Recombination and convergent evolution led to the emergence of 2019 Wuhan coronavirus

Juan Ángel Patiño-Galindo[1,2+], Ioan Filip[1,2+], Mohammed AlQuraishi[3, 4], Raul Rabadan[1,2*]

[1]Program for Mathematical Genomics,
[2]Departments of Systems Biology and Biomedical Informatics,
Columbia University, New York, NY, USA
[3]Department of Systems Biology
[4]Laboratory of Systems Pharmacology
Harvard University, Boston, MA, USA

[+]These two authors contributed equally to this work.
*Correspondence: rr2579@cumc.columbia.edu.

Abstract

The recent outbreak of a new coronavirus (2019-nCoV) in Wuhan, China, underscores the need for understanding the evolutionary processes that drive the emergence and adaptation of zoonotic viruses in humans. Here, we show that recombination in betacoronaviruses, including human-infecting viruses like SARS and MERS, frequently encompasses the Receptor Binding Domain (RBD) in the Spike gene. We find that this common process likely led to a recombination event at least 11 years ago in an ancestor of the 2019-nCoV involving the RBD. Compared with bat isolates, the recent ancestors of 2019-nCoV accumulated a high number of amino acid substitutions in the RBD and likewise in a region of polyprotein Orf1a that is critical for viral replication and transcription. Among these recent mutations, we identify amino acid substitutions common to the SARS 2003 outbreak isolates in positions 427N and 436Y, indicating potential adaptive convergent evolution. Both 427N and 436Y belong to a helix that appears to interact with the human ACE2 receptor. In sum, we propose a two-hit scenario in the emergence of the 2019-nCoV virus whereby the 2019-nCoV ancestors in bats first acquired genetic characteristics of SARS by incorporation of a SARS-like RBD through recombination before 2009, and subsequently, those recombinants underwent convergent evolution.

Introduction

In less than two months since initially reported in mid-December 2019, the recent coronavirus outbreak in Wuhan has caused more than 900 fatalities associated with severe respiratory disease. The causative agent was identified as a previously unknown RNA coronavirus virus, dubbed 2019-nCoV (for 2019 novel Coronavirus), of the betacoronavirus genus[1], with 80% similarity at nucleotide level to SARS coronavirus[2]. SARS and 2019-nCoV are the only members of *Sarbecovirus* subgenus of betacoronavirus that are known to infect humans. Other members of this subgenus are frequently found in bats, hypothesized to be the natural reservoir of many zoonotic coronaviruses[3]. In January 2020, a *Rhinolophus affinis* bat isolate obtained in 2013 from the Yunnan Province in China (named RaTG13) was reported to have 96% similarity to the 2019-nCoV[4], suggesting that the ancestors of the outbreak virus were recently circulating in bats. However, the specific molecular determinants that enable a virus like the recent ancestor of 2019-nCoV to jump species remain poorly characterized.

The capability of viral populations to emerge in new hosts can be explained by factors such as rapid mutation rates and recombination[5] which lead to both high genetic variability and high evolutionary rates (estimated to be between 10^{-4} and 10^{-3} substitutions per site per year) [6]. Previous genome-wide analyses in coronaviruses (CoV) have estimated that their evolutionary rates are of the same order of magnitude as in other fast-evolving RNA viruses[7,8]. Interestingly, evolution within the CoV Spike gene, whose encoded protein interacts with the host cell receptor and is a key determinant in host tropism, appears to occur faster than in most of the other CoV genes[8]. Recombination in RNA viruses, known to be frequent in coronaviruses, can lead to the acquisition of genetic material from other viral strains[9]. Indeed, recombination has been proposed to play a major role in the generation of new coronavirus lineages such as SARS-CoV[9]. Interestingly, a recent study suggests that 2019-nCoV was involved in a potential recombination event between different members of the *Sarbecovirus* subgenus[1].

In this work, we investigate the evolutionary events that have led to the emergence of the 2019-nCoV virus. In particular, we identify amino acid changes differentiating the 2019-nCoV genome from the next most closely related betacoronavirus strain, RaTG13, and furthermore, we find evidence of convergent evolution with human SARS-CoV. We also perform recombination analyses testing whether such distinctive patterns could be explained by point mutations and/or recombination. In short, we establish how recombination and subsequent convergent evolution in betacoronavirus likely led to the emergence of the 2019-nCoV strain.

Results

Recombination hotspots in betacoronavirus

To understand how recombination contributes to the evolution of betacoronaviruses across different viral subgenera and hosts, we analyzed 45 betacoronavirus sequences from the five major subgenera infecting mammals (*Embevovirus, Merbecovirus, Nobecovirus, Hibecovirus* and *Sarbecovirus*)(Supplementary Table 1)[10]. Using the RPDv4 package[11] to identify recombination breakpoints, we identified 103 recombination events (Figure 1a, Methods). Enrichment analysis indicates that recombination often involves the n-terminus of the Spike protein that includes the Receptor Binding Domain (RBP) (adjusted *p*-val. < 10^{-4}, binomial test on sliding window of 800 nucleotides) (Figure 1b, Supplementary Figure 1). This enrichment of recombination events persisted after restricting the analysis to the most common host (bats), suggesting that the recombination is not driven by sampling of multiple human sequences (Supplementary Figure 2). In all, we find that recombination in betacoronavirus frequently involves the Spike protein across viral subgenenera and hosts.

MERS recombination frequently involves the Spike gene

To study how recombination affects emerging human betacoronaviruses viruses at the viral species level, we focused our attention to MERS, due to the more extensive sampling both in humans and camels, the source of the recent zoonosis[12]. Using 381 MERS sequences (170 from human, 209 from camel and 2 from bat) (Supplementary Table 2) we show that the Spike region overlaps with the majority of recombination segments (83%, 20 of 24 identified events) (Figure 2a) with an enrichment of recombination breakpoints detected in the Spike and Membrane genes (Figure 2b, Supplementary Figure 3). This effect was not observed when restricting the analysis to human MERS samples only (n=170) possibly due to the lower number and diversity of sequences available (Supplementary Figure 4). We thus show that the enrichment of recombination events involving the Spike gene is also observed at a viral species level.

Recombination event in an ancestor of 2019-nCoV encompassing the RBD of the Spike gene

We then asked if this enrichment of recombination could be found in the recent history of 2019-nCoV. We first perform sliding phylogenetics showing topological incongruences between phylogenies involving three sections of the Spike gene: the 5', the RBD and the 3' (Figure 3a), supporting potential recombination within the Spike gene involving the RBD domain. In addition, recombination analysis performed with the RDP4 package detected a significant recombination event (genome positions 22614-23032) affecting 2019-nCoV and the human SARS in the RBD (*p*-val. < 0.003, RDP, Bootscan, Maxchi and Chimaera), recapitulating the result of Wu et al.[1] Interestingly, the clade of the full genome phylogeny

that includes RaTG13 and Wuhan 2019-nCoV shared 47 amino acids with human SARS in the RBD that were distinct from bat SARS-like CoV sequences (Supplementary Table 3). These results do not only demonstrate that 2019-nCoV displays a genotype similar to human SARS in the RBD, but also that it seems to have originated through recombination. To trace back the potential time of this recombination event, we used a Bayesian Phylogenetic approach[13] in the recombinant region (codons 200-500 in the Spike gene) and compared with the whole genome phylogeny (Figure 3b). As in the full genome phylogeny, Wuhan 2019-nCoV and RaTG13 were in the same clade, with human SARS-CoV as an outgroup. From our, results we propose two different scenarios. The first is that Wuhan 2019-nCoV derives from a recombination event between human SARS-CoV and another (unsampled) SARS-like CoV. The second is that there occurred at least two different recombination events, one leading to human SARS-CoV and another one leading to Wuhan 2019-nCoV. In either of these two scenarios, we inferred the time to the most recent common ancestor in the recombinant region of the clade leading to RaTG13 and Wuhan 2019-nCoV to have occurred no later than 2009 (2003-2013, 95% HPD limit).

Significant accumulation of amino acid substitutions in the RBD domain since the recombination event

We then investigated the mutational events that occurred in the Wuhan 2019-nCoV since its divergence from its most recent common ancestor with RaTG13 around 2009 (Figures 3b, 4a). A comparison of the distribution of nonsynonymous versus 4-fold degenerate site changes across the viral genome highlighted two regions with significant enrichment of nonsynonymous changes leading to amino acid substitutions (adjusted p-val. < 10^{-5} and p-val. < 10^{-3} for the first and second regions respectively, binomial test on sliding windows of 267 amino acids) (Figure 4b). The first region, with windows starting between positions 801 and 1067 in the Orf1a gene in our analysis, spans the non-structural proteins (nsp) 2 and 3 that were previously reported to accrue a high number of mutations between bat and SARS coronaviruses[14] and includes the ubiquitin-like domain 1, a glutamic acid-rich hypervariable region, and the SARS-unique domain of nsp3[15] that is critical to the replication and transcription of SARS-like CoV[16,17]. The second region that we found with high divergence from the RaTG13 bat virus contained 27 substitutions in the Spike protein, of which 20 were located in the RBD (Supplementary Table 4). There was no enrichment in mutations observed at 4-fold degenerate sites, suggesting that those regions do not correspond to any further recent recombination events (Supplementary Figures 5-8).

We found only two amino acid substitutions associated with the recent evolution of the 2019-nCoV virus that were also present in all human-infecting SARS-CoV but absent from the recent bat isolate from Yunnan (RaTG13) and likewise absent from any other bat isolate, except for isolates such as Rs7327,

Rs4874 and Rs4231 that are known to co-opt the human ACE2 receptor [18]. These two sites, 436Y and 427N, are located in the RBD of the Spike protein, suggesting potential adaptive convergent evolution for *Sarbecovirus* to infect humans. More specifically, the two sites belong to the short helix (427-436) of Spike (Figure 4c, Supplementary Figure 9) which lies at the interface of the human ACE2 receptor with the Spike protein. Furthermore, site 436Y appears to form a hydrogen bond with 38D in ACE2 (Figure 4c), likely contributing to the stability of the complex, which is disrupted by the mutation Y436F[19] (that is present in RaTG13). The second mutation that we identified, K427N, may disrupt the short helix and cause the loop to shift, further affecting stability (Supplementary Figure 9). These two positions, 436Y and 427N, which are present in all isolates from the 2009-nCoV viruses, are also found in viruses from other hosts, including civets (*Paguma larvata*) (Supplementary Table 5). Interestingly, a mouse-adapted SARS virus showed a mutation at position 436 (Y436H) that enhanced the replication and pathogenesis in mice[20,21], indicating that this change may have an effect in host tropism. We did not find any other amino acids shared by all human SARS-CoV and 2019-nCoV that were absent from bat SARS-like CoV samples in the rest of the genome. To summarize, we show here that after the recombination event in an ancestor of 2019-nCoV, amino acid substitutions accumulated in the RBD of the Spike protein, including changes that were found in human SARS-CoV.

Discussion

In this work, we have analyzed the recent evolution of Wuhan 2019-nCoV. We present a two-hit scenario involving two major mechanisms of genetic evolution in coronavirus, recombination followed by point mutations, leading to convergent evolution to human SARS virus. We show that the recombination events preferentially affect the RBD region in the Spike gene, both at the order of genus (betacoronavirus) and related species (MERS). We show that a recombination in this region occurred before 2009 involving a recent ancestor of 2019-nCoV. The recombinant ancestor would have acquired more than 40 amino acids characteristic of human SARS. Subsequent substitutions in the RBD domain led to the acquisition of human SARS amino acids that may have contributed to adaptation to humans. It has been hypothesized that recombination[7,22] and rapid evolution was observed between bat, civet and human SARS-CoVs[14], including in the nsp3 and Spike gene regions identified here in the Wuhan strain. The two-step process that we are proposing of the evolutionary adaptation of coronavirus to infect human hosts could have also occurred in the emergence of the 2003 SARS outbreak.

Author Contributions

J.P., I.F. and R.R. designed the study and prepared the manuscript. J.P. and I.F. performed computational analysis. M.A. helped with protein structure analyses.

Acknowledgements

We would like to thank Karen Gomez and Andrew Chen for their help editing the manuscript, and Zixuan Wang for her help on the figures. This work has been funded by NIH grants R01 GM117591, U54-CA225088 and DARPA/DOD grant W911NF-14-1-0397.

Disclosure of Potential Conflicts of Interest

R.R. is a member of the SAB of AimedBio in a project unrelated to the current manuscript.

References

1 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*, doi:10.1038/s41586-020-2008-3 (2020).

2 Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* **348**, 1967-1976, doi:10.1056/NEJMoa030747 (2003).

3 Banerjee, A., Kulcsar, K., Misra, V., Frieman, M. & Mossman, K. Bats and Coronaviruses. *Viruses* **11**, doi:10.3390/v11010041 (2019).

4 Peng Zhou, X.-L. Y., Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Gengfu Xiao, Zheng-Li Shi. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. *bioRxiv 2020.01.22.914952* **doi:** https://doi.org/10.1101/2020.01.22.914952 (2020).

5 Holmes, E. C. The phylogeography of human viruses. *Mol Ecol* **13**, 745-756, doi:10.1046/j.1365-294x.2003.02051.x (2004).

6 Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* **54**, 156-165, doi:10.1007/s00239-001-0064-3 (2002).

7 Hon, C. C. *et al.* Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J Virol* **82**, 1819-1826, doi:10.1128/JVI.01926-07 (2008).

8 Xiong, C., Jiang, L., Chen, Y. & Jiang, Q. Evolution and variation of 2019-novel coronavirus. *bioRxiv* (2020).

9 Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J Virol* **84**, 3134-3146, doi:10.1128/JVI.01394-09 (2010).

10 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, doi:10.1016/S0140-6736(20)30251-8 (2020).

11 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, doi:10.1093/ve/vev003 (2015).

12 de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* **14**, 523-534, doi:10.1038/nrmicro.2016.81 (2016).

13 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016, doi:10.1093/ve/vey016 (2018).

14 Chinese, S. M. E. C. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666-1669, doi:10.1126/science.1092002 (2004).

15 Neuman, B. W. Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. *Antiviral Res* **135**, 97-107, doi:10.1016/j.antiviral.2016.10.005 (2016).

16 Kusov, Y., Tan, J., Alvarez, E., Enjuanes, L. & Hilgenfeld, R. A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex. *Virology* **484**, 313-322, doi:10.1016/j.virol.2015.06.016 (2015).

17 Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res* **149**, 58-74, doi:10.1016/j.antiviral.2017.11.001 (2018).

18 Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* **13**, e1006698, doi:10.1371/journal.ppat.1006698 (2017).

19 Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864-1868, doi:10.1126/science.1116480 (2005).

20 Roberts, A. *et al.* A mouse-adapted SARS-coronavirus causes disease and mortality in BALB/c mice. *PLoS Pathog* **3**, e5, doi:10.1371/journal.ppat.0030005 (2007).

21 Becker, M. M. *et al.* Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. *Proc Natl Acad Sci U S A* **105**, 19944-19949, doi:10.1073/pnas.0808116105 (2008).

22 Holmes, E. C. & Rambaut, A. Viral evolution and the emergence of SARS coronavirus. *Philos Trans R Soc Lond B Biol Sci* **359**, 1059-1065, doi:10.1098/rstb.2004.1478 (2004).

23 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

24 Korber, B. & Myers, G. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res Hum Retroviruses* **8**, 1549-1560, doi:10.1089/aid.1992.8.1549 (1992).

25 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).

26 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

27 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007, doi:10.1093/ve/vew007 (2016).

28 Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188, doi:10.1214/aos/1013699998 (2001).

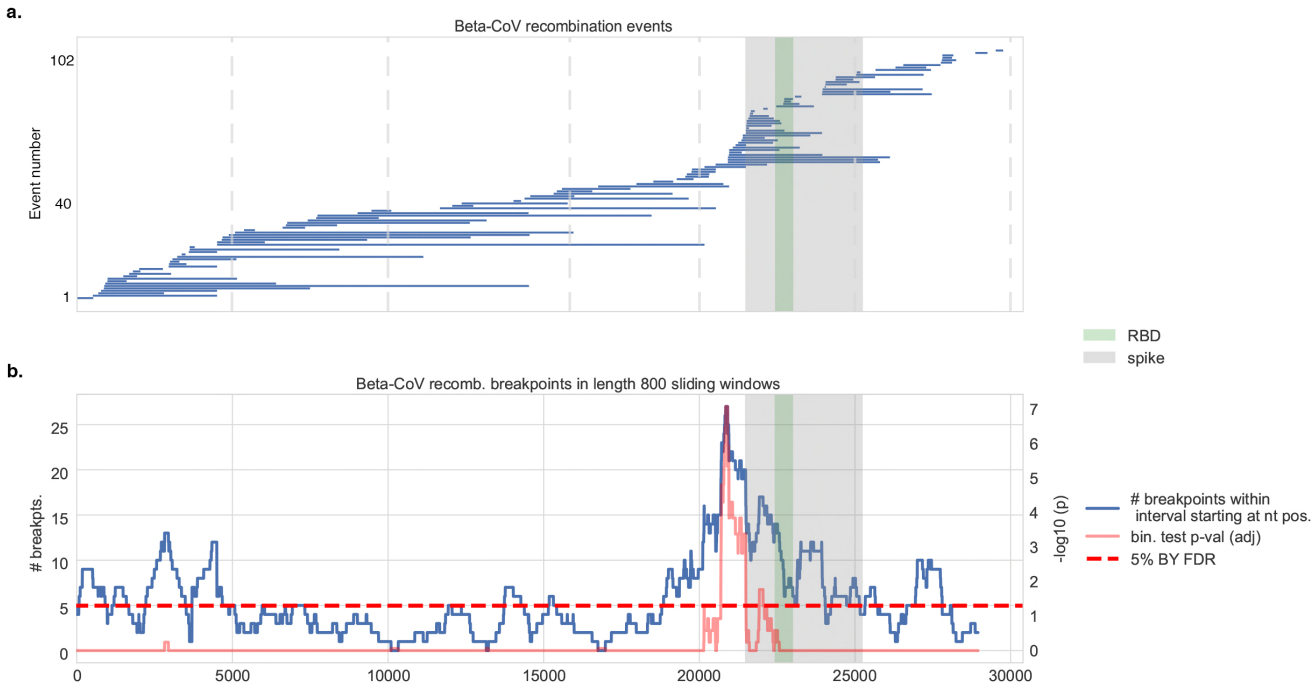

Fig. 1 | Recombination analysis of beta coronaviruses. a. Distribution of 103 inferred recombination events among human and non-human Beta-CoV isolates showing the span of each recombinant region along the viral genome with respect to SARS-CoV coordinates. The spike protein and its RBD are highlighted. **b.** Sliding window analysis shows (blue curve) the distribution of recombination breakpoints (either start or end) in 800 nucleotide (nt) length windows upstream (namely, in the 5' to 3' direction) of every nt position along the viral genome. The spike protein, and in particular the RBD and its immediate downstream region, are significantly enriched in recombination breakpoints in betacoronaviruses. Benjamini-Yekutieli (BY) corrected p-values are shown (red curve), and the 5% BY FDR is shown for reference (dotted line).

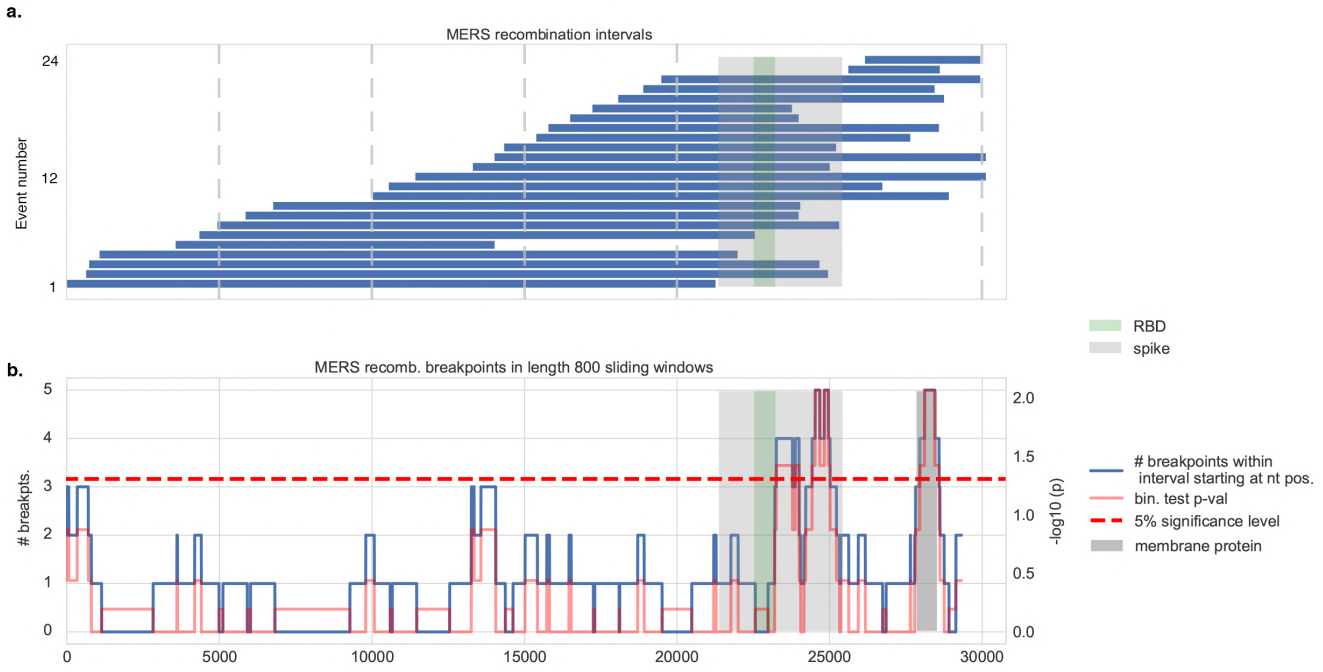

Fig. 2 | Recombination analysis in MERS coronaviruses. **a.** Distribution of 24 recombination events among human and non-human MERS-CoV isolates. The spike protein and its RBD are highlighted. **b.** Sliding window analysis shows (blue curve) the distribution of recombination breakpoints (either start or end) in 800 nucleotide (nt) length windows upstream (namely, in the 5' to 3' direction) of every nt position along the viral genome. The spike protein, and the RBD in particular, overlap with widows that are enriched in recombination breakpoints. Binomial test p-values (red curve) and the 5% significance level are shown (dotted line). The SARS membrane protein is highlighted (dark gray); it also shows an enrichment of recombination breakpoints.

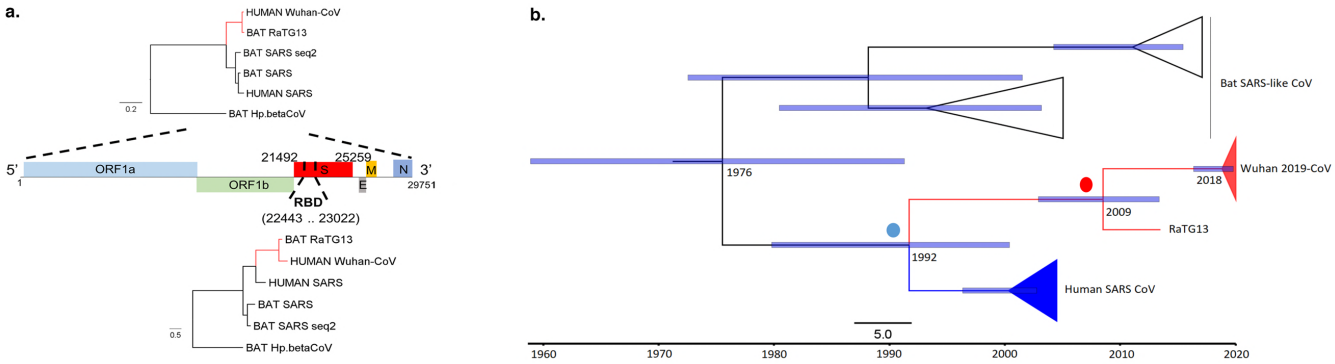

Fig. 3 | Recombination event in an ancestor of 2019-nCoV encompasses the RBD of the Spike gene. a. Full genome phylogenetic tree inference (top) shows 2019-nCoV sequence (HUMAN Wuhan-CoV) and its nearest bat-infecting RaTG13 strain clustering in a clade separate from the SARS isolates from human and bats (BAT and HUMAN SARS lineages) with respect to an outlying and more distant BAT sequence (Hp.betaCoV). Phylogenetic reconstruction (only 3rd codon positions) in the region of spike containing the RBD, on the other hand (bottom), shows the 2019-nCoV lineage clustering together with the HUMAN SARS strain. This change in tree topology gives evidence of a likely recombination in Beta-CoV encompassing the RBD which leads to the appearance of the Wuhan strain. Accession numbers: BAT Hp.betaCoV (KF636752); HUMAN SARS (FJ882963); BAT SARS (DQ071615); BAT SARS seq2 (DQ412043); RaTG13 (MN996532); HUMAN Wuhan-CoV (isolate 403962). **b**. Dated phylogeny of the RBD including RaTG13, 3 sequences from 2019-nCoV (red), 6 Bat SARS-like CoV (black) and 9 Human SARS CoV sequences (blue). The inference suggests possible scenarios with one recombination at least 11 years ago (highlighted as red dot), but possibly as long as 28 years ago (blue dot).

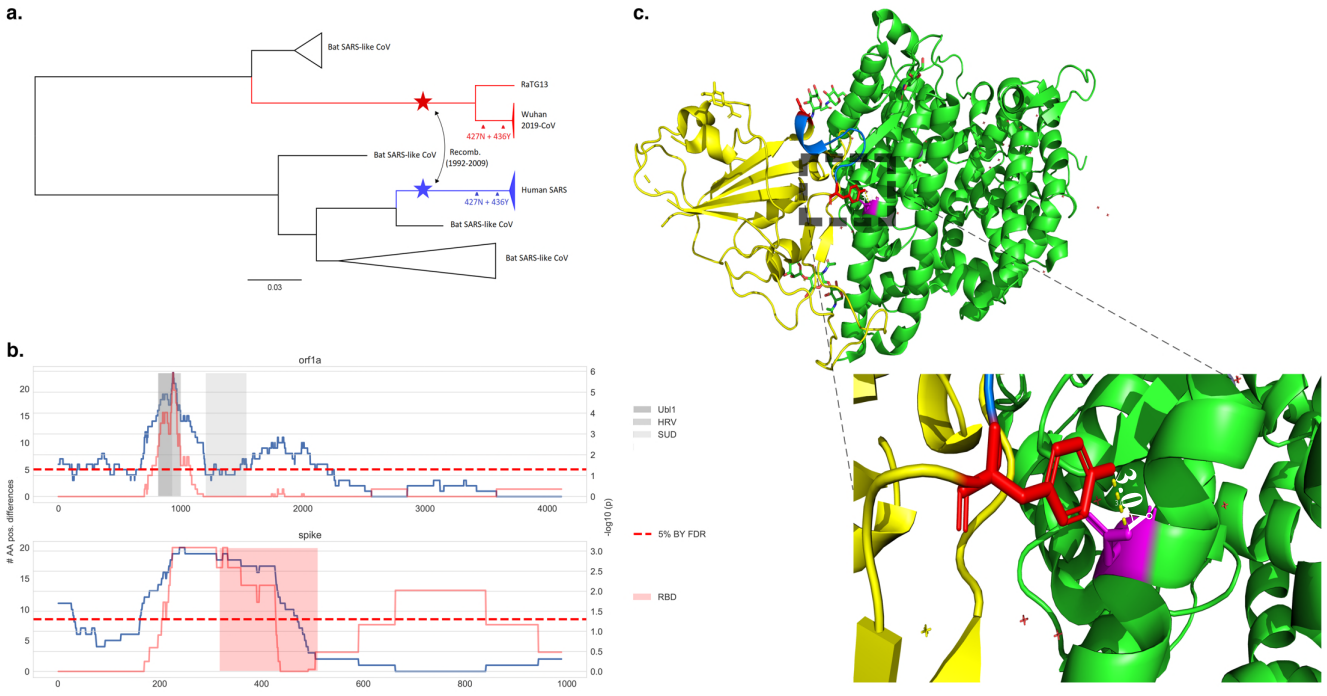

Fig. 4 | Higher rate of amino acid substitutions from the recent recombination event. a. Dated phylogenetic representation summarizing the successive evolutionary events that likely led to the emergence Wuhan 2019-nCoV strain: 1) Recombination of the RBD of the spike protein between the lineage ancestral to both 2019-nCoV and RaTG13 (red star) and the ancestral lineage of the human SARS (blue star); 2) More recent conserved amino acid (aa) substitutions at positions 427N and 436Y in spike that appeared since the recombination event in both the Wuhan 2019-nCoV lineage (red arrows) and in the Human SARS (blue arrows), strongly suggesting convergent evolution. **b.** Sliding window analysis (length 267 aa) identifies specific regions with high divergence from the RaTG13 bat virus in the RBD of spike (including 427N and 436Y), as well as in the Ubl1, HRV and SUD domains of nsp3 (non-structural protein 3) within the orf1a polyprotein. **c.** Interaction between the human ACE2 receptor (green) and the spike protein (yellow) based on SARS coronavirus (PDB accession code: 2AJF). Substitutions in 2019-nCoV at positions 427N and 436Y belong to a helix (blue) situated at the interface of the interaction with ACE2. **Detail:** site 436Y (red) forms a hydrogen bond (dashed yellow line) with 38D (purple) in ACE2, likely contributing to the stability of the complex.

Online Methods:

Sample collection

Wuhan 2019-nCoV and SARS/SARS-like-CoV

A set of 71 genome sequences derived from Wuhan 2019-nCoV (which represent all genome availability at GISAID on February 7, 2020; gisaid.org) was analyzed together with its closest animal-infecting relative, RaTG13 (accession number MN996532), and other genome sequences from human SARS (n=72) and bat SARS-like CoV (n=19), publicly available in Genbank (ncbi.nlm.nih.gov/genbank/) (Supplementary Table 6). Alignment was performed either at genome wide nucleotide level or at each CDS independently (Orf1a, Orf1b, S, E, M, N; at amino acid level) with MAFFTv7 ("auto" strategy)[23]. We used the program VESPA [24] to find the amino acid signatures that define Wuhan 2019-nCoV with respect RaTG13 and other bat SARS-like CoV. Concretely, two types of comparisons were performed: i) Wuhan 2019-nCoV (query set) vs RaTG13 (background set) to detect amino acid changes that define Wuhan 2019-nCoV, compared to RaTG13. ii) Wuhan 2019-nCoV + Human SARS CoV (query) vs bat SARS-like CoV (including RATG13) to detect amino acid changes that define both human infections, compared to bat SARS-like CoV virus. Only those mutations totally fixed in the query set and absent in the background alignment were further considered.

For each CDS we also compared the distribution of nonsynonymous substitutions with that of synonymous changes (considering 4-fold degenerate sites, detected with MEGA software 7 [25]) occurring between the earliest sampled Wuhan 2019 nCoV sequence (EPI_ISL_404227) and RaTG13.

Recombination analysis

Seven recombination detection methods implemented in the RDP4 software package (RDP, Geneconv, Bootscan, Maxchi, Chimaera, SiScan, 3seq)[11] were used to detect evidence of recombination with default parameters (*p*-value = 0.05, Bonferroni corrected), and depict the distribution of recombination events, in different CoV alignments:

1- The following selection of viral strains was used in order to find breakpoints involving 2019-nCoV: KF636752 (bat), FJ882963 (human SARS), DQ071615 (bat SARS-like CoV), DQ412043 (bat SARS-like CoV), RaTG13 and isolate 403962 from Wuhan 2019-nCoV.

2- A MERS-CoV genome alignment (n= 381; n= 170 human, n= 209 camel, and 2 bat sequences).

3- A betacoronavirus alignment (n=45 sequences, covering the genus diversity as in *Lu, R. et al., 2020*[10]).

Phylogenetic analysis

The evolutionary relationships between 2019-nCoV and other SARS/SARS-like viruses was inferred from genome alignment using PhyML (GTR + GAMMA 4CAT)[26]. The same program and model were used to reconstruct the phylogenetic tree of the (potentially recombinant) RBD, using only 3rd codon positions. Dated phylogeny of the RBD was obtained with BEAST v1.8.4 (Supplementary Table 7), after assessing the molecular clock signal of the selected sequences with TempEst [27]. The analysis was performed with the GTR+ GAMMA (4 cat) substitution model combined with an uncorrelated lognormal relaxed clock model and the Bayesian Skyline Plot demographic model. We used as prior distribution for the time to most recent common ancestor (tMRCA) a normal distribution with mean 40 years (standard deviation 10 years), as previously inferred[7]. Two independent runs of BEAST were performed, with MCMC chain lengths of $5x10^7$ states. Convergence of the estimated parameters was confirmed with Tracer http://tree.bio.ed.ac.uk/software/tracer/.

Statistical analysis

Sliding window analysis was performed in order to test for enrichment of recombination breakpoints (including both start and end breakpoints) along the viral genome in the following settings: 1) all Beta-CoV recombinations; 2) recombinations within non-human lineages for Beta-CoV; 2) all MERS recombinations; and 3) both human-specific and non-human MERS lineage recombinations separately. There were too few human-specific recombinations in Beta-CoV for in-depth analysis. For Beta-CoV analyses, the SARS CoV genomic coordinates were used as reference (accession NC_004718), whereas for MERS CoV, we used a MERS CoV sequence (accession NC_019843) as reference. Windows of 800 nucleotides were selected and binomial tests for the number of breakpoints in each window were performed under the null hypothesis that recombination breakpoints are distributed uniformly along the genome. Given the co-dependence structure of our statistical tests, adjustments were performed using the Benjamini-Yekutieli (BY) procedure[28] which provides a conservative multiple hypothesis correction that applies in arbitrary dependence conditions. For statistical significance, we chose 5% BY FDR. Our discoveries hold with different choices for window length, provided the window size is sensitive to the scale of interest (namely CoV proteins and the length of specific domains such as the RBD in the Spike).

We used the same sliding window approach to test for enrichment of gene-specific amino acid differences between the 2019-nCoV sequence and the bat virus RaTG13. For consistency, we selected 267 length windows of amino acids (corresponding to approximately 800 nucleotides) and performed *p*-value correction using the same procedure.