Massacci et al. J Transl Med (2020) 18:494
https://doi.org/10.1186/s12967-020-02675-4

Journal of Translational Medicine

RESEARCH Open Access

Design of a companion bioinformatic tool to detect the emergence and geographical distribution of SARS-CoV-2 Spike protein genetic variants

Alice Massacci[1], Eleonora Sperandio[2], Lorenzo D'Ambrosio[3], Mariano Maffei[4], Fabio Palombo[1], Luigi Aurisicchio[1], Gennaro Ciliberto[5†] and Matteo Pallocca[2*†]

Abstract

Background: Tracking the genetic variability of Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) is a crucial challenge. Mainly to identify target sequences in order to generate robust vaccines and neutralizing monoclonal antibodies, but also to track viral genetic temporal and geographic evolution and to mine for variants associated with reduced or increased disease severity. Several online tools and bioinformatic phylogenetic analyses have been released, but the main interest lies in the Spike protein, which is the pivotal element of current vaccine design, and in the Receptor Binding Domain, that accounts for most of the neutralizing the antibody activity.

Methods: Here, we present an open-source bioinformatic protocol, and a web portal focused on SARS-CoV-2 single mutations and minimal consensus sequence building as a companion vaccine design tool. Furthermore, we provide immunogenomic analyses to understand the impact of the most frequent RBD variations.

Results: Results on the whole GISAID sequence dataset at the time of the writing (October 2020) reveals an emerging mutation, S477N, located on the central part of the Spike protein Receptor Binding Domain, the Receptor Binding Motif. Immunogenomic analyses revealed some variation in mutated epitope MHC compatibility, T-cell recognition, and B-cell epitope probability for most frequent human HLAs.

Conclusions: This work provides a framework able to track down SARS-CoV-2 genomic variability.

Keywords: SARS-CoV-2 genome, SARS-CoV-2 mutation, COVID mutations, SARS-CoV-2 vaccine, Bioinformatic workflow, Docker

Background

With more than 50 million infected people and 1.3 million deaths, SARS-CoV-2 is likely to have found a reservoir in the human population, as demonstrated by the current 2020 boreal autumn outbreak. To prevent the viral spread, an effective vaccine is needed.

More than 100 different vaccines are in development worldwide, including China, India, the USA, and Europe. Most, if not all of them, target the Spike protein, the viral product able to bind the human receptor angiotensin-converting enzyme 2 (ACE2). These designs use different formulations or platforms, such as a vectored vaccine or nucleic acids, RNA, and DNA. The Spike protein is the

*Correspondence: matteo.pallocca@ifo.gov.it
†Gennaro Ciliberto and Matteo Pallocca contributed equally to this work
[2] Biostatistics, Bioinformatics and Clinical Trial Center, IRCCS Regina Elena National Cancer Institute, Rome, Italy
Full list of author information is available at the end of the article

best candidate for historical reasons [1] and recent preclinical evidence in non-human primates [2].

One of the main challenges is that 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 framework to address this potential issue.

A few preliminary reports already described both the main differences with the SARS-CoV virus [3] and the variant landscape of SARS-CoV-2 in different clades [4–6] Both pointed out the limited presence of functional mutations in critical regions of the genome; some groups also released open bioinformatic web applications to browse the virus variants [7]. Nevertheless, some other reports stressed a possible selective pressure on the D614G variant, the only frequent variation of the spike protein [8], recently providing in-vivo evidence of its increased fitness [9] . A bioinformatic analysis of SARS-CoV-2 epitopes showed high homology with SARS-CoV [10] and, thus, characterized many possible B- and T-cell epitopes, providing 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 [11]. Of note, it was recently reported from a study involving the plasma of 650 SARS-CoV-2 exposed patients that 90% of the serum or plasma activity targets the Receptor Binding Domain (RBD) of the Spike protein, the central point of SARS-CoV-2/ACE2 contact.

To provide an evidence-based approach, we present a freely available protocol to enable all the bioinformatic community to dissect SARS-CoV-2 genomic variability and its functional impact on DNA vaccines and an openly accessible web portal to browse the current most frequent viral mutations. This framework can support researchers to automatically mine data 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.

Methods

Bioinformatic framework

Following the overall workflow (Fig. 1a), a multi-FASTA file of viral sequences is aligned against the Wuhan strain (NC_045512.2) using NUCmer from the MUMmer package [12]. Filters used to fetch the sequences analyzed in this manuscript were Host = Human, Virus = hCov-19, flagged "complete", "high coverage," and "low coverage

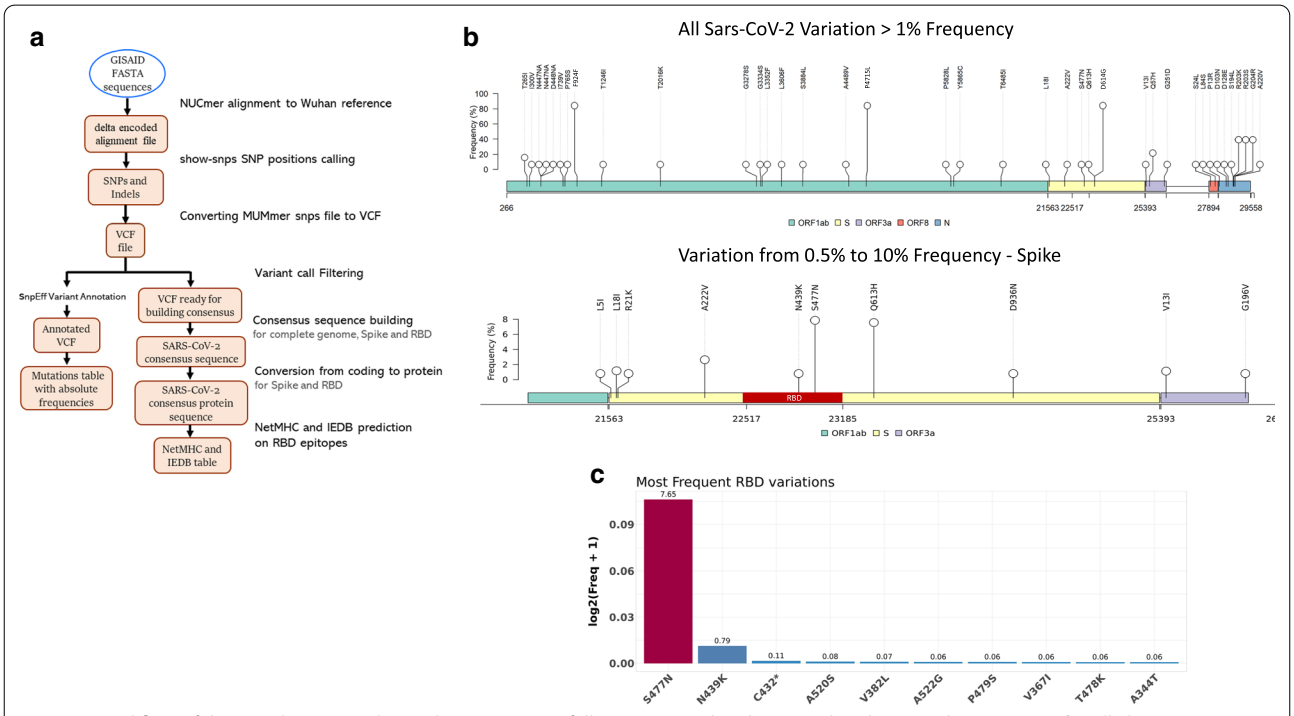

Fig. 1 **a** Workflow of the Covid-Miner package. The execution is fully automatized and encapsulated in a Docker container. **b** LolliPlot representation of SARS-CoV-2 variants at 1% frequency for the whole SARS-CoV-2 genome and from 0.5 to 10% for the Spike protein locus, showing the most frequent RBD variants at codon 447 (7%) and 439 (0.7%). **c** Barplot from occurrences of most frequent RBD variations

exclusion". Nucmer generates a delta encoded alignment file, which is then parsed using the show-snps utility. This produces a catalog of all the SNPs and indels internal to the alignments contained in the delta encoded file. Show-snps output is then converted into standard VCF format using code adapted from [13].

At this point, it is necessary to annotate 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 complete viral 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 [14] that embeds the NC_045512.2 genome assembly annotation in its standard package. A human-readable (and easily parsable) table is formatted thanks to the *SnpSift* jar package [15] of *snpEff*. 2D visualizations of the variants on the viral genome are generated via the LolliPlot function of the trackVignette's R package [16].

The whole toolset 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 for portability and reproducibility purposes, is available on https://gitlab.com/bioinfo-ire-release/covid-miner.

Downstream immunogenomic analyses were carried out by NetMHC for Class 1 recognition [17], IEDB for T-cell immunogenicity [18] and BepiPred for B-cell epitopes [19].

The web portal has been implemented in Angular (version 10.1) over the Bootstrap CSS framework (version 4.5) by leveraging on D3js library (version 6.2) for the graphical representations. The backend is written in Python language by adopting the Flask web framework (version 1.1.2).

Results

Analysis results on 93,930 SARS-CoV-2 sequences

We extracted 93,930 high-quality sequences (available on October 2, 2020) from the the EpiCov™ section of the GISAID portal [20] that acts as a worldwide repository of viral isolates. Every viral isolate contains zero or more variants with respect to the Wuhan strain. The fully annotated table of 20,640 variants is available in Additional file 1: Table S1.

Considering the total number of sequences, the most frequent Spike protein variation is confirmed to be D614G (Fig. 1b). Next, there are a few variants with a frequency of 2–7% that are slowly rising in the viral populations, namely S477N, Q613H and A222V. The S477N mutation lies in the RBD locus, while all the other RBD variants are below the 1% penetrance threshold (Fig. 1c, Additional file 1: Table S2).

We asked whether the RBD variants are strongly associated with a certain geographical region, to this purpose we propose to measure "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.

The geographical distribution shows how the S477N variant is strongly rooted in Australia, and the N439K is associated with clusters starting from the United Kingdom (Scotland). However, none of the most frequent variants are uniquely associated with one country as expected from the worldwide virus distribution, and clusters of co-occurring variants lie mostly in countries with the highest number of available sequences (i.e., USA, England) (Additional file 1: Table S2, Fig. 2a, b). In order to better understand the evolution of variants over time and space, we tracked the location of all the source isolates carrying these two variations (Fig. 2c). The S477N has been firstly identified in Colombia and is harbored in more than 60% of the isolates sequenced in Australia from June 2020. On the other hand, the N439K is dominating the isolates landscape in Ireland and England from August 2020.

Immunogenomic analysis

These novel RBD variants may have several biological and putatively clinical impacts on the virus functions. For instance, every protein-coding variant changes several epitope sequences presented on the human cell's surface from the Major Histocompatibility Complex (MHC). This, in turn, can have an impact on the human immune system recognition by T and B-cells. To shed light on these processes, we computationally modeled

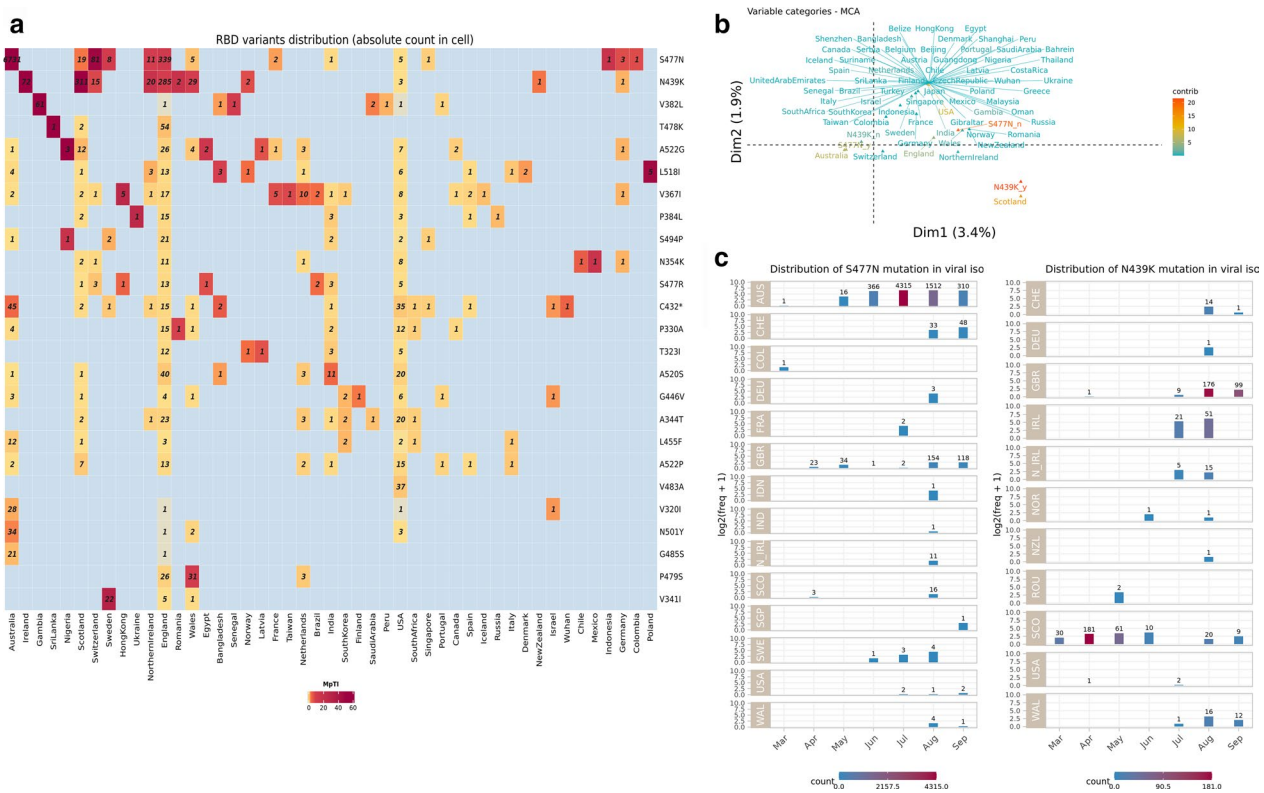

Fig. 2 **a** Heatmap of Receptor Binding Domain variants geographical distribution by normalized frequency. Countries with low occurrences of RBD variants (less than 5) were removed. **b** Multiple Correspondence Analysis (MCA) of Country-RBD variant association. **c** Evolution of S447N and N439K in space and time. Y-axis represents the relative frequency of that mutation in every country isolate set for each month. X-axis: months in 2020, as annotated in GISAID sequences. Every bar is annotated with absolute counts

the immunological impact of SARS-CoV-2 epitopes in terms of (1) MHC Class 1 presentation of antigens (2) T-cell immunogenicity (3) B-cell epitope prediction.

Considering the only RBD mutations above 0.1% frequency, we performed MHC class 1 binding prediction for both the wild-type (Wuhan strain) and the mutated strain, choosing the most frequent HLAs in public databases [21]. The software generated predictions for all possible 9-mers resulting from the full RBD sequence, and we computed how *binders* were calculated for all the considered HLAs. Out of 18 predictions, only one epitope had binding affinity in mutated or wild-type, with no change in binding affinity caused by the mutation (Additional file 1: Table S3).

Mutation	WT	Mutant	Wild-Type binders	Mutated binders
S477N	IYQAGSTPC	IYQAGNTPC	1	1

Then, 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) [18]. When considering T-cells with the class 1 immunogenicity tool [22], 54/189 (29%) 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, IYQAGNTPC (S477N) shows increased immunogenicity for all considered HLAs (Fig. 3a). These results point out to a putative variability in T-cell response mediated by these mutations.

Testing epitopes for putative B-cell recognition remains an analytical challenge; only a few algorithms have been developed for this purpose [23]. When focusing to all the mutated epitopes caused by S477N and N439K, none cause a shift of the amino acid exposition, as they are both predicted to be in the *exposed* status. The overall Epitope score, that takes into account a variable amino acid window surrounding the mutation, is slightly increased, from 0.535/0.516 in the WT sequence to 0.561 and 0.548 for S477N and N439K, respectively (Additional file 1: Table S5).

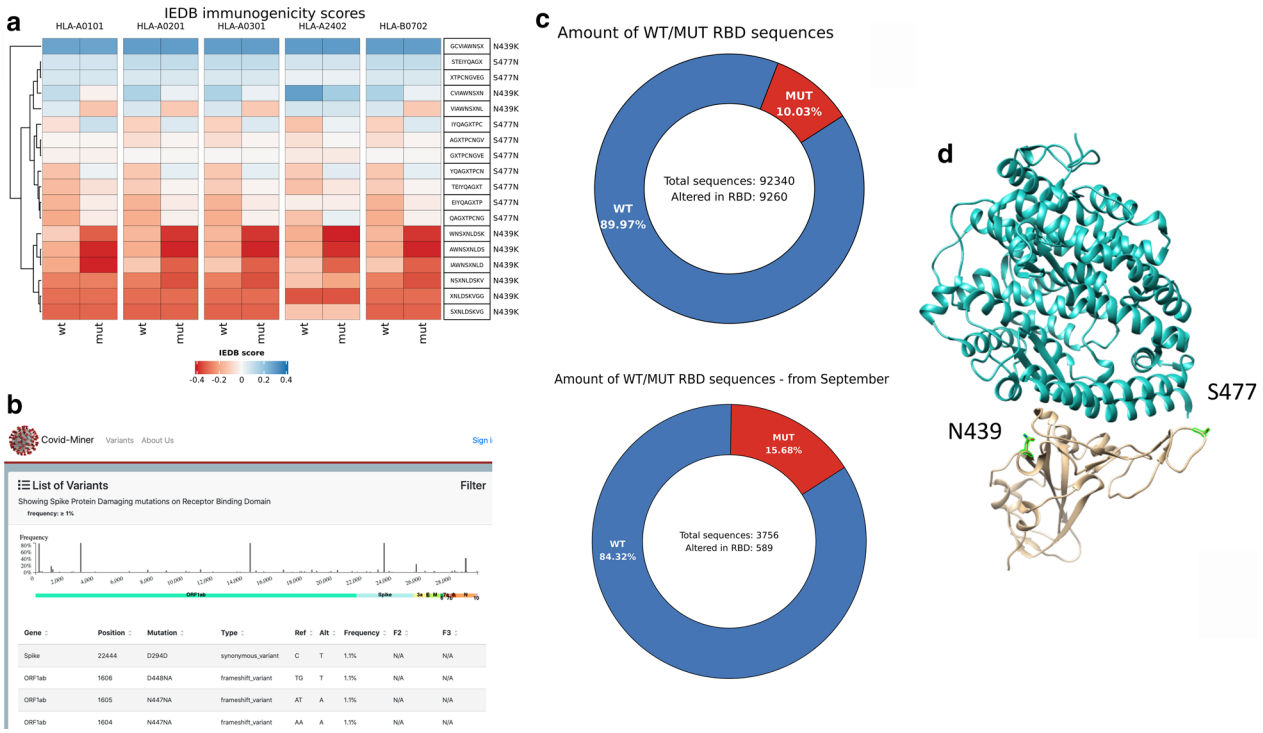

Fig. 3 **a** Heatmap representing T-cell immunogenicity scores for wild-type and mutated RBD epitopes. **b** Screenshot of covid-miner web portal homepage. **c** Donut plot of RBD mutational impact of overall sequences and September 2020 sequences. **d** Location of the most frequent RBD mutated amino acids N439 and S477

Covid-miner data portal

In order to share the results with a wider public, we created a frontend portal for the analysis results, freely accessible at https://covid-miner.ifo.gov.it. The web app features two sections; a *Variant* section is dedicated to browse and visualize the most frequent viral mutations over the genome, and a *Geographical* distribution heatmap that displays the association among countries and the most frequent RBD mutations. The home page displays the amount of wild type / mutated RBD variants as a main summarizing figure (Fig. 3b, c).

Discussion

This work presents a bioinformatic toolset, a confirmatory study, and a web portal focused on SARS-CoV-2 genetic drift and a framework to dissect viral genomic variability by focusing on single mutations. Our analysis revealed the strong emergence of one RBD variant, possibly crucial for the virus infectivity potential.

The RBD is located from aa319-541 (one of the two crystallography analysis reports it to be at aa333-526) [24, 25]. However, the major component of ACE2 contact lies in the Receptor Binding Motif (RBM), located in the aa438-506 central part. Both the S477N and the second-most frequent RBD variation N439K lie in the RBM, suggesting a selective pressure on this locus (Fig. 3d).

From the polarity point of view, only the N439K switch has a change from Neutral to Polar; a recent structural topology article provided an in-depth analysis of the Energy free change of the most frequent single and clustered mutations, showing that the N439K has a strong increase of Binding Free Energy (BFE) [26]. The S477N has only a slight BFE increase that does not seem to reflect its relative increase in frequency.

Another critical mutational effect is the putative change in antibody affinity. N439K was recently reported to reduce affinity to one of the described antibodies, H00S022, while no data is still available for S477N [27].

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

Conclusion

Tracking down and monitoring SARS-CoV-2 genomic evolution has a dramatic impact on disease severity and vaccine efficacy, even if these two variables are influenced by many factors, such as individual patient's clinical and genetic status. Nonetheless, these novel tools can create a framework to deal with the next viral pandemic waves.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-020-02675-4.

Additional file 1: Table S1. Full list of annotated variants found in the GISAID dataset. **Table S2.** Matrix of variant-country association, counts normalized by *mutations per thousand isolates*. **Table S3.** Epitope prediction binding via NetMHC4 analysis. **Table S4.** Epitope mutation fold-change ranking via IEDB analysis. **Table S5.** BepiPred results of B-cell epitope prediction for Wild Type, S477N and N439K Spike sequences.

Abbreviations

COVID-19: COronaVIrus Disease 2019; HLA: Human Leukocyte Antigen; MHC: Major Histocompatibility Complex; NGS: Next-Generation Sequencing; RBD: Receptor-Binding Domain; RBM: Receptor-Binding Motif; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; VCF: Variant Call Format; IEDB: Immune Epitope Database and analysis resource.

Acknowledgements

We thank Matteo Schiavinato on the valuable technical support on the Variant Calling files. We thank Dr. Frauke Goeman and Gabriele Lo Iudice for editorial assistance. The whole analysis was enabled by the GISAID portal.

Authors' contributions

Conceptualization, MP, GC, LA.; Methodology, MP, AM.; Software, AS, LDA.; Validation, LDA, ES.; Formal Analysis, MP, LDA, AM, ES, MM; Writing—Original Draft Preparation, MP.; Writing—Review & Editing, MP, FP, GC.; Visualization, AM, LDA,ES, MM; Supervision, MP; All authors read and approved the final manuscript.

Funding

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

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.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Takis srl, Rome, Italy. [2] Biostatistics, Bioinformatics and Clinical Trial Center, IRCCS Regina Elena National Cancer Institute, Rome, Italy. [3] Tumor Immunology and Immunotherapy Unit, IRCCS Regina Elena National Cancer Institute, Rome, Italy. [4] EvviVax srl, Rome, Italy. [5] Scientific Direction, IRCCS Regina Elena National Cancer Institute, Rome, Italy.

Received: 13 July 2020 Accepted: 11 December 2020
Published online: 30 December 2020

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. 2009;7:226–36.
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. 2020;69:806. https://doi.org/10.1126/science.abc6284.
3. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581:221.
4. Koyama T, Platt DE, Parida L. Variant analysis of COVID-19 genomes. J Bull World Heal Organ. 2020;2:1–21. https://www.researchgate.net/publication/339461351_Variant_analysis_of_COVID-19_genomes
5. Chiara M, Horner DS, Pesole G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. bioRxiv. 2020. https://doi.org/10.1101/2020.03.30.016790.
6. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18:179. https://doi.org/10.1186/s12967-020-02344-6.
7. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. coronapp: a Web Application to Annotate and Monitor SARS-CoV-2 Mutations. J Med Virol. 2020. https://doi.org/10.1002/jmv.26678.
8. 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. 2020. https://doi.org/10.1101/2020.04.29.069054.
9. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2020. http://www.nature.com/articles/s41586-020-2895-3
10. 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. 2020;27:671–80.
11. 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. 2020. https://doi.org/10.1016/j.cell.2020.05.015.
12. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLOS Comput Biol. 2018;14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
13. Schiavinato M. MatteoSchiavinato/Utilities: general purpose tools for every-day sequencing bioinformatics. If you use any of these tools, please acknowledge this repository (there are no publications). Let's all help each other). https://github.com/MatteoSchiavinato/Utilities
14. 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. 2012;6:80–92.
15. 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. 2012;3:35.
16. Ou J, Zhu LJ. trackViewer: a bioconductor package for interactive and integrative visualization of multi-omics data. Nat Methods. 2019. https://doi.org/10.1038/s41592-019-0430-y.
17. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan 4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. Netmhcpan-40 Improv Pept Cl I Interact Predict Integr eluted ligand Pept Bind Affin data. bioRxiv. 2017. https://doi.org/10.1101/149518.
18. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2018;47:D339–43. https://doi.org/10.1093/nar/gky1006.
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. 2017;45:W24–9. https://doi.org/10.1093/nar/gkx346.
20. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. Eurosurveillance. Solna Municipality: European Centre for Disease Prevention and Control (ECDC); 2017.
21. Gonzalez-Galarza FF, McCabe A, dos Santos EJM, 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. 2019;48:783–8. https://doi.org/10.1093/nar/gkz1029.
22. 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. 2013;9:1003266.
23. Galanis KA, Nastou KC, Papandreou NC, Petichakis GN, Iconomidou VA, Vassiliki A, et al. Linear B-cell epitope prediction: a performance review of currently available methods. Doi:https://doi.org/10.1101/833418
24. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nat Res. 2020;581:215–20. https://doi.org/10.1038/s41586-020-2180-5.
25. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020; 367:1260.
26. Chen J, Wang R, Wang M, Wei GW. Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol. 2020;432:5212–26.
27. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell. 2020;182:1284–94.

Publisher's Note