

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

### **Abstract**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a role in the mortality of more than 6 million people worldwide. This virus owns the genome, which contains four structural proteins, including spike (S), envelope (E), membrane (M), and nucleocapsid (N). The occurrence of structural mutations can induce the emergence of new variants. Depending on the mutations, the variants may display different patterns of infectivity, mortality, and sensitivity toward drugs and vaccines. In this study, we analyzed samples of amino-acid sequences (AASs) for structural proteins from the coronavirus 2019 (COVID-19) declaration as a pandemic to April 2022 among American countries. The analysis process included considering mutations' frequencies, locations, and evolutionary trends utilizing sequence alignment to the reference sequence. In the following, the results were compared with the same analyses among the samples of the entire world. Results displayed that despite samples of North America and international countries that own the region of 508 to 635 with the highest mutation frequency among S AASs, the region with the same characteristic was concluded as 1 to 127 in South America. Besides, the most frequent mutations in S, E, M, and N proteins from North America and worldwide samples were concluded as D614G, T9I, I82T, and R203M. In comparison, R203K was the first frequent mutation in N samples in South America. Widely comparing mutations between North America and South America and between the Americas and the world can help scientists introduce better drug and vaccine development strategies.

# Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries

## Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a member of the *Betacoronavirus* genus of the *Coronaviridae* family. Viruses of this family have positive-sense single-stranded enveloped RNA genomes, ranging from 26 to 32 kilobases in length [1]. Since the outbreak started in Wuhan, China, more than 6 million people have been killed by the Coronavirus 2019 (COVID-19) pandemic worldwide [2].

To date, 14 open reading frames (ORFs) encoding 27 structural and non-structural proteins (NSPs) have been identified for SARS-CoV-2. The 5' end of the SARS-CoV-2 genome contains genes that encode pp1ab and pp1a proteins that together include 15 NSPs (NSP1-NSP10 and NSP12-NSP16), while the 3' end of its genome contains genes that encode four structural proteins called spike (S), envelope (E), membrane (M) and nucleocapsid (N) as well as eight accessory proteins called 3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14 [3]. The S glycoprotein is embedded on the outer surface of the virus. By interacting with the angiotensin-converting enzyme 2 (ACE2) receptor, S protein trimers facilitate viral entry into host cells [4]. E protein plays an integral role in the production and maturation processes of the virus through interaction with the host cell membrane protein [5]. The M protein determines the shape of the E protein and contributes to the packaging of the viral RNA genome into a helical ribonucleocapsid in the process of virion formation [6]. The N protein, bound to the viral RNA genome, controls the viral replication cycle, the viral genome signaling, and the host cell reaction to the viral infection [5].

With the hope for a safe and effective strategy to tackle the COVID-19 problem, the interaction between drugs and structural proteins has been the subject of several therapeutic strategies [7-11]. Also, different vaccination pipelines have been developed mainly targeting the S protein [12, 13]. On the other hand, mutations of the structural proteins could increase the viral load, infectivity, and immune escape. For example, B.1.617.2 (Delta) and B.1.1.529 (Omicron) are more transmissible and resistant to neutralization due to mutations found in the structural proteins, especially the S protein [14-16]. Of note, the D614G, located in the S protein, is the most common variant of SARS-CoV-2, which increases infectivity [17]. These alterations may provide challenges in tackling the COVID-19 problem suggesting the importance of mutations in determining the quality of vaccinations and therapeutic strategies.

Our research aimed to determine the frequency of mutations in SARS-CoV-2 structural proteins in different American countries and also the entire world. We also investigated the evolutionary patterns of mutations based on their properties and ultimately, the mutations features will be compared between the Americas and worldwide data.

## Materials and Methods

# **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

## **Sequence gathering**

Mutation tracking was performed by gathering the SARS-CoV-2 AASs and comparison the data with Wuhan-2019 virus as reference sequence (access number: EPI\_ISL\_402124). Extraction of samples was conducted from GISAID database with the permission of Erasmus Medical Center [18-20]. All involved samples in the study were for human and samples with unknown time and location, different AAs length and non-specified AAs were kept out. Details of AA regions and the lengths of structural proteins were depicted in figure 1.

## **Sequences analyses**

Sequences alignment and analysis was performed in the form of FASTA files with Python 3.8.0 software. As it mentioned above, Wuhan-2019 was utilized as reference sequence and mutation/s was introduced via detection of the difference/s between samples and this sequence. Optimization of data was completed by applying 'Numpy' and 'Pandas' libraries.

Mutation tracking was conducted using following algorithm:

```
> For refitem, seqitem in zip (refseq, seq)
If (refitem! =seqitem)
Report a new mutant
```

Because of the equal lengths among all sequences, the following algorithm used 'Refseq,' and 'seq' refer to reference sequence and sample sequence, respectively. Ultimately, reports contain mutations, locations and subset AA and the figures were drawn using R 4.2.0.

## **Data normalization and statistical analysis**

In order to compare the data with more accuracy, normalization was conducted using R 4.0.3 and Microsoft Power BI. Therefore, for each country, the number of mutations was divided by the number of attributed sequences that were comparable in equal proportions.

## **Results**

### **Results related to the quantity of mutations**

Investigating the GISAID database led to extracting 9670364 qualified samples from North America and 639746 eligible samples of AASs from South America. Besides, 26090908 AASs sequences belonging to structural proteins of SARS-CoV-2 were extracted globally. Samples of North America included 427550 samples for S AASs, 3597547 samples for E AASs, 3182356 samples for M AASs and 2462911 samples for N AASs. In comparison, there were 64546,

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

217344, 202987 and 153869 samples from South America attributed to S, E, M and N AASs, respectively. Additionally, worldwide samples included 950459 samples for S AASs, 9914529 samples for E AASs, 8860463 samples for M AASs and 6365457 samples for N AASs.

2.13% of S samples attributed to North America did not carry any mutation and 3.18%, 20.58%, 9.99% and 35.42% of them displayed one, two, three and more than three mutations, respectively (Figure 2). Considering the S sequences of South America demonstrated different situation, in which 80.52% of samples carried more than three mutations and only 0.89% of S AASs did not have any mutation. Furthermore, 4.82% of worldwide results displayed no mutation and 26.31%, 25.31%, 13.75% and 29.79% of such samples included one, two, three and more than three mutations, respectively. More than 99% of E AASs from both North America and South America displayed no mutation or maximum one mutation. Among samples of North America, 72.85% and 27.01% of them carried no mutation and one mutation, respectively and in almost similar results from South America, 76.93% and 22.93% of samples included no mutation and one mutation, successively. In addition, 67.72% and 32.09% of global E samples demonstrated no mutation and one mutation, respectively.

Studying of M AASs demonstrated not incidence of mutations in 27.05%, 44.88% and 26.30% of samples belonging to North America, South America and entire world. Among sequences of North America, incidence rate of one, two, three and more than three mutations was concluded as 49.67%, 5.82%, 17.21% and 0.23%, successively. These former parameters for South America were observed as 35.14%, 3.74%, 16% and 0.23%, respectively. Moreover, analyzing the N AASs displayed no incidence of mutation in 2.21% of samples from North America and 5.46% and 73.03% of such samples included one and more than three mutations, respectively. Also 1.04% of samples belonging to South America did not display any mutation and almost 91% of such samples included three and more than three mutations, in which 35.24% had three mutations and 55.09% of them showed more than three mutations.

Results also were utilized in drawing a heat map for each of structural AASs to finding the regions with high predisposition toward mutation incidence. Although the region of 508 to 635 was concluded as the region with highest mutations frequency among S AASs from North America and also entire world, the most mutations relative to the total AASs among South America was occurred in the region of 1 to 127 (Figure 3). On the other hand, this different result was not observed in E samples, i.e. the region with highest mutations frequency was identical in all of them (7 to 14). Conversely, analyzing the data from North America and Worldwide samples introduced the region of 66 to 88 among M AASs as region with highest frequency rate. But highest frequency of mutations was occurred in the region of 1 to 22 in South America samples. On top of that, the region of 164 to 205 was concluded as region carrying highest frequency of mutation incidence in N samples from North America, South America and worldwide sources.

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

### **Discovered mutations among structural proteins**

D614G was found as the most frequent mutation among S AASs in both North America and South America with almost similar frequency (0.9837 in North America and 0.9957 in South America) (Figure 4). But arrangement of subsequent mutations was different between sequences of these two continents. E484K (0.1357 frequency) and L18F (0.1204 frequency) were detected as second and third frequent mutations in S AASs from North America. On the other hand, V1176F with 0.8466 frequency and E484K with 0.8118 frequency were introduced as second and third frequent mutations in S AASs belonging to South America. Worldwide data demonstrated that D614G (0.9764 frequency), E484K (0.1406 frequency) and L18F (0.1623 frequency) were top three frequent mutations, globally. Furthermore, T9I (0.2574 frequency), P71L/S (0.0016/0.0008 frequencies) and L21F/V/I (0.0013/0.0006/0.0001 frequencies) were first to third frequent mutations in E proteins attributed to North America. Additionally, first and second frequent mutations among such proteins of South America were similar to mutations of North America, but third frequent mutation was observed as V58F with 0.0015 frequency. T9I with 0.3064 frequency, P71L with 0.0046 frequency and V62F with 0.0022 frequency were observed as first to third frequent mutations entire world.

Interestingly, although first three frequent mutations of M AASs from both North America and South America were similar in types of mutations, their arrangements are noticeably different. I82T (0.5121 frequency in samples from North America and 0.3416 Frequency in samples from South America) was the most frequent M mutation in both continents but Q19E (0.1990 frequency) and A63T (0.1986 frequency) were second and third frequent mutations in M AASs from North America. In comparison, A63T with 0.1869 frequency and Q19E (0.1777 frequency) were found as second and third frequent mutations in South America; which had identical arrangement with top three frequent mutations globally. Also, considering the mutation frequencies among N AASs were indicated that R203M/K (0.6457/0.1689 frequencies), D377Y (0.6562 frequency) and D63G (0.6404 frequency) were first to third frequent mutations in North America. Further, R203K/M (0.5161/ 0.4128 frequencies), G204R (0.5147 frequency) and D377Y (0.4178 frequency) were concluded as first three frequent mutations in South America. Ultimately, R203M/K (0.6282/0.2359 frequencies), D377Y (0.6329 frequency) and D63G (0.6202 frequency) were showed as frequent mutations entire world with similar arrangements to North America. Additional data about the Americas have been listed in supplementary files S-FREQ-AMR, E-FREQ-AMR, M-FREQ-AMR and N-FREQ-AMR and data belonging to the global samples have been listed in supplementary files S-FREQ-WRD, E-FREQ-WRD, M-FREQ-WRD and N-FREQ-WRD.

### **Trends for emergence and distribution of structural mutations**

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

Timeline analyses of mutations displayed similar pattern of emergence and distribution trend of D614G among both continents and also entire world (Figure 5). D614G frequencies gained to increase from February 2020 and up to now saved its maximum frequency in almost fluctuated trend. What's more, from November 2021, T9I mutation gained to increase in both continents and received its maximum frequency in close to January 2022. Evolutionary patterns of M AASs showed that in January 2022, when I82T mutation was in next to minimum frequency, Q19E and A63T mutations had noticeably higher frequency than I82T and gained to increase in a growing trend started from November 2021. The similar trend was resulted in timeline of N AASs, i.e. when D377Y and D63G started decreasing trends of frequencies from December 2021, G204R gained to noticeable increase in both continents. Supplementary files S-EVL-AMR, E-EVL-AMR, M-EVL-AMR, N-EVL-AMR, S-EVL-WRD, E-EVL- WRD, M-EVL- WRD, N-EVL- WRD include supplemental data about the evolutionary patterns of mutations.

### **Discussion**

Keeping an eye on the evolution of SARS-CoV-2 and monitoring emerging mutations is essential for developing new diagnostic tests, vaccines, and treatments against COVID-19 in this ongoing pandemic. More than 90% of samples in each group had at least one mutation in the S or the N proteins in our study. On the other hand, the E and the M proteins had the least mutation rate. This is consistent with the data from Mutation Tracker [21]. According to Sato et al., the structural proteins had ~99% conservation until 2020, with the S protein having the highest proportion of AA changes along its sequence [22].

Substituting aspartic acid (D) to glycine (G) in residue 614 of the S protein has been one of the most frequent mutations in the viral genome. Evidence shows that the D614G facilitates cell entry, replication, infection, and fitness [23, 24]. D614G enhances binding to ACE2 [24, 25]. The aspartic acid in position 614 resides on the surface of the S protein. A hydrogen bond could form between the side chains of this aspartic acid and that of its neighboring protomer, linking residues of the S1 unit of one protomer with residues of the S2 unit of the neighboring protomer. Substitution of D614 with glycine would disrupt this side-chain hydrogen bond [26], resulting in a more stable structure with more S protein incorporation into the pseudovirion [27]. We found that T9I is now the most frequent among the E protein mutations, while in a previous study, S68F was recognized as the most common mutation in the E protein [22]. The prevalence of this mutation in Omicron strain has been determined as ~100% [28, 29]. The 9th position of E protein resides at the top of the transmembrane domain [30]. How the substitution of the hydrophilic threonine (T) to the hydrophobic isoleucine (I) affects the function of the virus is not precisely understood. However, it is suggested that this substitution could probably strengthen the interaction of the virion with membrane lipids because of isoleucine's hydrophobicity [31].

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

Although the M protein has been mainly conserved throughout the pandemic till February 2021, it underwent rapidly increasing mutations. We observed that I82T mutation is the most prevalent in North and South America, with 0.5121 and 0.3416 frequencies, respectively. This is similar to previous studies indicating the I82T mutation as one of the most frequent mutations in the M protein in the USA, with a 116-fold increase from October 2020 to February 2021 [32]. Suratekar et al. also reported I82T as one of the most prevalent mutations in the Delta variant in a systematic analysis of mutations from nearly 2 million SARS-CoV-2 genomes from 176 countries [33]. However, till 2020, D3G was reported as the most common mutation in this protein [22]. The change of a hydrophobic amino acid (Isoleucine) to a more polar one (Threonine) has not been shown to alter the structure of the M protein; however, the similarity of the M protein structure to glucose transporters suggests that this mutation may contribute to glucose uptake during viral replication [32].

Consistent with our analysis, alteration of the residue 203 in the N protein has been reported as the most frequent mutation in this protein. Our analysis consistently showed that substitution mutation in the region of 203 was the most common change in the N protein in North America, South America, and worldwide. Mutations in this residue have been seen in Alpha, Delta, and Omicron lineages [34, 35], increasing the SARS-CoV-2's infectivity and immune escape ability [34-37]. In an experiment with live a Coronavirus model, Delta variants with R203M could deposit 10 times more mRNA into host cells than wild-type viruses and generate 51 times more infectious viruses [37]. Also, another study showed that the R203M Delta variant had increased viral titer in patients [35]. This may be because R203 mutation increases the virion assembly by increasing the condensation of N-protein with viral genomic RNA [38].

Our analysis showed that the mutational hotspot regions in the S protein were positions 508 to 635 in North America and the whole world, the majority of which constitute the S1 region of the protein. In contrast, in South America, the N terminal domain (NTR), receptor-binding motif (RBM), and receptor-binding domain (RBD) were the locations with the most prominent mutations (1-127, 381-508). Consistent with this, Harvey et al. showed that residues at positions 18, 222, and 614 had the highest variation frequencies [39]. The RBD is essential as it is the region that binds to ACE2 and is targeted by vaccination platforms. The efficacy of vaccines against the Omicron variant was heavily attenuated, providing a significant concern for protection against the disease [40]. One of the responsible mutations for this evasion is R346K [41]. Surprisingly, the frequency of this mutation has bluntly decreased in the recent months in the North America, according to our data (from 0.79 to 0.16 within two months). Also, via an artificial intelligence (AI) model, researchers have found that K417N, E484A, and Y505H in the S protein could cause the most significant disruption of known antibodies to the S protein [42]. The frequency of these mutations has been rising until April 2022, but, in fact, we found that E484K, K471T, and Y505J are the most frequent amino acid changes for the 484, 417, and 505

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

positions of the S protein. This information may help provide a more precise prediction on how the functional mutations affect vaccination immunity.

A limitation of this study was not considering the nucleotide sequences of the virus and only focusing on the amino acid sequences, neglecting the aspects such as codon bias [43]. Also, the sample sizes in the dataset are biased towards Europe and North America, suggesting that some functionally significant variants may be overlooked. Overall, COVID-19 vaccination and treatment success is determined by the frequency of functional mutations and the rate at which they appear. Faster identification of mutations could speed up the adopted approaches for combating COVID-19.

### **Acknowledgments**

The authors thank all of the researchers who have shared genome data openly via the Global Initiative on Sharing All Influenza Data (GISAID).

### **Competing interests**

The authors declare that they have no conflicts of interest that might be relevant to the contents of this manuscript and the research was carried out regardless of commercial or financial relationships that may cause any conflict of interests.

### **Additional information**

**Supplementary Information:** Supplemental data associated with the current study have been gathered.



# Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries

## References

1. Naserghandi A, Allameh SF, Saffarpour R. All about COVID-19 in brief. *New Microbes New Infect.* 2020;35:100678.
2. Yang J, Yang G, Jin R, Song G, Yuan G. Changes in paediatric dental clinic after reopening during COVID-19 pandemic in Wuhan: A retrospective study. *BMJ Open.* 2022;12(1).
3. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe.* 2020;27(3):325-8.
4. Bakhshandeh B, Jahanafrooz Z, Abbasi A, Goli MB, Sadeghi M, Mottaqi MS, Zamani M. Mutations in SARS-CoV-2; Consequences in structure, function, and pathogenicity of the virus. *Microbial Pathogenesis.* 2021;154:104831.
5. Chang T-J, Yang D-M, Wang M-L, Liang K-H, Tsai P-H, Chiou S-H, Lin T-H, Wang C-T. Genomic analysis and comparative multiple sequences of SARS-CoV2. *Journal of the Chinese Medical Association.* 2020;83(6):537-43.
6. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics.* 2020;112(5):3588-96.
7. Chan WKB, Olson KM, Wotring JW, Sexton JZ, Carlson HA, Traynor JR. In silico analysis of SARS-CoV-2 proteins as targets for clinically available drugs. *Scientific Reports.* 2022;12(1):5320.
8. Strobelt R, Adler J, Paran N, Yahalom-Ronen Y, Melamed S, Politi B, Shulman Z, Schmiedel D, Shaul Y. Imatinib inhibits SARS-CoV-2 infection by an off-target-mechanism. *Scientific Reports.* 2022;12(1):5758.
9. León G, Herrera M, Vargas M, Arguedas M, Sánchez A, Segura Á, Gómez A, Solano G, Corrales-Aguilar E, Risner K, Narayanan A, Bailey C, Villalta M, Hernández A, Sánchez A, Cordero D, Solano D, Durán G, Segura E, Cerdas M, Umaña D, Moscoso E, Estrada R, Gutiérrez J, Méndez M, Castillo AC, Sánchez L, Sánchez R, Gutiérrez JM, Díaz C, Alape A. Development and characterization of two equine formulations towards SARS-CoV-2 proteins for the potential treatment of COVID-19. *Scientific Reports.* 2021;11(1):9825.
10. Peng Y, Du N, Lei Y, Dorje S, Qi J, Luo T, Gao GF, Song H. Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *Embo j.* 2020;39(20):e105938.
11. Oliveira SC, de Magalhães MTQ, Homan EJ. Immunoinformatic Analysis of SARS-CoV-2 Nucleocapsid Protein and Identification of COVID-19 Vaccine Targets. *Frontiers in immunology.* 2020;11:587615.
12. Berguer PM, Blaustein M, Bredeston LM, Craig PO, D'Alessio C, Elias F, Farré PC, Fernández NB, Gentili HG, Gándola YB, Gasulla J, Gudesblat GE, Herrera MG, Ibañez LI, Idrovo-Hidalgo T, Nadra AD, Nosedá DG, Paván CH, Pavan MF, Pignataro MF, Roman EA, Ruberto LAM, Rubinstein N, Sanchez MV, Santos J, Wetzler DE, Zelada AM, Argentinian AntiCovid C. Covalent coupling of Spike's receptor binding domain to a multimeric carrier produces a high immune response against SARS-CoV-2. *Scientific Reports.* 2022;12(1):692.
13. Shi R, Shan C, Duan X, Chen Z, Liu P, Song J, Song T, Bi X, Han C, Wu L. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature.* 2020;584(7819):120-4.
14. Laurini E, Marson D, Aulic S, Fermaglia A, Pricl S. Molecular rationale for SARS-CoV-2 spike circulating mutations able to escape bamlanivimab and etesevimab monoclonal antibodies. *Scientific Reports.* 2021;11(1):20274.

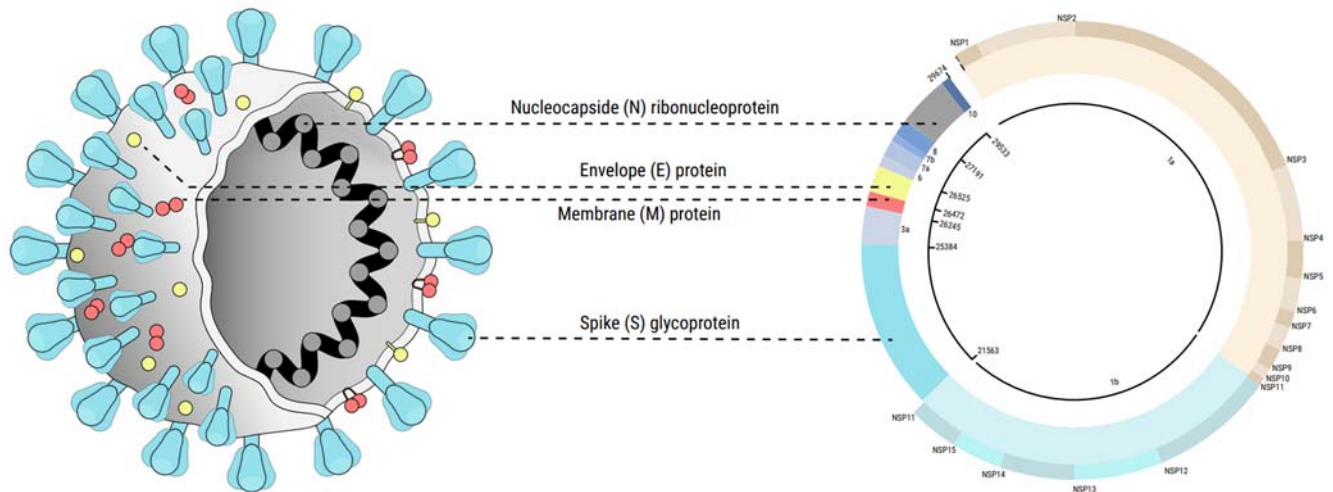
## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries

15. Gu H, Krishnan P, Ng DYM, Chang LDJ, Liu GYZ, Cheng SSM, Hui MMY, Fan MCY, Wan JHL, Lau LHK, Cowling BJ, Peiris M, Poon LLM. Probable Transmission of SARS-CoV-2 Omicron Variant in Quarantine Hotel, Hong Kong, China, November 2021. *Emerging infectious diseases*. 2022;28(2):460-2.
16. Tamimi F, Altigani S, Sanz M. Periodontitis and coronavirus disease 2019. *Periodontology* 2000. 2022.
17. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. 2020;182(4):812-27.e19.
18. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W. GISAID's Role in Pandemic Response. *China CDC Weekly*. 2021;3(49):1049.
19. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges*. 2017;1(1):33-46.
20. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. 2017;22(13).
21. Mutation Tracker. Available from: [https://users.math.msu.edu/users/weig/SARS-CoV-2\\_Mutation\\_Tracker.html](https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html), accessed [May 20, 2022].
22. Troyano-Hernández P, Reinoso R, Holguín Á. Evolution of SARS-CoV-2 Envelope, Membrane, Nucleocapsid, and Spike Structural Proteins from the Beginning of the Pandemic to September 2020: A Global and Regional Approach by Epidemiological Week. *Viruses*. 2021;13(2):243.
23. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinno KH, Leist SR, Schäfer A, Nakajima N, Takahashi K, Lee RE, Mascenik TM, Graham R, Edwards CE, Tse LV, Okuda K, Markmann AJ, Bartelt L, Silva Ad, Margolis DM, Boucher RC, Randell SH, Suzuki T, Gralinski LE, Kawaoka Y, Baric RS. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science*. 2020;370(6523):1464-8.
24. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, Rangarajan ES, Pan A, Vanderheiden A, Suthar MS, Li W, Izzard T, Rader C, Farzan M, Choe H. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communications*. 2020;11(1):6013.
25. Zhou B, Thao TTN, Hoffmann D, Taddeo A, Ebert N, Labroussaa F, Pohlmann A, King J, Steiner S, Kelly JN, Portmann J, Halwe NJ, Ulrich L, Trüb BS, Fan X, Hoffmann B, Wang L, Thomann L, Lin X, Stalder H, Pozzi B, de Brot S, Jiang N, Cui D, Hossain J, Wilson MM, Keller MW, Stark TJ, Barnes JR, Dijkman R, Jores J, Benarafa C, Wentworth DE, Thiel V, Beer M. SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature*. 2021;592(7852):122-7.
26. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. 2020;182(4):812-27.e19.
27. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izzard T, Farzan M, Choe H. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv : the preprint server for biology*. 2020:2020.06.12.148726.
28. Xia B, Wang Y, Pan X, Cheng X, Ji H, Zuo X, Jiang H, Li J, Gao Z. Why is the SARS-CoV-2 Omicron variant milder? *The Innovation*. 2022;3(4):100251.

## **Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries**

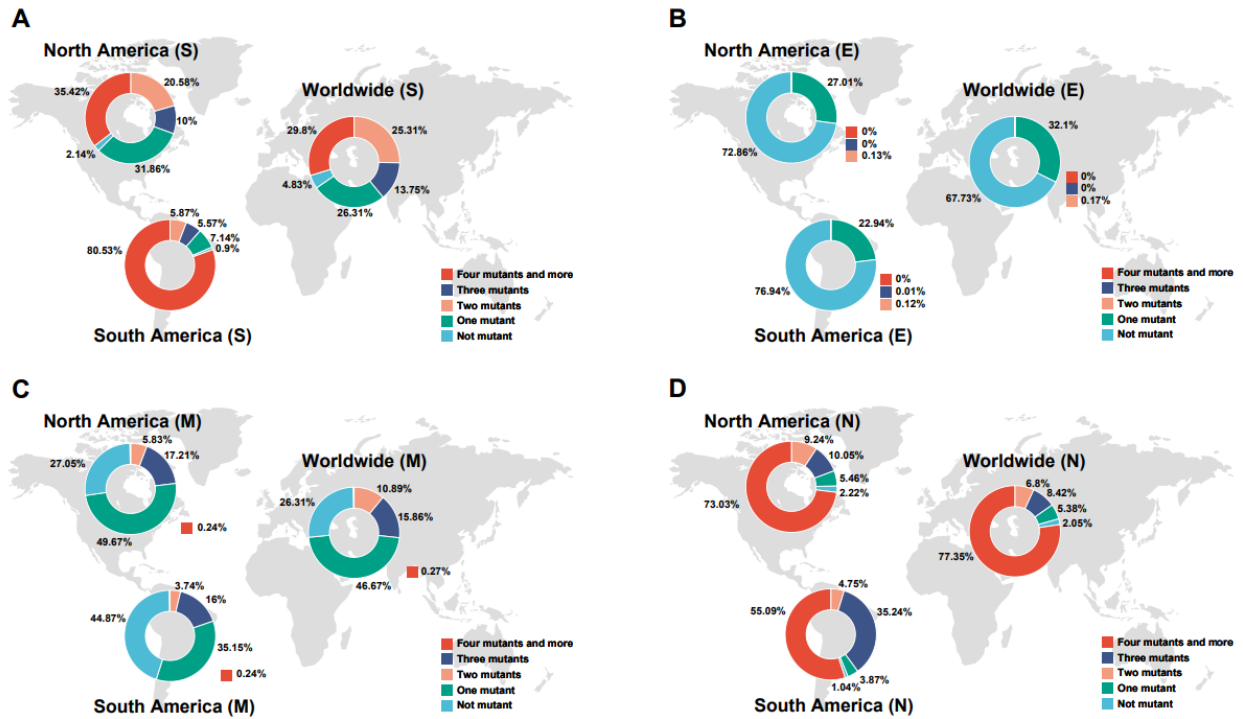
29. Kannan SR, Spratt AN, Sharma K, Chand HS, Byrareddy SN, Singh K. Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. *J Autoimmun.* 2022;126:102779-.
30. Mandala VS, McKay MJ, Shcherbakov AA, Dregni AJ, Kolocouris A, Hong M. Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nature Structural & Molecular Biology.* 2020;27(12):1202-8.
31. Timmers LFSM, Peixoto JV, Ducati RG, Bachega JFR, de Mattos Pereira L, Caceres RA, Majolo F, da Silva GL, Anton DB, Dellagostin OA, Henriques JAP, Xavier LL, Goetttert MI, Laufer S. SARS-CoV-2 mutations in Brazil: from genomics to putative clinical conditions. *Scientific Reports.* 2021;11(1):11998.
32. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. *Emerging Microbes & Infections.* 2021;10(1):885-93.
33. Suratekar R, Ghosh P, Niesen MJM, Donadio G, Anand P, Soundararajan V, Venkatakrishnan AJ. High diversity in Delta variant across countries revealed by genome-wide analysis of SARS-CoV-2 beyond the Spike protein. *Molecular Systems Biology.* 2022;18(2):e10673.
34. Lin X, Xue W, Zhang Y, Fu B, Trimpert J, Xing N, Kunec D, Tang W, Xiao Y, Meng K, Shi S, Wu H, Meng G, Zhu Z. Nucleocapsid 203 mutations enhance SARS-CoV-2 immune evasion. *bioRxiv.* 2021:2021.12.20.473471.
35. Li B, Deng A, Li K, Hu Y, Li Z, Shi Y, Xiong Q, Liu Z, Guo Q, Zou L. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *Nature Communications.* 2022;13(1):1-9.
36. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, Tang W, Xiao Y, Liu G, Luo H, Zhu W, Lin X, Meng G, Zhu Z. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host & Microbe.* 2021;29(12):1788-801.e6.
37. Syed AM, Taha TY, Tabata T, Chen IP, Ciling A, Khalid MM, Sreekumar B, Chen P-Y, Hayashi JM, Soczek KM. Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles. *Science.* 2021;374(6575):1626-32.
38. Zhao M, Yu Y, Sun LM, Xing JQ, Li T, Zhu Y, Wang M, Yu Y, Xue W, Xia T, Cai H, Han QY, Yin X, Li WH, Li AL, Cui J, Yuan Z, Zhang R, Zhou T, Zhang XM, Li T. GCG inhibits SARS-CoV-2 replication by disrupting the liquid phase condensation of its nucleocapsid protein. *Nat Commun.* 2021;12(1):2114.
39. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, Robertson DL, Consortium C-GU. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology.* 2021;19(7):409-24.
40. Hoffmann M, Krüger N, Schulz S, Cossmann A, Rocha C, Kempf A, Nehlmeier I, Graichen L, Moldenhauer AS, Winkler MS, Lier M, Dopfer-Jablonka A, Jäck HM, Behrens GMN, Pöhlmann S. The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell.* 2022;185(3):447-56.e11.
41. Iketani S, Liu L, Guo Y, Liu L, Chan JF, Huang Y, Wang M, Luo Y, Yu J, Chu H, Chik KK, Yuen TT, Yin MT, Sobieszczyk ME, Huang Y, Yuen KY, Wang HH, Sheng Z, Ho DD. Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *Nature.* 2022;604(7906):553-6.
42. Chen J, Wang R, Gilby NB, Wei G-W. Omicron (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance. *ArXiv.* 2021:arXiv:2112.01318v1.
43. Plotkin JB, Dushoff J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences.* 2003;100(12):7152-7.

## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries



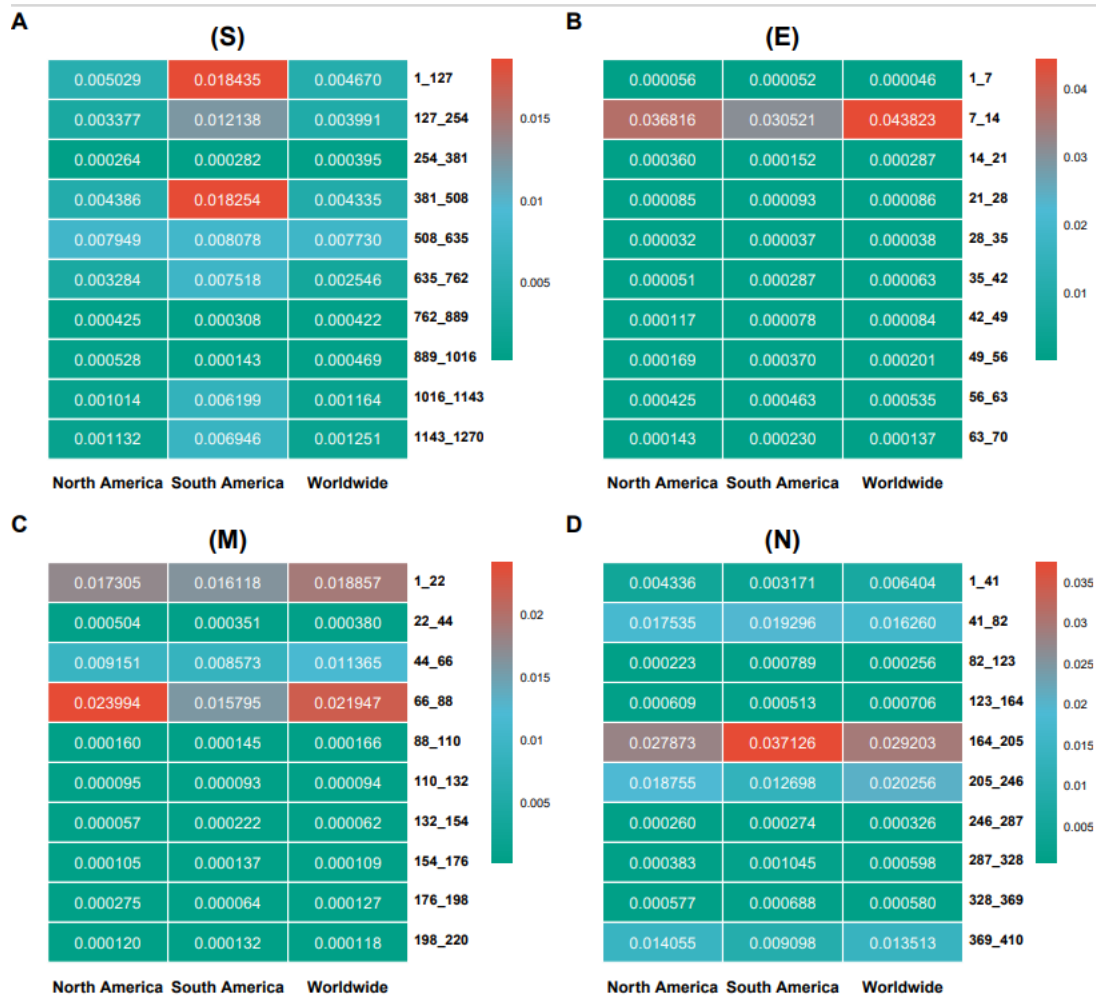
**Figure 1** the lengths and regions of structural proteins in SARS-CoV-2. E, M, N and S protein are 75, 222, 419 and 1273 in length, respectively. The regions of 26245-26472, 26523-27191, 28274-29533 and 21563-25384 are the genomic positions of E, M, N and S, respectively.

## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries



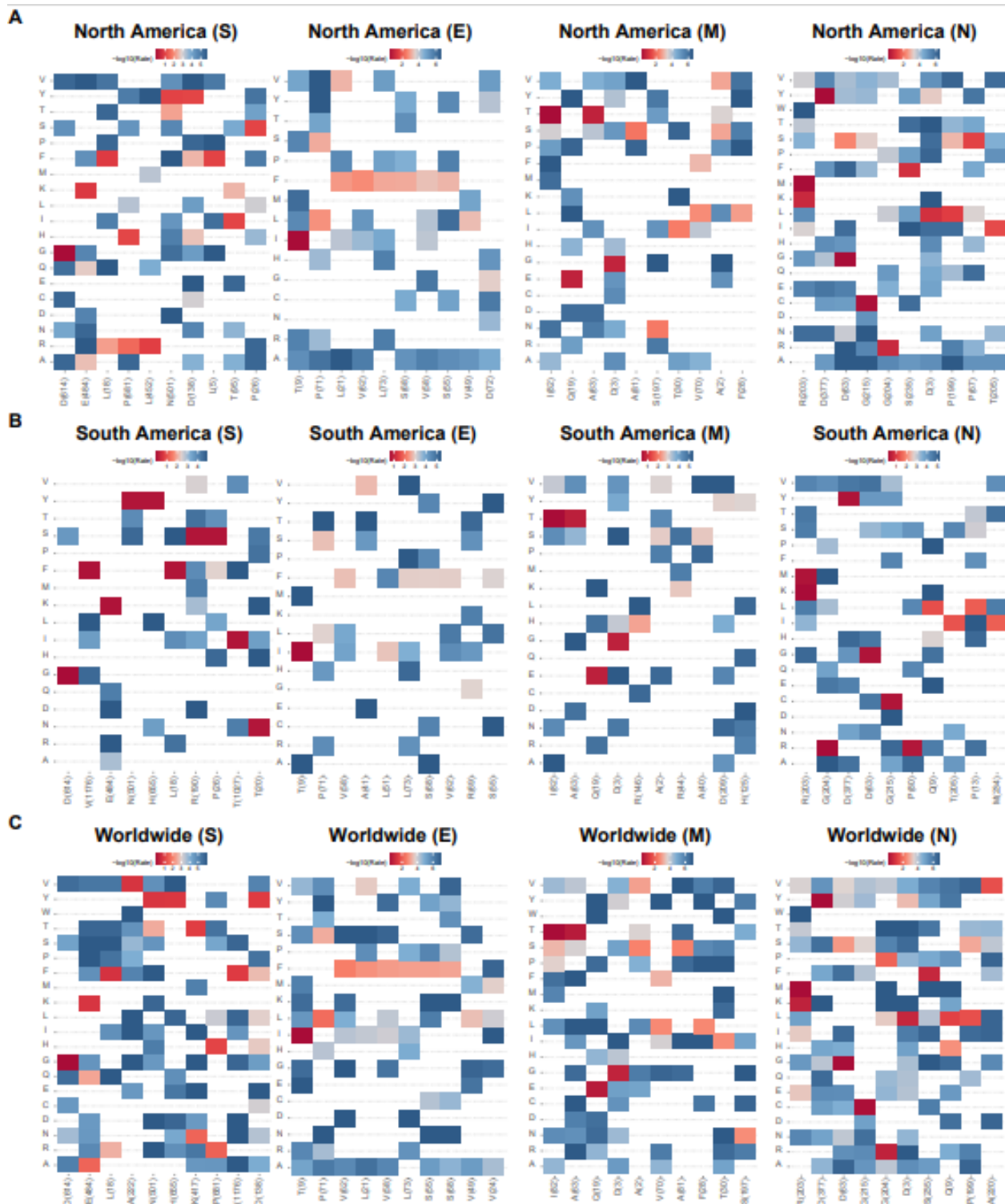
**Figure 2** Pie chart plot of the number of mutations in S, E, M, N proteins of SARS-CoV-2 as of April 2022. Results belonging to North America, South America and Worldwide samples were depicted in the sections of A, B and C, respectively.

## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries



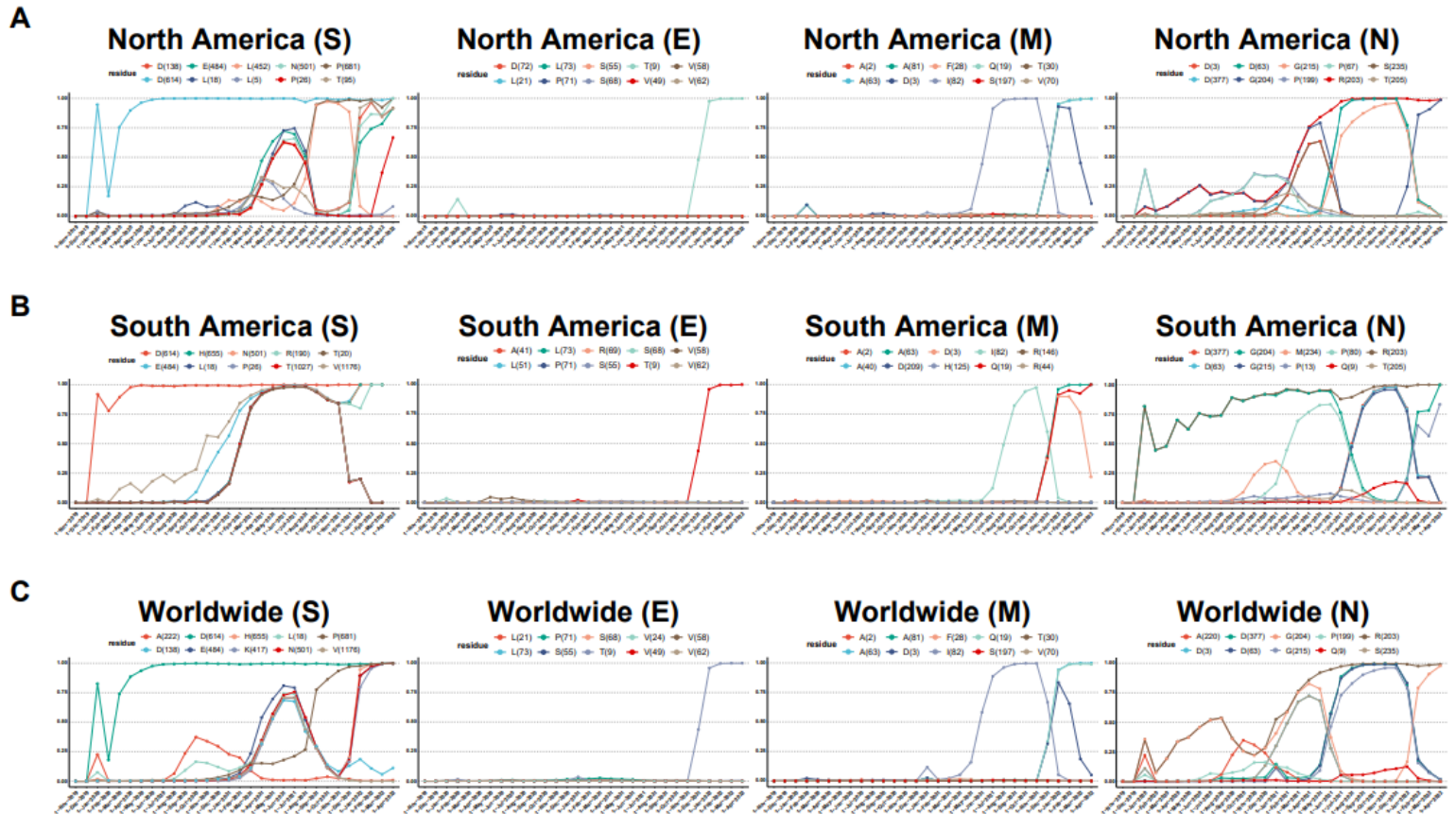
**Figure 3** Heat map of hotspot regions in S (A), E (B), M (C) and N (D) AAS of SARS-COV-2 up to April 2022 among North America, South America and Worldwide. The heat map indicates the rate of mutation per 100 AAs.

## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries



**Figure 4** Top 10 mutations in S, E, M, N of SARS-CoV-2 with the highest frequency in (A) North America, (B) South America and (C) Worldwide. The position of altered amino acids and substituted ones is shown differently based on the frequency percentage of substituted AA. The mutation frequency was estimated for each of them by normalizing the number of genomes carrying a given mutation in a desired geographic area.

## Mutational insights among the structural proteins of SARS-CoV-2: frequencies and evolutionary trends in American countries



**Figure 5** time evolution trajectories of top ten high-rate mutations of S, E, M, N of SARS-CoV-2 in different geographic areas including (A) North America, (B) South America, (C) Worldwide. According to the month of sample collection, data is computed as the number of AASs having a given mutation over the total number of AASs according to the month of sample collection.