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. Center for Biological Research Margarita Salas, CSIC. 2. National Biotechnology Center, CSIC. 3. Institute for Integrative Systems Biology, CSIC- University of Valencia. 4. Institute of Mathematical Sciences, 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, SNV coding for mutant residues that cause severe COVID-19 or lead to immune escape or vaccine failure are not well understood. We **aim** to identify **haplotypes throughout the virus proteome** associated with clinical phenotypes particularly with **vaccine failure**.

Methods: 386 and 4983 **aligned SARS-Cov-2 consensus genome sequences** and associated metadata from **Gregorio Marañon Hospital** (GMH386) and **GISAID** were used, respectively.
Pre-processing and analysis of **haplotypes** were carried out using scripts in **R** and **Python** programming languages, as well as python modules such as scipy.stats and others included in Biopython. Haplotype association was preformed by Chi-square test.
NCBI Reference Sequence NC_045512.2 (Wuhan) was used to retrieve the coding sequences.
Mutation is defined as mutational frequency for the minor nucleotide ≥ 10% and ≤ 90% for the major nucleotide at a particular gene locus. On the other hand, reference genome is made of the major nucleotide for each locus of the aligned sequences.

Results: A short description of whole population and haplotypes are shown in Tables 1 and 5. Using the nucleotide mutations, 85 and 45 non-synonymous amino acid mutations throughout the proteome of SARS-CoV-2 were identified for GMH386 and GISAID datasets, respectively**. Associations with vaccine failure** regardless of dose received were found in both set of sequences for two haplotypes. However, COVID-19 patients expected for haplotype 2 for GISAID was lower than expected significantly ($p = 5.88 \times 10^{-161)}$) (Tables 2 and 6). Haplotypes of mutations are outlined in the Tables 3-4 and Tables 7-8.
Figure 1 shows colored in red the shared mutations for both datasets located at one homotrimer of the Spike protein. The receptor-binding domain (RBD) is colored in yellow. In contrast, the N-terminal domain (NTD) is highlighted in cyan. These loci are noteworthy with asterisk in Tables 4 and 8.

Table 1. Brief description of the GM386 patients with a collection date from 2021-03-02 to 202204-24

	Female, n (%)	Male, n (%)	>= 65 y, n (%)	< 65 y, n (%)
GMH386 database, n = 386	199 (52)	187 (48)	250 (65)	136 (35)
h_GM386_1, n = 46	22 (48)	24 (52)	41 (89)	5 (11)
h_GM386_2, n = 39	24 (62)	15 (38)	35 (90)	4 (10)

Abbreviations: y = years; h_GM386 = haplotype for Gregorio Marañon Hospital 386 dataset

Table 2. Vaccination distribution for the haplotypes with a frequency ≥ 10 %

	Vaccinated Yes	Expected frequencies Yes	Vaccinated No	Expected frequencies No	*p* value*
h_GM386_1, n = 46	37	26	9	20	1.41E-03
h_GM386_2, n = 39	31	22	8	17	5.83E-03

Abbreviations: h_GM386 = haplotype for Gregorio Marañon Hospital 386 dataset
*Chi-square test

Table 3. Haplotypes with 85 non-synonymous mutations for GM386

Reference	STKAGAPVPPADVLVTLPTINGLPRIYATETATVGRSSSTDRKNGSEQGQNYTHNNDDQNLDSTTTDQAIFRDVTLTQPDMGGDS
h_GM386_1	...S....LS..L.....A..SIL...V..R.....................N....L.....T...AI.IL.G..CY.
h_GM386_2	RI..S.........F.IFH.....CV..I..I..GD.FPFANSNK.NAR.RYH.YKKY.HK....II.ET...L.....L.IR..R

Table 4. Protein mutations of SARS-Cov-2 for GM386

ORF1a:135	ORF1a:2710	ORF1b:591	S:19	S:405	S:501*	S:981*	M:112	N:203
ORF1a:842	ORF1a:2907	ORF1b:662	S:67*	S:408	S:505*	S:1146	ORF6:20	N:204
ORF1a:856	ORF1a:2930	ORF1b:829	S:95*	S:417	S:547	ORF3a:26	ORF6:61	N:215
ORF1a:1306	ORF1a:3027	ORF1b:1000	S:213	S:440	S:655*	ORF3a:64	ORF7a:82	N:377
ORF1a:1307	ORF1a:3053	ORF1b:1315	S:339*	S:446	S:679	ORF3a:223	ORF7a:120	N:413
ORF1a:1352	ORF1a:3090	ORF1b:1566	S:346	S:477	S:764	E:9	ORF7b:18	
ORF1a:1803	ORF1a:3201	ORF1b:1759	S:371*	S:484*	S:796*	M:3	ORF7b:40	
ORF1a:1887	ORF1a:3395	ORF1b:1918	S:373*	S:493*	S:950	M:19	N:9	
ORF1a:2046	ORF1a:3646	ORF1b:2163	S:375*	S:496*	S:954*	M:63	N:13	
ORF1a:2287	ORF1a:3758	ORF1b:2196	S:376	S:498*	S:969*	M:82	N:63	

Table 5. Brief description of the GISAID patients with a collection date from 2020-11-03 and 2022-02-03

	Female, n (%)	Male, n (%)	>= 65 y, n (%)	< 65 y, n (%)
GISAID database, n = 4983	2772 (56)	2200 (44)	1368 (28)	3615 (72)
h_GISAID_1, n = 902	496 (55)	405 (45)	268 (30)	634 (70)
h_ GISAID_2, n = 927	577 (62)	350 (38)	327 (35)	600 (65)
h_ GISAID_3, n = 607	347 (57)	260 (43)	132 (22)	475 (78)

Abbreviations: y = years; h_GISAID = haplotype for GISAID dataset dataset

Table 6. Vaccination distribution for the haplotypes with a frequency ≥ 10 %

	Vaccinated Yes	Expected frequencies Yes	Vaccinated No	Expected frequencies No	*p* value*
h_GISAID_1, n = 902	477	391	425	511	1.82E-10
h_GISAID_2, n = 927	33	402	894	526	5.88E-161
h_GISAID_3, n = 607	328	263	279	344	1.61E-08

Abbreviations: h_GISAID = haplotype for GISAID dataset dataset
*Chi-square test

Table 7. Haplotypes with 45 non-synonymous mutations for GISAID

Reference	KAAPPADVIGPAATGSSSEQGQNYHDQNLSTTDQRVTLTPDMGGD
h_GISAID_1	.S.LS..L.SLV.....L.....AI.I.G..CY
h_GISAID_2	R....T..V...VIDLPFARSRYHYYHKF..IGE.....L.IR..
h_GISAID_3	.S.LS..L.SLV.I.....L.....AI.I.G..CY

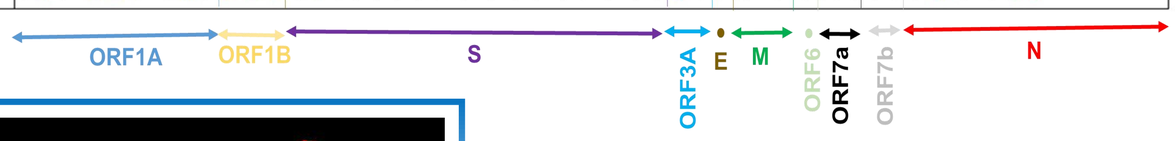

Table 8. Protein mutations of SARS-Cov-2 for GISAID

ORF1a:856	ORF1b:1000	S:493*	ORF3a:26	N:13
ORF1a:1306	ORF1b:1918	S:496*	ORF3a:64	N:63
ORF1a:1707	S:67*	S:498*	E:9	N:203
ORF1a:2046	S:95*	S:501*	M:3	N:204
ORF1a:2287	S:339*	S:505*	M:19	N:215
ORF1a:2710	S:371*	S:655*	ORF6:20	N:377
ORF1a:2907	S:373*	S:796*	ORF7a:82	
ORF1a:2930	S:375*	S:954*	ORF7a:120	
ORF1a:3758	S:484*	S:969*	ORF7b:18	
ORF1b:662		S:981*	ORF7b:40	

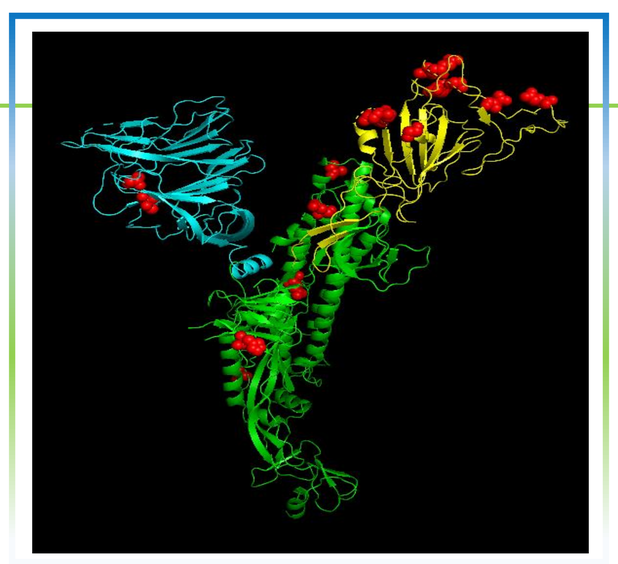

***Figure 1*. Common mutations located at the Spike protein**

***Conclusions:* Haplotype analysis suggest the relevance of the genetic variability throughout the SARS-CoV-2 whole-proteome in the response of COVID-19 patients to the vaccine.**

This research work was funded by the European Commission-NextGenerationEU (Regulation EU 2020/2094), through CSIC's Global Health Platform (PTI Salud Global)