Phylogenetic tree analysis of SARS-CoV-2 genomes

Zizhuo Wang, Shanghaitech University

Zhen Li, Shanghaitech University

This is a phylogenetic tree of SARS-CoV-2 genomes sampled from across the world during December 2019 to March 2020 together with a coronavirus gene from bat. We used distance-based neighbor-joining algorithnm to build the phylogenetic tree. The approach to distance matrix is Hamming distance.

1 Abstract

In a phylogenetic tree analysis of 160 complete human severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) genomes together with 1 bat coronavirus genomes considered to the probable origin of SARS-Cov-2 [1], we find three central variants, which we have named A, B, and C. The A type is mainly found in East Asia including China mailand, Taiwan, South Korea and Japan. The B type is found in significant proportions outside East Asia, that is, in Europeans and Singapore. The C type is found mainly in Europe and East Asia. In addition, most cases in China mainland have more variants. The tree shows the inovation of Sars-CoV-2 and help to traces routes of infections for documented coronavirus disease 2019 (COVID-19) cases. This indicates that phylogenetic tree can be used to help trace undocumented COVID-19 infection sources, which can then be quarantined to prevent recurrent spread of the disease worldwide.

2 Tree-building Method

Method of building a phylogenetic tree are usually classified to two kind: Distance-based and Character-based. UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and NJ belong to the distance-based method, MP (Maximum Parsimony) and ML(Maximum Likelihood)belongs to the character-based method.

Basically when we analyze populations that have close genetic relationship, the results given by above methods usually have similar structure in topology. Trees build by different method will exist partially difference in topology.

2.1 Comparison between NJ, UPGMA, MP and ML

UPGMA works on the assumption that the evolution rate in different germlines keeps the same, which means it may generate a substantial gap to the practical case when the evolution rate differs between germlines. [2]

MP do not rely on statistic method of any evolution model, it is decided by the minimum mutation number in all ancestral sequence reconstruction. Though whether the mutations are happened by the way of minimum nucleotide substitution is mysterious. Hence, when the number of nucleotides substitution is massive (i.e., the genetic relationships are not close), MP method are likely going wrong.

ML is a method highly depends on the selection of nucleotides substitution model, which usually contains parameters that reflect the evolution process. The rate of nucleotides substitution differs by different sites. It seems that the correctness of topology and the estimation precision of branch length could not be satisfied together. Additionally, ML performs worse when number of groups are big.

NJ is a derived method of ME (minimum evolution). ME choose a tree of minimum heights among all possible trees. Because dealing with all possible trees is a quite hard task, NJ start with a star tree to find the topology structure rather than finding out the optimal solution. [3]

2.2 Details of Neighbor-Join Algorithm

Assume each population are initially adjacent to every other populations, which would show as a star tree. The main idea of NJ is to merge two adjacent nodes which are most 'close' to a new node repeatedly, until a node contains all populations formed. Specific steps are shown below: [4]

0. Start with a "star" tree and a n by n distance matrix.

1. Calculate the Q matrix based on distance matrix.

2. Find the pair i and j for which Q(i, j) has its lowest value. Join i and j to create a node node, which is connected to the central node. Update lengths/weights of new branch (i.e., calculate distance from the pair members to the new node.

3. Update the distance matrix [(n-1) by (n-1)], i.e., calculate the distance from each node to the new node.

4. Repeat 1-3 with the updated distance matrix.

3 Tree Evaluating

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Bootstrapping is conducted using the columns of the character matrix. Each pseudoreplicate contains the same number of species (rows) and characters (columns) randomly sampled from the original matrix, with replacement. A phylogeny is reconstructed from each pseudoreplicate, with the same methods used to reconstruct the phylogeny from the original data. For each node on the phylogeny, the nodal support is the percentage of pseudoreplicates containing that node. Specific steps are shown below: [5]

1.Given an multi-sequence alignment, randomly pick columns from alignment with replacement.

2.Apply tree building to new data set.

3.Repeat selection a number of times.

4.The frequency with which each clade in the original tree is observed is measure of confidence for that clade.

4 Speed-up Method

Multiprocessing is a package that supports spawning processes using an API similar to the threading module. The multiprocessing package offers both local and remote concurrency, effectively side-stepping the Global Interpreter Lock by using subprocesses instead of threads. Due to this, the multiprocessing module allows the programmer to fully leverage multiple processors on a given machine.

Also, due to Global Interpreter Lock in Python, one process has only one thread. Multiprocessing usually have better performance on python. So we choose multiprocessing to speed-up the bootstrapping.

5 Result and Analysis

Zhou et al. recently reported a closely related bat coronavirus, with 96.2 percent sequence similarity to the human virus.Pairwise protein sequence analysis of seven conserved non-structural proteins.In addition, 2019-nCoV virus isolated from the bronchoalveolar lavage fluid of a critically ill patient could be neutralized by sera from several patients. Notably, we confirmed that 2019-nCoV uses the same cell entry receptor—angiotensin converting enzyme II (ACE2)—as SARS-CoV. [1] So we use this bat virus as an outgroup.

The output of the python script is in Newick format. iTOL(http://itol.embl.de/) is used to draw the phylogenetic tree. [6] The tree is shown below. 1

Overall, the tree, as expected in an ongoing outbreak, shows different versions of virus exist at the same time. The bat coronavirus is placed in in a cluster of cases which we have labeled "A." There are two subclusters of A. In the subcluster bat coronavirus placed, most cases are from East Asia. The cases of the A cluster outside East Asia are mainly from the United States and Australia.

There also two other main clusters. We have labeled them as B and C. The cases in B are mainly from Europe and Singapore. C includes cases from East Asia, Europe and Australia.

What is striking in the cluster B is that the phylogenetic tree can illustrate the well documented infection history of certain cases. On 25 February 2020, the first Brazilian was reported to have been infected following a visit to Italy. In the phylogenetic tree, 3 Italian cases are placed next to the Brazilian viral genome.

Similarlly the only case from Mexican in the tree,which is next to cases from Italy, is a documented infection diagnosed on 28 February 2020 in a Mexican traveler to Italy.

For the rest part of the phylogenetic tree, we can only draw a conclusion that the phylogeny becomes obscured by subsequent migration and mutation. The cases has been distant from the bat coronavirus outgroup rooting.

6 Data Availability

In early March 2020, the GISAID database contained a compilation of 253 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) complete and partial genomes contributed by clinicians and researchers from across the world since December 2019. The nucleotide sequences of the SARS-CoV-2 genomes used in this analysis are available, upon free registration, from the GISAID database (https://www.gisaid.org/).

The bat coranavirus genomes is available in the article of Zhou et al. [1]

7 Future Works

There several thing to do to improve our works. In our implement, the distace matrix uses simple Hamming distace. More complicated distance can be used to measure the distance of the sequences better. The Neighbor-Join Algorithm may also be improved to generate phylogenetic tree with higher accuracy and speed.

In addition, it would be more user-friendly if the process of drawing a tree from Newick format can be executed automaticlly.

8 My Thoughts and Gains

In this project, I take charge of the tree evaluating, speed-up and result analysis. I also wrote the main part of the final report.

I not only learn a lot of knowledge of building phylogenetic tree, but also how to analysis tree and trace routes of infections.

I was kind of panicky when our dear instructer Professor Zheng pointed out our analysis of the main trace routes may be wrong because the bat coranavirus origin are not credible. Now we realize that we didn't consider this problem very carefully. Though Zhou et al. gave various evidences to prove the bat coronavirus has a close relationship with Sars-CoV-2, this can be explain as the bat was actually infected by human beings. But I believe our analysis can still illustrate the infection route of several special cases whose infection history is well documented. If I had an opportunity to do more works on this topic, I would like to reconsider the whole process to make it more convincing and more logically rigorous.

References

[1] Zhou, P., Yang, X., Wang, X. et al.(2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273

[2] Edgar R C. (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets[J]. *Nucleic Acids Res* **32**,(1):380 385.

[3] Maier D. (1978) The complexity of some problems on subsequences and supersequences.*ACM* **25**, (7):322 336.

[4] Saitou N, Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-25.

[5] Felsenstein J (1985). "Confidence Limits on Phylogenies: An Approach Using the Bootstrap". *International Journal of Organic Evolution.* **39** (4): 783–791.

[6] Letunic I and Bork P (2019) "Interactive Tree Of Life (iTOL) v4: recent updates and new developments". *Nucleic Acids Res* **47** (W1): W256–W259

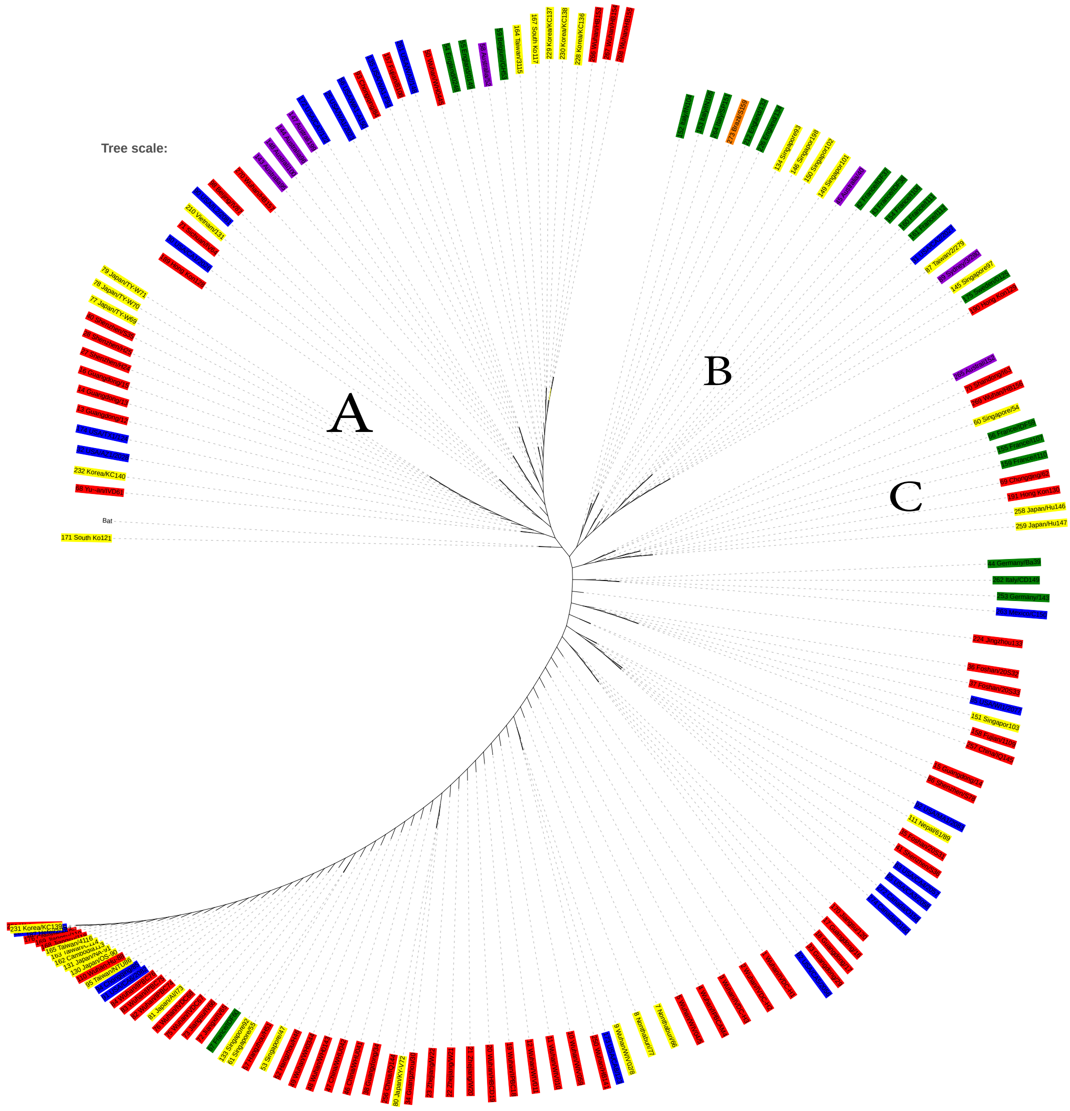
Tree scale:
A
B
C
171 South Ko121
Bat
68 Yu--an/IVD61
232 Korea/KC140
32 USA/AZ1/2028
174 USA/TX1/124
13 Guangdong/12
14 Guangdong/13
16 Guangdong/15
27 Shenzhen/H24
28 Shenzhen/H25
40 Shenzhen/S35
77 Japan/TY-W69
78 Japan/TY-W70
79 Japan/TY-W71
189 Hong Kon128
30 USA/CA1/2026
71 Sichuan/IV64
210 Vietnam/131
90 USA/IL2/2081
98 Beijing/IV87
270 Wuhan/HB157
143 Australia95
148 Australi100
144 Australia96
147 Australi199
172 USA/CA7/122
25 USA/WA1/2023
56 USA/WA1-A150
93 Chongqing/84
135 USA/WA1-F94
157 Fujian/8108
261 USA/WA2/148
50 Wuhan/WH0445
54 England/0248
55 England/0149
58 Australia/52
59 Belgium/GH53
164 Taiwan/3115
167 South Ko117
229 Korea/KC137
230 Korea/KC138
228 Korea/KC136
266 Wuhan/HB153
267 Wuhan/HB154
268 Wuhan/HB155
152 Italy/IN104
153 Italy/IN105
264 Italy/SP151
273 Brazil/S159
212 England/132
226 England/134
134 Singapore93
146 Singapor198
150 Singapor102
149 Singapor101
45 Australia/40
42 France/IDF37
43 France/IDF38
154 France/I106
160 France/I111
161 France/I112
31 USA/CA2/2027
87 Taiwan/2/279
89 Sydney/3/280
145 Singapore97
175 Sweden/0125
190 Hong Kon129
265 Australi152
70 Shandong/I63
269 Wuhan/HB156
60 Singapore/54
66 France/IDF59
155 France/I107
159 France/I110
69 Chongqing/62
191 Hong Kon130
258 Japan/Hu146
259 Japan/Hu147
44 Germany/Ba39
262 Italy/CD149
253 Germany/143
263 Mexico/C150
224 Jingzhou133
36 Foshan/20S32
37 Foshan/20S33
85 USA/WI1/2077
151 Singapor103
158 Fujian/1109
257 China/IQ145
15 Guangdong/14
86 Shenzhen/S78
92 USA/MA1/2083
111 Nepal/61/89
35 Foshan/20S31
41 Shenzhen/S36
63 USA/CA4/2057
65 USA/CA3/2058
272 Canada/O158
274 Canada/O160
170 Jiangsu/120
17 Guangdong/16
18 Guangdong/17
33 Guangdong/29
62 USA/CA5/2056
1 Wuhan/IVDC-H1
2 Wuhan/IVDC-H2
3 Wuhan/IVDC-H3
4 Wuhan/IPBCAM4
5 Wuhan/WIV04/5
7 Nonthaburi/66
8 Nonthaburi/77
9 Wuhan/WIV02/8
173 USA/CA8/123
250 Wuhan/HB141
10 Wuhan/WIV059
11 Wuhan/WIV010
12 Wuhan/WIV011
19 Wuhan/IPBC18
20 Wuhan/HBCD19
21 Zhejiang/W20
22 Zhejiang/W21
23 Zhejiang/W22
34 Guangzhou/30
80 Japan/KY-V72
256 China/IQ144
38 Guangdong/34
46 China/WHU041
47 China/WHU042
48 Wuhan/WH0143
49 Wuhan/WH0344
52 Hangzhou/H46
53 Singapore/47
57 Hangzhou/H51
61 Singapore/55
133 Singapore92
67 France/IDF60
72 Jiangxi/IV65
73 Jiangsu/IV66
75 Wuhan/IVDC67
76 Wuhan/IVDC68
81 Japan/AI/I73
82 Wuhan/IPBC74
83 Wuhan/IPBC75
84 Wuhan/IPBC76
91 USA/CA6/2082
94 Chongqing/85
95 Taiwan/NTU86
110 Wuhan-Hu-88
130 Japan/OS-90
131 Japan/NA-91
162 Cambodia113
163 Taiwan/C114
165 Taiwan/4116
168 Jiangsu/118
169 Jiangsu/119
176 China/
187 Hefei/
231 Korea/KC139

Figure 1