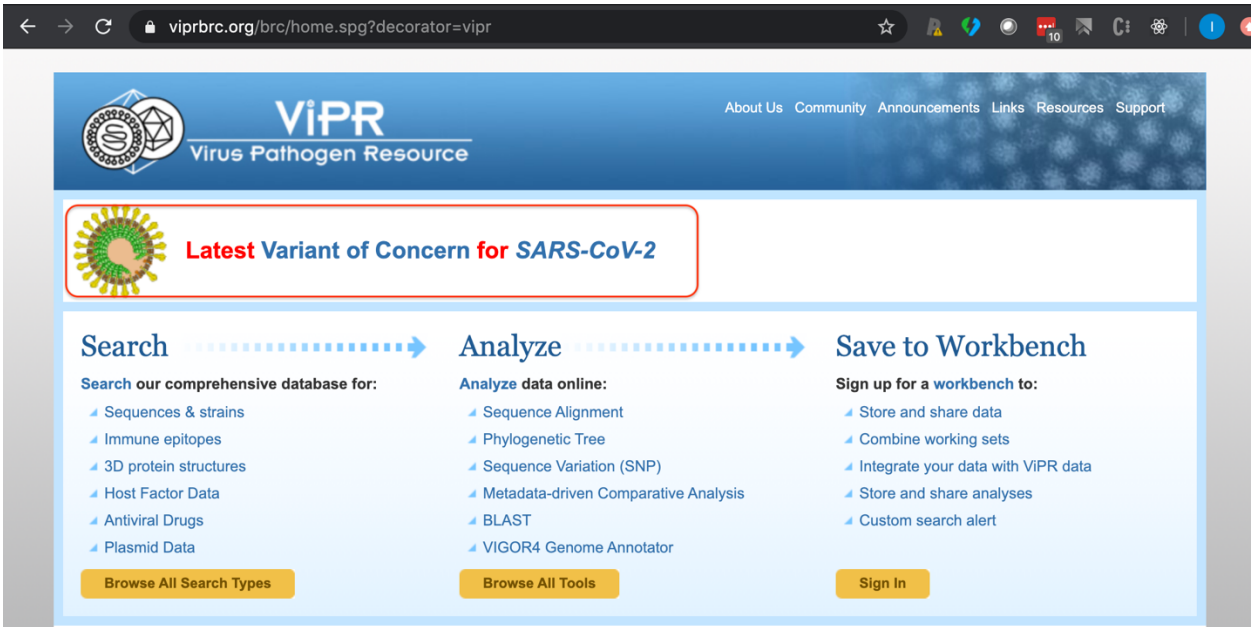
The SARS-CoV-2 Real-time Tracking and Early Warning System for SARS-CoV-2 Variants and Lineages of Concern (VoCs/LoCs)

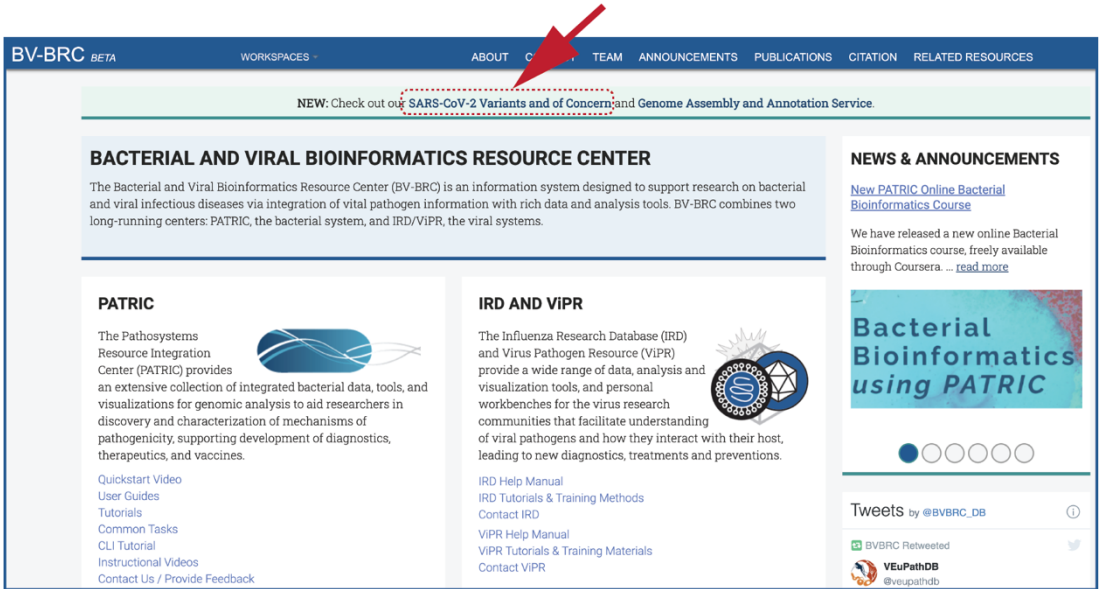
COVID-19 is currently a worldwide epidemic, with viral mutations creating new, time-critical concerns. The BRCs play an essential role in supporting infectious disease research and must respond to emerging diseases in a manner that best enables researchers working on countermeasures. Thus, a major goal of the system enhancement is to facilitate the rapid, near real-time, identification of these variants/lineages of concern (VoCs/LoCs):

- Variants affecting transmissibility
- Variants affecting virulence
- Variants affecting antibody neutralization (immune evasion)
- Variants affecting detection

The SARS-CoV-2 resource can be accessed through banners on the ViPR home page and *Coronaviridae* and SARS-CoV-2 portal pages:

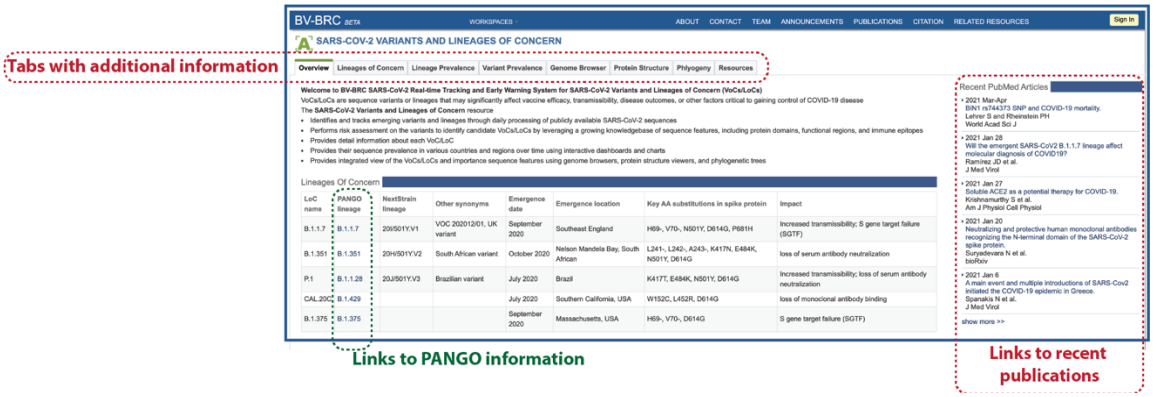


There is an additional way to access this information through the beta site for BV-BRC (https://beta.bv-brc.org/). To drill down and see more information about SARS-CoV-2 lineages and variants of concern, click on the SARS-CoV-2 Variants of Concern hyperlink that can be found at the top of the page.



Variants of Concern Overview

Clicking on the hyperlink will take you to an Overview page that shows summary information about the initial five lineages of concern (e.g., PANGO lineage information, key amino acid substitutions, functional impact. This page also contains a series of tabs that provide links to additional information about sequence prevalence, overlap with sequence feature through a genome browser, etc.

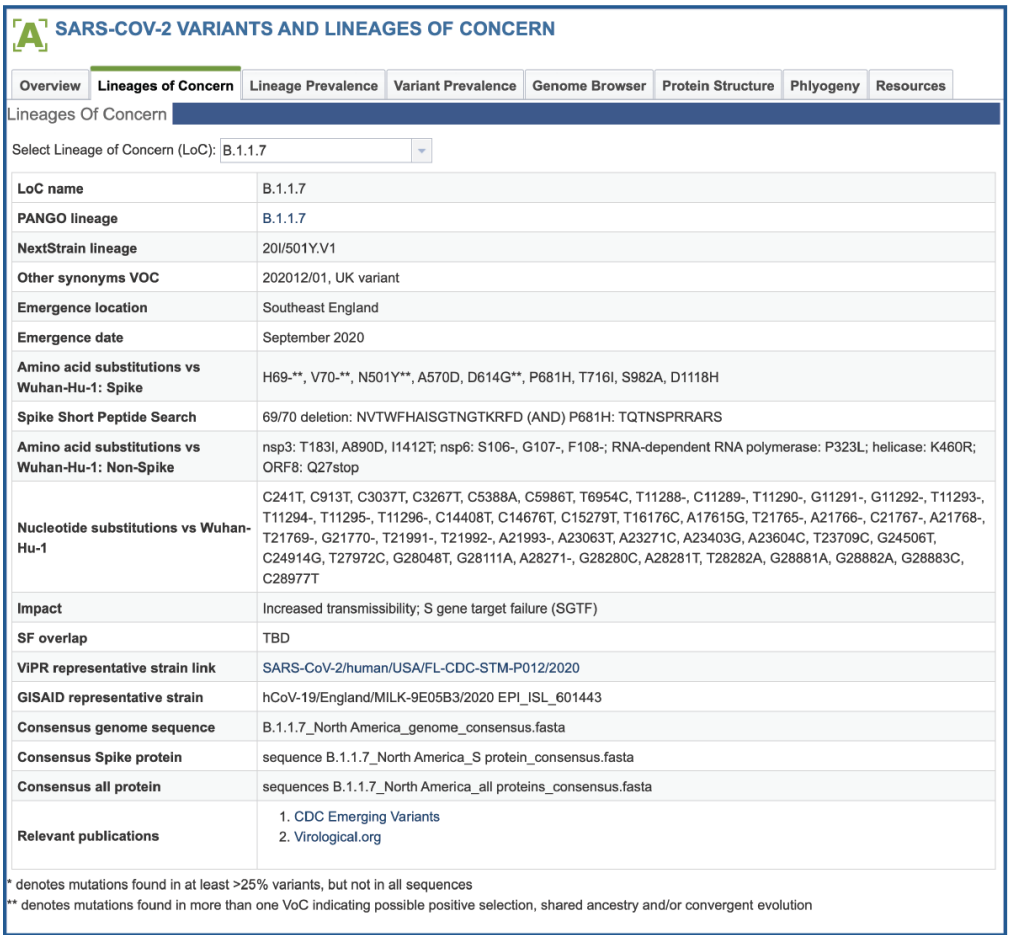


Lineage of Concern

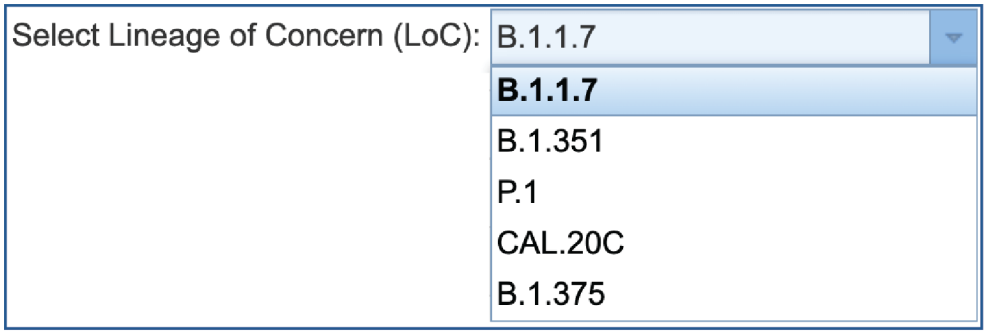
To see more information on the five lineages of concern, click on that tab.

This will open a page that shows detailed information on these lineages. Although the default shows the B.1.1.7 lineage, clicking on the down arrow below that box will allow you to view the detailed information about any of the defined LoCs. The information provided on these pages includes:

- The name of the LoC
- The PANGO lineage
- The NextStrain lineage
- Other synonyms VOC
- Emergence location: The place where the lineage/variant was first identified
- Emergence date: The date when the lineage/variant was first identified
- Amino Acid substitutions in the Spike protein: The mutations are defined by comparing a consensus sequence derived from available lineage isolates to the reference strain (see SOP) and are depicted as follows: Wuhan amino acid residue_Location in the Spike Protein_Amino acid replacement in the LoC.
    - D614G means that a D in the Wuhan strain was replaced by a G in the B.1.1.7 lineage.
    - H69- means that the H in the Wuhan strain was lost in the B.1.1.7 lineage
    - -69H would mean that an insertion of an H at site 69 would be found in the B.1.1.7 lineage.
    - An * means that the mutation is found in 25% of the variants
    - A ** means that mutations are found in more than on variant of concern.
- Amino acid substitutions in non-spike protein and whole genome nucleotide substitutions are also provided.
- Links to consensus sequences and representative strains.
- Reference to relevant citations regarding their functional impact.



To see information on additional lineages of concern, click on the drop-down box at the end of the lineage name. This will open the list of lineages, and after clicking one, the page will reload to show that data.



Lineage Prevalence

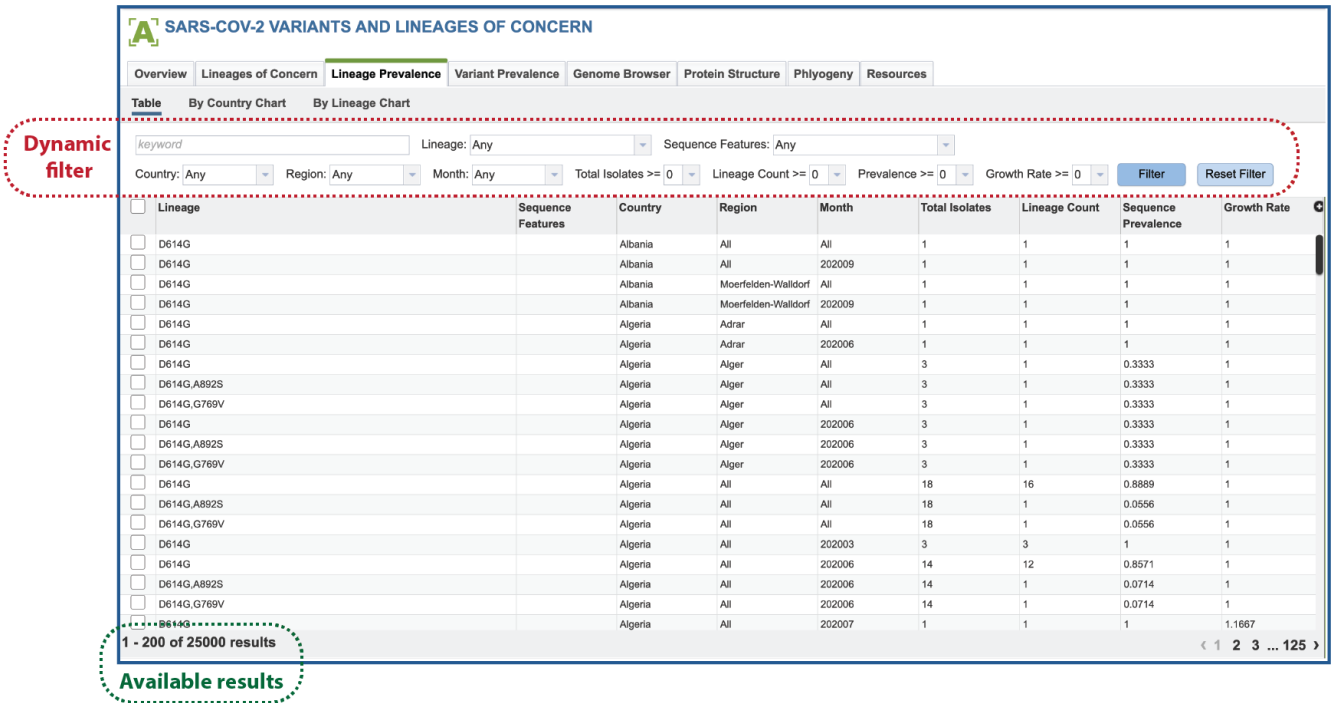
BV-BRC is currently focusing on five lineages of concern (B.1.1.7, B1.351, P.1, CAL.20C and B.1.375). The sequence characteristics that define these lineages are defined on the Lineages of Concern tab, but there is additional information available about these lineages that can be

seen by clicking on the Lineage Prevalence tab.

This will open a table that shows information about all non-synonymous substitutions identified in the Spike protein. The table has three inner tabs (Table, By County Chart, and By Lineage Chart). The table has a dynamic filter at the top, followed by the table that has the following columns:

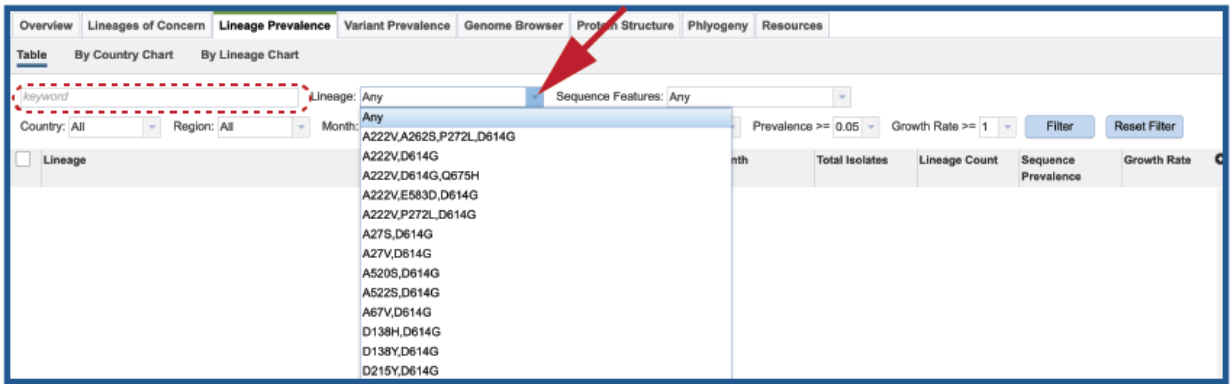
- Lineage: Information on all unique amino acid substitutions identified in the spike glycoprotein is provided.
- Sequence Features: These include protein domains, secondary structure elements, functional regions (e.g., receptor binding residues, enzyme active sites), sites of post-translational modifications, protein-protein interaction sites, site in which substitutions have been shown to alter virus characteristics, antibody and T cell epitopes.
- Country: The country where the lineage is found
- Region: The region of the country where the lineage is found
- Month: The year and month that the lineage has been found.
- Total Isolates: The number of isolates found that have the variant. The number will change based on the dynamic filter.
- Lineage Count: The number of isolate genomes in this lineage. This number will vary depending upon the filtering.
- Sequence Prevalence: The number in this column represents the total isolates of this lineage from the indicated country or region and month divided by the total number of all isolate sequences from that same country or region and month.
- Growth Rate: The growth rate is computed month over month. This number in this column is the prevalence of this variant for the current month divided by the past month.

A summary of the available data is provided at the bottom left of the table. This number changes when the filter is used.

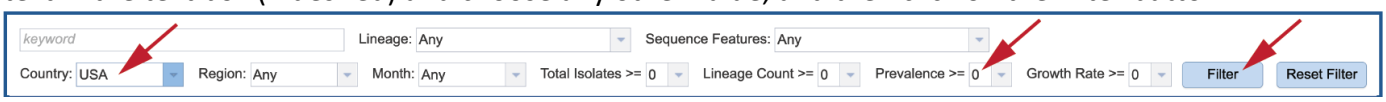


Filters are provided for each of the columns in in the table. An additional keyword filter is also provided and can be used for any term.

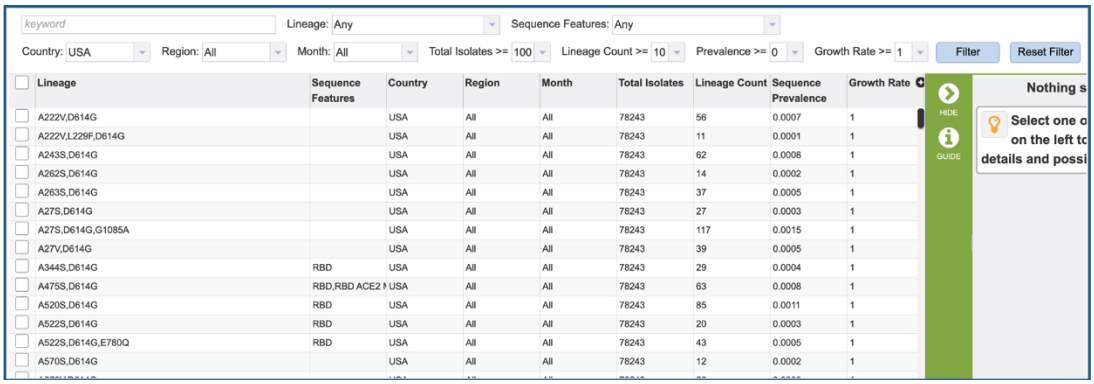
Text can be entered into the filter box. Click on the down arrow at the end of any of the other filters will open a drop-down box where selections can be made.



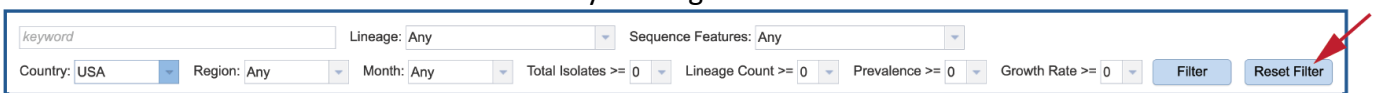
Any or all of the filtering facets can be used to generate Boolean AND operations. Simply enter text in the text box (if desired) and choose any other value, and then click on the Filter button.



This will reload the table to show the results based on that selection.



The filter can be reset to the default values by clicking on the Reset Filter button.



Variant Prevalence

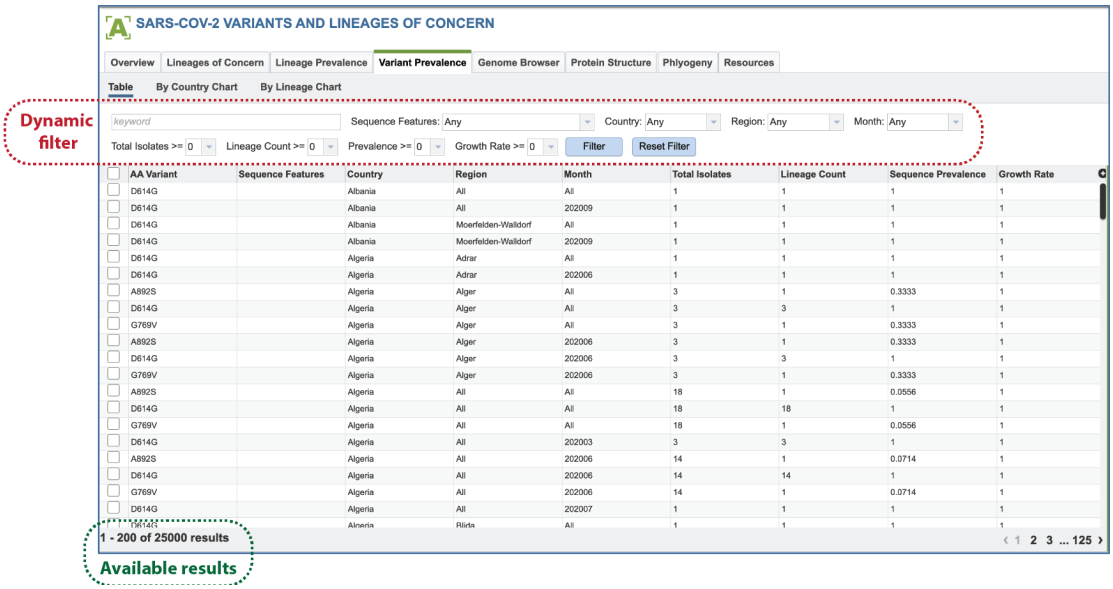
While Lineages of Concern describe constellations of mutations that share a common evolutionary ancestor, Variants of Concern track single amino acid substitutions that may share a common ancestor or have been generated through convergent evolution. The Variant Prevalence information is organized in a similar fashion to the LoC Prevalence described above. All variants are defined based on the reference Wuhan-hu-1 genome. To find information on the individual SARS-CoV-2 variants and their prevalence, which currently only include variations in the Spike protein (S), click on the Variant Prevalence tab.

This will open a table that shows all the variants in the spike protein. The table has three inner tabs (Table, By County Chart, and By Lineage Chart). The table has a dynamic filter at the top, followed ty the table that has the following columns:

- AA Variant: The amino acid variant (indicated by the amino acid called in the Wuhan spike protein, the location of that amino acid in the protein sequence, followed by the amino acid seen in the variant).
- Sequence Features: These include protein sequence features, and include protein domains, secondary structure elements, functional regions (e.g., receptor binding residues, enzyme active sites), sites of post-translational modifications, protein-protein interaction sites, site in which substitutions have been shown to alter virus characteristics, antibody and T cell epitopes.
- Country: The country where the variant is found
- Region: The region of the country where the variant is found
- Month: The year and month that the variant has been found
- Total Isolates: The number of isolates found that have the variant. The number will change based on the dynamic filter.

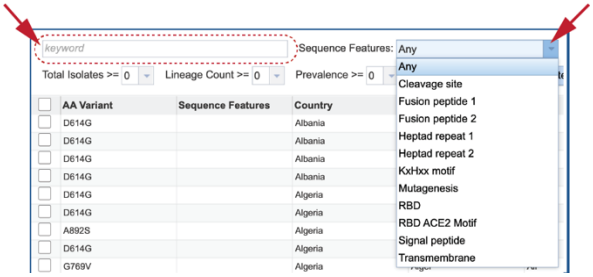
- Lineage Count: The number of lineages that has the variant.
- Sequence Prevalence: The number in this column represents the isolates with a given aa variant from the indicated country or region and month divided by the total number of all isolate sequences from that same country or region and month.
- Growth Rate: The growth rate is computed month over month. This number in this column is the prevalence of this variant for the current month divided by the past month.

A summary of the available data is provided at the bottom left of the table. This number changes when the filter is used.

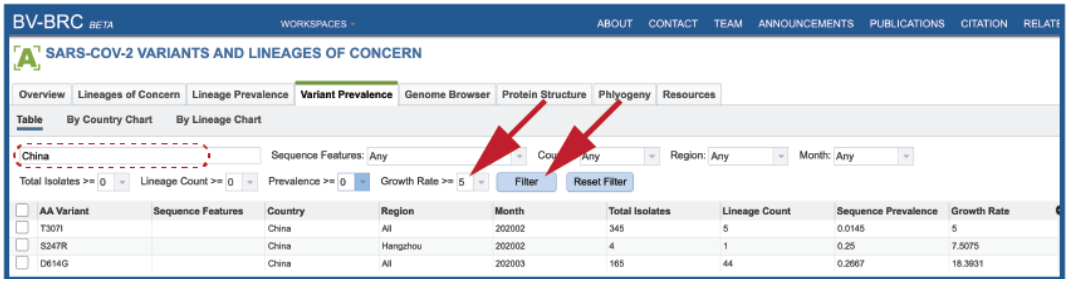


Filters are provided for each of the columns in in the table. An additional keyword filter is also provided and can be used for any term.

Text can be entered into the filter box. Click on the down arrow at the end of any of the other filters will open a drop-down box where selections can be made.



Any or all of the filtering facets can be used. Simply enter text in the text box (if desired) and choose any other value, and then click on the Filter button. This will reload the table to show the results based on that selection.



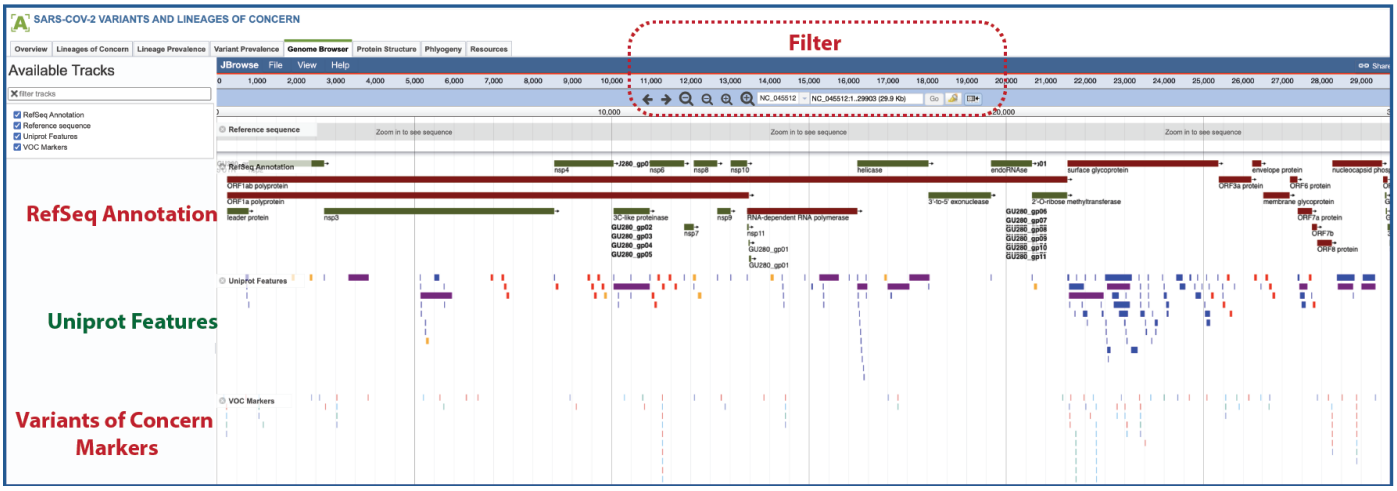
The filter can be reset to the default values by clicking on the Reset Filter button.

Genome Browser

As with other members of the *Coronaviridae* family, SARS-CoV-2 exhibits a relatively high mutation rate. As a result, any given SARS-CoV-2 sequence typically carries multiple nucleotide substitution in comparison with the initial outbreak strain; for this purpose the Wuhan-hu-1 reference strain (NC_045512) can be considered a close representative of the initial outbreak strain. The JBrowse genome browser[1] displays the sequence for Wuhan-hu-1, its RefSeq annotation, corresponding Uniprot sequence features and IEDB B cell/Ab epitopes. In addition, substitutions that result in increased transmissibility or immune escape are variants of concern (VoCs) from a public health perspective, and are also displayed on the browser in their relation to features of the Wuhan-hu-1 reference strain.

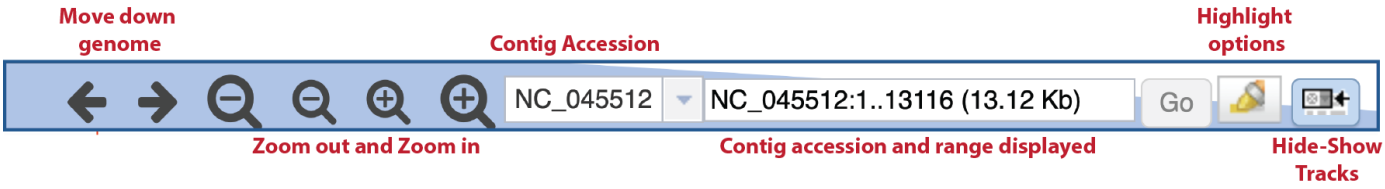
To locate the genome browser, click on the tab named Genome Browser

This will open the genome browser page. The page shows the annotations and the VoCs. The available tracks are shown to the right of the visualization.

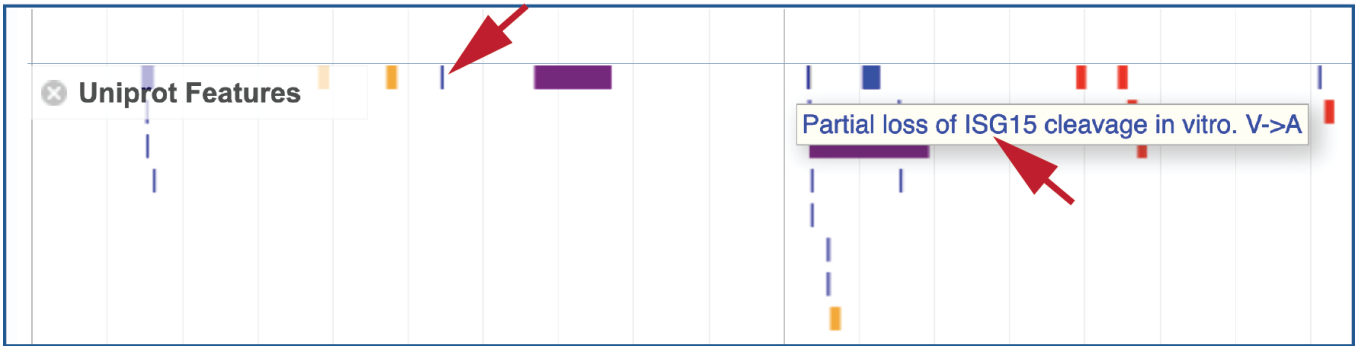


The browser view includes a dynamic filter. Researchers can use the arrow icons to move rapidly in a 5' or 3' direction along the reference genome. Magnifying icons with a minus (zoom out) or a plus (zoom in) sign within them provide an additional way to focus in on different aspects of the view.

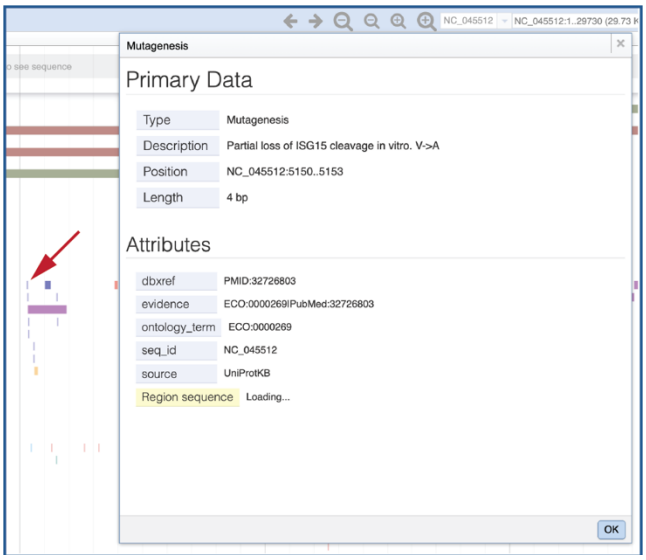
The text box following contig accession shows the contig and the start of the sequence, end of the sequence, and size of the sequence displayed. The display can be changed by adjusting the numbers following the contig (delineated by a ":"). Take care to maintain the format (1..13116) with the ".."between the start and end of the sequence. Researchers can use the highlight option to draw emphasis to a specific area of the genome, and also hide or show tracks using that particular icon.



Many features are displayed on the browser. Zooming in can reveal the names of particular features, but this can also be done by mousing over a feature of interest. This will open a pop-up window that provides the name of that particular feature. The pop-up window usually appears to the right of the feature, and, depending on the particular view, could be somewhat downstream from item.



More information about a specific feature can be seen by actively clicking on that feature from the browser view. This will open a pop-up window that shows the data on the specific annotation (Primary Data) as well as additional data linked to that feature (Attributes). The pop-up window can be closed by clicking on the OK button at the bottom right of the window.

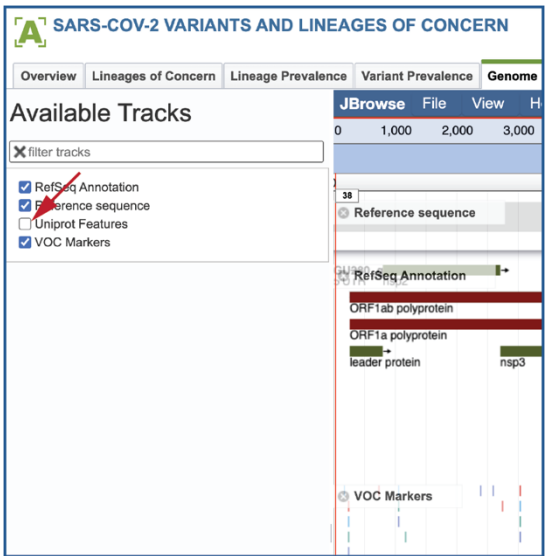


Clicking on one of the RefSeq annotations will load more information about that annotation.

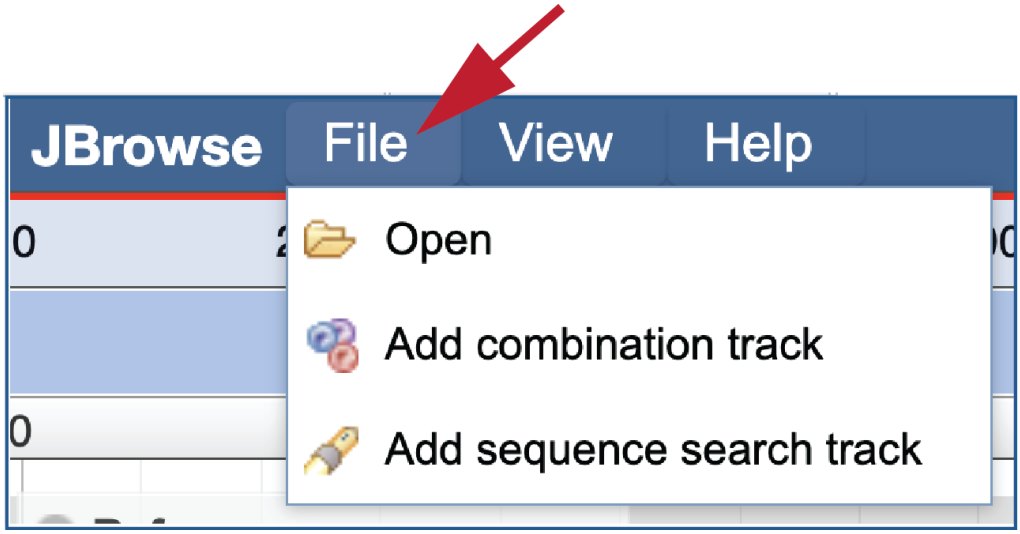
Removing individual tracks from a view is easily implemented by clicking on the “X” in front of the name of the specific track.

The track will be removed once this has happened.

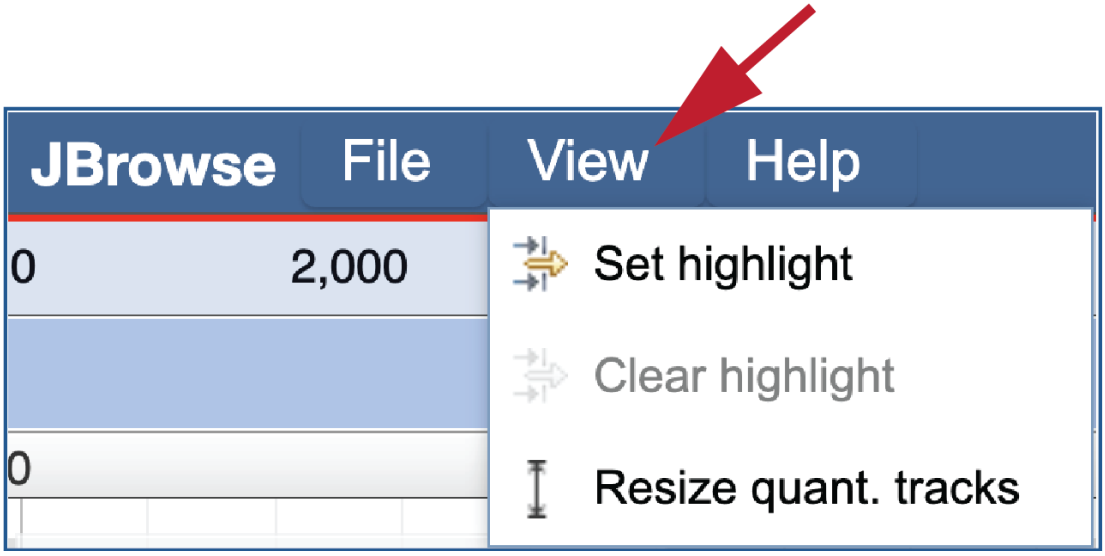
The removed track can be restored by clicking on the check box in front of it in the left side of the browser.



Researchers can also add their own data as an additional track, but it must match the reference genome (in this case the Wuhan-hu-1 genome). Clicking on the File tab and then selecting one of the options will enable this.



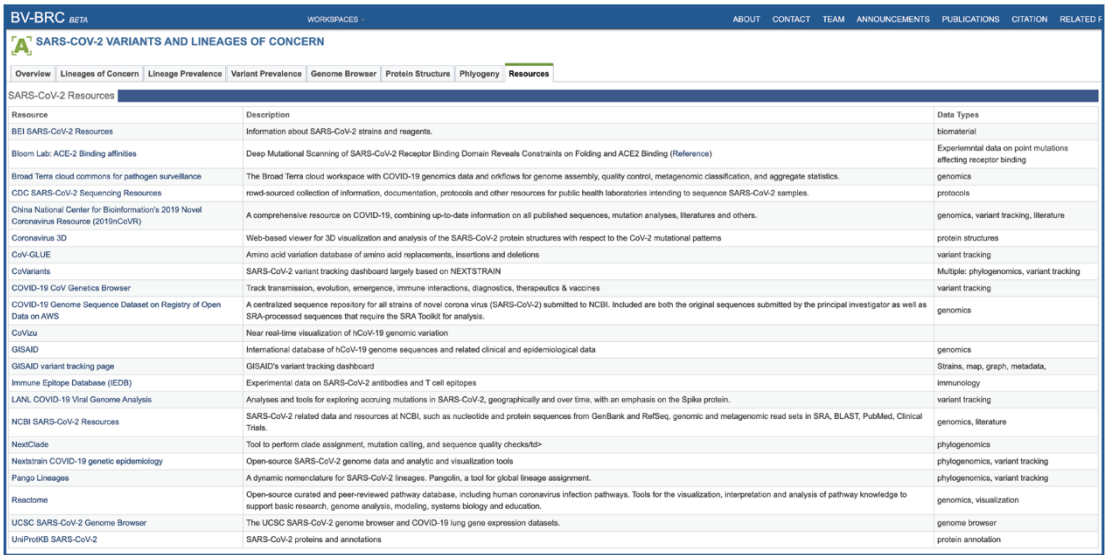
The overall view can also be adjusted by clicking on the View icon.



Additional Resources

Additional resources that provide information on the SARS-CoV-2 strains are provided under the Resources tab. This can be reached by clicking on that tab.

This will open the Resources page, where the name and brief description of the resource and the data types provided by the resource are included. The name of each resource is a hyperlink that will open a new tab to that page.



Resource	Description	Data Types
BEI SARS-CoV-2 Resources	Information about SARS-CoV-2 strains and reagents.	biomaterial
Bloom Lab: ACE-2 Binding affinities	Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding (Reference)	Experiemntal data on point mutations affecting receptor binding
Broad Terra cloud commons for pathogen surveillance	The Broad Terra cloud workspace with COVID-19 genomics data and orkflows for genome assembly, quality control, metagenomic classification, and aggregate statistics.	genomics
CDC SARS-CoV-2 Sequencing Resources	rowd-sourced collection of information, documentation, protocols and other resources for public health laboratories intending to sequence SARS-CoV-2 samples.	protocols
China National Center for Bioinformation's 2019 Novel Coronavirus Resource (2019nCoVR)	A comprehensive resource on COVID-19, combining up-to-date information on all published sequences, mutation analyses, literatures and others.	genomics, variant tracking, literature
Coronavirus 3D	Web-based viewer for 3D visualization and analysis of the SARS-CoV-2 protein structures with respect to the CoV-2 mutational patterns	protein structures
CoV-GLUE	Amino acid variation database of amino acid replacements, insertions and deletions	variant tracking
CoVariants	SARS-CoV-2 variant tracking dashboard largely based on NEXTSTRAIN	Multiple: phylogenomics, variant tracking
COVID-19 CoV Genetics Browser	Track transmission, evolution, emergence, immune interactions, diagnostics, therapeutics & vaccines	variant tracking
COVID-19 Genome Sequence Dataset on Registry of Open Data on AWS	A centralized sequence repository for all strains of novel corona virus (SARS-CoV-2) submitted to NCBI. Included are both the original sequences submitted by the principal investigator as well as SRA-processed sequences that require the SRA Toolkit for analysis.	genomics
CoVizu	Near real-time visualization of hCoV-19 genomic variation	
GISAID	International database of hCoV-19 genome sequences and related clinical and epidemiological data	genomics
GISAID variant tracking page	GISAID's variant tracking dashboard	Strains, map, graph, metadata,
Immune Epitope Database (IEDB)	Experimental data on SARS-CoV-2 antibodies and T cell epitopes	immunology
LANL COVID-19 Viral Genome Analysis	Analyses and tools for exploring accruing mutations in SARS-CoV-2, geographically and over time, with an emphasis on the Spike protein.	variant tracking
NCBI SARS-CoV-2 Resources	SARS-CoV-2 related data and resources at NCBI, such as nucleotide and protein sequences from GenBank and RefSeq, genomic and metagenomic read sets in SRA, BLAST, PubMed, Clinical Trials.	genomics, literature
NextClade	Tool to perform clade assignment, mutation calling, and sequence quality checks/td>	phylogenomics
Nextstrain COVID-19 genetic epidemiology	Open-source SARS-CoV-2 genome data and analytic and visualization tools	phylogenomics, variant tracking
Pango Lineages	A dynamic nomenclature for SARS-CoV-2 lineages. Pangolin, a tool for global lineage assignment.	phylogenomics, variant tracking
Reactome	Open-source curated and peer-reviewed pathway database, including human coronavirus infection pathways. Tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.	genomics, visualization
UCSC SARS-CoV-2 Genome Browser	The UCSC SARS-CoV-2 genome browser and COVID-19 lung gene expression datasets.	genome browser
UniProtKB SARS-CoV-2	SARS-CoV-2 proteins and annotations	protein annotation

References

1 Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a next-generation genome browser. *Genome research* **19**, 1630-1638 (2009).