

1 Multivariate Analyses of Codon Usage of SARS-CoV-2 and other  
2 betacoronaviruses

3

4 Haogao Gu<sup>1</sup>, Daniel Chu, Malik Peiris, Leo L.M. Poon<sup>1\*</sup>

5

6 <sup>1</sup>School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong  
7 Kong SAR

8

9 \*Corresponding author: E-mail: [llmpoon@hku.hk](mailto:llmpoon@hku.hk)

10

11 **Abstract**

12 Coronavirus disease 2019 (COVID-19) is a global health concern as it continues to spread  
13 within China and beyond. The causative agent of this disease, severe acute respiratory  
14 syndrome coronavirus 2 (SARS-CoV-2), belongs to the genus *Betacoronavirus* which also  
15 includes severe acute respiratory syndrome related coronavirus (SARSr-CoV) and Middle  
16 East respiratory syndrome related coronavirus (MERSr-CoV). Codon usage of viral genes are  
17 believed to be subjected to different selection pressures in different host environments.  
18 Previous studies on codon usage of influenza A viruses can help identify viral host origins  
19 and evolution trends, however, similar studies on coronaviruses are lacking. In this study,  
20 global correspondence analysis (CA), within-group correspondence analysis (WCA) and  
21 between-group correspondence analysis (BCA) were performed among different genes in  
22 coronavirus viral sequences. The amino acid usage pattern of SARS-CoV-2 was generally  
23 found similar to bat and human SARSr-CoVs. However, we found greater synonymous  
24 codon usage differences between SARS-CoV-2 and its phylogenetic relatives on spike and  
25 membrane genes, suggesting these two genes of SARS-CoV-2 are subjected to different  
26 evolutionary pressures.

27

28 **Keywords:** SARS-CoV-2; coronavirus; codon usage analysis; WCA.

29

## 30 **Introduction**

31 A novel coronavirus outbreak took place in Wuhan, Hubei province, China in December  
32 2019<sup>1</sup>. This novel coronavirus (SARS-CoV-2) causes pneumonia in patients<sup>2</sup> and it has  
33 rapidly spread to other provinces in China and other countries<sup>3</sup>. This novel coronavirus  
34 outbreak had raised global concern but current knowledge on the origin and transmission  
35 route of the pathogen is still limited. The SARS-CoV-2 belongs to the genus *Betacoronavirus*,  
36 which also includes two highly virulent human coronaviruses, SARS-CoV and MERS-CoV.  
37 Apart from human, many animal species, such as bat, rat, camel, swine and hedgehog, can be  
38 infected by different types of coronaviruses. Further sequence analyses of this novel and  
39 other betacoronaviruses might provide additional information to better understand the  
40 evolution of SARS-CoV-2.

41 Preferential codon usage is commonly seen in different organisms, and it has been evident  
42 that the uneven codon usage is not neutral but related to gene expression or other selection  
43 pressures<sup>4-6</sup>. There are two levels of codon usage biases, one is at amino acid level and the  
44 other is at synonymous codon level. The amino acid composition of proteins can be an  
45 important factor that explaining certain sequence traits. For example integral membrane  
46 proteins that are enriched in hydrophobic amino acids can create significant codon usage  
47 bias<sup>7</sup>. Amino acid composition sometime can also introduce confounding effects when one  
48 only focuses on studying the variations of synonymous codon usage. The use of global  
49 correspondence analysis (CA) and its derivatives within-group correspondence analysis  
50 (WCA) and between-group correspondence analysis (BCA) to analyze codon usages can  
51 overcome the above problem. In fact, WCA becomes “model of choice” for analyzing  
52 synonymous codon usage in recent years, as it is more robust than other traditional methods  
53 (e.g. CA with relative codon frequency or CA with RSCU values)<sup>7,8</sup>. This analytic approach,  
54 however, has not been used in studying viral sequences. As the natural history of the SARS-  
55 CoV-2 remains largely unknown, an in-depth codon usage analysis of this newly emerging  
56 virus might provide some novel insights.

57 In this study, we used both CA and WCA to analyses codon usage patterns of a vast number  
58 of betacoronavirus sequences. We found SARS-CoV-2 and bat SARSr-CoV have similar  
59 amino acid usage. However, our analyses suggested that the spike and member genes of  
60 SARS-CoV-2 have rather distinct synonymous codon usage patterns.

61 **Methods**

62 **Sequence data**

63 To construct a reference sequence dataset, available full-length complete genome sequences  
64 of coronavirus were collected through Virus Pathogen Resource database  
65 (<https://www.viprbrc.org/brc/home.spg?decorator=corona>, accessed 13 Jul 2019, ticket  
66 958868915368). The sequences were filtered by the following steps: (1) Remove sequences  
67 without protein annotation, (2) Keep only sequences with complete set of desired replicase  
68 and structural proteins (sequences coding for orf1ab, spike, membrane and nucleocapsid), (3)  
69 Filter out sequences that are unusually long and short (>130% or <70% of the median length  
70 for each group of gene sequences), (4) Limit our analysis to genus *Betacoronavirus* and (5)  
71 Concatenate orf1a and orf1b sequences to form orf1ab if necessary.

72 The final dataset comprised 769 individual strains (3076 individual gene sequences) that  
73 contain complete sets of coding regions for orf1ab, spike, membrane and nucleocapsid genes  
74 (see Supplementary Figure 1). The sequences for envelope gene were not included in the  
75 analysis because of the short length and potential bias in codon usage. Corresponding  
76 metadata for the sequences were extracted by the sequence name field. 24 complete genome  
77 sequences of the newly identified SARS-CoV-2 and its phylogenetically close relatives were  
78 retrieved from Genbank and GISAID (accessed 22 Jan 2020). Six genomes in this study were  
79 used as special references (BetaCoV/bat/Yunnan/RaTG13/2013|EPI\_ISL\_402131;  
80 BetaCoV/pangolin/Guangxi/PIE/2017|EPI\_ISL\_410539; MG772934.1\_Bat\_SARS-  
81 like\_coronavirus\_isolate\_bat-SL-CoVZXC21; MG772933.1\_Bat\_SARS-  
82 like\_coronavirus\_isolate\_bat-SL-CoVZC45;  
83 KY352407.1\_Severe\_acute\_respiratory\_syndrome-related\_coronavirus\_strain\_BtKY72 and  
84 GU190215.1\_Bat\_coronavirus\_BM48-31/BGR/2008), as they have previously been reported  
85 to have close phylogenetic relationship with SARS-CoV-2<sup>9-11</sup>. Detailed accession ID for the  
86 above data are provided in the Supplementary Table S1.

87 The codon count for every gene sequence input for the correspondence analysis was  
88 calculated by the SynMut<sup>12</sup> package. The implementation of the different correspondence  
89 analyses in this study was performed by functions in the package ade4<sup>13</sup>. Three stop codons  
90 (TAA, TAG and TGA) were excluded in the correspondence analysis.

## 91 **Global correspondence analysis (CA) on codon usage**

92 Correspondence analysis (CA) is a dimension reduction method which is well suited for  
93 amino acid and codon usage analysis. The concept in correspondence analysis is similar to  
94 Pearson's  $\chi^2$  test (i.e., the expected counts are calculated under the hypothesis of  
95 independence, based on the observed contingency table). With the deduced expected count  
96 table, the Euclidean distance or the  $\chi^2$  distance can be used to evaluate the difference between  
97 two observations. The  $\chi^2$  distance that we are using in the global correspondence analysis is  
98 applied for the row profile (adjusted for the size effect among difference genes) and the  
99 column profile (adjusted for the size effect among difference codons) and therefore the raw  
100 codon count rather than the Relative Synonymous Codon Usage (RSCU) values are more  
101 informative and suitable input for our model. The calculation of the  $\chi^2$  distance is included in  
102 the Supplementary Method.

103 All the correspondence analyses in this study were performed individually for each gene, to  
104 achieve better resolution on gene specific codon usage pattern.

## 105 **Within-group correspondence analysis and between-group correspondence analysis**

106 In contrast to the ordinary correspondence analysis, the within-block correspondence  
107 analysis<sup>14</sup> (WCA) can segregate the effects of different codon compositions in different  
108 amino acids. WCA has been recognized as the most accurate and effective CA method for  
109 studying the synonymous codon usage in various genomic profile<sup>8</sup>. WCA focuses on the  
110 within-amino acid variability, and it technically excludes the variation of amino acid usage  
111 differences. WCA was implemented based on the existing global CA, with additional  
112 information for factoring.

113 Between-group correspondence analysis (BCA) is complementary to WCA; BCA focuses on  
114 the between-group variability. BCA can be interpreted as the CA on amino acid usage. We  
115 used BCA in this study to investigate the amino acid usage pattern in different coronaviruses.

## 116 **Grand Average of Hydropathy (GRAVY) score**

117 Gravy score provides an easy way to estimate the hydropathy character of a protein<sup>15</sup>. It was  
118 used in this study as a proxy to identify proteins that are likely to be membrane-bound  
119 proteins. The GRAVY score was calculated in a linear form on codon frequencies as:

$$s = \sum_{i=1}^{64} \alpha_i f_i$$

120 Where  $\alpha_i$  is the coefficient for a particular amino acid (provided by data *EXP* in *Seqinr*  
121 package<sup>16</sup>) encoded by codon  $i$ ,  $f_i$  correspond to the relative frequency of codon  $i$ .

122

## 123 **Results**

### 124 **General sequence features in *Betacoronavirus***

125 A total of 3,076 individual gene sequences passed the filtering criteria and were included in  
126 this study. Viral sequences from 3 different species (*Middle East respiratory syndrome*  
127 *related coronavirus (MERSr-CoV)*, *Betacoronavirus 1*, *SARS related coronavirus (SARSr-*  
128 *Cov*)) were the three most dominant species (see Supplementary Figure S1) in the filtered  
129 dataset.

130 Four conserved protein sequence encoding regions of *Betacoronavirus* were analysed  
131 separately. The median lengths of the studied sequence regions were 21237 nt for orf1ab  
132 gene, 4062 nt for spike gene, 660 nt for membrane gene and 1242 nt for nucleocapsid gene.  
133 Spike gene has the lowest average and median G + C contents among these four genes  
134 (median: 37.45%, 37.31%, 42.60% and 47.22% for orf1ab, spike, membrane and  
135 nucleocapsid respectively). The G + C contents of the orf1ab and spike genes were found  
136 distributed in bi-modal patterns, and the G + C contents of SARS-CoV-2 were found located  
137 at the lesser half of the data of these two genes. The G + C contents for membrane and  
138 nucleocapsid genes of studied viral sequences were distributed in unimodal pattern (see  
139 Supplementary Figure S2).

140 The overall amino acid and codon usage of the dataset are plotted in an ascending order  
141 (Figure 1). We observed that leucine and valine were the two most frequently used amino  
142 acids in the four studied genes, while tryptophan, histidine and methionine were the three  
143 least used ones. We also found that codons ending with cytosine or guanine were generally  
144 less frequent than the codons ending with adenine or thymine. This pattern of uneven usage  
145 in synonymous codons is in accordance with the G + C content distribution results (codons  
146 ending with guanine or cytosine were less frequently observed).

147 We found a substantial bias in amino acid usage among these four genes, and this bias is well  
148 explained by the hydropathy of the encoded proteins (results from global correspondence  
149 analysis on all the four genes, collectively, data not shown). The GRAVY scores for every  
150 sequence were calculated to represent the degree of hydropathy. We discovered that the

151 nucleocapsid protein sequences had significantly lower GRAVY scores as compared to those  
152 from other genes, while the membrane protein sequences had highest GRAVY scores (see  
153 Supplementary Figure S3).

#### 154 **Correspondence analysis**

155 We first conducted a multivariate analysis of codon usage on the dataset by using global  
156 correspondence analysis. We also conducted WCA and BCA to study these sequences at  
157 synonymous codon usage and amino acid usage levels, respectively. Given that there were  
158 different amino acid usage biases among different genes (Supplementary Figure S3), we  
159 performed correspondence analyses of these genes separately.

160 Of all the four correspondence analyses for the four genes, the extracted first factors  
161 explained more than 50% of the total variance (see Supplementary Figure S4). The first two  
162 factors in orf1ab global CA represented 67.7% and 16.8% of total inertia. Similarly, the first  
163 two factors of the spike, membrane and nucleocapsid global CA represented 51.0% and  
164 18.5%, 52.6% and 20.2%, and 54.8% and 14.2%, respectively, of total inertia. With only  
165 these two factors, we could extract ~70% of the variability of the overall codon usage for  
166 each studied gene. These levels of representations were higher than or similar to those  
167 deduced from other codon usage analyses<sup>8,17,18</sup>.

#### 168 **The overall codon usage of SARS-CoV-2 in orf1ab, spike and membrane genes are** 169 **similar to those of bat and pangolin CoVs**

170 Based on the above CA analysis, the data points are shown in different colours that represent  
171 different features of the sequences (e.g. viral host or viral species). There were no  
172 neighbouring human viruses around SARS-CoV-2 in CA results of orf1ab, spike and  
173 membrane (Figure 2), suggesting that the overall codon usage of SARS-CoV-2 in the orf1ab,  
174 spike or membrane gene was significantly different from those of human betacoronaviruses.  
175 By contrast, the nucleocapsid genes of SARS coronavirus and SARS-CoV-2 are found to be  
176 relatively similar (Supplementary Figure S5A). Except for the nucleocapsid gene, virus  
177 sequences adjacent to the SARS-CoV-2 were all from bat coronaviruses (coloured in purple  
178 in Figure 2).

179 There are five groups of viral sequences of human origin in the dataset (SARS-CoV-2,  
180 Betacoronavirus 1, human coronavirus HKU 1, MERS-CoV and SARS-CoV). These five  
181 groups of viral sequences were well separated from each other in terms of codon usage,  
182 except the nucleocapsid gene sequences of SARS-CoV-2 and SARS-CoV as mentioned

183 above. There was no overlap between SARS-CoV-2 and human SARS-CoV in orf1ab, spike  
184 and membrane, yet SARS-CoV codon usage processed more similar to SARS-CoV-2  
185 compared to the other three types of human coronaviruses (i.e. yellow point always closest to  
186 SARS-CoV-2 in Supplementary Figure S5A).

187 Compared to human coronavirus sequences, the bat coronavirus sequences have more  
188 scattered codon usage, even within the same viral species (Supplementary S5B). Some viral  
189 species in bats formed their own clusters in all four genes (e.g. SARSr-CoV). SARSr-CoV is  
190 a group of coronavirus that can be found in both humans and bats. We observed that the data  
191 points of human SARSr-CoV are clustered with those of bat SARSr-CoV in all the four genes  
192 (by comparing the yellow points in Supplementary Figure S5A and S5B). The codon usage of  
193 SARS-CoV-2 in orf1ab, spike and membrane were slightly different from the SARS-CoV  
194 clusters and these data points are located in between SARSr-CoV and other coronavirus  
195 species (e.g. MERSr-CoV and bat coronavirus HKU9 etc.)

196 The global codon usages of bat RatG13 virus were found most similar to SARS-CoV-2 in  
197 orf1ab, spike and nucleocapsid genes, but not in membrane gene (Figure. 2). In the analysis  
198 of membrane protein, pangolin P1E virus had a more similar codon usage to SARS-CoV-2  
199 than all the other viruses. We found the similarity in codon usage between pangolin P1E and  
200 SARS-CoV-2 were also high in orf1ab, where P1E was the second closest data point to  
201 SARS-CoV-2. But this is not the case for spike and nucleocapsid genes.

202 We also observed that the codon usage pattern in spike gene was more complex than in other  
203 genes. For example, data points adjacent to the spike gene of SARS-CoV-2 were  
204 coronaviruses from bat, human and rodent hosts (Figure 2). The codon usage of rodent  
205 coronaviruses was generally distinct from human or bat coronaviruses in orf1ab, membrane  
206 and nucleocapsid gene sequences. By contrast, the spike gene sequences of murine  
207 coronaviruses were found located between SARSr-CoV and other coronaviruses, just like  
208 SARS-CoV-2 (Figure 2 and Supplementary Figure S6B). The codon usage from camel, swine  
209 and other coronaviruses were found to be well clustered and relatively distant to SARS-CoV-  
210 2 (see Supplementary Figure S6A, S5C, S5D).

### 211 **The codon usage at synonymous level suggested novel patterns of SARS-CoV-2 in spike** 212 **and membrane genes**

213 WCA and BCA were used to further differentiate codon usage of these betacoronaviruses at  
214 synonymous codon usage and amino acid usage levels, respectively. After applying the row-

215 block structure to the original global CA model, we found that most of the variability in  
216 codon usage can be explained at synonymous codon usage level (90.36% for orf1ab gene,  
217 85.29% for spike gene, 83.71% for membrane gene and 84.07% for nucleocapsid gene) (Table  
218 1).

219 Results from the BCA suggested that the amino acid usage of SARS-CoV-2 is closely related  
220 to bat and human SARS-CoVs in all four genes (Figure 3B and Figure 4B). Specifically, we  
221 discovered that the SARS-CoV-2 had amino acid usage pattern most similar to bat RaTG13  
222 virus, followed by pangolin P1E, bat CovVZC45 and bat CoVZXC21. The sequences of  
223 BtKY72 and BM48-31 were from a more phylogenetically distant clade, and, accordingly,  
224 they had relatively distinct amino acid usage to SARS-CoV-2 as expected in all four studied  
225 genes. This result agrees with the result in the full-genome phylogenetic analysis  
226 (Supplementary Figure S7).

227 The difference between SARS-CoV-2 and RaTG13 at synonymous codon usage level was  
228 marginal in orf1ab and nucleocapsid sequences. Interestingly, there were noticeable  
229 differences in the spike and membrane gene analyses. Our results suggest the synonymous  
230 codon usage patterns in the spike and membrane gene of SARS-CoV-2 are different from  
231 those of its genetically related viruses (i.e. RaTG13 and other reference relatives). For  
232 example, the synonymous codon usage pattern of SARS-CoV-2 was found to be closer to a  
233 cluster of rodent murine coronaviruses at the first two factorial levels (Figure 3A and Figure  
234 4A).

235 Further analysis on spike gene, however, suggested that the codon usage of SARS-CoV-2 and  
236 rodent murine coronaviruses were distinct at the third factorial level (Supplementary Figure  
237 S8A). The results show that although RaTG13 was not the point most adjacent to SARS-  
238 CoV-2 at the first and second dimension, it surpassed murine coronaviruses at the third  
239 dimension. Our results suggest a complex genomic background in the spike gene of SARS-  
240 CoV-2, which made its synonymous codon usage harder to differentiate from other genomic  
241 sequences in our WCA analysis. Despite the proximity between RaTG13 and SARS-CoV-2  
242 at three-dimensional level, they were still formed into two separated clusters (Supplementary  
243 Figure S8A). It is evident that the synonymous codon usage pattern of SARS-CoV-2 is  
244 distinct from other bat origin coronaviruses. The difference in synonymous codon usage is  
245 largely explained by the first factor (more than 50%), and our analysis on codon usages  
246 suggest that the first factor maybe highly related to the preferential usage of codons ending

247 with cytosine (Supplementary Figure S9). We also had similar observation for the membrane  
248 gene. Our three-dimensional analysis revealed that the synonymous codon usage of SARS-  
249 CoV-2 in membrane was most similar to P1E and CoVZXC21 (Supplementary Figure S8B).  
250 It is worth noting that comparing to RaTG13, P1E and CoVZXC21 had lower synonymous  
251 codon usage similarity to SARS-CoV-2 in the other three genes.

252 Overall, our WCA results support a more complex synonymous codon usage background on  
253 spike and membrane genes, though we identified unique codon usage patterns of SARS-CoV-  
254 2 on these two genes.

255

## 256 **Discussion**

257 Codon usage can be affected by many sequence features, including nucleotide composition,  
258 dinucleotide composition, amino acid preference, host adaption, etc<sup>8,19,20</sup>. The codon usages  
259 of viral sequences can vary by genes and host origins<sup>21-23</sup>. The bias in codon usage is a  
260 unique and distinctive characteristic that can reflect the “signature” of a genomic sequence.  
261 Codon usage analyses are often complementary to ordinary sequence alignment-based  
262 analyses which focus on the genetic distance at nucleotide level, whereas codon usage  
263 analyses enable capturing signals at different sequence parameters. Therefore, codon usage  
264 bias can be another good proxy for identifying unique traits (e.g. virus origin, host origin, or  
265 some functions of proteins) of a genome. The goal of this study was to investigate the codon  
266 usage bias of betacoronaviruses. By studying the codon usages of these viruses in a systematic  
267 manner, we identified viral sequences carrying traits similar to those of SARS-CoV-2, which  
268 provided useful information for studying the host origin and evolutionary history of SARS-  
269 CoV-2.

270 The codon usage of different genes in betacoronaviruses are very different. The G+C content,  
271 especially the GC3 content is known to be influential to the codon usage of some bacteria and  
272 viruses<sup>7,24,25</sup>. The GC3 content has pronounced effects on our WCA analysis of the orf1ab  
273 and spike genes. The GC3 content was found correlated with high WCA values on the first  
274 factor of orf1ab (Supplementary Figure S9). By contrast, codons ending with cytosine had  
275 lower factorial values in the spike gene analysis (Supplementary Figure S9). The G + C  
276 contents in membrane and nucleocapsid genes were less suppressed (Supplementary Figure  
277 S2). This can be partly explained by the fact that membrane and nucleocapsid are two genes  
278 with shorter lengths which may limit the flexibilities for mutation or codon usage adaptation.

279 In addition to global CA analysis, the application of WCA and BCA can eliminate the effects  
280 caused by amino acid compositions and synonymous codon usage, respectively. These  
281 alternative analytical tools were important to our study. It is because the amino acid  
282 sequences are expected to be more conserved such that they can preserve biological functions  
283 of the translated genes. By contrast, mutations at synonymous level tend to be more frequent,  
284 as most of these codon alternatives do not affect the biological function of a protein.

285 Of all the existing genomes in the dataset, RaTG13 best matched the overall codon usage  
286 pattern of the SARS-CoV-2. Although the SARS-CoV-2 had amino acid usage similar to bat  
287 and human SARSr-CoVs, the synonymous codon usages between them were relatively  
288 different, which indicates similar protein characteristics but maybe different evolutionary  
289 histories. The codon usage of bat coronaviruses are more scattered than coronaviruses of  
290 other hosts. This result agrees with the fact that bat is a major host reservoir of coronavirus<sup>26</sup>,  
291 thus it harbours coronaviruses with more complex genomic backgrounds.

292 SARS-CoV-2 was first identified in human, but its codon usage pattern is very different from  
293 those of other human betacoroanviruses (Supplementary Figure S5A). In fact, the codon  
294 usage at both the amino acid level and synonymous level denote that the orf1ab gene in  
295 SARS-CoV-2 had closest relationship to SARSr-CoV, especially RaTG13. The CoVZX45  
296 and CoVZXC21 had similar amino acid usage but relatively different synonymous codon  
297 usage to SARS-CoV-2 (Figure 3). Besides bat-origin SARSr-CoV, the pangolin P1E also had  
298 similar codon usage to SARS-CoV-2 both at amino acid and synonymous codon levels. The  
299 result in orf1ab is in accordance with the full-genome phylogenetic analysis (Supplementary  
300 Figure S7), showing a close relationship between SARS-CoV-2 and RaTG13 by the overall  
301 backbone of the genome.

302 The S protein is responsible for receptor binding which is important for viral entry. The  
303 genetic variability is extreme in spike gene<sup>27</sup>, and this highly mutable gene may possess  
304 valuable information about recent evolution history. In our results, the synonymous codon  
305 usage of SARS-CoV-2 in spike gene was distinct from those of RaTG13 and other  
306 phylogenetic relatives (Figure 3A), which was not observed in orf1ab or nucleocapsid gene.  
307 Although the codon usage in spike of SARS-CoV-2, RaTG13 and P1E were similar at amino  
308 acid level, the difference at synonymous codon usage level indicates that they are unlikely to  
309 share a very recent common ancestor. It is more likely that SARS-CoV-2, RaTG13 and P1E  
310 might have undergone different evolution pathways for a certain period of time. The amino

311 acid usage of SARS-CoV-2 in membrane was clustered with bat SARSr-CoV, however the  
312 synonymous codon usage of SARS-CoV-2 was still distinct to these bat coronaviruses.  
313 Notably, in membrane gene, pangolin PIE had a more similar synonymous codon usage to  
314 SARS-CoV-2 than RaTG13. These findings suggest that there may be different selection  
315 forces between genes. Our result supports different evolutionary background or currently  
316 unknown host adaption history in SARS-CoV-2. The codon usage of SARS-CoV-2 in  
317 nucleocapsid gene was similar to bat SARSr-CoV both at amino acid level and synonymous  
318 level, suggesting that no highly significant mutation happened in this gene.

319 Codon usage can be shaped by many different selection forces, including the influence from  
320 host factors. Some researchers have hypothesised that the codon usage in SARS-CoV-2  
321 maybe directly correlated to the codon usage of its host<sup>28</sup>. However our recent study on  
322 influenza A viruses implied that these may not be the most influential factors shaping the  
323 codon usage of a viral genome<sup>19</sup>. Our analysis took advantage of the existing genomes of  
324 *Betacoronavirus* to study the complex host effect on codon usage, which warrants more  
325 accurate but relatively conserved estimation.

### 326 **Acknowledgements**

327 We thank researchers for making viral sequences available for public access. We gratefully  
328 acknowledge the Originating and Submitting Laboratories for sharing genetic sequences and  
329 other associated data through the GISAID Initiative, on which this research is based. A list of  
330 the authors can be found in Supplementary Table S2.

### 331 **Funding**

332 This work was supported by Health and Medical Research Fund (Hong Kong) and National  
333 Institutes of Allergy and Infectious Diseases, National Institutes of Health (USA) (contract  
334 HHSN272201400006C).

### 335 **References**

- 336 1. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global  
337 health concern. *Lancet* (2020). doi:10.1016/S0140-6736(20)30185-9
- 338 2. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J.*  
339 *Med.* NEJMoa2001017 (2020). doi:10.1056/NEJMoa2001017
- 340 3. WHO. Novel coronavirus – Republic of Korea (ex-China). Geneva: World Health  
341 Organization. (2020).

- 342 4. Percudani, R. & Ottonello, S. Selection at the wobble position of codons read by the same  
343 tRNA in *Saccharomyces cerevisiae*. *Mol. Biol. Evol.* (1999).  
344 doi:10.1093/oxfordjournals.molbev.a026087
- 345 5. Pepin, K. M., Domsic, J. & McKenna, R. Genomic evolution in a virus under specific  
346 selection for host recognition. *Infect. Genet. Evol.* (2008). doi:10.1016/j.meegid.2008.08.008
- 347 6. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin.*  
348 *Genet. Dev.* (1998). doi:10.1016/S0959-437X(98)80038-5
- 349 7. Perriere, G. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids*  
350 *Res.* **30**, 4548–4555 (2002).
- 351 8. Suzuki, H., Brown, C. J., Forney, L. J. & Top, E. M. Comparison of correspondence analysis  
352 methods for synonymous codon usage in bacteria. *DNA Res.* **15**, 357–365 (2008).
- 353 9. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus:  
354 implications for virus origins and receptor binding. *Lancet (London, England)* (2020).  
355 doi:10.1016/S0140-6736(20)30251-8
- 356 10. Zhou, P. *et al.* Discovery of a novel coronavirus associated with the recent pneumonia  
357 outbreak in humans and its potential bat origin. *bioRxiv* 2020.01.22.914952 (2020).  
358 doi:10.1101/2020.01.22.914952
- 359 11. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins  
360 in southern China. *bioRxiv* 2020.02.13.945485 (2020). doi:10.1101/2020.02.13.945485
- 361 12. Gu, H. & Poon, L. L. Bioconductor - SynMut. (2019). Available at:  
362 <https://doi.org/doi:10.18129/B9.bioc.SynMut>. (Accessed: 24th January 2020)
- 363 13. Dray, S. & Dufour, A. B. The ade4 package: Implementing the duality diagram for ecologists.  
364 *J. Stat. Softw.* **22**, 1–20 (2007).
- 365 14. Benzécri, J. P. Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondance.  
366 *Cah. l'Analyse des données* **8**, 351–358 (1983).
- 367 15. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydrophobic character of a  
368 protein. *J. Mol. Biol.* (1982). doi:10.1016/0022-2836(82)90515-0
- 369 16. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical  
370 Computing Devoted to Biological Sequences Retrieval and Analysis. in *Structural approaches*  
371 *to sequence evolution: Molecules, networks, populations* (eds. Bastolla, U., Porto, M., Roman,  
372 H. E. & Vendruscolo, M.) 207–232 (Springer Verlag, 2007). doi:10.1007/978-3-540-35306-  
373 5\_10

- 374 17. Lobry, J. R. *Multivariate Analyses of Codon Usage Biases. Multivariate Analyses of Codon*  
375 *Usage Biases* (2018). doi:10.1016/c2018-0-02165-9
- 376 18. Zhou, T., Gu, W., Ma, J., Sun, X. & Lu, Z. Analysis of synonymous codon usage in H5N1  
377 virus and other influenza A viruses. *BioSystems* **81**, 77–86 (2005).
- 378 19. Gu, H., Fan, R. L. Y., Wang, D. & Poon, L. L. M. Dinucleotide evolutionary dynamics in  
379 influenza A virus. *Virus Evol.* **5**, (2019).
- 380 20. Hershberg, R. & Petrov, D. A. Selection on Codon Bias. *Annu. Rev. Genet.* (2008).  
381 doi:10.1146/annurev.genet.42.110807.091442
- 382 21. Wong, E. H., Smith, D. K., Rabadan, R., Peiris, M. & Poon, L. L. Codon usage bias and the  
383 evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **10**,  
384 253 (2010).
- 385 22. Cristina, J., Moreno, P., Moratorio, G. & Musto, H. Genome-wide analysis of codon usage  
386 bias in Ebolavirus. *Virus Res.* (2015). doi:10.1016/j.virusres.2014.11.005
- 387 23. Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its  
388 evolutionary origin. *Virus Res.* (2003). doi:10.1016/S0168-1702(02)00309-X
- 389 24. Gu, W., Zhou, T., Ma, J., Sun, X. & Lu, Z. Analysis of synonymous codon usage in SARS  
390 Coronavirus and other viruses in the Nidovirales. *Virus Res.* (2004).  
391 doi:10.1016/j.virusres.2004.01.006
- 392 25. Woo, P. C. Y., Huang, Y., Lau, S. K. P. & Yuen, K. Y. Coronavirus genomics and  
393 bioinformatics analysis. *Viruses* (2010). doi:10.3390/v2081803
- 394 26. Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V. & Schountz, T. Bats: Important  
395 reservoir hosts of emerging viruses. *Clinical Microbiology Reviews* (2006).  
396 doi:10.1128/CMR.00017-06
- 397 27. Gallagher, T. M. & Buchmeier, M. J. Coronavirus spike proteins in viral entry and  
398 pathogenesis. *Virology* (2001). doi:10.1006/viro.2000.0757
- 399 28. Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Homologous recombination within the spike  
400 glycoprotein of the newly identified coronavirus may boost cross-species transmission from  
401 snake to human. *J. Med. Virol.* jmv.25682 (2020). doi:10.1002/jmv.25682

402

403

404 Table 1. Variability explained by the synonymous codon usage level and the amino acid level.

	Orf1ab	Spike	Membrane	Nucleocapsid
WCA (synonymous codon level)	90.36%	85.29%	83.71%	84.07%
BCA (amino acid level)	9.64%	14.71%	16.29%	15.93%

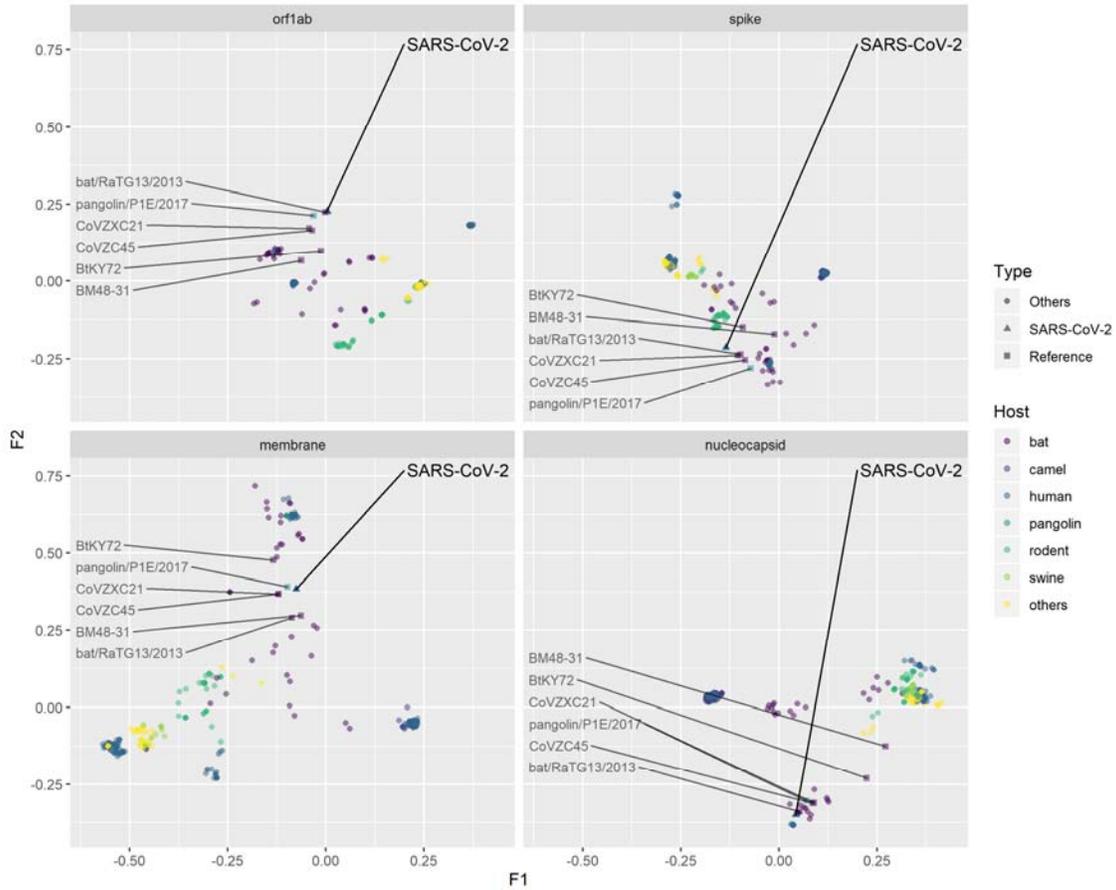
405

406

407 Figure 1. Codon usage in *Betacoronavirus* (Cleveland's dot plot). Points in green showed the count of  
 408 codons in a sample SARS-CoV-2 genome (MN908947).



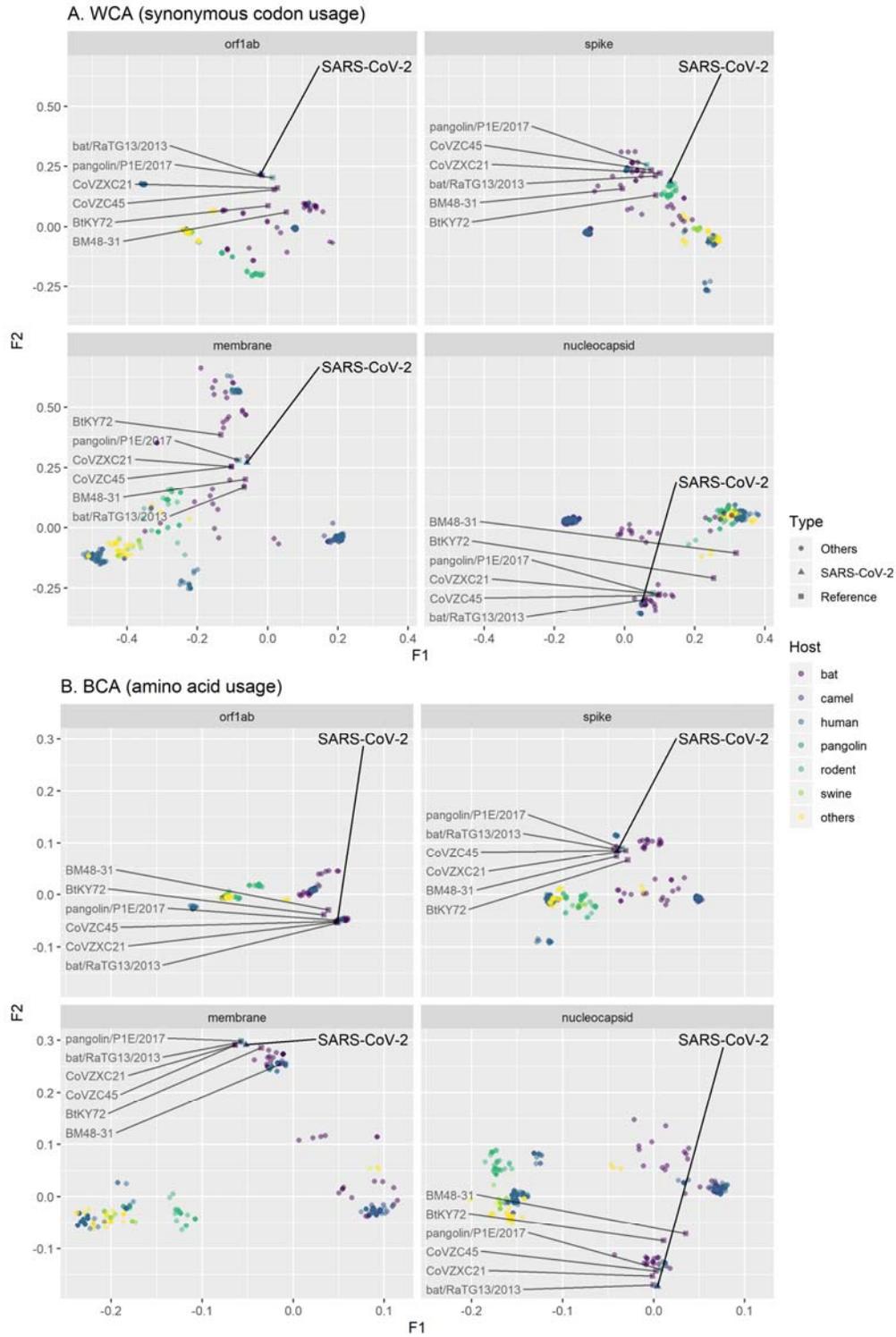
410 Figure 2. Factorial map of the first and second factors for global CA by different genes,  
411 coloured by different viral host. The SARS-CoV-2 and related reference data points were  
412 labelled.



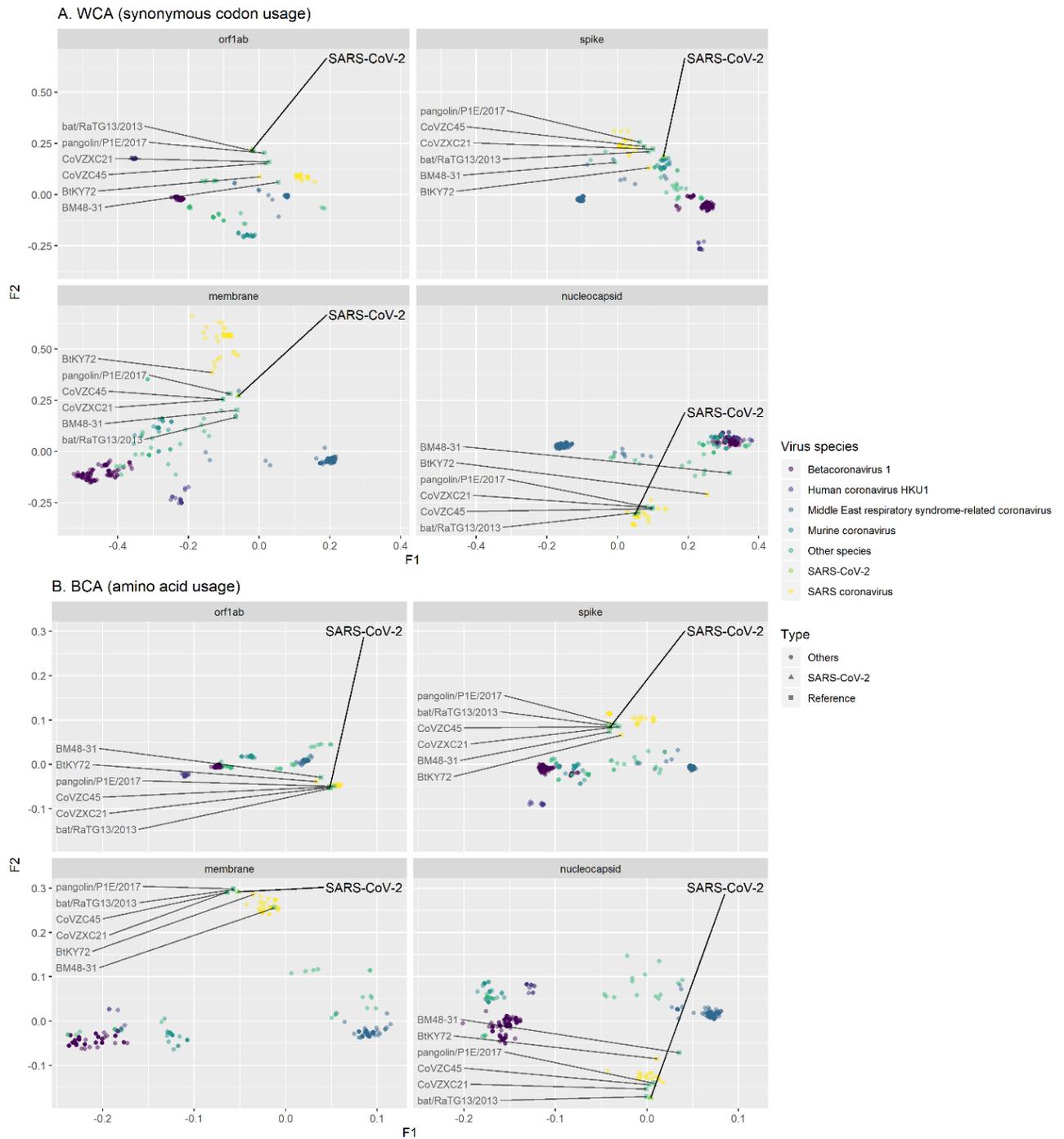
413

414

415 Figure 3. Factorial map of the first and second factors for WCA and BCA by different genes,  
416 coloured by different viral host. The SARS-CoV-2 and related reference data points were  
417 labelled.



419 Figure 4. Factorial map of the first and second factors for WCA and BCA by different genes,  
420 coloured by different viral species. The SARS-CoV-2 and related reference data points were  
421 labelled.



422

423