Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rapid spread of mutant alleles in worldwide COVID-19 strains revealed by genome-wide SNP analysis

Zhenglin Zhu (zhuzl@cqu.edu.cn)
Chongqing University
Gexin Liu
Chongqing University
Kaiwen Meng
China Agricultural University
Liuqing Yang
Chongqing Occupational Disease Prevention Hospital
Geng Meng (mg@cau.edu.cn)
China Agricultural University

Research Article

Keywords: Coronavirus, SNP, COVID-19, allele

Posted Date: April 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23205/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

The novel coronavirus (COVID-19) has become a pandemic and is threatening human health globally. Here, we report 14 newly evolved COVID-19 single nucleotide polymorphism (SNP) alleles those underwent a rapid increase (12 cases) or decrease (2 cases) in their frequency from between 10% and 50% in the last three months. The 14 SNPs are mostly (13/14) located in the coding region and are mainly (9/14) nonsynonymous substitutions. Out of the 14 SNPs, 12 SNPs showed a complete linkage in SNP pairs and clustered into 4 linkage groups, named LG_1 to LG_4. SNPs located in 514 and 27046 are independent events. Analyses in population genetics show that the increases in the new alleles result from genetic differentiation between Europe and America. We found that the mutants in LG_1 are driven by balancing selection and arose rapidly in Europe but not in America. The mutants in LG_2 and LG_3, also driven by balancing selection, arose rapidly in American but not in European strains. Based on analysis of geographic COVID-19 cases worldwide, we found that the mutants in LG_1 positively correlate the fatality rate of COVID-19 while those in LG_2 and LG_3 negatively correlate with the fatality rate. The correlations are statistically significant, suggesting that the virus strains possessing mutants in LG_1 are more aggressive, while those in LG_2 and LG_3 are in opposite. Further analysis revealed that mutants in LG_1 have been identified more frequently in European strains than in American strains, while mutants in LG_2 and LG_3 have been found more frequently in American strains. This may partially explain the higher fatality rates of COVID-19 infection in Italy, England and France, compared with the United States. These findings should be instructive for epidemiological surveys and disease control of COVID-19 in the future.

Introduction

COVID-19, also named 2019-nCoV, is a novel coronavirus that causes novel coronavirus pneumonia (NCP). The rapid spread of COVID-19 has become a global threat, since its first identification in Wuhan City, China, last December (Ralph, et al. 2020). To date, there have been more than one million confirmed NCP cases and more than 70000 cases have resulted in deaths around the world. To control the COVID-19 epidemic, research on the genomic epidemiology of the virus is important for the prediction of global evolutionary trends. Furthermore, the identification of the diversification in patterns and selection signatures in the COVID-19 genome (Lu, et al. 2020) during the evolution of the virus is essential for the early diagnosis and control of this disease.

Although the origin of the virus is still a mystery, COVID-19 has displayed higher divergence in genomic sequence to its possible origins (Bat-CoV-RaTG13, or Pangolin-CoV-2019)(Lam, et al. 2020) than previously anticipated (Tang, et al. 2020). The genome of COVID-19 is undergoing continuous evolution. Just two months after the virus was first reported (Lu, et al. 2020; Ralph, et al. 2020), there have been more than 100 substitution sites identified in COVID-19's protein coding region. Most of these mutations are located in the coding region of polyprotein 1ab (pp1ab, ORF1)(Namy, et al. 2006) and structural proteins (Fehr and Perlman 2015). It has also been reported that COVID-19 is undergoing active recombination (Yi 2020) , which is a common event among RNA viruses. A previous study has suggested

that the virus has evolved into two subtypes (L and S) classified by two complete linked single nucleotide polymorphisms (SNPs) at genome locations 8792 and 28144 (Tang, et al. 2020). SNP 29144 leads to an amino acid change from LEU (L) to SER (S) in ORF8, which is supposed to be related to viral replication (Muth, et al. 2018). In the study (Tang, et al. 2020), it was predicted that S is less aggressive but more adaptive than L and may increase in frequency in the future (Tang, et al. 2020). With more COVID-19 genomes sequenced and deposited, we are now able to reevaluate the performance of the polymorphic alleles.

Methods

Identification of rapidly changing mutants

We collected genomic sequences and related information of coronavirus from GISAID (www.gisaid.org), NCBI, CoVdb (Zhu, et al. 2020) and ViralZone (Hulo, et al. 2011). The whole genome alignments were calculated using CUDA ClustalW (Hung, et al. 2015). To assess genomic differences between months, we grouped coronavirus genomes into three groups according to their collection dates. We combined the genomes collected in December, 2019 and January 2020, considering that threre are few samples in 2019 (sixteen cases). Based on this approach, the number of the samples in December-January, Febuary and March were 202, 299 and 463, respectively. Using libraries in BioPerl, we wrote scripts to extract mutations with a change in frequency by near or more than 10% between any two months. Finally we performed a manual check and identified 14 cases complying with the request. The significance of the change in frequency was evaluated using the chi square test by R.

Linkage disequilibrium analysis

According to the published algorithms (Morton 1955; Lewontin 1964), we counted the possibility of coupling and repulsion gametes. The coupling gametes are alleles on the same chromosome that remain together, while the repulsion gametes are alleles on the same chromosome that are repulsed by each other and pair with alleles on the opposite strand. Then we wrote Perl scripts to calculate D', ρ^2 and LOD, which refer to a normalized basic linkage disequilibrium parameter, a squared correlation coefficient (Lewontin 1964), and a statistical test to infer the likelihood of obtaining the test data if the two loci are indeed linked (Morton 1955), respectively. They all positively correlate with the degree of genetic linkage.

Evolutionary analysis

Using LASTZ (Harris 2007), we performed genome-genome alignments between any two coronavirus strains and outputted the results in AXT format. From the results, we retrieved the corresponding sequences of other strains of one coronavirus gene, and realigned these sequences using MUSCLE (Edgar 2004a, b). To detect selection signals, the sliding widow analysis was used with a window size of 200 bp and a step size of 50 bp. For each sliding window, we calculated the scores of Pi (Watterson 1975) and Tajima's D (Tajima 1989) by VariScan 2.0 (Vilella, et al. 2005; Hutter, et al. 2006). The fixation index (Fst) was calculated according to published algorithms (Fumagalli, et al. 2013). We further

calculated the allele frequencies and used SweepFineder2 (DeGiorgio, et al. 2016) to calculate the composite likelihood ratio (CLR, step size = 50)(Nielsen, et al. 2005; Zhu and Bustamante 2005). The figures were generated using the R libraries “gdata” and “ggplot2”.

To test whether a negative Tajima’s D is biased caused because of a genetic bottleneck, a simple model was built (Figure S2) assuming that the population size of COVID-19 shrunk from N1 = 2558 to N2 = 450 and then expanded to N3 = 12196, based on the number of daily confirmed infections in Febuary 15, 2020, in Asia, in February 24, 2020, in Asia, and in March 16, 2020, in Europe. We assume that on Feb 29, 2020, the infected population started to expand. We used the software ms (Hudson 2002) to generate simulation data according to the model with the parameter “-G 11.189 -eG 0.3 0.0 -eN 0.5 0.2“. We tested the significance by ranking in distribution (Figure S2).

Statistical analysis

Most samples in March are from Europe and America, while most samples in January are from Asia. The countries where COVID-19 was collected were classified into three groups, ‘America’, ‘Europe’ and ‘Asia-Pacific’. ‘America’ mainly refers to the North America. ‘Europe’ refers to the continental Europe. ‘Asia-Pacific’ refers to all regions or countries in Asia and the Pacific Ocean. The fatality rates in European and American countries were calculated based on the data provided by the European Centre for Disease Prevention and Control from March 20, 2020, to March 26, 2020. The daily exponential growth rates (λ) were calculated as $\lambda = \ln[Y(t)]/t$ (Lipsitch, et al. 2003). A two sided test was used to assess the significance of Pearson's product-moment correlation in the correlation analysis.

Results

SNP frequency, linkage disequilibrium and haplotype analyses

Inspired by a previous effort (Tang, et al. 2020), we performed comprehensive analyses on 964 full-length COVID-19 genomes (collected from 2019-12-24 to 2020-03-18, Table S1) with a focus on evolutionary dynamics, selection and gene function. We searched for all SNPs and found SNPs with a monthly identification frequency variation near or higher than 0.1 during the past three months. A total of 986 polymorphic sites were identified, and 14 of those SNPs were highlighted in such cases (Figure 1A, 1B). From January to March, 12 mutants increased from a monthly frequency of 0%~3% to 9.6%~50.6%, while 2 decreased from a monthly frequency of 38% and 38.5% to 26% and 25.7% (Table 1). These changes in frequency are statistically significant (Chisq-Test, *P-value* < 0.05). Linkage disequilibrium analyses show that 12 of the 14 SNPs can be clustered into 4 linkage groups. SNPs in each group showed significant complete ($D' > 0.97$, $\rho^2 > 0.96$) or nearly complete ($D' = 0.8$, $\rho^2 = 0.78$) linkage (Figuure 1C, 1D) with a median LOD score of 146.62 (Table S2). For convenience, we named SNPs using the “SNP_Location” format, while the genome of the strain MN908047 is used as the reference. For example, we use SNP_241 to represent the SNP at location 241. Linkage groups are named using the format of ‘LG_’ plus a serial number (Table 1), from LG_1 to LG_4. We classified all COVID-19 genomes into 16

haplotypes based on the 14 SNPs. From a network view, we observed that there are only two major haplotypes in January but seven major haplotypes in March (Figure S1). This suggests that 5 newly evolved haplotypes became major in the last 2 months. This change is rapid, and is reflected by a significantly high Tajima's D (3.4, *P-value* = 0.002) of the haplotypes in March. The two major haplotypes in January correspond to the two subtypes L and S referred to previously(Tang, et al. 2020), which are now clustered into LG_2 in this study.

Population genetics and evolutionary patterns

We performed sliding window analyses in population genetics to address recent evolutionary patterns in the 14 SNP sites. To avoid oscillation caused by time scale, we only targeted the COVID-19 strains collected in March, 2020. These SNPs are mostly located in the regions with high values in Pi and Tajima's D (Figure 2A). The differentiation and diversity may result from population structure or balancing selection.

In haplotype analysis, we observed an increase in diversity from 0.079 to 0.372. A geographic bias for some major haplotypes or subtypes (Figure S1) was also detected. We further compared the genetic polymorphisms between European and American isolates. The locations of LG_2 and LG_3 are identified in the regions with a significant fixation index (Fst, *P-value* < 0.05, Figure 2B) while the location of LG_1 has an Fst with weak significance (*P-value* < 0.1). These results indicate differentiation between American and European virus in these linkage groups.

We performed population genetic tests on the strains collected from America (151 samples) and Europe (300 samples) separately. A parallel simulation test was performed to assess effects caused by genetic bottleneck on mutation sites showing directional selection signals (for details, see Materials and Methods, Figure S2). The results (Figure 2C, 2D) show that LG_1 is under balancing selection in European strains (SNP_14408, Tajima's D = 1.736, *P-value* < 0.05, Figure 3A, 3D), while the signal of balancing selection is absent in American strains (SNP_14408, Tajima's D = -0.2652, Figure 3A).

In contrast, LG_2 and LG_3 display signatures of balancing selection in American strains, revealed by SNP_28144 (Tajima's D = 1.265, *P-value* < 0.05, Figure 3C) and SNP_17858 (Tajima's D = 1.499, *P-value* < 0.05, Figure 3B), but present directional selection signals in European strains (SNP_28144, Tajima's D = -1.075, *P-value* = 0.3385 and SNP_17858, Tajima's D = -0.912, *P-value* = 0.4976). According to the evolutionary process of COVID-19 in the last three months, the simulation test found that these signals of directional selection are not significant, which indicates that the negative Tajima's D may have resulted from a genetic bottleneck (Figure S2). To verify the result concluded from Tajima's D value, we further performed a composite likelihood ratio test (CLR)(Nielsen, et al. 2005; Zhu and Bustamante 2005) to test the genomic regions where LG_2 and LG_3 are located. We found that the CLR signals were significantly higher in European viruses than in American viruses in the nearby sequences of SNP_28144 (LG_2) (Wilcox signed rank test, *P-value* = 0.0074) and the three SNPs in LG_3 (*P-value* = 0.0053). The CLR peaks adjacent to SNP_18060 (LG_3) and SNP_28144 (LG_2) were also observed in European strains (Figure 3E, 3F).

Both single independent SNPs (SNP_514 and SNP_27046) showed reduced diversity in American strains compared with European strains. These coincide with their negative Tajima's D in American strains. In the nearby region of the SNP_27046, we observed a peak and a significant increase in CLR in American strains compared with European strains (Figure 3F). The simulation test also indicates that the Tajima's D value of SNP_27046 (-1.219) is of weak significance (*P-value* = 0.0712). These results suggest that SNP_27046 is located in a directionally selected region. This is in accordance with the fact that the minor allele (T) accounts for a low percentage of 2% (Table S3) in American strains, suggesting that the major allele (C) is positively selected in America. In comparison, T increased from 0% to 22% in Europe strains, displaying a relaxation in selection (Figure S3). Although SNP_514 has a lower Tajima's D (-1.325) than SNP_27046, it is short of confident evidence in CLR analysis.

In the end, we did not find any significant selection signatures in the nearby sequence of LG_4.

Correlation coefficiency analysis between SNPs and fatality rate of the COVID-19

Based on the global COVID-19 cases reported by the European Centre for Disease Prevention and Control (www.ecdc.europa.eu), we calculated the daily fatality rates and the daily exponential growth rate (λ) (Lipsitch, et al. 2003) of COVID-19 in different countries (Table S4). We also calculated the percentages of minor alleles in different countries in March (Table S5). With these parameters, correlation analysis was performed to deduce the possible phenotype of the 14 alleles. We performed analysis with the data in the UK, Netherlands and US, for the adequacy of sample size (> 50). The result shows that the mutant in LG_1 is positively correlated (*P-value* = 0.0016) with the fatality rate, which suggests that the virus possessing mutants in LG_1 is more aggressive than others. In Italy (fatality rate = 9.3%, Table S4), the percentage of COVID-19 with the LG_1 mutant is 100%. Mutants in LG_2 and LG_3 are mild, showing a significant negative correlation (*P-value* = 0.0102) with the fatality rate. In accordance with previous findings, the mutant of the S type in LG_2 is predicted to be less aggressive than the L type in LG_2 (Tang, et al. 2020). The mutant in LG_4 is negatively correlated with λ but the significance is weak (*P-value* = 0.057).

Functional prediction of the SNPs

As shown in Figure 4, most SNPs (13/14) are in the coding region. SNP_251 (LG_1) is the only SNP located in the noncoding region. It is located at the leader sequence (Sawicki, et al. 2007) in front of ORF1 and may affect the protein-to-protein interaction(Pasternak, et al. 2004, 2006; Sawicki, et al. 2007) in the assembly into membrane bound replication-transcription complexes (RTCs)(Curtis, et al. 2004; Zuniga, et al. 2004; Sola, et al. 2005; Enjuanes, et al. 2006; Yount, et al. 2006). Seven SNPs are located in ORF1ab. Four are synonymous mutations, of which 3 are transitions from 'C' to 'T'. This is consistent with coronavirus codon usage bias towards U-ending (Castells, et al. 2017). SNP_14408 (LG_1) is located at the RNA-directed RNA polymerase and the amino acid (AA) mutation from PRO to LEU may lead to a change in protein flexibility and may further influence viral replication. SNP_17747 and SNP_17858 (both in LG_3) are located within the 2B domain of the helicase C-terminus, which is important for RNA recognition (Jia, et al. 2019). Thus, the change in AA at these two sites may influence RNA binding. The mutation SNP_23403 (LG_1) located in the surface glycoprotein (S) is 60 AA to the 3' end of the receptor

binding domain (RBD). This mutation changes AA side chain polarity, from a negative ASP to a neutral LEU, which may reduce protein rigidity and increase the membrane fusion efficiency during the virus infection. SNP_27046 changes the poplar AA THR to a nonpolar MET in the membrane glycoprotein (M), which potentially affects the membrane association of the protein. SNP_28881, SNP_28882 and SNP_28883 (in LG_4) are three consecutive mutations located in the nucleoprotein (N). They make two AA changes from ARG to LYS and from GLY to ARG. These changes will modify the protein charge distribution, which may affect the recognition of genomic RNA during virus replication and assembly. SNP_28144 (LG_2) is located in ORF8, which is also related to viral replication (Muth, et al. 2018).

Geological evolutionary patterns of COVID-19

COVID-19 shows a divergent evolutionary pattern between Europe and America. In accordance, we observed that the frequency changes of the 14 new alleles are also related to geological differentiation (Figure S3). As a potentially aggressive subtype, the identification frequency of the minor haplotype in LG_1 increased more rapidly in Europe than in America (Chisq-Test, *P-value* < 2.2e-16). In contrast, two mild subtypes in LG_2 and LG_3 arose more rapidly in America than in Europe (Chisq-Test, *P-value* < 2.2e-16). The minor major allele in SNP_27046 is mild and shows directional selection signals in American strains but no selection signals in European strains. Considering our analytic result that the minor haplotype in LG1 has strong correlation with the high fatality rate of the virus, the aggressive virus is exhibiting more adaptive behavior than mild ones in Europe and vice versa in America.

Discussion

With the purpose of predicting the evolutionary pattern of the pandemic COVID-19, we performed a thorough analysis of the virus' genome sequences. Our results show that COVID-19 has geological bias in genomic SNPs correlated with its aggressiveness. One possible explanation for this is that the genetic diversity of the human population in America is higher than that in Europe. For resistance to virus, genes involved in immune responses are polymorphic and have more self-incompatibility loci under balancing selection, such as the major histocompatibility complex (MHC)(Watkins, et al. 1990; Meyer and Thomson 2001; Kwiatkowski 2005). The higher the divergence of immune related genes, the more pathogens the immune cells are able to deal with. The overall genetic polymorphism of the immune system for American people may be higher than that for European people, causing a more severe selection on COVID-19 in America than in Europe.

The increments of most mutants are rapid, linked and simultaneous. This may result from the potential benefit of the mutant in new environments and a successive increasing population size. Beneficial mutants are common in a large population and can dramatically alter the genetic diversity at linked sites (Desai and Fisher 2007). Moreover, the viral population easily experiences evolution via multiple concurrent mutations (Desai and Fisher 2007). Balancing selection in LG_1 is positively maintained in Europe, while that in LG2 and LG3 are positively maintained in America. A similar case was recently reported in a psychiatric disorder-associated human gene (Sato and Kawata 2018). Considering the

nearby sequences of SNP_14408, SNP_28144 and SNP_17858 show the strongest selection signals with geological bias in LG_1, LG_2 and LG_3 respectively, they may be the causative mutation in these linkage groups. The increments of other mutants may be driven by hitchhiking effects (Smith and Haigh 2007). Thus, our results suggest paying more attention to these causative SNP sites in further epidemiological investigations of COVID-19.

Declarations

Author contributions

Z.Z. collected, analyzed and compiled the data. G.L., K.M. and L.Y. performed the analysis in statistics. Z.Z. and G.M. conceived the idea, coordinated the project and wrote the manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgments

This work was supported by grants from the National Key Research and Development Program (2019YFC1604600), the National Natural Science Foundation of China (31200941), the Fundamental Research Funds for the Central Universities (106112016CDJXY290002), the National Natural Science Foundation of HeBei province (19226631D).

References

Castells M, Victoria M, Colina R, Musto H, Cristina J. 2017. Genome-wide analysis of codon usage bias in Bovine Coronavirus. Virol J 14:115.

Curtis KM, Yount B, Sims AC, Baric RS. 2004. Reverse genetic analysis of the transcription regulatory sequence of the coronavirus transmissible gastroenteritis virus. J Virol 78:6061-6066.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics 32:1895-1897.

Desai MM, Fisher DS. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176:1759-1798.

Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Enjuanes L, Almazan F, Sola I, Zuniga S. 2006. Biochemical aspects of coronavirus replication and virus-host interaction. Annu Rev Microbiol 60:211-230.

Fehr AR, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol 1282:1-23.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sanchez E, Albrechtsen A, Nielsen R. 2013. Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195:979-992.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. [Ph.D. thesis]: Pennsylvania State University.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. 2011. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res 39:D576-582.

Hung CL, Lin YS, Lin CY, Chung YC, Chung YF. 2015. CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. Comput Biol Chem 58:62-68.

Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics 7:409.

Jia Z, Yan L, Ren Z, Wu L, Wang J, Guo J, Zheng L, Ming Z, Zhang L, Lou Z, et al. 2019. Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. Nucleic Acids Res 47:6538-6550.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77:171-192.

Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, Wei W, Cheung WY-M, Li W-J, Li L-F, et al. 2020. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. bioRxiv.

Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49:49-67.

Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, et al. 2003. Transmission dynamics and control of severe acute respiratory syndrome. Science 300:1966-1970.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor

binding. Lancet 395:565-574.

Meyer D, Thomson G. 2001. How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65:1-26.

Morton NE. 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318.

Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, Gloza-Rausch F, Balboni A, Battilani M, Rihtaric D, et al. 2018. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci Rep 8:15177.

Namy O, Moran SJ, Stuart DI, Gilbert RJ, Brierley I. 2006. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature 441:244-247.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15:1566-1575.

Pasternak AO, Spaan WJ, Snijder EJ. 2006. Nidovirus transcription: how to make sense...? J Gen Virol 87:1403-1421.

Pasternak AO, Spaan WJ, Snijder EJ. 2004. Regulation of relative abundance of arterivirus subgenomic mRNAs. J Virol 78:8102-8113.

Ralph R, Lew J, Zeng T, Francis M, Xue B, Roux M, Toloue Ostadgavahi A, Rubino S, Dawe NJ, Al-Ahdal MN, et al. 2020. 2019-nCoV (Wuhan virus), a novel Coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness. J Infect Dev Ctries 14:3-17.

Sato DX, Kawata M. 2018. Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits. Evol Lett 2:499-510.

Sawicki SG, Sawicki DL, Siddell SG. 2007. A contemporary view of coronavirus transcription. J Virol 81:20-29.

Smith JM, Haigh J. 2007. The hitch-hiking effect of a favourable gene. Genet Res 89:391-403.

Sola I, Moreno JL, Zuniga S, Alonso S, Enjuanes L. 2005. Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. J Virol 79:2506-2516.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21:2791-2793.

Watkins DI, Chen ZW, Hughes AL, Evans MG, Tedder TF, Letvin NL. 1990. Evolution of the MHC class I genes of a New World primate from ancestral homologues of human non-classical genes. Nature 346:60-63.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-276.

Yi H. 2020. 2019 novel coronavirus is undergoing active recombination. Clinical Infectious Diseases.

Yount B, Roberts RS, Lindesmith L, Baric RS. 2006. Rewiring the severe acute respiratory syndrome coronavirus (SARS-CoV) transcription circuit: engineering a recombination-resistant genome. Proc Natl Acad Sci U S A 103:12546-12551.

Zhu L, Bustamante CD. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics 170:1411-1421.

Zhu Z, Meng K, Meng G. 2020. A database resource for Genome-wide dynamics analysis of Coronaviruses on a historical and global scale. bioRxiv.

Zuniga S, Sola I, Alonso S, Enjuanes L. 2004. Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. J Virol 78:980-994.

Tables

Table 1. Overview of the 14 new SNPs. “Loc”, “Mutat”, “Chg” and “AA” are simplified expressions for location, mutation, change and amino acids, respectively. The locations of SNPs are according to MN908947. “Perc” refers to the percentage of minor alleles. ‘a/A’ refers to the number of minor alleles (a) and the number of major alleles (A) in January, February and March. “Pos in Codon” refers to the position of the mutation in the codon.

Loc	Mutat	Dec-Jan		Feb		Mar		Chg in Perc.	Chisq-Test	Protein	Pos in Codon	AA Mutat	Linkage Group
		Perc (a)	a/A	Perc (a)	a/A	Perc (a)	a/A						
241	C->T	1.5%	3/197	17.5%	51/240	52.1%	226/208	50.6%	< 2.2e-16				LG_1
514	T->C	0.0%	0/197	1.0%	3/292	9.6%	44/416	9.6%	4.033e-4	nsp1	3	Synonymous	
3037	C->T	1.5%	3/197	17.4%	52/246	49.2%	227/234	47.7%	< 2.2e-16	nsp3	3	Synonymous	LG_1
8782	C->T	38.0%	76/124	21.5%	64/233	26.0%	119/339	-12.0%	0.002593	nsp4	3	Synonymous	LG_2
14408	C->T	0.0%	0/202	17.1%	51/247	48.9%	226/236	48.9%	< 2.2e-16	RdRp	2	PRO->LEU	LG_1
17747	C->T	0.0%	0/202	1.7%	5/292	22.1%	101/355	22.1%	8.50E-13	Helicase	2	PRO->LEU	LG_3
17858	A->G	0.0%	0/202	1.7%	5/294	22.7%	105/358	22.7%	3.87E-13	Helicase	2	TYR->CYS	LG_3
18060	C->T	3.0%	6/196	2.3%	7/292	22.9%	106/357	19.9%	6.17E-13	Helicase	3	Synonymous	LG_3
23403	A->G	1.5%	3/199	17.7%	53/246	49.0%	227/236	47.5%	<2.2e-16	S	2	ASP->GLY	LG_1
27046	C->T	0.0%	0/202	0.3%	1/292	14.7%	68/394	14.7%	1.95e-08	M	2	THR->MET	
28144	T->C	38.5%	77/123	20.9%	62/234	25.7%	119/344	-12.8%	1.27e-03	ORF8	2	LEU->SER	LG_2
28881	G->A	0.0%	0/201	8.1%	24/274	22.7%	105/358	22.7%	4.40e-13	N	2	ARG->LYS	LG_4
28882	G->A	0.0%	0/201	8.1%	24/274	22.7%	105/358	22.7%	4.40e-13	N	3		LG_4
28883	G->C	0.0%	0/201	8.1%	24/274	22.7%	105/358	22.7%	4.40e-13	N	1	GLY->ARG	LG_4

Table 2. Correlation analysis results between the ratios of mutants and the fatality rate or the daily exponential growth rate (λ) of COVID-19. The *P-value* was calculated by Spearman's test. Coefficients with significance are marked by '**', while those with weak significance are marked by '*'.

	Fatality Rate		λ	
	Pearson Coefficient	*P-value*	Pearson Coefficient	*P-value*
LG_1	0.999**	0.0016	-0.934	0.2329
LG_2	-0.999**	0.0102	0.939	0.2244
LG_3	-0.999**	0.0269	0.947	0.2077
LG_4	0.961	0.1778	-0.996*	0.0568
SNP_514	0.677	0.5259	-0.897	0.2913
SNP_27046	0.684	0.5208	-0.901	0.2862

Figures

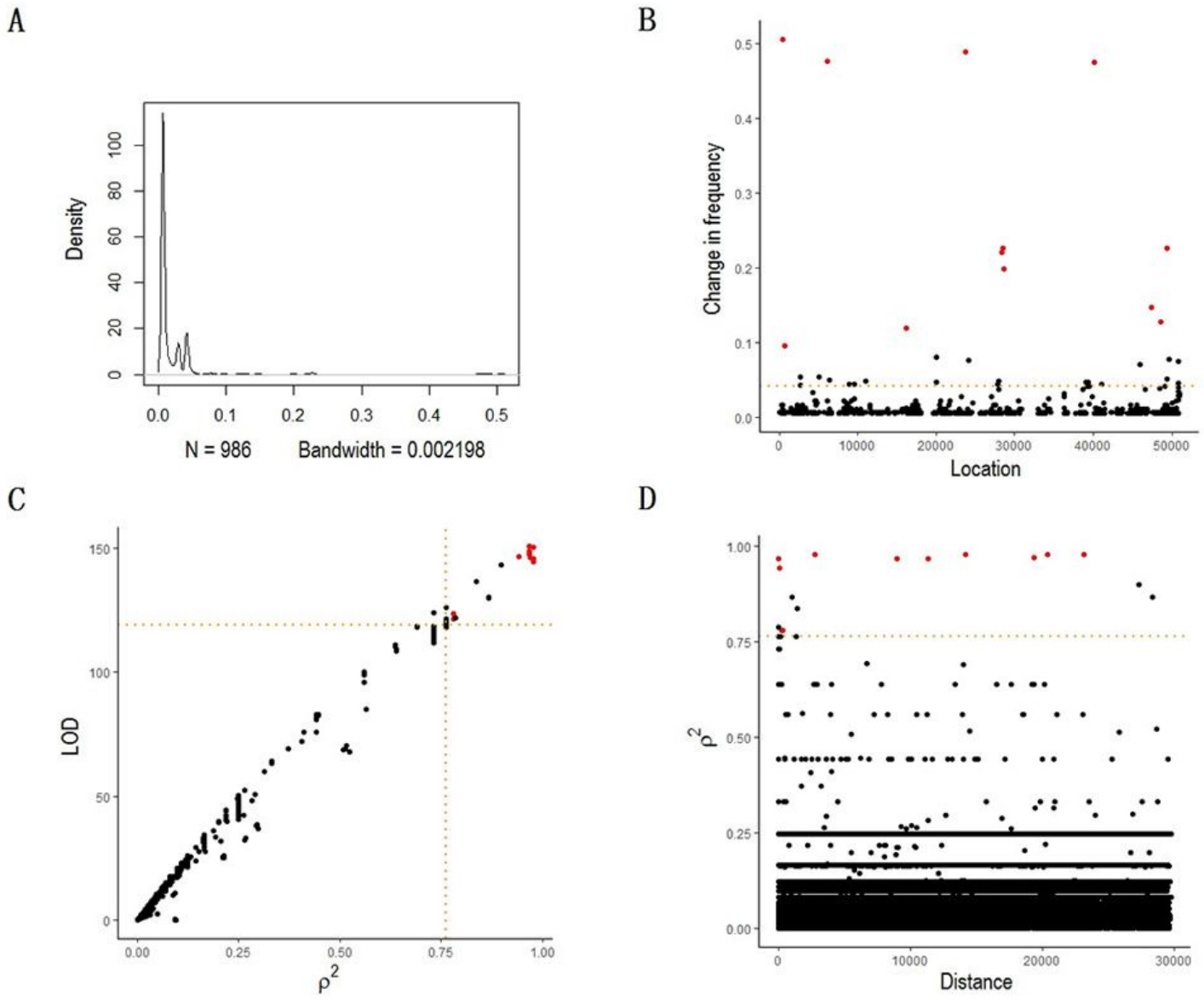

Figure 1

A is the density curve showing the distribution of frequency changes in 986 SNPs. B is the change in frequency of SNPs (y-axis) along different chromosome positions (x-axis), with the 14 new SNPs are marked by red dots.. C shows the LOD score of each pair of SNPs (y-axis) against the squared correlation coefficient ρ2 between that pair (x-axis). D shows ρ2 of each pair of SNPs (y-axis) against the genomic distance between that pair. In C and D, the pairs of the new 14 alleles showing complete or near complete linkages are marked in red. The orange dotted lines in B, C and D are the top 5% positions in the corresponded axes.

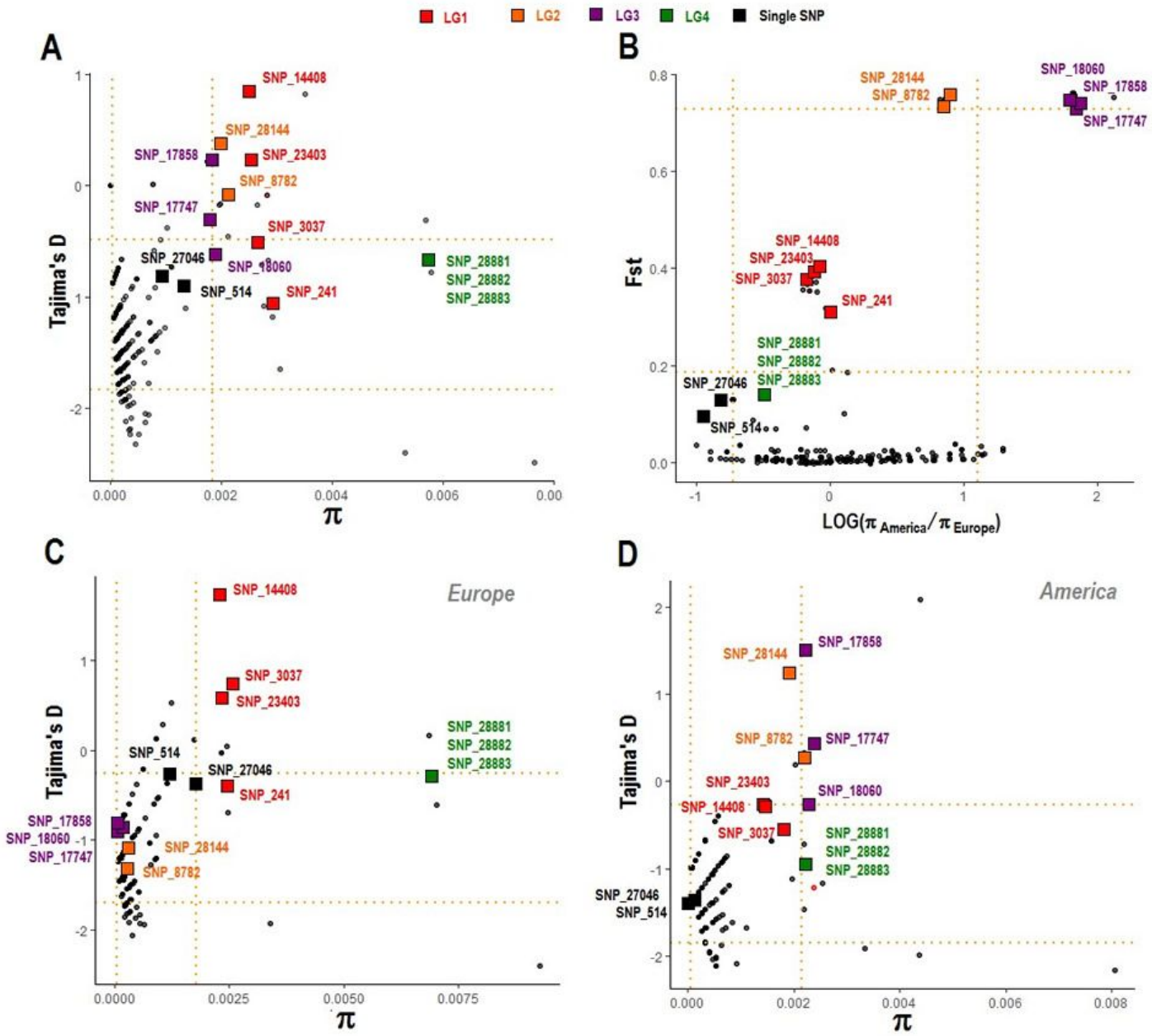

Figure 2

Population genetic analyses of the 14 new SNPs. A, C and D are distributions of Tajima's D and π of 200 bp windows with 50 bp steps for all strains (A), European strains (C) and American strains (D). B is a distribution of ln ratios (π America/π Tibetan) and Fst values, which were calculated in 200 bp windows sliding in 50 bp steps. Positions of the new SNPs are marked by rectangles in different colors, which denote different linkage groups. Dotted lines in orange point out the positions in the top 5% and in the bottom 5%.

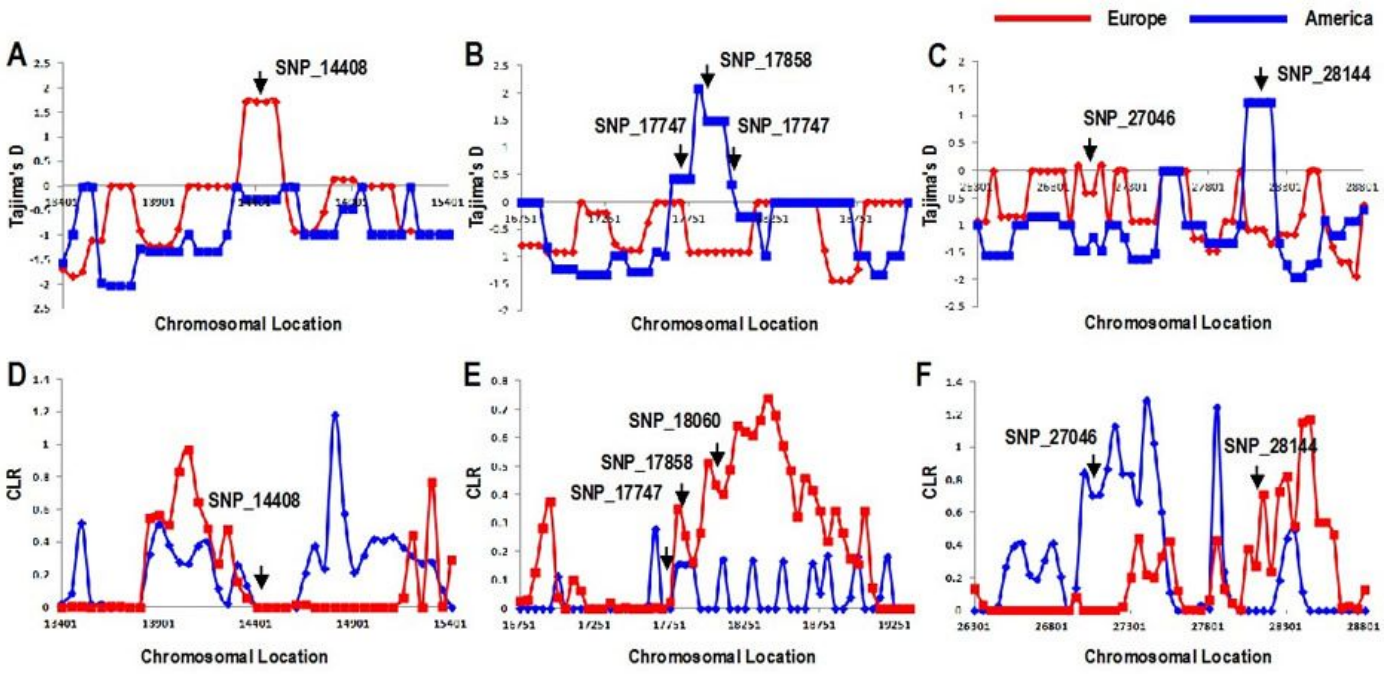

Figure 3

Sliding window analysis view of Tajima's D and CLR at 6 SNP sites, indicated by black arrows.

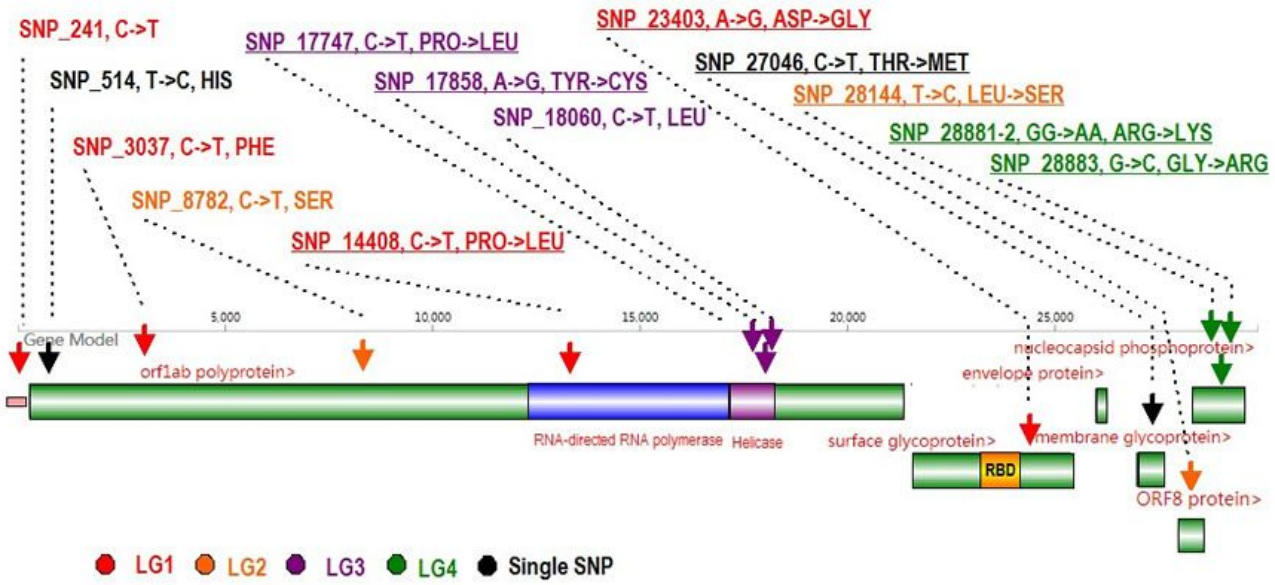

Figure 4

A diagram showing an overview of the 14 SNPs in the virus genome. SNP sites are marked by upside-down arrows and colored according to the linkage groups to which they belong. ‘RBD’ denotes the receptor binding domain, marked in yellow. A pink rectangle ahead of ORF1 denotes the leader sequence for transcription.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf