The most exposed regions of SARS-CoV-2 structural proteins are subject to under positive selection and gene overlap may locally modify this behavior

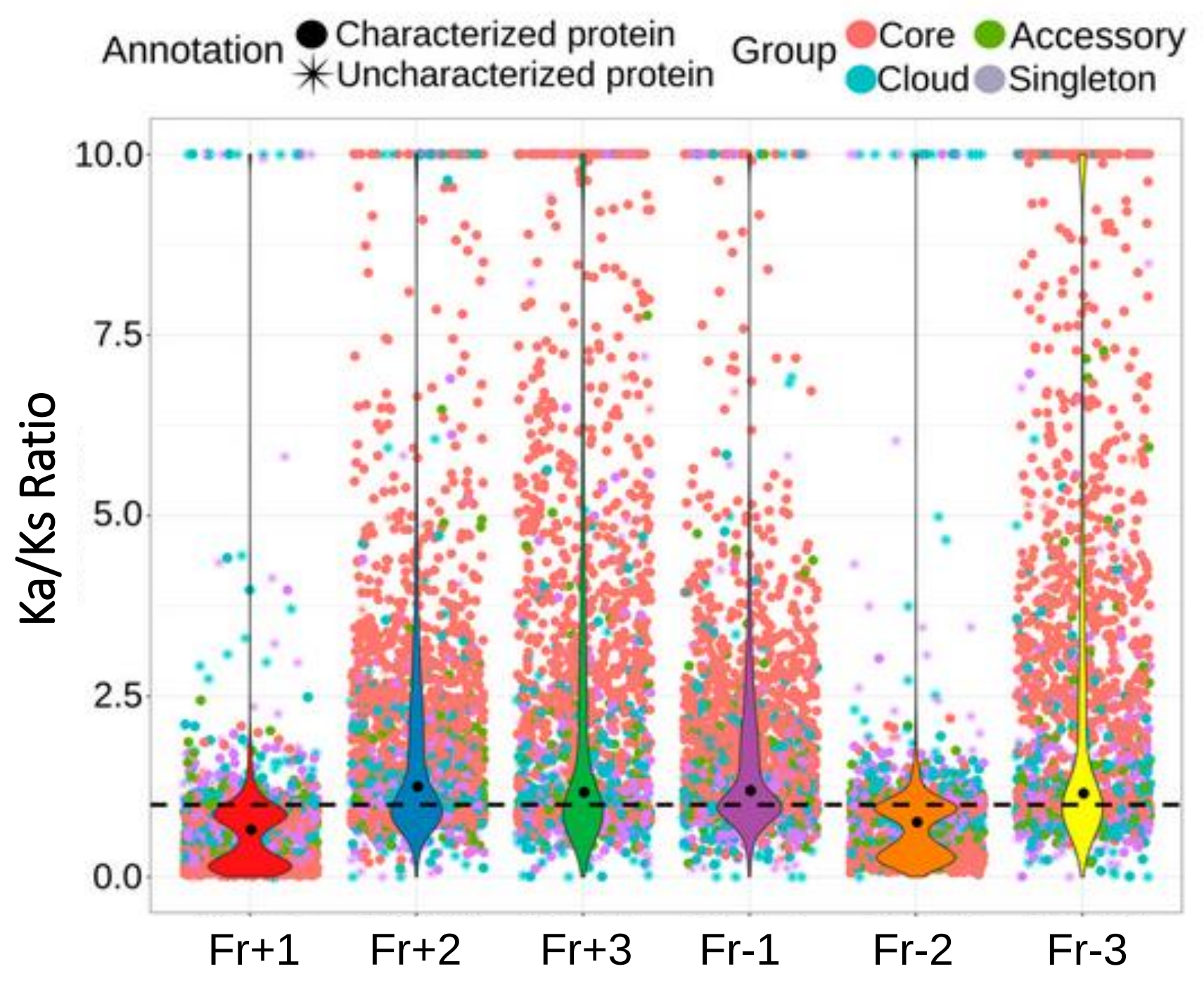
Alejandro Rubio[1], **María de Toro[2]**, **Antonio J. Pérez-Pulido[1]**

[1] Andalusian Centre for Developmental Biology (CABD, UPO-CSIC-JA). Faculty of Experimental Sciences (Genetics Dept.), University Pablo de Olavide, 41013, Seville, Spain
[2] Genomics and Bioinformatics Core Facility. Center for Biomedical Research of La Rioja, 26005, Logroño, Spain

JABI
JORNADAS ANDALUZAS DE BIOINFORMÁTICA
SEVILLA, 3-5 DE JULIO 2023

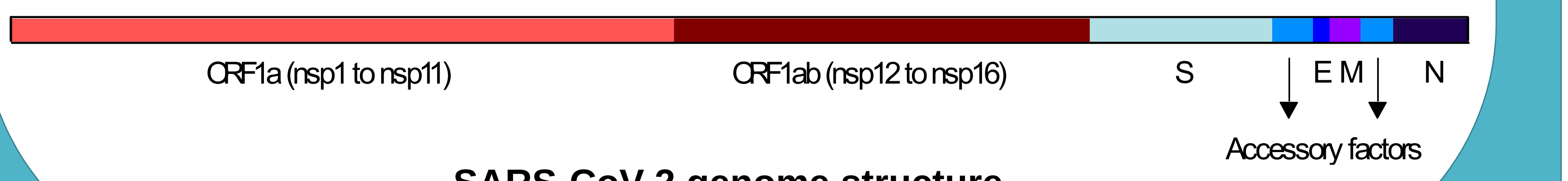
The SARS-CoV-2 virus pandemic that emerged in 2019 has been an unprecedented event in international science, as it has been possible to sequence millions of genomes, tracking their evolution very closely. This has enabled various types of secondary analyses of these genomes, including the measurement of their sequence selection pressure. In this work we have been able to measure the selective pressure of all the described SARS-CoV-2 genes, even analyzed by sequence regions, and we show how this type of analysis allows us to separate the genes between those subject to positive selection (usually those that code for surface proteins or those exposed to the host immune system) and those subject to negative selection because they require greater conservation of their structure and function. We have also seen that when another gene with an overlapping reading frame appears within a gene sequence, the overlapping sequence between the two genes evolves under a stronger purifying selection than the average of the non-overlapping regions of the main gene. We propose this type of analysis as a useful tool for locating and analyzing all the genes of a viral genome, when an adequate number of sequences are available.

The **Ka/Ks ratio** is used to measure the pressure selection. This is the number of **non-synonymous substitutions (Ka) per synonymous substitution site (Ks)**. We have previously shown that this ratio can measure evolutionary pressure in bacteria such as *Helicobacter pylori* or *Acinetobacter baumannii*. As expected, the Ka/Ks ratio is lowest at frame +1, where protein sequences are conserved.



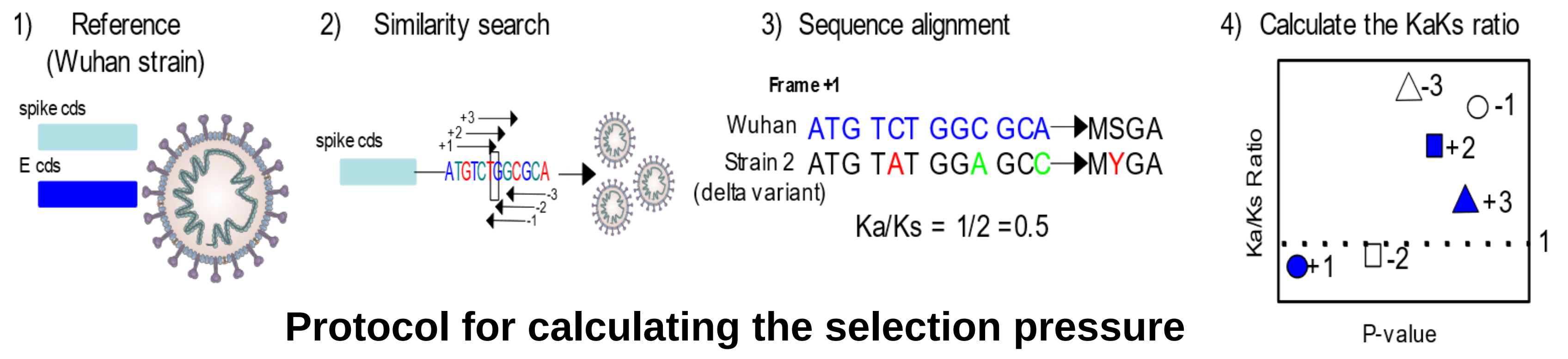
Distribution of the Ka/Ks ratio in all the protein-coding genes from the *Helicobacter pylori* pangenome

In this work, we used the genome of the **SARS-CoV-2 virus**, which caused the pandemic that began in 2019. Its genome codes for non-structural proteins (nsp), structural proteins (S, M, E and N) and accessory factors that help correct assembly.



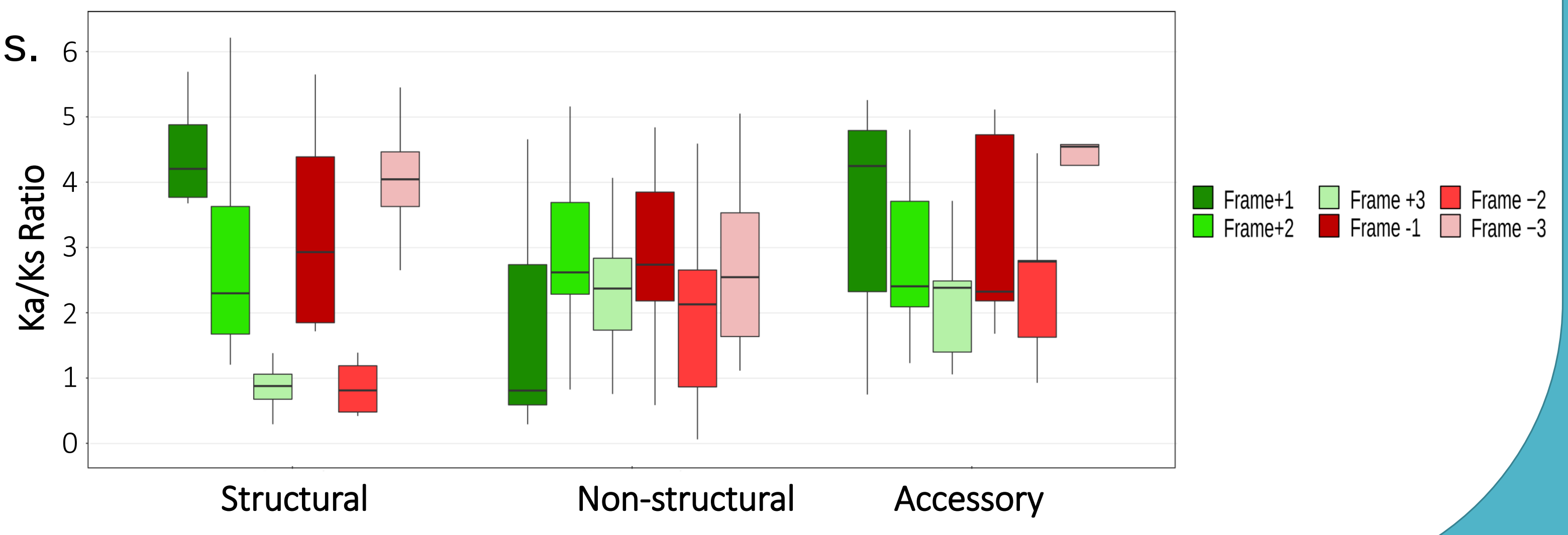
SARS-CoV-2 genome structure

The protocol has been adapted for use with approximately 2000 strains sequenced in a standardized manner at the Hospital Universitario San Pedro.



Protocol for calculating the selection pressure

The **structural and accessory genes** of the virus, which are the proteins most exposed to the host immune system and interact with proteins of the infected cell, **show positive selection**. However, **non-structural proteins**, which are involved in the processes that take place once the viral genome has entered the cell, **show negative selection**, suggesting that these proteins have essential functions.



Ka/Ks ratio distribution separated by gene type

The next step was to determine whether these variations in the Ka/Ks ratio were global or localized to specific points. A sliding window protocol was designed for this purpose.

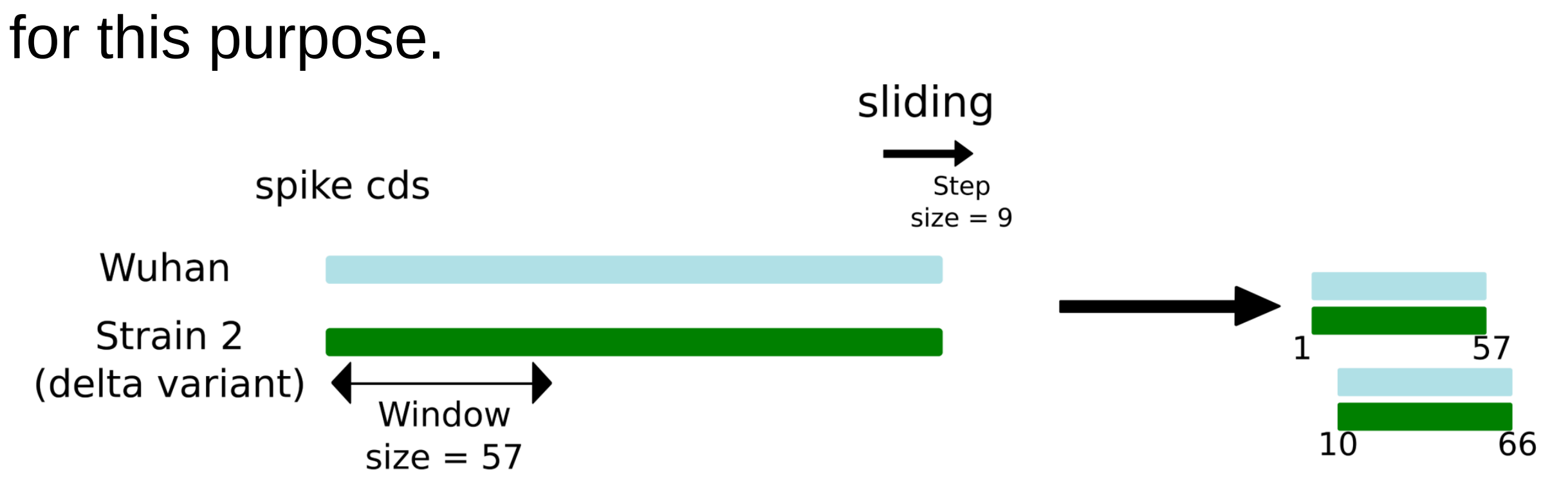
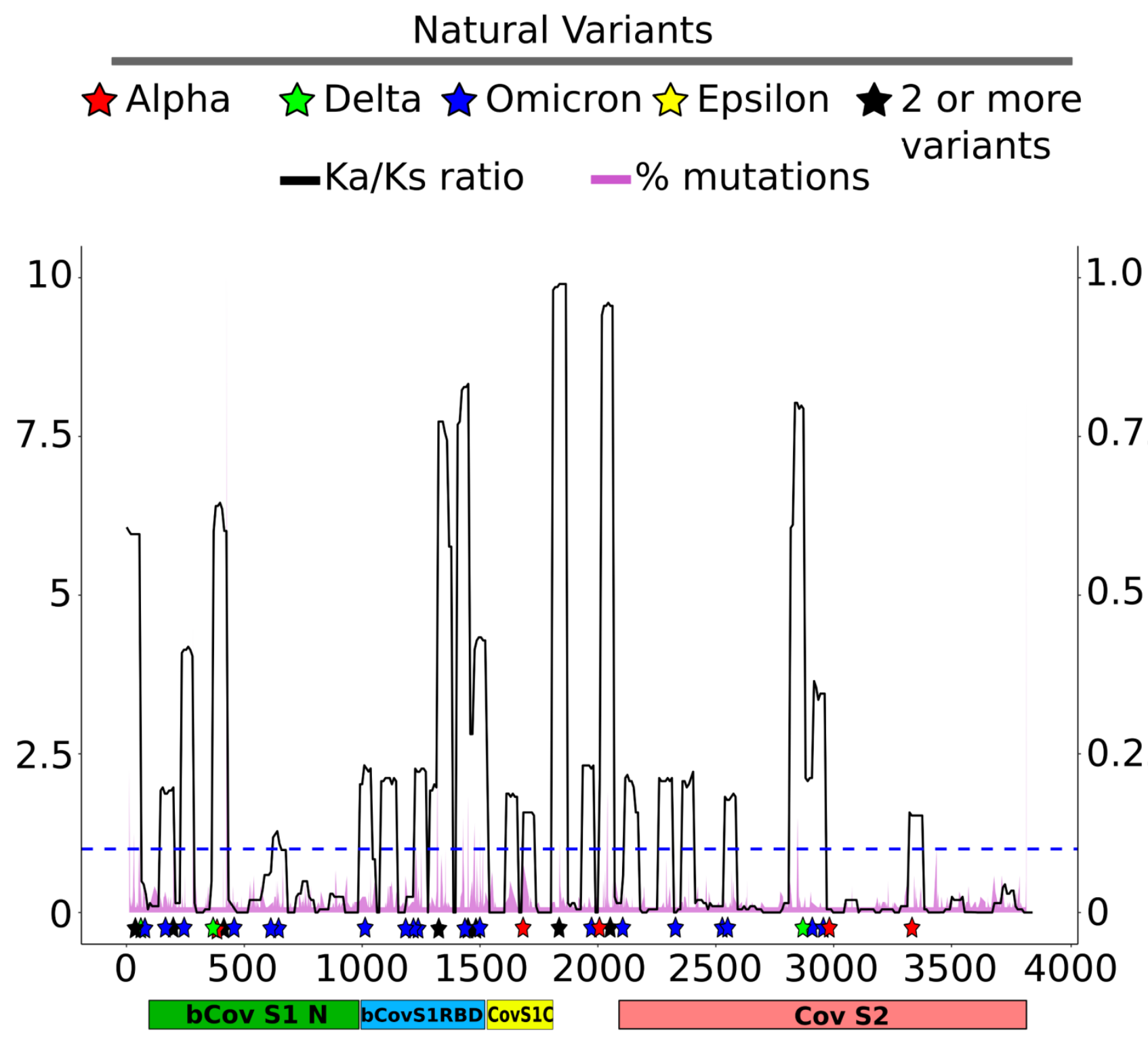
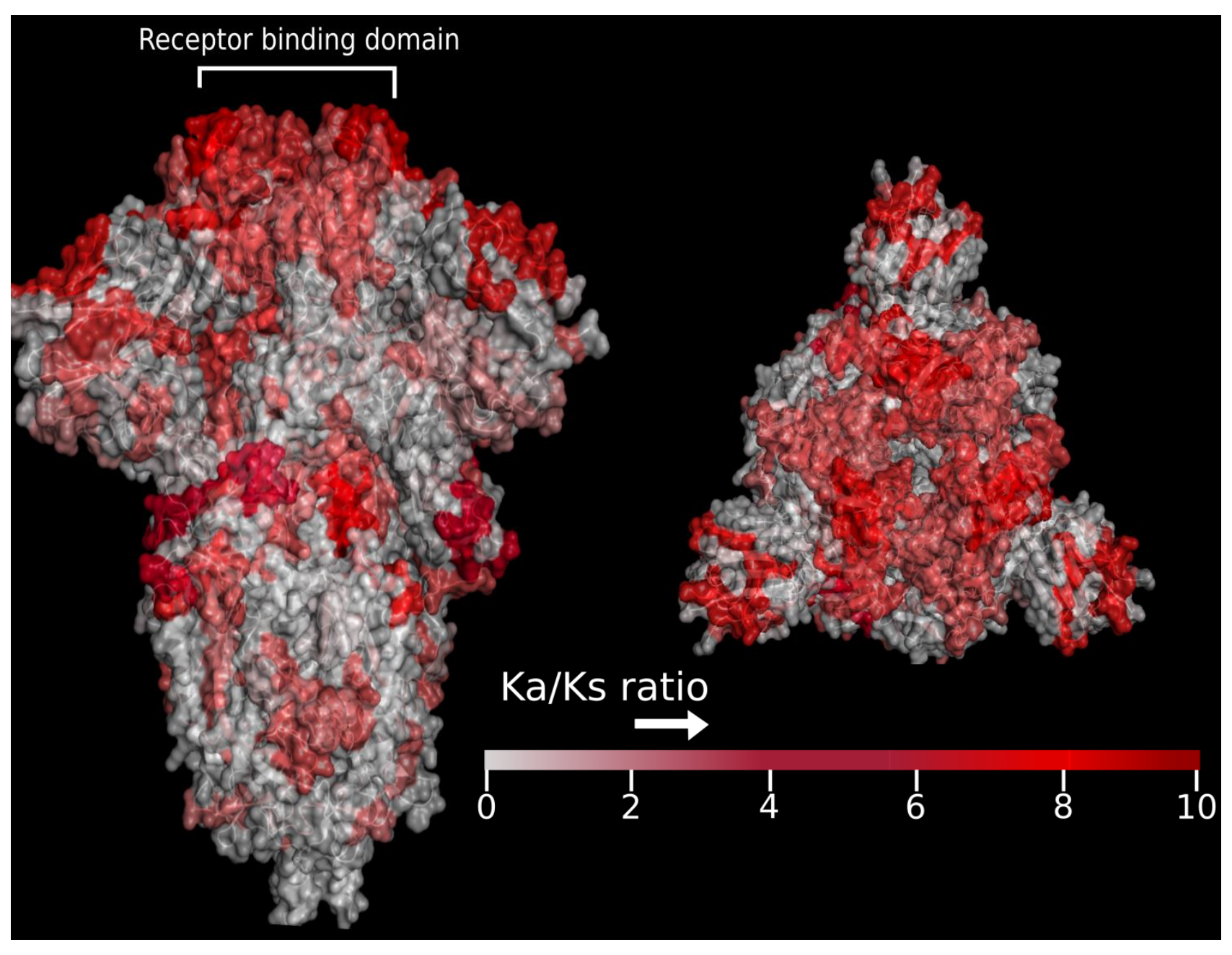


Diagram of the Sliding Window Method

Most structural genes show global changes in the Ka/Ks ratio. However, the spike (S) gene **has higher ratios** in the regions that interact with host cells at protein level (**ACE2 receptor, receptor binding domain**).

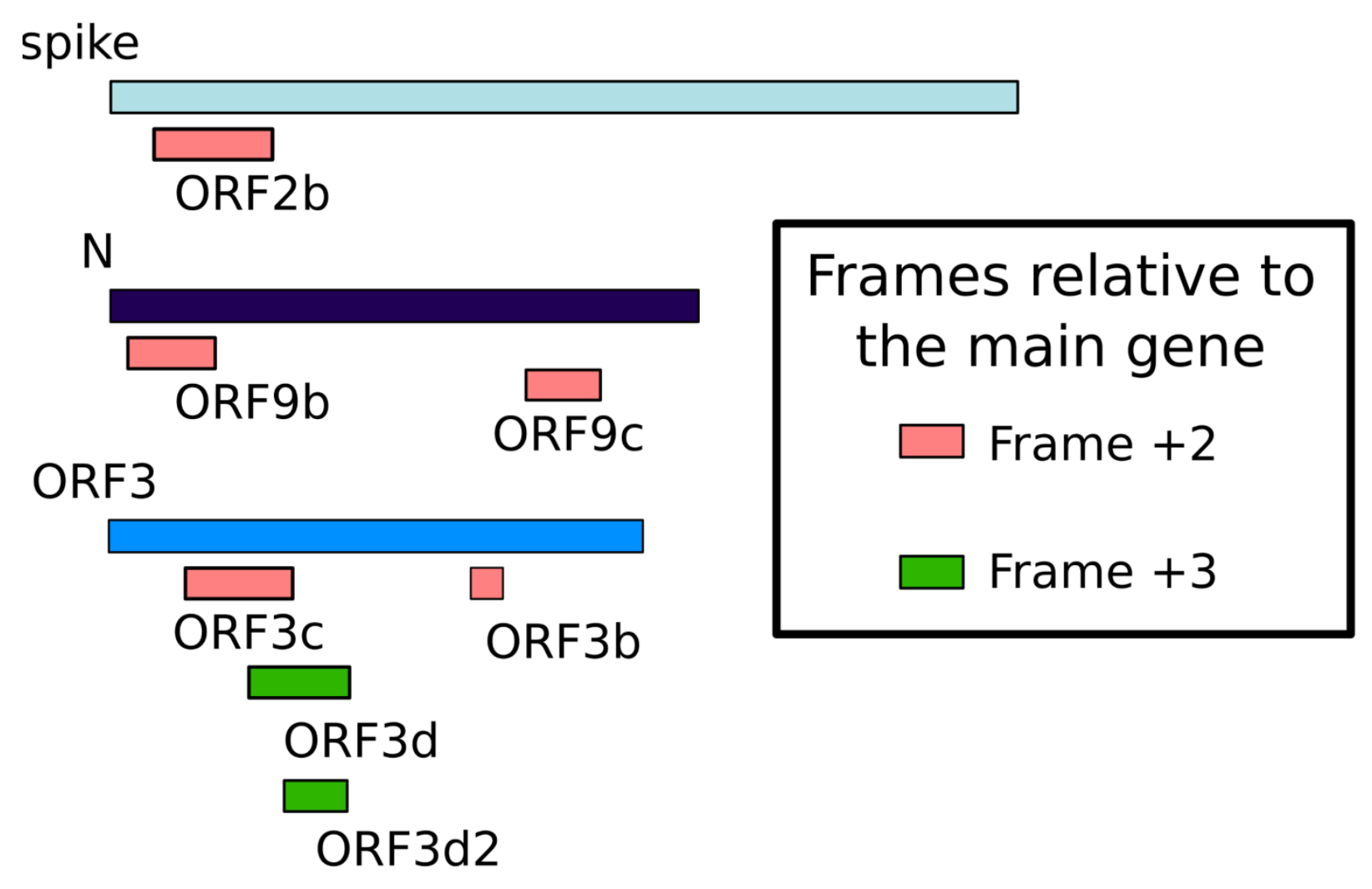


Ka/Ks ratio distribution along the sequence of S gene

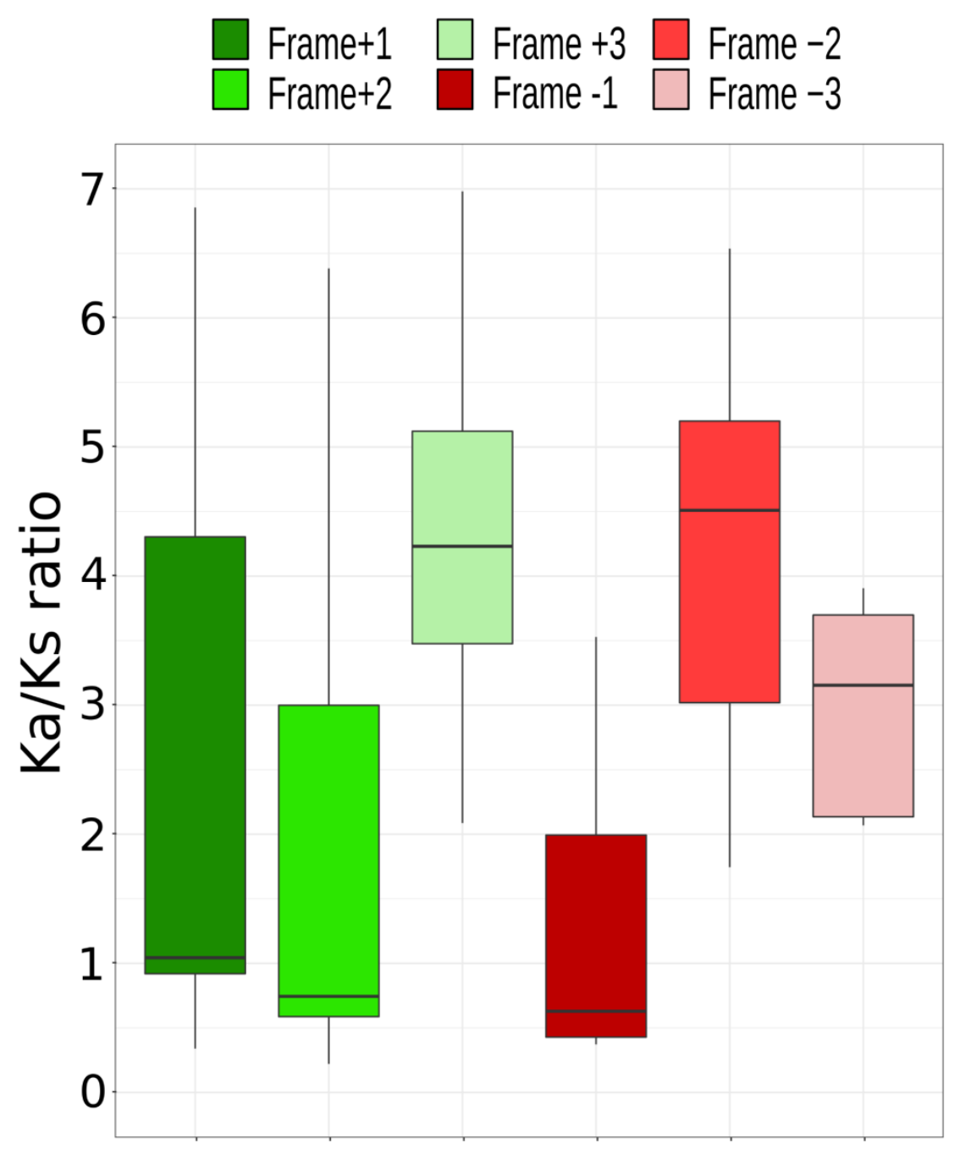


SARS-CoV-2 Spike protein structure highlighting regions with a higher Ka/Ks ratio

The Ka/Ks ratio was analyzed in **ORFs overlapping** important structural genes. In general, a **low ratio was obtained**, which could partly explain the low Ka/Ks ratio values of structural genes in alternative reading frames (frame +2 and frame +3) with respect to the value given by their reading frame +1.



Overlapping regions of *S*, *N* and *ORF3* genes



Ka/Ks ratio distribution in overlapping genes

The results presented here show how to **analyze the selection pressure** on the genes of a viral genome, which is not only useful for locating highly conserved regions and drug targets, but also allows the analysis of overlapping genes

SCAN ME