Analysis and application of microbiological data: Variants monitoring, lineage assignment and risk-prewarning for viral pathogens

Shuhui Songshh

CNCB-NGDC

Oct 14 2025 Beijing

国家生物信息中心版权所有

Biosketch

- Principal investigator, 2022-Current, CNCB/BIG, CAS
- Director of Bioinformatics Services Department of the China National Center for Bioinformation
- Deputy Director of National Genomics Data Center
- Focus: **Multi-omics data integration, analysis and mining**
- Achievements: ~70 publications on *PNAS, NAR, Mol Plant, Genome Biol, GPB, etc.*

基因组变异库GVM

https://ngdc.cncb.ac.cn/gvm/

Nucleic Acids Research (2018, 2021)#

图像表型库OPIA

https://ngdc.cncb.ac.cn/opia

Nucleic Acids Research (2024)#

知识库GWAS Atlas

https://ngdc.cncb.ac.cn/gwas/

Nucleic Acids Research (2020, 2023)#

Research Direction and Achievements

Genomic Variation

Data Resources

- GVM: Genome variation map (*Nucleic Acids Research* 2018#,*, 2021*)
- GWAS Atlas: Genome-wide association (*Nucleic Acids Research* 2020*, 2023*)
- RCoV19 (*YiChuan* 2020#, *Genomics Proteomics Bioinformatics* 2020#,2023*)
- CGI: Chloroplast Genome Information Resource (*Plant Biotechnol J* 2022#)
- OPIA: Open plant image archive (*Nucleic Acids Research* 2024*)

Algorithm and Tools

- Heterogeneity analysis of allele expression regulation (*PNAS* 2016#,*)
- Integration analysis of genotype and multidimensional phenotypic traits (*Genome Biol* 2020*)
- Variation based machine learning classification and identification (*Genomics Proteomics Bioinformatics* 2020*)
- Algorithm software for rapid construction of haplotype networks (*Brief Bioinform* 2023*)
- High risk variants warning model based on machine learning (*Adv Sci 2024**)

Application Platform

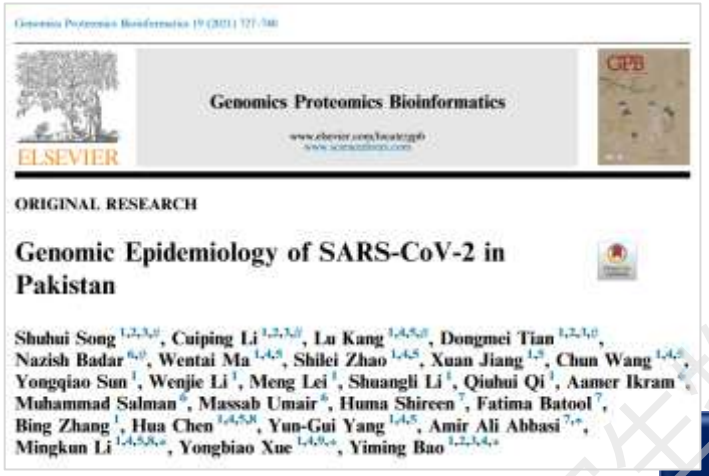
- Spatiotemporal dynamic monitoring platform for COVID-19 (*Genomics Proteomics Bioinformatics* 2020#, 2023*)
- Species identification based on chloroplast molecular markers (*Mol Ecol Resour. 2023#*)
- SoyOmics (*Mol Plant* 2023*)
- SugancaneOmics (*Plant Communication* 2025*)
- Immune big data resource management platform

#First Author，*Corresponding Author

Outline

国家生物信息中心版权所有

Studies about pathogens


Genomics Proteomics Bioinformatics

ORIGINAL RESEARCH

Genomic Epidemiology of SARS-CoV-2 in Pakistan

Shuhui Song [1,2,3,#], Cuiping Li [1,2,3,#], Lu Kang [1,4,5,#], Dongmei Tian [1,2,3,#], Nazish Badar [6,#], Wentai Ma [1,4,5], Shilei Zhao [1,4,5], Xuan Jiang [1,5], Chun Wang [1,4,5], Yongqiao Sun [1], Wenjie Li [1], Meng Lei [1], Shuangli Li [1], Qiuhui Qi [1], Aamer Ikram [6], Muhammad Salman [6], Massab Umair [6], Huma Shireen [7], Fatima Batool [7], Bing Zhang [1], Hua Chen [1,4,5,8], Yun-Gui Yang [1,4,5], Amir Ali Abbasi [7,*], Mingkun Li [1,4,5,8,*], Yongbiao Xue [1,4,9,*], Yiming Bao [1,2,3,4,*]


RESEARCH ARTICLE

ADVANCED SCIENCE

Machine Learning Early Detection of SARS-CoV-2 High-Risk Variants

Lun Li, Cuiping Li, Na Li, Dong Zou, Wenming Zhao, Hong Luo, Yongbiao Xue,* Zhang Zhang,* Yiming Bao,* and Shuhui Song*


OXFORD

Briefings in Bioinformatics, 2023, 24(3), 1–9

Advance access publication date 12 May 2023

Problem Solving Protocol

McAN: a novel computational algorithm and platform for constructing and visualizing haplotype networks

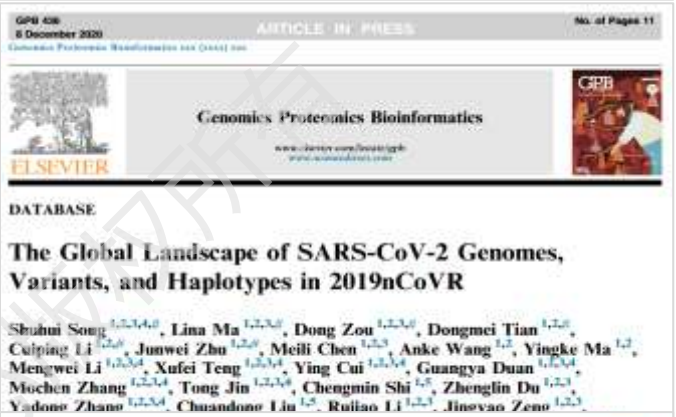
Lun Li[†], Bo Xu[†], Dongmei Tian[†], Anke Wang[†], Junwei Zhu[†], Cuiping Li, Na Li, Wei Zhao, Leisheng Shi, Yongbiao Xue, Zhang Zhang, Yiming Bao, Wenming Zhao and Shuhui Song


Genomics Proteomics Bioinformatics

DATABASE

RCoV19: A One-stop Hub for SARS-CoV-2 Genome Data Integration, Variant Monitoring, and Risk Pre-warning

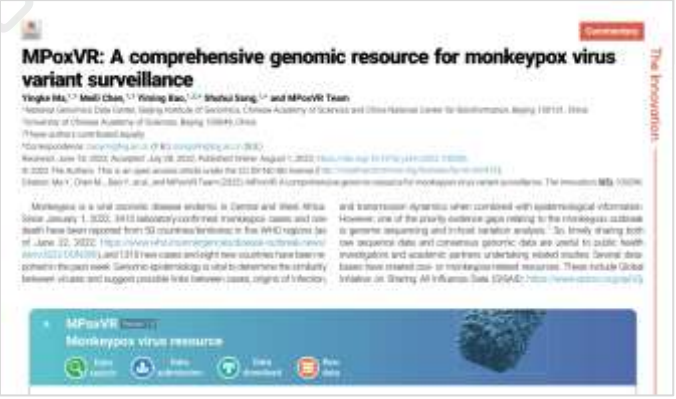
Cuiping Li [1,#], Lina Ma [1,2,3,#], Dong Zou [1,2,#], Rongqin Zhang [1,2,3,#], Xue Bai [1], Lun Li [1], Gangao Wu [1,2,3], Tianhao Huang [1,2,3], Wei Zhao [1,2,3], Enhui Jin [1,2,3], Yiming Bao [1,2,3,*], Shuhui Song [1,2,3,4,*]


Genomics Proteomics Bioinformatics

DATABASE

The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR

Shuhui Song [1,2,3,4,#], Lina Ma [1,2,3,#], Dong Zou [1,2,3,#], Dongmei Tian [1,2,#], Cuiping Li [1,2,#], Junwei Zhu [1,2,#], Meili Chen [1,2,3], Anke Wang [1,2], Yingke Ma [1,2], Mengwei Li [1,2,3,4], Xufei Teng [1,2,3,4], Ying Cui [1,2,3,4], Guangya Duan [1,2,3,4], Mochen Zhang [1,2,3,4], Tong Jin [1,2,3,4], Chengmin Shi [1,5], Zhenglin Du [1,2,3], Yadong Zhang [1,2,3,4], Chuandong Liu [1,5], Ruilan Li [1,2,3], Jinsyao Zeng [1,2,3]


MPoxVR: A comprehensive genomic resource for monkeypox virus variant surveillance

Yingke Ma[†], Meili Chen[†], Yiming Bao[*], Shuhui Song[*] and MPoxVR Team

Monkeypox virus resource

Analysis and Application

Database

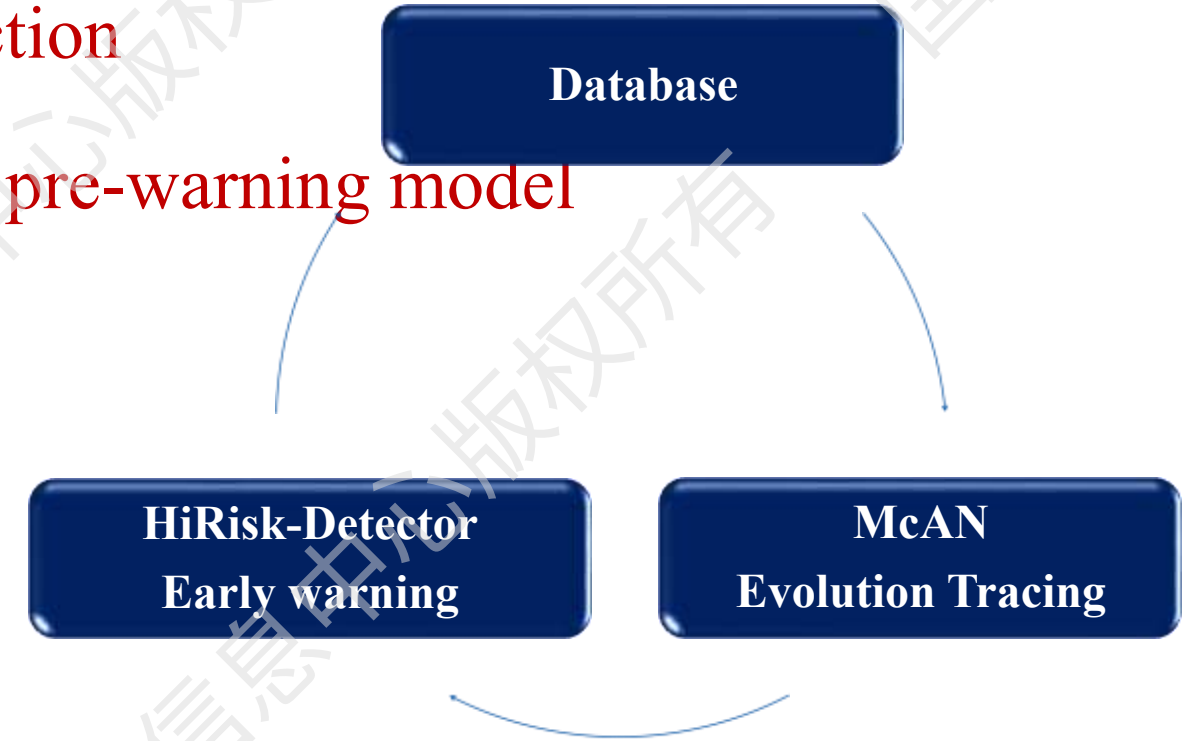
Subtype identification

Evolution network

Pre-warning

Outline

- Two algorithms
  - McAN: haplotype network construction
  - HiRisk-Detector: high-risk variants pre-warning model



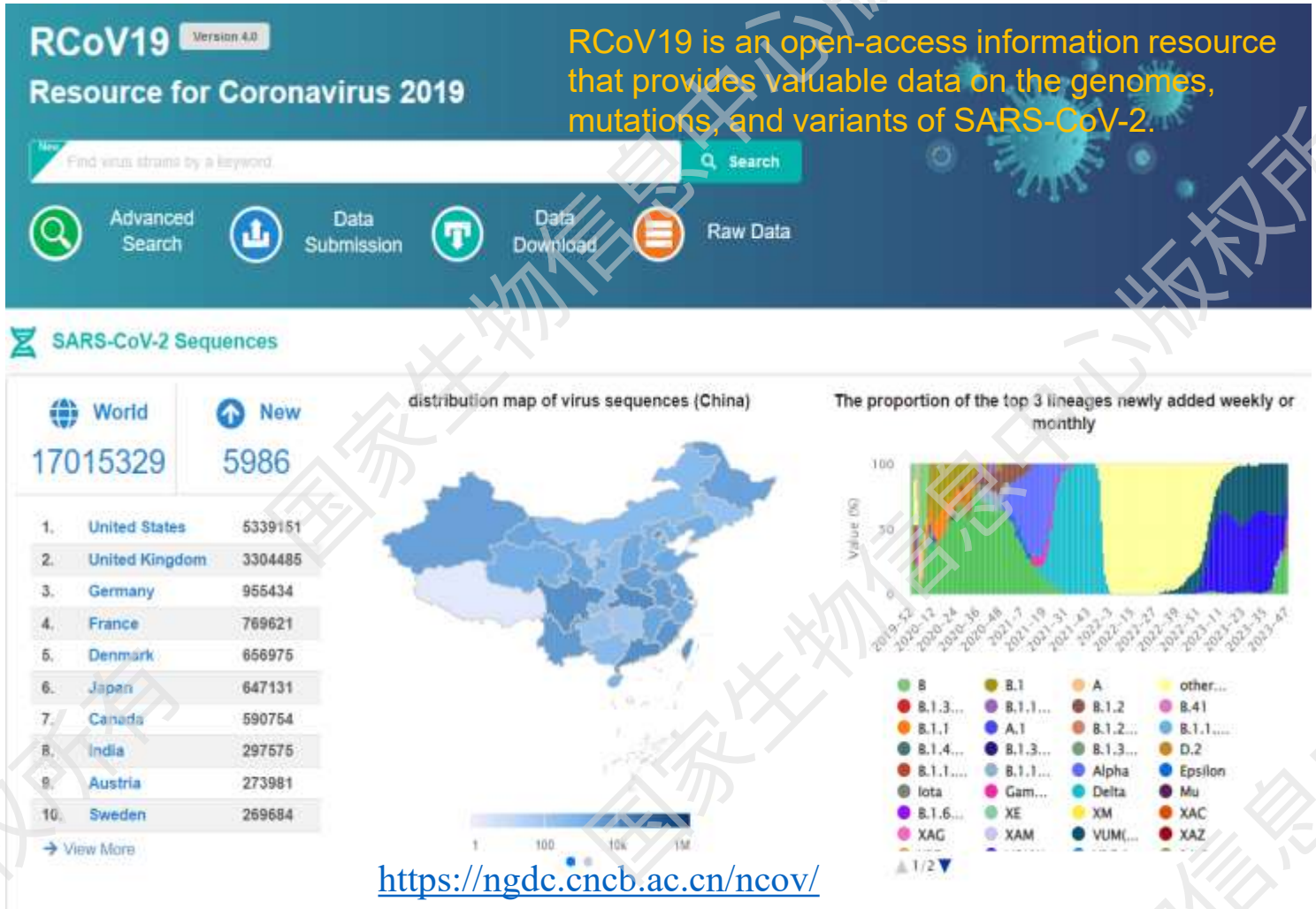
国家生物信息中心版权所有

Part 1

Data Resources for SARS-CoV-2(RCoV19) and Monkeypox Virus (MpoxVR)

国家生物信息中心版权所有

RCoV19: Resource for Coronavirus 2019

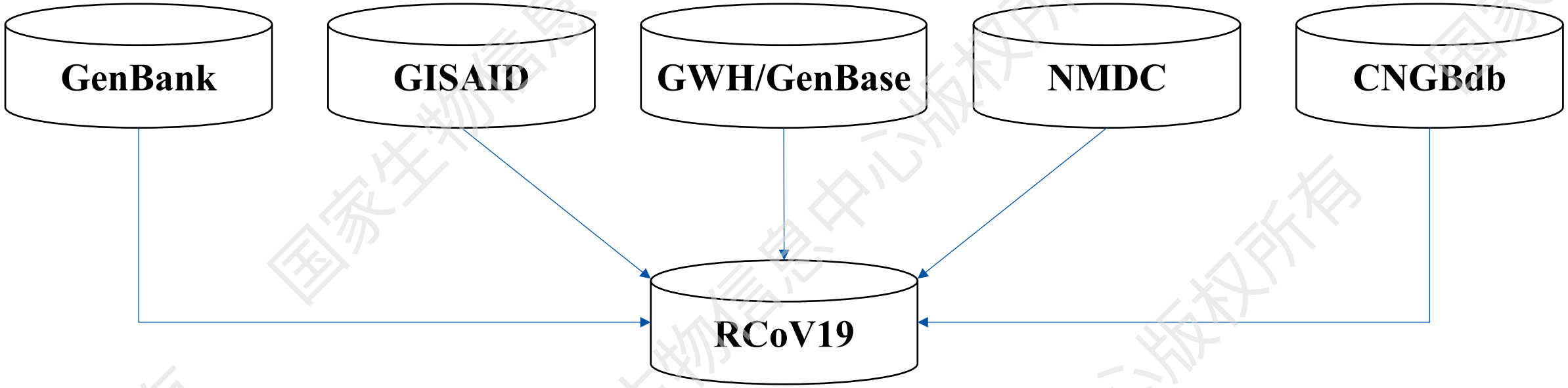


https://ngdc.cncb.ac.cn/ncov/

Users across the world

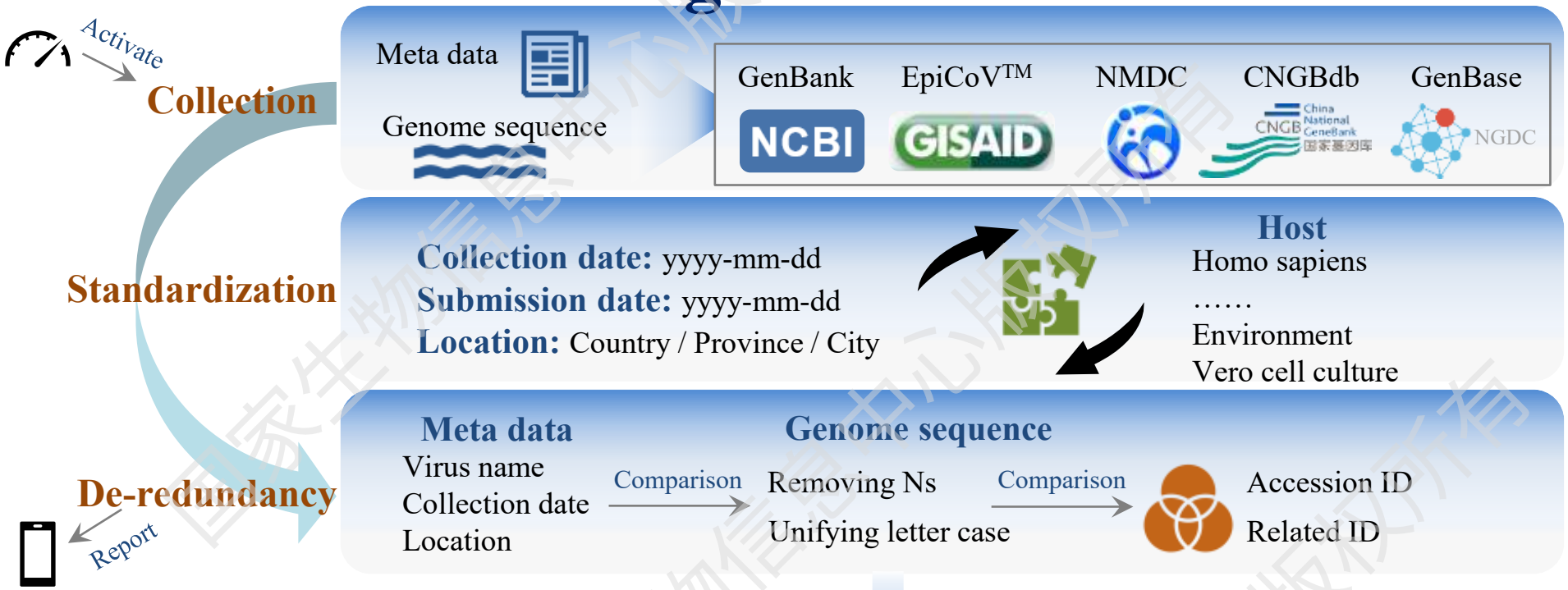
Zhao et al., 2020, *Yichuan;* Song et al., 2020, *Genomics Proteomics & Bioinformatics;* Gong et al. 2020, *Zoological Research*

Standardized integration of global SARS-CoV-2 genome information



As of May 28, 2025, more than 17 million non-redundant genome sequences have been released globally, involving 182 countries/regions and covering more than 9100 sampling sites. The genome sequence data of SARS-CoV-2 grew rapidly in the past years, and continues to increase slowly (increasing by ~1000 sequences every day) .

Data integration: automated curation and standardization of integration models



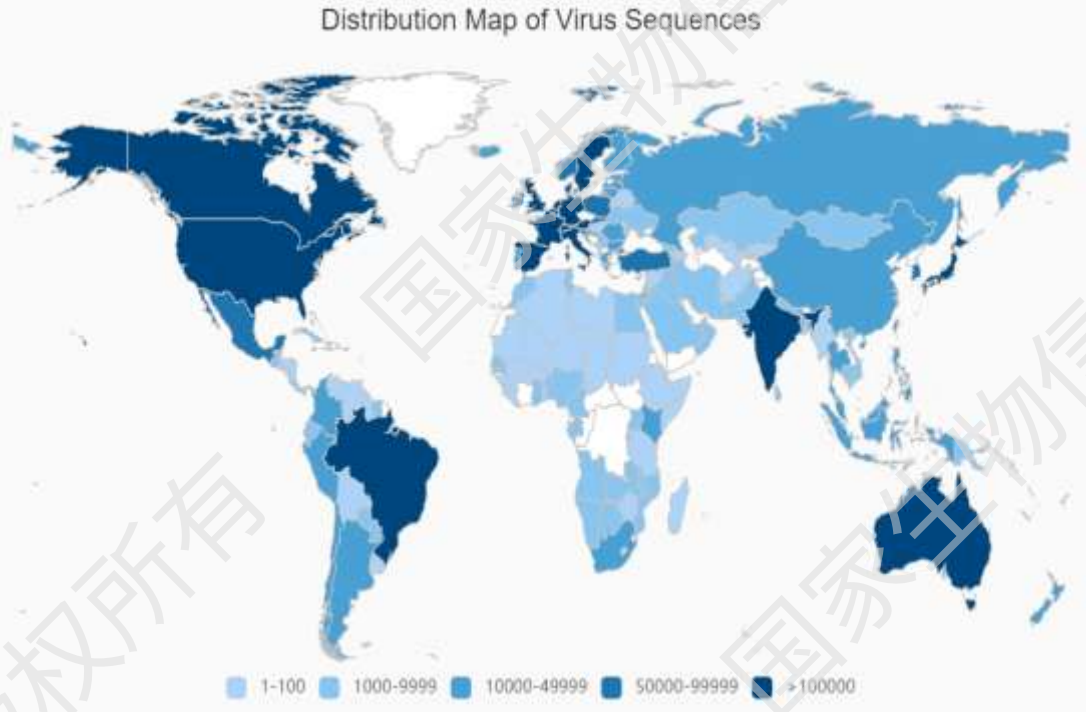
Assessment

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

国家生物信息中心版权所有

Global distribution of SARS-CoV-2 sequences

50004651	16929831	17056997	3337	13077	9444
CORONAVIRUS SEQUENCES	ISOLATES	SARS-COV-2 SEQUENCES	SUBMITTING LABS	ORIGINATING LABS	LOCATIONS



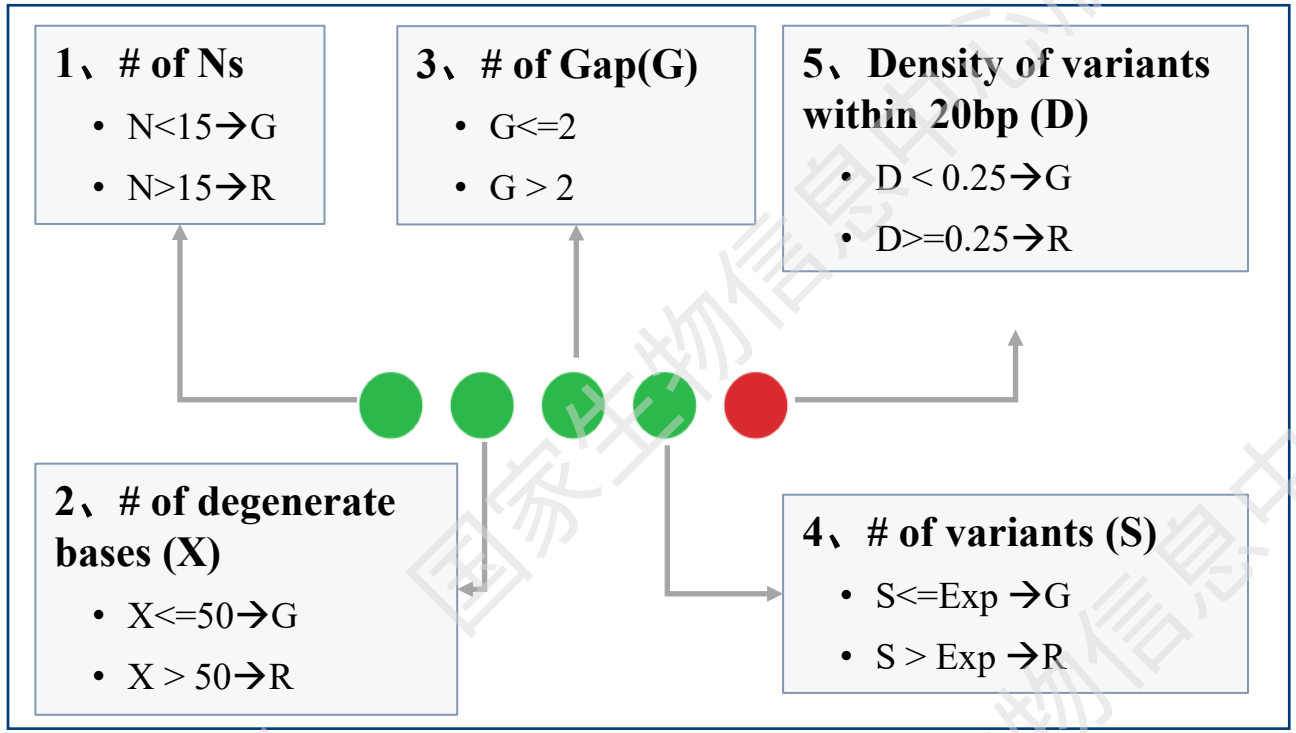
Continent	Country/Region	Genome Sequences	Complete Genome Sequences	Human Genome Sequences	Human Complete Genome Sequences	Monitoring report
North America	United States	4273589	3786596	4271821	3784867	
Europe	United Kingdom	2985057	2733257	2985049	2733253	
Europe	Germany	780764	720943	780753	720932	
Europe	France	598028	496370	598001	496347	
Europe	Denmark	571947	552836	571476	552365	
North America	Canada	430788	412245	430775	412232	
Asia	Japan	416096	413816	416087	413807	
Asia	India	255402	206052	255357	206025	
Europe	Sweden	212763	196109	212760	196106	
South America	Brazil	181412	171236	181394	171233	

Showing 1 to 10 of 192 entries

Previous 1 2 3 4 5 Next

https://ngdc.cncb.ac.cn/ncov/release_genome

Assessment of genome integrity and quality



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

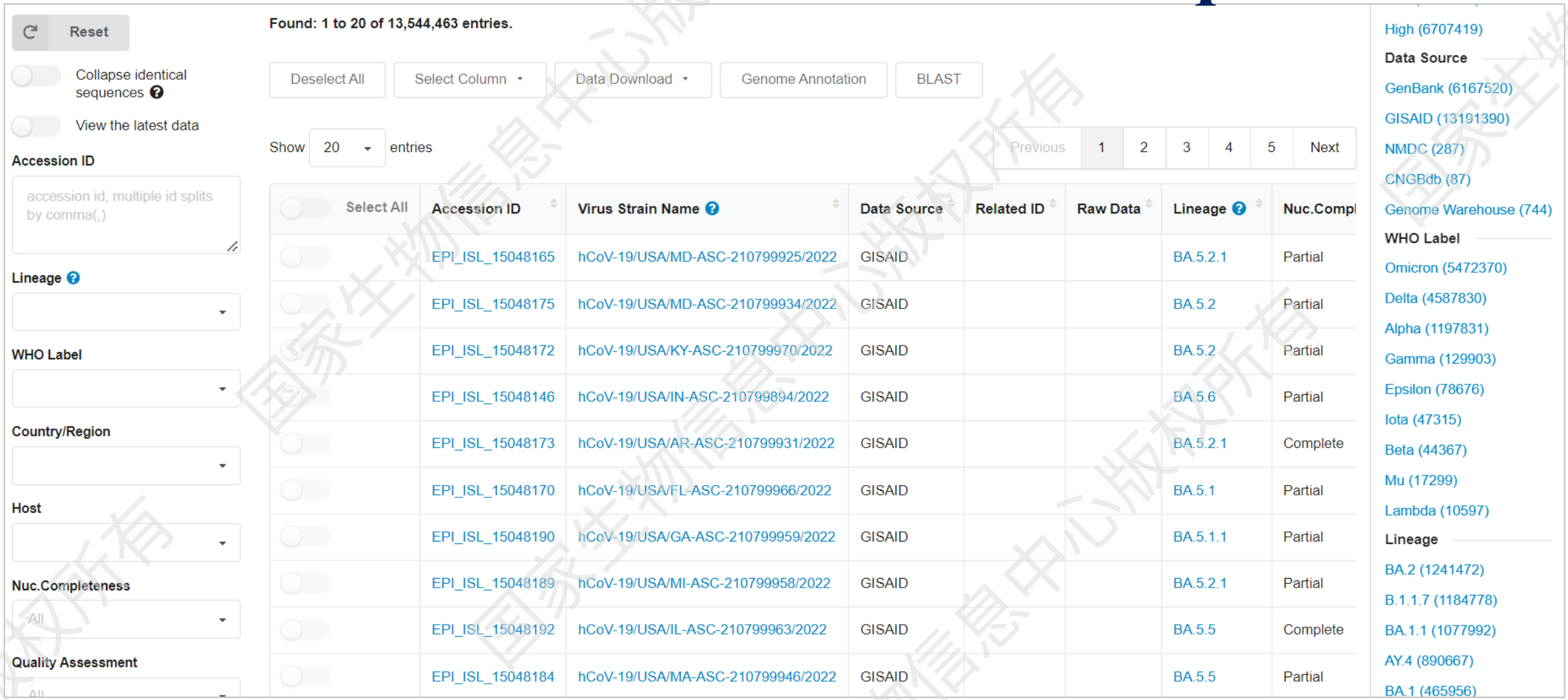
~48% complete & high quality

Data Release Statistics

41266539	13485509	13544463	3214	13104	8803
CORONAVIRUS SEQUENCES	ISOLATES	SARS-COV-2 SEQUENCES	SUBMITTING LABS	ORIGINATING LABS	LOCATIONS

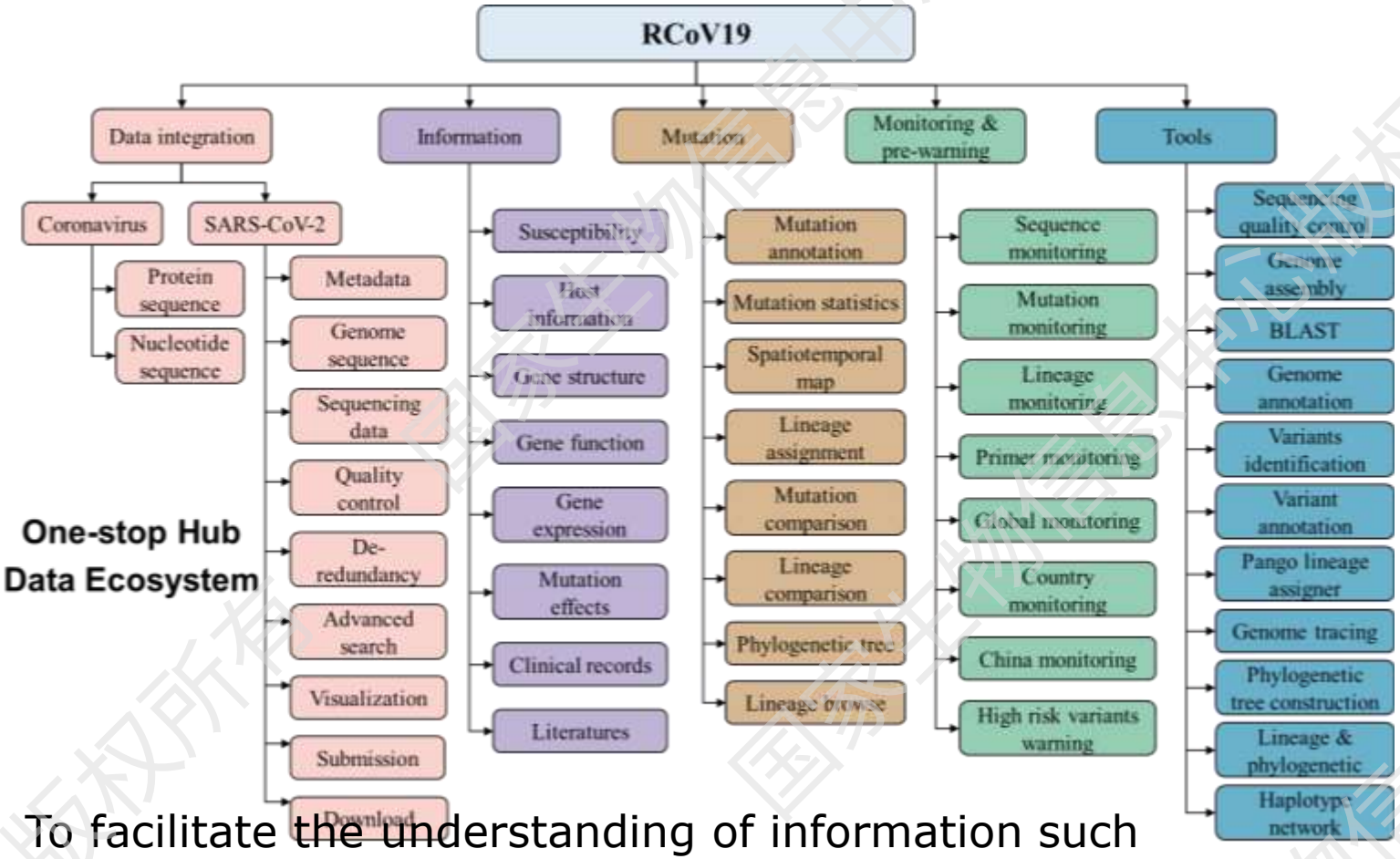
Genomics Proteomics & Bioinformatics (2020)

Search module for all released sequences

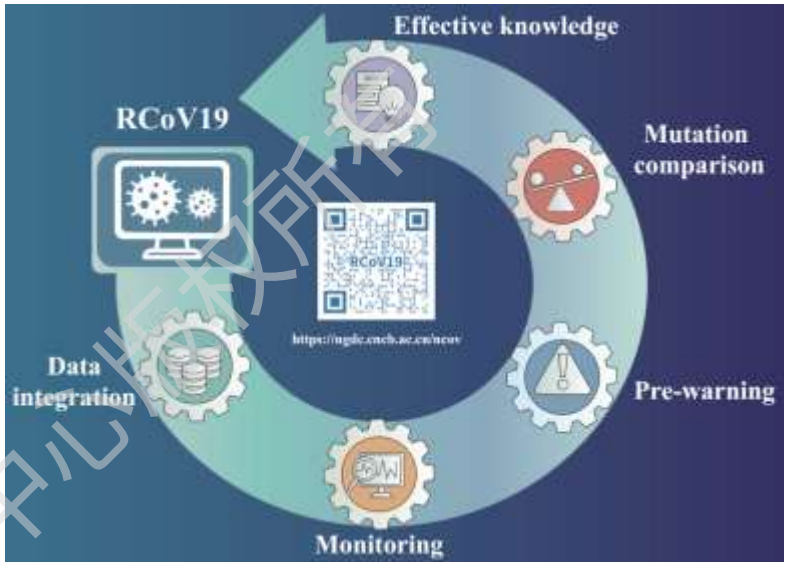


https://ngdc.cncb.ac.cn/ncov/genome/search

World's Largest and Most Comprehensive Public Resource for SARS-CoV-2



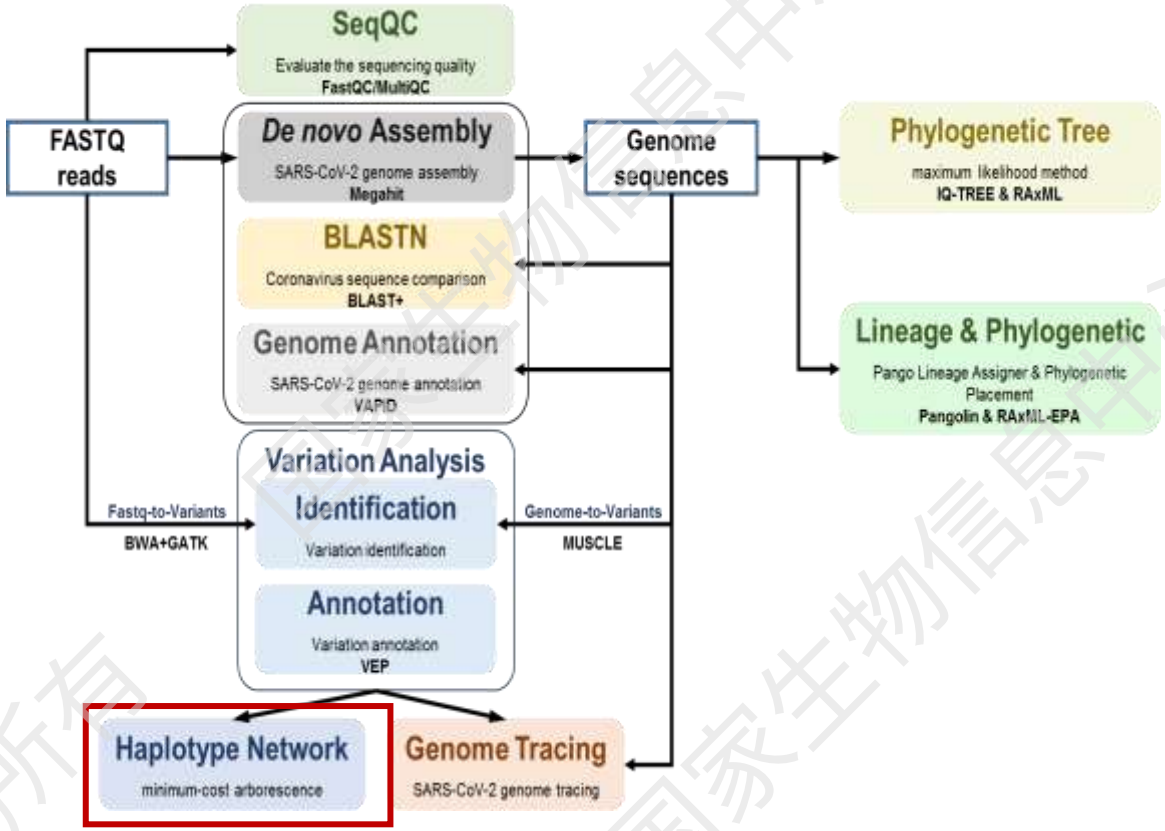
- 5 Key Functional Categories
- 50 Functional Modules
- 17M Data Integration



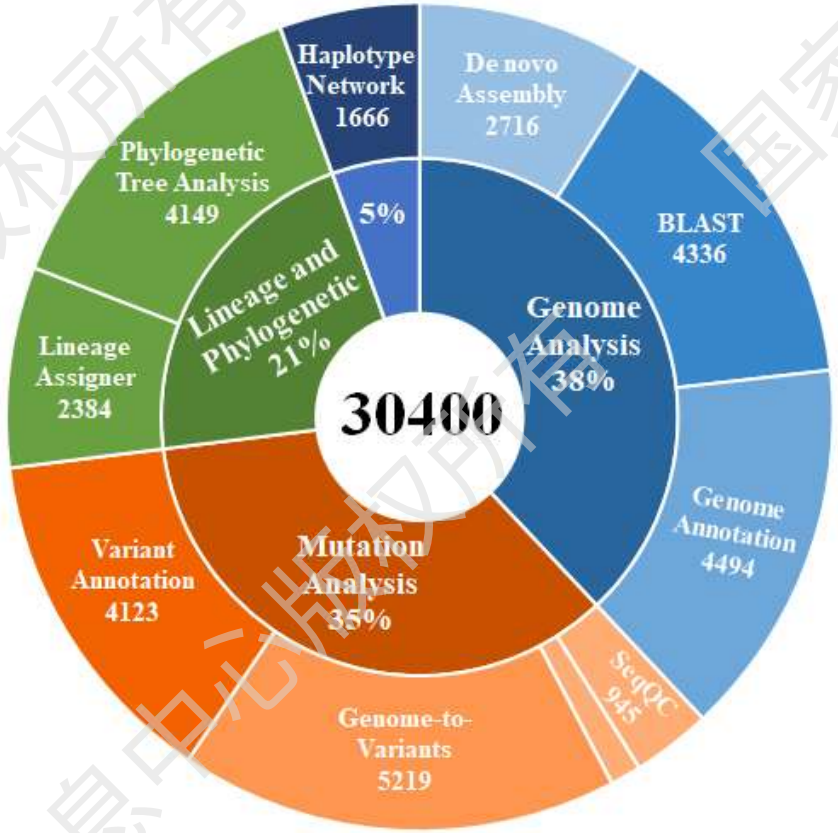
Data, Information, Knowledge, & Application

To facilitate the understanding of information such as mutation and evolution

Processing workflow and services



Processing workflow of the resource for coronavirus 2019



Statistics of platform users

* as of February 16, 2023

Zoological Research (2020)

猴 痘

国家生物信息中心版权所有

Part 2

McAN: A Novel Algorithm for Haplotype Network Construction

国家生物信息中心版权所有

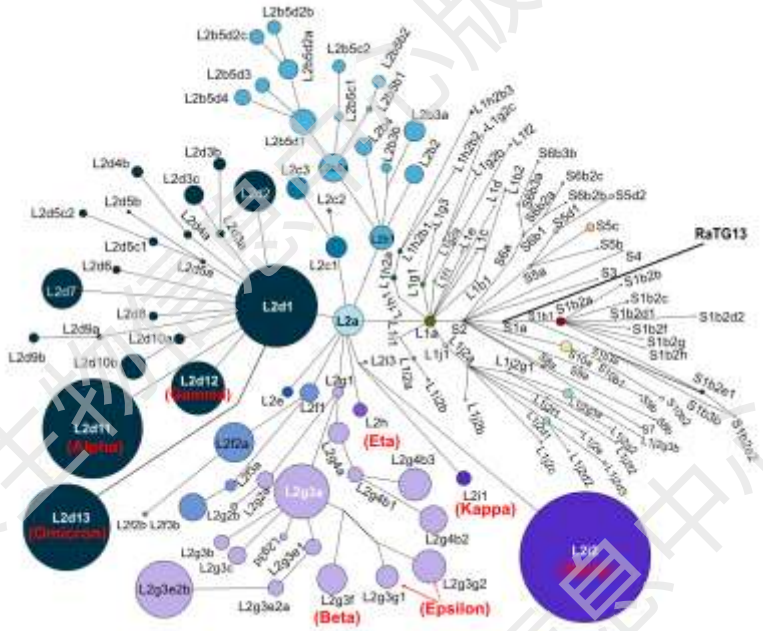
Haplotype network

Haplotype network is a powerful tool to trace the origin and evolution of epidemics.

A haplotype is a group of variants/genes within an organism that was inherited together from a single parent

Haplotype network	Biological meaning
Vertex/node	Haplotype
Edge	Evolutionary relationships
Size of vertex	# sequences
Length of edge	# different mutations

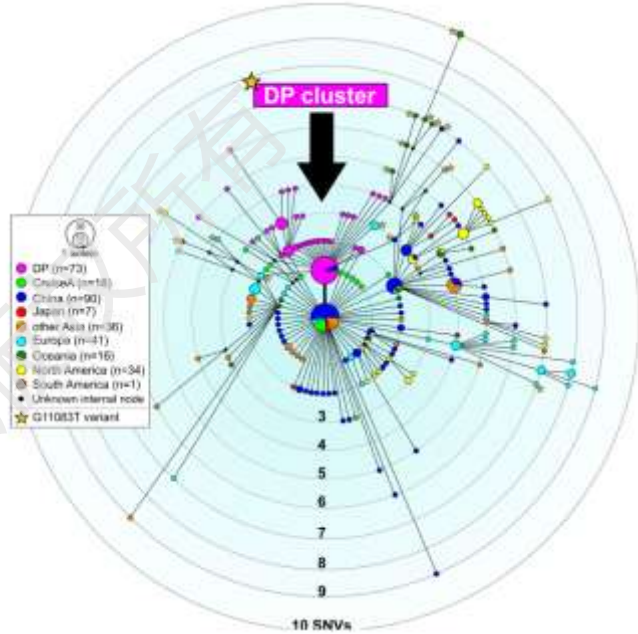
Tracking the evolution of SARS-CoV-2 (global)



Haplotype network of L/S clades of SARS-CoV-2

***Medical Review* (2022)**

Tracking the evolution of SARS-CoV-2 (local)



Haplotype network of SARS-CoV-2 from Diamond Princess

***PNAS* (2020)**

Haplotype network

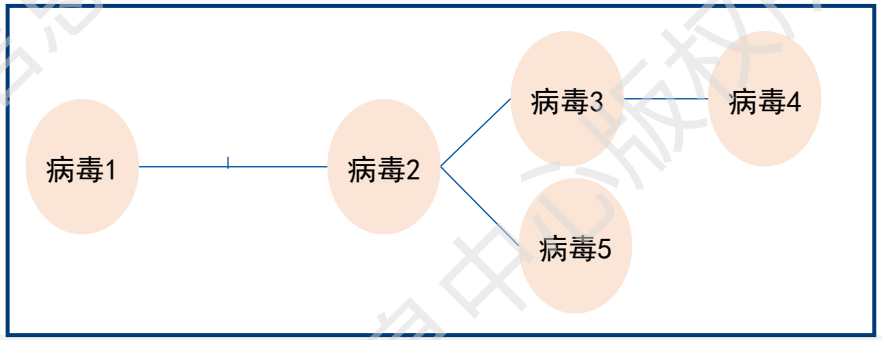
假如有如下5株病毒序列，变异信息如下：

```
病毒1：TCAGACTTACCTACGTACCTATTGCACCCAATTTACGTACTAGACGA
病毒2：TCAAACTTACCTACTTACCTATTGCACCCAATTTACGTACTAGACGA
病毒3：TCAAACTTACCTACTTACCTATTGCACCCATTTTACGTACTAGACGA
病毒4：TCAAACTTACCTACTTACCTATTGCACCCATTTTACGTAGTAGACGA
病毒5：TCAAACTTACCTACTTACCTATTGCACCCAATTTACCTACTAGACGA
```

单倍体型

```
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
```

单倍体型图演化关系



顶点：单体型
边：演化关系
点的大小：序列数
边长：不同碱基数

- “A 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。

The state-of-the-art algorithms

Challenge: The traditional haplotype network construction algorithms, such as MJN, TCS, MSN, and RMST, can not quickly process massive virus genome data.

For example, it takes more than 3 days to construct a haplotype network of 1000 SARS-CoV-2 sequences via MJN.

Algorithm	Characteristic		Time complexity	Release time
MJN (Median Joining Network)	Add unsampled nodes	• Running time • Memory costs • Collection date	$O(m^{2.2})$	1999 *Mol Biol Evol*
TCS (Templeton Method)	Add unsampled nodes		$O(m^5)$	1992 *Genetics*
MSN (Minimum Spanning Network)	Faster than other 3 algorithm		$O(m^2)$	1999 *Mol Biol Evol*
RMST (Randomized minimum spanning tree)	less alternative links than MSN		$O(m^2)$	2018 *Methods Ecol Evol*

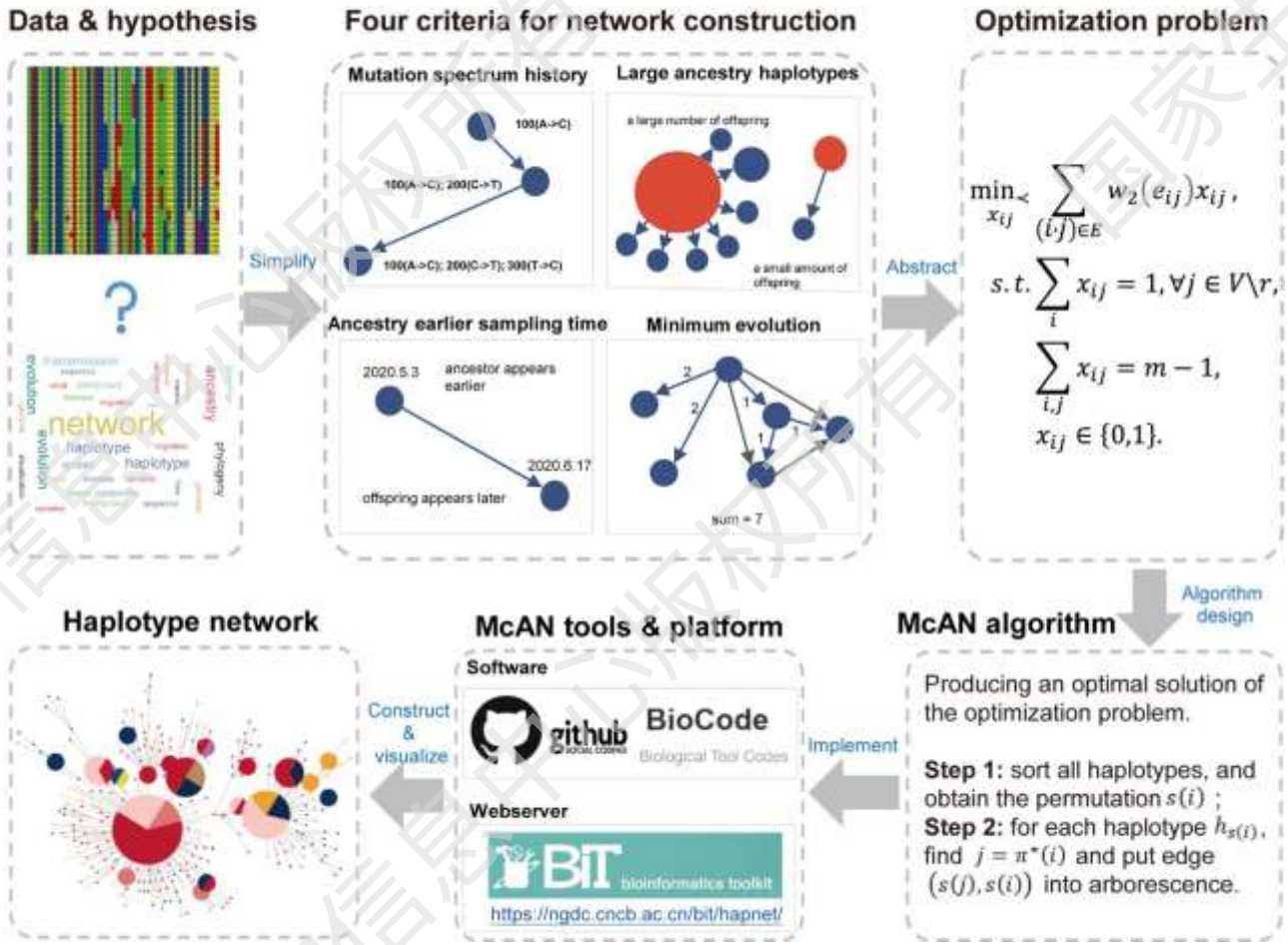
Overview of McAN algorithm and platform

- **McAN algorithm**

Minimum cost arborescence network (McAN) algorithm constructs haplotype network by considering mutation spectrum history, node size and sampling time simultaneously.

- **Platform**

McAN platform provides a user-friendly web server as well as an off line tool, allowing users to construct haplotype network, detect lineages and visualize the results interactively online.

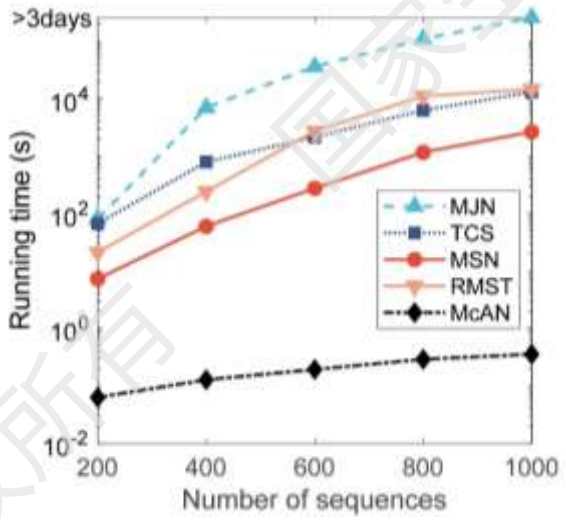


Schematic illustration for haplotype network construction by McAN.

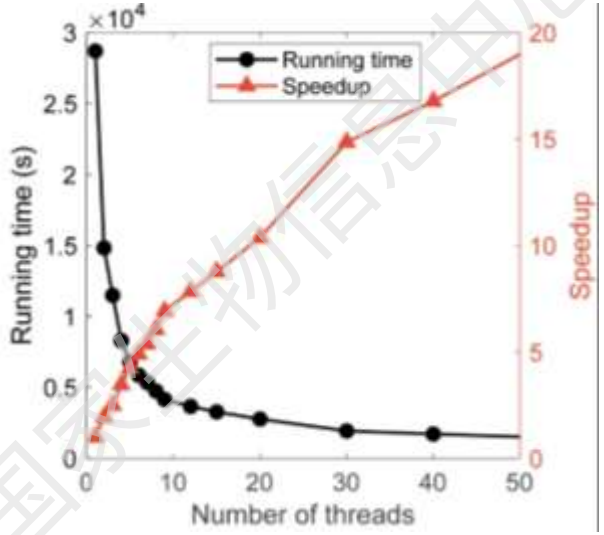
Briefings in Bioinformatics (2023)

Performance on simulated and real data sets

- **Speed up：** McAN is at least **100 times faster** than the state-of-the-art algorithms; takes only about 25 minutes (50 threads) for 5 million SARS-CoV-2 sequences.
- **Memory saving：** McAN **consumed less than 10% of the memory** of the other four state-of-the-art algorithms.
- **High accuracy：** McAN's accuracy is almost as high as that of the state-of-the-art methods.



Runtime of McAN, MJN, TCS, MSN, MJN and RMST on SARS-CoV-2 dataset (single thread).



Running time of McAN on 4,990,399 SARS-CoV-2 sequences (multi-threaded).

Algorithms	AUC	Memory cost (KB)
McAN	0.997	5,844
MSN	0.998	59,960
MJN	0.997	63,676
TCS	0.997	62,928
RMST	0.998	114,725

Accuracy and memory consumption of McAN, MJN, TCS, MSN, MJN, and RMST on a dataset containing 1001 simulated sequences.

Briefings in Bioinformatics (2023)

Comparison with traditional algorithm

- Comparison of McAN algorithm with traditional methods in terms of output network properties, time complexity, etc

Table S1. Comparison of different haplotype network construction algorithms.

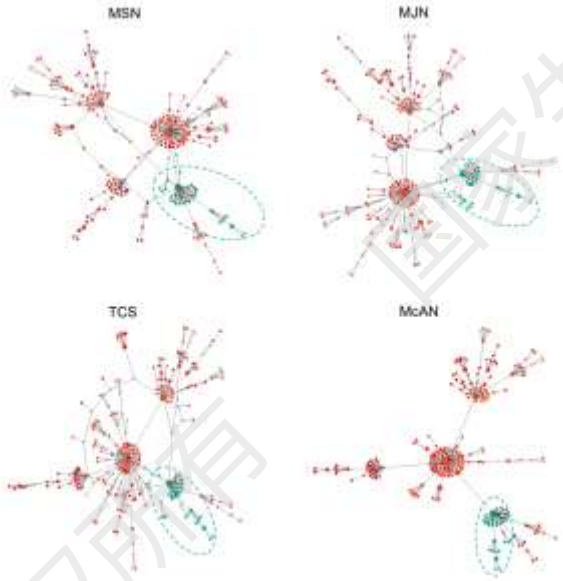
Algorithms[a]	Output	Distance matrix[b]	Time complexity[c]	
			Worst- or average-case	Best-case
MJN	Undirected graph	Yes	$O(n^{2.2})$ (average-case)	$\geq O(m^2)$
Dist + MSN	Undirected graph	Yes	$O(m^2)$ (worst-case)	$O(m^2)$
Dist + MST	Undirected tree	Yes	$O(m^2)$ (worst-case)	$O(m^2)$
Dist + RMST	Undirected graph	Yes	$O(m^2)$ (worst-case)	$O(m^2)$
TCS	Undirected graph	Yes	$O(m^5)$ (worst-case)	$\geq O(m^2)$
McAN	Arborescence	No	$O(m^2)$ (worst-case)	$O(m \log m)$

Note: [a] MJN, median-joining networks; Dist, the step of calculating distance matrix; MSN, minimum spanning network; MST, minimum spanning tree; RMST, randomized minimum spanning tree; TCS, Templeton Crandall and Sing algorithm; McAN, minimum-cost arborescence haplotype network. [b] whether the method need to calculate and store the distance matrix of haplotype. [c] n is the number of sequences, and m is the number of haplotypes. The time complexity of each method includes the calculation of distance matrix of haplotypes if the method takes distance matrix of haplotypes as input.

Tracing the evolution via McAN

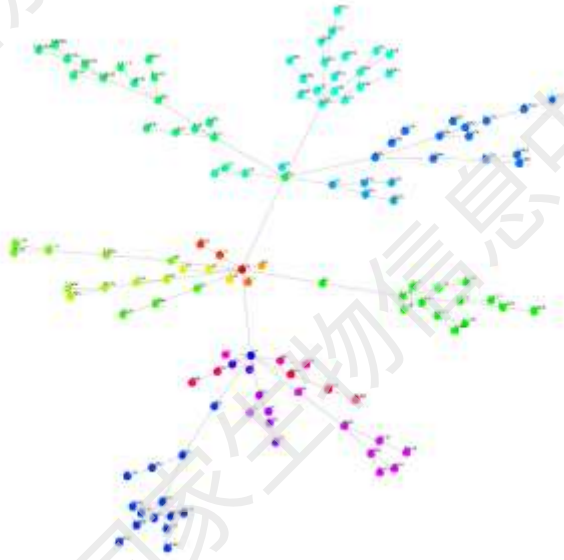
McAN can be used to trace the evolution of a variety of viruses, including SARS-CoV-2, Monkeypod virus, and Influenza A virus.

SARS-CoV-2 (local)



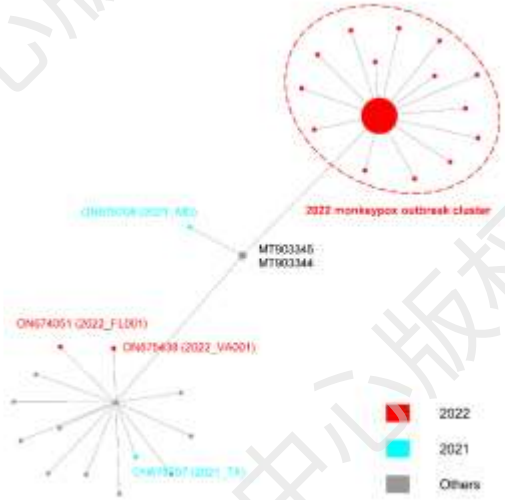
Haplotype networks of SARS-CoV-2 sequences from Diamond Princess (DP) dataset.

SARS-CoV-2 (global)



Haplotype network of 130 SARS-CoV-2 lineages constructed by McAN.

Monkeypox virus



The haplotype network on 44 whole genome sequences of monkeypox viruses by McAN.

Influenza A virus

Haplotype network of 21 human IAV sequences constructed by McAN.

Briefings in Bioinformatics (2023)

Service and sharing

- **Platform**: The McAN algorithm has been applied to the RCoV19 to provide services for researchers.
- **Open source**: The McAN source code is available at BioCode in April 2022 and has been downloaded about seven thousand times (as of May 17, 2024).

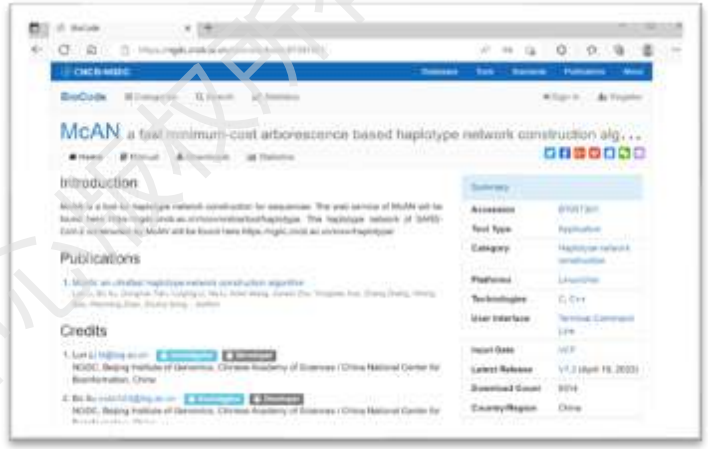
Haplotype network visualization

https://ngdc.cncb.ac.cn/ncov/haplotype

Haplotype network platform

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

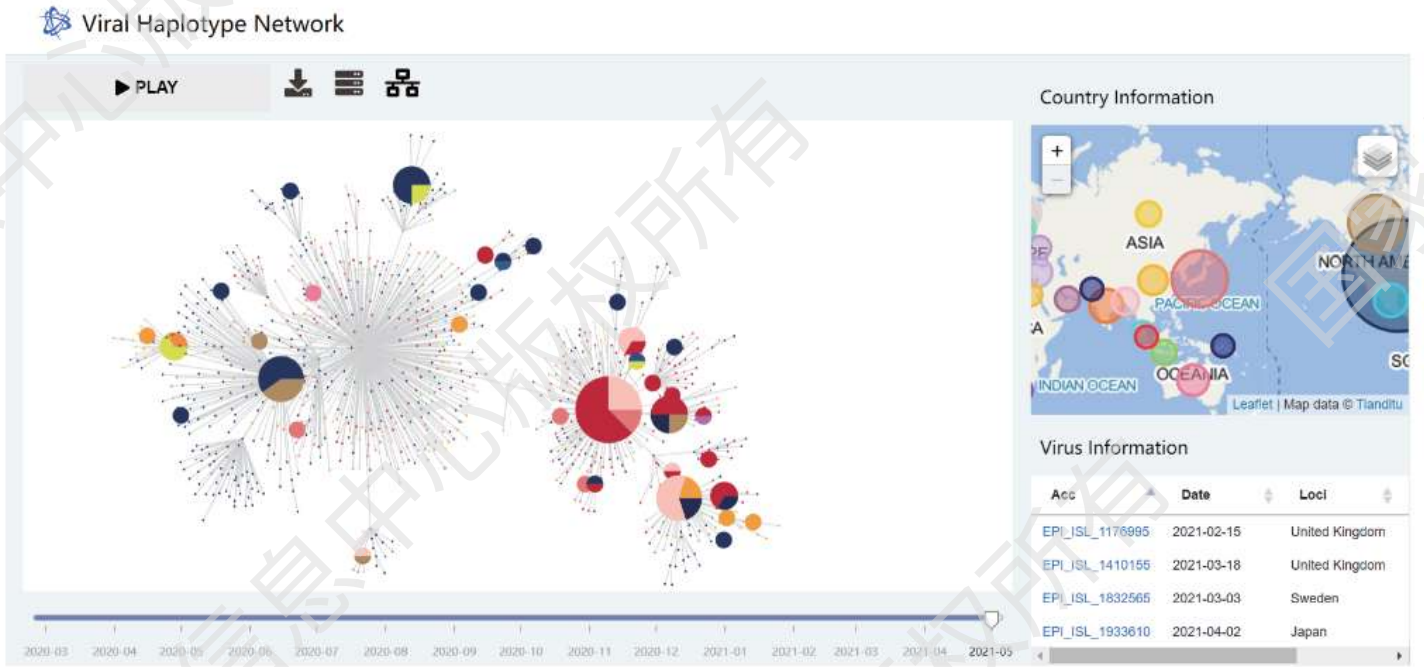
Source code at BioCode



https://ngdc.cncb.ac.cn/biocode/tools/BT007301

Briefings in Bioinformatics (2023)

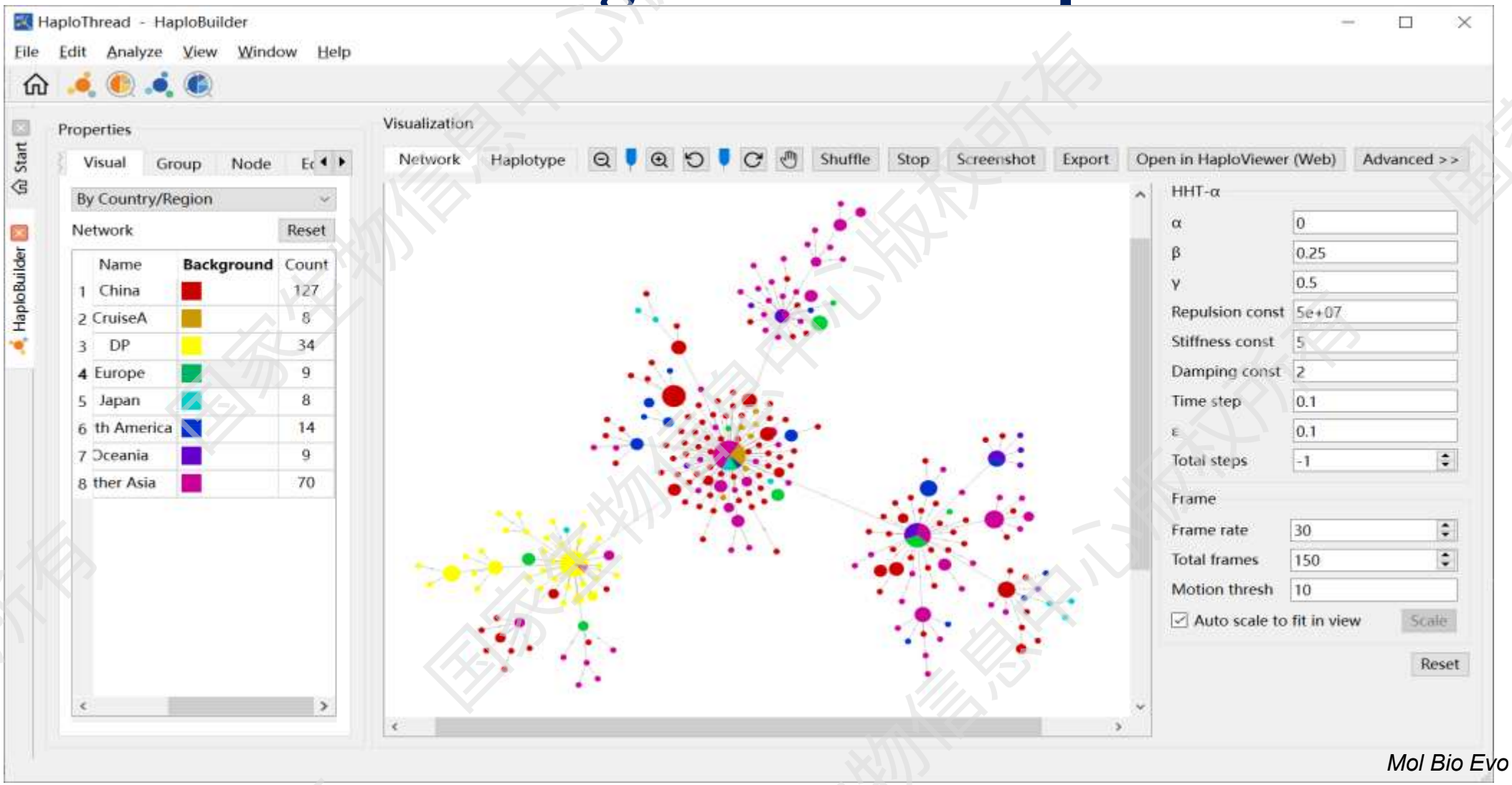
Online Service for McAN



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

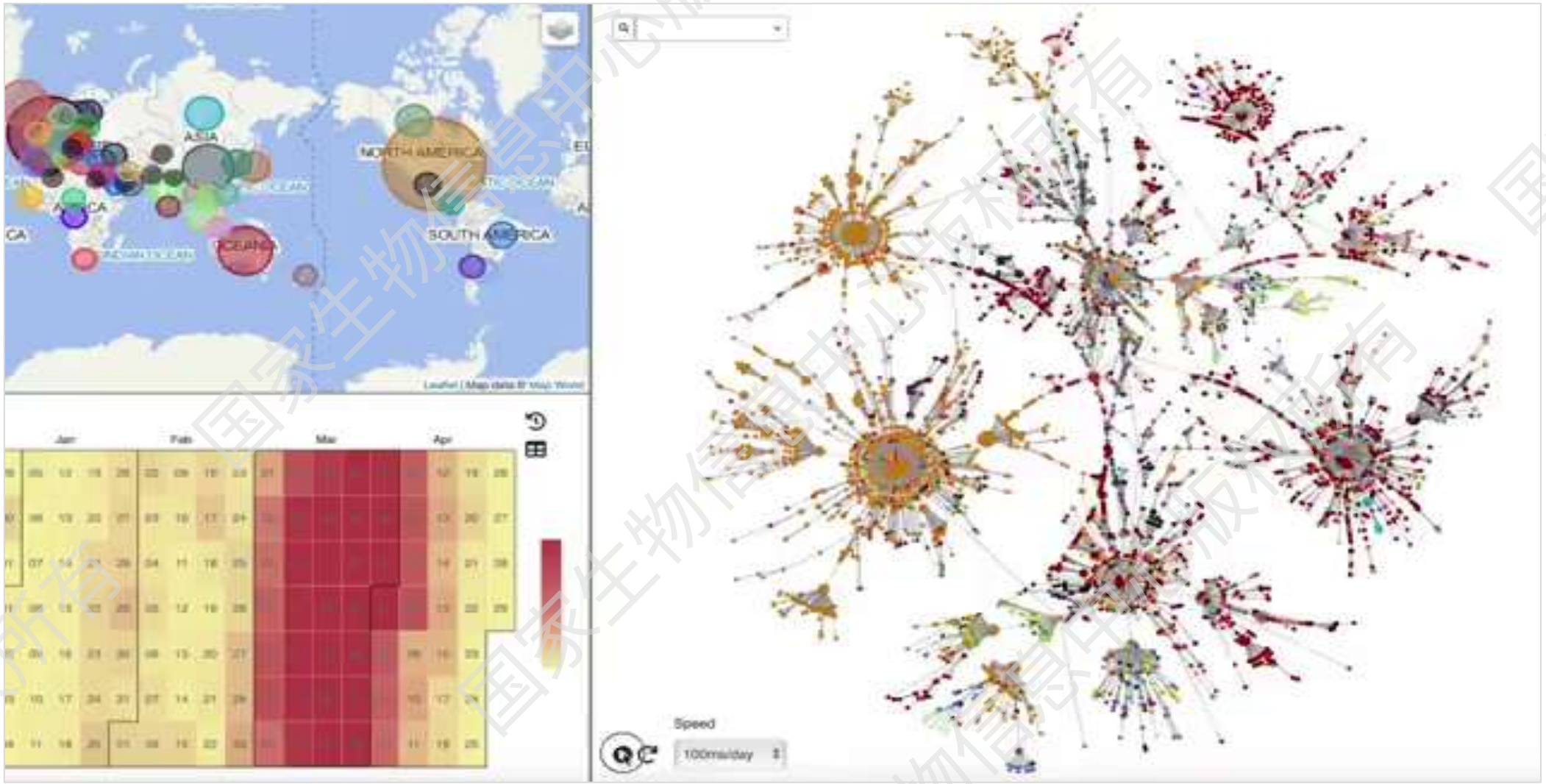
McAN源代码下载量超1.2万次，在线工具服务1730次

HaploThread:
A Scalable Integrated Desktop Platform



Mol Bio Evo (under review)

Haplotye network evolution dynamic tracking platform



Part 3

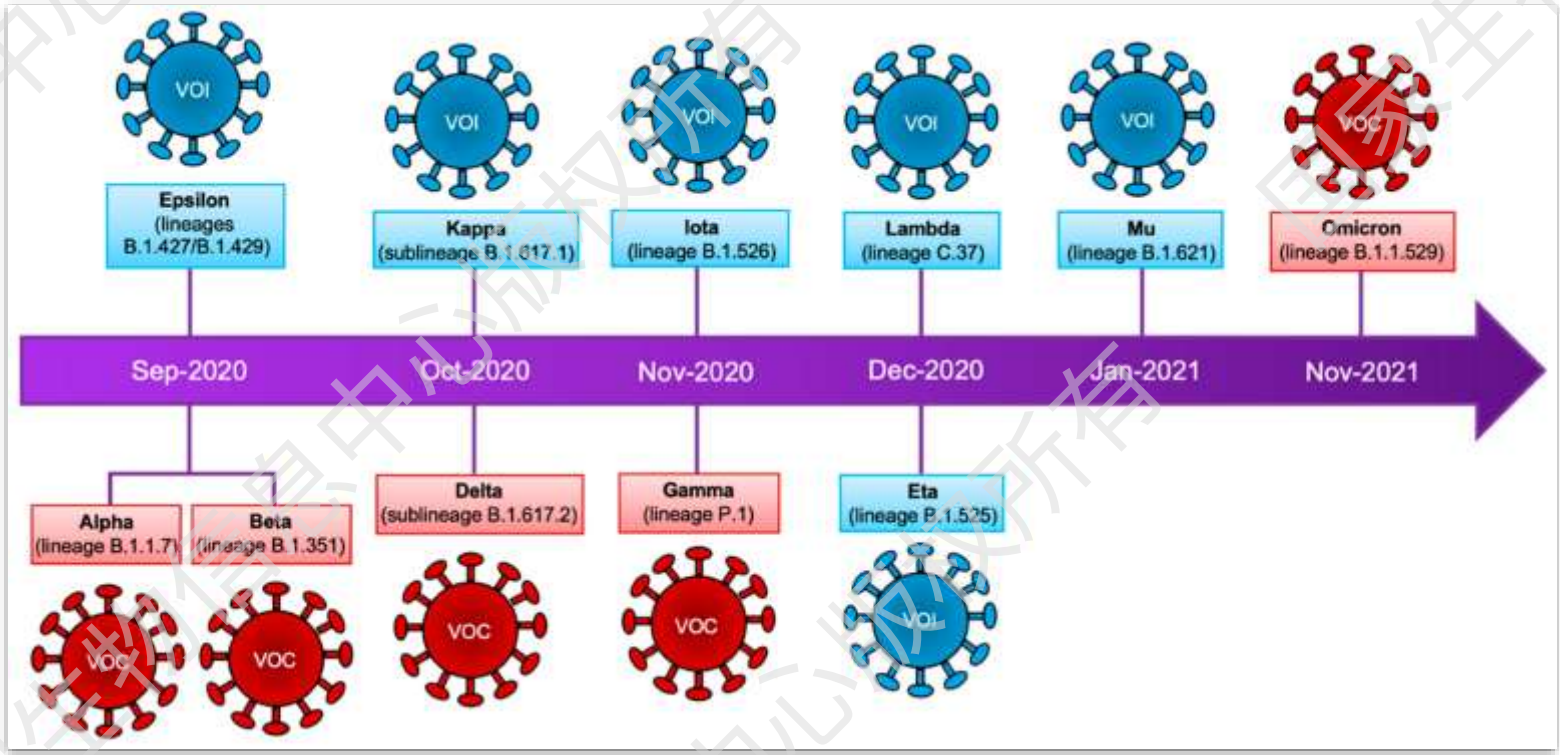
HiRisk-Detector: High-risk Variants Pre-warning

国家生物信息中心版权所有

High-risk variants of SARS-CoV-2

- **Variants of the SARS-CoV-2 continue to emerge**

some may affect the virus's properties, such as how easily it spreads, the associated disease severity, or the performance of vaccines, therapeutic medicines, diagnostic tools, or other public health and social measures.

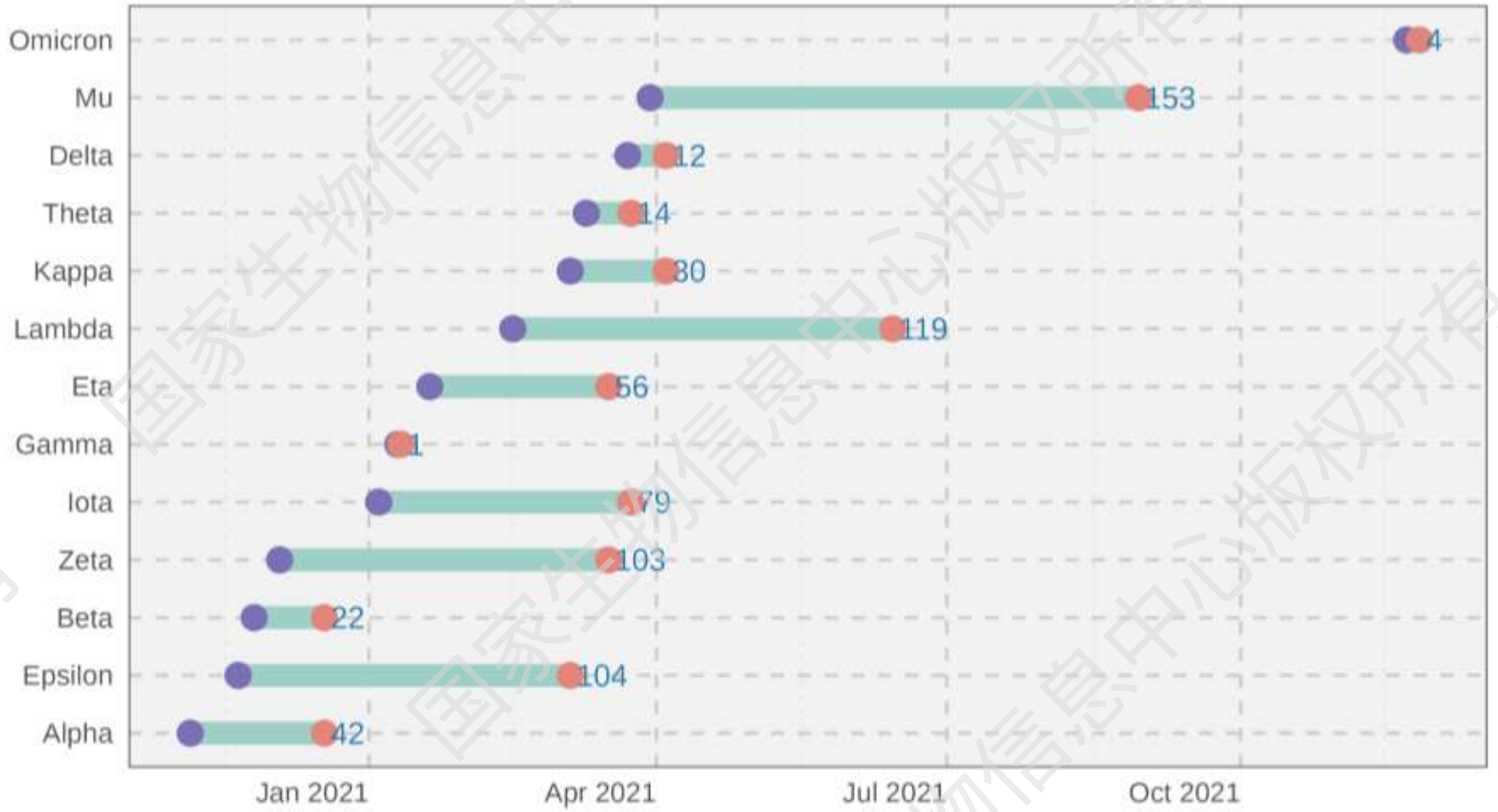


Timeline that summarizes the emergence of SARS-CoV-2 variants

- **Problem**: automatically detect high-risk variants of SARS-CoV-2 e.g., VOC, VOI, and VUM
- **Purpose**: prioritize global monitoring and research, and inform the COVID-19 response

Viruses (2022)

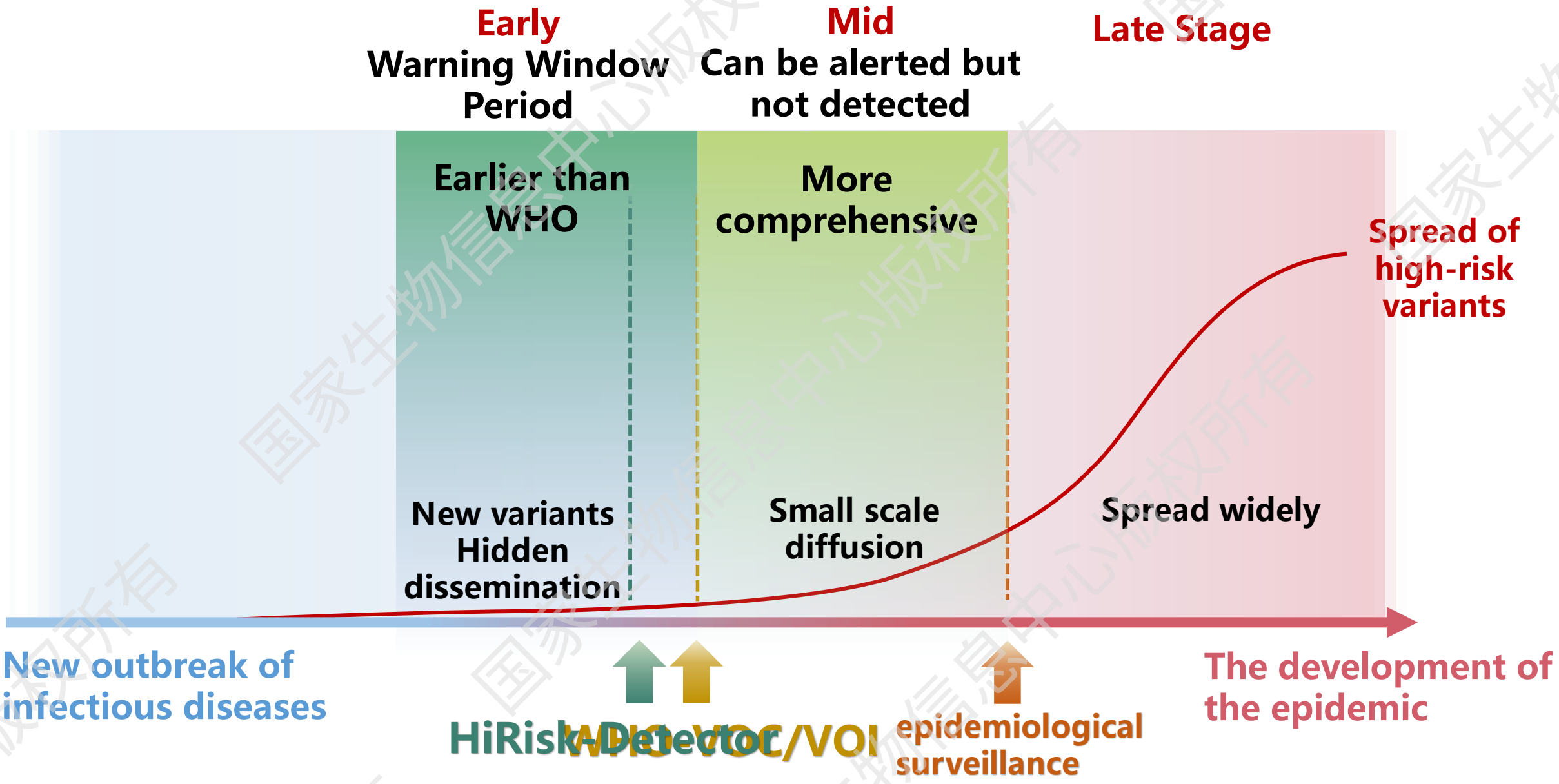
The lagged time for WHO determined high risk variants



国家生物信息中心版权所有

Methods of high-risk variants detection

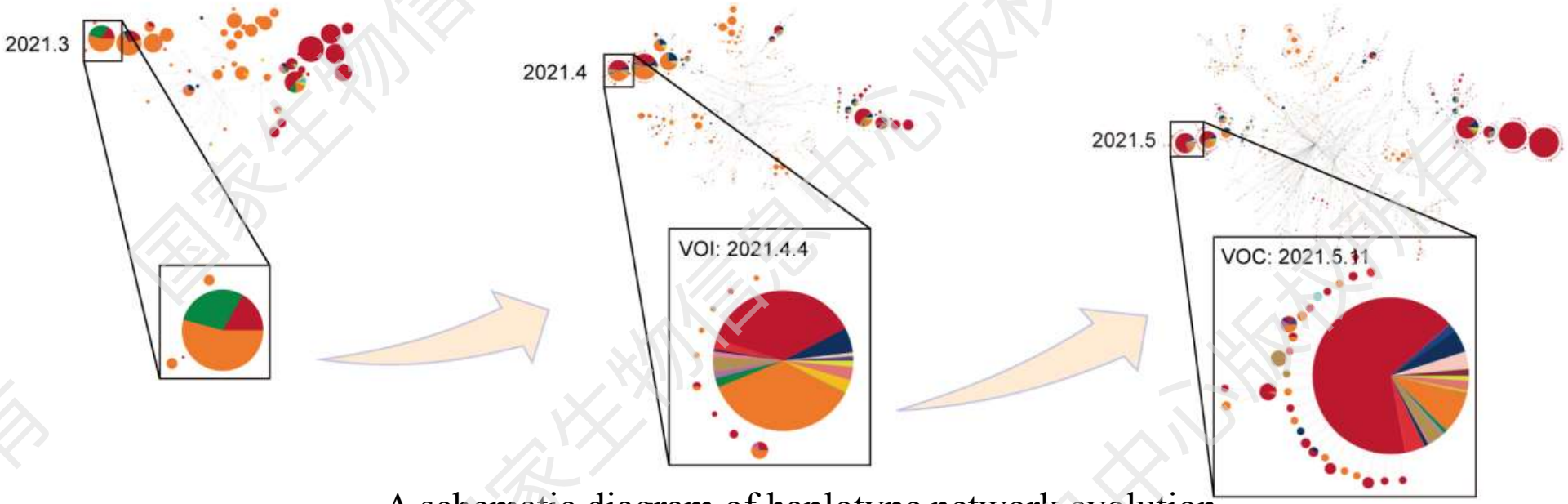
Categories	Name of organization/method	Characteristic
Semi-automatic method	WHO	• Reliance on experts • Relying on experts judgment
	CDC	
	ECDC	
Automatic methods	VarEPS	• Without the need for human intervention • Machine learning • No platform is available
	PyR_0	
	Predicting the mutational drivers of future SARS-CoV-2 variants of concern	
Wet experiment	Enhanced fitness of SARS-CoV-2 variant of concern Alpha but not Beta	• High costs • High reliability • It will take a lot of time



国家生物信息中心版权所有

Haplotype network and high-risk variants

- **Since WHO defined Delta as VOI**, the haplotype networks of Delta exhibit unique characteristics, e.g., out-degree and number of countries are increasing rapidly

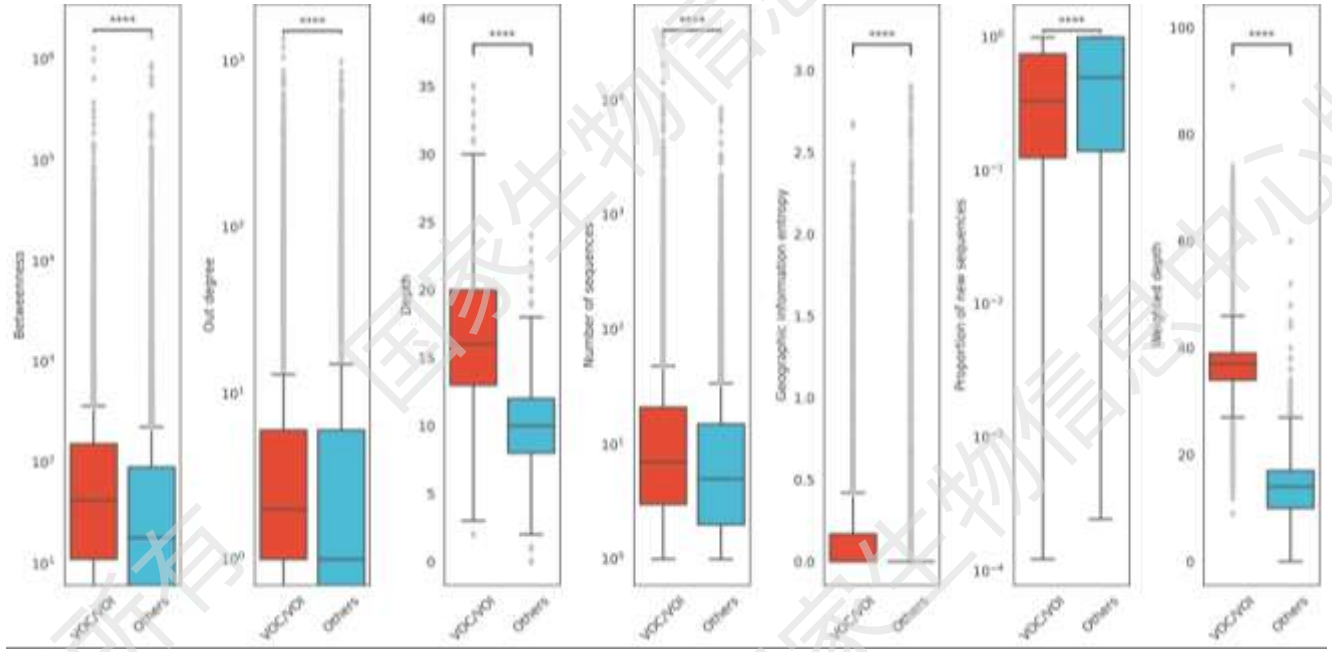


A schematic diagram of haplotype network evolution

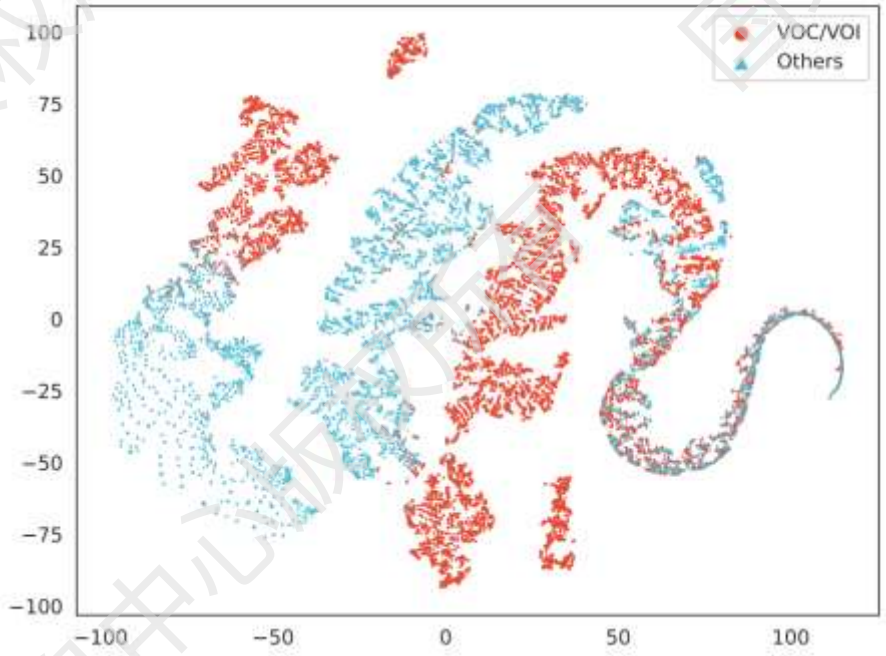
- **Hypothesis: haplotype network features have the potential to help predict the risk of SARS-CoV-2 variants**

Verify the hypothesis

- **Hypothesis confirmed: haplotype network features have the potential to help predict high-risk variants**



Differential analysis between features of VOC/VOI and other haplotypes

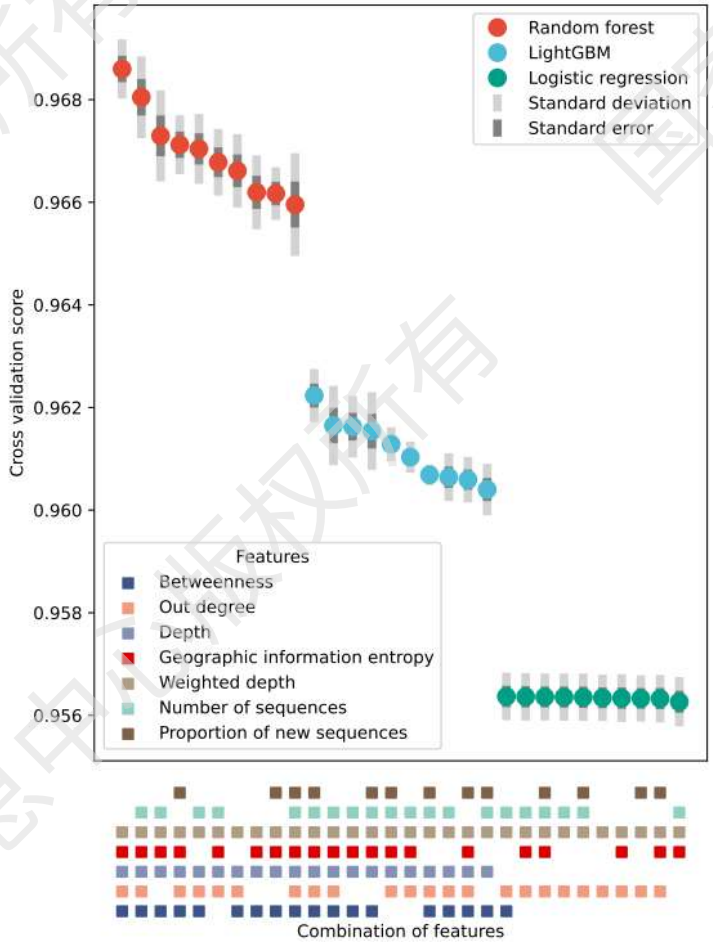


Dimensionality reduction with tSNE.

3/4 of the samples show clear boundaries between VOC/VOI and others in 2D projections

Detecting high-risk variants based on haplotype network features and machine learning

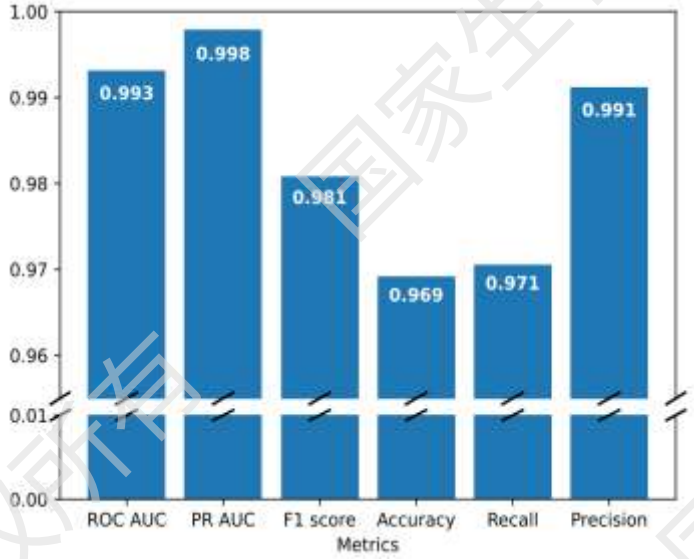
- **Label assignment:** VOC/VOI (high-risk), and other variants (low-risk)
- K-fold cross-validation (k=5)
- Select the AI method and features that achieve the best performance
- **Method:** random forest, LightGBM etc.
- **Features:** five of the seven features, including betweenness, out-degree, depth, geographic information entropy, and weighted depth



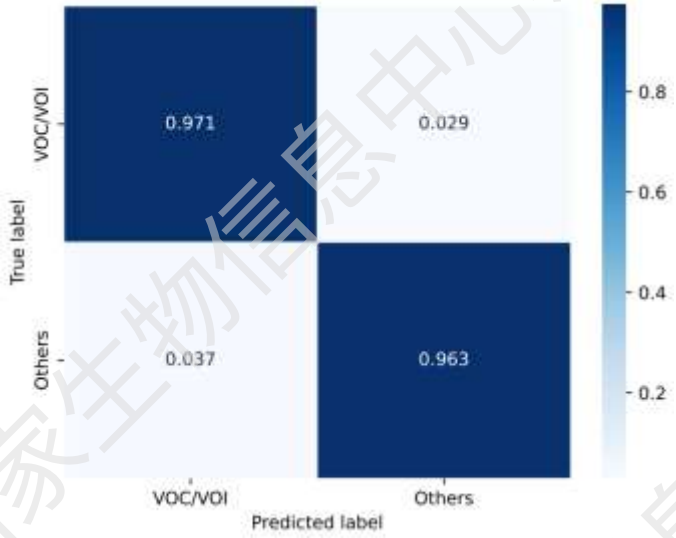
Performance on testing set

The five haplotype network features are powerful to help detect high-risk variants

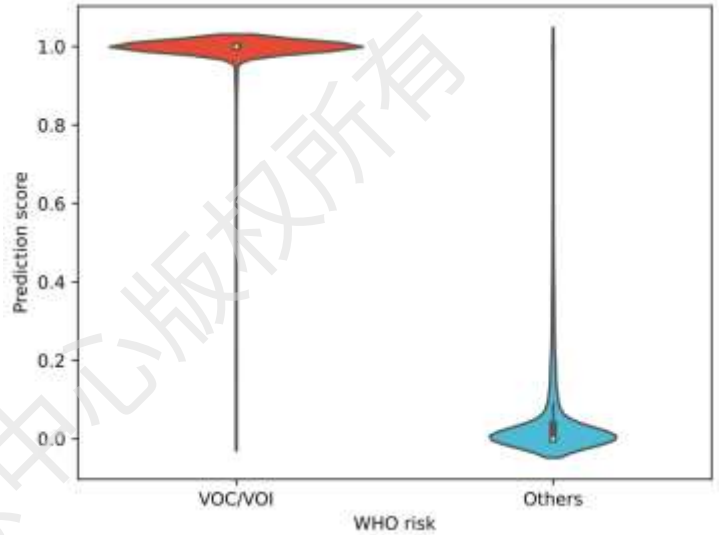
- ROC-AUC, PR-AUC, F1 score, accuracy, recall, and precision > 0.96
- TPR and TNR > 0.96, and FNR and FPR < 0.04
- Average prediction score: VOC/VOI = 0.97, others = 0.07



Six metrics including **ROC-AUC, PR-AUC, F1 score, accuracy, recall, and precision** on the testing set



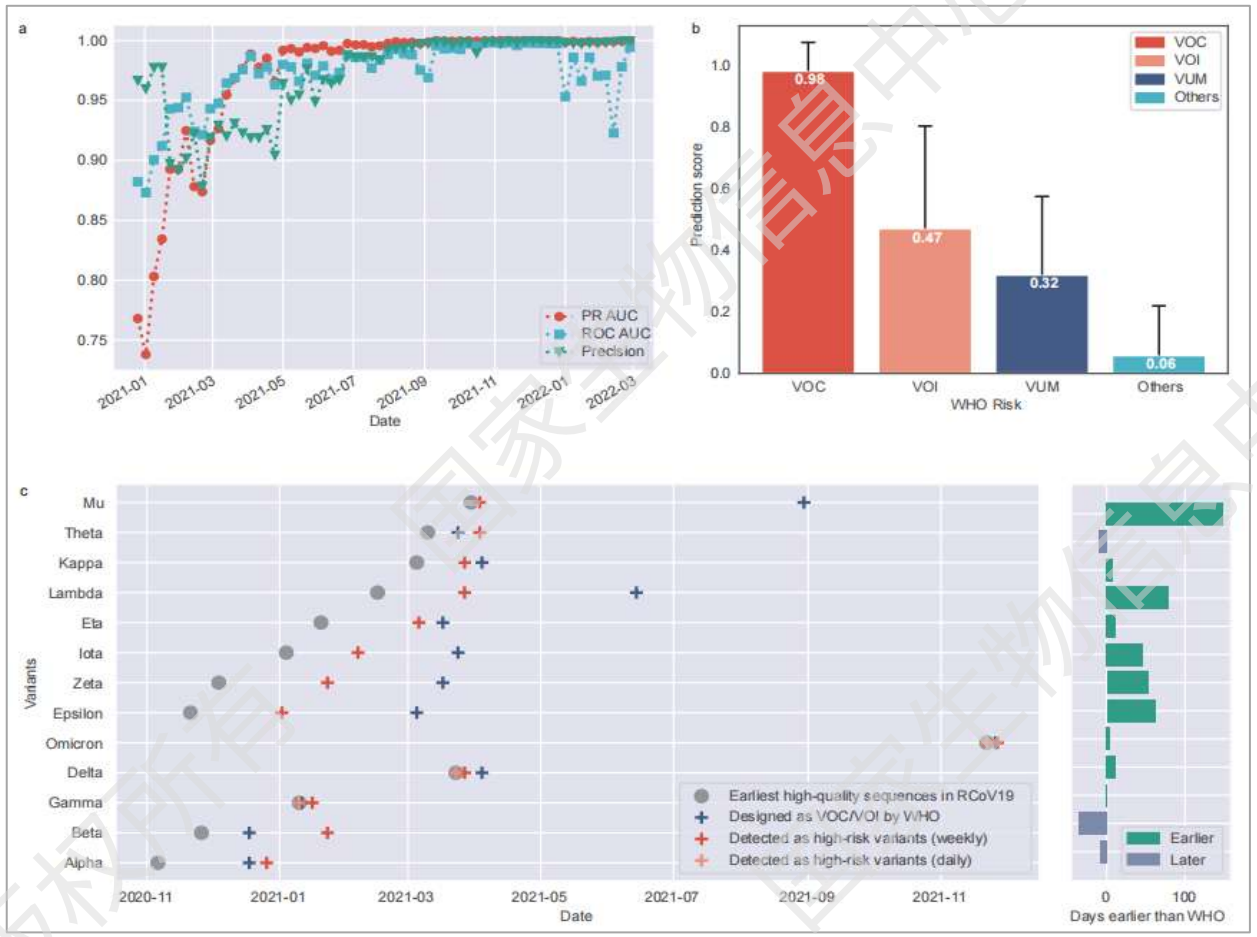
The normalized **confusion matrix**.



Prediction score of VOC/VOI and the other variants.

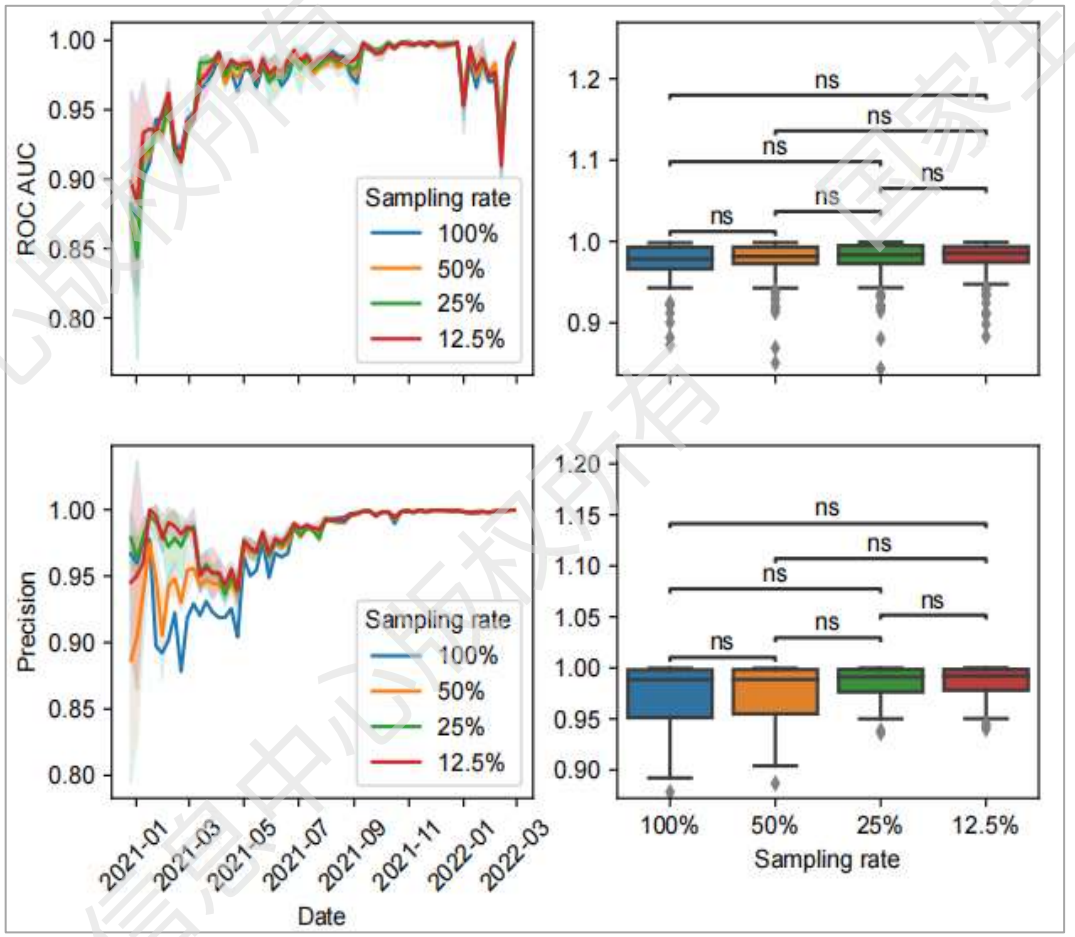
The Performance of HiRisk-Detector

High accuracy and precision



Successfully identified all 13 types of VOC/VOI, with an average warning time 27 days earlier than the time announced by the WHO; The average risk scores of VOC, VOI, VUM, and other variants decrease sequentially; All time point accuracy values are greater than 0.85

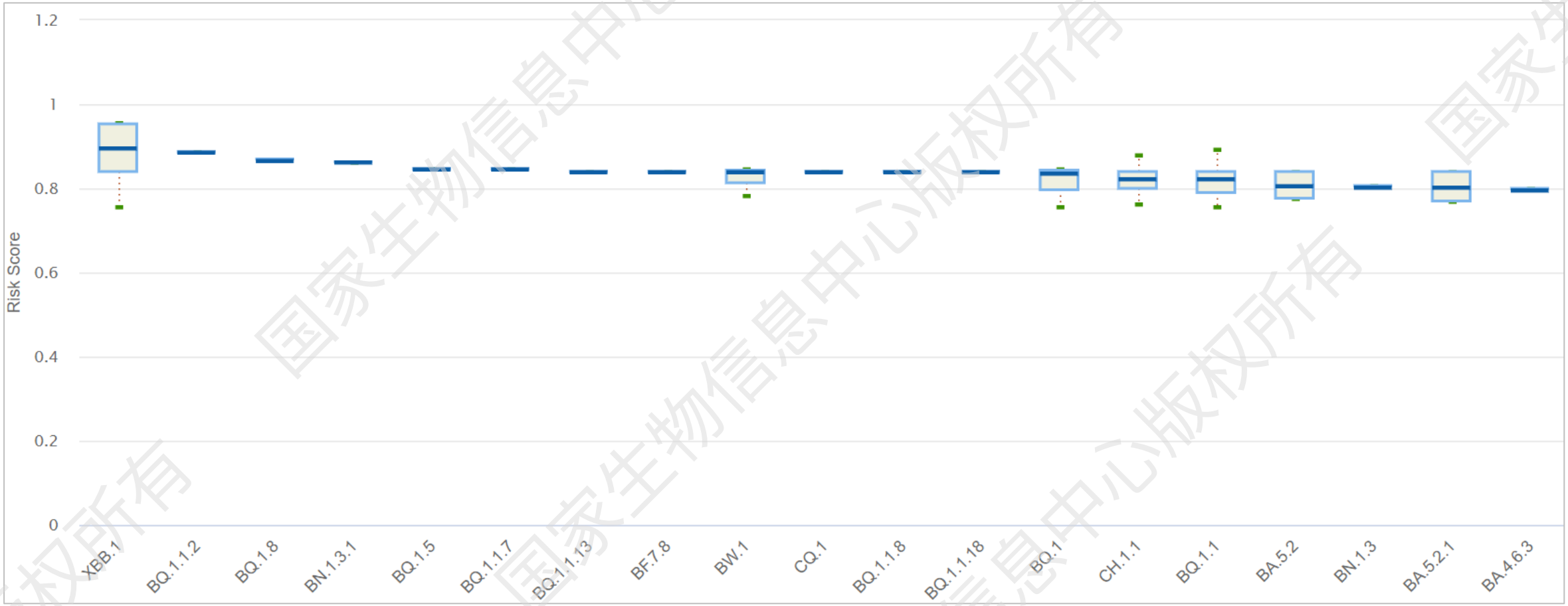
Robust to sequencing intensity



When the sequencing intensity decreases to actual values of 50%, 25%, and 12.5%, there is no significant change in performance indicators such as PR-AUC, ROC-AUC, and accuracy

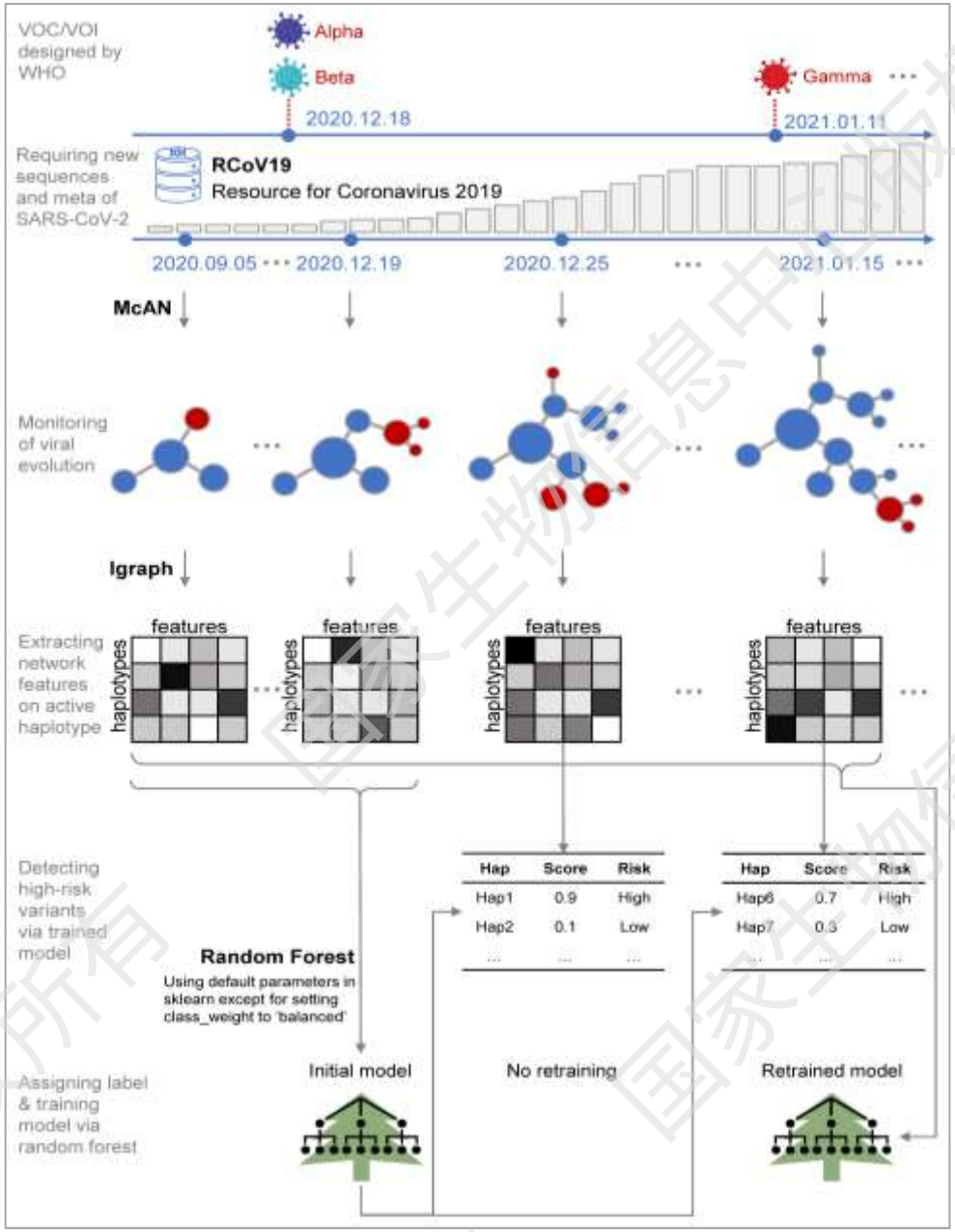
Advanced Science 2024

An example of the HiRisk-Predictor Results



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

国家生物信息中心版权所有



RESEARCH ARTICLE

ADVANCED SCIENCE
www.advancedscience.com

Machine Learning Early Detection of SARS-CoV-2 High-Risk Variants

Lun Li, Cuiping Li, Na Li, Dong Zou, Wenming Zhao, Hong Luo, Yongbiao Xue,* Zhang Zhang,* Yiming Bao,* and Shuhui Song**

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has evolved many high-risk variants, resulting in repeated COVID-19 waves over the past years. Therefore, accurate early warning of high-risk variants is vital for epidemic prevention and control. However, detecting high-risk variants through experimental and epidemiological research is time-consuming and often lags behind the emergence and spread of these variants. In this study, HiRisk-Detector a machine learning algorithm based on haplotype network, is developed for computationally early detecting high-risk SARS-CoV-2 variants. Leveraging over 7.6 million high-quality and complete SARS-CoV-2 genomes and metadata, the effectiveness, robustness, and generalizability of HiRisk-Detector are validated. First, HiRisk-Detector is evaluated on actual empirical data, successfully detecting all 13 high-risk variants, preceding World Health Organization announcements by 27 days on average. Second, its robustness is tested by reducing sequencing intensity to one-fourth, noting only a minimal delay of 3.8 days, demonstrating its effectiveness. Third, HiRisk-Detector is applied to detect risks among SARS-CoV-2 Omicron variant sub-lineages, confirming its broad applicability and high ROC-AUC and PR-AUC performance. Overall, HiRisk-Detector features powerful capacity for early detection of high-risk variants, bearing great utility for any public emergency caused by infectious diseases or viruses.

1. Introduction

The COVID-19 pandemic was characterized by recurrent waves of cases, driven by the emergence of multiple high-risk SARS-CoV-2 variants with increased capacity for transmission and evasion of existing immunity by infection or vaccine.[1] Therefore, rapid and early detection of such high-risk variants is critical to guide outbreak response.

To effectively address this challenge, the global health community has developed several surveillance systems to identify and monitor these variants. Currently, there are several methods for monitoring potential high-risk variants of SARS-CoV-2, which are primarily based on early warning systems established by global public health organizations,[2,3] such as the World Health Organization (WHO), the United States Centers for Disease Control and Prevention, and the European Centers for Disease Control and Prevention. Among them, one representative method is that WHO classifies the potential high-risk variants of SARS-CoV-2 into three categories—variants of concern (VOC), variants of interest (VOI), and variants under monitoring (VUM), with high risk to low risk according to transmissibility, virulence, immune escape, etc. Although this pre-warning system is highly recognized and widely used, it requires the involvement of the WHO's Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE) to divide the risk of different variants. In other words, existing methods require extensive human knowledge and intervention and cannot fully automate the risk assessment. Critically, they have a certain degree of pre-alarm lag, e.g., approximately one month by the WHO method.

L. Li, C. Li, N. Li, D. Zou, W. Zhao, H. Luo, Y. Xue, Z. Zhang, Y. Bao, S. Song
China National Center for Bioinformation
Beijing 100101, China
E-mail: lil@big.ac.cn; ybxue@big.ac.cn; zhangzhang@big.ac.cn; baoym@big.ac.cn; songshh@big.ac.cn

L. Li, C. Li, N. Li, D. Zou, W. Zhao, H. Luo, Y. Xue, Z. Zhang, Y. Bao, S. Song
National Genomics Data Center
Beijing Institute of Genomics
Chinese Academy of Sciences
Beijing 100101, China

W. Zhao, Z. Zhang, Y. Bao, S. Song
CAS Key Laboratory of Genome Sciences and Information
Beijing Institute of Genomics
Chinese Academy of Sciences
Beijing 100101, China

W. Zhao, Y. Xue, Z. Zhang, Y. Bao, S. Song
University of Chinese Academy of Sciences
Beijing 100049, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202405058

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202405058

Advanced Science 2024

Part 4

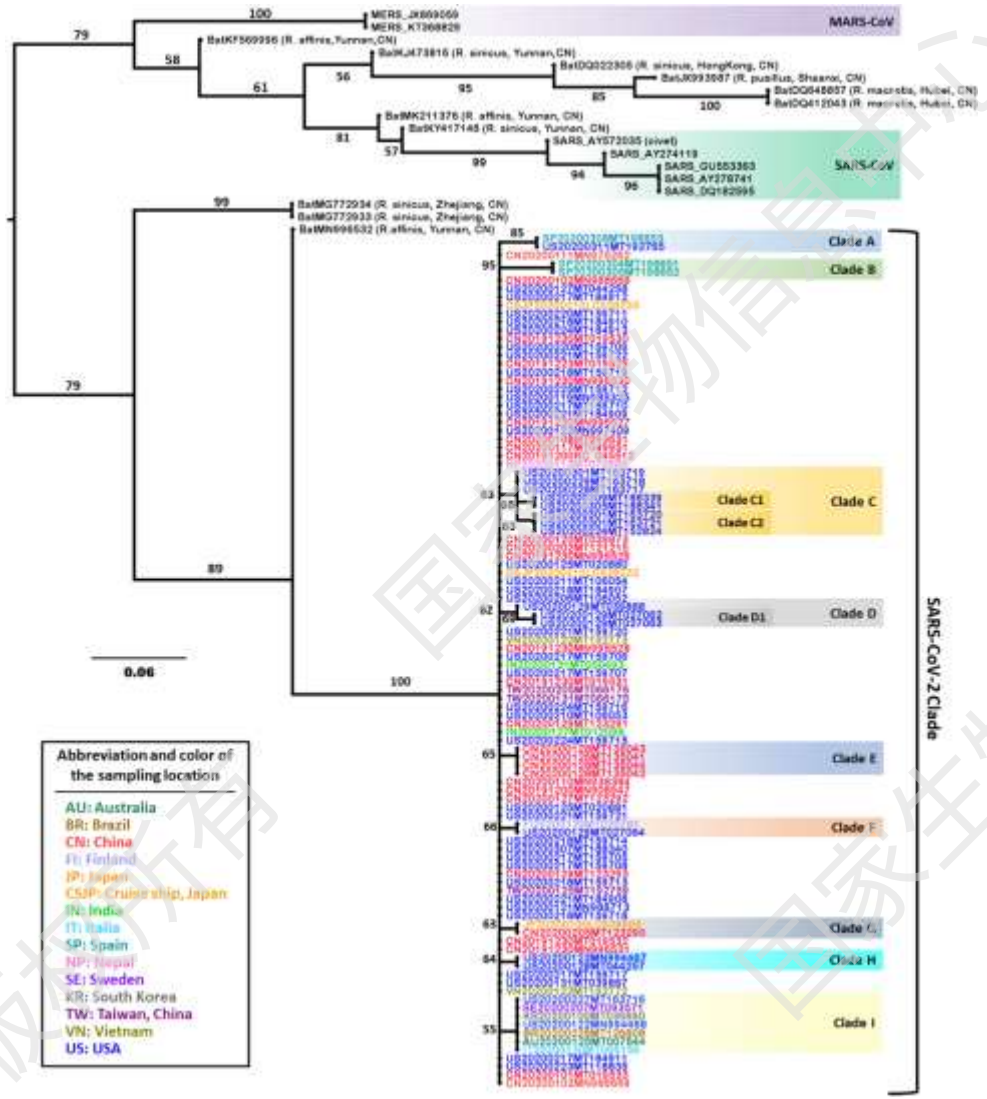
HaploNet-PathoLin: an automated Lineage Assignment and Nomenclature System for Viral Pathogens Based on Haplotype Network

国家生物信息中心版权所有

Traditional virus classification and naming system

Lefkowitz EJ, Dempsey DM, Hendrickson RC et al. *Nucleic Acids Res, 2018*

Virus lineage recognition and classification

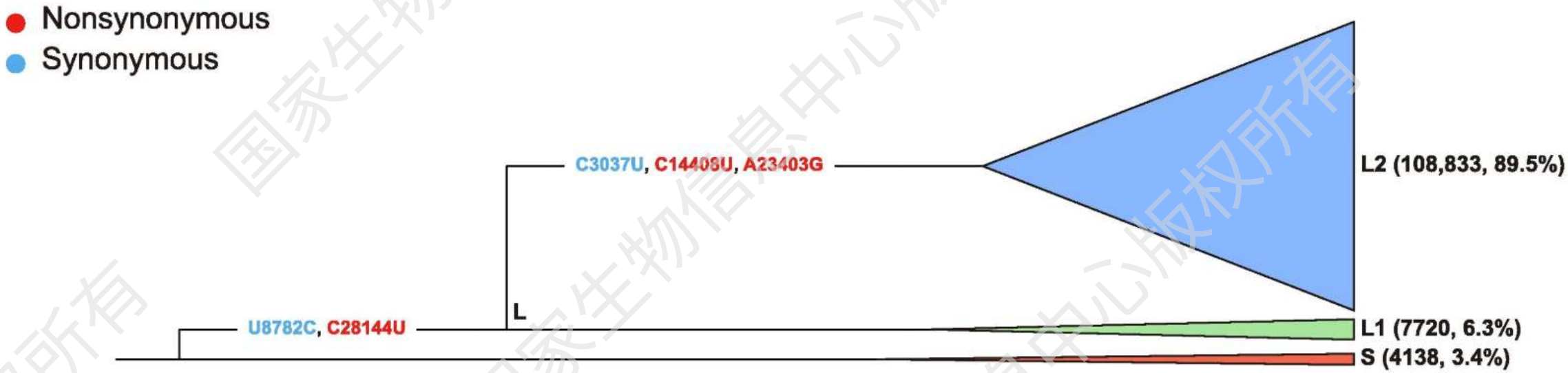


The virus undergoes complex mutations and differentiation during transmission, leading to the emergence of numerous new mutant strains. It is crucial to conduct scientifically accurate intraspecific lineage identification and classification for these constantly evolving and mutating viruses.

Li, T., Liu, D., Yang, Y. *et al. Sci Rep, 2020*

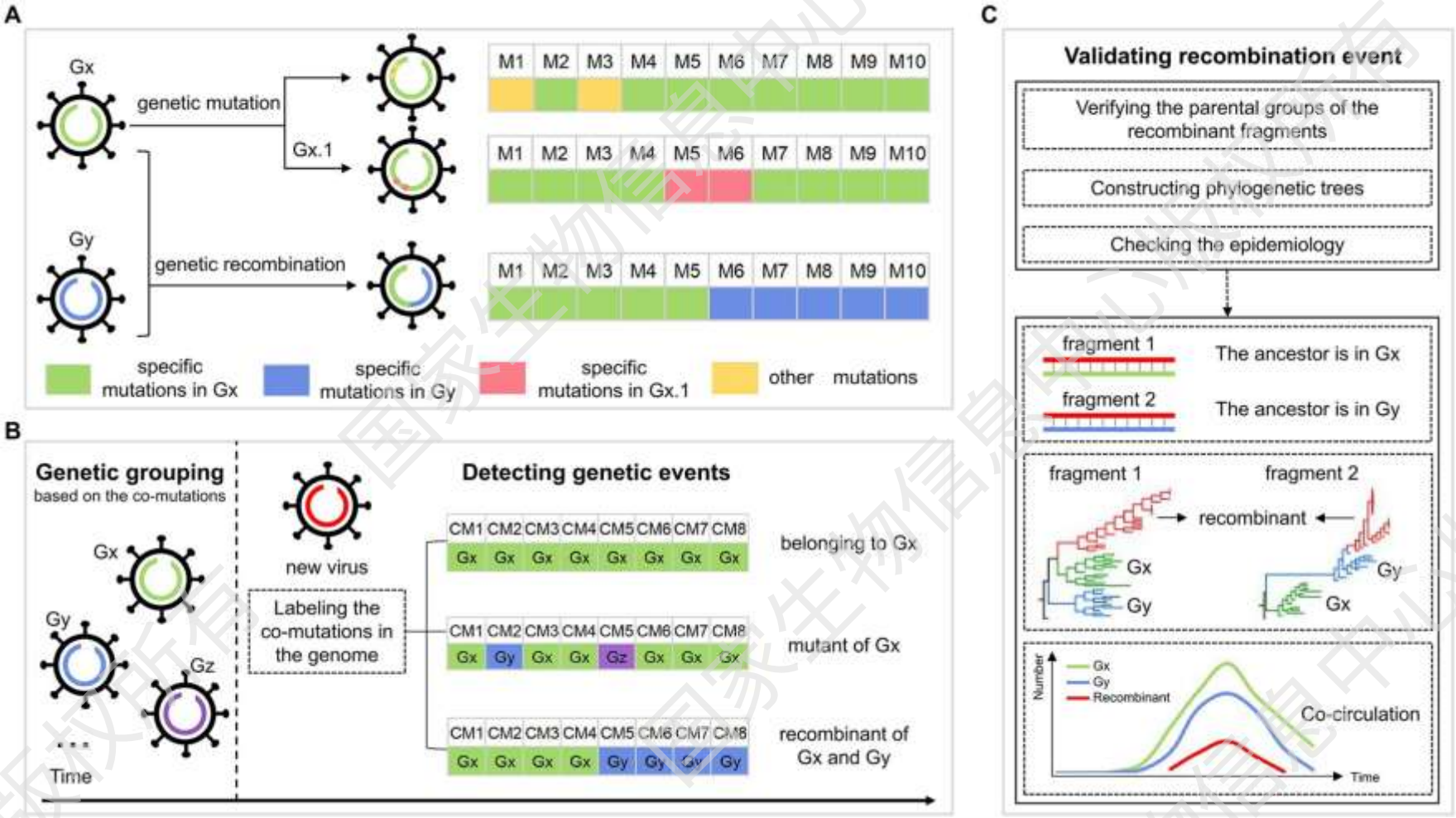
Research progress on SARS-CoV-2 lineages

Peking University Lu Jian team: Based on early virus sequence single nucleotide variations (SNVs), identified three main lineages and 130 sub lineages, and named them using a combination of letters and numbers.



Xiaolu Tang, Ruochen Ying, Xinmin Yao et al. *Science Bulletin, 2021*

Research progress on SARS-CoV-2 lineages



CMM grouping (Jiang Taijiao team of the Chinese Academy of Medical Sciences): based on the common mutation site, the early COVID-19 is divided into 43 pedigrees.

Qin Luyao , Meng Jing , Ding Xiao et al. *Frontiers in Microbiology, 2022*

Research progress on SARS-CoV-2 lineages

Pangolin system (Oliver team, Oxford University): a binary evolutionary tree based on viral sequences, widely used for SARS-CoV-2 lineage partitioning

WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern

31 May 2021 | Departmental update |Reading time: Less than a minute (198 words)

WHO has assigned simple, easy to say and remember labels for key variants of SARS-CoV-2, the virus that causes COVID-19, using letters of the Greek alphabet.

These labels were chosen after wide consultation and a review of many potential naming systems. WHO convened an expert group of partners from around the world to do so, including experts who are part of existing naming systems, nomenclature and virus taxonomic experts, researchers and national authorities.

WHO will assign labels for those variants that are designated as Variants of Interest or Variants of Concern by WHO. These will be posted on the WHO website.

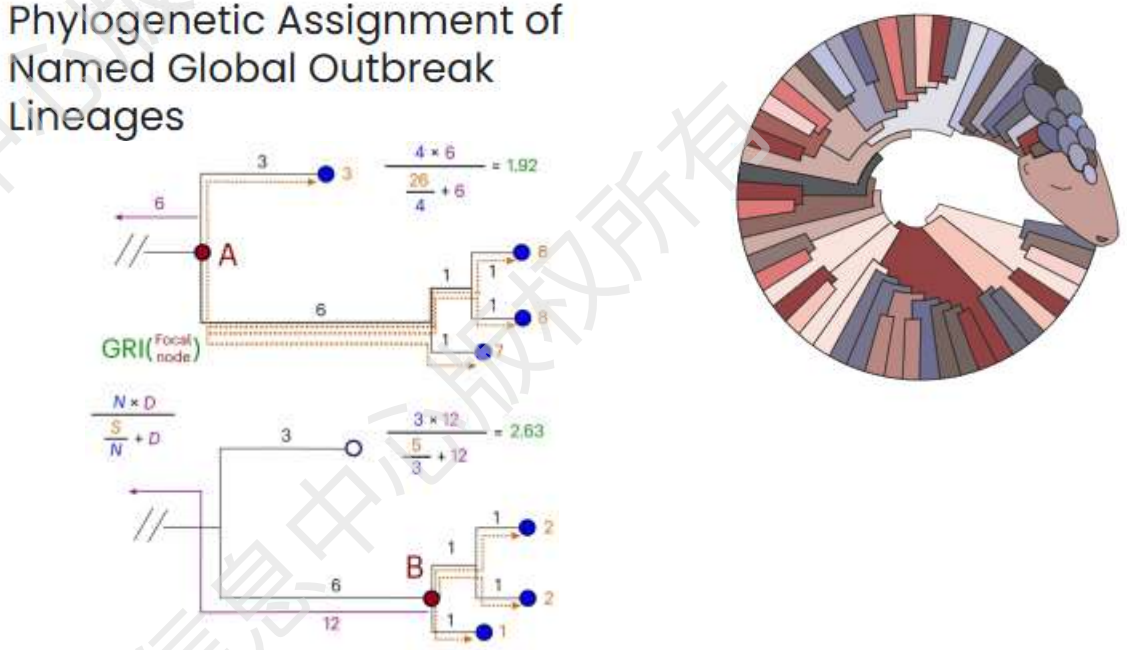
These labels do not replace existing scientific names (e.g. those assigned by GISAID, Nextstrain and Pango), which convey important scientific information and will continue to be used in research.

While they have their advantages, these scientific names can be difficult to say and recall, and are prone to misreporting. As a result, people often resort to calling variants by the places where they are detected, which is stigmatizing and discriminatory. To avoid this and to simplify public communications, WHO encourages national authorities, media outlets and others to adopt these new labels.

See the new labels here.

pangolin

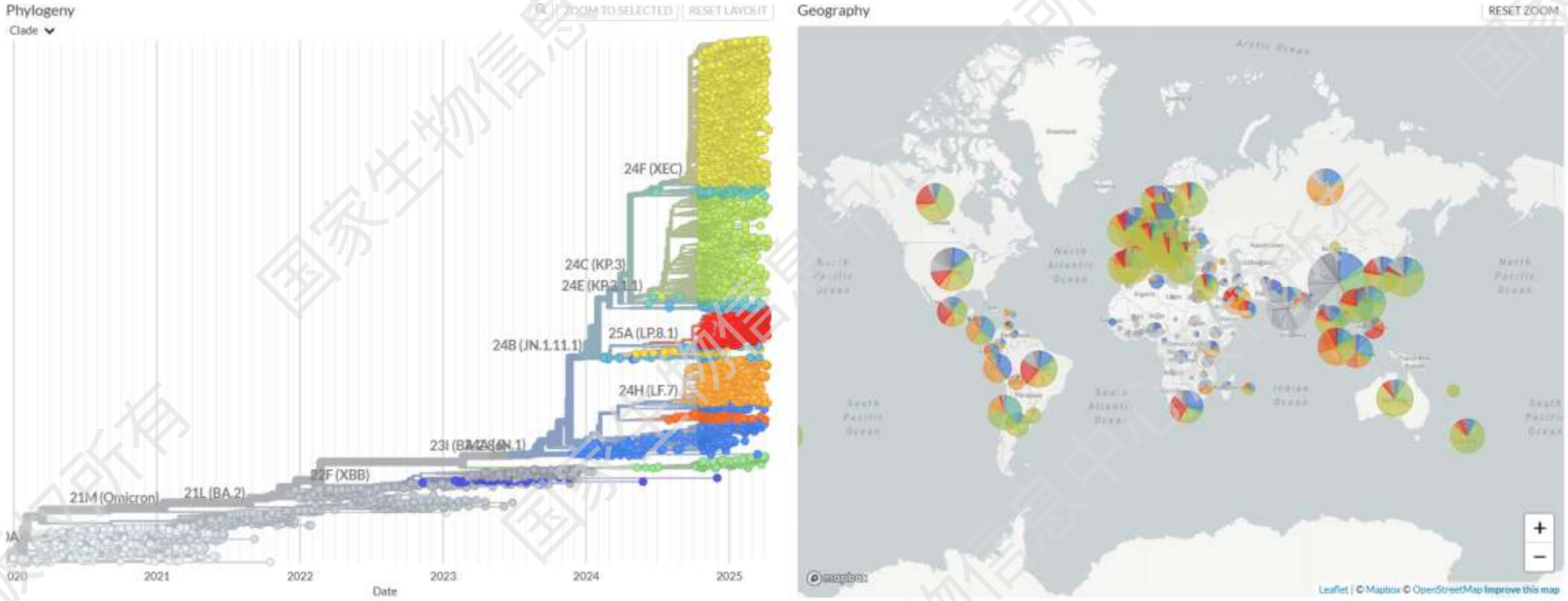
Phylogenetic Assignment of Named Global Outbreak Lineages



Áine O'Toole, Emily Scher, Anthony Underwood et al. *Virus Evolution, 2021*
McBroome, J., de Bernardi Schneider, A., Roemer, C. *et al. Nat Microbiol, 2024*

Research progress on SARS-CoV-2 lineages

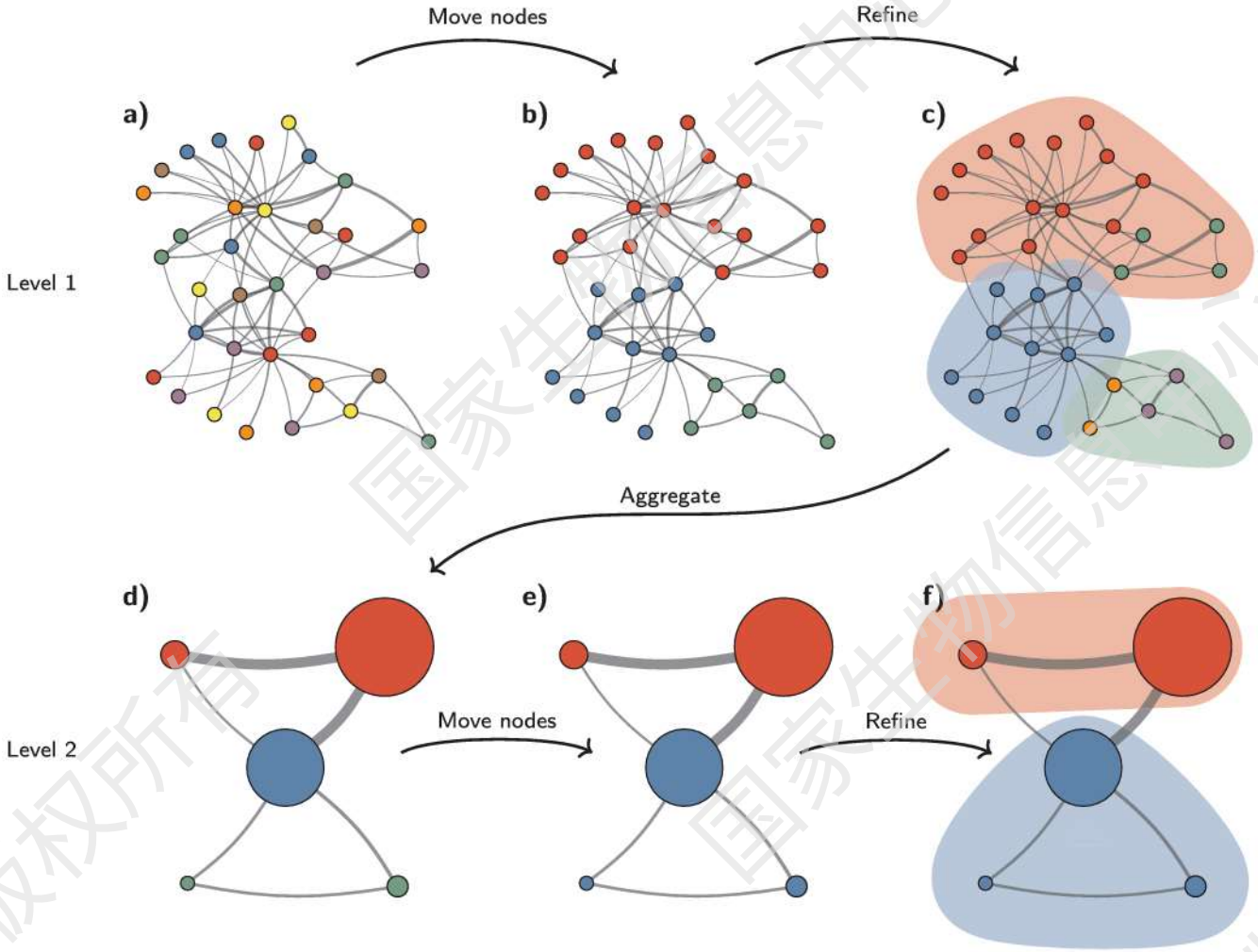
Nextstrain provides global and regional virus evolution dynamics through real-time analysis of GISAID data, supporting lineage classification and visualization.



https://nextstrain.org/sars-cov-2

国家生物信息中心版权所有

HaploNet-PathoLin



Coarse grained clustering based on monolithic of haplotype networks

$$Q = \frac{1}{2m}\sum_{uv}\left[A_{uv} - \frac{k_u k_v}{2m}\right] \cdot \delta(c_u, c_v)$$

$$\delta(c_u, c_v) = \begin{cases} 1, & if\ c_u = c_v \\ 0, & otherwise \end{cases}$$

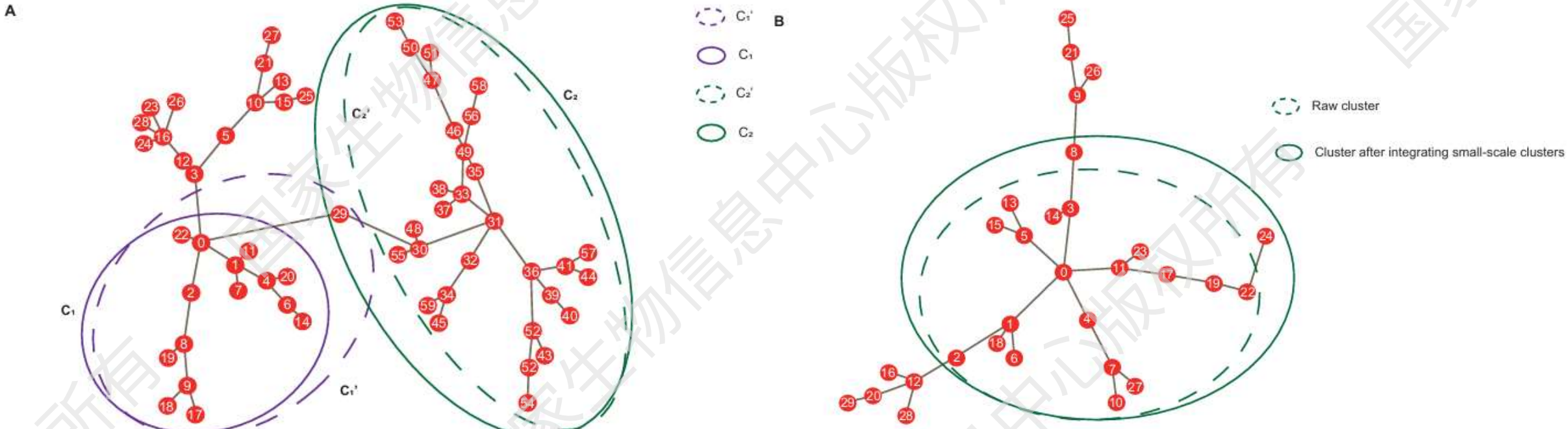
unpublish

国家生物信息中心版权所有

HaploNet-PathoLin

Fine-grained partitioning by breakpoint reassignment

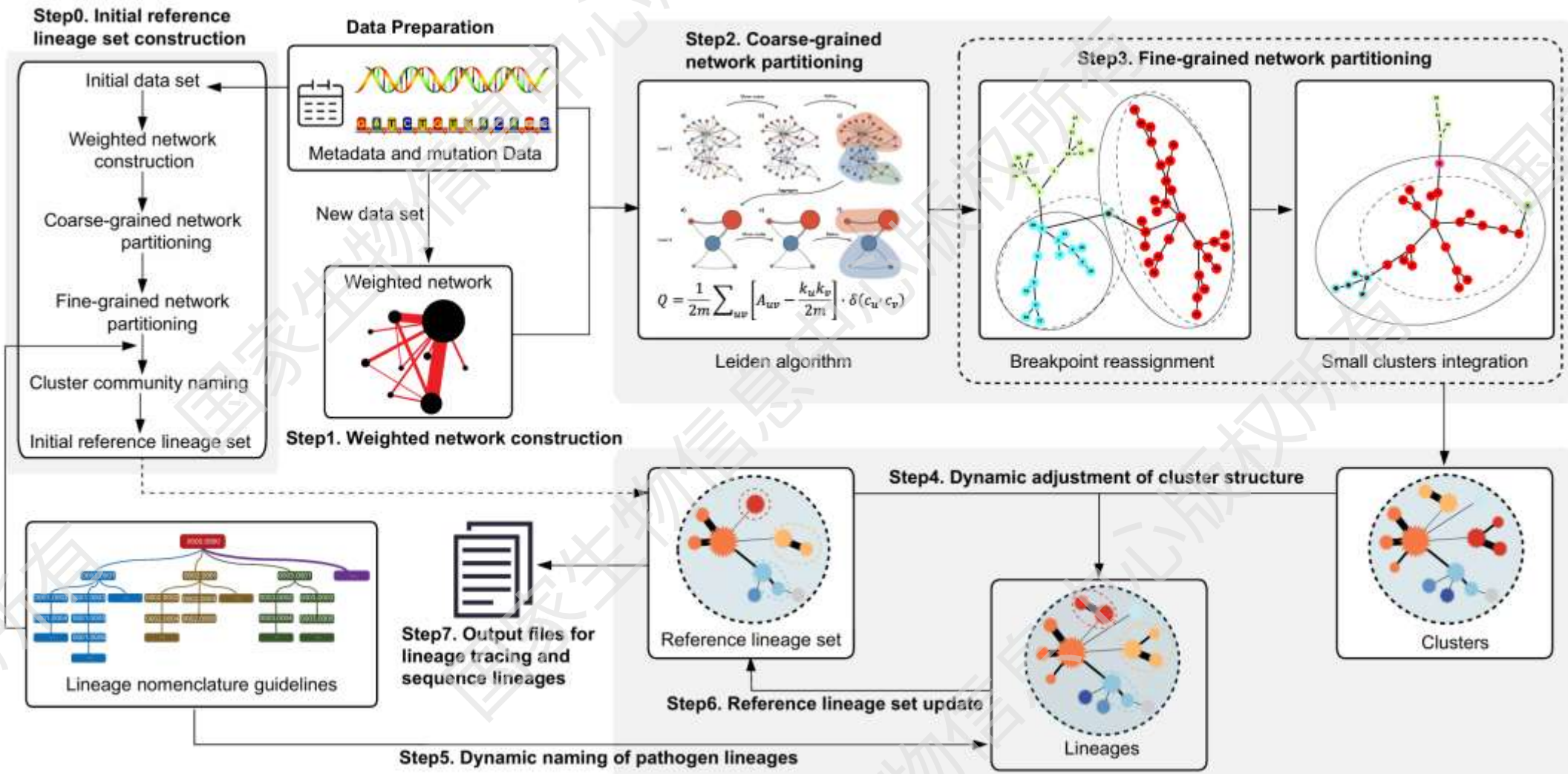
Fine-grained partitioning by small cluster integration.



Coarse grained clustering based on monolithic networks

unpublish

HaploNet-PathoLin



unpublish

HaploNet-PathoLin

HaploNet-PathoLin谱系划分结果评估

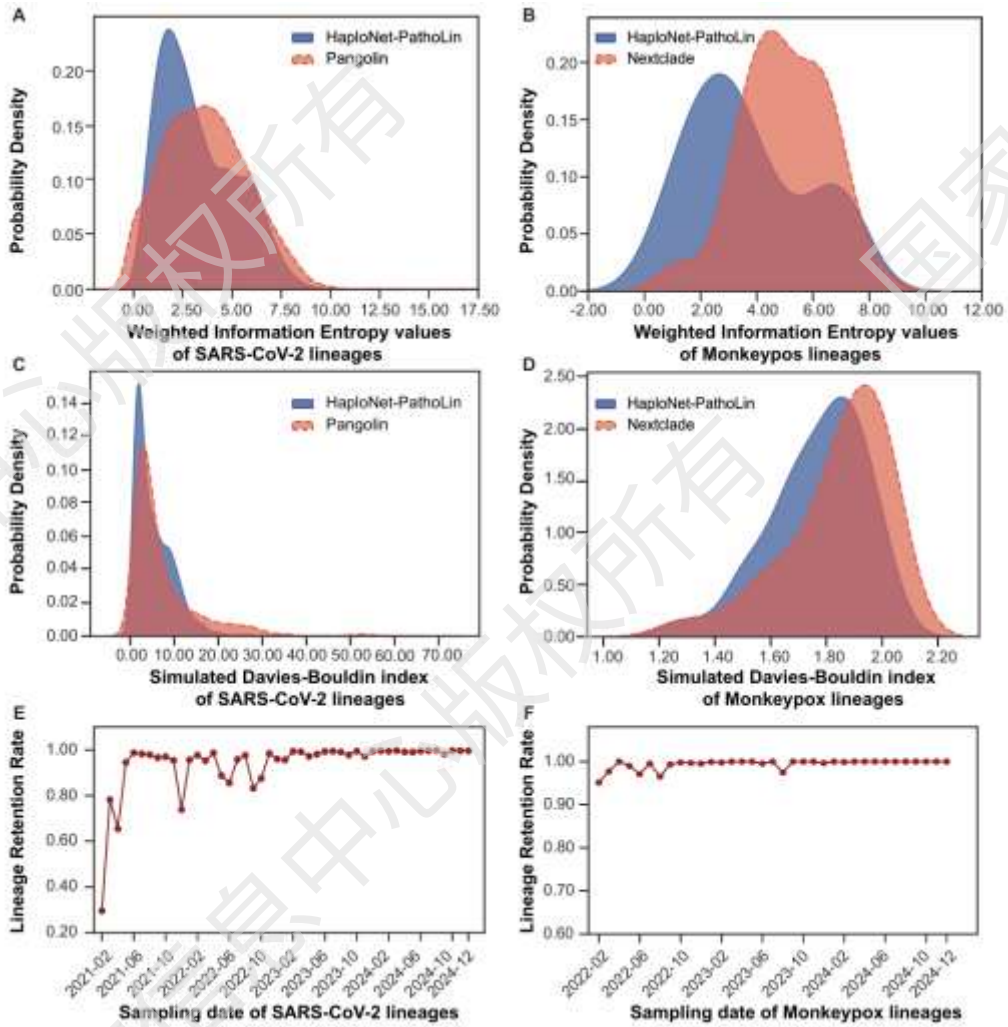
与Pangolin的比较:

$$E = -\sum_i \left(p_i \log_2 p_i * (1 + u_i)\right)$$

$$SDBI_{ij} = \frac{d_i + d_j}{d_{ij}}$$

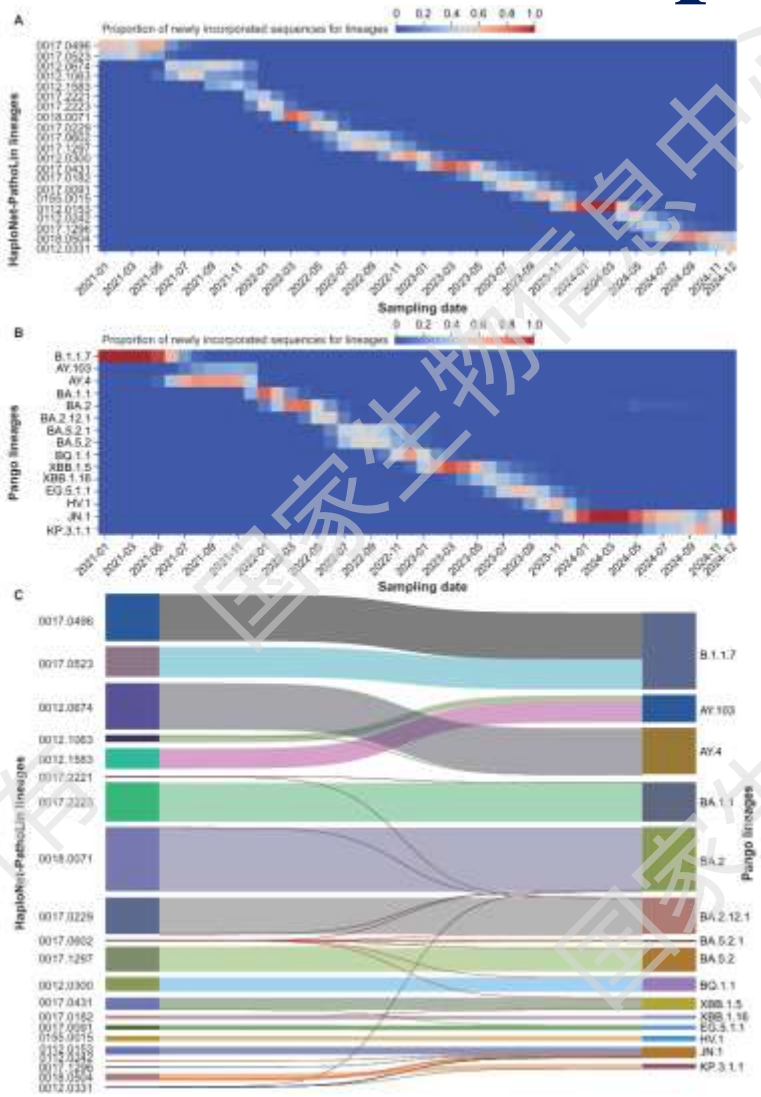
Dynamic partition stability assessment :

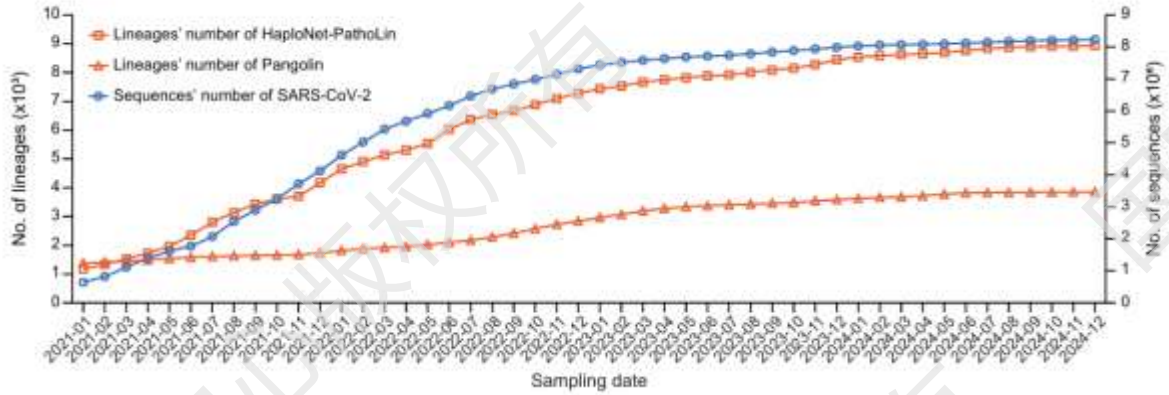
$$r_t = \frac{n_{t-c,t}}{|\cap(S_{t-c}, S_t)|}$$



unpublish

HaploNet-PathoLin





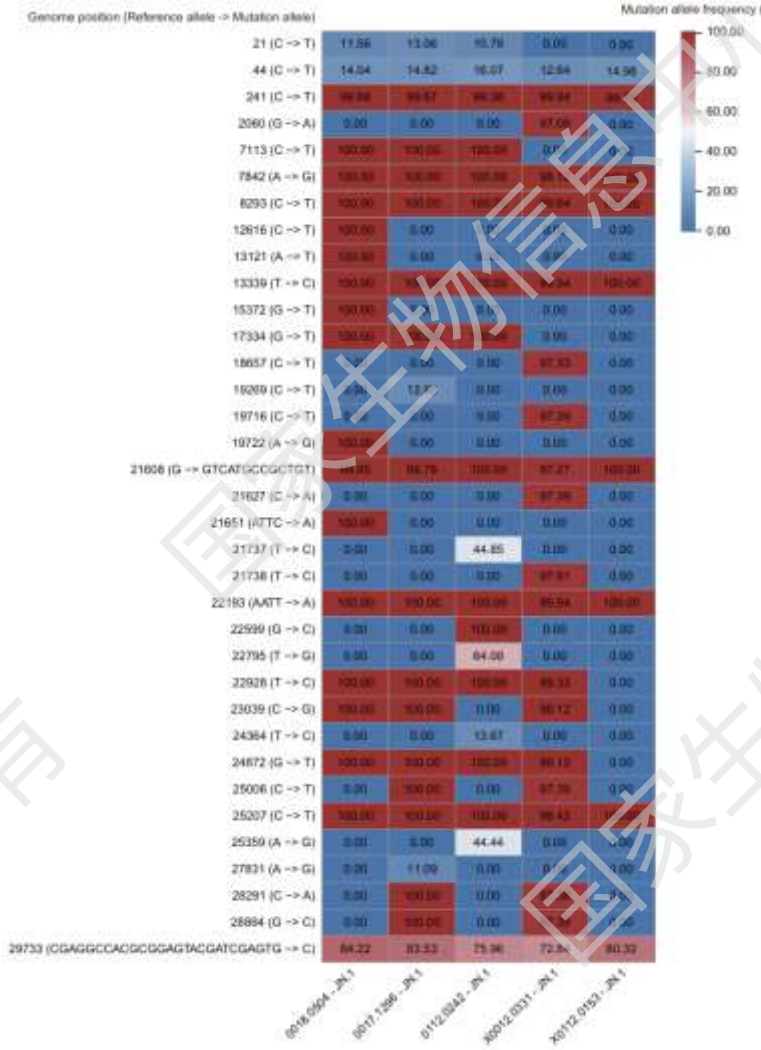
- HaploNet PathoLin successfully resolved the hidden evolutionary branches within the "single" transmission lineage (such as JN.1) in traditional classification through high-resolution lineage subdivision, revealing the "differential divergence" in virus adaptive evolution that is difficult to detect by traditional methods.
- These sub lineages exhibit a typical 4-6 month transmission cycle, which is highly consistent with the real epidemiological dynamics.

unpublish

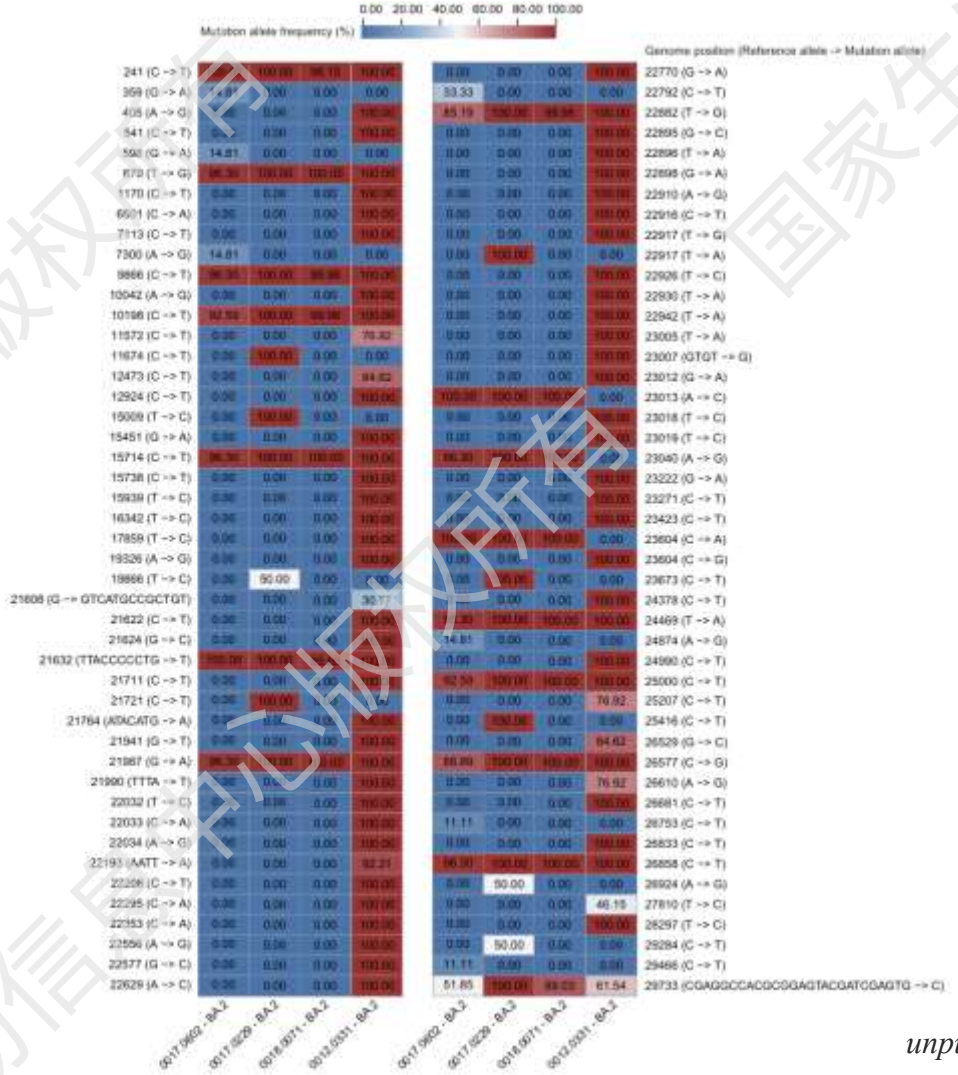
国家生物信息中心版权所有

HaploNet-PathoLin

Pangolin JN.1 vs HaploNet-PathoLin lineages



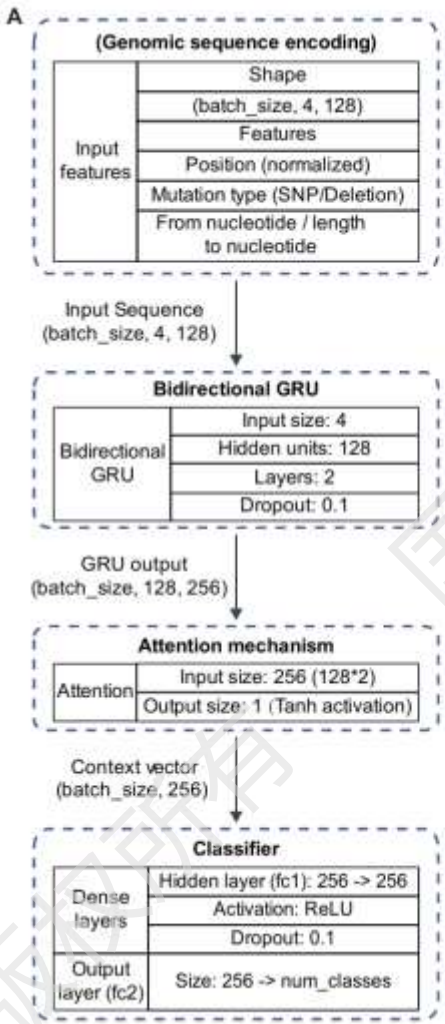
Pangolin BA.2 vs HaploNet-PathoLin lineages

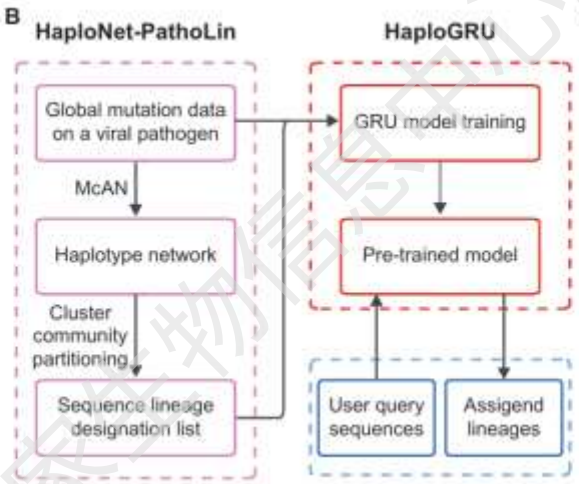


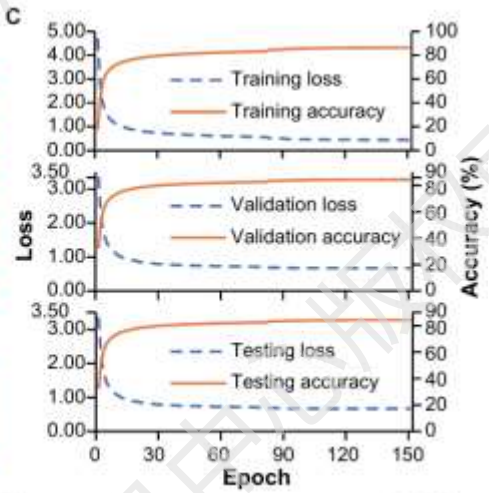
unpublish

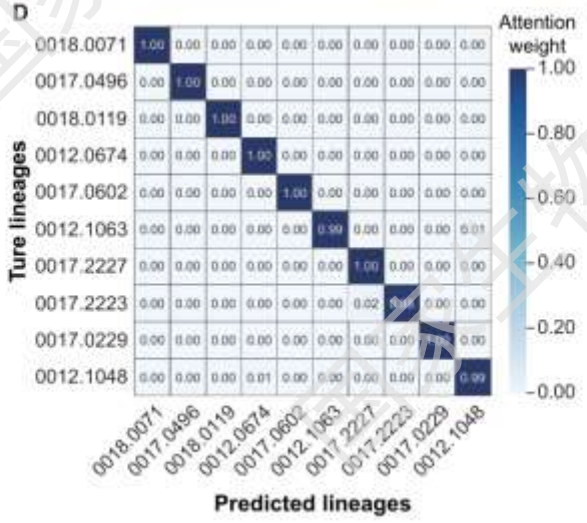
国家生物信息中心版权所有

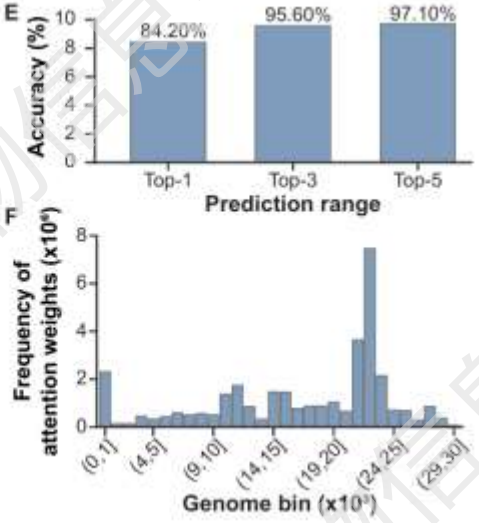
HaploGRU

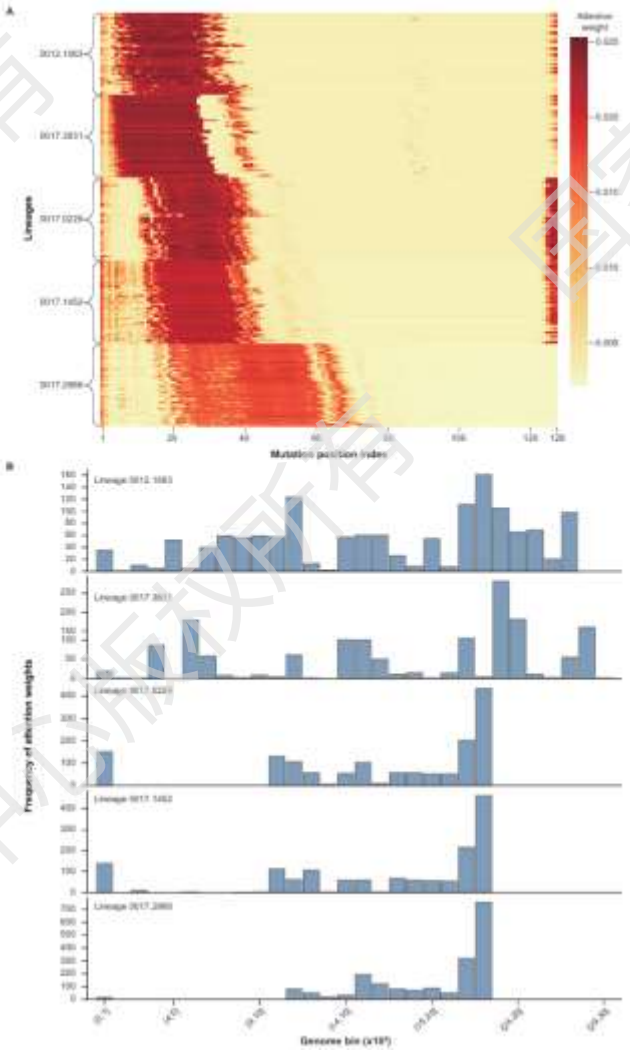

A
(Genomic sequence encoding)
Input features
Shape
(batch_size, 4, 128)
Features
Position (normalized)
Mutation type (SNP/Deletion)
From nucleotide / length to nucleotide
Input Sequence (batch_size, 4, 128)
Bidirectional GRU
Bidirectional GRU
Input size: 4
Hidden units: 128
Layers: 2
Dropout: 0.1
GRU output (batch_size, 128, 256)
Attention mechanism
Attention
Input size: 256 (128*2)
Output size: 1 (Tanh activation)
Context vector (batch_size, 256)
Classifier
Dense layers
Hidden layer (fc1): 256 -> 256
Activation: ReLU
Dropout: 0.1
Output layer (fc2)
Size: 256 -> num_classes


B
HaploNet-PathoLin
HaploGRU
Global mutation data on a viral pathogen
McAN
Haplotype network
Cluster community partitioning
Sequence lineage designation list
GRU model training
Pre-trained model
User query sequences
Assigend lineages


C
Training loss
Training accuracy
Validation loss
Validation accuracy
Testing loss
Testing accuracy
Loss
Accuracy (%)
Epoch


D
Ture lineages
Predicted lineages
Attention weight
0018.0071
0017.0496
0018.0119
0012.0674
0017.0602
0012.1063
0017.2227
0017.2223
0017.0229
0012.1048


E
Accuracy (%)
84.20%
95.60%
97.10%
Top-1
Top-3
Top-5
Prediction range
F
Frequency of attention weights (x10⁶)
Genome bin (x10⁵)



Part 5

International Collaboration

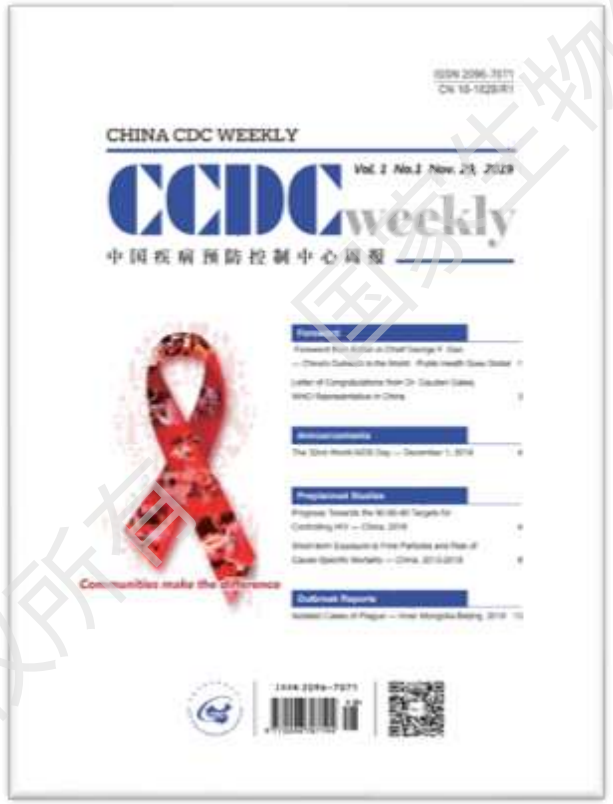
国家生物信息中心版权所有

CHINA CDC WEEKLY

CCDC weekly

中国疾病预防控制中心周报

ISSN 2096-7071 (Print)
ISSN 2096-3101 (Online)
CN 10-1629/R1



IF 2022
4.7
SCIE 58/207 Q2
SSCI 37/180 Q1

Editor-in-Chief

Hongbing Shen

General Director, China CDC

Academician, Chinese Academy of Engineering

Characteristic

China National Public Health Bulletin

Missions

Global data sharing

Building national public health leadership

Editorial Office

Telephone：010-63150701

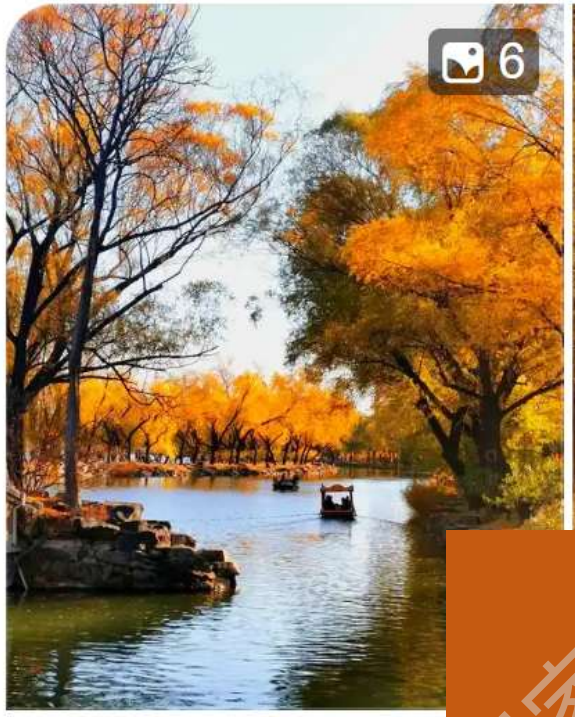
Email：weekly@chinacdc.cn

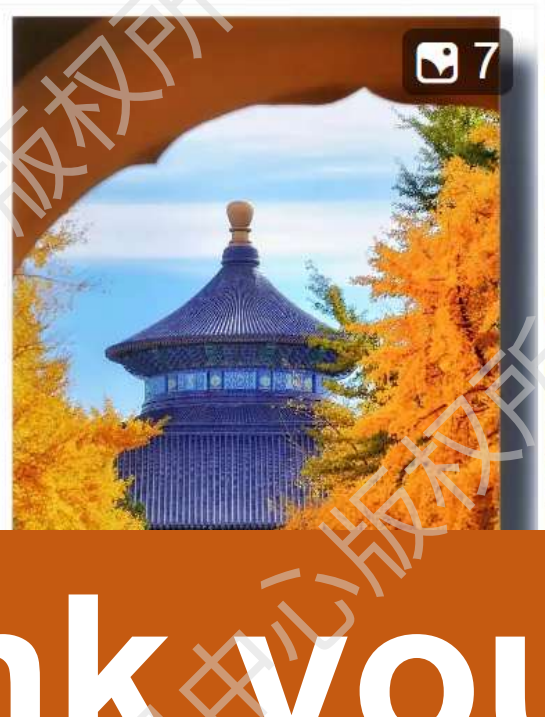
国家生物信息中心版权所有

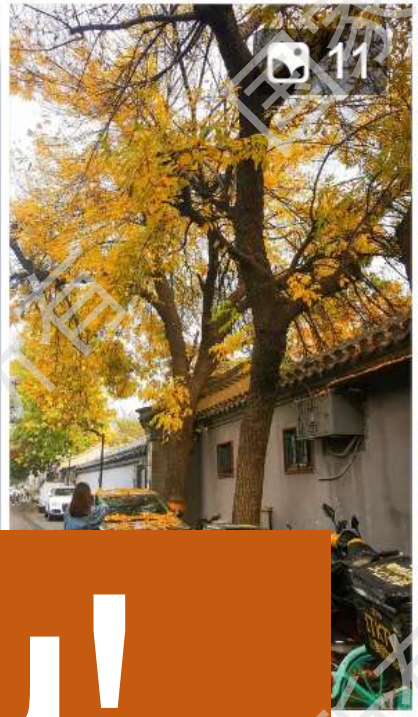
Collaboration Interests

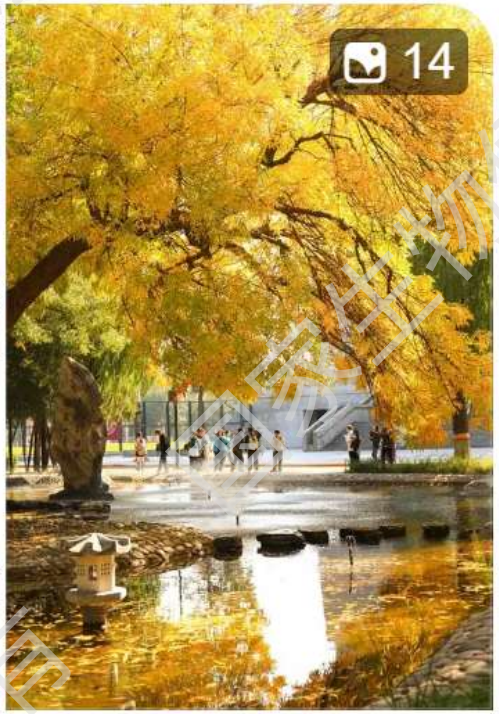
- Promote **exchanges and visits of academic staff and students**, sharing of non-confidential publications and materials, as well as scientific research cooperation
- Enhance cooperation in building bioinformatics and life **big data sharing networks**
- Explore opportunities for collaborating on **funded research programs**
- Discuss **cooperation in the frontier** and interdisciplinary field of life sciences (e.g. AI for health)

国家生物信息中心版权所有

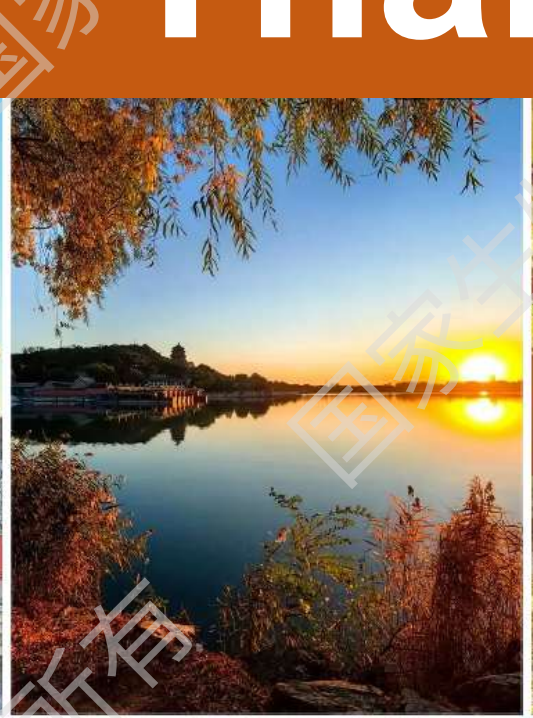
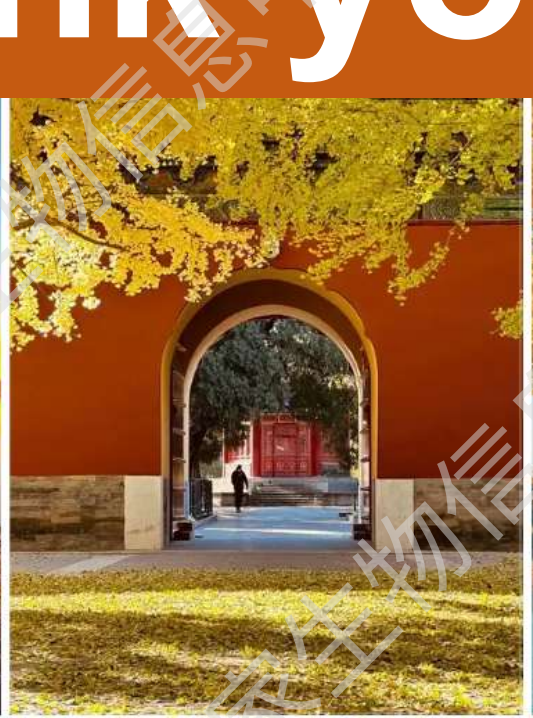
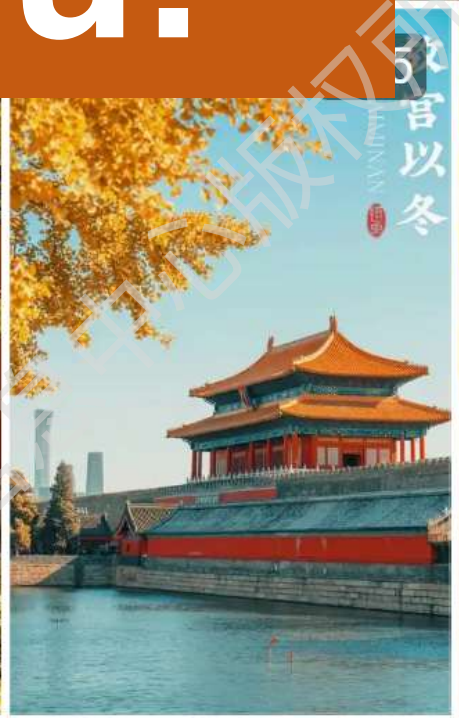








Thank you!


国家生物信息中心版权所有