BV-BRC SARS-CoV-2 Genome Browser - Methods

Last updated: April 29, 2021

Questions, comments: zwallace@jcvi.org

The BV-BRC SARS-CoV-2 Genome Browser is built upon the JBrowse platform. A wide variety of SARS-CoV-2 related data have been compiled and made available as individual data tracks:

- Epitopes
- Functional Features
- Gene and Protein
- Mutation Impact
- Primers and Probes
- Structural Features
- Variants of Concern

Epitopes

Data source: Experimentally determined B-cell epitopes for SARS-CoV-2 imported from the Immune Epitope Database and Analysis Resource (IEDB). The epitope data is also available via the Immune Epitope Search interface in ViPR: https://www.viprbrc.org/brc/vipr_virusEpitope_search.spg?method=ShowCleanSearch&decorator=corona

Data processing: Epitopes identified from different protein accessions were mapped to the RefSeq strain (Wuhan-Hu-1, NC_045512) and then transformed into a Jbrowse GFF track.

Functional Features

Data source: UniProtKB.
https://www.uniprot.org/uniprot/?query=proteome:UP000464024%20reviewed:yes

Data processing: All feature types except for beta strand, coiled coil, helix, and turn were used and transformed into a Jbrowse GFF track.

Gene and Protein

Data source: RefSeq Wuhan-Hu-1, NC_045512.2.

Data processing: Gene, protein, and mature peptide annotations from the above record were transformed into a Jbrowse track.

Mutation Impact

Data source:

1. Mutational effects of SARS-CoV-2 receptor binding domain (RBD) on monoclonal antibody and sera binding were derived from multiple experimental studies by the Bloom lab: https://raw.githubusercontent.com/jbloomlab/SARS2_RBD_Ab_escape_maps/main/processed_data/escape_data.csv
2. Mutational affects on ACE2 Binding Affinity analysis were also obtained from the Bloom lab: https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS/blob/master/results/single_mut_effects/single_mut_effects.csv

Data version: 2021-04-15

Background: All of these were constructed from the human convalescent sera/monoclonal antibody/moderna vaccine elicited antibody escape data for the Spike protein RBD mutant library (PMID: 32841599, 33592168, 33495308). The mutant library was constructed such that each site in the RBD was mutated with 19 different substitutions in the genetic background of Wuhan-Hu-1. The resulting library covers 3804 of the 3819 possible amino acid mutations in the RBD. Mutated RBD proteins were displayed on yeast cells and binding to antibody preparations or ACE2 protein measured. Barcodes were used to label the RBD variants and after the yeast cells were sorted into bins based on binding affinity/antibody escape, sequencing of the barcodes was used to derive raw counts of RBD variants that were escaping antibody binding or leading to changes in ACE2 binding. Statistical approaches were used to process these to derive their “escape fraction” for antibody escape and “bind_avg” for ACE2 binding.

Data processing: For mutational impact analysis for antibody escape we used the values under the “mut_escape” and “normalized_site_total_escape” columns. The values were used to construct the overlaid bar graphs as well as heat maps that can be seen under the mutation impact column. The following gives the specifics into how the antibody escape mutation impact tracks were designed.

1. Therapeutic antibody tracks (AZD8895, AZD1061, etesivimab, casirivimab, bamlanivimab, imdevimab, VIR-7831) → All of these data came from the CSV file by filtering for the therapeutic antibodies. Therapeutic antibodies are those prepared by the companies Regeneron, Eli Lilly, AstraZeneca, and Vir that have either been FDA

approved or are being evaluated through ongoing clinical trials. Each of these tracks represented as overlaid bar graphs used the values from the “mut_escape” column, which are the “escape fraction” values as defined by Bloom et al. and were made up of two BigWig files. One of the BigWig files had the maximum mut_escape value for all possible mutations at each site and the other BigWig file had the median mut_escape value for each site. Using the MultiBigWig Jbrowse plugin (https://github.com/elsiklab/multibigwig) the BigWig files were combined to design the overlaid bar graphs to display max and median values at each site.

2. Moderna vaccine antibodies track → This track used the same method as the therapeutic antibody tracks except the CSV file was filtered for “Moderna vaccine serum”.

3. Polyclonal sera track → All of these data came from the CSV file by filtering for “convalescent serum”. The serum is polyclonal antibody plasma isolated from convalescent COVID individuals. This track, represented as an overlaid bar graph, used the escape fraction values found under the “mut_escape” column. The track was made up of two BigWig files, one file containing the maximum mut_escape value across all convalescent serum samples at each site and the other file containing the median mut_escape value across all convalescent serum samples at each site. The MultiBigWig plugin was used to combine BigWig files into one overlaid bar graph track.

4. Antibody class tracks (Bloom Lab Antibodies by Class/Classes 1-4 Ab Escape) → These data came from the CSV file by filtering for “antibody” within the “condition_type” column. These tracks are represented as heatmaps. The track titled Bloom Lab Antibodies by Class is made up of 35 BigWig files, one for each monoclonal antibody, including those from the therapeutic antibody data. Each BigWig file for each antibody contains the “normalized_site_total_escape” value at each site. The MultiBigWig plugin was used to combine all 35 BigWig files into a

heatmap analysis that color codes the row names by antibody class. The track titled Classes 1-4 Ab Escape uses the same 35 antibodies to make up four BigWig files, one representing each class, where each class BigWig file is made up of the maximum normalized_site_total_escape value across all antibodies within a class at each site. he MultiBigWig plugin was used to combine the four BigWig files into another heatmap analysis track.

5. Values from the "bind_avg" of the associated CSV file was used for the ACE2 Binding Affinity track. Each site in the track denotes the maximum bind_avg value, only if the maximum bind_avg value was greater than 0.1, otherwise the site is given the minimum binding average. The methodology leaves it so that we only represent sites that truly lead to an increase in ACE2 binding affinity.

Primers and Probes

Data sources:

1. Diagnostic primers and probes used by public health agencies: Centers for Disease Control, China; Ministry of Public Health, Thailand; National Institute of Infectious Diseases, Japan; Pasteur Institute, France; United States Centers for Disease Control; World Health Organization.
2. ARTIC sequencing protocol v3 https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3
3. Metagenomic sequencing protocol https://www.protocols.io/view/sars-cov-2-enrichment-sequencing-by-spiked-primer-beshjeb6
4. CRISPER-Cas12-based diagnosis: PMID 32300245 https://pubmed.ncbi.nlm.nih.gov/32300245/

Data processing: Reagent sequences or the reverse-complement sequences were mapped to the reference strain (Wuhan-Hu-1, NC_045512) to obtain the coordinates in the reference. The mapped results were subsequently transformed into a Jbrowse GFF track.

Structural Features

Data source: UniProtKB.
https://www.uniprot.org/uniprot/?query=proteome:UP000464024%20reviewed:yes

Data processing: The following feature types: beta strand, coiled coil, helix, and turn were transformed into a Jbrowse GFF track.

Variants of Concern

Data source: ViPR phylogenetic analysis was used to identify the variant strains, and the Wuhan-Hu-1 representative strain from GenBank was used to map the variants across the strain sequences. ViPR strains aligned to PANGO lineages to identify the lineage name. From there BV-BRC team aggregated all the lineages along with their variants and came to a consensus with a group established by the NIH.

Data processing: Variants were mapped to the reference strain (Wuhan-Hu-1, NC_045512) based on the protein location and the amino acid coordinate. The mapped results were transformed into two Jbrowse GFF tracks, one for amino acid variations and one for nucleotide variations.

References

Dong, 2021. bioRxiv. 2021 Jan 28;2021.01.27.428529. doi: 10.1101/2021.01.27.428529. PMID: 33532768.
Greaney, 2021a. Cell Host Microbe. 2021 Jan 13;29(1):44-57.e9. doi: 10.1016/j.chom.2020.11.007. PMID: 33259788.
Greaney, 2021b. Cell Host Microbe. 2021 Mar 10;29(3):463-476.e6. doi: 10.1016/j.chom.2021.02.003. PMID: 33592168.
Greaney, 2021c. bioRxiv. 2021 Mar 18;2021.03.17.435863. doi: 10.1101/2021.03.17.435863. PMID: 33758856.
Greaney, 2021d. bioRxiv. 2021 Apr 14;2021.04.14.439844. doi: 10.1101/2021.04.14.439844. PMID: 33880474
Starr, 2021a. Science. 2021 Feb 19;371(6531):850-854. doi: 10.1126/science.abf9302. PMID: 33495308.
Starr, 2021b. bioRxiv. 2021 Feb 22;2021.02.17.431683. doi: 10.1101/2021.02.17.431683. PMID: 33655250.
Starr 2021c. Cell Reports Medicine. 2021 Apr 20;2(4):100255. doi: 10.1016/j.xcrm.2021.100255. PMID: 33842902
Tortorici, 2021. bioRxiv. 2021 Apr 8;2021.04.07.438818. doi: 10.1101/2021.04.07.438818. PMID: 33851169