



中国科学院北京基因组研究所（国家生物信息中心）

BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION



国家基因组科学数据中心

National Genomics Data Center

新冠病毒信息库与工具资源的使用

李翠萍

2024年10月21日

纲要

1. 新冠病毒**数据与信息库**
2. 大规模单体型网络快速**构建工具与追踪系统**
3. 高风险株系精准**预警模型及应用平台**
4. 新冠病毒**在线分析平台**

纲要

1. 新冠病毒**数据与信息库**

2. 大规模单体型网络快速构建工具与追踪系统

3. 高风险株系精准预警模型及应用平台

4. 新冠病毒在线分析平台

全球首个公开发布的新冠病毒信息库



Yi Chuan 2020; Zoological Research 2020; Genomics Proteomics Bioinformatics 2020, 2023

2020年1月22日上线，支撑全球新冠数据汇交整合、质控分析、监测预警



每日更新，服务全球

截至2024年10月20日，已收录全球**1778万**多条新冠病毒基因组序列信息，为全球**184**个国家/地区**46.6万**余访客提供数据服务，累计数据下载超过**206亿**条次，**国外访客占比高达60%**。

新冠病毒信息库RCoV19迭代更新

全球首个

- 数据收集整合
- 变异鉴定注释
(人工)

2020.1



- 变异鉴定注释
(自动化)
- 质量评估
- 单体型网络构建
- 系统演化树

2020.3



- 突变率
- 群体频率时空监测
- 演化支系监测
- 在线分析服务

2020.6



- 国家监测
- 核酸引物监测
- 多维知识审编
- 宿主监测

2022.1



- 高风险株系预警

2022.10

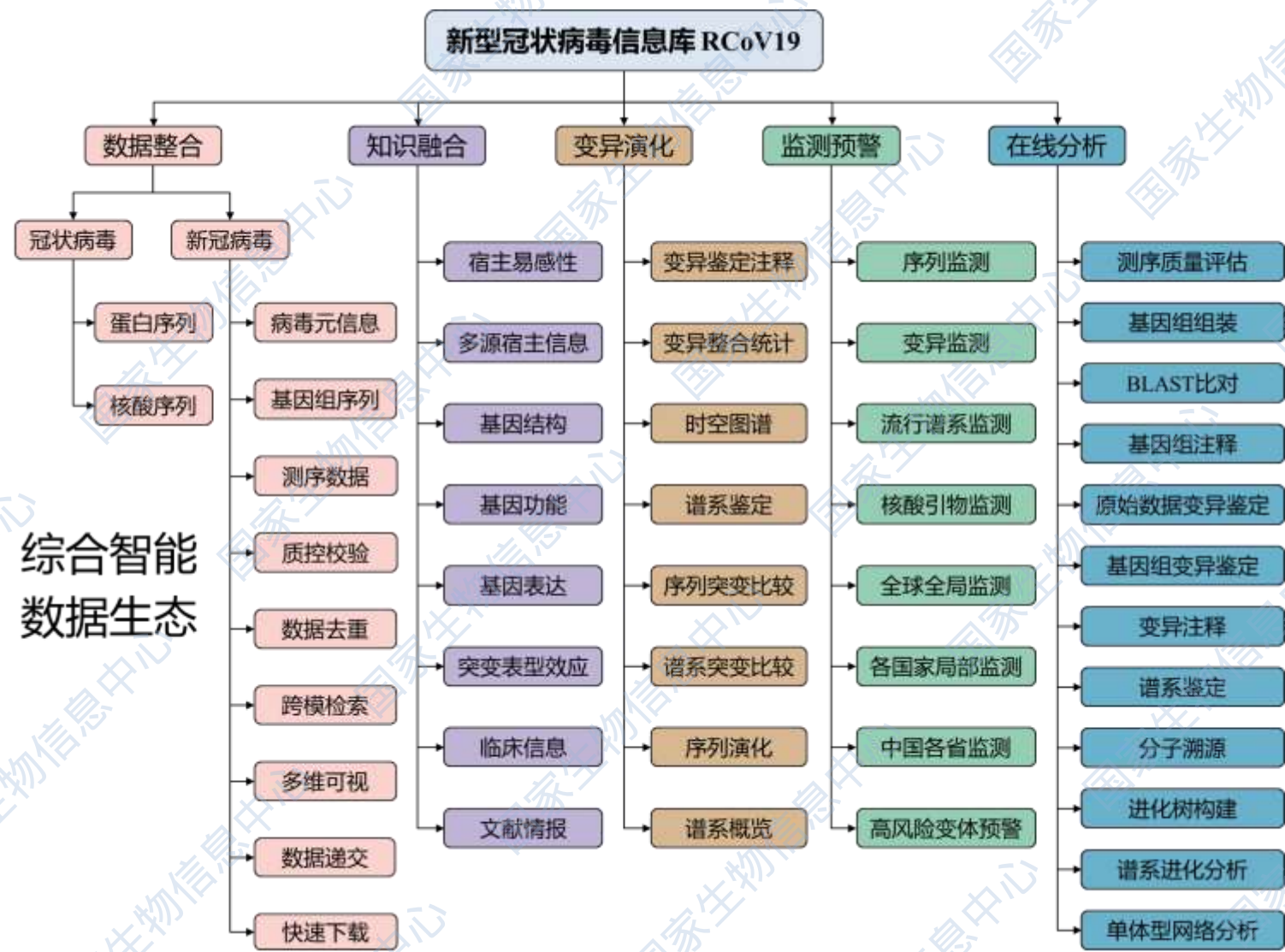


迭代更新

变异信息丰富
实时动态更新

<https://ngdc.cncb.ac.cn/ncov/>

当前规模最大、资源最丰富的新冠病毒研究公共平台



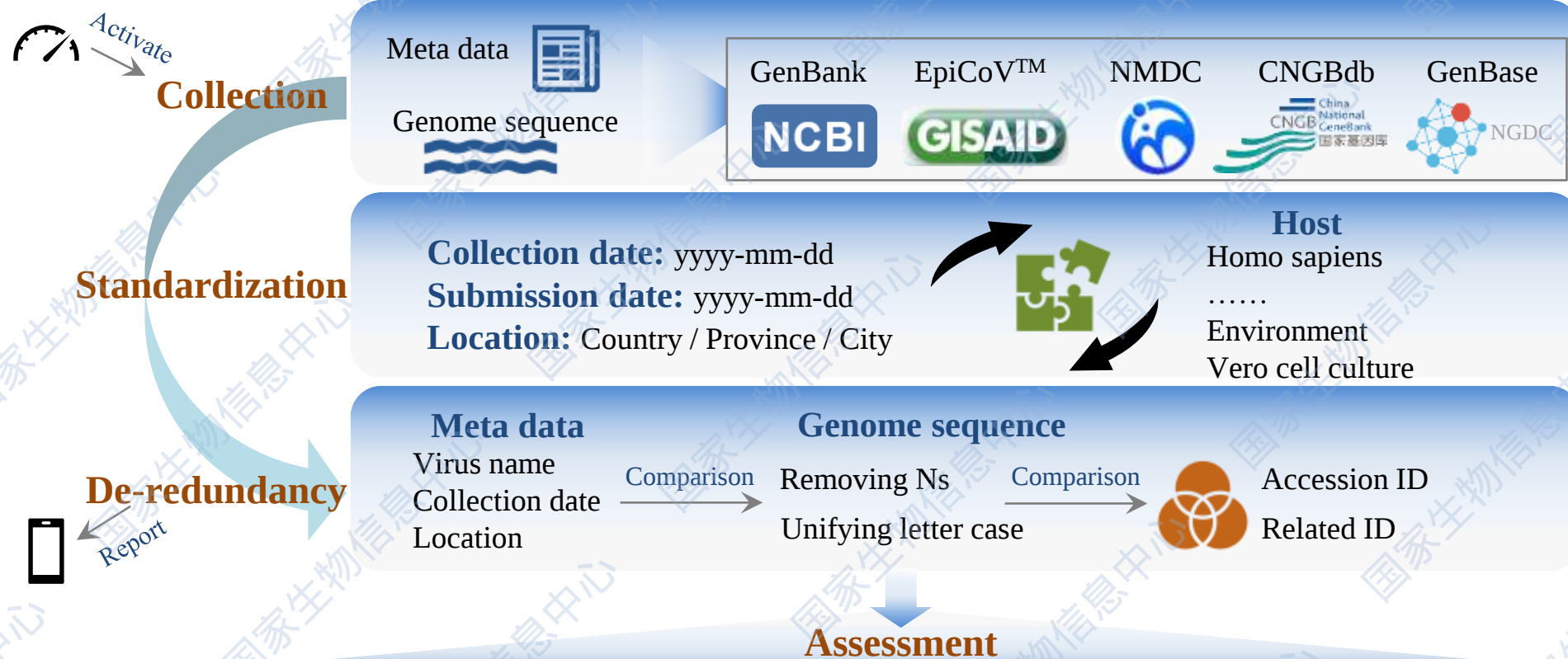
- 五大功能类
- 五十个功能组件
- 千万量级数据融合



“数据-信息-知识-应用”全链条

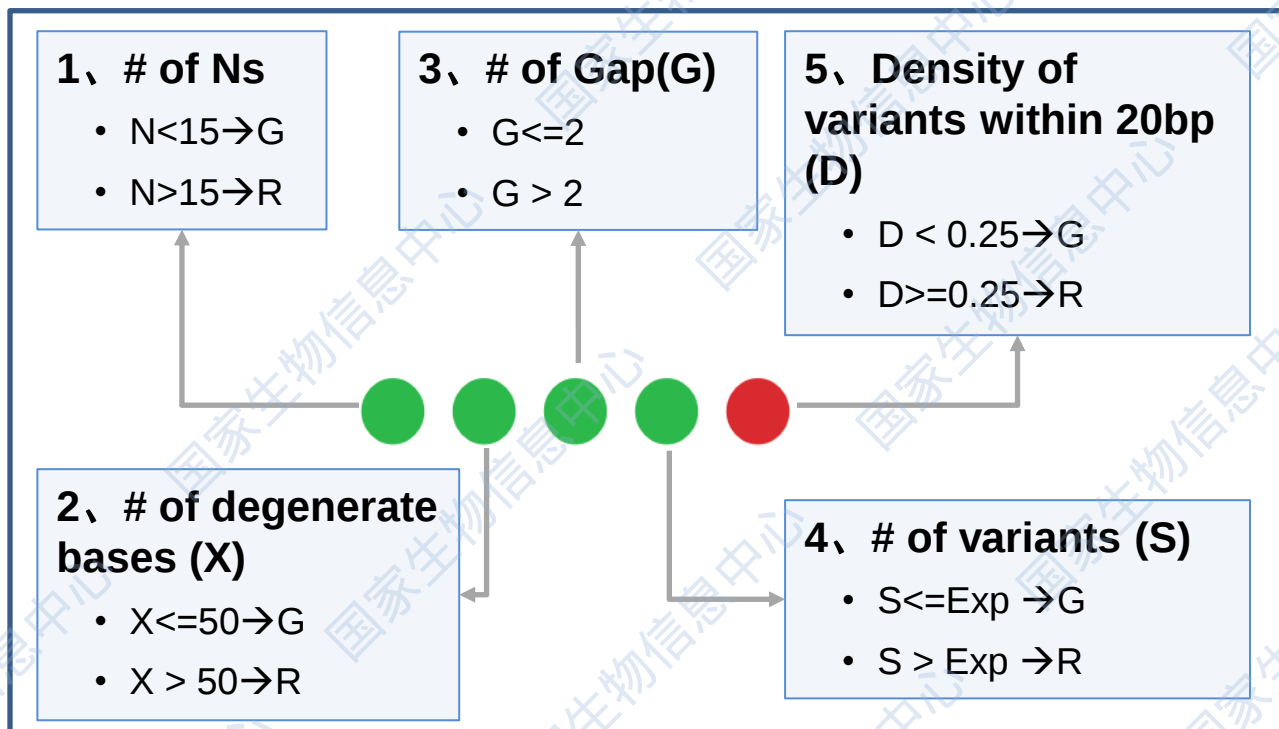
Genomics Proteomics Bioinformatics 2023

自动化审编、标准化整合模型



Genome completeness		N(s)	Degenerate base(s)	Gap(s)	Mutation(s)	Mutation Density	Sequence quality
Complete	●	<=15	<=50	<=2	<=Expected+1	<=0.25	High quality
Partial	●	>15	>50	>2	>Expected+1	>=0.25	Low quality

序列质量完整性评估



Nuc.Completeness	Sequence Length	Sequence Quality ?	Quality Assessment ?
Complete	29834	High	● ● ● ● ●
Complete	29782	High	● ● ● ● ●
Complete	29782	Low	● ● ○ ○ ○
Complete	29782	Low	● ● ○ ○ ○
Complete	29782	High	● ● ● ● ●

~48% complete & high quality

■ 数据统计与发布

50004651

冠状病毒序列

17682698

新冠病毒株

17787159

新冠病毒序列

3341

数据递交单位

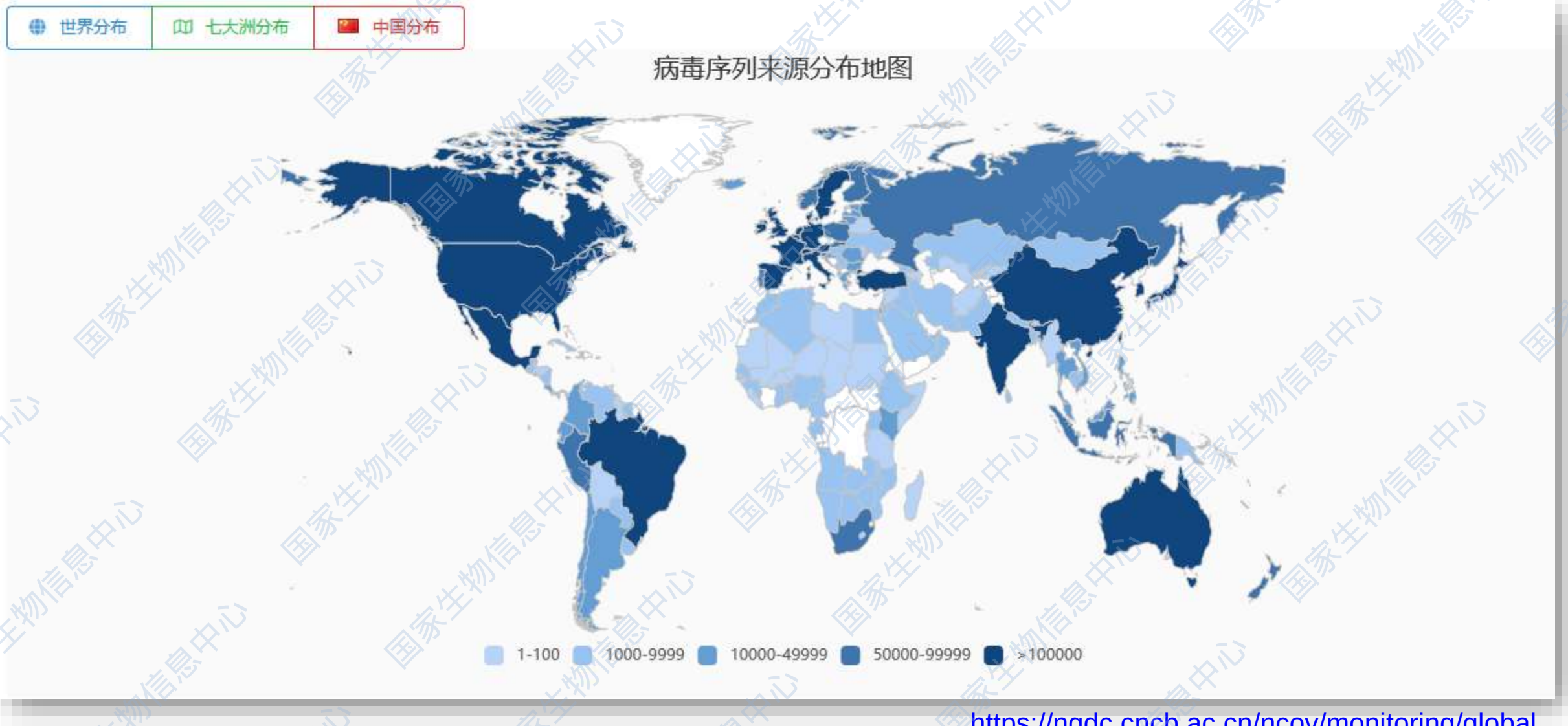
13072

样本采集单位

9676

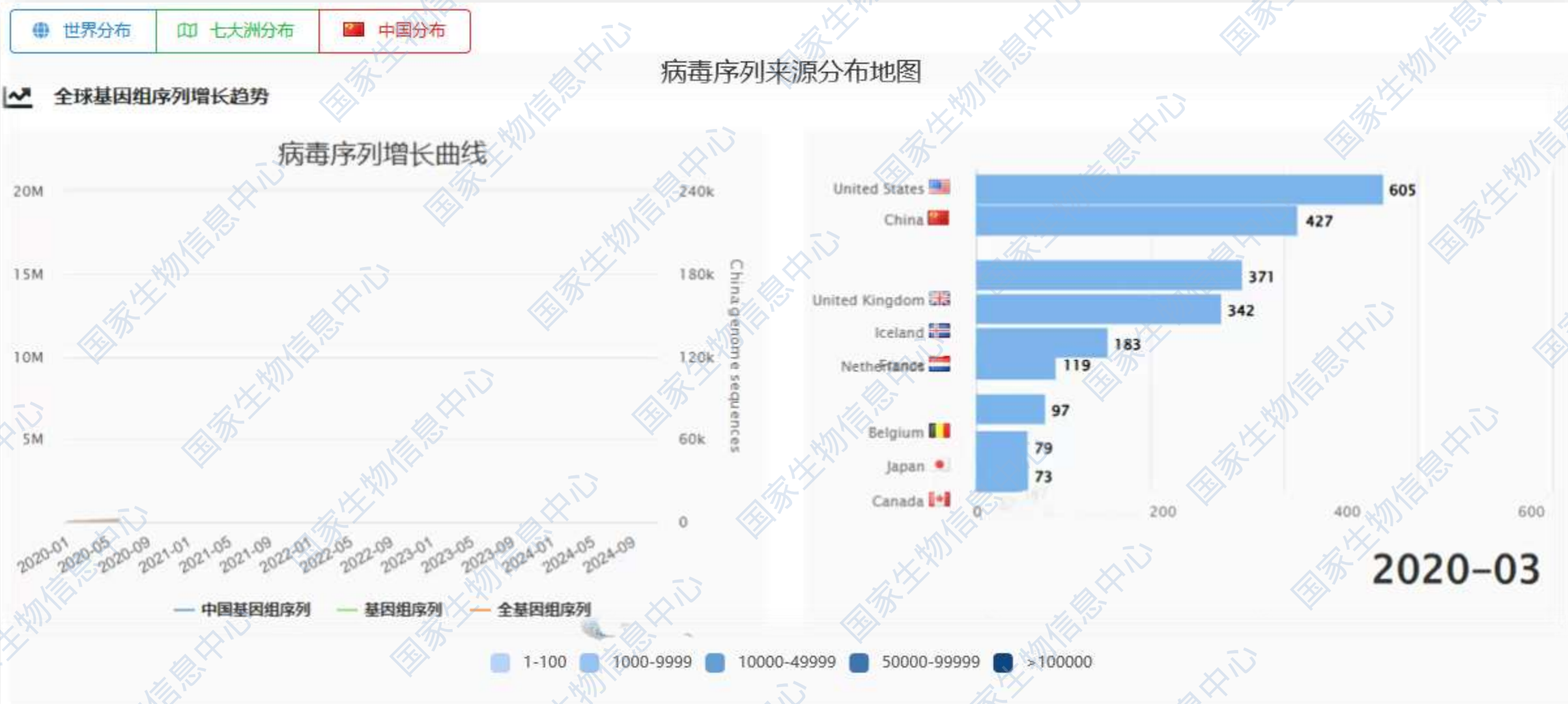
采样地点

序列分布及增长



<https://ngdc.cncb.ac.cn/ncov/monitoring/global>

序列分布及增长



<https://ngdc.cncb.ac.cn/ncov/monitoring/global>

序列综合检索

☐ 合并相同序列 ⓘ

☐ 查看最近更新

序列号

accession id, multiple id splits by comma(,)

谱系 ⓘ

WHO谱系命名

国家/地区

宿主

序列完整度

全部

质量评估

全部

数据库

全部

采样日期 (开始)

📅 采样日期 (开始)

取消选择

选择列 ▾

数据下载 ▾

基因组注释

BLAST比对

序列突变位点比较

下载选定序列

下载所有表格数据

下载检索结果数据

FTP序列下载

Show 20 entries

Previous

1

2

3

4

5

...

889358

Next

☐ 全选	序列号	病毒株名 ⓘ	数据来源	相关ID	原始数据	谱系 ⓘ	序列完整度	序列长度	序列质量 ⓘ	质量评估 ⓘ
☐	C_AA087738.1	SARS-CoV-2/human/CHN/GSTSCDC-101715/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087737.1	SARS-CoV-2/human/CHN/GSTSCDC-101714/2024	GenBase			JN.1	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087736.1	SARS-CoV-2/human/CHN/GSTSCDC-101713/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087735.1	SARS-CoV-2/human/CHN/GSTSCDC-101712/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087734.1	SARS-CoV-2/human/CHN/GSTSCDC-101711/2024	GenBase			JN.1	Complete	29757	🟢	🟢🟢🟢🟢
☐	C_AA087733.1	SARS-CoV-2/human/CHN/GSTSCDC-101710/2024	GenBase			JN.1	Complete	29771	🟢	🟢🟢🟢🟢
☐	C_AA087732.1	SARS-CoV-2/human/CHN/GSTSCDC-101709/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087731.1	SARS-CoV-2/human/CHN/GSTSCDC-101708/2024	GenBase			JN.1	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087730.1	SARS-CoV-2/human/CHN/GSTSCDC-101707/2024	GenBase			JN.1	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087729.1	SARS-CoV-2/human/CHN/GSTSCDC-1011706/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087728.1	SARS-CoV-2/human/CHN/GSTSCDC-101705/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087727.1	SARS-CoV-2/human/CHN/GSTSCDC-101704/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087726.1	SARS-CoV-2/human/CHN/GSTSCDC-101703/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢
☐	C_AA087725.1	SARS-CoV-2/human/CHN/GSTSCDC-101702/2024	GenBase			BA.2	Complete	29772	🟢	🟢🟢🟢🟢

序列完整度

Complete (15530543)

Partial (2256616)

序列质量

Low (8962459)

High (8701249)

数据来源

GenBank (8856474)

GISAI (16977517)

NMDC (287)

CNCBdb (87)

Genome Warehouse (744)

GenBase (60947)

WHO谱系命名

Omicron (9702500)

Delta (4730391)

Alpha (1247648)

Gamma (139461)

Epsilon (83166)

Iota (51860)

Beta (45520)

Mu (18726)

Lambda (11500)

谱系

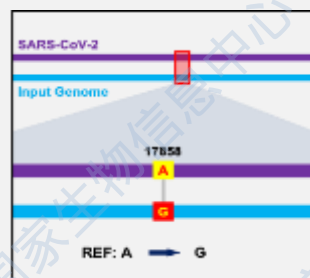
BA.2 (1351255)

B.1.1.7 (1227603)

<https://ngdc.cncb.ac.cn/ncov/genome/search>

基因组变异自动化鉴定与注释

快速、自动化

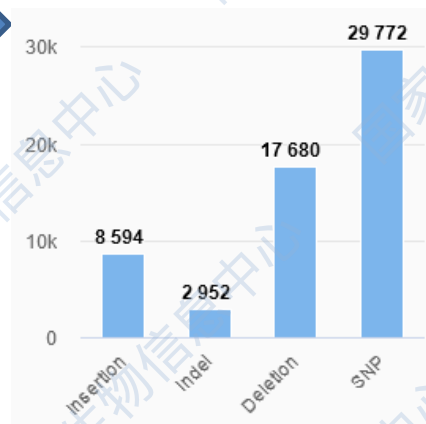


变异鉴定

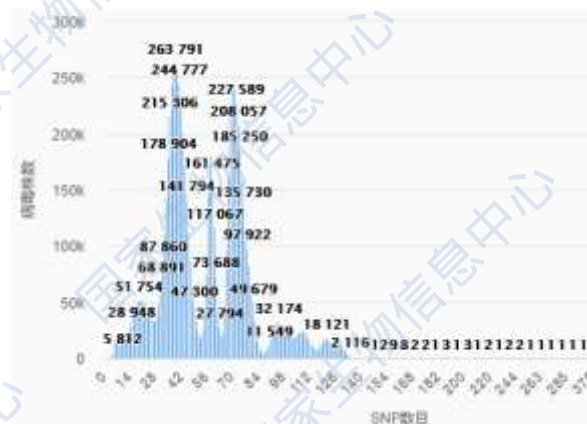
Reference	Start Position	End Position	Base
SARS-CoV-2	1	1	T
SARS-CoV-2	10	10	C
SARS-CoV-2	1	2	TT
SARS-CoV-2	1	2	A
SARS-CoV-2	490	490	C
...	...	...	...

变异注释

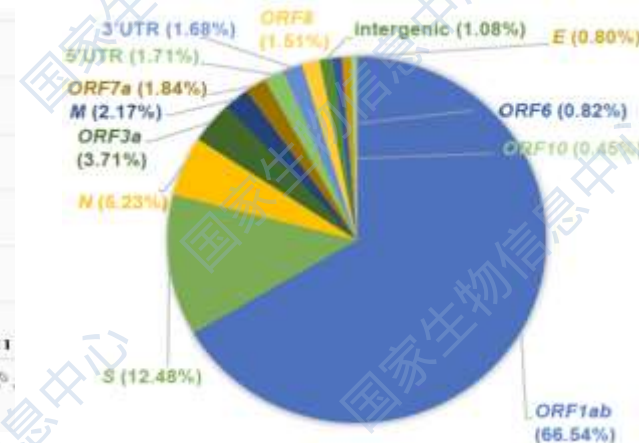
采样日期	病毒株名	SNP 数目	Insertion 数目	Deletion 数目	Indel 数目	相似度	质量评估分级	注释统计	氨基酸数目	最后更新	文件下载 (GFF3)
2020-11-27	hCoV-19/USA/NY-NYCPHL-001467/2020 (EPI_ISL_683826)	19	0	2	0	99.93	C	21(synonymous_variant:8; intergenic_variant:2; missense_variant:10; upstream_gene_variant:1)	10	2020-12-13 20:40:32	2019-nCoV_EPI_ISL_683826_variants.gff3



变异类型



变异数量分布



变异注释统计

数据分析结果被下载46,350次，服务用户在线分析18000+次 (截至2024年10月20日)

突变位点注释与监测

Q 搜索

基因/区域名称

变异病毒株数

1

8%

搜索

重置

CNCB-NGDC Public Download Server

Current Directory: Home / GVM / Coronavirus / vcf

Name

2019-nCoV_total.vcf.gz

Mutation.gz

效应类型

所有序列突变文件

↓

Show 5 entries

基因组文件

下载所有表格数据

Search:

基因组位置	基因/区域名称	变异病毒株数	碱基变化及病毒株数	变异类型	变异注释类型	基因名.CDS位置.序列变化	蛋白名称.位置.氨基酸变化	效应类型	时间方差	地域方差	最后更新
23403	S	8217971	A>M:1	SNP	coding_sequence_variant	S:c.1841gAt>gMt	S:p.614-	MODIFIER	0.023434	0.0034831	2024-10-21 08:37:39
23403	S	8217971	A>C:20	SNP	missense_variant	S:c.1841gAt>gCt	S:p.614D>A	MODERATE	0.023434	0.0034831	2024-10-21 08:37:39
23403	S	8217971	A>G:8216352	SNP	missense_variant	S:c.1841gAt>gGt	S:p.614D>G	MODERATE	0.023434	0.0034831	2024-10-21 08:37:39
23403	S	8217971	AT>G:1	Indel	frameshift_variant	S:c.1841-1842gAT>gG	S:p.614D>X	HIGH	0.023434	0.0034831	2024-10-21 08:37:39
23403	S	8217971	A>R:1589	SNP	coding_sequence_variant	S:c.1841gAt>gRt	S:p.614-	MODIFIER	0.023434	0.0034831	2024-10-21 08:37:39

注：变异注释是基于8242613条高质量人源序列的29799个变异位点计算得出。

<https://ngdc.cncb.ac.cn/ncov/variation/annotation>

突变位点信息追踪

时间频率

地域频率

变异位点动态曲线

重 相关病毒株

由于使用权限的原因，GISAID数据库的序列不能在此被下载，请登录GISAID的网站获取。

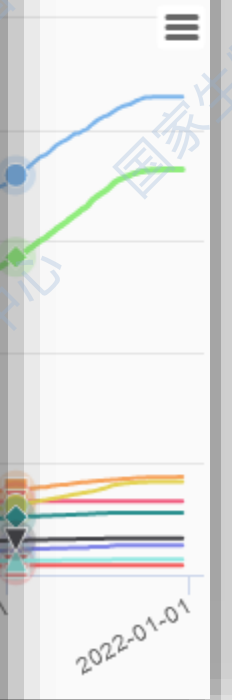
Show 5 entries

选择列

Search:

Isolates

病毒株名	序列号	数据来源	相关ID	谱系	序列完整度	序列长度	序列质量	质量评估	宿主	采样日期	采样地
hCoV-19/Canada/ON-KHS-04354-v1/2024	EPI_ISL_19435261	GISAID		KP.3.1.1	Complete	29703			Homo sapiens	2024-09-12	Canada Ontario
AK-USAFSAM-S22671	PP977224.1	GenBank	EPI_ISL_19226133	JN.1	Complete	29754			Homo sapiens	2024-06-10	United States Alaska
hCoV-19/USA/TN-SPHL-2079/2024	EPI_ISL_19198614	GISAID	PQ095704.1	JN.1.9	Complete	29694			Homo sapiens	2024-02-14	United States Tenne
hCoV-19/USA/TN-SPHL-2049/2024	EPI_ISL_19198596	GISAID	PQ095685.1	JN.1.9	Complete	29694			Homo sapiens	2024-02-15	United States Tenne



<https://ngdc.cncb.ac.cn/ncov/variation/annotation/variant/23403>

突变群体发生频率监测

频率

默认 0.0005 0.01 0.1 0.5

按方差排序

默认 升序 降序

基因/区域名称

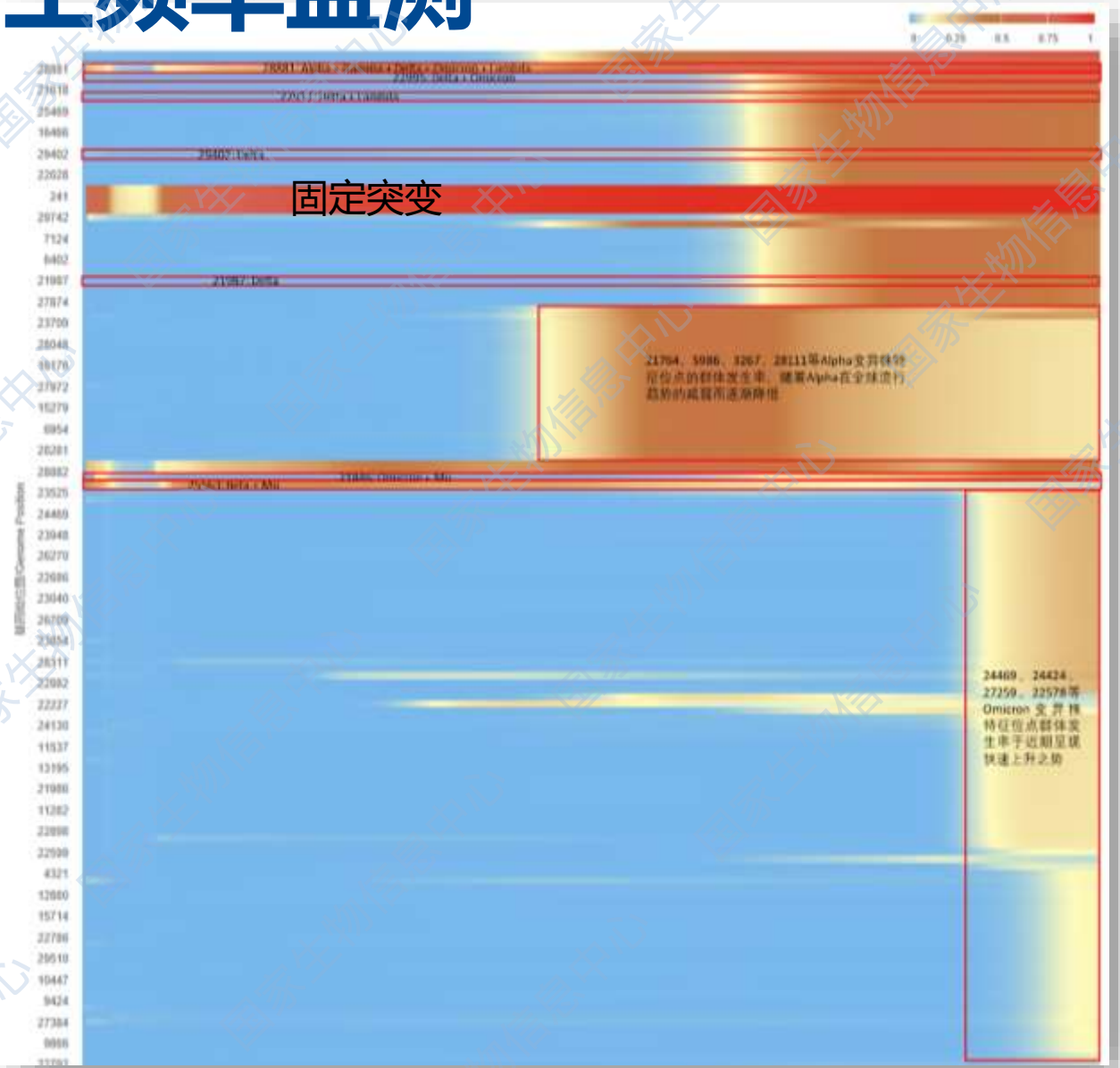
全部 5'UTR Intergenic ORF1ab S ORF3a E M ORF6 ORF7a ORF10 3'UTR

变异类型

全部 intergenic_variant upstream_gene_variant start_lost coding_sequence_inframe_deletion frameshift_variant missense_variant stop_gained synonymous_variant protein_altering_variant inframe_insertion - stop_lost stop_retained_variant downstream_gene_variant

转录调控及酶切位点

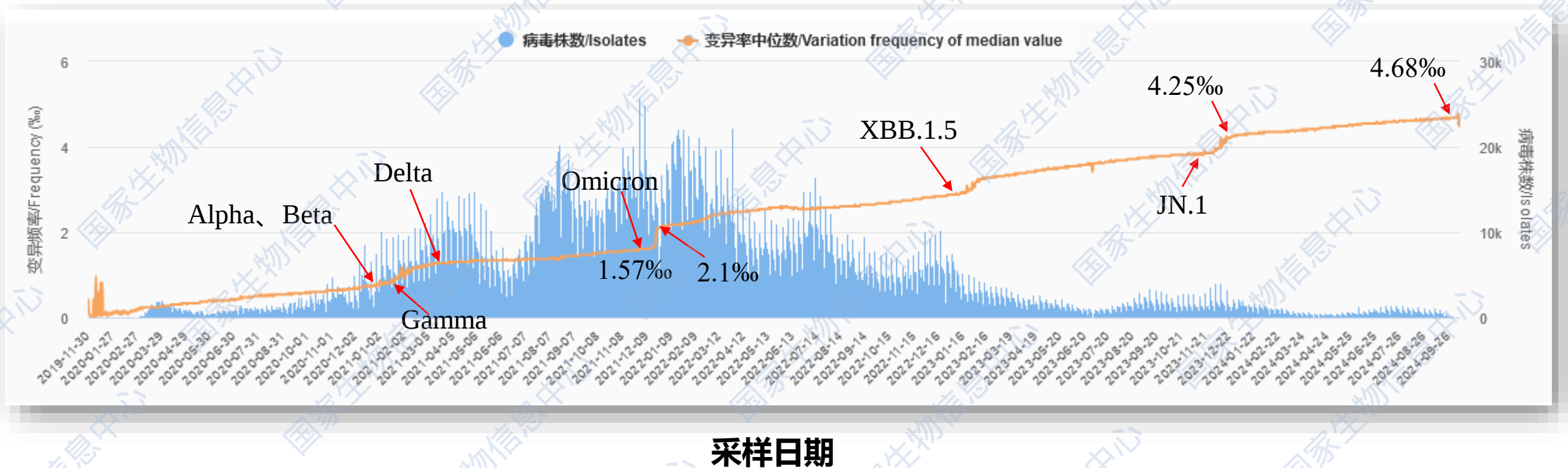
全部 TRS:orf1ab TRS:S TRS:ORF3a TRS:E TRS:M TRS:ORF6 TRS:ORF7a IRS:ORF8 TRS:N TRS:ORF10 Furin:S



<https://ngdc.cncb.ac.cn/ncov/variation/heatmap>

病毒基因组平均突变率监测

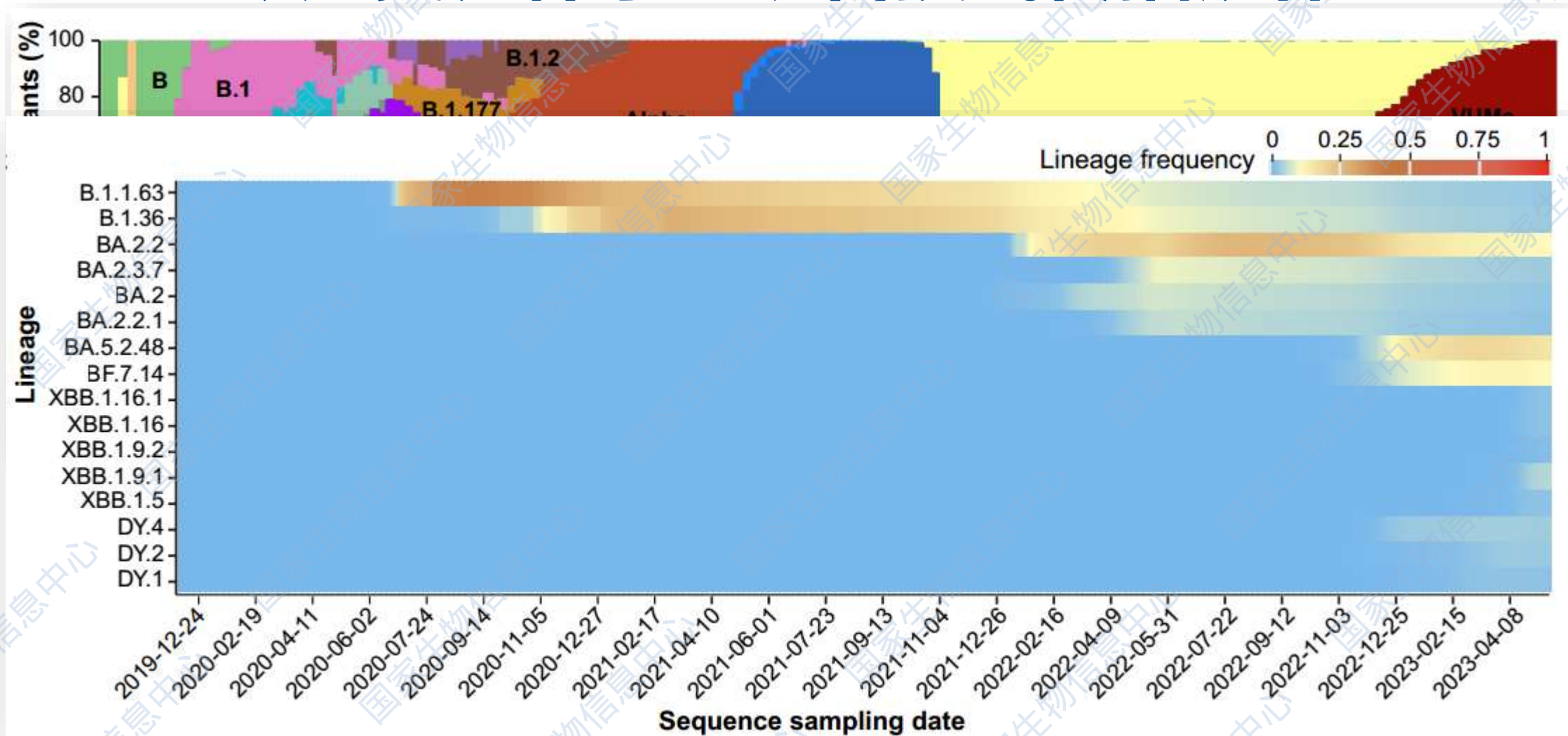
变异解析与监测



<https://ngdc.cncb.ac.cn/ncov/monitoring/global>

新冠变异株系呈现替换式传播规律

变异解析与监测



<https://ngdc.cncb.ac.cn/ncov/monitoring/global>

中国全域新冠病毒序列及支系分布

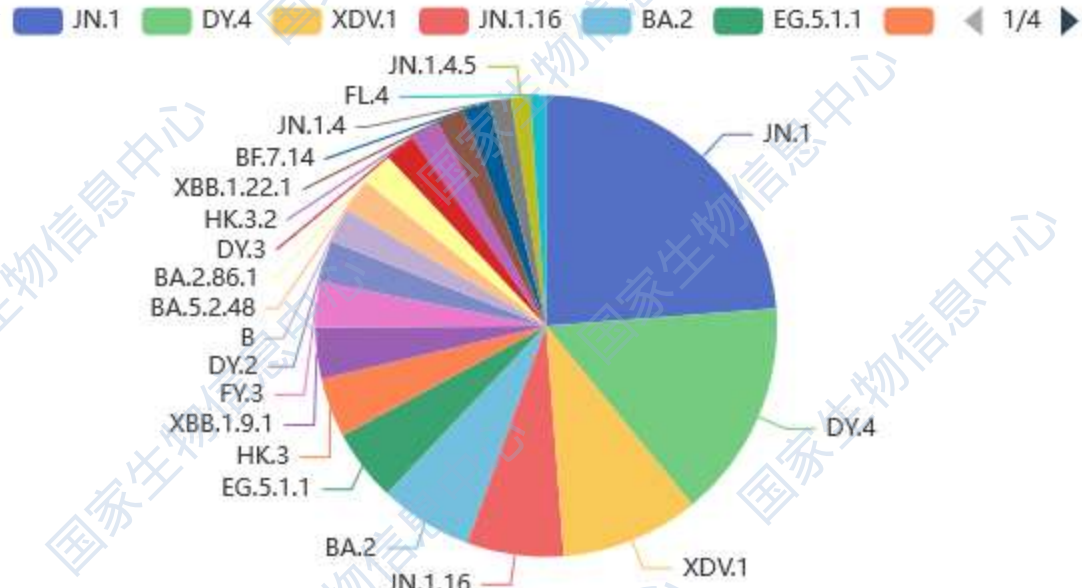
变异解析与监测

采样日期 (开始) 2019-12-01 采样日期 (结束) 2024-10-23 省市区 Hubei 重置

中国 病毒序列来源分布地图 (133273)

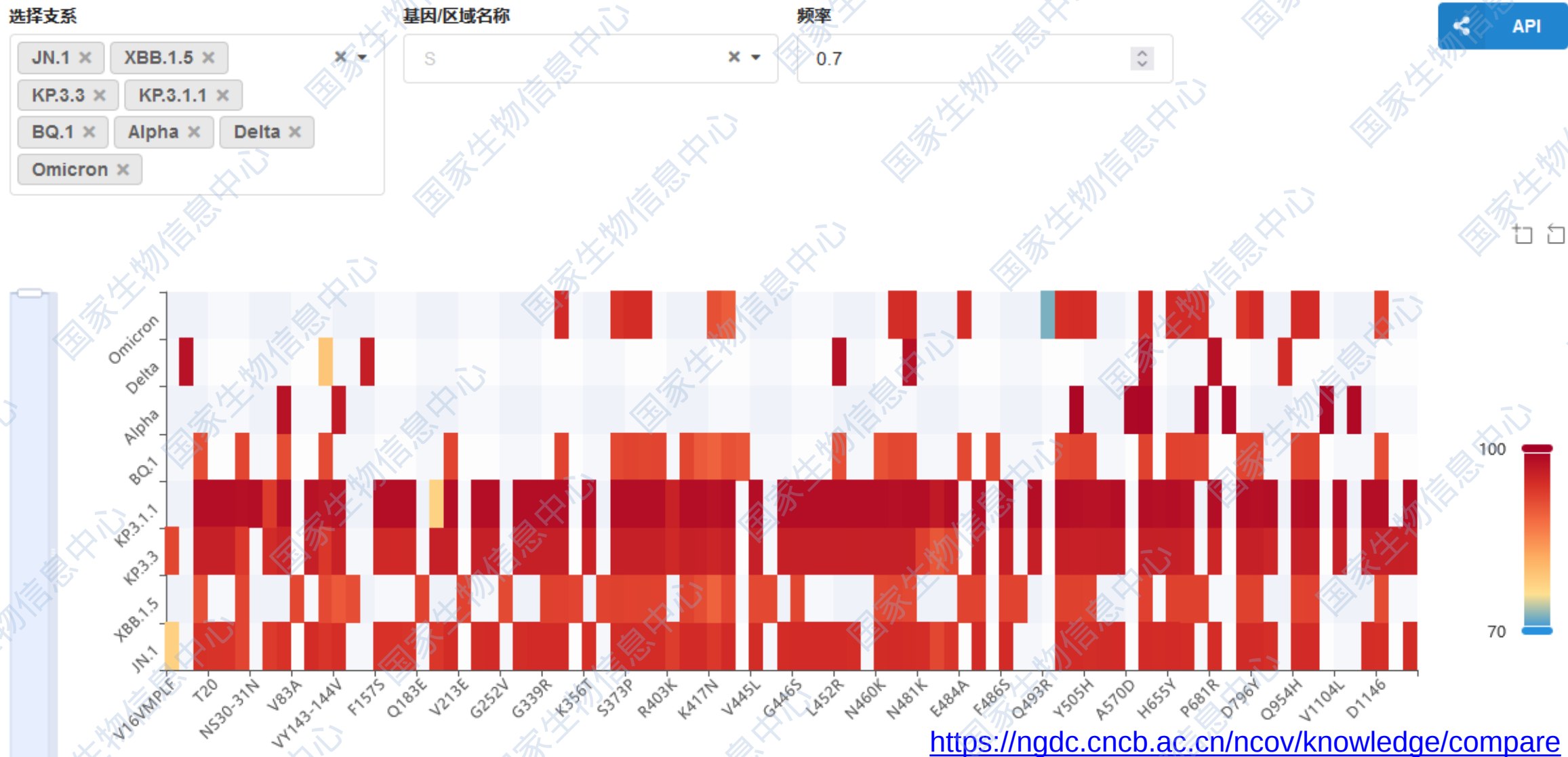


谱系分布 (Hubei: 12189)

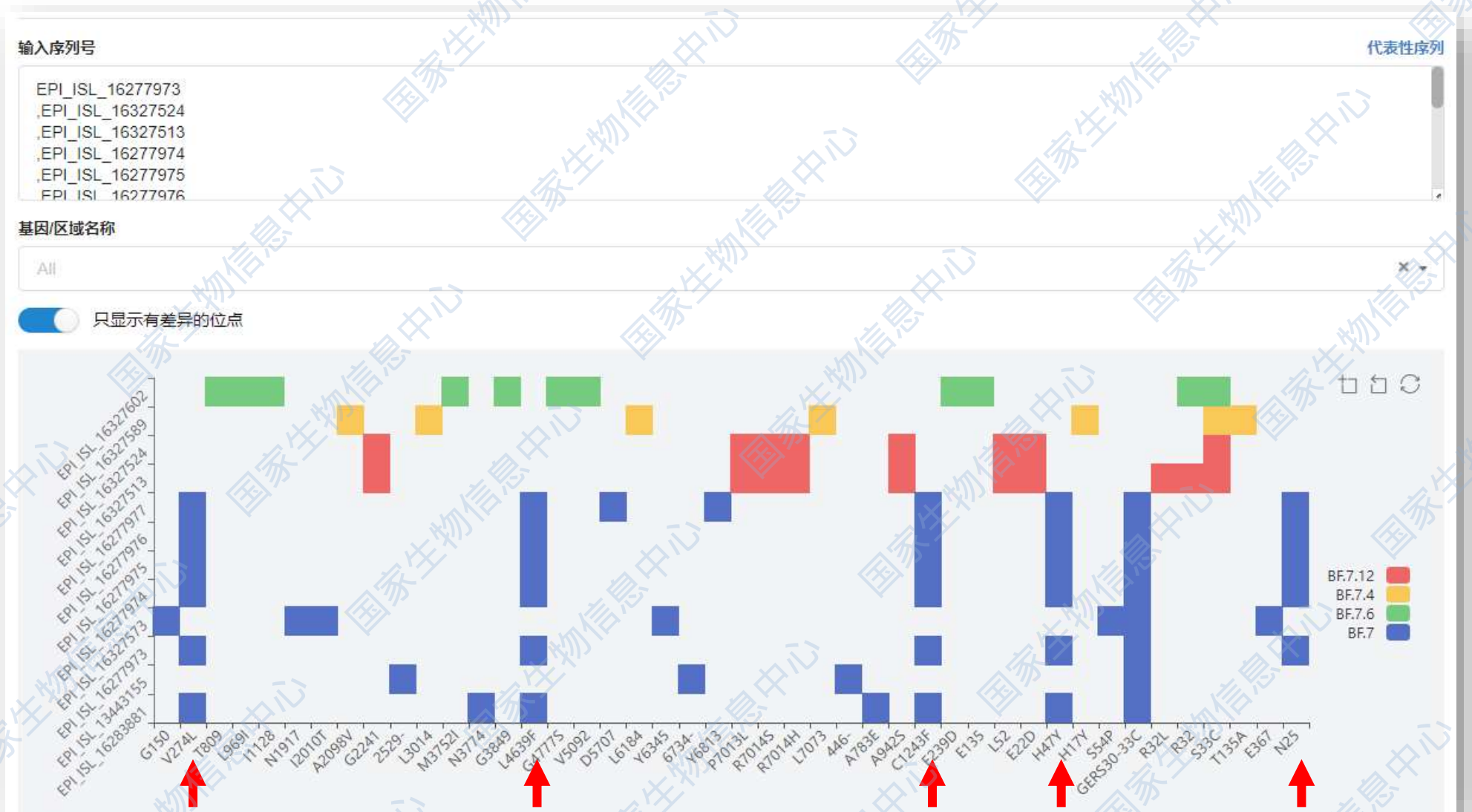


<https://ngdc.cncb.ac.cn/ncov/monitoring/country/China>

重要病毒株突变谱快速比较



重要病毒株突变谱快速比较



国内“BF.7.14”株共性变异位点

Orf1a: V247L

Orf1ab: L4639F

S: C1243F

Orf7a: H47Y

Orf10: N25

<https://ngdc.cncb.ac.cn/ncov/variation/sequence/compare>

重要突变效应知识融合

概览

感染力/传播力

抗体抗性

药物抗性

T细胞表位

致病性

结构稳定性

E

M

N

S

ORF10

ORF1ab

ORF3a

ORF6

ORF7a

ORF8

氨基酸突变

基因

效应

资源

Search 氨基酸突变

Search 基因

Search 效应

Search 资源

A1015D

S

I

S

P

B

概览

感染力/传播力

抗体抗性

药物抗性

T细胞表位

致病性

结构稳定性

Show 10 entries

Search:

氨基酸突变	注释基因	基因组位置	碱基变化	是否ACE2结合	是否关键互作氨基酸	效应	描述	验证方法	文献
F486L	S	23018 23020	T>C T>G;T>A	Yes	Yes	Drug resistance	Resistant to Ensovibep	Cell Assay	35502202
F486V	S	23018	T>G	Yes	Yes	Drug resistance	Resistant to Ensovibep	Cell Assay	35502202
A3436T	ORF1ab	10571	G>A			Drug resistance	Mutation associated with potential small molecule inhibitor resistance		36334368
A3436V	ORF1ab	10572	C>T			Drug resistance	Mutation associated with potential small molecule inhibitor resistance		36334368
A3454T	ORF1ab	10625	G>A			Drug resistance	Mutation associated with potential small molecule inhibitor resistance		36334368
A3454V	ORF1ab	10626	C>T			Drug resistance	Mutation associated with potential small molecule inhibitor resistance		36334368

<https://ngdc.cncb.ac.cn/ncov/knowledge/mutation>

被国内外同行高度认可

Brister, James (NIH/NLM/NCBI) [E] <jamesbr@ncbi.nlm.nih.gov>
RE: Re: SARS-CoV-2 data

Sew Yun Shin <sewyshin@mardi.gov.my>
Re: BHBD 2019nCoV Resource

Mohammadreza Mohammad Abadi <mrrn...>
Re: BHBD 2019nCoV Resource

Hello Yiming.

Sorry for the delayed reply. Just as this publication was being discussed, a new consortium of US SARS-CoV-2 data submitters was put together. This is a very large effort that includes CDC, their partners, and a variety of other data submitters. Anyway, this has been a major lift from my end, and I have been focused on organizing NCBI better support NCBI submissions and presentation of data.

I will follow up later in the week.

BTW, I have been very impressed by your web resource, particularly the genome variation calculations. I have a somewhat naïve question: Have you spoken directly to GISAID about your analysis and your GenBank to GISAID accession mapping? I'm wondering if they are upset by this type of usage (it is really helpful obviously).

Sofia KOSSIDA <sofia.kossida@igh.cnrs.fr>
Re: BHBD 2019nCoV Resource

Thanks a lot Yiming for the message and the neat resource !

I forwarded it to my colleagues and more. We were supposed to have our 3-day long evaluation which takes place once per 4 years starting tomorrow, but it was cancelled yesterday. Yes, we are all, more or less, concerned and we will fight it.

Ana Tereza R. de Vasconcelos <atriv@lncc>
Re: BHBD 2019nCoV Resource

Thank you.

Takashi Gojobori <tgojobori@kaust.edu.sa>
Re: BHBD 2019nCoV Resource

Dear Yiming,

Good morning!

Such a valuable for the scientific community.

Stay safe

Best regards

Ana

tion.

n

know.

Best regards

Ana

I hope that you and your colleagues as well as their families are safe in this Covid-19 situation.

The BHBD 2019nCoV Resource is very impactful achievements in such a short period.

Congratulations!

I am going to spread this information over my colleagues and friends.

Thanks much!

Sincerely,
Takashi

--
Takashi Gojobori
Distinguished Professor,
CBRC (Computational Bioscience Research Center),
BESE (Biological and Environmental Sciences and Engineering),
Room 4220, Level 4, Building 3 (Ibn Sina Build –Sea Side)
KAUST (King Abdullah University of Science and Technology),

I have been very impressed by your web resource, particularly the genome variation calculations. —Rodney Brister, NCBI

The BHBD 2019nCoV Resource is very impactful achievements in such a short period. —Takashi Gojobori, KAUST

整体运行效果：被NCBI、Elsevier等链接

贡献及影响

The screenshot displays the SARS-CoV-2 Portal website. The top navigation bar includes links for PSC, USERS, TOOLS, ABOUT BRIDGES, BRIDGES HOME, BRIDGES VIRTUAL TOUR, BRIDGES AI, and COVID-19 RESEARCH. The main header features the SARS-CoV-2 Portal logo and a search bar. Below the header, there are links for HOME, DOCUMENTATION, COMMUNITY, and CONTACT. The central content area is titled 'DOCUMENTATION' and features a large virus icon. A text box provides information about the MMP Team and the SARS-CoV-2 database. The database section mentions that content has been compiled from publicly available resources, including EMBL-EBI, NCBI, DDBJ, and the China National Center for Bioinformation. A list of related publications and other resources is also visible on the left side.

SARS-CoV-2 PORTAL

HOME DOCUMENTATION COMMUNITY CONTACT

DOCUMENTATION

The MMP Team
The SARS-CoV-2 Portal was developed as a part of the ELIXIR2 project, funded by the Research Council of Norway, and made online in April 2020. The Portal is maintained by The Center for Bioinformatics (SfB) at the UiT The Arctic University of Norway. SfB is hosting the UiT node of ELIXIR Norway. Below you will find a short description of who we build and implemented the SARS-CoV-2 database.

The SARS-CoV-2 database
The content of SARS-CoV-2 database has been compiled from a number of publicly available resources such as sequence databases at EMBL-EBI, NCBI, DDBJ and China National Center for Bioinformation. In addition we have used public available data from Johns Hopkins University Coronavirus Resource Centre and literature databases such as Europe PMC and PubMed for curation.

Related publications

- Nature News graphic
- Preprints of research
- COVID-19 Open Research Database
- NIH Office of Data Science
- European Molecular Biology Laboratory
- China National Center for Bioinformation
- CDC SARS-CoV-2
- Nextstrain open-source
- NCBI sequence data

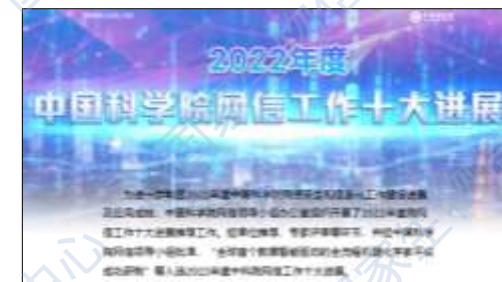
Other

- The Center for Bioinformatics
- Information
- 2019-nCoV
- You are here
- GETTING

Data collections

获奖情况

- 入选2022中科院网信工作十大进展和世界互联网大会—科技之魅领先科技成果
- 北京基因组所新冠肺炎抗疫科研工作入选中科院2021年度科技创新亮点成果、作为亮点工作写入《“一带一路”创新发展报告2021》
- 数据库论文获2020年度中国生物信息十大进展
- 科技攻关团队入选科技部《全国科技系统抗击新冠肺炎疫情先进集体》
- 新冠病毒变异监测与分子溯源研究团队荣获中科院第三届“科苑名匠”称号



纲要

1. 新冠病毒数据与信息库

2. 大规模单体型网络快速**构建工具与追踪系统**

3. 高风险株系精准预警模型及应用平台

4. 新冠病毒在线分析平台

病毒单体型网络

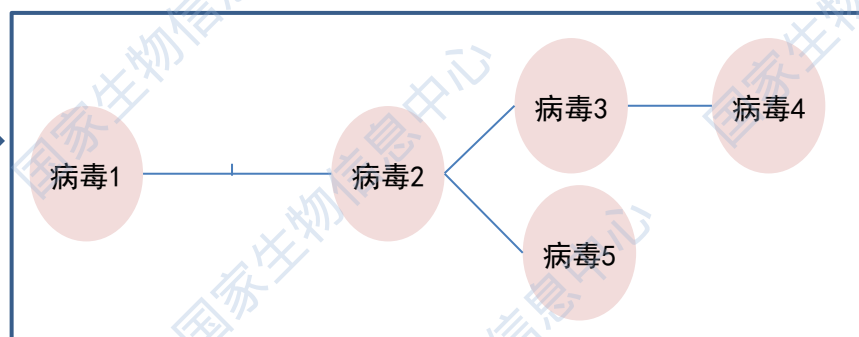
假如有如下5株病毒序列，变异信息如下：

病毒1：TCAGACTTACCTACGTACCTATTGCACCCAATTTACGTACTAGACGA
病毒2：TCA~~A~~ACTTACCTAC~~T~~TACCTATTGCACCCAATTTACGTACTAGACGA
病毒3：TCA~~A~~ACTTACCTAC~~T~~TACCTATTGCACCCA~~T~~TTTACGTACTAGACGA
病毒4：TCA~~A~~ACTTACCTAC~~T~~TACCTATTGCACCCA~~T~~TTTACGTAG~~G~~TAGACGA
病毒5：TCA~~A~~ACTTACCTAC~~T~~TACCTATTGCACCCAATTTAC~~C~~TACTAGACGA

单体型

病毒1（变异0个）：G-G-A-G-C
病毒2（变异2个）：A-T-A-G-C
病毒3（变异3个）：A-T-T-G-C
病毒4（变异4个）：A-T-T-G-G
病毒5（变异3个）：A-T-A-C-C

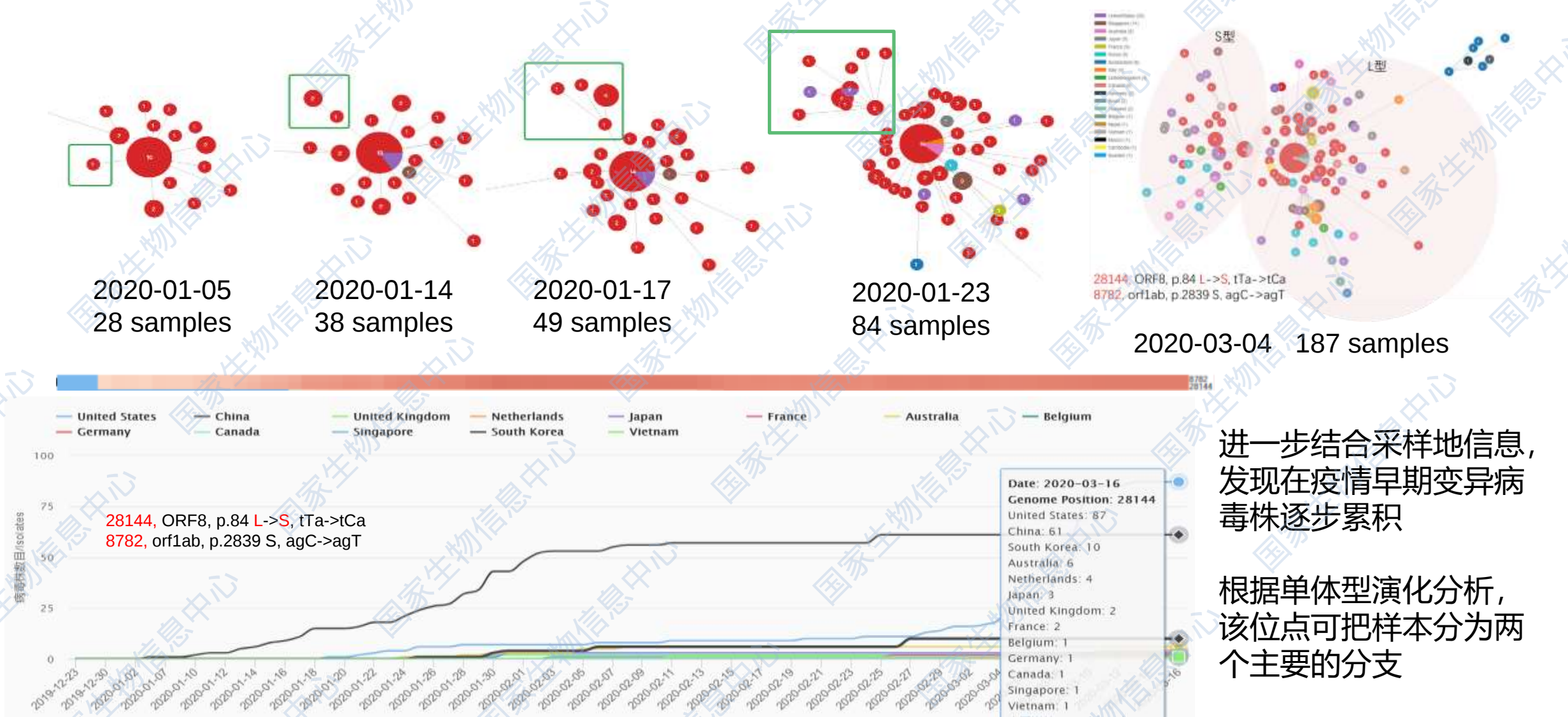
单体型网络演化关系



顶点：单体型
边：演化关系
点的大小：序列数
边长：不同碱基数

- haplotype is a group of variants/genes within an organism that was inherited together from a single parent". Haplotype networks are often used in phylogeographic studies to display genetic variation **within a species** or a group of **closely related species**
<https://www.nature.com/scitable/definition/haplotype-haplotypes-142>

单体型网络演化分析



进一步结合采样地信息，发现在疫情早期变异病毒株逐步累积

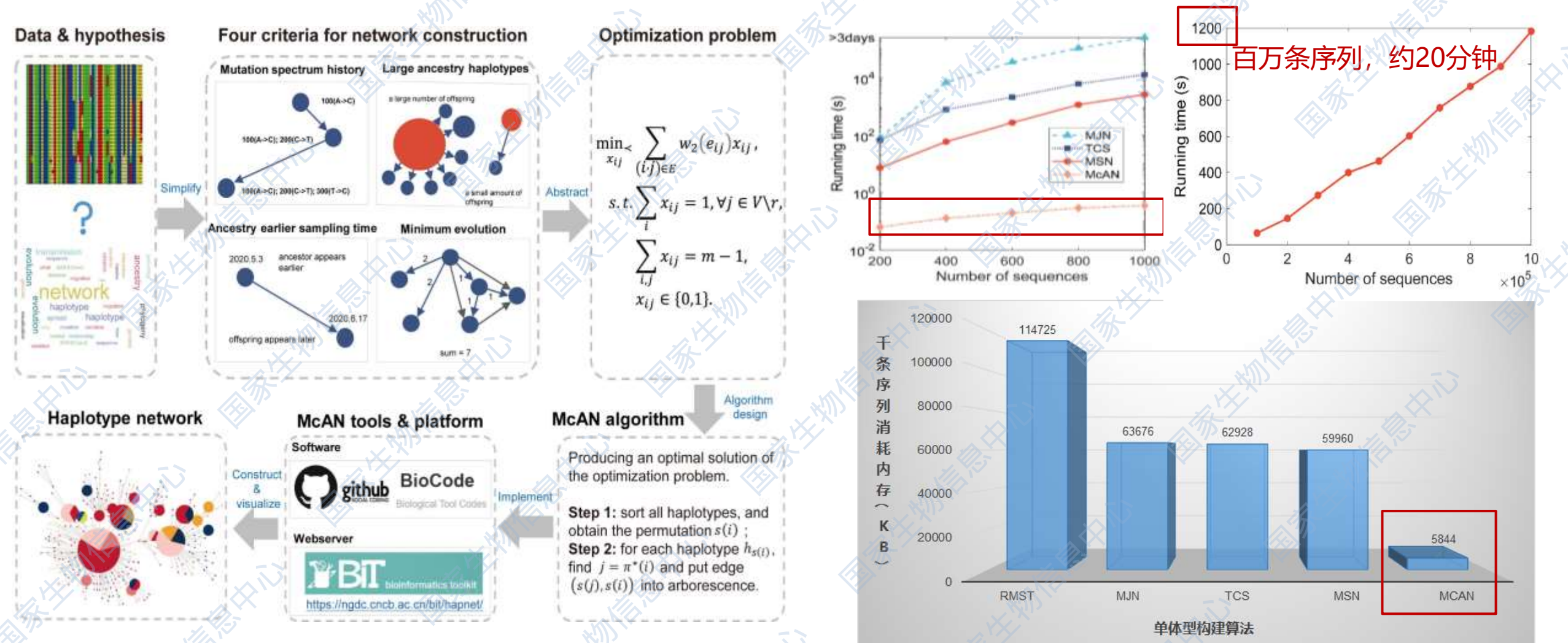
根据单体型演化分析，该位点可把样本分为两个主要的分支

单倍体型网络构建算法软件

- 相对成熟并广泛应用的单体型网络构建方法：MJN、MSN和TCS
- 无法快速地处理海量的病毒基因组数据

方法	优点	缺点	时间复杂度	提出时间	描述
MJN (Median Joining Network)	添加未采样节点	• 运行速度慢 • 内存消耗大 • 没有充分利用样本的采样时间	$O(m^{2.2})$	1999 <i>Mol Biol Evol.</i>	结合最小生成树方法和最大简约法
TCS (Templeton Method)	添加未采样节点		$O(m^5)$	1992 <i>Genetics</i>	计算序列之间的距离，根据距离将序列依次聚合
MSN (Minimum Spanning Network)	与MJN、TCS比相对较快		$O(m^2)$	1999 <i>Genetics</i>	在Kruskal最小数生成树算法的每轮迭代中，添加多条边
McAN (Minimum-cost Arborescence Network)	实现千万级数据网络构建		$O(m^2)$	2023 <i>BiB</i>	基于最小树形图构建单体型网络，无需存储单体型的距离矩阵和计算每对单体型之间的距离

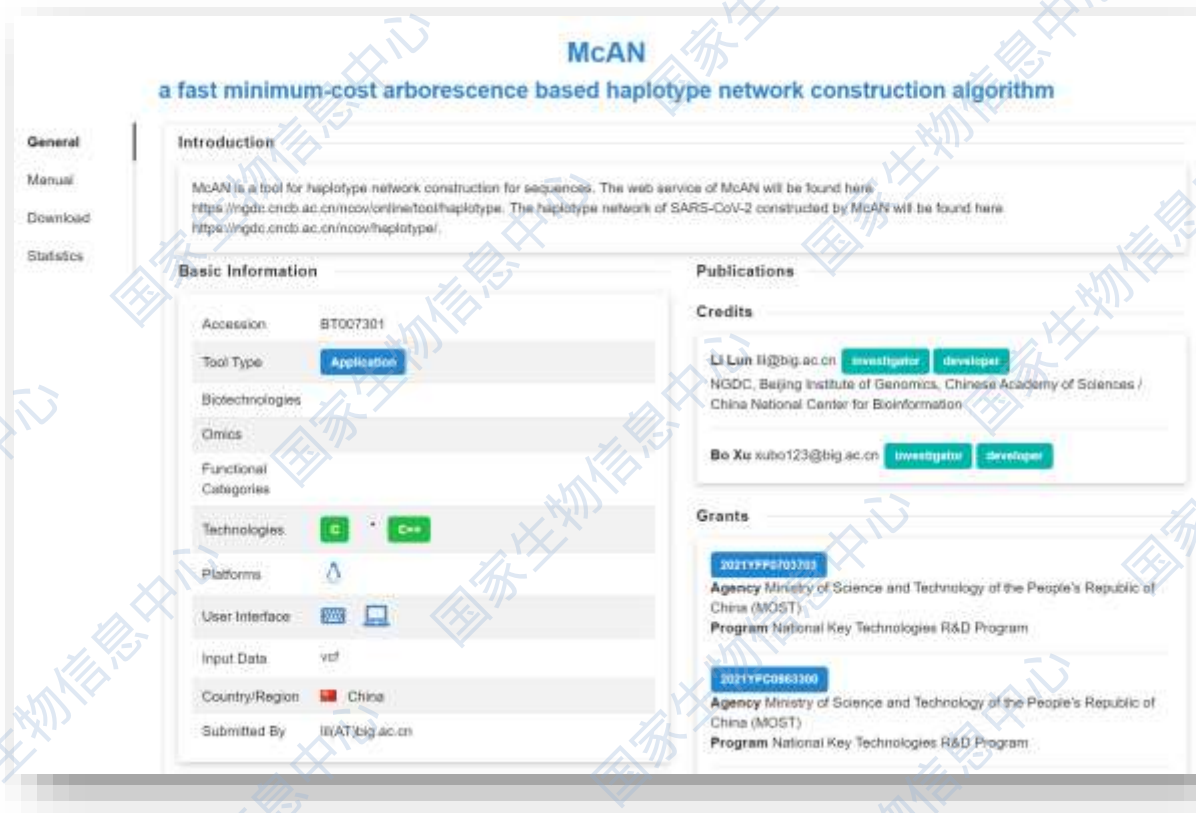
基于最小树形图的单体型网络快速构建算法McAN



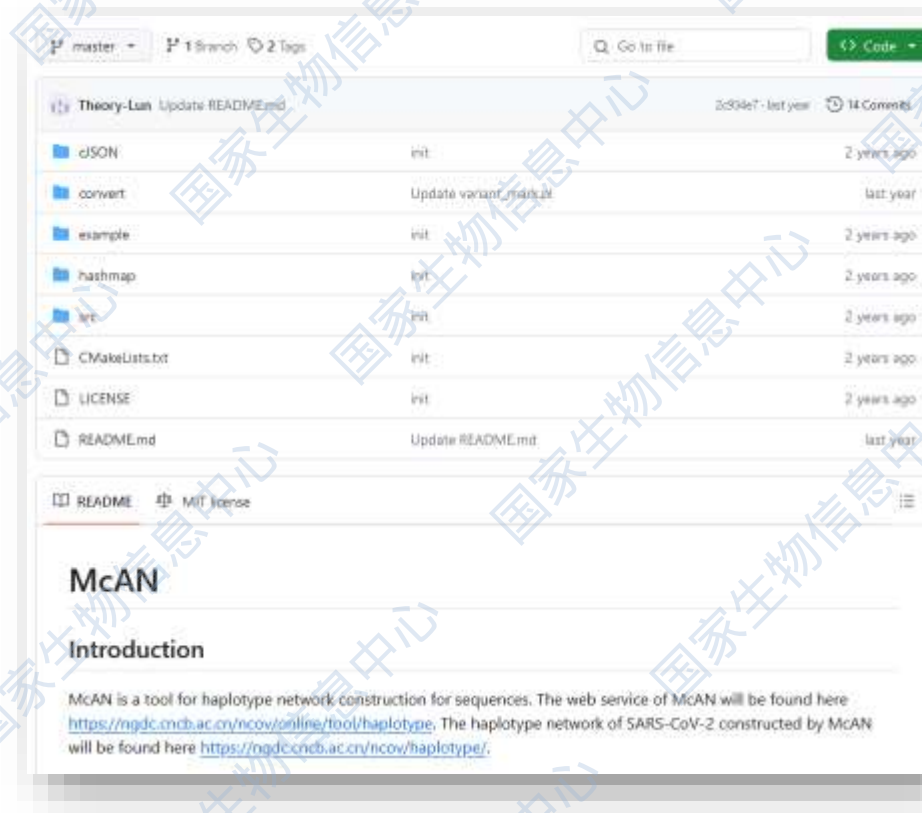
Briefings in Bioinformatics 2023

McAN算法代码开源共享

McAN源代码已公开发布至Github、BioCode，下载量超1.2万次



<https://ngdc.cncb.ac.cn/biocode/tool/7301>



<https://github.com/Theory-Lun/McAN>

单体型网络构建及可视化在线平台

McAN在线工具已部署在BIT为科研人员提供服务超过1700次

功能描述

该工具使用“基于最小树形图的单体型网络构建方法”(McAN)构建单体型网络。需要上传SARS-CoV-2序列变异文件(采用GF或VCF格式)和对应的元信息文件。或者从我们的2019新型冠状病毒信息库(RCoV19)选择已有样本。你可以从我们的信息库中通过设置起止采样时间,采样国家或地区来过滤已有样本,或者以随机方式选择样本。基于构建好的单体型网络,采用Newman方法对单体型划分网络。注意变异文件中3'端非编码区和5'端非编码区的突变将被自动删除。

选择文件格式

☒ 变异文件格式 ☐ VCF文件格式

变异文件 ? 必选项 [变异文件示例](#)

上传变异文件 No file chosen

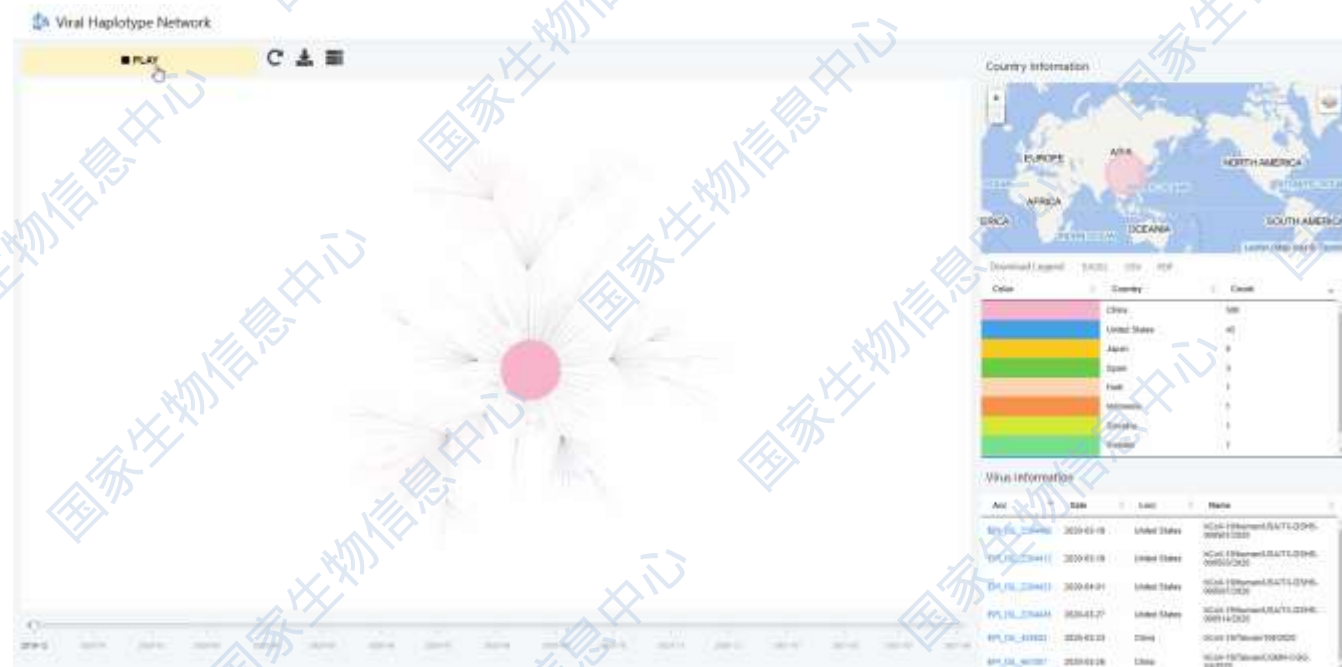
元信息文件 ? 必选项 [元信息示例](#)

上传元信息文件 No file chosen

选择特定数据集 可选项

开始采样时间 结束采样时间 国家

2019-01-01 2021-05-31 China

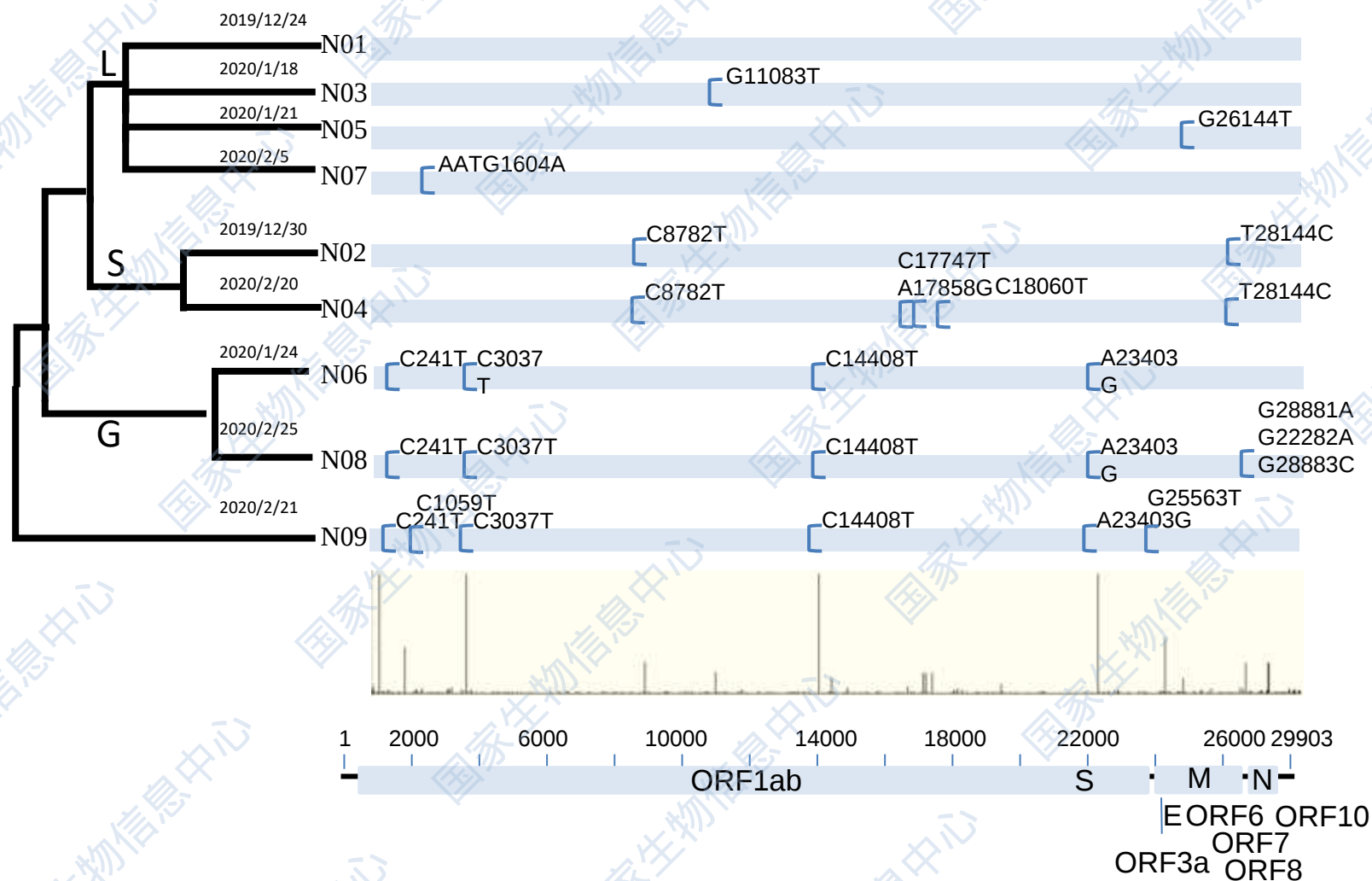
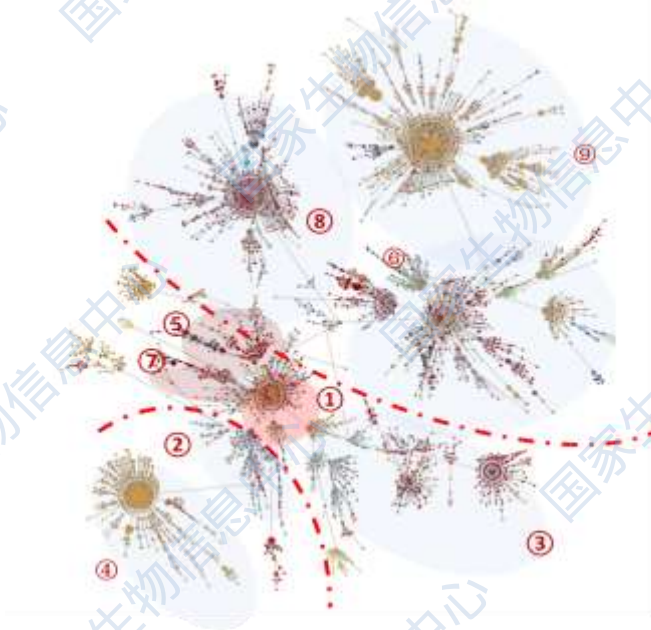


<https://ngdc.cncb.ac.cn/ncov/online/tool/haplotype>

全球疫情早期监测：单体型网络的团簇支系划分

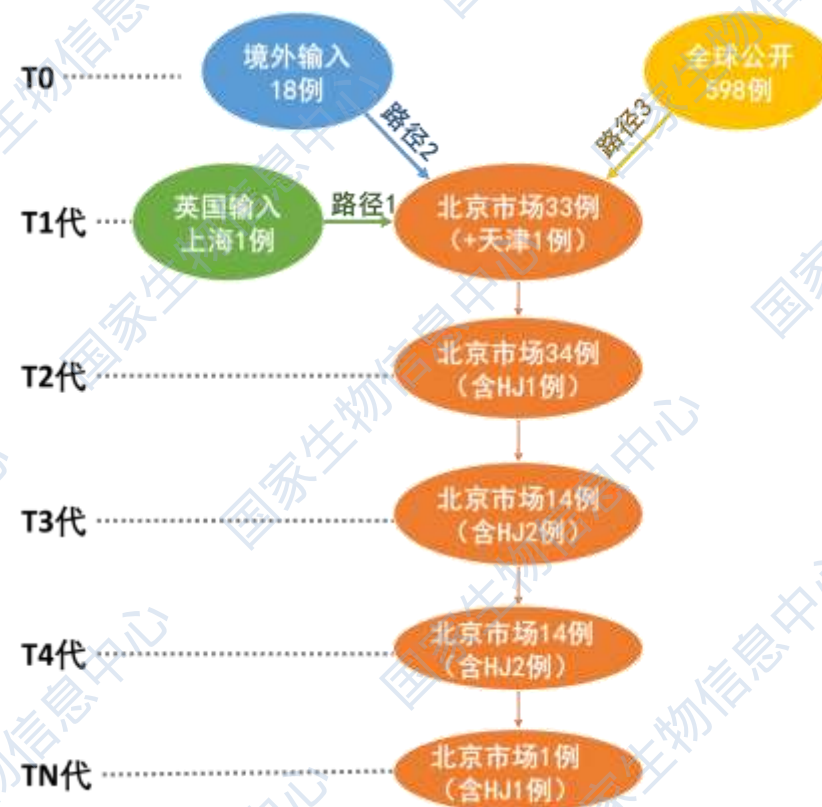
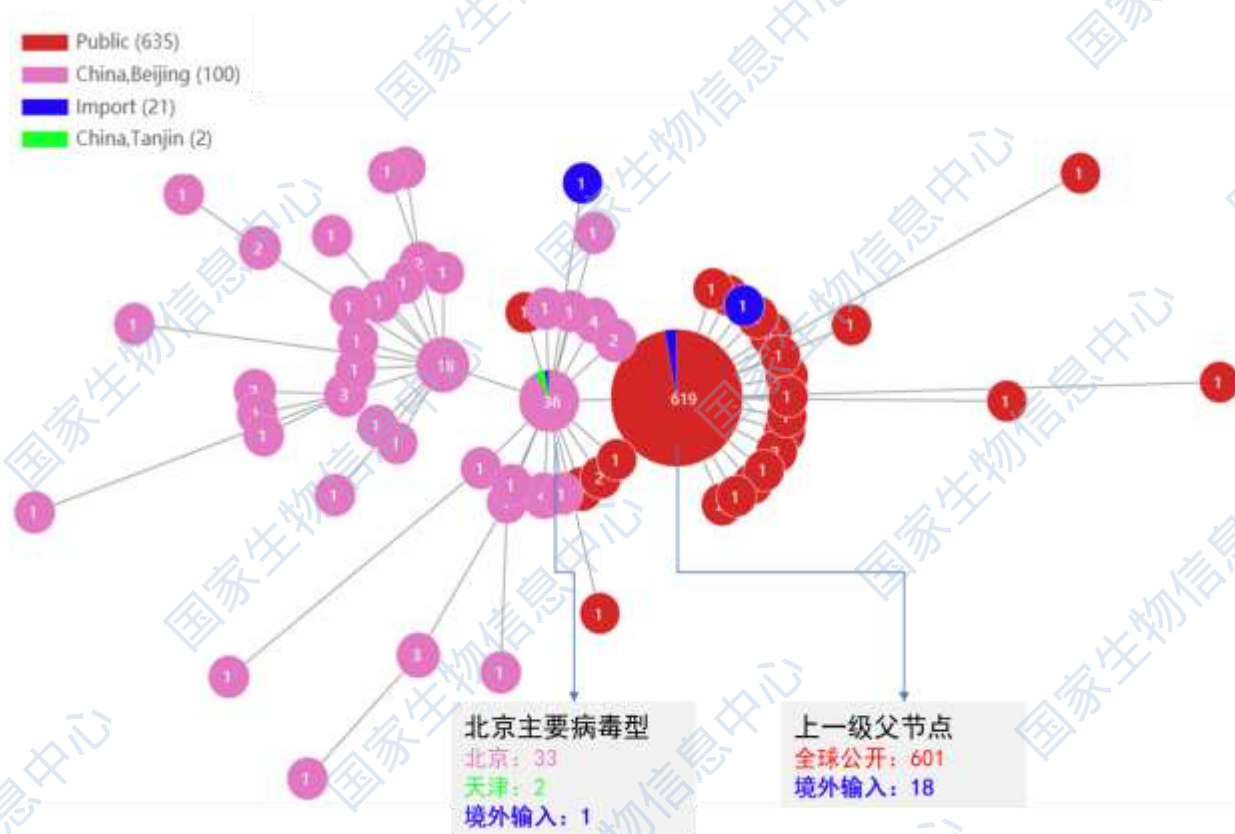
单体型网络分簇依据：

- 结合序列演变进化关系
- 位点突变发生时间
- 共性突变位点重叠情况
- 突变位点群体发生频率



Genomics Proteomics & Bioinformatics 2020

北京市新发地市场新冠病毒分子溯源研究



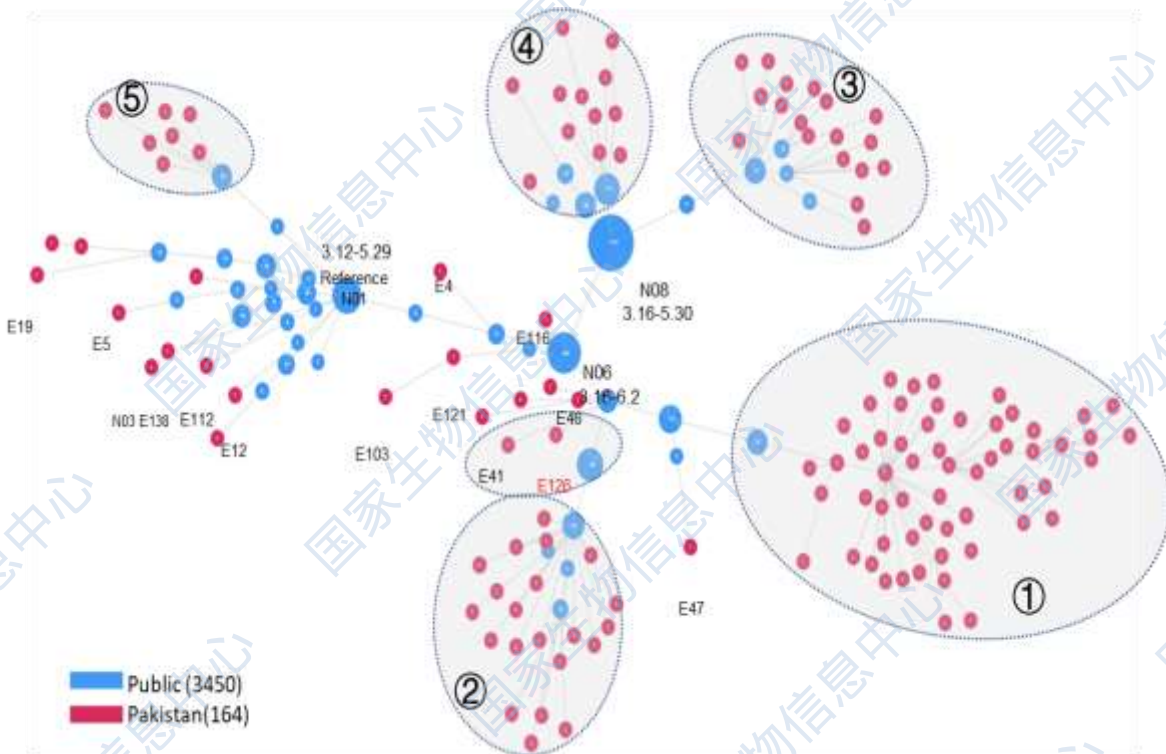
对北京市新发地市场相关序列进行变异分析与解读研判，为北京新发地疫情的精准防控提供决策支撑

- 路径1（人）：和北京序列100%一致，怀疑英国输入的1病例，可能与北京市场、市场确诊者/密接者有关系。
- 路径2（人）：和北京T1代序列比，少1个突变（6026），进化上更古老，怀疑境外输入的18例，可能在某个环节有关系。
- 路径3（食品）：全球公开序列中有598条序列和北京的T1代有序列层面的父子关系，怀疑进口冷链污染，造成传播。

NSR 2020

国际合作研究：巴基斯坦疫情早期病毒演化规律

(与巴基斯坦NIH/真纳大学合作)



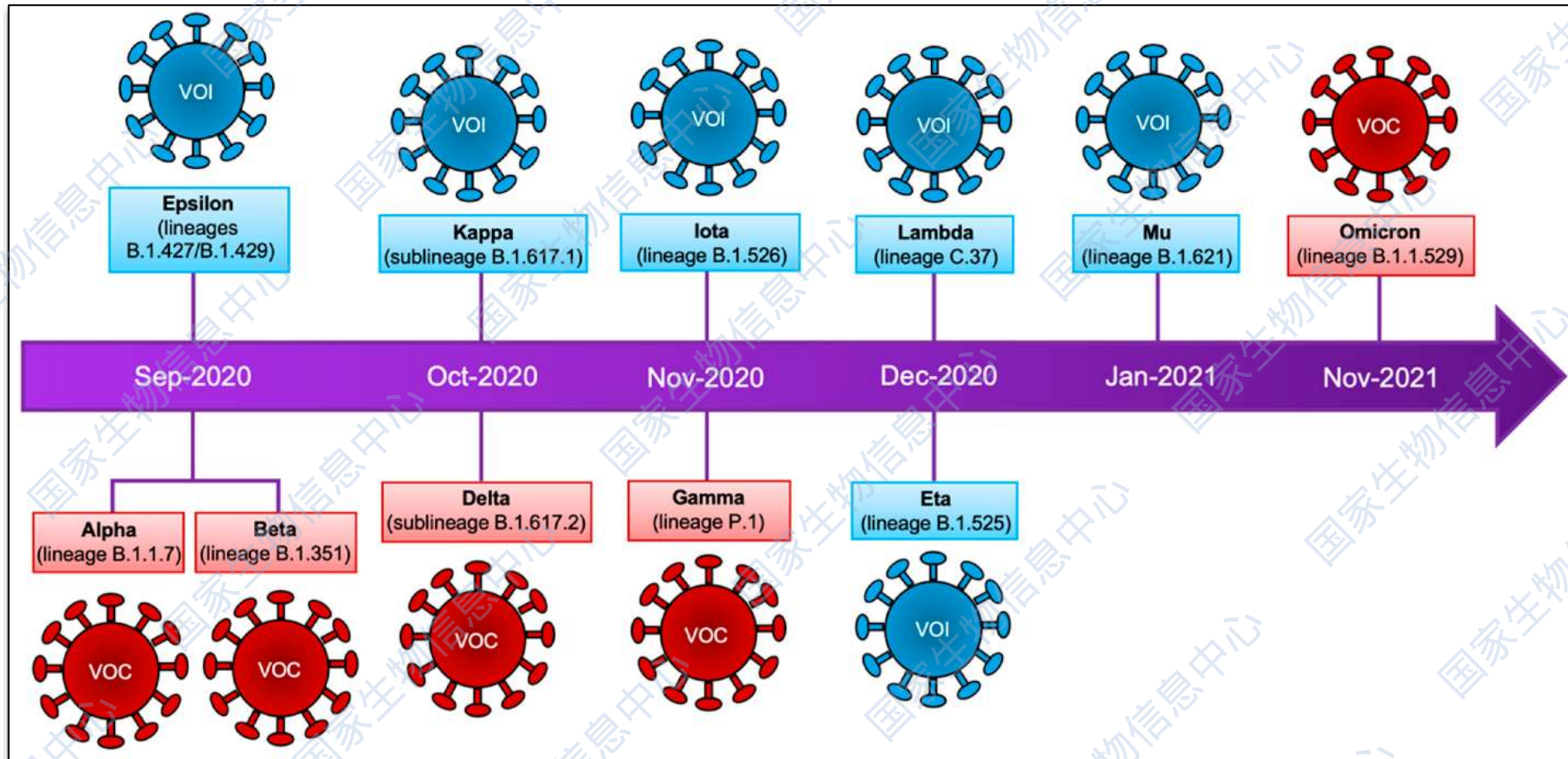
分析发现早期新冠传播与演化特征, 揭示巴基斯坦新冠病毒可能的输入路径

Genomics Proteomics & Bioinformatics 2021

纲要

1. 新冠病毒数据与信息库
2. 大规模单体型网络快速构建工具与追踪系统
3. 高风险株系精准**预警模型及应用平台**
4. 新冠病毒在线分析平台

新冠病毒不断演化，WHO发布重点关注株系

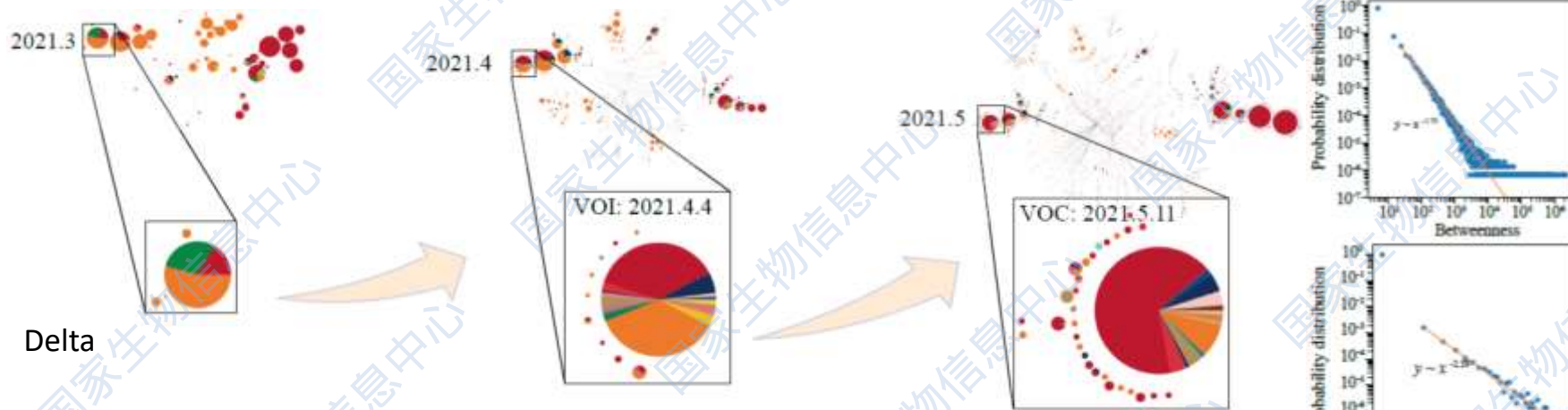


Flores-Vega et al., *Viruses*, 14(4), 653 (2022)

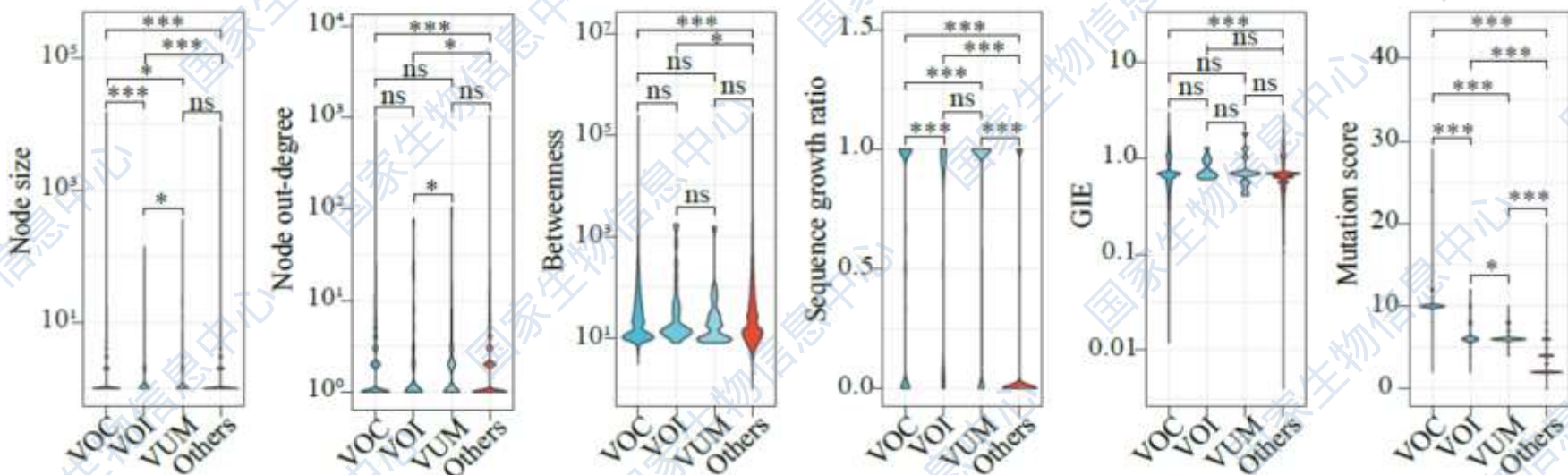
研究现状

类别	名称	方法或部门	国家或组织	特点
人工方法	世卫组织 (WHO)	病毒演化技术顾问组	联合国	分为不同风险等级 (VOC、VOI、VUM和其他) 依赖专家, 无法自动分类
	美国疾控中心 (CDC)	新冠病毒跨部门组	美国	
	欧洲疾控中心 (ECDC)	欧洲疾病控制中心和世卫组织欧洲区域办事处联合病毒特征工作组	欧盟	
自动方法	VarEPS	随机森林	中国	预测 单一指标 (传播力、抗体亲和力、适应度等), 没有综合的风险等级 没有把不同株系之间的演化关系作为特征
	PyR ₀	层次贝叶斯回归模型	美国	
	Predicting the mutational drivers of future SARS-CoV-2 variants of concern	逻辑回归	美国、英国	
生物实验	Enhanced fitness of SARS-CoV-2 variant of concern Alpha but not Beta	动物模型 (仓鼠、雪貂、小鼠), 人类细胞	德国、瑞士、美国	成本高、效率低 可信度高

病毒传播力相关的单体型网络特征



- ✓ 单体型网络特征与风险等级有关
- ✓ 特征在不同单体型上有差异
- ✓ 演化网络特征在WHO命名的不同风险等级的株系间有显著差别



Advanced Science 2024

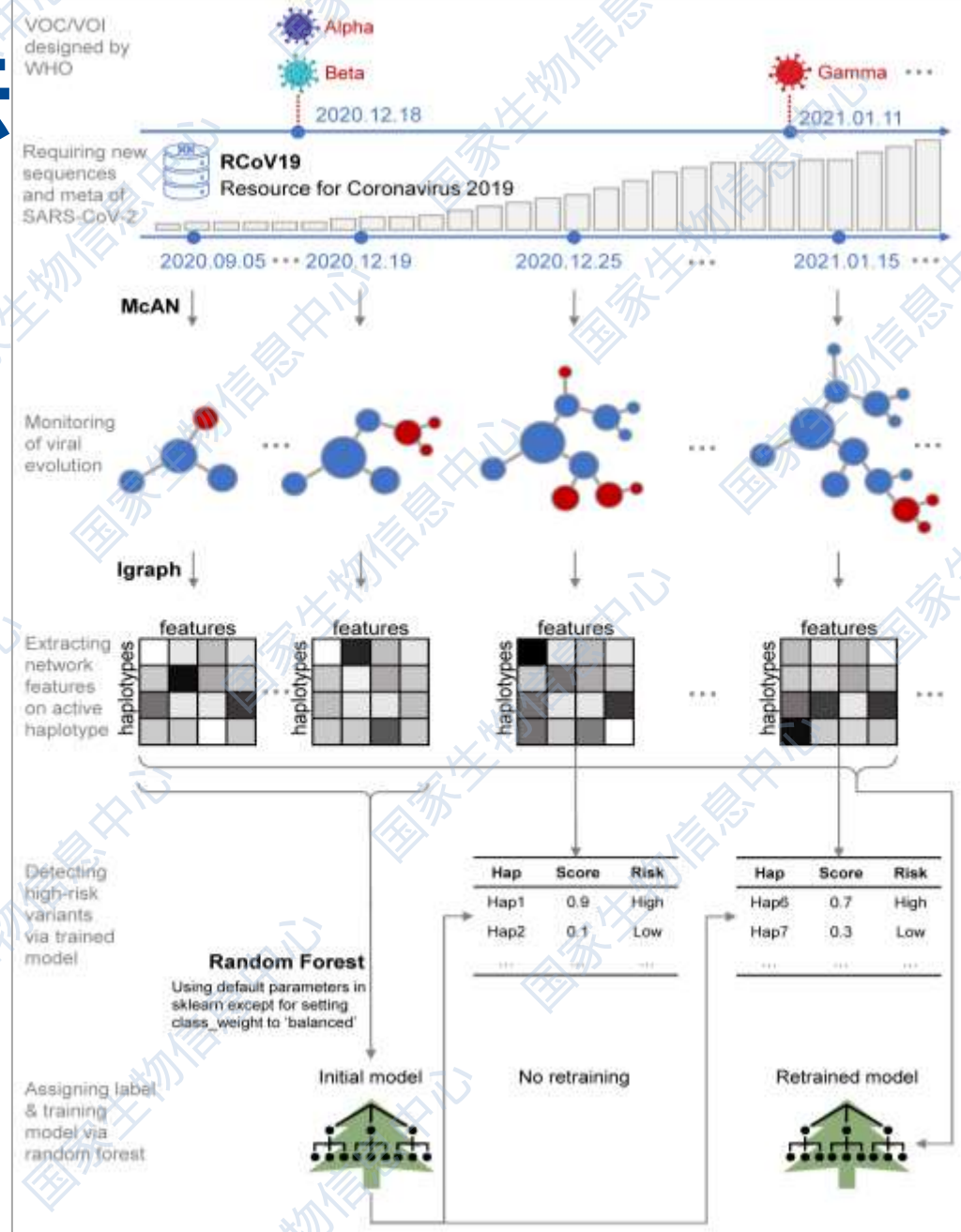
基于单体型网络特征的预警算法

HiRisk-Detector

利用单体型网络特征训练预警模型

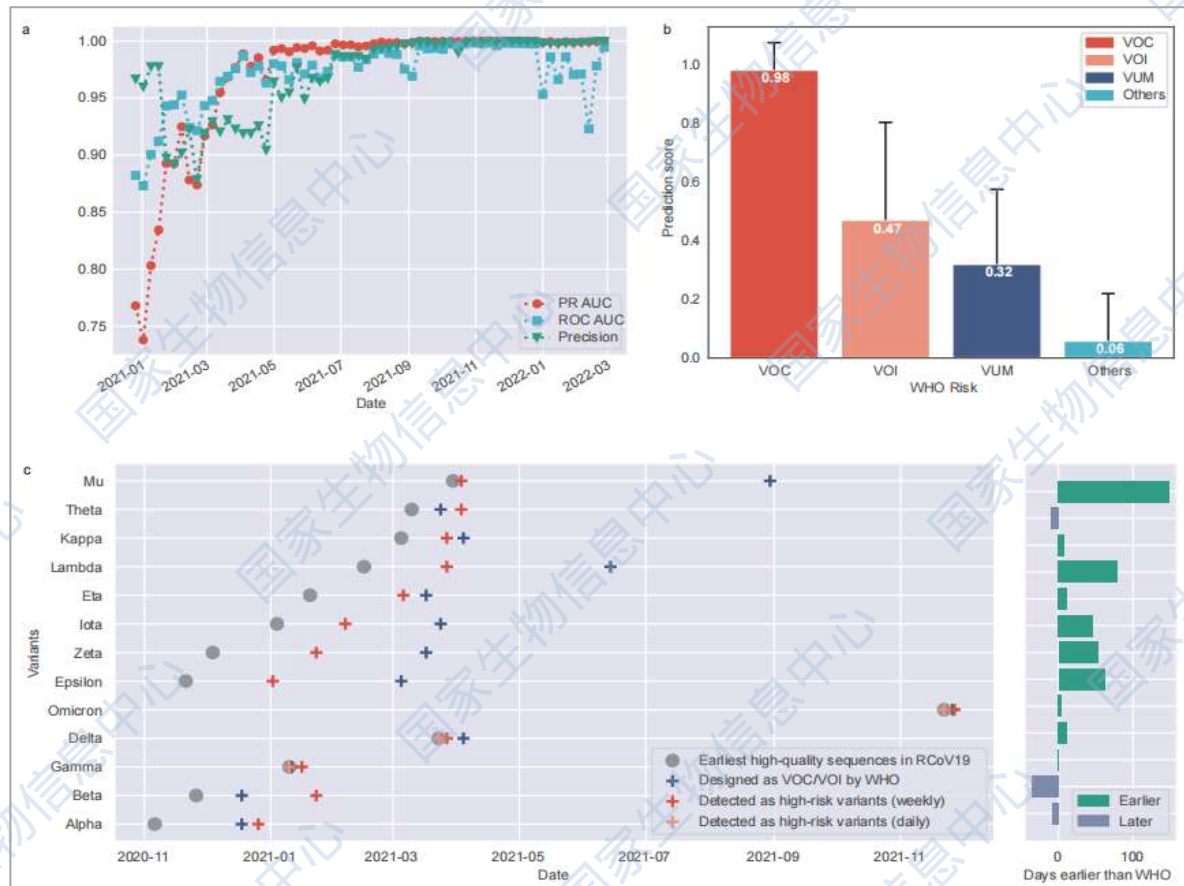
1. 构建单体型网络（不同时间点）
2. 在单体型网络中提取特征
 - 演化相关特征：出度、介数、连通性
 - 其他：节点大小、增长率、地理信息熵等
3. 根据WHO的分类标准分配标签
 - 高风险：VOC和VOI
 - 低风险：VUM和其他
4. 利用机器学习方法建立分类器

Advanced Science 2024

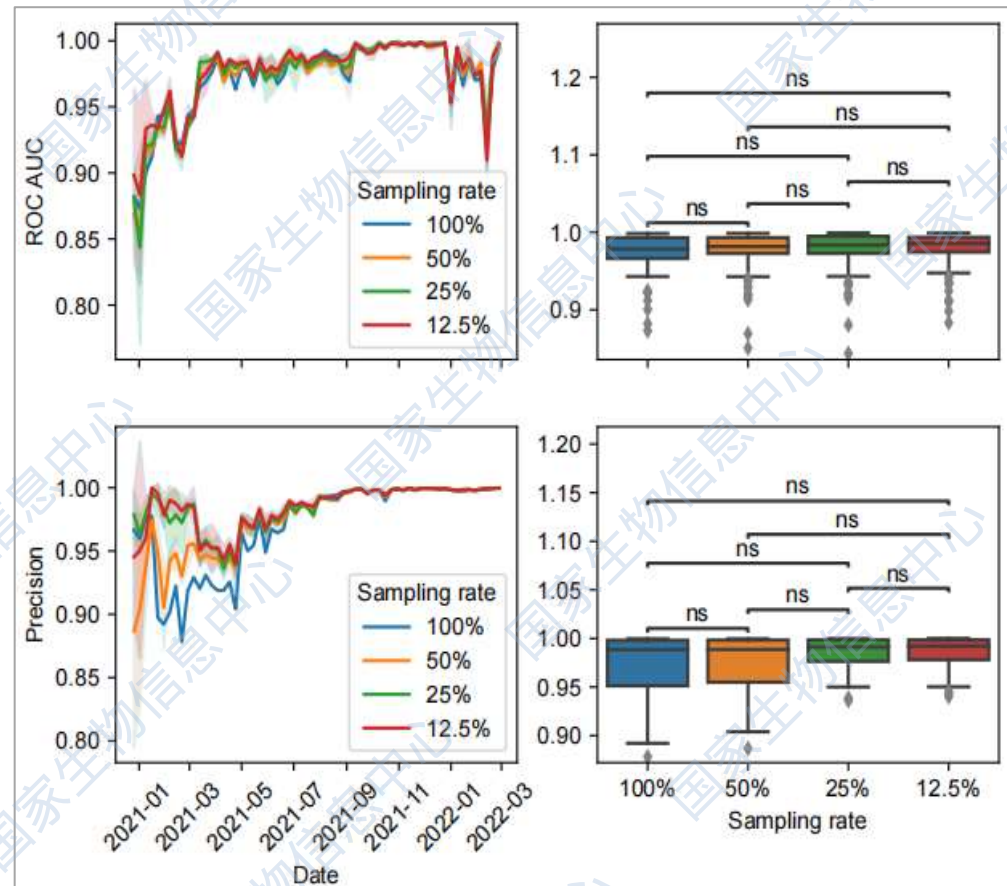


HiRisk-Detector算法性能

高精度和高灵敏度



对测序强度具有鲁棒性



成功识别了所有13种VOC/VOI，平均预警时间比世卫组织宣布的时间早27天；VOC、VOI、VUM和其他变异的平均风险得分依次下降；所有时间点精度值都大于0.85

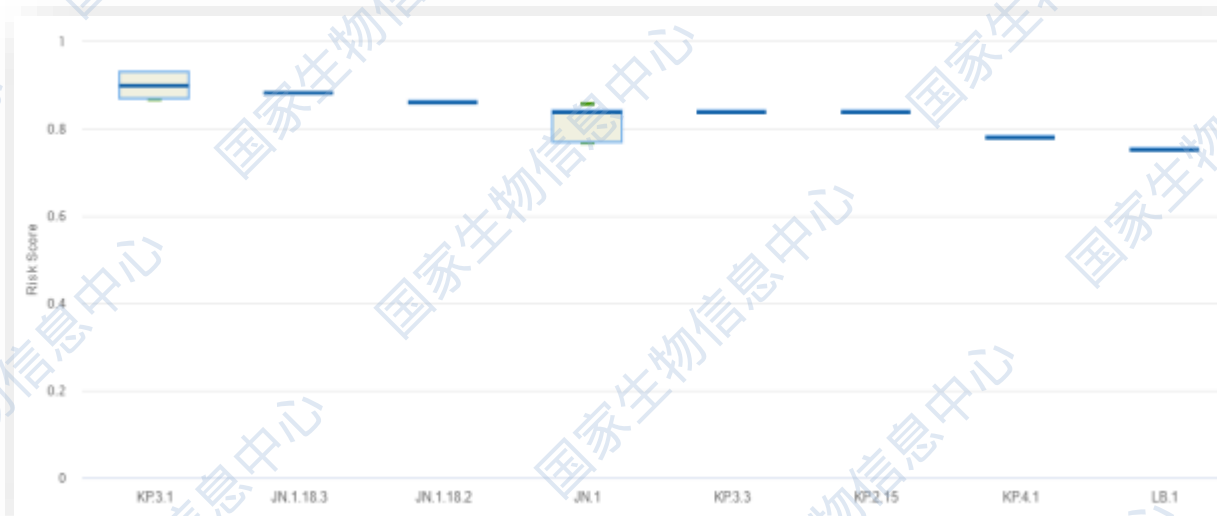
当测序强度降低到实际值50%、25%和12.5%时，PR-AUC、ROC-AUC和精度等性能指标无显著变化

Advanced Science 2024

服务于新冠疫情的防控

- 高风险株系检测方法已应用于2019新冠信息库 (RCoV19) 中的“高风险变体预警”模块
- 相关预警结果已写入新冠病毒监测与研判报告, 供各部委决策参考
- 源代码已于2023年10月公开至BioCode, 累计下载超1400次

高风险变体预警模块



<https://ngdc.cncb.ac.cn/ncov/monitoring/risk>

源代码已公开发布至BioCode



<https://ngdc.cncb.ac.cn/biocode/tool/7386>

实时预测及监测应用

2023年11月7日的风险检测结果表明，JN.1将成为流行株系。比WHO预警早41天。

数据集筛选

数据递交时间

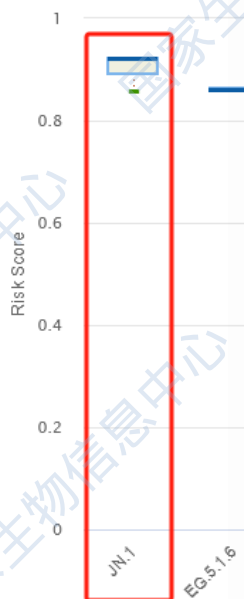
2023-11-07

当期统计， JN.1序列最早递交时间： 2023年10月22号

(二) 新冠病毒危险株系风险评估监测预警情况

利用中科院北京基因组所（国家生物信息中心）自主研发的新冠病毒高风险支系预警系统，基于2023年11月7日全球公开新冠病毒基因组序列数据，预警到13个支系的19个单体型具有较高风险分数的支系（大于0.8，图3），但是每个单体型的序列数较少，可继续观测追踪其后续发展情况。

需要注意的是，支系JN.1的3个单体型在本次预警中预警分数最高（超过0.85），目前该支系全球共享序列222条，可持续监测后续发展情况。

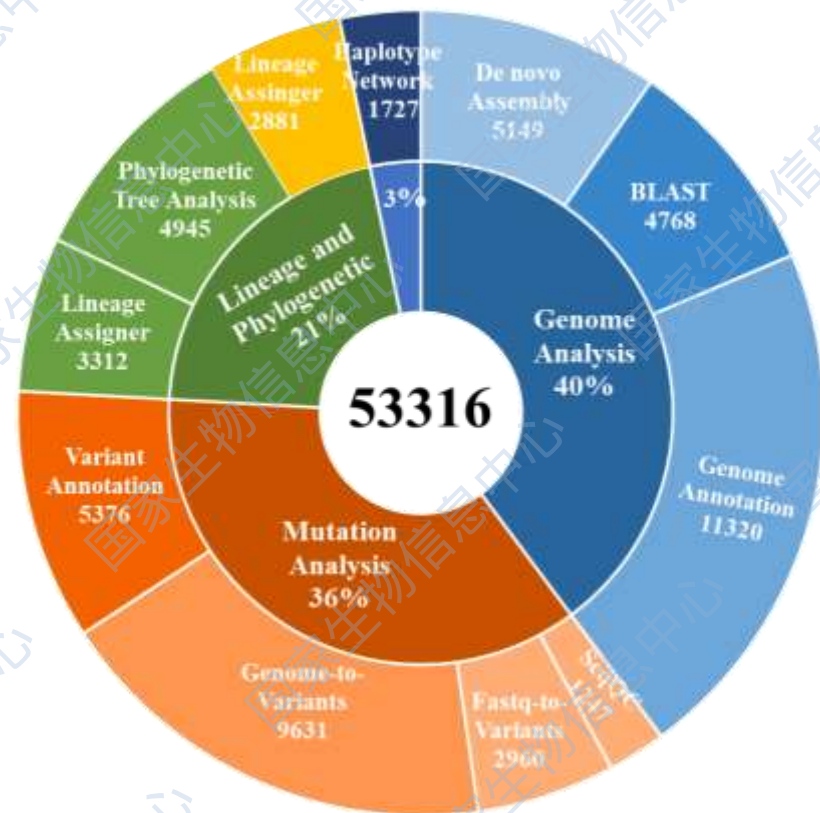
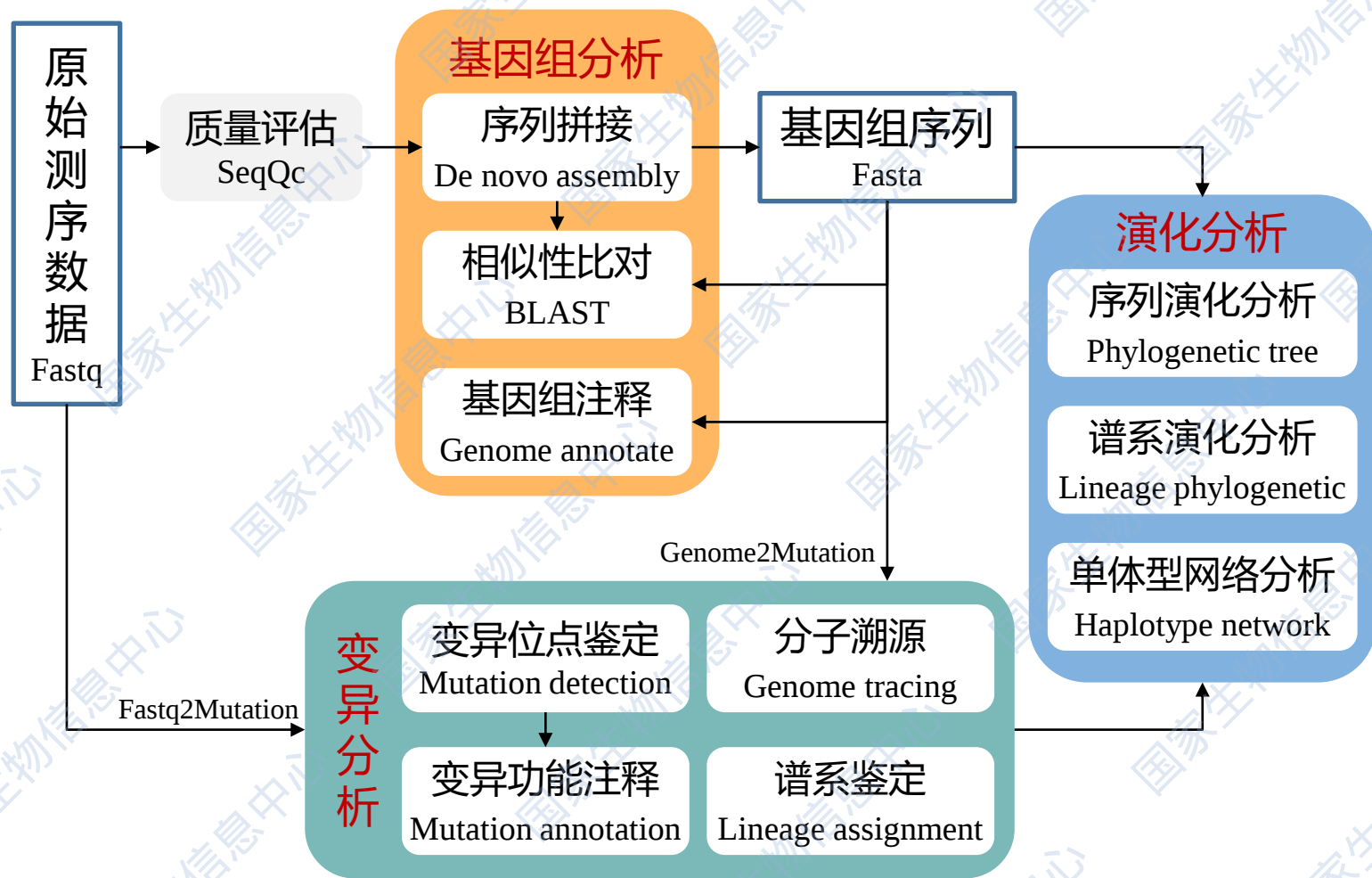


Pango lineage	Next strain clade	Genetic features	Earliest documented samples	Date of designation and risk assessments
BA.2.86 ^s	23I	Mutations relative to BA.2	24-07-2023	21-11-2023 BA.2.86 Initial Risk Evaluation, 21 November 2023
JN.1 [#]	24A	BA.2.86 + S:L455S	25-08-2023	18-12-2023 JN.1 Initial Risk Evaluation 18 December 2023 JN.1 Updated Risk Evaluation 9 February 2024 JN.1 Updated Risk Evaluation 15 April 2024

纲要

1. 新冠病毒数据与信息库
2. 大规模单体型网络快速构建工具与追踪系统
3. 高风险株系精准预警模型及应用平台
4. 新冠病毒**在线分析平台**

新冠病毒在线分析平台



✓快速 ✓批量分析
✓自动化 ✓全时在线服务

在线分析工具



Genome Tracing

Genome tracing工具可根据新冠病毒基因组或变异数据计算出其在数据库中最接近的序列，并显示其在全球时间空间的分布。

Denovo Assembly

Denovo Assembly工具可对二代测序原始数据进行序列拼接，并检测样本组装结果中含有的新冠病毒序列。

BLAST

BLAST工具可对用户在线提交的核酸序列进行BLAST比对，可选择新冠病毒基因组、新冠病毒数据库、冠状病毒基因组数据库。

Genome-to-Variants

Genome-to-Variants工具可将用户提交的新冠病毒基因组序列与参考基因组进行比对，检测SNP/indel变异位点。

Variant Annotation

Variant Annotation工具对用户输入的变异位点进行功能注释，显示变异位点所在基因、密码子和氨基酸的改变。

Position	Start	End	Position	Base
1000000	1	1	1	T
1000001	10	10	10	C
1000002	1	1	1	A
1000003	1	1	1	G

Genome Annotation

Genome Annotation工具可对用户提交的新冠病毒序列进行基因注释。

Fastq-to-Variants

Fastq-to-Variants工具可将二代测序数据与新冠病毒基因组进行序列比对，检测SNP/indel变异位点并进行功能注释。

SeqQC

SeqQC可以对FASTQ格式和BAM格式的文件进行序列质量评估，它是使用FastQC和MultiQC程序进行计算的。

Haplotype Network

单倍型网络构建工具可以实现对用户提交的变异数据文件和Meta数据文件结合我们库中已有数据进行整体网络图的计算和展示。

进化树

使用最大似然法构建系统发育树，该工具可将用户提交的序列使用IQ-TREE或RAxML构建系统发育树。

谱系和进化分析

该工具可使用Pangolin软件进行谱系预测，并将确认为新冠病毒的序列使用RAxML-EPA算法放置于新冠病毒参考系统演化树进行可视化展示。

Pangolin COVID-19 谱系分配

Pangolin COVID-19 Lineage Assigner 可将用户提交的序列使用Pangolin软件进行Pango谱系分配。



<https://ngdc.cncb.ac.cn/ncov/online/tools>

Zoological Research 2020

基因组拼接

功能描述

序列组装 (De novo-Assembly) 可将NGS二代测序数据进行BLASTN比对, 检测样本组装结果中含有的新冠病毒片段、测序深度等。

测序类型:

☒ 单向
☐ 双向

1. 选择测序类型

上传单端文件 ?

浏览... 未选择文件。

2.

*此流程目前仅支持illumina测序数据, 后续计划部署三代测序以减少上传时间。

邮件地址 (当计算时间较长时, 通过邮件通知)

邮箱

3. (optional) 填写邮件信息

运行

4.运行

提示: 目前队列运行的任务有0个, 排队任务有0个;

任务信息

作业号: 2020050964569

提交日期: 2020-05-09

运行状态: 完成

测序类型: Pair-end

下载结果文件

Download

BLASTN description

Show 5 entries

Query seq-id	Query length(bp)	Subject seq-id	Subject length(bp)	Identity(%)	Alignment length(bp)	Query coverage per subject(%)	Query coverage per hsp(%)	Query coverage per unique subject(%)	Query aligned	Subject aligned
k141_39	526	NC_045512.2	29903	100.000	140	27	27	27	387-526	12197-12336
k141_31	455	NC_045512.2	29903	100.000	58	13	13	13	1-58	19834-19891
k141_114	29937	NC_045512.2	29903	99.987	29855	99	99	99	10-29862	13-29867

Assembly statistics(QUAST)

Statistics type	Number	Statistics type	Number
# contigs	177	# contigs (>= 0 bp)	177
Largest contig	29937	# contigs (>= 1000 bp)	24
Total length	157071	# contigs (>= 5000 bp)	2
GC (%)	44.31	# contigs (>= 10000 bp)	1
N50	843	# contigs (>= 25000 bp)	1
N75	542	Total length (>= 1000 bp)	67464
L50	37	Total length (>= 5000 bp)	35464
L75	96	Total length (>= 10000 bp)	29937
# N's per 100 kbp	0.00	Total length (>= 25000 bp)	29937

<https://ngdc.cncb.ac.cn/ncov/online/tool/denovo-assembly>

基因组注释

功能描述

基因组注释 (Genome Annotation) 将从以下三个方面准

1. 基因 (S, M, E, N, ORF1ab, ORF1a, ORF3a, ORF6, O
 2. 基因的结构蛋白 (structured protein) ;
 3. ORF1ab和ORF1a的非结构蛋白 (non-structured prot
- 注: 该服务基于VADR工具 (Alejandro et al. 2020) (将删
看结果文件中的序列文件和注释文件); 相关基因和蛋白
其它基因组注释在线工具: ZCURVE_CoV 2.0.

Fasta序列

必选项

Fasta示例

输入Fasta序列

1. 直接输入序列

或者选择本地文件

浏览...

未选择文件。

邮件地址 (当计算时间较长时, 通过邮件通知计算结

邮箱

2. (optional) 填写邮件

运行

3.运行

任务信息

作业号: 2023051212688

提交日期: 20230512

运行状态: Finished

工具: VADR

下载结果文件

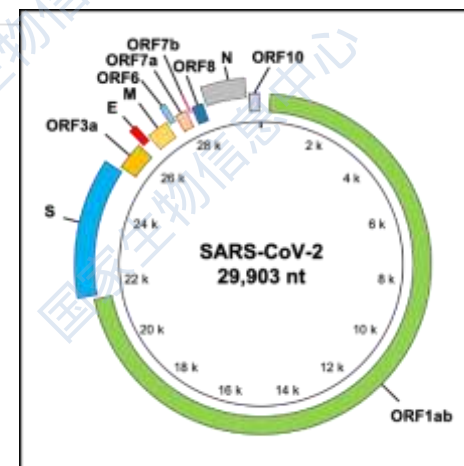
2023051212688.results.tar.gz

该压缩包包含所有结果文件, 注释前进行数据预处理: 删除头尾的N字符, 注释的基因坐标是基于处理后的序列得到的, 请同时参看结果文件中的序列文件和注释文件。

结果文件预览

```
>Feature MN908947.3
266      21555      gene
                        gene      ORF1ab
266      13468      CDS
13468     21555

product ORF1ab polyprotein
exception      ribosomal slippage
protein_id      MN908947.3_1
266      13483      CDS
product ORF1a polyprotein
protein_id      MN908947.3_2
```



<https://ngdc.cncb.ac.cn/nCoV/online/tool/genome-annotation>

基因组变异鉴定

功能描述

病毒基因组在线变异鉴定工具，可对用户提交的病毒基因组序列进行在线变异鉴定。该工具支持以SARS-CoV-2作为参照基因组序列(NC_045512)，同时支持用户提交参照基因组序列(fasta格式)。以SARS-CoV-2作为参照基因组序列时，工具会对变异

输入文件格式 (FASTA)

>BetaCoV/Wuhan/IVDC-HB-01/2019 EPI_ISL_402119
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGA
TCTGTTCTCTAAACGAACCTTTAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAG

VCF文件

GFF文件

统计文件

##reference=file://2019-nCoV

##contig=<ID=2019-nCoV,length=29903>

#CHROM> POS> ID> REF> ALT> QUAL> FILTER> INFO> FORMAT> SARS-CoV-2

2019-nCoV> 8782> 2019-nCoV_8782> C> T> .> .> RefRegion:orf1ab||VEP=synonymous_vari

2019-nCoV> 28144> 2019-nCoV_28144> T> C> .> .> RefRegion:ORF8||VEP=missense_variant,

2019-nCoV> 29803> 2019-nCoV_29803> CT> C> .> .> RefRegion:3'UTR||VEP=intergenic_varia

2019-nCoV> 29811> 2019-nCoV_29811> TAAA> T> .> .> RefRegion:3'UTR||VEP=intergenic_varia

##gff-version 3

##reference-region 2019-nCoV 1 29903

##sequence-region SARS-CoV-2 1 29815

2019-nCoV> SNP> SARS-CoV-2> 8782> 8782> .> .> ID=2019-nCoV_8782;Name=2019-r

2019-nCoV> SNP> SARS-CoV-2> 28144> 28144> .> .> ID=2019-nCoV_28144;Name=2019-

2019-nCoV> Deletion> SARS-CoV-2> 29803> 29804> .> .> ID=2019-nCoV_29803;Na

2019-nCoV> Deletion> SARS-CoV-2> 29811> 29814> .> .> ID=2019-nCoV_29811;Na

Reference> ReferenceLength>Query> QueryLength> Base (N) Counts> DegenerateBasesCounts> Analyzed?>

TotalVariants> SNP> Insertion> Deletion> Indel> PercentIdentity(%)

2019-nCoV> 29903> SARS-CoV-2> 29815> 0> 0> YES> 4> 2> 0> 2> 0>

99.987%

邮箱 4. (optional) 填写邮件信息

运行 5.运行

<https://ngdc.cncb.ac.cn/ncov/online/tool/variation>

原始测序数据变异鉴定

功能描述

变异检测 (Fastq-to-Variants) 可将NGS二代测序原始数据与新冠病毒病毒序列, 分析测序数据对新冠病毒基因组的覆盖度、测序深度、错误率等。

上传参照基因组

1. 设定参考基因组

默认使用如下序列

NC_045512

OR

上传一条FASTA参照序列

浏览... 未选择文件。

上传fastq格式文件

测序类型:

☒ 单向

☐ 双向

2. 选择测序 上传文件 填写样本

上传单端文件

浏览... 未选择文件。

样品名称:

样品名称 sample name

*此流程目前仅支持illumina测序数据, 后续计划部署三代测序数据分析文件以减少上传时间。

或者填写GSA中的相关CRR编号

2'. 或者直接调用GS

CRR Accession

CRR Accession

*SARS-CoV-2的CRR数据列表在GSA如下链接中: <https://ngdc.cncb.ac.cn/gsa/>

邮件地址 (当计算时间较长时, 通过邮件通知计算结果)

邮箱

Your

3. (optional) 填写邮

运行

4. 运行

统计信息

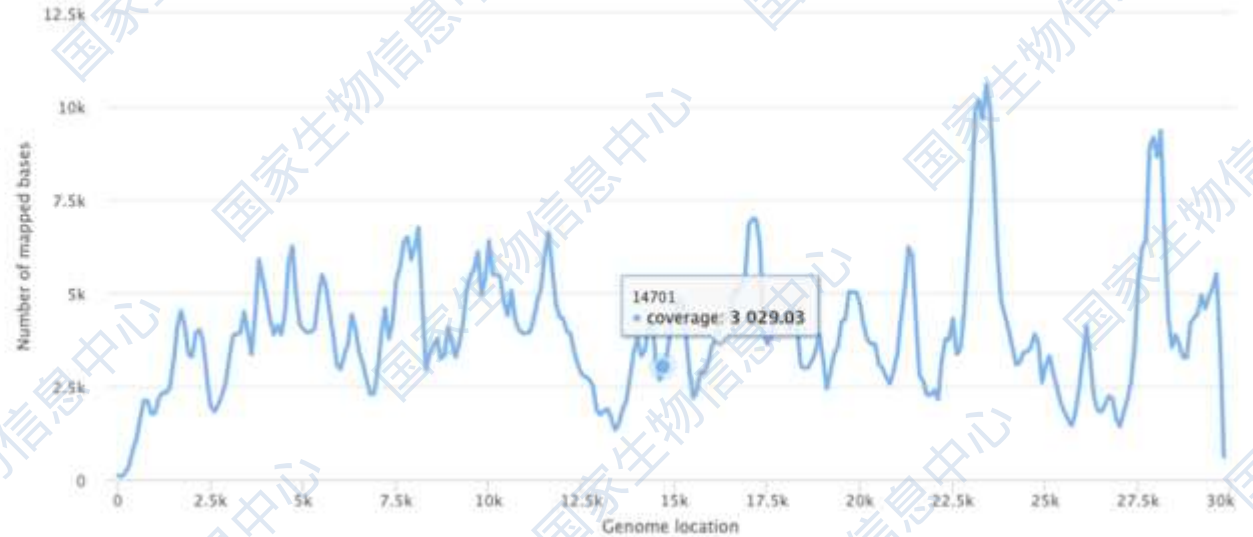
Statistics type	Number
Aligned_bases	115931610
Aligned_reads_number	642377
Coverage%	100
Duplication_rate%	0
Indel_rate%	0.0072
Mean_read_length	183.8
Mismatch_rate%	0.1
Sequencing_depth	3816.13
Variants_number	6

结果文件

1. Variant VCF file

2. Variant annotation table

基因组覆盖度



注释信息

Show 5 entries

Search:

Reference	Start position	End position	Base	Ref base	Ref annotation position	Annotation type	Protein.position.amino acids change	Gene.position.codons	Impact
2019-nCoV	8782	8782	T	REF:C	START:gene-orf1ab	synonymous_variant	QHD43415.1:p.2839S	gene-orf1ab:c.8517agC>agT	LOW
2019-nCoV	17747	17747	T	REF:C	START:gene-orf1ab	missense_variant	QHD43415.1:p.5828P>L	gene-orf1ab:c.17483cC>cTt	MODERATE
2019-nCoV	17858	17858	G	REF:A	START:gene-orf1ab	missense_variant	QHD43415.1:p.5865Y>C	gene-orf1ab:c.17594tA>tGt	MODERATE
2019-nCoV	18060	18060	T	REF:C	START:gene-orf1ab	synonymous_variant	QHD43415.1:p.5932L	gene-orf1ab:c.17796ctC>ctT	LOW
2019-nCoV	20281	20281	G	REF:T	START:gene-orf1ab	missense_variant	QHD43415.1:p.6673F>L	gene-orf1ab:c.20017Ttc>Ctc	MODERATE

Showing 1 to 5 of 6 entries

Previous 1 2 Next

<https://ngdc.cncb.ac.cn/ncov/online/tool/variation>

变异功能注释

功能描述

此工具可对用户输入的变异位点进行功能注释，显示变异位点所在基因、密码中和氨基酸的改变，以及对基因功能影响的程度。目前仅支持NC_045512作为reference的在线注释。

输入文件格式

#Reference	Start position	End position	Base
------------	----------------	--------------	------

下载结果文件

 results.tar.gz

该压缩包包含所有结果文件

Show 5 entries

Search:

Reference	Start position	End position	Base	Ref base	Ref annotation position	Annotation Type	Protein.Position.Amino acids change	Gene.Position.Codons	Impact
NC_045512	241	241	T	REF:C	START:5'UTR	upstream_gene_variant	QHD43415.1	gene-orf1ab	MODIFIER;DISTANCE=25
NC_045512	8782	8782	T	REF:C	START:gene-orf1ab	synonymous_variant	QHD43415.1:p.2839S	gene-orf1ab:c.8517agC>agT	LOW
NC_045512	23403	23403	G	REF:A	START:gene-S	missense_variant	QHD43416.1:p.614D>G	gene-S:c.1841gAt>gGt	MODERATE
NC_045512	28144	28144	C	REF:T	START:gene-ORF8	missense_variant	QHD43422.1:p.84L>S	gene-ORF8:c.251tTa>tCa	MODERATE
NC_045512	29803	29804	C	REF:CT	START:3'UTR	intergenic_variant	-	-	MODIFIER

<https://ngdc.cncb.ac.cn/ncov/online/tool/annotation>

Pango谱系判定

功能描述

Pangolin COVID-19 Lineage Assigner 可将用于
当前版本Pangolin version: 4.3.1, Pangolin-dat

Fasta序列 ? 必选项

Fasta示例

输入Fasta序列

1. 输入序列

或者选择本地文件 ?

选择文件

未选择

参数设置 可选项

2. 设置参数

analysis-mode

accurate (USHER) (default)

邮件地址 (当计算时间较长时, 通过邮件通

邮箱

3. (optional) 填写

运行

4. 运行

任务信息

作业号: 2023060187613

提交日期: 20230601

运行状态: Finished

工具: Pangolin

下载结果文件

2023060187613.results.tar.gz

结果文件预览

结果文件预览

Show 10 entries

Search:

Sequence Name	Lineage	Pangolin Version	Conflict
MW064593.1 SARS-CoV-2/human/USA/GA-QDX-594/2020	A.3	4.3	0.0
MW985569.1 SARS-CoV-2/human/USA/NJ-CDC-LC0019257/2021	B.1.526	4.3	0.0
OU194399.1 COG-UK/ALDP-142BAFB	B.1.1.7	4.3	0.0
OU256551.1 COG-UK/QEUEH-150C752	B.1.1.7	4.3	0.0
OU289933.1 COG-UK/ALDP-15C07F0	AY.4	4.3	0.0

可一次对多条序列进行谱系划分

<https://ngdc.cncb.ac.cn/ncov/online/tool/lineage-assigner>

基因组溯源

演化分析

两种格式
输入

Query sequence

Mutation

test_seq:241C/T,3037C/T,14408C/T,23403A/G,28881G/A,28882G/A,28883G/C,2091C/T,5128A/G,8360A/G,13860C/T,19839T/C,19999G/T,28905C/T

example

点击计算最近源序列

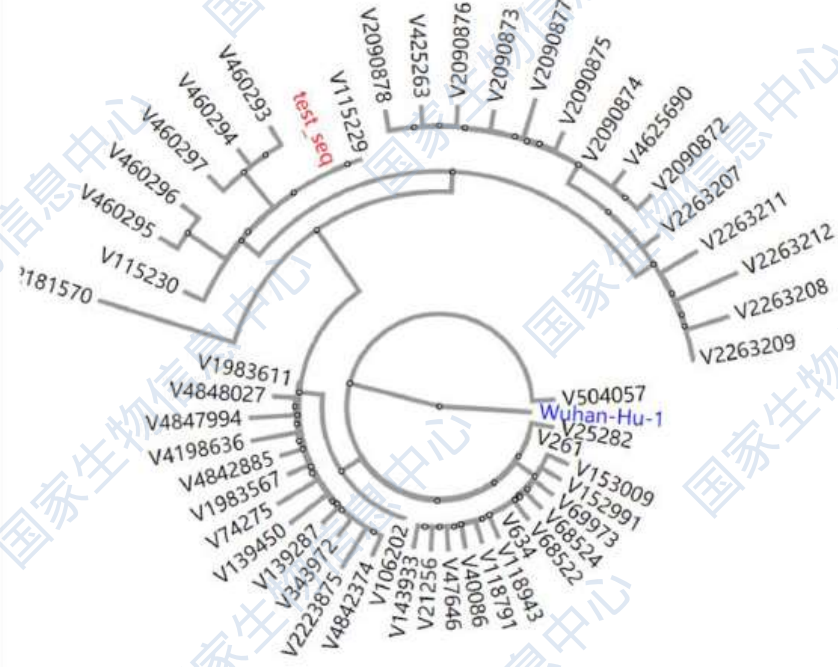
Matched sequences ?

Showing 1 to 50 of 50 entries

Copy Excel CSV

SeqID ?	Accession ?	Score ?	Lineage ?	Mutation ?	Su
V115229	EPI_ISL_498692	1	B.1.1.317	0	
V115230	EPI_ISL_498693	0.966667	B.1.1.317	1	3
V2263209	EPI_ISL_9301315	0.928571	B.1.1.317	2	2
V460294	EPI_ISL_2170890	0.911765	B.1.1.317	3	2
V460295	EPI_ISL_2170891	0.911765	B.1.1.317	3	8
V460296	EPI_ISL_2170888	0.911765	B.1.1.317	3	2
V460297	EPI_ISL_2170889	0.911765	B.1.1.317	3	2
V460293	EPI_ISL_2170887	0.911765	B.1.1.317	3	8
V2090874	OM098414.1	0.897619	B.1.1.317	3	1
V2263207	EPI_ISL_9301319	0.897619	B.1.1.317	3	8
V2090872	OM098425.1	0.870536	B.1.1.317	4	1
V2090875	OM098416.1	0.870536	B.1.1.317	4	1
V2090876	OM098417.1	0.870536	B.1.1.317	4	7

Phylogeny of matched sequences ?



reset

2 435 103

2 500k 2 750k

Briefings in Bioinformatics 2023

<https://ngdc.cncb.ac.cn/nCoV/online/tool/genome-tracing/>

系统发育树构建

功能描述

使用最大似然法构建系统发育树。该工具可将用户提交的序列使用IQ-TREE或RAxML构建系统发育树。

默认工作流程为：① 输入序列：新冠病毒基因组序列；② 修剪方法：以新冠参考病毒基
③ 替代模型：IQ-TREE使用ModelFinder自动选择（RAxML使用GTRGAMMA）；④
序列类型、修剪模式（refseq、TrimAI、无操作）、建树软件（IQ-TREE和RAxML）、
UFBoot - Ultrafast Bootstrap Approximation）。

Fasta序列 ? 必选项

[Fasta示例](#)

输入Fasta序列

或者选择本地文件 ?

[选择文件](#)

未选择任何文件

参数设置

2. 设置参数

software

IQ-TREE

▼

model

MFP

▼

trim ?



Whether Use Bootstrap

邮件地址 (当计算时间较长时, 通过邮件通知计算结果) 可选项

邮箱

Your Email

邮件主题

Email

运行

4. 运行

任务信息

作业号: 2021020834483

提交日期: 2021-07-26

运行状态: Finished

所用软件: iqtree

建树模型: HKY+F

截取方法: refseq

使用bootstrap: true

MAFFT比对策略: FFT-NS-2 (Fast but rough)

软件版本: MAFFT version 7.475, TrimAI version v1.4.rev15, IQ-TREE version 2.0.3

下载结果文件

results.tar.gz

该压缩包包含所有结果文件

结果文件预览

导出PNG图片

下载nwk文件



<https://ngdc.cncb.ac.cn/ncov/online/tool/tree>

基于Pango谱系的序列演化分析

演化分析

功能描述

Pango Lineage Assigner & Phylogenetic Placement工具可将病毒 (SARS-CoV-2) 的序列使用RAxML-EPA算法放置于新冠展示字段。

Fasta序列 ? 必选项 **Fasta示例**

输入Fasta序列

1. 输入序列或上传文件

或者选择本地文件 ? **浏览...** 未选择文件。

邮件地址 (当计算时间较长时, 通过邮件通知计算结果)

邮箱 Your Email Address 2. (optional) 填写邮件信息

运行

3. 运行

任务信息

作业号: 2021073032852

提交日期: 2021-07-30

序列条数: Waiting

软件版本: Pangolin version 3.1.7, line

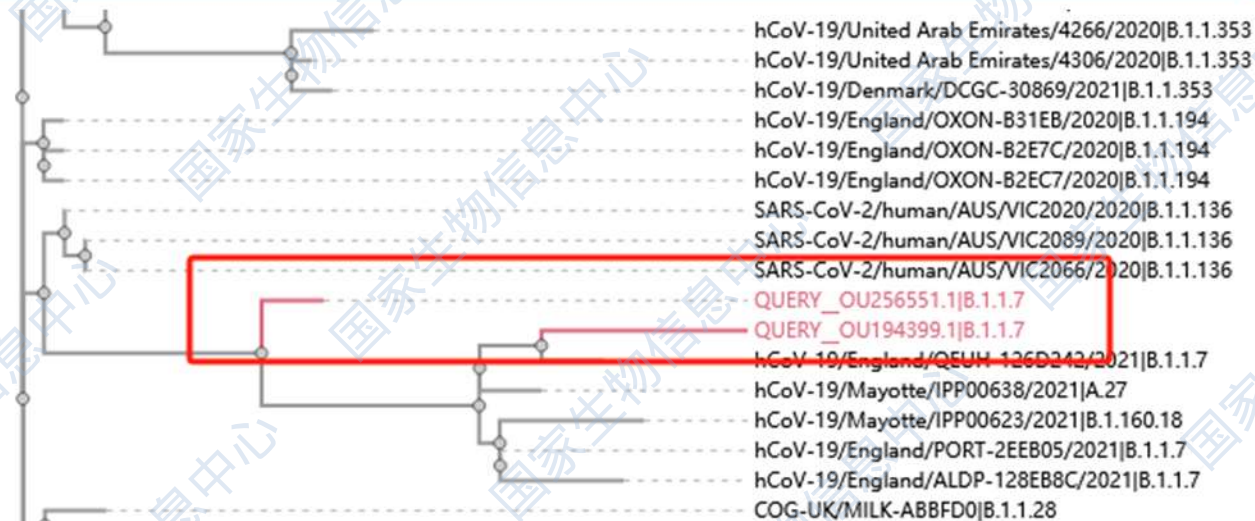
参考树版本: lineage version 2021-07-09

运行状态: Finished

结果文件预览

Show 10 entries

Show Name	Defline	Length(bp)	GC Content	Total N	Other Degenerate Bases	SARS-CoV-2?	Lineage	Conflict	Status
MW064593.1	MW064593.1 SARS-CoV-2/human/USA/GA-QDX-594/2020	29782	38.01%	0	0	Yes	A.3	nan	passed_qc
MW985569.1	MW985569.1 SARS-CoV-2/human/USA/NJ-CDC-LC0019257/2021	29704	37.98%	1	0	Yes	B.1.526	0.0	passed_qc
OU194399.1	OU194399.1 COG-UK/ALDP-142BAFB	29884	37.83%	121	0	Yes	B.1.1.7	0.0	passed_qc
OU256551.1	OU256551.1 COG-UK/QEUEH-150C752	29884	37.83%	121	0	Yes	B.1.1.7	nan	passed_qc



<https://ngdc.cncb.ac.cn/ncov/online/tool/lineage-tree>

单体型网络构建工具

功能描述

该工具使用“基于最小树形图的单体型网络构建方法”(MCAT)构建单体型网络。需要上传SARS-CoV-2序列变异文件(采用GF或VCF格式)和对应的元信息文件,或者从我们的2019新型冠状病毒信息库(RCoV19)选择已有样本。你可以从我们的信息库中通过设置起止采样时间、采样国家或地区来过滤已有样本,或者以随机方式选择样本。基于构建好的单体型网络,采用Newman方法对单体型划分团簇。注意变异文件中3'端非编码区和5'端非编码区的突变将被自动删除。

选择文件格式

☒ 变异文件格式
☐ VCF文件格式

1. 选择变异文件格式

变异文件

必选项

变异文件示例

上传变异文件

浏览... 未选择文件。

2. 上传变异文件

元信息文件

必选项

元信息示例

上传元信息文件

浏览... 未选择文件。

3. 上传对应的元信息文件

选择特定数据集

可选项

开始采样时间

2019-01-01

至

结束采样时间

2021-05-31

国家

China

缺失日期类型

month

随机取样个数(推荐 $0 < N < 10000$)

500

加入在RCoV19中选择的特定数据集, 构建单体型网络 (可选项)

构建单体型网络参数设置

4. (optional) 参数设置

最小突变率 (为了选择高突变率位点)

0.01

邮件地址 (当计算时间较长时, 通过邮件通知计算结果) 可选项

5. (optional) 填写邮件信息

邮箱

your_email

邮件主题

运行

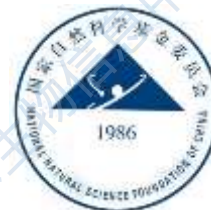
6. 运行

<https://ngdc.cncb.ac.cn/ncov/online/tool/haplotype>

致谢



Members in CNCB-NGDC
<https://ngdc.cncb.ac.cn/people>



中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

北京市科学技术委员会
Beijing Municipal Science & Technology Commission



“一带一路”国际科学组织联盟
Alliance of International Science Organizations

Contact us

National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation

Email:

Raw sequencing data submission gsa@big.ac.cn

Sequence submission genbase@big.ac.cn

Variants problem licp@big.ac.cn

Address: No.1 Beichen West Road, Chaoyang District, Beijing 100101, China



THANKS