Massimo Ciccozzi ORCID iD: 0000-0003-3866-9239

Domenico Benvenuto ORCID iD: 0000-0003-3833-2927

Silvia Angeletti ORCID iD: 0000-0002-7393-8732

Response to Carletti et al., "About the origin of the first two Sars-CoV-2 infections in Italy: inference not supported by appropriate sequence analysis."

Massimo Ciccozzi[1], Marta Giovanetti[2,3], Domenico Benvenuto[1], Silvia Angeletti *[4].

[1]Unit of Medical Statistics and Molecular Epidemiology, University Campus Bio-Medico of Rome, Rome, Italy;

[2]Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil;

[3]Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil;

[4]Unit of Clinical Laboratory Science, University Campus Bio-Medico of Rome, Rome, Italy

Corresponding author

Silvia Angeletti, Unit of Clinical Laboratory Science,

University Campus Bio-Medico of Rome, 00128 Rome, Italy.

email: s.angeletti@unicampus.it

Running head: COVID-2019 Italian epidemic

Dear Editor,

We are writing regarding the Commentary entitled “About the origin of the first two Sars-CoV-2 infections in Italy: inference not supported by appropriate sequence analysis.” from Carletti et al., published in Journal of Medical Virology.

We welcome the discussion on published article, which could improve the science and clarify any controversial issues.

The preliminary results we have shown in our manuscript (doi:10.1002/jmv.25699) entitled “The first two cases of 2019-nCoV in Italy: where they come from?” have been confirmed by our recent publication[1], as well as by the results shared with the scientific committee by the Nextstrain (Real-time tracking of pathogen evolution) Coronavirus report available on https://nextstrain.org/ncov.

I have actually two main points, both methodological, that appear to be crucial talking about.

The first one, as the authors should know, that in order to estimate phylogenies on a natural timescale a ‘molecular clock’ analysis has to be applied. A molecular clock is a theory attributed to Emile Zuckerkandl and Linus Pauling (1962), describing the relationship between observed genetic distances and time. Molecular clock models allow inference to proceed even when the alignments being analysed contain little or no temporal information. Before to start any kind of molecular analysis, researchers must to explore the degree of temporal signal in heterochronous sequences, to justify the use of the molecular clock models. This can be achieved using a simple regression-based approach. At this aim before to perform a Bayesian reconstruction in phylogenetic analysis the temporal signal have to be checked. A statistical regression analysis of genetic divergence from root to tip versus sampling dates must be conducted, to verify if the dataset uses is suitable to perform a Bayesian analysis. By this way, accordingly with Rambaut et al.[2] (doi: 10.1093/ve/vew007) they referred R2 = 0.80 as a 'strong' value for association between genetic distances and sampling dates, also referring to positive correlation, such as R2 = 0.21 and R2 = 0.13, as appearing "to be suitable for phylogenetic molecular clock analysis". In our analysis, despite only the partial sequences of the Chinese couple admited in italian hospital were available, we found a positive correlation in the test to justify the molecular clock analysis (R2= 0.46).

The second point that the authors probably do not know is the use of the Xia test to check the phylogenetic signal before starting the analysis [3]. It is well known, by expert in this field, that the accuracy of phylogenetic reconstruction depends mainly on the correct identification of homologous sites by sequence alignment, absence of heterotachy and little variation in the substitution rate over sites, consistency, efficiency and little bias in the estimation method, e.g. not plagued by the long-branch attraction problem, and sequence divergence, i.e. neither too conserved as nor too diverged to avoid substitution saturation. Substitution saturation is a serious problem in phylogenetic analysis decreasing the phylogenetic information contained in sequences[3,4]. When sequences have experienced saturation, the similarity between the sequences can depend on the similarity in nucleotide frequencies[5-7], which often does not reflect phylogenetic relationships. Xia's method is an index based on computer simulation with different sequence lengths, different number of OTUs. The critical value enables researchers to quickly judge whether a set of aligned sequences can be use in phylogenetic. A simple index of substitution saturation is used and is defined as Iss. When the sequences have experienced severe substitution saturation the index approaches to 1. In theory we need to find the critical Iss value (referred to as Iss.c) at which the sequences will begin to fail to recover the true tree. Once Iss.c is known for a set of sequences, then we can simply calculate the Iss value from the sequences and compare it against the Iss.c. If Iss is not smaller than Iss.c, then we can conclude that the sequences have experienced severe substitution saturation and should not be used for phylogenetic reconstruction. In our case, the Test of substitution saturation, gave us the values of 0.016 and 0.782 for Iss and Issc respectively, indicating absence of saturation and strong phylogenetic signal. Moreover, we did not indicate in the figure and in the method section the label of the sequences because we used all the available sequences in a database at the time of the analysis (as described in the text) to enforce the data set avoiding an eventual scarce phylogenetic signal. At the end, we believe that it is better to write and publish a research article giving new and fruitful information to the scientific community rather than criticized ones.

References

1. Giovanetti M, Angeletti S, Benvenuto D, Ciccozzi M. A doubt of multiple introduction of SARS-CoV-2 in Italy: a preliminary overview. *J Med Virol*. 2020 Mar 19. doi: 10.1002/jmv.25773

2. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2016; 2(1): vew007

3. Xia XH, Xie Z, Kjer KM. 18S ribosomal RNA and tetrapod phylogeny. *Syst Biol.*

2003 52(3): 283-95.

4. Lopez P, Forterre P, Philippe H. The root of the tree of life in the light of the covarion model. *J Mol Evol.* 1999; 49(4):496-508.

5. Lockhart PJ, Penny D, Hendy MD, Howe CJ, Beanland TJ, Larkum AW. Controversy on chloroplast origins. *FEBS Letters* 1992; 301(2): 127-31.

6. Steel MA, Lockhart PJ, Penny D. Confidence in evolutionary trees from biological sequence data. *Nature* 1993; 364(6436): 440-2.

7. Xia X, Xie Z. DAMBE: Software package for data analysis in molecular biology and evolution. *J Hered.* 2001; 92(4):371-3.