**Genomic epidemiology of the main SARS-CoV-2 variants circulating in Italy in 2020 and 2022 period**

Annalisa Bergna1, Alessia Lai1\*, Carla Della Ventura1, Bianca Bruzzone2, Alessandro Weisz3, Morena d'Avenia4, Sophie Testa5, Carlo Torti6, Caterina Sagnelli7, Angela Menchise 8, Gaetano Brindicci9, Daniela Francisci10, Ilaria Vicenti11, Nicola Clementi12,13, Annapaola Callegaro14, Emmanuele Venanzi Rullo15, Sara Caucci16, Vanessa De Pace2, Andrea Orsi2,17, Stefano Brusa18, Francesca Greco19, Vittoria Letizia20, Emilia Vaccaro21, Gianluigi Franci3, Francesca Rizzo3, Fabio Sagradi5, Leonardo Lanfranchi5, Nicola Coppola7, Annalisa Saracino9, Michela Sampaolo13, Silvia Ronchiadin22, Massimo Galli1, Agostino Riva1, Gianguglielmo Zehender1, and SARS-CoV-2 ITALIAN RESEARCH ENTERPRISE – (SCIRE) collaborative Group\*\*

1 Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy; annalisa.bergna@unimi.it, alessia.lai@unimi.it, carla.dellaventura@unimi.it, massimo.galli@unimi.it, agostino.riva@unimi.it, gianguglielmo.zehender@unimi.it

2 Hygiene Unit, IRCCS AOU San Martino-IST, Genoa, Italy; bianca.bruzzone@hsanmartino.it, Vanessa.DePace@hsanmartino.it, andrea.orsi@unige.it

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; aweisz@unisa.it, gfranci@unisa.it, frizzo@unisa.it

4 UOSVD of Cytopathology and Screening, Department of Laboratory Medicines, Ospedale di Venere, Asl Bari, Bari, Italy; morena.davenia@asl.bari.it

5 Unit of Infectious Diseases, Azienda Socio Sanitaria Territoriale Cremona, Cremona, Italy; sophie.testa@asst-cremona.it, fabio.sagradi@asst-cremona.it, leonardo.lanfranchi@asst-cremona.it

6 Infectious and Tropical Disease Unit, Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy; torti@unicz.it

7 Department of Mental Health and Public Medicine, University of Campania "Luigi Vanvitelli", Naples, Italy; caterina.sagnelli@unicampania.it, nicola.coppola@unicampania.it

8 Microbiology and Virology Laboratory, A.O.R. San Carlo Potenza, Potenza, Italy; angela.menchise@ospedalesancarlo.it

9 Infectious Diseases Unit, University of Bari, Bari, Italy; gaetanobrindicci@gmail.com, annalisa.saracino@uniba.it

10 Department of Medicine and Surgery, Clinic of Infectious Diseases, "Santa Maria della Misericordia" Hospital, University of Perugia, Perugia, Italy; daniela.francisci@unipg.it

11 Department of Medical Biotechnologies, University of Siena, Siena, Italy; ilariavicenti@gmail.com

12 Laboratory of Microbiology and Virology, Università "Vita-Salute" San Raffaele, Milan, Italy; clementi.nicola@hsr.it

13 IRCCS Ospedale San Raffaele, Milan, Italy; clementi.nicola@hsr.it , sampaolo.michela@hsr.it

14 Ospedali Riuniti, Department of Infectious Diseases, Bergamo, Italy; anna.callegaro@asst-bergamoest.it 15 Unit of Infectious Diseases, Department of Experimental and Clinical Medicine, University of Messina, Messina, Italy; evenanzirullo@gmail.com

16 Department of Biomedical Sciences and Public Health, Virology Unit, Polytechnic University of Marche, Ancona, Italy; s.caucci@staff.univpm.it

17 Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy; andrea.orsi@unige.it

18 Department of Translational Medical Sciences, Federico II University, Naples, Italy; stefano.brusa3@hotmail.it

19 UOC Microbiology and Virology, PO Cosenza, Cosenza, Italy; francesca.greco@aocs.it

20 UOSD Genetic and Molecular Biology, AORN Sant’Anna and San Sebastiano di Caserta, Caserta, Italy; vittoria.letizia@alice.it

21 Molecular Biology Units, AOU ‘S. Giovanni di Dio e Ruggi d'Aragona’ Università di Salerno, Salerno, Italy; emilia.vaccaro@sangiovannieruggi.it

22 Intesa Sanpaolo Innovation Center- Artificial Intelligence Lab, Turin, Italy; silvia.ronchiadin@intesasanpaolo.com

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

\*Corresponding author: Dr. Alessia Lai, Department of Biomedical and Clinical Sciences, University of Milan, via G.B. Grassi 74, 20157, Milan; +390250319775, alessia.lai@unimi.it

RUNNING HEAD: **20E.EU1, Alpha and Delta variants dynamics**

**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, two datasets were evaluated for each variant, the former comprising international genomes and the latter focusing on clusters containing at least 70% of Italian sequences.

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.

International clusters presented tMRCAs between 13/06/2020-27/07/2021. Re values showed the highest level between May-June until autumn 2020 in 20E.EU1 clade. The Alpha variant showed an increase in the Re in December 2020, when the highest mean value was estimated. Delta variant presented two peaks: the first between March-May, and the second between June-July 2021.

The present work highlighted a different evolutionary dynamic of studied lineages with a 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.

**KEY WORDS**

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

**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](#_ENREF_1) and driven by the widespread and rapid circulation of the virus in the human population.[2](#_ENREF_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](#_ENREF_3),[4](#_ENREF_4)

At the beginning of the pandemic, the ancestral lineages (in particular, the most widespread B lineage) entered Italy multiple times remaining confined to a few regions,[5](#_ENREF_5),[6](#_ENREF_6) being then quickly replaced by the B.1 lineage, characterized by the amino acid mutation D614G in the Spike protein. This lineage rapidly spread from Lombardy throughout the whole Italy and other European countries since the end of February and March 2020,[6](#_ENREF_6) quickly becoming the dominant lineage throughout the world.[7](#_ENREF_7) 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.[8](#_ENREF_8)

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.

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.[9](#_ENREF_9)

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 (Re or Rt, respectively), which represents the mean number of secondary infections due to an infected subject in a population not entirely susceptible and in the presence of containing measures (https://royalsociety.org/-/media/policy/projects/set-c/set-covid-19-R-estimates.pdf).[10](#_ENREF_10)

Although the real meaning of Rt 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.[11](#_ENREF_11)

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.[12](#_ENREF_12),[13](#_ENREF_13) 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.[14](#_ENREF_14)

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**

Among sequences collected between August 2020 and January 2022, it were selected 602 new whole genome sequences (WGS) belonging to the 20E.EU1 clade and the Alpha (B.1.1.7), and Delta variants (B.1.617.2). Samples were obtained by more than 74 clinical centers and laboratories located throughout Italy and participating in the collaborative SCIRE group. Viral RNA samples were collected from nasopharyngeal swabs of SARS-CoV-2-positive subjects including hospitalized subjects and outpatients. All participants 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**

For each selected variant (20E.EU1 clade, Alpha, and Delta) two different datasets were constructed, one Italian and one international. The Italian datasets combined Italian genomes characterized by SCIRE group with SARS-CoV-2 genomes obtained from GISAID (https://gisaid.org) selected according to the following criteria: 10 whole genomes for each Italian 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 with more than 5% of gap.

For the international datasets, 5 genomes for European and non-European countries and sampling month were included according to the criteria adopted for Italian dataset. The compositions of the national and international datasets for the variants under study are summarized in **Tables S1** and **S2**.

The sequences included in the Italian and international datasets were collected between August 2020 and October 2021, November 2020 and August 2021, and May 2021 and January 2022 for 20E.EU1 clade, Alpha and Delta variants, respectively.

Alignment of multiple sequences was obtained using Nextclade and edited by BioEdit v. 7.2.6.1 program (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Sequences were trimmed at the ends to obtain genomes of the same length (29,774 bp).

**2.4 Phylogenetic analysis**

The maximum likelihood trees of both the Italian and international datasets were estimated using IQ-TREE v. 1.6.12 (http://www.iqtree.org/)[15](#_ENREF_15) 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).

Italian clusters (including more than 2 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. In the international datasets, 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. Clusters including more than 30 sequences from Italian datasets and clusters characterized by >70% of Italian sequences from international datasets, were used for the Bayesian analyses.

**2.5 Phylodynamic analysis**

To characterize the epidemiological and evolutionary history of the different SARS-CoV-2 studied lineages, the coalescence and the birth-death models were used.

Bayesian analysis was performed by BEAST v. 2.7 (https://beast.community/)[16](#_ENREF_16) the same substitution model and molecular clock employed for the previously described analyses. The evolutionary rate was set at 8\*10-4 substitutions/site/year using a relaxed clock under the exponential coalescent model as transformed into units per year.

MCMC analyses were run for 100 million generations and sampled every 3,000. 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/),[17](#_ENREF_17) accepting ESS of at least 200. The uncertainty of estimates was indicated with 95% highest prior density (HPD) intervals.

The final tree is 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 2.48 was used to infer changes in the effective reproductive number (Re), 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).[18](#_ENREF_18) Given that the samples were collected during a short period of time, a “birth-death serial” model was used.

For the birth-death analysis, one and two intervals and a lognormal prior to Re, with a mean (M) of 0.0 and a variance (S) of 1.5 were chosen, which allows the Re 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. [19](#_ENREF_19) 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 = λ − δ), and the doubling time was estimated by the equation: doubling time = ln(2)/r.[20](#_ENREF_20)

**3 RESULTS**

**3.1 Mutation analyses of Italians dataset**

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 (**Table S3**). Almost all the sequences had the distinctive mutations of this lineage such as A220V (97.4%, n=1,043) in the N gene, A222V (97.8%, n=1,048) in the S gene, and V30L (99.3%, n=1,064) and L67F (99.3%, n=1,064) 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 20A/S:98F clade, while one sequence carried the D80Y mutation characteristic of the 20C/S:80Y clade.

In the Alpha variant dataset, only mutations/deletions typical of this lineage have been found, as shown in **Table S4**, with the only exception of the two consecutive mutations in the N gene, R203K (99.1%, n=1,312) and G204R (96.8%, n=1,281), 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 (**Table S5**).

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 Italian clusters in the international framework**

***3.2.1 20E.EU1 clade***

The phylogenetic analysis conducted using *Maximum Likelihood* approach on the international dataset of 20E.EU1 showed that 13.2% (321/2,425) of sequences were dispersed in the tree. Among the 255 identified clusters, 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**).

**Figure 1** shows the dated tree obtained by Bayesian analysis on 11 Italian clusters (5 mixed including more than 70% of Italian genomes and 6 pure Italian). Mean tMRCAs spanned a period from June to September 2020 (**Table S6**). 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=.01*) and were predominantly mixed, the late clusters, with tMRCA between August and September 2020, had the shortest duration (7-14 months) and the smallest size (12 median genomes) and were predominantly Italians.

Considering sampling locations (**Table S6**), 5 clusters included more than 50% of sequences from Northern Italy, 4 clusters from Southern Italy and 1 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=.003*).

***3.2.2 Alpha Variant***

A total of 467 clusters were found in Alpha variant dataset. Of 3,174 sequences, 892 (28.1%) were included in 40 clusters containing at least 1 Italian sequence (from 11 to 65). Twenty-four clusters (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 2 mixed, were analyzed for tMRCA estimations (**Figure 2, Table S7**). Mean tMRCAs of clusters was between November 2020 and late February 2021, with the majority (6/17, 35.3%) originating in December 2020. 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 of 5 months for the later ones (*p=.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=.009*).

***3.2.3 Delta Variant***

Phylogenetic analysis of Delta international dataset showed a total of 364 clusters of which 42 (n=1,536 isolates) containing at least one Italian sequence, classified as follows: 4 singleton (9.5%), 36 mixed (85.7%) and 2 pure Italian (4.8%) (**Table 1**).

Clusters involved 50%, 40.7% and 9.3% of Italian, European and international sequences.

The median size of 9 clusters including >70% of Italians was 16 isolates (11-79). Among these, the largest two clusters (#48 and #68), originated in March 2021, encompassed 62% (n=142) of all genomes and showed a median duration of 10 months *versus* 7 months for the other clusters (*p=.03*) (**Figure 3**, **Table S8**). Seven clusters (87.5%) 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. Cluster #193 contained a high proportion of sequences from Northern (44.4%) and Central (55.5%) Italy.

Sequences from the Islands were more frequently involved in clusters than genomes from Southern, Northern and Central Italy (60% *vs.* 54.2%, 49.4%, 47.9%; *p=.0002*).

**3.3 Phylodynamics of SARS-CoV-2 lineages and VOCs**

The skyline plot of 20E.EU1 dataset (Figure 4a) showed a rapid increase of the effective number of infections (Ne) 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.

The Re estimates (Figure 4b) showed a higher value since the beginning of the 20E circulation, around July 2020 (peak of Re value=1.125, *95%HPD:* 1.08-1.18). The Re value started to decline in autumn 2021 (October 2021), falling around 1 between December 2020-January 2021 and decreasing below 1 in April 2020.

The number of cases infected by Alpha variant rapidly increased in December 2020, reaching a plateau during the winter 2021. A decrease in infections was observed from April 2021 (**Figure 5a**).

An initial value of Re above 1 was observed in the summer-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 2021reaching values below the unit in May 2021 (**Figure 5b**).

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 6a**).

Similarly, the Re value (**Figure 6b**) 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**

At the beginning of the COVID-19 pandemic, Italy experienced a wide epidemic wave whose actual size and extent (at least in terms of the real number of infections) remain largely unknown but that was clearly demonstrated by numbers of hospitalizations and deaths. This was due to the lack of an effective surveillance system and limited availability of testing in these early times.[21](#_ENREF_21),[22](#_ENREF_22),[23](#_ENREF_23)

The genomic epidemiology studies showed that after the entry in Northern Italy of lineage B.1, the first SARS-CoV-2 variant at increased transmissibility,[7](#_ENREF_7) in late January or early February 2020,[24](#_ENREF_24) several descendent lineages widely spread to the whole Italy and worldwide.[25](#_ENREF_25),[6](#_ENREF_6) The size of this first wave of the Italian epidemic decreased from May 2020 and remained low until August, when the number of COVID-19 cases began to rise again, probably driven by the lineage 20E circulating in Europe from the early summer 2020, which caused a second wave of pandemic in Italy in autumn 2020, the most intense in terms of number of cases and deaths reported in the pre-Omicron era.

At the end of 2020 the first variant of concern (VOC) Alpha entered Italy and quickly replaced lineage 20E.EU1,[26](#_ENREF_26) causing a further wave reaching its peak in March-April 2021. Subsequently, it was rapidly replaced by the VOC Delta, which caused a resurgence of cases during the summer 2021 (between July and August 2021), and the largest epidemic in the autumn 2021, immediately before the rapid spread of the VOC Omicron which became prevalent in January 2022.[27](#_ENREF_27)

Since the early months of the COVID-19 pandemic in Italy, a large number of clinical centers for patient management and laboratories capable of performing genomic characterization of the virus, distributed throughout all regions of Italy, formed the SCIRE network with the aim of contributing to the common effort of obtaining the highest possible number of complete SARS-CoV-2 genomes.

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.

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.[6](#_ENREF_6)

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 of 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 abandoned.

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, probably due to the founder effect.

In support of the determining 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, particularly in the ORF1a gene.

Phylodynamic analysis allowed us to estimate the trends in a number of cases and changes in Re 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 Re estimation by BD skyline. As it is known, during an epidemic disease, Re changes mainly in proportion to the number of subjects who have immunized or died. When the number of infections increased, the Re value was higher than 1 becoming lower than 1 when the number of infections started to decrease. These estimations were in agreement 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 Rt estimates.

Frequently was observed the highest values of Re 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 Re values go back to the origin out of Italy of the epidemics. Delta variant showed a Re value peak higher than those of 20E and Alpha variants, in agreement with other studies[28](#_ENREF_28),[29](#_ENREF_29) according to the highest transmissibility of this variant.

The estimated Re values were apparently lower than the published R0 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 Re 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](#_ENREF_28) showing Re estimates obtained on growth rates of VOC genomes stored in public databases from different countries around the world, showed similar values to those estimated in this study.

The typical fluctuation of Rt observed during the pandemic related to the dynamics may explain the replacement mechanism of the previous variant (with Re value at its minimum, lower than 1) by the newly circulating variant (being at its maximum Re>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 season, but also the subsistence of control measures and the level of cross-immunization into the population.

**FIGURE LEGENDS**

**Figure 1**. Bayesian phylogenetic tree of the International 20E.EU1 Italian clusters. The different size of the circles on the nodes indicates the posterior probability between 1 and 0.9. Different clusters are represented with different colors, numbers of clusters are reported. The scale below the tree represents the time expressed in fractions of a year.

**Figure 2**. Bayesian phylogenetic tree of the International Alpha Italian clusters. The different size of the circles on the nodes indicates the posterior probability between 1 and 0.9. Different clusters are represented with different colors, numbers of clusters are reported. The scale below the tree represents the time expressed in fractions of a year.

**Figure 3.** Bayesian phylogenetic tree of the International Delta Italian clusters. The different size of the circles on the nodes indicates the posterior probability between 1 and 0.9. Different clusters are represented with different colors, numbers of clusters are reported. The scale below the tree represents the time expressed in fractions of a year.

**Figure 4.** (**a**) Bayesian skyline plot of 20E.EU1 clade. The y-axis indicates the effective population (Ne), 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 (Re) (y-axis).

**Figure 5.** (**a**) Bayesian skyline plot of the Alpha variant. The y-axis indicates the effective population (Ne), 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 (Re) (y-axis).

**Figure 6.** (**a**) Bayesian skyline plot of the Delta variant. The y-axis indicates the effective population (Ne), 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 (Re) (y-axis).

**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.

**COLLABORATIVE GROUP**

SCIRE collaborative Group: Claudia Balotta1, Spinello Antinori1, Chiara Resnati1, Mario Corbellino1, Stefano Rusconi1, Massimo De Paschale21, Valentina Ricucci2, Federica Stefanelli2, Nadia Randazzo2, Giada Garzillo17, Giorgio Giurato3, Teresa Rocco3, Maurizio Fumi24, Maddalena Schioppa20, Giovanni Matera6, Enrico Maria Trecarichi6, Alessandro Russo6, Angela Quirino6, Nadia Marascio6, Salvatore Rotundo6, Mario Starace7, Carmine Minichini7, Alessia Di Fraia7, Eugenio Milano9, Antonella Lagioia9, Anna Gidari10, Maurizio Zazzi11, Lia Fiaschi11, Massimo Clementi12, Nicasio Mancini25, Matteo Castelli12, Rea Valaperta14, Ludovica Varisano14, Giuseppe Nunnari15, Giuseppe Mancuso15, Stefano Menzo16, Carla Acciarri16, Arianna Miola22, Valeria Ricci22, Laura Li Puma22, Luigi Ruggerone22

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

**FUNDING**

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) 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).

**CONFLICTS OF INTEREST STATEMENT**

The authors declare no conflict of interest.

**DATA AVAILABILITY STATEMENT**

Whole-genome sequences were submitted to GISAID.

**REFERENCES**

1. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. *Journal of virology.* 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 (New York, NY).* 2020;370:564-570.

3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America.* 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 immunology.* 2022;3:iqac003.

5. 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 surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin.* 2020;25.

6. 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. *Scientific reports.* 2022;12:5736.

7. 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 e819.

8. 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.

9. 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 microbiology.* 2022;17:219-221.

10. 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.

11. Adam DC, Wu P, Wong JY, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine.* 2020;26:1714-1719.

12. Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. *Nature microbiology.* 2019;4:10-19.

13. Rife BD, Mavian C, Chen X, et al. Phylodynamic applications in 21(st) century global infectious disease research. *Global health research and policy.* 2017;2:13.

14. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature reviews Genetics.* 2022;23:547-562.

15. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular biology and evolution.* 2020;37:1530-1534.

16. Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology.* 2019;15:e1006650.

17. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic biology.* 2018;67:901-904.

18. 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 of the United States of America.* 2013;110:228-233.

19. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *The New England journal of medicine.* 2020;382:1199-1207.

20. 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. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases.* 2005;5:199-208.

21. Alicandro G, Remuzzi G, La Vecchia C. COVID-19 pandemic and total mortality in the first six months of 2020 in Italy. *La Medicina del lavoro.* 2020;111:351-353.

22. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? *Lancet (London, England).* 2020;395:1225-1228.

23. Usuelli M. The Lombardy region of Italy launches the first investigative COVID-19 commission. *Lancet (London, England).* 2020;396:e86-e87.

24. 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.

25. 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. *Nature communications.* 2021;12:434.

26. Lai A, Bergna A, Menzo S, et al. Circulating SARS-CoV-2 variants in Italy, October 2020-March 2021. *Virology journal.* 2021;18:168.

27. 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.

28. Manathunga SS, Abeyagunawardena IA, Dharmaratne SD. A comparison of transmissibility of SARS-CoV-2 variants of concern. *Virology journal.* 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.