BRIEF REPORT

REVISED Interactive SARS-CoV-2 mutation timemaps [version 2; peer review: 3 approved]

René L. Warren, Inanc Birol

Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, V5Z 4S6, Canada

v2 **First published:** 03 Feb 2021, **10**:68
https://doi.org/10.12688/f1000research.50857.1
Latest published: 03 Jun 2021, **10**:68
https://doi.org/10.12688/f1000research.50857.2

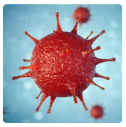
Abstract

As the year 2020 came to a close, several new strains have been reported of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the agent responsible for the coronavirus disease 2019 (COVID-19) pandemic that has afflicted us all this past year. However, it is difficult to comprehend the scale, in sequence space, geographical location and time, at which SARS-CoV-2 mutates and evolves in its human hosts. To get an appreciation for the rapid evolution of the coronavirus, we built interactive scalable vector graphics maps that show daily nucleotide variations in genomes from the six most populated continents compared to that of the initial, ground-zero SARS-CoV-2 isolate sequenced at the beginning of the pandemic.

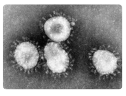
Availability: The tool used to perform the reported mutation analysis results, ntEdit, is available from GitHub. Genome mutation reports are available for download from BCGSC. Mutation time maps are available from https://bcgsc.github.io/SARS2/.

Keywords

SARS-CoV-2, COVID-19, Mutation time maps, GISAID, Interactive SVG



This article is included in the Disease Outbreaks gateway.



This article is included in the Coronavirus collection.

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 03 Jun 2021	✓ report	✓ report	✓ report
version 1 03 Feb 2021	? report	? report	? report

1. **Ingo Ebersberger**, Goethe-University Frankfurt, Frankfurt, Germany
Ruben Iruegas, Goethe-University Frankfurt, Frankfurt, Germany
2. **Takahiko Koyama**, IBM, Scarsdale, USA
3. **Jale Moradi**, Kermanshah University of Medical Sciences, Kermanshah, Iran

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: René L. Warren (rwarren@bcgsc.ca)

Author roles: Warren RL: Conceptualization, Formal Analysis, Resources, Software, Visualization, Writing – Original Draft Preparation; **Birol I**: Funding Acquisition, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Genome BC and Genome Canada [281ANV]; and the National Institutes of Health [2R01HG007182-04A1]. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Warren RL and Birol I. **Interactive SARS-CoV-2 mutation timemaps [version 2; peer review: 3 approved]** F1000Research 2021, **10**:68 https://doi.org/10.12688/f1000research.50857.2

First published: 03 Feb 2021, **10**:68 https://doi.org/10.12688/f1000research.50857.1

Introduction

In the last few weeks of 2020, new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) mutations in the United Kingdom (UK) were reported.[1] Although coronavirus genome mutations have been previously discovered and announced throughout the year, including the widely discussed D614G missense change in the spike protein,[2,3] the latest recurring surface protein mutations to be identified (e.g. N501Y, P681H) are cause for concern. The SARS-CoV-2 viral *S* gene encodes a surface glycoprotein, which upon interaction with host ACE-2 receptors, makes it possible for the coronavirus to gain entry to host cells and propagate. The reported changes to its sequence may be associated with increased virulence,[4] infectivity[3] and overall fitness.[5] The global response to those recent reports has been swift, with several countries shutting down air travel from the UK. This highlights the severity of the situation and the importance to track genomic variations and their predicted effects over time and space.

The rapid evolution of the SARS-CoV-2 genome in human hosts has prompted us to map all nucleotide changes that have appeared in 2020, since the first genome sequence of a COVID-19 patient isolate from the outbreak epicentre in Wuhan, China was made public.[6] For this, we leveraged the collaborative efforts of hundreds of institutions worldwide who, as of January 23rd 2021, have graciously shared over 260,000 SARS-CoV-2 genome sequences with the GISAID central repository since early January 2020.[7] Our mutation timemaps show the staggering number of nucleotide variants that have accumulated on the whole viral genome throughout the year, and especially since fall 2020, and in the six most populated continents. Here we present key features of these maps and how they may be of utility to researchers.

Methods

We first downloaded all complete, high-coverage SARS- CoV-2 genomes from GISAID[7] on January 23rd 2021 (human hosts samples collected). We then ran a genome polishing pipeline, which consists of ntHits[8] (v0.1.0 -b 36 -outbloom -c 1 -p seq -k 25) followed by ntEdit[9] (v1.3.4 -i 5 -d 5 -m 1 -r seq_k25.bf) and required at most 0.5 GB RAM and executed in ~1 sec. per genome on a single CPU. We used the first published SARS-CoV-2 genome isolate[6] (WH- Human 1 coronavirus, GenBank accession: MN908947.3) as the reference and each individual GISAID genome in turn as source of kmers to identify base variation relative to the former. The variant call format (VCF) output files from ntEdit were parsed and we tallied, for each submitted GISAID genome, the complete list of nucleotide variations. We next organized each nucleotide variant by sample collection date, continent of origin and, when applicable, evaluated its effect on the gene product that harbours the change to output an interactive scalable vector graphics (SVG) file. The script we developed to generate the maps is written in PERL and distributed under GPLv3. Users wishing to generate custom maps can download the script from Zenodo.[10] The full breadth of (unfiltered) SARS-CoV-2 nucleotide variations identified by this pipeline are updated on a weekly basis and are available for public download from https://www.bcgsc.ca/downloads/btl/SARS-CoV-2/mutations/.

Results and discussion

We analyzed nucleotide variations over time in over 260,000 SARS-CoV-2 viral genomes, submitted to the GISAID initiative[7] from around the globe, relative to that of the ground zero COVID-19 clinical isolate.[6] We mapped each mutation that was observed in five or more genomes each day. The 2020 calendar year from January 1st 2020 (day 1) to December 31st 2020 (day 366) is organized in a circle where each radius represents a day (1 day = 0.98 degree) and data points represent mutations along the reference genome sequence from 1 (closest to center) to 29,903 bp (near the outer rim). The size of each point is in log10 scale of the number of contributing viral genomes collected on that day that has the mutation, with colour assignments indicating the continent of origin where the mutation is observed. A mouse over each data point reveals the collection date, the nucleotide variant, the continent and associated number of contributing genome sequences (including daily sample fraction) and, when applicable, the gene product and predicted amino acid change.

From the SARS-CoV-2 genome mutation timemap (Figure 1A), we observe the first persistent mutations (≥5 genomes/day) appearing in late February 2020, including the prevalent D614G mutation in Europe on February 22nd (albeit since late January in fewer samples, Figure 1B). From there, the original coronavirus genome sustained many changes overtime (5,468 distinct variants mapped in 2020 as of January 23rd, 2021), including a sizeable proportion (56.8 %) of missense mutations. It is immediately evident from Figure 1A that variations from Europe account for a larger share (71.2%) of the variants mapped. Further, there appears to be a surge in variations identified in late summer/throughout fall 2020 in this continent. This may be explained by a disproportionate number of submissions with samples originating from this geographic location as the second wave hit hard. Thus, caution in interpreting the map is warranted. Of note, the spike protein gene variant N501Y, observed on our maps in Wales UK in late September 2020 (n = 2, 1.5% of Wales samples) (Figure 1B), is consistent with an earlier study reporting on its recurrent emergence within this time frame.[1] From the map, we clearly observe its emergence as it increases in frequency by late December 2020 (n = 13, 15.5% of Wales samples) and spreads to different regions. We also note the emergence of several additional mutations in the spike protein gene,

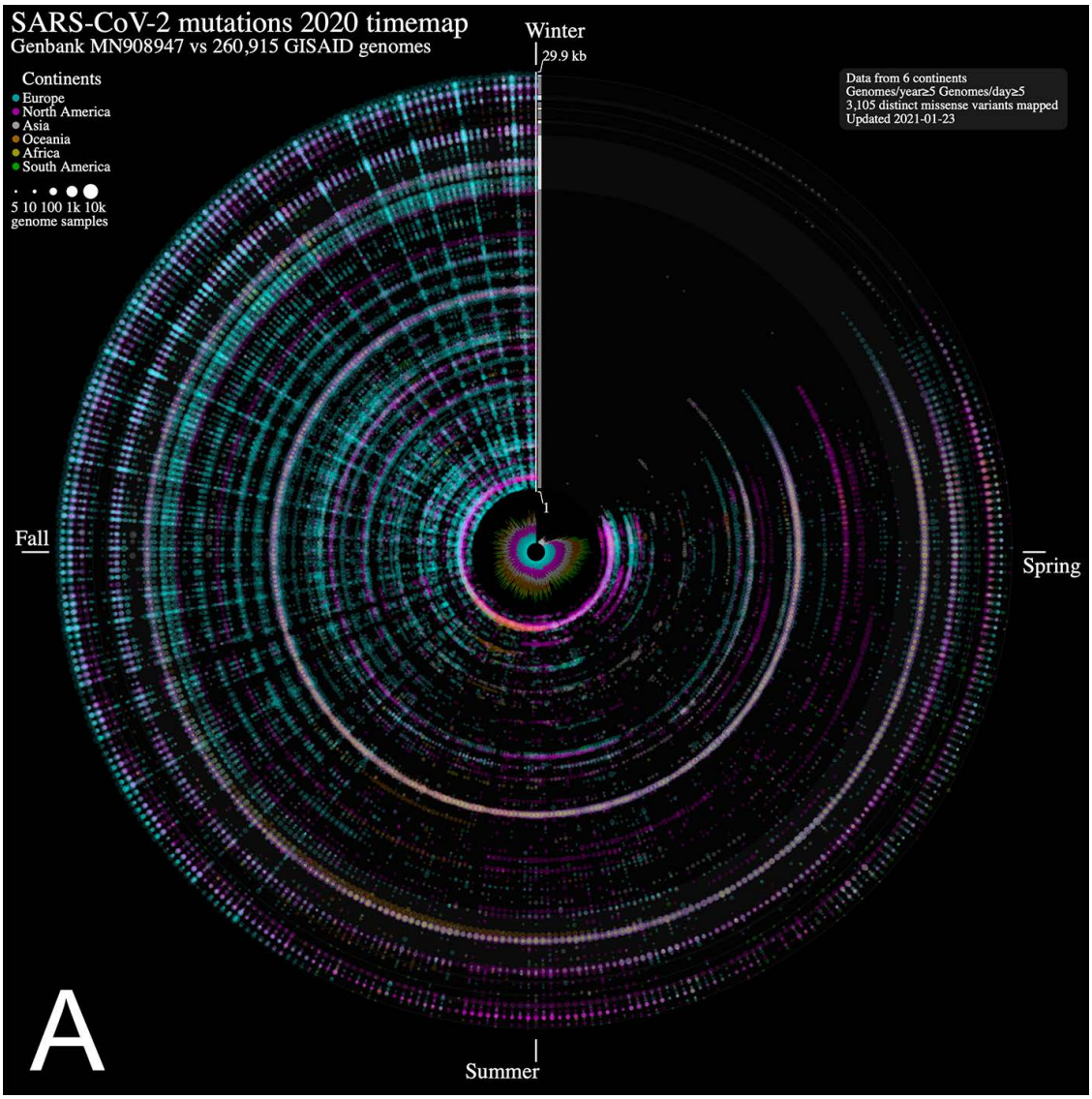


Figure 1. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) evolution in human hosts. ntEdit was used to map nucleotide variations between the first published coronavirus isolate from Wuhan, China in early January and over 260,000 SARS-CoV-2 genomes sampled from around the globe during the 2020 coronavirus disease 2019 (COVID-19) pandemic. The maps show missense mutations arising daily (A) in the world within the whole viral genome, with the reference genome represented by the vertical axis from bases 1 to 29.9 kbp and (B) in Europe within the spike protein gene. Alternating dark/light grey vertical rectangles and associated tracks depict, starting from the center, SARS-CoV-2 genes *ORF1AB*, *S, ORF3A, E, M, ORF6, ORF7A, ORF8, N,* and *ORF10.* Mutations identified daily are represented by circles in a given radius and are coloured by regions and sized relative to raw count (panel A) or ratio (panel B) of the daily samples. A stacked bar plot (center) shows sample count. The 2020 calendar year mutations are organized clockwise from the upper vertical. Hovering the mouse cursor over each data point reveals additional insights about the mutation, and each map offers a navigation wheel allowing to pan inall direction and zoom in/out (panel B). Panel B shows an annotated timemap of Europe, highlighting the detection of the first D614G spike protein gene mutation on January 28th 2020 (Germany, n = 3, upper right). We also highlight the N501Y spike mutation first observed on September 20th 2020 (panel B, inset) in only 1.7% (n = 2) of the Wales, UK daily genome samples, and at the end of the year on December 28th 2020 in 15.5% (n = 13) of the daily collected Wales, UK samples (data updated on May 24th, 2021). White arrows near the genome axis are used to draw attention to the emergence of spike protein gene mutations (from top to bottom) D1118H, S982A, T716I, P681H, A570D and N501Y.

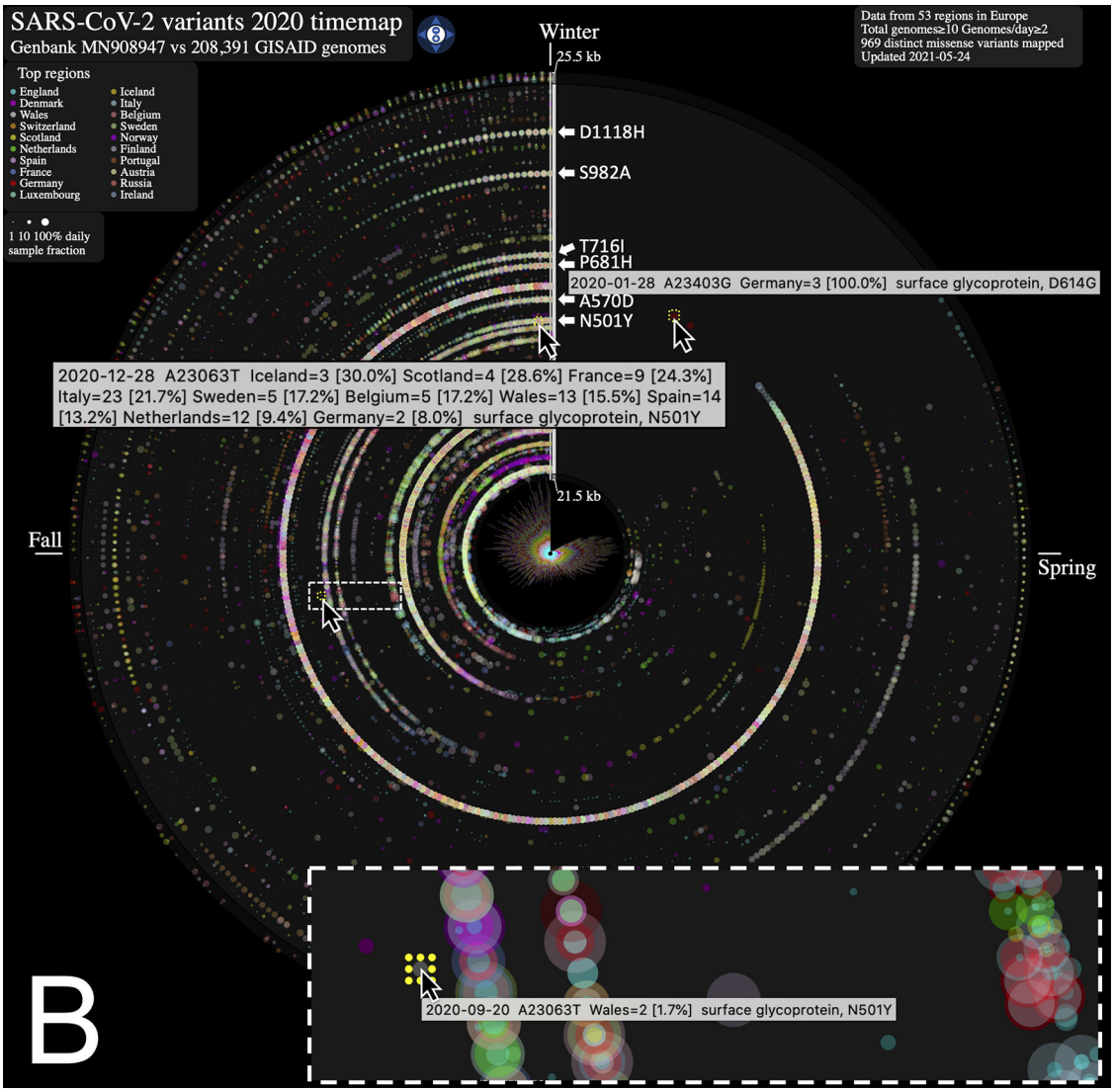


Figure 1. *(continued)*

including D1118H, S982A, T716I, P681H andA570D, all visible in late 2020 as they rose to prominence in the GISAID genome catalogue (Figure 1B).

Fuelled by the raging COVID-19 pandemic, GISAID's data is enabling more than a dozen SARS-CoV-2 variant web-based visualizations including those hosted by NextStrain,[11] CovMT[12] and CoVariants.[13] Those portals offer feature-rich and intuitive interfaces to navigate a comprehensive collection of graphs of SARS-CoV-2 variant lineages and compositions in key geographic locations. In some cases, the analysis results presented at these online portals is based on limited genome and nucleotide variation data subsets and the raw mutation call prediction for each sample is not readily available for download. With our project, we make all nucleotide variation calls public for each GISAID genome in the collection that is complete, high coverage, and with a complete associated sample collection date. We also provide tabulated data analysis results that are mutation-centric (https://www.bcgsc.ca/downloads/btl/SARS-CoV-2/mutations/), which is useful to evaluate mutation frequency overtime – data we have used in other SARS-CoV-2 related work to monitor emergence (not shown). With our timemaps, we offer an alternative visual display to what we have been accustomed to seeing this past year and a perspective that is not already covered by the aforementioned tools. We accomplished that by generating a comprehensive bird's eye view of all mutations that have accumulated in each GISAID genome since the beginning of the pandemic, to show the sheer scale of viral genome transformation that has happened – and still occurring – in human hosts. Of course, some of these displays have become dense as institutes worldwide submit new data to GISAID (7-fold more data in the catalogue since initial manuscript submission) and more nucleotide variations are detected overtime, but the maps still serve a

purpose in illustrating the staggering accumulation of variations in time, from around the globe, and to identify mutation hotspots. Since our initial release of the maps, we have generated additional timemaps for all SARS-CoV-2 genes and some of the emerging variants of concerns (e.g. lineages B.1.1.7, B.1.351, B.1.617, P.1) that have come to dominate the landscape in certain jurisdictions, due to the advantages conferred by their associated mutation signatures (available from: https://bcgsc.github.io/SARS2). These alternate views are useful in more clearly identifying new nucleotide variations arising in time and in certain jurisdictions, within specific variants. Taken as a whole, our timemaps offer a fairly qualitative, but still all-encompassing and comprehensive, view ofSARS-CoV-2 genome evolution in human hosts, less than two years since the ground zero strain genome was first characterized. We note that, importantly, the maps also offer quantitative and actionable information, which can be accessed by interactive navigation. Interactive visualization features such as mouse hover reveal the variant effect/product, sample frequency and origin for a given mutation. The SVG platform used offers pan, zoom, tilt, highlight, click and drag functionality to inspect variants in detail, including the detection of possible emergence at a specific time and geographical location. Further, the software built to make our SVG timemaps is freely available to scientists interested in generating custom and flexible views of SARS-CoV-2 genes not yet offered by our interface. With our periodically updated SARS-CoV-2 timemaps totalling over 120 individual SVG displays, we offer unique longitudinal views of strain development in real time. Each timemap provides an extensive yearly panorama of SARS-CoV-2 nucleotide variations and the means to follow variant evolution in human hosts, over time and space.

Data availability

Source data

The SARS-CoV-2 genome sequences can be accessed via the GISAID central repository. Processed single nucleotide variant (SNV) data is available from https://www.bcgsc.ca/downloads/btl/SARS-CoV-2/mutations/.

Maps availability

- Maps are available from: https://bcgsc.github.io/SARS2
- SNV detection source code is available from: https://github.com/bcgsc/ntedit
- Archived source code at time of publication: https://doi.org/10.5281/zenodo.4469840.[10]

Author contributions

Study design: RLW. Analysis: RLW. Both authors wrote the manuscript.

Acknowledgements

We acknowledge Cecilia (Lingyu) Yang for her early work on SARS-CoV-2 variants.

References

1. Rambaut A, *et al.*: **Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations.** *Virological.* 2020.
Reference Source
2. Dey T, *et al.*: **Identification and computational analysis of mutations in SARS-CoV-2.** *Comput Biol Med.* 2021; **129**: 104166.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Korber B, *et al.*: **Tracking changes in SARS- CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus.** *Cell.* 2020; **182**: 812–827.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Gu H, *et al.*: **Adaptation of SARS-CoV-2 in BALB/c Mice for Testing Vaccine Efficacy.** *Science.* 2020; **369**: 1603–1607.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Plante JA, *et al.*: **Spike mutation D614G alters SARS-CoV-2 fitness.** *Nature.* 2020.
PubMed Abstract | Publisher Full Text
6. Wu F, *et al.*: **A new coronavirus associated with human respiratory disease in China.** *Nature.* 2020; **579**: 265–269.
PubMed Abstract | Publisher Full Text | Free Full Text
7. re3data.org: **GISAID; editing status.** *re3data.org - Registry of Research Data Repositories.* 2020-02-03.
Publisher Full Text
8. Mohamadi H, *et al.*: **ntHits: de novo repeat identification of genomics data using a streaming approach.** *BioRxiv.* 2020.
Publisher Full Text
9. Warren RL, *et al.*: **ntEdit: scalable genome sequence polishing.** *Bioinformatics.* 2019; **35**: 4430–4432.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Warren R, Birol I: **Interactive SARS-CoV-2 mutation timemaps (Version v1.1).** *Zenodo* 2021, January 26.
Publisher Full Text
11. Hadfield J, *et al.*: **Nextstrain: real-time tracking of pathogen evolution.** *Bioinformatics.* 2018; **34**: 4121–4123.
Publisher Full Text
12. Alam I, *et al.*: **CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants.** *Lancet Infect Dis.* 2021; **21**: 602.
Publisher Full Text
13. Hodcroft EB, *et al.*: CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 21 June 2021

https://doi.org/10.5256/f1000research.57062.r86694

✔ **Ingo Ebersberger**
Applied Bioinformatics Group, Institute for Cell Biology and Neuroscience, Goethe-University Frankfurt, Frankfurt, Germany
Ruben Iruegas
Applied Bioinformatics Group, Institute for Cell Biology and Neuroscience, Goethe-University Frankfurt, Frankfurt, Germany

I have no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: BioSequence-Informatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 14 June 2021

https://doi.org/10.5256/f1000research.57062.r86693

✔ **Jale Moradi**
Department of Microbiology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran

The new version of the manuscript is improved in all sections, more details are provided in the result part by some examples that show the utility of the maps. The results indicate that the script

using to generate maps is changed, since, the interactive function of the maps has improved and it is possible to zoom on each spot. I believe this interactive evolutionary vector graphic map could be generated and used by other researchers in different projects to monitor the evolution of the SARS-CoV-2 virus.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical Microbiology, Genomics, Microbial Evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 June 2021

https://doi.org/10.5256/f1000research.57062.r86695

Takahiko Koyama
TJ Watson Research Center, IBM, Scarsdale, NY, USA

The tool is now more interactive with zooming to look into more details. Although I personally do not understand utility of concentric representation, there is not much to add before my approval for the indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, bioinformatics, oncology, immunology, virology, and stem cell biology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 01 June 2021

https://doi.org/10.5256/f1000research.53946.r85512

Jale Moradi
Department of Microbiology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran

The authors have geographically shown the nucleotide variations for global SARS-CoV-2 sequences in a time map. The sequences have been downloaded, polished and analyzed with ntHit and ntEdit. The Wuhan-Hu-1-NC_045512/MN908947 was set as the reference sequence, then the variations output was mapped based on the sample collection time by a script written in PERL. The results have shown in two circle maps including “whole viral genome” and “spike protein gene” variations over time from January 1st 2020 as day 1 to December 31st 2020 as day 366. Each radius in these circles represents a day and each spot on this radius shows a variation. Also, the spots are shown in different colours that each colour is indicating a specific geographical region (continent or country).

It is a useful tool to overview the evolution of the virus since the beginning of the epidemic. Furthermore, it can be concluded which part of the genome has more variations, also, the colour appearance of the map helps us to understand approximately how many mutations there are in different regions or from which ones the mutations originated. If it were possible to identify the relevant mutation (exact mutation type) by clicking on each spot, it could help more. Also, different spots have overlaps in many parts, which would provide better information if it was possible to determine which spots this overlap includes.

Overall, the developed script provides a useful map for viewing the pattern of virus evolution globally, although it would be more informative if the authors could improve this script to solve the mentioned issues.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical Microbiology, genomics, immunology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Jun 2021

René Warren, Canada's Michael Smith Genome Sciences Centre, Canada

We thank our Reviewer for their support of our work and insights. We also value the suggestions, as it helps us improve upon the work and broaden the interest.

We have just published a revised version of the manuscript (v2), which expands on the utility of the maps, situates them in context of other similar work, and introduces new map features to increase interactivity and overall experience.

Some of the maps' new features (since original submission):

Interactivity

1. Maps are draggable.

2. Zoom/pan.

3. Tilt 90 degrees to make axis horizontal (this and above features implemented in a navigation wheel).

4. Colour highlight on mutation tooltip.

5. Gene/variant views have additional colour highlight (by region) on certain maps*.

*The added functionality comes at a cost, making them sluggish when views are too dense, which is why this feature is currently only used to display individual genes/variant displays and not the whole genome

Improvements:

1. Over 120 individual displays, all SARS-CoV-2 genes are now presented.

2. Better discrimination of close high-frequency mutations allows more information to show through by adjusting the spot ratio (r=sqrt(freq*factor/pi) and no longer plots on a log10 in ratio mode.

3. When same %, adjust a secondary sort such that the colour matches the first region labelled.

4. Better grouping/sorting of overlapping points.

5. Added ability to explore switch year from the current view 2020<->2021 and between ratio(%) and raw (#) counts without having to go to main menu and use drop-down.

The mutation "spots" are also plotted incrementally (by coordinates) and by decreasing

order of frequency, allowing most mutations to interactively show (and not be obscured by overlaps). But overlaps are unavoidable with displays that are too dense, and some data points may still be out of reach, but other individual maps (eg. variant/gene levels) may provide a better visual of the most important mutations.

Improvements 2), 3) and 4) in particular are in response to our Reviewer's comment on spot overlap, and calculating the ratio in such a fashion (instead of log10) enables a better resolution on close-by high-frequency mutations (such as the D614G). Most displays will show missense mutation to minimize display density, but we also offer representations by types (missense vs silent) and all-encompassing. With tooltip, the mutation type is shown as either its effect in amino acid space (eg. N501Y) or silent when the nucleotide variation has no predicted effect.

Competing Interests: no competing interests to declare

Reviewer Report 17 May 2021

https://doi.org/10.5256/f1000research.53946.r85294

?

Takahiko Koyama
TJ Watson Research Center, IBM, Scarsdale, NY, USA

Authors have developed a web based visualization tool for longitudinal evolution of SARS-CoV-2 genomes.

Although they have made unique representation of longitudinal strain developments, it is not clear the utility of the tool. For instance, while concentric circle representation of daily genomes is visually appealing, it limits the duration to a year and inner part inevitably becomes crowded compared with outer area.

Lack of interactivity is also an issue. There must have been a way to magnify the area.

Furthermore, in mutation prone loci, the dots are overlapped and not easy to see what is going on. For these reasons, utility of the tool is limited; more improvements need to be done before it gains large user base.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, bioinformatics, oncology, immunology, virology, and stem cell biology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2021

René Warren, Canada's Michael Smith Genome Sciences Centre, Canada

Authors have developed a web based visualization tool for longitudinal evolution of SARS-CoV-2 genomes.

Although they have made unique representation of longitudinal strain developments, it is not clear the utility of the tool. For instance, while concentric circle representation of daily genomes is visually appealing, it limits the duration to a year and inner part inevitably becomes crowded compared with outer area.

We thank our Reviewer for the valuable insights provided and spending the time to review our work. We acknowledge limitations of the display, and we stress that our original work on this was done in December, on 200,000 GISAID genomes and one year's worth of data. Our preprint became public January 2021 and we subsequently submitted this work to F1000Research, summarizing the 2020 pandemic-associated SARS-CoV-2 variants for year 2020. A circular representation is an aesthetic choice, allowing to get a bird's eye view of the breadth of mutations.

Lack of interactivity is also an issue. There must have been a way to magnify the area.

This is a great suggestion. We have now added the ability to pan and zoom on each map, making the maps more interactive.

Furthermore, in mutation prone loci, the dots are overlapped and not easy to see what is going on. For these reasons, utility of the tool is limited; more improvements need to be done before it gains large user base.

The maps were first built to visually quantify the appreciable variability that exists in rapidly evolving SARS-CoV-2 genomes. Since, we have added spike-specific views, and variants of concerns (VOCs) to the list of maps available to the community. We also provide the tools to generate the maps, such that advanced users may customize and generate additional views of interest, as needed

Competing Interests: no competing interests

Author Response 03 Jun 2021

René Warren, Canada's Michael Smith Genome Sciences Centre, Canada

We wanted to add to our previous response to our Reviewer. Once again, we are grateful for your suggestions to improve upon interactivity of the maps. Since your Review, we have worked to improve the user experience and we list below some of the new features:

Interactivity

1. Maps are draggable.

2. Zoom/pan.

3. Tilt 90 degrees to make axis horizontal (this and above features implemented in a navigation wheel).

4. Colour highlight on mutation tooltip.

5. Gene/variant views have additional colour highlight (by region) on certain maps*.

*The added functionality comes at a cost, making them sluggish when views are too dense, which is why this feature is currently only used to display individual genes/variant displays and not the whole genome

Overall improvements

1. Over 120 individual displays, all SARS-CoV-2 genes are now presented.

2. Better discrimination of close high-frequency mutations allows more information to show through by adjusting the spot ratio (r=sqrt(freq*factor/pi) and no longer plots on a log10 in ratio mode.

3. When same %, adjust a secondary sort such that the colour matches the first region labelled.

4. Better grouping/sorting of overlapping points.

5. Added ability to explore switch year from the current view 2020<->2021 and between ratio(%) and raw (#) counts without having to go to main menu and use drop-down.

Thanks again for spending the time to review our work.

Competing Interests: no competing interests to declare

Reviewer Report 14 May 2021

https://doi.org/10.5256/f1000research.53946.r83795

?

Ingo Ebersberger
Applied Bioinformatics Group, Institute for Cell Biology and Neuroscience, Goethe-University Frankfurt, Frankfurt, Germany
Ruben Iruegas
Applied Bioinformatics Group, Institute for Cell Biology and Neuroscience, Goethe-University Frankfurt, Frankfurt, Germany

The authors present interactive mutation time maps for SARS-CoV-2, which provide a highly resolving view of when, where and how frequent a particular mutation was detected in the sampled SARS-CoV-2 genome sequences provided via GISAID. The manuscript itself is rather short. It is briefly describing the methodological approach of how the mutations have been detected and mapped to the reference genome. The combined Results and Discussion section is equally concise and comprises a description of what is seen in the interactive maps together with few example observations that can be made with these maps. The Discussion section ends with the expression of the hope that the maps presented here “will help researchers in their exploration of SARS-CoV-2 mutations and their predicted effect over time.”

Overall, the topic that is touched in this manuscript is highly relevant, as variations of SARS-CoV-2 is something that currently is and will be of major concern in the future. Here, the graphs present a very nice access to the information that is represented by the ever-increasing amount of viral genome sequences world-wide. The data presentation is appealing, and it allows to overview the general trends of SARS-CoV-2 evolution. However, we see considerable room for (essential) improvement.

Major issues:
The authors end the manuscript with the belief that the interactive maps will be of help for the research community working on SARS-CoV-2 variation. We miss two things here:

First, it would be great if the authors show how the data provided by the maps can be used to indeed come up with new conclusions, in particular with respect to the ‘predicted effect over time’.

For us, it is entirely unclear how such an analysis should be performed. Exploring the data, this is something that one nicely can do while looking at the plots, some clear signals, e.g. the fate of D614G, can also be extracted. But how to work with the data beyond this simple and straightforward 'looking' at the plots? Please, don't get us wrong here, we consider looking at data a very important aspect of data analysis. Still, the sheer amount of information, which results in very dense plots with many overlapping data points, makes it, in our opinion, very hard to identify emerging variants that should be monitored right from the start. Just to give you an example: D614G is represented by a very prominent circle in the plots. What would be the authors approach to identify and monitor a novel variant, say at position 615 of the reference strain? By looking at the plots, we consider this almost impossible, since the signal will be entirely covered by the prominent mutation at position 614.

The analysis is presented using the "ground-zero" strain as a reference. But is this still timely? Numerous variants have now frequencies that go far beyond that of the original nucleotide at a certain position, again, for example the D614G variant. This would allow to 'purge' the signal of very successful variants, helping to direct the focus on emerging variants.

When it comes to the website itself, we see some room for improvement:

- First and foremost, we think the plots are overcrowded with information. Although it is nice to see a global overview of the data across the entire genome, 365 days, and 6 continents, it is impossible (at least for me) to explore this information other than randomly clicking individual data points, as we have outlined above. we think, this approach would benefit from providing the information in more digestible data fractions. Thus far, the user can choose to focus on the spike, but not on the other proteins. It would be helpful, just as a suggestion, to focus also on variants with a certain prevalence. But we are sure that the authors will have way better ideas than our proposals here, once they specify how a user should work with the plots and the data. Looking at https://nextstrain.org, which also provides a very nice overview of SARS-CoV-2 variation, may give some hints.

- It would be very convenient, if the interactive plots would be designed such that the user can toggle the information for display, instead of having to go back to the main menu and select a different display mode.

- Trend lines that show the prevalence of a certain variant in a certain region over time would help a lot and should be easy to implement.

- The orientation of where in a genome a certain variant exists is very hard. Although the vertical bars at 12 h in the circular plot should indicate in what ORF a variant is located, this is really hard to track across the full plot. In particular, because the bar-ORF assignment is not visible.

- Animation of daily variant emergence is again a nice feature. However, it is a gif and not interactive. The time lapse does not allow the user to pause, fast forward, or skip to a particular time. Moreover, x-axis labels overlap in particular for the spike. This makes the plot nice to look at, but the information that can be retrieved is only limited.

- Graph of weekly spike protein variant emergence is not interactive and difficult to read, as the lines overlap with each other and some have similar colors. Some functionalities could

be implemented such as being able to toggle strains from the right menu, selecting a time range and continent/country, and being able to hover over to display the information. 2020 and 2021 plots have layout inconsistencies and could be merged into a single graph.

- The variant emergence graph heavily competes with the information in https://nextstrain.org, which claims to be updated daily.

In the outermost ring, we detected a variant that is assigned to na/na. What is this supposed to mean?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: BioSequence-Informatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 14 May 2021

René Warren, Canada's Michael Smith Genome Sciences Centre, Canada

The authors present interactive mutation time maps for SARS-CoV-2, which provide a highly resolving view of when, where and how frequent a particular mutation was detected in the sampled SARS-CoV-2 genome sequences provided via GISAID. The manuscript itself is rather short. It is briefly describing the methodological approach of how the mutations have been detected and mapped to the reference genome. The combined Results and Discussion section is equally concise and comprises a description of what is seen in the interactive maps together with few example observations that can be made with these maps. The Discussion section ends with the expression of the hope that the maps presented here "will help researchers in their

exploration of SARS-CoV-2 mutations and their predicted effect over time."

Overall, the topic that is touched in this manuscript is highly relevant, as variations of SARS-CoV-2 is something that currently is and will be of major concern in the future. Here, the graphs present a very nice access to the information that is represented by the ever-increasing amount of viral genome sequences world-wide. The data presentation is appealing, and it allows to overview the general trends of SARS-CoV-2 evolution. However, we see considerable room for (essential) improvement.

We thank our Reviewers for their comments, suggestions and diligence with their extensive report. Our response can be found below, in bold face

Major issues:
The authors end the manuscript with the belief that the interactive maps will be of help for the research community working on SARS-CoV-2 variation. We miss two things here:

First, it would be great if the authors show how the data provided by the maps can be used to indeed come up with new conclusions, in particular with respect to the 'predicted effect over time'. For us, it is entirely unclear how such an analysis should be performed. Exploring the data, this is something that one nicely can do while looking at the plots, some clear signals, e.g. the fate of D614G, can also be extracted. But how to work with the data beyond this simple and straightforward 'looking' at the plots? Please, don't get us wrong here, we consider looking at data a very important aspect of data analysis. Still, the sheer amount of information, which results in very dense plots with many overlapping data points, makes it, in our opinion, very hard to identify emerging variants that should be monitored right from the start. Just to give you an example: D614G is represented by a very prominent circle in the plots. What would be the authors approach to identify and monitor a novel variant, say at position 615 of the reference strain? By looking at the plots, we consider this almost impossible, since the signal will be entirely covered by the prominent mutation at position 614.

We greatly really appreciate community feedback on the potential usefulness of this work, and not only the maps, but additional analysis we were able to provide after we submitted the paper (our Reviewers made mentioned of them below), using the wealth of information we were able to mine from the GISAID genomes (these secondary analysis results, which consists of nucleotide variants and their effect, are tallied each week from each individual SARS-CoV-2 genome). We originally built the maps to be fairly qualitative, to simply gain a [visual] appreciation for the rapid coronavirus evolution on a year scale/factoring sample regions of origin, and this is what we presented in the manuscript. In our conclusion we give an example of a mutation that is observable from the GISAID genomes, on our maps, at the time reported in published papers; Since submission, the GISAID catalogue has more than doubled in size and maps quickly became dense, as our Reviewer indicated. To help remedy the problem and make the maps more useful, we have since started to provide additional genome and spike views of variants of concerns (VOCs) and have added visualizations for 2021 (a more digestible data fraction, indicated below by our Reviewer). Another type of information that can be extracted from the maps is the speed at which mutations in VOCs have appeared and spreading in additional

jurisdictions, which can be readily observed without too much effort. Our Reviewers are correct that variations in close proximity are difficult to see, which is why we provide views for the spike-encoding gene. Still, it would be difficult to differentiate between positions 614 and 615, which is why we provide the SVG-generating script such that interested parties would be able to generate custom views should they chose to (Ideally a more flexible website could help, see response below).

The analysis is presented using the "ground-zero" strain as a reference. But is this still timely? Numerous variants have now frequencies that go far beyond that of the original nucleotide at a certain position, again, for example the D614G variant. This would allow to 'purge' the signal of very successful variants, helping to direct the focus on emerging variants.

Our Reviewer is correct that the comparison is relative. When we started this project in December 2020, it made sense to use the "ground zero" strain genome. We could make the case for selecting another set of references to compare against, but it may lead to disagreements in scientific circles, on the base genome sequence to use. Additional maps may be produced in the future to see evolution within each VOCs, which may be an acceptable proposition.

When it comes to the website itself, we see some room for improvement:

- *First and foremost, we think the plots are overcrowded with information. Although it is nice to see a global overview of the data across the entire genome, 365 days, and 6 continents, it is impossible (at least for me) to explore this information other than randomly clicking individual data points, as we have outlined above. we think, this approach would benefit from providing the information in more digestible data fractions. Thus far, the user can choose to focus on the spike, but not on the other proteins. It would be helpful, just as a suggestion, to focus also on variants with a certain prevalence. But we are sure that the authors will have way better ideas than our proposals here, once they specify how a user should work with the plots and the data. Looking at https://nextstrain.org, which also provides a very nice overview of SARS-CoV-2 variation, may give some hints.*

- *It would be very convenient, if the interactive plots would be designed such that the user can toggle the information for display, instead of having to go back to the main menu and select a different display mode.*

- *Trend lines that show the prevalence of a certain variant in a certain region over time would help a lot and should be easy to implement.*

- *The orientation of where in a genome a certain variant exists is very hard. Although the vertical bars at 12 h in the circular plot should indicate in what ORF a variant is located, this is really hard to track across the full plot. In particular, because the bar-ORF assignment is not visible.*

- *Animation of daily variant emergence is again a nice feature. However, it is a gif and not interactive. The time lapse does not allow the user to pause, fast forward, or skip to a particular time. Moreover, x-axis labels overlap in particular for the spike. This makes the plot nice to look at, but the information that can be retrieved is only limited.*

- *Graph of weekly spike protein variant emergence is not interactive and difficult to read, as the lines overlap with each other and some have similar colors. Some functionalities could be implemented such as being able to toggle strains from the right menu, selecting a time range and continent/country, and being able to hover over to display the information. 2020 and 2021 plots have layout inconsistencies and could be merged into a single graph.*

- *The variant emergence graph heavily competes with the information in https://nextstrain.org, which claims to be updated daily.*

We thank our Reviewers for spending the time to navigate the website, which originally, wasn't part of the project (built as a means to share the maps). We agree that a more modern and flexible web design would help with the customization and eventual uptake of these maps. Some of the plots were added to the website for convenience, to show users what is possible to do with the extensive mutation data we are compiling for this project (and available for download here)

In the outermost ring, we detected a variant that is assigned to na/na. What is this supposed to mean?

These variants fall in UTR regions. Thank you for the feedback, in our next release of the maps, we will replace NA by UTR to indicate that this nucleotide variant compared to the reference is found outside coding regions. The last position indicates possible effect in the protein space, which is not applicable in this case.

Competing Interests: No competing interests to declare