Received: 13 July 2023 | Accepted: 12 October 2023

DOI: 10.1002/jmv.29193

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Genomic epidemiology of the main SARS-CoV-2 variants in Italy between summer 2020 and winter 2021

Annalisa Bergna[1] | Alessia Lai[1] | Carla Della Ventura[1] | Bianca Bruzzone[2] | Alessandro Weisz[3] | Morena d'Avenia[4] | Sophie Testa[5] | Carlo Torti[6] | Caterina Sagnelli[7] | Angela Menchise[8] | Gaetano Brindicci[9] | Daniela Francisci[10] | Ilaria Vicenti[11] | Nicola Clementi[12,13] | Annapaola Callegaro[14] | Emmanuele Venanzi Rullo[15] | Sara Caucci[16] | Vanessa De Pace[2] | Andrea Orsi[2,17] | Stefano Brusa[18] | Francesca Greco[19] | Vittoria Letizia[20] | Emilia Vaccaro[21] | Gianluigi Franci[3] | Francesca Rizzo[3] | Fabio Sagradi[5] | Leonardo Lanfranchi[5] | Nicola Coppola[7] | Annalisa Saracino[9] | Michela Sampaolo[13] | Silvia Ronchiadin[22] | Massimo Galli[1] | Agostino Riva[1] | Gianguglielmo Zehender[1] | SARS-CoV-2 ITALIAN RESEARCH ENTERPRISE – (SCIRE) collaborative Group

[1]Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy

[2]Hygiene Unit, IRCCS AOU San Martino-IST, Genoa, Italy

[3]Laboratory of Molecular Medicine and Genomics, Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno and Genome Research Center for Health, Baronissi, Italy

[4]UOSVD of Cytopathology and Screening, Department of Laboratory Medicines, Ospedale di Venere, Asl Bari, Bari, Italy

[5]Unit of Infectious Diseases, Azienda Socio Sanitaria Territoriale Cremona, Cremona, Italy

[6]Infectious and Tropical Disease Unit, Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy

[7]Department of Mental Health and Public Medicine, University of Campania "Luigi Vanvitelli", Naples, Italy

[8]Microbiology and Virology Laboratory, A.O.R. San Carlo Potenza, Potenza, Italy

[9]Infectious Diseases Unit, University of Bari, Bari, Italy

[10]Department of Medicine and Surgery, Clinic of Infectious Diseases, "Santa Maria della Misericordia" Hospital, University of Perugia, Perugia, Italy

[11]Department of Medical Biotechnologies, University of Siena, Siena, Italy

[12]Laboratory of Microbiology and Virology, Università "Vita-Salute" San Raffaele, Milan, Italy

[13]Laboratory of Microbiology and Virology, IRCCS San Raffaele Scientific Institute, Milan, Italy

[14]Ospedali Riuniti, Department of Infectious Diseases, Bergamo, Italy

[15]Unit of Infectious Diseases, Department of Experimental and Clinical Medicine, University of Messina, Messina, Italy

[16]Department of Biomedical Sciences and Public Health, Virology Unit, Polytechnic University of Marche, Ancona, Italy

[17]Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

[18]Department of Translational Medical Sciences, Federico II University, Naples, Italy

[19]UOC Microbiology and Virology, PO Cosenza, Cosenza, Italy

[20]UOSD Genetic and Molecular Biology, AORN Sant'Anna and San Sebastiano di Caserta, Caserta, Italy

[21]Molecular Biology Units, AOU 'S. Giovanni di Dio e Ruggi d'Aragona' Università di Salerno, Salerno, Italy

[22]Artificial Intelligence Laboratory, Intesa Sanpaolo Innovation Center, Turin, Italy

A list of authors and their affiliations appears at the end of the paper.

Correspondence
Alessia Lai, Department of Biomedical and Clinical Sciences, University of Milan, via G.B. Grassi 74, 20157, Milan, Italy.
Email: alessia.lai@unimi.it

Funding information
European Union's Horizon Europe Research and Innovation Actions; Ministero dell'Istruzione, dell'Università e della Ricerca; Regione Campania

Abstract

Since the beginning of the pandemic, SARS-CoV-2 has shown a great genomic variability, resulting in the continuous emergence of new variants that has made their global monitoring and study a priority. This work aimed to study the genomic heterogeneity, the temporal origin, the rate of viral evolution and the population dynamics of the main circulating variants (20E.EU1, Alpha and Delta) in Italy, in August 2020–January 2022 period. For phylogenetic analyses, three datasets were set up, each for a different main lineage/variant circulating in Italy in that time including other Italian and International sequences of the same lineage/variant, available in GISAID sampled in the same times. The international dataset showed 26 (23% Italians, 23% singleton, 54% mixed), 40 (60% mixed, 37.5% Italians, 1 singleton) and 42 (85.7% mixed, 9.5% singleton, 4.8% Italians) clusters with at least one Italian sequence, in 20E.EU1 clade, Alpha and Delta variants, respectively. The estimation of tMRCAs in the Italian clusters (including >70% of genomes from Italy) showed that in all the lineage/variant, the earliest clusters were the largest in size and the most persistent in time and frequently mixed. Isolates from the major Italian Islands tended to segregate in clusters more frequently than those from other part of Italy. The study of infection dynamics showed a positive correlation between the trend in the effective number of infections estimated by BSP model and the R_e curves estimated by birth-death skyline plot. The present work highlighted different evolutionary dynamics of studied lineages with high concordance between epidemiological parameters estimation and phylodynamic trends suggesting that the mechanism of replacement of the SARS-CoV-2 variants must be related to a complex of factors involving the transmissibility, as well as the implementation of control measures, and the level of cross-immunization within the population.

KEYWORDS
effective reproductive number, international contest, phylodynamic, SARS-CoV-2 variants

1 | INTRODUCTION

Since the beginning of the pandemic, SARS-CoV-2 has exhibited considerable genomic variability, despite the proofreading activity associated with the coronavirus polymerase,[1] and driven by the widespread and rapid circulation of the virus in the human population.[2]

Early isolates belonged to two distinct lineages (A and B) that then independently evolved, at a rate that has been estimated to be around two mutations per month, producing a plethora of viral lineages, descending from the original strains.[3,4]

From the first waves, in the absence of an effective surveillance and testing system, Italy experienced a wide epidemic wave whose actual size and extent remain largely unknown but that was clearly demonstrated by numbers of hospitalizations and deaths.[5–8]

The genomic epidemiology studies showed that the ancestral lineages (in particular, the most widespread B lineage) entered Italy multiple times remaining confined to a few regions,[9,10] being then quickly replaced by the B.1 lineage, characterized by the amino acid mutation D614G in the Spike protein. This lineage entered in Italy in late June or early February 2020 [11] rapidly spread from Lombardy throughout the whole Italy and other European countries,[10] quickly becoming the dominant lineage throughout the world.[12] Subsequently, several lineages, all characterized by this and additional mutations, have replaced each other all over the world. In the summer 2020, when the European borders were at least partially opened and the international travels resumed, the lineage 20E.EU.1, firstly identified in Spain, spread throughout Europe causing the second pandemic wave in Italy in the autumn 2020.[13,14]

In late 2020 was reported the first variants of concern (VOCs), indicating with this term viral strains characterized by multiple mutations in the Spike protein, causing significant effects in the viral phenotype, such as higher transmissibility, pathogenicity and/or ability to escape neutralizing antibodies, that were designated by

Greek letters according to the WHO nomenclature (https://www.who.int/activities/tracking-SARS-CoV-2-variants). Interestingly these variants did not derive one from each other, but directly descended from different ancestral lineages.

Two main variants circulated worldwide during the period covered by the present study (summer 2020-autumn 2021); the first, VOC Alpha, was identified as a variant of increasing incidence in UK in December 2020 and became the most prevalent variant (over 80%) in Europe and all over the world (about 70% of the characterized genomes), until summer 2021, when it was replaced by the VOC Delta.[15]

The VOC Delta was first identified in India, where it caused the second dramatic wave in that country in March-May 2021, and started to rise in Europe in late spring 2021, becoming the most prevalent (over 90%-100% of the infections) lineage worldwide from summer 2021 (https://gisaid.org/hcov19-variants/). The spread of this VOC was so broad that several derived sublineages raised all over the world.[16]

At the same time, the vaccination campaign against COVID-19, which began in many European and Western countries between December 2020 and early 2021, was proceeding apace, resulting in immunization of most of the population in those countries with at least two doses by the spring 2021 (https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html#uptake-tab).

The waves due to the replacement of a viral variant with the next one resulted in a typical fluctuating pattern of the effective or net reproductive number (R_e or R_t, respectively).[17]

Although the real meaning of R_t has been questioned, its estimate has proved useful in understanding the course of the pandemic and has been used to assess the effectiveness and/or the need for more stringent control measures in many countries, including Italy.[18]

Phylodynamic analysis, which is based on the combination of evolutionary, epidemiological and immunological characteristics influencing the shape of a viral phylogeny, has become an increasingly important tool in the molecular surveillance of infections, particularly those due to emerging viruses.[19,20]

The COVID-19 pandemic presented a unique opportunity to apply such analyses to molecular surveillance due to the availability of SARS-CoV-2 genomic sequences since early January 2020, which has reached the impressive number of more than 15 million complete genomic sequences stored in one of the most widely used databases (GISAID, https://gisaid.org/, as of April 2023).

These kinds of approaches together with genomic epidemiology have made it possible to trace international pandemic flows, reconstruct outbreaks and transmission networks, estimate transmissibility, and identify variants with higher transmissibility or immune escape capacity.[21]

In the present study, it was investigated the genomic epidemiology and phylodynamics of SARS-CoV-2 main lineages and VOCs, being prevalent in Italy during the period between summer 2020 and winter 2021, using genomes sampled at that time by the SCIRE (SARS-CoV-2 Italian Research Enterprise) Italian study network, along with a selection of national and international genomes available on public databases with known date and sampling location.

2 | MATERIALS AND METHODS

2.1 | Specimen collection

The nasopharyngeal swabs of a consecutive series of COVID-19 patients attending more than 70 clinical centers distributed throughout whole Italy, participating in the collaborative SCIRE group, were collected and processed between August 2020 and January 2022. All participants including both hospitalized inpatients and outpatients gave the written informed consent to the storage of their anonymized data in a protected database. All the data used in this study were previously anonymized as required by the Italian Data Protection Code (Legislative Decree 196/2003) and the general authorizations issued by the Data Protection Authority. The study was approved by the Sacco Hospital Ethics Committee (protocol n. 47866, 9 September 2020) and conducted in compliance with Good Clinical Practice guidelines and with the principles of the 1964 Declaration of Helsinki.

2.2 | Virus genome sequencing

SARS-CoV-2 RNA was extracted using the Kit QIAsymphony DSP Virus/Pathogen Midi kit on the QIAsymphony automated platform (QIAGEN, Hilden, Germany) (n = 218), the NucleoMag 96 Virus (Macherey–Nagel, Dueren, Germany) on automated KingFisher ml Magnetic Particle Processors (Thermo Fisher Scientific, Waltham, MA, USA) (n = 156) and manually with QIAamp Viral RNA Mini Kit (QIAGEN, Hilden, Germany) (n = 228).

Full genome sequences were obtained with different protocols: (i) by a modified version of the Artic Protocol (https://artic.network/ncov-2019) using the Illumina DNA Prep and the IDT ILMN DNA/RNA Index kit (Illumina, San Diego, CA, USA) or (ii) by the CleanPlex® SARS-CoV-2 Panel (Paragon Genomics Inc., Hayward, CA, USA). Sequencing was performed on the Illumina iSeq (n = 205), MiSeq (n = 247), and NextSeq (n = 150) platforms for all samples. The results were mapped and aligned to the reference genome obtained from GISAID (https://www.gisaid.org/, accession ID: EPI_ISL_406800) using the Geneious Prime software v. 11.1 (http://www.geneious.com, Biomatters, Auckland, New Zealand) or BWA-mem, and rescued using Samtools alignment/Map (Hinxton, UK) (v. 1.9).

The SARS-CoV-2 lineage was attributed to all sequences using the Pangolin COVID-19 Lineage Assigner v. 4.1.1 (https://pangolin.cog-uk.io/) and Nextclade v. 2.4.1 (https://clades.nextstrain.org/). Mutations were identified using Nextclade.

2.3 | SARS-CoV-2 data sets

A total of 847 SARS-CoV-2 whole genome sequences (WGS) were characterized by the network in the period covered by this study and made available on public databases. Patients' isolates belonging to the most widely circulating variants/lineages in that period in Italy

(n = 602) were than selected and aligned with other Italian and International sequences of the same lineage/variant, available in GISAID (https://gisaid.org).

Three datasets were set up, for each different lineage/variant: 20E.EU1 lineage (including a total of 2425 sequences), Alpha (3174 sequences) and Delta (3248 isolates) variant. Due to the large number of SARS-CoV-2 whole genome sequences available in GISAID public databases during the period under consideration, genomes were selected on the basis of the following criteria: for the Italian genomes: 10 whole genomes for each region and sampling month, in accordance with the circulation period of each variant, with a maximum of two sequences for region/week, excluding identical genomes and those with more than 5% of gap; for the international strains: five genomes for each European and non-European countries and sampling month. The composition of the dataset is summarized in Supporting Information (Table S1).

Alignment of multiple sequences was obtained using MAFFT (https://mafft.cbrc.jp/alignment/server/) and the alignment was manually cropped using BioEdit v. 7.2.6.1 (https://bioedit.software.informer.com/) at the same length (29,774 bp).

A root-to-tip regression analysis was made using TempEst (http://tree.bio.ed.ac.uk/software/tempest/) to investigate the temporal signal of the datasets.

2.4 | Phylogenetic analysis

The maximum likelihood trees of the three datasets were estimated using IQ-TREE v. 1.6.12 (http://www.iqtree.org/).[22] The GTR + F + R4 (General time reversible + empirical base frequencies + four number of categories) model, selected by the program, was used. 1,000 parametric bootstrap replicates were performed to support the nodes (≥60% bootstrap support).

The statistically significant clusters (including more than two sequences) were identified in the ML tree by Cluster Picker v.1.2.3 using 90% bootstrap support and a mean genetic distance of 1% as thresholds. Epidemiological characteristics of the identified clusters were further investigated using Cluster Matcher v.1.2 23 which allows the identification of clusters meeting given criteria. Only clusters including at least one Italian sequence were selected and classified as mixed (M), containing both Italian and non-Italian isolates in different proportions, pure Italian (IT), including only Italian genomes, or singleton (S), containing only a single Italian genome interspersed within non-Italian sequences.

2.5 | Phylodynamic analysis

To characterize the epidemiological and evolutionary history of the different SARS-CoV-2 variants/lineages in Italy, it were considered for each dataset only clusters including at least 70% of Italian sequences, having sufficient size for the analysis (>30 sequences), by using the coalescent Bayesian Skyline Plot and the birth-death models.

Bayesian analysis was performed by BEAST v. 2.7 (https://beast.community/)[23] with the same substitution model and molecular clock employed for the previously described analyses. The evolutionary rates were estimated by using a Log Normal (M = 8E-4, S = 1.25) prior distribution in real space using a strict clock under the Bayesian Skyline plot.

MCMC analyses were run for 60 million generations and sampled every 3000. Convergence was assessed by estimating Effective Sampling Size (ESS) after applying a 10% burn-in through Tracer v.1.7 software (http://tree.bio.ed.ac.uk/software/tracer/),[24] accepting ESS of at least 200. The uncertainty of estimates was indicated with 95% highest prior density (HPD) intervals.

The final tree was selected based on the maximum posterior probability (pp) value after performing a 10% burn-in using Tree Annotator v.10.4 software (included in the BEAST package). Posterior probabilities greater than 0.7 were considered significant. Finally, all trees were visualized and edited in FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

The birth-death skyline model implemented in Beast 2.7 was used to infer changes in the effective reproductive number (R_e), and other epidemiological parameters such as the death/recovery rate (δ), the transmission rate (λ), the origin of the epidemic, and the sampling proportion (ρ).[25] Given that the samples were collected during a short period of time, a "birth-death serial" model was used.

For the birth-death analysis, it were used four equidistant intervals and a Log Normal prior for the estimation of the effective reproductive number (R_e) with a mean (M) of 0.0 and a variance (S) of 1.5, which allows the R_e values to change between less than 1 and more than 7. A normal prior with M = 48.8 and S = 15 (corresponding to a 95% interval from 24.0 to 73.4) was used for the rate of becoming uninfectious. These values are expressed as units per year and reflect the inverse of the time of infectiousness (5.3–19 days; mean, 7.5) according to the serial interval estimated by Li et al.[26] Sampling probability (ρ) was estimated assuming a prior β (α = 1.0 and β = 3.547), estimated based on available genomes in the analyses and numbers of COVID-19 active cases at pick of the studied period.

The origin of the epidemic was estimated using a normal prior with M = 0.1 and S = 0.2 in units per year for 20E.EU1 clade, and a lognormal prior with M = 0.1 and S = 0.3, for Alpha and Delta variants. The mean growth rate was calculated based on the birth and recovery rates ($r = \lambda - \delta$), and the doubling time was estimated by the equation: doubling time = ln(2)/r.[27]

3 | RESULTS

3.1 | Mutation analyses of the Italian sequences

The comparison between Italian genomes and reference sequence in the 20E.EU1 dataset showed the presence of 10 amino acid substitutions in more than 10% of Italian isolates (Supporting Information: Table S2). Almost all the sequences had the distinctive

mutations of this lineage such as A220V (97.4%, $n = 1043$) in the N gene, A222V (97.8%, $n = 1048$) in the S gene, and V30L (99.3%, $n = 1064$) and L67F (99.3%, $n = 1064$) in the ORF10 and ORF14 regions, respectively. The D614G mutations in the S gene and P314L in the ORF1b, characteristics of lineage B.1, were also present almost in all isolates (99.1% and 98.3%, respectively).

Additional mutations at lower frequency were detected, ranging from 13.9% to 34.6% of isolates: A262S, P272L, Q675H in the S gene and A3623S in the gene ORF1a. Only two sequences showed the S98F mutation in the S gene, observed in the 20 A/S:98 F clade, while one sequence carried the D80Y mutation characteristic of the 20 C/S:80Y clade.

In the Alpha variant dataset, only mutations/deletions typical of this lineage have been found, as shown in Supporting Information: Table S3, with the only exception of the two consecutive mutations in the N gene, R203K (99.1%, $n = 1312$) and G204R (96.8%, $n = 1281$), characteristics of the B.1.1 lineage and descendants.

The E484K mutation, characteristics of Beta and Gamma variants, was observed in less than 1% of the sequences. Two (0.15%) sequences had the K417N mutation present in the Beta variant.

In Delta variant sequences, 34 mutations and two deletions have been identified in more than 10% of the genomes. Sixteen mutations and deletions were distinctive of this lineage, of which 14 (87.5%) were present in more than 87% of isolates (ranging from 87.2% to 99.6%). The G142D mutation in the S gene was found in just over half of cases (63.5%, $n = 948$), while the P681R mutation, also in the S gene, did not reach 10% of cases (9.2%, $n = 138$). The only substitution characteristic of the variant but present in a limited number of isolates was I82T (3.5%). A large number of mutations ($n = 12$), not identifying this variant, were found in ORF1a (Supporting Information: Table S4).

The mutations N501Y, typical of the Alpha variant, and E484K, present in the Beta and Gamma variant, were present in 0.13% of the analyzed sequences. The mutation E484Q, present in the B.1.617.1 lineages (Kappa variant) and B.1.617.3, was observed in 0.33% of cases.

3.2 | Phylogenetic analysis and dating of the Italian clusters

3.2.1 | 20E.EU1 clade

The phylogenetic analysis conducted using Maximum Likelihood approach on 20E.EU1 dataset showed that 2104 out of 2425 (87.6%) sequences were included in a total of 255 clusters, among which, 26 ($n = 917$ isolates) included at least one Italian sequence and were classified as follows: 6 (23%) pure Italians, 6 (23%) singleton (S), and 14 (54%) mixed (M) (Table 1).

Supporting Information: Figure S1 shows the dated tree obtained by Bayesian analysis on 11 Italian clusters (five mixed with >70% Italian genomes and six pure Italian). Root-to-tip regression analysis of the temporal signal revealed an association between genetic distances and sampling days (a correlation coefficient of 0.44 and a coefficient of determination [R2] of 0.2). The estimation of evolutionary rate gave a mean of 3.8×10^{-4} s/s/y (95%HPD: 3.35×10^{-4}–4.39×10^{-4}) and mean tMRCAs spanning a period from June to September 2020 (Supporting Information: Table S5). The median size of clusters was 29 (11–95) with a median duration of 11 months (ranging from 7 to 16 months). While the earlier clusters (tMRCA between June and July 2020) showed the longest duration (9–16 months), the largest size (median number of genomes 48; $p = 0.01$) and were predominantly mixed, the late clusters, with tMRCA between August and September

TABLE 1 Type and composition of clusters in the international dataset of 20E.EU1 clade, Alpha and Delta variant.

	cluster	Total clusters	n sequences	Origin of sequences		
				ITA	EU	noEU
20E.EU1	**IT**	**6**	133	133	0	0
	M	**14**	646	415	181	50
	S	**6**	138	6	91	41
	Total	**26**	917	554	272	91
Alpha	**IT**	**15**	235	235	0	0
	M	**24**	633	245	315	73
	S	**1**	24	1	22	1
	Total	**40**	892	481	337	74
Delta	**IT**	**2**	27	27	0	0
	M	**36**	1434	737	611	86
	S	**4**	75	4	14	57
	Total	**42**	1536	768	625	143

Abbreviations: EU, European sequences; IT, Italian; ITA, Italian seqeunces; M, mixed; noEU, International sequenze; S, Singleton.

2020, had the shortest duration (7–14 months) and the smallest size (12 median genomes) and were predominantly Italians.

Considering the Italian sampling locations (Supporting Information: Table S5), five clusters included more than 50% of sequences from Northern Italy, four clusters from Southern Italy and one cluster (#40), involved 92% of isolates from the largest Islands Sardinia and Sicily.

Globally, strains from Islands more frequently grouped into clusters than sequences from South, North, and Central Italian areas (73.8% vs. 55.3%, 50.6%, and 33.3%; $p = 0.003$).

3.2.2 | Alpha variant

A total of 467 clusters were found in Alpha variant dataset including 2489 of the 3147 (79.1%) SARS-CoV-2 genomes. Of the 40 clusters containing at least 1 Italian sequence24 (60%) were mixed, 15 (37.5%) pure Italian and only 1 (2.5%) singleton (Table 1).

A total of 17 Italian clusters, 15 pure Italian and two mixed, were analyzed and a mean 4.87×10^{-4} s/s/y (95%HPD: 4.18×10^{-4}–5.54×10^{-4}) evolutionary rate estimates were obtained. Root-to-tip regression analysis of the temporal signal revealed an association between genetic distances and sampling days (a correlation coefficient of 0.45 and a coefficient of determination [R^2] of 0.2).

Mean tMRCAs of clusters were between November 2020 and late February 2021, with the majority (6/17, 35.3%) originating in December 2020 (Supporting Information: Figure S2, Supporting Information: Table S6). Clusters originating in November-December included a median number of genomes of 18.5 while a median of 13 genomes was observed in those originating between January and February. Clusters persistence was between 5 and 10 months, with a median duration of 7.5 months for the earlier and 5 months for the later ones ($p = 0.002$).

An equal proportion of clusters were characterized by more than 50% of sequences from Northern and Southern Italy ($n = 6$, 35.3%, each).

Sequences from Islands were more frequently included in clusters than those from Northern, Central, and Southern Italy (53.3% vs. 38.7%, 36.7%, and 29.6%; $p = 0.009$).

3.2.3 | Delta variant

Phylogenetic analysis of Delta international dataset showed a total of 364 clusters encompassing a total of 2650 genomes (81.6%), 42 of which containing at least one Italian sequence, classified as follows: four singleton (9.5%), 36 mixed (85.7%), and two pure Italian (4.8%) (Table 1).

The median size of the nine clusters including >70% of Italians was 16 isolates (11–79). Root-to-tip regression analysis of the temporal signal revealed an association between genetic distances and sampling days (a correlation coefficient of 0.6 and a coefficient of determination [R^2] of 0.4). The estimation of evolutionary rate gave a mean of 7.57×10^{-4} s/s/y (95%HPD: 6.65×10^{-4}–8.48×10^{-4}) and mean tMRCAs spanning from March to July 2021 (Supporting Information: Figure S3, Supporting Information: Table S7). The largest two clusters (#48 and #68), were both mixed, originating in March 2021, and showed a median duration of 10 months versus 7 months for the other clusters ($p = 0.03$).

Seven clusters were characterized by a majority of sequences from Northern Italy (up to 100%, #331), and one (#137) included 100% of sequences sampled in Southern Italy. The last cluster (#193) contained a high proportion of sequences from Northern (44.4%) and Central (55.5%) Italy.

Sequences from the Islands were more frequently included in clusters than genomes from Southern, Northern and Central Italy (60% vs. 54.2%, 49.4%, 47.9%; $p = 0.0002$).

3.3 | Phylodynamics of SARS-CoV-2 lineages and VOCs in Italy

The skyline plot of 20E.EU1 dataset (Figure 1A) showed a rapid increase of the effective number of infections (N_e) during July 2020, followed by a second increase in October 2020, when the curve reached the plateau. The decrease of infections started in February 2021 reaching the lowest values in April 2021.

Consistent with this dynamics of infection, estimates of R_e showed a value greater than 1 from the beginning of the circulation of 20E, (Figure 1B), reaching a peak around July 2020 (R_e =1.125, 95% HPD: 1.08–1.18). The R_e value started to decline in autumn 2020 (October 2020), falling around 1 between December 2020–January 2021 and decreasing below 1 in April 2021.

The effective number of infections due to Alpha variant, estimated by BSP, rapidly increased in December 2020, reaching a plateau during the winter 2021. A decrease in infections was observed from April 2021 (Figure 2A).

Correspondingly a value of R_e above 1 was observed since the origin of the Alpha epidemic in the autumn 2020, with a peak around a value of 1.13 (95% HPD = 1.01–1.22) in winter 2020–2021. A decrease in the curve was observed between March and April 2021 reaching values below the unit in May 2021 (Figure 2B).

The skyline plot analyses showed a slow rise of the Delta variant in spring 2021 (between March and June 2021), followed by a rapid increase of infections in July 2021. From August the number of infections reached a plateau that persisted until the end of the year 2021 (Figure 3A).

Similarly, the R_e value (Figure 3B) shows a growth above 1 in spring 2021, reaching a value of 1.17 (95% HPD 1.1–1.3), followed by a gradual decrease to a value below 1 in the summer of the same year (August 2021).

4 | DISCUSSION

This study refers to the characterization and analysis of the genomic sequences obtained between summer 2020 and early 2022, the period in which the main viral lineages/variants circulating in Italy were 20E.EU1, Alpha and Delta.

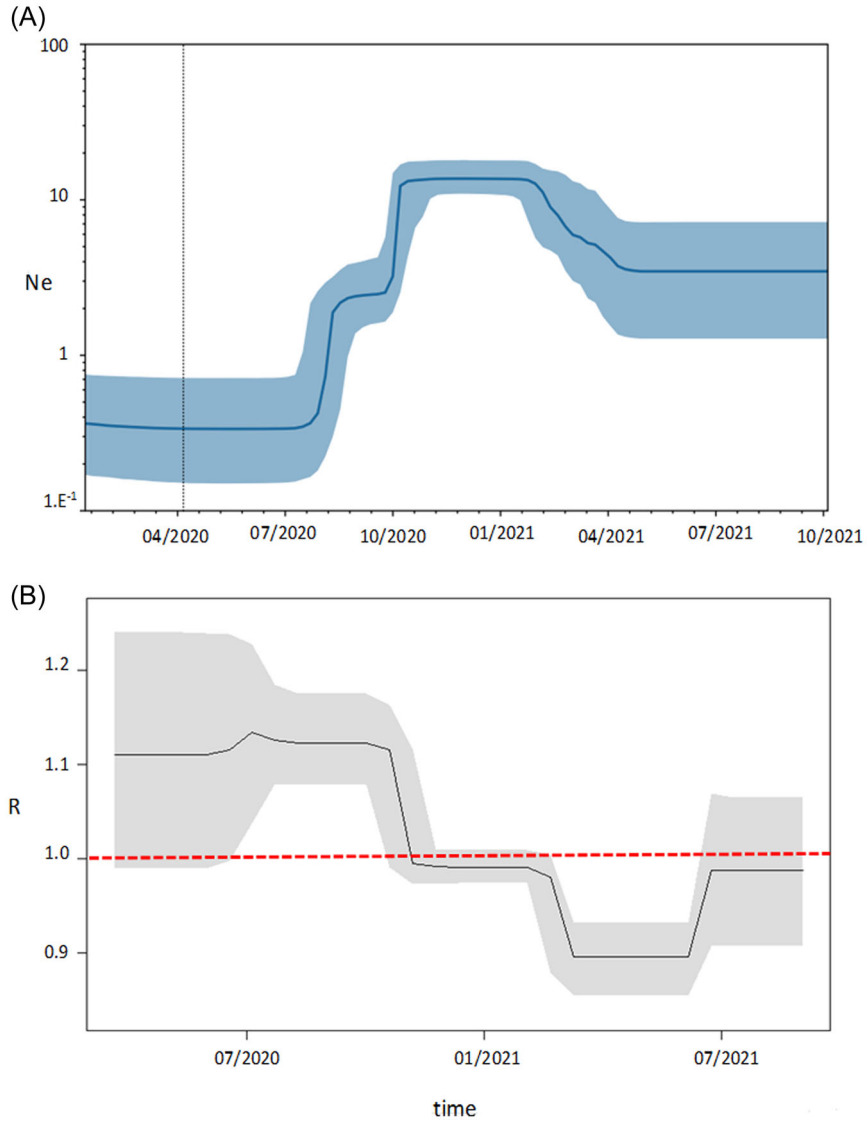

FIGURE 1 (A) Bayesian skyline plot of 20E.EU1 clade. The y-axis indicates the effective population (N_e), the x-axis shows the time expressed in dates. The thick line in the graph indicates the median of the value of the estimate, while the blue area indicates 95% HPD. (B) Birth-death skyline plot of 20E.EU1 clade, in relation to time (x-axis) and the effective reproduction rate (R_e) (y-axis).

The analyzed genomes of SARS-CoV-2 were included in over 100 clusters with a high frequency of mixed clusters (around 70%), which included strains circulating in different regions of the world, suggesting multiple introductions of these lineages/variants in Italy, probably due to international travels.[10]

The fact that the earliest clusters were the largest in size, the most persistent, and most frequently mixed could be related to pandemic containment measures, such as travel restrictions, which were relaxed during the summer 2020 but were then reintroduced in autumn, with the arrival of variants of concern, and gradually relaxed from May 2021. Pure Italian clusters, suggesting a local circulation of the virus, were more prevalent during periods in which restrictive measures were in place while, with the easing of containment measures, they have become increasingly less frequent. This could also be the reason why pure Italian clusters were observed with higher frequency for 20E.EU1 and Alpha variant, compared to Delta variant (<5% of all observed clusters) which circulated in Italy only later, when the restriction measures were largely relaxed.

The analysis of clusters including more than 70% of Italian genomes allowed an estimation of the times of entry and local circulation of the main lineages/variants into Italy. The majority of Italian 20E clusters had tMRCAs in the summer 2020, those Alpha mainly in winter between 2020 and 2021 and Delta clusters between springtime and summer 2021. A partial regional segregation of the isolates was also observed; regardless of the lineage, samples from major Islands formed more frequently clusters than the isolates obtained in other parts of Italy, highlighting the important role of genetic drift in the evolution of the virus.

In support of the important role of genetic drift, many of the identified mutations are those typical of each variant, particularly for Alpha, with only a few minority mutations in the S gene of the 20E lineage, not subject to selective pressure. In contrast, for Delta variant, several mutations that are not typical of the viral genotype have been identified in Italian strains, particularly in the ORF1a gene, confirming the higher genomic heterogeneity of this variant in comparison with the previous one.

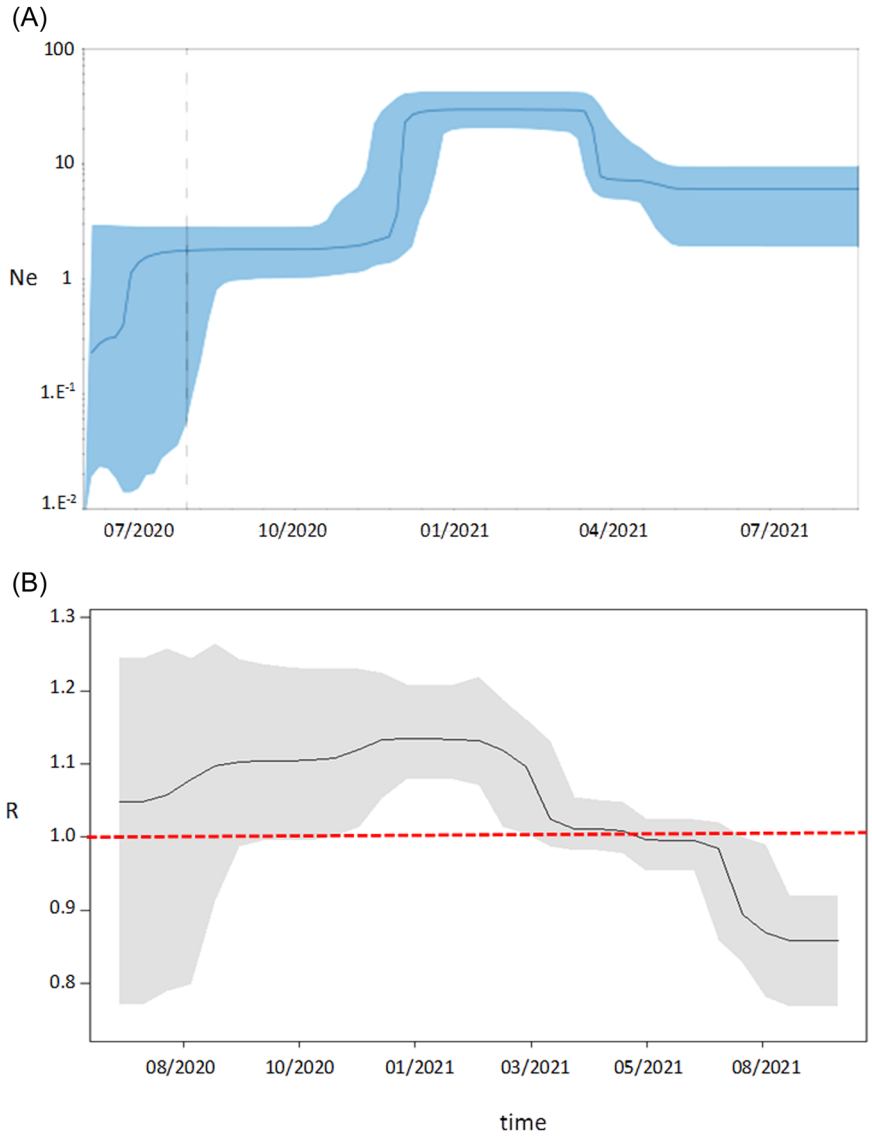

FIGURE 2 (A) Bayesian skyline plot of the Alpha variant. The y-axis indicates the effective population (N_e), the x-axis shows the time expressed in dates. The thick line in the graph indicates the median of the value of the estimate, while the blue area indicates 95% HPD. (B) Birth-death skyline plot of the Alpha variant, in relation to time (x-axis) and the effective reproduction rate (R_e) (y-axis).

Phylodynamic analysis allowed us to estimate the trends in a number of cases and changes in R_e parameter.

For all variants considered, it was observed a temporal correspondence between the trend in the number of infections estimated by the Skyline plot and the R_e estimation by BD skyline. As it is known, during an epidemic disease, R_e changes mainly in proportion to the decrease of number of susceptible subjects. These estimations agreed with the epidemiological data reported by the Italian surveillance (https://www.iss.it/coronavirus/-/asset_publisher/1SRKHcCJJQ7E/content/faq-sul-calcolo-del-rt), in particular those related to hospitalized cases and R_t estimates.

Frequently it was observed the highest values of R_e before the circulation of each VOC in Italy (summer 2020 for 20E, autumn 2020 for Alpha and March 2021 for Delta), probably because the initial R_e values go back to the origin out of Italy of the epidemics. Delta variant showed a R_e value peak higher than those of 20E and Alpha variants, in agreement with other studies[28,29] according to the highest transmissibility of this variant.

The estimated R_e values were apparently lower than the published R_0 estimates for the different variants (https://ibz-shiny.ethz.ch/covid-19-re-international/). This could be due to the persistence in Italy of control measures (such as social distancing, extended use of masks in public, etc.) at least in the first period considered, and the implementation of the vaccination campaign that began in Italy at the end of 2020 and was completed by more than 75% of the population in autumn/winter 2021 (https://ourworldindata.org/). Moreover, only some clusters were considered, representing limited outbreaks, and the R_e estimated is the average value of the clusters considered in the analysis. This value was not necessarily representative of the entire Italian population, considering that only a limited proportion of the cases have been sampled in Italy, particularly before 2021.

A recently published paper,[28] showing R_e estimates obtained on growth rates of VOC genomes stored in public databases from different countries around the world, showed values in Italy similar to those estimated in this study.

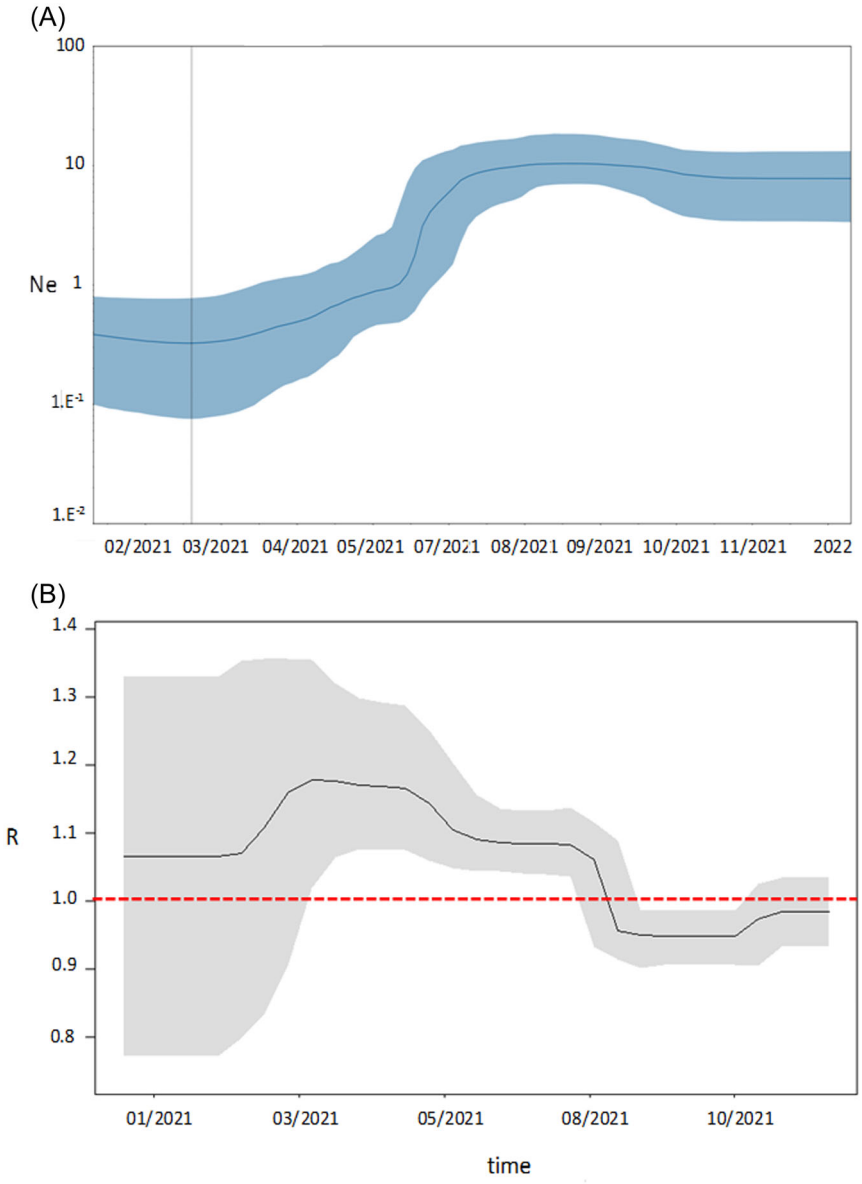

FIGURE 3 (A) Bayesian skyline plot of the Delta variant. The y-axis indicates the effective population (N_e), the x-axis shows the time expressed in dates. The thick line in the graph indicates the median of the value of the estimate, while the blue area indicates 95% HPD. (B) Birth-death skyline plot of the Delta variant, in relation to time (x-axis) and the effective reproduction rate (R_e) (y-axis).

The typical fluctuation of R_e observed during the pandemic may explain the replacement mechanism of the previous variant, with R_e value at its minimum, (lower than 1) by the newly circulating variant, being at its maximum (R_e>1).

In conclusion, these results suggest that the mechanism of replacement of the SARS-CoV-2 lineages and variants must be related to a complex of factors involving the transmissibility of the variant, the seasonality, but also the subsistence of control measures and the level of cross-immunization into the population.

COLLABORATIVE GROUP

SCIRE collaborative Group: Claudia Balotta[1], Spinello Antinori[1], Chiara Resnati[1], Mario Corbellino[1], Stefano Rusconi[1], Massimo De Paschale[21], Valentina Ricucci[2], Federica Stefanelli[2], Nadia Randazzo[2], Giada Garzillo[17], Giorgio Giurato[3], Teresa Rocco[3], Maurizio Fumi[24], Maddalena Schioppa[20], Giovanni Matera[6], Enrico Maria Trecarichi[6], Alessandro Russo[6], Angela Quirino[6], Nadia Marascio[6], Salvatore Rotundo[6], Mario Starace[7], Carmine Minichini[7], Alessia Di Fraia[7], Eugenio Milano[9], Antonella Lagioia[9], Anna Gidari[10], Maurizio Zazzi[11], Lia Fiaschi[11], Massimo Clementi[12], Nicasio Mancini[25], Matteo Castelli[12], Rea Valaperta[14], Ludovica Varisano[14], Giuseppe Nunnari[15], Giuseppe Mancuso[15], Stefano Menzo[16], Carla Acciarri[16], Arianna Miola[22], Valeria Ricci[22], Laura Li Puma[22], Luigi Ruggerone[22]

[23]Unit of Microbiology, Legnano Hospital, ASST Ovest Milanese, Legnano, Italy; massimo.depaschale@asst-ovestmi.it

[24]UOC Patologia Clinica, AO San Pio Benevento, Benevento, Italy; maurizio.fumi@tiscali.it

[25]Laboratory of Medical Microbiology and Virology, Department of Medicine and technological innovation, University of Insubria Varese, Italy; Laboratory of Medical Microbiology and Virology, Fondazione Macchi University Hospital Varese, Italy; nicasio.mancini@uninsubria.it

AUTHOR CONTRIBUTIONS

Annalisa Bergna, Alessia Lai, and Gianguglielmo Zehender conceived and designed the study. Annalisa Bergna, Alessia Lai and Gianguglielmo Zehender analyzed the data and wrote the first draft of the manuscript. Annalisa Bergna, Alessia Lai, Carla Della Ventura, Bianca Bruzzone, Alessandro Weisz, Morena d'Avenia, Sophie Testa, Caterina Sagnelli, Angela Menchise, Gaetano Brindicci, Daniela Francisci, Ilaria Vicenti, Nicola Clementi, Annapaola Callegaro, Emmanuele Venanzi Rullo, Sara Caucci, Vanessa De Pace, Andrea Orsi, Stefano Brusa, Francesca Greco, Vittoria Letizia, Emilia Vaccaro, Gianluigi Franci, Francesca Rizzo, Fabio Sagradi, Leonardo Lanfranchi, and Michela Sampaolo performed and analyzed whole genomes and collected data. Annalisa Bergna, Alessia Lai, Silvia Ronchiadin, and Gianguglielmo Zehender analyzed the data and performed phylogenetic analyses. Carlo Torti, Nicola Coppola, Annalisa Saracino, Massimo Galli, Agostino Riva collected data and participated in manuscript revision. Massimo Galli, Agostino Riva, and Gianguglielmo Zehender provided study oversight. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the UNIMI GSA-IDEA project. This project has received funding from the European Union's Horizon Europe Research and Innovation Actions under grant no. 101046041, and from the Ministero dell'Università e della Ricerca (PRIN 202022GZEHE_01), from EU funding within the NextGeneration EU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. PE00000007, INF-ACT) and from Regione Campania (grants: 'Monitoring the spread and genomic variability of the Covid 19 virus in Campania using NGS technology', POR Campania FESR 2014/2020, CUP: B14I20001980006, and 'GENOMAeSALUTE', POR Campania FESR 2014/2020, azione 1.5; CUP: B41C17000080007).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. Whole-genome sequences were submitted to GISAID. Datasets and performed analyses are available upon request.

ORCID

Annalisa Bergna http://orcid.org/0000-0003-1653-1917
Alessia Lai http://orcid.org/0000-0002-3174-5721
Carlo Torti http://orcid.org/0000-0001-7631-5453
Ilaria Vicenti http://orcid.org/0000-0002-4306-2960
Nicola Clementi http://orcid.org/0000-0002-1822-9861
Nicola Coppola http://orcid.org/0000-0001-5897-4949
Agostino Riva http://orcid.org/0000-0003-3171-3049
Gianguglielmo Zehender http://orcid.org/0000-0002-1886-2915

REFERENCES

1. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. *J Virol*. 2010;84:9733-9748.
2. Worobey M, Pekar J, Larsen BB, et al. The emergence of SARS-CoV-2 in Europe and North America. *Science*. 2020;370:564-570.
3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Nat Acad Sci*. 2020;117:9241-9243.
4. Balloux F, Tan C, Swadling L, et al. The past, current and future epidemiological dynamic of SARS-CoV- 2. *Oxford Open immunol*. 2022;3:iqac003.
5. Alicandro G, Remuzzi G, La Vecchia C. COVID-19 pandemic and total mortality in the first six months of 2020 in Italy. *Med Lav*. 2020;111:351-353.
6. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? *The Lancet*. 2020;395:1225-1228.
7. Usuelli M. The Lombardy region of Italy launches the first investigative COVID-19 commission. *The Lancet*. 2020;396:e86-e87.
8. Alteri C, Cento V, Piralla A, 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;12:434.
9. Stefanelli P, Faggioni G, Lo Presti A, 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. *Euro Surveill*. 2020;25:2000305.
10. Lai A, Bergna A, Toppo S, et al. Phylogeography and genomic epidemiology of SARS-CoV-2 in Italy and Europe with newly characterized Italian genomes between February-June 2020. *Sci Rep*. 2022;12:5736.
11. Lai A, Bergna A, Caucci S, et al. Molecular tracing of SARS-CoV-2 in Italy in the first three months of the epidemic. *Viruses*. 2020;12:798.
12. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182:812-827.e19.
13. Hodcroft EB, Zuber M, Nadeau S, et al Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*. 2021;595:707-712.
14. Lai A, Bergna A, Menzo S, et al. Circulating SARS-CoV-2 variants in Italy, October 2020-March 2021. *Virol J*. 2021;18:168.
15. Lai A, Bergna A, Della Ventura C, et al. Epidemiological and clinical features of SARS-CoV-2 variants circulating between April-December 2021 in Italy. *Viruses*. 2022;14:2508.
16. Focosi D, Maggi F, McConnell S, Casadevall A. Spike mutations in SARS-CoV-2 AY sublineages of the Delta variant of concern: implications for the future of the pandemic. *Future Microbiol*. 2022;17:219-221.
17. Billah MA, Miah MM, Khan MN. Reproductive number of coronavirus: a systematic review and meta- analysis based on global level evidence. *PLoS One*. 2020;15:e0242128.
18. Adam DC, Wu P, Wong JY, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Med*. 2020;26:1714-1719.
19. Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*. 2019;4:10-19.
20. Rife BD, Mavian C, Chen X, et al. Phylodynamic applications in 21(st) century global infectious disease research. *Glob Health Res Policy*. 2017;2:13.
21. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat Rev Genet*. 2022;23:547-562.

22. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the Genomic Era. *Mol Biol Evol*. 2020;37:1530-1534.
23. Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2019;15:e1006650.
24. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol*. 2018;67:901-904.
25. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*. 2013;110:228-233.
26. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus- infected pneumonia. *N Engl J Med*. 2020;382:1199-1207.
27. Walker PR, Pybus OG, Rambaut A, Holmes EC. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol*. 2005;5:199-208.
28. Manathunga SS, Abeyagunawardena IA, Dharmaratne SD. A comparison of transmissibility of SARS- CoV-2 variants of concern. *Virol J*. 2023;20:59.
29. Suzuki R, Yamasoba D, Kimura I, et al. Attenuated fusogenicity and pathogenicity of SARS-CoV-2 omicron variant. *Nature*. 2022;603:700-705.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bergna A, Lai A, Ventura CD, et al. Genomic epidemiology of the main SARS-CoV-2 variants in Italy between summer 2020 and winter 2021. *J Med Virol*. 2023;95:e29193. doi:10.1002/jmv.29193