Rapid longitudinal SARS-CoV-2 intra-host emergence of novel haplotypes regardless of immune deficiencies

Laura Manuto^{1,§}, Marco Grazioli^{1,§}, Andrea Spitaleri^{2,§}, Paolo Fontana³, Luca Bianco³, Luigi Bertolotti⁴, Martina Bado¹, Giorgia Mazzotti¹, Federico Bianca¹, Francesco Onelia¹, Giovanni Lorenzin⁵, Fabio Simeoni², Dejan Lazarevic², Elisa Franchin^{1,6}, Claudia Del Vecchio^{1,6}, Ilaria Dorigatti⁷, Giovanni Tonon^{2,8}, Daniela Cirillo^{5,*}, Enrico Lavezzo^{1,*}, Andrea Crisanti^{1,6,9,**}, Stefano Toppo^{1,10,**}

¹ Department of Molecular Medicine, University of Padova, Padua, Italy

² Center for Omics Sciences, IRCCS San Raffaele Institute, Milano, Italy

³ Research and Innovation Center, Edmund Mach Foundation, S. Michele all'Adige, Italy

⁴ Department of Veterinary Sciences, University of Torino, Grugliasco, Turin, Italy

⁵ Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy

⁶ Microbiology and Virology Diagnostic Unit Padua University Hospital, Via Giustiniani, Padova, Italy

⁷ MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK

⁸ Functional Genomics of Cancer Unit, Division of Experimental Oncology, IRCCS San Raffaele Scientific Institute, Milano, Italy

⁹ Department of Life Sciences, Imperial College London, London, UK

¹⁰ CRIBI Biotech Center, University of Padova, Padua, Italy

§ These authors contributed equally

* These authors jointly supervised this work

**Corresponding author: stefano.toppo@unipd.it; a.drcrisanti@imperial.ac.uk

ABSTRACT

On February 2020, the municipality of Vo', a small town near Padua (Italy), was quarantined due to the first coronavirus disease 19 (COVID-19)-related death detected in Italy. The entire population was swab tested in two sequential surveys. Here we report the analysis of the viral genomes, which revealed that the unique ancestor haplotype introduced in Vo' belongs to lineage B and, more specifically, to the subtype found at the end of January 2020 in two Chinese tourists visiting Rome and other Italian cities, carrying mutations *G11083T* and *G26144T*. The sequences, obtained for 87 samples, allowed us to investigate viral evolution while being transmitted within and across households and the effectiveness of the non-pharmaceutical interventions implemented in Vo'. We report, for the first time, evidence that novel viral haplotypes can naturally arise intra-host within an interval as short as two weeks, in approximately 30% of the infected individuals, regardless of symptoms severity or immune system deficiencies. Moreover, both phylogenetic and minimum spanning network analyses converge on the hypothesis that the viral sequences evolved from a unique common ancestor haplotype, carried by an index case. The lockdown extinguished both viral spread and the emergence of new variants, confirming the efficiency of this containment strategy. The information gathered from household was used to reconstructs possible transmission events.

AUTHOR SUMMARY

It is of great interest and importance to understand SARS-CoV-2 ability to mutate generating new viral strains, and to assess the impact of containment strategies on viral transmission. In this study we highlight the rapid intra-host haplotype evolution regardless of symptom severity and immune deficiencies that we observed during the first wave of the pandemic in the municipality of Vo' in Italy. The confirmation that all the haplotypes found in this small community derive from a common ancestor haplotype, has allowed us to track the rapid emergence of new variants but lockdown and mass testing efficiently prevented their spread elsewhere.

52 INTRODUCTION

53 Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection continues to spread world-
54 wide, with over 5 million of deaths out of more than 250 million of positive cases reported since the
55 beginning of the pandemic (1). There is extensive evidence that SARS-CoV-2 genome has evolved
56 and has acquired several mutations conferring higher viral fitness, with almost all genomic sites being
57 affected by mutation events (2–4).

58 Here we provide novel insights into the generation of diversity of SARS-CoV-2 genomes in a close
59 community.

60 On February and March 2020, preceding and following a two-weeks lockdown respectively, two
61 mass swab testing campaigns were conducted in Vo' to trace and isolate all the positive subjects. In
62 previous work, we investigated the effectiveness of the implemented non-pharmaceutical
63 interventions (NPIs), the role of asymptomatic infections (5) and the tracking of T-cell signatures of
64 the whole population (6). In May and November 2020, two follow-ups were carried out to assess the
65 antibody dynamics following infections (7).

66 Here we report on the results from the sequencing of SARS-CoV-2 genomes extracted from oro-
67 nasopharyngeal swabs of positive subjects. We obtained the sequences for 87 samples, representing
68 the vast majority of the positive subjects detected in Vo'. Sequencing data has allowed us to identify
69 unequivocally the ancestor haplotype of the virus circulating in Vo' at the time of the outbreak. This
70 evidence provided the opportunity to infer the evolution of SARS-CoV-2 in an isolated community,
71 from an intra-host, a household and a general perspective, tracing the generation of novel mutations
72 from the index haplotype. Information on date of symptoms onset allowed us to reconstruct a
73 temporal expanding network of infected subjects. The uniqueness of this study lies in the feature of
74 this isolated and small community that, since the very beginning of the SARS-CoV-2 pandemic, was
75 monitored over time, enabling us to investigate the virus evolution and transmission dynamics.

76

RESULTS

Viral haplotypes circulating in Vo'

During the first surveys, we collected oropharyngeal swab samples of nearly the entire population of Vo' at two consecutive time points, (21-29 February 2020 and 7 March 2020, respectively). Out of the tested subjects, 100 individuals had a positive result, with 12 of them testing positive again at the second time point. We sequenced the viral genome of all the available samples. As a result, 87 consensus sequences were obtained and uploaded in GISAID (8).

All the viral sequences were characterised by two mutations, *G11083T* and *G26144T* (from here on reported as Ancestor Haplotype, AH). Interestingly, these mutations were previously identified in a couple of Chinese tourists who, after visiting Verona, Parma and Florence, were diagnosed with SARS-CoV-2 infection in Rome on 31st January 2020, and correspond to the first reported case of COVID-19 in Italy (9). Surprisingly, this haplotype, which is classified as lineage B according to Rambaut's nomenclature (10), accounted for 43.8% of total sequences in the February-March 2020 in the Veneto region. In the surrounding regions, such as Lombardia and Trentino Alto Adige (11), we found evidence of a different source of introduction, given that the B.1 lineage was dominant (12) (Fig 1).

While the AH was found only in Vo' and in the province of Belluno, some AH's derivative subtypes were observed in other provinces of Veneto region.

The sequencing data indicate that the B lineage haplotypes introduced in the Veneto region, as the ones identified in Vo', derive from the AH, carrying G11083T and G26144T mutations, that further developed new additional mutations.

Fig 1. Lineages circulating in Northern Italian regions

101 *Percentages of SARS-CoV-2 lineages circulating in Emilia Romagna (n=2), Friuli Venezia Giulia*
 102 *(n=13), Lombardia (n=400), Trentino Alto Adige (n=58) and Veneto (n=226) regions in February*
 103 *and March 2020 according to GISAID database. The size of the pie charts reflects the number of*
 104 *sequences available per region. The different lineages are colored according to the legend. Black*
 105 *circles represent haplotypes identical to those found in Vo', diamonds indicate haplotypes with one*
 106 *additional mutation (edit distance +1 from Vo') and triangles correspond to haplotypes with two*
 107 *additional mutations (edit distance +2 from Vo').*

108

109 According to the sequenced data, the AH was the most prevalent in Vo' (38 out of 87, 43.7%), while
 110 additional 26 unique haplotypes derive from the ancestor one. A total of 42 unique point mutations
 111 defining the different haplotypes, (including *G11083T* and *G26144T*) mapped along different regions
 112 of the viral sequence, 1 was located in 5'-UTR (2.4%), 15 were synonymous (35.7%) and 26 were
 113 non-synonymous (61.9%). Moreover, a 9-nucleotide deletion was observed in two samples, resulting
 114 in 3 amino acid deletion affecting the non-structural protein 1 (nsp1) (*ORF1a:K141-*,*ORF1a:S142-*
 115 *,ORF1a:F143-*), a mutation already observed in other sequenced SARS-CoV-2 genomes belonging
 116 to different lineages (13) and different countries, suggesting homoplasy events.

117 Although the short time transmission window hampered the detection of a natural or purifying
 118 selection, most of the non-synonymous mutations found in Vo' emerged independently later on
 119 during the pandemic in different countries (S1 Table). The two mutations defining the AH behaved
 120 differently during the pandemic. *G26144T* was mostly associated with *G11083T* in the lineage B and
 121 disappeared in April 2021 after reaching a peak at the beginning of the pandemic (February-April
 122 2020). Conversely, the *G11083T* mutation appears in recently collected sequences belonging to
 123 different lineages, suggesting to be another example of homoplasy (14) (S1 Fig).

124 Interestingly, the mutation *Spike:L5F (C21575T)*, characterising the viral haplotype of two subjects
 125 in Vo', emerged from January 2021 to August 2021 as part of the mutation pattern defining the
 126 B.1.526 lineage (Iota variant), prevalent in USA (15) (S2 Fig).

127

128 *Vo' haplotypes in Europe*

129 We compared the viral genomes collected at Vo' with a selection of contemporary European
130 sequences from different locations. At the beginning of the pandemic the AH was circulating also in
131 other European countries, with UK, Italy (90% in Veneto region) and Spain reporting the highest
132 occurrences. To analyse the data, we used two different approaches based on phylogenetic analysis
133 (Maximum Likelihood phylogenetic tree, ML) and on the construction of a Minimum Spanning
134 Network, MSN (see Methods).

135 In Fig 2, we provide a ML phylogenetic tree and a MSN limited to Vo' haplotypes and to closely
136 related European haplotypes, where all the European sequences carrying the AH, including the Italian
137 ones, are represented as a big green node in the MSN, while they are collapsed in the phylogenetic
138 tree. The same analysis, extended to all the European sequences collected at the time period
139 considered, are provided in S3 and S4 Figs for the phylogenetic tree and the MSN, respectively.
140 According to phylogenetic and MSN reconstruction, only one sequence, “VO_SR_65” (*haplotype*
141 *G11083T, G26144T, C22088T*), shares a private mutation with a Polish sequence, *EPI_ISL_455452*
142 (*haplotype C4338T, T9743C, G11083T, G26144T, C22088T*). However, the Polish sequence (dated
143 2020-03-28) relates to a sample collected one month later than the Vo' sample “VO_SR_65” (dated
144 2020-2-21), suggesting an independent evolution and possibly a homoplasmy event.

145 These results confirmed that the Vo' haplotypes were not observed in other countries and did not
146 generate descendants, as a consequence of mass testing and lockdown strategy implemented in Vo'
147 (16–18).

148

149

150 **Fig 2. Maximum Likelihood phylogenetic tree and Minimum Spanning Network of sequences**
151 **related to Vo' viral genomes.**

152 *European sequences closely related to Vo' viral genomes and with collection date corresponding to*
 153 *the beginning of the pandemic were retrieved from GISAID and utilised for the ML phylogenetic tree*
 154 *and MSN.*

155 *a) Maximum Likelihood phylogenetic tree. The phylogenetic tree is collapsed in correspondence*
 156 *of the AH, found both in Vo' and other European countries. Sequences from Vo' are reported*
 157 *in red.*

158 *b) Minimum Spanning Network. The AH is represented as a big green node, containing both Vo'*
 159 *and European sequences. The size of the nodes reflects the abundance of identical sequences.*
 160 *The distance of the edges from the central node reflects the number of accumulated mutations.*
 161 *European/Italian sequences are reported in yellow, Vo' sequences are reported in red, and*
 162 *Wuhan reference node is represented as a blue square. The only extra-Vo' haplotype*
 163 *genetically related to a Vo' haplotype is indicated by the red arrow.*

164

165 *Intra-host viral evolution*

166 We obtained the viral sequence of 12 individuals positive at two sequential swab tests. As reported
 167 in Fig 3, in 4 out of 12 subjects (33.3%) the virus acquired at least a novel mutation in an average
 168 time of 11 days (range 9-16 days). While 3 out of 4 individuals acquired a single mutation at the
 169 second time point, one subject (samples *VO_SR_7* and *VO_SR_93*) accumulated 4 mutations in 13
 170 days, highlighting the rapidity of viral evolution in some subjects. No significant differences in days
 171 after the first swab test, disease severity, antibody production or T cell receptor breadth or depth (6)
 172 were found comparing individuals with a constant viral haplotype to individuals with a mutated viral
 173 haplotype (Figs S5 and S6). Giving that all the haplotypes identified in Vo' derive from the AH, and
 174 that within approximately two weeks of positivity, the viral genomes accumulated at least a novel
 175 mutation in 33.3% of cases, we further investigated this event among all the available subjects.
 176 Considering that in Vo' we identified 26 unique haplotypes (AH excluded) out of 87 subjects and the

177 two-weeks lockdown imposed by the local authorities, we can assume that intra-host evolutionary
 178 events occurred in 26 subjects out of 87 (29.9%), within this time window. Of these 26 cases, 4 were
 179 proven by the sequential swab tests, as described above, while other 10 cases could be inferred
 180 according to viral haplotype evolution within family households, virus haplotypes and the symptoms
 181 onset dates (Table 1).
 182

	Subjects	Notes
Households	SR_38, SR_122, SR_64, SR_12, SR_77	Subjects infected by family members carrying a different haplotype
Haplotypes	SR_56, SR_99	Subjects carrying a haplotype that evolved from a subtype of the AH
Symptoms onset dates	SR_65, SR_61, SR_30	First subjects contracting the infection, likely being infected with the AH

183 **Table 1. Inferred intra-host viral evolution**

184

185

186 **Fig 3. Within host variation.**

187 *Longitudinal analysis of SARS-CoV-2 intra-host variation in 12 subjects in an average of 11-day time*
 188 *window. The viral haplotypes of the two consecutive swab tests of each subject are reported on the*
 189 *same line. Each haplotype is depicted as a circle with each slice representing a mutation,*
 190 *characterised by a colour according to the legend. Minor variants are represented as triangles.*

191

192 *Within-household variability*

193 The different haplotypes identified in Vo' were analysed in the context of the household structure to
 194 investigate potential transmission chains and within household viral evolution.

195 The intra-family transmission chains reported in Fig 4 were based on: a) swab test results and relative
 196 dates, b) viral consensus genome, c) minor variants, considered as alternative nucleotides with a
 197 frequency from 5% to 49%. We assumed that the subjects who negativized earlier were the infectors
 198 and assumed no reversion.

199 Interestingly, we found evidence that a few mutations were household specific (not present in any
 200 other sample collected in Vo' outside the housed). Households 838 and 219, where the 2 out of 2 and
 201 3 out of 3 family members shared the same haplotype (*G11083T*, *G26144T*, *T2248C*, *C21575T*, and
 202 *G11083T*, *G26144T*, *G1944A*, *T9731C*, *G15957T*, *G20378T* respectively), exemplify this
 203 phenomenon.

204 Household 456, where one household member is characterised by the AH (haplotype *G11083T*,
 205 *G26144T*), the second one acquired a novel private mutation (haplotype *G11083T*, *G26144T*,
 206 *C27972A*) and the third one further acquired another private mutation (haplotype *G11083T*,
 207 *G26144T*, *C27972A*, *C26936T*), likely reflects the direction of transmission and the sequential viral
 208 evolution during transmission among household members.

209 By enriching genetic data with metadata, such as swab tests and contacts information, it was possible,
 210 for 8 out of 13 (61.5%) households, to reconstruct the household transmission chains, depicted with
 211 arrows in Fig 4 panels a and b. Notably, while 7 out of 8 of the reconstructed transmissions occurred
 212 within the same household Fig 4a, in one case we found evidence of household members contracting
 213 the infection outside of the household. This is confirmed by the difference in viral haplotypes between
 214 the two household members, with one of them sharing the same haplotype with a declared direct
 215 contact Fig 4c. In the remaining 5 households, although the transmission probably occurred within
 216 family members, we did not have enough information to reconstruct the transmission chains (Fig 4b).

217

218

219

220 **Fig 4. Within family diversity and transmission**

221 *Sequences were grouped according to the households and analysed separately. Each haplotype is*
 222 *depicted as a circle with each slice representing a mutation, characterised by a colour according to*
 223 *the legend. For each household (identification number provided at the top of each diagram) genetic*
 224 *information of family members (labelled with an identification number) and relative metadata (date*
 225 *of positive swab tests reported on x axis) were utilised to investigate the diversity and the transmission*
 226 *chain. Households were then subdivided into within household transmission (panel a), uncertain*
 227 *transmission (panel b), and extra-household transmission (panel c).*

228

229

230

231 *Transmission chain reconstruction*

232 In Fig 5, contact (panel on the left) and genetic (panel on the right) data are reported and compared.
 233 Given the incompleteness of the contact information and of the reported dates of symptom onset, it
 234 was not possible to identify unequivocally the chains of transmission events in Vo', not even with the
 235 support of genetics. However, the data suggest that the outbreak initiated from an indoor space that
 236 fuelled a superspreading event. The time lapse animation (see S7 movie) shows the reported contacts
 237 according to the reported times of symptoms onset.

238

239

240 **Fig 5. Network of contacts based on post-infection interviews**

241 *a) samples are connected according to the contacts declared during the post-infection interviews.*
 242 *Each contacts chain is coloured according to the legend, with nodes not declaring any contact being*
 243 *coloured in grey. Subjects who declared informative contacts, but without an available viral*
 244 *sequence, are depicted as white triangles.*

245 *b) Vo' haplotypes are clustered according to their mutations and to the genetic distance from the AH*
 246 *(edges at distance 1 are not drawn for graphical reasons)*

247 **DISCUSSION**

248 The population of Vo' was the first in Italy and Europe to be placed under lockdown by the regional
 249 authorities after the first ascertained COVID-19 death in the country. This situation has allowed us to
 250 study in details different aspects of the epidemic in a controlled epidemiological setting.

251 The sequencing data identified the Vo' Ancestor Haplotype (AH), characterised by *G11083T* and
 252 *G26144T* mutations, from which all the other sequenced haplotypes evolved. This haplotype, initially
 253 observed in two Chinese tourists visiting Italy in January 2020, was dominant in Veneto region at the

beginning of the pandemic, while lineage B.1 was prevalent in the surrounding regions, revealing multiple viral introductions in Italy (19).

Moreover, MSN and phylogenetic analysis including European sequences of the time confirmed that, although the AH was circulating also in other countries, none of the descendant haplotypes evolved in Vo' was observed elsewhere. Thus, the lockdown was efficient in containing the outbreak in Vo' and the new virus variants emerged from it. Interestingly, we observed three cases of putative homoplasmy, with: a) *C22088T*, present in a Polish sequence collected one month later than Vo' outbreak, b) *G11083T*, found in several sequences from different countries and lineages, c) *C21575T*, shared by two Vo's household members and appearing at a global level only in 2021, as a mutation defining the B.1.526 lineage.

According to present literature, the accumulation of viral mutations at the consensus level was previously observed in treated (20,21) or immunocompromised subjects (22,23) experiencing prolonged infections whilst other studies reported only longitudinal changes of intra-host minor variants (24,25). We here provide the first evidence of the rapid intra-host accumulation of mutations at the consensus level in a time window as short as two weeks regardless of symptoms severity or immunodeficiencies. The generation of new haplotypes occurred in average time of 11 days in 33% of samples (4 out of 12) in otherwise healthy individuals.

In an exemplary case, a total of four mutations accumulated in 13 days, pointing out the rapidity of viral evolution. Based on the collected data, we could infer a frequency of intra-host evolution of 30% (26 out of 87). Although no correlation with symptoms severity, length of the infection and serological data was found, other mechanisms could be responsible for within-host viral evolution, such as host immune response (26,27).

Given the frequency of intra-host viral evolution, it was difficult to define a transmission chain based on genetics, since the differences in the haplotypes can derive either from the haplotype carried by the infector or by an intra-host evolution occurred in the infectee. The integration of sequencing data

with metadata allowed us to infer the transmission chains among some household members indicating within household transmissions as the most frequent ones (7 out of 13). Though contact tracing analysis was limited to symptoms onset data to reconstruct a temporal expanding network of infected subjects, it enabled us to identify the first infections and a gathering place which had a main role in the transmission of the infection. To sum up, this study sheds light around the effectiveness of contact tracing based on interviews, which showed limitations in the coverage and traceability of the contacts among the identified infections. For these reasons, in the absence of effective digital contact tracing, mass testing followed by case isolation represents the best approach, when applicable, for identifying and isolating infections thus giving the best chances to achieve epidemic control (28–30). As a matter of fact, the variants generated in the Vo' population, did not spread outside Vo' showing that effective control strategies not only curb transmission, but also control the emergence and spread of new variants. The number of mutations observed in the Vo' population over just two weeks should warn about the velocity of adaptation that may be occurring in different subjects and the potential ability of SARS-CoV-2 to develop immune evasion.

MATERIALS AND METHODS

Sample collection, library construction and sequencing methods

The presence of SARS-CoV-2 genomes was evaluated with an in-house real-time RT–PCR method targeting the envelope gene (E) (31). Total nucleic acids were purified from nasopharyngeal swab samples using a MagNA Pure 96 System (Roche Applied Sciences). All samples were treated with DNase I (Promega, Madison, WI, USA) to eliminate residues, of the host, and bacterial gDNA. Quantity and quality checks were performed on the HSRNA chip for TapeStation 4100 (Agilent, Santa

304 Clara, CA, USA). Samples with low RNA concentration were concentrated using Vacufuge plus
305 (Eppendorf, Hamburg, Germany). Starting with 50 ng of total RNA, NEBNext Ultra II First Strand
306 and Non-Directional RNA Second Strand Synthesis Module (NEB, Ipswich, Massachusetts) were
307 used to convert RNA into cDNA. Before proceeding with the next steps, we performed RTqPCR on
308 all samples in order to determine the relative abundance of viral RNA. RTqPCR was performed by
309 amplification regions for the following target genes: RNA-dependent RNA-polymerase (RdRP),
310 nucleocapsid (N), envelope (E), and Internal Control.

311 The enrichment step was performed according to the manufacturer instructions. We used the Illumina
312 Nextera Flex kit for Enrichment/Respiratory Virus Oligos Panel Detection of SARS-CoV-2 RNA.
313 Briefly, samples were fragmented and barcoded with Nextera Flex, followed by PCR amplification
314 (17 cycles). Before hybridizations with Respiratory Virus Oligos Panel, samples were pooled
315 according to the viral load (previously determined by RTqPCR). Each pool was composed by 8
316 samples with similar Ct (threshold cycle), to avoid bias competition between samples with low and
317 high Ct during hybridization and sequencing.

318 After a quality check on HSDNA chip for TapeStation 4100(Agilent), libraries were diluted to 2.2pM,
319 mixed with Pfix (20%) and then loaded on NextSeq 500 sequencer (Illumina, San Diego, CA).
320 Sequencing was performed in PE (pair-end) mode, generating 149nt length reads, and 1-5M clusters
321 for each sample.

322

323 *Viral genome assembly: quality check and mapping of the reads*

324 Raw sequences were filtered for length and quality with Trimmomatic v0.40 according to the
325 following parameters: ILLUMINACLIP:TruSeq3-PE-2:2:30:10 LEADING:30 TRAILING:30
326 SLIDINGWINDOW:4:20 MINLEN:90. High quality reads were aligned on the SARS-CoV-2
327 reference genome (genbank ACC: NC_045512) with BWA-MEM v0.7.1. Duplicated reads were then
328 removed with Picard tool v2.25.0 (<http://broadinstitute.github.io/picard/>). Consensus sequence were

329 generated using a combination of SAMtools v1.11 and VarScan v2.4.1 variant caller. Consensus
330 sequences were reconstructed from VarScan output with an in-house script that automatically
331 introduces ‘N’ in low quality or uncertain/uncovered regions of the reference sequence.

332

333 *Sequence selection from GISAID*

334 The Vo’ genome sequences of SARS-CoV-2, divided by lineage, were used as input of cd-hit-est-2d
335 (32) and compared with the GISAID database. The first 20 most similar genomes of every cluster
336 were used as input of MAFFT (33) to produce a multiple alignment with the Wuhan genome
337 (GenBank ID MN908947.3) as reference. This dataset has been used to build the minimum spanning
338 network (MSN) and the phylogenetic tree.

339

340

341 *Minimum spanning network (MSN)*

342 The MSN(34) was built to find out the most probable path of contagion starting from a distance matrix
 343 calculated by the genome sequences of lineage B as classified by the Pangolin tool
 344 (<https://github.com/cov-lineages/pangolin>). The nodes of the MSN represent the genomes that share
 345 the same pattern of SNPs (with reference to the Wuhan genome MN908947.3), while an edge
 346 connects two nodes if they have compatible genomes (e.g. one can be obtained by the other by adding
 347 some mutations). Nodes are labelled with the sampling location and date of the corresponding
 348 sequences, edges with the number of mutations that differentiate the two sets of connected sequences.
 349 The minimum spanning network algorithm used to reconstruct the MSN from the distance matrix
 350 connects the virus sequences minimizing the edit distance (Hamming distance). In the reconstructed
 351 network, we defined a possible introduction as a node containing also sequences sampled outside the
 352 Vo' area. Additionally, the sampling date of the Vo' node (i.e. the introduction) must be later than
 353 that of the other GISAID sequences. Similarly, a possible exit is a node containing only Vo' sequences
 354 that are the source of an edge leading to a node containing only GISAID sequences (i.e. and no
 355 sequence from Vo'). Moreover, the sampling date of the Vo' node must precede that of the GISAID
 356 sequences. The MSNs were finally pruned to remove nodes not directly connected to a node
 357 containing sequences from Vo'.

358

359 *Phylogenetic analysis*

360 Two different sequence sets were used for depicting the phylogenetic relationships between Vo'
 361 samples and the European epidemiological framework. The first samples set included a total of 1252
 362 full genome sequences from Europe, collected in the same pandemic period. The second data set
 363 included a selection of sequences from the first set (n = 148). Both sets included SARS-CoV-2
 364 Wuhan-Hu-1 isolate (Genbank accession number MN90894) as reference (GenBank accession

number: MN908947.3). The 90 bp end at both 5' and 3' were eliminated, to avoid possible sequencing errors. *Phylogenetic* trees were constructed using RAxML (35) software. Each tree is statistically supported by bootstrap process (100 replicates) and Wuhan isolate was used as outgroup.

ETHICAL APPROVAL STATEMENT

The study was approved by the Ethics Committee for Clinical Research of the province of Padua under protocol number 0022591 - 7/4/2020. Study participation was by consent. For participants under 18 years of age, consent was provided by a parent or legal guardian.

ACKNOWLEDGEMENTS

This work was supported by the Veneto Region. E.L. and S.T. acknowledge research funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No 874735 (VEO). E.L. acknowledges funding from the University of Padova and the Department of Molecular Medicine (STARS-CoG ISS-MYTH and PRID/SID 2020). S.T. acknowledges funding from the University of Padua (TOPP_PRIV20_01 and TOPP_SID19_01). G.T. and A.C. acknowledge funding from Fondazione Umberto Veronesi, Misura Ricerca Covid 19, year 2020. We gratefully acknowledge the Authors from the Originating Laboratories and the Submitting Laboratories who generated and shared via GISAID the data on which part of this research is based.

384

385 REFERENCES

- 386 1. Coronavirus disease (COVID-19) – World Health Organization [Internet]. [citato 22 novembre 2021].
387 Disponibile su: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
- 388 2. Islam MR, Hoque MN, Rahman MS, Alam ASMRU, Akther M, Puspo JA, et al. Genome-wide
389 analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci Rep. 19
390 agosto 2020;10(1):14004.
- 391 3. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, et al. Large scale
392 genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic
393 variations of hotspots mutations. PLOS ONE. 10 novembre 2020;15(11):e0240345.
- 394 4. Roy C, Mandal SM, Mondal SK, Mukherjee S, Mapder T, Ghosh W, et al. Trends of mutation
395 accumulation across global SARS-CoV-2 genomes: Implications for the evolution of the novel
396 coronavirus. Genomics. novembre 2020;112(6):5331–42.
- 397 5. Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, et al.
398 Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature. 1 agosto
399 2020;584(7821):425–9.
- 400 6. Gittelman RM, Lavezzo E, Snyder TM, Zahid HJ, Elyanow R, Dalai S, et al. Diagnosis and Tracking
401 of SARS-CoV-2 Infection By T-Cell Receptor Sequencing [Internet]. 2021 feb [citato 8 novembre
402 2021] pag. 2020.11.09.20228023. Disponibile su:
403 <https://www.medrxiv.org/content/10.1101/2020.11.09.20228023v2>
- 404 7. Dorigatti I, Lavezzo E, Manuto L, Ciavarella C, Pacenti M, Boldrin C, et al. SARS-CoV-2 antibody
405 dynamics and transmission from community-wide serological testing in the Italian municipality of Vo'.
406 Nat Commun. 19 luglio 2021;12(1):4383.

- 407 8. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global
408 health. *Glob Chall*. 2017;1(1):33–46.
- 409 9. Capobianchi MR, Rueca M, Messina F, Giombini E, Carletti F, Colavita F, et al. Molecular
410 characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin Microbiol Infect*. 1
411 luglio 2020;26(7):954–6.
- 412 10. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature
413 proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. novembre
414 2020;5(11):1403–7.
- 415 11. Bianco L, Moser M, Silverj A, Micheletti D, Lorenzin G, Collini L, et al. On the origin and
416 propagation of the Covid-19 outbreak in the Italian province of Trento, a tourist region of Northern
417 Italy. *Viruses*.
- 418 12. Lai A, Bergna A, Caucci S, Clementi N, Vicenti I, Dragoni F, et al. Molecular Tracing of SARS-CoV-
419 2 in Italy in the First Three Months of the Epidemic. *Viruses*. agosto 2020;12(8):798.
- 420 13. Benedetti F, Snyder GA, Giovanetti M, Angeletti S, Gallo RC, Ciccozzi M, et al. Emerging of a SARS-
421 CoV-2 viral strain with a deletion in nsp1. *J Transl Med*. 31 agosto 2020;18(1):329.
- 422 14. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic
423 diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol*. 1 settembre 2020;83:104351.
- 424 15. Annavajhala MK, Mohri H, Wang P, Nair M, Zucker JE, Sheng Z, et al. Emergence and expansion of
425 SARS-CoV-2 B.1.526 after identification in New York. *Nature*. settembre 2021;597(7878):703–8.
- 426 16. Pachetti M, Marini B, Giudici F, Benedetti F, Angeletti S, Ciccozzi M, et al. Impact of lockdown on
427 Covid-19 case fatality rate and viral mutations spread in 7 countries in Europe and North America. *J*
428 *Transl Med*. 2 settembre 2020;18(1):338.
- 429 17. Gill M, Gray M. Mass testing for covid-19 in the UK. *BMJ*. 16 novembre 2020;371:m4436.

- 430 18. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of
431 SARS-CoV-2 in the Icelandic Population. *N Engl J Med*. 11 giugno 2020;382(24):2302–15.
- 432 19. Lai A, Bergna A, Toppo S, Morganti M, Menzo S, Ghisetti V, et al. Evolutionary Dynamics of SARS-
433 CoV-2 in Space and Time During the First Phase of the Epidemic in Italy. 29 luglio 2021 [citato 8
434 novembre 2021]; Disponibile su: <https://www.preprints.org/manuscript/202107.0654/v1>
- 435 20. Hensley MK, Bain WG, Jacobs J, Nambulli S, Parikh U, Cillo A, et al. Intractable Coronavirus Disease
436 2019 (COVID-19) and Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)
437 Replication in a Chimeric Antigen Receptor-Modified T-Cell Therapy Recipient: A Case Study. *Clin*
438 *Infect Dis*. 2 agosto 2021;73(3):e815–21.
- 439 21. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution
440 during treatment of chronic infection. *Nature*. aprile 2021;592(7853):277–82.
- 441 22. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of
442 SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med*. 3 dicembre 2020;383(23):2291–3.
- 443 23. Khatamzas E, Rehn A, Muenchhoff M, Hellmuth J, Gaitzsch E, Weiglein T, et al. Emergence of
444 multiple SARS-CoV-2 mutations in an immunocompromised host [Internet]. 2021 gen [citato 24
445 novembre 2021] pag. 2021.01.10.20248871. Disponibile su:
446 <https://www.medrxiv.org/content/10.1101/2021.01.10.20248871v1>
- 447 24. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2
448 within-host diversity and transmission. *Science*. 16 aprile 2021;372(6539):eabg0821.
- 449 25. Voloch CM, da Silva Francisco Jr R, de Almeida LGP, Brustolini OJ, Cardoso CC, Gerber AL, et al.
450 Intra-host evolution during SARS-CoV-2 prolonged infection. *Virus Evol* [Internet]. 1 novembre 2021
451 [citato 25 novembre 2021];7(2). Disponibile su: <https://doi.org/10.1093/ve/veab078>
- 452 26. Wang R, Hozumi Y, Zheng Y-H, Yin C, Wei G-W. Host Immune Response Driving SARS-CoV-2
453 Evolution. *Viruses*. ottobre 2020;12(10):1095.

27. Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. *Viruses*. marzo 2021;13(3):394.
28. Pearce N, Lawlor DA, Brickley EB. Comparisons between countries are essential for the control of COVID-19. *Int J Epidemiol*. 1 agosto 2020;49(4):1059–62.
29. Peto J. Covid-19 mass testing facilities could end the epidemic rapidly. *BMJ*. 22 marzo 2020;368:m1163.
30. Bosetti P, Kiem CT, Yazdanpanah Y, Fontanet A, Lina B, Colizza V, et al. Impact of mass testing during an epidemic rebound of SARS-CoV-2: a modelling study using the example of France. *Eurosurveillance*. 7 gennaio 2021;26(1):2001978.
31. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance*. 23 gennaio 2020;25(3):2000045.
32. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 1 dicembre 2012;28(23):3150–2.
33. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 15 luglio 2002;30(14):3059–66.
34. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 1 gennaio 1999;16(1):37–48.
35. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 1 maggio 2014;30(9):1312–3.

SUPPORTING INFORMATION

S1 Fig. Mutations defining Vo' haplotype in time

Number of sequences carrying G11083T (NSP6:L37F) and G26144T (NS3:G251V) mutations deposited in the GISAID database from December 2019 to September 2021. Mutations were analysed separately (green and pink line) and as part of the same haplotype (black line).

S2 Fig. S:L5F appeared in Iota variant in 2021

Number of sequences deposited in the GISAID database and collected from December 2019 to September 2021 carrying C21575T (S:L5F) mutation (yellow line) compared to sequences belonging to 21F clade (Iota variant, green line) and the combination of the two cases (black line).

S3 Fig. ML phylogenetic tree of European lineage B SARS-CoV-2 sequences at the beginning of the pandemic

Maximum likelihood phylogenetic tree of all European sequences collected at the beginning of the pandemic and uploaded in GISAID. Vo' sequences are coloured in red. All sequences carrying the Ancestor Haplotype are collapsed. A, B and C panels zoom in the subtypes of the ancestor haplotype observed in Vo'.

495 **S4 Fig. Minimum Spanning Network of European lineage B SARS-CoV-2 sequences at the**
 496 **beginning of the pandemic**

497 *MSN of all the European sequences collected at the beginning of the pandemic and uploaded in*
 498 *GISAID. Each dot represents a unique haplotype, with Vo' haplotypes coloured in red. Haplotypes*
 499 *are clustered according to the lineage they belong to, with each cluster being circled with a black*
 500 *dotted line, and showing the basic haplotype in the central node.*

501

502 **S5 Fig. Antibody titres of subjects positive to two sequential swab tests with an average time**
 503 **interval of 11 days**

504 *Antibody titres observed in subjects with stable (black) and unstable (blue) viral genome in an*
 505 *average time span of 11 days according to DiaSorin, Neutralisation, Abbott and Roche assays (Mann*
 506 *Whitney test).*

507

508 **S6 Fig. TCR features of subjects positive to two sequential swab tests with an average time**
 509 **interval of 11 days**

510 *TCR breadth class I (a) and class II (b) of patients with stable (black) and unstable (blue) viral*
 511 *genome in an average time span of 11 days (Mann Whitney test).*

512 *TCR depth class I (c) and class II (d) of patients with stable (black) and unstable (blue) viral genome*
 513 *in an average time span of 11 days (Mann Whitney test).*

514

515 **S7 Movie. Expanding Network based on symptoms onset data**

516 *Movie based on symptoms onset data reported by the infected individuals regardless of the*
 517 *availability of the viral sequences. Subjects reporting contacts are depicted as rectangles, while*
 518 *circles represent individuals reporting no contacts with the subjects traced in the main chain. Nodes*
 519 *with an available sequence are coloured in green, otherwise they appear in red. Individuals who did*
 520 *not report a symptom onset date appear in the last frame all together.*

Fig 1. Lineages circulating in Northern Italian regions

Fig 2. Maximum Likelihood phylogenetic tree and Minimum Spanning Tree (MST) showing relationships between various SARS-CoV-2 sequences. The tree is rooted on the left and branches out to the right. Sequences are color-coded: red for Vo', yellow for Extra Vo', and green for Vo' + extra Vo'. A scale bar of 0.4 is shown at the bottom left. A red arrow points to a specific node (Vo_SR_21). A blue square highlights a node (EPI_ISL_402125) labeled 'Wuhan reference'.

Fig 3. Within host variation.

a

Intra-family transmission

b

Likely Intra-family transmission

c

Extra-family transmission

Fig 4. Within family diversity and transmission

a

b

Fig 5. Network of contacts based on post-infection interviews

S1 Fig. Mutations defining Vo' haplotype in time

S2 Fig. S:L5F appeared in Iota variant in 2021

S3 Fig. ML phylogenetic tree of European lineage B SARS-CoV-2 s

S4 Fig. Minimum Spanning Network of European lineage B SARS-

S5 Fig. Antibody titres of subjects positive to two sequential swa

S6 Fig. TCR features of subjects positive to two sequential swab tests