Virus Evolution, 2024, **10(1)**, veae085

DOI: https://doi.org/10.1093/ve/veae085
Advance Access Publication 17 October 2024
Research Article

Genomic epidemiology reveals the variation and transmission properties of SARS-CoV-2 in a single-source community outbreak

Ning Zhao[1,‡], Min He[1,2,‡], HengXue Wang[1,‡], LiGuo Zhu[3,‡], Nan Wang[1], Wei Yong[1], HuaFeng Fan[1], SongNing Ding[1], Tao Ma[1], Zhong Zhang[1], XiaoXiao Dong[1], ZiYu Wang[1], XiaoQing Dong[1], XiaoYu Min[1], HongBo Zhang[1], Jie Ding[1,2,*]

[1]Microbiology Laboratory, Nanjing Medical University Affiliated Nanjing Municipal Center for Disease Control and Prevention, 2 Zizhulin Road, Nanjing, Jiangsu 210003, China
[2]School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China
[3]Department of Acute Infectious Disease Control and Prevention, Jiangsu Provincial Center for Disease Control and Prevention, 172 Jiangsu Road, Nanjing, Jiangsu 210009, China

[‡]These authors contributed equally to this work and share first authorship.
[*]Corresponding author. Microbiology Laboratory, Nanjing Medical University Affiliated Nanjing Municipal Center for Disease Control and Prevention, 2 Zizhulin Road, Nanjing, Jiangsu 210003, China. E-mail: yu2an2002@163.com

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused the coronavirus disease 2019 (COVID-19) pandemic, which is still a global public health concern. During March 2022, a rapid and confined single-source outbreak of SARS-CoV-2 was identified in a community in Nanjing municipal city. Overall, 95 individuals had laboratory-confirmed SARS-CoV-2 infection. The whole genomes of 61 viral samples were obtained, which were all members of the BA.2.2 lineage and clearly demonstrated the presence of one large clade, and all the infections could be traced back to the original index case. The most distant sequence from the index case presented a difference of 4 SNPs, and 118 intrahost single-nucleotide variants (iSNVs) at 74 genomic sites were identified. Some minor iSNVs can be transmitted and subsequently rapidly fixed in the viral population. The minor iSNVs transmission resulted in at least two nucleotide substitutions among all seven SNPs identified in the outbreak, generating genetically diverse populations. We estimated the overall transmission bottleneck size to be 3 using 11 convincing donor–recipient transmission pairs. Our study provides new insights into genomic epidemiology and viral transmission, revealing how iSNVs become fixed in local clusters, followed by viral transmission across the community, which contributes to population diversity.

Keywords: SARS-CoV-2; genomic epidemiology; phylogeny; variant; bottleneck

Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of coronavirus disease 2019 (COVID-19). SARS-CoV-2 variants with increased transmissibility, increased virulence, or decreased vaccine effectiveness were designated as previously circulating variants of concern (VOCs), namely, Alpha, Beta, Gamma, Delta, and Omicron (Zhao et al. 2022, Carabelli et al. 2023). Among them, the SARS-CoV-2 Omicron variant exhibits striking immune evasion ability and is rapidly spreading worldwide (Hong et al. 2022).

With the widespread use of deep sequencing methods, genomic epidemiology has become a powerful tool for determining the public health response to communicable disease outbreaks (Lu et al. 2020, Komissarov et al. 2021, Aggarwal et al. 2022, Gu et al. 2022, MacCannell et al. 2022). This is particularly true for diseases such as COVID-19, which shows a high spreading speed and a high proportion of asymptomatic infections. SARS-CoV-2 infections have been associated with outbreaks in many types of settings. Whole-genome sequencing (WGS) has been used to monitor the emergence of circulating strains and identify contentious links in outbreaks, enabling accurate clustering.

Moreover, SARS-CoV-2 mutates between and within hosts to alter viral infectivity, disease severity, or interactions with host immunity. Intrahost single-nucleotide variants (iSNVs) that emerge in the course of a virus epidemic could provide valuable information about the sizes of transmission bottlenecks, chains of person-to-person transmission, viral diversity, and the process of virus evolution (Wang et al. 2021b). Viral transmission bottlenecks determine the amount of genetic diversity produced in one host that could be passed on to another during transmission and generally constrain the evolution of viruses (Markov et al. 2023). Although current studies have mostly shown that SARS-CoV-2 has a narrow transmission bottleneck, further exploration may need to identify whether variants with certain characteristics are more likely to be transferred or whether the narrow bottleneck means that the passages of viral particles occur in a more random way such that minor iSNVs also have similar chances. Whether early Omicron BA.2.2 transmission exhibits characteristics similar to those of the Delta variant or the previous VOCs with respect to iSNV transmission and bottleneck sizes remain to be further elucidated.

In this study, we investigated the transmission characteristics of the Omicron BA.2.2 variant during a community-level outbreak that occurred in early March 2022. Prompt responses, including population nucleic acid screening in high-risk areas, intensive contact tracing, quarantining, and genomic analysis, were implemented, and the outbreak occurred over a short period. Our results provide genetic and epidemiological evidence and identify the infection source and variant profiles of this community transmission event.

Methods

Case definition, sample collection, and epidemiological investigation

Samples were collected, and SARS-CoV-2 testing was performed on oropharyngeal swabs during this outbreak. The PCR screening tests were performed by third-party institutions or the local Center for Disease Prevention and Control (CDC). Epidemiological surveys were implemented for all the cases, and the exposure history of positive cases and their close contacts was obtained through field investigations. Interviews were conducted, and public video surveillance systems were used to identify those who had direct or indirect contact with the cases. The contacts of the index cases were quarantined centrally or at home.

Genomic sequencing of SARS-CoV-2

Combined with epidemiological information, some of the positive samples were sequenced. The Target Capture Kit for SARS-CoV-2 Whole Genome (Baiyi Technology Co., Ltd, China) was used for the reverse transcription and genome amplification of the extracted RNA samples. The sequencing libraries were prepared via the Nextera XT Library Prep Kit (Illumina, USA) and subjected to end repair, A-tailing, and adaptor ligation. Negative controls were prepared with nuclease-free water. A high-throughput sequencing protocol according to the Illumina MiniSeq High Output Reagent Kit (300 cycles) was used. The genomic data were analyzed via the BAIYI MicroGeno Platform (v4.1, Hangzhou Baiyi Technology Co., Ltd, http://www.baiyi-tech.cn/, China). Fastp (Chen et al. 2018) (v0.23.2, https://github.com/OpenGene/fastp) was used to control the quality of the original data, and bwa (0.7.17-r1188, https://github.com/lh3/bwa) was used to compare the data to the SARS-CoV-2 reference genome Wuhan-Hu-1 (GenBank Accession No: MN908947.3). The whole-genome sequence was assembled via bcftools (Danecek et al. 2021) (V1.12, https://github.com/samtools/bcftools). Samples with coverage less than 95% were filtered out. Some cases were sampled and sequenced twice, and the higher coverage sequences were retained. Finally, 61 filtered SARS-CoV-2 samples were used for subsequent analysis.

Phylogenetic tree construction

From January 2022 to 30 April 2022, genomes of the Pango lineage BA.2 and its subbranches were randomly selected from the GISAID database (http://gisaid.org/), and a total of 4642 SARS-CoV-2 sequences were obtained. After mafft (v7.487, https://github.com/GSLBiotech/mafft) was used to compare the sequences from the GISAID database with 61 samples from this study, a phylogenetic tree was constructed via FastTree (Price et al. 2009) (http://www.microbesonline.org/fasttree/). A time-related phylogenetic tree was constructed via nextstrain build (v7.0.1, https://github.com/nextstrain/ncov).

SNP and iSNV analysis

Snippy (v4.6.0, https://github.com/tseemann/snippy) was used to calculate the differences in mutation sites between 61 samples and obtain the difference matrix. SNP sites were analyzed via the sns.clustermap (V0.11.1) in Python on the basis of the Euclidean distance. Nucleotide variations in relation to the reference sequence (the consensus sequence of NJ01) were identified and classified as iSNVs, which coexisted with the reference allele at an identical position. Mutational sites were detected via freeBayes (v1.3.2, https://github.com/freebayes/freebayes), and sites with a mutation frequency greater than 5% were retained. We defined iSNVs as those with alternative allele frequencies (AAFs) between 5% and 95% (Wang et al. 2021a). The iSNVs were identified only in samples with a minimum coverage of 60X of the read data for 95% of the genomic regions. The minor allele frequency was accepted by at least five reads. The consensus sequence was generated according to the majority alleles (more than 50%) at each position.

Bottleneck estimation

We excluded the head and tail sequences of the viral genome (positions 1–100 and 29 803 to 29 903) and the 23 “highly shared” sites (Lythgoe et al. 2021, Li et al. 2022). On the basis of the defined chains of transmission, we identified 11 donor–recipient pairs. We applied the approximate version and exact version of the beta-binomial sampling method (Leonardl et al. 2017), using a 5% minimum variant frequency cutoff to call the variants. The error bars denote the 95% confidence intervals. We estimated the bottleneck sizes of each transmission pair individually and calculated the overall transmission bottleneck sizes across transmission pairs.

Results

Outbreak description and epidemiological information

A local outbreak was declared in Jiangning district when the index patient NJ01 was identified via regular PCR screening for returning populations on the morning of 10 March 2022. We discovered 95 people who had SARS-CoV-2 infections between 10 and 26 March 2022, with a time frame of 16 days. The date of diagnosis of the first positive case, NJ01, was referred to as Day 0. On the basis of the discovered transmission connections and epidemiological links involving residences and time frames, these cases were grouped into six epidemiological clusters (A–F). The six epidemiological clusters of patients were connected by a presumed SARS-CoV-2 transmission network (Fig. 1 and Supplementary material). On the evening of March 9, the index case, NJ01, in Cluster A wandered through the neighborhood without wearing a mask, potentially infecting cases NJ06 in Cluster B, NJ07 in Cluster C, and NJ38 in Cluster D, who were in the same public spaces at exactly the same time as NJ01 was. Before recognition and quarantine decisions were made, these secondary cases initiated subsequent transmission and formed the corresponding epidemiological clusters. The

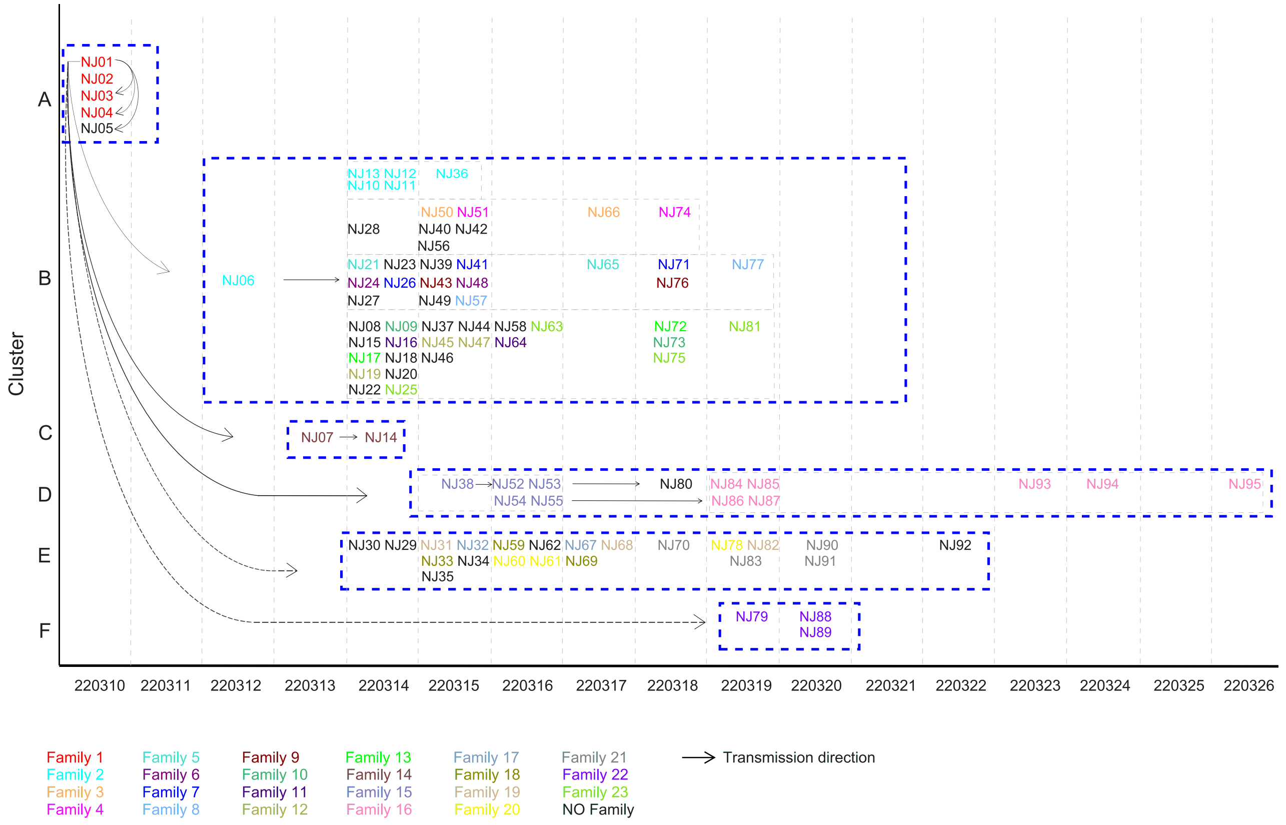

Figure 1. Schematic diagram of the assumed transmission of epidemiological Clusters A–F. *x*-axis: time; *y*-axis: epidemiological cluster. The solid arrows refer to direct epidemiologically confirmed transmission, and the dashed arrows indicate the putative direction of transmission. The same epidemiological clusters are within the blue dashed boxes. The same color indicates the same family.

initial infected individuals in Clusters E and F were most likely also infected by the index case, according to their geographical proximity to where NJ01 resided. However, unlike cases NJ06, NJ07, and NJ38, there was no additional evidence to support this conclusion, indicating the existence of undetected chains of transmission in the neighborhood via epidemiological methods.

Just 2 days after NJ06 was identified, 20 cases in Cluster B were revealed simultaneously in routine screening tests, among which 14 were located in the same building as NJ06. Another characteristic of the outbreak was the transmission propensities in confined spaces, with obvious apartment building and family clustering (Supplementary material).

Phylogenetic analysis

By using the Illumina sequencing platform, 61 high-quality whole-genome sequences were obtained, with an average sequencing depth of 5746.65 and a coverage depth of 95.58%–99.37% (Supplementary Table S1). WGS revealed two closely related lineages, namely, BA.2.2.1 (52/61) and BA.2.2 (9/61). The minor branch BA.2.2 comprised the index sample sequence NJ01 and eight other sequences that were 100% similar to NJ01, including 1 sequence in Cluster A, 2 sequences in Cluster C, and 5 sequences in Cluster D.

As shown in Fig. 2, all 61 successfully sequenced samples formed a compact clade, and these sequencing data were compatible with a single introduction of the Omicron virus and confirmed the linkage of onward case transmission in the community. The concentrated branch provided strong evidence for the epidemiologically suspected common source of the 6 clusters, characteristic of a unique signature, T14 034C. The additional general signature of Clusters B, E, and F was the mutation T6226C, with all the cases having this alteration, and 1 strain in Cluster A also contained this mutation. We identified one additional consensus variation, T358A (NJ77), in Cluster B. For Clusters A and D, there was no common consensus alteration that was shared by all the strains. One strain in each of Clusters A and D presented a mutation at locus 12 655. The four patients from one family in Cluster D carried an extra signature combination of G2129A, T17863C, and A27691C, including one patient with a further mutation, C25611T. Compared with NJ01, Cluster D had the greatest variety and presented five mutations at a consensus level, with one strain containing four substitutions, which was the most distant sequence from the index case NJ01. Unlike the other clusters, among which most patients were no more than tertiary cases, the four cases with 3–4 SNPs in Cluster D were quaternary and quinary cases. We assume that more generations are the main explanation for the greater SNP occurrence within this cluster.

The presence and transmission of iSNVs

The mutational sites associated with the consensus level of the SARS-CoV-2 genome data from this event were analyzed (reference genome: Wuhan-Hu-1, MN908947.3), and the results revealed that there were 82 nucleotide mutational sites (Fig. 3a and Supplementary Table S2), of which 14 034 was the unique mutational site of this outbreak. In Wuhan-Hu-1, the most frequent nucleotide substitutions (SNPs) were C > T (36.84%), G > A (11.94%), A > G (10.52%) and T > C (8.97%), and 68.27% of the substitutions were transitions. However, compared with the original strain of NJ01, all seven mutations at a consensus level were T > C (2 sequences), C > T, G > A, A > C, T > G and T > A, and 57.14% (4/7)

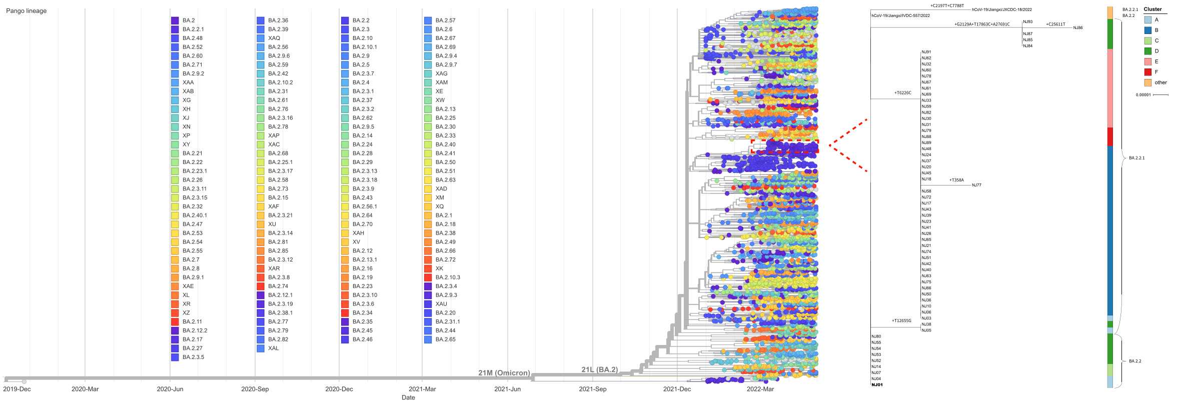

Figure 2. Phylogenetic analyses of SARS-CoV-2 genome sequences. A total of 4642 sequences were randomly selected from GISAID, and 61 sequences were selected for phylogenetic analysis. The zoomed-in section shows the strains in this study and those with close phylogenetic relationships (right). The sublineages are marked with different colors.

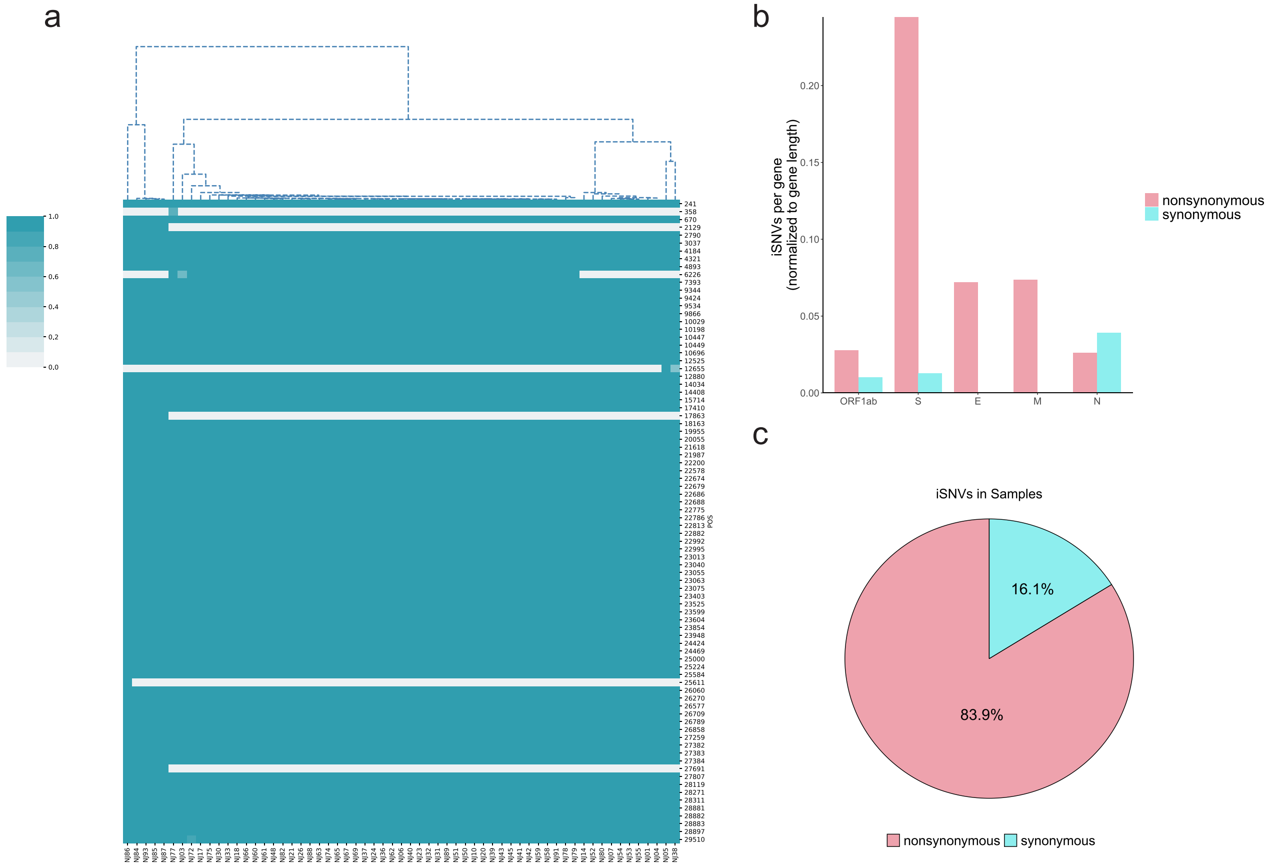

Figure 3. Nucleotide mutations in SARS-CoV-2 genomes. (a) Hierarchical clustering of nucleotide mutations (SNPs). The reference genome was Wuhan-Hu-1 (GenBank Accession No: MN908947.3). The color scale shows the degree of variation in each genome, from lowest (white, no mutation) to highest (blue, complete mutation). (b) Ratio of nonsynonymous to synonymous mutations (iSNVs). The reference genome was NJ01. (c) Distributions of synonymous and nonsynonymous mutations (iSNVs). The reference genome was NJ01. Pink indicates nonsynonymous mutations, and green indicates synonymous mutations.

of the substitutions were transitions. When we altered the reference (ancestor) strains, the nucleotide variation profile differed, possibly related to the length of the evolutionary time and the sample size.

Compared with NJ01, 74 identified iSNV sites were distributed across genomic regions, with nonsynonymous iSNV sites accounting for 77.03% (57/74) and synonymous iSNV sites (17/74) accounting for 22.97%. In addition, 118 iSNVs occurred in the 61

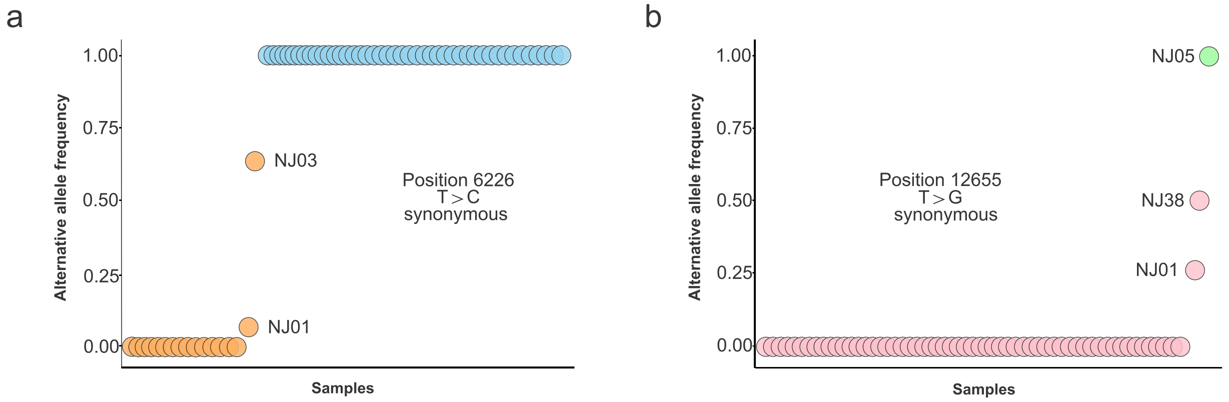

Figure 4. Low-frequency and fixed mutations. (a) Allele frequency of the synonymous mutation T > C at site 6226. (b) Allele frequency of the synonymous mutation T > G at site 12,655. Orange and pink represent no mutation or low-frequency mutations, and blue and green represent complete mutations.

individuals (Supplementary Table S3). We further analyzed the iSNVs/kb in the 61 samples, and an overall relatively low density of iSNVs (0.065 iSNVs/kb) was identified, which is comparable with the number of iSNVs identified in a previous study (0.0041 mean iSNVs per 100 sites) (Lythgoe et al. 2021). The highest frequency of iSNVs/kb was identified at S (0.26), followed by M (0.07) and E (0.07) (Fig. 3b), which was consistent with the frequency distribution of iSNVs in a previous report, indicating an uneven genomic distribution (Gu et al. 2023). The most abundant mutational patterns of iSNVs were T > A (34.75%), C > T (13.56%), G > T (11.02%), A > G (9.32%), T > C (6.78%), A > T (5.93%), and T > G (5.08%), and 33.05% of the iSNVs were transitions. The ratio of nonsynonymous to synonymous variants in all patients was 5.21 (Fig. 3c). Most (81.08%) iSNV sites were present in a single patient, whereas a small number of the iSNVs (18.92%) were present in at least two patients or were identical to consensus mutations. This suggested that most iSNVs are randomized mutations that occur at different positions rather than recurrent changes that occur frequently at specific regions under selection pressure. In addition, 93.10% of the shared variants were nonsynonymous, and 6.9% were synonymous. We compared the iSNV sites associated with this outbreak with the 82 accumulated SNP sites associated with the Wuhan strain, and only 4 sites overlapped.

If the donor sequences with minor iSNVs had corresponding recipients, the transmission of the iSNVs could be observed regarding the emergence and fixation of variants in the recipients, or the fade of the iSNVs, in contrast, could be observed in some transmission pairs. The minor iSNV T6226C, with a frequency of 7.98%, was delivered from the index case to NJ03 in Cluster A and NJ06 in Cluster B, resulting in an increased frequency of 66.80% in NJ03 and the fixed substitution of T6226C in NJ06 in one generation of transmission. This substitution of T6226C was eventually present in 44 out of the 61 (72.1%) subjects (Fig. 4a). We also observed a low allele frequency (26.85%) at site T12 655 G in the index case, which increased to a higher frequency of 50.21% in NJ38 in Cluster D and was fixed (99.89%) in NJ05 in Cluster A (Fig. 4b). However, T12655G was confined to only two samples in this outbreak.

Bottleneck size estimation

We calculated the transmission bottleneck size among the 11 epidemiologically defined donor–recipient transmission pairs via the beta binomial method. We identified a stringent bottleneck size, which is consistent with previous studies (Lythgoe et al. 2021, Wang et al. 2021a, Hannon et al. 2022). Finally, the maximum likelihood estimate for the overall transmission bottleneck size was 3. The transmission bottleneck sizes for the defined epidemiological pairs are shown in Fig. 5a and Supplementary Table S4. We observed how the minor iSNVs carried by the donor host were transmitted when one donor had multiple recipients (Fig. 5b). Our results revealed that the transmission bottleneck of SARS-CoV-2 was generally narrow, with most donor iSNVs not found in the recipients. However, the three relatively larger bottlenecks of 11, 23, and 23, which were derived from three transmission pairs with the original case as the donor, might have contributed to the diversity of the virus population during the outbreak.

Discussion

Here, we reconstructed the transmission mode of SARS-CoV-2 in this community outbreak. At two polymorphic nucleotide sites (6226 and 12655), we analyzed the process of low-frequency iSNV fixation, which has great implications for understanding viral transmission and evolutionary directions. It was evident that some variants emerged as iSNVs when the infection started and became consensus level variations in the secondary cases. The fast fixation of the minor iSNV T6226C suggested that the subsequent mutations might offer the virus certain selective advantages when the virus spreads rapidly. Interestingly, the T12655G variant was lost in Cluster D. The higher frequency donor iSNVs were not seemingly more convenient to successful transmission, or this data might suggest the possible occurrence of the purifying selection of T12655G in some circumstances, which was a transversion instead of a transition. Moreover, if a variation could not be fixed rapidly within a single host, the opportunity for the variation to be passed and maintained in the virus population might be reduced, which was likely the case for the T12655G variation, despite the fact that NJ38 had four recipients. We did not observe cotransmission of the two variants T6226C and T12655G simultaneously, which is consistent with a narrow bottleneck.

Our results from the single-source community outbreak further support the low within-host diversity of the SARS-CoV-2 genome reported in previous studies (Markov et al. 2023). Compared with a proofreading exoribonuclease, SARS-CoV-2 is more replicative, which results in slow mutational accumulation. Compared with the index case, the most distant sequence presented only four SNPs, which could be explained by the properties of SARS-CoV-2 described above, including its prompt responses and

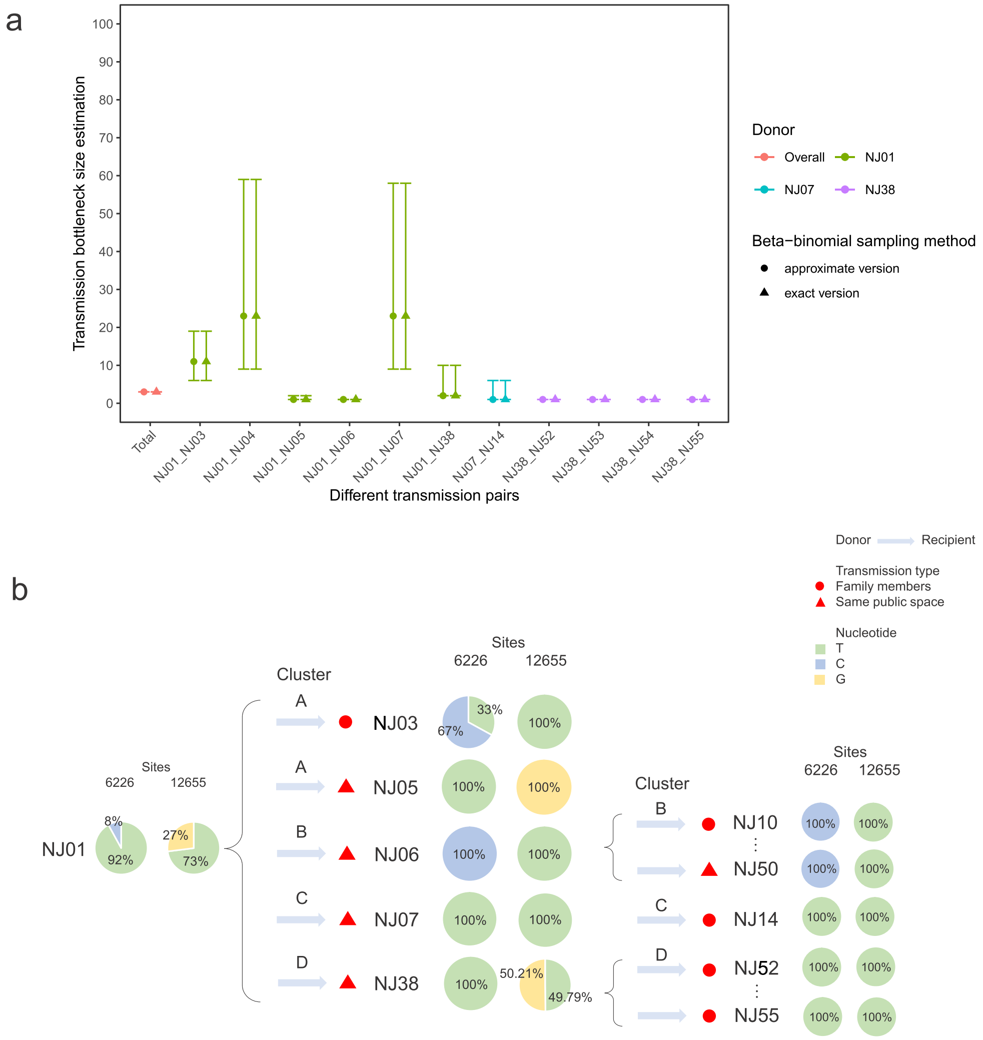

Figure 5. Estimated transmission bottleneck sizes and transmissions of intrahost variants. (a) Transmission bottleneck estimation via the approximate version and exact version of the beta-binomial sampling method. The bars show the means and 95% confidence intervals. (b) Minor iSNV transmission leads to diverse virus populations. The finding of limited shared intrahost viral diversity was demonstrated. Alternative alleles were observed to survive, transmit, and fix. The transmission of minor iSNVs explains some of the fixed substitutions observed in the virus population during the outbreak. The pie charts show the frequency of iSNVs. The arrows show the direction of transmission of the case pairs. Different colors represent different bases.

short duration (16 days) of transmission. A narrow bottleneck size might play a role in reducing the population size and viral genetic diversity during transmission. However, compared with the results of the 26-day outbreak caused by the Delta virus, which also presented at most 4 nucleotide substitutions from the index sample (Li et al. 2022), the level of viral diversity over time was likely greater in this outbreak caused by the Omicron virus.

The mutational properties of SARS-CoV-2 are influenced by host cell deamination by apolipoprotein B mRNA editing catalytic polypeptide-like proteins (APOBEC), adenosine deaminase acting on RNA proteins (ADAR), and oxidation by reactive oxygen species (ROS). Generally, APOBEC deamination of cytosine results in C > U, ADAR deamination of adenine drives an increase in A > G, and ROS oxidation of guanine leads to G > U (along with reciprocal G > A, U > C, and C > A mutations, respectively) (Azgari et al. 2021, Mourier et al. 2021). Our results indicated that among the nucleotide substitutions in the SNPs, transitions (A ↔ G or C ↔ T) were more prevalent than transversions. With respect to Wuhan-Hu-1, the proportion of the accumulated C > U substitution was the highest, which was consistent with previous reports (Sexton et al. 2023). However, the accumulated mutations in iSNVs across the genome might provide additional information regarding viral evolution and diversity. Xi et al. (2023) reported that the main substitution of iSNVs was a transition (C > U), and Armero et al. (2021) reported that the main substitution was a transversion (G > T). In our study, the substitution was also dominated by a transversion (T > A), but the types of transversions were inconsistent. San et al. (2021) reported that the most frequent iSNVs substitutions in the SARS-CoV-2 outbreak (CH1) in South Africa were A > G, C > U, U > C, and U > A. The most abundant mutational patterns of the iSNVs in the S gene in the report (San et al. 2021) were A > G, C > U, and U > A, whereas our data presented the prevailing patterns of U > A, A > G, and C > U in the S gene. These results suggest that there are differences in iSNV substitution patterns across different scenarios. The dominant iSNV substitution type was identified as a T > A transversion in our SARS-CoV-2 genomic data, but the reasons for these observations are currently unknown apart from the small sample size, which requires further study. Mutations, such as A > T and T > A transversions, are mediated through an as-yet-uncharacterized mechanism, suggesting the complexity of host effects on virus sequence changes (Giorgio et al. 2020, Simmonds and Schwemmle 2020). Moreover, the different mutational types between the population-level SNPs and within-host-level iSNVs further demonstrates that most iSNV mutations in a genetic pool tend to be unfixed in the process of evolution. The ratio of nonsynonymous to synonymous variants was 5.21 in iSNVs, which was completely different from that in the SNP mutations, which also supports these observations.

The size of the transmission bottleneck is a key factor in determining the possibility of the spread of a new within-host variation in a population (Zwart and Elena 2015). Braun et al. reported that during acute SARS-CoV-2 infection, diversity within the host is low, transmission bottlenecks are narrow, and in-host variation is rarely transmitted (Braun et al. 2021). Bendall et al. identified a per clade bottleneck of 1 for Alpha, Delta, and Omicron and 2 for non-VOCs, and that these tight bottlenecks reflect the low diversity at the time of transmission (Bendall et al. 2023). They are similar in size to the transmission bottleneck in this study. Estimates of the viral bottleneck size might be influenced by multiple factors, such as virus-specific differences, viral dynamics, routes of infection, molecular interactions at the virus–host interface, and the stochastic evolutionary processes (McCrone et al. 2018, Bendall et al. 2023).

The limitations of this study are presented in the Supplementary materials.

Conclusions

The identification of a single viral lineage among all sequenced samples in this outbreak suggested a single introduction of the Omicron BA.2.2 virus into the community, which was transmitted through community contact. The low level of genetic variation in this outbreak is further supported by an estimated stringent transmission bottleneck. The mutations might have initiated at the iSNV level, with most changes being nonsynonymous at a low frequency before a small fraction of the minor iSNVs could finally be fixed and identified as synonymous SNPs. In addition, transversions were more common than transitions during iSNV accumulation, whereas transitions were more prevalent than transversions among the SNP nucleotide substitutions, possibly implying the outcomes of purifying selection during the emergence of a new variant at the population level.

Acknowledgements

We gratefully acknowledge the efforts of local CDCs and hospitals in epidemiological investigations, sample collection, and screening. We thank Jiangsu Provincial CDC for providing guidance in dealing with the outbreak and sequencing services. We thank all the authors for submitting the consensus sequences to GISAID.

Author contributions

Conceived and designed the experiments: J.D., M.H., and N.Z. Performed the experiments: L.G.Z., H.X.W., N.W., W.Y., X.X.D., Z.Y.W., X.Q.D., X.Y.M., and H.B.Z. Analyzed the data: N.Z., M.H., H.X.W., N.W., H.F.F., S.N.D., T.M., and Z.Z. Contributed analysis tools: H.X.W. Wrote the manuscript: N.Z. and J.D. Reviewed and revised: J.D. and L.G.Z.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Conflict of interest: None declared.

Funding

This study was supported by the funds of Science and Technology Major Project of Nanjing Municipal City (No. 2021-11005), the Doctoral Program of Entrepreneurship and Innovation in Jiangsu Province (No. JSSCBS20211609), Nanjing Medical Science and Technology Development (No. YKK22193), and Nanjing Municipal Key Laboratory for Public Health Laboratory Technology (2023).

Data availability

The SARS-CoV-2 genome sequences in this study have been deposited in GenBank and the accession numbers were shown in Supplementary Table S1.

References

Aggarwal D, Warne B, Jahun AS *et al*. Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission. *Nat Commun* 2022;**13**:751.

Armero A, Berthet N, Avarre JC. Intra-host diversity of SARS-Cov-2 should not be neglected: case of the state of Victoria, Australia. *Viruses* 2021;**13**:133.

Azgari C, Kilinc Z, Turhan B *et al*. The mutation profile of SARS-CoV-2 is primarily shaped by the host antiviral defense. *Viruses* 2021;**13**:394.

Bendall EE, Callear AP, Getz A *et al*. Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. *Nat Commun* 2023;**14**:272.

Braun KM, Moreno GK, Wagner C *et al*. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog* 2021;**17**:e1009849.

Carabelli AM, Peacock TP, Thorne LG *et al*. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* 2023;**21**:162–77.

Chen S, Zhou Y, Chen Y *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–i890.

Danecek P, Bonfield JK, Liddle J *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:1–4.

Giorgio SD, Martignano F, Torcia MG *et al*. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* 2020;**6**:eabb5813.

Gu H, Quadeer AA, Krishnan P *et al*. Within-host genetic diversity of SARS-CoV-2 lineages in unvaccinated and vaccinated individuals. *Nat Commun* 2023;**14**:1793.

Gu H, Xie R, Adam DC *et al*. Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong. *Nat Commun* 2022;**13**:736.

Hannon WW, Roychoudhury P, Xie H *et al*. Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat. *Virus Evol* 2022;**8**:veac052.

Hong Q, Han WY, Li JW *et al*. Molecular basis of receptor binding and antibody neutralization of Omicron. *Nature* 2022;**604**:546–52.

Komissarov AB, Safina KR, Garushyants SK *et al*. Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. *Nat Commun* 2021;**12**:649.

Leonardl AS, Weissman DB, Greenbaum B *et al*. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *J Virol* 2017;**91**:e00171–17.

Li B, Deng A, Li K *et al*. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *Nat Commun* 2022;**13**:460.

Lu J, Plessis L, Liu Z *et al*. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* 2020;**181**:997–1003.

Lythgoe KA, Hall M, Ferretti L *et al*. SARS-CoV-2 within-host diversity and transmission. *Science* 2021;**372**:eabg0821.

MacCannell T, Batson J, Bonin B *et al*. Genomic epidemiology and transmission dynamics of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in congregate healthcare facilities in Santa Clara County, California. *Clin Infect Dis* 2022;**74**:829–35.

Markov PV, Ghafari M, Beer M *et al*. The evolution of SARS-CoV-2. *Nat Rev Microbiol* 2023;**21**:361–79.

McCrone JT, Woods RJ, Martin ET *et al*. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife* 2018;**7**:e35962.

Mourier T, Sadykov M, Carr MJ *et al*. Host-directed editing of the SARS-CoV-2 genome. *Biochem Biophys Res Commun* 2021;**538**:35–39.

Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**:1641–50.

San JE, Ngcapu S, Kanzi AM *et al*. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol* 2021;**7**:veab041.

Sexton NR, Cline PJ, Gallichotte EN *et al*. SARS-CoV-2 entry into and evolution within a skilled nursing facility. *Sci Rep* 2023;**13**:11657.

Simmonds P, Schwemmle M. Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other Coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories. *mSphere* 2020;**5**:e00408–20.

Wang D, Wang Y, Sun W *et al*. Population bottlenecks and intra-host evolution during human-to-human transmission of SARS-CoV-2. *Front Med Lausanne* 2021a;**8**:585358.

Wang Y, Wang D, Zhang L *et al*. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med* 2021b;**13**:30.

Xi B, Zeng X, Chen Z *et al*. SARS-CoV-2 within-host diversity of human hosts and its implications for viral immune evasion. *mBio* 2023;**14**:e0067923.

Zhao N, Zhou N, Fan HF *et al*. Mutations and phylogenetic analyses of SARS-CoV-2 among imported COVID-19 from abroad in Nanjing, China. *Front Microbiol* 2022;**13**:851323.

Zwart MP, Elena SF. Matters of size: genetic bottlenecks in virus infection and their potential impact on evolution. *Annu Rev Virol* 2015;**2**:161–79.

Virus Evolution, 2024, 10(1), veae085 , DOI: https://doi.org/10.1093/ve/veae085, Advance Access Publication 17 October 2024, Research Article