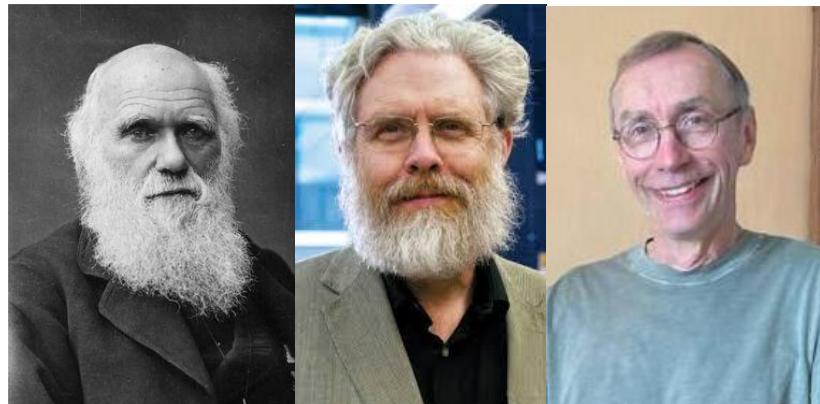


# 生物统计学： 生物信息中的概率统计模型

2022年秋



# 有关信息

- 授课教师: 宁康
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927
- 课程网页
  - <http://www.microbioinformatics.org/Biostatistics.html>
  - QQ群: 648612956



2022生物统计学

群号: 648612956



扫一扫二维码, 加入群聊。



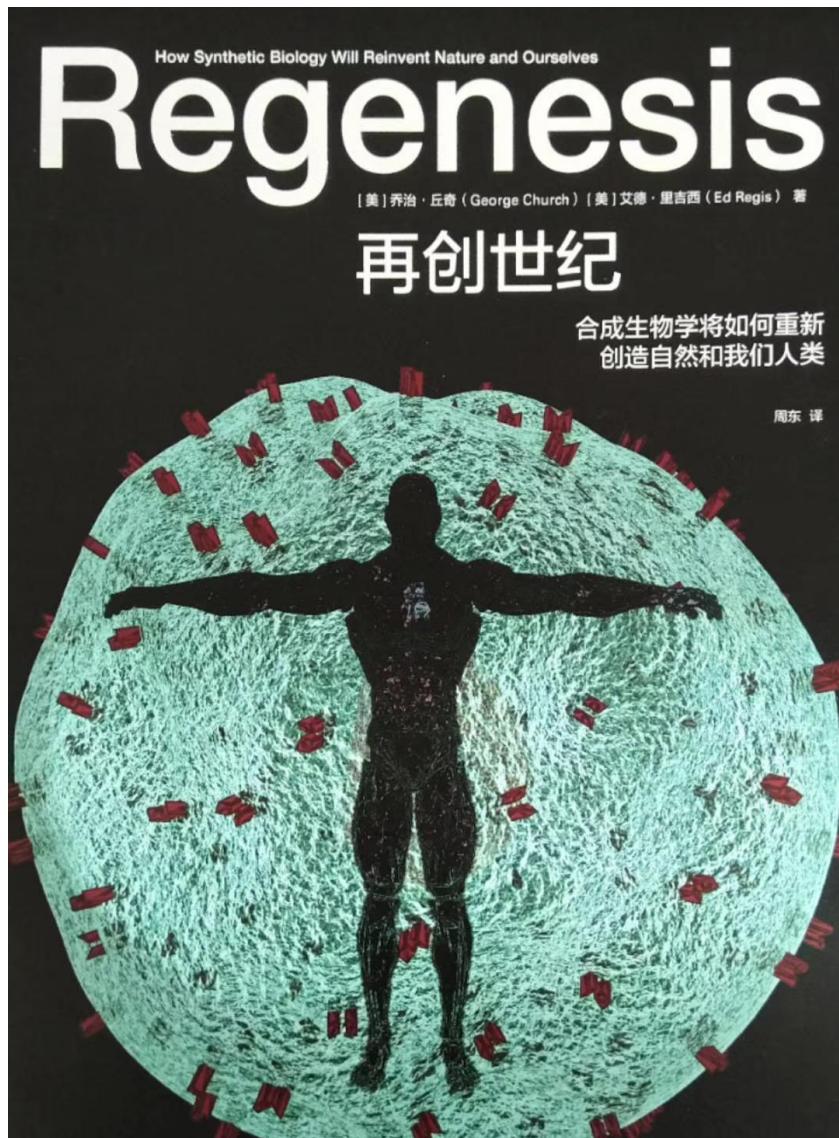
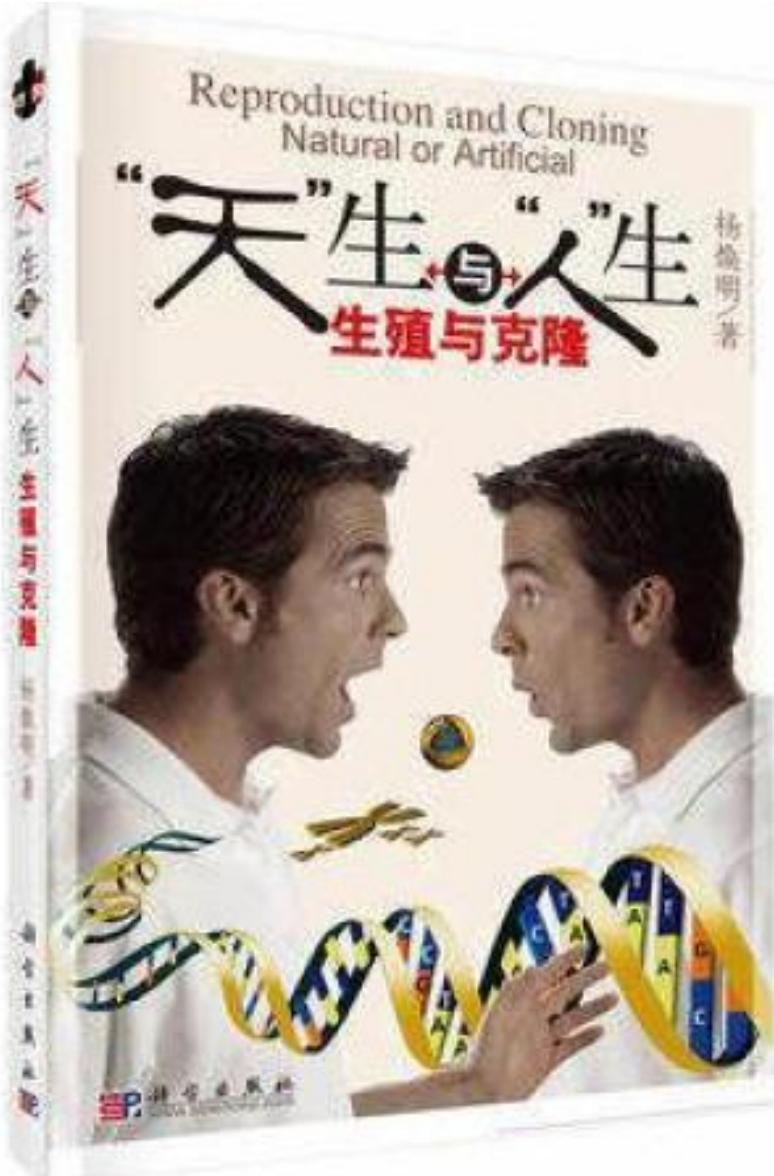
# 课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III )
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

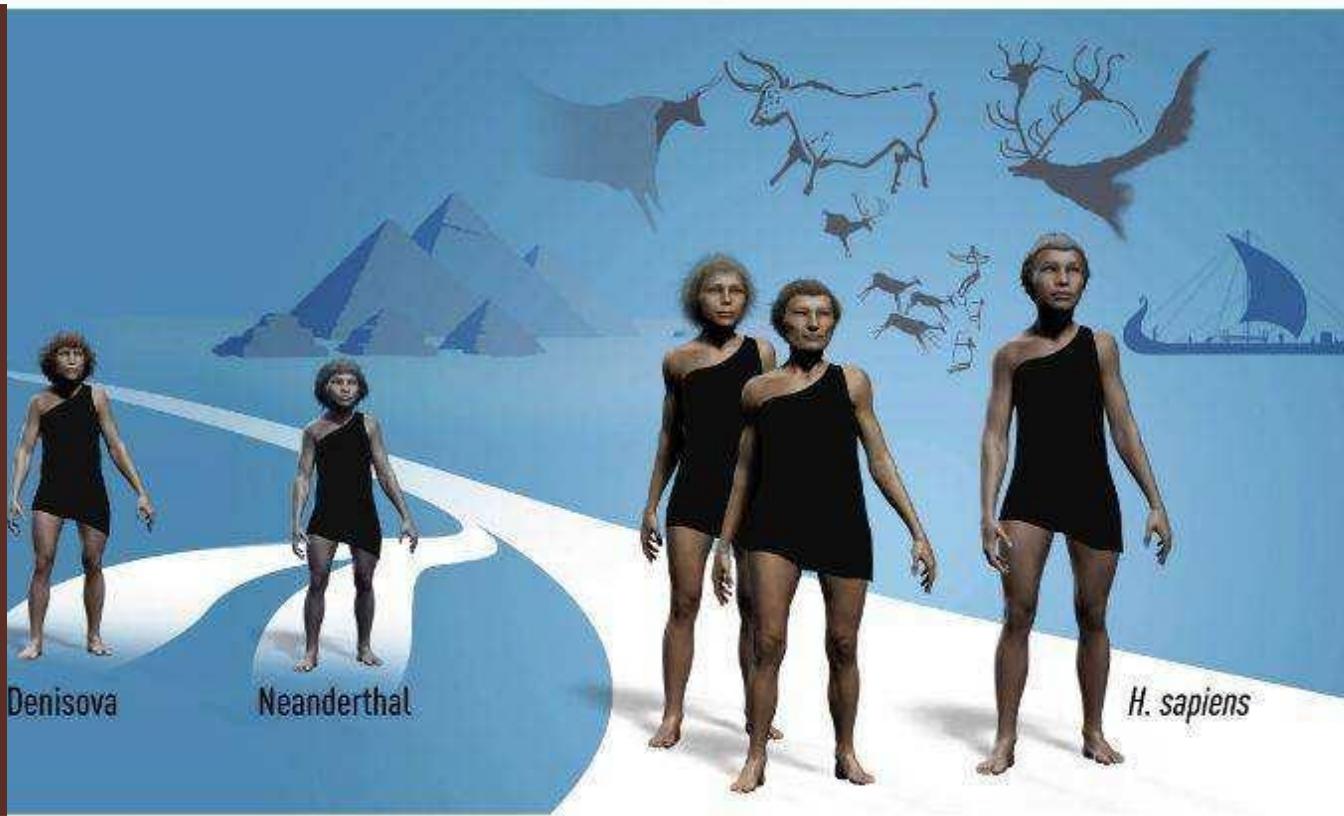
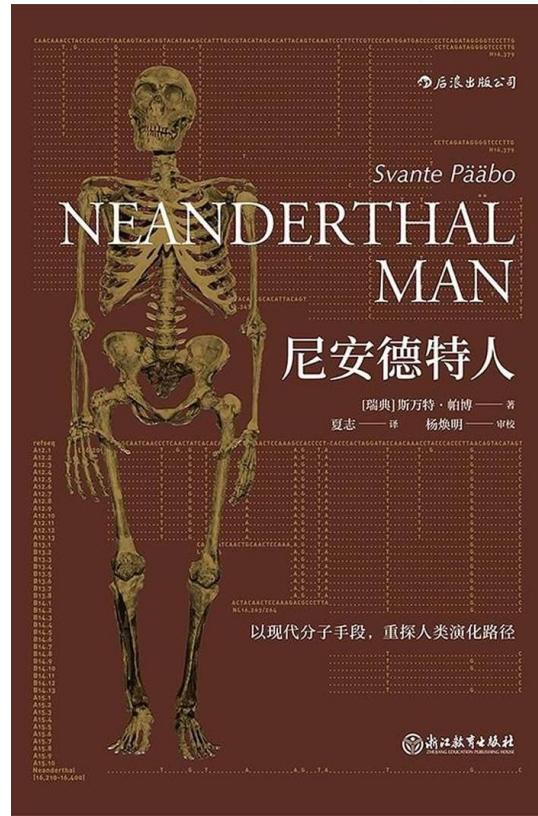
研究对象：  
生物序列，  
进化树，  
生物网络，  
基因表达  
...

方法：  
生物计算与生物统计

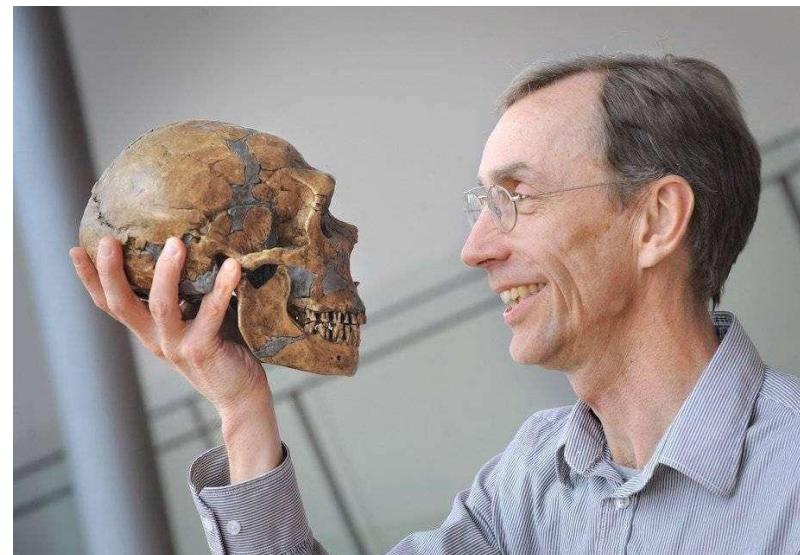
# 读物推荐



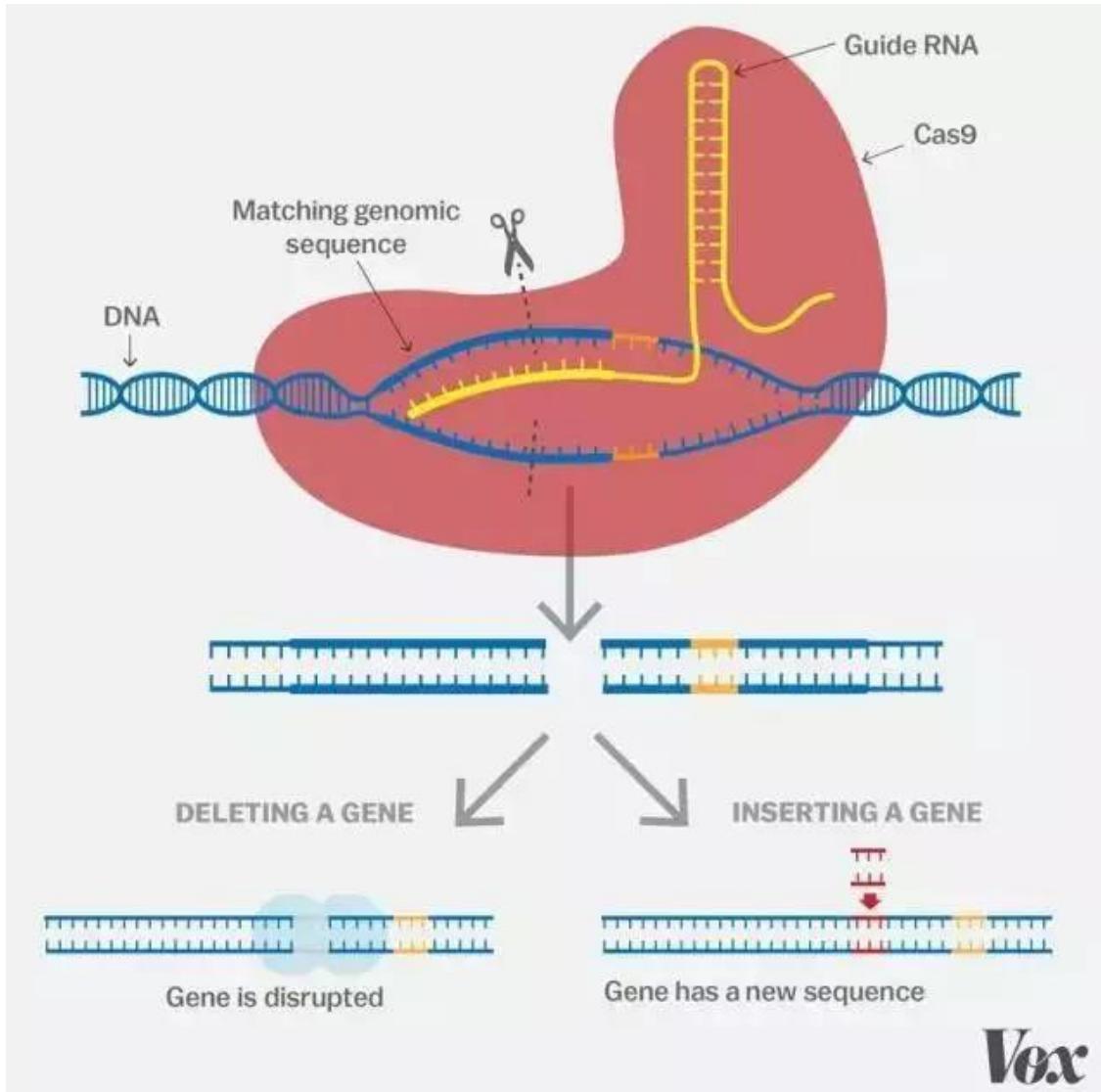
# 人类的进化



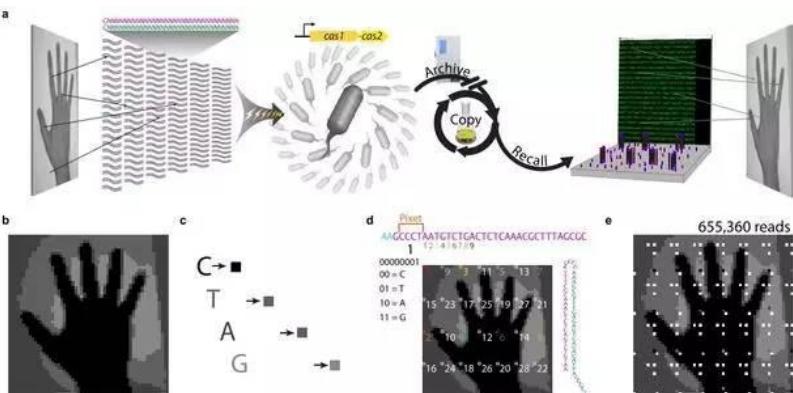
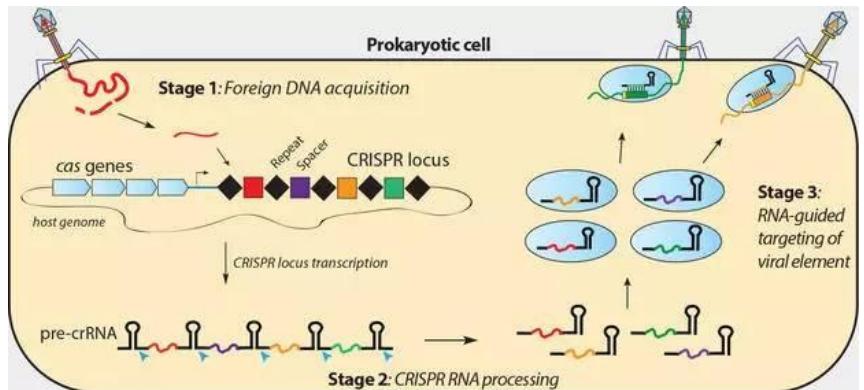
# 人类的进化



# CRISPR和基因编辑技术



# Understand it, create it!



Original Image

原始图像

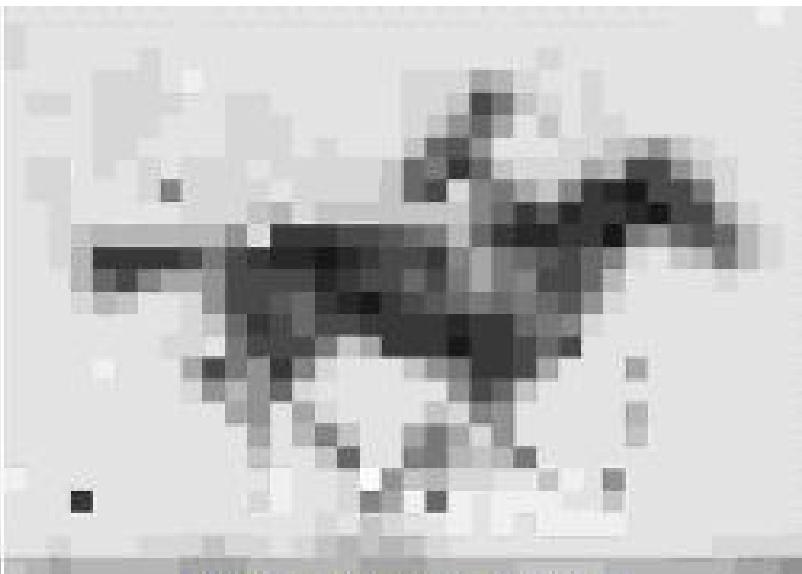
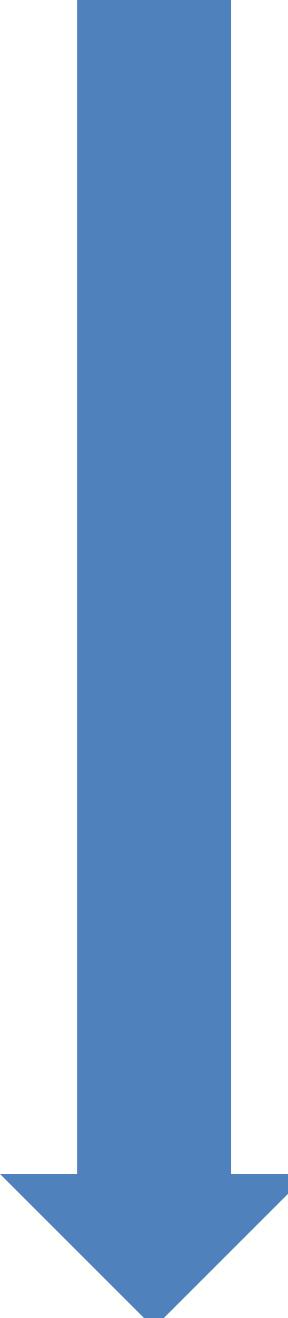
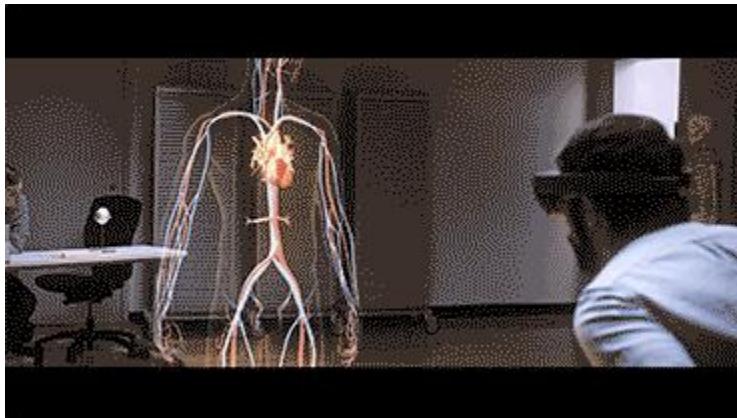


Image Reconstructed From Bacteria

从细菌DNA还原的图像



See it!



Understand it!



Create it!



# CRISPR和基因编辑技术

要清楚什么可以做，什么不可以做！

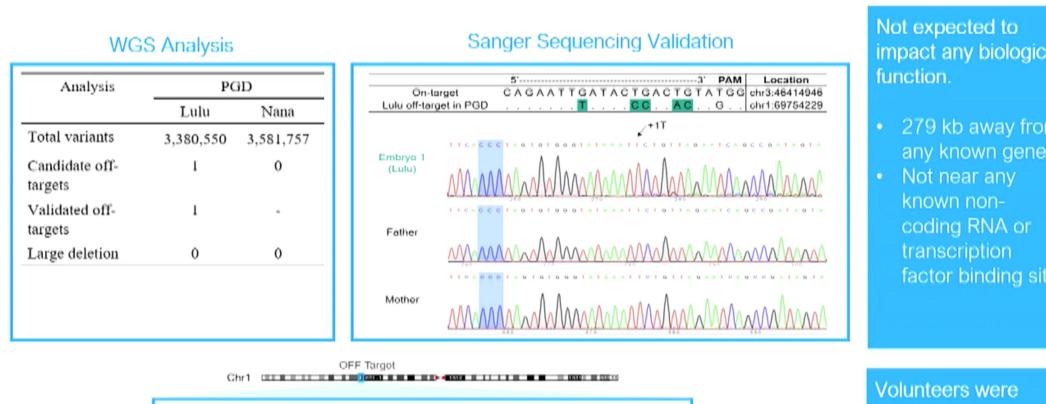
The slide title is "Embryo setting presents fewer editing events, but consequences pose very serious, whole-body risks". It features a graphic showing the number of cells in an adult human (783 billion) and the number impacted by off-targets (37.5 million). A table lists various adult tissues and their cell counts:

Tissue	Cells in an adult	Cells impacted by off-target of 0.0005% frequency*
Blood	783 billion	37.5 million
Liver	361 billion	18 million
Heart muscle	2 billion	100,000
Pancreas and cells	3 billion	150,000
Neurons	60 - 180 billion	3 - 15 million
Urat cells	3 billion	150,000

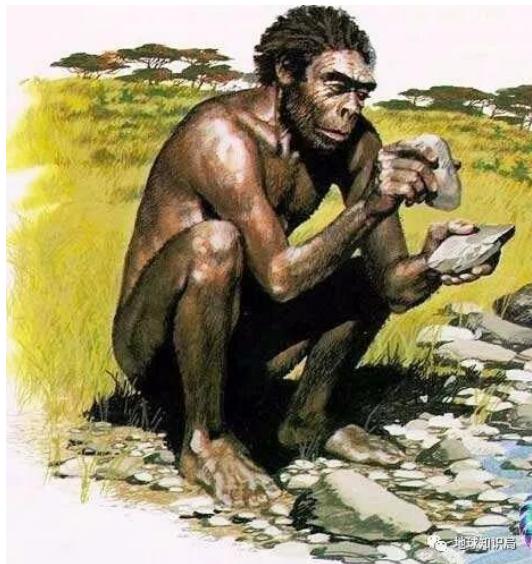
A large bracket indicates that adults have 37 trillion cells. On the right, a video frame shows a speaker at the podium.

In humans, after advice in Feb last year; found could achieve with indications of safety within 14 days

PGD identified one potential intergenic off-target



# CRISPR和基因编辑技术



OR



# CRISPR和基因编辑技术

Are there compelling medical indications?

The diagram features a horizontal bar divided into three color-coded sections: green on the left, white in the middle containing a question mark, and orange on the right. On either side of the bar are stylized traffic lights. The green section is associated with 'Disease prevention' and 'Consider alternatives...'. The orange section is associated with 'Modifying Disease Risk' and 'Enhancements'. A link is provided for 'Enhancements': <https://www.dnalc.org/view/1390-Genes-for-Learning-and-Memory.html>.

Disease prevention	Modifying Disease Risk	"Enhancements"
<ul style="list-style-type: none"><li>Huntington's</li><li>Tay Sach's</li><li>Cystic Fibrosis</li><li>Sickle cell anemia</li></ul>	<ul style="list-style-type: none"><li>HIV resistance (CCR5)</li><li>Heart disease (PCSK9)</li><li>Alzheimer's (APP A673T/+)</li><li>Cancer (BRCA1/2)</li><li>Resistance to global pandemics...</li></ul>	<ul style="list-style-type: none"><li>Muscularity (MSTN)</li><li>Height, skin color</li><li>Learning and memory</li></ul> <p><a href="https://www.dnalc.org/view/1390-Genes-for-Learning-and-Memory.html">https://www.dnalc.org/view/1390-Genes-for-Learning-and-Memory.html</a></p>

Consider alternatives...  
IVF, genetic diagnosis  
Somatic therapy  
When no alternative...  
Couples, both affected  
Infertility

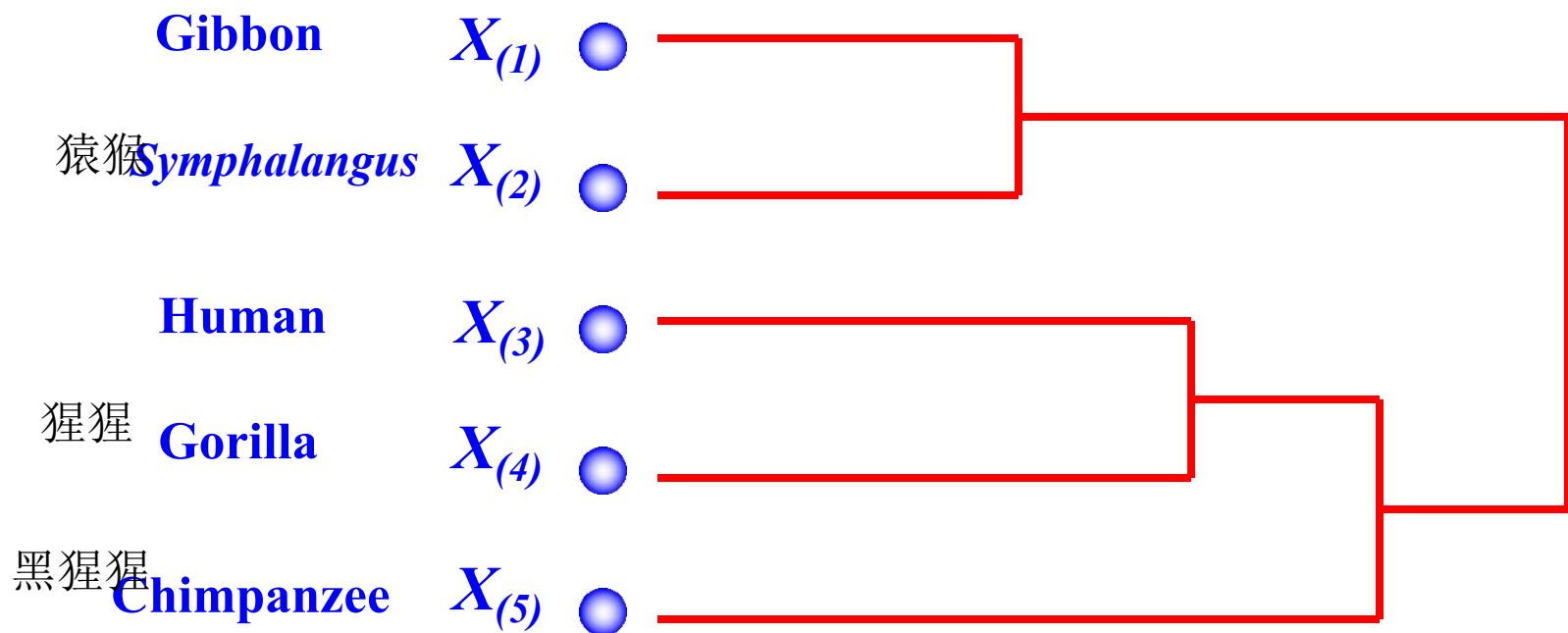
Permissible vs impermissible applications?

# 第4章：进化树构建的概率方法

- 问题介绍
- 进化树构建方法的概率方法

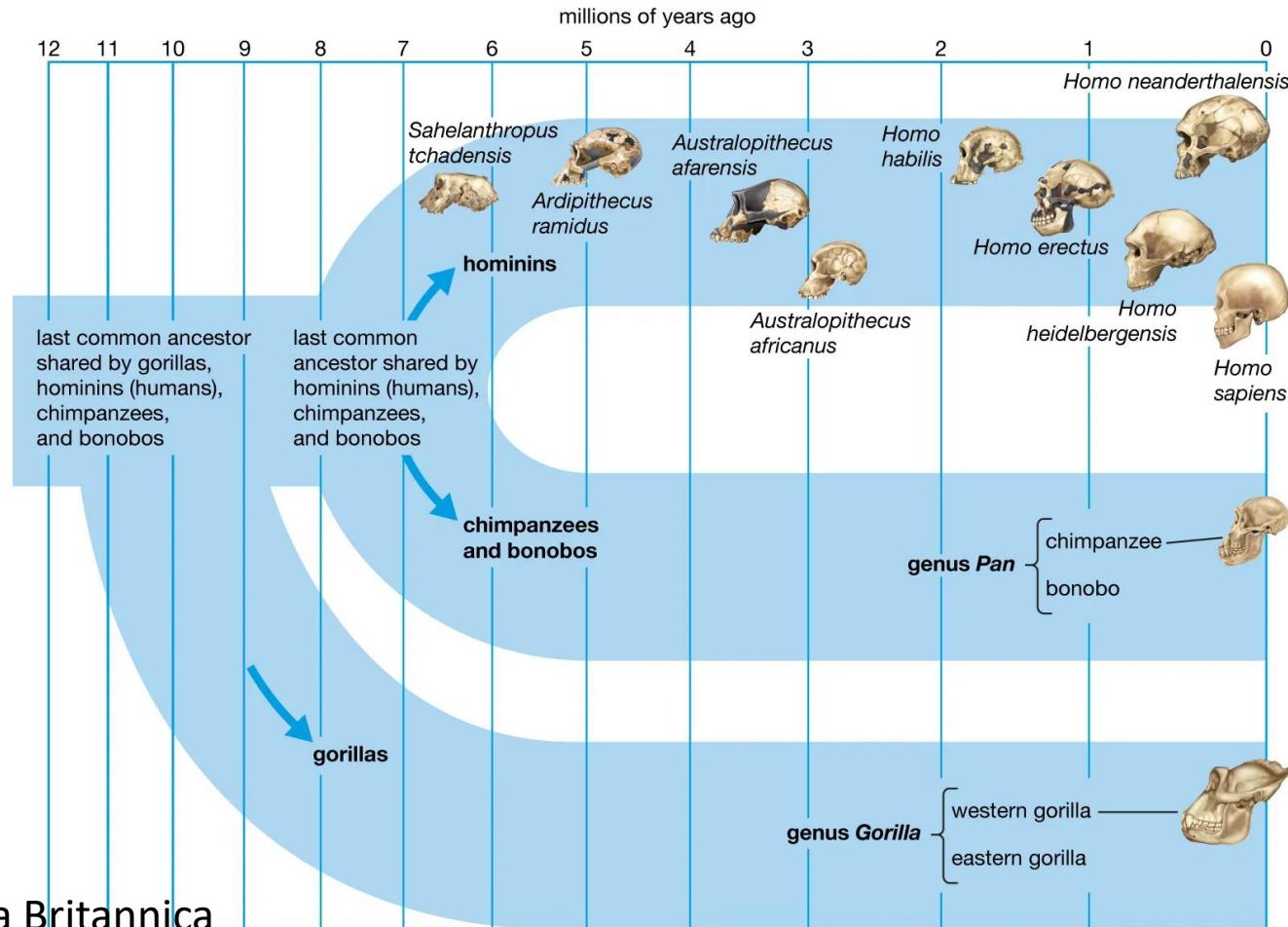
# Phylogenetic Tree

A toy example: primates



# Phylogenetic Tree

The evolution of primates as of 2020



# Phylogenetic Tree

The evolution of primates as of 2020

现存人科动物的分化关系：



3 1 0 0 万年前，人猿类跟猴猱类分离；

2 0 4 0 万年前，长臂猿跟“红+大+黑+人”分离；

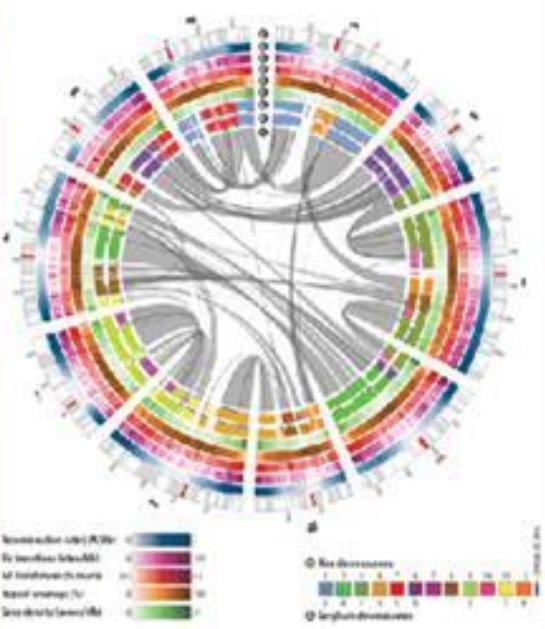
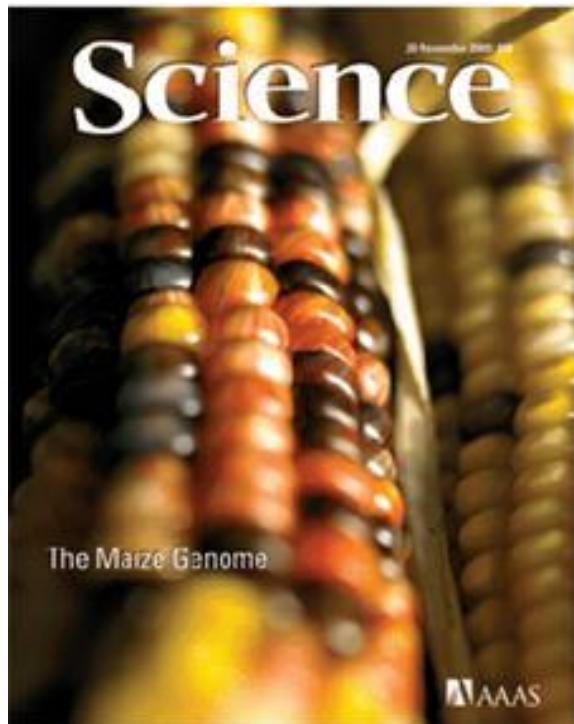
1 5 7 0 万年前，红猩猩跟“大+黑+人”分离；

8 8 0 万年前，大猩猩跟“黑+人”分离；

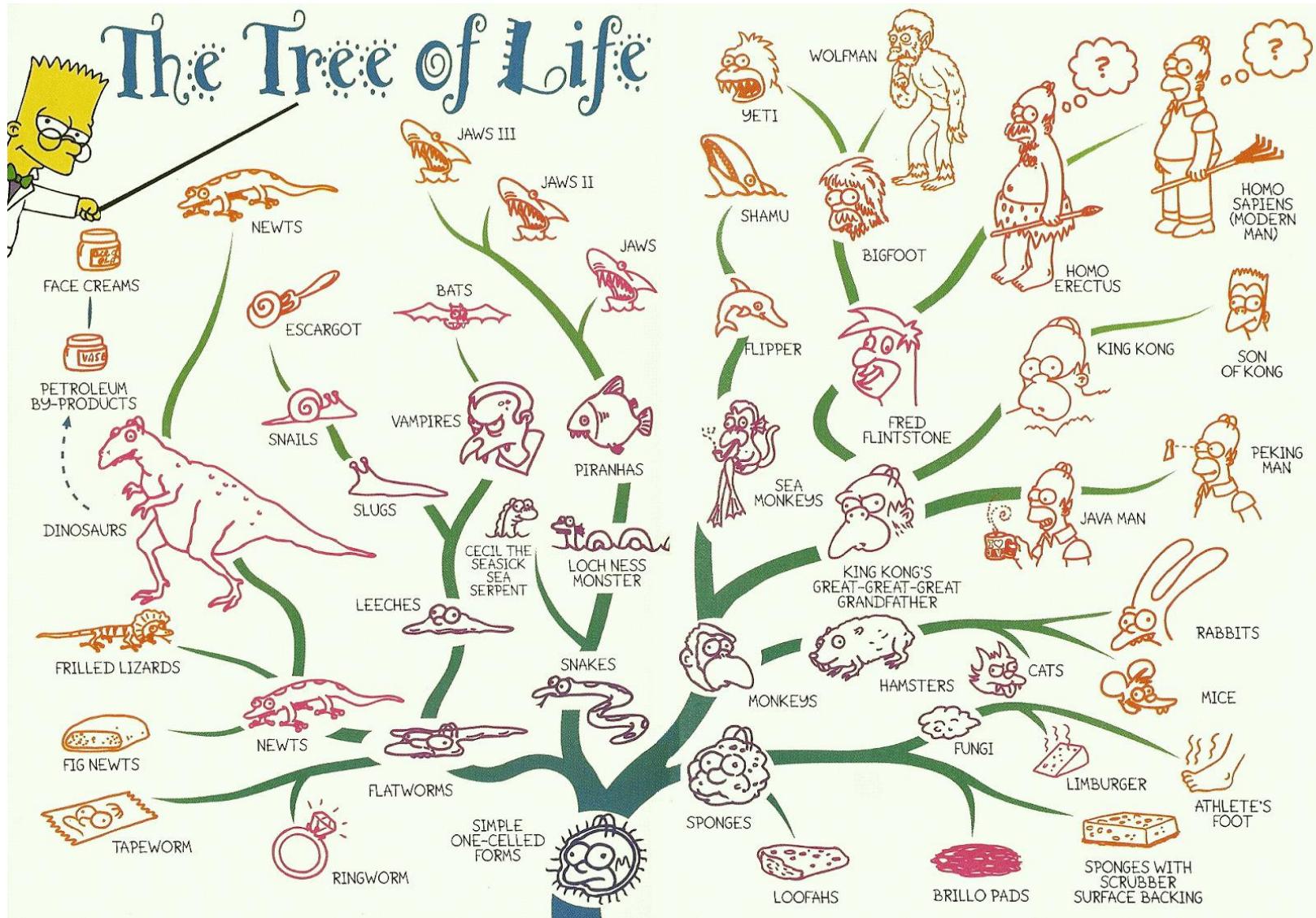
6 3 0 万年前，黑猩猩跟人类分离.....

# Phylogenetic Tree

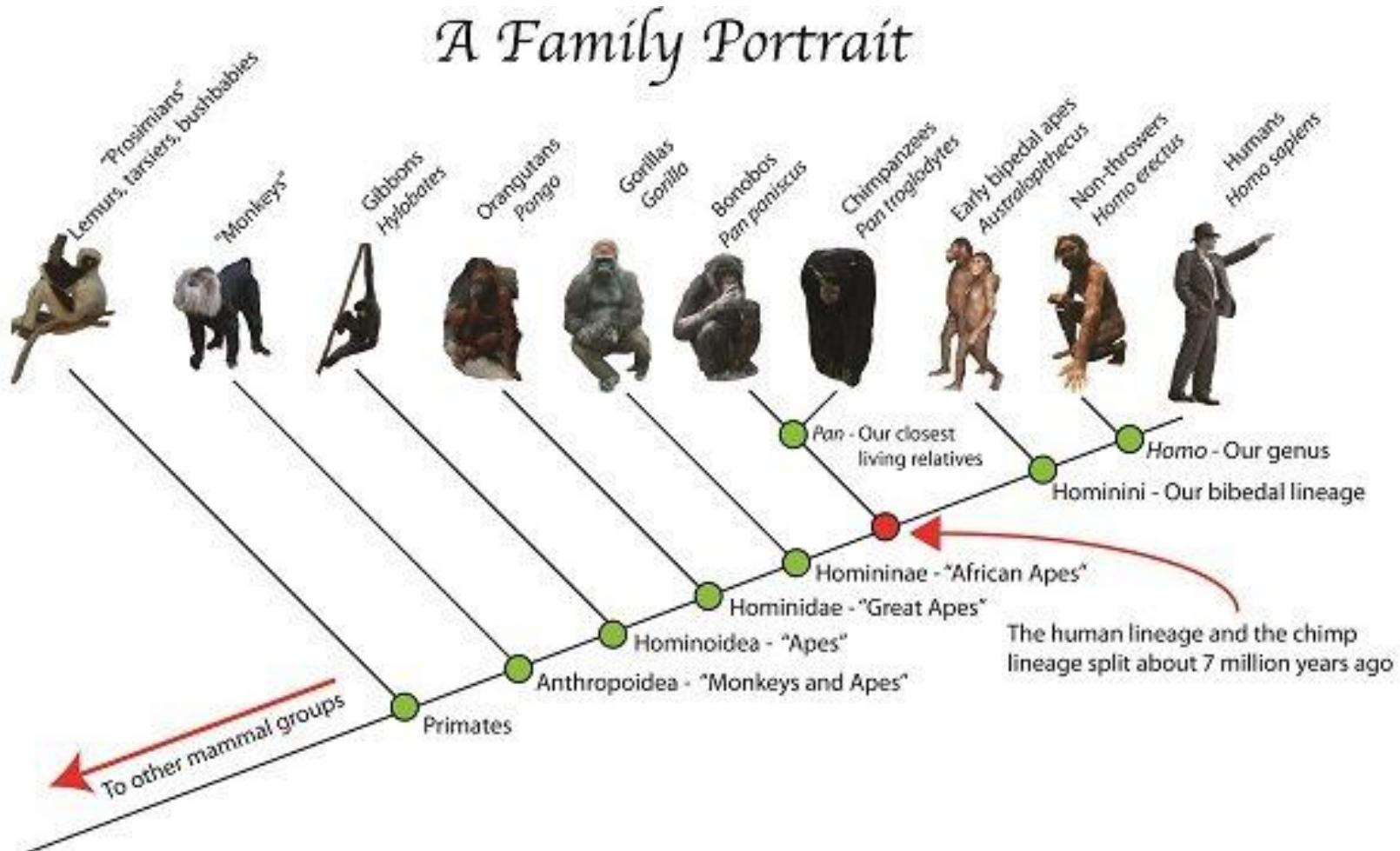
Everyone what to know how this tree  
look like...



# Phylogenetic Tree

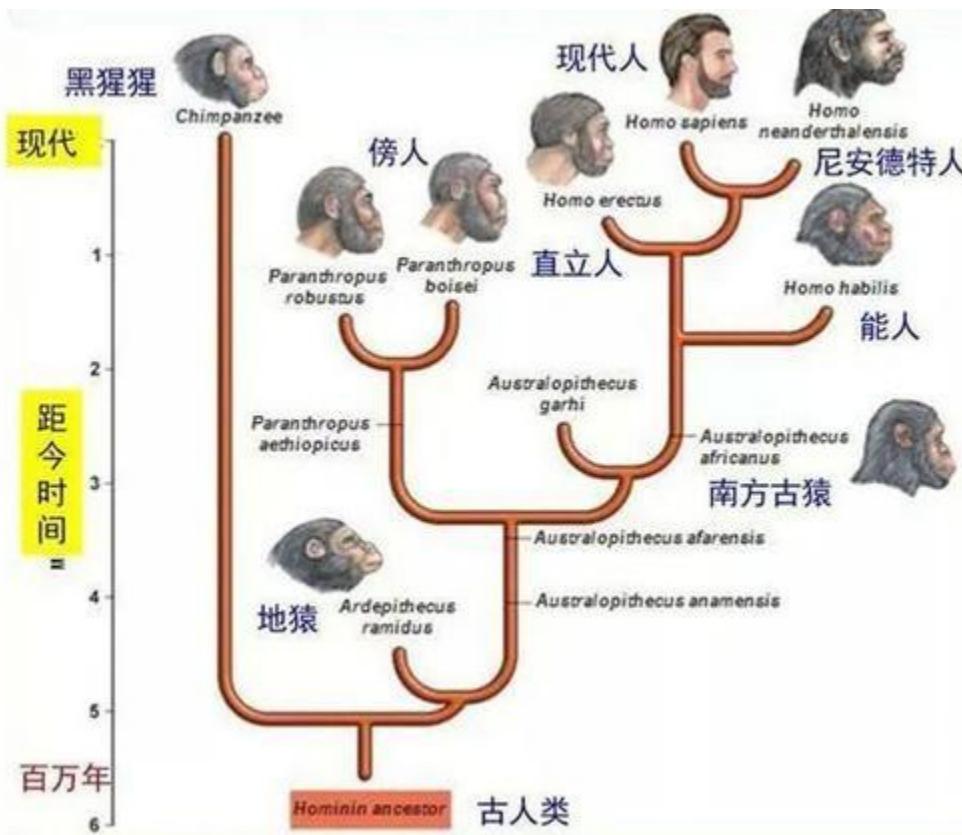


# Evolution everywhere



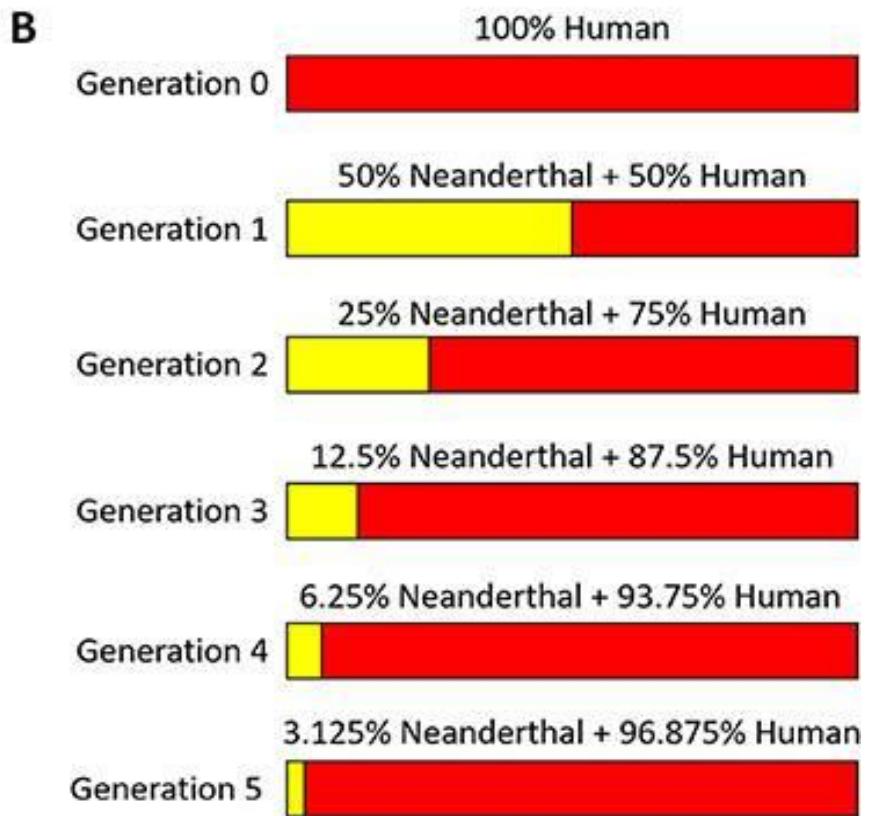
# Evolution everywhere

## Human evolution



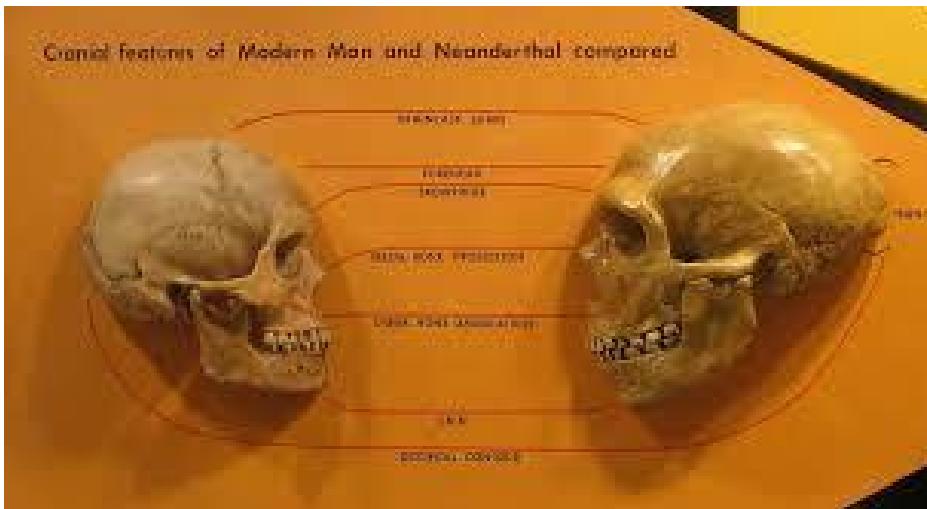
# Evolution everywhere

## Human evolution



# Evolution everywhere

## Human evolution



**ELONGATED SKULL**  
The Neanderthal face tended to be larger, with a brain case set back in a longer skull. An elongated skull may hint at a Neanderthal inheritance and is particularly common in the British Isles, Scandinavia and Iberia.

**SUPRAORBITAL RIDGE**  
The suprorbital ridge is a bony brow above the eye sockets which reinforces the weaker bones of the face. The pronounced brow ridge that Neanderthals shared with other archaic human species reduced when modern humans evolved, but did not disappear entirely.

**STRAIGHT, RED, THICK HAIR**  
70% of modern East Asians inherited mutations in genes which may be responsible for straightening and thickening hair. Between 2% and 6% of modern northwestern Europeans have red hair, a trait inherited from Neanderthals, compared with a global average of around 0.6%.

**FAIR SKIN AND FRECKLES**  
Fair skin is an advantage at northern latitudes because it is more efficient at generating vitamin D from weak sunlight.

**ROSY CHEEKS**  
Neanderthals had a large mental foramen in their mandible for facial blood supply, resulting in a reddening of the cheeks in cold weather or while doing physical exercise.

**SPACE BEHIND THE WISDOM TEETH**  
Neanderthals had jaws large enough to comfortably house all of their teeth. The jaw of the modern human doesn't have the space to cope with these vestiges of our foliage-chewing past which is why some of us need wisdom teeth removed.

**BROAD, PROJECTION NOSE**  
The angle of the Neanderthal nose bone projected out with a wide opening, making it a large and prominent facial feature. It could be an influence on the modern human aquiline nose prevalent in the Neanderthal hotspots of southern Europe and the Near East.

**LITTLE OR NO PROTRUDING CHIN**  
The Neanderthals' large jaw and protruding mid-face meant that they had a weak, or receding chin. The receding chin in modern humans is normally a congenital condition.

**INSULATING SKIN**  
The same Neanderthal mutations which affect hair also affect skin, making it more insulating and better adapted to colder environments.



# Evolution everywhere

## Human evolution

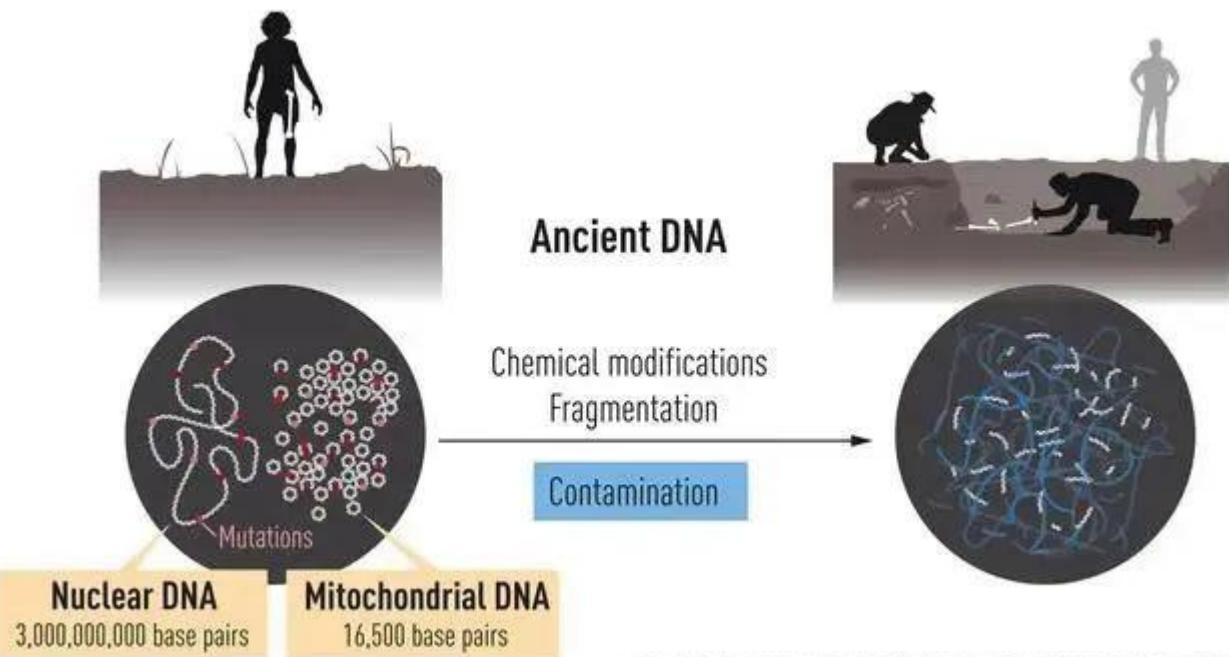
THE NOBEL PRIZE  
IN PHYSIOLOGY OR MEDICINE 2022



Svante Pääbo

"for his discoveries concerning the genomes of extinct hominins and human evolution"

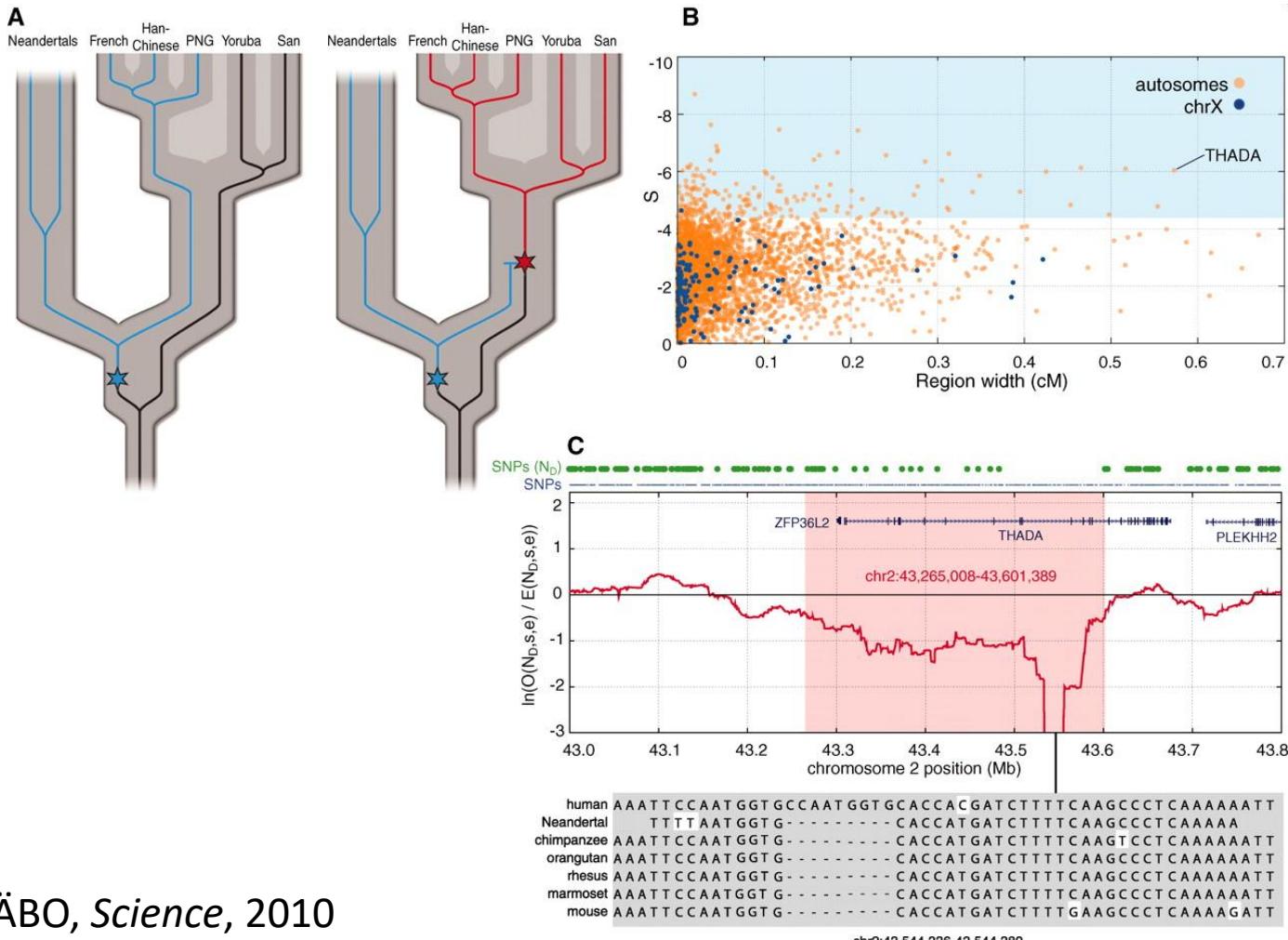
THE NOBEL ASSEMBLY AT KAROLINSKA INSTITUTET



© The Nobel Committee for Physiology or Medicine. Illustration: Mattias Karlén

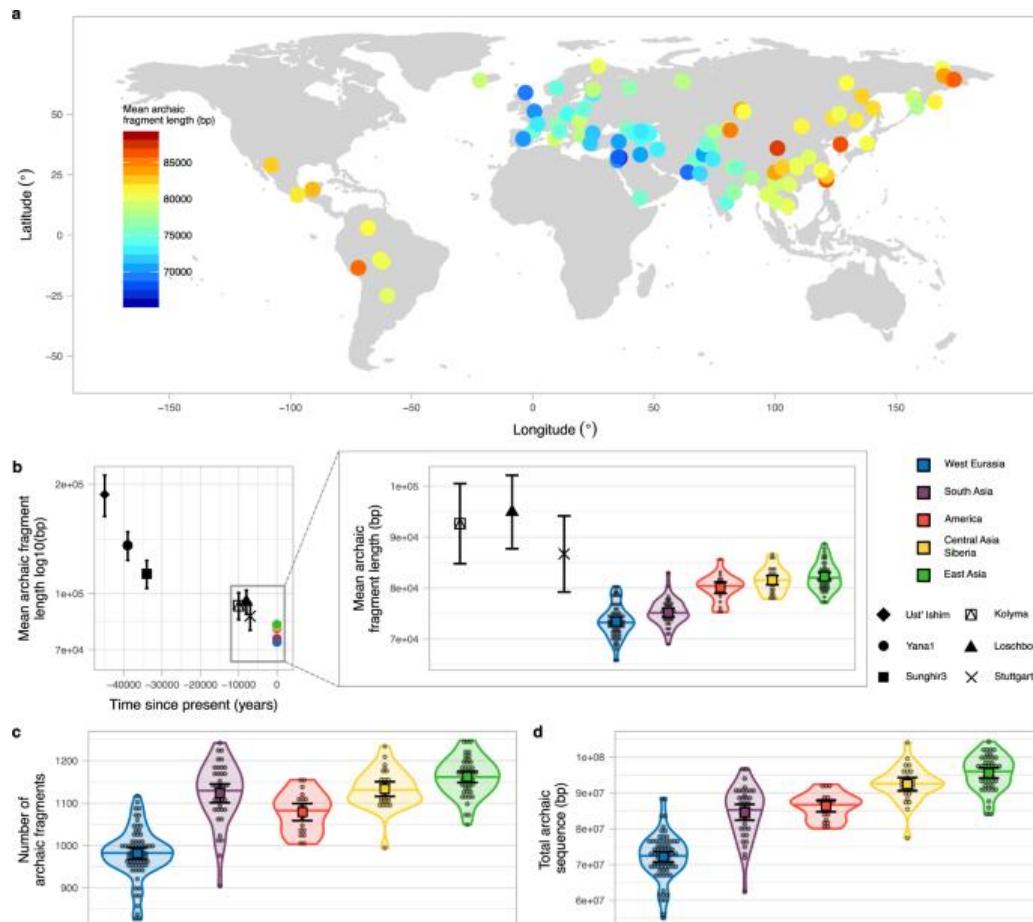
# Evolution everywhere

## Human evolution

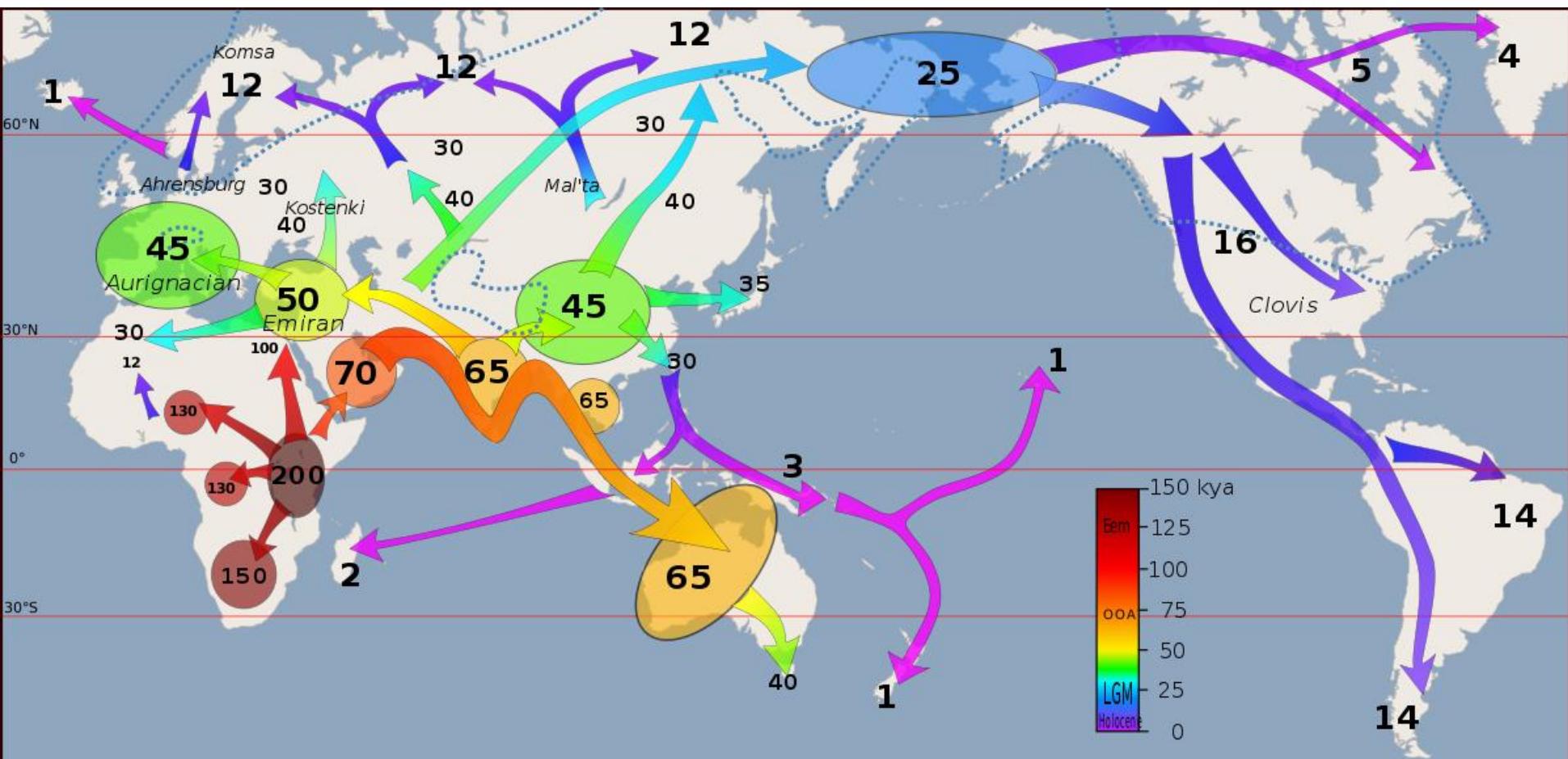
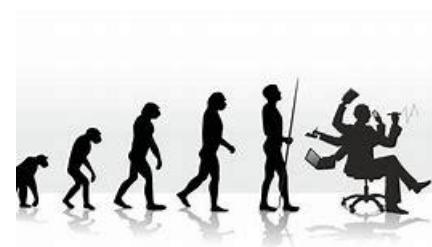


# Evolution everywhere

## Human evolution

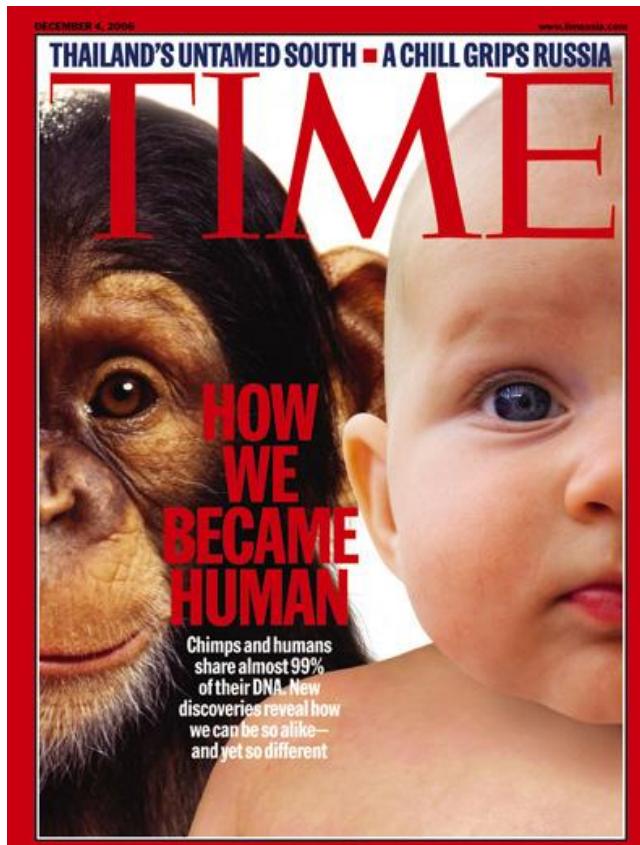


# Evolution everywhere



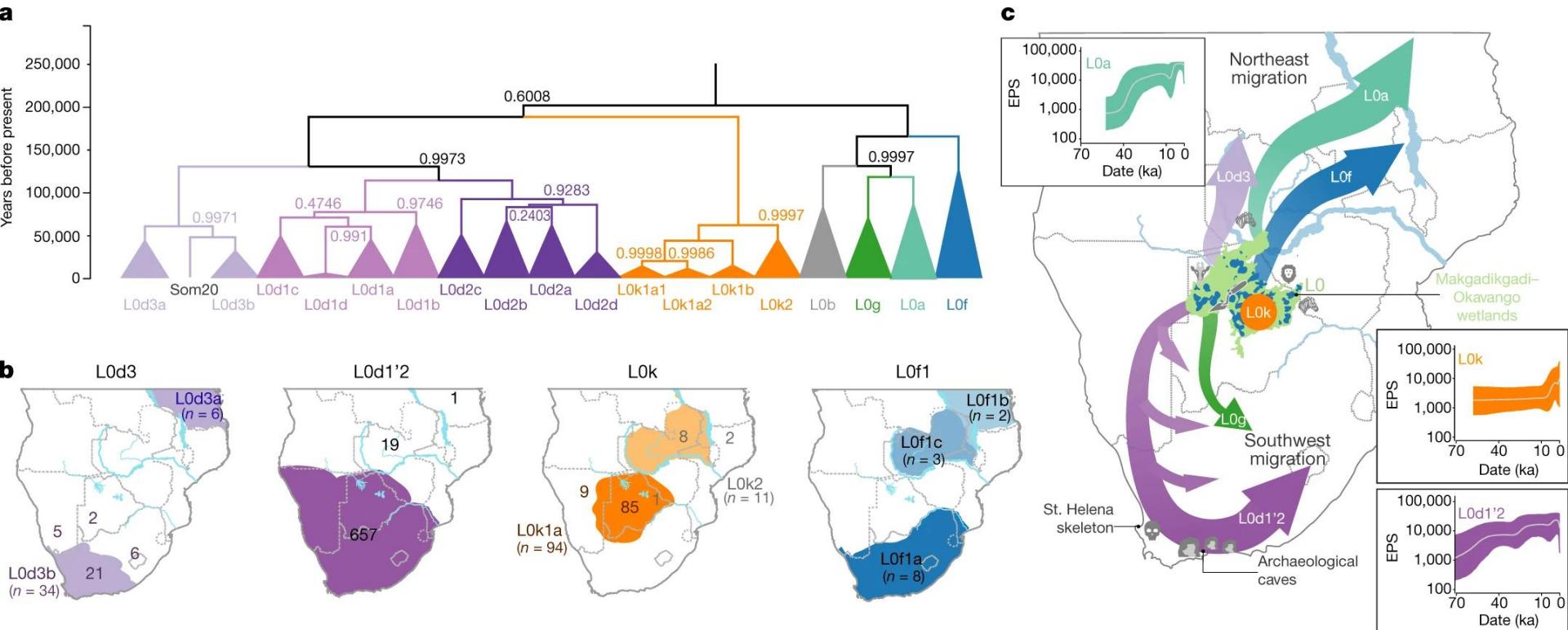
# Evolution everywhere

ISSUE 3185 | MAGAZINE COVER DATE: 7 July 2018



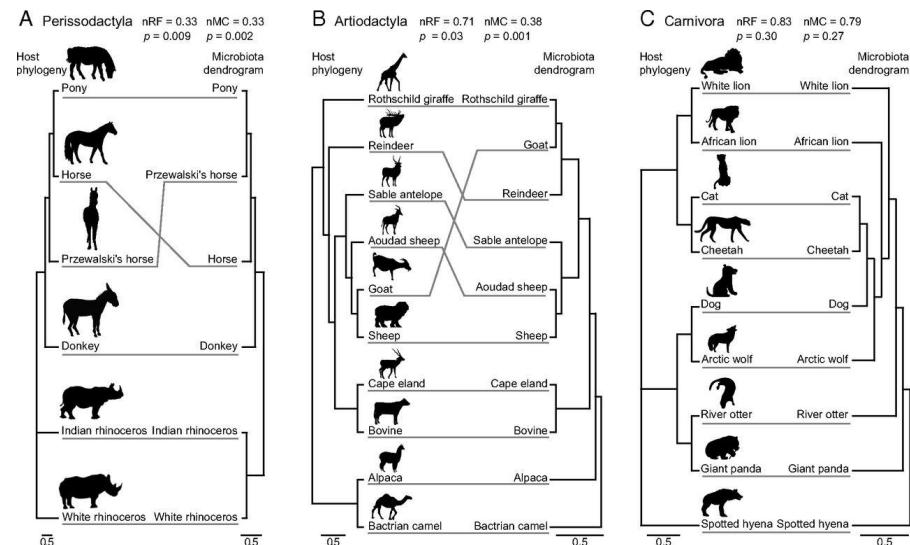
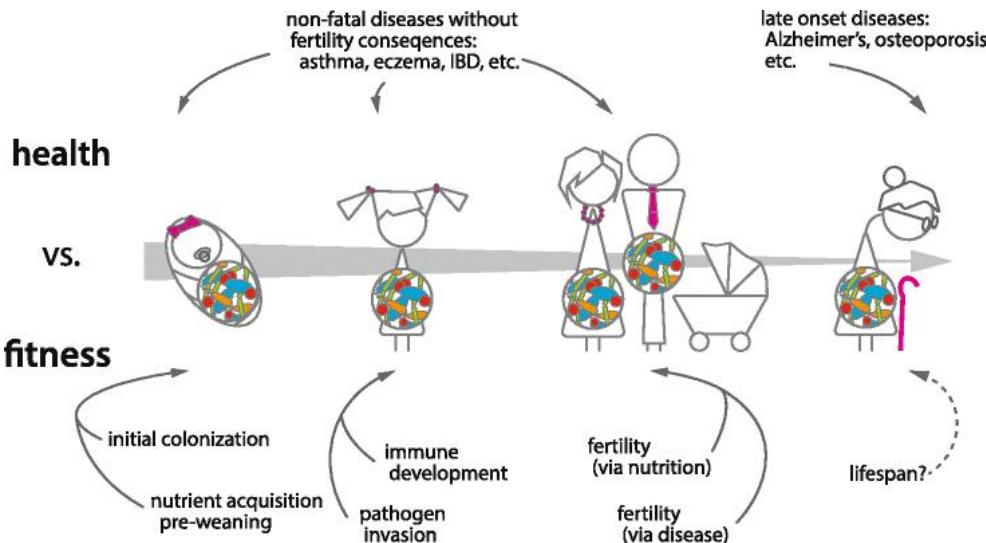
# Evolution everywhere

Out of “homeland”?

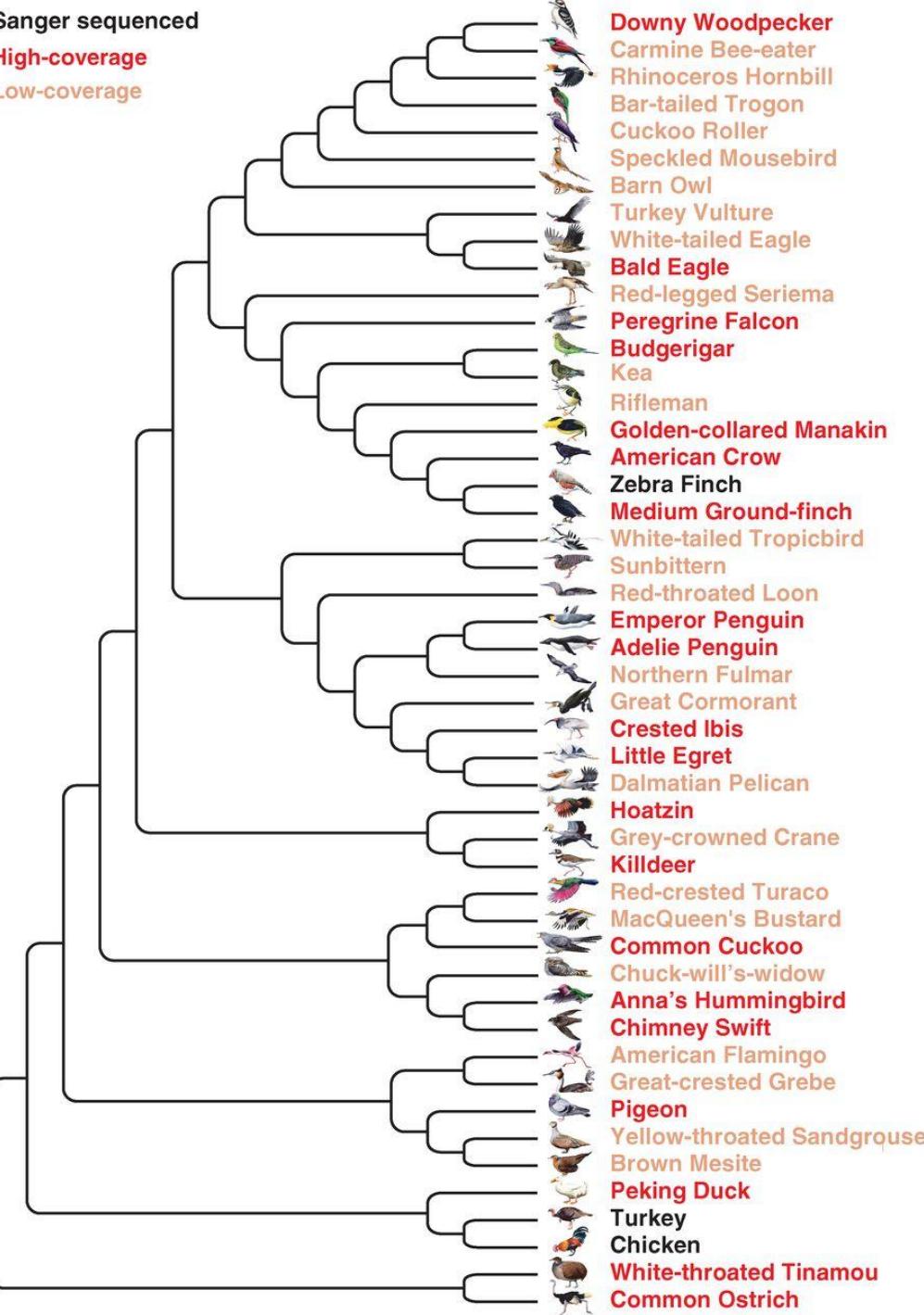
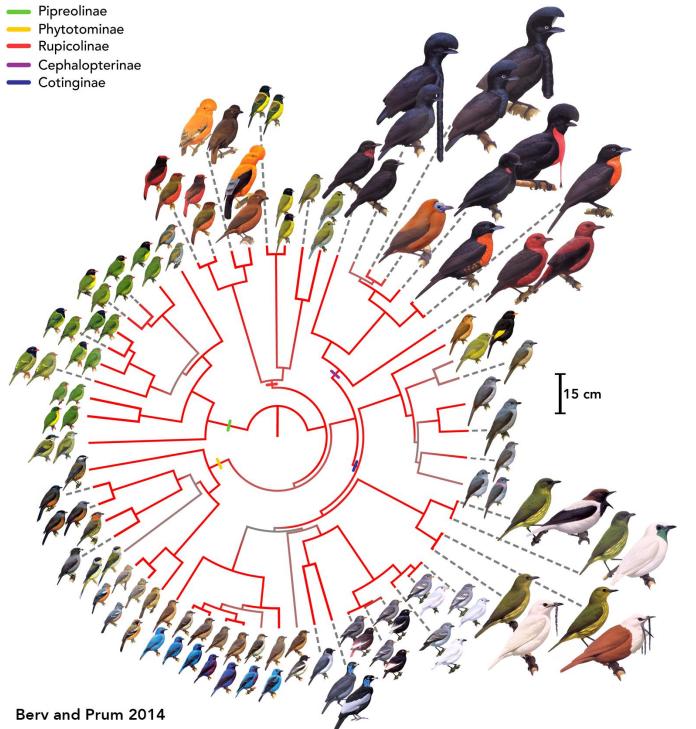


# Evolution everywhere

Microbiome evolution also exists



# Phylogenetic Tree

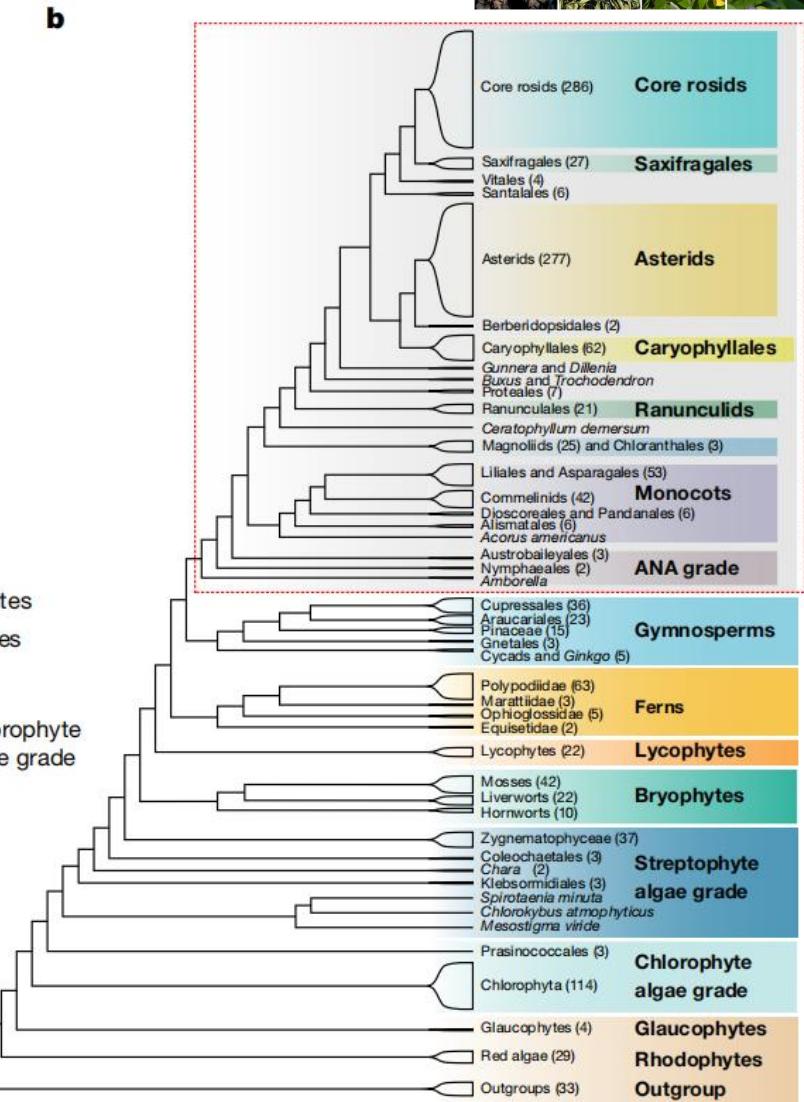
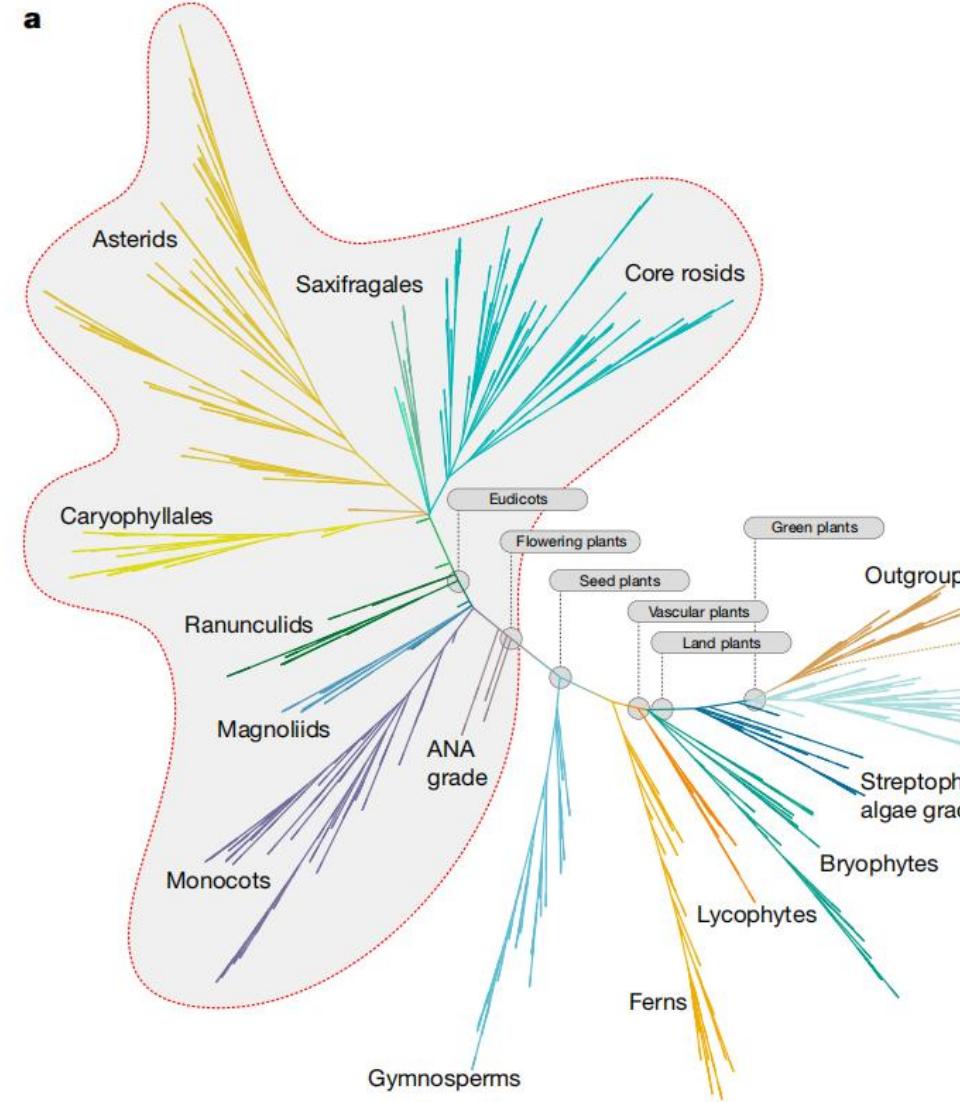


Reference:  
<https://b10k.genomics.cn/>

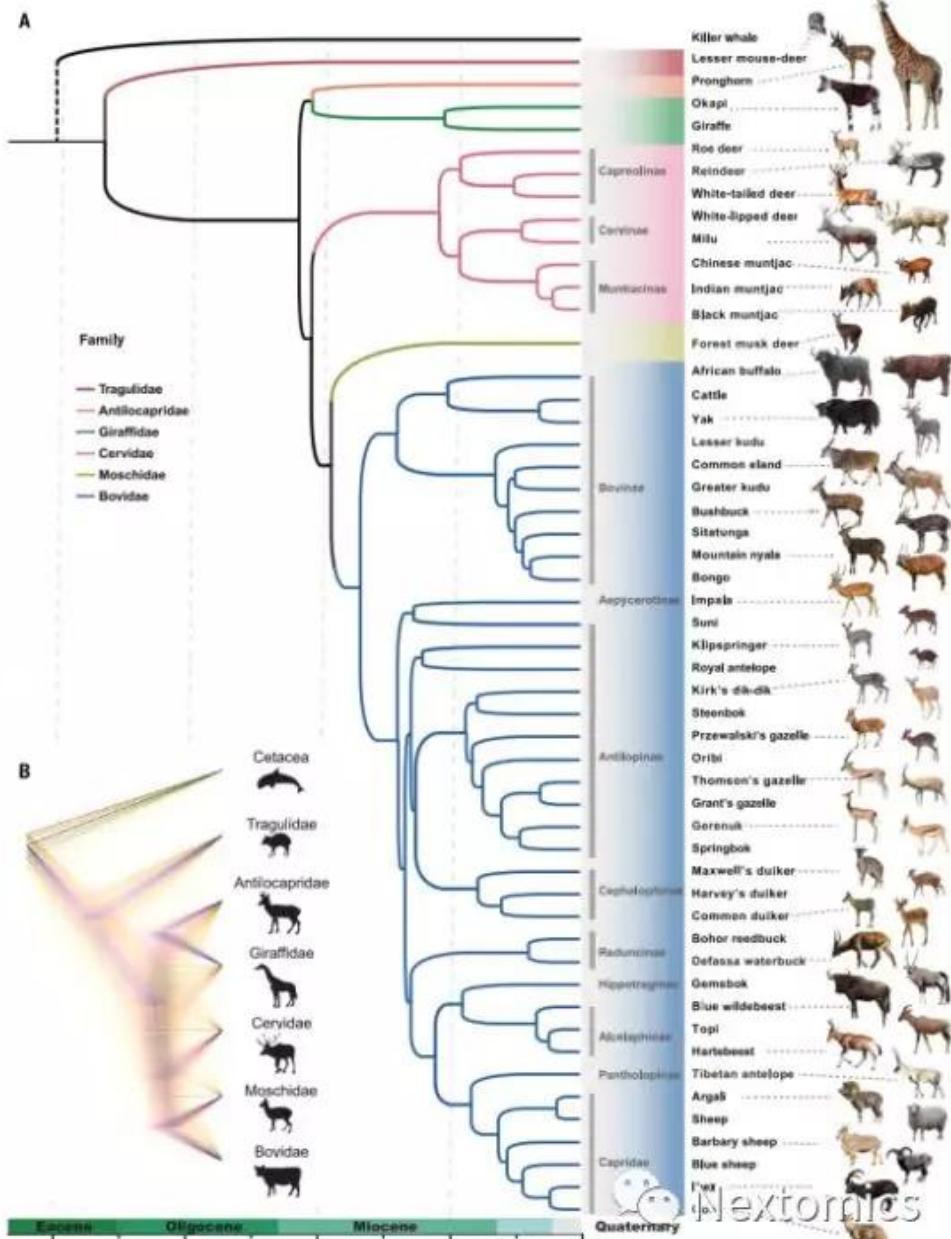
# Phylogenetic Tree

## Reference:

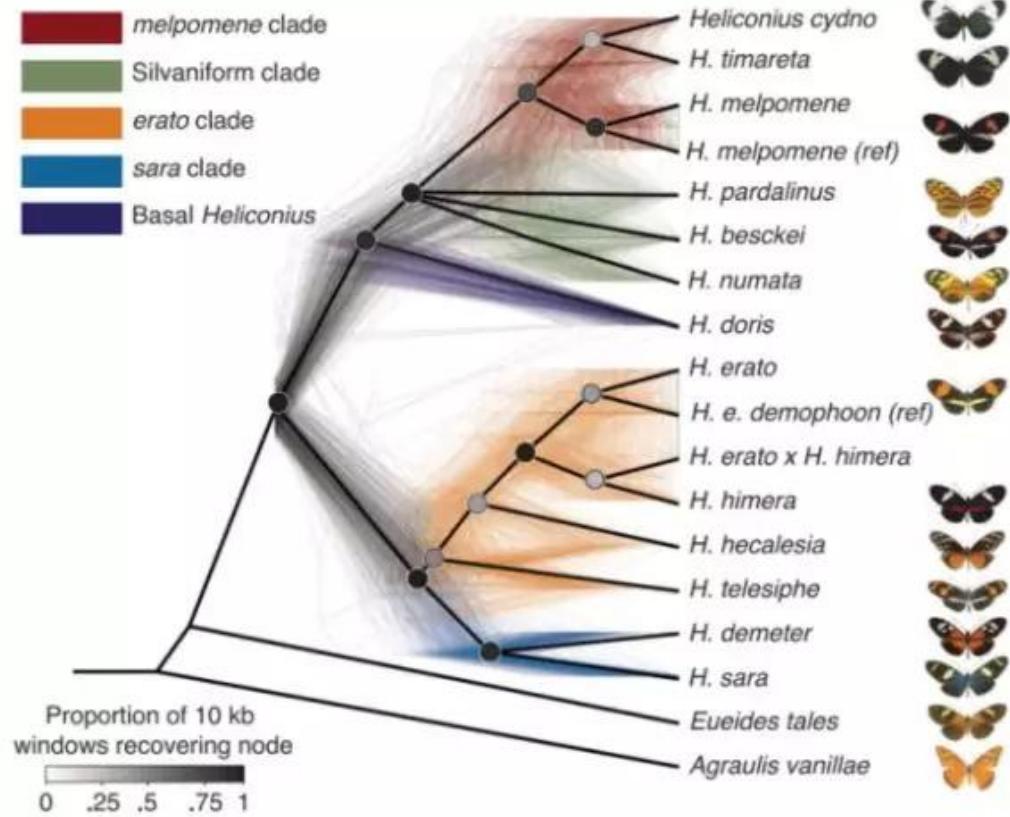
One thousand plant transcriptomes and the phylogenomics of green plants,  
*Nature*, 2019



# Phylogenetic Tree

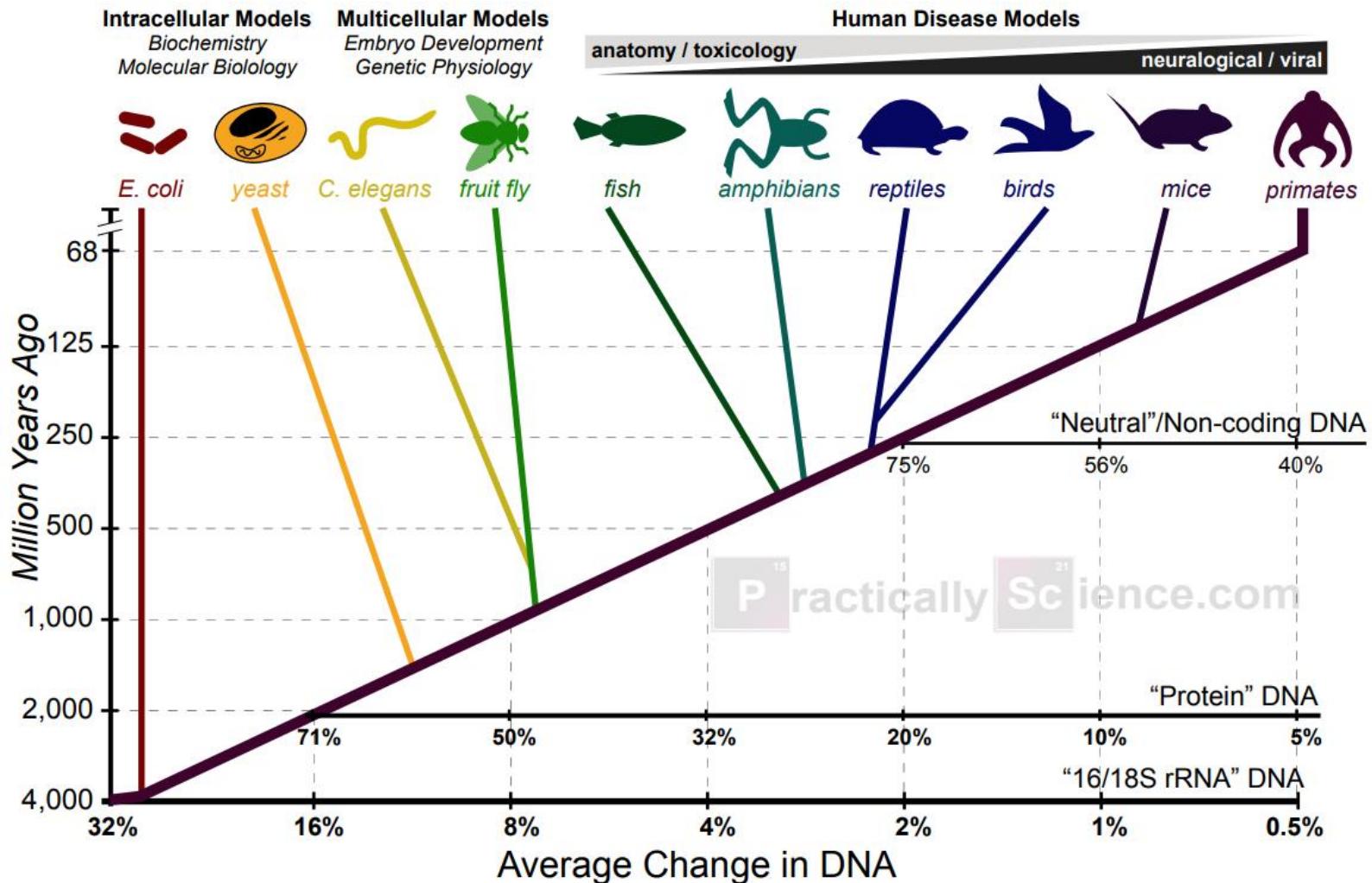


# Phylogenetic Tree

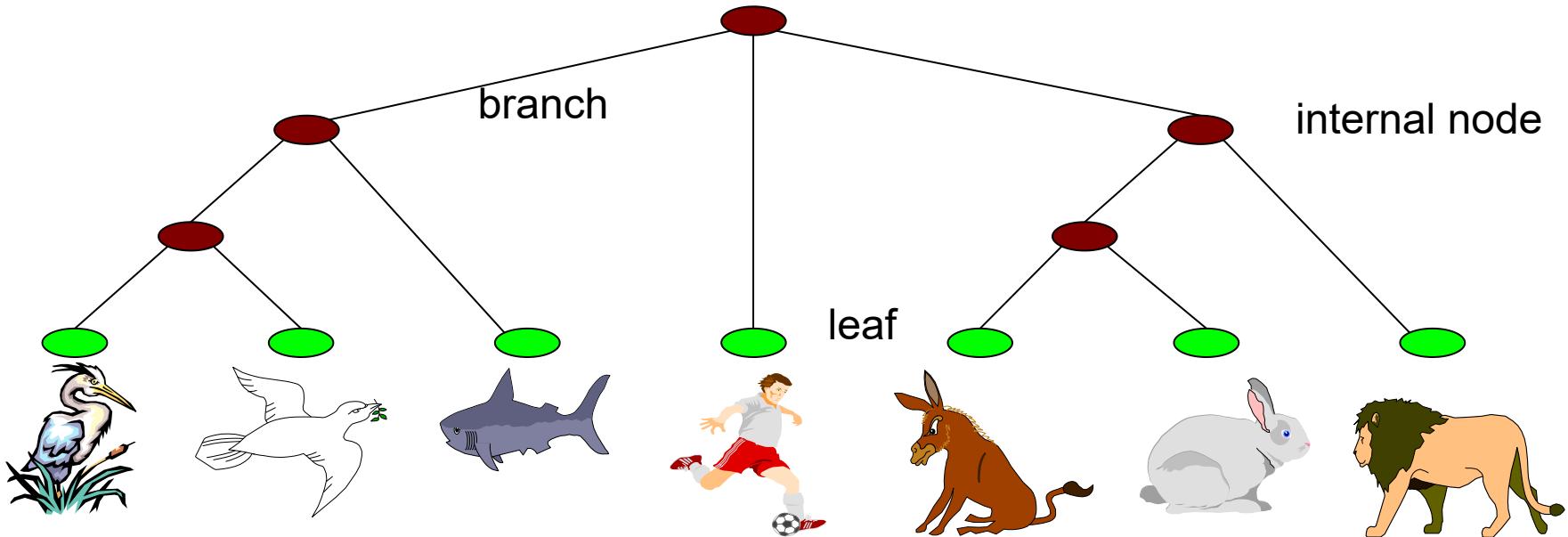


# Phylogenetic Tree

## Evolution of Model Organisms and the DNA Molecular Clock



# Phylogenetic Tree



- Topology: bifurcating
  - Leaves -  $1 \dots N$
  - Internal nodes  $N+1 \dots 2N-2$
- Branch length

Reference:  
<https://itol.embl.de/>

# 构建进化树算法

- 基于距离的构建方法：
  - UPGMA (Unweighted pair group method with arithmetic mean, 平均连接聚类法)
  - ME (Minimum Evolution, 最小进化法)
  - NJ (Neighbor-Joining, 邻接法)
- 基于特征的构建方法：
  - 最大简约法 (MP法)
  - 最大似然法 (ML法)
  - 进化简约法 (EP法)
  - 相容性方法

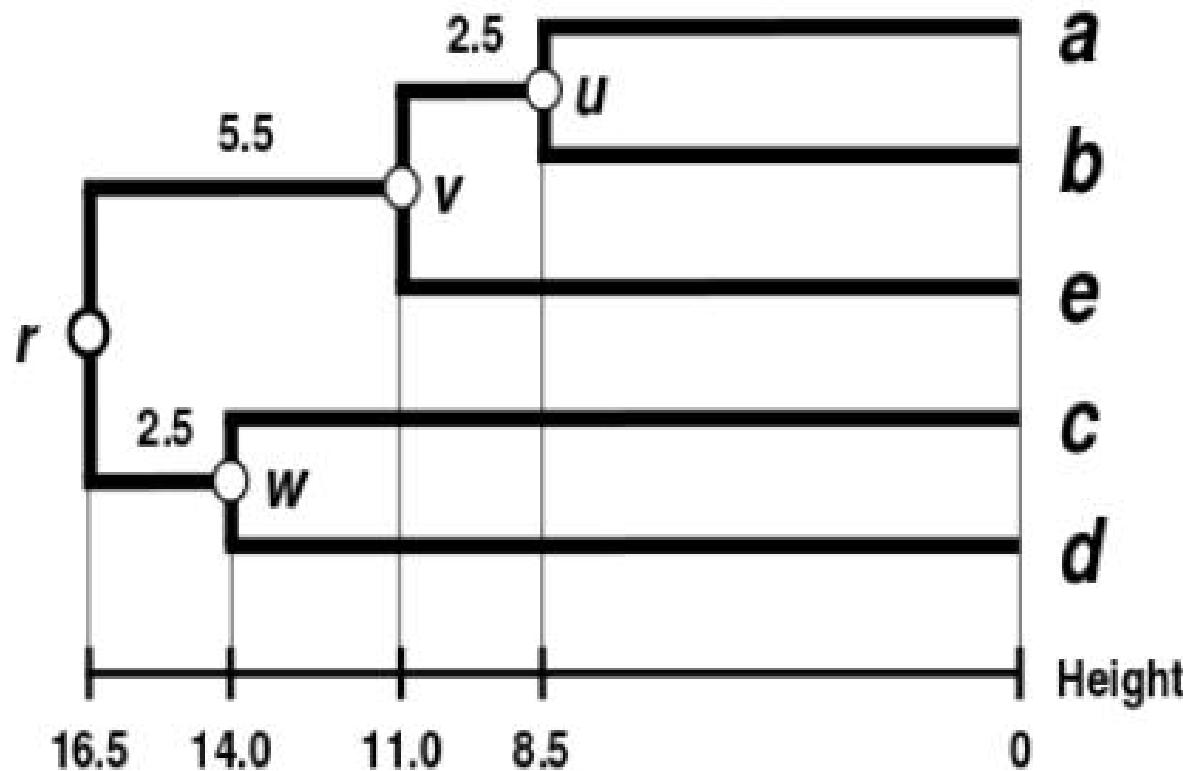
# 基于距离的构建方法

- UPGMA

- ①以已求得的距离系数,所有比较的分类单元的成对距离构成一个 $t \times t$ 方阵,即建立一个距离矩阵M。
- ②对于一个给定的距离矩阵,寻求最小距离值 $D_{pq}$ 。
- ③定义类群p和q之间的分支深度 $L_{pq}=D_{pq}/2$ 。
- ④若p和q是最后一个类群,侧聚类过程完成,否则合并p和q成一个新类群r。
- ⑤定义并计算新类群r到其他各类群i( $i \neq p$ 和q)的距离 $D_{ri}=(D_{pi}+D_{qi})/2$ 。
- ⑥回到第一步,在矩阵中消除p和q,加入新类群r,矩阵减少一阶,重复进行直至达到最后归群。

# 基于距离的构建方法

- UPGMA



# 基于距离的构建方法

## • Neighbor-Joining

① 对于给定距离矩阵中的每一端结*i*,用下式计算与其它分类单元之间的净趋异量( $R_i$ ) ( $t$ :矩阵中的分类单元数)

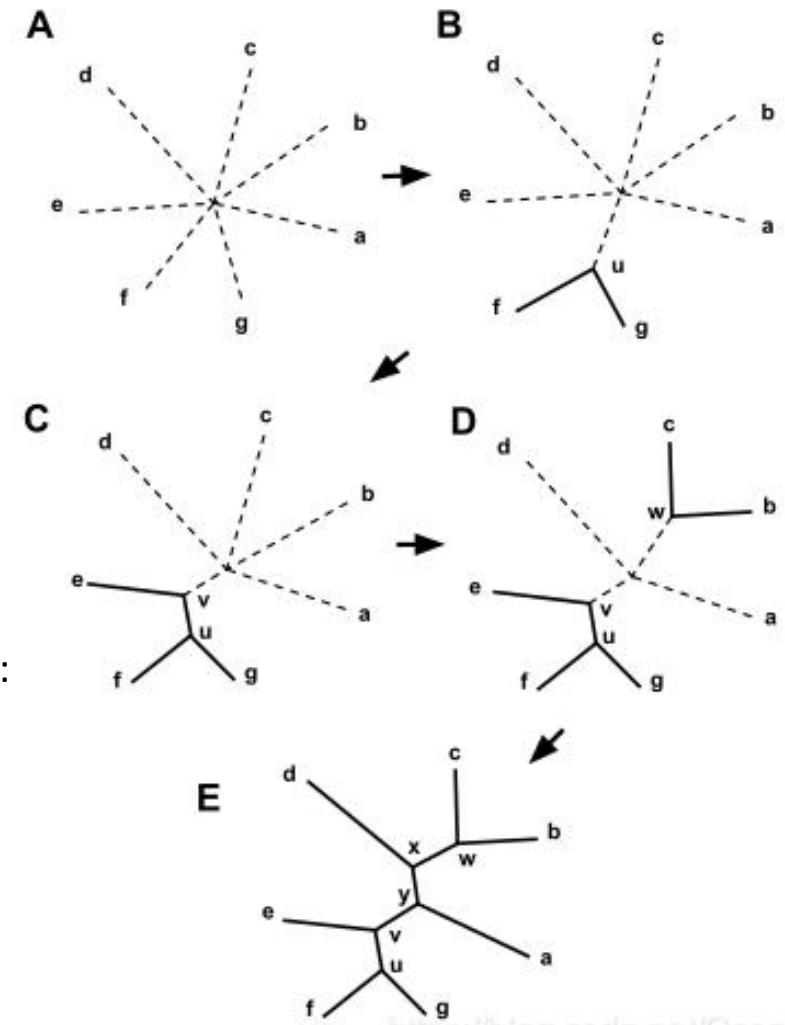
② 建立一个速率校正距离矩阵 $M$ ,其元素由下式确定:

③ 定义一个新节点,u的三个分支分别与节点*i,j*和树的其余部分相连,并且 $D_{ij}$ 为矩阵中距离最小者,u到节点*i*和*j*的分支长度定义为

④ 定义到树的其它节点*k*( $k \neq i$ 和*j*外的所有节点)的距离:

⑤ 从距离矩阵中删除*i*和*j*的距离,矩阵减少一阶。

⑥ 如果矩阵仍然多于两个的节点,重复第①-⑤步,否则删除最外两个节点的分支长度来确定外,树上其余节点都确定,最后是剩余的2个的分支长度 $S_y=D_{ij}$

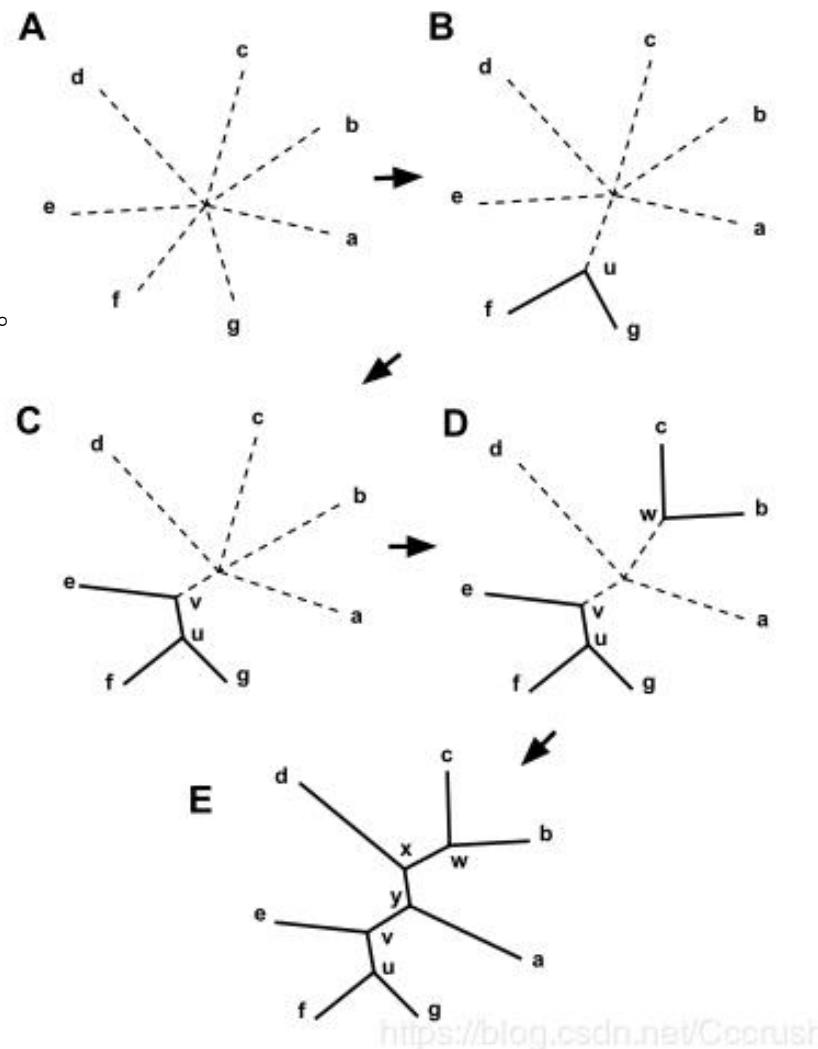


# 基于距离的构建方法

## • Neighbor-Joining

在概念上与UPGMA法相同,但是有四点区别

- a. NJ法不要求距离符合超度量特性,但要求数据应非常接近或符合叠加性条件,即该方法要求对距离进行校正。
- b. 邻接法在成聚过程中连接的是分类单元之间的节点(node),而不是分类单元本身。
- c. NJ法中原始距离数据用于估算系统树上所有端结分类单元之间的距离矩阵,校正后的距离用于确定节点之间的连接顺序。
- d. 在重建系统发育树时,NJ法取消了UPGMA法所做的假定,认为在此进化分支上,发生趋异的次数可以不同。



# 构建进化树算法

- 基于距离的构建方法：
  - UPGMA (Unweighted pair group method with arithmetic mean, 平均连接聚类法)
  - ME (Minimum Evolution, 最小进化法)
  - NJ (Neighbor-Joining, 邻接法)
- 基于特征的构建方法：
  - 最大简约法 (MP法)
  - 最大似然法 (ML法)
  - 进化简约法 (EP法)
  - 相容性方法

# 最大简约法（Maximum Parsimony）

最大简约法的理论基础是奥卡姆（Ockham）哲学原则，这个原则认为：解释一个过程的最好理论是所需假设数目最少的那一个。

方法：

计算所有可能的拓扑结构；

计算出所需替代数最小的那个拓扑结构，作为最优树。

Occam's Razor

The simplest explanation is  
usually the correct one.

# occam's razor method

## The Role of Ockham's Razor: To Screen for Plurality at the Beginning of the Scientific Method

1. OBSERVATION
2. NECESSITY
3. INTELLIGENCE/AGGREGATION OF DATA (The Three Key Questions)
4. CONSTRUCT FORMULATION
5. SPONSORSHIP/PEER INPUT (Ockham's Razor)



Ockham's  
Razor

6. HYPOTHESIS DEVELOPMENT
7. PREDICTIVE TESTING
8. COMPETITIVE HYPOTHESES FRAMING (ASKING THE RIGHT QUESTION)

9. FALSIFICATION TESTING

10. HYPOTHESIS MODIFICATION

11. FALSIFICATION TESTING/REPEATABILITY

12. THEORY FORMULATION/REFINEMENT

13. PEER REVIEW (Community Vetting)

14. PUBLISH

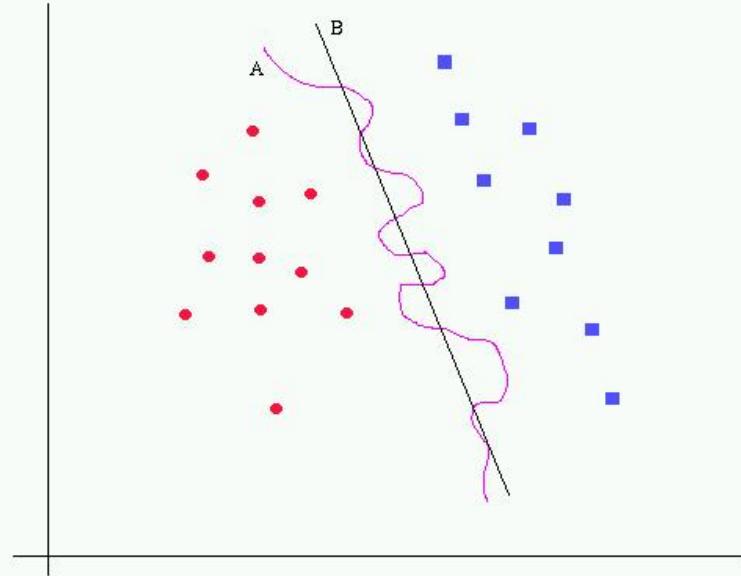
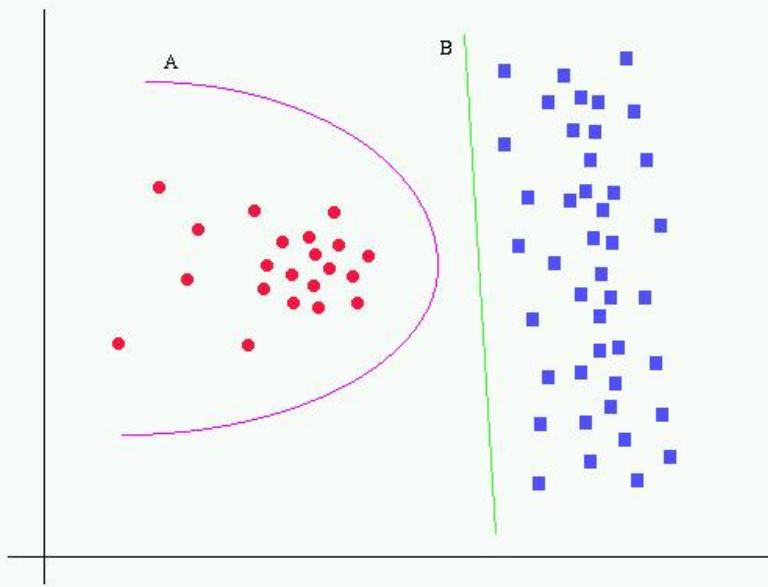
15. ACCEPTANCE

False Skeptics  
seek to block this step  
at all costs.

Plurality

Proof

# occam's razor method



# 最大简约法（Maximum Parsimony）

**依据** 基于奥卡姆（Ockham）哲学原则，这个原则认为：解释一个过程的最好理论是所需假设数目最少的那一个。

**方法** 计算所有可能的拓扑结构，计算出所需替代数最小的那个拓扑结构，作为最优树。

**特点** 用于分析如插入、缺失等序列。在分析序列上存在较多的回复突变或平行突变，而被检验的序列位点数又比较少的时候，最大简约法可能会给出一个不合理的或者错误的进化树推导结果。

# 最大似然法 (Maximum Likelihood)

ML法对所有可能的系统发育树都计算似然函数，似然函数值最大的那棵树即为最可能的系统发育树。

利用最大似然法来推断一组序列的系统发生树，需首先确定序列进化的模型，如Jukes—Cantor模型、Kimura二参数模型及一般二参数模型等。在进化模型选择合理的情况下，ML法是与进化事实吻合最好的建树算法。其缺点是计算强度非常大，极为耗时。

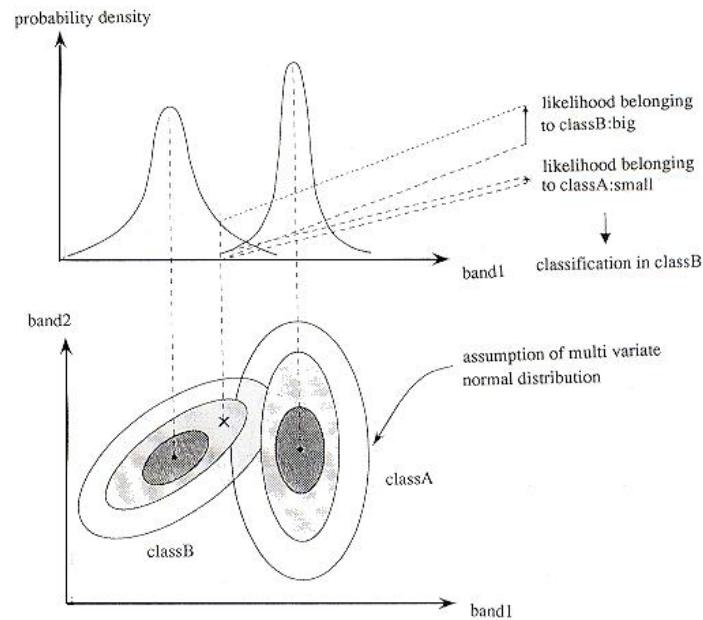


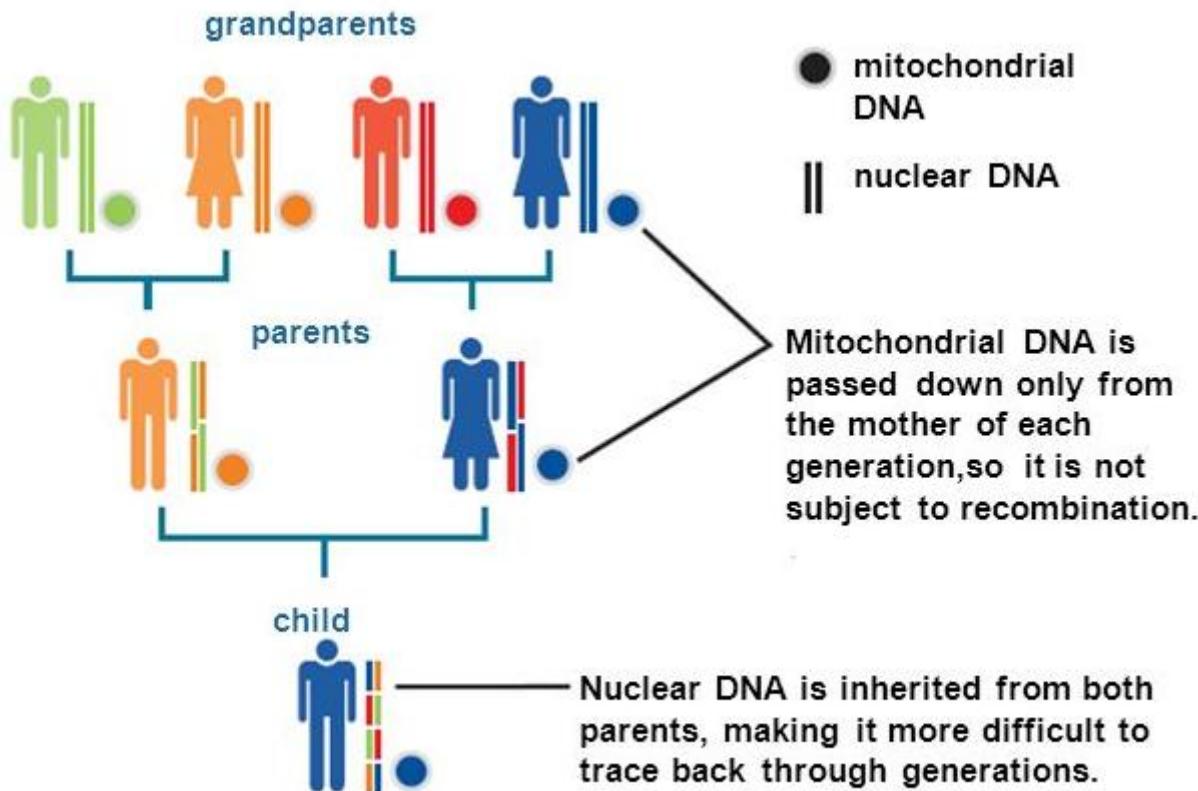
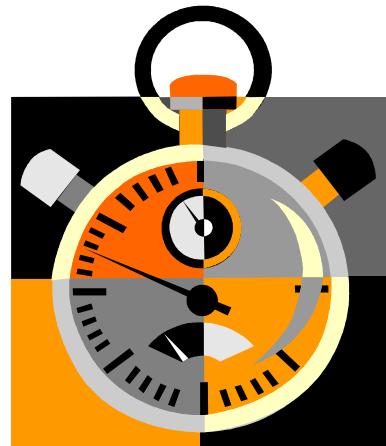
Figure 11.7.1 Concept of Maximum Likelihood Method

# Molecular Clock Hypothesis



- Amount of genetic difference between sequences is a function of time since separation.
- Rate of molecular change is constant (enough) to predict times of divergence

# Molecular Clock Hypothesis



# Likelihood of a Tree

- Given:
  - $n$  aligned sequences  $M = X_1, \dots, X_n$
  - A tree  $T$ , leaves labeled with  $X_1, \dots, X_n$
- Reconstruction  $t^*$ :
  - Labeling of internal nodes
  - Branch lengths

Goal: Find optimal reconstruction  $t^*$  : One maximizing the likelihood  $P(M | T, t^*)$

# Probabilistic Methods

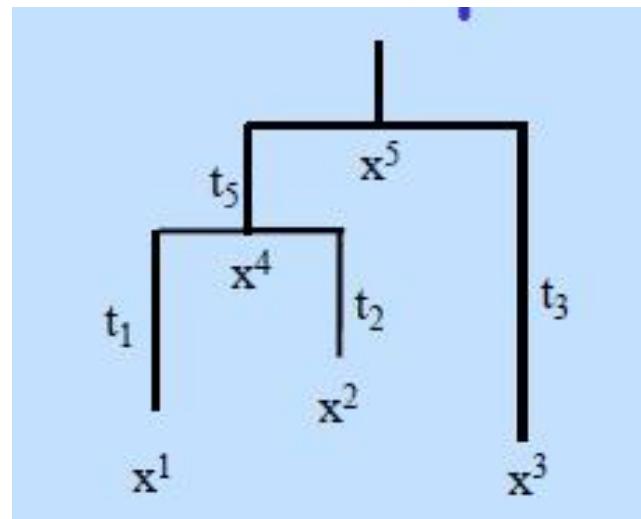
- The phylogenetic tree represents a generative probabilistic model (like HMMs) for the observed sequences.
- Background probabilities:  $q(a)$
- Mutation probabilities:  $P(a|b, t)$
- Models for evolutionary mutations
  - Jukes Cantor
  - Kimura 2-parameter model
- Such models are used to derive the probabilities

# Probabilistic Model

- Assumptions:
  - Each character is independent
  - The branching is a Markov process: The probability that a node  $x$  has a specific label is only a function of the parent node  $y$  and the branch length  $t$  between them
  - The probabilities  $P(x|y,t)$  are known

# Example

- Given the tree



$$\begin{aligned} & P(x_1, x_2, x_3, x_4, x_5 | T, t^*) \\ &= P(x_1 | x_4, t_1) P(x_2 | x_4, t_2) P(x_3 | x_5, t_3) P(x_4 | x_5, t_5) \end{aligned}$$

# Molecular Evolution

- Q: How can we model evolution on nucleotide level? (ignore gaps, focus on substitutions)
- A: Consider what happens at a specific position for small time interval  $\Delta t$
- $P(t)$  = vector of probabilities of {A,C,G,T} at time  $t$
  - $\mu_{AC}$  = rate of transition from A to C per unit time
  - $\mu_A = \mu_{AC} + \mu_{AG} + \mu_{AT}$  rate of transition out of A
  - $p_A(t+\Delta t) = p_A(t) - p_A(t) \mu_A \Delta t + p_C(t) \mu_{CA} \Delta t + \dots$

# Molecular Evolution

In matrix/vector notation, we get

$$P(t + \Delta t) = P(t) + QP(t)\Delta t$$

where  $Q$  is the substitution rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$

# Molecular Evolution

- This is a differential equation:

$$P'(t) = Q P(t)$$

- A substitution rate matrix  $Q$  implies a probability distribution over {A,C,G,T} at each position, including stationary (equilibrium) frequencies  $\pi_A, \pi_C, \pi_G, \pi_T$
- Each  $Q$  is an evolutionary model (some work better than others)

# Mutation Probabilities

$P(t)$  satisfy the following two property:

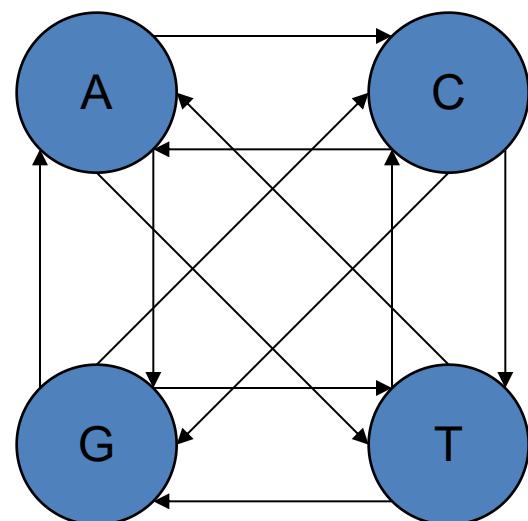
- **Lack of memory:**

$$- P_{a \rightarrow c}(t + t') = \sum_b P_{a \rightarrow b}(t) P_{b \rightarrow c}(t')$$

- **Reversibility:**

- Exist stationary probabilities  
 $\{P_a\}$  s.t.

$$P_a P_{a \rightarrow b}(t) = P_b P_{b \rightarrow a}(t)$$



# PAM矩阵

- Point accepted mutation (Dayhoff et al 1978)
- Given an tree of protein family, the frequence matrix  $A_{ab}$  counting the occurrence of an “a” in the ancestral sequence was replaced by a “b” in the descendant.
- Estimate the conditional probability  $p(b|a)$

$$P(b|a) = B_{a,b} = \frac{A_{ab}}{\sum_c A_{ac}}$$

# PAM矩阵

- Scaling B

$$C_{ab} = \sigma B_{ab}, C_{aa} = \sigma B_{aa} + (1 - \sigma)$$

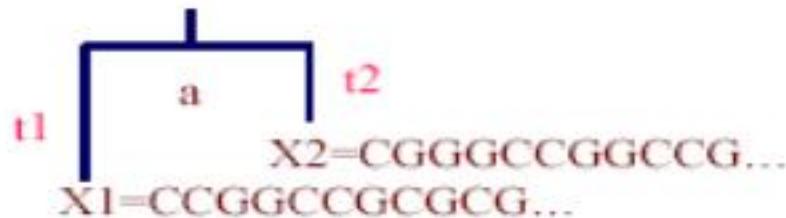
- Such that the expected number of substitution is 1%, i.e.

$$\sum_{ab} q_a q_b C_{ab} = 0.01$$

- Then the PAM(1) matrix is given by

$$S(1) = (C_{ab})$$

# Calculating the Likelihood for Ungapped Alignments



$$P(x^1, x^2 | T, t_1, t_2) = \prod_{n=1}^N P(x_n^1, x_n^2 | T, t_1, t_2)$$

$$P(x_n^1, x_n^2 | T, t_1, t_2) = \sum_a q_a P(x_n^1 | a, t_1) P(x_n^2 | a, t_2)$$

Assuming Jukes-Cantor model &  $q_C = q_G = q_A = q_T = \frac{1}{4}$  :

$$P(C, C | T, t_1, t_2) = q_C r_{t_1} r_{t_2} + q_G s_{t_1} s_{t_2} + q_A s_{t_1} s_{t_2} + q_T s_{t_1} s_{t_2} = \frac{1}{4}(r_{t_1} r_{t_2} + 3s_{t_1} s_{t_2})$$

$$P(C, G | T, t) = P(G, C | T, t) = \frac{1}{4}(r_{t_1} s_{t_2} + s_{t_1} r_{t_2} + 2s_{t_1} s_{t_2})$$

$$\Rightarrow P(x^1, x^2 | T, t_1, t_2) = 16^{-(n1+n2)} \left(1 + 3e^{-4\alpha(t_1+t_2)}\right)^{n1} \left(1 - e^{-4\alpha(t_1+t_2)}\right)^{n2}$$

where n1=matches, n2=mismatches

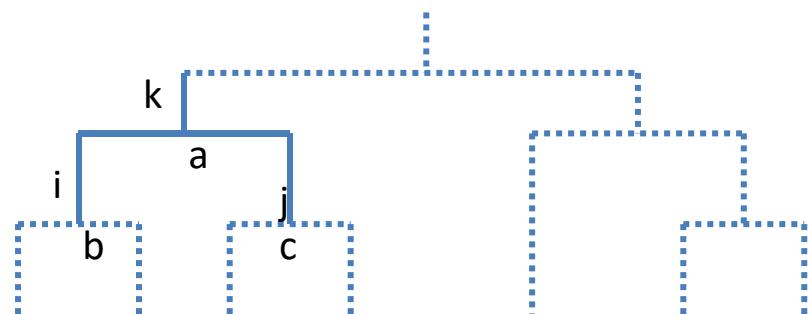
# Calculating the Likelihood for Ungapped Alignments

- $n$  sequences of length  $N$ , site  $u=1 \dots N$
- Given a rooted tree contains  $2n - 1$  nodes,  $1 \dots n$  being the leaf nodes,  $n+1 \dots 2n-1$  non-leaf, tree lengths  $t_1, \dots, t_{2n-1}$ .
- Let  $a(i)$  denote the ancestor of node  $a^i$

$$P(x^1, \dots, x^n | T, t) = \prod_{u=1}^N P(x_u^1, \dots, x_u^n | T, t)$$
$$P(x_u^1, \dots, x_u^n | T, t) = \sum_{a^{n+1}, \dots, a^{2n-1}} q_{a^{2n-1}} \prod_{i=n+1}^{2n-2} P(a^i | a^{\alpha(i)}, t_i)$$
$$\times \prod_{i=1}^n P(x_u^i | a^{\alpha(i)}, t_i)$$

# Felsenstein's Recursive Algorithm

- Let  $P(L_k|a)$  denote the probability of all the leafs below node  $k$  given that the residue at  $k$  is  $a$ .
- Then we compute  $P(L_k|a)$  from the probabilities  $P(L_i|b)$  and  $P(L_j|c)$  for all  $b$  and  $c$ , where  $i$  and  $j$  are the daughter nodes of  $k$ .



# Felsenstein's Recursive Algorithm

- Initialization: set  $k=2n-1$
- Recursion: Compute  $P(L_k | a)$  for all  $a$  as follows:
  - If  $k$  is leaf node:  $P(L_k | a)=1$  only if  $a = x_u^k$ .
  - If  $k$  is not a leaf node:
    - Compute  $P(L_i | a)$ ,  $P(L_j | a)$  for all  $a$  at the daughter nodes  $i, j$ , and set  $P(L_k | a) = \sum_{bc} P(b|a, t_i)P(L_i|b)P(c|a, t_j)P(L_j|c)$
- Termination: Likelihood at site  $u$ ,

$$P(x_u | T, t) = \sum_a P(L_{2n-1} | a) q_a$$

# Reversibility & Independence of Root Position

- The score of the optimal tree is independent of the root position if and only if:
  - the substitution matrix is **multiplicative**
  - the substitution matrix is **reversible**
- A substitution matrix is reversible if for all a,b and t:
$$P(b|a, t)q_a = P(a|b, t)q_b$$

# Maximum Likelihood (ML)

- Score each tree by
  - Assumption of independent positions “m”
- Branch lengths  $t$  can be optimized
  - Gradient Ascent
  - EM
- We look for the highest scoring tree
  - Exhaustive
  - Sampling methods (**Metropolis**)

# Computational Problem

- Such procedures are computationally expensive!
- Computation of optimal parameters, per candidate, requires non-trivial optimization step.
- Spend non-negligible computation on a candidate, even if it is a low scoring one.
- In practice, such learning procedures can only consider small sets of candidate structures

# 最大似然法 (Maximum Likelihood)

**基本思想** 当从模型总体随机抽取n组样本观测值后，最合理的参数估计量应该使得从模型中抽取该n组样本观测值的概率最大。

**方法** 选取一个特定的替代模型来分析给定的一组序列数据，使得获得的每一个拓扑结构的似然率都为最大值，然后再挑出其中似然率最大的拓扑结构作为最优树（所以分析时间比较长）

**特点** 最大似然法具有很好的统计学理论基础，是一个比较成熟的统计学方法。选择合理的模型后，最大似然法可以推导出一个效果很好的进化树结果。

# 构建进化树算法

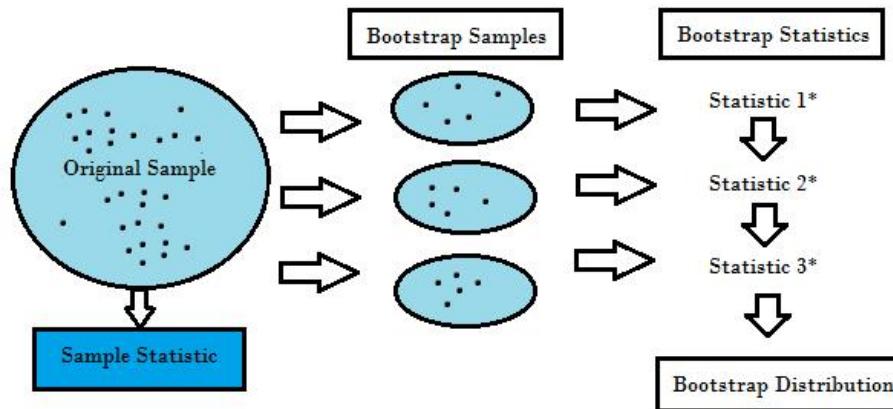
- 如果序列的相似性较高，各种方法都会得到不错的结果，模型间的差别也不大.
- 若有合适的分子进化模型可供选择，用最大似然法构树获得的结果较好.
- 对于近缘物种序列，通常情况下使用最大简约法.
- 而对于远缘物种序列，一般使用邻接法或最大似然法.
- 对于相似度很低的序列，邻接法往往出现长枝吸引(branch attraction)现象，有时严重干扰进化树的构建.

邻接法和最大似然法是需要选择模型的:

蛋白质序列的构树模型一般选择Poisson correction(泊松修正).

核酸序列的构树模型一般选择Kimura 2-parameter (Kimura-2参数).

# 构建进化树算法



在重建进化树过程中，均需选择**bootstrap**进行树的检验：

- 一般**bootstrap**的值>70，则认为重建的进化树较为可靠。
- 如果**bootstrap**的值太低，则有可能进化树的拓扑结构有错误，进化树是不可靠的。
- 一般推荐用两种以上不同的方法构建进化树，如果所得到的进化树类似，且**bootstrap**值总体较高，则得到的结果较为可靠。
- 通常情况下，只要选择了合适的方法和模型，构出的树均是有意义的，研究者可根据自己研究的需要选择最佳的树进行分析。

# 构建分子进化树相关的软件

## 软件 网址 说明

ClustalX	<a href="http://bips.u-strasbg.fr/fr/Documentation/ClustalX/">http://bips.u-strasbg.fr/fr/Documentation/ClustalX/</a>
ClustalW	<a href="http://www.cf.ac.uk/biosi/research/sequence-analysis/clustalw.html">http://www.cf.ac.uk/biosi/research/sequence-analysis/clustalw.html</a>
GeneDoc	<a href="http://www.psc.edu/biomed/genedoc/">http://www.psc.edu/biomed/genedoc/</a>
BioEdit	<a href="http://www.mbio.ncsu.edu/BioEdit/bioedit.html">http://www.mbio.ncsu.edu/BioEdit/bioedit.html</a>
MEGA	<a href="http://www.megasoftware.net/">http://www.megasoftware.net/</a>
PAUP	<a href="http://paup.csit.fsu.edu/">http://paup.csit.fsu.edu/</a>
PHYLIP	<a href="http://evolution.genetics.washington.edu/phylip.html">http://evolution.genetics.washington.edu/phylip.html</a>
PHYML	<a href="http://atgc.lirmm.fr/phym/">http://atgc.lirmm.fr/phym/</a>
PAML	<a href="http://abacus.gene.ucl.ac.uk/software/paml.html">http://abacus.gene.ucl.ac.uk/software/paml.html</a>
Tree-puzzle	<a href="http://www.tree-puzzle.de/">http://www.tree-puzzle.de/</a>
MrBayes	<a href="http://mrbayes.csit.fsu.edu/">http://mrbayes.csit.fsu.edu/</a>
MAC5	<a href="http://www.agapow.net/software/mac5/">http://www.agapow.net/software/mac5/</a>
TreeView	<a href="http://taxonomy.zoology.gla.ac.uk/rod/treeview.html">http://taxonomy.zoology.gla.ac.uk/rod/treeview.html</a>

图形化的多序列比对工具

命令行格式的多序列比对工具

多序列比对结果的美化工具

序列分析的综合工具

图形化、集成进化分析工具，不包括ML

商业软件，集成的进化分析工具

免费的、集成的进化分析工具

最快的ML建树工具

ML建树工具

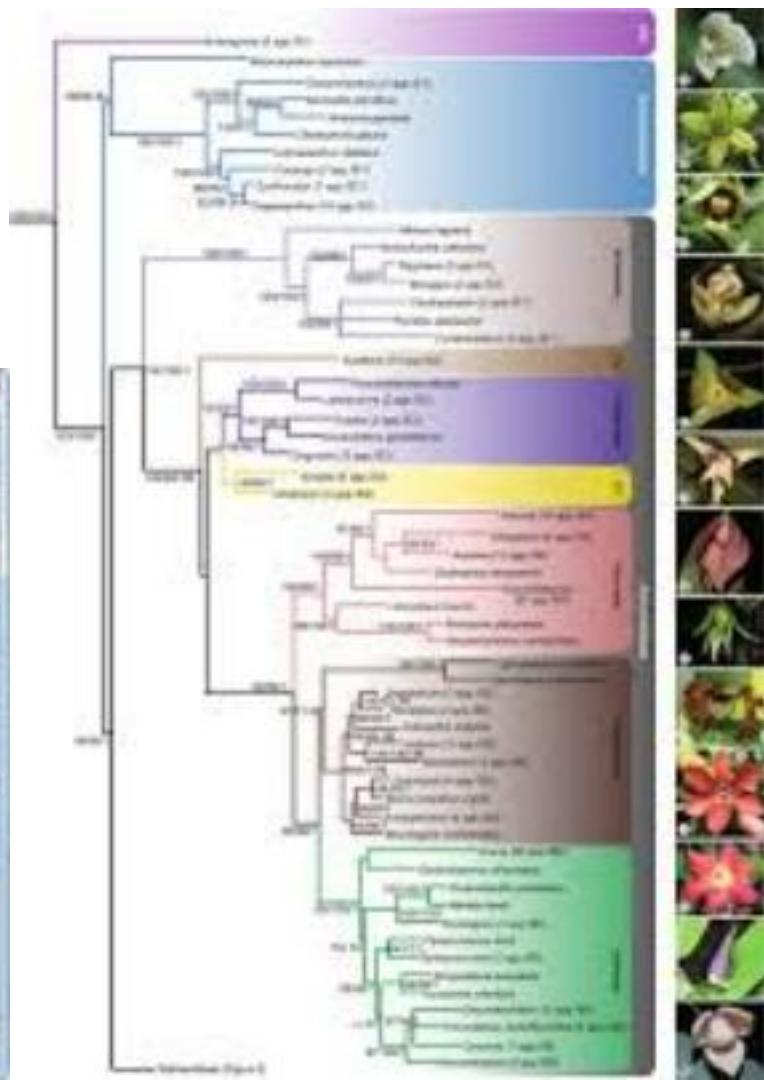
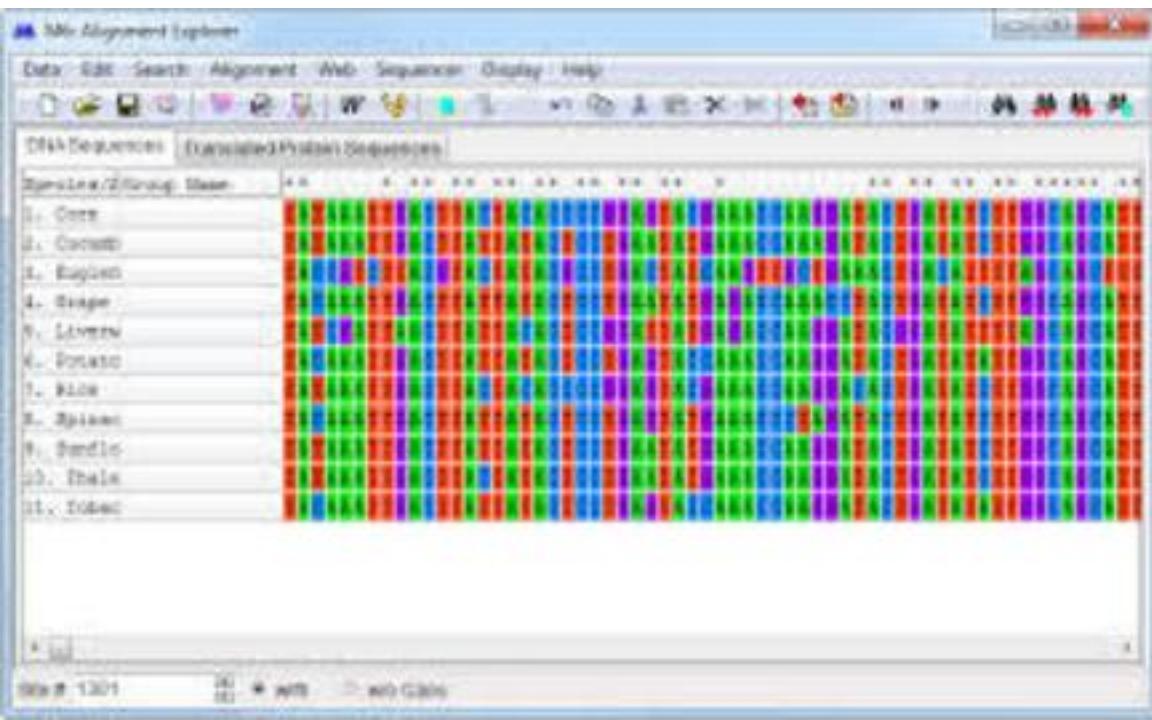
较快的ML建树工具

基于贝叶斯方法的建树工具

基于贝叶斯方法的建树工具

进化树显示工具

# MEGA: 建树平台



# FastTree

## 2

# FastTree: 建树平台

### Speed

Computing maximum-likelihood trees				
	# Distinct Sequences	#Positions	Settings	FastTree 2.0.0
Alignment				Hours
Efflux permeases (COG2814)	8,362	394 a.a.	JTT+CAT	0.25 0.35
ABC transporters (PF00005)	39,092	214 a.a.	JTT+CAT	1.0 0.96
16S ribosomal RNAs, distinct families	15,011	1,287 nt.	GTR+CAT	0.66 0.56
16S ribosomal RNAs, distinct families	15,011	1,287 nt.	JC+CAT	0.49 0.36
16S ribosomal RNAs	237,882	1,287 nt.	JC+CAT, -fastest	21.8 5.8

All of the timings are on a single CPU. The FastTree times include the [SH-like local support values](#). For huge alignments, FastTree 2.1 with -fastest is about twice as fast as 2.0, and the multi-threaded alignment I ran RAxML 6 with the fast hill-climbing option (not RAxML 7), and I ran PhyML was run with the fastest settings (no variation in rates across sites and no SPR moves).

In theory, FastTree takes  $O(N L a + N^{1.5})$  space and  $O(N^{1.5} \log(N) L a)$  time, where  $N$  is the number of unique sequences,  $L$  is the width of the alignment, and  $a$  is the size of the alphabet. While time complexity are dominated by initializing the top-hits lists and maintaining them during the neighbor-joining phase. The minimum-evolution NNIs and SPRs take  $O(N L a)$  time per round trip ( $O(N \log(N) L a^2)$  time total), and  $O(N L a)$  space. Similarly, the local supports take  $O(N L a^2)$  time and  $O(N L a)$  space. In practice, the maximum likelihood NNIs are usually the [slowest step](#).

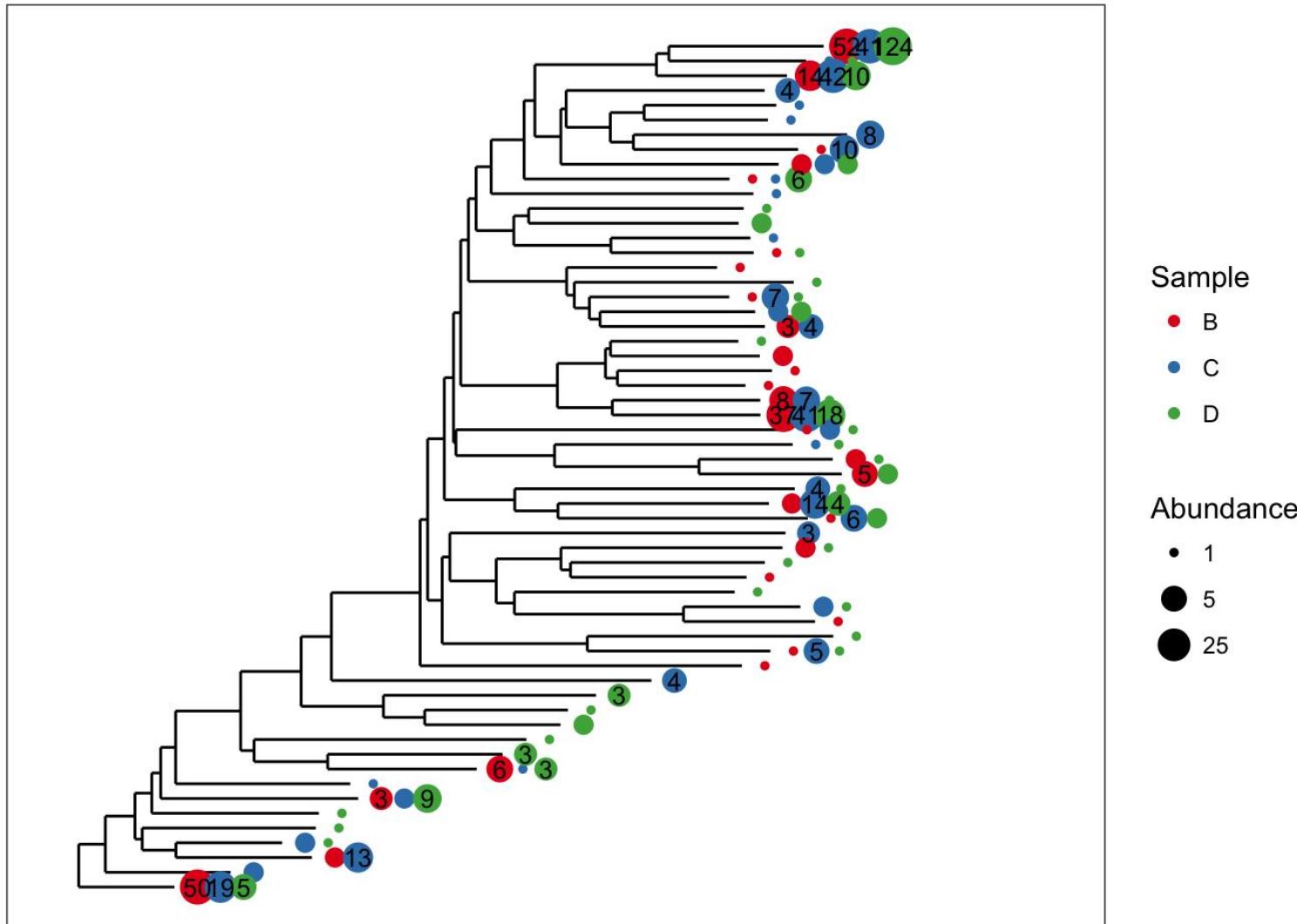
### Accuracy

FastTree is slightly more accurate than PhyML with NNI moves because it has a better starting tree (thanks to the minimum-evolution SPR moves). FastTree is much more accurate than most likelihood methods that do a more intensive search of topology space, such as PhyML with SPR moves or RAxML. However, for large alignments, the more accurate methods are orders-of-magnitude faster.

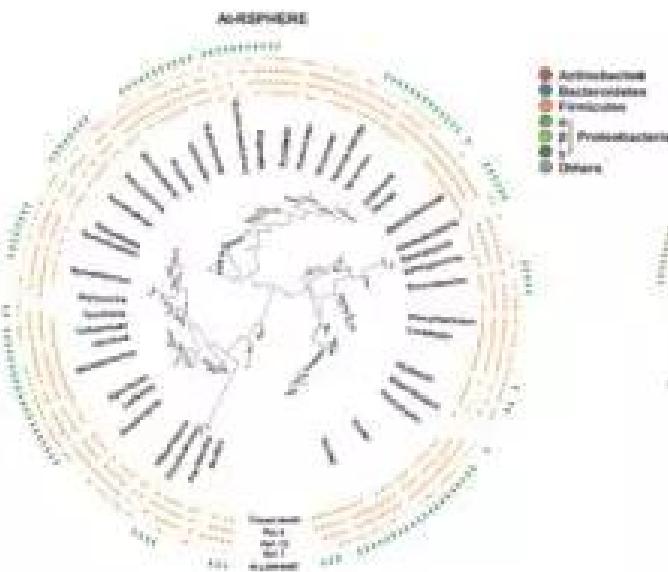
#### Topological accuracy for simulated alignments with varying numbers of sequences

	#Sequences	250	1,250	5,000	78,132
Type	a.a.	a.a.	a.a.	nt	
RAxML 7 (JTT+CAT + SPRs)		90.5%	88.4%	88.4%	--
PhyML 3.0 ( $\Gamma_4$ + SPRs)		89.9%	--	--	--
<b>FastTree 2.0.0 (JTT+CAT or JC+CAT)</b>		86.9%	83.7%	84.3%	92.1%
PhyML 3.0 ( $\Gamma_4$ , no SPR)		86.0%	--	--	--
PhyML 3.0 (no gamma, no SPR)		81.7%	80.1%	--	--
FastME 1.1 (log-corrected distances)		79.6%	77.7%	75.3%	--
BIONJ (max-lik. distances)		77.7%	73.7%	73.1%	--
Parsimony (RAxML)		76.8%	76.5%	69.4%	--
BIONJ (log-corrected distances)		76.6%	73.0%	72.3%	--
Neighbor-Joining (log-corrected distances)		76.0%	72.6%	71.6%	66.1%
<b>Clearcut 1.0.8</b> (log-corrected distances)		75.5%	72.3%	71.5%	58.1%

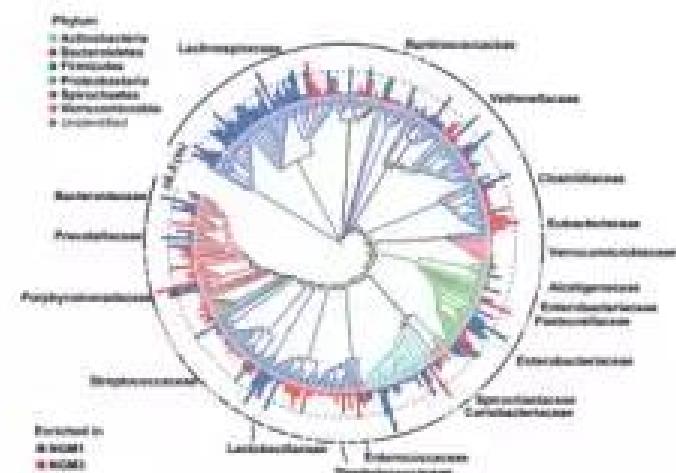
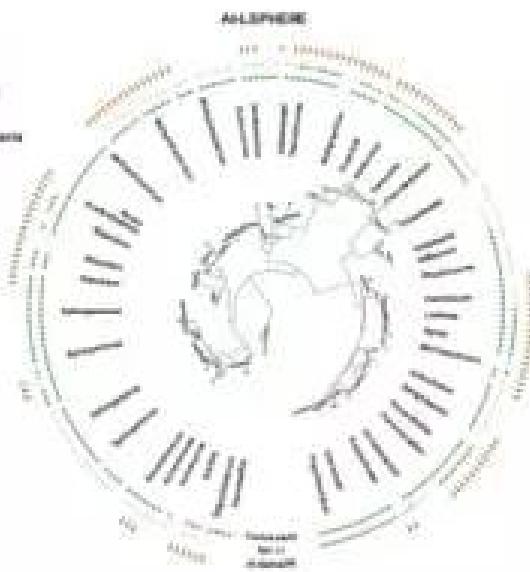
# ggplot: 物种进化关系分析



# iTOL: 物种进化关系分析

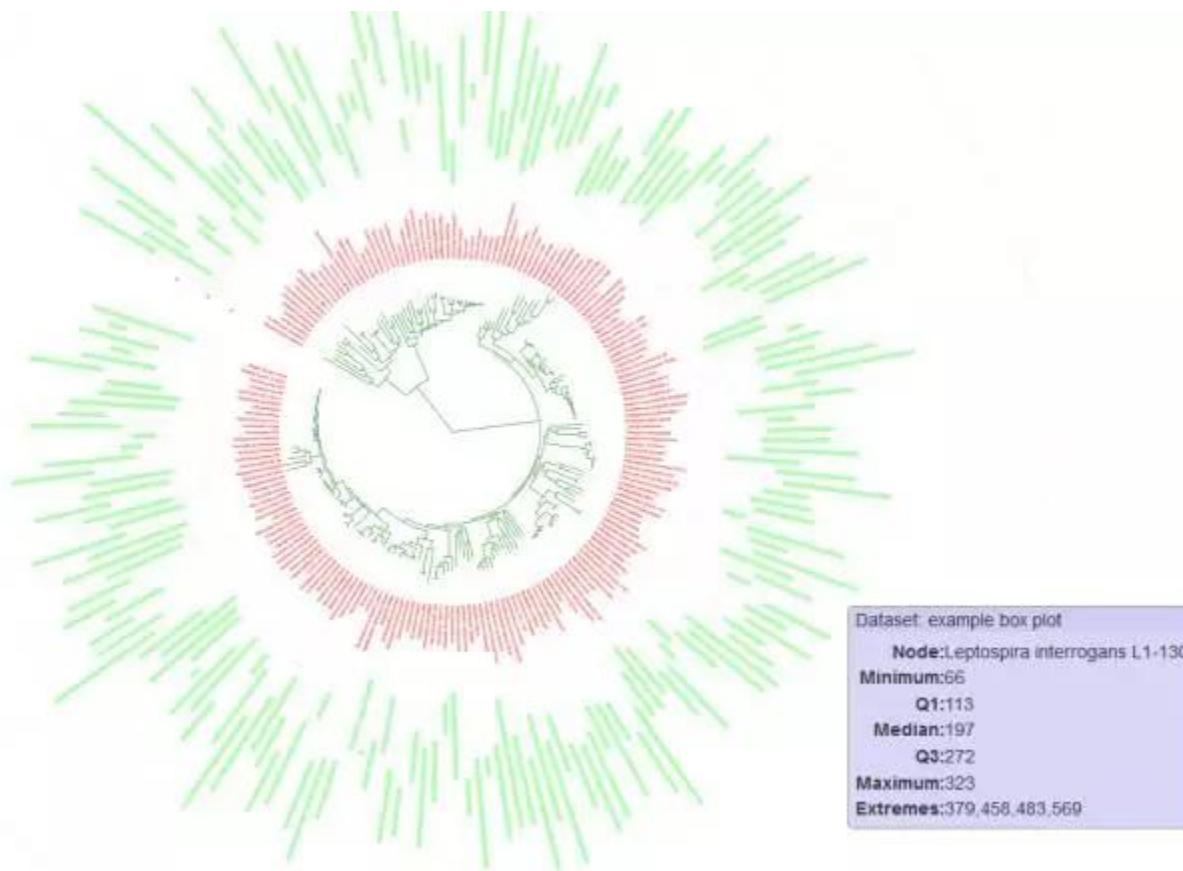


Nature 528, 364–369 (2015)

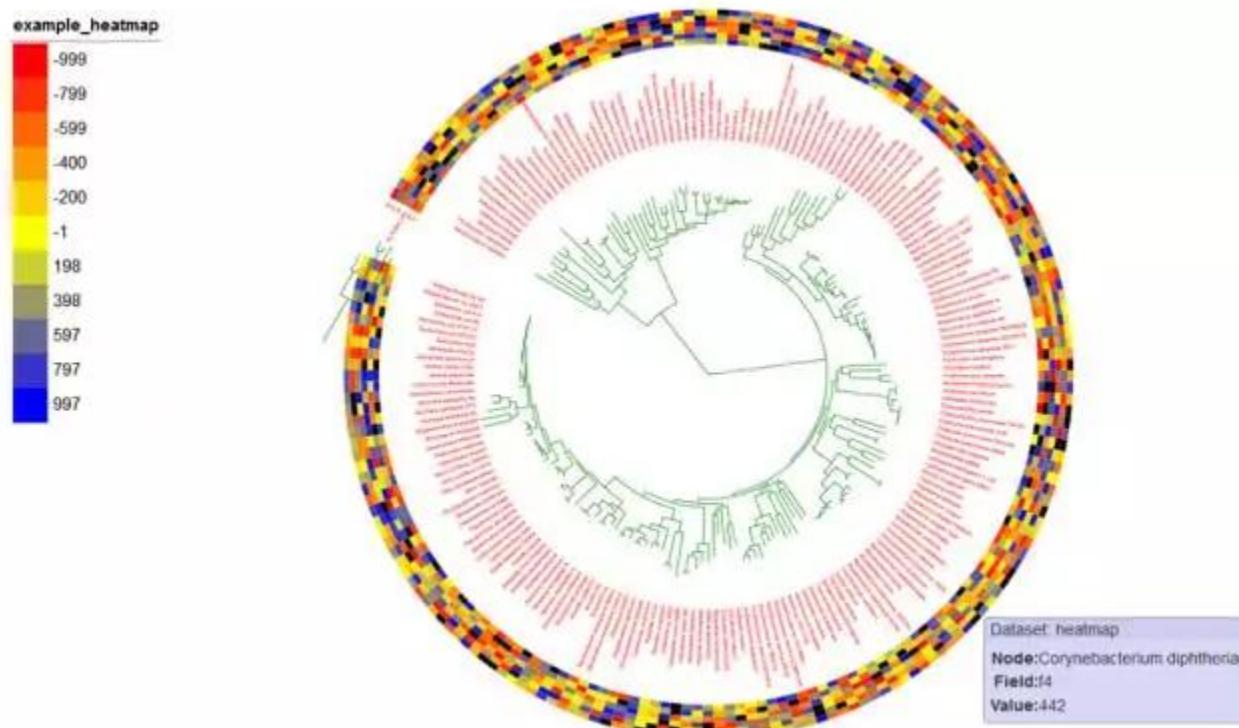


Nature Medicine 22, 1187–1191 (2016)

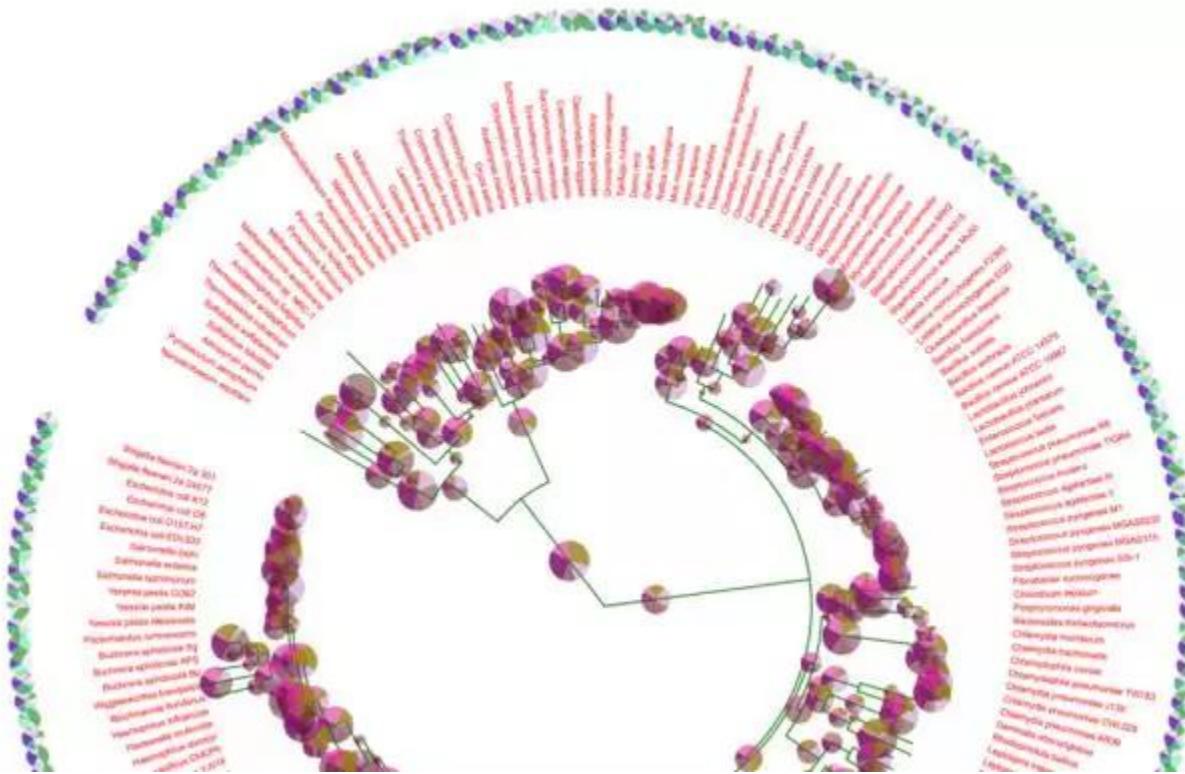
# iTOL: 物种进化关系分析



# iTOL: 物种进化关系分析



# iTOL: 物种进化关系分析

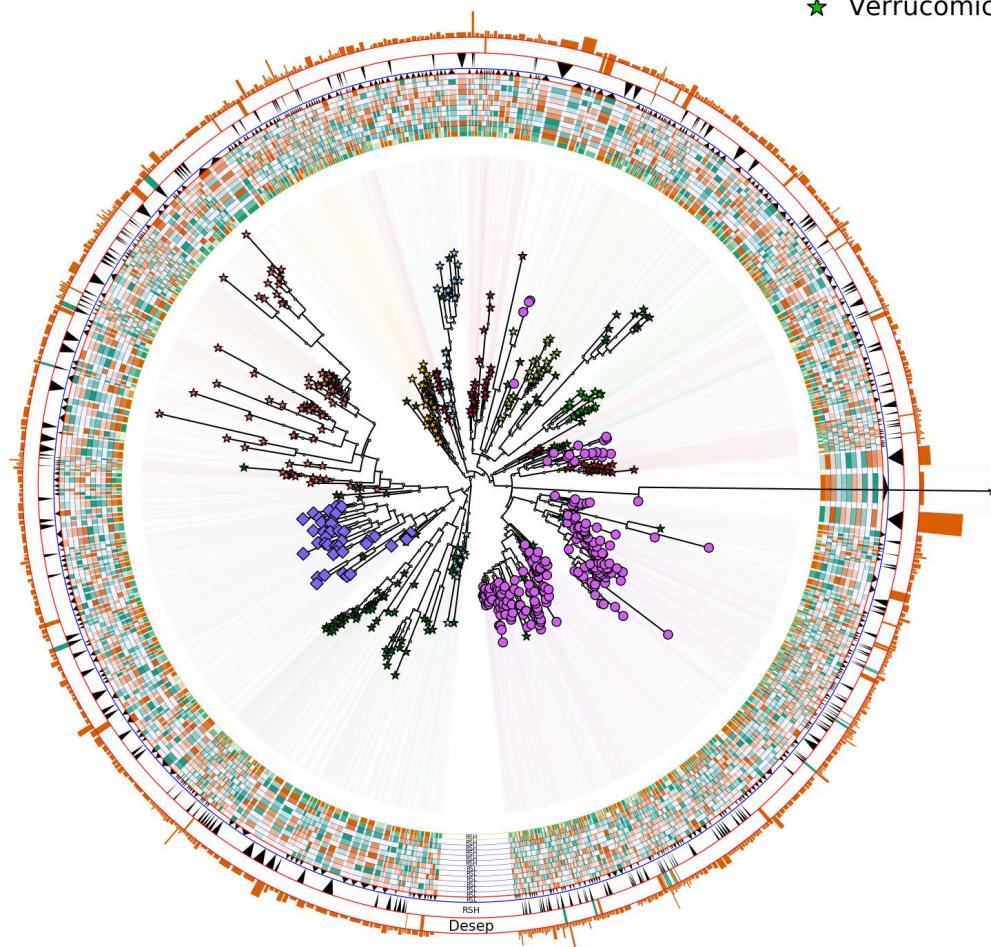


# iTOL: 物种进化关系分析

# GraPhIAn: 分类树分析

Metagenomic

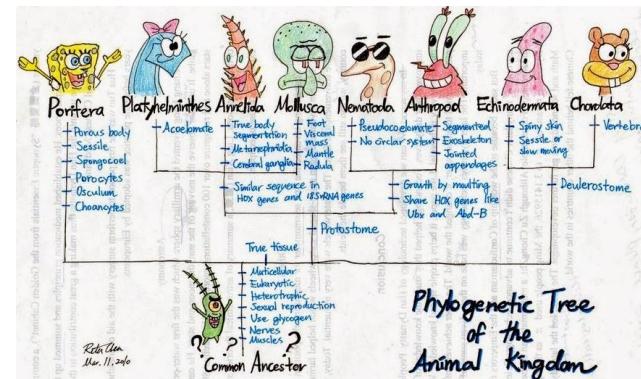
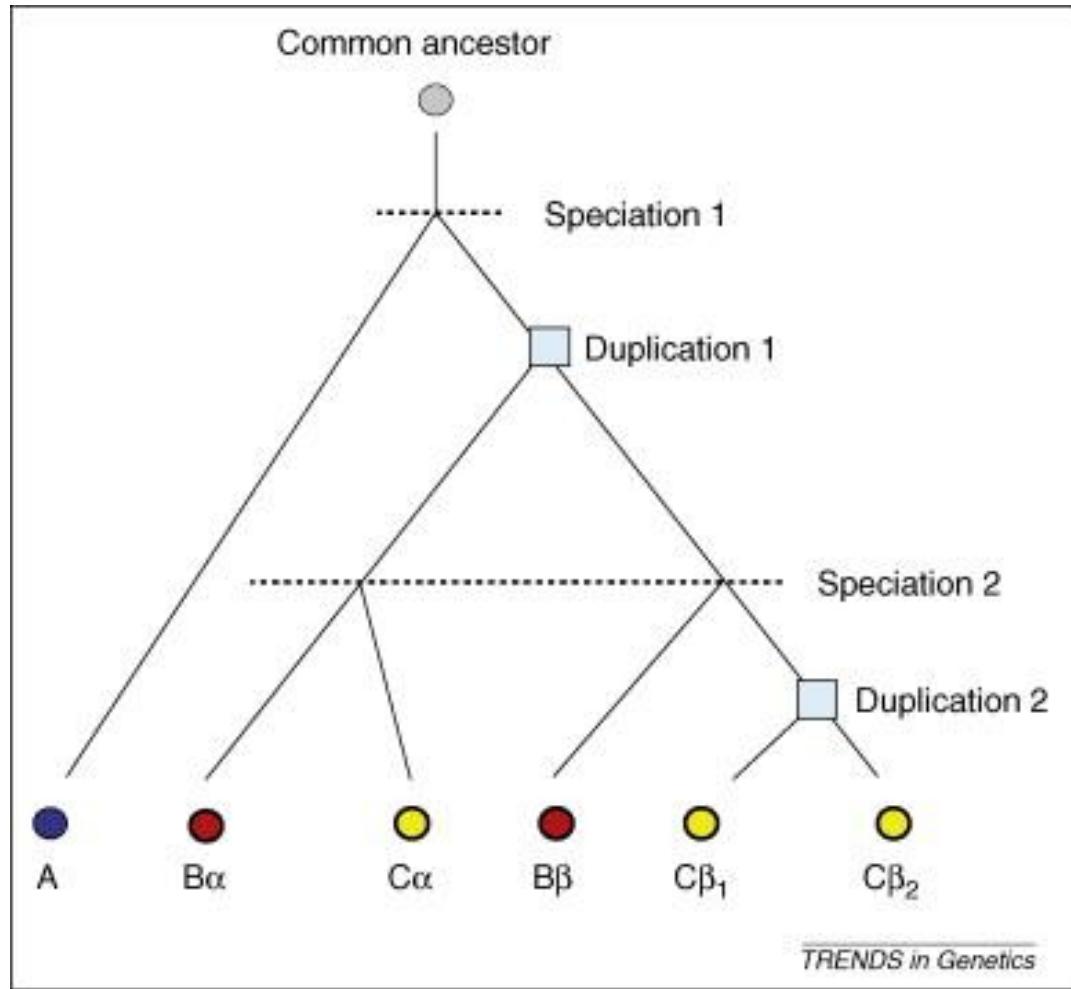
- ★ Acidobacteria
- ★ Actinobacteria
- ★ Bacteroidetes
- ◆ Chloroflexi
- ◆ Cyanobacteria
- ◆ Firmicutes
- ◆ Gemmatimonadetes
- ◆ Others
- ◆ Proteobacteria
- ◆ Verrucomicrobia



# 构建分子进化树相关的软件

- 就进化树而言，iTOL功能最为全面。iTOL无限制添加的数据集，外环可以制作各种图形包括箱线图等，可以使用多种符号填充外环。但Graphlan就不能这么随便了，只能使用两个符号填充外环。再多的环属性也就是设置环数量和颜色，透明度了。
- ggtree最容易上手，但是就一张圈图来说，它不能添加除了热图以外的其他图形，但是在非圈图的模式下，可以对多种数据进行合并，方法将更为简单，操作也容易一些。
- Graphlan可以制作分类树，是它不同于R包ggtree和iTOL的地方。

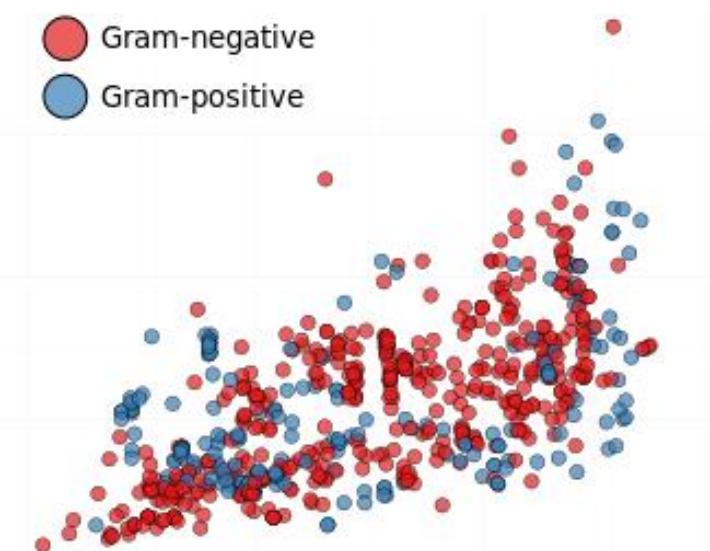
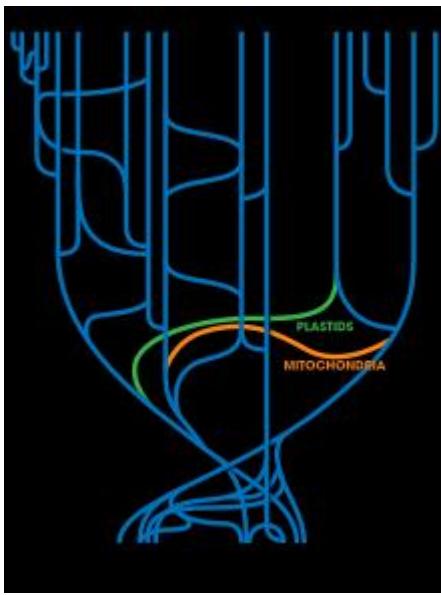
# Gene: ortholog and paralog



# Gene: HGT

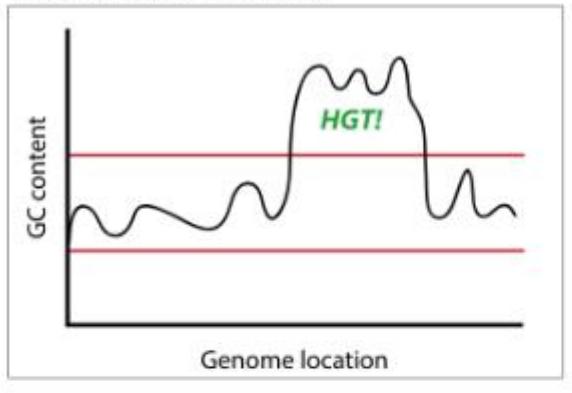
# Gene: HGT

Think about it:  
how does HGT affect the molecular clock calculation?

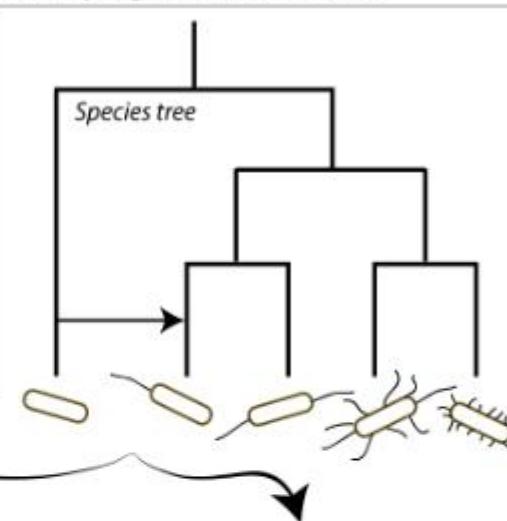


# HGT identification methods (phylogeny analysis)

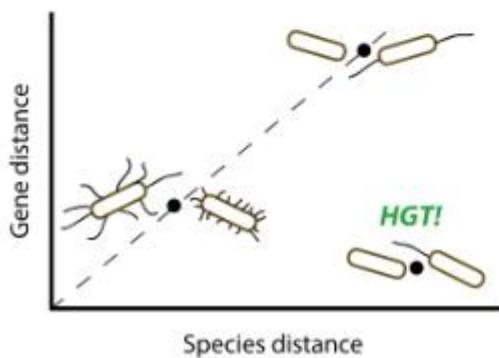
## 1. Parametric methods



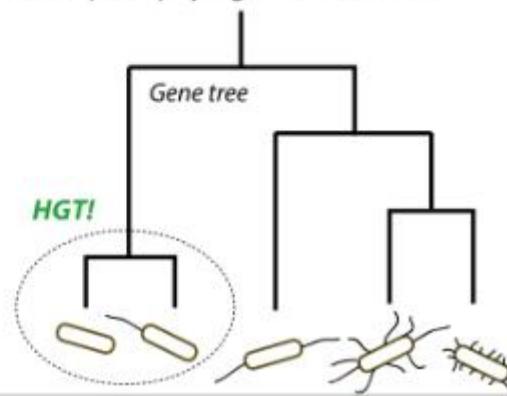
## 2. Phylogenetic methods



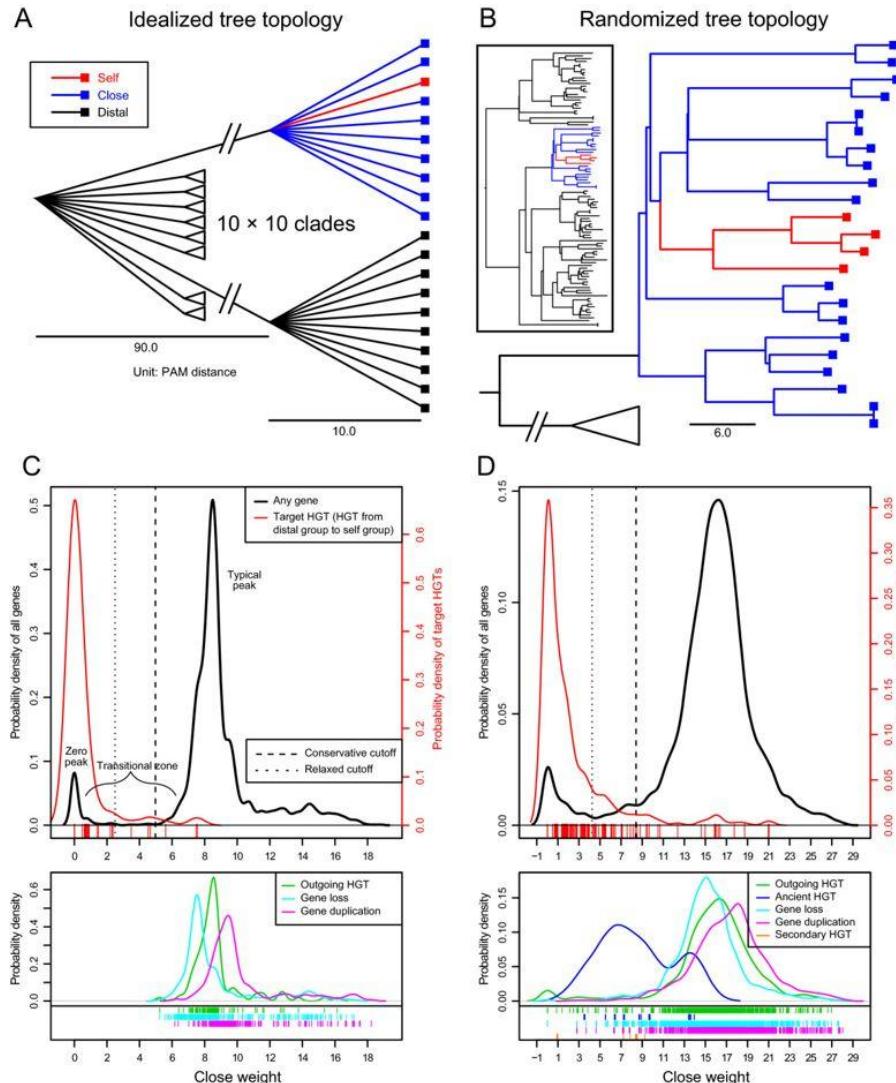
### 2a. Implicit phylogenetic methods



### 2b. Explicit phylogenetic methods



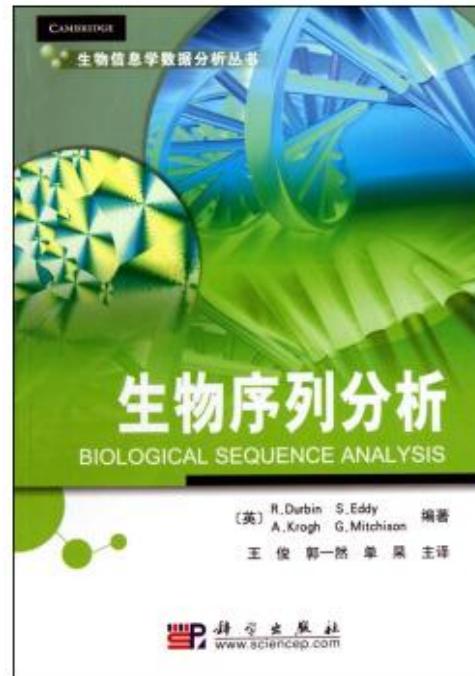
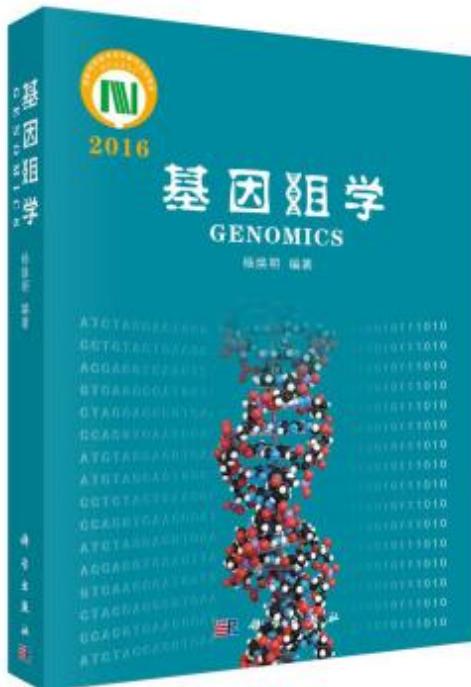
# HGT identification methods (phylogeny analysis)



# 参考文献

- R. Durbin, S. Eddy, A. Krogh and G. Mitchison.  
Biological Sequence Analysis—Probabilistic  
Models of Proteins and Nucleic Acids. 1998,  
Cambridge University Press.

# References



# References



# Slides credits

- 生物信息学研究方法概述：北京大学生物信息中心
- 生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学
- 神经网络与深度学习：邱锡鹏@复旦大学
- Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Harvard University
- Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS
- Deep Learning in the Life Sciences: MIT
- Probabilistic Graphical Models: Eric Xing@CMU
- Numerous other leading researchers and leading labs.....





