SARS-CoV-2 shifting transmission dynamics and hidden reservoirs limited the efficacy of public health interventions in Italy

Marta Giovanetti[1,2,3]*, Eleonora Cella[4]*, Francesca Benedetti[5]*, Brittany Rife Magalis[6]*, Vagner Fonseca[2,7,8], Silvia Fabris[3], Giovanni Campisi[9], Alessandra Ciccozzi[3], Silvia Angeletti[10], Alessandra Borsetti[11], Vittoradolfo Tambone[12], Caterina Sagnelli[13], Stefano Pascarella[14], Alberto Riva[15], Giancarlo Ceccarelli[16], Alessandro Marcello[17], Taj Azarian[4], Eduan Wilkinson[7], Tulio de Oliveira[7], Luiz Carlos Junior Alcantara[1,2], Roberto Cauda[18], Arnaldo Caruso[9], Natalie E Dean[19], Cameron Browne[20], Jose Lourenco[21], Marco Salemi[6^], Davide Zella[5^], Massimo Ciccozzi[3^]

[1]Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil; [2]Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [3]Medical Statistic and Molecular Epidemiology Unit, University of Biomedical Campus, Rome, Italy; [4]Burnett School of Biomedical Sciences, University of Central Florida, Orlando (Florida, USA); [5]Institute of Human Virology, Department of Biochemistry and Molecular Biology, School of Medicine, University of Maryland (Baltimore, USA); [6]Emerging Pathogens Institute & Department of Pathology, College of Medicine, University of Florida, Gainesville, FL 32610, USA; [7]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa; [8]Coordenação Geral dos Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, (CGLAB/SVS-MS) Brasília, Distrito Federal 70719-040, Brazil; [9]Department of Molecular and Translational Medicine, Section of Microbiology, University of Brescia, Brescia, Italy; [10]Unit of Clinical Laboratory Science, University Campus Bio-Medico of Rome, Rome, Italy; [11]National HIV/AIDS Researh Center, Istituto Superiore di Sanità, Rome, Italy; [12]Anthropology, and Applied Ethics, Campus Bio-Medico University, Rome, Italy; [13]Department of Mental Health and Public Medicine, University of Campania "Luigi Vanvitelli", Naples, Italy; [14]Department of Biochemical Sciences "A. Rossi Fanelli", University of Rome “La Sapienza”, Rome, Italy; [15]ICBR, University of Florida, Gainesville, FL 32610, USA; [16] Department of Public Health and Infectious Diseases, Policlinico Umberto I Università ‘Sapienza’, Rome; [17]Laboratory of Molecular Virology, International Centre for Genetic Engineering and Biotechnology

(ICGEB), Trieste, Italy; [18]Department Infectious Diseases, – Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Rome, Italy; [19] Department of Epidemiology, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610, USA; [20] Department of Mathematics, University of Lafayette, LA, USA; [21]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.

*These authors contributed equally to this article.

^Correspondence and requests for materials should be addressed to

m.ciccozzi@unicampus.it; Dzella@ihv.umaryland.edu; salemi@pathology.ufl.edu

Abstract

We investigated SARS-CoV-2 transmission dynamics in Italy, one of the countries hit hardest by the pandemic, using phylodynamic analysis of viral genetic and epidemiological data. We observed the co-circulation of at least 13 different SARS-CoV-2 lineages over time, which were linked to multiple importations and characterized by large transmission clusters concomitant with a high number of infections. Subsequent implementation of a three-phase nationwide lockdown strategy greatly reduced infection numbers and hospitalizations. Yet we present evidence of sustained viral spread among sporadic clusters acting as "hidden reservoirs" during summer 2020. Mathematical modelling shows that increased mobility among residents eventually catalyzed the coalescence of such clusters, thus driving up the number of infections and initiating a new epidemic wave. Our results suggest that the efficacy of public health interventions is, ultimately, limited by the size and structure of epidemic reservoirs, which may warrant prioritization during vaccine deployment.

Keywords: SARS-CoV-2; Italy; pandemic; genomic epidemiology.

Main text

On December 31st 2019, the World Health Organization (WHO) China Country Office was informed of pneumonia cases of unknown aetiology detected in Wuhan City, Hubei Province[1,2]. By January 11th – 12th 2020, Chinese authorities identified a novel single stranded, positive-sense enveloped RNA *Betacoronavirus*, with genome of 30,000 nucleotides in length, belonging to the *Coronaviridae* family, related to the severe acute respiratory syndrome coronavirus (SARS-CoV) that caused a global outbreak in 2002–2004 [3]. Initially named nCoV-2019 (novel Coronavirus 2019), the virus likely emerged from several recombination events in bats and pangolins [4], and was subsequently introduced in the human population through zoonotic transmissions [1,5]; it was later renamed SARS-CoV-2, and recognized as the etiologic agent of Coronavirus Disease 2019 (COVID-19) [6]. Epidemiological investigations and phylogenetic analysis promptly confirmed airborne SARS-CoV-2 human-to-human transmission [3,7]. Following its worldwide spread, the WHO declared the outbreak as a Public Health Emergency of International Concern on January 30th, 2020, and a pandemic on March 11th, 2020. As of December 16th, 2020, SARS-CoV-2 has spread to 216 countries with nearly 74 million confirmed cases and over 1.6 million fatalities [8].

Italy was one of the first and most affected countries in the world. By October 31st 2020, the Italian Ministry of Health and the Civil Protection Department reported 1.38 million total SARS-CoV-2-related cases, and 49,261 deaths [9]. The first confirmed imported cases dated back to January 30th 2020 when two tourists from Wuhan, China, were tested positive for SARS-CoV-2 in Rome (**Figure 1A**). On February 17th 2020, the Italian government confirmed the first locally acquired case in a small city in Northern Italy (Codogno, Lombardy region) [10]. Three days later, the first COVID-19-related death in Italy, a 78-year old male, was reported in the city of Padova. As the epidemic quickly spread throughout the country, establishing Italy as one of the major SARS-CoV-2 hotspots [11], the Italian government declared a Public Health Emergency of National Importance, enabling the introduction of restriction measures to limit new infections [12]. In the effort to flatten the epidemic curve, Phase I lockdown measures were first introduced on March 7th – 8th 2020 in

11 municipalities of Northern Italy, where most cases had occurred, and extended by March 11th to the whole country (**Figure 1A**). Described as the largest lockdown in the history of Europe [13], citizen mobility was restricted, except for "well grounded" work- or health-related reasons. A universal mask mandate was required at all times outdoors. Schools, university activities, public/cultural events, and sport competitions were also suspended nationwide, as well as non-essential commercial activities. Borders with other states were closed, and within the country public transport was limited or shut down.

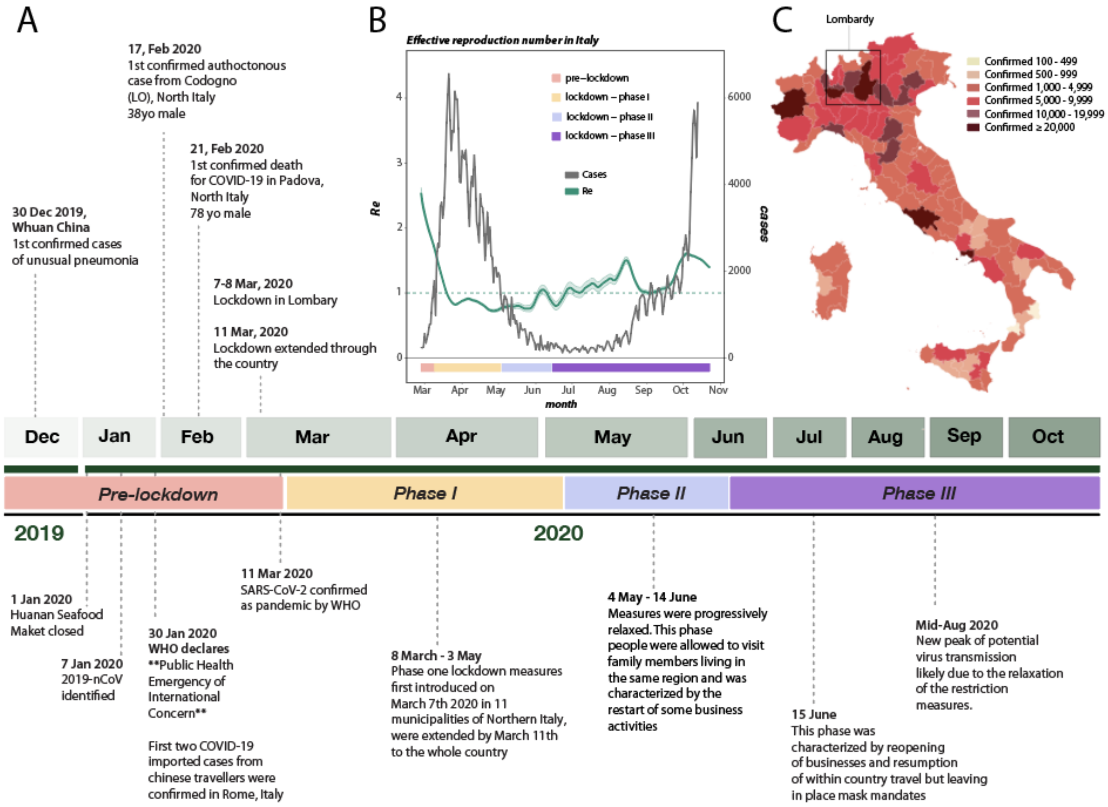


Figure 1. History of SARS-CoV-2 epidemic in Italy. A) Timeline of key events following the first confirmed cases of SARS-CoV-2 infection in Italy. **B**) Epidemic curve showing the progression of reported daily viral infection numbers in Italy from the beginning of the epidemic in March (black) and changes in *Re* estimations in the same period (green), with lockdown phases indicated along the bottom. **C**) Map of cumulative SARS-CoV-2 cases per 100,000 inhabitants in Italy up to Oct 2020.

As daily viral infection numbers decreased, public health measures were progressively relaxed through a Phase II (May 4th), which allowed visits to family members living in the same region and the restart of some business activities, and a Phase III (June 15th), which allowed reopening of businesses and resumption of within country travel, but left in place mask mandates and bans on large-scale meetings. A significant slowdown in the number of infections since the beginning of May 2020 (**Figure 1B**) validated the effectiveness of Phase I restrictions.

After a period of seemingly stable epidemic recession, with very few new cases detected between June-August, a new epidemic wave hit the country, resulting in higher incidence than before. Superimposition of the reported epidemic curve and dynamic estimates of the effective reproduction number, *Re*, throughout the three major periods (first wave, recess, second wave) of the Italian epidemic, revealed an interesting pattern (**Figure 1 B**). *Re* provides a measure of the average number of secondary infections caused by a single infected person: a growing epidemic is typically characterized by *Re* > 1, while *Re* < 1 indicates no growth. As expected, *Re* values were estimated to be > 2 at the beginning of SARS-CoV-2 exponential spread in Italy, and quickly fell to values < 1 after the start of Phase one lockdown measures. Yet, between end of June and end of August, through Phase II and III lockdowns, *Re* values showed an oscillating behaviour, with progressively higher peaks (>1), despite the consistently low number of newly detected infections. As infections and hospitalizations began to climb in September, *Re* temporarily decreased close to 1, to increase again by mid-October, just before the beginning of the new exponential growth of infected cases, currently ongoing. Indeed, by October 31st, all Italian regions, albeit with different rates were hit by the epidemic (**Figure 1C**). The rapid increase of COVID-19 patients requiring hospitalization during the early months of 2020, as well as *Re* oscillations during the period of epidemic recession, suggest the virus was circulating cryptically among undetected transmission clusters. During this time, there were possibly thousands of mild or asymptomatic infections among undetected (hidden) reservoirs that preceeded each exponential growth phase of each epidemic wave [14,15]. Indeed, dramatic resurgences in cases after easing stringent public health interventions (i.e., stay-at-home orders) that temporarily curtailed epidemic spread was also observed in several other European countries (e.g., UK, France, and Germany, among others).

To investigate further, we coupled epidemiological data with phylodynamic analysis of 714 viral sequences currently available from Italian patients, sampled between January 30th to October 1st, 2020 (see Methods). Viral population dynamics were assessed using non-parametric coalescent estimates of the effective population size (*Ne*) over time (a measure of viral diversity representing the number of diverse genomes contributing to the next generation), given a collection of plausible maximum likelihood (ML) evolutionary histories inferred from viral sequence data. Although distinct patterns could be observed in *Ne* estimates, all reconstructions agree on a rise in *Ne* until the end of March 2020, matching the rise in number of reported cases (**Supplementary Figure 1**). The best-fit model (i.e. the collection of trees with the highest likelihood, *log*L > -49,120) also depicts a steady, continuous decline in *Ne* until October (**Supplementary Figure 1**, pattern A), possibly reflecting the impact of lockdown measures on the viral population. As *Ne* is related to viral genetic diversity, this pattern may indicate that, despite the rapid rise of cases in late summer, the viral population maintained lower diversity relative to the earlier months of the epidemic. This is consistent with a reduction of viral importations, likely resulting from global public health intereventions such as travel bans. Two alternative patterns inferred from trees similar in likelihood value to the best-fit model, show either a similar downward trend followed by a pronounced increase in *Ne* between September and October (pattern B), or a slower but steady increase in *Ne* between April and October (pattern C). Both reconstructions are in agreement with an increase of viral *Ne*, corresponding with an exponential increase of SARS-CoV-2 infections during the second epidemic wave. Together, with the inferred oscillations of *Re* values following the first epidemic wave, the analyses suggest the persistence of complex transmission dynamics throughout the epidemic recession period, involving undetected asymptomatic or mildly affected individuals. Even considering the *Ne* values inferred from ML trees with lower likelihood, we arrive at an analogous conclusion – overall reduction in *Ne* after April but repeated fluctuations throughout recession and second epidemic wave (**Supplementary Figure 1**, black curves).

Longitudinal comparison of SARS-CoV-2 dissemination patterns over time among different Italian regions shows that the pre-lockdown phase was characterized by an exponential growth of the number of daily-confirmed COVID-19 cases and deaths, with highest incidence in the Northwest, followed by a significant decrease across all regions in the aftermath of lockdown measures (**Supplementary Figure 2)**. By the end of August 2020, epidemiological data also show increased and sustained transmission in the South and Insular regions, possibly driven by interregional spreading through small family/social network clusters. These regions are the main touristic destination for Italians, and most of the restrictions on international travel were still in place during Phase III [16]. Lineages proportion and regional-specific distribution in different parts of the country are indicative of several independent founder events (**Figure 2B**). For example, lineage A, predominant in Sicily, has been detected in epidemiologically linked transmission chains that appear to be related to immigrants arrived from North Africa during the late Phase III [17]. Interestingly, the number of circulating lineages have changed over time (**Supplementary Figure 3**). Sub lineage B.2 was the first one identified in January, marking the primary introduction of imported cases from China (no shown in Figure 2). Between February and April, additional sub lineages, such as

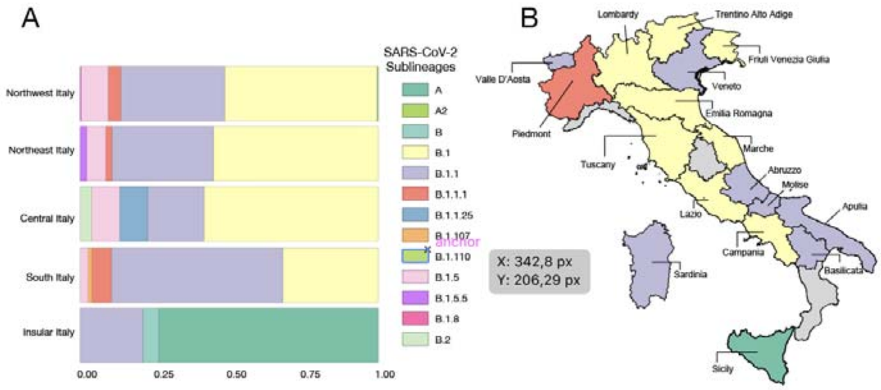


Figure 2. Frequency and distribution of SARS-CoV-2 lineages and sub lineages in Italy. **A**) Frequency of the lineages and sub lineages of SARS-CoV-2 among Italian macro regions. **B**) Distribution of the most prevalent lineage and sub lineage across the country.

B.1, B.1.1, and B.1.5 emerged in Northern and Central Italy, the epicenter of the first epidemic wave, likely reflecting subsequent importations. At the beginning of Phase II lockdown in May, which followed a dramatic decrease in cases, only B.1. and B.1.1 sub lineages were detected. During the period of epidemic recession between June and July, multiple sub lineages co-circulated again. However, the subsequent second wave was dominated by B.1.1 (September) and B.1. (October) (**Supplementary Figure 3**). Since Phase II and III measures permitted intra- and then inter-regional travel, respecitively, while country borders remained mostly closed (except with European countries part of the Shengen agreement), it is plausible that in the first epidemic wave lineages' heterogeneity resulted from intial founder events associated with international travel, and then propagated through within state mobility during epidemic recession. Such sequence-based inferences, however, should be interpreted with caution because of the inherent sampling bias in SARS-CoV-2 full-length genomes currently available from Italian patients, which could affect results and limit their generalizability [18].

In our sequence dataset, only Lombardy (Northwest, most affected region so far), has provided a robust number of viral genomes (n=405), which in turn corresponds approximatively to one genome available every 450 positive cases. Abruzzo in Central Italy is the second most represented region in terms of available genomes (n=87), while many other regions, including Liguria (Northwest), Umbria (Central) and Calabria (South) are not comprehensively represented (**Figure 3A**), thus affecting our ability to characterize in-depth SARS-CoV-2 molecular epidemiology at a regional level. Nevertheless, phylogeny-inferred virus evolutionary patterns are useful to corroborate epidemiological data, test hypotheses regarding factors driving epidemic dynamics, and assess public health interventions such as stay-at-home orders. To this end, we time-scaled the best 100 ML trees of all available SARS-CoV-2 full genomes from Italian patients, and inferred the most likely location of each internal node (ancestral sequence) in the trees (see Methods for details). The overall topologies of the inferred trees were highly similar, and linear regression of root-to-tip genetic distances against sampling dates indicated sufficient temporal signal in the sequence

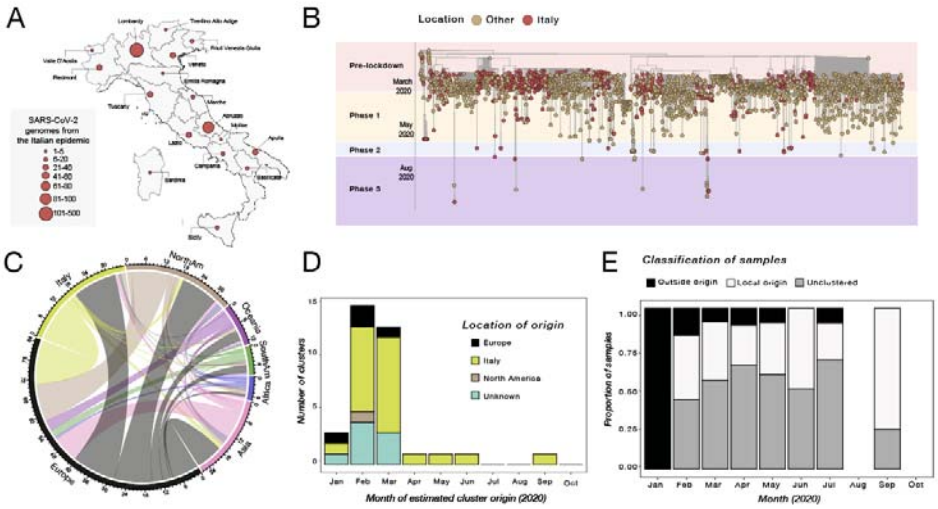


Figure 3. Phylogenetic characterization of Italian SARS-CoV-2 sequences. A) Map of Italy showing the number of SARS-CoV-2 genome sequences by region. The size of the circles indicates the number of new genomes available since the beginning of the epidemic in Italy; B) Time-resolved maximum likelihood tree of 1421 SARS-CoV-2 sequences including 714 from Italy (red circles). C) Chord diagram of estimated numbers of migration flows between the geographic areas. D) Frequency of estimated geographical origins for identified transmission clusters involving Italy and originating in the months of January through October of 2020. E) Frequency of Italian sequences (sampled from January through October) classified as unclustered (grey) or belonging to clusters with Italian (white) or non-Italian origins (black).

data (**Supplementary Figure 4**). Although SARS-CoV-2 evolutionary rate in Italy was somewhat lower (1.44 10^{-04} nucleotide substitutions/site/year) than values obtained for the worldwide epidemic [19,20], the most recent common ancestor (TMRCA) of the available Italian sequences, ranged between January 2nd and January 26th (mean Jan 14th) 2020, consistent with the date of the first confirmed case (Jan 30th). Similarly, the root node (origin) of a time-scaled ML tree including both Italian (n=714) and worldwide reference sequences (n=1,421) was placed in China (99.8% probability), with a TMRCA dating back to early December 2019, in agreement with available epidemiology data [21,22], further validating our phylogeny inference. The tree (**Figure 3B**) consistently shows most of the Italian sequences interspersed with virus strains collected in other countries. This pattern, alike the one observed elsewhere [23], confirms that emergence of SARS-CoV-2 strains during the first epidemic wave was primarily fostered by travel exposure during the pre-lockdown phase, rather than interregional spreading. According to the estimation of migration flows, we further examined the potential Italian role as an exporter of SARS-CoV-2. The number of state transitions into and from Italy (**Figure 3 panel C**) heavily relies on the number and nature of the sequences that are included from other locations. Independently of the dataset, and in line with the epidemiological information, most of the geographical sources of the introductions are attributed to Europe (**Figure 3C**). Well supported (bootstrap values > 90%) putative transmission clusters within the phylogeny were identified based on a pre-defined genetic distance threshold likely to detect epidemiologically linked sequences (see Methods). Clusters containing at least one Italian sequence were considered of interest for the estimation of temporal and spatial origins of the transmission. Temporal origins of each cluster were derived from the clock-estimated age of the MRCA of all sequences belonging to the cluster. Spatial origins were inferred using joint likelihood ancestral state reconstruction, given known country of sampling of tip nodes (sampled sequences) within the tree. As expected, the number of (well supported) clusters formed over the course of the epidemic was largely influenced by the number of contemporaneous samples (**Figure 3D**), limiting conclusions regarding the rate of cluster formation over time. The estimated geographic origins of each cluster reflected the distribution of samples among the reference sequences, largely limited to Europe and North America. However, after April, transmission clusters could only be traced back to Italy, suggesting highly localized transmission following the implementation of Phase one lockdown measures. Each Italian sequence was then classified either as unclustered (i.e. no cluster with any other sequence with bootstrap >90%), or belonging to a local (all Italian) cluster. Italian sequences within well supported clusters including and originating from non-

Italian strains were classified as belonging to “outside” clusters. Finally, each well supported cluster for which a single country could not be assigned with >90% probability as the one at the origin of that cluster, was also considered to be an outside (albeit unknown in origin) cluster. This revealed distinct patterns between January, February-July, and August (**Figure 3E**). All Italian sequences obtained in January belonged to clusters of foreign origin, demonstrating the influence of outside introductions before lockdowns were put into place. The predominant fraction was quickly replaced by sequences belonging to clusters of local origin and unclustered sequences, which suggests potential undersampling. The month of September, when the second epidemic wave was increasing, sequences of local origin, with no sequences of foreign origin, largely dominated. The fraction of sequences sampled in August (75%) was outside the 95% confidence interval (~50%) for the fraction in remaining months, emphasizing the significant contribution of local transmission on sequences sampled in September. However, the specific mutational profile of the Italian sequences (**Figure 4A**), relative to the Wuhan reference (NC_045512), also provided some evidence of recently imported strains during Phase III lockdown, when travel bans began to be eased. In particular, 97.34% (n=695) of the available sequences carried the

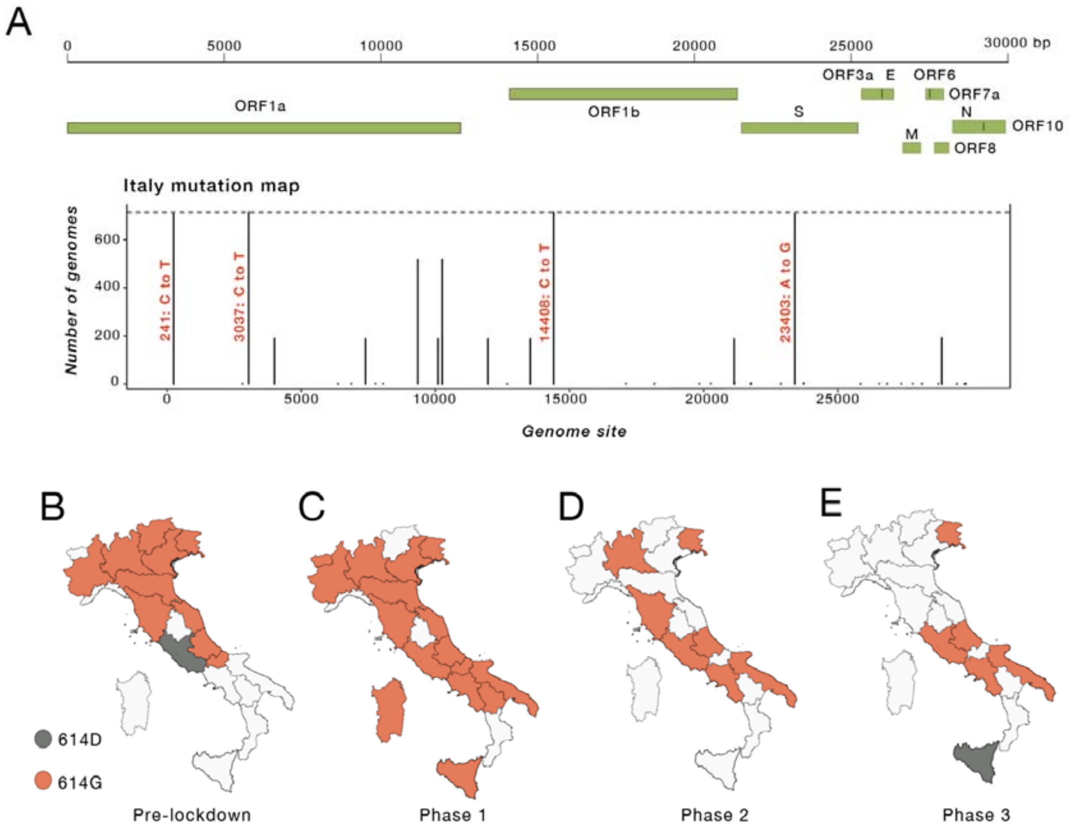


Figure 4. Italian strains mutations pattern. A) Variant maps of the most common mutations mapped against the SARS-CoV-2 genomes. Most common mutations defined as mutations present in >90% of the genomes in that group (black lines). B-E) Change in frequency of the D614G mutation in the Spike Protein across Italian regions during epidemic phases.

mutation encoding for the amino acid change D614G (genomic coordinate: *23403A>G*) in the Spike protein of SARS-CoV-2, while the remaining 2.66% (n=19) sequences displayed the nucleotide sequence encoding for the D614 wild type. The D614G mutation has been associated with higher infectivity and greater transmissibility with no effects on disease severity outcomes [24–26], although some of these findings have recently been questioned [27]. The frequency of the D614G polymorphism among Italian regions over time (**Figure 4B**) shows that while G614 quickly became rapidly dominant during the first epidemic wave, and remained the only one detected in the available sequences through the first two lockdown phases. Thereafter, the D614 variants re-emerged following the relaxation of Phase III measures in Sicily (Insular Italy), possibly due to the epidemiologically-linked transmission chains related to immigration flow from North Africa [17], a scenario reinforced by SARS-CoV-2 lineage A prevalence in that region (**Figure 2B**).

Results of cluster analysis indicate maintained local transmissions fostered by relatively small transmission chains during the months of low case reports and through the beginning of the second wave. This observation suggests that epidemic resurgence was associated with a relaxation of lockdown measures that led to increased local transmission, rather than a large number of virus re-introductions into the country. Such a scenario is also supported by surveys showing a significant reduction in the number of foreign tourists (about -65.9%), but an increase, albeit small (1.1%), of domestic tourism during the summer season after restrictions on interregional travel were relaxed [16]. In order to explore whether increased mobility could explain the second surge of cases in Italy, we carried out stochastic agent-based epidemic simulations. Mobility data across three different modes of transportation (walking, public, and personal vehicle), derived from Apple Mobility trends reports, were used as a proxy for the number of individuals with whom an infected individual comes into contact, which was allowed to vary over time (see Methods). As the number of hospitalizations also dropped drastically (and stayed low) following the first surge in cases, the role of removal of infected individuals from the population *via* hospitalization was also tested, by allowing the probability of an infected individual exiting the simulation to be proportional to the standardized hospitalization rates (also varying in time). The simulated number of active infections over time using the mobility data alone, hospitalization data alone, and combined were then compared to the empirical case data. Whereas all three models produced a similar rate of new infections during the first epidemic wave (**Supplementary Figure 5**), time-varying rate of removal based on hospitalization rates (without mobility data) produced a continuing exponential growth of infections (**Figure 5A**). As in empirical data, time-varying number of contacts based on mobility produced two distinct waves, which were the most similar to the epidemic curve (**Figure 5B**). The model incorporating both mobility and hospitalization rates produced a first wave that was of too large a magnitude and a second wave too early in origin than the previous model (**Figure 5C**). The model incorporating mobility data alone resulted in the lowest mean absolute error (**Figure 5D**), producing a first wave of similar timing and magnitude and a delayed second wave, closer to the empirical epidemiology data (**Figure 1B**).

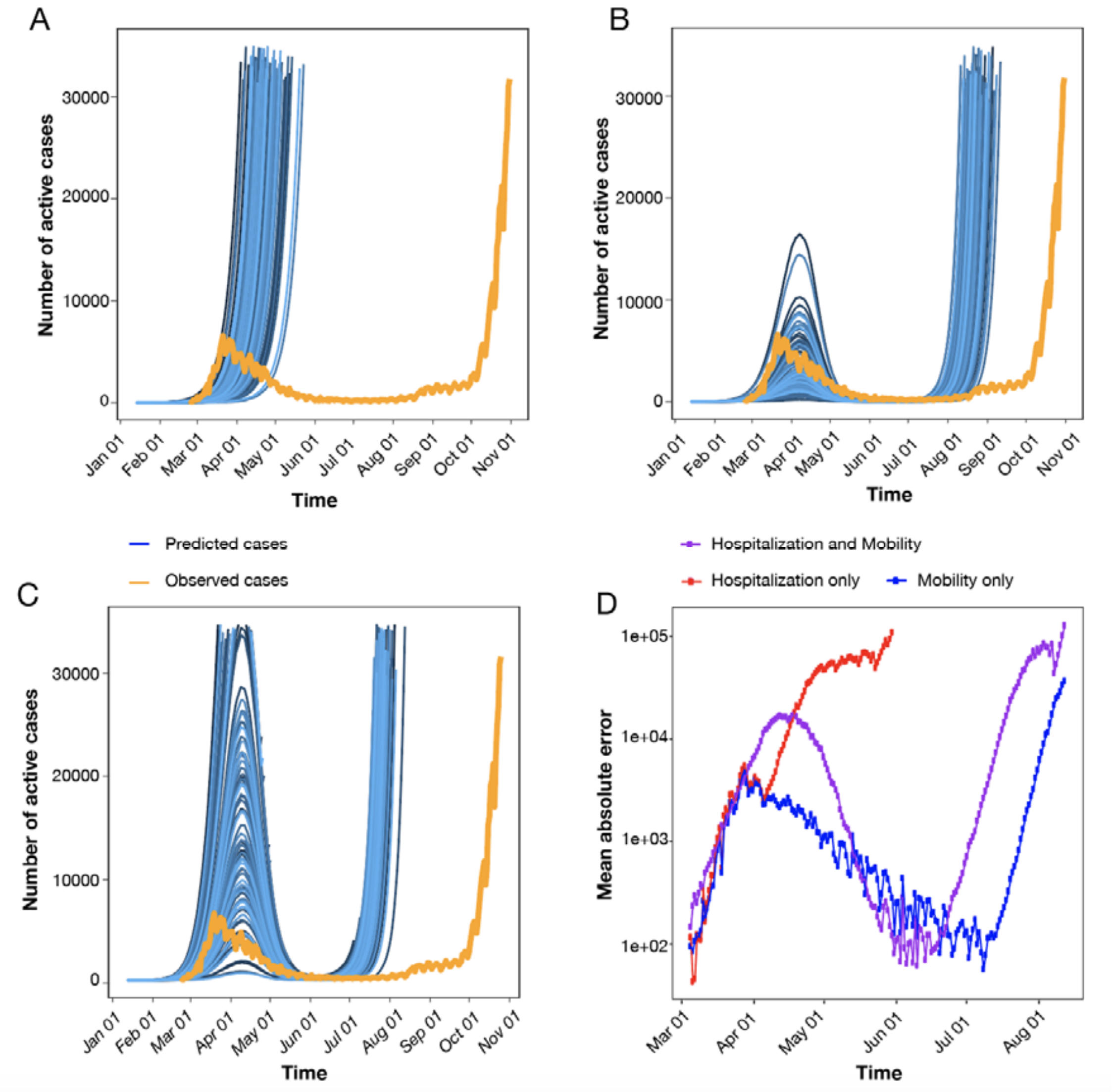


Figure 5. Simulated epidemics under scenarios involving time-varying mobility and hospitalization rates. A) The probability of removal of an infected individual was proportional to the empirical rate of hospitalization in the simulation of active infected cases over one year (blue). B) The number of individuals with whom an infected individual comes in contact was proportional to the empirically determined number of individuals utilizing walking, as well as public and personal modes of transportation, as primary means of mobility in the simulation of active infected cases over one year (blue). C) Hospitalization rates (as in A) and mobility data (as in B) were combined in the simulation of active infected cases over one year (blue). In A-C, orange represents the number of empirically observed infections. D) The absolute error was calculated for each time point of collected observations between simulated infections using hospitalization rates only (red), mobility data only (blue), and the combination (purple). Mean absolute error was calculated as the averaged error across 1,000 simulations. Simulation assumed a single outside introduction.

While results indicate that mobility data could reproduce epidemic wave patterns in Italy, we cannot exclude additional factors, that might have played a role delaying the second epidemic wave, not fully captured by our simulations, such as retention of restriction measures, different "types" of mobility between first and second wave, or higher temperatures in the summer [28].

By coupling phylodynamic analysis of viral genetic and epidemiology data, we show how the interplay between public health intervention and shifting SARS-CoV-2 transmission dynamics in Italy may explain the oscillation between times of relatively stable epidemic recession and dramatic resurgences, as it is currently being observed. This pattern of "rubberbanding" or "snapping back" after public health restrictions are lifted has, unfortunately, been followed by several other European countries. Overall, we show the critical role played by small transmission clusters, acting as "hidden reservoirs" during epidemic recession following aggressive lockdown measures, in maintaining SARS-CoV-2 low-level circulation in Italy, which eventually seeded a new epidemic wave. Despite the consistent agreement between different viral phylogeny-based and epidemiology data analyses, however, limitations of our work need to be acknowledged. Availability of a large number of viral sequences, collected over an extended period of time and sufficiently representative of the ongoing epidemic, is crucial for prompt genomic surveillance, and the evaluation and planning of effective and opportune control strategies. The number of Italian SARS-CoV-2 full genomes currently deposited in public databases represents a very small fraction (0.05%) of the documented number of confirmed cases in Italy, and sampling bias across regions differently affected by the epidemic further limits generalizability of the results. Moreover, our definition of putative transmission clusters (see Methods) does not require the sampling and inclusion of all the strains involved in a transmission chain, although it does allow for detection of monophyletic clades that likely comprise sequences epidemiologically linked through a transmission chain, whilebeit not directly. Nevertheless, epidemiology observations, corroborated by phylodynamic analyses based on available sequences, depicted a coherent picture. The first epidemic wave in Italy appears to have largely been linked to outside introductions leading to large transmission clusters, concomitant with high number of infections. Subsequent implementation of a three-phase nationwide lockdown strategy greatly mitigated numbers of infection and hospitalization during summer 2020. Yet, once mobility increased and social distancing decreased due to the progressive easing of lockdown measures, a sudden spike of infectious cases was observed, promptly followed by new hospitalizations. Our agent-based mathematical model recapitulates this phenomenon, further supporting the hypothesis that the small clusters observed during the summertime were acting, essentially, as "hidden reservoirs" that likely merged following the increased in mobility and reduction of social distancing measures. This in turn provided the "spark" for the sudden increase of infections observed at the end of summer, which led to the subsequent second wave of exponential grow. In other words, the drivers of SARS-CoV-2 transmission dynamics shifted from high levels of community transmission, likely involving mass super spreader events, in the early Italian epidemic, to sustainment by smaller family/social network clusters later in the epidemic. Unfortunately, this also suggests that no amount of community level interventions may be sufficient to curb the epidemic as long as people do not adhere to individual level measures such as mask use, hand hygiene, and social distancing. New lockdown measures are likely to provide only temporary relief, as has already happened in the first months of the epidemic in Italy and many other countries. Indeed, an important debate is currently ongoing about vaccine deployment, given financial and logistic restrictions mandating a very long phased deployment, based on prioritization policies. In this context, our results suggest that hidden transmission reservoirs may continue to sustain local outbreaks into late 2021, as vaccine rollout will likely take months before reaching the necessary herd-immunity threshold. Ultimately, our ability to curb successfully the current pandemic, may be linked to our ability to determine number and structure of such reservoirs within the social and behavioral context of specific locales.

Acknowledgments

We thank all the authors, originating and submitting laboratories that have kindly deposited and shared genome data on GISAID EpiCoV database, on which this research is based. An acknowledgment table can be found in **Supplementary Table 3.** We thank all personnel from the Italian Health Surveillance System of the *Protezione Civile* that assisted with epidemiological data collection. MS, BDR and CB contribution was funded, in part, by the U.S. National Science Foundation RAPID grant (DMS-2028728). MS and BDR are also supported by the Stephany W. Holloway University Chair in AIDS Research.

Authors' contributions

Conception and design: MG, EC, BRM, MS and MC; Data collection: MG, EC; Investigations: MG, EC, FB, BRM, VF, SF, and JL; Data Analysis: MG, EC, BRM, EW, VF, NED, CB and JL; Writing – Original: MG, EC, FB, BRM, MS, DZ and MC; Draft Preparation: MG, EC, FB, BRM, JL, VT, MS, DZ and MC; Revision: GC, SP, AR, AB, VT, CS, AM, TA, EW, TdO, LCJA, GC, RC, AC, JL, MS, DZ, and MC.

Competing Interests Statement

The authors declare no competing interests.

REFERENCES

1. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* vol. 26 450–452 (2020).

2. Yang, X.-L. *et al.* Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **90**, 3253–3256 (2016).

3. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

4. Massimiliano S. Tagliamonte *et al.* Recombination and purifying selection preserves covariant movements of mosaic SARS-CoV-2 protein S | bioRxiv. (2020) doi:http://doi.org/10.1101/2020.03.30.015685.

5. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

6. Giovanetti, M., Angeletti, S., Benvenuto, D. & Ciccozzi, M. A doubt of multiple introduction of SARS-CoV-2 in Italy: A preliminary overview. *J. Med. Virol.* **92**, 1634–1636 (2020).

7. The 2019-new coronavirus epidemic: Evidence for virus evolution - Benvenuto - 2020 - Journal of Medical Virology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25688.

8. World Health Organization (WHO). Coronavirus disease (COVID-19) Situation Report – 103, 02 May 2020 [Internet]. 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200502-covid-19-sitrep-103.pdf?sfvrsn=d95e76d8_4.

9. Civil Protection Department. COVID-19 Italia. http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1.

10. La Repubblica. Coronavirus, i contagi nel Lodigiano sono 15: i primi sono un 38enne di Codogno e sua moglie. In isolamento 250 persone. https://milano.repubblica.it/cronaca/2020/02/21/news/coronavirus_a_milano_contagg iato_38enne_e_un_italiano_ricoverato_a_codogno-249121707/.

11. Percivalle, E. *et al.* Prevalence of SARS-CoV-2 specific neutralising antibodies in blood donors from the Lodi Red Zone in Lombardy, Italy, as at 06 April 2020. *Eurosurveillance* **25**, (2020).

12. Civil Protection Department. State of epidemiological emergency. http://www.protezionecivile.gov.it/media-comunicazione/news/dettaglio/-/asset_publisher/default/content/coronavirus-dichiarato-lo-stato-di-emergenza.

13. Jason Horowitz & Emma Bubola. On Day 1 of Lockdown, Italian Officials Urge Citizens to Abide by Rules. *NYTimes.com* (2020).

14. The first two cases of 2019-nCoV in Italy: Where they come from? - Giovanetti - 2020 - Journal of Medical Virology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25699.

15. Stefanelli, P. *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* **25**, 2000305 (2020).

16. CST Firenze for Assoturismo Confesercenti. *Turismo Estate 2020 Italia: mancano gli stranieri, calo della domanda del -30,4%*.

http://centrostudituristicifirenze.it/blog/turismo-estate-2020-italia-mancano-stranieri-calo-della-domanda/.

17. Irene Schöfberger & Marzia Rango. Migration in West and North Africa and across the Mediterranean - COVID-19 and migration in West and North Africa and across the Mediterranean - | IOM Online Bookstore. (2020).
18. Mavian, C., Marini, S., Prosperi, M. & Salemi, M. A Snapshot of SARS-CoV-2 Genome Availability up to April 2020 and its Implications: Data Analysis. *JMIR Public Health Surveill.* **6**, (2020).
19. Andrew Rambaut. Phylodynamic Analysis | 176 genomes | 6 Mar 2020. *Virological* https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356 (2020).
20. Su, Y. C. *et al.* Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. *bioRxiv* 2020.03.11.987222 (2020) doi:10.1101/2020.03.11.987222.
21. Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003.e9 (2020).
22. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, (2020).
23. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582 (2020).
24. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
25. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2895-3.
26. Benvenuto, D. *et al.* Evidence for mutations in SARS-CoV-2 Italian isolates potentially affecting virus transmission. *J. Med. Virol.* **92**, 2232–2237 (2020).

27. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).

28. Benedetti, F. *et al.* SARS-CoV-2: March toward adaptation. *J. Med. Virol.* (2020) doi:10.1002/jmv.26233.

29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

30. Müller, H. *et al.* VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. *Nucleic Acids Res.* **45**, W567–W572 (2017).

31. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).

32. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

33. O'Toole Á & McCrone J. Phylogenetic Assignment of Named Global Outbreak LINeages. https://github.com/hCoV-2019/pangolin (2020).

34. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

35. Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3**, (2017).

36. R Core Team. R: A language and environment for statistical computing. in (2019).

37. Volz, E. M. & Didelot, X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Syst. Biol.* **67**, 719–728 (2018).

38. Prosperi, M. C. F. *et al.* A novel methodology for large-scale phylogeny partition. *Nat. Commun.* **2**, 321 (2011).

39. Liam Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* (2012).

40. Pupko, T., Pe'er, I., Shamir, R. & Graur, D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.* **17**, 890–896 (2000).

41. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

42. Wickham, H., François, R., Henry, L., Müller, K. & RStudio. *dplyr: A Grammar of Data Manipulation*. (2020).

43. Henry, L., Wickham, H. & RStudio. *purrr: Functional Programming Tools*. (2020).

44. Kun Ren. rlist: A Toolbox for Non-Tabular Data Manipulation version 0.4.6.1 from CRAN. https://rdrr.io/cran/rlist/.

45. Yu, G., Jones, B. & Arendsee, Z. *tidytree: A Tidy Tool for Phylogenetic Tree Data Manipulation*. (2020).

46. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* (2016).

47. Dowle, M. *et al. data.table: Extension of 'data.frame'*. (2020).

48. Wickham, H. Reshaping Data with the reshape Package. *J. Stat. Softw.* **21**, 1–20 (2007).

49. Grolemund, G. & Wickham, H. Dates and Times Made Easy with lubridate. *J. Stat. Softw.* **40**, 1–25 (2011).

50. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

51. Wickham, H. & RStudio. *tidyr: Tidy Messy Data*. (2020).

52. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).

53. Jérémie Scire *et al. A method to monitor the effective reproductive number of SARS-CoV-2*.

54. Gostic, K. M. *et al.* Practical considerations for measuring the effective reproductive number, Rt. *medRxiv* (2020) doi:10.1101/2020.06.18.20134858.

55. Goldstein, E. *et al.* Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21825–21829 (2009).

56. Lequime, S., Bastide, P., Dellicour, S., Lemey, P. & Baele, G. nosoi: A stochastic agent-based transmission chain simulation framework in r. *Methods Ecol. Evol.* **11**, 1002–1007 (2020).

Methods

Sequence data collection

To perform a comprehensive analysis of the genomic epidemiology of SARS-CoV-2 in Italy, after excluding low-quality genomes (> 10% of ambiguous positions), we dowloaded all Italian full-length viral genomes available on GISAID (https://www.gisaid.org/) (n=714) up to October 31h 2020. Appropriate acknowledgement was given to the sequencing laboratories (**Supplementary Data S1**). Sampling locations of available genomes in this dataset included 17 of 20 regions in Italy, and collection dates spanned from January 30th (the first two imported cases in Italy) to October 1st 2020. Each Italian sequence was used in a local alignment (BLAST) [29] search for the most (genetically) similar non-Italian sequence in the GISAID database as of Oct 31st, 2020, and linked to two reference sequences including the best match (highest E-value) with a date occurring within one month following, as well as one month prior to the sampling date of the Italian sequence (although, in some cases, only a single non-Italian reference sequence fulfilling one of the inclusion criteria could be found for multiple Italian query sequences). After removing duplicate sequences and masking mutations potentially associated with common sequencing errors, using a vcf filter [30], a final dataset of 1,421 reference sequences was assembled (**Supplementary Table 2**).

Sequence alignments and phylogenetic analysis

Sequences (Italian + reference strains) were aligned using MAFFT (FF-NS-2 algorithm) employing default parameters [31]. The alignment was manually curated to optimize number and location of gaps using Aliview [32]. A site-specific mutational comparison of the 714 Italian genomic sequences obtained from the GISAID database was made with the MAFFT-aligned SARS-CoV-2 reference genome (RefSeq: NC_045512.2), obtained from the GenBank database. Lineage assessment was conducted using the Phylogenetic Assignment of Named Global Outbreak LINeages tool available at https://github.com/hCoV-2019/pangolin [33]. Phylogenetic analysis of was performed using the maximum likelihood (ML) method implemented in IQ-TREE (version 1.6.10), employing the best-fit model of nucleotide substitution according to the Bayesian Information Criterion (BIC), as indicated by the Model Finder application implemented in IQ-TREE [34]. The statistical robustness of individual nodes was determined using 1000 bootstrap replicates.

Molecular clock calibration and estimation of virus effective population size

ML trees were inspected in TempEst v1.5.3 for the presence of temporal signal (i.e., linear relationship between genetic distance and sampling time in the available sequences) [21]. The treedater package in R v3.6.0 [35,36] was used for molecular clock calibration of the Italy-only data, as well as the combined Italy and reference data. The top 100 maximum likelihood (ML) trees (i.e. the trees with the 100 lowest -*log*[likelihood] values), were chosen for calibration according to a strict clock (no branch specificity) among the Italy-only data, whereas a single-best ML tree was chosen for the combined dataset. Individual *taxa* sampling times were used to rescale branch lengths to time in each tree using a starting value of $8x10^{-4}$ substitutions/site/year. The skygrowth non-parametric demographic model [37] was then used in R with time-scaled trees to estimate of median virus effective population size (*Ne*) and

95% high posterior density intervals for each week during the epidemic in Italy (Italy-only dataset) using the default smoothing parameter value (tau) of 0.1.

SARS-CoV-2 transmission cluster identification and characterization

Transmission clusters were identified using Phylopart v2 [38] applied to the ML tree of combined sequence data (scaled in substitutions/site). A range of percentile thresholds spanning 10^{-6} % – 15% of the whole-tree patristic distance distribution was used to choose an optimal threshold point and to verify robustness of cluster composition. The minimum percentile threshold that maximized the number of clusters was chosen as the optimal threshold by performing multiple clustering runs on randomly sampled patristic distance distributions (1 million for each run). Well-supported sub-trees (bootstrap values > 90%) with mean pairwise patristic distances among *taxa* within the chosen threshold were considered putative transmission clusters (i.e. clusters comprising sequences epidemiologically linked through a transmission chain, although some of the direct links may be missing). Only clusters containing at least 1 Italian sequence were considered in downstream analyses. The phytools package [39] in R was used for joint likelihood reconstruction of discrete ancestral origins [40] according to country (and associated uncertainty) for the most recent common ancestor (MRCA) of each transmission cluster within the ML tree (scaled in substitutions/site) for the combined dataset. Transition rates among discrete states (countries) along tree nodes were considered to be equal. The tree scaled in time was used to attribute temporal origins to each cluster, or time of MRCA (TMRCA). The following R packages were used in the manipulation of data for cluster characterization and visualization: ape [41], dplyr [42], purr [43], rlist [44], tidytree [45], ggplot2 [46], data.table [47], reshape2 [48], lubridate [49], ggtree [50], tidyr [51], parallel [36].

Estimation of basic reproduction number

Estimates for daily basic reproduction number, *Re*, of SARS-CoV-2 in Italy were obtained from the COVID-19-re data repository (https://github.com/covid-19-Re/dailyRe-Data) as at 20th September 2020. The effective reproductive number describes the average number of secondary infections caused by an infected individual. The relevant method of calculation of Re builds upon another method developed by Cori et al.[52], accessible through EpiEstim R package. Instead of using a time series of infection incidence, which cannot be observed directly, the relevant method infers the infection incidence time series based on secondary sources of information such as COVID-19 confirmed case data, hospital admissions, and deaths. This was considered in combination with two other sets of time variables: i) the duration of SARS-CoV-2 incubation period and ii) the time delays between the onset of the symptoms and a positive test, a hospital admission or the death of a patient. The relevant method infers infection time series from the stated observed incidence data by deconvolution [52–55].

Epidemiology data assembly

We analysed COVID-19 cases counts in Italy from publicly released data up to October 31st 2020 from the Italian Civil Protection Department repository (https://github.com/pcm-dpc/COVID-19) that releases daily updates on the number of new confirmed cases, deaths and recoveries, with a breakdown by region. To illustrate the epidemic progression, the daily number of confirmed cases of people infected with SARS-Cov-2 in Italy was plotted alongside a timeline of lockdown phases and variation in estimated virus reproduction number until October 31st 2020. For convenience the geographical locations were aggregated by Italian macro regions: Northeast, Northwest, Central, South, and Insular, which are basic regions for the application of regional policies (Italian regions). Mobility data over time, combining data on three different forms of transportation - personal vehicle, public, and walking - were obtained from Apple Mobility Trends Reports (https://covid19.apple.com/mobility).

Agent-based stochastic model simulation of the Italian epidemic

The Italian epidemic was simulated using the forward-time, agent-based stochastic transmission chain simulator, nosoi [56], which allows for time-varying parameterization. The simulation was initiated with a single infected individual with a probability of transmission of 0.02 per day following an incubation period, which was set to a mean of 5 days and standard deviation of 2 days. The rate of transmission was fixed throughout the simulation at this value. Following the incubation period, each individual was considered infectious for approximately 9 days, exiting the simulation at a mean time of 14 days (standard deviation equal to 2) after infection. Three different simulation scenarios were tested, assuming a direct relationship between 1) rate of mobility and the number of other individuals with which each infected individual comes into contact, 2) hospitalization rate and the rate at which an infected individual was removed from the simulation during the infectious period (approximately 5 to 14 days), or both. Hospitalization and mobility data were standardized by (observed-minimum)/(maximum-minimum) to a range of 0-1 and modeled using a Fourier series periodic function, with mobility data comprised of 2 sine/cosine terms (linear model regression R2=0.8484) and hospitalization data of 4 sine/cosine terms (linear model regression R^2=0.984). In scenarios 1 and 3, the probability of exiting the simulation was allowed to vary over time proportionally to the standardized rate of hospitalization, resulting in a maximum of ~35% of infected individuals hospitalized during peak hospitalization of the epidemic. In scenarios 2 and 3, the number of contacts for each infected individual was allowed to vary over time proportionally to the mobility rate, resulting in a mean of approximately 15 individuals in contact with each infected individual. For scenario 1, a static mean removal rate (during the infectious period described above) over time was set to 0.04 (standard deviation of 0.01). For scenario 2, the mean number of contacts per individual was set to 15 (standard deviation of 8). The mean absolute error for each time point was calculated to assess the deviation of the simulated number of actively infected individuals for each of the three scenarios from the true number of cases, provided by Italian Ministry of Health and the Civil Protection Department.

Data Availability

Sequence and epidemiology raw data utilized, generated or analyzed during these studies are available from the authors upon request (including sequence alignment and R scripts for the phylodynamic analyses).

Supplementary Figures Legend

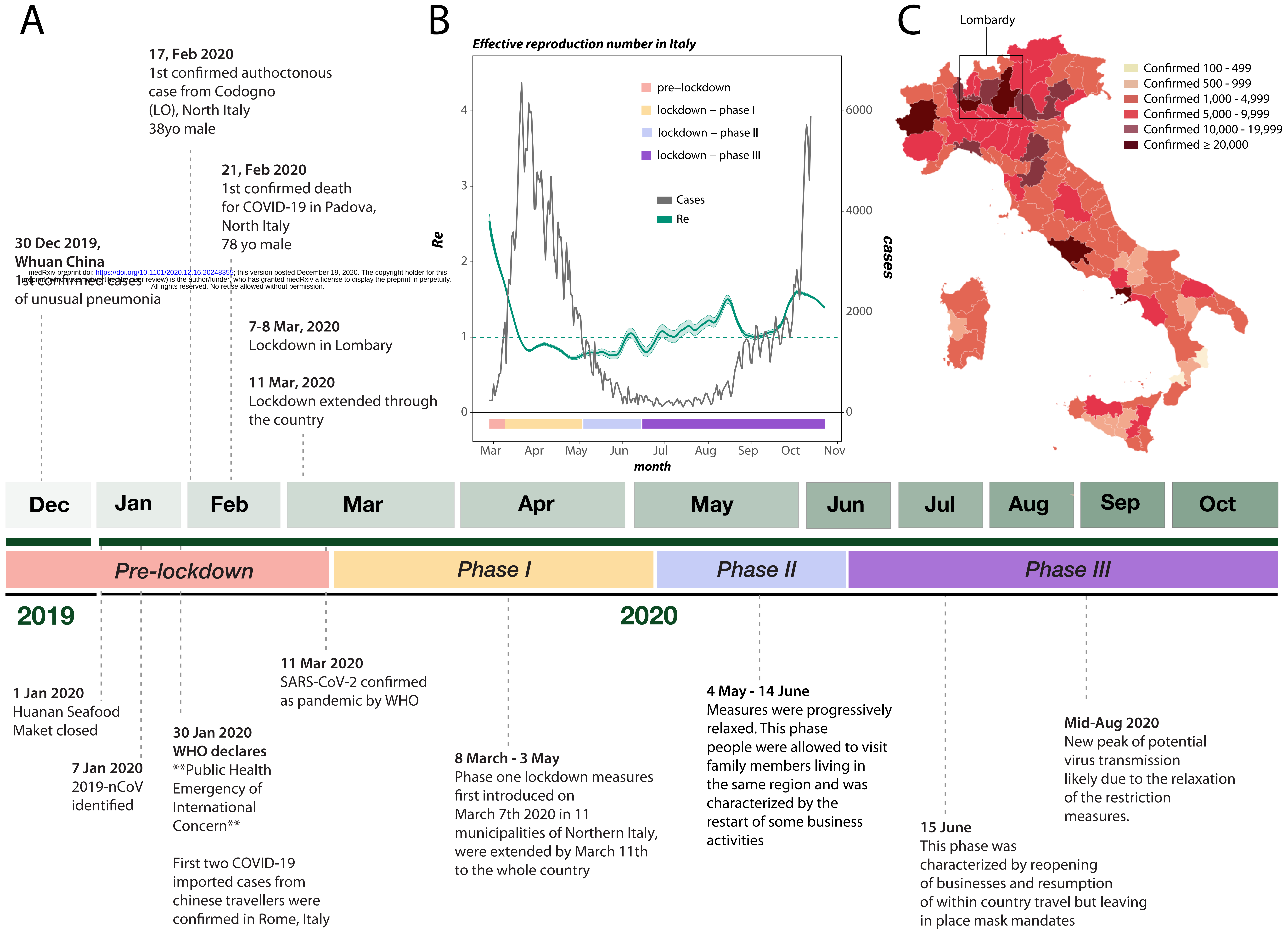
Supplementary Figure 1. Estimates of the viral effective population size (Ne) in the Italian epidemic. Estimates of the viral effective population size (Ne) using a sample of 100 phylogenetic trees with the highest log-likelihood values. Encircled letters (A, B and C) label the three major patterns inferred from collection of trees with the highest likelihood values.

Supplementary Figure 2. Spatial and temporal distribution of SARS-CoV-2 trough Italian regions. Italian regions were aggregated into five macro regions (NUTS, Nomenclature of territorial units for statistics): Northeast, Northwest, Central, South and Insular. The left-hand Y-axis (left) represents incidence (cases per 100K population, black curve) of COVID-19, while the secondary Y-axis (right) represents the number of deaths (red curve) related to COVID-19 through the country. Y-axis numbers are represented as log10 for visual purposes. Colours at the bottom represent the epidemic phases in Italy: pre-lockdown (early transmission) in light red, phase 1 in yellow, phase 2 in light violet and phase 3 in purple. Maps of Italy were superimposed to showing to exact location of each region: Northwest; Northeast; Central; South and Insular.

Supplementary Figure 3. Frequency and distribution of SARS-CoV-2 lineages in Italy over time.

Supplementary Figure 4. Analysis of temporal structure. **A.** Root-to-tip genetic divergence of Italian sequences against time of sampling. **B.** Root-to-tip genetic divergence for the whole dataset (Italian strains + reference sequences) against time of sampling.

Supplementary Figure 5. Empirical observations (points) and chosen model (lines) for hospitalization rates (red) and number of mobile individuals (blue) over time. Hospitalization rates represent the number hospitalized per positive case, whereas mobility represent the number of individuals utilizing walking, as well as public and personal modes of transportation, as primary means of mobility (blue). Values were standardized for comparison. Empirically observed number of new infections (standardized) are also shown for comparison in orange.



30 Dec 2019, Whuan China
1st confirmed cases of unusual pneumonia

17, Feb 2020
1st confirmed authoctonous case from Codogno (LO), North Italy
38yo male

21, Feb 2020
1st confirmed death for COVID-19 in Padova, North Italy
78 yo male

7-8 Mar, 2020
Lockdown in Lombary

11 Mar, 2020
Lockdown extended through the country

1 Jan 2020
Huanan Seafood Maket closed

7 Jan 2020
2019-nCoV identified

30 Jan 2020
WHO declares
Public Health Emergency of International Concern

First two COVID-19 imported cases from chinese travellers were confirmed in Rome, Italy

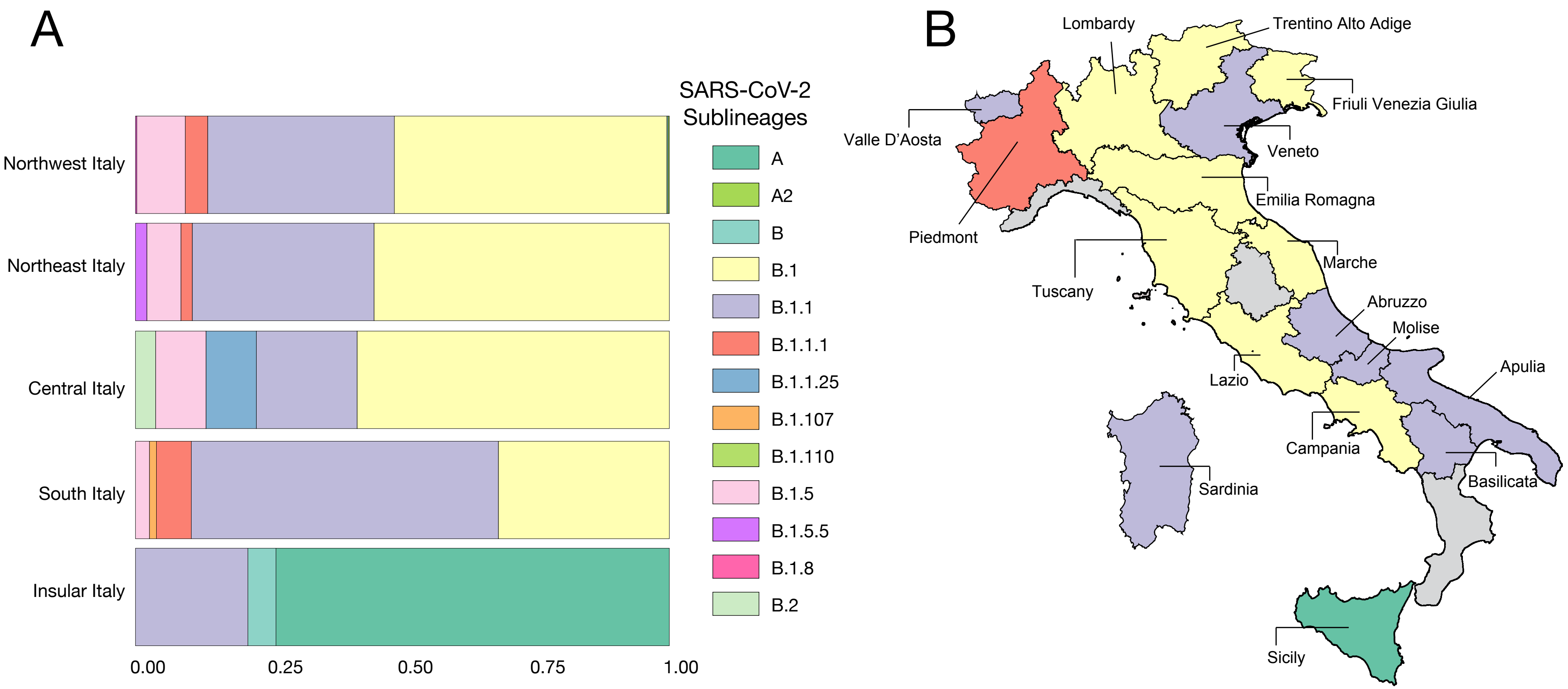
11 Mar 2020
SARS-CoV-2 confirmed as pandemic by WHO

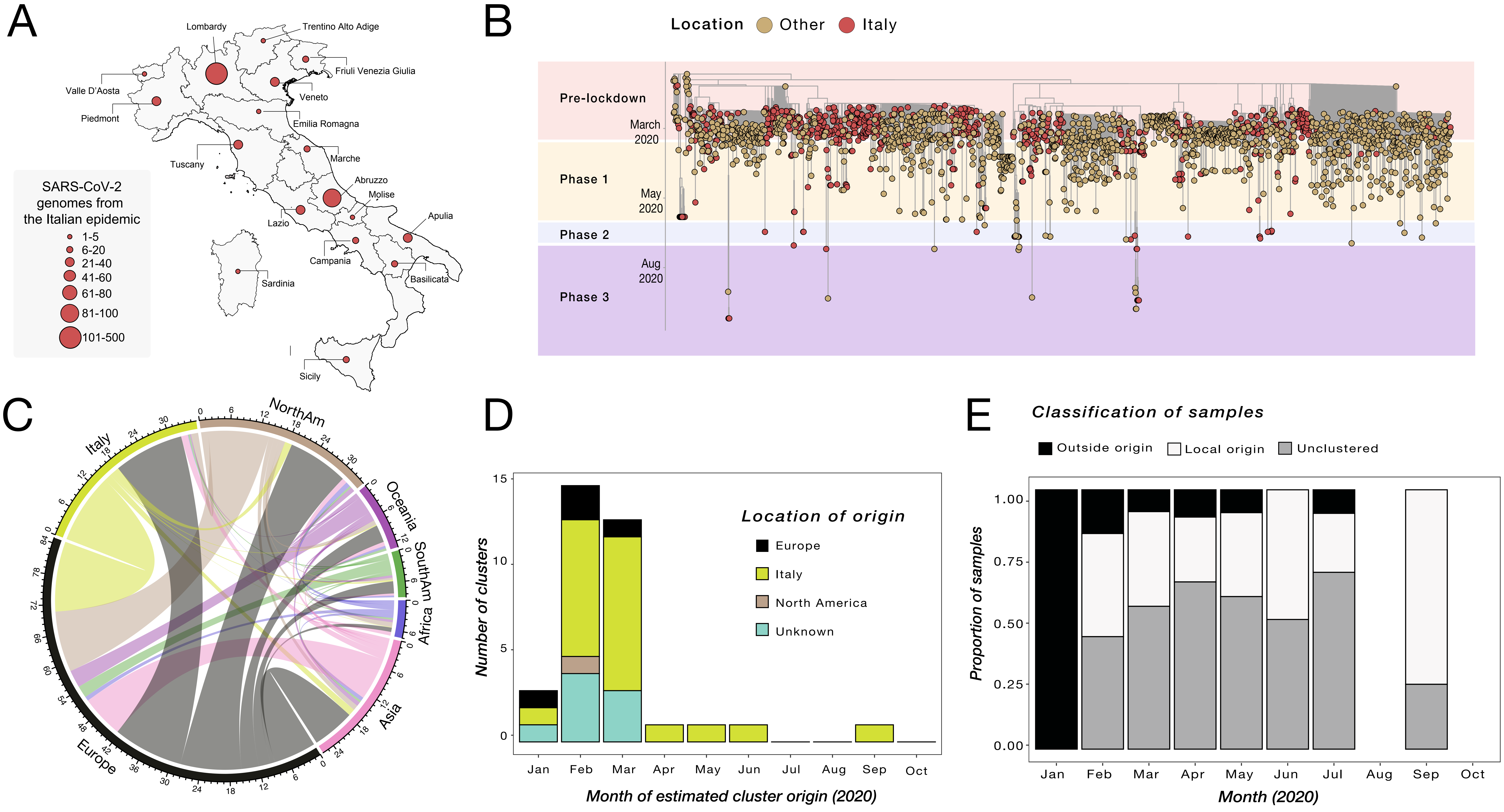
8 March - 3 May
Phase one lockdown measures first introduced on March 7th 2020 in 11 municipalities of Northern Italy, were extended by March 11th to the whole country

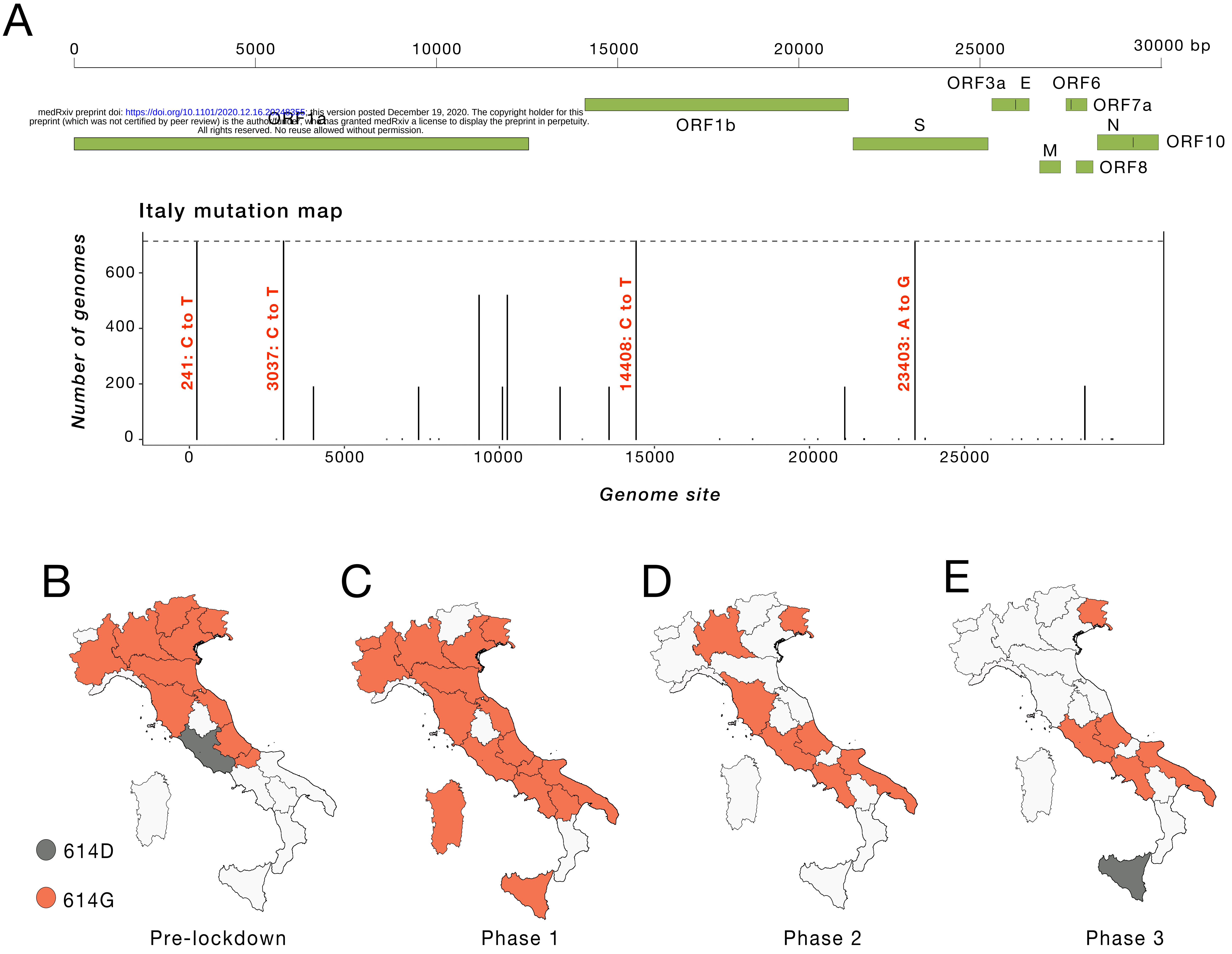
4 May - 14 June
Measures were progressively relaxed. This phase people were allowed to visit family members living in the same region and was characterized by the restart of some business activities

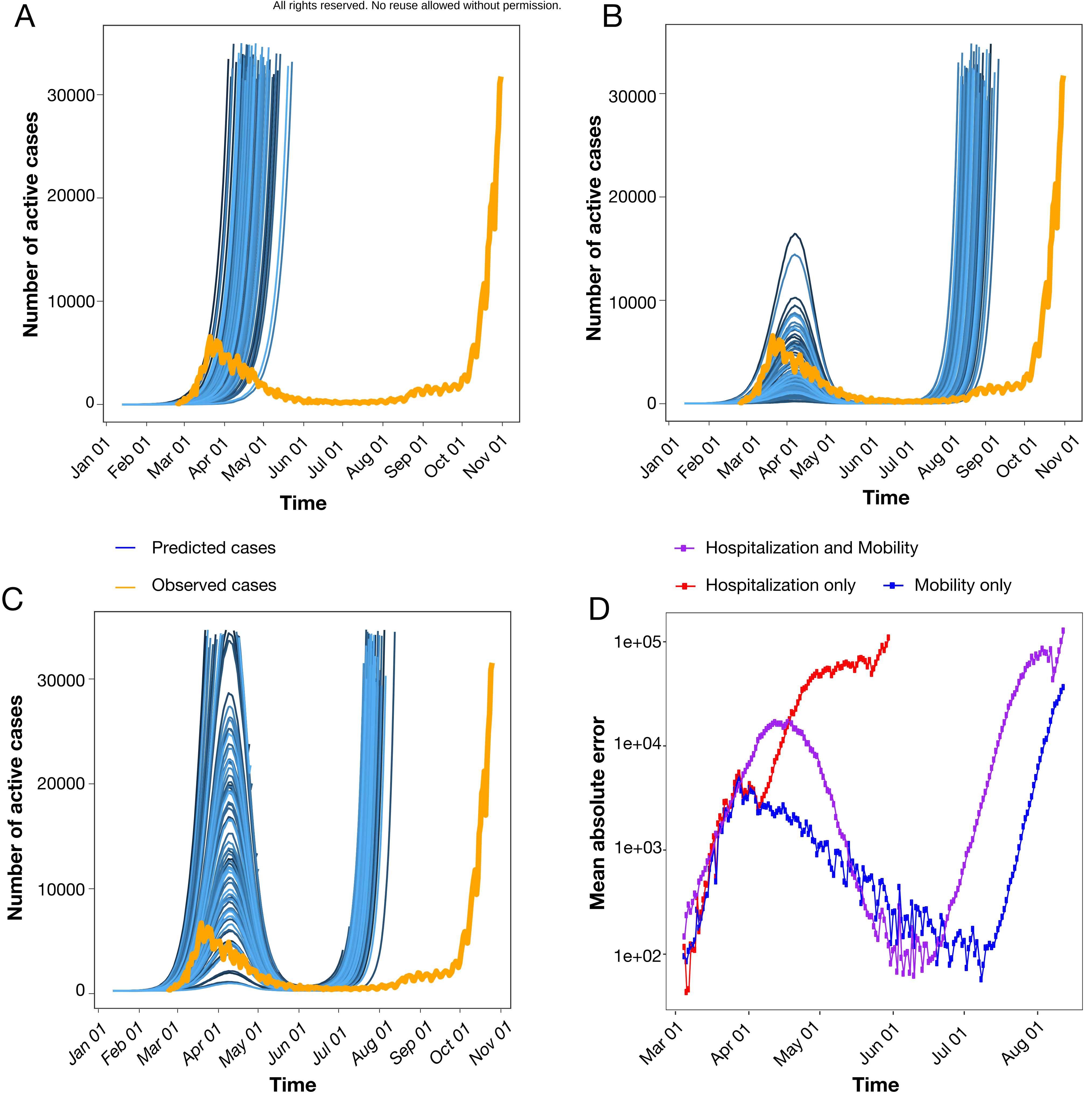
15 June
This phase was characterized by reopening of businesses and resumption of within country travel but leaving in place mask mandates

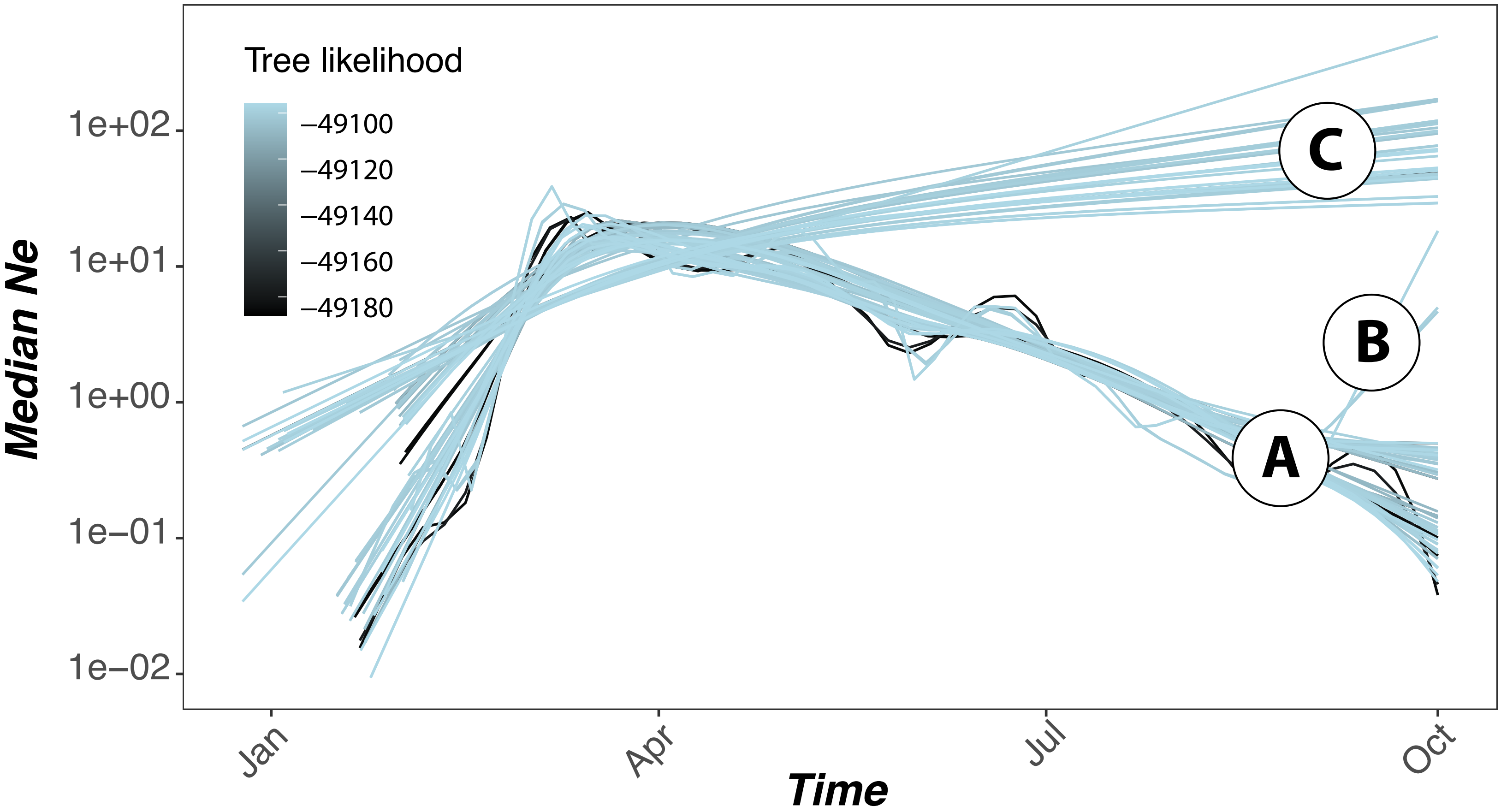
Mid-Aug 2020
New peak of potential virus transmission likely due to the relaxation of the restriction measures.

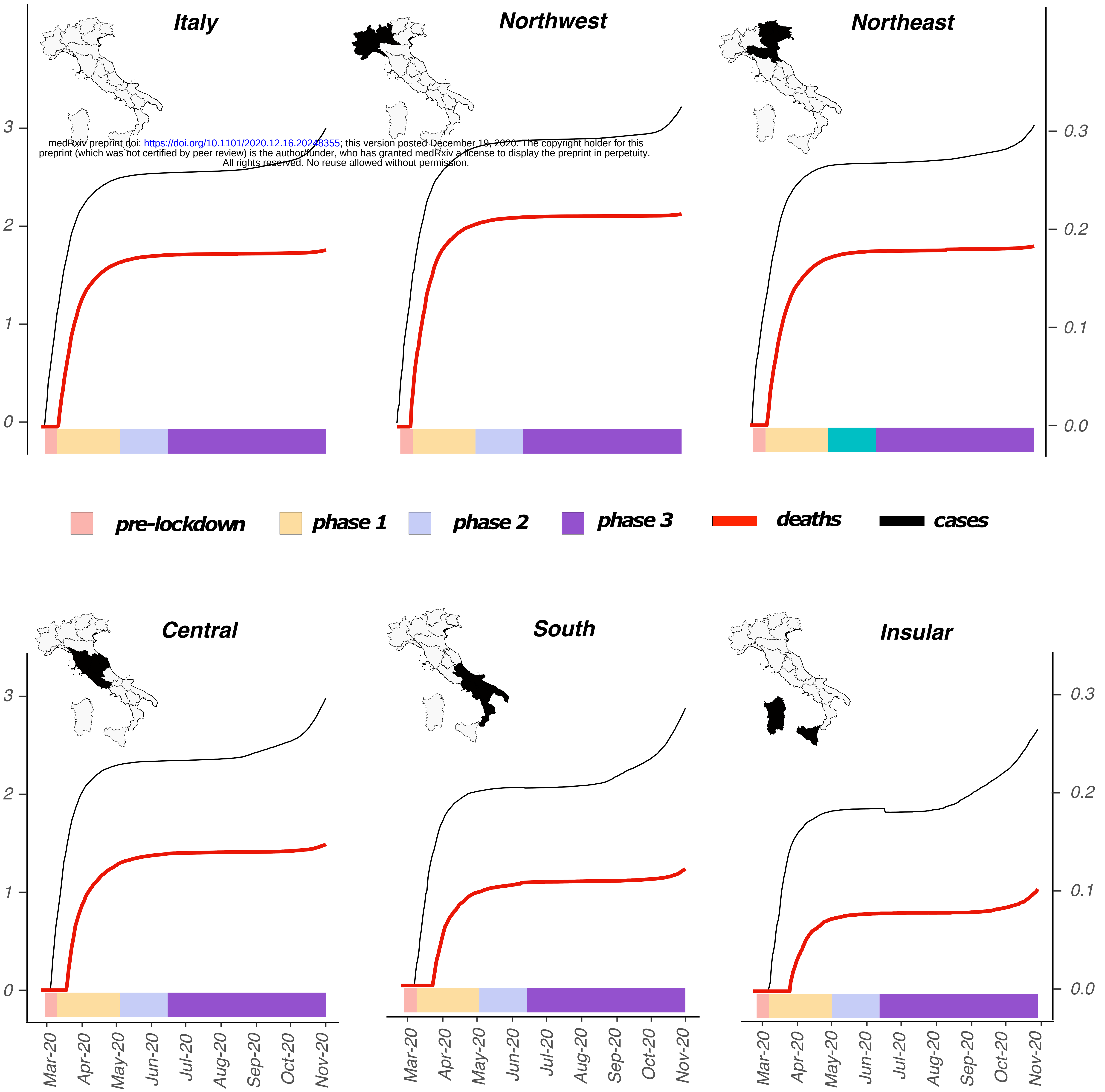

A
B
SARS-CoV-2 Sublineages
A
A2
B
B.1
B.1.1
B.1.1.1
B.1.1.25
B.1.107
B.1.110
B.1.5
B.1.5.5
B.1.8
B.2
Northwest Italy
Northeast Italy
Central Italy
South Italy
Insular Italy
0.00
0.25
0.50
0.75
1.00
Lombardy
Trentino Alto Adige
Friuli Venezia Giulia
Valle D'Aosta
Veneto
Emilia Romagna
Piedmont
Marche
Tuscany
Abruzzo
Molise
Apulia
Lazio
Campania
Sardinia
Basilicata
Sicily

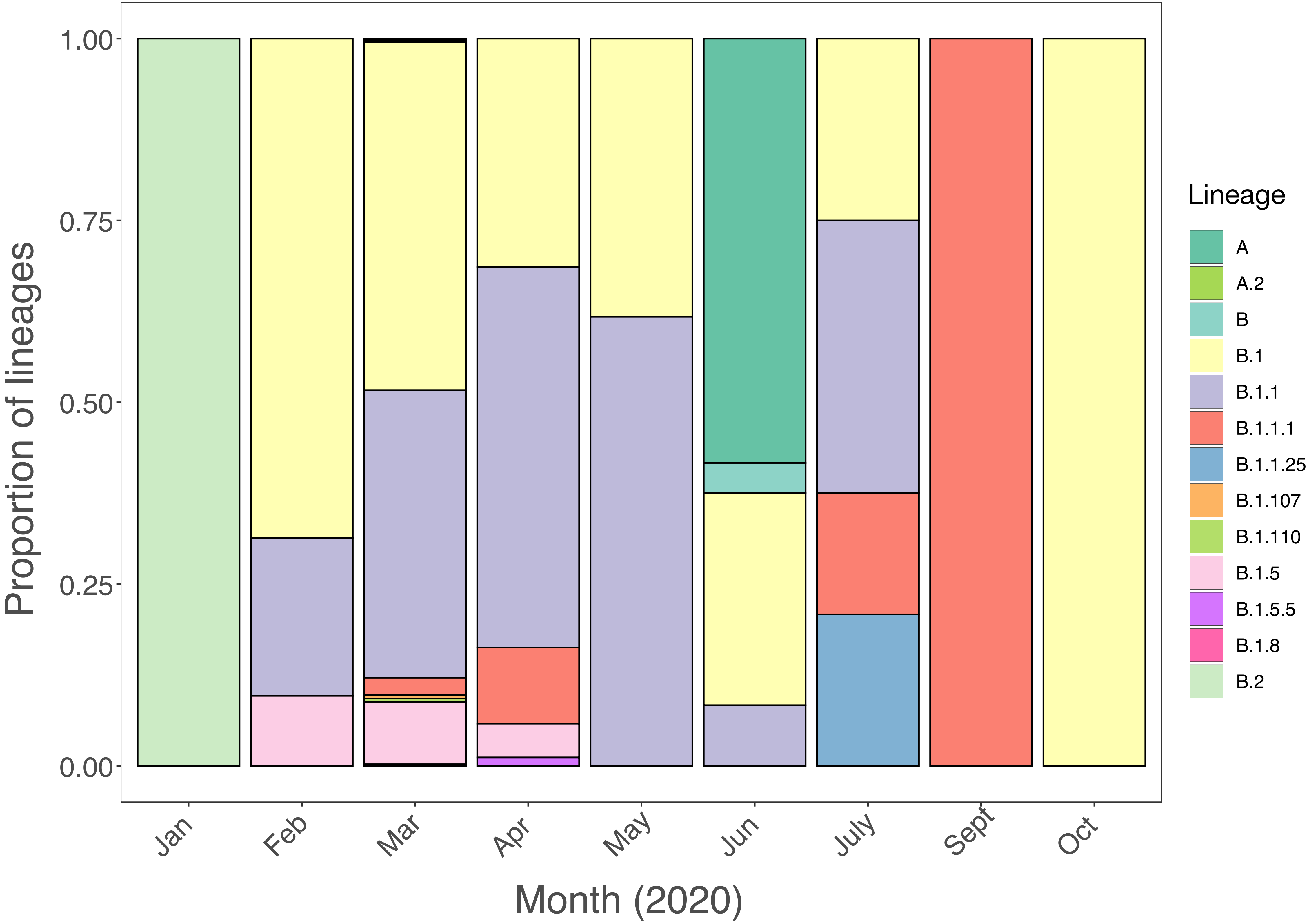

A
Lombardy
Trentino Alto Adige
Friuli Venezia Giulia
Valle D'Aosta
Veneto
Piedmont
Emilia Romagna
Tuscany
Marche
Abruzzo
Molise
Lazio
Apulia
Campania
Basilicata
Sardinia
Sicily
SARS-CoV-2 genomes from the Italian epidemic
1-5
6-20
21-40
41-60
61-80
81-100
101-500
B
Location
Other
Italy
Pre-lockdown
March 2020
Phase 1
May 2020
Phase 2
Aug 2020
Phase 3
C
Italy
NorthAm
Oceania
SouthAm
Africa
Asia
Europe
D
Location of origin
Europe
Italy
North America
Unknown
Number of clusters
0
5
10
15
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Month of estimated cluster origin (2020)
E
Classification of samples
Outside origin
Local origin
Unclustered
Proportion of samples
0.00
0.25
0.50
0.75
1.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Month (2020)

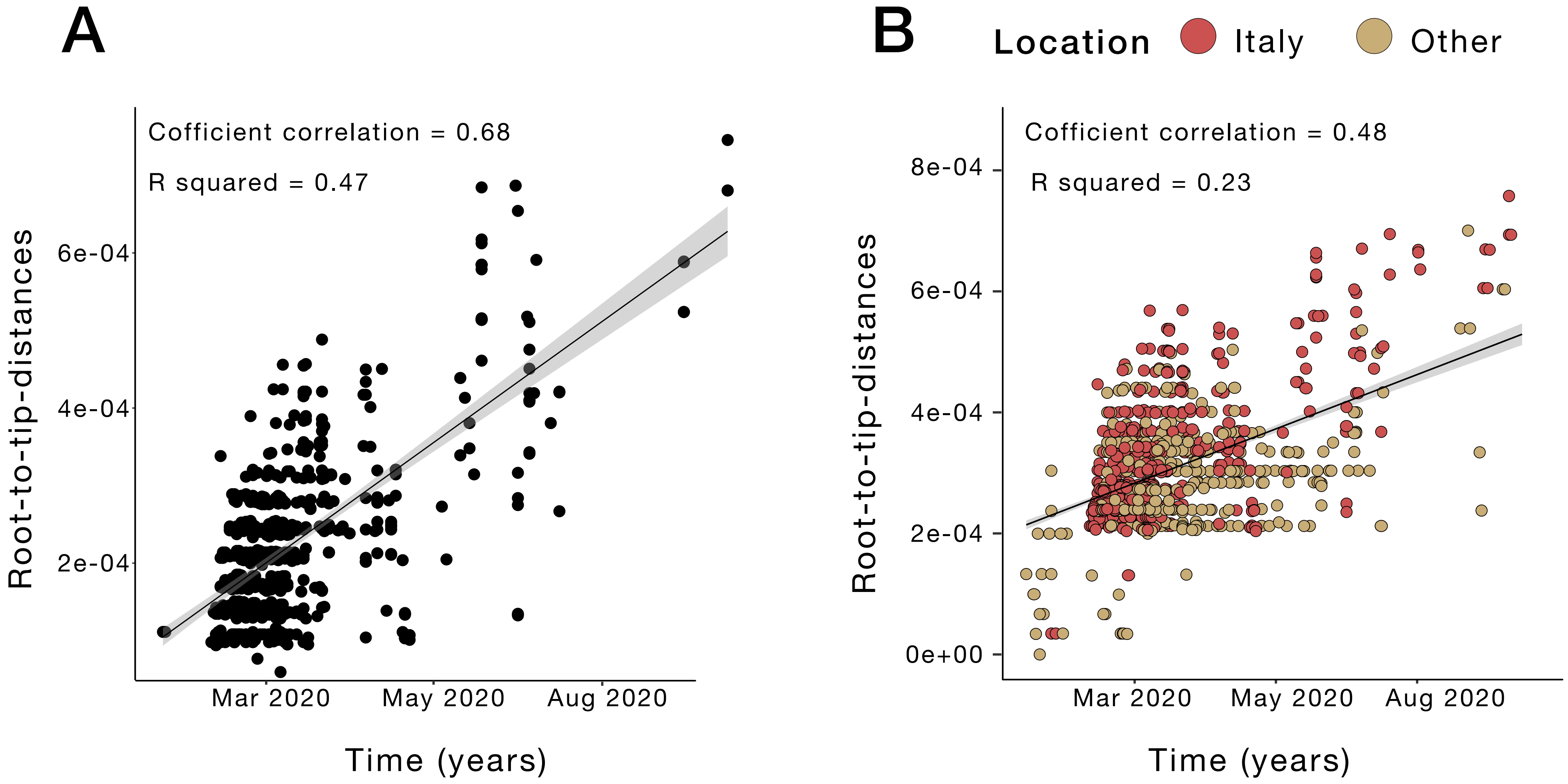

A
0
5000
10000
15000
20000
25000
30000 bp
ORF1a
ORF1b
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
Italy mutation map
Number of genomes
0
200
400
600
241: C to T
3037: C to T
14408: C to T
23403: A to G
0
5000
10000
15000
20000
25000
Genome site
B
C
D
E
614D
614G
Pre-lockdown
Phase 1
Phase 2
Phase 3

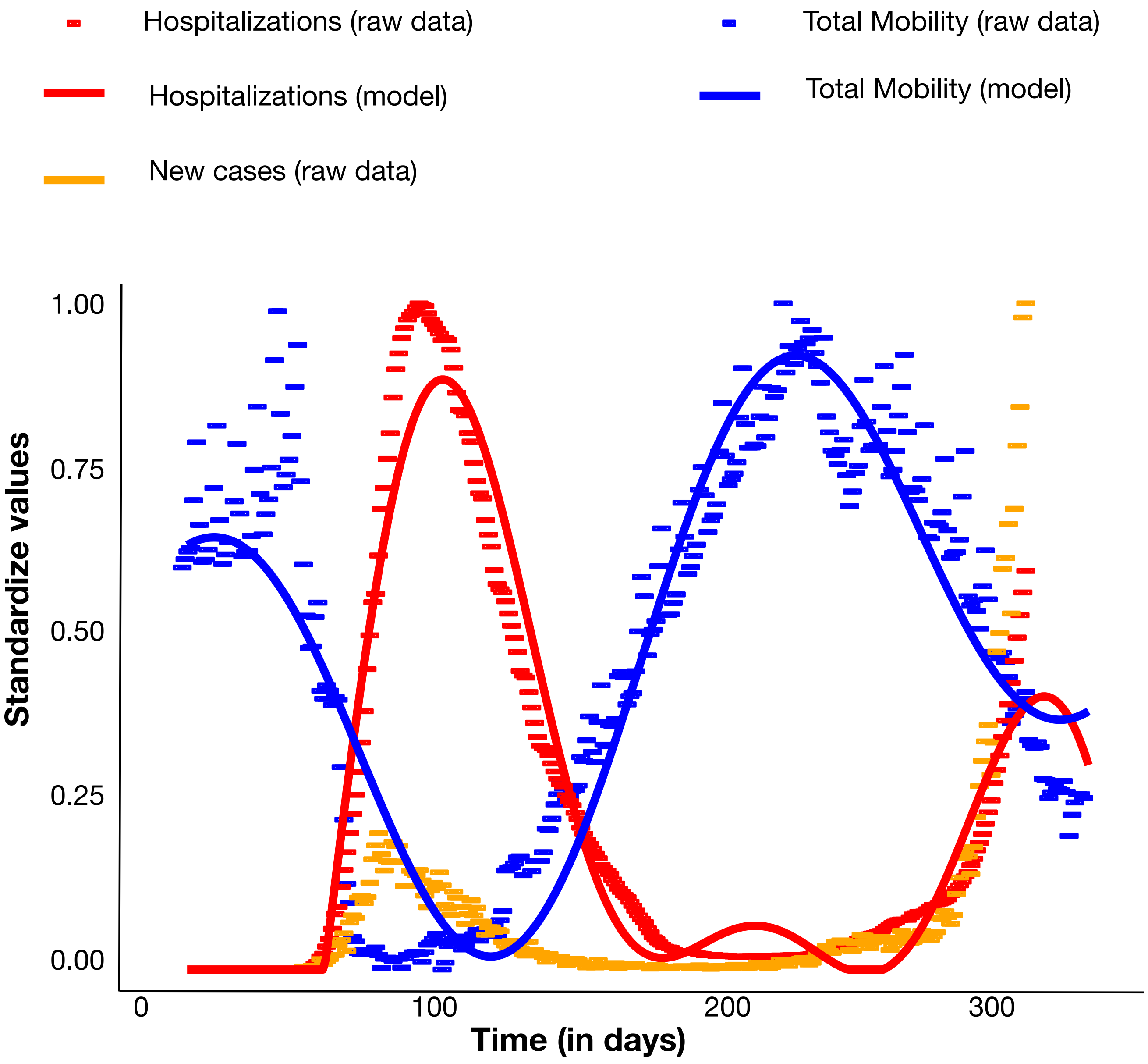

A
Number of active cases
30000
20000
10000
0
Jan 01
Feb 01
Mar 01
Apr 01
May 01
Jun 01
Jul 01
Aug 01
Sep 01
Oct 01
Nov 01
Time
B
Number of active cases
30000
20000
10000
0
Jan 01
Feb 01
Mar 01
Apr 01
May 01
Jun 01
Jul 01
Aug 01
Sep 01
Oct 01
Nov 01
Time
Predicted cases
Observed cases
Hospitalization and Mobility
Hospitalization only
Mobility only
C
Number of active cases
30000
20000
10000
0
Jan 01
Feb 01
Mar 01
Apr 01
May 01
Jun 01
Jul 01
Aug 01
Sep 01
Oct 01
Nov 01
Time
D
Mean absolute error
1e+05
1e+04
1e+03
1e+02
Mar 01
Apr 01
May 01
Jun 01
Jul 01
Aug 01
Time


Tree likelihood
−49100
−49120
−49140
−49160
−49180
Median Ne
1e+02
1e+01
1e+00
1e−01
1e−02
Time
Jan
Apr
Jul
Oct
A
B
C


Italy
Northwest
Northeast
Central
South
Insular
0
1
2
3
0.0
0.1
0.2
0.3
pre-lockdown
phase 1
phase 2
phase 3
deaths
cases
Mar-20
Apr-20
May-20
Jun-20
Jul-20
Aug-20
Sep-20
Oct-20
Nov-20


Proportion of lineages
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
July
Sept
Oct
Month (2020)
Lineage
A
A.2
B
B.1
B.1.1
B.1.1.1
B.1.1.25
B.1.107
B.1.110
B.1.5
B.1.5.5
B.1.8
B.2


A
B
Location
Italy
Other
Cofficient correlation = 0.68
R squared = 0.47
Cofficient correlation = 0.48
R squared = 0.23
Root-to-tip-distances
Root-to-tip-distances
6e-04
4e-04
2e-04
8e-04
6e-04
4e-04
2e-04
0e+00
Mar 2020
May 2020
Aug 2020
Mar 2020
May 2020
Aug 2020
Time (years)
Time (years)


Hospitalizations (raw data)
Total Mobility (raw data)
Hospitalizations (model)
Total Mobility (model)
New cases (raw data)
Standardize values
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (in days)