大數據分析與人類疾病

Big data analysis and human diseases

Trees-Juen Chuang (莊樹諄)

中央研研院 基因體研究中心 研究員/教授
台灣大學 基因體與系統生物學學程 核心教授

http://idv.sinica.edu.tw/trees/
Email: trees@gate.sinica.edu.tw

Academia Sinica
Genomics Research Center

最低的水果摘完之後

—顏擇雅

- **Comparative & Evolutionary Genomics (dry)**
- **Causative Biology (dry/wet)**
- **Systems Biology (dry/wet)**
- **Neuropsychiatric Disease (wet)**
- **Machine Learning & Omics Data Analysis (dry)**

- 比較與演化基因體學（資訊統計）
- 調控因果生物學（資訊統計/分生實驗）
- 系統生物學（資訊統計/分生實驗）
- 神經精神疾病（分生實驗）
- 機器學習與多體學資料分析（資訊統計）

基因、演化及大數據分析

Trees-Juen Chuang (莊樹諄)

http://idv.sinica.edu.tw/trees/

Email: trees@gate.sinica.edu.tw

Academia Sinica

Genomics Research Center

猩球崛起　劇情簡介

美國科學家威爾・羅德曼利用黑猩猩進行實驗，研究藥物治療阿茲海默症。猩猩主角凱撒因此智力大增，牠眼見人類對待猩猩的種種不人道，憤而領導猩猩族群，反抗人類世界，希望解救所有受到不人道待遇的猩猩們，並帶領大家重返真正的家—森林。

http://zh.wikipedia.org/wiki/%E7%8C%BF%E4%BA%BA%E7%88%AD%E9%9C%B8%E6%88%B0%EF%BC%9A%E7%8C%A9%E5%87%B6%E9%9D%A9%E5%91%BD#.E6.95.85.E4.BA.8B.E7.B0.A1.E4.BB.8B

<學院報告> (短篇小說)

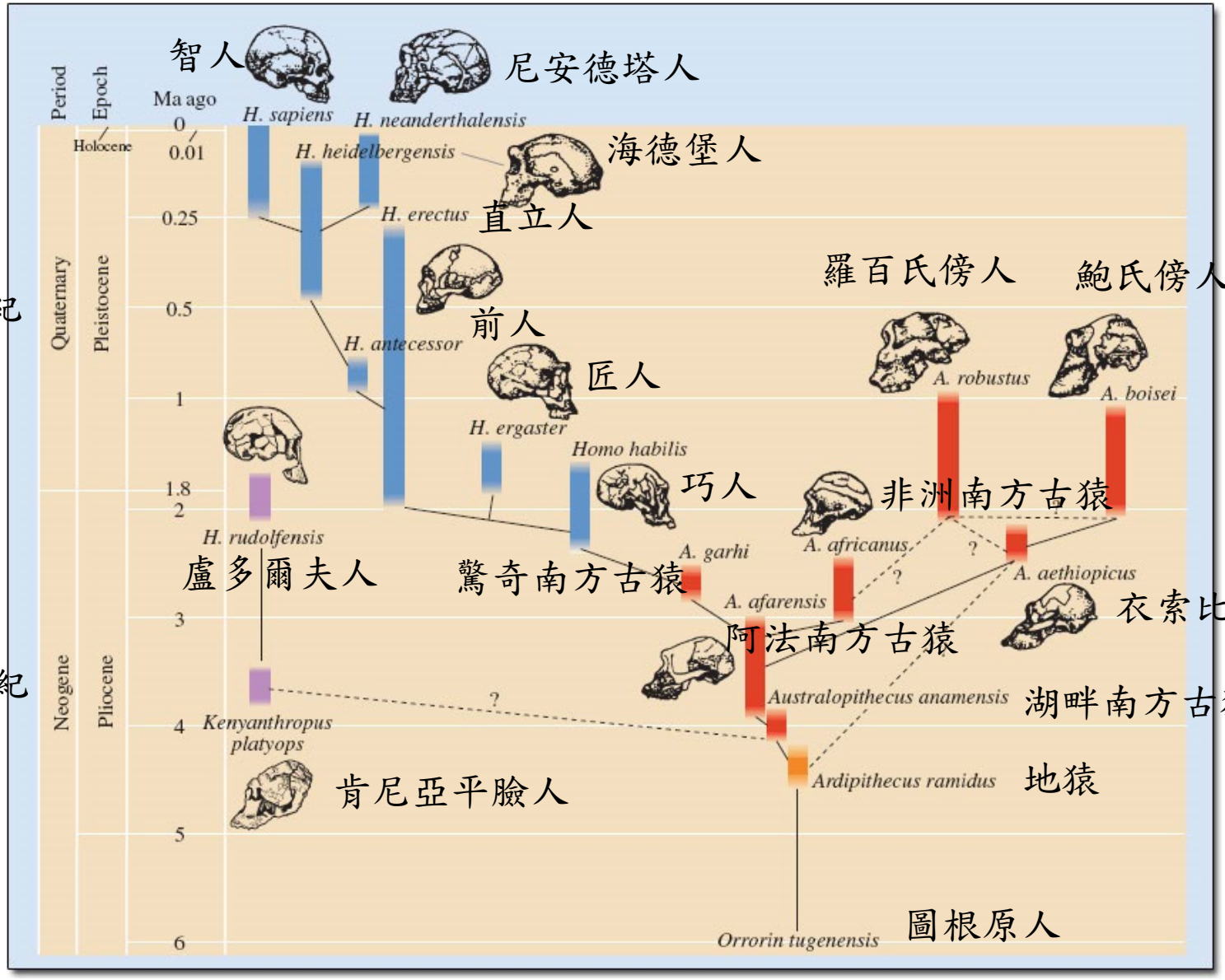
法蘭茲．卡夫卡

卡夫卡以黑猩猩的觀點來看自由與身為人類的意義。

內容描述一隻猩猩在非洲被人類抓走並關在籠內運往歐洲去，這隻猩猩在籠內便開始思考，他如果要活下，他就必須找尋“出路”。

而他的"出路"就是必須向籠子外面的那些“人”一樣，所以他開始學習“人類”的行為、語言，甚至達到一般歐洲人的教育水準，並受邀到學院裡作一場演講。

http://goods.ruten.com.tw/item/show?213021173077 26
http://www.wretch.cc/blog/teruhydehele/11062302

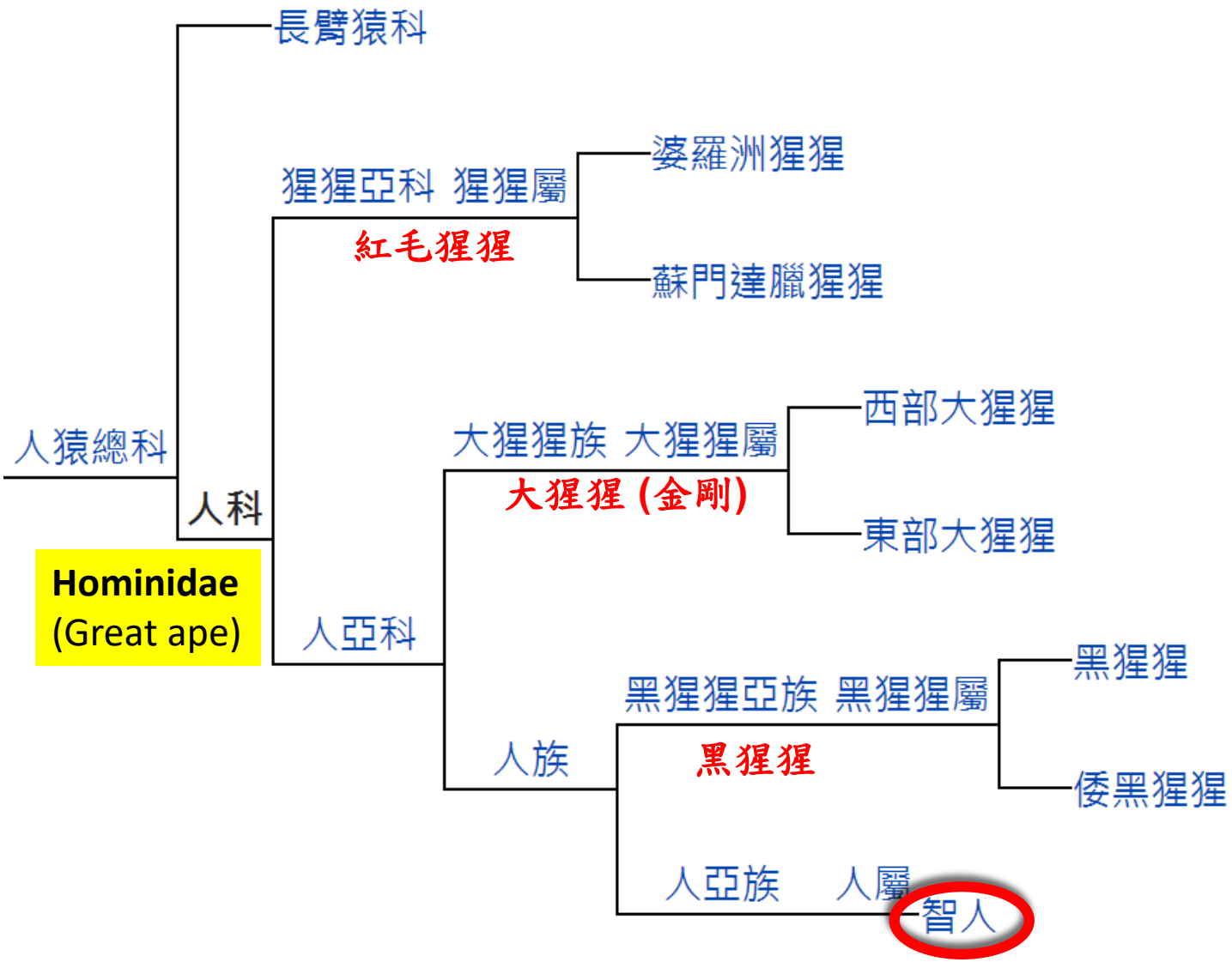


第四紀

晚第三紀

(http://davidpratt.info/ape1.htm)

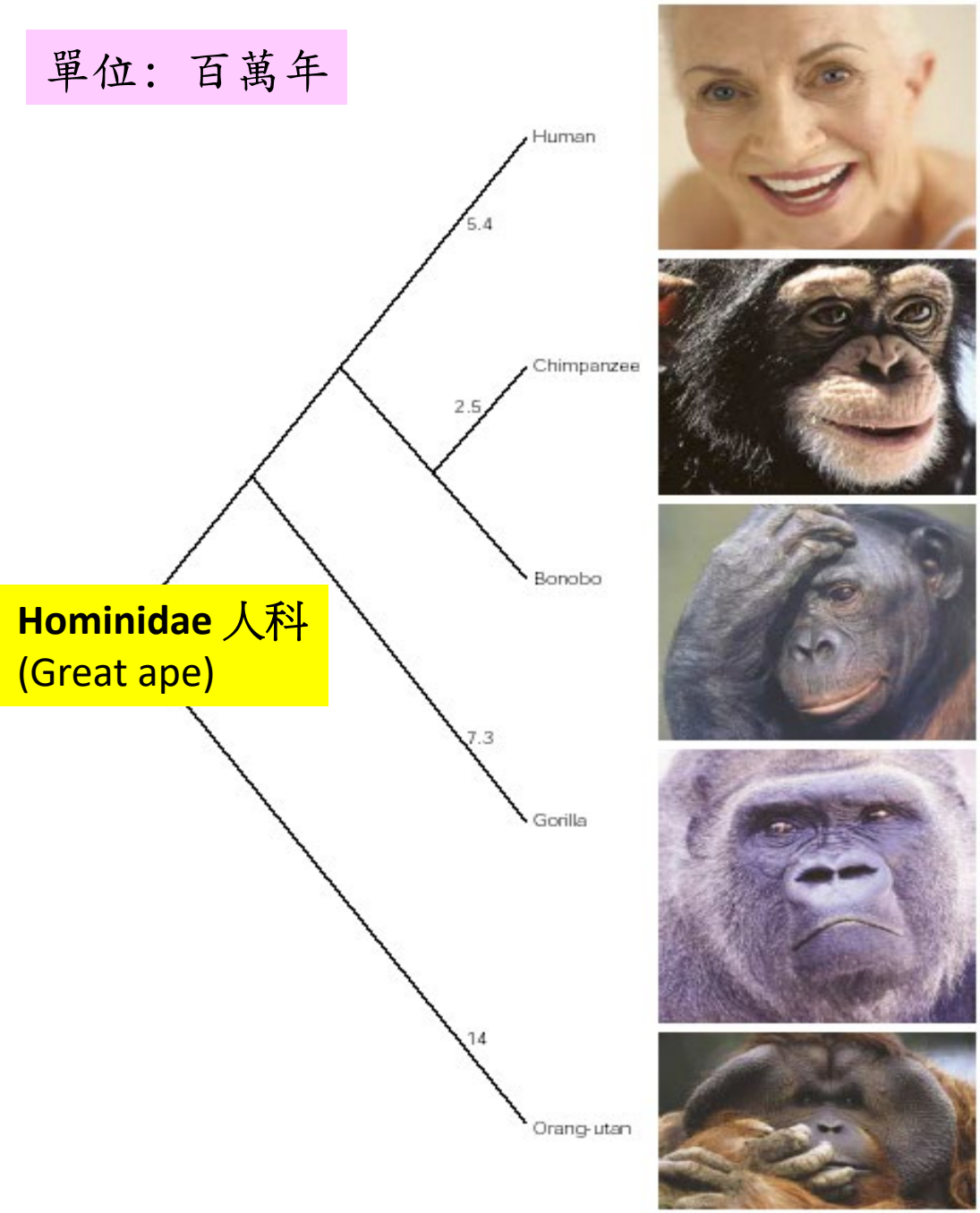
界、門、綱、目、科、屬、種



國際自然保護聯盟分類中，智人是人科物種中唯一無危物種。

維基百科

單位：百萬年



Hominidae 人科
(Great ape)

黑猩猩
(Chimpanzee)

侏儒黑猩猩
(Bonobo)

非洲

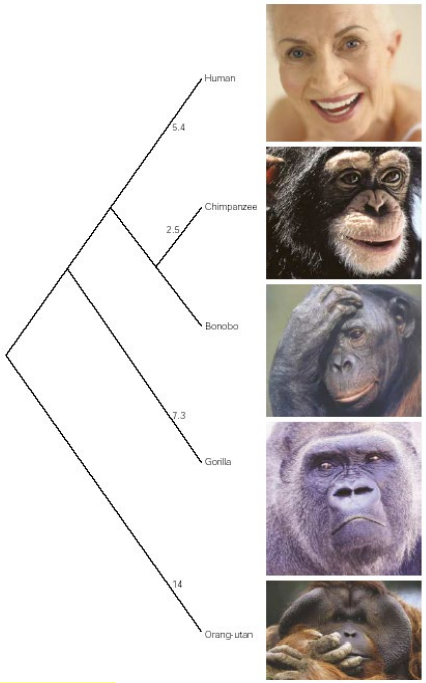
大猩猩
(Gorilla)

紅毛猩猩　婆羅洲跟蘇門達臘
(Orangutan)

Common ancestor

共同祖先

最近的共同祖先



人跟黑猩猩的基因非常相像

➔有人說人是「第三種黑猩猩」。

(The Third Chimpanzee--Jared Diamond)

醫藥上的應用:研究為何有些疾病，在人身上會致命，在黑猩猩身上卻不發病或僅產生輕微病症，如愛滋病、阿茲海莫症、B/C型肝炎等。

- 人有23對染色體: 1, 2, …, 22, X, Y
- 黑猩猩有24對染色體: 1, 2, …, 23, X, Y

- 人的第2號染色體相當於黑猩猩的第12和13號染色體
- 人和黑猩猩的基因體大小差不多

- 人和黑猩猩在DNA上的相似度超過98% (1.23%的差異)

- 幾乎所有的人的基因和黑猩猩的基因皆具很高的相似度

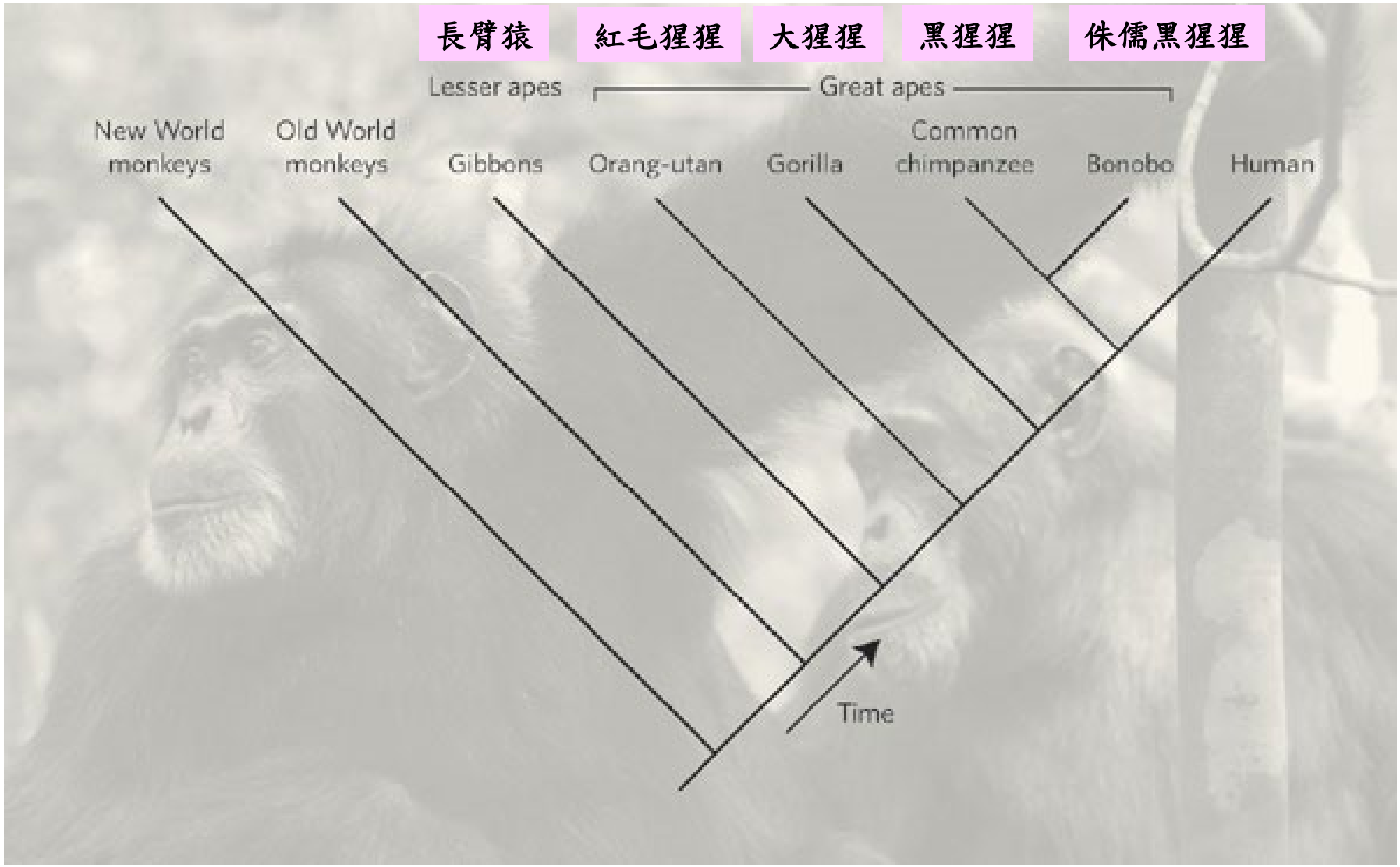
人 vs. 老鼠 ➔ 88%

人 vs. 雞 ➔ 60%

靈長目

Nature 2005, 437: 50-51

新世界猴 舊世界猴 小猿 大猿

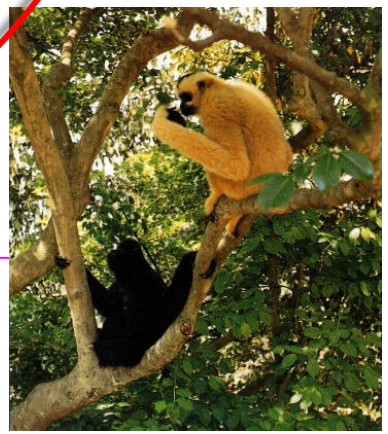


界、門、綱、目、科、屬、種

Hominidae 人科
(Great ape)

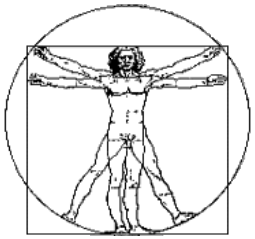
靈長目

- 原猴科（樹鼩ㄑㄩˊ、狐猴、眼鏡猴）
- 猴科
  - 新大陸猴（蜘蛛猴）
  - 舊大陸猴（獼猴、狒狒）
- 猿科
  - 小猿（長臂猿）
  - 大猿
- 人科



單位：百萬年

Human

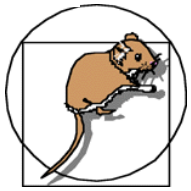


75 Myr

Rat

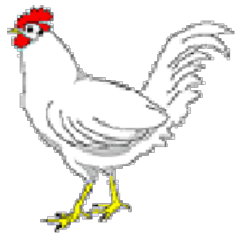
12-24 Myr

Mouse



310 Myr

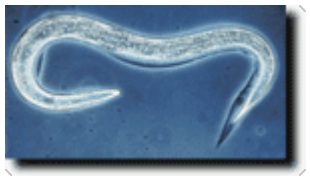
Chicken



1,000 Myr

Invertebrate

無脊椎動物



基因體序列 (DNA 序列):

aagtacgatatgccgagtcccatatgcgtagtcatgagc

由四種核苷酸 (nucleotide acid) 排列組合而成

A, G, C, T

A（adenine，腺嘌呤）
G（guanine，鳥嘌呤）

T（thymine，胸腺嘧啶）
C（cytosine，胞嘧啶）

人類的DNA序列由30億個A、G、T、C核苷酸排列組合而成

A（adenine，腺嘌呤）　T（thymine，胸腺嘧啶）
G（guanine，鳥嘌呤）　C（cytosine，胞嘧啶）

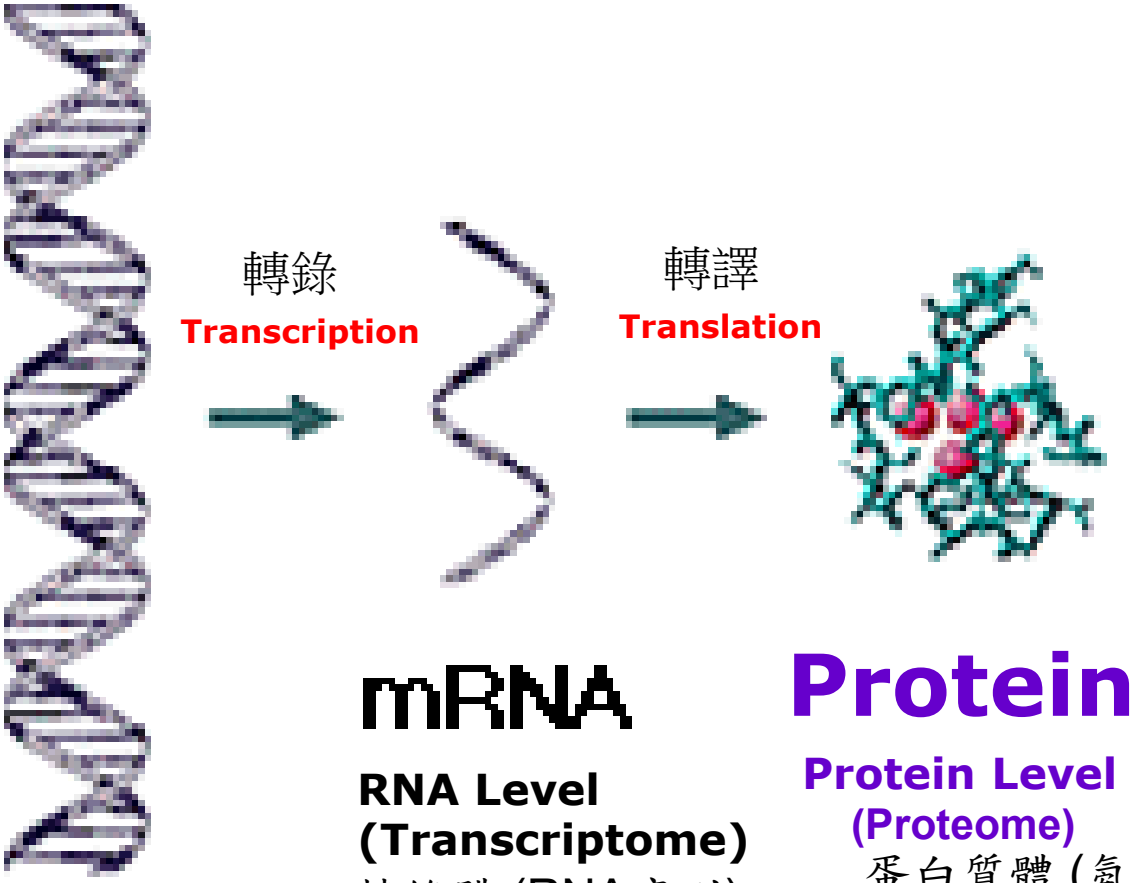
DNA ➔ nucleotide acid (核苷酸)
- ① Phosphoric acid(磷酸)
- ② Deoxyribose (去氧核糖)
- ③ Nitrogenous base (含氮鹽基)

Nitrogenous base (含氮鹽基)
- ① Purines：
  - Adenine (A, 腺嘌呤)
  - Guanine (G, 鳥糞嘌呤)
- ② Pyrimidine：
  - Cytosine (C, 胞嘧啶)
  - Thymine (T, 胸腺嘧啶)

- **DNA sequence: A, C, G, T --- 4 letters**
- **RNA sequence: A, C, G, U (Uracil, (U), 尿嘧啶) --- 4 letters**

分子生物學的中心法則 (central dogma)

克拉克; 1958



Chromosome

DNA Level
(genome)
基因體 (DNA序列)

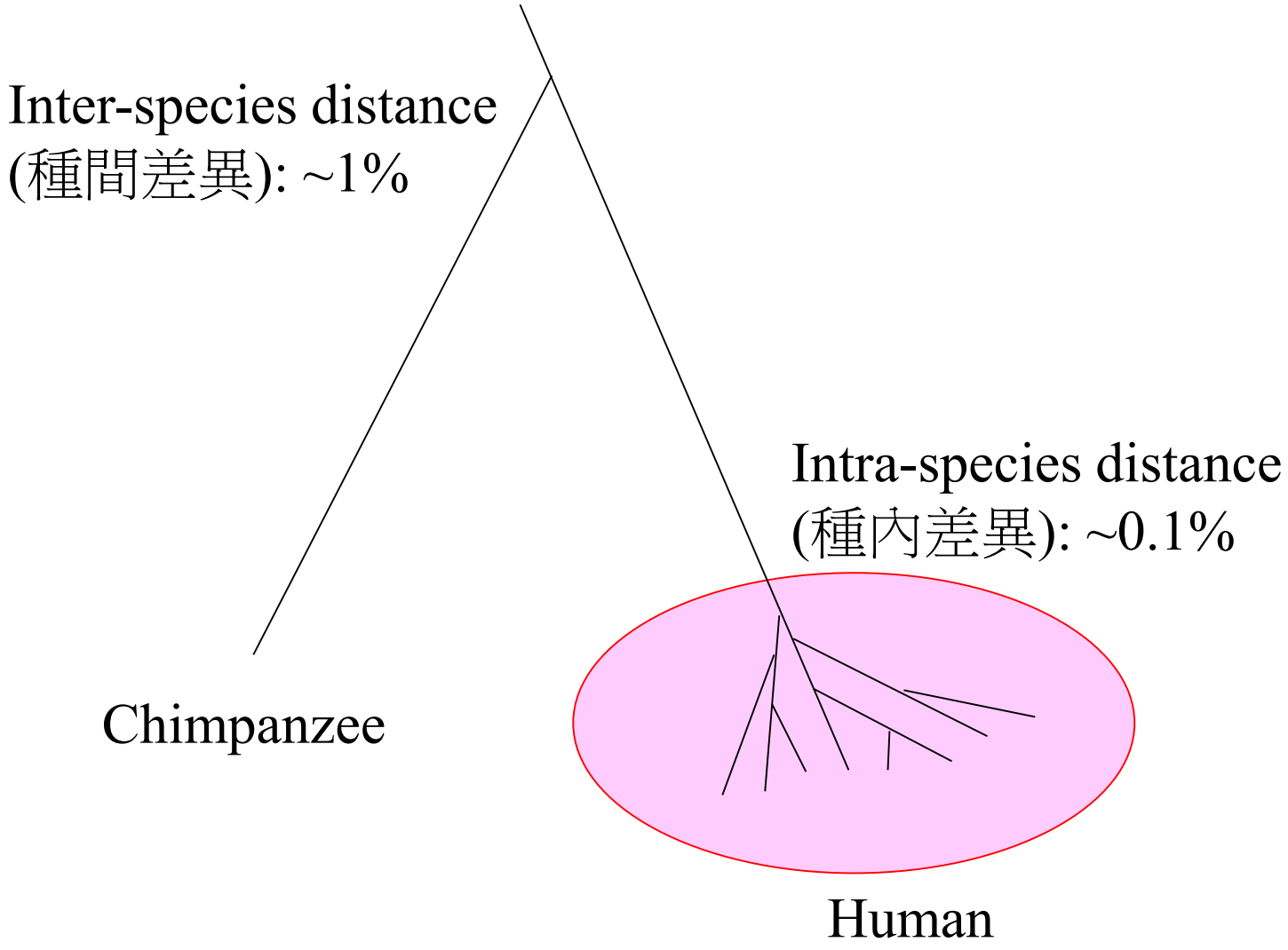
mRNA

RNA Level
(Transcriptome)
轉錄體 (RNA序列)

Protein

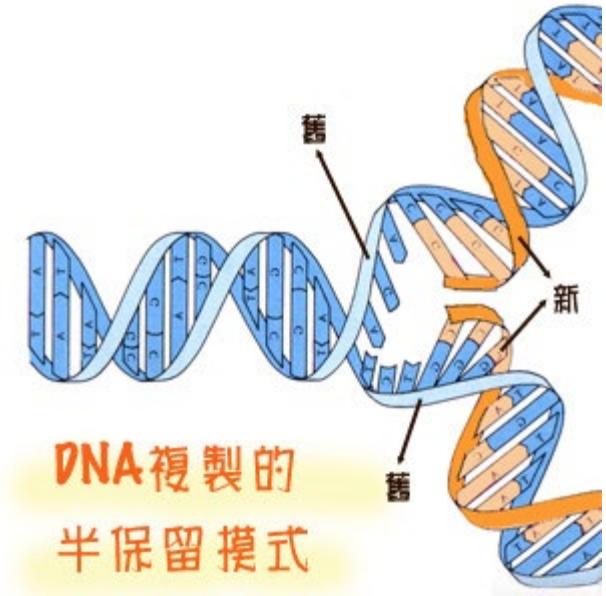
Protein Level
(Proteome)
蛋白質體 (氨基酸序列)

Function

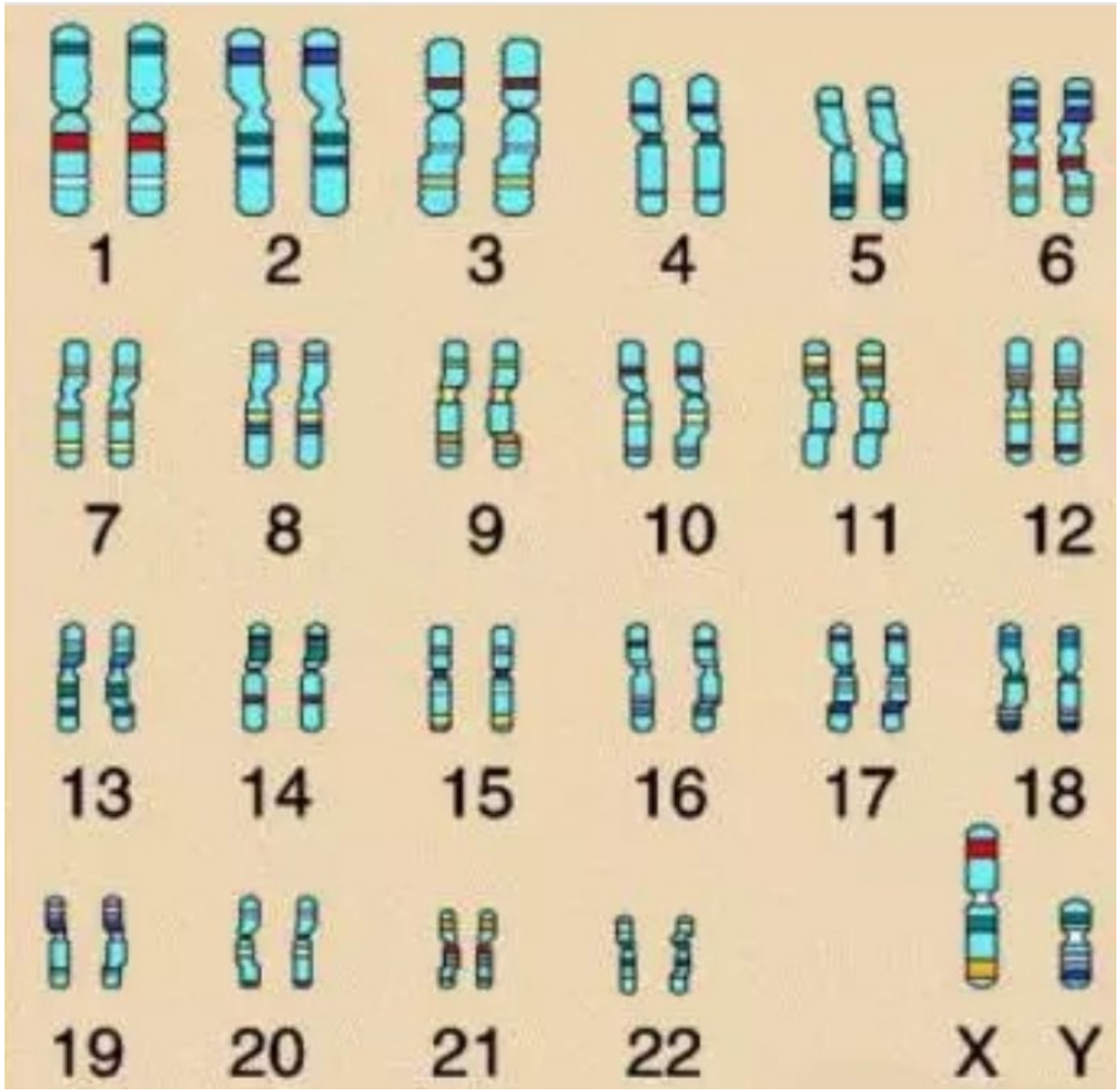

Inter-species distance
(種間差異): ~1%
Intra-species distance
(種內差異): ~0.1%
Chimpanzee
Human

DNA的雙螺旋模型

1953年，華生與克拉克發表了DNA的雙螺旋模型，它的結構以及所揭發的生命密碼系統，堪稱上一個世紀生物學最重大的發現。



1955年12月:**印尼裔美國籍遺傳學家 Joe Hin Tjio,** 人有23對染色體 (chromosomes)
➔ 所有 DNA 序列的總稱叫做: 基因體 (genome)



科學家經過50年的努力，在本世紀的初期，也完成了人類基因組的定序計畫，寫出了組成23對染色體的30億個字母A、G、C、T的序列。

人類的DNA序列由30億 (3×10^9) 個A、G、C、T核苷酸排列組合而成

一對染色體片段序列:


5’
3’
aagtacgatatgccga
ttcatgctatacggct
3’
5’


5’
3’
aagtacgatatgccga
ttcatgctatacggct
3’
5’

演化的幾個力量

- **Mutation (突變)**
- **Selection (選擇)**
- **Genetic drift (基因漂變)**
- **Gene flow (基因流動)**

突變是隨機發生的。

選擇則讓有利於生存與繁殖的遺傳性狀在族群間擴大，有害的性狀在族群間漸漸減少甚至消失。

天擇 (natural selection)

- 負向選擇 (negative selection)
- 正向選擇 (positive selection)
- 中性選擇 (neutral selection)
- 平衡選擇 (balancing selection):血紅蛋白基因

人擇 (artificial selection)

天擇

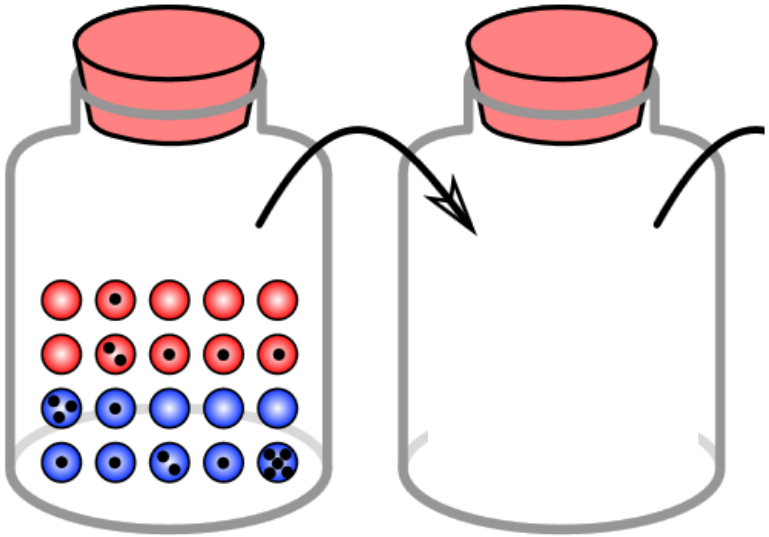
平衡選擇

鐮刀型細胞貧血 vs. 瘧原蟲侵襲

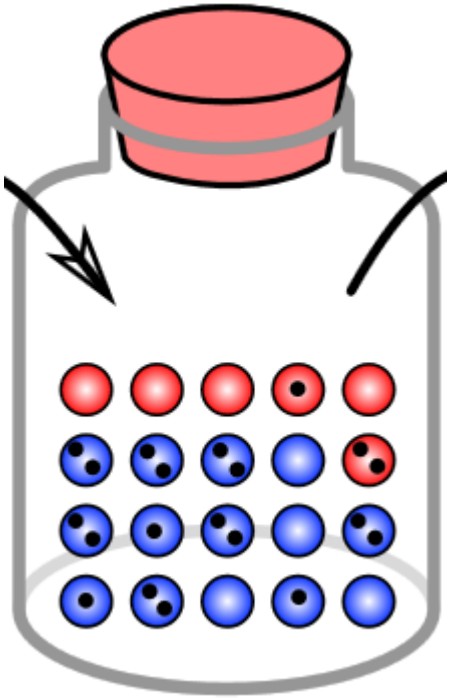
血紅素: 紅血球中負責攜帶氧氣的蛋白質。
鐮型血球貧血症患者: 轉譯成血紅素的基因(HbB)發生點突變。
➔紅血球的攜氧量降低
➔嚴重貧血

非洲瘧疾流行的區域，瘧原蟲無法在鐮刀形紅血球內成長。

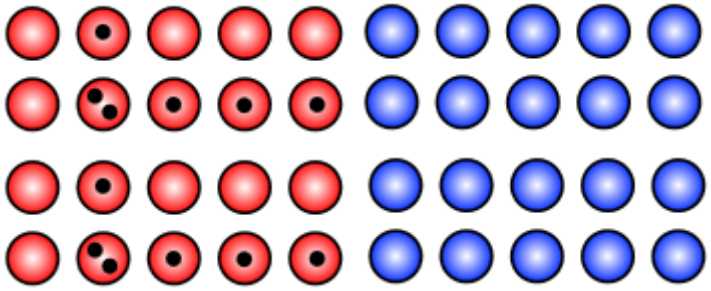
•Genetic drift (基因漂變)



第一代

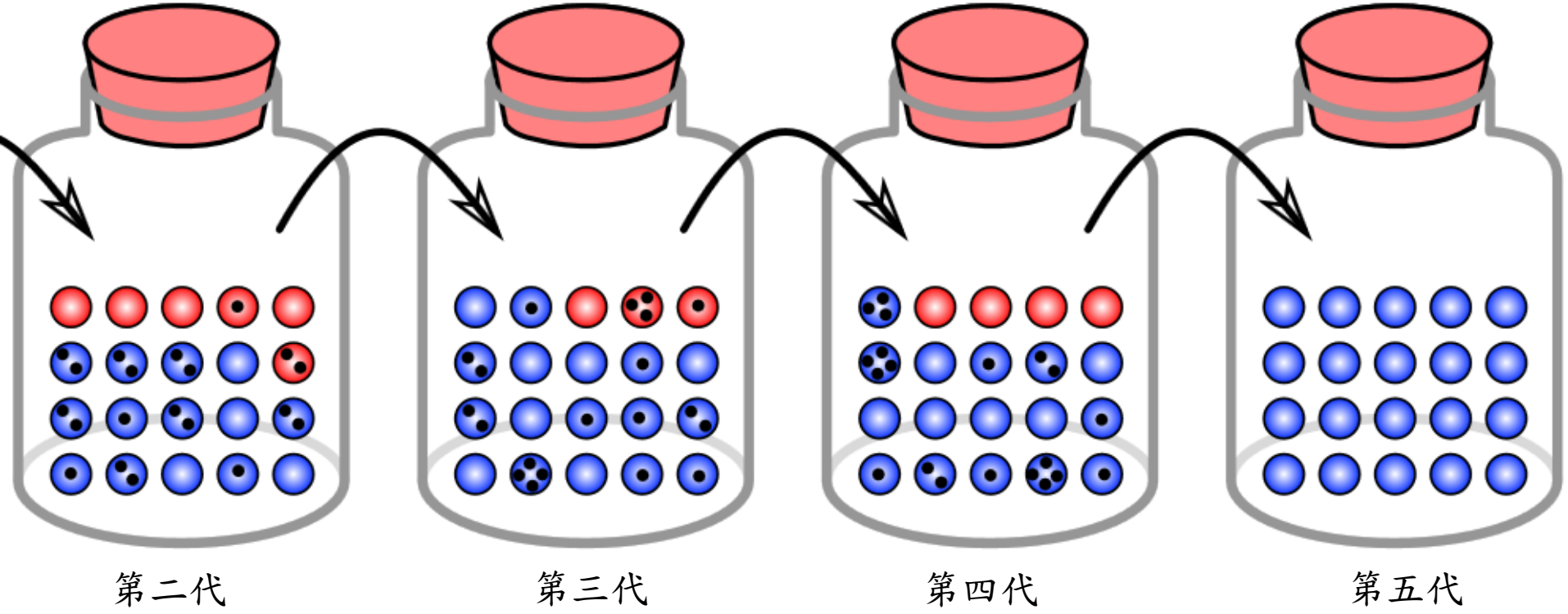


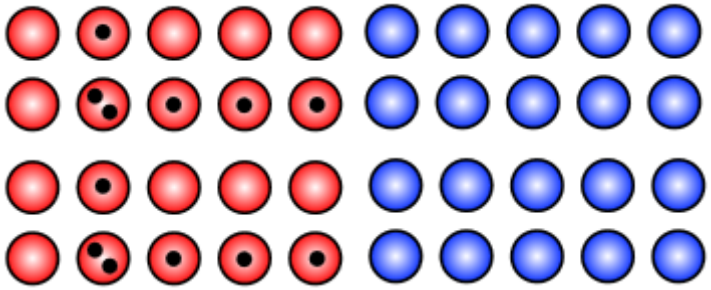
第二代



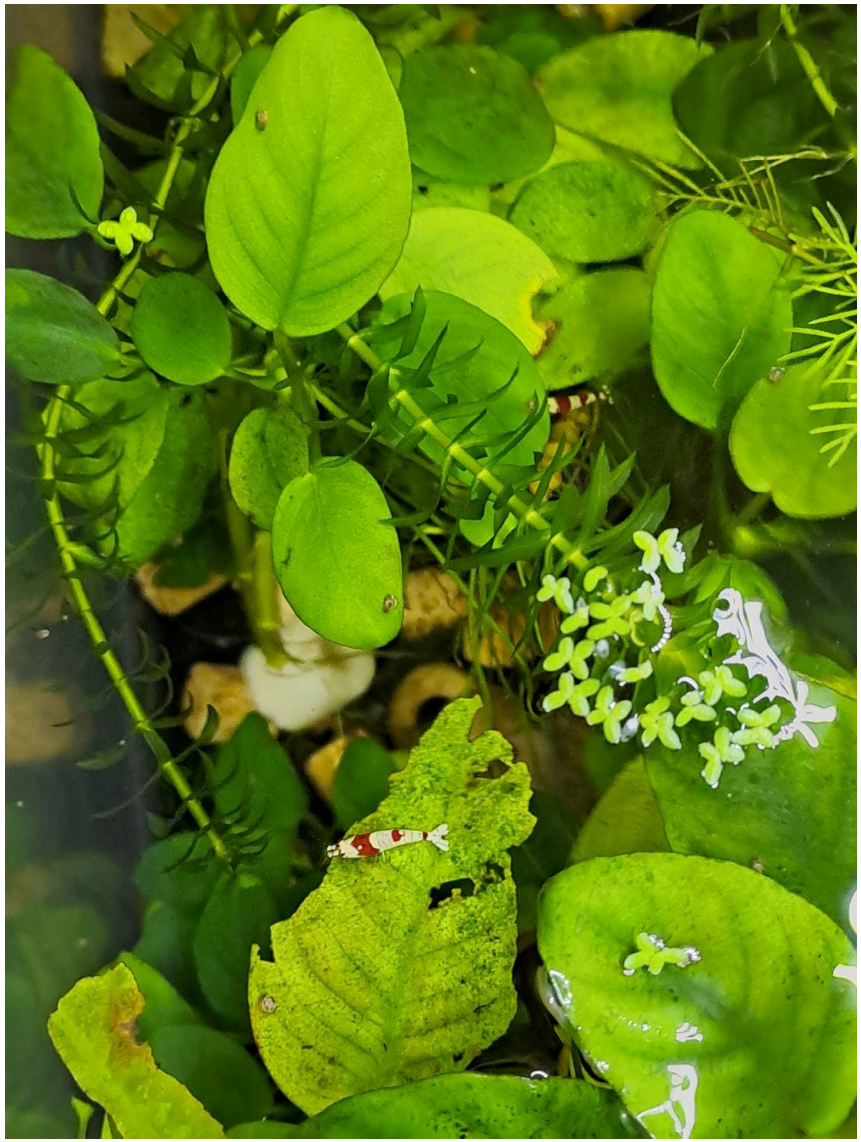
維基百科

•Genetic drift (基因漂變)

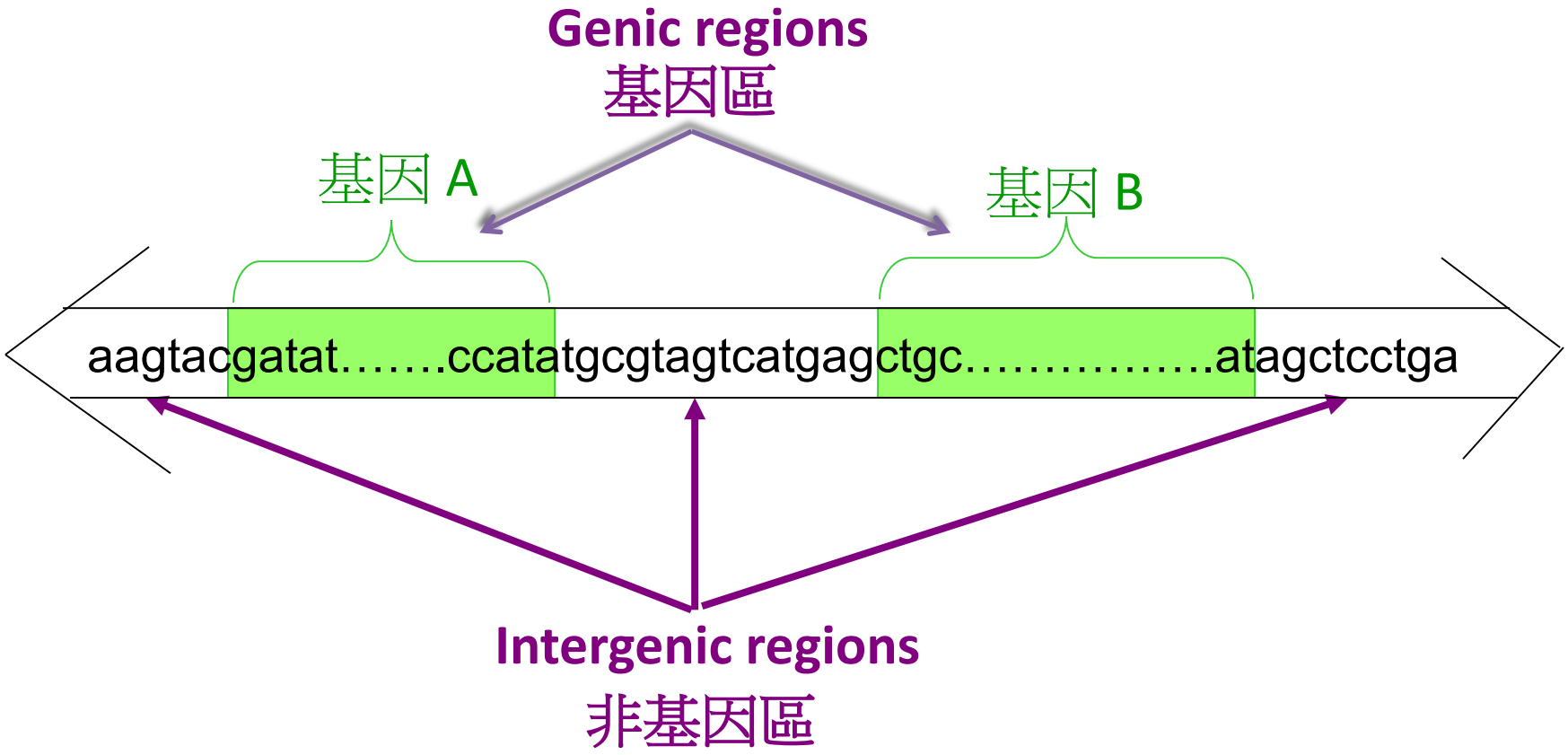




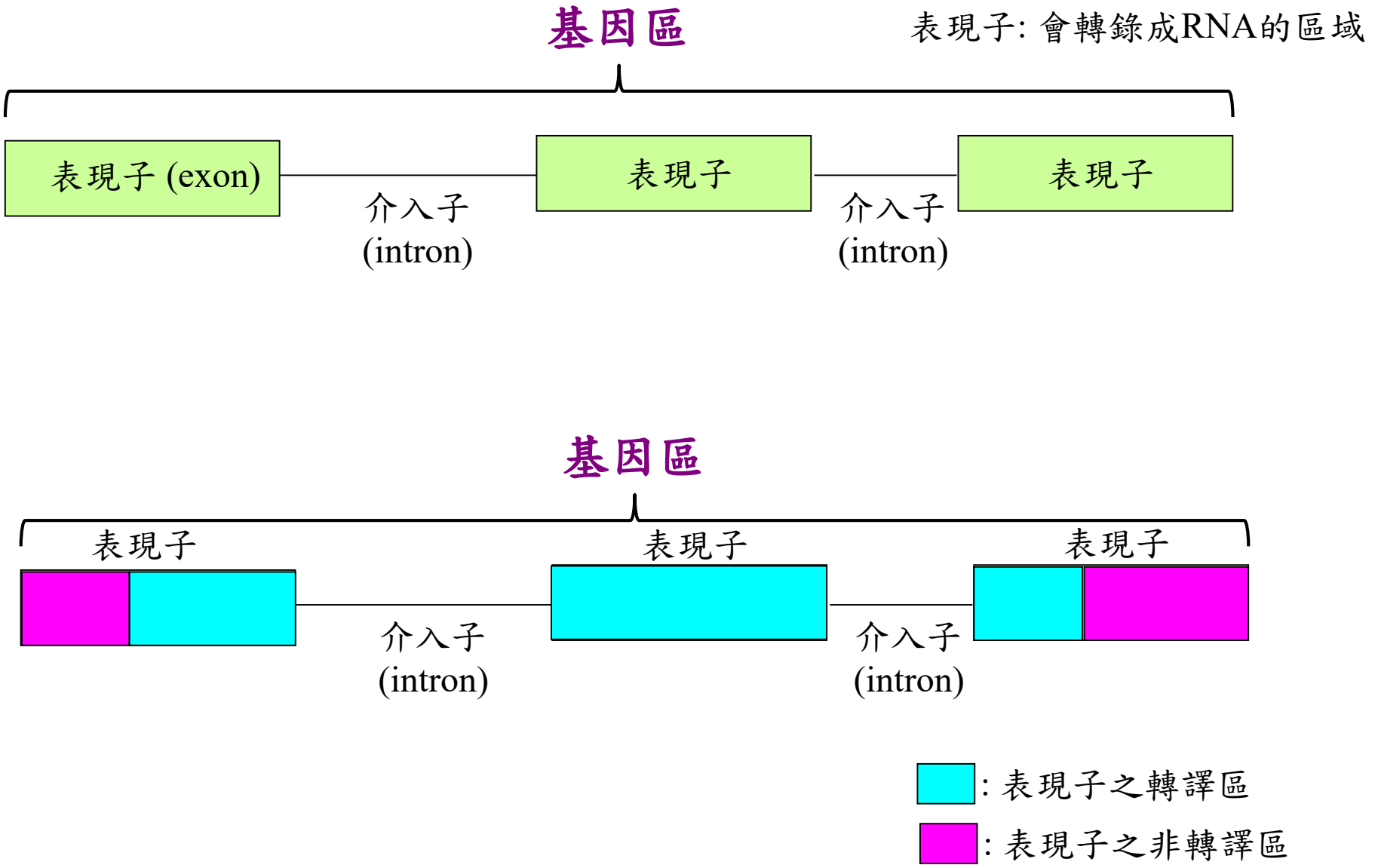
Gene flow (基因流動)



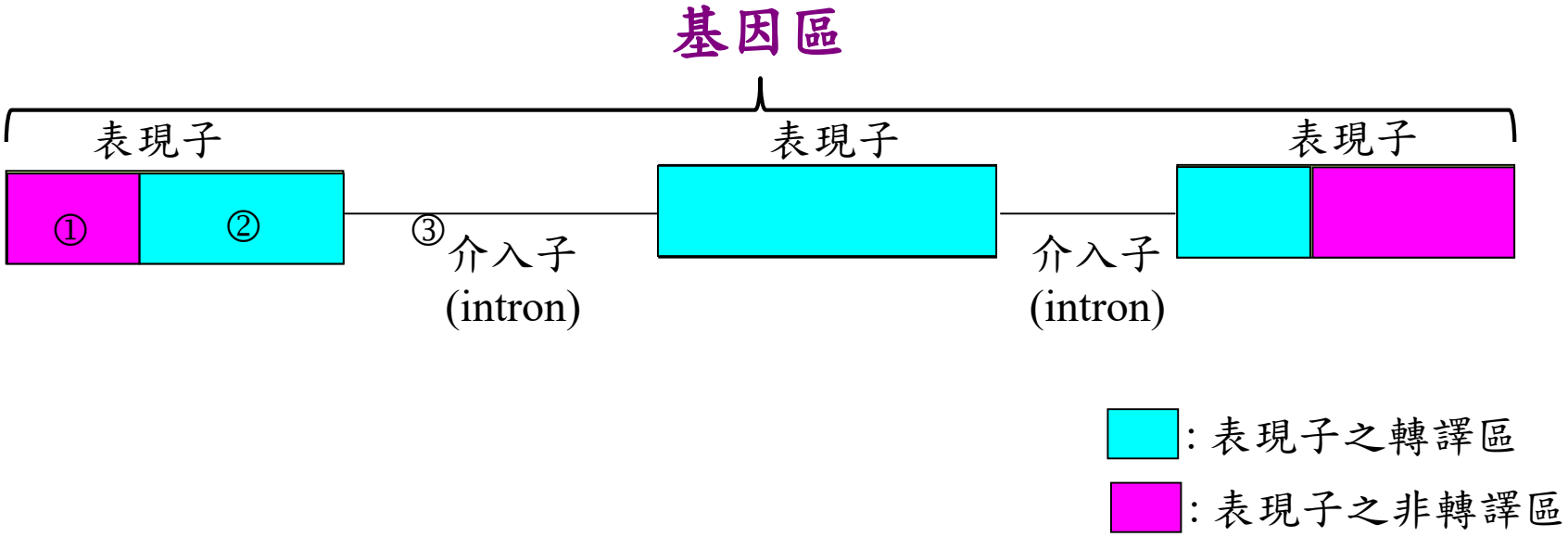
突變是隨機發生的。


Genic regions
基因區
基因 A
基因 B
aagtacgatat…….ccatatgcgtagtcatgagctgc……..atagctcctga
Intergenic regions
非基因區

突變是隨機發生的。


基因區
表現子: 會轉錄成RNA的區域
表現子 (exon)
介入子
(intron)
表現子
介入子
(intron)
表現子
基因區
表現子
介入子
(intron)
表現子
介入子
(intron)
表現子
: 表現子之轉譯區
: 表現子之非轉譯區

突變是隨機發生的。



基因區 (genic region)

- Exon (表現子;外顯子)
  - ✓ORF (open reading frame; 轉譯區)
  - ✓UTR (untranslated region; 非轉譯區)

- Intron (介入子; 內含子)

Genetic code 基因密碼

DNA sequence: A, C, G, T --- 4 letters
RNA sequence: A, C, G, U --- 4 letters
Amino acid sequence: --- 20 letters

64種密碼子➔20種胺基酸
(鴿籠原理)

First Position (5′)	Second position U	C	A	G	Third Position (3′)
U	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	U
	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	C
	Leu (**L**)	Ser (**S**)	**Stop**	**Stop**	A
	Leu (**L**)	Ser (**S**)	**Stop**	Trp (**W**)	G
C	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	U
	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	C
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	A
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	G
A	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	U
	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	C
	Ile (**I**)	Thr (**T**)	Lys (**K**)	Arg (**R**)	A
	Met (**M**)	Thr (**T**)	Lys (**K**)	Arg (**R**)	G
G	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	U
	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	C
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	A
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	G

開始編碼的密碼子(codon): AUG ← 必要但非充分條件
結束編碼密碼子: UAA, UAG, UGA ← 必要且充分條件

充分條件

A: 中樂透頭獎。 B: 覺得很快樂。
A: 下雨。 B: 無遮雨篷的操場濕掉。

A 是 B 的充分條件 (A發生一定導致B)

必要條件

A: 空氣的存在。 B: 人可以生存。
A: 努力。 B: 成功。
A: 父母。 B: 小孩。

A 是 B 的必要條件 (有B一定要有A)

充分且必要條件

A: 可被2整除的整數。 B: 偶數。
A: 可燃物、助燃物、溫度達到燃點。 B: 燃燒。

A: 吃某種食品或藥。 B: 病好了

Genetic code 基因密碼

64種密碼子➔20種胺基酸
(鴿籠原理)

First Position (5′)	Second position U(T)	Second position C	Second position A	Second position G	Third Position (3′)
U(T)	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	U
	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	C
	Leu (**L**)	Ser (**S**)	Stop	Stop	A
	Leu (**L**)	Ser (**S**)	Stop	Trp (**W**)	G
C	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	U
	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	C
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	A
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	G
A	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	U
	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	C
	Ile (**I**)	Thr (**T**)	Lys (**K**)	Arg (**R**)	A
	Met (M)	Thr (**T**)	Lys (**K**)	Arg (**R**)	G
G	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	U
	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	C
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	A
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	G

aagtacgatatg aatagtaacataaaagtagtcatgagctgg……

M N S N I K V V M S W …….

First Position (5′)	Second position U(T)	C	A	G	Third Position (3′)
U(T)	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	U
	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	C
	Leu (**L**)	Ser (**S**)	**Stop**	**Stop**	A
	Leu (**L**)	Ser (**S**)	**Stop**	Trp (**W**)	G
C	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	U
	Leu (**L**)	Pro (**P**)	His (**H**)	Arg (**R**)	C
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	A
	Leu (**L**)	Pro (**P**)	Gln (**Q**)	Arg (**R**)	G
A	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	U
	Ile (**I**)	Thr (**T**)	Asn (**N**)	Ser (**S**)	C
	Ile (**I**)	Thr (**T**)	Lys (**K**)	Arg (**R**)	A
	Met (M)	Thr (**T**)	Lys (**K**)	Arg (**R**)	G
G	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	U
	Val (**V**)	Ala (**A**)	Asp (**D**)	Gly (**G**)	C
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	A
	Val (**V**)	Ala (**A**)	Glu (**E**)	Gly (**G**)	G

沉默突變（silent mutation）：密碼改變，但對應的胺基酸不變。

aagtacgatatg aatagtaacataaaagtagtcatgagctgg……

M N S N I K V V M S W

aag

K

First Position (5′)	Second position U(T)	C	A	G	Third Position (3′)
U(T)	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	U
	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	C
	Leu **(L)**	Ser **(S)**	**Stop**	**Stop**	A
	Leu **(L)**	Ser **(S)**	**Stop**	Trp **(W)**	G
C	Leu **(L)**	Pro (**P**)	His (**H**)	Arg (**R**)	U
	Leu **(L)**	Pro (**P**)	His (**H**)	Arg (**R**)	C
	Leu **(L)**	Pro (**P**)	Gln **(Q)**	Arg **(R)**	A
	Leu **(L)**	Pro (**P**)	Gln **(Q)**	Arg **(R)**	G
A	Ile **(I)**	Thr (**T**)	Asn **(N)**	Ser (**S**)	U
	Ile **(I)**	Thr (**T**)	Asn **(N)**	Ser (**S**)	C
	Ile **(I)**	Thr (**T**)	Lys **(K)**	Arg **(R)**	A
	Met (M)	Thr (**T**)	Lys **(K)**	Arg **(R)**	G
G	Val **(V)**	Ala (**A**)	Asp (**D**)	Gly (**G**)	U
	Val **(V)**	Ala (**A**)	Asp (**D**)	Gly (**G**)	C
	Val **(V)**	Ala (**A**)	Glu **(E)**	Gly **(G)**	A
	Val **(V)**	Ala (**A**)	Glu **(E)**	Gly **(G)**	G

錯義突變（missense mutation）：使密碼所對應的胺基酸改變。

aagtacgatatg aatagtaacataaaagtagtcatgagctgg……

M N S N I K V V M S W ……

aat

N

First Position (5′)	Second position U(T)	C	A	G	Third Position (3′)
U(T)	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	U
	Phe (**F**)	Ser (**S**)	Tyr (**Y**)	Cys (**C**)	C
	Leu **(L)**	Ser **(S)**	**Stop**	**Stop**	A
	Leu **(L)**	Ser **(S)**	**Stop**	Trp **(W)**	G
C	Leu **(L)**	Pro (**P**)	His (**H**)	Arg (**R**)	U
	Leu **(L)**	Pro (**P**)	His (**H**)	Arg (**R**)	C
	Leu **(L)**	Pro (**P**)	Gln **(Q)**	Arg **(R)**	A
	Leu **(L)**	Pro (**P**)	Gln **(Q)**	Arg **(R)**	G
A	Ile **(I)**	Thr (**T**)	Asn **(N)**	Ser (**S**)	U
	Ile **(I)**	Thr (**T**)	Asn **(N)**	Ser (**S**)	C
	Ile **(I)**	Thr (**T**)	Lys **(K)**	Arg **(R)**	A
	Met (M)	Thr (**T**)	Lys **(K)**	Arg **(R)**	G
G	Val **(V)**	Ala (**A**)	Asp (**D**)	Gly (**G**)	U
	Val **(V)**	Ala (**A**)	Asp (**D**)	Gly (**G**)	C
	Val **(V)**	Ala (**A**)	Glu **(E)**	Gly **(G)**	A
	Val **(V)**	Ala (**A**)	Glu **(E)**	Gly **(G)**	G

無義突變（nonsense mutation）：使原本可製造蛋白質的密碼變成停止密碼。

aagtacgatatg aatagtaacataaaagtagtcatgagctgg……

M N S N I K V V M S W …….

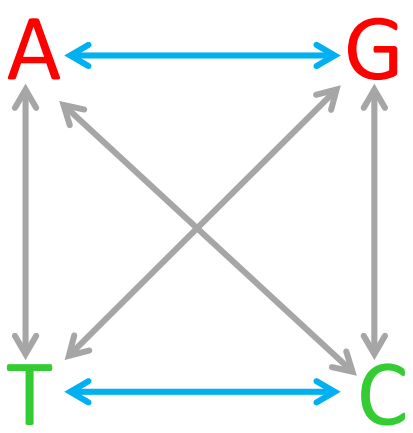
tag

Stop codon

突變是隨機發生的。

A（adenine，腺嘌呤）
G（guanine，鳥嘌呤）

T（thymine，胸腺嘧啶）
C（cytosine，胞嘧啶）



嘌呤類

嘧啶類

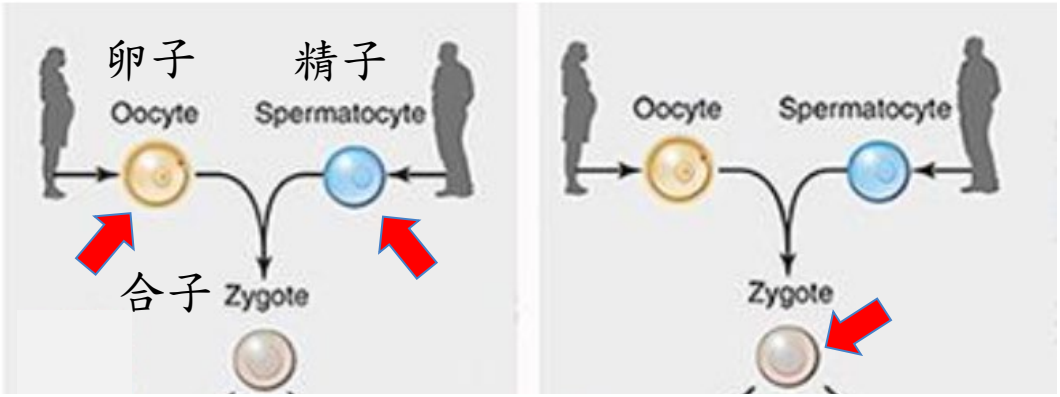
突變是隨機發生的。

A（adenine，腺嘌呤）
G（guanine，鳥嘌呤）

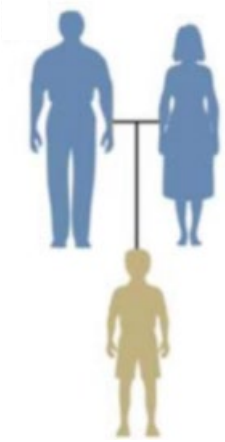
T（thymine，胸腺嘧啶）
C（cytosine，胞嘧啶）

地址：
115201臺北市南港區研究院路一段128號

de novo mutation (新發生的突變)

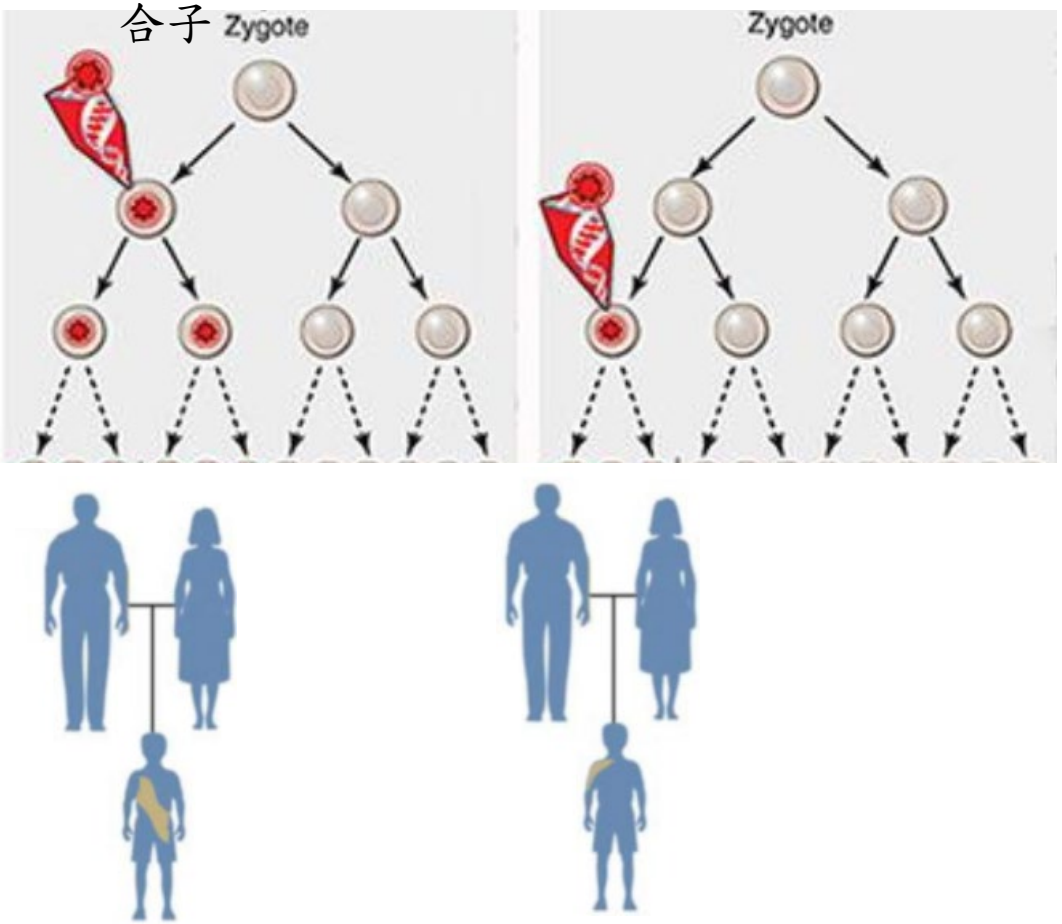


[Courtesy of Science/AAAS, Poduri et al., 2013.]



Modified from *Genes* **2014**, *5*(4), 1064-1094.

Mosaicism (鑲嵌現象)



[Courtesy of Science/AAAS, Poduri et al., 2013.]

Modified from *Genes* **2014**, *5*(4), 1064-1094.

人有多少個基因?

	Organism	# of protein-coding genes	Genome size
人類免疫缺陷病毒	HIV 1	9	10
甲型流感病毒	*Influenza A* virus	10-11	14
噬菌體	Bacteriophage λ	66	49
人類皰疹病毒第四型	Epstein Barr virus	80	170
鬼羽箭屬菌	*Buchnera sp.*	610	640
海棲熱袍菌	*T. maritima*	1,900	1,900
金黃色葡萄球菌	*S. aureus*	2,700	2,900
霍亂弧菌	*V. cholerae*	3,900	4,000
枯草桿菌	*B. subtilis*	4,400	4,200
大腸桿菌	*E. coli*	4,300	4,600
釀酒酵母	*S. cerevisiae*	6,600	12,000
秀麗隱桿線蟲	*C. elegans*	20,000	100,000
阿拉伯芥	*A. thaliana*	27,000	140,000
黑腹果蠅	*D. melanogaster*	14,000	140,000
紅鰭東方鮭	*F. rubripes*	19,000	400,000
玉米	*Z. mays*	33,000	2,300,000
小鼠	*M. musculus*	20,000	2,800,000
智人	*H. sapiens*	21,000	3,200,000
普通小麥	*T. aestivum* (hexaploid)	95,000	16,800,000

viruses

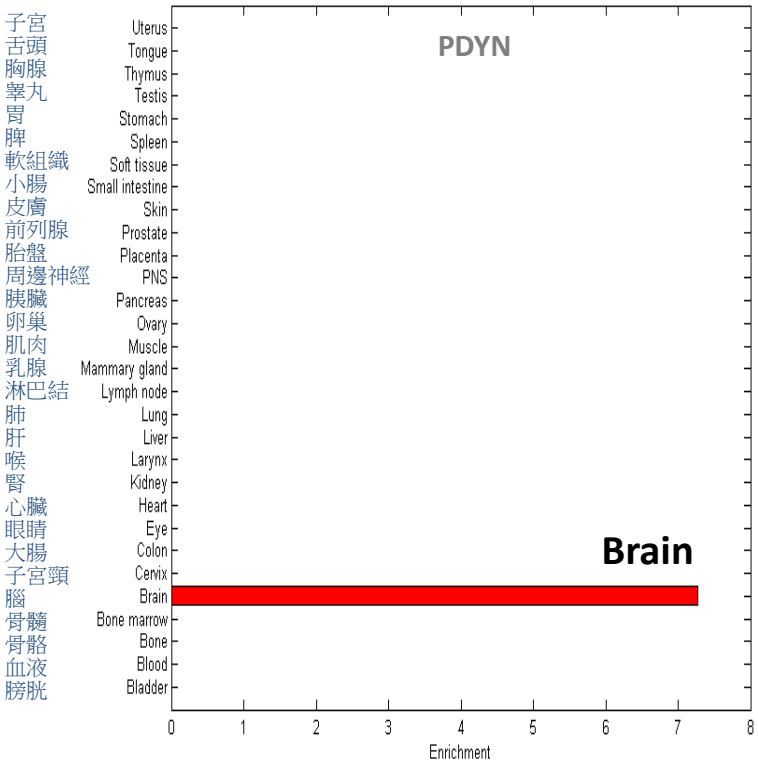
prokaryotes

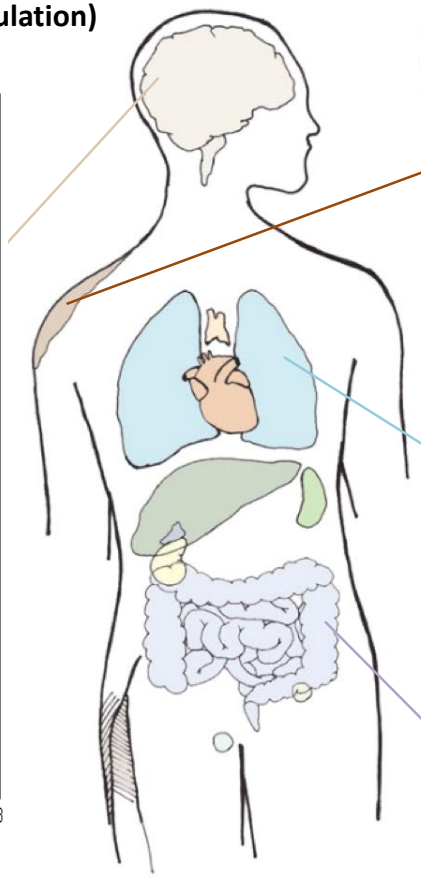
eukaryotes

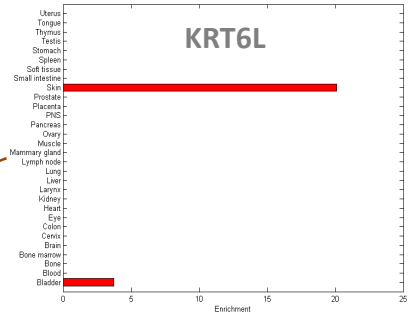
($\times 10^3$)

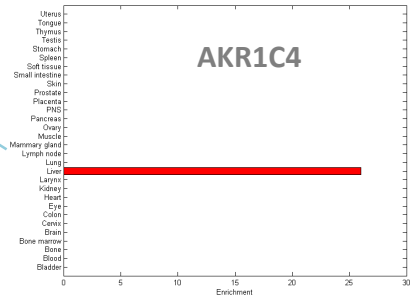
(http://book.bionumbers.org/how-many-genes-are-in-a-genome/)

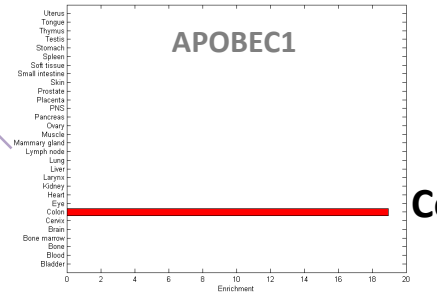
TiGER database (Tissue-specific Gene Expression and Regulation)

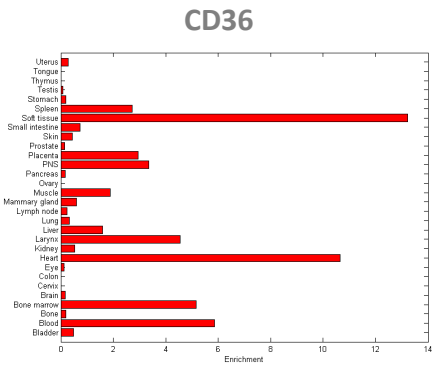


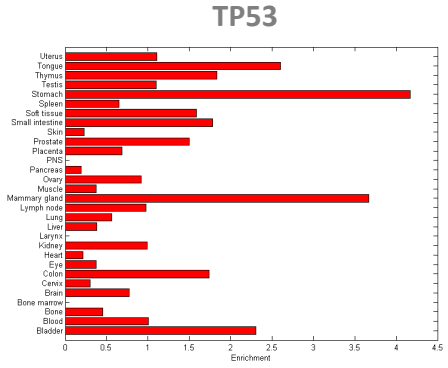


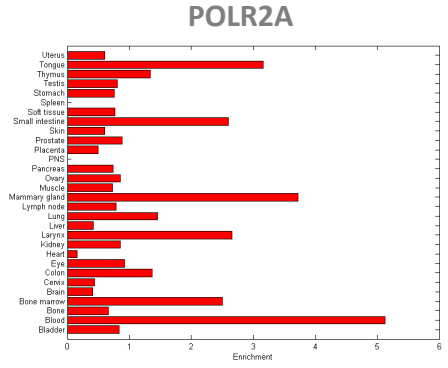




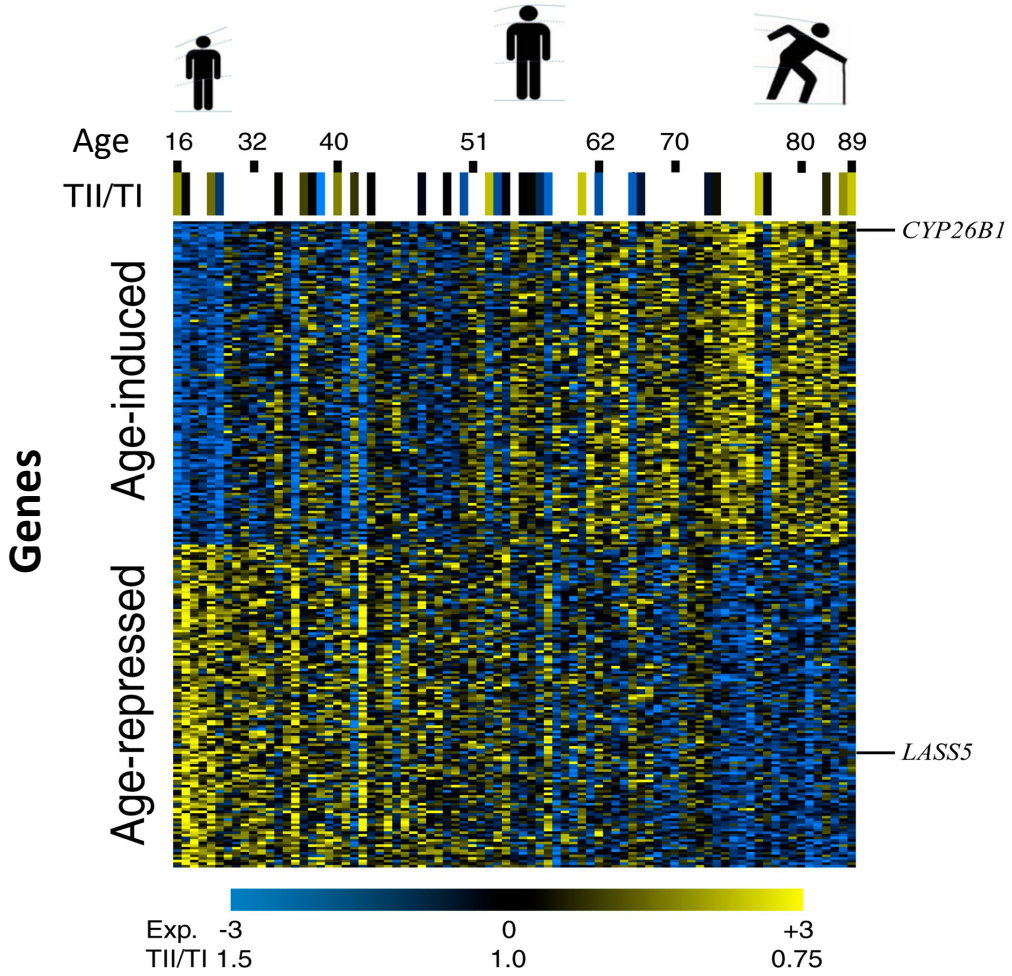








年輕人的基因與老年人的基因



Zahn, Jacob M., et al. "Transcriptional profiling of aging in human muscle reveals a common aging signature." PLoS Genet 2.7 (2006).

DNA 間的差異和疾病

GWAS: Genome-Wide *Association* Study
全基因體關聯性分析研究

單核苷酸多態性 (Single Nucleotide Polymorphism; SNP)

aagtacgatcggccga

aagtgcgattggccga

aagtacgatcggccga
aagtacgattggccga

aagtgcgatcggccga
aagtgcgattggccga

基因型 (genotype): aa, gg, ag

同型合子（homozygous）➔ aa, gg
異型合子（heterozygous）➔ ag

單核苷酸多態性 (Single Nucleotide Polymorphism; SNP)

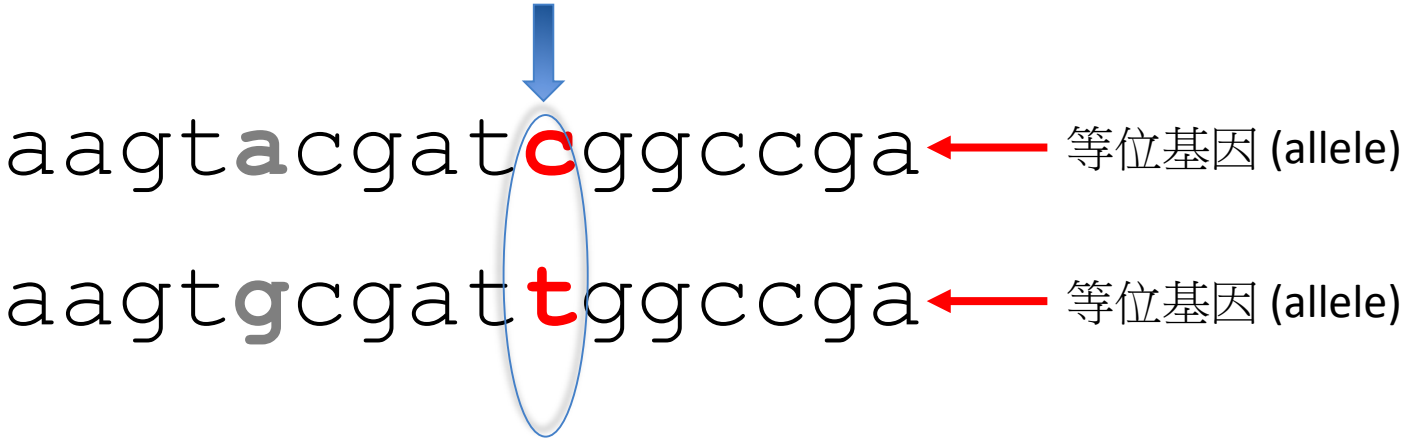
aagt**a**cgat**c**ggccga

aagt**g**cgat**t**ggccga

基因型 (genotype): cc, ct, tt

同型合子（homozygous）
異型合子（heterozygous）

單核苷酸多態性 (Single Nucleotide Polymorphism; SNP)

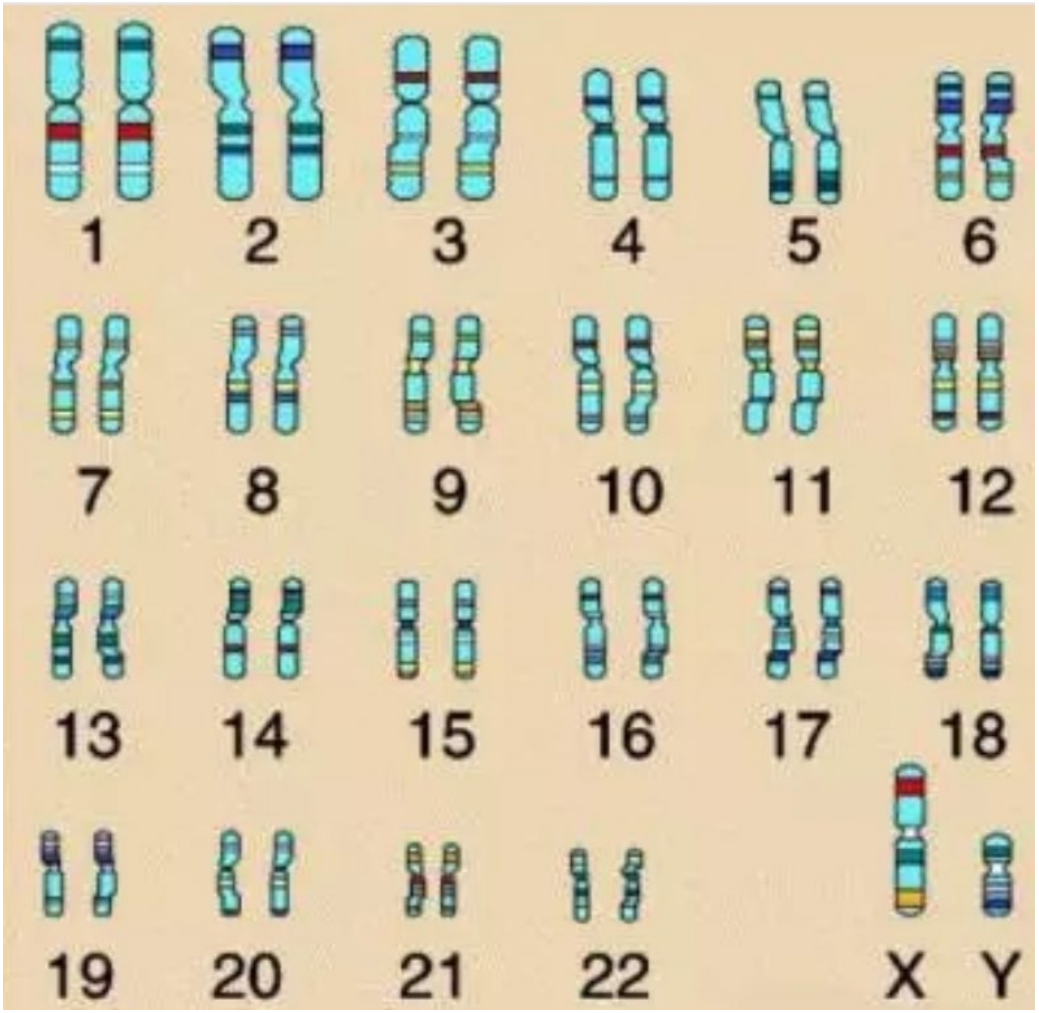


基因型 (genotype): cc, ct, tt

同型合子（homozygous）
異型合子（heterozygous）

人是二倍體
只有精子和卵子是單倍體
其他體細胞都是雙倍體

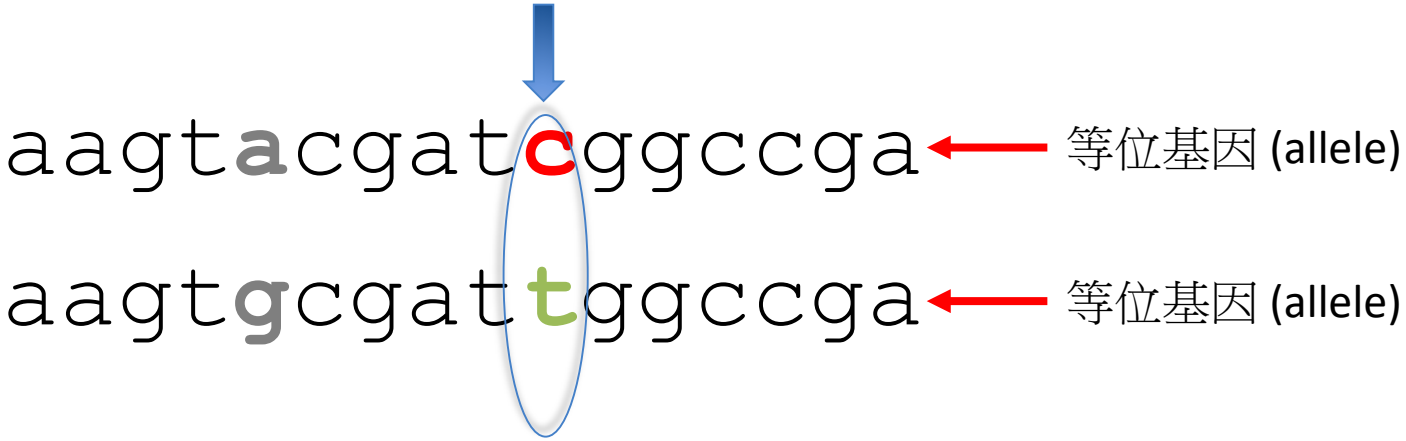
所有 DNA 序列的總稱叫做: 基因體 (genome)



常染色體 (autosomes)
性染色體 (sex chromosome)

人類的DNA序列由30億 (3×10^9) 個A、G、T、C核苷酸排列組合而成

單核苷酸多態性 (Single Nucleotide Polymorphism; SNP)



基因型 (genotype): cc, ct, tt

同型合子（homozygous）
異型合子（heterozygous）

100個人共有幾條alleles?

90個人基因型為 cc
3個人基因型為 ct
7個人基因型為 tt

c的allele frequency=? (90×2+3)/(100×2)=91.5%
t的allele frequency=? (7×2+3)/(100×2)=8.5%

DNA 間的差異 (基因型差異)
(自閉症患者 vs. 無自閉症健康者)

GWAS: Genome-Wide *Association* Study
全基因體關聯性分析研究

Grove et al. *Nature Genetics* (2019)

自閉症患者（18,381人）

A是主要等位基因➔ ~100%
G是次要等位基因➔ ~0%

健康者（27,969人）

G是主要等位基因➔ ~100%
A是次要等位基因➔ ~0%

遺傳性疾病

是指以基因為主要致病原因的疾病。

單一基因缺陷造成的遺傳疾病

- 鐮刀型貧血症
- 白化症

多重基因共同影響所造成的遺傳疾病

- 冠狀心臟疾病
- 高血壓
- 中風
- 許多種類的癌症

膠質瘤 (glioma)

WHO 分級: Grades I, II, III, IV

腦瘤分級

	第一級	第二級	第三級	第四級
特色	惡性度最低，透過手術可能治癒	生長速度慢，局部治療後仍可能復發，部分腫瘤易進展為高惡性	有浸潤能力，且具有核異形性和分裂能力增加	分裂能力快、侵略性強，容易轉移跟復發
類型	腦膜瘤、神經節膠質細胞瘤等	中心性神經細胞瘤、室管膜瘤等	分化不良星狀細胞瘤、分化不良寡樹突膠質瘤等	神經膠質母細胞瘤、髓母細胞瘤等
存活時間	通常時間較長	約六至八年	約三年	不超過兩年

Low Grade Glioma (LGG)

yahoo! 新聞

Grade IV ➔ 多型性神經膠母細胞瘤（glioblastoma multiforme，GBM）
➔ 高惡性腫瘤

膠質瘤 (glioma)

多型性神經膠母細胞瘤 (GBM)

平均存活期約15個月
五年存活率只有約5%

台灣每年約有1000至2000人罹患腦瘤

檢查方式包括有：
電腦斷層掃瞄（Computed Tomography；CT）、磁共振成像（Magnetic Resonance Imaging；MRI）、腦波圖（Electroncephalography；EEG）以及腦血管攝影。

膠質瘤 (glioma)

多型性神經膠母細胞瘤 (GBM)

原發性 (primary GBM) →……..→ 復發性 (recurrent GBM)

原發到再次被診斷復發時間 (Time to relapse; TTR)

疾病無進展期間 (progression-free interval; PFI)

膠質瘤 (glioma) 多型性神經膠母細胞瘤 (GBM)

建立復發時間預測模型

Training set (訓練資料集)
Validation set (驗證資料集)
➔以訓練資料集來建立預測模型。
➔同一個來源的資料，以8:2比率分訓練與驗證資料集。

Testing set (測試資料集)
➔以另一組不同來源的資料測試所建立的模型。

樣本數量太小➔ 訓練資料集+測試資料集

膠質瘤 (glioma) 多型性神經膠母細胞瘤 (GBM)

建立復發時間預測模型

Training set (訓練資料集)
Validation set (驗證資料集)
➔以訓練資料集來建立預測模型。
➔同一個來源的資料，以8:2比率分訓練與驗證資料集。

Testing set (測試資料集)
➔以另一組不同來源的資料測試所建立的模型。

建立復發時間預測模型的方法

統計模型
機器學習—含類神經網路(深度學習)
統計模型+機器學習

最怕的事 ➔ garbage in garbage out

三件最重要的事:

1. 作弊、抄襲、造假: 害人害己
   (遲交、缺交、抄襲)

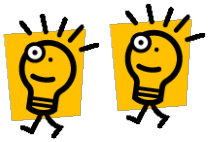
2. 謀定而後動、不要害怕失敗

3. 探索未知的喜悅

- 比較與演化基因體學（資訊統計）
- 調控因果生物學（資訊統計/分生實驗）
- 系統生物學（資訊統計/分生實驗）
- 神經精神疾病（分生實驗）
- 機器學習與多體學資料分析（資訊統計）

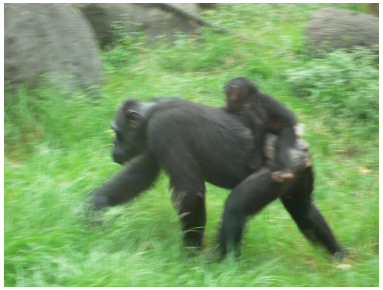
生活的樂趣
在於創造希望

Homepage:
http://idv.sinica.edu.tw/trees/



基因、演化及大數據分析

黑猩猩 *Pan troglodytes*



人 *Homo sapiens*

大數據分析、演化、神經科學實驗室

主持人: Trees-Juen Chuang 莊樹諄

http://idv.sinica.edu.tw/trees/

Email: trees@gate.sinica.edu.tw

Academia Sinica 中央研究院 基因體研究中心

Genomics Research Center