ABC

Applied Bioinformatics Course

Molecular Phylogeny and Phylogeny Tree Construction

分子系统发生学及系统发生树的构建

汇报人：孙田舒

sunts521@gmail.com

Self-introduction

2012-2018

生命科学学院
School of Life Sciences

- 顾红雅 教授

2018-2020

中国农业科学院
CHINESE ACADEMY OF AGRICULTURAL SCIENCES

- 黄三文 教授

2020-

UNIVERSITY OF CAMBRIDGE

- Julian Hibberd 教授

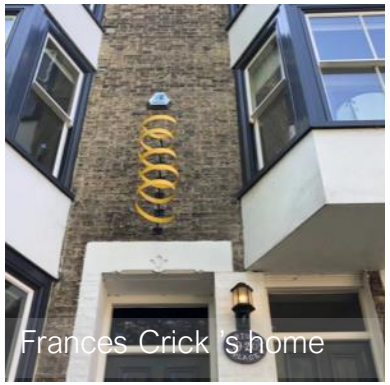
Cambridge

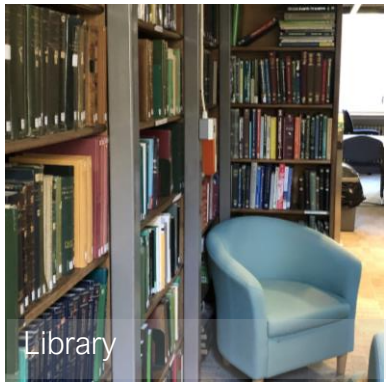
Study at Cambridge

History

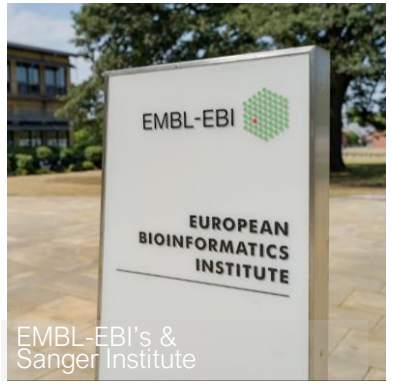

121 Nobel laureates

DNA helix announced


Frances Crick 's home


Library

Innovation



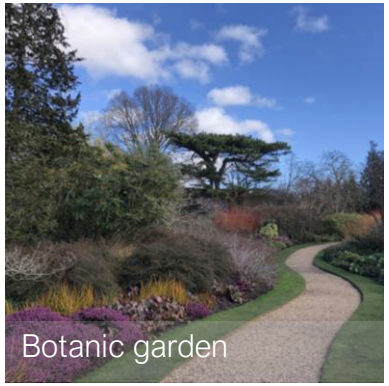
EMBL-EBI's & Sanger Institute


The Babraham institute

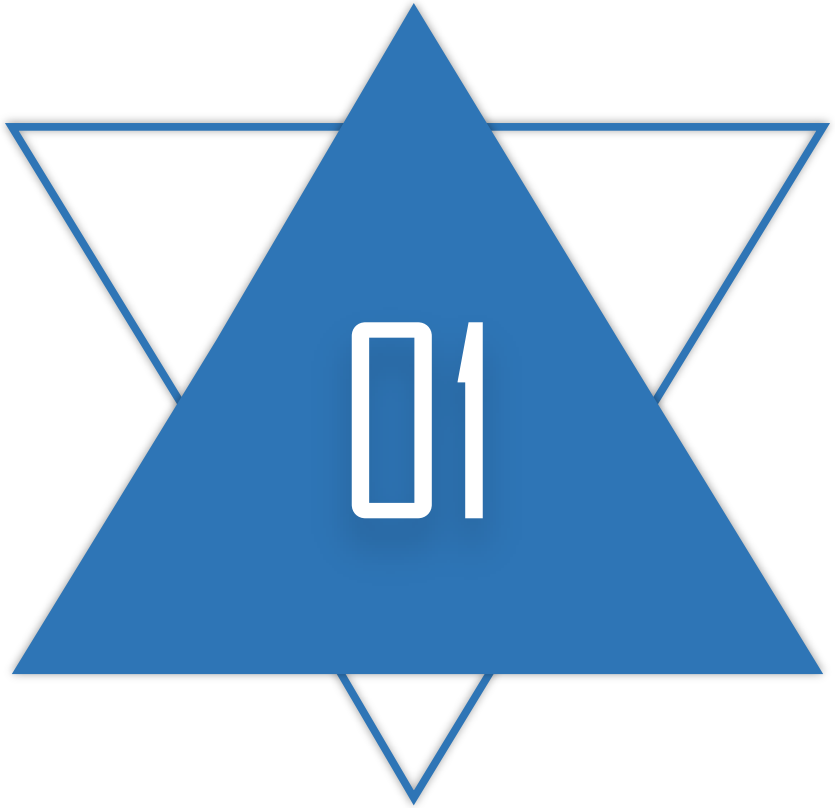
Science park

Leisure


Cam river


Botanic garden

目录
CONTENTS



系统发生学的概念

What is phylogeny?

系统发生学的概念

Phylogeny 系统发生

系统发生是指任何生物实体（基因、个体、种群、物种和种上阶元）的起源和演化关系

Molecular phylogeny 分子系统发生

是利用各种分子性状构建的生物实体之间起源和演化关系，采用分子数据主要是DNA和蛋白质序列，也包括其它类型的分子数据

Phylogenetics 系统发生学

是研究利用各种性状构建基因、个体、种群、物种和种上阶元之间系树树和网络的原理和方法的学科。系统发生学重建演化历史依赖于对取样物种的性状分布进行数学推论，这种重建涉及不同类群共享的同源性状，并以此推断系统树。这种数学推断的准确性完全依赖于对性状演化的假设和模型。

系统发生关系的含义

表征关系

Phenetic

不考虑进化关系，仅以所有可利用的性状为基础的全面相似性程度排列关系。以表表征关系为基础的分类学研究即表征分类学

分支关系

Cladistic

分支关系指物种或种群之间与共同祖先相对近度的关系，以此为基础的系统学研究即支序系统学

遗传关系

Patristic

遗传关系是生物在遗传组成方面的关系，在群体遗传学中采用遗传相关性系数来度量，在种上阶元采用亲缘距离来度量。一次为基础的系统学即分子系统学

生物间的关系

生物之间在系统发生学上的相关性称为系统发生关系。生物之间存在着以下各种的复杂关系，这些关系从不同角度反映了生物之间的相关性。这些关系都可以用分支图解来表示，即树状图（dendrogram）。

系统发生关系

Phyletic

广义的系统发生或种系关系包括任何生物实体的起源和演化关系。系统发生关系一般无法重建完整的生物演化历史，是一种对历史的假设，可以随着研究的深入而无限逼近真实的演化历史

年代关系

Chronistic

年代或时序关系在演化时间标度上指示有机体之间的关系，即系统发生树垂直轴上有机体之间的关系，这种关系对研究演化速率和趋势有一定的意义

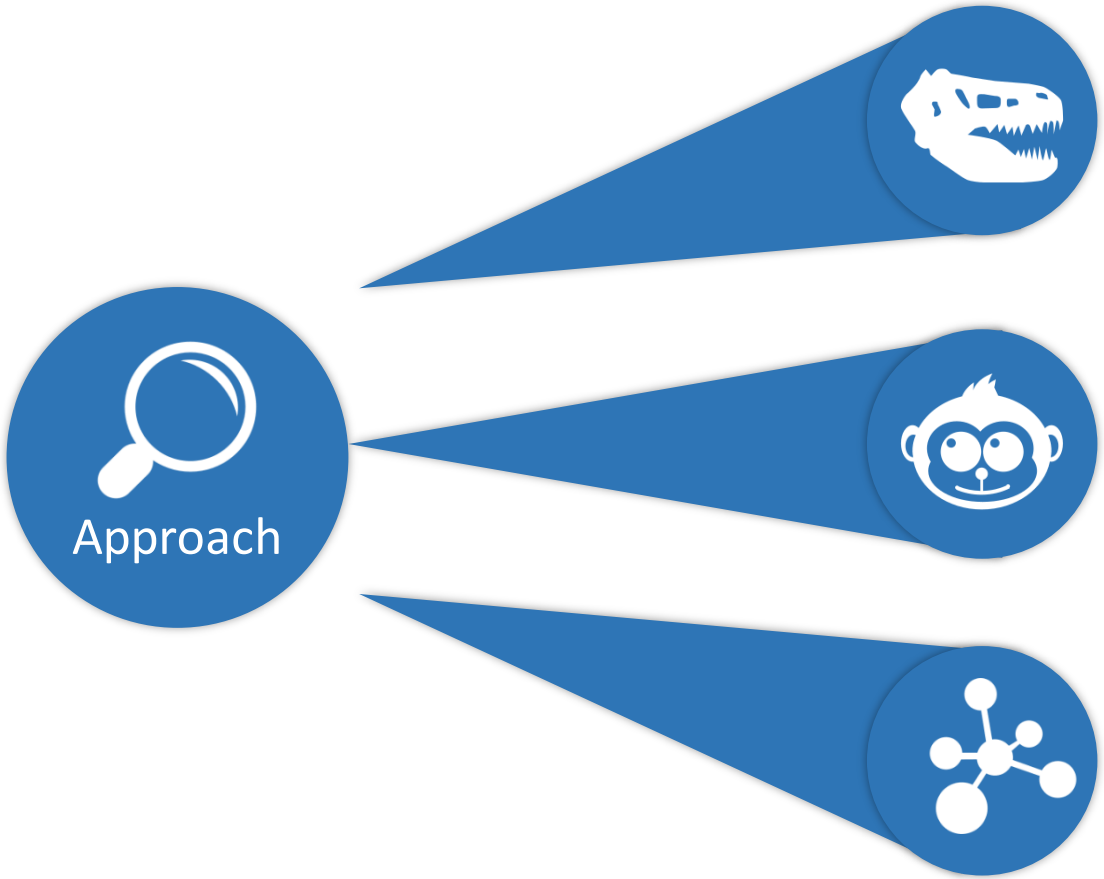
地理分布关系

Chorologic

地理分布关系是指生物在演化过程中形成的空间分布关系。

系统发生研究、推论或重建都是指通过现存或化石生物性状的比较分析来建立生物类群的系统发生关系。

如何研究系统发生



化石证据 Fossil records

- Available for certain species
- Limited by abundance, habitat
- Fragmentary and ambiguous

表型证据 Comparative morphology and comparative physiology

- Species with distant phylogenetic relationships may have similar morphology or physiology phenotype
- Difficult to find comparable phenotype, especially for distant species

分子证据 Molecular approach: DNA or Protein sequences

- Since 1980s

利用分子数据研究系统发生

分子数据的优点

- 可以识别的同源性质**范围广泛**，核糖体蛋白质，rRNA等可以进行最宽范围的系统发生重建
- 可以提供**大量性状**，如人类基因组可以提供32亿个核苷酸位点
- 数据**不依赖于化石记录**（活体或灭绝）
- 分子数据容易处理，可以**客观定量**和可检验地处理并确定建树方法；分子形状的演化机制和模型研究深入，可以借助这些模型设计有效避免系统误差、比较可靠的统计学建树方法
- 不同分子性状演化速率变化大，适用于**不同分类阶元**的系统发生分析
- 获得数据**成本低廉**，测定一个全基因组的价格和时间可以被接受

20世纪80年代以来，分子数据迅速取代形态学特征，成为重建系统发生关系的基本数据。分子系统学的大量研究没有排斥形态学数据建立的系统发生关系，而是将**分子与形态结合起来**，从而更好的对生物多样性进行描述和解释。

分子系统发生关系的假设

对分子数据的通常假设

系统发生分析方法及其结果的准确性与演化过程的模型密切相关，不同点方法有不同的假设，这些假设能够保证方法的有效性

同源性

分子序列是同源的

Molecular sequences are homologous

代表性

分子序列的差异能够代表分歧时间

The differences among molecular sequences are in relationship with divergence time

随机性和独立性

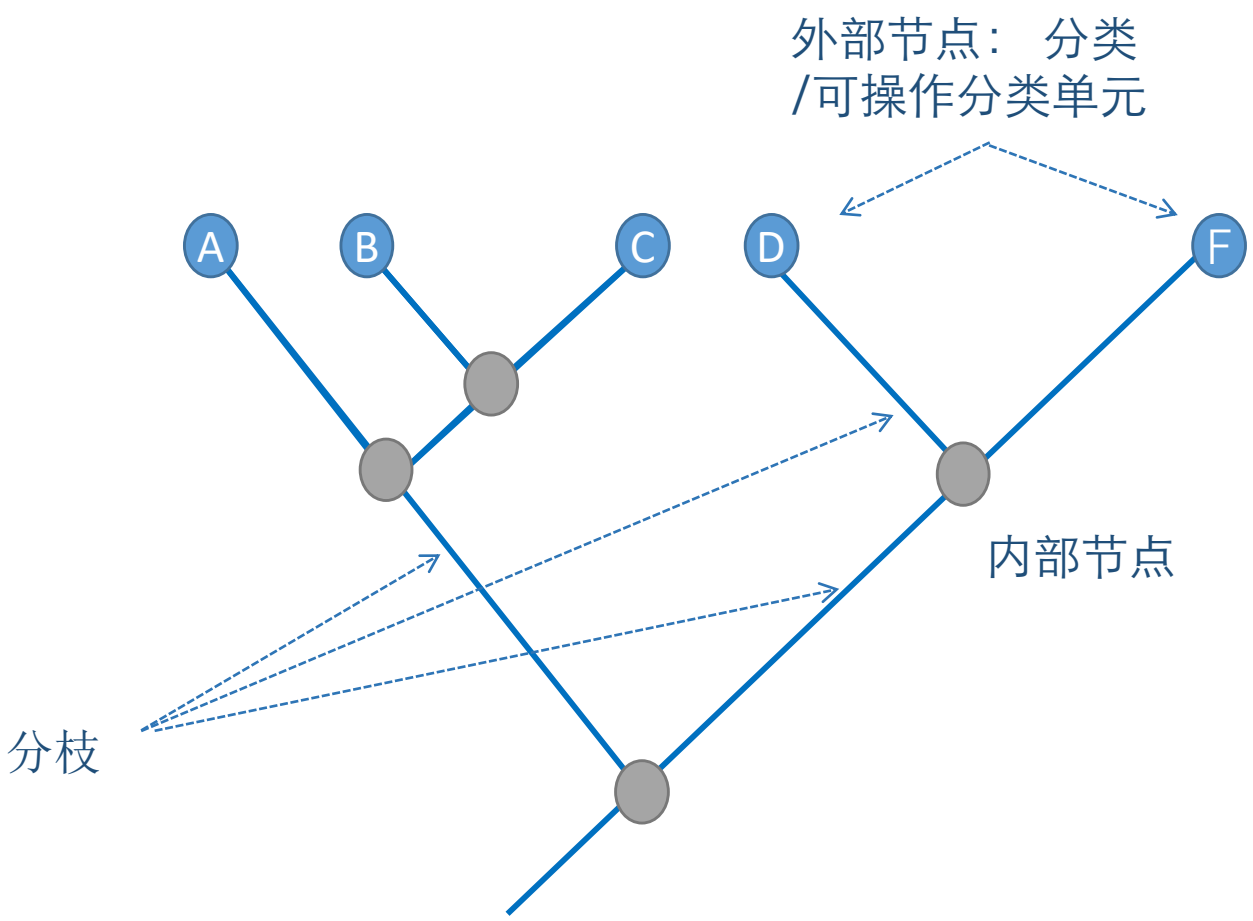
序列中的位点是的演化是随机和独立的

Each position in a sequence evolved randomly and independently

系统发生树的基本概念

Basic concepts of phylogenetic tree

系统树的要素



- 拓扑结构 topology（分支形式branch pattern）

指分类单元在系统树上的分支情况、分支排列的相对位置。所有的分支分类单元真正的分支形式只有一种，即这些分类单元的演化历史。系统发生分析的目的是在所有可能的分支形式中找到最符合观测数据的一种拓扑结构。

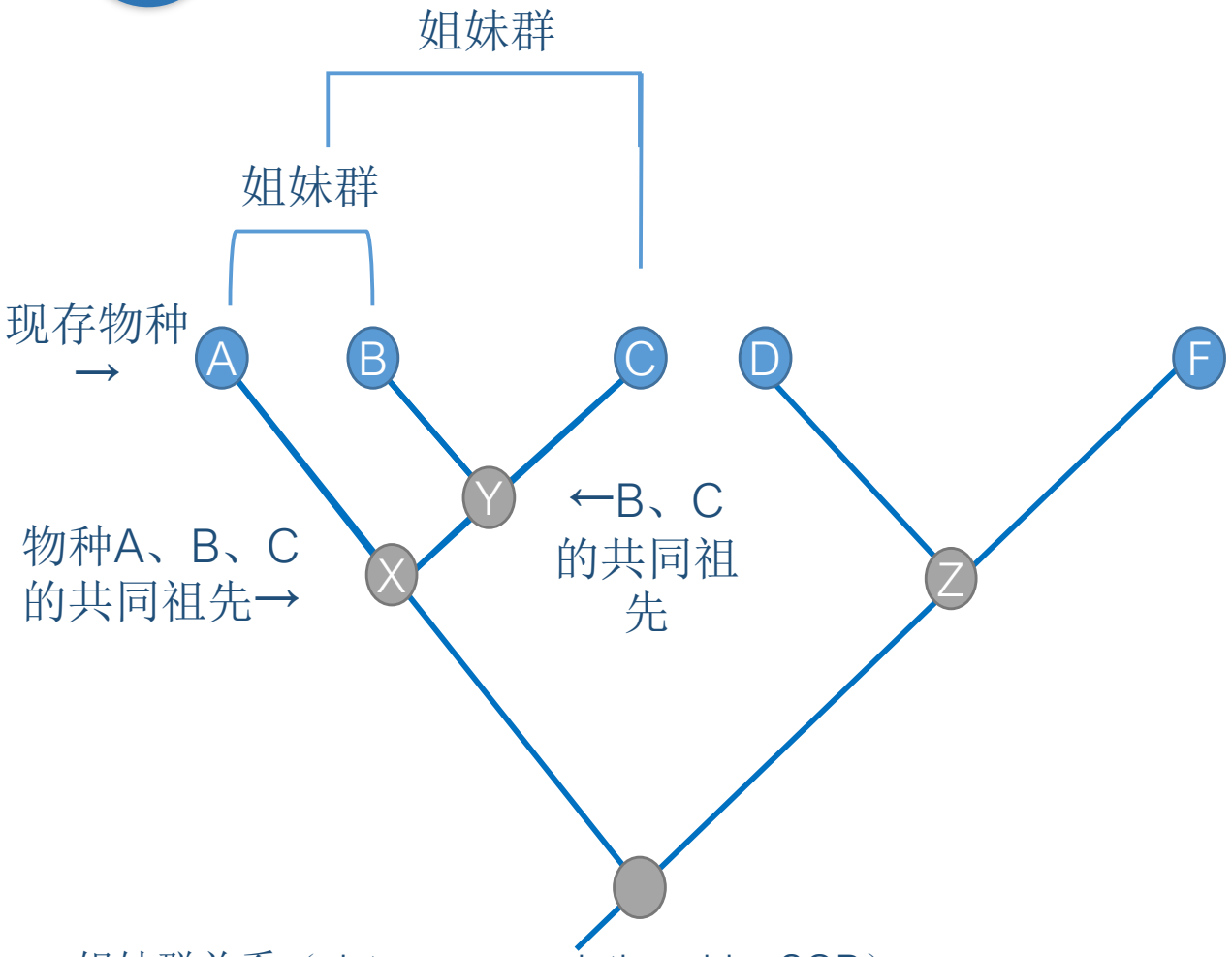
- 节点（node）

系统树末端节点成为操作分类单元（OTU），一般是现存生物，可以是基因、个体、种群等。物种树内部节点是现存祖先，而群体系统树和基因树内部节点可以是假象祖先也可以是现存的个体或等位基因。

- 分枝（branch）

一个分枝只连接相邻的两个节点。标度系统树每个分枝都有分枝长度表示该分枝的演化改变数量，即分枝长度是时间与演化速率的乘积。

系统树表达的信息



- 姐妹群关系（sister-group relation ship, SGR）

当两个分类单元共同拥有一个不为第三者所有的祖先时，此二者称为一个姐妹群。SGR是系统发生的核心关系，SGR即分支图（cladogram）。

- 祖裔关系（ancestor-descendant relationship, ADR）

系统树上从根到末端节点的传承关系，末端节点就是现存的分类单元，内部节点可以是假象祖先，也可以是化石祖先。

- 相对祖先的近度关系（relative recency of common ancestor）

反映分类单元的血缘关系（geneological relationship）的最确切方法为共同祖先的相对近度，即对于任何第三个类群，如果其中两个类群拥有比其它类群更近的祖先的话，则这两个类群就拥有比另一个类群更近的亲缘关系。

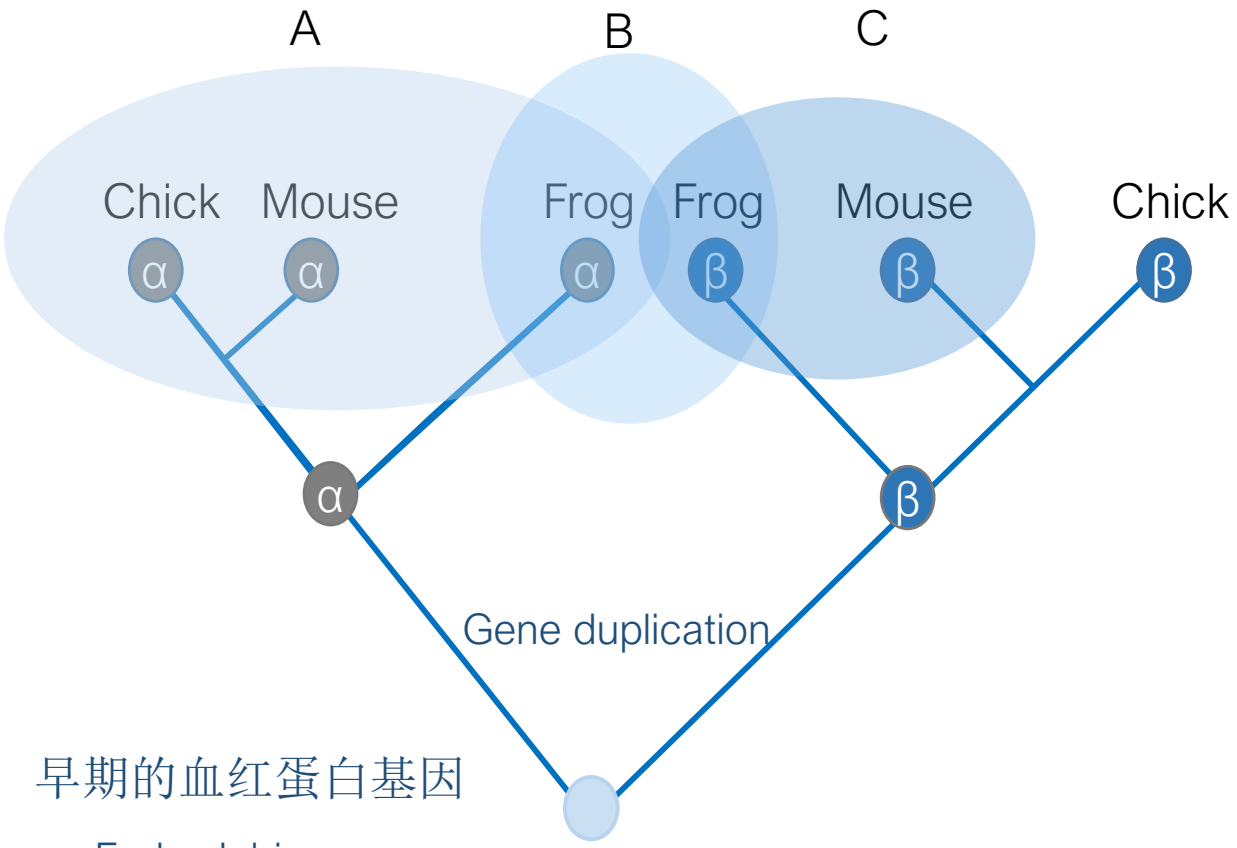
- 相对演化速率

当系统树每个节点的分歧时间已知时，可将系统树各个分枝的演化速率计算出来，以速率图（ratogram）方式显示。

- 分歧年代关系

借助部分内部节点的化石时间矫正信息，可以采用分子测年的方法估计出系统树各个分枝节点的分歧时间，以时序图（chronogram）的方式显示

单系类群、旁系类群和多系类群



早期的血红蛋白基因

Early globin gene

- 单系类群（monophyly）

是拥有一个共同祖先、且这一个共同祖先的后裔全部包括在内的所有分类单元组成的类群，如A。

A monophyletic group is a taxon which forms a clade, meaning that it consists of an ancestral species and all its descendants.

- 并系类群（paraphyly ）

具有一个共同祖先但不包括所有后裔的类群，即并系类群包含了共同祖先的部分（而非全部）后代，如C

A paraphyletic group consists of all of the descendants of a common ancestor minus one or more monophyletic groups. A paraphyletic group is thus 'nearly' monophyletic

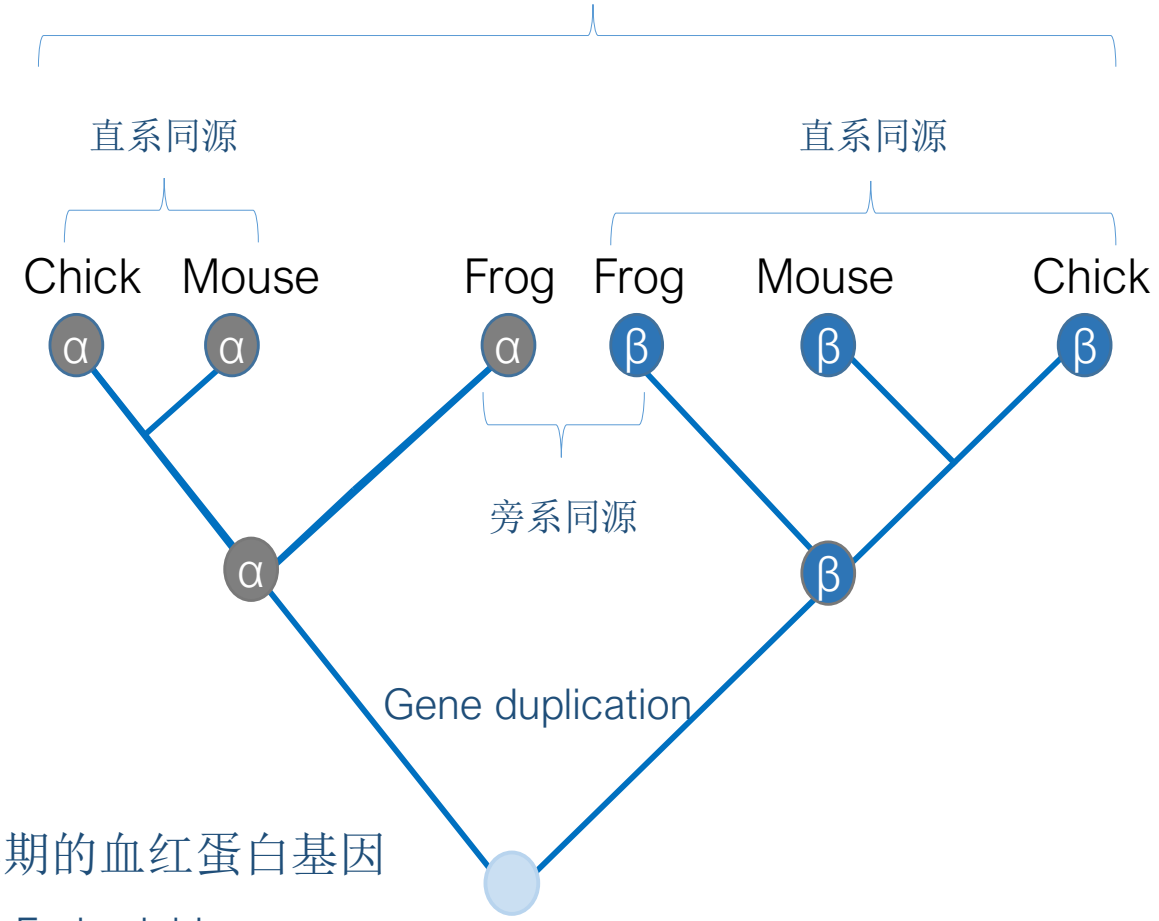
- 多系类群（polyphyly）

一个分类群当中的成员，在系统树上分别位于相隔着其它分枝的分支上，即该类群不包括所有成员的最近共同祖先，如B。

A polyphyletic group is characterized by convergent features or habits ; the features by which the group is differentiated from others are not inherited from a common ancestor.

直系同源和旁系同源

在生物学种系发生理论中，若两个或多个结构具有相同的祖先，则称它们同源（Homology）。



- 直系同源（orthology）

直系同源的序列因物种形成（speciation）而被区分开（separated）：若一个基因原先存在于某个物种，而该物种分化为了两个物种，那么新物种中的基因是直系同源的；

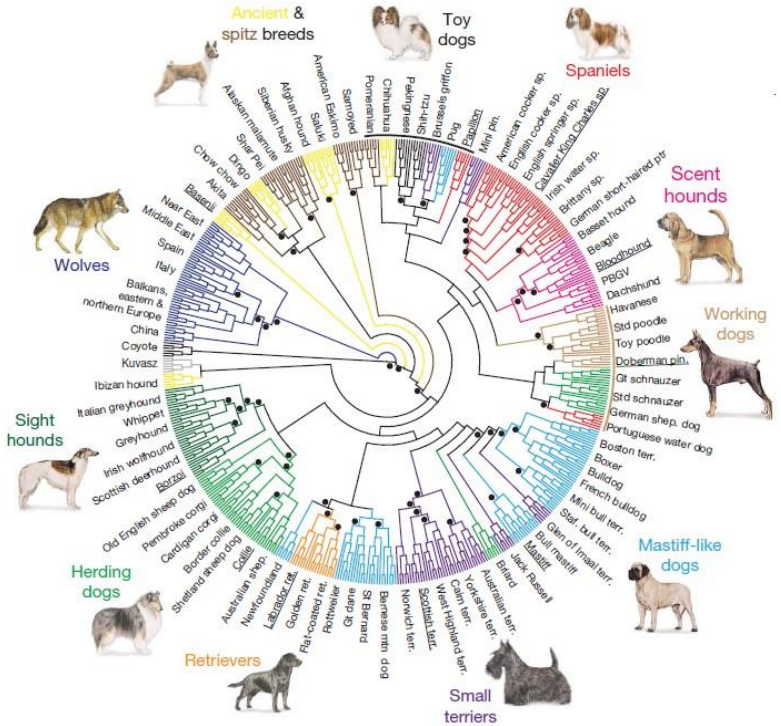
Homologous sequences are orthologous if they are inferred to be descended from the same ancestral sequence separated by a speciation event: when a species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologous.

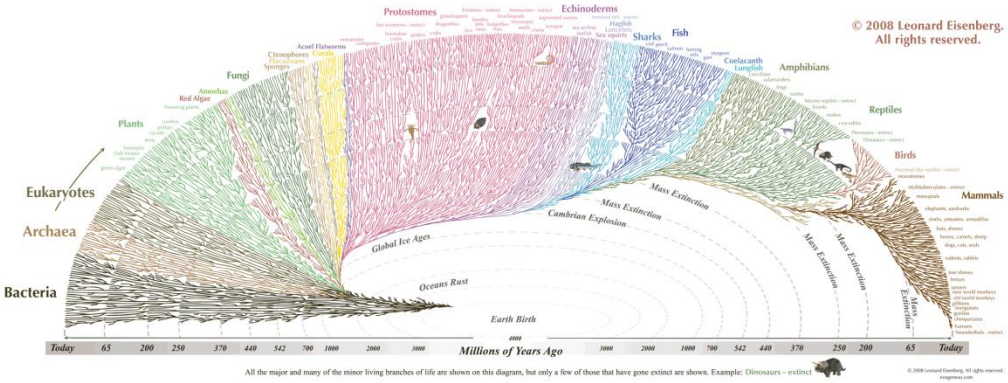
- 旁系同源（paralogy）

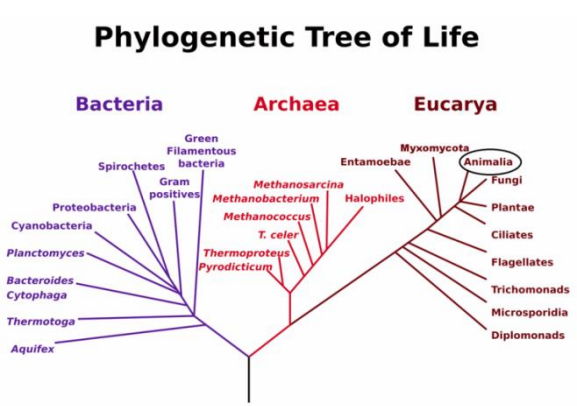
旁系同源的序列因基因复制（gene duplication）而被区分开（separated）：若生物体中的某个基因被复制了，那么两个副本序列就是旁系同源的。

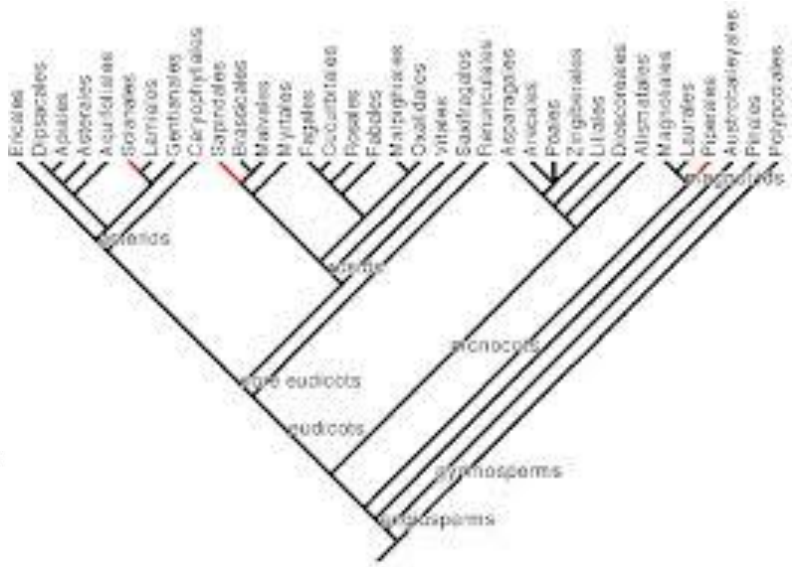
Homologous sequences are paralogous if they were created by a duplication event within the genome. If this was a gene duplication event: if a gene in an organism is duplicated to occupy two different positions in the same genome, then the two copies are paralogous.

系统树的表现形式









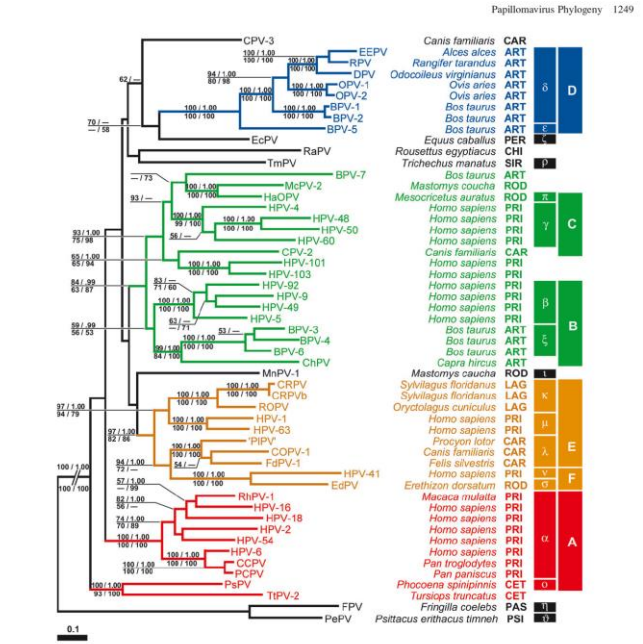
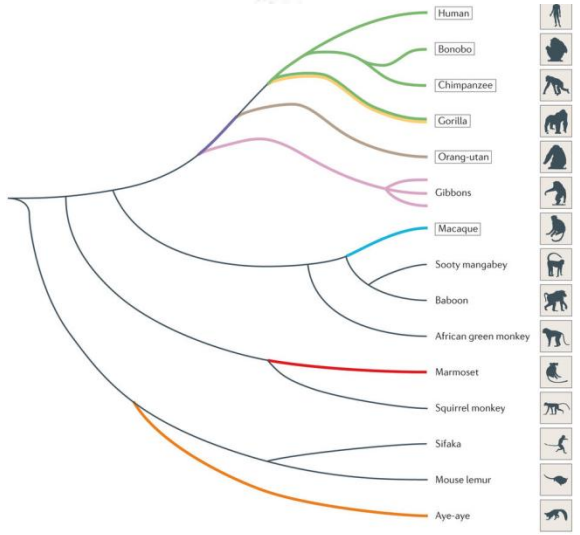


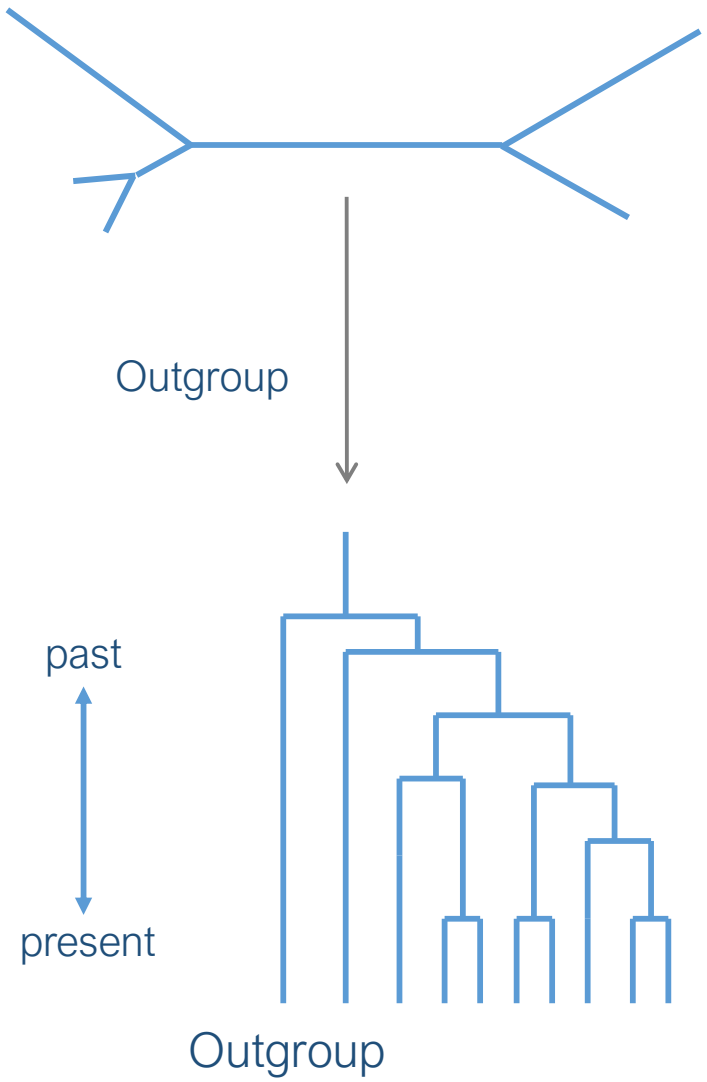
Fig. 2.—ML tree of 53 phylogenetically representative PVes as inferred from a combined E1–E2–L1 amino acid sequence analysis (1,082 parsimony-informative positions) justified by PHTs (table 2). All non-human PVes and 18 representative HPV types were used for analyses. PV genera (de Villiers et al. 2004) are indicated by Greek lettering, upper case lettering follow an alternative E1–E2 classification of Bravo and Alonso (2007). Higher order host taxa are abbreviated as follows: ART, Artiodactyla; CAR, Carnivora; CET, Cetacea; CHI, Chiroptera; LAG, Lagomorpha; PAS, Passeriformes; PER, Perissodactyla; PRI, Primates; PSI, Psittaciformes; ROD, Rodentia; and SIR, Sirenia. The supertaxa are colored blue (δ+ ε), ocher (κ+ μ+λ+ ν+ σ), green (π+ γ+ β+ ξ), and red (θ+ ο), respectively. Branch lengths are drawn to scale, with the scale bar indicating the number of amino acid substitutions per site. Numbers on branches are bootstrap support values to clusters on the right of them (above: criteria = ML/Bayesian probabilities; below: criteria = MP/distance; values under 50 are not shown).



Nature Reviews | Genetics

Different in the way it's presented!

系统树的类型:有根树 VS. 无根树



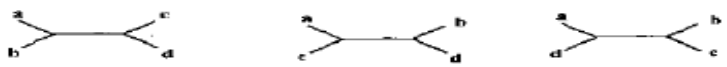
- 无根树（unrooted tree）

一棵系统树如果不指定最初的共同祖先即无根树，无根树没有方向，它仅知名OUT之间的相对分支关系，而并未指明演化路线，从而任何一个节点都有三种可能的演化方向，因而不能定义单系类群。

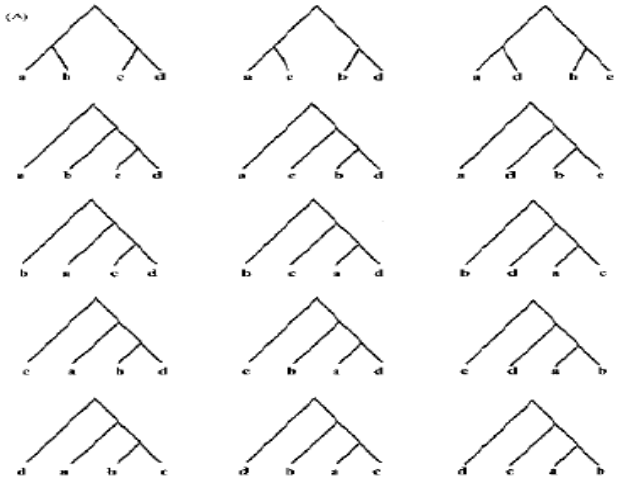
- 有根树（rooted tree）

如果系统树中一个节点代表了在时间上早于其他所有节点，则该节点称为系统树的根，这样的系统树称为有根树，有根树是有方向的。有根树的根向外定义了时间、分支的先后次序和各级分支的共同祖先。有根树的根是系统树上所有的OUT的共同祖先，二歧有根树每一个节点有两个演化方向。

- 对于一定数目的OUT，可以构建许多可能的具有不同分支型式的有根树和无根树



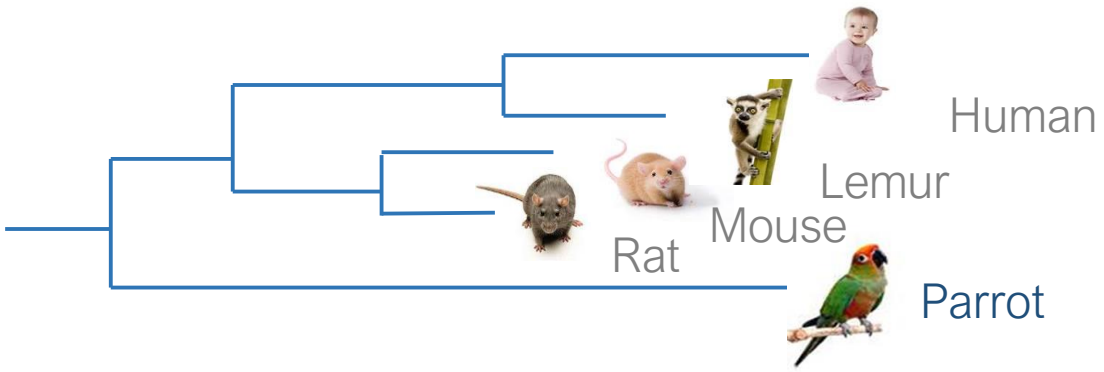
$N_U=(2n-5)!/^{2n-3}(n-3)!$



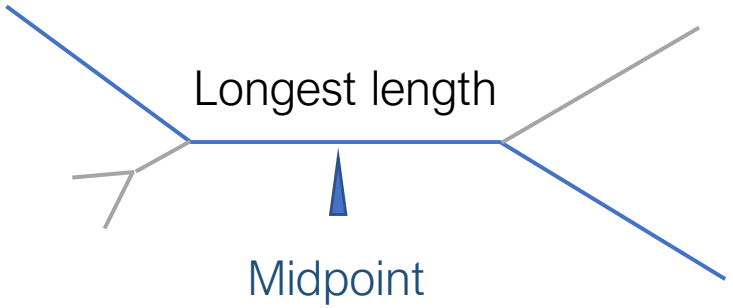
$N_R=(2n-3)!/^{2n-2}(n-2)!$

Species	Unrooted	Rooted
4	3	15
10	2027025	34459425
50	2.84×10^{74}	2.75×10^{78}

系统树的赋根方法



Outgroup is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time.



- 外类群赋根（outgroup rooting）

内群（ingroup）:所感兴趣而深入研究的分类单元

外类群（outgroup）:外类群是内群的同源序列，但与内群相比，演化上与母群分离的更早。

外类群赋根是目前应用最广也是最有效的赋根方法，其方法的质量取决于选择的外类群是否合适。实践中可以用多个外类群，也可以先分析建立了稳定的内群系统时后加入外类群。

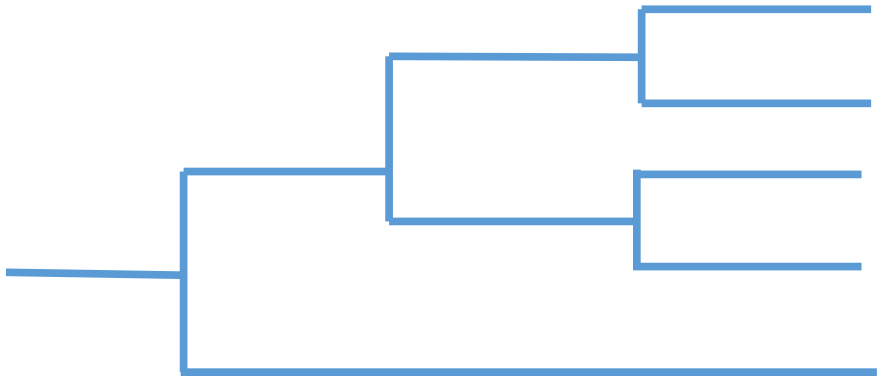
- 分子中赋根法/中点赋根法（midpoint rooting）

是将系统树上通径最长的两个OUT中点作为该树的根。

按照分子钟的原理，序列在不同时间和不同支系的演化速率相同，则系统树最长分支意味着分歧年代最早，该分支中点是这一分支两端所有支系的根。

其它方法：贝叶斯赋根法、内群限制树赋根法、并源基因赋根法等。

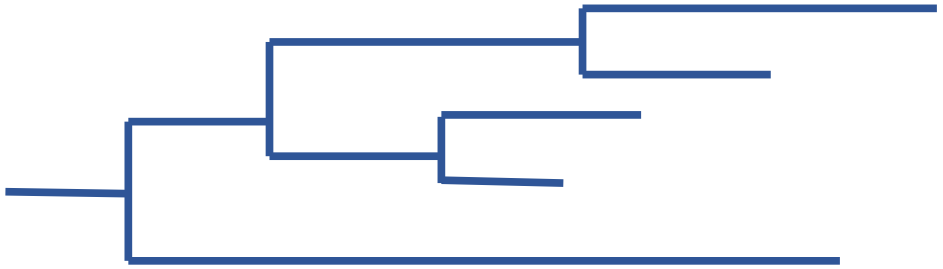
系统树的类型:标度树 vs. 未标度树



- 未标度树（unscaled tree）

各分支的长度不表示性状状态变异的量，但在有根树上节点的位置仍可与时间相对应

All branches in the tree are the same length. (only topology)



- 标度树（scaled tree）

系统树上各分支的长度代表了性状状态变异的量

Branches will be different lengths based on the number of evolutionary changes or distance. (branch length & topology)

系统树的类型:基因树 vs. 物种树

- 基因树（gene tree）

是根据DNA或蛋白质序列构建的系统树

- 物种树（species tree）

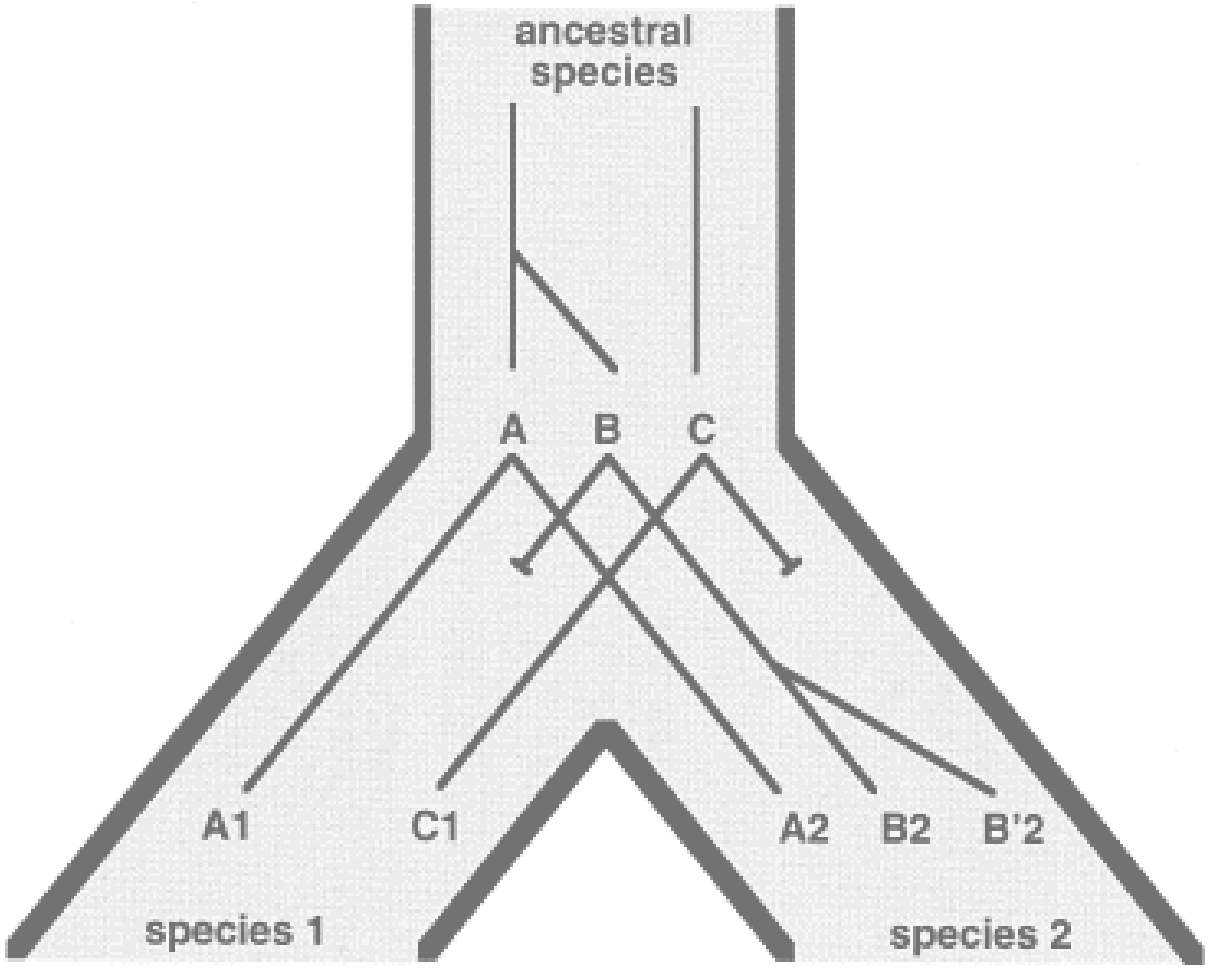
是表达生物类群演化路径的系统树

有序列构建物种树的误差：

- 序列的数据结构杂乱或系统误差
- 基因深度溯祖、基因重复、水平转移等

注意：

- 来自不同物种的两个基因分化时间可以早于物种分化时间
- 基因树的拓扑结构可能与物种树不完全一致



系统树的演化解释误区

类群演化程度的高级与低级

- 演化程度的高级与低级（原始与进化）只能用于不同历史时期的生物比较上。
- 无论生物简单还是复杂，拥有共同祖先的现存所有物种都经历了相同的演化时间，在演化上处于同等水平（虽然复杂性程度有差别）。
- 系统树根部不等于原始，支部不等于高等。正确理解是基部分枝拥有更多祖先特征，发生分歧时间较早，特征与祖先更接近。

系统树上排列次序和亲疏程度

- 判断系统树上分类单元关系的密切程度是共享祖先在时间上的远近

同胞与祖先

- 系统树上应该作为同胞（姐妹群）的分支往往被当作祖先，比如现存的单细胞生物和多细胞生物是姐妹群。

系统发生树的构建

How to construct a phylogenetic tree

构建系统发生树的流程

- 选择恰当的分子数据
- Choosing molecular marker

- 多序列比对
- Performing multiple sequence alignment

- 确定建树方法
- Determining a tree building method

- 对系统树的可靠性进行评估
- Assessing tree reliability

选择恰当的分子数据

蛋白序列

- More conserved: 61 codons→20 AAs
- No different evolutionary rates
- No preferential codon usage
- More sensitive alignment
- Gap doesn’t cause frameshift errors

The decision depends on the properties of the sequences and the purposes of the study. Generally, for studying very closed organisms, nucleotide sequences can be used. While if the phylogenetic relationships to be delineated are at the deepest level using protein sequences makes more sense.

核酸序列

- Rapid evolutionary rates can be informative for closely related sequences.
- Depict synonymous and nonsynonymous substitutions, revealing evidence of positive or negative selection.

其它考量

- Domain/Motif Vs. Full length
- Coding Vs. Non-coding
- Genomic Vs. Mitochondrial/ Chloroplastic DNA

多序列比对

序列比对是序列数据分析中最关键的环节，因为系统发生分析是建立在序列位点同源性基础上的。

What can we do

- Choosing substitution models 选择替代模型

Nucleotide: Juke-Cantor Model and Kimura Model

Protein: PAM or JTT amino acid substitution matrix

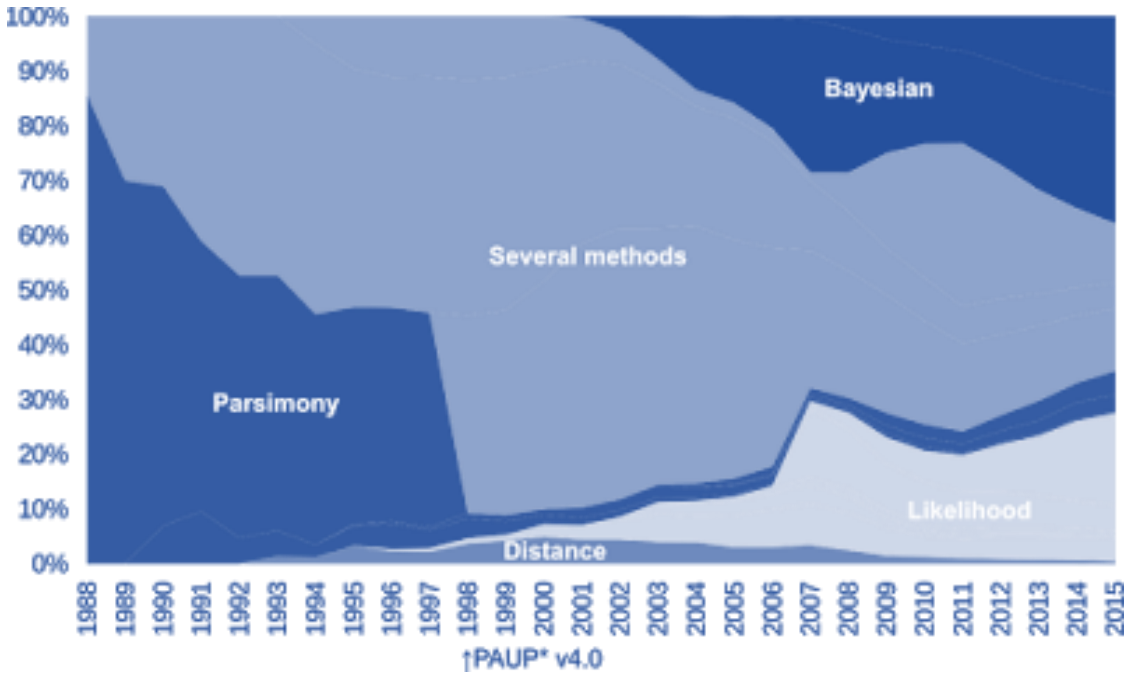
- Manual editing: correcting mismatching of key cofactor residues and residues of similar physicochemical properties 手动调节比对结果
- Full alignment or parts of it (domain only) 根据结构域比对情况进行调节
- Remove ambiguously aligned regions(subjective process) 删除可信度低的比对区域

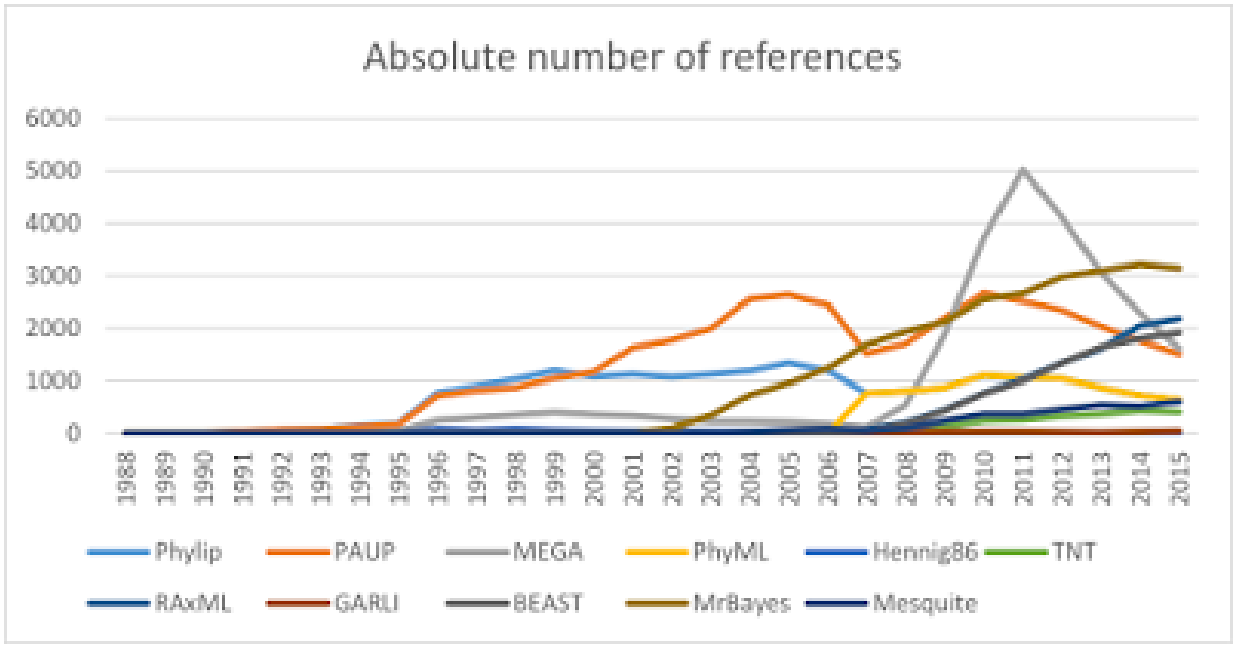
确定建树方法

NJ 邻接法Neighbor-joining	• 根据所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度
MP 最大简约法(Maximum Parsimony)	• 最大简约法根据序列的多重比对结果，对所有可能正确的拓扑结构进行计算并挑选出所需替代数最少的拓扑结构作为最优树，即能够利用最少的步骤去解释多重比对中的碱基差异。
ML 最大似然法 Maximum Likelihood	• 最大似然法以一个特定的替代模型分析一组序列数据的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，使所获得的每一个拓扑结构的似然率均为最大，挑选似然率最大的拓扑结构作为最优树。
Bayes 贝叶斯	• 贝叶斯方法比最大似然法能表示更多的可信进化模型替代率的变异可以在各个点建模贝叶斯方法有一个非常宽的先验分布后验概率分布用Gibbs 样本和MCMCM方法计算。。

一般来讲，如果模型合适，ML的效果较好。对近缘序列，有人喜欢MP，因为用的假设最少。MP一般不用在远缘序列上，这时一般用NJ或ML.对相似度很低的序列，NJ往往出现Long-branch attraction（LBA，长枝吸引现象），有时严重干扰进化树的构建。贝叶斯的方法则太慢。对于各种方法构建分子进化树的准确性，有综述认为贝叶斯的方法最好，其次是ML，然后是MP。

确定建树方法



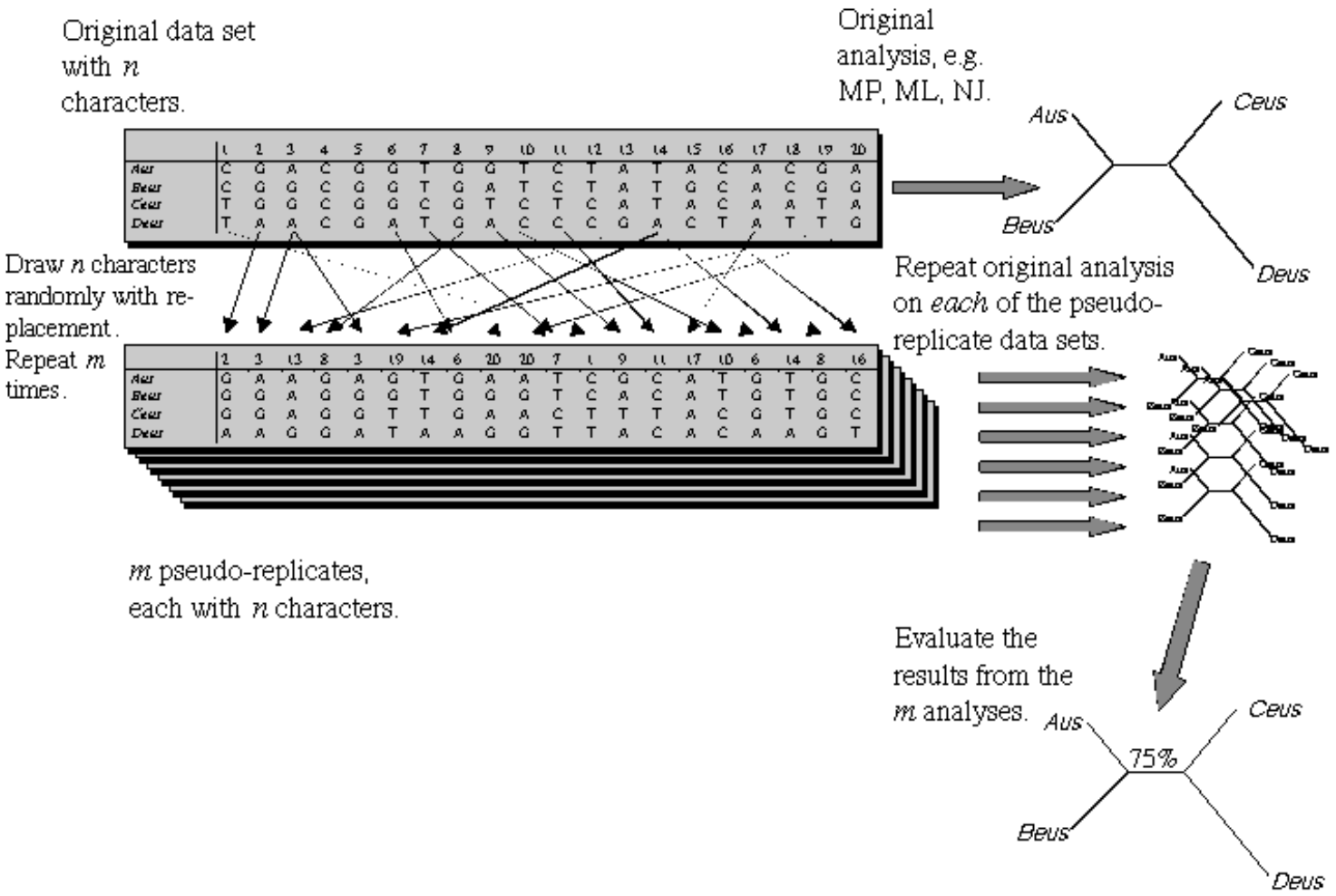


- 一般推荐用两种不同的方法构建进化树，如果所得到的进化树类似，则结果较为可靠。

http://phylobotanist.blogspot.co.uk/2016/04/the-popularity-of-phylogenetic-programs.html

系统树的可靠性评估——自举法（Boot strapping）

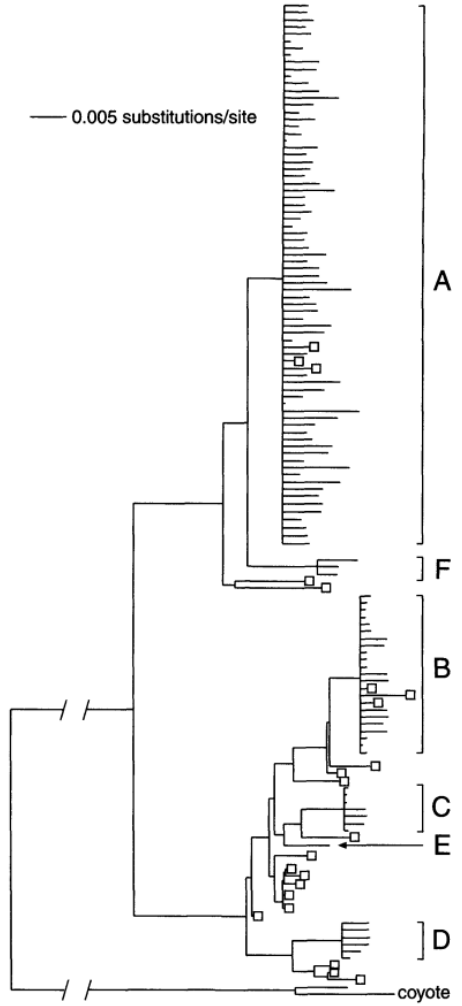
自举法：是对所比较序列上的替换位点作多次随机取样，根据每次取样的数据可以得到新的树形图，相同的组合出现在某一个节点上的次数占总取样次数的百分比就是该节点的bootstrap值。



应用实例

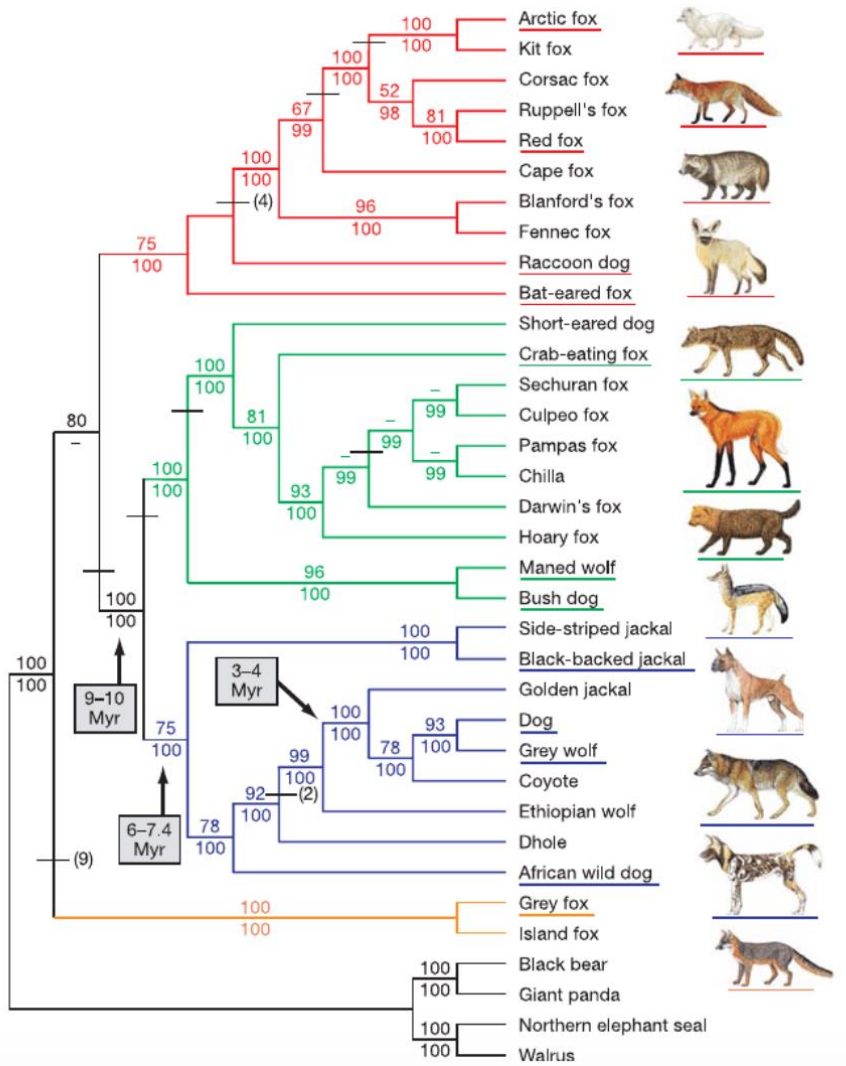
Practical cases

实例1：物种起源



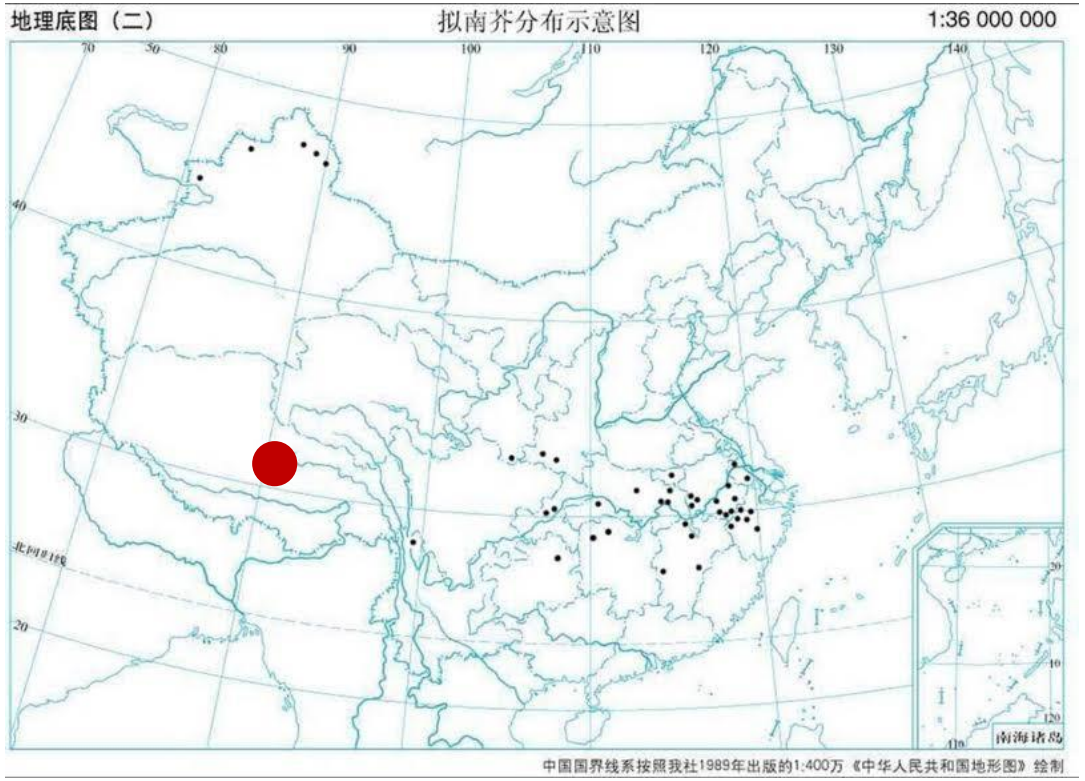
We analyzed the genetic variation in 582 base pairs (bp) of mtDNA in 654 domestic dogs from Europe, Asia, Africa, and Arctic America and in 38 Eurasian wolves (*13*, *14*)

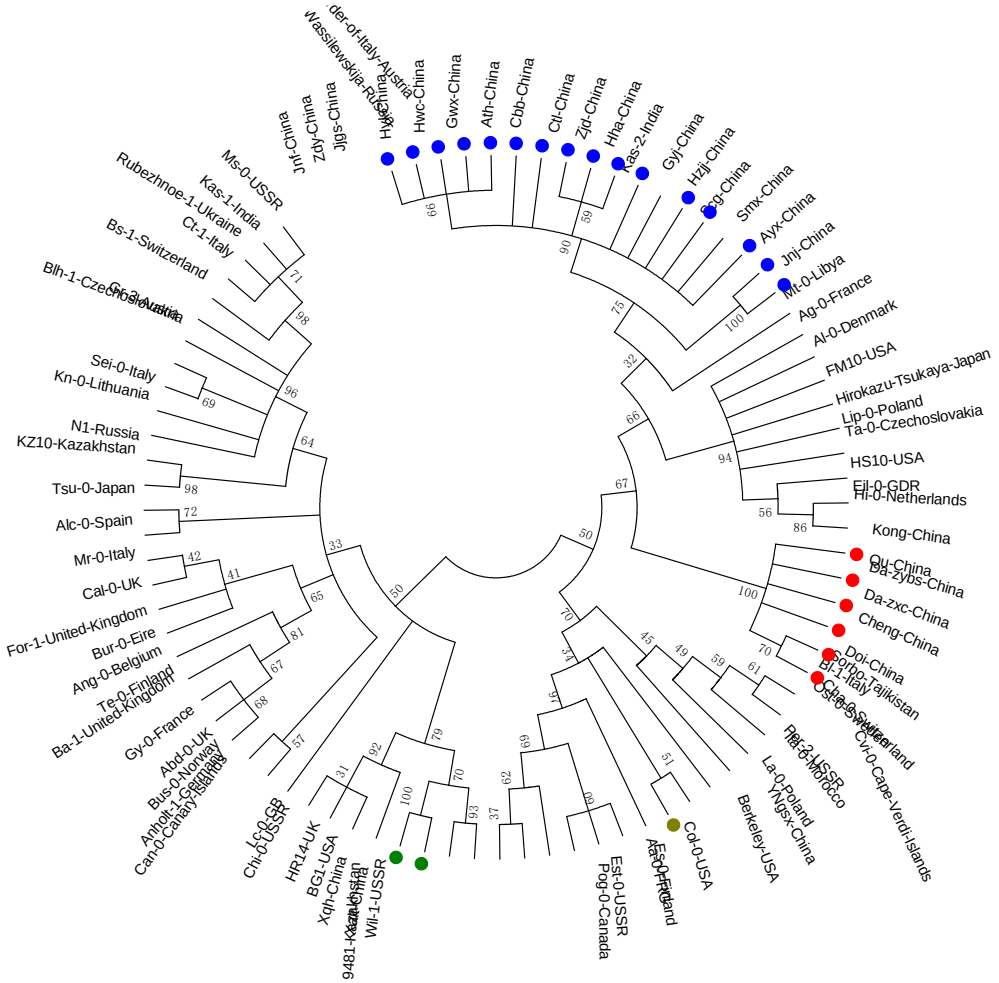
Savolainen et. al Science 2002



Lindblad-Toh et. al Nature 2005

实例II：适应性演化的研究

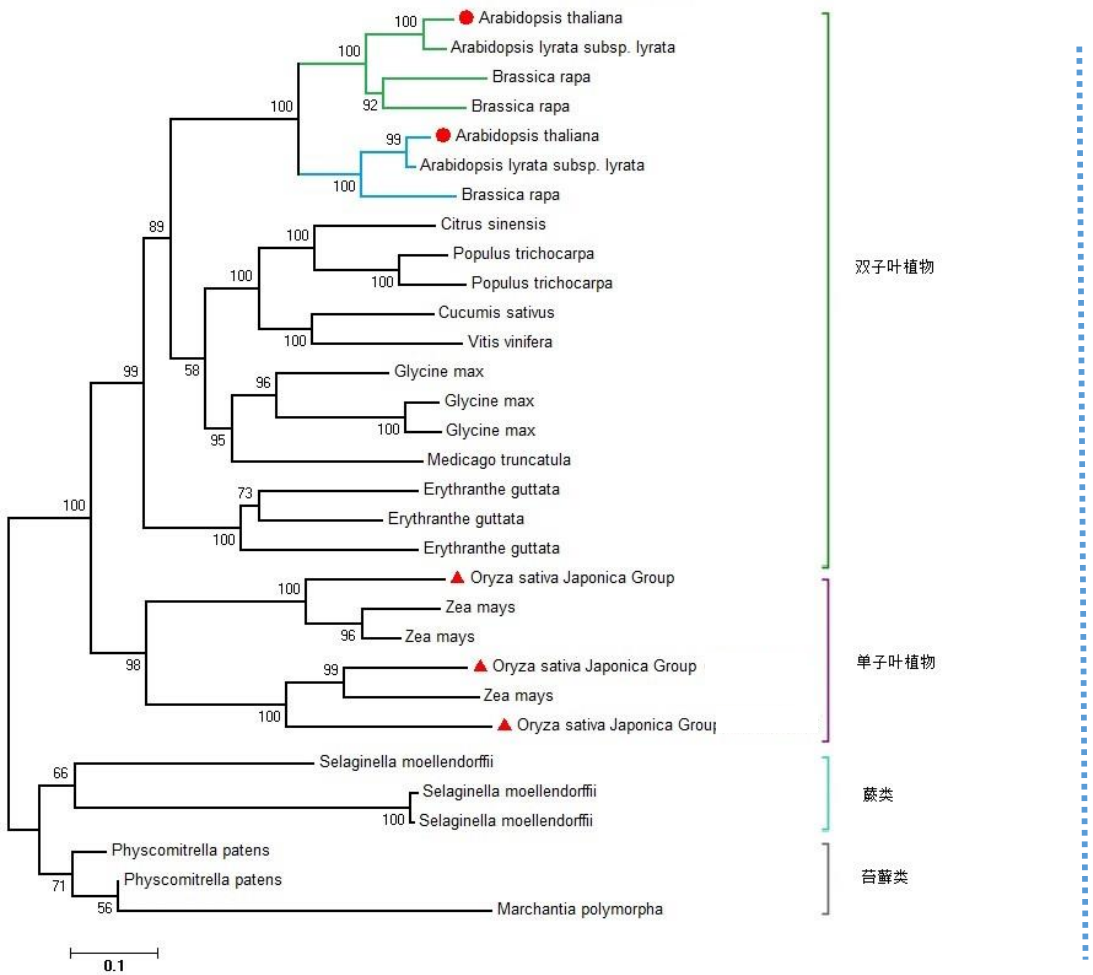




- 通过构建系统树确定西藏野生拟南芥居群起源，认为西藏居群是独立起源，且可能是其它长江流域居群的祖先。进而对其抗冻性进行研究（正向遗传）。
- 数据选择：不受选择的叶绿体基因间区

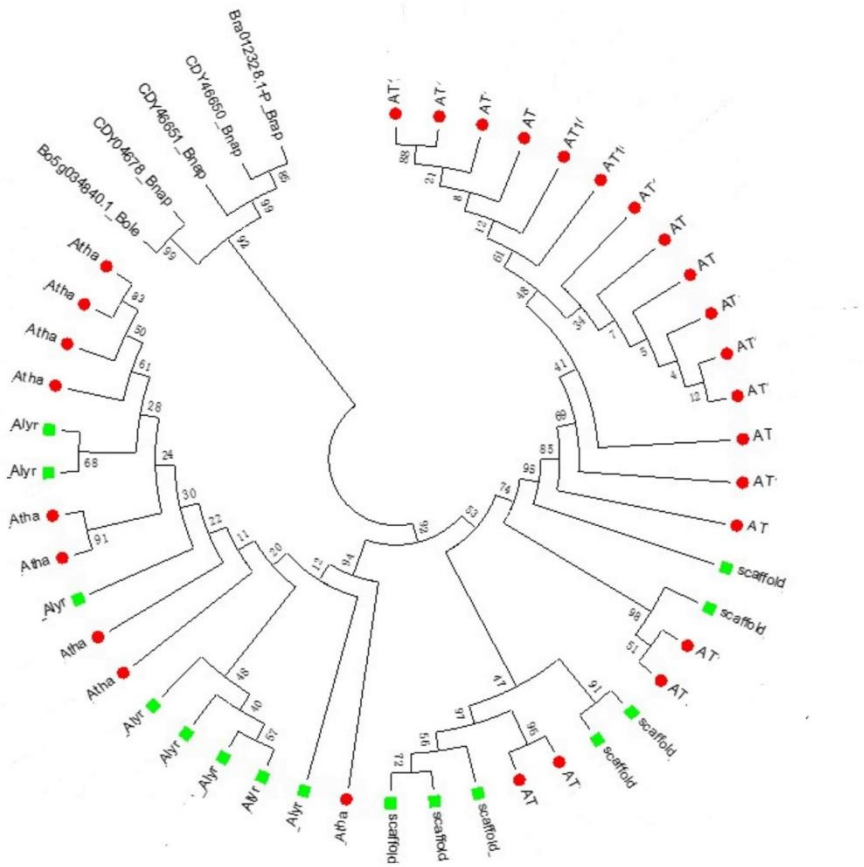
Tianshu Sun, unpublished data

实例III：基因功能的研究



Meiying Wu, unpublished data

- 研究拟南芥中已知功能的转录因子在不同物种演化作用，以及在水稻中的功能。
- 数据选择：蛋白序列

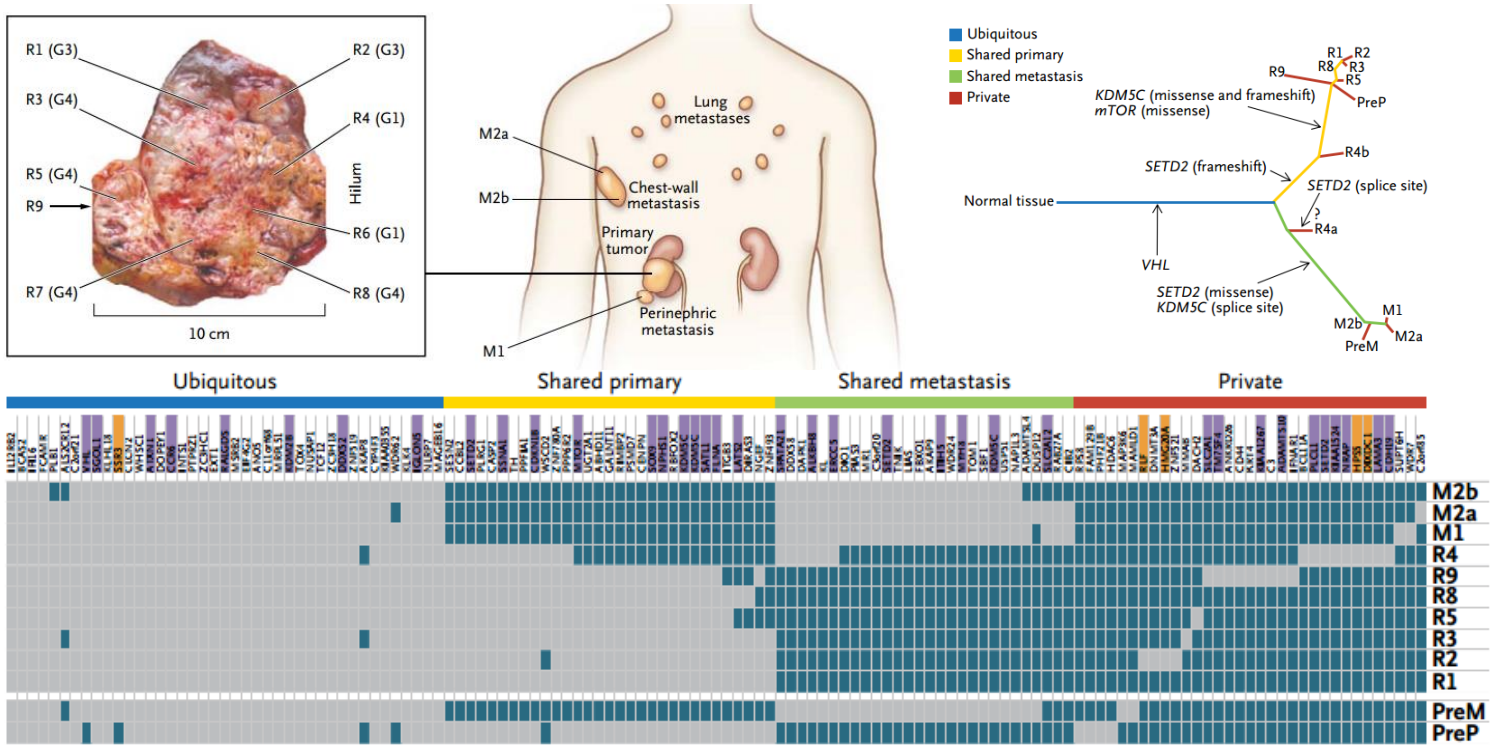


Yihao Shi, unpublished data

- 研究一组在配子体内特异表达的基因的分子功能（反向遗传）
- 数据选择：蛋白序列

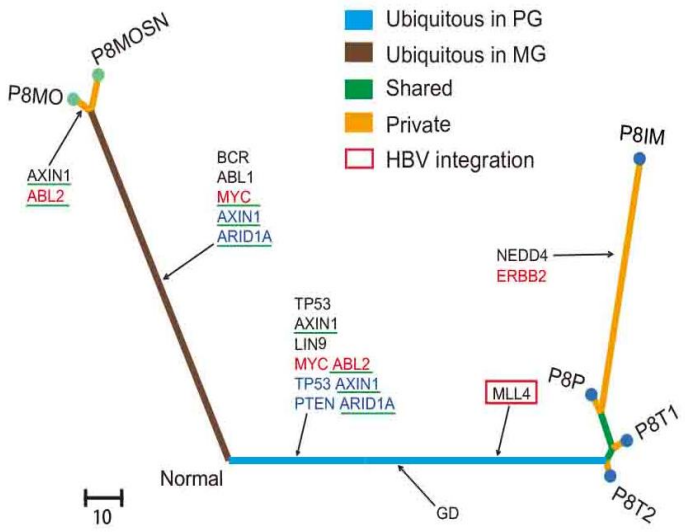
实例IV：肿瘤内部基因组层面异质性及克隆演化

原位vs.转移位



Gerlinger, M., *et.al.* N Engl J Med 2012;366:883-92.

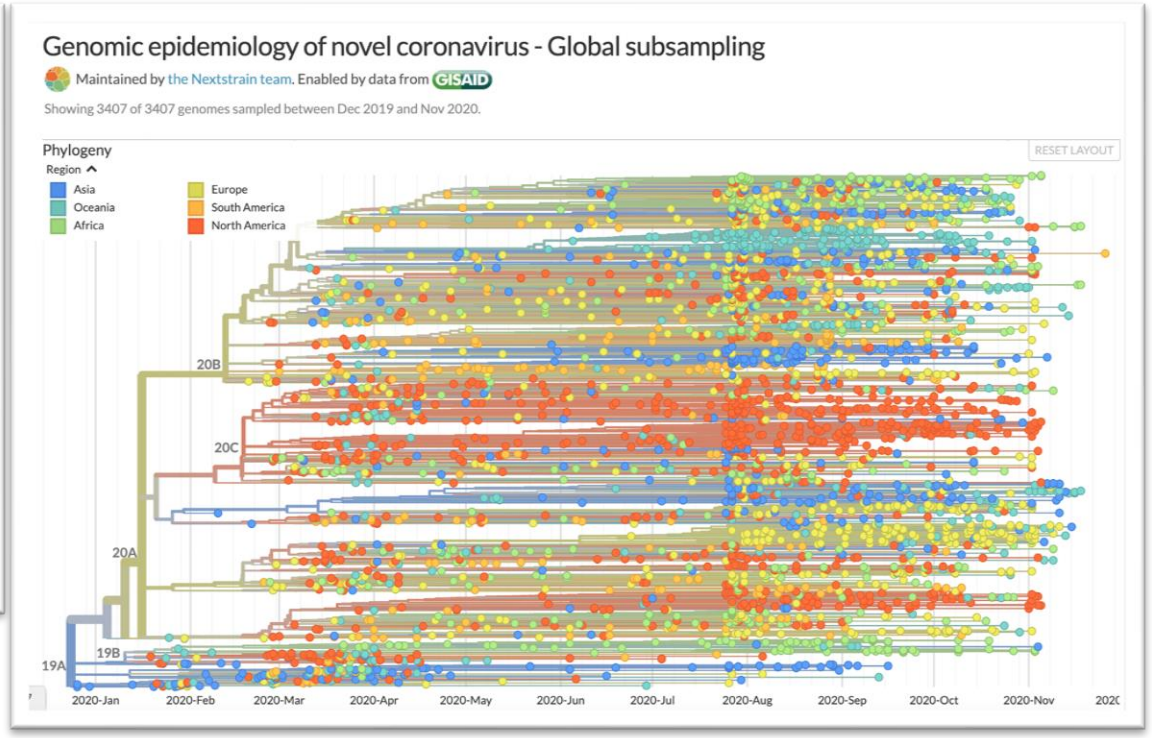
多源发病灶



Ruoyan Li, doctoral thesis

- 肿瘤内部基因组层面异质性及克隆演化，为临床诊断和治疗提供指导
- 数据选择：基因组（DNA）序列

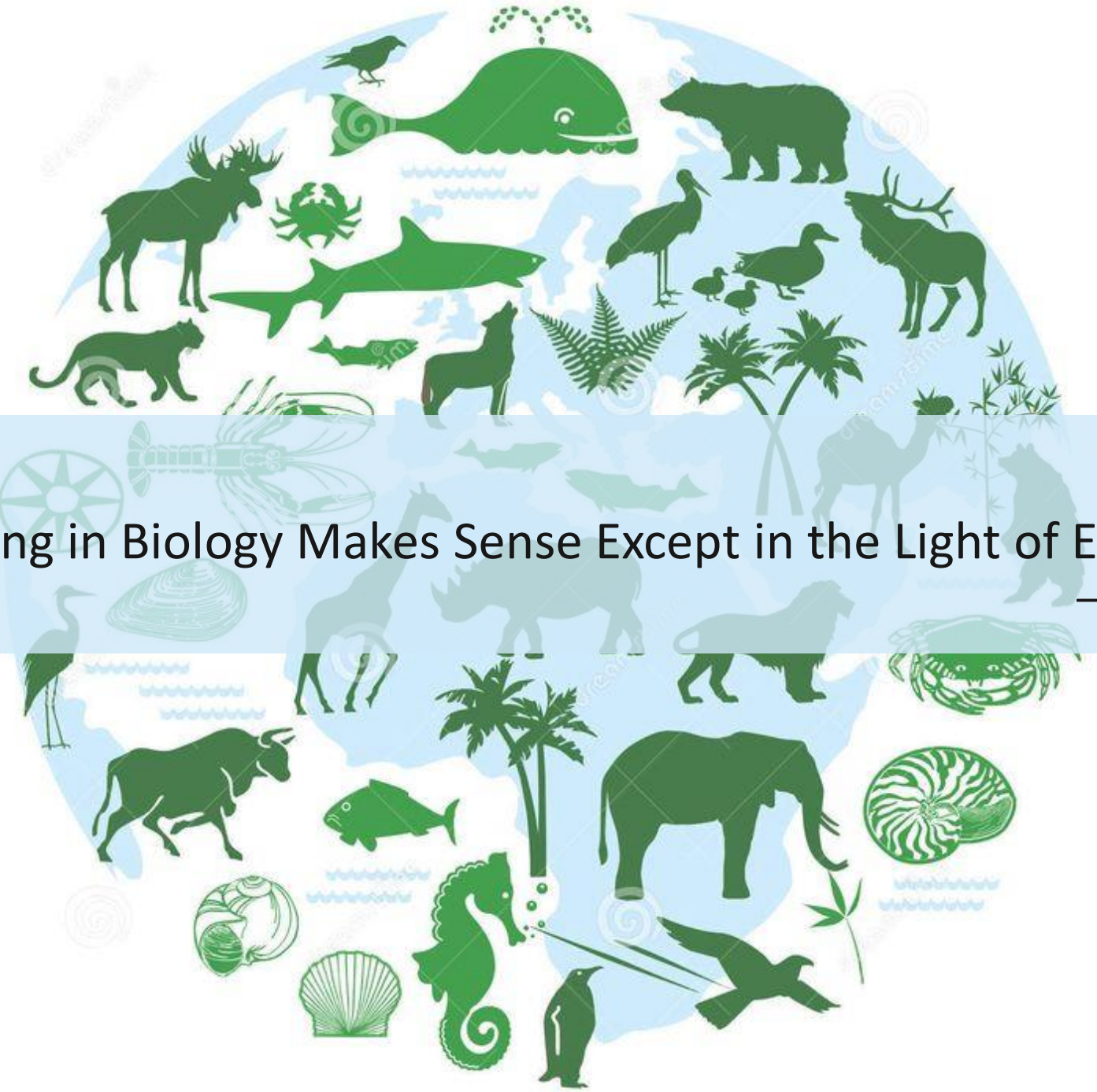
实例V: Covid-19



- development of diagnostics and vaccines
- assess patterns of transmission and spread

Nothing in Biology Makes Sense Except in the Light of Evolution

——Theodosius Dobzhansky



MOOC：生物演化

coursera Explore What do you want to learn? For Enterprise For Students Log In Join for Free

顾红雅
Peking University

Bio

顾红雅，北京大学生命科学学院教授，博士生导师。1982年获南京大学理学学士，1987年获美国华盛顿大学理学博士学位。现兼任 "*Journal of Molecular Evolution*"的 associate editor， "*Journal of Systematics and Evolution*"和"*Journal of Integrative Plant Biology*"的编委。顾红雅教授长期从事植物系统发育与演化、植物遗传多样性与环境适应性相关性研究，以及植物基因家族的演化研究。先后承担863、国家自然科学重点基金、国际合作等多项研究项目。在生物遗传及演化相关的国际刊物上发表论文60余篇；参与编著教材3部，科普专著1部，翻译教材3部。获北京大学第十七届"我爱我师·最受学生爱戴的老师"金葵奖，国家教委科技进步三等奖、中国科技协青年科技奖、国家教委优秀留学回国人员等奖项。

Courses

中国大学MOOC 课程 学校 学校云 慕课堂 下载APP 搜索感兴趣的课程

顾红雅 北京大学 – 教授
关注0人 | 粉丝330人
+ 关注

顾红雅，北京大学生命科学学院教授，博士生导师，主要从事植物遗传多样性和演化、基因家族的功能和演化，以及植物适应环境的分子机制的科学研究和教学工作。承担863、国家自然科学重点基金、国际合作等多项研究项目。在相关领域的国际刊物上发表论文80余篇。主编《生物演化》数字教材1部，参与编著教材3部，科普专著2部，参与翻译植物生物学、生物演... 查看全部

主讲课程(3) 讨论(247)

大学生物学
教育部大学生物学课程教 821

生物演化
北京大学 8877

青少年慕课——带你感受科技之光
爱课程 3849

Thanks!

感谢各位的聆听

Questions are welcomed!