No detectable signal for ongoing genetic recombination in SARS-CoV-2

Damien Richard[1,2*], Christopher J. Owen[1], Lucy van Dorp[1], François Balloux[1*]

[1] UCL Genetics Institute, University College London, UK;

[2] Institute of Child Health, University College London, UK;

[*] Correspondence: richarddamienfr@gmail.com (Damien Richard), f.balloux@ucl.ac.uk (François Balloux).

Abstract

The COVID-19 pandemic has led to an unprecedented global sequencing effort of its viral agent SARS-CoV-2. The first whole genome assembly of SARS-CoV-2 was published on January 5 2020. Since then, over 150,000 high-quality SARS-CoV-2 genomes have been made available. This large genomic resource has allowed tracing of the emergence and spread of mutations and phylogenetic reconstruction of SARS-CoV-2 lineages in near real time. Though, whether SARS-CoV-2 undergoes genetic recombination has been largely overlooked to date. Recombination-mediated rearrangement of variants that arose independently can be of major evolutionary importance. Moreover, the absence of recombination is a key assumption behind the application of phylogenetic inference methods. Here, we analyse the extant genomic diversity of SARS-CoV-2 and show that, to date, there is no detectable hallmark of recombination. We assess our detection power using simulations and validate our method on the related MERS-CoV for which we report evidence for widespread genetic recombination.

Introduction

Genetic recombination is widely recognised as an important force in evolution, as it allows for the combination, within a single genome, of variants that arose independently in different genetic backgrounds [1]. Viruses are no exception to this pattern [2]. For example, recombination between its eight genomic segments (reassortment) is the fundamental mechanism behind the emergence of pandemic influenza A strains [3]. Recombination is also key for many viruses to generate new antigenic combinations that allow host immune systems evasion [4]. Moreover, the absence of recombination is also a prerequisite for phylogenetic inference [5]. Indeed, phylogenetic trees are limited to represent a single realisation of the past demography of the samples analysed. In the presence of genetic recombination, different regions of the sequence under scrutiny will support different evolutionary histories for the samples analysed, and hence result in conflicts in the topology of the phylogenetic reconstruction [5]. Not accounting for recombination can also lead to false positive detection of sites undergoing positive selection [6, 7].

Repositories of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes have grown at an unprecedented pace, with over 150,000 high-quality complete genome assemblies currently available on the Global Initiative on Sharing All Influenza Data (GISAID) repository as of 17/11/2020 [8, 9]. This allows for the near-real-time monitoring of the emergence and spread of novel mutations [10, 11] and the description of emerging lineages [12, 13]. Most of these studies rely on phylogenetic reconstructions of genome-wide Single Nucleotide Polymorphisms (SNPs), and implicitly assume the absence of pervasive genetic recombination in SARS-CoV-2. So far there has been limited effort to assess the extent of ongoing recombination in SARS-CoV-2 ([14, 15], https://observablehq.com/@spond/linkage-disequilibirum-in-sars-cov-2) despite its potential relevance to understanding the duration and propensity of co-infections in host.

Conflicts (incongruence) between phylogenies inferred from different genome segments can be indicative of recombination. Some of the numerous methods developed to detect genetic recombination rely on this concept [16-19]. The so-called “compatibility test” checks if all four combinations of alleles of a pair of biallelic sites (00, 01, 10, 11) are present among the sequences. More refined methods relying on this principle have been developed including the pairwise homoplasy index (PHI) [20]. Recombination also has the effect of decorrelating allele frequencies, with this effect increasing with physical distance along the genome. In a population undergoing frequent recombination, this causes linkage disequilibrium of alleles to decay with physical distance on the sequence [4]. The r^2 metric [21] is commonly used to measure linkage disequilibrium [22].

In this work, we aimed to detect signals of genetic recombination within the SARS-CoV-2 global population. We assembled a curated alignment of 6,546 available SARS-CoV-2 genomes enriched for those collected more recently to maximise genetic diversity in the dataset, and hence our ability to detect recombination. We applied two different statistical methods for the detection of genetic recombination and assessed their power using bespoke simulations. We validate our methodology on the related *Betacoronavirus* Middle East respiratory syndrome-related coronavirus (MERS-CoV) [23] responsible for the MERS outbreaks beginning in 2012, for which we find evidence for recombination, consistent with previous reports [24]. Our results do not identify detectable evidence for recombination in the SARS-CoV-2 population as of September 2020.

Results

No signal of recombination in SARS-CoV-2

We compiled an alignment of 6,546 SARS-CoV-2 isolates sampled across six continental regions. In order to maximize our detection power, whilst keeping the dataset computationally manageable, we chose to restrict our analysis to the most recently collected genomes during the month of September 2020. This is expected to maximise the genetic diversity in the dataset, and hence increase power to detect recombination events. We detect over 4,000 polymorphic positions following masking of sites putatively suggested as artefactual ([25] https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 29/10/2020). These include a large fraction of homoplasies (29.6%), thought to largely be induced by host immune system RNA editing [15, 26, 27].

A PHI test applied to the SARS-CoV-2 alignment reported a *p*-value of 0.78 suggesting no signal of recombination. Consistently, LD decay regression coefficients and R squared statistics did not fall outside of the distributions obtained after randomly permuting genome coordinates (Figure 1).

To test our ability to detect low levels of recombination, we simulated alignments with levels of genetic diversity matching that observed in the true SARS-CoV-2 alignment, but using varying recombination rates. We detected recombination in 100% of the datasets simulated with 3e-3 recombination events per genome per viral replication (60% for a rate of 3e-4, Supplementary Table S1). This low detection power is linked to the high homogeneity of the SARS-CoV-2 population, reflected by the mean pairwise distance of 19.4 (95%HPD 3-30) SNPs in the alignment analysed.

In addition, we searched the global SARS-CoV-2 phylogeny for isolates displaying root-to-tip distances in the upper 5% quartile of the distribution. These may offer some of the best candidates for isolates having experienced recombination, which is expected to increase terminal branch length. We detected 24 phylogenetic outliers which grouped into 13 phylogenetic clades. Localisation of their mutations and those of their phylogenetic neighbours in matrices did not support a recombinant origin. Indeed, rather than displaying syntenic groups of private mutations that would suggest a recombination-mediated origin, the 24 outliers mostly displayed a randomly distributed excess of mutations (Supplementary Figures S1-S13). Our results therefore suggest that recombination in SARS-CoV-2 is either absent, or occurring at a rate too low relative to mutation to be detectable under the genetic diversity charactering the SARS-CoV-2 population at this stage.

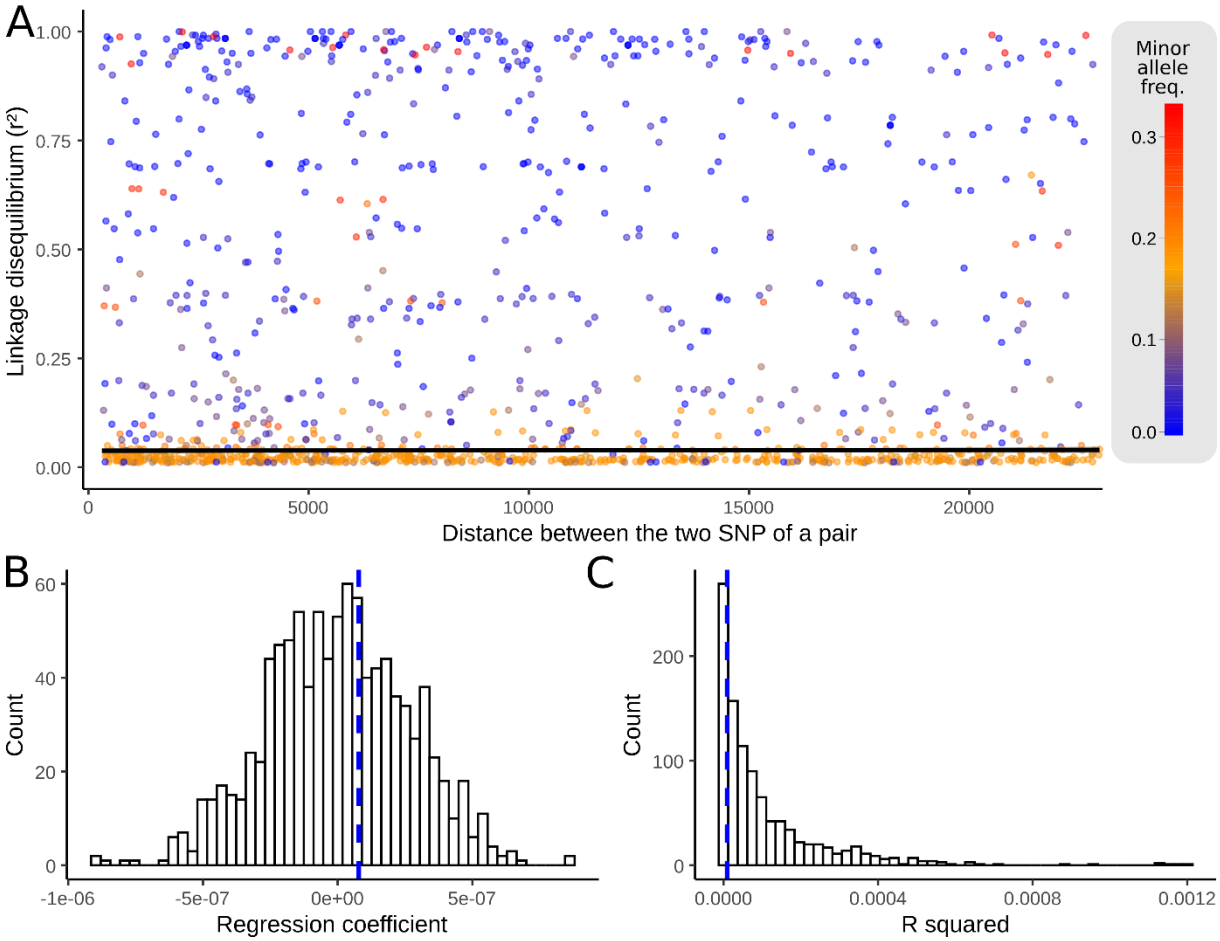

Figure 1: Linkage disequilibrium (r^2) as a function of physical distance on the SARS-CoV-2 genome. (A) Linkage disequilibrium measured by r^2 (y-axis) for all pairs of SNPs represented as a function of the genetic distance separating the SNPs of each pair. Black line: fitted linear model (regression coefficient: 7.84e-8; R-squared: 9.33e-6). (B) Distribution of regression coefficients of the linear models obtained following consideration of 1000 position permuted datasets. Blue dashed line: value obtained for the true SARS-CoV-2 alignment. (C) Distribution of the R squared values of the linear models of 1000 position permuted datasets. Blue dashed line: value of the SARS-CoV-2 true alignment.

Recombination occurs in MERS-CoV

To validate our approach, we applied the same method to an alignment of MERS-CoV, a related *Betacoronavirus* thought to be widely recombining [24, 28]. Counter to observations for SARS-CoV-2, the MERS-CoV dataset yielded detectable evidence of recombination. Beside the decay of the r^2 ~ distance regression slope, values of both R-squared and the regression coefficient largely fall outside of the distributions of the same parameters of the permuted datasets (Figure 2). The PHI test reported a p-value <1e-12. Of note, these tests were repeated after discarding C to T mutations, mostly caused by the host immune RNA editing systems, that might produce an artefactual signal of recombination. Tests on the pruned alignment still provide evidence of significant signals of recombination in MERS-CoV (PHI test p-value <1e-12 and significant decay of r^2, Supplementary Figure S14).

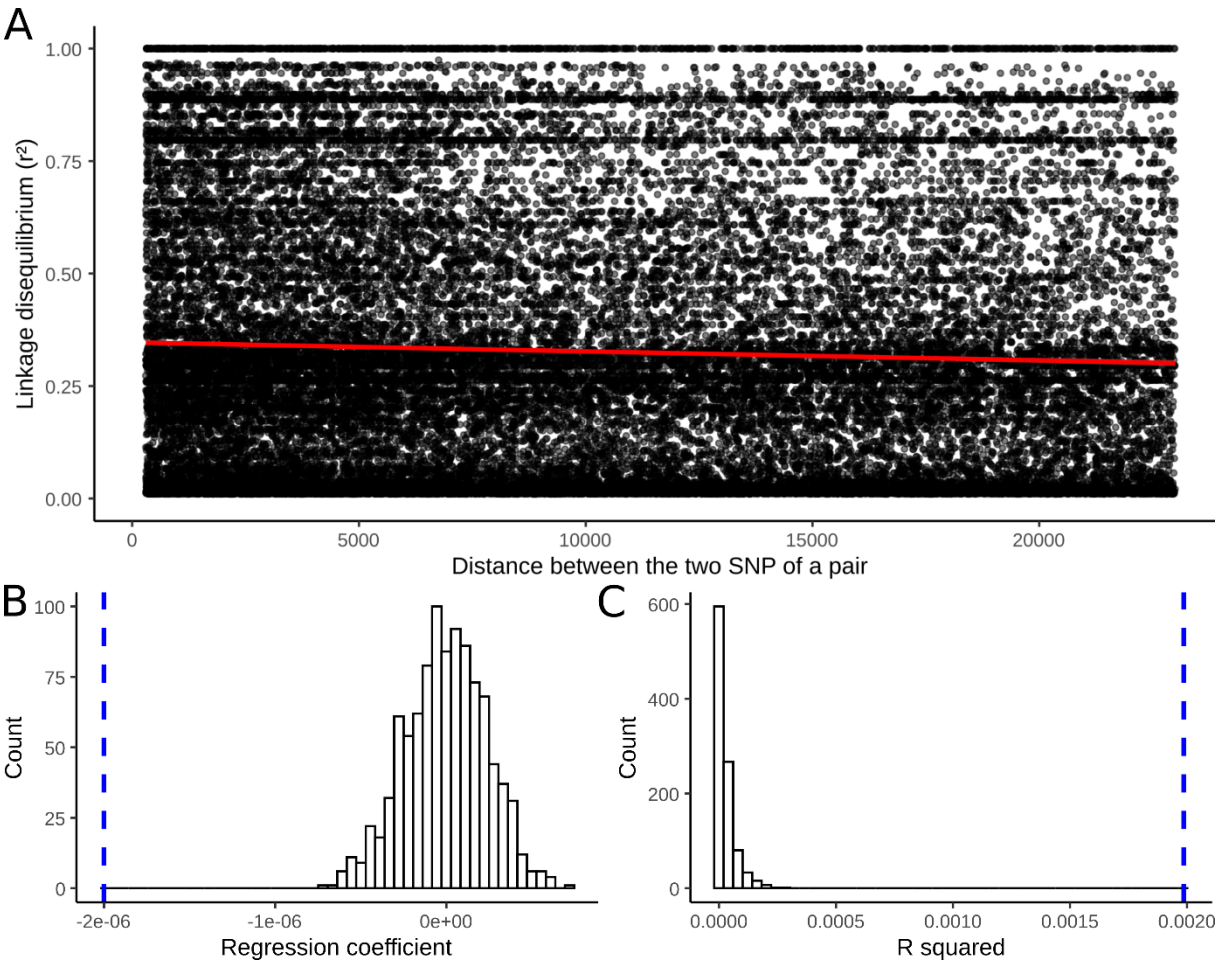

Figure 2: Linkage disequilibrium (r^2) as a function of physical distance on the MERS-CoV genomes. Pairs comprising SNP differing by a frequency >=0.1 have been discarded, lowering the number of pairs from 261,090 to 212,342. (A) Linkage disequilibrium (y-axis) for all pairs of SNPs is represented as a function of the distance separating the SNPs of each pair. Red line: fitted linear model (regression coefficient: -2.00e-6; R-squared: 1.99e-3). (B) Distribution of the regression coefficients of the linear models obtained following consideration of 1000 position permuted datasets. Blue dashed line: value of the MERS-CoV true alignment. (C) Distribution of the R squared values of the linear models of 1000 position permuted datasets. Blue dashed line: value of the MERS-CoV true alignment.

Discussion

In this study, we analyzed a dataset of 6,546 SARS-CoV-2 assemblies sampled during the month of September 2020 across six continental regions. Applying two distinct detection methods, we did not find signal of recombination among the SARS-CoV-2 population tested. Conversely, we detected evidence of recombination in SARS-CoV-2-like simulated recombining datasets as well as in a set of 459 MERS-CoV coronaviruses assemblies, known for being prone to recombination.

A priori it is highly plausible that SARS-CoV-2 has the potential to recombine [28]. Recombination has been suggested to be common in coronaviruses, including both for human [29-33] and animal [34-36] associated lineages, as inferred from genomic approaches [37], observed in cell culture [38, 39] and *in vivo* [40]. It has been claimed that all of the human epidemic coronaviruses: SARS-CoV-1 [41-43], MERS-

CoV [44], and SARS-CoV-2 [45-47] may have evolved through recombination events leading to some genome mosaicism, particular over receptor binding regions.

At this stage we do not detect a genetic hallmark for recombination in SARS-CoV-2. This does not necessarily imply that SARS-CoV-2 lacks the ability to recombine. For genetic recombination to leave a measurable signal in the genetic data, there needs to be sufficient genetic differentiation between the recombining viruses. Given the low intra-host genetic diversity of epidemic viruses such as SARS-CoV-2, this requires mixed infections (i.e. coinfection of the same host by distinct SARS-CoV-2 lineages). While such events are expected to be rare, there have been reports of mixed infections [48]. Though, given the limited genetic diversity of SARS-CoV-2 strains currently in circulation, even mixed infections may often not involve sufficiently differentiated strains to leave a detectable signal following a recombination event. The recent host jump into humans of SARS-CoV-2, most likely through a single transmission to humans from an unknown animal reservoir, created an essentially genetically invariant viral population, with genetic diversity building up through the accumulation of mutations since the begininning of the pandemic. The genomic diversity of SARS-CoV-2 is still far below its mutation-drift equilibrium and remains very low at this stage [10]. As a result, putative recombination events would only be supported by a limited number of SNPs, and would require high detection sensitivity to be identified.

In contrast, we detected recombination in MERS-CoV despite the far smaller sample size of the alignment analysed. Besides the higher genetic diversity of MERS-CoV at this stage, this may also be due to major epidemiological differences between MERS and COVID-19. MERS is mainly a disease of dromedary camels, with spillover events into humans [28]. In camels, the high prevalence of the disease and the mostly mild symptoms it causes is suggested to favour co-infection [24]. On the contrary, the severity of the symptoms in human lowers the probability of co-infection. The camel host, which is known to harbour other coronaviruses, could provide a hub of genetic diversity creation in MERS-CoV through recombination [49]. Human MERS-CoV infection was first documented in 2012, but it is thought the virus had been previously circulating for at least a few years in camels [50]. Our MERS-CoV alignment comprises samples spanning from 2012 to 2019. Mutations accumulating over this time-scale provide more diverse genetic markers that facilitate the detection of putative recombination events.

While we did not detect evidence of genetic recombination in SARS-CoV-2 to date, it remains of importance to repeat such analyses as the genetic diversity of the SARS-CoV-2 population will increase, and to consider its possible impact when conducting phylogenetics studies in the future.

Materials and Methods

SARS-CoV-2 dataset

All 6,546 SARS-CoV-2 high quality genomes (containing less than 5% of "N" and being >29,000 bp long) sampled during the month of September 2020 available on GISAID (as of October 15th 2020) were downloaded and profile aligned to the Wuhan-Hu-1 reference genome (GenBank accession MN908947; GISAID ID EPI_ISL_402125) using MAFFT v7.471 [51]. A full list of acknowledgements together with submitting and originating laboratories is provided in Supplementary Table S2. SNPs flagged as putative sequencing errors were discarded (https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 29/10/2020). The final dataset comprised 4,199 SNPs and a mean pairwise SNP difference of 19.4 (95%HPD 3-30). Following construction of a maximum likelihood tree using IQTree Covid-release [52], 29.6% of SNPs were identified as homoplasic by HomoplasyFinder [53].

MERS dataset

456 high-quality MERS-CoV genomes isolated from both camels and human were downloaded from the NCBI Virus database and profile aligned to the HCoV-EMC/2012 reference genome (GenBank accession NC_019843) using MAFFT v7.471 [51] (Supplementary Table S3). We detected 8,788 SNPs in the alignment, with a mean pairwise SNP count of 123.36 (95%HPD 18-234). Homoplasies were identified as described above and represented 12.4% of the polymorphic positions.

Detection of recombination

Two recombination tests were performed on each dataset. First, a pairwise homoplasy index (PHI) test was used to detect recombination setting the number of permutations to 100 and the window size set to 300 bp with otherwise default parameters [20]. Additionally, we computed the linkage disequilibrium (r^2) for all pairs of bi-allelic SNPs occurring in >=1% of the isolates using tomahawk (https://mklarqvist.github.io/tomahawk/). 90% of the SNP pairs grouped <= 23,000 nucleotides apart. Linkage disequilibrium estimation over distances larger than 23,000 rely on a few SNP pairs only, so we restricted the dataset to those 90% pairs. A linear model was fitted to the distribution of r^2 values as a function of the distance separating the two SNPs in each pair. The regression coefficient of this linear model indicates whether linkage disequilibrium decays with physical distance or not. To formally test for the presence of recombination, we produced 1000 permuted datasets (randomly associating r^2 values with distance values) and fitted a linear model to each one of the permuted datasets. We then assessed whether the real R-squared and regression coefficients values fell either inside or outside of the distributions of the parameters generated by the randomly permuted datasets. A limitation of the use of the r^2 metrics as an estimator of linkage disequilibrium is its dependency on allele frequencies, causing a possible reduction in statistical power [54]. It has therefore been proposed to compute r^2 only for pairs of SNPs that do not differ markedly in frequency in the studied population [55]. Discarding pairs of SNPs is suboptimal in the context of SARS-CoV-2's already restricted genetic diversity. However, we still implemented this approach which led to similar results to tests on non-frequency filtered SNPs (Supplementary Figure S15).

We performed a third test for recombination in SARS-CoV-2 by focusing specifically on isolates that were flagged as phylogenetic outliers in the global phylogeny. Recombinant isolates are expected to

be located at the tip of long terminal branches if there is phylogenetic incongruency between the mutations they carry. We applied TreeShrink to identify the accessions displaying root-to-tip distances in the upper 5% quartile (-q 0.05 parameter) of the root-to-tip distance distribution [56]. The mutations carried by those outliers were visualy compared to that of their neighbours in the phylogeny.

Power of recombination detection

In order to characterise the statistical power of the recombination detection methods employed, we simulated in silico SARS-CoV-2 alignments using MSprime [57]. The simulated mutation rate was set to match that of the real dataset (Supplementary Figure S16). We generated datasets with numbers of recombination events per genome per viral replication of 0, 3e-7, 3e-6, 3e-5, 3e-4, 3e-3 and 3e-2 (ten replicates each). PHI tests and linkage disequilibrium decay tests were performed on those simulated datasets as described previously.

References

1. Muller HJ: **The relation of recombination to mutational advance.** *Mutat Res* 1964, **106:**2-9.
2. Koonin EV, Dolja VV, Krupovic M: **Origins and evolution of viruses of eukaryotes: The ultimate modularity.** *Virology* 2015, **479-480:**2-25.
3. Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, Chen H, Webster RG, Peiris JSM, Guan Y: **Dating the emergence of pandemic influenza viruses.** 2009, **106:**11709-11712.
4. Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F: **Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences.** *Infection, Genetics and Evolution* 2015, **30:**296-307.
5. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny estimation.** *J Mol Evol* 2002, **54:**396-402.
6. Anisimova M, Nielsen R, Yang Z: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164:**1229-1236.
7. Shriner D, Nickle DC, Jensen MA, Mullins JI: **Potential impact of recombination on sitewise approaches for detecting positive natural selection.** *Genet Res* 2003, **81:**115-121.
8. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** 2017, **1:**33-46.
9. Shu Y, McCauley J: **GISAID: Global initiative on sharing all influenza data – from vision to reality.** 2017, **22:**30494.
10. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al: **Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.** *Infection, Genetics and Evolution* 2020, **83:**104351.
11. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva Filipe A, Wojcechowskyj JA, Davis C, Piccoli L, Pascall DJ, Dillen J, et al: **The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity.** 2020**:**2020.2011.2004.355842.
12. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG: **A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.** *Nature Microbiology* 2020, **5:**1403-1407.
13. Hodcroft EB, Zuber M, Nadeau S, Comas I, González Candelas F, Stadler T, Neher RA: **Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020.** 2020**:**2020.2010.2025.20219063.
14. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, et al: **On the origin and continuing evolution of SARS-CoV-2.** *National Science Review* 2020, **7:**1012-1023.
15. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F: **No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2.** *Nature Communications* 2020, **11:**5986.
16. Posada DJMb, evolution: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** 2002, **19:**708-717.
17. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22:**3096-3098.
18. Boni MF, de Jong MD, van Doorn HR, Holmes EC: **Guidelines for identifying homologous recombination events in influenza A virus.** *PLOS ONE* 2010, **5:**e10434.
19. Lam HM, Ratmann O, Boni MF: **Improved algorithmic complexity for the 3SEQ recombination detection algorithm.** *Molecular Biology and Evolution* 2017, **35:**247-251.
20. Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172:**2665-2681.
21. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38:**226-231.

22. Haydon DT, Bastos ADS, Awadalla P: **Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments.** 2004, **85:**1095-1100.
23. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia.** *N Engl J Med* 2012, **367:**1814-1820.
24. Dudas G, Rambaut A: **MERS-CoV recombination: implications about the reservoir and potential for adaptation.** *Virus evolution* 2016, **2:**vev023-vev023.
25. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G, et al: **Stability of SARS-CoV-2 phylogenies.** *PLOS Genetics* 2020, **16:**e1009175.
26. Simmonds P: **Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories.** 2020, **5:**e00408-00420.
27. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG: **Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2.** 2020, **6:**eabb5813.
28. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF: **Epidemiology, genetic recombination, and pathogenesis of coronaviruses.** *Trends in Microbiology* 2016, **24:**490-502.
29. Lau SKP, Lee P, Tsang AKL, Yip CCY, Tse H, Lee RA, So L-Y, Lau Y-L, Chan K-H, Woo PCY, Yuen K-Y: **Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination.** 2011, **85:**11325-11337.
30. Kin N, Miszczak F, Lin W, Gouilh MA, Vabret A, Consortium E: **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* 2015, **7:**2358-2377.
31. Pyrc K, Dijkman R, Deng L, Jebbink MF, Ross HA, Berkhout B, van der Hoek L: **Mosaic structure of human coronavirus NL63, one thousand years of evolution.** *Journal of Molecular Biology* 2006, **364:**964-973.
32. Woo PCY, Lau SKP, Yip CCY, Huang Y, Tsoi H-W, Chan K-H, Yuen K-Y: **Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1.** 2006, **80:**7136-7145.
33. Graham RL, Baric RS: **Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission.** 2010, **84:**3134-3146.
34. Terada Y, Matsui N, Noguchi K, Kuwata R, Shimoda H, Soma T, Mochizuki M, Maeda K: **Emergence of pathogenic coronaviruses in cats by homologous recombination between feline and canine coronaviruses.** *PLOS ONE* 2014, **9:**e106534.
35. Decaro N, Mari V, Campolo M, Lorusso A, Camero M, Elia G, Martella V, Cordioli P, Enjuanes L, Buonavoglia C: **Recombinant canine coronaviruses related to transmissible gastroenteritis virus of swine are circulating in dogs.** 2009, **83:**1532-1537.
36. Tian P-F, Jin Y-L, Xing G, Qv L-L, Huang Y-W, Zhou J-Y: **Evidence of recombinant strains of porcine epidemic diarrhea virus, United States, 2013.** *Emerging infectious diseases* 2014, **20:**1735-1738.
37. Herrewegh AA, Smeenk I, Horzinek MC, Rottier PJ, de Groot RJ: **Feline coronavirus type II strains 79-1683 and 79-1146 originate from a double recombination between feline coronavirus type I and canine coronavirus.** *J Virol* 1998, **72:**4508-4514.
38. Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA: **Recombination between nonsegmented RNA genomes of murine coronaviruses.** *J Virol* 1985, **56:**449-456.
39. Makino S, Keck JG, Stohlman SA, Lai MM: **High-frequency RNA recombination of murine coronaviruses.** *J Virol* 1986, **57:**729-737.
40. Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM: **In vivo RNA-RNA recombination of coronavirus in mouse brain.** *J Virol* 1988, **62:**1810-1813.

41. Hon C-C, Lam T-Y, Shi Z-L, Drummond AJ, Yip C-W, Zeng F, Lam P-Y, Leung FC-C: **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.** 2008, **82:**1819-1826.
42. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, et al: **Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus.** *PLOS Pathogens* 2017, **13:**e1006698.
43. Lau SKP, Li KSM, Huang Y, Shek C-T, Tse H, Wang M, Choi GKY, Xu H, Lam CSF, Guo R, et al: **Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events.** 2010, **84:**2808-2819.
44. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF: **Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat.** *J Virol* 2014, **88:**11297-11303.
45. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF: **The proximal origin of SARS-CoV-2.** *Nature Medicine* 2020, **26:**450-452.
46. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL: **Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.** *Nature Microbiology* 2020, **5:**1408-1417.
47. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al: **A new coronavirus associated with human respiratory disease in China.** *Nature* 2020, **579:**265-269.
48. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, et al: **Shared SARS-CoV-2 diversity suggests localised transmission of minority variants.** 2020**:**2020.2005.2028.118992.
49. Sabir JSM, Lam TT-Y, Ahmed MMM, Li L, Shen Y, E. M. Abo-Aba S, Qureshi MI, Abu-Zeid M, Zhang Y, Khiyami MA, et al: **Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia.** 2016, **351:**81-84.
50. Lau SKP, Wong ACP, Lau TCK, Woo PCY: **Molecular evolution of MERS coronavirus: dromedaries as a recent intermediate host or long-time animal reservoir?** *Int J Mol Sci* 2017, **18**.
51. Katoh K, Standley DM: **MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.** *Molecular Biology and Evolution* 2013, **30:**772-780.
52. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R: **IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.** *Molecular Biology and Evolution* 2020, **37:**1530-1534.
53. Crispell J, Balaz D, Gordon SV: **HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny.** *Microbial genomics* 2019, **5:**e000245.
54. VanLiere JM, Rosenberg NA: **Mathematical properties of the r2 measure of linkage disequilibrium.** *Theoretical population biology* 2008, **74:**130-137.
55. Eberle MA, Rieder MJ, Kruglyak L, Nickerson DA: **Allele Frequency Matching Between SNPs Reveals an Excess of Linkage Disequilibrium in Genic Regions of the Human Genome.** *PLOS Genetics* 2006, **2:**e142.
56. Mai U, Mirarab S: **TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees.** *BMC Genomics* 2018, **19:**272.
57. Kelleher J, Etheridge AM, McVean G: **Efficient coalescent simulation and genealogical analysis for large sample sizes.** *PLoS Comput Biol* 2016, **12:**e1004842.

Data and Code Availability

All analysed SARS-CoV-2 data is available on registration to GISAID with the accession IDS and acknowledgements provided in Table S2. MERS-CoV assemblies are freely available on NCBI with the included accessions provided in Table S3. Scripts used in this study are available at https://github.com/DamienFr/LD_SARS-CoV-2.

Competing Interests

The authors have no competing interests to declare.

Acknowledgements

D.R. is supported by a NIHR Precision AMR award. C.O is funded by a NERC-DTP studentship. L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). L.v.D. is supported by a UCL Excellence Fellowship. We wish to particularly acknowledge all of the large number of contributing and submitting laboratories sharing SARS-CoV-2 assemblies via the GISAID platform, including the UK (COG-UK) consortium (a full list of consortium names and affiliations can be found at https://www.cogconsortium.uk). COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute.