

Cite as: A. Popa *et al.*, *Sci. Transl. Med.*
10.1126/scitranslmed.abe2555 (2020).

CORONAVIRUS

Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2

Alexandra Popa^{1†}, Jakob-Wendelin Genger^{1†}, Michael D. Nicholson^{2†}, Thomas Penz^{1†}, Daniela Schmid^{3†}, Stephan W. Aberle^{4†}, Benedikt Agerer^{1†}, Alexander Lercher^{1†}, Lukas Endler⁵, Henrique Colaço¹, Mark Smyth¹, Michael Schuster¹, Miguel L. Grau⁶, Francisco Martínez-Jiménez⁶, Oriol Pich⁶, Wegene Borena⁷, Erich Pawelka⁸, Zsófia Keszei¹, Martin Senekowitsch¹, Jan Laine¹, Judith H. Aberle⁴, Monika Redlberger-Fritz⁴, Mario Karolyi⁸, Alexander Zoufaly⁸, Sabine Maritschnik³, Martin Borkovec³, Peter Hufnagl³, Manfred Nairz⁹, Günter Weiss⁹, Michael T. Wolfinger^{10,11}, Dorothee von Laer⁷, Giulio Superti-Furga^{1,12}, Nuria Lopez-Bigas^{6,13}, Elisabeth Puchhammer-Stöckl⁴, Franz Allerberger³, Franziska Michor^{2,14}, Christoph Bock^{1,15}, Andreas Bergthaler^{1,*}

¹CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria. ²Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; and Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA. ³Austrian Agency for Health and Food Safety (AGES), 1220 Vienna, Austria. ⁴Center for Virology, Medical University of Vienna, 1090 Vienna, Austria. ⁵Bioinformatics and Biostatistics Platform, Department of Biomedical Sciences, University of Veterinary Medicine, 1210 Vienna, Austria. ⁶Institute for Research in Biomedicine (IRB), 08028 Barcelona, Spain. ⁷Institute of Virology, Medical University Innsbruck, 6020 Innsbruck, Austria. ⁸Department of Medicine IV, Kaiser Franz Josef Hospital, 1100 Vienna, Austria. ⁹Department of Internal Medicine II, Medical University of Innsbruck, 6020 Innsbruck, Austria. ¹⁰Department of Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria. ¹¹Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, 1090 Vienna, Austria. ¹²Center for Physiology and Pharmacology, Medical University of Vienna, 1090 Vienna, Austria. ¹³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ¹⁴The Broad Institute of MIT and Harvard, Cambridge, MA, USA; The Ludwig Center at Harvard, Boston, MA, USA; and the Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA, USA. ¹⁵Department of Laboratory Medicine, Medical University of Vienna, 1090 Vienna, Austria.

[†]Equal contributions

[‡]Equal contributions

*Corresponding author. Email: abergthaler@cemm.oew.ac.at

Superspreading events shaped the Coronavirus Disease 2019 (COVID-19) pandemic, and their rapid identification and containment are essential for disease control. Here we provide a national-scale analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) superspreading during the first wave of infections in Austria, a country that played a major role in initial virus transmissions in Europe. Capitalizing on Austria's well-developed epidemiological surveillance system, we identified major SARS-CoV-2 clusters during the first wave of infections and performed deep whole-genome sequencing of more than 500 virus samples. Phylogenetic-epidemiological analysis enabled the reconstruction of superspreading events and charts a map of tourism-related viral spread originating from Austria in spring 2020. Moreover, we exploited epidemiologically well-defined clusters to quantify SARS-CoV-2 mutational dynamics, including the observation of a low-frequency mutation that progressed to fixation within the infection chain. Time-resolved virus sequencing unveiled viral mutation dynamics within individuals with COVID-19, and epidemiologically validated infector-infectee pairs enabled us to determine an average transmission bottleneck size of 10^3 SARS-CoV-2 particles. In conclusion, this study illustrates the power of combining epidemiological analysis with deep viral genome sequencing to unravel the spread of SARS-CoV-2, and to gain fundamental insights into mutational dynamics and transmission properties.

INTRODUCTION

The SARS-CoV-2 pandemic has already infected more than 20 million people in 188 countries, causing 737,285 deaths globally as of August 11, 2020 and extraordinary disruptions to daily life and national economies (1, 2). The international research community rapidly defined pathophysiological characteristics of the Coronavirus Disease

2019 (COVID-19), established diagnostic tools, assessed immunological responses and identified risk factors for a severe disease course (3–6). Clustered outbreaks and superspreading events of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pose a particular challenge to pandemic control (7–10). However, we still know comparatively little about fundamental properties of SARS-CoV-2 genome

evolution and transmission dynamics within the human population.

Acquired fixed mutations enable phylogenetic analyses and have already led to insights into the origins and routes of SARS-CoV-2 spread (11–14). Conversely, low-frequency mutations and their changes over time within individual patients can provide insights into the dynamics of intra-host evolution. The resulting intra-host viral populations represent groups of variants with different frequencies, whose genetic diversity contributes to fundamental properties of infection and pathogenesis (15, 16).

Austria is located in the center of Europe and has a population of 8.8 million. It operates a highly developed health care system, which includes a national epidemiological surveillance program. As of August 7, 2020, contact tracing had been performed for all 21,821 reported SARS-CoV-2 positive cases. Out of these, 10,385 cases were linked to epidemiological clusters, whereas no infection chains were identified for the remaining cases (17). Linked to Austria's prominent role in international winter tourism, the country emerged as a potential superspreading transmission hub across the European continent in early 2020. During the first phase of the pandemic in Europe (February to May 2020), winter tourism-associated spread of SARS-CoV-2 from Austria may have been responsible for up to half of the imported cases in Denmark and Norway and a considerable share of imported cases in several other countries including Iceland and Germany (11, 18, 19).

In this study, we reconstructed major SARS-CoV-2 infection clusters in Austria and analyzed their role in international virus spread by combining phylogenetic and epidemiological analysis. Moreover, we analyzed our deep viral genome sequencing data from epidemiologically identified transmission chains and family clusters using biomathematical models, in order to infer genetic bottlenecks and the mutation dynamics of SARS-CoV-2 genome evolution. Our results provide fully integrated genetic and epidemiological evidence for continental spread of SARS-CoV-2 from Austria and establish fundamental transmission properties in the human population.

RESULTS

Genomic epidemiology reconstruction of SARS-CoV-2 infection clusters in Austria

We selected and analyzed SARS-CoV-2 virus samples from geographical locations across Austria, with a focus on the provinces of Tyrol and Vienna, given that these two regions were initial drivers of the pandemic in Austria (fig. S1A) (17). We sequenced 572 SARS-CoV-2 RNA samples from 449 unique SARS-CoV-2 cases spanning a time frame between February 24 until May 7. This captured both the onset and the peak of the initial COVID-19 outbreak in Austria (Fig. 1A).

The selected samples covered multiple epidemiological and clinical parameters including age, sex, and viral load (fig. S1B–C). Samples from both swabs (nasal, oropharyngeal) and secretions (tracheal, bronchial) were included (fig. S1D) in order to investigate not only the evolutionary dynamics within the population, but also within individuals.

Out of the 572 samples, 427 passed our sequencing quality controls (>96% genome coverage, >80% aligned viral reads, ≤1,500 uncalled nucleotides in the consensus sequences), and following the removal of cell culture samples, 420 samples were considered for low frequency analysis. Out of the 420 samples, 345 corresponded to unique SARS-CoV-2 cases and were further integrated in our phylogenetic analyses, as they corresponded to unique patient identifiers with complete sample annotation at the time of the analysis (fig. S1E). For these 345 samples, we assembled SARS-CoV-2 genome sequences, constructed phylogenies and identified low-frequency mutations based on high-quality sequencing results with >5 million reads per sample and >80% of mapped viral reads (fig. S2A–B).

To obtain robust quantifications of minor variants in all 420 samples, we validated our sample processing workflow and pipeline with additional experimental controls including synthetic SARS-CoV-2 genome titrations, technical replicates for sample preparation and sequencing runs, and dilution experiments (data file S1). Matched controls were highly consistent with each other, indicative of excellent assay performance and a highly reproducible analysis pipeline (fig. S2C–F). For an alternative allele frequency of 0.01, we obtained an average accuracy of 90.92% (range from 68–97%). In addition, the shared percentage of detected variants between control pairs ranged from 50% to 90.97% for a cutoff of 0.02 of the allele frequencies. The high specificity of detection even at low frequencies, as well as the large overlap of detected variants, supported the choice of a 0.02 frequency cutoff for calling high-confidence variants (data file S1).

To investigate the link between local outbreaks in Austria and the global pandemic, we performed phylogenetic analysis of 345 SARS-CoV-2 genomes from Austrian cases and 7,666 global genomes from the GISAID (Global Initiative on Sharing All Influenza Data) database (data file S2). Similar mutation profiles, together with information of geographical proximity of the samples and time of infection are strong indicators of possible transmission links. Therefore, groups of virus sequences were annotated as phylogenetic clusters when they all shared a homogenous mutation pattern and originated from the same geographical location and time period. Among the distinct phylogenetic clusters identified, six could be linked to specific geographic locations of the probable region of infection (Fig. 1B). Three out of these six clusters comprised samples with a geographical location mainly in the Tyrol region (hereafter named Tyrol-1, Tyrol-2, and Tyrol-

3), whereas the other three originated in Vienna (hereafter named Vienna-1, Vienna-2, and Vienna-3). These clusters are related to the global clades 19A, 20A, 20B and 20C of the widely used Nextstrain classification (fig. S3A).

Independently, contact tracing surveillance assigns SARS-CoV-2 cases to epidemiological clusters based on the identification of transmission lines. In Austria, an extensive centralized tracing program was implemented during the COVID-19 outbreak. This program facilitated grouping of positive cases with a common exposure history and a comparable time frame of infection into epidemiological clusters. Integration of the phylogenetic analysis of Austrian SARS-CoV-2 sequences with epidemiological data resulted in strong overlap of these two lines of evidence, with 199 out of the 345 sequences (65%) assigned to epidemiological clusters (data file S3). All sequenced samples from epidemiological cluster A mapped to the relatively homogenous phylogenetic cluster Vienna-1 (Fig. 1C) with an index patient that had returned from Italy.

Our largest phylogenetic cluster, “Tyrol-1” (fig. S3B), contained samples originating mainly from Austria’s Tyrol region (73/90 samples) and overlapped with epidemiological cluster S (44/53 epidemiologically annotated samples). This phylogenetic cluster included resident and travel-associated cases to the ski-resort Ischgl or the related valley Paznaun (Fig. 1C). Although different SARS-CoV-2 strains circulated in the region of Tyrol, this data suggests that epidemiological cluster S originated from a single strain with a characteristic mutation profile leading to a large outbreak in this region. To elucidate the possible origin of the SARS-CoV-2 strain giving rise to this cluster, we searched for sequences matching the viral mutation profile among global SARS-CoV-2 sequences (Fig. 1D-E). Using phylogenetic analysis, we found that the mutation profile defining the Tyrol-1 cluster matched the definition of the global clade 20C of the Nextstrain classification (fig. S3C). This clade is predominantly populated by strains from North America.

To reveal possible transmission lines specifically between European countries in February and March 2020, we performed phylogenetic analysis using all 7,731 European high-quality SARS-CoV-2 sequences sampled before March 31 that were available in the GISAID database (data file S2). Using this approach, we identified several samples matching the Tyrol-1 cluster mutation profile from a local outbreak in the region Hauts-de-France in the last week of February (20). Introduction of this SARS-CoV-2 strain to Iceland by tourists with a travel history to Austria was reported as early as March 2 (Fig. 1E, fig. S3C) (11), indicating that viruses with this mutational profile were already present in Ischgl in the last week of February. These findings suggest that the emergence of cluster Tyrol-1 coincided with the local outbreak in France and with the early stages of the severe outbreak in

Northern Italy (21). The viral genomes observed in the Tyrol-1 cluster were closely related to those observed among the Icelandic cases with a travel history to Austria (fig. S3D-E) (11). Vice versa, many of the Icelandic strains with a “Tyrol-1” mutation profile had reported an Austrian or Icelandic exposure history (fig. S3F). Together, these observations and epidemiological evidence support the notion that the SARS-CoV-2 outbreak in Austria propagated to Iceland. Moreover, the emergence of these strains coincided with the emergence of the global clade 20C. One week after the occurrence of SARS-CoV-2 strains with this mutation profile in France and Ischgl, an increasing number of related strains based on the same mutation profile could be found across continents (Fig. 1E), for example in New York City (12). As a popular skiing destination attracting thousands of international tourists, Ischgl may have played a critical role as transmission hub for the spread of clade 20C in Europe and beyond (fig. S3G-H) (12). However, due to the lack of global epidemiological surveillance programs, it is rarely possible to infer direct transmission lines between countries.

Our results integrating epidemiological and sequencing data emphasize that phylogenetic analyses of SARS-CoV-2 sequences empower robust tracing from inter-individual to local and international spreading events (12). Of note, both clusters Tyrol-1 and Vienna-1 originated from crowded indoor events (an *Apré Ski* bar and a sports class, respectively), which are now appreciated as high-risk situations for super-spreading events.

Dynamics of low-frequency and fixed mutations in clusters

Next, we sought to uncover the mutational dynamics of SARS-CoV-2 during its transmission through the human population. We investigated the mutation profiles of our samples in terms of both fixed mutations (that drive the phylogenetic analyses) and the pool of low frequency variants of each one of our samples. More than half of the fixed mutations in the Austrian SARS-CoV-2 genomes were non-synonymous (Fig. 2A), most frequently occurring in non-structural protein 6 (*nsp6*), open reading frame 3a (*ORF3a*), and *ORF8* (Fig. 2B-C). An analysis of mutational signatures in the 7,666 global strains and the Austrian subset of SARS-CoV-2 isolates showed a heterogeneous mutational pattern dominated by C>U, G>U and G>A substitutions (Fig. 2D).

We assessed the pool of variants for both low-frequency and fixed mutations (Fig. 3A-B) and observed similar mutation patterns among these two sets of variants, which supports the accuracy of low-frequency mutation calling (Fig. 3C-D). However, this pattern was lost for variants with an alternative frequency less than 0.01, which appear prone to false-positive variant calls. These results suggest that the same biological and evolutionary forces are at work for low-frequency and fixed mutations. Although the functional impact

of variants across the genomes will need further research, we found that regions such as the 5' UTR which contains multiple stable RNA secondary structures were subject to an increased mutation rate (Fig. 3D). Variants in the 5' UTR region are mainly localized along the stem-loop secondary structures (Fig. 3E). We found that 31% of the positions in the genome (9,391 total positions) harbored variants (alternative allele frequency ≥ 0.02) among the 420 sequenced strains from Austria and identified mutational hotspots for both high (≥ 0.5) and low-frequency (< 0.5) mutations (Fig. 4A). Among these, 9,034 positions exhibited only low-frequency mutations (< 0.50), whereas four positions (241, 3037, 14408, 23403) demonstrated fixation of the alternative allele in more than 50% of samples. We also identified 31 positions with alternative alleles being fixed in more than three samples and exhibiting a frequency < 0.5 in at least two other samples (for example 15380, 20457).

Based on our phylogenetic analysis, we identified a subcluster inside the phylogenetic Tyrol-1 cluster that was defined by a fixed non-synonymous G>U mutation at position 15,380 (Fig. 4B). This mutation was absent from all other Austrian cases but was detected at low and intermediary frequencies in other cases of the Tyrol-1 cluster. Around the time of emergence of this mutation, sequences sharing the same mutational profile (Tyrol-1 haplotype and G>U at position 15,380) appeared in other European countries including Denmark and Germany (Fig. 4C). Similarly, a synonymous fixed C>U mutation at position 20,457 defined a subcluster inside the phylogenetic Vienna-1 cluster (Fig. 4D). The cases from this subcluster intersected with members of two families (families 1 and 7) (Fig. 4E). Four members of family 1 tested positive for SARS-CoV-2 on March 8 and were epidemiologically assigned to cluster A. Yet, their viral sequences exhibited a wide range of C>U mutation frequencies at position 20,457 (0.00, 0.036, 0.24, and 1.00 respectively) (Fig. 4D-E). Conversely, four members of family 7, who tested positive for SARS-CoV-2 between March 16 and 22, were epidemiologically assigned to cluster AL and harbored viral genomes with a fixed U nucleotide at position 20,457 (Fig. 4D-E).

Through several telephone interviews, we followed up with the members of both families to reconstruct the timeline of the infection events (data file S4). Both grandparents of family 1 were exposed to infected case CeMM1056 (node N13; sampling date March 3) during a recreational indoor event on February 28 and subsequently tested positive for SARS-CoV-2 (Fig. 4D-E, Fig. 5A). The woman, CeMM0176 (node N16; sampling date March 8), did not present a mutation at position 20,457, whereas her husband, CeMM1057 (node N15; sampling date March 6), had the U allele at this position with a frequency of 0.036. The chain of transmission continued in family 1 with the infection of the couple CeMM0175 (node N18; sampling date March 8) and CeMM0177 (node N17;

sampling date March 8), who had the U mutation at frequencies of 0.25 and 1, respectively. All further transmissions from CeMM1057 (node N15) resulted in a fixed mutation at position 20,457. CeMM1058 (node N25; sampling date March 8) was in contact with CeMM1057 on March 2 and attended a funeral on March 5 with CeMM1059 (node N27; sampling date March 11). On March 8 multiple persons participated at a birthday party, which included case CeMM1059 together with CeMM1062 (node N29; sampling date March 13). Case CeMM1062 was part of a choir with multiple members of family 7 (CeMM0218 [node N31], CeMM0219 [node N32] and CeMM0217 [node N33]) on March 10 (Fig. 4D-E, Fig. 5A). Given our phylogenetic analysis and epidemiological reconstruction of transmission chains, we thus provide strong evidence for the emergence of a fixed mutation within a family and its spreading across previously disconnected epidemiological clusters. Together, these results from two super-spreading events (Tyrol-1, Vienna-1) demonstrate the power of deep viral genome sequencing in combination with detailed epidemiological data for observing viral mutation on their way from emergence at low frequency to fixation.

Impact of transmission bottlenecks and intra-host evolution on SARS-CoV-2 mutational dynamics

The emergence and potential fixation of mutations in the viral populations within a patient depend on inter-host bottlenecks and intra-host evolutionary dynamics (22, 23). An examination of the individual contributions of these forces requires pairs of samples from validated transmission events. For this purpose, we combined intra-family cases, known epidemiological transmission chains, and subsequent telephone investigations to track the index cases as well as the context, date and nature of each transmission event (Fig. 5A, data file S4) (22, 24). Our set of SARS-CoV-2 positive cases comprised thirty-nine epidemiologically confirmed infector-infectee pairs (Fig. 5A, fig. S4A, data file S4).

One particularly well-defined network of SARS-CoV-2 transmission events linked cases from epidemiological cluster A and AL (Fig. 4E, Fig. 5A). The index case of cluster A is CeMM0003 (node N1) who contracted the virus during a visit to the north of Italy, further infecting his family members, and later, case CeMM0146 (node N3) during a dinner meeting (17). Multiple infections were linked to case CeMM0146 through an indoor sports activity. Among these cases was CeMM1056 (node N13), who further transmitted the virus to case CeMM1057 (node N15) as previously described for the 20,457 mutation linking cluster A and AL (Fig. 4E) (17). Based on these data, we investigated the transmission dynamics between known pairs of donors and recipients by inferring the number of virions initiating the infection, also known as the genetic bottleneck size (22, 24). The quality of the samples and the underlying low-frequency variants are critical for computing robust bottleneck sizes. In our data, samples with

low Ct values (≤ 28) resulted in the detection of 38.6 variants (cutoff of 0.02) on average. Samples with high Ct values (> 28) had on average 109.1 variants. The samples in the transmission chain were of high quality, with an average Ct value of 22.2, and only 9 out of the 43 samples were higher than 28 (fig. S4A, data file S4).

Bottleneck size estimates were calculated by comparing the frequency of detected variants in each transmission pair (fig. S4B-E). In particular, we computed bottleneck size using the beta-binomial method (24), and on three sets of alternative frequency cutoffs: [0.01, 0.95], [0.02, 0.95], and [0.03, 0.95] (fig. S4F, data file S4). Although the absolute values of the estimates were influenced by these cutoffs, their underlying average bottleneck sizes were comparable: 1227.59 (25 and 75% quartile: 21 - 2053.5; standard deviation (SD): 1692.235), 1110.513 (25 and 75% quartile: 2.5 - 2115; SD: 1661.183), and 1319.41 (25 and 75% quartile: 3.5-1763; SD: 1685.378) for 0.01, 0.02, and 0.03 cutoffs, respectively (Fig. 5B, fig. S4G). In conclusion, taking advantage of a well-described and independently confirmed transmission network with thirty-nine transmission events, we found that the number of viral particles transmitted from one individual to another that contributed productively to the infection was on average higher than 1000.

Last, we investigated the dynamics of intra-host evolution by using time-resolved viral sequences from thirty-one longitudinally sampled patients. These patients were subject to different medical treatments and five of them succumbed to COVID-19-related complications (data file S5). To analyze intra-host viral dynamics, we focused on variants observed in at least two samples from the same patient. This approach resulted in a pool of high-confidence mutations (> 0.02) with high coverage across same-patient samples (mean = 42,099 reads) (fig. S5A). Same-patient samples shared more variants than unrelated sample pairs (defined as non-same-patient, nor from the transmission chains) (fig. S5B). In addition, variants shared between samples from the same patient were unlikely to be found in unrelated samples (fig. S5C).

We observed diverse mutation patterns across individual patients and over time. Most patient samples showed a small number of stable low-frequency mutations (≥ 0.02 and ≤ 0.50), whereas cases CeMM0108, CeMM0172, CeMM0251, CeMM0269, CeMM0299, and CeMM0221 exhibited higher variability, including the fixation and loss of individual mutations (Fig. 5C, fig. S5D). The patient-specific dynamics of viral mutation frequencies may reflect the effect of host-intrinsic factors such as immune responses or the patients' overall health, and extrinsic factors such as different treatment protocols. We also examined the genetic distance between samples obtained across donor–recipient pairs and serially acquired patient samples. However, the difference between increased genetic divergence of the virus within

individual patients over the course of infection compared to inter-host transmission was not significant (p-value: 0.075) (Fig. 5D).

DISCUSSION

Unprecedented global research efforts are underway to counter the COVID-19 pandemic around the globe and its pervasive impact on health and socioeconomics. These efforts include the genetic characterization of SARS-CoV-2 to track viral spread and to investigate the viral genome as it undergoes changes in the human population. Here, we leveraged deep viral genome sequencing in combination with national-scale epidemiological workup to reconstruct Austrian SARS-CoV-2 clusters that played a substantial role in the international spread of the virus. Our study describes how emerging low-frequency mutations of SARS-CoV-2 became fixed in local clusters, followed by viral spread across countries, thus connecting viral mutational dynamics within individuals and across populations. Exploiting our well-defined epidemiological clusters, we determined the inter-human genetic bottleneck size for SARS-CoV-2 – which is the number of virions that start the infection and produce progeny in the viral population – at around 10^1 - 10^3 . Our estimated bottlenecks are based on a substantial number of defined donor-recipient pairs and in agreement with recent studies implying larger bottleneck sizes for SARS-CoV-2 compared to estimates for the influenza A virus (22, 25–28). These bottleneck sizes correlated inversely with higher mutation rates of influenza virus as compared to SARS-CoV-2.

In agreement with our experimentally determined bottleneck sizes, a recent preprint describing a dose-response modeling study estimated 3×10^2 - 2×10^3 SARS-CoV-2 virions necessary to initiate an infection (29). The dynamics of super-spreading events seem to be driven by the number of inter-individual contacts and the quantity of transmitted virus over time (29). Accordingly, our relatively large observed bottleneck size could be the result of patient exposure to high virus accumulations in shared and closed space and may have been influenced by a lack of protective measures in the early phase of the first COVID-19 wave in spring 2020. Although we inferred an average bottleneck size of 10^3 viral particles on average, the broad range of these values indicates that lower numbers of transmitted particles may also lead to a successful infection.

Our sequencing approach resulted in high-confidence variant calling and robust genome-wide coverage, hence it is unlikely that technical limitations constituted a major source of bias. However, estimates of viral bottleneck sizes are likely influenced by many parameters not covered in this study, including virus-specific differences and stochastic evolutionary processes (28). Successful viral transmission also depends on other factors including the rate of decay of viral particles,

frequency of susceptible cells, the host immune response, and co-morbidities (22, 30). The cases we analyzed were subject to different clinical contexts and treatments as well as disease outcomes. To better understand the mechanisms at work during infection, future investigations will need to probe these factors in the context of viral intra-host diversity across body compartments and time (31–34).

This study underscores the value of combining epidemiological approaches with virus genome sequencing to provide critical information to help public health experts to track pathogen spread. Our genomic epidemiology analysis enabled the retrospective identification of SARS-CoV-2 chains of transmission and international hotspots such as the phylogenetic cluster Tyrol-1 (14, 35–37). We also found that the Tyrol clusters were heterogeneous in regards to the S protein D614G mutation, which has been reported to contribute to viral transmissibility and fitness (38–41). Moreover, our phylogenetic analysis of the Vienna-1 cluster demonstrated the practical utility of viral genome sequencing data for uncovering previously unknown links between epidemiological clusters. This result was subsequently confirmed by follow-up contact tracing. We presented this case as an example of how the integration of contact tracing and sequencing information supports tracking the emergence and development of clusters. This demonstrates that deep viral genome sequencing can contribute directly to public health efforts by enhancing epidemiological surveillance.

Since the onset of the SARS-CoV-2 outbreak, many pandemic containment strategies have been implemented across the world. Where effective, these measures led to the reduction in the number of positive cases and limited superspreading events such as those investigated in this study. We found that most of the investigated infections likely involved the effective transmission of at least 1000 viral particles between individuals, suggesting that social distancing and mask wearing may be effective even when they cannot prevent the spread of all viral particles. As a future perspective, our study supports the relevance of investigating viral genome evolution of SARS-CoV-2 in order to enable informed decision-making by public health authorities (42).

MATERIALS AND METHODS

Study design

The goal of this study was to analyze mutational patterns in the SARS-CoV-2 genome to infer transmission in the human population from inter-individual to global scale. For this purpose, isolated viral RNA from 572 Austrian samples (February to May 2020) was processed for genome consensus sequence reconstruction and variant calling as approved by the ethics committee of the Medical University of Vienna. Additional analyses on subsets of samples consisted of the profiling of the mutational patterns across the genome and

bottleneck size estimates based on transmission pairs. Data presented in this study are based on epidemiological and contact tracing data from the Austrian Department of Infection Epidemiology & Surveillance at AGES.

Sample collection and processing

Patient samples were obtained from the Medical Universities of Vienna Institute of Virology, Medical University of Innsbruck Institute of Virology, Medical University of Innsbruck Department of Internal Medicine II, Central Institute for Medical-Chemical Laboratory Diagnostics Innsbruck, Klinikum Wels-Grieskirchen and the Austrian Agency for Health and Food Safety (AGES). Samples were obtained from suspected or confirmed SARS-CoV-2 cases or contact persons of these. Sample types included oropharyngeal swabs, nasopharyngeal swabs, tracheal secretion, bronchial secretion, serum, plasma and cell culture supernatants. RNA was extracted using the following commercially available kits by adhering to the manufacturer's instructions: MagMax (Thermo Fischer), EasyMag (bioMérieux), AltoStar Purification Kit 1.5 (Altona Diagnostics), MagNA Pure LC 2.0 (Roche), MagNA Pure Compact (Roche) and QIAlymphony (Qiagen). Viral RNA was reverse-transcribed with Superscript IV Reverse Transcriptase (ThermoFisher). The resulting cDNA was used to amplify viral sequences with modified primer pools from the Artic Network Initiative (43). PCR reactions were pooled and subjected to high-throughput sequencing.

Sample sequencing

Amplicons were cleaned up with AMPure XP beads (Beckman Coulter) with a 1:1 ratio. Amplicon concentrations were quantified with the Qubit Fluorometric Quantitation system (Life Technologies) and the size distribution was assessed using the 2100 Bioanalyzer system (Agilent). Amplicon concentrations were normalized, and sequencing libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to manufacturer's instructions. Library concentrations again were quantified with the Qubit Fluorometric Quantitation system (Life Technologies) and the size distribution was assessed using the 2100 Bioanalyzer system (Agilent). For sequencing, samples were pooled into equimolar amounts. Amplicon libraries were sequenced on the NovaSeq 6000 platform (Illumina) using SP flowcell with a read length of 2x250-base-pairs in paired-end mode.

Sequencing data processing and analysis

Following demultiplexing, fastq files containing the raw reads were inspected for quality criteria (base quality, N and GC content, sequence duplication, over-represented sequences) using FASTQC (v. 0.11.8) (44). Trimming of adapter sequences was performed with BBDUK from the BBtools suite (<https://jgi.doe.gov/data-and-tools/bbtools>). Overlapping read sequences within a pair were corrected for using

BBMERGE function from BBTools. Read pairs were mapped on the combined Hg38 and SARS-CoV-2 genome (GenBank: MN908947.3, RefSeq: NC_045512.2) using the BWA-MEM software package with a minimal seed length of 17 (v 0.7.17) (45). BWA-MEM accounts for mismatches, insertions and deletions in the alignment score and the mapping quality. Only reads mapping uniquely to the SARS-CoV-2 viral genome were retained. Primer sequences were removed after mapping by masking with iVar (46). From the viral reads BAM (binary alignment map) file, the consensus FASTA file was generated using Samtools (v 1.9) (47), mpileup, Bcftools (v 1.9) (47), and SEQTK (<https://github.com/lh3/seqtk>). For calling low-frequency variants the viral read alignment file was realigned using the Viterbi method provided by LoFreq (v 2.1.2) (48). After adding InDel qualities low frequency variants were called using LoFreq. Variant filtering was performed with LoFreq and Bcftools (v 1.9) (49). Only variants with a minimum coverage of 75 reads, a minimum phred-value of 90 and indels (insertions and deletions) with a HRUN of minimum 4 were considered. All analyses except for the control analysis in Fig. 3C were performed on variants with a minimum alternative frequency of 0.01. The cutoff for the alternative frequency mainly used in this study was set to 0.02, except for Fig. 5B. Annotations of the variants were performed with SnpEff (v 4.3) (50) and SnpSift (v 4.3) (51).

Epidemiological analyses and identification of SARS-CoV-2 infection clusters

The investigation of transmission chains (contact tracing) was conducted by the Department of Infection Epidemiology & Surveillance at the AGES. Epidemiological clusters were defined as accumulations of cases within a certain time-period in a defined region and with common source of exposure. The required information for cluster annotation and resolution in chains of transmission was collected during the official case-contact tracing by the public health authorities, resulting in identification of the most likely source cases and successive cases of the index cases. Contact tracing was performed according to technical guidance relating to this measure produced by the European Centre for Disease Prevention and Control (ECDC) (52). For refinement and validation of contact tracing data for cluster A and cluster AL, we contacted 17 cases for 15-min interviews. The interviews comprised 10 questions concerning the most likely source, time, place, and setting of transmission, contact persons, and the course of disease (start and end of symptoms, kind of symptoms, severity, and hospitalization).

Phylogenetic analysis and inference of transmission lines

Phylogenetic analysis was conducted using the Augur package (version 7.0.2) (53). We compiled a randomly subsampled dataset of 7666 full length viral genomes with high coverage (<1% Ns) that were available from GISAID

(<https://www.gisaid.org/>, June 2) and the 345 sequences obtained in this publication. GISAID sequences were filtered for entries from human hosts with complete sampling dates. Metadata information for patient age and sex were excluded from the analysis. Multiple sequence alignments (MSA) were performed using mafft (54). A masking scheme for homoplasic and highly ambiguous sites was applied to avoid bias in the following phylogenetic analysis as discussed elsewhere (55). We reconstructed the phylogeny with the augur pipeline using IQ-Tree (54) and further processed the resulting trees with treetime to infer ancestral traits of the nodes (56). Phylogenetic trees were rooted with the genome of “Wuhan-Hu-1/2019”. The same workflow was repeated for phylogenetic reconstruction of all high-quality European strains before March 31, 2020 available in the GISAID database by June 7, 2020 (7,731). Clade annotations for global trees were adapted from nextstrain.org

(<https://github.com/nextstrain/ncov/blob/master/defaults/clades.tsv>; <https://clades.nextstrain.org/>), clusters of Austrian strains were identified based on shared mutation profiles and patient location from epidemiological data.

Bottleneck estimation

Our analysis to estimate the transmission bottleneck sizes for each donor-recipient pair was based on the beta-binomial method presented in (24). For a given variant present in the donor, this method assumes that the number of transmitted virions carrying the variant is binomially distributed with the bottleneck size as the number of trials and success probability as the variant frequency in the donor. Following transmission, the viral population during early infection is modeled as a linear birth-death process, implying that the proportion of the viral population descended from any virion in the bottleneck population is beta-distributed. Using this model for the change in variant frequencies between donor and recipient pairs and assuming independence of mutations leads to the likelihood model of (24). Maximum likelihood analysis then provides the bottleneck statistics. Error bars denote 95% confidence intervals, determined by a likelihood ratio test. This method was applied to variants in the following frequency ranges [0.01,0.95], [0.02,0.95], [0.03,0.95]. Due to the high sequencing depth of our study we used the approximate version of the beta-binomial method.

Intra-patient time series analyses

Among our 420 high-quality SARS-CoV-2-positive samples, we had 31 unique cases with multiple time-point samplings (a total of 106 samples). 19/31 cases had only two samples per patient. For each of the 31 cases, we only considered variants with an alternative frequency greater than 0.02 and that were shared across at least two of the intra-patient samples. We retrieved the depth of coverage of the selected variants for each sample for each patient. To compare how many variants were shared intra-patient as opposed to

unrelated samples, we first identified potentially unrelated cases by eliminating all samples from the same patient, as well as all the samples in the transmission chains in Fig. 5A, resulting in 281 samples hereafter termed “unrelated”. We then enumerated all 39,340 unordered pairs of the 281 unrelated samples. Only variants between [0.02, 0.5] were considered. We computed the percentage of variants shared by each pair out of the total number detected across the two samples. We then compared the percentage of variant sharing between intra-patient and unrelated pairs of samples with a Wilcoxon test. To test how widely the intra-patient variants ([0.02, 0.5]; 173 positions) were detected in other samples, we examined how often they were detected in the pool of 218 unrelated samples.

Genetic distance

For shared mutations with defined donor-to-recipient transmission, we determined those mutations present in both samples and calculated their absolute difference in frequency. Similarly, we performed the same computations between time consecutive pairs for serially sampled patients. If multiple samples were obtained on the same day, the sample with lowest Ct value was considered. Note that the time-consecutive pairs had a differing number of days between samples. To these genetic distances obtained from the shared variants we added the sum of the frequencies of the variants detected in only one of the pairs of shared samples; that is, we calculated the l1-norm of the variant frequencies. Statistical difference between the genetic distances from transmission pairs versus consecutive pairs from serially sampled patients, was determined by a Wilcoxon (one-sided) rank-sum test.

Statistical methods

Control samples were compared with a linear regression method and the corresponding R^2 reported. For mutational patterns analyses, a statistical test was devised to compare the deviation of the observed number of mutations from the expected distribution as detailed in the Materials and Methods. The frequency of mutations in overlapping windows across the genome was statistically assessed with a log-likelihood test. For bottleneck size computations a maximum likelihood approach was applied. The comparison of genetic diversity between groups was performed with a standard Wilcoxon test. Significance was inferred for p-values ≤ 0.05 .

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/scitranslmed.abe2555/DC1

Fig. S1: Data overview.

Fig. S2: Technical pipeline and controls.

Fig. S3: Phylogenetic analysis of SARS-CoV-2 sequences from Austrian COVID-19 patients in global context.

Fig. S4: Bottleneck size estimations.

Fig. S5: Viral intra-host diversity in individual patients.

Data file S1: Sample and sequencing information of the 572 samples and the controls.

Data file S2: Acknowledgements for SARS-CoV-2 genome sequences derived from

GISAID.

Data file S3: Epidemiological clusters referred to in this study.

Data file S4: Transmission chain and sample information for ClusterA/ClusterAL and family-related cases.

Data file S5: Clinical information of patients with COVID-19 relating to Fig 5 and fig S5.

REFERENCES AND NOTES

1. P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L. Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shen, X. Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G. F. Xiao, Z. L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020). [doi:10.1038/s41586-020-2012-7](https://doi.org/10.1038/s41586-020-2012-7) [Medline](#)
2. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020). [doi:10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1) [Medline](#)
3. N. Vabret, G. J. Britton, C. Gruber, S. Hegde, J. Kim, M. Kuksin, R. Levantovsky, L. Malle, A. Moreira, M. D. Park, L. Pia, E. Risson, M. Saffern, B. Salomé, M. Esai Selvan, M. P. Spindler, J. Tan, V. van der Heide, J. K. Gregory, K. Alexandropoulos, N. Bhardwaj, B. D. Brown, B. Greenbaum, Z. H. Gümüş, D. Homann, A. Horowitz, A. O. Kamphorst, M. A. Curotto de Lafaille, S. Mehndru, M. Merad, R. M. Samstein, M. Agrawal, M. Aleynick, M. Belabed, M. Brown, M. Casanova-Acebes, J. Catalan, M. Centa, A. Charap, A. Chan, S. T. Chen, J. Chung, C. C. Bozkus, E. Cody, F. Cossarini, E. Dalla, N. Fernandez, J. Grout, D. F. Ruan, P. Hamon, E. Humblin, D. Jha, J. Kodysh, A. Leader, M. Lin, K. Lindblad, D. Lozano-Ojalvo, G. Lubitz, A. Magen, Z. Mahmood, G. Martinez-Delgado, J. Mateus-Tique, E. Meritt, C. Moon, J. Noel, T. O'Donnell, M. Ota, T. Plitt, V. Pothula, J. Redes, I. Reyes Torres, M. Roberto, A. R. Sanchez-Paulete, J. Shang, A. S. Schanoski, M. Suprun, M. Tran, N. Vaninov, C. M. Wilk, J. Aguirre-Ghisso, D. Bogunovic, J. Cho, J. Faith, E. Grasset, P. Heeger, E. Kenigsberg, F. Krammer, U. Laserson; Sinai Immunology Review Project. Immunology of COVID-19: Current State of the Science. *Immunity* **52**, 910–941 (2020). [doi:10.1016/j.immuni.2020.05.002](https://doi.org/10.1016/j.immuni.2020.05.002) [Medline](#)
4. D. Mathew, J. R. Giles, A. E. Baxter, D. A. Oldridge, A. R. Greenplate, J. E. Wu, C. Alanio, L. Kuri-Cervantes, M. B. Pampena, K. D'Andrea, S. Manne, Z. Chen, Y. J. Huang, J. P. Reilly, A. R. Weisman, C. A. G. Ittner, O. Kuthuru, J. Dougherty, K. Nzingha, N. Han, J. Kim, A. Pattekar, E. C. Goodwin, E. M. Anderson, M. E. Weirick, S. Gouma, C. P. Arevalo, M. J. Bolton, F. Chen, S. F. Lacey, H. Ramage, S. Cherry, S. E. Hensley, S. A. Apostolidis, A. C. Huang, L. A. Vella, M. R. Betts, N. J. Meyer, E. J. Wherry; UPenn COVID Processing Unit, Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**, eabc8511 (2020). [doi:10.1126/science.abc8511](https://doi.org/10.1126/science.abc8511) [Medline](#)
5. X. Zhang, Y. Tan, Y. Ling, G. Lu, F. Liu, Z. Yi, X. Jia, M. Wu, B. Shi, S. Xu, J. Chen, W. Wang, B. Chen, L. Jiang, S. Yu, J. Lu, J. Wang, M. Xu, Z. Yuan, Q. Zhang, X. Zhang, G. Zhao, S. Wang, S. Chen, H. Lu, Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**, 437–440 (2020). [doi:10.1038/s41586-020-2355-0](https://doi.org/10.1038/s41586-020-2355-0) [Medline](#)
6. D. Ellinghaus, F. Degenhardt, L. Bujanda, M. Buti, A. Albillos, P. Invernizzi, J. Fernández, D. Prati, G. Baselli, R. Asselta, M. M. Grimsrud, C. Milani, F. Aziz, J. Kässens, S. May, M. Wendorff, L. Wienbrandt, F. Uellendahl-Werth, T. Zheng, X. Yi, R. de Pablo, A. G. Chercoles, A. Palom, A.-E. Garcia-Fernandez, F. Rodriguez-Frias, A. Zanella, A. Bandera, A. Protti, A. Aghemo, A. Lleo, A. Biondi, A. Caballero-Garralda, A. Gori, A. Tanck, A. Carreras Nolla, A. Latiano, A. L. Fracanzani, A. Peschuck, A. Julià, A. Pesenti, A. Voza, D. Jiménez, B. Mateos, B. Nafria Jimenez, C. Quereda, C. Paccapelo, C. Gassner, C. Angelini, C. Cea, A. Solier, D. Pestaña, E. Muñoz-Diaz, E. Sandoval, E. M. Paraboschi, E. Navas, F. García Sánchez, F. Ceriotti, F. Martinelli-Boneschi, F. Peyvand, F. Blasi, L. Téllez, A. Blanco-Grau, G. Hemmrich-Stanisak, G. Grasselli, G. Costantino, G. Cardamone, G. Foti, S. Aneli, H. Kurihara, H. ElAbd, I. My, I. Galván-Femenia, J. Martín, J. Erdmann, J. Ferrusquía-Acosta, K. Garcia-Etxebarria, L. Izquierdo-Sanchez, L. R. Bettini, L. Sumoy, L. Terranova, L. Moreira, L. Santoro, L. Scudeller, F. Mesonero, L. Roade, M. C. Rühlemann, M. Schaefer, M. Carraaba, M. Riveiro-Barciela, M. E. Figueroa Basso, M. G. Valsecchi, M. Hernandez-Tejero, M. Acosta-Herrera, M. D'Angiò, M. Baldini, M. Cazzaniga, M. Schulzky, M. Cecconi, M. Wittig, M. Ciccirelli, M. Rodríguez-Gandía, M. Boccione, M. Miozzo, N. Montano, N. Braun, N. Sacchi, N. Martínez, O. Özer, O. Palmieri, P. Faverio, P. Preatoni, P. Bonfanti, P. Omodei, P. Tentorio, P. Castro, P. M. Rodrigues, A. Blandino Ortiz, R. de Cid, R. Ferrer, R.

- Gualtierotti, R. Nieto, S. Goerg, S. Badalamenti, S. Marsal, G. Matullo, S. Pelusi, S. Juzenas, S. Aliberti, V. Monzani, V. Moreno, T. Wesse, T. L. Lenz, T. Pumarola, V. Rimoldi, S. Bosari, W. Albrecht, W. Peter, M. Romero-Gómez, M. D'Amato, S. Duga, J. M. Banales, J. R. Hov, T. Folseraas, L. Valenti, A. Franke, T. H. Karlsen; Severe Covid-19 GWAS Group, Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020). [Medline](#)
7. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005). [doi:10.1038/nature04153](#) [Medline](#)
8. T. M. McMichael, D. W. Currie, S. Clark, S. Pogosjans, M. Kay, N. G. Schwartz, J. Lewis, A. Baer, V. Kawakami, M. D. Lukoff, J. Ferro, C. Brostrom-Smith, T. D. Rea, M. R. Sayre, F. X. Riedo, D. Russell, B. Hiatt, P. Montgomery, A. K. Rao, E. J. Chow, F. Tobolowsky, M. J. Hughes, A. C. Bardossy, L. P. Oakley, J. R. Jacobs, N. D. Stone, S. C. Reddy, J. A. Jernigan, M. A. Honein, T. A. Clark, J. S. Duchin; Public Health–Seattle and King County, EvergreenHealth, and CDC COVID-19 Investigation Team, Epidemiology of covid-19 in a long-term care facility in King County, Washington. *N. Engl. J. Med.* **382**, 2005–2011 (2020). [doi:10.1056/NEJMoa2005412](#) [Medline](#)
9. D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, Clinical Characteristics of 138 Hospitalized Patients with, 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020). [doi:10.1001/jama.2020.1585](#) [Medline](#)
10. L. Hamner, P. Dubbel, I. Capron, A. Ross, A. Jordan, J. Lee, J. Lynn, A. Ball, S. Narwal, S. Russell, D. Patrick, H. Leibrand, High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice - Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 606–610 (2020). [doi:10.15585/mmwr.mm6919e6](#) [Medline](#)
11. D. F. Gudbjartsson, A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, B. Eiríksdottir, R. Fríðriksdottir, E. E. Gardarsdottir, G. Georgsson, O. S. Gretarsdottir, K. R. Gudmundsson, T. R. Gunnarsdottir, A. Gylfason, H. Holm, B. O. Jónsson, A. Jonasdottir, F. Jónsson, K. S. Jónsdóttir, T. Kristjánsson, D. N. Magnúsdóttir, L. le Roux, G. Sigmundsdóttir, G. Sveinbjörnsson, K. E. Sveinsdóttir, M. Sveinsdóttir, E. A. Thorarensen, B. Thorbjörnsson, A. Löve, G. Masson, I. Jónsdóttir, A. D. Möller, T. Gudnason, K. G. Kristinsson, U. Thorsteinsdóttir, K. Stefánsson, Spread of SARS-CoV-2 in the Icelandic Population. *N. Engl. J. Med.* **382**, 2302–2315 (2020). [doi:10.1056/NEJMoa2006100](#) [Medline](#)
12. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksha, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, H. van Bakel, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020). [doi:10.1126/science.abc1917](#) [Medline](#)
13. R. Pung, C. J. Chiew, B. E. Young, S. Chin, M. I. C. Chen, H. E. Clapham, A. R. Cook, S. Maurer-Stroh, M. P. H. S. Toh, C. Poh, M. Low, J. Lum, V. T. J. Koh, T. M. Mak, L. Cui, R. V. T. P. Lin, D. Heng, Y. S. Leo, D. C. Lye, V. J. M. Lee; Singapore 2019 Novel Coronavirus Outbreak Research Team, Investigation of three clusters of COVID-19 in Singapore: Implications for surveillance and response measures. *Lancet* **395**, 1039–1046 (2020). [doi:10.1016/S0140-6736\(20\)30528-6](#) [Medline](#)
14. X. Deng, W. Gu, S. Federman, L. du Plessis, O. G. Pybus, N. R. Faria, C. Wang, G. Yu, B. Bushnell, C.-Y. Pan, H. Guevara, A. Sotomayor-Gonzalez, K. Zorn, A. Gopez, V. Servellita, E. Hsu, S. Miller, T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, H. Y. Chu, J. Shendure, K. R. Jerome, C. Anderson, K. Gangavarapu, M. Zeller, E. Spencer, K. G. Andersen, D. MacCannell, C. R. Paden, Y. Li, J. Zhang, S. Tong, G. Armstrong, S. Morrow, M. Willis, B. T. Matyas, S. Mase, O. Kasiye, M. Park, G. Masinde, C. Chan, A. T. Yu, S. J. Chai, E. Villarino, B. Bonin, D. A. Wadford, C. Y. Chiu, Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020). [doi:10.1126/science.abb9263](#) [Medline](#)
15. M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, R. Andino, Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006). [doi:10.1038/nature04388](#) [Medline](#)
16. R. Andino, E. Domingo, Viral quasispecies. *Virology* **479–480**, 46–51 (2015). [doi:10.1016/j.virol.2015.03.022](#) [Medline](#)
17. P. Kreidl, D. Schmid, S. Maritschnik, L. Richter, W. Borena, J.-W. Genger, A. Popa, T. Penz, C. Bock, A. Bergthaler, F. Allerberger, Emergence of coronavirus disease 2019 (COVID-19) in Austria. *Wien. Klin. Wochenschr.* (2020). [doi:10.1007/s00508-020-01723-9](#) [Medline](#)
18. A. Bluhm, E. Al, M. Christandl, F. Gesmundo, F. R. Klausen, L. Mančinska, V. Steffan, D. S. França, A. H. Werner, SARS-CoV-2 Transmission Chains from Genetic Data: A Danish Case Study. *bioRxiv* (2020). [doi:10.1101/2020.05.29.123612](#)
19. C. L. Correa-Martínez, S. Kampmeier, P. Kumpers, V. Schwierzeck, M. Hennies, W. Hafezi, J. Kühn, H. Pavenstädt, S. Ludwig, A. Mellmann, A Pandemic in Times of Global Tourism: Superspreading and Exportation of COVID-19 Cases from a Ski Area in Austria. *J. Clin. Microbiol.* **58**, e00588-20 (2020). [doi:10.1128/JCM.00588-20](#) [Medline](#)
20. H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, S. Cauchemez, Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208–211 (2020). [doi:10.1126/science.abc3517](#) [Medline](#)
21. A. R. Tuite, V. Ng, E. Rees, D. Fisman, Estimation of COVID-19 outbreak size in Italy. *Lancet Infect. Dis.* **20**, 537 (2020). [doi:10.1016/S1473-3099\(20\)30227-9](#) [Medline](#)
22. M. P. Zwart, S. F. Elena, Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annu. Rev. Virol.* **2**, 161–179 (2015). [doi:10.1146/annurev-virology-100114-055135](#) [Medline](#)
23. J. L. Geoghegan, A. M. Senior, E. C. Holmes, Pathogen population bottlenecks and adaptive landscapes: Overcoming the barriers to disease emergence. *Proc. Biol. Sci.* **283**, 20160727 (2016). [doi:10.1098/rspb.2016.0727](#) [Medline](#)
24. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *J. Virol.* **91**, e00171-17 (2017). [doi:10.1128/JVI.00171-17](#) [Medline](#)
25. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, A. Justice, A. Green, M. A. Ansari, L. Abeler-Dorner, C. E. Moore, T. E. A. Peto, R. Shaw, P. Simmonds, D. Buck, J. A. Todd, on behalf of the OSVG Analysis Group, D. Bonsall, C. Fraser, T. Gol'ubchik, Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv* (2020). [doi:10.1101/2020.05.28.118992](#)
26. S. Pfefferle, T. Günther, R. Kobbe, M. Czech-Sioli, D. Nörz, R. Santer, J. Oh, S. Kluge, L. Oestereich, K. Peldschus, D. Indenbirken, J. Huang, A. Grundhoff, M. Aepfelbacher, J. K. Knobloch, M. Lütgehetmann, N. Fischer, SARS-CoV-2 variant tracing within the first COVID-19 clusters in Northern Germany. *Clin. Microbiol. Infect.* **S1198-743X(20)30587-5** (2020). [doi:10.1016/j.cmi.2020.09.034](#) [Medline](#)
27. L. L. M. Poon, T. Song, R. Rosenfeld, X. Lin, M. B. Rogers, B. Zhou, R. Sebra, R. A. Halpin, Y. Guan, A. Twaddle, J. V. DePasse, T. B. Stockwell, D. E. Wentworth, E. C. Holmes, B. Greenbaum, J. S. M. Peiris, B. J. Cowling, E. Ghedin, Quantifying influenza virus diversity and transmission in humans. *Nat. Genet.* **48**, 195–200 (2016). [doi:10.1038/ng.3479](#) [Medline](#)
28. J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, A. S. Lauring, Stochastic processes constrain the within and between host evolution of influenza virus. *eLife* **7**, e35962 (2018). [doi:10.7554/eLife.35962](#) [Medline](#)
29. M. Prentiss, A. Chu, K. K. Berggren, Superspreading Events Without Superspreaders: Using High Attack Rate Events to Estimate N_0 for Airborne Transmission of COVID-19. *medRxiv*, 2020.10.21.20216895 (2020).
30. X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020). [doi:10.1038/s41591-020-0869-5](#) [Medline](#)
31. R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirgmaier, C. Drosten, C. Wendtner, Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469 (2020). [doi:10.1038/s41586-020-2196-x](#) [Medline](#)
32. Y. Wang, D. Wang, L. Zhang, W. Sun, Z. Zhang, W. Chen, A. Zhu, Y. Huang, F. Xiao, J. Yao, M. Gan, F. Li, L. Luo, X. Huang, Y. Zhang, S. Wong, X. Cheng, J. Ji, Z. Ou, M.

- Xiao, M. Li, J. Li, P. Ren, Z. Deng, H. Zhong, H. Yang, J. Wang, X. Xu, T. Song, ` Mok, M. Peiris, N. Zhong, J. Zhao, Y. Li, J. Li, J. Zhao, Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Population in COVID-19 Patients, *bioRxiv*, 2020.05.20.103549 (2020).
33. S. L. Díaz-Muñoz, R. Sanjuán, S. West, Sociovirology: Conflict, Cooperation, and Communication among Viruses. *Cell Host Microbe* **22**, 437–441 (2017). [doi:10.1016/j.chom.2017.09.012](https://doi.org/10.1016/j.chom.2017.09.012) [Medline](#)
 34. M. A. Nowak, C. E. Tarnita, T. Antal, Evolutionary dynamics in structured populations. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 19–30 (2010). [doi:10.1098/rstb.2009.0215](https://doi.org/10.1098/rstb.2009.0215) [Medline](#)
 35. J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. F. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L. Wyllie, C. C. Kalinich, I. M. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger, K. R. Jerome, P. Roychoudhury, H. Xie, L. Shrestha, M. L. Huang, V. E. Pitzer, A. Iwasaki, S. B. Omer, K. Khan, I. I. Bogoch, R. A. Martinello, E. F. Foxman, M. L. Landry, R. A. Neher, A. I. Ko, N. D. Grubaugh, Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990–996.e5 (2020). [doi:10.1016/j.cell.2020.04.021](https://doi.org/10.1016/j.cell.2020.04.021) [Medline](#)
 36. M. M. Böhmer, U. Buchholz, V. M. Corman, M. Hoch, K. Katz, D. V. Marosevic, S. Böhm, T. Woudenberg, N. Ackermann, R. Konrad, U. Eberle, B. Treis, A. Dangel, K. Bengs, V. Fingerle, A. Berger, S. Hörmansdorfer, S. Ippisch, B. Wicklein, A. Grahl, K. Pörtner, N. Müller, N. Zeitmann, T. S. Boender, W. Cai, A. Reich, M. An der Heiden, U. Rexroth, O. Hamouda, J. Schneider, T. Veith, B. Mühlemann, R. Wölfel, M. Antwerpen, M. Walter, U. Protzer, B. Liebl, W. Haas, A. Sing, C. Drosten, A. Zapf, Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet Infect. Dis.* **20**, 920–928 (2020). [doi:10.1016/S1473-3099\(20\)30314-5](https://doi.org/10.1016/S1473-3099(20)30314-5) [Medline](#)
 37. J. F. W. Chan, S. Yuan, K. H. Kok, K. K. W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Y. Yip, R. W. S. Poon, H. W. Tsoi, S. K. F. Lo, K. H. Chan, V. K. M. Poon, W. M. Chan, J. D. Ip, J. P. Cai, V. C. C. Cheng, H. Chen, C. K. M. Hui, K. Y. Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet* **395**, 514–523 (2020). [doi:10.1016/S0140-6736\(20\)30154-9](https://doi.org/10.1016/S0140-6736(20)30154-9) [Medline](#)
 38. L. Zhang, C. B. Jackson, H. Mou, A. Ojha, E. S. Rangarajan, T. Izard, M. Farzan, H. Choe, The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020.06.12.148726 (2020). [doi:10.1101/2020.06.12.148726](https://doi.org/10.1101/2020.06.12.148726) [Medline](#)
 39. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDaniel, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles; Sheffield COVID-19 Genomics Group, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020). [doi:10.1016/j.cell.2020.06.043](https://doi.org/10.1016/j.cell.2020.06.043) [Medline](#)
 40. Q. Li, J. Wu, J. Nie, L. Zhang, H. Hao, S. Liu, C. Zhao, Q. Zhang, H. Liu, L. Nie, H. Qin, M. Wang, Q. Lu, X. Li, Q. Sun, J. Liu, L. Zhang, X. Li, W. Huang, Y. Wang, The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284–1294.e9 (2020). [doi:10.1016/j.cell.2020.07.012](https://doi.org/10.1016/j.cell.2020.07.012) [Medline](#)
 41. P.-Y. Shi, J. Plante, Y. Liu, J. Liu, H. Xia, B. Johnson, K. Lokugamage, X. Zhang, A. Muruato, J. Zou, C. Fontes-Garfias, D. Mirchandani, D. Scharf, B. Kalveram, J. Billello, Z. Ku, Z. An, A. Freiberg, V. Menachery, X. Xie, K. Plante, S. Weaver, Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility, *Res. Sq.*, rs.3.rs-70482 (2020).
 42. S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**, 860–868 (2020). [doi:10.1126/science.abb5793](https://doi.org/10.1126/science.abb5793) [Medline](#)
 43. K. Itokawa, T. Sekizuka, M. Hashino, R. Tanaka, M. Kuroda, Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLOS ONE* **15**, e0239403 (2020). [doi:10.1371/journal.pone.0239403](https://doi.org/10.1371/journal.pone.0239403) [Medline](#)
 44. S. Andrews, FastQC - A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, Babraham Bioinform., <http://www.bioinformatics.babraham.ac.uk/projects/> (2010).
 45. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). [doi:10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) [Medline](#)
 46. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019). [doi:10.1186/s13059-018-1618-7](https://doi.org/10.1186/s13059-018-1618-7) [Medline](#)
 47. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). [doi:10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) [Medline](#)
 48. A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, N. Nagarajan, LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012). [doi:10.1093/nar/gks918](https://doi.org/10.1093/nar/gks918) [Medline](#)
 49. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011). [doi:10.1093/bioinformatics/btr509](https://doi.org/10.1093/bioinformatics/btr509) [Medline](#)
 50. P. Cingolani, A. Platts, L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012). [doi:10.4161/fly.19695](https://doi.org/10.4161/fly.19695) [Medline](#)
 51. P. Cingolani, V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, X. Lu, Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012). [doi:10.3389/fgene.2012.00035](https://doi.org/10.3389/fgene.2012.00035) [Medline](#)
 52. European Centre for Disease Prevention and Control, Contact tracing: public health management of persons, including healthcare workers, having had contact with COVID-19 cases in the European Union, (2020).
 53. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). [doi:10.1093/bioinformatics/bty407](https://doi.org/10.1093/bioinformatics/bty407) [Medline](#)
 54. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015). [doi:10.1093/molbev/msu300](https://doi.org/10.1093/molbev/msu300) [Medline](#)
 55. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Słodkiewicz, N. Goldman, Issues with SARS-CoV-2 sequencing data. *virological.org* (2020).
 56. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). [doi:10.1093/ve/vex042](https://doi.org/10.1093/ve/vex042) [Medline](#)
- Acknowledgments:** We thank the Biomedical Sequencing Facility at CeMM for assistance with next-generation sequencing. We thank Peter Obrist, Rainer Gatringer, Christian Paar, and Gregor Hörmann for providing samples and Tobias Pahlke for support with the computing cluster. We thank the Tourism office Paznaun – Ischgl for statistical data. **Funding:** AL was supported by a DOC fellowship of the Austrian Academy of Sciences. Z.K. was supported by a fellowship of the Marie Skłodowska-Curie Actions (MSCA) Innovative Training Network H2020-MSCA-ITN-2019 (grant agreement No 813343). C.B. and A.B. were supported by ERC Starting Grants (European Union's Horizon 2020 research and innovation program, grant agreement numbers 679146 respectively 677006). This project was funded in part by the Vienna Science and Technology Fund (WWTF) as part of the WWTF COVID-19 Rapid Response Funding 2020 (A.B.). **Author contributions:** AP, JWG, CB and AB designed the study design and wrote the manuscript. AP, JWG, MN, DS, BA, AL, LE, HC, MSm, MSc, MG, FM, OP, ZK, MS, SM, MB, MTW, GSF, NLB, FA, FM, CB, AB performed data analysis for this study. TP, BA, AL, MSe and JL designed assays and processed experimental samples. SWA, WB, EP, JHA, MRF, MK, AZ, PH, MN, GW, DvL, EPS provided samples and collected data. AB coordinated the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data associated with this study are in the paper or supplementary materials. An online repository of all study-related

data, results, and the interactive Nextstrain Austria database is provided on the website <http://www.sarscov2-austria.org>. Raw BAM files were submitted for inclusion in the COVID-19 Data Portal hosted by the European Bioinformatics Institute under project number PRJEB39849. Virus sequences (data file S2) are deposited in the GISAID database. All phylogenetic trees used in this study are available for visualization under the following URLs: (1) Global build: <https://nextstrain.org/community/bergthalerlab/SARS-CoV-2/NextstrainAustria>, with raw data available at <https://zenodo.org/record/4247401>; (2) Build with European strains before 31 March: <https://nextstrain.org/community/bergthalerlab/SARS-CoV-2/EarlyEurope>, raw data available at: <https://zenodo.org/record/4247401>; (3) Build with Austrian strains used for phylogenetic analysis: <https://nextstrain.org/community/bergthalerlab/SARS-CoV-2/OnlyAustrian>, with raw data available at <https://zenodo.org/record/4247401>. Code for sample processing and phylogenetic analyses is available at <https://zenodo.org/record/4247401>. The time-dynamics frequency of variants in each patient is available at: <https://zenodo.org/record/4247401>. The pairwise comparison of variants between pairs of samples in the transmission lines Fig. 5A are available at <https://zenodo.org/record/4247401>. The code to reproduce the mutational profile and genome-wide mutation rate analysis is available at <https://zenodo.org/record/4275398>. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using this material.

Submitted 12 August 2020
 Accepted 16 November 2020
 Published First Release 23 November 2020
 10.1126/scitranslmed.abe2555

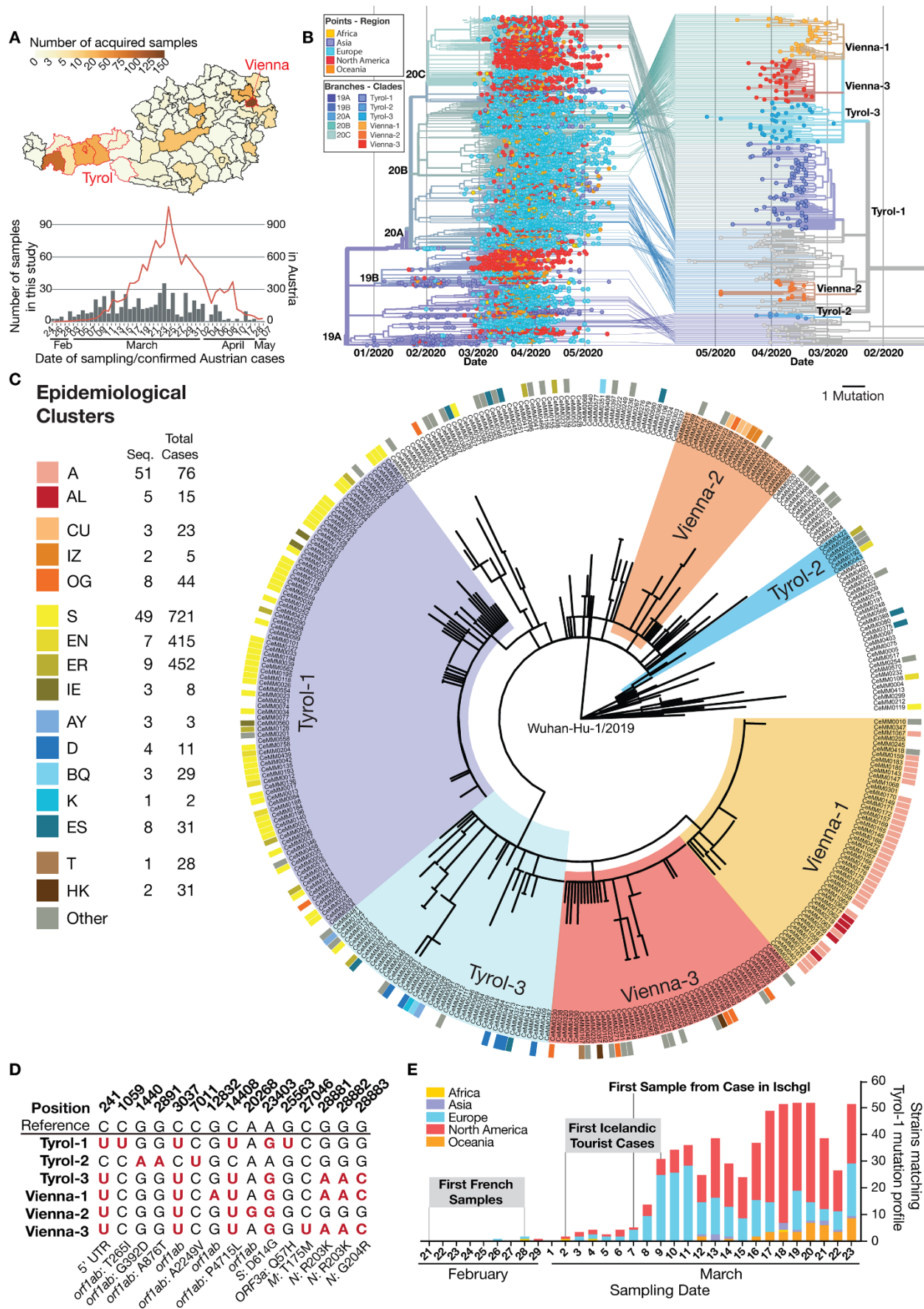
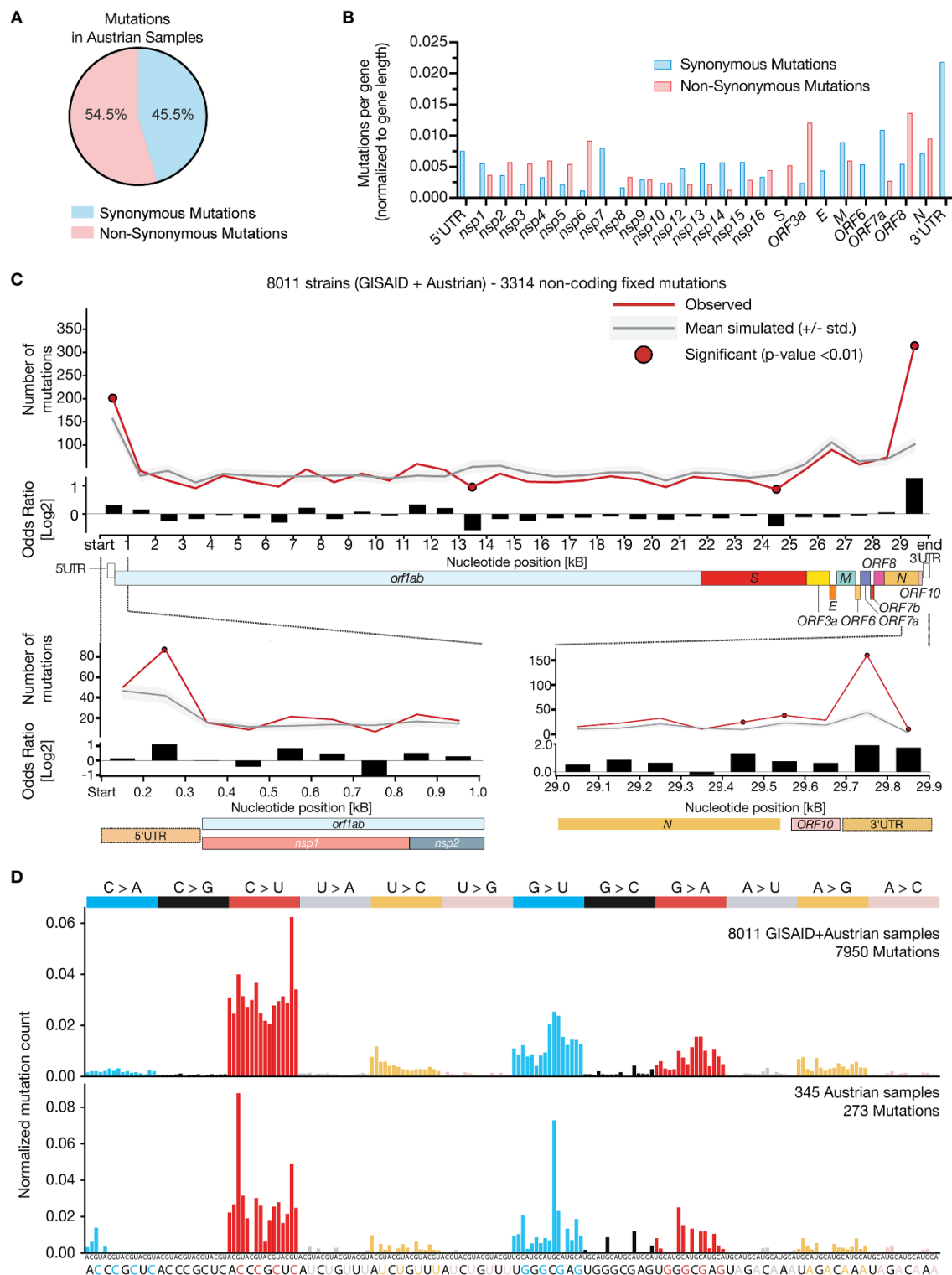


Fig. 1. Phylogenetic-epidemiological reconstruction of SARS-CoV-2 infection clusters in Austria. (A) Number of acquired samples per district in Austria (top) and sampling dates of samples that underwent viral genome sequencing in this study (bottom), plotted in the context of all confirmed cases (red line) in Austria. (B) Connection of Austrian strains to global clades of SARS-CoV-2. Points indicate the regional origin of a strain in the time-resolved phylogenetic tree from 7,666 randomly subsampled sequences obtained from GISAID including 345 Austrian strains sequenced in this study (left). Lines from global phylogenetic tree (left) to phylogenetic tree of all Austrian strains obtained in this study (right) indicate the phylogenetic relation and Nextstrain clade assignment of Austrian strains. Color schemes of branches represent Nextstrain clade assignment (left) or phylogenetic clusters of Austrian strains (right). (C) Phylogenetic tree of SARS-CoV-2 strains from Austrian patients with COVID-19 sequenced in this study. Phylogenetic clusters were identified based on characteristic mutation profiles in viral genome sequences of SARS-CoV-2 positive cases in Austria. Cluster names indicate the most abundant location of patients based on epidemiological data. The circular color code indicates the epidemiological cluster assigned to patients based on contact tracing. (D) Mutation profiles of phylogenetic clusters identified in this study. Positions with characteristic mutations compared to reference sequence “Wuhan-Hu-1” (GenBank: MN908947.3) are highlighted in red. Details regarding the affected genes or genomic regions and the respective codon and amino acid change are given below the table. (E) Timeline of the emergence of strains matching the mutation profile of the Tyrol-1 cluster in the global phylogenetic analysis by geographical distribution with additional information from European phylogenetic reconstruction.



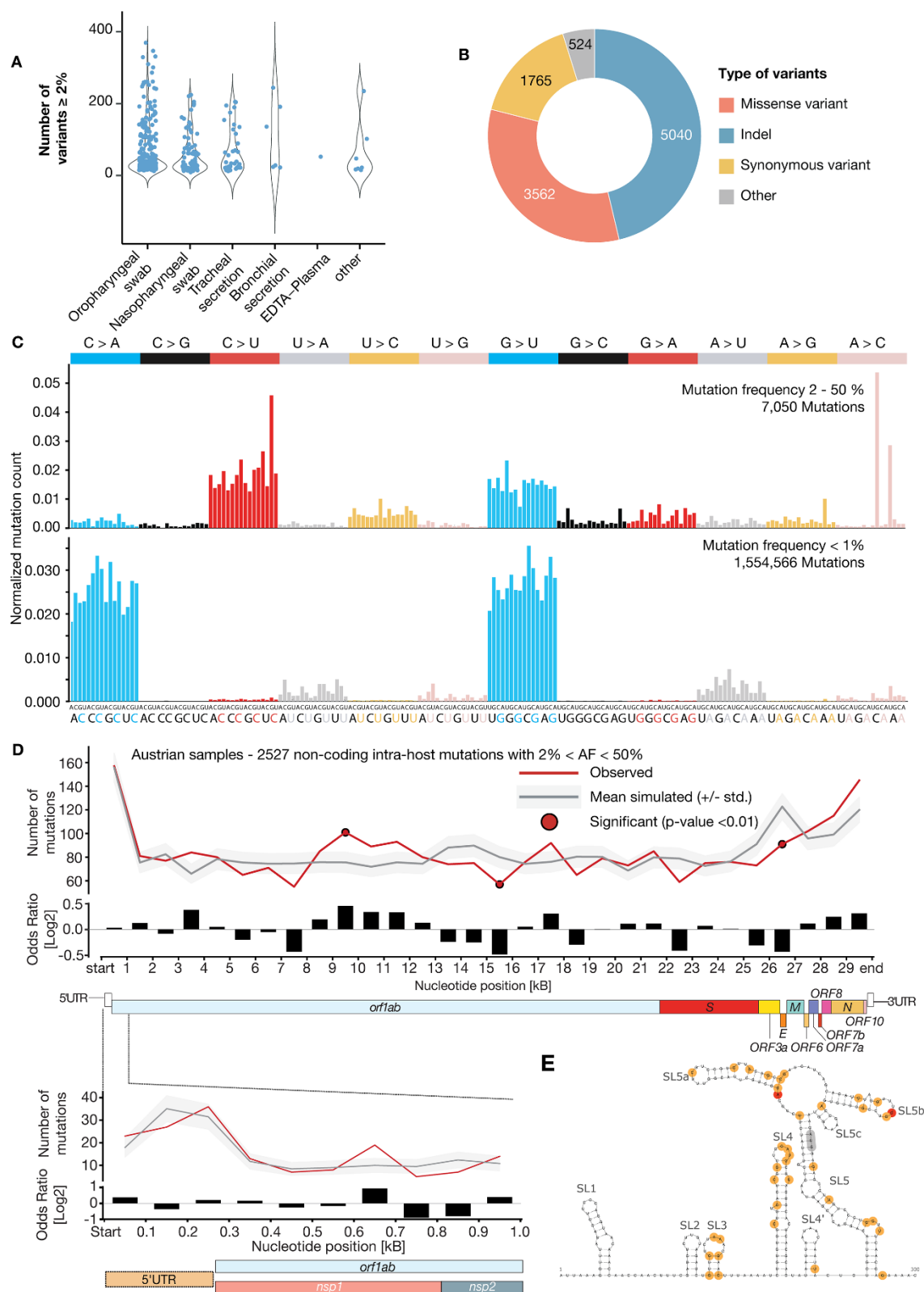


Fig. 3. Analysis of low-frequency mutations.

(A) Number of variants detected across different sample types. (B) Number of variants per variant class. (C) Mutational profile (relative probability of each trinucleotide) of 7,050 intra-host mutations across Austrian samples (alleles frequencies between 0.02 and 0.05) (upper panel). Mutational profile (relative probability of each trinucleotide) of 1,554,566 intra-host mutations across Austrian samples (alleles frequencies < 0.01) (lower panel). (D) Analysis of the mutation rate (analogous to the interhost mutation rate panel) across the SARS-CoV-2 genome using 2,527 intra-host non-protein affecting mutations with alleles frequencies between 0.02 and 0.5. (E) RNA secondary structure prediction of the upstream 300 nt of the SARS-CoV-2 reference genome (NC 045512.2), comprising the complete 5' untranslated region (UTR) and parts of the nsp1 protein nucleotide sequence. The canonical AUG start codon is located in a stacked region of SL5 (highlighted in gray). Mutational hotspots observed in the Austrian SARS-CoV-2 samples are highlighted: two fixed mutations at positions 187 and 241, respectively, are marked in red, and low-frequency variants with an abundance between [0.02, 0.5] in individual samples are shown in orange. Insertion and deletion variants are not shown.

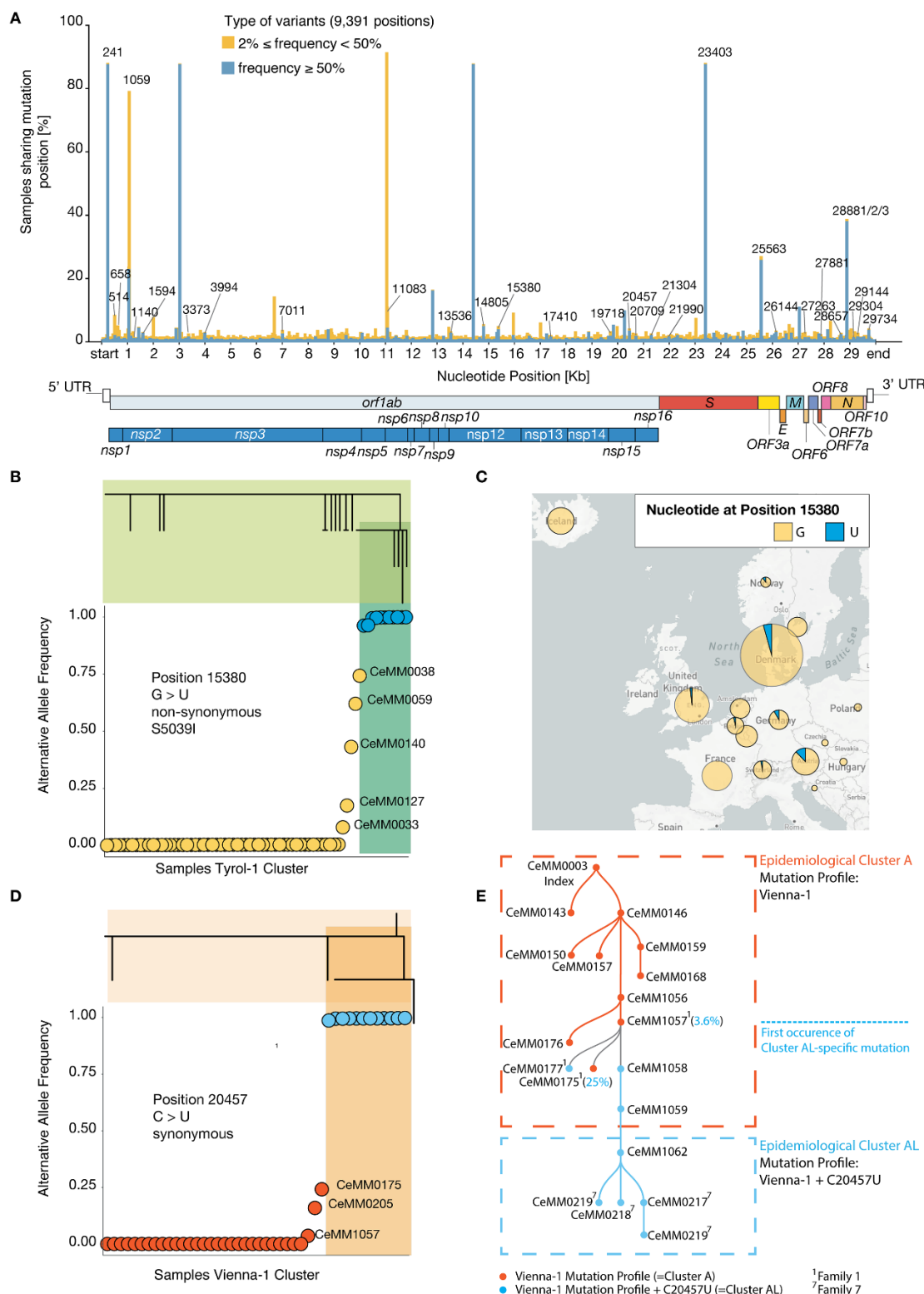


Fig. 4. Dynamics of low-frequency and fixed mutations in superspreading clusters. (A) Percentage of samples sharing detected (≥ 0.02) mutations across genomic positions. For each of the 9,391 positions harboring an alternative allele, the percentage of samples with high (≥ 0.50) or low $[0.02, 0.50]$ frequency are reported in dark blue and orange, respectively. (B) Allele frequency of non-synonymous mutation G > U at position 15,380 across samples in the phylogenetic cluster Tyrol-1. This variant has been observed both as low frequency variant and as fixed mutation, the latter defining a phylogenetic subcluster (dark green). (C) Proportion of European samples with a reference (yellow) or alternative (blue) allele at position 15,380. (D) Allele frequency of synonymous mutation C > U at position 20,457 across samples of the Vienna-1 phylogenetic cluster. This variant is fixed and defines a phylogenetic subcluster (dark orange) as part of the broader Vienna-1 cluster. (E) Schematic representation of the transmission lines between epidemiological cluster A and cluster AL reconstructed based on results from deep viral sequencing and case interviews. The transmission scheme is overlaid with epidemiological clusters and family-related information.

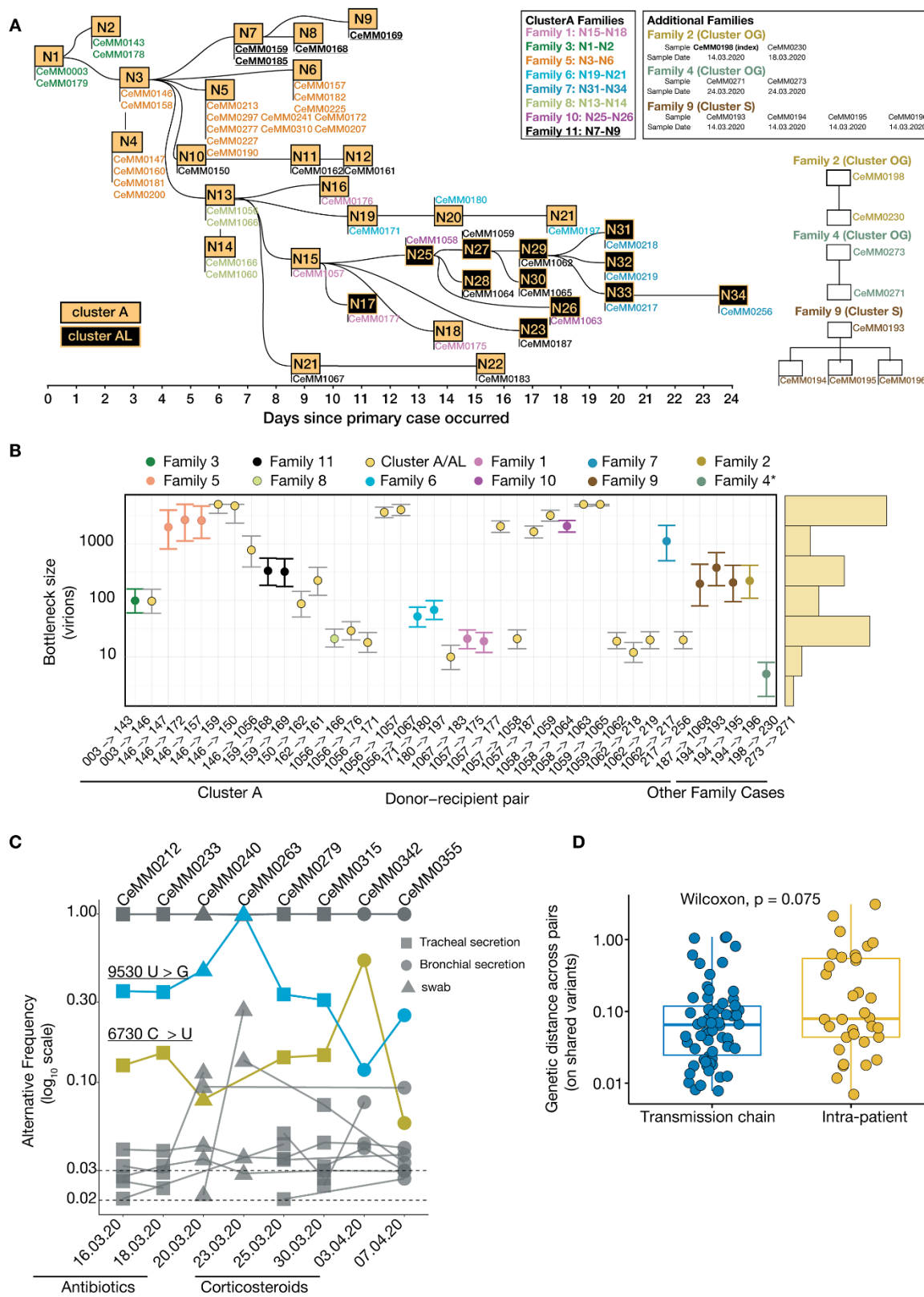


Fig. 5. Impact of transmission bottlenecks and intra-host evolution on SARS-CoV-2 mutational dynamics. (A) Schematics of time-related patient interactions across epidemiological clusters A and AL. Each node represents a case and links between the nodes are epidemiologically confirmed direct transmissions. Samples sequenced from the same individual are reported under the corresponding node. Cases corresponding to the same family are color-coded accordingly. Additional families, unrelated to clusters A/AL, and their epidemiological transmission details are also reported. (B) Bottleneck size (number of virions that initiate the infection in a recipient) estimation across donor-recipient pairs based on the transmission network depicted in Fig. 5A, ordered according to the timeline of cluster A for the respective pairs, and with a cutoff of [0.01, 0.95] for alternative allele frequency. For patients with multiple samples, the earliest sample was considered for bottleneck size inference. Centered dots are maximum likelihood estimates, with 95% confidence intervals. A star (*) for

family 4 indicates that the transmission line was inferred as detailed in the Methods. The histogram (yellow bars) of all the bottleneck values is provided on the right side of the graph. (C) Alternative allele frequency (y-axis) of mutations across available time points (x-axis) for patient 5. Only variants with frequencies ≥ 0.02 and shared between at least between two time points are shown. Two mutations increasing in frequency are color coded. (D) Genetic distance values of mutation frequencies between donor-recipient pairs (Fig. 5A-B) (transmission chains) and intra-patient consecutive time points (Fig. 5C, fig. S5D). Only variants detected in two same-patient samples were considered.

Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2

Alexandra Popa, Jakob-Wendelin Genger, Michael D. Nicholson, Thomas Penz, Daniela Schmid, Stephan W. Aberle, Benedikt Agerer, Alexander Lercher, Lukas Endler, Henrique Colaço, Mark Smyth, Michael Schuster, Miguel L. Grau, Francisco Martínez-Jiménez, Oriol Pich, Wegene Borena, Erich Pawelka, Zsófia Keszei, Martin Senekowitsch, Jan Laine, Judith H. Aberle, Monika Redlberger-Fritz, Mario Karolyi, Alexander Zoufaly, Sabine Maritschnik, Martin Borkovec, Peter Hufnagl, Manfred Nairz, Günter Weiss, Michael T. Wolfinger, Dorothee von Laer, Giulio Superti-Furga, Nuria Lopez-Bigas, Elisabeth Puchhammer-Stöckl, Franz Allerberger, Franziska Michor, Christoph Bock and Andreas Bergthaler

Sci Transl Med published online 23 November 2020

ARTICLE TOOLS

<http://stm.sciencemag.org/content/early/2020/11/20/scitranslmed.abe2555>

SUPPLEMENTARY MATERIALS

<http://stm.sciencemag.org/content/suppl/2020/11/20/scitranslmed.abe2555.DC1>

RELATED CONTENT

<http://science.sciencemag.org/content/sci/early/2020/11/23/science.abe2424.full>
<http://stm.sciencemag.org/content/scitransmed/12/570/eabd3876.full>
<http://stm.sciencemag.org/content/scitransmed/12/568/eabe0948.full>
<http://stm.sciencemag.org/content/scitransmed/12/554/eabc1126.full>
<http://stm.sciencemag.org/content/scitransmed/12/564/eabd5487.full>
<http://stm.sciencemag.org/content/scitransmed/12/559/eabc3103.full>

REFERENCES

This article cites 50 articles, 10 of which you can access for free
<http://stm.sciencemag.org/content/early/2020/11/20/scitranslmed.abe2555#BIBL>

PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.

Copyright © 2020, American Association for the Advancement of Science