Identification of super-transmitters of SARS-CoV-2

Xuemei Yang[1#], Ning Dong[1#], Edward Wai-Chi Chan[2], Sheng Chen [1*]

[1]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

[2]State Key Lab of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong;

#contribute equally to the work.

*Corresponding author: Sheng Chen, City University of Hong Kong, Kowloon, Hong Kong;
Email: shechen@cityu.edu.hk

Keywords: SARS-CoV-2, COVID-19, Origin, phylodynamic analysis, Super-transmitter

Abstract

A newly emerged coronavirus, SARS-CoV-2, caused severe outbreaks of pneumonia in China in December 2019 and has since spread to various countries around the world. To probe the origin and transmission dynamics of this virus, we performed phylodynamic analysis of 247 high quality genomic sequences of viruses available in the GISAID platform as of March 05, 2020. A substantial number of earliest sequences reported in Wuhan in December 2019, including those of viruses recovered from the Huanan Seafood Market (HNSM), the site of the initial outbreak, were genetically diverse, suggesting that viruses of multiple sources were involved in the original outbreak. The viruses were subsequently disseminated to different parts of China and other countries, with diverse mutational profiles being recorded in strains recovered subsequently. Interestingly, four genetic clusters defined as Super-transmitters (STs) were found to become dominant and were responsible for the major outbreaks in various countries. Among the four clusters, ST1 is widely disseminated in Asia and the US and mainly responsible for outbreaks in the states of Washington and California in the US as well as those in South Korea at the end of February and early March, whereas ST4 contributed to the pandemic in Europe. Each ST cluster carried a signature mutation profile which allowed us to trace the origin and transmission patterns of specific viruses in different parts of the world. Using the signature mutations as markers of STs, we further analysed 1539 genome sequences reported after February 29, 2020. We found that around 90% of these genomes belonged to STs with ST4 being the dominant one and their contribution to pandemic in different continents were also depicted. The identification of these super-transmitters provides insight into the control of further transmission of SARS-CoV-2.

Introduction

A number of newly emerged coronaviruses such as the highly pathogenic severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) have caused serious respiratory and intestinal infections in human within the past two decades [1]. In December 2019, another new coronavirus, SARS-CoV-2, has emerged and caused outbreaks of lower respiratory tract infections, often with poor clinical outcome, in Wuhan, China. The virus, which has since spread to other cities in China and various countries worldwide [2], exhibited a high potential to undergo human-to-human transmission [3]. As of March 26, 2020, a total of 692 thousand infections were documented worldwide, among which 68 thousand occurred in China (https://www.gisaid.org/epiflu-applications/global-cases-betacov/). As a result, WHO declared the risk of SARS-CoV-2 as "Very High" in China (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf).

The genomic characteristics of SARS-CoV-2 have been elucidated using phylogenetic, structural and mutational analyses by scientists across the globe [4]. High-throughput sequencing revealed that SARS-CoV-2 was a novel betacoronavirus which resembled SARS-CoV at around 79.5% sequence identity [5, 6]. A recent study indicated that SARS-CoV-2 was 96% identical to a bat coronavirus RaTG13 (accession: MN996532) at the genomic level, suggesting that bat might be a natural host of SARS-CoV-2 [7]. GISAID is a platform for sharing genetic data of influenza. Currently, a rapidly increasing number of SARS-CoV-2 genomic sequences are being deposited into this database from laboratories around the world [8]. On the other hand, some recent studies also suspected that Malayan pangolins (*Manis javanica*) could be the intermediate host of this new coronavirus, since the amino acid sequence of the S protein of coronaviruses derived from Malayan pangolins illegally imported to

Guangdong Province of China, as well as coronaviruses harboured by pangolins in Guangxi province of China, exhibited very high homology with the S protein sequence of SARS-CoV-2, even though the overall homology between SARS-CoV-2 and RaTG13 is still the highest[9]. However, due to the inability to detect or isolate SARS-CoV-2 from pangolins in the Wuhan Huanan seafood wholesale market, the site in which the first batch infected patients had commonly visited, the theory of pangolins being the culprit of the Wuhan pneumonia outbreak is not substantiated. The intermediate host for SARS-CoV-2 therefore remains a mystery. In fact, it remains unclear if the Huanan Seafood Wholesale Market is the origin of this outbreak as some of the earliest cases were confirmed to have no linkage with this market. It is urgent to identify the source(s) of viruse(s) that caused this outbreak to design more effective control measures to stop the continuous worldwide transmission of these highly contagious viruses. With more sequences released, it is very important to provide more insights of this virus through indepth sequence analysis. One recent study has analysed over 100 available genome sequences and revealed that sequences belonging to different genetic clusters have evolved [10]. In this study, we retrieved and analysed the publicly shared genome sequences as of March 25, 2020 to investigate the genetic diversity and phylodynamics of these SARS-CoV-2 viruses. We identified notable four clusters of genomes with high transmission and mutation rate and have become the most dominant viruses in the later stage of transmission. Results in this study should provide insight into the control of further transmission of SARS-CoV-2.

Materials and Methods

Sequence analysis, alignment and mutation identification

A total of 343 full-length SARS-CoV-2 genomes available in the GISAID platform (https://platform.gisaid.org/) as of March 5, 2020 were downloaded [8]. A total of 247 sequences with high sequence quality as noted in the GISAID database were included for further analysis after removing sequences containing little temporal signal and thus are not unsuitable for inference using phylogenetic molecular clock models. Information regarding the date and country of isolation were also retrieved from the GISAID platform. The annotated reference genome sequence of the SARS-CoV-2 isolate Wuhan-Hu-1 (accession: NC_045512.2) was downloaded from the NCBI GenBank database. All genomes were annotated by GATU Genome Annotator [11] using the SARS-CoV2 isolate Wuhan-Hu-1 (NC_045512.2) as reference [12]. Nucleotide and amino acid mutations of all genome and separate proteins were analyzed by blast (https://blast.ncbi.nlm.nih.gov/) using the sequence of strain Wuhan-Hu-1 as reference.

Phylodynamic analysis

Global genomic surveillance of SARS-CoV-2 was implemented by means of an automated phylogenetic analysis pipeline using Nextstrain, which generates an interactive visualization integrating a phylogeny with sample metadata such as geographic location or host age [13]. The pipeline involved the sequence alignment module with MAFFT [14], phylogenetic analysis with IQ-TREE [15], maximum-likelihood phylodynamic analysis with Treetime [16], identification of nucleotide and amino acid mutations with Augur, and result visualization with Auspice [13]. The outputs were edited by Inkscape 0.91 [17].

Phylogenetic analysis

Alignment of the complete genome sequences was conducted with MAFFT v7.310 [14]. Phylogenetic tree of all SARS-CoV-2 was built with RAxML version 8.2.4 [18]. The tree was edited by iTOL [19].

Quick identification of the types of SARS-CoV-2 genomes in the database

All complete genomes available as March 28 on the GISAID database were downloaded. Single Nucleotide Polymorphisms (SNPs) calling were performed by Snippy (https://github.com/tseemann/snippy) using Wuhan-Hu-1 as reference. Super-transmitter clusters were classified according relative variants. A total of 1956 qualified genomes submitted after February 29, 2020 were included.

Results

Phylodynamics analysis of genome sequences of SARS-CoV-2 strains collected worldwide

To trace the evolutional process and identify the common ancestor of 247 strains of SARS-CoV-2 collected worldwide, root-to-tip regression scatter plots was conducted among all SARS-CoV-2 genomes, with R^2 being found to be 0.23, suggesting that these 247 viral sequences shared a common recent ancestor **(Fig 1a)**. The date of the most recent common ancestor (tMRCA) of all reported SARS-CoV-2 viruses was 2019-Nov-12, suggesting that this virus emerged recently **(Fig 1a)**. A total of 379 nucleotide mutations were identified among these 247 sequences based on the sequence alignment, among which $G^{11083}T$ (n=5), T3G (n=3), $G^{29864}A$ (n=3), $C^{29870}A$ (n=3), A1T (n=2), A4T (n=2), $T^{4402}C$ (n=2), $G^{5062}T$ (n=2), $T^{18603}C$ (n=2) and $G^{22661}T$ (n=2) were the most homoplasic mutations (Fig 2, supplementary file 1). A total of 147 strains were found to contain single amino acid change, with the majority of such changes being located within ORF1ab (n=104). The $L^{3606}F$ change was detected in two viral sequences, while other mutations occurred only once. Mutations that result in amino acid changes include single substitution in the S protein (n=19, $D^{614}G$, $L^{752}F$, $F^{32}I$, $H^{655}Y$, $V^{483}A$, $F^{157}L$, $V^{615}L$, $K^{202}N$, $S^{939}F$, $F^{797}C$, $A^{93}0V$, $R^{408}I$, $V^{367}F$, $Q^{409}E$, $S^{254}F$, A435S, $D^{1146}E$, $S^{247}R$ and $P^{1143}L$), ORF3a (n=8, $E^{191}G$, $G^{76}S$, $K^{61}N$, $V^{259}L$, $T^{176}I$, $L^{140}V$, $T^{269}M$ and $V^{88}L$), N protein (n=6, $K^{247}I$, $S^{194}L$, $P^{46}S$, $S^{327}L$, $E^{378}Q$ and $D^{343}V$), ORF8 (n=4, $T^{11}I$, $L^{84}S$, $S^{97}N$ and $S^{67}F$), ORF7a (n=3, $P^{34}S$, Q62* and $H^{73}Q$), ORF10 (n=2, $P^{10}S$ and $I^{13}M$) and E protein (n=1, $S^{6}L$) (**Fig 1b**). Identification of single amino acid substitutions in SARS-CoV-2 isolates consistently showed that these isolates shared a recent common ancestor but entered diverse evolution paths. The estimated substitution rate of SARS-CoV-2 was 8.90e-04 subs/site/year, which was similar to that of other RNA viruses including SARS-CoV, Ebola virus, Zika virus, and others, which was found to be at ~ 1e-3 subs/site/year (http://virological.org/t/phylodynamic-analysis-93-genomes-15-feb-2020/356). Based on this mutation rate, a genome of 29kb of SARS-CoV-2 will

end up with ~26 mutations per genome per year, suggesting that within the two months' study period, the number of mutations in each genome should not exceed five if all test isolates emerged as a result of natural evolution of a single SARS-CoV-2 strain.

Multiple origins of SARS-CoV-2 in Wuhan, China

To shed light on the evolution trend of SARS-CoV-2, we analysed the time-dependent changes in mutation profiles of the test strains in detail. A total of 16 viral genomes collected before Jan. 01, 2020, were included (**Table 1**). All of these 16 genomes were obtained from Wuhan, with half of them from Huanan Seafood Wholesale Market (HNSM) where the original outbreak occurred. Six genomes contained identical sequences, four of which belonged to isolates obtained from HNSM. Compared to these six viral genomes, others displayed various mutation profiles which comprised 1 to 6 mutations in the genomes. We therefore set these six genomes as reference genome for subsequent analyses. Two earliest viral genomes reported on Dec. 24 and 26 were found to harbour two and three mutations when compared to the reference viral genome, respectively. Four viral genomes from HNSM also contained two mutations with different profiles, suggesting that the original SARS-CoV-2 strain might have been circulating in HNSM for a certain period of time and underwent mutational changes in different intermediate hosts (**Table S1**). These observations suggested that HNSM was not the only origin of the COVID-19 outbreak, instead the market might only serve as a medium in which transmission of this virus to human first occurred. The original virus seemed to have transmitted to various provinces in China subsequently, including Guangdong, Zhejiang, Anhui, Jiangshu and Chongqing, and then to other countries including Japan, Taiwan, Thailand and USA in the following month (January 2020). The viral genome reported initially from USA were those of viruses recovered from the patients in Princess Diamond Cruise, confirming that the original virus was the one that caused the outbreak in this cruise; such view is consistent

with the finding that identical genomes were reported in Japan, where the cruise ship was docked. A total of 26 out of the 247 genome sequences tested contain one mutation. Unless isolated from the same location, most of these genomes exhibit unique mutational profile. Five sequences from the Princess Diamond cruise ship were found to exhibit unique mutation profiles, thus further suggesting that random mutations occurred during viral evolution. It should be noted that these genome sequences were also reported in Wuhan, other parts of China and various other countries, confirming that the transmission of the original virus to different parts of the world was accompanied by active but random mutational changes during the process (**Table S1**).

Phylogenetic analysis of genome sequences of SARS-CoV-2

Phylogenetic analysis of the 247 SARS-CoV-2 genomes was also performed, with results showing that such viral genomes exhibited highly diverse genetic profiles and that random mutations occurred during the evolutional process within the first two months. Interestingly, four clusters of genome sequences were observed among the 247 genomes, with the rest exhibiting more diverse profiles. These results were consistent with the data of maximum-likelihood phylodynamic analysis shown in Figure 1. Comparison of the mutation profile of each cluster enabled us to discover that all viral genomes in the same cluster were derived from one parental viral genome sequence which bears a signature mutation profile, as such profile could be identified in all offsprings (**Fig 2**). The first cluster contained two mutations, $C^{8782}T$ and $T^{28144}C$; the second cluster contained the mutation $G^{26144}T$; the third cluster contained the mutation $G^{11083}T$; the fourth cluster contained three mutations, $C^{241}T$, $C^{3037}T$ and $A^{23403}G$. Tracing the changes in mutation profiles of these viral genomes over time allowed us to visualize the transmission and evolution dynamics of SARS-CoV-2. Since viruses of all of

these four clusters exhibited very high potential to undergo global transmission, we define viruses in these four clusters as super-transmitter cluster 1 (ST1), 2 (ST2), 3(ST3) and 4(ST4).

Evolution and transmission of super-transmitter cluster 1 (ST1)

ST1 carried the signature mutation profile of $C^{8782}T$ and $T^{28144}C$. The $C^{8782}T$ change is a silent mutation, whereas $T^{28144}C$ is associated with the amino acid substitution $L^{84}S$ in the Orf8ab protein. The ST1 viruses were transmitted very efficiently and a total of 85 out of 247 (34%) genome sequences belonging to this cluster as of March 03, 2010. The earliest sequence of in this cluster was reported in Wuhan, China on Jan 05, 2020 and seven were subsequently reported in January and February in different parts of China and Australia, suggesting that widespread transmission of this cluster of viruses occurred (**Table 2**). The viruses in ST1 were mainly transmitted among Asian countries area such as China, Japan, South Korea, Taiwan and Singapore, as well as North America, in particular the states of California and Washington in USA (**Table 2**). The viruses in ST1 were also found to be able to rapidly mutate along the transmission paths. Three genome sequences that were reported in Australia, Vietnam and USA on Feb. 28, 24 and Mar. 03, 2020 respectively, were found to harbour a total of 11 mutations. An additional nine mutations were acquired by the parental virus within 50 days (from Jan 05 to Feb. 24, 2020), with a mutation rate of 2.3e-3 subs/site/year (29kb genome size), which was much higher than the predicted mutation rate of SARS-CoV-2 (4.057 e-4 subs/site/year) and other coronaviruses such as SARS-CoV-1 and MERS virus. Among viral genomes in this cluster, 43 out of 85 genomes exhibited five or more mutations (**Table 2**).

Detailed analysis of mutation profiles of the genome sequences in ST1 enables us to trace the evolution routes of these viruses in specific region. In Washington State, USA, a genome sequence with three mutations, $C^{18060}T$, $C^{8782}T$ and $T^{28144}C$, was reported on Jan 25, 2020. A

virus carrying these three mutations was reported in Fujian, China on Jan 21, 2020, suggesting that the virus might have originated from Fujian, China (**Table 2**). This virus was then further transmitted in the Washington area and continued to acquire mutations. Twelve genome sequences reported between March 01 - 05, 2020 in the states of Washington and California contained 6 to 11 mutations. The data provided direct evidence of active evolution that results in a large number of mutational changes during the process of transmission of a single virus within a short period of time. In addition, a genome sequence which has two additional mutations when compared to the original virus in this cluster, but were different from those in genome sequences in Washington, was reported in Sichuan, China, suggesting that the same parental virus was also transmitted across China during this period (**Table 2**).

Evolution and transmission of super-transmitter 2 (ST2)

ST2 carried the signature mutation $G^{26144}T$, which resulted in the $G^{251}V$ amino acid substitution in Orf 3 protein of SARS-CoV-2. The first viral genome in this cluster was reported on Jan. 25, 2020 in Australia and a total of 28 out of the 247 (11.2%) sequences tested were found to belong to ST2. The parental virus had acquired different mutations and had been disseminated to various Asian countries, North America (USA), Europe, South America (Brazil) and transmission in Australia. Viruses in this cluster seemed to be extensively transmitted by the end of January and lasted till early February. By the end of February, however, transmission efficiency of such viruses seemed to have dropped, as only 4 oof 28 sequences reported during the period Feb. 26 to Mar. 03. 2020 belong to this cluster. Viruses in this cluster were also found to have significantly mutated. Examples are three genomes reported from Switzerland on Feb. 29, 2020, in which 9, 10 and 11 mutations were identified respectively (**Table 3**). Our data showed that as many as eight additional mutations were acquired by the parental virus within 30 days (from Jan 28 to Feb. 29, 2020), representing a mutation rate of 3.3e-3

subs/site/year (29kb genome size), which was much higher than the predicted mutation rate of SARS-CoV-2 (8.0e-4 subs/site/year).

Evolution and transmission of super-transmitter 3 (ST3)

ST3 carried the signature mutation $G^{11083}T$, which caused the $L^{3606}F$ amino acid substitution in the Orf 1 protein of SARS-CoV-2. The first viral genome in this cluster was reported on Jan. 18, 2020 in Chongqing, China and a total of 22 (9%) genome sequences were reported so far. It has since been transmitted to some Asian countries including Singapore, as well as Japan, Europe, USA and Australia (**Table 4**). Like ST1 and ST2, viruses in this cluster were also found to mutate efficiently, with one genome reported on Feb. 27, 2020 from Washington, USA, carrying 12 mutations. Our data showed that a total of eleven mutations were acquired by the parental virus within a 40 days period (from Jan 18 to Feb. 27, 2020), with a mutation rate of 2.8e-3 subs/site/year (29kb genome size), which was again much higher than the predicted mutation rate of SARS-CoV-2 (8.0e-4 subs/site/year). Curiously, there is no virus of this cluster being reported in Iran, a country with one of the highest incidence of SARS-CoV-2 infections. However, two genome sequences from Australia, which belong to viruses recovered from patients with travel history to Iran, were reported, suggesting that this cluster of virus might also contribute to the outbreaks in Iran. In addition, the first genome sequence from Brazil, which might have originated from Italy, also belonged to this cluster (**Table 4**).

Evolution and transmission of super-transmitter 4 (ST4)

ST4 carried a signature mutation profile that consists of three mutations: $C^{241}T$, $C^{3037}T$ and $A^{23403}G$. The $C^{241}T$ and $C^{3037}T$ changes are silent mutations, whereas $A^{23403}G$ results in the $D^{614}G$ substitution in the spike (S) protein of SARS-CoV-2. ST4 viruses were found to be transmitted only in Europe, with the exception of one genome from Mexico with travel history

from Italy, and contributed to the current explosive increase in incidence of COVID-19 in Europe (**Table 5**). Genomes in ST4 were reported more recently, from the end of February to early March. A total of 21 out of 247 (8.4%) genome sequences have been reported so far. Among the 247 genome sequences of the four clusters of super-transmitters, there is no genome containing either one of these three mutations or a combination of two of these three mutations, suggesting that the parental viral genome of ST4 could not be identified. The first virus of this cluster was reported on Jan 28, 2020 in Germany. This virus acquired another silent mutation, $C^{14408}T$, and further spread to other countries in Europe. This virus was also found to mutate efficiently, with three genomes reported on Feb. 29, 2020 from Switzerland carrying 9, 10 and 11 mutations respectively. Eight additional mutations were acquired by the parental virus within 30 days (from Jan 28 to Feb. 29, 2020), with a mutation rate of 3.3e-3 subs/site/year (29kb genome size), which was much higher than the predicted mutation rate of SARS-CoV-2 (8.0e-4 subs/site/year). At a later stage, ST4 viruses became the most efficient in causing transmission in Europe. Among the 28 genomes reported between Feb. 20 to March 03, 2020, 20 (71%) belonged to this transmitter (**Table 5**).

Temporal and spatial distribution of super-transmitters of SARS-CoV-2

To better understand the temporal and spatial distribution of these super-transmitters, we plot variation in the types of genome sequences recovered from different continents against time. The original viruses were transmitted in the first week before the emergence of these super-transmitters. ST1 was the first batch of viruses that emerged and dissemination continued throughout the study period. Other STs emerged at different time points and transmission also peaked at different dates. Transmission of ST2 and ST3 mainly occurred between mid January to mid February. Transmission of ST4 viruses mainly began at the end of February. Viruses of the four clusters exhibited much higher mutation rate than those which exhibited diverse

genetic profiles and could not be allocated into specific genetic cluster when compared to the original genome (**Fig 3a**). ST1 viruses were those which were disseminated extensively in China, in particular in the later stage of the outbreak (**Fig 3b**). ST1, ST2 and ST3 were prevalent in other Asian countries (**Fig 3c**). All the four clusters were involved in the outbreaks in Europe in the early stage, but ST4 was the cluster that eventually transformed the outbreaks in Europe to the pandemic level (**Fig 3d**). In Oceania, ST1 was involved mainly in the early stage of outbreak, yet ST2 became dominant at the later stage (**Fig 3e**). ST1 and other types of viruses were the major transmitters in the US. ST1 was shown to be transmitted mainly in the states of Washington and California, whereas the other types were mainly transmitted in other states, (**Fig 3f, Table S1**).

Distribution of different types of most recent SARS-CoV-2 in different parts of the world

Upon finishing our manuscript, we went to check the available genome sequences in the database and found a rapid increase of numbers of sequences. A total of 1539 genome sequences reported after February 29, 2020 were included for a quick analysis to identify the type of these most recent genomes. As shown in Table 6, most of the genomes were reported from USA (968 / 63%) and Europe (441 / 29%), where the pandemics were the most server. It is good to see some genomes from Africa (20 / 1%) and South America (23 / 2%), which were minimally reported before March 01, 2020. Among these genomes, 89% of the genomes belonged to ST1-4 with ST4 being the most dominant (56%), while the original derivatives accounted for only 11%, which were mainly reported in UK and Netherland. In Africa, ST4 (18/20, 90%) was the major type with some of the cases showing travel history to Europe; in Asia, the major types became ST3 (17/33, 52%) and ST4 (16/33, 48%); in Europe, all types were presence with ST4 being the dominant one (668/968, 69%); all the types were reported in the US with ST1 (282/441, 62%) and ST4 (137/441, 14%) being the dominant; in Canada,

all types except for ST1 were present; in Oceania, all four STs were present with ST3 being the dominant; in South America, ST1 and ST4 were the most dominant types (**Table 6**).

Discussion

We conducted detailed and comprehensive analyses of sequences of SARS-CoV-2 reported from December to March 05, 2020 and deposited in the GISAID database. The detailed analysis of 247 high quality genome sequences of SARS-CoV-2 provides insight into the evolution and transmission of this novel virus (**Fig 5**). The ancestor of SARS-CoV-2 could have emerged at a date as early as November, 2019 based on results of phylodynamics analysis of these genome sequences. According to the time line of outbreaks, the original virus from Wuhan city and HNSM was responsible for the widespread transmission of SARS-CoV-2 in different parts of the world in January. The origin of the outbreak was not limited to HNSM, those which occurred in multiple sites in Wuhan city might have contributed to the early transmission events and subsequent dissemination to different parts of China and various countries around the world. These data also implied that wild animals in HNSM may not be the intermediate host of SARS-CoV-2 as sources other than HNSM are also considerd the origin of this virus. Given the fact that multiple patients in Wuhan were simultaneously infected by viruses of different genetic composition in the initial outbreak, we hypothesize that a common wild animal such as wild rat would be the most likely intermediate host. Alternatively, a common environmental factor, such as a faulty sewage system, may be involved. It is therefore necessary to investigate the role of a common animal vector or dissemination route in eliciting the SARS-CoV-2 outbreak.

Interestingly, as the original virus continued to transmit in China and over the world, it has evolved into four major genetic clusters, namely super-transmitter clusters, along with other non-cluster variants derived from the original virus. Each ST cluster carried its unique signature mutation(s), which enable us to trace its origin and transmission dynamics. In the early transmission stage (December and January, 2020), variants from the original virus were

dominant, yet by the end of February and early March, these four super-transmitters became dominant, with different STs being prevalent in different regions of the world. ST1 was prevalent in China and other parts of Asia and became the major virus that caused severe outbreaks in Washington and California states in the US and South Korea; ST2 and ST3 were extensively transmitted in other parts of Asia and Europe during the end of January and early February but its prevalence dropped at the end of February and early March, and was replaced by ST4 which contributed to the pandemic in Europe. Mapping the mutational profile of viral genomes enables us to trace the transmission of viruses of different clusters in different parts of the world. Interestingly, ST4 was not reported in China or other parts of the world. The first genome of this cluster was reported in Germany and contributed to the rapid dissemination of SARS-CoV-2 in Europe. Mutation profile with ST4 is unique, with three mutations being observed in the first viral genome. Genomes with one or two of these three mutations were not reported anywhere. These data do not simply imply that ST4 originated from Europe. One limitation of the study is that we can only utilize currently available genome sequences. The lack of genome sequence of ST4 in other continent does not necessarily mean that ST4 viruses are not present in other continents. A second limitation of this study is the lack of data to explain the mechanisms underlying the evolution of these genetic clusters into super-transmitters. Every ST cluster carried at least one amino acid mutation in different protein. Whether such mutational changes were the key step that enabled SARS-CoV-2 to evolve into super-transmitters must be investigated in future research studies.

Lastly, our data also provided insight into the major transmitting viruses in current pandemic areas in the world. For example, in Italy, ST2, ST3 were reported in the end of January, while ST4 was reported in February and early March. Similar trends were seen in other countries with exception that a higher proportion of the original viral genomes were reported in

Netherland. In the US, the original viruses were reported in other states, while ST1 was the major virus that caused outbreak in Washington and California States. Other ST genomes were also sporadically reported in the US. Although data from Iran is not available, two genomes reported from Australia with travel history from Iran were shown to belong to ST3, suggesting that this cluster was responsible for the pandemic in Iran. In Australia, all genomes were reported except ST4.

Using the signature mutations as markers for different STs, we were able to analysed 1539 genomes reported in March. The data further confirmed that four STs became dominant in March with around 90% of the genomes belonging to these four STs, among which ST4 became the dominant cluster transmitting in Europe. It also stared transmitting to other parts of world including Africa, Asia, North America and Oceania. ST1 was still the major type transmitting in the US and has transmitted to South America in particular Brazil. These data confirmed that ST1 and ST4 would become worldwide transmitter and dominate the future transmission of SARS-CoV-2 in the world.

In conclusion, this study show that four major genetic clusters of viruses evolved from the original SARS-CoV-2 and have transmitted extensively over the world, each becoming dominant in different parts of the world, and that viruses without any signature mutation of the four super-transmitters appear to be transmitted much less efficiently. These super-transmitters exhibit not only high transmission efficiency, but also high mutation rate without compromising infectivity, compromising effectiveness of current infection control effort. Our findings therefore provide important insight into the molecular features of the highly transmissible variants of SARS-CoV-2.

References

1. Cui J, Li F, Shi Z-L: **Origin and evolution of pathogenic coronaviruses**. *Nature reviews Microbiology* 2019, **17**(3):181-192.
2. Gorbalenya AE: **Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group**. *bioRxiv* 2020.
3. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, Liu L, Shan H, Lei C-l, Hui DS: **Clinical characteristics of 2019 novel coronavirus infection in China**. *medRxiv* 2020.
4. Sardar R, Satish D, Birla S, Gupta D: **Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis**. *bioRxiv* 2020.
5. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N: **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding**. *The Lancet* 2020.
6. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R: **A novel coronavirus from patients with pneumonia in China, 2019**. *New England Journal of Medicine* 2020.
7. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y *et al*: **Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study**. *Lancet* 2020.
8. Shu Y, McCauley J: **GISAID: Global initiative on sharing all influenza data–from vision to reality**. *Eurosurveillance* 2017, **22**(13).
9. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020.
10. X Tang CW, X Li, Y Song, X Yao, X Wu, Y Duan, H Zhang, Y Wang, Z Qian, J Cui, J Lu: **On the origin and continuing evolution of SARS-CoV-2**. *National Science Review, nwaa036, https://doiorg/101093/nsr/nwaa036* 2020.
11. Tcherepanov V, Ehlers A, Upton C: **Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome**. *BMC Genomics* 2006, **7**:150.
12. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY *et al*: **A new coronavirus associated with human respiratory disease in China**. *Nature* 2020.
13. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA: **Nextstrain: real-time tracking of pathogen evolution**. *Bioinformatics* 2018, **34**(23):4121-4123.
14. Katoh K, Asimenos G, Toh H: **Multiple alignment of DNA sequences with MAFFT**. In: *Bioinformatics for DNA sequence analysis.* edn.: Springer; 2009: 39-64.
15. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Molecular biology and evolution* 2015, **32**(1):268-274.
16. Sagulenko P, Puller V, Neher RA: **TreeTime: Maximum-likelihood phylodynamic analysis**. *Virus evolution* 2018, **4**(1):vex042.
17. Bah T: **Inkscape: guide to a vector drawing program**: prentice hall press; 2007.
18. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.
19. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees**. *Nucleic Acids Research* 2016, **44**(Web Server issue):W242-W245.

Acknowledgments

We are grateful for the discussion and comments from Prof. Mengsu Yang from City University of Hong Kong. We acknowledge the use of the genome sequences in the GISAID platform. This study was not supported by any research grant.

Conflicts of interest

We declare that we have no conflict of interest.

459 **Table 1. Mutational profile analysis of SARS-CoV-2 genome sequences obtained in**
460 **December of 2019.**

Accession ID	Location	Collection date	Host	Origin	Number of mutations / % of sequence homology with reference sequence	Nucleotide changes
EPI_ISL_402119	China / Hubei / Wuhan	2019-12-30	Human	HNSM	0 / 100%	/
EPI_ISL_402124	China / Hubei / Wuhan	2019-12-30	Human	HNSM	0 / 100%	/
EPI_ISL_402125	China	2019-12-31	Human	HNSM	0 / 100%	/
EPI_ISL_402129	China / Hubei / Wuhan	2019-12-30	Human	HNSM	0 / 100%	/
EPI_ISL_403929	China / Hubei / Wuhan	2019-12-30	Human	WH	0 / 100%	/
EPI_ISL_412899	China / Hubei / Wuhan	2019-12-30	human	WH	0 / 100%	/
EPI_ISL_402132	China / Hubei / Wuhan	2019-12-30	Human	WH	1 / 99%	T21656A
EPI_ISL_403930	China / Hubei / Wuhan	2019-12-30	Human	WH	1 / 99%	T6996C
EPI_ISL_412898	China / Hubei / Wuhan	2019-12-30	human	WH	1 / 99%	A24325G
EPI_ISL_402121	China / Hubei / Wuhan	2019-12-30	Human	HNSM	2 / 99%	G20670A; G20679A
EPI_ISL_402127	China / Hubei / Wuhan	2019-12-30	Human	HNSM	2 / 99%	G21316A; A24325G
EPI_ISL_402128	China / Hubei / Wuhan	2019-12-30	Human	HNSM	2 / 99%	G7016A; A21137G
EPI_ISL_402130	China / Hubei / Wuhan	2019-12-30	Human	HNSM	2 / 99%	A8001C; C9534T
EPI_ISL_406798	China / Hubei / Wuhan	2019-12-26	Human	WH	2 / 99%	C6968A; T11764A
EPI_ISL_402123	China / Hubei / Wuhan	2019-12-24	Human	WH	3 / 99%	A3778G; A8388G; T8987A
EPI_ISL_403931	China / Hubei / Wuhan	2019-12-30	Human	WH	6 / 99%	T104A; T111C; T112G; C119G; T120C; G124A

461

462 HNSM, Hua Nan Seafood Wholesale Market.

463

464 **Table 2. Mutational analysis of genome sequences in super-transmitter cluster 1.**
465

Accession ID	Location	Collection date	Origin	Number of mutations / % of sequence homology with reference sequence	Nucleotide changes
EPI_ISL_406801	Asia / China / Hubei / Wuhan	2020-01-05	NC	2 / 99%	☆ (C8782T, T28144C)
EPI_ISL_407893	Oceania / Australia / New South Wales / Sydney	2020-01-24	NC	2 / 99%	☆
EPI_ISL_412979	China / Hubei / Wuhan	2020-01-18	NC	2 / 99%	☆
EPI_ISL_413691	China	2020-01	NC	2 / 99%	☆
EPI_ISL_413729	China	2020-02	NC	2 / 99%	☆
EPI_ISL_413746	China	2020-02	NC	2 / 99%	☆
EPI_ISL_413748	China	2020-02	NC	2 / 99%	☆
EPI_ISL_413750	China	2020-02	NC	2 / 99%	☆
EPI_ISL_403932	China / Guandong / Shenzhen	2020-01-14	NC	3 / 99%	☆ ╣ (C8782T, T28144C, C29095T)
EPI_ISL_403933	China / Guandong / Shenzhen	2020-01-15	NC	3 / 99%	☆ ╣
EPI_ISL_403935	China / Guangdong / Shenzhen	2020-01-15	NC	3 / 99%	☆ ╣
EPI_ISL_406030	China / Guangdong / Shenzhen	2020-01-10	WH	3 / 99%	☆ ╣
EPI_ISL_406593	Asia / China / Guandong / Shenzhen	2020-01-13	NC	3 / 99%	☆ ╣
EPI_ISL_406223	USA / Arizona / Phoenix	2020-01-22	NC	4 / 99%	☆ ╣G11083T
EPI_ISL_413809	China	2020-02	NC	4 / 99%	☆ ╣G18686T
EPI_ISL_408666	Japan / Tokyo	2020-01-31	NC	4 / 99%	☆ ╣C2662T
EPI_ISL_408665	Japan / Tokyo	2020-01-29	NC	5 / 99%	☆ ╣C2662T; C3792T
EPI_ISL_408667	Japan / Tokyo	2020-01-31	NC	5 / 99%	☆ ╣C2662T, G29705T
EPI_ISL_405839	China / Guangdong / Shenzhen	2020-01-11	WH	5 / 99%	☆ ╣C9561T; T15607C
EPI_ISL_411956	North America / USA / Texas	2020-02-11	NC	7 / 99%	☆ ╣ T18603C; T18975A; A19175C; C27925T

EPI_ISL_404895	USA / Washington / Snohomish County	2020-01-19	NC	3 / 99%	☆ T28144C
EPI_ISL_407976	Europe / Belgium / Leuven	2020-02-03	WH	3 / 99%	☆ A29863T
EPI_ISL_408480	China / Yunnan / Kunming	2020-01-17	NC	3 / 99%	☆ G11083T
EPI_ISL_408489	Taiwan / Taipei	2020-01-31	WH	3 / 99%	☆ G11528S
EPI_ISL_410535	Singapore	2020-02-03	NC	3 / 99%	☆ G28878A
EPI_ISL_411926	Taiwan / Taipei	2020-01-24	NC	3 / 99%	☆ A29889G
EPI_ISL_413854	China / Guangdong	2020-01-30	NC	3 / 99%	☆ C6501T
EPI_ISL_411060	China / Fujian	2020-01-21	NC	3 / 99%	☆ ╧ (C8782T, T28144C, C18060T)
EPI_ISL_407214	USA / Washington	2020-01-25	NC	3 / 99%	☆ ╧
EPI_ISL_407215	USA / Washington	2020-01-25	NC	3 / 99%	☆ ╧
EPI_ISL_408478	China / Chongqinq / Yongchuan	2020-01-21	NC	5 / 99%	☆ ╧ C29200T, C1342T
EPI_ISL_413456	USA / Washington / King County	2020-02-20	NC	5 / 99%	☆ ╧ ╤ (C8782T, T28144C, C18060T C17747T, A17858G)
EPI_ISL_413560	USA / Washington	2020-02-28	NC	5 / 99%	☆ ╧ ╤
EPI_ISL_412970	USA / Washington / Snohomish County	2020-02-24	NC	6 / 99%	☆ ╧ ╤ C5784T,
EPI_ISL_413457	USA / Washington	2020-02-29	NC	6 / 99%	☆ ╧ ╤ C20S,
EPI_ISL_413458	USA / Washington	2020-03-01	NC	6 / 99%	☆ ╧ ╤ T20281C
EPI_ISL_413563	USA / Washington	2020-03-03	NC	6 / 99%	☆ ╧ ╤ C9430A,
EPI_ISL_413650	USA / Washington	2020-03-05	NC	6 / 99%	☆ ╧ ╤ T23010C
EPI_ISL_413651	USA / Washington	2020-03-05	NC	6 / 99%	☆ ╧ ╤ T23010C
EPI_ISL_413653	USA / Washington	2020-03-05	NC	6 / 99%	☆ ╧ ╤ A6T
EPI_ISL_413455	USA / Washington	2020-02-28	NC	8 / 99%	☆ ╧ ╤ T29867A, G29868A, C29870A
EPI_ISL_413486	USA / Washington	2020-03-01	NC	8 / 99%	☆ ╧ ╤ A3406C, C5784T, C23525T,

EPI_ISL_413925	USA / California / San Francisco	2020-03-05	GPCS	8 / 99%	☆ ╧╤ C23185T, A3046G, A16467G,
EPI_ISL_413931	USA / California / San Francisco	2020-03-05	NC	9 / 99%	☆ ╧╤ A3046G, A16467G, G16975T, C23185T,
EPI_ISL_413562	USA / Washington	2020-03-02	NC	10 / 99%	☆ ╧╤ C313del, C9180T, G29864A, T29867A, G29868A
EPI_ISL_413652	USA / Washington	2020-03-05	NC	11 / 99%	☆ ╧╤ T23010C, G29861A, G29864C, T29867A, G29868C, C29870A
EPI_ISL_407193	South Korea / Gyeonggi-do	2020-01-25	NC	4 / 99%	☆ ┤ (C8782T, T28144C, T4402C; G5062T)
EPI_ISL_412870	South Korea/ Seoul	2020-01-30	NC	4 / 99%	☆ ┤
EPI_ISL_413513	South Korea	2020-02-27	NC	4 / 99%	☆ ┤
EPI_ISL_413514	South Korea	2020-02-27	NC	4 / 99%	☆ ┤
EPI_ISL_413515	South Korea	2020-02-27	NC	4 / 99%	☆ ┤
EPI_ISL_413516	South Korea	2020-02-27	NC	4 / 99%	☆ ┤
EPI_ISL_413518	China / Beijing	2020-01-26	NC	4 / 99%	☆ ┤
EPI_ISL_413519	China / Beijing	2020-01-28	NC	4 / 99%	☆ ┤
EPI_ISL_413521	China / Beijing	2020-01-28	NC	4 / 99%	☆ ┤
EPI_ISL_413520	China / Beijing	2020-01-28	NC	5 / 99%	☆ ┤ A29301T
EPI_ISL_412871	South Korea / Seoul	2020-01-31	NC	6 / 99%	☆ ┤ C1779T, C15017T
EPI_ISL_410718	Queensland / Gold Coast	2020-02-05	NC	4 / 99%	☆ ╠ (C8782T, T28144C, G28878A; G29742A)
EPI_ISL_411954	USA / California	2020-02-06	NC	4 / 99%	☆ ╠
EPI_ISL_413853	China / Guangdong	2020-01-30	NC	4 / 99%	☆ ╠
EPI_ISL_410717	Australia / Queensland / Gold Coast	2020-02-05	NC	5 / 99%	☆ ╠ T28144C
EPI_ISL_412978	China / Hubei / Wuhan	2020-01-17	NC	4 / 99%	☆ C12141A, C23816T
EPI_ISL_413711	China	2020-02	NC	4 / 99%	☆ C6501T, C16887T

EPI_ISL_413523	India / Kerala	2020-01-31	China	6 / 99%	☆ A1691G, C6501T, C16877T, C24351T
EPI_ISL_413749	China	2020-02	NC	4 / 99%	☆ C14768T, A17805T
EPI_ISL_413858	China / Guangdong	2020-01-30	NC	4 / 99%	☆ A27749N, G27750N
EPI_ISL_412980	China / Hubei / Wuhan	2020-01-18	NC	5 / 99%	☆ T18996C, C24370T, T29029C
EPI_ISL_407071	Europe / England	2020-01-29	NC	5 / 99%	☆ T22586Y; T23605G; T28144C
EPI_ISL_407073	Europe / England	2020-01-29	NC	5 / 99%	☆ T23605G; T18488C, A29596G
EPI_ISL_412982	China / Hubei / Wuhan	2020-02-07	NC	5 / 99%	☆ G5657A, A23403G, A25725G,
EPI_ISL_413697	China	2020-02	NC	5 / 99%	☆ C207T, T946C, A11430G
EPI_ISL_413751	China	2020-02	NC	5 / 99%	☆ TTT27792-27794del
EPI_ISL_413761	China	2020-02	NC	5 / 99%	☆ C207T, T946C, A11430G
EPI_ISL_408484	China / Sichuan / Chengdu	2020-01-15	NC	6 / 99%	☆ ▌T19190A; C24034T
EPI_ISL_406034	USA / California / Los Angeles	2020-01-23	NC	7 / 99%	☆ ▌G1548A; C24034, A28792T
EPI_ISL_410045	USA / Illinois	2020-01-28	NC	7 / 99%	☆ ▌T490A; C3177T, C24034T;
EPI_ISL_412028	Hong Kong	2020-01-22	NC	7 / 99%	☆ ▌C1663T, G22661T, G29862T
EPI_ISL_408668	Vietnam / Thanh Hoa	2020-01-24	NC	11 / 99%	☆ ▌A27T; C28del; C24034T; T29858C; G29861C; G29864del; T29867A
EPI_ISL_407896	Queensland / Gold Coast	2020-01-30	NC	7 / 99%	☆ A21949M; C24790T; C25587T; G28878A; G29742A

EPI_ISL_412873	South Korea / Chungcheongnam-do	2020-02-06	NC	7 / 99%	☆ T3086C, C6255T, G11083T, G17122A, A29871G
EPI_ISL_413791	China	2020-02	NC	7 / 99%	☆ C207T, T946C, A11430G, A16474G, C25000A
EPI_ISL_404253	USA / Illinois / Chicago	2020-01-21	NC	8 / 99%	T490W; C3177Y; C24034Y; T26729Y; G28077Y; C28854Y
EPI_ISL_412869	South Korea /Seoul	2020-01-30	NC	8 / 99%	☆ A1740C, C8782T, C17104T, G26167T, G29593A, A29869G
EPI_ISL_412983	China / Hubei / Tianmen	2020-02-08	NC	9 / 99%	☆ A3175G, G3179A, C14422T, C14585T, G23405C, C28315T, T29680K
EPI_ISL_413485	China / Anhui / Suzhou	2020-01-24	NC	9 / 99%	☆ A4T, C2189T, T3086C, A5094G, G11083del, C16049T, G17122A,
EPI_ISL_407894	Queensland / Gold Coast	2020-01-28	NC	11 / 99%	☆ A6604R; C13681M; A13682M; C13684M; T13686K; G13687K; A13693W; G28878A; G29742A

☆, C8782T, T28144C; ╣, C29095T; ╧ C18060T; ╤ C17747T, A17858G; ┥, T4402C; G5062T; ╠, G28878A; G29742A; ▌, T26729C;

NC, Not confirmed;

469 **Table 3. Mutational analysis of genome sequences in super-transmitter cluster 2.**
470

Accession ID	Location	Collection date	Origin	Number of mutations / % of sequence homology with reference sequence	Nucleotide changes
EPI_ISL_408977	Australia / Sydney	2020-01-25	NC	1 / 99%	○ (G26144T)
EPI_ISL_406036	USA / California	2020-01-22	NC	2 / 99%	○ C17000T
EPI_ISL_412029	Hong Kong	2020-01-30	NC	2 / 99%	○ T13929C
EPI_ISL_413863	China / Guangdong	2020-02-01	NC	2 / 99%	○ C22787G
EPI_ISL_406596	France / Paris	2020-01-23	NC	2 / 99%	○ ▽ (G26144T, G22661T)
EPI_ISL_406597	France / Paris	2020-01-23	WH	2 / 99%	○ ▽
EPI_ISL_410720	France / Paris	2020-01-23	NC	2 / 99%	○ ▽
EPI_ISL_411219	France / Paris	2020-01-28	NC	2 / 99%	○ ▽
EPI_ISL_411220	France / Paris	2020-01-28	NC	2 / 99%	○ ▽
EPI_ISL_410713	Singapore	2020-01-27	NC	2 / 99%	○ C28849T
EPI_ISL_410714	Singapore	2020-02-03	NC	2 / 99%	○ C21859T
EPI_ISL_410536	Singapore	2020-02-06	NC	2 / 99%	○ ▼(G26144T, C21859T)
EPI_ISL_410715	Singapore	2020-02-04	NC	2 / 99%	○ ▼
EPI_ISL_410716	Singapore	2020-02-04	NC	2 / 99%	○ ▼
EPI_ISL_410546	Italy / Rome	2020-01-31	HB	2 / 99%	○ ◇ (G26144T, G11083T)
EPI_ISL_412974	Italy / Rome	2020-01-29	NC	2 / 99%	○ ◇
EPI_ISL_410545	Italy / Rome	2020-01-29	HB	3 / 99%	○ A2269T; G11083N;
EPI_ISL_413603	Finland / Helsinki	2020-03-03	NC	4 / 99%	○ ◇ C14805T, G29405C
EPI_ISL_413016	Brazil / Sao Paulo	2020-02-28	Italy	5 / 99%	○ ◇ C2388T, C14805T, T17247C,
EPI_ISL_413019	Switzerland / Zurich	2020-02-26	NC	9 / 99%	○ ◇, G11084TTTin, C14805T, T17247C, C24378T, C26894T
EPI_ISL_413025	USA / Washington	2020-02-27	NC	12 / 99%	○ ◇, A35T, C36T, T2446C, C3411T, G5572T, C14805T, G29864A, T29867A, G29868A, C29870A

EPI_ISL_406031	Taiwan / Kaohsiung	2020-01-23	NC	4 / 99%	○ G16188T; A25964G; 29877Tin
EPI_ISL_413018	South Korea	2020-02-06	NC	4 / 99%	○ A2707G, G26640T, T26677C
EPI_ISL_412116	England	2020-02-09	NC	5 / 99%	○ A2470G, C2558T, G11083N, C14805T,
EPI_ISL_413017	South Korea	2020-02-06	NC	6 / 99%	○ T4402C, G5062T, G26640T, T26677C, T28144C
EPI_ISL_411951	Sweden	2020-02-07	NC	7 / 99%	○ G2717A; A9274G; C13225G; T13226C; A17376G; T23952G;
EPI_ISL_411929	South Korea	2020-01	WH	9 / 99%	○ G2971T; C6031T; C12115T; T15597C; C20936T; C22224G; G25775T; T26354A
EPI_ISL_406844	Australia / Victoria	2020-01-25	NC	13 / 99%	○ T19065C; T22303G; 29750-29759del

471 ○, G26144T; ▽, G22661T; ▼, C21859T; ◇, G11083T;

472

473 **Table 4. Mutational analysis of genome sequences in super-transmitter cluster 3.**
474

Accession ID	Location	Origin	Collection date	Number of mutations / % of sequence homology with reference sequence	Nucleotide changes
EPI_ISL_408481	China / Chongqing	NC	2020-01-18	1 / 99%	⊙ (G11083T)
EPI_ISL_407988	Singapore	NC	2020-02-01	1 / 99%	⊙
EPI_ISL_412968	Japan	NC	2020-02-10	1 / 99%	⊙
EPI_ISL_410546	Italy / Rome	HB	2020-01-31	2 / 99%	⊙ G26144T
EPI_ISL_412030	Hong Kong	NC	2020-02-01	2 / 99%	⊙ G29841A
EPI_ISL_412969	Japan	NC	2020-02-10	2 / 99%	⊙ C29635T
EPI_ISL_412974	Italy / Rome	NC	2020-01-29	2 / 99%	⊙ G26144T
EPI_ISL_408430	France / Paris	NC	2020-01-29	3 / 99%	⊙ ◎ (G11083T, C1190T, C9438T)
EPI_ISL_410984	France / Paris	NC	2020-01-29	3 / 99%	⊙ ◎
EPI_ISL_411218	France / Paris	NC	2020-02-02	3 / 99%	⊙ ◎
EPI_ISL_408480	China / Yunnan / Kunming	NC	2020-01-17	3 / 99%	⊙ ♁ (G11083T, C8782T, T28144C)
EPI_ISL_406223	USA / Arizona / Phoenix	NC	2020-01-22	4 / 99%	⊙ ♁ C29095T
EPI_ISL_412873	South Korea	NC	2020-02-06	7 / 99%	⊙ ♁ T3086C, C6255T, G17122A, A29871G
EPI_ISL_413485	China / Anhui / Suzhou	NC	2020-01-24	9 / 99%	⊙ ♁ A4T, C2189T, T3086C, A5094G, C16049T, G17122A
EPI_ISL_413603	Finland / Helsinki	NC	2020-03-03	4 / 99%	⊙ $ G26144T, G29405C
EPI_ISL_413016	Brazil / Sao Paulo	Italy	2020-02-28	5 / 99%	⊙ $ C2388T, T17247C, G26144T
EPI_ISL_413025	USA / Washington	NC	2020-02-27	12 / 99%	⊙ $ A35T, C36T, T2446C, C3411T, G5572T, G26144T, G29864A, T29867A, G29868A, C29870A
EPI_ISL_413214	Australia Sydney	NC	2020-02-29	5 / 99%	⊙ % G29374A

EPI_ISL_412975	Australia Sydney	Iran	2020-02-28	6 / 99%	⊙ % G4255A, , A20047G
EPI_ISL_413213	Australia / Sydney	Iran	2020-02-29	7 / 99%	⊙ % C884T, G8653T, C24704T
EPI_ISL_408482	Shandong / Qingdao	NC	2020-01-19	7 / 99%	⊙ % C7299T; C27612G; T28688C
EPI_ISL_413589	Netherlands / Utrecht	NC	2020-03-01	8 / 99%	⊙ C241T, G2527T, C3037T, C6428T, C14408T, A23403G, A25575C

475

476 ⊙, G11083T; ◎, C1190T, C9438T; ♁, C8782T, T28144C; $, C14805T; %, G1397A, T28688C, G29742T.

477 **Table 5. Mutational analysis of genome sequences in super-transmitter cluster 4.**
478

Accession ID	Location	Origin	Collection date	Number of mutations / % of sequence homology with reference sequence	Nucleotide changes
EPI_ISL_406862	Germany / Bavaria / Munich	NC	2020-01-28	3 / 99%	△ (C241T, C3037T, A23403G)
EPI_ISL_413555	United Kingdom / Wales	NC	2020-02-27	4 / 99%	△ # (C241T, C3037T, A23403G, C14408T)
EPI_ISL_413566	Netherlands / Blaricum	NC	2020-03-02	4 / 99%	△ #
EPI_ISL_413591	Netherlands / Zeewolde	NC	2020-03-02	4 / 99%	△ #
EPI_ISL_413593	Luxembourg	NC	2020-02-29	5 / 99%	△ # C23575T
EPI_ISL_413648	Portugal	Spain	2020-03-01	5 / 99%	△ # C29144T
EPI_ISL_412973	Italy	NC	2020-02-20	6 / 99%	△ # T29867N, G29868N
EPI_ISL_413489	Italy / Milan	NC	2020-03-03	8 / 99%	△ # A187G, A6956C, T29867N, G29868N
EPI_ISL_413602	Finland / Helsinki	NC	2020-03-03	6 / 99%	△ # G22865T, C29585T
EPI_ISL_413572	Netherlands / Haarlem	NC	2020-03-01	7 / 99%	△ # T1666C, C3037T, G25563T
EPI_ISL_413589	Netherlands / Utrecht	NC	2020-03-01	8 / 99%	△ # G2527T, C6428T, G11083T, A25575C
EPI_ISL_413022	Switzerland / Zurich	NC	2020-02-29	7 / 99%	△ # & (C241T, C3037T, A23403G, C14408T, G28881A, G28882A, G28883C)
EPI_ISL_413579	Netherlands / Nootdorp	NC	2020-03-03	7 / 99%	△ # &
EPI_ISL_413587	Netherlands / Tilburg	NC	2020-03-03	7 / 99%	△ # &
EPI_ISL_412912	Germany/Baden-Wuerttemberg	Italy	2020-02-25	8 / 99%	△ # & G10265A
EPI_ISL_413584	Netherlands / Rotterdam	NC	2020-03-03	8 / 99%	△ # & C27046T
EPI_ISL_413647	Portugal	Germany	2020-03-01	8 / 99%	△ # & C27046T
EPI_ISL_413604	Finland / Helsinki	NC	2020-03-03	9 / 99%	△ # & C27046T, T29807C
EPI_ISL_413023	Switzerland / Zurich	NC	2020-02-29	9 / 99%	△ # & A22168C,

EPI_ISL_412972	Mexico / Mexico City	Italy	2020-02-27	10 / 99%	△ # & C13206G, A15807del, G24268del,

△, C241T, C3037T, A23403G; #, C14408T; &, C14408T, G28881A, G28882A, G28883C.

482 **Table 6. Distribution of different types of SARS-CoV-2 over the world after February 29, 2020.**

Types of viruses	Numbers of genomes (%)	Major areas	Locations / numbers of genomes (%)						
			Africa	**Asia**	**Europe**	**NA (USA)**	**NA (Canada)**	**Oceania**	**South America**
Total	1539		20(1)	33(2)	968(63)	441(29)	11(1)	27(2)	23(2)
Original derivatives	173 (11)	Europe (UK and Netherland)	0(0)	0(0)	158(16)	11(3)	2(2)	0(0)	2(9)
ST1	340 (21)	USA, Europe, South America	1(5)	1(3)	30(3)	282(64)	0(0)	5(19)	8(35)
ST2	172 (11)	Europe	1(5)	0(0)	111(12)	9(2)	2(2)	3(11)	0(0)
ST3	132 (9)	Asia, Europe	0(0)	17(52)	119(12)	12(3)	5(5)	15(56)	0(0)
ST4	856 (56)	Africa, Asia, Europe, Oceania	18(90)	16(48)	663(69)	137(14)	4(4)	5(19)	12(52)

483

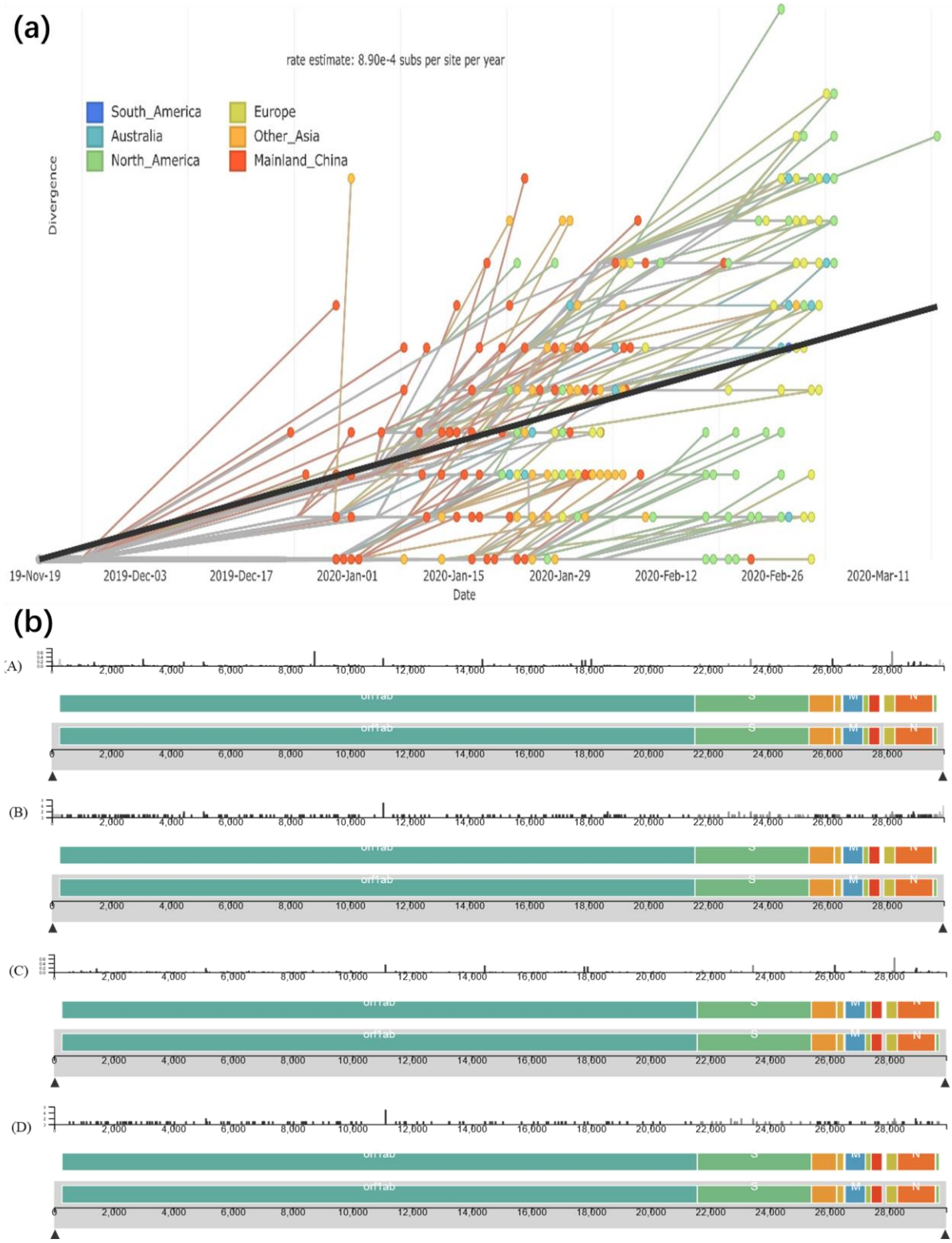


Figure 1. Root-to-tip regression scatter plots and distribution of mutations across the SAR-CoV-2 genomes. (a) Root-to-tip regression scatter plots of different strains of SARS-CoV-2. (b) Distribution of mutations across the SAR-CoV-2 genome. (A) and (B) illustrated the entropy and events of nucleotide mutations. (C) and (D) illustrated the entropy and events of amino acid mutations. Figure 1. Root-to-tip regression scatter plots of different strains of SARS-CoV-2. Dots in the plot indicate the SARS-CoV-2 isolates used in this study. The color

491 of each dot represents the region of isolation of the corresponding isolate. Ancestral state
492 reconstruction and branch length timing were performed with TreeTime [16].

493

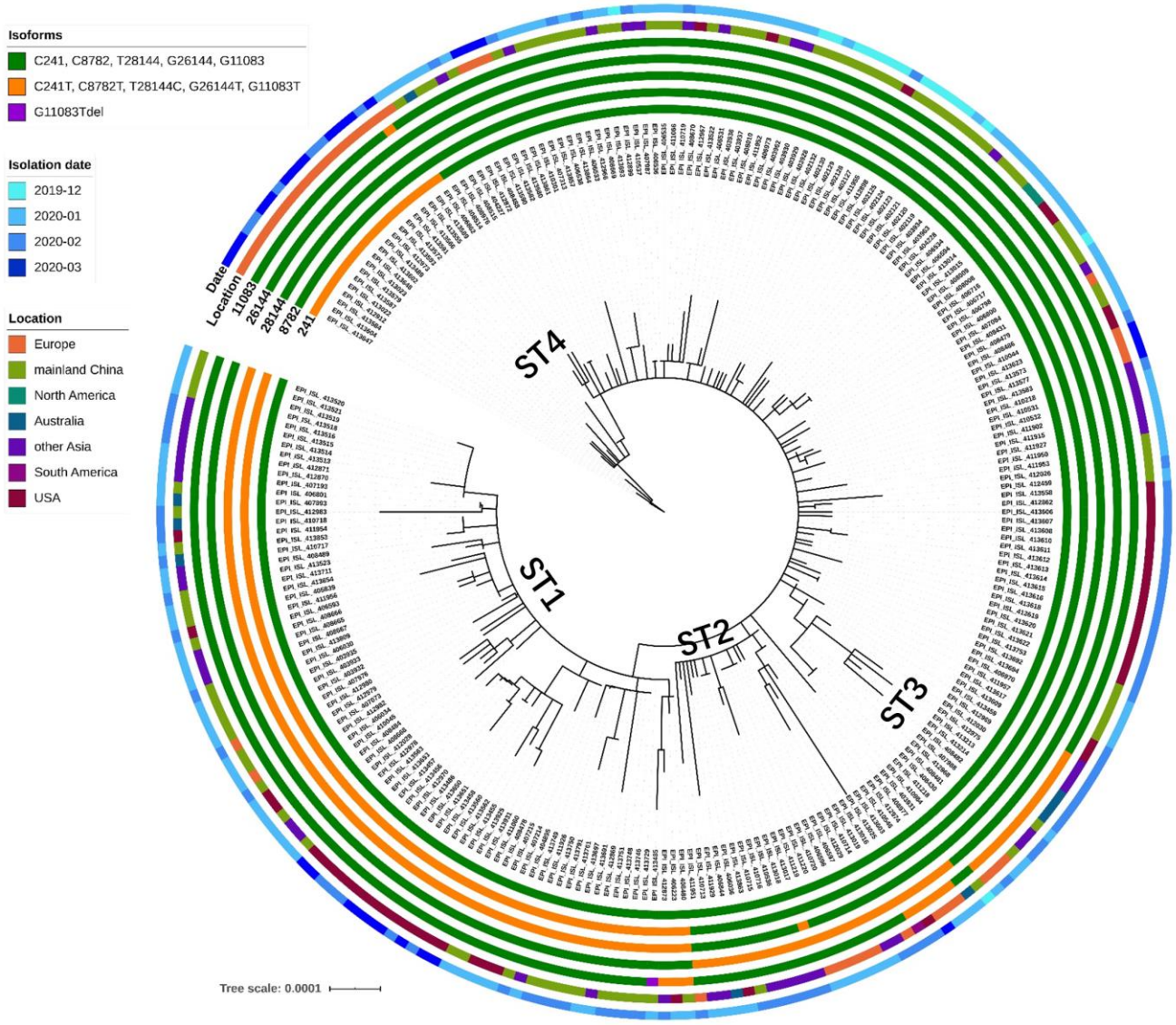


Figure 2. Phylogentic analysis of SARS-CoV-2 genomes. Four super-transmitter clusters (STs) were identified. Each ST was found to exhibit a signature mutation profile.

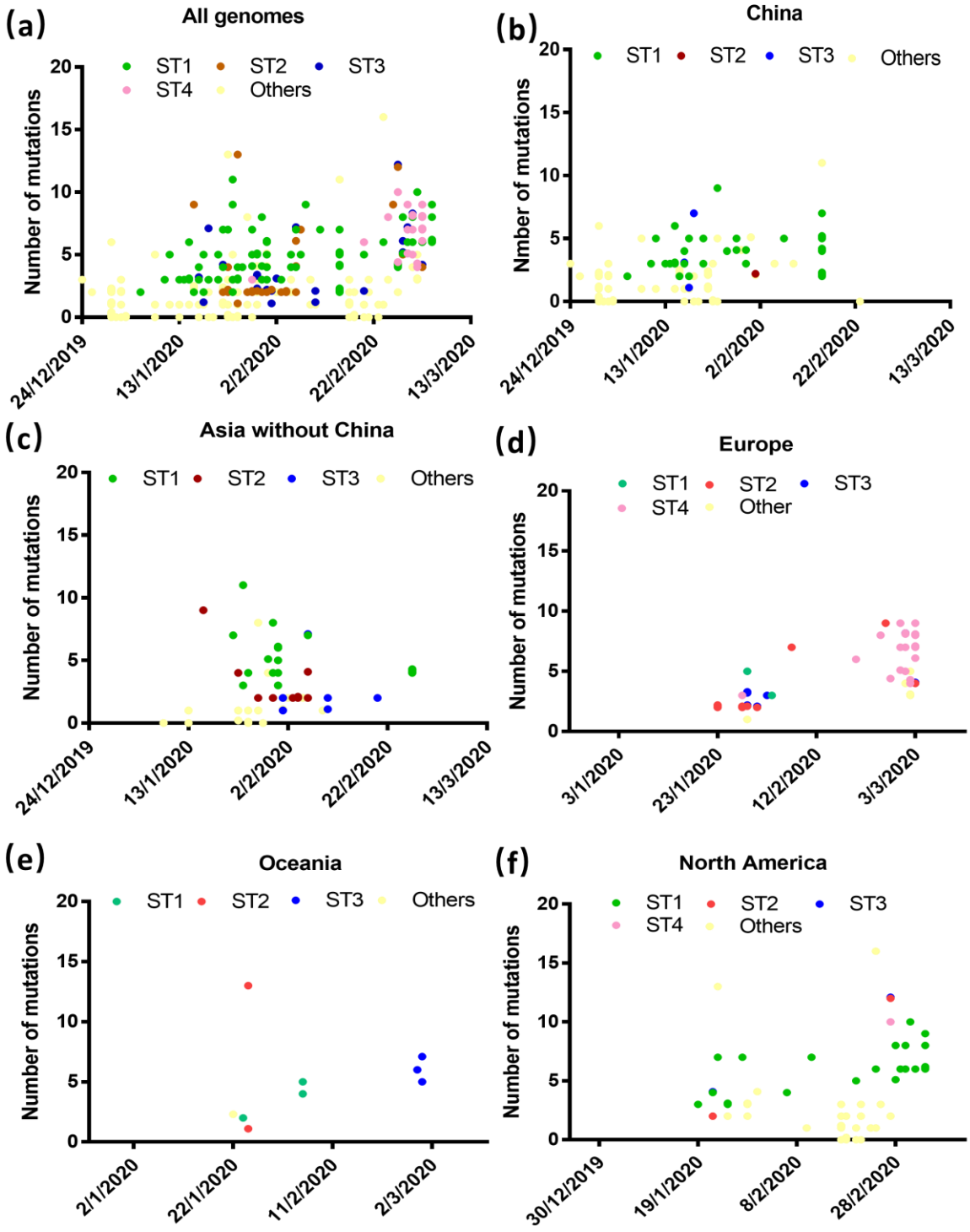
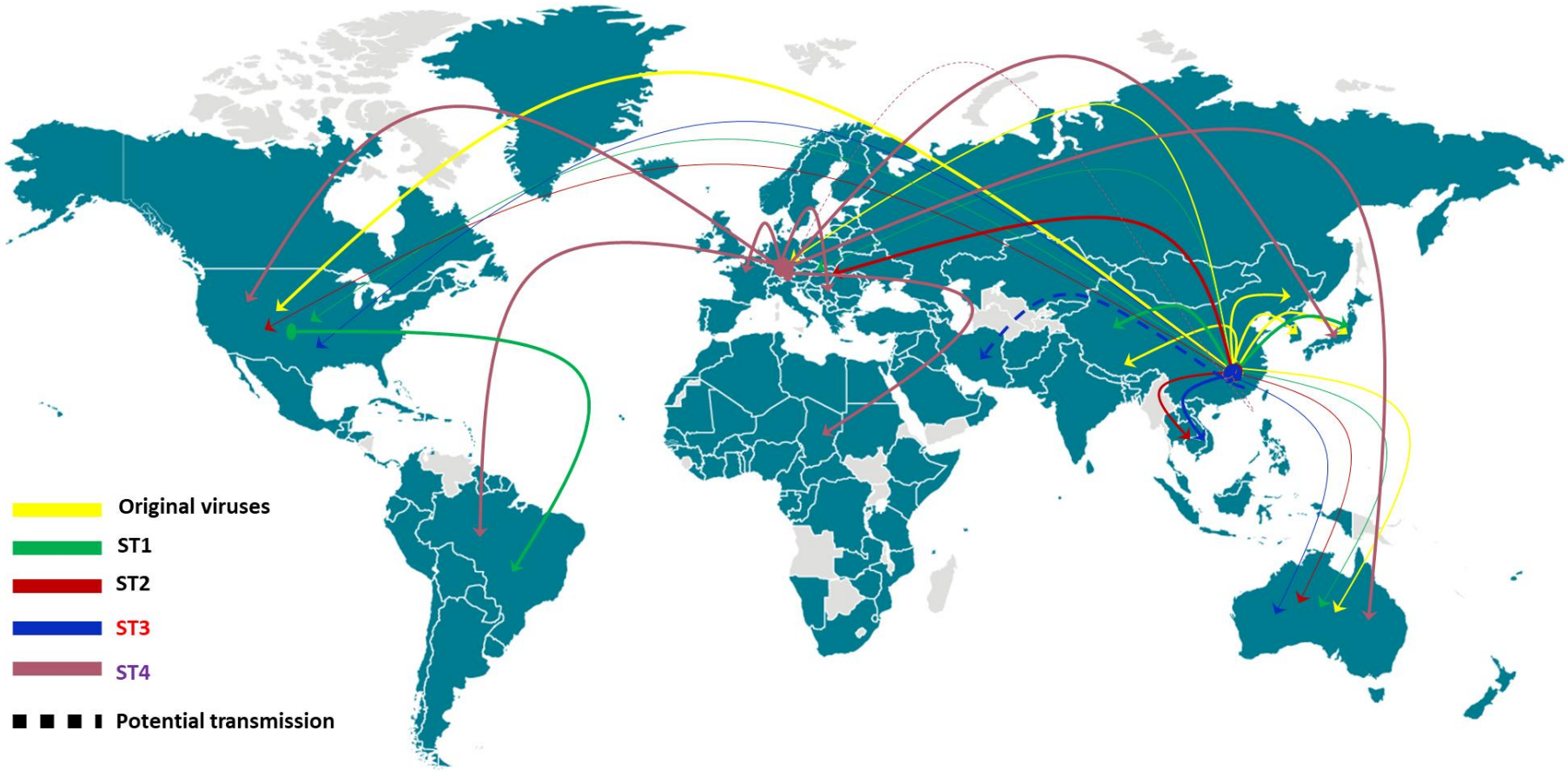


Figure 3. Changes in the distribution pattern and mutation rate of different super-transmitter clusters in various continents over time. Distribution of different STs and their mutations (a) Overall, (b) in China, (c), in Asian countries excluding China, (d) in Europe, (e) in Oceania, and (f) in North America. Two genomes with over 20 mutations were not included to facilitate easy visualization of the graphs.



505

506 **Figure 4. Transmission of super-transmitters and other derivatives of the original SARS-CoV-2 in different areas of world.** The derivatives
507 of the original virus have been transmitted worldwide and contributed to the early outbreak of COVID-19. ST1 transmits mainly in Asia and the
508 US but was less prevalent in other parts of the world. ST2 and ST3 was transmitted mainly in Asian countries other than China, as well as Europe
509 from mid of January to mid of February. ST4 was transmits mainly in Europe in the beginning and then transmitted to all over the world.