Bioinformatics analysis of mutations in SARS-CoV-2 and clinical phenotypes

E. Ulzurrun[1,2], D. del Hoyo[2], M. Álvarez-Herrera[3], P. Ruiz-Rodríguez[3], B. Navarro-Domínguez[3], S. Rodríguez Santana[4], D. Garcia Rasines[4], R. Naveiro[4], C. Gil[1], J.M. Carazo[2], M. Coscollá[3], C.O.S. Sorzano[2], N. Campillo[1,4]

1. Centro de Investigaciones Biológicas Margarita Salas, CSIC
2. Centro Nacional de Biotecnología, CSIC
3. Instituto de Biología Integrativa de Sistemas, CSIC-Universidad de Valencia
4. Instituto de Ciencias Matemáticas, CSIC

Background: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), initially reported in Wuhan (China) has spread worldwide. Like other viruses, SARS-CoV-2 accumulates mutations with each cycle of replication by continuously evolving a viral strain with one or more single nucleotide variants (SNVs). However, SNVs that cause severe COVID-19 or lead to immune escape or vaccine failure are not well understood. We aim to identify SNVs associated with severe clinical phenotypes.

Methods: In this study, 27429 whole-genome aligned consensus sequences of SARS-CoV-2 were collected from genomic epidemiology of SARS-CoV-2 project in Spain (SeqCOVID)[1]. These samples were obtained from patients who required hospitalization and/or intensive care unit admission (ICU) , excluding those registered in the first pandemic wave. Besides, 248 SARS-CoV-2 genomes were isolated from COVID-19 hospitalized patients from Gregorio Marañon General University Hospital (GMH) of which 142 were fully vaccinated. Bioinformatics tools using R and Python programming languages were developed and implemented comparing those to SARS-CoV-2 Wuhan-Hu-1 (reference genome).

Results: Using a selection threshold mutational frequency 10%, 27 SNVs were expected to have association with hospitalization and ICU risk. The reference haplotype differing at the SNV coding for lysine at the residue 203 (N:R203K) was found to have negative association with COVID-19 hospitalization risk ($p = 5.37 \times 10^{-04}$). Similarly a negative association was observed when the residue at 501 is replaced by tyrosine (S:N501Y) ($p = 1.33 \times 10^{-02}$) . The application of a Chi-square test suggested that SNV-haplotypes coding for mutants residues such as (S:A222V, N:A220V, ORF10:V30L) and (ORF1a:T1001I, ORF1a:I2230T, S:N501Y, S:T716S, S:S982A, ORF8:Q27*, N:R203K, N:S235F) have negative associations with COVID-19 hospitalization risk ($p = 6.58 \times 10^{-07}$ and $p = 2.27 \times 10^{-16}$, respectively) and COVID-19 ICU risk ($p = 1.15 \times 10^{-02}$ and $p = 2.51 \times 10^{-02}$, respectively). Focusing on the SNV-haplotype coding the mutations (S:A222V, N:A220V, N:D377Y, ORF10:V30L) were observed to increase the risk of COVID-19 hospitalization ($p = 2.71 \times 10^{-04}$). Results from SARS-CoV-2 genomes analysis from GMH showed 63 coding SNVs which met the established threshold value. Applying a Chi-square test, the SNV-haplotype carrying coding variants for mutant residues in 5 ORF proteins and surface and membrane glycoprotein and nucleocapsid phosphoprotein was significantly associated with vaccine failure in hospitalized COVID-19 patients ($p = 7.91 \times 10^{-04}$).

Conclusions: SNV-haplotypes carrying variants lead to non-synonymous mutations located along SARS-CoV-2 whole-proteome may influence COVID-19 severity and vaccine failure suggesting a functional role in the clinical outcome for COVID-19 patients.

[1] https://seqcovid.csic.es/