SARS-CoV2 genomic statistical analysis to study hospitalization and vaccine failure

S. Rodríguez Santana[1], R. Naveiro[1], D. García Rasines[1], C. B. Guevara Maldonado[1], E. Ulzurrun[23], M. Álvarez-Herrera[4], P. Ruiz-Rodríguez[4], B. Navarro-Domínguez[4], C.O.S. Sorzano[3], M. Coscollá[4], N. Campillo Martín[12], D. Ríos Insua[1]

1 - Instituto de Ciencias Matemáticas, ICMAT-CSIC
2 - Centro de Investigaciones Biológicas Margarita Salas, CIB-CSIC
3 - Centro Nacional de Biotecnología, CNB-CSIC
4 - Instituto de Biología Integrativa de Sistemas I2SysBio, CSIC y Universidad de Valencia

Background: The study of properties of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) through its genome has become a very relevant topic to provide a better understanding of the evolution of the pandemic. The availability of sequenced genomes of SARS-Cov-2, including its different mutations, has led to an attempt to characterize the mutations may be associated with infection severity or vaccine failure. Most of the existing research is centred on individual mutations, with a minor focus on models that account for interactions between mutations.

Methods: We developed a framework for processing and studying genomic data to characterize the effect of individual and pairs of mutations in the spike protein on SARS-CoV-2 infection severity and vaccine failure. To this end, we employed different datasets with the relevant information available: we used over 25.000 sequences obtained from patients with information about the acuteness of the infection (hospitalization, ICU admission, and death of the patient). On the other hand, we studied the prevalence of certain mutations or pairs of mutations depending on the vaccination status of almost 5.000 patients using their respective SARS-CoV-2 sequences. These analyses were conducted fitting a hierarchical group Lasso model that allows for the presence of individual effects and pairwise interaction terms. In addition, individual studies of the relevant mutations and interactions were conducted using propensity score matching to control for possible confounding factors.

Results: In both cases, we identified some of the most prevalent individual positions already present in the literature, which serves as a good benchmark. Moreover, we found other relevant individual mutations, as well as some important interaction effects between them, which sometimes play a more important role in the model than previously studied individual mutations. We constructed a graph representing the pairwise interactions selected by the model that provides insight into the disjoint community structure for important mutations. Finally, through a detailed individual study, we extracted further information about the relevance of the selected mutations and interactions, which are of major importance in vaccine failure studies.

Conclusions: Our model allows identification and characterization of novel individual mutations and interaction terms that are associated with infection severity and vaccine failure. The results show different structures of important mutations and interactions in the SARS-CoV-2 spike protein. These have not been reported in the literature yet and, in some cases, have stronger effects on the final outcome for the patient than the originally studied individual mutations.