Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy

Claudia Alteri, Valeria Cento, Antonio Piralla, Valentino Costabile, Monica Tallarita, Luna Colagrossi, Silvia Renica, Federica Giardina, Federica Novazzi, Stefano Gaiarsa, Elisa Matarazzo, Maria Antonello, Chiara Vismara, Roberto Fumagalli, Oscar Massimiliano Epis, Massimo Puoti, Carlo Federico Perno & Fausto Baldanti

From February to April 2020, Lombardy (Italy) reported the highest numbers of SARS-CoV-2 cases worldwide. By analyzing 346 whole SARS-CoV-2 genomes, we demonstrate the presence of seven viral lineages in Lombardy, frequently sustained by local transmission chains and at least two likely to have originated in Italy. Six single nucleotide polymorphisms (five of them non-synonymous) characterized the SARS-CoV-2 sequences, none of them affecting N-glycosylation sites. The seven lineages, and the presence of local transmission clusters within three of them, revealed that sustained community transmission was underway before the first COVID-19 case had been detected in Lombardy.
Since coronavirus disease 2019 (COVID-19) was initially reported in China on 30 December 2019 (refs. 1,2), SARS-CoV-2 has been spreading worldwide and, as of June 14, 2020, there have been 7.55 million confirmed infections and 423,000 deaths reported worldwide (World Health Organization, 2020).

In Italy, the first case of evident SARS-CoV-2 transmission emerged on February 20, in Codogno, Lombardy, when a young man affected by interstitial pneumonia was diagnosed for SARS-CoV-2. From that date, the number of diagnosed COVID-19 cases exponentially increased, and Lombardy became the area most affected by the COVID-19 pandemic, totaling 89,018 infections by June 2, out of a total of 233,197 cases in Italy. Thereafter, the COVID-19 epidemic grew exponentially during the first days of March 2020, peaking on March 21, 2020, with 6,557 newly confirmed cases. Two months later, reported COVID-19 cases in Italy dropped to ~600 per day, indicating the first wave of epidemic was nearing containment.

With a population of 10 million, Lombardy is the most densely populated region in Italy, and one of the largest. Milan is the largest metropolitan area in Italy and the third most densely populated region in Italy, and one of the largest. Milan is the hub of many transportation links to Europe and beyond. This scenario creates the conditions to host and favor the spread of a highly transmissible virus such as SARS-CoV-2.

In this study, an integrated approach comprising epidemiological and viral genetic data was used to reconstruct the pattern of SARS-CoV-2 spread in this region. Whole-genome sequencing was performed for 346 SARS-CoV-2 strains obtained from individuals from various geographical areas in a time span of 2 months. The main aim of this study was to trace local transmission chains and the temporal and geographical evolution of the virus in Lombardy in relation to epidemiological data and measures implemented to contain the outbreak.

**Results**

**Patient characteristics.** From February 22 through April 4, 2020, nasopharyngeal swabs taken from a total of 25,082 patients were screened for SARS-CoV-2 infection at two major hospitals in Lombardy (Supplementary Fig. 1). A diagnosis of COVID-19 was made for 11,445 of them. Whole-genome sequencing was performed in 371 samples collected from 371 patients residing in all 12 provinces of Lombardy and with varying disease symptoms ranging from mild to critical. Sampling selection criteria for these samples and the comparison of their demographic and clinical characteristics with SARS-CoV-2-infected general population are illustrated in Supplementary Fig. 1, Supplementary Table 1, and Supplementary Results.

Twenty-five samples were excluded due to failed amplification \((n = 9)\) or poor genomic coverage \((<60\%, n = 16)\). The final study population thus consisted of 346 patients, whose geographical, demographic, and clinical characteristics are reported in Fig. 1 and Table 1.

One-hundred and ninety-five \((56.4\%)\) were male, and the median age was 72 \((IQR: 54–84)\) years. Fever was the most common COVID-19 symptom at admission, followed by cough and dyspnea. Chest radiographs or CT scan confirmed a classical bilateral interstitial pneumonia for 49.6% of them. Patients with critical or severe COVID-19 manifestation more frequently suffered from at least one chronic comorbidity, compared to those with moderate and mild manifestations \((P = 0.002\) by Chi-square test for trend), with greater impact played by hypertension and chronic kidney disease \((P < 0.001\), respectively by Chi-square test for trend). Regarding contact tracing, 25 reported to be COVID-19 contact while only four patients out of 237 declared an international travel history. For three of them, the destination was not specified, while one individual reported travel in Singapore from January 12 to February 5, 1 month before receiving a SARS-CoV-2 diagnosis of a B.1.1 strain infection (March 10).

**Genome coverage.** SARS-CoV-2 sequence reads were able to cover from 94.0% to 99.7% of the reference genome (GenBank: NC_045512.2), independently of SARS-CoV-2 load (Supplementary Fig. 2). The few genome regions \((N = 4)\) with lower reads coverage were consistently limited to no more than 35 nt.

**Lineages and single nucleotide polymorphisms characterizing SARS-CoV-2 sequences.** By reassigning sequences to the lineages according to the PANGOLIN application3, we found that the majority of sequences \((n = 179, 51.7\%)\) belonged to lineage B.1, 126 \((36.4\%)\) belonged to lineage B.1.1, and 34 \((9.8\%)\) were assigned to lineage B.1.5. Lineage B.1.1.1 was found in four north Italian sequences, while A.2, B, and B.1.8 lineages were detected in one patient each (Fig. 2). Demographic and clinical characteristics of patients infected with SARS-CoV-2 from such lineages are reported in Table 1. While no relevant differences in demographic and clinical characteristics were found among lineages, some differences were found in the geographical distribution of the three most represented lineages in Lombardy (B.1, B.1.1, and B.1.5). Lineage B.1 mostly affected southern Lombardy, including Lodi, Cremona, and Mantua, with 90.2% of these provinces’ sequences belonging to this lineage (55/61, Fig. 1 and Table 1), while in the north of Lombardy, mostly in Como, Lecco, Bergamo, and its adjacent territories such as Alzano and Nembro, 62.7% of sequences belonged to B.1.1 lineage (84/134, Fig. 1 and Table 1). Finally, lineage B.1.5 was mainly detected in Pavia, with a local cluster in this area composed of 23 B.1.5 sequences.

The composition of sequences did not change substantially from the early testing phase (before lockdown, March 8) to later time points (after March 8), even if B.1.1 lineage, which had made up 27 out of 106 sequences \((25.5\%)\) in the early phase of testing, increased to 99 out of 240 haplotypes \((41.2\%)\) in the later phases of testing, mostly because of the decreased frequency of the B.1 and other B.1 derived lineages (Table 1 and Fig. 3a). Looking at the genetic distance with respect to SARS-CoV-2 reference strain (belonging to lineage B), it was naturally lower for isolates in lineage B.1 with respect to isolates in lineage B.1.1 and B.1.5 \((1.4 \times 10^{-4} [1.0 \times 10^{-4}, 1.7 \times 10^{-4}] vs 1.7 \times 10^{-4} [1.4 \times 10^{-4}, 1.7 \times 10^{-4}] vs 2.4 \times 10^{-4} [2.1 \times 10^{-4}, 2.7 \times 10^{-4}], P < 0.001)\) by Kruskal–Wallis test, Fig. 3b), thus indicating a closer genetic relatedness to original B strains for lineage B.1 with respect to the others. Concordantly with the increase of lineage B.1.1 over time, the genetic pairwise distance of the 346 sequences indicated that the SARS-CoV-2 sequences evolved progressively during time \((r = 0.585 and 0.582\) respectively, Fig. 3c and Supplementary Fig. 3).

Six single nucleotide polymorphisms (SNPs) were shared among >10 Italian genomes, with a prevalence ranging from 3.2% (20268, A to G, syn in nsp15) to 99.4% of the SNP 23403, A to G, non-syn D to G in S and 14408, C to T, non-syn P to L in RdRp (Fig. 4a, b). All SNP were detected in at least one previously published SARS-CoV-2 sequence in GISAID. Five out of these six SNPs were non-synonymous, and three were mapped within SARS-CoV-2 structural proteins. Notably, two non-synonymous SNPs reside in S protein: the C to T at position 23575 (S protein; amino acid T to I; intra-patient prevalence ~100%) found in 24 north Italian sequences and previously detected in <1% of samples in China, the rest of Europe and North America, and the A to G at position 23403 (S protein;
amino acid D to G; intra-patient prevalence ~54%), present in 344 north Italian sequences and firstly detected in China with a prevalence of 1.7%, then rapidly selected and spreading in all countries with a prevalence ranging from ~17% in Asia to ~70% in Europe and North America.

Emergence of lineages in Lombardy. The Bayesian molecular clock analysis estimated a mean evolutionary rate of $1.16 \times 10^{-3}$ subs/site/year (95% HPD, $1.01 \times 10^{-3}$ to $1.32 \times 10^{-3}$).

The current low genetic diversity of SARS-CoV-2 genomes worldwide implies very low posterior probabilities (<0.50) for most of the nodes. Nevertheless, it was possible to identify clear transmission chains and clusters with posterior probability support ≥0.98, able to clarify the dynamic of viral lineages in Lombardy (Fig. 5, Table 2).

As stated above, two sequences from Lombardy clustered with A.2 and B sequences, respectively (Fig. 2). The patient infected by lineage A.2 was diagnosed on March 15 in Lodi. Unfortunately, travel information and possible COVID-19 contact were
Disease severity
g
Sequences in cluster 60 (17.3) 25 (14.0) 12 (9.5) 23 (67.6) 0 (0.0) <0.001

syn RG to KR (N)
28881-28883, GGG to AAC, non-syn
23575, C to T, non-syn
24 (6.9) 1 (0.6) 0 (0.0) 23 (67.6) 0 (0.0) <0.001

(RdRp)
SNP in SARS-CoV-2 genome
–
Cluster C 6 (1.7) 0 (0.0) 6 (4.8) 0 (0.0) 0 (0.0)

Age, years
72 (54.8–84) 68 (54–83) 74 (55–84) 74 (60–85) 72 (46–84) 0.815

18–39
30 (8.7) 17 (9.5) 9 (7.1) 4 (11.8) 0 (0.0) 0.484

40–49
36 (10.4) 20 (11.2) 11 (8.7) 3 (8.8) 2 (28.6) 0.374

50–59
48 (13.9) 31 (17.3) 16 (12.7) 1 (2.9) 0 (0.0) 0.082

60–69
46 (13.3) 23 (12.8) 16 (12.7) 6 (17.6) 1 (2.2) 0.915

70–79
64 (18.5) 24 (13.4) 30 (23.8) 8 (23.5) 2 (28.6) 0.101

≥80
122 (35.3) 64 (35.8) 44 (34.9) 12 (35.3) 2 (28.6) 0.978

Sex, male
195 (56.4) 87 (48.6) 80 (63.5) 23 (67.6) 5 (71.4) 0.018

Residency
Milan
75 (21.7) 46 (25.7) 24 (19.0) 3 (8.8) 2 (28.6) 0.108

Como
68 (19.7) 26 (14.5) 40 (31.7) 2 (5.9) 0 (0.0) 0.001

Pavia
61 (17.6) 20 (11.2) 15 (11.9) 23 (67.6) 3 (42.9) <0.001

Bergamo
36 (10.4) 12 (6.7) 23 (18.3) 1 (2.9) 0 (0.0) 0.003

Lecco
30 (8.7) 9 (5.0) 21 (16.7) 0 (0.0) 0 (0.0) 0.001

Lodi
33 (9.5) 30 (16.8) 1 (0.8) 0 (0.0) 2 (28.6) <0.001

Cremona
26 (7.5) 23 (12.8) 0 (0.0) 3 (8.8) 0 (0.0) 0.001

Othera
17 (4.9) 13 (7.3) 2 (1.6) 2 (5.9) 0 (0.0) 0.135

Chronic comorbiditiesa
155 (52.4) 80 (51.3) 57 (53.3) 14 (46.7) 4 (66.7) 0.759

Hypertension
107 (35.8) 52 (33.3) 42 (39.3) 11 (36.7) 2 (33.3) 0.530

Obesity
23 (7.8) 9 (5.8) 11 (10.3) 2 (6.7) 1 (16.7) 0.477

Diabetes
31 (10.5) 14 (9.0) 14 (13.1) 2 (6.7) 1 (16.7) 0.591

Cardiovascular disease
98 (32.3) 46 (29.5) 40 (37.4) 10 (33.3) 2 (33.3) 0.635

Chronic obstructive lung disease
42 (12.4) 16 (10.3) 19 (17.8) 5 (16.7) 2 (33.3) 0.175

Malignancies
35 (11.8) 21 (13.5) 11 (10.3) 2 (6.7) 1 (16.7) 0.639

Chronic kidney disease
23 (7.8) 8 (5.1) 9 (8.4) 3 (10.0) 3 (50.0) 0.868

Chronic liver disease
4 (1.4) 3 (1.9) 1 (0.9) 0 (0.0) 0 (0.0) 0.788

Othera
28 (9.7) 15 (9.7) 9 (8.7) 4 (13.3) 0 (0.0) 0.765

Symptoms at admission
Fever
150 (65.2) 73 (59.3) 52 (68.4) 19 (67.9) 6 (100.0) 0.174

Cough
106 (46.1) 51 (41.5) 35 (46.1) 15 (53.6) 15 (53.6) 5 (83.3) 0.204

Dyspnea
93 (40.4) 44 (35.8) 32 (42.1) 14 (50.0) 3 (50.0) 0.571

Time from symptoms-onset to hospital admission
0.29 (0.14–0.57) 0.29 (0.14–0.50) 0.33 (0.17–0.69) 0.26 (0.00–0.37) 0.17 (0.00–0.57) 0.244

SARS-CoV-2 diagnosis, weeks
March, 14 March, 06–March, 20 March, 11 March, 05–March, 21 March, 15 March, 09–March, 20 March, 10 March, 05–March, 20 March, 18 March, 10–March, 23 0.138

Before March, 8 106 (30.6) 63 (35.2) 27 (21.4) 15 (44.1) 1 (14.3) 0.987

B.1, N = 179 B.1.1, N = 126 B.1.5, N = 34 Othera, N = 7

P-valueb

Overall, N = 346

Table 1 Demographic and clinical findings of the 346 SARS-CoV-2-infected patients according to lineages.
unavailable for this patient, thus hampering the tracing of the exact dynamic of the infection. Looking at the topology of the trees, the European transmission chain probably originated in Spain (Fig. 5 and Supplementary Fig. 4) and continued in France in the second half of February. The patient infected by lineage B was a 92-year-old male with no travel history. The tree topology suggested that the European transmission chain probably started in the UK, and arrived in Italy by means of untracked individuals (Fig. 5 and Supplementary Fig. 4).

Most of SARS-CoV-2 sequences from Lombardy (gray taxa with red dots, 344/346 [99.4%]) are interspersed within B sub-lineages (Fig. 2).

The molecular clock analysis revealed that the most recent ancestor of these viruses dates back to the second half of January (January 20, 95% HPD intervals: Jan, 16–Jan, 23, Supplementary Figs. 4 and 5), in line with evidence of initial outbreaks in other European countries in the second half of January6,7. The Bayesian reconstruction also permitted estimation of the median tMRCA of the COVID-19 pandemic as December 3, 2019 (95% HPD November 19 to December 15, 2019; Fig. 5), consistent with previous estimates6,7.

By both ML and Bayesian tree, lineages B.1, B.1.1, B.1.8, B.1.1.1, and B.1.5 seem descendant from a sequence belonging to lineage B and collected on January 28 in Germany (known as BavPat1 and named as Jan, 28– Germany in the trees). BavPat1 sequence was part of a B local clade described in Bavaria, resulting from a single travel-associated Chinese primary case5,8, and singled out as the initial outbreak triggering the Italian one. However, none of the 346 Italian SARS-CoV-2 sequences here analyzed belonged to lineage B, with the exception of one. This sequence differs from BavPat1 for 6 SNP. Thus, according to recent published evidence9, the lack of a local cluster involving BavPat1 suggested that the initial outbreak in Bavaria was unlikely to have been responsible for directly seeding Lombardy.

Lineage B.1 was initially detected in Lombardy in the second half of February, and in early March was present in the Netherlands, UK, Central Europe, and USA (Supplementary Figs. 4 and 5). Viral sequences were characterized by SNPs at positions 14408 (C to T, non-syn P to L in RdRp) and 23403 (C to T, non-syn D to G in S), with an intra-patient prevalence >95%.

Lineage B.1 was characterized by sustained local transmission, with no evidence of foreign origins, as suggested by the topology of Cluster A. This cluster was characterized by a posterior probability of 0.99 and a tMRCA dated February 14 (Feb, 5–Feb, 20). It was almost exclusively composed of sequences from Lombardy, with the exception of a Swiss 25-year-old patient diagnosed on February 27, a sequence obtained from a Lithuanian patient, whose collection date is partially reported in GISAID (2020-Feb), and a 75-year-old Italian patient diagnosed on March 3 in the Italian Marche region. All Cluster A sequences were characterized by the SNP 26530 (A to G, non-syn D to G in M protein; intra-patient prevalence >90%). The relatively small number of foreign sequences belonging to this cluster, and the strain at the bases of this transmission chain (a sequence collected in Bergamo on March 01), support the assumption, therefore, that the transmission chain started in Lombardy. None of patients involved in this cluster reported international or domestic travels. Two of them (both from Como) stated they had been in contact with a confirmed COVID-19 case, thus confirming the evidence of a local clade.

By taking into account the timing of diagnosis, 40% of the north Italian sequences involved in Cluster A were collected before National lockdown, thus in the early-targeted testing, and 60% in the later-targeted testing (Table 2). The high-risk area of Bergamo allowed this clade to seed in the time-frame preceding lockdown. On the contrary, Milan, where the perception of infection risk was probably lower and where mobility was permitted until March, was mainly involved at a later time-point, after national lockdown.

Overall, these data suggest that potential origin of lineage B.1 was in Lombardy.

Lineage B.1.1 is a clear descendant of lineage B.1 (posterior probability = 1.00) and was simultaneously detected in late February in Denmark, Germany, UK, United States, and Lombardy (Fig. 5 and Supplementary Fig. 4). Two SNPs in N (28881–28883; GGG to AAC, intra-patient prevalence >99%) further characterized the 126 sequences that belong to this lineage. The exact origin of the entire B.1.1 lineage remains difficult to determine, in the light of its simultaneous circulation in such a vast territory. Two transmission chains probably starting in late February/early March were detected within this lineage (Clusters C and D, Fig. 5, Supplementary Fig. 4, and Table 2). Clusters C and D contained sequences characterized by the non-syn SNP A8072G in NSP3 (N to D), and the syn T4579A in pp1a plus the non-syn G18898T in NSP14 (V to F), respectively. One sequence collected in Como in the first days of March is at the basis of Cluster C, which also included two strains isolated in Belgium thereafter. A sequence diagnosed in Atlanta on February 29 carrying the mutation G18898T but not T4579A is at the origin of Cluster D. Unfortunately, no information regarding recent travel history or contacts are available for these GISAID sequences, neither did patients involved in these clusters report international travel (with the exception of one in Cluster C who had traveled to an unknown destination), making it difficult to reconstruct the initial contact tracing contributing to the definition of this clade.

Overall, the simultaneous circulation of lineage B.1.1 in different countries and the presence of transmission chains with probable foreign origin support the hypothesis that this lineage may have seeded Lombardy by means of diverse geographically distinct access sources.

Thirty-four sequences from Lombardy belonged to lineage B.1.5. This lineage can be divided into two main chains, one of them supported by a posterior probability ≥0.98 (Cluster B). Cluster B involved a complex tracing network of 24 persons, who were infected in Lombardy (mainly in Pavia), and diagnosed after March 8. All sequences were characterized by the SNP 23575 (C to T, non-syn T to I in S protein; intra-patient prevalence: >99%). The earliest and most closely related strain was a sequence from China collected on January 24 in the province of Zhejiang. In line with this, the most recent ancestor of this cluster dates back to January 22 (Jan, 18–Jan, 23), while the transmission chain probably started on February 12 (Feb, 6–Feb, 23) (Fig. 5, Supplementary Fig. 4, and Table 2).

The other chain was composed of a total of 26 sequences (11 from Lombardy), all of them characterized by the SNP at position 20268, A to G, syn (ns15, intra-patient prevalence: 50%). Sequences from Lombardy were intermixed with sequences from central and east Europe (n = 10), eastern and west Asia (n = 2), North America (n = 1), UK and the Netherlands (n = 2). Looking at the topology of the trees these sequences seem to be closely related to a subgroup of B.1 sequences collected in late February in Lombardy (e.g. Pavia, February 23, and Cremona, February 24 even if the support value is <0.10). The low posterior probability and the absence of information regarding recent travel history makes it hard to understand if a common origin really exists or if their clustering is due to multiple introductions of genetically similar viruses from geographically distinct sources.

Sub-lineages B.1.8 and B.1.1.1 were detected in one and four individuals, respectively. The patient infected by lineage B.1.8 was
an 84-year-old male from Lodi with no travel history, diagnosed on March 5. This infection was preceded by another detected in the Netherlands on March 4 (Fig. 5 and Supplementary Fig. 4, posterior probability = 0.99). This lineage probably originated from B.1 sequences detected in Lombardy in late February (Fig. 5 and Supplementary Fig. 4). Patients infected by lineage B.1.1.1 come from Milan and Pavia, and did not report any travel history. A strain collected in UK on March 13 is at the basis of the B.1.1.1 cluster (Fig. 5 and Supplementary Fig. 4, posterior probability = 1.00), thus suggesting a probable foreign origin for this lineage.

Discussion
These data on the genomic epidemiology of SARS-CoV-2 in Lombardy, based on a consistent number of whole genome SARS-CoV-2 sequences circulating in a single region, indicate the simultaneous circulation of at least three widely represented lineages of SARS-CoV-2 (B.1, B.1.1, and B.1.5), supported by transmission chains occurring since the first half of February (Table 2, Fig. 5, Supplementary Fig. 4).

Even if these lineages were sizeable and spread rapidly, they showed different geographical distributions. Lineage B.1, probably characterized by origin in Lombardy, mostly affected Lodi, Cremona, and Mantua in the south, and accounted for 90.2% of sequences coming from these provinces (55/61, Fig. 1 and Table 1). This is a less densely populated area, with less intense interconnections with Europe and beyond, making a modest lineage diversification plausible. Lineage B.1.1 predominated in the north of Lombardy, mostly in Como, Lecco, Bergamo, and its adjacent territories such as Alzano and Nembro, and accounted...
for 62.7% of sequences coming from these provinces (84/134, Fig. 1 and Table 1). This is a densely populated area, served by dense motorway and railway networks, and by two international airports. Thus, we cannot exclude that foreign introductions contributed to the spread of this lineage, a hypothesis supported by the characteristics of the local clusters identified within it. Lastly, lineage B.1.5 was mainly detected in Pavia, because of a local cluster detected in this area which involved 23 B.1.5 sequences. Of note, Milan, the most important urban area of Lombardy, is the only province in which all lineages are equally represented (Table 1, \( P = 0.108 \) by Chi-square test for trend).

Given the central location of Milan and its intense economical and transportation links to Europe and beyond, the simultaneous circulation of the different lineages in this area would seem plausible.

These data are in line with recent accounts of the dynamics of the SARS-CoV-2 pandemic in densely populated areas like New York State, Belgium, and (more recently) the Netherlands\(^1\). The same scenario was also observed in less densely populated areas characterized by regular international travel, such as Iceland\(^2\), where the government managed to contain the spread of the virus thanks to a strategy of aggressive testing, contact tracing, and quarantine. To contain and mitigate the COVID-19 epidemic, Italy has been the first European country to implement unprecedented measures to restrict individual mobility, and to promote social distancing, with the aim of interrupting transmission of the SARS-CoV-2 virus. Following the detection of the first case of COVID-19 cases in Lombardy on February 21, the government implemented an ever-increasing number of emergency decrees, which culminated in the complete lockdown on March 8.

Due to the intense commercial relations between China and the rest of the world, thousands of Chinese and Europeans traveled from China to Europe and vice versa during January. In Italy, direct flights from China were stopped on January 31, while indirect access from China through other routes was suspended only later (see below). However, SARS-CoV-2-infected people could enter Italy long before this date, as was proven by the diagnosis of SARS-CoV-2 in Chinese tourists who arrived in Italy on January 23. In addition, the suspension of direct flights from China could not prevent the entrance of people though transit flights from other countries\(^3\). The analysis of the COVID-19
outbreak and the modeling assessment of the effects of travel limitations also confirmed that, despite the comprehensive restrictions on travel to and from China since January 23, 2020, many individuals exposed to SARS-CoV-2 have been traveling internationally without being detected. Lombardy, with three international airports (e.g. Milano, Linate, Bergamo), is Italy’s greatest and most prosperous international business hub, constantly hosting international events. It is thus very likely that a number of infected people, both directly from China and also through other countries, entered Lombardy in that period. The untraced entry of SARS-CoV-2 might have facilitated the diffusion of the virus at population level and its spread at local level, as suggested by the epidemiological characteristics of the identified local clusters, mainly sustained by patients with no travel history. Our data allow accurate interpretation of the COVID-19 outbreak in Lombardy and its transmission dynamics also, thanks to a revised phylogenetic approach, although some national-level data have been published (with limited sampling and details). In particular, the circulation of SARS-CoV-2 in the country (an area of 300,000 km², and a population of 60 million) was described by analyzing a lower number of whole-genome sequences. Here, we focused the attention on a territory of 24,000 km², with a population of 10 million, which accounted for 37% of COVID-19 cases and 53% of deaths in the country. Lombardy is also the first region to have been affected by the COVID-19 outbreak, with a rate of 112.9 deaths per 100,000 population, almost six times higher than in the rest of Italy. The first phase of epidemic also saw the emergence of many cases within an extremely short period of time, a scenario that does not necessarily favor the simultaneous and massive circulation of different lineages. The effort to sequence and implement with epidemiological data 346 SARS-CoV-2 sequences endemic in this territory made it possible to describe the heterogeneity of the virus circulating in Lombardy, and ultimately to appreciate local sustained transmissions and to estimate the time of their probable introduction.

Looking at the correlation between lineages and COVID-19, no trend of association between disease severity (including evidence

### Fig. 4 Prevalence of most representative single nucleotide polymorphisms (SNPs) in SARS-CoV-2 genomes isolated in Lombardy according to geographical locations and their intra-patient prevalence.

**a** Frequency of single nucleotide polymorphisms (with respect to the Wuhan reference genome NC_045512.2) among SARS-CoV-2 sequences according to different geographical location. Italian sequences include the 6 and 8 sequences from North and Central Italy present in Gisaid at April 10.

**b** Intra-patient prevalence of most representative SNPs in North Italian SARS-CoV-2 sequences according to sample date. Box and whiskers plots indicate the median and interquartile range of intrapatient prevalence. SNP: single nucleotide polymorphism; syn: synonymous; non-syn: non-synonymous; nsp15: non-structural protein 15; S: spike; M: matrix; N: nucleocapsid; RdRp: RNA-dependent RNA polymerase.
Fig. 5 Bayesian phylogeographic reconstruction incorporating location and travel history data of the most informative sequences for virus spread and clustering in Lombardy. SARS-CoV-2 genomes from Lombardy were shown by red circles. Local clusters of SARS-CoV-2 sequences supported by a posterior probability $\geq 0.98$ by the time-scaled maximum clade credibility tree were highlighted in red (A), in blue (B), in cyan (C), and in orange (D). Two independent chains were run for 25 million states. Parameters and trees were sampled every 1000 states.
of interstitial pneumonia, and severe or critical presentation) and lineages was detected (Table 1).

Looking at the overall variability of SARS-CoV-2, we found that only six SNPs (five out of six non-synonymous)\(^18\)\(^{--}\)\(^20\) characterized our consensus sequences, highlighting a good conservation rate of the virus over time. This conservation rate is confirmed within the spike structural protein, where only two mutations, one of them at low prevalence (i.e., C23575T, corresponding to the amino acid variant T671I), were detected. None of these mutations have a role in altering pre-existing N-glycosylation sites or in creating new ones\(^2^1\), a factor which may be beneficial for the development of vaccines strategies.

Worthy of mention is the SNP A to G at position 23403, corresponding to the variant D614G in spike protein, detected in 99.4% of our SARS-CoV-2 sequences (all with an intra-patient prevalence >99%). As is already known, this variant is observed frequently in European countries, such as the Netherlands, Switzerland, and France, but seldom observed in China. As recently reported, this variant, located within a B-epitope, causes the substitution of a large acidic residue (aspartic acid), with a small hydrophobic residue (glycine). Its rapid fixation at population level might thus suggest a role in viral entry, and enhancement of interaction between receptor-binding-domain of the S protein with the entry receptor ACE2 (ref.\(^2^2\)). In line with this evidence, D614G variant was recently associated with lower Ct values, indicative of potentially higher in vivo viral loads\(^2^3\). The presence of D614G in almost all the sequences sampled in Lombardy suggests the dominance of this mutation since the first phases of epidemic in Lombardy, and in Europe, confirming its rapid spread and persistence. Unfortunately, the fixation of this mutation in our region prevented us from defining any association with disease status and Ct values. However, the quite low Ct values observed in our population (median, IQR: 18.8, 16.8–20.1) could be in line with the association of D614G variant with lower Ct values, indicative of potentially higher in vivo viral loads\(^2^3\), and thus related to severe outcome in COVID-19 (refs.\(^2^4\)\(^--\)\(^2^6\)).

Our study has some limitations. The analysis of phylogenetic structures during such an early phase of the pandemic should be interpreted carefully, as the number of mutations that define phylogenetic lineages is small and may be similar to the rate of potential errors introduced during reverse transcription, PCR amplification, or sequencing\(^2^7\). To overcome these problems, Bayesian approach, recognized as a powerful way to estimate species divergence\(^2^8\), and thus expected to provide more robust results, was applied. Moreover, the integration of host characteristics (such as geographical location, collection date, and clinical manifestations) aided phylogenetic interpretation.

The intra-host variability of SARS-CoV-2, and the role of potential existing minority variants, has not been investigated here. Initial evidence suggests that intra-host variation of SARS-CoV-2 can be frequently found among clinical samples (median number of intra-host variants: 1–4), but at the same time these variants were not observed in the population as polymorphisms, probably suggesting a bottleneck or purifying selection was involved\(^2^9\)\(^,\)\(^3^0\). Thus, ad hoc designed studies are necessary to provide an extensive overview of SARS-CoV-2 intra-host variability and minority variants description, if and how these minority variants can spread in the population, and their potential role in virulence and transmissibility.

In the peak of SARS-CoV-2 epidemic, diagnosis was mainly addressed to symptomatic cases (prevalently older people with no travel history) or subjects at high risk of exposure (i.e. health care workers exposed to positive patients without adequate protection). This approach may have caused a substantial underestimation of positive subjects, precluding the possibility

### Table 2 Distribution of SARS-CoV-2 Local Clusters, according to lineages, timing of diagnosis, and reported risk.

| Lineage | Before March 8 | After March 8 | Tot |
|---------|---------------|---------------|-----|
| B       | 10 (46.6)     | 66 (66.4)     | 76  |
| C       | 9 (39.1)      | 70 (70.2)     | 79  |
| D       | 2 (8.8)       | 30 (30.1)     | 32  |

### Table 3 Details of SARS-CoV-2 Lineages, according to sample.

| Lineage | Before March 8 | After March 8 | Tot |
|---------|---------------|---------------|-----|
| B       | 10 (46.6)     | 66 (66.4)     | 76  |
| C       | 9 (39.1)      | 70 (70.2)     | 79  |
| D       | 2 (8.8)       | 30 (30.1)     | 32  |
extending our analysis to asymptomatic infections, whose role in the transmission chains and in influencing the evolution of epidemic remains a challenge to investigate.

Moreover, even if the number of samples here analyzed is noteworthy, the epidemic in the east (i.e. Brescia and Mantua) and the valleys of the north (i.e. Valtellina and Valcamonica, Fig. 1) is poorly represented. Yet, with the sole exception of Brescia, which resembled Bergamo, Cremona, and Lodoi in how SARS-CoV-2 spread, Mantua and the valleys accounted for only the 6.7% of confirmed COVID-19 cases in Lombardy, a percentage that does not constitute a major issue for the good representability of SARS-CoV-2 epidemic in the region. Even if the sampling scheme used in this study ( Supplementary Fig. 1, Supplementary Table 1, and Supplementary Results) suggests a good representativeness of the selected population in relation to the general population, we cannot exclude that denser sampling (both temporally and spatially) would reveal novel dispersal patterns not observed here and further address lineages movement.

In conclusion, this study contributes to the identification and circulation pattern of seven SARS-CoV-2 lineages in an area highly affected by COVID-19, which caused more than 16,000 deaths in Lombardy in a number of weeks. Of these seven lineages, three were widely represented and at least two likely originated in this region, suggesting that the virus was circulating undetected for some time before first detection and confirming a central role of Lombardy in the SARS-CoV-2 epidemic. We cannot exclude that this multiple and simultaneous circulation of SARS-CoV-2 strains can have exacerbated the transmissibility potential of the virus and thus created a real viral storm in such a densely populated region.

Methods
Sample collection and epidemiological data. This retrospective observational study included 371 SARS-CoV-2-positive nasopharyngeal-swabs obtained from adult patients hospitalized or referred for the diagnosis at two major hospitals in Lombardy from February 22 to April 4. Demographics, epidemiological and clinical data were obtained retrospectively by pseudonymized electronic medical records. The study protocol was approved by local Research Ethics Committee of the Niguarda and San Matteo hospitals (prot. 92-15032020 and P_2020002940).

Virus genome assembly. Reference-based assembly of the raw data was performed as follows: Illumina adapters were removed, and reads were filtered for quality (average q28 threshold and read length >135 nt) using FASTP43. First and last 15 nucleotides were then removed from all reads. The mapping of cleaned reads was performed against the GenBank reference genome NC_045512.2 (Wuhan, collection date: December 2019) using BWA-mem46. Reads mapping on SARS-CoV-2 reference genome were a median of 99.7% (IQR: 99.5%-99.8%), and were able to cover from 94.0% to 99.7% of the SARS-CoV-2 reference genome (GenBank: NC_045512.2), independently of SARS-CoV-2 load (Supplementary Fig. 2a, b). The few genome regions (N=4) with lower reads coverage were consistently limited to no more than 35 nt. SNP variants were called through a pipeline based on samtools/bcftools47, and all SNPs having a minimum supporting read frequency of 40% with a depth ≥50 were retained. Consensus was generated using the github freely distributed software vcf_consensus_builder (https://github.com/peterk87/vcf_consensus_builder).

Phylogenetic analysis. SARS-CoV-2 lineages of the 346 SARS-CoV-2 consensus sequences obtained were assigned according to the PANGOLIN application (Pangolin. https://pangolin.cog-uk.io)48. In order to represent the global diversity of the lineages by the end of April 2020 while minimizing the impact of sampling bias, 395 GISAID deposited sequences were added to the consensus sequences obtained by our samples. Sequences were aligned using ClustalX and manually inspected in Bioedit. The final alignment was composed of 741 sequences 29,159 nucleotides long.

To account for regions which might potentially be the result of hypervariability or sequencing artifacts, alignment positions showing significant homoplasy were identified by a combined approach. Homoplases were firstly identified using HomoplasyFinder, and then confirmed by Treetime (homoplaspy setting)49,50.

In order to explore the phylogenetic structure of SARS-CoV-2, the 741 sequences were analyzed by the maximum likelihood (ML)40,41 and Bayesian coalescent43 methods, under the best-fit model of nucleotide substitution GTR+I4. The Bayesian coalescent tree analysis was undertaken with BEAST v1.10.5 (ref. 43), using the GTR+I substitution model with an exponential population growth tree prior and strict molecular clock, under a noninformative continuous-time Markov chain (CTMC) reference prior44.

The information regarding location and recent travel history (less than 1 month before the SARS-CoV-2 diagnosis) of the most informative sequences for virus spread and clustering identified in the first Bayesian tree were incorporated in a second Bayesian tree inference42, in order to yield more robust reconstructions of virus spread.

For further details regarding the homoplaspy checking, the criteria used for the selection of GISAID sequences, and phylogenetic methods, refer to Supplementary Methods.

Data availability
Nature Research reporting summary, source data, and supplementary information were uploaded as related manuscript files. The 346 SARS-CoV-2 sequences obtained in this study are openly available on GISAID portal and European Nucleotide Archive under the accession numbers EPI_ISL_542096–EPI_ISL_542443 and ERX4545164–ERX4548732, respectively. The list of accession numbers is available in the Supplementary Data 3. The list of the whole-genome SARS-CoV-2 sequences (n=3,244) retrieved from GISAID (gsaids.org) on 3 May 2020 and their corresponding accession numbers and lineages are available in the Supplementary Data 1. The list of the 395 selected GISAID SARS-CoV-2 sequences considered for phylogenetic analyses and their corresponding accession numbers and lineages are available in Supplementary Data 2. The list of the significant homoplasies positions identified by HomoplasyFinder and TreeTime are available in the Supplementary Data 4. Lombardy map was retrieved from the https://raw.githubusercontent.com/blackmad/neighborhoods/master/lombardy.geojson. COVID-19 cumulative case data were obtained from the World Health Organization (WHO) COVID-19 database (http://www.safet.gov.it/portal/novocoronavirus/dettagliocontenuti/Novocoronavirus.jsp?lingua=english&id=5367&area=nuovoCoronavirus&menu=nuovo).

General population data were retrieved from the Italian National Institute of Statistics (http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1). The de-identified data regarding demographic and clinical features related to each patient are available on reasonable request from the corresponding author. Data are not publicly available in order to fully comply to privacy guarantee. Source data are provided with this paper.
References
1. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269 (2020).
2. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273 (2020).
3. Eurostat. The statistical office of the European Union statistics. https://ec.europa.eu/eurostat/statistical-abstracts/2020/virus-coronavirus-2-in-patients-with-coronavirus-disease-2019 (2020).
4. Rambaut, A. Phylodynamic Analysis, 176 genomes. Preprint at https://doi.org/10.5281/zenodo.4306802 (2020).

Received: 17 June 2020; Accepted: 10 December 2020; Published online: 19 January 2021

ARTICLE
