

Human endemic coronavirus emergence in the context of past and recent zoonotic outbreaks

Diego Forni (✉ diego.forni@lanostrafamiglia.it)

IRCCS Eugenio Medea

Rachele Cagliani

Bioinformatics, Scientific Institute, IRCCS E. MEDEA

Federica Arrigoni

University of Milan-Bicocca

Martino Benvenuti

University of Milan-Bicocca

Chiara Pontremoli

Bioinformatics, Scientific Institute, IRCCS E. MEDEA <https://orcid.org/0000-0002-3022-502X>

Uberto Pozzoli

Bioinformatics, Scientific Institute, IRCCS E. MEDEA

Mario Clerici

University of Milan

Luca De Gioia

University of Milan-Bicocca

Manuela Sironi

Eugenio Medea

Article

Keywords: SARS-CoV-2 evolution, coronaviruses, viral disease emergence

Posted Date: January 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-134999/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Understanding the evolutionary dynamics of the four human endemic coronaviruses might provide insight into the future trajectories of SARS-CoV-2 evolution. We re-assessed the timing of endemic coronavirus emergence and we show that all viruses entered human populations in a time-frame ranging from ~500 to 55 years ago. Because the three highly pathogenic coronaviruses (SARS-CoV, MERS-CoV, and SARS-CoV-2) spilled-over in tight temporal succession, the pattern of coronavirus emergence, in analogy to that of influenza pandemics, is highly irregular. To contextualize this observation in a wider perspective of viral diseases emergence since 1945, we mined epidemiology database information. After controlling for reporting bias, we find that, contrary to widespread beliefs, the occurrence of viral diseases (either zoonotic or not) has not increased over the last decades. Analysis of the recent and ongoing evolution of HCoV-229E and HCoV-OC43 indicated that positive selection most likely contributed to fine-tune the interaction with the human interferon/inflammatory response. Conversely, integration of evolutionary inference and molecular dating provided evidence that these viruses are not undergoing antigenic drift and the temporal emergence of spike protein variants is best explained by optimization of receptor binding affinity. These data provide a fresh look on viral disease emergence and on coronavirus evolution.

Introduction

Coronaviruses (order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae*) are a diverse group of positive-sense, single stranded RNA viruses with high zoonotic potential¹⁻³. In 2002, a highly pathogenic coronavirus, severe acute respiratory syndrome coronavirus (SARS-CoV), spilled-over from palm civets to humans and caused ~8,000 cases in several countries⁴. These events were followed by the appearance, in 2012, of Middle East respiratory syndrome coronavirus (MERS-CoV), a camel-derived pathogen that caused multiple outbreaks of respiratory disease mainly in the Arabic Peninsula⁵. Containment and surveillance strategies allowed the control of these viruses, which have never (SARS-CoV) or only occasionally (MERS-CoV) reappeared in human populations⁶. However, at the end of 2019, a novel coronavirus, designated SARS-CoV-2 by the ICTV⁷, emerged in China and is now recognized as the cause of COVID-19⁸. The virus rapidly spread worldwide and the World Health Organization declared the SARS-CoV-2 pandemic in early March 2020. Most likely, SARS-CoV-2 originated and evolved in bats, eventually spilling over to humans, either directly or through an intermediate host⁹⁻¹⁵. To date, more than 72 million COVID-19 cases have been confirmed (<https://covid19.who.int/>, as of 18th December 2020), suggesting that, until an effective vaccination campaign is implemented, the virus will continue to circulate among people and, possibly, other animals¹⁶⁻¹⁹.

The epidemic behavior of SARS-CoV, MERS-CoV, and SARS-CoV-2, as well as their clinical severity, have clearly raised awareness on the potential danger posed by coronaviruses, which were considered relatively harmless to humans before 2002. In fact, four other coronaviruses (HCoV-OC43, HCoV-HKU1, HCoV-NL63, and HCoV-229E), sometimes referred to as “common cold coronaviruses”, have been

circulating in human populations for decades, usually causing mild symptoms^{2,20,21}. All these viruses are seasonal and generate short-term immunity, with reinfections being common within one year²²⁻²⁵.

Like the highly pathogenic coronaviruses, the endemic coronaviruses have a zoonotic origin^{2,3,26}. Phylogenetic analyses indicated that bats most likely represent the animal reservoirs from which the HCoV-NL63 and HCoV-229E alphacoronaviruses emerged²⁷⁻³⁰. It is presently unknown whether HCoV-NL63 was transmitted to humans via an intermediate host, as the most closely related viruses were detected in bats from Kenya²⁷. Conversely, viruses highly similar to HCoV-229E were identified in camelids (dromedary camels and alpacas), strongly suggesting that, in analogy to MERS-CoV, these animals represented the zoonotic source of human infection^{2,3,31-33}. The other two endemic coronaviruses, HCoV-OC43 and HCoV-HKU1, belong to the *Betacoronavirus* genus and most likely have their animal reservoirs in rodents^{2,3}. Whereas it is widely accepted that bovines were the intermediate hosts mediating the transmission of HCoV-OC43 to humans, the zoonotic source of HCoV-HKU1 is presently unknown^{2,3,26,34,35}. Given the commensal behavior of several rodents, it cannot be excluded that the virus was directly transmitted to our species by mice or related animals.

Although with some controversy³⁶, most previous estimates indicated that the endemic coronaviruses entered human populations in the last 1000 years^{2,28,30,34,35,37,38}. However, early analyses were often based on a small number of sequences and did not control for effects that are now recognized to affect molecular dating. Also, little is known about the past and ongoing selective events that accompanied the emergence and spread of endemic coronaviruses in human populations. In the wake of the COVID-19 pandemic, a better understanding of the evolutionary dynamics of endemic coronaviruses, as well as of the tendency of viral disease emergence, might provide valuable insight into the possible trajectories of SARS-CoV-2 evolution, especially in case the virus should become endemic.

Results

Time-frame of human endemic coronavirus emergence

As mentioned above, all endemic coronaviruses were estimated to have recently emerged as human pathogens^{2,28,30,34,35,37,38}. However, besides being generally based on a limited number of sequences, most previous analyses did not include some of the viruses that are now recognized to be closely related to endemic human coronaviruses (e.g., the dromedary camel alphacoronaviruses related to HCoV-229E). Also, it is now recognized that the presence of recombination, the lack of a temporal signal in the sequence data, and the pervasive effect of purifying selection can affect molecular dating³⁹⁻⁴¹. Accounting for these effects has become common practice only in recent years. Thus, we decided to reassess the timing of endemic coronavirus emergence and to date the time when circulating strains last shared a common ancestor. To this purpose, we retrieved all available sequences with known sampling date for the four coronaviruses (HCoV-OC43, n=167; HCoV-229E, n=31; HCoV-NL63, n=37; HCoV-HKU1,

n=68) (Supplementary Table 1). The animal viruses most closely related to each human coronavirus were also included in the analyses.

Because recombination is known to be frequent in all coronavirus genera⁴²⁻⁴⁴, we used 3SEQ to identify recombination events, which were detected in all datasets (Fig. 1)⁴⁵. Based on the location of breakpoint positions, we then selected the longest non-recombining region for each alignment. Specifically, we obtained relatively long regions for HCoV-OC43, HCoV-229E, and HCoV-HKU1, whereas two short regions (NL63_reg1 and NL63_reg2) were available for the HCoV-NL63 alignment (Fig.1). For all the selected non-recombining regions, maximum likelihood phylogenetic trees were constructed and we checked for the presence of a temporal signal by performing regression of root-to-tip genetic distances against sampling dates. These analyses indicated a strong temporal signal for all regions, with the exclusion of NL63_reg1 and of the HCoV-HKU1 long region (Fig. 1). In this latter case, the lack of a temporal signal is most likely due to the short time span among virus sampling dates, with the earliest sequences collected in 2003 (Supplementary Table 1).

Before performing molecular dating, we evaluated whether natural selection strongly affected branch length estimates in the viral phylogenies. In fact, it is now recognized that purifying selection and saturation effects contribute to the time-dependent substitution rate variation in viruses, which in turn affects molecular dating^{40,46}. We thus estimated branch lengths using the aBS-REL (adaptive branch-site random effects likelihood) model, which accounts for different selective pressures among lineages and is relatively robust to substitution saturation⁴⁷. For all phylogenies, branch lengths estimated with aBS-REL were comparable to those obtained with a GTR (General Time Reversible) model (Fig. 1), suggesting that molecular dating can be performed with minor effects related to the time dependency of substitution rates.

Thus, for the three phylogenies (HCoV-OC43, HCoV-229E, and NL63_reg2) showing a temporal signal, we used a Bayesian approach to estimate substitution rates and time-measured evolutionary histories. Substitution rates in the range of 1.6×10^{-4} to 1.5×10^{-3} were obtained, in line with previous analyses⁴³. For the HCoV-HKU1 phylogeny, date estimates were obtained by using the substitution rate of HCoV-OC43 (another betacoronavirus) as a prior. For the circulating strains of all coronaviruses, we obtained similar tMRCA (time to the most recent common ancestor) estimates, which ranged from 71 (HCoV-229E) to 55 (HCoV-HKU1) years ago (Fig. 2, Supplementary Fig. 1). The splits of human coronaviruses from their most closely related animal viruses were more variable. Specifically, we estimated that HCoV-OC43 split from the bovine coronavirus (BCoV) lineage around 1891 (IC:1876-1905), whereas HCoV-229E separated from the camel alphacoronavirus in the 18th century (1754, IC:1713-1791) (Fig. 2). Because bovines and camels are plausible zoonotic sources for the human infections, these split dates may be considered as good proxies for the time when HCoV-OC43 and HCoV-229E entered human populations. The long time span separating the split times and the tMRCA are most likely accounted for either by extinct ancestral lineages or by unsampled viral diversity. With respect to HCoV-HKU1 and HCoV-NL63, they were estimated to have diverged from the related rodent or bat viruses in 1651 (IC:1469-1793) and 1871

(IC:528-1984), respectively (Supplementary Fig. 1). However, confidence intervals for HCoV-NL63 were very large. This effect is most likely due to the short non-recombining region we used for dating and indicates that the split from the bat virus can be estimated with substantial uncertainty. Also, because there is no indication whether rodents and bats represented the hosts from which the spillovers occurred, the time when HCoV-NL63 and HCoV-HKU1 entered human populations remains highly uncertain. Nonetheless, these data support a recent introduction of endemic coronaviruses in human populations and indicate that all zoonotic transmissions occurred in a time frame ranging from ~500 to 55 years ago (Fig. 3).

Human coronavirus emergence in the context of viral outbreaks

The molecular dating analyses reported above indicate that all endemic human coronaviruses emerged as human pathogens earlier than 55 years ago, and most likely in a more distant past (Fig. 3). This implies that, at least between ~ 1965 and 2002 (when SARS-CoV appeared), no coronavirus gained the ability to spread widely in our species. Thus, the pattern of coronavirus emergence seems to be highly irregular and to have intensified in recent years. To compare the timing of coronavirus emergence to that of another respiratory virus, we recorded the occurrence of known influenza pandemics since 1500. As previously noted, this pattern is also irregular and there is no clear relationship between pandemic occurrence and human population size⁴⁸. Notably, influenza pandemics were also uncommon from the 70s to 2009 (when H1N1pdm09 caused a pandemic) (Fig. 3). Overall, these data are not in full agreement with the notion that emerging infectious diseases (EIDs) have progressively increased in frequency in the last decades, with a peak in the 80s (possibly secondary to AIDS diffusion)⁴⁹. However, this might be due to chance and to the fact that we only analyzed few viruses. We thus retrieved data from the original work that analyzed EID occurrence (from 1940 to 2004). Specifically, Jones and co-workers defined the timing of EIDs based on the first description of the original case(s) of a novel human infectious disease⁴⁹. We focused on viral EIDs and we separated zoonotic diseases that are not transmitted by an arthropod vector, vector-borne diseases, and non-zoonotic viruses. Results did not suggest a strong increasing trend with time, with the exclusion of a higher frequency of zoonoses in 1990-2000 (Fig. 4b,d,f).

We next explored the annual occurrence of viral disease outbreaks as recorded in GIDEON (Global Infectious Diseases and Epidemiology Network), the most comprehensive epidemiology database at a global scale⁵⁰. We removed viruses for which an effective vaccine has been developed from the total record of viral disease outbreaks, to avoid geographic and timing biases. Diseases that are mainly sexually transmitted (e.g., AIDS, hepatitis B) or whose spread was largely due to the use of human-contaminated material (e.g., AIDS, hepatitis C) were also excluded. This originated a dataset of 1252 outbreaks for 69 viral diseases. We considered either all outbreaks (that involved humans and/or animals) or outbreaks with at least 50 recorded human cases that occurred between 1945 and 2017 (to avoid border effects). For both arthropod-borne and non-arthropod-borne zoonoses (but not for non-zoonotic viruses), we observed a tendency to increase with time, which was particularly evident only since

~1995-2000 (Fig. 4a,c,e). However, these trends are likely to be strongly influenced by increased reporting efforts and the small number of outbreaks in the 40s and 50s is in line with previous observations that the number of discovered viruses abruptly increased around 1954, roughly corresponding to the advent of tissue culture techniques for virus detection⁵¹. As previously suggested⁴⁹, we thus used the annual number of articles published in the Journal of Infectious Diseases (JID) to account for this effect (Fig. 4g). After Bonferroni correction for multiple tests and controlling for reporting effort (see Materials and Methods), no significant increase with time was observed for zoonotic disease outbreaks (Fig. 4a). A significant decrease with time was instead observed for non-zoonotic diseases and for all outbreaks of arthropod-borne diseases (Fig. 4c,e).

Overall, these data suggest that the timing of large epidemics is episodic and seems to be unrelated from human population growth. In general, the number of outbreaks of zoonotic and non-zoonotic viral diseases has not increased significantly over the last 80 years.

Recent and ongoing evolution of HCoV-OC43 and HCoV-229E

To investigate the selective patterns acting on the coding regions of HCoV-OC43 and HCoV-229E since their separation from bovine/camel viruses, we applied gammaMap⁵², a method that combines analysis of within-population variation and divergence from an outgroup to estimate codon-wise selection coefficients (γ).

Coronaviruses have large and complex genomes which encode 16 non-structural (nsps) and four structural proteins (spike, envelope, membrane, and nucleoprotein), as well as a variable number of accessory molecules. Embecoviruses (e.g., HCoV-OC43, HCoV-HKU1, and BCoV) encode an additional structural protein, a hemagglutinin-esterase (HE) which serves as a receptor-destroying enzyme^{2, 53}. In line with data on several other viruses^{47, 54-56}, we found that most codons evolved under strong to moderate purifying selection ($\gamma < -5$) (Supplementary Fig. 2). However, sites with robust evidence of positive selection (posterior probability > 0.75 of $\gamma \geq 1$) could also be detected. The majority of these sites are located in a restricted number of proteins with mainly structural functions (Fig. 5, 6, and Supplementary Fig. 3). Whereas most selected sites in the spike proteins and in HE are polymorphic in circulating viral populations, those located in other regions are not (Supplementary Table 2). Importantly, the positively selected sites in the spike proteins of both HCoV-OC43 and HCoV-229E are clustered within regions that interact with the cellular receptors (9-O-acetylated sialoglycans and aminopeptidase N, ANPEP) and that were previously shown to modulate binding (Fig. 5 and 6). Thus, most positively selected sites in HCoV-229E map to the three loops that contact human ANPEP (hANPEP) (see below)^{57, 58}. Likewise, several positively selected sites are located within the sialoglycan-binding site of the HCoV-OC43 spike protein, and changes at sites 22 and 24 in other embecoviruses largely affect binding affinity⁵⁹. Similarly, the HE positively selected sites map to the lectin domain. Mutations at positions 114, 177,

and 178 determine the loss of sialoglycan binding, which is thought to have contributed to the shift to the human host⁶⁰. This indicates that gammaMap reliably identified relevant selection signatures.

Analysis of the polymorphic positively selected sites in the S and HE proteins of viruses sampled at different time intervals indicated that the evolution of HCoV-OC43 and HCoV-229E is ongoing and that new amino acid combinations have progressively emerged (Fig. 5c and 6c). Indeed, the amino acid status at the positively selected sites broadly corresponds to the receptor binding domain (RBD) classes of HCoV-229E and to HCoV-OC43 genotypes. Based on our dating analyses, the most recent RBD class of HCoV-229E (class VI) and the HCoV-OC43 genotypes (F/G/H) emerged ~ 15 years ago (Fig. 2). Interestingly, analysis of the RBD region in 92 BCoV sequences revealed limited variability with no clear temporal pattern (Supplementary Fig. 4). The same comparison could not be performed for camelid viruses as most of them were sampled in 2014-2015.

The HCoV-229E spike protein evolves to optimize receptor binding, not as a result of antigenic drift

Because, in analogy to SARS-CoV-2, HCoV-229E binds a protein receptor, we further investigated the positively selected sites in the spike protein. The specificity of HCoV-229E for hANPEP was previously ascribed to an extended tandem of H-bonds involving the 314-320 segment of RBD loop 1 and the 287-292 portion of a beta-strand surface-exposed of hANPEP domain II⁵⁷. Most of these interactions involve backbone atoms, reducing the dependency on sequence variations. In fact, the camel alphacoronavirus can use hANPEP as a receptor³². It was however suggested that changes in loop regions might accommodate species-specific differences among ANPEP orthologs and optimize receptor binding affinity⁵⁸. We thus compared the HCoV-229E RBD crystal structure and the corresponding model for camel alphacoronavirus (Fig. 7a). We also modeled camel ANPEP (cANPEP) based on the structure of the human ortholog. Overall, cANPEP features fewer charged residues at the interface than the human protein. In particular, T287 and I314 are replaced by D288 and D315 in the human receptor, whereas G291 is replaced by K292. Analysis of the contact interface indicated that the positively selected sites 316 (R or K, depending on RBD class), 407 (S in class I, H in classes V and VI), and 408 (K in classes I, V and VI) contribute additional interactions with the human protein than those established by the camel virus (Fig. 7a). These are made possible by the presence of the charged residues in the human receptor. Overall, these observations suggest that HCoV-229E can interact with hANPEP more efficiently than the camel virus and that positively selected sites contribute to increase affinity.

Previous investigations showed that the affinity of the six RBD classes of HCoV-229E for hANPEP varies in a range of K_d from ~430 (class I) to ~30 nM (classes V and VI) (Fig. 7c)⁵⁷. In particular, a strong increase in affinity is observed for classes V and VI. Some of the positively selected sites contribute to this increased affinity by changing loop conformation and by establishing additional interactions (Supplementary Fig. 5). For instance, H407 in classes V and VI forms an additional polar interaction with

the spatially close D315 of hANPEP, and K408 in the same classes intercepts the E291 backbone in the receptor. (Fig. 7a).

Variations in the RBD loops, which progressively emerged over the last 50 years (Fig. 7c), were previously proposed to derive from immune selection⁵⁷. Inspection of the IEDB database revealed that no experimental epitope for the spike protein of HCoV-229E has been described. We thus used the sequences of RBDs belonging to different classes to predict epitope positions for using BepiPred-2. Results indicated that epitopes do differ among RBD classes (Supplementary Table 3) and map to different structural regions (Fig. 7b) (data for HCoV-OC43 are shown in Supplementary Fig. 6). This is in line with the observation that antibodies against classes I and IV show no cross-neutralization and that HCoV-229E is undergoing antigenic drift^{57,58}. Nonetheless, the hypothesis of antigenic drift is difficult to reconcile with the evidence that reinfection with HCoV-229E is common and humoral immunity might be short-lived²²⁻²⁵. To clarify these issues, we used an extended set of spike protein sequences to date the temporal emergence of RBD classes. Results indicated that classes II, III, and IV, which have about two-fold higher affinity than class I, emerged 1-8 years apart. However, since the appearance of class V (with much higher affinity) about 44 years ago, no class emerged for 26 years (Fig. 7c). In fact, class VI split from class V about 17 years ago and the two classes show very similar sequence and binding properties. These different time intervals are poorly consistent with antigenic drift. Instead, these results suggest that strains with higher affinity for the cellular receptor have out-competed strains with lower affinity, and that HCoV-229E has evolved to optimize binding to the cellular receptor.

Discussion

Zoonotic diseases have been constantly emerging during human history, accounting for a large number of outbreaks, epidemics and pandemics. In recent years, several emerging and re-emerging pathogens have spread locally or globally. In addition to the three pathogenic coronaviruses, we have experienced the H1N1 pandemic (2009), as well as two pandemics caused by arthropod-borne viruses (chikungunya virus in 2014 and Zika virus in 2015)⁶¹. A cross-country Ebola epidemic has caused thousands of deaths in 2014-2016 in West Africa (<https://www.who.int/health-topics/ebola/>), and Nigeria reported the largest ever number of Lassa fever cases in 2018 (<https://www.who.int/csr/don/23-march-2018-lassa-fever-nigeria/>). These events have led to the conclusion that the occurrence of viral zoonoses is escalating⁶¹⁻⁶⁷. However, a survey of emerging infectious diseases (EIDs) caused by zoonotic viruses indicates no remarkable increase over eight decades and the occurrence of influenza pandemics in the last 3 centuries reveals no temporal pattern, nor an association with human population size (Fig. 3 and 4). Our data on the timing of endemic coronavirus emergence show that the four viruses entered human populations more than 50 years ago (and most likely much earlier). This indicates that, for several decades, no coronavirus spilled over to humans (or at least caused a registered outbreak) until three highly pathogenic coronaviruses emerged in tight temporal succession. To contextualize this observation in a wider perspective of viral diseases emergence over the last 8 decades, we mined data from GIDEON, a comprehensive epidemiology database. Because they have very different ecological characteristics, we

separately analyzed zoonoses that, in analogy to coronaviruses, are not transmitted by arthropod vectors, arthropod-borne infections, and non-zoonotic diseases. An analysis of raw data - i.e. the number of outbreaks (irrespective of the number of cases) and the number of sizable (more than 50 human cases) outbreaks per year - showed a weak tendency for zoonoses to increase over time, with a possible upswing since 1995. However, this trend was abolished, if not reversed, when outbreak number was corrected for a measure of reporting effort, indicating that the occurrence of viral diseases has not been increasing over the last decades.

Of course, the data used to reach this conclusion have a number of limitations. First, outbreak counts are clearly subject to reporting biases depending on geographical location, misdiagnosis (especially before the widespread use of molecular testing), and concurrent historical events (e.g., wars, famines). Second, zoonotic events resulting in a small number of cases may remain unreported (especially in the past or in developing regions), a bias we tried to overcome by performing the analyses for outbreaks with at least 50 human cases. Third, the annual number of JID articles only represents a proxy for reporting effort. Whereas it is difficult to weight the relevance of these possible biases, they are unlikely to affect our results to such a degree that we failed to detect a strong increase of outbreaks if it really occurred. Moreover, as mentioned above, no strong escalation of EIDs caused by viruses was observed when we analyzed the data curated by Jones and coworkers, who did not rely on GIDEON⁴⁹. In this respect, it is worth mentioning that our data do not contradict the overall conclusion reached in their work- i.e. that EIDs have risen in frequency since 1940. In fact, Jones and colleagues found that less than 26% of EIDs are caused by viruses, indicating that the increase they observed over time is mostly due to other pathogens⁴⁹. We also stress that our findings concerning the timing of outbreaks by no means imply that anthropic changes previously associated to infectious disease emergence (e.g., deforestation, climate change, wildlife habitat encroachment, loss of biodiversity)^{64, 67-77} have no role in favoring the occurrence of viral zoonoses.

If the three highly pathogenic coronaviruses do not fit within a more general trend for zoonotic viruses to increasingly infect human populations, we are left with the question of which factors can explain the irregular pattern of coronavirus emergence. This same issue has remained unanswered for decades in the case of influenza, as virological and non-virological factors have been associated to the occurrence of pandemics with poor explanatory power⁷⁸⁻⁸⁴. Thus, it is presently impossible to predict the timing of such events and their severity. In general, as noted elsewhere, we have very little ability to anticipate which viruses will emerge and how pathogenic they will be^{85, 86}. In this respect, it is also worth mentioning that, because they have now circulated in (and adapted to) human populations for decades, if not centuries, we cannot exclude that the endemic coronaviruses were once more pathogenic than they are now. Indeed, it was previously suggested that the 1889-1890 flu pandemic, which was characterized by pronounced central nervous system symptoms, was actually caused by HCoV-OC43, as the dates corresponds to the time when the virus entered human populations and HCoV-OC43 displays some neurotropism³⁴.

Although we cannot go back in time and infer the original phenotype of endemic coronaviruses, nor can we have a full picture of their ancestral genetic diversity, analysis of their evolution is potentially very informative to understand the future trajectories of SARS-CoV-2, and of coronaviruses in general. Analysis of bat coronaviruses indicated that, in analogy to SARS-CoV, SARS-CoV-2 required limited adaptation to gain the ability to infect and spread in our species^{56,87}. As HCoV-OC43 and HCoV-229E most likely emerged from bovine and camelid coronaviruses, we investigated which selective events accompanied the divergence of these human viruses from the animal ones and their diffusion in humans. We note, however, that because of the lack of information on early isolates (as evident in the phylogenetic gaps in Fig. 2), it is formally impossible to distinguish between the initial events associated with the optimization for human infection and the ongoing adaptation resulting from immune selection or other pressures.

Our results indicate that the spike protein and other structural proteins of both viruses represented the major targets of selection. An interesting exception is the strong signature of selection we observed for HCoV-OC43 ORF5 (also known as ns12.9). The encoded protein functions as a viroporin and its deletion reduces viral replication, inflammatory response, and virulence in mouse models⁸⁸. Positive selection also drove the evolution of the membrane proteins of both viruses, as well as of the envelope protein of HCoV-OC43. This latter, besides having structural roles, acts as a viroporin and represents a neurovirulence factor^{89,90}. Likewise, the membrane proteins of several coronaviruses, including HCoV-OC43, in addition to their role in virion maturation, are capable of antagonizing interferon responses⁹¹⁻⁹⁵. Overall, these data suggest that positively selected sites in these proteins might contribute to fine-tune the interaction between coronaviruses and human immune responses.

Clearly, the spike protein, as well as HE in the case of HCoV-OC43, have a major interest as targets of selection, as they represent major determinants of host range and infectivity¹⁻³. Most selected sites were found to be located in the receptor binding domains of the spike proteins, as well as in the lectin domain of HE. However, additional sites mapped to other regions of the spike proteins and were mostly fixed in frequency. These include three sites in the heptad repeat region of the spike protein of HCoV-229E and one site in the fusion peptide of HCoV-OC43 (Fig. 5 and 6). Notably, the heptad repeat region was previously described as major target of selection in MERS-CoV and related camel viruses^{96,97} and variants within this region and/or the fusion peptide were shown to modulate viral tropism and host range in several viruses, including animal coronaviruses⁹⁸⁻¹⁰¹.

Coronaviruses can use very different cellular receptors and their spike proteins display a remarkable ability to adapt to different cellular receptors^{2,42}. Embecoviruses such as HCoV-OC43, HCoV-HKU1 and BCoV attach to 9-O-acetylated sialoglycans via the spike protein, with HE acting as a receptor-destroying enzyme^{53,59}. Conversely, HCoV-229E and HCoV-NL63, in analogy to other betacoronavirus, use a protein receptor^{102,103}. Biochemical and crystallographic analyses indicated that, since the shift to the human host, the spike and HE proteins of HCoV-OC43 have co-evolved to optimize the balance between binding and release from sialoglycans in human airways^{60,104}. We confirm herein the previously observed

emergence of spike and HE variants over time and the replacement of earlier variants with the more recent ones (Fig. 5). However, the relative binding affinity and esterase activity of such variants have not been extensively investigated, yet. Conversely, binding assays have shown that different classes of the HCoV-229E spike protein RBD have very different binding affinity for hANPEP. The appearance of variants with increased affinity has clearly occurred progressively in time (Fig. 7c) and changes at the positively selected sites have most likely facilitated the initial spill-over from camels (Fig. 7a).

Whereas these data suggest that HCoV-OC43 and HCoV-229E have been adapting to optimize infection and spread in human populations, an alternative possibility is that the selective pressure towards amino acid replacements is exerted by the host immune system. Indeed, it was previously suggested that antigenic drift was responsible for the emergence of distinct HCoV-229E RBD classes⁵⁷. This assumption was based on the observation that antibodies raised against class I RBD do not neutralized viruses with RBDs belonging to different classes^{57,58}. Nonetheless, growing evidence suggests that the humoral immune response against endemic coronaviruses wanes in a few months²²⁻²⁵. As a consequence, natural reinfection is possible and the antibody response can hardly be regarded as a strong selective pressure for these viruses. In line with this observation, three lines of evidence indicate that antigenic drift is not primary responsible for selecting changes within the RBDs of HCoV-OC43 and HCoV-229E. First, limited variation with no temporal pattern was evident in the RBD region of BCoV sequences sampled over 34 years. Second, dating of the emergence of HCoV-229E RBD classes indicated an initial rapid turnover of classes I to IV followed by a 26-year time during which no variant turned up after the emergence of class V. Class V and the closely related class VI RBDs differ in binding affinity from the other classes by almost an order of magnitude⁵⁷. Third, neutralizing antibodies against SARS-CoV, SARS-CoV-2 and MERS-CoV can bind region of the spike protein other than the RBD or can recognize different viral proteins¹⁰⁵⁻¹⁰⁸. However, the polymorphic selected sites we detected were almost exclusively located within the RBDs.

These results, together with the remarkable seasonality of endemic coronaviruses, suggest that selection has been acting to optimize binding to the cellular receptor and that strains with increasing affinity have replaced those with lower binding ability and, possibly, lower infectivity. These observations have relevance for the evolution of SARS-CoV-2. Recent work has indicated that the spike protein of SARS-CoV-2 can tolerate a substantial number of substitutions, with some of them even increasing receptor binding¹⁰⁹. Thus, it is in theory possible that, as observed for HCoV-229E, variants that enhance ACE2 binding will emerge. We however note that even class V and VI RBDs have much lower affinity for hANPEP ($K_d \sim 30\text{nM}$) than most other human coronaviruses to their respective cellular receptors (K_d in the range of 1 to 5 nM for SARS-CoV-2, SARS-CoV, and HCoV-NL63)^{110,111}. It is thus possible that the selective pressure for increased binding is low for SARS-CoV-2. Nonetheless, even minor changes in binding affinity might confer a selective advantage over other strains, as preliminary data on a newly evolved variant in the UK suggest (<https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>).

Whether SARS-CoV-2 will evolve variants that alter its antigenic properties is also, of course, of great interest. Our data on HCoV-229E and HCoV-OC43 suggest that antigenic drift is not a major evolutionary driving force for these viruses and available data on SARS-CoV-2 suggest that variants affecting antigenic properties are present at low frequency in the circulating viral population^{112,113}. However, the evolution of the virus is likely to depend on the duration and efficacy of natural immunity. Although immunity against SARS-CoV-2 might be relatively short-lived, additional data over longer time-frames will be required to definitely address this issue¹¹⁴. Moreover, in rare cases of long-term SARS-CoV-2 infection, either or not in association with convalescent plasma treatment, the virus might experience within-host selective pressures that favor the appearance of immune escape variants¹¹⁵⁻¹¹⁷. If such variants do not impair viral fitness in naive hosts, they may spread in the population, especially if they are associated with changes that increase infectivity, as it seems to be the case for the recently emerged UK variant (<https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>). Likewise, the imminent mass deployment of vaccines against SARS-CoV-2 will subject the virus to a selective pressure that the endemic coronaviruses have never experienced.

Methods

Sequences and alignments

Complete or almost complete genome sequences for all four endemic coronaviruses were downloaded from the NCBI database (National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>). Only sequences with known sampling dates were included in the analyses (Supplementary Table 1). The HCoV-OC43 Paris strain was excluded as its sampling date is uncertain¹¹⁸. For each human coronavirus, the closest phylogenetically related animal virus was also retrieved: camel alphacoronavirus, BCoV, murine coronavirus, and a bat coronavirus from Kenya for HCoV-229E, HCoV-OC43, HCoV-HKU1, and HCoV-NL63, respectively (Supplementary Table 1).

All complete genome sequences with sampling year of BCoV were also retrieved from NCBI, along with all HCoV-229E spike sequences, again with collection date (Supplementary Table 4 and Supplementary Table 1).

Sequence alignments were generated using MAFFT (v7.427)¹¹⁹, with default parameters.

Recombination analysis

Recombination can affect phylogenetic tree branch length estimates and, consequently, molecular evolution analyses³⁹. Thus, each coronavirus alignment was tested for evidence of recombination signals using the 3SEQ software (v.1.7)⁴⁵. This method scans a given alignment searching for mosaic recombination signals in all possible sequence triplets. The result is the identification of genomic regions in which one of the three sequences is the recombinant (child) of the other two (parental). 3SEQ full scans were run with a recombination significance threshold of 0.01. All significant recombination events were mapped onto coronavirus alignments and the longest non-recombinant genomic regions, defined as

the genomic region comprised between two recombination breakpoints, were selected for subsequent analyses (Fig. 1). Unique recombination events identified in genome alignments are shown in Figure 1. No breakpoint was detected in the extended set of HCoV-229E spike proteins.

Phylogenetic trees and temporal signal

Phylogenetic trees for the non-recombinant regions of all endemic coronaviruses were reconstructed using the phyML software under a general time reversible (GTR) model plus gamma-distributed rates and 4 substitution rate categories¹²⁰. Substitution models were estimated using JmodelTest 2^{121, 122}. Internal GTR estimated branch lengths were compared to branch lengths calculated using a model that accounts for different selective pressures among lineages. This model is implemented in the aBSREL (adaptive Branch-Site Random Effects Likelihood¹²³) tool from the HYPHY suite (version 2.5)¹²⁴. To evaluate whether the non-recombinant genomic regions selected for the analyses carried sufficient temporal signal, we calculated the correlation coefficients (r) of regressions of root-to-tip genetic distances against sequence sampling years¹²⁵. We applied a method that minimizes the residual mean squares of the models and calculated p values by performing clustered permutations (1,000) of the sampling dates^{125, 126}. We considered significant a regression with $p < 0.05$. Non-recombinant regions of HCoV-OC43, HCoV-229E, and HCoV-NL63 (region 2) showed evidence of temporal signal (Fig. 1). The same result was obtained for the extended set of HCoV-229E RBD spike proteins (Supplementary Fig. 7).

Molecular dating

Time estimate phylogenetic reconstruction was performed using a Bayesian approach implemented in the Bayesian Evolutionary Analysis by Sampling Trees (BEAST, v.1.10.4) software¹²⁷.

To select the best-fit molecular clock and tree prior, we ran the path sampling tool implemented in BEAST to choose between a constant size, an exponential growth, or a coalescent Bayesian skyline tree prior and between a strict and a relaxed log normal clock (50 steps, 1,000,000 iterations each). A Bayes factor test was applied to compare the different likelihoods (Supplementary Table 4).

A constant size population tree prior with a strict clock model was favored for HCoV-OC43 and HCoV-229E, whereas an exponential growth population tree prior with a relaxed clock model with log-normal distribution was preferred for HCoV-NL63.

For the HCoV-HKU1 phylogeny, we used the mean rates estimated for the other betacoronavirus HCoV-OC43 (mean= 1.6×10^{-4} , standard deviation= 8.8×10^{-6}) as informative rate prior following a normal distribution.

We performed two different Markov chain Monte Carlo runs for all four endemic coronaviruses, two hundred million iterations each, and sampled every 10,000 steps after a 10% burn-in. Runs were combined after checking for convergence and for heaving effective sampling sizes > 100 .

We generated a maximum clade credibility tree using TreeAnnotator¹²⁸, which was visualized with FigTree (<http://tree.bio.ed.ac.uk/>).

Population genetics-phylogenetic analysis

Selective events that accompanied the appearance of the human viruses were investigated for HCoV-OC43 and HCoV-229E, the two endemic coronaviruses for which the closest related animal virus is almost certain (i.e. the bovine and the camelid coronaviruses) (Supplementary Table 1).

Analyses were performed with gammaMap, that uses intra-specific variation and inter-specific diversity to estimate, along coding regions, the distribution of selection coefficients (γ). Thus, for the two coronaviruses, all ORF coding region sequences (cds) were retrieved from the same set of strains analyzed before, all possible overlapping regions were masked and single ORF cds alignments were generated using MAFFT, with codons sequence type as parameter.

In the gammaMap framework, the selection coefficient is defined as $2PNes$, where P is the ploidy, N_e is effective population size, and s is the fitness advantage of any amino acid-replacing derived allele. The method categorizes selection coefficients into 12 predefined classes ranging from -500 (inviable) to 100 (strongly beneficial), with 0 indicating neutrality⁵². We also assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary within genes following log-normal distributions, whereas p (probability of adjacent codons to share the same selection coefficient) following a log-uniform distribution. Finally, for the selection coefficients, we considered a uniform Dirichlet distribution with the same prior weight for each selection class. We performed 2 runs with 100,000 iterations each and with a thinning interval of 10 iterations. Runs were merged after checking for convergence. Codon positions were defined as positively selected if they showed a posterior probability > 0.75 of having $\gamma \geq 1$.

Sequence logos were generated using WebLogo¹²⁹ (<https://weblogo.berkeley.edu/>).

EID events, influenza pandemics, and viral disease outbreaks

The timing of influenza pandemics was obtained from a previous work⁸³, as well as from references therein^{80, 130-138}.

EID events caused by viral diseases (1940 to 2004) were retrieved from a previous work⁴⁹.

We compiled a list of viral disease outbreaks from data stored in the Global Infectious Disease and Epidemiology Network (GIDEON) database⁵⁰ (last accessed December 15th, 2020) (Supplementary Table 6). GIDEON is updated monthly, and collates information from a wide range of sources, including Health Ministry publications, peer-review journals, textbooks, as well as the WHO and CDC websites. Information of individual outbreaks was manually inspected to determine whether the overall number of human cases

(aggregated over countries in cases of multi-regional outbreaks) was higher than 50 (Supplementary Table 7).

To analyze the relationship between viral disease outbreaks and time, we fitted generalized linear models with Poisson error using $\log(\text{number JID articles})$ as an offset. The number of JID articles was selected as an offset because, as previously indicated ⁴⁹, the journal is a major, dedicated venue for the publication of human infectious disease research since 1945. Computations were performed in R version 3.1.2 using the `glm` function.

Molecular modeling and epitope prediction

The structure of HCoV-229E RBD of class I in complex with hANPEP was retrieved by the Protein Data Bank (PDB ID: 6AKT). Such structure was also used as template to model the interaction between the S-protein RBD of HCoV-229E camel ortholog and cANPEP, using the webserver HOMCOS ¹³⁹. The HOMCOS webserver performs `blastp` searches ¹⁴⁰ to look for complexes formed by proteins, which are homologous to the query proteins. We then selected one of these complexes as a template and launched the program MODELLER ¹⁴¹ that models the interaction between the query proteins with a script provided by HOMCOS. On the basis of sequence similarity at the binding interface, we chose structures 6U7E, 6U7F and 6U7G as templates to model the interaction of RDB Class II, Class IV and Class V/VI, respectively, with hANPEP. The same templates have been used to model the sole RBDs using SWISS-MODEL ¹⁴². These RBDs structures were then used to map epitopes on the molecular surface. VADAR 1.8 (Volume, Area, Dihedral Angle Reporter) ¹⁴³, was used to assess the accuracy of all models. VADAR uses a combination of more than 15 specific algorithms to calculate different parameters for each residue and of the overall protein structure. We used such parameters to verify i) the agreement of observed structural parameters (such as ϕ and ψ dihedral and buried charges) of the newly predicted structures with the expected values calculated on the corresponding sequences and ii) the presence of a low number of packing defects. Structures were then analyzed with the software PyMOL ¹⁴⁴, that was also used to create proteins figures.

Epitope positions were predicted using the BepiPred-2.0 method with default parameters and accessed through the IEDB server (<http://tools.iedb.org/bcell/help/#Bepipred-2.0>) ¹⁴⁵.

Declarations

Acknowledgments

We are grateful to Prof. Elio Antonello, Fabrizio Nicastro, and Giovanni Pareschi for valuable discussion and constructive input. This work was supported by the Italian Ministry of Health (“Ricerca Corrente 2019-2020” to MS, “Ricerca Corrente 2018-2020” to DF).

Author Contributions

Conceptualization, DF and MS; Formal Analysis, DF, RC, FA, MB, CP, UP, and MS; Investigation, DF, RC, FA, MB, CP, UP, and MS; Visualization, DF, RC, FA, MB; Writing –Original Draft, MS, DF, FA; Writing –Review & Editing, MS, MC, LDG; Funding Acquisition MS and DF; Supervision, MS, MC and LDG.

Declaration of Interests

The authors declare no competing interests.

Data availability

Lists of virus accession IDs are reported in Supplementary Tables 1 and 4. Predicted epitopes are reported in Supplementary Table 2.

References

1. Ye, Z. W. *et al.* Zoonotic origins of human coronaviruses. *Int. J. Biol. Sci.* **16**, 1686-1697 (2020).
2. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol.* **25**, 35-48 (2017).
3. Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181-192 (2019).
4. Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967-1976 (2003).
5. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. & Fouchier, R. A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814-1820 (2012).
6. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966-1970 (2003).
7. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**, 536-544 (2020).
8. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727-733 (2020).
9. Killerby, M. E., Biggs, H. M., Midgley, C. M., Gerber, S. I. & Watson, J. T. Middle East Respiratory Syndrome Coronavirus Transmission. *Emerg. Infect. Dis.* **26**, 191-198 (2020).
10. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
11. Lam, T. T. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv*, 2020.02.13.945485 (2020).
12. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286-289 (2020).

13. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv*, 2020.02.07.939207 (2020).
14. Liu, P. *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog.* **16**, e1008421 (2020).
15. Sironi, M. *et al.* SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective. *Infect. Genet. Evol.* **84**, 104384 (2020).
16. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**, 860-868 (2020).
17. Olival, K. J. *et al.* Possibility for reverse zoonotic transmission of SARS-CoV-2 to free-ranging wildlife: A case study of bats. *PLoS Pathog.* **16**, e1008758 (2020).
18. Oude Munnink, B. B. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* (2020).
19. Shaman, J. & Galanti, M. Will SARS-CoV-2 become endemic? *Science* **370**, 527-529 (2020).
20. Bucknall, R. A., King, L. M., Kapikian, A. Z. & Chanock, R. M. Studies with human coronaviruses. II. Some properties of strains 229E and OC43. *Proc. Soc. Exp. Biol. Med.* **139**, 722-727 (1972).
21. Woo, P. C. *et al.* Clinical and molecular epidemiological features of coronavirus HKU1-associated community-acquired pneumonia. *J. Infect. Dis.* **192**, 1898-1907 (2005).
22. Callow, K. A., Parry, H. F., Sergeant, M. & Tyrrell, D. A. The time course of the immune response to experimental coronavirus infection of man. *Epidemiol. Infect.* **105**, 435-446 (1990).
23. Edridge, A. W. D. *et al.* Seasonal coronavirus protective immunity is short-lasting. *Nat. Med.* **26**, 1691-1693 (2020).
24. Galanti, M. & Shaman, J. Direct Observation of Repeated Infections With Endemic Coronaviruses. *J. Infect. Dis.* (2020).
25. Schmidt, O. W., Allan, I. D., Cooney, M. K., Foy, H. M. & Fox, J. P. Rises in titers of antibody to human coronaviruses OC43 and 229E in Seattle families during 1975-1979. *Am. J. Epidemiol.* **123**, 862-868 (1986).
26. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. *Adv. Virus Res.* **100**, 163-188 (2018).
27. Tao, Y. *et al.* Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. *J. Virol.* **91**, e01953-16. doi: 10.1128/JVI.01953-16. Print 2017 Mar 1 (2017).
28. Huynh, J. *et al.* Evidence supporting a zoonotic origin of human coronavirus strain NL63. *J. Virol.* **86**, 12816-12825 (2012).
29. Corman, V. M. *et al.* Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. *J. Virol.* **89**, 11858-11870 (2015).
30. Pfeifferle, S. *et al.* Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. *Emerg. Infect. Dis.* **15**, 1377-1384 (2009).

31. Crossley, B. M., Mock, R. E., Callison, S. A. & Hietala, S. K. Identification and characterization of a novel alpaca respiratory coronavirus most closely related to the human coronavirus 229E. *Viruses* **4**, 3689-3700 (2012).
32. Corman, V. M. *et al.* Link of a ubiquitous human coronavirus to dromedary camels. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9864-9869 (2016).
33. Sabir, J. S. *et al.* Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* **351**, 81-84 (2016).
34. Vijgen, L. *et al.* Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J. Virol.* **79**, 1595-1604 (2005).
35. Vijgen, L. *et al.* Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. *J. Virol.* **80**, 7270-7274 (2006).
36. Brandão, P. E. Could human coronavirus OC43 have co-evolved with early humans? *Genet. Mol. Biol.* **41**, 692-698 (2018).
37. Al-Khannaq, M. N. *et al.* Molecular epidemiology and evolutionary histories of human coronavirus OC43 and HKU1 among patients with upper respiratory tract infections in Kuala Lumpur, Malaysia. *Virol. J.* **13**, 33-016-0488-4 (2016).
38. Bidokhti, M. R. M. *et al.* Evolutionary dynamics of bovine coronaviruses: natural selection pattern of the spike gene implies adaptive evolution of the strains. *J. Gen. Virol.* **94**, 2036-2049 (2013).
39. Schierup, M. H. & Hein, J. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**, 1578-1579 (2000).
40. Duchene, S., Holmes, E. C. & Ho, S. Y. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.* **281**, 10.1098/rspb.2014.0732 (2014).
41. Rieux, A. & Balloux, F. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol. Ecol.* **25**, 1911-1924 (2016).
42. Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J. Virol.* **84**, 3134-3146 (2010).
43. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* (2020).
44. Forni, D., Cagliani, R. & Sironi, M. Recombination and Positive Selection Differentially Shaped the Diversity of Betacoronavirus Subgenera. *Viruses* **12**, 1313. doi: 10.3390/v12111313 (2020).
45. Lam, H. M., Ratmann, O. & Boni, M. F. Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Mol. Biol. Evol.* **35**, 247-251 (2018).
46. Aiewsakun, P. & Katzourakis, A. Time-Dependent Rate Phenomenon in Viruses. *J. Virol.* **90**, 7184-7195 (2016).
47. Wertheim, J. O., Chu, D. K., Peiris, J. S., Kosakovsky Pond, S. L. & Poon, L. L. A case for the ancient origin of coronaviruses. *J. Virol.* **87**, 7039-7045 (2013).

48. Morens, D. M. & Taubenberger, J. K. The Mother of All Pandemics Is 100 Years Old (and Going Strong)! *Am. J. Public Health* **108**, 1449-1454 (2018).
49. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* **451**, 990-993 (2008).
50. Edberg, S. C. Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. *Clin. Infect. Dis.* **40**, 123-126 (2005).
51. Woolhouse, M. E. *et al.* Temporal trends in the discovery of human viruses. *Proc. Biol. Sci.* **275**, 2111-2115 (2008).
52. Wilson, D. J., Hernandez, R. D., Andolfatto, P. & Przeworski, M. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* **7**, e1002395 (2011).
53. de Groot, R. J. Structure, function and evolution of the hemagglutinin-esterase proteins of corona- and toroviruses. *Glycoconj. J.* **23**, 59-72 (2006).
54. Ho, S. Y. *et al.* Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087-3101 (2011).
55. Wertheim, J. O. & Kosakovsky Pond, S. L. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* **28**, 3355-3365 (2011).
56. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2. *J. Virol.* (2020).
57. Wong, A. H. M. *et al.* Receptor-binding loops in alphacoronavirus adaptation and evolution. *Nat. Commun.* **8**, 1735-017-01706-x (2017).
58. Li, Z. *et al.* The human coronavirus HCoV-229E S-protein structure and receptor binding. *Elife* **8**, 10.7554/eLife.51230 (2019).
59. Hulswit, R. J. G. *et al.* Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2681-2690 (2019).
60. Bakkers, M. J. *et al.* Betacoronavirus Adaptation to Humans Involved Progressive Loss of Hemagglutinin-Esterase Lectin Activity. *Cell. Host Microbe* **21**, 356-366 (2017).
61. Morens, D. M., Daszak, P., Markel, H. & Taubenberger, J. K. Pandemic COVID-19 Joins History's Pandemic Legion. *mBio* **11**, e00812-20. doi: 10.1128/mBio.00812-20 (2020).
62. Cunningham, A. A., Daszak, P. & Wood, J. L. N. One Health, emerging infectious diseases and wildlife: two decades of progress? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 20160167. doi: 10.1098/rstb.2016.0167 (2017).
63. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9-20 (2018).
64. Karesh, W. B. *et al.* Ecology of zoonoses: natural and unnatural histories. *Lancet* **380**, 1936-1945 (2012).
65. Pike, J., Bogich, T., Elwood, S., Finnoff, D. C. & Daszak, P. Economic optimization of a global strategy to address the pandemic threat. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18519-18523 (2014).

66. The, L. Zoonoses: beyond the human-animal-environment interface. *Lancet* **396**, 1-6736(20)31486-0 (2020).
67. Weiss, R. A. & McMichael, A. J. Social and environmental risk factors in the emergence of infectious diseases. *Nat. Med.* **10**, S70-6 (2004).
68. Allen, T. *et al.* Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* **8**, 1124-017-00923-8 (2017).
69. Epstein, J. H., Field, H. E., Luby, S., Pulliam, J. R. & Daszak, P. Nipah virus: impact, origins, and causes of emergence. *Curr. Infect. Dis. Rep.* **8**, 59-65 (2006).
70. Johnson, C. K. *et al.* Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proc. Biol. Sci.* **287**, 20192736 (2020).
71. Keesing, F. *et al.* Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468**, 647-652 (2010).
72. Keesing, F. *et al.* Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468**, 647-652 (2010).
73. Murray, K. A. & Daszak, P. Human ecology in pathogenic landscapes: two hypotheses on how land use change drives viral emergence. *Curr. Opin. Virol.* **3**, 79-83 (2013).
74. Pulliam, J. R. *et al.* Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *J. R. Soc. Interface* **9**, 89-101 (2012).
75. Smith, K. F. *et al.* Ecology. Reducing the risks of the wildlife trade. *Science* **324**, 594-595 (2009).
76. Wilkinson, D. A., Marshall, J. C., French, N. P. & Hayman, D. T. S. Habitat fragmentation, biodiversity loss and the risk of novel infectious disease emergence. *J. R. Soc. Interface* **15**, 20180403. doi: 10.1098/rsif.2018.0403 (2018).
77. Zinsstag, J. *et al.* Climate change and One Health. *FEMS Microbiol. Lett.* **365**, fny085. doi: 10.1093/femsle/fny085 (2018).
78. Dowdle, W. R. Influenza A virus recycling revisited. *Bull. World Health Organ.* **77**, 820-828 (1999).
79. Hayes, D. P. Influenza pandemics, solar activity cycles, and vitamin D. *Med. Hypotheses* **74**, 831-834 (2010).
80. Morens, D. M. & Taubenberger, J. K. Pandemic influenza: certain uncertainties. *Rev. Med. Virol.* **21**, 262-284 (2011).
81. Snyder, M. R. & Ravi, S. J. 1818, 1918, 2018: Two Centuries of Pandemics. *Health. Secur.* **16**, 410-415 (2018).
82. Taubenberger, J. K. & Morens, D. M. Pandemic influenza—including a risk assessment of H5N1. *Rev. Sci. Tech.* **28**, 187-202 (2009).
83. Towers, S. Sunspot activity and influenza pandemics: a statistical assessment of the purported association. *Epidemiol. Infect.* **145**, 2640-2655 (2017).
84. Viboud, C. & Lessler, J. The 1918 Influenza Pandemic: Looking Back, Looking Forward. *Am. J. Epidemiol.* **187**, 2493-2497 (2018).

85. Holmes, E. C., Rambaut, A. & Andersen, K. G. Pandemics: spend on surveillance, not prediction. *Nature* **558**, 180-182 (2018).
86. Morse, S. S. *et al.* Prediction and prevention of the next pandemic zoonosis. *Lancet* **380**, 1956-1965 (2012).
87. MacLean, O. A. *et al.* Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans. *bioRxiv* (2020).
88. Zhang, R. *et al.* The ns12.9 Accessory Protein of Human Coronavirus OC43 Is a Viroporin Involved in Virion Morphogenesis and Pathogenesis. *J. Virol.* **89**, 11383-11395 (2015).
89. Stodola, J. K., Dubois, G., Le Coupanec, A., Desforgues, M. & Talbot, P. J. The OC43 human coronavirus envelope protein is critical for infectious virus production and propagation in neuronal cells and is a determinant of neurovirulence and CNS pathology. *Virology* **515**, 134-149 (2018).
90. Torres, J., Wang, J., Parthasarathy, K. & Liu, D. X. The transmembrane oligomers of coronavirus protein E. *Biophys. J.* **88**, 1283-1290 (2005).
91. Beidas, M. & Chehadeh, W. Effect of Human Coronavirus OC43 Structural and Accessory Proteins on the Transcriptional Activation of Antiviral Response Elements. *Intervirology* **61**, 30-35 (2018).
92. Siu, K., Chan, C., Kok, K., Woo, P. C. & Jin, D. Suppression of innate antiviral response by severe acute respiratory syndrome coronavirus M protein is mediated through the first transmembrane domain. *Cellular & molecular immunology* **11**, 141-149 (2014).
93. Lui, P. *et al.* Middle East respiratory syndrome coronavirus M protein suppresses type I interferon expression through the inhibition of TBK1-dependent phosphorylation of IRF3. *Emerging microbes & infections* **5**, 1-9 (2016).
94. Yang, Y. *et al.* The structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of Middle East respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists. *Protein & cell* **4**, 951-961 (2013).
95. Fang, X. *et al.* The membrane protein of SARS-CoV suppresses NF- κ B activation. *J. Med. Virol.* **79**, 1431-1439 (2007).
96. Forni, D. *et al.* The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses. *Sci. Rep.* **5**, 14480 (2015).
97. Cotten, M. *et al.* Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *MBio* **5**, 10.1128/mBio.01062-13 (2014).
98. Yamada, Y., Liu, X. B., Fang, S. G., Tay, F. P. & Liu, D. X. Acquisition of cell-cell fusion activity by amino acid substitutions in spike protein determines the infectivity of a coronavirus in cultured cells. *PLoS One* **4**, e6130 (2009).
99. Navas-Martin, S., Hingley, S. T. & Weiss, S. R. Murine coronavirus evolution in vivo: functional compensation of a detrimental amino acid substitution in the receptor binding domain of the spike glycoprotein. *J. Virol.* **79**, 7629-7640 (2005).

100. de Haan, C. A. *et al.* Cooperative involvement of the S1 and S2 subunits of the murine coronavirus spike protein in receptor binding and extended host range. *J. Virol.* **80**, 10909-10918 (2006).
101. McRoy, W. C. & Baric, R. S. Amino acid substitutions in the S2 subunit of mouse hepatitis virus variant V51 encode determinants of host range expansion. *J. Virol.* **82**, 1414-1424 (2008).
102. Hofmann, H. *et al.* Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7988-7993 (2005).
103. Yeager, C. L. *et al.* Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357**, 420-422 (1992).
104. Lang, Y. *et al.* Coronavirus hemagglutinin-esterase and spike proteins coevolve for functional balance and optimal virion avidity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25759-25770 (2020).
105. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450-456 (2020).
106. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, 1-6 (2020).
107. Flehmig, B. *et al.* Persisting Neutralizing Activity to SARS-CoV-2 over Months in Sera of COVID-19 Patients. *Viruses* **12**, 1357 (2020).
108. Jiang, S., Hillyer, C. & Du, L. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. *Trends Immunol.* (2020).
109. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
110. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19970-19974 (2009).
111. Yi, C. *et al.* Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cellular & Molecular Immunology*, 1-10 (2020).
112. Forni, D. *et al.* Antigenic variation of SARS-CoV-2 in response to immune pressure. *Mol. Ecol.* (2020).
113. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell. Host Microbe* (2020).
114. Saad-Roy, C. M. *et al.* Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science* **370**, 811-818 (2020).
115. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383**, 2291-2293 (2020).
116. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* (2020).
117. Kemp, S. A. *et al.* Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion. *medRxiv* (2020).
118. Vijgen, L., Lemey, P., Keyaerts, E. & Van Ranst, M. Genetic variability of human respiratory coronavirus OC43. *J. Virol.* **79**, 3223-4; author reply 3224-5 (2005).

119. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
120. Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113-137 (2009).
121. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
122. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).
123. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342-1353 (2015).
124. Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679 (2005).
125. Murray, G. G. *et al.* The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7**, 80-89 (2016).
126. Duchene, S., Duchene, D., Holmes, E. C. & Ho, S. Y. The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data. *Mol. Biol. Evol.* **32**, 1895-1906 (2015).
127. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
128. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
129. Crooks, G., Hon, G. & Chandonia, J. M, Brenner SE. *WebLogo: a sequence logo generator. Genome Res* **14**, 1188-1190 (2004).
130. Hampson, A. W. & Mackenzie, J. S. The influenza viruses. *Med. J. Aust.* **185**, S39-43 (2006).
131. Mamelund, S. Influenza, historical. *Medicine* **54**, 361-371 (2008).
132. Lattanzi, M. in *Influenza Vaccines for the Future* 245-259 (Springer, 2008).
133. Potter, C. Chronicle of influenza pandemics, Nicholson KG, Webster R, Hay A: Textbook of Influenza. (1997).
134. Garrett, L. in *The coming plague: newly emerging diseases in a world out of balance* (Macmillan, 1994).
135. Beveridge, W. I. The chronicle of influenza epidemics. *Pubbl. Stn. Zool. Napoli.*, 223-234 (1991).
136. Kilbourne, E. D. in *Influenza* 3-22 (Springer, 1987).
137. Pyle, G. F. in *The diffusion of influenza: patterns and paradigms* (Rowman & Littlefield, 1986).
138. Patterson, K. D. in *Pandemic influenza, 1700-1900: a study in historical epidemiology* (Rowman & Littlefield Totowa, NJ, USA:, 1986).
139. Kawabata, T. HOMCOS: an updated server to search and model complex 3D structures. *Journal of structural and functional genomics* **17**, 83-99 (2016).

140. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
141. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815 (1993).
142. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201 (2006).
143. Willard, L. *et al.* VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31**, 3316-3319 (2003).
144. Schrödinger, L. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (2017). *Google Scholar There is no corresponding record for this reference.*
145. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* **45**, W24-W29 (2017).
146. Walls, A. C. *et al.* Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531**, 114-117 (2016).

Figures

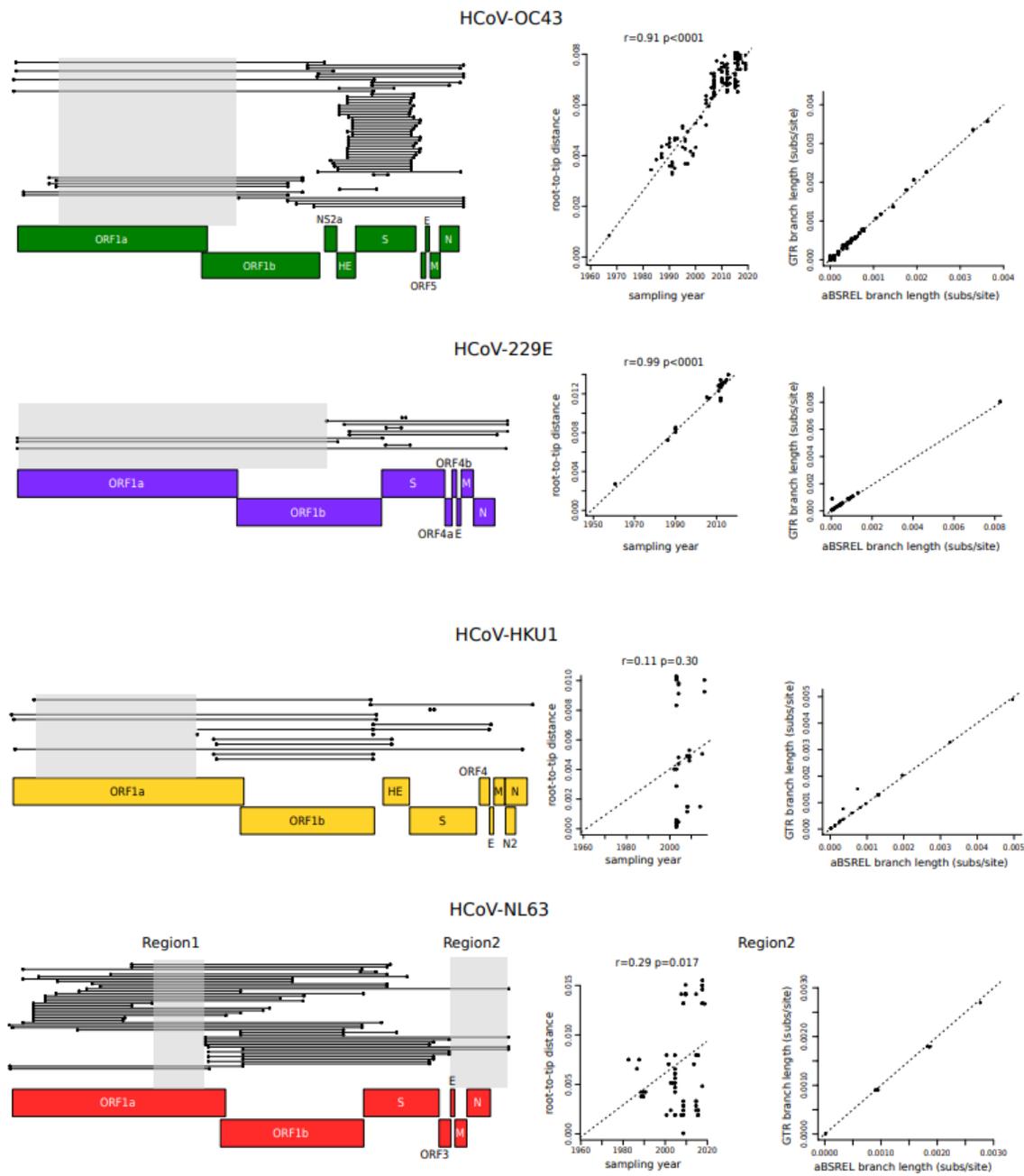


Figure 1

Recombination events and temporal signal. Unique recombination events in endemic coronaviruses (left panels). Each event is shown as a line with dots representing the start and the end. The non-recombinant regions used in the analyses are indicated with gray shadows. Schematic representations of coronavirus genomes are also reported. Plots of the root-to-tip distance as a function of sampling years are shown in the central panels. Each point corresponds to a viral sequence and the dotted line is the linear regression

calculated using a method that minimizes the residual mean squares. The r coefficient and the corresponding p value are also shown. Right panels report comparisons of branch lengths obtained using the aBS-REL and the GTR models. Each dot represents an internal branch of a phylogenetic tree calculated using the longest non-recombinant regions of each endemic coronavirus.

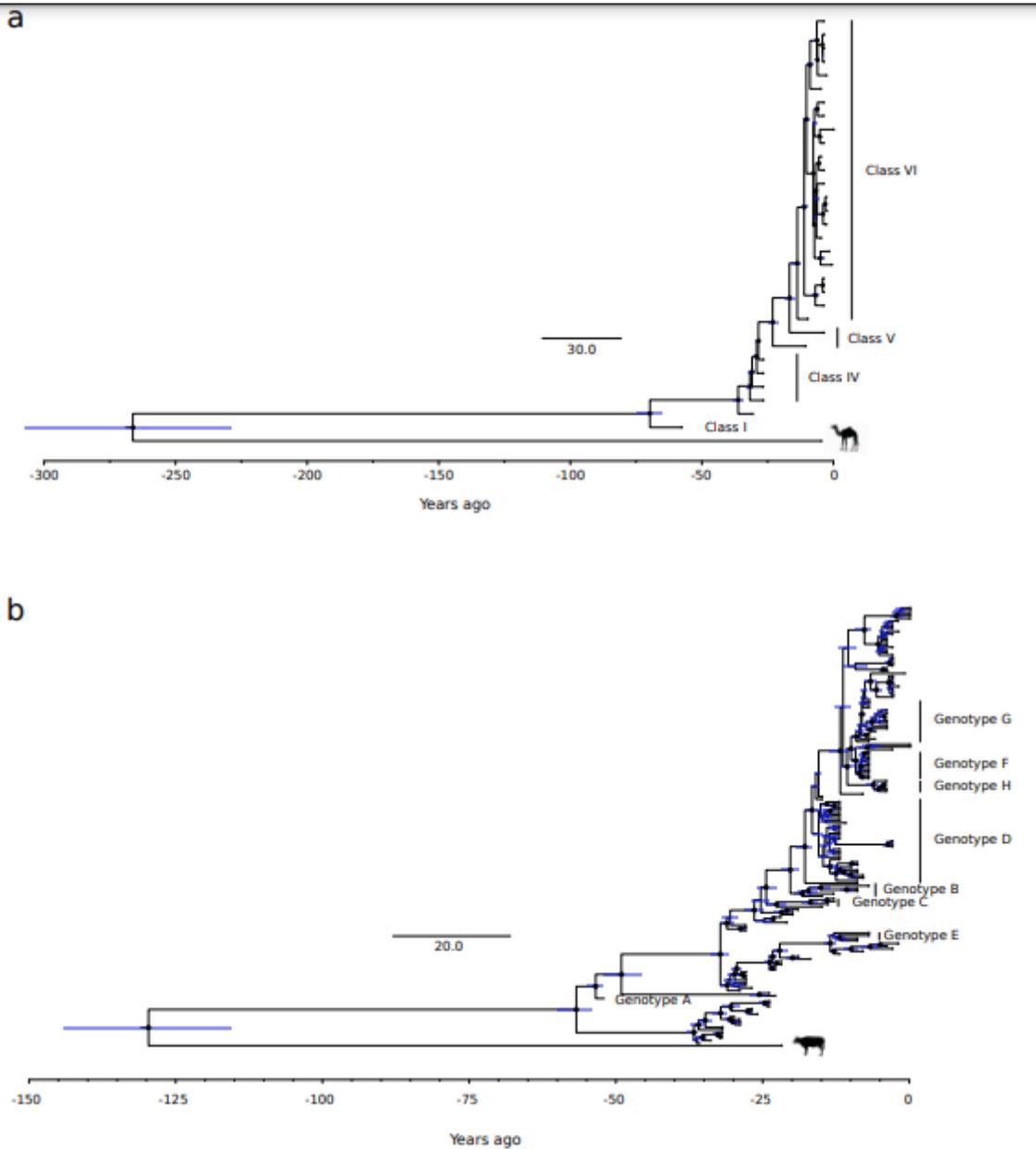


Figure 2

HCoV-229E and HCoV-OC43 timescaled phylogenetic trees. Maximum likelihood trees estimated for the non-recombinant region of HCoV-229E (a) and HCoV-OC43 (b). Branch lengths represent the evolutionary time measured by the grids corresponding to the timescale shown at the tree base (in years). For internal nodes 95% credible interval bars are shown and black dots indicate a posterior probability >0.80 for that node.

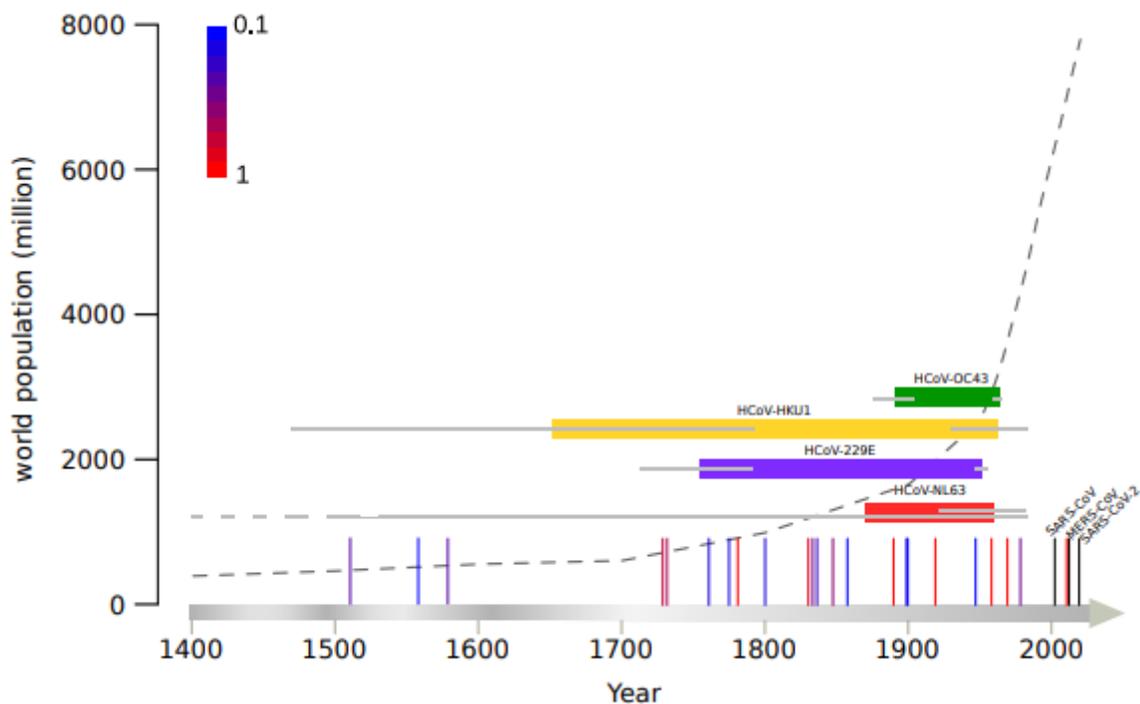


Figure 3

Timeline of endemic coronavirus emergence. Colored horizontal bars represent the time span between the divergence of each coronavirus from the closest known animal virus and the tMRCA (time to the most recent common ancestor) of circulating strains. Gray bars indicate 95% credible intervals. Vertical lines represent influenza pandemic events and the scaled colors (from blue to red, see legend on the plot) indicate the proportion of articles reporting that event as pandemic 80, 130-138. The three recent coronavirus zoonoses are shown as black vertical lines. The world human population count from 1400 C.E. is reported as a dotted gray line.

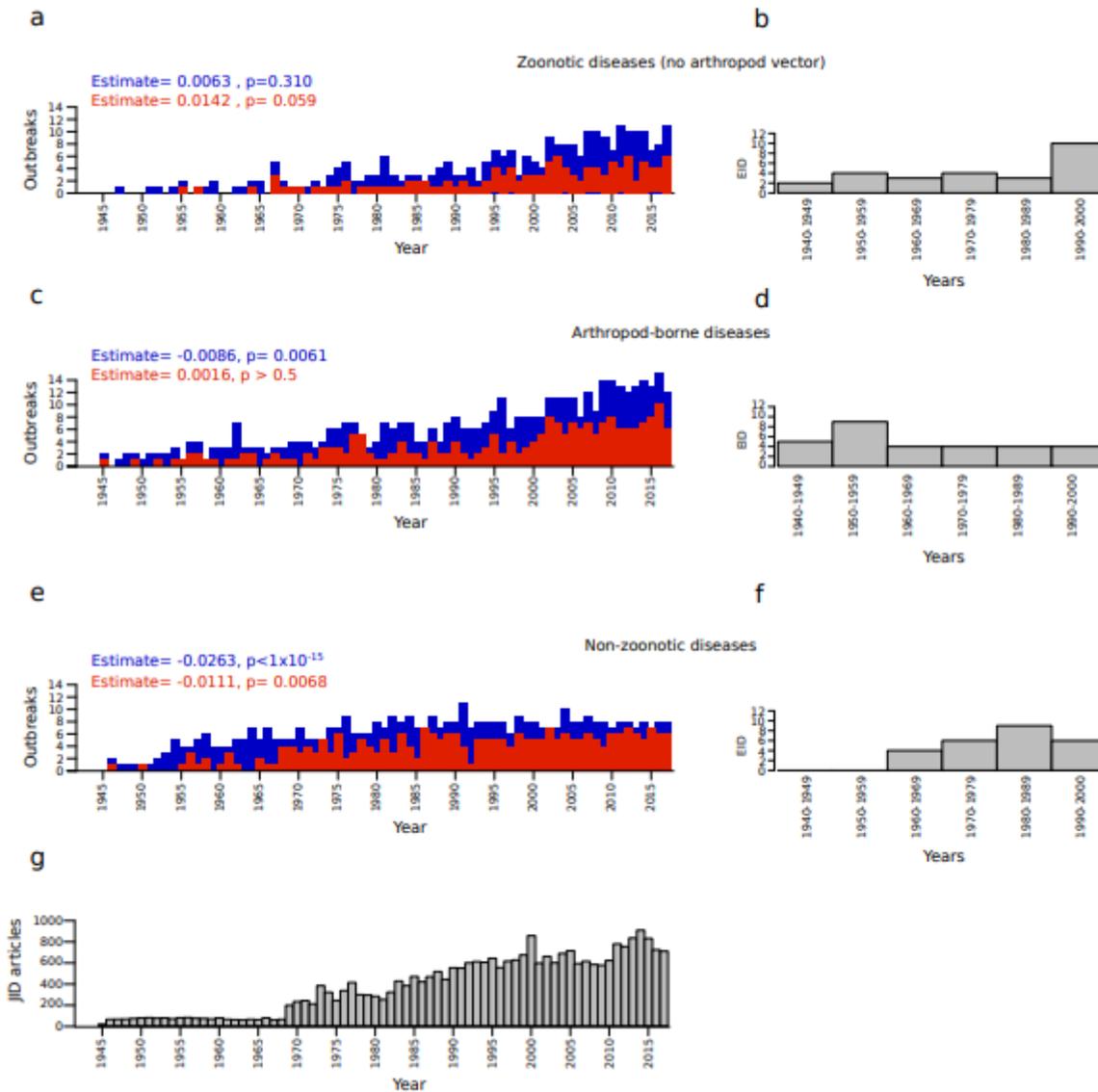


Figure 4

Analysis of viral disease outbreaks and EID events in the last decades. Diseases were divided based on their origin and transmission route: zoonotic with no arthropod vector (a, b), arthropod-borne (c, d), and non-zoonotic (e, f). The overall number of disease outbreaks recorded in the GIDEON database (blue), as well the number of outbreaks involving at least 50 human cases (red), are shown in the left panels. The results of generalized linear models (Glm) offset by $\log(\text{number of JID articles})$ are reported. P values are corrected for three tests (number of disease types we analyzed). Panels b, d, and f show the number of EID events as previously recorded 49. (g) Number of annual articles published in JID.

with a 9-O-acetylated sialic acid. For clarity, only the positively selected sites (orange) on one monomer of the spike protein are shown. Structural details of portions of the spike protein where positively selected sites were identified are also reported. In particular, positively selected sites at domain A, together with the 9-O-acetylated sialic acid binding site are shown. Loop1 and loop2, defining the pocket, are also indicated (L1 and L2, blue labels). Secondary structure features are depicted as grey cartoons, while side chains of positively selected sites are represented as sticks: C atoms are colored in orange, N atoms in blue and O atoms in red. Glycans are rendered as dark blue spheres except for the 9-O-acetylated sialic with its C atoms in green, O atoms in red and N atoms in blue. (c) Sequence logos of the RBD loop1 and loop2 of the spike protein and 177-196 region of the HE protein. Positively selected sites are shown in orange (irrespective of their amino acid frequency). Sequence logos are grouped by collection date.

positively selected sites (in orange) on one monomer of the spike protein are shown. Structural details of portions of the spike protein are also reported, in particular three RBD loops (L1, L2 and L3, blue labels). Secondary structure features are depicted as light-blue cartoons, while side chains of positively selected sites are represented as sticks: C atoms are colored in orange, N atoms in blue and O atoms in red. Glycans are rendered as dark blue spheres. (c) Sequence logos of the three RBD loops. Positively selected sites are shown in orange (irrespective of their amino acid frequency). Sequence logos are grouped by collection date.

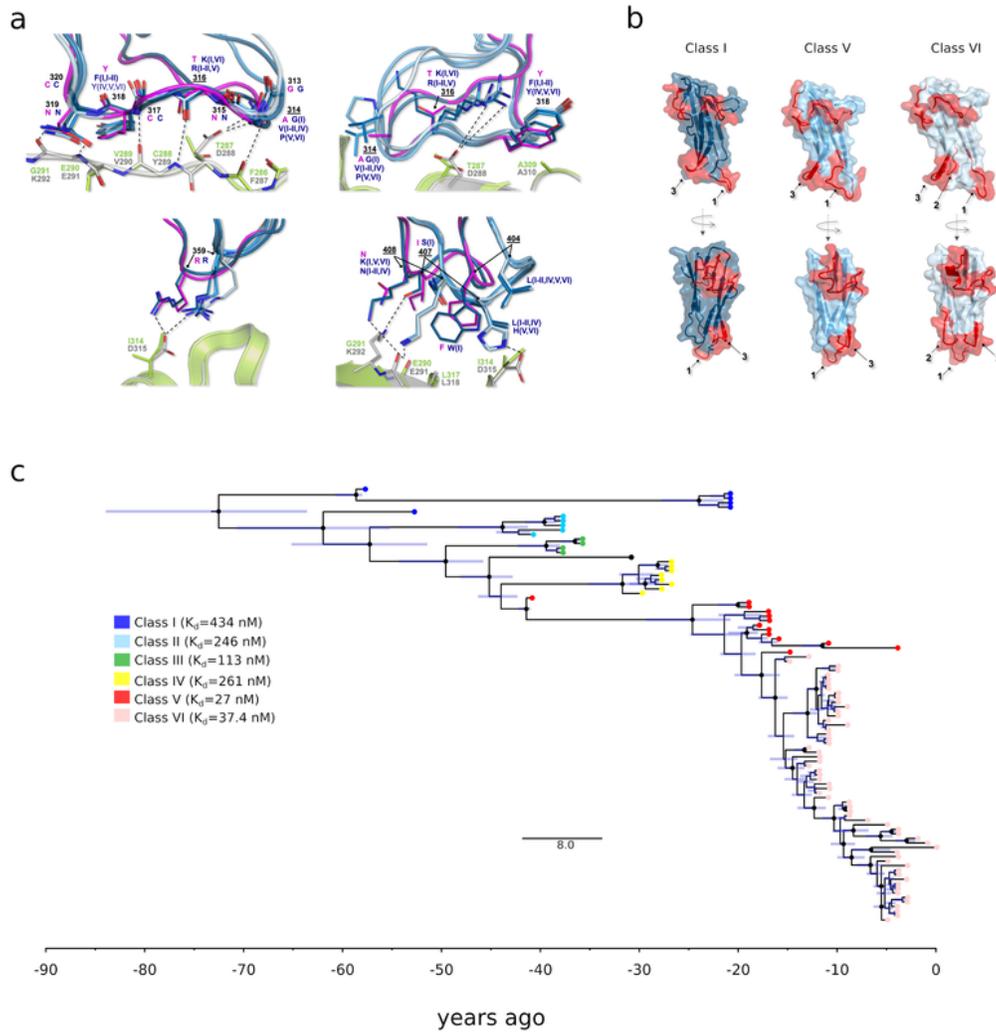


Figure 7

Molecular evolution of HCoV-229E spike protein. (a) Atomic details of the interactions between HCoV-229E (six RBD classes, blue shades) and the corresponding model for camel alphacoronavirus (pink) with both hANPEP (grey) and cANPEP (light green). RBDs-hANPEP conserved interaction pattern, involving loop 1, is shown in the upper-left panel. Relevant structural differences among the various RBDs and the details on their interactions at the binding interface are represented in the other three panels. Positively selected sites are underlined. N atoms are colored in blue, O atoms in red. Salt bridges and H-bonds are represented by dashed black lines. (b) Epitopes mapped on the spike protein RBDs of three different HCoV-E229 classes. The RBDs are in different shades of blue, whereas the epitopes are in red. Labels 1, 2 and 3 refer to loop 1, loop 2 and loop 3. (c) Timescaled maximum likelihood tree of the spike protein RBD. Branch lengths represent the evolutionary time measured by the grids corresponding to the timescale shown at the tree base (in years). For internal nodes 95% credible interval bars are shown and black dots indicate a posterior probability >0.80 for that node. Tip nodes are colored based on the figure legend, where the K_d of all six RBD classes for hANPEP interaction calculated by 57 is also reported.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.pdf](#)