JVI Accepted Manuscript Posted Online 16 September 2015
J. Virol. doi:10.1128/JVI.01755-15

Evidence for an ancestral association of human coronavirus 229E with bats

Victor Max Corman[1,2#], Heather J. Baldwin[3,4#], Adriana Fumie Tateno[1,5#], Rodrigo Melim Zerbinati[1,5], Augustina Annan[6], Michael Owusu[6], Evans Ewald Nkrumah[6], Gael Darren Maganga[7], Samuel Oppong[8], Yaw Adu-Sarkodie[8], Peter Vallo[3,9], Luiz Vicente Ribeiro Ferreira da Silva Filho[5,10], Eric M. Leroy[7,11], Volker Thiel[12,13], Lia van der Hoek[14], Leo L. M. Poon[15], Marco Tschapka[3,16], Christian Drosten[1,2*], Jan Felix Drexler[1,2*]

1: Institute of Virology, University of Bonn Medical Centre, Bonn, Germany
2: German Centre for Infection Research (DZIF), partner site Bonn-Cologne, Germany
3: Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm, Ulm, Germany
4: Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia
5: Laboratório de Virologia (LIMHC-FMUSP), Instituto de Medicina Tropical de São Paulo, Universidade de São Paulo
6: Kumasi Centre for Collaborative Research in Tropical Medicine (KCCR), Kumasi, Ghana
7: Centre International de Recherches Médicales de Franceville, Franceville, Gabon
8: Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
9: Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Brno
10: Instituto da Criança, Hospital das Clínicas da FMUSP, São Paulo, Brazil
11: Institut de Recherche pour le Développement, UMR 224 (MIVEGEC), IRD/CNRS/UM1, Montpellier, France
12: Department of Infectious Diseases and Pathobiology, Vetsuisse Faculty Bern, University of Bern, Switzerland
13: Federal Department of Home Affairs, Institute of Virology and Immunology, Bern and Mittelhäusern, Switzerland

14: Department of Medical Microbiology, Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands

15: School of Public Health, The University of Hong Kong, Hong Kong, Special Administrative Region, People's Republic of China

16: Smithsonian Tropical Research Institute, Balboa, Panama

#These authors contributed equally

*Addresses for correspondence:

Professor Dr. Christian Drosten, Institute of Virology, University of Bonn Medical Centre, 53127 Bonn, Germany, Tel.: +49-228-287-11055/Fax: +49-228-287-14433, e-mail: drosten@virology-bonn.de

Professor Dr. Jan Felix Drexler, Institute of Virology; University of Bonn Medical Centre, 53127 Bonn, Germany, Tel.: +49-228-287-11697/Fax: +49-228-287-14433, e-mail: drexler@virology-bonn.de

Running head: HCoV-229E-related bat coronaviruses

Abstract word count: 247

Text Body word count: 4,372

One-sentence summary: HCoV-229E-related bat coronaviruses are genetically highly diversified and suggest HCoV-229E acquired major genomic deletions upon host switching, potentially involving camelids as intermediate hosts.

Keywords: Africa, Coronavirus, Bats, Camelids, HCoV-229E, Zoonoses

Abstract

We previously showed that close relatives of human coronavirus (HCoV)-229E exist in African bats. The small sample and limited genomic characterizations prevented further analyses so far. Here, we tested 2,087 fecal specimens from 11 bat species sampled in Ghana for HCoV-229E-related viruses by RT-PCR. Only hipposiderid bats tested positive. To compare the genetic diversity of bat viruses and HCoV-229E, we tested historical isolates and diagnostic specimens sampled globally over 10 years. Bat viruses were five- to sixfold more diversified than HCoV-229E in *RNA-dependent RNA polymerase* (*RdRp*) and *Spike* genes. In phylogenetic analyses, HCoV-229E strains were monophyletic and not intermixed with animal viruses. Bat viruses formed three large clades in close and more distant sister relationship. A recently described 229E-related alpaca virus occupied an intermediate phylogenetic position between bat and human viruses. According to taxonomic criteria, human, alpaca and bat viruses form a single CoV species showing evidence for multiple recombination events. HCoV-229E and the alpaca virus showed a major deletion in the *Spike* S1 region compared to all bat viruses. Analyses of four full genomes from 229E-related bat CoVs revealed an eighth open reading frame (ORF8) located at the genomic 3'-end. ORF8 also existed in the 229E-related alpaca virus. Re-analysis of HCoV-229E sequences showed a conserved transcription regulatory sequence preceding remnants of this ORF, suggesting its loss after acquisition of a 229E-related CoV by humans. These data suggested an evolutionary origin of 229E-related CoVs in hipposiderid bats, hypothetically with camelids as intermediate hosts preceding the establishment of HCoV-229E.

Importance

The ancestral origins of major human coronaviruses (HCoV) likely involve bat hosts. Here, we provide conclusive genetic evidence for an evolutionary origin of the common cold virus HCoV-229E in hipposiderid bats by analyzing a large sample of African bats and

characterizing several bat viruses on a full genome level. Our evolutionary analyses show that animal and human viruses are genetically closely related, can exchange genetic material and form a single viral species. We show that the putative host switches leading to the formation of HCoV-229E were accompanied by major genomic changes including deletions in the viral spike glycoprotein gene and loss of an open reading frame. We re-analyze a previously described genetically related alpaca virus and discuss the role of camelids as potential intermediate hosts between bat and human viruses. The evolutionary history of HCoV-229E likely shares important characteristics with that of the recently emerged highly pathogenic MERS-Coronavirus.

87 **Introduction**

88 Coronaviruses (CoV) are enveloped viruses with a single-stranded, positive-sense contiguous
89 RNA genome of up to 32 kilobases. The subfamily *Coronavirinae* contains four genera
90 termed *Alpha*-, *Beta*-, *Gamma*- and *Deltacoronavirus*. Mammals are predominantly infected
91 by alpha- and betacoronaviruses, while gamma- and deltacoronaviruses mainly infect avian
92 hosts (1, 2).
93
94 Four human coronaviruses (HCoVs) termed HCoV-229E, -NL63, -OC43 and -HKU1
95 circulate in the human population and mostly cause mild respiratory disease (3). HCoV-229E
96 is frequently detected in up to 15% of specimens taken from individuals with respiratory
97 disease (4-6). Although HCoV-229E can be detected in fecal specimens, HCoVs generally
98 don't seem to play a role in acute gastroenteritis (7-9). Severe respiratory disease with high
99 case-fatality rates is caused by *severe acute respiratory syndrome* (SARS)-CoV and *Middle*
100 *East respiratory syndrome* (MERS)-CoV which emerged recently. HCoV-229E and HCoV-
101 NL63 belong to the genus *Alphacoronavirus*, while HCoV-OC43, HCoV-HKU1, SARS- and
102 MERS-CoV belong to the genus *Betacoronavirus* (1, 10).
103
104 In analogy to major human pathogens including Ebola virus, rabies virus, mumps virus and
105 hepatitis B and C viruses (11-16), the evolutionary origins of SARS- and MERS-CoV were
106 traced back to bats (17-22). The genetic diversity of bat CoVs described over the last decade
107 exceeds the diversity in other mammalian hosts (2). This has led to speculations on an
108 evolutionary origin of all mammalian CoVs in bat hosts (23). Bats share important ecological
109 features potentially facilitating virus maintenance and transmission, such as close contact
110 within large social groups, longevity, and the ability of flight (13, 24).
111

How humans become exposed to remote wildlife viruses is not always clear (25). Human infection with SARS-CoV and MERS-CoV was likely mediated by peri-domestic animals. For SARS-CoV, the suspected source of infection were carnivores (26). Preliminary evidence suggested that these carnivore hosts may also have adapted SARS-CoV for human infection (27). For MERS-CoV, camelids are likely intermediate hosts, supported by circulation of MERS-CoV in camel herds globally and for prolonged periods of time (28-30). Whether MERS-CoV only recently acquired the capacity to infect humans in camelids is unclear.

The evolutionary origins of HCoV-229E are uncertain. In 2007, a syndrome of severe respiratory disease and sudden death was recognized in captive alpacas from the U.S. (31) and an alphacoronavirus genetically closely related to HCoV-229E was identified as the causative agent (32).

In 2009, we detected viruses in fecal specimens from 5 of 75 hipposiderid bats from Ghana and showed that these bat viruses were genetically related to HCoV-229E by characterizing their partial *RNA-dependent RNA polymerase* (*RdRp*) and *Nucleocapsid* genes (33). Lack of specimens containing high CoV RNA concentrations so far prevented a more comprehensive characterization of those bat viruses to further address their relatedness to HCoV-229E. Here, we tested more than 2,000 bats from Ghana for CoVs related to HCoV-229E. We describe highly diversified bat viruses on a full genome level and analyze the evolutionary history of HCoV-229E and the genetically related alpaca CoV.

Materials and Methods

Bat and human sampling

Bats were caught in the Ashanti region, central Ghana, during 2009-2011 as described previously (21). Archived anonymized respiratory specimens derived from patients sampled between 2002-2011 were obtained from Hong Kong/China, Germany, The Netherlands, Brazil and Ghana.

RNA purification, coronavirus detection and characterization

RNA was purified from approximately 20 mg of fecal material suspended in 500 µL RNAlater stabilizing solution using the MagNA Pure 96 system (Roche Penzberg, Germany). Elution volumes were 100 µL. Testing for CoV RNA was done using a real time RT-PCR assay designed to allow detection of HCoV-229E and all genetically related bat CoVs known from our pilot study (33). Oligonucleotide sequences were CoV229Elike-F13948m TCYAGAGAGGTKGTTGTTACWAAYCT, CoV229Elike-P13990m FAM (6-Carboxyfuorescein)-TGGCMACTTAATAAGTTTGGIAARGCYGG-BHQ1 (Black Hole Quencher 1) and CoV229Elike-R14138m CGYTCYTTRCCAGAWATGGCRTA. Testing used the SSIII RT-PCR Kit (Life Technologies, Karlsruhe, Germany) with the following cycling protocol in a LightCycler 480 (Roche, Penzberg, Germany): 20 min. at 50 ºC for reverse transcription, followed by 3 min. at 95 ºC and 45 cycles of 15 sec. at 95 ºC, 10 sec. at 58 ºC and 20 sec. at 72 ºC. CoV quantification relied on cRNA *in vitro* transcripts generated from TA-cloned peri-amplicons using the T7-driven Megascript (Life technologies, Heidelberg, Germany) kit as described previously (34). Partial *RdRp* gene sequences from real time RT-PCR-positive specimens were obtained as described previously (18). Full CoV genomes and *Spike* gene sequences were generated for those specimens containing highest CoV RNA concentrations using sets of nested RT-PCR assays (primers available upon request) located along the HCoV-229E genome and designed to amplify small sequence

islets. Sequence islets were connected by bridging long-range nested PCR using strain-specific primers (available upon request) and the Expand High Fidelity kit (Roche) on cDNA templates generated with the Superscript III reverse transcriptase (Life Technologies).

Phylogenetic analyses

Bayesian phylogenetic reconstructions were made using MrBayes V3.1 (35) under assumption of a GTR+G+I nucleotide substitution model for partial *RdRp* sequences and the WAG amino acid substitution model for translated open reading frames (ORFs). Two million generations were sampled every 100 steps, corresponding to 20,000 trees of which 25% were discarded as burn-in before annotation using TreeAnnotator V1.5 and visualization using FigTree V1.4 from the BEAST package (36). Neighbor-joining phylogenetic reconstructions were made using MEGA5.2 (37) and a percentage nucleotide distance model, the complete deletion option and 1,000 bootstrap replicates. Genome comparisons were made using MEGA5.2 (37); SSE V1.1 (38) and recombination analyses were made using SimPlot V3.5 (39).

176 **Results**

177 Specimens from 2,087 bats belonging to 11 species were available for PCR testing. **Table 1**
178 provides details on the overall sample composition and detection rates in individual bat
179 species. Only bats belonging to the family Hipposideridae tested positive in 81 of 1,853
180 specimens (4.4%). All positive-testing bats had been morphologically identified in the field as
181 either *Hipposideros* cf. *ruber* or *H. abae*. Those were the most abundant species within the
182 sample. No HCoV-229E-related RNA was detected in the 17 available specimens from *H.*
183 *jonesi* and *H.* cf. *gigas*.
184
185 An 816 nucleotide (nt) fragment from the *RdRp* gene was obtained from 41 of the 81 positive
186 specimens (GenBank accession nos. KT253259-KT253299). This fragment was used for
187 further analysis as the 816 nt sequence yields improved resolution in inference of phylogeny
188 as compared to shorter sequences derived from RT-PCR screening of field-derived samples
189 (2). To expand the available genomic data for HCoV-229E, the 816 nt *RdRp* fragment was
190 also sequenced from 23 HCoV-229E strains from patients sampled between 2002-2011 in
191 China, Germany, The Netherlands, Brazil, and Ghana. In addition, the 816 nt *RdRp* fragment
192 was sequenced from two historical HCoV-229E strains isolated in 1965 and the 1980ies (40)
193 (GenBank accession nos. KT253300-KT253323). In analogy to the official taxonomic
194 designation *SARS-related CoV* including human SARS-CoV and related CoVs from other
195 animals (1), we hereafter restrict usage of the term HCoV-229E to the human virus and refer
196 to the animal viruses as 229E-related CoV. **Figure 1A** shows a Bayesian phylogeny of the
197 partial *RdRp* gene. The bat virus diversity we observed in our pilot study (represented by
198 viruses Buoyem344 and Kwamang19) was expanded greatly. A phylogenetically basal virus
199 termed Kwamang8 obtained within our pilot study was not detected again, although the
200 present study contained specimens from the same cave and bat species. All human strains
201 occupied an apical phylogenetic position and were not intermixed with any of the animal

202 viruses. The recently described alpaca 229E-related CoV (32) clustered with two viruses
203 obtained from hipposiderid bats in a parallel study from our groups in the Central African
204 country Gabon (41). The two Gabonese bat-associated viruses differed from the alpaca 229E-
205 related CoV by only 3.2% nucleotide content within the *RdRp* fragment. Hipposiderid bat
206 CoVs were neither sorted by sampling sites, nor by their host species in their *RdRp* genes.
207 Overall, bat 229E-related CoVs sampled over 3 years differed up to 13.5% in their nt and
208 3.3% in their amino acid (aa) sequences. Although the HCoV-229E dataset used for
209 comparison was sampled over 50 years, the human-associated viruses showed 5-10fold less
210 genetic diversity than bat viruses with only 1.4% nt and 0.7% aa variation. Because of the
211 small sequence variation in HCoV-229E, **Figure 1A** contains only nine representative HCoV-
212 229E strains. The neighbor-joining phylogeny shown in **Figure 1B** represents the high
213 sequence identity between all HCoV-229E strains determined in this study.
214
215 To analyze to which extent bat 229E-related CoV show genetic variation, the *Spike* gene
216 encoding the viral glycoprotein was characterized from 15 representative bat viruses (labeled
217 with a triangle in **Figure 1A**). **Figure 1C** shows a Bayesian phylogenetic tree of the bat 229E-
218 related CoV *Spike* gene sequences and HCoV-229E full *Spike* sequences sampled over 50
219 years. The bat viruses formed three genetically diverse lineage, of which two phylogenetically
220 basal lineages contained bat viruses only. These lineages were sorted according to their
221 sampling sites Kwamang (abbreviated KW) and Akpafu Todzi (abbreviated AT). A third
222 lineage contained closely related bat viruses obtained from three different sample sites
223 separated by several hundred kilometers (Buoyem, Kwamang and Forikrom) (21). These data
224 suggested co-circulation of different *Spike* gene lineages within sampling sites as well as the
225 existence of separate lineages between sites. However, the small number of viruses
226 characterized from the phylogenetically basal bat clades 1 and 2 implies that caution should
227 be taken in assertions on geographically separated *Spike* gene lineages. The alpaca 229E-

related CoV and all HCoV-229E strains clustered in apical phylogenetic position compared to the bat viruses. The most closely related bat viruses from lineage 1 differed from HCoV-229E by 8.4-13.7%. The two other bat virus lineages were less related to HCoV-229E with 30.6-33.0% aa sequence distance.

Topologies of the Bayesian phylogenetic reconstructions of *RdRp* and *Spike* genes from bats and the alpaca were not congruent, compatible with past recombination events across animal 229E-related CoVs. The high similarity of the *RdRp* gene of human HCoV-229E strains did not allow comparisons of the *RdRp*-based with the *Spike*-based topology. To further investigate the genomic relationships of bat 229E-related CoVs and HCoV-229E, the full genomes were determined directly from fecal specimens from four representative bat viruses (labeled with circles in **Figures 1A** and **C**). **Figure 2A** shows that bat 229E-related CoV genomes comprise 28,014-28,748 nt, which exceeds the length of known HCoV-229E strains by 844-1,479 nt. As shown in **Figure 2B**, HCoV-229E and all bat viruses were closely related within the putative *ORF1ab*. This allowed the delineation of non-structural proteins (nsp) 1-16 for all bat viruses in analogy to HCoV-229E. **Table 2** provides details on length and cleavage sites of the predicted nsp. Sequence identity in seven concatenated nsp is used by the International Committee for the Taxonomy of Viruses (ICTV) for CoV species designation (1). As shown in **Table 3**, the four fully sequenced bat viruses showed translated aa sequence identities of 93.3-97.1% with HCoV-229E. This was well above the 90% threshold established by the ICTV, indicating all bat 229E-related CoVs and HCoV-229E form a single species. Bat virus Kwamang8, which formed a phylogenetically basal sister-clade to the other bat viruses and HCoV-229E, could not be sequenced on a full genome level. The aa sequence of the partial *RdRp* gene of Kwamang8 differed by only 3.3% from other bat viruses and HCoV-229E. Based upon previous comparisons of CoV *RdRp* sequences for tentative species delineation (2, 18), Kwamang8 forms part of the same species as the other bat viruses and

HCoV-229E. This CoV species would also include the recently described alpaca 229E-related CoV (32), which showed 96.9-97.2% aa sequence identity with HCoV-229E and 94.2-97.8% with the bat viruses in the seven concatenated nsp domains.

As shown in **Figure 2A**, all seven open reading frames (ORFs) known from HCoV-229E were found in bat 229E-related CoVs in the sequence *ORF1a/1b-Spike-ORF4-Envelope-Membrane-Nucleocapsid*. Amino acid identities between predicted ORFs of the bat viruses and HCoV-229E ranged from the 67.2-91.6% described above for the translated *Spike* genes to 88.3-94.6% (*ORF1ab*), with bat virus lineage 1 consistently showing highest aa sequence identities. **Table 4** provides details for all sequence comparisons.

We looked for additional support for the existence of these predicted ORFs by analyzing the sequence context at their 5'-termini. This is because in CoVs, ORFs are typically preceded by highly conserved transcription regulatory sequence (TRS) elements (42). All putative ORFs from bat-229E related CoVs showed high conservation of the typical HCoV-229E TRS core sequence UCU C/A AACU and adjacent bases. **Table 5** provides details on all putative TRS elements within bat 229E-related CoV genomes.

Figure 3A shows Bayesian phylogenetic trees reconstructed for all individual ORFs. The alpaca 229E-related CoV clustered in intermediate position between HCoV-229E and the bat viruses in the *ORF1ab* and *Spike*, but with bat viruses only in *Membrane*, *Envelope*, *Nucleocapsid*, and *ORF4*. The divergent topologies again suggested recombination events in 229E-related CoVs. To find further evidence for recombination events and identify genomic breakpoints, 229E-related CoVs were analyzed by bootscanning. As shown in **Figure 3B**, bootscanning supported multiple recombination events involving HCoV-229E, bat 229E-related CoVs and the alpaca 229E-related CoV. Major recombination breakpoints occurred within the *ORF1ab* and the beginning of the *Spike* gene, compatible with previous analyses of

CoV recombination patterns (2) and the divergent topologies between the *RdRp* and *Spike* genes noted above. Bootscanning also suggested a potential genomic breakpoint within the *Spike* gene, mapping to the borders of the S1 (associated with for receptor binding) and S2 domains (associated with membrane fusion). This would be consistent with previous evidence supporting intra-*Spike* recombination events in bat-associated CoVs (43). To obtain further support for potential intra-*Spike* recombination events, separate phylogenetic reconstructions for the S1 and the S2 domains were made. As shown in **Figure 3B**, these phylogenetic reconstructions supported recombination events involving the alpaca 229E-related CoV and HCoV-229E, but not the bat 22E-related CoVs. In the S1 domain, the alpaca 229E-related CoV clustered with clinical HCoV-229E strains, while the HCoV-229E reference strain inf-1 isolated in 1962 clustered in phylogenetically basal sister relationship. Only in the S2 domain, the intermediate position of the alpaca compared to bat and human 229E-related CoVs noted before in comparisons of the full *Spike* was maintained. These data may hint at recombination events between HCoV-229E and the alpaca virus and further supported genetic compatibility between these two viruses belonging to one CoV species.

Three major differences existed between HCoV-229E, the alpaca 229E-related CoV and the bat 229E-related CoVs. The first of these differences occurred in the putative *ORF4*. Similar to HCoV-229E strains characterized from clinical specimens, a contiguous *ORF4* existed in all bat viruses that was 156-164 aa residues longer than the alpaca 229E-related CoV *ORF4*. Re-analysis of the putative *ORF4* sequence of the alpaca 229E-related CoV showed that this apparently shorter *ORF4* was due to an insertion of a single cytosine residue at position 181. Without this putative insertion, the alpaca 229E-related CoV *ORF4* showed the same length as homologous ORFs in bat 229E-related CoVs and HCoV-229E. Since the HCoV-229E *ORF4* is known to accumulate mutations in cell culture (40), the apparently truncated ORF in the alpaca 229E-related CoV isolate may thus not occur *in vivo*. The extended *ORF4* of the

alpaca 229E-related CoV would be most closely related to bat viruses from clade 1 with 5.5% aa sequence distance, compared to at least 8.8% distance from HCoV-229E strains.

The second difference was a considerably longer S1 portion of the bat 229E-related CoV *Spike* genes compared to HCoV-229E. **Figure 4** shows that the three bat lineages contained 185-404 additional aa residues upstream of the putative receptor binding domain (44, 45) compared to HCoV-229E. Bat lineage 1 which was phylogenetically most closely related to HCoV-229E carried the smallest number of additional aa residues. Of note, the alpaca 229E-related CoV was identical to HCoV-229E in the number of aa residues within this region of the *Spike* gene.

The third major difference was the existence of an additional putative ORF downstream of the *Nucleocapsid* gene in all bat viruses. Non-homologous ORFs of unknown function downstream the *Nucleocapsid* occur in several alpha- and betacoronaviruses, including *Feline infectious peritonitis virus* (FIPV), *Transmissible gastroenteritis virus of swine* (TGEV), *Rhinolophus bat CoV HKU2*, *Scotophilus bat CoV 512*, *Miniopterus bat CoV HKU8* (23), the *Chaerephon* bat CoVs BtKY22/BtKY41, the *Cardioderma* bat CoV BtKY43 (46) and bat CoV HKU10 from Chinese *Hipposideros* and *Rousettus* species (47). In the genus *Betacoronavirus*, only *Bat CoV HKU9* from *Rousettus* and the genetically related *Eidolon* bat CoV BtKY24 (46) carry additional ORFs at this genomic position. No ORF in the 3'-terminal genome region is known from HCoV-229E. The alpaca 229E-related CoV contains an ORF at this position termed *ORFX* by Crossley et al. (32). In analogy to consecutive numbers used to identify HCoV-229E ORFs, we refer to this ORF as *ORF8* hereafter. The putative TRS context preceding *ORF8* was conserved in all bat 229E-related CoV and in the alpaca 229E-related CoV, suggesting that a corresponding subgenomic mRNA8 may exist. The 3'-UTR of bat 229E-related CoVs immediately followed the putative *ORF8*. This was supported by the

existence of a conserved octanucleotide sequence and highly conserved stem elements forming part of the pseudo-knot typically located at the 5'-end of alphacoronavirus 3'-UTRs (48). As shown in **Figure 5**, HCoV-229E shows a high degree of sequence conservation compared to bat 229E-related CoVs and the alpaca 229E-related CoV in this genomic region, including a highly conserved putative TRS. Bioinformatic analyses (49-51) provided evidence for the presence of two transmembrane domains in the predicted proteins 8 of the alpaca and the genetically related bat 229E-related viruses. This may imply a role of the predicted protein 8 in coronaviral interactions with cellular or viral membranes.

As shown in **Figure 5**, one of the bat 229E-related CoV lineages represented by virus KW2E-F56 contained a highly divergent *ORF8*. In protein BLAST comparisons, the KW2E-F56 *ORF8* showed limited similarity to the putative *ORF7b* of HKU10 and to the putative *ORF8* located upstream the *Nucleocapsid* of a Nigerian *Hipposideros* betacoronavirus termed Zaria bat CoV (47, 52). This may hint at cross-genus recombination events between different hipposiderid bat CoVs in the past. However, overall aa sequence identity between these bat CoV ORFs was very low with maximally 28.2%. As shown in **Figure 6**, only the central part of these ORFs contained a stretch of 46 more conserved aa residues showing up to 39.1% sequence identity and 47.8% similarity (Blosum62 matrix). The origin and function of the divergent *ORF8* thus remain to be determined.

352 **Discussion**

353 We characterize highly diverse bat CoVs on a full genome level and show that these viruses
354 form one species together with HCoV-229E and a recently described virus from alpacas (32).
355 We analyze the genomic differences between human, bat and alpaca 229E-related CoVs to
356 elucidate potential host transitions during the formation of HCoV-229E.
357
358 A major difference between bat 229E-related CoVs and HCoV-229E was the *Spike* deletion
359 in HCoV-229E compared to the bat viruses. Interestingly, the bat 229E-related CoV lineage 1
360 which was phylogenetically most related to HCoV-229E also carried the smallest number of
361 additional aa residues. Most chiropteran CoVs are restricted to the gastrointestinal tract,
362 whereas HCoVs mainly replicate in the respiratory tract (2). The Spike deletion in HCoV-
363 229E compared to ancestral bat viruses is thus noteworthy, since deletions in this protein have
364 been associated with changes in coronaviral tissue tropism. This is best illustrated by TGEV,
365 whose full-length Spike variants are associated with a dual tropism for respiratory and enteric
366 tract, whereas the deleted variant termed *porcine respiratory CoV* (PRCV) mainly replicates
367 in the respiratory tract (53). One could hypothesize that adaptation of bat 229E-related CoV
368 lineage 1 to both non-chiropteran hosts and to respiratory transmission may have been easier
369 compared to the other bat 229E-related CoV lineages.
370 Because the exact aa residues of the HCoV-229E RBD conveying cell entry are not known, it
371 is difficult to predict whether the bat viruses may interact with the HCoV-229E cellular
372 receptor Aminopeptidase N (45) or its *Hipposideros* homologue. Characterization of this bat
373 molecule and identification of permissive cell culture systems may allow initial susceptibility
374 experiments for chimeric viruses. Of note, although the alpaca 229E-related CoV was
375 successfully isolated (32), no data on receptor usage and cellular tropism are available so far
376 (2, 53).

Another major difference was the existence of an *ORF8* downstream the *Nucleocapsid* gene in bat 229E-related viruses and the detection of putative sequence remnants of this ORF in HCoV-229E. Hypothetically, deterioration of *ORF8* in HCoV-229E could have occurred due to loss of gene function in human hosts after zoonotic transmission from bats or intermediate hosts. This may parallel gradual deletions in the SARS-CoV accessory *ORF8* during the human epidemic compared to bat SARS-related CoVs (54) and is consistent with characterizations of HCoV-229E clinical strains showing high variability of this genomic region (55).

The virus-host association between 229E-related CoVs and the bat genus *Hipposideros* is strengthened by our virus detections in *Hipposideros* species in Ghana and in Gabon (41), which is separated from Ghana by about 1,800 km. The observed link between 229E-related alphacoronaviruses and hipposiderid bats is paralleled by the detections of genetically closely related betacoronaviruses in different *Hipposideros* species from Ghana, Nigeria, Thailand and Gabon (33, 41, 52, 56), suggesting restriction of these CoVs to hipposiderid bat genera. Due to their proofreading capacity, CoVs show evolutionary rates of 10E-5 to 10E-6 substitutions per site per replication cycle, which is much slower than rates observed for other RNA viruses (57, 58). Our data thus suggest a long evolutionary history of 229E-related CoVs in Old World hipposiderid bats that greatly exceeds that of HCoV-229E in humans, confirming previous hypotheses from our group (33).

The putative role of the alpaca 229E-related CoV in the formation of HCoV-229E is unclear. Our data enable new insights into the evolutionary history of HCoV-229E. First, the alpaca 229E-related CoV contained an intact *ORF8* which was genetically related to the homologous gene in bat 229E-related CoVs. Second, genes of the alpaca CoV clustered either with bat viruses only or in intermediate position between bat viruses and HCoV-229E. Because the

alpaca 229E-related CoV showed the same deletion in its *Spike* gene as HCoV-229E compared to bat 229E-related CoVs, it may be possible that alpacas represent a first host switch from bats followed by a second inter-host transfer from alpacas to humans. The relatedness of the alpaca 229E-related CoV to older HCoV-229E strains rather than to contemporary ones reported by Crossley et al. would be compatible with this scenario (32). However, the alpaca 229E-related CoV was reported only from captive animals in the U.S. and whether this virus is indeed endemic in New World alpacas is unclear. Additionally, the apparent intra-*Spike* recombination event may speak against a role of the alpaca virus as the direct ancestor of HCoV-229E. Further analyses will be required to confirm this putative recombination event, ideally including additional sequence information from old HCoV-229E strains. Furthermore, a hypothetical direct transfer of Old World bat viruses to New World alpacas appears geographically unfeasible. It would be highly relevant to investigate Old World camelids for 229E-related CoVs that may have been passed on to captive alpacas and that may represent direct ancestors of HCoV-229E.

Additional constraints to consider in the hypothetical role of camelids for the evolutionary history of 229E-related CoVs is the time and place of putative host switches from bats. Camels were likely introduced to Africa not earlier than 5,000 years ago from the Arabian Peninsula (59, 60) and could not possibly come into direct contact with West African *H.* cf. *ruber* or *H. abae* of the Guinean savanna. The majority of CoV species seems to be confined to host genera (2). Therefore, it may be possible that 229E-related CoV transmission was mediated through closely related species like *H. tephrus*, which occurs in the Sahel zone and comes into contact to populations of *H.* cf. *ruber* distantly related to those from the Guinean savanna (61). This bat species should be analyzed for 229E-related CoVs together with other genera of the family Hipposideridae, like *Asellia* or *Triaenops*, which are desert-adapted bats sharing their habitat with camelids both in Arabia and Africa and may harbor genetically related CoVs. An important parallel to this evolutionary scenario is the role of camelids for

429 the emerging MERS-CoV (30, 62), whose likely ancestors also occur in bats (20, 21).
430 However, we cannot rule out that the alpaca 229E-related CoV and HCoV-229E represent
431 two independent zoonotic acquisitions from 229E-related CoVs existing in hipposiderid bats
432 and potentially yet unknown intermediate hosts.
433
434 The existence of different serotypes in the expanded 229E-related CoV species is unclear.
435 CoV neutralization is mainly determined by antibodies against the S protein, and particularly
436 the S1 domain (63). The phylogenetic relatedness of the S1 domains from the alpaca 229E-
437 related CoV and HCoV-229E suggests that these viruses form one serotype. The most closely
438 related bat 229E-related CoV lineage showed 8.4% aa sequence distance in the translated
439 *Spike* gene from HCoV-229E. This was comparable to the 7.8-18.6% aa distance between
440 FIPV, TGEV und *canine CoV*, which belong to one CoV species (*Alphacoronavirus 1*) and
441 for which cross-neutralization was observed (64). The about 30% Spike aa sequence distance
442 between the other bat 229E-related lineages and HCoV-229E were comparable to the distance
443 between HCoV-NL63 and HCoV-229E, which form two different serotypes (65). HCoV-
444 229E thus likely forms one serotype that includes the alpaca 229E- and potentially the most
445 closely related bat 229E-related lineage, while the other bat 229E-related lineages may form
446 different serotypes. In our study, lack of bat sera and absence of bat 229E-related CoV
447 isolates prevented serological investigations. The generation of pseudotyped viruses carrying
448 bat 229E-related *Spike* motifs may allow future serological studies. Of note, our joint analyses
449 of Ghanaian patients with respiratory disease in this study and previous work from our group
450 investigating Ghanaian villagers (66) showed that Ghanaians were infected with the globally
451 circulating HCoV-229E, whereas no evidence of bat 229E-related CoV infecting humans was
452 found. If serotypes existed in 229E-related CoVs, serologic studies may thus aid to elucidate
453 putative exposure of humans and potential camelid intermediate hosts to these bat viruses.

454 It should be noted that throughout Africa, bats are consumed as wild game (67) and humans
455 frequently live in close proximity of bat caves (68), including usage of bat guano as fertilizer
456 and drinking water from these caves (21). These settings potentially facilitate the exposure of
457 humans and their peri-domestic animals, including camelids, to these previously remote bat
458 viruses.
459 In summary, HCoV-229E may be a paradigmatic example of the successful introduction of a
460 bat CoV into the human population, possibly with camelids as intermediate hosts.
461

462 **Acknowledgement**

463 We thank Monika Eschbach-Bludau, Sebastian Brünink, Tobias Bleicker, Fabian Ebach and
464 Thierno Diawo Dallo at the Institute of Virology, Bonn, for technical assistance and Ebenezer
465 Kofi Badu, Priscilla Anti, Olivia Agbenyega, Florian Gloza-Rausch, Stefan Klose and
466 Thomas Kruppa for their help during the organization and conducting of field work. For all
467 capturing, sampling, and exportation of bat specimens, we obtained permission from the
468 respective countries' authorities. This study was supported by the European Union FP7
469 projects EMPERIE (contract number 223498) and ANTIGONE (contract number 278976)
470 and by the German Research Foundation (DFG grants DR 772/3-1, KA1241/18-1 and DR
471 772/7-1, TH 1420/1-1).

References

1. **de Groot RJ, Baker, S.C., Baric, R., Enjuanes, L., Gorbalenya, A.E., Holmes, K.V., Perlman, S., Poon, L., Rottier, P.J.M.,, Talbot PJ, Woo, P.C.Y. and Ziebuhr, J.** 2012. Family *Coronaviridae*, p 806-820. *In* King A, Adams M, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press, Amsterdam; Boston.
2. **Drexler JF, Corman VM, Drosten C.** 2014. Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. Antiviral Res **101:**45-56.
3. **van der Hoek L.** 2007. Human coronaviruses: what do they cause? Antivir Ther **12:**651-658.
4. **Walsh EE, Shin JH, Falsey AR.** 2013. Clinical impact of human coronaviruses 229E and OC43 infection in diverse adult populations. J Infect Dis **208:**1634-1642.
5. **Berkley JA, Munywoki P, Ngama M, Kazungu S, Abwao J, Bett A, Lassauniere R, Kresfelder T, Cane PA, Venter M, Scott JA, Nokes DJ.** 2010. Viral etiology of severe pneumonia among Kenyan infants and children. JAMA **303:**2051-2057.
6. **Gaunt ER, Hardie A, Claas EC, Simmonds P, Templeton KE.** 2010. Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method. J Clin Microbiol **48:**2940-2947.
7. **Esper F, Ou Z, Huang YT.** 2010. Human coronaviruses are uncommon in patients with gastrointestinal illness. J Clin Virol **48:**131-133.
8. **Paloniemi M, Lappalainen S, Vesikari T.** 2015. Commonly circulating human coronaviruses do not have a significant role in the etiology of gastrointestinal infections in hospitalized children. J Clin Virol **62:**114-117.

498 9. **Risku M, Lappalainen S, Rasanen S, Vesikari T.** 2010. Detection of human
499 coronaviruses in children with acute gastroenteritis. J Clin Virol **48:**27-30.
500 10. **de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, Fouchier**
501 **RA, Galiano M, Gorbalenya AE, Memish Z, Perlman S, Poon LL, Snijder EJ,**
502 **Stephens GM, Woo PC, Zaki AM, Zambon M, Ziebuhr J.** 2013. Middle East
503 Respiratory Syndrome Coronavirus (MERS-CoV); Announcement of the Coronavirus
504 Study Group. J Virol **87:**7790-7792.
505 11. **Drexler JF, Geipel A, Konig A, Corman VM, van Riel D, Leijten LM, Bremer**
506 **CM, Rasche A, Cottontail VM, Maganga GD, Schlegel M, Muller MA, Adam A,**
507 **Klose SM, Carneiro AJ, Stocker A, Franke CR, Gloza-Rausch F, Geyer J, Annan**
508 **A, Adu-Sarkodie Y, Oppong S, Binger T, Vallo P, Tschapka M, Ulrich RG,**
509 **Gerlich WH, Leroy E, Kuiken T, Glebe D, Drosten C.** 2013. Bats carry pathogenic
510 hepadnaviruses antigenically related to hepatitis B virus and capable of infecting
511 human hepatocytes. Proc Natl Acad Sci U S A **110:**16151-16156.
512 12. **Drexler JF, Corman VM, Muller MA, Lukashev AN, Gmyl A, Coutard B, Adam**
513 **A, Ritz D, Leijten LM, van Riel D, Kallies R, Klose SM, Gloza-Rausch F, Binger**
514 **T, Annan A, Adu-Sarkodie Y, Oppong S, Bourgarel M, Rupp D, Hoffmann B,**
515 **Schlegel M, Kummerer BM, Kruger DH, Schmidt-Chanasit J, Setien AA,**
516 **Cottontail VM, Hemachudha T, Wacharapluesadee S, Osterrieder K,**
517 **Bartenschlager R, Matthee S, Beer M, Kuiken T, Reusken C, Leroy EM, Ulrich**
518 **RG, Drosten C.** 2013. Evidence for novel hepaciviruses in rodents. PLoS Pathog
519 **9:**e1003438.
520 13. **Drexler JF, Corman VM, Muller MA, Maganga GD, Vallo P, Binger T, Gloza-**
521 **Rausch F, Rasche A, Yordanov S, Seebens A, Oppong S, Adu Sarkodie Y,**
522 **Pongombo C, Lukashev AN, Schmidt-Chanasit J, Stocker A, Carneiro AJ, Erbar**
523 **S, Maisner A, Fronhoffs F, Buettner R, Kalko EK, Kruppa T, Franke CR, Kallies**

R, Yandoko ER, Herrler G, Reusken C, Hassanin A, Kruger DH, Matthee S, Ulrich RG, Leroy EM, Drosten C. 2012. Bats host major mammalian paramyxoviruses. Nat Commun **3:**796.

14. **Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, Delicat A, Paweska JT, Gonzalez JP, Swanepoel R.** 2005. Fruit bats as reservoirs of Ebola virus. Nature **438:**575-576.

15. **Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T.** 2006. Bats: important reservoir hosts of emerging viruses. Clin Microbiol Rev **19:**531-545.

16. **Quan PL, Firth C, Conte JM, Williams SH, Zambrana-Torrelio CM, Anthony SJ, Ellison JA, Gilbert AT, Kuzmin IV, Niezgoda M, Osinubi MO, Recuenco S, Markotter W, Breiman RF, Kalemba L, Malekani J, Lindblade KA, Rostal MK, Ojeda-Flores R, Suzan G, Davis LB, Blau DM, Ogunkoya AB, Alvarez Castillo DA, Moran D, Ngam S, Akaibe D, Agwanda B, Briese T, Epstein JH, Daszak P, Rupprecht CE, Holmes EC, Lipkin WI.** 2013. Bats are a major natural reservoir for hepaciviruses and pegiviruses. Proc Natl Acad Sci U S A **110:**8194-8199.

17. **Wacharapluesadee S, Lumlertdacha B, Boongird K, Wanghongsa S, Chanhome L, Rollin P, Stockton P, Rupprecht CE, Ksiazek TG, Hemachudha T.** 2005. Bat Nipah virus, Thailand. Emerg Infect Dis **11:**1949-1951.

18. **Drexler JF, Gloza-Rausch F, Glende J, Corman VM, Muth D, Goettsche M, Seebens A, Niedrig M, Pfefferle S, Yordanov S, Zhelyazkov L, Hermanns U, Vallo P, Lukashev A, Muller MA, Deng H, Herrler G, Drosten C.** 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. J Virol **84:**11336-11349.

19. **Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C, Zhang YJ, Luo CM, Tan B, Wang N, Zhu Y, Crameri G,**

Zhang SY, Wang LF, Daszak P, Shi ZL. 2013. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature **503:**535-538.

20. **Ithete NL, Stoffberg S, Corman VM, Cottontail VM, Richards LR, Schoeman MC, Drosten C, Drexler JF, Preiser W.** 2013. Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. Emerg Infect Dis **19:**1697-1699.

21. **Annan A, Baldwin HJ, Corman VM, Klose SM, Owusu M, Nkrumah EE, Badu EK, Anti P, Agbenyega O, Meyer B, Oppong S, Sarkodie YA, Kalko EK, Lina PH, Godlevska EV, Reusken C, Seebens A, Gloza-Rausch F, Vallo P, Tschapka M, Drosten C, Drexler JF.** 2013. Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and Europe. Emerg Infect Dis **19:**456-459.

22. **Woo PC, Wang M, Lau SK, Xu H, Poon RW, Guo R, Wong BH, Gao K, Tsoi HW, Huang Y, Li KS, Lam CS, Chan KH, Zheng BJ, Yuen KY.** 2007. Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. J Virol **81:**1574-1585.

23. **Woo PC, Lau SK, Huang Y, Yuen KY.** 2009. Coronavirus diversity, phylogeny and interspecies jumping. Exp Biol Med (Maywood) **234:**1117-1127.

24. **Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, Mills JN, Timonin ME, Willis CKR, Cunningham AA, Fooks AR, Rupprecht CE, Wood JLN, Webb CT.** 2013. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proceedings of the Royal Society B: Biological Sciences **280**.

25. **Wolfe ND, Dunavan CP, Diamond J.** 2007. Origins of major human infectious diseases. Nature **447:**279-283.

26. **Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF,**

Yuen KY, Peiris JS, Poon LL. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science **302:**276-278.

27. **Graham RL, Baric RS.** 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol **84:**3134-3146.

28. **Chu D, Poon L, Gomaa MM, Shehata MM, Perera R, Abu Zeid D, El Rifay AS, Siu LY, Guan Y, Webby RW, Ali MA, Peiris M, Kayali G.** 2014. MERS Coronaviruses in Dromedary Camels, Egypt Emerg Infect Dis **20**.

29. **Muller MA, Corman VM, Jores J, Meyer B, Younan M, Liljander A, Bosch BJ, Lattwein E, Hilali M, Musa BE, Bornstein S, Drosten C.** 2014. MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983-1997. Emerg Infect Dis **20:**2093-2095.

30. **Reusken CB, Haagmans BL, Muller MA, Gutierrez C, Godeke GJ, Meyer B, Muth D, Raj VS, Smits-De Vries L, Corman VM, Drexler JF, Smits SL, El Tahir YE, De Sousa R, van Beek J, Nowotny N, van Maanen K, Hidalgo-Hermoso E, Bosch BJ, Rottier P, Osterhaus A, Gortazar-Schmidt C, Drosten C, Koopmans MP.** 2013. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. Lancet Infect Dis **13:**859-866.

31. **Crossley BM, Barr BC, Magdesian KG, Ing M, Mora D, Jensen D, Loretti AP, McConnell T, Mock R.** 2010. Identification of a novel coronavirus possibly associated with acute respiratory syndrome in alpacas (Vicugna pacos) in California, 2007. J Vet Diagn Invest **22:**94-97.

32. **Crossley BM, Mock RE, Callison SA, Hietala SK.** 2012. Identification and characterization of a novel alpaca respiratory coronavirus most closely related to the human coronavirus 229E. Viruses **4:**3689-3700.

33. **Pfefferle S, Oppong S, Drexler JF, Gloza-Rausch F, Ipsen A, Seebens A, Muller MA, Annan A, Vallo P, Adu-Sarkodie Y, Kruppa TF, Drosten C.** 2009. Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. Emerg Infect Dis **15:**1377-1384.

34. **Drexler JF, Kupfer B, Petersen N, Grotto RM, Rodrigues SM, Grywna K, Panning M, Annan A, Silva GF, Douglas J, Koay ES, Smuts H, Netto EM, Simmonds P, Pardini MI, Roth WK, Drosten C.** 2009. A novel diagnostic target in the hepatitis C virus genome. PLoS Med **6:**e31.

35. **Ronquist F, Huelsenbeck JP.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:**1572-1574.

36. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **29:**1969-1973.

37. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:**2731-2739.

38. **Simmonds P.** 2012. SSE: a nucleotide and amino acid sequence analysis platform. BMC Res Notes **5:**50.

39. **Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol **73:**152-160.

40. **Dijkman R, Jebbink MF, Wilbrink B, Pyrc K, Zaaijer HL, Minor PD, Franklin S, Berkhout B, Thiel V, van der Hoek L.** 2006. Human coronavirus 229E encodes a single ORF4 protein between the spike and the envelope genes. Virol J **3:**106.

41. **Maganga GD, Bourgarel M, Vallo P, Dallo TD, Ngoagouni C, Drexler JF, Drosten C, Nakoune ER, Leroy EM, Morand S.** 2014. Bat distribution size or shape as determinant of viral richness in african bats. PLoS One **9:**e100172.

42. **Thiel V, Herold J, Schelle B, Siddell SG.** 2001. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. J Gen Virol **82:**1273-1281.

43. **Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF.** 2014. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. J Virol **88:**11297-11303.

44. **Breslin JJ, Mork I, Smith MK, Vogel LK, Hemmila EM, Bonavia A, Talbot PJ, Sjostrom H, Noren O, Holmes KV.** 2003. Human coronavirus 229E: receptor binding domain and neutralization by soluble receptor at 37 degrees C. J Virol **77:**4435-4438.

45. **Hofmann H, Simmons G, Rennekamp AJ, Chaipan C, Gramberg T, Heck E, Geier M, Wegele A, Marzi A, Bates P, Pohlmann S.** 2006. Highly conserved regions within the spike proteins of human coronaviruses 229E and NL63 determine recognition of their respective cellular receptors. J Virol **80:**8639-8652.

46. **Tao Y, Tang K, Shi M, Conrardy C, Li KS, Lau SK, Anderson LJ, Tong S.** 2012. Genomic characterization of seven distinct bat coronaviruses in Kenya. Virus Res **167:**67-73.

47. **Lau SK, Li KS, Tsang AK, Shek CT, Wang M, Choi GK, Guo R, Wong BH, Poon RW, Lam CS, Wang SY, Fan RY, Chan KH, Zheng BJ, Woo PC, Yuen KY.** 2012. Recent transmission of a novel alphacoronavirus, bat coronavirus HKU10, from Leschenault's rousettes to pomona leaf-nosed bats: first evidence of interspecies

transmission of coronavirus between bats of different suborders. J Virol **86:**11906-11918.

48. **Williams GD, Chang RY, Brian DA.** 1999. A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication. J Virol **73:**8349-8355.

49. **Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Honigschmid P, Schafferhans A, Roos M, Bernhofer M, Richter L, Ashkenazy H, Punta M, Schlessinger A, Bromberg Y, Schneider R, Vriend G, Sander C, Ben-Tal N, Rost B.** 2014. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res **42:**W337-343.

50. **Krogh A, Larsson B, von Heijne G, Sonnhammer EL.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol **305:**567-580.

51. **Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ.** 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc **10:**845-858.

52. **Quan PL, Firth C, Street C, Henriquez JA, Petrosov A, Tashmukhamedova A, Hutchison SK, Egholm M, Osinubi MO, Niezgoda M, Ogunkoya AB, Briese T, Rupprecht CE, Lipkin WI.** 2010. Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. MBio **1**.

53. **Sanchez CM, Gebauer F, Sune C, Mendez A, Dopazo J, Enjuanes L.** 1992. Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. Virology **190:**92-105.

54. **Anonymous.** 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science **303:**1666-1669.

674 55. **Farsani SM, Dijkman R, Jebbink MF, Goossens H, Ieven M, Deijs M,**
675 **Molenkamp R, van der Hoek L.** 2012. The first complete genome sequences of
676 clinical isolates of human coronavirus 229E. Virus Genes **45:**433-439.
677 56. **Gouilh MA, Puechmaille SJ, Gonzalez JP, Teeling E, Kittayapong P,**
678 **Manuguerra JC.** 2011. SARS-Coronavirus ancestor's foot-prints in South-East Asian
679 bat colonies and the refuge theory. Infect Genet Evol **11:**1690-1702.
680 57. **Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S,**
681 **Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR.** 2010. Infidelity of
682 SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete
683 genome sequencing. PLoS Pathog **6:**e1000896.
684 58. **Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR.** 2007. High fidelity of murine
685 hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J Virol
686 **81:**12135-12144.
687 59. **Mburu DN, Ochieng JW, Kuria SG, Jianlin H, Kaufmann B, Rege JE, Hanotte**
688 **O.** 2003. Genetic diversity and relationships of indigenous Kenyan camel (Camelus
689 dromedarius) populations: implications for their classification. Anim Genet **34:**26-32.
690 60. **Epstein H, Mason IL.** 1971. The origin of the domestic animals of Africa. Africana
691 Pub. Corp.
692 61. **Vallo P, Guillén-Servent A, Benda P, Pires DB, Koubek P.** 2008. Variation of
693 mitochondrial DNA in the Hipposideros caffer complex (Chiroptera: Hipposideridae)
694 and its taxonomic implications. Acta Chiropterologica **10:**193-206.
695 62. **Perera RA, Wang P, Gomaa MR, El-Shesheny R, Kandeil A, Bagato O, Siu LY,**
696 **Shehata MM, Kayed AS, Moatasim Y, Li M, Poon LL, Guan Y, Webby RJ, Ali**
697 **MA, Peiris JS, Kayali G.** 2013. Seroepidemiology for MERS coronavirus using
698 microneutralisation and pseudoparticle virus neutralisation assays reveal a high

prevalence of antibody in dromedary camels in Egypt, June 2013. Eurosurveillance **18:**8-14.

63. **Meyer B, Drosten C, Muller MA.** 2014. Serological assays for emerging coronaviruses: challenges and pitfalls. Virus Res **194:**175-183.

64. **Horzinek MC, Lutz H, Pedersen NC.** 1982. Antigenic relationships among homologous structural polypeptides of porcine, feline, and canine coronaviruses. Infect Immun **37:**1148-1155.

65. **Hofmann H, Pyrc K, van der Hoek L, Geier M, Berkhout B, Pohlmann S.** 2005. Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. Proc Natl Acad Sci U S A **102:**7988-7993.

66. **Owusu M, Annan A, Corman VM, Larbi R, Anti P, Drexler JF, Agbenyega O, Adu-Sarkodie Y, Drosten C.** 2014. Human coronaviruses associated with upper respiratory tract infections in three rural areas of ghana. PLoS One **9:**e99782.

67. **Mickleburgh S, Waylen K, Racey P.** 2009. Bats as bushmeat: a global review. Oryx **43:**217-234.

68. **Anti P, Owusu M, Agbenyega O, Annan A, Badu EK, Nkrumah EE, Tschapka M, Oppong S, Sarkodie YA, Drosten C.** 2015. Human–Bat Interactions in Rural West Africa. Emerg Infect Dis **21**.

69. **da Silva Filho LV, Zerbinati RM, Tateno AF, Boas LV, de Almeida MB, Levi JE, Drexler JF, Drosten C, Pannuti CS.** 2012. The differential clinical impact of human coronavirus species in children with cystic fibrosis. J Infect Dis **206:**384-388.

70. **Chibo D, Birch C.** 2006. Analysis of human coronavirus 229E spike and nucleoprotein genes demonstrates genetic drift between chronologically distinct strains. J Gen Virol **87:**1203-1208.

Figure legends

Figure 1. Phylogenetic relationships of the genus *Alphacoronavirus,* HCoV-229E strains and the novel bat viruses

A, Bayesian phylogeny of an 816 nucleotide *RdRp* gene sequence fragment corresponding to positions 13,891-14,705 in HCoV-229E prototype strain inf-1 (GenBank accession no. NC002645) using a GTR+G+I substitution model. SARS-coronavirus (CoV) was used as an outgroup. Viruses with additional sequence information generated in this study were marked with circles (full genome) or marked with triangles (*Spike* gene). Bat viruses detected in our previous studies from Ghana (33) and Gabon are given in cyan (41). B, Neighbour-joining phylogeny of the same *RdRp* gene fragment with a nucleotide percentage distance substitution model and the complete deletion option. The tree was rooted against HCoV-NL63. Viruses were coloured according to their origin. C. Bayesian phylogeny of the full *Spike* gene of bat 229E-related CoVs, the alpaca 229E-related CoV and HCoV-229E strains identified with GenBank accession numbers and year of isolation, using a WAG amino acid substitution model and HCoV-NL63 as an outgroup. The novel bat 229E–related CoVs are shown in boldface and red. Branches leading to the outgroup were truncated for graphical reasons as indicated by slashed lines. Values at nodes show support of grouping from posterior probabilities or 1,000 bootstrap replicates (only values above 0.7 were shown).

Figure 2. Genome organization of 229E-related coronaviruses and relationships between viruses from bats and humans

A, 229E-related CoV genomes represented by black lines; ORFs are indicated by grey arrows. Locations of transcription-regulatory core sequences (TRS) are marked by black dots. HCoV-NL63 is shown for comparison. B, Similarity plots generated using SSE V1.1 (38) using a sliding window of 400 and a step size of 40 nucleotides (nt). The HCoV-229E prototype strain inf-1 was used with animal viruses identified in the legend.

Journal of Virology

Figure 3. Bayesian phylogenies of major open reading frames and recombination analysis of HCoV-229E and related animal viruses

A, Phylogenies were calculated with a WAG amino acid substitution model. The novel bat viruses are shown in red. The alpaca CoV is shown in cyan. Filled circles, posterior probability support exceeding 0.95, scale bar corresponds to genetic distance. Details on the origin of HCoV-229E strain VFC408 which was generated for this study can be retrieved from (69). Branches leading the outgroup HCoV-NL63 were truncated for graphical reasons. B, Bootscan analysis using the Jukes-Cantor algorithm with a sliding window of 1,500 and a step size of 300 nt. The HCoV-220E inf-1 strain was used with animal 229E-related viruses as identified in the legend. C. Phylogenies of the S1 and S2 subunit were calculated according to A. One representative HCoV-229E strain was selected per decade according to (70); GenBank accession nos. DQ243974, DQ243964, DQ243984, DQ243967.

Figure 4. Amino acid sequence alignment of the 5'-end of the *Spike* gene of HCoV-229E and related animal viruses

Amino acid alignment of the first part of the *Spike* gene of 229E-related CoVs including four bat 229E-related CoVs, the alpaca 229E-related CoV and the HCoV-229E inf-1 strain. Conserved amino acid residues are marked in black, sequence gaps are represented by hyphens.

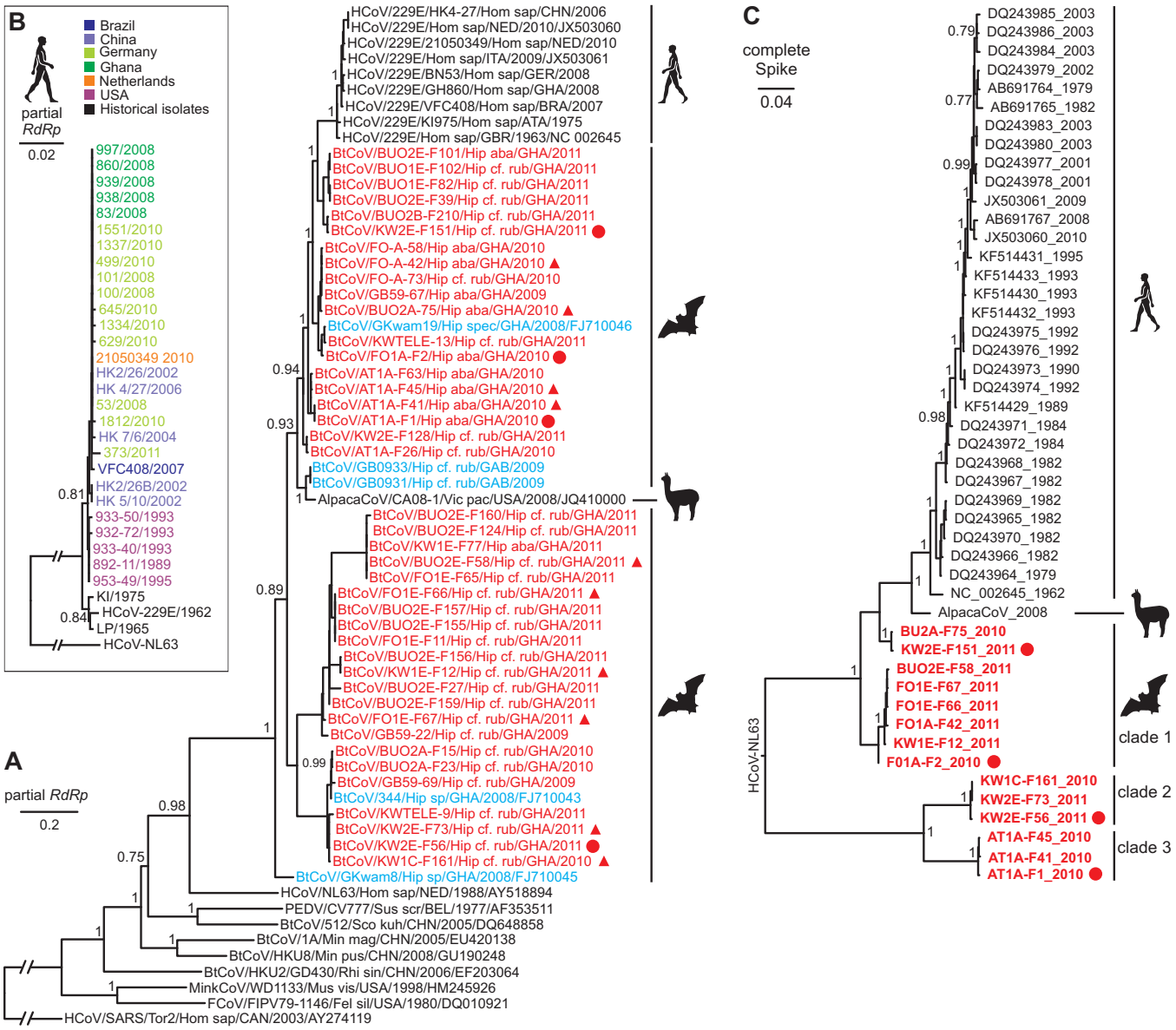
Figure 5. Nucleotide sequence alignment of the genomic 3'-end of HCoV-229E and related animal viruses

Nucleotide alignment of the genome region downstream the *Nucleocapsid* gene including four bat 229E-related CoV, the alpaca 229E-related CoV and representative HCoV-229E full genomes identified with GenBank accession number or strain name. Dots represent identical

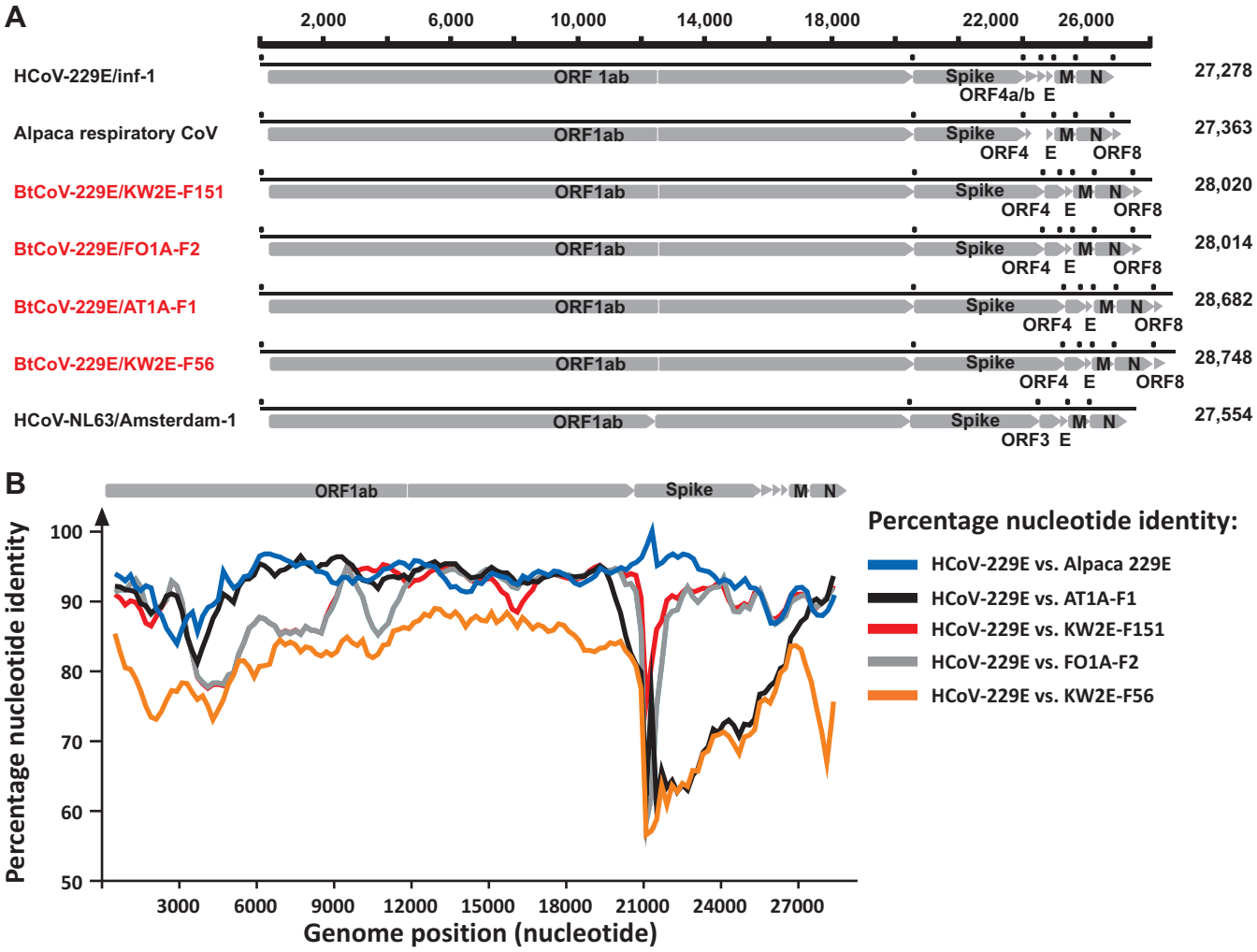
nucleotides, hyphens represent sequence gaps. Grey bars above alignments indicate open reading frames and the beginning of the poly-A tail. The putative start and stop codon of *ORF8* is labelled lime green, the corresponding putative TRS element is marked blue. The conserved genomic sequence elements and the highly conserved stem elements forming part of the pseudo-knot (PK) were marked with grey and purple background.

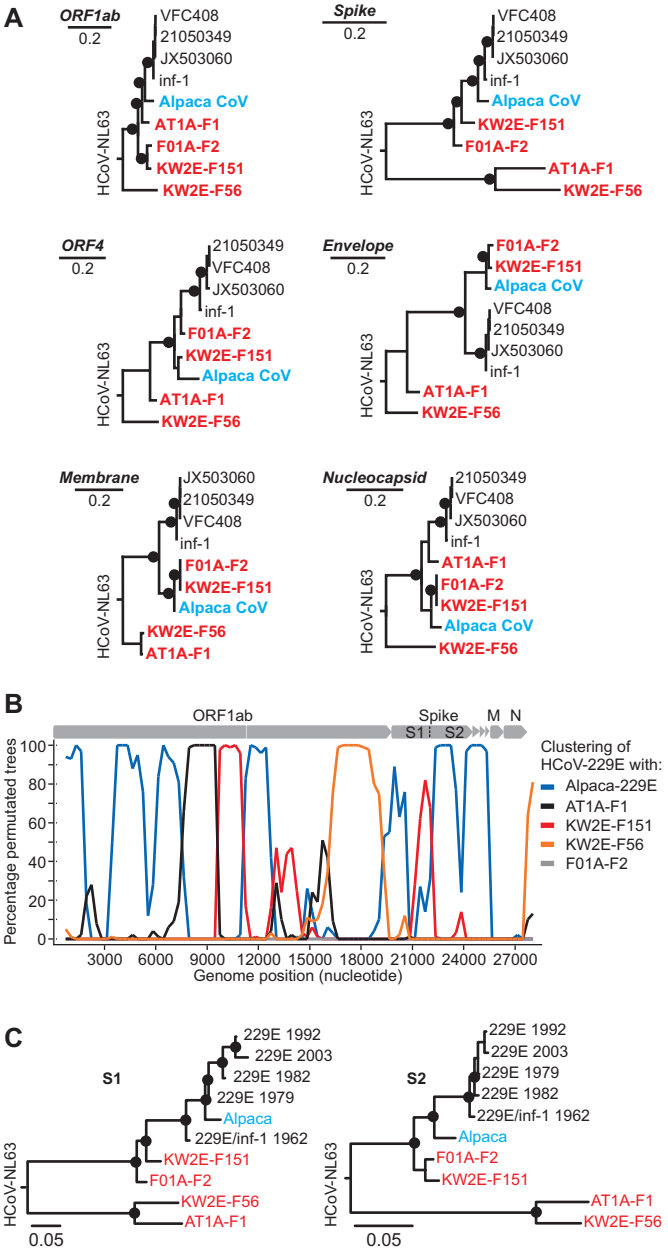
Figure 6. Amino acid sequence alignment of the putative *ORF8* from a bat 229E-related coronavirus and closest hits from two other hipposiderid bat coronaviruses

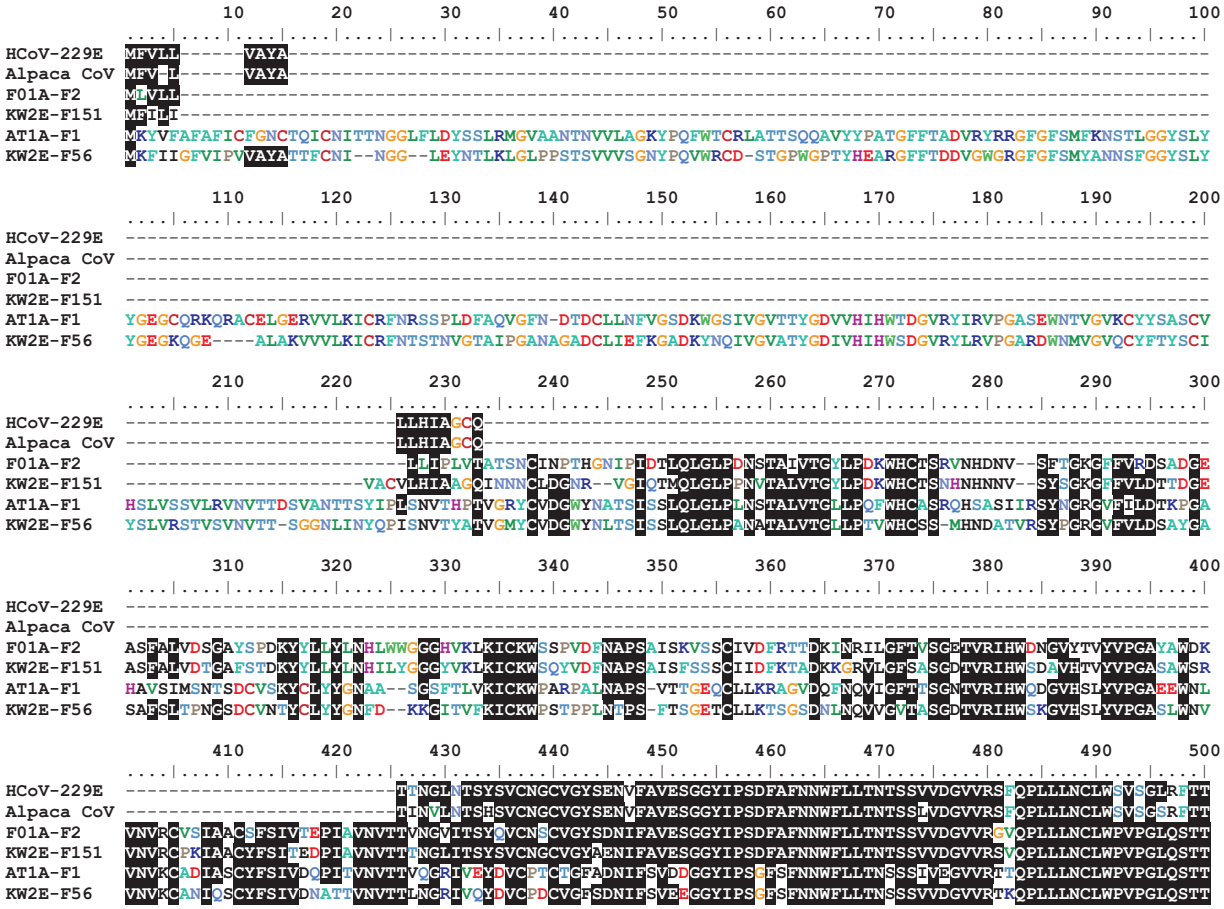
Conserved amino acid residues between sequence pairs are highlighted in color according to amino acid properties, sequence gaps are represented by hyphens. The central domain showing higher sequence similarity between compared viruses is boxed for clarity. The 229E-related alphacoronavirus KW2E-F56 from a *Hipposideros* cf. *ruber* detected in this study is given in red, the alphacoronavirus HKU10 originated from a Chinese *H. pomona*, the betacoronavirus Zaria originated from a Nigerian *H. gigas*.

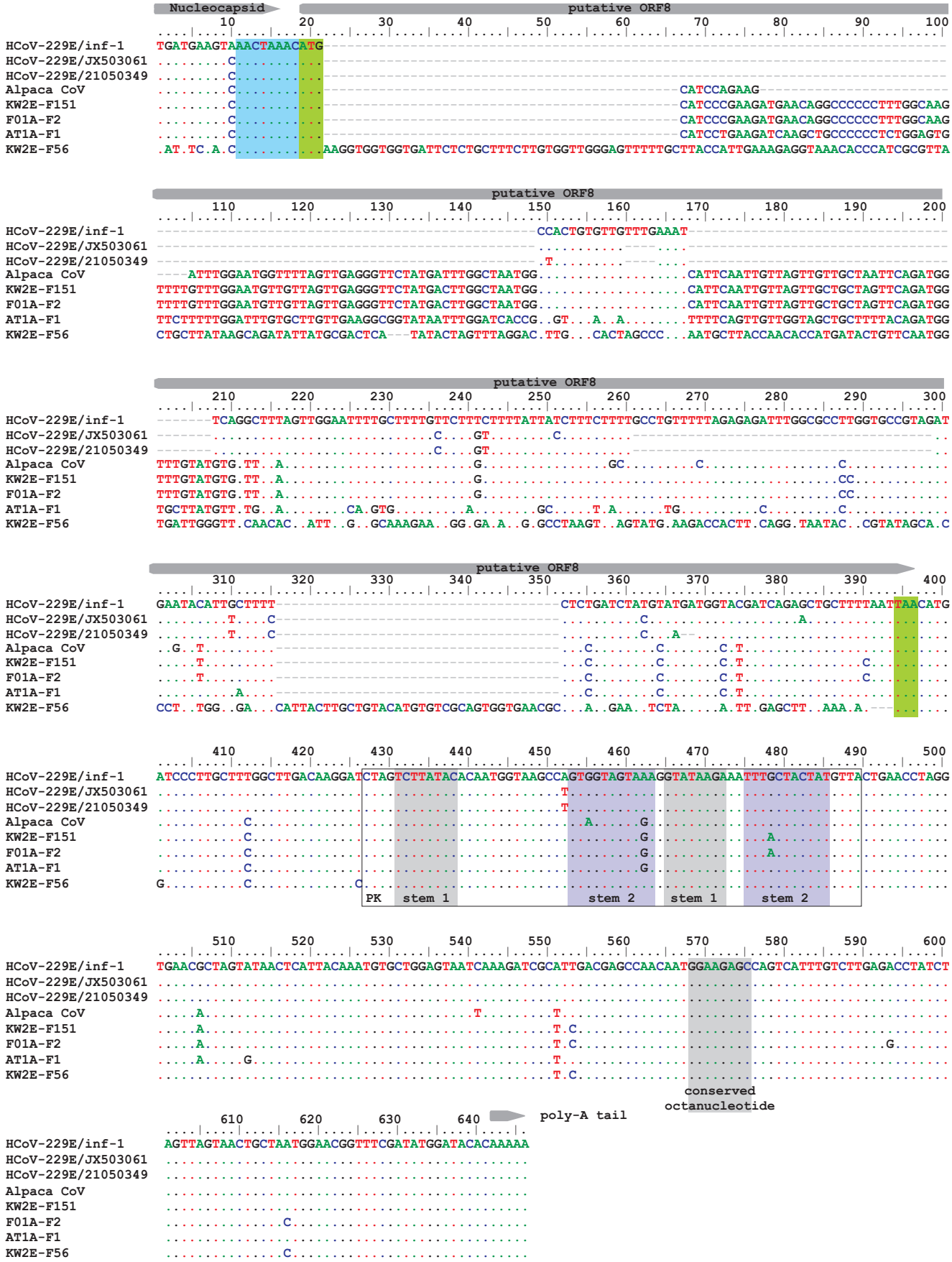


Journal of Virology









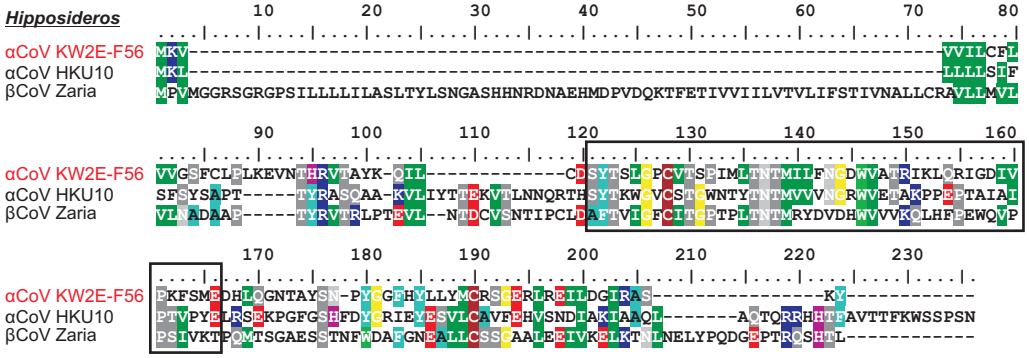


Table 1. Overview of bats tested for 229E-related coronaviruses in Ghana

Species	*n*	Positives (%)
Coleura afra	68	0
Hipposideros abae	242	19 (7.8)
H. cf. *gigas*	12	0
H. jonesi	5	0
H. cf. *ruber*	1611	62 (3.8)
Nycteris cf. *gambiensis*	91	0
Rhinolophus alcyone	4	0
R. landeri	9	0
Taphozous perforatus	21	0
Lissonycteris angolensis	20	0
Rousettus aegyptiacus	4	0
Total	**2,087**	**81 (3.9)**

Table 2. Coding capacity for the putative non-structural proteins of the novel bat 229E-related coronaviruses

	KW2E-F151		F01A-F2		AT1A-F1		KW2E-F56	
	1st to last amino acid	Protein size	1st to last amino acid	Protein size	1st to last amino acid	Protein size	1st to last amino acid	Protein size
NSP1	Met^{1}-Gly^{111}	111	Met^{1}-Gly^{111}	111	Met^{1}-Gly^{111}	111	Met^{1}-Gly^{109}	109
NSP2	Asn^{112}-Gly^{897}	786	Asn^{112}-Gly^{897}	786	Asn^{112}-Gly^{897}	786	Asn^{110}-Gly^{895}	786
NSP3	Gly^{898}-Ala^{2494}	1597	Gly^{898}-Ala^{2494}	1597	Gly^{898}-Ala^{2492}	1595	Gly^{896}-Ala^{2489}	1594
NSP4	Gly^{2495}-Gln^{2975}	481	Gly^{2495}-Gln^{2975}	481	Gly^{2493}-Gln^{2973}	481	Gly^{2490}-Gln^{2970}	481
NSP5	Ala^{2976}-Gln^{3277}	302	Ala^{2976}-Gln^{3277}	302	Ala^{2974}-Gln^{3275}	302	Ala^{2971}-Gln^{3272}	302
NSP6	Ser^{3278}-Gln^{3556}	279	Ser^{3278}-Gln^{3556}	279	Ser^{3276}-Gln^{3553}	278	Ser^{3273}-Gln^{3551}	279
NSP7	Ser^{3557}-Gln^{3639}	83	Ser^{3557}-Gln^{3639}	83	Ser^{3554}-Gln^{3636}	83	Ser^{3552}-Gln^{3634}	83
NSP8	Ser^{3640}-Gln^{3834}	195	Ser^{3640}-Gln^{3834}	195	Ser^{3637}-Gln^{3831}	195	Ser^{3635}-Gln^{3829}	195
NSP9	Asn^{3835}-Gln^{3943}	109	Asn^{3835}-Gln^{3943}	109	Asn^{3832}-Gln^{3940}	109	Asn^{3830}-Gln^{3938}	109
NSP10	Ala^{3944}-Gln^{4078}	135	Ala^{3944}-Gln^{4078}	135	Ala^{3941}-Gln^{4075}	135	Ala^{3939}-Gln^{4073}	135
NSP11	Ser^{4079}-Glu^{4097}	19	Ser^{4079}-Glu^{4097}	19	Ser^{4076}-Glu^{4094}	19	Ser^{4074}-Glu^{4092}	19
NSP12	Ser^{4079}-Gln^{5005}	927	Ser^{4079}-Gln^{5005}	927	Ser^{4076}-Gln^{5002}	927	Ser^{4074}-Gln^{5000}	927
NSP13	Ala^{5006}-Gln^{5602}	597	Ala^{5006}-Gln^{5602}	597	Ala^{5003}-Gln^{5599}	597	Ala^{5001}-Gln^{5597}	597
NSP14	Ser^{5603}-Gln^{6120}	518	Ser^{5603}-Gln^{6120}	518	Ser^{5600}-Gln^{6117}	518	Ser^{5598}-Gln^{6115}	518
NSP15	Gly^{6121}-Gln^{6468}	348	Gly^{6121}-Gln^{6468}	348	Gly^{6118}-Gln^{6465}	348	Gly^{6116}-Gln^{6463}	348
NSP16	Ser^{6469}-Lys^{6768}	300	Ser^{6469}-Lys^{6768}	300	Ser^{6466}-Lys^{6766}	301	Ser^{6464}-Lys^{6763}	300

Table 3. Comparison of amino acid identities of seven conserved replicase domains of the bat 229E-related coronaviruses, HCoV-229E and the alpaca 229E-related coronavirus for species delineation

	Percentage amino acid sequence identity					
		Human Coronavirus 229E[a] vs.				
Domains	**within Bat 229E[b]**	**KW2E-F56**	**AT1A-F1**	**KW2-F151**	**F01A-F2**	**ACoV[c] vs Bat 229E[b]**
ADRP	75.6-100	75-75.6	91.1-92.9	84.5-85.1	84.5-85.1	76.8-90.5
NSP5(3CLpro)	90.7-100	90.4-90.7	97.4-97.7	96.4-96.7	97.4-97.7	90.4-97.4
NSP12 (RdRp)	97.5-100	95.7-96	97.3-97.6	96.9-97.3	97.2-97.7	97.3-98.9
NSP13 (NTPase/Hel)	97.2-100	96.5-97.2	97.2-97.8	97.3-98	98-98.7	97.8-99.3
NSP14 (ExoN/N7-MTase)	96.1-100	95-95.6	97.5-98.1	97.3-97.9	96.9-97.5	96.3-99.2
NSP15 (NendoU)	92.8-100	92.2	96.3-96.6	96.6-96.8	96.8-97.1	91.4-96.8
NSP16 (O-MT)	91.7-100	90.7-91	91.7-92	97.3-97	97.3-97.7	90.7 – 98.0
Concatenated domains	94.5-100	93.3-93.6	96.4-96.8	96.4-96.7	96.7-97.1	94.2-97.8

[a]including - HCoV 229E - Inf-1, HCoV 229E - 0349, HCoV 229E - J0304;

[b]including - Bat CoV KW2E-F56, AT1A-F1, KW2E-F151 and F01A-F2;

[c]ACoV - Alpaca Coronavirus

GenBank accession numbers of reference sequences: HCoV-229E - Inf-1: NC_002645.1; HCoV-229E - 0349: JX503060; HCoV-299E - J0304: JX503061; Alpaca CoV (ACoV): JQ410000

Table 4. Amino acid identity between open reading frames of human, bat and camelid 229-related coronaviruses

Percentage Amino Acid Sequence Identity							
	Human Coronavirus 229E[a] vs.						
	KW2E-F151	F01A-F2	AT1A-F1	KW2E-F56	AcoV	within Bat CoV[b]	ACoV[c] vs Bat CoV[b]
ORF1a	89.5 - 89.9	89.5 - 89.8	92.6 - 93.1	84.1 - 84.6	92.9 - 93.3	83.8 - 97.9	85.1 - 93.5
ORF1ab	92.5 - 92.9	92.6 - 93	94.2 - 94.6	88.3 - 88.8	94.6 - 95	88.7 - 98.3	89.3 - 95.2
Spike	87.5 - 91.6	87.4 - 91.4	67.2 - 68.9	67.2 - 69.1	92.8 - 94.4	66.8 - 92.4	69.7 - 90.8
ORF4	92.4 - 93.1	92.6 - 93.2	77.3 - 78.8	71.2 - 73.6	79.7 - 78.1	75.7 - 96.4	67.2 - 82.8
Envelope	89.6 - 90.9	89.6 - 90.9	77.6 - 78.9	78.7 - 80	89.6 - 90.9	77.3 - 98.7	77.3 - 100
Membrane	90.2 - 90.7	89.3 - 89.9	86.2 - 86.7	87.1 - 87.6	89.8 - 90.2	86.7 - 98.7	86.3 - 99.1
Nucleocapsid	90.7 - 92	90.2 - 91.5	88.6 - 90.4	75.8 - 76.6	88.4 - 89.7	78.7 - 99.5	78.2 - 94
ORFX/8	-	-	-	-	-	12.5 - 100	15.2 - 83.9

[a]including - HCoV 229E - Inf-1, HCoV 229E - 0349, HCoV 229E - J0304

[b]including - Bat CoV KW2E-F56, AT1A-F1, KW2E-F151, F01A-F2;

[c]ACoV - Alpaca Coronavirus

Table 5. Putative transcription regulatory sequences of the novel bat 229E-related coronaviruses and HCoV-229E

	HCoV-229E/ inf-1	KW2E-F151	F01A-F2	AT1A-F1	KW2E-F56
Leader	(62) UCUCAACUAAACN$_{220}$ (293) AUG	(62) UCUCAACUAAACN$_{220}$ (293) AUG	(62) UCUCAACUAAACN$_{220}$ (293) AUG	(62) UCUCAACUAAACN$_{220}$ (293) AUG	(62) UCUCAACUAAACN$_{220}$ (293) AUG
Spike	(20571) UCUCAACUAAAUAA A (20586) AUG	(20585) UCUCAACUAAAUAA A (20600) AUG	(20585) UCUCAACUAAAUAA A (20600) AUG	(20576) UCUCAACUAAAAA (20589) AUG	(20570) UCUCAACUAAGUA (20583) AUG
ORF4	(24054) UCAACUAAAN$_{38}$ (24101) AUG	(24644) UCAACUAAACN$_{38}$ (24691) AUG	(24638) UCAACUAAACN$_{38}$ (24685) AUG	(25290) UCAACUAAACN$_{38}$ (25337) AUG	(25258) UCAACUAAACN$_{38}$ (25304) AUG
Envelope	(24599) UCUCAACUAAN$_{152}$ (24762) AUG	(25190) UCUCAACUAACN$_{149}$ (25349) AUG	(25184) UCUCAACUAACN$_{149}$ (25343) AUG	(25836) UCUCAACUAACN$_{149}$ (25992) AUG	(25805) UCAACUAACN$_{131}$ (25962) AUG
Membrane	(24991) UCUAAACUAAACG ACA (25007) AUG	(25578) UCUAAACUAAACGA CA (25594) AUG	(25572) UCUAAACUAAACGA CA (25588) AUG	(26224) UCUAAACUAAACG (26237) AUG	(26185) UCUAAACUAAACG (26198) AUG
Nucleocapsid	(25680) UCUAAACUGAACGA AAAG (25698) AUG	(26270) UCUAAACUGAACGA AAAG (26288) AUG	(26264) UCUAAACUGAACGA AAAG (26282) AUG	(26934) UCUAAACUGAACGA AAACC (26953) AUG	(26874) UCUAAACUGAACGA AAACC (26893) AUG
ORF8		(27468) UCAACUAAAC (27478) AUG	(27462) UCAACUAAAC (27472) AUG	(28130) UCAACUAAAC (28141) AUG	(28124) UCAACUAAAC (28134) AUG

First bracket: Genome position of the first residue of the putative TRS sequence, second bracket: genome position of the first base of the start codon; $N_{lower\ case}$: number of base residues between end of the putative TRS sequence and start codon (where applicable)

Journal of Virology