Viral Instant Mutation Viewer (VIMVer) : a tool to speed up the identification and analysis of new SARS-CoV2 emerging variant and beyond

Vincent Wilde, Bruno Canard, François Ferron

Viral Instant Mutation Viewer (VIMVer) : a tool to speed up the identification and analysis of new SARS-CoV2 emerging variant and beyond

Wilde Vincent[1], Canard Bruno [1,2] and Ferron François [1,2,*]

[1] Aix-Marseille Université and Laboratoire Architecture et Fonction des Macromolécules Biologiques (AFMB), CNRS – UMR-7257, 13288 Marseille, France.
[2] European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany.

* To whom correspondence should be addressed. Tel: +33 (0)491828628 ; Email: francois.ferron@univ-amu.fr

ABSTRACT

The appearance of genetic variants impacts vaccination efficiency and therapeutic options, generating a need to map and relate mutations observed in the proteome and the genome. We develop an user-friendly web-service software (Viral Instant Mutation Viewer or VIMVer) which allows a direct identification of mutations in the genome and its counterpart in the viral proteome. Since its emergence in 2019, the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), responsible for the COVID-19 pandemic, has generated an overwhelming amount of data while becoming one of the most studied viruses of the *Nidovirales* family. We originally develop this tool during the COVID pandemic, thus for any SARS-CoV-2 nucleotide sequence the web-service gives a fast identification, mapping, and display of new mutations simultaneously at the nucleotide and amino acid level in comparison to a reference sequence (Wuhan-1). Furthermore, the lineage or the relative position to known lineage of the variant of interest is available on the link to Phylogenetic Assignment of Named Global Outbreak LINeages (Pangolin COVID-19). The workflow presented here is available through a web service (https://vimver.afmb.univ-mrs.fr/). The source code is released under public license and can be easily adapted for further development to other viruses.

1. INTRODUCTION

Coronaviruses (CoV) are large genome, positive-strand RNA viruses of the order *Nidovirales* that have recently attracted global attention due to the ongoing COVID-19 pandemic. Despite significant efforts to control its spread, through worldwide vaccination campaign strategies, SARS-CoV-2 is still causing substantial health and economic burden, emphasizing the need for a continuous monitoring of its genome evolution to identify the circulating and emerging SARS-CoV-2 lineages (1). Thus, the scale of the COVID-19 pandemic has led to unprecedented efforts by the research community to rapidly identify variant sequences, to test therapeutics

and vaccines, to understand the molecular basis of SARS-CoV-2 entry, pathogenesis, and immune targeting in light of these emerging mutations. This effort resulted in the availability of more than 15 million genomic sequences available through the GISAID data portal (https://www.gisaid.org/) in June 2023. Whole genome sequencing allows to discriminate from one lineage to the other, and spot variant of concerns (VOCs). Next generation sequencing (NGS) methods are widely used, from hospitals to academic research labs, and yet the process to rapidly identify mutant sequences and its impact at the protein level is still tedious maybe less so since the development of tools like nextclade (2). In this work, we present a workflow that easily allows the user to identify the mutation hot spots on the genome, to assign its lineage and instantly visualize its potential impact at the protein level. This resource can help practitioners, and researchers to quickly discriminate the new emergence of mutations at the genome level and its potential impact at the protein level. VIMVer is publicly available at : https://vimver.afmb.univ-mrs.fr/.

2. MATERIAL AND METHODS

Our goal is to propose a tool to quickly analyze viral genomes VIMVer rapidly detects mutations and their consequences on the protein structure relative to a reference strain present in the database. VIMVer workflow is developed under Ubuntu 22.04 LTS and is an early stage web-service developed in python 3.8+ using the Django framework, and meant to be friendly to the virology community. Django has a heavy and rigorous but adaptable architecture which simplifies any integration of new functions during the development.

The workflow is presented in Figure 1 and can be summarized as follow : a nucleic sequence is submitted as an input to a search engine currently a blastn (3) to query the VIMVer reference database. The multiple sub-alignments generated by blastn are optimized through a realign procedure using Muscle version (4) to proof the alignment. All alignments are done using the nucleotide sequences and are dynamically translated to generate the corresponding protein alignment. The protein alignment reflects the genomic variability and is independent of amino acid composition. The final step is the dynamic visualization of both alignments (nucleotide / amino acids) which allows to spot mutations silent of not. User can submit the resulting output to pangolin webserver to assess lineage.

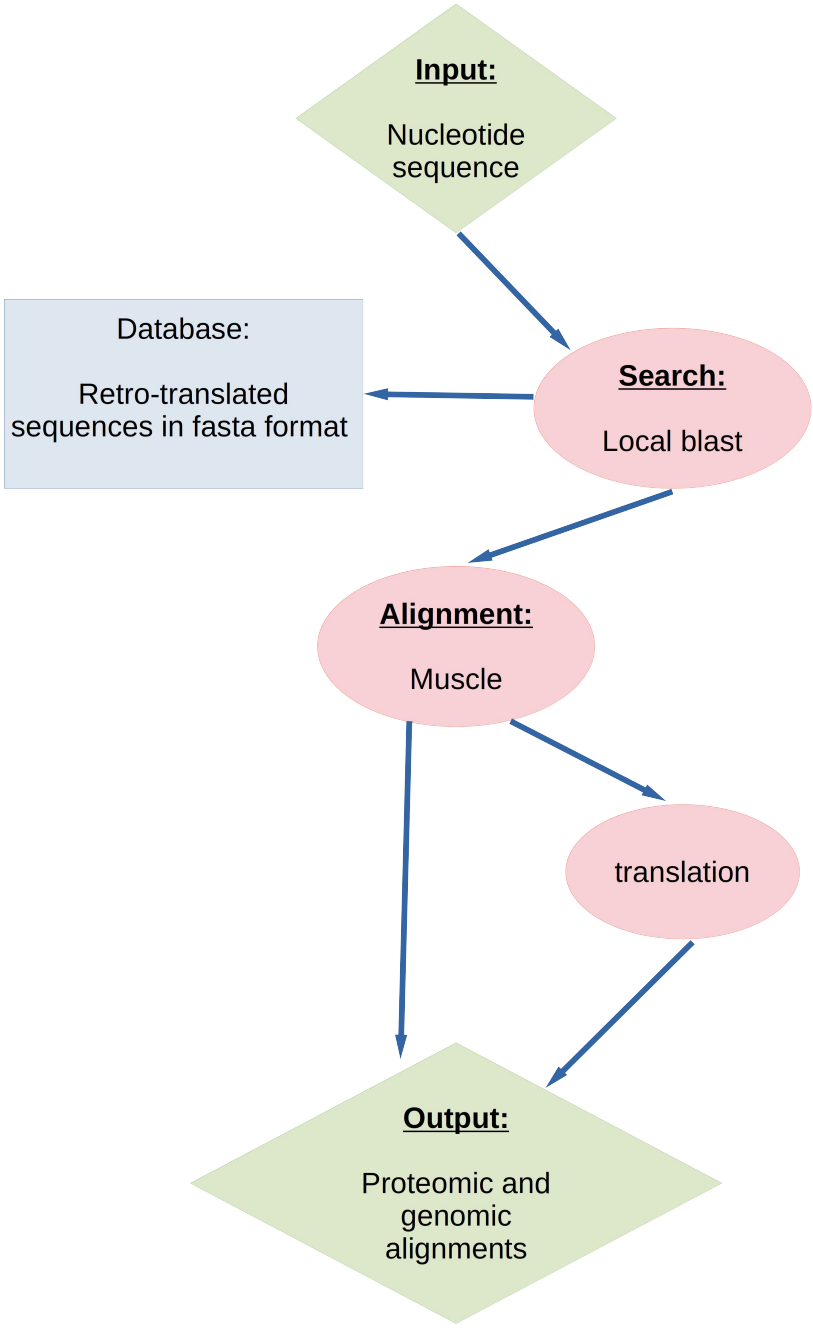

Figure 1. The workflow scheme of VIMVer.

2.1. Database

Any viral reference sequence could be implemented in the database and constituted of the retro-translated genomic sequence(s) from the corresponding viral proteins sequences. In this context, viral genome sequences are treated as modular sequences representing their translated states. Therefore, genomes coding for polyproteins, which are subsequently processed into multiple proteins, are pre-annotated at the genomic level. Currently, the development focus of VIMVer lies on

the SARS-CoV-2 genome. Consequently, the database currently consists of the complete SARS-CoV-2 genome (Wuhan-1: NC_045512.2 / YP_009725297.1), encompassing 26 proteins in both nucleotide and amino acid sequences. Each sequence type is organized into separate files in FASTA format. To accomplish this, the entire genome was divided into sub-sequences to match each functional protein. These nucleotide sub-sequences were meticulously retro-translated to ensure the preservation of codon-wise reading frames and to address ribosomal frameshifts, a characteristic feature of CoV genomes. Notably, the first open reading frame (ORF) of CoV involves a ribosomal frameshift, leading to the synthesis of two polyproteins, pp1a and pp1ab (8). This frameshift has implications for the translation of the landmark protein nsp12, as its N-terminal is encoded at the end of ORF1a, overlapping nsp11, while the remaining coding sequence is shifted by +1 frameshift. In our local database, we have merged nsp11 at the beginning of nsp12 to align with the coding reading frame, resulting in the duplication of the corresponding retro-translated sequence's n^{th} nucleotide. Therefore, both genomic and protein nsp12 sequences are accurately assigned and numbered in our database.

Furthermore, Django utilizes the MySQL database to store all items from the created models, including user request results, which are only accessible for 24 hours. VIMVer generates FASTA files containing the resulting alignments.

2.2 Search engine and Alignments

As shown in Figure 1, we are using Blastn (3) mainly to confront the query to our database. Blastn parameters are left by default. In case that the inquiry is too distant from the reference or that the inquiry's sequence has too many ‘N’s (unspecified nucleotide(s)), blastn produces multiple sub-alignments for a hit. Muscle (v3.8) is used (4) to correct the alignment (5). All alignments are done in respect to the reference nucleotide sequences to serve as anchors for the following codon-wise co-translation. Here Blastn was selected for it ease to deploy and parametrization considering the size of the database.

The output is a list of as many hits blastn found for a maximum sequences of the prepared dataset. Each, hit opens a two frames window that shows on the left panel the proteins alignments and on the right panel the nucleotides alignments. Sequences are interactive so one can follow its position on the sequence using mouse cursor on the corresponding sequence. Mutation are highlighted by an asterisk under each alignment. Outputs are downloadable in multi-fasta format.

2.3. Accessibility

Each job runs under few seconds and is identified by a key made of 6 printable characters, generated when a query is launched. This job name allows the users to retrieve their run within 24 hours. Additional information is available on the project’s GitHub.

2.4. Deployment

Viral Instant Mutation Viewer is deployed on a apache2 server hosted from our laboratory server. The *ssl* protocol was obtained with Encrypt's services. For users or developers who want to install VIMVer locally the source code is freely available, however it will not be functional straight away after download. Indeed the settings and configurations file has to be protected. A setting file (setting.py) including an encryption key used to run Django functionality needs to be present in the path This setting file is generated once by the administrator when creating the VIMVer project within django. Moreover, it is also within the setting file that the administrator will define its server setting prior deployment Details for installation are provided in the GitHub repository .

3. RESULTS

3.1 web interface and practical example

VIMVer is a bioinformatics tool that quickly shows mutations on the user's query with the proper numbering both in genome and protein. Our goal is to propose a tool to quickly analyze viral genomes, and for any SARS-CoV2 sequence of interest, VIMVer rapidly detects mutations and their consequences on the protein structure relative to the Wuhan-1 reference strain. This is particularly important in the case of the frameshift which obscures the recognition of the 1b fragment. Our web tool is composed of 3 pages respectively named : query, finder, viewer. VIMVer is accessible through the internet using any web navigator.

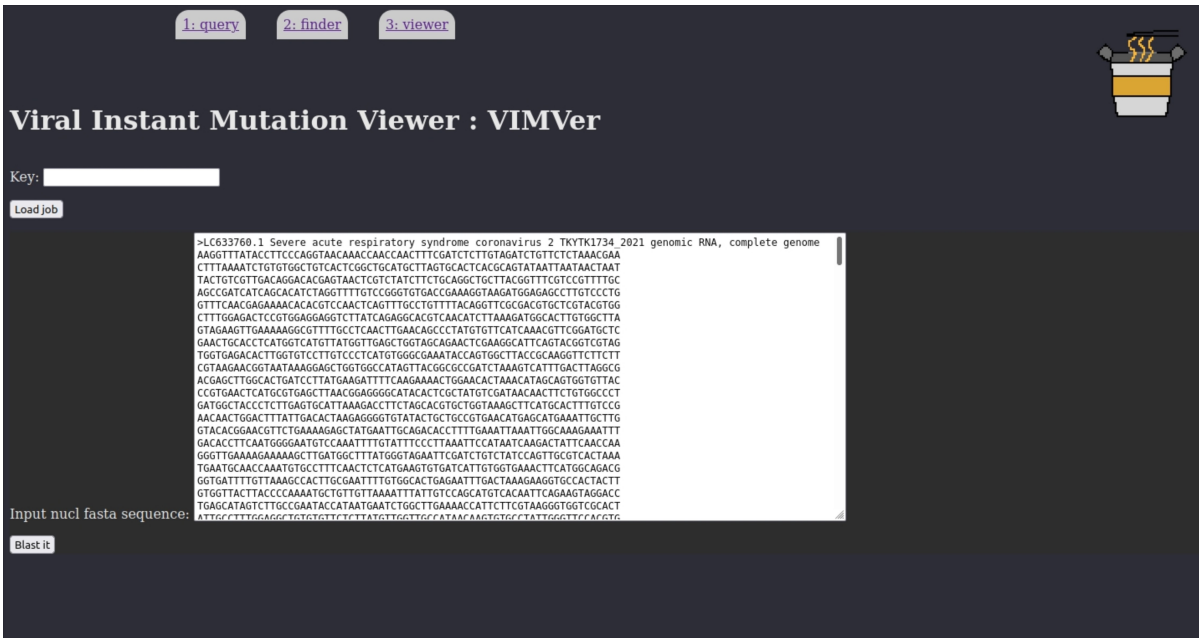

Figure 2. Snapshot of the submit form page of VIMVer: A fasta sequence with the header can be pasted in the form section as query.

The user can paste his query in fasta fomat in the homepage shown in Figure 2. The inquiry must be a nucleotide sequence possibly from a sequencing result in fasta format, the length can range from a random recommended minimum of 30 nucleotides, or a single coding sequence, to a whole viral genome. In this version queries have to be SARS-CoV or SARS-CoV2 related to retrieve hits from the blastn. When the query is submitted : VIMVer generates a job-key, which is only shown once. On the same page, there is a link to the PANGOLIN web-service (6) (link to:pangolin), to retrieve information to determine the lineage of the query sequence.

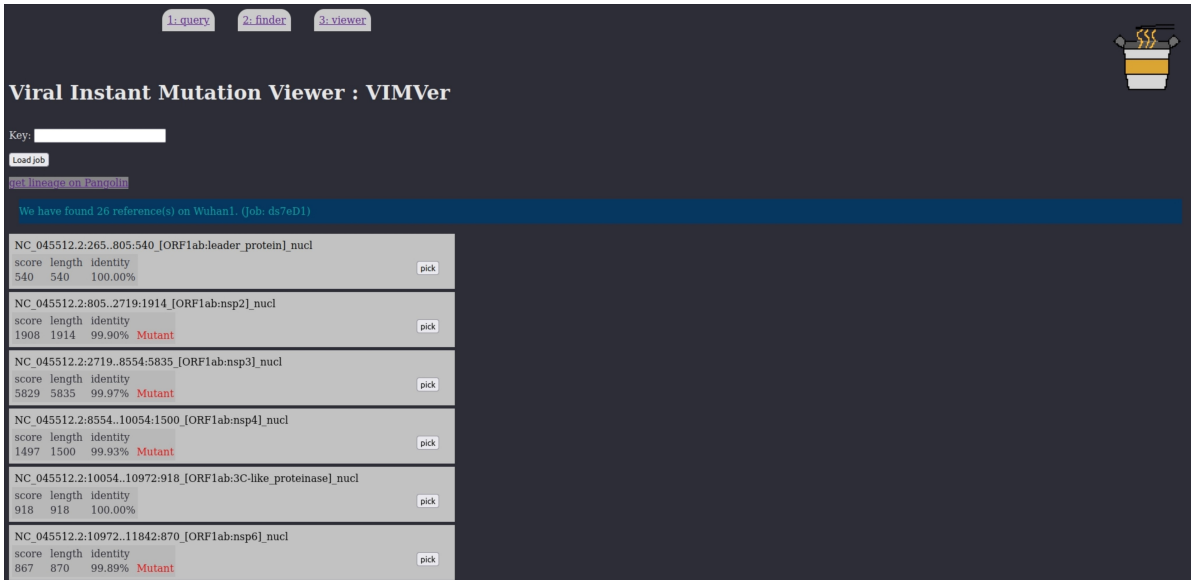

Figure 3. Snapshot of the resulting VIMVer search : The hits are presented as a list with an id number corresponding to the identified domain. The score, the matching length and the sequence identity to the reference sequence are shown. If non silent mutation are detected then a Mutant mention appears in red. A pick button allows to access individual results.

The result, screened in Figure 3 is returned as a table for which each entry corresponds to a “blastn” hit, described by the score, length and identity percentage. Considering that our database is small, the E-value is not relevant and therefore not shown, albeit calculated, instead the score is shown as it represents the number of aligned residues which allows to infer the identity percentage. If the query nucleotide sequence is not a silent mutation, then the corresponding hit is marked as “Mutant” in red. At the end of the page, the user can download the blastn results in fasta format. To access the resulting alignment sampled in Figure 4, click on the ‘pick’ button on the same line.

As proof of concept, we tested our project in a real-life scenario. To do so, we are looking for two well-known proteins: the domain RdRp and the spike protein S

Indeed, the S protein is the third ORF of the SARS-CoV2 genome, and the gene most prone to accumulate mutations that influence the infectivity of emerging variant (7). If one wants to observe and analyze the spike nucleotide and protein sequences evolution, straight out of sequencing campaign or from database analysis VIMVer makes the task easy and quick.

The nsp12 gene (the main replicative RdRp core) is also a point of interest for our project since it is often reported with incorrect numbering due to the *nidovirales* frameshift (8, 9).

As an example, a random entry in GISAID (EPI_ISL ID : 14769247) is selected and examined for mutations. We selectively focus on the RdRp domain for the frameshift management and the spike protein for its high mutation susceptibility . The goal is to process a whole genome, that is why the test sequence has to be a complete genome with a high coverage rate. On the query page of VIMVer, a single fasta sequence was pasted in the input text box.

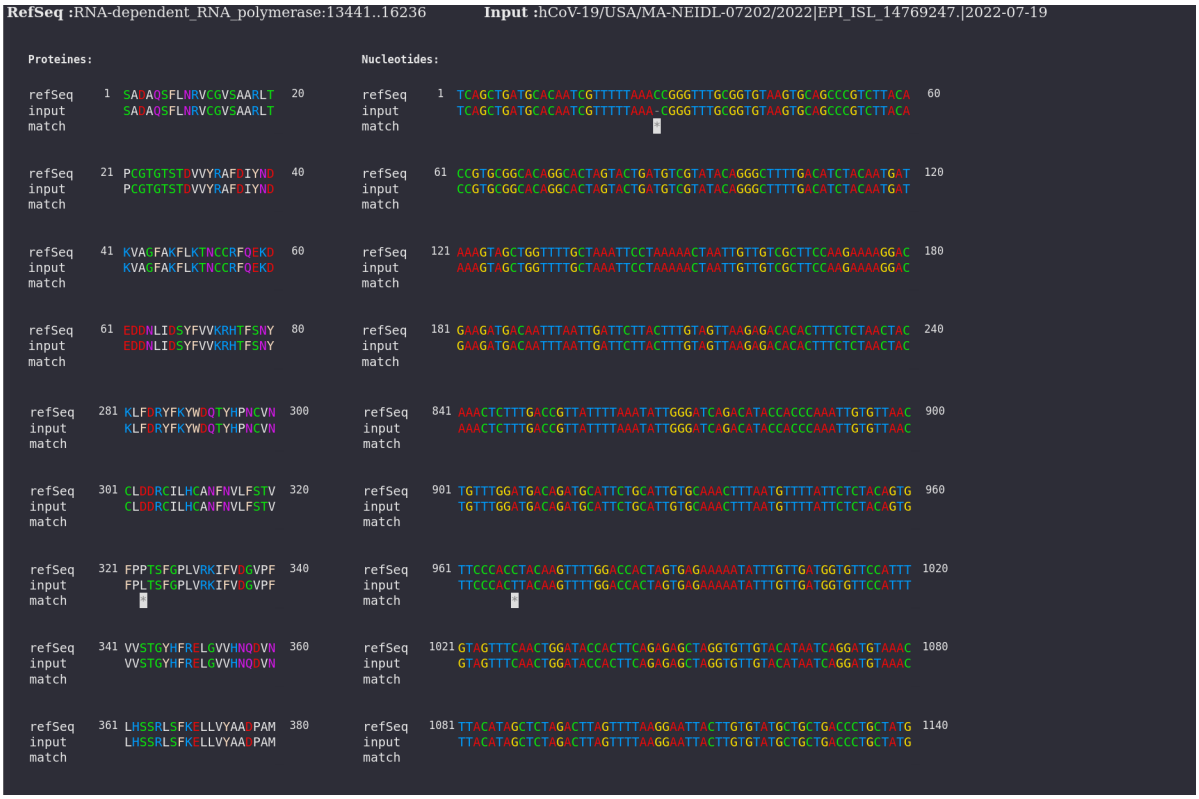

Figure 4. VIMVER result page presenting a dual alignment of the Gisaid entry EPI_ISL_14769247 on nsp12 Wuhan reference protein (left) / genomic (right). The alignment is partial due to space reason and was shorten from positions 81 to 281. Shown protein numbering 1 to 80 and 281 to 380. Shown nucleotide numbering 1 to 240 and 841 to 1140. On the left we have the protein side (20 residues per line), on the right the nucleotide side (60 residues per line). Any mutations are marked by '*'. The

mouse cursor is dynamic on both alignments so the user can highlights the couple amino acids/codon pair.

In the alignments of the RdRp domain and the spike protein respectively in Figure 4 & supplementary Figure1, the numbering is respected and each mutation is easily identified as silent or otherwise. In the RdRp example the frameshift is discretely managed and one non-synonymous mutation (P323L) is found in the RdRp domain. In the spike example several non-synonymous mutations were found : T19I, L24-, P25-, P26-, A27S, H69-, V70-, G142D, V213G, G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452R, S477N, T478K, E484A, F486V, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K. All of which indicates that this spike is deviant from the original strain.

A quick request to the Pangolin web-service confirms the lineage: BF10, corresponding to 22B-like (Omicron variant). Indeed, accordingly to nextstrain.org the clade 22B includes the lineage BF10, as presented in Figure 5.

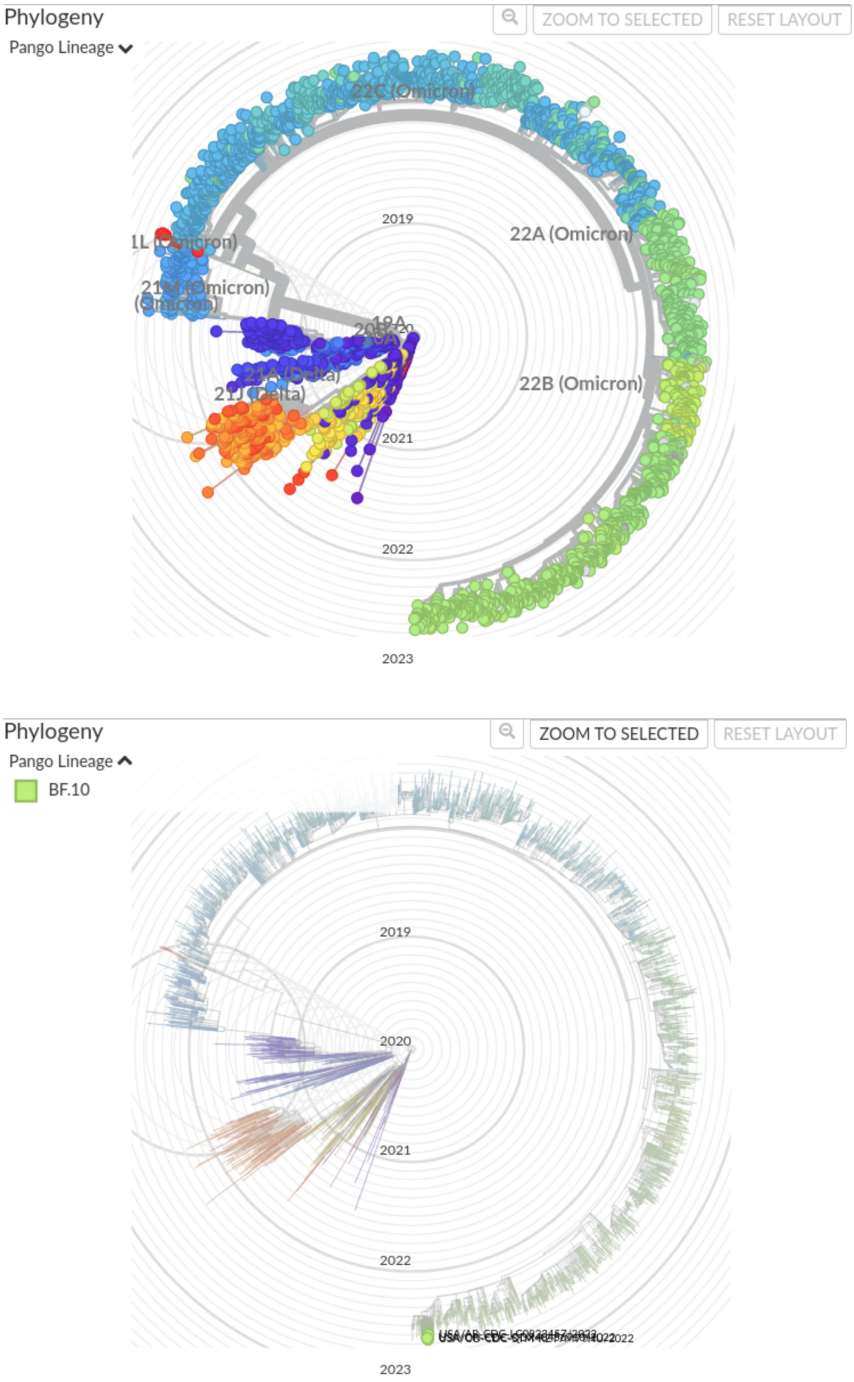

Figure 5. Position of lineage BF10 on the most recent SARS-Cov2 phylogeny (at the date of the writing of the paper): "Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months" (Dataset: ncov, open, global, 6m; Data range: from 2019-12-17 to 2022-09-04; data form GenBank) done on nextstrain.org the 2022-09-09 (https://nextstrain.org/ncov/open/global/6m?f_pango_lineage=BF.10). (top) all major clades; (bottom) focus on BF10 lineage.

4. DISCUSSION

In its final stage, VIMVer, designed as a transversal tool, should target audience from research studies to clinical applications. It could be implemented as an automatized survey workflow synergized with wet-lab applications and resources, such as the database on RNA viruses described in the VAZyMolO DB (9) used in structural and functional studies (11, 12, 12–18). Further developments concerning the online VIMVER is to extend the reference database, to the annotated sequences of VAZyMolO. The latter is a highly curated database of soluble functional or structural protein domains. The combination of both would then allow not only to spot an emerging mutation but to correlated it with structural information and compared it within a defined domain family.

The connection of VIMVer to protein structure analysis software, such as AlphaFold (19) or other structural prediction and analysis tools, should facilitate the connection of genotype to phenotype for emerging virus gene products. The main spirit of VIMVer is to propose a deep analysis of each mutation found in the user's genomic sequence of interest against the corresponding completely annotated reference. Considering its specific conception, VIMVer diverges from existing web-services like the remotely close web-service like NextClade (2) that does not provide specific annotated protein related dataset.

5. Conclusions

Viral Instant Mutation Viewer (VIMVer) is a sequence analysis tool of potential use to the community deciphering genomic sequences and connecting them to phenotype. Our workflow allows instant visualization of SARS-CoV-2 mutations on both genomic and protein sequences with proper numbering. It delivers an interactive alignment, and allows rapid lineage identification that could be easily used in genomic and epidemiological reports. VIMVer can be easily implemented in larger infrastructure, or adapted to all RNA viruses and is interoperable with already existing workflows as it is able to deliver results in a few seconds, even using a modest computational infrastructure. At the time of writing, the pandemic is now over but still drives us to better prepare ourselves to the next episode. Thus VIMVer contributes to existing services to better serve the scientific community and allow to analyze quickly mutational changes detected in the enormous wealth of data generated daily.

AVAILABILITY

VIMVer is available through a web service (https://vimver.afmb.univ-mrs.fr/).

The code and workflow documentation is available on https://github.com/wildevince/VIMVer. accessed on 19 April 2022 .

The project is under a free software license CeCILL (see more: http://www.cecill.info).

FUNDING

This research was funded by Fondation de la Recherche Médicale (Aide aux équipes), and the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101005077 (IMI-CARE).

ACKNOWLEDGMENTS

The authors acknowledge Dr. F. Debarre from “Ecole Normale Supérieur” IESS, Paris for usefull discussion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Sohrabi,C., Alsafi,Z., O’Neill,N., Khan,M., Kerwan,A., Al-Jabir,A., Iosifidis,C. and Agha,R. (2020) World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int J Surg*, **76**, 71–76.

2. Aksamentov,I., Roemer,C., Hodcroft,E.B. and Neher,R.A. (2021) Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, **6**, 3773.

3. BLAST: Basic Local Alignment Search Tool.

4. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.

5. Nuin,P.A., Wang,Z. and Tillier,E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.

6. O’Toole,Á., Pybus,O.G., Abram,M.E., Kelly,E.J. and Rambaut,A. (2022) Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics*, **23**, 121.

7. Harvey,W.T., Carabelli,A.M., Jackson,B., Gupta,R.K., Thomson,E.C., Harrison,E.M., Ludden,C., Reeve,R., Rambaut,A., Peacock,S.J., *et al.* (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*, **19**, 409–424.

8. Bhatt,P.R., Scaiola,A., Loughran,G., Leibundgut,M., Kratzel,A., Meurs,R., Dreos,R., O’Connor,K.M., McMillan,A., Bode,J.W., *et al.* (2021) Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science*, **372**, 1306–1313.

9. Plant,E.P. and Dinman,J.D. (2008) The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front Biosci*, **13**, 4873–4881.

10. Ferron,F., Rancurel,C., Longhi,S., Cambillau,C., Henrissat,B. and Canard,B. (2005) VaZyMolO: a tool to define and classify modularity in viral proteins. *J Gen Virol*, **86**, 743–749.

11. Uversky,V.N. and Dunker,A.K. (2010) Understanding Protein Non-Folding. *Biochim Biophys Acta*, **1804**, 1231–1264.

12. Gorbalenya,A.E., Lieutaud,P., Harris,M.R., Coutard,B., Canard,B., Kleywegt,G.J., Kravchenko,A.A., Samborskiy,D.V., Sidorov,I.A., Leontovich,A.M., *et al.* (2010) Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antiviral Res*, **87**, 95–110.

13. Imbert,I., Snijder,E.J., Dimitrova,M., Guillemot,J.-C., Lécine,P. and Canard,B. (2008) The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. *Virus Res*, **133**, 136–148.

14. Coutard,B., Gorbalenya,A.E., Snijder,E.J., Leontovich,A.M., Poupon,A., De Lamballerie,X., Charrel,R., Gould,E.A., Gunther,S., Norder,H., *et al.* (2008) The VIZIER project: Preparedness against pathogenic RNA viruses. *Antiviral Res*, **78**, 37–46.

15. John,S.P., Wang,T., Steffen,S., Longhi,S., Schmaljohn,C.S. and Jonsson,C.B. (2007) Ebola Virus VP30 Is an RNA Binding Protein. *J Virol*, **81**, 8967–8976.

16. Radivojac,P., Iakoucheva,L.M., Oldfield,C.J., Obradovic,Z., Uversky,V.N. and Dunker,A.K. (2007) Intrinsic Disorder and Functional Proteomics. *Biophys J*, **92**, 1439–1456.

17. Habchi,J., Mamelli,L., Darbon,H. and Longhi,S. (2010) Structural Disorder within Henipavirus Nucleoprotein and Phosphoprotein: From Predictions to Experimental Assessment. *PLoS One*, **5**, e11684.

18. Ogola,E.O., Kopp,A., Bastos,A.D.S., Slothouwer,I., Marklewitz,M., Omoga,D., Rotich,G., Getugi,C., Sang,R., Torto,B., *et al.* (2022) Jingmen Tick Virus in Ticks from Kenya. *Viruses*, **14**, 1041.

19. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.