SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to mass gathering events

Madlen Stange[1,2,3+], Alfredo Mari[1,2,3+], Tim Roloff[1,2,3+], Helena MB Seth-Smith[1,2,3+], Michael Schweitzer[1,2], Myrta Brunner[4], Karoline Leuzinger[5,6], Kirstine K. Søgaard[1,2], Alexander Gensch[1], Sarah Tschudin-Sutter[7], Simon Fuchs[8], Julia Bielicki[9], Hans Pargger[10], Martin Siegemund[10], Christian H Nickel[11], Roland Bingisser[11], Michael Osthoff[12], Stefano Bassetti[12], Rita Schneider-Sliwa[4], Manuel Battegay[7], Hans H Hirsch[5,6,7], Adrian Egli[1,2,*]

[1] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[2] Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[3] Swiss Institute for Bioinformatics, Basel, Switzerland

[4] Human Geography, University of Basel, Basel, Switzerland

[5] Clinical Virology, University Hospital Basel, Basel, Switzerland

[6] Transplantation & Clinical Virology, Department of Biomedicine, University of Basel, Basel, Switzerland

[7] Infectious Diseases and Hospital Epidemiology, University Hospital Basel and University of Basel, Basel, Switzerland

[8] Health Services for the City of Basel, Basel, Switzerland

[9] Pediatric Infectious Diseases, University Children's Hospital Basel, Basel, Switzerland

[10] Intensive Care Unit, University Hospital Basel, Basel, Switzerland

[11] Emergency Medicine, University Hospital Basel, Basel, Switzerland

[12] Internal Medicine, University Hospital Basel, Basel, Switzerland

+these four authors contributed equally to this work

*** correspondence**

Adrian Egli, MD PhD

University Hospital Basel

Petersgraben 4

4031 Basel, Switzerland

Email: adrian.egli@usb.ch

Phone: +41 61 556 5749

Abstract

Background. The first case of SARS-CoV-2 in Basel, Switzerland was detected on February 26th 2020. We present a phylogenetic study to explore viral introduction and evolution during the exponential early phase of the local COVID-19 outbreak from February 26th until March 23rd.

Methods. We sequenced SARS-CoV-2 naso-oropharyngeal swabs from positive 746 tests that were performed at the University Hospital Basel in the timeframe of our study. We successfully generated 468 high quality genomes from unique patients and called variants with our COVID-19 Pipeline (COVGAP). We analysed viral genetic diversity using PANGOLIN taxonomic lineages. To identify introduction and dissemination events we incorporated global SARS-CoV-2 genomes and inferred a time-calibrated phylogeny. We used epidemiological data to aid interpretation of phylogenetic patterns.

Findings. The early outbreak in Basel was dominated by lineage B.1 (83·6%), detected from March 2nd, although the first lineage identified was B.1.1. Within B.1, a clade defined by the SNP C15324T contains 68·2% of our samples ('Basel cluster'), including 157 identical sequences at the root of the 'Basel cluster', suggesting local spreading events. We infer the origin of the 'Basel cluster' defining mutation to mid-February in our tri-national region. The remaining genomes map broadly over the global phylogenetic tree, evidencing several events of introduction from and/or dissemination to other regions of the world via travellers. We also observe family transmission events.

Interpretation. A single lineage variant dominated the outbreak in the City of Basel while other lineages such as the first (B1.1) did not propagate. We identify mass gathering events and less so travel returners and family transmission as causes for the local outbreak. We highlight the importance of adding specific questions to the epidemiological questionnaires that are collected, to obtain data on attendance of large gathering events and locations as well as travel history to effectively identify routes of transmissions in up-coming outbreaks. This phylogenetic analysis enriches epidemiological and contact tracing data, allowing, even retrospectively, connection of seemingly unconnected events, and can inform public health interventions.

63 **Introduction**

64 The COVID-19 pandemic has rapidly spread around the globe during the first six months of 2020.
65 The causative coronavirus, SARS-CoV-2, is the subject of many studies using genomic analysis
66 providing key insights into viral diversity across cities[1], provinces[2-5], countries[6-11], and globally[12].
67 SARS-CoV-2 has an estimated mutation rate of $0.71\text{-}1.40 \times 10^{-3}$ [13], which translates to 21-42 mutations
68 per year. Due to the accumulation of mutations, phylogenetic analysis of SARS-CoV-2 is becoming
69 more granular over time[14], providing increasing resolution of transmission dynamics and events.
70 Comparisons of single nucleotide polymorphisms (SNPs) allows us to explore transmission events
71 with highest resolution across communities. The identification of transmission routes is important,
72 especially with various public health measures being introduced, such as lockdown policies, which
73 have been implemented on country or regional levels to limit viral transmission. The impact of public
74 health measures can be monitored through phylogenies[3]. Genomic data can also deliver insights into
75 mutations and whether they alter virulence, or aid adaptation to novel hosts. The spike protein D614G
76 mutation, for example, has been implicated in more effective transmission[15], although the actual
77 impact may be through fixation in an expanding lineage rather than conferring increased
78 transmissibility per se[16].

79 The scope of this study is to provide a more granular picture of the phylogenetic diversification and
80 propagation of the early-stage SARS-CoV-2 pandemic on a local scale. The City of Basel has a
81 population of 175,350 inhabitants (median over the past five years) with half a million people in the
82 Basel area. Situated in North-Western Switzerland, directly bordering both Germany and France,
83 Basel has almost 34,000 workers commuting daily across the international borders[17]. Given this large
84 exchange of people in this tri-national region, the fact that the neighbouring region Alsace, France,
85 was already experiencing an intense epidemic[18], and a low threshold testing strategy implemented
86 weeks before the first case, we aim to explore the early stage of SARS-CoV-2 transmission dynamics
87 in Basel and the surrounding area from the first case to one week post border closure.

Results

Characteristics of the longitudinal study

This cohort study includes all patient samples from Basel-City and the surrounding area during the initial 26 days of the local outbreak, between February 26th and March 23rd. Only single, non-repeated tests per patient were considered eligible for phylogenetic analysis. This timeframe covers the first two positively tested cases in Basel on February 26th, which we were able to capture via early implementation of PCR-based detection by routine diagnostics, until the date of border closure plus seven days (March 23rd).

From the first case on February 26th 2020 until March 23rd we had performed 6,943 PCR tests. Of these, 746 samples (10.7%) were SARS-CoV-2 positive (**Figure 1**). March 23rd had the maximum number of positive tests, with 66 cases. Of all PCR tested patients and positively tested patients during the study period, a majority were female with a median age of 42 and 49 years, respectively (**Table 1**). Of the PCR-confirmed cases, only 17 (2.3%) were in patients younger than 18 years. (**Table 1**, **Figure S1**). 418 (56%) were living in the canton of Basel-City (City of Basel, Riehen, Bettingen) and 328 (44%) were from the surrounding area.

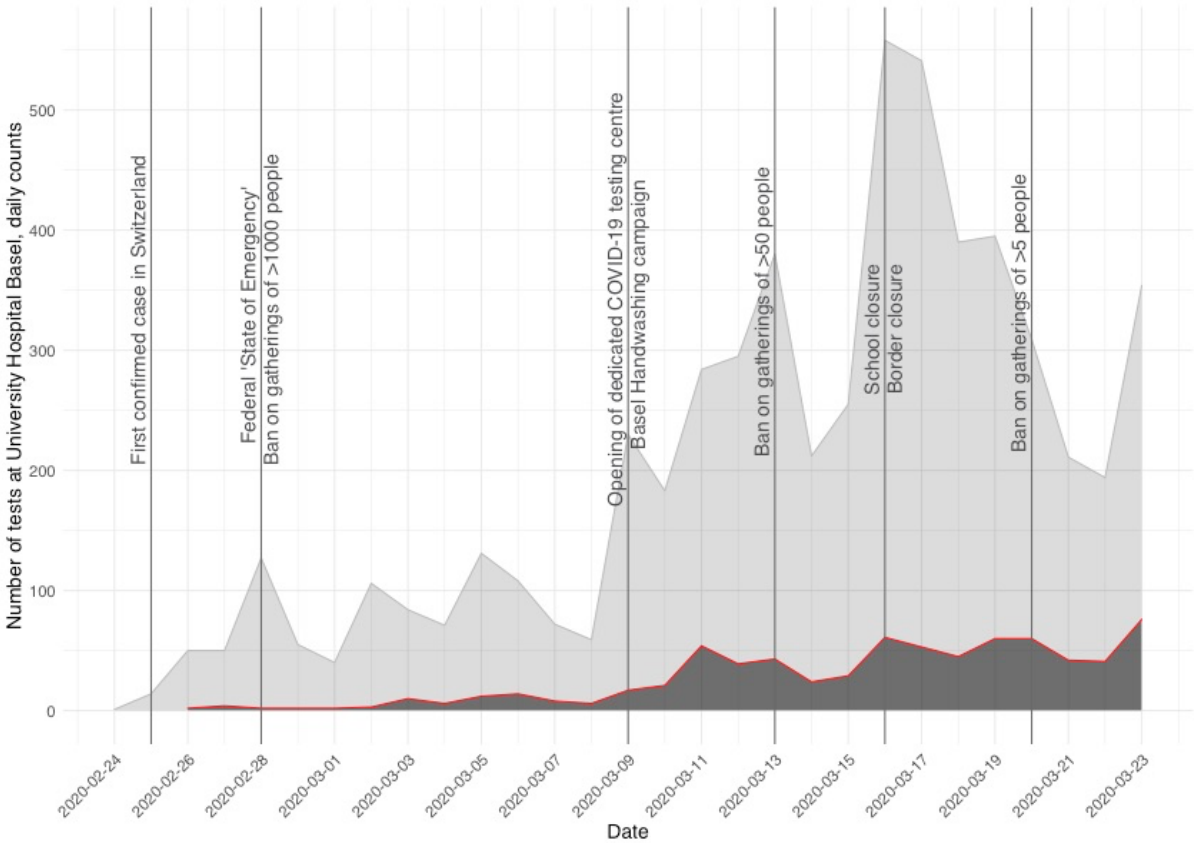

Figure 1. Epidemiological curve of the first COVID-19 wave in the city of Basel and hinterland, Switzerland. Positive (red line, dark grey area) and negative (light grey area) SARS-CoV-2 PCR tests are depicted from the beginning of the outbreak in February to March 23, 2020. Major events and imposed restrictions are marked by horizontal lines. First confirmed cases in Switzerland and Basel were on February 25th and February 26th, respectively.

COVGAP pipeline validation

No false positive variants were called (**Figure S4**, **Table 2**, **Table S1**). Ambiguous mapping was responsible for the failure to call the three indels, resulting in insufficient (<70%) coverage to be reliably called as a variant. This validation allowed us to determine the specificity (100%), sensitivity (94.2%), and accuracy (100%) of COVGAP, thus confirming its accuracy in SNP detection, as well as its ability to call the majority of indels from short read data.

The COVGAP pipeline produced 468 (63%) high quality genomes for subsequent analysis, of the 746 samples taken. The remaining samples were either not available (N = 57), did not pass sequencing quality control (N = 156), or were duplicates from the same patient (N = 65). These 468 samples are

118 subsequently referred to as the Basel area cohort. Of these, 240 (51.9%) were from female patients
119 (**Table 1**, **Figure S1**), and 12 (2.6%) were from patients younger than 18 years.

120 **Phylogenetic lineages observed over time in the Basel area cohort**

121 Over the 26-day study period, 13 out of 91 globally circulating phylogenetic lineages were recorded
122 in the Basel area; only one additional lineage was recorded in all Swiss sequences.
123 Lineage B.1 dominated the cases during the initial phase of the outbreak (**Figure 2**), with 83.6% (N =
124 391) of sequenced samples (**Table 3**), being recorded for the first time in Basel on March 2nd. The
125 first patient diagnosed at our hospital, on February 26th, had a virus belonging to lineage B.1.1. This
126 lineage is seen sporadically through the outbreak with a maximum of six sequenced cases from March
127 16th. Lineage B.1 is associated with the Italian outbreak[14], yet both B.1.1 (35.7%) and B.1 (51.0%)
128 were the most prevalent lineages in Italy during this time span (**Figure 3**). From March 13th, rarer
129 lineages are seen in the Basel area, such as B.1.1.6, a lineage that is associated with an Austrian
130 origin[14]. Only two cases from the A lineage and sub-lineages were sequenced in the Basel area cohort
131 (**Table 3**).

132 **Lineage diversity in Basel-City, Switzerland, and neighbouring countries**

133 Comparison with lineages found in neighbouring countries over this period shows that lineage B.1
134 dominates in all, but the numbers of lineages identified, and the proportions vary (**Figure 3**).
135 Switzerland (77.1%) has a similarly large proportion of B.1 lineage to France (90.0%). We describe
136 the viral diversity based on abundance of lineages (retrieved from GISAID, for details see
137 **Supplementary material)** using a range of diversity indices (**Table S5**). Simpson diversity, which
138 accounts for differences in sample abundance between countries, was highest in Germany (3.87) and
139 Austria (3.73), followed by Italy (2.52); it was lowest in Switzerland (1.62) and France (1.23). The
140 share of our samples that originates from Basel-City residents (376 samples) excluding sequences that
141 were obtained from commuters mirrors the lineage proportions of Switzerland (**Figure 3**), while
142 contributing 56% of all available sequencing data for Switzerland within this timeframe (GISAID
143 database as of June 22nd,[19,20]).

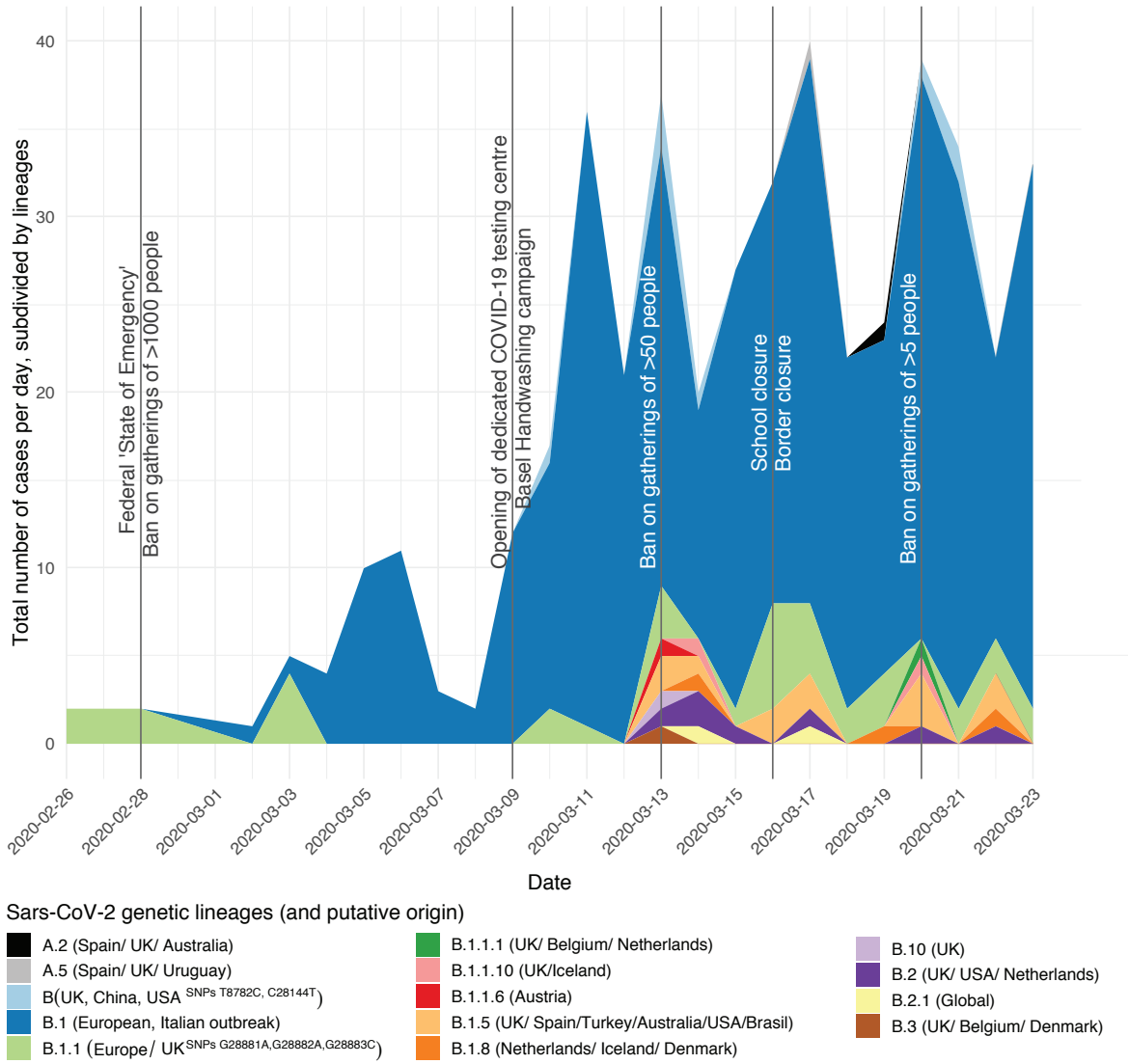

Figure 2. Detection of SARS-CoV-2 lineages found in the Basel area cohort from the first detected case on February 26th to March 23rd 2020. Major events and imposed restrictions are marked by horizontal lines. Low abundant lineages increase after the end of winter school vacation (March 8th) and are introduced by travel returners.

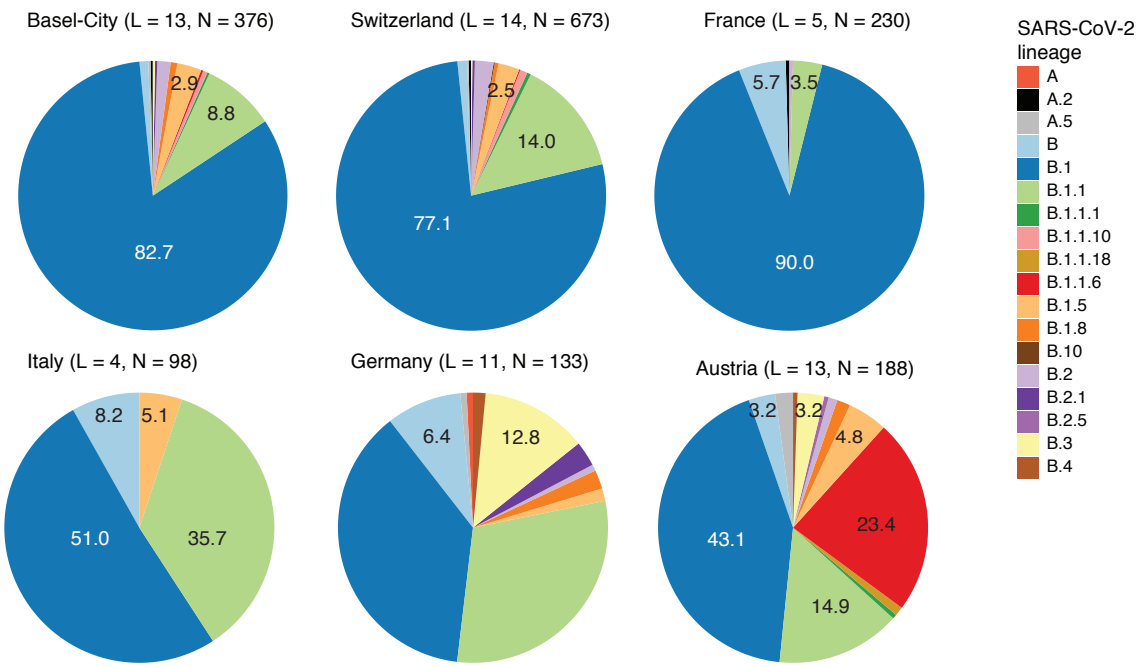

Figure 3. SARS-CoV-2 lineage diversity in neighbouring countries to Switzerland from first detected case until March 23rd, 2020. Number of lineages (L) and total number of genomes (N) per country in brackets, values within charts represent percentages. France was the first country in Europe that had confirmed COVID-19 cases on January 24th, followed by Germany on January 27th, Italy on January 31st, and some weeks later Austria and Switzerland followed on February 25th. Simpson diversity based on the available genomes and PANGOLIN lineage assignments is largest in Germany (3.87) and Austria (3.73), followed by Italy (2.52), it is smallest in Switzerland (1.62) and France (1.23). Basel-City mirrors the lineage proportions of Switzerland, while contributing half of Switzerland's sequence data.

Basel samples in global phylogenetic context

In order to better contextualize our findings, we analysed our virus genomes phylogenetically with a subset of global publicly available sequences (see **Supplementary material**; **Figure 4**). While phylogenetic lineages may show some geographical signal, lineages do not exclusively correspond to continents (**Figure 4A**), illustrating the degree of global interconnectivity and speed of spreading. The phylogenetic lineages recorded in the Basel area are distributed across the global phylogenetic tree

(**Figure 4B**). Mismatch of 'taxonomic' assignment and phylogenetic origin of genomes assigned to B.1 can be seen, with several as yet unnamed sub-lineages apparent in the phylogeny (**Figure 4B**). We can identify a major clade, within lineage B.1, comprising 68.2% of our samples (319/468 samples with 264 (82.8%) from patients from cantons Basel-City and Basel-Landschaft; **Figure 5A**). The remaining Basel area sequences (31.8%) (**Figure 5 B-C, Figure S6 B-C**) are spread throughout the phylogeny and cluster with global genomes. Introductions and features of some of the clades are analysed in the following sections and in **Supplementary material.**

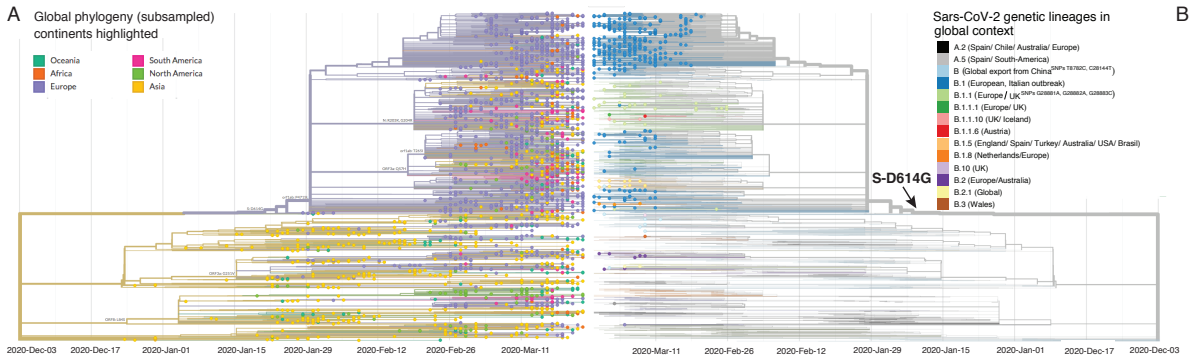

Figure 4. SARS-CoV-2 phylogeny of Basel area samples and genetic lineages (PANGOLIN) in a global context. A. Time tree of SARS-CoV-2 genomes from the Basel area cohort as well as subsampled global genomes (30 genomes per country and month), coloured by continent of origin. Amino acid mutations at internal nodes representing clade defining mutations are shown. **B.** Mirrored time tree coloured by genetic lineages sensu PANGOLIN v.May19 (https://github.com/cov-lineages/). Each tip with a circle represents a genome from the Basel area cohort, branches without circled tips represent global genomes, included to confer the global context of the Basel genomes.

187 **The first identified introduction of SARS-CoV-2 to Basel**

188 The first two positively diagnosed patients with SARS-CoV-2 in Basel and first identified cases of
189 COVID-19, *Patient 1* and *Patient 2*, travelled together to Italy. In our analysis, both *Patient 1* and
190 *Patient 2* carried viruses from the B.1.1 lineage, the second most prevalent lineage in Italy at that time
191 (**Figure 3**). Interestingly, the two virus genomes from *Patient 1* and *Patient 2* are separated by two
192 SNPs, suggesting two independent infections (**Figure 5B**). Moreover, we did not identify minor
193 alleles within these samples that would hint at double infection of the patients. In this case, the
194 epidemiological cluster of *Patient 1* and *Patient 2* is not congruent with the phylogenetic inference.

195 The virus genome of *Patient 1* carried a synonymous mutation at C313T in *ORF1ab*, which is found
196 in samples from Israel, Hungary, Japan, USA, Argentina, Greece, India, Brazil, Morocco, and
197 Netherlands among others (nextstrain.org), all sharing an unsampled common ancestor that emerged
198 around February 25th (CI February 23-26th). This SNP is also found in ten other Basel area cohort
199 genomes: eight of these were from a family and social friends cluster unrelated to *Patient 1* that were
200 diagnosed between March 13th and March 22nd.

201 The virus genome of *Patient 2* carried a synonymous mutation at T19839C in *ORF1ab* and not
202 C313T, and clusters together with eight identical virus samples sampled between February 26th and
203 March 23rd. Two of these are from family members of *Patient 2*, who tested positive two and six days
204 later, suggesting a family route of infection. The epidemiological data of the other six patients
205 suggests no direct transmission via *Patient 2*. One of the six returned from a Swiss ski resort three
206 days prior to onset of symptoms. Two additional samples forming a family cluster (*Family 1*, **Figure**
207 **5B**) diagnosed on March 3rd, carried the T19839C plus non-synonymous mutation G28179A leading
208 to amino acid change ORF8-G96S. One member of *Family 1* travelled with *Patients 1* and *2* to Italy
209 and possibly got infected there with a yet different virus variant.

210 **Introduction of the Basel cluster**

211 The clade within lineage B.1, into which 68.2% (N = 319) of our Basel area cohort sequences fall, is
212 characterized by a synonymous SNP C15324T in *ORF1ab*, henceforth referred to as the "Basel
213 cluster". *Patient 1* and *Patient 2* are not linked to this Basel cluster. The first sample within the Basel

214 cluster, from March 2nd, was from a patient residing in Central Switzerland, who was transferred to a
215 care facility in Basel where six further people tested positive between March 5th and March 17th with
216 identical viral genomes. The source of the first infection is unknown. The second sample within the
217 Basel cluster was from a patient diagnosed on March 3rd, who had attended a religious event in
218 Alsace, France, that took place between February 17th and 21st. One other patient, who tested positive
219 on March 9th and had symptoms 14 days prior to testing, also attended this event.

220 **Description of the Basel cluster**

221 The 319 genomes in the Basel cluster (**Figure 5A**) show a divergence of up to five SNPs up until
222 March 23rd, although 157 genomes are identical and located at the root of the clade. This infers that
223 the common ancestor originated between February 9th and February 17th. That the clade defining SNP
224 C15324T (GISAID emerging clades label 20A/15324T) was registered globally for the first time on
225 March 2nd simultaneously in the Basel area sample 42173111 and GISAID sample
226 Germany/FrankfurtFFM7/2020 suggests unsampled circulation of this variant from mid-February. By
227 searching all genomes available on GISAID (N = 80,189; as of August 12th, 2020), filtering for
228 genomes belonging emerging clade 20A/15324 and sampled until March 23rd (N = 2,856), we find
229 that subsequently the C15324T mutation has also been observed in other countries but remains most
230 prevalent in Switzerland (N_{GISAID} =57/213, 26.8%; $N_{GISAID+this\ study}$ = 386/675, 57.2%; first genome
231 42173111 from March 2nd) (**Table S3**). Of the 57 Swiss genomes with this mutation that were already
232 deposited on GISAID, 26 also originate from the cantons Basel-City and Basel-Landschaft (**Table**
233 **S4**). The mutation is also found in higher proportions in genomes from France (N = 69/369, 18.7%,
234 first from March 3rd sample France/HF1870/2020), Luxembourg (N = 24/116, 20.7%, from March 8th
235 sample Luxembourg/LNS2614631/2020), and Belgium (N = 40/268, 14.9%, from March 6th sample
236 Belgium/NKR-030645/2020), but date later than the first recorded occurrence from Switzerland and
237 Germany (**Table S3**). Subsequently, as of March 9th, descendants of this variant were recorded
238 outside Europe (**Table S4**) in smaller proportions than in Basel (**Table S3**), which suggests
239 dissemination from the Basel area.

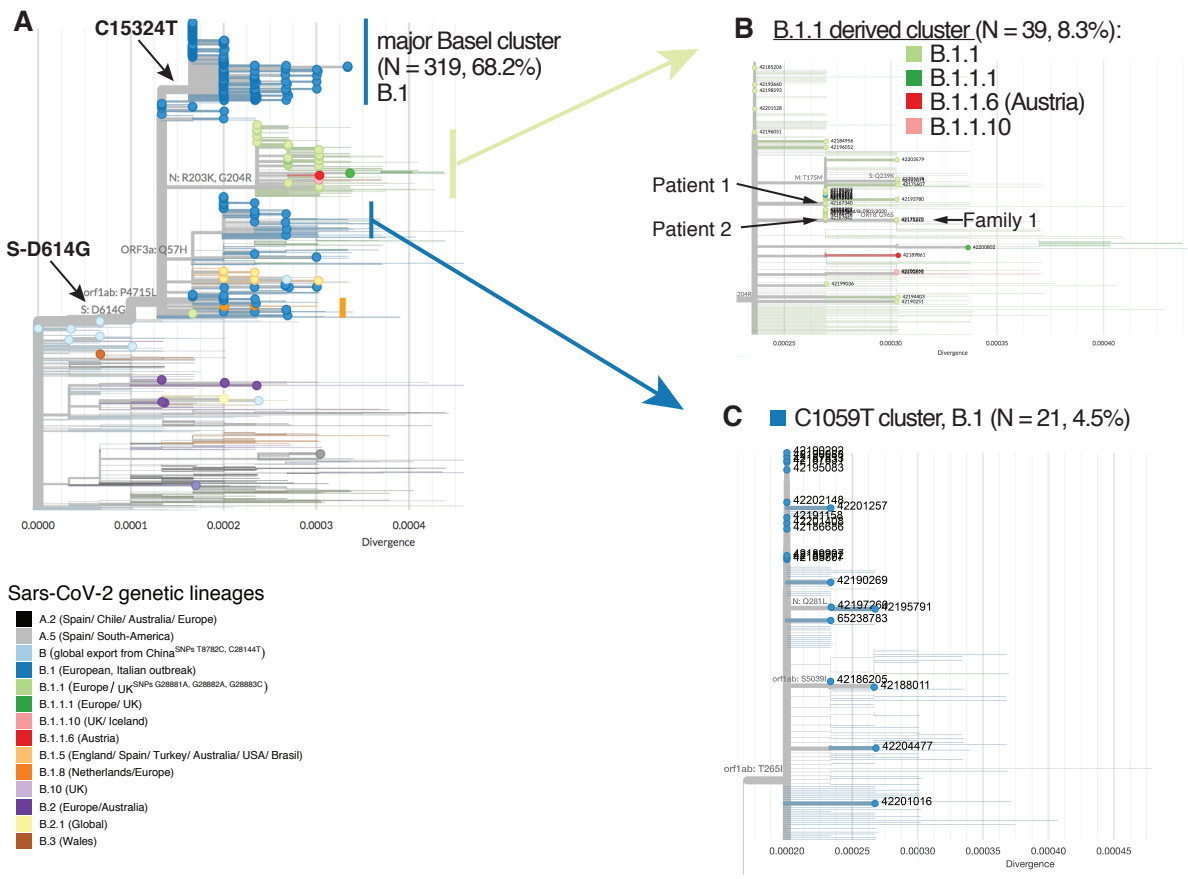

Figure 5. Divergence trees plotting nucleotide divergence between 468 genomes and expanded clusters of genomes in selected phylogenetic lineages. A. Genomes from Basel area cohort in global context. Tree composition is identical to the time tree from Figure 4. Branches with circles at the tip represent genomes from the present study; branches without circles represent global genomes from GISAID. The major Basel cluster contains samples with up to five mutations. **B**. Zoom into a mixed cluster derived from B.1.1 with seven to nine mutations difference to the root. A single genomes assigned to lineage B.1.1.6 has an assumed origin in Austria; two genomes (B.1.1.10) most likely originate from the UK. **C**. Potential ski-holiday related cluster (C1059T) with seven samples having a confirmed association with skiing destinations. Note: Divergence can be translated to number of mutations difference to the root (Wuhan-Hu-1) by multiplication by with the SARS-CoV-2 genome size (29903 bases).

252 **Potential ski-holiday related cluster (C1059T)**

253 Previous reports have identified viruses in lineage B.1 carrying SNP C1059T (amino acid change
254 ORF1a-T265I) in travel returners from ski holidays in Ischgl, Austria[9,21]. We detected 21 (4.5%) viral
255 samples within our cohort with these features (**Figure 5C**), divergent by up to two SNPs from the
256 common ancestor. Samples date from March 11th to March 23rd with an inferred internal node age of
257 February 21st (CI: February 20th-February 21st, 2020), fitting with possible infections in Ischgl from
258 end of February to March 13th [22]. Epidemiological data confirms that five patients from that cluster
259 had returned from Austria, and three specifically from Ischgl, the fourth from Tyrol, before testing
260 positive. Two additional patients returned from skiing in Swiss ski resorts.

261 **Spike protein mutation prevalent in Basel patients**

262 The spike protein S-D614G mutation is associated with the B.1 lineage and all those derived from this
263 (**Figure 4**). As such, it occurs in 448 of the 468 (95.7%) samples from the Basel area. The SNP
264 responsible has not been lost once in our sub-sampled dataset, but it is not present in B.2 or B.3 or
265 other sister lineages to B.1 (**Table 3**). Among our samples, we found no significant difference in viral
266 loads between patients with and without the S-D614G mutation ($z = -0.881$, $p = 0.38$). However, our
267 cohort is biased to samples with higher viral load, as these were those that were successfully
268 sequenced.

269 **Discussion**

270 We reconstruct the early events focusing on introduction and spread of SARS-CoV-2 in Basel and the
271 surrounding area, Switzerland, from a phylogenetic perspective. We present COVGAP, a new
272 combination of existing tools to effectively and efficiently mine SARS-CoV-2 genomes from Illumina
273 paired end reads. Unlike other currently available tools[23], COVGAP shows higher sensitivity levels in
274 SNP calling from raw reads (100%), failing only in ambiguously mapped deletions and insertions. In
275 such cases, it adopts a coverage-conservative approach, needed to reliably call variants in real world
276 scenarios.

277 The majority of genome variants in Basel are similar to those from France, Italy, and Germany. We
278 found the presence of 13 SARS-CoV-2 lineages in our samples, with the beginning of the Basel
279 outbreak being powered by the European B.1 lineage. In particular, a B.1 lineage variant with the
280 C15324T mutation dominated the early phase of the local spread with 70% of samples forming a large
281 Basel cluster. Compared to Victoria, Australia[5], the UK[24], or Austria (this study) the diversity seen
282 arriving in Switzerland and Basel as determined using Simpson diversity is more limited, reflecting
283 European rather than intercontinental connections. This diversity measure can be used to monitor viral
284 introductions as an effect of travel restrictions in the future.

285 The Basel cluster virus variant 20A/C15324T was first detected in Europe on March 2nd in Germany
286 and Switzerland simultaneously. We locate its geographic origin to our tri-national region between
287 February 9th and February 17th. Our epidemiologically-informed phylogenetic analysis indicates that
288 the Basel cluster represents a larger transmission chain that was unchecked and spread effectively
289 among unrelated people throughout Basel and eventually outside of Europe. The first recognized case
290 in the large Basel cluster goes back to a patient in a care facility, in which several more infections
291 occurred.

292

293 The beginning of the COVID-19 outbreak described here coincided with the winter school holidays in
294 Basel, from February 22nd to March 8th. During this time, many residents take the opportunity to
295 travel, in particular to skiing resorts. Viral introductions from ski resorts are known from contract
296 tracing data to have affected Germany[25], Denmark[21], Iceland[9], France, Spain, and UK[26], with Ischgl,
297 Austria being a described source of many cases. Our data supports the finding that skiing resorts in
298 Austria and Switzerland served as dissemination hotspots. This school holiday also provided the
299 opportunity for the first identified SARS-CoV-2 introductions to Basel through two jointly returned
300 travellers, notably each with different viral variants. Overall, however travel returners did not drive
301 the outbreak in Basel evidenced by the low diversity of variants and proportion of such variants in our
302 sample. A second likely source represents the many workers travelling daily across borders from
303 France and Germany, particularly from heavily affected areas such as Alsace[18]. As the B.1 and B.1.1

304 lineages were dominant in France and Germany, these may be some sources of cases and
305 transmissions.
306
307 The timing of the epidemic in Basel also coincided with three major events. Firstly, a religious event
308 from February 17th to 21st in Alsace that was described as a super-spreading event in France[27]. We
309 confirmed that the virus genomes of two patients known to have attended are indeed situated at the
310 root of the clade that constitutes the Basel outbreak. Secondly, carnival in the Basel area is celebrated
311 over several weeks from early January, with numerous events, and thousands of active participants.
312 The culmination is the UNESCO world-heritage 'Basler Fasnacht', scheduled this year (2020) for
313 March 2nd-4th, but cancelled due to COVID-19. Notably, the active participants practice the piccolo
314 and drums over weeks in closed rooms as a preparation for their performance during the carnival, and
315 unofficial events are likely to have taken place. Thirdly, Basel hosted three international soccer events
316 at the St. Jakob Stadium on February 15th (20,675 spectators), 23rd (20,265 spectators), and 27th
317 (14,428 spectators). All three major events fell around the inferred date of origin of the Basel
318 mutation C15324T and subsequent dissemination phase.
319
320 A limitation of the current study is that we are likely to have missed some cases as not all
321 symptomatic people were advised to be tested, especially children younger than 18 years old.
322 Nevertheless, our cohort represents a very high sequencing density per detected case for a city (468
323 genomes from 746 PCR-confirmed cases in Basel area (62.7%) and from 10,680 PCR-confirmed
324 cases nationwide (4.4%)) for this early phase of the pandemic[28].
325
326 The availability and integration of epidemiological data in the interpretation of phylogenetic clades
327 underlines the validity of instrumentalising those tools for improving the understanding of SARS-
328 CoV-2 outbreak dynamics. Utilizing the clades as the backbone for targeted epidemiological analysis
329 of specific cases helped in grasping how mass gatherings, travel returners, and care facilities may
330 influence an outbreak within a city. The epidemiological data that was collected for the Federal Office

331 of Public Health (FOPH), as requested by law, helped tremendously to verify travel related links;
332 however it was not designed to obtain data on local super-spreading events such as attendance to
333 soccer games, visiting clubs, restaurants, bars, and concerts and future versions could be improved.
334 The classical epidemiological context is very important to further explain molecular epidemiological
335 links especially in a still not very diversified virus.
336
337 In conclusion, the start of the outbreak of SARS-CoV-2 in the Basel area was characterized by a
338 dominant variant, C15324T, within the B.1 lineage, which we infer to have arisen in mid-February in
339 our tri-national region. Large gatherings (potential super spreading events) could have had profound
340 effects on outbreak dynamics. Improved surveillance measures are needed in the management of an
341 outbreak, including large-scale, active screening in the broader public, including more children to
342 assess their role in transmission. Our analysis shows the potential of molecular epidemiology to
343 support classical contact tracing, even retrospectively, in order to evaluate and improve measures to
344 contain epidemics like COVID-19.
345

346 ## Materials and Methods

347 **Patients, samples, and diagnosis**

348 Respiratory samples from the University Hospital Basel and the University Children's Hospital Basel
349 (UKBB) patients were tested for SARS-CoV-2: from January 23rd 2020 testing was based on current
350 case definitions from the Federal Office of Public Health (FOPH); from 27th February additionally, all
351 respiratory samples negative for other respiratory pathogens were tested. Patient samples which tested
352 positive for SARS-CoV-2 [29,30] up to and including March 23rd were considered eligible for the present
353 study. In total 6,943 diagnostic tests were performed during the study period. The 746 positively
354 tested cases came predominantly from the administrative unit of Basel-City, Riehen, and Bettingen
355 (418, 58%), while the remaining patients were from Basel-Landschaft and neighbouring cantons and
356 countries.

357 For diagnosis, swabs from the naso- and oropharyngeal sites (NOPS) were taken, and combined into
358 one universal transport medium tube (UTM, Copan). Total nucleic acids (TNAs) were extracted using
359 the MagNA Pure 96 system and the DNA and viral RNA small volume kit (Roche Diagnostics,
360 Rotkreuz, Switzerland) or using the Abbott m2000 Realtime System and the Abbott sample
361 preparation system reagent kit (Abbott, Baar, Switzerland). Aliquots of extractions were sent for
362 diagnosis to Charité, Berlin, Germany from January 23rd - 29th, and to Geneva to the National
363 Reference Centre (NAVI) in Switzerland from January 29th. In-house analysis started February 27th as
364 part of the hospital routine diagnostics as previously described [30].

365 **Whole genome sequencing (WGS)**

366 SARS-CoV-2 genomes were amplified following the amplicon sequencing strategy of the ARTIC
367 protocol (https://artic.network/ncov-2019) with V.1 or V.3 primers [31]. In detail, real-time reverse
368 transcriptase (RT) reactions were run to a total volume of 10µl extracted total nucleic acid. After
369 some optimization, PCR used 25 cycles for samples with a diagnostic cycle threshold (C_t) value lower
370 than 21 (viral loads higher than 8.2 log10 Geq/ml); 40 cycles for all other samples (lower viral load
371 samples) and repeats. Purified amplicons were converted into Illumina libraries with Nextera Flex

372 DNA library prep kit (Illumina) automated on a Hamilton STAR robot, using 5ng input DNA. 96
373 libraries were multiplexed and sequenced paired-end 150 nucleotides on an Illumina NextSeq 500
374 instrument.

375 *Consensus sequence generation and detection of mutations*

376 After demultiplexing using bcl2fastq software version v.2.17 (Illumina), COVGAP (COVid-19
377 Genome Analysis Pipeline) was used (**Figure S2**). This incorporates: quality filtering using
378 trimmomatic software version v.0.38 [32] to remove Illumina adaptors and PCR primer sequences from
379 read ends; removal of reads smaller than 127 bases, and removal of reads with a phred score under 20
380 (calculated across a 4-base sliding window). Quality filtered reads were mapped to the Wuhan-Hu-1
381 reference MN908947.3 [33] using the BWA aligner [34]. Reads flagged as mapping to the reference were
382 retained [35], and are deposited under project PRJEB39887. SNPs and indels with respect to the
383 reference sequence were called using pilon version 1.23 [36]. Pilon summary metrics 'alternative allele
384 fraction' (AF) and 'depth of valid reads in pileup' (DP) were used to identify major and minor alleles
385 across all bases, which is not implemented in pilon itself. Major alleles were called if supported by
386 70% of the reads covering the variant locus (AF) for any locus with a minimum of 50x coverage
387 (DP). Variants were applied to the reference to produce a consensus sequence; any base position with
388 less than 50x coverage was masked with ambiguous characters (Ns) using BCFTools version 1.10.2
389 [37]. Consensus sequences were accepted for further analysis when containing up to 10% Ns. Summary
390 statistics, logs, coverage plots, and genome stack plots were generated using R version 3.6.0 and
391 packages Gviz v1.30 [38], Sushi v1.23 [39], seqinr v3.6.1 [40], and ggplot2 v3.11 [41]. COVGAP also provides
392 per genome quality control visual outputs (**Figure S5**) and is available at
393 https://github.com/appliedmicrobiologyresearch.

394 Quality control statistics such as the relationship between C_t-value and number of mapped reads and
395 coverage are presented in **Figure S3**. In general, we observed a negative trend linking C_t values and
396 percentage of ambiguous bases (Ns) being called as a result of low coverage. Sequences passing the
397 quality filter (n=533) showed a lower C_t value (median: 22.4±5.14) than the ones that failed (n =156;
398 median: 35.75.9±5.75).

399 *COVGAP Validation*

400 We used a set of 15 randomly *in silico* mutated SARS-CoV-2 mock genomes for the validation of the
401 specificity (identification of true negatives) and accuracy (identification of true negatives and true
402 positives) of COVGAP. Additionally, the genome MT339040, which harbours an 81 nucleotide
403 deletion in the ORF7a gene and a further seven SNPs relative to the reference [42] was used. Together,
404 the mock genomes possess 38 mutations including 30 SNPs, six deletions and two insertions across
405 the reference genome MN908947.3. The genomes were then shredded to artificial paired-end 150
406 nucleotide reads using SAMtools wgsim [37] and processed by COVGAP. For validation purposes,
407 original mock genomes and the COVGAP generated genomes from the shredded reads were aligned
408 using Seaview v4.6 [43] and clustalw [44]. A phylogeny was built using PhyML within Seaview with
409 default parameters.

410 **Phylogenetic lineage assignment of Basel samples**

411 To assess the phylogenetic diversity of SARS-CoV-2 samples during the early phase of the pandemic
412 we inferred the lineage assignment for each consensus sequence derived from the COVGAP pipeline
413 using PANGOLIN ver. May 19th (Phylogenetic Assignment of Named Global Outbreak LINeages) [14]
414 available at github.com/hCoV-2019/pangolin. Details on lineage summaries, describing which
415 countries lineages have been reported from and where transmission events have been recorded, can be
416 found at https://github.com/hCoV-2019/lineages. Lineage assignments were used to aid visualization
417 of phylogenetic diversity in Basel in a global context. For global sequences we used the PANGOLIN
418 lineage assignments as provided by GISAID (https://www.gisaid.org/; [19,20]) (details next section),
419 which were used for plotting purposes on phylogenetic trees.

420 To compare the lineage diversity in Switzerland and Basel-City to neighbouring European countries
421 (Austria, France, Germany, and Italy) during the early phase of the pandemic, we visualized relative
422 abundances of lineages using all high-quality, on GISAID (downloaded June 22nd, 2020) available
423 consensus sequences for the time until March 23rd from Austria (N = 188), France (N = 230),
424 Germany (N = 133), and Italy (N = 98). For Switzerland (N = 673), we combined our sequences (N =
425 468) with other sequence data from Switzerland [48] published on GISAID. To infer the diversity for

426 canton Basel-City (including Bettingen and Riehen) excluding sequences that were obtained from
427 commuters, we used the Basel-City portion (N = 376) of the Basel area cohort excluding samples
428 from patients from cities outside of the administrative district of Basel-City. We calculated Simpson
429 diversity (inverse Simpson concentration) as implemented in the SpadeR package v.0.0.1 [45-47], which
430 controls for lineage abundance differences between the countries, which is dependent on available
431 sequence data, and which ranges from 0 (no diversity) to indefinite (large diversity).

432 **Analysing Basel SARS-CoV-2 genomes in global phylogenetic context**

433 High-quality and full-length consensus sequences and corresponding metadata (sample ID, date of
434 sample, geographic location of sampling, PANGOLIN lineage) from Swiss [48] and global viruses were
435 downloaded from GISAID on June 22nd, 2020, making 49,284 individual genome sequences. 43,252
436 sequences were retained after filtering for genomes with under 10% ambiguous characters (Ns)
437 (author Genivaldo Gueiros Z. Silva) [49]. Metadata and consensus sequences of the Basel samples and
438 global data from GISAID were combined for further joint analysis, which were performed using
439 custom R scripts and the nextstrain command line interface analysis pipeline v.2.0.0 (nextstrain.org)
440 and augur v.8.0.0 [50].
441 Dates in our study samples correspond to date of sampling. Sequences were filtered by date from
442 December 1st 2019 to March 23rd 2020 using an R custom script in R version 4.0.0 [51] and packages
443 tidyr ver.1.1.0 [52], dplyr ver. 1.0.0. [53], and readr ver. 1.3.1. [54]: 15,973 consensus sequences, including
444 the Basel area sequences, remained. These time-filtered sequences were sub-sampled by geographic
445 location to 30 sequences per country and month. Non-human derived viruses as well as sequences
446 with other ambiguous characters (Us), as well as those from cruise ships, and duplicated sequences
447 defined by the nextstrain team as of June 24th (https://github.com/nextstrain/ncov/) were excluded
448 using *augur filter* [50] resulting in 2,485 sequences for the final phylogenetic analysis dataset.
449 Consensus sequences were aligned to the NCBI Refseq sequence Wuhan-Hu-1 reference
450 MN908947.3 using mafft v7.467 with method FFT-NS-fragment [55] and options --reorder --keeplength
451 --mapout --kimura 1 -- addfragments --auto. The resulting alignment was end-trimmed to remove
452 low-quality bases (bases 1-55; 29804-29903). We masked homoplasic sites (**Table S2**) that have no

453 phylogenetic signal [56] (deposited at https://github.com/W-L/ProblematicSites_SARS-CoV2). Please
454 note, that this list is under constant development as number and diversity of sequence data evolves;
455 we retrieved the data on June 19th, 2020. Masking was done using *augur mask*.
456 The resulting alignment was analysed in IQ-TREE 2 [57] for tree inference using *augur tree* with
457 substitution model GTR+G. The tree in Newick format was then subjected, together with the date
458 information of each genome and the initial sequence alignment, to an estimation of the evolutionary
459 rate by a regression of the divergence (number of mutations) against the sampling date using
460 TreeTime [58] implemented in *augur refine*. Genomes or branches that deviated more than four
461 interquartile ranges from the root to the tip versus the time tree were removed as likely outliers. The
462 resulting time-calibrated and divergence trees were re-rooted to MN908947.3 and MT291826.1, the
463 first official cases and published genomes of SARS-CoV-2 from Wuhan, China.
464 Ancestral trait reconstruction of each patient's viral genome was done for region (continent) and
465 country as well as region and country of exposure using *augur* traits with a sampling bias correction
466 of 2.5. Internal nodes and tips (actual genomes) were annotated regarding their nucleotide and amino
467 acid changes in relation to the reference using *augur ancestral* and *augur translate*, respectively. All
468 data were exported as json files (supplementary files) using *augur export v2* to be visualized in
469 *auspice* v2 [50].
470 Identified clades of interest were further inspected for existing epidemiological links using data
471 collected by the University Hospital.

472 **Identifying genomes belonging to GISAID emerging clade A20/15324T**

473 To identify a possible geographic origin of the synonymous C15324T mutation in *ORF1ab*, we
474 performed a search on all available GISAID genomes as of August 12th, 2020. We downloaded all
475 high quality and complete genomes that were assigned do GISAID legacy clade G (corresponds to
476 clade 20A) and PANGOLIN lineage B.1 (all three are mostly congruent [59]) with a collection date
477 between December 2019 and March 23rd, 2020 (N = 2,856). We used Nextclade version 0.3.5
478 (https://clades.nextstrain.org) to infer genomic mutations and filtered for sequences that contained
479 C15324T. This procedure allowed avoidance of homoplasic mutations at this site. Further, we

480 downloaded metadata for all high quality and complete genomes (as of August 12th, 2020)
481 irrespective of clade to calculate summary statistics of number of genomes sequenced per country.

482 **Identification of S-gene D614G mutation in Basel sequences**

483 We screened the early phase Basel sequences for the mutation at nucleotide position 23,403 based on
484 the alignment to the ***Wuhan-Hu-1*** reference sequence MN908947.3. Viral load (C_t-value) of patients
485 that carried lineages with a mutated S-D614G gene (N = 274) were compared to patients that carried
486 the ancestral allele (N = 12) using a Mann-Whitney U test.

487 **Data accessibility**

488 Sequencing data (viral reads only) was submitted to European Nucleotide Archive (ENA) under
489 accession number PRJEB39887, consensus sequences were submitted to GISAID, bioinformatic
490 pipelines are accessible on Github.

491 **Ethics**

492 The study was conducted according to good laboratory practice and in accordance with the
493 Declaration of Helsinki and national and institutional standards and was approved by the ethical
494 committee (EKNZ 2020-00769). The clinical trial accession number is NCT04351503
495 (clinicaltrials.gov).
496

References

1 Tayoun, A. *et al.* Genomic surveillance and phylogenetic analysis reveal multiple introductions of SARS-CoV-2 into a global travel hub in the Middle East. *BioRxiv* **2020.05.06.080606** (2020).

2 Banu, S. *et al.* A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates. *bioRxiv*, 2020.2005.2031.126136, doi:10.1101/2020.05.31.126136 (2020).

3 Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003.e1009, doi:10.1016/j.cell.2020.04.023 (2020).

4 Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis*, doi:10.1016/s1473-3099(20)30562-4 (2020).

5 Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv*, 2020.2005.2012.20099929, doi:10.1101/2020.05.12.20099929 (2020).

6 Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, eabd2161, doi:10.1126/science.abd2161 (2020).

7 Díez-Fuertes, F. *et al.* Phylodynamics of SARS-CoV-2 transmission in Spain. *bioRxiv*, 2020.2004.2020.050039, doi:10.1101/2020.04.20.050039 (2020).

8 Gámbaro, F. *et al.* Introductions and early spread of SARS-CoV-2 in France. *bioRxiv*, 2020.2004.2024.059576, doi:10.1101/2020.04.24.059576 (2020).

9 Gudbjartsson, D. F. *et al.* Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med* **382**, 2302-2315, doi:10.1056/NEJMoa2006100 (2020).

10 Kumar, P. *et al.* Integrated genomic view of SARS-CoV-2 in India. *bioRxiv*, 2020.2006.2004.128751, doi:10.1101/2020.06.04.128751 (2020).

11 Zehender, G. *et al.* Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J Med Virol*, doi:10.1002/jmv.25794 (2020).

12 team, T. N. *Genomic epidemiology of novel coronavirus - Global subsampling*, <https://nextstrain.org/ncov/global> (2020).

13 Hill, V. & Rambaut, A. *Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06*, <https://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420> (2020).

14 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*, doi:10.1038/s41564-020-0770-5 (2020).

15 Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*, doi:10.1101/2020.06.12.148726 (2020).

16 van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *bioRxiv*, 2020.2005.2021.108506, doi:10.1101/2020.05.21.108506 (2020).

17 *Grenzgänger*, <https://www.statistik.bs.ch/haeufig-gefragt/arbeiten/grenzgaenger.html> (

18 Swissinfo.ch. *Swiss hospitals to take French coronavirus patients*, <https://www.swissinfo.ch/eng/cross-border-care_swiss-hospitals-take-french-coronavirus-patients/45634674> (

19 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

20 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, 30494, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).

21 Bluhm, A. *et al.* SARS-CoV-2 Transmission Chains from Genetic Data: A Danish Case Study. *bioRxiv*, 2020.2005.2029.123612, doi:10.1101/2020.05.29.123612 (2020).

22 Versteeg, B. *et al.* Genomic analyses of the *Chlamydia trachomatis* core genome show an association between chromosomal genome, plasmid type and disease. *BMC genomics* **19**, 130, doi:10.1186/s12864-018-4522-3 (2018).

23 Xing, Y., Li, X., Gao, X. & Dong, Q. MicroGMT: A Mutation Tracker for SARS-CoV-2 and Other Microbial Genome Sequences. *Front Microbiol* **11**, 1502, doi:10.3389/fmicb.2020.01502 (2020).

24 Pybus, O. G. *et al. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages*, <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507> (2020).

25 Felbermayr, G., Hinz, J. & S, C. *Après-ski: The Spread of Coronavirus from Ischgl through Germany*, <https://www.ifw-kiel.de/fileadmin/Dateiverwaltung/IfW-Publications/Gabriel_Felbermayr/Apres-ski__The_Spread_of_Coronavirus_from_Ischgl_through_Germany/coronavirus_from_ischgl.pdf> (2020).

26 Hodcroft, E. B. Preliminary case report on the SARS-CoV-2 cluster in the UK, France, and Spain. *Swiss Med Wkly* **150**, doi:10.4414/smw.2020.20212 (2020).

27 Zeitung, B. *Wir haben in Mulhouse die Epidemie-Phase erreicht*, <https://www.bazonline.ch/basel/region/gottesdienst-in-mulhouse-entwickelt-sich-zu-moeglichem-coronaherd/story/29069253> (

28 Joseph, S. J., Didelot, X., Gandhi, K., Dean, D. & Read, T. D. Interplay of recombination and selection in the genomes of Chlamydia trachomatis. *Biol Direct* **6**, 28, doi:10.1186/1745-6150-6-28 (2011).

29 Goldenberger, D. *et al.* Brief validation of the novel GeneXpert Xpress SARS-CoV-2 PCR assay. *J Virol Methods* **284**, 113925, doi:10.1016/j.jviromet.2020.113925 (2020).

30 Leuzinger, K. *et al.* Epidemiology of SARS-CoV-2 Emergence Amidst Community-Acquired Respiratory Viruses. *J Infect Dis*, doi:10.1093/infdis/jiaa464 (2020).

31 Quick, J. *nCoV-2019 sequencing protocol. protocols.io* <dx.doi.org/10.17504/protocols.io.bdp7i5rn> (2020).

32 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

33 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

34 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

35 Wala, J., Zhang, C. Z., Meyerson, M. & Beroukhim, R. VariantBam: filtering and profiling of next-generational sequencing data using region-specific rules. *Bioinformatics* **32**, 2029-2031, doi:10.1093/bioinformatics/btw111 (2016).

36 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).

37 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

38 Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol* **1418**, 335-351, doi:10.1007/978-1-4939-3578-9_16 (2016).

39 Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**, 2808-2810, doi:10.1093/bioinformatics/btu379 (2014).

40 D., C. & J.R., L. in *Structural Approaches to Sequence Evolution. Biological and Medical Physics, Biomedical Engineering* (eds Bastolla U., Porto M., Roman H.E., & Vendruscolo M.) (Springer, 2007).

41 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, <https://ggplot2-book.org/> (

42 Holland, L. A. *et al.* An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J Virol* **94**, doi:10.1128/jvi.00711-20 (2020).

43 Gouy, M., Guindon, S. & Gascuel, O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* **27**, 221-224, doi:10.1093/molbev/msp259 (2009).

44 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).

45 Chao, A., Chiu, C. H. & Hsieh, T. C. Proposing a resolution to debates on diversity partitioning. *Ecology* **93**, 2037-2051, doi:10.1890/11-1817.1 (2012).

46 Chao, A. & Jost, L. Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution* **6**, 873-882, doi:10.1111/2041-210X.12349 (2015).

47 Chao, A., Wang, Y. & Jost, L. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol Evol* **4**, 1091-1100, doi:doi.org/10.1111/2041-210X.12108 (2013).

48 Nadeau, S. *et al.* Quantifying SARS-CoV-2 spread in Switzerland based on genomic sequencing data. *medRxiv*, 2020.2010.2014.20212621, doi:10.1101/2020.10.14.20212621 (2020).

49 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

50 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

51 Team, R. C. *R: A language and environment for statistical computing*, <https://www.r-project.org/> (

52 Wickham, H. & Henry, L. *tidyr: Tidy Messy Data. R package version 1.1.0.* , <https://CRAN.R-project.org/package=tidyr> (2020).

53 Wickham, H., François, R., Henry, L. & Müller, K. *dplyr: A Grammar of Data Manipulation. R package version 1.0.0.* , <https://CRAN.R-project.org/package=dplyr> (2020).

54 Wickham, H., J, H. & Francois, R. *readr: Read Rectangular Text Data. R package version 1.3.1.* , <https://CRAN.R-project.org/package=readr> (2018).
55 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).
56 De Maio, N., C, W. & N, G. <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/10> (
57 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).
58 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042, doi:10.1093/ve/vex042 (2018).
59 *GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses*, <https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/> (

Acknowledgements

We thank Daniel Gander, Christine Kiessling, Magdalena Schneider, Elisabeth Schultheiss, Clarisse Straub, and Rosa-Maria Vesco (University Hospital Basel) for excellent technical assistance with sequencing. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel, the support from the sciCORE team for the analysis is greatly appreciated. Support for the creation of schematic figures (S2) was provided by BioRender.com. We thank all authors who have shared their genomic data on GISAID especially the Stadler Lab from ETHZ for sharing other Swiss sequences. A full table outlining the originating and submitting labs is included as a supplementary file. No dedicated funding was used for this work.

Authors contributions

AE and HH devised the project. KL and AG collected and prepared samples and associated data. MS performed the phylogenetic analysis and interpretation, and led the writing and revising of the report. AM constructed the COVGAP bioinformatic pipeline and released it on Github. TR and HSS prepared viral RNA for sequencing, directed the phylogenetic analysis and deposited genomic data to GISAID and ENA. MSch and KKS collected clinical and epidemiological data. MyB and RSS provided geographical expertise. JB, STS and SF provided public health and epidemiological expertise. HP, MSi, CHN, RB, MO, SB, and MB provided clinical expertise and valuable discussion on the results.

All authors commented on the draft report and contributed to the final version.

Competing interests statement

The authors declare no competing interests.

649 ## Tables

650 **Table 1. Number and age summary of all tested patients, positively tested patients, and patients with**
651 **successfully sequenced SARS-CoV-2 genomes, by sex.**

		Number	%	Median Age [years]	IQR [years]	< 18 years old
All tests*	Males	3067	44.2	44	31-60	396 (5.7%)
	Females	3867	55.8	42	29-56	
Positive tested*	Males	363	48.7	49	33-61.5	17 (2.3%)
	Females	383	51.3	47	32-60	
In study cohort*	Males	222	48.1	49	34-60	12 (2.6%)
	Females	240	51.9	47	33-60	

652 ** six patients with no information regarding sex*

653

654 **Table 2. Sensitivity, specificity, and accuracy of COVGAP.**

	MOCK POSITIVE	MOCK NEGATIVE
COVGAP POSITIVE	TP=180	FN=11
COVGAP NEGATIVE	FP=0	TN=2541564

655 *TP: true positive, FP: false positive, FN: false negative. Numbers represent cumulative counts of bases that were*
656 *or were not mutated over the 16 test genomes.*

657

658

659

660

661

662

663

664

665 **Table 3. Number of cases harbouring the S-D614G mutation in spike protein encoding gene in each**
666 **phylogenetic lineage (PANGOLIN definition ver. May 19) and total count, in Basel area cohort by March**
667 **23rd 2020.**

Phylogenetic lineage	Number of samples S^{G614} (derived)	Number of samples S^{D614} (ancestral)	Total counts
A.2	0	1	1
A.5	0	1	1
B	2	6	8
B.1	391	0	391
B.1.1	36	0	36
B.1.1.1	1	0	1
B.1.1.10	2	0	2
B.1.1.6	1	0	1
B.1.5	12	0	12
B.1.8	3	0	3
B.10	0	1	1
B.2	0	8	8
B.2.1	0	2	2
B.3	0	1	1
Sum	**448**	**20**	**468**

668