Recombination and lineage-specific mutations led to the emergence of SARS-CoV-2

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 (SARS-CoV-2) 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-CoV and MERS-CoV, 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 SARS-CoV-2 involving the RBD. As a result of this recombination event, SARS-CoV and SARS-CoV-2 share a similar genotype in RBD, including two insertions (positions 432-436 and 460-472), and alleles 427N and 436Y. Both 427N and 436Y belong to a helix that interacts with the human ACE2 receptor. Ancestral state analyses revealed that SARS-CoV-2 differentiated from its most recent common ancestor with RaTG13 by accumulating a significant number of amino acid changes in the RBD. In sum, we propose a two-hit scenario in the emergence of the SARS-CoV-2 virus whereby the SARS-CoV-2 ancestors in bats first acquired genetic characteristics of SARS-CoV by incorporation of a SARS-like RBD through recombination before 2009, and subsequently, the lineage that led to SARS-CoV-2 accumulated further, unique changes specifically in the RBD.

Introduction

In the three months since the initial reports in mid-December 2019, the recent COVID-19 pandemic has caused close to 10,000 fatalities associated with severe respiratory disease worldwide[1]. The causative agent of COVID-19 was identified as a previously unknown RNA coronavirus (CoV) virus, dubbed SARS-CoV-2, of the betacoronavirus genus[2], with 80% similarity at nucleotide level to the Severe Acute Respiratory Syndrome (SARS) coronavirus[3]. SARS-CoV and SARS-CoV-2 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[4]. 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 SARS-CoV-2[5], suggesting that the ancestors of the outbreak virus were recently circulating in bats. However, the specific molecular and evolutionary determinants that enable a virus like the recent ancestor of SARS-CoV-2 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[6] 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)[7]. Previous genome-wide analyses in coronaviruses have estimated that their evolutionary rates are of the same order of magnitude as in other fast-evolving RNA viruses[8,9]. Recombination in RNA viruses, known to be frequent in coronaviruses, can lead to the acquisition of genetic material from other viral strains[10]. Indeed, recombination has been proposed to play a major role in the generation of new coronavirus lineages such as SARS-CoV[10]. Furthermore, a recent study suggests that SARS-CoV-2 was involved in a potential recombination event between different members of the *Sarbecovirus* subgenus[2].

In this work, we investigate the evolutionary events that characterize the emergence of SARS-CoV-2. In particular, we identify one recombination event in a region that determines host tropism, the Receptor Binding Domain (RBD) of the Spike gene, which led human SARS and SARS-CoV-2 to share a similar haplotype in the RBD. Through ancestral reconstruction analyses, we also observe that, subsequent to this recombination, a significant enrichment of nonsynymous changes occurred in SARS-CoV-2 after the split from its most recent common ancestor (MRCA) with RaTG13.

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)[11]. Using the RPDv4 package[12] 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-CoV 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 the Middle East respiratory syndrome coronavirus (MERS-CoV), due to the extensive sampling both in humans and in camels (recognized as the source of the recent zoonosis[13]). Using 381 MERS-CoV 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-CoV 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.

Identification of a recombination event involving SARS-CoV-2 in RBD of Spike gene

We then asked if any signal of recombination could be found in the recent history of SARS-CoV-2. 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 SARS-CoV-2 and the human SARS-CoV in the RBD (*p*-val. < 0.003, RDP, Bootscan, Maxchi and Chimaera), recapitulating the result of Wu et al.[2]

We then analyzed the consensus amino acid sequences within the RBD region involved in this recombination event, comparing the clade affected by the recombination to the clade consisting of the closest bat SARS-like CoVs (including KY770859, KJ473816, MG772933, MG772934 and KY417145 – Supplementary Figure 5). The two consensus sequences differed in 62 codon positions including two insertions spanning the 432-436 and 466-472 regions. Interestingly, out of these 62 changes, 23 amino acids are conserved in human SARS-CoV and SARS-CoV-2 (Supplementary Table 3), including four positions that are conserved in the human-infecting viral strains but that are not conserved in those infecting bats: 427N, 436Y, 455I and 466N. In summary, these results demonstrate, first of all, that SARS-CoV-2 displays a genotype similar to SARS-CoV in the RBD, and secondly, that these two lineages share a common history of recombination in the RBD.

To trace back the potential time of the recombination event involving ancestral lineages of SARS-CoV and SARS-CoV-2, we used a Bayesian Phylogenetic approach[14] 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, SARS-CoV-2 and RaTG13 were in the same clade, with human SARS-CoV as an outgroup. In light of these results, there two possible scenarios to consider. The first is that SARS-CoV-2 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 SARS-CoV-2. In either of these two scenarios, the inferred the time to the most recent common ancestor in the recombinant region of the clade leading to RaTG13 and SARS-CoV-2 is no later than 2009 (2003-2013, 95% HPD limit).

Conserved mutations in SARS-CoV and SARS-CoV-2 RBD

We highlight two mutations, 427N and 436Y, conserved in human SARS-CoV, SARS-CoV-2 and human-infecting strains, but that are not conserved in other CoVs that only infect bats. Both mutations belong to the short helix (427-436) of Spike (Figure 4c, Supplementary Figure 6) 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[15] (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 6). These two positions, 436Y and 427N, which are present in all SARS-CoV-2 isolates, are also found in viruses from other hosts, including civets (*Paguma larvata*) (Supplementary Table 4). Interestingly, a mouse-adapted SARS virus showed a mutation at position 436 (Y436H) that enhanced the replication and pathogenesis in mice[16,17], indicating that this change may have an effect in host tropism. It is noteworthy that unlike 436Y, 455I and 466N, the 427N mutation was present in the closest strains not

involved in the recombination event (KY770859 and KJ473816), suggesting that it appeared in their MRCA through point mutation. Our ancestral state reconstructions also suggest that 427N has appeared at two other independent times in bat SARS-like CoVs, and only at external branches (sequences JX993988 and JX993987; Supplementary Figure 5). Other bat isolates with the 427N allele, such as Rs7327, Rs4874 and Rs4231, are known to co-opt the human ACE2 receptor[18], further reinforcing the role of 427N as an adaptive mutation for the interaction with ACE2.

Recent substitutions in SARS-CoV-2 RBD

We compared the SARS-CoV-2 sequence to that of the genome-wide nearest bat CoV, namely RaTG13. The distributions of nonsynonymous and 4-fold degenerate site changes between SARS-CoV-2 and RaTG13 across the viral genome highlighted two regions with significant enrichment of nonsynonymous changes (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 CoVs[19] and includes the ubiquitin-like domain 1, a glutamic acid-rich hypervariable region, and the SARS-unique domain of nsp3[20] that is critical to replication and transcription[21,22]. 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 5). There was no significant enrichment in mutations observed at 4-fold degenerate sites (Supplementary Figures 7-10).

Trough ancestral state reconstruction, we then inferred the entire Spike sequence of the most recent common ancestor (MRCA) of RaTG13 and SARS-CoV-2. We assessed the changes that were specific to the lineage leading to SARS-CoV-2 and identified 32 substitutions (of which 2 only were in the RBD) at 4-fold degenerated sites, and 10 amino acid changes (of which 5 were in the RBD) which have likely occurred in the lineage leading to SARS-CoV-2 (Supplementary Table 6). Thus, we report a significant enrichment of nonsynonymous changes in the RBD (Fisher Exact Test: *p*-value =0.005; Odds Ratio = 13.6, 95% Confidence Interval = 1.7 – 180.0). The five amino acid changes in RBD were I428L, A430S, K465T and Y484Q. Interestingly, position 484 has already been reported to directly interact with ACE2[23].

Discussion

In this work, we have analyzed the evolution of SARS-CoV-2 and its closest relatives. We suggest a two-hit scenario involving both ancestral recombination and recent mutational events in the lineage that led to SARS-CoV-2. It has been hypothesized previously that recombination[8,24] and rapid evolution was observed between bat, civet and human SARS-CoVs[19]. 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-CoV). We show that a recombination in this region occurred before 2009 involving a common ancestor of SARS-CoV and SARS-CoV-2. In this recombination event, SARS-CoV-2 acquired more than 20 amino acids characteristic of SARS-CoV and its closest relatives. The similarities between both SARS-CoVs suggest that they share the same cell tropism[25,26].

Interestingly, positions 427N and 436Y are retained in human sarbecoviruses but are not conserved among non-human strains, suggesting that they might be relevant for human infection but not necessary for infecting other hosts. Indeed, position 436 in Spike is part of the RBD-ACE2 interface, while position 427 is proximal to others such as 426 which are key for the strength of the RBD-ACE2 interaction[23, 26].

The lineage that gave rise to SARS-CoV-2 underwent further differentiation from its MRCA with RaTG13, accumulating mutations in key positions of the RBD, such as 484Q, and the insertion of a four amino acids long polybasic cleavage site. The insertion of the polybasic cleavage site is unique to SARS-CoV-2 among other Sarbecoviruses, and occurred between Spike positions 666-667 (using SARS-CoV NC_004718 reference sequence coordinates). The presence of polybasic cleavage sites has been associated in other viruses such as Influenza with high pathogenicity[26, 27]. Indeed, experimental analyses that introduced such sites in the S1-S2 junction of human SARS-CoV have observed an increase of cell-cell fusion without affecting viral entry[28, 29].

In conclusion, our evolutionary analyses have revealed that the evolutionary processes leading to SARS-CoV-2 can be explained by a two-hit scenario of recombination and recent mutational events in the Spike.

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 thank GISAID and all the laboratories where the data used in this study was collected and processed (Supplementary Table 9). We would also 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 W.H.O. Coronavirus disease 2019 (COVID-19) Report No. 60, (2020).

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

3 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).

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

5 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).

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

7 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).

8 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).

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

10 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).

11 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).

12 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).

13 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).

14 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).

15 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).

16 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).

17 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).
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 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).
20 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).
21 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).
22 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).
23 Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science*, doi:10.1126/science.abb2762 (2020).
24 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).
25 Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, doi:10.1016/j.cell.2020.02.058 (2020).
26 Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, doi:10.1016/j.cell.2020.02.052 (2020).
27 Schrauwen, E. J. *et al.* The multibasic cleavage site in H5N1 virus is critical for systemic spread along the olfactory and hematogenous routes in ferrets. *J Virol* **86**, 3975-3984, doi:10.1128/JVI.06828-11 (2012).
28 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine*, doi:10.1038/s41591-020-0820-9 (2020).
29 Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358-369, doi:10.1016/j.virol.2006.02.003 (2006).
30 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).
31 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).
32 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).
33 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528, doi:10.1093/bioinformatics/bty633 (2018).
34 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).

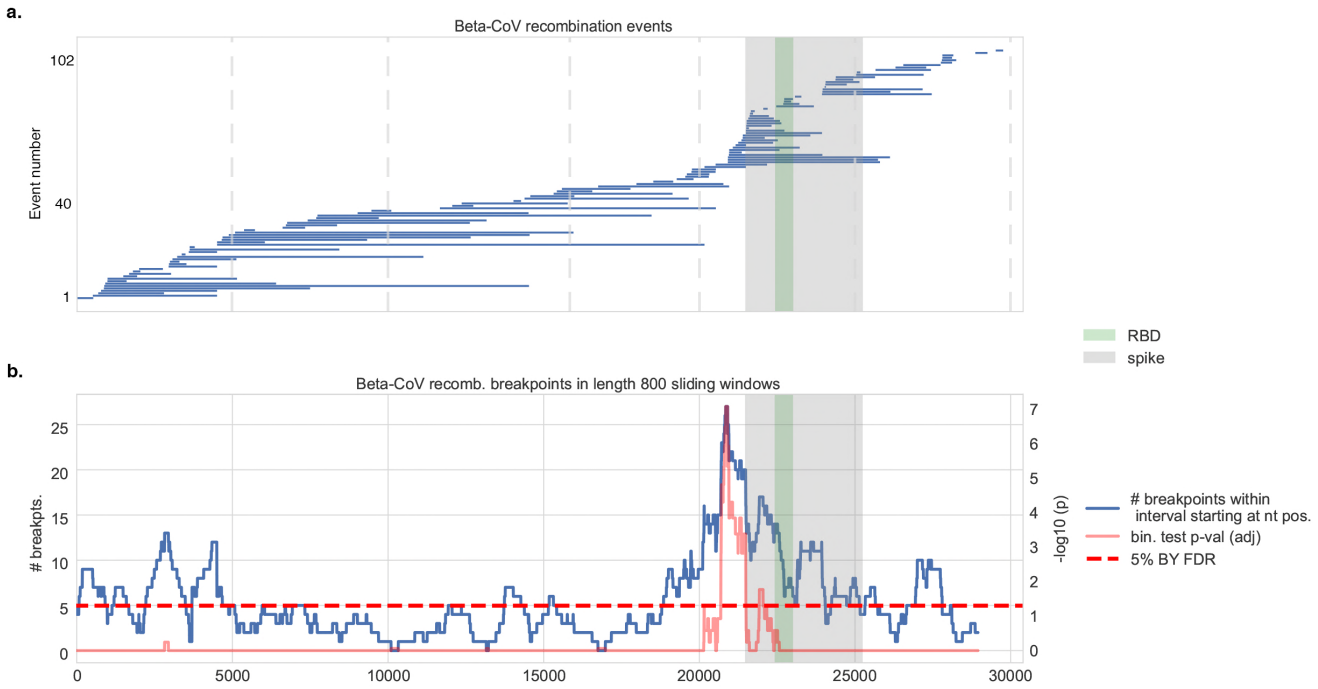

Fig. 1 | Recombination analysis of betacoronaviruses. a. Distribution of 103 inferred recombination events among human and non-human beta-CoV isolates showing the span of each recombinant region along the viral genome with respect to SARS-CoV coordinates. The spike protein and its RBD are highlighted. **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).

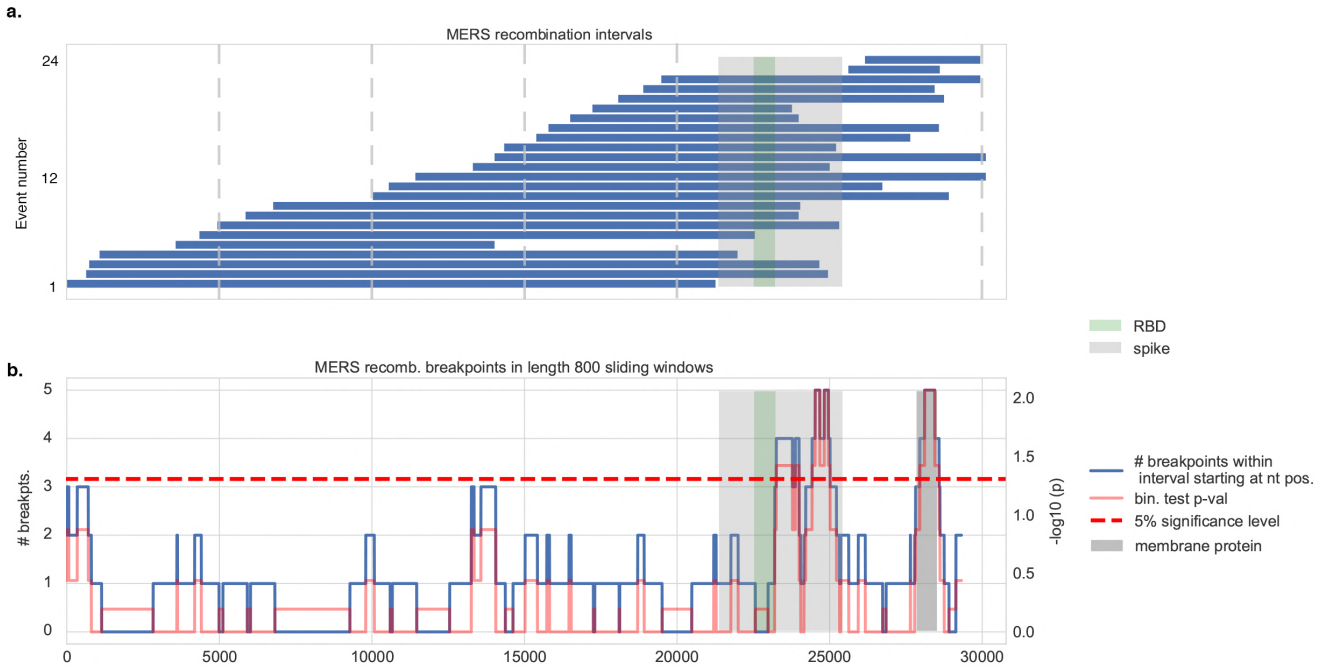

Fig. 2 | Recombination analysis in MERS coronaviruses. a. Distribution of 24 recombination events among human and non-human MERS-CoV isolates. The spike protein and its RBD are highlighted. **b.** Sliding window analysis shows (blue curve) the distribution of recombination breakpoints (either start or 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-CoV membrane protein is highlighted (dark gray); it also shows an enrichment of recombination breakpoints.

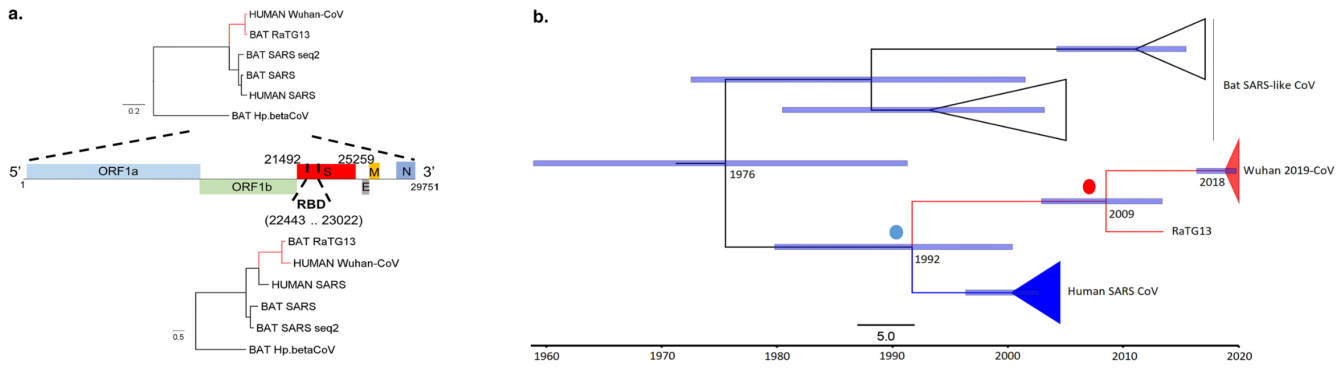

Fig. 3 | Recombination event in an ancestor of SARS-CoV-2 encompasses the RBD of the Spike gene. a. Full genome phylogenetic tree inference (top) shows SARS-CoV-2 sequence (HUMAN Wuhan-CoV) and its nearest bat-infecting RaTG13 strain clustering in a clade separate from the SARS-CoV isolates from human 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 SARS-CoV-2 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); SARS-CoV-2 (isolate 403962). **b**. Dated phylogeny of the RBD including RaTG13, 3 sequences from SARS-CoV-2 (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).

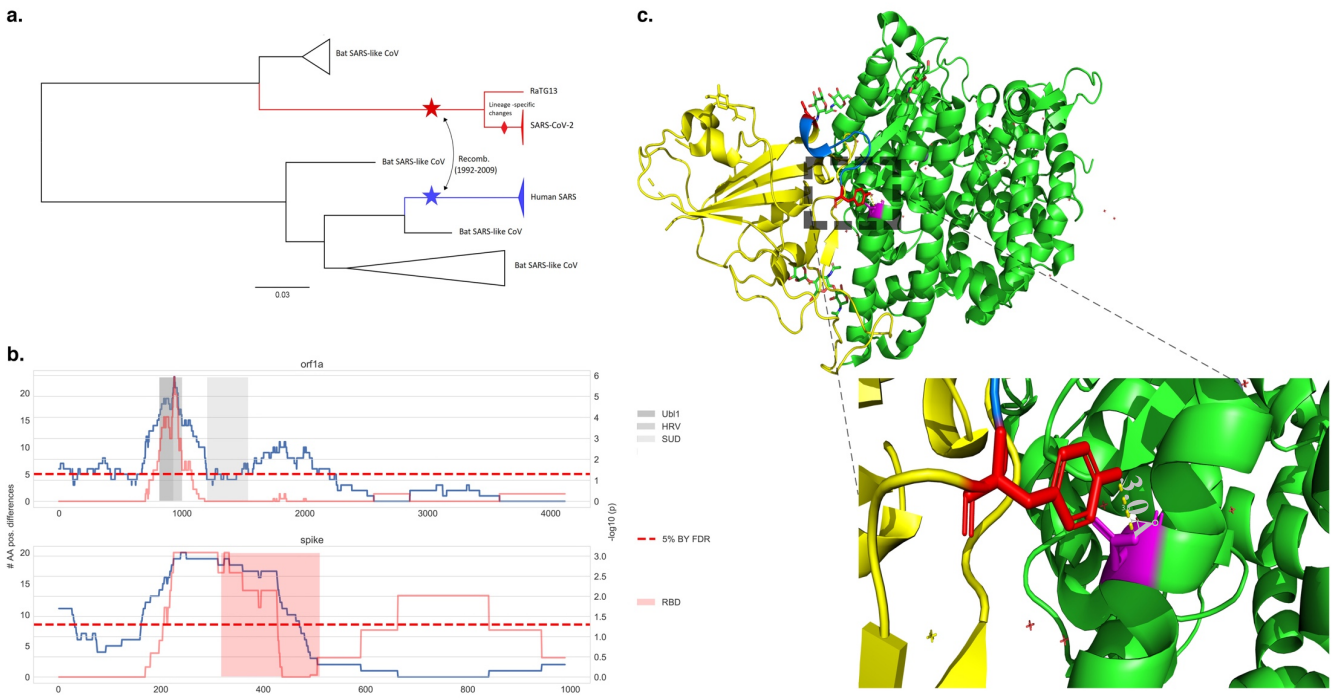

Fig. 4 | Two-hit scenario for the emergence of SARS-CoV-2. a. Phylogenetic representation summarizing the evolutionary events that likely led to the emergence of SARS-CoV-2: hit 1) Recombination of the RBD of the Spike protein between the lineage ancestral to both SARS-COV-2 and RaTG13 (red star) and the ancestral lineage of the human SARS-CoV (blue star); hit 2) SARS-CoV-2 accumulated five AA mutations in RBD since its divergece from its MRCA. **b.** Sliding window analysis (length 267 aa) identifies specific regions of SARS-CoV-2 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 SARS-CoV-2 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

SARS-CoV-2 and SARS/SARS-like-CoV

A set of 71 genome sequences derived from SARS-CoV-2 (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-CoV (n=72) and bat SARS-like CoV (n=39), publicly available in Genbank (ncbi.nlm.nih.gov/genbank/) (Supplementary Table 7). Alignment was performed either at genome wide nucleotide level or at the Spike CDS (at amino acid level)independently with MAFFTv7 ("auto" strategy)[30].

Seven recombination detection methods implemented in the RDP4 software package (RDP, Geneconv, Bootscan, Maxchi, Chimaera, SiScan, 3seq)[12] 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 SARS-CoV-2: KF636752 (bat), NC_004718 (human SARS), DQ071615 (bat SARS-like CoV), DQ412043 (bat SARS-like CoV), MG772934 (bat SARS-like CoV, MN996532 (RaTG13), NC_045512 (SARS-CoV-2).
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*[11]).

Phylogenetic analysis

The evolutionary relationships between SARS-CoV-2 and other SARS/SARS-like CoVs was inferred from genome alignment using PhyML (GTR + GAMMA 4CAT)[31]. 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 8), after assessing the molecular clock signal of the selected sequences with TempEst[32]. 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/.

Ancestral state reconstruction

The evolutionary changes that occurred in the Spike gene in Sarbecoviruses, either at amino acid level or 4-fold degenerated sites, were traced using a continuous-time Markov chain model for ancestral state reconstruction implemented in the Ape R package[33]. As input, the ML tree derived from 3rd codon positions at the RBD recombinant region was used. We then compared the distribution of synonymous and nonsynonymous substitutions along the branch specific to SARS-CoV-2.

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-CoV recombinations; and 3) both human-specific and non-human MERS-CoV 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 CoVs, 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[34] which provides a conservative multiple hypothesis correction that applies in arbitrary dependence conditions. For statistical significance, we chose 5% BY false discovery rate (FDR). Our discoveries are valid with different choices of window length, provided the window length is sensitive to the scale CoV proteins and the length of specific domains such as the RBD in the Spike gene.

We used the same sliding window approach to test for enrichment of gene-specific nonsynonymous as well as synonymous differences between SARS-CoV-2 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.