1   **Extensive recombination-driven coronavirus diversification expands the pool of**

2   **potential pandemic pathogens**

3

4   Stephen A. Goldstein[1*†], Joe Brown[*1], Brent S. Pedersen[1], Aaron R. Quinlan[1], Nels C. Elde[1†]

5

6   [1]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA.

7   [*]Contributed equally.

8   [†]Corresponding authors.

9   Email: sgoldstein@genetics.utah.edu (S.A.G)

10   nelde@genetics.utah.edu (N.C.E)

11

12 **Abstract**

13 The ongoing SARS-CoV-2 pandemic is the third zoonotic coronavirus identified in the last

14 twenty years. Enzootic and epizootic coronaviruses of diverse lineages also pose a significant

15 threat to livestock, as most recently observed for virulent strains of porcine epidemic diarrhea

16 virus (PEDV) and swine acute diarrhea-associated coronavirus (SADS-CoV). Unique to RNA

17 viruses, coronaviruses encode a proofreading exonuclease (ExoN) that lowers point mutation

18 rates to increase the viability of large RNA virus genomes, which comes with the cost of limiting

19 virus adaptation via point mutation. This limitation can be overcome by high rates of

20 recombination that facilitate rapid increases in genetic diversification. To compare dynamics of

21 recombination between related sequences, we developed an open-source computational

22 workflow (IDPlot) to measure nucleotide identity, locate recombination breakpoints, and infer

23 phylogenetic relationships. We analyzed recombination dynamics among three groups of

24 coronaviruses with noteworthy impacts on human health and agriculture: *SARSr-CoV*,

25 *Betacoronavirus-1*, and SADSr-CoV. We found that all three groups undergo recombination with

26 highly diverged viruses from sparsely sampled or undescribed lineages, which can disrupt the

27 inference of phylogenetic relationships. In most cases, no parental origin of recombinant regions

28 could be found in genetic databases, suggesting that much coronavirus diversity remains

29 unknown. These patterns of recombination expand the genetic pool that may contribute to future

30 zoonotic events. Our results also illustrate the limitations of current sampling approaches for

31 anticipating zoonotic threats to human and animal health.

32

33

Introduction

34

35   In the 21st century alone three zoonotic coronaviruses have caused widespread human

36 infection: SARS-CoV in 2002 [1, 2], MERS-CoV in 2012 [2], and SARS-CoV-2 in 2019 [3]. Four

37 other coronaviruses, OC43, 229E, NL63, and HKU1 are endemic in humans and cause mild-to-

38 moderate respiratory disease with low fatality rates, though they may cause outbreaks of severe

39 disease in vulnerable populations [4-7]. Like SARS-CoV-2, SARS-CoV, and MERS-CoV, these

40 endemic viruses emerged from animal reservoirs. The origins of 229E and NL63 have been

41 convincingly linked to bats, much like the 21st century novel coronaviruses [8-10]. In a striking

42 parallel, both MERS-CoV and 229E appear to have emerged from bats into camelids,

43 established a new persistent reservoir, and then spilled over into humans [11-14]. In contrast,

44 the viral lineages that include OC43 and HKU1 originated in rodents [15,16], though very limited

45 rodent sampling leaves us with a poor understanding of the deep evolutionary history of these

46 viruses. Given the short infectious period of human coronavirus infections, the establishment of

47 endemicity was likely preceded by a period of intense and widespread transmission on regional

48 or global scales. In other words, SARS-CoV-2 is likely the fifth widespread coronavirus epidemic

49 or pandemic involving a still-circulating virus, though the severity of the previous four cannot be

50 reliably ascertained.

51   Livestock are similarly impacted by spillover of coronaviruses from wildlife reservoirs.

52 Three viruses closely related to OC43, bovine coronavirus (BCoV), equine coronavirus (ECoV)

53 and porcine hemagglutinating encephalomyelitis virus (PHEV) are enzootic or epizootic in cows,

54 horses, and pigs respectively [17]. Since 2017, newly emerged swine acute diarrhea syndrome-

55 associated coronavirus (SADS-CoV) has caused significant mortality of piglets over the course

56 of several outbreaks [18,19]. Sampling of bats proximal to impacted farms determined that

57 SADS-CoV outbreaks are independent spillover events of SADSr(elated)-CoVs circulating in

58 horseshoe bats [20]. Molecular studies of SADS-CoV have identified the potential for further

59 cross-species transmission, including the ability to infect primary human airway and intestinal

60 cells [21,22].

61   Emergence of novel viruses requires access to new hosts, often via ecological

62 disruption, and the ability to efficiently infect these hosts, frequently driven by adaptive

63 evolution. Uniquely among RNA viruses, coronavirus genomes encode a proofreading

64 exonuclease that results in a significantly lower mutation rate for coronaviruses compared to

65 other RNA viruses [23,24]. This mutational constraint is necessary for maintaining the stability of

66 the large (27-32 kb) RNA genome but limits the evolution of coronaviruses via point mutation.

67 The high recombination rate of coronaviruses compensates for the adaptive constraints

68   imposed by high-fidelity genome replication [24,25]. The spike glycoprotein in particular has

69   previously been identified as a recombination hotspot [26]. Acquisition of new spikes may

70   broaden or alter receptor usage, enabling host-switches or expansion of host range.

71   Additionally, it may result in evasion of population immunity within established host species,

72   effectively expanding the pool of susceptible individuals. Recombination in other regions of the

73   genome is less well-documented but may also influence host range, virulence, and tissue

74   tropism, and likely contributed to the emergence of SARS-CoV [27,28].

75   To study the dynamics of recombination among clinically significant coronavirus lineages

76   we developed a novel web-based software, IDPlot, that incorporates multiple analysis steps into

77   a single user-friendly workflow. Analyses performed by IDPlot include multiple sequence

78   alignment, nucleotide similarity analysis, and tree-based breakpoint prediction using the GARD

79   algorithm from the HyPhy genetic analysis suite [29]. IDPlot also allows the direct export of

80   sequence regions to NCBI Blast to ease identification of closest relatives to recombinant regions

81   of interest.

82   Using IDPlot, we analyzed recombination events in three clinically significant lineages of

83   coronaviruses with sufficient samplings to conduct robust analyses: SARS-CoV-2-like viruses,

84   OC43-like viruses (*Betacoronavirus-1*) in the *Betacoronavirus* genus, and the SADSr-CoV group

85   of alphacoronaviruses. In all three groups, we found clear evidence of recombination resulting in

86   viruses with high overall nucleotide identity but exhibiting substantial genetic divergence in

87   discrete genomic regions. Recombination was particularly enriched around and within the spike

88   gene and 3' accessory genes. Within all three groups, recombination has occurred with

89   undescribed lineages, indicating that coronavirus diversity, even within these consequential sub-

90   groups of viruses, remains considerably undersampled. The potential for viruses to rapidly

91   acquire novel phenotypes through such recombination events underscores the importance of a

92   more robust and coordinated ecological, public health, and research response to the ongoing

93   pandemic threat of coronaviruses.

94

95   **Results**

96

97   ***Coronavirus phylogenetic relatedness is variable across genomes***

98   Coronavirus genomes, at 27-32 kilobases (kb) in length, are among the largest known

99   RNA genomes, surpassed only by invertebrate viruses in the same *Nidovirales* order [30,31].

100   The 5' ~20 kb of the genome comprises open reading frames 1a and 1b, which are translated

101   directly from the genome as polyproteins pp1a and pp1ab and proteolytically cleaved into

102 constituent proteins (**Figure 1A**) [32]. Orf1ab is among the most conserved genes and encodes

103 proteins essential for replication, including the RNA-dependent RNA-polymerase (RdRp), 3C-

104 like protease (3ClPro), helicase, and methyltransferase. Given the high degree of conservation

105 in this region, coronavirus species classification is typically determined by the relatedness of

106 these key protein-coding regions [33]. The 3' ~10 kb of the genome contains structural genes

107 including those encoding the spike and the nucleocapsid proteins, as well as numbered
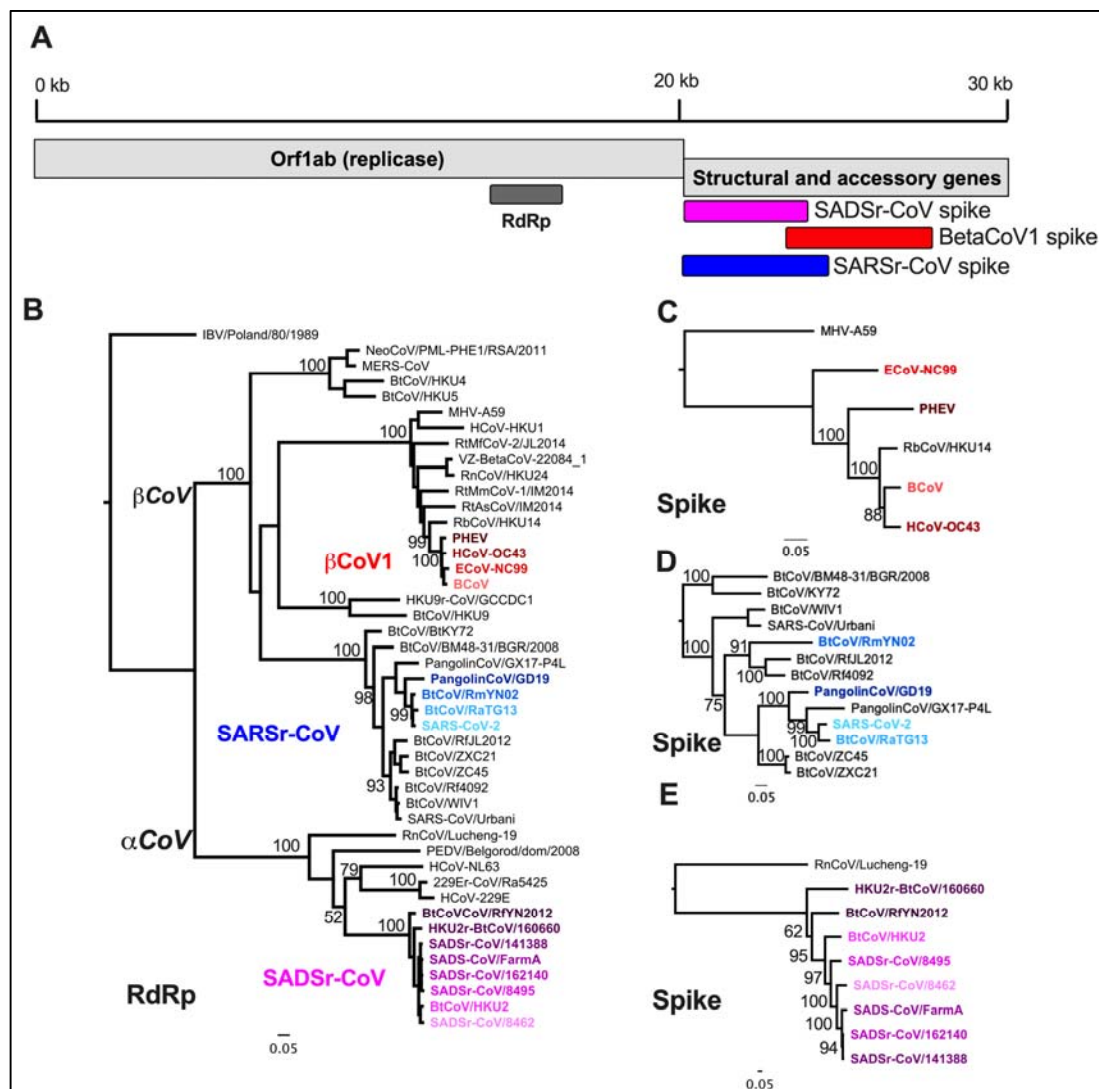

Figure 1. AlphaCoV and BetaCoV phylogenetic relationships are genome region-dependent. A) Basic coronavirus genome organization with the 5' ~20 kb comprising the replicase gene that is proteolytically processed into up to 16 individual proteins. The 3' 10 kb comprises structural and genus-specific accessory genes. B) Maximum-likelihood (ML) phylogenetic tree of alpha and betaCoVs full-length RNA-dependent RNA-polymerase encoding region of Orf1ab. C) ML phylogenetic tree of full-length spike genes from SADS-related CoVs (magenta), rooted with the distantly related alphacoronavirus HCoV-229E D) ML phylogenetic tree of spike genes from viruses in the species *Betacoronavirus 1* (red) rooted with the distantly related betacoronavirus mouse hepatitis virus. E) ML phylogenetic tree of spike genes of SARSr-CoVs, with SARS-CoV-2-like viruses further analyzed in the paper highlighted in blue.

108   accessory genes that are unique to coronavirus genera and subgenera [34]. In contrast to the

109   relative stability of the replicase region of the genome, the structural and accessory region, and

110   in particular the spike glycoprotein, have been identified as recombination hotspots [26].

111   We set out to characterize the role of recombination in generating diversity across the

112   coronavirus phylogeny. A classic signature of recombination is differing topology and/or branch

113   lengths of phylogenetic trees depending on what genomic regions are analyzed. To identify

114   lineages of interest for recombination analysis, we built a maximum-likelihood phylogenetic tree

115   of full-length RdRp-encoding regions of representative alpha and betacoronaviruses, which

116   contain all human and most mammalian coronaviruses (**Figure 1B**). To further test whether

117   comparisons of RdRp sequence reflected ancestral relatedness, we conducted the same

118   analysis for the 3ClPro and Helicase-encoding regions of Orf1ab (**Figure S2**). Phylogenetic

119   relationships were generally maintained in these trees and genetic relatedness remains very

120   high (90-99% within groups), which leads to some reshuffling with low bootstrap support. From

121   these trees we chose to further investigate the evolutionary dynamics of three clinically

122   significant groups of coronaviruses: SARS-CoV-2 like viruses (blue) from within *SARSr(elated)-*

123   *CoV*, among which recombination has been reported though not characterized in detail,

124   endemic and enzootic OC43-like viruses of *Betacoronavirus-1* (*BetaCoV1*) (red), and SADSr-

125   CoVs (magenta). Although other coronavirus lineages are of public health interest, such as

126   those including the human coronaviruses HKU1, NL63, and 229E there is a relative paucity of

127   closely related sequences to these viruses, limiting our current ability to analyze these

128   genomes.

129   Within each group there is modest diversity revealed by comparing RdRp sequence: 94-

130   99% nt identity among the SADSr-CoVs, >97% nt identity within *Betacoronavirus-1*, and 91-

131   99% among the SARS-CoV-2-like viruses (**Figure S3**). Similar results were observed for 3ClPro

132   and Helicase-encoding regions (**Figure S4**). In contrast, spike gene phylogenetic trees of each

133   group show greater diversity as illustrated by extended branch lengths and/or changes in tree

134   topology, suggesting either rapid evolution and/or recombination diversifies this region (**Figure**

135   **1B-D**). To analyze these possibilities, we developed a new pipeline to better study these

136   evolutionary patterns.

137

138   ***IDPlot Facilitates Nucleotide Identity and Recombination Analysis***

139   To investigate possible recombination-driven diversity among these viruses we

140   developed IDPlot, which incorporates several distinct analysis steps into a single Nextflow

141   workflow [35] and generates a comprehensive HTML report to facilitate interpretation and

142   downstream analysis. IDPlot combines existing algorithms into a single pipeline and provides a

143   statistically supported means to adjust recombination prediction and phylogenetic analysis in a

144   quickly interpretable visual display.

145   First, IDPlot generates a multiple sequence alignment using MAFFT (**Figure 2A**) [36]

146   with user-assigned reference and query sequences. In its default configuration, IDPlot then

147   generates a sliding window average nucleotide identity (ANI) plot, also displaying the multiple

148   sequence alignment with differences to the reference sequence (colored vertical lines) and gaps

149   (gray boxes) clearly highlighted. The plot is zoomable, and selected sequence regions can be

150   exported directly to NCBI BLAST. Users can also choose to run GARD, the recombination

151   detection program from the HyPhy suite of genomic analysis tools [29]. If GARD is implemented

152   (**Figure 2B**), distinct regions of the multiple sequence alignment are depicted between the

153   alignment and the ANI plot, and phylogenetic trees for each region are generated using

154   FastTree2 (**Figure 2C**) [37] and displayed (**Figure 2E**).

155   A significant barrier to effective use of GARD is that because it ultimately presents
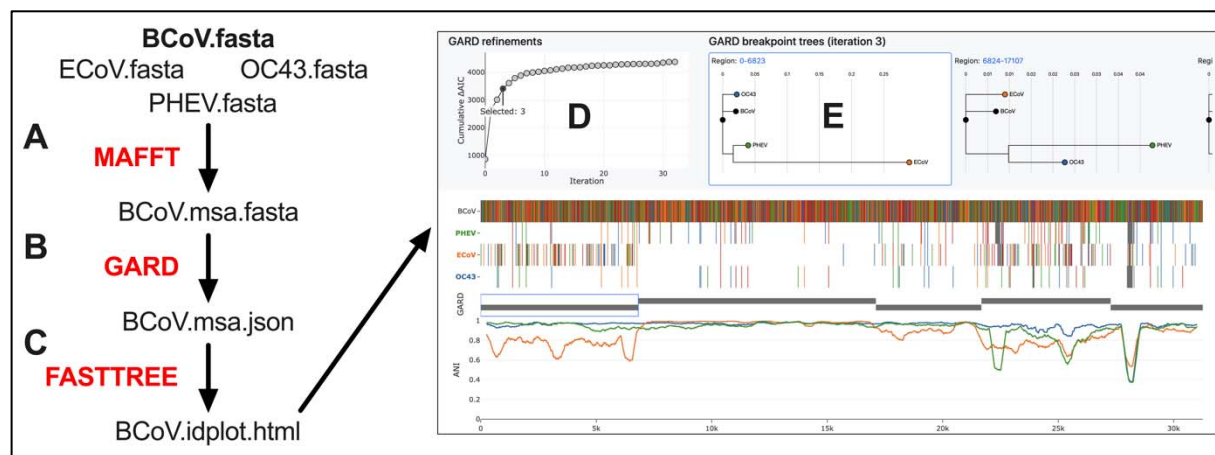

Figure 2. IDPlot workflow. IDPlot workflow. A) Reference and query sequences are aligned using MAFFT. B) Breakpoint detection is performed using GARD, capturing breakpoints across iterative refinements. C) Phylogenetic trees based on breakpoints from each iteration and are created using FastTree 2. D) Improvement in ΔAIC-c is plotted against the iteration E) Phylogenetic trees associated with the selected GARD iteration are displayed

156   multiple (sometimes dozens of) iterations, model choice and therefore the selection of

157   breakpoints for further analysis can be challenging. To alleviate this issue the IDPlot output

158   includes a graph showing a cumulative count of GARD's statistical iterations (Akaike information

159   criterion (AIC-c) on the y-axis) and the GARD model number on the x-axis (**Figure 2D**). GARD

160   uses ΔAIC-c for each proposed model to indicate the degree of fit improvement over the

161   preceding model, and this graph allows the user to easily determine when improvements

162   become increasingly marginal, which is often accompanied by prediction of spurious

163 breakpoints. Upon selection of a GARD iteration, the display switches to show the associated

164 phylogenetic trees (**Figure 2E**). Genomic regions are clickable, immediately bringing the

165 appropriate phylogenetic tree to the center of the display. Finally, the ability to export sequences

166 directly to BLAST enables the user to search for related sequences in GenBank, useful when

167 defined regions are highly divergent from the reference sequence or others included in the data

168 set under analysis.

169

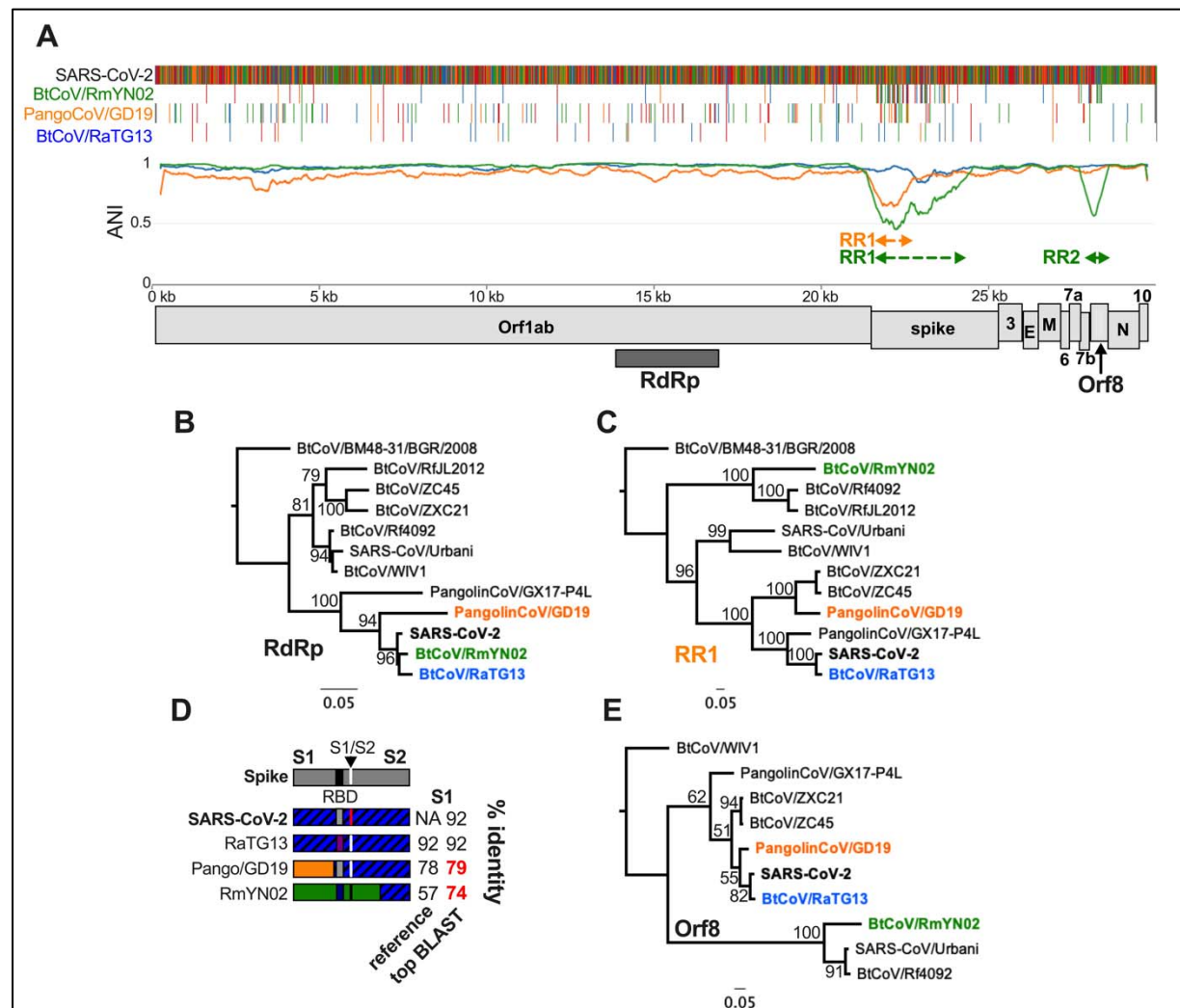170 **SARS-CoV-2-like virus recombination with distant SARSr-CoVs**

171 To test and validate IDPlot as a tool for examining the recombination dynamics of

172 coronaviruses, we initially conducted an analysis of SARS-CoV-2-like viruses within *SARSr-*

173 *CoV.* We chose these viruses as our initial IDPlot case study because recombination has been

174 previously described [38,39], though not characterized in great phylogenetic detail. This

175 provided the opportunity to evaluate IDPlot against a known framework but also advance our

176 understanding of the role recombination has played in the evolution of these clinically significant

177 viruses.

178 Prior to 2019 the SARS-CoV-2 branch within *SARSr-CoV* was known only from a single,

179 partial RdRp sequence published in 2016 [40]. Upon the discovery of SARS-CoV-2 this RdRp

180 sequence was extended to full genome-scale [3] and additional representatives from bats and

181 pangolins have since been identified [39] [41,42]. However, this singularly consequential lineage

182 remains only sparsely sampled and its evolutionary history largely obscured. Most attention on

183 these viruses to date has focused on the recent evolutionary history of SARS-CoV-2 with

184 respect to possible animal reservoirs and recombinant origins. Much less attention has been

185 paid to analyzing the evolution of known close relatives, the bat viruses RaTG13 and RmYN02,

186 and PangolinCoV/GD19.

187 Our IDPlot analysis does not support an emergence of SARS-CoV-2 via recent

188 recombination, consistent with previously published work [3] [38]. RaTG13 shows consistently

189 high identity across the genome with the only notable dip comprising the receptor-binding

190 domain in the C-terminal region of spike S1 (**Figure 3A**), which is proposed to originate via

191 either recombination or diversifying selection [38]. However, the still limited sampling in the

192 SARS-CoV-2-like lineage results in weak phylogenetic signals unable to distinguish between

193 rapid mutational divergence and recombination producing the low ANI in the RaTG13 receptor

194 binding domain.

195 In contrast, PangolinCoV/GD19 and RmYN02 show one and two significant drops in

196 ANI, respectively. Phylogenetic analysis of the PangolinCoV/GD19 recombinant region captures

197   the signal for both that virus (**Figure 3A, 3C, S5C**) and RmYN02 RR1, showing that both

198   viruses fall onto separate branches highly divergent from SARS-CoV-2 and RaTG13 (**Figures**

199   **3C**) with only 81% and 74% nucleotide identity to the closest sequences in GenBank,

200   respectively (**Figure 3D, S5A**). These findings identify three unique spike genes among SARS-

201   CoV-2 and its three closest known relatives (**Figure 3D**), indicative of recombination with

202   *SARSr-CoV* lineages that remain to be discovered despite being the focus of intense virus

203   sampling efforts over the last eighteen years, since the emergence of SARS-CoV.

Figure 3. SARSr-CoV IDPlot Analysis. A) IDPlot analysis of SARS-CoV-2-like SARSr-CoVs with color-coded dashed lines defining divergent regions arising from recombination events with ancestral viruses. B) ML tree of the RdRp-encoding region of SARS-2-like and other SARSr-CoVs showing close relationship between the SARS-CoV-2-like viruses. C) ML tree of PangolinCoV/GD19 RR1 (which overlaps with BtCoV/RmYN02 RR1) showing different topology than the RdRp tree. D) Schematic of spike proteins indicating divergent regions and nucleotide identity to the reference sequence and closest related sequence in GenBank. E) ML tree of ORf8 showing that RmYN02 Orf8 is a divergent member of the SARS-CoV-like Orf8 branch.

204   In addition to spike, RmYN02 contains a second recombinant region that encompasses

205   the 3' end of Orf7b and the large majority of Orf8 (**Figure 3A, S5A**). Orf8 is known to be highly

206 dynamic in SARSr-CoVs. SARS-CoV underwent an attenuating 29 nt deletion in Orf8 in 2002-

207 2003 [43] and Orf8 deletions have been identified in numerous SARS-CoV-2 isolates as well

208 [44-46]. In bat SARSr-CoVs intact Orf8 is typically though not always present but exhibits a high

209 degree of phylogenetic incongruence. Additionally, the progenitor of SARS-CoV encoded an

210 Orf8 gene gained by recombination [28,47]. The BtCoV/RmYN02 Orf8 has only 50% nt identity

211 to SARS-CoV-2 Orf8 and groups as a distantly related member of the branch containing SARS-

212 CoV (**Figure 3E**), exhibiting just 80% nucleotide identity to the closest known sequence.

213 Although the precise function of Orf8 is unknown, there is some evidence that like other

214 accessory proteins it mediates immune evasion [43]. Therefore, recombination in Orf8 has the

215 potential to alter virus-host interactions and may, like spike recombination, impact host range

216 and virulence.

217 This analysis confirmed that IDPlot allows us to characterize recombination events in

218 detail with a single workflow. We demonstrate that multiple SARS-CoV-2-like viruses have

219 recombined with unsampled SARSr-CoV lineages, limiting our ability to assess sources of

220 genetic diversification for these viruses. Under-sampling has implications limiting the

221 incisiveness of both laboratory and field investigations of these viruses.
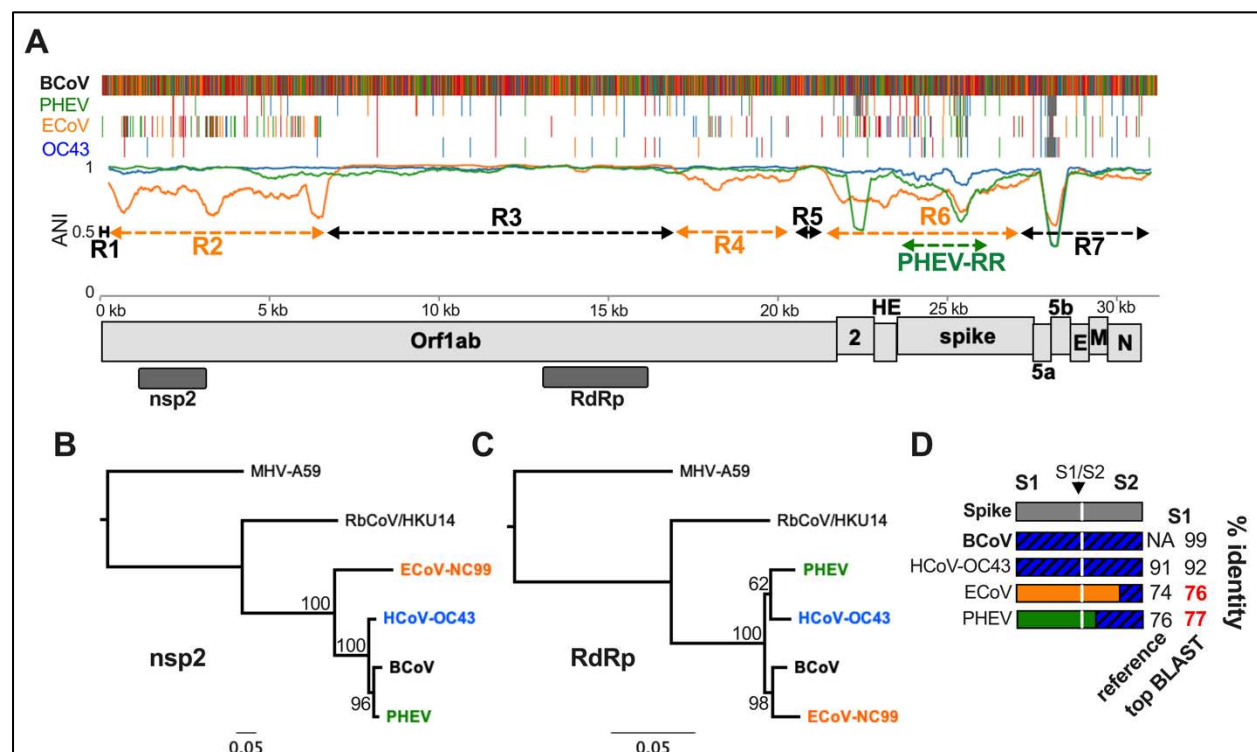
222

223 ***OC43-like viruses encode divergent spikes acquired from unsampled betacoronaviruses***

224 After validating IDPlot for recombination analysis of coronaviruses, we used it to

225 characterize recombination among the viruses in the *Betacoronavirus-1* (*BetaCov1*) group,

226 which includes the human endemic coronavirus OC43 and closely related livestock pathogens

227 bovine coronavirus (BCoV), equine coronavirus (ECoV), porcine hemagglutinating

228 encephalomyelitis virus (PHEV), and Dromedary camel coronavirus HKU23 (HKU23). Due to

229 the apparent low virulence of OC43 and limited sampling of the lineage, these viruses receive

230 relatively little attention outside agricultural research. However, this lineage has produced a

231 highly transmissible human virus, can cause severe disease in vulnerable adults, and is poorly

232 sampled [4]. An ancestral BCoV is believed to be the progenitor of the other currently

233 recognized *BetaCoV1* viruses with divergence dates estimated at 100-150 years ago for

234 OC43/PHEV [48] and 50 years ago for HKU23 [49]. Recombination with other

235 betacoronaviruses has been previously described for HKU23, so we excluded it from our

236 analysis [50]. The most closely related known virus to *BetaCoV1*, rabbit coronavirus HKU14

237 (RbCoV/HKU14) was reported to associate with ECoV in some regions [51], but no detailed

238 recombination analysis of the relationship between these viruses has been previously

239 described.

240   We conducted IDPlot analysis of OC43 and these related enzootic viruses of livestock

241   (**Figure 4A**) and identified at least six major recombination breakpoints in the ECoV genome.

242   The largest divergent region (Region 2) is >6 kilobases (**Figure 4A**). This region encompassing

243   ~20% of the genome exhibits only ~75% nt identity to the reference sequence, just ~81%

244   identity to any known sequence, and occupies a distant phylogenetic position relative to RdRp

245   (**Figure 4B-C, S6A, S6C-D**). In contrast to previous reports that ECoV clusters closely with

246   RbCoV/HKU14 in this region [51], our analysis reveals that this region of ECoV was acquired

247   via recombination from a viral lineage not documented in GenBank.

248   Striking variability in ANI within Region 2 led us to conduct a more detailed analysis.

249   IDPlot did not predict internal Region 2 breakpoints, so we conducted a manual analysis guided

250   by the IDPlot multiple sequence alignment, phylogenetic trees for each proposed sub-region,

251   and BLAST analysis to further dissect differing evolutionary relationships for sub-regions. We

252   found at least six and possibly seven distinct sub-regions (**Figure S7**). Nucleotide identity to top

253   BLAST hits of these sub-regions is highly variable (<70% to >90%), as is identity of the hits

254   themselves, with genetic contribution from RbCoV/HKU14-like viruses, BCoV-like viruses, and

255   more distant, uncharacterized lineages within the *Embecovirus* genus (**Figure S5**). Together,

256   this demonstrates that Region 2 was not acquired via a single recombination event but rather

257   represents a mosaic of known and unknown viral lineages that share an overlapping ecological

258   niche with ancestral ECoV.

259   Another major recombinant ECoV region, Region 6, includes the entire NS2 and HE

260   genes as well as the majority of the spike gene (**Figure 4A, S6A**). Within this region on the

261   multiple sequence alignment, we also identified a recombination event encompassing the

262   majority of the PHEV spike gene, though this required downsampling (removing ECoV) to

263   simplify the analysis (**Figure 4A, S6A**). Both ECoV Region 6 and the PHEV recombinant region

264   occupy relatively distant nodes on a phylogenetic tree (**Figure S6G, J**) and exhibit <80%

265   sequence identity to the reference sequence or any sequence in GenBank (**Figure S6A**),

266   indicating they are derived from independent recombination events or diverged via repeated

267   mutations. Recombination appears most likely given the even dispersion of low identity

Figure 4. Recombination analysis of *Betacoronavirus-1*. A) Nucleotide identity plot and multiple sequence alignment of *BetaCoV-1* viruses. Orange dashed lines indicate divergent regions of the ECoV-NC99 genome while black dashed lines are regions with high identity to the reference sequence bovine coronavirus (BCoV). B) ML tree of nsp2-encoding region of Orf1ab, which falls within the divergent ECoV-NC99 Region 2. C) ML tree of the RdRp-encoding region of Orf1ab. D) Schematic depicting the spike gene diversity of *BetaCoV1* demonstrating the divergence of ECoV-NC99 and PHEV. Top BLAST hits in bolded red indicate no GenBank entries with >80% nucleotide identity.

268   throughout the region, including in portions of the spike S2 domain which is otherwise highly

269   conserved. Additional sampling might more definitively resolve these possibilities. Finally, we

270   identified a third recombinant region, Region 4, in which ECoV exhibited high nucleotide identity

271   with RbCoV/HKU14 (**Figure S4A, E**), further demonstrating the highly mosaic nature of the

272   ECoV genome.

273   Our analysis of equine coronavirus offers a remarkable example of the degree and
274   speed of divergence facilitated by the high recombination rates among coronaviruses. Previous
275   genomic characterization of ECoV suggested that it is the most divergent member of *BetaCoV1*
276   based on nucleotide identity and phylogenetic positioning of full-length Orf1ab. However, in the
277   >10 kilobase Region 3 that accounts for ~1/3 of the entire genome (**Figure 4A**) ECoV exhibits
278   the highest nucleotide identity to BCoV in our dataset (98.5%) (**Figure 4A, S6D**), which is
279   inconsistent with it having diverged earlier than OC43 and PHEV. The latter viruses are
280   estimated to have shared a common ancestor with BCoV 100-150 years ago [48], suggesting
281   that all of the observed ECoV recombination has occurred more recently. Our discovery of
282   recombinant regions of unknown betacoronavirus origin suggest that unsampled, distantly
283   related lineages occupy overlapping ecological niches with ECoV and may continue to circulate
284   and participate in recombination events. Basal members of the subgenus that includes
285   *BetaCoV1* have been identified exclusively in rodents (**Figure 1B**), suggesting they are a
286   natural reservoir for these viruses. Although relatively little attention has been directed to these
287   viruses, studies of BCoV and ECoV cross-neutralization suggest population immunity to OC43
288   may provide only limited protection against infection mediated by these novel spikes [52]. No
289   recent zoonotic infections from this lineage have been documented, but the genomic collision of
290   these viruses with yet-undiscovered, presumably rodent viruses warrants a reassessment of
291   their potential threat to human health.
292
293   **SADSr-CoVs encode highly diverse spike and accessory genes**
294   In 2017 a series of highly lethal diarrheal disease outbreaks on Chinese pig farms were
295   linked to a novel alphacoronavirus, swine acute diarrhea syndrome-associated coronavirus
296   (SADS-CoV) [20,53], which is closely related to the previously described BtCoV/HKU2 [54].
297   Sampling of horseshoe bats nearby affected farms revealed numerous SADSr-CoVs with >95%
298   genome-wide nucleotide identity, suggesting porcine outbreaks were due to spillover from local
299   bat populations. To gain a better view of the genetic diversity among these viruses, we
300   conducted IDPlot analysis of a prototypical SADS-CoV isolate (FarmA) and seven bat SADSr-
301   CoVs sampled at different times before and after the first outbreaks in livestock (**Figure 5A**)
302   using bat SADSr-CoV/162140 as a reference sequence. Three notable observations emerged
303   from the identity plot: 1. Like ECoV, BtCoV/RfYN2012 exhibits evidence of recombination in the
304   5' end of Orf1ab 2. the spike region of the genome is highly variable as previously reported [20],
305   and 3. The 3' end of the genome also exhibits considerable diversity (**Figure 5A**).

306   To confirm the recent common ancestry of SADSr-CoVs in our data set we conducted

307   nucleotide identity and phylogenetic analyses of the RdRp, 3ClPro, helicase, and

308   methyltransferase NTD-encoding regions of Orf1ab. All viruses exhibit exhibit 94-100%

309   nucleotide identity to the reference SADSr-CoV/162140 in these regions of the genome (**Figure**

310   **S4E-F, S8B-E, S9B-C**). In contrast, BtCoV/RfYN2012 recombinant region 1 (RR1) has <70%

311   identity to the reference or any known sequence (**Figure S8F, S9A**), providing evidence that an

312   uncharacterized alphacoronavirus lineage circulates in horseshoe bats, which frequently

313   recombines with SADSr-CoVs.

314   The spike gene is a striking recombination hotspot among SADSr-CoVs. Due to the

315   clustering of putative breakpoints surrounding the 5' end, 3' end, and middle of spike, we ran

316   IDPlot on subsets of three viruses – SADSr-CoV/162140 (reference), SADSr-CoV/141388 or

317   SADS-CoV/FarmA, and a virus of interest from the larger dataset. We found breakpoints

318   delineating six distinct and highly divergent spike genes among the eight analyzed viruses

319   (**Figure 5B**), which reflects recombination events encompassing either the entire spike or the

320   S1 subunit that mediates receptor binding. There are 3 unique full-length spikes

321   (BtCoV/RfY2012, HKU2r-BtCoV/160660, BtCoV/HKU2) with 63-73% nucleotide identity to the

322   reference sequence and two unique S1 domains (SADSr-CoVs/8462 and 8495) with <80%

323   identity to the reference (**Figure 5B, S9A**). Some of these regions match with high identity to

324   partial sequences in GenBank (indicated by an asterisk in Figure 5B) which may be either the

325   parent virus of the recombinant spike or different isolates of the same virus for which a full-

326   length genome is available. Other spikes in this dataset are clearly divergent from any other

327   known sequence.

328   In addition to spike, accessory proteins that target innate immunity can play important

329   roles in host range and pathogenesis [34]. We found a second recombination hotspot

330   surrounding the accessory gene Orf7a, which rivals spike gene diversification. Specifically, our

331   dataset contained five distinct Orf7a genes, some of which lack any closely related sequences

332   in GenBank (**Figure 5C, S9A**).

333   Finally, we mapped each inferred occurrence of a recombination event onto a SADSr-

334   CoV phylogenetic tree. SADSr-CoVs 141388 and 8495 share an Orf7a recombination event,

335   suggesting a recent common ancestor for these two viruses. The tree based on 3ClPro was

336   most consistent with this evolutionary scenario (**Figure 5D**), while the other trees exhibit slightly

337   different topology with minimal diversity, likely due to cryptic recombination events among very

338   closely related viruses. Considering the 3ClPro tree, it is evident that many independent

339   recombination events occurred in the very recent past given that few of the events are shared

340   among the viruses in our dataset (**Figure 5D**).

341   The SADSr-CoV lineage is rapidly diversifying via recombination, particularly in the spike

342   and ORF7a accessory genes. We observed that numerous viruses with >95-99% identity in

343   conserved Orf1ab regions contain highly divergent spike and accessory genes which may shift

344   host range and virulence in otherwise nearly isogenic viruses. These findings highlight how

345   viruses sampled to date represent only a sliver of circulating SADSr-CoV coronavirus diversity

346   and that coronaviruses can change rapidly, drastically, and unpredictably via recombination with

347   both known and unknown lineages. The SADSr-CoVs exemplify the potential of coronaviruses

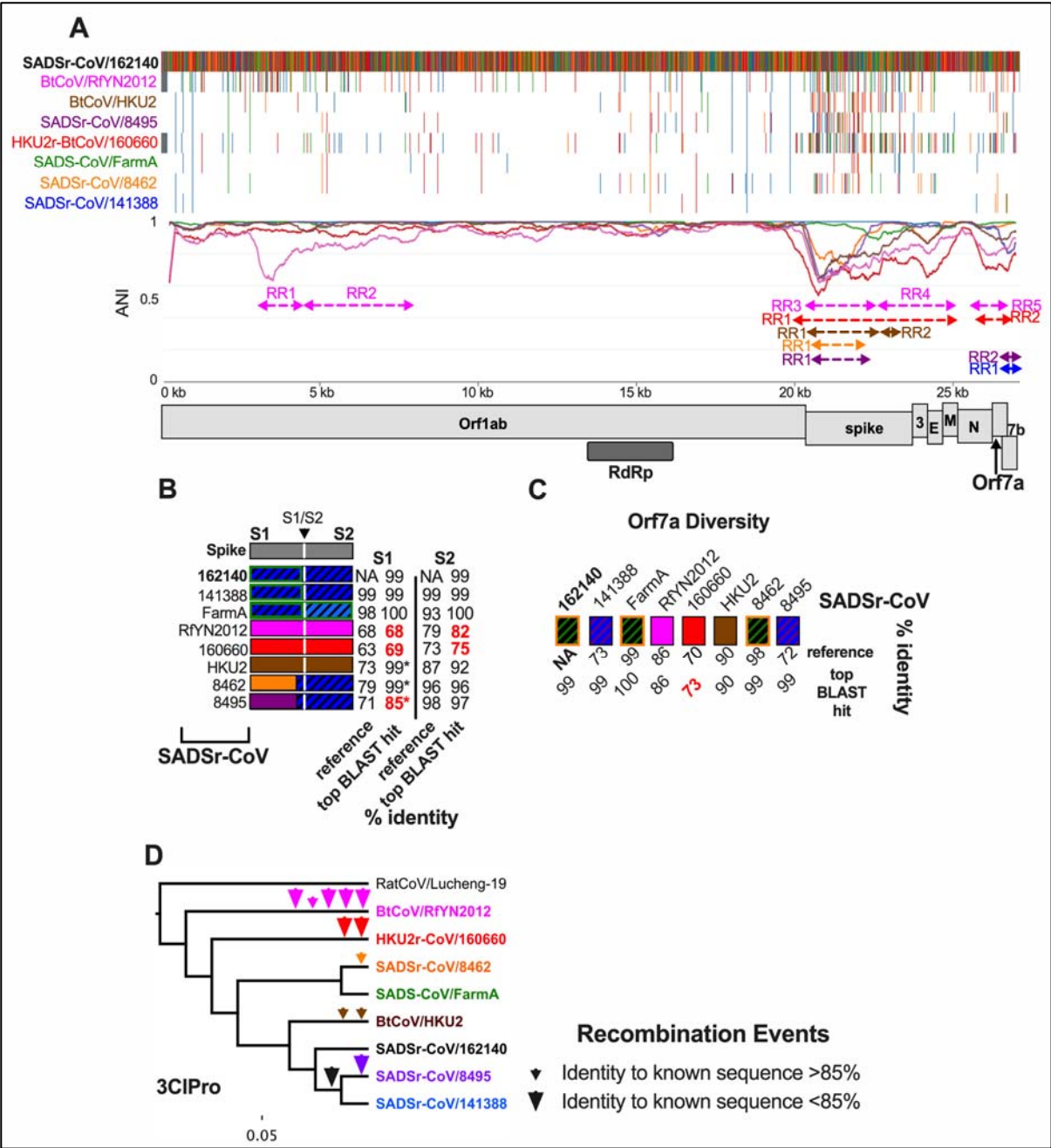348   to rapidly evolve through promiscuous recombination.

349

350

Figure 5. SADSr-CoV IDPlot Analysis. A) IDPlot nucleotide identity and multiple sequence alignment of eight SADSr-CoVs. Color-coded dashed lines indicate divergent regions in corresponding viruses owing to recombination events. B) Schematic of spike genes of SADSr-CoVs along with nucleotide identity to the reference sequence and closest related sequences in GenBank for S1 and S2 domains. C) Schematic of Orf7a diversity with nucleotide identity to the reference sequence and closest related sequences in GenBank D) Phylogenetic tree of SADSr-CoVs based on 3ClPro sequence illustrating the history of inferred recombination events indicated by arrowheads.

351

Discussion

352

353   We developed IDPlot to explore the role of recombination in the diversification of

354   coronaviruses. Coronaviruses are ubiquitous human pathogens with vast and underexplored

355   genetic diversity. SARS-CoV-2 is the second SARSr-CoV known to infect humans and the fifth

356   zoonotic coronavirus known to sweep through the human population following HCoVs 229E,

357   NL63, HKU1, and OC43 [9,10,15,48,55,56]. Most effort in evaluating the threat to human health

358   posed by coronaviruses has been dedicated to discovery of novel SARSr-CoVs in wildlife, yet

359   prior to the SARS-CoV-2 pandemic this group of viruses went largely undetected. Much less

360   attention has been paid to other groups that have produced human coronaviruses such as the

361   sparsely sampled *Betacoronavirus-1* and emerging livestock viruses such as the SADSr-CoVs,

362   which exhibit potential to infect humans and already have significant economic impacts.

363   Recombination detection can be difficult when parental viruses are unknown, as was revealed

364   with our analysis, due to difficulty in distinguishing between true recombination events versus

365   repeated mutations under strong selective pressure. Rapid evolution is most evident for spike

366   receptor binding domains, leading to polymorphism at critical residues [57,58]. Multiple

367   sequence alignments generated by IDPlot demonstrate that even in divergent spikes, the low

368   nucleotide identity is evenly distributed throughout putative recombinant regions. Additionally,

369   we see high divergence even in conserved regions such as Orf1ab and the spike S2 domain. In

370   all of these regions, including accessory genes, reshuffling of phylogenetic trees described in

371   our analysis provides strong evidence that recombination, not repeated individual mutations of

372   critical amino acid residues, accounts for the observed diversity.

373   We initially used the SARS-CoV-2-like viruses to test and validate IDPlot and in the

374   process characterized recombination among these viruses in greater detail than previously

375   reported. The observed variability in arrangements of PangolinCoV/GD19 and RmYN02 on a

376   *SARSr-CoV* phylogenetic tree (**Figure 3B-C, 3E, S5**) depending on the region being sampled is

377   a classic recombination signal easily observed in the IDPlot output. We also analyzed

378   recombination dynamics for viruses in *BetaCoV1* and among SADSr-CoVs. Broad similarities

379   emerge from these studies. Most recombination appears to involve the spike gene and/or

380   various accessory genes. However, in both *BetaCoV1* and among SADSr-CoVs we detected

381   recombination events in Orf1ab as well. Spike and accessory gene recombination events are

382   particularly notable given the potential to influence host range and pathogenesis.

383   This preliminary analysis showed that IDPlot is a powerful new pipeline for sequence

384   identity analysis, breakpoint prediction, and phylogenetic analysis. Existing workflows for

385   nucleotide similarity analysis are proprietary, lack the ability to identify phylogenetic

386 incongruence that is a signature of recombination and do not support direct export of genomic

387 regions for BLAST analysis. This automates and streamlines multi-step analysis with few

388 barriers to use. Nevertheless, there are opportunities for further improvement. Analysis of

389 recombination breakpoints implemented in GARD are of limited value for resolving unique

390 breakpoints in close proximity, as observed surrounding and within SADSr-CoV and other spike

391 genes, necessitating the use of small sets of sequences. Second, GARD is computationally

392 intensive and best suited to small data sets. It is configured as an optional step in IDPlot, so

393 multiple sequence alignments and nucleotide identity plots can be rapidly generated in a local

394 environment. However, for GARD analysis we relied on a high-performance computing cluster

395 to expedite the process. In the future, we anticipate adding other, less intensive breakpoint

396 prediction algorithms to the IDPlot options menu. Future advances in computational methods

397 may also improve the ability to resolve unique breakpoints clustered in genomic regions that are

398 recombination hotspots, most notably the spike gene.

399   Our IDPlot analyses revealed new evidence of extensive recombination-driven evolution

400 in other coronavirus groups. Wildlife sampling indicates that SADSr-Covs are a large pool of

401 closely related viruses circulating in horseshoe bat populations at high frequency. This is the

402 same genus of bats that include SARSr-CoVs suggesting that the ecological conditions for

403 SADSr-CoV spillover into humans may be in place. The relatedness of these viruses means

404 they have had little time to diverge via mutation, but we find they are rapidly diversifying due to

405 recombination, acquiring spike and accessory genes from unsampled viral lineages. These

406 findings demonstrate that rather than a single threat to human health posed by SADS-CoV,

407 there is a highly diverse reservoir of such viruses in an ecological position and with diversity

408 reminiscent of SARSr-CoVs. We found a similar dynamic at play among *BetaCoV1* which are

409 under-sampled to an even greater degree and receive far less attention. Nevertheless, these

410 viruses are involved in genetic exchange with unsampled lineages, with unpredictable

411 consequences.

412   Our findings bear on strategies for anticipating and countering future zoonotic events.

413 SARSr-CoVs garner considerable attention, with an intense focus on viruses able to infect

414 human cells using ACE-2 as an entry receptor. However, RmYN02 demonstrates that viruses

415 can toggle between spikes that recognize ACE-2 or different entry receptors but still infect the

416 same hosts and continue to undergo recombination. Work to prepare for future zoonotic SARSr-

417 CoVs must account for the possibility that the threat will come from coronaviruses only distantly

418 related to SARSr-CoVs undergoing frequent recombination and distributing genetic diversity

419 across the phylogenetic tree of coronaviruses.

420   More attention to the evolutionary dynamics of *BetaCoV1* and SADSr-CoVs is also

421 warranted. Both groups originate in wildlife: rodents and horseshoe bats respectively, and are

422 enzootic or epizootic in livestock. *BetaCoV1* includes a pandemic virus that swept the human

423 population, OC43, while SADS-CoV efficiently infects primary human respiratory and intestinal

424 epithelial cells [22]. Our ability to anticipate threats from both groups would benefit from

425 additional sampling, with *BetaCoV1* being particularly undersampled. Increased surveillance at

426 wildlife-livestock interfaces, including agricultural workers is needed for early detection of novel

427 viruses coming into contact with humans. Due to recombination, prior infection with a virus such

428 as OC43 cannot be presumed to be protective against even closely related viruses that can

429 encode highly divergent spikes, as demonstrated in our analysis. Similarly, efforts to develop

430 medical countermeasures against SADS-CoV should consider the full breadth of diversity

431 among related viruses, while aiming for broadly effective vaccines and therapeutics.

432   Using IDPlot, we identified extensive diversity among coronavirus spike and accessory

433 genes with potential implications for future pandemics. From the standpoint of understanding

434 coronavirus evolution, frequent recombination events often reshuffle phylogenetic trees and can

435 obscure evolutionary relationships. The extent to which viruses in current databases contain

436 genomic regions with no known close relatives makes clear that coronavirus diversity is vast

437 and poorly sampled, even for viruses circulating in well-studied locations. This proximity raises

438 the possibility of recurrent zoonoses of coronaviruses encoding divergent spike and accessory

439 genes. Therefore, preparedness efforts should consider a broad range of virus diversity rather

440 than risk a more narrow focus on close relatives of coronaviruses that most recently impacted

441 human health.

442

443 **Methods**

444

445 **Virus Sequences.** All sequences were downloaded from GenBank with the exception of

446 PangolinCoV/GD19 and BtCoV/RmYN02, which were acquired from the Global Initiative on

447 Sharing All Influenza Data (GISAID) database (https://www.gisaid.org).

448

449 **IDPlot**. IDPlot is initiated by the user designating reference and query sequences. A .gff3

450 annotation file can also be included in the input. The first step of IDPlot is multiple sequence

451 alignment using MAFFT [36] with default parameters. Size of the sliding window is customizable

452 and set to 500 for all of our analyses. For recombination analysis we ran GARD [29] as an

453 optional step, utilizing the multiple sequence alignment generated by MAFFT. Trees for each

454   GARD iteration are generated and displayed using Fast Tree 2 [37]. The entire output is then

455   exported into a chosen directory as idplot.html as well .json files containing raw GARD data.

456   More detailed information on IDPlot is available in the GitHub repository at

457   https://github.com/brwnj/idplot.

458

459   **Phylogenetic validation of breakpoints.** Putative breakpoints were further tested by

460   maximum-likelihood phylogenetic analysis using PhyML [59]. For *Betacoronavirus-1*,

461   RbCoV/HKU14 and MHV (as a root) were aligned with the four viruses in the IDPlot dataset. For

462   SADSr-CoVs we chose HCoV-229E as the root, with the exception of the spike gene, and

463   aligned it with the eight viruses in our dataset. We rooted the SARSr-CoVs with BtCoV/BM48-

464   31/BGR/2008. Given the better sampling of SARSr-CoV, we included more diversity in that

465   alignment to enhance phylogenetic signal. The signal for *BetaCoV1* and SADSr-CoV is

466   constrained by sampling limitations. We extracted breakpoint-defined regions from the

467   alignment and generated ML-phylogenetic trees using a GTR substitution model and 100

468   bootstraps. "Up" and "Dn" regions are the 500 nucleotides upstream or downstream of a

469   proposed 5' or 3' breakpoint, respectively. In the case of SADSr-CoV the clustering of

470   breakpoints around the 5' and 3' ends of spike precluded using unique Up and Dn regions for

471   each recombination event. Instead, we used the N-terminal section of nsp16 (MTase) and the M

472   gene, respectively. For BtCoV/RmYN02 RR2 and ORf8 phylogenetic testing we excluded

473   SARSr-CoVs that have a deletion in Orf8. RmYN02 UpRR2 also does not include BtCoV/WIV1

474   because it has a unique open reading frameed insert in this region and so does not align with

475   SARSr-CoVs lacking this Orfx.

476

477   **BLAST analysis.** To identify the source of recombinant regions we used NCBI Blastn with

478   default parameters, excluding the query sequence from the search. For SADSr-CoVs partial

479   spike sequences frequently appear as top hits. We included these, denoted by an asterisk in

480   reporting the results.

481

482   **Acknowledgements.** We thank E.C. Holmes and co-authors for the use of BtCov/RmYN02 and

483   PangolinCoV/GD19 genome sequences in our analysis. We thank Z.A. Hilbert for manuscript

484   assistance.

485

486

487

References

1. Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, et al. Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. The New England Journal of Medicine. 2003;348: 1967–1976. doi:10.1056/NEJMoa030747

2. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. The New England Journal of Medicine. 2012;367: 1814–1820. doi:10.1056/NEJMoa1211721

3. Zhou P, Yang X-L, Wang X-G, Ben Hu, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. Nature Publishing Group; 2020;579: 270–273. doi:10.1038/s41586-020-2012-7

4. Patrick DM, Petric M, Skowronski DM, Guasparini R, Booth TF, Krajden M, et al. An Outbreak of Human Coronavirus OC43 Infection and Serological Cross-reactivity with SARS Coronavirus. Can J Infect Dis Med Microbiol. Hindawi; 2006;17: 330–336. doi:10.1155/2006/152612

5. Hand J, Rose EB, Salinas A, Lu X, Sakthivel SK, Schneider E, et al. Severe Respiratory Illness Outbreak Associated with Human Coronavirus NL63 in a Long-Term Care Facility. Emerging Infectious Diseases. 2018;24: 1964–1966. doi:10.3201/eid2410.180862

6. Zeng Z-Q, Chen D-H, Tan W-P, Qiu S-Y, Xu D, Liang H-X, et al. Epidemiology and clinical characteristics of human coronaviruses OC43, 229E, NL63, and HKU1: a study of hospitalized children with acute respiratory tract infection in Guangzhou, China. Eur J Clin Microbiol Infect Dis. 2018;37: 363–369. doi:10.1007/s10096-017-3144-z

7. Killerby ME, Biggs HM, Haynes A, Dahl RM, Mustaquim D, Gerber SI, et al. Human coronavirus circulation in the United States 2014-2017. J Clin Virol. 2018;101: 52–56. doi:10.1016/j.jcv.2018.01.019

8. Pfefferle S, Oppong S, Drexler JF, Gloza-Rausch F, Ipsen A, Seebens A, et al. Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. Emerging Infectious Diseases. 2009;15: 1377–1384. doi:10.3201/eid1509.090224

9. Huynh J, Li S, Yount B, Smith A, Sturges L, Olsen JC, et al. Evidence supporting a zoonotic origin of human coronavirus strain NL63. Journal of Virology. 5 ed. American Society for Microbiology Journals; 2012;86: 12816–12825. doi:10.1128/JVI.00906-12

10. Tao Y, Shi M, Chommanard C, Queen K, Zhang J, Markotter W, et al. Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. Perlman S, editor. Journal of Virology. American Society for Microbiology Journals; 2017;91: 85. doi:10.1128/JVI.01953-16

11. Khalafalla AI, Lu X, Al-Mubarak AIA, Dalab AHS, Al-Busadah KAS, Erdman DD. MERS-CoV in Upper Respiratory Tract and Lungs of Dromedary Camels, Saudi Arabia, 2013–2014. Emerging Infectious Diseases. 2015;21: 1153–1158. doi:10.3201/eid2107.150070

528 12. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, et al. Rooting
529   the phylogenetic tree of middle East respiratory syndrome coronavirus by
530   characterization of a conspecific virus from an African bat. Journal of Virology. American
531   Society for Microbiology Journals; 2014;88: 11297–11303. doi:10.1128/JVI.01498-14

532 13. Corman VM, Eckerle I, Memish ZA, Liljander AM, Dijkman R, Jonsdottir H, et al. Link of a
533   ubiquitous human coronavirus to dromedary camels. Proc Natl Acad Sci USA. National
534   Academy of Sciences; 2016;113: 9864–9869. doi:10.1073/pnas.1604472113

535 14. Crossley BM, Mock RE, Callison SA, Hietala SK. Identification and characterization of a
536   novel alpaca respiratory coronavirus most closely related to the human coronavirus 229E.
537   Viruses. Multidisciplinary Digital Publishing Institute; 2012;4: 3689–3700.
538   doi:10.3390/v4123689

539 15. Lau SKP, Woo PCY, Li KSM, Tsang AKL, Fan RYY, Luk HKH, et al. Discovery of a novel
540   coronavirus, China Rattus coronavirus HKU24, from Norway rats supports the murine
541   origin of Betacoronavirus 1 and has implications for the ancestor of Betacoronavirus
542   lineage A. Sandri-Goldin RM, editor. Journal of Virology. American Society for
543   Microbiology Journals; 2015;89: 3076–3092. doi:10.1128/JVI.02420-14

544 16. Wang W, Lin X-D, Guo W-P, Zhou R-H, Wang M-R, Wang C-Q, et al. Discovery, diversity
545   and evolution of novel coronaviruses sampled from rodents in China. Virology. Academic
546   Press; 2015;474: 19–27. doi:10.1016/j.virol.2014.10.017

547 17. Zhang J, Guy JS, Snijder EJ, Denniston DA, Timoney PJ, Balasuriya UBR. Genomic
548   characterization of equine coronavirus. Virology. 2007;369: 92–104.
549   doi:10.1016/j.virol.2007.06.035

550 18. Li K, Li H, Bi Z, Gu J, Gong W, Luo S, et al. Complete Genome Sequence of a Novel
551   Swine Acute Diarrhea Syndrome Coronavirus, CH/FJWT/2018, Isolated in Fujian, China,
552   in 2018. Matthijnssens J, editor. Microbiol Resour Announc. American Society for
553   Microbiology; 2018;7: 466. doi:10.1128/MRA.01259-18

554 19. Zhou L, Li QN, Su JN, Chen GH, Wu ZX, Luo Y, et al. The re-emerging of SADS-CoV
555   infection in pig herds in Southern China. Transbound Emerg Dis. John Wiley & Sons, Ltd;
556   2019;66: 2180–2183. doi:10.1111/tbed.13270

557 20. Zhou P, Fan H, Lan T, Yang X-L, Shi W-F, Zhang W, et al. Fatal swine acute diarrhoea
558   syndrome caused by an HKU2-related coronavirus of bat origin. Nature. 1st ed. Nature
559   Publishing Group UK; 2018;556: 255–258. doi:10.1038/s41586-018-0010-9

560 21. Yang Y-L, Qin P, Wang B, Liu Y, Xu G-H, Peng L, et al. Broad Cross-Species Infection of
561   Cultured Cells by Bat HKU2-Related Swine Acute Diarrhea Syndrome Coronavirus and
562   Identification of Its Replication in Murine Dendritic Cells In Vivo Highlight Its Potential for
563   Diverse Interspecies Transmission. Gallagher T, editor. Journal of Virology. American
564   Society for Microbiology Journals; 2019;93: 3134. doi:10.1128/JVI.01448-19

565 22. Edwards CE, Yount BL, Graham RL, Leist SR, Hou YJ, Dinnon KH, et al. Swine acute
566   diarrhea syndrome coronavirus replication in primary human cells reveals potential
567   susceptibility to infection. Proceedings of the National Academy of Sciences. National
568   Academy of Sciences; 2020;4: 202001046. doi:10.1073/pnas.2001046117

569 23. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. High fidelity of murine hepatitis virus
570   replication is decreased in nsp14 exoribonuclease mutants. Journal of Virology. American
571   Society for Microbiology Journals; 2007;81: 12135–12144. doi:10.1128/JVI.01296-07

572 24. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA
573   proofreading machine regulates replication fidelity and diversity. RNA Biol. Taylor &
574   Francis; 2011;8: 270–279. doi:10.4161/rna.8.2.15013

575 25. Smith EC, Denison MR. Implications of altered replication fidelity on the evolution and
576   pathogenesis of coronaviruses. Curr Opin Virol. 2012;2: 519–524.
577   doi:10.1016/j.coviro.2012.07.005

578 26. Graham RL, Baric RS. Recombination, Reservoirs, and the Modular Spike: Mechanisms
579   of Coronavirus Cross-Species Transmission. Journal of Virology. 2010;84: 3134–3146.

580 27. Yang X-L, Hu B, Wang B, Wang M-N, Zhang Q, Zhang W, et al. Isolation and
581   Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of
582   Severe Acute Respiratory Syndrome Coronavirus. Perlman S, editor. Journal of Virology.
583   American Society for Microbiology; 2016;90: 3253–3256. doi:10.1128/JVI.02582-15

584 28. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, et al. Discovery of a rich gene pool of
585   bat SARS-related coronaviruses provides new insights into the origin of SARS
586   coronavirus. PLOS Pathogens. Public Library of Science; 2017;13: e1006698.
587   doi:10.1371/journal.ppat.1006698

588 29. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic
589   algorithm for recombination detection. Bioinformatics. 2006;22: 3096–3098.

590 30. Saberi A, Gulyaeva AA, Brubacher JL, Newmark PA, Gorbalenya AE. A planarian
591   nidovirus expands the limits of RNA genome size. PLOS Pathogens. Public Library of
592   Science; 2018;14: e1007314. doi:10.1371/journal.ppat.1007314

593 31. Debat HJ. Expanding the size limit of RNA viruses: Evidence of a novel divergent
594   nidovirus in California sea hare, with a ~35.9 kb virus genome. 2018. doi:10.1101/307678

595 32. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis.
596   Methods Mol Biol. New York, NY: Springer New York; 2015;1282: 1–23. doi:10.1007/978-
597   1-4939-2438-7_1

598 33. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The
599   species Severe acute respiratory syndrome-related coronavirus : classifying 2019-nCoV
600   and naming it SARS-CoV-2. Nature Microbiology. Nature Publishing Group; 2020;5: 536–
601   544. doi:10.1038/s41564-020-0695-z

602 34. Liu DX, Fung TS, Chong KK-L, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV
603   and other coronaviruses. Antiviral Research. 2014;109: 97–109.
604   doi:10.1016/j.antiviral.2014.06.013

605 35. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow
606   enables reproducible computational workflows. Nat Biotechnol. Nature Publishing Group;
607   2017;35: 316–319. doi:10.1038/nbt.3820

608 36. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple
609   sequence alignment based on fast Fourier transform. nar. Oxford University Press;
610   2002;30: 3059–3066.

611 37. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees
612   for Large Alignments. Poon AFY, editor. PLoS ONE. Public Library of Science; 2010;5:
613   e9490. doi:10.1371/journal.pone.0009490

614 38. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins
615   of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.
616   Nature Microbiology. Nature Publishing Group; 2020;382: 1–10. doi:10.1038/s41564-020-
617   0771-4

618 39. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A Novel Bat Coronavirus Closely
619   Related to SARS- CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the
620   Spike Protein. Current Biology. Elsevier Ltd; 2020;: 1–12. doi:10.1016/j.cub.2020.05.023

621 40. Ge X-Y, Wang N, Zhang W, Hu B, Li B, Zhang Y-Z, et al. Coexistence of multiple
622   coronaviruses in several bat colonies in an abandoned mineshaft. Virol Sin. Springer
623   Singapore; 2016;31: 31–40. doi:10.1007/s12250-016-3713-9

624 41. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, et al. Identifying SARS-
625   CoV-2-related coronaviruses in Malayan pangolins. Nature. Nature Publishing Group;
626   2020;583: 282–285. doi:10.1038/s41586-020-2169-0

627 42. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related
628   coronavirus from Malayan pangolins. Nature. Nature Publishing Group; 2020;583: 286–
629   289. doi:10.1038/s41586-020-2313-x

630 43. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, et al. Attenuation of
631   replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early
632   stages of human-to-human transmission. Sci Rep. Nature Publishing Group UK; 2018;8:
633   990–15177. doi:10.1038/s41598-018-33487-8

634 44. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, et al. Discovery and
635   Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the
636   Early Evolution of SARS-CoV-2. Schultz-Cherry S, editor. mBio. 2020;11: e01610–20.

637 45. Pereira F. Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. Infect
638   Genet Evol. Elsevier B.V; 2020;85: 104525–104525.

639 46. Young BE, Fong S-W, Chan Y-H, Mak T-M, Ang LW, Anderson DE, et al. Effects of a
640   major deletion in the SARS-CoV-2 genome on the severity of infection and the
641   inflammatory response: an observational cohort study. The Lancet. Elsevier; 2020;396:
642   603–611. doi:10.1016/S0140-6736(20)31757-8

643 47. Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, et al. Severe Acute Respiratory
644   Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related
645   Coronavirus from Greater Horseshoe Bats through Recombination. Perlman S, editor.
646   Journal of Virology. 2015;89: 10532–10547. doi:10.1128/JVI.01048-15

647 48.   Vijgen L, Keyaerts E, Lemey P, Maes P, Van Reeth K, Nauwynck H, et al. Evolutionary
648   history of the closely related group 2 coronaviruses: porcine hemagglutinating
649   encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. Journal of
650   Virology. 2006;80: 7270–7274. doi:10.1128/JVI.02675-05

651 49.   Woo PCY, Lau SKP, Wernery U, Wong EYM, Tsang AKL, Johnson B, et al. Novel
652   betacoronavirus in dromedaries of the Middle East, 2013. Emerging Infect Dis. 2014;20:
653   560–572. doi:10.3201/eid2004.131769

654 50.   So RTY, Chu DKW, Miguel E, Perera RAPM, Oladipo JO, Fassi-Fihri O, et al. Diversity of
655   Dromedary Camel Coronavirus HKU23 in African Camels Revealed Multiple
656   Recombination Events among Closely Related Betacoronaviruses of the Subgenus
657   Embecovirus. Pfeiffer JK, editor. Journal of Virology. American Society for Microbiology
658   Journals; 2019;93: 255. doi:10.1128/JVI.01236-19

659 51.   Lau SKP, Woo PCY, Yip CCY, Fan RYY, Huang Y, Wang M, et al. Isolation and
660   characterization of a novel Betacoronavirus subgroup A coronavirus, rabbit coronavirus
661   HKU14, from domestic rabbits. Journal of Virology. 6 ed. 2012;86: 5481–5496.
662   doi:10.1128/JVI.06927-11

663 52.   Nemoto M, Kanno T, Bannai H, Tsujimura K, Yamanaka T, Kokado H. Antibody response
664   to equine coronavirus in horses inoculated with a bovine coronavirus vaccine. J Vet Med
665   Sci. JAPANESE SOCIETY OF VETERINARY SCIENCE; 2017;79: 1889–1891.
666   doi:10.1292/jvms.17-0414

667 53.   Gong L, Li J, Zhou Q, Xu Z, Chen L, Zhang Y, et al. A New Bat-HKU2-like Coronavirus in
668   Swine, China, 2017. Emerging Infectious Diseases. 2017;23: 1607–1609.
669   doi:10.3201/eid2309.170915

670 54.   Lau SKP, Woo PCY, Li KSM, Huang Y, Wang M, Lam CSF, et al. Complete genome
671   sequence of bat coronavirus HKU2 from Chinese horseshoe bats revealed a much
672   smaller spike gene with a different evolutionary lineage from the rest of the genome.
673   Virology. Elsevier Inc; 2007;367: 428–439. doi:10.1016/j.virol.2007.06.009

674 55.   Maganga GD, Pinto A, Mombo IM, Madjitobaye M, Mbeang Beyeme AM, Boundenga L,
675   et al. Genetic diversity and ecology of coronaviruses hosted by cave-dwelling bats in
676   Gabon. Sci Rep. Nature Publishing Group; 2020;10: 7314–13. doi:10.1038/s41598-020-
677   64159-1

678 56.   Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, et al. Complete genomic
679   sequence of human coronavirus OC43: molecular clock analysis suggests a relatively
680   recent zoonotic coronavirus transmission event. Journal of Virology. American Society for
681   Microbiology Journals; 2005;79: 1595–1604. doi:10.1128/JVI.79.3.1595-1604.2005

682 57.   Guo H, Hu B-J, Yang X-L, Zeng L-P, Li B, Ouyang S, et al. Evolutionary Arms Race
683   between Virus and Host Drives Genetic Diversity in Bat Severe Acute Respiratory
684   Syndrome-Related Coronavirus Spike Genes. Pfeiffer JK, editor. J Virol. 2020;94:
685   e00902–20.

686 58.   Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol.
687   Annual Reviews; 2016;3: 237–261. doi:10.1146/annurev-virology-110615-042301

688   59.   Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online—a web server for fast
689   maximum likelihood-based phylogenetic inference. nar. 2005;33: W557–W559.

690

691