scientific reports

Check for updates

OPEN

A comparative recombination analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic

Simon Pollett[1,2,3], Matthew A. Conte[1], Mark Sanborn[1], Richard G. Jarman[1], Grace M. Lidl[1], Kayvon Modjarrad[4] & Irina Maljkovic Berry[1]✉

The SARS-CoV-2 pandemic prompts evaluation of recombination in human coronavirus (hCoV) evolution. We undertook recombination analyses of 158,118 public seasonal hCoV, SARS-CoV-1, SARS-CoV-2 and MERS-CoV genome sequences using the RDP4 software. We found moderate evidence for 8 SARS-CoV-2 recombination events, two of which involved the spike gene, and low evidence for one SARS-CoV-1 recombination event. Within MERS-CoV, 229E, OC43, NL63 and HKU1 datasets, we noted 7, 1, 9, 14, and 1 high-confidence recombination events, respectively. There was propensity for recombination breakpoints in the non-ORF1 region of the genome containing structural genes, and recombination severely skewed the temporal structure of these data, especially for NL63 and OC43. Bayesian time-scaled analyses on recombinant-free data indicated the sampled diversity of seasonal CoVs emerged in the last 70 years, with 229E displaying continuous lineage replacements. These findings emphasize the importance of genomic based surveillance to detect recombination in SARS-CoV-2, particularly if recombination may lead to immune evasion.

The emergence of SARS-CoV-2 has generated interest in role of recombination in the evolution of this and other human coronaviruses (hCoV). Recombination has been observed in many RNA viruses and is noted to occur at a higher frequency in positive-sense RNA viruses, a category that includes SARS-CoV-2 and other medically important coronaviruses[1–3]. From an evolutionary biology perspective, it remains unclear why recombination occurs in RNA viruses. Hypotheses include recombination being an incidental outcome of RNA polymerase function, or an evolutionary favorable purge of deleterious genotypes and/or generation of advantageous genotypes[1].

While several studies have examined the putative role of recombination in the zoonotic emergence of SARS-CoV-2[4–6], few have focused on the emergence of recombination during the first year of the SARS-CoV-2 pandemic. To date, none have systematically examined the genetic propensity of recombination across all human coronaviruses in order to predict the possible evolutionary future of SARS-CoV-2, despite prior observations of recombination in OC43-hCoV, HKU1-hCoV, NL63-hCoV, and MERS-CoV[7–19].

RNA virus recombination has been associated with changes in host range, host response and virulence[1]. Identifying the presence of recombination, or predicting the risk of recombination, in viral populations of SARS-CoV-2 is critical for several reasons. First, circulating recombinants may complicate molecular diagnostic performance. Second, recombinants may cause rapid escape from naturally acquired immunity, as has been observed in the norovirus genus, which has caused pandemics due to the rapid emergence of new genotypes generated by recombination of structural genes[20]. For SARS-CoV-2, such an event may have major implications, especially if circulating recombinant results in escape from both natural and vaccine induced immunity[21]. Finally, genomic epidemiology has increasingly been shown to be an important public health tool for SARS-CoV-2 and failing to accommodate for recombinant data may lead to incorrect epidemiological inference because of possible phylogenetic incongruence[22,23].

We undertook a comparative recombination analysis of all published SARS-CoV-1, SARS-CoV-2, MERS-CoV and seasonal hCoV (OC43, 229E, NL63 and HKU1) genomes. Specifically, we aimed to identify the frequency

[1]Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, MD, USA. [2]Infectious Disease Clinical Research Program, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, MD, USA. [3]Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc, Bethesda, MD, USA. [4]Emerging Infectious Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, MD, USA. ✉email: irina.maljkovicberry.ctr@mail.mil

Coronavirus species	*n* genome sequences	Recombination events detected by any method	Recombination events detected by ≥ 3 methods	Recombination events detected by ≥ 3 methods and without another evolutionary process possibly explaining the recombination signal (high evidence)	Recombination events with high level evidence seen in multiple genomes
hCoV-229E	22	4	3	1	1
hCoV-NL63	65	31	24	14	7
hCoV-OC43	138	23	16	9	6
hCoV-HKU1	37	14	9	1	1
MERS-CoV	365	12	10	7	6
SARS-CoV-1	49	1	0	0	0
SARS-CoV-2	100296[a]	33	8	0	0

Table 1. Frequency of recombination events detected in 229E, NL63, OC43, HKU1, MERS-CoV, SARS-CoV1 and SARS-CoV-2, stratified by level of evidence. [a]Randomly subsampled into 100 × n = 300 independent datasets.

and genomic location of recombination events in these hCoV types, and estimate the impact of these recombinants on hCoV emergence dates (TMRCA). As part of this analysis, we reconstructed the time-scale of circulating seasonal hCoV evolution and lineage replacement to provide insights into the possible future evolutionary trajectory of SARS-CoV-2.

Results

There is moderate evidence for recombination in several SARS-CoV-2 genomes through October 2020. Among SARS-CoV-2 genomes, we detected a total of 8 recombination events detected by at least three detection methods, though these events were possibly caused by another non-recombinant process (Table 1). Some of these events, while not supported by a high level of evidence, were noted in multiple sequences and may therefore represent circulating recombinant forms (Supplementary Table S1). However, these recombination events were not found across all subsampled datasets. For those recombination events with moderate evidence, we noted that half of the events (4/8) comprised breakpoints in the non-ORF1 genes, and one quarter (2/8) occurred within the spike gene (Supplementary Table S2, Fig S6). None were positioned in the furin cleavage site. The GC content corresponding to these breakpoints was variable (range 0 to 100%, IQR = 30–60%) but typically low (median GC content = 40%); breakpoints occurred at locations with variable predicted RNA secondary structure with icSHAPE in vivo scores ranging from 0 to 1 (Table S6).

Recombination is relatively frequent in seasonal endemic coronaviruses and has a propensity for the non-ORF1 genes. Within the 229E, OC43, NL63 and HKU1 datasets, we noted 1, 9, 14, and 1 high confidence recombination events, respectively (Table 1). These recombination events were found in 13.6%, 20.3%, 89.2%, and 35.1% of the analyzed sequences, respectively. In the OC43 and NL63 datasets, we noted significantly more breakpoints in the non-ORF1 region of the genome containing structural genes than in the ORF1 ($p = 0.0004$, $p = 0.0032$ respectively) (Fig. 1).

Recombinants were noted across entire clades of HKU1, NL63 and OC43 viruses, with clusters of genomes sharing identical recombination patterns, indicating spread of a recombinant virus following its emergence through a recombination event. The OC43 and NL63 datasets also contained genomes with unique recombination patterns (singletons) (Figs. 2, 3). Furthermore, we noticed some singletons falling within already recombinant clades, indicating presence of successful second generation recombination (Figs. 2, 3). A 10th OC43 event with moderate recombination evidence based on RDP4 results showed topological incongruence in subgenomic phylogenetic trees and was therefore removed for further time-scaled analyses.

MERS-CoV but not SARS-CoV-1 is characterized by frequent recombination with a propensity for non-ORF1 genes. Within the MERS-CoV dataset, we detected 7 recombination events with a high level of confidence (detected by 3 or more methods and not explained by another evolutionary process) (Table 1). These recombination events were found in 14.5% of all analyzed MERS-CoV genomes. Of these, 6 were found across clades, suggesting recombinants were sufficiently fit for onward transmission (Fig. 4). We noted recombinant clades defined by camel hosts, as well as camel and human hosts (Fig. 4), suggestive of inter-host recombinant spread. Moreover, we noted significantly more breakpoints in the non-ORF1 region of the genome containing structural genes compared to ORF1 ($p < 0.001$) (Fig. 1). We noted only a single low evidence recombination event involving the structural region of the SARS-CoV1 genome (4503 -25,998 nt), but this was not confirmed by multiple methods (Table 1).

Recombination in seasonal coronaviruses substantially alters estimates of temporal structure. Our approach to identification of recombinants and their subsequent removal from the datasets led to major improvements in estimated temporal structure of hCoV-OC43, hCoV-NL63 and hCoV-HKU1 (Table 2, Fig S7–S10). These changes involved both the regression coefficient and the regression intercept, which serve as crude estimates of evolutionary rates and TMRCA, and are often used to assess the clock-likeness (linear

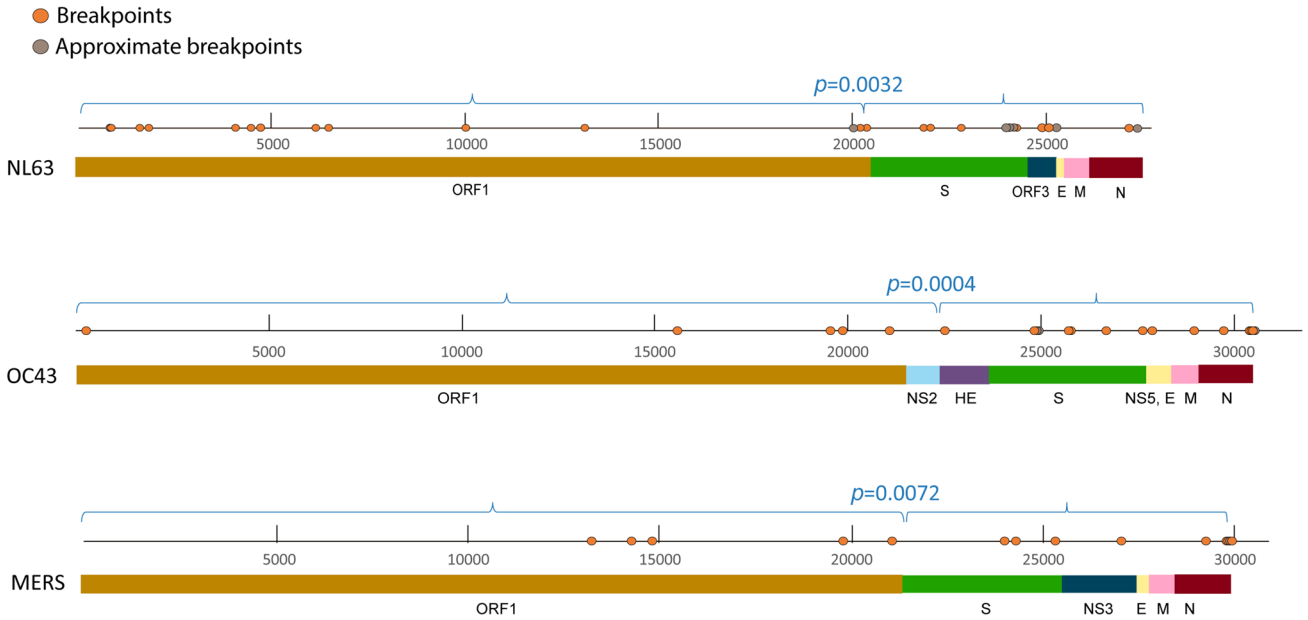


Figure 1. Estimated recombination breakpoint positions NL63, OC43 and MERS-CoV whole genomes. *p* values for the frequency of recombination breakpoints in the non-ORF1 region (containing the structural genes) versus the ORF1 region are derived by the χ^2 test. Approximate breakpoints are breakpoints that could not be placed with certainty due to overlapping recombination or other reasons.

relationship of genetic distance across sequence sampling times) of the data for further analyses (Table 2, Fig S7–S10). In contrast, removal of the single 229E recombinant did not cause substantive change in estimated temporal structure as estimated by regression coefficient and TMRCA (Table 2).

The current global diversity of seasonal hCoVs arose across the last 70 years. We estimated a TMRCA date (year A.D) of 1989, 1970, 1964, and 1951 for the 229E, OC43, NL63 and HKU1 lineages, respectively, indicating relatively recent emergence of the current seasonal hCoV lineages (Table 3, Supplementary Figures S1–S4). Importantly, these do not necessarily represent de novo emergence of these viruses from animal origins but rather may represent the divergence from older OC43, NL63 and HKU1 lineages, respectively. Indeed, the TMRCA for 229E is preceded by historical descriptions of the circulation of this virus in humans, which was discovered in 1962[24]. We therefore extended our 229E full genome analysis and inferred TMRCA estimates from partial genome datasets of the complete N gene (N = 101, length = 1167nt), the complete S gene (Supplementary Figure S5) (N = 78, length = 3522nt), the RBD (S1) domain S gene (N = 89, length = 1650nt), and concatenated S and N genes (N = 63, length = 4689nt). These datasets excluded the original 229E strain isolate (ACTT-VR-740, GenBank accession DQ243963.1) due to likely lab adaptation changes. These yielded TMRCA point estimates between 1966 and 1975 (229E RBD S1 gene TMRCA = 1966.2; 229E S gene TMRCA = 1975; 229E N gene TMRCA = 1973.9; 229E N and S gene TMRCA = 1974.1).

Discussion

We performed a comprehensive recombination analysis across all medically important human coronaviruses, with an overarching aim to identify the current and future risk of recombinant emergence in the SARS-CoV-2 pandemic. We note moderate evidence for SARS-CoV-2 recombination during the first year of the COVID-19 pandemic. In other hCoV species, we note that recombination has a predilection for the non-ORF1 region of the genome containing structural genes, and is relatively frequent in most medically significant hCoV over a relatively short evolutionary timescale. These findings are timely with the announcement of a more recent unpublished recombinant SARS-CoV-2 strain[25], as well as the recent focus on how insights can be learned from the functional implications of antigenic evolution now demonstrated in seasonal coronaviruses[26].

From more than 100,000 SARS-CoV-2 genomes, we noted 8 instances of recombination with moderate confidence. Two of these events were noted in the spike gene; none involved the furin cleavage site. In our analysis the SARS-CoV-2 estimated breakpoints were found in regions with typically low but variable GC content (median = 40%, IQR = 30–60%, range 0–100%), and breakpoints occurred at locations with variable predicted RNA secondary structure with icSHAPE scores ranging from 0 to 1.

However, all SARS-CoV-2 recombination events were flagged by the RDP4 software as being possibly driven by other processes despite support by three or more recombination detection methods. This may reflect the relatively lower viral diversity across the first year of the pandemic. Similarly, we did not detect a high-confidence recombination signal in SARS-CoV-1, a virus with a limited temporal distribution. In contrast, we show that recombination was relatively frequent in seasonal coronavirus and MERS-CoV datasets comprising a longer period of sampling, including recombinants sufficiently fit for onward transmission. Furthermore, our analyses

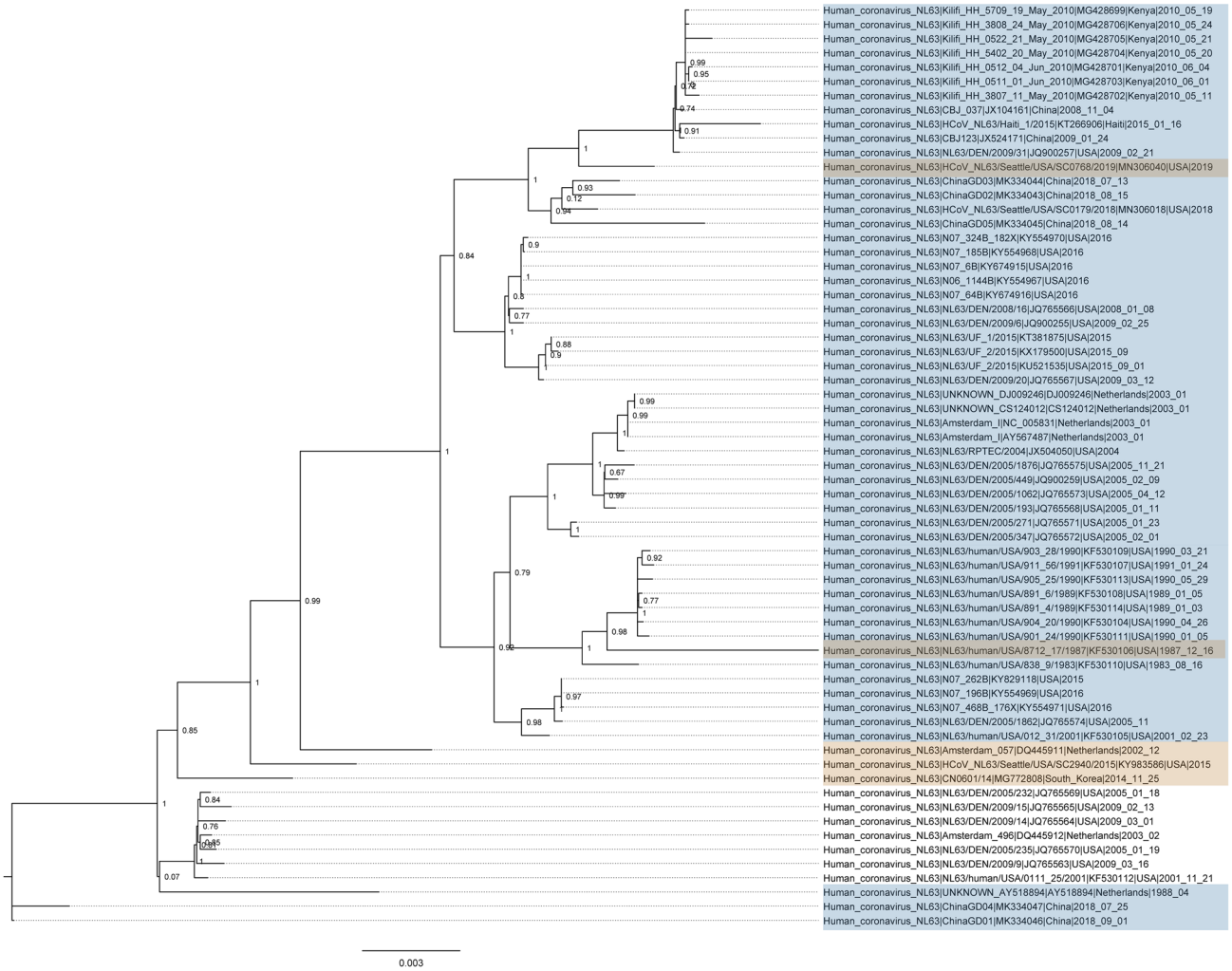


Figure 2. Maximum likelihood phylogeny of recombinants in NL63. Scale represents nucleotides per site. Recombinant events with multiple genomes are marked in blue, or as singletons are marked in yellow. Phylogeny was rooted with a 229E outgroup (removed for clarity).

show that three of the coronaviruses (MERS, OC43 and NL63) had breakpoint propensity for the non-ORF1 region of the genome containing structural genes. Several reasons for this may exist, such as inherent structural similarities of the coronaviruses in this region causing enhanced enzyme slippage, or positive selection pressure. The latter may be correlated with evasion of the human immune system, although recombination in RNA viruses is not generally thought as an evolutionary process which is driven by natural selection to favor advantageous genotypes[1].

Our endemic coronavirus analyses also highlighted that recombination affected the estimated temporal structure of coronavirus sequence datasets, particularly in the OC43 and NL63 types. This serves as a caution for genomic epidemiology studies which do not identify and account for recombinants in SARS-CoV-2 and other coronavirus analyses. Indeed, recombination has long been known to be a cause of phylogenetic incongruence for other viruses[23]. Our time-scaled evolutionary analyses, adjusted for recombination to restore a molecular clock signal, yielded insights into the recent epidemiology of seasonal coronaviruses. We noted that the current sampled diversity of seasonal coronaviruses has emerged within a 70 year period, punctuated by new lineage emergence at intervals ranging from 5 to 20 years. For certain seasonal coronaviruses the uncertainty interval of these TMCRA estimates did overlap with the first reported cases. However, historical epidemiological data on the time-scales of human coronavirus emergence also have uncertainty. For instance, while OC43 was first isolated from a human case in 1967, potentially cross reactive sera have been identified as far back as 1965[27,28]. Moreover, caution is required in inferring that these viruses spilled over into humans at these timepoints, however, as this may reflect the divergence of new lineages from prior, unsampled older viral populations circulating in humans. In the case of 229E hCoV, we noted that the full genome TMRCA estimates were preceded by the clinical reports of 229E infection in the 1960s[24], and a spike gene analysis incorporating older partial genome sequences from the earlier twentieth century showed lineage extinction and replacement. This is consistent with previous analyses suggesting that 229E evolution is characterized by prior lineage extinction and new lineage emergence[29].

It is important to note that our RNA genomic recombination detection in sequence data remains a statistical estimation only, as previously discussed in detail for dengue viruses[30]. In addition, the size of our SARS-CoV-2

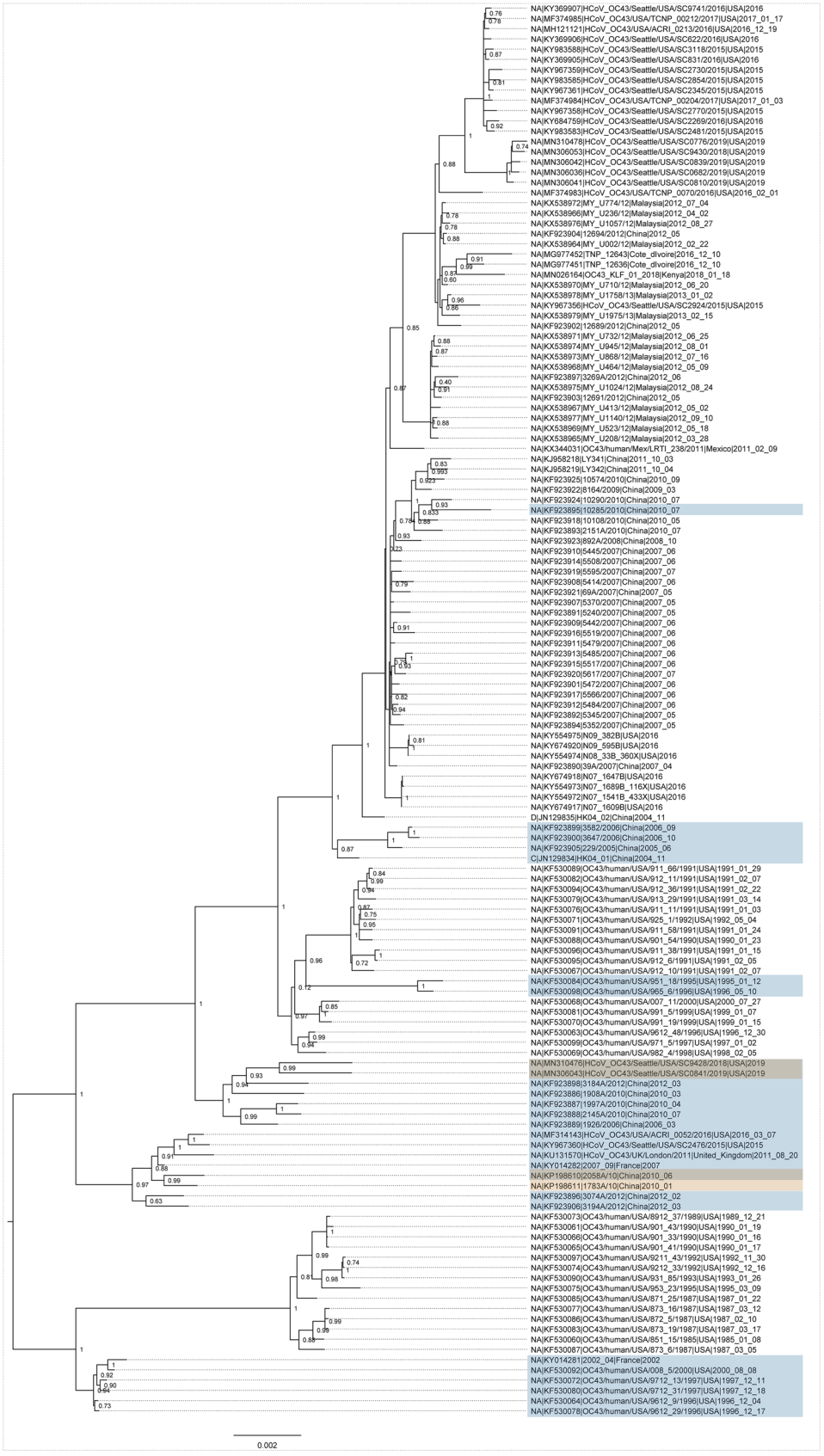


Figure 3. Maximum likelihood phylogenies of recombinants in OC43. Scale represents nucleotides per site. Recombinant events with multiple genomes are marked in blue, or as singletons are marked in yellow. Phylogeny was rooted with an HKU1 outgroup (removed for clarity).

Figure 4. Maximum likelihood phylogeny of recombinants in MERS-CoV. Scale represents nucleotides per site. (**a**) Taxa colored by host (camel = black, human = green). (**b**) Colored taxa indicate confirmed recombinant clades. ▶

dataset was computationally prohibitive to perform recombination detection across all data, which might have resulted in missing additional recombination events. Also, unsampled data are a pervasive technical risk for recombination detection, as demonstrated by the variable finding of recombinant events across subsampled SARS-CoV2 data. Finally, our strict criteria for identifying recombination events with high confidence may have resulted in the removal of some true recombinant events.

Still, these findings provide critical insights into the possible projected evolution of SARS-CoV2. Recombination in other RNA viruses has been associated with changes in host tropism, virulence or epidemiology[1]. Ongoing genomic-based COVID-19 surveillance has recently been highlighted as a critical public health tool to detect novel SARS-CoV-2 variants, such as the B.1.1.7, B.1.351, and P.1 variants[31,32]. Robust genomic surveillance will be essential for the timely detection of recombination in SARS-CoV-2, which may have implications for the diagnosis of and immunity to this pandemic pathogen.

Methods

Data curation and alignment. Full genomes of endemic seasonal hCoVs, MERS-CoV and SARS-CoV-1 were downloaded from the NIAID Virus Pathogen Database and Analysis Resource (ViPR)[33,34]. Specifically, we obtained n = 22 hCoV-229E whole genomes, n = 138 hCoV-OC43 whole genomes, n = 68 hCoV-NL63 whole genomes, n = 37 HKU1 whole genomes, n = 365 MERS-CoV whole genomes (including human and camel host), and n = 49 SARS-CoV-1 whole genomes. We excluded laboratory constructs and sequences without host, collection date and location history. Datasets were aligned using MAFFT[35], with manual alignment thereafter in MEGA v6.0[36]. Alignment lengths were adjusted such that all genomes from one alignment were of approximately equal length (OC43 = 30,639 nt; NL63 = 27,483 nt; HKU1 = 29,892 nt; 229e = 27,292 nt; MERS = 29,985 nt; SARS1 = 29,719 nt). Partial genomes were excluded. Accession numbers for these data are presented in Supplementary Table S3.

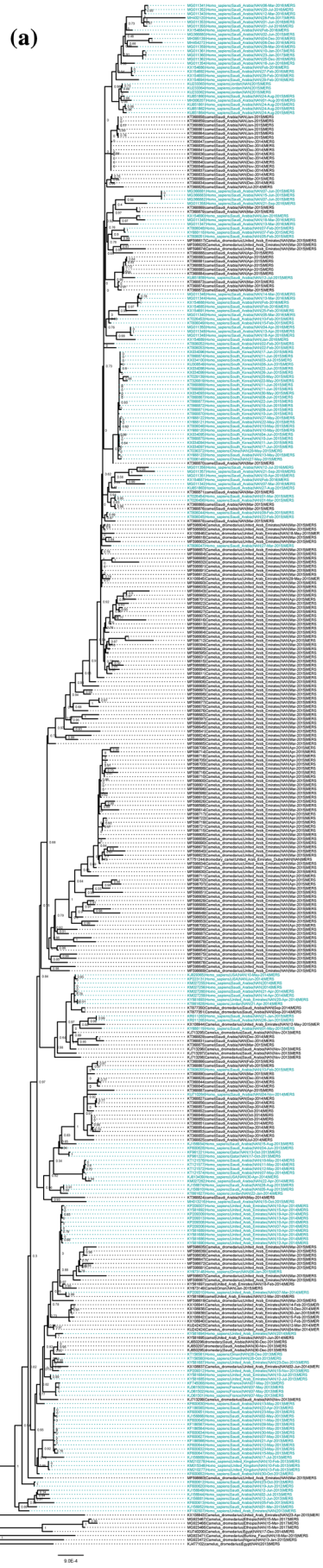
All available full SARS-CoV-2 genomes (n = 157,439) up to October 23, 2020 were downloaded from the GISAID database[37]. These data were curated by removing (i) any partial genomes (< 90% full genome length), (ii) any genomes with > 100 continuous ambiguous base calls (Ns), and (iii) removing bat and pangolin non-human genomes. A single genome (Italy/CAM-IZSM-45946/2020) was removed due to the presence of non-IUPAC nucleotide codes. The remaining 100,296 genomes were aligned to the MN908947.3 reference using mafftc (7.471) in the MAFFT software[35]. Alignment ends (first 265 and last 229 nt) were trimmed using the trimal command[38]. This yielded a final alignment of length 29,409 bp.

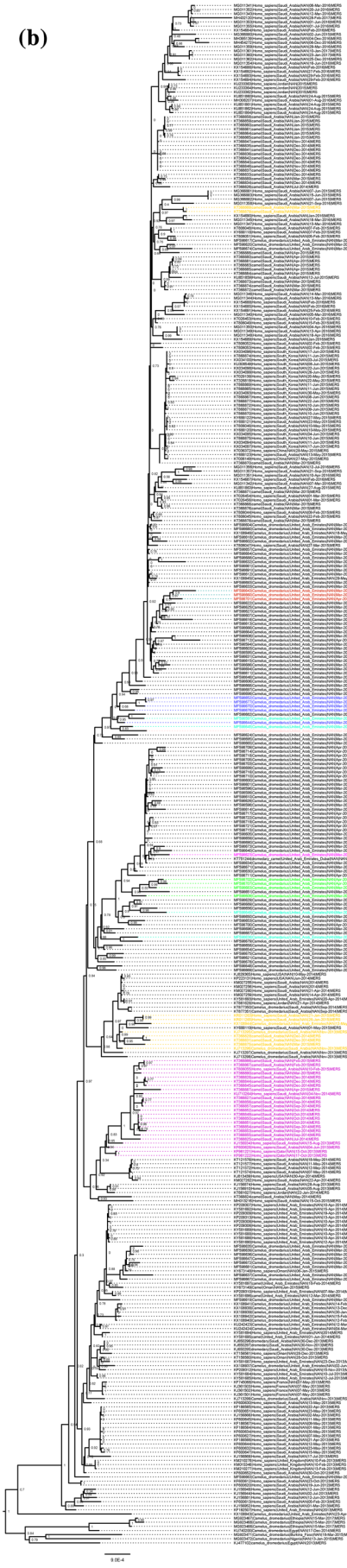
Recombination detection and determination of breakpoints. Datasets underwent recombination detection using the RDP4 software[39]. Recombination signal detection was performed with a suite of methods: the original RDP method[40], BOOTSCAN[41], MAXCHI[42], CHIMAERA[43], 3SEQ[44], GENECONV[45], SISCAN[46]. Following the detection of a 'recombination signal' with these methods, the approximate breakpoint positions were determined using a hidden Markov model, BURT, and the recombinant sequence identified using the PHYLPRO[47], and VISRD[48] methods.

For SARS-CoV2, due to the prohibitive computational demand of recombination detection analysis in a dataset of this size, we randomly subsampled, with replacement, n = 300 sequences with 100 × iterations (n = 30,000 full genomes) to perform recombination detection.

As individual recombination detection methods may have limited specificity, we developed a customized framework to ascertain the level of evidence for those recombination events detected in these datasets. Recombination events identified by only one or two methods in RDP4 were assigned a 'low' level of confidence, and those identified by at least three methods in RDP4 were assigned at least a 'moderate' level of confidence. We further assigned a 'high' level of confidence for those recombination events identified by at least three RDP4 methods with no other identified process which may have explained the recombination signal[39]. The evolutionary processes that might be misinterpreted as recombination are typically caused by a combination of mutation rate variation between sites superimposed on mutation rate variation along lineages (M. Darren, personal communication, Dec 4, 2020). For those recombination events with a high level of evidence, the location of breakpoints were plotted across the whole genome and a χ^2 test used to compare frequency of breakpoints at non-ORF1 region of the genome containing structural genes, versus non-structural ORF1. Breakpoints that could not be placed with certainty due to overlapping recombination or other reasons were not included in the analyses. We determined both the number of recombination events by hCOV type, in addition to whether such events were found in multiple genomes for that hCoV type. A single recombination event can contain several genomes, meaning that the recombination occurred within the ancestor of these genomes, followed by its successful spread. Therefore, the number of recombinant genomes may be higher than the number of observed recombination events.

Estimation of temporal structure of hCoV with and without recombinant strains. Datasets for each hCoV species underwent nucleotide model substitution selection using JModelTest2[49], with model selection as follows: 229E = SYM + I + G, HKU1 = GTR + G + I, NL63 = GTR + I, OC43 = GTR + G + I. A maximum likelihood phylogeny was inferred using the PhyML software[50], with aLRT for node support and tips labeled by date of collection. Root-to-tip regression was performed using the Temp-Est tool[51], with slope coefficient and


(a)


(b)

Lineage	Sequences (n)	Date range (years)	Slope coefficient[a]	Intercept (TMRCA)[b]
Recombinants not removed				
hCoV-229E	22	26.36	2.69×10^{-4}	1990 A.D
hCoV-NL63	65	35.38	-1.00×10^{-4}	2149 A.D
hCoV-OC43	138	33.98	−0.00	27,359 A.D
hCoV-HKU1	37	14.17	4.52×10^{-4}	1941 A.D
Recombinants removed				
hCoV-229E	19	26.36	2.65×10^{-4}	1990 A.D
hCoV-NL63*	56	35.04	7.68×10^{-5}	1944 A.D
hCoV-OC43	110	34.00	2.83×10^{-4}	1967 A.D
hCoV-HKU1	24	13.41	1.00×10^{-3}	1978 A.D

Table 2. Root-to-tip regression coefficient and intercept of seasonal hCoV phylogenies with and without recombinants removed. [a]Approximates evolutionary rate (substitutions/site/year). [b]Approximates TMRCA. *Non-recombinant region was used, with genomes containing breakpoints in this region removed (N = 9).

	TMRCA (A.D)	Lower 95% HPD	Upper 95% HPD	Nucleotide Subst Model	Clock model	Demographic model
229E[a]	1989	1988	1990	SYM + I + G	Strict	Constant
HKU1[a]	1951	1842	1998	GTR + G + I	UCLN	Bayesian Skyline
NL63[b]	1964	1945	1978	HKY + I	UCLN	Bayesian Skyline
OC43[a]	1970	1960	1978	GTR + I	UCLN	Bayesian Skyline

Table 3. Bayesian TMRCA estimates for 229E, HKU1, NL63 and OC43[a]. [a]Recombinant genomes removed. [b]Non-recombinant region 13093-20198. UCLN, uncorrelated lognormal; TMRCA, time to most recent common ancestor.

intercept values used as preliminary estimates of evolutionary rate and time-to-most-common recent ancestor (TMRCA), respectively. Confirmed recombinant sequences were annotated on these phylogenies to identify recombinant clades and singletons. Phylogenetic analyses were repeated with the recombinants removed to estimate the impact of recombination on time-scales and temporal structure of hCoV evolution. For NL63, removal of recombinant genomes resulted in a small dataset and a small remaining fraction of the initial phylogenetic tree. Therefore, NL63 genomes were screened for a common region without any recombination signal. A region of 7105 nts (region 13093-20198 of the alignment) was found in which most genomes (n = 56) had no recombination signal, and this region was used for subsequent time-scaled phylogenetic analyses, with the 9 genomes with recombination breakpoints in this region removed from the dataset.

Estimating time-scales of emergence of the currently circulating seasonal hCoV lineages. We leveraged recombination-free endemic seasonal hCoV datasets, in addition to removing other root-to-tip regression outlier genomes (n = 3 + 4 identical genomes for OC43, n = 1 for NL63, n = 0 for 229e, n = 2 for HKU1), to reconstruct the time-scale of emergence of currently circulating seasonal hCoV lineages across the 229E, OC43, NL63 and HKU1 types. We focused on these coronaviruses because they are well established viruses within human populations and may serve as a model for the projected evolutionary future of SARS-CoV-2. This is in contrast to the now extinct SARS-CoV-1 virus, and MERS-CoV, the latter which continues to be defined by more sporadic and discrete spillover events[52].

Time-scaled genealogies of these viruses were inferred using the BEAST software 1.8.4[53]. To minimize the risk of model misspecification, we inferred maximum clade credibility phylogenies with combinations of demographic models (constant, exponential and skyline population models) and molecular clock models (strict versus relaxed) (Supplementary Table S4). For each hCoV dataset, the optimal combination of demographic and molecular clock model was selected by logarithmic marginal likelihoods inferred by the path-sampling/ stepping-stone method[54].

Predicted features of underlying RNA template near putative SARS-CoV-2 recombination breakpoints. GC content around putative SARS-CoV-2 breakpoints was determined by calculating the percentage of G (guanine) and C (cytosine) bases in 5-base windows via the UCSC SARS-CoV-2 genome browser tool using 5 nt windows[55,56]. Secondary RNA structure at putative SARS-CoV-2 breakpoints was predicted using selective 2-hydroxyl acylation and profiling experiment (SHAPE) reactivity[57], SHAPE Shannon entropy (comSuperFold)[57], and icSHAPE in vivo scores via the UCSC SARS-CoV-2 genome browser[58].

Disclaimer. Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The View(s) expressed are those of the authors and do not necessarily reflect the official views of the Uniformed Services University of the Health Sciences, Henry M. Jackson Founda-

tion for the Advancement of Military Medicine, Inc., Department of Health and Human Services, the National Institutes of Health, the Departments of the Army, Navy or Air Force, the Department of Defense, or the U.S. Government.

Data availability

All genomes are available in public sequence repositories (GISAID and ViPR) and their accession numbers are reported in Supplementary Information.

Received: 17 March 2021; Accepted: 9 August 2021
Published online: 30 August 2021

References

1. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine?. *Nat. Rev. Microbiol.* **9**, 617–626. https://doi.org/10.1038/nrmicro2614 (2011).
2. Dhama, K. *et al.* Coronavirus disease 2019-COVID-19. *Clin. Microbiol. Rev.* **33**, e0002820. https://doi.org/10.1128/cmr.00028-20 (2020).
3. Cheng, V. C., Lau, S. K., Woo, P. C. & Yuen, K. Y. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clin. Microbiol. Rev.* **20**, 660–694. https://doi.org/10.1128/cmr.00023-07 (2007).
4. Neches, R. Y., McGee, M. D. & Kyrpides, N. C. Recombination should not be an afterthought. *Nat. Rev. Microbiol.* **18**, 606. https://doi.org/10.1038/s41579-020-00451-1 (2020).
5. Paraskevis, D. *et al.* Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect. Genet. Evol.* **79**, 104212. https://doi.org/10.1016/j.meegid.2020.104212 (2020).
6. Wu, A. *et al.* Mutations, Recombination and Insertion in the evolution of 2019-nCoV. *bioRxiv* https://doi.org/10.1101/2020.02.29.971101 (2020).
7. 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. https://doi.org/10.1128/jvi.79.3.1595-1604.2005 (2005).
8. Zhang, Y. *et al.* Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination. *J. Infect.* **70**, 641–650. https://doi.org/10.1016/j.jinf.2014.12.005 (2015).
9. Zhang, Z., Shen, L. & Gu, X. Evolutionary dynamics of MERS-CoV: Potential recombination, positive selection and transmission. *Sci. Rep.* **6**, 25049. https://doi.org/10.1038/srep25049 (2016).
10. Zhang, X. W., Yap, Y. L. & Danchin, A. Testing the hypothesis of a recombinant origin of the SARS-associated coronavirus. *Arch. Virol.* **150**, 1–20. https://doi.org/10.1007/s00705-004-0413-9 (2005).
11. Liu, P. *et al.* Prevalence and genetic diversity analysis of human coronaviruses among cross-border children. *Virol. J.* **14**, 230. https://doi.org/10.1186/s12985-017-0896-0 (2017).
12. Kin, N., Miszczak, F., Lin, W., Gouilh, M. A. & Vabret, A. Genomic analysis of 15 Human coronaviruses OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes. *Viruses* **7**, 2358–2377. https://doi.org/10.3390/v7052358 (2015).
13. Lau, S. K. *et al.* Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. *J. Virol.* **89**, 10532–10547. https://doi.org/10.1128/jvi.01048-15 (2015).
14. Lau, S. K. *et al.* Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. *J. Virol.* **85**, 11325–11337. https://doi.org/10.1128/jvi.05512-11 (2011).
15. Dominguez, S. R. *et al.* Genomic analysis of 16 Colorado human NL63 coronaviruses identifies a new genotype, high sequence diversity in the N-terminal domain of the spike gene and evidence of recombination. *J. Gen. Virol.* **93**, 2387–2398. https://doi.org/10.1099/vir.0.044628-0 (2012).
16. Pyrc, K. *et al.* Mosaic structure of human coronavirus NL63, one thousand years of evolution. *J. Mol. Biol.* **364**, 964–973. https://doi.org/10.1016/j.jmb.2006.09.074 (2006).
17. Woo, P. C. *et al.* Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1. *J. Virol.* **80**, 7136–7145. https://doi.org/10.1128/jvi.00509-06 (2006).
18. Sabir, J. S. *et al.* Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* **351**, 81–84. https://doi.org/10.1126/science.aac8608 (2016).
19. Wang, Y. *et al.* Origin and Possible Genetic Recombination of the Middle East Respiratory Syndrome Coronavirus from the First Imported Case in China: Phylogenetics and Coalescence Analysis. *MBio* **6**, e01280-01215. https://doi.org/10.1128/mBio.01280-15 (2015).
20. Eden, J. S., Tanaka, M. M., Boni, M. F., Rawlinson, W. D. & White, P. A. Recombination within the pandemic norovirus GII.4 lineage. *J. Virol.* **87**, 6270–6282. https://doi.org/10.1128/jvi.03464-12 (2013).
21. Dearlove, B. *et al.* A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. USA* **117**, 23652–23662. https://doi.org/10.1073/pnas.2008281117 (2020).
22. Montoya, V. *et al.* Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing. *Infect. Genet. Evol.* **43**, 329–337. https://doi.org/10.1016/j.meegid.2016.06.015 (2016).
23. Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F. & González-Candelas, F. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.* **30**, 296–307. https://doi.org/10.1016/j.meegid.2014.12.022 (2015).
24. Hamre, D. & Procknow, J. J. A new virus isolated from the human respiratory tract. *Proc. Soc. Exp. Biol. Med.* **121**, 190–193. https://doi.org/10.3181/00379727-121-30734 (1966).
25. https://www.newscientist.com/article/2268379-two-coronavirus-variants-have-merged-heres-what-you-need-to-know/, cited Feb 20 2021.
26. Eguia, R. T. *et al.* A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog.* **17**, e1009453. https://doi.org/10.1371/journal.ppat.1009453 (2021).
27. McIntosh, K., Becker, W. B. & Chanock, R. M. Growth in suckling-mouse brain of "IBV-like" viruses from patients with upper respiratory tract disease. *Proc. Natl. Acad. Sci. USA* **58**, 2268–2273. https://doi.org/10.1073/pnas.58.6.2268 (1967).
28. McIntosh, K. *et al.* Seroepidemiologic studies of coronavirus infection in adults and children. *Am. J. Epidemiol.* **91**, 585–592. https://doi.org/10.1093/oxfordjournals.aje.a121171 (1970).
29. Kahn, J. S. & McIntosh, K. History and recent advances in coronavirus discovery. *Pediatr Infect Dis J* **24**, S223-227. https://doi.org/10.1097/01.inf.0000188166.17324.60 (2005) (**discussion S226**).
30. Chen, R. & Vasilakis, N. Dengue–quo tu et quo vadis?. *Viruses* **3**, 1562–1608. https://doi.org/10.3390/v3091562 (2011).
31. 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. Accessed 31 Dec 2020.
32. CDC. SARS-CoV-2 Variant Classifications and Definitions. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html.

33. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879. https://doi.org/10.1093/molbev/msu121 (2014).
34. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593-598. https://doi.org/10.1093/nar/gkr859 (2012).
35. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131–146. https://doi.org/10.1007/978-1-62703-646-7_8 (2014).
36. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729. https://doi.org/10.1093/molbev/mst197 (2013).
37. https://www.gisaid.org/. Accessed 9 Jan 2021.
38. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348 (2009).
39. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1**, vev003. https://doi.org/10.1093/ve/vev003 (2015).
40. Martin, D. & Rybicki, E. RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563. https://doi.org/10.1093/bioinformatics/16.6.562 (2000).
41. Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* **21**, 98–102. https://doi.org/10.1089/aid.2005.21.98 (2005).
42. Smith, J. M. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129. https://doi.org/10.1007/bf00182389 (1992).
43. Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA* **98**, 13757–13762. https://doi.org/10.1073/pnas.241370698 (2001).
44. Lam, H. M., Ratmann, O. & Boni, M. F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol. Biol. Evol.* **35**, 247–251. https://doi.org/10.1093/molbev/msx263 (2018).
45. Padidam, M., Sawyer, S. & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225. https://doi.org/10.1006/viro.1999.0056 (1999).
46. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582. https://doi.org/10.1093/bioinformatics/16.7.573 (2000).
47. Weiller, G. F. Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.* **15**, 326–335. https://doi.org/10.1093/oxfordjournals.molbev.a025929 (1998).
48. Lemey, P., Lott, M., Martin, D. P. & Moulton, V. Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. *BMC Bioinformatics* **10**, 126. https://doi.org/10.1186/1471-2105-10-126 (2009).
49. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772. https://doi.org/10.1038/nmeth.2109 (2012).
50. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704. https://doi.org/10.1080/10635150390235520 (2003).
51. Rambaut, A., Lam, T. T., MaxCarvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007. https://doi.org/10.1093/ve/vew007 (2016).
52. https://apps.who.int/iris/bitstream/handle/10665/326126/WHO-MERS-RA-19.1-eng.pdf. Accessed 31 Dec 2020.
53. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016. https://doi.org/10.1093/ve/vey016 (2018).
54. Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol* **30**, 239–243. https://doi.org/10.1093/molbev/mss243 (2013).
55. Fernandes, J. D. *et al.* The UCSC SARS-CoV-2 genome browser. *Nat. Genet.* **52**, 991–998. https://doi.org/10.1038/s41588-020-0700-8 (2020).
56. https://genome.ucsc.edu/covid19.html.
57. Tavares, R. C. A., Mahadeshwar, G., Wan, H., Huston, N. C. & Pyle, A. M. The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. *J. Virol.* https://doi.org/10.1128/JVI.02190-20 (2020).
58. Sun, L. *et al.* In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell* **184**, 1865-1883 e1820. https://doi.org/10.1016/j.cell.2021.02.008 (2021).

Acknowledgements

The authors would like to acknowledge all the authors and originating and submitting laboratories of the sequences from GISAID's EpiCov that were used in our analyses. A full acknowledgments list is submitted as Supplementary Table S5. This study was funded by the Global Emerging Infections Surveillance (GEIS) Branch (ProMIS ID: P0140_20_WR) and the US Department of Defense Health Agency. Support for this work was provided by the Infectious Disease Clinical Research Program (IDCRP), a Department of Defense program executed through the Uniformed Services University of the Health Sciences, Department of Preventive Medicine and Biostatistics through a cooperative agreement with The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. (HJF).

Author contributions

S.P., M.A.C., M.S. and I.M.B. designed and performed the analyses. S.P. and I.M.B. created the Figures. S.P., M.A.C., M.S., R.G.J., G.M.L., K.M. and I.M.B. wrote and edited the manuscript, I.M.B., R.G.J., G.M.L. and K.M. obtained the funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-96626-8.

Correspondence and requests for materials should be addressed to I.M.B.

Reprints and permissions information is available at www.nature.com/reprints.