

## SARS-COV-2 GenBrowser 数据交换接口文件格式定义(1.1 版)

因为 GenBrowser 由单机版和网络版两部分组成，两者的功能不尽相同，所需接口文件也不相同。本文件仅为接口文件提供格式说明，具体如何使用软件请参考使用手册（[http://www.egps-software.net/egpscloud/eGPS\\_Desktop.html](http://www.egps-software.net/egpscloud/eGPS_Desktop.html)）。

当前数据接口文件获取地址：<https://ngdc.cncb.ac.cn/ncov/apis/data-latest/>

历史数据接口文件获取地址：<https://ngdc.cncb.ac.cn/ncov/apis/archives/>

单机版与网路版所需文件见下面列表：

表 1、接口文件列表

分类	文件名	需求对象	包含信息
进化树	accessionNumbers.zip	两者	进化树叶子(菌株)状态信息
可视化	mainDataFile.zip	两者	树拓扑结构，菌株信息等
所需要	finalLeafStates.zip	网页版	菌株基因组突变状态
的文件	countries.zip	两者	国家和地区的名称与编号
基因组	processed_aligned.6.virus.refined.zip	网页版	处理过后的 DNA 联配文件
可视化	processed_protein.aligned.6.virus.zip	网页版	处理过后的蛋白质联配文件
所需要	similarity_window_20_100.zip	网页版	相似性点图所需数据
的文件	refGenomeInfor.zip	两者	新冠病毒参考基因组的结构
	aligned.6.virus.refined.fas.zip	单机版	六条代表性的基因组多重联配
	key_domains.zip	两者	基因组关键结构域
	firstSubmitter.zip	两者	最先发现该新突变的人或组织
	selectionCof.zip	单机版	关于选择系数的计算
	primers.zip	两者	引物 track 的默认接口文件

其中历史数据接口文件包括最重要的 mainDataFile.zip 和 accessionNumbers.zip 两个数据文件。这两个文件包括 annotated evolutionary tree, mutations, tip-dated leaves, 国家和地区、采样时间、病人年龄和性别。

说明：

1. 网络版只需打开网页，输入网址即可使用（<https://www.biosino.org/genbrowser/> 或 <https://ngdc.cncb.ac.cn/genbrowser/>）。

2. 使用单机版可选择自动从网络获取数据或者从本地读取数据，从本地读取时至少需包括“mainDataFile”和“accessionNumbers”两个文件。
3. “mainDataFile”为核心文件，包含了毒株的突变信息和相关 meta 信息，足以重建序列比对（需使用者自己编写程序进行解析）。
4. “mainDataFile”和“accessionNumbers”文件会不断更新，为树可视化的必须文件，有这两个文件即可启动单机版程序。
5. “selectionCof”为正选择检测的结果文件，非必须，会随着“mainDataFile”和“accessionNumbers”文件同步更新。该文件可预先提供以加快软件响应速度，也可以不提供直接自行计算。当用户选择从网络读取文件时，软件会从网络获取该文件。当用户选择从本地读取文件时，软件会自动识别并处理用户输入的文件。文件存在时，软件将读取该文件内容并显示正选择检测结果。文件不存在时，软件仍可以启动，选择系数计算模块将不显示结果，用户可点击计算按钮实时计算。
6. “firstSubmitter”为存储提交者信息的文件，非必须，将会随着“mainDataFile”和“accessionNumbers”文件同步更新。该文件存储的信息将会提供给 Mutation freq. track，当存在该文件时，用户可以在 track 中查看到是哪个提交者率先提交的序列包含了这个位点的突变。当文件不存在时，软件仍然可以正常启动，但无法显示提交者信息。
7. 其余单机版所需文件会随软件 eGPS 一起分发，无需用户准备。网络版所需文件均从网络获取。

## 1 进化树可视化所需要的文件

---

### 1.1 主要数据文件

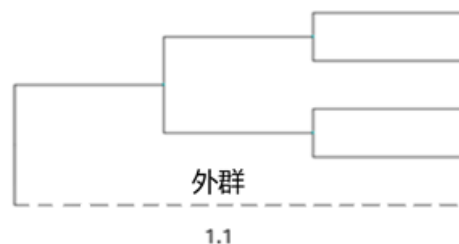
文件名： mainDataFile.txt

总体说明：

本格式拓展自 Newick tree format (nwk)格式。原格式为 ((1,2),(3,4))，编码后呈进化树的形状，如右图 1.1 所示。

最下方分支代表外群，以虚线表示。

本文件包含信息：树拓扑结构，菌株信息等。



逐行说明：

第 1 行：#SARS-Cov-2 format eGPS v1.1

此行说明 数据格式 的版本号。

第 2 行：Updated on 062820

此行说明数据更新的时间，格式为【月、日、年】。

第 3 行：Genome size 29903 | considered from 100 to 29800

Genome size 29903

此行说明新冠病毒的标准序列(accession number: NC\_045512) 长度为 29903，但是纳入数据分析的区域限制在第 100 到 29800 之间的序列。其中标准序列可以从下述链接获取：[https://www.ncbi.nlm.nih.gov/nucleotide/NC\\_045512](https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512)

considered from 100 to 29800

此行说明的是实际使用的基因序列。在基因组的起始两端，进行测序拼接时可能出现的问题。因此在起始两端各忽略大约 100bp 长度碱基。

100; // 1-based, inclusive

29800; // 1-based, inclusive

因此，如果用每个位点的情况衡量枝长，该枝的枝长 = 发生在该枝的突变数目 / (29800 - 100 + 1)

第 4 行：Mutation rate per year per gene:

此行说明每个基因每年的突变速率。该信息为针对不同基因估算的突变速率，在估计的时候纳入了各种不同的突变类型，包括 insertion 与 deletion。

第 5 行：本格式的主体内容

记录了进化树的拓扑结构，内节点与叶子结点对应的各种信息等。下文分块重点阐述里面的内容

## 外群的定义方式

定义外群: `outgroup*`

用 `outgroup` 作为关键词, 使用此关键词作为名字开头的叶子节点, 即为外群。

比如 `outgroup_RaTG13` 或者 `outgroup-RaTG13`。关键词不区分大小写。

此外, 允许定义多个外群。当存在多个外群时, 显示所有外群的名字, 但是代表外群的虚线只需要一条。

## 软件内外群的名字

读取文件后, **软件内显示的外群名字** 为 **节点信息中外群的名字** 去除关键词以及中间的连接字符后的字符串。

例如: 节点信息中的外群名字 “`outgroup_RaTG13`” 去除关键词 `outgroup`, 以及中间的连接字符 “`_`”, 得到最终显示在**软件内的外群名字** “**`RaTG13`**

## 叶子节点定义

格式:

节点索引号 : 突变 : 采样日期 : 病人性别 : 病人年龄 : 国家代码 : 省份代码或省份

例如: `1:C29200T:012120:M:46:86:Chongqing/Yongchuan`

序号	列	类型	举例	备注
1	节点索引号	int	1,2,3...	用于通过节点索引号, 获取病毒毒株的名字, 以及数据库的 <b>accession number</b> , 外群索引为 -1,-2,-3....
2	突变	string	A1313T T134A	允许存在多个突变, 使用“ ”分隔。 允许为空。若为空则表示没有突变, 而不是数据缺失。 下面是如何表示插入与缺失。 <b>Deletion:</b> * <b>A15093-</b> * <b>AAG12147-</b> * <b>TGCTG19482-</b> <b>A15093-</b> 表示在15,093的碱基A被删除了。

				<p>AAG12147-表示在12,147的碱基A和后面两个碱基AG被删除了。</p> <p>如果删除序列过长，比如TGCTG19482-，则客户端显示为T..G19482-。在接口文件中则为完整的删除序列。</p> <p><b>Insertion:</b></p> <ul style="list-style-type: none"> <li>* G11527GT</li> <li>* G11071GTTC</li> <li>* T11554CAG</li> <li>* T12877TCTACGG</li> </ul> <p>G11527GT表示在11,527的碱基G后面插入了一个碱基T。</p> <p>G11071GTTC表示在11,071的碱基G后面插入了三个碱基TTC。</p> <p>T11554CAG表示11,554的碱基T，被替换为三个碱基CAG。</p> <p>如果插入序列过长，比如T12877TCTACGG，则客户端显示为T12877T..G。在接口文件中则为完整的插入序列。</p>
3	采样信息	string (6 chars)	031120	格式为: 月、日、年。 允许为空。若为空则表示数据缺失。
4	病人性别	char	F	F (female 女性)或 M (Male 男性)。允许为空。若为空则表示数据缺失。
5	病人年龄	double	45 或 0.5	允许为空。若为空则表示数据缺失。
6	国家和地区代码	int	86 (中国)	参考长途电话国家和地区代码，在软件中已定义。因加拿大与美国共享同一个长途电话国家代码，所以这里定义加拿大的国家

				和地区代码为一个未分配的号码 887。国家和地区代码的定义见 <code>countries.json</code> 文件 在外群中不需要这个信息，故为空。
7	省份	string	Wuhan	仅出现一次的省份不需要专门定义。多个样本中出现的省份，应在数据文件中定义。 例如 <b>Wuhan</b> 出现了多次，则定义 1 代表武汉。该定义可以随时在变化，需在文件中指定。 在外群中不需要这个信息，故为空。 <b>注意：需要确认地区的名称中，禁止使用括号。</b>

#### 内节点定义

格式:

突变:推测日期:置信区间下限偏移量:置信区间上移偏移量

列	类型	举例	备注
突变	string	A1313T T134A	允许存在多个突变。 有外群时，根节点此项为空。 即有外群存在时，根节点与外群所在节点此项都为空。
日期	string	031120	格式为: 月、日、年。 外群为空，即直接显示的是“:”，前面没有任何信息。 有外群存在时，根节点与外群所在节点都为空。
置信区间下限偏移量	int	0 或 正值	比上述时间更早多少天。 有外群时，根节点此项为 0。
置信区间上移偏移量	int	0 或 正值	比上述时间更晚多少天。 有外群时，根节点此项为 0。

## 1.2 信息补充文件

文件名: accessionNumbers.txt

总体说明: 样本的补充信息文件

包含信息: 样本所对用的数据库中的序列号

逐行说明:

第 1 行

#SARS-CoV-2 format eGPS v1.1

说明版本号

第 2 行: Updated on 042920, 写明数据更新的时间, 格式为: 月、日、年。

后续行的格式: seqIndex isolate accessionNumber

项目	类型	举例	备注
节点索引号	int	1, 2, 3, .....	通过节点索引号与其他文件的叶子节点索引号对应。用来获取病毒毒株的名字, 以及数据库的 accession number 与其对应的叶子节点。
病毒株的名称	string	NanChang/JX216/2020	病毒株的名称(isolate)。
序列号, 即 accession number	string	MT039888 或者 GWHABKJ00000000	基因组序列在数据库中的索引号。对于在多个数据库中存在记录的序列, 排除重复记录之后, 所用的序列号。

特别说明: 外群不需要上述序列号和病毒株信息。然而为了一致性, 以及解读标准数据文件中的外群, 定义外群的名称和索引号的关联。外群的索引号均为负整数。

### 1.3 叶子节点最终状态文件

文件名: finalLeafStates.txt

总体说明:

用于协助计算突变频率, 供网页版使用, 减轻浏览器所需要的计算负担。

格式:

第一列为突变位置;

第二列开始为叶子的索引号

例如:

514	59	
17410	59	
28166	153	
9223	67	
29705	22	
27147	96	97

从第 2 列开始就是突变的叶子(病毒株)的索引号, 此索引号同 accessionNumbers.txt 文件。

### 1.4 国家代码文件

文件名为 countries.json

总体说明:

以 JSON 格式存储了国家的名称与其对应的代号, 参考了长途电话国家代码。例如中国的代号是 86。



## 2 基因组可视化所需要的文件

---

### 2.1 六条代表性基因组联配文件

文件名: aligned.6.virus.refined.fas.zip

总体说明: 该文件以 fasta 格式存储了 6 条具有代表性物种的基因组序列联配。该文件供单机版使用, 单机版将会从中计算出蛋白质的联配信息。这一文件包括完整的序列比对文件, 比如在新冠病毒标准基因组中不存在的、但是在其他冠状病毒中存在的序列片段。

### 2.2 处理后的 DNA 联配文件

文件名: processed\_aligned.6.virus.refined.zip

总体说明: 以 fasta 格式存储了直接用来可视化的核苷酸序列信息。这一序列比对, 仅仅包括新冠病毒标准基因组中相应位置的序列比对的情况。

### 2.3 处理后的蛋白质联配文件

文件名: processed\_protein.aligned.6.virus.zip

总体说明: 以 fasta 格式存储了直接用来可视化的氨基酸序列信息。这一序列比对, 仅仅包括新冠病毒标准基因组中相应位置的氨基酸序列比对的情况。

### 2.4 基因组相似性文件

文件名: similarity\_window\_20\_100.zip

总体说明: 以滑动窗口形式展示的六条具有代表性基因组的相似性, 用 SARS-COV-2 基因组作为参考序列。该文件保存了 window step 为 20, window size 为 100 时的计算结果。该文件仅供网页版可视化所需, 而单机版会根据六条代表性基因组联配文件进行实时计算。故该文件仅提供了 json 文件格式。

### 2.5 新冠病毒参考基因组的结构文件

文件名: refGenomeInfor.zip

总体说明: 该文件记录了新冠病毒参考基因组的结构。文件格式为 tsv 格式, 第一列为 ORF 名称, 后两列是起止位置。

## 2.6 基因组关键结构域文件

文件名: key\_domains.zip

总体说明: 该文件格式为 tsv 格式, 以 Tab 键分隔一条记录的信息, 包括表头与具体的信息。数据来自 NCBI (所有位置都是对应参考基因组的)。

格式 (各个列分别为):

- 1) name: domain 的名字
- 2) gene:domain 所在的基因名
- 3) aa\_start:domain 的氨基酸起始位置
- 4) aa\_end:domain 的氨基酸结束位置
- 5) nt\_start:domain 的核苷酸起始位置
- 6) nt\_end:domain 的核苷酸结束位置
- 7) Pfam\_ID/CDD\_ID: Pfam 数据库的 ID 或者 CDD 数据库的 ID (有的没有提供 Pfam ID, 所以这里是一个或的关系)
- 8) Note: domain 的一些相关信息
- 9) website: 链接到 Pfam 或 CDD 数据库所需要的对应网址

## 2.7 引物信息文件

文件名: primers.zip

总体说明: 该文件格式为 tsv 格式, 以 Tab 键分隔一条记录的信息, 包括表头与具体的信息。

格式 (各个列分别为):

- 1) Institution: 引物来源
- 2) Gene: 引物所在位置对应的基因
- 3) Index: 同一个引物来源, 并且针对用一个基因设计的引物序号
- 4) F\_Start: 正向引物的核苷酸起始位置
- 5) F\_End: 正向引物的核苷酸终止位置
- 6) R\_Start: 反向引物的核苷酸起始位置

7) R\_End: 反向引物的核苷酸终止位置。

## 2.8 新突变的首次发现者（个人或组织）

文件名：firstSubmitter.zip

总体说明：该文件为 json 格式，用于给“Allele freq track”提供新突变的首次发现者（个人或组织）和首次提交时间的信息。

格式：

```
{  
  "Mutation2SubmitterInfMap": {  
    "突变后的状态": {  
      "accessionNum": "第一次发现该突变的样本的数据库编号",  
      "submitDate": "第一次发现该突变的样本的提交时间",  
      "submitter": "第一次发现该突变的样本的提交者"  
    }, {} , ...  
  }  
}
```

例子:

```
{
  "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 选择系数

文件名: selectionCof.zip

总体说明: 该文件为 json 格式, 存储使用现有树中的全部样本计算的潜在正选择位点的结果, 用于“Non-neutral evolution”面板的“Allele frequency based”板块的默认数据展示。

格式:

```
[
  {
    derivedAA": ["氨基酸突变"],
    "endFreq": 突变固定下来或到已有采样时间时该突变的频率,
    "endTime": "突变固定的时间或已有样本的最晚时间",
```

"freq": [从突变出现开始该突变的频率变化数据（天为单位）],  
"gene": "突变所在的基因名",  
"mutatedState": "核苷酸突变",  
"pValue": 计算选择系数时的线性拟合的  $p$  值,  
"position": 突变的位置,  
"r2": 计算选择系数时线性拟合的  $R^2$ ,  
"selCoeff": 选择系数,  
"startFreq": 突变的起始频率,  
"startTime": "突变第一次被检测到的时间"  
},  
{}, ...,  
{  
]

例子:

```
[
{
  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"
},
{, ... ,
}
]
```