ISSN 2618-1584 Vol. 2, Issue 3: 35-41, DOI: 10.25236/FMSR.2020.020306

The genomic evolution of SARS-CoV-2 during the spread in the world

Qiling Geng[1], Bomin Wei[1], Zhangcheng Ji[1], Tianyang Wu[2], Zhinan Wang[4] and Shenghong Chen[3]

[1] Princeton International School of Mathematics and Science; PRISMS 19 Lambert Dr, Princeton NJ , USA 08540
[2] The Harker School – Upper School, 500 Saratoga Ave., San Jose, CA, USA 95129
[3] Shanghai American School, 258 Jinfeng Road, Minhang District, Shanghai, China 201107
[4] Shenzhen Middle School, Renmin North Road Shenzhong Street No.18, LuoHu District Shenzhen, Guangdong Province, China 518001

Abstract. *Under the strike of a global pandemic, SARS-CoV-2 has caused a serious global health crisis that challenges worldwide reaction to contain the spread of virus. Numerous studies have been done about SARS-CoV-2 mutation; yet, questions about the correlation between mutation with factors as genders and geographical location are not comprehensively and timely studied. Thus, in this research, we aim to explore the association of the frequency of several mutation hotspots and these factors by generating graphs with R language based on the data of 93625 SARS-CoV-2 genomes analyzed through Python script and conducting statistical tests, in which we failed to reject the null hypothesis that there’s no significant association. Our research has offered insights for the way of mutation of SARS-CoV-2 in different regions and genders, contributing to the understanding of the evolution of SARS-CoV-2, which is of vital importance to the development of a vaccine.*

Keywords: *SARS-CoV-2, Coronavirus, vaccine*

1. Introduction

The emergence of SARS-CoV-2 in late 2019 and its subsequent global spread has led to millions of infections, posing a global health emergency (Wang et al., 2020a; Yang et al., 2020). Research has scurried in isolating, sequencing, and cloning the virus, developing diagnostic kits, and testing candidate vaccines. Coronavirus disease can range from mild self-limiting disease to acute respiratory distress syndrome and death (Wang et al., 2020b). However, significant questions remain about the mutation hotspots in the COVID-19 genome based on geographical location or other factors that remain empty. The rate of spontaneous mutation is a crucial parameter in modeling the genetic structure and evolution of populations (Huang et al., 2020). The impact of the accumulated load of mutations and the consequences of increasing the mutation rate are essential in assessing the genetic health of people.

SARS-CoV-2 is an enveloped, +ssRNA virus, belonging to the Beta coronavirus genus. RNA viruses are flanked by highly structured 5′- and 3′-untranslated regions (UTRs), indispensable for translation and replicating of the viral genome (Latinne et al., 2020). Standard RNA secondary structure prediction tools such as mold and RNAfold (Table 4) are based on the calculation of the minimum free energy (MFE). They can fold reliably on small local windows of up to 300 nt. Secondary structures of larger genomic segments or interactions spanning larger regions, including pseudogenes, are still bioinformatically challenging. Based on the folding of multiple sequences, these sequences are generally more reliable due to following the footsteps of evolution by compensatory mutations. Viruses usually come along with many similar sequences perfect for a large alignment and predicted secondary structures (Zhu et al., 2020). RNA viruses have high mutation rates—up to a million times higher than their hosts—and these high rates are correlated with enhanced virulence and evolvability, traits considered beneficial for viruses. However, most mutations are synonymous mutations, and even for nonsynonymous mutations, most of them will not benefit the survival of their hosts. Many mutations cause organisms to leave fewer descendants over time, so the action of natural selection on these mutations is to purge them from the population. While a small percentage of mutations are helpful, and some are inconsequential (neutral or nearly neutral in effect), a large portion of mutations are harmful (Koyama et al., 2020).

To find out how the new viral variants were spreading across the countries, we focused our study on assessing the SARS-CoV-2 mutation. These variations may contribute to improving medical prognosis, and vaccine design

2. Materials and Methods

The experimental design was focused around the goal of identifying mutation hotspots based on sex and region in the SARS-CoV-2 genome. In order to do this, we used a Python script to retrieve the data of 93625 COVID genomes from six major regions (Africa, Asia, Europe, North America, South America) from the GenBank database. Our samples only included human hosts with lengths between 28011-34692. To find the mutations of each genome, we compared them to the original "Wuhan seafood market pneumonia virus" (WSM, NC_045512) (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512) using MUSCLE (Edgar, 2004; Koyama et al., 2020). We then found the number of mutations in each region using self-made scripts, then the SNPs with a frequency of over 10% in the genomes were identified and plotted. We then used ggplot in R to plot the mutation rate vs. sex and mutation count vs. position graphs.

Sample selection criteria:

Age: We select middle-aged individuals, as the data in the GISAID database contains samples mostly from this age demographic.

Gender: Since the number of deceased patients is 2.4x more male than female, we will use male patients so as not to arrive at false conclusions based on gender. Thought the result has shown that gender did not affect the mutation rates in any region.

Region division: Out of the six major continents, we will choose genomes from patients residing in the top 3 most infected countries on each continent to represent the COVID gene for that continent. The reason is that the mutation rate is directly correlated with the infectivity rate. 100 most recent genomes per country=3*6*100 = 1800 samples.

3. Results

3.1 Gender did not affect the mutation in any region

After removing outlier points from the Asia region, we calculated the mutation rates among female, male and unknown gendered COVID-19 patients. Generally speaking, the mutation frequency of data in Asia and North America is relatively low, and it is impossible to distinguish gender differences (Figure 1). It may be because the data we downloaded is not sufficient for the analysis. In Africa, Europe, and South America, on the other hand, higher mutation frequencies are detected, and some differences can also be seen (Figure 1). Yet, there are no significant differences among groups, indicating gender does not affect the mutation rates in any region (Figure 1).

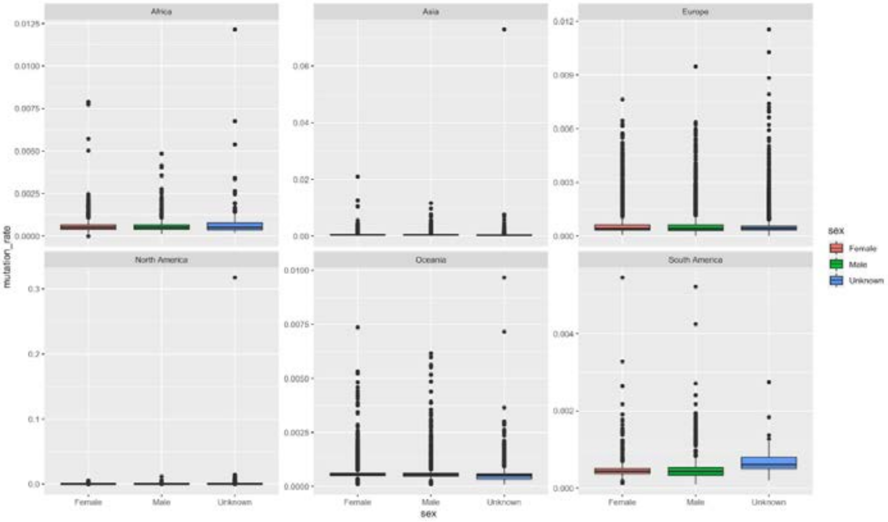

Figure. 1 The effect of sex on the mutation rate in six different regions

3.2 SNP distributed with a pattern

Additionally, we retained SNPs with frequency > 0.1 and displayed the mutation frequency of each site. Interestingly, a bimodal distribution towards the start and end of the genome is observed, especially in South America and Africa (Figure 2). It indicates that both ends of the virus sequence are more likely to mutate, consistent with previous reports.

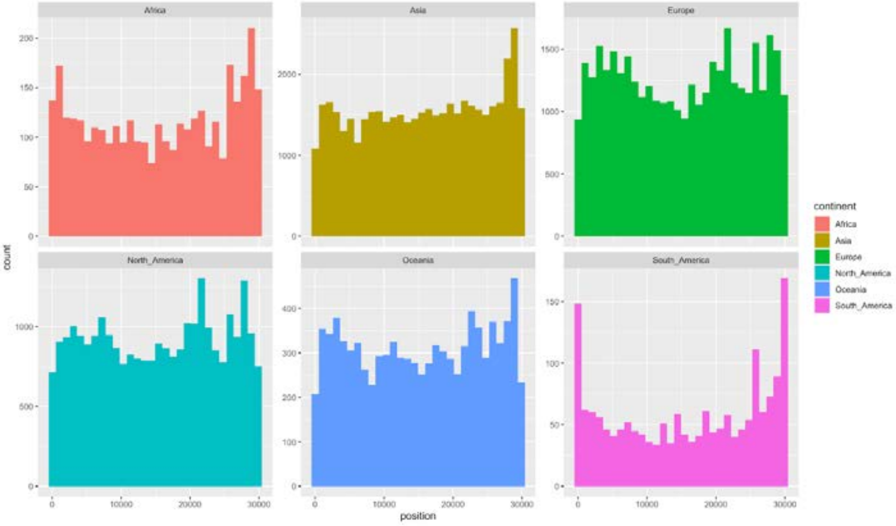

Figure. 2 The distribution of SNPs that occur with a frequency > 0.1 among six regions.

If we considered all the SNPs in six regions, the distribution shows three peaks in the SARS-CoV-2 genome (Figure 3). They appear at the position of 250-750 nt, 2,000-2,250 nt, and 3,000 nt, indicating there are patterns in the mutation loci. As the frequency of gene mutation increases, the corresponding protein functions could be altered. Additionally, the highest mutation locates on the 3’ end, which may be due to the specificity of the RNA virus. Besides, the total mutation is higher in Asia, probably because it is the birthplace of COVID-19, and provides a longer time for mutation of the virus (Figure 3).

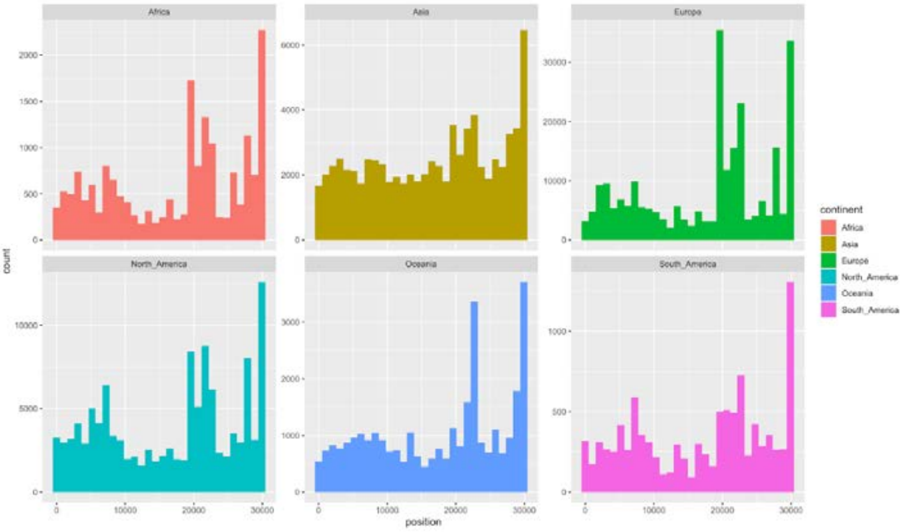

Figure. 3 The distribution of all SNPs in six regions.

We also observed mutations at position 241 (C → T), 3037 (C → T), 14408 (C → T), 28881 (G → A), 28882 (G → A), and 28883 (G → C). They occurred throughout all the regions. From a previous study, mutation at position 14408 caused an amino acid substitution in the RdRp protein. Nucleotide 14408 C-to-T results in a P-to-L amino acid change at 323, an interface domain in SARS-CoV-2. It is speculated to be involved in interactions with other proteins that may modulate the activity of RdRp. Additionally, mutation 28881 is also reported in the previous study, in which locates in a nucleocapsid phosphoprotein (ORF9a), indicating the robustness of our results.

4. Discussion

Currently, we have identified mutation hotspots based on sex and region in the SARS-CoV-2 genome. Our SARS-CoV-2 genome came from the GenBank database. We divided the genome in gender, age, and region, which are the three primary factors we consider. By writing a Python script, we can categorize the different data of sex, the difference data of age, and the difference data of the region, which are Africa, Asia, Europe, North America, South America. Our samples only included human hosts with lengths between 28011-34692.

We observed that in each region, the male, female, and unknown gendered patients' COVID genomes have similar mutation rates. Therefore, we concluded that gender did not affect the mutation rates in any region. Mostly, we observed that mutations at position 241 (C → T), 3037 (C → T), 14408 (C → T), 28881 (G → A), 28882 (G → A), and 28883 (G → C) occurred throughout all the regions. In a previous study, two mutations at positions 14408 and 28881 were reported. The SNPs with a frequency of over 10% in the genomes were identified and plotted, as mentioned above.

5. Conclusion

We focused on the goal of identifying mutation hotspots based on sex and region in the SARS-CoV-2 genome. Using the python script and other alignment methods, we found that gender does not affect the mutations rates in regions. We find that Asia has the highest mutations rate; more than 80% of positions have mutations. We use the SNP data to show the distribution of mutation frequency in the virus genome. Since there are indels at the end of the sequences, the plot has a bulge. For the exact common mutation positions, we find that some mutations occurred throughout all the regions. For example, 241 (C → T), 3037 (C → T), 14408 (C → T), 28881 (G → A), 28882 (G → A), and 28883 (G → C). From recent studies, the mutations at 14408 and 28881 could cause amino acid substitution in the RdRp protein and in the nucleocapsid phosphoprotein (ORF9a), respectively (Pachetti et al., 2020; Yin, 2020). These mutations are present throughout the area in our research.

Author Information

Corresponding Author

Correspondence should be addressed to Zhe Leng at email lengzhe@ibowu.com and Dr. Long Lin at linlong@genomics.cn.

Author Contributions

‡All six authors contributed equally to the conception of the research idea and research work, and all contributed to the writing of the manuscript.

Notes

The authors declare no competing financial interests.

Acknowledgements

The authors received kind and extensive help from the JSA (Juveniles Science Academy) established by iBowu, which provided student mentorship and all activity cooridinations throughout the research. The authors also owe a debt to the BGI Genomics Co Ltd for providing the necessary database access and computational resources. The authors also found inspirations from the discussion with Mr. Zixiao Sui from Beijing National Day School.

References

[1] Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

[2] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., *et al.* (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497-506.

[3] Koyama, T., Platt, D., and Parida, L. (2020). Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ *98*, 495-504.

[4] Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., *et al.* (2020). Origin and cross-species transmission of bat coronaviruses in China. Nat Commun *11*, 4235.

[5] Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., *et al.* (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med *18*, 179.

[6] Wang, L., He, W., Yu, X., Hu, D., Bao, M., Liu, H., Zhou, J., and Jiang, H. (2020a). Coronavirus disease 2019 in elderly patients: Characteristics and prognostic factors based on 4-week follow-up. J Infect *80*, 639-645.

[7] Wang, T., Du, Z., Zhu, F., Cao, Z., An, Y., Gao, Y., and Jiang, B. (2020b). Comorbidities and multi-organ injuries in the treatment of COVID-19. Lancet *395*, e52.

[8] Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., *et al.* (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med *8*, 475-481.

[9] Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics *112*, 3588-3596.

[10] Zhu, N., Wang, W., Liu, Z., Liang, C., Wang, W., Ye, F., Huang, B., Zhao, L., Wang, H., Zhou, W., *et al.* (2020). Morphogenesis and cytopathic effect of SARS-CoV-2 infection in human airway epithelial cells. Nat Commun *11*, 3910.