

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

3 Claudia Alteri^{1§}, Valeria Cento^{2§}, Antonio Piralla^{3§}, Valentino Costabile⁴, Monica Tallarita³, Luna
4 Colagrossi⁵, Silvia Renica¹, Federica Giardina³, Federica Novazzi³, Stefano Gaiarsa³, Elisa
5 Matarazzo², Maria Antonello¹, Chiara Vismara⁶, Roberto Fumagalli⁷, Oscar Massimiliano Epis⁸,
6 Massimo Puoti⁹, Carlo Federico Perno^{1,6✉}, Fausto Baldanti^{3,10}

7 ¹Department of Oncology and Hemato-oncology, University of Milan, Milan, Italy

8 ²Residency in Microbiology and Virology, University of Milan, Milan, Italy

9 ³Molecular Virology Unit, Microbiology and Virology Department Fondazione IRCCS Policlinico San
10 Matteo, Pavia, Italy

11 ⁴Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy

12 ⁵Department of Laboratories, Bambino Gesù Children's Hospital, Rome, Italy

13 ⁶Chemico-clinical and Microbiological Analyses, ASST Grande Ospedale Metropolitano Niguarda,
14 Milan, Italy

15 ⁷Department of Anesthesiology, Critical Care and Pain Medicine, ASST Grande Ospedale
16 Metropolitano Niguarda, 20162, Milan, Italy

17 ⁸Rheumatology Unit, ASST Grande Ospedale Metropolitano Niguarda, 20162, Milan, Italy

18 ⁹Infectious Diseases, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy

19 ¹⁰Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia,
20 Italy

21 § These authors contributed equally: CA, VC, AP

22 ✉ email: cf.perno@uniroma2.it

23 **Running head: Genomic epidemiology of SARS-CoV-2 in Lombardy, Italy**

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

24 **Abstract**

25 From February to April, 2020, Lombardy (Italy) was the area who worldwide registered the highest
26 numbers of SARS-CoV-2 infection. By extensively analyzing 346 whole SARS-CoV-2 genomes, we
27 demonstrated the simultaneous circulation in Lombardy of two major viral lineages, likely derived
28 from multiple introductions, occurring since the second half of January. Seven single nucleotide
29 polymorphisms (five of them non-synonymous) characterized the SARS-CoV-2 sequences, none of
30 them affecting N-glycosylation sites. These two lineages, and the presence of two well defined
31 clusters inside Lineage 1, revealed that a sustained community transmission was ongoing way
32 before the first COVID-19 case found in Lombardy.

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47 **Introduction**

48 Since coronavirus disease 2019 (COVID-19), was initially reported in China on 30th December
49 2019^{1,2}, SARS-CoV-2 spreads world-wide and, as of 14th June 2020, there have been 7.55 million
50 confirmed infections and 423,000 deaths reported worldwide (World Health Organization, 2020).

51 In Italy, the first case of evident SARS-CoV-2 transmission occurred at February 20th, in Codogno,
52 Lombardy, when a young man affected by an interstitial pneumonia was diagnosed for SARS-CoV-
53 2. Since that day, the number of diagnosed COVID-19 cases exponentially increased and Lombardy,
54 become the area most affected by the COVID-19 pandemic, with a total number of 89,018 infections
55 at June 2, out of 233,197 total cases in Italy. Thereafter, the COVID-19 epidemic grew exponentially
56 during the first days of March 2020, peaking on 21 March 2020 with 6,557 newly confirmed cases
57 per day. Two months later, reported COVID-19 cases in Italy dropped to ~600 per day, indicating
58 the epidemic is close to be contained.

59 Lombardy is the most populated region in Italy (10 million inhabitants) and one of the widest. Milan
60 represents the largest metropolitan area in Italy and the third most populated functional urban area
61 in Europe with a strong economical and transportation links to Europe and outside. This scenario
62 makes Lombardy the perfect place to host and favor the spread of a highly transmissible virus, such
63 as SARS-CoV-2.

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

70 **Methods**

71 **Sample collection, and epidemiological data**

72 This retrospective cohort study included 371 SARS-CoV-2-positive nasopharyngeal-swabs of adult
73 patients hospitalized or referred for the diagnosis at two major Hospitals in Lombardy since February

74 22, to April 4. Clinical data were obtained retrospectively by pseudonymised electronic medical
75 records. The study protocol was approved by local Research Ethics Committee of the two hospitals
76 (prot. 92-15032020 and P_20200029440). This study was conducted in accordance with the
77 principles of the 1964 Declaration of Helsinki.

78 The severity of the disease was classified into mild, moderate, or severe, if showing i) mild clinical
79 symptoms without sign of pneumonia on imaging, ii) fever and respiratory symptoms with radiological
80 findings of pneumonia, iii) respiratory distress, with oxygen saturation $\leq 93\%$ at rest, mechanical
81 ventilation, or presence of multiorgan failure (septic shock) and/or admission to intensive care unit
82 (ICU) hospitalization.

83 **Virus amplification and sequencing**

84 Total RNAs were extracted from nasopharyngeal swabs by using QIAamp Viral RNA Mini Kit,
85 followed by purification with Agencourt RNAClean XP beads. Both the concentration and the quality
86 of all isolated RNA samples were measured and checked with the Nanodrop. Virus genomes were
87 generated by using version 1 of the CleanPlex SARS-CoV-2 Research and Surveillance Panel
88 (<https://www.paragongenomics.com/product/cleanplex-sars-cov-2-panel/>), according to the
89 manufacturer's protocol starting with 50 ng total RNA and followed by Illumina sequencing on a
90 NextSeq 500. Briefly, the multiplex PCR was performed with 2 pooled primer mixture and the cDNA
91 reverse transcribed with random primers was used as a template. After 10 rounds of amplification,
92 the two PCR products were pooled and purified. Then the digestion reaction was performed to
93 remove non-specific PCR products, followed with second PCR reaction for barcoding with 24 rounds
94 of amplification. Libraries were checked using High Sensitivity Labchip and quantified with Qubit.
95 Equimolar quantity of libraries was pooled, and the obtained run library mix was loaded at 1.5pM
96 into NextSeq500 for sequencing in the Mid Output format with paired-end 2 x 150 bp. The Illumina
97 sequencing platform takes less than 26 hours to obtain 30.2 Gb of sequencing data (compressed
98 format), achieving between 170.000 and 856.000 paired-end fragments per sample (340.000 and
99 1.712.000 sequences), with a mean coverage depth of 2.500.

100 **Virus genome assembly**

101 Reference-based assembly of the metagenomic raw data was performed as follows. Illumina
102 adapters were removed, and reads were filtered for quality (average q28 threshold and read length
103 > 135 nt) using FASTP.³ First and last 15 nucleotides were then removed from all reads. The
104 mapping of cleaned reads was performed against the GenBank reference genome NC_045512.2
105 (Wuhan, collection date: December 2019) using BWA-mem⁴, and consensus sequences were
106 generated using samtools 1.10.⁵ Single nucleotide polymorphisms (SNP variants) were called
107 through a pipeline based on samtools/bcftools⁶, and all SNPs having a minimum supporting read
108 frequency of 40% with a depth ≥ 100 were retained in the consensus sequence.

109 **Phylogenetic analysis**

110 The consensus sequence obtained merging the most represented SNP (intra-patient prevalence
111 >40%) was retained per patient. Moreover, all available whole-genome SARS-CoV-2 sequences (n
112 = 3244) on GISAID (gisaid.org) on 3 May 2020 were downloaded. Sequences from GISAID that
113 were error-rich, those without a date of sampling and identical sequences from each country
114 outbreak were removed. Finally, the dataset was reduced by only retaining the earliest, and the most
115 recently sampled sequences from each country outbreak (range of dates: December, 24 2019 –
116 April, 4 2020). The resulting dataset of 205 GISAID sequences therefore represents the global
117 diversity of the virus while minimizing the impact of sampling bias. The 205 GISAID deposited
118 sequences were added to the consensus sequence obtained by our samples. Sequences were
119 aligned using ClustalX and manually inspected in Bioedit. The final alignment length was 29,282
120 nucleotides. We used both the maximum likelihood (ML) and Bayesian coalescent methods to
121 explore the phylogenetic structure of SARS-CoV-2. The ML phylogeny was estimated with RAxML⁷
122 using the under the best-fit model of nucleotide substitution GTR+I⁸ with gamma-distributed rate
123 variation⁹. Tree topology was assessed with the fast bootstrapping function with 1000 replicates. The
124 ML tree was inspected in TempEst¹⁰, in order to define the correlation between genetic diversity
125 (root-to-tip divergence) and time of sample collection (Supplementary Figure 1). In order to obtain a
126 corresponding time-scaled maximum clade credibility tree, a Bayesian coalescent tree analysis was
127 undertaken with BEAST v1.10.4,¹¹ using the HKY+Q4 substitution model with gamma-distributed

128 rate variation with an exponential population growth tree prior and an uncorrelated relaxed
129 molecular clock, under a noninformative continuous-time Markov chain (CTMC) reference prior¹².
130 Taxon sets were defined and used to estimate the posterior probability of monophyly and the
131 posterior distribution of the tMRCA of observed phylogenetic clusters. Four independent chains were
132 run for 50 million states and parameters and trees were sampled every 1,000 states. Upon
133 completion, chains were combined using LogCombiner after removing 10% of states as burn-in and
134 convergence was assessed with Tracer. The maximum clade credibility (MCC) tree was inferred
135 from the Bayesian posterior tree distribution using TreeAnnotator, and visualized with FigTree 1.4.4
136 (<http://tree.bio.ed.ac.uk/software/figtree/>). Monophyly and tMRCA (time to the most recent common
137 ancestor) statistics were calculated for each taxon set from the posterior tree distribution.

138 **Statistical Analysis**

139 Data were analyzed using Rgui and the statistical software package SPSS (v32.0; SPSS Inc.,
140 Chicago, IL).

141 **Data availability**

142 Sequences are in the process to be deposited in GISAID (<https://www.gisaid.org/>) and will be
143 completed by acceptance of the manuscript. Additional data that support the findings of this study
144 are available from the authors upon request.

145 **Results**

146 **Patients' characteristics**

147 From February 22 through April 4, 2020, nasopharyngeal swabs of a total of 25,082 patients were
148 screened for SARS-CoV-2 infection at two major Hospitals in Lombardy. Of them, 11,445 received
149 a diagnosis of COVID-19. Whole genome sequencing was performed in a total of 371 samples
150 collected from 371 patients residing in all 12 provinces of Lombardy and with varying disease
151 symptoms, ranging from mild to severe. Twenty-five samples were excluded due to failed
152 amplification (n=9), or poor genomic coverage (<60%, n=16). The final study population thus
153 consisted of 346 patients, whose demographic and clinical characteristics are reported in Table 1.
154 This study population was well representative of the whole Lombardy region, with exception of the

155 East part and North valleys (i.e. Brescia, Mantua, Valtellina and Valcamonica, Figure 1). Two-
156 hundred patients (57.8%) were male, and the median age was of 72 (IQR: 53-83) years. Fever was
157 the most common COVID-19 symptom at admission, followed by cough and dyspnea. Chest
158 radiographs or CT scan confirmed a classical bilateral interstitial pneumonia for 24.1% of them.
159 Patients with severe COVID-19 manifestation more frequently suffered from at least one chronic
160 comorbidity, compared to those with moderate and mild manifestations ($p=0.002$), with greater
161 impact played by hypertension and chronic kidney disease ($p<0.001$).

162 **Genome Coverage**

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

167 **Single nucleotide polymorphisms characterizing North Italian sequences**

168 The genetic pairwise distance of the 346 sequences indicated that the SARS-CoV-2 sequences
169 evolved progressively during time ($\rho=0.465$, Supplementary Figure 1 and 3). Seven single
170 nucleotide polymorphisms (SNP) were shared among >5 genomes, with a prevalence ranging from
171 3.6% (20268, A to G, syn in nsp15) to 99.2% (14408, C to T, non-syn P to L in RdRp) (Figure 2 A
172 and B). All SNPs have been detected in at least one previously published SARS-CoV-2 sequence
173 in GISAID. Five out of these seven SNPs were non-synonymous, and three mapped within SARS-
174 CoV-2 structural proteins. Notably, 2 non-synonymous SNPs reside in S protein: the C to T at
175 position 23575 (S protein; amino acid T to I; intra-patient prevalence $\sim 100\%$) found in 23 North Italian
176 sequences and previously detected in less than 1% of samples in China, Europe and North America,
177 and the A to G at position 23403 (S protein; amino acid D to G; intra-patient prevalence $\sim 54\%$),
178 present in 67.8% of North Italian sequences and firstly detected in China with a prevalence of 1.7%,
179 then rapidly selected and spreading in all countries with a prevalence ranging from $\sim 17\%$ in Asia to
180 $\sim 70\%$ in Europe and North America.

181 **Phylogenetic estimates and lineages characteristics**

182 To explore the distribution of SARS-CoV-2 sequences in Lombardy we performed both an estimated
183 maximum likelihood (ML) phylogeny (Figure 3), and a Bayesian molecular clock analysis (Figure 4).
184 ML tree showed that most of SARS-CoV-2 sequences from Lombardy (grey taxa with red dots,
185 342/346 [98.8%]) are interspersed within two major viral lineages (1, 2) containing previously isolated
186 Italian strains, and strains from other European, American and Asian countries (gray branches, no
187 dots) (Figure 3). Demographic and clinical characteristics of patients infected with SARS-CoV-2 from
188 such lineages are reported in Table 1.

189 According to the topology of the ML tree, the genetic distance from SARS-CoV-2 reference strain
190 was lower for isolates in Lineage 1 respect to isolates in Lineage 2 (2.7×10^{-4} [2.0×10^{-4} ; 3.7×10^{-4}] vs
191 3.4×10^{-4} [2.7×10^{-4} ; 4.1×10^{-4}], $P < 0.001$, Supplementary Figure 3B), thus indicating a closer genetic
192 relatedness with original strains for Lineage 1 respect to Lineage 2. The mean genetic pairwise
193 distance inside Lineages was broadly comparable (3.9×10^{-4} vs 3.5×10^{-4} , $P = 0.118$), suggesting
194 similar evolutionary characteristics and replication profiles.

195 These lineages did not contain viral strains isolated in the first months of the outbreak in China (black
196 branches, no dots); this let us hypothesize a transmission chain not directly involving China (i.e., the
197 country where the pandemic originated). Notably, the most closely related viral isolate that clustered
198 outside such Lineages was isolated in Central Europe, in the second half of January. The 61.0% of
199 Lombardy sequences (N=211) were positioned in Lineage 1, interspersed within sequences from
200 West-North Europe, North and South America, and Asia. Of note, most of isolates collected in the
201 South of Lombardy, mainly in Lodi and Cremona, were involved in this lineage (the 93.7% [30/32]
202 and 88.5% [23/26], respectively). In line with this, sequences from Lodi and Cremona with a
203 collection date between February 20 and 22 were the earliest isolates detected in Lineage 1, together
204 with isolates from Central Europe, the Netherlands and South America, at the end of February
205 (February 25 and 29).

206 The remaining 37.9% of Lombardy sequences (N=131) composed the Lineage 2, together with
207 sequences from European, South American and Asian countries. As Lineage 1 contained most of
208 sequences from the South of Lombardy, Lineage 2 contained the 61.2% of isolates collected in the
209 North, mainly in Lecco (66.7%), Bergamo (61.1%) and Como (58.8%). In line with this, earliest

210 sequences in Lineage 2 included isolates from Bergamo collected between 25 and 29 February, and
211 sequences from Central and North Europe (February 25 and March 2).

212 This suggests that the initial spread of viral strains belonging to the two Lineages involved mainly
213 the territory in which the first access occurred, with little communication between the two highly
214 affected areas.

215 The Bayesian molecular clock analysis estimated a mean evolutionary rate of 1.47×10^{-3}
216 ³subs/site/year (95% HPD, 1.34×10^{-3} - 1.62×10^{-3}). The current low genetic diversity of SARS-CoV-2
217 genomes worldwide implies a very low posterior probabilities (<0.50) for most of the nodes.
218 Concordantly, SARS-CoV-2 sequences in Lineage 1 did not form a monophyletic group (Figure 4).
219 Nevertheless, it was possible to identify 2 transmission clusters with posterior probability support of
220 1 (Cluster A and B), characterized by the SNP A26530G in M (intra-patient prevalence ~98%) and
221 A20268G in nsp15 (intra-patient prevalence ~54%), respectively. In both clusters, Lombardy
222 sequences (24 for Cluster A [Bergamo: 9, Lecco: 5, Milan: 5; Como: 3; Lodi: 1; Cremona: 1], and 6
223 for Cluster B [Pavia: 2; Como: 1; Milan: 1; Cremona: 1; Brescia: 1]) were intermixed with sequences
224 sampled by other countries (Europe for Cluster A, Europe and West Asia for Cluster B). This further
225 supports the hypothesis that Lineage 1 viruses might be imported into Lombardy on multiple
226 occasions.

227 Differently by Lineage 1, the Lineage 2 was characterized by a posterior probability of 1 in the
228 maximum credibility tree (i.e., SARS-CoV-2 sequences grouped monophyletically in 100% of trees
229 in the posterior sample, Figure 4). Two SNPs in N (28881-28883:GGG>AAC) characterized the 131
230 sequences involved in Lineage 2. Of note, these mutations were firstly identified in China, and then
231 spread in all countries, reaching the 37.6% in Italy (intra-patient prevalence:100%) (Figure 2A).
232 Single nucleotide polymorphisms (SNPs), the non-syn A26530G in M (intra-patient prevalence
233 ~98%) and the syn A20268G in nsp15 (intra-patient prevalence ~54%), firstly identified in South Asia
234 and Europe (Figure 2A), characterized sequences in Cluster A by Cluster B.

235 From the molecular clock analysis, we were able to estimate the times of the most recent common
236 ancestor (tMRCA) of all monophyletic traces. Lineage 2 has earlier tMRCAs that coincide with 24th
237 January 2020 (95% HPD 20 January to 28 January 2020). Clusters A and B have few days later

238 tMRCAs, around 27th January 2020 (95% HPD 23 January to 31 January 2020) and 31th January
239 2020 (95% HPD 26 January to 5 February 2020) (Figure 4). By looking at the distribution of
240 Lombardy sequences in the Bayesian Tree, it is thus conceivable that the circulation of SARS-CoV-
241 2 started in this region with at least two different and nearly contemporaneous foci in the second half
242 of January, one month before the first COVID-19 diagnosis in Codogno (Lodi). This is in line with
243 evidences of initial outbreaks in other European countries, including Germany, in the second half of
244 January.¹³

245 The Bayesian reconstruction also allowed to estimate the median tMRCA estimate of the COVID-19
246 pandemic that was 4 December 2019 (95% HPD 29th November to 21th December 2019; Figure 4),
247 consistent with previous estimates.¹⁴⁻¹⁵

248 **Discussion**

249 These data on the genomic epidemiology of SARS-CoV-2 in Lombardy, based on the largest number
250 of whole genome SARS-CoV-2 sequences generated in a single study, indicate the simultaneous
251 circulation of two lineages of SARS-CoV-2 likely starting from multiple introductions, which occurred
252 in the second half of January (Figure 3 and 4), one month before the first COVID-19 case detected
253 in Codogno (Lodi). Of note, the ML tree showed that these two chains of transmission were wide in
254 size and fast in spreading.

255 The MRCAs of these two SARS-CoV-2 lineages were closely related in time, as they dated just one
256 month before the first diagnosed case, yet they showed different preferential geographical
257 circulations. Lineage 1 mostly interested the Southern Lombardy, including Lodi and Cremona (Table
258 1), while Lineage 2 predominated in the North of Lombardy, mostly in Bergamo and its adjacent
259 territories (such as Alzano and Nembro, that represented a major focus of the epidemics). The
260 predominance of these lineages in different territories, the lack of a monophyletic signal for Lineage
261 1, and the detection of two well supported clusters inside it (Figure 4), supports the hypothesis of
262 multiple and dislocated introductions of SARS-CoV-2 with different route of viral transmission. These
263 data are in line with recent description of dynamics of the SARS-CoV-2 pandemic in densely
264 populated areas,^{14,16} where SARS-CoV-2 started its spread by multiple, independent, and frequently
265 undetected, introductions.

266 Altogether, the spread of two lineages in the two most populated cities of the Lombardy region (Milan
267 and Pavia), along with a broadly comparable level of genetic pairwise distance, and similar viral load
268 levels within them, support their similar fitness, and hence exclude the occurrence of a strong
269 competition between them. Consistent with this, using a Chi-squared test, no trend of association
270 between disease severity (including evidence of interstitial pneumonia, and severe presentation) and
271 lineages was detected (Table 1). Nevertheless, if specific clades or mutational variants might
272 modulate the clinical presentation and spread of the disease should be prudently evaluated, until
273 new and extensive data focused on this setting will be available.

274 At this regard, sequence analysis defined the presence of founding mutations characterizing these
275 two transmission chains. Lineage 2 is defined by sequences characterized by the 3 SNP 28881-
276 28883:GGG>AAC, resulting in two amino acid 203-204:RG>KR changes in the nucleocapsid (N)
277 protein of the SARS-CoV-2. These mutations are localized at the end of serine-arginine (SR)
278 dipeptide of the SR-rich motif (aa 183-195: SSRSSSRSRNSSR [NSTPGSSRG]) characterizing the
279 N protein of SARS-CoV-2, and introduce a lysine between a SR-dipeptide. By previous study on
280 SARS-CoV, which shares the 75% amino acid similarity with SARS-CoV-2, the deletion of the SR-
281 rich motif, or its mutation, significantly reduce viral genomic transcription, the levels of the infectious
282 virions, and the rate of host cell translational inhibitory activity.^{17,18} Whether the 203-204:RG>KR
283 changes would be able to affect SARS-CoV and/or SARS-CoV-2 virus replication is intriguing, yet
284 currently unknown.

285 Cluster A and B, in Lineage 1, are characterized by the presence of two SNPs, the non-syn A26530G
286 (intra-patient prevalence ~98%), leading to the D3G mutation within a B-cell epitope of M protein¹⁹,
287 and the syn A20268G in nsp15 (intra-patient prevalence ~54%).

288 Looking at the overall variability of SARS-CoV-2, we found that only 7 SNPs (2 out of 7 synonymous)
289 characterized our consensus sequences, highlighting a good conservation rate of this virus along
290 time. This conservation rate is confirmed within the spike structural protein, where only 2 mutations,
291 one of them at low prevalence (i.e, C23575T, corresponding to the amino acid variant T671I), were
292 detected. None of these mutations have a role in altering pre-existing N-glycosylation sites or in
293 creating new ones,²⁰ which may be beneficial for the development of vaccines strategies.

294 Worth of mention is the SNP A to G at position 23403, corresponding to the variant D614G in spike
295 protein, detected in the 67.8% of our SARS-CoV-2 sequences. As already known, this variant is
296 observed frequently in European countries, such as the Netherlands, Switzerland, and France, but
297 seldom observed in China. Its rapid fixation at population level might suggest a role in viral entry,
298 and enhancement of interaction between receptor-binding-domain of the S protein with the entry
299 receptor ACE2. This variant, located within a B-epitope, causes the substitution of a large acidic
300 residue (aspartic acid), with a small hydrophobic residue (glycine). Further investigations are
301 required to define if such marked differences in both size and hydrophobicity in the middle of the
302 epitope might compromise the binding affinity to antibodies against wild-type spike protein, elicited
303 by vaccines.²¹

304 Our study may have some limitations. The analysis of phylogenetic structures during such an early
305 phase of the pandemic should be interpreted carefully, as the number of mutations that define
306 phylogenetic lineages is small and may be similar to the rate of potential errors introduced during
307 reverse transcription, PCR amplification, or sequencing.²² To overcome these problems, Bayesian
308 approach, known to be a powerful way to estimate species divergence, and thus expected to provide
309 more robust results, was applied. Moreover, the integration of host characteristics (such as
310 geographical location, collection date and clinical manifestations) aided phylogenetic interpretation.
311 Moreover, the intra-host variability of SARS-CoV-2, and the role of potential existing minority variants
312 has not been investigated here. Initial evidences suggest that intra-host variation of SARS-CoV-2
313 can be frequently found among clinical samples (median number of intra-host variants: 1–4), but at
314 the same time these variants were not observed in the population as polymorphisms, probably
315 suggesting a bottleneck or purifying selection involved.^{23,24} Thus, ad hoc designed studies are
316 necessary to provide an extensive overview of SARS-CoV-2 intra-host variability and minority
317 variants description, if and how these minority variants can spread in the population, and their
318 potential role in virulence and transmissibility.

319 In the peak of epidemic, SARS-CoV-2 diagnosis was mainly addressed to symptomatic cases or
320 subjects at high risk of exposure (i.e. health care workers exposed to positive patients without
321 adequate protection). This approach may have caused a substantial underestimation of positive

322 subjects, preventing to include in our analysis asymptomatic infections, whose role in the
323 transmission chains and in influencing the evolution of epidemic remains a challenge to investigate.
324 Moreover, even if the number of samples here analysed is noteworthy, the epidemic in the East (i.e.
325 Brescia and Mantua) and the valleys of the North (i.e. Valtellina, and Valcamonica, Figure 1) is poorly
326 represented. With exception of Brescia, that represents today one of the most COVID-19 affected
327 area after Bergamo, Cremona and Lodi, Mantua and the valleys account today for the 6.7% of
328 confirmed COVID-19 cases in Lombardy, a percentage that does not represent a major issue for the
329 good reproducibility of SARS-CoV-2 epidemic in the region.

330 In conclusion, this study allows the identification of SARS-CoV-2 lineages circulating in the most
331 affected COVID-19 area at the beginning of 2020, representing a huge reserve of genetic information
332 of a virus that became able to pandemic spread, and caused in Lombardy more than 16,000 deaths
333 in some weeks. We cannot exclude that this multiple and simultaneous circulation of SARS-CoV-2
334 strains can have exacerbated the transmissibility potential of the virus and thus create a real viral
335 storm in such high densely populated region. Only the large-scale surveillance and intervention
336 measures implemented in the early March in Italy were effective in reducing community transmission,
337 ultimately containing the epidemic and limiting the dissemination to other regions.

338 **References**

- 339 1 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-
340 269, doi:10.1038/s41586-020-2008-3 (2020).
- 341 2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin.
342 *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
- 343 3 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
344 **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).
- 345 4 Houtgast, E. J., Sima, V. M., Bertels, K. & Al-Ars, Z. Hardware acceleration of BWA-MEM genomic
346 short read mapping for longer read lengths. *Computational biology and chemistry* **75**, 54-64,
347 doi:10.1016/j.compbiolchem.2018.03.024 (2018).
- 348 5 Li H, A statistical framework for SNP calling, mutation discovery, association mapping and population
349 genetical parameter estimation from sequencing data, *Bioinformatics* (2011) 27(21) 2987-93.
- 350 6 Danecek, P., Schiffels, S. & Durbin, R. Multiallelic calling model in bcftools. Available at
351 <https://samtools.github.io/bcftools/call-m.pdf>, (2016).
- 352 7 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
353 phylogenies. *Bioinformatics* 30, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
- 354 8 Tavaré S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.
355 *Lectures Math Life Sci.* 17, 57–86.

- 356 9 Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over
357 sites: approximate methods. *Journal of molecular evolution* **39**, 306-314, doi:10.1007/BF00160154
358 (1994).
- 359 10 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of
360 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* **2**, vew007,
361 doi:10.1093/ve/vew007 (2016).
- 362 11 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.
363 *Virus evolution* **4**, vey016, doi:10.1093/ve/vey016 (2018).
- 364 12 Ferreira, M.A., and Suchard, M.A. (2008). Bayesian analysis of elapsed times in continuous-time
365 Markov chains. *Can. J. Stat.* **36**, 355–368.
- 366 13 Böhmer MM, Buchholz U, Corman VM, et al. Investigation of a COVID-19 outbreak in Germany
367 resulting from a single travel-associated primary case: a case series [published online ahead of print,
368 2020 May 15]. *Lancet Infect Dis.* 2020;S1473-3099(20)30314-5. doi:10.1016/S1473-3099(20)30314-
369 5
- 370 14 Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell*,
371 doi:10.1016/j.cell.2020.04.023 (2020).
- 372 15 Rambaut, A. (2020). Phylodynamic Analysis, 176 genomes. [http://virological.org/t/phylodynamic-
373 analysis-176-genomes-6-mar-2020/356](http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356).
- 374 16 Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, et al. Introductions and early spread of SARS-CoV-2
375 in the New York City area [published online ahead of print, 2020 May 29]. *Science.* 2020;eabc1917.
376 doi:10.1126/science.abc1917
- 377 17 Kumar, S., Maurya, V. K., Prasad, A. K., Bhatt, M. L. B. & Saxena, S. K. Structural, glycosylation and
378 antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV).
379 *Virusdisease* **31**, 13-21, doi:10.1007/s13337-020-00571-5 (2020).
- 380 18 Tylor S, Andonov A, Cutts T, et al. The SR-rich motif in SARS-CoV nucleocapsid protein is important
381 for virus replication. *Can J Microbiol.* 2009;55(3):254-260. doi:10.1139/w08-139
- 382 19 Peng TY, Lee KR, Tarn WY. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe
383 acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization,
384 translation inhibitory activity and cellular localization. *FEBS J.* 2008;275(16):4152-4163.
385 doi:10.1111/j.1742-4658.2008.06564.x
- 386 20 Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-
387 CoV-2 spike [published online ahead of print, 2020 May 4]. *Science.* 2020;eabb9983.
- 388 21 Koyama T, Weeraratne D, Snowdon JL, Parida L. Emergence of Drift Variants That May Affect COVID-
389 19 Vaccine Development and Antibody Treatment. *Pathogens.* 2020;9(5):E324. Published 2020 Apr
390 26.
- 391 22 Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires
392 caution [published online ahead of print, 2020 May 19]. *Nat Microbiol.* 2020;10.1038/s41564-020-
393 0738-5.
- 394 23 Shen Z, Xiao Y, Kang L, et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients
395 [published online ahead of print, 2020 Mar 4]. *Clin Infect Dis.* 2020;ciaa203. doi:10.1093/cid/ciaa203
- 396 24 Rose R, Nolan DJ, Moot S, Feehan A, Cross S, Garcia-Diaz J, Lamers SL. Intra-host site-specific
397 polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. *MedRxiv*
398 2020

399

400 **Conflict of Interest:** The authors have no financial and non-financial competing interests that might
401 be perceived to influence the results and/or discussion reported in this paper.

402 **Funding:** This work was financially supported by an unrestricted grant from Cariplo foundation.

403 **Acknowledgments:** We thank Biodiversa s.r.l. for providing technical support, particularly Dr. Marco
404 Dotto, for his valuable assistance in the implementation of the study. The authors also thank Dr.
405 Silvia Nerini and all the staff of the Microbiology and Virology Laboratory of ASST Grande Ospedale
406 Metropolitano Niguarda and IRCCS San Matteo for outstanding technical support in processing swab
407 samples, performing laboratory analyses and data management.

408 **Conflict of Interest:** The authors have no financial and non-financial competing interests that might
409 be perceived to influence the results and/or discussion reported in this paper.

410 **Author contributions:** CA, VCe, AP equally contributed in study design, data collection, data
411 analysis, data interpretation, writing; VCo performed bioinformatic analysis and data processing; MT,
412 LC, SC, SR, FG, FN processed samples and collected data; SG helped in bioinformatic analysis;
413 EM and MA helped in sample processing and data collection; CV, RF, OME, MP recruited samples
414 and enrolled patients; CFP and FB conceived and directed the study, and critically revised the
415 manuscript.

416 **Materials & Correspondence:** Corresponding author is Carlo Federico Perno, MD, PhD, University
417 of Milan, email: cf.perno@uniroma2.it. The data that support the findings of this study are available
418 from the corresponding author upon reasonable request.

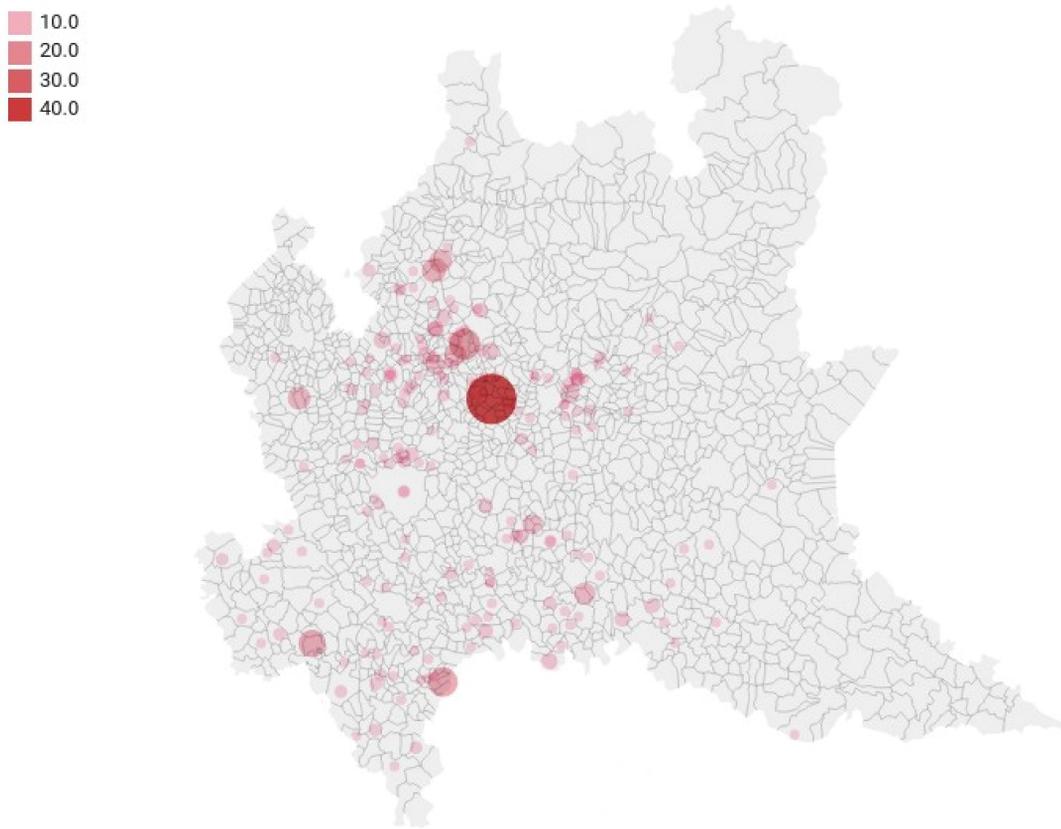
Table 1. Demographic, and clinical findings of the 346 SARS-CoV-2 infected patients according with lineages observed by ML tree.

	Overall, N=346	Lineages ^a		P-value ^b
		1, N=211	2, N=131	
Demographics and clinical characteristics				
Age, years	72 (53-83)	70 (53-83)	73 (55-83)	0.737
Sex, Male	196 (57.8)	109 (51.6)	85 (64.9)	0.045
Residency				
<i>Milan</i>	75 (21.7)	49 (23.2)	25 (19.1)	0.442
<i>Como</i>	68 (19.7)	28 (13.3)	40 (30.5)	<0.001
<i>Pavia</i>	61 (17.6)	42 (19.9)	19 (14.5)	0.246
<i>Bergamo</i>	36 (10.4)	14 (6.6)	22 (16.8)	0.005
<i>Lecco</i>	30 (8.7)	10 (4.7)	20 (15.3)	0.002
<i>Lodi</i>	32 (9.2)	30 (14.2)	1 (0.8)	<0.001
<i>Cremona</i>	26 (7.5)	23 (10.9)	2 (1.5)	0.002
<i>Other^c</i>	18 (5.2)	15 (7.1)	2 (1.5)	0.040
Chronic comorbidities ^d	163 (52.8)	96 (50.8)	65 (55.6)	0.488
<i>Hypertension</i>	33 (10.7)	22 (11.6)	10 (8.5)	0.505
<i>Obesity</i>	21 (6.8)	11 (5.8)	10 (8.5)	0.363
<i>Diabetes</i>	37 (12.1)	16 (8.5)	21 (17.9)	0.022
<i>Cardiovascular disease</i>	96 (31.1)	53 (28.0)	41 (35.0)	0.199
<i>Chronic obstructive lung disease</i>	43 (13.9)	22 (11.6)	21 (17.9)	0.169
<i>Malignancies</i>	39 (12.6)	24 (12.7)	14 (12.0)	0.992
<i>Chronic kidney disease</i>	24 (7.8)	12 (6.3)	11 (9.4)	0.447
<i>Chronic liver disease</i>	5 (1.6)	3 (1.6)	2 (1.7)	0.935
<i>Other^e</i>	28 (9.1)	18 (9.5)	10 (8.5)	0.933
Symptoms at admission ^f				
<i>Fever</i>	107 (59.4)	70 (56.0)	35 (67.3)	0.220
<i>Cough</i>	64 (27.0)	42 (32.6)	21 (40.0)	0.268
<i>Dyspnea</i>	48 (26.7)	29 (23.2)	18 (34.6)	0.168
Time from symptoms-onset to SARS-CoV-2 diagnosis, weeks	0.29 (0.0-0.57)	0.28 (0.14-0.57)	0.43 (0.07-0.71)	0.157
Collection date (month, day)	03-14 (03-06; 03-20)	03-11 (03-05; 03-20)	03-16 (03-09; 03-21)	0.023
Disease severity^g				
<i>Mild</i>	168 (70.9)	112 (74.3)	54 (65.1)	0.187
<i>Moderate</i>	38 (13.9)	25 (11.8)	13 (15.7)	0.709
<i>Severe</i>	31 (11.3)	14 (6.6)	16 (19.3)	0.081
Evidence of Interstitial Pneumonia ^h	56 (24.1)	30 (20.3)	25 (30.9)	0.103
SARS-CoV-2 rtPCR				
Mean cycle thresholds ⁱ	18.7 (16.7-20.0)	19.0 (17.2-20.2)	18.1 (16.3-19.9)	0.060
SNP in SARS-CoV-2 genome				
20268, A to G, syn (nsp15)	12 (3.5)	12 (5.7)	0 (0.0)	0.013
23575, C to T, non-syn T to I (S)	23 (6.6)	23 (10.9)	0 (0.0)	0.001
26530, A to G, non-syn D to G (M)	20 (5.8)	20 (9.5)	0 (0.0)	<0.001
28881-28883, GGG to AAC, non-syn RG to KR (N)	130 (3.8)	2 (0.9)	128 (97.7)	<0.001

Data are expressed as median (IQR), or N (%). ^aA total of 343 SARS-CoV-2 isolates are involved in lineages. ^bP-values were calculated by Mann-Whitney test, or Chi² test, as appropriate. ^cOther includes Brescia, Mantua, Monza and Brianza, Sondrio and Varese. ^dData available for 309 patients. ^eIncluding: Crohn's disease (n=1), Hashimoto's thyroiditis (n=3), familial lipid disorders (n=9), rheumatoid

arthritis (n=3), Amyotrophic Lateral Sclerosis (n=1), cognitive disorders (n=11). ^dData available for 180 patients. ^eData available for 237 patients. ^hDiagnosed by X Ray or CT Scan. Data available for 232 patients. ⁱReal-time reverse transcription PCR Ct (cycle threshold) values of these samples ranged from 9 to 35 (GeneFinder™ COVID-19 Plus RealAmp Kit, ELITech; Allplex™ 2019-nCoV Assay, Seegene; Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25(3):2000045. doi:10.2807/1560-7917.ES.2020.25.3.2000045; <https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf>).

Figure 1. Geographic distribution of COVID-19 cases among the 12 provinces of Lombardy.



Milan: 75; Como: 68; Pavia: 61; Bergamo: 36; Lecco:30; Lodi:32; Cremona: 26; Other (including Brescia, Mantua, Monza and Brianza, Sondrio and Varese): 18

Figure 2. Prevalence of most representative single nucleotide polymorphisms (SNP) 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. (B) Intra-patient prevalence of most representative SNPs in North Italian SARS-CoV-2 sequences according to sample date. Italian sequences in panel A include the 6 and 8 sequences from North and Central Italy present in GISAID at May 05.

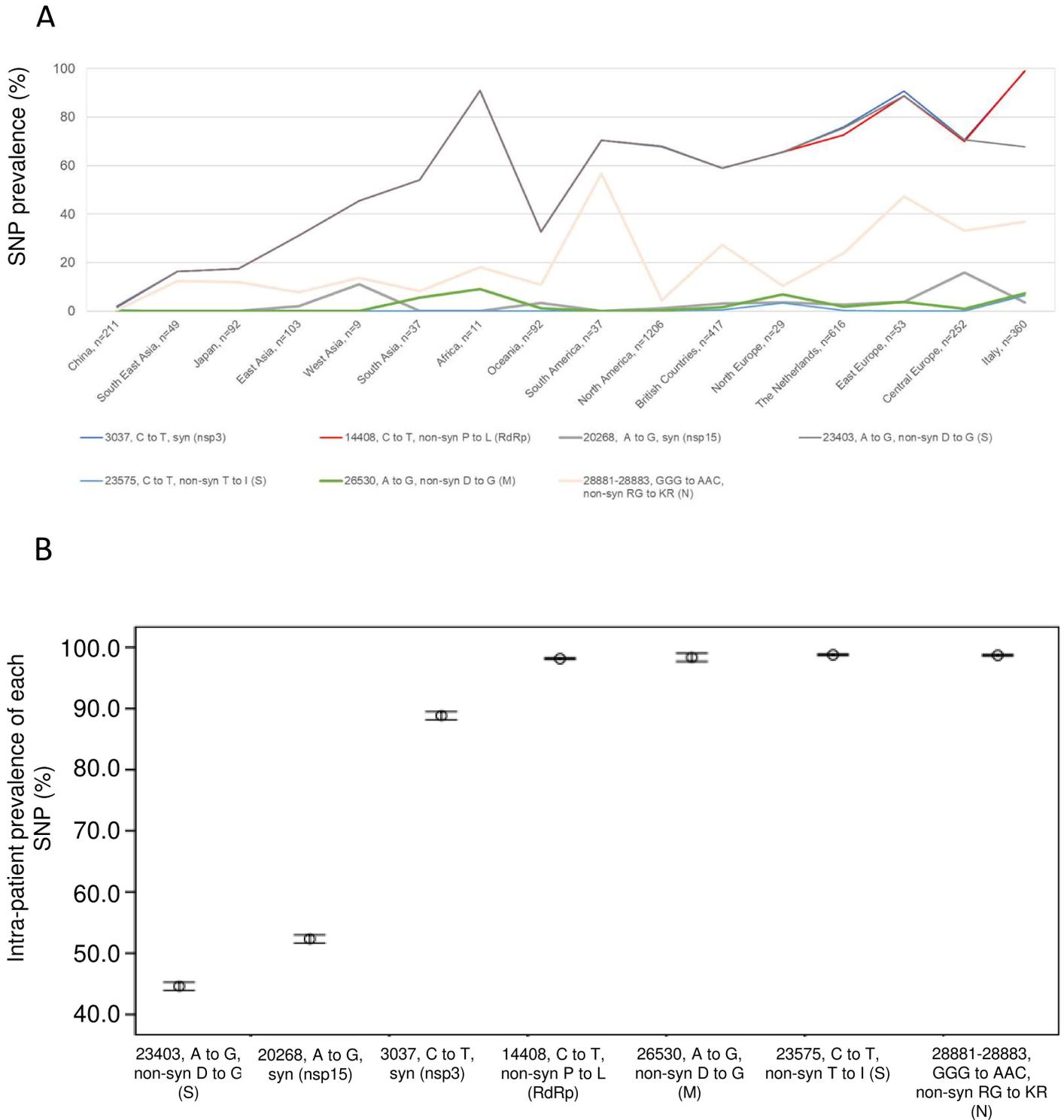
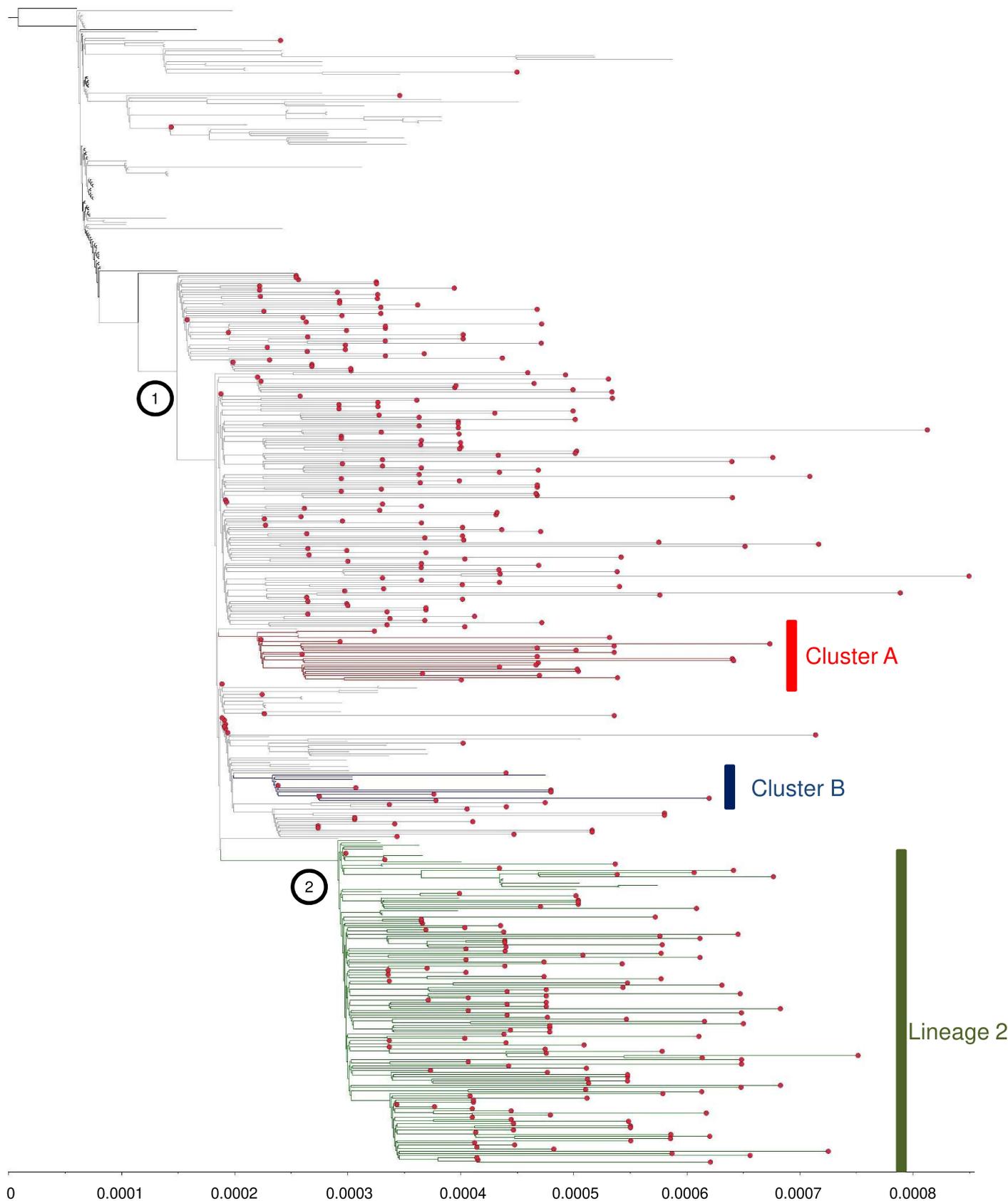
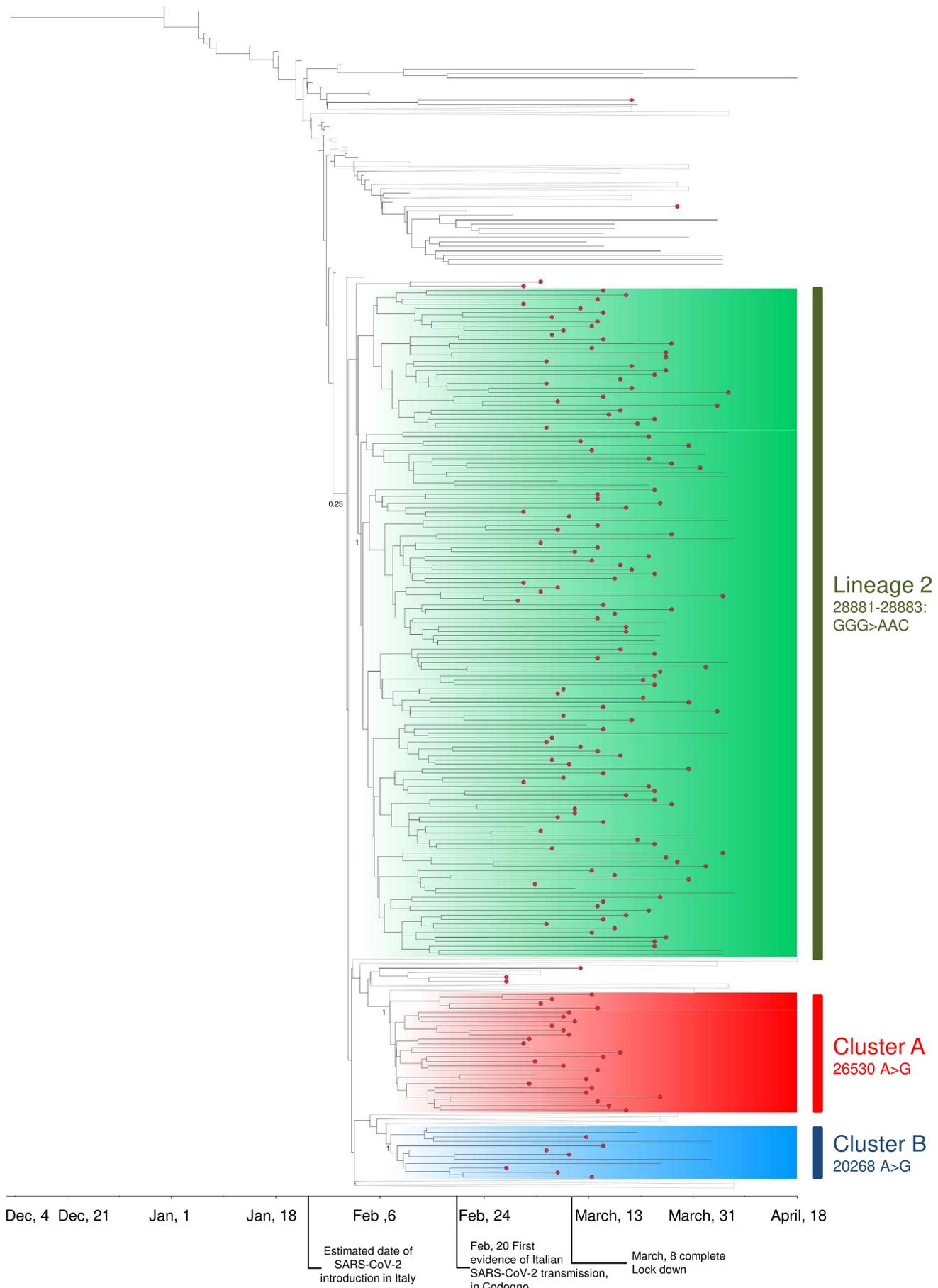


Figure 3. Estimated maximum likelihood phylogeny of SARS-CoV-2 sequences from Lombardy (gray taxa with red dots) and genomes from China (black taxa without dots) and other countries (gray taxa without dots). The two lineages (1 and 2) were reported by black dots at corresponding nodes. SARS-CoV-2 clades supported by a posterior probability of 1.00 in the maximum clade credibility tree were highlighted in green (Lineage 2), in red (cluster A), in blue (Cluster B).



Gisaid sequences come from Central Italy (n=6), East Europe (n=6), North Europe (n=6), South America (n=6), Africa (n=7) Japan (n=7), Oceania (n=7), West Asia (n=7), North Italy (n=9), South Asia (n=9), Central Europe (n=11), East Asia (n=11), European Low Countries (n=12), South East Asia (n=13), North America (n=18), British Countries (n=26), China (n=36).

Figure 4. Time-scaled maximum clade credibility tree with SARS-CoV-2 sequences from Lombardy (red circle) and other countries (no circle). SARS-CoV-2 sequences with a posterior probability of 1.00 were highlighted in green (lineage 2), in red (cluster A), in blue (cluster B). All nodes with posterior probabilities <0.8 have been collapsed.



Gisaid sequences come from Central Italy (n=6), East Europe (n=6), North Europe (n=6), South America (n=6), Africa (n=7) Japan (n=7), Oceania (n=7), West Asia (n=7), North Italy (n=9), South Asia (n=9), Central Europe (n=11), East Asia (n=11), The Netherlands (n=12), South East Asia (n=13), North America (n=18), British Countries (n=26), China (n=36).