Review: Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations (Laamarti *et al*, 2020)

by Fiona Walter

In the paper "Large scale genomic analyses of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations", Laamarti *et al*. (2020) comparatively analyse mutations in 3067 complete SARS-CoV-2 genomes to understand phylogenetic and geographical correlations within the virus' population[1]. The paper was submitted for publishing in May 2020, only two months after the SARS-CoV-2 epidemic was declared a pandemic by the World Health Organisation (WHO)[2]. The authors describe that two similar but smaller studies[3,4] of <100 genomes were a valuable source to understand the viruses' genomic structure, phylogeny and how genetic diversity evolved in this pathogen. Further studies[5] had found that the 30kb SARS-CoV-2 genome consist to two thirds of the ORF (ORF1ab) coding for several non-structural elements, the spike protein, and the viruses RNA polymerase. The authors argue that due to these previous studies being a successful tool in the evolving pandemic, and further information about the virus having become available, a genomic analysis with a significantly larger sample set is justified. With this analysis they aim to provide a more accurate prediction of evolutionary capacity, and determine negatively selected residues to inform the design of therapeutic targets.

The full sequences of 3067 SARS-CoV-2 genomes were collected from two public databases, mapped against the original isolated SARS-CoV-2 genome (Wuhan-Hu-1/2019), and compared using computational genomics. To identify mutations, the authors generated a phylogenetic tree using maximum-likelihood (ML) and then produced a heatmap to correlate hotspot mutations with geographical locations. This showed that two-thirds of mutations were non-synonymous and distributed in six different SARS-CoV-2 genes, with most mutations found within ORF1ab, in the receptor-binding-domain (RBD) of spike protein. Ten hyper-variable genomic hotspots were identified, four of which were again found in ORF1ab. Geographically, most mutations and variability were found in the US and China. However, nearly 40% of samples were taken from both of these countries. The geographical mutational hotspot heatmap showed that mutations were geographically linked with two distinct clusters, a Europe-America-Africa cluster and an Asia-Australia cluster. Aotearoa was found to be the country with the highest mutational rate.

Next, the authors analysed the selective pressure of the collected SARS-CoV-2 genomes. To do so, the genomes were aligned codon-by-codon to generate a phylogenetic tree using ML and then four different algorithms were run to analyse site-specific selection. The results showed a relatively constant mutational rate with an increase in mutational rate in the most recently sequenced genomes. The authors identified the spike protein gene and ORF1ab as mostly being under purifying selection, whilst in both genes 15 specific sites were found to be under negative selection. In the spike protein most genes were under positive selection, however, the sites under negative selection were in the RBD region. Lastly, the authors constructed and analysed the pangenome of SARS-CoV-2 by comparing the collected SARS-CoV-2 genomes to 115 *Betacoronavirus* proteomes and using BLAST to identify orthologous genes. This revealed that the SARS-CoV-2 genome is comprised of 11 genes in the pan-genome, 9 of which are part of the SARS-CoV-2 core genome. One of the two accessory genes is a gene only found in SARS-CoV-2.

In the discussion, the authors argue that their in-depth phylogenetic and phylogeographic analysis of the SARS-CoV-2 genome is paramount to the investigation into the viruses' pathogenesis, prevention and treatment. Characterizing such a large quantity of genomes showed significant results that can be used in future studies. They discuss that the results showed mutations predominantly occurring in a region in

ORF1ab that codes for the endosome-associated protein, and thus provides an explanation for why SARS-CoV-2 is more contagious than its precursor SARS-CoV. Importantly, the results showed that SNP mutations are not random and highly represented in critical genes, such as the ORF1ab. The authors claim that identification of both, specific sites subject to negative selective pressure and positive selective pressure, will inform vaccine development. They then conclude that their study provides a valuable starting point for future studies.

I agree with the authors' argument that their study is important to inform future actions in regards to prevention and treatment of SARS-CoV-2. However, I think the results and discussion are lacking analysis into the underlying reasons for the genetic diversity observed in the geographical locations. I believe that deeper analysis of this topic would have helped distinguish between people traveling and carrying the virus with them, and the virus independently developing particular mutations. Ignoring the former may have resulted in misinterpretation of the data, therefore the claim that ORF1ab is a key mutational hotspot is perhaps incorrect. Even though there was a lack of insight in some parts of the discussion, I do believe that, at the time, this study was relevant and had a significant impact on future research. It was published just 2 months after the WHO announced the global pandemic and, since then has been cited 38 times[1], with much of the resulting research relating to vaccine development[6] and further genomic analyses[7,8]. Finally, the authors stated that they added all their analyses to a website (http://covid-19.medbiotech.ma) with the intention to make this data accessible to the public and scientists. However, the website does not seem to work anymore, which is a little disappointing considering how important data sharing and communication has become during the pandemic.

References

1. Laamarti, M., Alouane, T., Kartti, S., Chemao-Elfihri, M. W., Hakmi, M., Essabbar, A., Laamarti, M., Hlali, H., Bendani, H., Boumajdi, N., Benhrif, O., Allam, L., El Hafidi, N., El Jaoudi, R., Allali, I., Marchoudi, N., Fekkak, J., Benrahma, H., Nejjari, C., ... Ibrahimi, A. (2020). Large scale genomic analysis of 3067 SARS-COV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. *PLOS ONE*, *15*(11). https://doi.org/10.1371/journal.pone.0240345
2. world health organisation. (n.d.). *Who situation report 32 - who | world health organization*. Coronavirus disease (COVID-19) Situation Report – 102. Retrieved October 7, 2021, from https://www.who.int/docs/default-source/searo/indonesia/covid19/who-situation-report-32.pdf?sfvrsn=3d4913aa_2.
3. Li, L., Huang, T., Wang, Y., Wang, Z., Liang, Y., Huang, T., Zhang, H., Sun, W., & Wang, Y. (2020). Covid-19 patients' clinical characteristics, discharge rate, and Fatality Rate of meta-analysis. *Journal of Medical Virology*, *92*(6), 577–583. https://doi.org/10.1002/jmv.25757
4. Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., & Lu, J. (2020). On the origin and continuing evolution of SARS-COV-2. *National Science Review*, *7*(6), 1012–1023. https://doi.org/10.1093/nsr/nwaa036
5. Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3
6. Alouane, T., Laamarti, M., Essabbar, A., Hakmi, M., Bouricha, E. M., Chemao-Elfihri, M. W., Kartti, S., Boumajdi, N., Bendani, H., Laamarti, R., Ghrifi, F., Allam, L., Aanniz, T., Ouadghiri, M., El Hafidi, N., El Jaoudi, R., Benrahma, H., Attar, J. E., Mentag, R., ... Ibrahimi, A. (2020). Genomic diversity and hotspot mutations in 30,983 SARS-COV-2 genomes: Moving toward a universal vaccine for the "confined virus"? *Pathogens*, *9*(10), 829. https://doi.org/10.3390/pathogens9100829
7. Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., Rangarajan, E. S., Pan, A., Vanderheiden, A., Suthar, M. S., Li, W., Izard, T., Rader, C., Farzan, M., & Choe, H. (2020). SARS-COV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19808-4
8. Huang, S.-W., Miller, S. O., Yen, C.-H., & Wang, S.-F. (2020). Impact of genetic variability in ACE2 expression on the evolutionary dynamics of SARS-COV-2 spike D614G mutation. *Genes*, *12*(1), 16. https://doi.org/10.3390/genes12010016