1 **Genomic and protein structure modelling analysis depicts the origin and infectivity of**

2 **2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China**

3

4 Ning Dong[1#], Xuemei Yang[1#], Lianwei Ye[1#], Kaichao Chen[1#], Edward Wai-Chi Chan[2], Mengsu

5 Yang[3], Sheng Chen [1*]

6

7 [1]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary

8 Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

9

10 [2]State Key Lab of Chirosciences, Department of Applied Biology and Chemical Technology,

11 The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong;

12

13 [3]Department of Biomedical Science, Jockey Club College of Veterinary Medicine and Life

14 Sciences, City University of Hong Kong, Kowloon, Hong Kong

15

16 #contribute equally to the work.

17

18 *Corresponding author: Sheng Chen, City University of Hong Kong, Kowloon, Hong Kong;

19 Email: shechen@cityu.edu.hk

20

21

22 **Keywords:** 2019-nCoV, Genomics, Protein modelling, Origin, Infectivity, Wuhan

23   **Abstract**

24   Detailed genomic and structure-based analysis of a new coronavirus, namely 2019-nCoV,

25   showed that the new virus is a new type of bat coronavirus and is genetically fairly distant from

26   the human SARS coronavirus. Structure analysis of the spike (S) protein of this new virus

27   showed that its S protein only binds weakly to the ACE2 receptor on human cells whereas the

28   human SARS coronavirus exhibits strongly affinity to the ACE receptor. These findings suggest

29   that the new virus does not readily transmit between humans and should theoretically not able to

30   cause very serious human infection. These data are important to guide design of infection control

31   policy and inform the public on the nature of threat imposed by 2019-nCov when results of direct

32   laboratory tests on this virus are not expected to be available in the near future.

33   A cluster of pneumonia cases of unknown cause were reported in Wuhan, the capital City of

34   Hubei Province of China in December 2019. As of 18th January 2020, a total of 44 such cases

35   were documented, among whom two patients have died, five in critical condition, and six have

36   been discharged from hospital. Most patients had visited or worked in a seafood wholesale

37   market in Wuhan. An exception being one woman who had not been to the market but has close

38   contact with her husband, one of the patients who worked in the market before falling sick,

39   suggesting the possibility that the agent can at least undergo a limited degree of human to human

40   transmission. Apart from this couple, there is no strong evidence which suggests that the

41   unknown agent is highly infectious as no health care personnel in the hospitals where the patients

42   were admitted were infected. Three suspicious cases were reported outside China, two in

43   Thailand and one in Japan. These three patients were known to originated from or have been to

44   Wuhan.

45

46   On 6th January 2020, the Chinese authority released the sequence (accession#: MN908947) of a

47   novel coronavirus, designated as 2019-nCoV, which was isolated from one of the pneumonia

48   patients and confirmed to be the causative agent for this outbreak. Coronaviruses are a large

49   family of viruses, most of which cause mild infections such as the common cold, but some such

50   as the SARS and MERS (Middle East Respiratory Syndrome) viruses cause severe and potential

51   fatal respiratory tract infections. Some coronaviruses are known to be transmitted easily between

52   humans, while others do not. Based on currently available information, the 2019-nCoV virus

53   belongs to a category that can cause severe illness in some patients but does not transmit readily

54   between people. It is necessary to investigate the genetic and functional data of this new virus

55   and compare to other coronaviruses so as to guide future research and design of appropriate

56 infection control policy to prevent widespread dissemination of another potentially deadly

57 coronavirus since the emergence of the SARS and MERS viruses. In this study, we performed

58 in-depth genetic analysis of 2019-nCoV and generated data which provide timely and valuable

59 insight into the potential origin of this virus, its ability to cause human infection, and its genetic

60 relatedness with SARS and MERS.

61

62 Phylogenetic analysis of genomic sequences of coronaviruses deposited in the GenBank revealed

63 that 2019-nCoV belonged to betacoronavirus and exhibited the closest linkage with two SARS-

64 like coronavirus from bat (bat-SL-CoVZX45 and bat-SL-CoVZX21) (**Fig 1a**). According to the

65 phylogenetic tree, human SARS viruses were closest to bat SARS-like viruses but with a lesser

66 degree to bat coronaviruses, and was least related to other coronaviruses. The 2019-nCoV stands

67 in a position between bat SARS-like viruses and bat coronaviruses, suggesting it is less related to

68 the human SARS virus than other bat SARS-like viruses, and is likely a new type of bat

69 coronavirus. Nevertheless, all coronaviruses that exhibit close linkage with 2019-nCoV

70 originated from bat, strongly suggesting that this new coronavirus originated from bat (**Fig 1a**).

71 Coronaviruses of other species including the murine coronavirus are genetically distant from this

72 new coronavirus, indicating that 2019-nCoV did not originate from other animal hosts. As bats

73 are not sold in the Wuhan market, animals that serve as the transmission vehicle remains to be

74 identified (*1, 2*).

75

76 The sequence of 2019-nCoV was annotated and aligned with several representative

77 coronaviruses selected according to the degree of genetic relatedness depicted by the

78 phylogenetic tree (**Fig 1b**). These included two highly homologous human SARS coronavirus:

79   SARS CoV P2 (FJ882963) and SARS CoV ZJ02 (EU371559), one bat SARS virus that exhibits

80   high homology with human SARS virus and similar potential to infect human as human SARS

81   coronavirus: bat SARS CoV W1V1 (KF367457) (*3*), two bat SARS-like viruses that were not

82   able to infect human: (bat-SL-CoVZX45 and Rp Shaanxi2011), and two un-related

83   coronaviruses, the MERS virus MERS CoV (NC019843) and the Avian Infectious Bronchitis

84   (IBV) virus IBV CoV (AY646283). The new 2019-nCoV was annotated slightly different from

85   the human SARS virus and other coronaviruses, but the functionally important ORFs, ORF1a

86   and ORF1b, and major structural proteins including the spike (S), membrane (M) and envelop (E)

87   and nucleic capsid (N) proteins are well annotated (**Fig 1b**). Consistent with the phylogenetic

88   tree data, 2019-nCoV did not align well with the MERS and IBV virus (**Fig S1**). Among the

89   SARS viruses, bat coronaviruses and 2019-nCoV, the non-structural proteins generally aligned

90   well but variations were observable in major structural proteins and some small ORFs (**Fig 1b**).

91   Detailed sequence alignment showed that 2019-nCoV exhibited significant sequence variation at

92   several regions with the human SARS coronavirus including the N-terminal region of ORF1a

93   and S protein, ORF3, E, ORF6, 7 and 8, and the middle part of N protein. Bat SARS-like virus

94   W1V1 aligned well with human SARS virus P2, with some variations at ORF8 and an insertion

95   between ORF6 and 7. Coronavirus 2019-nCoV aligned best with the bat SARS-like virus bat-

96   SL-CoVZX45, with the majority of genetic variations being seen at the N-terminal part of S

97   protein. The bat-SL-CoVZX45 virus itself exhibited a high degree of variation with another bat

98   SARS-like CoV Rp Shaanxi2011 at ORF1a, the N-terminus of S protein and other structural

99   proteins. However, bat SARS-like CoV Rp Shaanxi2011 exhibited high homology with human

100   SARS virus ZJ02, with variation being seen at the N-terminus of S protein and the middle part of

101   N protein (**Fig 1b**). To check if 2019-nCoV is more close to bat SARS like virus or bat

102 coronavirus, we selected two adjacent viruses from each group, bat SARSCoV Rf1/2004 and bat

103 coronavirus BM48-31/BGR/2008, to perform alignment with the new virus with result showing

104 that 2019-nCoV showed similar diversity from these two strains (**Fig S1b**). When 2019-nCoV

105 was aligned with bat CoV HKU9-1, another bat coronavirus with further distance, it showed that

106 2019-nCoV was very different from this virus (**Fig S1c**). These sequence alignment data were

107 consistent with results of phylogenetic tree analysis and indicated that 2019-nCoV exhibited

108 stands in between bat SARS like viruses and bat coronaviruses, but is genetically distant from

109 the human SARS virus. It should be considered a new type of bat coronavirus.

110

111 Since the S protein is the protein that exhibits the highest degree of genetic variations among

112 different coronaviruses, we performed phylogenetic analysis of the S protein of different

113 coronaviruses (**Fig S2**). Our data showed that the S protein of 2019-nCoV exhibited high

114 homology with bat SARS-like coronaviruses such as bat-SL-CoVZX45 and bat-SL-CoVZX21,

115 human SARS virus and bat coronaviruses. Homology of S protein of 2019-nCoV with

116 representative human SARS virus, bat SARS like viruses and bat coronaviruses was determined

117 as shown in Table S2. S protein of 2019-nCoV showed about 76% homology to human SARS

118 virus P2 and high homology to bat SARS like viruses, while it showed 72% homology to closest

119 bat coronavirus, bat coronavirus BM48-31, even lower homology with other bat coronaviruses.

120 These data further suggested that 2019-nCoV is more likely a new type of bat coronavirus with

121 only loose linkage with the SARS virus. Interestingly, different regions of the S proteins

122 exhibited different levels of homology among the known coronaviruses (**Fig 1b**). Amino acid

123 sequence alignment showed consistently that the N-terminal regions were far more diverse than

124 the C-terminus, which seemed to be highly conservative (**Fig S3a**). Aligning these regions to the

125   structure of S protein indicated that the structurally conserved C-terminal aligned well to the

126   transmembrane domain which consists of a double helix, whereas the most variable region

127   aligned to the N-terminal domain; on the other hand, the receptor binding domain exhibited

128   intermediate level of sequence variation (**Fig S3b**).

129

130   Due to the high amino acid sequence homology of the S protein in 2019-nCoV and the SARS

131   virus which can cause severe human infection, we analyzed the structural similarity of this

132   protein in various viruses. Protein structure modeling was performed to obtain high quality

133   structure of S proteins from different coronaviruses (**Table S1**). The high level similarity

134   observable between the structures of S protein from different viruses implied that the S protein of

135   2019-nCoV and bat coronaviruses would most likely use the same human cell receptor as SARS

136   virus. It was shown that angiotensin-converting enzyme 2 (ACE2) was the cellular receptor of

137   the SARS virus S protein (*4*). Complex structure of ACE2 with the receptor binding domain

138   (RBD) of S protein of SARS virus has been resolved and demonstrated tight interaction between

139   these two proteins at the interaction interface (*5*). These data prompted us to determine the level

140   of interaction between the S protein of 2019-nCoV with its potential cellular receptor ACE2.

141   Using modeled S protein structure and by further performing structure-based alignment, we

142   obtained the complex structure of RBD of the S protein of 2019-nCoV and several bat

143   coronaviruses with human ACE2, using the complex structure of RBS of S protein from human

144   SARS virus and ACE2 (2ajf) as reference (*5*). Structural analysis of potential interactions

145   between RBD of S protein from human SARS virus and ACE2 protein depicted several

146   interaction points including four hydrophobic interactions: ACE2(Y^{41})/ RBD(Y^{484}), ACE2(L^{45})/

147   RBD(Y^{484}), ACE2(L^{79}, M^{82})/ RBD(L^{472}), ACE2(Y^{83})/ RBD(Y^{475}), one salt-bridge: ACE2(E^{329})/

148   RBD(R^{426}) and one cation-π interaction: ACE2(K^{353})/ RBD(Y^{491}) (**Fig 2a, 2e**). However,

149   examination of interaction between RBD of 2019-nCoV and human ACE2 depicted only one

150   potential hydrophobic interaction between ACE2(L^{79}, M^{82}) and RBD(F^{486}) and one cation-π

151   interaction interaction ACE2(K^{353})/ RBD(Y^{492}). Further examination of interactions between

152   RBD from bat SARS like coronaviruses, bat-SL-CoVZX45 and Rp Shaanxi2011 that do not

153   infect human, showed only one cation-π interaction interaction, ACE2(K^{353})/ RBD(Y^{481}).

154   Another bat-originated coronavirus, bat SARS CoV W1V1 that displayed strong binding to

155   ACE2 and exhibited potential to cause human infection was also included for analysis. Binding

156   affinity of RBD from this virus to ACE2 was as tight as that of the human SARS virus, involving

157   four hydrophobic interactions: ACE2(Y^{41})/ RBD(Y^{485}), ACE2(L^{45})/ RBD(Y^{485}), ACE2(L^{79},

158   M^{82})/ RBD(F^{473}), ACE2(Y^{83})/ RBD(Y^{476}), one salt-bridge: ACE2(E^{329})/ RBD(R^{427}) and one

159   cation-π interaction site ACE2(K^{353})/ RBD(Y^{492}). These data suggested that the higher binding

160   affinity of RBD of coronavirus to ACE2 will confer the virus higher infectivity and

161   pathogenicity. The fact that the RBD of 2019-nCoV exhibited much lower affinity to ACE2

162   implies that the virulence potential of 2019-nCoV should be much lower than that of human

163   SARS virus, but is nevertheless stronger than viruses that do not cause human infection; such

164   finding is also consistent with the current epidemiological data in that 2019-nCoV only caused

165   severe pneumonia in patients with weaker immune system such as the elderlies and people with

166   underlying diseases. The weaker binding affinity of 2019-nCoV to human cell might also explain

167   the limited human to human transmission potential of this virus observed to date.

168

169   In this study, we utilized the whole genome sequence of the newly discovered coronavirus, 2019-

170   nCoV, that caused an outbreak of pneumonia in Wuhan, China to perform comparative genetic

8

171   and functional analysis with the human SARS virus and coronaviruses recovered from different

172   animals. Phylogenetic analysis of coronavirus of different species indicated that 2019-nCoV

173   might have originated from bat, but the intermediate transmission vehicle is not known at this

174   stage. Genetic linkage analysis showed that 2019-nCoV lied at the interface between bat SARS

175   like coronavirus and bat coronavirus and should belong to a novel type of bat coronavirus owing

176   to high degree of variation from the human SARS virus. Analysis of the potential interaction of

177   RBD of 2019-nCoV with human ACE2 receptor protein indicated that its affinity to human cell

178   is much lower than that of human SARS virus due to the loss of several important interaction

179   sites, implying that the infectivity and pathogenicity of this new virus should be much lower than

180   the human SARS virus. These data facilitate design of appropriate policies to control further

181   dissemination of this new coronavirus.

182

183 **Acknowledgments**

184 We acknowledge generation and provision of the genome sequence of 2019-nCoV (deposited in

185 GenBank with the accession number MN908947) by Professor Yong-Zhen Zhang from The

186 Shanghai Public Health Clinical Center & School of Public Health, and scientists from the

187 partner institutions including the Central Hospital of Wuhan, Huazhong University of Science

188 and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for

189 Communicable Disease Control and Prevention, Chinese Center for Disease Control, and the

190 University of Sydney, Sydney, Australia. We also appreciate contribution by authors who

191 deposited the coronavirus sequences that we have used in this study in GenBank. This study has

192 no intention to claim any credit of the whole genome sequence of 2019-nCoV and was designed

193 with the only purpose of providing insight into the genetic and functional characteristics of this

194 newly identified coronavirus through the utilization and analysis of sequences generated by the

195 above acknowledged person and institutions.

196

197 **Conflicts of interest:** We declare that we have no conflicts of interest.

198

199

10

Reference:

1. S. K. Lau *et al.*, Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A* **102**, 14040-14045 (2005).
2. K. K. To, I. F. Hung, J. F. Chan, K. Y. Yuen, From SARS coronavirus to novel animal and human coronaviruses. *J Thorac Dis* **5 Suppl 2**, S103-108 (2013).
3. X. Y. Ge *et al.*, Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535-538 (2013).
4. W. Li *et al.*, Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450-454 (2003).
5. F. Li, W. Li, M. Farzan, S. C. Harrison, Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864-1868 (2005).
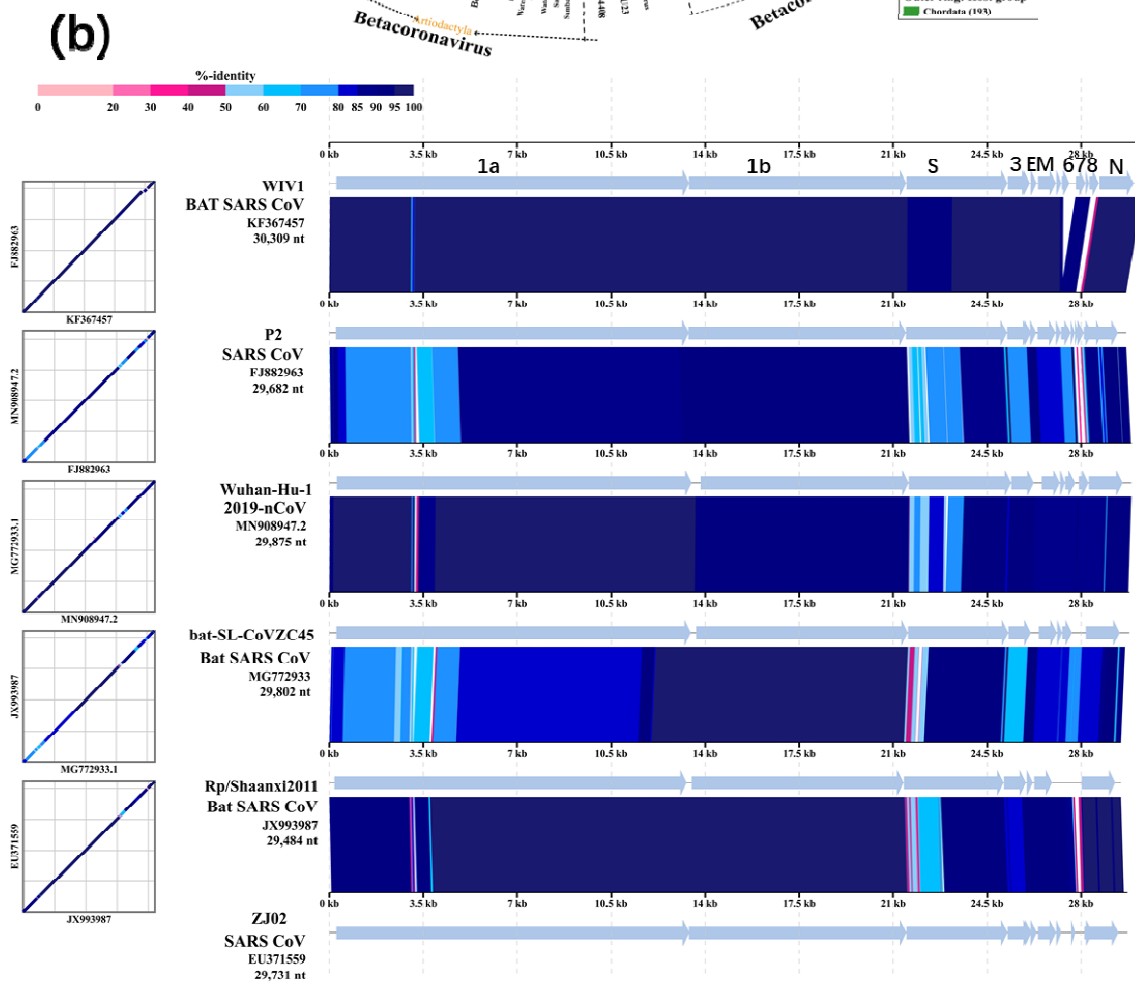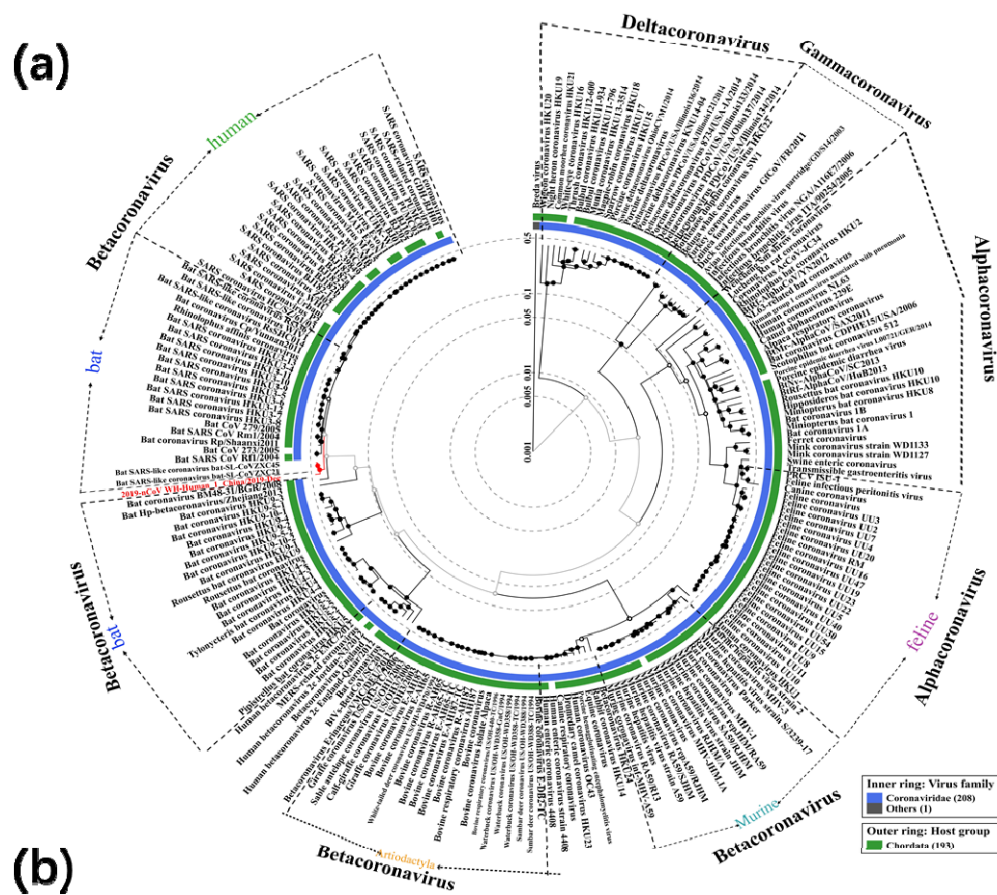
(a)

(b)

212

213 **Figure 1. Phylogenetic analysis and sequence alignment of coronoviruses of different**

214 **species.** (a) Phylogenetic tree of coronaviruses from different species. The type of coronovirus

215 and the host were labelled. Virus labeled with red is the newly discovered coronovirus 2019-

216 nCoV. (b) Sequence alignment of representative the new 2019-nCoV, bat SARS like

217 coronoviruses and bat coronoviruses. These included two highly homologous human SARS

218 coronaviruses: SARS CoV P2 (FJ882963) and SARS CoV ZJ02 (EU371559), one bat SARS

219 virus showing high homology with human SARS virus and similar potential to infect human as

220 human SARS coronavirus: bat SARS CoV W1V1 (KF367457), two bat SARS-like viruses that

221 are not able to infect human: (bat-SL-CoVZX45 and Rp Shaanxi2011) and the newly discovered

222 2019-nCoV from Wuhan.

13

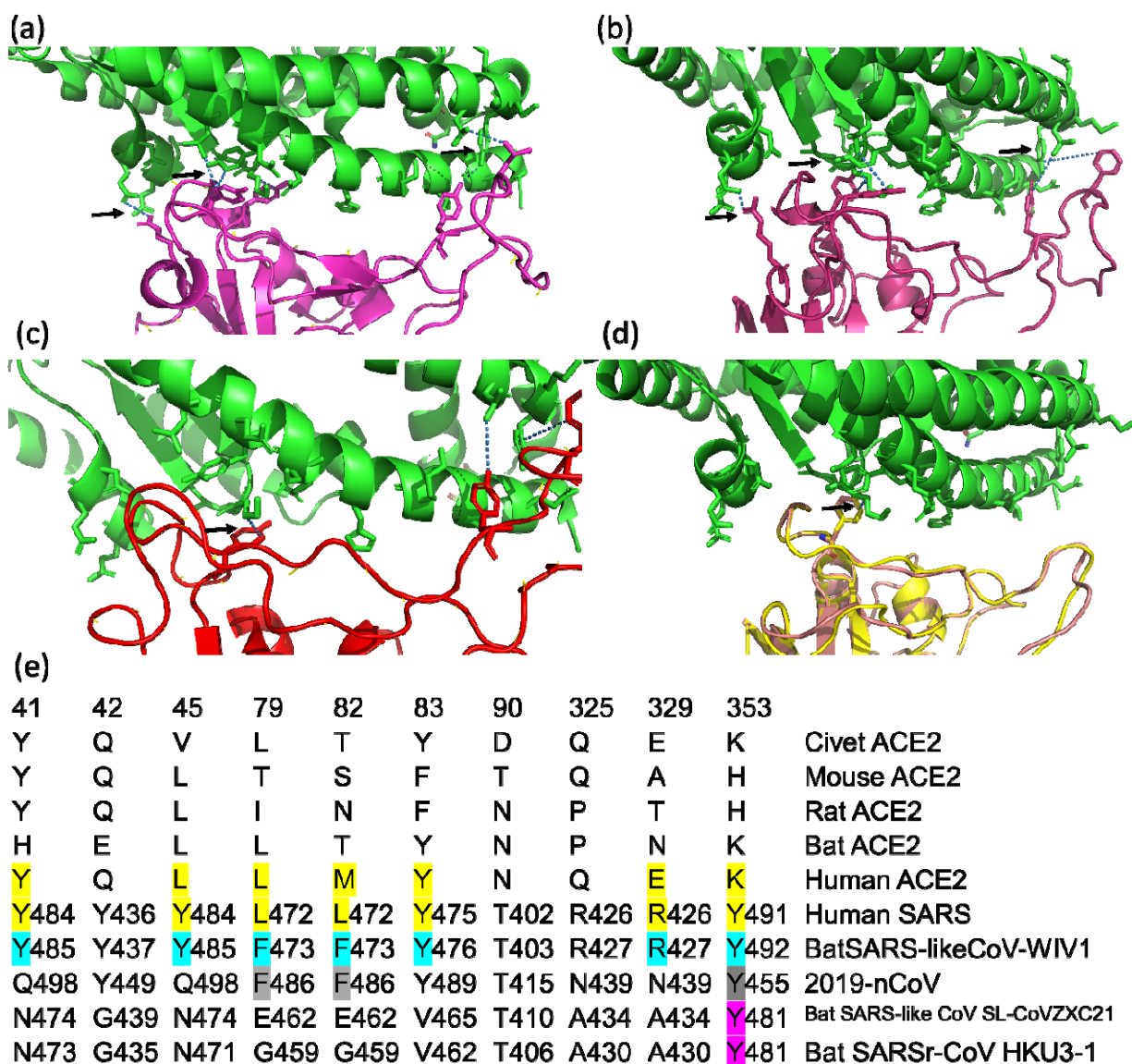| 41	42	45	79	82	83	90	325	329	353	
Y	Q	V	L	T	Y	D	Q	E	K	Civet ACE2
Y	Q	L	T	S	F	T	Q	A	H	Mouse ACE2
Y	Q	L	I	N	F	N	P	T	H	Rat ACE2
H	E	L	L	T	Y	N	P	N	K	Bat ACE2
Y	Q	L	L	M	Y	N	Q	E	K	Human ACE2
Y484	Y436	Y484	L472	L472	Y475	T402	R426	R426	Y491	Human SARS
Y485	Y437	Y485	F473	F473	Y476	T403	R427	R427	Y492	BatSARS-likeCoV-WIV1
Q498	Y449	Q498	F486	F486	Y489	T415	N439	N439	Y455	2019-nCoV
N474	G439	N474	E462	E462	V465	T410	A434	A434	Y481	Bat SARS-like CoV SL-CoVZXC21
N473	G435	N471	G459	G459	V462	T406	A430	A430	Y481	Bat SARSr-CoV HKU3-1

Figure 2. Potential interactions between receptor binding domain (RBD) of S proteins from different coronaviruses and the human cell receptor ACE2. Interactions between ACE2 with RBD of human SARS virus (a), highly similar bat SARS like coronavirus, CoV-W1V1 that can infect human (b), new type of coronavirus, 2019-nCoV (c), bat SARS like coronavirus CoV SL-CoVZXC21 and bat coronavirus HKU3-1 (d) are shown. The detailed amino acid interaction sites between these two proteins are shown in (e). Arrows showed the areas with interacted residues from both proteins. Amino acid highlighted with different colors indicated the potential interaction residues between different proteins, which was highlighted with different colors.