Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The dominant strain of SARS-CoV-2 is a mosaicism

Wei Wang 1, Cheng-Peng Li 1, Mei He, Sheng-Wen Li, Lin Cao, Nai-Zheng Ding, Cheng-Qiang He *  
Shandong Provincial Key Laboratory of Animal Resistance Biology, College of Life Science, Shandong Normal University, Jinan, Shandong 250014, China

A R T I C L E   I N F O  
Keywords: COVID-19  
SARS-CoV-2  
Homologous recombination  
Dominant strain

A B S T R A C T

COVID-19 is seriously threatening human health all over the world. A comprehensive understanding of the genetic mechanisms driving the rapid evolution of its pathogen (SARS-CoV-2) is the key to controlling this pandemic. In this study, by comparing the entire genome sequences of SARS-CoV-2 isolates from Asia, Europe and America, and analyzing their phylogenetic histories, we found a lineage derived from a recombination event that likely occurred before March 2020. More importantly, the recombinant offspring has become the dominant strain responsible for more than one-third of the global cases in the pandemic. These results indicated that the recombination might have played a key role in the pandemic of the virus.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, also known as 2019-nCoV) is the pathogen of the coronavirus disease 2019 (COVID-19). First noticed in Wuhan, China in December 2019, the virus soon hit the entire globe badly. The pandemic crisis is even worsening now. As of August 2021, the virus has affected more than 220 countries, areas or territories, caused more than 209 million infections, and killed more than 4.4 million people (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). To control SARS-CoV-2, scientists are concentrating on characterizing the virus and its replication dynamics (Abuin et al., 2020), tracking its movement through the human population (Lee et al., 2020; Worobey et al., 2020), exploring virus origin (Zhou et al., 2020), and developing vaccines (Sharma et al., 2020). The results of these works are based on the genetics and evolution of the virus. Continuous evolution of SARS-CoV-2 leads to the emergence of new variants, which has brought challenges to vaccine development and other control measures against the virus (Khurshid et al., 2020). Therefore, a comprehensive understanding of the genetic mechanisms underlying its evolution is the basic issue of controlling the virus. It is unknown yet whether homologous recombination, an important genetic mechanism, influences SARS-CoV-2 evolution. Here, we reported the discovery of a dominant SARS-CoV-2 lineage with a mosaic genome, revealing that homologous recombination is a notable evolutionary power of the virus.

Results and discussion In order to explore the origin of SARS-CoV-2 strains circulating in China and its neighboring countries in the first half of this year, we selected and analyzed their genome sequences deposited in the Genbank database. When comparing their genomic sequences with those of isolates from Germany and the United States (US), it was found that the genomes of a group from Bangladesh exhibited mosaic characteristic (Fig. 1A). Taking the nucleotide position 7434 of the viral genome as the boundary, the Bangladeshi group (e.g., NIB-1) shared higher genome sequence similarity with one US group (e.g., UNC_200428) forward, but with the German group (e.g., NRW-04) afterward. We also compared all variable genomic sites of three representatives of the Bangladeshi group with those of their putative parents (Fig. 1B). The inflexion of sequence similarity change was clearer (Fig. 1C). According to Fisher’s exact statistics, the putative recombination breakpoint with the maximum chi-square value was located at the region around the position 7000. Delimited by the putative breakpoint, there was significant difference in the similarity of the recombinant virus to the two parents (p < 0.01).  

Phylogenetic reconstruction based on the representatives of these virus groups (listed in Table 1) showed that they constituted three parallel lineages (Fig. 1D). Lineage I covers isolates from China, Germany, and the US, with Wuhan-Hu-1 collected in December 2019 being in an ancestral status, lineage II is mainly endemic to the US, and lineage III is composed of the mosaic isolates. Thus, one parent is a member of lineage I type and the other from lineage II type. Further, to figure out when and where this recombination event took place, a recombinant isolate was used as the query to perform BLAST in GenBank to seek its sisters. It was discovered that many isolates collected in March 2020 were clustered with the Bangladeshi group into lineage III. Notably,
Virus Research 305 (2021) 198553

some US isolates were in the ancestral status of this mono-phylogenetic lineage (Fig. 1D). Given that both lineages I and II were prevalent in the US early in the virus outbreak, it is more likely that the recombination event occurred in the US before March 2020.

The discordant phylogenetic pattern of a virus genome is a golden indicator for judging the origins of the recombination. To further confirm this recombination event, different genomic parts of the recombinant and parent lineages were used to reconstruct their phylogenetic histories through Maximum Likelihood and Neighbor-Joining methods. For each tree, the two methods obtained the same topology except for the different bootstrap values (Figs. 1 and S1). As expected, the recombinant viruses fell into lineage II before the putative breakpoint (Fig. 1E), but were clustered with lineage I as a mono-phylogenetic group after the breakpoint (Fig. 1F). Each mono-phylogenetic group with the recombinant lineage was supported by the robust bootstrap value. Thus, the mosaic genome indeed has double origins involving the I- and II-type viruses.

To understand the impact of the recombination event on the pandemic of SARS-CoV-2, we also analyzed the virus information deposited in the GISAID database (https://www.gisaid.org/). Based on the phylogenetic history, the SARS-CoV-2 isolates are divided into 8 clades in the database: S, O, L, V, G, GH, GR, and GV (Fig. S2). Among them, GR is the largest clade accounting for more than 30% of the available isolates in total, 65% in South America, 56% in Asia, 54% in Oceania, 36% in Africa, 30% in Europe, and 12% in North America (Fig. 2). Therefore, the GR group should be the dominant strain in the pandemic. By December 2020, GR had differentiated into a new sublineage GRY (Fig. S4). Interestingly, we found that these GR isolates are the offspring of the recombinant virus (Figs. S3 and S4), indicating that the recombination might have played a key role in the SARS-CoV-2 pandemic.

High mutation rates are generally considered deleterious for RNA virus with asexual reproduction (Chao, 1990). A decrease in the mean fitness of its population is continually driven by the evolutionary mechanism known as Muller’s ratchet. This is where the load of deleterious mutations increases in a ratchet-like manner with successive loss of the fittest viruses (Donis, 1991; Muller, 1964). As a form of sexual reproduction, genetic recombination is thought the power to stop Muller’s ratchet in asexual reproduction biology (Naito and Pawlowska, 2010). This process enables some viruses to acquire key adaptive mutations to fill a major fitness gap in a single step. Usually, a dominant strain represents the highly adaptable variant. The recombinant SARS-CoV-2 has become the largest dominant strain, suggesting that the recombination might have conferred high adaptability on the virus.

For coronavirus, its replication is mediated by the polymerase Nsp12 without proofreading function (Robson et al., 2020). Although its non-structure protein Nsp14 has a limited proofreading function (Robson et al., 2020), its transcrip-
to carefully exclude the recombination event when molecular epidemiology of COVID-19 is surveyed using SARS-CoV-2 genome sequences.

Recombination is an important evolution power of CoVs. Genetic recombination has been previously documented for different CoVs, including SARS-CoV (Lau et al., 2010) and MERS-CoV (Sabir et al., 2016). For the occurrence of recombination, co-infection of SARS-CoV-2 in a host cell is necessary. Long-term existence in COVID-19 patients (Wang et al., 2020) and high prevalence of asymptomatic infection of SARS-CoV-2 might increase the chance of co-infection. To date, virus co-infection has not been reported in the COVID-19 patients; however, re-infection has been found repeatedly. The emergence of the recombinant lineage suggested that a COVID-19 patient might be subjected to co-infection of two distinct SARS-CoV-2 strains.

Currently, there is an urgent need to develop SARS-CoV-2 vaccines to control the ongoing pandemic. However, recombination between viruses might result in their antigen shift and fitness change in the host (Lowen, 2017; Ludwig-Begall et al., 2020). Thus, the recombination between SARS-CoV-2 strains might bring a challenge to the development of effective vaccines against the virus.

In all, our analyses demonstrated that SARS-CoV-2 can undergo

### Table 1
SARS-CoV-2 representative used in this study.

| Access_number | Isolate     | Isolation_source | Country | Collection_date |
|---------------|-------------|------------------|---------|-----------------|
| MT646074      | MD-HP0006   | Human            | USA     | 2020-03         |
| MT551604      | UNC_200459  | Human            | USA     | 2020-04         |
| MT397929      | UF-12       | Environment      | USA     | 2020-04         |
| MT345876      | ID-UW-4378  | Human            | USA     | 2020-04         |
| MT477855      | AK151       | Human            | USA     | 2020-04         |
| MT321950      | LA-EVTLD1   | Human            | USA     | 2020-04         |
| MT565496      | UNC_200407  | Human            | USA     | 2020-04         |
| MT565497      | UNC_200428  | Human            | USA     | 2020-04         |
| MT654999      | UNC_200465  | Human            | USA     | 2020-04         |
| MT654988      | UNC_200431  | Human            | USA     | 2020-04         |
| MT397929      | UF-12       | Environment      | USA     | 2020-04         |
| MT432195      | LA-EVTLD1   | Human            | USA     | 2020-04         |
| MT477855      | AK151       | Human            | USA     | 2020-04         |
| MT345876      | ID-UW-4378  | Human            | USA     | 2020-04         |
| MT551604      | UNC_200459  | Human            | USA     | 2020-04         |

Fig. 2. The distribution of recombinant offspring in each continent n: available SARS-CoV-2 genome number in GISAID.
recombination during its natural infection and circulation. The dominant strain of the virus is mosaic, suggesting that homologous recombination might have played a key role in the pandemic. And also, SARS-CoV-2 may adopt recombination for rapid evolution, leading to the emergence of novel variants with higher adaption, which we should be aware of and pay more attention to.

2. Materials and methods

SARS-CoV-2 genomes were collected from GenBank and GISAID, and aligned with CLUSTAL W. The representatives of the recombinant and parent lineages were listed in Table 1. Phylogenetic histories were reconstructed employing Maximum Likelihood and Neighbor-Joining methods implemented in MEGA X with the optimal substitution models and rates among sites (Kumar et al., 2018). The robustness of lineage was tested by bootstrap method with 1000 replicates. A lineage models and rates among sites (Kumar et al., 2018). The robustness of similarity comparison and similarity graphical representation were

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.virusres.2021.198553.

References

Abuin, P., Anderson, A., Ferramosca, A., Hernandez-Vargas, E.A., Gonzalez, A.H., 2020. Characterization of SARS-CoV-2 dynamics in the host. Annu. Rev. Control 50, 457–468. https://doi.org/10.1016/j.jacrcontrol.2020.09.008.

Chao, L., 1990. Fitness of RNA virus decreased by Muller’s ratchet. Nature 348, 454–455.