Review of: "Sequence evidence that the D614G clade of SARS-CoV-2 was already circulating in northern Italy in the fall of 2019"

Ales Kovarik[1]

1 Institute of Biophysics ASCR

Potential competing interests: No potential competing interests to declare.

General

This manuscript reports on the origin of human SARS-CoV-2 virus believed to be a cause of Covid-19 disease. It is reported that a major D614G variant which spread rapidly worldwide during the pandemics have already been circulating in European clinical samples collected before the Covid-19 outbreak. This is certainly a highly significant finding while there are also other interpretations of results – please see specific comments as below.

Major issues

1. Interpretation of the Table 1 results is tricky. Virus strains are defined by haplotypes based on a combination of mutations at different positions of virus genomes. Unfortunately, each sequence collected in Italy 2019 contains just a single SNP and haplotypes cannot be reconstructed. Therefore, it is difficult to interpret these sequences as a particular virus lineage, such as the one of D614G.

2. The data are slightly overinterpreted. I would suggest to town down strong statements such as the one in title “Sequence evidence that the D614G clade of SARS-CoV-2 was already circulating in northern Italy in the fall of 2019“. Perhaps, the title could be modified as follows: „Sequence variants specific for major SARS-Cov-2 coronavirus lineages may have been circulating in northern Italy as early as the fall of 2019“. (I am not sure about the English grammar, please check)

3. My other concern are mutations listed in Table 1. All but one mutations are from C-to-T substitutions. It could be mentioned that this substitution is a characteristic feature of coronavirus mutation spectra likely resulting from cytosine deamination events. The C-to-T bias have been observed in early SARS-CoV-2 sequenced genomes at an early phase of world pandemics (G*enes* **2020**, *11*(7), 761; https://doi.org/10.3390/genes11070761). Possibly the observed CCCA>TTTG switch is a gradual process involving several independent C-to-T mutations. The China/Germany TTCG haplotype in the bottom of the phylogeny tree (Figure 1) supports the evolutionary trajectory. If so, one should also consider the reference sequences which has the CCCG haplotype. The claimed Italian 2019 variants actually cast doubts about the reference sequence being the ancestral one.

4. Table 1. Please include nucleotides in the reference*NC_045512. Also missing data /nucleotides- should be marked*

with some symbol, e.g. a dash "-".

5. I understand that the author used published sequences in the data base for the bioinformatic analysis. However, in studies like this, one has to be hundred percent sure that the data base sequences are correct. Samples were stored probably for months before the extraction of RNA and downstream RT-PCR analyses. Can the author exclude contamination of 2019 samples with PCR products? I wish to mention that contaminations are very common problem in molecular PCR analyses. Contaminations of 2019 samples with PCR products carrying virus sequences essentially nullify the results of the study. This is just my assumption while the proof should come from the Italian lab.

6. The phylogeny tree is nice containing data which may illuminate the origin of the virus. I agree. I just wonder what the 0.2 value on a Bar scale represents? Usually, it is defined as a distance calculated from pairwise alignment. If so, the value seems me to too high. Typically, SARS-CoV-2 genomes differ by less than 10-30 mutations per c. 30 kb in pairwise comparisons the corresponds to divergence of less than 0.1% which means a distance value of less than 0.001. Please check.

Minor issues

Table 2 title. G614G should read D614G