Federico Di Lello ORCID iD: 0000-0001-9771-9705

Title Page:

Title: Evolutionary analysis of SARS-CoV-2 spike protein for its different clades

Running title: SARS-CoV-2 spike molecular diversification

Authors: Matías J. PERESON[a,b], Diego M. FLICHMAN[b,c], Alfredo P. MARTÍNEZ[d], Patricia BARÉ[b,e], Gabriel H. GARCIA[a], Federico A. DI LELLO[a,b]

Affiliations:

[a]Universidad de Buenos Aires. Facultad de Farmacia y Bioquímica. Instituto de Investigaciones en Bacteriología y Virología Molecular (IBaViM). Buenos Aires, Argentina.

[b]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Argentina.
[c]Instituto de Investigaciones Biomédicas en Retrovirus y Síndrome de Inmunodeficiencia Adquirida (INBIRS) – Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires, Buenos Aires, Argentina.
[d]Virology Section, Centro de Educación Médica e Investigaciones Clínicas Norberto Quirno "CEMIC". Buenos Aires, Argentina.
[e]Instituto de Medicina Experimental (IMEX) – Academia Nacional de Medicina. Buenos Aires, Argentina.

Corresponding Author:

Dr. Federico Alejandro Di Lello, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956, 4º piso, (1113), Ciudad de Buenos Aires, Argentina. Phone: +54 11 5287 4472

Fax: +54 5287 4662, E-mail: fadilello@ffyb.uba.ar

Abstract

The spike protein of SARS-CoV-2 has become the main target for antiviral and vaccine development. Despite its relevance, there is scarce information about its evolutionary traces. The aim of this study was to investigate the diversification patterns of the spike for each clade of SARS-CoV-2 through different approaches.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jmv.26834.

Two thousand and one hundred sequences representing the seven clades of the SARS-CoV-2 were included. Patterns of genetic diversifications and nucleotide evolutionary rate were estimated for the spike genomic region.

The haplotype networks showed a star shape, where multiple haplotypes with few nucleotide differences diverge from a common ancestor. Four hundred seventy nine different haplotypes were defined in the seven analyzed clades. The main haplotype, named Hap-1, was the most frequent for clades G (54%), GH (54%), and GR (56%) and a different haplotype (named Hap-252) was the most important for clades L (63.3%), O (39.7%), S (51.7%), and V (70%). The evolutionary rate for the spike protein was estimated as 1.08×10^{-3} nucleotide substitutions/site/year. Moreover, the nucleotide evolutionary rate after nine months of pandemic was similar for each clade.

In conclusion, the present evolutionary analysis is relevant since the spike protein of SARS-CoV-2 is the target for most therapeutic candidates; besides, changes in this protein could have consequences on viral transmission, response to antivirals and efficacy of vaccines. Moreover, the evolutionary characterization of clades improves knowledge of SARS-CoV-2 and deserves to be assessed in more detail since re-infection by different phylogenetic clades has been reported.

Keywords: SARS-CoV-2; Spike protein; Evolution; Clades

Introduction

In December 2019, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged and shocked the entire world [1]. After a year of worldwide circulation, more than 98 million cases and 2 million deaths have been reported globally [2]. Seven genetic clades (S, L, O, V, G, GR, and GH) have been described over time that are spread throughout different countries [3]. These clades represent a challenge for public health as re-infection cases with different clade strains have been reported [4,5,6]. In fact, more than 200 candidates for vaccines against SARS-CoV-2 and several antivirals are already being developed [7,8]. Most of vaccines and therapeutic drugs are directed towards the spike glycoprotein (S) that is responsible for entering the host cell through recognition of the receptor ACE2 with the receptor binding protein (RBD) [8,9,10,11,12]. Therefore, knowing the evolutionary rate of the S is relevant since changes in this protein could affect the efficacy of the vaccine and the antivirals directed to S. Even though some studies have determined the nucleotide evolutionary rate of SARS-CoV-2 using the entire genome [13,14], those values are

slower and do not represent the real mutation capacity of the S region alone. Only one study has reported the nucleotide evolutionary rate of the S genomic region in the first four months of the pandemic, but without differentiating the seven viral clades, which can be relevant in therapeutics and re-infections [15]. Thus, the aim of this study was to determine the nucleotide evolutionary rate and the haplotype network of the S region for SARS-CoV-2 in general and for each of the seven genetic clades during the first nine months of pandemic.

Materials and Methods

Datasets

In order to generate datasets representing different geographic regions and time evolution for each of the seven clades of SARS-CoV-2, from December 2019 to September 2020, data of complete genome sequences available at GISAID (https: //www.gisaid.org /) on September 2020 were randomly monthly collected for several geographic regions. Data inclusion criteria were: a.- complete genomes, b.- high coverage level, and c.- human host only (no other animals or environmental samples). Complete genomes were aligned using MAFFT against the Wuhan-Hu-1 reference genome (NC_045512.2, EPI_ISL_402125). The resulting multiple sequence alignments were split in a dataset corresponding to the S region [3,822nt (21,563-25,384)] and RBD (included in S) [762nt (22,550-23,311)].

Phylogenetic and statically analysis / Genetic characterization

Patterns of genetic diversifications for both genomic regions S and RBD for each clade were analyzed using the median-joining reconstruction method at the PopART v1.7.2 software [16]. Haplotypes shared among all clades were analyzed in Arlequin 3.5.2.2 software [17]. Polymorphism indices were calculated separately for each clade with DnaSPv. 6.12.01 [18].

Nucleotide evolutionary rate

The estimation of the nucleotide evolutionary rate for the entire S-coding region datasets were carried out with the Beast v1.8.4 program package [19] at the CIPRES Science Gateway server [20]. The temporal calibration was established by the samples' date of sampling. The best nucleotide substitution model was selected according to the Bayesian information criterion (BIC) method in IQ-TREE v1.6.12 software [21]. The analysis was performed under a relaxed (uncorrelated lognormal) molecular clock model recommended previously by Duchene & col. [22] with an exponential demographic model [23]. Analyses were run for $8x10^6$ generations and sampled every $8x10^5$ steps. The convergence of the "meanRate" and "allMus" parameters [effective sample size (ESS) ≥ 200, burn-in 10%] was verified with Tracer v1.7.1 [24]. The obtained substitution rate was probed against 10 independent

replicates of the analysis with the time calibration information (date of sampling) randomized as described by Rieux & Khatchikian, 2017 [25].

Results

Datasets

Three-hundred sequences were randomly selected for each clade. Two thousand and one hundred sequences were curated and selected for the analysis. Table 1 shows the SARS-CoV-2 sequences included for every month and clade.

Phylogenetic and statically analysis / Genetic characterization

The haplotype networks (Figure 1) reflect the diversity indices results as a star shape with multiple haplotypes with few nucleotide differences that diverge from a common ancestor. In all cases, the RBD diversification is lower than the spike one, being the lowest for clades S and V. For the S-coding region, 479 different haplotypes were defined in the seven analyzed clades. The number of haplotypes observed among clades ranged from 53 for the V clade to 89 for the GH and GR clades (Table 2). The major haplotype 1 (Hap-1), defined by amino acids S12, L18, R21, A222, N439, S477, T478, A522, E583, G614, Q675, E780, D936, V1068, and P1263 was the most frequent for clades G (54%), GH (54%), and GR (56%). However, other 10 haplotypes with amino acid changes respect to the Hap-1 were also observed. On the other hand, haplotype 252 (Hap-252), defined by amino acids L5, L8, H49, V367, A575, D614, A829, A846, D1084, and A1087 was the most frequent for clades L (63.3%), O (39.7%), S (51.7%), and V (70%). In addition, other 10 haplotypes showed one amino acid change respect to Hap-252. Table 3 shows the frequency of each haplotype with amino acid changes.

The haplotype diversity was moderate to high in every clade, ranging from Hd=0.507 to 0.793 (Table 2). In contrast, nucleotide diversity was relatively low for each clade, ranging between π=0.0018 for V and π=0.0040 for O (Table 2). Although overall diversity was similar among different clades, haplotype and nucleotide diversities were both the lowest for V. On the other hand, haplotype and nucleotide diversity was higher for G, GH, GR, and O (Table 2). The RBD region showed indices with a similar trend but with lower values compared to the S region.

Evolutionary rate

After nine months of pandemic, the estimated evolutionary rate for the S genomic region of SARS-CoV-2 was 1.08×10^{-3} nucleotide substitutions per site per year (s/s/y) (95% HPD interval 7.94×10^{-4} to 1.41×10^{-3} s/s/y). Additionally, the nucleotide evolutionary rate for the different genetic clades ranged between 1.06×10^{-3} and 1.69×10^{-3} s/s/y (Table 4). A date-randomization analyses showed no overlapping between the 95% HPD substitution-rate intervals obtained from real data and from date-randomized datasets for all clades (Figure 2).

Dataset for the clade L did not reach convergence (ESS<200). To verify the reliability of the result, 10 independent runs were performed. All of them converged in a similar posterior distribution. Likewise, for many of the random sample datasets, convergence was not achieved (ESS between 100 and 200). For those datasets that

did not reach convergence, two independent runs were carried out and concatenated [26].

When the evolutionary rate was analyzed according to the emergence of each clade, founding clades (L, O, S, and V) tended to present evolutionary rates slightly slower than the more recent clades (G, GH, and GR), ($p = 0.157$).

Discussion

The evolutionary characterization of spike genomic region of SARS-CoV-2 is crucial to estimate the course that re-infections, vaccines, and therapeutics would have in the pandemic's future. In this study, the evolutionary rate of the most important SARS-CoV-2 protein for vaccine development was estimated in general and separately for each genetic clade described in GISAID. In this context, the spike haplotype network showed a founding central paternal haplogroups from which multiple sequences with modest changes derived. Overall, the nucleotide evolutionary rate after nine months of pandemic was similar for each clade.

At the beginning of the pandemic, the most prevalent clades were L, O, V, and S. Later, with the appearance of the D614G mutation in the S protein, clade G emerged and remained with a high and stable prevalence. After this initial step, the GR clade has emerged and grown until it became the most prevalent. Finally, the GH clade peaked at 30% in May 2020 and then began to decrease [3]. In this sense, it is important to highlight that clades with the mutation D614G in the S protein (clades G, GH, and GR) have been suggested to present a higher transmission efficiency although they would not be associated with a more severe pathogenesis [27].

Therefore, in order to describe the evolution of the S protein variants, the study of haplotypes network in all seven clades and for both regions (S and RBD alone) was performed. This analysis showed several identical sequences grouped together resulting in a star-shaped network, which is characteristic of viral outbreaks [28]. For the spike, this general analysis was supported by statistics that show a large number of haplotypes with a small number of nucleotide changes (low nucleotide diversity). However, for the RBD region, an increase in identical haplotypes was observed, which translates into a decrease in other parameters (H, Hd, and Π). This may be due to the conserved nature of the cell receptor-binding region, which is necessary for the infection of target cells. It is noteworthy that the lowest gene and nucleotide diversities observed for clade V, in both S and RBD, could be the result of fewer sequences available for this clade during the nine months analyzed here. In this way, it can be observed that more than 90% of the V clade sequences were distributed in four months (February to May). On the other hand, the highest nucleotide diversity observed in clade O is the result of a less clearly defined pattern of mutations [29].

Several amino acid changes detected in the haplotypes present in our analysis are part of the RBD (V367F, S477N, N439K, T478I, and A522S). From these amino acid changes, positions 367 and 439 were associated with the binding affinity of RBD [30,31]. Additionally, the mutation L5F in the signal peptide was present in 3.3% of members belonging clade V [27]. Other changes associated to relevant functions [27,30] such as H49Y in clade L (associated with monomer stability), A829T in clade S (fusion peptide), D936Y in clade GH [Heptad repeat 1 (HR1) associated with

monomer stability], and P1263 in clade G (present in the cytoplasmic tail), were also detected in 1% to 3.4%.

The evolutionary characterization of the wide spectrum of haplotypes contributes to determine the haplotype significance and its association with disease severity, response to antivirals, development of vaccines, and host genetic factors.

The evolutionary rate of S protein estimated for all together clades was significantly higher than that previously reported by analyzing the entire genome [14,28]. This is expected since the complete genome includes several genomic regions with a high degree of conservation, while the S region is one of the most rapidly evolving in the SARS-CoV-2 genome [15]. Nonetheless, the spike evolution rate was quite similar to that obtained by analyzing this region during the first four months of the pandemic [15]. Although the evolutionary rate of all clades was similar, the founding clades (L, O, V, and S) showed evolutionary rates slightly lower than the most recent and currently more distributed ones (G, GH, and GR). This could be endorsed to the spread process in human populations since they are the most widely disseminated clades around the world.

This study provides substantial data on the evolutionary process of S protein in the different clades of a virus that infects a susceptible population where a massive active immunization process has not yet been carried out. However, as it was aforementioned, the evolutionary rate of the S region remained stable throughout the nine considered months. In coming months, this scenario may modify and it would be necessary to re-evaluate the results from this study. In fact, a new clade named GV was described last months [32]. The inclusion in the study of only 2,100 of the 73,393 available sequences on September 2020 is a limitation that implies a bias in the obtained results, although the sequence selection process was carefully carried out in order to generate a representative dataset from different time courses and a wide geographic range.

Conclusions

Since the S protein of SARS-CoV-2 mediates the entry in the host cell and is the target for most therapeutic candidates, it is essential to know the way this genomic region is evolving, given that changes in this protein could have consequences on viral transmission, response to antivirals, and efficacy of vaccines. On this basis, the results obtained in this work about the evolutionary rate of the spike protein during the first nine months of pandemic are very significant. Furthermore, the evolutionary study of each separate clade ads to the virus knowledge and deserves to be assessed in more detail since re-infection by a different phylogenetic clade has been reported.

Competing interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding: None

Author contributions:

MJP: Data curation, acquisition of data, analysis and interpretation of data, drafting the article, final approval of the version to be submitted.

DMF: Data curation, Validation, drafting the article, final approval of the version to be submitted.

APM: Data curation, Validation, revising the article critically for important intellectual content, final approval of the version to be submitted.

PB: Data curation, acquisition of data, analysis and interpretation of data, revising the article critically for important intellectual content, final approval of the version to be submitted.

GG: Data curation, acquisition of data, analysis and interpretation of data, drafting the article, final approval of the version to be submitted.

FAD: Conception and design of the study, acquisition of data, analysis and interpretation of data, drafting the article, final approval of the version to be submitted.

Acknowledgements

MJP, DMF, PB and FAD are members of the National Research Council (CONICET). We would like to thank to the researchers who generated and shared the sequencing data from GISAID (https://www.gisaid.org/) and Mrs. Silvina Heisecke from CEMIC-CONICET for providing language assistance.

DATA AVAILABILITY STATEMENT

Data derived from public domain resources

REFERENCES

[1] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England journal of medicine*, *382*(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

[2] World Healt Organization, 2020a. Coronavirus disease (COVID-19) Weekly Operational Update on COVID-19. January 27, 2021. Retrieved from: https://www.who.int/publications/m/item/weekly-epidemiological-update---27-january-2021 (27 January 2021, date last accessed).

[3] Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., et al. (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 25(32), 2001410. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410

[4] Gupta, V., Bhoyar, R. C., Jain, A., Srivastava, S., Upadhayay, R., Imran, M., et al. (2020). Asymptomatic reinfection in two healthcare workers from India with genetically distinct SARS-CoV-2. Clinical infectious diseases: an official publication of

the Infectious Diseases Society of America, ciaa1451. Advance online publication. https://doi.org/10.1093/cid/ciaa1451

[5] To, K. K., Hung, I. F., Ip, J. D., Chu, A. W., Chan, W. M., Tam, A. R., et al. (2020). COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America, ciaa1275. Advance online publication. https://doi.org/10.1093/cid/ciaa1275

[6] Van Elslande, J., Vermeersch, P., Vandervoort, K., Wawina-Bokalanga, T., Vanmechelen, B., Wollants, E., et al. (2020). Symptomatic SARS-CoV-2 reinfection by a phylogenetically distinct strain. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America, ciaa1330. Advance online publication. https://doi.org/10.1093/cid/ciaa1330

[7] Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2020). Characteristics of SARS-CoV-2 and COVID-19. Nature reviews. Microbiology, 1–14. Advance online publication. https://doi.org/10.1038/s41579-020-00459-7

[8] World Healt Organization, 2020b. Draft landscape of COVID-19 candidate vaccines. November 3, 2020. Retrieved from: https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines (9 November 2020, date last accessed).

[9] Alexpandi, R., De Mesquita, J. F., Pandian, S. K., & Ravi, A. V. (2020). Quinolines-Based SARS-CoV-2 3CLpro and RdRp Inhibitors and Spike-RBD-ACE2 Inhibitor for Drug-Repurposing Against COVID-19: An in silico Analysis. Frontiers in microbiology, 11, 1796. https://doi.org/10.3389/fmicb.2020.01796

[10] Olaleye, O. A., Kaur, M., Onyenaka, C., & Adebusuyi, T. (2020). Discovery of Clioquinol and Analogues as Novel Inhibitors of Severe Acute Respiratory Syndrome Coronavirus 2 Infection, ACE2 and ACE2 - Spike Protein Interaction In Vitro. bioRxiv: the preprint server for biology, 2020.08.14.250480. https://doi.org/10.1101/2020.08.14.250480

[11] Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., et al. (2020). Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences of the United States of America, 117(21), 11727–11734. https://doi.org/10.1073/pnas.2003138117

[12] Trezza, A., Iovinelli, D., Santucci, A., Prischi, F., & Spiga, O. (2020). An integrated drug repurposing strategy for the rapid identification of potential SARS-CoV-2 viral inhibitors. Scientific reports, 10(1), 13866. https://doi.org/10.1038/s41598-020-70863-9

[13] Giovanetti, M., Benvenuto, D., Angeletti, S., & Ciccozzi, M. (2020). The first two cases of 2019-nCoV in Italy: Where they come from?. Journal of medical virology, 92(5), 518–521. https://doi.org/10.1002/jmv.25699

[14] van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases, 83, 104351. https://doi.org/10.1016/j.meegid.2020.104351

[15] Pereson, M. J., Mojsiejczuk, L., Martínez, A. P., Flichman, D. M., Garcia, G. H., & Di Lello, F. A. (2020). Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence. Journal of medical virology, 10.1002/jmv.26545. Advance online publication. https://doi.org/10.1002/jmv.26545

[16] Leigh, J. W., & Bryant, D. (2015). POPART: Full-feature software for haplotype network construction. Methods in Ecology and Evolution, 6(9), 1110–1116. https://doi.org/10.1111/2041-210X.12410
[17] Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources, 10(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x
[18] Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Molecular biology and evolution, 34(12), 3299–3302. https://doi.org/10.1093/molbev/msx248
[19] Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus evolution, 4(1), vey016. https://doi.org/10.1093/ve/vey016
[20] Miller, M. A., Pfeiffer W. & Schwartz, T. (2010). "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," *2010 Gateway Computing Environments Workshop (GCE)*, New Orleans, LA, 1-8. https://doi.org/10.1109/GCE.2010.5676129
[21] Kalyaanamoorthy, S., Minh, B. Q., Wong, T., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods, 14(6), 587–589. https://doi.org/10.1038/nmeth.4285
[22] Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. bioRxiv, [Preprint]. https://doi.org/10.1101/2020.05.04.077735
[23] Grassly, N. C., & Fraser, C. (2008). Mathematical models of infectious disease transmission. Nature reviews. Microbiology, 6(6), 477–487. https://doi.org/10.1038/nrmicro1845
[24] Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic biology, 67(5), 901–904. https://doi.org/10.1093/sysbio/syy032
[25] Rieux, A., & Khatchikian, C. E. (2017). tipdatingbeast: an r package to assist the implementation of phylogenetic tip-dating tests using beast. Molecular ecology resources, 17(4), 608–613. https://doi.org/10.1111/1755-0998.12603
[26] Lemey, P., Salemi, M., & Vandamme, A. (Eds.). (2009). The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing (2nd ed.). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511819049
[27] Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell, 182(4), 812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043
[28] Liu, Q., Zhao, S., Shi, C. M., Song, S., Zhu, S., Su, Y., et al. (2020). Population Genetics of SARS-CoV-2: Disentangling Effects of Sampling Bias and Infection Clusters. Genomics, proteomics & bioinformatics, S1672-0229(20)30062-0. Advance online publication. https://doi.org/10.1016/j.gpb.2020.06.001
[29] Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in microbiology, 11, 1800. https://doi.org/10.3389/fmicb.2020.01800
[30] Teng, S., Sobitan, A., Rhoades, R., Liu, D., & Tang, Q. (2020). Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-

binding affinity. Briefings in bioinformatics, bbaa233. Advance online publication. https://doi.org/10.1093/bib/bbaa233

[31] Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., et al. (2020). Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cellular & molecular immunology, 17(6), 621–630. https://doi.org/10.1038/s41423-020-0458-z

[32] Nextstrain: real-time tracking of pathogen evolution. Retrieved from: www.gisaid.org (14 December 2020, date last accessed).

Table 1. Number of SARS-CoV-2 sequences from GISAID database on September 18th, by month and clade as per the selection criteria (Temporal structure).

Clade	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Total
G	0	8 (2)*	20 (3)	55 (7)	52 (7)	47 (7)	39 (6)	39 (6)	20 (6)	20 (6)	300
GH	0	0	18 (3)	53 (7)	50 (7)	44 (7)	40 (7)	40 (7)	35 (6)	20 (6)	300
GR	0	0	35 (3)	45 (7)	50 (7)	40 (7)	43 (7)	35 (7)	32 (7)	20 (6)	300
L	17 (8)	43 (5)	53 (5)	65 (6)	55 (5)	49 (4)	14 (4)	4 (2)	0	0	300
O	0	35 (2)	40 (4)	55 (6)	46 (6)	42 (5)	40 (5)	24 (5)	14 (5)	4 (4)	300
S	1 (1)	50 (5)	50 (5)	70 (6)	68 (6)	31 (5)	25 (5)	4 (4)	1 (1)	0	300
V	0	4 (2)	44 (4)	101 (6)	97 (6)	33 (5)	18 (4)	2 (2)	1 (1)	0	300
Total	18 (9)	140 (16)	260 (27)	444 (45)	418 (44)	286 (40)	219 (38)	148 (33)	103 (26)	64 (22)	2100 (300)

*The number of sequences selected for the general dataset (N=300), at each moment and clade, are shown in parentheses

Table 2. Summary of the haplotype and nucleotide diversity indices for the entire Spike and the Receptor Binding-Domain coding regions for each clade of SARS-COV2.

SPIKE

Clade	S	H	Hd	П
G	100	86	0.704 ± 0.030	0.00037 ± 0.00003
GH	102	89	0.704 ± 0.030	0.00038 ± 0.00003
GR	112	89	0.683 ± 0.031	0.00038 ± 0.00003
L	87	76	0.598 ± 0.035	0.00023 ± 0.00002
O	81	68	0.793 ± 0.019	0.00040 ± 0.00002
S	72	60	0.716 ± 0.027	0.00031 ± 0.00002
V	56	53	0.507 ± 0.036	0.00018 ± 0.00002
General	134	107	0.857 ± 0.015	0.00052 ± 0.00003

RBD

Clade	S	H	Hd	П
G	15	15	0.183 ± 0.030	0.00028 ± 0.00005
GH	17	19	0.196 ± 0.031	0.00032 ± 0.00006
GR	23	23	0.281 ± 0.034	0.00041 ± 0.00006
L	15	14	0.104 ± 0.024	0.00016 ± 0.00004
O	9	9	0.193 ± 0.030	0.00027 ± 0.00004
S	3	4	0.027 ± 0.013	0.00003 ± 0.00002
V	3	4	0.020 ± 0.011	0.00003 ± 0.00001
General	17	17	0.166 ± 0.029	0.00026 ± 0.00005

S= Number of variable sites; H = Number of haplotypes; Hd = Haplotype diversity; π = Nucleotide diversity (per site)

Table 3. Frequency of haplotypes with amino acid changes in the spike for each clade of SARS-COV2.

Clade/ Haplotype (aa Change respect to Hap-1)	G N (%)	GH N (%)	O N (%)	GR N (%)	Clade/ Haplotype (aa Change respect to Hap-252)	L N (%)	O N (%)	S N (%)	V N (%)
Hap-1 (*)	162 (54)	162 (54)	25 (8.3)	168 (56)	Hap-252 (#)	190 (63.3)	119 (39.7)	154 (51.3)	210 (70)
Hap-7 (S477N)	5 (1.7)			10 (3.4)	Hap-254 (H49Y)	3 (1)			
Hap-34 (N439K)	4 (1.3)				Hap-256 (D1084Y)	3 (1)			
Hap-67 (P1263L)	4 (1.3)				Hap-282 (NC)	6 (2)	62 (20.7)		
Hap-86 (L18F, A222V)	20 (6.8)				Hap-291 (L5F)				10 (3.3)
Hap-90 (A522S, E780C)		5 (1.7)			Hap-320 (A575S)	8 (2.6)			
Hap-91 (E780C)		6 (2)			Hap-324 (A1087S)	3 (1)			
Hap-105 (D936Y)		10 (3.4)			Hap-367 (L8V)			17 (5.7)	
Hap-137 (E583D)		3 (1)			Hap-382 (V367F)		5 (1.7)		

Hap-171 (Q675R)			3 (1)	Hap-384 (D614A)			3 (1)	
Hap-187 (S12F)			3 (1)	Hap-415 (A829T)			5 (1.7)	
Hap-226 (T478I)			8 (2.6)	Hap-437 (A846S)				
Total	**300 (100)**	**300 (100)**	**300 (100)**	**Total**	**300 (100)**	**300 (100)**	**300 (100)**	**300 (100)**

N: Number, Hap= haplotype, aa= Amino acid, NC= No amino acid changes

* Hap-1: S12, L18, R21, A222, N439, S477, T478, A522, E583, G614, Q675, E780, D936, V1068, and P1263.

Hap-252: L5, L8, H49, V367, A575, D614, A829, A846, D1084, and A1087.

Table 4. Mean rates of the Spike-coding region (nt = 3822) for each clade of SARS-COV2.

Clade	N	Model	Mean Rate	HPD 95% inteval
G	300	TIM2+f	1.47×10^{-3}	1.05×10^{-3} - 1.95×10^{-3}
GH	300	TIM2+f+I	1.42×10^{-3}	9.67×10^{-4} - 1.94×10^{-3}
GR	300	TIM2+f+I	1.69×10^{-3}	1.11×10^{-3} - 2.30×10^{-3}
L	300	TIM2+f	1.11×10^{-3}	5.90×10^{-4} - 1.61×10^{-3}
O	300	TIM2u+f	1.06×10^{-3}	7.20×10^{-4} - 1.50×10^{-3}
S	300	TN+F	1.33×10^{-3}	8.41×10^{-4} - 1.83×10^{-3}
V	300	HKY+F	1.15×10^{-3}	6.51×10^{-4} - 1.64×10^{-3}
General	300	GTR+F+I	1.08×10^{-3}	7.94×10^{-4} - 1.41×10^{-3}

N: Number of sequences

FIGURES

Figure 1 Me**dian-joining haplotype networks.** The seven clades of SARS-CoV 2 described to date are compared to both the entire Spike and the RBD coding region. The diameters of the spheres are proportional to the frequency of haplotypes. The main haplogroups are indicated.

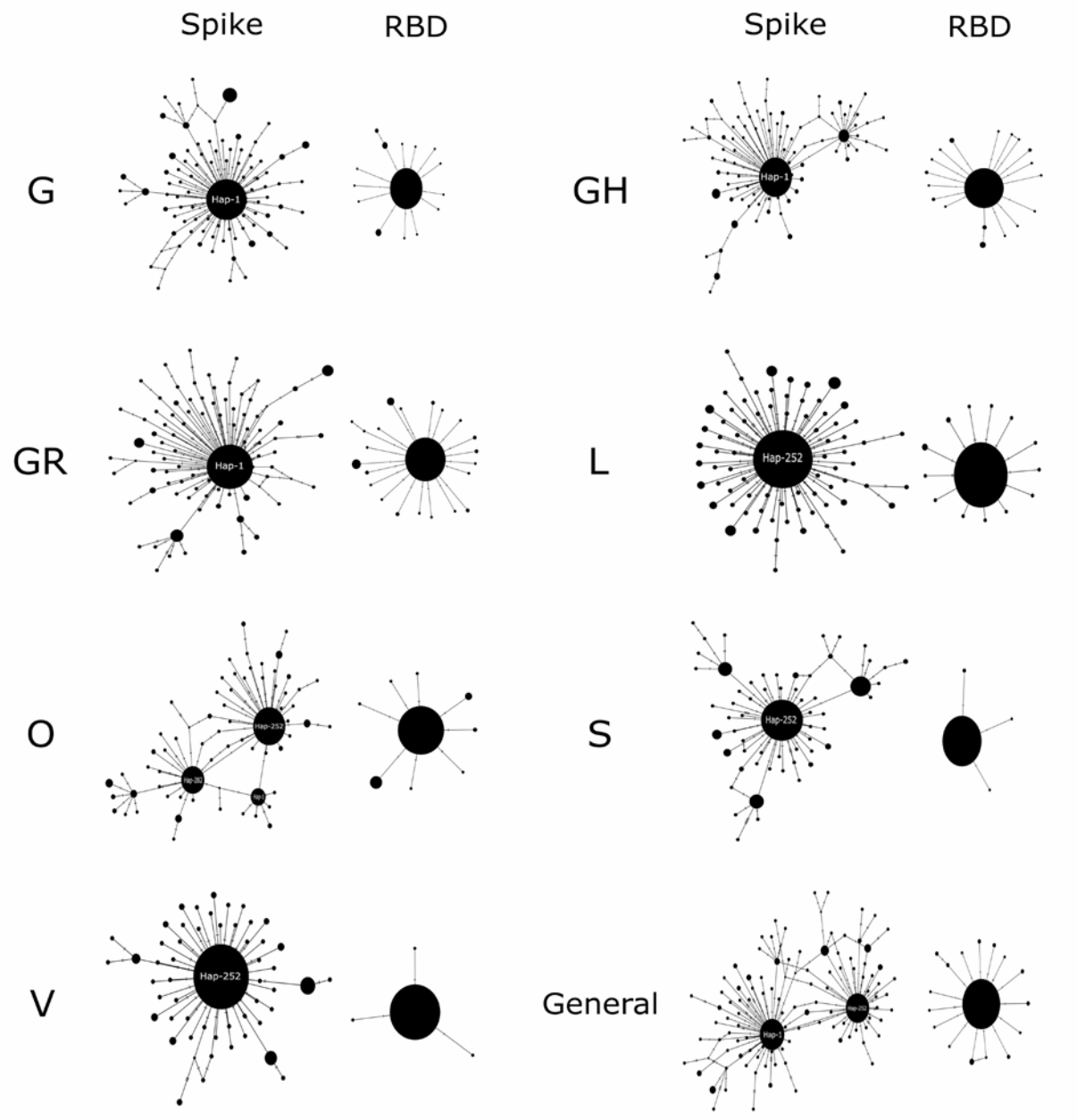

Figure 2 Te**st of temporal structure.** Comparison of the evolutionary rates estimated for the original dataset vs. the date-randomized ones. This analysis was performed for the Spike-coding region (3822nt) of each clade. s.s.y = substitutions/site/year.