

SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes

Irwin Jungreis (✉ ILJungr@csail.mit.edu)

Massachusetts Institute of Technology <https://orcid.org/0000-0002-3197-5367>

Rachel Sealfon

Flatiron Institute

Manolis Kellis

MIT <https://orcid.org/0000-0001-7113-9630>

Article

Keywords: Protein-coding Gene Set, Evolutionary Constraint, Functional Mutations, Immune Avoidance, Spike-protein Mutations

Posted Date: October 1st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-80345/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on May 11th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-22905-7>.

1 SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 2 Sarbecovirus genomes

3 Irwin Jungreis^{1,2*}, Rachel Sealfon³, Manolis Kellis^{1,2*}

4 ¹MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA;

5 ²Broad Institute of MIT and Harvard, Cambridge, MA;

6 ³Center for Computational Biology, Flatiron Institute, New York, NY

7 *Corresponding authors: iljungr@csail.mit.edu and manoli@mit.edu

8

9 **Summary**

10 Despite its overwhelming clinical importance, the SARS-CoV-2 gene set remains unresolved, hindering
11 dissection of COVID-19 biology. Here, we use comparative genomics to provide a high-confidence
12 protein-coding gene set, characterize protein-level and nucleotide-level evolutionary constraint, and
13 prioritize functional mutations from the ongoing COVID-19 pandemic. We select 44 complete
14 Sarbecovirus genomes at evolutionary distances ideally-suited for protein-coding and non-coding
15 element identification, create whole-genome alignments, and quantify protein-coding evolutionary
16 signatures and overlapping constraint. We find strong protein-coding signatures for all named genes
17 and for 3a, 6, 7a, 7b, 8, 9b, and also ORF3c, a novel alternate-frame gene. By contrast, ORF10, and
18 overlapping-ORFs 9c, 3b, and 3d lack protein-coding signatures or convincing experimental evidence
19 and are not protein-coding. Furthermore, we show no other protein-coding genes remain to be
20 discovered. Cross-strain and within-strain evolutionary pressures largely agree at the gene, amino-acid,
21 and nucleotide levels, with some notable exceptions, including fewer-than-expected mutations in nsp3
22 and Spike subunit S1, and more-than-expected mutations in Nucleocapsid. The latter also shows a
23 cluster of amino-acid-changing variants in otherwise-conserved residues in a predicted B-cell epitope,
24 which may indicate positive selection for immune avoidance. Several Spike-protein mutations, including

25 D614G, which has been associated with increased transmission, disrupt otherwise-perfectly-conserved
26 amino acids, and could be novel adaptations to human hosts. The resulting high-confidence gene set
27 and evolutionary-history annotations provide valuable resources and insights on COVID-19 biology,
28 mutations, and evolution.

29 **Introduction**

30 SARS-CoV-2, the virus responsible for COVID-19¹, is a betacoronavirus in the subgenus Sarbecovirus,
31 which also includes SARS-CoV, responsible for the 2003 severe acute respiratory syndrome (SARS)
32 outbreak. Its large 29,903-nucleotide positive-strand RNA genome encodes ~30 known and
33 hypothetical mature proteins (**Fig. 1a, Fig. 2, Extended Data Fig. 1**). Despite SARS-CoV-2's extreme
34 medical importance, its gene content remains surprisingly unresolved, with several hypothetical open
35 reading frames (ORFs) whose function or even protein-coding status is unknown. Moreover, no
36 systematic resource exists for interpreting the functional impact of SARS-CoV-2 mutations and
37 prioritizing candidate drivers that may underlie phenotypic differences between strains.

38 A large open reading frame spans two thirds of the genome, and results in non-structural proteins nsp1-
39 nsp10 and nsp12-nsp16 when an internal programmed translational frameshift² occurs (ORF1ab), or
40 nsp1-11 otherwise (ORF1a) with translation terminating four codons past the frameshift site. ORF1ab
41 encodes Pol (polymerase, RNA-dependent replication), Hel (helicase), ExoN (exonuclease,
42 proofreading), 3CL-PRO (polyprotein cleavage), and other proteins involved in host-cell suppression,
43 immune suppression, and diverse viral functions (Supplementary Table S2).

44 The last third of the genome encodes named proteins S (Spike surface glycoprotein), composed of S1
45 (viral attachment to host-cell ACE2 receptor) and S2 (membrane fusion, viral entry), E (Envelope
46 protein), M (Membrane glycoprotein), and N (Nucleocapsid, RNA genome packaging), which are
47 present in all coronaviruses, and several unnamed proteins. Their host-cell translation requires
48 subgenomic RNAs of varying lengths, such that each functional ORF is first (or early) on its own
49 transcript³. These subgenomic RNAs result from synthesis of negative-sense intermediates by

50 transcription starting from the 3' end of the genomic RNA, extending to one of several internal
51 transcription-regulatory sequences (TRS), and looping to a common 5' leader sequence; the negative-
52 sense intermediates are then used as templates for synthesis of positive-sense subgenomic RNAs⁴.

53 The remaining unnamed ORFs are Sarbecovirus-specific and subject to disagreement on which
54 encode functional proteins (Supplementary Table S2). NCBI annotates SARS-CoV-2 (NC_045512.2)
55 with 3a, 6, 7a, 7b, 8, and 10. UniProt also annotates 9b and 9c (which they name 14), both overlapping
56 N (in an alternate frame). The paper introducing SARS-CoV-2 also shows 3b (which overlaps 3a in
57 SARS-CoV but is truncated in SARS-CoV-2, with several in-frame stop codons)¹. Other publications⁵⁻¹³
58 include different subsets, use different names, or propose additional ORFs (including 3c and 3d
59 overlapping 3a). NCBI annotates SARS-CoV (NC_004718.3) orthologs of 3a, 6, 7a, 7b, and 9b, but 8 is
60 split into 8a and 8b, 3b is included, and neither 9c nor 10 are included. (ORF nomenclature details in
61 **Supplementary Text S1.**)

62 High-throughput experiments provide some evidence on SARS-CoV-2 gene content, though they
63 sometimes disagree, cannot prove non-functionality of non-detected ORFs (as they only capture
64 specific conditions), and cannot distinguish incidental transcriptional/translational events from selected
65 function. Proteomics identified peptides for 1ab, S, 3a, M, 6, 7a, 8, N, and 9b, but not E, 3b, 7b, 9c, or
66 10^{14,15}. Direct-RNA sequencing found subgenomic RNAs for a different subset: S, 3a, E, M, 6, 7a, 7b, 8,
67 and N, but limited or no support for 3b, 3d, 3c, 9b, 9c, and 10¹⁵⁻¹⁸, with 3c, 7b¹⁹, and 9b possibly
68 translated by leaky ribosomal scanning from 3a, 7a, and N subgenomic RNAs, respectively. Ribosome
69 profiling predicted translation of 1ab, S, 3a, E, M, 6, 7a, 7b, 8, N, and 10, and eleven alternate-frame
70 ORFs including 3c and 9b, but not 3d, 3b, or 9c¹⁰.

71 Here, we use comparative genomics of 44 Sarbecovirus strains to resolve the SARS-CoV-2 protein-
72 coding gene set (**Fig. 1**), and to distinguish genetic variants more likely to have functional importance.
73 We select 44 closely-related and complete coronavirus genomes, generate whole-genome alignments,
74 evaluate protein-coding and nucleotide-level constraint, and annotate synonymously-constrained

75 codons. We show that five hypothetical ORFs are not functional proteins and confirm protein-coding
76 status for seven accessory ORFs, including novel alternate-frame ORF3c within 3a. We use protein-
77 level and nucleotide-level inter-strain constraint to analyze 1875 mutations from 2544 pandemic
78 isolates, show gene-level and codon-level agreement between within-strain and across-strain selective
79 pressures, reveal recent adaptive acceleration for N and surprising deceleration for S1 and nsp3, and
80 flag mutations disrupting evolutionarily-conserved positions that may represent novel adaptations to
81 human hosts, including Spike D614G.

82 **Results**

83 **Strain selection, alignment, constraint**

84 We selected and aligned 44 complete Sarbecovirus genomes (SARS-CoV-2, SARS-CoV, and 42 bat-
85 infecting strains, **Extended Data Fig. 2, Supplementary Table S1**) at evolutionary distances well-
86 suited for identifying protein-coding genes and non-coding purifying selection, spanning ~3
87 substitutions per 4-fold degenerate site on average (comparable to 29-mammals/12-flies projects^{20,21}),
88 and ranging from 1.2 (E) to 4.8 (nsp16) and higher (**Supplementary Table S2**). Betacoronaviruses
89 outside Sarbecovirus (including MERS-CoV) are too distant (eg. no detectable homology across ORFs
90 6-7a-7b-8), and SARS-CoV-2/SARS-CoV isolates are too proximal for reliable evolutionary signatures.

91 To distinguish regions evolving under protein-coding constraint, we used their codon substitution
92 patterns across Sarbecoviruses, quantified using codon-resolution PhyloCSF²² scores in all three
93 reading frames, and smoothed using a hidden Markov model to create genome browser tracks^{1,23,24}
94 (**Fig. 1b, Fig. 2**). We also computed gene-resolution PhyloCSF scores for each known protein and
95 hypothetical ORF, and generated CodAlignView²⁵ visualizations highlighting protein-coding vs. non-
96 coding features for manual exploration of their alignments in all reading frames (**Fig. 1c,**
97 **Supplementary Table S2**). These tools are widely-accepted standards for protein-coding gene
98 annotation and for distinguishing protein-coding vs. non-coding genes in human and other species²⁰⁻
99 ^{22,26-28}.

100 Beyond protein-coding constraint for amino-acid translation, we also evaluated nucleotide-level
101 overlapping constraint within protein-coding regions indicative of dual-coding regions, RNA structures,
102 RNA-binding protein sites, etc, using reduced synonymous-substitution rate estimated using FRESCO,
103 which we previously developed and applied to viruses²⁹ and human³⁰. We annotated 1394
104 synonymously-constrained codons (14% of 9744, FDR=0.125) and defined 92 synonymous-constraint
105 elements (SCEs) (covering 1555 codons), using 9-codon-resolution significantly-decreased
106 synonymous rate relative to gene average^{29,31}.

107 Coding constraint on non-overlapping genes

108 As expected, E, M, N, S2, nsp1-nsp10, and nsp12-nsp16 showed clear protein-coding constraint
109 (**Supplementary Table S2**), with a change in constrained reading frame at the known programmed
110 frameshift (**Fig. 2, Extended Data Fig. 1**). Beyond its first 9 codons that match Pol, the 13-codon
111 nsp11 showed no nucleotide changes in Sarbecovirus, but stop-codon gain/loss across
112 betacoronaviruses indicates it is not separately functional (**Supplementary Fig. S1**).

113 S1 shows extremely rapid nucleotide evolution (near-zero phyloP³² and phastCons³³) but strong
114 PhyloCSF scores, indicating unambiguous protein-coding evolution and highlighting the power of
115 PhyloCSF to recognize protein-coding evolutionary signatures despite rapid nucleotide evolution.

116 ORFs 3a, 7a, 7b, and 8 show clear positive PhyloCSF scores, indicating conserved protein-coding
117 regions functional at the amino acid level (**Fig. 2b**). The first half and last quarter of ORF6 show strong
118 PhyloCSF signal, indicating that it encodes a functional protein, despite a less-constrained intermediate
119 portion, and an overall near-zero average score per codon (-0.3, **Fig. 1c**).

120 ORF8 shows near-zero nucleotide-level conservation (phyloP/phasCons), lacks well-established
121 functions, and was split into 8a/8b in SARS-CoV, suggesting at first glance that it might be non-
122 functional. However, it shows strongly-positive protein-coding PhyloCSF score (4.61/codon), and long
123 stretches of strong protein-coding constraint, indicating unambiguous protein-coding function. Its high

124 nucleotide-level rate is inflated by past recombination, but remains high even using an ORF8-specific
125 phylogeny (**Supplementary Fig. S2**).

126 By contrast, ORF10 shows no protein-coding constraint anywhere along its length, contains in-frame
127 stop codons in all but four Sarbecoviruses truncating the last third of its already-short length (38 amino
128 acids), includes a frame-shifting deletion in one of those four strains, and shows near-perfect
129 nucleotide-level conservation (phyloP/phastCons) extending beyond the ORF on both sides, indicating
130 it is not protein-coding but instead has non-coding functions (**Fig. 2b, Extended Data Fig. 3a**). (This
131 region overlaps the 3'-UTR pseudoknot RNA structure³⁴ involved in RNA synthesis, providing a likely
132 explanation for its high nucleotide-level constraint). Moreover, ribosome footprints in the region occur in
133 an overlapping upstream ORF or in a truncated ORF rather than uniquely in ORF10, consistent with
134 incidental-initiation events rather than functional translation (**Extended Data Fig. 3b**), and previously-
135 used comparative evidence for protein-coding function ignored a frameshifting deletion and was
136 insufficiently-powered (**Extended Data Fig. 3c**).

137 N-overlapping ORF 9b is coding, 9c is not

138 Evolutionary evidence for/against overlapping ORFs is harder to resolve, as protein-coding signatures
139 in the primary reading frame heavily influence scores in alternate frames: they skew the signal as
140 protein-preserving mutations in one frame are typically protein-disruptive in the other, and they
141 compress the signal as there are fewer substitutions. However, their dual-coding nature leads to a
142 depletion of synonymous substitutions in the primary ORF localized over the overlapping segment,
143 resulting in a strong signal of overlapping-constraint²⁹⁻³¹, used next to investigate ORFs 9c and 9b
144 overlapping N.

145 The 73-amino-acid-long ORF9c/ORF14 shows no localized synonymous constraint in N (**Fig. 3**), calling
146 its protein-coding status into question. Moreover, its start codon is lost in one strain, most strains have
147 a three-codons-earlier stop (**Extended Data Fig. 4**), its start codon is 460 nucleotides after N's with 9
148 intervening AUG codons (thus unlikely to be translated via leaky ribosomal scanning), direct-RNA

149 sequencing found no ORF9c-specific subgenomic RNAs^{16–18} (and no TRS is appropriately positioned to
150 create one), shows no ribosome footprint¹⁰ or proteomics^{14,15} evidence, and many SARS-CoV-2
151 isolates³⁵ contain stop-introducing mutations⁷. We conclude ORF9c does not encode a functional
152 protein.

153 The 97-amino-acid-long ORF9b shows high amino-acid substitution rate in its central portion but
154 significant localized synonymous constraint in N for its start and end regions (**Fig. 3**), even relative to
155 the overall low synonymous rate of N, consistent with dual-coding functions. Moreover, its start and
156 stop codons are perfectly conserved and its 97 codons are stop-free in all Sarbecoviruses. Its Kozak
157 context is stronger than N's and perfectly-conserved and its start codon is only 10 nucleotides
158 downstream of N's, allowing it to be translated from N's subgenomic RNA via leaky scanning
159 (**Extended Data Fig. 5**). ORF9b's negative PhyloCSF score is consistent with dual-coding signal
160 biases. ORF9b also has proteomics support^{15,36,37} (including evidence of viral-RNA binding³⁸), and
161 alternate-frame translation support by ribosome profiling¹⁰. In SARS-CoV, ORF9b protein (and
162 antibodies to it) was detected in SARS patients^{39,40}, localized in mitochondria, and interfered with host
163 cell antiviral response when overexpressed⁴¹. We conclude ORF9b encodes a conserved functional
164 protein with rapidly-changing portions.

165 ORF3c is a novel functional protein

166 We next searched for additional protein-coding genes by computing PhyloCSF scores for all 67
167 hypothetical non-NCBI-annotated AUG-to-stop SARS-CoV-2 ORFs ≥ 25 codons long that are not
168 contained in a longer same-frame ORF (locally-maximal). None had positive PhyloCSF scores, but
169 some may be coding as overlapping-ORF scores are reduced by alternative-frame protein-coding
170 constraint, so we investigated near-zero top candidates for evidence of localized synonymous
171 constraint, start/stop-codon conservation, and absence of in-frame stops or frameshifting indels.

172 The highest-scoring candidate, which we call ORF3c, overlaps ORF3a near its start (**Fig. 4**), with 38 of

173 its 41 codons overlapping synonymous constraint elements in ORF3a, localized nearly-perfectly on the
174 dual-coding region. Despite the score biases of dual-coding regions, ORF3c has PhyloCSF score
175 closer to non-overlapping protein-coding ORFs than to hypothetical non-coding ORFs (**Fig. 1c**),
176 indicating Sarbecovirus selection for protein-coding function. Strikingly, ORF3c also has many
177 synonymous substitutions that are non-synonymous in ORF3a, indicating ORF3c may be an equally-
178 strong driver of constraint in the dual-coding region (both frames show similar scores in the dual-coding
179 region). ORF3c also has conserved start and stop codons except for near-cognate GUG start in one
180 strain and a one-codon extension in SARS-CoV-2 and RaTG13, with no in-frame stop codons or indels.
181 We conclude ORF3c encodes a functional, conserved protein.

182 Previous studies proposed four ORFs overlapping 3a^{6,8-13}: 3c (41 codons), 3d (57 codons), 3b (22
183 codons, a truncated ortholog of SARS-CoV ORF3b), and a subset of 3d (33 codons). ORF3c was
184 proposed using synonymous constraint across 6 closely-related strains⁸ and a broader set of
185 Sarbecoviruses⁹, although on its own such evidence could also stem from other overlapping functional
186 elements (and is abundant in SARS-CoV-2 even outside dual-coding regions), and using ribosome
187 footprinting¹⁰, although such signal can also result from incidental, non-functional translation (and the
188 other 8 such candidates lacked any conservation); it was predicted to contain a viroporin-like
189 transmembrane domain⁸ and to be translated via leaky scanning⁹. The other three ORF3a-overlapping
190 candidates are not conserved and show variable length, premature stop codons, and other evidence
191 indicating they are not protein-coding (**Extended Data Fig. 6, Extended Data Fig. 7, Supplementary**
192 **Text S2**).

193 We examined all next-best-scoring candidates, and expanded the search to include shorter ORFs,
194 near-cognate start codons, non-locally-maximal ORFs, and ORFs on the negative strand, but found no
195 other convincing candidates (**Supplementary Text S3, Supplementary Fig. S4**), concluding our
196 protein-coding gene catalog is complete.

197 [A new reference gene set for SARS-CoV-2](#)

198 Altogether, our revised reference gene set consists of 1a, 1ab, S, 3a, 3c, E, M, 6, 7a, 7b, 8, N, and 9b,
199 including novel ORF 3c and previously-ambiguous 9b, and excluding 3b, 3d, 9c, and 10. These genes
200 are unambiguously translated into conserved functional proteins across Sarbecoviruses, and our
201 decisions are supported by a wealth of experimental evidence^{10,14–18}, including subgenomic RNAs^{15–18}
202 (or leaky scanning), ribosome profiling¹⁰, and proteomics experiments^{14,15}(Supplementary Text S4).
203 This high-confidence reference gene set can form the basis for understanding viral biology and the
204 functional roles of pandemic mutations (**Supplementary Text S5**).

205 Sarbecovirus conservation informs SARS-CoV-2 variant impact

206 We next used the evolutionary history of each codon across Sarbecoviruses to annotate 1875 single-
207 nucleotide variants (SNVs) across 2544 SARS-CoV-2 isolates sequenced during the current COVID-19
208 pandemic, including 1142 amino-acid-changing (missense), 628 amino-acid-preserving (synonymous),
209 and 104 non-coding substitutions (**Supplementary Table S3**).

210 We classified all amino acid positions as “conserved” (no change in any of the 44 Sarbecovirus
211 genomes) or “non-conserved/changed” (at least one change) for each of the mature proteins and
212 hypothetical ORFs (**Supplementary Table S2**), a definition independent of the phylogenetic tree, and
213 thus resilient to recombination events common in coronavirus phylogenies⁴².

214 Within-strain vs cross-strains evolution

215 The fraction of changed amino acids varied greatly across ORFs (17%-80%, **Fig. 5a, x-axis**), indicating
216 dramatically different evolutionary pressures. Unnamed accessory ORFs had more changed amino
217 acids (average 57%) than named and well-characterized ORFs (average 28%). ORF1ab mature
218 proteins varied from 57% changed (nsp2) to <17% (3CL-PRO, Pol, Hel, ExoN, nsp7-10) and Spike
219 subunits from 61% changed (S1) to 25% (S2).

220 Faster-evolving proteins across Sarbecoviruses showed more amino-acid-changing mutations within
221 SARS-CoV-2 (Spearman correlation 0.70), indicating Sarbecovirus evolutionary pressures still apply

222 during the current pandemic (**Fig. 5a**). This inter-vs-within-strain agreement also held at codon
223 resolution, with amino-acid-changing mutations preferentially disrupting non-conserved residues (535
224 mutations in 3264 positions, 16.4%) vs. conserved residues (607 in 6480, 9.4%, $p < 10^{-10}$) (**Extended**
225 **Data Fig. 9a**).

226 Accelerated and decelerated evolution

227 Notable deviations from this general agreement may reflect recent accelerated/decelerated evolution.
228 S1 showed significantly-fewer mutations than expected from its extremely-high inter-strain rate (13%
229 amino-acid-changing mutations observed vs. 17% expected, nominal $p = 0.0017$, depletion: 28);
230 additional SNVs (N=2696, May 9, 2020) further strengthened the statistical significance of this result
231 ($p = 0.00033$). Nsp3 also showed significantly fewer mutations than expected (10% vs. 15%, nominal
232 $p < 10^{-9}$, depletion: 90) and Nucleocapsid significantly more (21% vs. 11%, nominal $p < 10^{-8}$, excess: 42).

233 The lower-than-expected number of mutations in S1/nsp3 might indicate recent mutation-rate or
234 selective-pressure changes, possibly stemming from different phases of host-adaptive evolution, with
235 pre-pandemic earlier-adapting S1/nsp3 (eg. via non-human-host transmission or undetected human
236 transmission) requiring fewer pandemic-phase human-adaptive mutations than other later-adapting
237 genes (noting that only a subset of mutations are adaptive). Alternatively, S1/nsp3 may have more
238 positions in which deleterious mutations would be strongly-deleterious (purified-out even in shorter
239 timescales) vs. mildly-deleterious (purified-out only over larger timescales). Lastly, frequent S1
240 recombination could inflate inter-strain rate estimates, but probably insufficiently to account for the
241 observed discrepancies. (**Supplementary Text S6**).

242 The higher-than-expected number of variants in N might be explained by positive selection for host
243 adaptation. We investigated whether such positively-selected variation might be clustered in specific
244 segments, and searched the entire genome for clusters of variants disrupting conserved amino acid
245 residues. We found no significantly-depleted regions and only one region significantly-enriched
246 (**Supplementary Text S7**) relative to gene-specific variant density ($p < 0.012$ after conservative

247 genome-wide multiple-hypothesis correction), which was indeed localized in N, and contained 14
248 variants disrupting conserved residues (out of the observed excess of 29 such variants in N)
249 concentrated in 20-amino-acid region R185-G204 (noting this enrichment is relative to the already-high
250 enrichment of such variants in N). This region overlaps a predicted B-Cell epitope⁴³, suggesting positive
251 selection for immune system avoidance (**Fig. 5b, Extended Data Fig. 9c**).

252 Spike SNV prioritization

253 We next investigated whether we can help prioritize candidate driver SNVs associated with phenotypic
254 differences between SARS-CoV-2 strains, using the evolutionary history of each amino acid across
255 Sarbecoviruses to provide position-specific estimates of evolutionary constraint, thus taking into
256 account the biological context and precise functions that each amino acid plays in coronavirus biology
257 (beyond position-independent general estimates from general amino acid properties).

258 As proof-of-principle, we focused on 16 amino-acid-changing variants in Spike with high frequency
259 and/or epitope proximity^{44,45} (**Supplementary Table S3**). Among them, radical-amino-acid-change
260 D614G, which rose in frequency across multiple cities and increases infectivity *in vitro*⁴⁵⁻⁴⁷, disrupts a
261 perfectly-conserved residue (across Sarbecoviruses), and lies in a stretch of 11 perfectly-conserved
262 amino acids (**Fig. 5c**), indicating its disruption is deleterious in bat-host contexts, and likely represents a
263 novel human-host adaptation.

264 Of the other 15 Spike variants, two are in perfectly-conserved residues (V615I/F, P1263L) and two in
265 mostly-conserved residues in highly-conserved regions (A831V, A829T/S), indicating likely-functional
266 changes. Another three are in moderately-conserved contexts (V367F, D839Y/N/E, D936Y/H) less
267 likely to be functional, and eight lie in repeatedly-altered amino acids in poorly-conserved regions and
268 likely-neutral.

269 Lastly, Sarbecovirus evolutionary context helps prioritize likely drivers among co-inherited mutations.
270 Spike D614G was nearly always co-inherited with Pol P4715L (also radical and altering a perfectly-

271 conserved residue in a highly-conserved context, but potentially-deleterious given Pol's slow evolution
272 and less-likely-to-be-adaptive function), nsp3 nucleotide change C3037T (repeatedly-observed
273 synonymous change, outside synonymously-constrained elements, likely-neutral), and nucleotide
274 change C241T (perfectly-conserved, non-coding, in a loop of six unpaired bases in the conserved 5'-
275 UTR SL5B secondary structure³⁴ 25 nucleotides upstream of ORF1ab).

276 Synonymous and non-coding substitutions

277 Even for synonymous SNVs we found agreement between cross-strain and within-strain constraint,
278 with synonymously-constrained codons showing fewer synonymous variants (73 of 1394, 5.2%) than
279 non-synonymously-constrained codons (555 of 8350 positions, 6.6%, binomial $p=0.029$, **Extended**
280 **Data Fig. 9b**).

281 We also classified 643 intergenic and 5'/3'-UTR positions as "conserved" (N=432, 67%) or "non-
282 conserved" (**Supplementary Table S3**), and found a surprising (but non-significant) SNV excess in
283 conserved positions (17.4% vs. 13.7%, $p=0.17$).

284 Discussion

285 We used comparative genomics to determine the conserved functional protein-coding genes of SARS-
286 CoV-2, resulting in a new high-confidence evolutionarily- and experimentally-supported reference gene
287 set, including ORFs 1a, 1ab, S, 3a, 3c, E, M, 6, 7a, 7b, 8, N, and 9b, but excluding 3b, 3d, 9c, and 10.

288 We show that novel ORF 3c is functional and conserved, and that no other conserved genes remain to
289 be discovered.

290 Our comparative genomics evidence complements experimental approaches by providing a
291 comprehensive function-centric view of protein constraint, summed over all environmental conditions
292 and hosts spanned by the strains compared here, while experimental methods only profile a single
293 environmental and host condition in each experiment. Moreover, while experimental methods can suffer
294 from incidental transcriptional or translational events, evolutionary signatures specifically measure

295 functional constraint for a given function. While in principle our methods may miss recently-evolved
296 genes that only function in a subset of strains, we found that our Sarbecovirus cross-strain evolutionary
297 evidence agreed with SARS-CoV-2/SARS-CoV within-strain experimental evidence, suggesting it is
298 unlikely that we may have missed newly-evolved genes.

299 It is important to note that comparative genomics methods that focus on nucleotide-level constraint
300 such as phyloP and phastCons, as valuable as they are, would have mistakenly rejected S1 and ORF8
301 as seemingly non-conserved (given their extremely-rapid evolutionary rate and recombination history),
302 and conversely included ORF10 as seemingly-conserved (given high nucleotide-level conservation in
303 the overlapping RNA structure). Instead, our methods were able to correctly distinguish the protein-
304 coding status of these genes because they use protein-coding evolutionary signatures that: (a) focus on
305 the patterns of change characteristic of protein-coding constraint (specific codon substitution
306 frequencies and reading frame conservation) rather than the overall number of substitutions; and (b)
307 are less sensitive to the specific phylogenetic tree relating the genomes compared, and thus resilient to
308 the frequent recombination events that characterize coronavirus genomes.

309 We found that both protein-coding and non-coding constraint agree between cross-strain Sarbecovirus
310 substitutions and within-strain SARS-CoV-2 mutations, enabling us to classify SARS-CoV-2 variants
311 into likely-functional vs. likely-neutral according to their evolutionary constraint. This revealed that the
312 Spike D614G substitution likely represents a new adaptation to human hosts, as it disrupts a
313 Sarbecovirus-conserved residue in a strongly-conserved region of S1, and to interpret the likely
314 functional impact of genetic variants co-inherited with D614G based on their evolutionary history.
315 Beyond the specific examples cited here, our annotations are broadly useful for interpreting SARS-
316 CoV-2 variants and inferring causal relationships between viral mutations and disease phenotype. For
317 interpreting future variants, we also created a genome browser track hub to facilitate SARS-CoV-2
318 variant interpretation based on their evolutionary context, and based on our revised gene annotations.
319 We found three notable exceptions to the otherwise-strong agreement between inter-strain and within-

320 strain variation: N showed significantly more amino-acid-changing mutations than expected, and nsp3
321 and S1 showed significantly fewer. For N, the acceleration is consistent with positive selection for
322 human-host adaptation across many variants, including a 20-amino-acid region enriched for conserved-
323 residue-disrupting variants in a predicted B-cell epitope. For nsp3 and S1, the deviation raises the
324 possibility they may represent pioneer proteins that adapt to new-host transmission prior to its
325 pandemic phase, then require fewer mutations while other proteins ‘catch up’, an observation that may
326 be more generally true across different proteins showing acceleration/deceleration in different phases
327 of host adaptation and pandemic spread. Another possibility is that the space of deleteriousness across
328 all possible mutations is differently-distributed for nsp3 and S1 compared to other proteins, with more
329 deleterious mutations in the strongly-deleterious end of the distribution, thus explaining the discrepancy
330 in the number of observed amino-acid-changing substitutions between the short timescales captured in
331 the recent pandemic SNVs vs. the longer timescales captured in cross-Sarbecoviruses comparative
332 genomics. We discuss these and other possibilities in **Supplementary Text S6**.

333 Overall, our new reference gene set provides a solid foundation for systematically dissecting the
334 function of SARS-CoV-2 proteins, and focusing experimental work on high-confidence uncharacterized
335 ORFs, which can be guided in part by their evolutionary dynamics (such as the rapid evolution and
336 recombination history of ORF6 and ORF8, indicating possible adaptive roles). In addition, our gene-
337 level, codon-level, and nucleotide-level Sarbecovirus constraint, and the classification of all existing and
338 potential SNVs into likely-functional vs. likely-neutral based on their evolutionary history, provide
339 important foundations for elucidating SARS-CoV-2 biology, understanding its evolutionary dynamics,
340 prioritizing candidate driver mutations among co-inherited mutations, and prioritizing candidate regions
341 for vaccine design and refinement.

342 **References**

- 343 1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**,
344 265–269 (2020).

- 345 2. Baranov, P. V. *et al.* Programmed ribosomal frameshifting in decoding the SARS-CoV genome.
346 *Virology* **332**, 498–510 (2005).
- 347 3. Miller, W. A. & Koev, G. Synthesis of subgenomic RNAs by positive-strand RNA viruses. *Virology*
348 **273**, 1–8 (2000).
- 349 4. Sawicki, S. G., Sawicki, D. L. & Siddell, S. G. A contemporary view of coronavirus transcription. *J.*
350 *Viol.* **81**, 20–29 (2007).
- 351 5. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for
352 virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
- 353 6. Chan, J. F.-W. *et al.* Genomic characterization of the 2019 novel human-pathogenic coronavirus
354 isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes &*
355 *Infections* vol. 9 221–236 (2020).
- 356 7. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.
357 *Nature* (2020) doi:10.1038/s41586-020-2286-9.
- 358 8. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Coding potential and sequence conservation of
359 SARS-CoV-2 and related animal viruses. *Infection, Genetics and Evolution* vol. 83 104353 (2020).
- 360 9. Firth, A. E. A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. *Journal*
361 *of General Virology* (2020) doi:10.1099/jgv.0.001469.
- 362 10. Finkel, Y., Mizrahi, O. & Nachshon, A. The coding capacity of SARS-CoV-2. *bioRxiv* (2020).
- 363 11. Nelson, C. W., Ardern, Z., Goldberg, T. L., Meng, C. & Kuo, C. H. A previously uncharacterized
364 gene in SARS-CoV-2 illuminates the functional dynamics and evolutionary origins of the COVID-19
365 pandemic. *bioRxiv* (2020).
- 366 12. Pavesi, A. New insights into the evolutionary features of viral overlapping genes by discriminant
367 analysis. *Virology* **546**, 51–66 (2020).
- 368 13. Konno, Y., Kimura, I., Uriu, K., Fukushi, M. & Irie, T. SARS-CoV-2 ORF3b is a potent interferon
369 antagonist whose activity is further increased by a naturally occurring elongation variant. *bioRxiv*
370 (2020).

- 371 14. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*
372 **583**, 469–472 (2020).
- 373 15. Davidson, A. D. *et al.* Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals
374 a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein.
375 *Genome Med.* **12**, 68 (2020).
- 376 16. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
- 377 17. Taiaroa, G. *et al.* Direct RNA sequencing and early evolution of SARS-CoV-2.
378 doi:10.1101/2020.03.05.976167.
- 379 18. Nomburg, J., Meyerson, M. & DeCaprio, J. A. Pervasive generation of non-canonical subgenomic
380 RNAs by SARS-CoV-2. 2020.04.28.066951 (2020) doi:10.1101/2020.04.28.066951.
- 381 19. Schaecher, S. R., Mackenzie, J. M. & Pekosz, A. The ORF7b protein of severe acute respiratory
382 syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into
383 SARS-CoV particles. *J. Virol.* **81**, 718–731 (2007).
- 384 20. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals.
385 *Nature* **478**, 476–482 (2011).
- 386 21. Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly
387 genomes. *Genome Res.* **17**, 1823–1836 (2007).
- 388 22. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish
389 protein coding and non-coding regions. *Bioinformatics* **27**, i275–82 (2011).
- 390 23. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- 391 24. Raney, B. J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations
392 on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014).
- 393 25. I Jungreis, MF Lin, CS Chan, M Kellis. CodAlignView. *CodAlignView: The Codon Alignment Viewer*
394 <https://data.broadinstitute.org/compbio1/cav.php> (2016).
- 395 26. Mudge, J. M. *et al.* Discovery of high-confidence human protein-coding genes and exons by whole-
396 genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.* **29**, 2073–2087 (2019).

- 397 27. McCorkindale, A. L. *et al.* A gene expression atlas of embryonic neurogenesis in *Drosophila*
398 reveals complex spatiotemporal regulation of lncRNAs. *Development* **146**, (2019).
- 399 28. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic*
400 *Acids Res.* **47**, D766–D773 (2019).
- 401 29. Sealfon, R. S. *et al.* FRESCo: finding regions of excess synonymous constraint in diverse viruses.
402 *Genome Biol.* **16**, 38 (2015).
- 403 30. Khan, Y. A. *et al.* Evidence for a novel overlapping coding sequence in POLG initiated at a CUG
404 start codon. *BMC Genet.* **21**, 25 (2020).
- 405 31. Lin, M. F. *et al.* Locating protein-coding sequences under selection for additional, overlapping
406 functions in 29 mammalian genomes. *Genome Research* vol. 21 1916–1928 (2011).
- 407 32. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution
408 rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
- 409 33. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.
410 *Genome Res.* **15**, 1034–1050 (2005).
- 411 34. Rangan, R., Zheludev, I. N. & Das, R. RNA genome conservation and secondary structure in
412 SARS-CoV-2 and SARS-related viruses: a first look. *RNA* (2020) doi:10.1261/rna.076141.120.
- 413 35. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID’s innovative contribution to
414 global health. *Glob Chall* **1**, 33–46 (2017).
- 415 36. Bojkova, D. *et al.* SARS-CoV-2 infected host cell proteomics reveal potential therapy targets.
416 *Preprint available at Research Square* (2020).
- 417 37. Bezstarosti, K., Lamers, M. M., Haagmans, B. L. & Demmers, J. A. A. Targeted Proteomics for the
418 Detection of SARS-CoV-2 Proteins. *bioRxiv* 2020.04.23.057810 (2020)
419 doi:10.1101/2020.04.23.057810.
- 420 38. Schmidt, N. *et al.* A direct RNA-protein interaction atlas of the SARS-CoV-2 RNA in infected
421 human cells. *bioRxiv* 2020.07.15.204404 (2020) doi:10.1101/2020.07.15.204404.
- 422 39. Chan, W. S. *et al.* Coronaviral hypothetical and structural proteins were found in the intestinal

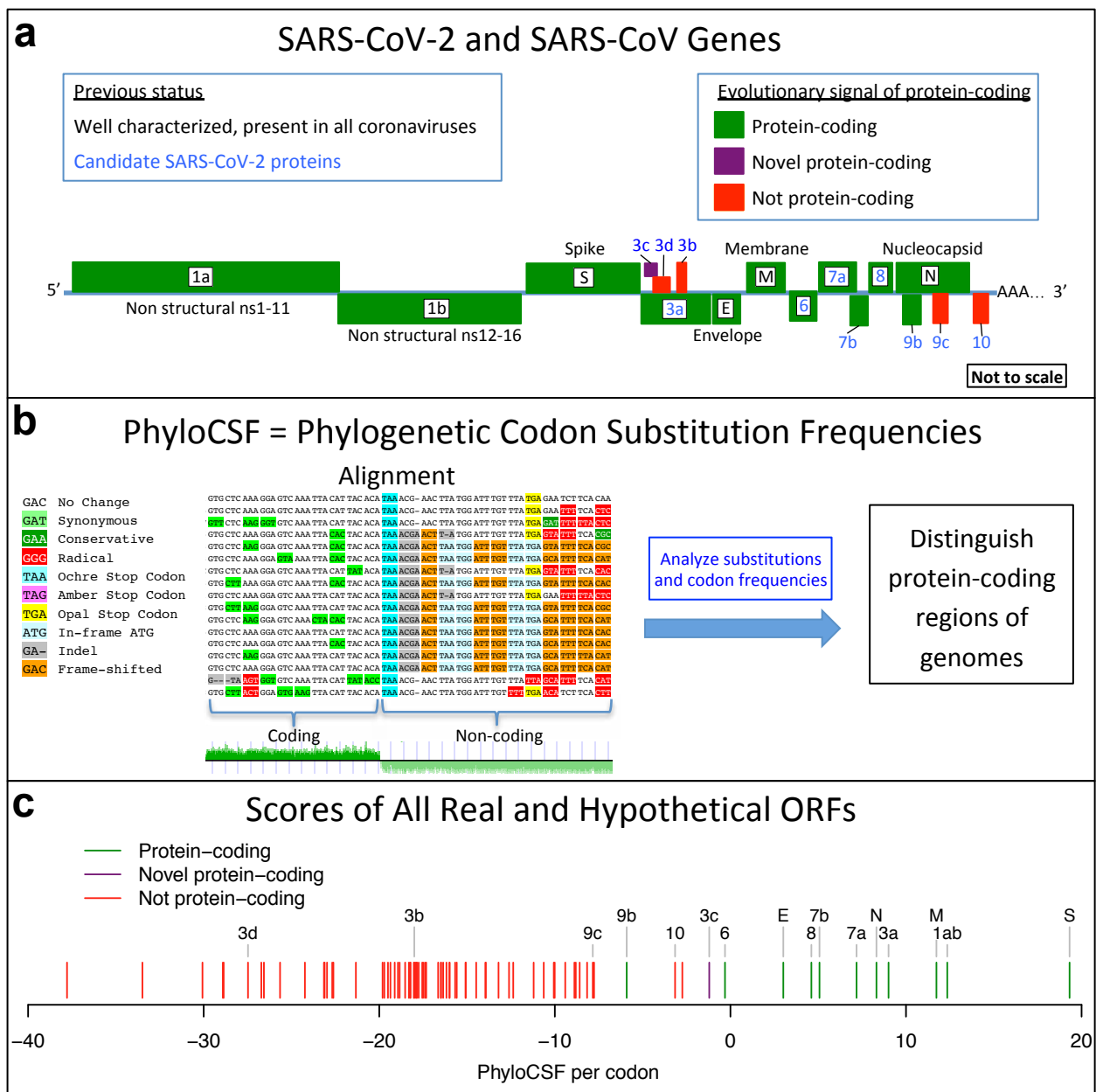
- 423 surface enterocytes and pneumocytes of severe acute respiratory syndrome (SARS). *Mod. Pathol.*
424 **18**, 1432–1439 (2005).
- 425 40. Qiu, M. *et al.* Antibody responses to individual proteins of SARS coronavirus and their
426 neutralization activities. *Microbes Infect.* **7**, 882–889 (2005).
- 427 41. Shi, C.-S. *et al.* SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting
428 mitochondria and the MAVS/TRAF3/TRAF6 signalosome. *J. Immunol.* **193**, 3080–3089 (2014).
- 429 42. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection.
430 *Science Advances* eabb9153 (2020).
- 431 43. Grifoni, A. *et al.* A Sequence Homology and Bioinformatic Approach Can Predict Candidate
432 Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* **27**, 671–680.e2 (2020).
- 433 44. Korber, B., Fischer, W., Gnanakaran, S. G. & Yoon, H. Spike mutation pipeline reveals the
434 emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* (2020).
- 435 45. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases
436 infectivity of the COVID-19 virus. *Cell* (2020) doi:10.1016/j.cell.2020.06.043.
- 437 46. Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and
438 increases infectivity. *bioRxiv* 2020.06.12.148726 (2020) doi:10.1101/2020.06.12.148726.
- 439 47. Yurkovetskiy, L. *et al.* SARS-CoV-2 Spike protein variant D614G increases infectivity and retains
440 sensitivity to antibodies that target the receptor binding domain. *bioRxiv* (2020)
441 doi:10.1101/2020.07.04.187757.
- 442 48. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**,
443 D853–D858 (2019).
- 444 49. Lab, Z. NW-align. *NW-align* <http://zhanglab.ccmb.med.umich.edu/NW-align> (2-Apr-2012).
- 445 50. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein
446 sequences. *Protein Sci.* **27**, 135–145 (2018).
- 447 51. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
448 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

- 449 52. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner.
450 *Genome Res.* **14**, 708–715 (2004).
- 451 53. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology
452 and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
- 453 54. Fernandes, J. D. *et al.* The UCSC SARS-CoV-2 Genome Browser. 2020.05.04.075945 (2020)
454 doi:10.1101/2020.05.04.075945.
- 455 55. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–
456 4123 (2018).
- 457 56. Dinan, A. M., Lukhovitskaya, N. I., Olendraite, I. & Firth, A. E. A case for a negative-strand coding
458 sequence in a group of positive-sense RNA viruses. *Virus Evol* **6**, veaa007 (2020).
- 459 57. DeRisi, J. L. *et al.* An exploration of ambigrammatic sequences in narnaviruses. *Sci. Rep.* **9**, 17982
460 (2019).
- 461

462

Figures:

463



464

465

466

467

468

469

470

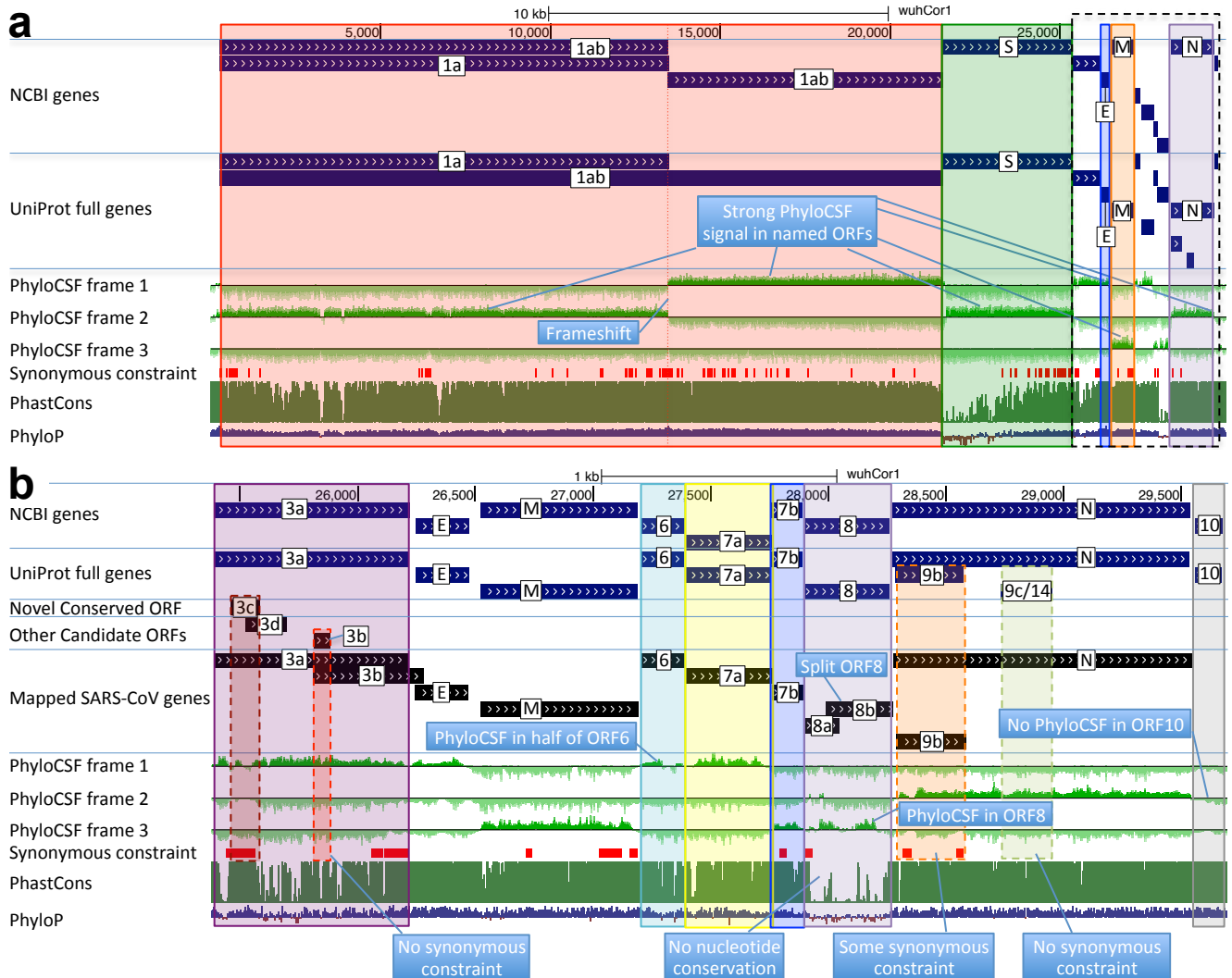
471

472

Figure 1. Overview. **a.** Previously annotated named (black font) and unnamed or proposed (blue font) SARS-CoV-2 genes, with confirmed protein-coding (green), rejected (red), or novel protein-coding (purple) classification, using evolutionary and experimental evidence. **b.** Phylogenetic Codon Substitution Frequencies (PhyloCSF) scores distinguish protein-coding (left) vs. non-coding (right) using evolutionary signatures, including distinct frequencies of amino-acid-preserving (green) vs. amino-acid-disruptive (red) substitutions, and stop codons (cyan/magenta/yellow) in frame-specific alignments, and additional features. **c.** PhyloCSF score (x-axis) for all confirmed (green) and rejected (red) ORFs, showing annotated/hypothetical/novel (labeled) and all AUG-initiated ≥ 25 -codons-long locally-maximal ORFs (unlabelled). Novel ORF3c (purple) clusters with protein-coding. Only-modestly-negative ORF9c/ORF10 scores are artifacts of score

473 compression in high-nucleotide-constraint regions, and substantially drop when nucleotide-conservation-scaled (see
474 **Extended Data Fig. 8**).

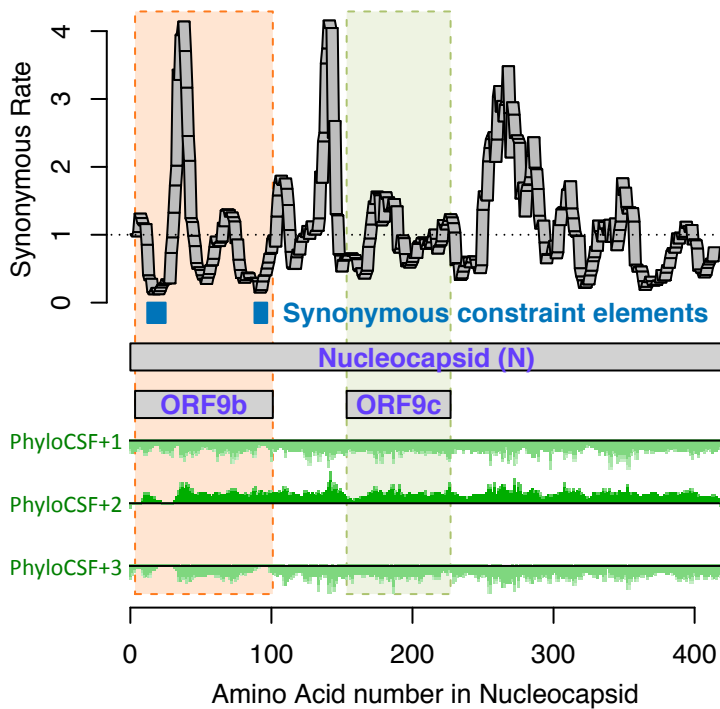
475



476

477 **Figure 2. Genome-wide protein-coding signatures.** SARS-CoV-2 NCBI/UniProt genes (blue), unannotated proposed
 478 genes and mapped SARS-CoV genes (black, panel b only), frame-specific protein-coding PhyloCSF scores (green),
 479 Synonymous Constraint Elements (SCEs) (red), and phastCons/phyloP nucleotide-level constraint (green/blue/red)
 480 across genomic coordinates (x-axis) for entire genome (panel a) and final 4-kb subset (panel b, dashed black box),
 481 highlighting (light blue boxes): **(a)** strong protein-coding signal in correct frame for each named gene; conservation-signal
 482 frame-change at programmed frameshift site; strong protein-coding signal throughout S despite lack of nucleotide
 483 conservation in S1; **(b)** unambiguous and frame-specific protein-coding signal for unnamed ORFs 3a (despite only partial
 484 nucleotide conservation), 7a, 7b, and 8 (despite lack of nucleotide conservation); clear protein-coding signal in first half
 485 and last quarter of ORF6; no protein-coding signal for 10 (despite high nucleotide conservation); synonymous constraint
 486 (red) in novel-ORF 3c and confirmed-ORF 9b; no synonymous constraint in rejected ORFs 9c, 3b, 3d.

487



488

489

Figure 3. Synonymous constraint in Nucleocapsid overlaps 9b but not 9c/14. Synonymous substitution rate in 9-

490

codon windows (y-axis) across N (x-axis), normalized to gene-wide average (dotted black line). Synonymous constraint

491

elements (blue) expected for dual-coding constraint localize in overlapping ORF9b (dashed orange rectangle) indicating it

492

is protein-coding, but not 9c (dashed purple rectangle) indicating it is not protein-coding. PhyloCSF protein-coding signal

493

(green) in frame3 (encoding 9b and 9c/14) remains strongly negative throughout the length of 9c/14 (green box),

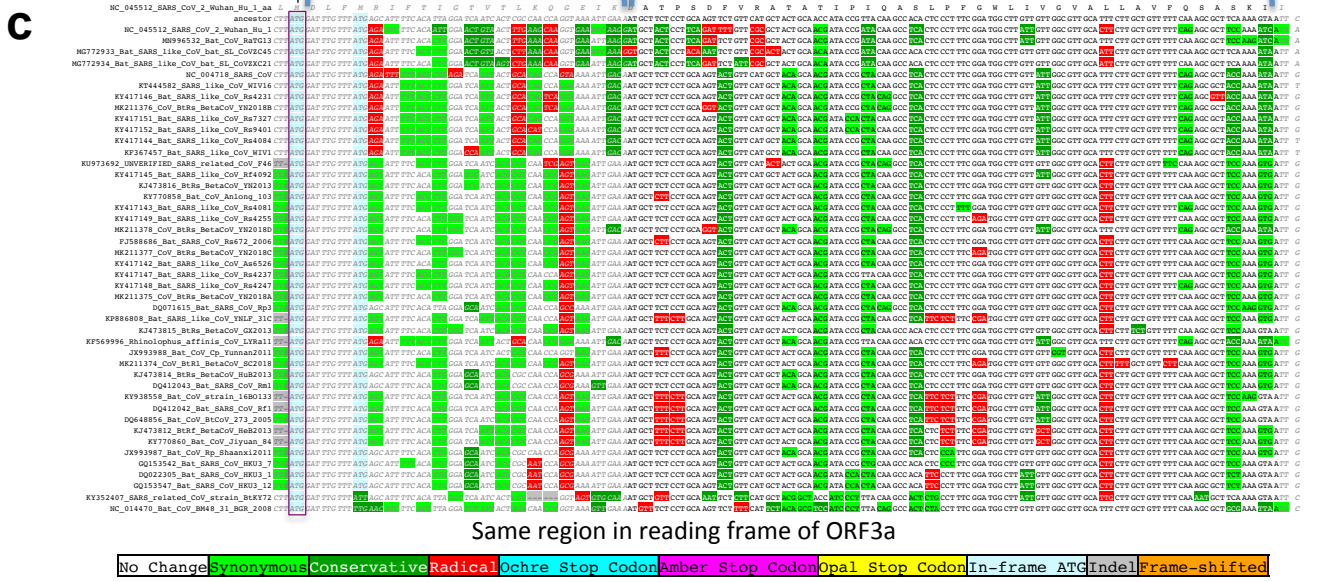
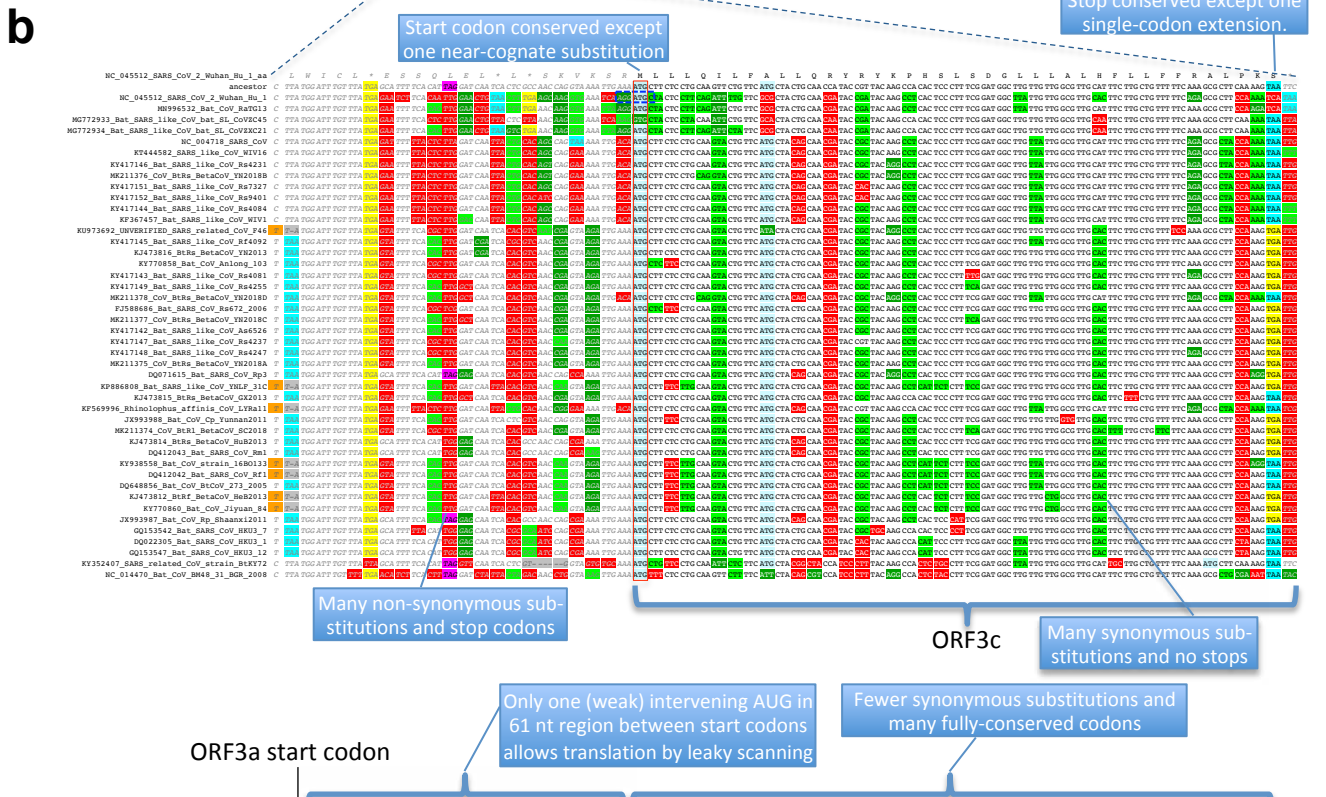
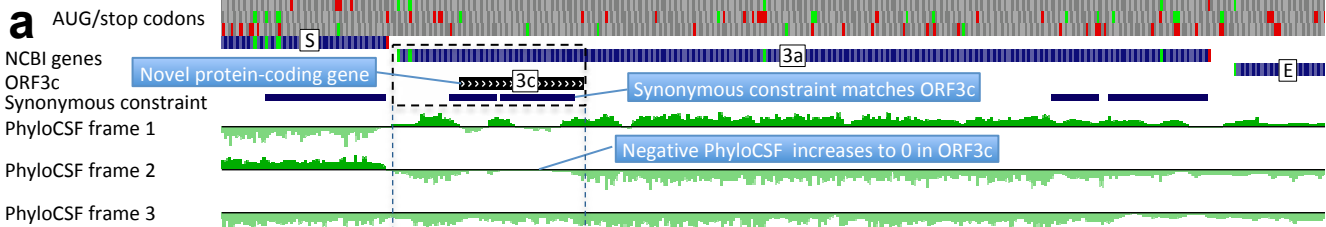
494

indicating 9c/14 is non-coding, but rises to near-zero values for two regions of 9b, indicating protein-coding selection,

495

while PhyloCSF signal frame 2 (encoding N) remains consistently high throughout the length of ORF9c.

496



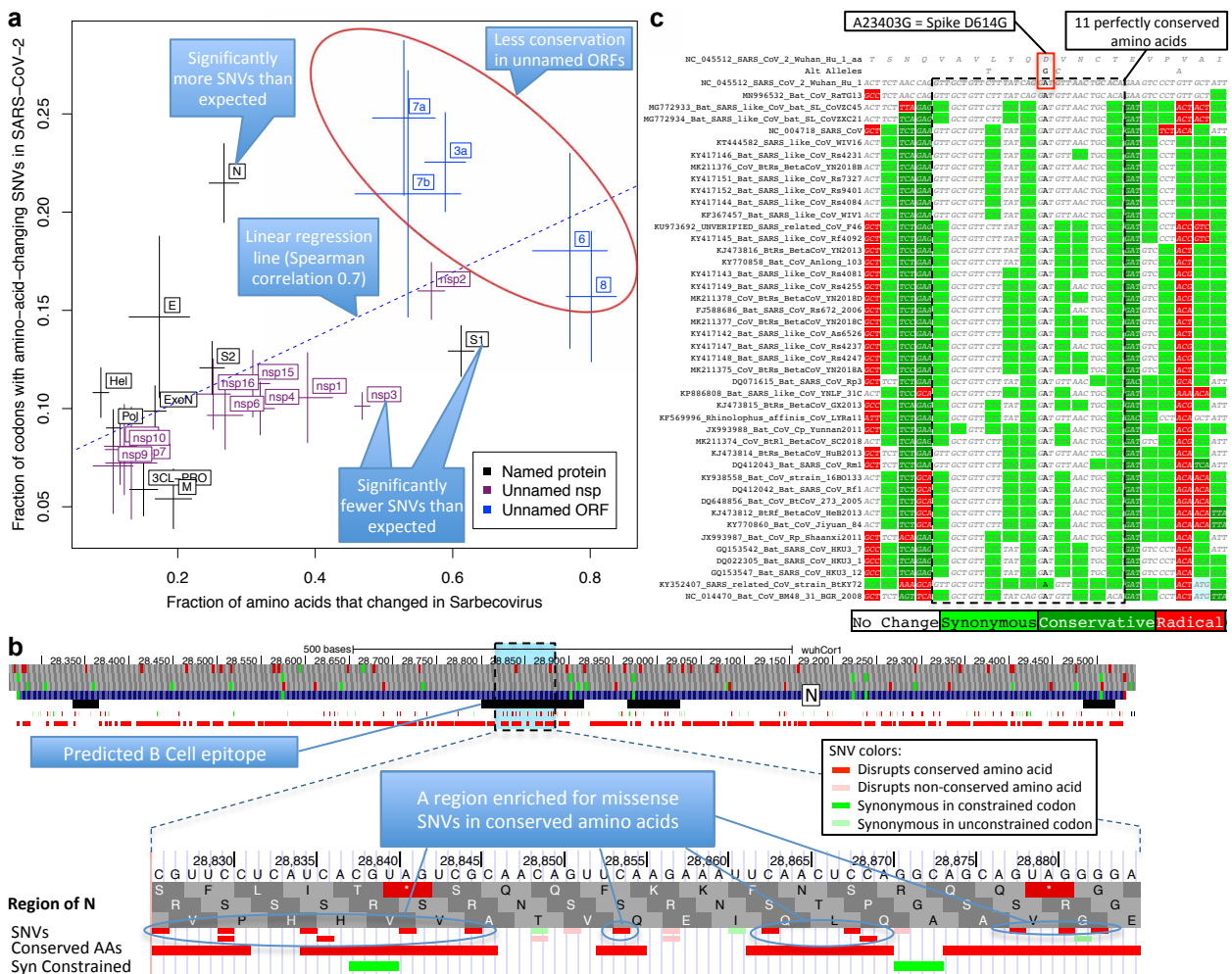
497

498

Figure 4. Novel gene 3c overlapping 3a is protein-coding. a. Synonymous-constraint elements (blue) match nearly-

499 perfectly 41-codon ORF_c dual-coding region boundaries (black), and protein-coding evolutionary signatures (green)
500 switch between frame 1 and 2 (rows) in the dual-coding region, with frame-2 signal (negative flanking ORF_{3c}) increasing
501 to near-zero, and frame-1 signal (high flanking the dual-coding region) dropping to near-zero. **b,c.** Codon-resolution
502 evolutionary signatures (colors, CodAlignView²⁵) annotating genomic alignment (letters) spanning ORF_{3a} start and dual-
503 coding region, in frame-1 (top) and frame-2 (bottom), highlighting (blue boxes): (**b**, frame-2, ORF_{3c}) radical codon
504 substitutions (red) and stop codons (yellow, magenta, cyan) prior to ORF_{3c} start; synonymous (light green) and
505 conservative (dark green) substitutions in ORF_{3c}; ORF_{3c}'s start codon is conserved, except in one strain (row 4) with
506 near-cognate GUG; ORF_{3c}'s stop codon is conserved except for one-codon extension in two strains (rows 2-3); no
507 intermediate stop codons in ORF_{3c}; (**c**, frame-1, ORF_{3a}) abundant synonymous and conservative substitutions in 3a prior
508 to dual-coding region; increase in fully-conserved codons (white) over dual-coding region. Short interval (61nt) with only
509 one weak-Kozak-context intervening start codon indicates ORF_{3c} may be translated from ORF_{3a}'s subgenomic RNA via
510 leaky scanning.

511



512

513

514

515

516

517

518

519

520

521

522

523

524

Figure 5. Within-strain variation vs. inter-strain divergence. a. Gene-level comparison. Long-term inter-strain evolutionary divergence (x-axis) and short-term within-strain variation (y-axis) show strong agreement (linear regression dotted line, Spearman-correlation=0.70) across mature proteins (crosses, denoting standard error of mean on each axis), indicating that Sarbecovirus-clade selective pressures persist in the current pandemic. Well-characterized genes (black) show fewer changes in both timescales (bottom left) and less-well-characterized ORFs (blue) show more in both (top right). Significantly-deviating exceptions are: nsp3 and S1 (bottom right) showing significantly-fewer amino-acid-changing SNVs than expected from their cross-Sarbecovirus rapid evolution, and N (top left), showing significantly-more, possibly due to accelerated evolution in the current pandemic. **b. Rapidly-evolving Nucleocapsid region.** Top: Nucleocapsid context showing B-cell epitope predictions (black, "IEDB Predictions" track), and our annotation track-hub showing: conserved amino acids (red blocks), synonymously-constrained codons (green blocks), and SNV classification (colored tick-marks) as conserved/non-conserved (dark/light) and missense/synonymous (red/green); top 3 tracks show AUG

525 codons (green) and stop codons (red) in three frames. Bottom: Focus on 20-amino-acid region R185-G204 (dotted box) in
526 predicted B-cell epitope (black) significantly-enriched for amino-acid-changing variants (red) disrupting perfectly-
527 conserved residues, indicative of positive selection in SARS-CoV-2 for immune system avoidance. **c. Spike D614G**
528 **evolutionary context.** Sarbecovirus alignment (text) surrounding Spike D614G amino-acid-changing SNV, which rose in
529 frequency in multiple geographic locations suggesting increased transmissibility. This A-to-G SNV disrupts a perfectly-
530 conserved nucleotide (bold font, A-to-G), which disrupts a perfectly-conserved amino-acid (red box, D-to-G), in a
531 perfectly-conserved 11-amino-acid region (dotted black box, light-green=synonymous-substitutions) across bat-host
532 Sarbecoviruses, indicating D614G represents a human-host-adaptive mutation.

533

534

535

536

537 **Methods**

538 **Genomes and Alignments**

539 Genome sequences were obtained from <https://www.ncbi.nlm.nih.gov/>. The genomes and NCBI
540 annotations for SARS-CoV-2 and SARS-CoV were obtained from the records for accessions
541 NC_045512.2 and NC_004718.3, respectively. The UniProt annotations for SARS-CoV-2 were
542 obtained from the UCSC Genome Browser ⁴⁸ on April 5, 2020.

543 The 44 Sarbecovirus genomes used in this study were selected starting from all betacoronavirus and
544 unclassified coronavirus full genomes listed on ncbi via searches
545 [https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694002\[Organism:exp\]](https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694002[Organism:exp]) and the same with txid1986197
546 and txid2664420 on 5-Mar-2020, excluding any that differed from NC_045512.2 in more than 10,000
547 positions in a pairwise alignment computed using NW-align ⁴⁹, that cutoff being chosen so as to
548 distinguish Sarbecovirus genomes among those that were classified, and removing near duplicates,
549 including all SARS-CoV and SARS-CoV-2 genomes other than the reference. Coronavirus genomes in
550 the left half of Extended Data Fig. 2 were those listed by
551 <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=11118> on 11-Feb-2020.

552 The genomes were aligned using clustalo⁵⁰ with the default parameters. The Phylogenetic tree was
553 calculated using RAxML ⁵¹ using the GTRCATX model.

554 **PhyloCSF, FRESCo, and other conservation metrics**

555 PhyloCSF (Phylogenetic Codon Substitution Frequencies)²² determines whether a given nucleotide
556 sequence is likely to represent a functional, conserved protein-coding sequence by determining the
557 likelihood ratio of its multi-species alignment under protein-coding and non-coding models of evolution
558 that use pre-computed substitution frequencies for every possible pair of codons, and codon
559 frequencies for every codon, trained on whole-genome data. PhyloCSF was run using the `29mammals`
560 empirical codon matrices but with the Sarbecovirus tree substituted for the mammals tree. Input
561 alignments were extracted from the whole-genome alignment and columns containing a gap in the

562 reference sequence were removed. Browser tracks were created as described previously²⁶. Scores
563 listed in Supplementary Table S2 were calculated on the local alignment for each ORF or mature
564 protein, excluding the final stop codon, using the default PhyloCSF parameters, including --
565 `strategy=mle`.

566 FRESCo²⁹ was run using HYPHY version 2.220180618beta(MP) for Linux on x86_64 on 9-codon
567 windows in each of the NCBI annotated ORFs. Alignments were extracted for the ORF excluding the
568 final stop codon, and gaps in the reference sequence were removed. SCEs were found by taking all
569 windows having synonymous rate less than 1 and nominal p-value<10⁻⁵, and combining overlapping or
570 adjacent windows. For the variant analysis, FRESCo was also run on 1-codon windows using codon
571 alignments as described previously²⁹.

572 Substitutions per site and per neutral site for each annotated ORF and mature protein were calculated
573 by extracting the alignment column for each site or, respectively, 4-fold degenerate site, from the
574 whole-genome alignment and determining the parsimonious number of substitutions using the whole-
575 genome phylogenetic tree. For columns in which some genomes did not have an aligned nucleotide,
576 the number of substitutions was scaled up by the branch length of the entire tree divided by the branch
577 length of the tree of genomes having an aligned nucleotide in that column.

578 PhastCons and phyloP tracks shown in Fig. 2 are the Comparative Genomics tracks from the UCSC
579 Genome Browser, which were constructed from a multiz⁵² alignment of the list of 44 Sarbecovirus
580 genomes that we supplied to UCSC.

581 Analysis of Single Nucleotide Variants

582 Single nucleotide variants were downloaded from the “Nextstrain Vars” track in the UCSC Table
583 Browser on 2020-04-18 at 11:46 AM EDT. Table S3 includes one additional mutation, G24047A, from a
584 later download, in order to represent Korber variant A829T/S. We defined an amino acid to be
585 “conserved” if there were no amino-acid-changing substitutions in the Sarbecovirus alignment of its

586 codon. We defined codons to be “synonymously constrained” if the synonymous rate at that codon
587 calculated by FRESCo using 1-codon windows was less than 1.0 with nominal p-value<0.034,
588 corresponding to a false discovery rate of 0.125. We defined an intergenic nucleotide to be “conserved”
589 if there were no substitutions of that nucleotide in the Sarbecovirus alignment. We classified SNVs as
590 Synonymous, Nonsynonymous, or Noncoding, relative to the NCBI annotations, so SNVs within ORF10
591 were classified as coding, and SNVs within overlapping ORFs 3c and 9b were classified relative to the
592 longer containing ORFs 3a and N, respectively. However, in Supplementary Table S3, we also
593 classified variants according to our proposed reference gene annotations (fields beginning with New_);
594 when classifying variants in overlapping ORFs 3a/3c and N/9b we classify SNVs relative to the ORF in
595 which the variant is non-synonymous if that is true for only one of the frames, or the ORF for which the
596 amino acid change is more radical (as defined by the blosum62 matrix obtained from biopython version
597 1.58⁵³) if it is non-synonymous in both frames, or the larger ORF if the variant is synonymous in both
598 frames.

599 We determined mature proteins for which the density of amino-acid-changing SNVs differed
600 significantly from the density that would be expected from their level of conservation, by calculating the
601 residual of a linear regression of amino-acid-changing SNV density as a function of the fraction of
602 conserved amino acids, for all mature proteins. The regression line was $y=0.235-0.165x$. We
603 determined significance using a binomial p-value with a false discovery rate cutoff of 0.05. To further
604 test significance of the SNV depletion in S1, we downloaded a larger set of SNVs from the UCSC Table
605 Browser as above on 2020-05-09.

606 The 16 Spike-protein variants prioritized were those reported by Korber et al. in their bioRxiv preprint or
607 later *Cell* publication (ones at greater than 0.3% frequency, or 0.1% if near certain epitopes).

608 To find regions that were significantly enriched for missense variants in conserved amino acids, we first
609 defined a null model as follows. For each mature protein, we counted the number of missense variants
610 and the number of conserved amino acids and randomly assigned each SNV to a conserved amino

611 acid in the same mature protein, allowing multiplicity. For any positive integer n, we found the largest
612 number of variants that had been assigned to any set of n consecutive conserved amino acids within
613 the same mature protein across the whole genome. Doing this 100,000 times gave us a distribution of
614 the number of missense variants in the most enriched set of n consecutive conserved amino acids in
615 the genome. Comparing the number of actual missense variants in any particular set of n consecutive
616 conserved amino acids to this distribution gave us a nominal p-value for that n. We applied this
617 procedure for each n from 1 to 100 and multiplied the resulting p-values by a Bonferroni correction of
618 100 to calculate a corrected p-value for a particular region to be significantly enriched. We note that
619 these 100 hypotheses are correlated because enriched regions of different lengths can overlap, so a
620 Bonferroni correction is overly conservative and our reported p-value of 0.012 understates the level of
621 statistical significance. To find significantly depleted regions we applied a similar procedure with every n
622 from 1 to 1000, but did not find any depleted regions with nominal p-value less than 0.05 even without
623 multiple hypothesis correction.

624 Miscellaneous

625 Ribosome footprints shown in Extended Data Fig. 3 are from the track hub at [ftp://ftp-](ftp://ftp-igor.weizmann.ac.il/pub/hubSARSRibo.txt)
626 [igor.weizmann.ac.il/pub/hubSARSRibo.txt](ftp://ftp-igor.weizmann.ac.il/pub/hubSARSRibo.txt)¹⁰.

627 Data Access

628 The PhyloCSF tracks and FRESCo synonymous constraint elements are available for the SARS-CoV-
629 2/wuhCor1 assembly in the UCSC Genome Browser at <http://genome.ucsc.edu> as public track hubs
630 ^{1,23,24,48} named “PhyloCSF” and “Synonymous Constraint”. The alignments and phylogenetic tree used
631 here are provided as supplementary materials. The alignments may be viewed, color coded to indicate
632 protein-coding signatures, using CodAlignView (<https://data.broadinstitute.org/compbio1/cav.php>) with
633 alignment set wuhCor1_c and chromosome name NC_045512v2.

634 Our proposed reference gene set for SARS-CoV-2 and the set of previously proposed genes that we
635 have rejected are included in BED format in Supplementary materials and are available as the

636 “PhyloCSF Genes” track in the UCSC Genome Browser (the track showing the genes we have rejected
637 may be displayed using the configuration page).

638 A browser track showing SARS-CoV-2 single nucleotide variants, color coded by whether they are non-
639 coding, synonymous, or amino-acid-changing, and whether they are in conserved codons, as well as
640 tracks showing all codons that are conserved at the amino acid or synonymous level, may be viewed in
641 the UCSC Genome Browser using the track hub at [https://data.broadinstitute.org/compbio1/SARS-
642 CoV-2conservation/trackHub/hub.txt](https://data.broadinstitute.org/compbio1/SARS-CoV-2conservation/trackHub/hub.txt). The details page for each SNV includes information about
643 Sarbecovirus conservation and a link to view the alignment of a neighborhood of the SNV in
644 CodAlignView. It is our intention to update this track hub as the list of variants in the UCSC Table
645 Browser is updated. [Note to reviewers: classification is currently with respect to NCBI annotations; we
646 will add a track classifying SNVs with respect to our PhyloCSF Genes annotations once our paper is
647 accepted.]

648 In this resource, we have augmented variant data made available by UCSC⁵⁴ with our own
649 annotations. UCSC data came from nextstrain.org⁵⁵, which was derived from genome sequences
650 deposited in GISAID³⁵. Right of use and publication of the underlying sequences is entirely controlled
651 by the authors of the original resource and the contributors of individual sequences, who are
652 acknowledged in the Nextstrain metadata file included with supplementary materials. Our analysis
653 provides an additional layer of annotation on their work rather than replicating or replacing it.

654 Original data usage policy as provided by UCSC: “The data presented here is intended to rapidly
655 disseminate analysis of important pathogens. Unpublished data is included with permission of the data
656 generators, and does not impact their right to publish. Please contact the respective authors (available
657 via the Nextstrain metadata.tsv file) if you intend to carry out further research using their data. Derived
658 data, such as phylogenies, can be downloaded from nextstrain.org (see "DOWNLOAD DATA" link at
659 bottom of page) - please contact the relevant authors where appropriate.”

660

661 **Acknowledgements**

662 We thank the UCSC genome browser staff and Maximilian Haeussler in particular for sharing our gene
663 annotations with the community. We thank all contributors to the GISAID database for sharing primary
664 sequences, and nextstrain.org/ucsc.edu for making variant data available. We thank Jeremy Luban,
665 Robert Garry, and Mark Diekhans for helpful input. This work was supported by the National Human
666 Genome Research Institute of the National Institutes of Health under Award Number U41HG007234.
667 The content is solely the responsibility of the authors and does not necessarily represent the official
668 views of the National Institutes of Health. Additional support was provided by the Wellcome Trust grant
669 number WT108749/Z/15/Z and NIH grant R01 HG004037.

670

671 **Author Contributions**

672 I.J. and M.K. conceived and designed the study and carried out all analyses. R.S. calculated
673 synonymous constraint. I.J. and M.K. wrote the manuscript.

674

675 **Competing interest declaration**

676 The authors declare no competing interests.

677 **Data Availability and Code Availability**

678 All data generated or analysed during this study are included in this published article and its
679 supplementary information files.

680 **Additional info**

681

682 Supplementary Information is available for this paper.

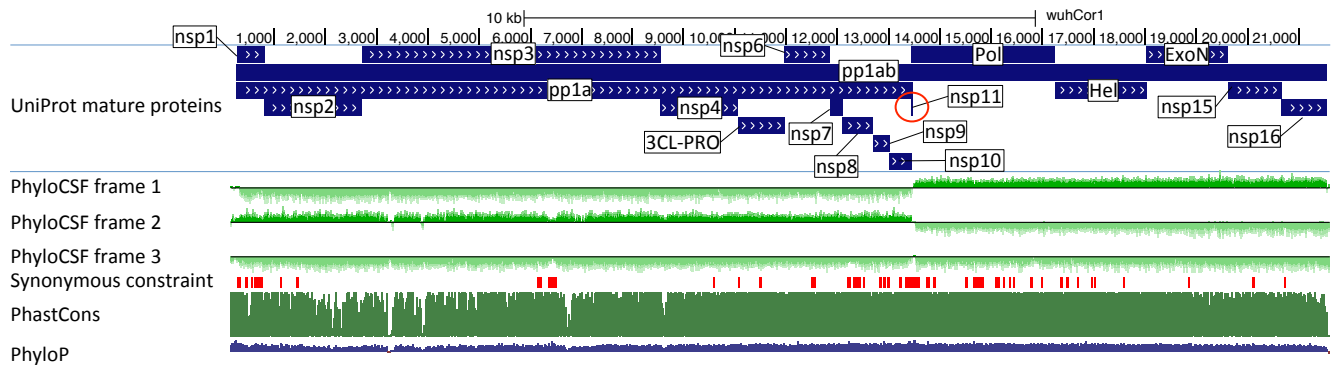
683

684 Correspondence and requests for materials should be addressed to Manolis Kellis
685 manoli@mit.edu.

686 **Extended Data Figures**

687

688



689

690 **Extended Data Figure 1. PhyloCSF signal for polyprotein ORF1.** UCSC Genome Browser view SARS-CoV-2 genome
691 for polyprotein ORF1ab showing UniProt gene annotations for individual non-structural proteins (nsp), PhyloCSF tracks
692 (green) in each of 3 reading frames, and Synonymous Constraint Elements (SCEs, red), along with phastCons/phyloP
693 nucleotide-level constraint (green/blue). Polyprotein 1ab is processed into 16 mature peptides nsp1-nsp16. PhyloCSF
694 signal shows clear protein-coding signal for all proteins, indicating clearly that all are functional proteins (except nsp11,
695 red circle, discussed in the main text). PhyloCSF signal captures the correct frame throughout the entire length of each
696 protein (except nsp3, where some small regions show reduced frame-2 signal and/or increased frame-3 signal, but upon
697 inspection these are only stop-codon-free in frame-2 and do not represent dual-coding candidates).



698

55 Coronaviridae

44 Sarbecoviruses

699

Extended Data Figure 2. Phylogenetic tree of 44 Sarbecovirus genomes and larger phylogenetic context. Left:

700

Phylogenetic tree of a selection of Coronaviridae genomes, including the seven that infect humans (red asterisks). Right:

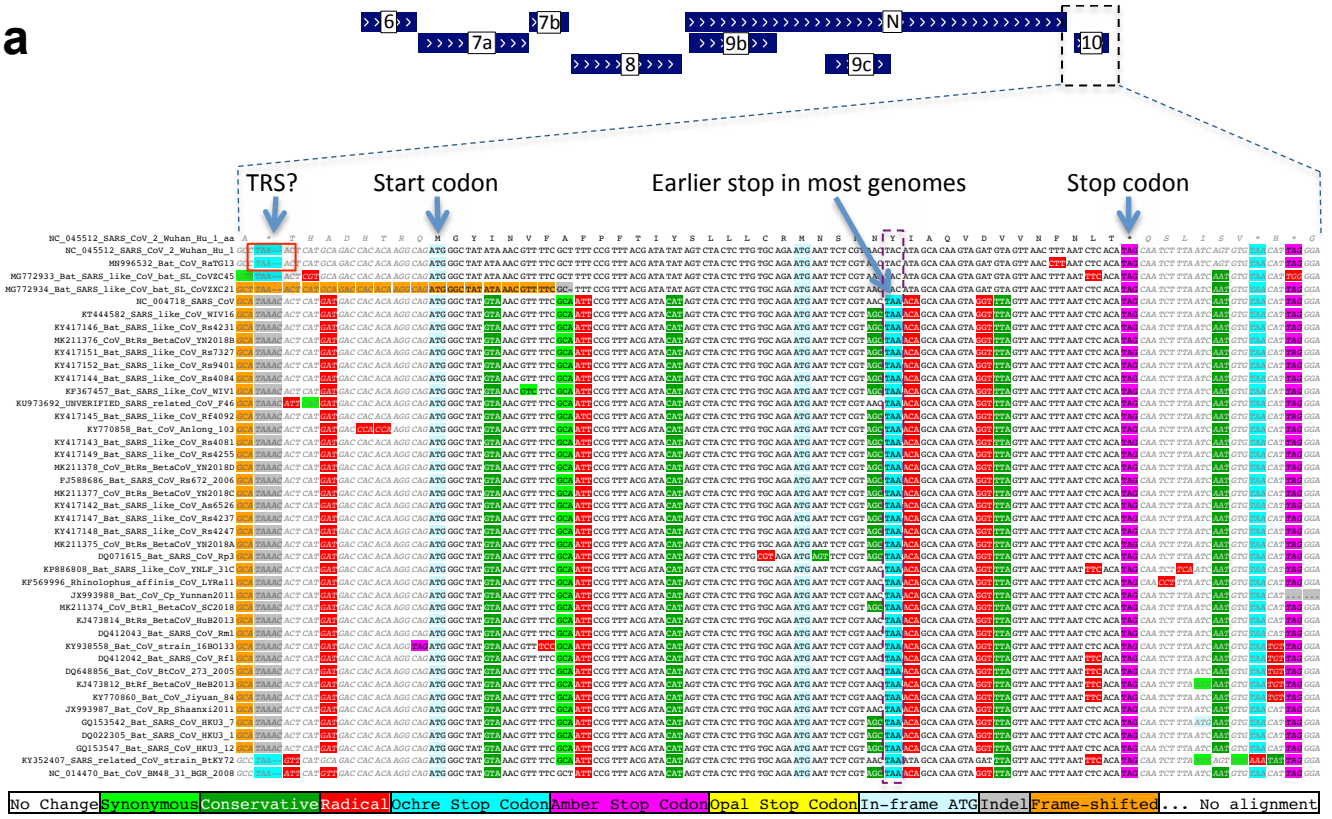
701

Phylogenetic tree of the 44 Sarbecovirus genomes used in this study. Trees are based on whole-genome alignments and

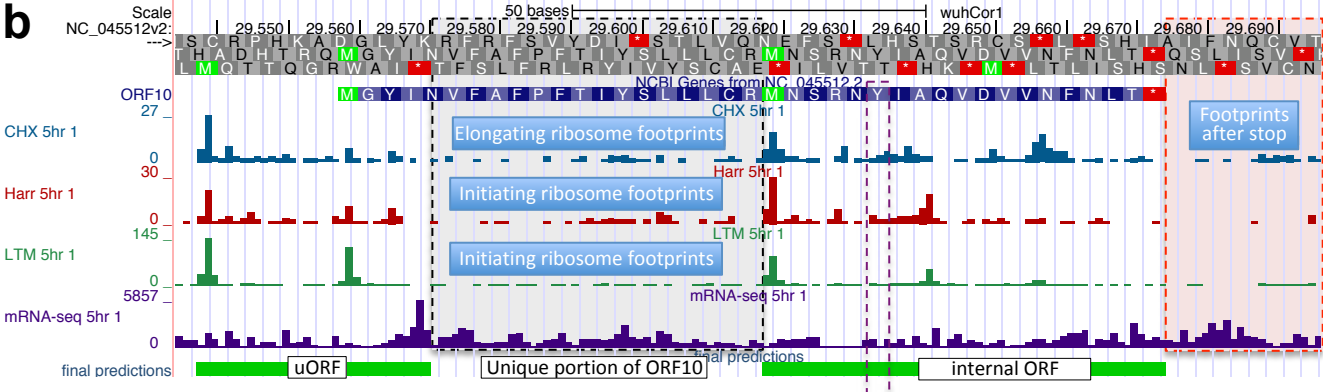
702

might be different from the history at particular loci, due to recombination.

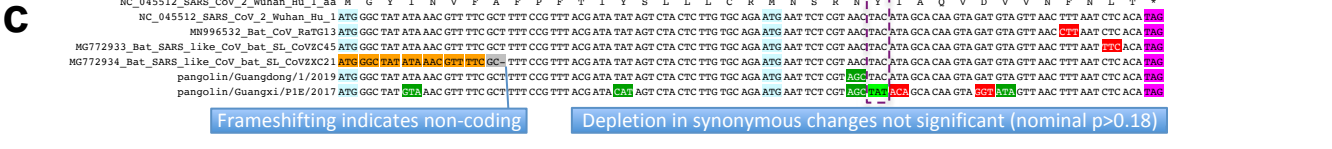
a



b



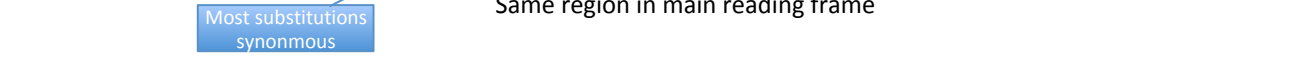
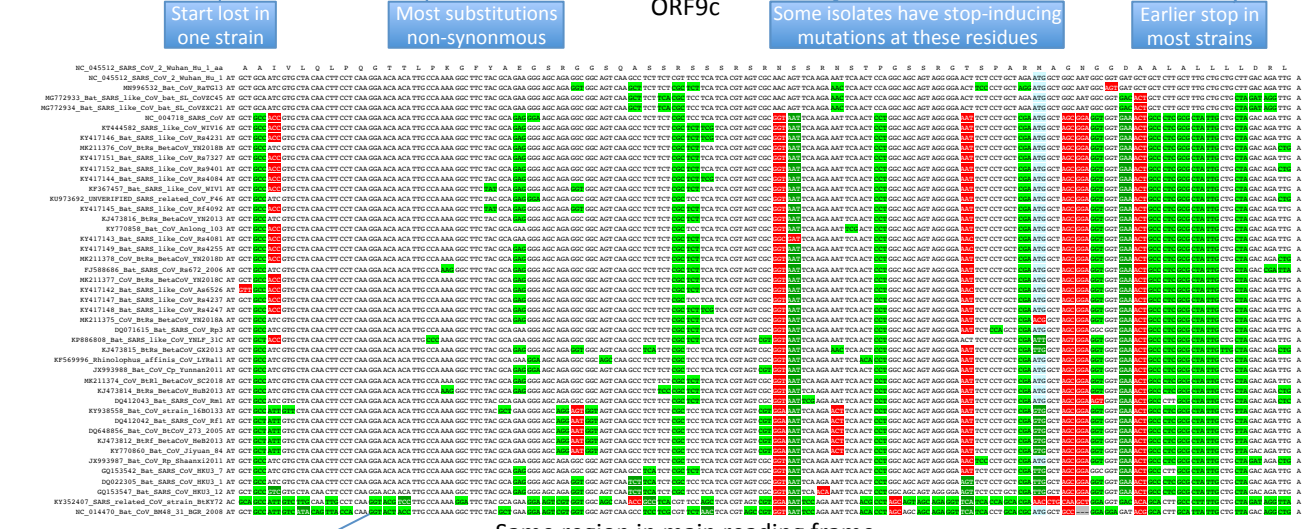
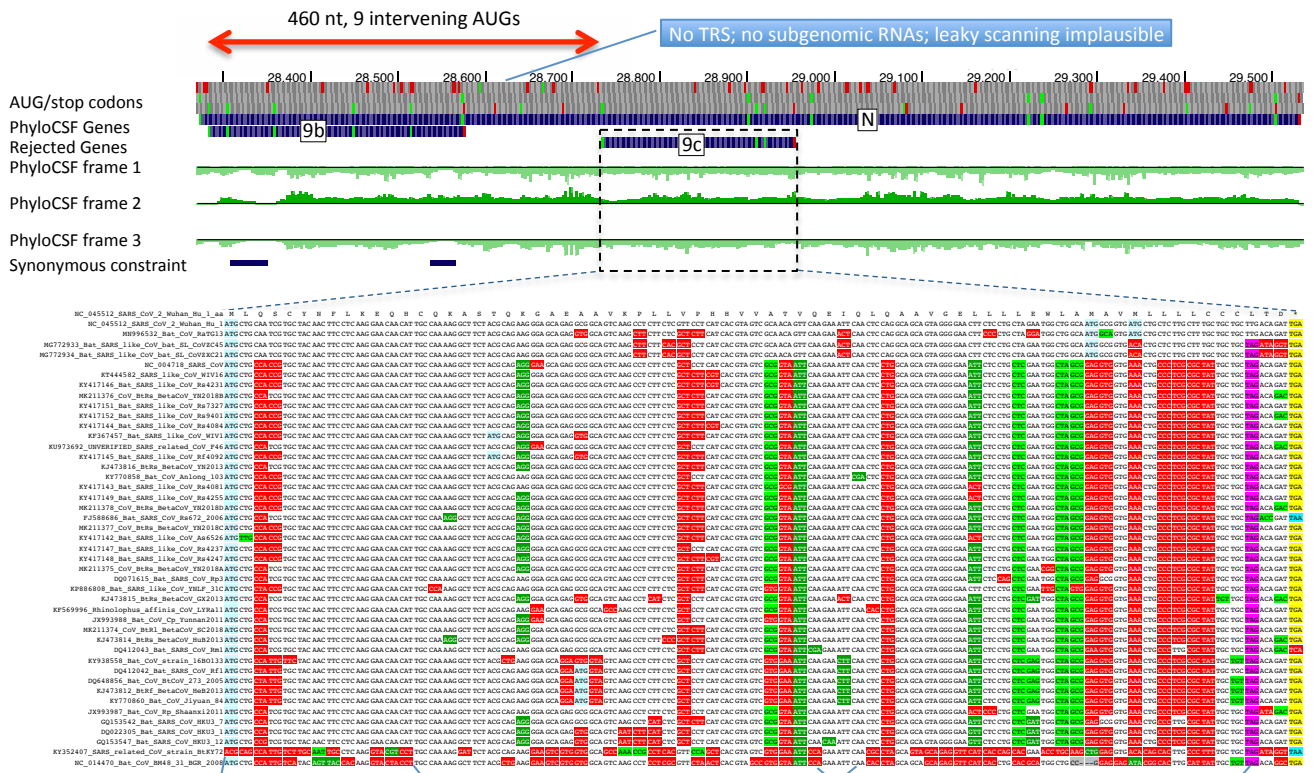
c



Extended Data Figure 3. ORF10 is not protein-coding.

a. Alignment of Sarbecovirus genomes at ORF10, including 30 additional flanking nucleotides on each side. Most substitutions are amino-acid-changing, either radical (red) or conservative (dark green), with only two synonymously-changing positions (light green), indicating this is not a protein-coding region. In addition, nearly all strains show an earlier stop codon (cyan), further reducing the length of this already-short ORF from 38 amino-acids to 25, and one of the four strains lacking the earlier stop includes a frame-shifting deletion. The putative partial transcription-regulatory sequence (TRS) present in SARS-CoV-2 and its closest relative (Bat

710 CoV RaTG13) is not conserved in any other strains. The region surrounding ORF10 shows very high nucleotide-level
711 conservation, which spans ORF10 and extends beyond its boundaries in both directions, indicating that this portion of the
712 genome is functionally important even though it does not code for protein (indeed, this region is part of a pseudoknot RNA
713 structure involved in RNA synthesis). **b.** Ribosome footprints previously used to suggest ORF10 translation¹⁰ in fact
714 localize either in an upstream ORF (uORF, green) or in an internal ORF (green, “final predictions” track¹⁰), but not in the
715 unique portion of ORF10 (dashed black box), indicating they are less likely to reflect functional translation of ORF10, and
716 more likely to represent incidental translation initiation events. We note that the density of elongating footprints in the
717 unique portion (black box) is no greater than the density after the stop codon (red box), consistent with incidental events.
718 We also note that the internal ORF is only 18 codons long in 4 strains, and only 5 codons long in the other 40
719 Sarbecovirus strains, given the early stop codon (purple box) and unlikely to be functional. Footprint tracks show
720 elongating ribosome footprints in cells treated with cycloheximide (blue, CHX), and footprints enriched for initiating
721 ribosomes using harringtonine (Harr, red), and lactimidomycin (LTM, green). “mRNA-seq” track shows RNA-seq reads. **c.**
722 CodAlignView²⁵ of alignment previously used to argue that a high dN/dS ratio in ORF10 indicated positive selection for
723 protein-coding-like rapid evolution⁸, based on only six closely-related strains (SARS-CoV-2, three bat viruses, two
724 pangolin viruses). The authors noted a frameshifting deletion (orange/grey) in one of the bat viruses, which provides
725 strong evidence against conserved protein-coding function, but they interpreted it (without evidence) as a potential
726 sequencing error and excluded the strain from consideration. Even ignoring the frameshift-containing strain, the evidence
727 used is insufficient to reach statistical significance: the alignment includes only 9 substitutions, of which 4 are radical, 4
728 are conservative, and 1 is synonymous. In a neutrally-evolving region with 9 substitutions, we would expect 2-3
729 synonymous changes, depending on the evolutionary model used, and a depletion to only 1 synonymous change is not
730 statistically significant (nominal p-value>0.18 even in the most generous evolutionary model). This already-non-significant
731 nominal p-value would move even further from significance with the necessary multiple-hypothesis corrections.
732



733 **No Change****Synonymous****Conservative****Radical****Ochre****Stop Codon****Amber Stop Codon****Opal Stop Codon****In-frame****ATG****Indel****Frame-shifted**

734 **Extended Data Figure 4. Nucleocapsid-overlapping ORF9c is not protein-coding** Sarbecovirus alignment of frame2-

735 encoded ORF9c (top), which overlaps frame3-encoded Nucleocapsid (bottom). ORF9c start codon is lost in one,

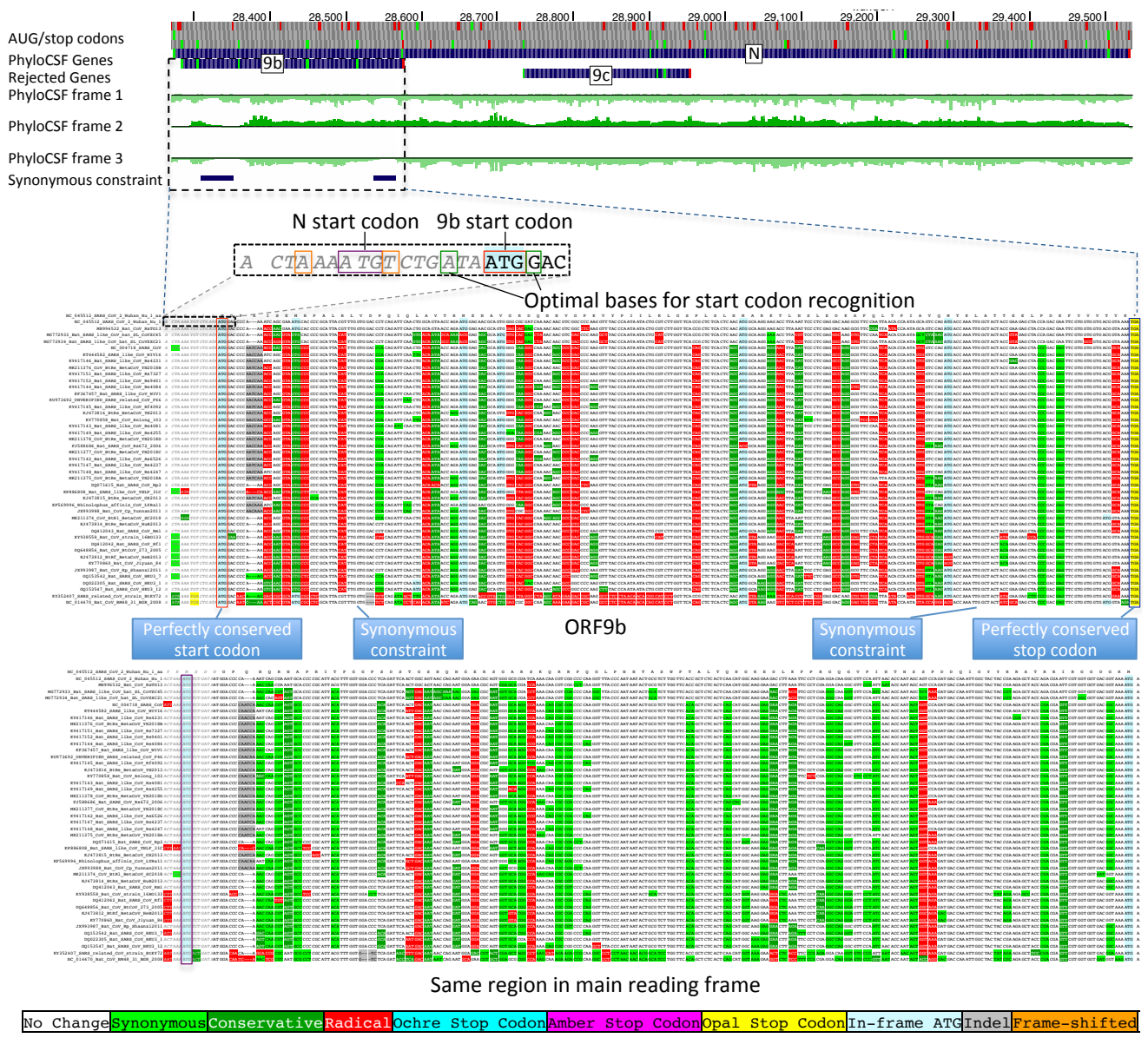
736 and most strains have an earlier UAG stop codon (magenta) 3 codons before the end. In Nucleocapsid-encoding frame 2

737 (bottom), nearly all nucleotide substitutions are amino-acid-preserving (synonymous, light green), indicating strong

738 purifying selection for protein-coding function. By contrast, in ORF9c-encoding frame 3 (top), nearly all nucleotide

739 substitutions result in function-disrupting (radical) amino acid changes (red), and very few result in synonymous (light
740 green) or function-preserving (conservative, dark green) substitutions, indicating lack of purifying selection for protein-
741 coding function for ORF9c, so it does not play conserved protein-coding functions. In addition, ORF9c is unlikely to be
742 translated via leaky ribosomal scanning because its start codon is 460 nucleotides after N's (red arrow) with 9 intervening
743 AUG codons (green dots), direct-RNA sequencing found no ORF9c-specific subgenomic RNAs¹⁶⁻¹⁸, no TRS is
744 appropriately positioned to create one, and several SARS-CoV-2 isolates³⁵ contain stop-introducing mutations⁷, indicating
745 that ORF9c is not a recently-evolved strain-specific gene either. We conclude 9c is not protein-coding.

746



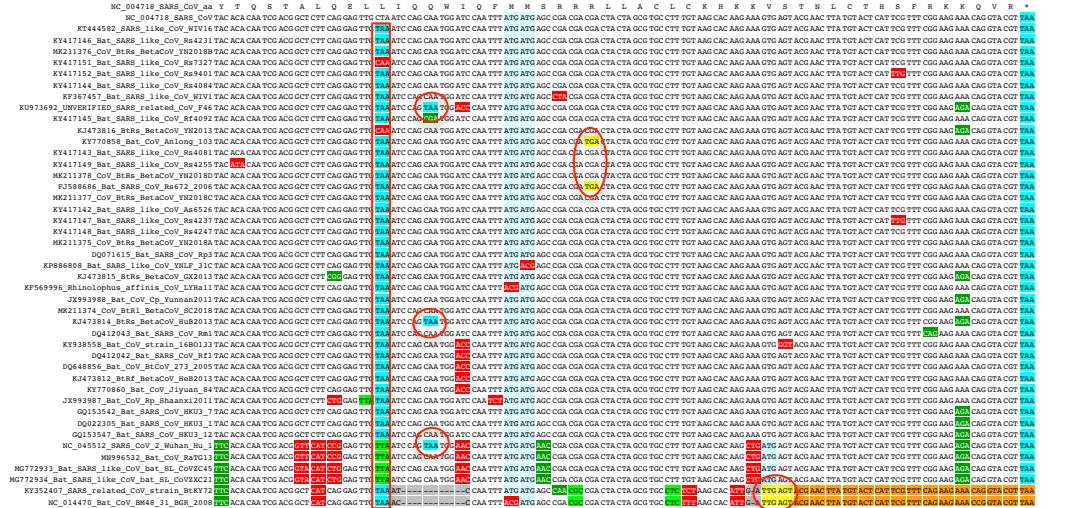
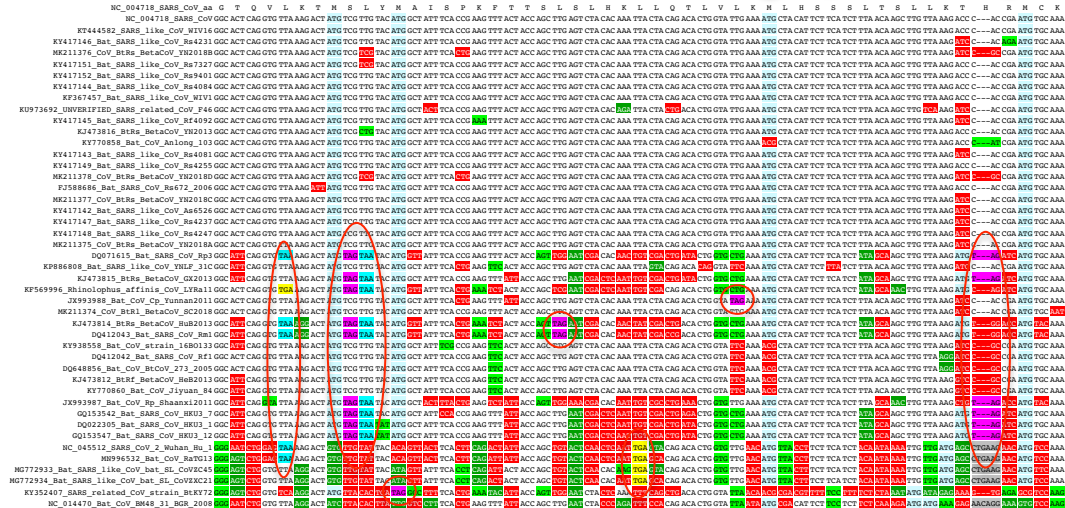
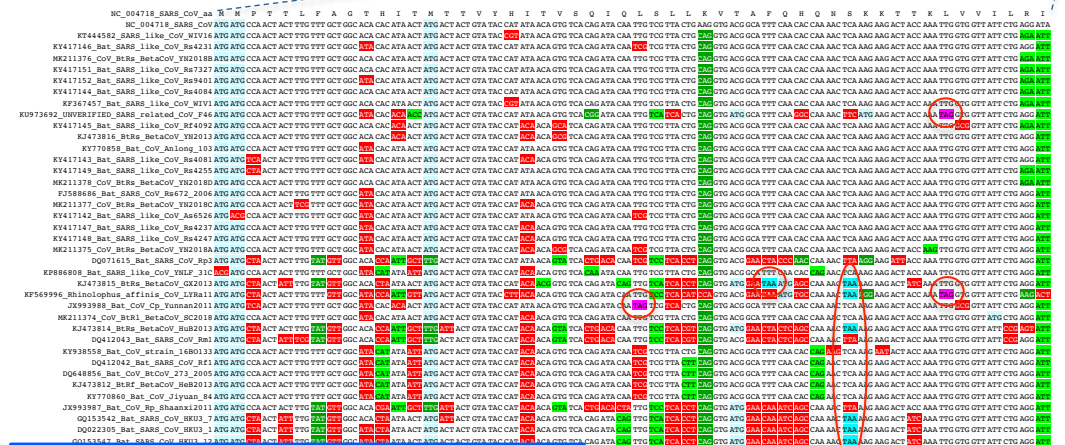
750 **Extended Data Figure 5. Nucleocapsid-overlapping ORF9b is protein-coding.** Sarbecovirus alignment of frame3-
 751 encoded ORF9b (top), which overlaps frame2-encoded Nucleocapsid (bottom). Although ORF9b-encoding frame3 shows
 752 many function-disrupting (radical, red) substitutions, its start codon (red box) is perfectly conserved, its stop codon (blue
 753 box) is perfectly conserved, and there are no intermediate stop codons in any strain. Moreover, its Kozak start-codon
 754 context (dashed black box) is optimal for ribosomal start codon recognition, with A in position -3 and G in position +4
 755 (green boxes), while the start codon context of N is less optimal, with an A in -3 and T in +4 (orange boxes), making it

756 likely that ORF9b can be translated by leaky scanning from the same subgenomic RNA as N, as it is only ~2 codons
757 downstream of N's start. Moreover, both the optimal 9b start-codon context, and the less-optimal N start-codon context
758 are fully-conserved features across all Sarbecovirus strains, indicating that leaky-scanning translation may be a
759 conserved feature throughout Sarbecoviruses. In addition, ORF9b shows significant localized synonymous constraint in N
760 in its start and end regions (**Fig. 3**), even relative to the overall low synonymous rate of N, consistent with dual-coding
761 functions. ORF9b also has proteomics support^{15,36,37} in SARS-CoV-2, including evidence of viral-RNA binding³⁸, and
762 alternate-frame translation support by ribosome profiling¹⁰. In SARS-CoV, ORF9b protein (and antibodies to it) was
763 detected in SARS patients^{39,40}, localized in mitochondria, and interfered with host cell antiviral response when
764 overexpressed⁴¹. We conclude ORF9b encodes a conserved functional protein-coding gene.

AUG/stop codons

SARS-CoV-2 ORFs

SARS-CoV ORFs

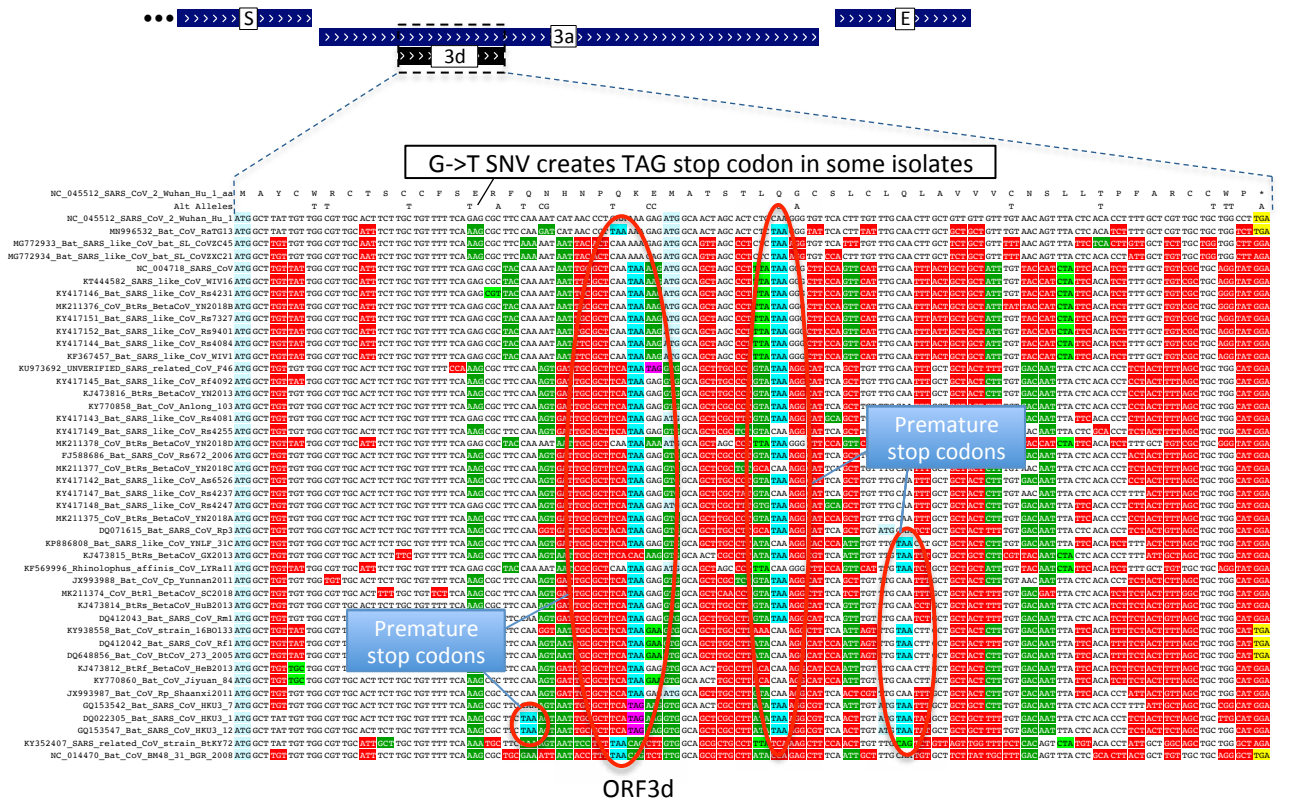


No Change Synonymous Conservative Radical Ochre Stop Codon Amber Stop Codon Opal Stop Codon In-frame ATG Indel Frame-shifted

766

767 **Extended Data Figure 6. ORF3b is not protein-coding.** Sarbecoviruses alignment of SARS-CoV 154-codon ORF3b
768 overlapping ORF3a, (reordered with SARS-CoV and related strains on top). Although start codon is conserved in all but
769 one strain, ORF length is highly variable due to numerous in-frame stop codons (red ovals and red rectangle). The 22-
770 codon ORF in SARS-CoV-2 has strongly negative PhyloCSF score, does not overlap any SCEs, and even among the four
771 strains sharing its stop codon (blue rectangle) all six substitutions are radical amino acid changes, providing no evidence
772 of amino-acid-level purifying selection. Ribosome profiling did not find translation of ORF3b, transcription studies did not
773 find substantial transcription of an ORF3b-specific subgenomic RNA, and translation by leaky scanning would implausibly
774 require ribosomal bypass of eight AUG codons (green rectangles, top panel), some with strong Kozak context.
775 (Supplementary Fig. S3 has comparison to reading frame of ORF3a.)

776



777 No Change Synonymous Conservative Radical Ochre Stop Codon Amber Stop Codon Opal Stop Codon In-frame ATC Indel Frame-shifted... No alignment

778 **Extended Data Figure 7. ORF3d is not protein-coding.** Sarbecovirus alignment of 57-codon ORF3d (referred to by

779 some authors as 3b) overlapping ORF3a shows mostly function-altering radical amino-acid substitutions (red columns),

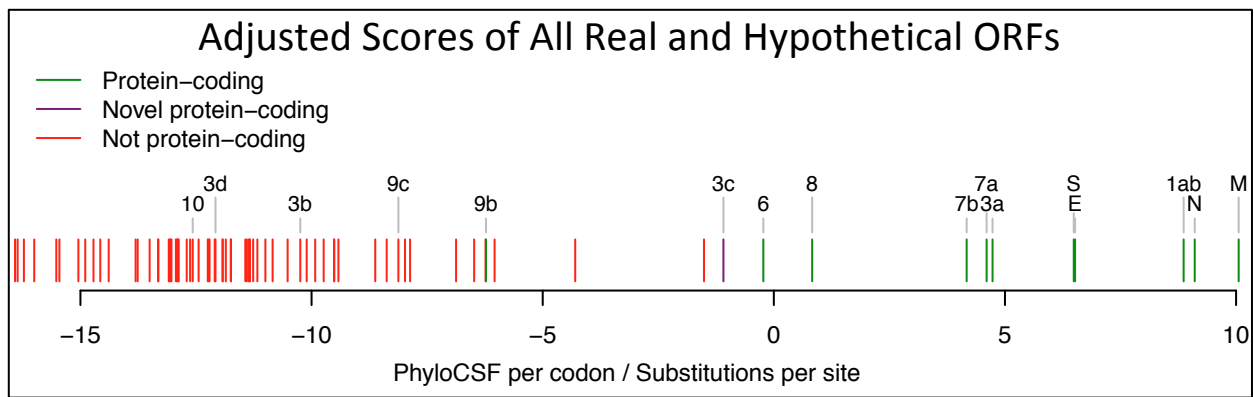
780 and repeated interruption of by one or more premature stop codons in all other strains (red ovals), unambiguously

781 indicating that ORF3d is not a conserved protein-coding gene. A substantial fraction of SARS-CoV-2 isolates have stop-

782 introducing mutations, and ribosome profiling did not identify ORF3d as a translated ORF¹⁰, indicating that it is not a

783 recently-evolved strain-specific gene either.

784



785

786

787

788

789

790

791

792

793

794

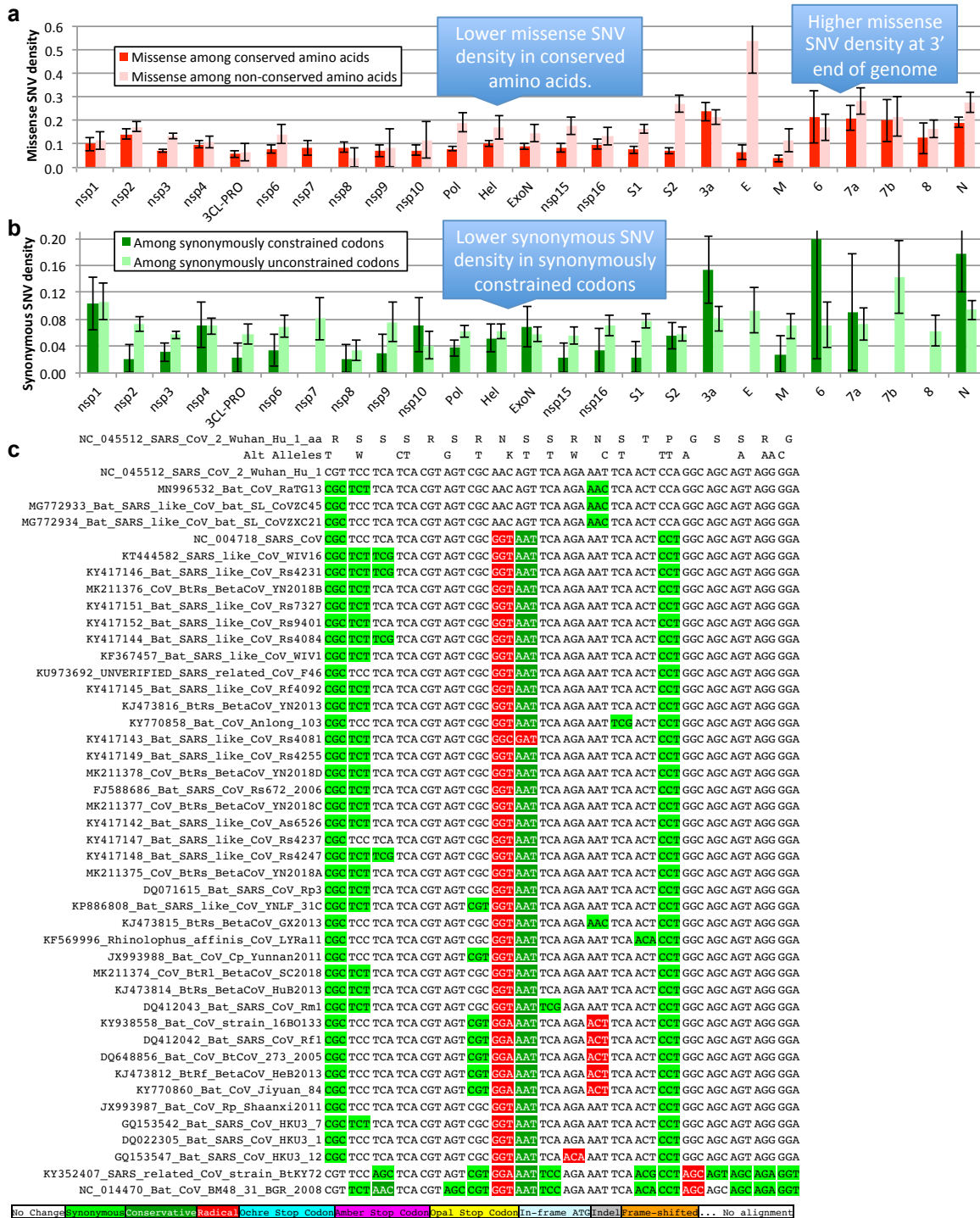
795

796

797

798

Extended Data Figure 8. Branch-length-adjusted PhyloCSF score strongly rejects ORF10. Similar to Fig. 1c, but showing PhyloCSF scores per codon divided by the average number of substitutions per site, to adjust for the fact that high-nucleotide-conservation regions show compressed unscaled PhyloCSF scores (closer to zero) because there are fewer nucleotide substitution events. The branch-length-scaled score distribution further separates the scores of confirmed protein-coding genes (green) from non-protein-coding segments (red). The very low score of ORF10 with this adjustment indicates that its only-slightly-negative unscaled-PhyloCSF score in Fig. 1c stems from the high nucleotide conservation of the region, rather than protein-coding constraint. The scores of N-overlapping ORFs 9b and 9c are both reduced, consistent with the high nucleotide conservation of N. Notably, the branch-length-adjusted score for 3c remains high, consistent with its protein-coding nature, and despite the higher overall nucleotide conservation of its dual-coding region. We have manually inspected all other candidates with adjusted scores higher than 9c, and all are rejected (as not protein-coding): two are discussed in Supplementary Figure S4 (and are not protein-coding), and the remaining all show internal stop codons (and are not protein-coding).



799

800

801

802

803

804

Extended Data Figure 9. Single nucleotide variants and conservation. Error bars indicate standard error of mean. **a.** Density of SNVs disrupting conserved amino acids (dark red) is significantly lower than disrupting non-conserved amino acids (light red). Both densities are higher near the 3' end of the genome, indicating higher mutation rate or less purifying selection even among amino acids that are perfectly conserved in Sarbecovirus. **b.** Density of synonymous variants in synonymously constrained codons (dark green) is significantly lower than among synonymously unconstrained codons (light green), a depletion seen in most genes. Overall, conservation

805 in the Sarbecovirus clade at both the amino acid level and nucleotide level is associated with purifying selection on variants in the SARS-
806 CoV-2 population. **c.** Alignment of 20 amino acid Nucleocapsid region that is highly enriched for variants disrupting perfectly conserved
807 amino acids (alternate alleles shown in second row, W = A or T, K = G or T). There are 14 non-synonymous variants among the 14
808 perfectly conserved amino acids (columns with no red or dark green). This region is contained within a predicted B Cell epitope,
809 suggesting positive selection for immune system avoidance.

810

811

Figures

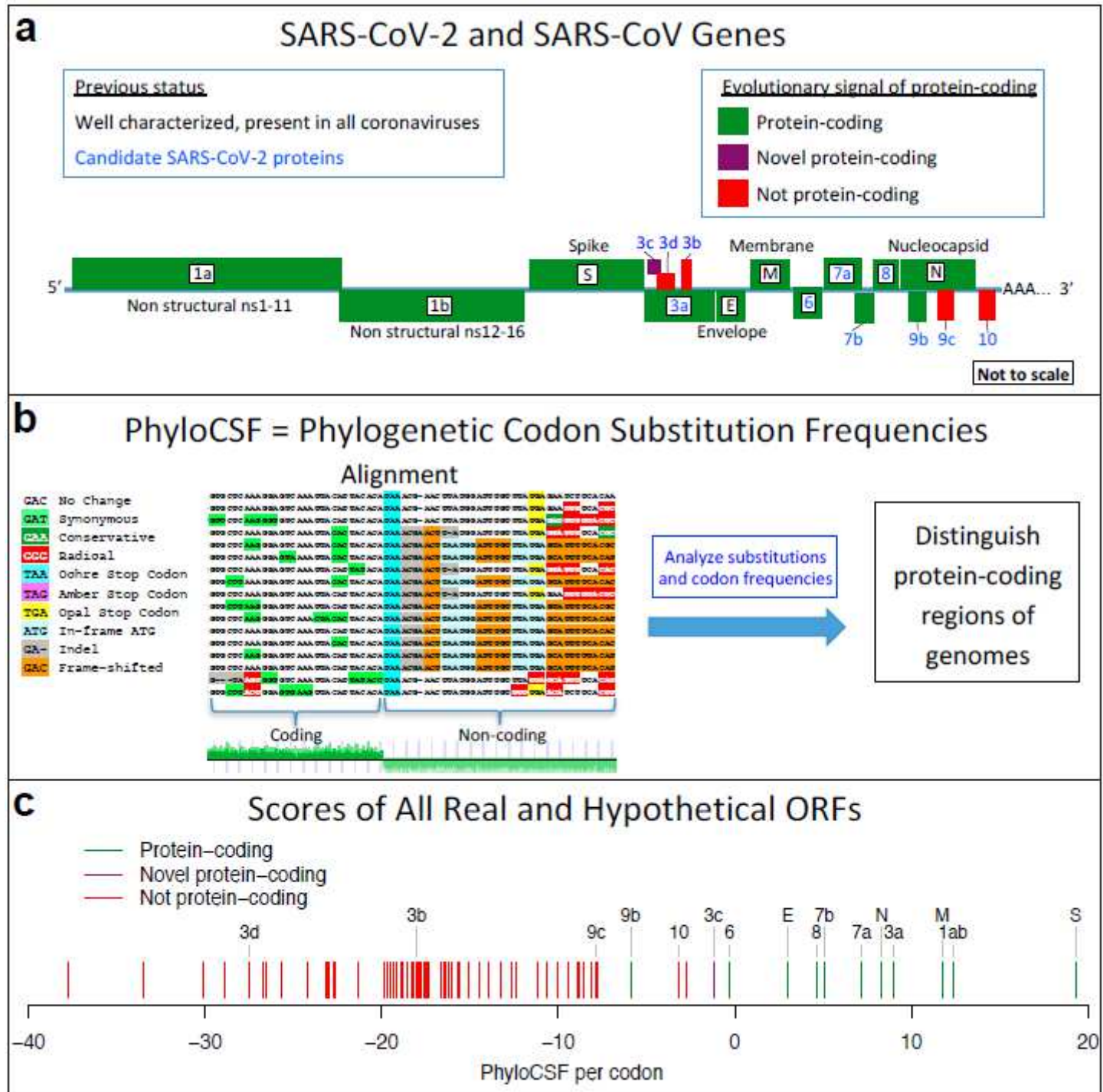


Figure 1

Overview. a. Previously annotated named (black font) and unnamed or proposed (blue font) SARS-CoV-2 genes, with confirmed protein-coding (green), rejected (red), or novel protein-coding (purple) classification, using evolutionary and experimental evidence. b. Phylogenetic Codon Substitution Frequencies (PhyloCSF) scores distinguish protein-coding (left) vs. non-coding (right) using evolutionary signatures, including distinct frequencies of amino-acid-preserving (green) vs. amino-acid-disruptive (red)

substitutions, and stop codons (cyan/magenta/yellow) in frame-specific alignments, and additional features. c. PhyloCSF score (x-axis) for all confirmed (green) and rejected (red) ORFs, showing annotated/hypothetical/novel (labeled) and all AUG-initiated ≥ 25 -codons-long locally-maximal ORFs (unlabelled). Novel ORF3c (purple) clusters with protein-coding. Only-modestly-negative ORF9c/ORF10 scores are artifacts of score compression in high-nucleotide-constraint regions, and substantially drop when nucleotide-473 conservation-scaled (see Extended Data Fig. 8).

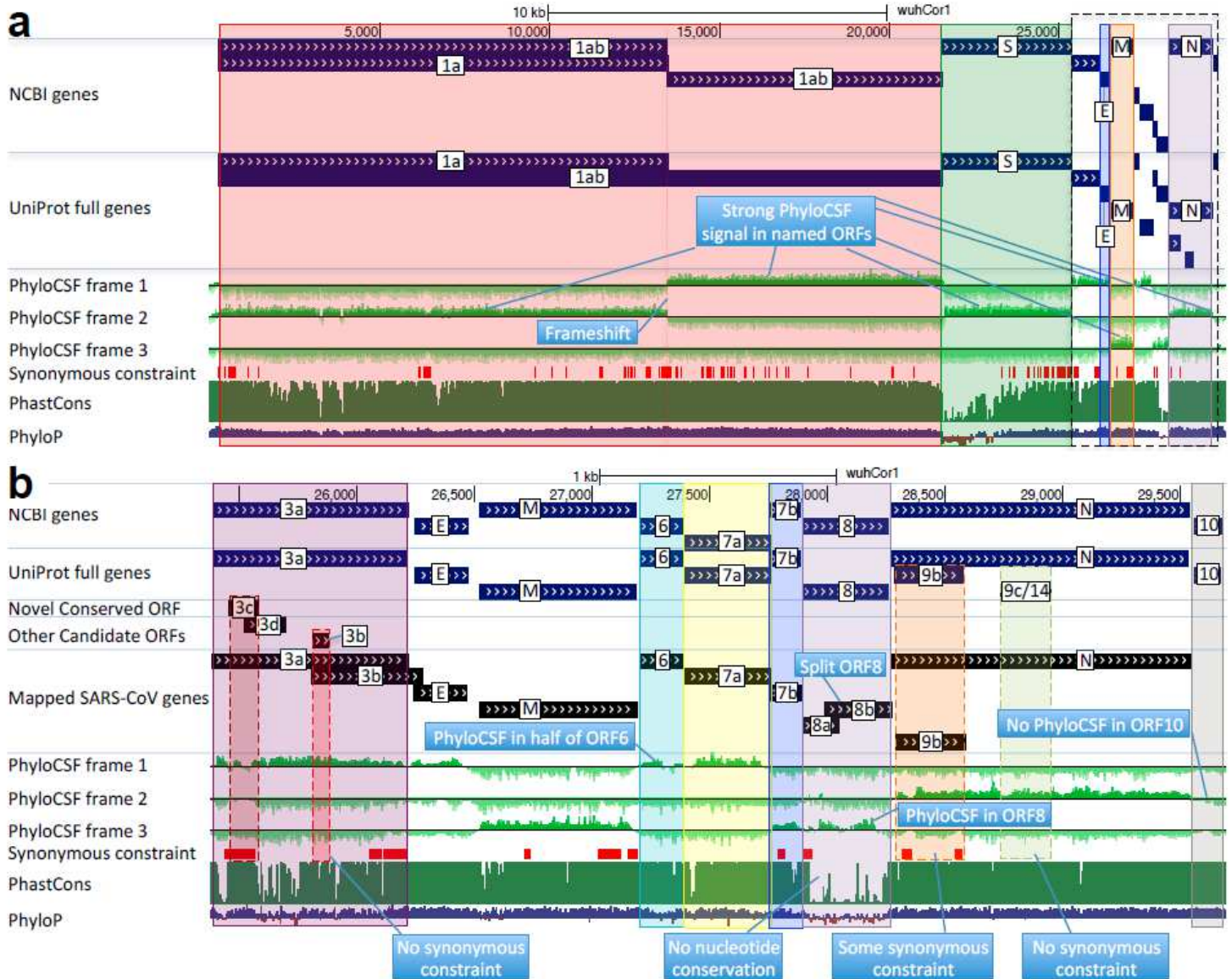


Figure 2

Genome-wide protein-coding signatures. SARS-CoV-2 NCBI/UniProt genes (blue), unannotated proposed genes and mapped SARS-CoV genes (black, panel b only), frame-specific protein-coding PhyloCSF scores (green), Synonymous Constraint Elements (SCEs) (red), and phastCons/phyloP nucleotide-level constraint (green/blue/red) across genomic coordinates (x-axis) for entire genome (panel a) and final 4-kb subset (panel b, dashed black box), highlighting (light blue boxes): (a) strong protein-coding signal in correct frame for each named gene; conservation-signal frame-change at programmed frameshift site;

strong protein-coding signal throughout S despite lack of nucleotide conservation in S1; (b) unambiguous and frame-specific protein-coding signal for unnamed ORFs 3a (despite only partial nucleotide conservation), 7a, 7b, and 8 (despite lack of nucleotide conservation); clear protein-coding signal in first half and last quarter of ORF6; no protein-coding signal for 10 (despite high nucleotide conservation); synonymous constraint (red) in novel-ORF 3c and confirmed-ORF 9b; no synonymous constraint in rejected ORFs 9c, 3b, 3d.

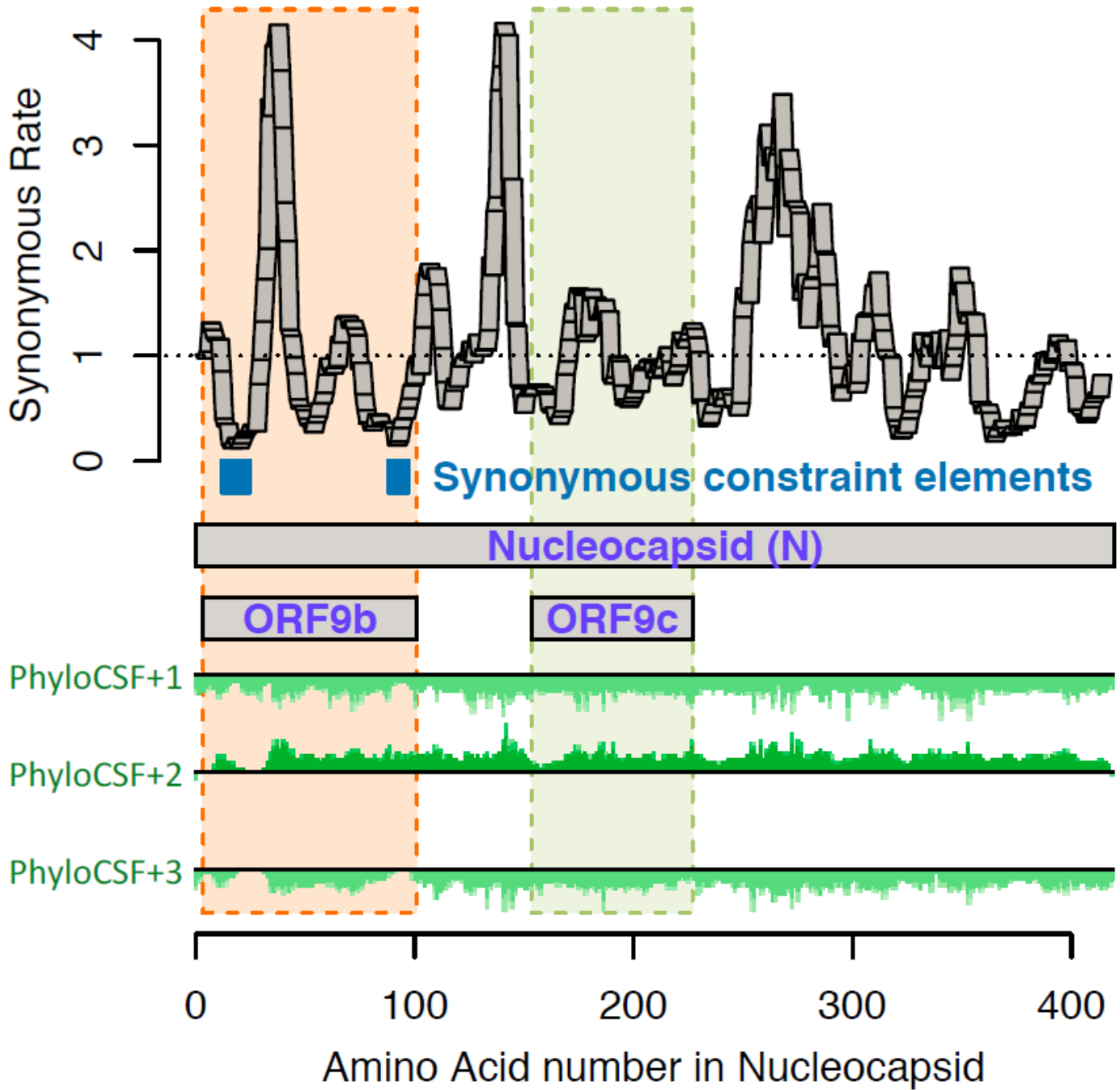


Figure 3

Synonymous constraint in Nucleocapsid overlaps 9b but not 9c/14. Synonymous substitution rate in 9-codon windows (y-axis) across N (x-axis), normalized to gene-wide average (dotted black line). Synonymous constraint elements (blue) expected for dual-coding constraint localize in overlapping ORF9b (dashed orange rectangle) indicating it is protein-coding, but not 9c (dashed purple rectangle) indicating it is not protein-coding. PhyloCSF protein-coding signal (green) in frame3 (encoding 9b and 9c/14) remains strongly negative throughout the length of 9c/14 (green box), indicating 9c/14 is non-coding, but rises to near-zero values for two regions of 9b, indicating protein-coding selection, while PhyloCSF signal frame 2 (encoding N) remains consistently high throughout the length of ORF9c.

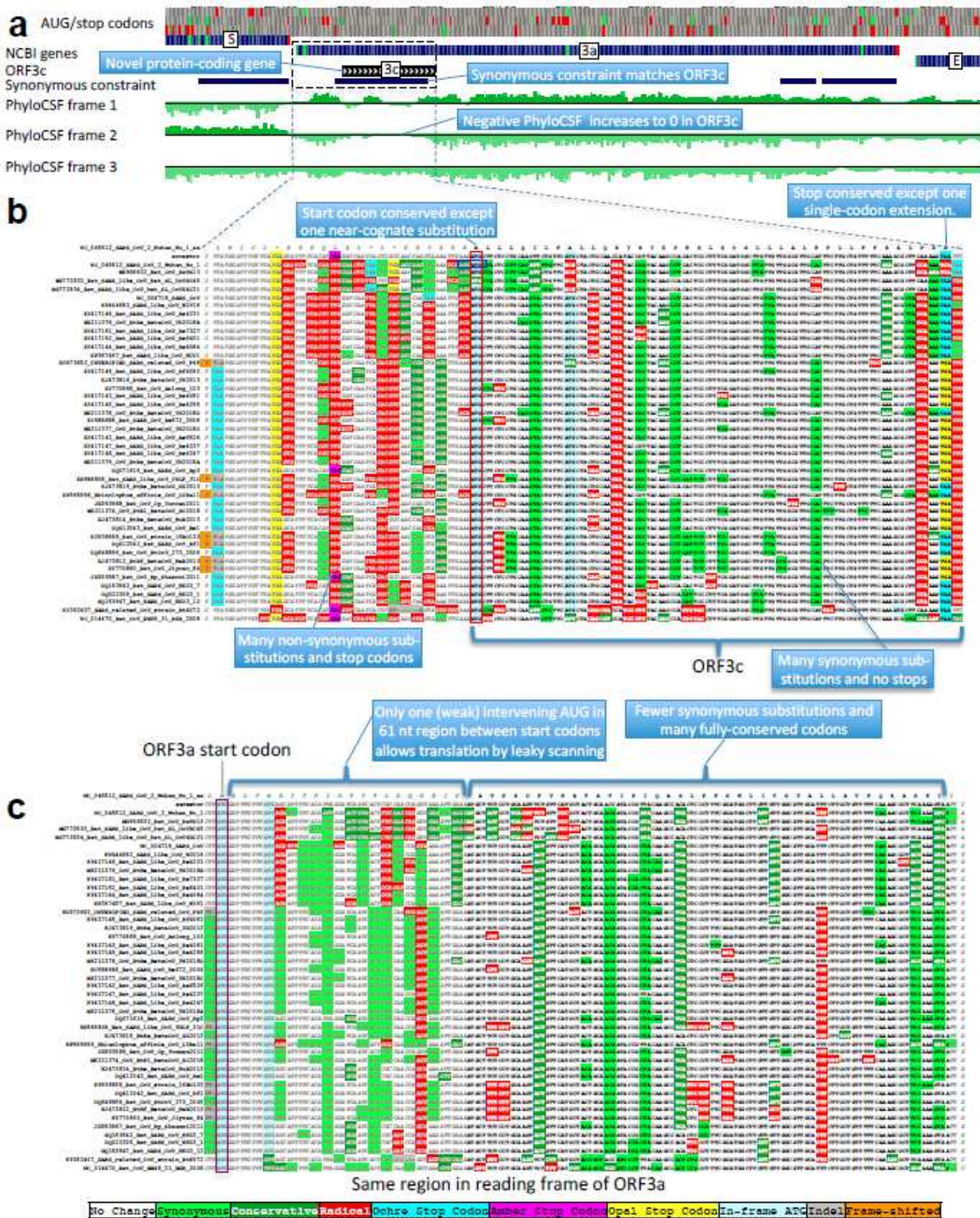


Figure 4

Novel gene 3c overlapping 3a is protein-coding. a. Synonymous-constraint elements (blue) match nearly perfectly 41-codon ORF3c dual-coding region boundaries (black), and protein-coding evolutionary signatures (green) switch between frame 1 and 2 (rows) in the dual-coding region, with frame-2 signal (negative flanking ORF3c) increasing to near-zero, and frame-1 signal (high flanking the dual-coding region) dropping to near-zero. b,c. Codon-resolution evolutionary signatures (colors, CodAlignView25)

annotating genomic alignment (letters) spanning ORF3a start and dual-coding region, in frame-1 (top) and frame-2 (bottom), highlighting (blue boxes): (b, frame-2, ORF3c) radical codon substitutions (red) and stop codons (yellow, magenta, cyan) prior to ORF3c start; synonymous (light green) and conservative (dark green) substitutions in ORF3c; ORF3c's start codon is conserved, except in one strain (row 4) with near-cognate GUG; ORF3c's stop codon is conserved except for one-codon extension in two strains (rows 2-3); no intermediate stop codons in ORF3c; (c, frame-1, ORF3a) abundant synonymous and conservative substitutions in 3a prior to dual-coding region; increase in fully-conserved codons (white) over dual-coding region. Short interval (61nt) with only one weak-Kozak-context intervening start codon indicates ORF3c may be translated from ORF3a's subgenomic RNA via leaky scanning.

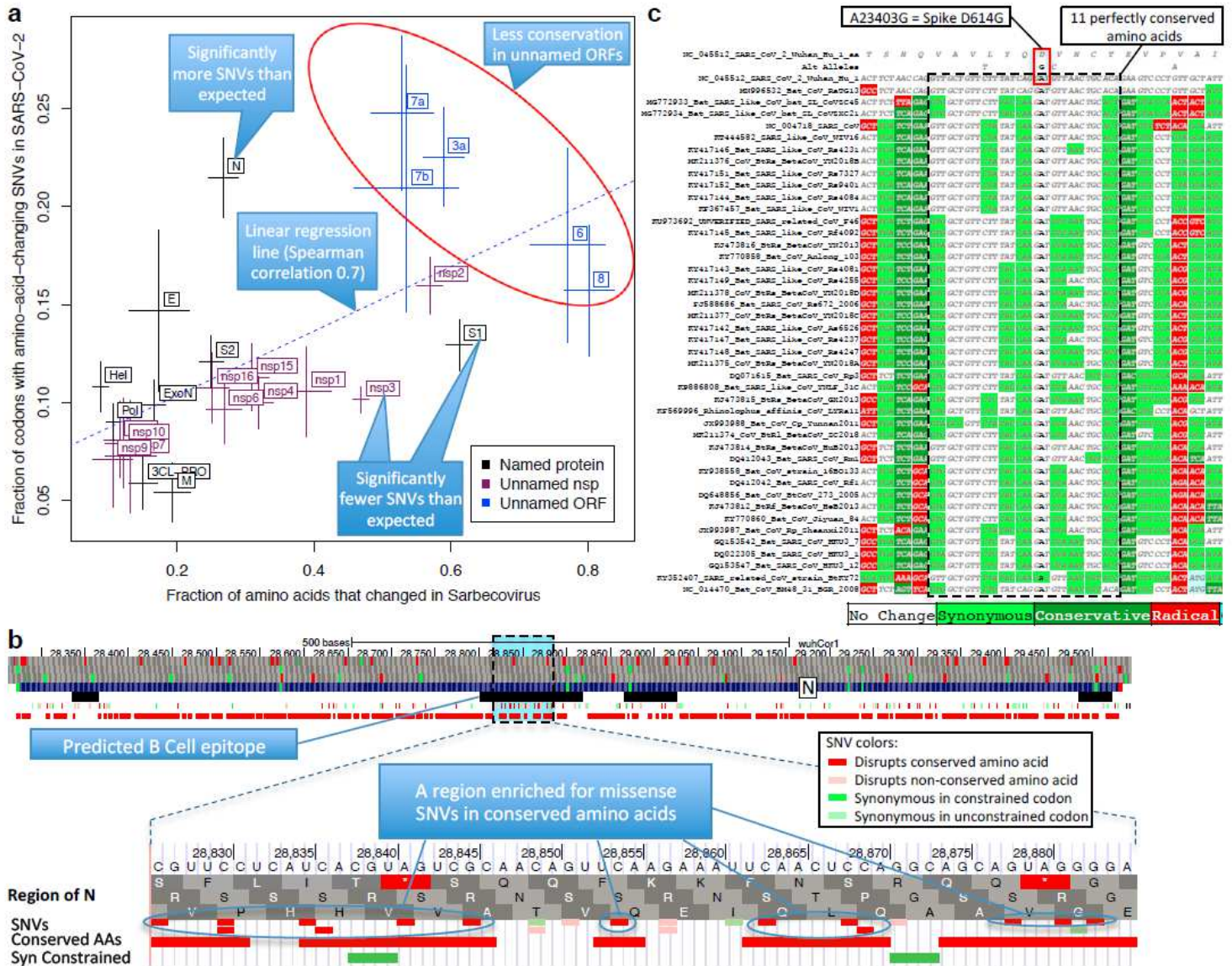


Figure 5

Within-strain variation vs. inter-strain divergence. a. Gene-level comparison. Long-term inter-strain evolutionary divergence (x-axis) and short-term within-strain variation (y-axis) show strong agreement (linear regression dotted line, Spearman-correlation=0.70) across mature proteins (crosses, denoting

standard error of mean on each axis), indicating that Sarbecovirus-clade selective pressures persist in the current pandemic. Well-characterized genes (black) show fewer changes in both timescales (bottom left) and less-well-characterized ORFs (blue) show more in both (top right). Significantly-deviating exceptions are: nsp3 and S1 (bottom right) showing significantly-fewer amino-acid-changing SNVs than expected from their cross-Sarbecovirus rapid evolution, and N (top left), showing significantly-more, possibly due to accelerated evolution in the current pandemic. b. Rapidly-evolving Nucleocapsid region. Top: Nucleocapsid context showing B-cell epitope predictions (black, “IEDB Predictions” track), and our annotation track-hub showing: conserved amino acids (red blocks), synonymously-constrained codons (green blocks), and SNV classification (colored tick-marks) as conserved/non-conserved (dark/light) and missense/synonymous (red/green); top 3 tracks show AUGcodons (green) and stop codons (red) in three frames. Bottom: Focus on 20-amino-acid region R185-predicted B-cell epitope (black) significantly-enriched for amino-acid-changing variants (red) disrupting perfectly-conserved residues, indicative of positive selection in SARS-CoV-2 for immune system avoidance. c. Spike D614G evolutionary context. Sarbecovirus alignment (text) surrounding Spike D614G amino-acid-changing SNV, which rose in frequency in multiple geographic locations suggesting increased transmissibility. This A-to-G SNV disrupts a perfectly-conserved nucleotide (bold font, A-to-G), which disrupts a perfectly-conserved amino-acid (red box, D-to-G), in a perfectly-conserved 11-amino-acid region (dotted black box, light-green=synonymous-substitutions) across bat-host Sarbecoviruses, indicating D614G represents a human-host-adaptive mutation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SIGuide.doc](#)
- [SupplementaryData.pdf](#)
- [44Sarbecoviruses.nh.txt](#)
- [44Sarbecoviruses.fa.txt](#)
- [PhyloCSFGenes.bed.txt](#)
- [PhyloCSFRejectedGenes.bed.txt](#)
- [nextstrainncovglobalmetadata.txt](#)
- [SupplementaryTableS1.txt](#)
- [SupplementaryTableS2.txt](#)
- [SupplementaryTableS3.txt](#)
- [SupplementaryTableS4.txt](#)