
WashU Virus Genome Browser

Release 1.0

Sep 10, 2020

Contents:

1	Navigating the WashU Virus Genome Browser	1
1.1	Virus Browser Home Page	1
1.2	Genomic Track View	3
1.2.1	Tracks	4
1.2.2	Apps	10
1.2.3	Settings	10
1.2.4	Help	11
1.3	Phylogenetic Tree View	12
1.3.1	Coloring Tree by Metadata	12
1.3.2	Identifying Strains of Interest	15
2	SNV track	19
2.1	Understanding the SNV track	19
2.2	Density mode and the “zoomed-out” view	20
2.3	Full mode and the “zoomed-in” view	20
2.3.1	Nomenclature:	20
2.3.2	Color code:	21
2.4	Behind the SNV track: the “pairwise” format	21
2.4.1	Uploading interface	21
2.4.2	The pairwise format	22
2.4.3	To generate the pairwise format	23
2.4.4	Batch alignment from FASTA to pairwise format	23
2.4.5	Batch upload as json files	24
2.4.6	Upload json-formatted datahub	24
3	SNV2 track	27
3.1	Functionality of SNV2 tracks	27
3.2	Defining the format of SNV2	28
3.3	Scripts for generating snv2 tracks	30
4	Public Data Hubs	31
4.1	Loading in a public data hub	31
4.2	Introducing currently available public data hubs (As of May 24, 2020)	34
4.2.1	NCBI database	34
4.2.2	Nextstrain database	34
4.2.3	GISAID database	35
4.2.4	Diagnostics	35

4.2.5	Putative SARS-CoV-2 Immune Epitopes	35
4.2.6	Recombination Events	37
4.2.7	Viral RNA modifications	37
4.2.8	Viral RNA expression	38
4.2.9	SARS-CoV-2 host transcriptional responses database	38
5	Contact Us	41
6	Cite Us	43
7	Indices and tables	45

Navigating the WashU Virus Genome Browser

1.1 Virus Browser Home Page

The WashU Virus Genome Browser hosts hundreds to thousands of genomic sequence pairwise alignments to the reference for 4 virus species: SARS-CoV-2, severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and Ebola. When users first navigate to the browser landing page (<http://virusgateway.wustl.edu/>), they are provided with a drop-down menu of the 4 hosted viruses (red arrow below), to select which virus they would like to view as their reference.

WashU Virus Genome Browser

About Featured Datahubs

Choose a Virus Reference:

SARS-CoV-2

Ebola

SARS

MERS

SARS-CoV-2

★ Tree View

📄 Data Table

🛒 0 files

📁 Browser View

Resources

📺 Video tutorials

📄 Documentation

📄 Github

WashU Virus Genome Browser

Explore the many ways to visualize viral sequencing data on our openly available platform.

Get started

Browser View

Compare -omics data across multiple SARS-CoV-2 strains or between SARS-CoV-2 and related viral sequences

Upon selecting a reference species and clicking “Data Table” (brown arrow above), the user is then taken to a table populated with sortable and searchable metadata for all available strains of the species. The user can then select strains of interest (examples are highlighted in purple below), and continue to view those particular strains in either our genomic track browser view or see how their selected strains relate to all other available strains of the species in our phylogenetic tree view. Please note that for SARS-CoV-2, only strains housed in Nextstrain (PMID:29790939) are displayed in the data table as well as in the tree view.

WashU Virus Genome Browser

Choose a Virus Reference:

SARS-CoV-2

Resources

- Video tutorials
- Documentation
- Github

Use header row to filter data, click on rows to add to Cart

ID	Accession	Isolate	Molecule Type	Country	Collection Date
1	EPI_ISL_406798	Wuhan/WH01/2019	genomic RNA	China: Hubei, Wuhan	2019-12-26
2	EPI_ISL_402130	Wuhan/WIV07/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
3	EPI_ISL_416397	Shanghai/SH0107/2020	genomic RNA	China: Shanghai,	2020-02-02
4	EPI_ISL_416425	Hangzhou/ZJU-07/2020	genomic RNA	China: Zhejiang, Hangzhou	2020-02-03
5	EPI_ISL_429103	Guangzhou/GZMU0060/2020	genomic RNA	China: Guangdong, Guangzhou	2020-02-09
6	EPI_ISL_402121	Wuhan/IVDC-HB-05/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
7	EPI_ISL_402120	Wuhan/IVDC-HB-04/2020	genomic RNA	China: Hubei, Wuhan	2020-01-01
8	EPI_ISL_408511	env/Wuhan/IVDC-HBF13/2020	genomic RNA	China: Hubei, Wuhan	2020-01-01
9	EPI_ISL_416348	Shanghai/SH0039/2020	genomic RNA	China: Shanghai,	2020-02-06
10	EPI_ISL_413857	Guangdong/2020XN4448-P0002/2020	genomic RNA	China: Guangdong,	2020-01-31
11	EPI_ISL_402123	Wuhan/IPBCAMS-WH-01/2019	genomic RNA	China: Hubei, Wuhan	2019-12-24
12	EPI_ISL_413578	Netherlands/Nieuwendijk_1363582/2020	genomic RNA	Netherlands: North Brabant, Nieuwendijk	2020-03-01
13	EPI_ISL_413581	Netherlands/Oss_1363500/2020	genomic RNA	Netherlands: North Brabant, Oss	2020-02-29
14	EPI_ISL_402127	Wuhan/WIV02/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
15	EPI_ISL_411955	USA/CA/8/2020	genomic RNA	USA: California,	2020-02-10
16	EPI_ISL_434534	Wuhan/IVDC-HB-GX02/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
17	EPI_ISL_412898	Wuhan/HB/CDC-HB-02/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
18	EPI_ISL_402128	Wuhan/WIV05/2019	genomic RNA	China: Hubei, Wuhan	2019-12-30
19	EPI_ISL_430016	USA/CA-SR016/2020	genomic RNA	USA: California, San Diego	2020-03-24
20	EPI_ISL_430739	Beijing/BJ625/2020	genomic RNA	China: Beijing,	2020-02-06

1.2 Genomic Track View

When users select the “Browser View” (orange arrow above), they are taken to our standard browser view layout, adapted from the WashU Epigenome Browser (<https://epigenomegateway.wustl.edu/>). The WashU Virus Genome Browser maintains the functionality of the Epigenome Browser while also incorporating new features particularly useful for probing virus genomics data.

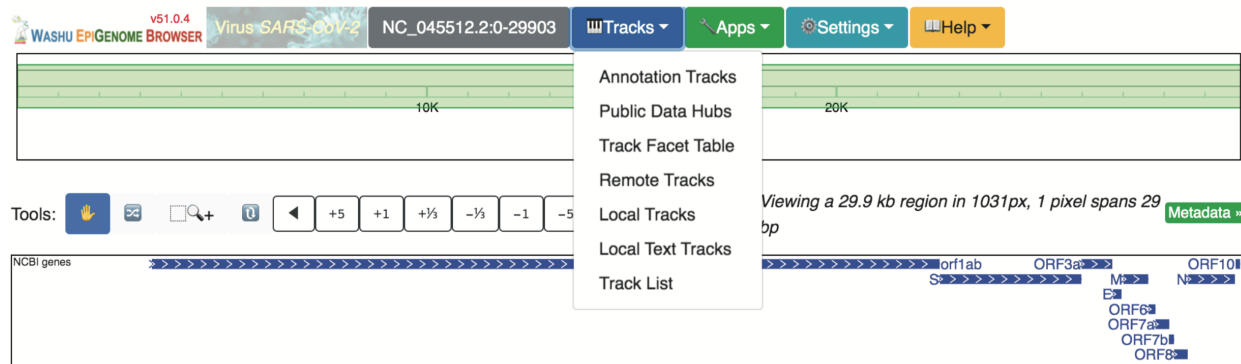


To the right of the WashU Virus Genome Browser logo is a utility bar (circled above in red), similar to the one provided in the WashU Epigenome Browser. From left to right, this bar lists the reference species the user is currently viewing, provides a platform for navigating the genome to particular regions of interest, provides a drop-down menu for selecting data tracks to be viewed on the browser, provides several applications for viewing, sharing, or saving region views and tracks of interest, and provides several customizable options in “Settings”, such as the ability to highlight a region of interest and change the track legend width, among others. Detailed information regarding the use of the region locator can be found within the Epigenome Browser tutorial (<https://eg.readthedocs.io/en/latest/usage.html#the-top-menu>).


Details regarding the default tracks loaded into view for SARS-CoV-2, which include a sequence diversity track, a mutation frequency track, RNA expression, putative SARS immune epitopes, transcription regulatory sequence locations, recombinations, RNA modifications, and SNV tracks of any strains in the user’s cart, are discussed further below.

1.2.1 Tracks

The “Tracks” tab lists several functionalities, from loading existing tracks hosted by the browser to user-specified uploads.



From this drop-down menu, if the user selects “Annotation Tracks”, an expandable menu of pre-existing tracks appears for the user to select from. For SARS-CoV-2, these tracks include gene annotations, genome comparison tracks, a GC density track, as well as a sequence diversity track and a mutation alert track, all of which can be loaded into the current browser session. Of particular clinical importance, the “Sequence diversity (Shannon Entropy)” track comprehensively displays the divergence across all GISAID strains at each genomic position. Additionally, the “Mutation Alert” track displays the number of strains with a mutation at each position. Together, these tracks provide an overview of accumulating mutations and their abundance, aiding in monitoring diagnostic primers for expected effectiveness as the virus evolves. Because of this utility, both the “Sequence diversity (Shannon Entropy)” track and the “Mutation Alert” track are loaded into view by default.

- 
- ▼ SARS-CoV-2
 - ▼ Ruler
 - Ruler (Added)
 - ▼ Genes
 - NCBI genes (Added)
 - ▼ Assembly
 - GC Percentage Add
 - ▼ Diversity
 - Sequence diversity (Shannon Entropy) (Added)
 - Mutation Alert (Added)
 - ▼ Genome Comparison
 - MERS to SARS-CoV-2 alignment Add
 - SARS to SARS-CoV-2 alignment Add
 - pangolin CoV to SARS-CoV-2 alignment Add
 - bat CoV to SARS-CoV-2 alignment Add

The “Remote Tracks” selections allow for user-upload of individual tracks or data hubs from a hosted url. Upon selection, the user is prompted to select whether he/she would like to upload an individual track (default tab) or upload a data hub (right-hand tab). If uploading an individual track, the user can then select the drop-down arrow under “Track type” to view all tracks supported by the WashU Virus Browser and select the track type matching their data.

[Add Remote Track](#)[Add Remote Data Hub](#)

Add remote track

Track type [track format documentation](#)

- Numerical
 - ✓ bigWig - numerical data
 - bedGraph - numerical data, processed by tabix in .gz format
 - qBED - quantized numerical data, processed by tabix in .gz format
- Annotation
 - bed - annotation data, processed by tabix in .gz format
 - bigBed - annotation data
 - refBed - gene annotation data, processed by tabix in .gz format
- Categorical
 - categorical - categorical data, processed by tabix in .gz format
- Methylation
 - methyIC - methylation data, processed by tabix in .gz format
- Interaction
 - hic - long range interaction data in hic format
 - cool - long range interaction data in cool format, use data uuid instead of URL
 - bigInteract - long range interaction data in bigInteract format
 - longrange - long range interaction data in longrange format
- Repeats
 - repeatmasker - repeats annotation data in bigBed format
- Alignment
 - bam - reads alignment data
 - pairwise - pairwise alignment data
- 3D Structure
 - g3d - 3D structure in .g3d format
- Dynamic
 - dbedgraph - Dynamic bedgraph data

A comprehensive list and description of these tracks can be found here: <https://eg.readthedocs.io/en/latest/tracks.html>.

To upload a track (in text format, such as a .bed.txt, .bedgraph.txt, or .longrange.txt, etc.) directly from one's compute, the user would select "Local Text Tracks". Upon selection, the user can select from the drop-down menu the file type that matches their data. In the default view of this pop-up window, the text file format is "bed" and an example of the text file format is below. Optionally, the user can configure track options, such as metadata, track height, track color, etc. by filling in the text box below.

You can upload track data in text file without formatting them to the binary format. Check more at [text tracks](#).



1. Choose text file type

bed

text file in BED format, each column is separated by tab

Example:

```
chr1 13041 13106 reg1 1 +
chr1 753329 753698 reg2 2 +
chr1 753809 753866 reg3 3 +
chr1 754018 754252 reg4 4 +
chr1 754361 754414 reg5 5 +
chr1 754431 754492 reg6 6 +
chr1 755462 755550 reg7 7 +
chr1 761040 761094 reg8 8 +
chr1 787470 787560 reg9 9 +
chr1 791123 791197 reg10 10 +
```

(Optional) Configure track options below in JSON format: [Example](#) [available properties for tracks](#)

1

Use a Worker thread: ☐ [\(Check if your file is huge.\)](#)

2. Choose text files:

Choose Files No file chosen

if you choose more than one file, make sure they are of same type.

To upload a track (in binary format, such as .bigwig, .hic, .g3d, .bedgraoh.gz(.tbi), .bed.gz, ect.) directly from one's computer, the user would select "Local Tacks". Here, the user can choose whether to upload a track (default tab) or data hub (right-handed tab) from his/her computer directly. After selecting "Add Local Tracks", the user can then select the track type matching their file via the drop-down menu shown above. Users can optionally specify in the accompanying text box display preferences for their added track, as demonstrated below by selecting "Example".

Add Local Track

Add Local Hub



1. Choose track file type:

bigWig

(Optional) Configure track options below in JSON format: [Example](#) [available properties for tracks](#)

1 {"height": 100, "color": "red"}

2. Choose track file:

Choose Files No file chosen

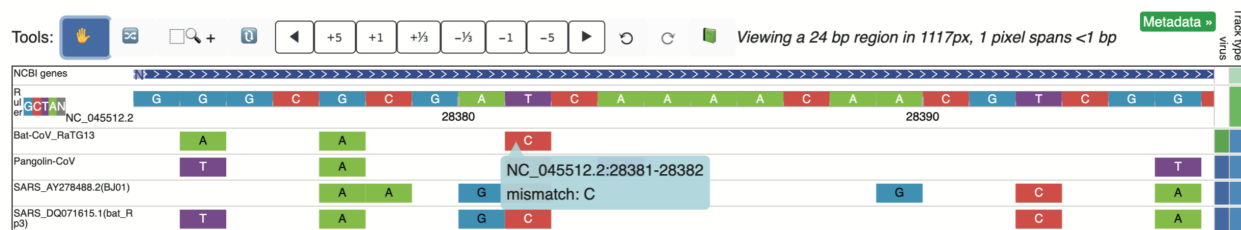
All tracks loaded onto the browser can be easily managed by selecting “Track List”, as demonstrated below.

Displayed tracks



Label	Track type	Re...
NCBI genes	geneannotation	
Ruler	ruler	
Sequence diversity (Shannon Entropy)	bedgraph	
Mutation Alert	pairwise	
Viral RNA expression (nanopore)	bigwig	
Transcription regulatory sequences (TRSs)	categorical	
TRS-L-dependent recombination	longrange	
Viral RNA Modifications	dbedgraph	
EPI_ISL_406798 pairwise alignment	pairwise	
EPI_ISL_416397 pairwise alignment	pairwise	
EPI_ISL_402120 pairwise alignment	pairwise	
Previous	Page 1 of 1	20 rows Next

In addition to pre-existing track types hosted on the WashU Epigenome Browser, the WashU Virus Browser also introduces the user to a new additional track type, called a “SNV” track, which displays alignment results in “pairwise” format. The SNV track type displays any genomic variations the strain has from the chosen reference (3 such tracks are pictured below). If the variant is a mismatch, the track will display the deviated nucleotide following the same color scheme as the reference. If a user selects a particular sequence variation, a pop-up window will show the details of the variation. A very detailed description and tutorial regarding the SNV tracks are in the section “SNV Track” below.



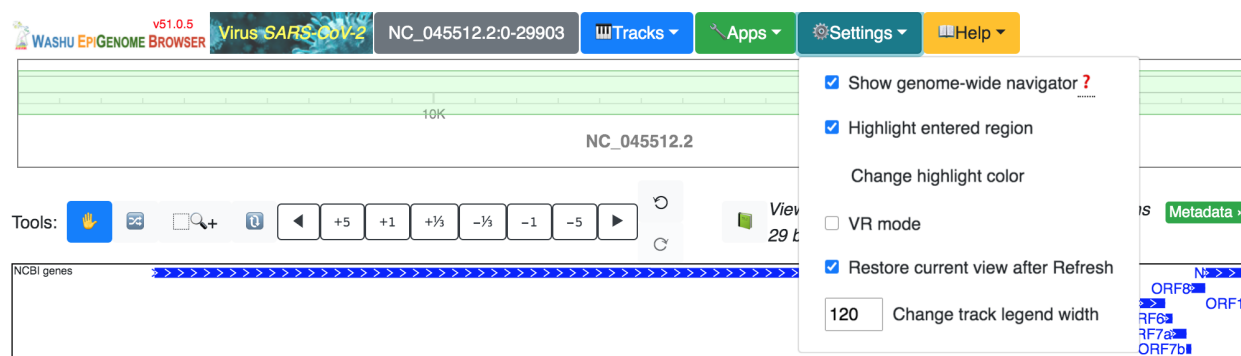
In addition to annotation tracks and user-uploaded tracks, our browser hosts genomics data of various types organized according to studies, pre-prints, or data providers, in a effort to rapidly integrate data. Detailed information pertaining to navigating existing data hubs is provided in the section “Public Data Hubs” below.

1.2.2 Apps

All applications available on the WashU Virus Genome Browser are also available on the WashU Epigenome Browser, and are described in detail in the Epigenome Browser tutorial (<https://eg.readthedocs.io/en/latest/usage.html#apps>). Of the available apps, “Region Set View” allows the user to visualize several distant genomic regions in the same viewing window. Selecting “Session” allows the user to save their current browser status, generating a session ID that can be shared with collaborators and allowing for easily resuming at a later time. “Fetch Sequence” allows the user to quickly obtain the reference sequence spanning the current view in a fasta format. Selecting “Screenshot” allows the user to generate publication-quality SVGs or PDFs of the current frame of view, with the option to highlight genomic regions of interest. Additional apps include “Gene Plot”, “Scatter Plot”, and “Go Live”, which are explained in detail in the Epigenome Browser tutorial.

1.2.3 Settings

Several browser settings have customizable options which the user may define. When selecting the “Settings” drop-down menu, several functions are provided as shown below (and as described in the Epigenome Browser tutorial: (<https://eg.readthedocs.io/en/latest/usage.html#settings>))

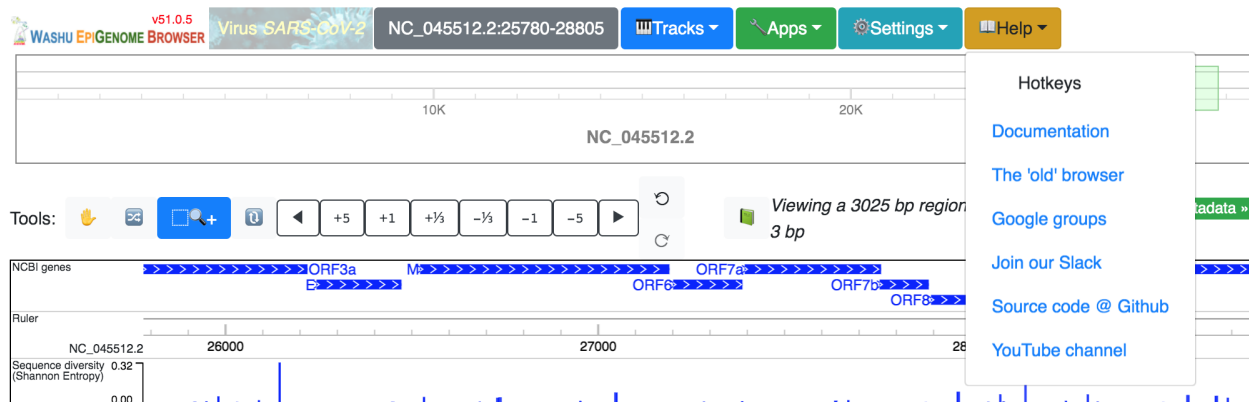


The first option “Show genome-wide navigator” is a toggle option, allowing the user to hid or show the complete genome layout at the top of the browser (circled in red below), highlighting in green the section fo the genome currently being viewed in the browser.



1.2.4 Help

The “Help” drop-down menu provides several links to browser resources:



These resources include a list of keyboard shortcuts for commonly used tools:

- **Alt + H** or **Alt + D** : Drag tool
- **Alt + S** or **Alt + R** : Reorder/Swap Tool
- **Alt + M** : Magnify Tool
- **Alt + Z** and **Alt + X** : Pan one full panel left or right.
- **Alt + I** and **Alt + O** : Zoom In and Out 1 fold.
- **Alt + G** : Toogle the re-order many tracks interface.

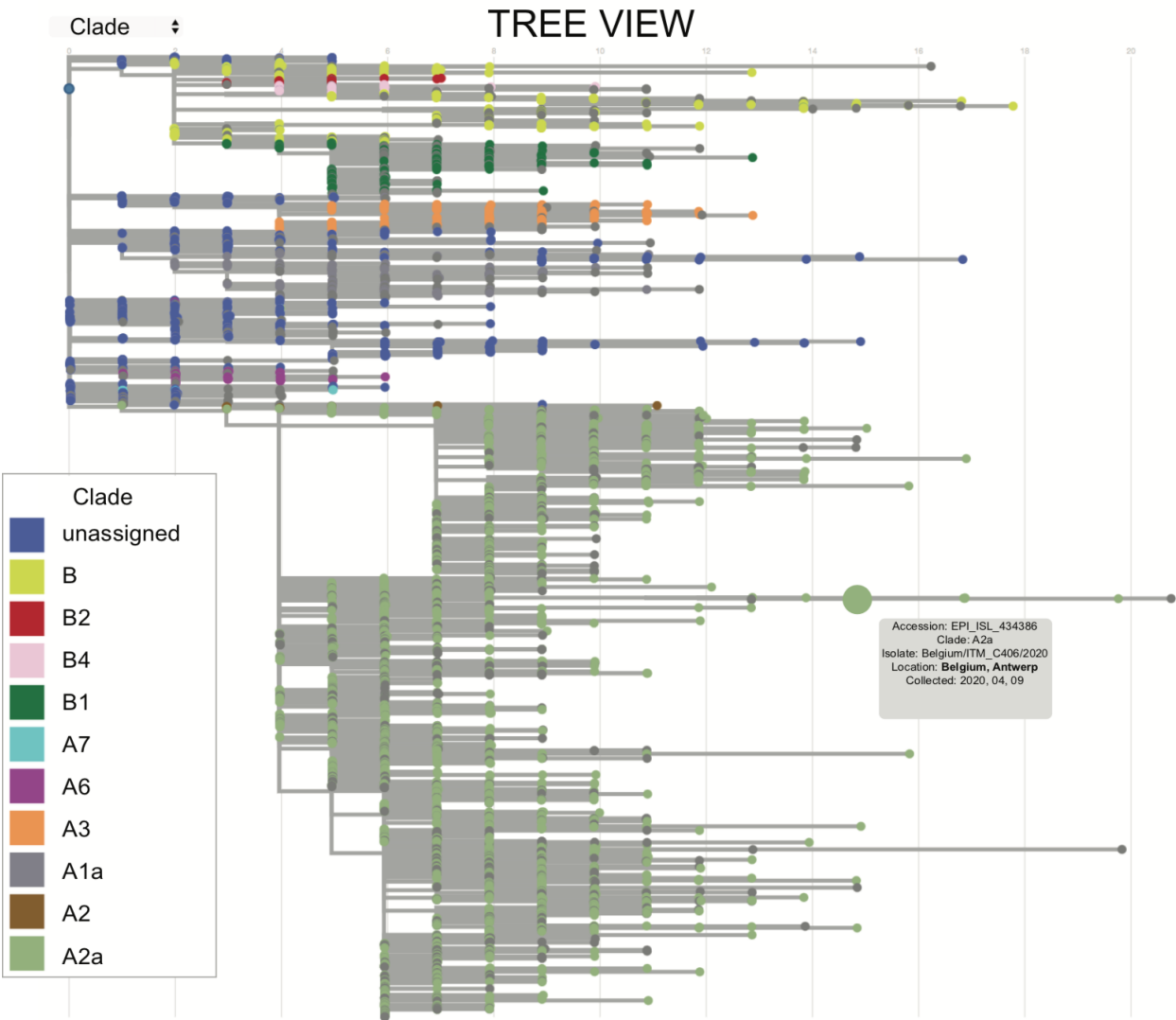
Additional links include one to our documentation page, with helpful tutorials, examples, and suggestions for customizing your browsing experience, a link to the original or ‘old’ Epigenome Browser, a google groups page populated with questions and answers, a link to our Slack page, where users can directly communicate with the WashU Virus Genome Browser team in real time, a link to our Github page, which houses our repository of all available scripts, and a link to our YouTube channel, where walk-through clips can be viewed.

1.3 Phylogenetic Tree View

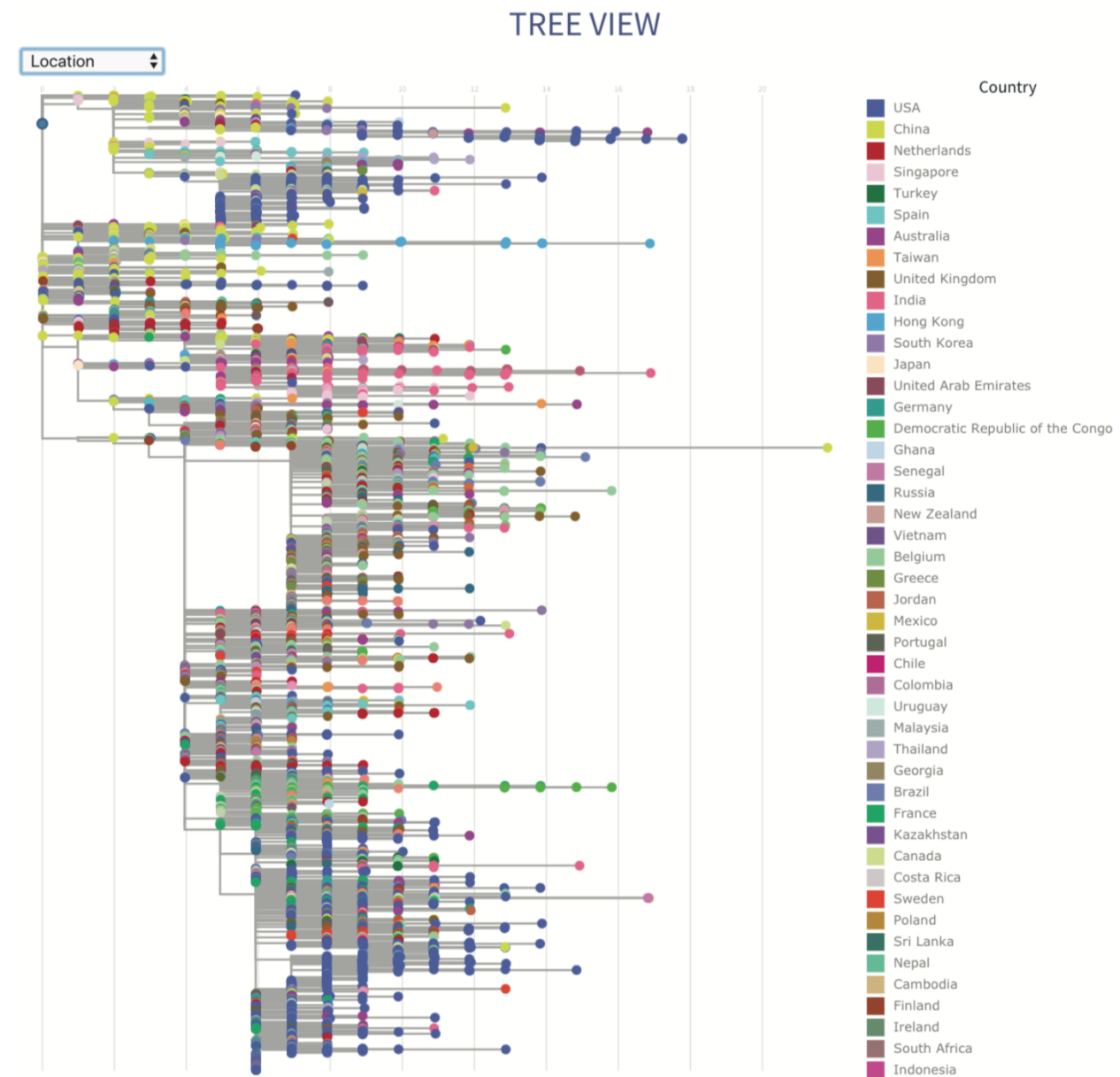
When users select the “Tree View” (blue arrow in second figure shown above), they are taken to our tree visual representation of the strains from the selected reference species. If the selected reference is SARS-CoV-2, the browser will load in a phylogenetic tree parsed from Nextstrain (http://data.Nextstrain.org/ncov_global.json), and therefore includes only strains available from Nextstrain. If the user selects any of the remaining three viruses (SARS, MERS, or Ebola), they are directed to an approximately-maximum-likelihood tree consisting of all available strains hosted by NCBI (<https://www.ncbi.nlm.nih.gov/nuccore>), built using FastTree with the GTR substitution model.

1.3.1 Coloring Tree by Metadata

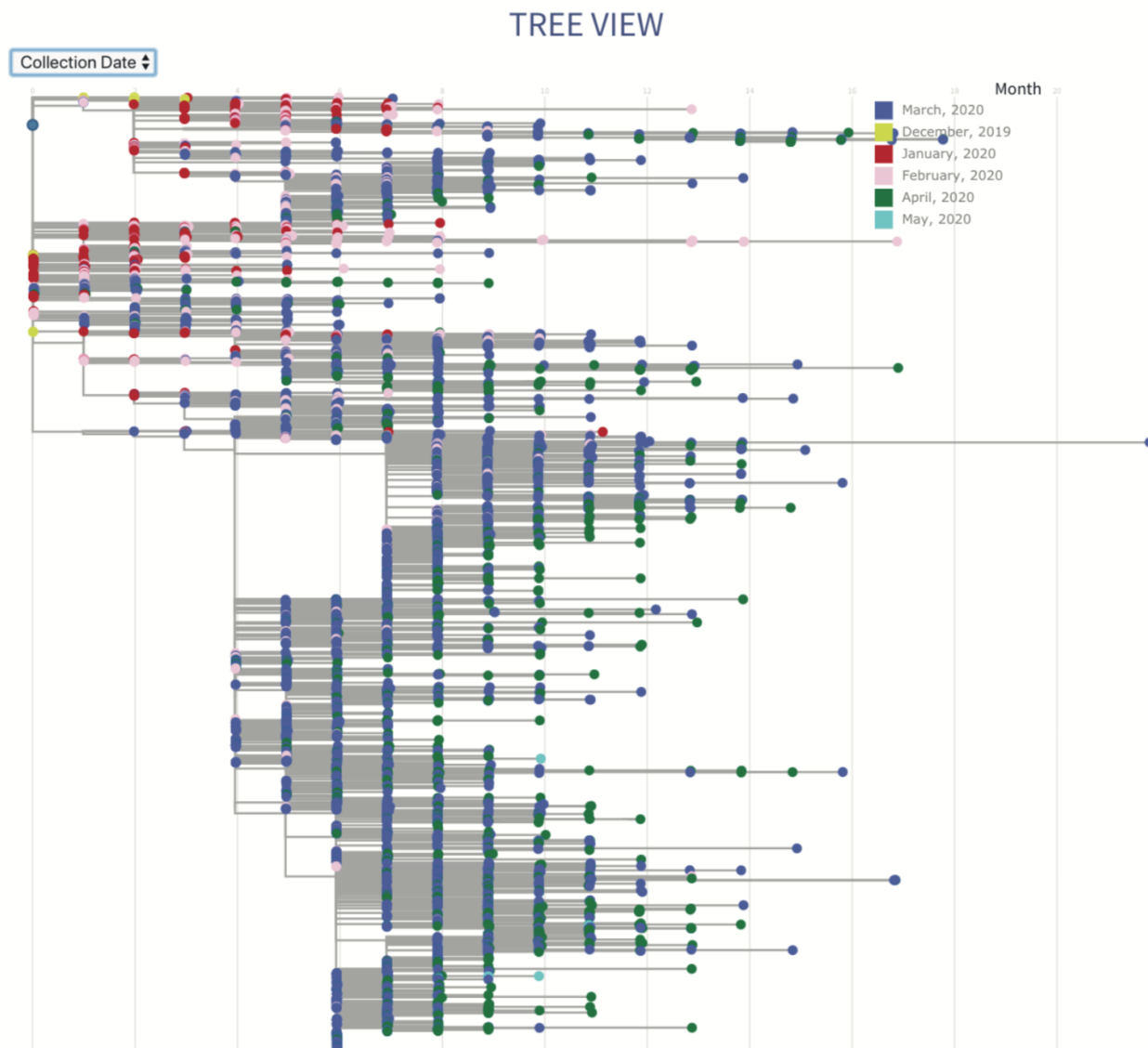
In any of the four trees available, the user is able to organize the tree by color by selecting either “Clade” (as is the default):



By “Location”:



Or by “Collection Date”:



1.3.2 Identifying Strains of Interest

If users pre-select strains and add them to their cart from the data table, they can see where their strains of interest fall within the tree (please keep in mind that for SARS-CoV-2, only strains housed in Nextstrain will be available for viewing in the tree view). As an example below, the following tracks have been added to cart:

Data

EPI_ISL_422624

SARS-CoV-2

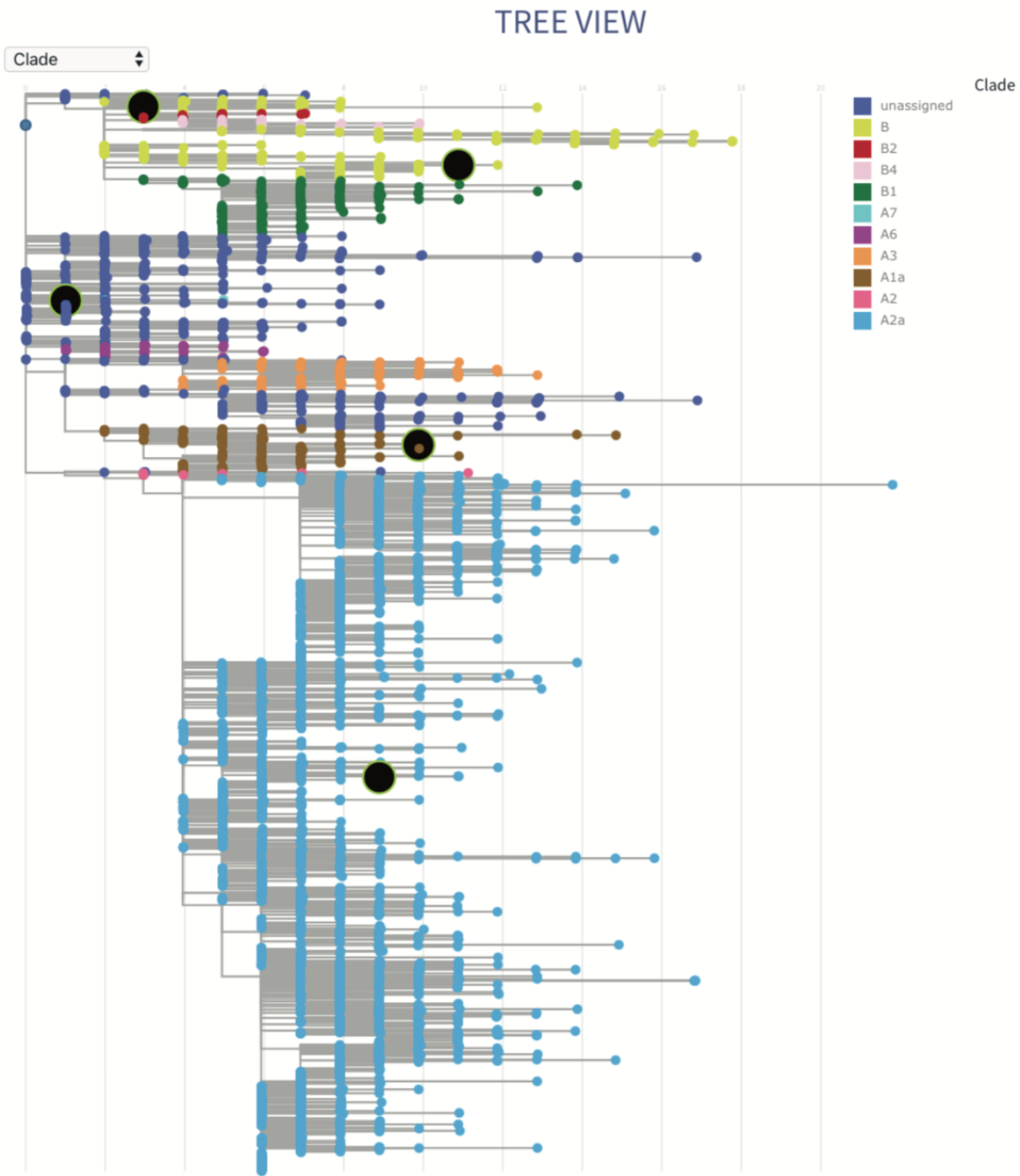


EPI_ISL_417498

SARS-CoV-2



Upon loading in the tree view, all selected strains are enlarged and colored black, as shown below.

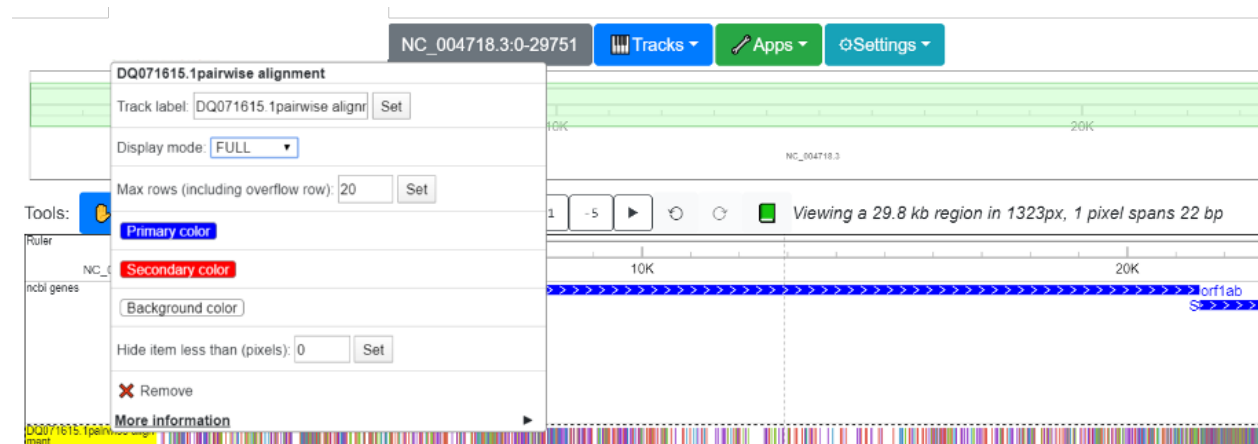


SNV track

This section shows you how to display variations with using “SNV” tracks.

2.1 Understanding the SNV track

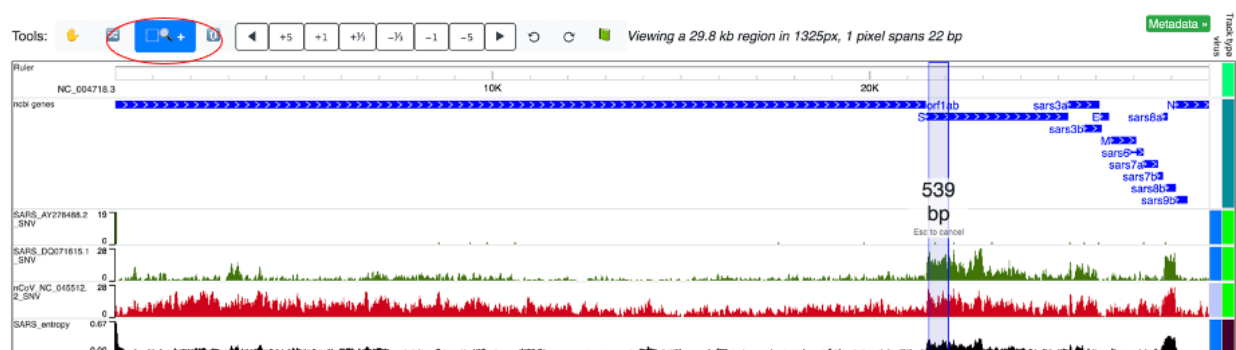
The SNV track is a new track type added to the WashU Virus Genome Browser to view sequence variations from the reference. The track supports 2 display modes: “density” mode for a “zoomed out” view and “full” mode for a “zoomed in” view. The “density” mode displays the density of variation, suitable for a genomic view, whereas the “full” mode has a color code for the detailed information of each variation, suitable for viewing an individual locus. To switch between density mode and full mode, right click on the head of the track and use the “display mode” drop-down menu.



2.2 Density mode and the “zoomed-out” view

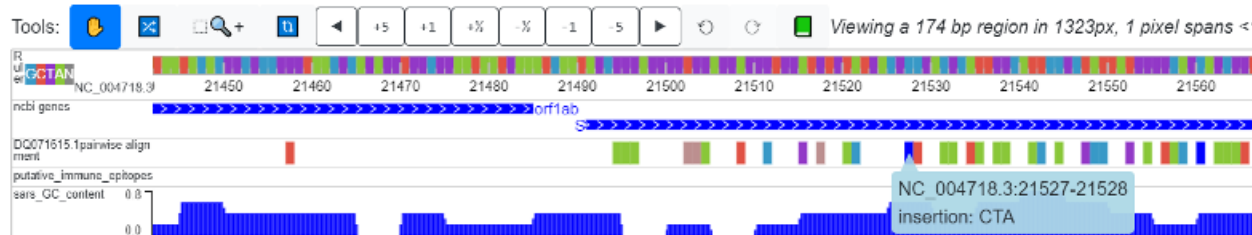
The density mode was implemented because when viewing the entire genome, individual variations are impossible to see. Instead, the density mode depicts the frequency of mutations across the genome, averaging over genomic intervals, as illustrated in the screenshot below. 2 SARS strains (AY278488.2 (“BJ01”) and DQ71615.1 (“bat rp3”)) and one SARS-CoV-2 strain (the reference strain, NC_045512.2) are aligned to the SARS reference genome. Sequence variation displayed in density mode shows that the divergence between the SARS-CoV-2 reference genome (red, below) and the SARS reference genome is higher than the divergence between the two additional SARS strains (green, below) and the SARS reference genome.

For AY278488.2, the variation from reference is mainly confined to the beginning of the genome, while the remainder of the genome is relatively consistent with the reference. However, for DQ071615.1 (bat-derived), the 5’ end of gene S displays high variation from the reference genome. Likewise, the SARS Shannon track shows that the SARS genome is highly diverse across different strains across gene S. Once a region of interest is identified, the standard magnification tool (circled in red) of the browser can be used to quickly zoom into the region.



2.3 Full mode and the “zoomed-in” view

When zoomed in to the nucleotide-level and displayed in “Full” mode, a color-coded track indicating all the variation from reference will be shown. The “Full” mode is further detailed below.



2.3.1 Nomenclature:

Reference: the “reference” is the sequence corresponding to the viral species selected by the user. It is the one completely color coded according to nucleotides and is shown as the “ruler” at the top of the browser view. The references hosted on the browser for the 4 virus species (SARS-CoV-2, SARS, MERS, and Ebola) are NCBI reference sequences.

Query: The “query” is the sequence being aligned to the “reference”.

Variation refers to events where the nucleotide of the query at a certain position is different from the reference. It can be a mismatch, insertion or deletion.

2.3.2 Color code:

Mismatches:

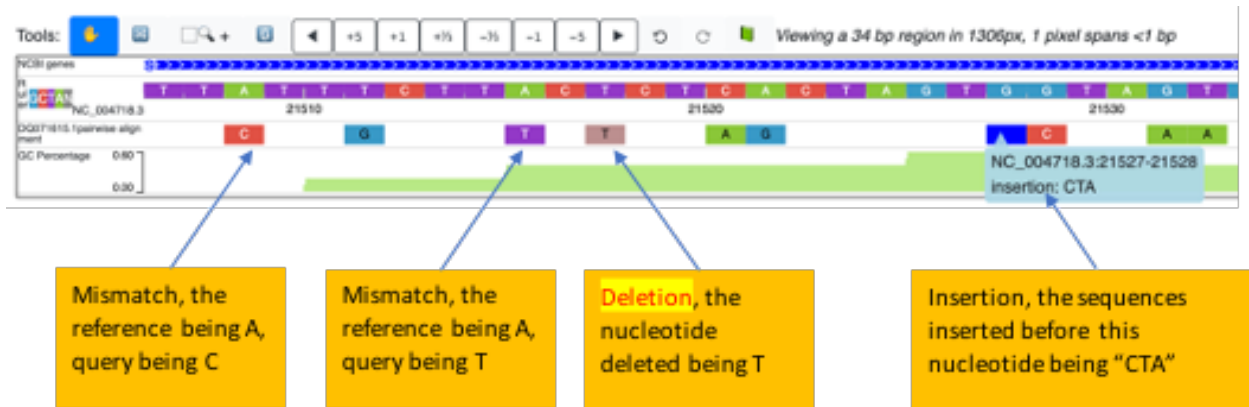
1. A mismatch from the reference observed at a specific nucleotide, the query being “A”: green (#89C738)
2. the query being “T”: purple (#9238C7)
3. the query being “C”: orange (#E05144)
4. the query being “G”: light blue (#3899C7)
5. the query being “N”: grey52 (#858585)

Deletions:

1. If a deletion is present in the query (a gap for the query in the pairwise alignment), the nucleotide will be colored “Rosy brown (#BC8F8F)”

Insertions:

1. The reference on the browser is always ungapped. If an insertion into the query happens (which signifies a gap for the reference in a pairwise alignment), the nucleotide proceeding the insertion will be colored “blue”. Details of inserted sequences will be revealed if you click on the nucleotide colored blue.

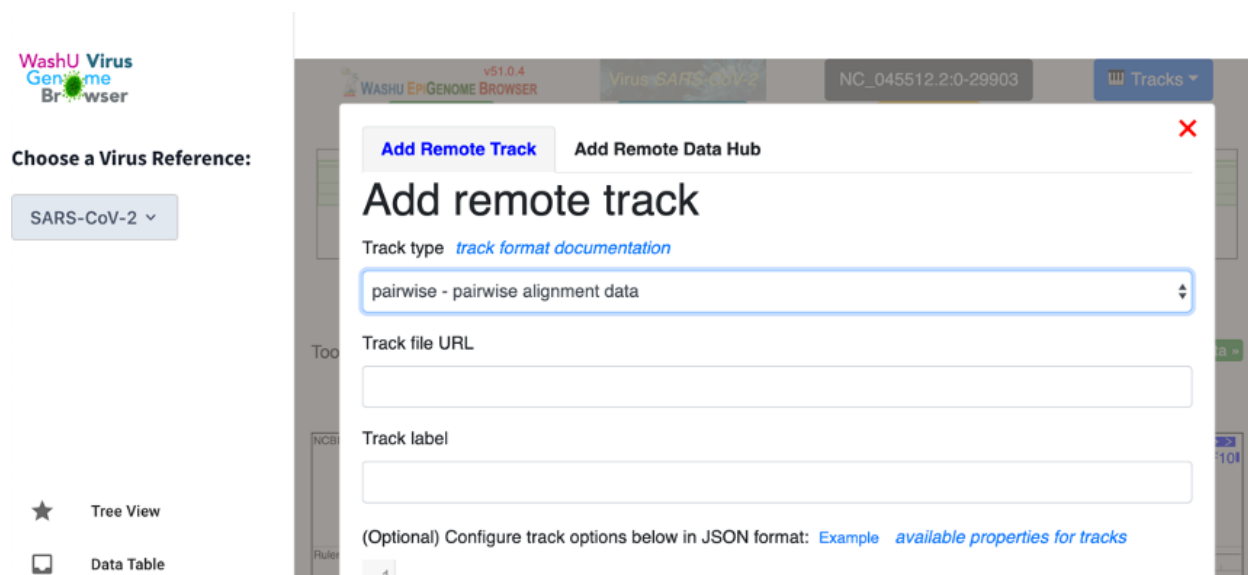


2.4 Behind the SNV track: the “pairwise” format

Alignment results can differ significantly when different aligners and different parameters are used. However, most aligners return alignments in the same format: FASTA (or markx3 in the case of EMBL aligners). Therefore, the WashU Virus Genome Browser offers scripts for the user to convert their own alignment results into “pairwise”-formatted files, which can be directly uploaded and displayed on the browser. Scripts can be found on our GitHub page: <https://github.com/debugpoint136/WashU-Virus-Genome-Browser>.

2.4.1 Uploading interface

In the browser view, click “Tracks” and then “Remote Tracks”. This will lead the user to a upload interface. Once there, select “pairwise” as the track type and enter the track’s URL.



Alternatively, if the track is stored on the user's local computer, he/she can upload the track by selecting "Tracks" > "Local Tracks". See the "Tracks" section under "Navigating the WashU Virus Genome Browser" above for more details.

2.4.2 The pairwise format

The pairwise format is an extension of the .bed format, where the 4th column contains variations from the reference. Specifically:

Column 1: Name of the reference that the query is aligned to

Column 2: The start position on the reference

Column 3: The end position on the reference

Column 4: Variation type and details

For columns 2 and 3, since the pairwise format is per-nucleotide, "end" usually equals "start" + 1 (the only exception is a deletion, and when several consecutive nucleotides are deleted, they can be merged into 1 line). Also, the browser is 0-based, which means that if there is a mismatch at the first nucleotide, column 2 would be "0" while column 3 would be "1".

For column 4, the format is "variation_type:detail". Variation types are: "insertion", "deletion", and "mismatch". For "insertion", "detail" is the sequence inserted before this nucleotide. For "deletion", the "detail" is the nucleotide of the reference at this position that was deleted. For "mismatch", "detail" is the nucleotide of the query.

Note: Matches between the query and the reference are not coded in the pairwise format.

NC_004718.3	96	97	mismatch: T
NC_004718.3	140	141	mismatch: C
NC_004718.3	142	143	mismatch: G
NC_004718.3	258	260	mismatch: A
NC_004718.3	3089	3090	insertion: GG
NC_004718.3	3093	3094	insertion: CTCA
NC_004718.3	21527	21528	insertion: CTA

(continues on next page)

(continued from previous page)

NC_004718.3	21560	21561	insertion: C
NC_004718.3	3059	3062	deletion: AGA
NC_004718.3	3223	3224	deletion: A

Note: The files need to be sorted, zipped, and indexed before uploading, just as you would need to do for any .bed file. To zip the files:

2.4.3 To generate the pairwise format

We offer a script (“publicConvertMarkx3.py”) to generate a pairwise-formatted file from any pairwise alignment result in markx3 or FASTA format. The requirements are as follows:

1. The reference should be the first sequence.
2. Only 2 sequences should be present in the file (as is the case for pairwise alignment).

An example FASTA-formatted pairwise alignment result is shown below:

```
>test_reference ..
ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG-----A
>test_query ..
CTG--TCTC-CTG---CCGACAAGACCAACGTCAAGGCCGCTGGGGTAAGA
```

The script used to convert this to our pairwise format is: “publicConvertMarkx3.py”, which calls “convert_tsv_to_bed_and_cat.sh” (both located on our GitHub page: <https://github.com/debugpoint136/WashU-Virus-Genome-Browser>). To see help, use:

```
$ python publicConvertMarkx3.py
usage: python publicConvertMarkx3.py <markx3> <out_pairwise> <script_dir>

[Required]
    <markx3>                                pairwise alignment output in_
↪markx3 format. the first sequence should be the reference

    <out_pairwise>                            output pairwise formatted file that can_
↪be displayed directly on wash u virus browser as SNV track

    <script_dir>                               the directory where all our_
↪scripts are stored

contact: changxu.fan@gmail.com for help
```

The output files will be automatically zipped and ready to go!!

2.4.4 Batch alignment from FASTA to pairwise format

We offer another script (“publiAlignment.py”) that can perform pairwise alignments in batch (using EMBL aligners “stretcher” or “water”) and directly generate files in pairwise format that can be directly uploaded as SNV tracks.

The script is: publicAlignment.py and is located here: <https://github.com/debugpoint136/WashU-Virus-Genome-Browser/blob/master/scripts/publicAlignment.py>

```
$ python publicAlignment.py
Batch pairwise sequence alignment using "stretcher" or "water".
outputs "pairwise" format files that can be directly displayed on the wash u viral_
↳ browser as SNV tracks.

[Required (for job submission)]
  --script_dir                the directory where all our scripts are_
↳ stored
  --ref_fa                    fasta file containing reference sequence._
↳ All other sequences will be aligned to it. Should contain only one sequence
  --strain_fa                 fasta file containing sequences of_
↳ individual strains. Can contain multiple sequences. They will be aligned to ref_fa_
↳ in a pair-wise manner separately
  --tempt_dir                 tempt_dir to store intermediate files
  --SNV_dir                   the directory to store generated pairwise_
↳ files
  --aligner                   aligner to use. currently support
↳ "stretcher" for global alignment and "water" for local alignment
  --email                     required by the embo aligners. you will_
↳ not receive junk from them
contact: changxu.fan@gmail.com for help
```

2.4.5 Batch upload as json files

We offer another script ("publicJsonGen.py", located here: <https://github.com/debugpoint136/WashU-Virus-Genome-Browser/blob/master/scripts/publicJsonGen.py>) that takes in a tab-delimited text file (.tsv file) listing the web location and track type of individual files, and outputs a .json file that can be used to upload multiple tracks in batch.

```
$ python publicJsonGen.py
usage: python publicJsonGen.py <tsv> <json>

[Required]
  <tsv>                        a tsv file with 4 columns: name, url,_
↳ track_type, virus. one line per track
                                the file should contain header.
                                order of the columns doesn't matter.
                                virus means virus type, used for metadata.
  <json>                       output json file that can be directly_
↳ uploaded onto wash u virus browser as custom datahub
contact: changxu.fan@gmail.com for help
```

The .tsv file should have a format similar to that shown below in order for a successful conversion:

name	url	track_type	virus
SARS_AY278488.2_SNV	https://your.url.to.file1	pairwise	SARS
SARS_DQ071615.1_SNV	https://your.url.to.file2	pairwise	SARS
SARS_AY278488.2_SNV	https://your.url.to.file3	pairwise	SARS

2.4.6 Upload json-formatted datahub

To upload a json-formatted data hub, in the browser view, select "Tracks" and then select either "Remote Tracks" or "Local Tracks" (depending on whether the .json file is stored remotely to locally, see above documentation under

“Navigating the WashU Virus Genome Browser” for further details).

WashU Virus
Genome
Browser

Choose a Virus Reference:

SARS-CoV-2 ▾

★ Tree View

📄 Data Table

🛒 0 files

WASHU EPIGENOME BROWSER v51.0.4 Virus-SARS-CoV-2 NC_045512.2:0-29903 Tracks ▾

Add Remote Track Add Remote Data Hub

Add remote data hub

Remote hub URL [data hub documentation](#)

Load from URL

Or

Choose datahub file

Row: Track type ↔ Column: Not used ▾

Track type	5/6
------------	-----

3.1 Functionality of SNV2 tracks

SNV2 tracks are an extension of SNV tracks. It's equipped with the ability to show amino acid level mutations. It also supports everything that SNV tracks support: color code mutations, zoomed-in and zoomed-out views, revealing detailed info upon mouse click, etc. SNV2 format is specified with extreme flexibility.

Essentially, the SNV2 format is a combination of categorical tracks (color coding) and bed tracks (text display). You can imagine it as “color coded bed tracks”. This enables it to be reused for many other purposes. We provide scripts to generate SNV2 tracks reporting the potential amino acid mutations, but we encourage you to customize it with your own scripts.



3.2 Defining the format of SNV2

SNV2 tracks can have as many columns as necessary. The first 3 columns encode the genomic positions. The 4th column encodes category. All columns starting from the 4th column will be shown as text upon mouse click. The mapping between categories and colors is specified in Json.

col- umn	detail
1	chromosome name for epigenome browser, reference name for virus genome browser (such as NC_045512.2 for SARS-CoV-2)
2	start position on the reference genome, 0 based, inclusive
3	stop position on the reference genome, 0 based, not inclusive
4	category. Controls color code, will also show as text in popup window upon clicking
5, 6, 7...	text columns. will show as text in popup window upon clicking

The 4th column is mapped to colors through specifying “segmentColors” in the “options” part of the json of datahubs. The detailed tutorial on how to use json is here: <https://epigenomegateway.readthedocs.io/en/latest/datahub.html?highlight=json>

However, if you are using the SNV2 track to show AA mutations, you don’t need to upload as json, because we have default color code mapping:

```
"options": {
  "segmentColors": {
    "un_sequenced": "Linen",
    "noncoding_insertion": "LightGrey",
    "noncoding_deletion": "LightGrey",
    "noncoding_mismatch": "LightGrey",
```

(continues on next page)

(continued from previous page)

```

    "silent": "DimGrey",
    "frameshift": "FireBrick",
    "missense": "CornflowerBlue",
    "AA_deletion": "CornflowerBlue",
    "AA_insertion": "CornflowerBlue",
    "N_mask": "Linen",
    "deletion_mask": "Linen"
  }
}

```

For a quick demo:

```
NC_045512.2    10000    10001    duck    cyberduck    cyberduck quit unexpectedly
```

zip it and index it using bgzip and tabix (<https://epigenomegateway.readthedocs.io/en/latest/tracks.html?highlight=tabix#prepare-track-files>). Then put it into a Json like this:

```

[ {
  "name": "duck",
  "type": "snv2",
  "url": "http://your.url.to.duck.file/duck.snv2.gz",
  "options": {
    "segmentColors": {
      "duck": "red"
    }
  }
} ]

```

upload the track through Tracks -> Remote Tracks -> Add Remote Data Hub You will see:



One of our snv2 tracks for SARS-CoV-2 is coded like this:

NC_045512.2	0	16	un_sequenced	un_sequenced	
NC_045512.2	240	241	noncoding_mismatch	mismatch: T	NC_045512.
↪2:240-241 ORF:noncoding C > T noncoding_mismatch					
NC_045512.2	3036	3037	silent mismatch: T	NC_045512.2:3034-3037	
↪ORF1a:F924 TTC > TTT F > F silent ; NC_045512.2:3034-3037 ORF1a:F924 TTC					
↪> TTT F > F silent					
NC_045512.2	14407	14408	missense	mismatch: T	NC_045512.2:14406-
↪14409 ORF1a:P4715 CCT > CTT P > L missense					
NC_045512.2	23402	23403	missense	mismatch: G	NC_045512.2:23401-
↪23404 S:D614 GAT > GGT D > G missense					
NC_045512.2	29872	29903	un_sequenced	un_sequenced	

3.3 Scripts for generating snv2 tracks

All of our premade SNV2 tracks that you can see in Tracks -> Public Data Hubs are generated through a set of scripts that can be found at <https://github.com/debugpoint136/WashU-Virus-Genome-Browser/tree/master/scripts/snv2/>

The main function `tsv2snv2.2()`, which we used to generate all snv2 files in the public data hubs, is in `snv2_public_7_2_20.R`, while the helper functions are in `snv2_helper_7_2_20.R`. We used `snv2_orf_7_2_20.R` to generate the tsv file with ORF information required in the main function.

The arguments of `tsv2snv2.2()`:

argument	detail
<code>tsv.vec</code>	a vector of tsv files generated by <code>publicAlignment.py</code> in <code>tempt_dir</code> (an argument for <code>publicAlign.py</code>).
<code>ref.fasta</code>	name of the fasta file for the reference. The one for SARS-CoV-2 is here
<code>ref.orf.table</code>	a dataframe or the name of a tsv file containing this dataframe. the ORF information for the reference. the one for SARS-CoV-2 is already generated. Refer to it for formatting.
<code>min.contig.head.tail</code>	integer used to mask the beginning and the end of the sequenced region, so that unsequenced regions won't be treated as deletions. default is 15, which means the first 15 continuous non-deletions marks the beginning of the sequenced region. The same applies for the end of the sequenced region
<code>out.snv2</code>	a vector of output snv2 file name
<code>hier.df</code>	a dataframe containing the 'hierarchy' of different types of mutations, can also be the name of a tsv file containing this dataframe. Sometimes one nucleotide can be used by more than one ORF. The mutation at this nucleotide might cause different types of amino acid mutations for different ORFs. For example, a mutation can be silent for ORF A, but missense for ORF B. In this case, 'missense' will override 'silent' because of the settings in <code>hier.df</code> . Refer to hier.df.6.18.tsv for formatting
<code>thread.number</code>	the number of snv2 files to generate in parallel
<code>thread.size</code>	the number of orfs to process in parallel
<code>bed-tools.path</code>	deprecated. Just pass it something
<code>return.df</code>	bool. if the entire snv2 track should be returned as a dataframe.

CHAPTER 4

Public Data Hubs

As individual research groups continue to rapidly study SARS-CoV-2, there has been a surge in available SARS-CoV-2 genomics data, ranging from thousands of sequenced strains, to host immune responses, and viral expression and modifications. As this data is becoming available, we are integrating relevant findings onto our browser in the form of public data hubs for efficient upload of several related tracks for easy visual comparisons. Here, we will introduce how to load a specific data hub of interest and also briefly introduce all existing public data hubs (as of May 24, 2020). Please note that we will continue to update this documentation as additional public data hubs are added.

4.1 Loading in a public data hub

From the main browser view, information regarding existing data hubs can be viewed by selecting “Tracks” > “Public Data Hubs”. This will populate a table with information regarding data hub collection names, hub names, and numbers of tracks.

Public data hubs



Collection	Hub name	Tracks	Add
▶ NCBI database	All NCBI SARS-CoV-2 isolates	Updating	+
▶ Nextstrain database	All Nextstrain SARS-CoV-2 isolates	Updating	+
▶ GISAID database	All GISAID SARS-CoV-2 isolates	Updating	+
▶ Diagnostics	Primers	Updating	+
▶ Diagnostics	CRISPR-based diagnostic tests	2	+
▶ Putative SARS-CoV-2 Immune Epitopes	SARS-CoV-2 Epitopes Predicted to Bind HLA Class 1 Proteins Databa...	1	+
▶ Putative SARS-CoV-2 Immune Epitopes	Congeneric (or Closely-related) Putative SARS Immune Epitopes Loc...	1	+
▶ Putative SARS-CoV-2 Immune Epitopes	Putative SARS-CoV-2 Epitopes	14	+
▶ Recombination events	Recombination events (Kim et al., 2020)	3	+
▶ Viral RNA modifications	Viral RNA modifications (Kim et al., 2020)	10	+
▶ Viral RNA expression	Viral RNA expression (Kim et al., 2020)	1	+
▶ Sequence variation	D614G prevalence across time	1	+
Previous	Page 1 of 1	20 rows	Next

No tracks from data hubs yet. Load a hub first.

Users can get more information regarding a data hub of interest, such as the data source and track descriptions, by selecting the collection, as demonstrated below.

Public data hubs



Collection	Hub name	Tracks	Add
<div></div>	<div></div>		
▼ NCBI database	All NCBI SARS-CoV-2 isolates	Updating	<div>+</div>

Collection details

SNV tracks of all SARS-CoV-2 strains on NCBI Genbank displaying their sequence variation from reference

Hub details

hub built by	Changxu Fan (fanc@wustl.edu)
hub info	All SARS-CoV-2 strains available on NCBI. Aligned to reference genome (NC_045512.2) using EMBL 'stretcher'.
data source	https://www.ncbi.nlm.nih.gov/nuccore
white space	Matching the reference
colored bars	Variation from the reference. Details are color coded. Zoom in to click on the bar to see detail
long stretches of rosy brown	Unsequenced regions

To load tracks from a given data hub onto the browser, the user can select the “+” button under the “Add” column. This will populate a sortable metadata table where the user can select specifically which tracks to add from the data hub. In the example below, data are being sorted by location, as highlighted in the purple box (although, please note that the majority of the locations are not pictured here, but are available on the browser).

Public data hubs

Collection	Hub name	Tracks	Add
▶ NCBI database	All NCBI SARS-CoV-2 isolates	Updating	✓
▶ Nextstrain database	All Nextstrain SARS-CoV-2 isolates	Updating	+
▶ GISAID database	All GISAID SARS-CoV-2 isolates	Updating	+
▶ Diagnostics	Primers	Updating	+
▶ Diagnostics	CRISPR-based diagnostic tests	2	+
▶ Putative SARS-CoV-2 Immune Epitopes	SARS-CoV-2 Epitopes Predicted to Bind HLA Class 1 Proteins Database	1	+
▶ Putative SARS-CoV-2 Immune Epitopes	Congeneric (or Closely-related) Putative SARS Immune Epitopes Locations (this publication)	1	+
▶ Putative SARS-CoV-2 Immune Epitopes	Putative SARS-CoV-2 Epitopes	14	+
▶ Recombination events	Recombination events (Kim et al., 2020)	3	+
▶ Viral RNA modifications	Viral RNA modifications (Kim et al., 2020)	10	+
▶ Viral RNA expression	Viral RNA expression (Kim et al., 2020)	1	+
▶ Sequence variation	D614G prevalence across time	1	+

Previous Page 1 of 1 20 rows Next

Row: location Column: Not used

location		
location		
Puerto Rico		0/13
Netherlands		0/1
Germany		0/1

After selecting a metadata term of interest, such as “Czech Republic” above, the user can add the tracks to their browser view by selecting the “+” to the right of the track(s), as shown in green below.

Track table

Search tracks

H1 or H3K4me3, etc...

Free text search over track labels and metadata.

Name	Data hub	location	URL	Format	Add
MT371568.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	✓
MT371569.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+
MT371570.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+
MT371571.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+
MT371572.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+
MT371573.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+
MT371574.1_SARS-CoV-2/human/CZE...	All NCBI SARS-CoV-2 isolates	Czech Republic	https://wangftp.wustl.ed...	pairwise	+

Data tracks from all existing data hubs or user-uploaded tracks can be managed in a similar way by selecting “Tracks” and then “Track Facet Table”.

4.2 Introducing currently available public data hubs (As of May 24, 2020)

Here, we introduce all available public data hubs as of May 24, 2020, organized by Collection. Please note that this selection will continue to expand as new studies are made available regarding SARS-CoV-2 and related host genomic data.

4.2.1 NCBI database

As of May 24, 2020, 3896 SARS-CoV-2 sequences are available from NCBI (<https://www.ncbi.nlm.nih.gov/nucleotide>). Pairwise alignments between all available strains and the reference have been added to the NCBI database public data hub and are available for viewing on the WashU Virus Genome Browser in the form of SNV tracks (detailed above in the “SNV” section). All strains can be added to the browser at once, or the user can sort the strains by location and collection date to pre-filter for specific strains of interest. Similarly, metadata information regarding location and collection date can be displayed on the right side of the tracks once loaded into the browser view, by selecting “Metadata” and specific terms of interest, as shown below.



4.2.2 Nextstrain database

As of May 24, 2020, 4415 SARS-CoV-2 sequences are available from Nextstrain (<https://data.nextstrain.org/ncov.json>). Pairwise alignments of all available strains to the reference have been added to the Nextstrain database public data hub and are available for viewing on the browser in the form of SNV tracks (detailed above in the “SNV” section). As described above, all strains can be added to the browser at once, or a subset can be selected for upload by pre-filtering the strains by specific metadata terms of interest.

4.2.3 GISAID database

As of May 24, 2020, 30612 SARS-CoV-2 sequences are available from GISAID. Pairwise alignments between all available strains and the reference sequence have been added to the GISAID database public data hub and are available for viewing on the browser in the form of SNV tracks (detailed above in the “SNV” section). As described above, all strains can be added to the browser at once (not recommended, as there are several thousand), or a subset can be selected for upload by pre-filtering the strains by specific metadata terms of interest.

4.2.4 Diagnostics

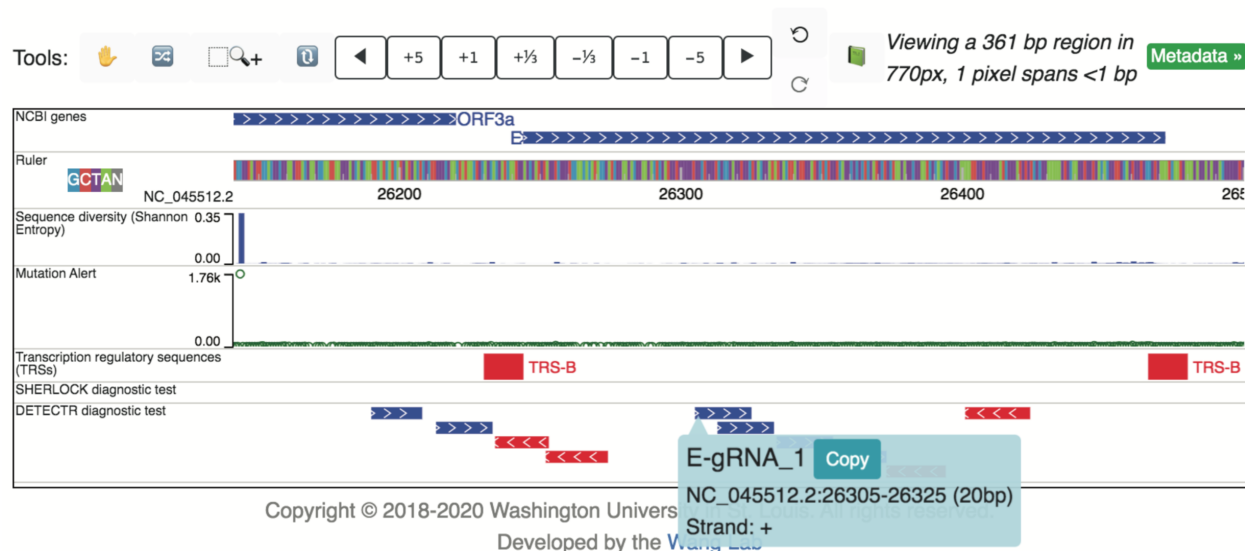
The Diagnostics Collection currently houses two separate data hubs, both of which are described below. All data hubs encompassed in this collection contain relevant annotations pertaining to diagnostic testing.

Primers

As of May 24, 2020, the locations of primers for SARS-CoV-2 testing have been made available for the USA (3 sets of CDC primers), China, Hong Kong, France, Germany, Japan, and Thailand. All primer sequence locations are made available in the Primers database public data hub, and can be loaded in all at once or based on country of interest.

CRISPR-based diagnostic tests

The CRISPR-based diagnostic tests data hub consists of two tracks: a SHERLOCK diagnostic test track displaying primer and guide RNA sequence locations used in the CRISPR-Cas13a-based SHERLOCK assay for detecting SARS-CoV-2 ([https://www.broadinstitute.org/files/publications/special/COVID-19%20detection%20\(updated\).pdf](https://www.broadinstitute.org/files/publications/special/COVID-19%20detection%20(updated).pdf)), and a DETECTR diagnostic test track displaying the primary and guide RNA sequence locations used in the CRISPR-Cas12-based DETECTR assay for detecting SARS-CoV-2 (PMID: 32300245). An example highlighting the location of a DETECTR E guide-RNA is shown below, revealing low sequence diversity at this location.

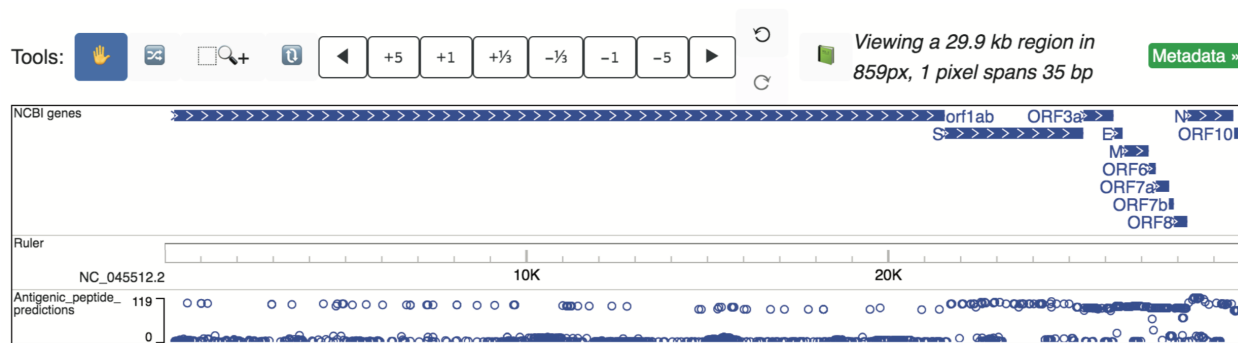


4.2.5 Putative SARS-CoV-2 Immune Epitopes

The collection “Putative SARS-CoV-2 Immune Epitopes” currently consists of three data hubs: “SARS-CoV-2 Epitopes Predicted to Bind HLA Class 1 Proteins”, “Congeneric (or Closely-related) Putative SARS Immune Epitopes”, and “Putative SARS-CoV-2 Epitopes”. All three data hubs feature the locations of immune epitopes, some of which were identified in SARS-CoV-1 and maintain sequence conservation in SARS-CoV-2.

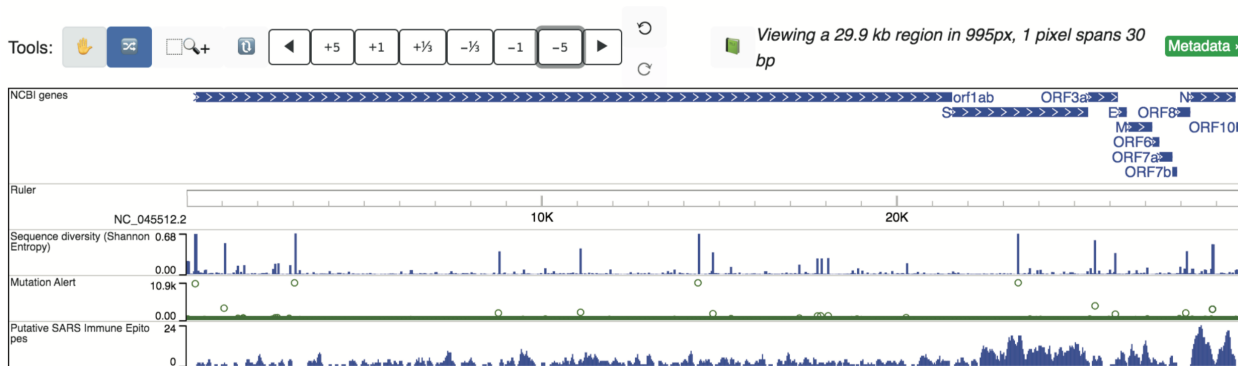
SARS-CoV-2 Epitopes Predicted to Bind HLA Class 1 Proteins

Predicted SARS-CoV-2 epitopes likely to bind class 1 MHC proteins were made available in the pre-print Campbell, et al., 2020 (DOI: 10.1101/2020.03.30.016931). Locations of the predicted sequences within the SARS-CoV-2 genome were identified (those with 100% sequence similarity and on the positive strand) and their locations (x-axis) as well as the number of unique strain IDs reporting the peptide (y-axis) are shown when loading in the track “Antigenic_peptide_predictions”, as shown below.

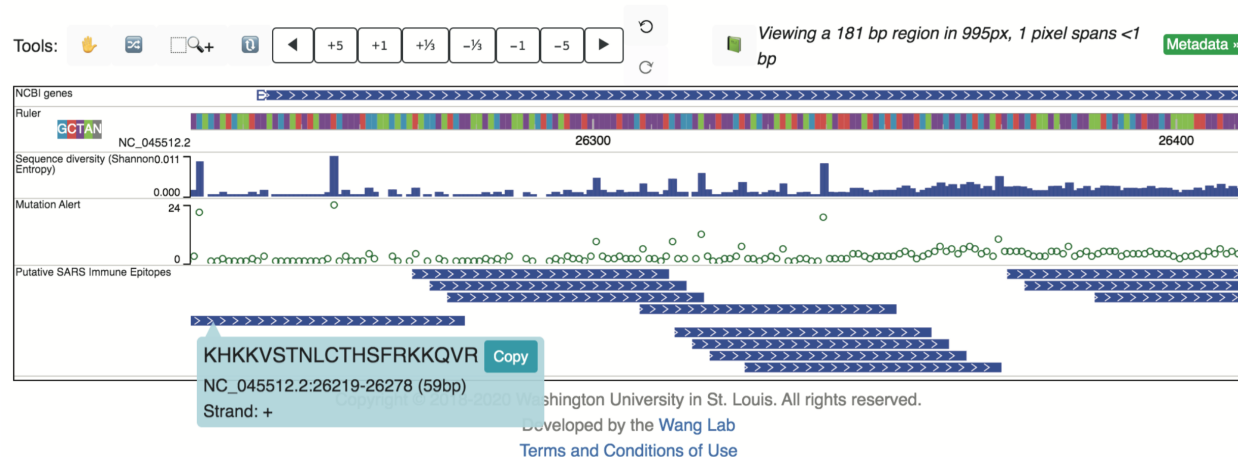


Congeneric (or Closely-related) Putative SARS Immune Epitopes

Linear immune epitopes identified in SARS-CoV-1 cataloged in the Immune Epitope Database and Analysis Resource (IEDB) that retain 100% sequence identity in SARS-CoV-2 are displayed in a single track, which can be set to either “Density” mode to view the abundance of epitopes over large portion of the genome:



Or can be set to “Full” mode to visualize individual epitopes, whose sequences are displayed upon selection:



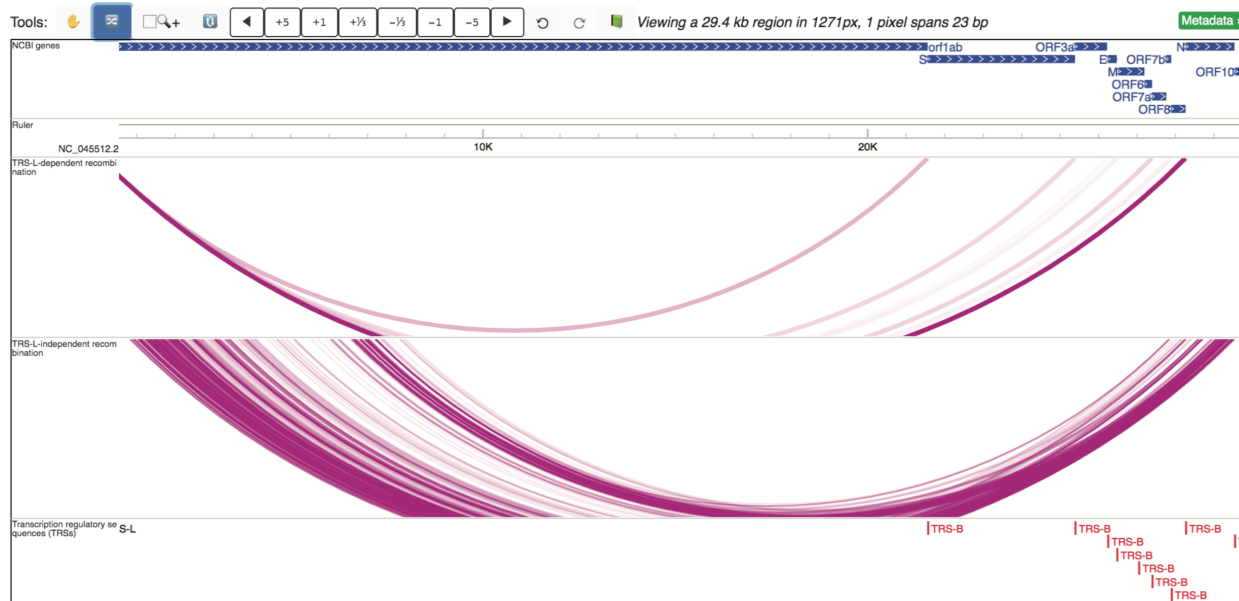
Putative SARS-CoV-2 Epitopes

This data hub hosts several (14) different tracks, pertaining to an assortment of different studies, and in-

cludes tracks displaying CD8 epitopes restricted to HLA-A*02:01 (DOI: 10.1101/2020.03.23.004176), B cell immune epitope predictions (DOI: 10.1101/2020.02.12.946087), CD4 T-cell immune epitope predictions (DOI: 10.1101/2020.02.12.946087), CD8 T-cell immune epitope predictions (DOI: 10.1101/2020.02.12.946087), putative epitopes for CD8+ T cells with widespread HLA binding properties (DOI: 10.1101/2020.04.06.027805), and N-terminal SARS-CoV-2 putative MHC-II epitopes (DOI: 10.1101/2020.04.17.20061440).

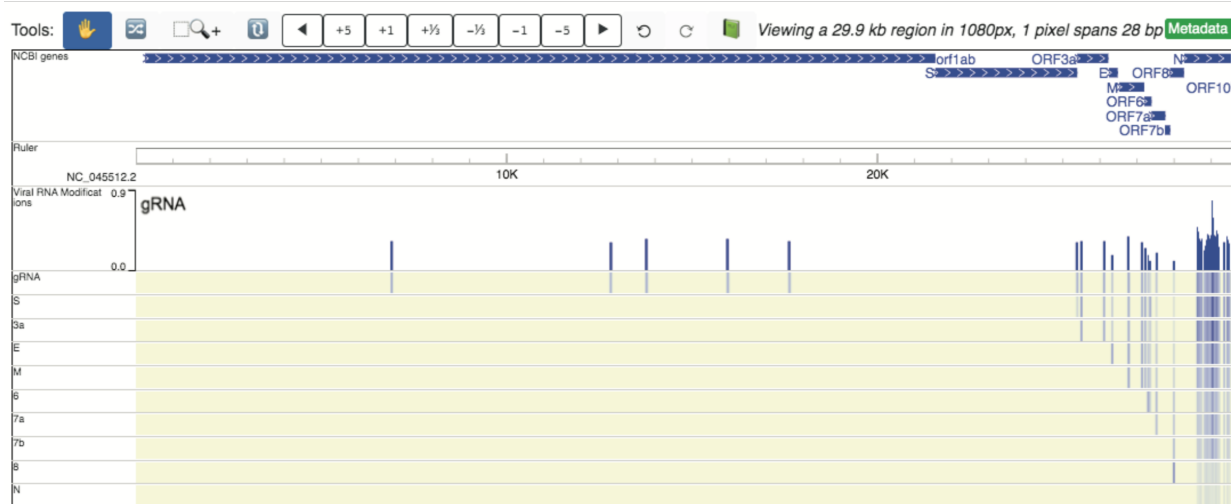
4.2.6 Recombination Events

Recombination events in the SARS-CoV-2 transcriptome were detected by junction-spanning RNA-seq reads generated by Kim, et al., 2020 (PMID: 32330414), and comprise three tracks collectively making up the Recombination events data hub. Of the three included tracks, two are longrange interaction tracks, displaying TRS-L dependent recombination events and TRS-L-independent recombination events, respectively. Locations of predicted recombination sites (TRSs or transcription regulatory sequences) are also available as an additional track. All three are shown below.



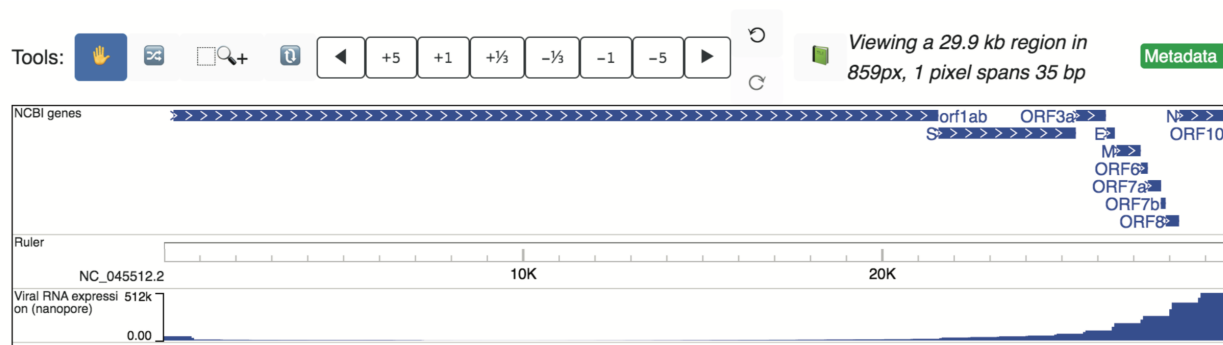
4.2.7 Viral RNA modifications

RNA modifications detected using Nanopore direct sequencing are reported in Kim et al., 2020 (PMID: 32330414), and comprise the 11 tracks available in the Viral RNA modifications data hub. Modification states include: gRNA, S, 3a, E, M, 6, 7a, 7b, 8, and N. Each modification has a static track that can be loaded in individually. In addition, a dynamic track is available (and also loaded in the default SARS-CoV-2 browser view) which rotates through displaying the modification signal across the genome for each modification.



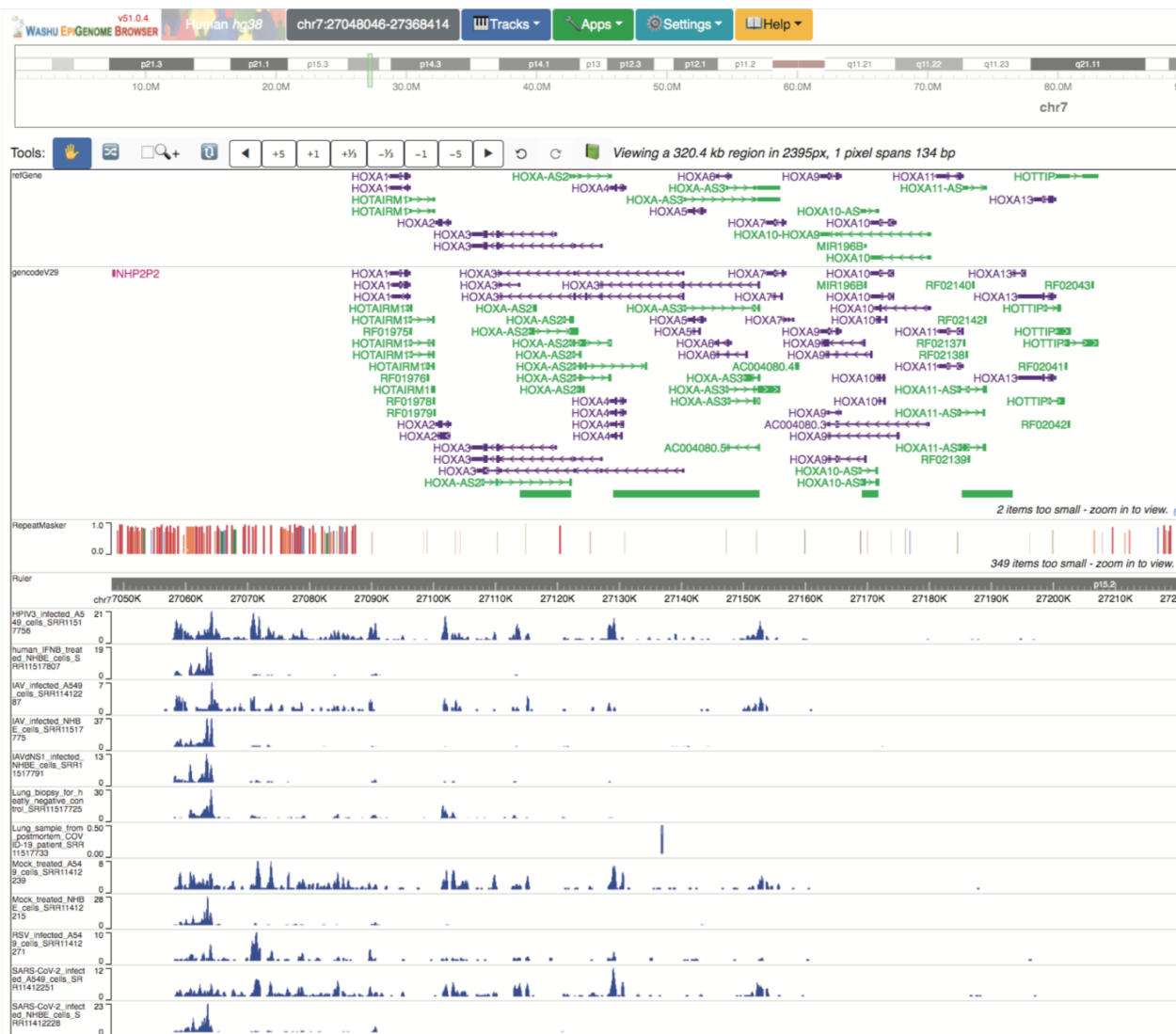
4.2.8 Viral RNA expression

Nanopore expression data was collected for SARS-CoV-2-infected Vero cells, and reported in Kim et al., 2020 (PMID: 32330414). In our data hub “Viral RNA expression”, we have added a bigwig file which displays the raw nanopore read counts at each genomic position. This track, shown below, is also one of the tracks displayed by default for SARS-CoV-2.



4.2.9 SARS-CoV-2 host transcriptional responses database

In addition to viral genomics pairwise alignments hosted by the browser, the WashU Virus Browser offers a unique view of host transcriptional responses to SARS-CoV-2 infection through partnership with the WashU Epigenome Browser. When navigating to the WashU Virus Genome Browser landing page, the user can opt to view data hubs containing host responses by selecting the link “Host transcriptional responses to SARS-CoV-2” under the “Featured Datahubs” drop-down menu. Selecting this link redirects the user to the hg38 genome hosted in the WashU Epigenome Browser, as shown below.



As demonstrated above, 12 RNA-seq tracks are pre-loaded into view from the pre-print Blanco-Melo, et al., 2020 (PMID: 32416070). However, the user can choose to look at additional tracks available within the data hub by selecting “Tracks” > “Public Data Hubs” > “SARS-CoV-2 Host Transcriptional Responses (Blanco-Melo, et al. 2020) Data Hub.” The data hub houses 195 RNA-seq tracks which can be either directly loaded into view, or can be pre-filtered based on several metadata terms. Once desired tracks are loaded into view, associated metadata can be displayed and includes the options shown below.

Current terms

Suggested terms

- + Track type
- + Run
- + BioProject
- + BioSample
- + Experiment
- + GEO_Accession
- + Sample_Name
- + Source_Name
- + SRA_Study
- + Treatment
- + Strain

CHAPTER 5

Contact Us

You can contact us by submitting an issue request at our github repo: <https://github.com/twlab/virusbrowser/issues>.

CHAPTER 6

Cite Us

If you find the Browser is useful for your research, please help us by citing the following paper:

Flynn, J.A., Purushotham, D., Choudhary, M.N.K. et al. Exploring the coronavirus pandemic with the WashU Virus Genome Browser. *Nat Genet* (2020). <https://doi.org/10.1038/s41588-020-0697-z>

CHAPTER 7

Indices and tables

- `genindex`
- `modindex`
- `search`