

国家生物信息中心系列培训

全基因组测序在病原溯源中的经典案例分享

报告人：李论

2025 年 6 月 26 日

提纲

- 01 | 背景介绍
- 02 | 基于系统发育树的病原溯源
- 03 | 基于单体型网络的病原溯源
- 04 | 基于序列相似性的病原溯源
- 05 | 实操练习

提纲

- 01 | 背景介绍
- 02 | 基于系统发育树的病原溯源
- 03 | 基于单体型网络的病原溯源
- 04 | 基于序列相似性的病原溯源
- 05 | 实操练习

病原微生物威胁国家生物安全

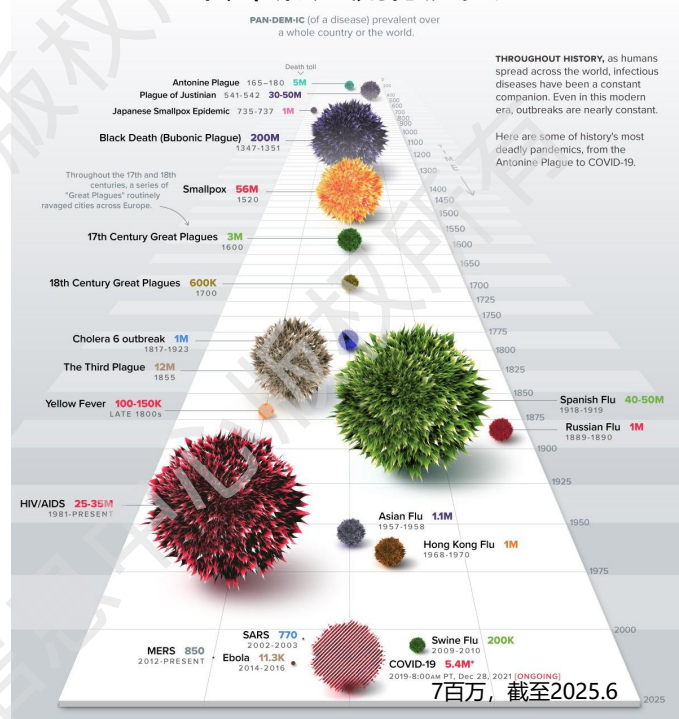
- 病原微生物反复威胁人类健康，夺去了数十亿的生命
- 新冠病毒的大流行不仅带来了前所未有的公共卫生危机，也让病原体的来源与传播路径成为全球关注的焦点，推动溯源研究走向前台

国务院新闻办公室4月30日发布《关于新冠疫情防控与病毒溯源的中方行动和立场》白皮书

白皮书说，《世卫组织召集的SARS-CoV-2全球溯源研究：中国部分——世卫组织—中国联合研究报告》等研究通过系统开展流行病学、分子溯源、动物宿主排查与冷链输入研究，排除了武汉作为新冠病毒自然起源地的可能性，认为“武汉实验室泄漏病毒”途径极不可能，并为全球科学界提供了关键的实证数据与研究范式。

http://www.scio.gov.cn/zd gz/gzdt1/202504/t20250430_893961.html

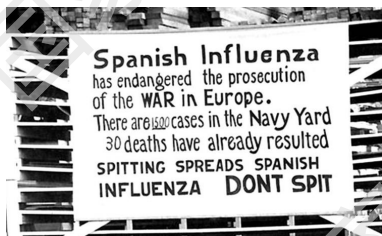
传染病大流行历史



病原微生物溯源

- **定义：**病原体溯源（Pathogen tracing or source tracking）是指通过基因组学、流行病学等技术手段，追踪病原体的来源、传播路径和演化过程
- **意义：**有助于阻断传播链，降低疫情风险，提升传染病防控能力；以客观数据击碎谣言，避免污名化

1918大流感 vs “西班牙大流感”



位于费城海军飞机工厂外的一个标识

- **历史背景：**第一次世界大战（1914.7.28 至 1918.11.11）
- **战时新闻管制：**交战国（英、法、德等）为维持士气，严格封锁疫情消息
- **污名化：**西班牙作为中立国，未实施新闻审查，国内疫情被广泛报道，国际社会误以为西班牙是源头

以史為鑒



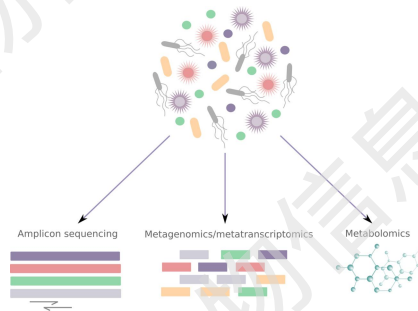
用于病原微生物溯源的测序技术

全基因组测序



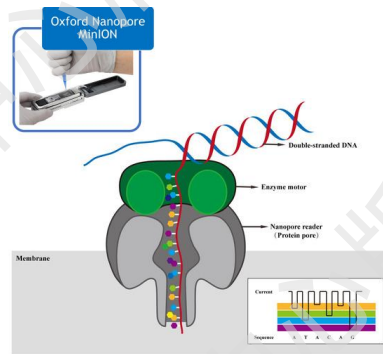
- 高分辨率、完整基因组
- 需分离培养

宏基因组测序



- 无需培养、广谱性
- 分辨率较低

纳米孔测序



- 现场测序、设备便携
- 单碱基错误率偏高

靶向基因测序

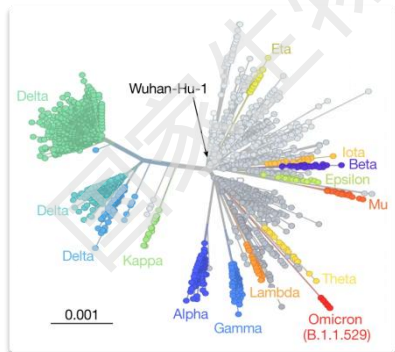


- 快速识别已知类群
- 对未知病原体无效

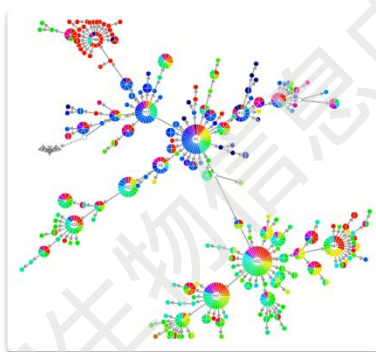
基于全基因组测序的病原溯源方法

- 分为两大类：基于演化的病原溯源和基于序列相似性的病原溯源

基于演化的病原溯源



系统发育树



单体型网络

基于序列相似性的病原溯源

Position	1	2	3	4	5	6	
Query	○	○	—	—	—	—	Score
Hit 1	○	○	○	—	—	—	0.833
Hit 2	○	○	—	○	—	—	0.833
Hit 3	○	○	—	—	○	○	0.750

SNP相似性分析

提纲

01 | 背景介绍

02 | **基于系统发育树的病原溯源**

03 | 基于单体型网络的病原溯源

04 | 基于序列相似性的病原溯源

05 | 实操练习



钻石公主号——“新冠肺炎制造厂”

钻石公主号，2020年1月20日出发，途经6个港口，共有3711人登船，712人被感染新冠。神户大学医学系教授岩田健太郎称这艘船为“新冠肺炎制造厂”。



2020年2月7日，在抵达日本东京南部横滨的大国码头游轮码头后，穿着防护装备的工作人员进入钻石公主号游轮，将一名病人转移到医院。

单次引入 vs 多次引入？

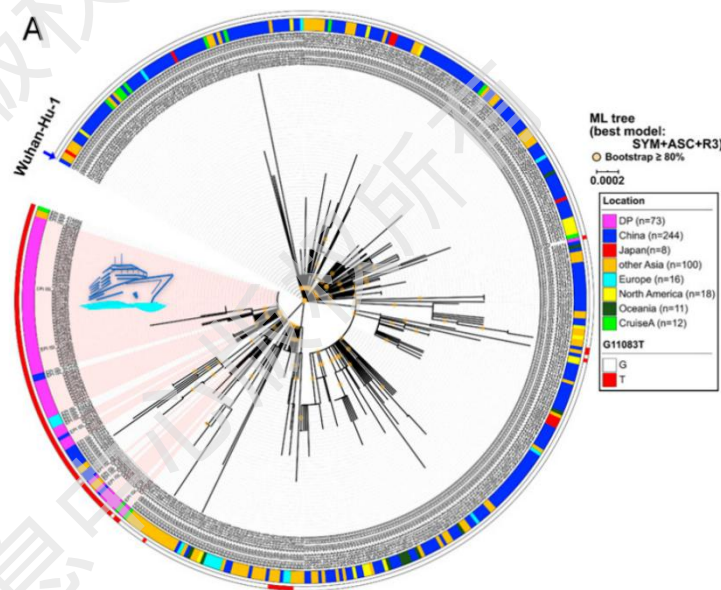


钻石公主号游轮的行程



系统发育树在溯源中的作用

- 数据
 - 73条序列，来自钻石公主号
 - 412条序列，GISAID，2020.5.14
- 方法
 - 最大似然树 (Maximum Likelihood Tree)
- 结果
 - 钻石公主号的新冠序列聚在一簇
 - 都带有G11083T
- 结论
 - 单次引入

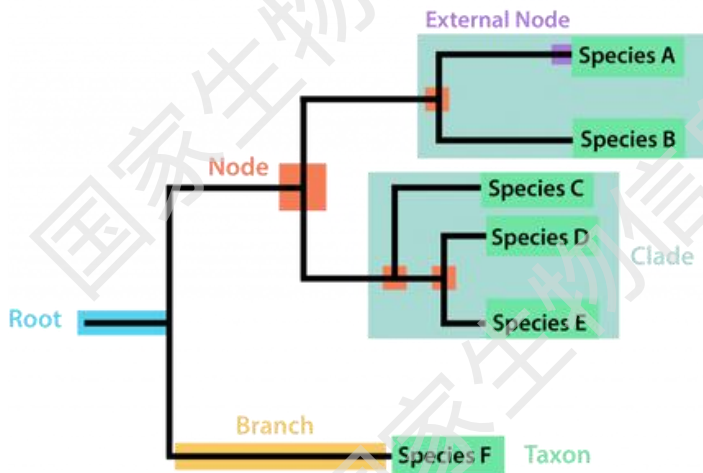


Tsuyoshi Sekizuka, PNAS, 2020



系统发育树背景知识

- **系统发育树** (Phylogenetic tree)，也称为进化树或谱系树，是一种用来表示不同物种或同一物种不同个体之间进化关系的树状图。它通过分支结构展示了物种或基因序列的共同祖先以及它们随时间进化的路径。



<https://jaquishgen564s19.weebly.com/phylogenetics.html>

- **根**：所有序列的共同祖先
- **叶子节点**：样本
- **内部节点**：代表共同祖先
- **进化支**：由一个共同祖先及其所有后代组成
- **外群**：在系统发育树中，与研究对象亲缘关系最远的物种，用作比较和参考



系统发育树构建步骤



系统发育树构建准备

• 数据准备

数据来源：

- 测序平台：获取原始序列数据
- 公共数据库：NCBI、GISAID、CNCB等

序列质量与特性：

- 确保序列质量，避免噪声或错误
- 能代表目标物种的遗传特征
- 基因长度和覆盖范围

• 多序列比对

序列比对示例：

序列 1: A T G C A T G C A
序列 2: A T - C A T G C A
序列 3: A T G C - T G C A

注：相同颜色表示相同核苷酸，“-”表示空位(gap)

常用多序列比对软件：

软件	特点
Clustal系列	老牌软件，操作简单。Clustalw 速度较慢但结果可靠；Clustal omega适用于大数据集
MAFFT	针对大数据集（超千条，万条以上序列）设计，速度快，功能强大，能快速处理复杂任务
Muscle	平衡了速度和准确性，在速度和精度上优于ClustalW



系统发育树构建算法

可分为三大类，每种方法都对进化过程做出了特定的假设。

类别	方法	原理简介	特点
距离法	邻接法 (NJ)	基于距离矩阵，聚类合并距离最近的分支	快速，适合大规模数据
	UPGMA	假设分子钟，按照相同速率聚类	适合分子钟假设的数据
最大似然 (ML)		基于进化模型，寻找最大可能性的树结构	结果可靠，耗时较长
最大简约 (MP)		寻找最少进化步骤的树	对序列变化假设最少化



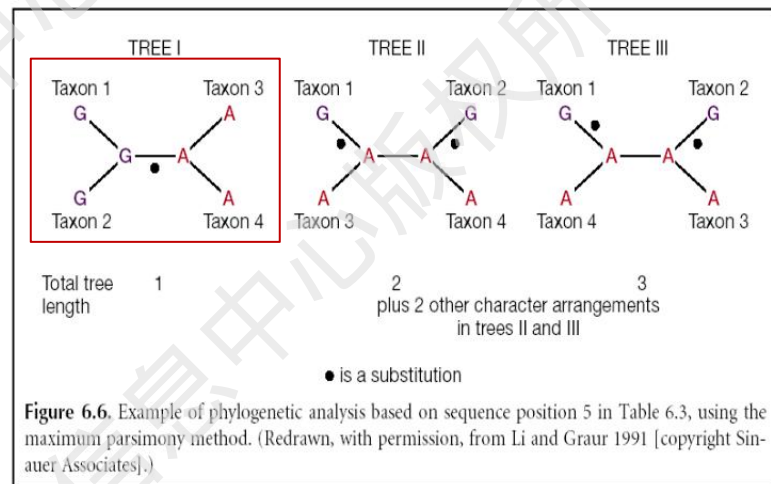
系统发育树构建算法

• 最大简约法 (Maximum Parsimony)

最大简约法基于奥卡姆哲学原则 (Ockham's Razor) 即解释一个过程的最好理论是所需假设数目最少的那一个。

- 寻找需要**最少进化事件** (如核苷酸替换或氨基酸替换) 来解释序列数据的系统发育树
- 评估所有可能的树拓扑结构, 计算每种拓扑下进化改变的总数
- 选择改变总数**最小**的树作为最优树

Taxa	Sequence position (sites) and character								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

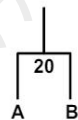


系统发育树构建算法

- 距离法-非加权组平均法
(UPGMA, Unweighted Pair-Group Method with Arithmetic Mean)

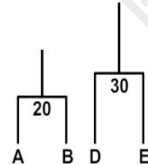
- 假设所有谱系以恒定的速率进化 (符合分子钟假说)
- 迭代地将距离最近的两个单元合并, 并计算新聚类与其余聚类之间的平均距离

	A	B	C	D	E
A	0				
B	20	0			
C	60	50	0		
D	100	90	40	0	
E	90	80	50	30	0



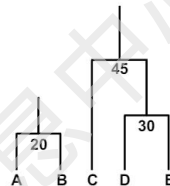
- New average distance between C and AB is:
 $C \text{ to } AB = (60 + 50) / 2 = 55$
- Distance between D and AB is:
 $D \text{ to } AB = (100 + 90) / 2 = 95$
- Distance between E and AB is:
 $E \text{ to } AB = (90 + 80) / 2 = 85$

	AB	C	D	E
AB	0			
C	55	0		
D	95	40	0	
E	85	50	30	0



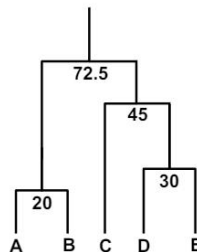
- New average distance between DE and AB is:
 $DE \text{ to } AB = (95 + 85) / 2 = 90$
- Distance between DE and C is:
 $DE \text{ to } C = (40 + 50) / 2 = 45$

	AB	C	DE
AB	0		
C	55	0	
DE	90	45	0



New Average distance between CDE and AB is:
 $CDE \text{ to } AB = (90 + 55) / 2 = 72.5$

	AB	CDE
AB	0	
CDE	72.5	0



Mission accomplished!

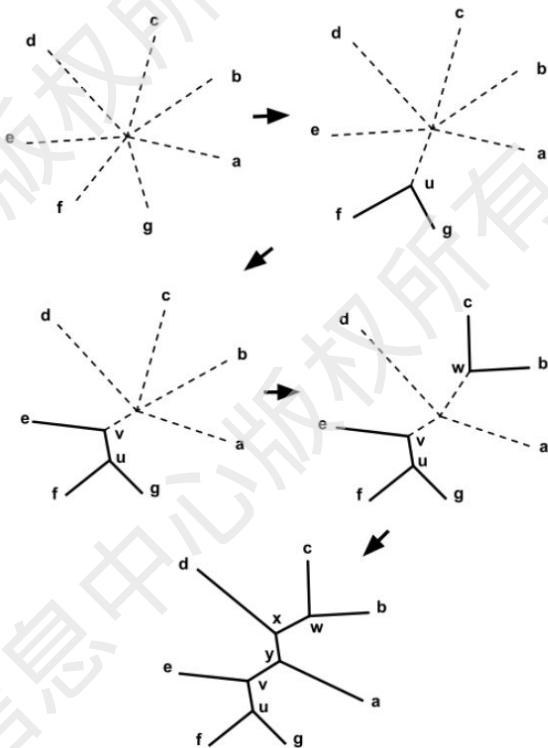


系统发育树构建算法

- 距离法-邻接法

(NJ, Neighbor-Joining)

- 通过寻找使树的**总枝长最小化**的方式来构建树，不要求进化速率恒定
- 从星状树开始，迭代地将距离最近的“邻居”配对，并定义连接它们的新内部节点，逐步构建树的拓扑结构。



系统发育树构建算法

• 最大似然法

(ML, Maximum Likelihood)

- 最大似然法是一种基于统计学的系统发育树构建方法，评估在给定的进化模型和树拓扑结构下，观察到当前序列数据的概率（即似然值）。

$$L(T, \theta) = P(D|T, \theta)$$

- ML法的目标是找到一个树拓扑结构和一组模型参数(包括枝长和替换速率等)，使得观察到的序列数据出现的概率最大化。
- 是目前最佳的建树方法之一。

步骤：

- 选择一个进化模型
- 对于每个位置，生成所有可能的树结构
- 在进化模型的基础上，计算这些树的似然度，并将它们相加得到列似然度。
- 通过乘以每个位置的可能性来计算树的可能性
- 选择可能性最大的树



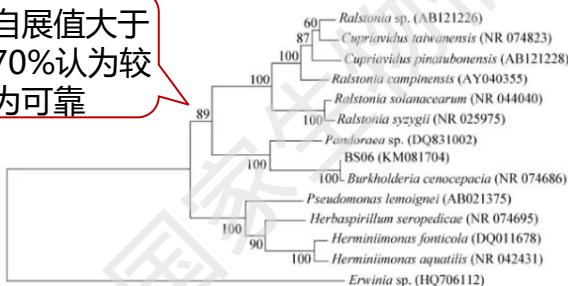
系统发育树评估与美化

• 树的评估：自展法

(Bootstrap method)

- 从排列的多序列中随机有放回的抽取某几列，构成相同长度的新的排列序列
- 重复上面的过程，得到多组新的序列
- 对这些新的序列进行建树，再观察这些树与原始树是否有差异，以此评价建树的可靠性

自展值大于
70%认为较
为可靠



• 树的美化

MEGA

内置树查看和编辑功能，支持多种布局

FigTree

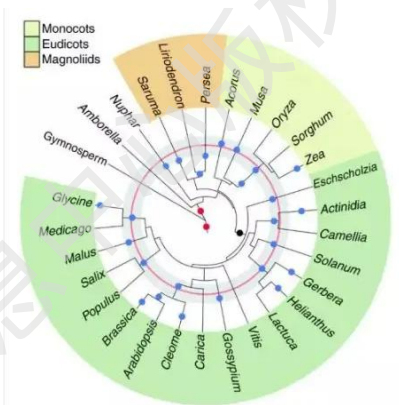
桌面应用，生成高质量图形

iTOL

在线交互式工具，支持定制化可视化

ggtree

R语言包，提供灵活性和可定制性



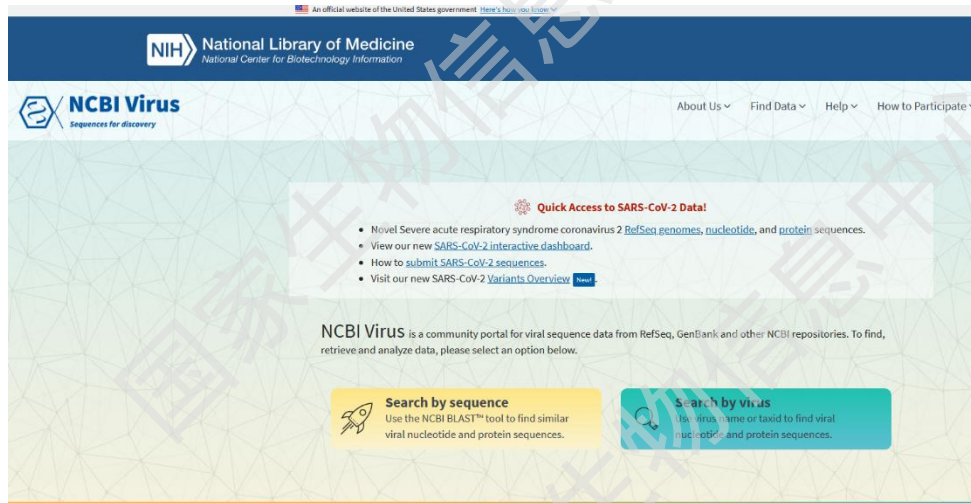
常用的系统发育树构建工具

分类	软件	说明	URL
桌面软件	PHYLIP	免费的、集成的进化分析工具	http://evolution.genetics.washington.edu/phylip.html
	MEGA	图形化、集成的进化分析工具	http://www.megasoftware.net/
	PAUP	商业软件，集成的进化分析工具	http://paup.csit.fsu.edu/
在线工具	NCBI Virus	支撑病毒序列数据集的检索、展示和分析	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
	NextStrain	提供用于揭示病原体的传播规律和进化特征的工具	https://nextstrain.org/



系统发育树常用工具

- 在线工具: 病毒序列信息库 NCBI Virus



<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>

- NCBI Virus是一种整合资源, 旨在支撑病毒序列数据集的检索、展示和分析。
- 建立时间: 2009年
- 机构: 美国国家生物技术信息中心 (NCBI)



系统发育树常用工具

- 在线工具: 病毒序列信息库 NCBI Virus

SARS-CoV-2 Data Hub

Quick Links: Betacoronavirus BLAST, CDC Outbreak Information, SARS-CoV-2 Articles in PubMed, SRA Data, Datasets command line

Advanced Filters for GenBank Sequences, Visual Filters for GenBank Sequences, Outbreak Sequence Statistics, Lineages Frequency and Location of GenBank + SRA Data, Search GenBank + SRA Data by Mutation

Refine Results: Virus/Taxonomy (1) x

Selected Results: 200 / 9,120,970

Build Phylogenetic Tree

Nucleotide (9,120,970) Protein (47,665,451)

Accession	Organism Name	GenBank
NC_045512.2	Severe acute respir...	
PV771649.1	Severe acute respir...	
PV771650.1	Severe acute respir...	
PV771651.1	Severe acute respir...	
PV771652.1	Severe acute respir...	

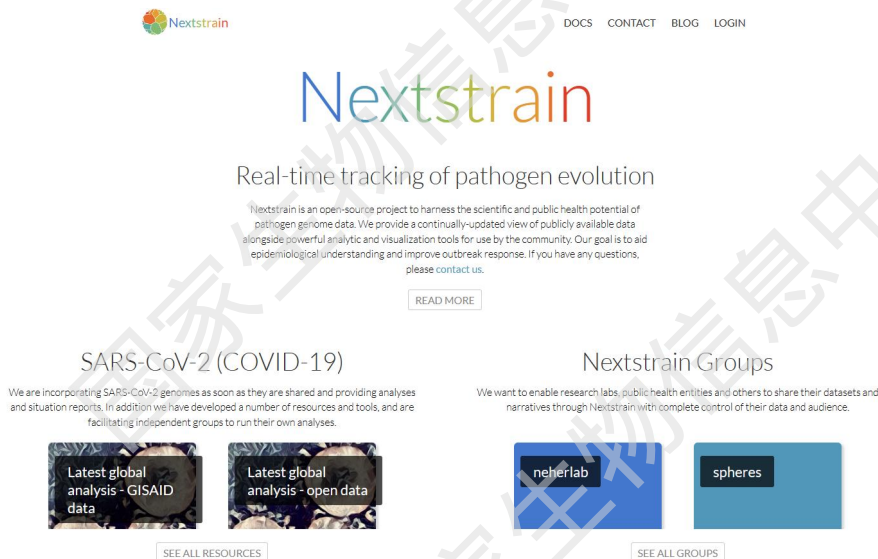
选择序列

Phylogenetic Tree



系统发育树常用工具

- 在线工具: **NextStrain**



<https://nextstrain.org/>

- 提供了工具和平台，帮助研究者对传染病病原体进行基因组学分析，揭示病原体的传播规律和进化特征。
- 建立时间：2018年
- 机构：Fred Hutchinson Cancer Research Center



系统发育树常用工具

- 在线工具: **NextStrain**

SARS-CoV-2: 用自己的数据集进行枝系、突变、质量分析

Nextclade^{v3.3.1}

Clade assignment, mutation calling, and sequence quality checks

Provide sequence data

File Link Text Example

Drag & drop files or folders

Select files

Selected reference dataset

☒ Suggest automatically Reset Suggest

SARS-CoV-2 (XBB)
official
Reference: Wuhan-Hu-1 with XBB SNPs
Updated at: 2024-02-16 04:00:32 (UTC)
Dataset name: nextstrain/sars-cov-2/XBB

Load example

Change reference dataset Run

选择数据集

选择参考数据集

点击分析

<https://clades.nextstrain.org/>

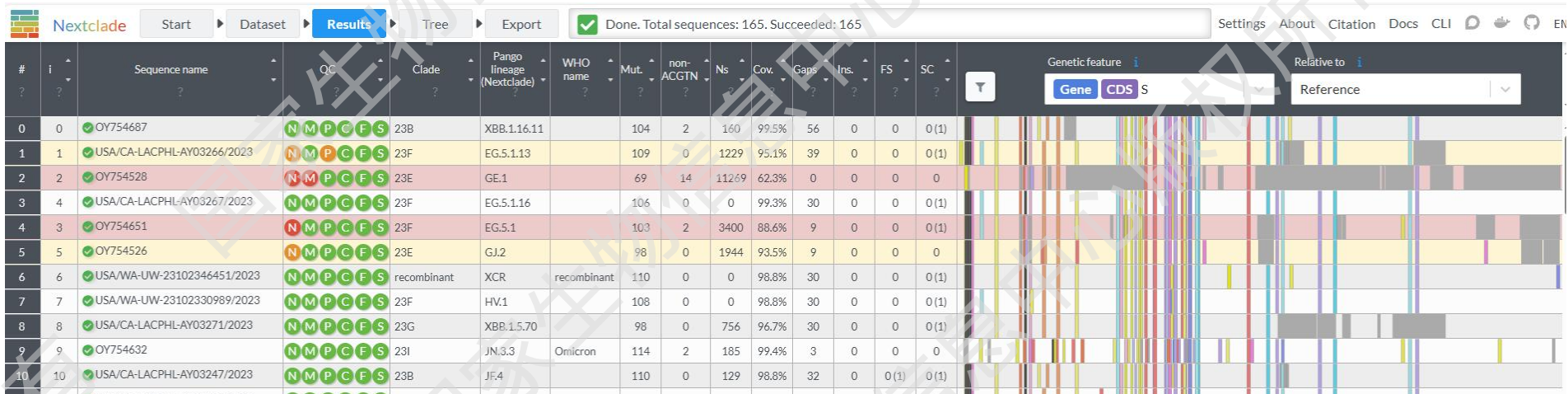


系统发育树常用工具

- 在线工具: **NextStrain**

SARS-CoV-2: 用自己的数据集进行枝系、突变、质量分析

每条序列的质量及突变信息

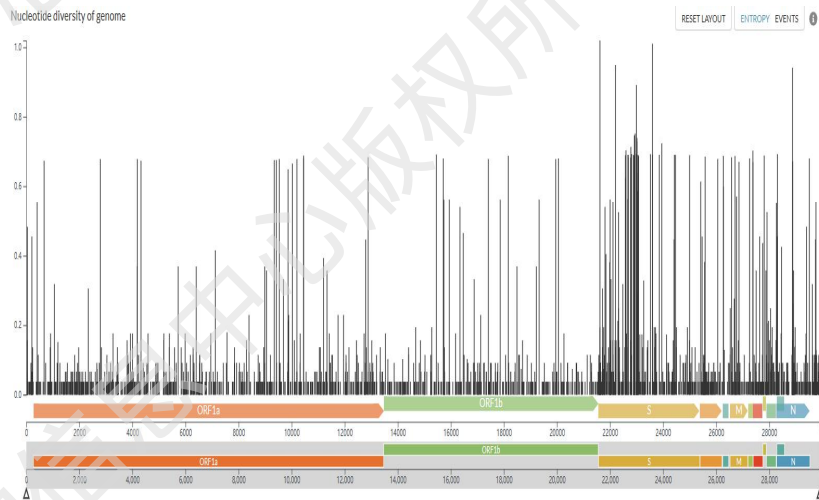
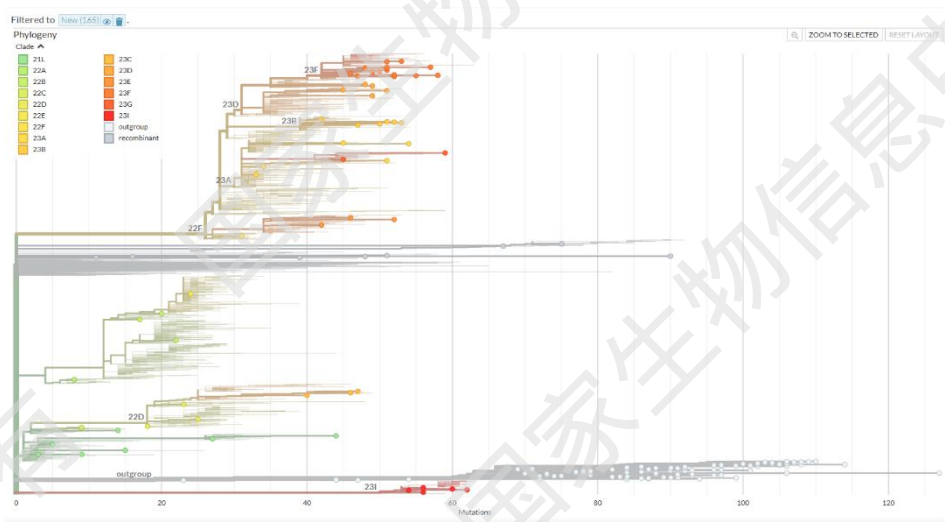


系统发育树常用工具

- 在线工具: **NextStrain**

SARS-CoV-2: 用自己的数据集进行枝系、突变、质量分析

支系划分及突变分布



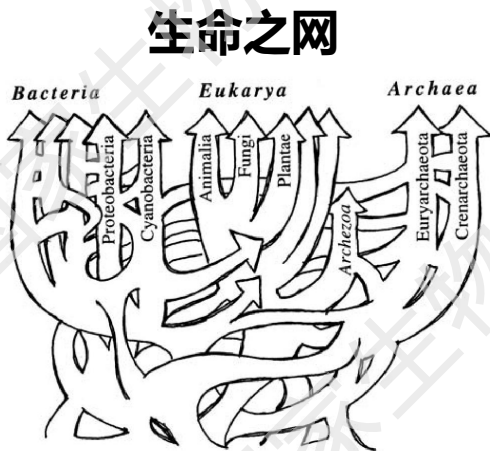
提纲

- 01 | 背景介绍
- 02 | 基于系统发育树的病原溯源
- 03 | 基于单体型网络的病原溯源**
- 04 | 基于序列相似性的病原溯源
- 05 | 实操练习



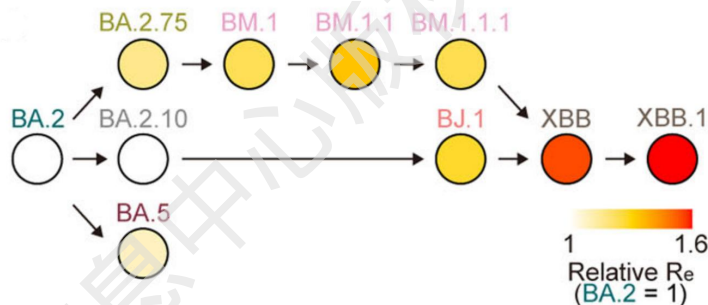
从“生命之树”到“生命之网”

- 进化过程中存在大量由杂交、水平基因转移、重组、基因转换或基因复制与丢失等环节引起的网状事件。



The concept of a rooted phylogenetic “network of life”, as proposed in [61]. Reprinted by permission of AAAS: Science, 284(5423):2124-2128 ©1999.

新冠病毒重组事件

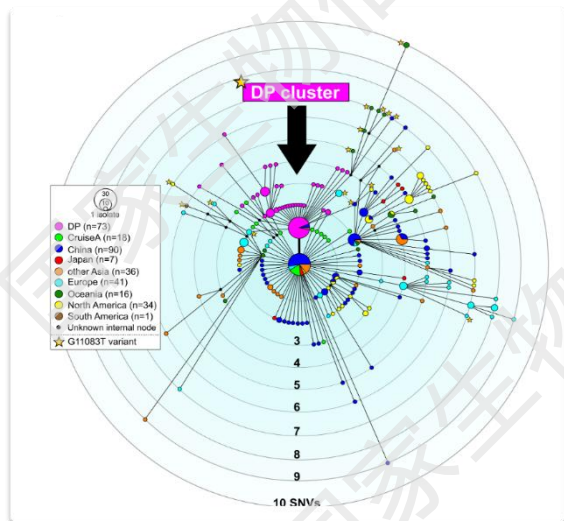


Tamura T et al. *Nat Commun*, 2023

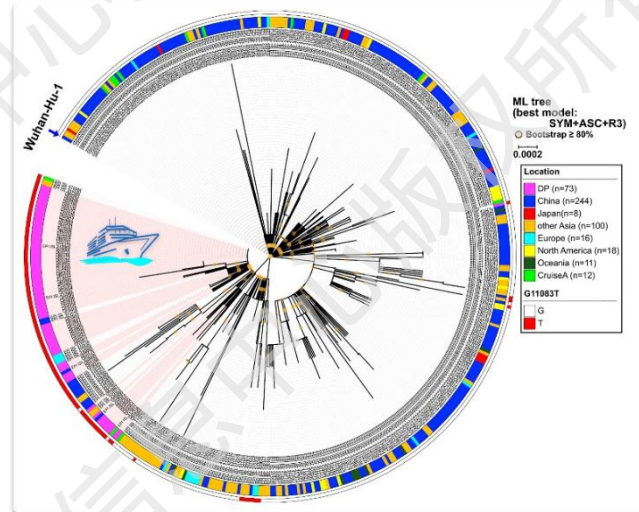


单体型网络 vs 系统发育树

- **共同点：**描述病原体演化传播路径
- **单体型网络特点：**有环，样本不仅在叶子节点上；通常用于种内演化关系的推断



钻石公主号单体型网络

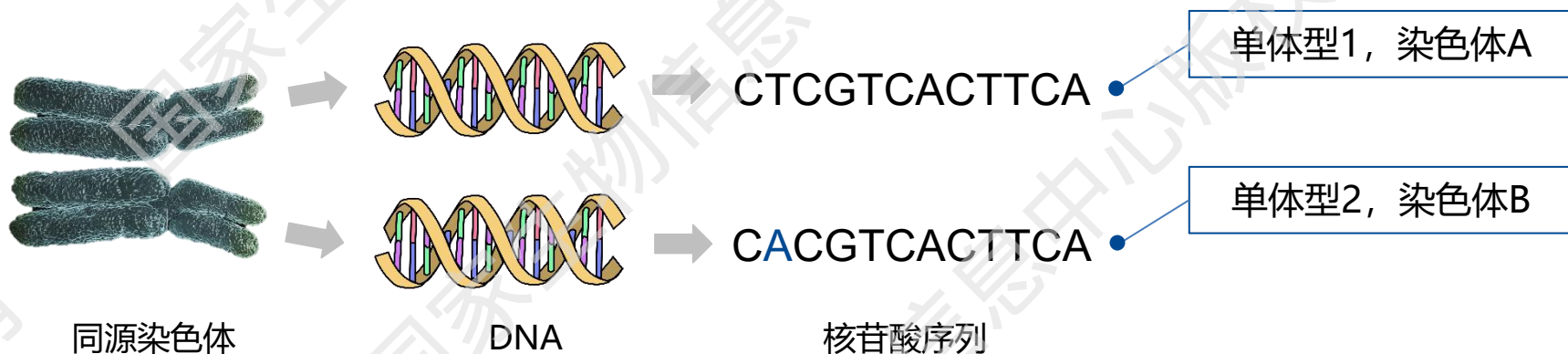


钻石公主号系统发育树

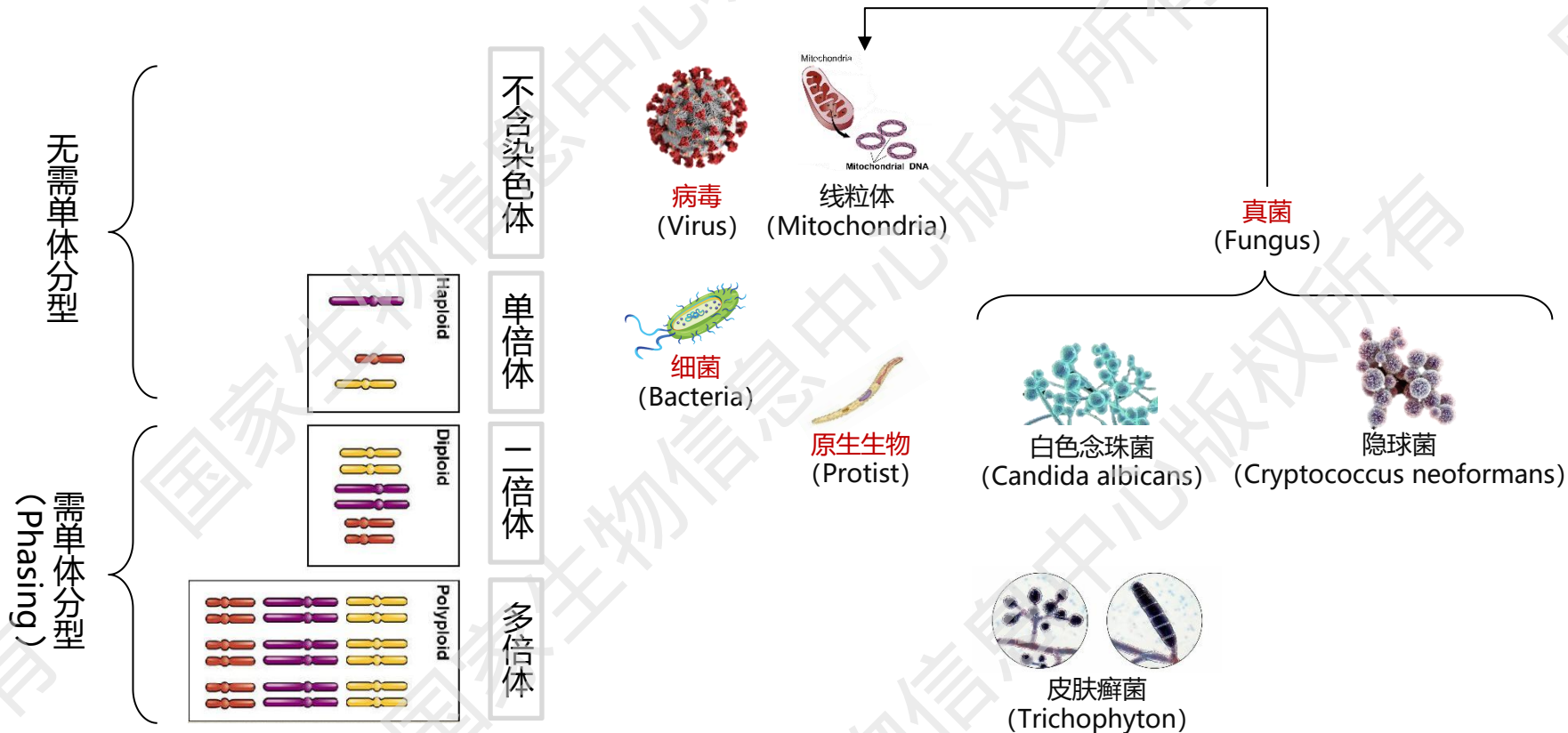


单体型的定义

单体型 (Haplotype, 也可称为单倍型) 是指生物体内由单亲遗传而来的一组基因。单体型 (haplotype) 一词来源于单倍体 (haploid) 和基因型 (genotype), 前者描述的是只有一组染色体的细胞, 后者指的是生物体的基因组成。



病原微生物与单体型示意图



单体型网络的定义

单体型网络 (Haplotype Network) 是一种用于遗传学研究的分析工具。通过构建单体型网络，研究者可以展示遗传变异之间的模式和连接，从而更好地理解种群内部的遗传多样性以及种群间的遗传差异。

单体型网络的组成

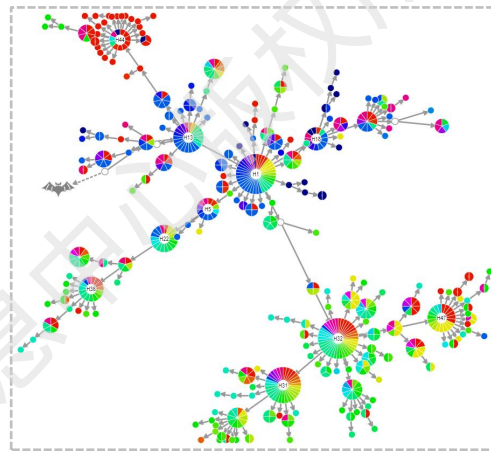
顶 点 : 单体型 (包含一些序列)

连接 (边) : 单体型之间的演化关系

点的大小 : 单体型中包含的序列数

边 长 : 差异位点数

新冠病毒单体型网络

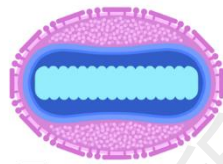
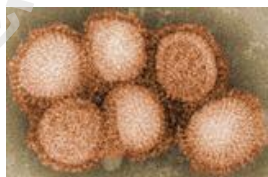


单体型网络的应用

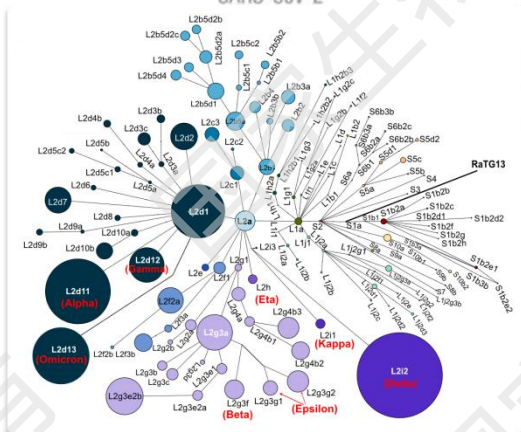
单体型网络可用于描述多种病原体的演化路径



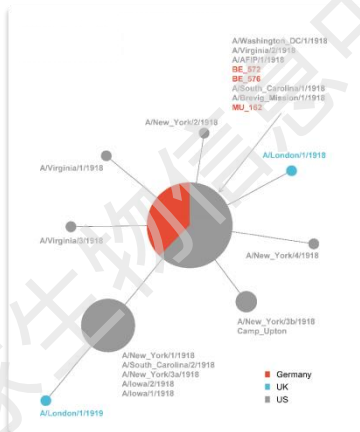
SARS-CoV-2



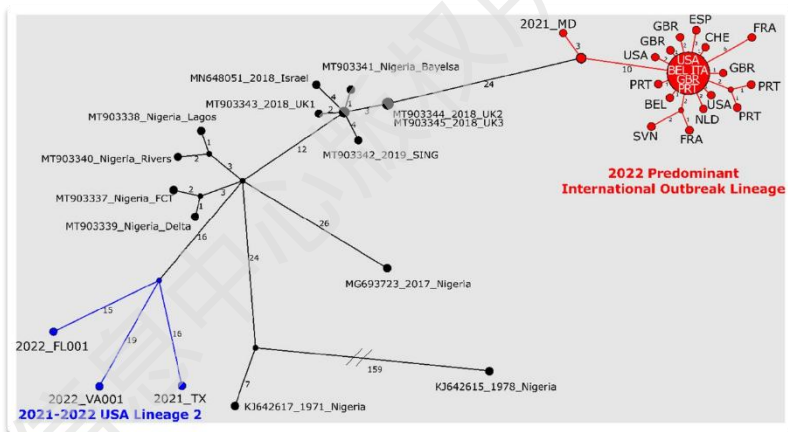
62 nanometers Monkeypox Virus



新冠病毒变体单体型网络



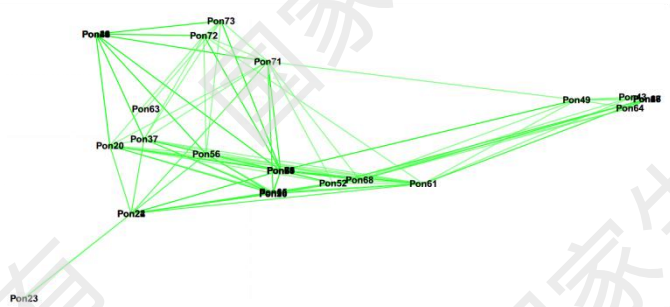
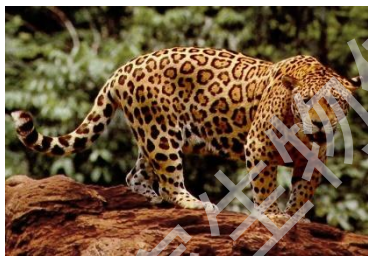
1918流感病毒单体型网络



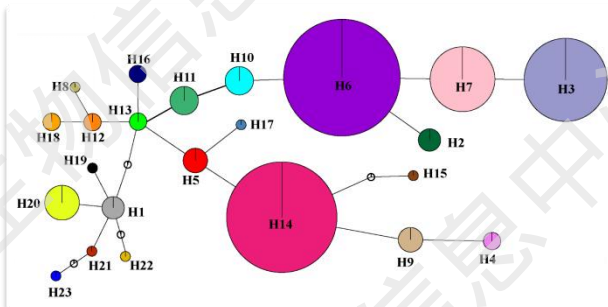
猴痘病毒单体型网络

单体型网络的应用

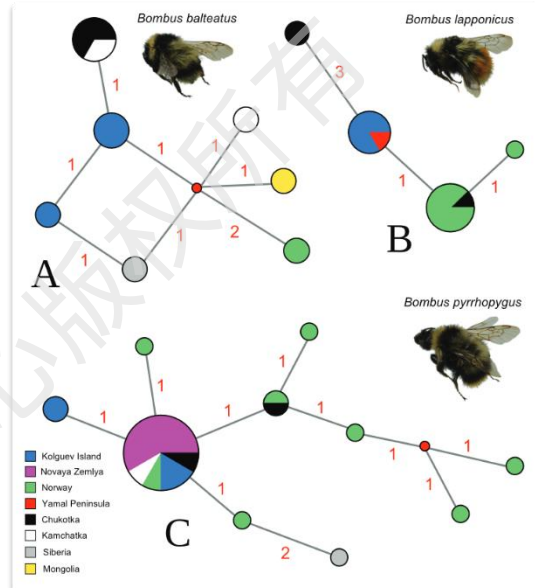
- 还可应用于动物、植物等多种生物类型
- 通常用线粒体或叶绿体DNA，有时用Y染色体



美洲虎 (*Panthera onca*) 线粒体DNA单体型网络



亚洲梨叶绿体DNA单体型网络

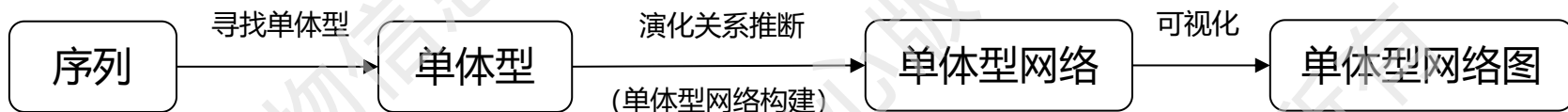


大黄蜂线粒体DNA单体型网络



单体型网络构建算法

- 单体型网络构建步骤



- 两类算法



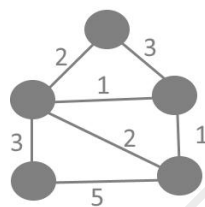
最小生成树 (MST)

- **目标**: 将单体型连接成一个树, 使边长之和最小
- **算法**: Kruskal算法、Prime算法
- **局限**: 无法有效处理病毒重组等典型的网状演化事件 (reticulate evolution)

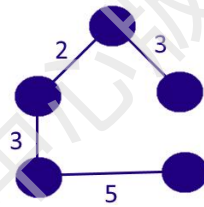
最小进化原则

The *minimum-evolution* (ME) method of phylogenetic inference is based on the assumption that the tree with *the smallest sum of branch length* estimates is most likely to be the true one.

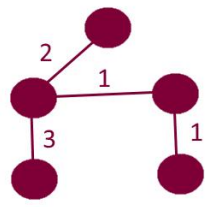
最小生成树示意图



Graph



Spanning Tree
Cost = 13



Minimum Spanning
Tree, Cost = 7

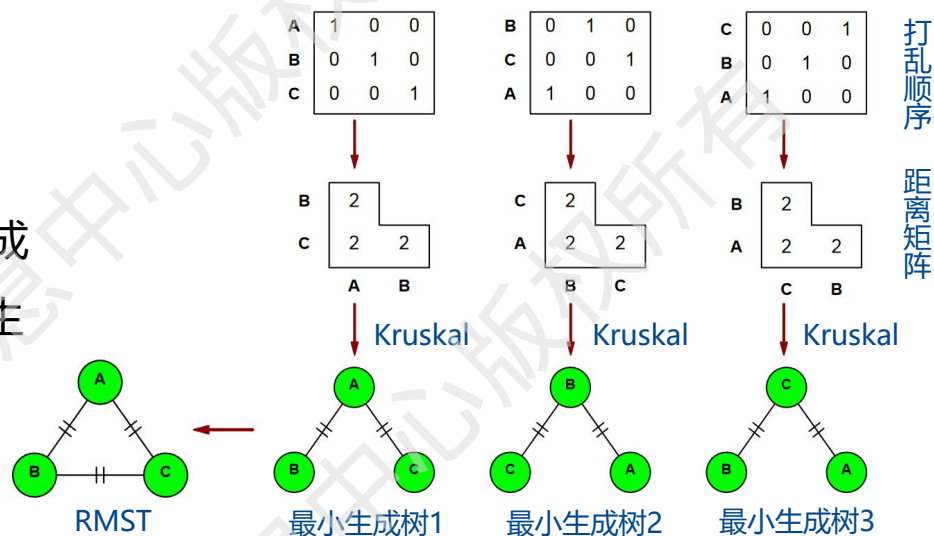
多种演化路径的生成： MSN与RMST

- 生成多种演化路径

- 最小生成网络算法 (MSN)：改进Kruskal算法，生成多种演化路径
- 随机最小生成树 (RMST)：随机生成多个最小生成树，并合并所有最小生成树

- 局限性

- 难以应对采样缺失

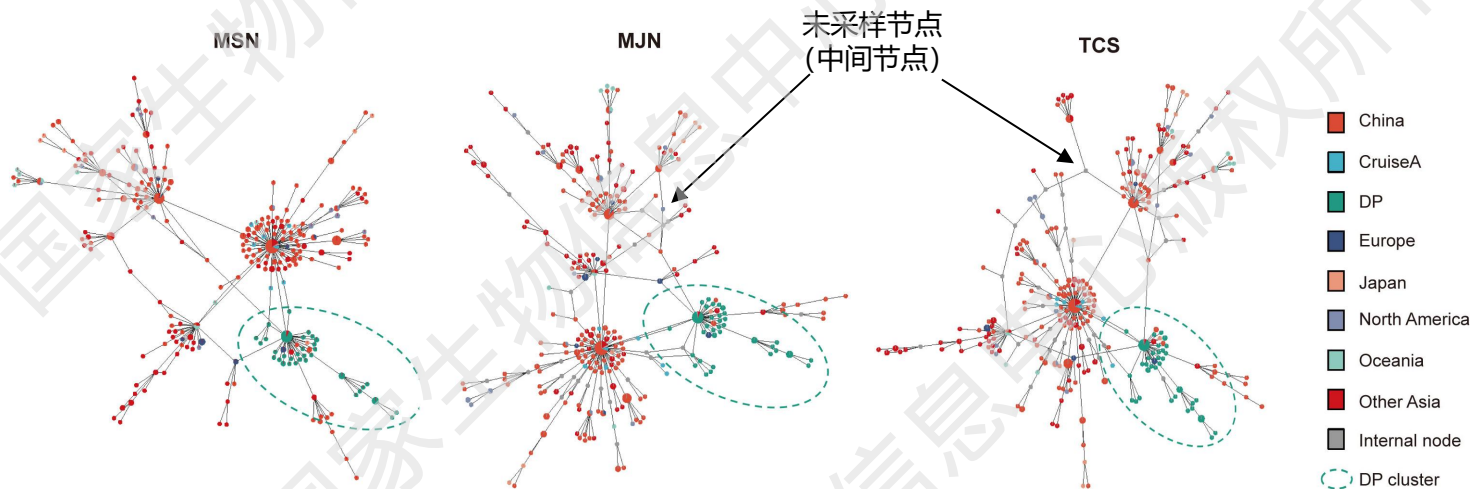


RMST算法示意图



推测未采样的样本

- 推测未采样样本
 - TCS (Templeton Crandall and Sing algorithm) 算法
 - MJN (median-joining network) 算法



钻石公主号新冠病毒单体型网络



MJN推断中间节点

- 推断中间节点
- 考虑稍复杂的情况
 - 3条序列
 - 字母表中有多个值 (如A、T、C和G)

U: **AGCGA**

V: **CAAGC**

W: **GCCAC**

Majority

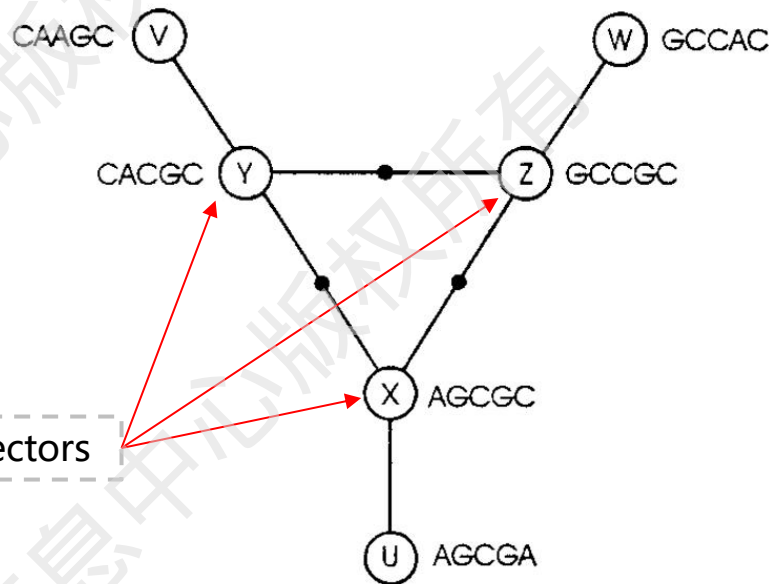
consensus: ??CGC

X: **AGCGC**

Y: **CACGC**

Z: **GCCGC**

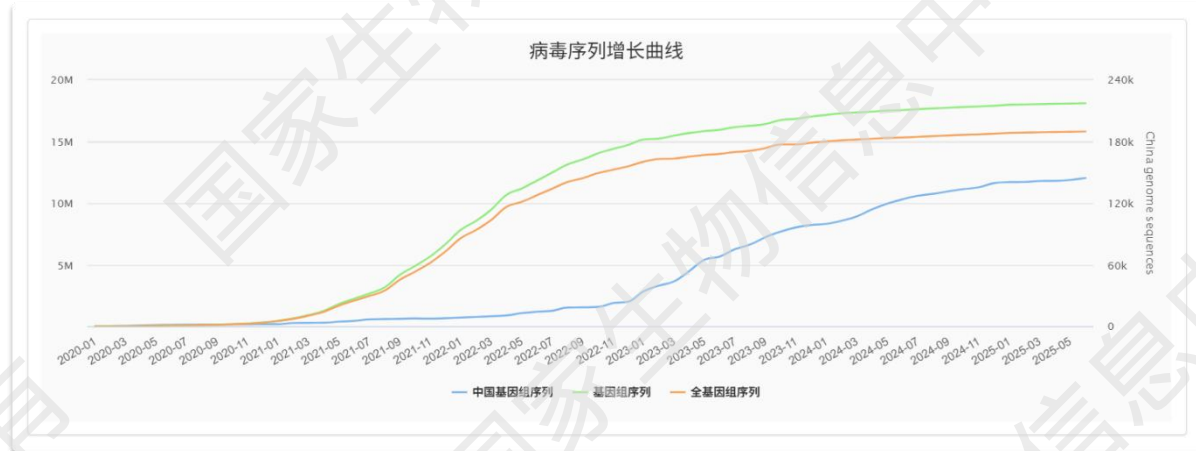
median vectors



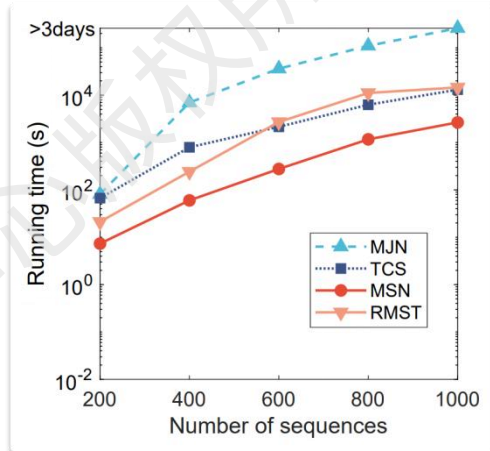
大数据时代的挑战

- 随着新冠疫情的暴发，产生了海量的新冠病毒基因组数据
- 对单体型网络构建带来了巨大挑战

RCoV19共收录新冠病毒序列1800余万条 (2025年6月24日)

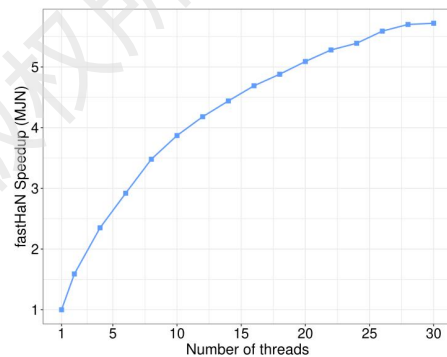
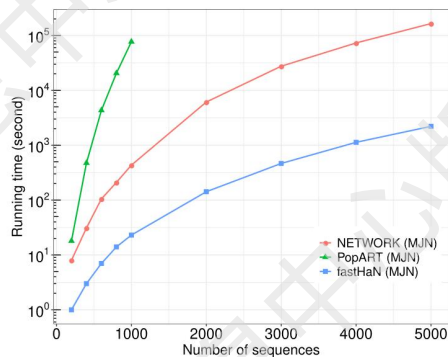
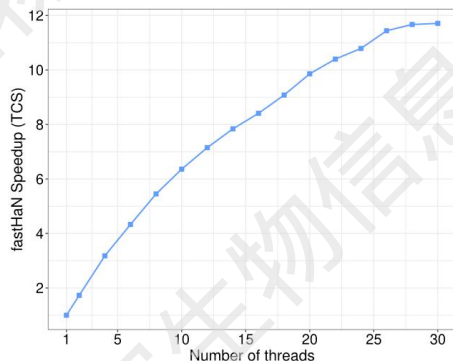
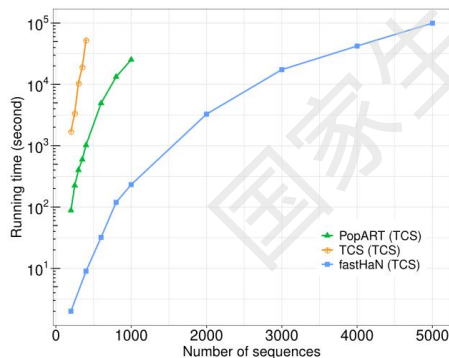


1000条新冠序列，MJN运行时间超过3天



fastHaN (fast Haplotype Network)

- 对传统算法TCS、MSN和MJN更高效的实现
- 处理样本量为十万量级的新冠病毒全基因组数据

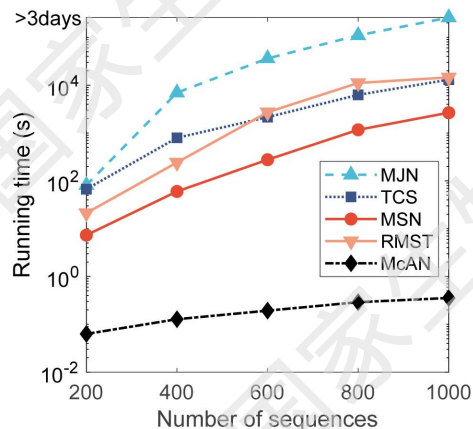


Chi et al., 2023, Molecular Ecology Resources



McAN (Minimum-cost Arborescence Network)

- 新颖的病原微生物演化关系快速推断算法
- 处理样本量为百万级的新冠病毒全基因组数据
 - McAN的运行速度提升两个数量级，内存消耗降低90%以上，同时不损失精度
 - 500万条新冠病毒序列，McAN算法仅耗时约25分钟（50线程）



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

Li et al., 2023, BIB



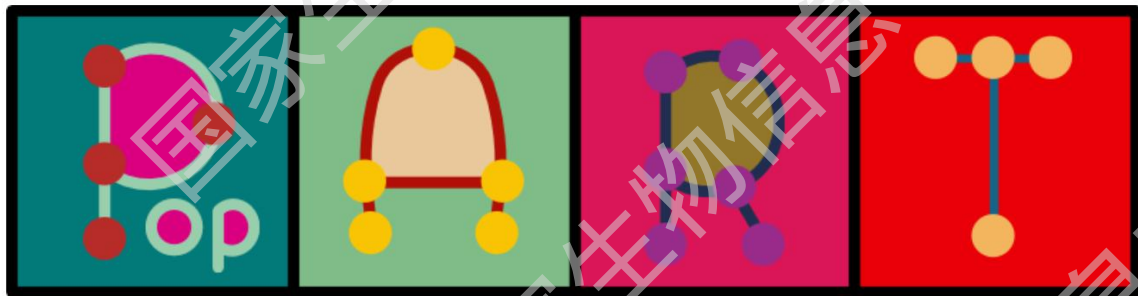
单体型构建软件 and 平台

类型	优点	代表性软件	实现的算法									
			TCS	MJN	MST	MSN	RMST	McAN	RM	Integer NJ	Ancestral MP Network	Tight Span Walker
命令行或库	灵活性和可自动化 : 可以轻松集成到脚本和其他软件项目中, 适合自动化任务和批处理。 资源利用高效 : 通常比图形界面软件更加轻量, 消耗更少的系统资源。	pegas	✓	✓	✓	✓	✓					
		fastHaN-cmd	✓	✓		✓						
		McAN-cmd						✓				
在线工具	无需安装 : 直接通过网页浏览器使用, 无需下载和安装。 自动更新 : 服务端更新软件, 用户无需手动操作, 始终使用最新版本。 跨平台 : 只要有网页浏览器, 可以在任何操作系统上使用, 用户体验一致。	fastHaN-platform	✓	✓		✓						
		tcsBU	✓									
		McAN-platform						✓				
桌面版软件	安全性 : 数据处理在本地完成, 不需要通过网络传输, 降低了数据泄露的风险。 易用性 : 用户能够轻松地学习、使用、理解和掌握软件。	POPART	✓	✓	✓	✓				✓	✓	✓
		HapNetworkView	✓	✓								
		HaploThread	✓	✓		✓		✓				



POPART桌面版软件

- **作者:** Jessica Leigh 等
- **单位:** Department of Mathematics and Statistics, University of Otago, New Zealand (奥塔哥大学)



POPART logo



Jessica Leigh

POPART桌面版软件

- **期刊:** Methods in Ecology and Evolution (IF: 6.6, 2022年)
- **引用:** 6256 (google学术, 截至2025年6月24日)

Methods in Ecology and Evolution



British Ecological Society

Methods in Ecology and Evolution 2015, **6**, 1110–1116

doi: 10.1111/2041-210X.12410

APPLICATION

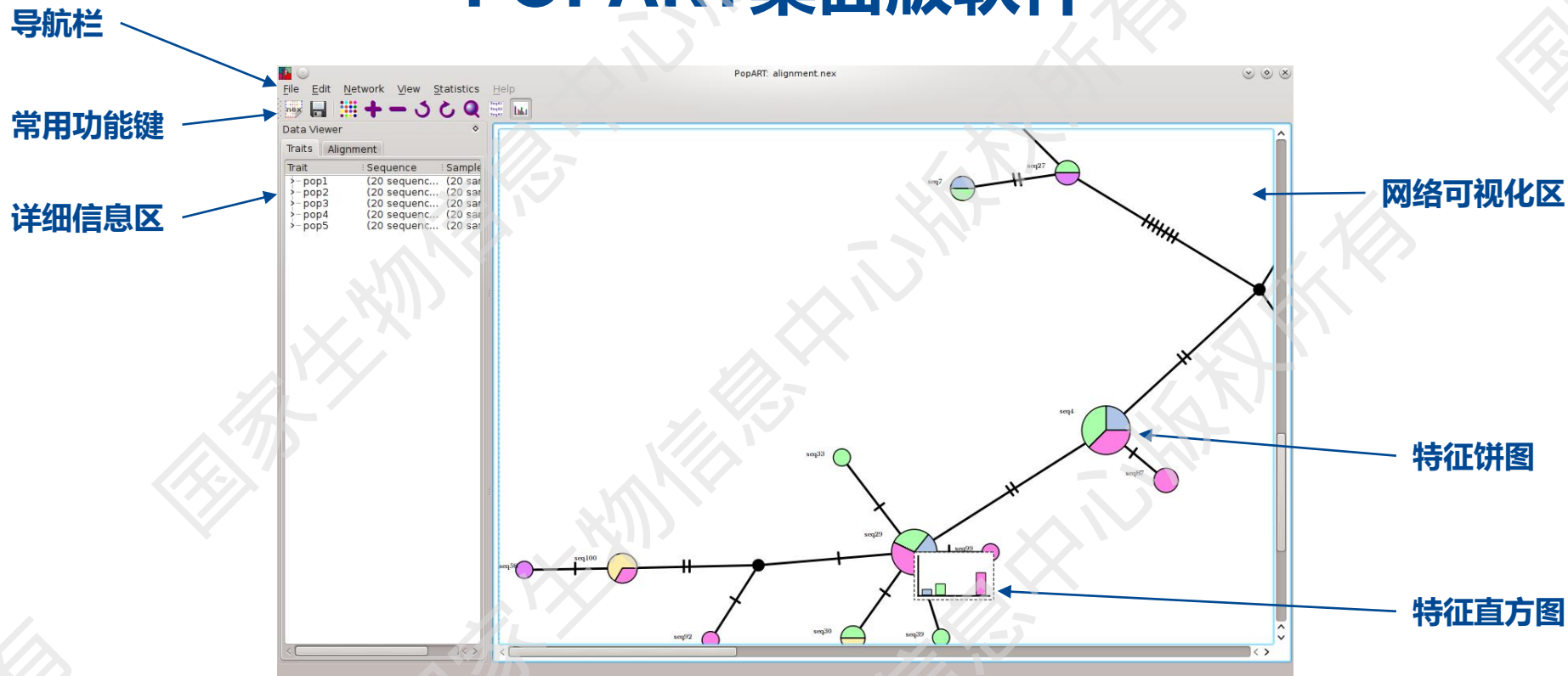
POPART: full-feature software for haplotype network construction

Jessica W. Leigh* and David Bryant

Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin 9054, New Zealand



POPART桌面版软件



POPART 桌面版软件截屏



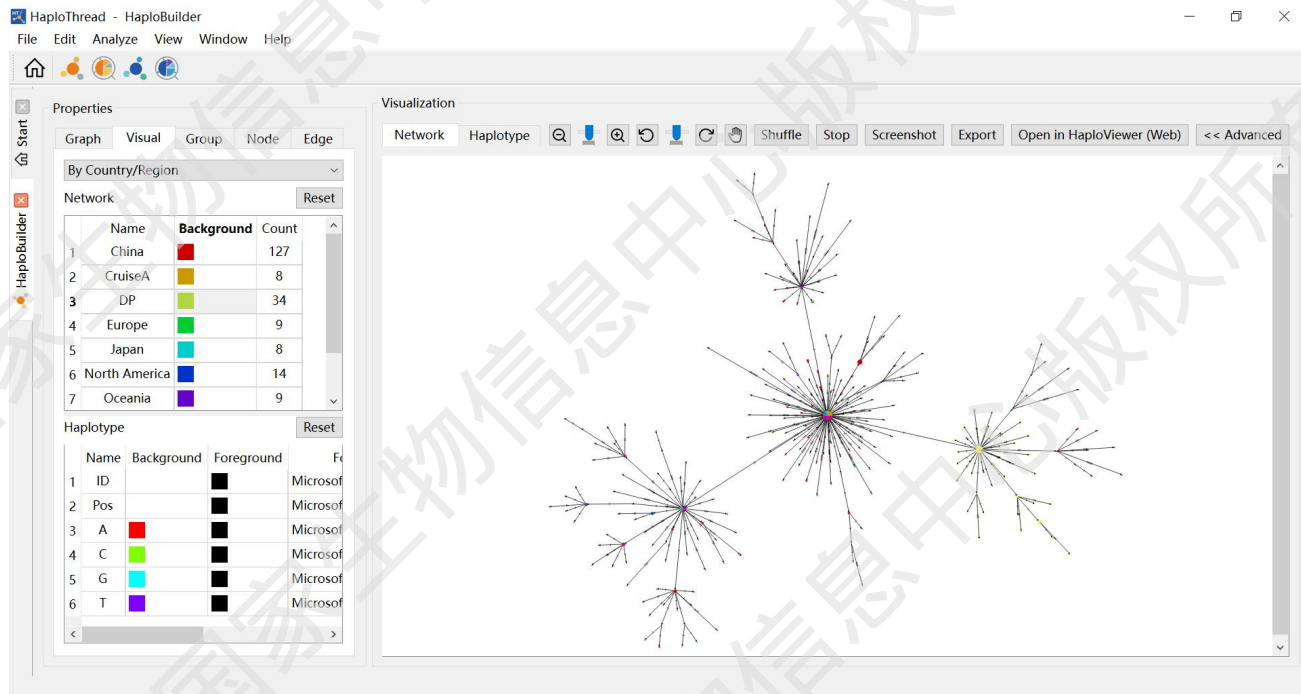
POPART桌面版软件

- **文件格式：**支持NEXUS文件格式
- **操作系统：**支持Windows、Mac OSX (≥ 10.6)、Linux/Unix/FreeBSD
- **算法：**集成了多种单体型网络构建算法
 - TCS、MJN、MSN、Integer NJ、Ancestral MP Network 和 Tight Span Walker



HaploThread桌面版软件

- 集成了多种多线程单体型网络构建算法



McAN在线分析平台

- **期刊**: Briefings in Bioinformatics (IF: 7.7, 2024)
- **单位**: 国家生物信息中心



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

<https://doi.org/10.1093/bib/bbad174>

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



McAN在线分析平台

全球地理分布

网络可视化区

单体型地理分布

序列详细信息

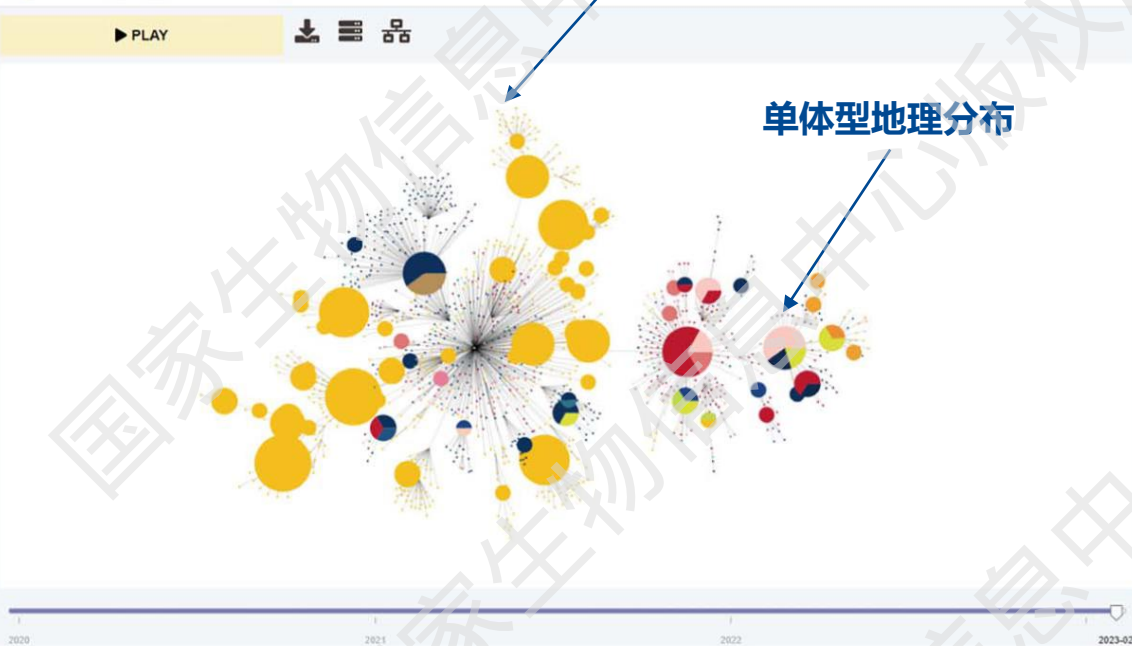
跳转到RCoV19数据库

McAN在线分析平台展示

时间轴

Viral Haplotype Network

PLAY



Country Information



Virus Information

Acc	Date	Locl	Name
EPI_ISL_16508442	2022-12-21	China	hCoV-19/Henan/HNCDCBT-619/2022
EPI_ISL_16635569	2022-12-18	China	hCoV-19/Zhejiang/VDC-865/2022



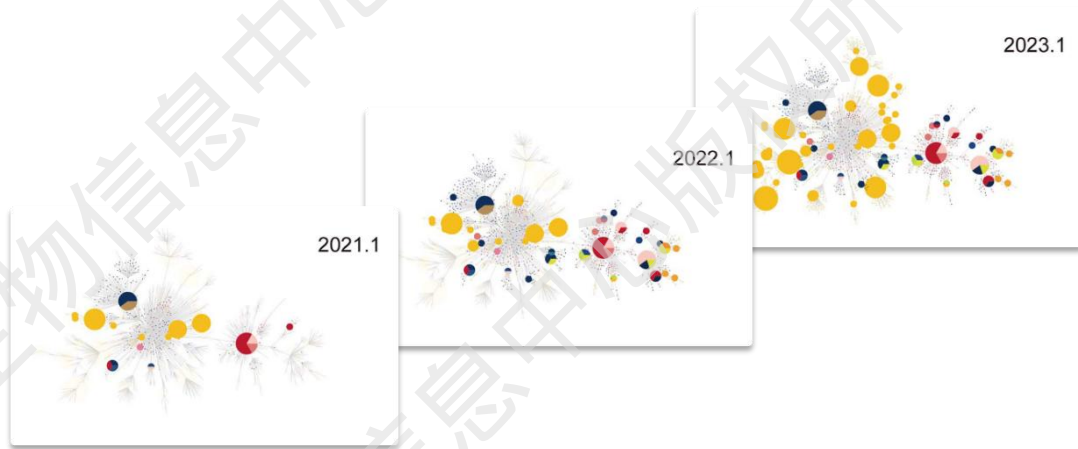
CNCB-NGDC

McAN在线分析平台

- **输入格式**: VCF (The Variant Call Format)、customized genovar formats
- **可视化功能**: 按时间展示单体型, 按团簇展示单体型



按团簇展示单体型



按时间展示单体型

提纲

- 01 | 背景介绍
- 02 | 基于系统发育树的病原溯源
- 03 | 基于单体型网络的病原溯源
- 04 | 基于序列相似性的病原溯源**
- 05 | 实操练习



序列相似性方法

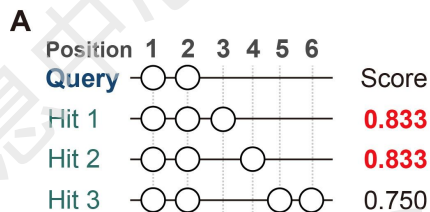
- 序列相似性反映了流行病学关联，是分子溯源中的重要步骤
- 与系统发育树、单体型网络相比速度较快

方法	特点
Civet	• 寻找与查询序列差异最小的序列 • 未考虑突变发生概率的差异
GISAID/AudacityInstant	
Genome-tracing	• 采用加权距离，提高准确性

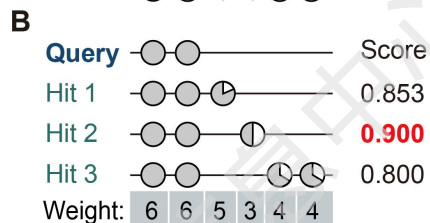


Genome-tracing算法

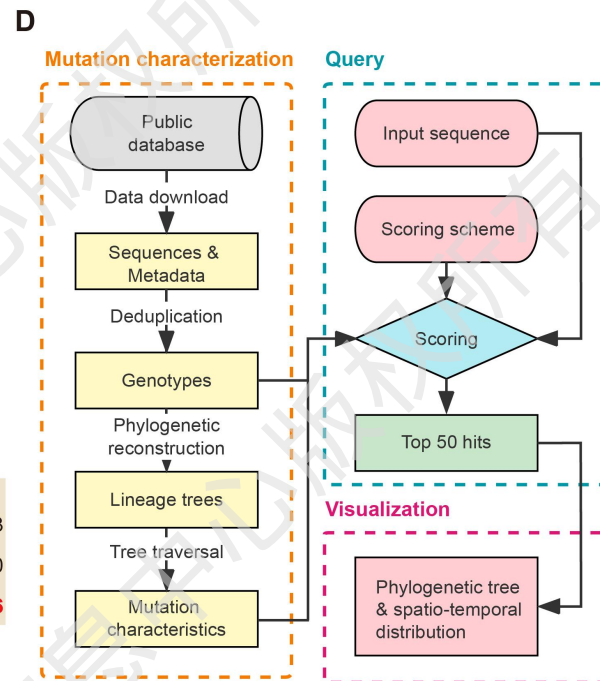
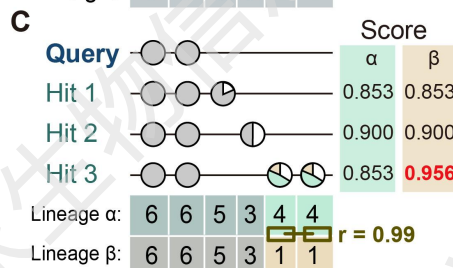
不加权



按发生率加权



按发生率加权且考虑突变间的连锁不平衡



Ma W et al., Brief Bioinform, 2023



Genome-tracing算法

距离计算公式:

$$\text{Rank}_{p,q} = \frac{1}{2} \times \frac{\sum_{i=1}^k W_i}{\sum_{i=1}^m W_i} + \frac{1}{2} \times \frac{\sum_{i=1}^k W_i}{\sum_{i=1}^n W_i}$$

其中， W_i 为突变权重， k 为共有的突变， m 和 n 为两条序列各自的突变。



Genome-tracing页面

Query sequence

Mutation test_seq:241C/T,3037C/T,14408C/T,23403A/G,28881G/A,28882G/C,2091C/T,5128A/G,8360A/G,1386C/T,19839T/C,19999G/T example

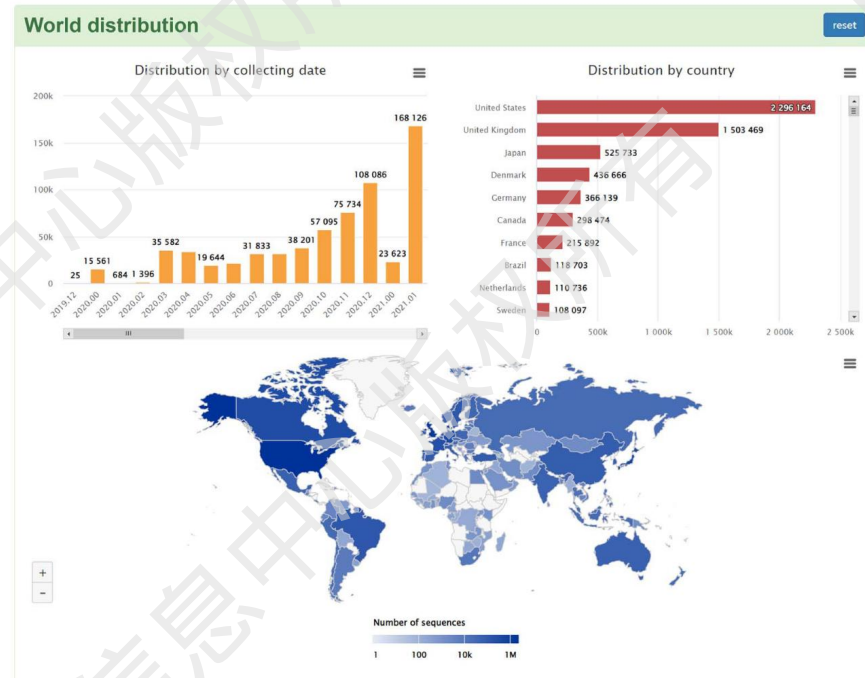
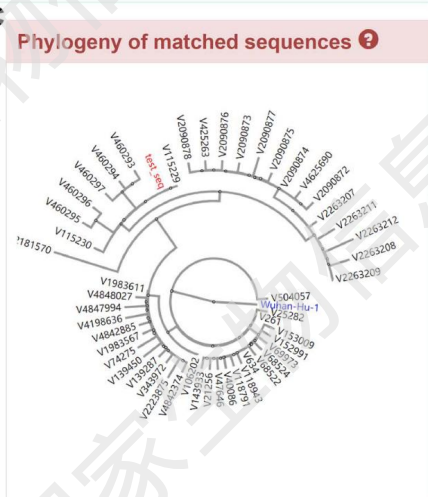
Fasta **Method** UM **Server** server01

Please ensure that you adhere to the mutation format specified as seq_ID:POS1Allele1/Allele2.POS2Allele1/Allele2, or alternatively, you can use the fasta format sequence as input on the webserver.

Matched sequences

Showing 1 to 50 of 50 entries

SeqID	Accession	Score	Lineage	Mutation	Sub
V115229	EPI_ISL_498892	1	B.1.1.317		0
V115230	EPI_ISL_498893	0.986667	B.1.1.317		1
V2263209	EPI_ISL_9301315	0.928571	B.1.1.317		2
V460294	EPI_ISL_2170890	0.911765	B.1.1.317		3
V460295	EPI_ISL_2170891	0.911765	B.1.1.317		3
V460296	EPI_ISL_2170888	0.911765	B.1.1.317		3
V460297	EPI_ISL_2170889	0.911765	B.1.1.317		3
V460293	EPI_ISL_2170887	0.911765	B.1.1.317		3
V2090874	OM098414.1	0.897619	B.1.1.317		3
V2263207	EPI_ISL_9301319	0.897619	B.1.1.317		3
V2090872	OM098425.1	0.870536	B.1.1.317		4
V2090875	OM098416.1	0.870536	B.1.1.317		4
V2090876	OM098417.1	0.870536	B.1.1.317		4
V2090878	OM098419.1	0.870536	B.1.1.317		4
V425263	EPI_ISL_2131805	0.870536	B.1.1.317		4
V2090873	OM098426.1	0.870536	B.1.1.317		4



提纲

- 01 | 背景介绍
- 02 | 基于系统发育树的病原溯源
- 03 | 基于单体型网络的病原溯源
- 04 | 基于序列相似性的病原溯源
- 05 | 实操练习



实操练习1：基于MEGA软件进行病原溯源

- 用MEGA与FigTree软件复现钻石公主号系统发育树

论文

ARTICLES ▾ FRONT MATTER AUTHORS ▾ TOPICS +

RESEARCH ARTICLE | MICROBIOLOGY

Haplotype networks of SARS-CoV-2 infections in the *Diamond Princess* cruise ship outbreak

Tsuyoshi Sekizuka, Kentaro Itokawa, Tsutomu Kageyama, and Makoto Kuroda

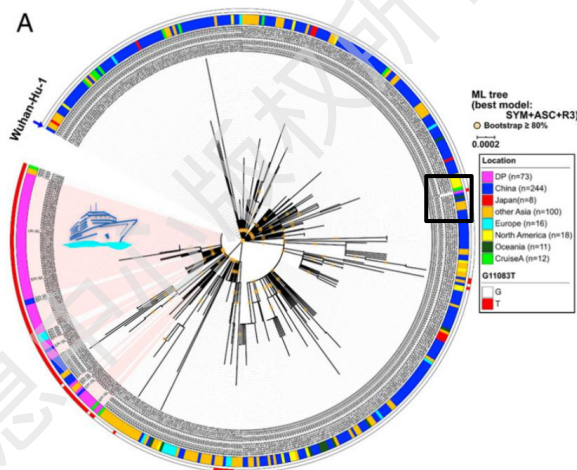
Edited by Y. Kawaoka, University of Wisconsin–Madison, Madison, WI, and approved July 2, 2020 (received for review April 10, 2020)

July 28, 2020 | 117 (33) 20198–20201 | <https://doi.org/10.1073/pnas.2006824117>

28,735 | 53

PDF/EPUB

论文中的系统发育树

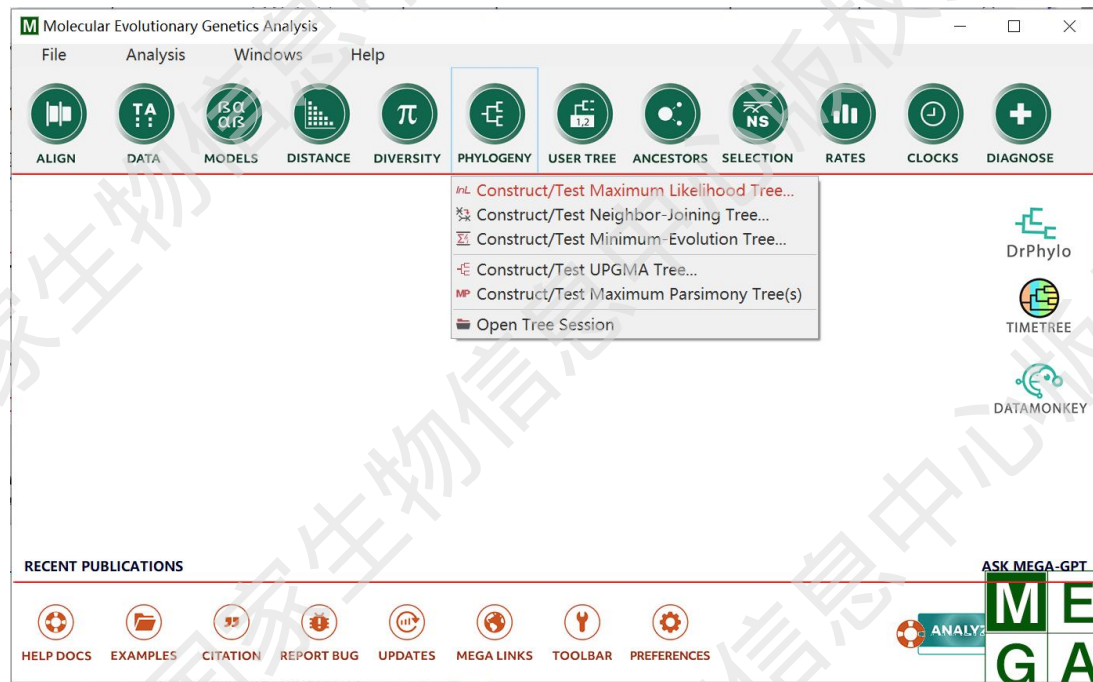


钻石公主号样本形成一个簇，支持单次引入



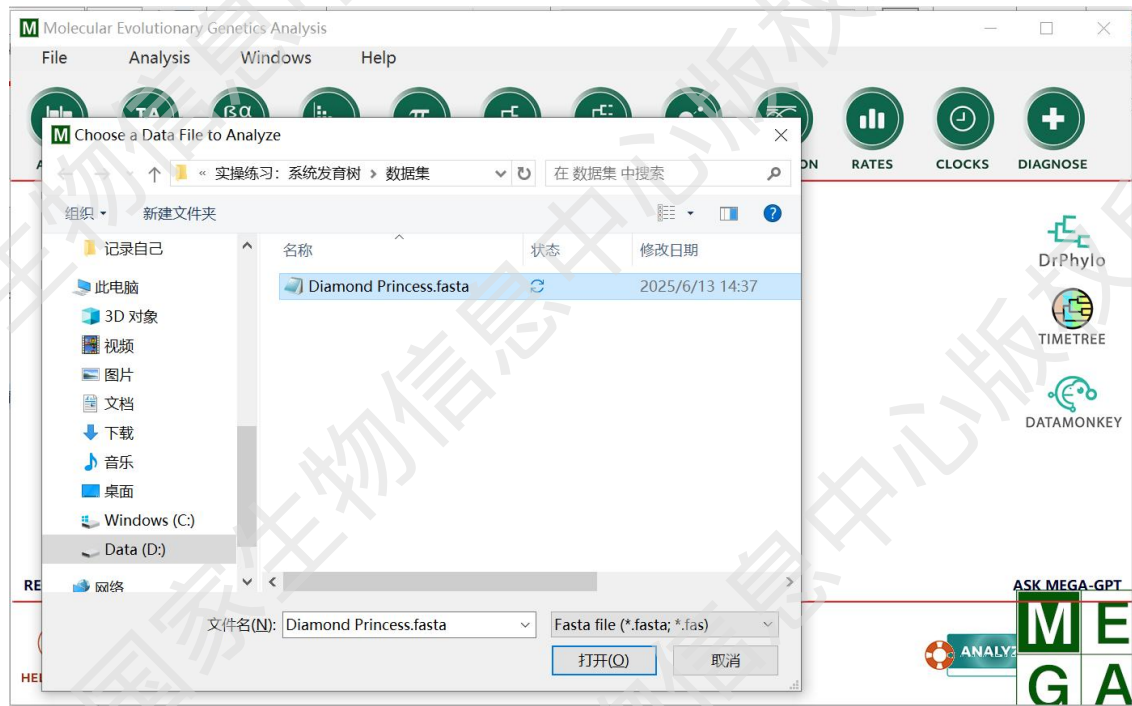
实操练习1：基于MEGA软件进行病原溯源

- 点击PHYLOGENY -> Construct/Test Maximum Likelihood Tree...



实操练习1：基于MEGA软件进行病原溯源

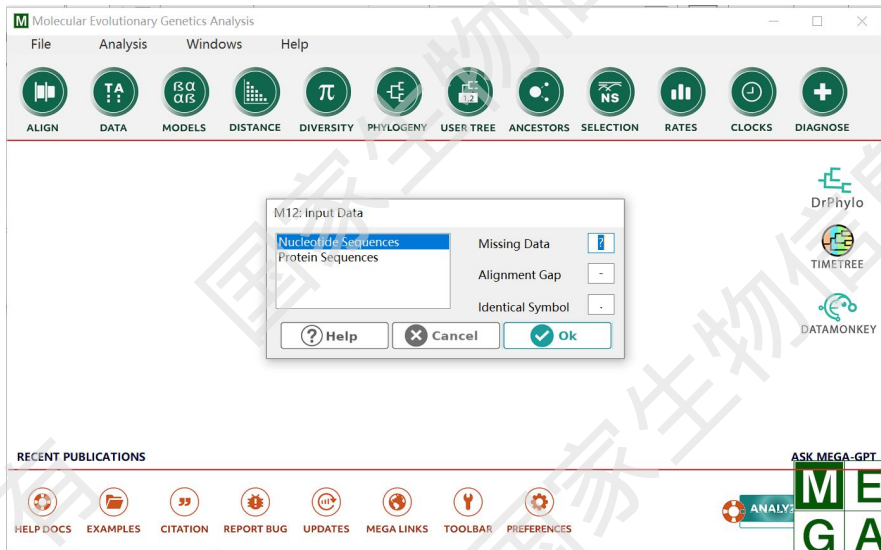
- 导入数据：practice\1_practice_phylogenetic_tree\dataset\Diamond_Princess.fasta



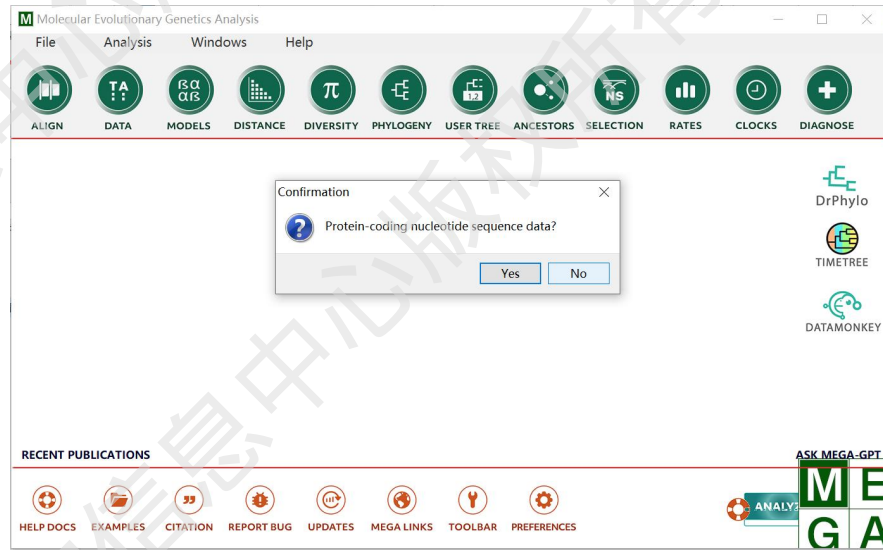
实操练习1：基于MEGA软件进行病原溯源

- 选择数据类型

选Nucleotide Sequences点击ok



点击No



实操练习1：基于MEGA软件进行病原溯源

- 点击OK

M12: Analysis Preferences

Phylogeny Reconstruction

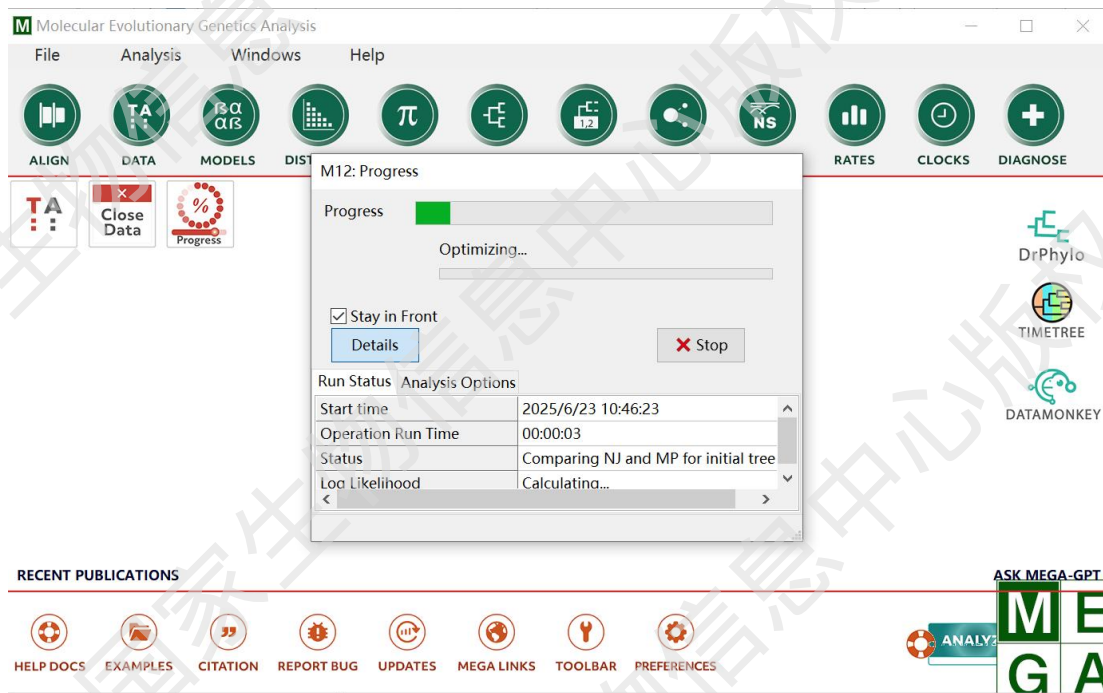
Option	Setting
ANALYSIS	
Statistical Method →	Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny →	None
SUBSTITUTION MODEL	
Substitutions Type →	Nucleotide
Model/Method →	Tamura-Nei model
RATES AND PATTERNS	
Rates among Sites →	Uniform Rates
DATA SUBSET TO USE	
Gaps/Missing Data →	Use all sites
TREE INFERENCE OPTIONS	
ML Heuristic Method →	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML →	Make initial tree automatically (Default - NJ/M)
Branch Swap Filter →	None
SYSTEM RESOURCE USAGE	
Number of Threads →	7

? Help X Cancel ✓ Ok



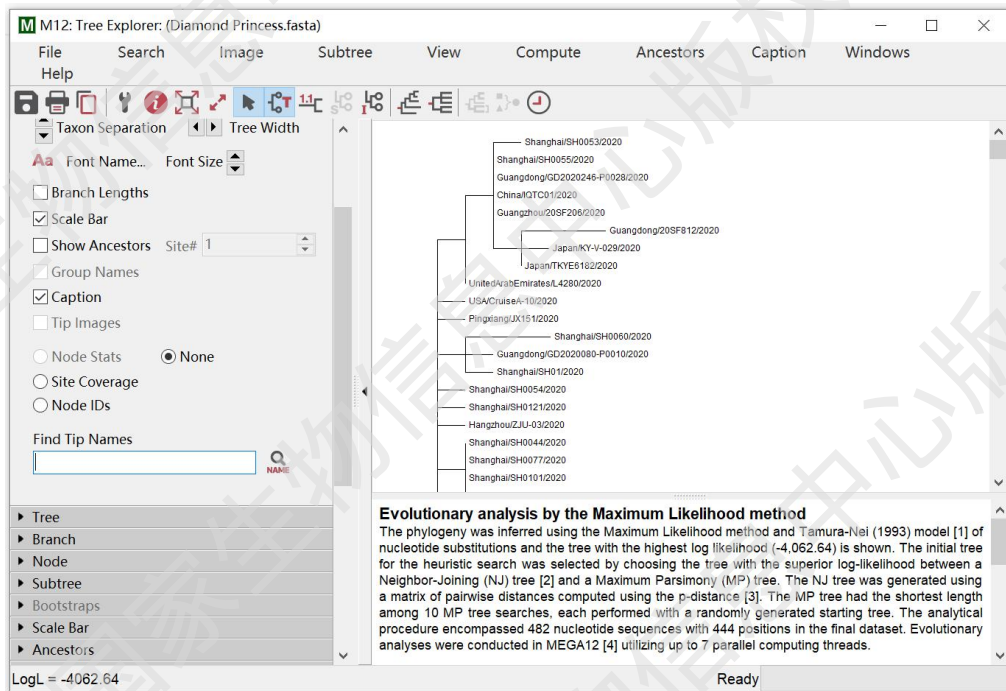
实操练习1：基于MEGA软件进行病原溯源

- 开始构建系统发育树



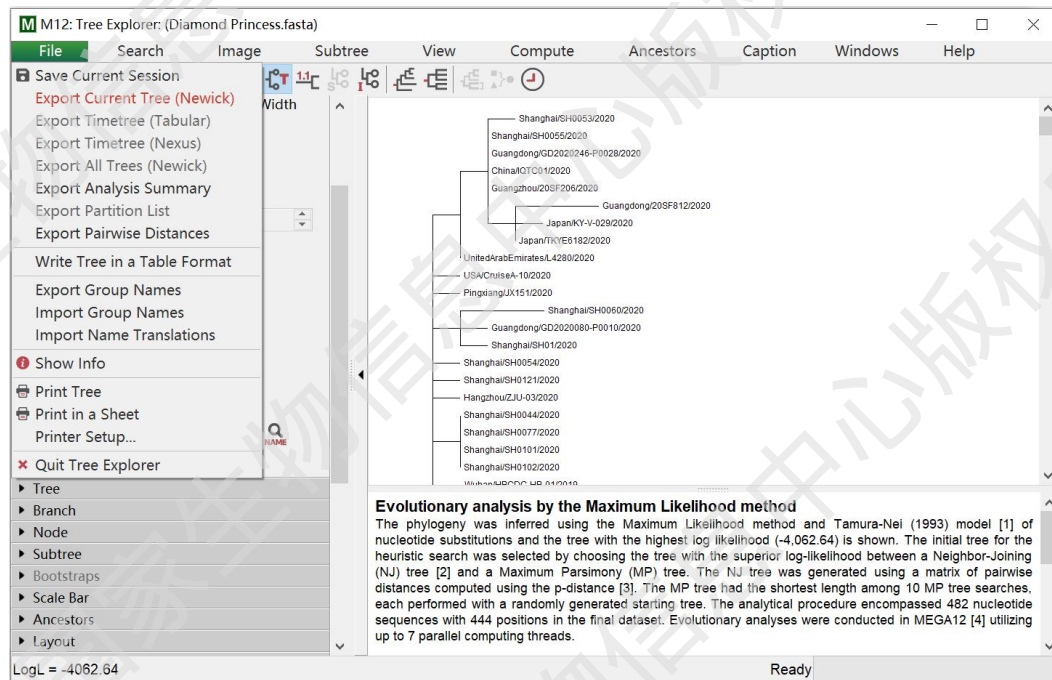
实操练习1：基于MEGA软件进行病原溯源

- 程序运行完成后，会展示构建完成的系统发育树



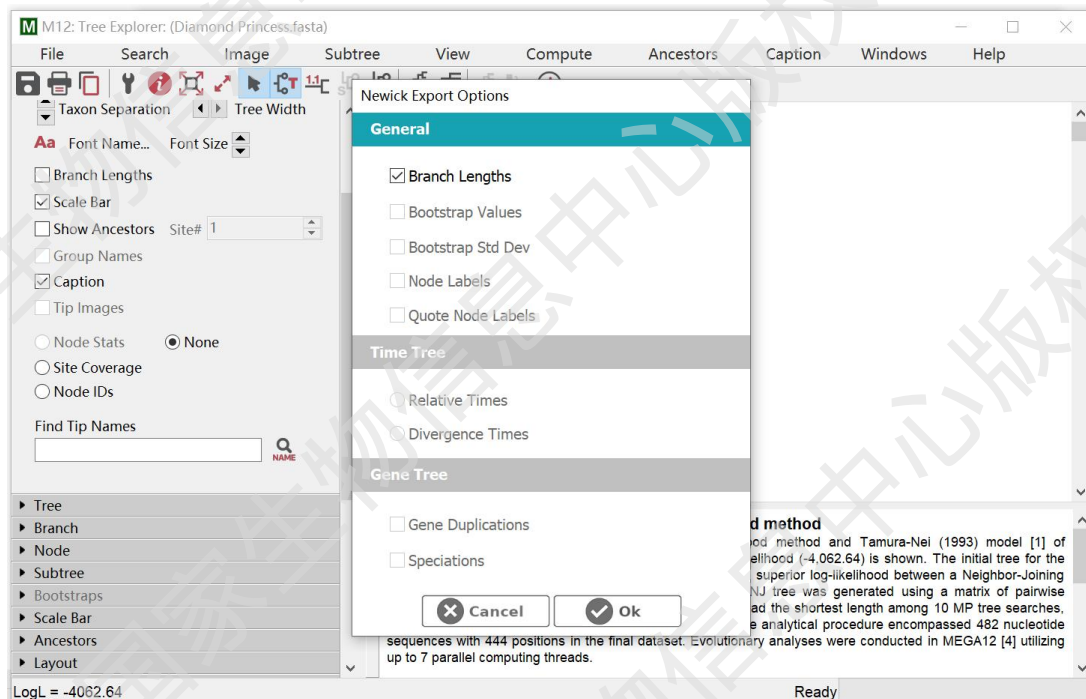
实操练习1：基于MEGA软件进行病原溯源

- 导出系统发育树：点击File -> Export Current Tree (Newick)



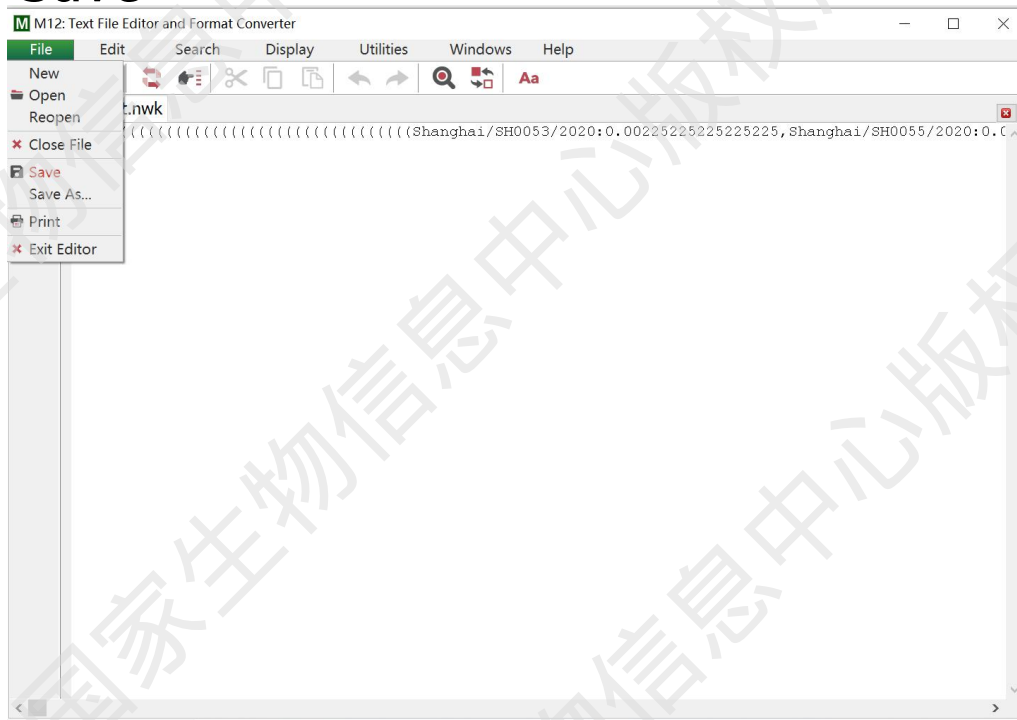
实操练习1：基于MEGA软件进行病原溯源

- 点击OK



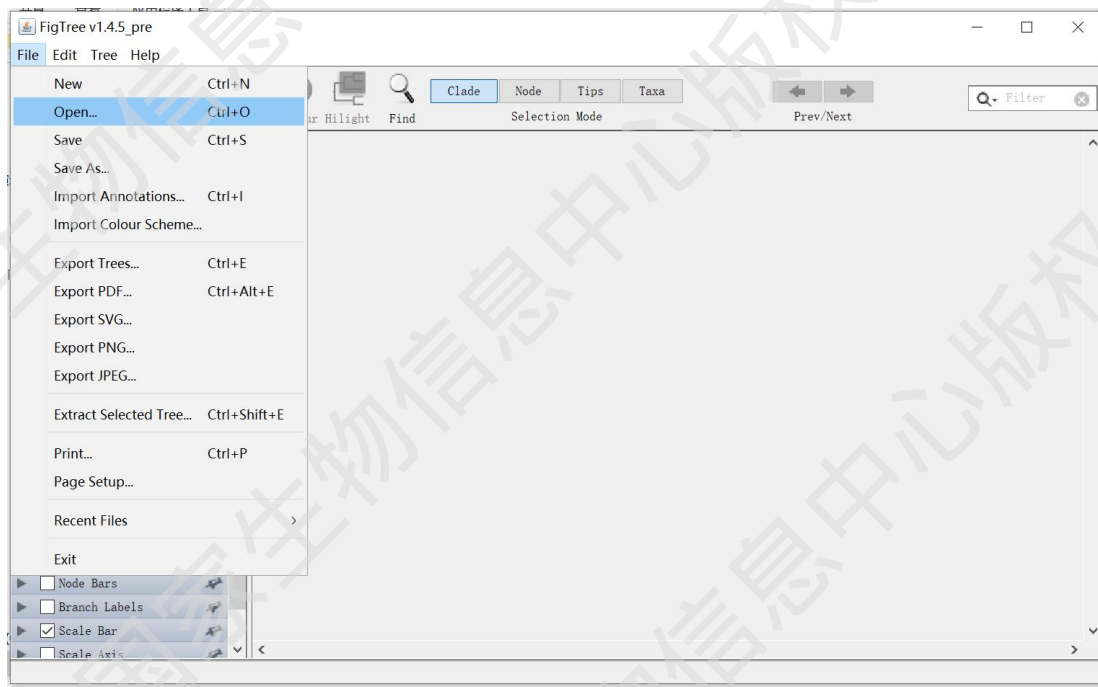
实操练习1：基于MEGA软件进行病原溯源

- 点击File -> Save



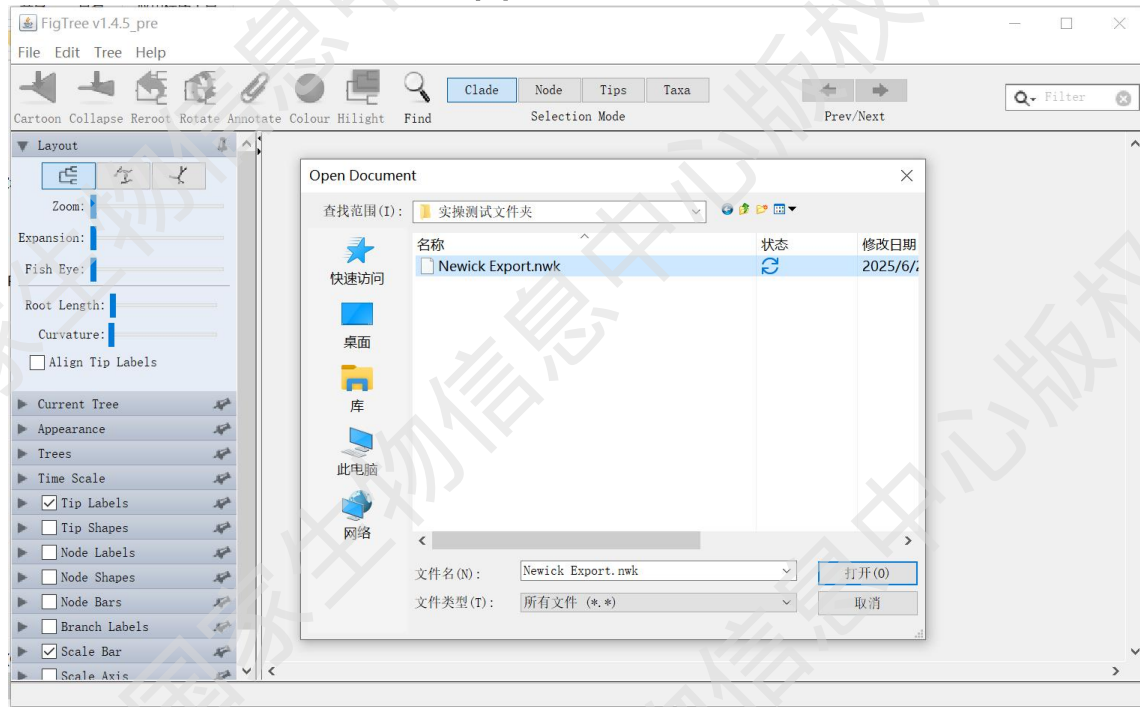
实操练习1：基于MEGA软件进行病原溯源

- 运行FigTree软件。点击 File -> Open。



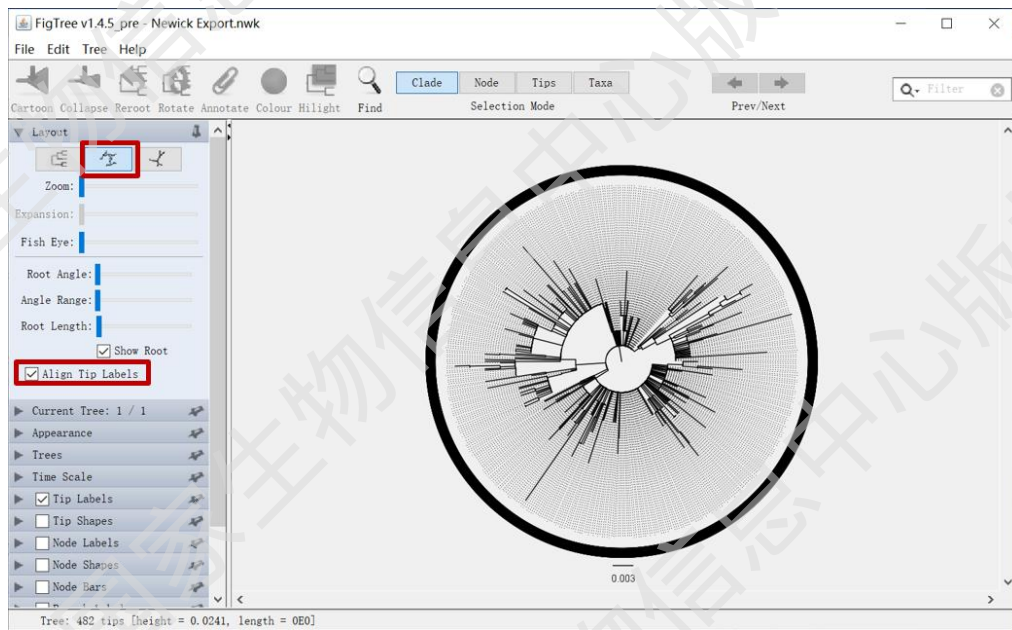
实操练习1：基于MEGA软件进行病原溯源

- 打开刚刚导出的Newick文件



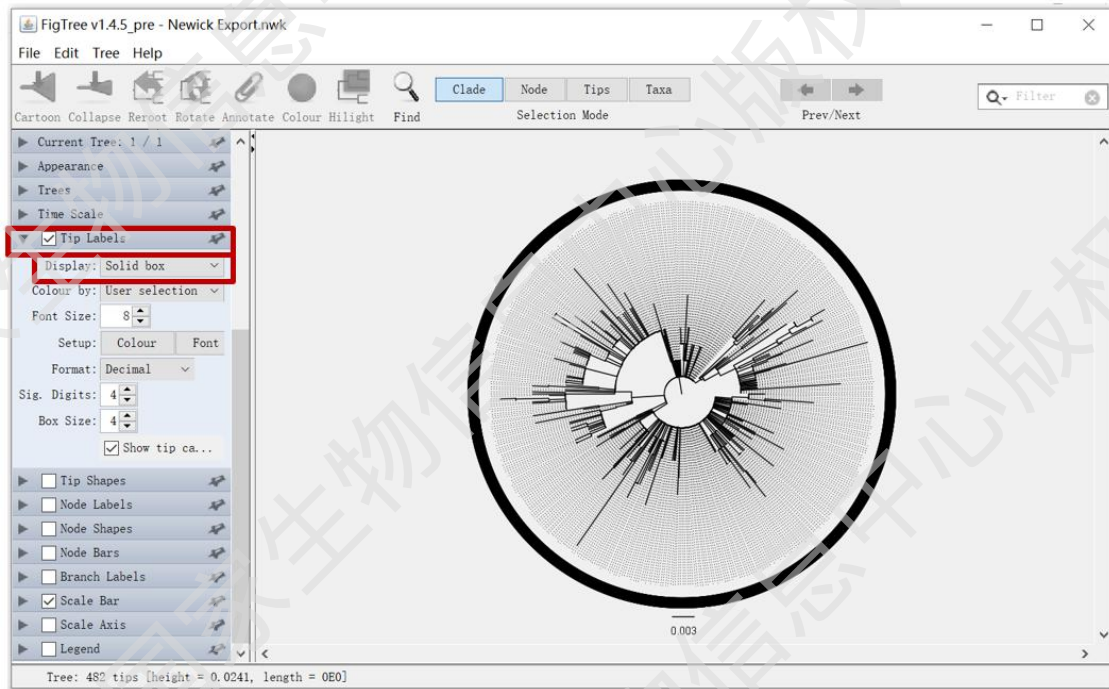
实操练习1：基于MEGA软件进行病原溯源

- 参数设置：在Layout中，点击Polar tree layout，并勾选Align Tip Labels（见红框）



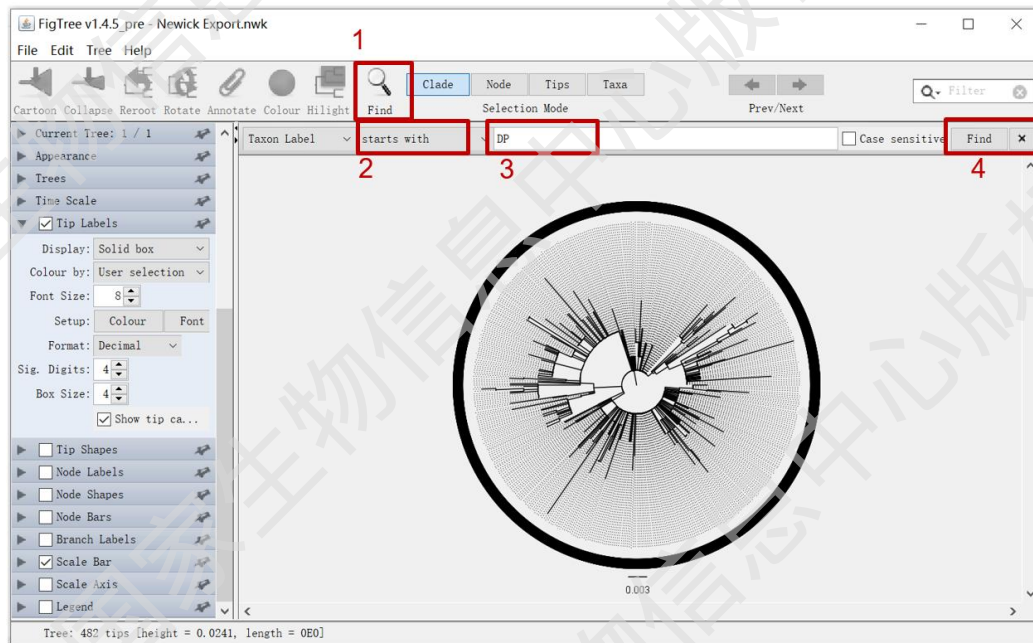
实操练习1：基于MEGA软件进行病原溯源

- 参数设置：点击Tip Labels。在Display中选择Solid Box。



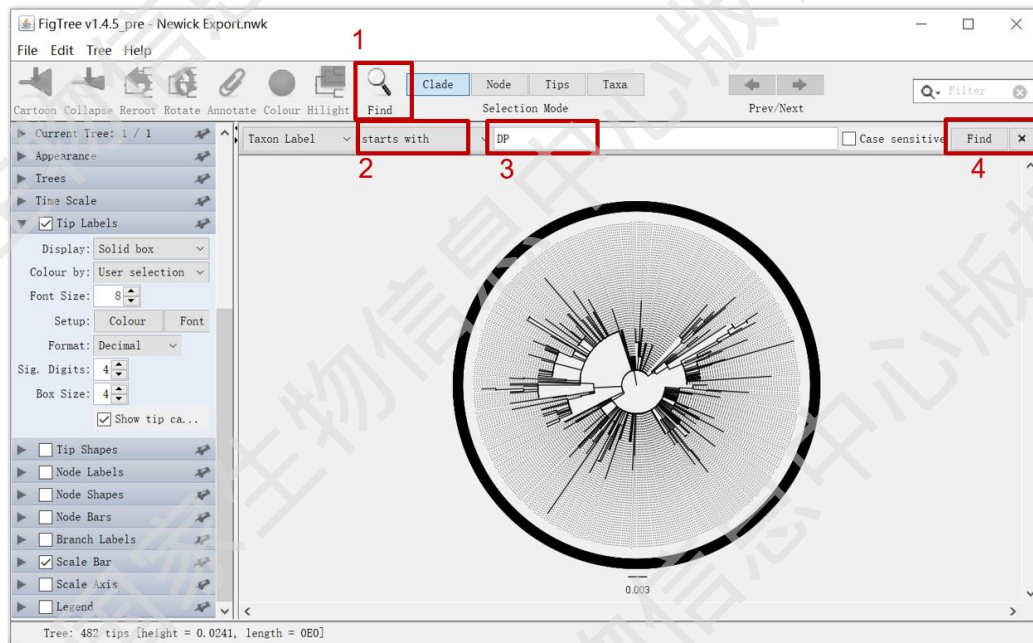
实操练习1：基于MEGA软件进行病原溯源

- 给钻石公主号样本着色：点击Find，选择starts with，输入DP，点击Find



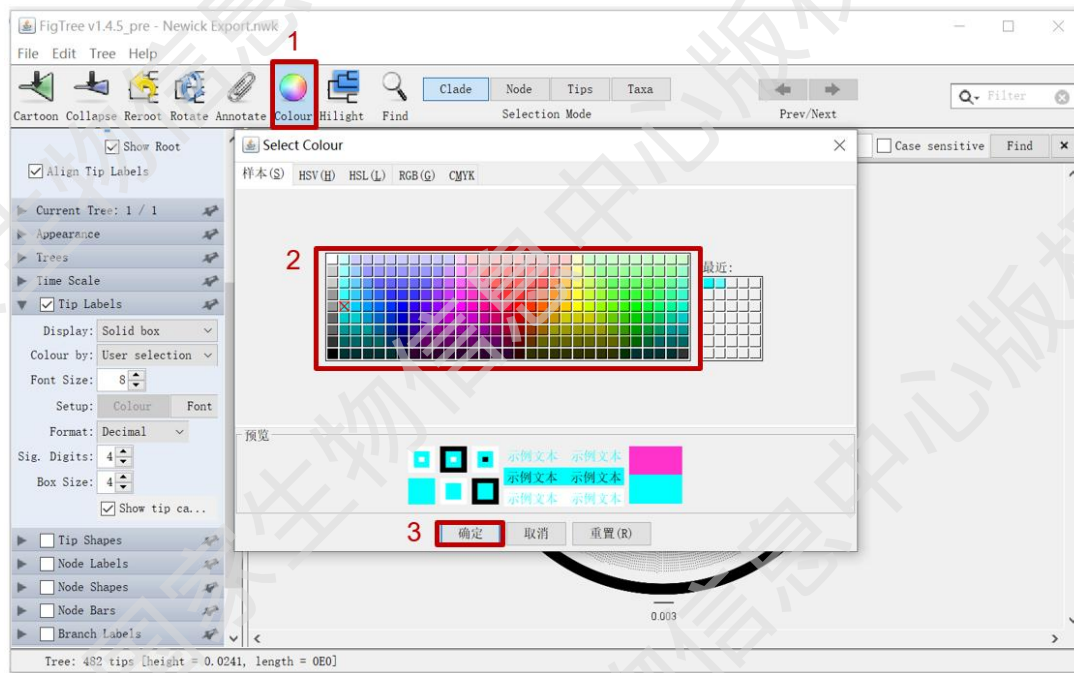
实操练习1：基于MEGA软件进行病原溯源

- 给钻石公主号样本着色：点击Find，选择starts with，输入DP，点击Find



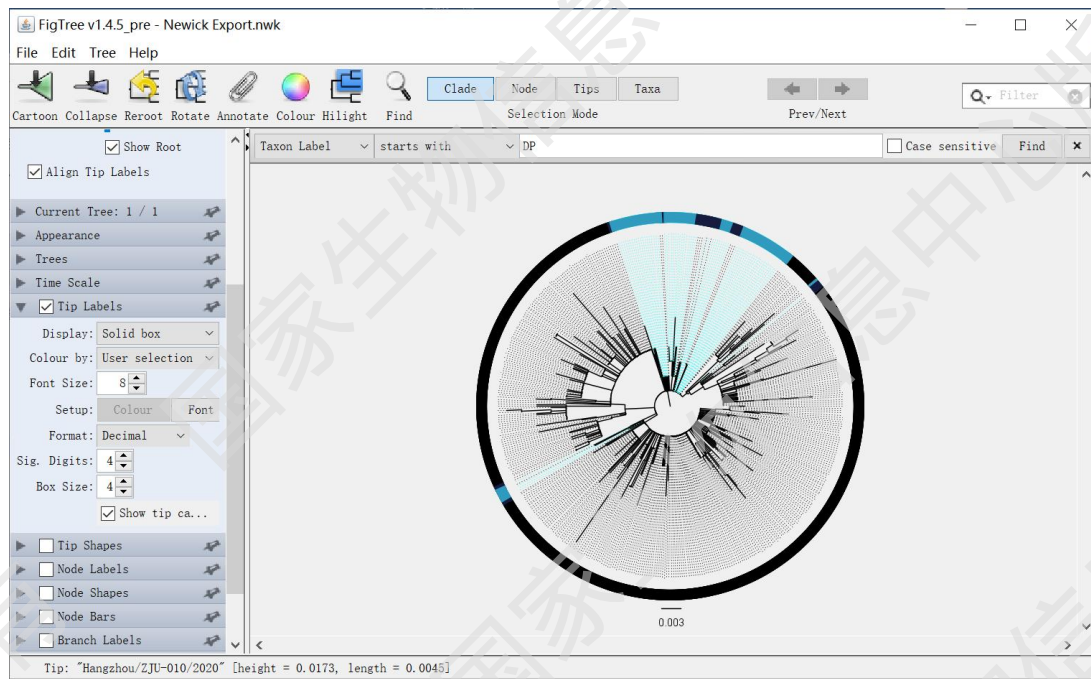
实操练习1：基于MEGA软件进行病原溯源

- 给钻石公主号样本着色：点击Colour，选择合适的颜色，点击确定

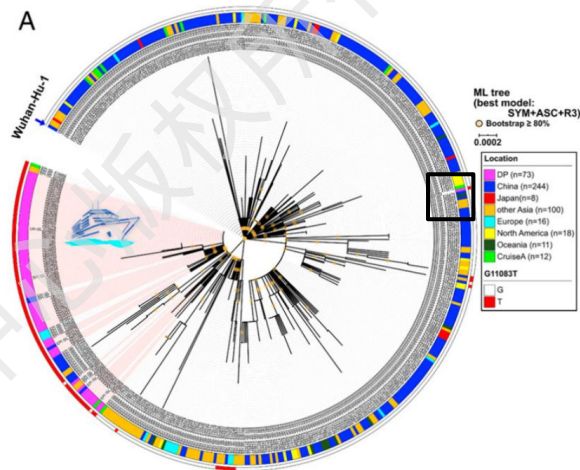


实操练习1：基于MEGA软件进行病原溯源

- 钻石公主号样本被着色

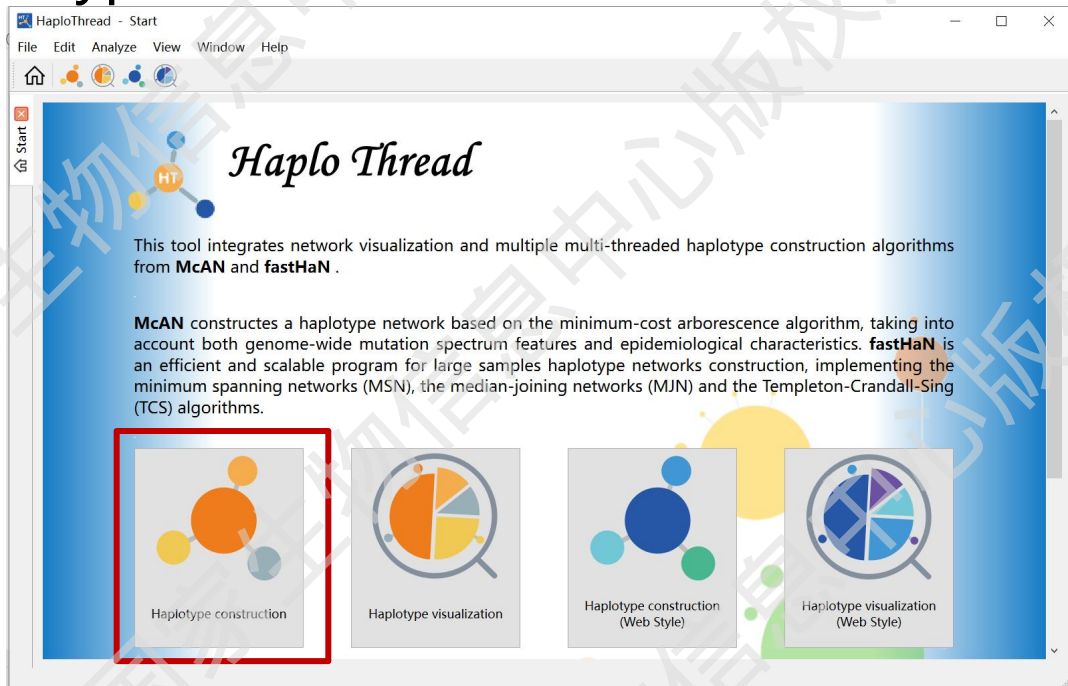


论文中的系统发育树

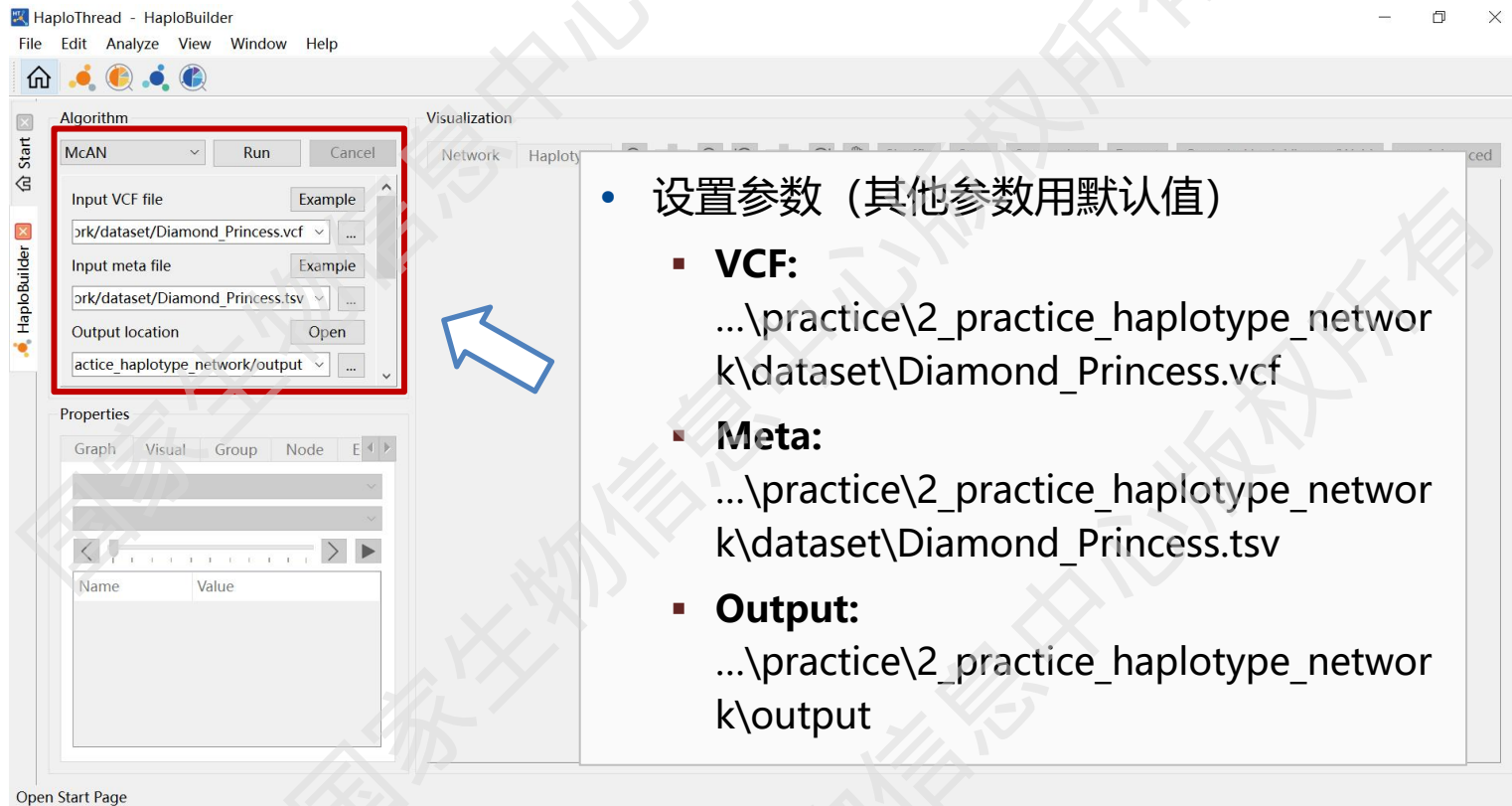


实操练习2: HaploThread构建单体型网络

- 点击Haplotype construction



实操练习2: HaploThread构建单体型网络



Algorithm: McAN

Input VCF file: ork/dataset/Diamond_Princess.vcf

Input meta file: ork/dataset/Diamond_Princess.tsv

Output location: actice_haplotype_network/output

Run

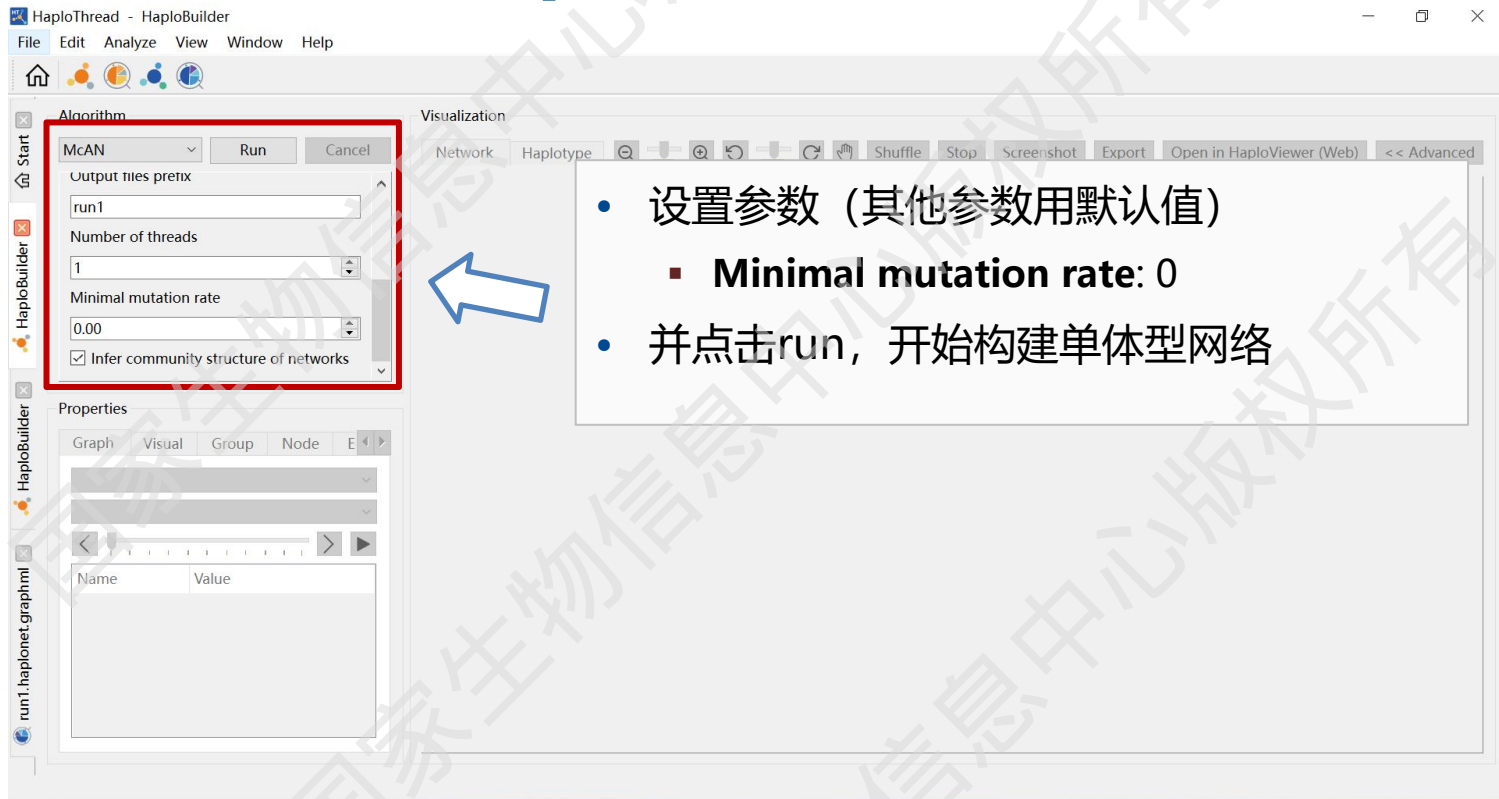
Properties

Name	Value
------	-------

- 设置参数 (其他参数用默认值)
 - VCF:**
...\practice\2_practice_haplotype_network\dataset\Diamond_Princess.vcf
 - Meta:**
...\practice\2_practice_haplotype_network\dataset\Diamond_Princess.tsv
 - Output:**
...\practice\2_practice_haplotype_network\output



实操练习2: HaploThread构建单体型网络



Algorithm

McAN

Run

Cancel

Output files prefix

run1

Number of threads

1

Minimal mutation rate

0.00

☒ Infer community structure of networks

Visualization

Network

Haplotype

Shuffle

Stop

Screenshot

Export

Open in HaploViewer (Web)

<< Advanced

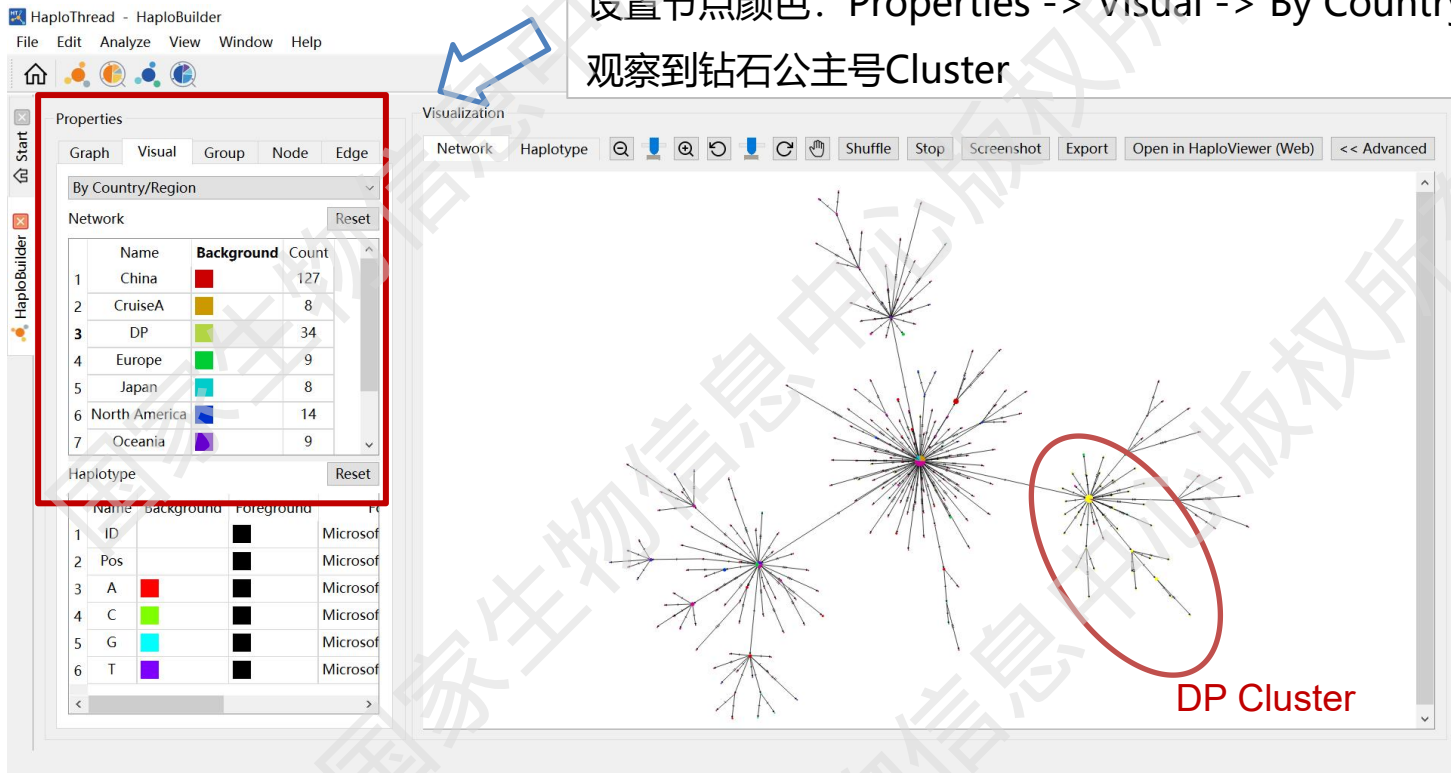
- 设置参数 (其他参数用默认值)
 - **Minimal mutation rate: 0**
- 并点击run, 开始构建单体型网络



实操练习2: HaploThread构建单体型网络

设置节点颜色: Properties -> Visual -> By Country/Region

观察到钻石公主号Cluster



实操练习3：使用基因组溯源工具对新冠病毒进行溯源

1. 病毒数据：practice\3_practice_sequence_similarity\query_sample_mutation.txt
2. 打开在线平台：<https://ngdc.cncb.ac.cn/ncov/online/tool/genome-tracing>
3. **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
4. 方法：WM
5. 点击“Find the closest sequences”：运行程序

Query sequence

Mutation [example](#)

Fasta **Method**

 Please ensure that you adhere to the mutation format specified as seq_ID:POS1Allel1/Allele2,POS2Allel1/Allele2, or alternatively, you can use the fasta format sequence as input on the webserver.



实操练习3：使用基因组溯源工具对新冠病毒进行溯源

Query sequence

Mutation example

Fasta Choose File **Method** Find the closest sequences reset

Please ensure that you adhere to the mutation format specified as seq_ID:POS1Allel1/Allele2.POS2Allel1/Allele2, or alternatively, you can use the fasta format sequence as input on the webserver.

✱ 00:00:01

Expected running time: ① UM: < 2 min; ② WM: < 2 min; ③ WM+CL: < 4 min; ④ WM+LD+CL: < 4 min. Please note that the actual running time may exceed the expected time due to heavy task burden. If the processing time is longer than expected, you can stop the task by **refreshing** or **closing the website** and resubmit the task later or use the standalone version. All scripts could be found at <https://github.com/tpplol/GenomeTracing>

任务开始运行后会
计时

任务运行完成后会自动
输出最相似的50个
样本

得分代表样本间的相
似性

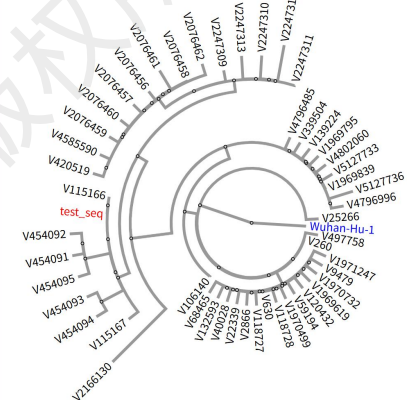
最相似的50个样本
构建系统发育树

Matched sequences ②

Showing 1 to 50 of 50 entries

SeqID	Accession	Score	Lineage	Mutation
V115166	EPI_ISL_498692	1	8.1.1.317	0
V115167	EPI_ISL_498693	0.966475	8.1.1.317	1
V2247311	EPI_ISL_9301315	0.928219	8.1.1.317	2
V454094	EPI_ISL_2170888	0.915172	8.1.1.317	3
V454093	EPI_ISL_2170891	0.914392	8.1.1.317	3
V454095	EPI_ISL_2170889	0.912734	8.1.1.317	3
V454092	EPI_ISL_2170890	0.912166	8.1.1.317	3
V454091	EPI_ISL_2170887	0.911903	8.1.1.317	3
V2076458	OM098414.1	0.897627	8.1.1.317	3
V2247309	EPI_ISL_9301319	0.89712	8.1.1.317	3
V2076459	OM098416.1	0.87038	8.1.1.317	4
V420519	EPI_ISL_2131805	0.870287	8.1.1.317	4
V2076462	OM098419.1	0.87027	8.1.1.317	4
V2076456	OM098425.1	0.870242	8.1.1.317	4
V4585590	OG586441.1	0.870241	8.1.1.317	4
V2076460	OM098417.1	0.870227	8.1.1.317	4

Phylogeny of matched sequences ③



实操练习4：使用McAN在线平台追踪钻石公主号疫情

1. 准备DP数据：
practice\4_practice_haplotype_network_online\dataset
2. 打开在线平台：
<https://ngdc.cncb.ac.cn/bit/hapnet/>
3. 物种类型：其他物种
4. 文件格式：变异文件格式
5. 变异文件：mutation.txt
6. 元信息文件：meta.txt
7. 最小突变率：0
8. 点击运行：跳转到下一个页面

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或者其它物种变异文件(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。

选择物种类型

☐ 新冠病毒

☒ 其它物种

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

Choose File

No file chosen

元信息文件 ? 必选项

[元信息文件示例](#)

上传元信息文件

Choose File

No file chosen

构建单倍型网络图参数设置 可选项

最小突变率 (为了选择高突变率位点)

0

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 Title

运行



实操练习4：使用McAN在线平台追踪钻石公主号疫情

🏠 首页 > 📁 单体型分析 > 📊 结果

📁 单体型分析

任务信息

作业号	2024041525062
提交日期	2024-04-15
运行状态	Finished
选择物种类型	Other species
选择文件格式	Mutation file
最小突变率	0

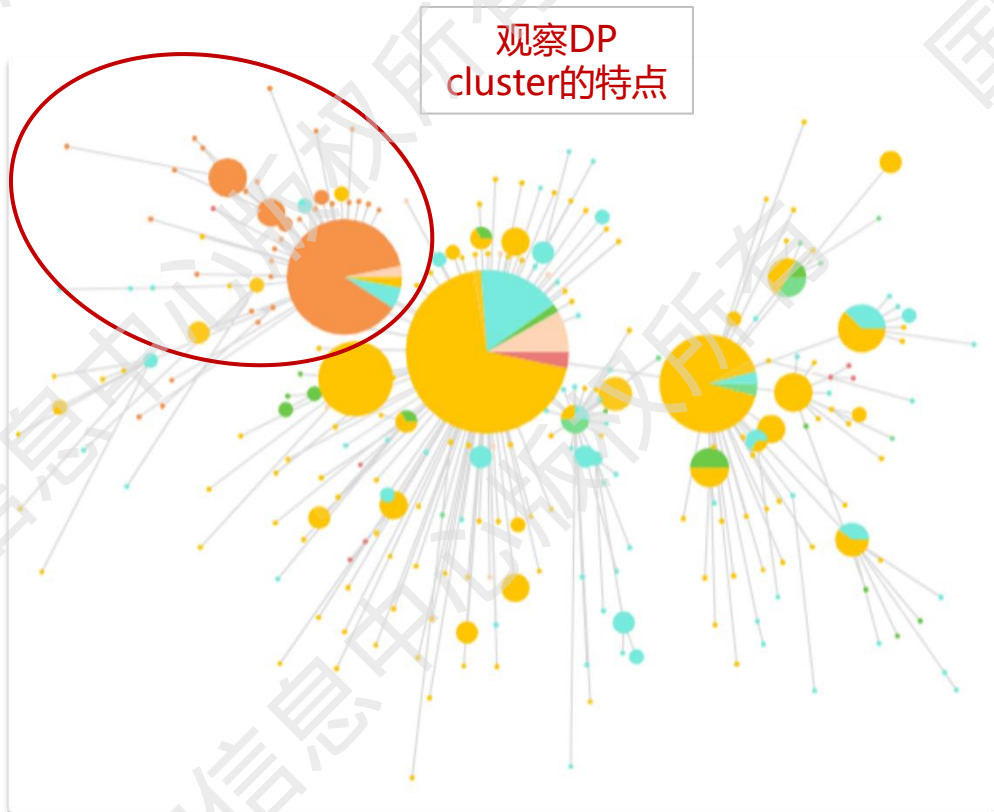
下载结果文件

Download 

在线查看单倍型网络图

[View online](#)

任务完成后，点击
View online，查
看单体型网络



致谢

感谢CNCB全体老师同学



感谢课题组全体老师同学





THANKS

1918流感溯源

- 地理起源存在争议
- 假说
 - 美国起源说
 - 欧洲起源说
 - 亚洲起源说

