中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

国家基因组科学数据中心
National Genomics Data Center

第8章 比较基因组学与分子演化

章张 博士

中国科学院大学
University of Chinese Academy of Sciences

基因组学 - Genomics

2024春季学期

内容提纲

- 生命演化：比较基因组学与分子演化
- 中性突变理论与分子钟假说
- 分子序列演化
- 密码子使用偏好
- 选择压力、同义与非同义替换率
- 系统发育树
- 群体基因组学

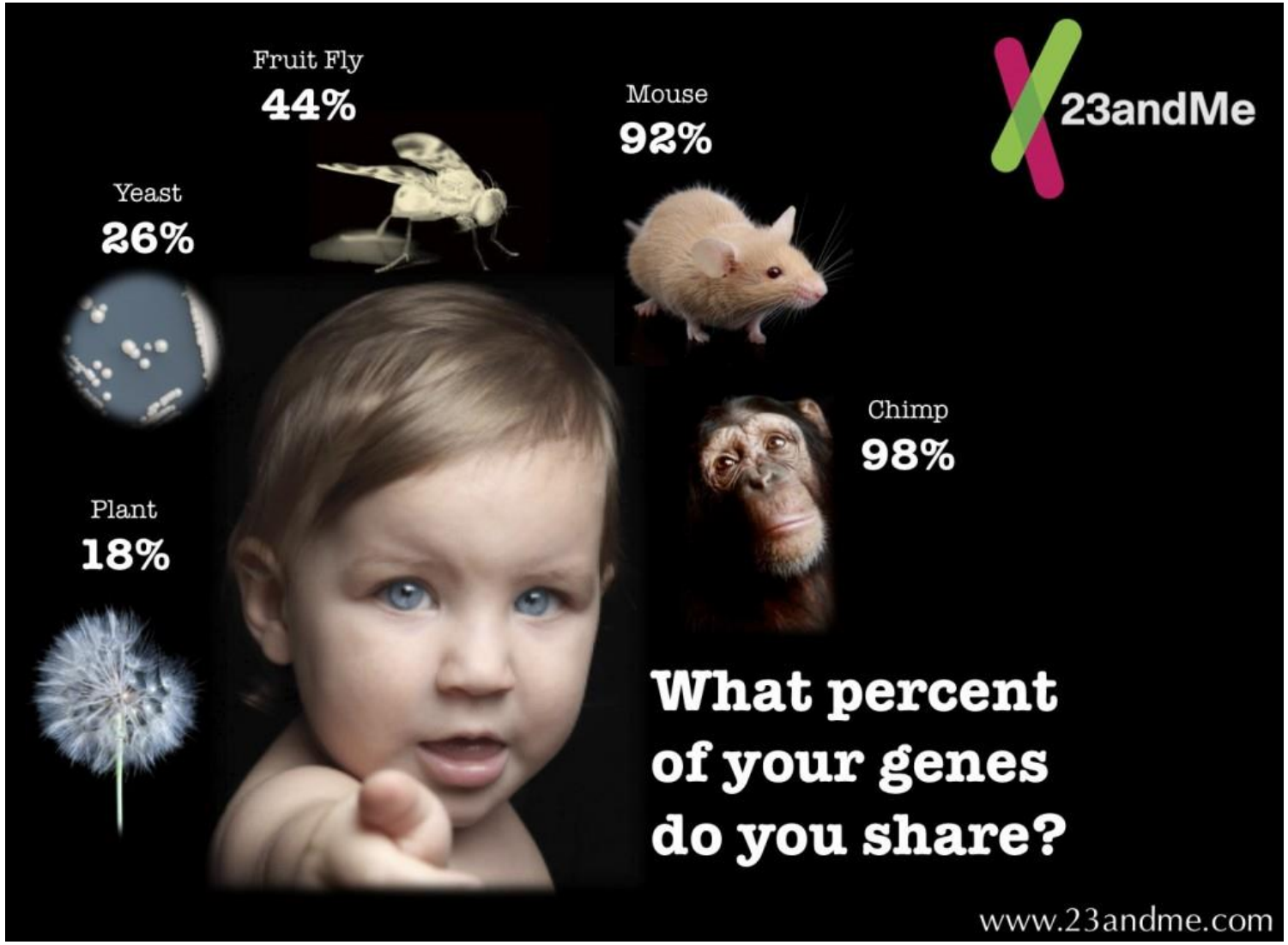
Nothing in biology makes sense except in the light of evolution.——Theodosius Dobzhansky

明进化之光，格生命之谛。

生命演化

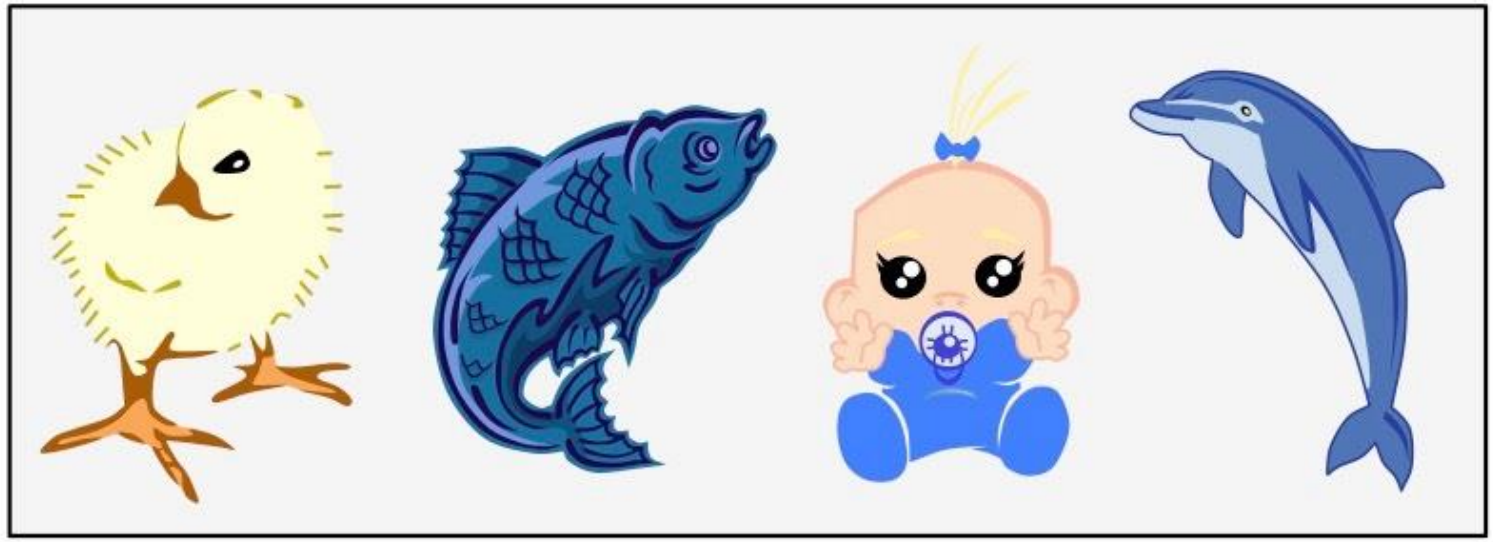
《墨子·经说上第四十二》：“**化，若蛙为鹑**”，蛙变鹑，古代传说，两栖变鸟类的过程叫做“化”，虚幻的生物进化之例

How do different species differ?



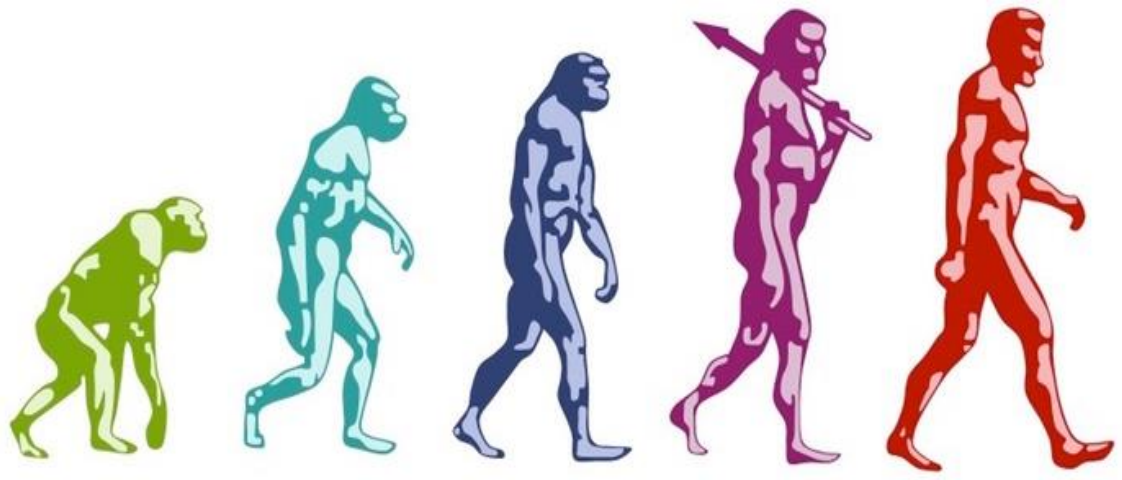
生命演化

- “同宗同祖”：基因组序列是所有生命体演化的“活化石”，其基因组序列相似性显示所有已知物种皆是从共同祖先或是祖先基因池逐渐分化产生。
- “由简到繁”：从原始简单物种演化成为复杂有智慧的物种，从单细胞物种到多细胞物种，生命体基因组序列可用于重新构建数据化的“生命之树”(Tree of Life)。



Evolution = change over time

- Evolution at its most fundamental level simply describes **a *change over time***.
- Heritable characteristics are encoded by **genes** and may be transferred *between generations* as **alleles.**
- A concise definition is a change in the *allele frequency* of a population's ***gene pool*** over successive generations.



Evolution译为演化（变演），遵循严复先生的翻译，避免“进化”被误解为生命过程只能向一个方向

严复（1854–1921）
《天演论》 Evolution and Ethics

托马斯·赫胥黎（Thomas Huxley，1825–1895）
英国著名博物学家，达尔文进化论最杰出的代表

物竞天择
适者生存

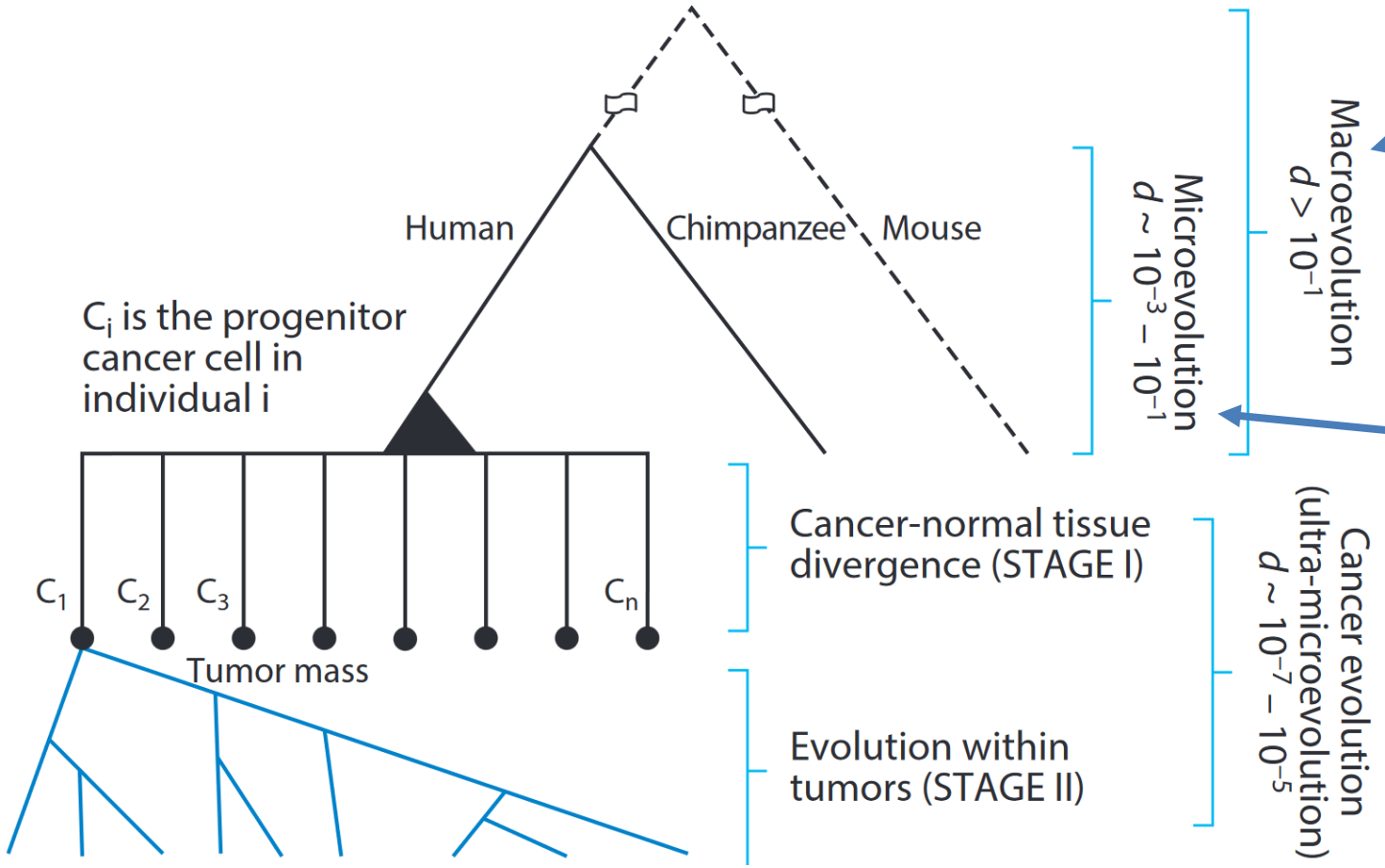
翻译三原则
信 达 雅

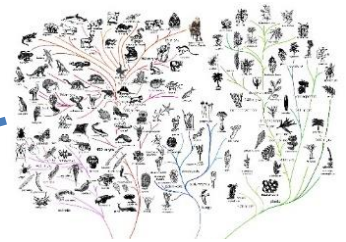
Microevolution vs Macroevolution

Microevolution and macroevolution describe fundamentally identical processes on different evolutionary scales.

	Microevolution	Macroevolution
Change	small evolutionary changes	evolutionary changes resulting in speciation
Species / Population	within a species or population	in large populations
Time	within a short period of geological time	over relatively long geological periods
Association	due to mutation, selection, and genetic drift	associated with significant environmental change

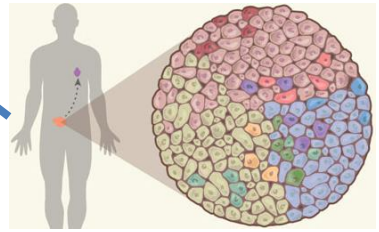
宏进化、微进化与超微进化





宏进化
(macroevolution)
$d > 10^{-1}$

微进化
(microevolution)
$d \sim 10^{-3}$ —10^{-1}



超微进化
(ultra-microevolution)
$d \sim 10^{-7}$ — 10^{-5}

The scales of organismal evolution and cancer evolution.
evolutionary distance *d*: number of differences per nucleotide

The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process. *Annual Review of Genetics* 2016 https://doi.org/10.1146/annurev-genet-112414-054842

Studying evolution at the molecular level

"*Nothing in biology makes sense except in the light of evolution*."

明德格物，语出《礼记·大学》：“大学之道，在明明德”，“致知在格物，格物而后知至”。

明德，即注重道德修养及内化。
格物，即探究事物本质及原理。

明进化之光，格生命之谛。

Theodosius Dobzhansky (1900–1975)
Russian-American geneticist and evolutionary biologist

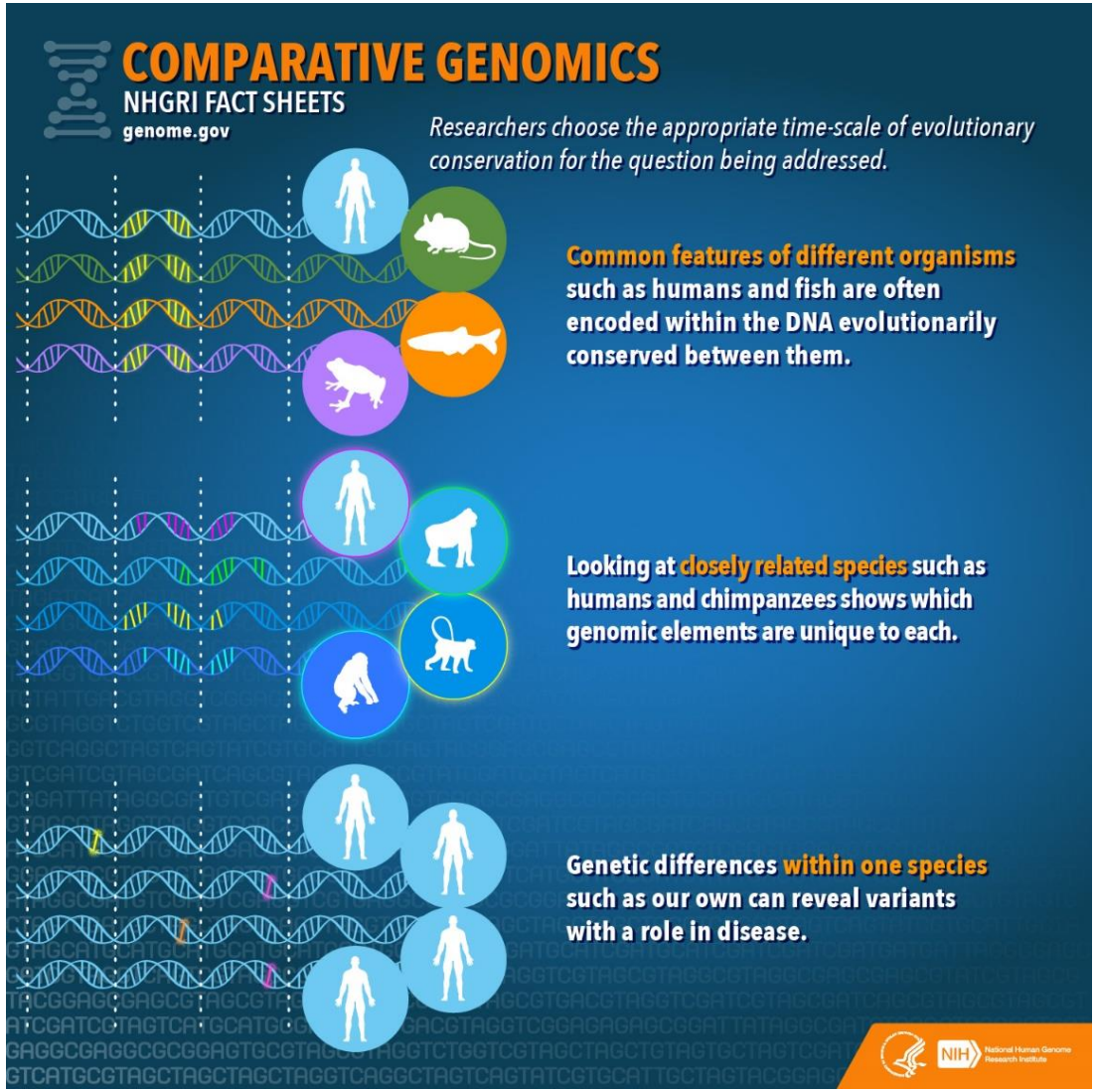
- **Morphological (Mendelian) approach:**
  Confined to within-species genetic changes
- **Molecular approach:**
  Removal of species barrier and extension to between-species

DNA sequences are more informative than protein sequences, due to the degeneracy of the genetic code.

比较基因组学（Comparative Genomics）

比较基因组学是通过比较基因序列特征和基因组结构了解基因功能、表达机制和不同物种亲缘关系的生物学研究。

- 种间比较：通过对不同亲缘关系物种的基因组序列比较，研究**进化保守区域**、**物种特有区域**以及**物种进化关系**等。
- 种内比较：鉴定同种群体内基因组存在的**变异和多态性**，基于这些变异研究个体与群体对疾病的**易感性**和对药物与环境因子的**不同反应**。



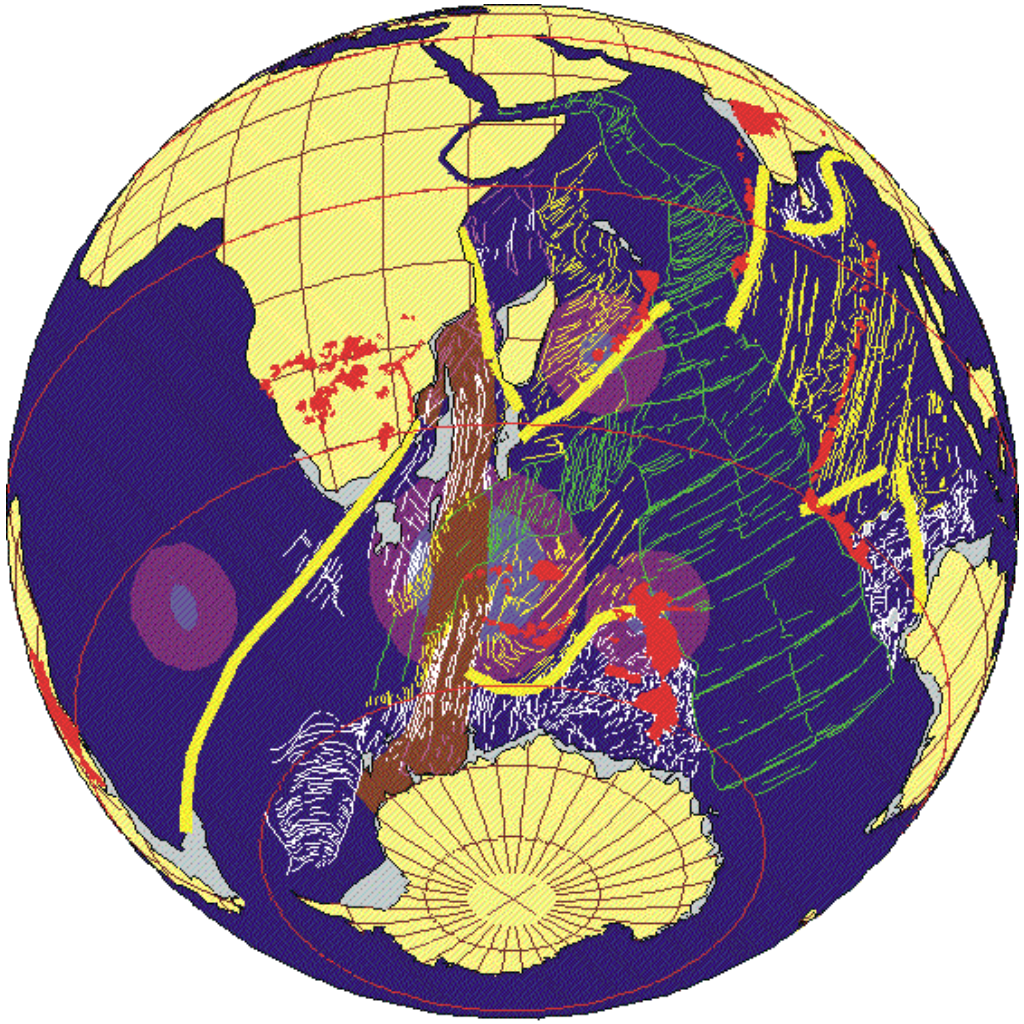
https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet

分子演化（Molecular Evolution）

how sequences change over time

- Molecular evolution is the area of evolutionary biology that **studies evolutionary change at the level of the DNA/protein sequence**.
- Molecular evolution is **the process of change** in the sequence composition across generations.
- **Major topics** in molecular evolution
  - rates and patterns of sequence changes
  - relative importance of adaptive and neutral changes
  - neutral evolution vs. natural selection
  - origins of new genes
  - the genetic basis of speciation
  - evolutionary forces on genomic and phenotypic changes

生命演化与地球运动



生命起源：

生物学
化学
地质学
考古学
地球化学
天文学
天体物理学
……

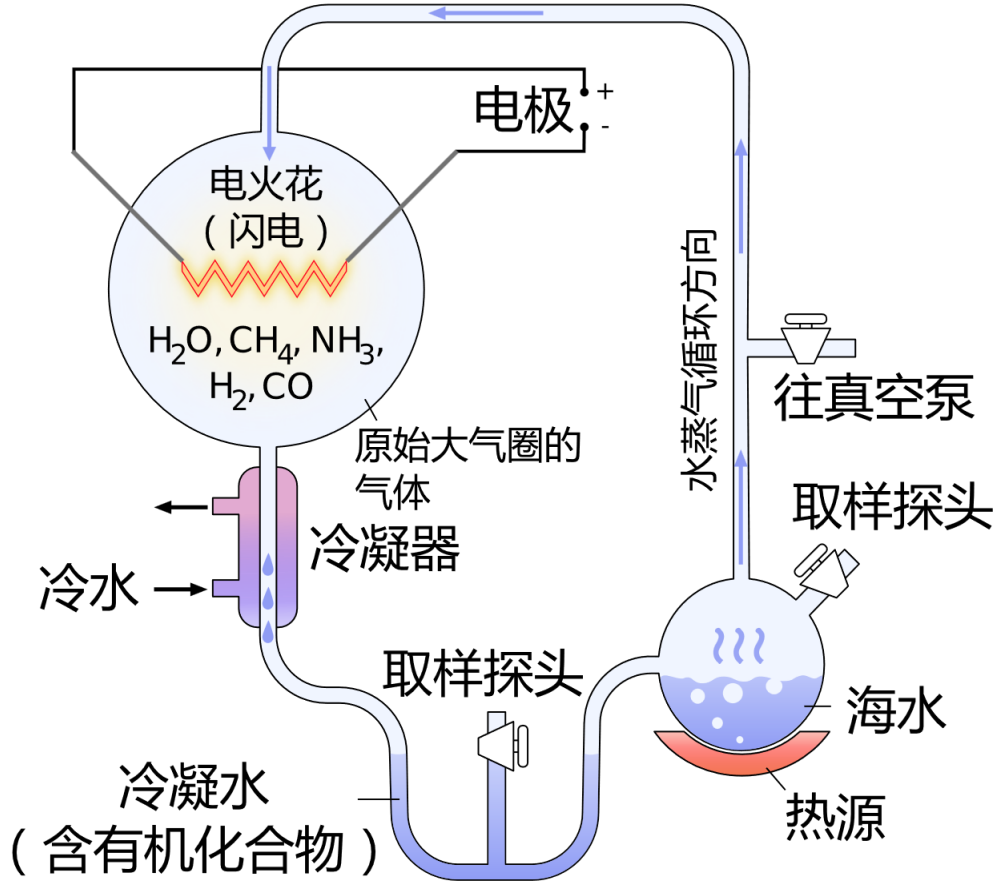
https://kartoweb.itc.nl/gondwana/gondwana.html

地球上的生命起源

米勒-尤里实验（Miller-Urey experiment）

生命起源的经典实验之一，模拟早期地球环境，研究目的是测试化学演化的发生情况，认为早期地球环境使无机物合成有机化合物的反应较易发生。

Stanley Miller (1930–2007). This experiment was conducted in 1952, with assistance from Harold Urey, at the University of Chicago.



右下烧瓶模拟海洋环境，左上烧瓶则模拟闪电

From Wikipedia

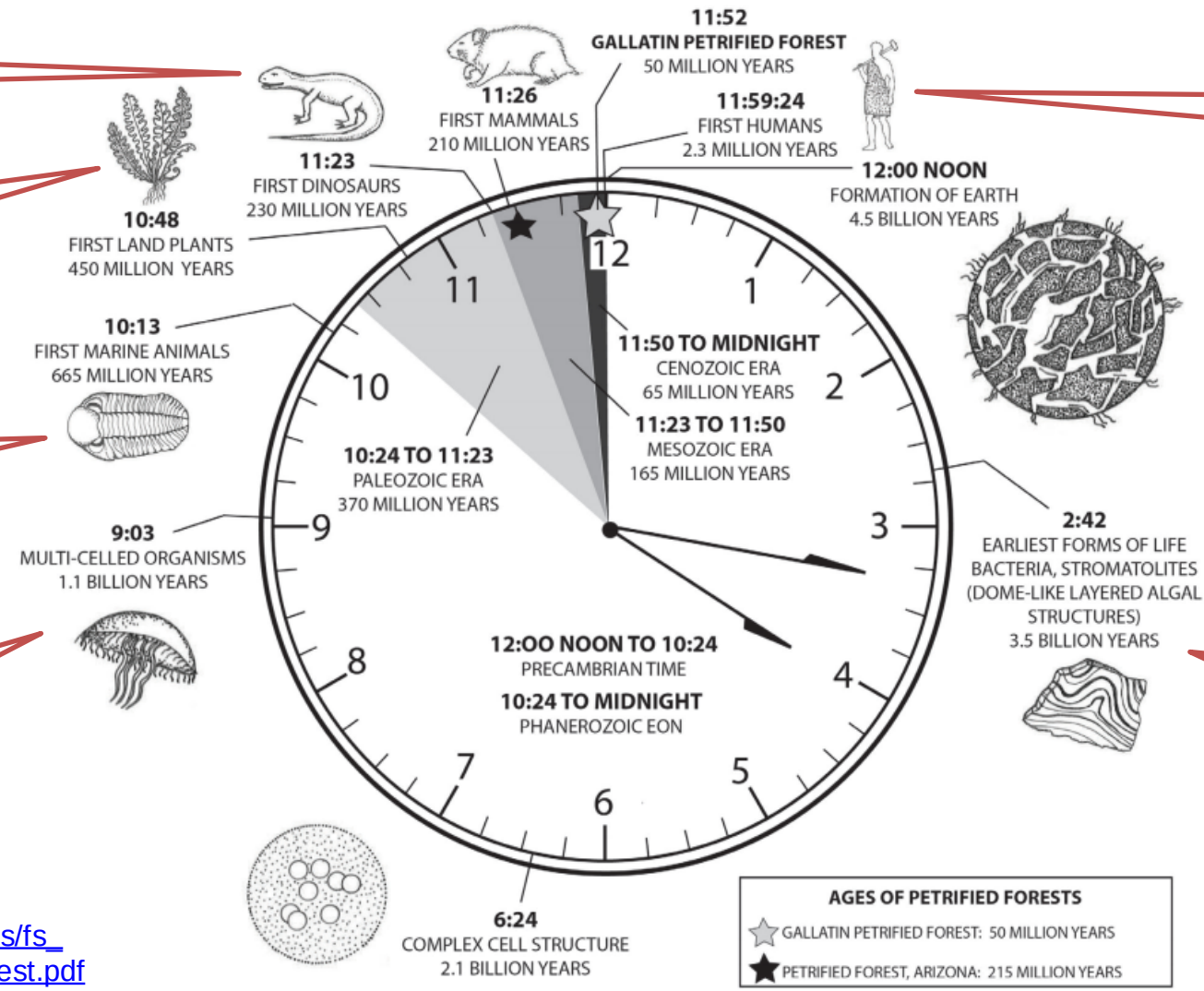
History of Earth in 12-hour clock

恐龙

陆生植物

水生动物

多细胞生物

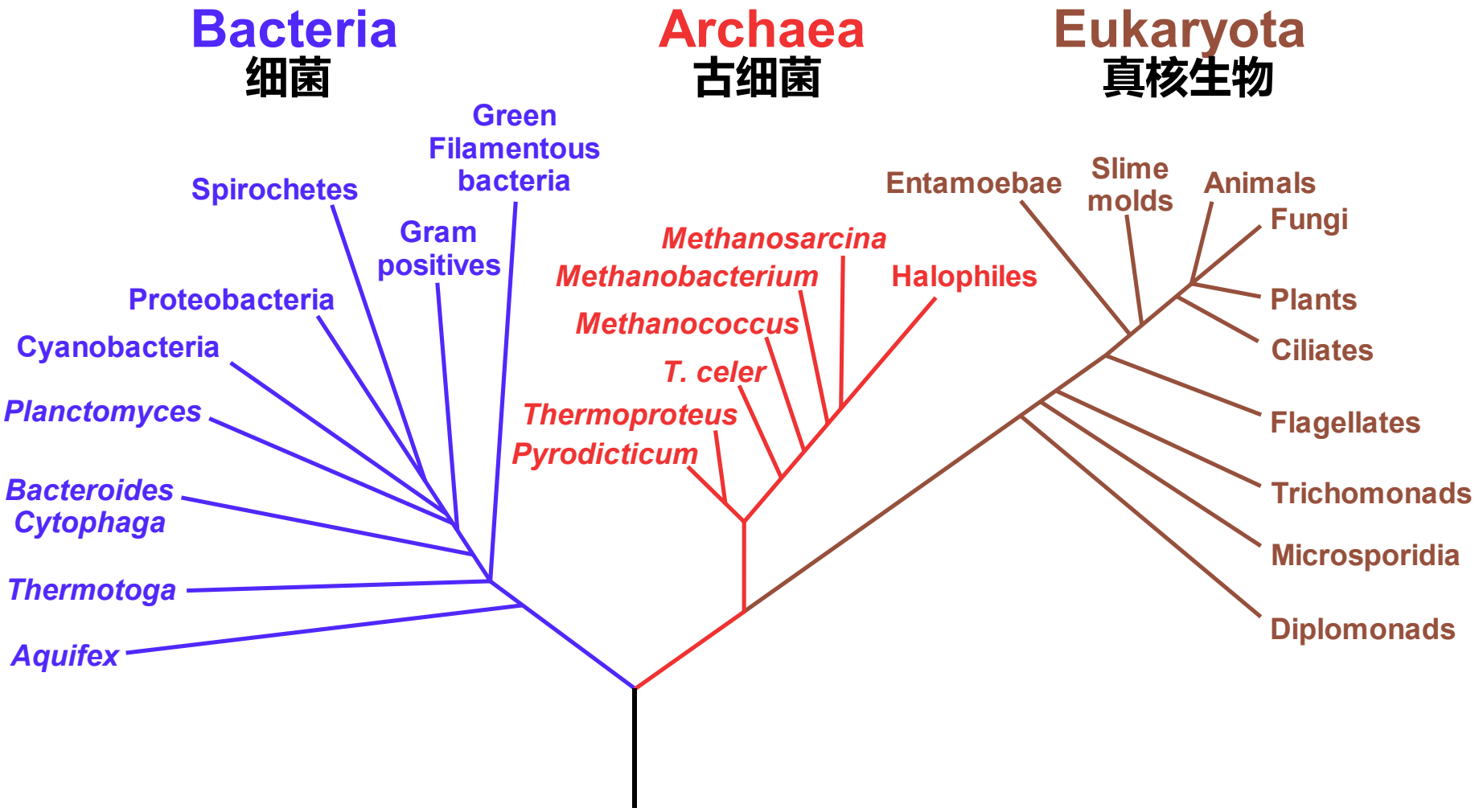


人类

早期生命迹象

https://www.fs.usda.gov/sites/default/files/fs_media/fs_document/gallatin-petrified-forest.pdf

生命之树（Tree of Life）



演化学说与分子钟

重要演化学说及历史人物

Jean-Baptiste Lamarck (1744–1829): 法国博物学家，进化论奠基人

1761-1768年在军队服役，1768年与卢梭相识，法国著名的思想家、哲学家、教育家、文学家，系统地研究了植物学

1778年出版了3卷集的《法国植物志》，有成就的植物学家

1793年应聘为巴黎博物馆无脊椎动物学教授

1801年完成《无脊椎动物的系统》

1809年发表《动物学哲学》，提出用进废退与获得性遗传

1817年完成《无脊椎动物自然史》

First law: use and disuse
Second law: inheritance of acquired characteristics

Gregor Johann Mendel (1822–1884): 奥地利帝国(捷克)，现代遗传学创始人

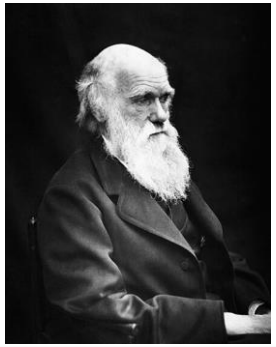
1840年进入帕拉茨基大学哲学学院学习理论哲学以及物理学

1843年因家贫而辍学，同年10月到圣托马斯修道院做修士，1847年被任命为神父

1849年受委派到茨纳伊姆中学任希腊文和数学代课教师

1851-1853年在维也纳大学学习物理、化学、数学、动物学和植物学

1856-1863年，8年豌豆杂交实验，1865年提出遗传因子概念，建立孟德尔定律



Charles Robert Darwin (1809–1882): 英国博物学家、地质学家和生物学家

1828年剑桥大学，1831-1836年，“贝格尔号”，环球考察

1838年10月，达尔文进化论的框架基本形成

1844年钱伯斯激进的进化论问世; 50年代中期斯宾塞的社会进化观诞生；1854年华莱士《论控制新种引入的法则》1857年《论变种无限地离开其原始模式的倾向》

1859年发表《物种起源》，提出生物进化论学说

Lamarck's Theory of Evolution

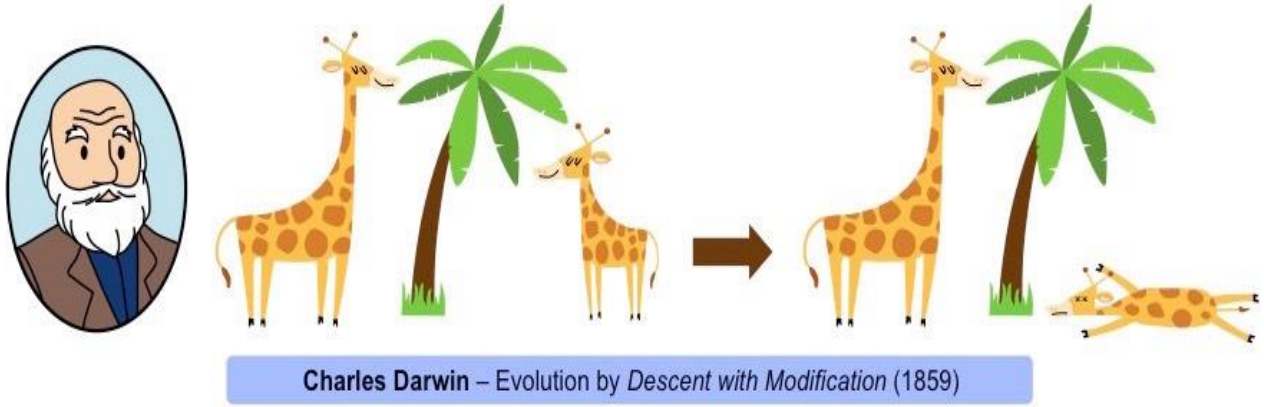
- Giraffes all had SHORT necks originally
- Giraffe's Necks got LONGER from stretching for food
- "Acquired" trait (long necks) then passed to offspring
- Giraffe population became long-necked

Darwin's Theory of Evolution

- A combination of Lamarckian ideas and recent fossil discoveries
- Species living today had been changed over time and stemmed from a single (or few) ancestral organism(s).
- Organisms which possess traits better suited to conditions would have an adaptive advantage and be more likely to reproduce.
- These traits would hence become more common within the population and the species would gradually change over time.

Neo Darwinism
Mendelian inheritance
DNA structure

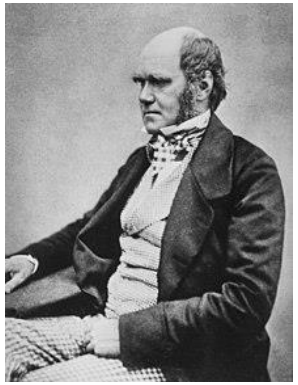
拉马克Jean-Baptiste Lamarck (1744—1829）：进化论奠基人
1809年发表《动物学哲学》，
提出用进废退与获得性遗传



Charles Darwin – Evolution by *Descent with Modification* (1859)

Long-necked giraffes are randomly born and have more offspring due to their competitive advantage

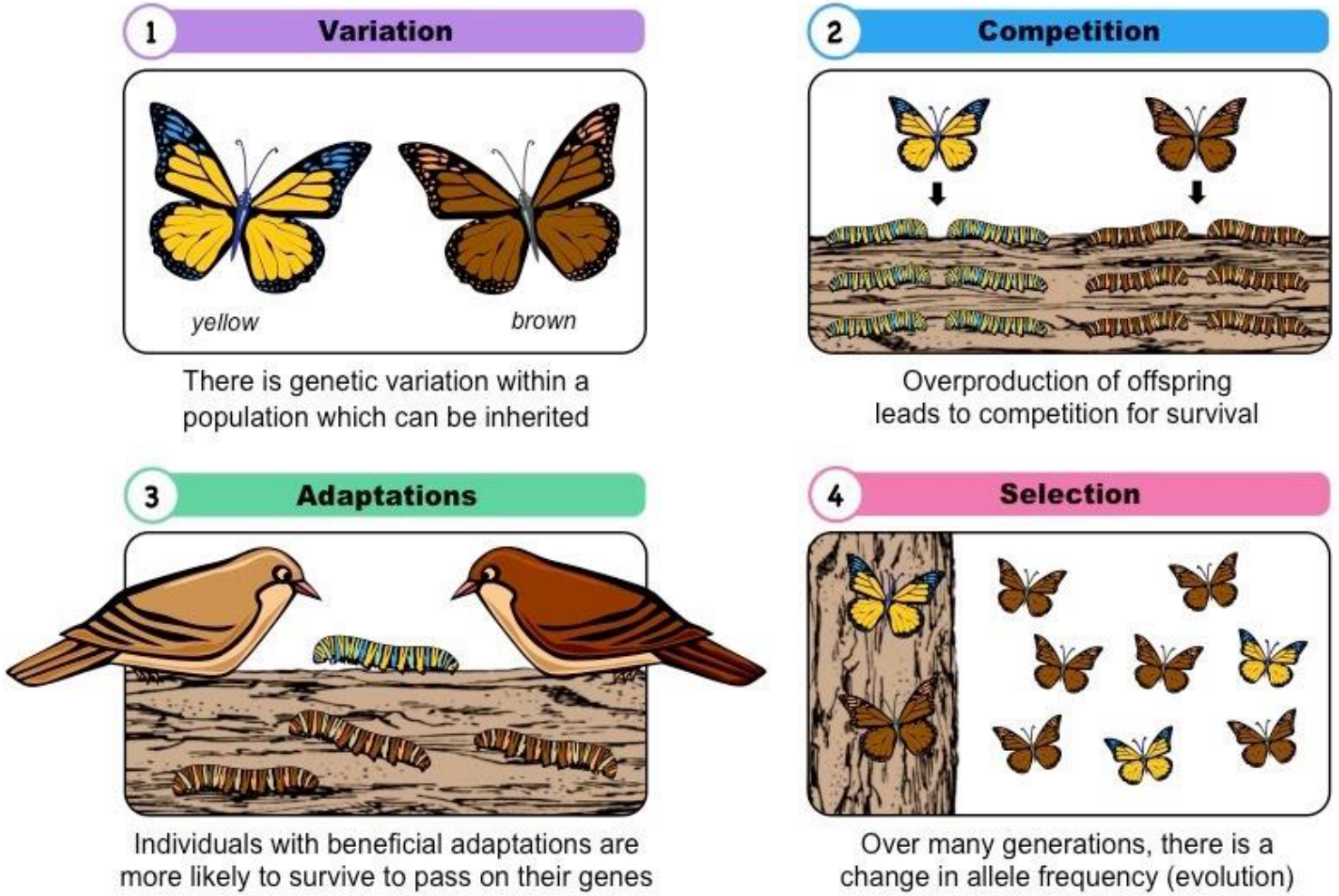
https://ib.bioninja.com.au/standard-level/topic-5-evolution-and-biodi/52-natural-selection/theories-of-evolution.html



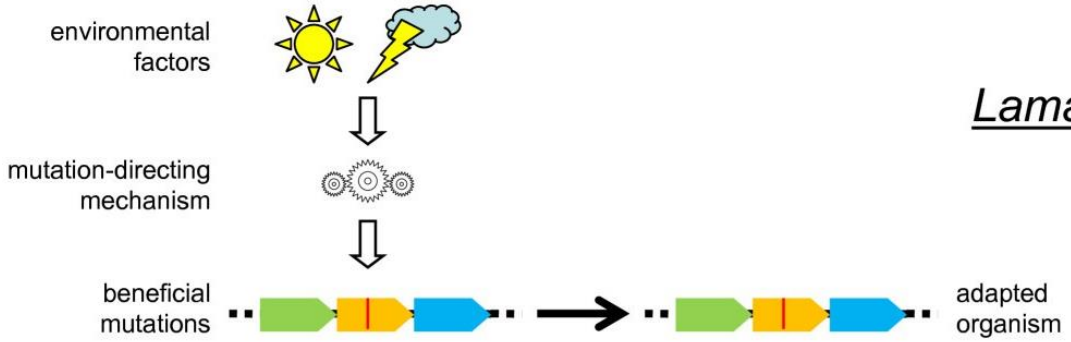
Charles Darwin
(1809—1882)
1859年《物种起源》

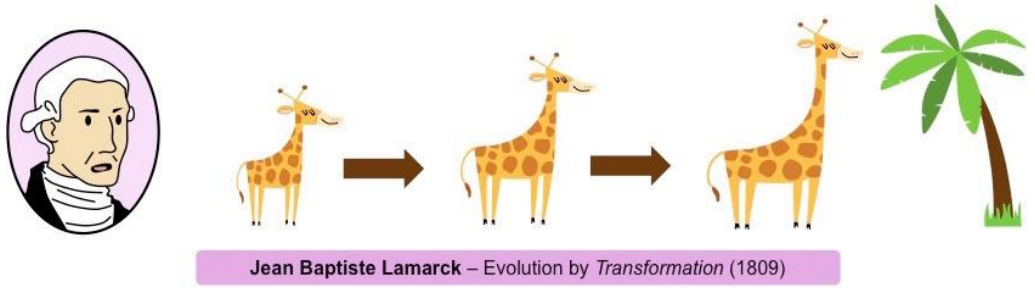
Natural Selection

The theory of natural selection—‘survival of the fittest’ (适者生存)

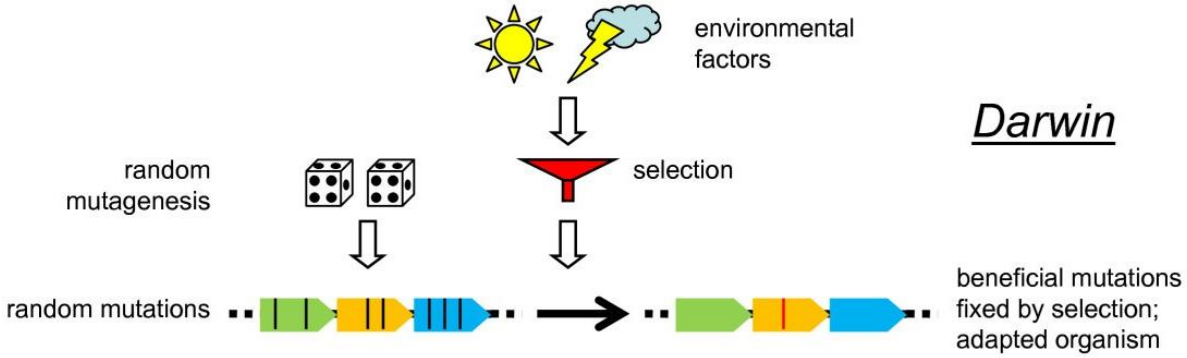


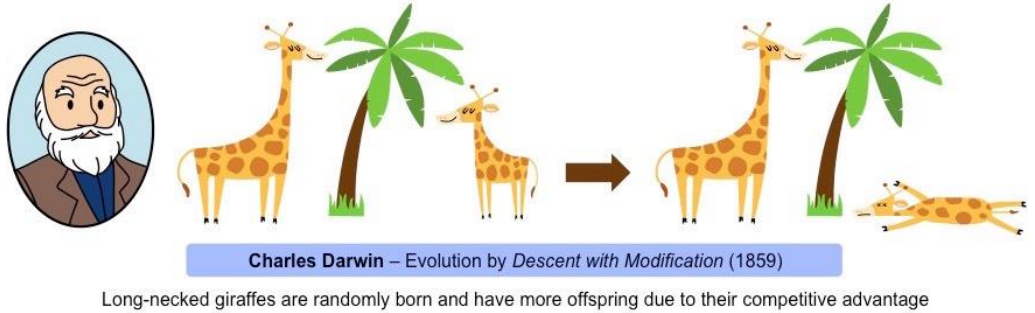
Darwinian vs. Lamarckian





Lamarck's Theory of Evolution (Use and Disuse)





Darwin's Theory of Evolution (Natural Selection)

Is evolution Darwinian or/and Lamarckian? *Biology Direct* 2009
https://doi.org/10.1186/1745-6150-4-42

https://ib.bioninja.com.au/standard-level/topic-5-evolution-and-biodi/52-natural-selection/theories-of-evolution.html

The Neutral Theory of Molecular Evolution

The theory was introduced by the Japanese biologist Motoo Kimura in **1968** and described in detail in his **1983** monograph ***The Neutral Theory of Molecular Evolution***. The theory has been widely accepted and is the **guiding principle for studying evolutionary genomics and the molecular basis of phenotypic evolution**.

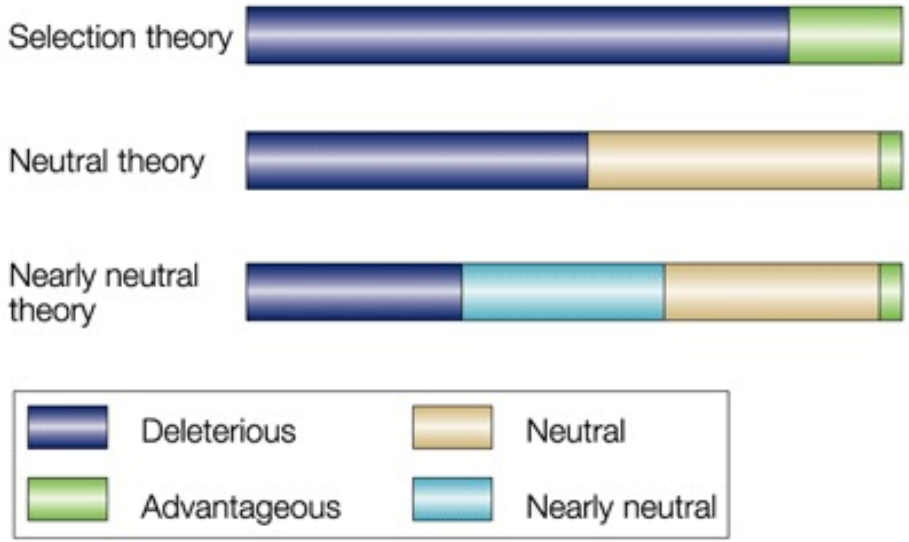
Motoo Kimura 木村资生
(1924–1994)

- A neutral mutation **does not affect an organism's ability** to survive and reproduce.
- Most evolutionary changes at the molecular level are **caused by random genetic drift** through mutations that are **selectively neutral**, giving rise to genetic polymorphisms.
- Most evolutionary changes and polymorphisms within species are not caused by natural selection, but **by random genetic drift**.

***Motivated by the Haldane's dilemma**: it takes about 300 generations for a beneficial mutation to become fixed in a mammalian lineage, meaning that **the number of substitutions (1.5 per year)** in the evolution between humans and chimpanzees was too high to be explained by beneficial mutations.*

Neutral vs Selection

- The neutral theory is **not an anti-Darwinian theory**.
- Both recognize that **natural selection** is responsible for the adaptation of organisms to their environment.
- Both recognize that **most new mutations in functionally important regions are deleterious** and that **purifying selection** quickly removes these deleterious mutations from populations.
- The evolutionary fate of genetic variation is best explained by **stochastic processes**.
- The **null hypothesis** of molecular evolution is that **sequence evolves neutrally**.
- It provides a theoretical framework for developing methods that **detect the forces** acting on sequences.



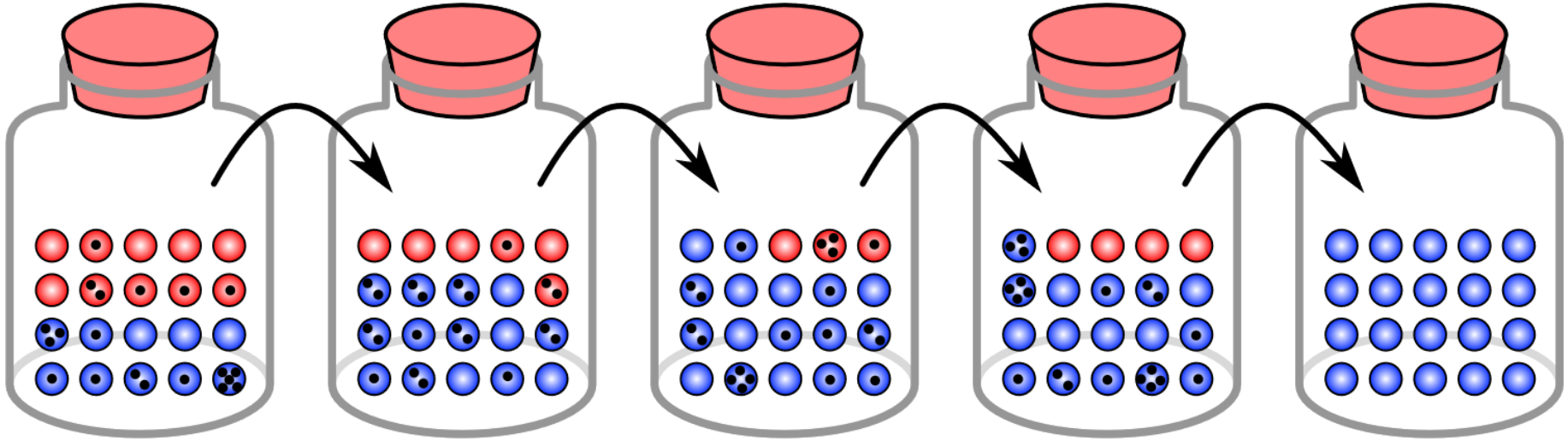
The modern molecular clock
Nature Reviews Genetics 2003

Forces in Molecular Evolution

- **Selection** occurs when organisms with greater fitness are favored in subsequent generations, thereby **increasing the instance of underlying genetic variants in a population**.
  - **Natural selection** is any selective process that occurs due to the **fitness of an organism to its environment**.
  - **Artificial selection**, also known as **selective breeding**, is imposed by an outside entity, typically humans, in order to increase the frequency of **desired traits**.
- **Mutations** are **stochastic** and typically occur **randomly** across genes. Most mutations that occur are single nucleotide polymorphisms, resulting in point mutations. Most organisms display a strong bias in the types of mutations that occur with **strong influence in GC content**.
- **Genetic drift** is the change of allele frequencies from one generation to the next due to stochastic effects of random sampling in finite populations.

遗传漂变（Genetic drift)

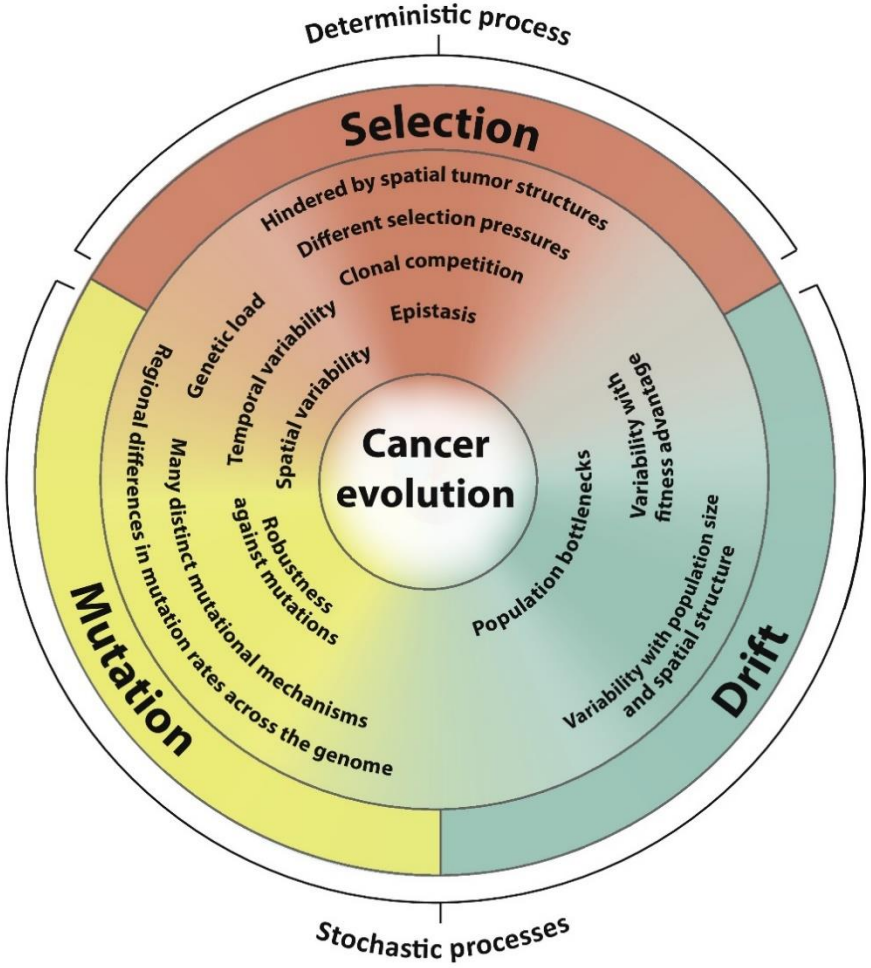
- 遗传漂变：是指**种群中基因库在代际发生随机改变的一种现象**。由于任何一个个体的生存与繁殖都受到随机因素影响，繁殖过程可看做一种抽样，**子代携带的等位基因即是对亲代抽取的一种样本**。这一过程中的抽样误差使**子代中的等位基因频率与亲代并不相等**，尤其是在**小种群中**。
- 遗传漂变可能改变某一等位基因的频率，甚至致其完全消失，进而降低种群的遗传多样性。一般情况下，**种群的生物个体数量越少，遗传漂变的效应就越强**。遗传漂变是生物进化的关键机制之一。



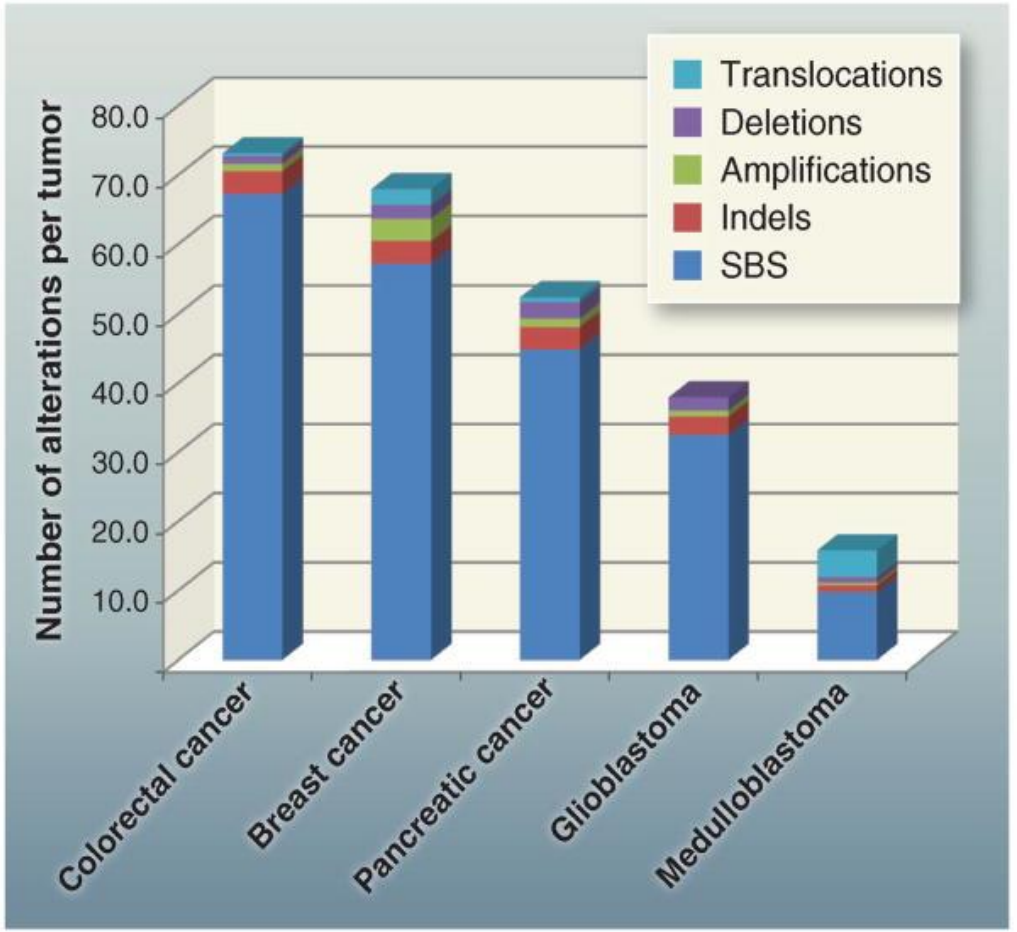
遗传漂变模拟模型（20个小球代表某种群。红蓝小球代表种群中的两种等位基因）
在较小的种群中，某一等位基因可能在几代之内就被固定下来。

https://en.wikipedia.org/wiki/Genetic_drift

癌症演化



Precision Cancer Medicine. *Trends in Cancer* 2015
DOI: https://doi.org/10.1016/j.trecan.2015.11.003



Single-base substitutions (SBS)

Cancer Genome Landscapes. *Science* 2013

分子钟（Molecular Clock）

The rate of evolution in any protein (or later, DNA) molecule is **approximately constant over time and over evolutionary lineages**.

The notion of the existence of a so-called "molecular clock" was first attributed to Émile Zuckerkandl and Linus Pauling who, in 1962, noticed that **the number of amino acid differences in hemoglobin between different lineages changes roughly linearly with time**, as estimated from fossil evidence.

- Rates of molecular evolution can be remarkably **constant over time**, **producing a molecular clock**.
- The constancy of rates was **explained by the neutral theory** by assuming that most changes to DNA or protein sequences are neutral — that is, driven by drift not selection.
- Variation in rate between lineages can cause **substantial bias in molecular date estimates.**
- Stochastic fluctuations in substitution rate over time in lineages make **molecular date estimates imprecise**.

分子钟假说

- 序列之间的遗传差异的数量是自分化以来的时间的函数。
- 分子序列的变化速率稳定，可以用来预测分化的时间。

物种分化时间：化石证据

1. 灵长目-啮齿动物： ~80 Myr ago
2. 哺乳动物-鸟类： ~310 Myr ago
3. 哺乳动物-两栖类： ~350 Myr ago
4. 四肢动物-硬骨鱼： ~430 Myr ago
5. 脊椎动物-果蝇(昆虫)： ~830 Myr ago

Nature Genetics 31, 205 - 209 (2002)

Beta globins: human-mouse

- Beta globin chains of closely related species are highly similar:
- Observe simple alignments below:

```
Human β chain: MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLL
Mouse β chain: MVHLTDAEKAAVNGLWGKVNPDDVGGEALGRLL

Human β chain: VVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
Mouse β chain: VVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVIN

Human β chain: AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGN
Mouse β chain: AFNDGLKHLDNLKGTFAHLSELHCDKLHVDPENFRLLGN

Human β chain: VLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Mouse β chain: MIVIVLGHHLGKEFTPCAQAAFQKVVAGVASALAHKYH
```

There are a total of **27** mismatches, or (147 – 27) / 147 = **81.7 %** identical.

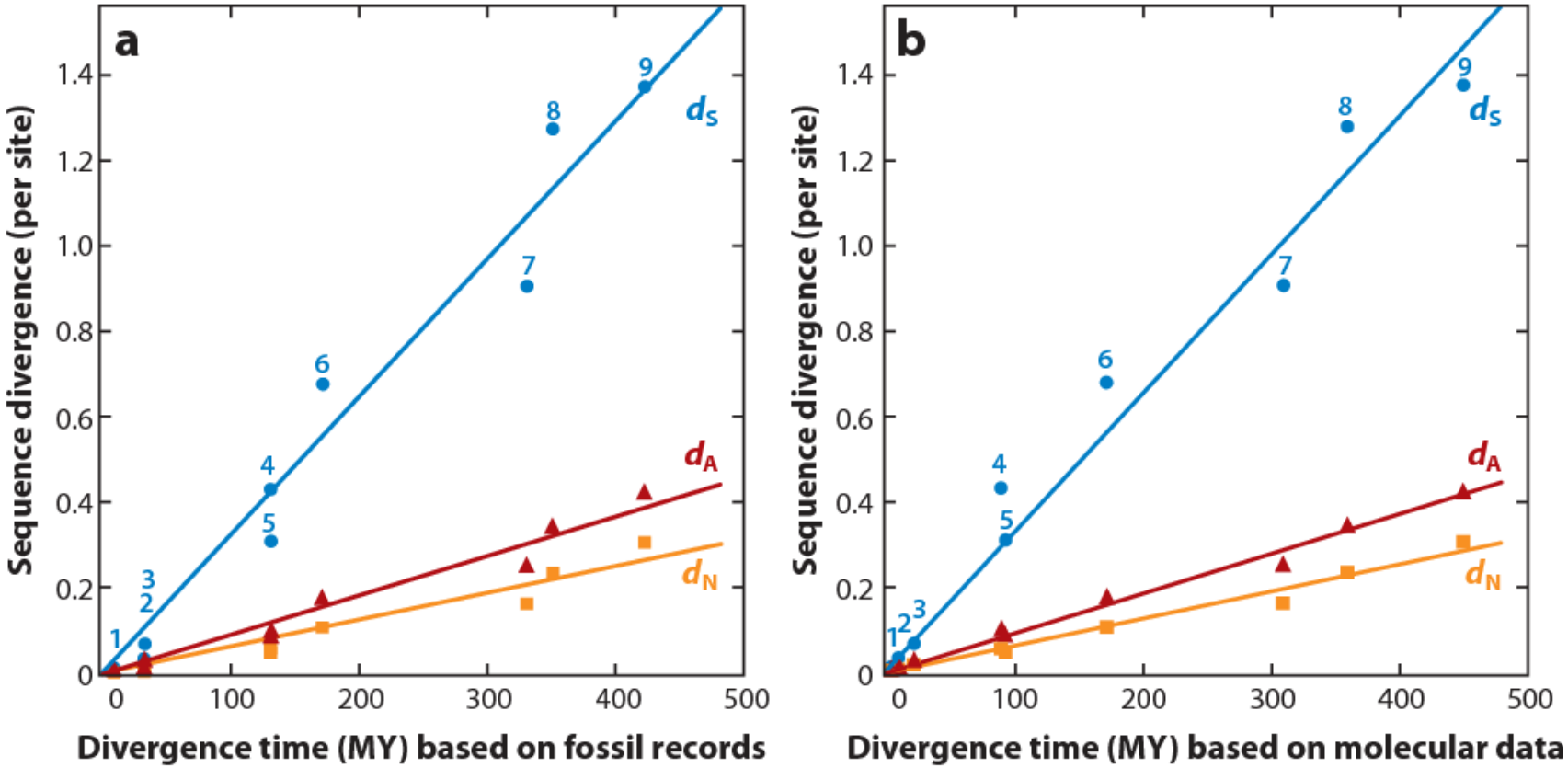
Beta globins: human-chicken

```
Human    β chain: MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLL
Chicken  β chain: MVHWTAEEKQLITGLWGKVNVAECGAEALARLL

Human    β chain: VVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
Chicken  β chain: IVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLT

Human    β chain: AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGN
Chicken  β chain: SFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGD

Human    β chain: VLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Chicken  β chain: ILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
```

There are a total of **44** mismatches, or (147 – 44) / 147 = **70.1 %** identical.

As expected, mouse β chain is '*closer*' to that of human than chicken's.

Evidence of the molecular clock

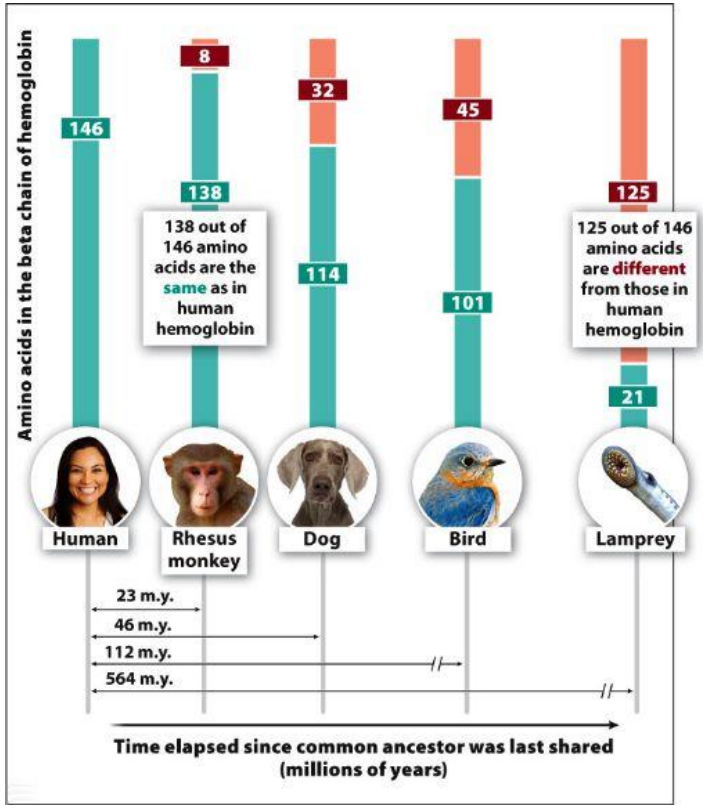


Linear relationships of the number of amino acid substitutions per residue (**d**A) and the numbers of synonymous (**d**S) and nonsynonymous (**d**N) nucleotide substitutions per site, with divergence times based on the fossil record (**a**) and molecular data from 10 vertebrates (**b**).

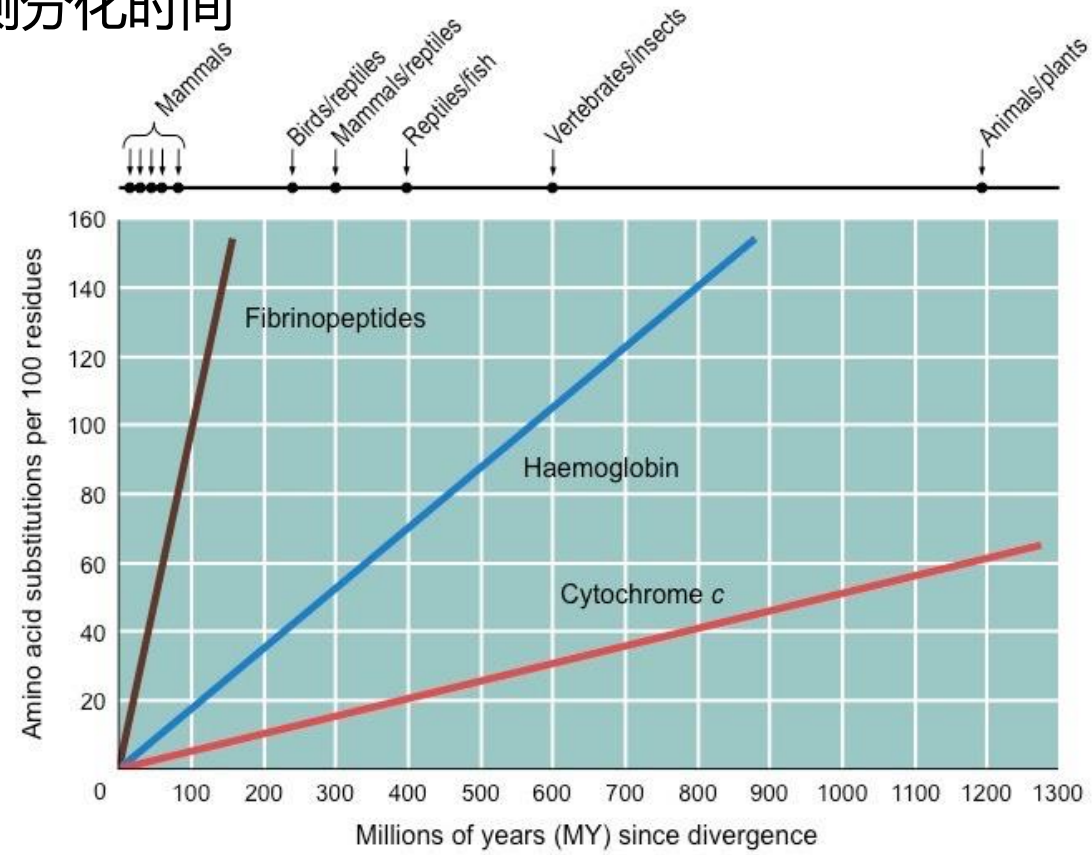
The neutral theory of molecular evolution in the genomic era. *Annu Rev Genom Hum Genet* 2010

分子钟推测分化时间

- 物种分化时间推断：化石证据（放射性同位素衰变）
- 由于化石证据零散不完整，可采用分子钟推测分化时间

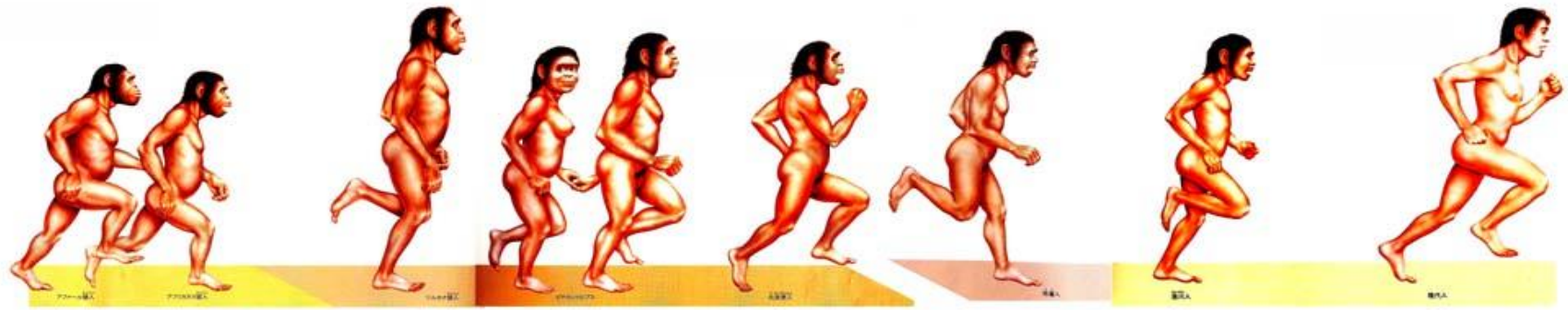


If a gene which mutates at a rate of 1 bp per 100,000 years has 6 bp different, divergence occurred 600,000 years ago.



Limitation: Different genes may change at different rates.
Haemoglobin mutates more rapidly than cytochrome c.

人类演化

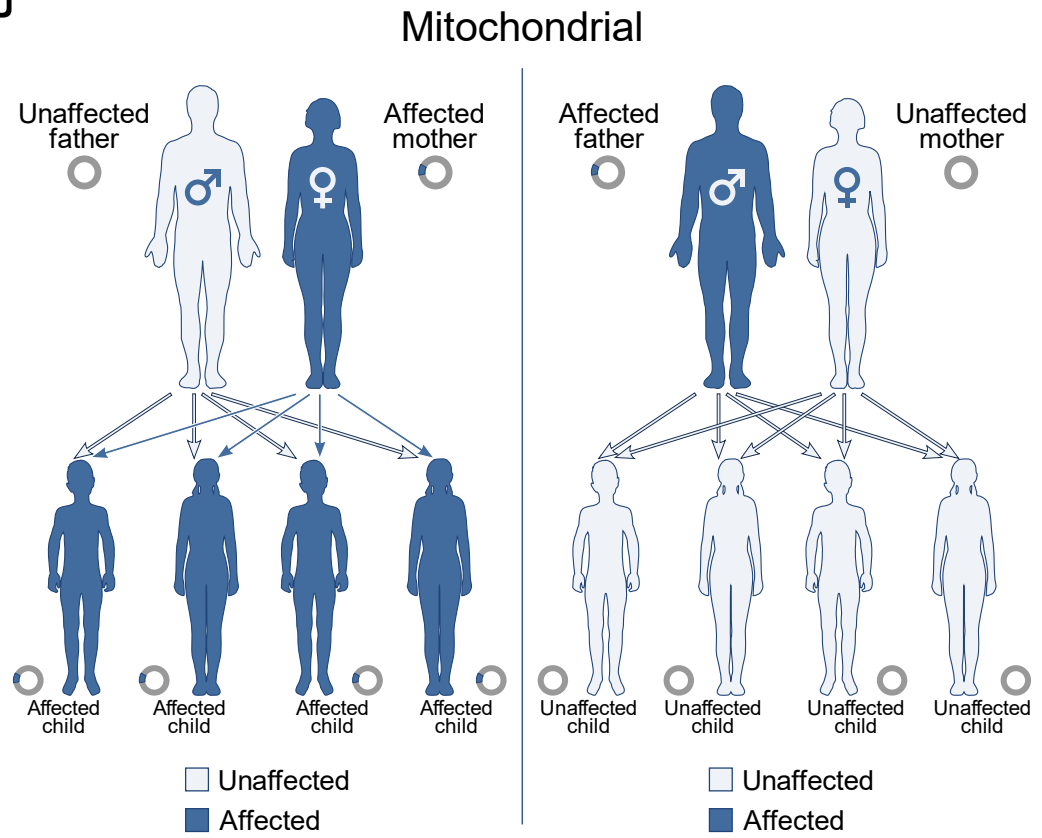


	线粒体	Y染色体	常染色体
碱基数	16,569	57,227,415	~3,000,000K
遗传方式	母系	父系	父母系
是否重组	否	否	是
突变速率	快	慢	慢
最近共同祖先MRCA	线粒体夏娃Mitochondrial Eve	Y染色体亚当 Y-chromosomal Adam	-

分子钟与人类演化

由于线粒体DNA（mtDNA）母系遗传，不存在重组且突变率高（是核DNA的10倍左右），可用于推断较短时期内分化时间

- 线粒体夏娃：其mtDNA存在于任何一位现存人类体内，根据分子时钟估算，距今大约14万年前，与相对的父系祖先“Y染色体亚当”，是生存在不同年代。
- Y染色体亚当：推断发现所有今天的人类都是生存于6万年前非洲大陆的男人的后裔。



https://www.ncbi.nlm.nih.gov/nuccore/NC_012920.1

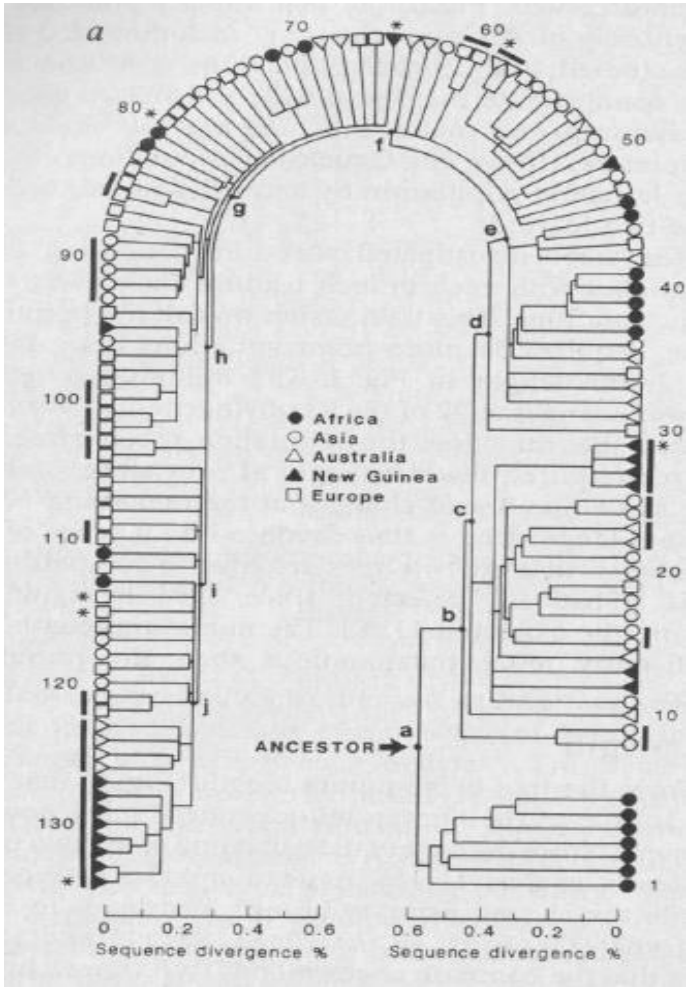
Evolutionary Tree of Humans (mtDNA)

mtDNA analysis supports the “Out of Africa” Hypothesis.

- The evolutionary tree separates one group of Africans from a group containing all five populations.
- African population was the most diverse.
- Tree was rooted on branch between groups of greatest difference.

Out of Africa:

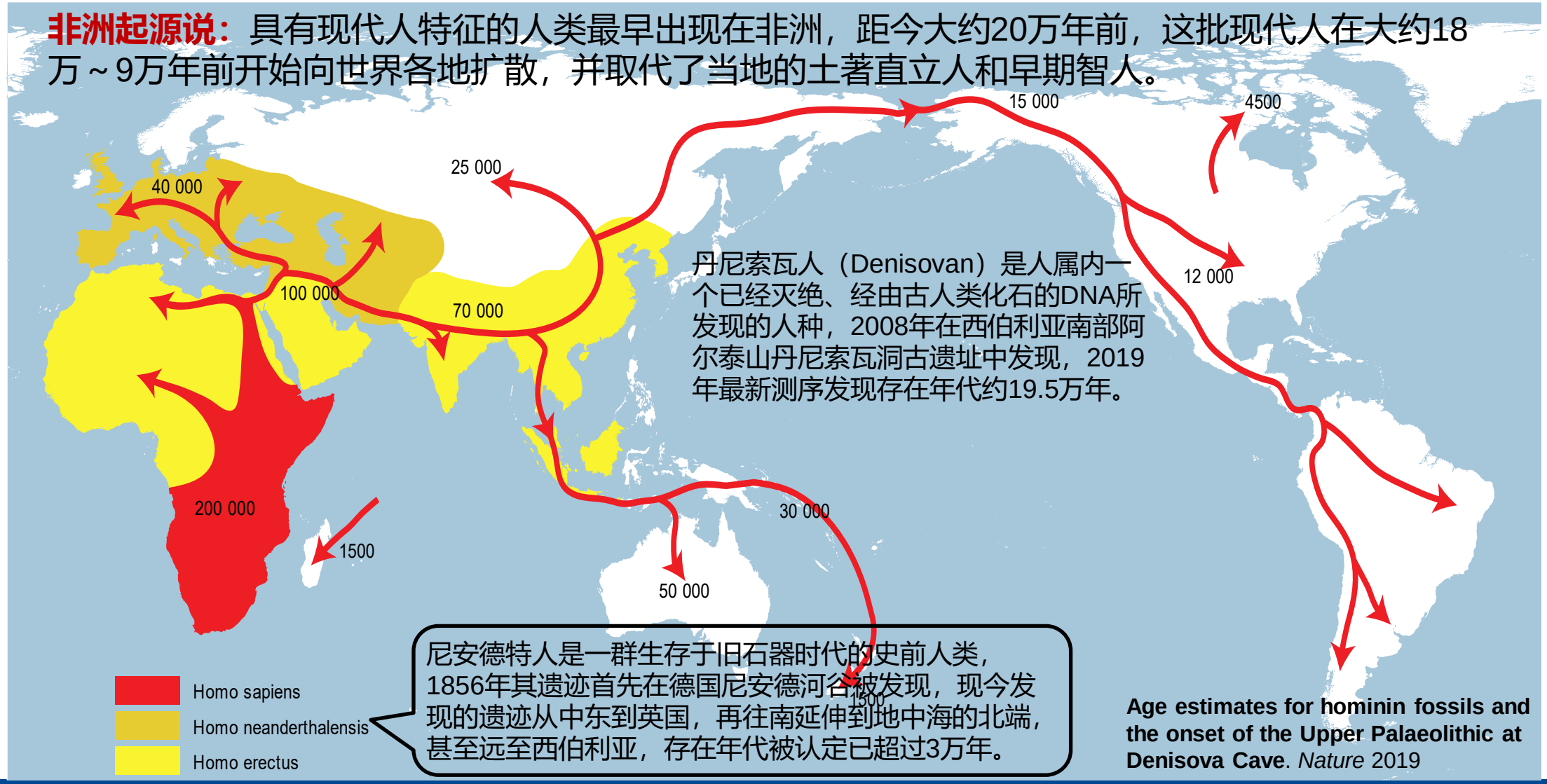
- Humans evolved in Africa ~150,000 years ago.
- Humans migrated out of Africa, replacing other humanoids around the globe.
- There is no direct descendence from Neanderthals.



African populations and the evolution of human mitochondrial DNA. *Science* 1991. DOI: 10.1126/science.1840702

人类起源与迁徙

非洲起源说：具有现代人特征的人类最早出现在非洲，距今大约20万年前，这批现代人在大约18万～9万年前开始向世界各地扩散，并取代了当地的土著直立人和早期智人。



丹尼索瓦人（Denisovan）是人属内一个已经灭绝、经由古人类化石的DNA所发现的人种，2008年在西伯利亚南部阿尔泰山丹尼索瓦洞古遗址中发现，2019年最新测序发现存在年代约19.5万年。

尼安德特人是一群生存于旧石器时代的史前人类，1856年其遗迹首先在德国尼安德河谷被发现，现今发现的遗迹从中东到英国，再往南延伸到地中海的北端，甚至远至西伯利亚，存在年代被认定已超过3万年。

Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature* 2019

古基因组学（Paleogenomics）

是什么奠定了我们成为今天独特的人类?

2022年诺贝尔生理学/医学奖授予斯万特·帕博（Svante Pääbo）

- 表彰他“在已灭绝的古人类基因组和人类进化方面的发现”
- 开创性研究催生了“**古基因组学**”这一全新学科

尼安德特人基因组测序

- 1997年，完成对尼安德特人线粒体DNA的测序，这是人类测出的第一个已灭绝的古人类DNA序列，证明尼安德特人是人类进化过程中的死胡同
- 2001年，人类基因组计划宣告完成，得到了现代人基因组序列的帕博，开发了辨别古DNA和现代人DNA的方法，能够将古DNA比对到现代人基因组中
- 2010年，帕博成功完成了尼安德特人的全基因组序列测序，测序精度达到了50层，证明人类祖先离开非洲后曾与尼安德特人有基因交流

丹尼索瓦人的发现

- 人类历史上第一个仅凭DNA证据就命名的人类新物种，付巧妹负责提取DNA并测序，比对发现与尼安德特人线粒体序列不同

分子序列演化

DNA序列演化

对于两条长度为n的DNA序列，不同的碱基对为n_d，且转换为（transition）P，颠换为（transversion）Q，则

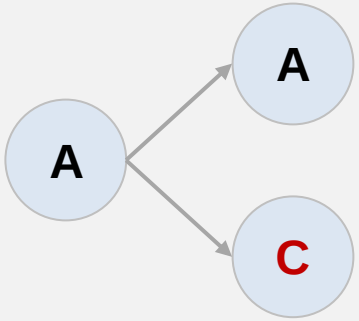
- $n_d = P + Q$
- 通常情况下，转换出现频率高于颠换，转换/颠换比kappa=P/Q
- 当序列分化程度较小时，且核苷酸替代是随机发生时，则Q=2P
- 两条序列的分化或差异程度p可表示为：$p = n_d / n$

然而，分化参数p未考虑：**观测到的替换数小于真实发生的替换数。**
因此，需要构建数学模型进行模拟演化变异。

Why observed substitutions < real ones?

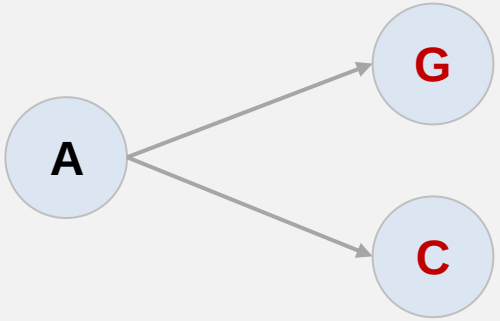
1 difference

Single substitution



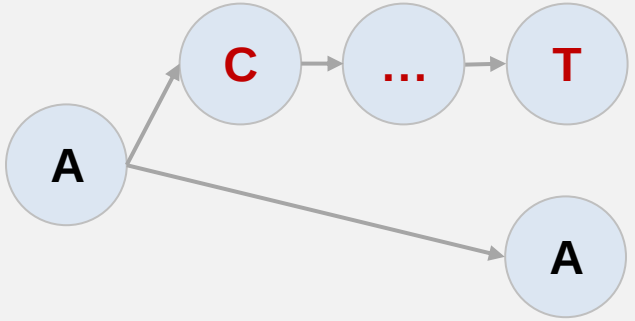
1 change

Coincidental substitution



2 changes

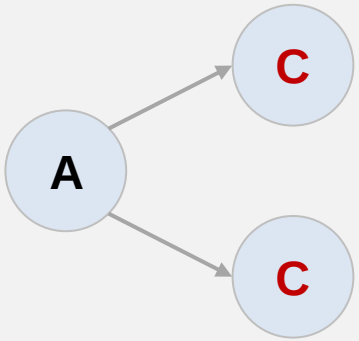
Multiple substitution



n changes

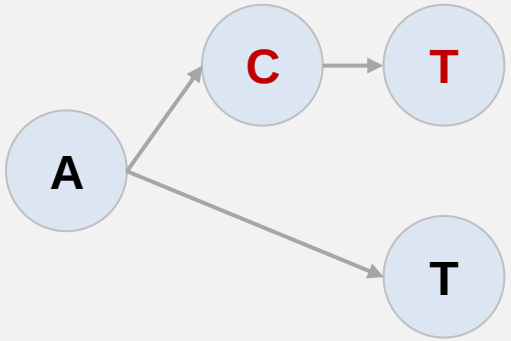
No difference

Parallel substitution



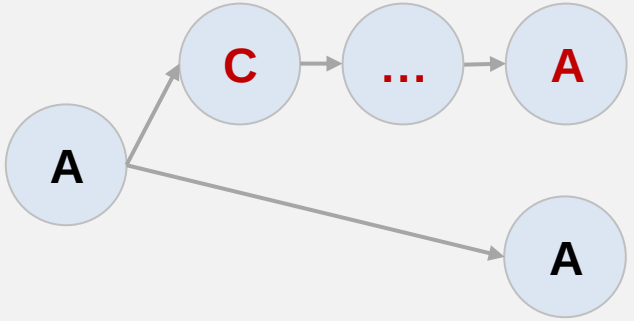
2 changes

Convergent substitution



3 changes

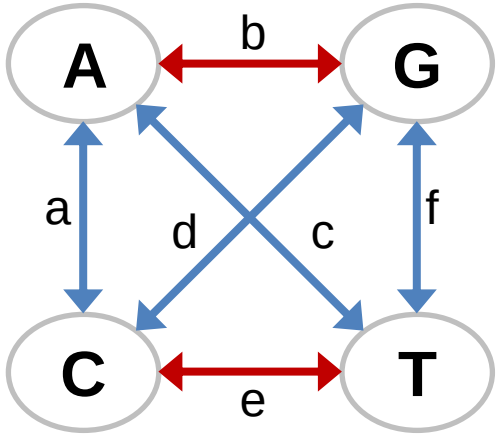
Back substitution



n changes

DNA substitution models

DNA substitutions models are frequently used in molecular phylogenetic analyses and evolutionary distance estimations.



transitions 转换

transversions 颠换

transitions 转换

- **Transitions**: changes within purines (A↔G) or pyrimidines (T↔C)
- **Transversions**: changes between a purine and a pyrimidine

Model	Exchangeability parameters	Base frequency parameters	Reference
JC69 (or JC)	a=b=c=d=e=f	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$	Jukes and Cantor (1969)
F81	a=b=c=d=e=f	all π_I values free	Felsenstein (1981)
K2P (or K80)	a=c=d=f, b=e	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$	Kimura (1980)
HKY85	a=c=d=f, b=e	all π_I values free	Hasegawa et al. (1985)
K3ST (or K81)	a=f, c=d, b=e	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$	Kimura (1981)
TN93	a=c=d=f, b≠e	all π_I values free	Tamura and Nei (1993)
SYM	all exchangeability parameters free	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$	Zharkikh (1994)
GTR (or REV)	all exchangeability parameters free	all π_I values free	Tavaré (1986)

Modeling sequence evolution by Jukes-Cantor Model

假定：任一位点的核苷酸替换频率相等，且每一位点的核苷酸每年以α的概率演变为其他三种核苷酸的一种，那么

- 一个核苷酸替换为其它三种核苷酸的概率为$\gamma = 3\alpha$;
- 假设在t年前分化出两条核苷酸序列X和Y:
  - q_t表示X和Y之间**相同**核苷酸的比例值
  - p_t表示X和Y之间**不同**核苷酸的比例值
- 因此，$p_t = 1 - q_t$

Jukes-Cantor Model (cont.)

对于X序列和Y序列，在时间t+1时(过了一年)，核苷酸依然相同的概率q_{t+1}，则来自两部分：

- 相同位点(q_t)：γ是替换为其它核苷酸的概率，则X和Y以$(1-\gamma)^2$的概率保持不变，当γ较小时，γ^2可以忽略，则$q_{t+1}=1-2\gamma$
- 不同位点($1-q_t$)：对于X和Y之间不同位点，假设在时间t时，X序列上的位点为i，Y序列上为j：
  - 如果X的i变j，而Y的j不变，概率为$\alpha(1-\gamma)=\gamma(1-\gamma)/3$;
  - 如果X的i不变，而Y的j变i，概率为$\alpha(1-\gamma)=\gamma(1-\gamma)/3$;

因此，X和Y相同的总概率为：$2\gamma(1-\gamma)/3$，当γ较小时，γ^2可以忽略，则近似为：$2\gamma/3$

Jukes-Cantor Model (cont.)

因此，在时间t+1时核苷酸相同的概率q_{t+1}概率为：

$$q_{t+1} = (1-2\gamma)q_t + (2/3)\gamma(1-q_t)$$

$$\Leftrightarrow q_{t+1} = q_t - 2\gamma q_t + \frac{2}{3}\gamma - \frac{2}{3}\gamma q_t$$

$$\Leftrightarrow q_{t+1} = q_t + \frac{2}{3}\gamma - \frac{8}{3}\gamma q_t$$

$$\Leftrightarrow q_{t+1} - q_t = \frac{2}{3}\gamma - \frac{8}{3}\gamma q_t$$

令 $\frac{d_q}{d_t} = q_{t+1} - q_t$

解其微分方程

$$\frac{d_q}{d_t} = \frac{2}{3}\gamma - \frac{8}{3}\gamma q$$

当初始条件$t = 0$且$q = 1$时

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{8}{3}\gamma t})$$

How do different species differ?

因此，每一位点替换的期望值d=2γt, 代入

$$q = 1 - \frac{3}{4}\left(1 - e^{-\frac{8}{3}\gamma t}\right), \text{且} d = 2\gamma t, \text{则}$$

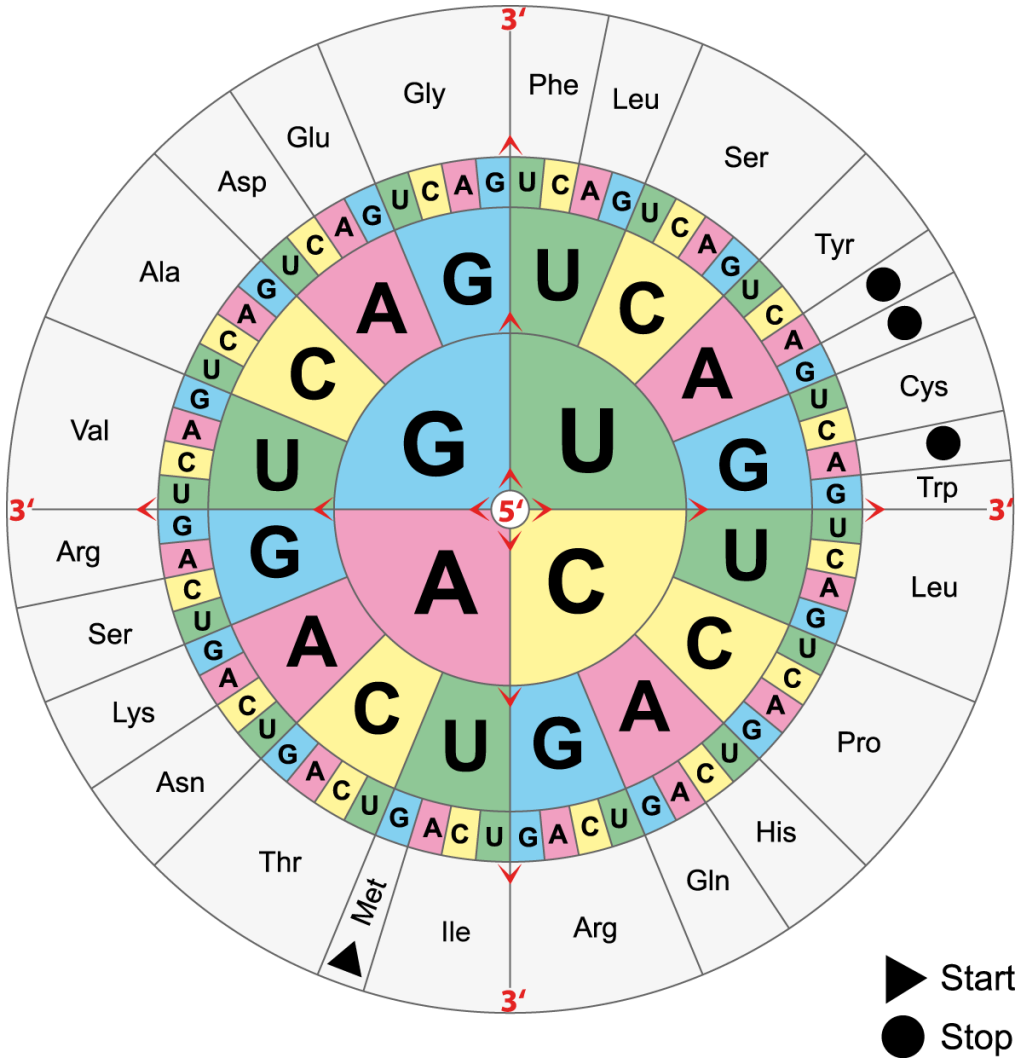
$$1 - p = 1 - \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right) \Leftrightarrow \frac{4}{3}\mathrm{p} = 1 - \mathrm{e}^{-\frac{4}{3}\mathrm{d}}$$

$$\Leftrightarrow e^{-\frac{4}{3}d} = 1 - \frac{4}{3}p \Leftrightarrow d = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

p为观测值，d为校正值

密码子使用偏性
Codon Usage Bias

密码子简并性（Codon Degeneracy)

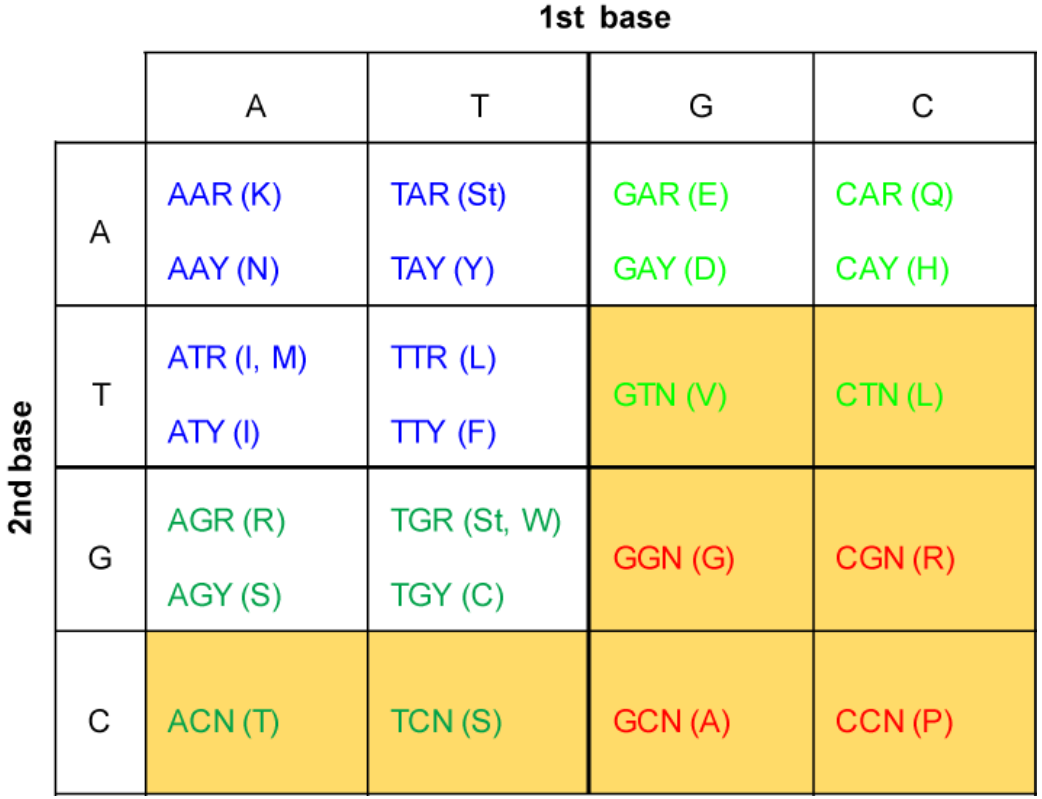


64个密码子编码20个氨基酸（标准密码子表）

Degeneracy is the redundancy of the genetic code

密码子简并性

n-fold degenerate site if only *n* of four possible nucleotides (A, C, G, T) at this position specify the same amino acid



2nd base \ 1st base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

_n_-fold degenerate site

_n_重简并位点

例如：

TTA的第一位：2重简并位点

TTA的第二位：0重简并位点

TTA的第三位：2重简并位点

CGA的第一位：2重简并位点

CGA的第二位：0重简并位点

CGA的第三位：4重简并位点

ATA的第三位：3重简并位点

Yu, J. (2007) *Genomics Proteomics Bioinformatics*, 5:1–6.

Zhang, Z. and Yu, J. (2011) *Genomics Proteomics Bioinformatics*, 9(1-2):21-29.

遗传密码表（Genetic code）

传统编排方式

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U
	UUA Leu UUG Leu		UAA Stop UAG Stop	UGA Stop UGG Trp	C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
			CAA Gln CAG Gln		
A	AUU Ile AUC Ile AUA Ile	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn	AGU Ser AGC Ser	U C
	AUG Met		AAA Lys AAG Lys	AGA Arg AGG Arg	A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G
			GAA Glu GAG Glu		

基于GC含量和嘌呤含量的编排方式

1st base

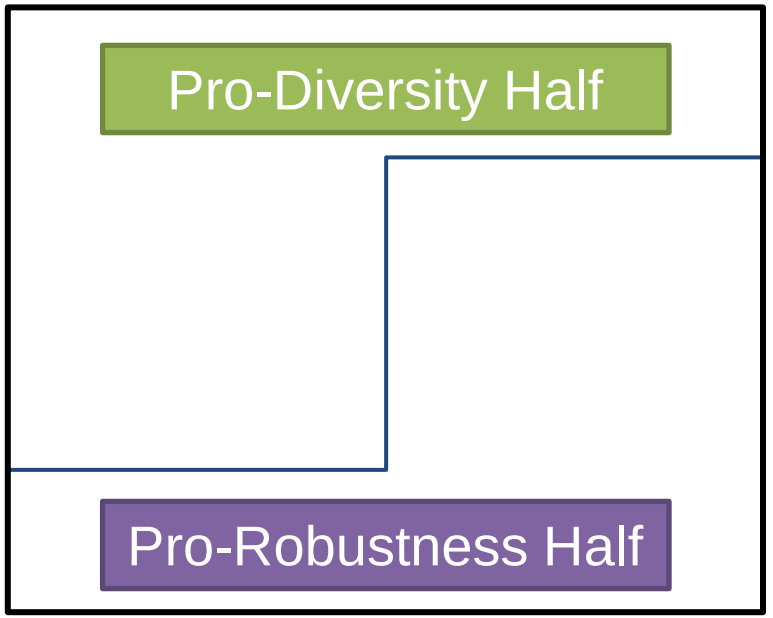
2nd base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

N=任何核苷酸，R=嘌呤（A/G），Y=嘧啶（T/C），St=终止密码子

遗传密码表：二分法

2nd base \ 1st base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

Yu, J. (2007) *Genomics Proteomics Bioinformatics*, 5:1–6.
Zhang, Z. and Yu, J. (2011) *Genomics Proteomics Bioinformatics*, 9(1-2):21-29.



Pro-Diversity Half: 32 codons, 15 amino acids as well as start and stop codons
Prof-Robustness Half: 32 codons, 8 amino acids; any substitution at third codon position will not lead to amino acid change

遗传密码表：四分法

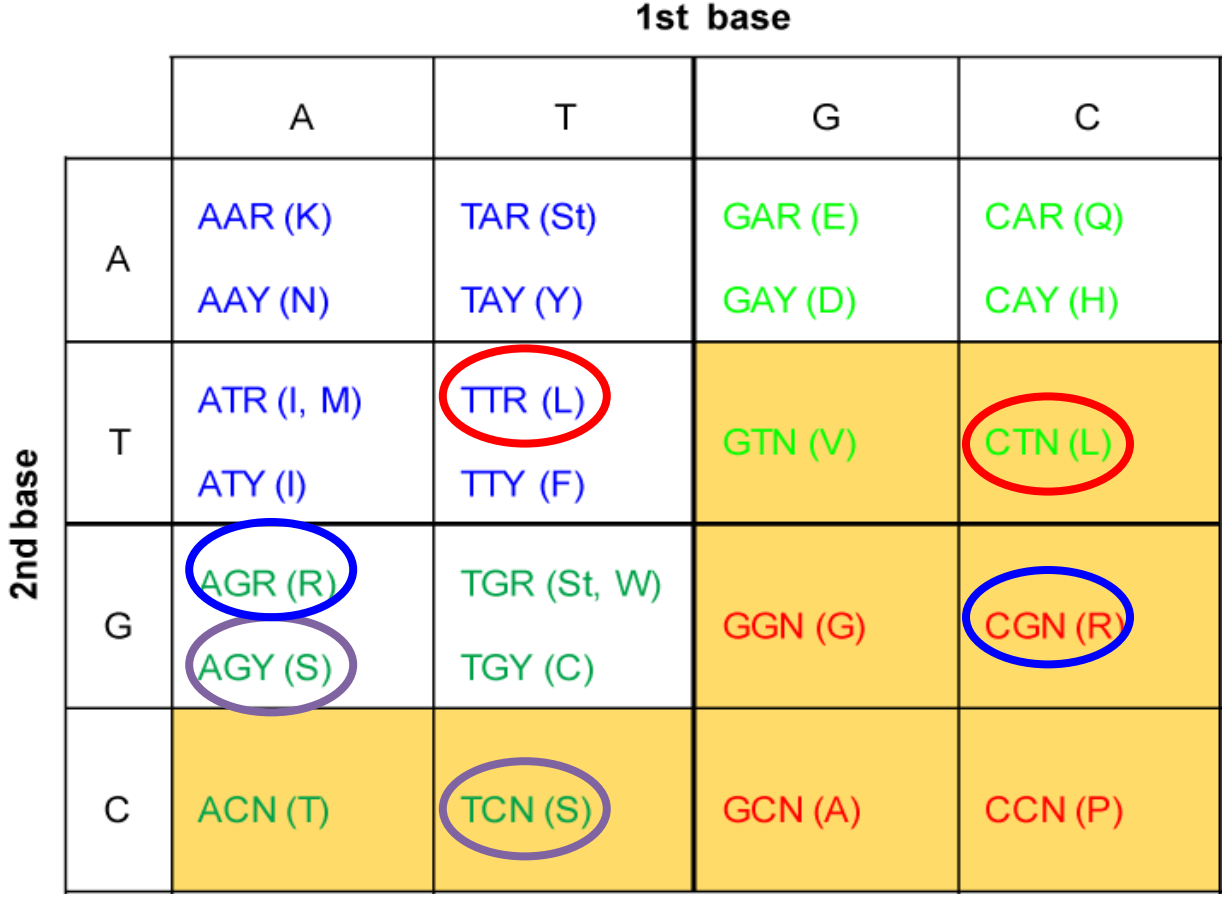
2nd base \ 1st base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

Yu, J. (2007) *Genomics Proteomics Bioinformatics*, 5:1–6.
Zhang, Z. and Yu, J. (2011) *Genomics Proteomics Bioinformatics*, 9(1-2):21-29.

AT Rich	**GCp1**
GCp2	**GC Rich**

AT Rich: AT at codon positions 1 and 2
GC Rich: GC at codon positions 1 and 2
GCp1: GC at codon position 1
GCp2: GC at codon position 2

Three amino acids with six-fold degenerate codons: Leu, Ser, Arg



Three amino acids with six-fold degenerate codons (Ser, Leu and Arg) are distributed:

- across PDH and PRH;
- among all four quarters.

Zhang, Z. and Yu, J. (2011) *Genomics Proteomics Bioinformatics*, 9(1-2):21-29.

Codon Usage

- There are 64 codons (61 sense codons plus 3 stop codons) that code for 20 amino acids.
- Theoretically, if there is no selection or mutation bias, it is expected that the codons encoding the same amino acid are on average in equal frequencies in protein coding regions of DNA.
- In practice, the frequencies of different codons for the same amino acid are usually different, and some codons are used more often than others.

Codon Usage Bias

- Codon usage bias is a phenomenon in which **synonymous codons** (that encode the same amino acid) **are used at different frequencies** across different species and between different genes within one species.
- Codon usage bias is **a result of a balance of mutation and selection** (Bulmer, M. 1991. Genetics).

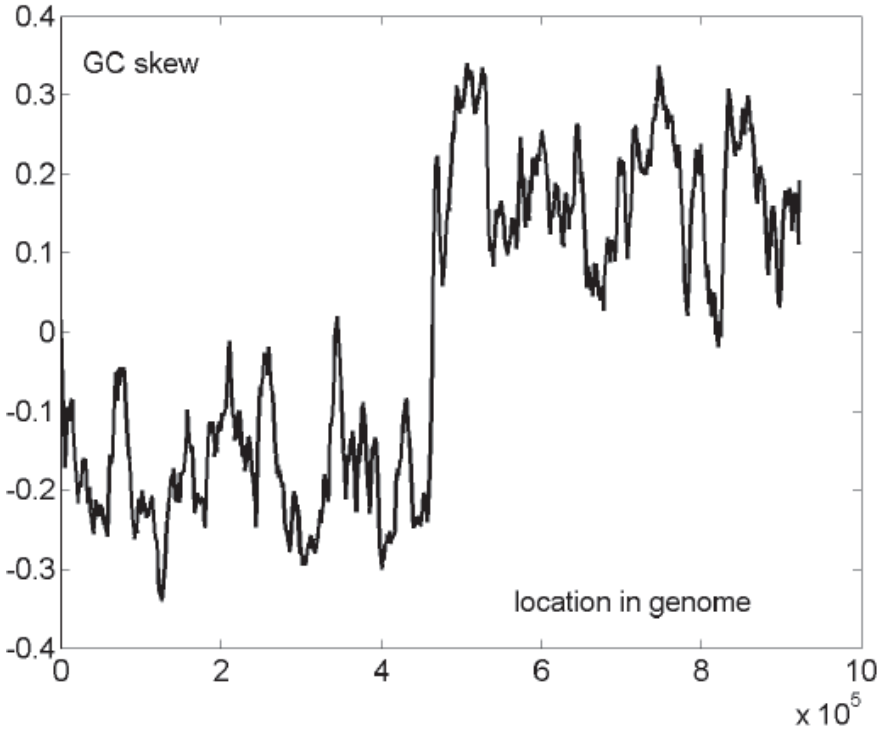
Codon	*E. coli*	*D. melanogaster*
GGG	0.02	0.03
GGA	0.00	0.28
GGT	0.59	0.26
GGC	0.39	0.43

Hershberg, R, Petrov, DA. 2008. Selection on codon bias. Annu Rev Genet 42:287-299.
Plotkin, JB, Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12:32-42.

Causes of Codon Usage Bias

Mutation

- GC content: ~20% to ~80% in bacteria (Sueoka, N. 1962. PNAS 48:582-591)
- Strand-related bias (Karlin, S, Campbell, AM, Mrazek, J. 1998. Annu Rev Genet 32:185-225)
- Horizontal gene transfer (Lawrence, JG, and Ochman, H. 1998. PNAS 95:9413–9417)
- Isochore structure (Bernardi, G. et al 1985. Science 228:953–958).

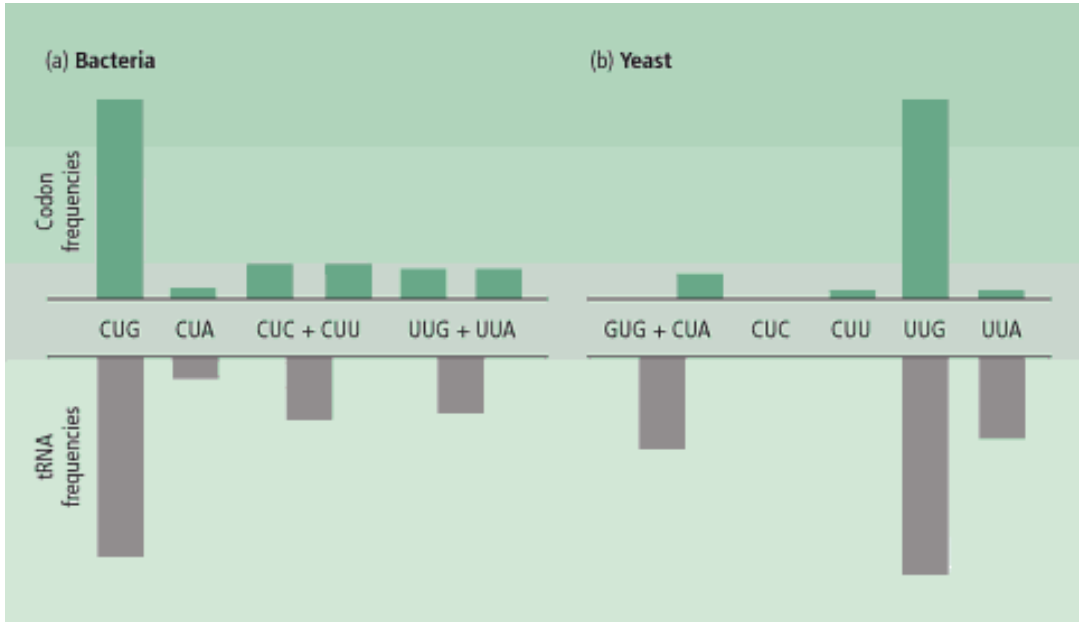


GC skew for the *B. burgoferri* genome (Ermolaeva, MD. 2001. Curr Issues Mol Biol 3:91-97)

Causes of Codon Usage Bias (cont.)

Selection

- Translational selection favors codons that correspond to the most abundant tRNA species (Akashi, H. *Genetics* 1994; Ikemura, T. *J Mol Biol* 1981)
- Positive correlation between codon usage bias and gene expression (Gouy, M and Gautier, C. *Nucleic Acids Res* 1982)
- Highly conserved sites have more biased codon usage (Stoletzki, N and Eyre-Walker, A. *Mol Biol Evol* 2007).



Relative frequencies of codons match tRNA abundances (Ridley M. 2004. Evolution (3rd ed.). Oxford: Blackwell publishing).

Codon Usage Bias (CUB)

- Natural selection on codon usage is more efficient in species with large effective population, such as, prokaryotes and unicellular eukaryotes.
- Codon bias can be used to predict protein expression level.

Goetz, R.M. and Fuglsang, A. (2005) Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli, *Biochem Biophys Res Commun,* ***327, 4-7.***

Henry, I. and Sharp, P.M. (2007) Predicting gene expression level from codon usage bias, *Mol Biol Evol,* ***24, 10-12.***

Roymondal, U., Das, S. and Sahoo, S. (2009) Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome, *DNA Res,* ***16, 13-30.***

Measures for estimating CUB

- Reference-based measures: highly-expressed genes as reference datasets
  - ✓ **CAI** (Codon Adaptation Index; Sharp and Li 1987)
  - ✓ Fop (Frequency of Optimal Codons; Ikemura 1981)
  - ✓ RSCU (Relative Synonymous Codon Usage; Sharp et al 1986)
  - ✓ *E(g)* (Karlin et al 2001)
- Distribution-based measures: without requiring any prior knowledge
  - ✓ **CDC** (Codon Deviation Coefficient; Zhang and Yu 2012)
  - ✓ N_c (Effective Number of Codons or ENC; Wright 1990)
  - ✓ N_c' (a variant of N_c; Novembre 2002)
  - ✓ CBI (Codon Bias Index; Morton 1993)

Reference-based measure: CAI

- CAI: range from 0 (no bias) to 1 (maximum bias)

$$\mathrm{CAI} = \left(\prod_{i=1}^{L} w_i\right)^{1/L}$$

w_i = relative adaptation value for codon i
L = number of codons

$$w_i = \frac{\mathrm{RSCU}_i}{\mathrm{RSCU}_{i\,\max}}$$

RSCU_i = Relative synonymous codon usage for codon i
$\mathrm{RSCU}_{i\max}$ = maximum RSCU_i in the reference set

$$\mathrm{RSCU}_i = \frac{x_i}{\frac{1}{N}\sum_{j=1}^{N} x_j}$$

X_i = Number of occurrences of codon i
N = 1, 2, 3, 4 or 6

Distribution-based measure: CDC

CDC adopts a model factoring GC and purine contents to obtain the expect ed codon frequencies and compare the expected against the observed.

- The universal set U is defined as U = {A, T, G, C}
- GC ($S=G \cup C$) and purine ($R=A \cup G$) contents are two subsets of U.

Thus,

- $G = (G \cup C) \cap (A \cup G) = S \cap R = SR$
- $C = (G \cup C) \cap (A \cup G)^c = S \cap R^c = S(1 - R)$
- $A = (G \cup C)^c \cap (A \cup G) = S^c \cap R = (1 - S)R$
- $T = (G \cup C)^c \cap (A \cup G)^c = S \cap R^c = (1 - S)(1 - R)$

Apply to nucleotide contents at three codon positions (i=1,2,3), so:

$A_i = (1-S_i)R_i, \quad T_i = (1-S_i)(1-R_i), \quad G_i = S_iR_i, \quad C_i = S_i(1-R_i)$

$\therefore$ Expected frequency of codon xyz = $x_1y_2z_3$, where x, y, z $\in$ {A, T, G, C}

Zhang and Yu (2010)
Biology Direct 5:63

Distribution-based measure: CDC

- CDC (Codon Deviation Coefficient) employs the cosine distance metric to render a distance coefficient based on the cosine of the angle (θ) between observed and expected codon vectors.

$$\mathrm{CDC} = 1 - \cos(\theta) = 1 - \frac{\sum_{xyz} \pi_{xyz} \times \hat{\pi}_{xyz}}{\sqrt{\sum_{xyz} \pi_{xyz}^{\ 2} \times \sum_{xyz} \hat{\pi}_{xyz}^{\ 2}}}$$

$\hat{\pi}$: observed codon usage

π : expected codon usage

- CDC represents codon usage bias, ranging from 0 (no bias) to 1 (maximum bias).

选择压力、同义和非同义替换率

遗传密码表（Genetic code）

2nd base \ 1st base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

N=任何核苷酸，R=嘌呤（A/G），Y=嘧啶（T/C），St=终止密码子

同义与非同义替换

- **同义替换—Synonymous** (silent) substitution: do not change amino acid

  TTT (Phe) <-> TTC (Phe)

- **非同义替换—Nonsynonymous** substitution: change amino acid

  TTT (Phe) <-> TTA (Leu)

$$\text{Ka or dN} \approx \frac{\text{nonsynonymous substitution (Nd)}}{\text{nonsynonymous site (N)}}$$

$$\text{Ks or dS} \approx \frac{\text{synonymous substitution (Sd)}}{\text{synonymous site (S)}}$$

Ka: nonsynonymous substitution rate, or nonsynonymous substitutions per nonsynonymous site
Ks: synonymous substitution rate, or synonymous substitutions per synonymous site

同义与非同义替换率

- Estimating nonsynonymous and synonymous substitution rates (denoted as Ka and Ks) is important in **understanding the dynamics of molecular sequence evolution**.
- Synonymous (silent) mutations are largely **invisible to natural selection**, while nonsynonymous (amino-acid replacing) mutations may be under **strong selective pressure.**
- Comparison of the rates of fixation of those two types of mutations provides a powerful tool for understanding the mechanisms of **DNA sequence evolution**.
- Variable nonsynonymous/synonymous rate ratios among lineages may indicate **adaptive evolution or relaxed selective constraints** along certain lineages.

选择压力（Selective Pressure）

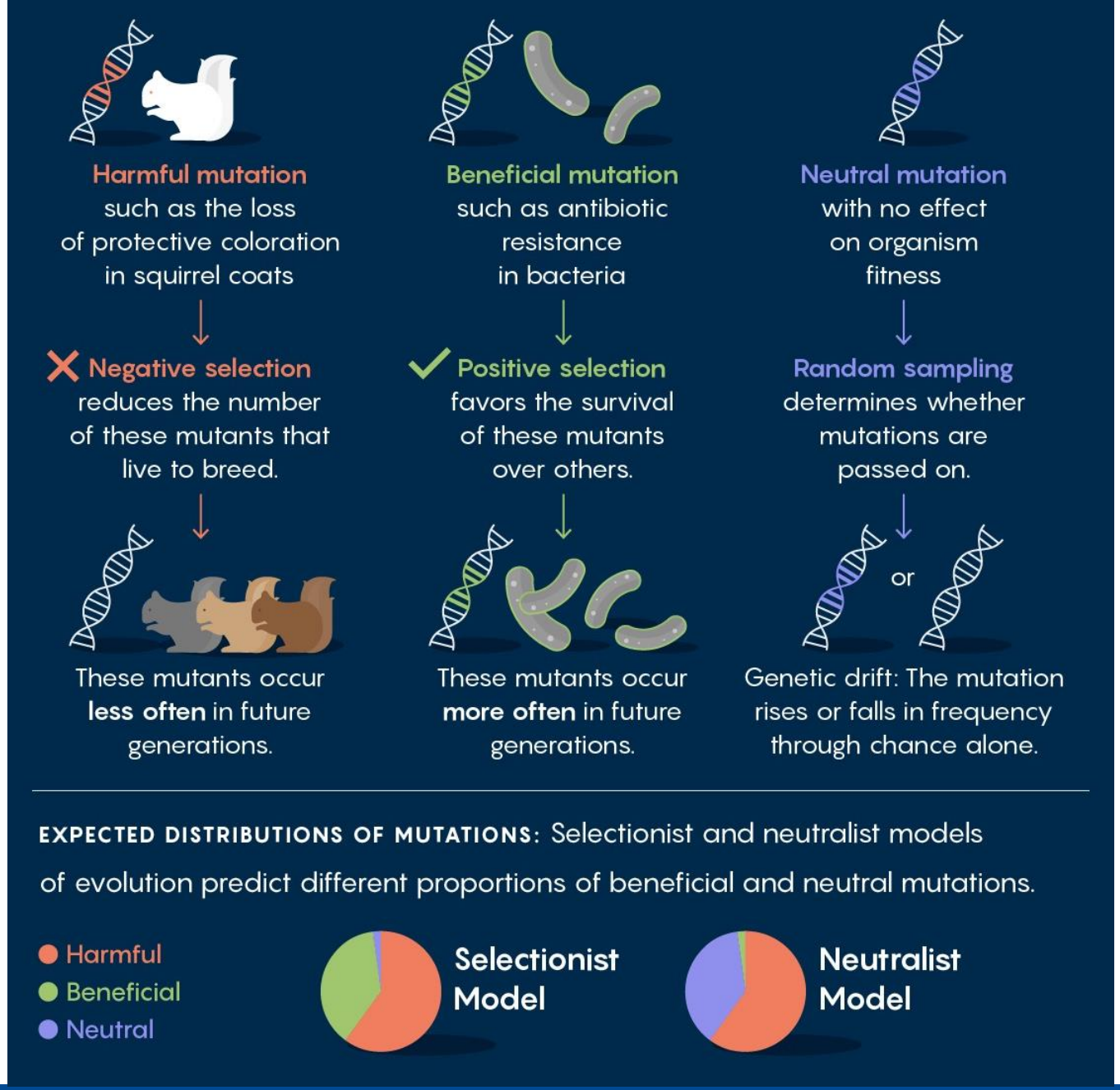
- Ks, synonymous substitution rate, can be thought as the background mutation rate of evolution.
- Selective Pressure is defined as Ka/Ks
  - ✓Ka>Ks：**正选择/适应性进化**，positive/adaptive selection, indicating a functional benefit to diversify the amino acid sequence and possibly obtaining new functions
  - ✓Ka=Ks：**中性突变**，neutral mutation
  - ✓Ka<Ks：**负选择/纯化选择**，negative/purifying selection, eliminating deleterious mutations and keeping the protein as it is.

Molecular Evolution

Negative Selection

Positive Selection

Neutral Mutation



https://www.quantamagazine.org/neutral-theory-of-evolution-challenged-by-evidence-for-dna-selection-20181108/

Three steps to estimate Ka and Ks

- **counting sites:**

  synonymous sites (S) and nonsynonymous (N) sites
- **counting substitutions**:

  synonymous (Sd) and nonsynonymous (Nd) substitutions
- **correcting for multiple substitutions:**

  due to observed substitutions < real ones

Key problems in estimation

Example:

seq1: **ATC CTA CGA GGC CGC GTA**

seq2: **ATC TTA CGA AGA CGC TCT**

	同义	非同义	总数
位点	synonymous sites (S)	nonsynonymous sites (N)	S + N = 18
替换	synonymous substitutions (Sd)	nonsynonymous substitutions (Nd)	Sd + Nd = 6

(1) How to determine S, N, Sd and Nd?

Ka ≈ Nd / N, nonsynonymous substitutions per nonsynonymous site

Ks ≈ Sd / S, synonymous substitutions per synonymous site

Key problems in estimation (cont.)

Take **counting substitutions** as an example:

seq1: **ATC CTA CGA GGC CGC GTA**

seq2: **ATC TTA CGA AGA CGC TCT**

- **CTA (Leu) <-> TTA (Leu):**

One difference: 1 evolutionary pathway

- **GGC (Gly) <-> AGA (Arg)**

Two differences: 2 evolutionary pathways

I: GGC (Gly) <-> AGC (Ser) <-> AGA (Arg)

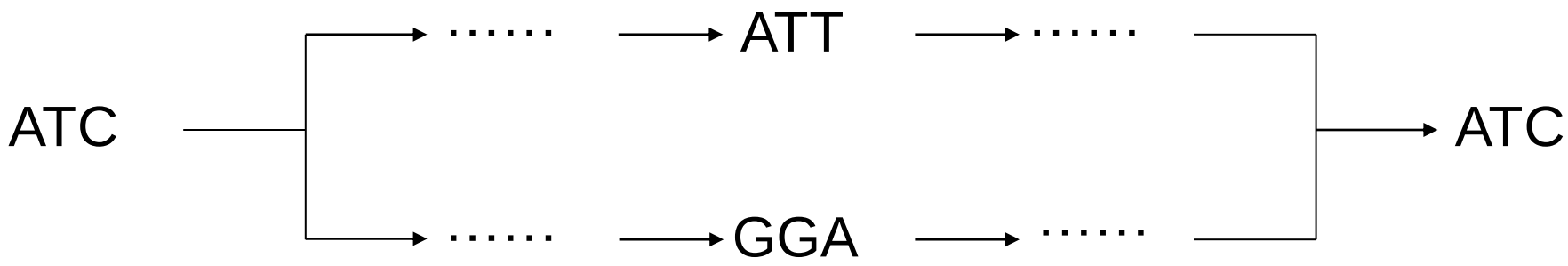
II: GGC (Gly) <-> GGA (Gly) <-> AGA (Arg)

- **GTA (Val) <-> TCT (Ser)**

Three differences: 6 evolutionary pathways……

Key problems in estimation (cont.)

The observed number of substitutions **underestimates** the real number of substitutions, as sequences diverge over time.



(3) How to correct accurately for multiple substitutions?

Ka (the expected) > Nd / N (the observed)

Ks (the expected) > Sd / S (the observed)

Hurst, L. D. (2002) *Trends in Genetics*, 18(9): 486-487.

Methods for estimating Ka and Ks

Method	Model			Reference
	Sites[a]	Substitutions[b]	Correcting[c]	
Approximate Methods				
NG	JC	JC	JC	Nei and Gojobori (1986)
LWL	JC	K2P	K2P	Li, Wu, and Luo (1985)
MLWL	K2P	K2P	K2P	Tzeng, Pan, and Li (2004)
LPB	–	–	K2P	Li (1993) and Pamilo and Bianchi (1993)
MLPB	–	–	K2P	Tzeng, Pan, and Li (2004)
YN	HKY	HKY	HKY	Yang and Nielsen (2000)
MYN	TN	TN	TN	Zhang, Li, and Yu (2006)
Maximum-Likelihood Methods				
GY		HKY		Goldman and Yang (1994)
MS		Model Selection		Zhang et al. (2006)
MA		Model Averaging		Zhang et al. (2006)

NG

- **Counting Sites**

seq1: **ATC CTA CGA GGC CGC GTA**

seq2: **ATC TTA CGA AGA CGC TCT**

	ATC (Ile)		
All possible substituted codons	TTC (Phe)	ACC (Thr)	ATT (Ile)
	CTC (Leu)	AAC (Asn)	ATA (Ile)
	GTC (Val)	AGC (Ser)	ATG (Met)
Synonymous sites (s)	0	0	2/3
Nonsynonymous sites (n)	1	1	1/3

Synonymous sites: S = ∑s

Nonsynonymous sites: N = ∑n

Nei, M. and Gojobori, T. (1986) *Mol. Biol. Evol.*, **3**, 418-426.

NG (cont.)

- **Counting Substitutions**

seq1: **ATC CTA CGA GGC CGC GTA**

seq2: **ATC TTA CGA AGA CGC TCT**

- **CTA (Leu) <-> TTA (Leu)**
  sd=1, nd=0
- **GGC (Gly) <-> AGA (Arg)**

```
I:  GGC (Gly) <->  AGC (Ser) <->  AGA (Arg)
II: GGC (Gly) <->  GGA (Gly) <->  AGA (Arg)
```

  sd = 0.5, nd = 1.5
- **GTA (Val) <-> TCT (Ser)**
  .....

Synonymous substitutions: Sd = ∑sd

Nonsynonymous substitutions: Nd = ∑nd

NG (cont.)

- **Correction: Jukes-Cantor formula**

$$d = -\frac{3}{4}\log e(1-\frac{4}{3}\times d0)$$

$$\mathrm{Ka} = -\frac{3}{4}\log e(1-\frac{4}{3}\times\frac{\mathrm{Nd}}{\mathrm{N}})$$

$$\mathrm{Ks} = -\frac{3}{4}\log e(1-\frac{4}{3}\times\frac{\mathrm{Sd}}{\mathrm{S}})$$

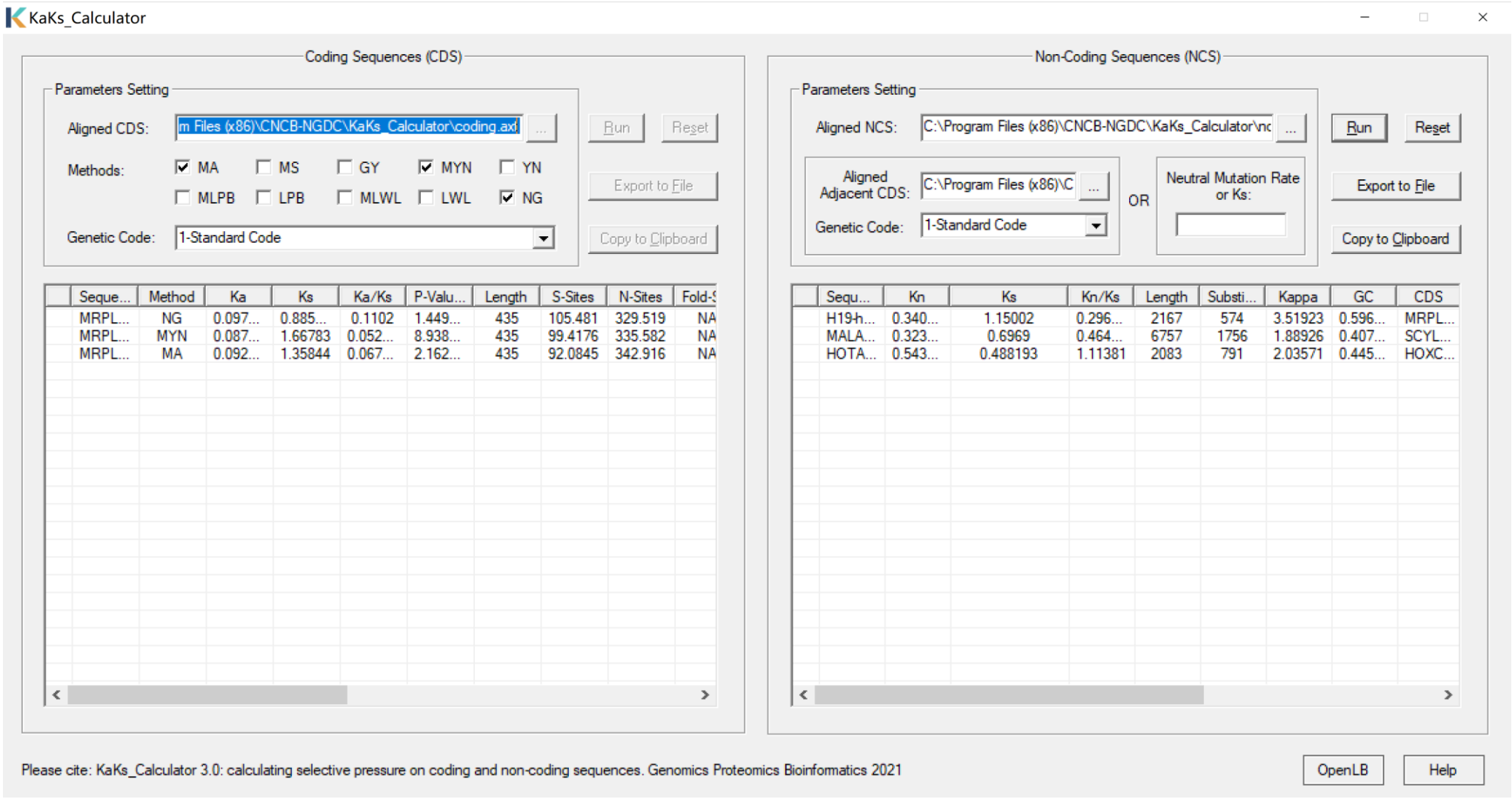
Assume:

(1) If Nd/N = 0.15, then Ka = 0.17.

(2) If Sd/S = 0.4, then Ks = 0.57.

Note: d_0 should be smaller than 3/4.

KaKs_Calculator 3.0



非编码序列的自然选择压力计算

非编码序列

G A G C C C G A T

G A G C C A G A T

Noncoding
Nucleotide substitution

$$Kn = \frac{\text{nucleotide substitution}}{\text{site}}$$

临近编码序列

T T T G C C T T T

Phe Phe Phe Leu

T T C G C C T T A

Synonymous substitution

Nonsynonymous substitution

$$Ks = \frac{\text{synonymous substitution}}{\text{synonymous site}}$$

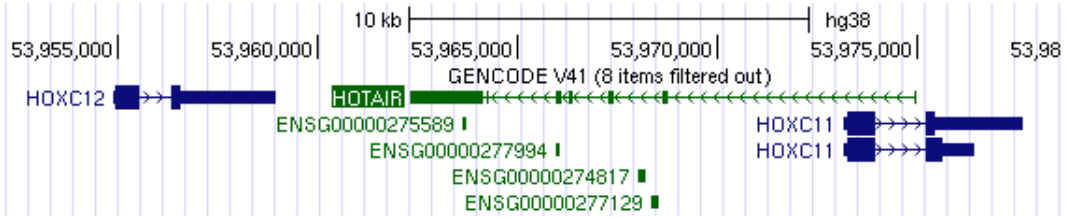
非编码序列自然选择压力计算模型：

$$\xi = \frac{Kn}{Ks} \begin{cases} \xi > 1 & \text{adaptive/positive selection} \\ \xi \approx 1 & \text{neutral mutation} \\ \xi < 1 & \text{purifying/negative selection} \end{cases}$$

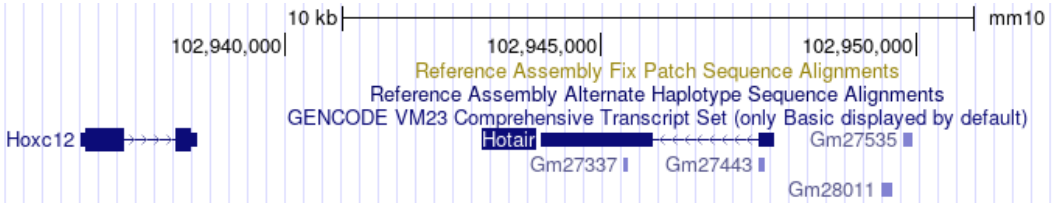
自然选择压力及其替换率计算

Non-coding			Coding			
Gene symbol	*Kn*	$\xi = Kn / Ks$	Gene symbol	*Ka*	*Ks*	$\omega = Ka / Ks$
H19	0.340	0.296	*MRPL23*	0.088	1.150	0.077
MALAT1	0.324	0.464	*SCYL1*	0.040	0.697	0.058
HOTAIR	0.544	1.114	*HOXC12*	0.020	0.488	0.041

Human



Mouse



- **正选择：*HOTAIR***
- **负选择：*H19*, *MALAT1***

KaKs_Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. *Genomics, Proteomics & Bioinformatics* 2022

Dating the divergence time

"*Estimates of the times since divergence of …… were made by* ***calculation of synonymous base substitution rates (Ks values)*** *and* ***assumption of a mutation rate of 1.4 × 10^{-8} substitutions per synonymous site per year*** *(Koch et al., 2000). **The median Ks values** between orthologous genes …… **(0.49 to 0.51)** are representative of the Brassica–Arabidopsis divergence, which we estimate at **17 to 18 MYA**.*"

$$\text{Divergence Time} = \frac{Ks/2}{1.4\times 10^{-8}}$$

$$= \frac{0.25 \text{ substitutions/site}}{1.4\times 10^{-8} \text{ substitutions/site/year}}$$

$$= 0.1785\times 10^{8} \text{ year} = 17.85\times 10^{6} \text{ year} = 17.85 \text{ MYA}$$

Yang, T.J, et al (2006) *Plant Cell*. 18(6): 1339–1347.

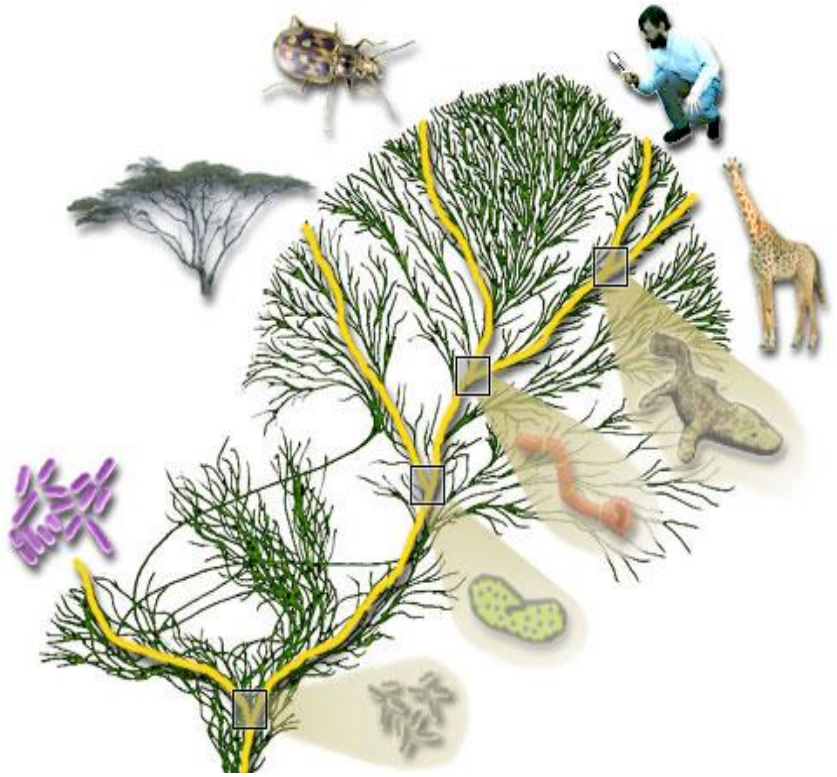
Statistical significance of Ka/Ks by Fisher exact test

	synonymous	nonsynonymous
substitution	Sd	Nd
site	S	N

The null hypothesis is Sd/S = Nd/N or neutral mutation.

Reject the null hypothesis if Sd/S is significantly greater (negative selection) or smaller (positive selection) than Nd/N, as indicated by small *P*-value (e.g., <0.05).

系统发育树：可视化演化关系
Visualization of evolutionary relation ships by phylogenetic tree

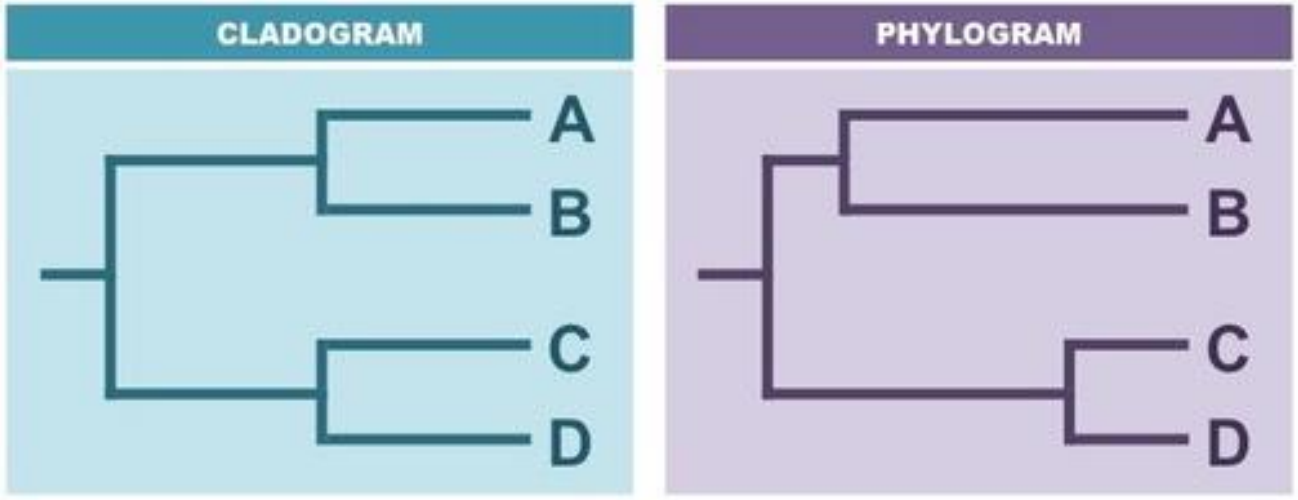


Phylograms vs Cladograms

- Both diagrams show relationships between lineages.
- Phylograms (phylogenetic trees) **show the amount of genetic change** over time calibrated by fossils or a molecular clock.
  - ✓ They **do infer the amount of evolutionary time** separating two species.
  - ✓ The greater the differences between the sequences, the longer the time span since the two species had a common ancestor.
- Cladograms **do not infer** an evolutionary time scale—**the length of each branch is the same**.

Cladogram vs Phylogram

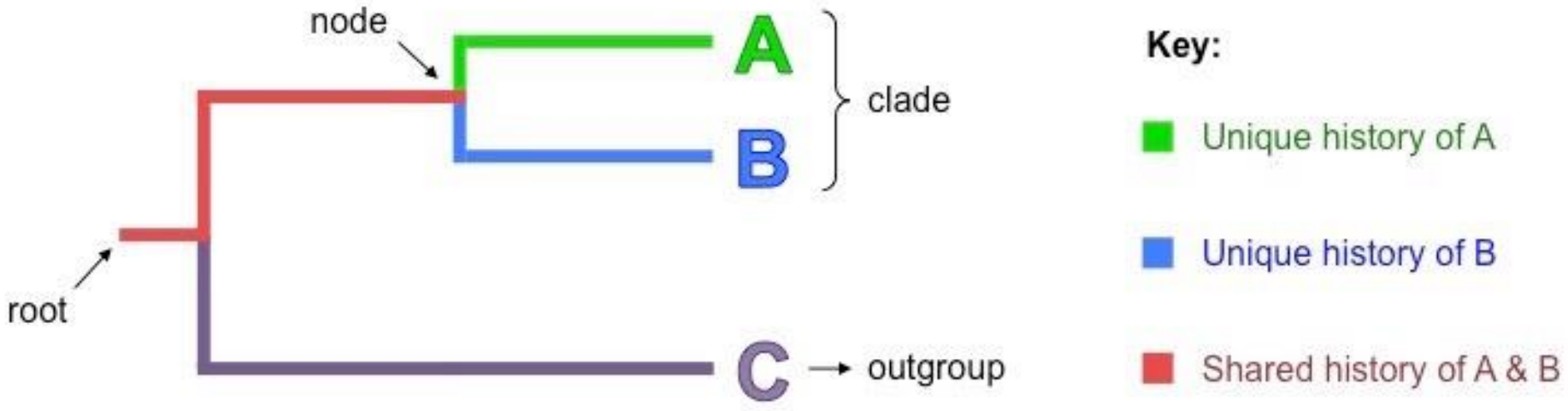
分支图



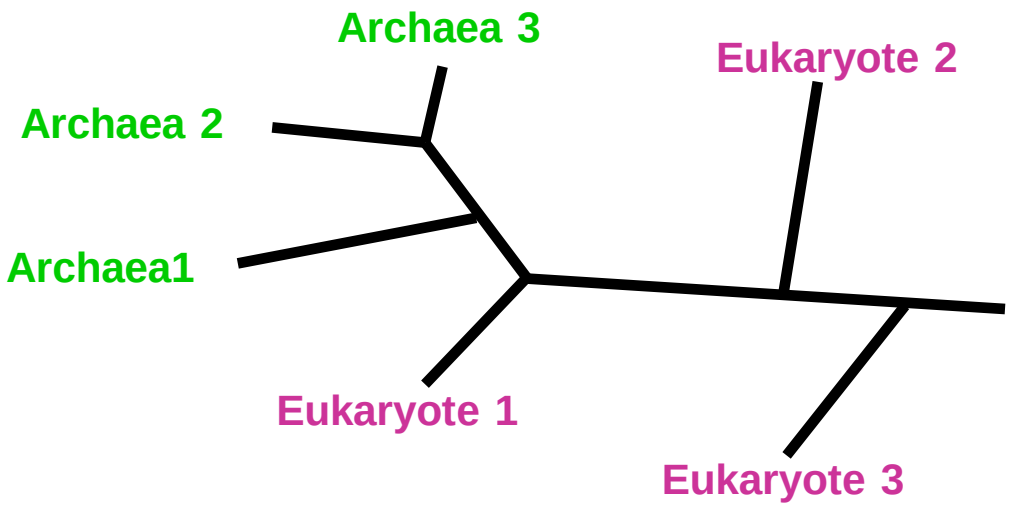
系统发生树

Phylogeny Terminology

- **Root** – The initial ancestor common to all organisms within the cladogram (incoming line shows it originates from a larger clade)
- **Nodes** – Each node corresponds to a hypothetical common ancestor that speciated to give rise to two (or more) daughter taxa
- **Outgroup** – The most distantly related species in the cladogram which functions as a point of comparison and reference group
- **Clades** – A common ancestor and all of its descendants (i.e. a node and all of its connected branches)

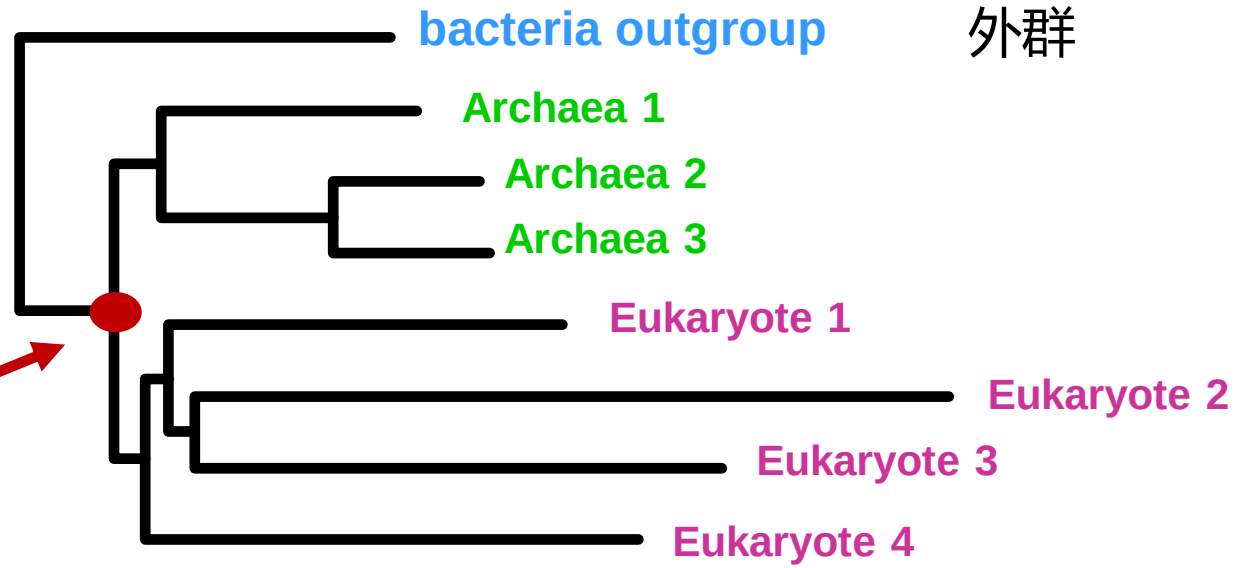


Rooted and Unrooted Trees



Unrooted tree
无根树

Rooted tree
有根树



通过外群
确定树根

Phylogenetic Tree

- 系统发育树：亦称系统发生树、分子演化树、分子进化树
- 依据分子序列分化程度，重构演化关系，预测分子序列功能
- 建树方法：
  - ✓距离法（Distance Matrix）：UPGAM、Neighbor-Joining（邻接法）
  - ✓最大简约法（Maximum Parsimony）：替换数最小的拓扑结构，作为最优树
  - ✓最大似然法（Maximum Likelihood）：需要替换模型，似然率最大的拓扑结构作为最优树

Construct a Tree with UPGMA

"Unweighted Pair Group Method with Arithmetic Mean"

- Originally developed for numeric taxonomy in 1958 by Sokal and Michener
- Simplest algorithm for tree construction, so it's fast!

非加权配对算术平均法

Algorithm procedure:

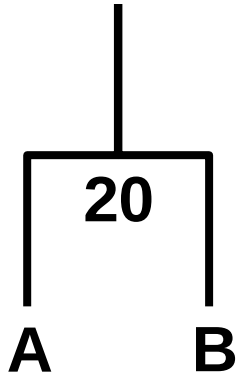
- Derive a distance Matrix from aligned sequences.
- Assume each leaf/sequence as a cluster.
- Repeat steps 1 and 2 until there are only two clusters.

  Step 1: Cluster a pair of leaves (taxa) by shortest distance.

  Step 2: Derive a new distance Matrix by recalculating a new average distance with the new cluster and other leaves.

Example of UPGMA

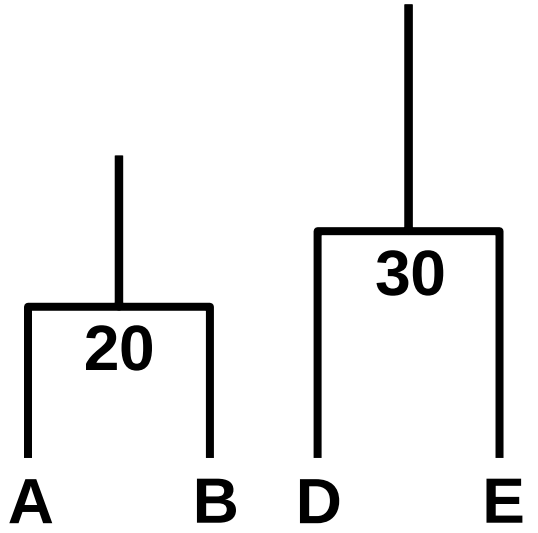
	A	B	C	D	E
A	0				
B	20	0			
C	60	50	0		
D	100	90	40	0	
E	90	80	50	30	0



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

Example of UPGMA (cont.)

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

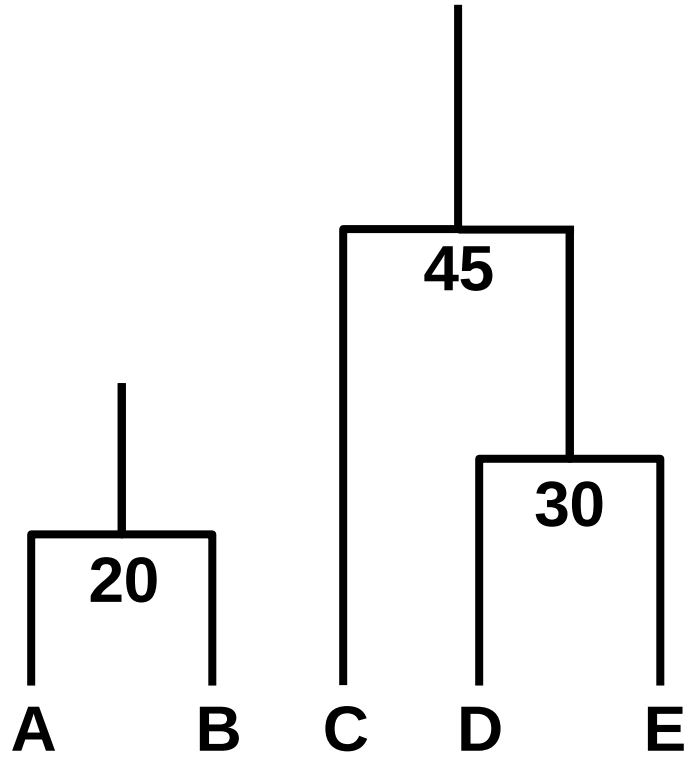


- New average distance between DE and AB is:
  DE to AB = (95 + 85) / 2 = 90
- Distance between DE and C is:
  DE to C = (40 + 50) / 2 = 45

Example of UPGMA (cont.)

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

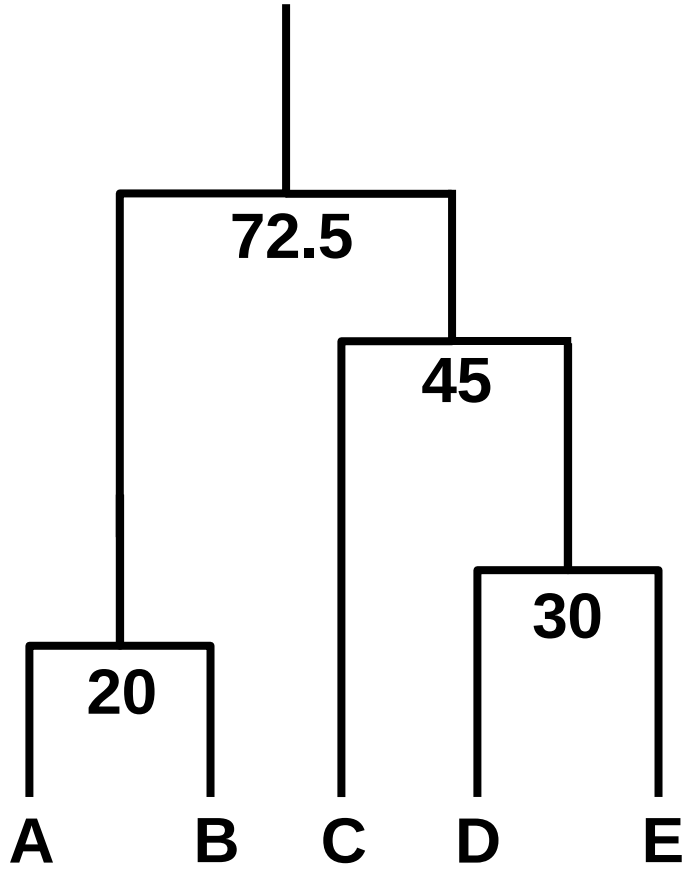
New Average distance between CDE and AB is:
CDE to AB = (90 + 55) / 2 = 72.5



Example of UPGMA (cont.)

	AB	CDE
AB	0	
CDE	72.5	0

Mission accomplished!



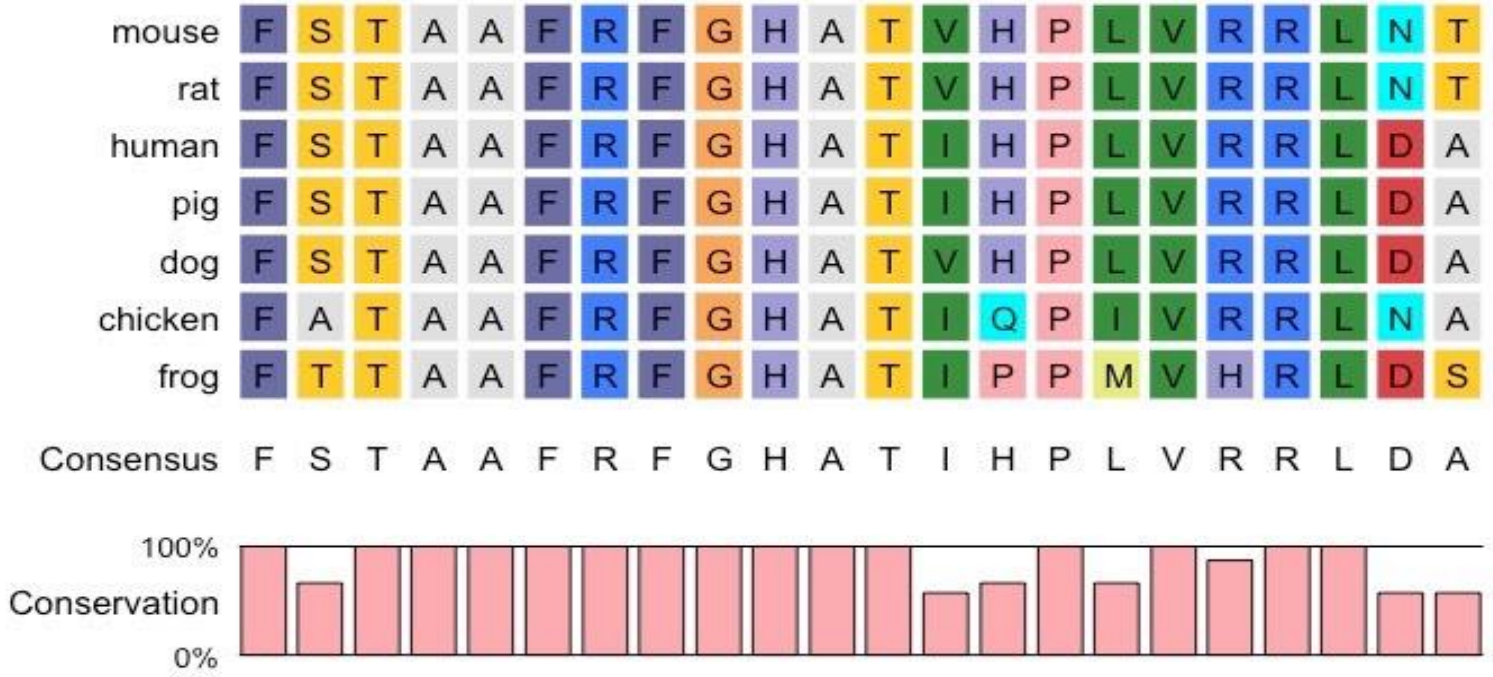
Steps for tree construction

Step 1: Select a gene/protein common to a range of selected organisms

Step 2: Obtain molecular sequences for all selected organisms

Step 3: Run a multiple alignment to compare molecular sequences

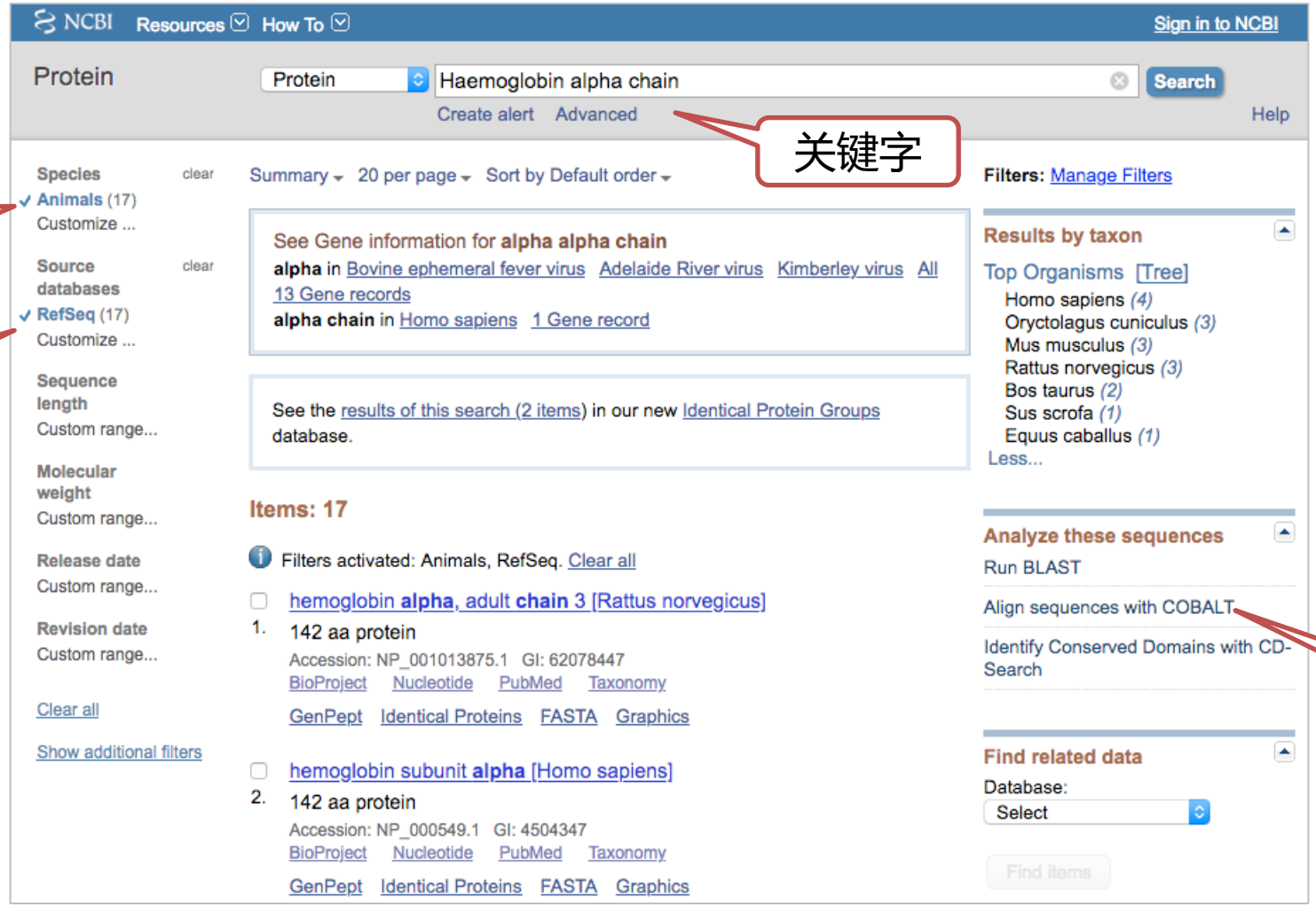
Step 4: Generate a phylogeny tree (cladogram) from aligned sequences



Hemoglobin subunit alpha

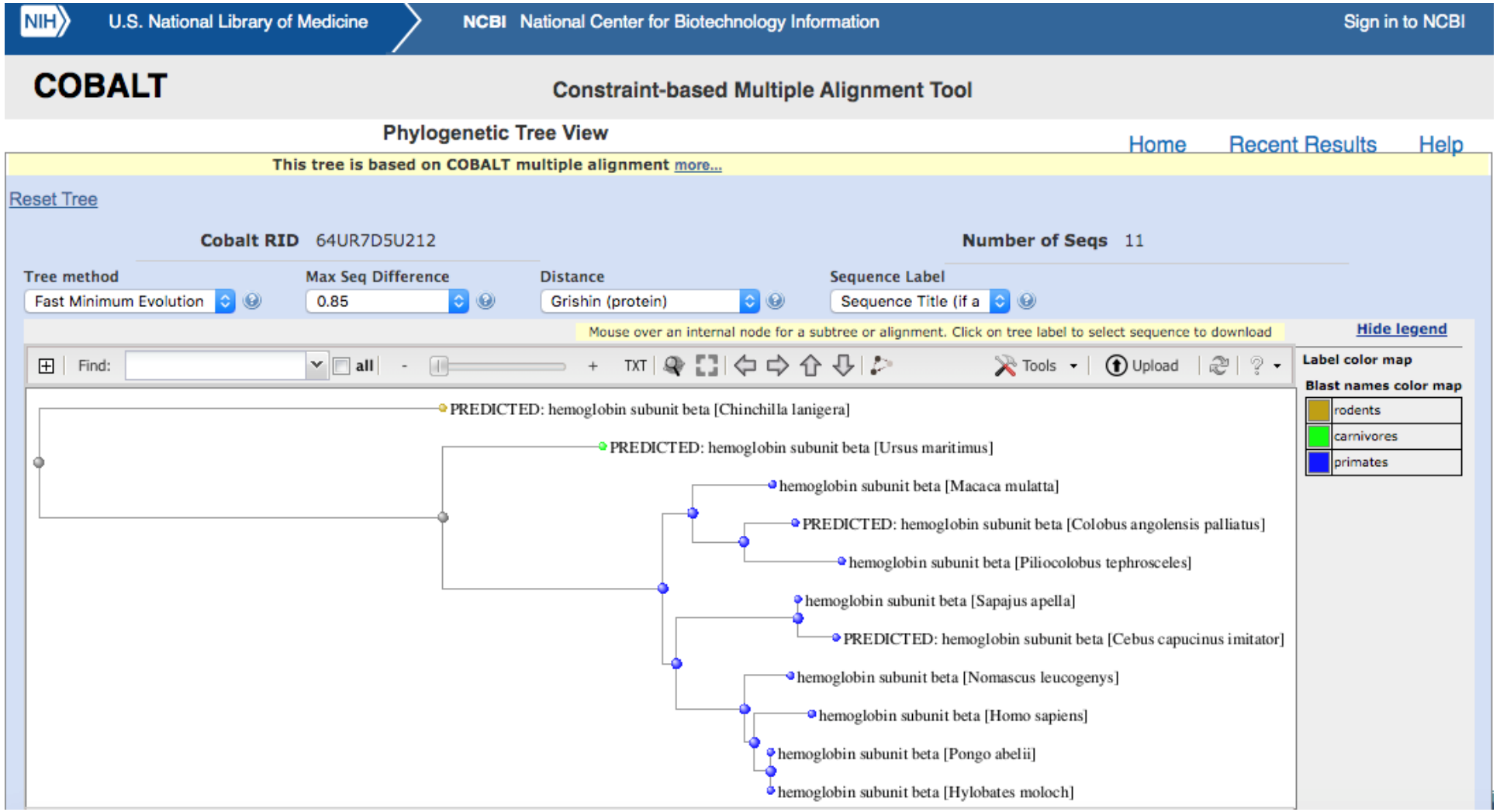
>NP_000549.1 hemoglobin subunit alpha [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>NP_001036091.1 hemoglobin subunit alpha [Pan troglodytes]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>XP_012661133.1 hemoglobin subunit alpha [Otolemur garnettii]
MVLSPTDKSNVKAAWEKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGKKVADALTNAVSHVVNFKLLSHCLLVTLACHHPAELT
PAVHASFDKFMASVSTVLTSKYR
>XP_004628004.1 hemoglobin subunit alpha [Octodon degus]
MVLSPADKTNVKTAWGKIGGHGAEYGAEALFRMFLSFPTTKTYFHHFDLSAGSAQIKSHGKKVSDALTTAVDHLDDLPTALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLSAHHPADFTPAVHASLDKFLATVSTVLTSKYR
>XP_006050021.1 hemoglobin subunit alpha-4 [Bubalus bubalis]
MVLSAADKSNVKAAWGKVGGHAADYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGAKVANALTKAVGHLDDLPGALSELSDLHAHKLRVDPV
NFKLLSHSLLVTLASHLPNDFTPAVHASLDKFLASVSTVLTSKYR
>XP_008155896.1 hemoglobin subunit alpha [Eptesicus fuscus]
MVLSPADKSNVKAAWDKVGGNAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVGDALGSAVAHMDDLPGALSALSDLHAYKLRVDPV
NFKLLSHCLLVTLACHNPAEFTPAVHASLDKFLASVSTVLVSKYR
>XP_010380159.1 hemoglobin subunit alpha [Rhinopithecus roxellana]
MVLSPADKTNVKAAWGKVGGHGGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>XP_022450006.1 hemoglobin subunit alpha [Delphinapterus leucas]
MVLSPADKTNVKGTWAKIGNHSAEYGAEALERMFISFPSTKTYFSHFDLGHGSAQIKGHGKKVADALTKAVGHIDNLPDALSELSDLHAHKLRVDPV
NFKLLSHCLLVTLALHLPADFTPSVHASLDKFLASVSTVLTSKYR

Hemoglobin sequences in NCBI



https://www.ncbi.nlm.nih.gov/protein/?term=Haemoglobin+alpha+chain

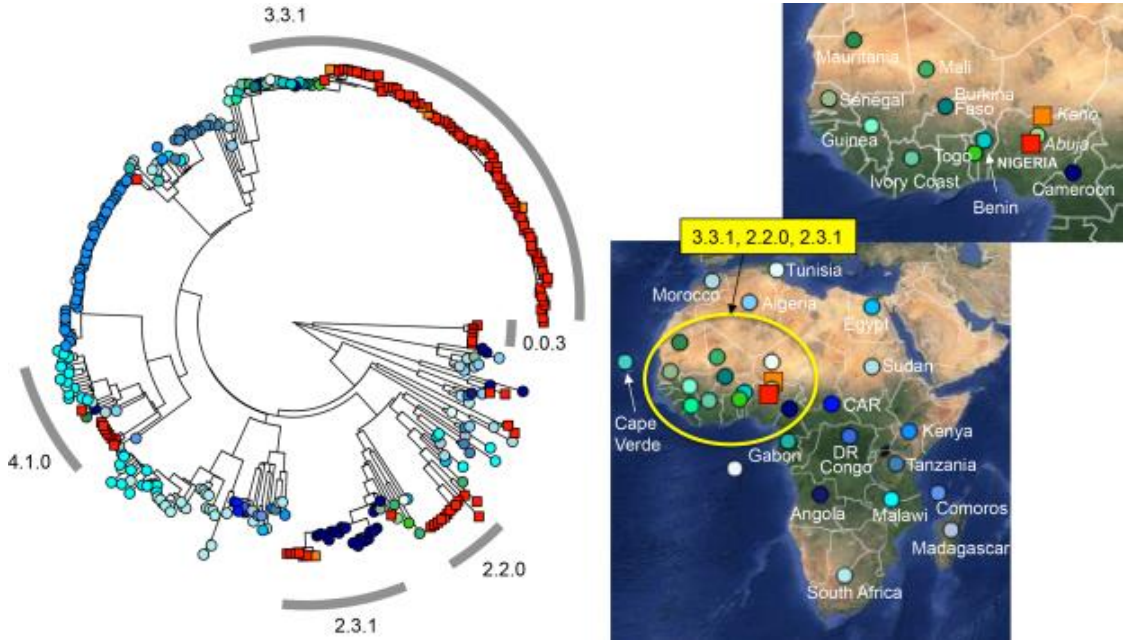
Phylogeny tree construction in NCBI

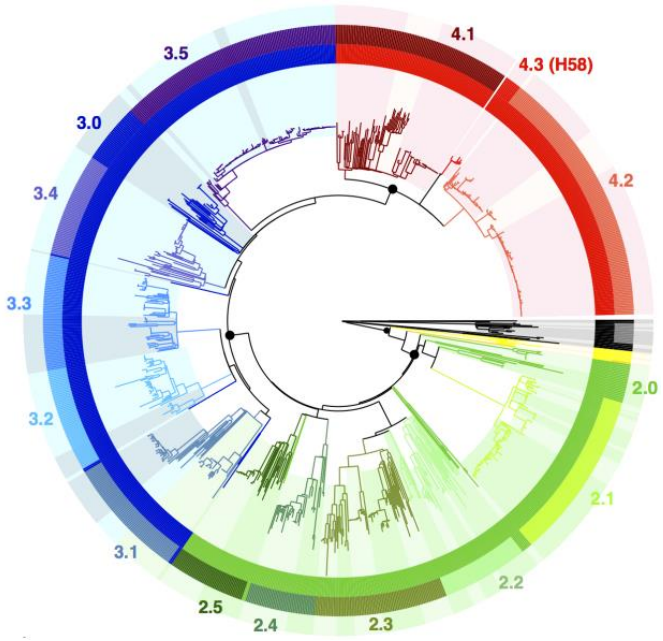


https://www.ncbi.nlm.nih.gov/gene/3043/ortholog

常用软件

- **MEGA** (Molecular Evolutionary Genetics Analysis): www.megasoftware.net
- **PAML** (Phylogenetic Analysis by Maximum Likelihood): http://abacus.gene.ucl.ac.uk/software/paml.html
- **iTOL** (Interactive Tree Of Life): an online tool for the display, annotation and management of phylogenetic trees, https://itol.embl.de
- 392 tools: http://evolution.gs.washington.edu/phylip/software.html





群体基因组学

Population Genomics: Advancing Understanding of Nature

Population genomics is the large-scale comparison of DNA sequences of populations.

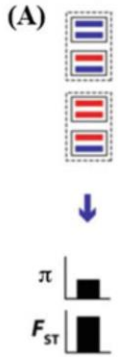
The simultaneous study of **numerous loci** and **genome regions** to better understand the roles of **evolutionary processes** (such as mutation, genetic drift, gene flow, and natural selection) that influence variation across **genomes and populations**.

传统群体遗传学

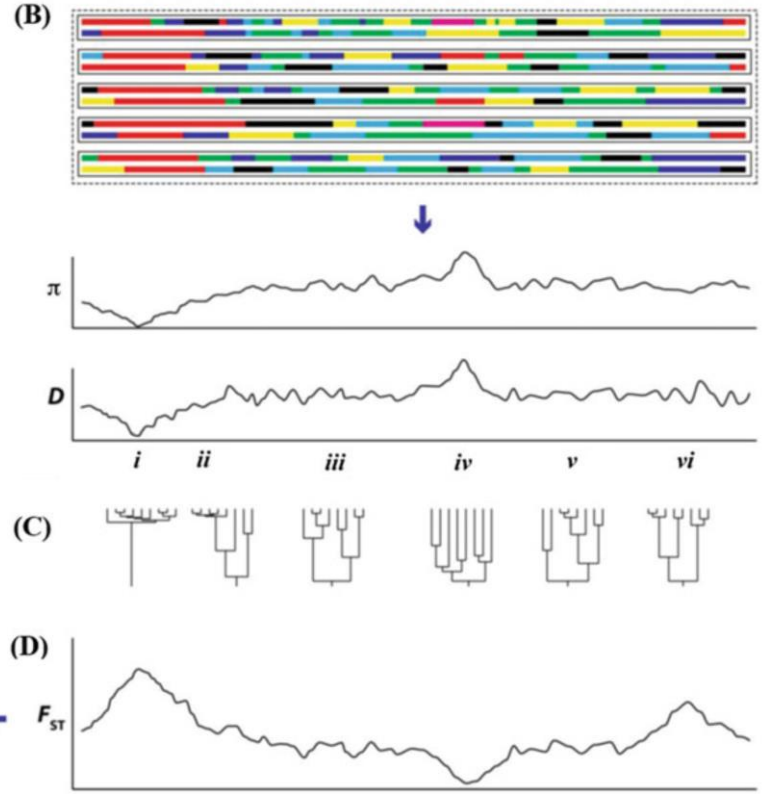
等位基因

核苷酸多样性（π）

种群分化（F_{st}）



群体基因组学



单倍型

等位基因频谱（Tajima's D）

不同进化过程

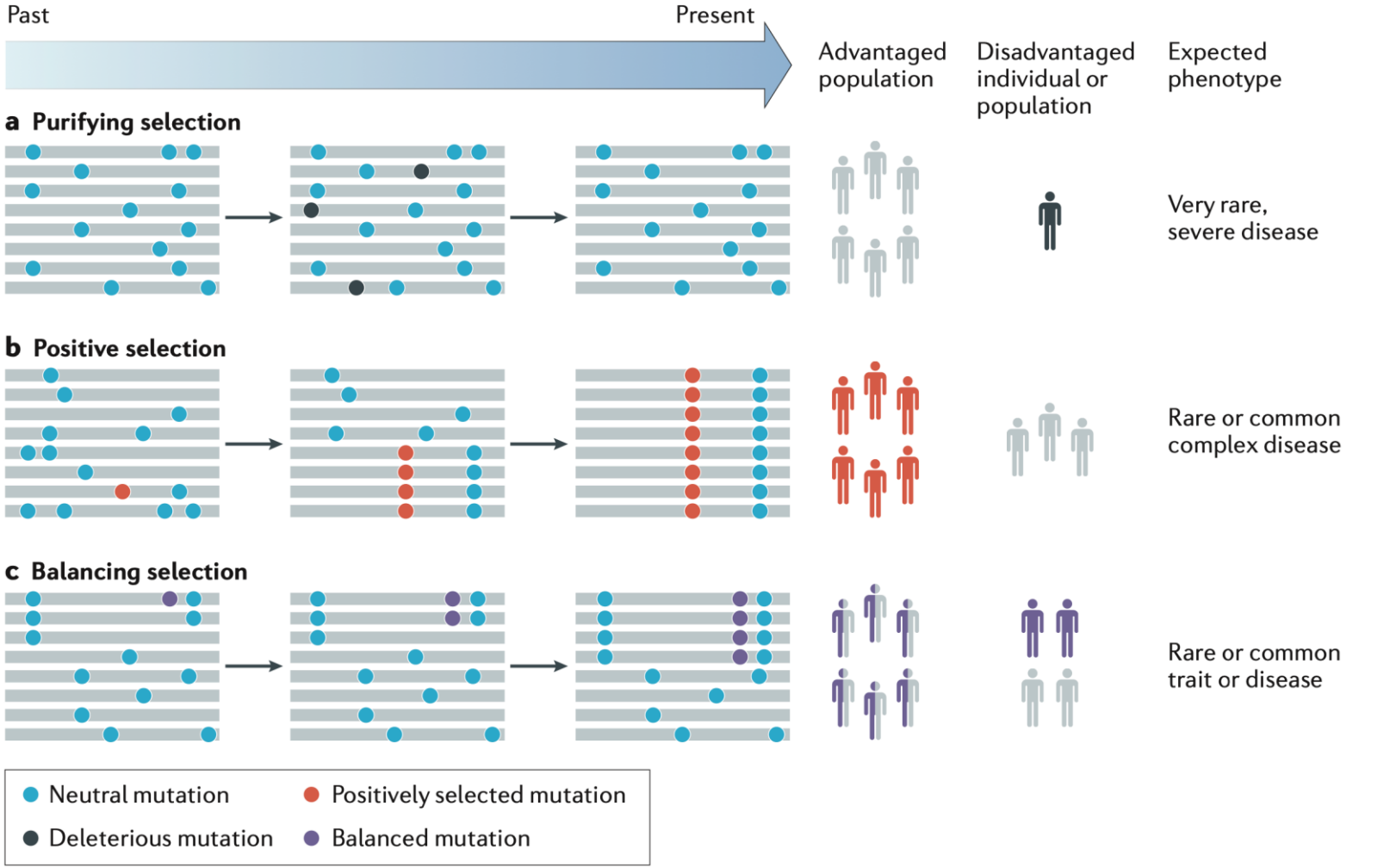
跨种群统计量（如F_{st}）比较

Population genomics is the large-scale application of genomic technologies to study populations of individuals. For example, population genomics research can be used to study human ancestry, migrations and health.

识别统计显著的异常值

Population Genomics: Advancing Understanding of Nature. Population Genomics. Springer 2018

群体基因组学中自然选择的分类



有害突变—清除

有利突变—选择—固定位点遗传多态性降低

平衡选择—多个等位基因被保留（如：杂合优势）

Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol* 2013 https://doi.org/10.1038/nri3421

群体内选择检验Tajima's D Test

鉴定目标DNA序列在进化过程中是否遵循中性进化模型。Tajima's D是由日本研究员田田文雄（Fumio Tajima）创建的群体基因检验统计数据。D的计算方法是两种遗传多样性测量值之间的差异：成对差异的平均数量和分离位点的数量，每个都按比例调整，以便在中等规模的恒定大小的群体中预期它们是相同的.

$$\theta_\pi = \frac{\sum_{i<j} k_{ij}}{\binom{n}{2}} \qquad \theta_w = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

群体多态性参数θ，在二倍体中定义为$\theta = 4N_e\mu$，N_e是有效群体大小，μ是突变率（通常未知）。常用分离位点数θ_W和两两序列核苷酸差异平均数θ_π来估计θ， n是序列数，S为分离位点数，k_{ij} 是序列i 和 j之间的差异位点数。

Tajima's D Test 是比较θ的两个估计值的差异

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{Var\,(\theta_\pi - \theta_W)}}$$

$D \begin{cases} = 0 & \text{符合中性进化} \\ < 0 & \text{存在自然选择} \end{cases}$

Value	interpretation 1	interpretation 2
D=0	Observed similar to expected	No evidence of selection
D<0	Rare alleles abundant	Recent selective sweep, population expansion after a recent bottleneck, linkage to a swept gene
D>0	Rare alleles scarce	Balancing selection, sudden population contraction

```
Position   12345 67890 12345 67890
Person Y   00000 00000 00000 00000
Person A   00100 00000 00100 00010
Person B   00000 00000 00100 00010
Person C   00000 01000 00000 00010
Person D   00000 01000 00100 00010
```

Y vs A: 3　Y vs B: 2　Y vs C: 2　Y vs D: 3　A vs B: 1
A vs C: 3　A vs D: 2　B vs C: 2　B vs D: 1　C vs D: 1

$$\theta_\pi = \frac{\sum_{i<j} k_{ij}}{\binom{n}{2}} = \frac{20}{10} = 2$$

$$\theta_w = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} = \frac{4}{1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = 1.92$$

$$d = 0.08$$

物种内/间选择检测 McDonald-Kreitman (MK) test

	Fixed	Polymorphic
Nonsynonymous	D_N	P_N
Synonymous	D_S	P_S

D_N : the number of nonsynonymous fixed differences
D_S : the number of synonymous fixed differences
P_N : the number of nonsynonymous polymorphisms differences
P_S : the number of synonymous polymorphisms differences

The neutrality index (NI) quantifies the direction and degree of departure from neutrality.

$$\mathrm{NI} = \frac{P_N/PS}{D_N/DS}$$

NI=1: neutral mutation
NI>1: negative selection
NI<1: positive selection

	Fixed	Polymorphic
Nonsynonymous	7	2
Synonymous	17	42

$$\frac{P_N}{P_S} = \frac{2}{42}, \frac{D_N}{D_S} = \frac{7}{17}$$

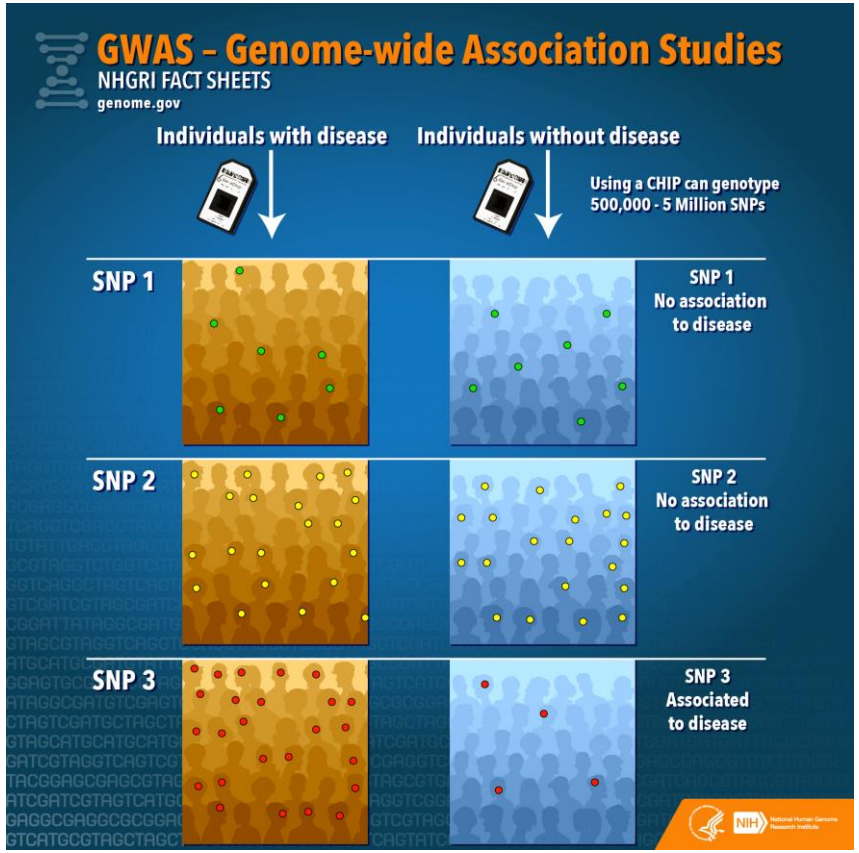
NI = (2/42) / (7/17) < 1

Fisher's exact test: $p = 0.007$

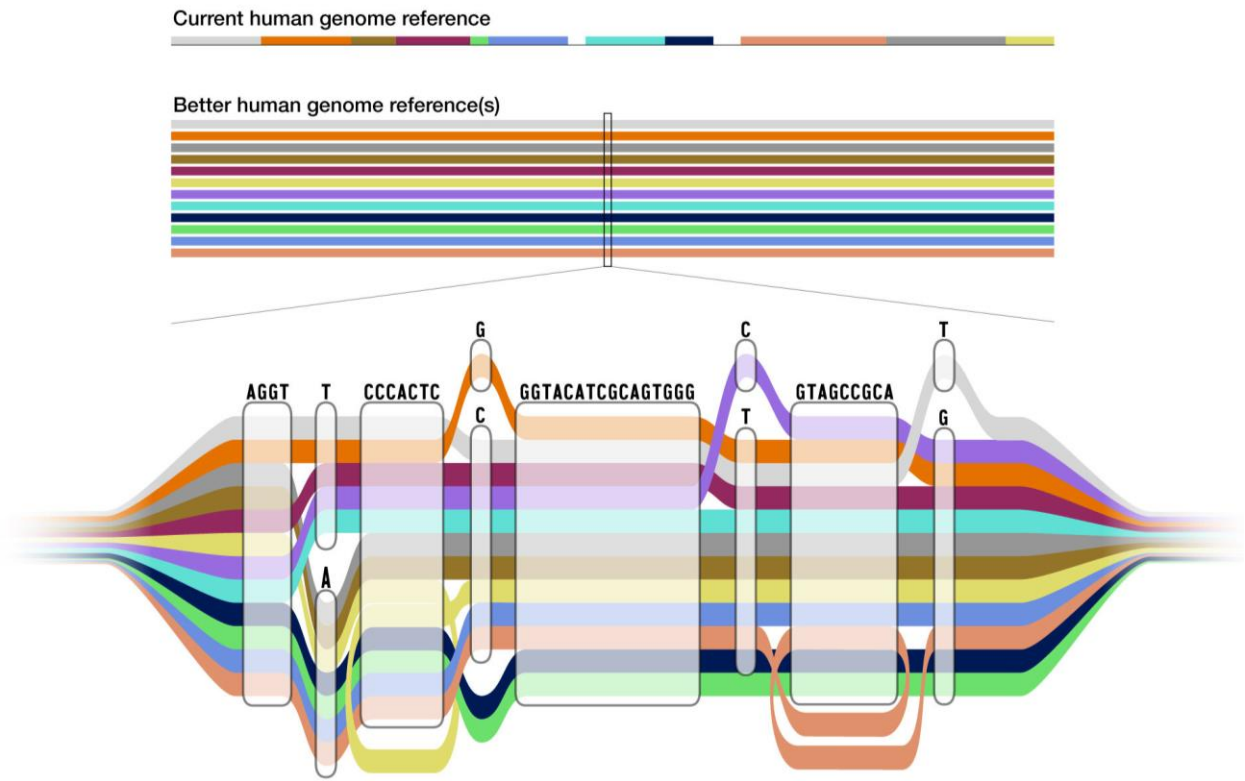
McDonald, J. H., and Kreitman, M. (1991). **Adaptive protein evolution at the Adh locus in *Drosophila***. *Nature* 351, 652–654.

群体基因组：从GWAS到Pangenome

Genome-Wide Association Study



Pangenome Graph



Summary

- 比较基因组学与分子演化
- 分子钟假说
  - 线粒体夏娃（母系遗传）、Y染色体亚当（父系遗传）
  - 非洲起源说与人类迁徙
- 拉马克进化理论、达尔文自然选择理论与中性突变理论
  - 突变：有益、中性、有害
  - 作用力：选择（Selection）、突变（Mutation）、遗传漂变
- 分子序列演化模型：JC模型，计算分化距离
- 遗传密码子表，密码子使用偏好，选择与突变共同作用、计算方法
- 非同义和同义替换率，选择压力Ka/Ks（正选择、中性突变、负选择）
- 系统发育树：根的类型、建树主要方法与流程
- 群体基因组：自然选择检测方法(Tajima'D、MK)、GWAS、Pangenome

Further Reading

- **Population Genomics: Advancing Understanding of Nature**. Population Genomics. Springer 2018 https://doi.org/10.1007/13836_2018_60
- **Adaptive Molecular Evolution.** *Handbook of Statistical Genomics* 2019, https://doi.org/10.1002/9781119487845.ch13
- **The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process.** *Annual Review of Genetics* 2016
- **Molecular phylogenetics: principles and practice.** *Nature Reviews Genetics* 2012
- **Synonymous but not the same: the causes and consequences of codon bias**. *Nature Reviews Genetics* 2011
- **Selection on codon bias**. *Annu Rev Genet* 2008
- **The neutral theory of molecular evolution in the genomic era**. *Annu Rev Genom Hum Genet* 2010
- **Selectionism and Neutralism in Molecular Evolution**. *Molecular Biology and Evolution* 2006
- **The Ka/Ks ratio: diagnosing the form of sequence evolution**. *Trends in Genetics* 2002
- **Molecular Evolution and Phylogenetics**. Oxford University Press 2000