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

Gene Reports 27 (2022) 101554

Contents lists available at ScienceDirect

Gene Reports

journal homepage: www.elsevier.com/locate/genrep

Insights into the mutation T1117I in the spike and the lineage B.1.1.389 of SARS-CoV-2 circulating in Costa Rica

Jose Arturo Molina-Mora

Centro de Investigación en Enfermedades Tropicales (CIET) & Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica

ARTICLE INFO

Keywords:
SARS-CoV-2
T1117I
Lineage B.1.1.389
Costa Rica
COVID-19

ABSTRACT

Emerging mutations and genotypes of the SARS-CoV-2 virus, responsible for the COVID-19 pandemic, have been reported globally. In Costa Rica during the year 2020, a predominant genotype carrying the mutation T1117I in the spike (S:T1117I) was previously identified. To investigate the possible effects of this mutation on the function of the spike, *i.e.* the biology of the virus, different bioinformatic pipelines based on phylogeny, natural selection, and co-evolutionary models, molecular docking, and epitopes prediction were implemented.

Results of the phylogeny of sequences carrying the S:T1117I worldwide showed a polyphyletic group, with the emergence of local lineages. In Costa Rica, the mutation is found in the lineage B.1.1.389 and it is suggested to be a product of positive/adaptive selection. Different changes in the function of the spike protein and more stable interaction with a ligand (nelfinavir drug) were found. Only one epitope out 742 in the spike was affected by the mutation, with some different properties, but suggesting scarce changes in the immune response and no influence on the vaccine effectiveness.

Jointly, these results suggest a partial benefit of the mutation for the spread of the virus with this genotype during the year 2020 in Costa Rica, although possibly not strong enough with the introduction of new lineages during early 2021 which became predominant later. In addition, the bioinformatic analyses used here can be applied as an *in silico* strategy to eventually study other mutations of interest for the SARS-CoV-2 virus and other pathogens.

1. Introduction

The SARS-CoV-2 virus, responsible for the COVID-19 pandemic, has been detected in more than 181 million people worldwide and more than 365,000 cases in Costa Rica (with at least 4650 deaths) until June 30th, 2021. As part of the epidemiological surveillance of the pandemic in Costa Rica, we have studied both genomic features of the virus and clinical and demographic patterns among diagnosed patients (Molina-Mora et al., 2021a; Molina-Mora et al., 2021b).

Emerging mutations in the spike and genotypes have been reported globally, including variants of interest (VOI) and variants of concern (VOC), which are still under study owed to possible or confirmed changes in transmission, severity, clinical manifestations, mortality, or vaccine effectiveness (Graham et al., 2021). In Costa Rica, during 2020 the predominant SARS-CoV-2 genome was a genotype carrying the mutation T1117I in the spike (S:T1117I), as we reported previously (Molina-Mora et al., 2021a). This local genotype, now classified as a Costa Rican PANGOLIN lineage B.1.1.389, reached up to 30% of cases in this country by December 2020 (Molina-Mora et al., 2021a). Furthermore, this lineage was distributed among all the distinct clusters or clinical profiles from Costa Rican cases of COVID-19, as we analyzed using machine learning (Molina-Mora et al., 2021b). For other geographic locations, other lineages carrying the mutation S:T1117I have been reported but, unlike Costa Rica, with a frequency $< 1\%$ (https://www.gisaid.org/).

The growing interest in the spread of SARS-CoV-2 genotypes, including the VOC/VOI worldwide and the lineage B.1.1.389 in Costa Rica, is mainly explained by the mutations in the spike protein. Spike protein is relevant not only because of its interaction with the receptor in human cells (ACE2, angiotensin-converting enzyme 2) (Zhou et al., 2020), but also it is implicated in the activation of the immune response against the virus by natural infection or vaccination (Graham et al., 2021; Koyama et al., 2020). Hence, changes in the spike protein sequence could affect the transmission and clinical manifestations (Graham et al., 2021; Toyoshima et al., 2020), as well as the vaccine effectiveness (Koyama et al., 2020), *i.e.*, the biology of the virus.

The spike protein is composed of 1273 amino acids and has two domains S1 (aa14–685) and S2 (aa686–1273) that are responsible for

E-mail address: jose.molinamora@ucr.ac.cr.

https://doi.org/10.1016/j.genrep.2022.101554
Received 13 September 2021; Received in revised form 29 January 2022; Accepted 4 February 2022
Available online 8 February 2022

the binding step to the ACE2 receptor. The recognition and binding to receptors are part of the activity of the S1 domain, specifically with the RBD (receptor-binding domain, aa319–541). A subsequent conformational change of the S2 domain, which harbors the putative fusion PF peptide (aa788–806) and the heptad repeats HR1 (aa912–984) and HR2 (aa1163–1213), facilitates the fusion between the envelope protein of the SARS-CoV-2 and the plasma membrane of the host cell (Astuti and Ysrafil., 2020; Xia et al., 2020).

In the case of the mutation S:T1117I, despite being located between the HR1 and HR2 regions, predictions suggested possible effects on viral oligomerization needed for cell infection, and more studies were demanded to investigate the possible changes in transmissibility, severity, or vaccine effectiveness (Molina-Mora et al., 2021a). To this end, this work aimed to study in-depth the biological effects of the mutation S:T1117I on the function of the spike protein. To address that, different bioinformatic pipelines and mathematical models were implemented to assess the phylogenetic relationships among genotypes, possible natural selection in the local spread, co-evolutionary models, interaction with molecules, as well as an immune activity using genome sequencing data. Furthermore, this work offers an integrative strategy using distinct bioinformatic analyses which are usually reported independently. Thus, this implementation can be eventually applied to other mutations in the spike, proteins of SARS-CoV-2, or other pathogens.

2. Methods

2.1. SARS-CoV-2 sequences

All the available SARS-CoV-2 sequences with the mutation S:T1117I, up to April 30th and from all the locations worldwide, were retrieved from the GISAID database (Global Initiative on Sharing All Influenza Data, www.gisaid.org). Details for each genome is summarized in the Supplementary Material. The same database was used to obtain all the sequences from Costa Rica during the same period. Epidemiological and genomic data for all the sequences by country or lineage were summarized using the Outbreak.info tool (https://outbreak.info/).

2.2. Phylogenetic analysis of sequences with the mutation S:T1117I

All the worldwide sequences with the mutation S:T1117I were aligned using MAFFT v7.471 (Katoh et al., 2002). The construction of the phylogenetic tree model was achieved with IQ-TREE v1.6.12 (Minh et al., 2020), including ModelFinder (Kalyaanamoorthy et al., 2017) to select the best nucleotide substitution model (using the Bayesian Information Criterion BIC, the best model was TN + F + I). Visualization was done using the iTOL tool v4 (Letunic and Bork, 2019).

2.3. Positive selection analysis of mutations

To detect sites in the spike protein of positive/adaptive and negative/purifying selection, we performed an analysis of natural selection using HyPhy (hypothesis testing using phylogenies) (Kosakovsky Pond et al., 2005). To this end, all the whole genome sequences from Costa Rica used for the selection of the genome location for the spike protein (ORF coordinates of the NC_045512.2 reference genome) with *getfasta*-Bedtools package (Quinlan and Hall, 2010). Sequences were aligned with MAFFT v7.471 (Katoh et al., 2002). The spike sequence-based phylogenetic tree model was built using IQ-TREE v1.6.12 (Minh et al., 2020) using ModelFinder (Kalyaanamoorthy et al., 2017) to infer the best nucleotide substitution model (using the Bayesian Information Criterion BIC, the best model was GTR + I). Subsequently, the analysis of sites under pervasive, *i.e.*, consistently across the entire phylogeny, diversifying selection was achieved the Bayesian inference FUBAR (Fast, Unconstrained Bayesian AppRoximation) model (Murrell et al., 2013). In detail, we calculated the ratio of non-synonymous to synonymous substitutions (dN/dS) values using the default parameters for the Bayesian model on a per-site basis with the alignment and the corresponding phylogeny files for the spike sequences. The evidence for selection was assessed using posterior probabilities ($\geq$0.9, with a range 0–1), in which a substitution was classified as a product of a positive if $dN/dS > 1$ or as negative if $dN/dS < 1$. We assumed that each variant was *de novo* generated in our sequences and then it was subsequently transmitted, as well as the selection pressure for each site was constant along the entire phylogeny (Ferrareze et al., 2021; Lythgoe et al., 2021; Murrell et al., 2013).

2.4. Coevolutionary analysis of spike sequences

All the spike sequences from SARS-CoV-2 genomes from Costa Rica were aligned using MAFFT v7.471 (Katoh et al., 2002). The reference sequence was also included. Subsequently, the MISTIC2 tool with default parameters (Colell et al., 2018; Simonetti et al., 2013) was used to assess co-evolutionary couplings and to identify functionally important residues in the spike protein, using mutual information and residues conservation models. The structural model of the spike (see Molecular Docking section) was incorporated to predict effects on the structure.

2.5. Molecular docking

To assess the effects of the mutations S:D614G and S:T1117I, the two spike mutations of the lineage B.1.1.389, on the interaction of the spike with a specific ligand, we performed a molecular docking analysis. For the ligand selection, because the 1117 position is between the HR1 and HR2 regions of the spike, we selected the nelfinavir drug which is known to bind spike in the HR1 region (Musarrat et al., 2020).

The structural model of the spike (ID: 6VXX) was obtained from PDB (https://www.rcsb.org/), and the molecular structure of the nelfinavir drug, which is known, was downloaded from Drugbank (https://go.drugbank.com/). The analysis was done using the reference spike PDB model (wild type, WT) as well as with the mutated version with the mutations S:D614G and S:T1117I. The Chimera software (Pettersen et al., 2004) was used not only to generate the mutations in the spike sequence of genomes from the lineage B.1.1.389, but also to minimize energy and visualize the molecules.

Afterward, molecular docking was implemented using the DockThor server (Guedes et al., 2021), in which a grid box parameters were standardized (center x = 221.9245, center y = 208.9935, center z = 195.6175, total size x = 40 Å, total size y = 40 Å, total size z = 40 Å, and discretization = 0.42 Å) in the regions FP, HR1 and HR2 of the spike, similar to (Sixto-López et al., 2021). Comparison between the WT and mutated proteins in the docking analysis was done using energy, in which the conformation with the lowest free energy values was chosen as the most stable for each case.

2.6. Epitope analysis

Spike protein-based epitopes were predicted using IEDB tool (Immune Epitope Database, https://www.iedb.org/), including the following parameters: lineal and discontinuous peptides, for T and B cells as well as MHC ligands, class I and II MHC molecules, human host, and COVID-19 disease. The spike protein sequence (NCBI ID: YP_009724390.1) of the SARS-CoV-2 reference genome (wild type, WT) was used as a model. After the prediction was achieved, all the candidate peptides covering the position T1117 in the spike were selected. Then, the modification to I1117 (mutated, S:T1117I) was done manually. Predictions of binding and processing (T and B cells, and the MHC) were re-run with the mutated peptide using specific tools of the Epitope Analysis Resource (part of the IEDB tool). Physicochemical properties and toxicity were evaluated using ToxinPred (Gupta et al., 2013), and allergenicity prediction was done with AllergenFP tool (Dimitrov et al., 2014). Scores for all the predictions were compared for the WT and mutated peptides.

3. Results

3.1. SARS-CoV-2 genomes harboring the mutation S:T1117I define a polyphyletic group with multiple lineages around the world, including the lineage B.1.1.389 as a local genotype

A total of 1155 SARS-CoV-2 sequences were found with the mutation S:T1117I in 54 countries worldwide until April 30th, 2020. United States (USA), England, and Costa Rica are the top locations in which this mutation has been reported, with 268 (23.2%), 266 (23.0%), and 126 (10.9%), respectively. However, among all the sequenced genomes for the United States and England, the cumulative prevalence for genomes carrying S:T1117I is <0.5%, whereas for Costa Rica represents 22% of the sequenced cases up to April 30th, 2020.

The phylogenetic analysis of these sequences (Fig. 1) defines a polyphyletic group, suggesting multiple and independent origins with a divergent profile. This observation is also supported by the diversity of lineages (89 in total), in which 68.7% of genomes belongs to six distinct genotypes: B.1.1.7 (240 sequences, most from European countries), B.1.1.1 (149, most from England), B.1.1.389 (141, most from Costa Rica), B.1 (135, most from the USA), B.1.177 (73, most from European countries) and B.1.1 (56, most from England). More details are in Supplementary Table S1.

Due to Costa Rica is part of the top 3 countries reporting more genomes with S:T1117I (with 126 sequences), those genomes are part of the local lineage B.1.1.389, and the last was the predominant group according to the cumulative prevalence in this country during 2020 and early 2021, we followed our analysis studying this genotype and this mutation. The lineage B.1.1.389 is characterized by the presence of at least eight mutations among the genomes (Fig. 2A), including the D614G and T1117I in the spike. This genotype has been reported in all the seven provinces of Costa Rica with a prevalence range between 10 and 41%, reaching up to 26% out of all sequences (Fig. 2B-C). In addition, new lineages started to circulate during 2021 in Costa Rica, including the A.2.5, A.2.5.1, A.2.5.2, B.1.1.7 (alpha variant) and P.1 (gamma variant), with the subsequent reduction of the B.1.1.389 in the first months of the year 2021 (Fig. 2D).

3.2. Mutation S:T1117I found in the Costa Rican lineage B.1.1.389 is product of natural selection, with some effects on the activity of the function and interactions of the spike protein

In order to analyze the evolutionary context of the mutation S:T1117I and its relation with all the available genomes, we aligned all the 407 spike sequences from SARS-CoV-2 genomes from Costa Rica (with or without the mutation). The reference sequence was also included,

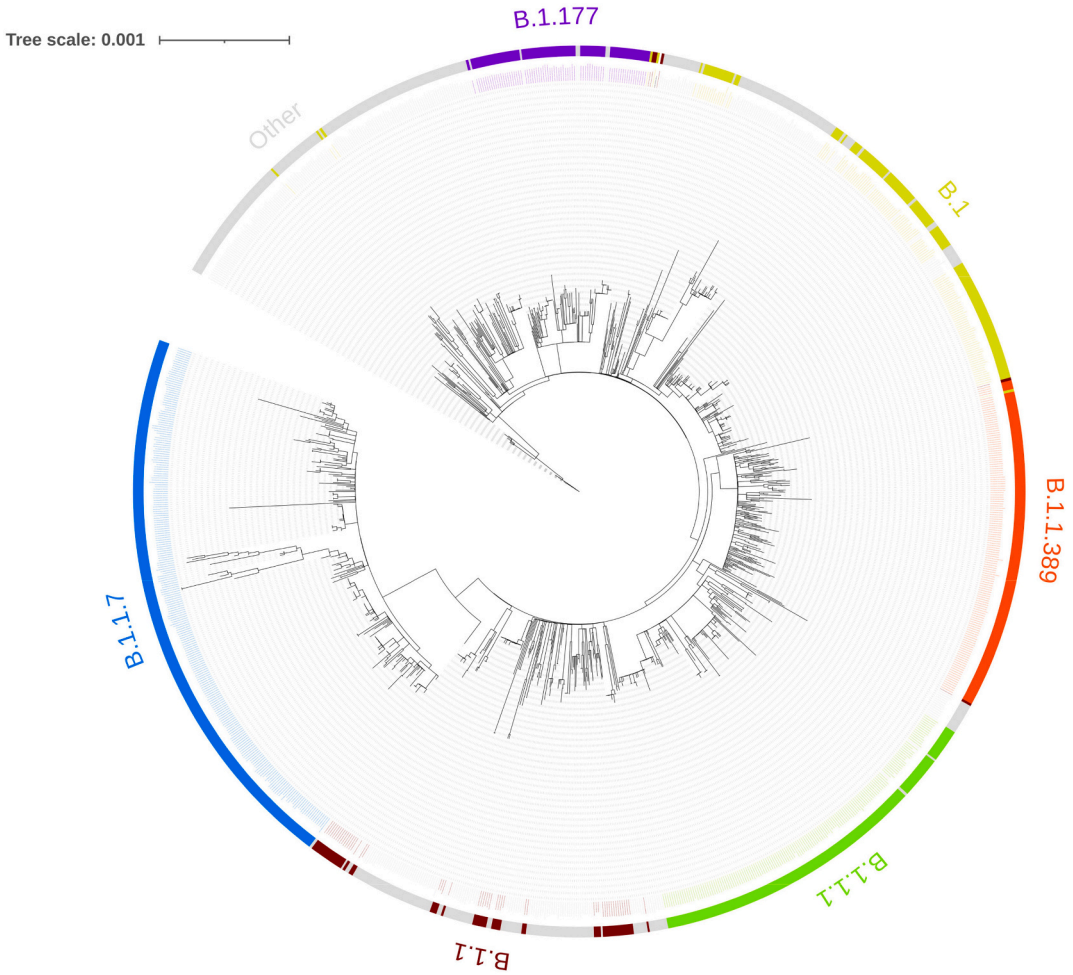
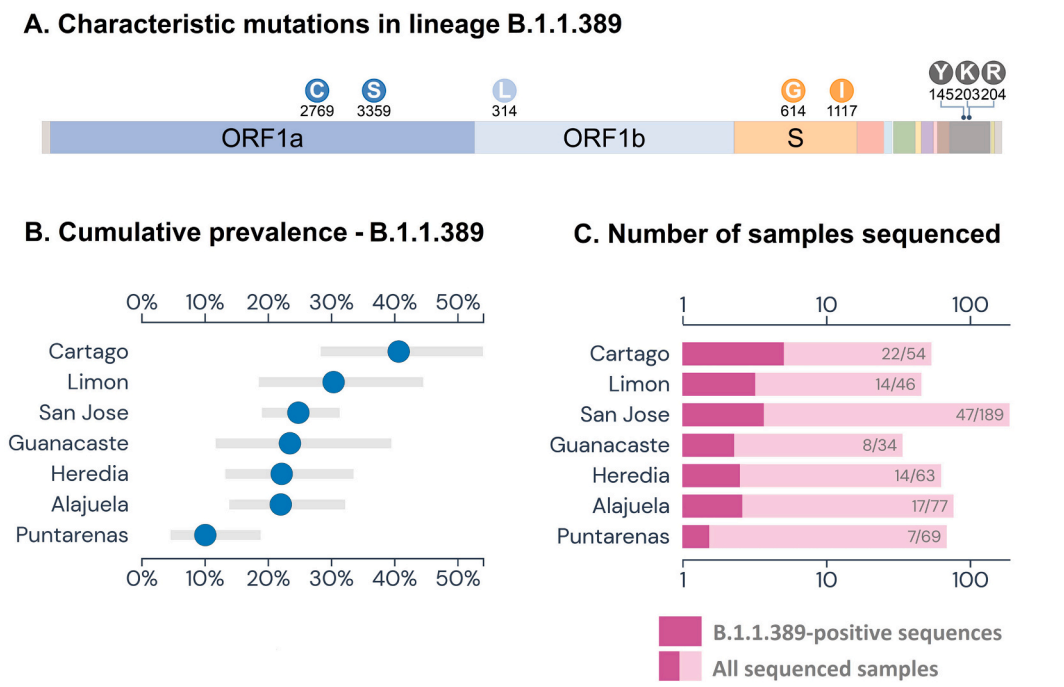


Fig. 1. Phylogenetic tree of SARS-CoV-2 genome sequence carrying the mutation T1117 in the spike (S:T1117I) of all around the world. The 1155 available sequences in GISAID database (until April 30th, 2021) with this variant are distributed according to PANGOLIN lineages. Six lineages are predominant with a frequency > 5% (colors), each one with a monophyletic origin. Rest of the lineages (frequency < 5%) were represented in gray color. The B.1.1.389 was the only lineage that carries the S:T1117I as a characteristic mutation (marker for the lineage), unlike the other groups in which this mutation is not widely found among the genomes of the lineage. In addition, B.1.1.389 is the only S:T1117I-carrying lineage with a relatively high prevalence of 22% in a specific location (Costa Rica), unlike other lineages with a prevalence <0.5% in other countries.



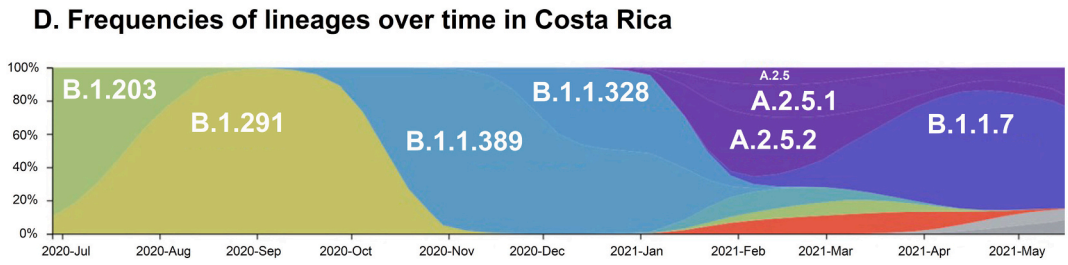


Fig. 2. Epidemiological and genomic determinants of the B.1.1.389 lineage. (A) The whole sequence of the SARS-CoV-2 reference genome is represented, including genes which are identified by different colors. Mutations of the B.1.1.389 are represented using circles. The lineage is characterized by the presence of eight mutations including two in the spike, the D614G and the T1117I variants. (B–C) The B.1.1.389 lineage has been found in all the seven provinces of Costa Rica (prevalence range 10–41%), reaching up to 22% out of all sequences from Costa Rica. (D) Relative frequencies of different lineages among all the sequences from Costa Rica over the time are represented using different colors. New lineages started to circulate during 2021 in Costa Rica, including the A.2.5, A.2.5.1, A.2.5.2, B.1.1.7 (alpha variant), and P.1 (gamma variant), with the subsequent reduction of the B.1.1.389 which was dominant during 2020. Details of the dominant lineages over time are found in the Supplementary Table S1.

completing 408 sequences.

We first studied the natural selection of mutations in the spike protein. The analysis of positive/adaptive selection of mutations among spike sequences found eight sites using the FUBAR model, in which the dN/dS was >1 (Table 1). Tellingly, the sites 501 and 1118, corresponding to the mutations S:N501Y and S:D1118H present in the lineage B.1.1.7, and 1117, site of the mutation of our interest, were included.

Table 1
Analysis of positive selection of mutations by a Fast Unconstrained Bayesian AppRoximation (FUBAR) among protein sequences of the spike of the SARS-CoV-2 from Costa Rican cases of COVID-19.

Codon	Probability [dN/dS > 1]	Empirical Bayes factor (EBF) [dN/dS] > 1	Potential scale reduction Factor (PSRF)	Effective sample size (N_eff)
26	0.96	31.57	1.01	220.44
5	0.95	29.79	1.01	218.56
80	0.95	28.63	1.01	239.95
1117	0.94	22.93	1.01	244.71
677	0.94	22.04	1.01	245.94
1118	0.93	18.41	1	639.92
501	0.91	14.36	1	337.09
1027	0.9	13.24	1	499.11

In contrast, as shown in Fig. 3, the analysis of residues found no drastic effects of the mutation S:T1117I on the spike in a co-evolutionary context using mutual information and residues conservation models. In this sense, the corrected Mutual Information (MI) was used to identify correlations between positions and the possible effect on the structure or function of the protein, revealing that the most impacted region (orange connections) are part of the RBD (positions 319–541), including the case of the mutation N501Y.

Afterward, we continued our study with a molecular docking approach assessing how the mutations in the spike are affecting the molecular interactions. The WT sequence and the mutated version with both mutations S:D614 and S:T1117I (spike sequence for the B.1.1.389 genomes) were considered for the docking with nelfinavir drug as ligand. After modeling, the drug was located in the HR1 region of S2 domain in the spike (Fig. 4), as expected. As shown in Table 2, the affinity of the nelfinavir increased because of lower energy was reported for the mutated version (−9.656 kcal/mol) in comparison to the WT spike sequence (−9.231 kcal/mol). The same pattern was observed for total and van der Waals energies, unlike the electrostatic energy.

Finally, to describe the effect of the mutation S:T1117I on the

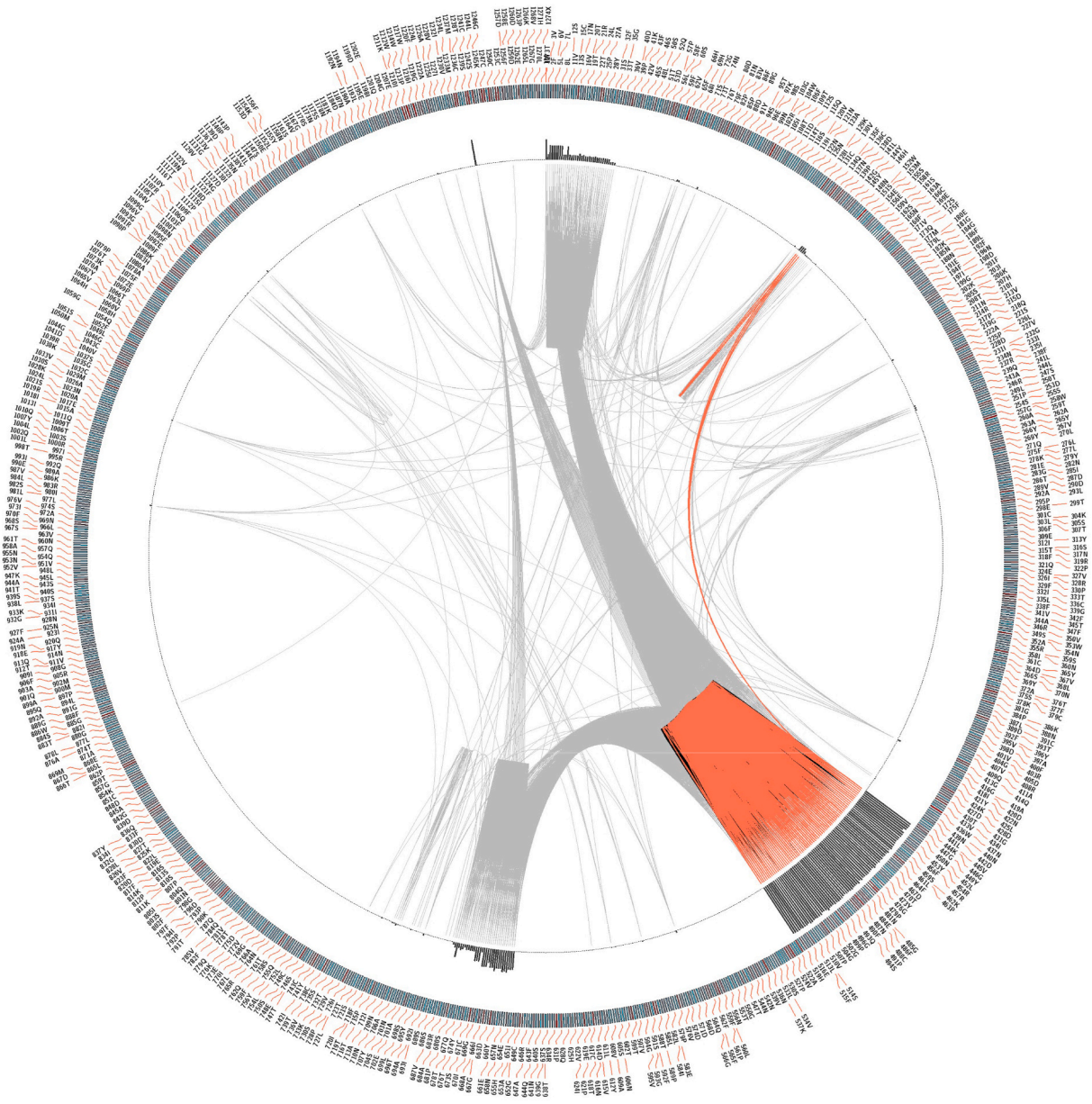


Fig. 3. Mutual coevolutionary relationship between residues in the spike protein of SARS-CoV-2 from Costa Rican cases of COVID-19. After a multiple sequence alignment of protein sequences was done, the corrected Mutual Information (MI) was used to identify correlations between positions and the possible effect on the structure or function of the spike protein. The protein sequence was presented as a circular plot, residue by residue. The most impacted region (residues with orange connections) are part of the RBD (position 319–541), including the case of the mutation N501Y. For the other residues, including the case of the mutation S:T1117I, no drastic effects are predicted according to the correlation metrics which are presented with gray connections.

immune activity, an immunoinformatic approach to study peptides was performed. Epitopes were predicted for the spike protein, including lineal or continuous peptides able to bind and be processed by T and B cells, and MHC molecules. Results are shown in Table 3. A total of 742 candidate peptides were identified but only one epitope overlapped the 1117 position (IEDB-ID: 1309561), which corresponded to a linear B cell peptide. The comparison of the WT and the mutated version of the sequence showed changes in physicochemical properties, but scores of recognition by B cells and binding to MHC molecules resulted more favorable for the WT peptide, unlike the processing by MHC-I.

4. Discussion

Genome sequencing has been a primordial step in the surveillance of the SARS-CoV-2 virus to identify mutations and new genotypes (Castillo et al., 2020b; Graham et al., 2021). According to GISAID database, more than 2 million cases have been sequenced around the world until June 30th, 2021.

With the emergence of new variants, the characterization of mutations is relevant because the virus can change transmission, severity, mortality, vaccine effectiveness, and others (Graham et al., 2021; Mercatelli and Giorgi, 2020; Tegally et al., 2020). In Costa Rica, the alpha (lineage B.1.1.7), beta (B.1.351) and gamma (P.1) VOCs have been reported in the first semester of the year 2021 (https://www.gisaid.org/), unlike the delta (B.1.617.2). However, during 2020, lineage B.1.1.389 was predominant in this country, with an accumulative prevalence of 30% by January 2021 (Molina-Mora et al., 2021a), and up to 22% by April 2021 (this study). Owed to the key mutation of the B.1.1.389 is the S:T1117I, we thereby analyzed to study the effects of this change on the spike function.

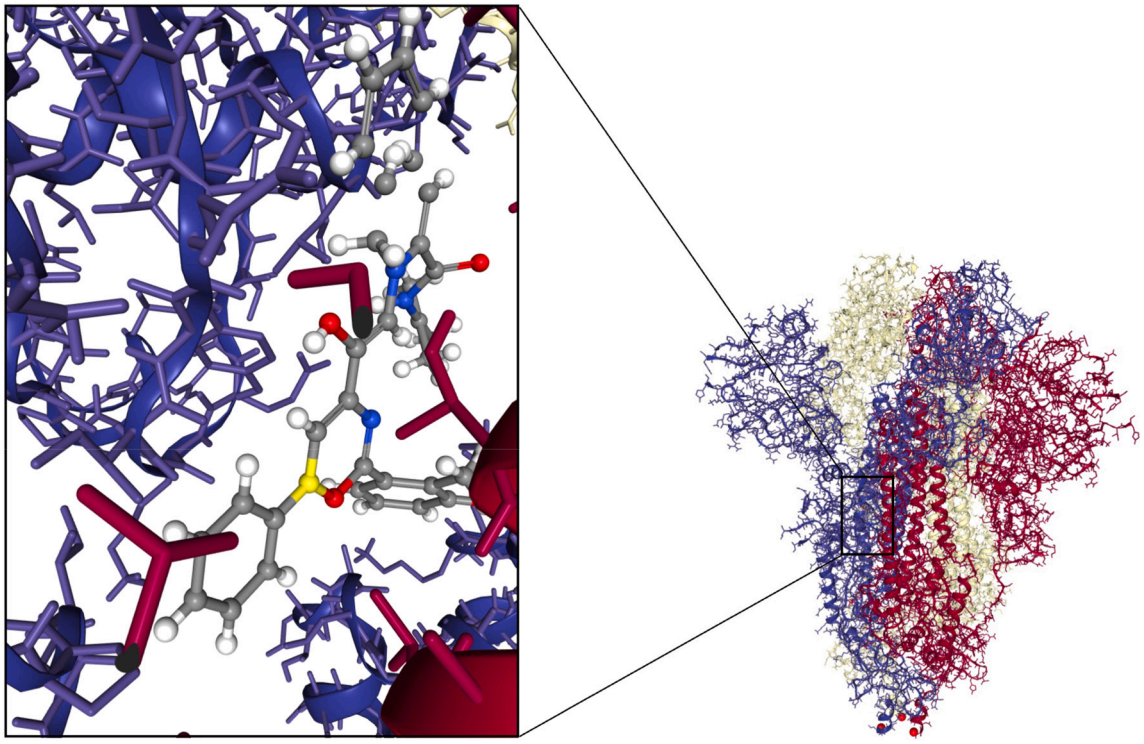


Fig. 4. Molecular docking of the complex protein-ligand using the mutated spike protein (lineage B.1.1.389) and nelfinavir drug. The drug was docked into the HR1 region of S2 domain in the spike, using the WT or the mutated (S:D614G and S:T1117I) proteins. Affinity energy predicted a more stable complex for the mutated protein in comparison with the wild type.

Table 2
Molecular docking between nelfinavir and the spike protein of the SARS-CoV-2 for WT and the mutated sequences.

Parameter	WT spike sequence (reference NC_045512.2)	Mutated spike from lineage B.1.1.389 (mutations S:D614G and S:T1117I)
Affinity (kcal/mol)	−9.231	−9.656[a]
Total energy (kcal/mol)	46,498.778	46,499.431
van der Waals energy (kcal/mol)	−22.798	−28.593
Electrostatic energy (kcal/mol)	−25.596	−19.015

[a] Best score for stability.

We started by comparing all the available sequences with the S:T1117I around the world. The 1155 sequences defined a polyphyletic tree using phylogeny, in which 89 distinct lineages were recognized, including six dominant ones. Most of the cases were identified in the USA, England, and Costa Rica, although only the last had a high cumulative prevalence of 22%, in contrast to the <0.5% for the others countries. Similar patterns with polyphyletic groups and different prevalence have been also reported for other mutations, including studies in other Latin American countries (Candido et al., 2021; Castillo et al., 2020a, 2020b; Ferrareze et al., 2021; Muñoz et al., 2021; Taboada et al., 2020).

The geographic and epidemiological analyses of COVID-19 cases in Costa Rica showed different transitions of lineages over time. Due to Costa Rica is a small country and all the main cities in each province are easily connected to the Central Valley (the most populated region of the country), the propagation patterns tend to be similar among the provinces. However, a bias of sample selection with fewer sequenced cases is recognized for far away provinces such as Guanacaste and Limón.

Afterward, we followed the analysis using natural selection models, with all the available spike sequences from Costa Rica, to determine the evolutionary role of mutations in the spike protein. Eight sites or mutations were under positive/adaptive selection, including S:T1117I. Interestingly, S:N501Y and S:D1118H from lineage B.1.1.7 (alpha variant) were also recognized. In general, these results are in line with other similar approaches worldwide, in which have been demonstrated that spike protein is under persistent positive selection (Berrio et al., 2020; Ferrareze et al., 2021; Lythgoe et al., 2021). Our analysis is the first one with the report of selection for the S:T1117I. Nevertheless, a different scenario for the S:T1117 was found using mutual information to study positions and the possible effect on structure or function. The investigation of the co-evolutionary context of the spike mutations found that the RBD is the most impacted region by the mutations, as may be expected. No drastic effects were recognized for the region associated with the mutation S:T1117I. Only a few works have implemented this approach to study mutations of SARS-CoV-2, in which a reduction of the conservation for the RBD region was also demonstrated, with implications in the interaction of the spike with antibodies (Verkhivker, 2020; Verkhivker and Di Paola, 2021). These contrasting results can be explained by the approaches, which are based on distinct modeling strategies. The natural selection models use *p-value* or the *posteriori probability* to select mutations based on the dN/dS ratio after the stochastic evolutionary models are implemented with sequence data (Kosakovsky Pond et al., 2005). In contrast, the co-evolutionary approach uses sequences and the structural model of the spike to characterize the impact of the mutation (Colell et al., 2018; Simonetti et al., 2013). Besides, although the selection analyses could suggest a selective advantage for a single mutation as in this case, a random founder effect cannot be discard.

Although sequences carrying S:T1117I from other latitudes could be included, we followed the cases from Costa Rica (lineage B.1.1.389) due to: (i) a local epidemiological interest in which B.1.1.389 represents a big group with a monophyletic origin in a particular geographic location

Table 3
Comparison of epitopes associated with the mutation S:T1117I of the SARS-CoV-2.

Peptides in the spike as candidate epitopes	742 peptides (lineal or conformational)	
Peptides overlapping the position 1117 in the spike: only one	Sequence: QRNFYEPQIITTDNTFVSGN Epitope ID: 1309561 Type: linear B cell peptide Positions in the spike protein: 1106–1125 More info: https://www.iedb.org/epitope/1309561	
Comparison of the selected epitope	Spike WT sequence (reference NC_045512.2) QRNFYEPQIITTDNTFVSGN	Spike with the mutation T1117I (lineage B.1.1.389) QRNFYEPQIITIDNTFVSGN
Hydrophobicity	−0.19	−0.15
Charge	−1	−1
Molecular weight	2344.82	2356.88
Toxicity and allergenicity analysis	Non-toxic and probable allergen	Non-toxic and probable allergen
Cell B epitope prediction: antigenic region and global score (average)	YEPQIITTDNTF Candidate based on length: yes Average score: 0.454[a]	YEP Candidate based on length: no Average score: 0.436
MHC-I processing prediction (HLA-A01:01): sequence with highest affinity	NFYEPQIITTDNTF Score: −1.69	NFYEPQIITIDNTF Score: −1.68[a]
MHC-I binding prediction (HLA-A01:01): sequence with highest affinity	TTDNTFVS Score: 2.2[a]	FYEPQIITI Score: 2.9
MHC-II binding prediction (HLA-DRB1*01:01): sequence with highest score	EPQIITTDNTFVSGN Score: 46[a]	EPQIITIDNTFVSGN Score: 20

[a] Best score in the comparison.

(Costa Rica) with a high prevalence, unlike the other cases; (ii) B.1.1.389 was the only lineage that carries the S:T1117I as a characteristic mutation (a marker for the lineage), while this mutation is not widely found among the genomes of the other lineages (for example, S:T1117I is present in a minority of cases of the B.1.1.7 lineage, although a large number of S:T1117I-carrying genomes are a lineage B.1.1.7); and (iii) the fact of considering a large dataset can create a complex scenario to get conclusive results for the analysis of selection, as found by (Volz et al., 2021) with >25,000 whole genome sequences from the United Kingdom.

On the other hand, according to the architecture of the spike protein (Xia et al., 2020), the S:T1117I is located between the HR1 and HR2 regions of the S2 domain. Preliminary predictions of the effects of this mutation on the function of the spike were suggested in the oligomerization needed for cell infection (Molina-Mora et al., 2021a). To gain more insights into the relevance of the mutation, molecular docking was performed using nelfinavir drug against the spike protein. For both, the WT and the mutated (S:D614 and S:T1117I) spike proteins, the drug was docked in the HR1 region of S2 domain, as reported before (Musarrat et al., 2020). Although the mutations are far from the RBD region, it is known that outside binding site mutations affect the ligand recognition (Sixto-López et al., 2021), as found here with changes in affinity energy. A higher affinity was found for the mutated spike in comparison to the WT version, with energy values that are in line with other studies including WT and mutations in the spike protein using nelfinavir (Musarrat et al., 2020; Sixto-López et al., 2021) or other drugs (Calligari et al., 2020; Hall and Ji, 2020). These results suggest that the mutated spike sequence defines a more stable complex with the ligand than the WT protein and it could be more affected by the inhibition mechanism of the drug, as previously reported for other mutations in the spike for this drug and other molecules (Sixto-López et al., 2021; Verkhivker and Di Paola, 2021).

Besides, we analyzed antigenic peptides from spike protein that could be affected by the mutation S:T1117I using immunoinformatics, similar to other approaches (Li et al., 2020; Lin et al., 2020). A total of 742 epitopes were predicted for the spike protein, but only one overlap the site of the mutation. The analysis of the effect of the mutation for that sequence revealed the best scores for the WT sequence, with some changes in the physicochemical properties. These results show that the S:T1117I changes the properties of the epitope to be recognized by the immune system. However, the peptide is located out of the RBD (S1 domain) and, more importantly, the peptide corresponds to only one peptide out of 742 candidates that are predicted to induce an immune response, suggesting that the vaccination effectiveness is not affected by this mutation.

Regarding the clinical relevance of the lineage B.1.1.389, our previous analyses with clinical data of COVID-19 cases from Costa Rica and a machine learning approach revealed that the seven clinical profiles, including an asymptomatic group, were not associated with a specific genome. The frequency of genomes carrying the S:T1117I was similar among the clinical profiles (Molina-Mora et al., 2021a). Other factors such as genetics and risk factors of the susceptible host contribute to the outcome of the disease.

Finally, although different epidemiological scenarios could explain the transition of SARS-CoV-2 genotypes in Costa Rica and around the world, an evolutionary explanation of the presence of distinct genotypes cannot be excluded. The fact that the lineage B.1.1.389 ended up disappearing at the beginning of the year 2021, could be a product of the Hill-Robertson effect (Hill and Robertson, 1966) when other advantageous genotypes arrived. This is, under the presence of two competing advantageous (from the point of view of the virus) genotypes, one (such as the lineage A.2.5 or the Alpha variant) completely took over and resulted dominant with a subsequent loss of the advantageous genotype, like B.1.1.389.

Altogether, the different effects of the mutation were found on the function of the spike protein, the interaction with molecules, and immunity, showing a complex scenario regarding the real value of this mutation in the spread of SARS-CoV-2 and the pandemic. Such a scenario and the fact that the lineage B.1.1.389 was predominant during the second semester of 2020 could suggest that the mutation represented a partial advantage in that period. Nevertheless, this benefit could be not strong enough with the introductions of other lineages such as the lineages A.2.5, A.2.5.1 and, A.2.5.2 (with high frequency in Central America) which were predominant during 2021, and later with the arrival of the alpha and gamma variants. In addition, our previous work demonstrated that the clinical manifestations of COVID-19 in Costa Rican cases were independent of the SARS-CoV-2 lineages, including no differences for the lineage B.1.1.389 and the mutation S:T1117I (Molina-Mora et al., 2021b).

COVID-19 disease, as an infection disease, depends not only on the viral agent (SARS-CoV-2 genotype), but also on the host conditions (genetics and risk factors) and the environment (social behavior) (Molina-Mora et al., 2021b; Tsui et al., 2020). Thus, this work represents another piece to understand the dynamics of the COVID-19 pandemic in Costa Rica, in this case with a focus on the virus. Further analyses are required to validate experimentally predictions and results, which is the main limitation of this *in silico* study. Although these analyses are reported in other studies, they appear in independent publications while here we provide them together for studying a particular mutation with an integrative approach. Thus, this *in silico* strategy can be eventually used to study mutations of interest for the SARS-CoV-2 virus and other pathogens.

5. Conclusions

In conclusion, this work analyzed the mutation T1117I in the spike protein of the SARS-CoV-2. Insights into the biological meaning of this change were gained, including the description of a polyphyletic pattern around the world, with the emergence of local lineages. S:T1117I is found in the lineage B.1.1.389 and its positive selection affected the spread of genomes carrying this mutation in Costa Rica. Different changes in the function of the spike protein, higher affinity in the interaction with molecules, and scarce changes in immunity were revealed for this mutation. This suggests a partial benefit for the spread of the virus with this genotype during the year 2020 in Costa Rica, although possibly not strong enough with the introduction of new lineages during early 2021 which became predominant later. In addition, the bioinformatic strategy followed here can be eventually used to analyze other mutations in SARS-CoV-2 virus or other pathogens.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.genrep.2022.101554.

Ethical approval and consent to participate

This study was approved by the scientific committee of CIET-UCR (No. 242–2020). Consent was not required for this study.

Availability of data and material

Data were retrieved from public databases, as described in Methods. See Supplementary Material for details of sequence ID, locations, lineages, and other data.

Funding

This work was funded by Vicerrectoría de Investigación – Universidad de Costa Rica, with the Project "C0196 Protocolo bioinformático y de inteligencia artificial para el apoyo de la vigilancia epidemiológica basada en laboratorio del virus SARS-CoV-2 mediante la identificación de patrones genómicos y clínico-demográficos en Costa Rica (2020-2022)".

CRediT authorship contribution statement

Jose Arturo Molina-Mora: Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Visualization, Data curation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author declares that there is no conflict of interest.

Acknowledgments

We thank all the researchers who sheared all sequences and other data into public databases, including the public laboratories from Costa Rica, as well as Meriyeins Sibaja-Amador and Carlos Martínez-Calderón for their assistance in different activities of the project.

References

Astuti, I., Ysrafil., 2020. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): an overview of viral structure and host response. Diabetes Metab. Syndr. Clin. Res. Rev. 14 (4), 407–412. https://doi.org/10.1016/j.dsx.2020.04.020.

Berrio, A., Gartner, V., Wray, G.A., 2020. Positive selection within the genomes of SARS-CoV-2 and other coronaviruses independent of impact on protein function. PeerJ 8, e10234. https://doi.org/10.7717/peerj.10234.

Calligari, P., Bobone, S., Ricci, G., Bocedi, A., 2020. Molecular investigation of SARS–CoV-2 proteins and their interactions with antiviral drugs. Viruses 12 (4), 445. https://doi.org/10.3390/V12040445, 2020, Vol. 12, Page 445.

Candido, S., Mishra, S., Crispim, M.A.E., Sales, F.C., Jesus, J.G.De, Andrade, P.S., Camilo, C.C., 2021. In: Genomics and Epidemiology of a Novel SARS-CoV-2 Lineage in Manaus, Brazil, pp. 1–45.

Castillo, A.E., Parra, B., Tapia, P., Acevedo, A., Lagos, J., Andrade, W., Fernández, J., 2020. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. J. Med. Virol. 92 (9), 1562–1566. https://doi.org/10.1002/jmv.25797.

Castillo, A.E., Parra, B., Tapia, P., Lagos, J., Arata, L., Acevedo, A., Fernández, J., 2020. Geographical distribution of genetic variants and lineages of SARS-CoV-2 in Chile. Front. Public Health 8, 562615. https://doi.org/10.3389/fpubh.2020.562615.

Colell, E.A., Iserte, J.A., Simonetti, F.L., Marino-Buslje, C., 2018. MISTIC2: comprehensive server to study coevolution in protein families. Nucleic Acids Res. 46 (W1), W323–W328. https://doi.org/10.1093/nar/gky419.

Dimitrov, I., Naneva, L., Doytchinova, I., Bangov, I., 2014. AllergenFP: allergenicity prediction by descriptor fingerprints. Bioinformatics 30 (6), 846–851. https://doi.org/10.1093/bioinformatics/btt619.

Ferrareze, P.A.G., Franceschi, V.B., Caldana, G.D., Zimerman, R.A., Thompson, C.E., Mayer, A.de M., 2021. E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. Infection, Genetics and Evolution 93, 104941. https://doi.org/10.1016/j.meegid.2021.104941.

Graham, M.S., Sudre, C.H., May, A., Antonelli, M., Murray, B., Varsavsky, T., Gunson, R. N., 2021. Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. The lancetPublic Health 6 (5), e335–e345. https://doi.org/10.1016/s2468-2667(21)00055-4.

Guedes, I.A., Costa, L.S.C., dos Santos, K.B., Karl, A.L.M., Rocha, G.K., Teixeira, I.M., Dardenne, L.E., 2021. Drug design and repurposing with DockThor-VS web server focusing on SARS-CoV-2 therapeutic targets and their non-synonym variants. Sci. Rep. 11 (1), 5543. https://doi.org/10.1038/s41598-021-84700-0.

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Raghava, G.P.S., 2013. In silico approach for predicting toxicity of peptides and proteins. PLoS ONE 8 (9). https://doi.org/10.1371/journal.pone.0073957.

Hall, D.C., Ji, H.F., 2020. A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. Travel Med. Infect. Dis. 35, 101646 https://doi.org/10.1016/J.TMAID.2020.101646.

Hill, W.G., Robertson, A., 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8 (3), 269–294. https://doi.org/10.1017/S0016672300010156.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14 (6), 587–589. https://doi.org/10.1038/nmeth.4285.

Katoh, K., Misawa, K., Kuma, K.I., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30 (14), 3059–3066. https://doi.org/10.1093/nar/gkf436.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21 (5), 676–679. https://doi.org/10.1093/bioinformatics/bti079.

Koyama, T., Platt, D., Parida, L., 2020. Variant analysis of SARS-cov-2 genomes. Bull. World Health Organ. 98 (7), 495–504. https://doi.org/10.2471/BLT.20.253591.

Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256–W259. https://doi.org/10.1093/nar/gkz239.

Li, W., Li, L., Sun, T., He, Y., Liu, G., Xiao, Z., Zhang, J., 2020. Spike protein-based epitopes predicted against SARS-CoV-2 through literature mining. In: Medicine in Novel Technology and Devices, 8. https://doi.org/10.1016/J.MEDNTD.2020.100048.

Lin, L., Ting, S., Yufei, H., Wendong, L., Yubo, F., Jing, Z., 2020. Epitope-based peptide vaccines predicted against novel coronavirus disease caused by SARS-CoV-2. Virus Res. 288, 198082 https://doi.org/10.1016/J.VIRUSRES.2020.198082.

Lythgoe, K.A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Golubchik, T., 2021. SARS-CoV-2 within-host diversity and transmission. Science 372 (6539). https://doi.org/10.1126/SCIENCE.ABG0821.

Mercatelli, D., Giorgi, F.M., 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. 11, 1800. https://doi.org/10.3389/fmicb.2020.01800.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Teeling, E., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37 (5), 1530–1534. https://doi.org/10.1093/molbev/msaa015.

Molina-Mora, J.A., Cordero-Laurent, E., Godínez, A., Calderón-Osorno, M., Brenes, H., Soto-Garita, C., Duarte-Martínez, F., 2021. SARS-CoV-2 genomic surveillance in Costa Rica: evidence of a divergent population and an increased detection of a spike T1117I mutation. Infect. Genet. Evol. 92, 104872 https://doi.org/10.1016/j.meegid.2021.104872.

Molina-Mora, J.A., González, A., Jiménez-Morgan, S., Cordero-Laurent, E., Brenes, H., Soto-Garita, C., Duarte-Martínez, F., Clinical profiles at the time of diagnosis of COVID-19 in Costa Rica during the pre-vaccination period using a machine learning approach, 2021. MedRxiv, 2021.06.18.21259157. https://doi.org/10.1101/2021.06.18.21259157.

Muñoz, M., Patiño, L.H., Ballesteros, N., Paniz-Mondolfi, A., Ramírez, J.D., 2021. Characterizing SARS-CoV-2 genome diversity circulating in south american countries: signatures of potentially emergent lineages? Int. J. Infect. Dis. 105, 329–332. https://doi.org/10.1016/j.ijid.2021.02.073.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K., 2013. FUBAR: a fast, unconstrained Bayesian AppRoximation for inferring selection. Mol. Biol. Evol. 30 (5), 1196–1205. https://doi.org/10.1093/molbev/mst030.

Musarrat, F., Chouljenko, V., Dahal, A., Nabi, R., Chouljenko, T., Jois, S.D., Kousoulas, K. G., 2020. The anti-HIV drug nelfinavir mesylate (Viracept) is a potent inhibitor of cell fusion caused by the SARSCoV-2 spike (S) glycoprotein warranting further evaluation as an antiviral against COVID-19 infections. J. Med. Virol. 92 (10), 2087–2095. https://doi.org/10.1002/jmv.25985.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF chimera - a visualization system for exploratory research and analysis. J. Comput. Chem. 25 (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6), 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Simonetti, F.L., Teppa, E., Chernomoretz, A., Nielsen, M., Marino Buslje, C., 2013. MISTIC: mutual information server to infer coevolution. Nucleic Acids Res. 41 (Web Serverissue), 8–14. https://doi.org/10.1093/nar/gkt427.

Sixto-López, Y., Correa-Basurto, J., Bello, M., Landeros-Rivera, B., Garzón-Tiznado, J.A., Montaño, S., 2021. Structural insights into SARS-CoV-2 spike protein and its natural mutants found in Mexican population. Sci. Rep. 11 (1), 4659. https://doi.org/10.1038/s41598-021-84053-8.

Taboada, B., Vazquez-Perez, J.A., Muñoz-Medina, J.E., Ramos-Cervantes, P., Escalera-Zamudio, M., Boukadida, C., Arias, C.F., 2020. Genomic analysis of early SARS-CoV-2 variants introduced in Mexico. J. Virol. 94 (18) https://doi.org/10.1128/jvi.01056-20.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., De Oliveira, T., 2020. In: Alisoltani-Dehkordi, Arghavan (Ed.), Emergence and Rapid Spread of a New Severe Acute Respiratory Syndrome-related Coronavirus 2 (SARS-CoV-2) Lineage With Multiple Spike Mutations in South Africa, 10. https://doi.org/10.1101/2020.12.21.20248640.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. 65 (12), 1075–1082. https://doi.org/10.1038/s10038-020-0808-9.

Tsui, B.C.H., Deng, A., Pan, S., 2020. COVID-19: Epidemiological factors during aerosol-generating medical procedures. September 1 Anesthesia and Analgesia. https://doi.org/10.1213/ANE.0000000000005063. Lippincott Williams and Wilkins.

Verkhivker, G., 2020. Coevolution, dynamics and allostery conspire in shaping cooperative binding and signal transmission of the SARS-CoV-2 spike protein with human angiotensin-converting enzyme 2. Int. J. Mol. Sci. 21 (21), 1–31. https://doi.org/10.3390/ijms21218268.

Verkhivker, G.M., Di Paola, L., 2021. Integrated biophysical modeling of the SARS-CoV-2 spike protein binding and allosteric interactions with antibodies. J. Phys. Chem. B 125 (18), 4596–4619. https://doi.org/10.1021/acs.jpcb.1c00395.

Volz, E., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O'Toole, Á., Connor, T.R., 2021. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 184 (1), 64. https://doi.org/10.1016/J.CELL.2020.11.020.

Xia, S., Zhu, Y., Liu, M., Lan, Q., Xu, W., Wu, Y., Lu, L., 2020. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. July 1 Cellular and Molecular Immunology. https://doi.org/10.1038/s41423-020-0374-2. Springer Nature.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7.