SARS-COV-2 GenBrowser file format (version 3.0)

There are two versions of GenBrowser: stand-alone version and online version. The two versions have different features. The interface files they need are also different. This document only provides format specification for the interface files, please read the user manual for how to use the software (http://www.egps-software.net/egpscloud/eGPS_Desktop.html).

You can download the latest interface files from the following website: https://ngdc.cncb.ac.cn/ncov/apis/data-latest/

Archives are available at:

https://ngdc.cncb.ac.cn/ncov/apis/archives/

In order to record mutations better, we designed the 3.0 format version. The 3.0 format version takes the same compression way as the2.0 format version.

Note:

1. The interface file format consists of two parts: the first is the way of compression and decompression, and the second is the way of encoding the specific content. V1.1 version use zip for compression/decompression; V2.0 and V3.0 use tar/xz for compression and decompression. Simply speaking, file format = compression/decompression method + specific content.

   How to decompression:

   Zip

   Windows operation system: Use winRAR and other software to decompress.

   Command line: unzip yourFile.zip

   Tar/xz

   Windows operation system: Use winRAR and other software to decompress.

   Command line: tar -xf yourFile.txz

The required interface files for both versions are listed in table 1:

Table 1: The required interface files.

Classification	Filename	Required for	Information
For tree visualization	accessionNumbers.txz	Both	Information of leaf nodes in the tree
	mainDataFile.txz	Both	Tree topology, sample information, etc.
	countries.zip	Both, the desktop version is self-contained	Name and code of the countries/regions
For genome visualization	processed_aligned.6.virus.refined.zip	Online version	Processed DNA alignment file
	processed_protein.aligned.6.virus.zip	Online version	Processed protein alignment
	similarity_window_20_100.zip	Online version	Data needed for similarity plot
	refGenomeInfor.zip	Both, the desktop version is self-contained.	Structure of reference genome of COVID-19
	aligned.6.virus.refined.fas.zip	Stand-alone version, self-contained	Multiple sequence alignment of six representative genome
	key_domains.zip	Both, the desktop version is self-contained	Key domains of genome
	firstSubmitter.txz	Both	Submitters who have made the first discoveries of novel SARS-CoV-2 variants
	selectionCof.txz	Stand-alone version	To calculate selection coefficient for a number of alleles
	mutationFreq.zip	Online-version	Global mutation frequency for per genomic locus
	primers.zip	Both, the desktop version is self-contained	File for primer track

mainDataFile.zip and accessionNumbers.zip are archived. The two files contain the annotated evolutionary tree, mutations, tip-dated leaves, collection date of strains, accession numbers, isolate names, patient sex and age.

Note:

1. To use the web version , users only need to open browser and access the following URL (https://www.biosino.org/genbrowser/ or https://ngdc.cncb.ac.cn/genbrowser/).

2. To use the stand-alone version, users can choose to automatically obtain data from the Internet or to read local data by entering the file path. When reading local data, the compressed file that at least contains “mainDataFile” and “accessionNumbers” are required.

3. “mainDataFile” is the core file, containing the mutation information and associated meta information of strains. With this file, users can rebuild the sequence alignments (Programming skill is required).

4. The files “mainDataFile” and “accessionNumbers” are constantly updated and are the necessary files for tree visualization. The stand-alone version can start with the two files.

5. The “selectionCof” is the file storing the results of positive selection. The file is not necessary and will update with “mainDataFile” and “accessionNumbers”. This file can be provided in advance to speed up software response, or results can be calculated instantly by the software when the file is not provided. When the user chooses to obtain data from the Internet, the software will get this file from the Internet. When the user chooses to obtain local data, the software will automatically recognize and process the file that was entered by user. When the file exists, the software will read the file and show the results of positive selection. When the file is lost, the software can still start normally, but the positive selection module will not show results and users can click the calculate button to get real-time results.

6. The “firstSubmitter” is the file storing the information of submitters. The file is not necessary and will update with “mainDataFile” and “accessionNumbers”. The information this file stored will be provided to the “Mutation freq track”. When the file exists, users can check the information of the submitter who first submitted the sequence that contains the mutation in certain site. When the file is lost, the software can still start normally, but the submitter information cannot be displayed.

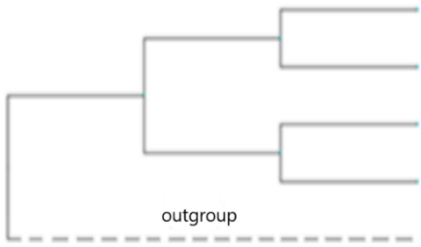
7. Other files that the stand-alone version need will be distributed with the eGPS software. Users do not need to prepare these files. All files required for the web version are obtained from the Internet.

1 FILES FOR TREE VISUALIZATION

1.1 MAIN DATA FILE

Filename： mainDataFile.txt

General description：
This format is extended from Newick tree format (nwk) format.



1.1

Original format is ((1,2), (3,4)). After coding, it takes the shape of evolutionary tree, as shown in picture 1.1.

The lowest branch represents the outgroup. The outgroup should be represented by a dotted line.

This file contains information of tree topology, sample information, etc.

Line by line description：
Line 1：#SARS-Cov-2 format eGPS v3.0

This line indicates the version number of the data format.

Line 2：Updated on 625:1670101

This line indicates the data update time，and December 1, 2019 is defined as 0. A positive integer indicates how many days later than this date; a negative integer indicates how many days earlier. 1670101 is the number of the strains (exclude outgroups).

Line 3：Genome size 29903 | considered from 100 to 29800

Genome size 29903

This line indicates the length of reference sequence of COVID-19 (accession number: NC_045512) is 29903. The sequence can be downloaded from the following link: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

considered from 100 to 29800

This line indicates the actual genome length we used. However, when assembling the sequences, error might occur at the beginning and end of the genome. Therefore, we ignore about 100bp at both ends. The actual genome sequence should be indicated as the red words shown.

100; // 1-based, inclusive

29800; // 1-based, inclusive

In this example, if the length of branch is measured by the condition of each site. The length of branch = the number of mutations in the branch / (29800 – 100 + 1).

Line 4：Mutation rate per year per gene

This line indicates the mutation rate of each gene per year. This information is the estimated mutation rate for each gene, and it includes different mutation types, i.e., SNPs, insertions and deletions.

From line 5：Main content of this format

This part records the topological structure of the evolutionary tree and the corresponding information between inner node and leaf node.

Definition of the outgroups
Define outgroup：outgroup*

Using outgroup as the keyword. The leaf node that use this keyword at the beginning of their name is an outgroup.

For example: outgroup_RaTG13 or outgroup-RaTG13. The keyword is case-insensitive.

In addition, define multiple outgroups are allowed. When there are multiple outgroups, the names of all outgroups should be displayed. However, only one dotted line is needed to represent multiple outgroups.

Names of outgroups in software
After the file is read, the name of outgroup displayed in the software is a string after removing the keywords and the connection string in the middle.

For example: If the name of outgroup in the node information is “outgroup_RaTG13”, we will remove the keyword “outgroup” and the connection string “_”and the name of outgroup displayed in software will be “RaTG13”.

Definition of leaf node
Format：

index:mutation:collection date:patient gender:patient age:country code:province code or province name

For example: 1:C29200T:012120:M:46:86:Chongqinq/Yongchuan

Serial number	Key	Type of value	Examples	Notes
1	index	int	1,2,3…	Index can be used for getting the name of the virus strain, as well as the access number of the database. The indexes for outgroups are -1, -2, -3, etc.
2	mutation	string	A1313T|T134A	Multiple mutations are allowed, using "|" as separator. The value is allowed to be empty. If the value is empty, it means there is no mutation instead of the data is missing. Mutation types include SNV, insertion, and deletion. For more details, see the table "Mutation record format" below.
3	sampling date	int	25	Define December 1, 2019 as 0. A positive integer indicates how many days later than this date; a negative integer indicates how many days earlier. The value is allowed to be empty. If the value is empty, it means the data is missing.
4	patient gender	char	F	F (female) or M (Male). The value is allowed to be empty. If the value is empty, it means the data is missing.
5	age of patient	double	45 or 0.5	The value is allowed to be empty. If the value is empty, it means the data is missing.
6	country and region code	int	86 (China)	Refer to the long-distance call country/region code, which has been defined in the software. Since Canada and the United

				States share the same long distance call country/region code, the country/region code of Canada is defined as an unassigned number 887. For the definition of country/region code, please see the file *countries.json*. This information is not needed in the outgroup, so this value is empty in the outgroup.
7	province code or province name	string	Wuhan	Provinces that appear only once do not need to be defined specifically. Provinces that appear in multiple samples should be defined in the data file. Provinces that appear multiple times are numbered from zero. The correspondence between the index and the province will be recorded from line 6 of this document. For example, if Wuhan appears several times, then define 1 as Wuhan. This definition can be changed at any time and needs to specified in the data file. This information is not needed in the outgroup, so this value is empty in the outgroup. Attention：brackets are not allowed in province names.

Table: Mutation record format

This format records the change from the ancestor state to the derived state. We define the record way of three mutation types as:

Type	Record format	Example	Annotation
SNV	Ancestral allele (A/T/C/G)+Position+ Derived allele	A5T, A23403G A1AT	We define SNV as a mutation that affects only a single

			genomic position (the position here refers specifically to the position of the reference genome). When the ancestor state is not '-', the insertion of single position is also classified as SNV here.
Deletion	Ancestral alleles + Position+ '-'	GTT5-	This type represents successive deletions starting from 'Position'. Please note that the genomic position to be deleted may have an insertion string.
Insertion	'-' + Position+Derived allele When the state of the last position (position of the reference genome) contains an "Insert string", we add a "`" before the "Insert string".	(1) Insertions as back mutations of deletions affecting multiple positions：ATG1-，-1ATG (2) Insertions of which the "Ancestral allele" is "-": ATTAA1-, -1A`GG,-2TTAA	The "Ancestral allele" of insertions must be '-'.

Definition of internal node
Format:

Mutation:inferred date:lower limit offset of confidence interval:upper limit offset of confidence interval

Key	Type of value	Examples	Notes
Mutation	string	A1313T|T134A	Multiple mutations are allowed. If there is an outgroup, this

			value in the root node is empty. That is, when there is an outgroup, this value in both the root node and the node of the outgroup is empty.
Inferred date	int	25	Define December 1, 2019 as 0. A positive integer indicates how many days later than this date; a negative integer indicates how many days earlier. In outgroup, this value is empty. It means the value direct display as ":". There is no information in front of the colon string. When there is an outgroup, this value in both the root node and the node of the outgroup is empty.
Lower limit offset of confidence interval	int	0 or positive value	Explain the number of days of possible error earlier than the date value. When there is an outgroup, this value of the root node is 0.
Upper limit offset of confidence interval	int	0 or positive value	Explain the number of days of possible error later than the date value. When there is an outgroup, this value of the root node is 0.

1.2 ADDITIONAL INFORMATION FILE

File name： accessionNumbers.txt

General description：additional information file for samples
Contained information: the serial number in the database used for samples

Line by line description：
Format of following lines ： isolate accessionNumber

Key	Type of value	Examples	Notes
Name of the sample	string	NanChang/JX216/2020	Name of the sample (isolate).
accession number	string	MT039888 or GWHABKJ00000000	The index number of the genome sequence in the database. For a sequence which has records in multiple databases, the sequence number used after excluding duplicate records.

Each sample also corresponds to a unique index number, and the node index number corresponds to the leaf node index number of the mainDataFile. It is used to obtain the name of the sample (i.e. virus isolate), accession number of the database for corresponding leaf node. The index number is not written in the accessionNumbers.txt, it is obtained by the line number, and the index of the first line is fixed to -2, and it is incremented line by line.

Special note: The sequence number and related information is not needed for the outgroup. However, for the consistency, and interpret the outgroups in standard data files, the association between the name of outgroup and index number is also defined. The index numbers of outgroups are all negative integers. We use two outgroups, so the index number starts from -2.

1.3 COUNTRY CODE FILE

File name: countries.json or countries.zip

General description：
This file stores the name of the country and its corresponding code in JSON format. The country code refers to the long-distance call country code. For example, the code for China is 86.

2.1 ALIGNMENT FILE FOR SIX REPRESENTATIVE GENOMES

File name: aligned.6.virus.refined.fas.zip

General description:

This file stores the alignment of genome sequences for 6 representative species. The file is prepared for stand-alone version and the stand-alone version GenBrowser will calculate the alignment for protein. This file includes the complete sequence alignment file, such as those that do not exist in the standard genome of the COVID-19, but exist in other coronaviruses.

2.2 DNA ALIGNMENT FILE AFTER PROCESSED

File name: processed_aligned.6.virus.refined.zip

General description:

This file stores the nucleotide sequence information that can be directly used for visualization. This file only includes the alignment of corresponding positions in the reference genome of COVID-19.

2.3 PROTEIN ALIGNMENT FILE AFTER PROCESSED

File name: processed_protein.aligned.6.virus.zip

General description:

This file stores the amino acid sequence information that can be directly used for visualization. This file only includes the amino acid alignment of corresponding positions in the reference genome of COVID-19.

2.4 GENOME SIMILARITY FILE

File name: similarity_window_20_100.zip

General description:

The similarity of six representative genomes is shown in the form of sliding window. SARS-COV-2 genome is used as reference sequence. The file saves the calculation

results when the window step is 20 and the window size is 100. The file is only needed for visualization of the online version, while the stand-alone version will perform real-time calculation based on alignment file for six representative genomes. Therefore, the file is only provided in the format of JSON file.

2.5 STRUCTURE FILE OF REFERENCE GENOME OF COVID-19

File name: refGenomeInfor.zip

General description:

This file records the structure of the reference genome of the COVID-19. The file format is TSV format. The first column is the ORF name, and the last two columns are the start and end positions.

2.6 KEY DOMAIN DOCUMENTS OF GENOME

File name: key_domains.zip

General description:

The format of the file is TSV, with tab key to separate the information of one record, including header and specific information. The data are from NCBI (all locations are corresponding to the reference genome).

Columns are:

1) name: the name of domain

2) gene: the gene name of domain

3) aa_start: amino acid starting position of domain

4) aa_end: amino acid ending position of domain

5) nt_start: nucleotide starting position of domain

6) nt_end: nucleotide ending position of domain

7) Pfam_ID/CDD_ID: Pfam database ID or CDD database ID (some files do not provide Pfam ID)

8) Note: Some related information of domain

9) website：the website address needed for linking to Pfam or CDD database

2.7 PRIMER INFORMATION FILE

File name: primers.zip

General description:

The format of the file is TSV format, with tab key to separate the information of one record, including header and specific information.

File format:

1) Institution: primer source, the institution that designed the primer
2) Gene: the gene corresponding to the primer location
3) Index: a number index for the primers that for same gene and designed by the same institution
4) F_Start: Nucleotide starting position of forward primer
5) F_End: Nucleotide ending position of forward primer
6) R_Start: Nucleotide starting position of reverse primer
7) R_End: Nucleotide ending position of reverse primer

2.8 SUBMITTERS WHO HAVE MADE THE FIRST DISCOVERIES OF NOVEL VARIANTS

File name: firstSubmitter.zip

General description: The file is in JSON format and is used in “Allele freq track” to provide information of submitters who have made the first discoveries of novel variants and the time of submission.

File format:

{

"Mutation2SubmitterInfMap":{

"The state after mutation":{

"accessionNum": "The accession number of the sample that first carried the mutation. ",

"submitDate": " The submission time of the sample that first carried the mutation. ",

"submitter": " The submitter of the sample that first carried the mutation. "

},{}, …

}

}

For example:

```
{
"Mutation2SubmitterInfMap":{
"22984A":{
"accessionNum": "EPI_ISL_416682",
"submitDate": "2020-03-23",
"submitter": "UW Virology Lab"
},
"6317T":{
"accessionNum": "EPI_ISL_417157",
"submitDate": "2020-03-24",
"submitter": "Seattle Flu Study"
}, {},{}…
}
}
```

2.9 GLOBAL MUTATION FREQUENCY

File name: mutationFreq.zip

General description: The file is in JSON format and is provided for the web version to display the global mutation frequency.

File format:

{"listOfLeafStates":[

{"ancestralState": " The ancestral state of the site ",

"derivedStates": ["The derived state of the site ","…","…"],

"freq": "The mutation frequency of this site",

"position": The position information

}, {},…

]

}

For example:

{"listOfLeafStates":[

{"ancestralState":"C","derivedStates":["A","T","-"],"freq":"0.001887","position":100},

{"ancestralState":"G","derivedStates":["A","C","T","-"],"freq":"0.000286","position":101},

{"ancestralState":"G","derivedStates":["A","T","-"],"freq":"0.00007","position":102},

{"ancestralState":"C","derivedStates":["A","T","-"],"freq":"0.000161","position":103},

{"ancestralState":"T","derivedStates":["C","-"],"freq":"0.000017","position":104},

{"ancestralState":"G","derivedStates":["A","C","T","-"],"freq":"0.000733","position":105},

{"ancestralState":"C","derivedStates":["A","T","G","-"],"freq":"0.002435","position":106},

{"ancestralState":"A","derivedStates":["C","G","-"],"freq":"0.000026","position":107}]

}

Special note: The Tooltip of the mutation frequency module of the web version only displays the above information only, and will not recalculate the allele frequency when the tree changes due to filtering operations.

2.10 SELECTION COEFFICIENT

File name: selectionCof.zip

General description: The file is in JSON format and is used to store the results of putative positively selected sites calculated with the whole sample in the tree. The results stored in this file is used as the default data presented in the “Allele frequency based” module in the “Non-neutral evolution” panel

File format:

[

{

derivedAA": ["Amino acid mutation "],

"endFreq": The frequency of the mutation at the mutation fixed time or final sampling time,

"endTime": " The mutation fixed time or final sampling time ",

"freq": [The data (in days) of the change in frequency of the mutation since its occurrence.],

"gene": "The name of the gene the mutation located in",

"mutatedState": " Nucleotide mutation ",

"pValue": P value for linear fitting when calculating selection coefficient,

"position": The position of mutation,

"r2": R square for linear fitting when calculating selection coefficient,

"selCoeff": The selection coefficient,

"startFreq": The start frequency of the mutation,

"startTime": " The time when the mutation was first detected "

},

{}, … ,

{}

]

For example:


```
[
{
derivedAA": ["T2007I"],
"endFreq": 0.08571428571428572,
"endTime": "20/10/16",
"freq": [0.000814663951120163,
        0.0007374631268436578,
        0.0006731740154830024, ….],
"gene": "ORF1a",
"mutatedState": "C6285T",
"pValue": 0.0,
"position": 6285,
"r2": 0.55172130669812,
"selCoeff": 0.01413648876353557,
"startFreq": 0.000814663951120163,
"startTime": "20/03/11"
},
{}, … ,
{}
]
```