Journal of Infection 85 (2022) 64–74

Contents lists available at ScienceDirect

Journal of Infection

journal homepage: www.elsevier.com/locate/jinf

Unraveling the hurdles of a large COVID-19 epidemiological investigation by viral genomics

Regina Sá[a,*], Joana Isidro[b], Vítor Borges[b], Sílvia Duarte[c], Luís Vieira[c], João P Gomes[b], Sofia Tedim[d], Judite Matias[a], Andreia Leite[e,f]

[a] Public Health Unit of the Baixo Vouga Health Center Grouping, Regional Health Administration of the Center Portugal (ARSC), Aveiro, Portugal
[b] Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal
[c] Innovation and Technology Unit, Department of Human Genetics, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal
[d] Department of Mathematics, University of Aveiro (UA), Aveiro, Portugal
[e] NOVA National School of Public Health, Public Health Research Center, Universidade NOVA de Lisboa, Lisbon, Portugal
[f] Comprehensive Health Research Center, Universidade NOVA de Lisboa, Lisbon, Portugal

ARTICLE INFO

Article history:
Accepted 17 May 2022
Available online 21 May 2022

Keywords:
Public health
SARS-CoV-2
Disease outbreaks
Epidemics
Genomics

SUMMARY

COVID-19 local outbreak response relies on subjective information to reconstruct transmission chains. We assessed the concordance between epidemiologically linked cases and viral genetic profiles, in the Baixo Vouga Region (Portugal), from March to June 2020. A total of 1925 COVID-19 cases were identified, with 1143 being assigned to 154 epiclusters. Viral genomic data was available for 128 cases. Public health authorities identified two large epiclusters (280 and 101 cases each) with a central role on the spread of the disease. Still, the genomic data revealed that each epicluster included two distinct SARS-CoV-2 genetic profiles and thus more than one transmission network. We were able to show that the initial transmission dynamics reconstruction was most likely accurate, but the increasing dimension of the epiclusters and its extension to densely populated settings (healthcare and nursing home settings) triggered the misidentification of links. Genomics was also key to resolve some sporadic cases and misidentified direction of transmission. The epidemiological investigation showed a sensitivity of 70%–86% to detect transmission chains. This study contributes to the understanding of the hurdles and caveats associated with the epidemiological investigation of hundreds of community cases in the context of a massive outbreak caused by a highly transmissible and new respiratory virus.

Introduction

The novel coronavirus disease (COVID-19) was first reported in patients with atypical pneumonia, in December 2019, in China.[1,2] The COVID-19 outbreak was declared a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and a pandemic on March 11, 2020.[3,4] Portugal had its first detected cases of the disease reported on March 2, 2020, reaching 42 171 cases and 1 576 deaths by 30th June 2020.[5] In Portugal, the Northern and Central Regions (NUTS II) were the most affected in the first wave. Baixo Vouga (Aveiro Region) (NUTS III), with a total of 370 394 inhabitants in 2011 and 11 municipalities, belongs to the Central Region (NUTS II) and borders the Northern Region, presenting as a buffer towards south. The first confirmed case in Baixo Vouga was on March 8th, in Aveiro municipality.[6] There were a total of 1925 COVID-19 cases from March 8th to the end of June in Baixo Vouga, 39% belonging to one municipality, Ovar.[6] With about 54 120 inhabitants in 2018, Ovar was the first Portuguese municipality to be declared to have active community transmission.[7] On the same day, a sanitary cordon was decreed in this municipality, lasting between March 17 and April 17, 2020.[8]

Local outbreak response is highly dependent on the epidemiologic investigation. However, epidemiological investigation relies on subjective information to reconstruct transmission chains and cannot guarantee total validity. Thus, viral whole-genome sequencing (WGS) facilitates the identification of phylogenetic profiles of SARS-CoV-2, allowing the surveillance of the viral lineages in circulation,[9,10] the occurrence of introduction events[11–13] and the reconstruction of transmission chains by informing the putative source of transmission.[14–20] To date, some studies have scrutinized the concordance between viral genomics and epidemiological data on the reconstruction of COVID-19 transmission chains,[21–25] but only one provided a quantitative estimate.[15] Here, we combine epidemiological and viral genomic data to understand COVID-19 transmission chains, from March to June 2020, in Baixo Vouga, an

* Corresponding author.
E-mail address: arsa3@arscentro.min-saude.pt (R. Sá).

https://doi.org/10.1016/j.jinf.2022.05.013
0163-4453/

area with a municipality in sanitary cordon and presenting as one of the regions with higher incidence rates in the first wave. We aim to assess the concordance between epidemiologically linked cases and genome sequencing of SARS-CoV-2, and further describe the context of the misidentifications.

Materials and methods

Study design, data sources and primary data collection

We conducted an outbreak analysis with previously collected data on epidemiological characteristics of all confirmed cases of COVID-19 from 8th March to 30th June, in Baixo Vouga. In Portugal, health authorities were informed of confirmed cases through three possible ways: (i) the medical or laboratorial notification of the case, through the national surveillance system platform (SINAVE); (ii) the positive test result sent directly from the laboratories; or, (iii) the national tool specifically developed for COVID-19 follow-up and contact tracking (Trace COVID-19®). As part of the epidemiological investigation, primary data was collected, by the Public Health Unit of Baixo Vouga, using an internal tool created for this purpose, since the national one (SINAVE) did not allow the effective identification of some information that was essential for outbreak control, as for example the epidemiological link. For this research, we used data on the characteristics of each case, including demographic (age, sex, and municipality), clinical (date of symptoms onset, clinical presentation, and past medical history), laboratorial (date of the diagnostic test, type of diagnostic test), and epidemiological (date and context of probable contagion, infector and classification (cluster or sporadic)).

Epidemiological investigation

Epidemiological link refers to the putative source of infection of a confirmed case using the epidemiological investigation, this duplet is hereby considered as the *infected/infector.* Epidemiologically linked cases were defined as (a) direct, when a confirmed case was identified as the probable infector OR when the infector was otherwise identified through contact tracing, and (b) indirect, when a confirmed case belonged to a context where a cluster was identified but there was no clear connection with one specific confirmed case as its infector. In the case of an indirect epidemiological link, the index case of the cluster in which the infected patient was included was considered as the infector. The context of the contagion was attributed according to the relationship between infector/infected. Cases were considered sporadic in the following situations: (i) data on the source of infection was available but no infector was found; (ii) the epidemiological link involved cases outside the geographical area of Baixo Vouga; (iii) insufficient or inexistent data referring to the epidemiological link.

Epicluster refers to the complete transmission chain of interconnected cases, based on the infected/infector duplet. The same epicluster could include smaller clusters (subclusters) identified by the health authorities during the epidemiological investigation. A subcluster was considered in the presence of two or more cases at the same time and space. All the reconstruction and analysis of the epiclusters were preformed using the R version 4.0.3 and the package Epicontacts.

Viral whole-genome sequencing and genomic investigation

In Portugal, RT-PCR positive samples for SARS-CoV-2 from country-spread laboratories (enrolled in the Portuguese network)[26] are regularly sent to the National Institute of Health Doutor Ricardo Jorge (INSA) for genomic analysis, in the context of the genomic surveillance of the virus. The genome sequences analysed in this study were retrieved from the nationwide SARS-CoV-2 genome collection and, as such, no specific inclusion criteria were established to define the sample set. The availability of genomes associated with the studied patients was therefore contingent on the laboratory/Institute's partnerships in the genomics network, sample shipment conditions and sequencing success (strictly linked with viral load).[27]

SARS-CoV-2 positive RNA samples were subjected to amplicon-based whole-genome amplification with tiled, multiplexed primers, following the Artic Consortium protocol.[28] Briefly, after Illumina NexteraXT library preparation, paired-end sequencing was performed either on Illumina MiSeq or NextSeq 550, targeting ~1 M reads per sample. Analysis of sequence read data was conducted using the bioinformatics pipeline implemented in INSaFLU,[29] which is a web-based (and also locally installable) platform for amplicon-based next-generation sequencing data analysis. Phylogenetic analysis and clade assignment was performed with Nextstrain pipeline[30] version from March 27, 2021 (https://github.com/nextstrain/ncov), using default settings. In brief, sequences were aligned against the reference Wuhan-Hu-1/2019 genome of SARS-CoV-2 (GenBank accession MN908947) using MAFFT v. 7.110,[31] and further used to build a maximum likelihood phylogenetic tree based on the GTR model using IQ-TREE v2.0.3.[32] Treetime v0.8.[33] is applied to infer a time-resolved phylogeny.

To evaluate the phylogenetic context of the studied genome sequences and assess their genetic relatedness, phylogenetic trees were built for both the whole national genome collection and the Baixo Vouga dataset. To facilitate the congruence analysis between genomic and epidemiological data, genome sequences were grouped into "main viral genetic profiles" (i.e., "clade-level" clusters, 1–4) and further hierarchically coded (HC) according to their ancestry within each main genetic profile.

Statistical analysis

Since sequencing was performed in a subset of the total confirmed COVID-19 cases, the characteristics of the included sample were compared to the not sequenced population, mainly to assess the representativeness of the samples in the genome dataset. Categorical data is presented as total and relative frequencies. Numerical data (age) is presented as median and interquartile range. Associations between categorical data were tested using a Chi squared or Fisher test, depending on the expected counts per cell (if <5 Fisher test was used). Associations between numerical (age) and categorical variables were assessed using an independent T-student test, if normally distributed data, or a Wilcoxon rank-sum test, if non-normally distributed.

Concordance between genomic and epidemiologic data was assessed by identifying the prevalence of the predominant viral genetic profile for each epicluster. Sensitivity was estimated using WGS as assumed Gold-Standard. Epiclusters with less than two individuals with available WGS data were excluded from this analysis. A meta-analysis of the sensitivities (proportions) in observational and non-comparative data was performed, using each epicluster as if it were a single study. The study design was not considered as a differentiating variable because the method of epidemiological investigation was always the same. Since participants come from a single common population and go through the same procedures, performed by the same professionals, under the same conditions, a fixed-effect model was considered. Therefore, the variance of observed effect sizes across the epiclusters is due to the random sampling error inherent in each one, namely, the within-epiclustrer variance.[34] Since the proportions did not present with a normal distribution, they were transformed into their natural logarithm for the analysis and then converted back to proportions,

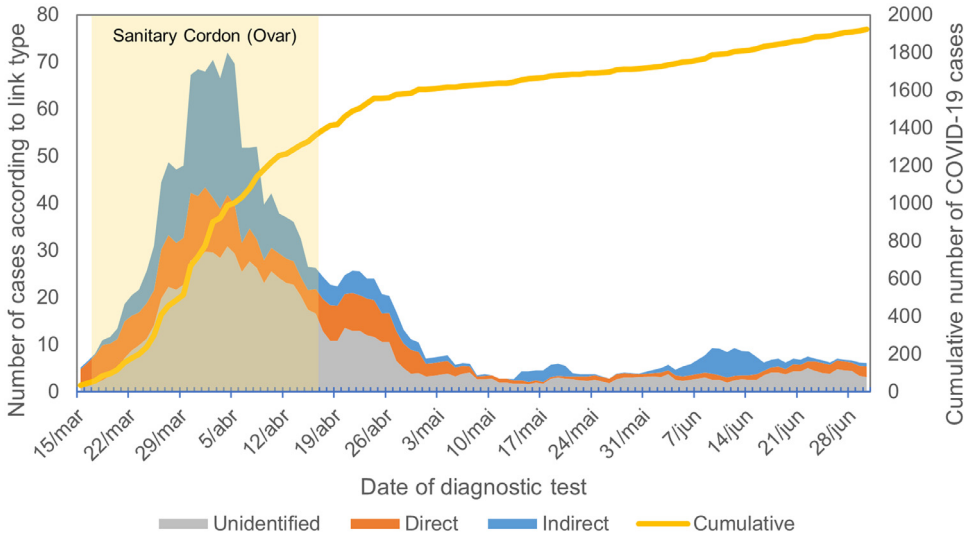

Fig. 1. Weekly evolution of COVID-19 cases, according to the link type (unidentified, direct or indirect).

together with its 95% confidence intervals, to present results easily. To analyze the study weight, the inverse of the variance of the transformed proportion was used. To pool the individual effect sizes and their sampling variances based on the inverse variance method, an analysis via linear (fixed-effects) models was performed. To test our choice on using fixed effects models (instead of random) we run a heterogeneity analysis using τ^2, I^2 and Q-statistic. The forest plot was built using the logit transformation. All statistical analyses were performed using R version 4.0.3. The level of significance was set at 5%.

Ethical considerations

Individual patient records were anonymized before analysis. This research did not involve the collecting of primary data. Secondary data, previously collected by the Public Health Authorities during epidemiological investigation, were used in accordance with the Ethical Standards for Public Health Research.

Results

Study population and epiclusters overview

A total of 1925 confirmed cases of COVID-19 were identified by the Baixo Vouga Public Health Unit from 8th March to 30th June 2020, based on the day of the sample collection of the diagnostic test (Fig. 1). Of these, 62.86% were females, and the median age was 56 years (IQR 39 - 76). The proportion of individuals in the total dataset that were diagnosed in a situation of sanitary cordon (according to time and geographical location) was 32.10%. Healthcare professionals accounted for 10,39% of the cases and the residents in nursing homes for 15.84%. Among the 424 (22.03%) cases for whom the context of contagion was available, the most common setting was cohabitant (51.65%), followed by work (30.66%).

A total of 1143 cases were assigned to epiclusters (59.38% of the total dataset), with 989 (86.53%) having an identified infector, for which 433 (43.78%) were considered direct (Table 1 and Fig. 1). Within an epicluster, 36.66% of the cases belonged to a sanitary cordon area, while for sporadics the corresponding value was 25.45%. Sporadics were more likely to present previous medical conditions (37.93%) then cases within an epicluster (28.86%). Also, certain clinical presentations were more frequent in cases within an epicluster, such as cough, myalgia, headaches, rhinorrhea, anosmia and dysgeusia. Additionally, individuals within epiclusters were more frequently associated with closed contexts of nursing homes (21.43%) or healthcare (12.60%).

A total of 154 epiclusters were detected, with a median of 2 cases (IQR = 2) (ranging from 2 to 280). A total of 10 epiclusters had more than 15 cases and the biggest had 280 (A) and 101 (B). Most epiclusters were detected during a restricted period coincident with the sanitary cordon. After this period, epiclusters were detected less frequently and had a shorter duration (Fig. 2). Viral whole-genome sequencing data was available for a total of 128 cases, representing 6.65% of the total dataset and 11.20% of the cases within an epicluster (Supplementary Table 1). The available WGS dataset mostly covers the early epidemics, reflecting the initial focus of the national surveillance (127/128 samples were collected until 17 April 2020, the end of the sanitary cordon).[26,35] This subset is thus enriched by cases from the sanitary cordon area (74.22%) (Supplementary Table 1) and involves a slightly higher proportion of symptomatic patients (78.12%), when compared to cases without WGS data.

Viral genomic diversity across epiclusters and sporadic cases

Genomic analysis of the 128 available sequences revealed four main viral genetic profiles: 95 sequences forming a subclade within Nextstrain clade 20A due to a shared mutation (G24077T, leading to the Spike D839Y amino acid change) (hereafter designated as genetic profile 1), 31 sequences belonging to Nextstrain clade 20B (genetic profile 2) and two genetically distant sequences belonging to Nextstrain clade 19A (genetic profiles 3 and 4) (Fig. 3, Supplementary Fig. 1 Phylogenetic tree, Microreact interactive figure: https://microreact.org/project/2E8H7Hw8mzdYTh3DhEck4q/3ea7fc98). SARS-CoV-2 with genetic profile 1 (Pango lineage B.1.91) was most likely imported from Italy in mid-late February and highly spread in Portugal during the first wave. It was highly prevalent in the Northern and Central regions, being likely responsible for about 25% of all COVID-19 cases in the early epidemics in Portugal.[26] As such, the Baixo Vouga cases here identified integrate transmission chains representing ramifications of this massive dissemination, which is further supported by the limited circulation of this variant in other countries. The observed genetic divergence within these transmission chains enhances the resolution power of genomics to corroborate or exclude epidemi-

Table 1
Demographic, clinical presentation, context of the contagion, epidemiological link type, sanitary cordon area and case settings data for Baixo Vouga COVID-19 sporadic and within-epicluster cases until 30 June 2020.

	Sporadics (N = 782)n (%)	Cases within an epicluster (N = 1143)n (%)
Demographic		
Age (at the day of the diagnostic test)	55.79 (21.73)	56.91 (24.27)
Sex (Male)	315 (40.28%)	400 (35.00%)
Anamnesis		
Symptomatic (Yes)	528 (67.52%)	736 (64.39%)
Previous medical conditions (Yes)	99 (37.93%)	101 (28.86%)
Epidemiological link type		
Direct	.	433 (43.78%)
Indirect	.	556 (56.22%)
Sanitary cordon area (at the date of diagnosis)		
Sanitary Cordon (Yes)	199 (25.45%)	419 (36.66%)
Case settings		
Nursing home resident or professional (Yes)	114 (14.58%)	245 (21.43%)
Healthcare professional (Yes)	56 (7.16%)	144 (12.60%)

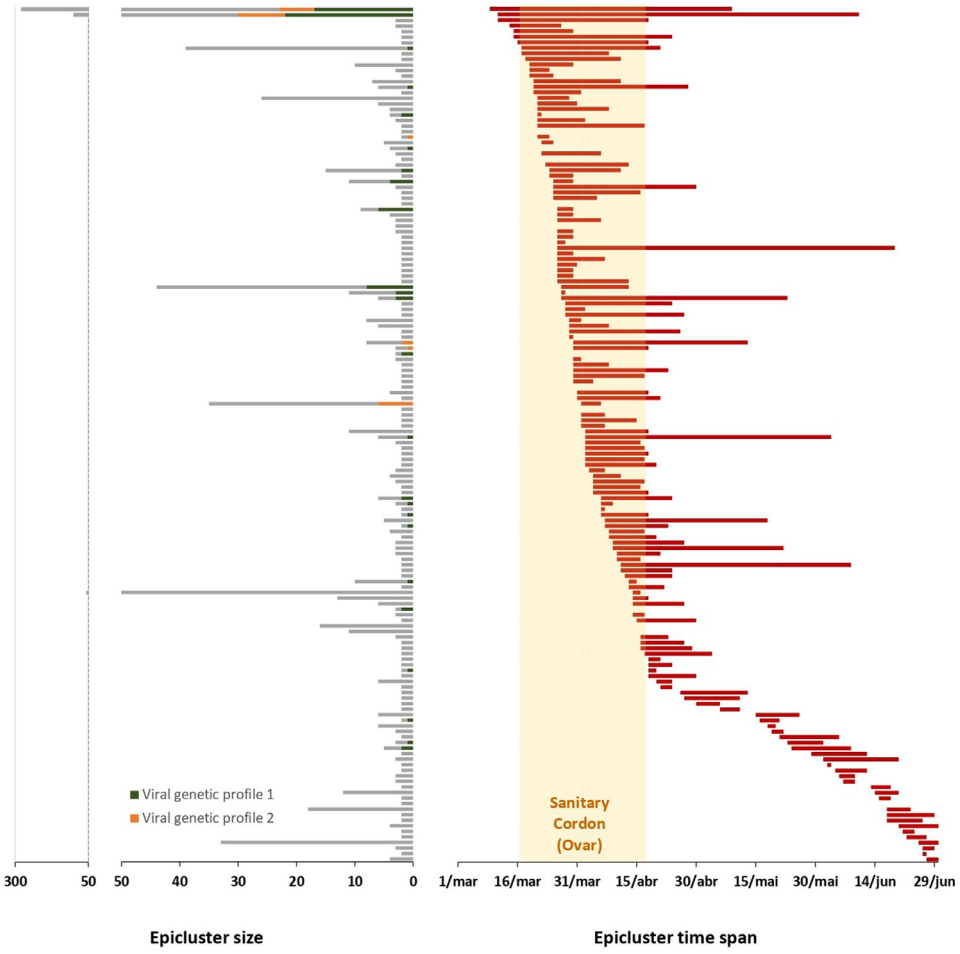

Fig. 2. Size, proportion of cases in each main viral genetic profile, and timespan of the COVID-19 epiclusters identified in Baixo Vouga from 8th March to 30th June. The left panel represents the size of each epicluster, including the number of cases in each main viral genetic profile. The right panel represents the timespan duration of the same epiclusters (using the date of diagnosis of the first and last case of the epicluster as limits).

ological links. In contrast, SARS-CoV-2 with the genetic profile 2 (including Nextstrain clade 20B root sequences and 1–3 SNPs descendants) was highly prevalent worldwide (especially in Europe) during the study period (March-June 2021),[36] and was introduced multiple times (and in multiple locations) in Portugal during the early epidemics.[35] Considering the relatively low diversification of clade 20B worldwide, the power of genomics to track the source of infection and disclose direct contacts within genetic profile 2 is reduced when compared with genetic profile 1. Regarding genetic profiles 3 and 4, both represent singleton sequences in the global national phylogenetic tree (Supplementary Figure 1).

From the 128 cases with available WGS data, 18 (14.06%) were classified as sporadic and 110 (85.94%) were associated with an epicluster, according to the epidemiological investigation (Supplementary Table 1). Sporadic cases were linked to sequences from the four main genetic profiles, including the singleton sequences of genetic profiles 3 and 4. From the 110 genotyped cases associated with epiclusters, 86 (78.18%) and 24 (21.82%) belong to

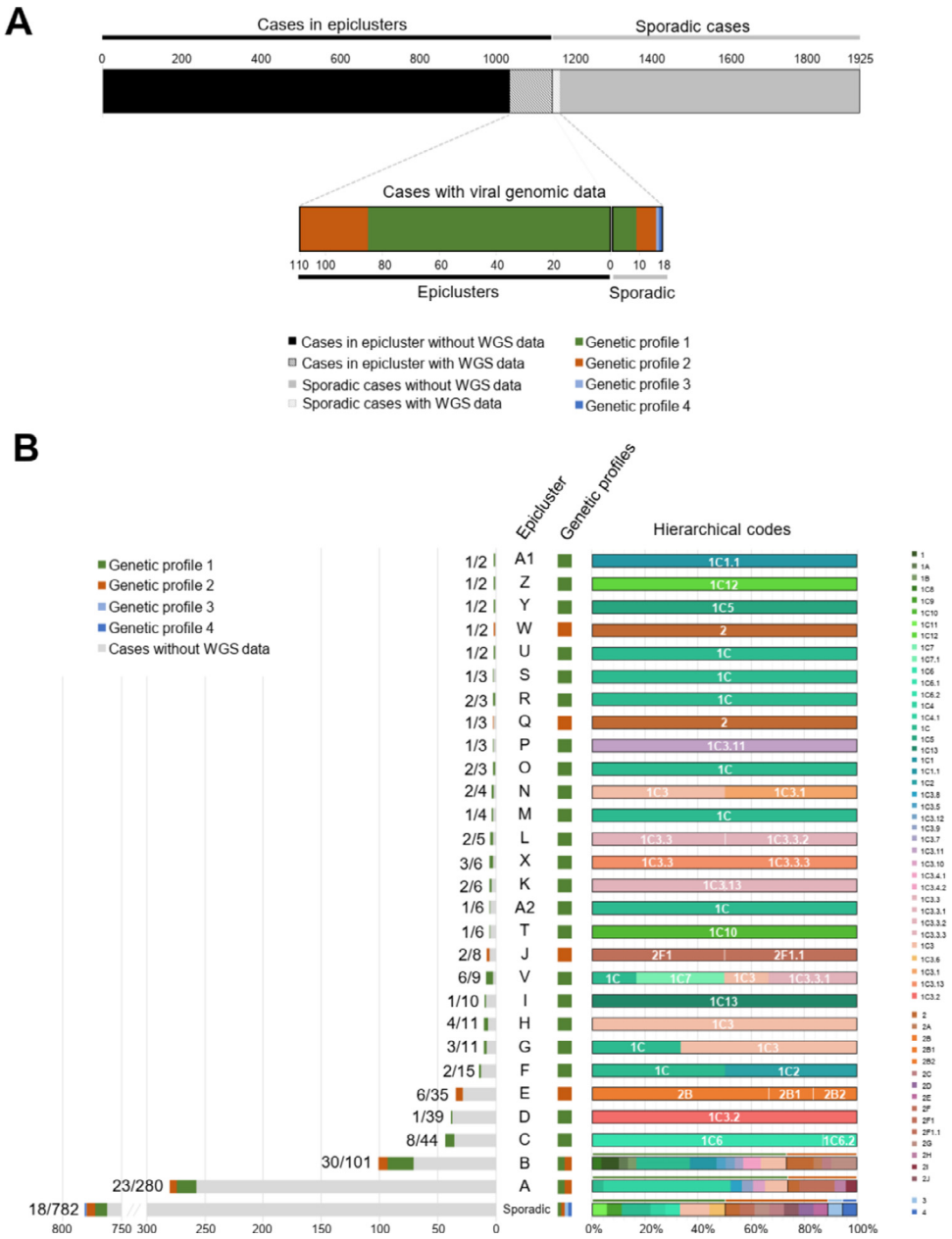

Fig. 3. Overview of the epidemiological characterization, WGS sampling and viral genomic diversity of the Baixo Vouga Region COVID-19 cases, from 8th March to 30th June 2020 ($N = 1925$). A –Proportion of cases in epiclusters and sporadics, and their distribution by viral genetic profile. B – Size and proportion of genotyped cases per epicluster and overview of their genetic diversity, represented by alphanumeric hierarchical codes (HC) classified according to their ancestry within the main genetic profile (detailed in Supplementary Figure 1).

genetic profiles 1 and 2, respectively. In total, there was WGS data available for cases epidemiologically linked to 28 distinct epiclusters (which together enrolled 625 COVID-19 cases), with 15 epiclusters including more than one genotyped case (number of genotyped cases ranged from two to 30) (Fig. 3). Twenty-six out of the 28 epiclusters were linked to a single main genetic profile (either 1 or 2), while only two epiclusters (A and B) incongruently mixed cases linked to distinct main genetic profiles. For the remaining epiclusters, the genetic profiles were highly congruent within each epicluster, i.e., each epicluster either contained 100% identical sequences or sequences with congruent ancestrality (i.e., with a hierarchically direct phylogenetic link) (Fig. 3 and Supplementary Figure 1). This integrative approach allowed, for example: i) the inclusion of sporadic cases into well-defined genetic clusters/epiclusters (e.g., PT1134 fell within tree branch 1C6, which perfectly overlapped with epicluster C; PT0539 revealed the hierarchical code (HC) 1C4, which is shared by other nine sequences, all of them linked to epicluster A); ii) the identification of matching sequences across several epiclusters, supporting that they are most likely part of the same transmission network (e.g. 1C was detected in patients linked to 10 distinct epiclusters and in 3 sporadic cases); iii) the identification of epiclusters or individual cases potentially linked to epiclusters that extended beyond the studied population, as achieved by integration of the studied sequences in the frame of the national SARS-CoV-2 genetic diversity (e.g., PT0232, which was linked to the 2-case epicluster Y, integrated a large genetic cluster in the global phylogenetic tree) (Microreact interactive figure: https://microreact.org/project/2E8H7Hw8mzdYTh3DhEck4q/3ea7fc98, Fig. 3, and Supplementary Fig. 1).[37]

Epidemiological investigation of large epiclusters: understanding incongruences by viral genomic analysis

Genomic investigation revealed that the two largest epiclusters (A and B) mixed cases linked to two distinct main SARS-CoV-2 genetic profiles, thus showing that each epicluster enrolls more than one source/transmission network. As such, we sought to reconstruct the transmission chains considering the integration of viral genetic data to understand the context that led to the observed incongruences.

The epidemiological investigation pointed that epicluster A (Supplementary Figs. 1, 2 and Fig. 3), with 280 cases (23 of which had WGS data) started with a patient that clinically presented with fever, dry cough, myalgias, headaches, and generalized weakness, that was then hospitalised. The most likely source of infection of this index case was an event that the patient attended two days prior to the symptoms' onset. Another raised possibility had been a subcluster in a local pharmacy, but this hypothesis was discarded since these cases were only diagnosed later. Indeed, the viral WGS data now shows that the index case (PT0017, HC 2) had a genomic profile that was ancestral to the one found in the pharmacy subcluster (PT2188, HC 2I), supporting that the index case might have been in the origin of this subcluster. The transmission chain then extended to an outpatient health service. The viral genome data obtained from two patients and a close personal contact of a healthcare worker also belonged to main genetic profile 2 (2F1, 2H, and 2F), thus supporting the epidemiological link with the pharmacy subcluster. Contact tracing follow-up indicated that this outpatient health service subcluster (with 36 cases) originated seven other clusters, two in nursing home settings and five in familiar settings. The two familiar subclusters with available viral genomic data belonged to the main genetic profile 1 (1C, 1C3, and 1C3.9), contradicting the epidemiological link with the outpatient health service subcluster. The smallest nursing home subcluster, with 85 cases, 11 of which had genome data available, included genomes from main profile 1 (1C4 ($n = 9$) and 1C4.1 ($n = 2$)). This subcluster was indexed to a worker that was also the domiciliary caregiver of a patient with a strong link to the outpatient health service (PT0343, 2F1). Genotyping data not only excludes this hypothesis, but also supports that this nursing home outbreak was not connected with the outpatient health service subcluster, as all the other genotyped cases were linked to SARS-CoV-2 main genetic profile 1 (clade 20A with the additional spike mutation D839Y), which was dominant in the community. The other nursing home subcluster, with 114 cases, had only one case with available WGS data (PT2190, 2F). Although the viral genome data was coherent with the epidemiological link connecting to the previous healthcare subcluster, we cannot discard a similar scenario as the one observed in the previous nursing home since there is no viral genotyping data for any of the residents.

Epicluster B (Supplementary Figs. 1, 3 and Fig. 3), as ascertained by epidemiology, included 101 cases (30 of which had WGS data) and started with a patient diagnosed in early March. Although the source of infection was unknown, members of this patient's school (in a neighbouring municipality) had returned from a trip to Italy a few days before the patient's symptoms onset (late February). Even though no viral genotyping data was available for this epicluster's index case, genomic data of some of its direct secondary cases belonged to main genetic profile 1 (PT0290 - 1C, through a personal contact and PT0189 - 1C8, through a contact in the school). Furthermore, the integrative analysis with the national collection of SARS-CoV-2 showed that a sequence outside the study population (PT0310, also a member of the same school) is the intermediary genome between PT0290 and PT0189, being the ancestor of PT0189. Since all these cases were associated with the main genetic profile 1 (clade 20A with the D839Y mutation in Spike), it is most likely that the index case was also infected with this SARS-CoV-2 variant. Although this variant has likely emerged in Italy, it was extensively disseminated in the community in Portugal in the early epidemics.[26] So, it is uncertain whether this transmission network originated in the community or whether it reflects an additional introduction of this SARS-CoV-2 variant in Portugal. According to the epidemiological investigation, epicluster B started in the municipality of Ovar (subjected to a sanitary cordon on March 17) but then extended to other territories, with only 25% of the cases belonging to the initial municipality. This transmission chain spread throughout cohabitants, family, friends, and healthcare workers that were in direct or indirect contact with the index case, then spreading into hospital settings and creating two subclusters with eleven cases (1C, 1C3.7, 1C3.4.1, 1C3.4.2) and seven cases (HC 1 ($n = 2$)), that were genetically coherent. Other subclusters involved a company (10 cases, linked to 1C ($n = 1$) and 1C1 ($n = 3$)), through a cohabitant of the index case, and a restaurant (3 cases, 1A and 1B), through a contact of the index case, with both transmissions being supported by viral genotyping. Epidemiological investigation assumed that this transmission chain extensively propagated through healthcare settings, leading to three other hospitals-related subclusters (56 cases) and one social-care related cluster (7 cases). However, the WGS data available for two of these hospital clusters (9 and 29 cases) revealed an association with either the main genetic profile 2 (2C ($n = 1$)) or a mix of both profiles 1 (1C3 ($n = 2$), 1C3.5, 1C3.12) and 2 (HC 2 ($n = 3$), 2A, 2 G ($n = 3$)), respectively.

As observed for epicluster A, the retrospective integration of genomic data showed that the initial epidemiological reconstruction of the transmission dynamics of epicluster B was most likely accurate. Still, likely due to the increasing dimension of the epiclusters, its extension to settings of higher population density, such as healthcare and nursing home settings (tightly connected with the community), the misidentification of epidemiological links inevitably occurred after a certain point of the progressively more complex epidemiological investigation.

Epidemiological versus viral genomic data: revisiting sporadic cases and misidentified direction of transmission

To resolve sporadic cases and misidentified direction of transmission, we revisited the epidemiological investigation behind these cases after integration of viral genomic data. This allowed the understanding of the hurdles and caveats associated with the epidemiological investigation during periods of high incidence rates (Fig. 1). Supplementary Table 2 summarizes the epidemiological data available for the "sporadic" cases and how genomics could indeed help placing most of these cases within well-established transmission chains. In fact, 15 out of the 18 sporadic cases could be linked either to the largest transmission network in Baixo Vouga (driven by the SARS-CoV-2 D839Y variant, i.e. the viral genetic profile 1), to specific epiclusters, or to COVID-19 cases identified outside the study population. The three remaining cases were either assumed as true sporadic (absence of epidemiological links or strong genomic links with any other case) or unresolved cases (lack of genomic resolution to infer a potential link to other cases).

WGS also allowed the identification of cases for whom the inferred direction of transmission is not supported by the genomics (Table 2). As such, by revisiting all infector-infected pairs with available genomic data, we could provide a probable explanation for the incongruence for most cases, contextualizing them within the transmission networks taking place within the geographic area of the study. We highlight five incongruences found within epicluster V, which were challenging to disentangle even with the integration of genomics. A single individual (PT0217, 1C7.1), who was presumed to have infected co-habitants and work colleagues, was

Table 2
Misidentifications of the direction of transmission between epidemiologically linked cases.

Infector - Infected pair	Epicluster	Cluster context	Epidemiological link context	Genomic incongruence	Phylogenetic distance	Probable explanation
654 (PT1207, 1B) - 397 (PT1211, 1A)	B	Restaurant	Cohabitant	Profiles 1A and 1B are two ramifications of the same ancestral profile.	4 SNPs	The patients were infected by two distinct individuals (with ancestral or matching profiles) or by the same individual (less likely).
561 (PT0189, 1C8) - 384 (PT0767, 1C)	B	Social care institution	Cohabitant	The genomic profile of the infected (1C) is ancestral to the one from the infector (1C8).	4 SNPs	The direction of transmission occurred from 1C to 1C8, being most likely mediated by another individual(s); in fact, an intermediary genetic profile (PT0310) was found in the national genome database.
662 (PT0217, 1C7.1) - 1234 (PT0265, 1C)	V	Police station	Non-cohabitant personal relationship	The genomic profile of the infected individuals is either ancestral (1C, 1C7) or represents a different ramification of the same ancestral profile (1C3, 1C3.3.1) in relation to the genetic profile collected from the infector (1C7.1).	3 SNPs	The individual with the profile 1C7.1 was most likely infected by its cohabitant (PT0218) with the profile 1C7 (congruent with symptoms onset in both patients), rather than being the infector of its cohabitants and work colleagues (as inferred by the epidemiological investigation).
662 (PT0217, 1C7.1) - 875 (PT1194, 1C3)	V	Police station	Work colleague		4 SNPs	
662 (PT0217, 1C7.1) - 1097 (PT0246, 1C3.3.1)	V	Police station	Work colleague		6 SNPs	
662 (PT0217, 1C7.1) - 1098 (PT0247, 1C3.3.1)	V	Police station	Cohabitant of case 1097		6 SNPs	
662 (PT0217, 1C7.1) - 665 (PT0218, 1C7)	V	Police station	Cohabitant of case 662		1 SNP	
34 (PT2189, 1C3.9) - 22 (PT0308, 1C3)	A	Familial	Patient/Caregiver	The genomic profile of the infected (1C3) is ancestral to the one from the infector (1C3.9), differing by 3 SNPs from each other.	3 SNPs	The direction of transmission occurred from 1C3 to 1C3.9, which is congruent with the symptoms' onset of both cases and their epidemiological context in a healthcare setting (caregiver transmitted to the patient).

indeed a secondary case of one of the two co-habitants. Intriguingly, two distinct ramifications (1C3 and 1C7) of the predominant genomic profile 1, representing two distinct sub-transmission chains, were found in both settings (familial and workplace). As such, the epidemiological link of PT0217 with the two settings led to the connection of all cases, jeopardizing the identification of more than one transmission chain solely by the contact tracing data.

Sensitivity of the epidemiological investigation

The overall sensitivity, for epiclusters with more than two cases with available viral WGS data, was 85.57% (Supplementary Table 3). Although almost all epiclusters (13/15) presented 100% of their cases within the same viral genetic profile, 14 cases out of 97 (14.43%) were discordant (i.e., presented the viral genetic profile that was not dominant within the epicluster).

Ninety-seven ($n = 97$, 75.78% of the total cases with viral WGS data) individuals were included in the pooled analysis of sensitivity, divided into 15 epiclusters, from which six presented more than three genotyped cases (the median). Low heterogeneity in the effect sizes was found ($\tau^2 = 0$; $I^2 = 0\%$; Q-statistic = 4.332 (p-value = 0.993)). For that reason, along with the fact that the methodology of epidemiological investigation was homogeneous for all epiclusters, a fixed effects model was used. The average sensitivity of the pooled analysis, and its 95% confidence interval, showed to be 79.55% (70.42% – 86.40%) (Fig. 4).

Discussion

Local outbreak response is highly dependent on the epidemiologic investigation to control transmission and reduce the spread of COVID-19. As such, knowing to what degree and in what contexts epidemiological investigation is failing to identify the source of infection is important to improve implemented measures of infection prevention and control and to strengthen the preparedness for future epidemics. In this study, we retrospectively integrated viral genomic data in the frame of a large epidemiological investigation, involving 1925 COVID-19 cases, that was carried out from March to June 2020, in Baixo Vouga Region, one of the regions with the highest incidence rates during the first epidemic wave in Portugal. Resulting from this extensive epidemiologic investigation, we detected 154 epiclusters that included 59% of the cases of the total dataset. Viral genotyping data was available for 128 cases, 18 of which had been previously considered sporadic. Four main viral genetic profiles were identified, revealing 74% of samples belonging to a sub-clade within Nextstrain clade 20A due to the shared D839Y mutation in the Spike protein (profile 1), 24% of sequences belonging to Nextstrain clade 20B (profile 2) and two genetically distant sequences belonging to Nextstrain clade 19A (profiles 3 and 4). This findings are coherent with the virus circulating both in Portugal and Europe in the initial phase of the epidemic.[26,35,36]

We provided a quantitative assessment of the epidemiological investigation validity, by assessing the agreement of epiclusters with the main viral genetic profiles. The overall sensitivity of epidemiological investigation to associate cases that belonged to the same epicluster was high (70%–86%), which is in line with previous findings.[15,17] While we found agreement regarding the main genetic profile for all cases within 13 out of the 15 epiclusters with at least two genotyped cases, these were small clusters, thus with highly uncertain values. The smaller sensitivity obtained in the metanalysis, when comparing to the overall sensitivity, points the importance of combining these results to obtain a more robust estimate. To our knowledge, the other study that quantified that estimative for each epidemiologically-linked group obtained a me-

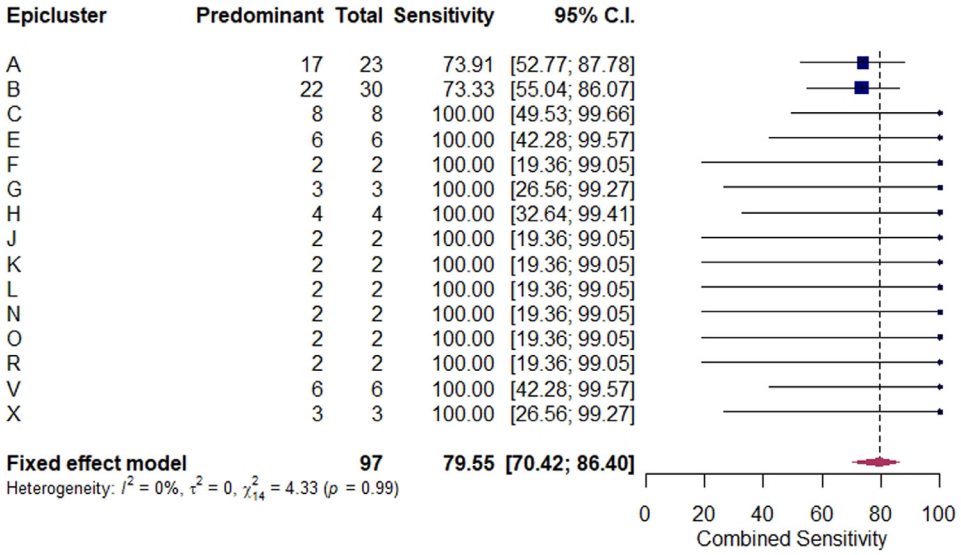

Fig. 4. Forest plot with proportions' metanalysis results of epidemiological investigation sensitivity. (Epicluster) Each epicluster with viral WGS data on more than two cases; (Predominant) The number of cases with viral WGS that belong to one predominant viral genetic profile; (Total) The total number of cases in each epicluster with viral WGS data; (Sensitivity) The proportion of the cases within the predominant viral genetic profile among the total of cases with WGS data. Horizontal lines depict CIs for each study. The surrounding box shows the contribution made by each epicluster's estimate to the overall pooled estimate, weighted by the standard error of that individual series. Estimates are the inverse of the logit transformation.

dian of 100% (IQR = 93%–100%) of cases associated with a single dominant viral genetic cluster.[15] However, the differences in sample size, context and methodology behind this study, as well as the low viral genetic diversity observed during the first months of SARS-CoV-2 expansion in human population, challenges a direct comparison. For instance, our investigation supports that the large majority of Baixo Vouga cases integrate epiclusters representing ramifications of a massive dissemination of a SARS-CoV-2 Spike D839Y variant,[26] as ascertained by the detection of a similar genetic profile across multiple epiclusters. This particular context, together with the high COVID-19 incidence, might have led to unrecognized or undocumented contacts between cases from distinct epiclusters, justifying why they were not linked.

This retrospective insight on a large-scale epidemiological investigation performed during the early spread of an emerging pandemic virus yielded important observations. A first remarkable observation is that two distinct SARS-CoV-2 genetic clades were mixed in the two largest epiclusters (A and B), as also observed in an epicluster traced in another region of Portugal.[38] By revisiting their expansion, we showed that the misidentification of epidemiological links occurred when the transmission chains extended to healthcare and nursing home settings. The public health authorities logically attributed all cases inside the same closed setting to the same epicluster. However, such settings (i.e., healthcare, nursing homes) can bias the contact tracing investigation due to their high population density and tight connection with the community, thus potentiating the misidentification of epidemiological links.[15,20,39–41] A second observation, in line with previous knowledge, is that we were able to identify, inside healthcare facilities, situations of probable nosocomial transmission,[42–46] situations of unlikely nosocomial transmission[47,48] and situations where two distinct transmission chains were developing in the same facility.[20,39,41,49] For example, in epicluster B, a healthcare service presented two subclusters with cases residing in Baixo Vouga, of whom viral sequences (HC 1(n = 2), 1C, 1C3.7, 1C3.4.1, 1C3.4.2) showed phylogenetic coherence between them. These cases belong to the same transmission chain, although they did not directly infect each other. In particular, 1C3.4.1 and 1C3.4.2 are exclusive to this setting and hence support nosocomial transmission. Another healthcare subcluster in the same epicluster (B) presented with a mix of two different main genetic profiles, indicating the presence of parallel transmission chains. Additionally, in an outpatient healthcare setting, we were able to confirm probable transmission, in a moment when infection prevention and control measures were not yet fully implemented in Portugal (e.g., in that period the use of masks was only recommended for symptomatic people). Another not unexpected observation was the confirmation of transmission inside nursing homes, in line with previous knowledge.[16,22,39,50–53] Indeed, transmission within this setting was supported by the clear phylogenetic segregation of 11 cases (1C4 and 1C4.1) (Supplementary Figure 1) with identical or near identical viral sequences (maximum 1 SNP difference). These findings are in line with previous knowledge and support the need of implementing effective measures to prevent and control infection in closed settings, such as nursing homes.[22] A fourth observation is that in a company we verified high genetic coherence between epidemiologically linked cases of epicluster B, supporting transmission within this setting.[41,54–56] Fifth, our investigation illustrates a large interaction between different settings (e.g., workplace, familial, healthcare services), which is in line with previous literature.[15,17,20,39,40,56]

In a context of community transmission and rapid spread, our findings show a decrease in the reliability of the epidemiological links over time. The retrospective integration of genomic data showed that the initial epidemiological reconstruction of the transmission dynamics of epicluster B was mostly accurate, with a clear dominance of the same main viral genetic profile. Still, likely due to the increasing dimension of the epiclusters and its extension to settings of higher population density (e.g., healthcare), the misidentification of epidemiological links inevitably occurred after a certain point and, in fact, viral sequencing reveals a mix of genetic profiles indicating parallel transmission chains. This was most likely due to the difficulty of rapidly adapting existing resources to a high speed of viral dissemination, for which the services were not prepared, especially in the context of an epidemic of a poorly characterized infectious disease.

In our retrospective investigation, genomics was also key to understand some sporadic cases and misidentified infector-infected

direction of transmission. Although only a few cases could be screened, they were very informative at several levels by unraveling distinct contexts where epidemiological investigation was limited either because not enough information could be collected for contact tracing or due to the lack of resolution to solve the epidemiological links and/or the direction of transmission (even when rich epidemiological data was available). For instance, genomics allowed 13 sporadic cases to be linked to Baixo Vouga transmission chains, while others were associated with cases and/or genetic clusters expanding beyond the study region (as assessed by the additional integration of the viral data into the nationwide genome collection). In what comes to misidentified directions of infection, we found a peculiar situation in a work setting subcluster, where a worker (1C7.1) was wrongly assumed as the infector of both colleagues and its cohabitants. In fact, a genotyped cohabitant presented a more ancestral profile (1C7), indicating a different direction of transmission. Intriguingly, two distinct ramifications (1C3 and 1C7) of the main viral genetic profile 1 were found. Although some intermediary cases are not represented in our dataset, two colleagues (1C3 and 1C3.3.1.) most certainly belong to the same transmission chain. Apart from the difficulties that are inherent to the epidemiological investigation (i.e., collected data is highly dependent on the information provided by the patient), these observations consolidate the need of having robust surveillance systems integrated with pathogen genomic sequencing. This addition offers a critical value not only to promote the early detection and enhanced monitoring of emergent and known pathogenic agents at the national and international levels, but also to support local/regional epidemiological investigation and outbreak resolution, as we were able to show.[15–17,39,40,57–59]

This study presented with some limitations. First, the subset with viral genomic data was slightly enriched with cases both symptomatic or from a specific sanitary cordon area within Baixo Vouga Region. The public health "focus" on the sanitary cordon area and the higher sequencing success for samples with higher viral load (which correlates with symptomatology[60]) likely justifies the overrepresentation of this kind of samples in the national genome collection (from where sequence data was retrospectively retrieved; https://insaflu.insa.pt/covid19/). Second, we present a relatively low proportion of sequenced viral samples (6.5%)[15–17,20], which is a direct consequence of the study involving a retrospective integration of the available SARS-CoV-2 national sequence database, rather than involving a "real-time" sequencing. Nonetheless, this might be an issue for the overall representativeness of the studied sample for the entire population, limiting the external validity of the inferences, and inter-study comparability. Limited sequencing can also lead to the underrepresentation in the phylogenetic tree of cases with epidemiological importance that may reveal a hidden transmission chain. This could also be a result of such epidemiologically important cases presenting low viral loads and, therefore, lower chances of successful sequencing. Finally, in the quantitative analysis, the 15 epiclusters had a small number of sequenced cases, thus associated with highly uncertain values. However, by preforming a metanalysis we weighted the sample size in the final estimates, overcoming this issue.

This study also contributes with relevant knowledge for future epidemics preparedness. The Portuguese epidemiological surveillance system and associated procedures are homogenously implemented at the national level. Therefore, we expect that the resources, context, and challenges associated with the epidemiological investigation in other regions, as well as the main hurdles, are possibly similar to the ones presented here for Baixo Vouga. Even though, previous capacity building and some specific actions taken by the Public Health Unit of Baixo Vouga might have facilitated the ability to conduct such massive investigation involving hundreds of epidemiological inquiries. Harmonized data collection and management, teams fully dedicated to epidemiological investigation and contact tracing, training and automatization of processes, and the ability of rapidly mobilizing trained professionals according to the necessities are examples of such actions. For instance, a previously created internal tool, that allows the integration of rich epidemiological data and a more comprehensive linkage of cases (features not fully covered by the national reporting system – SINAVE), turned out to be critical to better follow transmission chains. Furthermore, professionals and IT tools (with the support of the management bodies) were rapidly mobilized to support the investigation, also enhancing the capacity to perform timely contact tracing. Such improvements could be considered for implementation in other contexts. International health authorities are strongly promoting the urge integration of genome sequencing data as an essential component of surveillance systems.[61] In Portugal, such a framework is partially established, but still needs to be strengthened to automate the systematic integration of genomic and epidemiological data, and to promote timely communication of results back to local public health entities.[35] These efforts will not only require the re-design of currently implemented systems towards a more comprehensive and interactive data flow between all stakeholders, but also require that all public health teams and decision-makers are aligned with the need to embed genomics into the routine activities of infectious disease surveillance and outbreak resolution. The COVID-19 pandemic highlighted this field by establishing genomic surveillance as a key tool for Public Health decision-making towards outbreak control. Still, despite the great advances in this field for SARS-CoV-2, there are huge discrepancies observed between countries, in particular, in the establishment of frameworks for timely and robust integration of epidemiological and genomic data.[57,58,62,63,63–65] Also, the level of detail captured by field public health teams during epidemiological investigations and available records (which depend on data protection regulations) might vary from country to country with potential impact on reliability and power of the investigation, as well as on the robustness of outcome comparisons across studies/countries. Still, genomics-informed public health outbreak responses at country level theoretically demand multiple resources and logistics, involving wide participation of laboratories, (de)centralized data analysis (depending on the geographical organization of the health system: national, regional, or local), and the rapid sharing of integrated genomic and epidemiological data.

In conclusion, the present study identified some of the difficulties associated with the epidemiological investigation of hundreds of community cases in the context of massive epidemics. It constitutes an example of how the integration of genomic data can strongly support and complement epidemiological investigation and public health actions and guide preparedness actions to face future epidemics. This research is able to describe with detail transmission chains with great dimension, in an early phase of the epidemic, including a setting of sanitary cordon, in Portugal, and providing a quantitative measure of sensitivity of the epidemiological investigation and a framework for future systems.

Funding

This study did not receive any funding.

Author agreement

All authors have seen and approved the final version of the manuscript.

Ethical approval

Ethical approval for this study was granted from the Regional Health Administration of the Center Portugal (ARSC) Ethics Committee (23/2021) and from the Ethics Committee for Health of the National Institute of Health Doutor Ricardo Jorge (INSA) (March/2020).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jinf.2022.05.013.

References

1. Peng Z, Xing-Lou Y, Xian-Guang W, Ben H, Lei Z, Wei Z, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**(7798):270–3. doi:10.1038/s41586-020-2012-7.
2. Na Z., Dingyu Z., Wenling W., Xingwang L., Bo Y., Jingdong S., et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020. DOI: 10.1056/NEJMoa2001017.
3. World Health Organization. IHR emergency committee on novel coronavirus (2019-nCoV). Available at https://www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov). Accessed June 3, 2020, 2020.
4. World Health Organization. WHO director-general's opening remarks at the media briefing on COVID-19. Available at https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. Accessed June 3, 2020, 2020.
5. Saúde DG. *Novo Coronavírus COVID-19: Relatório da Situação Epidemiológica Em Portugal*. Lisboa; 2020.
6. Unidade de Saúde Pública do ACES Baixo Vouga. *Relatório De Situação Da COVID-19 Na Região Do Baixo Vouga*. Aveiro; 2020.
7. National Institute of Statistics. *Resident Population (No.) By Place of Residence (Ovar, NUTS III - 2013), Annual*. Statistics Portugal; 2018. Available at https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0008273&selTab=tab0. Accessed June 3, 2020, 2020.
8. Resolution of the Council of Ministers 10-D /2020, 2020-03-19. Electronic journal of the republic. Available at https://dre.pt/web/guest/home/-/dre/130413790/details/maximized?serie=I&day=2020-03-19&date=2020-03-01. Accessed June 3, 2020, n.d.
9. Global Initiative on Sharing All Influenza Data (GISAID). Available at https://www.gisaid.org/. Accessed May 20, 2021, n.d.
10. Nextstrain. Available at https://nextstrain.org/. Accessed January 23, 2021, n.d.
11. Paola S, Giovanni F, Lo PA, Stefano F, Antonella M, Eleonora B, et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance* 2020;**25**(13):2000305. doi:10.2807/1560-7917.ES.2020.25.13.2000305.
12. Joilson X, Marta G, Talita A, Vagner F, Vitor BC, Aparecida RA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg Microbes Infect* 2020;**9**(1):1824–34. doi:10.1080/22221751.2020.1803146.
13. Andreas W, Torsten H, Tobias W, Kohns VM, Daniel S, Tina S. Genetic structure of SARS-CoV-2 reflects clonal superspreading and multiple independent introduction events, North-Rhine Westphalia, Germany, February and March 2020. *Euro Surveill* 2020;**25**(22). doi:10.2807/1560-7917.ES.2020.25.22.2000746.
14. Peter F, Lucy F, Colin R, Michael F. Phylogenetic network analysis of SARS-CoV-2 genomes. *PNAS* 2020;**117**(17):9241–3. doi:10.1073/pnas.2004999117.
15. Torsten S, Lane RC, Sherry LN, Sebastian D, Anders GS, Leon C, et al. Tracking the COVID-19 pandemic in Australia using genomics. *Nat Commun* 2020;**11**(1):4376. doi:10.1038/s41467-020-18314-x.
16. Ana DSF, G SJ, Thomas W, Joseph H, Elihu AC, Patawee A, et al. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol* 2021;**6**(1):112–22. doi:10.1038/s41564-020-00838-z.
17. Samira AM, Adil AW, Latif KA, Amina AJ, Sajjad A, Hanan A, et al. Molecular epidemiology of COVID-19 in Oman: a molecular and surveillance study for the early transmission of COVID-19 in the country. *Int J Infect Dis* 2021;**104**:139–49. doi:10.1016/j.ijid.2020.12.049.
18. Francesca DG, Sebastian D, Ilaria P, Valentina C, Francesca P, Cesare C, et al. Genomic epidemiology of the first wave of SARS-CoV-2 in Italy. *Viruses* 2020;**12**(12). doi:10.3390/v12121438.
19. BöhmerMerle M, Udo B, M CV, Martin H, Katharina K, Marosevic Durdica V, et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. *Lancet Infect Dis* 2020;**20**(8):920–8. doi:10.1016/S1473-3099(20)30314-5.
20. Meredith LW, Hamilton LW, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020;**20**(11):1263–72. doi:10.1016/S1473-3099(20)30562-4.
21. Claudia A, Valeria C, Antonio P, Valentino C, Monica T, Luna C, et al. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat Commun* 2021;**12**(1):434. doi:10.1038/s41467-020-20688-x.
22. Joanne T. Serial testing for SARS-CoV-2 and virus whole genome sequencing inform infection risk at two skilled nursing facilities with COVID-19 outbreaks — minnesota, April–June 2020. *MMWR Morb Mortal Wkly Rep* 2020;**69**. doi:10.15585/mmwr.mm6937a3.
23. Karmarkar EN, Blanco I, Amornkul PN, DuBois A, Deng X, Moonan PK, et al. Timely intervention and control of a novel coronavirus (COVID-19) outbreak at a large skilled nursing facility—San Francisco, California. *Infect Control Hosp Epidemiol* 2020:1–8 n.d. doi:10.1017/ice.2020.1375.
24. Wallace M, James AE, Silver R, Koh M, Tobolowsky FA, Simonson S, et al. Rapid transmission of severe acute respiratory syndrome coronavirus 2 in detention facility, Louisiana, USA, May–June 2020. *Emerg Infect Dis* 2021;**27**(2):421–9. doi:10.3201/eid2702.204158.
25. Pedro N, Fernandes V, Cavadas B, Guimarães JT, Barros H, Tavares M, et al. Field and molecular epidemiology: how viral sequencing changed transmission inferences in the first portuguese SARS-CoV-2 infection cluster. *Viruses* 2021;**13**(6):1116. doi:10.3390/v13061116.
26. Borges V, Isidro J, Cortes-Martins H, Duarte S, Vieira L, Leite R, et al. Massive dissemination of a SARS-CoV-2 Spike Y839 variant in Portugal. *Emerg Microbes Infect* 2020;**9**(1):2488–96. doi:10.1080/22221751.2020.1844552.
27. Borges V, Sousa C, Menezes L, Gonçalves AM, Picão M, Almeida JP, et al. Tracking SARS-CoV-2 lineage B.1.1.7 dissemination: insights from nationwide spike gene target failure (SGTF) and spike gene late detection (SGTL) data, Portugal, week 49 2020 to week 3 2021. *Eurosurveillance* 2021;**26**(10):2100131. doi:10.2807/1560-7917.ES.2021.26.10.2100130.
28. Quick J. nCoV-2019 sequencing protocol 2020. DOI: 10.17504/protocols.io.bbmuik6w.
29. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP. INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. *Genome Med* 2018;**10**(1):46. doi:10.1186/s13073-018-0555-0.
30. Hadfield J, Megill C, BellSidney M, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;**34**(23):4121–3. doi:10.1093/bioinformatics/bty407.
31. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**(14):3059–66. doi:10.1093/nar/gkf436.
32. Nguyen LT, Schmidt HA, Haeseler VA, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**(1):268–74. doi:10.1093/molbev/msu300.
33. Sagulenko P, Puller V, Neher RA. Tree time: maximum-likelihood phylodynamic analysis. *Virus Evol* 2018;**4**(1) vex042. doi:10.1093/ve/vex042.
34. Wang N.. How to conduct a meta-analysis of proportions in R: a comprehensive tutorial. Available at https://www.researchgate.net/publication/325486099_How_to_Conduct_a_Meta-Analysis_of_Proportions_in_R_A_Comprehensive_Tutorial. Accessed June 29, 2021, 2018.
35. Borges V., Isidro J., Trovão N.S., Duarte S., Cortes-Martins H., Martiniano H., et al. The early dynamics of the SARS-CoV-2 epidemic in Portugal. MedRxiv 2021:2021.02.22.21252216. DOI: 10.1101/2021.02.22.21252216.
36. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance* 2020;**25**(32):2001410. doi:10.2807/1560-7917.ES.2020.25.32.2001410.
37. National Institute of Health (INSA) Dr. Ricardo Jorge. Genetic diversity of the novel coronavirus SARS-CoV-2 (COVID-19) in Portugal. Available at https://insaflu.insa.pt/covid19/. Accessed July 29, 2021, n.d.
38. Pedro N, Silva CN, Magalhães AC, Cavadas B, Rocha Ana M, Moreira Ana C, et al. Dynamics of a dual SARS-CoV-2 lineage co-infection on a prolonged viral shedding COVID-19 case: insights into clinical severity and disease duration. *Microorganisms* 2021;**9**(2):300. doi:10.3390/microorganisms9020300.
39. Page AJ, Mather AE, Le-Viet T, Meader EJ, Alikhan NF, Kay Gemma L, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb Genom* 2021;**7**(6). doi:10.1099/mgen.0.000589.
40. Snell LB, Fisher CL, Taj U, Stirrup V, Merrick B, Alcolea-Medina A, et al. Combined epidemiological and genomic analysis of nosocomial SARS-CoV-2 infection early in the pandemic and the role of unidentified cases in transmission. *Clin Microbiol Infect* 2021. doi:10.1016/j.cmi.2021.07.040.
41. Lehnertz NB, Wang X, Garfin J, Taylor J, Zipprich J, Von BB, et al. Transmission dynamics of severe acute respiratory syndrome coronavirus 2 in high-density settings, Minnesota, USA, March–June 2020. *Emerg Infect Dis* 2021;**27**(8):2052–63. doi:10.3201/eid2708.204838.
42. Stirrup O, Hughes J, Parker M, Partridge DG, Shepherd JG, Blackstone J, et al. Rapid feedback on hospital onset SARS-CoV-2 infections combining epidemiological and sequencing data. *Elife* 2021;**10**:e65828. doi:10.7554/eLife.65828.
43. Daniela L, Anna S, Marisa A, Angela L, Antonio S, Antonio P, et al. Investigation of an outbreak of symptomatic SARS-CoV-2 VOC 202012/01-lineage B.1.1.7 infection in healthcare workers, Italy. *Clin Microbiol Infect* 2021;**27**(8) 1174.e1-1174.e4. doi:10.1016/j.cmi.2021.05.007.
44. Chi-Chung CV, Sau-Chun FK, Kit-Hang SG, Shuk-Ching W, Shui-Kuen CL, Man-Sing W, et al. Nosocomial outbreak of coronavirus disease 2019 by possible airborne transmission leading to a superspreading event. *Clin Infect Dis* 2021;**73**(6):e1356–64. doi:10.1093/cid/ciab313.

45. Paltansing S, Sikkema RS, de Man SJ, Koopmans MPG, Oude Munnink BB, de Man P. Transmission of SARS-CoV-2 among healthcare workers and patients in a teaching hospital in the Netherlands confirmed by whole-genome sequencing. *J Hosp Infect* 2021;**110**:178–83. doi:10.1016/j.jhin.2021.02.005.
46. Lumley SF, Bede C, Nicholas S, Gillian R, Street TL, Jeremy S, et al. Epidemiological data and genome sequencing reveals that nosocomial transmission of SARS-CoV-2 is underestimated and mostly mediated by a small number of highly infectious individuals. *J Infect* 2021;**83**(4):473–82. doi:10.1016/j.jinf.2021.07.034.
47. Wong RCW, Lee MKP, Siu GKH, Lee LK, Leung JSL, Leung ECM, et al. Healthcare workers acquired COVID-19 disease from patients? An investigation by phylogenomics. *J Hosp Infect* 2021;**115**:59–63. doi:10.1016/j.jhin.2021.05.017.
48. Francis RV, Harriet B, Mitch C, Carl Y, Theocharis T, Louise B, et al. The impact of real-time whole genome sequencing in controlling healthcare-associated SARS-CoV-2 outbreaks. *J Infect Dis* 2021:jiab483. doi:10.1093/infdis/jiab483.
49. Sikkema RS, Pas SD, Nieuwenhuijse DF, Áine O'T, Jaco V, Anne VL, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *Lancet Infect Dis* 2020;**20**(11):1273–80. doi:10.1016/S1473-3099(20)30527-2.
50. Mary L, Guerrino M, Niamh M, Una SF, Gabriel G, Suzie C, et al. Whole-genome sequencing to track severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in nosocomial outbreaks. *Clin Infect Dis* 2021;**72**(11):e727–35. doi:10.1093/cid/ciaa1433.
51. Van Hensbergen M, Den Heijer Casper DJ, Petra W, Volker H, Ter Waarbeek Henriëtte LG, Oude Munnink Bas B, et al. COVID-19: first long-term care facility outbreak in the Netherlands following cross-border introduction from Germany, March 2020. *BMC Infect Dis* 2021;**21**:418. doi:10.1186/s12879-021-06093-9.
52. Ladhani SN, Yimmy CJ, Roshni J, Jonathan F, Emma CB, Amoolya V, et al. Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care Homes. *J Infect* 2020;**81**(4):621–4. doi:10.1016/j.jinf.2020.07.027.
53. Lemieux JE, Siddle KJ, Shaw BM, Christine L, Schaffner SF, Adrianne GY, et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 2021;**371**(6529). doi:10.1126/science.abe3261.
54. Chitra P, Farhat H, Harsha PK, Risha R, Pramada P, Vijayalakshmi R, et al. Genomic epidemiology reveals multiple introductions and spread of SARS-CoV-2 in the Indian state of Karnataka. *PLoS One* 2020;**15**(12):e0243412. doi:10.1371/journal.pone.0243412.
55. Dana W, Jürg B, Rampini SK, Verena K, Maryam Z, Schreiber Peter W, et al. Does respiratory co-infection facilitate dispersal of SARS-CoV-2? investigation of a super-spreading event in an open-space office. *Antimicrob Resist Infect Control* 2020;**9**(1):191. doi:10.1186/s13756-020-00861-z.
56. Boogaard Laura H, Sikkema Reina S, van Beek Janko HGM, BrockhoffHenricus J, Eva D, et al. A mixed-methods approach to elucidate SARS-CoV-2 transmission routes and clustering in outbreaks in native workers and labour migrants in the fruit and vegetable packaging industry in South Holland, the Netherlands, May to July 2020. *Int J Infect Dis* 2021;**109**:24–32. doi:10.1016/j.ijid.2021.06.021.
57. Oude MBB, Nieuwenhuijse DF, Mart S, Áine O'T, Manon H, Madelief M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat Med* 2020;**26**(9):1405–10. doi:10.1038/s41591-020-0997-y.
58. Lane CR, Sherry NL, Porter AF, Sebastian D, Kristy H, Patiyan A, et al. Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study. *Lancet Public Health* 2021;**6**(8):e547–56. doi:10.1016/S2468-2667(21)00133-X.
59. Philippe L, Nick R, Hong SL, Vittoria C, Chiara P, Van den Broeck F, et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* 2021;**595**(7869):713–17. doi:10.1038/s41586-021-03754-2.
60. Salvatore PP, Patrick D, Ashutosh W, Rabold EM, Sean B, Dietrich Elizabeth A, et al. Epidemiological correlates of polymerase chain reaction cycle threshold values in the detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Clin Infect Dis* 2021;**72**(11):e761–7. doi:10.1093/cid/ciaa1469.
61. European Centre for Disease Prevention and Control. *Expert Opinion On Whole Genome Sequencing For Public Health surveillance: Strategy to Harness Whole Genome Sequencing to Strengthen EU Outbreak Investigations and Public Health Surveillance*. LU: Publications Office; 2016.
62. National Academies of Sciences. *Engineering, and Medicine; Division on Earth and Life Studies; Board on Life Sciences; Health and Medicine Division; Board on Health Sciences Policy; Committee on Data Needs to Monitor Evolution of SARS–CoV-2. Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies*. Washington (DC): National Academies Press (US); 2020.
63. Dinesh A, Richard M, Hamilton WL, Tehmina B, Tumelty Niamh M, Brown Colin S, et al. The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. *Lancet Microbe* 2021;**0**(0). doi:10.1016/S2666-5247(21)00208-1.
64. Xianding D, Wei G, Scot F, Louis DP, Pybus Oliver G, Faria Nuno R, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 2020;**369**(6503):582–7. doi:10.1126/science.abb9263.
65. Villabona-Arenas Ch Julián H, William P, Tully Damien C. Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* 2020;**5**(7):876–7. doi:10.1038/s41564-020-0738-5.