Companion vaccine Bioinformatic design tool reveals limited functional genomic variability of SARS-Cov-2 Spike Receptor Binding Domain

Domain

Alice Massacci[1], Lorenzo D'Ambrosio[2], Eleonora Sperandio[1], Fabio Palombo[3], Luigi Aurisicchio[3], Gennaro Ciliberto^[4] and Matteo Pallocca^[1]*

[1]Biostatistics, Bioinformatics and Clinical Trial Center, IRCCS Regina Elena National Cancer Institute, Rome, Italy

[2]Tumor Immunology and Immunotherapy Unit, IRCCS Regina Elena National Cancer Institute, Rome, Italy.

[3]Takis srl, Rome, Italy

[4]Scientific Direction, IRCCS Regina Elena National Cancer Institute, Rome, Italy.

***Corresponding author:** Dr. Matteo Pallocca, Biostatistics, Bioinformatics and Clinical Trial Center, IRCCS Regina Elena National Cancer Institute, Via Elio Chianesi 53, 00144 Rome, Italy. Tel. +39-06-52666992, E-mail: matteo.pallocca@ifo.gov.it

^These authors contributed equally to this work

Abstract

Tracking the genetic variability of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is crucial for several reasons, such as to identify target sequences to generate robust vaccines and neutralizing monoclonal antibodies; to track viral genetic temporal and geographic evolution; and to mine for viral variants associated with reduced or increased severity. Several online tools and bioinformatic analyses have been released for this purpose; four main viral clades have been

described via phylogenetic analyses. Here, we present an open-source bioinformatic protocol focused on SARS-CoV-2 single mutations and minimal consensus sequence building as a companion vaccine design tool. Results on the whole GISAID sequence dataset at the time of the writing (April–June 2020) confirms no high variability pressure on the viral genome in terms of quality and quantity of mutations. We then focused our analysis on the receptor-binding domain region of the spike protein and only found a few variants associated with particular geographic locations that should be tracked over time. Finally, additional immunogenomic analyses revealed some variation in mutated epitope MHC compatibility and T-cell recognition for most frequent human HLAs.

Keywords: SARS-CoV-2 genome; SARS-CoV-2 mutation; COVID mutations; SARS-CoV-2 vaccine; bioinformatic workflow; docker

Introduction

With more than 6.5 million infected people, SARS-CoV-2 is likely to have found a reservoir in the human population, increasing the possibility of an outbreak in the next autumn. To prevent the viral spread, an effective vaccine is needed.

More than 100 different vaccines are in development all over the word, including China, India, the USA and Europe. Most, if not all of them, target the spike protein using different protein formulations or platforms, such as a vectored vaccine or nucleic acids, RNA and DNA. The spike protein is the best candidate for historical reasons [1], and for recent preclinical evidence in non-human primates [2].

One of the main questions to ask is the following: from preclinical vaccine testing to the first-in-man trial to confirmatory trials, regulatory approval and large-scale vaccine distribution, from 6 to 12 months will elapse. During this time frame, will the spike protein of the circulating virus mutate? If so, are these genetic variants going to escape protective immune responses induced by the vaccine? In this paper, we try to provide a bioinformatic toolbox to address this potential issue.

A few preliminary reports already described the variant landscape of COVID-19 in different clades [3-4]. Both pointed out the limited presence of functional mutations in key regions of the genome. Nevertheless, some other preprints stressed a possible selection pressure on the D614G variant, the only frequent variation of the spike protein [5]. A recent bioinformatic analysis of SARS-CoV-2 epitopes showed a high homology with SARS-CoV [6] and, thus, characterized many possible B- and T-cell epitopes, and the same group provided an *in vivo* characterization of the virus proteins that are most targeted by T cells, confirming the spike protein to be the first region of interest [7].

To provide an evidence-based approach, we looked at the public available sequences and provide some comments with respect to vaccines and immune therapy. We also present a freely available bioinformatic protocol in order to enable all the bioinformatic community to dissect hCov-19 genomic variability and its functional impact on DNA vaccines. This work presents a protocol able to automatically mine a list of profiles useful for vaccine design from a pool of viral sequences and a target protein or domain of interest: a consensus sequence, in terms of average or minimal identity and a list of mutations with their absolute frequencies. Furthermore, the tool provides insightful graphical representation of the mutation frequency focusing on the genetic position of certain crucial domain, such as the receptor-binding domain (RBD) of the spike protein.

Results

Bioinformatic framework

The overall workflow is depicted in Figure 1A. The only input of the analysis pipeline is a multi-sequence FASTA file representing the genome repertoire of many viral assemblies, which is available on the EpiCov™ section of the GISAID portal [8]. Filters used to fetch the 24,379 sequences analyzed in this manuscript were Host=Human, Virus=hCov-19, flagged “complete”, “high coverage” and “low coverage exclusion”.

As a first step, sequences are aligned against the Wuhan strain (NC_045512.2) with Bowtie2 [9] in order to produce a SAM/BAM alignment file. The said file is distributed across N files, containing one sequence each. This step is required for the bcftools [10] *pileup* command that computes all possible alleles found in every position, which is always with respect to the Wuhan strain reference. Variants are then called and indel-normalized via the bcftools *call* command. At this point, we have a list of genetic variants in a VCF file describing how every sample/viral individual differs from the original strain, and the next step would be describing and annotating the said variants in terms of functional effect and penetration in the viral population, while producing a consensus sequence that represents the most frequent allele for each position. The *consensus* command from bcftools builds the said consensus from the filtered VCF file. In addition to the consensus for the viral complete genome, we build the consensus sequence separately for the spike protein and RBD. Moreover, a consensus of minimal identity is constructed, applying the N character at variant sites instead of the most frequent base. Such consensus of minimal identity, by showing which residues are conserved and which residues are variable, would help in the design of a vaccine.

Variant annotation is employed via the *snpEff* package [11] that embeds the NC_045512.2 genome assembly annotation in its standard package. A human-readable (and easily parsable) table

is formatted thanks to *SnpSift* jar package [12] of *snpEff*. 2D visualizations of the variants on the viral genome are generated via the LolliPlot function of the trackVignette's R package.

The whole workflow and test sequence files have been included in a Docker image, enabling users to run the whole pipeline with just one command (e.g. *docker run covid-miner sequences.fasta*). The whole workflow, with sample test data and Dockerfile is available on https://gitlab.com/bioinfo-ire-release/covid-miner

All software first picked for this workflow are standard tools employed in an NGS human sequencing facility lab. For instance, bowtie2 is a software design to handle short sequences (up to ~1000 bp according to the authors [13]). Execution of the full pipeline on thousands of sequences resulted in 1-day computing time with 4-core and 16 GB RAM allocated. Testing and re-implementation of the workflow through optimized software, such as NUCMER [14], which, from preliminary tests, cuts the computing time down to minutes, is ongoing at the time of the writing (Fig. 1A).

Analysis results on 24,379 SARS-CoV-2 sequences

The annotated variants table of all the GISAID sequences (available on June 3, 2020), is available in Table 1. Looking at the total number of sequences the most frequent Spike protein variation is confirmed to be D614G (Figure 1B and C). After that, there is a set of variants with a frequency of 0.3–0.7% that occurred 100–200-times in the viral population. None of these variants lie in the Spike protein RBD locus. When focusing on the sole RBD, from amino acids 319–541, 13 variants arise, all with a relative frequency less than 0.1% and 10–20 absolute occurrences.

These RBD variants are represented in further detail in Figure 1B and C. We reasoned whether these variants belong to a certain geographical region and normalized the mutation count by "mutations per thousand isolates", that is

$$MpTI = \frac{M * 1e3}{I_c}$$

with M as the absolute mutation count and I_c the total isolates for that country.

This normalization smooths out the inter-country variability, but the resolution bias remains due to the high variability in isolate sequencing among countries, changing in order of magnitude from thousands to dozens. This bias will cause rare mutations to be hidden in countries with a few associated isolates, while the associations with highly frequent outliers will remain more robust. By inspecting RBD variants in 12 countries, a few events, such as A522V, are uniquely associated with their locations (i.e., Latvia), while clusters of co-occurring variants lie mostly in countries with the highest number of available sequences (i.e., USA, UK) (Table 2, Figure 2A-C).

Furthermore, we explored whether these RBD mutations were influencing the class 1 presentation of SARS-CoV-2 epitopes. From the top five mutations at 0.1% frequency, we performed MHC class 1 binding prediction with NetMHC4 [15], for both the wild-type (Wuhan strain) and the mutated strain, choosing the most frequent HLAs in public databases [16]. The software generated predictions for all possible 9-mers resulting from the full RBD sequence, and we computed how *binders* where calculated for all the considered HLAs. Out of 34 predictions, only three epitopes had binding affinity in mutated or wild-type, and none of them had an increased binding affinity caused by the mutation. Instead, one of them had a decreased binding affinity, namely GAEGFNCYF epitope caused by V483A mutation (Table 3).

WT	Mutant	Wild-type binders	Mutated binders
IYQAGSTPC	IYQAGNTPC	1	1
IYQAGSTPC	IYQAGSIPC	1	1
GVEGFNCYF	GAEGFNCYF	1	0

Finally, we asked whether these mutations had a significant impact on T- and B-cell recognition, and we ranked them via the Immune Epitope Database and analysis resource (IEDB) [17]. When considering T-cells with the class 1 immunogenicity tool [18], 117/250 (47%) epitopes showed a negative, non-immunogenic score in both wild-type and mutated forms. When focusing on the predicted class 1 binders for the same epitope/HLA combination in mutated peptides, both IYQAGSIPC and IYQAGNTPC epitopes do not show a strong induction in IEDB score, also starting with a low immunogenicity scores (Table 4, Figure 2C). These results point out to a limited variability in T-cell response mediated by these mutations.

Taken together, all these results confirm the both a limited variation pressure on critical SARS-CoV-2 loci, such as the RBD and on the whole Spike protein, and the scarce predicted immunological impact of the rare described variants.

Discussion

This work presents a confirmatory study of SARS-CoV-2 genetic drift and a framework to dissect viral genomic variability by focusing on single mutations.

As a limitation, we acknowledge that the consensus sequence generated by our workflow do not represent any particular clade nor viral isolate and do not take into account linkage among variations. However, the focus in specific mutational events can enable an easier constant for a virus that is undergoing millions of replications, for clinical severity and vaccine efficacy monitoring.

Regarding the immunogenomic analysis, it would be of great interest coupling the class 1 prediction with B- rather than T-cell recognition. This task can be carried out via prediction tools, such as BepiPred [19] and is currently being investigated.

At the time of the writing, the clinical impact of the pandemic in Italy has been decreasing in the whole country in terms of contagion rates and mortality absolute counts. Several doctors working in COVID centers reported an anecdotical decreased severity of disease symptoms. Many of them explained these findings as a critical consequence of the change in clinical protocols, contact tracing, and higher temperatures that should decrease viral load and symptoms. In other cases, a molecular impact of mutations is being appointed as a warm gun. Our analysis, according to other reports, shows no substantial genetic pressure on the 83 available Italian genomes.

List of abbreviations

BAM, Binary Alignment Map;COVID-19, COronaVIrus Disease 2019;HLA, Human Leukocyte Antigen; MHC, Major Histocompatibility Complex; NGS, Next-Generation Sequencing; RBD, Receptor-Binding Domain; SAM, Sequence Alignment Map; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; VCF, Variant Call Format; Immune Epitope Database and analysis resource (IEDB)

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All R and Bash/shell scripts developed are available at https://gitlab.com/bioinfo-ire-release/covid-miner along with test files and a walkthrough.

Conflicts of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by the Italian Ministry of Health (Ricerca Corrente 2019).

Authors' contributions

Conceptualization, MP, GC, LA.; Methodology, MP, AM.; Software, AS, LDA.; Validation, LDA, ES .; Formal Analysis, MP, LDA, AM, ES.; Writing – Original Draft Preparation, MP.; Writing – Review & Editing, MP, FP, GC.; Visualization, AM, LDA,ES.; Supervision, MP;

Acknowledgements

We thank Dr. Mattia D'Antonio for the valuable technical support.

Figure and Table Legends

Figure 1: A: Workflow of the Covid-Miner package. **B and C:** Lolliplop representation of SARS-CoV-2 variants at 5% frequency.

Figure 2: A: Heatmap of Receptor Binding Domain variants by normalized frequency. Countries with no occurrencies of RBD variants were removed. **B**: Multiple Correspondence Analysis (MCA) of Country-RBD variant association **C:** Heatmap representing immunogenicity scores for wild-type and mutated RBD epitopes.

Table 1: Full list of annotated variants found in the GISAID dataset.

Table 2: Matrix of variant-country association, counts normalized by *mutations per thousand isolates*.

Table 3: Epitope prediction binding via NetMHC4 analysis.

Table 4: Epitope mutation fold change ranking via IEDB analysis.

References

1. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV - A target for vaccine and therapeutic development. Nat. Rev. Microbiol. Nat Rev Microbiol; 2009. p. 226–36.

2. Yu J, Tostanoski LH, Peter L, Mercado NB, McMahan K, Mahrokhian SH, et al. DNA vaccine protection against SARS-CoV-2 in rhesus macaques. Science (80-) [Internet]. American Association for the Advancement of Science; 2020 [cited 2020 Jun 5];eabc6284. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.abc6284

3. Koyama T, Platt DE, Parida L. Variant analysis of COVID-19 genomes. J Bull World Heal Organ [Internet]. 2020 [cited 2020 Jun 5];2:1–21. Available from: https://www.researchgate.net/publication/339461351_Variant_analysis_of_COVID-19_genomes

4. Chiara M, Horner DS, Pesole G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. bioRxiv. Cold Spring Harbor Laboratory; 2020;2020.03.30.016790.

5. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2020 [cited 2020 Jun 5];2020.04.29.069054. Available from: https://doi.org/10.1101/2020.04.29.069054

6. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe. Cell Press; 2020;27:671-680.e2.

7. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Rydyznski Moderbacher C, et al. Journal Pre-proof Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell [Internet]. Cell Press; 2020 [cited 2020 Jun 15]; Available from: https://doi.org/10.1016/j.cell.2020.05.015

8. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. European Centre for Disease Prevention and Control (ECDC); 2017.

9. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. NIH Public Access; 2012;9:357–9.

10. bcftools [Internet]. [cited 2020 May 22]. Available from: http://samtools.github.io/bcftools/bcftools.html

11. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). Taylor and Francis Inc.; 2012;6:80–92.

12. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. Front Genet; 2012;3.

13. Bowtie 2: Manual [Internet]. [cited 2020 May 22]. Available from: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

14. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol [Internet]. BioMed Central; 2004 [cited 2020 Jun 5];5:R12. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-2-r12

15. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics [Internet]. 2015;32:511–7. Available from: https://doi.org/10.1093/bioinformatics/btv639

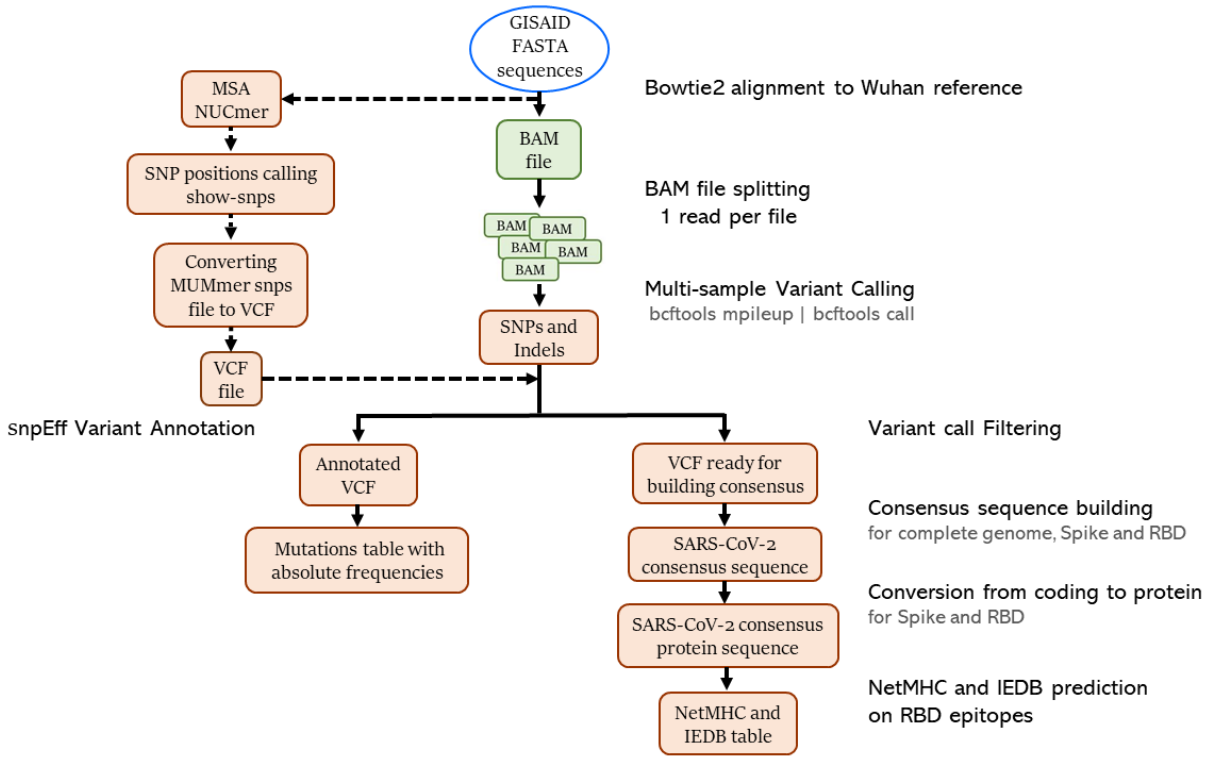
16. Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res [Internet]. 2019;48:D783–8. Available from: https://doi.org/10.1093/nar/gkz1029

17. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res [Internet]. 2018;47:D339–43. Available from: https://doi.org/10.1093/nar/gky1006

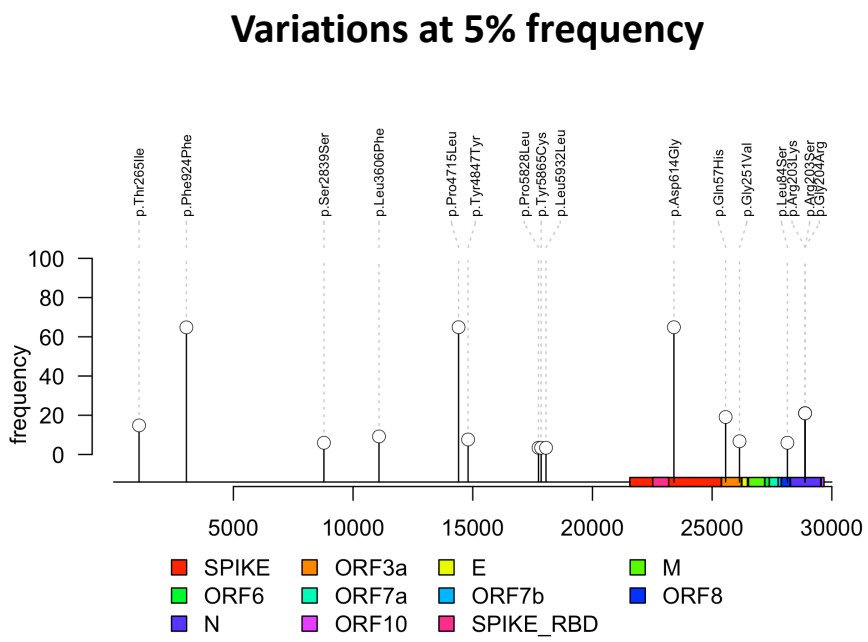
18. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. PLoS Comput Biol. PLoS Comput Biol; 2013;9.

19. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res [Internet]. 2017;45:W24–9. Available from: https://doi.org/10.1093/nar/gkx346

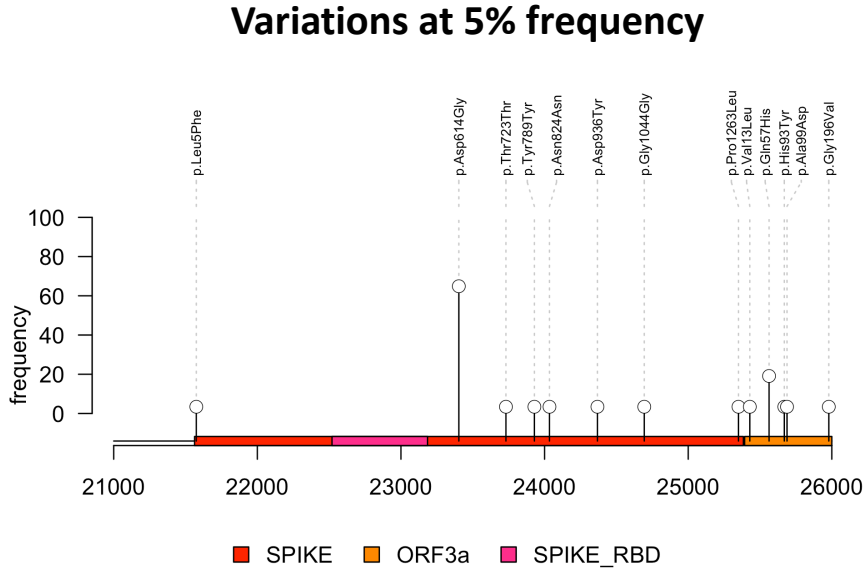
A


GISAID FASTA sequences
MSA NUCmer
Bowtie2 alignment to Wuhan reference
BAM file
SNP positions calling show-snps
BAM file splitting
1 read per file
BAM
Converting MUMmer snps file to VCF
Multi-sample Variant Calling
bcftools mpileup | bcftools call
SNPs and Indels
VCF file
snpEff Variant Annotation
Variant call Filtering
Annotated VCF
VCF ready for building consensus
Mutations table with absolute frequencies
Consensus sequence building
for complete genome, Spike and RBD
SARS-CoV-2 consensus sequence
Conversion from coding to protein
for Spike and RBD
SARS-CoV-2 consensus protein sequence
NetMHC and IEDB prediction on RBD epitopes
NetMHC and IEDB table

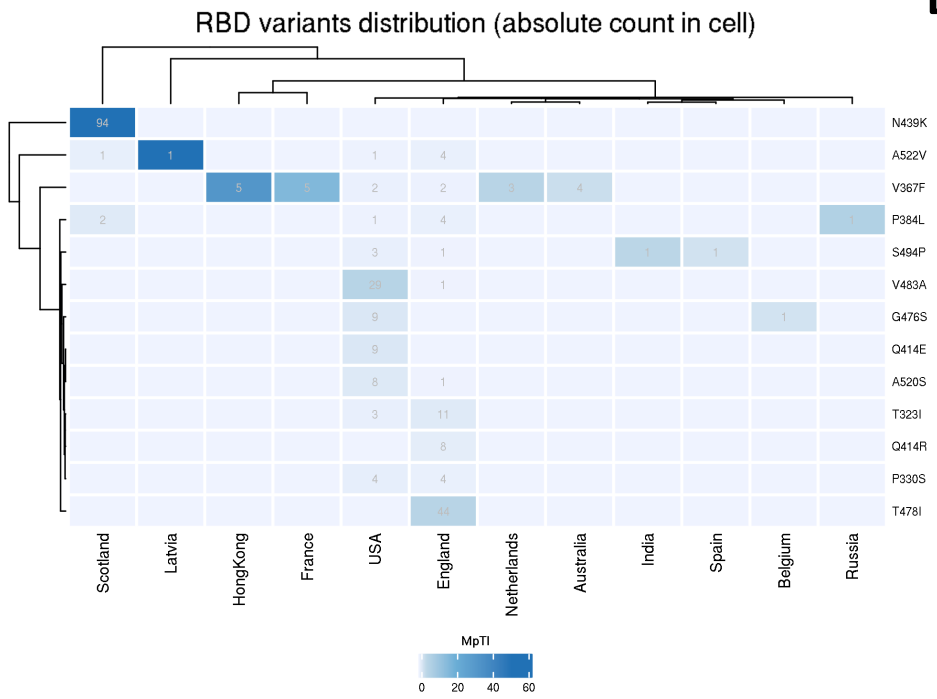
B


Variations at 5% frequency
p.Thr265Ile
p.Phe924Phe
p.Ser2839Ser
p.Leu3606Phe
p.Pro4715Leu
p.Tyr4847Tyr
p.Pro5828Leu
p.Tyr5865Cys
p.Leu5932Leu
p.Asp614Gly
p.Gln57His
p.Gly251Val
p.Leu84Ser
p.Arg203Lys
p.Arg203Ser
p.Gly204Arg
frequency
100
80
60
40
20
0
5000
10000
15000
20000
25000
30000
SPIKE
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
SPIKE_RBD

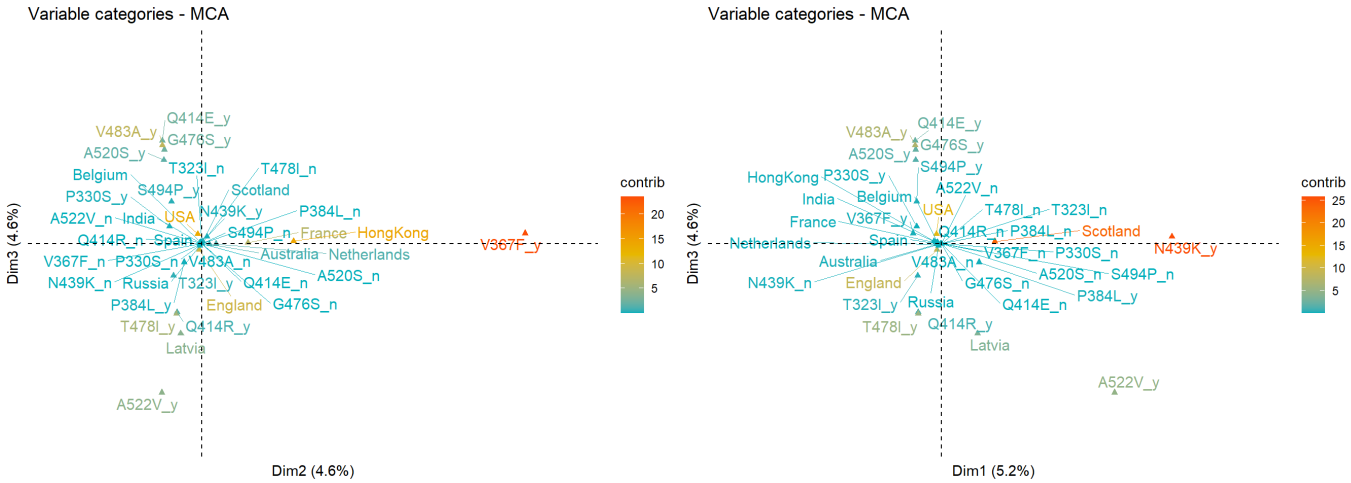
C


Variations at 5% frequency
p.Leu5Phe
p.Asp614Gly
p.Thr723Thr
p.Tyr789Tyr
p.Asn824Asn
p.Asp936Tyr
p.Gly1044Gly
p.Pro1263Leu
p.Val13Leu
p.Gln57His
p.His93Tyr
p.Ala99Asp
p.Gly196Val
frequency
100
80
60
40
20
0
21000
22000
23000
24000
25000
26000
SPIKE
ORF3a
SPIKE_RBD

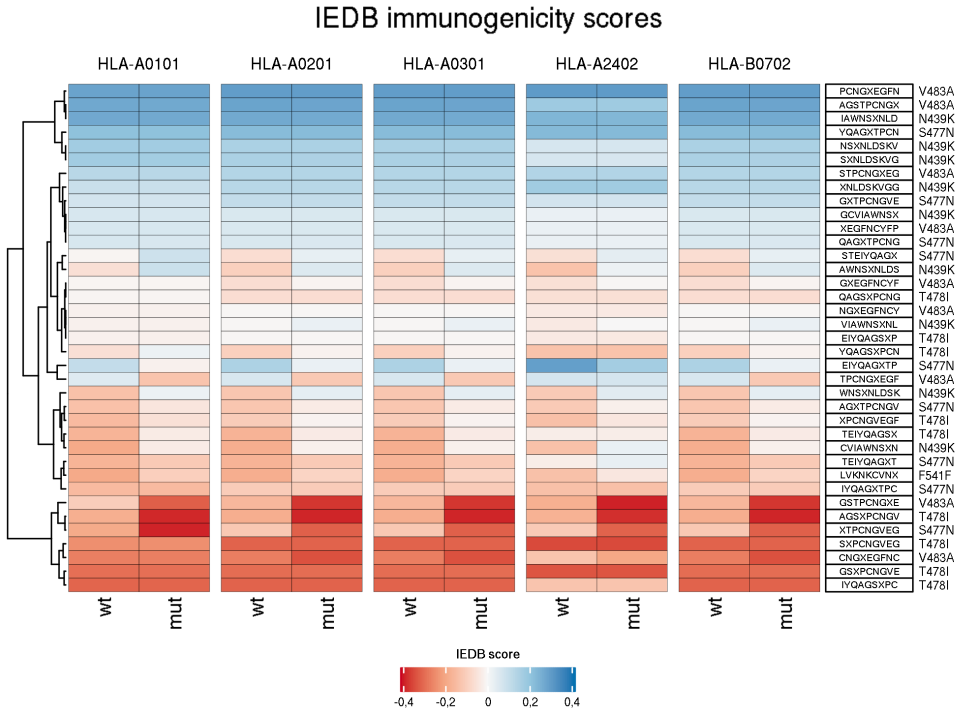
A


RBD variants distribution (absolute count in cell)
N439K
A522V
V367F
P384L
S494P
V483A
G476S
Q414E
A520S
T323I
Q414R
P330S
T478I
Scotland
Latvia
HongKong
France
USA
England
Netherlands
Australia
India
Spain
Belgium
Russia
MpTI
0
20
40
60

B


Variable categories - MCA
Dim3 (4.6%)
Dim2 (4.6%)
contrib
Variable categories - MCA
Dim3 (4.6%)
Dim1 (5.2%)
contrib

C


IEDB immunogenicity scores
HLA-A0101
HLA-A0201
HLA-A0301
HLA-A2402
HLA-B0702
wt
mut
IEDB score
-0.4
-0.2
0
0.2
0.4