Factors affecting SARS-CoV-2 sequence conservation: SARSNTdb database

By

JOHN ORGERA

A thesis submitted to the

Graduate School-Camden

Rutgers The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Master of Science

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Andrey Grigoriev

And approved by

_____

Dr. Andrey Grigoriev

_____

Dr. Sunil Shende

_____

Dr. Marien Solesio

Camden, New Jersey

August 2022

ABSTRACT OF THE DISSERTATION

Factors affecting SARS-CoV-2 sequence conservation: SARSNTdb database

By John Orgera

Dissertation Director:

Dr. Andrey Grigoriev

SARSNTdb offers a curated, nucleotide-centric database for users of varying SARS-CoV-2 knowledge. Its user-friendly interface enables querying coding regions and coordinate intervals to find out the various functional and selective constraints that act upon the corresponding nucleotides and amino acids. Users can easily obtain information about viral genes and proteins, functional domains, repeats, secondary structure formation, intragenomic interactions, and mutation prevalence. Currently, many databases are focused on the phylogeny and amino acid substitutions, mainly in the spike protein. While providing mutation data, SARSNTdb takes a more nucleotide-focused approach as RNA does more than just code for proteins and many insights can be gleaned from its study. For example, RNA-targeted drug therapies for SARS-CoV-2 are currently being developed and it is essential to understand the features only visible at that level. This database enables the user to identify regions that are more prone to forming secondary structures that drugs can target. Finally, the database allows for comparing SARS-CoV-2 and SARS-CoV domains and sequences. SARSNTdb can serve the research community by being a curated repository for information that gives a jump start to analysing a mutation's effect far beyond just determining synonymous/non-synonymous substitutions in protein sequences.

INTRODUCTION

Background

Coronavirus disease 2019 (COVID-19) pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus was first identified in December 2019 in Wuhan China(1). The WHO declared the outbreak a pandemic in March 2020(2) and since then the scientific community has generated a massive amount of data and that needs to be organized in order to understand the virus. This data includes sequences, papers, medical information, and proteomics data. Over time the virus has continually evolved and has branched into several variants, some of them being Variants of Concern (VOC) such as the Omicron and Delta variants(3). These variants are identified primarily by sequencing, further stressing the importance of competent SARS-CoV-2 databases that allow scientists to interrogate mutations understand the functions of the various viral proteins that the variants may affect.

The virus is a positive sense RNA virus with 16 non-structural proteins (NSPs) and 10 structural and accessory proteins(4). The 16 non-structural proteins are cleaved from two open reading frames (ORF) known as ORF1a and ORF1b(5). ORF1b is a polyprotein that is cleaved into NSP1 to NSP11. ORF1b is cleaved into NSP12-NSP16(5). ORF1a and 1b are separated a by a -1 ribosomal frameshift upstream of the ORF1a stop codon. The accessory proteins are transcribed to a negative sense template by the RNA-Dependent RNA Polymerase (RdRP), a complex of nsp12, 7, and 8(5). This is done in the 3'-5' direction from the 3' end of the genome. When the RdRP reaches the 5' start of a protein it is transcribing it encounters a transcription regulatory sequence B (TRS-B). This initiates a jump to the TRS-Leader at coordinate 70 on the 5' end, skipping the nucleotides between the TRS-B and TRS-L

(5). This attaches the leader sequence to the 5'end of the template, allowing the host's ribosomes to recognize and translate the RNA.

The virus has several unique evolutionary constraints placed upon it. For example, its relatively short genetic material caused it to evolve several unique methods to create novel proteins. As previously stated, the frameshift separating ORF1a and ORF1b is one of them. This frameshift is caused by a unique 7nt "slippery sequence" at the RNA level(5). Another example of short RNA motifs regulating the genome is the TRS-B. This 6nt long sequence is responsible for the creation of all the accessory and structural proteins. Several proteins overlap at the RNA level such as ORF7a and ORF7b or the Nucleocapsid (N) protein and ORF9b(4). ORF7b proteins are thought to be created through leaky scanning of ORF7a(4). It is likely features at the RNA level that are responsible for these efficient methods of translation. Finally, RNA secondary structures are often formed by SARS-CoV-2(6). RNA secondary structures are often targeted by host defences like RIG-I and are placed under evolutionary constraint(7).

Defining a Niche

Several databases related to SARS-CoV-2 have studied and reported on the evolution of the virus and identifying variants(8). For example, GISAID(9), the primary repository of assembled SARS-CoV-2 genomes at the time of writing, has over 11 million genome sequence samples submitted and list on their website several databases that use this new data to track the evolution of the virus over time. However, interpreting a reported nucleotide or amino acid substitution often requires sifting through pieces of information related to affected proteins of the virus. Such information is scattered throughout the web in many papers and databases. If a mutation is found at a certain coordinate, a thorough investigation delving into

multiple papers is required to understand the functional importance of that one nucleotide and its surroundings. To remedy this, we created SARSNTdb, a compact database of highly interlinked data records that can allow the user to rapidly navigate from genome positions to functional/selective constraints on the corresponding nucleotides and amino acids.

Genome databases typically list coordinates of coding and non-coding regions and provide their annotations per such region. In contrast, SARSNTdb is nucleotide-centric, it allows querying annotations for every position in the genome from the perspective of potential selection factors affecting the corresponding nucleotide. Public attention to SARS-CoV-2 virus has generally focused on mutations occurring in its genome (and their impact on vaccine efficacy and virus spread). Most frequently, SARS-CoV-2 mutations are viewed through the prism of immune system evasion (3, 10). While this is relevant for the (most widely known) viral spike protein, the general public and scientific community are often at a loss when other substitutions are considered, especially silent ones or short insertions/deletions (indels). Given the significant interest in variants of concern (VOC), strong focus on selection would provide complementary functional context for respective VOC mutations, beyond the trivial synonymous/non-synonymous designations. Examples of such context include repeats, secondary structure formation, intragenomic interactions, nucleotide and amino acid conservation, and mutation prevalence.

For example, it is known that repeats and their variations play a critical role in production of subgenomic mRNAs in coronaviruses (5). These repeats direct the RdRP to jump from one coordinate to the leader sequence at the 5’ end of the genome, creating a template recognizable by the host’s ribosomes (5). In addition, variations in the RNA declared as synonymous may alter the RNA secondary

structure formation. Currently, RNA secondary structure is being investigated for drug targeting and variations on the structure could affect therapeutic effects of drugs currently in development (11). Finally, VOCs, such as Omicron, display large number of both spike (12) and non-spike substitutions or indels. Hence, it is important to recognize the effect of mutations commonly passed over as non-synonymous or taking place in regions not under intense scrutiny.

To ensure the consistent cataloguing of the nucleotide and amino acid substitutions, we re-evaluated mutations across >22,000 patient samples, for which raw metatranscriptome datasets of sufficient quality were available in NCBI's SRA(13). We avoided taking mutations reported in GISAID since it contained already assembled genomes. In many cases, such genomes contained unresolved regions or segments of low genome coverage (jeopardizing mutation calling or producing massive sections of missing data) and it was not possible to tell how reliable these assemblies were. By calling mutations *de novo* we increased the consistency of the substitution data collected.

Perspectives

We feel that this database fills a niche yet unfilled by other SARS-CoV-2 databases. Other databases often have a focus on the amino acid coordinates and non-synonymous substitutions, passing over other important SNVs at the RNA level. In addition, data was scattered so that upon sequencing a novel mutation, an investigation delving into several papers and databases was required to understand the functional importance of affected nucleotide. Now, with SARSNTDB, one can simply input that coordinate into our *Genome Search* function and find what proteins it is included in and the selective factors that act upon it.

This database will continue to be useful to researchers or students new to SARS-CoV-2 as most of the data it provides remains relatively constant, even with new variants emerging. The data in the database is subject to change of course as our understanding of the virus continues to evolve. One example of this is data related to SNVs, of which there is a great deal available online already. However, instead of using GSAID's available SNV data we downloaded a large amount of data from NCBI's website and aligned them using GROM. This was done as we found excessive amounts of soft clipping with the commonly used Minimap2. GROM, on the other hand, had less. Overall, we realigned and called variants on >16000 samples and found ~33000 unique SNVs. These data are also available on our database.

Unique datasets also available on our website include both visualized comparisons of SARS-CoV-2 and SARS-CoV, and repeats of SARS-CoV-2. Both of these data were of interest to our lab and we wanted to make them available to others as well.

MATERIAL AND METHODS

Data Collection and Processing

We downloaded mutation data from the NCBI's SRA using the prefetch feature. Datasets from the UK had been originally aligned using minimap2(14). We found that minimap2 produced in these SARS-CoV-2 datasets excessive soft clipping, leading to lower accuracy when calling variants. To solve this, we unaligned the reads using Samtools (15) to convert them to fastq files. We then re-aligned the fastq files to the reference sequence using BWA mem, producing what we found to be more consistent

alignment patterns with less soft clipping. We did the same alignment step for the samples, where fastq files were available in SRA. We then used GROM (16) to find SNVs in the data. In total we found ~33,000 unique SNVs using GROM across >22,000 samples. The output VCF files were then converted to SQL files detailing the sequencing platform, coordinates, and alternate nucleotides for each sample.

To identify repeats in the SARS-CoV-2 genome, we analysed the Wuhan reference sequence(1) (NC_045512.2) using UGENE(17) with the default settings. We then organized the repeats by coordinate and identified repeats that were super-repeats of one another (superstrings of shorter repeat strings) using in-house scripts.

Data Type	**Tool Used**	**Source**
Protein Structure Visualizations	I-TASSER – M.L. based protein structure predictor	Zhang group(18)
SHAPE reactivities of RNA	SHAPE-MaP	Yang et al (19)
SHAPE reactivities of RNA	icSHAPE	Sun et al(11)
Normalized SHAPE reactivities of RNA	SHAPE-MaP and DMS-MaPseq	Manfredonia et al(6)
Intragenome RNA interactions	SPLASH	Yang et al(19)
Repeat Detection and Coordinates	UGENE	Scripts ran in-house
SNV Data	GROM	Produced in-house

Table 1: A display of the types, sources, and tools used to generate data that populate the database

Data on Protein and RNA Structure

Protein structures were obtained from the Zhang group who has used I-TASSER to predict protein structure for all SARS-CoV-2 proteins (18). Their predictions are highly accurate for the SARS-CoV-2 proteins despite relatively few homologous sequences with available protein structures.

To show the secondary structure of SARS-CoV-2 genomic RNA we collected several datasets from groups that have measured the viral RNA accessibility at a single base resolution. The first of these was taken from Manfredonia et al. (6) who has used SHAPE and DMS mutational profiling to find secondary structure maps with single base resolution. Yang et al.(19) has used SHAPE-MaP to find the reactivities of the reference sequence as well as a delta variant sequence. Finally, Sun et al (11) has used icSHAPE to map reactivities.

Data presented as *Intragenome Interaction Data* represent regions of pairwise RNA interactions across the genome. Such regions have been detected via proximity ligation sequencing was performed using SPLASH to find these regions in Vero-E6 infected cell (19).

Gene, Protein and Functional Domain data

We obtained the coordinates of viral non-coding regions, its genes and proteins, their respective nucleotide and amino acid sequences from the NCBI record of the SARS-CoV-2 (NC_045512.2). SARS-CoV's information was retrieved in the same way from the NCBI record of the Tor reference sequence (NC_004718.3). We then performed a thorough literature review (across hundreds of papers) of proteins in SARS-CoV-2 and SARS-CoV to obtain their functional descriptions. Next, we

identified the available coordinates of functional domains in both viruses. Using BLAST(20) and CLUSTAL-W(21), we further performed pairwise alignments of the proteins of SARS-CoV-2 and SARS-CoV to evaluate the levels of amino acid identity of the homologous functional domains. We manually curated mismatched coordinates of such homologous domains between different studies, produced reconciled coordinates and transferred the domain annotations, further accompanied on respective pages by the publications describing them.

RESULTS AND DISCUSSION

Succinctly, the data in the database is retrieved by users via two main query hubs. One is the *Genome Search* page and is comprised of several datasets and information retrieved from literature. The other is made available in the *Mutation Search* page (and *Repeat* page), presenting results of our re-analysis of >22,000 patient samples obtained from NCBI's publicly available SRA SARS-CoV-2 genomes. We interlinked these sections comprehensively in order to provide the user an easy way to carry over the findings gained in one section to another.

Web App Implementation and User Interface

Users can access SARSNTdb at https://grigoriev-lab.camden.rutgers.edu/sarsdb/

The website is implemented in PHP (version 7.4.29) and the SQL server through mysql (version 15.1) with MariaDB (distribution 10.3.34).

The interface of the database consists of several tabs. The *Search* tab has a dropdown menu that brings the user to a *Genome Search*, *Mutation Search*, and a *Repeat Search*. These searches are interconnected to allow the user to take information gleaned from one search into another. The *Help* tab instructs the user on how to use the website by

providing an example. The *Reference* tab brings the user to this article where they can learn about the data sources and how the website was constructed.

Accessing Gene, Protein and Functional Domain Details

The *Genome Search* page allows the user to specify nucleotide coordinate intervals and find information about functionally relevant regions of the SARS-CoV-2 virus that overlap or are contained between these coordinate pairs. Such regions most often correspond to genes and functional domains they encode. Also, this search reports about nearby repeats and intragenomic interactions obtained using a SPLASH technique (19).

One can also select a single ORF or Nsp from a menu to get to such genes. Their protein products are described on the *Protein Detail* page [Fig.1]. In addition to images of the predicted structure of the SARS-CoV-2 proteins, their functional domains, smaller motifs, and certain amino acid residues with annotated functionality are also displayed graphically. At the bottom of this page there are the relevant RNA and Protein sequences derived from the respective NCBI reference.

Since SARS-CoV-2 domains are typically derived from the previous body of work on SARS-CoV, we devoted a special page for each protein in both viruses for comparing domains. This page is linked form the *Protein* page and contains a table detailing the similarities of the two viruses and an alignment of both protein sequences created using CLUSTALW(21) and BLAST(20). The coordinates in the table are derived from primary literature and review papers (that can be accessed by clicking the hyperlinks on the coordinates) and sometimes they differ, despite being reasonably well aligned. The residue identities and positives were derived by performing a BLAST alignment of the genome sequences.

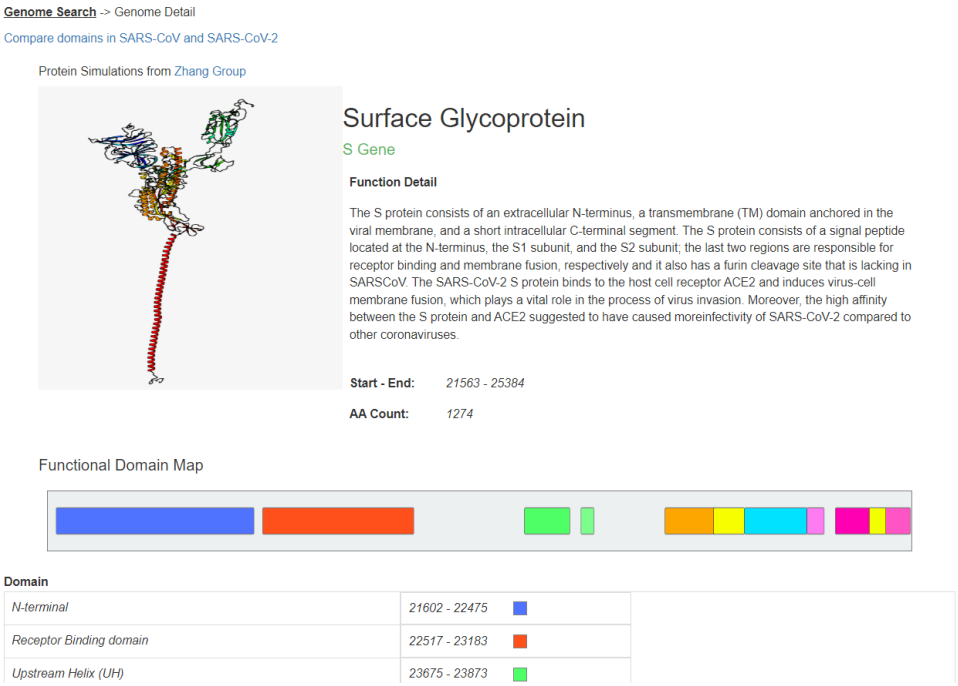

Fig.1: The Genome Detail page of the S protein

Visualization of Mutation and RNA Structure Details

The *Mutation Search* page allows the user to search for mutations in a nucleotide range or within a gene. The search results are bar graphs depicting the number and type of substitutions in the range. The bars are also subdivided by sequencing platforms. If a more granular view of the mutations is needed the user can click on *Mutation Detail* to see expanded information related to the mutations in the that nucleotide range. We also provided links “Try your luck in Covariant” for each non-synonymous substitution as some of the frequent mutations (mostly in the S gene) can be reflected in that database. Below the *Substitution Frequency* table there is a histogram that displays SNV frequency across the nucleotide range selected.

Also on this page is SHAPE data that may help inform the user why certain regions may be conserved due to the secondary structure constraints. When the size of the searched region is large, the SHAPE Data is displayed in intervals where the SHAPE value is averaged across that region. If the size of the searched interval is under 100

nucleotides, each position and the SHAPE value is displayed individually. If the shape value is above 0.5 it is displayed in blue indicating a high reactivity while below 0.5 is displayed in red and indicates a low reactivity. We selected several SHAPE datasets and displayed them in separate graphs. These data make up our *Mutation Search* page and it is visualized on the page using CanvasJS (22).

Repeats in the Genome

The *Repeat Page* [Fig.2] allows the user to search the SARS-CoV-2 reference sequence for repeats of size 6 nucleotides or greater. Displayed on this page is the genome schematic with proteins coloured distinctly. When a repeat is found red lines appear on the genome indicating repeat locations, and a table displaying the coordinates of the repeats as well as which protein they appear in is displayed. Also available are repeats, which are super-strings containing the searched repeat; these are deemed super-repeats. For example, the repeat AACAGGA is a super-repeat of AACAGG as the former is a super-string of (i.e., contains) the latter. Clicking on these super-repeats brings the user to a *Repeat Page* for the super-repeat (with their respective super-repeats, if available). For the default search on the *Repeat Page* and a clear biological example, we provide the minimal repeat of the transcription regulatory sequence (TRS) from the SARS-CoV-2 virus(5), with all locations of canonical TRS visualized throughout the genome for the user.

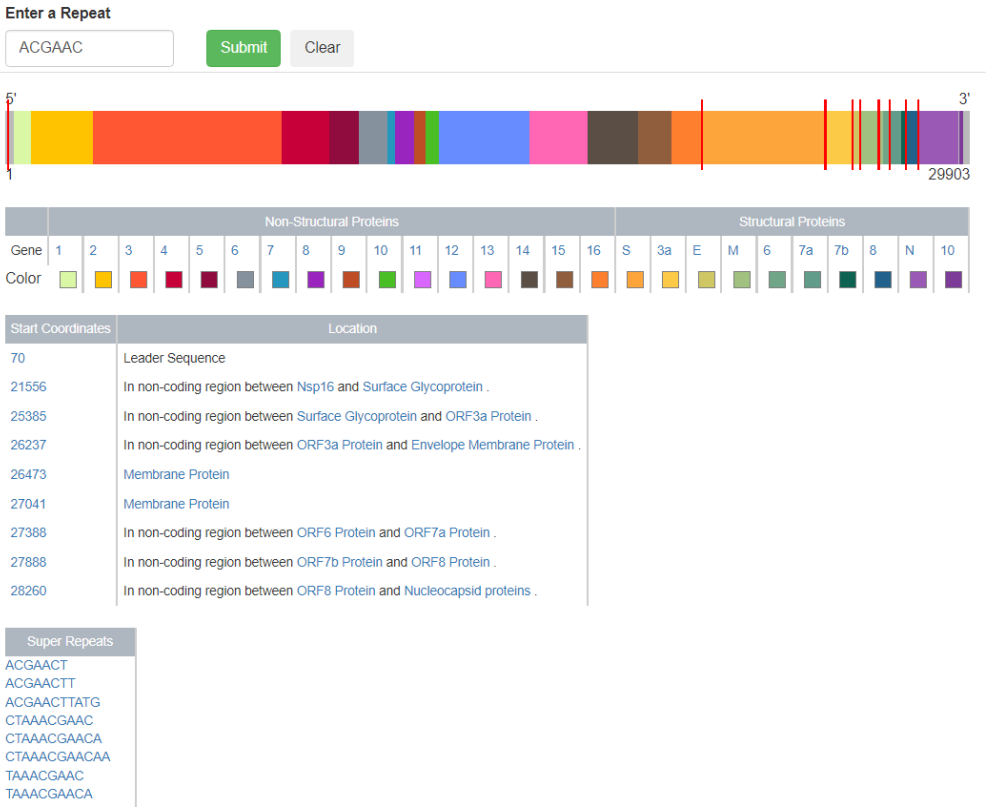

Figure 2. Visualization of the leader sequence repeat ACGAAC across the genome.

Case Study

As stated previously there are many databases tracking the waves of VOCs and their typical mutations. The virus continues to evolve, and even the general public is made aware of new substitutions in the best-annotated spike protein. When new mutations appear, it is important to be able to quickly identify where they occur and analyse their effects by detecting genome features nearby. Furthermore, substitutions take place not only the spike protein, yet those affecting other parts of the genome are typically ignored in the databases and many analyses.

In contrast, SARSNTdb could be an excellent starting point for such quick evaluation. Consider the mutation C28311T, found in Omicron. Let us first go to our *Genome Search* page and input the coordinate 28311 and find it is part of two overlapping genes, encoding the Nucleocapsid protein as well as ORF9b. In N it is part of the N-

terminal arm/Intrinsically disordered region. In ORF9b we see it is part of the site that interacts with NEMO. We also see that it is a part of some common repeats and has intragenomic interactions at the 5' end of the protein as well as a region 200nt away that it binds with. These close intragenomic interacting regions could form pockets that are often the targets of RNA based therapeutics (11). Clicking *View Details* for ORF9b, we find its function and see it supresses the innate immune system through regulating Mitochondrial Antiviral Signalling pathways (7). In comparing it to SARS-CoV we find this domain is not well conserved with only 63% similarity overall. In the paper linked via the domain coordinates in the table, we find that this region, when deleted, resulted in a loss of function of the protein and its interaction with NEMO (7). If this nucleotide change results in a non-synonymous mutation, it could affect the function of the protein. By clicking *Mutations* on the table, we are brought to the mutation page showing the NEMO interaction region's mutation frequencies, SHAPE scores and, if we click Detail, a breakdown if that specific mutation has been found. If it has been found the detail page will also show the type of variant it creates. In this case the SNP has been found previously and changes a proline to a serine. We find it has been found in >1500 samples but the link "Try your luck in Covariant" will return Error 404, as this variant is not yet annotated there. In addition, the SHAPE score of this nucleotide is low according to all datasets, indicating that it may be prone to forming secondary structures within the RNA. Overall, with all such results about this mutation we can conclude that it should be monitored as it has been persisting over time and now, with Omicron spreading rapidly, may be gaining increased prevalence. This mutation could affect the ability of ORF9b to supress the innate immune system through interacting with NEMO and its effects should be explored further.

Conclusion

SARSdb is a database for users of varying levels of knowledge about virology or genomics. It provides nucleotide-level functional information about various aspects of the SARS-CoV-2 genome. It features a quick and easy coordinate-based search for SARS-CoV-2 gene and protein functions, mutations found in patient samples, structural and sequence elements of the virus RNA and several other features. We reviewed, analysed, and provided visualization for data that could help users to better understand the virus, and to do this rapidly. We will continue to add mutation data as we process other samples with GROM.

CONFLICT OF INTEREST

The Authors declare no conflict of interest

FUNDING

The work in A.G.'s lab is supported by the National Science Foundation [MCB-2027611 to A.G.] and National Institutes of Health [R15CA220059 to A.G.].

REFERENCES

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.
2. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. 2020;91(1):157-60.
3. Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. The Lancet. 2021;398(10317):2126-8.
4. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature. 2021;589(7840):125-30.
5. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181(4):914-21.e10.
6. Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Research. 2020;48(22):12436-52.
7. Wu J, Shi Y, Pan X, Wu S, Hou R, Zhang Y, et al. SARS-CoV-2 ORF9b inhibits RIG-I-MAVS antiviral signaling by interrupting K63-linked ubiquitination of NEMO. Cell Reports. 2021;34(7):108761.
8. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021 [Available from: https://covariants.org/.
9. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021;3(49):1049-51.
10. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature. 2022;602(7898):657-63.
11. Sun L, Li P, Ju X, Rao J, Huang W, Ren L, et al. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. Cell. 2021;184(7):1865-83.e20.
12. Gobeil SMC, Henderson R, Stalls V, Janowska K, Huang X, May A, et al. Structural diversity of the SARS-CoV-2 Omicron spike. Molecular Cell. 2022;82(11):2050-68.e6.
13. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7-19.
14. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
16. Smith SD, Kawash JK, Grigoriev A. Lightning-fast genome variant detection with GROM. GigaScience. 2017;6(10):gix091.
17. Okonechnikov K, Golosova O, Fursov M, team tU. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166-7.
18. Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Reports Methods. 2021;1(3):100014.
19. Yang SL, DeFalco L, Anderson DE, Zhang Y, Aw JGA, Lim SY, et al. Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. Nature Communications. 2021;12(1):5113.
20. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
21. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-80.
22. Inc F. Canvas.js. non-commercial 3.6.6 ed: Fenopix Inc.