancestor for all lineages. Meanwhile, ancestral ranges of Europe (J), East Asia (D) and/or West Asia (C) were located more proximally to most Australian clusters, indicating these regions were additional minor sources of viral entry to Australia. We used BayArea analysis to reconstruct ancestral ranges on the FastTree phylogeny (Fig. **S3**). The finding was consistent with the previous analyses from IQ-tree and BEAST trees, except for a greater range of geographic locations with the addition of Africa (K), North Asia (F) as minor sources of viral dissemination to Australia (Fig. **S3** and Table **1**).

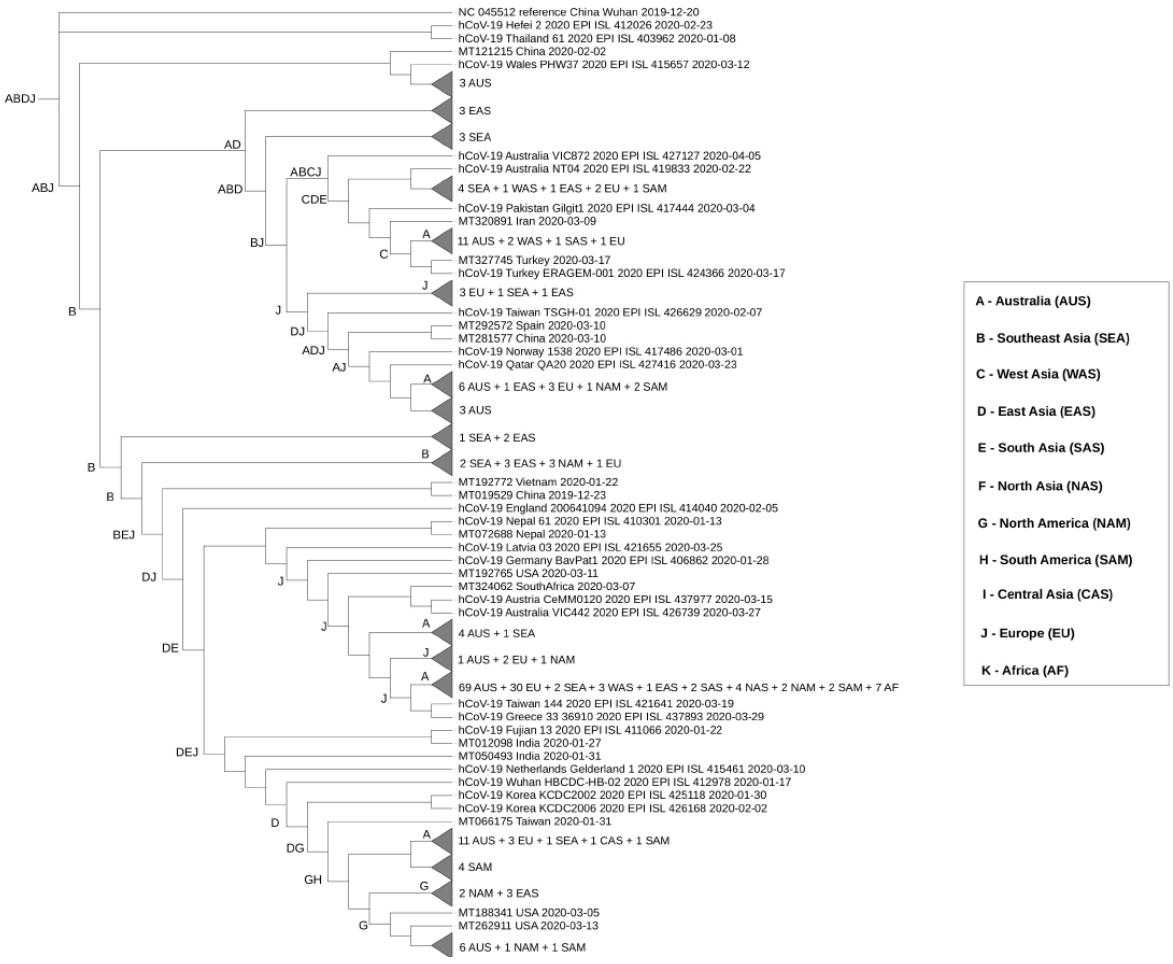

Fig. (3). Most likely ancestral ranges at nodes identified with dispersal events for the IQ-tree consensus maximum likelihood phylogeny.

4. DISCUSSION

Our phylogenetic study investigated the clonality of Australian SARS-CoV-2 isolates, up to the 14th March 2020, in a local and international context. The network analysis of 270 sequences representative of 1,499 SARS-CoV-2 complete genomes from both Australia and internationally indicates viral multiclonality in Australia. Specifically, we found two main clusters of Australian SARS-CoV-2 isolates: one cluster aligned mainly with isolates from Europe, while a second cluster aligned primarily with isolates from Asia, particularly East Asia. Minor clusters were aligned with those from the Americas. The finding is supported by the observation of two main ancestral ranges (BD/ABDJ and J) identified over our three ML and MCC trees. This differs somewhat to a study by Foster *et al.* (2020) [37] in which 160 SARS-CoV-2 complete genomes sampled worldwide were analysed through its phylogenetic network. This study identified three main clusters from Europe, the Americas and East Asia, distinguished by amino acid variations. We found most Australian sequences aligned with those from Asia and Europe, and very few sequences aligned with those from the Americas. This difference could be explained by the effectiveness of public health interventions in Australia [38] such as border closure on the 20th March 2020 [12] or enforced quarantine on the 28th March 2020 [39] before the number of confirmed cases in the United States started to accumulate exponentially from late March 2020 [40].

Similarly, a study by Yu *et al.* (2020) [41] categorised SARS-CoV-2 isolates from Australia into three groups (B, C and D) clustering with sequences mainly from China, Europe and the United States. In our study, the Americas was observed as a minor source of viral dissemination. This may be due to differences in the methods utilised and the number of Australian SARS-CoV-2 sequences used in our study compared to the study by Yu *et al*. (2020) [41]. The study of Yu *et al*. (2020). focused on the evolutionary history of SARS-CoV-2 in China and worldwide [41] instead of transmission patterns into Australia [41], and included only six Australian sequences to represent all Australian cases.

The analysis of ancestral reconstructions at internal nodes from the trees identified five (sub)continents as ancestral areas or points of dissemination from which SARS-CoV-2 spread to Australia: Southeast Asia (B), West Asia (C), East Asia (D), South America (H), and Europe (J). Meanwhile, the analyses of the ML and MCC trees conducted here were all congruent in identifying Southeast and East Asia (BD) as the most common recent ancestors of all lineages, ABDJ for the IQ-tree phylogeny and BD for the FastTree and BEAST trees. The majority of the sampled Australian SARS-CoV-2 isolates were descendants of those originating from Europe (*i.e.* J, CJ, GJ or DJ). This is concurrent with findings from a phylodynamic analysis by Seemann *et al*. (2020) [38] that showed Asia and Europe as two of three main sources for the clusters of returning Victorian travellers. Furthermore, the finding on IQ-tree and FastTree trees identified dispersal events from North America (G) and South Asia (E) to Australia while FastTree proposed additional dispersal routes from North Asia (F) and Africa (K). Based on the evolutionary relationships among SARS-CoV-2 isolates from different countries and event routes at reconstructed ancestral nodes, we proposed three main routes for SARS-CoV-2 dissemination into Australia: (1) viruses dispersed directly from East Asia and/or Southeast Asia to Australia (2) viruses dispersed to several geographic locations mostly within Europe before entering to Australia, (3) viruses dispersed to several geographic locations mostly outside Europe before entering to Australia.

Nextstrain reported five distinct clades of SARS-CoV-2, namely 19A, 19B, 20A, 20B and 20C [42, 43]. 20A, 20B and 20C were more frequent in Europe and South America [44] while 19A and 19B were more frequently observed in Asia [43]. However, the geographical distribution of the viral clades is approximate, depending on sampling and reporting systems, number of infected cases, application of travel restrictions and different public health responses among countries worldwide. As observed on Nextstrain, all five clades with greater numbers of clades 19B and 20C, were present in Australia during the first five months of the pandemic, followed by an increase in cases of clade 20 B in NSW and VIC from April 2020 [45]. Regarding viral transmissions into Australia on Nextstrain, SARS-CoV-2 were mostly disseminated from Asia, then Europe during the early stage of the pandemic with additional sources from America and Africa subsequently [45]. The observation on Nextstrain is generally congruent with the findings from our study during the early stage of the pandemic. The two main clusters observed through our network analysis could correspond to clades 19B and 20C found on Nextstrain, with generally similar observed dissemination routes mainly from Asia and Europe.

Community transmission was also demonstrated by the clusters of Australian sequences along the constructed phylogenies in this study. However, no relevant demographic information of the SARS-CoV-2 isolates other than their country of origin was available on GISAID and NCBI, their social risks of transmission for these cases are unknown. Meanwhile, the findings of Seeman *et al.* (2020) provided genomic evidence of local transmission which was highly associated with social gatherings [38].

We recognise that our study has limitations. The process of sequence retrieval, data curation and clustering resulted in 117 Australian isolates - 112 Australian from VIC and 5 from NT, compared to 1,346 sequences initially collected. This means all sequences from the other Australian states than VIC and NT were removed because of being clustered with representative local and overseas sequences in our dataset (Fig. **1**) meanwhile these isolates represented independent introductions. This may overestimate or underestimate the importance of certain routes of SARS-CoV-2 dissemination to Australia and does not give a representation of community transmission within Australia. For instance, there were fewer Australian isolates clustered with lineages dispersed from East/South East Asia than there were from Europe. This is likely due to the removal of all isolates from NSW, QLD and WA, 457 from VIC and 6 from NT, totalling over Australian 700 viral sequences from our initial dataset by CD-HIT clustering, with these sequences represented by MT121215_China|2020-02-02 from East Asia. Likewise, most of the sampled Australian SARS-CoV-2

isolates were from the state of VIC and no travel history data could be collected for these isolates. Therefore, this limits the ability to confirm the findings and determine possible interstate transmission of SARS-CoV-2 within Australia.

Although some studies on the transmission of SARS-CoV-2 in Australia have been conducted, the understanding of viral introduction into and between different Australian states is not well understood. A study on local transmission of SARS-CoV-2 in Australia identified 28% of 209 NSW isolates collected from 21st January to 28th March 2020 were from locally acquired infections, 18.6% of these were from household contacts [46]. The study also observed multiple overseas importations of the virus into NSW, from the same regions we identified in our study [46]. According to a study by Seeman *et al*. (2020), 61% of infected cases in VIC from 6th January to 14th April 2020 were acquired overseas [38]. Moreover, over 50% of locally acquired cases were reported to be associated with multiple SARS-CoV-2 introductions by these travellers [38].

CONCLUSION

In conclusion, this investigation contributes to our understanding of multiple clonality and dissemination of SARS-CoV-2 isolates circulating in Australia. We highlighted a number of geographic areas from which SARS-CoV-2 viruses circulating in Australia originated, with at least two major dispersals into this country. Using network and phylogenetic analysis, we demonstrate that at least two types of the same sources appear to be circulating within Australia, confirming the multiclonality of Australian SARS-CoV-2 isolates. We demonstrate that Europe (J) and East Asia-Southeast Asia (BD) appear to be the main geographical regions of dissemination into Australia, while confirming East and South-East Asian isolates as the most common recent ancestor to all circulating worldwide. The finding is considered genetic support for strict deferrals and restrictions to donors with recent travel or exposure history to minimise risk to the blood supply, although SARS-CoV-2 has not been known as transfusion transmissible.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used in the study that is the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The SARS-CoV-2 sequences used in this study are available in the GISAID and NCBI nucleotide databases (https://www.gisaid.org/ and https://www.ncbi.nlm.nih.gov /sars-cov-2/).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

FUNDING

None.

ACKNOWLEDGEMENTS

We acknowledge Queensland University of Technology for granting to N. M. H. P. Australian Government Research Training Program (RTP) Stipend (International) and QUT HDR Tuition Fee Sponsorship for her Doctor of Philosophy degree. The authors gratefully acknowledge Aoibhe Mulcahy for her advice on using iTOL to condense phylogenies.

SUPPLEMENTARY MATERIAL

Supplementary materials are available on the publishers' website along with the published article.

REFERENCES

[1] Zhu N, Zhang D, Wang W, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020; 382(8): 727-33. [http://dx.doi.org/10.1056/NEJMoa2001017] [PMID: 31978945]

[2] Wang D, Hu B, Hu C, *et al.* Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020; 323(11): 1061-9. [http://dx.doi.org/10.1001/jama.2020.1585] [PMID: 32031570]

[3] World Health Organization. 2020.https://covid19.who.int/

[4] Uddin M, Mustafa F, Rizvi TA, *et al.* SARS-CoV-2/COVID-19: viral genomics, epidemiology, vaccines, and therapeutic interventions. Viruses 2020; 12(5): 526. [http://dx.doi.org/10.3390/v12050526] [PMID: 32397688]

[5] Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 2020; 6(1): 14. [http://dx.doi.org/10.1038/s41421-020-0153-3] [PMID: 33723226]

[6] Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol 2020; 92(5): 522-8. [http://dx.doi.org/10.1002/jmv.25700] [PMID: 32027036]

[7] Woo PC, Huang Y, Lau SK, Yuen K-Y. Coronavirus genomics and bioinformatics analysis. Viruses 2010; 2(8): 1804-20. [http://dx.doi.org/10.3390/v2081803] [PMID: 21994708]

[8] Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep 2020; 19100682 [http://dx.doi.org/10.1016/j.genrep.2020.100682] [PMID: 32300673]

[9] Sun P, Qie S, Liu Z, Ren J, Li K, Xi J. Clinical characteristics of hospitalized patients with SARS-CoV-2 infection: A single arm meta-analysis. J Med Virol 2020; 92(6): 612-7. [http://dx.doi.org/10.1002/jmv.25735] [PMID: 32108351]

[10] Xydakis MS, Dehgani-Mobaraki P, Holbrook EH, *et al.* Smell and taste dysfunction in patients with COVID-19. Lancet Infect Dis 2020; 20(9): 1015-6. [http://dx.doi.org/10.1016/S1473-3099(20)30293-0] [PMID: 32304629]

[11] Sanche S, Lin YT, Xu C, *et al.* The novel coronavirus, 2019-nCoV, is highly contagious and more infectious than initially estimated medRxiv 2020. [http://dx.doi.org/10.1101/2020.02.07.20021154]

[12] Ting I, Palmer A. One hundred days of the coronavirus crisis. ABC News 2020.

[13] Handley E. From Wuhan to Australia: a timeline of key events in the spread of the deadly coronavirus. ABC News 2020.

[14] Australian Government, Department of Health. 2020.https://www.health.gov.au/news/health-alerts/ novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers

[15] Lamm KS, Redelings BD. Reconstructing ancestral ranges in historical biogeography: properties and prospects. J Syst Evol 2009; 47(5):

369-82.
[http://dx.doi.org/10.1111/j.1759-6831.2009.00042.x]

[16] Landis MJ, Matzke NJ, Moore BR, Huelsenbeck JP. Bayesian analysis of biogeography when the number of areas is large. Syst Biol 2013; 62(6): 789-804.
[http://dx.doi.org/10.1093/sysbio/syt040] [PMID: 23736102]

[17] Ree RH, Moore BR, Webb CO, Donoghue MJ. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. Evolution 2005; 59(11): 2299-311.
[http://dx.doi.org/10.1111/j.0014-3820.2005.tb00940.x] [PMID: 16396171]

[18] Wu F, Zhao S, Yu B, *et al.* A new coronavirus associated with human respiratory disease in China. Nature 2020; 579(7798): 265-9.
[http://dx.doi.org/10.1038/s41586-020-2008-3] [PMID: 32015508]

[19] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics 2010; 26(5): 680-2.
[http://dx.doi.org/10.1093/bioinformatics/btq003] [PMID: 20053844]

[20] Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22(13): 1658-9.
[http://dx.doi.org/10.1093/bioinformatics/btl158] [PMID: 16731699]

[21] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002; 30(14): 3059-66.
[http://dx.doi.org/10.1093/nar/gkf436] [PMID: 12136088]

[22] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013; 30(4): 772-80.
[http://dx.doi.org/10.1093/molbev/mst010] [PMID: 23329690]

[23] Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, *et al.* DnaSP 6: DNA sequence Ppolymorphism analysis of large data sets. Mol Biol Evol 2017; 34(12): 3299-302.
[http://dx.doi.org/10.1093/molbev/msx248] [PMID: 29029172]

[24] Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999; 16(1): 37-48.
[http://dx.doi.org/10.1093/oxfordjournals.molbev.a026036] [PMID: 10331250]

[25] http://fluxus-engineering.com

[26] Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009; 26(7): 1641-50.
[http://dx.doi.org/10.1093/molbev/msp077] [PMID: 19377059]

[27] Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 2010; 5(3)e9490
[http://dx.doi.org/10.1371/journal.pone.0009490] [PMID: 20224823]

[28] Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015; 32(1): 268-74.
[http://dx.doi.org/10.1093/molbev/msu300] [PMID: 25371430]

[29] Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods 2017; 14(6): 587-9.
[http://dx.doi.org/10.1038/nmeth.4285] [PMID: 28481363]

[30] Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 2013; 30(5): 1188-95.
[http://dx.doi.org/10.1093/molbev/mst024] [PMID: 23418397]

[31] Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 2018; 4(1)vey016
[http://dx.doi.org/10.1093/ve/vey016] [PMID: 29942656]

[32] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BioMed Central 2007; p. 214.

[33] Yu Y, Harris AJ, He X. S-DIVA (Statistical Dispersal-Vicariance Analysis): A tool for inferring biogeographic histories. Mol Phylogenet Evol 2010; 56(2): 848-50.
[http://dx.doi.org/10.1016/j.ympev.2010.04.011] [PMID: 20399277]

[34] Yu Y, Blair C, He X. RASP 4: Ancestral state reconstruction tool for multiple genes and characters. Mol Biol Evol 2020; 37(2): 604-6.
[http://dx.doi.org/10.1093/molbev/msz257] [PMID: 31670774]

[35] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res 2019; 47(W1): W256-9.
[http://dx.doi.org/10.1093/nar/gkz239] [PMID: 30931475]

[36] Kodandaramaiah U. Use of dispersal–vicariance analysis in biogeography – a critique. J Biogeogr 2010; 37(1): 3-11.
[http://dx.doi.org/10.1111/j.1365-2699.2009.02221.x]

[37] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 2020; 117(17): 9241-3.
[http://dx.doi.org/10.1073/pnas.2004999117] [PMID: 32269081]

[38] Seemann T, Lane CR, Sherry NL, *et al.* Tracking the COVID-19 pandemic in Australia using genomics. Nat Commun 2020; 11(1): 4376.
[http://dx.doi.org/10.1038/s41467-020-18314-x] [PMID: 32873808]

[39] Worthington B. Forced coronavirus quarantine for all people returning to Australia 2020.

[40] Roser M, Ritchie H, Ortiz-Ospina E, Hasell J. 2020. Available from: https://ourworldindata.org/coronavirus

[41] Yu W-B, Tang G-D, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. Zool Res 2020; 41(3): 247-57.
[http://dx.doi.org/10.24272/j.issn.2095-8137.2020.022] [PMID: 32351056]

[42] Hadfield J, Megill C, Bell SM, *et al.* Nextstrain: Real-time tracking of pathogen evolution. Bioinformatics 2018; 34(23): 4121-3.
[http://dx.doi.org/10.1093/bioinformatics/bty407] [PMID: 29790939]

[43] Bedford T, Neher R, Hadfield J, *et al.* Genomic epidemiology of novel coronavirus - global subsampling 2020. Available from: https://nextstrain.org/ncov/global

[44] Alm E, Broberg EK, Connor T, *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill 2020; 25(32): 1.
[http://dx.doi.org/10.2807/1560-7917.ES.2020.25.32.2001410] [PMID: 32794443]

[45] Bedford T, Neher R, Hadfield J, *et al.* Genomic epidemiology of novel coronavirus - Oceania-focused subsampling: Nextstrain 2020. Available from: https://nextstrain.org/ncov/oceania?p=full

[46] Rockett RJ, Arnott A, Lam C, *et al.* Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med 2020; 26(9): 1398-404.
[http://dx.doi.org/10.1038/s41591-020-1000-7] [PMID: 32647358]

© 2021 Phan *et al.*

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: https://creativecommons.org/licenses/by/4.0/legalcode. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.