

1 **Codon usage pattern reveals SARS-CoV-2 as a monomorphic pathogen of hybrid origin**
2 **with role of silent mutations in rapid evolutionary success**

3 **Kanika Bansal^{1*}, Sanjeet Kumar², Prabhu B. Patil^{1*}**

4 ¹CSIR- Institute of Microbial Technology, Chandigarh, India

5 ²Gangadhar Meher University, Odisha, India

6 *Corresponding author

7 Address for correspondence:

8 Kanika Bansal, PhD

9 Research Associate

10 Laboratory of Bacterial Genomics and Evolution

11 CSIR- Institute of Microbial Technology, Chandigarh, India, 160036

12 Email: kanikabansal@imtech.res.in

13

14 Prabhu B. Patil, PhD

15 Principal Scientist

16 Laboratory of Bacterial Genomics and Evolution

17 CSIR- Institute of Microbial Technology, Chandigarh, India, 160036

18 Email: pbpatil@imtech.res.in

19

20

21

22

23

24

25

26

27

28 **Abstract**

29 Viruses are dependent on the host tRNA pool, and an optimum codon usage pattern (CUP) is
30 the driving force in its evolution. Systematic analysis of CUP of the coding sequences (CDS)
31 of representative major pangolin lineages A and B of SARS-CoV-2 indicate a single
32 transmission event of a codon-optimized virus from its source into humans. Here, no direct
33 congruence could be detected in CUP of all CDS of SARS-CoV-2 with the non-human
34 natural SARS viruses further reiterating its novelty. Several CDS show similar CUP with bat
35 or pangolin, while others have distinct CUP pointing towards a possible hybrid nature of the
36 virus. At the same time, phylogenetic diversity suggests the role of even silent mutations in
37 its success by adapting to host tRNA pool. However, genomes of SARS-CoV-2 from primary
38 infections are required to investigate the origins amongst the competing natural or lab leak
39 theories.

40 **Introduction**

41 The origin and success of the novel SARS coronavirus (SARS-CoV-2) (betacoronavirus)
42 causing the COVID-19 disease pandemic has been a topic of intense discussion. In the past
43 two decades since the first outbreak of SARS in 2002, several SARS-related coronaviruses
44 were reported from the bat, which was speculated to be significant reservoir for future
45 possible outbreaks ¹⁻⁴. Bats are the only flying mammals representing 20% of known
46 mammalian species and are critical natural reservoirs of many zoonotic viruses like Nipah
47 virus, Hendra virus, rabies virus, Ebola virus, etc. ⁵⁻⁷. Besides bat, a considerable number of
48 wild animals have played a pivotal role in zoonotic transfers ⁸. According to reports before
49 SARS-CoV-2 pandemic, due to human interventions, there was a high-risk assessment of
50 SARS coronavirus infection from wild animals like bats, civets, pangolins, snakes, tiger and
51 primates in China ⁹⁻¹¹.

52 The human-wildlife interface as a part of culture or globalization poses risks for zoonotic
53 transfers followed by disease outbreaks like coronavirus outbreaks: SARS (2002,2003),
54 MERS (2012), and SARS-CoV-2 (2019). Animal reservoirs for such outbreaks are estimated
55 by their genome similarities with already reported SARS viruses from diverse animals. For
56 instance, the SARS 2003 outbreak virus had 99.6% genome similarity with palm civets
57 indicating it to be a direct source. Just 0.4% divergence from the animal reservoir stipulates
58 its recent transfer into the masked palm civet population ¹². Despite genetic diversity with bat
59 SARS-CoV they were ultimately found to be a source of the pandemic due to no pathogen

60 prevalence in wild civet population and clinical symptom manifestation in civets, unlike bats
61 ⁴. In the current pandemic, there are several theories of the origin of SARS-CoV-2 either
62 from bat, pangolin, dog, or some intermediate host, etc. ^{13, 14}. The closest match to SARS-
63 CoV-2 is RaTG13 (96% identity), isolated from *Rhinolophus affinis* bat ¹⁵, followed by
64 pangolin SARS viruses with 91% identity ¹⁶. As the closest match is just 96%, it has opened a
65 heated debate in the scientific community for its origin, and no direct animal source can be
66 detected.

67 According to genome similarities, SARS-CoV-2 differs from its closest SARS coronavirus by
68 4%, followed by 9% with its next closest relative pangolin. It indicates that the virus has
69 evolved before infecting humans, and there is a missing link between bat/pangolin and
70 humans, which further inflates the argument on the animal source. Nevertheless, another
71 study based on CpG island deficiency in SARS-CoV-2 and canine coronavirus
72 (alphacoronavirus) suggested that dogs may have provided a cellular environment for SARS-
73 CoV-2 evolution into a CpG deficient virus ¹⁷. Hence, they claim dog to be a direct source of
74 the current pandemic, raising a constant debate ¹⁸ ([https://www.linkedin.com/pulse/where-
75 dog-laymans-version-my-mbe-paper-xuhua-xia/](https://www.linkedin.com/pulse/where-dog-laymans-version-my-mbe-paper-xuhua-xia/)). But most other RNA viruses like pestivirus
76 in addition to bat or pangolin SARS-CoV are also depleted in CpG are not included in the
77 study. CpG island deficiency is not a unique feature of dog SARS-CoV and a later study
78 contradicted that there is no direct evidence for the role of dogs as intermediate hosts ¹⁸.

79 Usage patterns of synonymous codons are a critical feature in the adaptation of organisms as
80 viruses are dependent on the host tRNA pool for replication and disease manifestations. For
81 instance, codon adaptation indices were studied for retroviruses infecting humans, including
82 the HIV-1 virus ¹⁹. Once the viral genome is in the host translational mechanism, genes
83 having optimized codons according to the host translate faster, resulting in higher fitness of
84 the virus ²⁰. Hence, for host jump events, viral codon optimization based on the host tRNA
85 pool is critical ²¹⁻²³. In the present study, we have focused on the codon usage pattern (CUP)
86 of CDS of SARS coronavirus from different hosts under debate (bat, pangolin, and dog) as a
87 probable origin for SARS-CoV-2. An optimum CUP is vital in its evolution, and probable
88 host jumps, and this also results in synonymous changes in the viral genome, which are not
89 revealed by mutational studies at protein level. Population based mutational analysis of
90 SARS-CoV-2 at nucleotide level have revealed various silent mutations conserved in the
91 genome ^{24, 25}. These silent mutations may have consequent alteration in codon usage or
92 translation efficiency (Mercatelli and Giorgi 2020). Systematic insight into CUP is required

93 to trace the evolutionary trajectory, understand its origin, and remarkable success of emergent
94 viruses like SARS-CoV-2. For a virus to be successful, it should be able to efficiently
95 transmit to the host, i.e., recognize host (SARS-CoV-2 spike recognizing human ACE2), and
96 once inside the host, it should replicate (RNA dependent RNA polymerase rdrp (ORF 1ab))
97 its ORFs (ORF 1ab, spike (S), ORF 3a, envelope (E), membrane (M), ORF 6, ORF 7a, ORF
98 7b, ORF 8 and nucleocapsid (N)). The rdrp and spike are now considered important targets
99 for vaccine development for SARS-CoV-2²⁶⁻²⁸. Evolutionary studies till now suggest that the
100 spike receptor-binding domain of SARS-CoV-2 is more similar to pangolin SARS strains as
101 compared to bat SARS¹⁸.

102 Presently, we have analyzed CUP of coding regions in SARS coronavirus isolates reported
103 from humans, bat, pangolin, and dog. Here, we have calculated the percentage of GC biased
104 synonymous codons for amino acids having at least four synonymous codons (Glycine,
105 Valine, Threonine, Leucine, Arginine, Serine, Proline and Alanine)²⁹. Patil et. al, have
106 proven how CUP can detect horizontally acquired genes from a diverse background under
107 selection pressure by analyzing codon usage pattern of each amino acid in a particular gene in
108 a graphical way. Similarly, a host jump event may lead to codon optimization, which will be
109 reflected in the CUP. Ideally, for an organism, its crucial genes should have a similar pattern
110 of CUP. Nevertheless, genes pivotal for viral host jump and disease manifestations like spike
111 or rdrp may show deviation from the pattern. Hence, CUP graphs enable us to visually
112 inspect the patterns of synonymous changes across diverse hosts and be suitable for
113 addressing the surprising origin of the virus.

114 Interestingly, CUP for all the CDS for 134 SARS-CoV-2 genomes (supplementary table 1)
115 was not diversified irrespective of their diverse phylogenetic lineages known in the
116 population. This indicates recent and one-time introduction of an isolate into the human host.
117 While diversity in phylogeny as seen by major and minor lineages suggests that even silent or
118 synonymous mutations play an important role in the rapid emergence and spread. In this
119 context, it is pertinent to note that any mutation can have a consequence in virus. It is
120 dependent on tRNA pool of host that are biased towards a particular set of degenerate codons
121 for a particular amino acid. In fact, a silent mutation can be lethal for a virus if matching
122 tRNA is not encoded in the genome of host or absent in a particular cell or tissue. Further,
123 studies in this regard need of the hour to understand this silent co-evolution in viruses in
124 general and SARS-CoV-2 in particular. On the other hand, the CUPs for the CDSs for other
125 probable hosts i.e. bat, dog and pangolins were diversified (as depicted from standard

126 deviation bars in figure 1). Unlike SARS-CoV-2 isolates of humans, CUP of all CDS were
127 variable in isolates of non-human hosts like a bat, pangolin, and dog (supplementary table 2)
128 depicting ongoing adaptation and evolution of SARS in these hosts (figure 1). Amongst the
129 non-human hosts, bat has most varied CUP correlating with the well-known fact that bat is a
130 reservoir of SARS coronaviruses.

131 Overall, CUP of SARS-CoV-2 for ORF 1ab, envelope and ORF 6, were overlapping with
132 that of SARS from non-human hosts i.e., bat and pangolin, with some exceptions. For
133 instance, ORF 1ab has overlapping pattern for SARS of pangolin origin with human SARS-
134 CoV-2. Here, bat SARS also had similar pattern with SARS-CoV-2 with slightly higher
135 fractions of codon usage for leucine and proline. Envelope protein had overlapping patterns
136 of CUP of SARS-CoV-2 with SARS from pangolin and bat except for a slightly higher
137 fraction of serine (bat SARS) and valine (bat and pangolin SARS). In case of ORF 6 also
138 CUP of bat and pangolin have overlapping patterns with SARS-CoV-2. Here, pangolin SARS
139 ORF 6 did not have arginine codons ending with G or C while, bat and SARS-CoV-2 had the
140 maximum fraction (i.e. 1) of these. Further, CUP of SARS-CoV-2 for spike, ORF 7a, ORF 7b
141 and nucleocapsid proteins were having similar pattern with that of SARS from bat or
142 pangolin. However, CUP for ORF 3a, membrane and ORF 8 had distinct CUP patterns for
143 SARS-CoV-2 compared with that of bat and pangolin. However, CUP of SARS from dog had
144 distinct patterns for the CDS analysed in the study, clearly overruling dog as a probable
145 source compared to bat and pangolin.

146 CUP of all CDS among lineages A and B were not diversified, indicating a single event of
147 transmission of a codon-optimized SARS strain to the human population from its source.
148 Further, CUP of SARS-CoV-2 is not showing congruency with its non-human natural
149 counterparts. Hence, in the current study, we could not find closest relative of SARS-CoV-2
150 in natural settings which is in accordance to the previous genome similarity assessment. CUP
151 pattern of ORF 1ab, envelope protein and ORF 6 is overlapping and spike protein, ORF 7a,
152 ORF 7b and nucleocapsid protein is showing similar pattern, while, CUP of membrane
153 protein, ORF 3a and ORF 8 are distinct from SARS of non-human hosts (bat or pangolin). It
154 indicates that the evolution of all CDS is not linked. It can be depicted that either SARS-
155 CoV-2 is a hybrid virus or the closest relative in natural settings is not yet discovered.

156 However, lack of closely related natural source of SARS-CoV-2 have now shifted lab leak
157 theory to the mainstream from the conspiracy theory^{30, 31} Hence, the probable origin of

158 SARS-CoV-2 is a debate between two competing hypotheses of natural or lab leak. In order
159 to find the true origin, we need to include SARS-CoV-2 from primary infection cases.

160 **References**

- 161 1. Cui, J., Li, F. and Shi, Z.-L., Origin and evolution of pathogenic coronaviruses. *Nature Reviews*
162 *Microbiology*, 2019, **17**, 181-192.
- 163 2. Ge, X.-Y., Li, J.-L., Yang, X.-L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W. and
164 Peng, C., Isolation and characterization of a bat sars-like coronavirus that uses the ace2 receptor.
165 *Nature*, 2013, **503**, 535-538.
- 166 3. Hu, B., Zeng, L.-P., Yang, X.-L., Ge, X.-Y., Zhang, W., Li, B., Xie, J.-Z., Shen, X.-R., Zhang, Y.-Z. and
167 Wang, N., Discovery of a rich gene pool of bat sars-related coronaviruses provides new insights into
168 the origin of sars coronavirus. *PLoS pathogens*, 2017, **13**, e1006698.
- 169 4. Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Cramer, G., Hu, Z. and Zhang, H.,
170 Bats are natural reservoirs of sars-like coronaviruses. *Science*, 2005, **310**, 676-679.
- 171 5. Halpin, K., Young, P.L., Field, H. and Mackenzie, J., Isolation of hendra virus from pteropid bats: A
172 natural reservoir of hendra virus. *Journal of General Virology*, 2000, **81**, 1927-1932.
- 173 6. Leroy, E.M., Kumulungui, B., Pourrut, X., Rouquet, P., Hassanin, A., Yaba, P., Délicat, A., Paweska,
174 J.T., Gonzalez, J.-P. and Swanepoel, R., Fruit bats as reservoirs of ebola virus. *Nature*, 2005, **438**, 575-
175 576.
- 176 7. Mackenzie, J., Chua, K., Daniels, P., Eaton, B., Field, H., Hall, R., Halpin, K., Johansen, C., Kirkland, P.
177 and Lam, S., Emerging viral diseases of southeast asia and the western pacific. *Emerging infectious*
178 *diseases*, 2001, **7**, 497.
- 179 8. Bengis, R., Leighton, F., Fischer, J., Artois, M., Morner, T. and Tate, C., The role of wildlife in
180 emerging and re-emerging zoonoses. *Revue scientifique et technique-office international des*
181 *epizooties*, 2004, **23**, 497-512.
- 182 9. Bell, D., Robertson, S. and Hunter, P.R., Animal origins of sars coronavirus: Possible links with the
183 international trade in small carnivores. *Philosophical Transactions of the Royal Society of London*
184 *Series B: Biological Sciences*, 2004, **359**, 1107-1114.
- 185 10. Gottlieb, S., Chinese scientists must test wild animals to find the host of sars. 2003.
- 186 11. TANG, H.-w., HUANG, J.-c. and HE, J.-f., Analysis of risk factors of sars coronaryvirus infection
187 among population contacted with wild animals in guangdong province. *China Tropical Medicine*,
188 2006, **3**.
- 189 12. Shi, Z. and Hu, Z., A review of studies on animal reservoirs of the sars coronavirus. *Virus research*,
190 2008, **133**, 74-87.
- 191 13. Paraskevis, D., Kostaki, E.G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G. and Tsiodras,
192 S., Full-genome evolutionary analysis of the novel corona virus (2019-ncov) rejects the hypothesis of
193 emergence as a result of a recent recombination event. *Infection, Genetics Evolution*, 2020, **79**,
194 104212.
- 195 14. Zhang, T., Wu, Q. and Zhang, Z., Probable pangolin origin of sars-cov-2 associated with the covid-
196 19 outbreak. *Current biology*, 2020, **30**, 1346-1351. e1342.
- 197 15. Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B. and Huang,
198 C.-L., A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 2020,
199 **579**, 270-273.
- 200 16. Zhang, T., Wu, Q. and Zhang, Z., Pangolin homology associated with 2019-ncov. *BioRxiv*, 2020.
- 201 17. Xia, X., Extreme genomic cpg deficiency in sars-cov-2 and evasion of host antiviral defense.
202 *Molecular biology evolution*, 2020, **37**, 2699-2705.
- 203 18. Pollock, D.D., Castoe, T.A., Perry, B.W., Lytras, S., Wade, K.J., Robertson, D.L., Holmes, E.C., Boni,
204 M.F., Kosakovsky Pond, S.L. and Parry, R., Viral cpg deficiency provides no evidence that dogs were
205 intermediate hosts for sars-cov-2. *Molecular biology evolution*, 2020, **37**, 2706-2710.

- 206 19. RoyChoudhury, S. and Mukherjee, D., Complex codon usage pattern and compositional features
207 of retroviruses. *Computational mathematical methods in medicine*, 2013, **2013**.
- 208 20. Carbone, A., Codon bias is a major factor explaining phage evolution in translationally biased
209 hosts. *Journal of Molecular Evolution*, 2008, **66**, 210-223.
- 210 21. Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., Das, J., Munjal, A. and Singh,
211 R.K., Analysis of nipah virus codon usage and adaptation to hosts. *Frontiers in microbiology*, 2019,
212 **10**, 886.
- 213 22. Tian, L., Shen, X., Murphy, R.W. and Shen, Y., The adaptation of codon usage of+ ssrna viruses to
214 their hosts. *Infection, Genetics Evolution*, 2018, **63**, 175-179.
- 215 23. Van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L. and
216 Xia, X., Hiv-1 modulates the trna pool to improve translation efficiency. *Molecular biology evolution*,
217 2011, **28**, 1827-1834.
- 218 24. Kumar, S. and Bansal, K., Cross-sectional genomic perspective of epidemic waves of sars-cov-2: A
219 pan india study. *bioRxiv*, 2021.
- 220 25. Mercatelli, D. and Giorgi, F.M., Geographic and genomic distribution of sars-cov-2 mutations.
221 *Frontiers in microbiology*, 2020, **11**, 1800.
- 222 26. Aftab, S.O., Ghouri, M.Z., Masood, M.U., Haider, Z., Khan, Z., Ahmad, A. and Munawar, N.,
223 Analysis of sars-cov-2 rna-dependent rna polymerase as a potential therapeutic drug target using a
224 computational approach. *Journal of translational medicine*, 2020, **18**, 1-15.
- 225 27. Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.-J. and Jiang, S., The spike protein of sars-cov—a target
226 for vaccine and therapeutic development. *Nature Reviews Microbiology*, 2009, **7**, 226-236.
- 227 28. Huang, J., Song, W., Huang, H. and Sun, Q., Pharmacological therapeutics targeting rna-
228 dependent rna polymerase, proteinase and spike protein: From mechanistic studies to clinical trials
229 for covid-19. *Journal of clinical medicine*, 2020, **9**, 1131.
- 230 29. Patil, P.B. and Sonti, R.V., Variation suggestive of horizontal gene transfer at a lipopolysaccharide
231 (lps) biosynthetic locus in xanthomonas oryzae pv. Oryzae, the bacterial leaf blight pathogen of rice.
232 *BMC microbiology*, 2004, **4**, 1-14.
- 233 30. Thacker, P.D.J.b., The covid-19 lab leak hypothesis: Did the media fall victim to a misinformation
234 campaign? 2021, **374**.
- 235 31. Sallard, E., Halloy, J., Casane, D., Decroly, E. and van Helden, J.J.E.C.L., Tracing the origins of sars-
236 cov-2 in coronavirus phylogenies: A review. 2021, 1-17.

237

238 **Figure legends:**

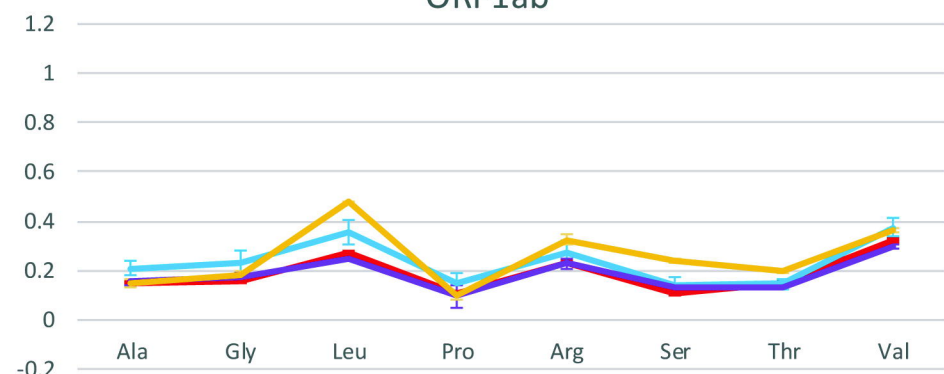
239 **Figure 1:** Codon usage pattern for all CDS of SARS genomes from human (SARS-CoV-2),
240 bat, pangolin and dog. Eight amino acids with at least four synonymous codons are
241 represented in the X-axis, and the percentage of codons ending with G/C for each amino acid
242 is represented on Y-axis. Standard deviations for each amino acid codon usage is represented
243 by vertical error bars.

244 **Supplementary material:**

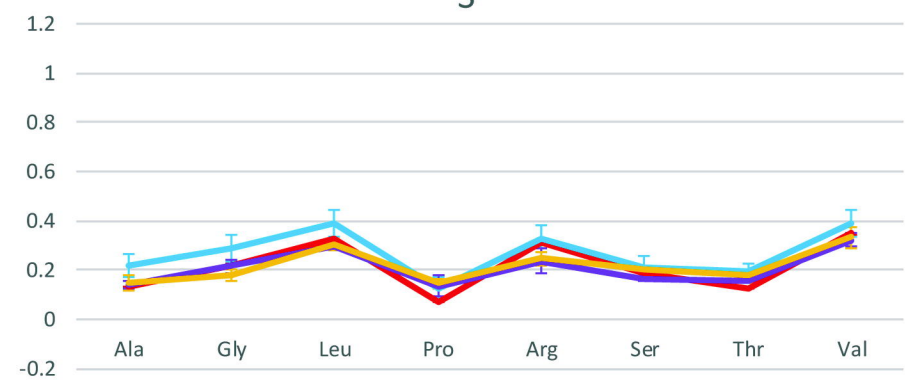
245 **Supplementary table 1:** Metadata for the human SARS-CoV-2 genomes used in the study

246 **Supplementary table 2:** SARS strains from non-human hosts used in the present study.

ORF1ab

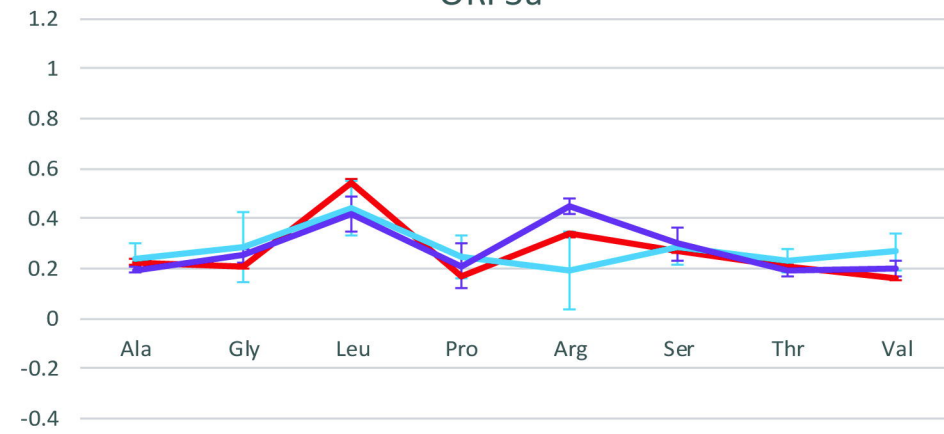


S

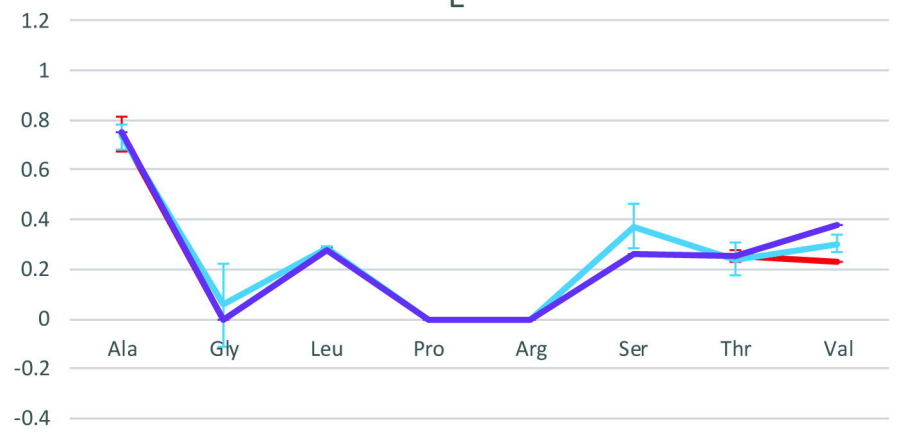


bioRxiv preprint doi: <https://doi.org/10.1101/2020.10.12.335521>; this version posted September 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

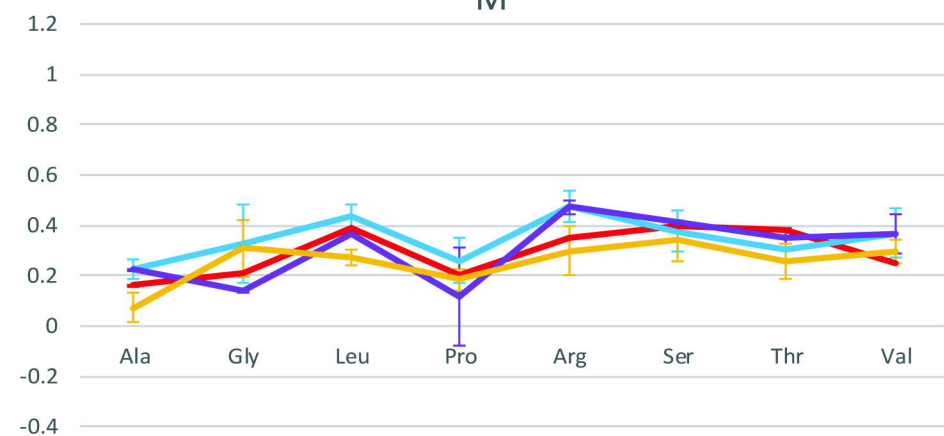
ORF3a



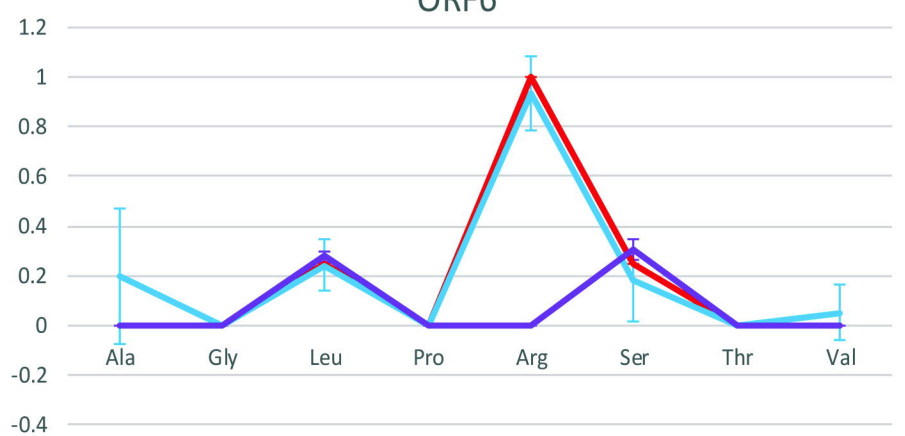
E



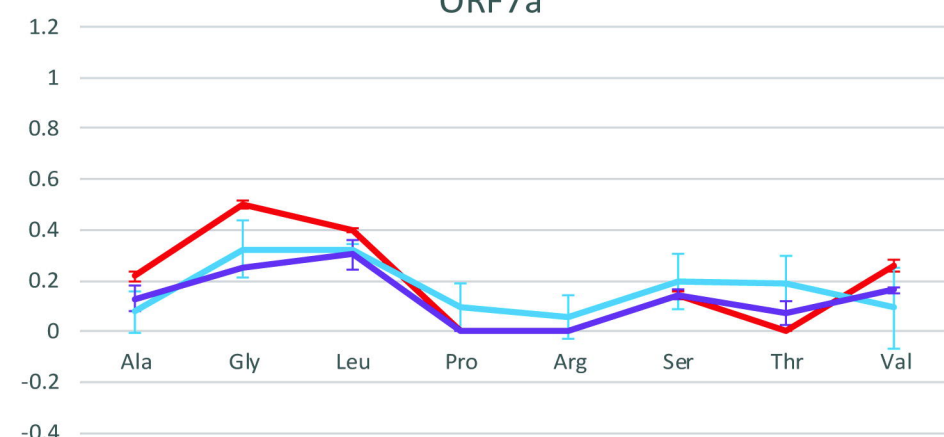
M



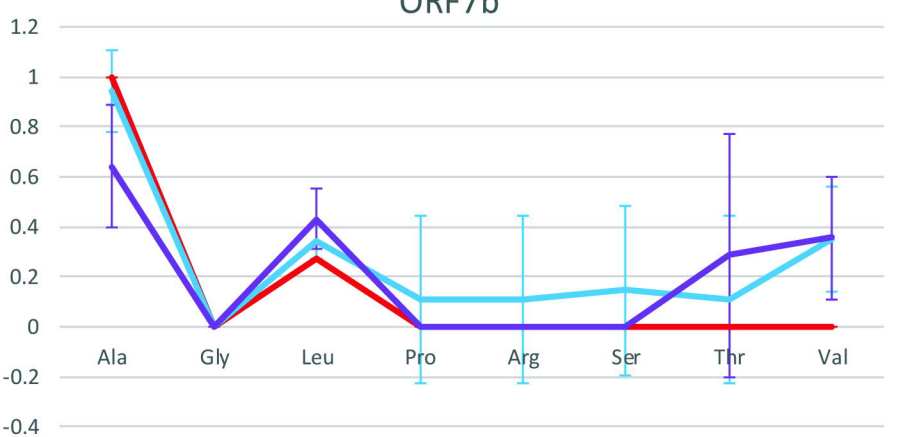
ORF6



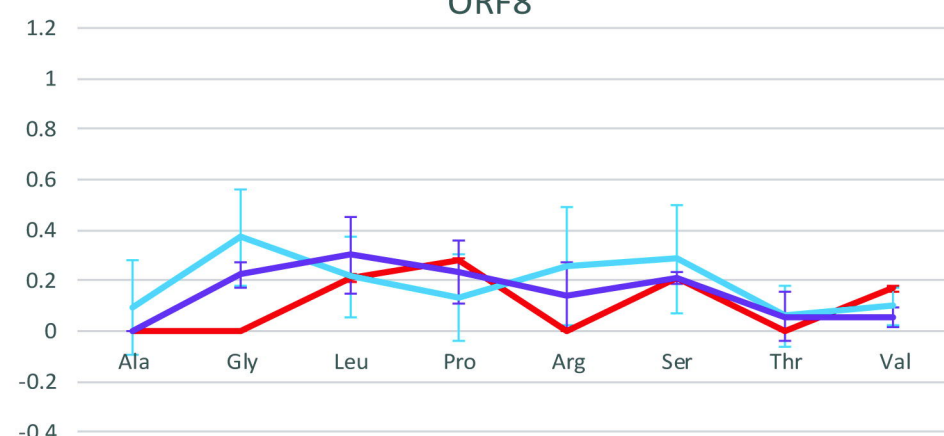
ORF7a



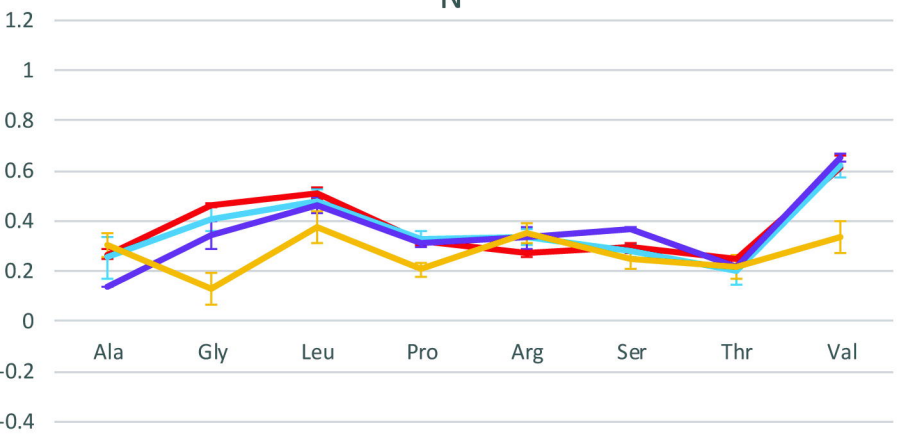
ORF7b



ORF8



N



Human Bat Pangolin Dog