

# Genetic Variation and Evolution of the 2019 Novel Coronavirus

Salvatore Dimonte<sup>a</sup> Muhammed Babakir-Mina<sup>b</sup> Taib Hama-Soor<sup>b</sup> Salar Ali<sup>b</sup>

<sup>a</sup>BioMolecular Lab, Barletta, Italy; <sup>b</sup>Technical College of Health, Sulaimani Polytechnic University, KGR, Sulaimani, Iraq

## Keywords

Biodiversity · Ecology · Mutation rate · Pandemics · Prevalence · Clinical genetics

## Abstract

**Introduction:** SARS-CoV-2 is a new type of coronavirus causing a pandemic severe acute respiratory syndrome (SARS-2). Coronaviruses are very diverting genetically and mutate so often periodically. The natural selection of viral mutations may cause host infection selectivity and infectivity. **Methods:** This study was aimed to indicate the diversity between human and animal coronaviruses through finding the rate of mutation in each of the spike, nucleocapsid, envelope, and membrane proteins. **Results:** The mutation rate is abundant in all 4 structural proteins. The most number of statistically significant amino acid mutations were found in spike receptor-binding domain (RBD) which may be because it is responsible for a corresponding receptor binding in a broad range of hosts and host selectivity to infect. Among 17 previously known amino acids which are important for binding of spike to angiotensin-converting enzyme 2 (ACE2) receptor, all of them are conservative among human coronaviruses, but only 3 of them significantly are mutated in animal coronaviruses. A single amino acid aspartate-454, that causes dissociation of the RBD of the spike and ACE2, and F486 which gives the strength of binding with ACE2 remain intact in all

coronaviruses. **Discussion/Conclusion:** Observations of this study provided evidence of the genetic diversity and rapid evolution of SARS-CoV-2 as well as other human and animal coronaviruses.

© 2021 S. Karger AG, Basel

## Introduction

Coronavirus disease (COVID-19) is the severe acute respiratory disease in humans that is caused by the emergence of a new type of a coronavirus. The outbreak of the disease was first identified in Wuhan, China, in December 2019. Wuhan city became an epicenter of the disease, and then the outbreak spread outside of China and caused the emergence of the disease in 213 countries and territories. The total number of infected cases (25,658,983) and death (855,186) has been recorded globally until September 1, 2020. The World Health Organization called it as a public health emergency of worldwide concern (<https://www.worldometers.info/coronavirus/>). The estimated mean time of incubation period of SARS-CoV-2 is 5.1 days. In general, all 3 emerged coronaviruses have relatively similar mean time of incubation period around 5 days. More precisely it is 5.1 in SARS-CoV-2 (ranged from 4.5 to 5.8 days) [1], 5 days in SARS (ranged from 2 to 14 days) [2], and 7 days in MERS (ranged from 2 to 14 days) [3].

The coronavirus family is an enveloped virus containing a single-strand RNA, and the diameter of the virus is about 80–120 nm. Coronaviruses are divided on 4 types:  $\alpha$ -coronavirus,  $\beta$ -coronavirus,  $\delta$ -coronavirus, and  $\gamma$ -coronavirus. Some common human coronaviruses, HKU1, NL63, and OC43 and 229E, infect humans and only cause mild respiratory disease, while MERS-CoV and SARS-CoV-1 cause devastating severe acute respiratory infection [4]. All devastating coronaviruses, 2019-nCoV (SARS-CoV-2) with MERS-CoV and SARS-CoV-1, belong to the same group of coronaviruses,  $\beta$ -coronavirus [5]. The genome sequence homology between SARS-CoV-2 and other viruses are noted: it is 79% homologous with SARS-CoV-1, and it is more homologous to BatCoV RaTG13 [6–9] which belongs to SARS-like bat coronaviruses. There are no recombination evidences found between them [10].

Both SARS-CoV-2 and SARS-CoV-1 exploit the same receptor for the binding process, angiotensin-converting enzyme 2 (ACE2) receptor, and this means that both viruses may share structure similarity and a way of attaching to the cell receptor [11]. SARS-CoV-2 recognizes and binds with ACE2 receptor through its spike protein to initiate infection. It has been found via structural model analysis that the new coronavirus SARS-CoV-2's affinity of binding to ACE2 is 10 times stronger than the previously known coronavirus, SARS-CoV-1. The surface spike glycoprotein helps viral entry into host cells via homotrimers projection from the viral surface. The spike protein has 2 main domains: S1 which is responsible to bind the host cell receptor and S2 is responsible for the fusion of both viral and host cellular membranes [12].

The most genomic variation part in SARS-CoV-1 and SARS-CoV-2 is the receptor-binding domain (RBD) in the spike protein [13, 14], and some locations in this protein sequence might be related to positive selection [15]. Due to the abundant variability in SARS-CoV-2 isolates, many questions require an answer to understand whether these mutations have a role in the pathogenicity of SARS-CoV-2. This is significant to understand the viral infection mechanisms and pave a way to find drug and vaccine to protect people from the next stage of the pandemic.

## Materials and Methods

NC\_045512.2 COVID-19/Wuhan-Hu-1CHN/2019/First Isolate was used as a reference strain for the definition of mutations. Multiple sequence alignments of sequences were performed using T-Coffee program (<https://www.ebi.ac.uk/Tools/msa/tcoffee/>)

and were manually edited with the BioEdit software. The T-coffee program was used to mitigate the pitfalls of progressive alignment methods and because it is suitable for small alignments. In the case of amino acid insertion or deletion, these events were not considered.

All amino acid viral sequences of the spike, nucleocapsid, membrane, and envelope proteins were screened in both animal and human viruses, and the frequency of mutations was calculated and statistically compared using the  $\chi^2$  test (based on a  $2 \times 2$  contingency table containing the number of isolates from animal and human viruses and the number of isolates with and without mutations) [16, 17]. Fisher exact tests were used to determine whether the differences in frequency between the 2 groups of samples (animal vs. human) were statistically significant.

The Benjamini-Hochberg method was used to identify results that were statistically significant in multiple hypothesis testing. A strong false discovery rate of 0.001 was used to determine statistical significance [18].

## Results

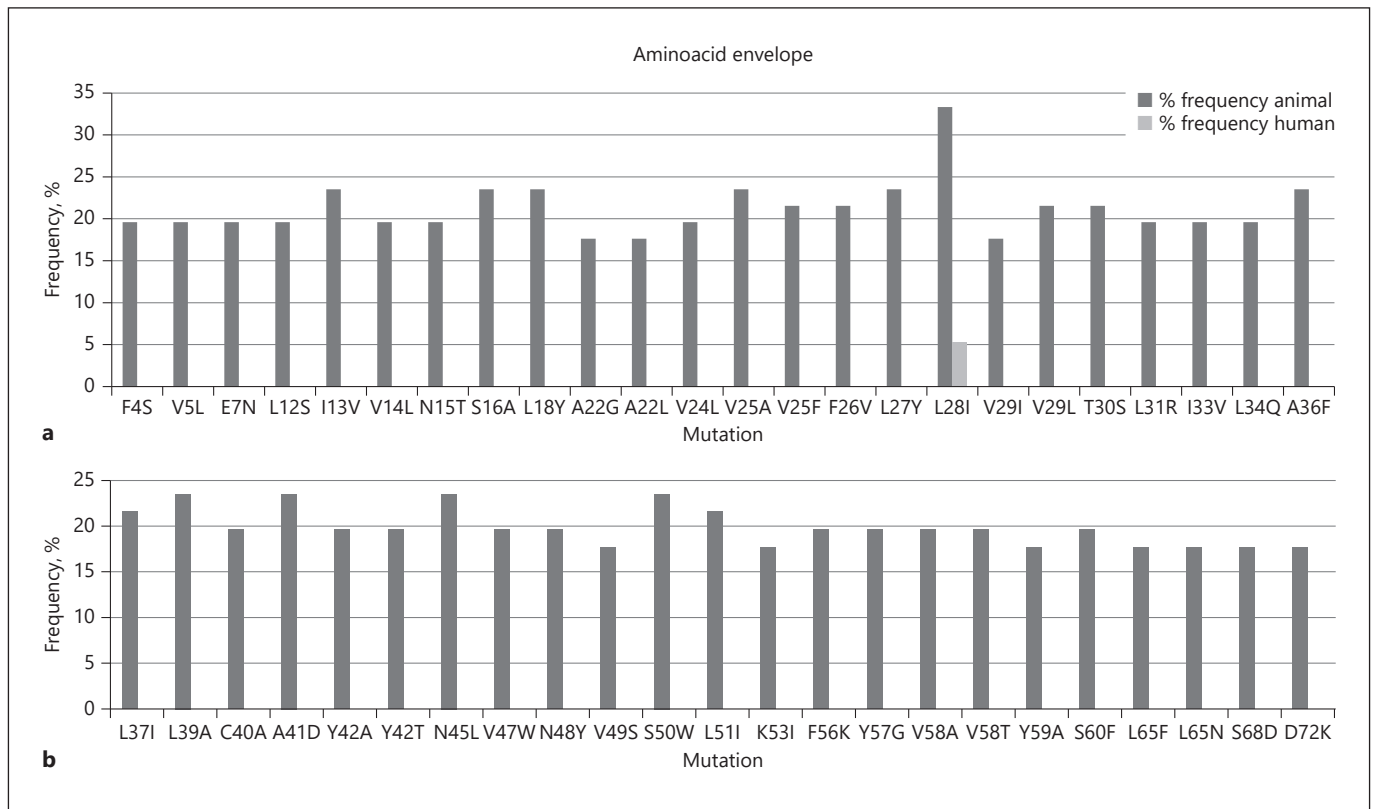
The pandemic COVID-19 is the third known outbreak to cause respiratory illness in the current century which has been caused by zoonotic coronaviruses, and it is thought to have jumped from bats or/and pangolin hosts to humans [14, 15, 19–21]. The disease spread mostly everywhere globally very rapidly mainly due to traveling. Several mutations have been observed in the virus genome during the rapid dissemination of the disease across different continents. Antigenic drift may occur as a result of accumulation of several mutations in the virus during seasonal outbreaks which helps the virus to survive due to natural selection process like it is common among influenza viruses [14]. Tracking mutations in the genome of the virus is necessary especially in spike protein which directly has a role in binding and infecting the cell. Vaccine development is mostly dependent on the spike protein, and this is aimed to find antibody against spike or any other surface proteins to neutralize viruses [22, 23].

Therefore, in this study, the occurrence of mutations in different animal and human coronaviruses and finding genetic diversity among human coronaviruses and/with animal coronaviruses were focused. For this purpose, over 50 protein sequences of each of the spike (S), nucleocapsid (N), envelope (E), and membrane (M) proteins from the online database GenBank was collected. The sum of SARS viruses' amino acid changes is listed in Table 1. The sequences were processed for each of them separately in both human and animal viruses as follows.

**Table 1.** Sum of SARS viruses' amino acid changes

ENV	Spike		Membrane				Nucleocapsid					
	<i>nAb</i>	<i>RDB</i>										
F4S	E340S	L387I	A2S	A68D	R131F	Y178V	S2A	I84H	Q163D	Q242S	F307L	Y360V
V5L	E340G	N388A	D3N	V70S	R131I	K180R	D3S	Y87W	G164Q	G243S	A308V	K361G
E7N	F342S	D389G	T7S	Y71F	P132A	L181R	N4G	R88N	T165G	Q244S	S310N	F363R
L12S	A344G	D389S	T9D	R72Q	L133I	G182Q	N4V	R89E	L167I	T245R	A311S	T366D
I13V	R346L	D389T	T9P	I73P	L134E	A183R	G5K	A90Q	K169D	V246T	A311V	E367Q
V14L	F347I	L390G	L13V	I73V	E135G	S184Y	G5N	R93W	F171N	T247P	S312A	P368E
N15T	F347Y	F392K	K14V	W75T	E135S	Q185L	N8D	G96M	F171V	S250K	S312H	K370R
S16A	A348P	F392T	K15E	I76G	S136A	R186G	N8G	G96P	Y172F	A252R	F314C	K373S
L18Y	A348Q	N394D	L17F	T77F	S136V	R186N	Q9K	D98K	E174P	E253K	F314L	K374R
A22G	S349T	N394S	E18R	T77G	E137P	V187A	R10T	K100G	E174V	A254M	F315L	A376S
A22L	A352G	Y396F	Q19N	G78F	L138M	V187D	N11P	K100Q	S176N	S255A	G316F	D377E
V24L	W353F	A397L	W20Y	G79A	V139G	A188Q	F17L	M101R	G179R	K256D	M317G	D377S
V25A	W353Y	V401R	L22F	G79V	G141S	G189S	S21K	L104Q	Q181G	K256H	S318G	E378R
V25F	N354K	D405T	V23F	I80A	A142L	D190G	R41P	L104V	Q181R	K257R	I320V	A381P
F26V	N354Y	T415A	G25N	I80F	A142P	S191K	Q43K	S105P	A182R	P258R	G321T	L382A
L27Y	K356S	G416T	G25T	A83L	V143I	S191T	Q43R	P106D	S183N	R259E	M322P	R385Q
L28I	I358H	K417C	F26A	M84L	V143T	G192R	G44V	P106S	S184T	R259Y	M322T	Q390K
V29I	S359P	I418C	F26I	C86V	I144L	F193W	L45G	R107A	R185A	Q260C	E323K	T391E
V29L	V362P	D467N	L27F	L87F	L145K	A195F	L45R	R107N	S187R	Q260W	E323R	T393D
T30S	A363F	T470Y	F28L	L87I	R146S	A195V	P46K	Y109H	N192A	A264P	V324E	L394V
L31R	D364R	Y473L	L29T	V88T	H148T	S197A	P46S	E118A	N192S	T265P	V324L	L395T
I33V	Y365F	T500C	T30F	G89C	R150L	S197V	N47R	E118H	S193P	K266P	P326G	P396D
L34Q	Y365L	T500G	T30I	G89L	R150Y	R198K	N48G	G120D	S197G	A267E	S327D	A397N
A36F	S366E	G504Y	W31F	L90V	I151C	Y199Q	T49N	L121A	S197R	A267G	T329L	A397Q
L37I	V367T	Y505F	W31L	M91L	I151F	Y199S	T49P	P122N	T198S	N269S	W330E	A398L
L39A	L368I	Q506G	I32L	M91S	I151V	R200K	A50L	P122R	P199R	T271D	W330H	D399E
C40A	L368V	P507E	C33V	W92F	A152E	R200S	W52L	Y123W	G200S	Q272A	L331I	L400F
A41D	Y369N	P507S	L34V	S94G	H154F	I201H	T54A	A125D	S201N	A273C	Y333F	F403E
Y42A	N370S	Y508L	F37Y	S94M	H154Q	I201V	T54Q	N126R	A208G	A273V	G335F	S404P
Y42T	S371G	R509Y	A38G	F96W	H155K	G202D	A55P	K127Q	A211S	R276P	G335Y	Q406V
N45L	A372L	V512D	Y39H	A98N	H155W	N203D	T57K	I130V	G214S	G278T	A336K	L407I
V47W	S373W	V512L	A40Y	F103Y	G157A	N203T	T57R	I131F	G215E	P279K	A336T	Q408N
N48Y	S375N	L513V	N41K	A104K	R158K	Y204G	Q58A	T135A	A217D	T282K	I337M	Q409W
V49S	T376K	S514Y	N41T	A104R	R158T	K205E	Q58V	E136K	A218I	Q283E	I337T	S410G
S50W	K378F	F515V	R42Y	T106C	C159V	K205S	H59K	L139D	A218L	G284K	K338V	S413A
L51I	C379S	E516Q	N43S	M109W	D160E	L206A	H59T	L139K	L219I	Q289A	L339V	A414L
K53I	Y380V	E516T	R44A	S111A	D160Q	N207E	G60D	N140T	L219V	Q289D	D340P	A414V
F56K	G381D		R44K	T116S	I161P	N207V	G60N	T141K	L221A	E290K	D341K	D415G
Y57G	G381S		L46I	N117D	I161V	D209N	E62P	K143R	L222A	L291M	K342D	S416E
V58A	V382I		I48G	I118A	D163H	D209V	D63F	D144S	L223A	I292V	K342S	Q418E
V58T	V382L		I49L	L119I	D163Q	S212T	L64P	H145N	L223V	R293E	K347D	A419L
Y59A	P384Y		I49V	L119V	E167D	S213D	L64W	I146L	L224K	Q294E	K347E	A419N
S60F	T385G		L51M	L120G	E167F	S213E	F66V	I146Q	D225K	T296I	D348N	
L65F	T385L		I52L	N121S	I168V	S214G	Q70N	T148V	D225Q	T296V	Q349F	
L65N	T385S		F53I	N121T	T169F	D215E	Q70S	A152D	R226A	D297K	Q349Y	
S68D	K386P		L56C	V122I	L16H	D215S	G71A	N153K	R226I	Y298A	I351K	
D72K			L57F	P123L	A171C	N216K	I74T	N154F	N228K	K299G	L352I	
			V60L	P123S	S173A	N216S	N75G	A155D	E231G	W301F	L353C	
			T61N	L124V	S173P	I217V	T76E	A155P	S232I	W301V	N354D	
			T61V	H125T	R174D	L220H	N77G	A156Q	K233Q	W301Y	K355E	
			L62I	T127Q	T175R	V221L	S79K	I157Y	M234Q	P302T	K355S	
			C64V	T127R	L176R	V221Y	S79N	V158P	K237Q	Q303A	H356C	
			F65G	I128Q	L176T	Q222T	P80K	Q160R	G238K	I304M	H356Q	
			L67F	L129C	S177N	Q222V	D81S	L161F	Q239S	A305L	I357V	
			L67I	T130N	Y178I		D82Q	P162S	Q241K	Q306N	A359G	

Sum of the mutations found and observed in a statistically significant manner ( $p \leq 0.001$ ) for each viral protein analyzed in the text. In the table were listed the changes, related to envelope (ENV), spike (nAb and RDB epitopes crucial and enough to bind ACE2 receptor), membrane, and nucleocapsid viral proteins. The ENV protein is responsible for protection of the interior parts of the virus and has a role in viral assembly during viral replication. The spike protein helps in viral attachment to its corresponding receptor and mediates fusion of the cell and viral membrane. The surface membrane is the most abundant structural protein and defines the shape of the viral envelope. The nucleocapsid is the pretentious structure inside the box of the coronaviruses: N structural protein and CoV RNA genome make up the nucleocapsid.



**Fig. 1.** Frequencies of SARS viruses' amino acid changes in envelope protein. Frequencies of envelope signatures in animal viral isolates (dark gray) and human viral isolates (light gray). The analysis was performed in sequences derived from 106 subjects; 51 re-

ported animal viral strains, and 55 reported human viral strains. Statistically significant differences were assessed by  $\chi^2$  tests of independence. All  $p$  values were calculated from 2-sided tests using 0.001 as the significance level ( $p \leq 0.001$ ).

### Envelope

The entire envelope protein sequences, derived from 51 animal viruses and 55 human viruses, were analyzed (Fig. 1). In Figure 1, frequencies of SARS-CoV-2 envelope amino acid signatures were shown, using the isolate NC\_045512.2 COVID-19/Wuhan as a reference.

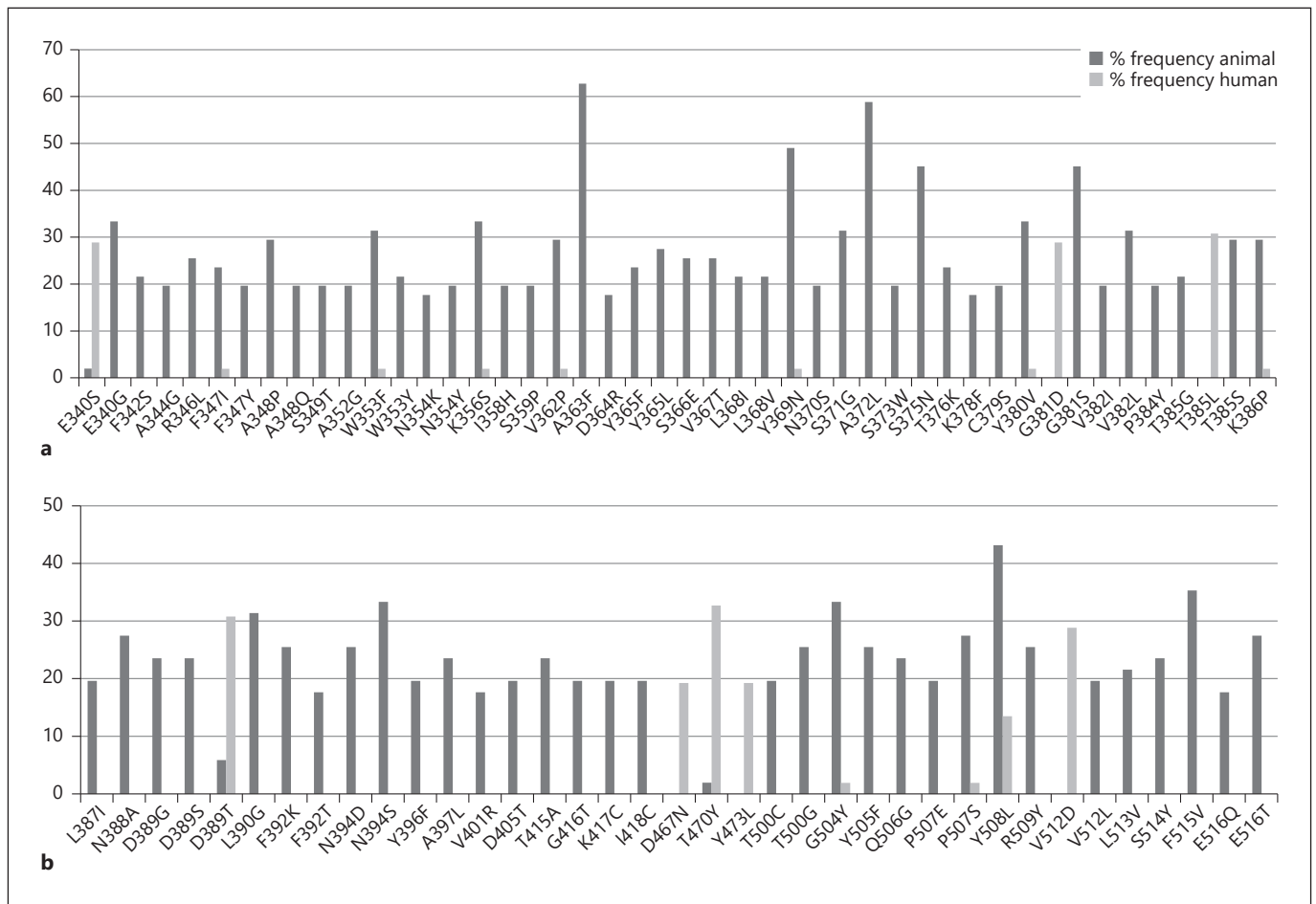
In envelope virus sequences, among the 76 residues, 365 mutations were observed; 47, in a statistically significant manner ( $p \leq 0.001$ ). In animal viruses' sequences, 333 mutations were observed; 47 of them (F4S, V5L, E7N, L12S, I13V, V14L, N15T, S16A, L18Y, A22G, A22L, V24L, V25A, V25F, F26V, L27Y, L28I, V29I, V29L, T30S, L31R, I33V, L34Q, A36F, MUT, L37I, L39A, C40A, A41D, Y42A, Y42T, N45L, V47W, N48Y, V49S, S50W, L51I, K53I, F56K, Y57G, V58A, V58T, Y59A, S60F, L65F, L65N, S68D, and D72K) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 1a, b). In envelope human virus sequences, 94 mutations were observed and just 1, L28I, was observed in a statistically significance manner ( $p \leq 0.001$ ) (3 isolates; 5.4%) (Fig. 1a).

### Spike

The entire spike protein sequences, derived from 51 animal viruses and 52 human viruses, were analyzed. This protein has several domains (1,273 amino acid residues): N-terminal domain, RBD, receptor-binding motif, subdomain-1, subdomain-2, fusion peptide, heptad repeat 1, heptad repeat 2, transmembrane region, and intracellular domain. The SARS-CoV-2 spike RBD bound to the cell receptor ACE2, and here this crucial region (domain located between amino acid residues 336–516) was analyzed [24].

In Figure 2, frequencies of SARS-CoV-2 spike RBD amino acid signatures were shown, using the isolate NC\_045512.2 COVID-19/Wuhan as a reference.

In spike viruses' sequences, among the 1,273 residues, 9,782 mutations were observed; 982, in a statistically significant manner ( $p \leq 0.001$ ). In animal sequences, 920 mutations were observed, and specifically in the RBD region, 83 of them (E340S, E340G, F342S, A344G, R346L, F347I, F347Y, A348P, A348Q, S349T, A352G, W353F,



**Fig. 2. a, b** Frequencies of SARS viruses' amino acid changes in neutralizing antibody (m396 and 80R) epitope (336–516 amino acid domain) spike protein. **b** The SARS-CoV-2 RBD/ACE2 interface corresponds to 387–516 amino acid domain. Frequencies of spike signatures in animal viral isolates (dark gray) and human viral iso-

lates (light gray). The analysis was performed in sequences derived from 103 subjects; 51 reported animal viral strains, and 52 reported human viral strains. Statistically significant differences were assessed by  $\chi^2$  tests of independence. All  $p$  values were calculated from 2-sided tests using 0.001 as the significance level ( $p \leq 0.001$ ).

W353Y, N354K, N354Y, K356S, I358H, S359P, V362P, A363F, D364R, Y365F, Y365L, S366E, V367T, L368I, L368V, Y369N, N370S, S371G, A372L, S373W, S375N, T376K, K378F, C379S, Y380V, G381D, G381S, V382L, V382I, P384Y, T385G, T385L, T385S, K386P, L387I, N388A, D389G, D389S, D389T, L390G, F392K, F392T, N394D, N394S, Y396F, A397L, V401R, D405T, T415A, G416T, K417C, I418C, D467N, T470Y, Y473L, T500C, T500G, G504Y, Y505F, Q506G, P507E, P507S, Y508L, R509Y, V512D, V512L, L513V, S514Y, F515V, E516Q, and E516T) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 2a, b).

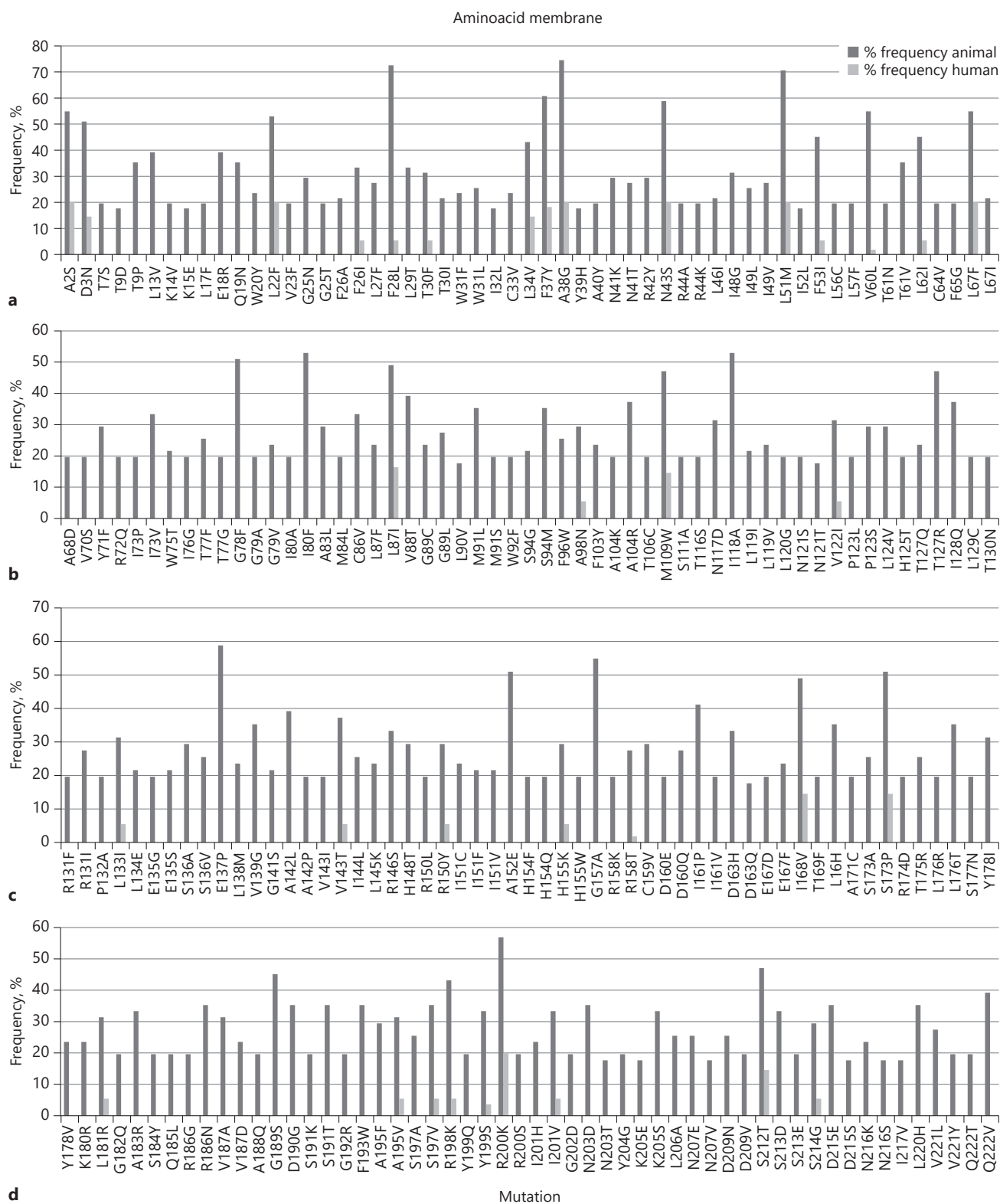
In spike human virus sequences, 222 mutations were observed, and specifically in the RBD region, 18 of them (E340S, F347I, W353F, K356S, V362P, Y369N, Y380V,

G381D, T385L, K386P, D389T, D467N, T470Y, Y473L, G504Y, P507S, Y508L, and V512D) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 2a, b).

To best study this domain, the geographical localization of sequences was analyzed. Homogeneity continental sources were observed: just the D405T animal virus mutation (10 Chinese isolates) shows a specific country (and continent, Asia).

### Membrane

The entire membrane protein sequences, derived from 51 animal viruses and 55 human viruses, were analyzed (Fig. 3). In Figure 3, frequencies of SARS-CoV-2 membrane amino acid signatures were shown, using the isolate NC\_045512.2 COVID-19/Wuhan as a reference.



(For legend see next page.)



In membrane viruses sequences, among the 222 residues, 851 mutations were observed; 219, in a statistically significance manner ( $p \leq 0.001$ ). In animal membrane sequences, 794 mutations were observed, and specifically 219 of them (A2S, D3N, T7S, T9D, T9P, L13V, K14V, K15E, L17F, E18R, Q19N, W20Y, L22F, V23F, G25N, G25T, F26A, F26I, L27F, F28L, L29T, T30F, T30I, W31F, W31L, I32L, C33V, L34V, F37Y, A38G, Y39H, A40Y, N41K, N41T, R42Y, N43S, R44A, R44K, L46I, I48G, I49L, I49V, L51M, I52L, F53I, L56C, L57F, V60L, T61N, T61V, L62I, C64V, F65G, L67F, L67I, A68D, V70S, Y71F, R72Q, I73P, I73V, W75T, I76G, T77F, T77G, G78F, G79A, G79V, I80A, I80F, A83L, M84L, C86V, L87F, L87I, V88T, G89C, G89L, L90V, M91L, M91S, W92F, S94G, S94M, F96W, A98N, F103Y, A104K, A104R, T106C, M109W, S111A, T116S, N117D, I118A, L119I, L119V, L120G, N121S, N121T, V122I, P123L, P123S, L124V, H125T, T127Q, T127R, I128Q, L129C, T130N, R131F, R131I, P132A, L133I, L134E, E135G, E135S, S136A, S136V, E137P, L138M, V139G, G141S, A142L, A142P, V143I, V143T, I144L, L145K, R146S, H148T, R150L, R150Y, I151C, I151F, I151V, A152E, H154F, H154Q, H155K, H155W, G157A, R158K, R158T, C159V, D160E, D160Q, I161P, I161V, D163H, D163Q, E167D, E167F, I168V, T169F, L16H, A171C, S173A, S173P, R174D, T175R, L176R, L176T, S177N, Y178I, Y178V, K180R, L181R, G182Q, A183R, S184Y, Q185L, R186G, R186N, V187A, V187D, A188Q, G189S, D190G, S191K, S191T, G192R, F193W, A195F, A195V, S197A, S197V, R198K, Y199Q, Y199S, R200K, R200S, I201H, I201V, G202D, N203D, N203T, Y204G, K205E, K205S, L206A, N207E, N207V, D209N, D209V, S212T, S213D, S213E, S214G, D215E, D215S, N216K, N216S, I217V, L220H, V221L, V221Y, Q222T, and Q222V) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 3a–d).

In human membrane virus sequences, 262 mutations were observed and specifically 35 of them (A2S, D3N, L22F, F26I, F28L, T30F, L34V, F37Y, A38G, N43S, L51M, F53I, V60L, L62I, L67F, L87I, A98N, M109W, V122I, L133I, V143T, R150Y, H155K, R158T, I168V, S173P,

L181R, A195V, S197V, R198K, Y199S, R200K, I201V, S212T, and S214G) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 3a–d).

### Nucleocapsid

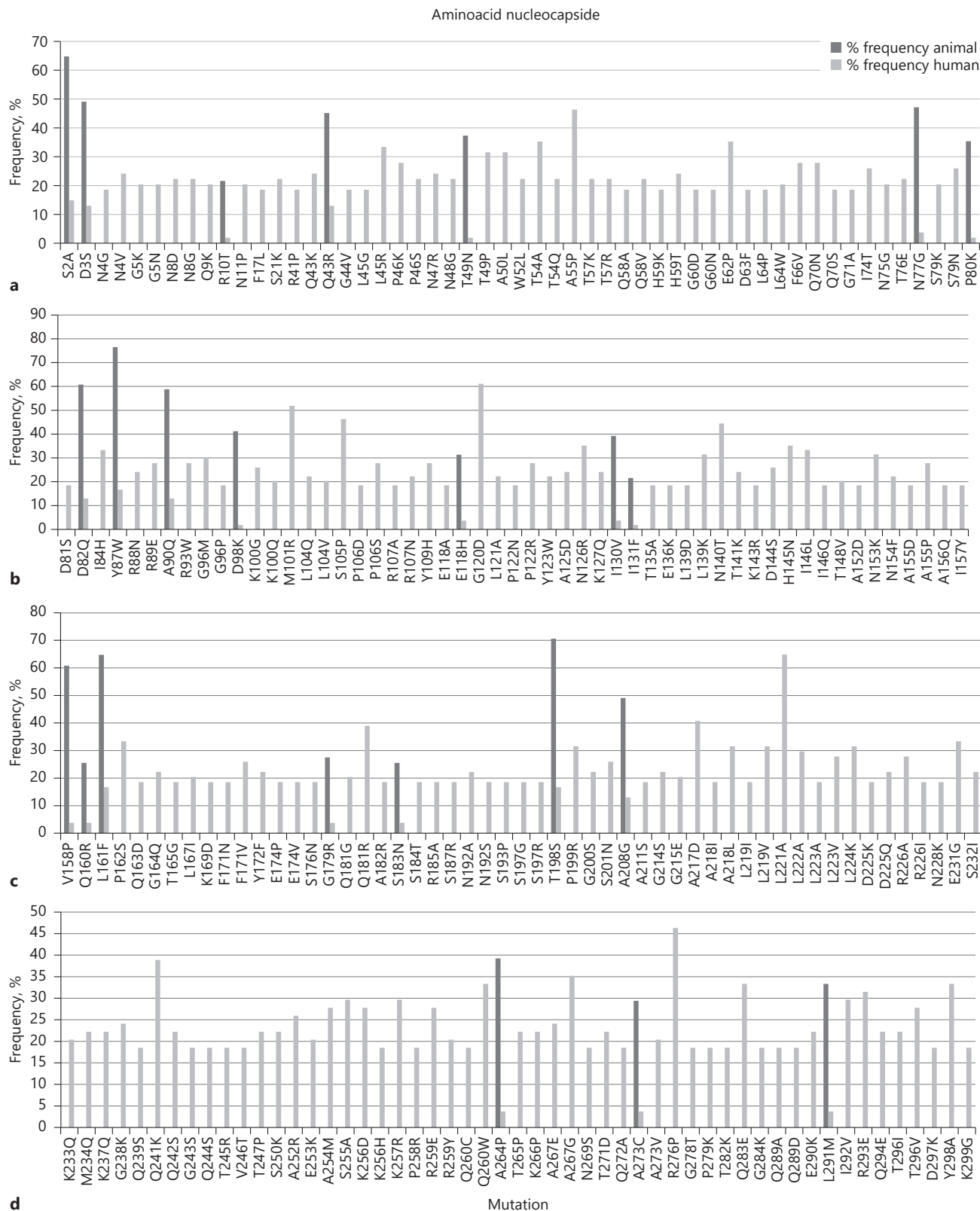
The entire nucleocapsid protein sequences, derived from 51 animal viruses and 54 human viruses, were analyzed (Fig. 4). In Figure 4, frequencies of SARS-CoV-2 nucleocapsid amino acid signatures were shown, using the isolate *NC\_045512.2 COVID-19/Wuhan* as a reference.

In nucleocapsid virus sequences, among the 419 residues, 1,885 mutations were observed; 317, in a statistically significance manner ( $p \leq 0.001$ ).

In animal nucleocapsid sequences, 613 mutations were observed and specifically 38 of them (D82Q, Y87W, A90Q, D98K, E118H, I130V, I131F, V158P, Q160R, L161F, G179R, S183N, T198S, A208G, A264P, A273C, L291M, W301F, F307L, A308V, G316F, M317G, E323R, S327D, W330E, K347E, I351K, K355E, K370R, A381P, and Q418E) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 4a–f).

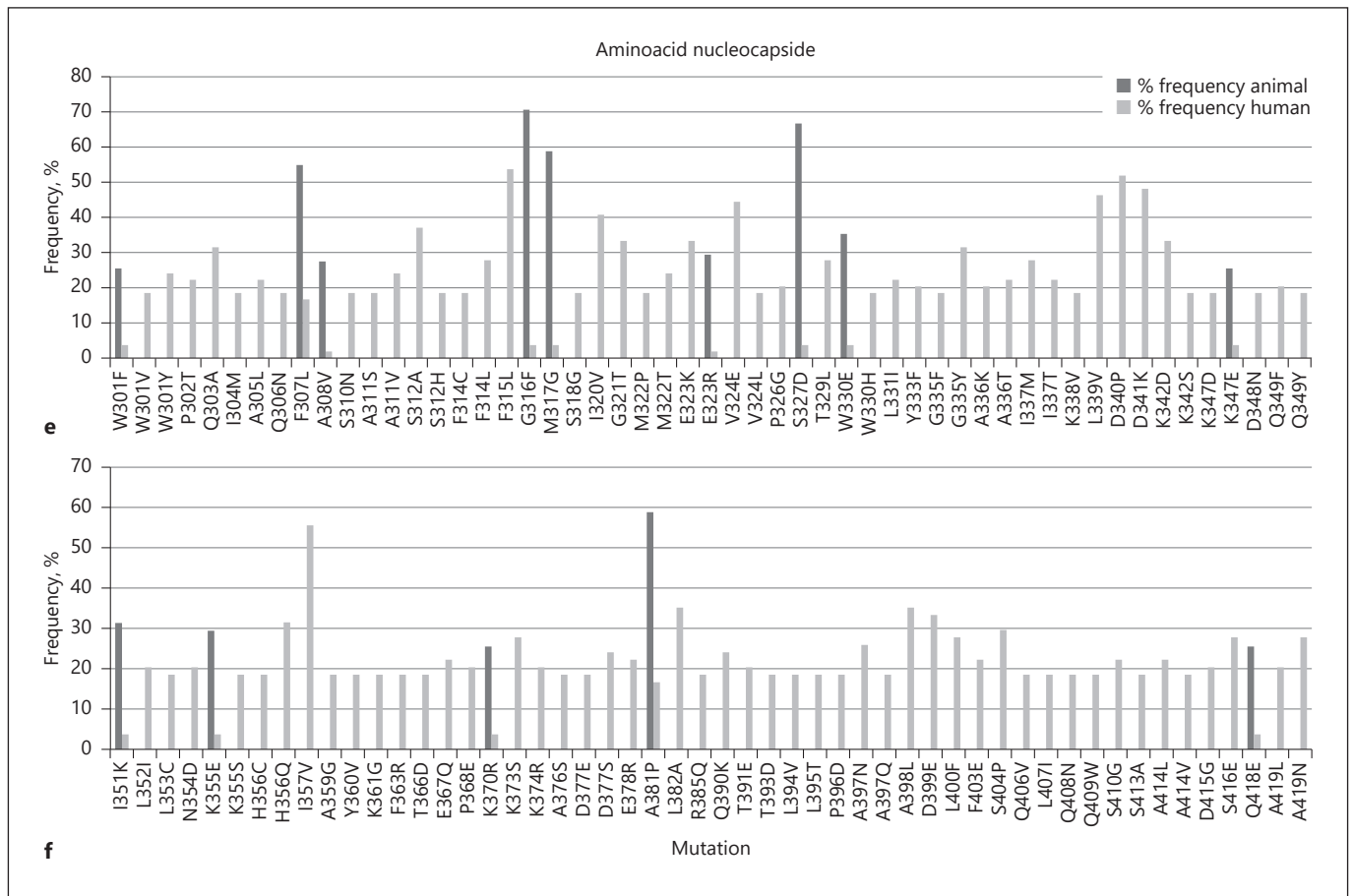
In human nucleocapsid sequences, 1,628 mutations were observed and specifically 317 of them (S2A, D3S, N4G, N4V, G5K, G5N, N8D, N8G, Q9K, R10T, N11P, F17L, S21K, R41P, Q43K, Q43R, G44V, L45G, L45R, P46K, P46S, N47R, N48G, T49N, T49P, A50L, W52L, T54A, T54Q, A55P, T57K, T57R, Q58A, Q58V, H59K, H59T, G60D, G60N, E62P, D63F, L64P, L64W, F66V, Q70N, Q70S, G71A, I74T, N75G, T76E, N77G, S79K, S79N, P80K, D81S, D82Q, I84H, Y87W, R88N, R89E, A90Q, R93W, G96M, G96P, D98K, K100G, K100Q, M101R, L104Q, L104V, S105P, P106D, P106S, R107A, R107N, Y109H, E118A, E118H, G120D, L121A, P122N, P122R, Y123W, A125D, N126R, K127Q, I130V, I131F, T135A, E136K, L139D, L139K, N140T, T141K, K143R, D144S, H145N, I146L, I146Q, T148V, A152D, N153K, N154F, A155D, A155P, A156Q, I157Y, V158P, Q160R, L161F, P162S, Q163D, G164Q, T165G, L167I, K169D, F171N, F171V, Y172F, E174P, E174V, S176N, G179R, Q181G, Q181R, A182R, S183N, S184T, R185A, S187R, N192A, N192S, S193P, S197G, S197R, T198S, P199R, G200S, S201N, A208G, A211S, G214S, G215E, A217D, A218I, A218L, L219I, L219V, L221A, L222A, L223A, L223V, L224K, D225K, D225Q, R226A, R226I, N228K, E231G, S232I, K233Q, M234Q, K237Q, G238K, Q239S, Q241K, Q242S, G243S, Q244S, T245R, V246T, T247P, S250K, A252R, E253K, A254M, S255A, K256D, K256H, K257R, P258R, R259E, R259Y, Q260C, Q260W, A264P, T265P, K266P, A267E, A267G, N269S, T271D, Q272A,

**Fig. 3.** Frequencies of SARS viruses' amino acid changes in membrane protein. Frequencies of membrane signatures in animal viral isolates (dark gray) and human viral isolates (light gray). The analysis was performed in sequences derived from 106 subjects; 51 reported animal viral strains, and 55 reported human viral strains. Statistically significant differences were assessed by  $\chi^2$  tests of independence. All  $p$  values were calculated from 2-sided tests using 0.001 as the significance level ( $p \leq 0.001$ ).



(Figure continued on next page.)





**Fig. 4.** Frequencies of SARS viruses' amino acid changes in nucleocapsid protein. Frequencies of nucleocapsid signatures in animal viral isolates (dark gray) and human viral isolates (light gray). The analysis was performed in sequences derived from 105 subjects; 51

reported animal viral strains, and 54 reported human viral strains. Statistically significant differences were assessed by  $\chi^2$  tests of independence. All  $p$  values were calculated from 2-sided tests using 0.001 as the significance level ( $p \leq 0.001$ ).

A273C, A273V, R276P, G278T, P279K, T282K, Q283E, G284K, Q289A, Q289D, E290K, L291M, I292V, R293E, Q294E, T296I, T296V, D297K, Y298A, K299G, W301F, W301V, W301Y, P302T, Q303A, I304M, A305L, Q306N, F307L, A308V, S310N, A311S, A311V, S312A, S312H, F314C, F314L, F315L, G316F, M317G, S318G, I320V, G321T, M322P, M322T, E323K, E323R, V324E, V324L, P326G, S327D, T329L, W330E, W330H, L331I, Y333F, G335F, G335Y, A336K, A336T, I337M, I337T, K338V, L339V, D340P, D341K, K342D, K342S, K347D, K347E, D348N, Q349F, Q349Y, I351K, L352I, L353C, N354D, K355E, K355S, H356C, H356Q, I357V, A359G, Y360V, K361G, F363R, T366D, E367Q, P368E, K370R, K373S, K374R, A376S, D377E, D377S, E378R, A381P, L382A, R385Q, Q390K, T391E, T393D, L394V, L395T, P396D, A397N, A397Q, A398L, D399E, L400F, F403E, S404P,

Q406V, L407I, Q408N, Q409W, S410G, S413A, A414L, A414V, D415G, S416E, Q418E, A419L, and A419N) in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 4a–f).

## Discussion/Conclusion

Different variants of coronaviruses have been identified infecting human and animals. The differences in genome structure and sequences make the coronaviruses to be different in severity of infection and host selectivity. Coronaviruses continuously change their structures via mutation, deletion, and/or insertion mutations. The most genomic variation part in SARS-CoV and SARS-CoV-2 is the RBD in the S protein [13, 14] and some locations in S protein sequence might be related to positive selection

[15]. Due to the several changes in the genome sequences of SARS-CoV-2 isolates, it is necessary to find the location of mutations and to understand the role of these mutations in the pathogenicity of SARS-CoV-2. This is significant to understand the viral infection mechanisms and pave a way to find drug and vaccine to protect people from the next stage of the pandemic.

### *Envelope Protein*

The surface envelope is the smallest but abundant structural protein. This protein participates to create envelope which is the important part of the virus, and it is responsible for protection of the interior parts of the virus and has a role in viral assembly during viral replication [25, 26]. Due to the envelope's roles in host cell infectivity, it is necessary to investigate the mutations in this protein in different animal and human coronaviruses, including SARS-CoV-2.

For this purpose, the full amino acid sequences of several envelope proteins were retrieved from GenBank and used in this study. Fifty-one animal viruses and 55 human viruses' envelope proteins were taken and analyzed (Fig. 1).

In envelope viruses' sequences, among the 76 residues, 365 mutations were observed and 47 in a statistically significant manner ( $p \leq 0.001$ ).

In animal viruses' sequences, 333 mutations were observed; 47, of them in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 1). In envelope human virus sequences, 94 mutations were observed and just 1, L28I, was observed in a statistically significant manner ( $p \leq 0.001$ ) (3 isolates; 5.4%) (Fig. 1a). These results show that there was high diversity seen in animal viruses than human viruses because animal viruses isolated in different species of animals, but for human viruses, only 1 host participated in the diversity. The rate of significant amino acid mutations (14.1%) is higher in animal coronaviruses than in human coronaviruses (0.9%). This means that human coronavirus preserves its envelope amino acids with the least number of amino acid mutations, and majority of its amino acids are conservative.

### *Spike Protein*

The surface spike glycoprotein is the most important part of the virus, and it is responsible in natural host selection to initiate infection. This trimmer protein helps in viral attachment to its corresponding receptor and mediates fusion of the cell and viral membrane [27–29]. The S protein has 2 main domains: S1 which is responsible to bind the host cell receptor, and S2 is responsible for the

fusion of membrane of both viral and host cellular membranes [12]. Due to the variability in sequences of the RBD SARS-CoV-1 and SARS-CoV-2 [13, 14], it is necessary to investigate the mutations in this protein in different animal and human coronaviruses including SARS-CoV-2.

For this purpose, the full amino acid sequences of several spike proteins were retrieved from GenBank and used in this study. Fifty-one animal viruses and 52 human viruses spike proteins were taken and analyzed. Due to the important role of RBD of the spike, only this domain was analyzed in this study. More accurately, amino acid residues 336–516 in SARS-CoV-2, which are crucial and enough to bind ACE2 receptor [24], were analyzed.

In general, in spike virus's sequences, among the 1,273 residues, 9,782 mutations were observed; 982, in a statistically significant manner ( $p \leq 0.001$ ). This large number of mutations indicates the wide variation between different host coronaviruses. Of which, 982 mutations are significant which means it is found in most of the virus's spikes of most of the coronaviruses. The same investigation was made in animal and human coronaviruses. In animal coronaviruses, 920 mutations were observed, and specifically in the RBD region, 83 of them in a statistically significant manner ( $p \leq 0.001$ ) (Fig. 2). The lower rate of significant number of mutations (8.1%) in human coronaviruses indicates the close relation between human coronaviruses more than in animal coronaviruses (9%). The variation in sequences is much less among human coronaviruses than animal coronaviruses. In spike human virus sequences, 222 mutations were observed, and specifically in the RBD region, 18 of them (E340S, F347I, W353F, K356S, V362P, Y369N, Y380V, G381D, T385L, K386P, D389T, D467N, T470Y, Y473L, G504Y, P507S, Y508L, and V512D) in a statistically significance manner ( $p \leq 0.001$ ) (Fig. 2). Only 222 amino acid mutations in human coronaviruses are different. This means that out of 1,273, 222 (17.4%) amino acids are different among different types of coronaviruses and 1,051 (82.5%) amino acids are conservative in all human coronaviruses. Of which, only 18 amino acids are significantly different which means they are the most variable amino acids prone to mutation in the spike proteins of human coronaviruses, and interestingly, most of them are located in RBD. The large proportion of significant amino acid mutations in RBD domain of spike protein indicating that this domain is very variable among human and animal coronaviruses and it may be because it is responsible for binding to the cell receptor in different hosts and it may indicate the strength of binding as well.

In RBD of SARS-CoV-2, many amino acids have a direct contact with ACE2 receptor and have a role in direct spike and corresponding receptor binding [24]. These amino acids are K417, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, Q493, G496, Q498, T500, N501, G502, and Y505. In our study, it is revealed that K417, T500, and Y505 amino acids are missing in most of animal coronaviruses and the differences are significant. The mentions amino acids are substituted to K417C, T500C, and Y505F in animal coronaviruses. According to ding [24], these 3 amino acids are part of 17 (17.6%) amino acids which are important in the binding of spike protein of SARS-CoV-2 to human ACE2 receptor. Therefore, significant substitution mutation in animal coronaviruses of these amino acids may have a role in host specific infectivity and are unable to infect human cells.

It is revealed in a different study that 193 amino acid residues (318–510) of RBD of SARS-CoV-1 spike protein are enough to bind to ACE2 receptor and among them, one-point mutation (aspartate-454) can abort the binding interaction between S1 domain of spike and ACE2 receptor [30]. In our study, this amino acid is not among significant animal and human coronaviruses mutations, and amino acids found previously are not required to have contact with ACE2 receptor [30]. Therefore, this amino acid may be very conservative in all coronaviruses and it has a big role in spike receptor binding but not host selectivity. F486 in RBD of the spike protein gives strength for RBD binding of spike protein of SARS-CoV-2 to the ACE2 receptor [31]. This mutation is also not present among significant amino acid mutations in animal and human coronaviruses; therefore, it may be one of the conservative amino acids.

Finally, a recent observation underlied that RDB plays a fundamental role as a damping element of the massive viral particle's motion prior to cell recognition, while also facilitating viral attachment, fusion, and entry [32].

### *Membrane Protein*

The surface membrane protein participates to create an outer layer of the virus which is the important part of the virus and it is responsible for giving a shape to the virus and protection of the interior parts of the virus. This membrane has a role in pathogenesis and virus entry into the cell. After binding of spike to the ACE2 receptor, it moves closer to its corresponding host cell membrane and mediates fusion of the cell viral membranes [6–9, 24, 31]. Due to its important role in pathogenesis, it is necessary to investigate the mutations in this protein in different animal and human coronaviruses, including SARS-CoV-2.

For this purpose, the full amino acid sequences of several membrane proteins were retrieved from GenBank and used in this study. Fifty-one animal viruses and 55 human viruses' membrane proteins were taken and analyzed.

In Figure 3, frequencies of SARS-CoV-2 membrane amino acid signatures were shown, using the isolate *NC\_045512.2 COVID-19/Wuhan* as a reference. In membrane viruses' sequences, among the 222 residues, 851 mutations were observed; 219, in a statistically significance manner ( $p \leq 0.001$ ). In animal membrane sequences, 794 mutations were observed, and specifically, 219 of them in a statistically significant manner ( $p \leq 0.001$ ). In animal viruses, rate of significant amino acid mutations is very high (27.5%) which means the diversity among them is very significant. This may have been related to the diversity of animal hosts ranging from birds, land animals, and sea animals where the virus lives and propagates because the virus needs to adapt themselves in new hosts and enjoynments to survive. On the other hand, the significant mutation rate (13.3%) is less in human hosts because there is no diversity between human hosts as seen in animals. In human membrane virus sequences, 262 mutations were observed, and specifically, 35 of them in a statistically significant manner ( $p \leq 0.001$ ).

### *Nucleocapsid*

Nucleocapsid is the pretentious structure inside the box of the coronaviruses and it combines with the nucleic acid, RNA. The function of nucleocapsid is to hang viral nucleic acid around itself by packaging the genomic viral genome into long, flexible, helical ribonucleoprotein complexes, the nucleocapsid. The nucleocapsid protects the RNA of the virus and ensures its replication and transmission [33, 34]. Due to large number of diversity between human and animal coronaviruses, it is necessary to investigate the rate of mutations of nucleocapsid between animal and human coronaviruses. For this purpose, the entire nucleocapsid protein sequences, derived from 51 animal viruses and 54 human viruses, were analyzed (Fig. 4).

In nucleocapsid virus's sequences, among the 419 residues, 1,885 mutations were observed; 317, in a statistically significance manner ( $p \leq 0.001$ ). In animal nucleocapsid sequences, 613 mutations were observed, and specifically, 38 of them in a statistically significant manner ( $p \leq 0.001$ ). In human nucleocapsid sequences, 1,628 mutations were observed, and specifically, 317 of them in a statistically significant manner ( $p \leq 0.001$ ). The number of significant amino acid mutations is lower in animal vi-

ruses at a rate of 6.1% than in human viruses, 19.4%. In all previously mentioned structural proteins, envelope, membrane, and spike, the rate of significant mutations was higher than in that of human viruses, whereas it is opposite in nucleocapsid protein.

A connection between conserved viral structure and possible target medicine treatment is mandatory. As recently reported, thousands of compounds including approved drugs and drugs in the clinical trial are available in the literature, and some anti-COVID-19 candidates based on computer-aided drug design can be followed up [35].

Here, in general, in all human structural proteins, significant amino acid mutation(s) was observed. The highest significant mutation (19.4%) was observed in nucleocapsid protein, whereas the lowest rate was in envelope protein (0.9%). This means that the structure of envelope and envelope protein is very stable and the amino acids are very conservative while it is opposite in nucleocapsid. The rate of significant amino acid mutations in spike protein (8.1%) is less than that in nucleocapsid (19.4%). In spite of the existing less-significant amino acid mutations in spike protein, the mutations in spike protein still have a crucial role in the viruses' host selectivity and infectivity because it is directly related to receptor binding and indicating the type of cell and host to infect. Therefore, the least mutation in RBD of spike may make a catastrophic effect on the human hosts than other structural proteins of the coronaviruses.

The structural proteins, spike, envelope, nucleocapsid, and membrane proteins are very genetically diverse among human and animal coronaviruses. The highest rate of significant mutation was found in nucleocapsid rather than in spike protein, but most of mutations in spike proteins are in the RBD part of S1 of the spike. The least rate of mutation was in envelope protein which means envelope protein amino acid residues are more stable and conservative. A single amino acid which dissociates spike binding with ACE2 receptor remains intact in all human and animal coronaviruses. In addition, among 17 amino acids that have a direct contact with ACE2 receptor, only 3 of them are significantly mutated and substituted in animal coronaviruses.

## Summary

Large number of significant mutations were recorded in animal coronaviruses than in human coronaviruses. The highest rate of significant amino acid substitution was found in nucleocapsid viral protein, but it is the low-

est in envelope protein which means envelope protein is more stable and less diverse. In spike protein, the highest number of significant amino acid mutations was found in RBD. Three out of 17 binding amino acids in RBD are significantly mutated in animal coronaviruses. A single amino acid of RBD, aspartate-454, which is essential for the binding of spike protein with ACE2 cell receptor remains intact in all human and animal coronaviruses.

## Acknowledgements

We acknowledge the president of Sulaimani Polytechnic University and especially Assistant Professor Dr. Alan Faraydoon Ali for his great help and academic support to performing this research. Also, we would like to appreciate Mam Humanitarian Foundation for its financial and logistic support.

## Statement of Ethics

No ethical approval was required, as it is not applicable to our research. According to the study design, neither medical treatments nor procedures involving humans or animals were performed.

## Conflict of Interest Statement

The authors have no conflicts of interest to declare.

## Funding Sources

This work was supported by Mam Humanitarian Foundation.

## Author Contributions

S.A. and M.B.-M. conceived the study, and T.H.-S. and S.D. participated in its design. All authors critically discussed and interpreted the results, drafted and critically reviewed this manuscript, and approved the final version.

## References

- 1 Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med.* 2020;172(9):577–82.
- 2 Varia M, Wilson S, Sarwal S, McGeer A, Gournis E, Galanis E, et al. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *CMAJ.* 2003;169(4):285–92.



- 3 Virlogeux V, Park M, Wu JT, Cowling BJ. Association between severity of MERS-CoV infection and incubation period. *Emerg Infect Dis*. 2016;22(3):526–8.
- 4 Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and sources of endemic human coronaviruses. *Adv Virus Res*. 2018;100:163–88.
- 5 Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727–33.
- 6 Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res*. 2017;27(1):119–29.
- 7 Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog*. 2018;14(8):e1007236.
- 8 Kirchdoerfer RN, Wang N, Pallesen J, Wrapp D, Turner HL, Cottrell CA, et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci Rep*. 2018;8(1):15701.
- 9 Yuan Y, Cao D, Zhang Y, Ma J, Qi J, Wang Q, et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat Commun*. 2017;8:15092.
- 10 Yu WB, Tang GD, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. *Zool Res*. 2020;41(3):247–57.
- 11 Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. 2020;181(2):281–e6.
- 12 Tortorici MA, Veesler D. Structural insights into coronavirus entry. *Adv Virus Res*. 2019;105:93–116.
- 13 Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–3.
- 14 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265–9.
- 15 Lu G, Wang Q, Gao GF. Bat-to-human: spike features determining “host jump” of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol*. 2015;23:468–78.
- 16 Dimonte S, Svicher V, Salpini R, Ceccherini-Silberstein F, Perno CF, Babakir-Mina M. HIV-2 A-subtype gp125c2-v3-c3 mutations and their association with CCR5 and CXCR4 tropism. *Arch Virol*. 2011;156:1943–51.
- 17 Dimonte S, Babakir-Mina M, Aquaro S, Perno CF. Natural polymorphisms of HIV-1 subtype-C integrase coding region in a large group of ARV-naïve infected individuals. *Infection*. 2013;41(6):1097–102.
- 18 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57(1):289–300.
- 19 Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*. 2019;17(3):181–92.
- 20 de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol*. 2016;14(8):523–34.
- 21 Fehr AR, Channappanavar R, Perlman S. Middle East respiratory syndrome: emergence of a pathogenic human coronavirus. *Annu Rev Med*. 2017;68:387–99.
- 22 Chen WH, Hotez PJ, Bottazzi ME. Potential for developing a SARS-CoV receptor-binding domain (RBD) recombinant protein as a heterologous human vaccine against coronavirus infectious disease (COVID)-19. *Hum Vaccin Immunother*. 2020;16(6):1239–42.
- 23 Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*. 2020;368(6491):630–3.
- 24 Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020;581(7807):215–20.
- 25 Venkatagopalan P, Daskalova SM, Lopez LA, Dolezal KA, Hogue BG. Coronavirus envelope (E) protein remains at the site of assembly. *Virology*. 2015;478:75–85.
- 26 Nieto-Torres JL, Dediego ML, Alvarez E, Jiménez-Guardeño JM, Regla-Nava JA, Llorente M, et al. Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein. *Virology*. 2011;415(2):69–82.
- 27 Siu YL, Teoh KT, Lo J, Chan CM, Kien F, Escriou N, et al. The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol*. 2008;82(22):11318–30.
- 28 Kirchdoerfer RN, Cottrell CA, Wang N, Pallesen J, Yassine HM, Turner HL, et al. Prefusion structure of a human coronavirus spike protein. *Nature*. 2016;531(7592):118–21.
- 29 Song HC, Seo MY, Stadler K, Yoo BJ, Choo QL, Coates SR, et al. Synthesis and characterization of a native, oligomeric form of recombinant severe acute respiratory syndrome coronavirus spike glycoprotein. *J Virol*. 2004;78(19):10328–35.
- 30 Wong SK, Li W, Moore MJ, Choe H, Farzan M. A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. *J Biol Chem*. 2004;279(5):3197–201.
- 31 Lone B, Madhurima V. Dielectric and conformational studies of 1-propanol and 1-butanol in methanol. *J Mol Model*. 2011;17(4):709–19.
- 32 Moreira RA, Chwastyk M, Baker JL, Guzman HV, Poma AB. Quantitative determination of mechanical stability in the novel coronavirus spike protein. *Nanoscale*. 2020;12(31):16409–13.
- 33 McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*. 2014;6(8):2991–3018.
- 34 de Haan CA, Rottier PJ. Molecular interactions in the assembly of coronaviruses. *Adv Virus Res*. 2005;64:165–230.
- 35 Boopathi S, Poma AB, Kolandaivel P. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. *J Biomol Struct Dyn*. 2020 Apr;30:1–10.