SUPPLEMENTARY INFORMATION

Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans

Oscar A. MacLean¹,ª, Spyros Lytras¹,ª, Joshua B. Singer¹, Steven Weaver², Sergei L. Kosakovsky Pond²,*, David L. Robertson²,*

¹MRC-University of Glasgow Centre for Virus Research, Scotland, UK.
²Temple University, Institute for Genomics and Evolutionary Medicine, Philadelphia, USA.

#Joint first authors.
*To whom correspondence should be addressed: david.l.robertson@glasgow.ac.uk, spond@temple.edu

Signals of positive selection in SARS-CoV-2

Searching for positive selection within the SARS-CoV-2 outbreak radiation. We used the FUBAR method¹ to identify sites exhibiting signatures of diversifying positive selection in the SARS-CoV-2 outbreak data. Such signatures could be suggestive of ongoing human adaptation of the virus in the pandemic. FUBAR detects positive selection by looking for codons which have an elevated rate of nonsynonymous protein coding substitutions (dN) relative to synonymous substitutions (dS). It allows both synonymous and non-synonymous rate variation across codons, as has been observed in SARS-CoV-2 evolution², but assumes that selective pressures are constant through time (no branch to branch variation). The input data for this software was a concatenated coding alignment of 396 sequences from GISAID up to March 16th, 2020.

This analysis reported ten sites as showing significant evidence of positive selection across the pandemic phylogeny (Table 1). Due to the low diversity in these 396 SARS-CoV-2 samples, there is limited power to confidently estimate the synonymous and nonsynonymous substitution rate for each codon. This means that statistical power to identify positive selection in the form of dN/dS>1 for any given codon is limited, and the posterior distribution should be flat. The presence statistically significantly signatures of positive selection is therefore somewhat surprising.

Supplementary Table 1. The location and posterior probabilities of the ten mutations detected by the FUBAR selection analysis.

| Codon in concatenated alignment | ORF  | mutation | Posterior probability |
|--------------------------------|------|----------|-----------------------|
| 476                            | NSP2 | I296V    | 0.935871674           |
| 1599                           | NSP3 | L781F    | 0.943645748           |
| 3606                           | NSP6 | L37F     | 0.975858139           |
| 7461                           | Spike| V367F    | 0.952007026           |
| 7708                           | Spike| D614G    | 0.939001013           |
| 7954                           | Spike| V860Q    | 0.919218684           |
|   |   |   |   |
|---|---|---|---|
| 7955 | Spike | L861K | 0.944083041 |
| 8720 | ORF-M | D3G | 0.938851732 |
| 9248 | ORF-8 | L84S | 0.939347212 |
| 9577 | ORF-N | I292T | 0.952894581 |

To understand the specific mutational patterns that might explain these significant results, we looked at where in the phylogeny these putatively positively selected mutations were occurring. For all but two of the ten positive selected codons, this signal was being driven by apparent convergent evolution (or homoplasy) in the tree, with the same mutation occurring in parallel across the phylogeny. To investigate whether this observation was truly due to independent events or because of recombination signatures in the SARS-CoV-2 outbreak tree, we firstly determined if the samples with these convergent mutations were geographically correlated. As selective pressure acting on an un treatable novel zoonotic virus is likely to be globally shared (adaptation to humans), but recombination requires co-localisation of viruses in the time and space, geographic clustering would be a good indication that these mutations are not independent.

The homoplasies driving ORF8 L84S and ORF1ab L1599S mutations were both found in South Korean isolates, and each of the two instances of ORF M D3G and ORF N I292T were found in the Netherlands. This geographic clustering was suggestive of recombination and was investigated further.

**Recombination or selection.** Two of the ten FUBAR-flagged sites, Spike codons 860 and 861, did not show any homoplasies. Both signals could be attributed to the same run of four neighbouring U to A mutations spanning the two codons. These mutations were found in only a single sample: EPI_ISL_408485 from Beijing and have not been observed since (to date 8/5/2020). This suggests that they were either sequencing errors, or a large single mutation spanning two codons, which has not subsequently spread. Multiple nucleotide changes within a single codon should be rare and sequencing error is a plausible explanation.

The positive selection signature at NSP6 codon 37 can be explained by multiple homoplasies of G to U mutations at nucleotide 11083. This mutation is found in four distinct haplotypes (Supplementary figure 1) across different areas of the phylogeny. There are flanking mutations on both sides of this site shared by sequences which both possess and do not possess the 11083 mutation. For this to occur under a recombination scenario, multiple breakpoints would be required for each homoplasy. These observations therefore are not most parsimoniously explained by recombination alone. The presence of this mutation across the tree could be driven by either positive selection, parallel sequencing error or hypermutability through polymerase slippage.
Supplementary Figure 1. Alignment of variable sites (removing invariant sites across these samples) surrounding the ORF1A L3606F recurrent G→T mutation. It appears to independently originate on four genetic backgrounds. The presence of ‘n’s in a few of the samples suggests that sequencing ambiguity is common in this region.

Both the ORF8 codon 84 and ORF1ab codon 1599 positive selection signals appear to be raised by a single South Korean sample (GISAID accession 413017). This sample possesses two derived mutations either side of a hypothesised breakpoint. These pairs of derived mutations belong to samples with different haplotypes (Supplementary figure 2). Therefore the 413017 sample appears to be a recombinant between sample 413018 and 413513 or 412871. As both 413017 and 413018 were sequenced by the same lab and released at the same time, this recombination event may be an artefactual product of lab cross-contamination.

Supplementary figure 2. Alignment of variable sites in the whole genome alignment, with base positions shown above. The putative recombinant South Korean sample 413017 shows mismatching topologies across its genome, clustering with different South Korean samples either side of the inferred breakpoint. The Wuhan sample 402124 was collected on 30/12/2019, and shows no unique mutations, it should thus represent the ancestral state of the four chosen samples.
Supplementary figure 3. A) Spike D614G mutation: sites either side of D614G show derived mutations in Wuhan 412982 congruent with the tree, suggesting that it is not a recombinant, or has multiple breakpoints. B) N ORF I292T: both Netherlands samples with the mutation were sequenced by the same Dutch lab and released at the same time. C) Spike V367F homoplasy at a single site in Hong Kong, unlikely to be a recombinant. D) An observed insertion homoplasy in newer data.

The Spike D614G signal was driven by apparent convergent evolution between one Wuhan sample (412982) in addition to the main lineage containing 86 samples. This sample shares mutations with its closest related sequence (Guangdong 413867) on both sides of this homoplasy (Supplementary figure 3A), suggesting it is not the result of recombination. No newly sequenced samples uploaded up to 27/4/2020 containing the D614G mutation clustered with the Wuhan 412982 sample, suggesting that this haplotype did not spread or that this homoplasy is driven by sequencing error. Additional sequences displaying apparent convergent evolution at this site have since been sequenced, these have been taken as evidence of positive selection. However, given that this mutation now occurs in 59% of sequenced samples (as of 27/4/20), it will be one of the mutations most likely to be variable if multiple viral genotypes are present following lab contamination or in mixed infections, and so most prone to being shuttled onto new backgrounds by recombination. Therefore, whilst high frequency mutations are the most important to study, they are also the most prone to misleading homoplasies, and must be analysed with the most caution.
The N ORF 292 site detected by FUBAR is driven by a similar convergent evolution event history. However, both samples exhibiting the same derived I to T mutation (Figure 3B; GISAID IDs 413570 and 413574) were sequenced by the same Dutch lab and released at the same time, again suggesting that lab cross contamination is a likely driver. However, unlike South Korean sample 413017, there is only one shared derived mutation (codon 292), and therefore the genomic evidence for recombination in these samples is weaker.

The Spike V367F signal was driven by apparent convergent evolution between four French samples sequenced in January and a Hong Kong sample 412028, which shows shared variation either side of the homoplasy suggesting it is not a recombinant (Supplementary figure 3C). Looking through more recent data shows additional homoplasies in a simple neighbour joining tree. Additionally, newly generated sequences since the FUBAR analysis cluster around the Hong Kong sample, further suggesting it is not a lab sequencing error.

Subsequent informal scans of newer data have revealed evidence of additional lab recombination events (see Supplementary figure 3D).
Frequency-based analysis of SARS-CoV-2 polymorphisms

In addition to searching for positive selection, we investigated if signatures of purifying selection on segregating variation in the current SARS-CoV-2 data could be observed (sequences as of 14/5/20). We compared the relative frequencies of nonsynonymous and synonymous mutations in the pandemic data. Codons with multiple mutations present were discarded from the analysis to avoid ambiguity in the order of mutations, and simplify synonymous/nonsynonymous classification.

Most mutations of both classes are at very low frequency (main text Figure 2), indicative of the viral population expansion that the pandemic has undergone. dN/dS was approximately 0.6 in singletons, suggesting that 40% of nonsynonymous mutations are strongly deleterious and therefore never observed in the population. There is a weak observable trend towards a higher proportion of mutations being synonymous at the highest frequency intervals, suggestive of some ongoing selection against circulating amino acid replacements in the pandemic. This observation may be partially driven by sequencing errors which are not transmitted and so are at low frequency. These sequencing errors are likely to have a dN/dS value of 1, which may make the estimate that 40% amino acid replacements are strongly deleterious an underestimate of the true value. However, the decline in nonsynonymous/synonymous ratio occurs across the range of frequencies, suggesting that sequencing errors alone are not driving the trend. It is important to consider that the observed frequencies are likely to differ from true global frequencies due to biased sampling of infections in the pandemic\(^4\), and so we caution against overinterpretation of specific mutation frequencies.
Supplementary figure 4. Synonymous rate of each of the three SRV classes estimated by BUSTED[S] + HMM and the proportion attributed to each class for all non-recombinant open reading frames.
Supplementary References
1. Murrell, B. et al. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.* (2013) doi:10.1093/molbev/mst030.
2. Nielsen, R., Wang, H. & Pipes, L. Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *bioRxiv* 2020.04.20.052019 (2020) doi:10.1101/2020.04.20.052019.
3. Phelan, J. et al. Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world. *bioRxiv* 2020.04.28.066977 (2020) doi:10.1101/2020.04.28.066977.
4. MacLean, O. A., Orton, R. J., Singer, J. B. & Robertson, D. L. No evidence for distinct types in the evolution of SARS-CoV-2. doi:10.1093/ve/veaa034.