Investigating the genomic landscape in novel coronavirus (2019-nCoV) genomes to identify non-synonymous mutations for use in diagnosis and drug design

Manish Tiwari[1+] and Divya Mishra[2*+]

[1]National Institute of Plant Genome Research, Jawaharlal Nehru University Campus, Aruna Asaf Ali Marg, New Delhi, 110067, India.

[2]Department of Plant Pathology, Kansas State University, 66506, Kansas, United State of America.

[+] -These authors contributed equally.

***Correspondence:** divyam@ksu.edu

Significance

This study presents a comprehensive phylogenetic analysis of SARS-CoV2 isolates to understand discrete mutations that are occurring between patient samples. This analysis will provide an explanation for varying treatment efficacies of different inhibitory drugs and a future direction towards a combinatorial treatment therapies based on the kind of mutation in the viral genome.

Abstract

Novel coronavirus has wrecked medical and health care facilities claiming ~5% death tolls globally. All efforts to contain the pathogenesis either using inhibitory drugs or vaccines largely remained futile due to a lack of better understanding of the genomic feature of this virus. In the present study, we compared the 2019-nCoV with other coronaviruses, which indicated that bat-SARS like coronavirus could be a probable ancestor of the novel coronavirus. The protein sequence similarity of pangolin-hCoV and bat-hCoV with human coronavirus was higher as compared to their nucleotide similarity denoting the occurrence of more synonymous mutations in the genome. Phylogenetic and alignment analysis of 591 novel coronaviruses of different clades from Group I to Group V revealed several mutations and concomitant amino acid changes. Detailed investigation on nucleotide substitution unfolded 100 substitutions in the coding region of which 43 were synonymous and 57 were of non-synonymous type. The non-synonymous substitutions resulting into 57 amino acid changes were found to be distributed over different hCoV proteins with maximum on spike protein. An important di-amino acid change RG to KR was observed in ORF9 protein. Additionally, several interesting features of the novel coronavirus genome have been highlighted in respect to various other human infecting viruses which may explain extreme pathogenicity, infectivity and simultaneously the reason behind failure of the antiviral therapies.

Keywords

bat-hCoV, coronavirus, pangolin-hCoV, phylogeny, SARS, synonymous and non-synonymous substitutions.

Introduction

Current and last two decades saw emergence of zoonotic coronavirus (CoV) crossing the species barrier ultimately infecting human species resulting in pandemics such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) (Drosten et al., 2003; Zaki et al., 2012). An apocalyptic threat is posed by a sempiternal pathogen ruining the health and the economies on global scale. A severe pneumonia outbreak starting in December, 2019 in the Wuhan city, Hubei Province, China was caused by novel CoV referred as "2019 novel coronavirus" or "2019-nCoV" (Huang et al., 2020; Zhu et al., 2020). CoVs are RNA viruses with wide host pathogenesis in mammals including humans, pangolins and birds. Genetically the CoVs were categorised into four major genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* (Li, 2016). The alpha and beta CoVs infect mammals whereas the gamma and delta CoVs infect birds (Tang et al., 2015). Primary symptoms associated with CoV infection include respiratory, hepatic, enteric and neurological diseases. Previous investigation showed that there are 6 type of CoVs (hCoV-NL63, hCoV-229E, hCoV-OC43, hCoV-HKU1, SARS-CoV, and MERS-CoV) which can infect the human species. HCoV-NL63, hCoV-229E belongs to alphaCoV genus while rest belongs to betaCoV genus. (Tang et al., 2015). The betaCoVs appears to be *prima-facie* genre of CoVs which will peril universal human civilization in upcoming decades. Recently, the 2019-nCoV outbreak spread from China to the intercontinental arena and already infected 0.3 million people globally claiming ~13000 (~4.3%) deaths till 21st March 2020 (https://www.worldometers.info/coronavirus/#countries). China and Italy were the epicentres until now and chances for more calamitous centres cannot be ruled out in near future. Genome sequence analysis of SARS, MERS and 2019-nCoV confirmed its presence in betaCoVs family and divergence from the other two viruses (Zhu et al., 2020). The 2019-nCoV is a positive-strand RNA viruses with ~29 Kb genome size, 125 nm in diameter and 6 to 11 open reading frames (ORFs) (Song et al., 2019). Viral genome encodes for 4 major structural proteins namely envelope (E), spike (S), membrane (M) and 3–5 nucleocapsid (N) proteins. The genome starts with short untranslated regions (5' UTR) followed by genes 5′-replicase (rep gene), S, E, M, N and 3' UTR (Song et al., 2019). Two-third of the genome is represented by the rep gene at 5' end which encodes for non-structural protein (Nsp). Spike

protein is responsible for receptor binding and corresponding viral entry into the host and hence important target for future drugs to restrict the viral titre (Du et al., 2009, 2017). Viral assembly relies primarily on M and E proteins and RNA synthesis is achieved by action of N protein (Song et al., 2019).

To mitigate the severity of 2019-nCoV, researchers around the world are trying to develop antibodies and vaccine against this deadly virus. The problem with the delay in antiviral medication is superficial understanding of the virus. A dire need is to unravel the mutations in the viral genome and concomitant amino acid changes occurring presumably due to varying geographical location or upon interaction with the diverse human immune system. Various reports compared the SARS, MERS, bat and pangolin coronaviruses and paved way for significant findings, still leaving a lacunae in terms of the variations in the hCoV genomes and comparison with the previous available viruses resources. The present study deals with the mutations in the hCoV genomes and resulting change in amino acids.

Material and Methods

To analyse the phylogenetic relationship between different coronaviruses, 591 genomes were downloaded from Global Initiative on Sharing All Influenza Database (GISAID) (https://www.gisaid.org/). The hCoV is an RNA virus and the deposited sequences are in DNA format. To prevent anomaly in the data represented, complete genomes and only high coverage datasets were utilized. The genomic sequences were aligned using MUSCLE program (v3.8.31) (Edgar, 2004). The alignments were utilized to deduce various nucleotide substitutions and maximum likelihood phylogenetic tree with 1000 bootstrap was constructed by RAxML program (Stamatakis, 2014). The alignment and tree were visualized using Jalview 2.11.0 (Waterhouse et al., 2009) and iTOL respectively (Letunic and Peer, 2007). Different substitutions and resulting amino acid changes were analyzed between human, bat, pangolin and SARS coronavirus genomes. To deduce a mutation or amino acid change only those confirmed in three individual genomes were considered (replicates for biological significance).

Result and Discussion

Comparative genomic analyses of human novel coronavirus with other coronaviruses

Genomic features may provide an important clue about the relatedness and evolution of the organism. In order to get an insight into the similitude and dissimilitude between human

novel coronavirus (hCoV) and other coronaviruses, the genome sequence of human novel coronavirus (hCoV) were compared with bat coronavirus (GU190215.1) (Drexler et al., 2010), severe acute respiratory syndrome-related coronavirus strain BtKY72 (KY352407.1) (Tao and Tong, 2019), bat SARS-like coronavirus isolate bat-SL-CoVZC45 (MG772933.1) (Hu et al., 2018), hCoV-19/pangolin/Guangdong/1/2019|EPI ISL 410721 (pangolin-hCoV) and hCoV-19/bat/Yunnan/RaTG13/2013|EPI ISL (bat-hCoV) which revealed approximately 81%, 81%, 89%, 90% and 96% similarity, respectively (Table 1).

To further assess the relationship between hCoV and other coronaviruses, alignment and phylogenetic analysis was carried out. Alignment of hCoV with above mentioned viruses showed that several nucleotide sites were unique in hCoV sequences when compared to other coronaviruses (Table 2). Among these sites, C :T (hCoV :other coronavirus) is the most prevalent substitution followed by T:C, G:A and A:G (Table 2). Many regions were absent in bat and SARS coronavirus genome when compared to hCoV, bat-hCoV and pangolin-hCoV. Among these regions, one of largest portion is of 391 nt (28026-28417) coding for ORF8 protein in hCoV putatively involved in interspecies transmission (Lau et al., 2015). Genomic similarities and alignment indicate that several mutation events over the time is responsible for emergence of human novel coronavirus. Further a phylogenetic analysis between these viruses displayed that hCoVs are closer to bat/SARS-like virus (MG772933.1) and distant from SARS coronavirus (KY352407.1) and bat coronavirus (GU190215.1) (Figure 1). These results demonstrate that SARS coronavirus and bat coronavirus (GU190215.1) could be apparent ancestor of other coronaviruses studied in the investigation.

Scrutiny of nucleotide and amino acid in coding region of the genome revealed that the hCoV genome share 92.67% and 96.92% similarity at nucleotide level with pangolins and bat hCoV genome, whereas the similarity level increased up to 97.82% and 98.67% at amino acid level (Table 3). This indicates that most substitutions taking place were of synonymous type. Among various protein coding genes Nsp4-10, Nsp12-14, Nsp16, M, E and ORF6 shared highly conserve amino acid composition between bat-hCoV and hCoV with >99% similarity, especially Nsp7-10, Nsp16, E and ORF6 share 100% amino acid similarity (Table 3). The 100% similarity in these regions across 591 hCoVs, bat and pangolin-hCoV mark them to be a probable target region for future antibodies and vaccine therapy. Notably, Nsp2 and Nsp14 region in hCoVs were most diversified in terms of nucleotide when compared to pangolin and bat-hCoV, whereas ORF10 and E regions were the least diverse (Supplementary table 2).

Phylogenetic analyses of human novel coronavirus

We investigated the phylogenetic analysis of 591 genomic sequences of hCoV obtained from GISAID database using RAxML methods. The phylogram was majorly divided into 5 groups based on their clade division. Bat and pangolin-hCoV were categorized in the group I and all other hCoVs were categorized in group II to V (Figure 2). Group II comprises of the human 2019-nCoV mainly belonging to different province of China. However, few exceptions were also from South Korea, Japan, Vietnam, Chile, USA, India, Belgium, Spain, Germany and France hCoVs. Group III mainly comprised of the hCoVs belonging to USA while Group IV represented the hCoVs of mixed type population belonging to several countries distributed over continents. Group V possess the hCoV from European countries along with few hCoVs of America and one from Taiwan. To understand mystery underlying the clustering pattern of the hCoVs, bat and pangolin-hCoV were used as a reference sequence to observe the nucleotide substitution in hCoV members in different groups. Interestingly, hCoV members (Group II and Group III) falling in proximity to Group I have less substitution in the genome sequences (Table 4). The T:C (GroupV-hCoV:bat-hCoV and GroupV-hCoV:pangolin-hCoV) substitution were frequent in Group V as compared to hCoV representing other groups (Table 4) . The genomic signature of USA-hCoVs present in Group V is very different from USA-hCoVs of Group III. This could be indicative of differences between direct and community transmission of the virus. Member belonging to each subgroup has distinct genomic features in terms of nucleotide substitution (Table 4).

Non-synonymous substitutions and associated amino acid changes

Genomic comparison of 591 hCoV sequences among the human as well as with pangolin-hCoV and bat-hCoV revealed several sites possessing substitutions which clearly indicated the mutation in viral genome either according to the geographical locations or upon interaction with the human immune system. A detailed investigation of the nucleotide substitution in the coding region of hCoVs genome with perspective of encoded amino acids revealed 43 synonymous and 57 non-synonymous substitutions (Table 5). The proteins Nsp1, Nsp5, Nsp7-10, Nsp14-16, ORF4, ORF7a, ORF7b and ORF10 mainly possessed synonymous substitutions and hence were mostly devoid of amino acid changes. The 57 amino acid changes were distributed over 12 regions in the ~30kb genome. The number of amino acids substitutions varied between different regions such as 7 in Nsp2, 10 in Nsp3, 5 in Nsp4, 3 in Nsp6, 1 in Nsp12, 4 in Nsp13, 11 in Spike, 3 in ORF3a, 2 in ORF5, 1 in ORF6, 2

in ORF8 and 8 in ORF9 (Figure 3). Intriguingly, various important non-synonymous mutations were observed majorly in European and US continent while the mutations were mostly synonymous in Asian continent. These interesting observations can be used to infer the reason behind larger infectivity and pathogenicity in these regions (Table 5). Further two type of amino acid change *viz*., conservative and radial replacements were intensively studied with respect to previous reports stating the effect of such changes on the enzymatic activities. Mutations were most prevalent in the spike region followed by Nsp2, Nsp3 and ORF9 (N) (Table 5). Spike region determines the specific binding to host receptor and initiation of viral replication. This region is reported to be the most potent and indispensable for viral attachment and entry into host system. The RRAR amino acids found only in the human CoVs spike region has proved to be essential for binding to host receptor (Walls et al., 2020). We observed similar region in the hCoV genomes studied (23713-23724 region in nucleotide alignment), although there was mutation in two hCoV-England nucleotide sequences (CTCC**G**CGGCGGG in place of CTCC**T**CGGCGGG) but the resulting amino acid remained same in all hCoV genomes. These findings corroborate the essentiality of RRAR sequence for viral infection to host system. We found different type of mutation in hCoV spike protein at different places such as leucine to valine (L8V), glutamine to histidine (Q675H and also found in ORF3a:Q57H), glutamine to lysine (Q239K) and aspartate to glycine (D614G and also found in ORF5, D3G) might have potential role to augment viral infection (Table 5). Previous investigations showed mutations such as leucine to valine change in retroviral envelope protein, glutamine to lysine in influenza virus, glutamine to histidine and aspartate to glycine in H1N1 had a severe impact in virus entry, replication and cross infectivity to other species (Côté et al., 2012; Glinsky, 2010; Yamada et al., 2010).

Additionally, mutations were present in structural proteins such as, glycine to valine mutation in ORF3a (G196V and G251V). Similar amino acid change imparts resistance against inhibitor drug saquinavir in the human immunodeficiency virus type 1 (HIV-1). This might provide an explanation why drugs used for treating HIV became a failure in case of hCoV infection (Hong et al., 1997). Notably, in ORF9 region the nucleotide sequence GGG changed to AAC in European and American continent resulting in a change of amino acid from RG to KR (AGGGGA coding for RG changed to AAACGA coding for KR, 28993-28995 in nucleotide alignment).

Furthermore, several amino acid changes were also observed in the non-structural proteins (Nsp) of the hCoVs which may affect the virulence and titer. Threonine to isoleucine

substitution was observed in different Nsp proteins (Nsp2:T85I, Nsp3:T127I and T1030I and Nsp4:T295I) mainly in European and US samples. Earlier reports established that threonine to isoleucine substitution increased viral infectivity of Ebola virus and resistance to ganciclovir in human cytomegalovirus (Kurosaki et al., 2018; Wolf et al., 1995). Importantly, alanine to valine substitution in non-structural protein, NS2A in Zika virus affects viral RNA synthesis and results in vivo viral attenuation (Marquez-Jurado et al., 2018). This mutated virus also induce a comprehensive protection against lethal challenge proposed by the wild type Zika virus. Falling in similar lines alanine to valine substitutions in non-structural proteins (Nsp3, A1187V, Nsp4, A457V and Nsp6, A46V) could reduce viral lethality of hCoVs (Table 5). These mutations might pave way towards identification of less lethal strains and help to raise immunity to counteract the noxious strains. An isoleucine to valine mutation (Nsp2, I559V and Nsp3, I797V) and methionine to isoleucine (Nsp4, M33I) were observed in hCoVs. Change of isoleucine to valine in polymerase subunit PB2 of influenza virus resulted in critically enhanced activity of reconstituted polymerase complex (Rolling et al., 2009) and M to I substitution in HIV-1 reverse transcriptase imparted resistance to nucleoside analog 2′,3′-dideoxy-3′thiacytidine (3TC) (Julias et al., 2004). Interestingly, presence of a non-synonymous substitution in RNA Dependent RNA Polymerase (RDRP) region in majority of European hCoV samples resulted in change of amino acid from proline to leucine (P314L). It will be quite interesting to validate the effect of this substitution on RDRP activity as one of the previous study established that similar change of proline-to-leucine substitution (P236L) of HIV-1 reverse transcriptase, imparts resistance against a highly specific inhibitor bisheteroarylpiperazines (BHAPs) (Fan et al., 1995). These examples clearly show that amino acid changes may significantly affect the functional competency of polymerase and the associated subunits.

In conclusion, present study enlightens about several types of mutation such as deletion, insertion and substitutions present in 2019-nCoV samples. These mutations may vary at different geographical distribution or interaction with different host systems. Few mutations also resulted in change of amino acid which may provide an explanation for failure of previously employed antiviral therapies. This research will better equip the researchers to utilize the mutated amino acid information for drug targets in particular geography and less cases of failure. Beside the substitution resulting into transformation to a more virulent strain there are number of highly conserved regions in the hCoV genome which can be used as target for inhibitory drugs and vaccine development for a large repertoire of strains. Finally,

we believe that our data provide useful information pertaining the changes in genomic and proteomic features which could serve as a guide to design the future antiviral therapies and diagnostics.

Acknowledgements

We kindly acknowledge National Institute of Plant Genome Research (NIPGR) and Department of Biotechnology, Govt. of India (http://www.dbtindia.nic.in).

Author contributions

M.T. performed the computational analysis, D.M. prepared all the figures and tables. M.T. and D.M. designed the project and wrote the article.

Conflict of interest

The authors declare no conflict of interest.

Figure legends

Figure 1. Phylogenetic relationship of hCoVs with other coronavirus. Phylogram of human novel coronavirus and other viruses. Phylogenetic analysis of bat coronavirus (GU190215.1), severe acute respiratory syndrome-related coronavirus strain BtKY72 (KY352407.1), bat SARS-like coronavirus isolate, bat-SL-CoVZC45 (MG772933.1), pangolin-hCoV, bat-hCoV and hCoV using the maximum-likelihood method (RAxML) keeping the bootstrap value 1000. Human coronavirus (hCoV, pangolin-hCoV, bat-hCoV) and bat SARS-like coronavirus falls in one clade while Severe acute respiratory syndrome-related coronavirus strain BtKY72 (KY352407.1) and bat coronavirus (GU190215.1) in another clade.

Figure 2. Phylogenetic relationship among 2019-nCoV. The phylogenetic tree of 519 hCoVs sequence were divided in 5 groups. Bat and pangolin hCoVs were categorised in group I and rest of the hCoVs in group II-V. The phylogram was constructed by maximum likelihood keeping the bootstrap value 1000.

Figure 3. The genomic landscape of the hCoV genome representing amino acid changes. The non-synonymous mutations resulting into 57 amino acid changes in respective proteins are marked. The numerals on the arrow represent the position corresponding to amino acid change.

Table 1: The percentage similarity of genomes of bat coronavirus, bat SARS like coronavirus, SARS coronavirus, pangolin-hCoV and bat-hCoV compared to 2019-hCoV genome

Coronavirus	Identity with hCoVs	Adenine	Guanine	Thymine	Cytosine
GU190215.1 Bat coronavirus BM48-31/BGR/2008	80.84%	8130	6151	9304	5691
KY352407.1 Severe acute respiratory syndrome-related coronavirus strain BtKY72	80.84%	8354	6052	9440	5428
MG772933.1 Bat SARS-like coronavirus isolate bat-SL-CoVZC45	89.12%	8740	6020	9468	5574
hCoV19/pangolin/Guangdong/1/2019|EPI ISL 410721	90%	8956	5862	9465	5542
hCoV19/bat/Yunnan/RaTG13/2013|EPI ISL 402131	96%	8912	5857	9587	5489

Table 2: The list of substitution in hCoV sample (CoV-19/Shandong/LY008/2020|EPI_ISL_414941/1-29868) with respect to all coronavirus [(found common nucleotide and variable nucleotide in SARS (KY352407.1 severe acute respiratory syndrome-related coronavirus strain BtKY7O), batCoV (GU190215.1 Bat coronavirus BM48-31/BGR/2008) , bat/SARS-CoV MG772933.1 (bat SARS-like

coronavirus isolate bat-SL-CoVZC45), pangolin CoV (hCoV-19/pangolin/Guangdong/1/2019|EPI ISL 410721) and bat CoV (hCoV-19/bat/Yunnan/RaTG13/2013|EPI ISL 402131]

Position	All coronavirus (common)	hCoV	Position	All coronavirus (variable	hCoV
378	C	T	29376	A, T	G
379	T	C	691	C	A
487	T	G	805	C,A,T	G
511	C	T	883	G,T	A
943	A	G	1186	T,C	G
1516	C	T	1387	T,G	C
1594	T	C	1420	T,A	C
1714	T	C	1480	a,T	c
2486	T	C	1627	T,G,A	C
2921	A	G	1843	A,C	G
3273	G	C	1948	C,A	T
3662	t	c	2062	a,T	c
3923	A	G	2107	a,c	t
4506	G	A	2110	t,g	c
4658	G	A	2143	t,g	c
4964	T	C	2170	c,g	t
5534	T	C	2227	a	c
5579	T	C	2648	t,g	c
5667	C	T	2847	A,T	G
5715	C	A	2888	T,A	C
5732	G	A	3476	A,T	G
6011	T	C	3530	T,A	C
6098	T	C	3623	G,C	A
6576	G	A	3626	T,A	C
6579	G	A	3808	C,A	T
6952	T	C	3881	A,T	G
7336	T	C	3989	C,T,G	A
7540	T	C	4009	T,C	A
7804	T	C	4226	A,C,T	G
8335	T	C	4349	A,T	G
8389	T	C	4487	T,G	C
9298	T	C	4520	T,A,G	C
9292	A	G	4958	T,A	C
9358	T	C	5530	A,C	G
9391	T	C	5615	T,A	C
9703	C	T	6095	G,C,T	A
9770	G	A	6335	T,A	C
10690	G	A	6353	T,G	C
11080	T	C	6605	T,A	C
11153	G	T	7010	T,A	C
11776	t	c	7462	A,G	G
11881	T	A	7489	T,A,G	C
11896	T	C	7548	C,T	A
11974	T	C	7597	T,G	C
12187	T	C	10186	T,C	a
12887	T	C	10351	T,A,G	C
13273	A	G	10354	A,T	G
15309	T	C	10384	G,T	A
15330	C	T	11272	A,T	C
15792	T	C	11830	A,T	C
16024	T	C	12511	A,G	C
16089	T	C	12643	T,C,G	A
16183	T	C	15999	G,C	T
16764	C	T	16032	T,A	C
17409	T	A	16554	G,T	A

17622	T	C	16938	C,A,G	T
17748	T	C	18006	T,A	C
17922	T	C	18339	T,A	C
18555	T	C	18420	T,A	C
18657	G	A	18441	C,G	T
20371	C	T	18558	T,A	C
20895	T	C	18692	A,T	G
21129	T	C	18729	T,G	C
21165	C	T	19818	A,C,T	G
21174	A	T	19947	G	T
21306	T	C	20033	A,G	C
21751	C	T	20034	T,C	G
21808	T	C	20163	T,A	C
22734	A	G	20250	A,T	C
22764	C	T	20700	A,G	C
22968	T	C	20706	T,G	A
23238	A	C	21195	A,T	G
23810	T	C	21707	A,G	T
24044	T	C	22578	A,T	G
24176	T	A	22593	T,C	A
24257	T	C	22797	T,A,G	C
24509	T	C	22857	T,G	C
24824	A	G	22902	T,G,A	C
25031	T	C	23103	A,T	G
25122	C	T	23666	T,G,C	A
25160	A	G	23753	A,C	T
25262	A	G	23834	T,A	C
25602	C	T	23861	T	A
25690	A	G	23885	A,T	C
26295	C	T	24899	G,T	A
26459	C	T	25379	T, C	G
26863	C	T	24616	g,a	t
26974	G	A	26149	T, A	C
27076	G	A	26663	G	T
28085	c	a	26664	A	C
28093	a	t	26950	T,G	C
28133	g	a	26983	A,T	G
29259	t	c	27587	T,A	C
29505	c	t	28539	A, T	G
29936	t	c	29232	C,A	T
29925	g	t	24326	A.C,T	G
28546	C	T	23765	A,T	C
25567	A	G	22332	T,A	G
23157	C	T	22175	C,T,G	A
20794	C	T	21378	A,T	C
14104	G	A	21327	A,T,C	G
13136	C	T	20616	A,G,T	C
6773	C	T	19596	A,G,T	C
2451	T	C	18945	A,T	G
2037	T	C	18900	G,T,A	C
1391	C	Y	14190	A,T	G
1094	C	T	12142	C,T	G
544	T	A	12013	G,A,T	C

Table 3: The comparison of nucleotide and amino acid similarity of pangolin-hCoV and bat-hCoV with hCoV

Gene(s)	**Nucleotide**		**Amino acid**	
	Pangolin hCoV/hCoV	**Bat hCoV/hCoV**	**Pangolin hCoV/hCoV**	**Bat hCoV/hCoV**
Nsp1	91.68	96.29	95	96.11
Nsp2	88.47	95.4	93.89	98.28

Nsp3	88.7	95.61	92.4	96.66
Nsp4	90.13	95.87	97.6	99.6
Nsp5	93.98	95.86	99.35	99.67
Nsp6	92.87	95.98	98.97	99.66
Nsp7	91.97	97.59	100	100
Nsp8	93.43	98.15	100	100
Nsp9	88.86	98.82	100	100
Nsp10	95.92	98.32	100	100
Nsp12	91.38	97.82	99.46	99.57
Nsp13	92.18	98.06	99.83	99.83
Nsp14	88.56	97.22	98.86	99.24
Nsp15	92	96.24	97.69	97.69
Nsp16	92.51	95.64	100	100
ORF3A	93.24	96.26	97.1	97.83
ORF4/E	99.12	99.56	100	100
ORF5/M	93.62	95.37	98.65	99.55
ORF6	95.7	98.38	96.72	100
ORF7A	93.44	95.63	97.54	97.54
ORF7B	92.42	99.24	95.35	97.67
ORF8	92.08	96.72	94.26	94.26
ORF9/N	96.11	96.83	97.62	98.81
ORF10	99.15	99.15	97.37	97.37
SPIKE	89.17	92.89	highly variable	97.41
average	92.6676	96.916	97.81916667	98.67

Table 4: The list of nucleotide substitution in Group II-V compared to Bat-hCoV and Group II-V compared to Pangolin-hCoV

Position (s)	**Bat-hCoV**	**Pangolin--hCoV**	**Presence of mismatches more than one in members of the groups as compared to other groups**			
	GroupI		**GroupII**	**GroupIII**	**GroupIV**	**GroupV**
28256	C	C	C	C	T	T
26642	A	G	A	A	A	G
26438	C	C	C	T	C	C
26256	G	G	G	T	T	G
26091	G	G	T	G	G	G
25922	C	C	C	G	C	C
25675	A	A	T	T	T	G
24974	A	G	A	A	A	G
24792	G	G	G	T	G	G
24493	C	C	C	T	C	C
24501	A	A	G	G	G	C
24502	G	G	G	G	G	C
24437	A	A	A	G	A	A
24146	C	T	T	C	C	C
23870	A	A	C	T	C	C
23843	C	C	C	C	C	T
23717	-	-	G	G	G	G
23687	C	T	C	C	C	T
23699	G	G	G	G	G	C
23681	A	A	T	T	G	T
23632	C	C	C	C	C	T
23515	A	A	A	A	A	G
23297	C	C	C	T	C	C
23122	T	A	T	C	T	T
22773	G	G	G	T	G	G
22435	C	-	C	C	T	C
22389	C	A	C	C	C	A
22131	T	G	T	T	C	T

21815	C	T	C	C	T	C
21680	T	T	T	G	T	T
21588	T	T	T	C	T	T
21482	C	C	C	T	C	C
21165	T	T	T	T	C	T
20290	A	A	A	A	A	G
20241	T	T	T	T	T	C
19932	T	T	T	T	T	C
18153	T	A	T	T	T	C
17591	A	A	G	A	A	A
17840	C	C	C	C	T	T
17503	C	C	C	C	T	C
17466	T	T	C	C	T	C
17340	T	C	T	T	C	T
16627	G	G	G	G	T	G
16559	A	G	A	A	G	A
16548	T	T	T	T	T	G
16473	G	G	A	G	G	G
15416	C	C	C	C	T	T
14897	C	C	T	C	T	C
14816	C	C	C	C	T	C
14500	C	C	C	C	C	T
14039	A	A	A	A	A	T
14021	T	T	T	T	T	C
13786	C	C	C	T	C	C
13628	C	T	C	C	C	T
13595	G	G	G	G	G	C
13494	T	T	T	T	T	G
13500	T	T	T	T	T	G/A
13268	C	C	C	C	C	T
13143	C	C	C	C	C	T
12837	G	G	G	G	G	A
12565	C	C	C	C	C	T
11796	G	G	G	G	T	G
11522	A	A	G	A	A	A
11530	A	A	A	A	A	G
11502	A	A	A	G	A	A
11201	C	C	C	C	C	T
11176	G	G	T	T	G	G
10833	T	C	C	C	C	T
10820	C	C	C	C	C	T
10357	A	A	A	A	A	G
10230	T	T	T	T	C	T
10189	G	G	G	G	G	A
10054	C	C	C	C	T	C
10016	C	C	C	C	T	C
10023	T	T	T	T	C	T
9606	A	G	A	A	G	A
9569	A	A	T	A	A	A
9571	G	G	G	G	T	G
9530	C	C	C	C	T	C
8874	C	C	T	T	C	C
8859	T	T	C	T	T	T
8745	G	G	G	G	T	G
8514	G	G	G	G	A	G
7890	G	G	G	T	G	G
6593	C	C	T	C	C	C
6520	C	C	C	C	C	T
6402	C	C	C	T	C	C
6347	C	C	C	C	C	T
5952	T	T	T	C	T	T
5876	C	C	C	T	C	C
5376	C	T	C	C	C	T
5176	A	A	A	A	T	A

5154	G	G	G	T	G	G
4747	C	C	C	C	T	C
4494	T	T	T	C	T	T
4347	A	T	G	G	A	G
4094	C	C	C	C	C	T
3817	G	G	G	G	G	T
3462	C	C	C	C	C	A
3254	C	C	C	T	T	T
3228	G	G	G	G	G	T
3129	A	A	A	G	A	A
3182	C	C	C	C	T	C
3120	T	C	C	C	C	T
2974	G	G	G	G	A	G
2745	C	T	T	C	C	C
2641	C	C	C	C	T	C
2610	G	G	G	G	G	T
2563	A	A	A	A	G	A
2499	T	T	C	C	C	T
1774	A	A	A	A	G	A
1749	T	T	T	T	T	C
1677	C	C	C	C	C	T
1653	T	T	T	T	C	T
1480	G	A	G	G	A	G
1523	G	G	G	G	A	G
1142	C	C	C	C	C	T
1273	C	C	C	C	C	T
1029	T	T	C	T	T	T
915	C	C	C	T	C	C
597	C	C	T	T	C	T
462	C	C	C	C	C	A
396	C	C	C	C	C	T
337	C	C	C	C	T	C
324			C	C	C	T
290	C	C	C	C	C	T
269	T	C	C	C	T	C
270	A	A	A	A	A	G
29747	C	C	C	C	T	C
29739	C	C	C	C	T	C
29705	G	G	A	G	G	G
29658	C	C	C	C	A	C
29665	G	G	G	G	G	A
29486	G	G	G	G	A	G
29415	C	C	C	C	T	C
29402	G	G	G	G	G	T
29283	T	T	T	T	T	C
29207	T	T	T	C	C	C
28995	G	G	G	G	G	C
28994	g	G	G	G	A	A
28993	g	G	G	G	A	A
28990	G	G	A	G	G	G
28975	C	C	T	C	C	C
28966	C	C	C	C	T	C
28938	C	C	C	C	C	T
28851	G	G	G	G	G	T
28800	T	T	T	T	C	T
28769	C	A	T	C	C	C
28692	G	G	T	G	G	G
28569	G	G	G	G	A	G
28490	A	A	G	G	T	G
28456	C	C	C	C	G	A
28189	G	G	C	G	G	G
27158	C	T	G	T	C	C
26871	T	T	C	T	T	T
26841	T	T	C	T	T	T

26789	T	T	T	T	C	T
26752	G	G	G	G	T	G
26642	A	G	A	A	A	G
26438	C	C	C	C	T	C
26200	C	C	T	C	C	C
25665	C	C	T	C	C	C
25462	C	C	C	C	C	T
25425	G	G	G	G	T	G
24229	C	C	C	C	C	T
20361	A	G	A	A	A	G
20327	C	C	C	C	C	T
20144	C	C	C	C	C	T
20140	A	A	A	A	G	A
18696	C	T	C	T	T	T
18661	C	T	T	C	C	C
18563	c	c	T	C	C	C
967	C	C	C	C	T	C
998	A	A	A	A	G	A
948	C	C	C	T	C	C
697	G	G	G	G	A	G

Table 5: Distribution of synoymous, non-synonymous nucleotide changes and associated amino acid changes in different protein coding genes of hCoV genome

		Nucleotide			**Amino Acid**			
Gene	Region	Pangolin	Bat	Human	Position of Amino Acid change	Original Amino Acid	Changed Amino Acid	Remark
5'UTR	324	C	C	C				T found in European countries and USA no Asian country except Taiwan
NSP1	597	T	T	T	no change			C in 5 Human, 3Netherland, 1 England, 1 USA
NSP2	967	C	C	C	27	R	C	T in 4 Australia
NSP2	1029	T	T	T	no change			C in 3 China
NSP2	1142	C	C	C	85	T	I	T in US and Europe samples
NSP2	1273	C	C	T	129	P	S	T IN 3 France

								samples
NSP2	1480	A	G	G	198	V	I	A in 19 Human
NSP2	1523	G	G	A	212	G	D	A in 10 Germany, 3 Netherland and 1 Wales
NSP2	1653	C	T	C	no change			C in 4 Netherland
NSP2	1677	T	T	C	no change			T in 4 Netherland
NSP2	1749	T	T	T	no change			C in 2 France and 1 Netherland
NSP2	2499	T	T	C	no change			T in France
NSP2	2563	A	A	A	559	I	V	G in 1 Brazil, 1 Netherland, 1 Scotland and 4 England
NSP2	2641	A	A	A	585	P	S	G in 1 Brazil, 1 Netherland, 1 Scotland and 4 England
NSP2	2745	T	C	C	no change			T in 3 Japan
NSP3	2974	G	G	G	58	A	T	A in 10 Germany, 1 Wales and 3 Netherland
NSP3	3120	C	T	C	no change			T in European and USA, no Asian country
NSP3	3129	A	A	A	127	T	I	G in 3 USA
NSP3	3228	C	C	C	142	L	F	T in 8 Netherland
NSP3	3254	C	C	T	151	L	P	C in 5 Shangdong
NSP3	4347	T	A	G	no change			2 T in Netherland and 2 A in Australia
NSP3	4494	T	T	T	no change			C in South Korea, Beijing and Vietnam

NSP3	4747	C	C	C	654	R	W	T in 4 Germany
NSP3	5154	G	G	G	789	L	F	T in 8 South Korea and 4 Biejing
NSP3	5176	A	A	A	797	I	V	G in 2 USA and 1 Canada
NSP3	5876	C	C	C	1030	T	I	T in 4 USA
NSP3	6347	C	C	C	1187	A	V	T in 3 Netherland and 1 South Korea
NSP3	6402	T	C	C	no change			T in 3 USA
NSP3	6593	C	C	C	1269	P	L	T in India, China and Guangdong
NSP4	8745	G	G	G	33	M	I	T in 4 Australia and 1 Georgia
NSP4	8874	T	T	C	no change			T in Asia and USA mainly
NSP4	9530	C	C	C	295	T	I	T in 3 France samples
NSP4	9569	T	T	T	308	F	Y	A in 2 France, Spain, 4 Chile and USA
NSP4	9606	G	A	A	no change			G in Germany, USA and England
NSP4	10016	C	C	C	457	A	V	T in 4 USA samples
NSP4	10023	C	T	T	no change			C in England and Germany
NSP4	10054	C	C	C	470	H	Y	T in 2 Hongkong and 1 cannine Hongkong
NSP5	10230	T	T	C	no change			T in 3 Singapore
NSP5	10833	C	T	C	no change			T in 2 Netherland and 1 England
NSP6	11175	G	G	G	37	L	F	T in almost every country

								representative
NSP6	11201	C	C	C	46	A	V	T in 4 Netherland
NSP6	11522	A	A	A	153	Y	C	G in China
NSP6	11796	T	T	C	no change			T in 2 Hongkong and 1 canine Hongkong
NSP8	12565	T	T	C	no change			T in Guangdong and Shanghai
NSP10	13494	T	T	T	no change			G in 4 Belgium
NSP12	14021	T	T	T	no change			C in 6 Hongkong 1 cannine
NSP12	14500	C	C	C	314	P	L	T in majority of European countries few USA, Georgia and no Asian
NSP12	14816	T	C	C	no change			T in 2 Hongkong, cannine
NSP12	14897	C	C	C	no change			T in European and USA
NSP12	15416	C	C	C	no change			T in France, China, Japan, Ghuangzhou, Guangdong and Switzerland
NSP13	16473	G	G	G	49	V	I	A in 3 Netherland samples
NSP13	16559	G	A	A	no change			G in 4 USA
NSP13	17340	C	T	T	no change			C in England, Netherland, Scotland and Brazil
NSP13	17466	T	T	C	no change			T in India, China, Fujian, 2 Foshan and Singapore
NSP13	17503	C	C	C	392	R	C	T in 2 Netherland and 1 USA

NSP13	17840	C	C	C	504	P	L	T in USA and 2 Canada
NSP13	17951	A	A	A	541	Y	C	G in 2 Canada and many USA
NSP14	18153	A	T	C	no change			T in various China, USA and 1 Canada no European or other Asian
NSP14	18696	C	T	T	no change			C inUSA and Australia
NSP15	19932	T	T	T	no change			C in Switzerland and 1 Netherland
NSP15	20241	C	C	C	no change			T in 2 Belgium and 1 USA
NSP15	20361	A	A	A	no change			G in 2 Switzerland, 1 Panama and 1 Spain
NSP16	21165	T	T	T	no change			C in 3 England
SPIKE	21680	T	T	T	8	L	V	G in 2 Hongkong and 1 canine
SPIKE	21815	C	C	C	49	H	Y	T in 1 Jiangsu, 1 USA and 3 Guangdong
SPIKE	22389	A	C	C	239	Q	K	A in 5 Netherland and 1 Finland
SPIKE	22773	G	G	G	367	V	F	T in 5 Franceand 1 Hongkong
SPIKE	23122	T	A	T	483	V	A	C in 3 USA
SPIKE	23297	C	C	C	no change			T in 4 USA
SPIKE	23515	A	A	A	614	D	G	G in Europe, USA continent and 1 Wuhan
SPIKE	23687	T	C	C	675	Q	H	T in Denmark, Luxembourg and

								Scotland
SPIKE	23713-24				682-85		RRAR	CTCCTCGGCGGG in Human, CTCCGCGGCGGG in 2 England
SPIKE	24146	C	T	C	no change			T in Beijing, 3USA, Sichuan, Vietnam and Nepal
SPIKE	24437	A	A	A	no change			G in 2 Wuhan and 1 USA
SPIKE	24493	C	C	C	940	S	F	T in Spain 2 Australia
SPIKE	24501	A	A	A	943	S	P/T	C in 3 Belgium/c in place of g in 4 Belgium
SPIKE	24974	G	A	A	no change			G in 5 Netherland
SPIKE	25462	C	C	C	1263	P	L	T in 3 England
ORF3A	25675	A	A	G	57	Q	H	T in mainly France, USA, some Europe and Taiwan
ORF3A	26091	G	G	G	196	G	V	T in France, Chile, Spain and USA
ORF3A	26256	G	G	G	251	G	V	T in various countries
ORF4/E	26438	C	C	C	no change			T in 4 USA
ORF5/M	26642	G	A	A	3	D	G	G in 2 England, USA, Switzerland, Finland and Scotland
ORF5/M	26841	T	T	T	no change			C in Beijing, 2 USA, Vietnam, Hongkong and Sichuan
ORF5/M	27158	C	C	C	175	T	M	T in European only mainly in

								Netherland
ORF6	27439	G	G	G	42	K	N	T in 3 USA
ORF8	28189	G	G	G	62	V	L	C in USA, Beijing, Hongkong, Sichuan and Vietnam
ORF8	28256	C	C	T	84	L	S	C in most Human all across
ORF9	28456	C	C	C	24	T	N	A in 4 Netherland
ORF9	28769	A	C	C	no change			T in 2 France, 4 Chile, 1 USA and 1 Spain
ORF9	28800	T	T	T	no change			C in Europe, USA and China
ORF9	28851	G	G	G	156	A	S	T in 2 England and 1 Netherland
ORF9	28966	C	C	C	194	S	L	T in 3 USA, Foshan, Canada and Shenzen
ORF9	28975	C	C	C	197	S	L	T in 2 France, Spain, 4Chile and USA
ORF9	28990	G	G	G	202	S	N	A in 4 Australia, Guangdong, Singapore, 3 Netherland and USA
ORF9	28993-95	GGG	GGG	GGG	203-4	RG	KR	AAC in place of GGG in various European countries and USA
ORF9	29207	T	T	C	no change			T in USA, China and Japan
ORF9	29260	T	T	T	292	I	T	C in 3 Switzerland and Netherland

ORF9	29415	C	C	C	344	P	S	T in 3 China and 2 Japan
ORF9	29486	G	G	G	no change			A in 3 Australia
ORF10	29665	G	G	G	no change			A in 3 USA
3'UTR	29854	G	G	G				T

References

Côté, M., Zheng, Y.-M., Li, K., Xiang, S.-H., Albritton, L.M., and Liu, S.-L. (2012). Critical role of leucine-valine change in distinct low pH requirements for membrane fusion between two related retrovirus envelopes. The Journal of Biological Chemistry *287*, 7640–7651.

Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth, D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yordanov, S., et al. (2010). Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. Journal of Virology *84*, 11336–11349.

Drosten, C., Günther, S., Preiser, W., van der Werf, S., Brodt, H.-R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R.A.M., et al. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. New England Journal of Medicine *348*, 1967–1976.

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.-J., and Jiang, S. (2009). The spike protein of SARS-CoV — a target for vaccine and therapeutic development. Nature Reviews Microbiology *7*, 226–236.

Du, L., Yang, Y., Zhou, Y., Lu, L., Li, F., and Jiang, S. (2017). MERS-CoV spike protein: a key target for antivirals. Expert Opinion on Therapeutic Targets *21*, 131–143.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research *32*, 1792–1797.

Fan, N., Evans, D.B., Rank, K.B., Thomas, R.C., Tarpley, W.G., and Sharma, S.K. (1995). Mechanism of resistance to U-90152S and sensitization to L-697,661 by a proline to leucine change at residue 236 of human immunodeficiency virus type 1 (HIV-1) reverse

transcriptase. FEBS Letters *359*, 233–238.

Glinsky, G. V (2010). Genomic analysis of pandemic (H1N1) 2009 reveals association of increasing disease severity with emergence of novel hemagglutinin mutations. Cell Cycle (Georgetown, Tex.) *9*, 958–970.

Hong, L., Zhang, X.-J., Foundling, S., Hartsuck, J.A., and Tang, J. (1997). Structure of a G48H mutant of HIV-1 protease explains how glycine-48 replacements produce mutants resistant to inhibitor drugs. FEBS Letters *420*, 11–16.

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., Yang, L., Ding, C., Zhu, X., Lv, R., et al. (2018). Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerging Microbes & Infections *7*, 154.

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

Julias, J.G., Boyer, P.L., McWilliams, M.J., Alvord, W.G., and Hughes, S.H. (2004). Mutations at position 184 of human immunodeficiency virus type-1 reverse transcriptase affect virus titer and viral DNA synthesis. Virology *322*, 13–21.

Kurosaki, Y., Ueda, M.T., Nakano, Y., Yasuda, J., Koyanagi, Y., Sato, K., and Nakagawa, S. (2018). Different effects of two mutations on the infectivity of Ebola virus glycoprotein in nine mammalian species. Journal of General Virology *99*, 181–186.

Lau, S.K.P., Feng, Y., Chen, H., Luk, H.K.H., Yang, W.-H., Li, K.S.M., Zhang, Y.-Z., Huang, Y., Song, Z.-Z., Chow, W.-N., et al. (2015). Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. Journal of Virology *89*, 10532 LP – 10547.

Letunic, I., and Peer, B. (2007). Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. Annual Review of Virology *3*, 237–261.

Marquez-Jurado, S., Nogales, A., Avila-Perez, G., Iborra, F.J., Martinez-Sobrido, L., and Almazan, F. (2018). An alanine-to-valine substitution in the residue 175 of Zika virus NS2A

protein affects viral RNA synthesis and attenuates the virus in vivo. Viruses *10*.

Rolling, T., Koerner, I., Zimmermann, P., Holz, K., Haller, O., Staeheli, P., and Kochs, G. (2009). Adaptive mutations resulting in enhanced polymerase activity contribute to high virulence of influenza A virus in mice. Journal of Virology *83*, 6673–6680.

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., Zhu, H., Zhao, W., Han, Y., and Qin, C. (2019). From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses *11*.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) *30*, 1312–1313.

Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., and Xia, X.-Q. (2015). Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Scientific Reports *5*, 17155.

Tao, Y., and Tong, S. (2019). Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. Microbiology Resource Announcements *8*.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell.

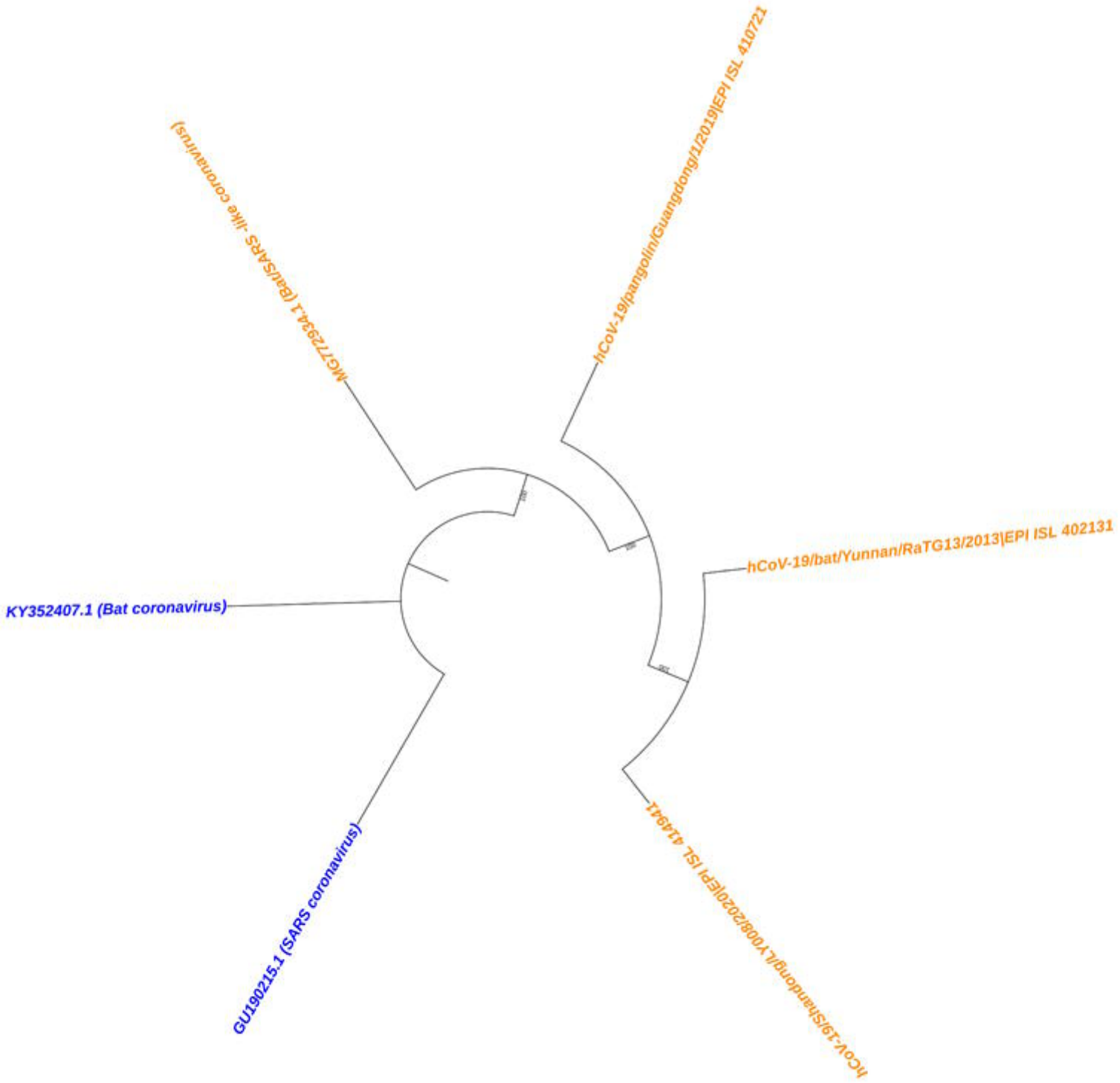
Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

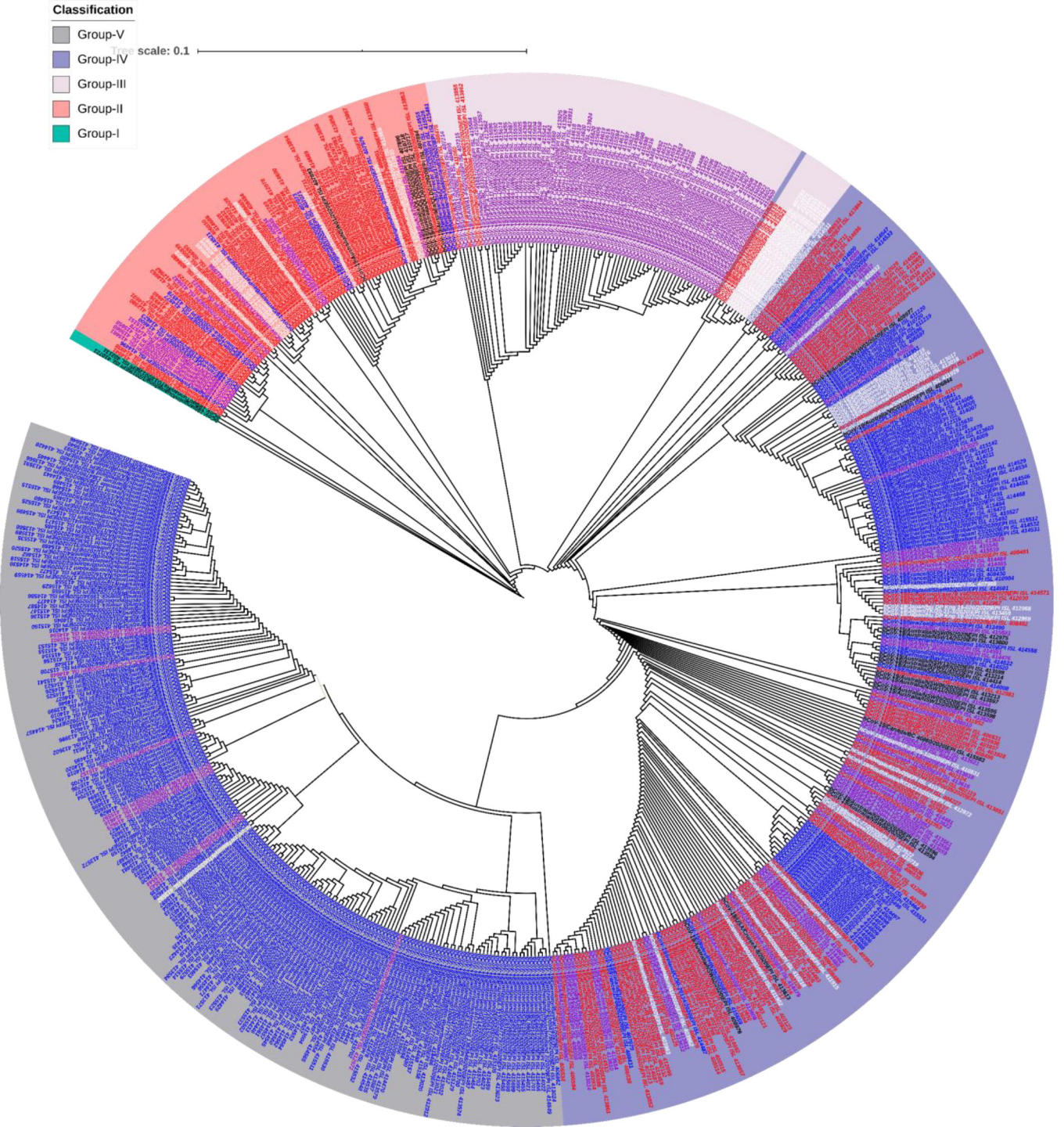
Wolf, D.G., Smith, I.L., Lee, D.J., Freeman, W.R., Flores-Aguilar, M., and Spector, S.A. (1995). Mutations in human cytomegalovirus UL97 gene confer clinical resistance to ganciclovir and can be detected directly in patient plasma. Journal of Clinical Investigation *95*, 257–263.

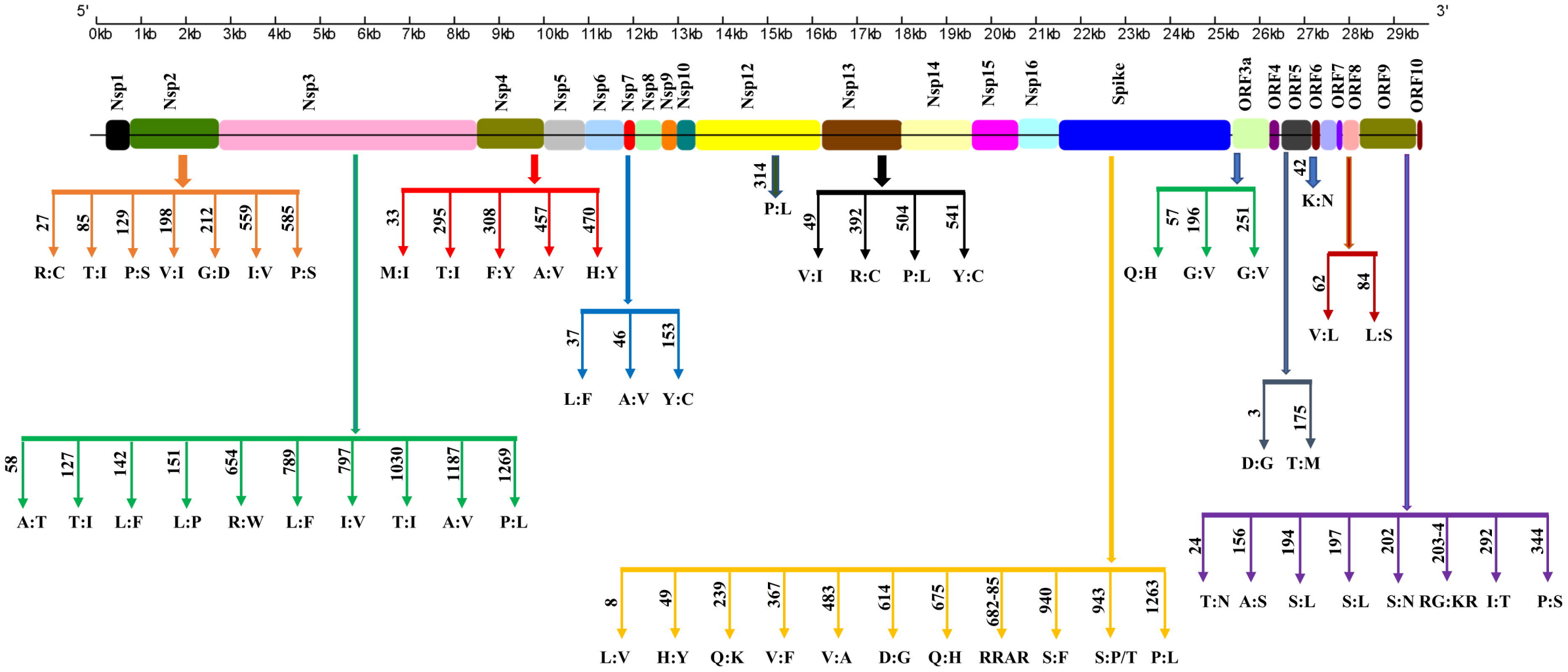
Yamada, S., Hatta, M., Staker, B.L., Watanabe, S., Imai, M., Shinya, K., Sakai-Tagawa, Y., Ito, M., Ozawa, M., Watanabe, T., et al. (2010). Biological and structural characterization of a host-adapting amino acid in Influenza virus. PLOS Pathogens *6*, e1001034.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New England Journal of Medicine *367*, 1814–1820.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine *382*, 727–733.


hCoV-19/pangolin/Guangdong/1/2019|EPI ISL 410721
MG772934.1 (Bat/SARS-like coronavirus)
hCoV-19/bat/Yunnan/RaTG13/2013|EPI ISL 402131
KY352407.1 (Bat coronavirus)
GU190215.1 (SARS coronavirus)
hCoV-19/Shandong/L-Y008/2020|EPI ISL 414941


Classification
Group-V
Group-IV
Group-III
Group-II
Group-I
Tree scale: 0.1


5'
3'
0kb 1kb 2kb 3kb 4kb 5kb 6kb 7kb 8kb 9kb 10kb 11kb 12kb 13kb 14kb 15kb 16kb 17kb 18kb 19kb 20kb 21kb 22kb 23kb 24kb 25kb 26kb 27kb 28kb 29kb
Nsp1
Nsp2
Nsp3
Nsp4
Nsp5
Nsp6
Nsp7
Nsp8
Nsp9
Nsp10
Nsp12
Nsp13
Nsp14
Nsp15
Nsp16
Spike
ORF3a
ORF4
ORF5
ORF6
ORF7
ORF8
ORF9
ORF10
27 85 129 198 212 559 585
R:C T:I P:S V:I G:D I:V P:S
33 295 308 457 470
M:I T:I F:Y A:V H:Y
37 46 153
L:F A:V Y:C
314
P:L
49 392 504 541
V:I R:C P:L Y:C
57 196 251
Q:H G:V G:V
42
K:N
62 84
V:L L:S
3 175
D:G T:M
58 127 142 151 654 789 797 1030 1187 1269
A:T T:I L:F L:P R:W L:F I:V T:I A:V P:L
8 49 239 367 483 614 675 682-85 940 943 1263
L:V H:Y Q:K V:F V:A D:G Q:H RRAR S:F S:P/T P:L
24 156 194 197 202 203-4 292 344
T:N A:S S:L S:L S:N RG:KR I:T P:S