

# Limited genomic reconstruction of SARS-CoV-2 transmission history within local epidemiological clusters

Pilar Gallego-García,<sup>1</sup> Nair Varela,<sup>1,2</sup> Nuria Estévez-Gómez,<sup>1,2</sup> Loretta De Chiara,<sup>1,2</sup> Iria Fernández-Silva,<sup>3</sup> Diana Valverde,<sup>1,2,3</sup> Nicolae Sapoval,<sup>4,†</sup> Todd J. Treangen,<sup>4,‡</sup> Benito Regueiro,<sup>2,5,6</sup> Jorge Julio Cabrera-Alvargonzález,<sup>2,5</sup> Víctor del Campo,<sup>2,7</sup> Sonia Pérez,<sup>2,5,§</sup> and David Posada<sup>1,2,3,\*,¶</sup>

<sup>1</sup>CINBIO, Universidade de Vigo, Vigo 36310, Spain, <sup>2</sup>Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, <sup>3</sup>Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, Vigo 36310, Spain, <sup>4</sup>Department of Computer Science, Rice University, Houston, TX 77005, USA, <sup>5</sup>Department of Microbiology, Complejo Hospitalario Universitario de Vigo (CHUVI), Sergas, Vigo 36213, Spain, <sup>6</sup>Microbiology and Parasitology Department, Medicine and Odontology, Universidade de Santiago, Santiago de Compostela 15782, Spain and <sup>7</sup>Department of Preventive Medicine, Complejo Hospitalario Universitario de Vigo (CHUVI), Sergas, Vigo 36213, Spain

<sup>†</sup><https://orcid.org/0000-0002-0736-5075>

<sup>‡</sup><https://orcid.org/0000-0002-3760-564X>

<sup>§</sup><https://orcid.org/0000-0003-1734-1904>

<sup>¶</sup><https://orcid.org/0000-0003-1407-3406>

\*Corresponding author: E-mail: [dposada@uvigo.es](mailto:dposada@uvigo.es)

## Abstract

A detailed understanding of how and when severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission occurs is crucial for designing effective prevention measures. Other than contact tracing, genome sequencing provides information to help infer who infected whom. However, the effectiveness of the genomic approach in this context depends on both (high enough) mutation and (low enough) transmission rates. Today, the level of resolution that we can obtain when describing SARS-CoV-2 outbreaks using just genomic information alone remains unclear. In order to answer this question, we sequenced forty-nine SARS-CoV-2 patient samples from ten local clusters in NW Spain for which partial epidemiological information was available and inferred transmission history using genomic variants. Importantly, we obtained high-quality genomic data, sequencing each sample twice and using unique barcodes to exclude cross-sample contamination. Phylogenetic and cluster analyses showed that consensus genomes were generally sufficient to discriminate among independent transmission clusters. However, levels of intrahost variation were low, which prevented in most cases the unambiguous identification of direct transmission events. After filtering out recurrent variants across clusters, the genomic data were generally compatible with the epidemiological information but did not support specific transmission events over possible alternatives. We estimated the effective transmission bottleneck size to be one to two viral particles for sample pairs whose donor-recipient relationship was likely. Our analyses suggest that intrahost genomic variation in SARS-CoV-2 might be generally limited and that homoplasy and recurrent errors complicate identifying shared intrahost variants. Reliable reconstruction of direct SARS-CoV-2 transmission based solely on genomic data seems hindered by a slow mutation rate, potential convergent events, and technical artifacts. Detailed contact tracing seems essential in most cases to study SARS-CoV-2 transmission at high resolution.

**Key words:** intrahost variants; shared variants; transmission bottleneck; viral contagion; local outbreak; contact trace.

## 1. Introduction

In recent years, genomic epidemiology has revealed itself as a powerful tool for tracking viral outbreaks (Grubaugh et al. 2019b). Particularly for diseases with a high proportion of asymptomatic infections like coronavirus disease 2019 (COVID-19), the use of genomic information might be especially relevant to understand their dissemination. Several methods have been developed to reconstruct infectious disease outbreaks using genomic information (e.g. Didelot, Gardy, and Colijn 2014; Jombart et al. 2014; Worby et al. 2014a; Hall, Woolhouse, and Rambaut 2015,

2016; Lumby, Nene, and Illingworth 2018; Didelot et al. 2021). However, these strategies rely on pathogen genomes mutating rapidly between infected individuals (Campbell et al. 2018). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), responsible for the COVID-19 pandemic, has spread globally in a very short time. SARS-CoV-2 has a mutation rate in the order of  $1 \times 10^{-3}$  mutations per site per year (Koyama, Platt, and Parida 2020; van Dorp et al. 2020b). For MERS-CoV-2, in principle with a similar mutation rate, the prediction is that in most cases, the consensus sequences sampled from a transmission pair (donor and receptor)

will be identical, precluding a complete reconstruction of the outbreak (Campbell et al. 2018). As a counterpart, for SARS-CoV-1, with a mutation rate four times higher, we expect to see several mutations between transmission pairs, which considerably augments the power to resolve transmission history (Campbell et al. 2018).

These considerations are based on consensus sequences that represent the dominant viral lineage within a host. However, pathogens with high rates of evolution, such as RNA viruses, accumulate new mutations more or less rapidly as they replicate within the individuals they infect, generating intrahost genomic variation. The generation of this genomic diversity enables viral populations to evade host immune responses (Hensley et al. 2009; Henn et al. 2012; Parameswaran et al. 2017), alter disease severity (Vignuzzi et al. 2006), and adapt to changing environments (Stapleford et al. 2014; Stern et al. 2017). Notably, the study of the shared intrahost genomic variation among individuals can be critical for identifying contagion events and transmission clusters (Didelot, Gardy, and Colijn 2014; Worby, Lipsitch, and Hanage 2014b, 2017; Park et al. 2015). Moreover, it also allows for estimating the size of the founding pathogen population transmitted from the donor to the recipient host (i.e. the transmission bottleneck size) (Frise et al. 2016; Sobel Leonard, Weissman, and Greenbaum 2017). Several studies have already shown that intrahost genomic variation can be detected in most SARS-CoV-2 infections, generally at low levels, but with some variation among individuals (Kuipers et al. 2020; Seemann et al. 2020; Shen et al. 2020; Wölfel et al. 2020; Butler et al. 2021; Lythgoe et al. 2021; Tonkin-Hill et al. 2021; Valesano et al. 2021; Braun et al. 2021b; Wang et al. 2021b). Most SARS-CoV-2 intrahost mutations appear at low frequencies, often less than 5 per cent, are primarily under purifying selection and display particular biochemical signatures (Graudenzi et al. 2021; Sapoval et al. 2021; Tonkin-Hill et al. 2021; Wang et al. 2021b).

A key question is whether SARS-CoV-2 intrahost variation can be transmitted during contagion. The answer is not straightforward, as shared intrahost variants among unrelated individuals can also result from convergent evolution or mutational hotspots (Tonkin-Hill et al. 2021; Valesano et al. 2021). So far, a few studies have used shared genomic variants between putative donor-receptor pairs to infer narrow transmission bottlenecks of one to ten virions, in SARS-CoV-2 (Li et al. 2022; Lythgoe et al. 2021; San et al. 2021; Wang et al. 2021a). Limited genomic diversity can prevent the reconstruction of disease outbreaks (Campbell et al. 2018). While distinct SARS-CoV-2 transmission clusters might be identified using consensus sequences (Letizia et al. 2020; Popa et al. 2020; Seemann et al. 2020), its moderate mutation rate and rapid transmission might prevent the detailed reconstruction of the transmission events within these clusters (Tonkin-Hill et al. 2021). Leveraging intrahost variation, San et al. (2021) studied two nosocomial SARS-CoV-2 outbreaks, showing that potential donor-recipient pairs are supported in some cases but not in others by shared intrahost variants.

All in all, it is not clear whether the observed levels of inter and intrahost variation in SARS-CoV-2 and the apparently small size of the transmission bottleneck could limit our capability to reconstruct local SARS-CoV-2 outbreaks in detail using only genomic information. Intrahost mutations, typically at very low frequencies, are sensitive to methodological artifacts like sequencing errors (Turakhia et al. 2020; De Maio et al. 2020a; Kubik et al. 2021) and cross-sample contamination, and the occurrence of mutational hotspots can confound the identification of transmission events. Here, we wanted to assess our ability to

reconstruct putative transmission chains and to infer reliable transmission bottleneck sizes in SARS-CoV-2. For this, we obtained high-quality genomic data from ten independent epidemiological clusters, with two replicates per sample and with unique oligonucleotide spike-ins to detect potential contamination, leveraging both interhost and intrahost variants and *ad hoc* phylogenetic techniques. Our results confirm the low levels of intrahost variability and the small transmission bottleneck of SARS-CoV-2, suggesting that genomic data alone might not be sufficient to fully resolve direct SARS-CoV-2 transmissions, revealing the need for additional sources of information like detailed contact tracing.

## 2. Material and methods

### 2.1 Sample collection

According to the epidemiological records, we identified forty-nine patients infected with SARS-CoV-2 conforming ten independent transmission clusters originated in nursing homes, family households, and birthday parties in the city of Vigo, NW Spain (Fig. 1; Table S1). After that, we recovered the corresponding diagnostic nasopharyngeal exudates collected at the Vigo University Hospital Complex (CHUVI). This study was conducted under the approval of the Galician Drug Research Ethics Committee (CEIm-G code 2020-301).

### 2.2 Epidemiological information

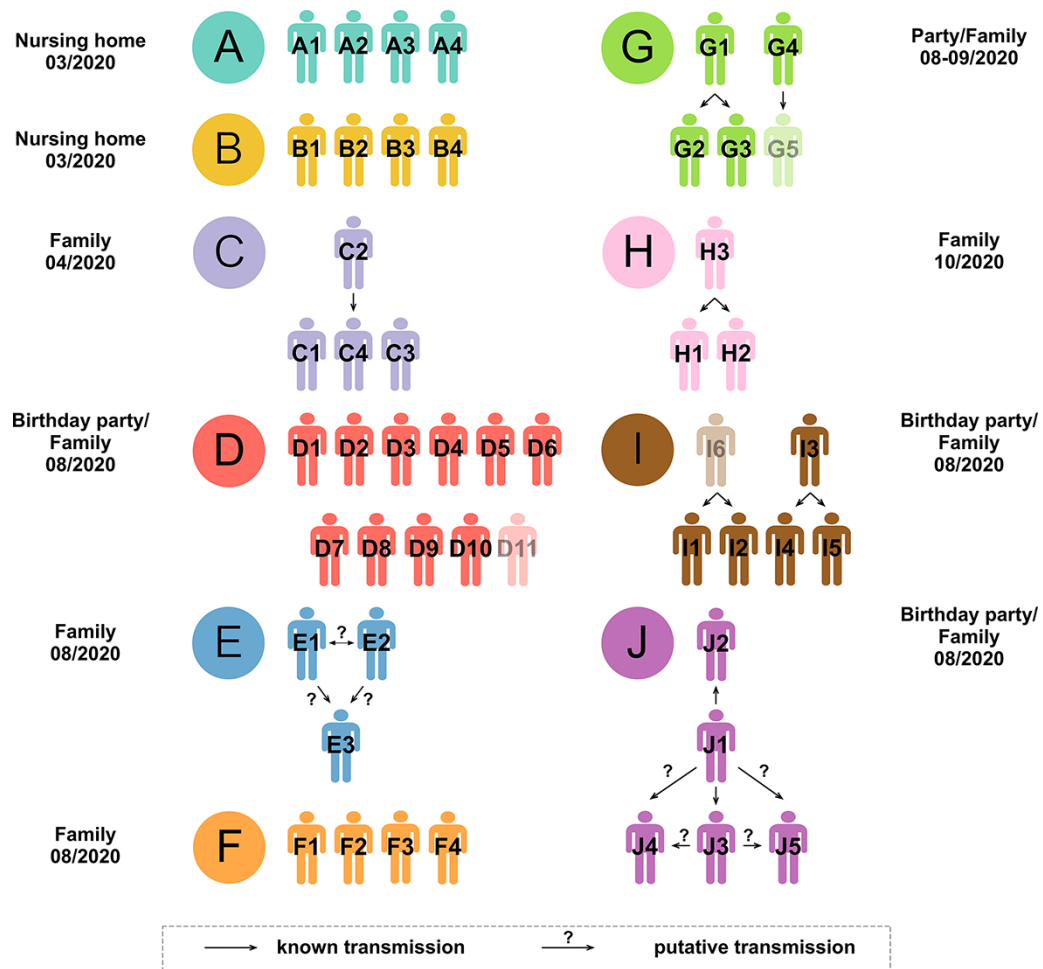
Clusters A and B belong to two different nursing homes, and in both cases, the primary case could not be established with confidence (Fig. 1). Cluster C is a family in which there was a probable transmission from C2 to C4. Cluster D is a large family spanning four different households. D1 came from another Spanish city and likely started the D transmission at a birthday party. Cluster E is a family in which brothers E1 and E2 were infected abroad before infecting their parent, E3. Cluster F is another family that was likely infected by an unsampled case from another city in Spain. Cluster G originates in two individuals (G1 and G4) that attended the same event and afterward infected their respective families, G1 to G2 and G3, and G4 to G5 (G5 failed at sequencing). Cluster H is another family in which H3 likely infected H1 and H2. Cluster I starts with two children (I6 and I3; I6 failed at sequencing) that got infected at the same birthday party before infecting their families, I6 to I1 and I2, and I3 to I4 and I5. Cluster J is a family in which J1 infected partner J2 and child J3. After that, either J1 or J3 infected J4 and J5.

### 2.3 RNA extraction

Following the manufacturer's recommendations, we extracted the viral RNA from the nasopharyngeal exudates using the MagNA Pure 24 Total NA Isolation kit (Roche Diagnostics, Basel, Switzerland). Different team members processed each RNA sample independently to obtain two technical replicates for each patient sample, from retrotranscription to library construction.

### 2.4 Viral load measurement

We measured SARS-CoV-2 genome copy concentration for each sample by real-time polymerase chain reaction (RT-PCR) of the E gene with the Sarbecovirus E-gene ModularDx (TIB Molbiol, Berlin, Germany) kit in a LightCycler® z480 System (Roche Molecular Systems Inc, Meylan, France). Viral load was estimated using linear regression ( $R^2 > 0.99$ ) from the standard curve generated with the Ct values obtained for serial dilutions (log) of RNA standards with known viral RNA genome equivalents/ $\mu$ l (Vogels et al. 2020).



**Figure 1.** Transmission clusters and epidemiological information. Black arrows indicate 'known' transmission events identified in the epidemiological records. Question marks highlight potential alternatives. Samples from patients in faded color failed at sequencing.

## 2.5 complementary DNA (cDNA) synthesis and multiplex amplification

We followed the ARTIC sequencing protocol (v.3) (Quick et al. 2017), a multiplex PCR-based target enrichment that produces 400-bp amplicons that span the SARS-CoV-2 genome, with slight modifications. First, we retrotranscribed the RNA samples to cDNA using the SuperScript IV reverse transcriptase (Invitrogen, MA, USA), starting with 10  $\mu$ l of RNA. Then we ran 30 PCR cycles for all the samples, independently of the Ct value, using the ARTIC primer Pool1 and Pool2 (IDT, CA, USA) and the Q5 Hot Start DNA polymerase (New England Biolabs, MA, USA). Next, we mixed the corresponding PCR products from each sample before cleaning (1.2:1 ratio beads to sample). We eluted the clean PCR products with 35  $\mu$ l nuclease-free water, recovered 33  $\mu$  and performed quantification with the Qubit 3.0 using the dsDNA HS or BR kit (Thermo Fisher Scientific, MA, USA), and checked amplicon size with the 2200 TapeStation D1000 kit (Agilent Technologies, CA, USA).

## 2.6 Addition of individual barcodes

We added 1  $\mu$ l of an X-mer single-stranded oligonucleotide with a unique barcode sequence at 38 fM to each retrotranscription reaction to detect potential sample cross-contamination. To prepare these barcode spike-ins, we used as a template the alcohol dehydrogenase 1 (*adh1*) mRNA (XM\_008650471.2) from *Zea mays*,

as described in the PrimalSeq v.4.0 protocol (Matteson et al. 2020). After a cleanup step (2:1 ratio beads to sample), we recovered a final volume of 22  $\mu$ l and performed QC (Qubit 3.0 and 2200 TapeStation). We added F and R primers with the same barcode sequence at the same concentration as the ARTIC primer pools to amplify the barcodes in the multiplex PCRs.

## 2.7 Library construction and genome sequencing

We built ninety-eight whole-genome sequencing libraries employing the DNA Prep (M) Tagmentation kit (Illumina, CA, USA) using  $\frac{1}{4}$  of the recommended volume, with approximately 125 ng of input DNA. Finally, we checked the size of the libraries and quantified them as described above. We sequenced the ninety-eight libraries in two high-output (7.5 Gb) runs (sixty and thirty-eight samples, respectively) on an Illumina MiniSeq (PE150 reads) at the sequencing facility of the University of Vigo.

## 2.8 Detection of potential cross-sample contamination

To assess the level of cross-sample contamination, we quantified the specific maize barcode content in each fastq file. For this, we aligned the raw reads against the *Zea mays adh1* sequence using BWA-mem (Li 2013) with default settings and demultiplexed the mapped reads with cutadapt (v.2.10) (Martin 2011), specifying a minimum overlap of 15 and a maximum error rate of 0.1. Then, we

counted the number of each type of reverse and forward barcodes present in each FASTQ file and checked if they were any different from the ones added to the corresponding sample.

## 2.9 Variant calling and consensus sequences

We assessed the quality of the fastq files using FastQC (Andrews 2010) and aligned the reads to the reference MN908947.3 from Wuhan using BWA-mem (Li 2013). We used iVar (Grubaugh et al. 2019a) to trim primer sequences and low-quality regions of the reads, as well as remove reads with less than 30 bp. We evaluated the quality of the aligned trimmed reads using Picard v2.21.8 (<http://broadinstitute.github.io/picard>; last accessed 20 December 2021). We used SAMtools depth v1.10 (Li et al. 2009) to calculate the sequencing coverage along the genome for each replicate. We only kept samples for which ten or more reads covered more than 75 per cent of the viral genome in the two replicates and with less than 2.5 per cent missing bases on the consensus sequence.

We used iVar (Grubaugh et al. 2019a) to identify single nucleotide changes and indels, with a minimum base quality threshold of 20 and a minimum read depth of 10. The calls obtained were confirmed with LoFreq (Wilm et al. 2012). We only retained variants that appeared in both replicates with a minimum overall variant allele frequency (VAF) of 2 per cent. Based on their frequency, we labeled the genomic changes detected as *fixed* ( $VAF \geq 0.98$ ; to account for potential sequencing errors) and *intrahost* variants ( $0.02 \leq VAF < 0.98$ ). We masked and removed from further analyses positions containing complex variants (i.e. nucleotide changes plus indels) or those deemed as homoplasic (De Maio et al. 2020b), including the sites immediately before and after.

To build a consensus sequence for each sample, we merged the reads from the two replicates with SAMtools *mpileup* and fed them to iVar *consensus* with a minimum VAF threshold of 0.5. We assigned the consensus sequences to a SARS-CoV-2 clade with Nextclade (<https://clades.nextstrain.org>; last accessed 20 December 2021) and to a SARS-CoV-2 PANGO lineage (Rambaut et al. 2020) with Pangolin (O’Toole et al. 2021).

## 2.10 Delimitation of epidemiological clusters

The simplest method for delimiting epidemiological clusters using genomic data alone is estimating a phylogenetic tree using the consensus sequences. For this, we aligned the consensus sequences with the reference using MAFFT v.7 (Kato and Standley 2013) (`mafft—maxiterate 500 <input>`) and ran IQ-TREE (v.2.0.6) (Nguyen et al. 2015) (`iqtree2 -T AUTO -s <alignment.fasta> -m TEST -b 1000 -o MN908947.3`) with the best-fit nucleotide substitution model and 1,000 bootstrap replicates. We also built a timetree based on the output tree of IQ-TREE and the dates of the samples using TreeTime (v.0.8.1) (Sagulenko, Puller, and Neher 2018) (`treetime—aln <alignment.fasta>—tree <treename>—dates <dates.csv>—max-iter 30`). In addition, we tried six heuristics developed explicitly for the reconstruction of epidemiological clusters described in Worby, Lipsitch, and Hanage (2017). The weighted distance tree and the minimum distance tree use the genetic distances among consensus sequences. On the other hand, the weighted and maximum variant tree strategies rely exclusively on shared intrahost variants, while the hybrid weighted tree and maximum tree procedures use intrahost variants and consensus genetic distances. Furthermore, we also estimated transmission clusters using the Transcluster algorithm (Stimson et al. 2019), assuming a mutation rate of  $1 \times 10^{-3}$  mutations/site/year. We explored four values for the transmission rate (10, 25, 50, 100 transmissions

per year) and six for the transmission cutoff (one to six transmission events). Finally, we also tried a probabilistic approach such as the one implemented in Phydely (Han et al. 2019), which leverages the inferred phylogenetic trees and can work without predefined genetic distance thresholds (`phydely.py—tree <treename>—collapse_zero_branch_length`).

## 2.11 Inference of transmission history

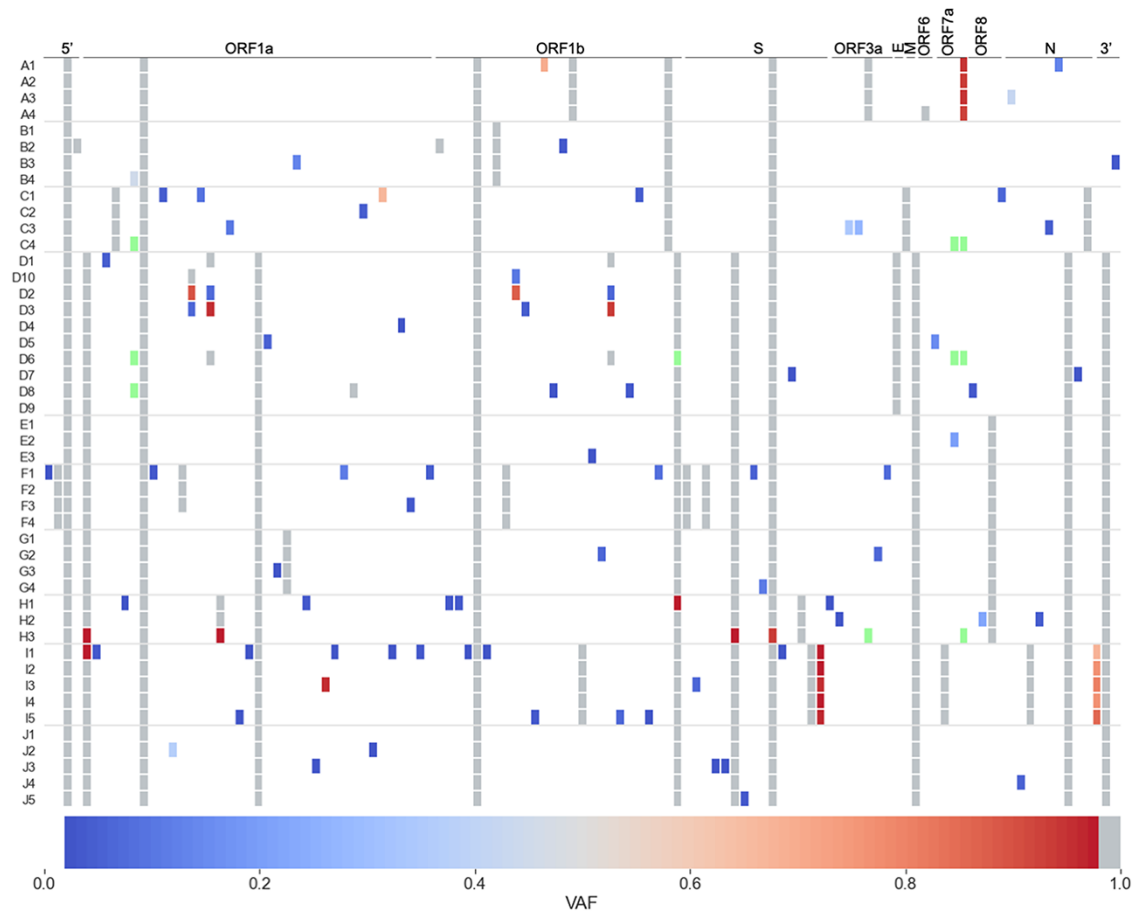
Within each cluster, we tried several approaches to estimate which individuals transmitted the virus and in which direction, that is, to learn who infected whom. First, we explored the Worby et al. heuristics, which assume that the donor for each sample has the most similar sequence or more shared intra-host variants. In addition, we implemented a simple approach that leverages the intrahost variation along a minimum spanning tree (MST). First, we computed Euclidean pairwise distances among all individuals within a cluster, with the *rdist* R package (<https://github.com/blasern/rdist>; last accessed 20 December 2021), using the VAF distributions. Afterward, we built the MSTs based on those distances with the function *mst* from the *ape* R package (Paradis and Schliep 2019). Then, assuming a single source for each cluster, we inferred the transmission direction that minimized the generation of novel variants in the receptor, meaning that in a pair of individuals, the donor should be the one with a higher number of private mutations. Finally, we also explored TransPhylo (Didelot et al. 2017), using the dated phylogeny obtained with TreeTime. We ran the algorithm for 150,000 Markov Chain Monte Carlo iterations and assumed a Gamma distribution for the generation time with shape 1 and scale 0.01917 (Perera et al. 2021).

## 2.12 Estimation of the transmission bottleneck size

To estimate the transmission bottleneck size of SARS-CoV-2 (i.e. the size of the viral population transferred from the donor to the recipient host), we used the beta-binomial method of Sobel Leonard, Weissman, and Greenbaum (2017). This method assumes that the intrahost variants detected did not arise *de novo* in different patients. This calculation includes only intrahost donor variants shared with the recipient (note that they can be fixed in the recipient but not in the donor). We identified putative donor–recipient pairs according to the available epidemiological information (Fig. 1). We lacked epidemiological information for clusters D and F, and we identified possible transmission pairs according to the genomic data (see Section 3). For the estimation of the transmission bottleneck size, we used the R code at [https://github.com/weissmanlab/BB\\_bottleneck](https://github.com/weissmanlab/BB_bottleneck) (last accessed 20 December 2021), under the *approximate* model (given that the sequencing depth per sample was very high, around 6,000X) and setting the maximum bottleneck size to an arbitrarily large value of 600, and the VAF cutoff to 0.02.

## 2.13 Assessment of selective pressures

The ratio of non-synonymous changes per non-synonymous site (*dN*) to the number of synonymous substitutions per synonymous site (*dS*) is one of the most popular statistics for detecting selective pressures at the molecular level. We estimated the *dN/dS* ratio for each sample using the *dNdScv* package (Martincorena et al. 2017), recently adapted for its application to SARS-CoV-2 (Tonkin-Hill et al. 2021). We used the default substitution model with 192 rate parameters.



**Figure 2.** VAFs per sample. VAFs were calculated after filtering recurrent variants. Fixed mutations ( $\text{VAF} \geq 0.98$ ) are in gray, fixed reference alleles ( $\text{VAF} < 0.02$ ) are in white, and positions with missing data (depth below 20) are in light green.

### 3. Results

#### 3.1 Viral load and sequencing

Twenty-seven out of the forty-nine samples had a viral load above  $10^3$  copies/ $\mu\text{l}$  (Table S1). Sequencing coverage and breadth were high (mean depth  $\pm$  sd:  $6316.71 \pm 2336.99$ ; breadth: 0.999) (Table S2), except for three samples (D11, G5, and I6, all with a Ct  $> 32$  for gene E), that we excluded from further analyses. We did not detect appreciable cross-contamination between samples (Table S2).

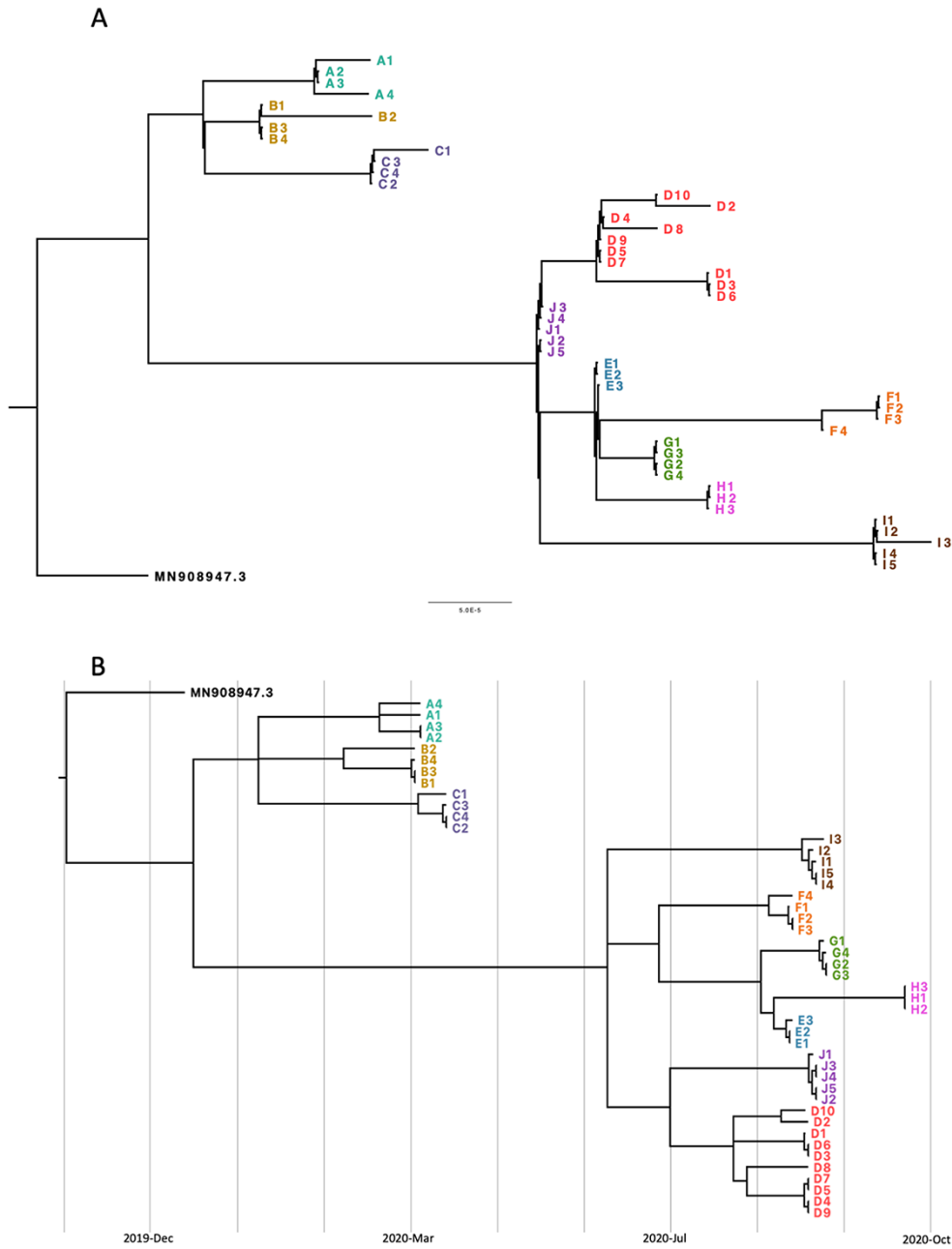
#### 3.2 Inter and intrahost variation

Most variants were fixed ( $\text{VAF} \geq 0.98$ ) (Fig. 2, Figure S1). The number of differences among consensus sequences was, on average, 2.12, 2.28, and 7.57 variants, within early clusters (A–C), late clusters (D–J), and among early and late clusters (see also Table S3). We observed on average 19.76 variants per sample, of which 8.17 were intrahost (Table S4). Both fixed and intrahost variants were shared among samples at different VAFs. Several intrahost variants appeared recurrently in multiple samples, often corresponding to indels at low frequency (Table S4). These recurrent variants may correspond to potential sequencing errors and mutational hotspots, which might confound our analyses. Therefore, we decided to filter out intrahost variants present in more than one cluster. After filtering, there were 2.13 intrahost variants per sample on average, with a maximum of 11 (Table S4). Before and after filtering, the number of intrahost variants detected per sample was unrelated to sequencing depth, Ct values, or viral load

(Figure S2). Furthermore, VAFs between sample replicates were significantly correlated (Pearson correlation coefficient = 0.99,  $P$ -value =  $5.6 \times 10^{-113}$ ) (Figure S3). All samples were assigned to two related clades/lineages (20A/B.1 and 20E(EU1)/B.1.177), which was not particularly surprising as these were the dominant lineages in the area at the time of sampling. We estimated  $dN/dS$  values for missense variants consistently below 1, suggesting a predominance of intrahost purifying selection across samples (Figure S4).

#### 3.3 Delimitation of transmission clusters

The maximum likelihood (ML) trees obtained with the consensus sequences showed the epidemiological clusters as distinct groups, mostly monophyletic and with relatively high bootstrap support (Fig. 3). Remarkably, adding the temporal information with Tree-Time improves the phylogenetic resolution of the clusters, which become all monophyletic (Fig. 3B). However, standard phylogenetic approaches do not explicitly inform about the number of clusters or the assignment of the different individuals to clusters. In the absence of additional epidemiological information (i.e. colors in our trees), researchers often infer putative transmission clusters using some kind of distance threshold. The weighted distance tree and the minimum distance tree, which use consensus sequences to explicitly delimit clusters, were identical and highly congruent with the epidemiological information (Fig. 4A). In this case, the only ‘error’ was that cluster D was divided into two, although we might expect it because D1, D3, and D6 share two



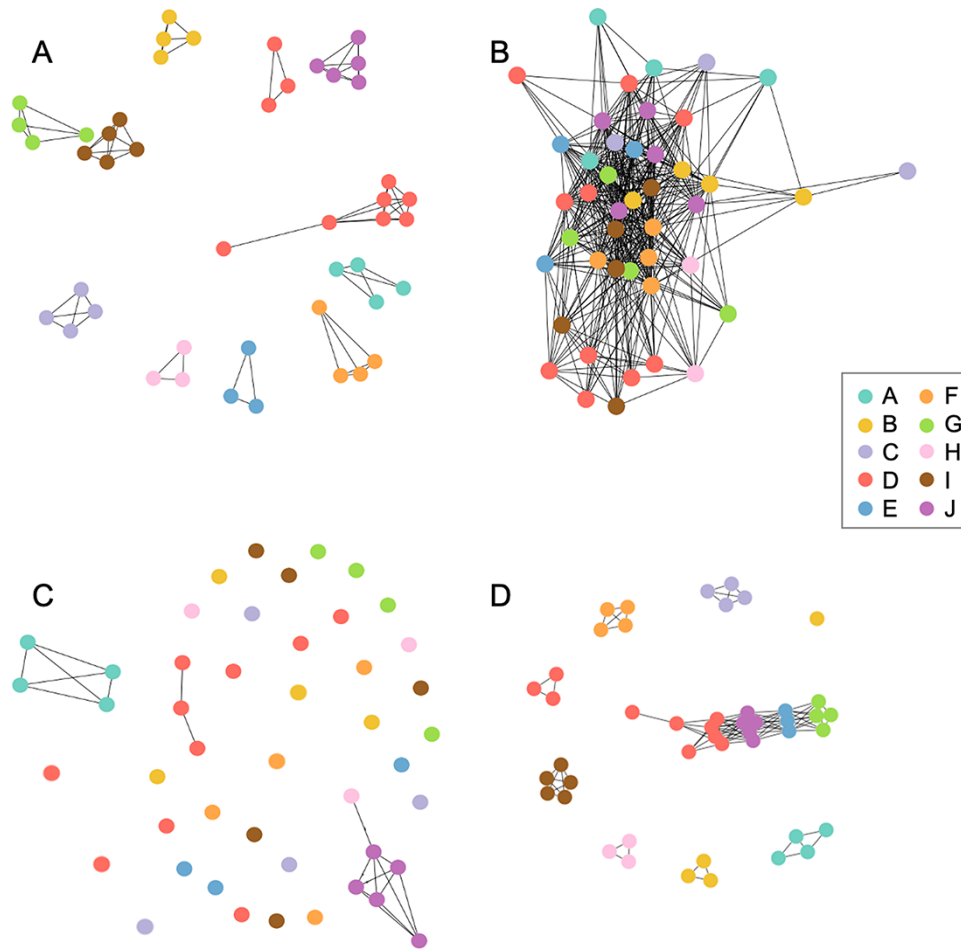
**Figure 3.** Consensus-sequence phylogenetic trees. (A) Maximum likelihood phylogenetic tree inferred with IQ-TREE. Numbers above branches are bootstraps values (%). Only bootstrap values above 50 are shown. (B) Time-scaled ML tree inferred with TreeTime using the dates of extraction.

consensus mutations that the rest of the D individuals do not present. Indeed, cluster D is large and phylogenetically diverse, and we might not have sampled all the infected individuals in this transmission chain.

The weighted variant and maximum variant trees, based exclusively on intrahost variants, were also identical and produced a very complex network in which all individuals seemed related to each other (Fig. 4B). After removing the recurrent intrahost variants common to multiple individuals and clusters (taking advantage of the epidemiological information), these methods identified three clusters primarily compatible with the epidemiological information, plus thirty-three unconnected individuals (Fig. 4C). Cluster A was perfectly delimited, while cluster I formed a group with a sample from cluster H. The only other three

clusterized samples were from cluster D (again D1, D3, and D6). The hybrid transmission methods, which use the connections established by the weighted variant and maximum variant trees and incorporate consensus information for those samples without a donor or recipient, did not result in any noticeable improvement compared to methods based on consensus sequences (data not shown). Finally, the transmission-based clustering method in Transcluster was able to identify some of the epidemiological clusters but not all (Fig. 4D). In this case, congruence with the epidemiological data was maximal after setting a transmission rate of 25 and a transmission cutoff of 1.

In the case of Phydelyty, which tries to delimitate transmission clusters upon a given phylogeny, the inferred clusters were often subsets of the epidemiological clusters, both when using



**Figure 4.** Clustering approaches. (A) Weighted distance/minimum distance tree. (B) Weighted variant/maximum variant tree after standard masking. (C) Weighted variant/maximum variant tree after removing recurrent low-frequency variants (D) Transcluster transmission network (transmission rate = 25, transmission cutoff = 1).

the ML (Figure S5A) and the TreeTime tree (Figure S5B). For example, Phylodelty divided cluster D into two subclusters (one of them composed again by D1, D3, and D6) and three other samples were left unassigned in both trees, while for most of the other clusters (A, B, C, E, F, and I) there were one or more unassigned samples. Phylodelty precisely identified clusters G and J when using the ML tree, while H is the only cluster that was correctly delimited in both the ML and the TreeTime tree.

### 3.4 Inference of transmission history

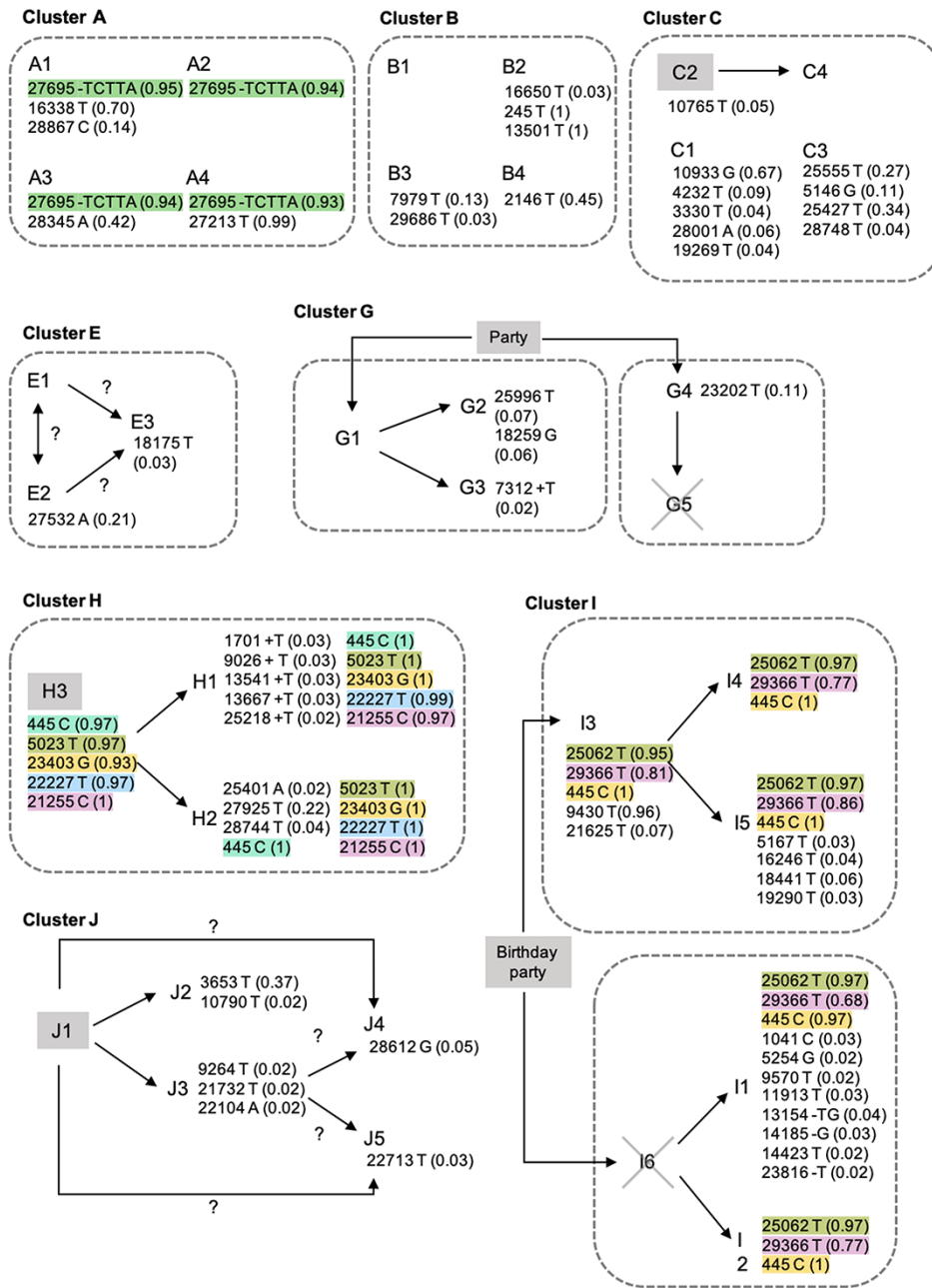
#### 3.4.1 Transmission in nursing homes

For clusters A and B, we had no epidemiological information other than the corresponding nursing home. In cluster A (Fig. 5), the four samples share what seems to be an intrahost variant (27695-TCTTA). However, given its high VAF (0.93–0.95) and the fact that the individuals do not share other variants, this deletion may be a truly fixed variant, where sequencing or calling errors prevented its identification in all reads. In any case, the genetic data does not help identify the different transmission events in this cluster with confidence. In cluster B, no shared variation was apparent. B2 has two private apparently fixed variants, suggesting it was infected later than the other cluster members or from a different source. Again, it was not possible to infer the transmission network for this cluster.

#### 3.4.2 Transmission in clusters with partial contact trace information

We had partial contact trace information for clusters C, E, G, H, I, and J. However, the lack of shared intrahost variants prevented a detailed reconstruction of their transmission history in most cases (Fig. 5). The epidemiological record suggests a transmission from C2, the index case, to C4 in cluster C. This event is compatible with the genetic data, as C2 has a single intrahost variant at low frequency (0.05), which could have been lost during transmission to C4, which has no intrahost variants. Private variants with low VAFs in C1 and C3 could have arisen *de novo* within each individual after transmission, but three of them with higher VAFs (0.27, 0.34, and 0.67) are more difficult to explain in the same way. In cluster E, the genetic information cannot resolve whether E1 or E2 infected E3. In cluster G, we did not observe shared intrahost variants. In contrast, the distribution of the private variants is compatible with the epidemiological information, and it does not help resolve further the transmission history.

In cluster H, the three samples share five fixed (VAF  $\geq 0.98$ ), or almost fixed, variants. The index case (H3) does not seem to have intrahost variants, contrary to H1 and H2. However, the quality of the sequencing data in the case of H3 is well below average, so it is possible that low-frequency variants were overlooked in this sample. All five members of cluster I share three variants with high VAF—or fixed in several cases. Variants 445C and 25062 could be



**Figure 5.** Variant sharing within clusters. Gray boxes indicate index cases or originating events. Gray dashed lines delimit households (nursery homes for clusters A and B). Shared variants are highlighted with the same color. Fixed variants (VAF  $\geq 0.98$ ) common to all members of a cluster are not shown. Crossed samples could not be sequenced.

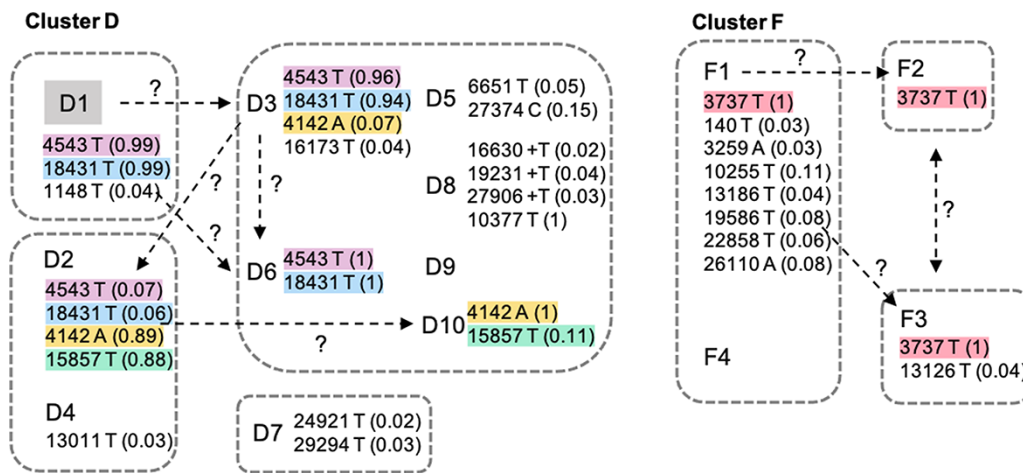
genuinely fixed in all samples, including cases where the apparent VAF is 0.95–0.97. The distribution of variant 29366T is remarkable, as it appears in all cases with a VAF of 0.68–0.86. Another salient observation is that I3, one of the index cases, has a well-supported variant (9430T) with a VAF of 0.96 that does not appear in the other samples from this cluster. Cluster J lacked shared intrahost variants, so the genetic data neither confirmed nor invalidated the contact tracing information.

### 3.4.3 Inferring transmission in the absence of contact trace information

**3.4.3.1 Ad hoc approaches.** We did not have detailed information about contacts in clusters D and F, so we tried to identify

transmission events considering just the genomic data (Fig. 6). In cluster D, the transmission started at a birthday party where the index case was D1. D1 shares two variants with D3 and D6 (4543T and 18431T), both fixed (VAF  $\geq 0.98$ ) in D1 and D6 and close to fixation in D3 (0.96 and 0.94, respectively). Therefore, we hypothesize that D1  $\rightarrow$  D3 and D1  $\rightarrow$  D6, but alternatively D3  $\rightarrow$  D6, could be transmission pairs. These two variants also appear in D2 but at a very low VAF (0.07 and 0.06, respectively). D3 and D2 further share 4142 A, but this variant has a low VAF in D3 (0.07) and a high VAF (0.89) in D2. Furthermore, D2 has 15857T at high VAF (0.88). Given that we assumed that D1 infected D3, we considered that D3 could have infected D2. However, the explanation for the observed VAF patterns might imply recombination and *de novo* mutation.





**Figure 6.** Shared variants and inferred transmission events for clusters D and F. Below each sample ID, we show variant site and allele, followed by its VAF in parenthesis. Fixed variants (VAF  $\geq 0.98$ ) shared by all members of a cluster are not shown. Dashed arrows indicate putative transmission events. Question marks highlight potential alternatives.

Finally, D10 shares with D2 variants 4142A and 15857T at high frequency, so we also considered D2  $\rightarrow$  D10 another likely transmission pair. In cluster F, where we do not have an index case, F1 and F3 share a fixed variant (3737 T). Given that F1 has seven private variants at low frequency, but F3 only two, F1 might be the pair's donor because it seems easier to lose these variants during the F1  $\rightarrow$  F3 bottleneck than to arise *de novo* in F1 after an F3  $\rightarrow$  F1 transmission. F2 also has 3737T fixed. Following the same logic, F2 could have been infected by F1, but also by F3.

**3.4.3.2 Statistical and graphical approaches.** The Worby et al. approaches were not as helpful in inferring transmission as for delimiting clusters. Assigning the source of each sample to the patient with the highest number of shared intrahost variants or the minimum genetic distance (using weights or absolute values) resulted in samples with multiple potential sources and pairs with bidirectional transmission (Fig. 4A, B). Relying only on intrahost shared variants proved inefficient, as most samples were not connected to any other (Fig. 4C). Consensus sequences from the same cluster were very similar, so multiple samples were often equidistant, preventing choosing one of them as the source. The MST analysis (Figure S6) was incompatible with the epidemiological information. Apart from tied transmission paths for some of the clusters (clusters D and E, with three and two options, respectively), the starting point of the transmission did not coincide with the epidemiological information in any of the cases. TransPhylo could differentiate the different clusters (Fig. 7, Figure S7); however, the inferred direct transmission events within clusters were often not compatible with the epidemiological information.

### 3.5 Transmission bottleneck size

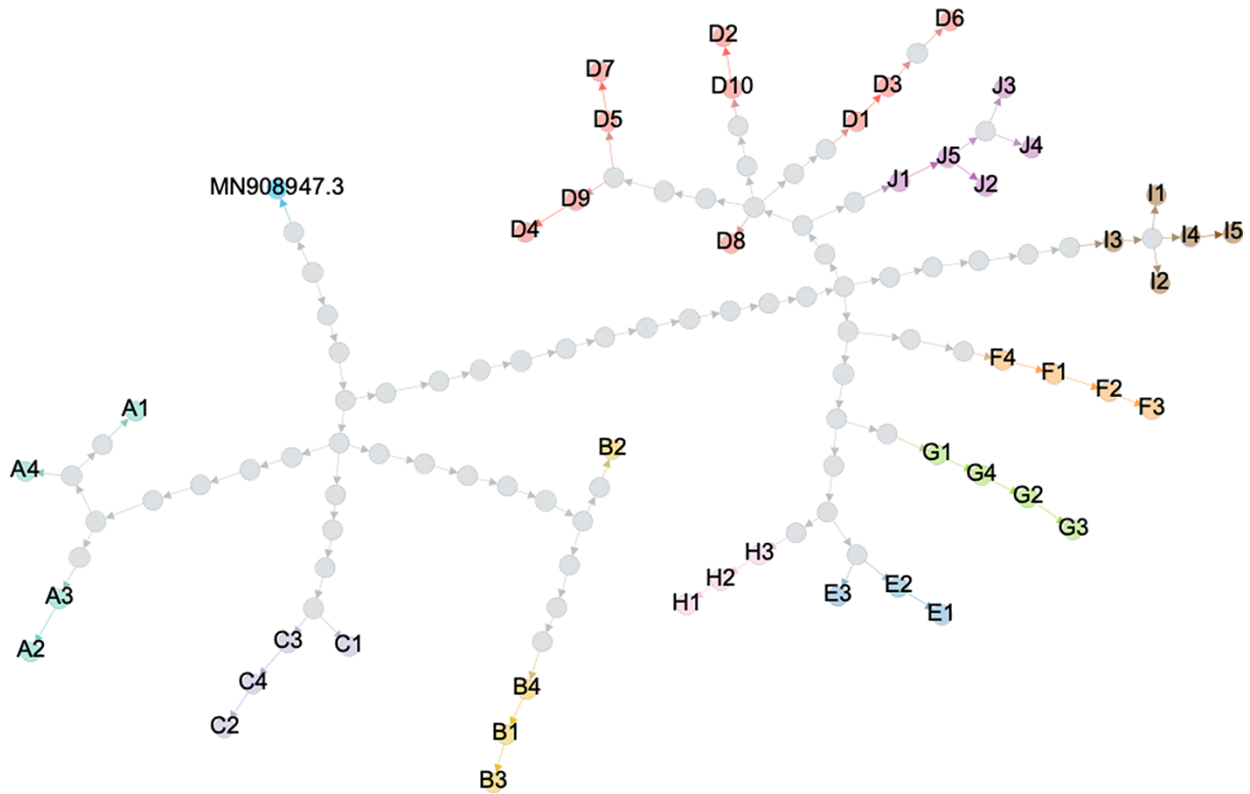
We selected individual pairs representing direct transmission events to estimate the transmission bottleneck size according to the epidemiological and genomic information. We discarded clusters A and B (nursing homes) from this analysis because it was impossible to identify likely transmission pairs in these cases. We had contact information about at least a transmission pair for clusters C, E, G, H, I, and J. The situation was more complex for clusters D and F, so we identified possible transmission pairs considering both the epidemiological and the genomic data, as described above.

Across the studied transmission pairs, we found an average of 0.38 (range 0–3) shared intrahost variants (Table S6). Accordingly, the estimated transmission bottleneck sizes were typically small (one to two viral particles) (Fig. 8, Table S6). To ensure that our selection of transmission pairs in clusters D and F was not biasing these estimates downwards, we also calculated the transmission bottleneck sizes for all potential pairs within these two clusters. The estimated bottlenecks were consistently one to two. Note that the bottleneck size can only be estimated when there is at least one variant in the donor (regardless of whether that variant is observed in the recipient). If none of the donor variants appear in the recipient, the estimated bottleneck size will be one, with a variable confidence interval depending on the variant calling threshold.

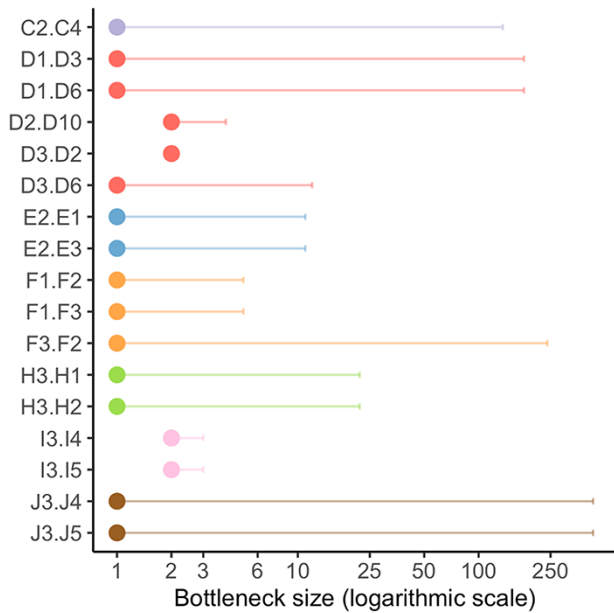
## 4. Discussion

Understanding SARS-CoV-2 transmission is crucial to identify which situations minimize or maximize the risk of infection and, therefore, implement more effective control strategies. Former studies have tried to reconstruct SARS-CoV-2 local transmission chains with more or less success using a combination of epidemiological and genomic data (Popa et al. 2020; Sekizuka et al. 2020; Shen et al. 2020; Hamilton et al. 2021; San et al. 2021). However, it is unclear whether, in situations for which contact tracing information is limited, we can use SARS-CoV-2 genomic information alone to understand who infected whom. Here, we show that while SARS-CoV-2 genomic variation can be helpful, at least in some cases, to delimit distinct transmission clusters, it is not enough to resolve with confidence transmission chains and direct transmission events. Using the most likely transmission pairs, we infer a narrow effective transmission bottleneck size for SARS-CoV-2 in the order of one to ten viral particles.

Unlike most of the previous studies of SARS-CoV-2 intrahost variation, we made use of substantial laboratory and bioinformatic controls to stress variant calling reproducibility, including the use of unique barcodes to discard cross-sample contamination, sequencing replicates, high sequencing depth, multiple variant callers, and curation of recurrent variants. While a few studies have included some of these controls to a different extent (Lythgoe et al. 2021; Tonkin-Hill et al. 2021; Braun et al. 2021b), this is the first study to consider contamination, replicates, and multiple



**Figure 7.** TransPhylo transmission graph. Gephi (Bastian, Heymann, and Jacomy 2009) depiction of TransPhylo’s consensus transmission tree. Gray dots represent inferred unsampled individuals.



**Figure 8.** Estimated transmission bottleneck sizes. Labels on the Y-axis represent donor–recipient pairs. Estimates were obtained with the beta-binomial ML method (Sobel Leonard, Weissman, and Greenbaum 2017). Horizontal lines represent 95 per cent confidence intervals. The X-axis is on a logarithmic scale.

variant calling strategies directly for each of the samples studied, plus the use of contact information. We found a limited number of intrahost variants (~8 before filtering recurrent variants and ~3

after filtering), as reported in earlier studies (Kuipers et al. 2020; Seemann et al. 2020; Shen et al. 2020; Wölfel et al. 2020; Butler et al. 2021; Tonkin-Hill et al. 2021; Valesano et al. 2021; Braun et al. 2021b; Wang et al. 2021b). Half of our samples (twenty-seven/forty-eight) had a viral load above  $10^3$  copies/ $\mu$ l, which is the threshold determined in Valesano et al. (2021) for reliable identification of intrahost variants with a VAF  $\geq 2$  per cent in single replicates.

Another novel aspect of this work is the explicit exploration of established methods that use viral interhost and/or intra-host genomic variation to delimitate transmission clusters and to identify transmission chains and direct contagions, and the comparison of these results with the available epidemiological information. In our samples, all from the same city and corresponding to two time points, the level of interhost genomic variation was generally low. However, this did not prevent the distinction among local clusters. When the sampling dates were taken into account, the concordance between genomic and epidemiological clusters was maximized, highlighting the relevance of the temporal information. Methods for cluster delimitation that rely exclusively on intrahost variants did not work well in this regard. In contrast, methods based on differences at the consensus level could differentiate the clusters near perfectly. These results suggest that in SARS-CoV-2, consensus sequences might be enough in some cases, to separate samples belonging to different clusters from the same area. But caution regarding sampling is always necessary. There might be unsampled individuals that do not belong to any of these clusters but that might have identical consensus sequences. At the same time, intra-host SARS-CoV-2 diversity does not seem sufficient for cluster delimitation.

Here, transmission history within nursing homes or households, where most SARS-CoV-2 infections occur (Lee et al. 2020), was complicated to decipher. In general, all the methods we tried, even those relying on intrahost variation, could not identify clear transmission patterns within clusters, as seen before in care homes (Hamilton et al. 2021) or in nosocomial outbreaks (Abbas et al. 2021). This poor resolution can be explained by a lack of genetic variation but also by homoplasy, as we observed several shared intrahost variants among apparently unrelated samples. In addition, we noticed that if VAF thresholds are relaxed, unique or uncommon mutations appear in many individuals, suggesting these variants are either recurrent artifacts or hotspot mutations (Tonkin-Hill et al. 2021). Deciding which shared intrahost variants in SARS-CoV-2 are the result of transmission events is not easy. Much care should be taken regarding reliable genotyping and identifying recurrent events, particularly for samples with low viral loads (van Dorp et al. 2020a,b; Kubik et al. 2021; Valesano et al. 2021; Braun et al. 2021b). Recurrent, low-frequency insertions in SARS-CoV-2 have already been detected elsewhere (Kuipers et al. 2020; Rayko and Komissarov 2020; Turakhia et al. 2020; Tonkin-Hill et al. 2021).

Although not addressed in this study, another potential complication regarding the identification of shared intrahost variants is the occurrence of significant intrahost evolution. Several studies have reported VAF changes within days in SARS-CoV-2 (Jary et al. 2020; Tonkin-Hill et al. 2021; Voloch et al. 2021; Wang et al. 2021b), even faster in immunocompromised individuals (Avanzato et al. 2020; Kemp et al. 2021). On the other hand, more rigorous studies report that diversity does not increase over time, although this does not imply that VAFs cannot change significantly among different time points (Valesano et al. 2021). Significant intrahost evolution would imply that the amount of sharing between samples could change depending on the exact sampling dates, so the inferences derived from it.

Consistent with previous studies, our estimates indicate that the SARS-CoV-2 transmission bottleneck size is small or very small (Gu et al. 2022; Li et al. 2022; Lythgoe et al. 2021; San et al. 2021; Wang et al. 2021a; Braun et al. 2021b), with only a few viral particles being responsible for the successful growth within the recipient. Notably, tight bottleneck estimates have also been obtained for a highly transmissible SARS-CoV-2 lineage like Delta (Li et al. 2021). In contrast, Popa et al. (2020) estimated an average transmission bottleneck size for SARS-CoV-2 of 1,000, but these estimates have been recently revised because of the inclusion of suspicious, highly shared intrahost variants (Martin and Koelle 2021; Nicholson et al. 2021). A strong transmission bottleneck reduces, in general, the efficacy of selection, impeding the contribution of intrahost diversity to viral adaptation at a global scale (McCrone and Luring 2018), although exceptions to this rule exist (Zwart and Elena 2015; Braun et al. 2021a). Still, interhost competition among SARS-CoV-2 variants can maintain and increase viral fitness. On the other hand, a small transmission bottleneck size for SARS-CoV-2 is compatible (but not proof of) with a dominance of aerosol transmission over direct contact, as seen in influenza (Varble et al. 2014; Frise et al. 2016; McCrone and Luring 2018). Further studies are necessary to elucidate whether this is really the case for SARS-CoV-2.

If only one or a few unique virions are passed during transmission, then most of SARS-CoV-2 intrahost variation has to be due to the accumulation of *de novo* mutations (Valesano et al. 2021; Voloch et al. 2021). These *de novo* mutations seem to be mainly deleterious. We inferred strong intrahost purifying

selection across the genome for missense variants, as in prior studies (Shen et al. 2020; Lythgoe et al. 2021; Tonkin-Hill et al. 2021).

Our results suggest that SARS-CoV-2 genomic diversity is helpful to delimitate different transmission clusters within a relatively small area, but that could be insufficient to fully resolve transmissions within a household or in the same social event. In other words, genomics alone cannot help identify who infected whom—but might discard putative contagions. Thus, contact tracing data will be essential to study direct SARS-CoV-2 transmission events, as it occurs in typical slow-evolving pathogens (Campbell et al. 2018, 2019).

Overall, the biological picture that has become apparent after this and preceding studies is that SARS-CoV-2 intrahost variation is low and mainly determined by genetic drift and purifying selection. The transmission bottleneck is very narrow, with only a few virions effectively contributing to the genomic diversity of the infection, so intrahost variants are infrequently transmitted from one host to another. Under this scenario, only a minority of infections, typically prolonged ones, should lead to the appearance of novel variants. Therefore, managing long-term SARS-CoV-2 infections should become a priority.

## Data availability

Raw FASTQ files have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (Leinonen, Sugawara, and Shumway 2011) (Project Accession No. PRJNA778931). Viral consensus genomes are available at the Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley 2017) (accession numbers in Supplementary Table S7).

## Supplementary Data

Supplementary data is available at *Virus Evolution* online.

## Funding

This project was funded by grant EPICOVIGAL FONDO SUPERA-COVID19 from Banco Santander, Consejo Superior de Investigaciones Científicas (CSIC) and Conferencia de Rectores de las Universidades Españolas (CRUE), grant CT850A-2 from the Axencia de Coñecemento en Saúde (ACIS) of the Servizo Galego de Saúde (SERGAS) from the Consellería de Sanidade Xunta de Galicia, and grant ED431C2018/54-GRC from the Consellería de Cultura, Educación e Ordenación Universitaria of Xunta de Galicia. NS and TT were supported in part by a C3.ai Digital Transformation Institute award. Funding for open access charge: Universidade de Vigo/CISUG.

**Conflict of interest:** None declared.

## Author contributions

D.P. conceived the study and designed the analyses. S.P., J.J.C., V.d.C., and B.R. obtained the patient samples and the epidemiological information. S.P., N.E.G., L.d.C., I.F.S., and D.V. planned and performed the laboratory experiments. P.G.G., N.V., N.S., and T.T. carried out the bioinformatic analyses. D.P. wrote the draft manuscript with the help of P.G.G. and N.V. All authors read the manuscript and contributed to interpretation and discussion.

## References

- Abbas, M. et al. (2021) 'Explosive Nosocomial Outbreak of SARS-CoV-2 in a Rehabilitation Clinic: The Limits of Genomics for Outbreak Reconstruction', *The Journal of Hospital Infection*, 117: 124–34.
- Andrews, S. (2010) 'FastQC: A Quality Control Tool for High Throughput Sequence Data'. [Online].
- Avanzato, V. A. et al. (2020) 'Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer', *Cell*, 183: 1901–12.e9.
- Bastian, M., Heymann, S., and Jacomy, M. (2009) 'Gephi: An Open Source Software for Exploring and Manipulating Networks', *Third International AAAI Conference on Weblogs and Social Media*.
- Braun, K. M. et al. (2021a) 'Transmission of SARS-CoV-2 in Domestic Cats Imposes a Narrow Bottleneck', *PLoS Pathogens*, 17: e1009373.
- et al. (2021b) 'Acute SARS-CoV-2 Infections Harbor Limited Within-host Diversity and Transmit via Tight Transmission Bottlenecks', *PLoS Pathogens*, 17: e1009849.
- Butler, D. et al. (2021) 'Shotgun Transcriptome, Spatial Omics, and Isothermal Profiling of SARS-CoV-2 Infection Reveals Unique Host Responses, Viral Diversification, and Drug Interactions', *Nature Communications*, 12: 1660.
- Campbell, F. et al. (2019) 'Bayesian Inference of Transmission Chains Using Timing of Symptoms, Pathogen Genomes and Contact Data', *PLoS Computational Biology*, 15: e1006930.
- et al. (2018) 'When Are Pathogen Genome Sequences Informative of Transmission Events?' *PLoS Pathogens*, 14: e1006885.
- De Maio, N. et al. (2020a), *Issues with SARS-CoV-2 Sequencing Data*. <<http://virological.org>> <<http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>> accessed 20 December 2021.
- et al. (2020b), *Masking Strategies for SARS-CoV-2 Alignments*. <<https://virological.org>> <<https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480>> accessed 12 May 2021.
- Didelot, X. et al. (2017) 'Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks', *Molecular Biology and Evolution*, 34: 997–1007.
- Didelot, X., Gardy, J., and Colijn, C. (2014) 'Bayesian Inference of Infectious Disease Transmission from Whole-genome Sequence Data', *Molecular Biology and Evolution*, 31: 1869–79.
- Didelot, X. et al. (2021) 'Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo', *Current Protocols*, 1: e60.
- Frise, R. et al. (2016) 'Contact Transmission of Influenza Virus between Ferrets Imposes a Looser Bottleneck than Respiratory Droplet Transmission Allowing Propagation of Antiviral Resistance', *Scientific Reports*, 6: 29793.
- Graudenzi, A. et al. (2021) 'Mutational Signatures and Heterogeneous Host Response Revealed via Large-scale Characterization of SARS-CoV-2 Genomic Diversity', *iScience*, 24: 102116.
- Grubaugh, N. D. et al. (2019a) 'An Amplicon-based Sequencing Framework for Accurately Measuring Intra-host Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.
- et al. (2019b) 'Tracking Virus Outbreaks in the Twenty-first Century', *Nature Microbiology*, 4: 10–9.
- Gu, H. et al. (2022) 'Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong', *Nature Communications*, 13: 736.
- Hall, M., Woolhouse, M., and Rambaut, A. (2015) 'Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set', *PLoS Computational Biology*, 11: e1004613.
- Hall, M. D., Woolhouse, M. E. J., and Rambaut, A. (2016) 'Using Genomics Data to Reconstruct Transmission Trees during Disease Outbreaks', *Revue Scientifique et Technique*, 35: 287–96.
- Hamilton, W. L. et al. (2021) 'Genomic Epidemiology of COVID-19 in Care Homes in the East of England', *eLife*, 10: e64618.
- Han, A. X. et al. (2019) 'Inferring Putative Transmission Clusters with Phylodelity', *Virus Evolution*, 5: vez039.
- Henn, M. R. et al. (2012) 'Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants upon Immune Recognition during Acute Infection', *PLoS Pathogens*, 8: e1002529.
- Hensley, S. E. et al. (2009) 'Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift', *Science*, 326: 734–6.
- Jary, A. et al. (2020) 'Evolution of Viral Quasispecies during SARS-CoV-2 Infection', *Clinical Microbiology and Infection*, 26: 1560.e1–1560.e4.
- Jombart, T. et al. (2014) 'Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data', *PLoS Computational Biology*, 10: e1003457.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Kemp, S. A. et al. (2021) 'SARS-CoV-2 Evolution during Treatment of Chronic Infection', *Nature*, 592: 277–82.
- Koyama, T., Platt, D., and Parida, L. (2020) 'Variant Analysis of SARS-CoV-2 Genomes', *Bulletin of the World Health Organization*, 98: 495–504.
- Kubik, S. et al. (2021) 'Recommendations for Accurate Genotyping of SARS-CoV-2 Using Amplicon-based Sequencing of Clinical Samples', *Clinical Microbiology and Infection*, 27: 1036.e1–1036.e8.
- Kuipers, J. et al. (2020) 'Within-patient Genetic Diversity of SARS-CoV-2', *BioRxiv*.
- Lee, E. C. et al. (2020) 'The Engines of SARS-CoV-2 Spread', *Science*, 370: 406–7.
- Leinonen, R., Sugawara, H., and Shumway, M. International Nucleotide Sequence Database Collaboration. (2011) 'The Sequence Read Archive', *Nucleic Acids Research*, 39: D19–21.
- Letizia, A. G. et al. (2020) 'SARS-CoV-2 Transmission among Marine Recruits during Quarantine', *The New England Journal of Medicine*, 383: 2407–16.
- Li, H. et al. (2009) 'The Sequence alignment/map (SAM) format and SAMtools', *Bioinformatics*, 25: 2078–9.
- Li, B. (2022) 'Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant', *Nature Communications*, 13: 460.
- Li, H. (2013) 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM', *arXiv [Q-bio.gn]*. arXiv.
- Lumby, C. K., Nene, N. R., and Illingworth, C. J. R. (2018) 'A Novel Framework for Inferring Parameters of Transmission from Viral Sequence Data', *PLoS Genetics*, 14: e1007718.
- Lythgoe, K. A. et al. (2021) 'SARS-CoV-2 Within-host Diversity and Transmission', *Science*, 372: eabg0821.
- Martin, M. (2011) 'Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads', *EMBnet.journal*, 17: 10–2.
- Martin, M. A., and Koelle, K. (2021) 'Comment on Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2', *Science Translational Medicine*, 13: eabh1803.
- Martincorena, I. et al. (2017) 'Universal Patterns of Selection in Cancer and Somatic Tissues', *Cell*, 171: 1029–41.e21.
- Matteson, N. et al. (2020) 'PrimalSeq: Generation of Tiled Virus Amplicons for MiSeq Sequencing V1 (Protocols.io.bez7j9n)',

- Protocols.io, ZappyLab, Inc. <<https://www.protocols.io/view/primal-seq-generation-of-tiled-virus-amplicons-for-bez7jf9n>> accessed 20 December 2021.
- McCrone, J. T., and Lauring, A. S. (2018) 'Genetic Bottlenecks in Intraspecies Virus Transmission', *Current Opinion in Virology*, 28: 20–5.
- Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
- Nicholson, M. D. et al. (2021) 'Response to Comment on Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2', *Science Translational Medicine*, 13: eabj3222.
- O'Toole, Á. et al. (2021) 'Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool', *Virus Evolution*, 7: veab064.
- Paradis, E., and Schliep, K. (2019) 'Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R', *Bioinformatics*, 35: 526–8.
- Parameswaran, P. et al. (2017) 'Intrahost Selection Pressures Drive Rapid Dengue Virus Microevolution in Acute Human Infections', *Cell Host & Microbe*, 22: 400–10.e5.
- Park, D. J. et al. (2015) 'Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone', *Cell*, 161: 1516–26.
- Perera, D. et al. (2021) 'Reconstructing SARS-CoV-2 infection dynamics through the phylogenetic inference of unsampled sources of infection', *PLoS One*, 16: e0261422.
- Popa, A. et al. (2020) 'Genomic Epidemiology of Superspreading Events in Austria Reveals Mutational Dynamics and Transmission Properties of SARS-CoV-2', *Science Translational Medicine*, 12: eabe2555.
- Quick, J. et al. (2017) 'Multiplex PCR Method for MinION and Illumina Sequencing of Zika and Other Virus Genomes Directly from Clinical Samples', *Nature Protocols*, 12: 1261–76.
- Rambaut, A. et al. (2020) 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.
- Rayko, M., and Komissarov, A. (2020) 'Quality Control of Low-frequency Variants in SARS-CoV-2 Genomes', *BioRxiv*.
- Sagulenko, P., Puller, V., and Neher, R. A. (2018) 'TreeTime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4: vex042.
- San, J. E. et al. (2021) 'Transmission Dynamics of SARS-CoV-2 Within-host Diversity in Two Major Hospital Outbreaks in South Africa', *Virus Evolution*, 7: veab041.
- Sapoval, N. et al. (2021) 'SARS-CoV-2 Genomic Diversity and the Implications for qRT-PCR Diagnostics and Transmission', *Genome Research*, 31: 635–44.
- Seemann, T. et al. (2020) 'Tracking the COVID-19 Pandemic in Australia Using Genomics', *Nature Communications*, 11: 4376.
- Sekizuka, T. et al. (2020) 'Haplotype Networks of SARS-CoV-2 Infections in the Diamond Princess Cruise Ship Outbreak', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 20198–201.
- Shen, Z. et al. (2020) 'Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients with Coronavirus Disease 2019', *Clinical Infectious Diseases*, 71: 713–20. Oxford Academic.
- Shu, Y., and McCauley, J. (2017) 'GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality', *Euro Surveillance*, 22.
- Sobel Leonard, A., Weissman, D. B., and Greenbaum, B. (2017) 'Transmission Bottleneck Size Estimation from Pathogen Deep-sequencing Data, with an Application to Human Influenza A Virus', *Journal of Virology*, 91: e00171–17.
- Stapleford, K. A. et al. (2014) 'Emergence and Transmission of Arbovirus Evolutionary Intermediates with Epidemic Potential', *Cell Host & Microbe*, 15: 706–16.
- Stern, A. et al. (2017) 'The Evolutionary Pathway to Virulence of an RNA Virus', *Cell*, 169: 35–46.
- Stimson, J. et al. (2019) 'Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions', *Molecular Biology and Evolution*, 36: 587–603.
- Tonkin-Hill, G. et al. (2021) 'Patterns of Within-host Genetic Diversity in SARS-CoV-2', *eLife*, 10: e66857.
- Turakhia, Y. et al. (2020) 'Stability of SARS-CoV-2 Phylogenies', *PLoS Genetics*, 16: e1009175.
- Valesano, A. L. et al. (2021) 'Temporal Dynamics of SARS-CoV-2 Mutation Accumulation within and across Infected Hosts', *PLoS Pathogens*, 17: e1009499.
- van Dorp, L. et al. (2020a) 'Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2', *Infection, Genetics and Evolution*, 83: 104351.
- et al. (2020b) 'No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2', *Nature Communications*, 11: 5986.
- Varble, A. et al. (2014) 'Influenza A Virus Transmission Bottlenecks are Defined by Infection Route and Recipient Host', *Cell Host & Microbe*, 16: 691–700.
- Vignuzzi, M. et al. (2006) 'Quasispecies Diversity Determines Pathogenesis through Cooperative Interactions in a Viral Population', *Nature*, 439: 344–8.
- Vogels, C. et al. (2020) 'Generation of SARS-COV-2 RNA Transcript Standards for qRT-PCR Detection Assays V1', *Protocols.io*. ZappyLab Inc.
- Voloch, C. M. et al. (2021) 'Intra-host Evolution during SARS-CoV-2 Prolonged Infection', *Virus Evolution*, 7: veab078.
- Wang, D. et al. (2021a) 'Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2', *Frontiers of Medicine*, 8: 585358.
- Wang, Y. et al. (2021b) 'Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Populations in COVID-19 Patients', *Genome Medicine*, 13: 30.
- Wilm, A. et al. (2012) 'LoFreq: A Sequence-quality Aware, Ultra-sensitive Variant Caller for Uncovering Cell-population Heterogeneity from High-throughput Sequencing Datasets', *Nucleic Acids Research*, 40: 11189–201.
- Wölfel, R. et al. (2020) 'Virological Assessment of Hospitalized Patients with COVID-2019', *Nature*, 581: 465–9.
- Worby, C. J. et al. (2014a) 'The Distribution of Pairwise Genetic Distances: A Tool for Investigating Disease Transmission', *Genetics*, 198: 1395–404.
- Worby, C. J., Lipsitch, M., and Hanage, W. P. (2014b) 'Within-host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data', *PLoS Computational Biology*, 10: e1003549.
- (2017) 'Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data', *American Journal of Epidemiology*, 186: 1209–16.
- Zwart, M. P., and Elena, S. F. (2015) 'Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution', *Annual Review of Virology*, 2: 161–79.