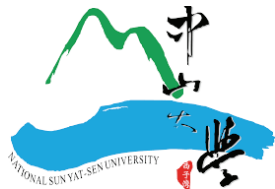


# Genomics- Human genome project

## 基因組學－人類基因體計劃及其應用



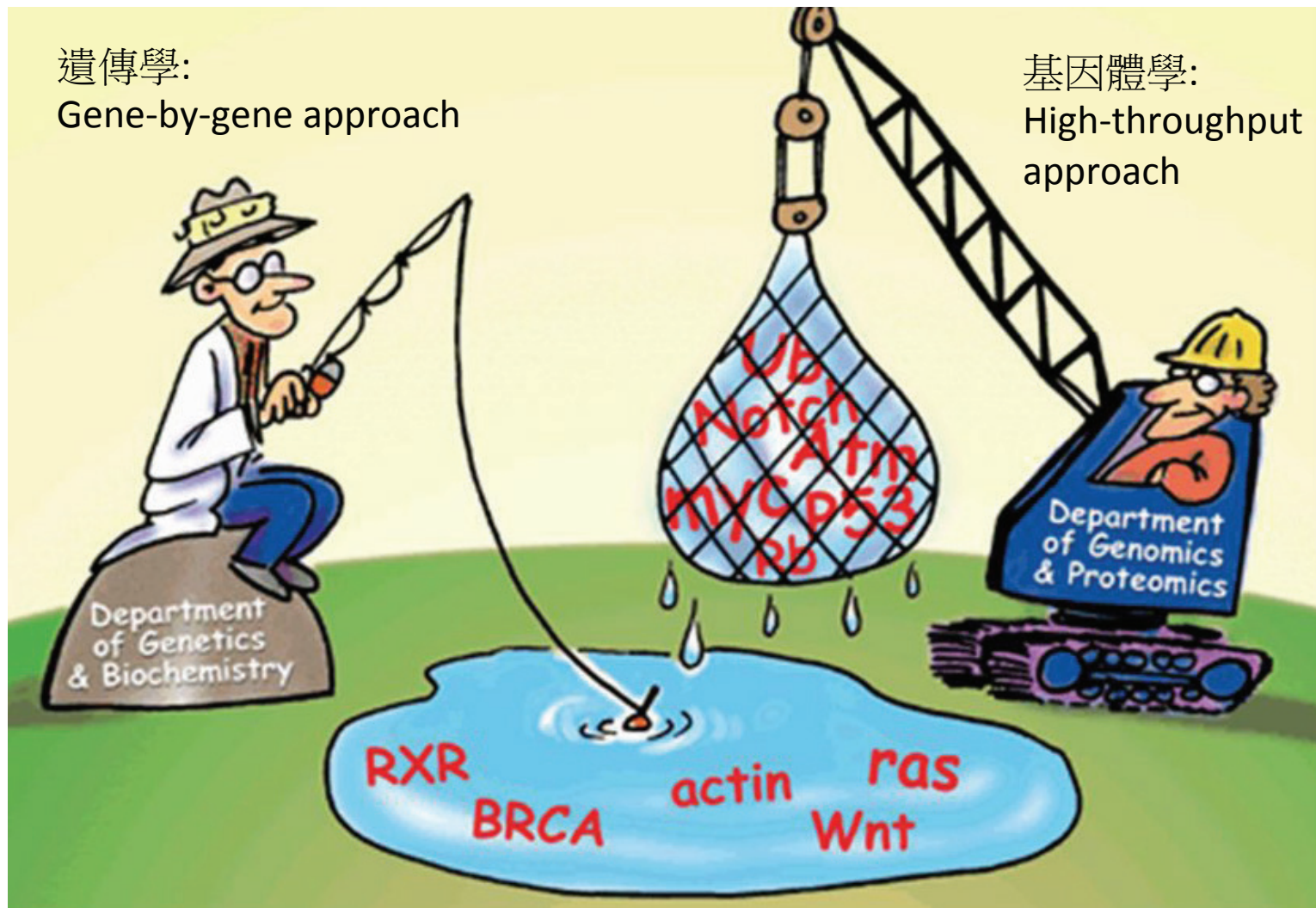
國立中山大學 生物科學系 黃明德

# 基因組學 (Genomics)

- 基因組：細胞內所有的DNA序列，包含核DNA、葉綠體DNA、粒腺體DNA
- 基因組學(基因體學)
  - 目的：研究基因組結構、基因功能、基因演化
  - 工具：
    - 基因組定序 – 基因組完整定序
    - 生物資訊學 – 序列組裝及分析
    - 遺傳學 - 基因功能分析
  - 流程：基因組定序 -> 基因註解 -> 基因功能分析

後基因體時代

- 基因學遺傳學相較，為以高通量(high-throughput)策略研究基因功能



如果你在研究一個性狀，你要如何找到標的基因？

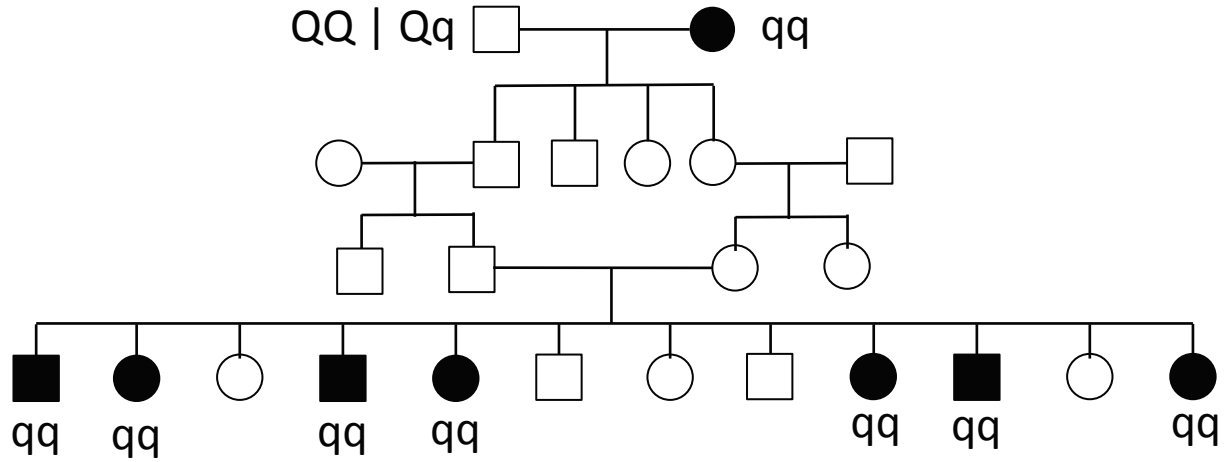




- 科學家利用其它性狀標定基因

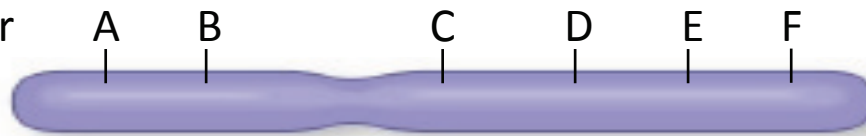


# 家族圖譜與基因定位



#1	aa	aa	AA	Aa			Aa	AA		aa
#2	Bb	Bb	bb	Bb			Bb	bb		bb
#3	cc	Cc	Cc	cc			Cc	Cc		cc
#4	dd	Dd	dd	dd			Dd	dd		dd
#5	ee	ee	ee	ee			ee	ee		ee
#6	ff	ff	Ff	ff			ff	Ff		ff

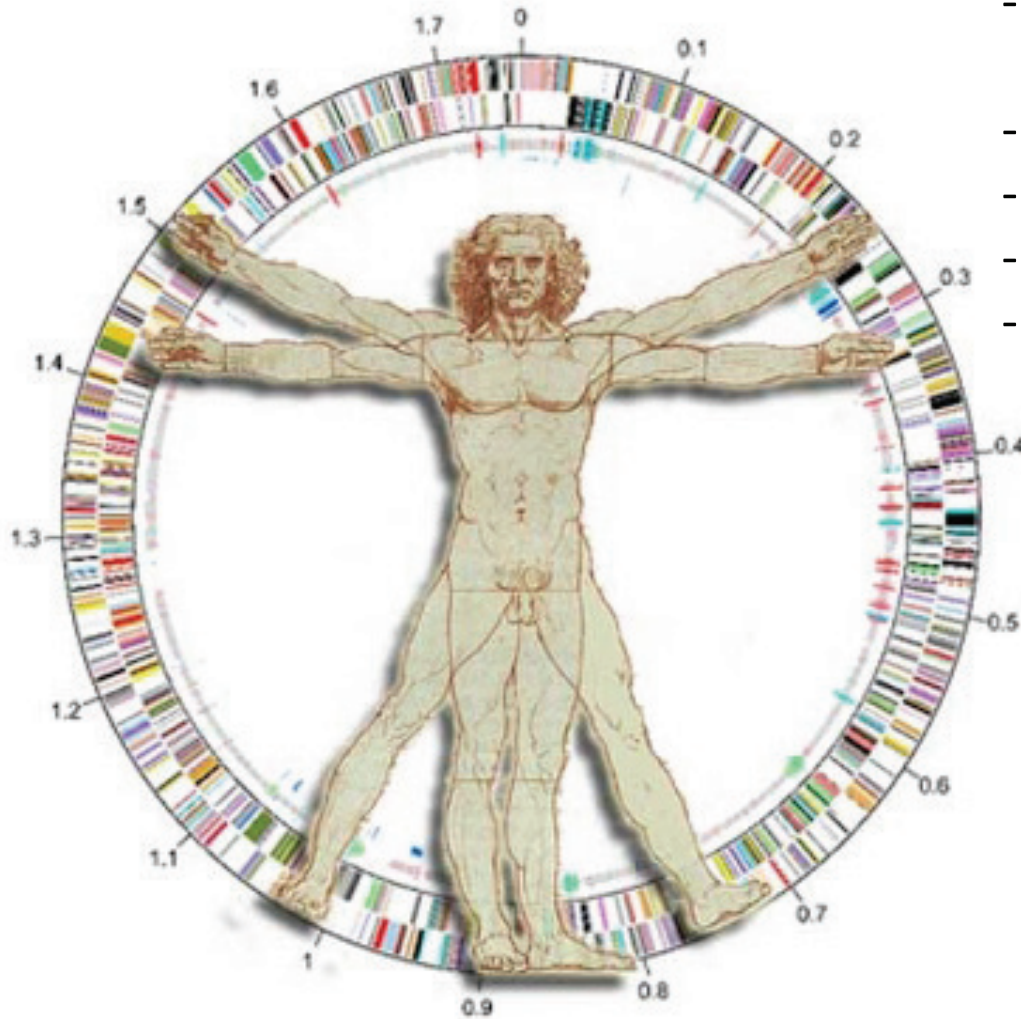
Marker



## Solomon islands blonde (索羅門群島金髮.)



# Human genome project



- 目的：將人類基因組序列完全定序並註解所有基因
- 1990計劃啟動
- 2003公佈草圖
- 總經費 \$3,000,000,000美元
- 共18個國家參與

# 人類基因組計劃/ Human genome project (HGP)

- 1984 – 科學家於美國能源部會議提出構想
- 1986 – 與會科學家再次強調該計劃重要性並討論
- 1988 – 與會科學家一致同意該計劃重要性並準備著手進行
- 1990 - 提出初步構想(為期15年, 經費美金 \$3,000,000,000, 採用階層式定序法)
- 1992 - 發布低解析度基因組草圖(genome map)
- 1998 – Celera公司宣佈將以霰彈槍定序法於五年內完成基因組定序, 經費 \$300,000,000, 完成後將註冊所有基因
- 1999 – 第一條染色體公布 (chromosome 22)
- 2000 - Celera公司宣佈已完成 ~97%
- 2003 - 人類基因組計劃完成



## 定序金額試算

- Sanger 定序法  
單次定序: 500-1000 bp  
單次定序費用: \$ 3
- Human genome: 3,200,000,000 bp  
 $3,200,000,000 / 500 = 6,400,000$  定序次數  
 $6,400,000 * 3 = \text{NT\$ } 19,200,000$







**Why need \$3,000,000,000?**

# 基因組大小

- 單位

1 bp = 1 bp , 1 kb = 1,000 bp, 1MB = 1,000,000 bp,  
1GB = 1,000,000,000 bp

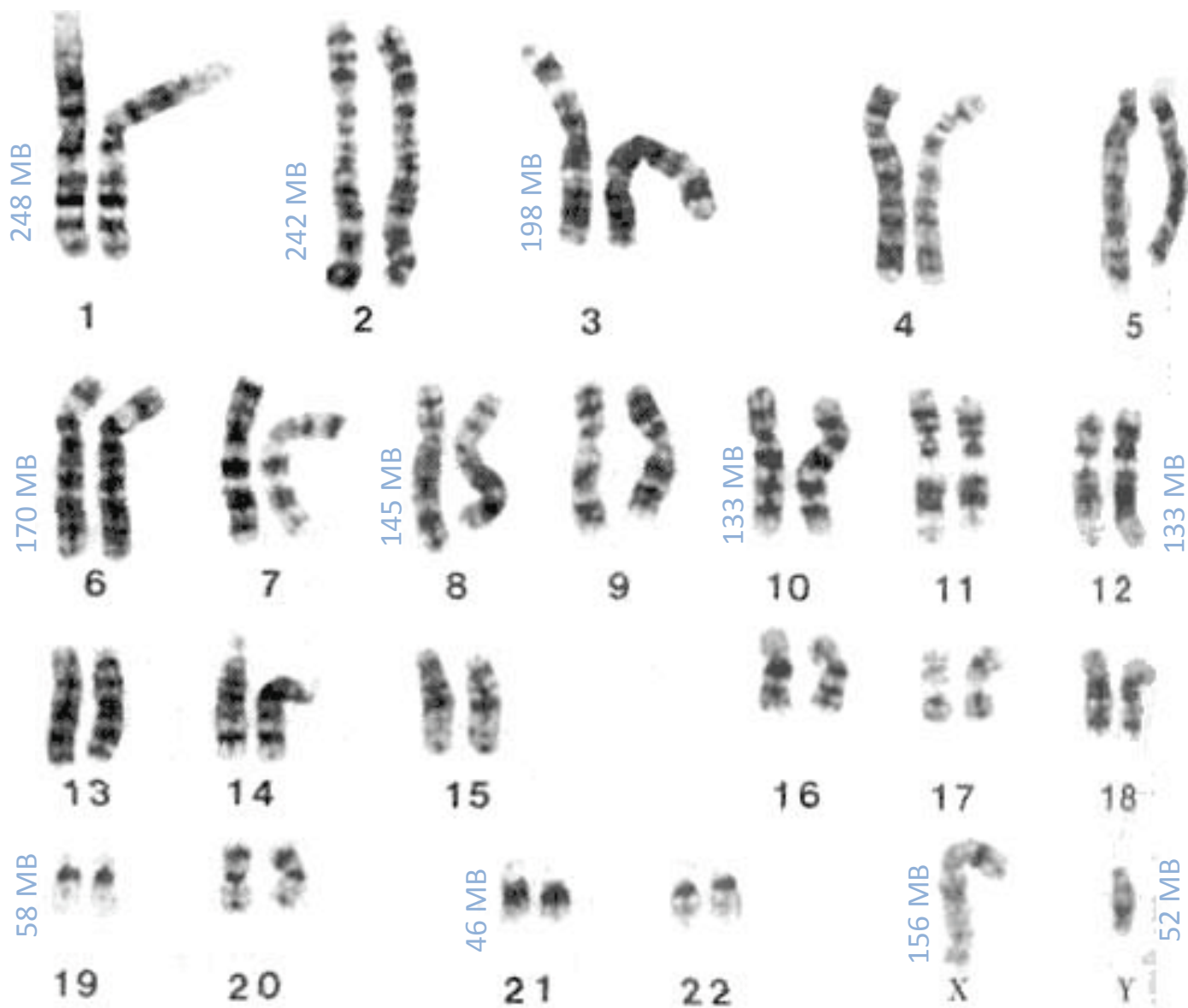
- 基因組大小

Species	<i>Porcine circovirus</i>	<i>Escherichia coli</i>	<i>Caenorhabditis elegans</i>	<i>Drosophila melanogaster</i>	<i>Homo sapiens</i>	<i>Amoeba dubia</i>
Genome Size	1759 bp	4.6 MB	100 MB	130 MB	3.2 GB	670 GB
Common Name	 Virus	 Bacteria	 Nematode	 Fruit fly	 Human	 Ameoba

基因數目                      3                      4288                      19,000                      13,600                      ~ 20,000                      ?

C值謎 (C-value enigma): 生物的C值 (或基因組大小) 並不與生物複雜程度相關的現象

# 人類染色體數目及其大小

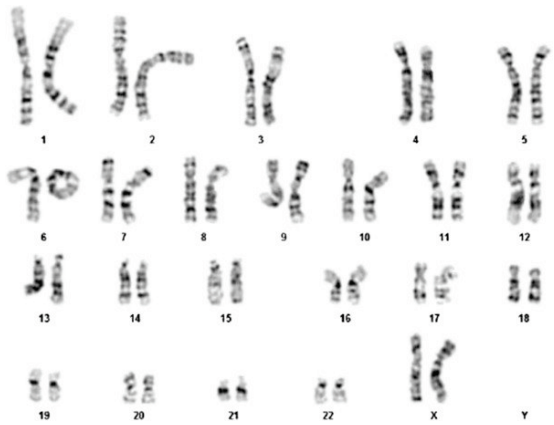


Total: 3,234.83 Mb

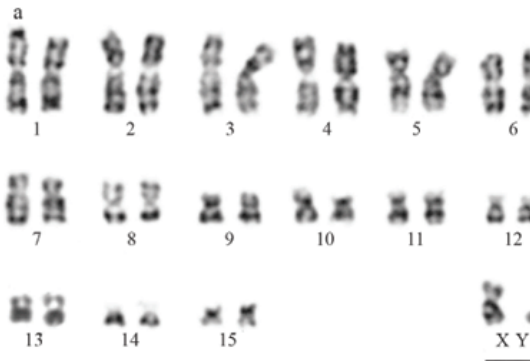
# 染色體條帶技術

- 利用染劑使染色體呈現各自獨特條帶形態，藉以區別染色體的不同

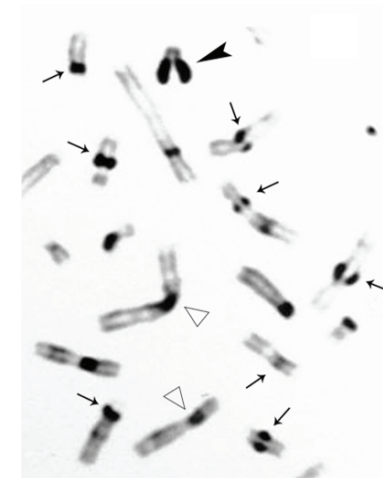
技術	方法	亮帶	暗帶
G 帶	胰酶 + Giemsa	GC rich	AT rich
R 帶	塩處理 + Giemsa	AT rich	GC rich
Q 帶	Quinacrine	GC rich	AT rich
C 帶	Ba(OH) <sub>2</sub> + Giemsa	著絲點以外	著絲點



G-banding

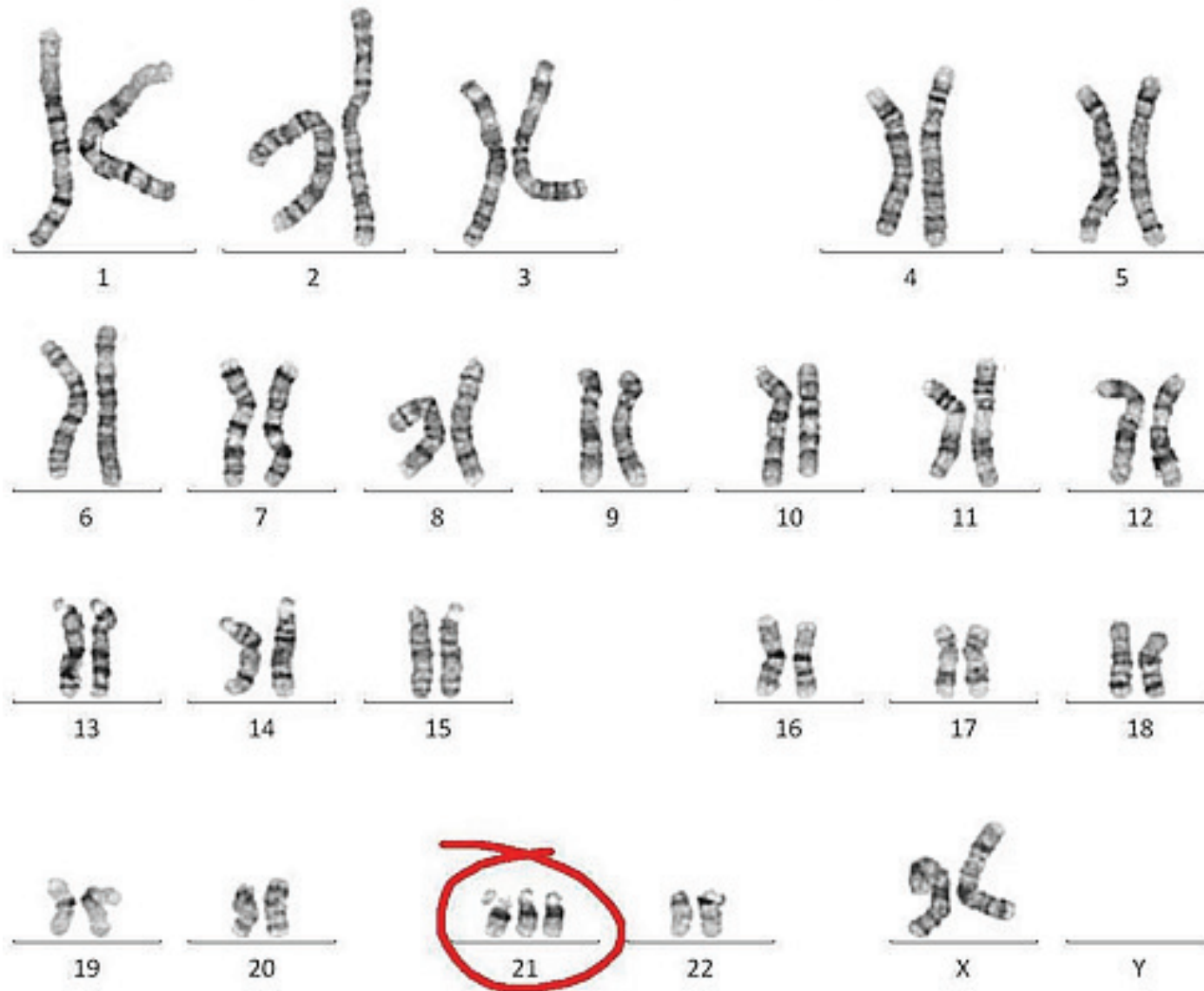


R-banding



C-banding

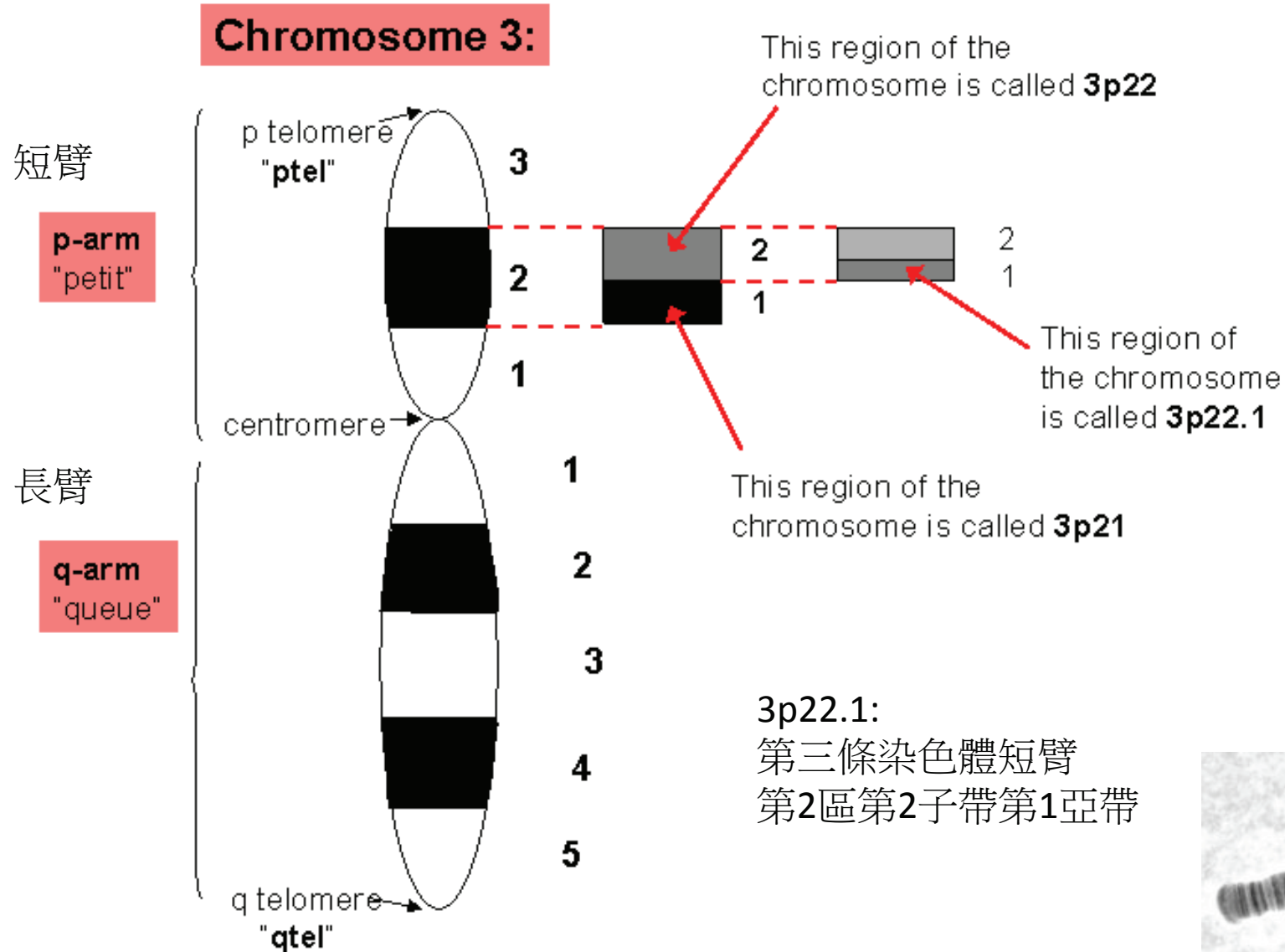
# 唐氏症-第21號染色體異常



<https://www.quora.com/How-does-having-an-extra-chromosome-cause-Down-syndrome>



# 細胞遺傳圖譜 (cytogenetic map)

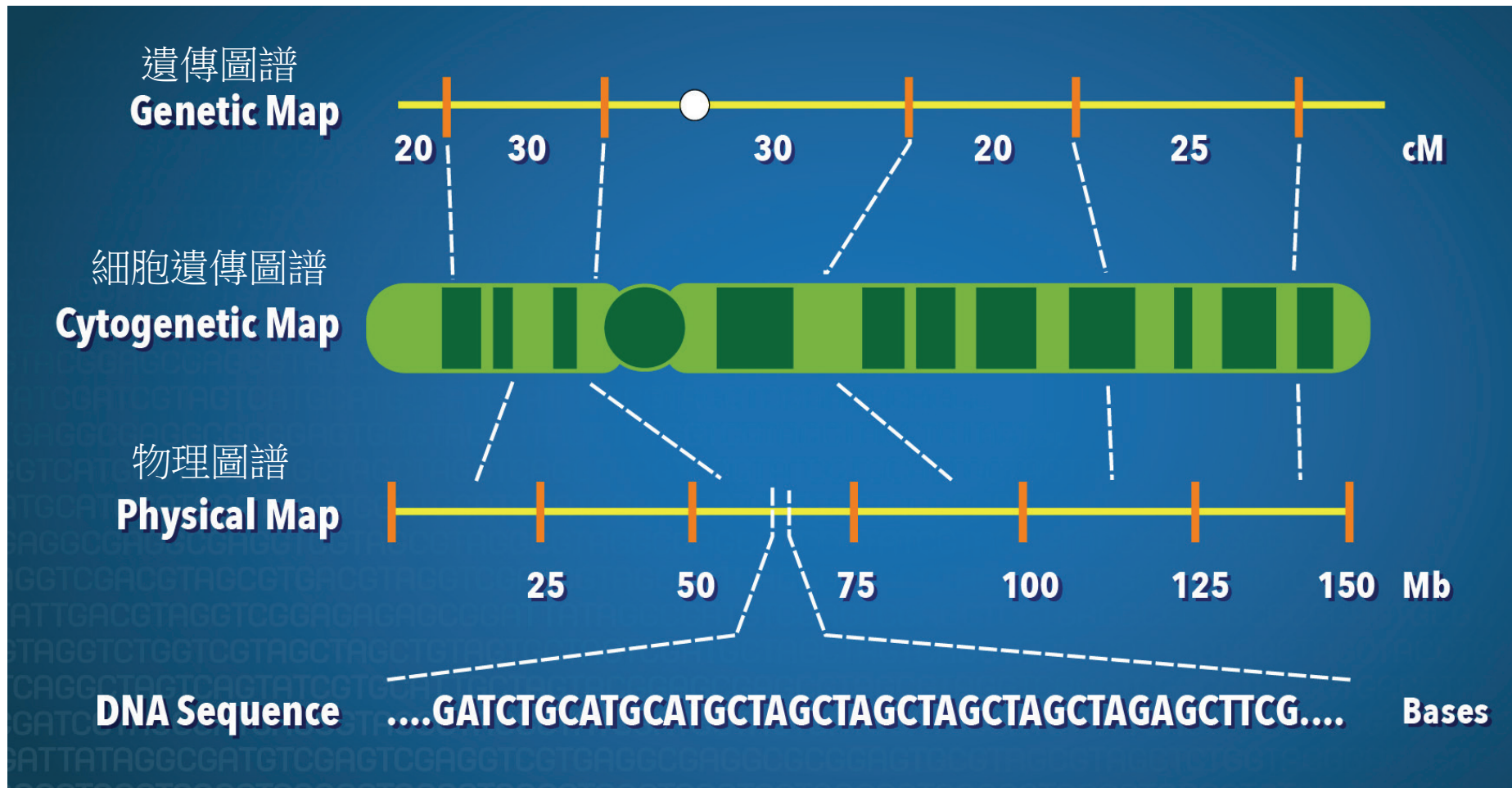


# Genome map (基因組圖譜)

Cytogenetic map: 由染色體染色而來，沒有單位，以區域劃分

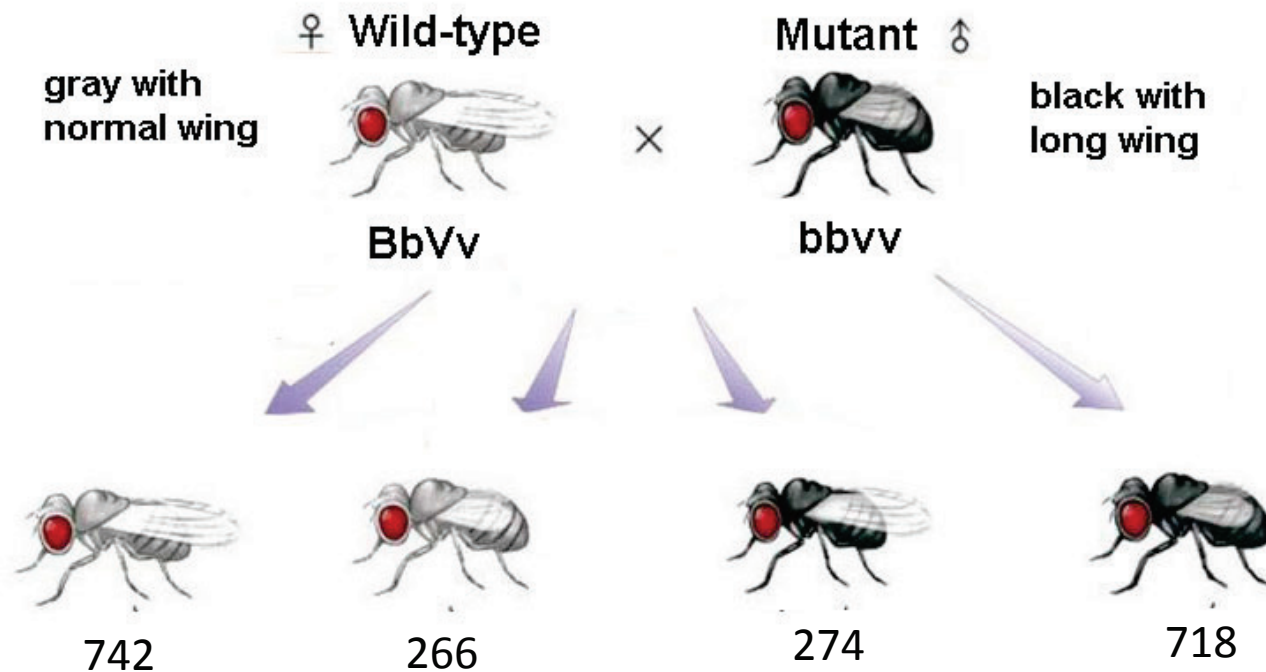
Genetic map: 由互換率計算而來，單位cM (centimorgan)

Physical map: 由序列定序而來, 單位bp

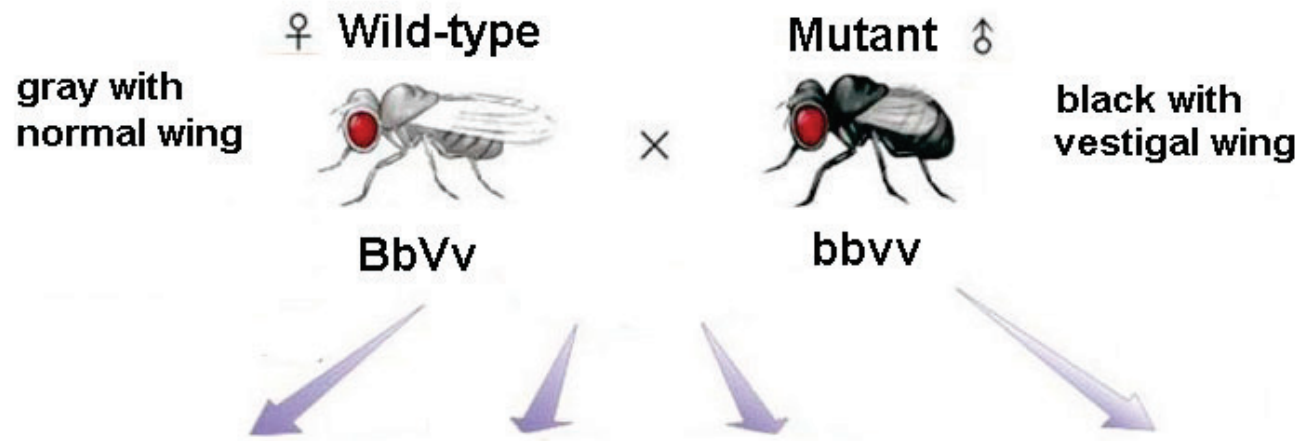


# Chromosome recombination (染色體重組)

野生型果蠅(灰身長翅)和突變果蠅(黑身短翅)交配產生2000隻後代，其中742隻為灰身長翅、266隻為灰身短翅、274隻為黑身長翅，以及718隻為黑身長翅。請問控制體色(B基因)及翅膀長度(V基因)兩基因是否為連鎖？如果是的話，請問其相距多少centimorgan?



# Chromosome recombination (染色體重組)



F<sub>1</sub> 子代



742      266      274      718

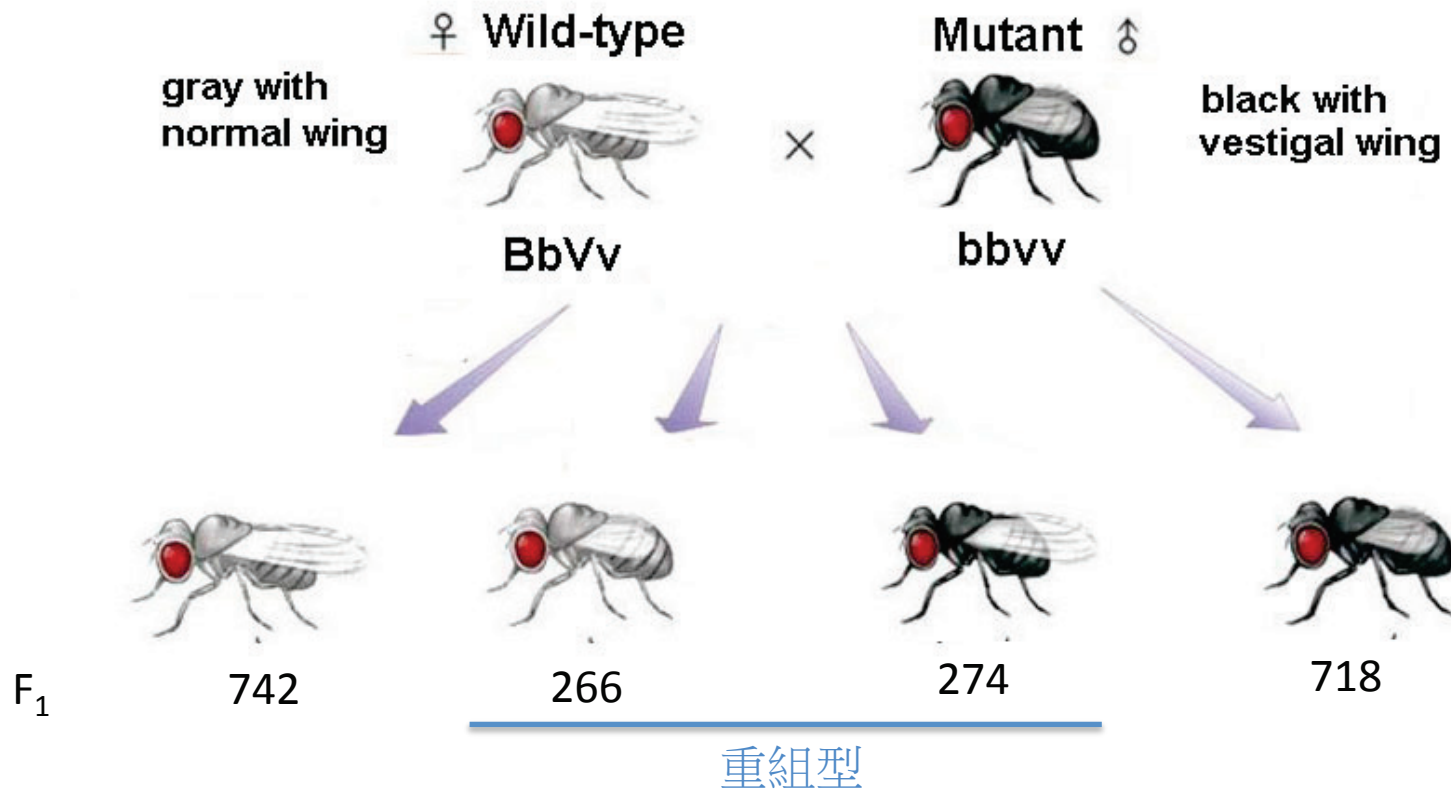
B, V 非連鎖基因  
之期望值

500      500      500      500

B, V 為完全連鎖  
之期望值

1000      0      0      1000

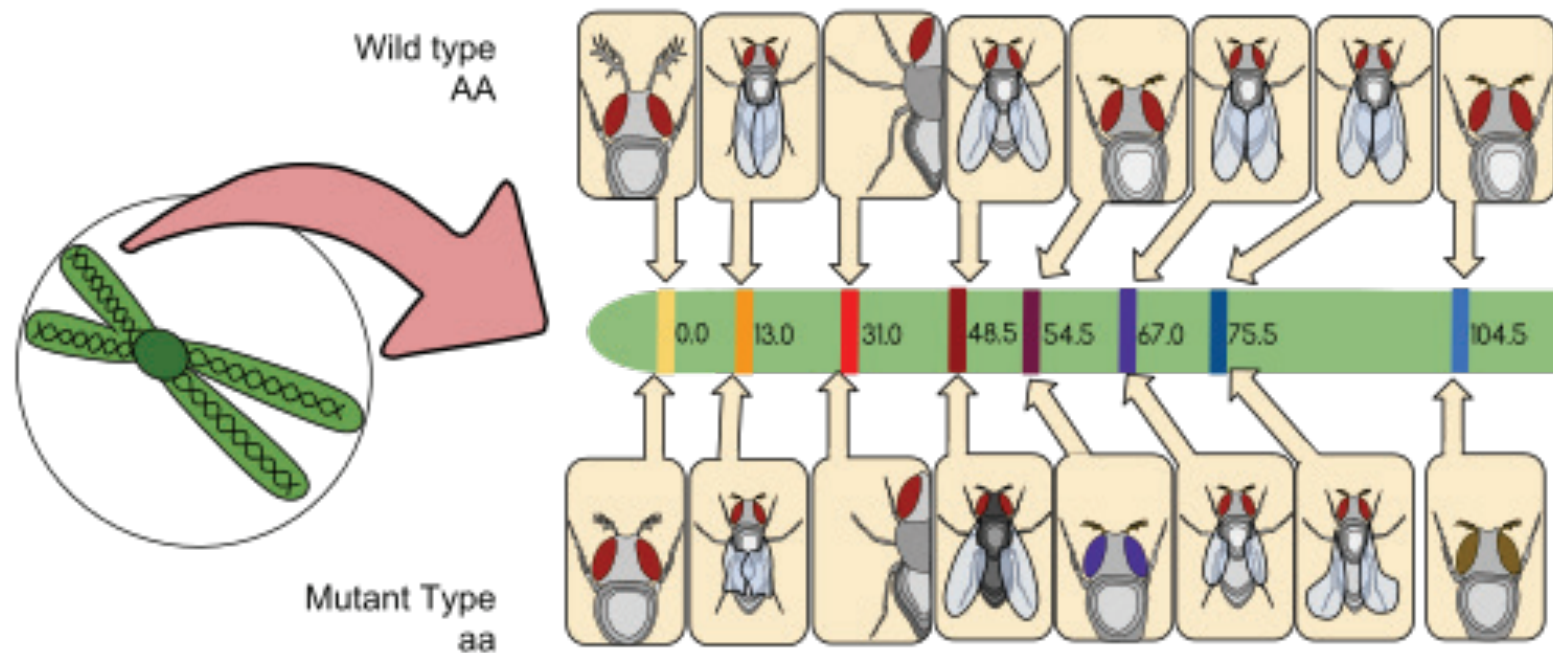
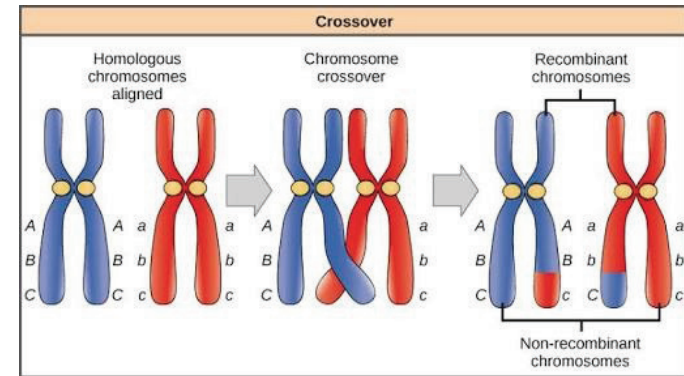
# Recombination rate (重組率)





# 遺傳圖譜

- 1% 重組率 = 1 cm (centimorgan)
- 重組率 < 50% --> "連鎖"
- 重組率 = 0% --> "完全連鎖"
- 在人類細胞中 1 cM 約 1 Mb

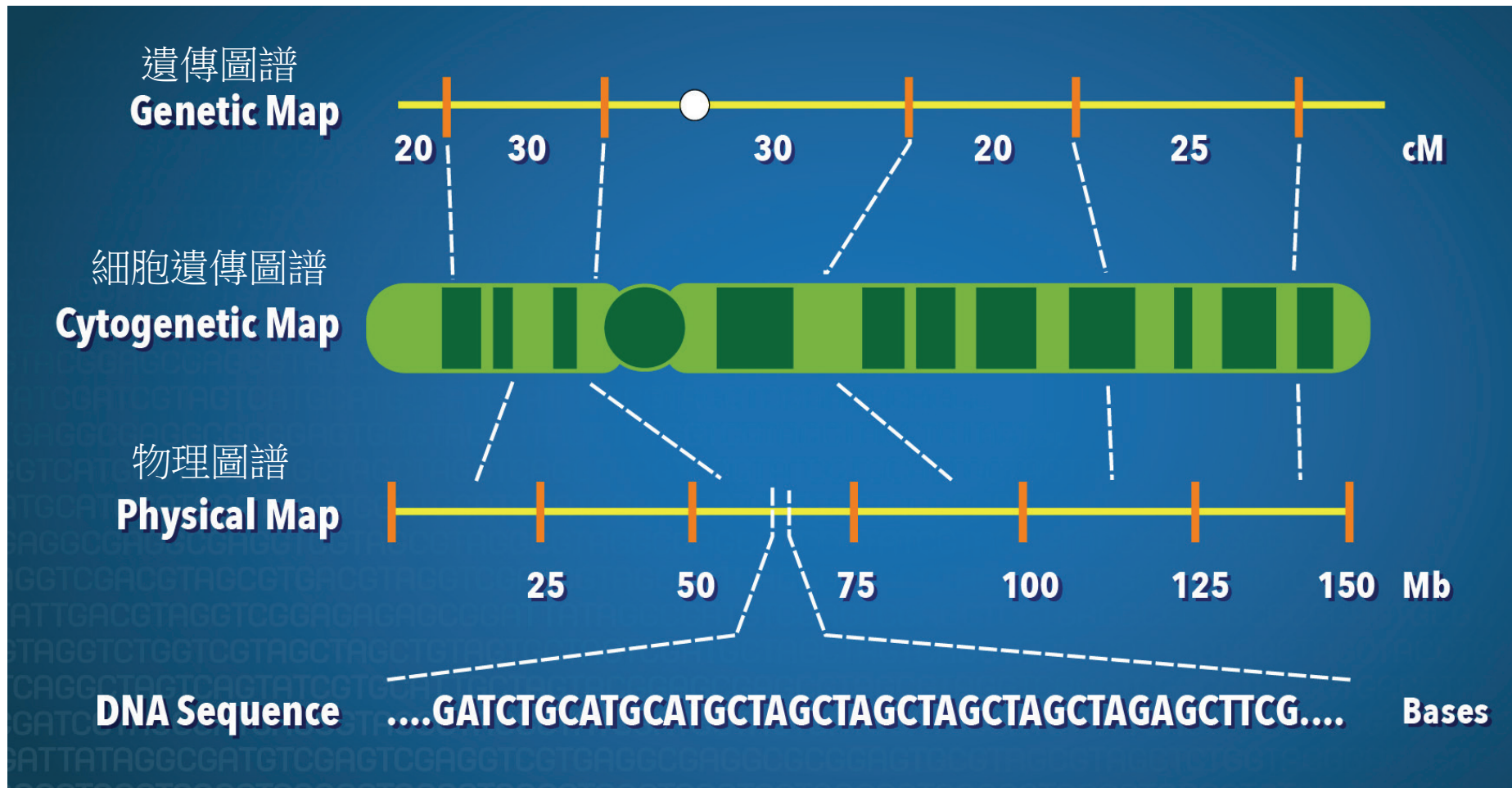


# Genome map (基因組圖譜)

Cytogenetic map: 由染色體染色而來，沒有單位，以區域劃分

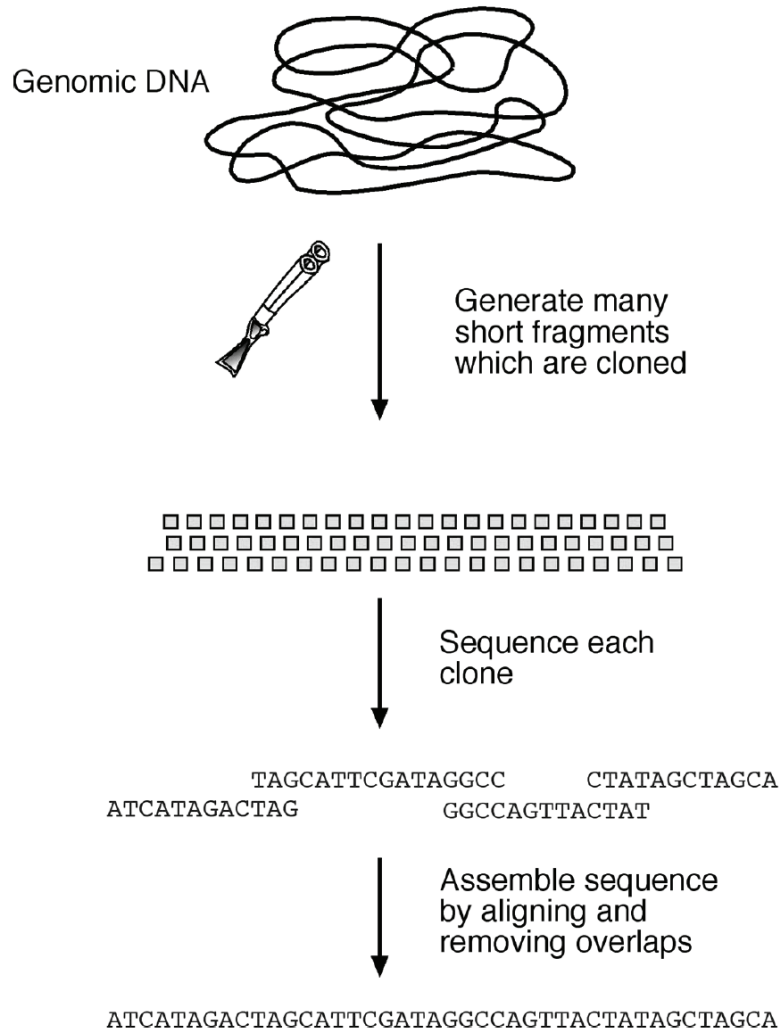
Genetic map: 由互換率計算而來，單位cM (centimorgan)

Physical map: 由序列定序而來, 單位bp

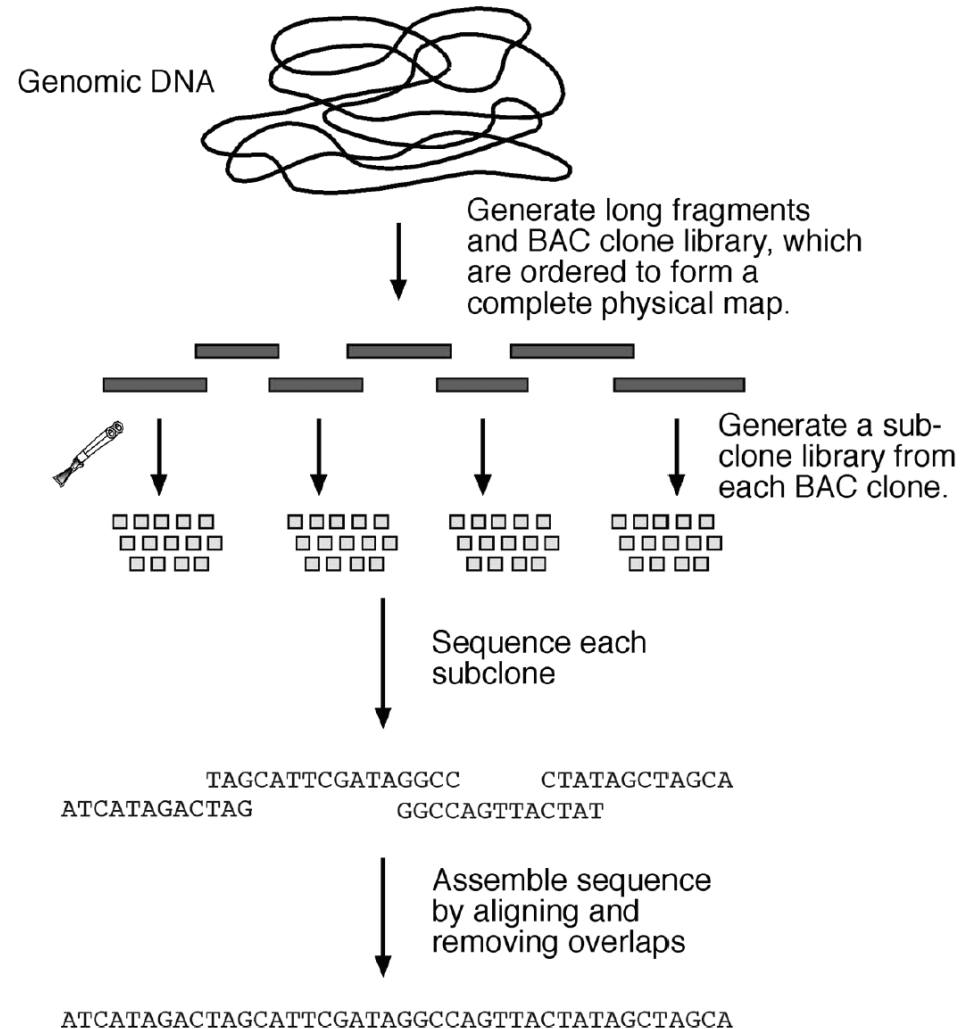


# 全基因組定序策略

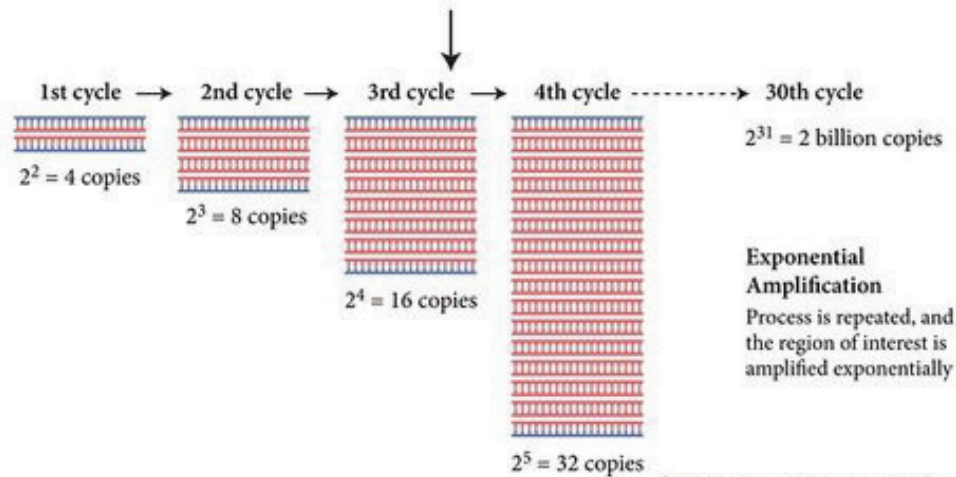
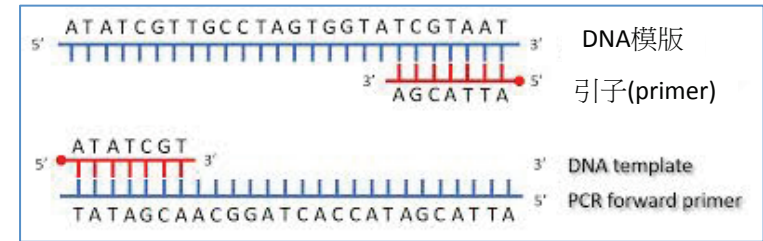
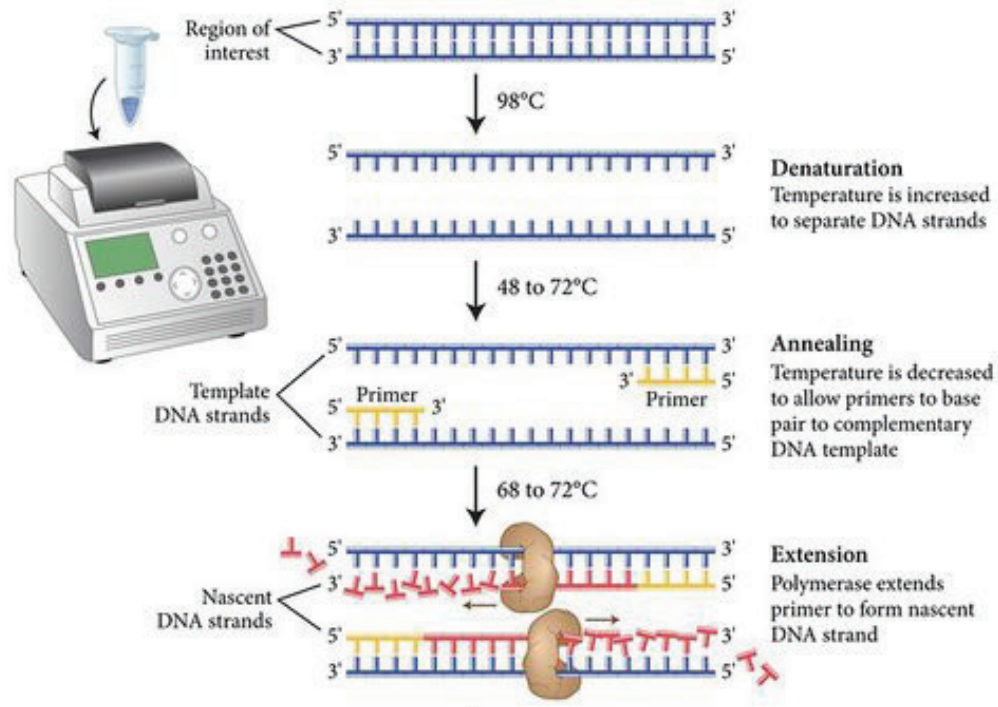
## Shotgun sequencing approach (霰彈槍定序法)



## Hierarchical sequencing approach (階層式定序法)



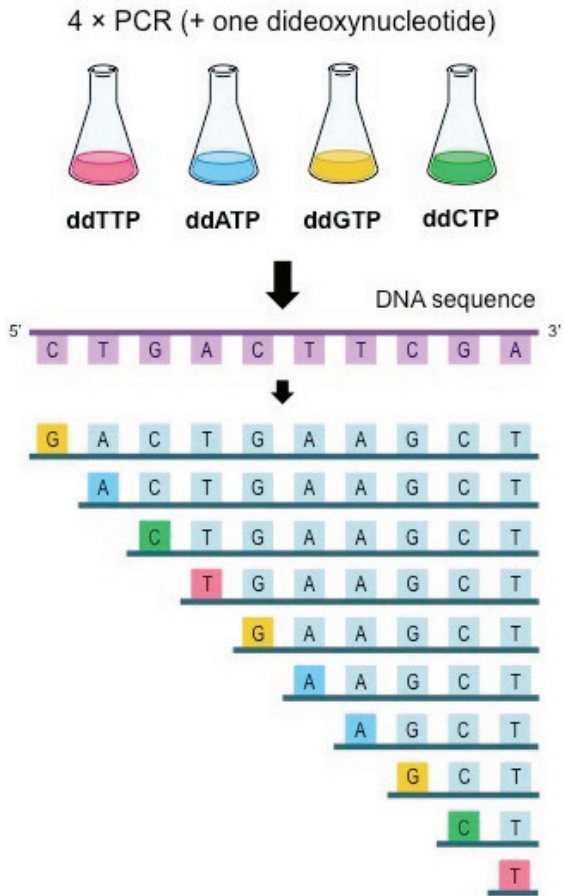
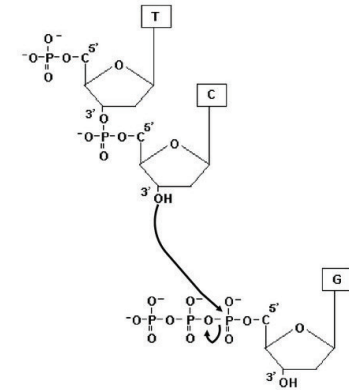
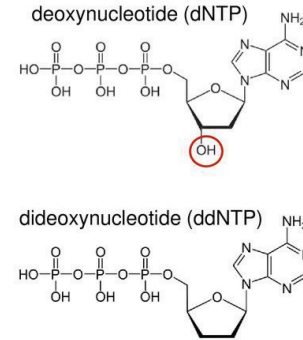
# PCR (聚合酶連鎖反應)





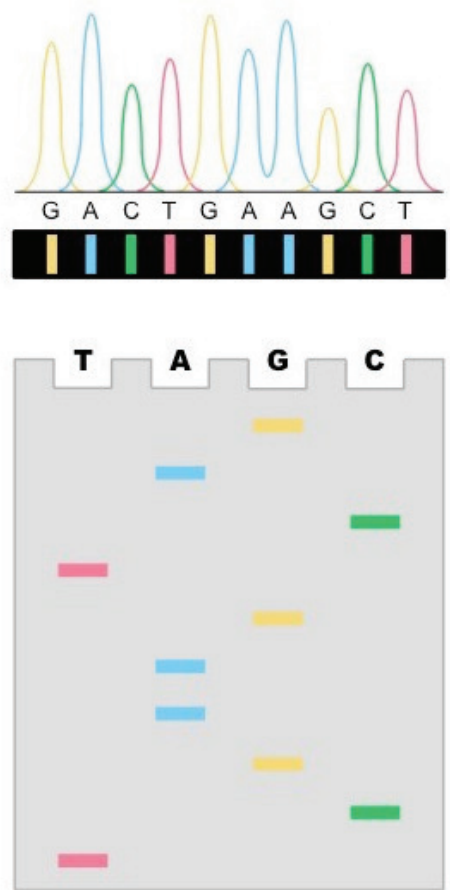
# Sanger sequencing

- 藉由螢光標定ddNTP(雙脫氧核苷酸,五碳糖缺乏3端OH基)使PCR反應停止
- 定序長度 500-1200 bp



Use a sequencing machine

Separate with a gel



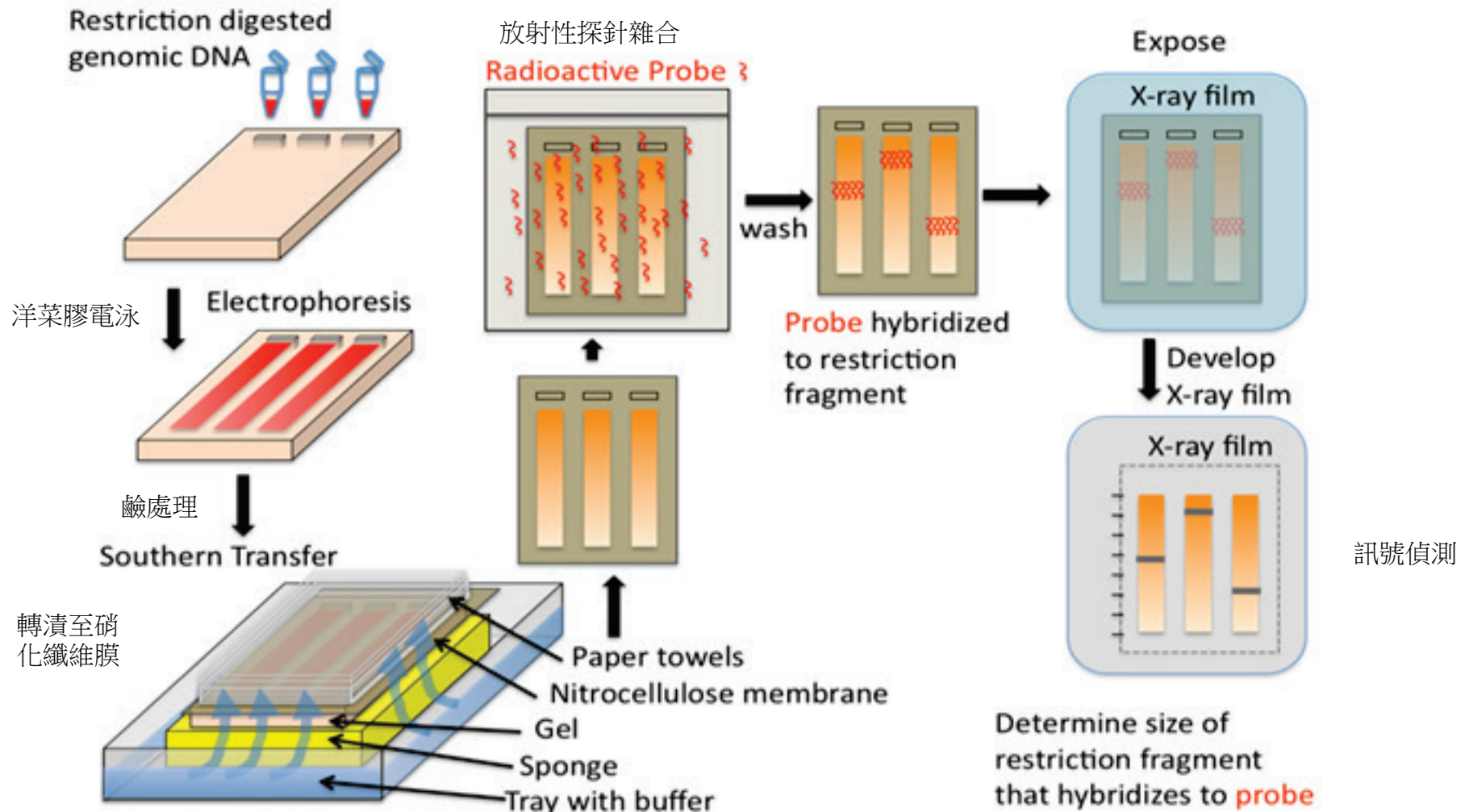
PCR

跑膠，螢光分析

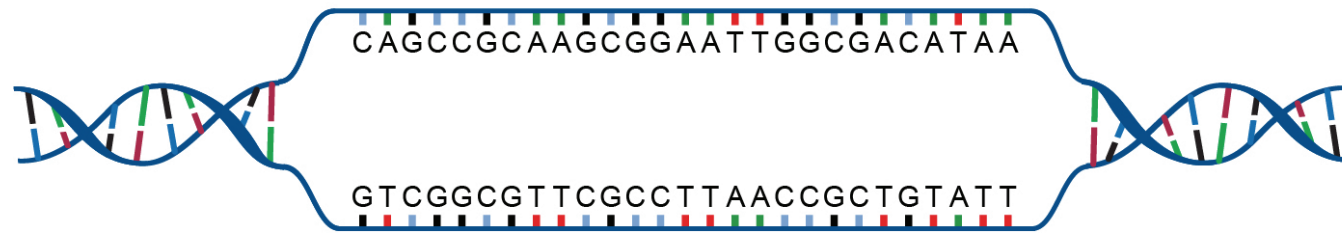


# 南方墨點法 (Southern blotting)

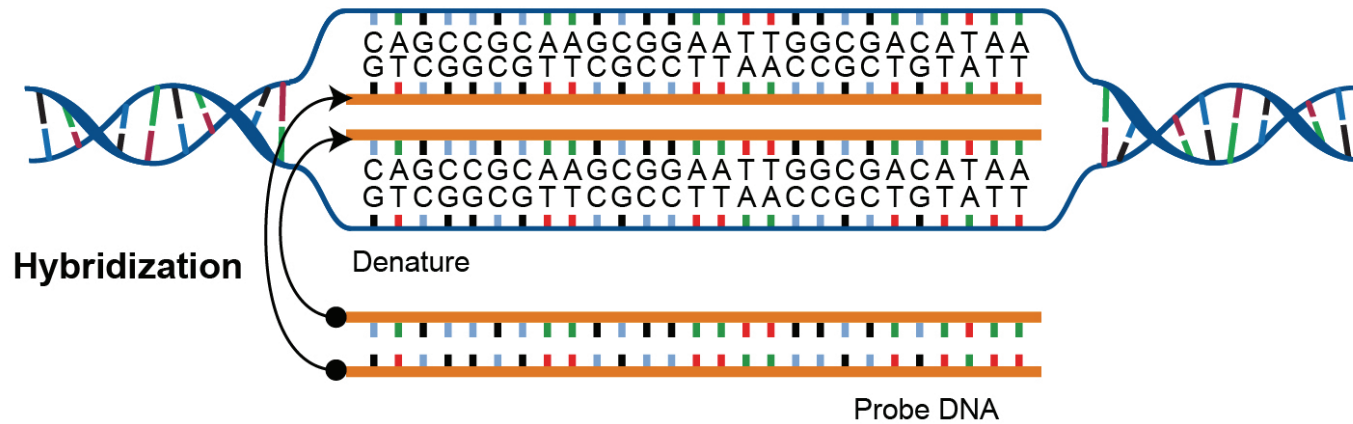
- 主要原理為單股DNA可和放射性探針(probe,單股DNA)結合
- 酵素作用後DNA -> 洋菜膠電泳 -> 鹼處理使DNA成單股 -> 轉漬至硝化纖維膜 -> 放射性探針雜合 -> 訊號偵測



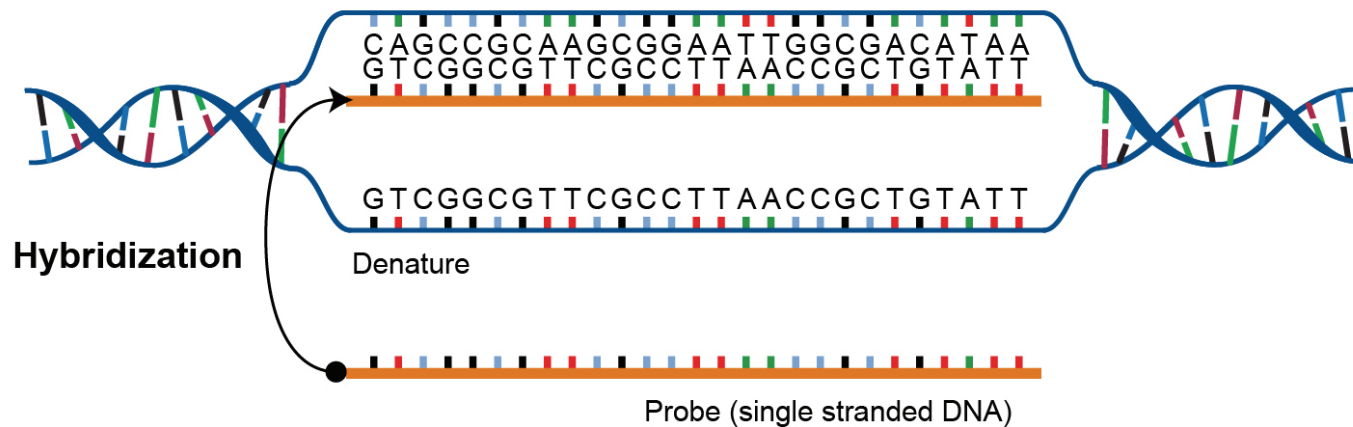
# 核酸雜合反應 (Nucleic acid hybridization)



熱處理、鹼破壞  
或其它化學溶液  
使雙股DNA變成  
單股(變性，  
denature)

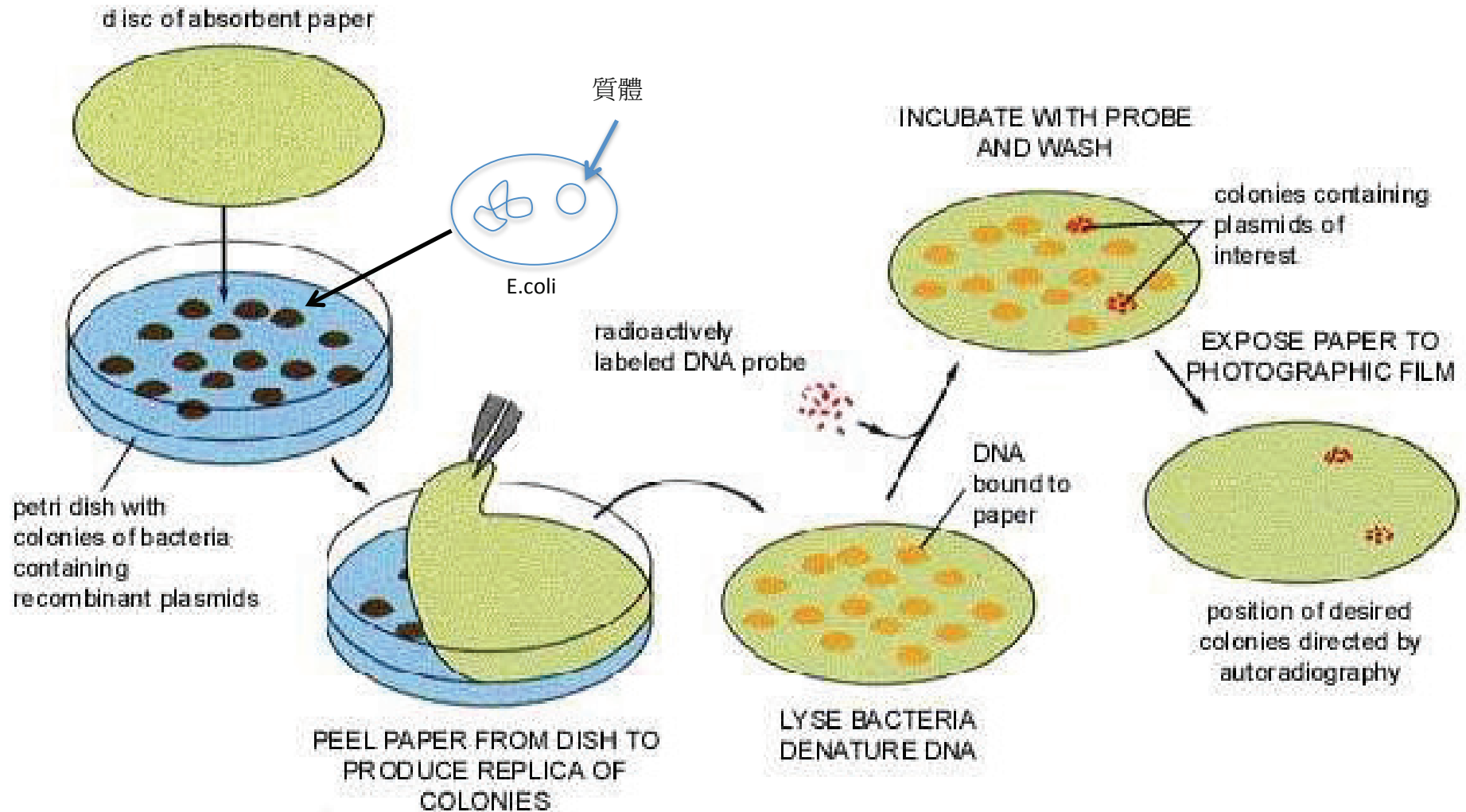


DNA探針來源若為  
雙股需先變成單股  
才能進行雜合，探  
針上可標定放射線  
或其它螢光物質



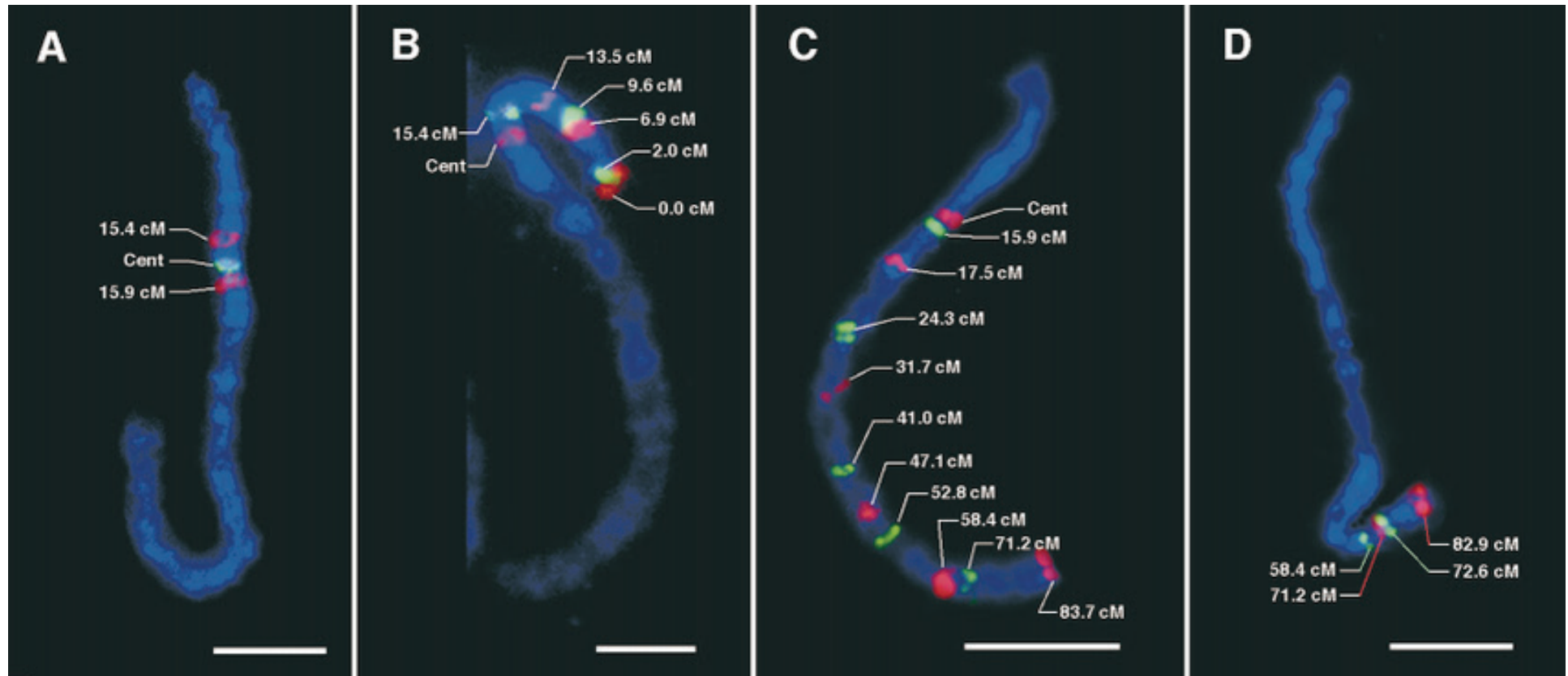
# Colonies hybridization(菌落雜交)

- 主要原理為單股DNA可和放射性探針(probe,單股DNA)結合
- 目的為尋找含有特定序列的菌落
- 培養菌落 → 拓印至硝化纖維膜 → 鹼破壞打破細胞並使DNA變性 → 放射性探針雜合 → 訊號偵測



# 螢光原位雜合FISH (Fluorescence in situ hybridization)

- 主要原理為單股DNA可和螢光標定探針(probe,單股DNA)結合
- 目的為確認目標序列在染色體上的位置
- 細胞固定於玻片 → 以formamide將染色體變性 → 螢光標定探針雜合 → 螢光顯微鏡觀察

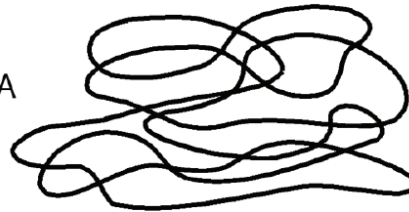


# 霰彈槍定序法 (Shotgun sequencing approach)

## Shotgun Sequencing Approach

基因組DNA

Genomic DNA



Generate many short fragments which are cloned

以物理、化學或酵素法將DNA打斷 ~500 bp



Sequence each clone

次世代定序

TAGCATTCGATAGGCC      CTATAGCTAGCA  
ATCATAGACTAG      GGCCAGTTACTAT

Assemble sequence by aligning and removing overlaps

基因組序列組裝

ATCATAGACTAGCATTCGATAGGCCAGTTACTATAGCTAGCA

## DNA定序 (DNA sequencing)

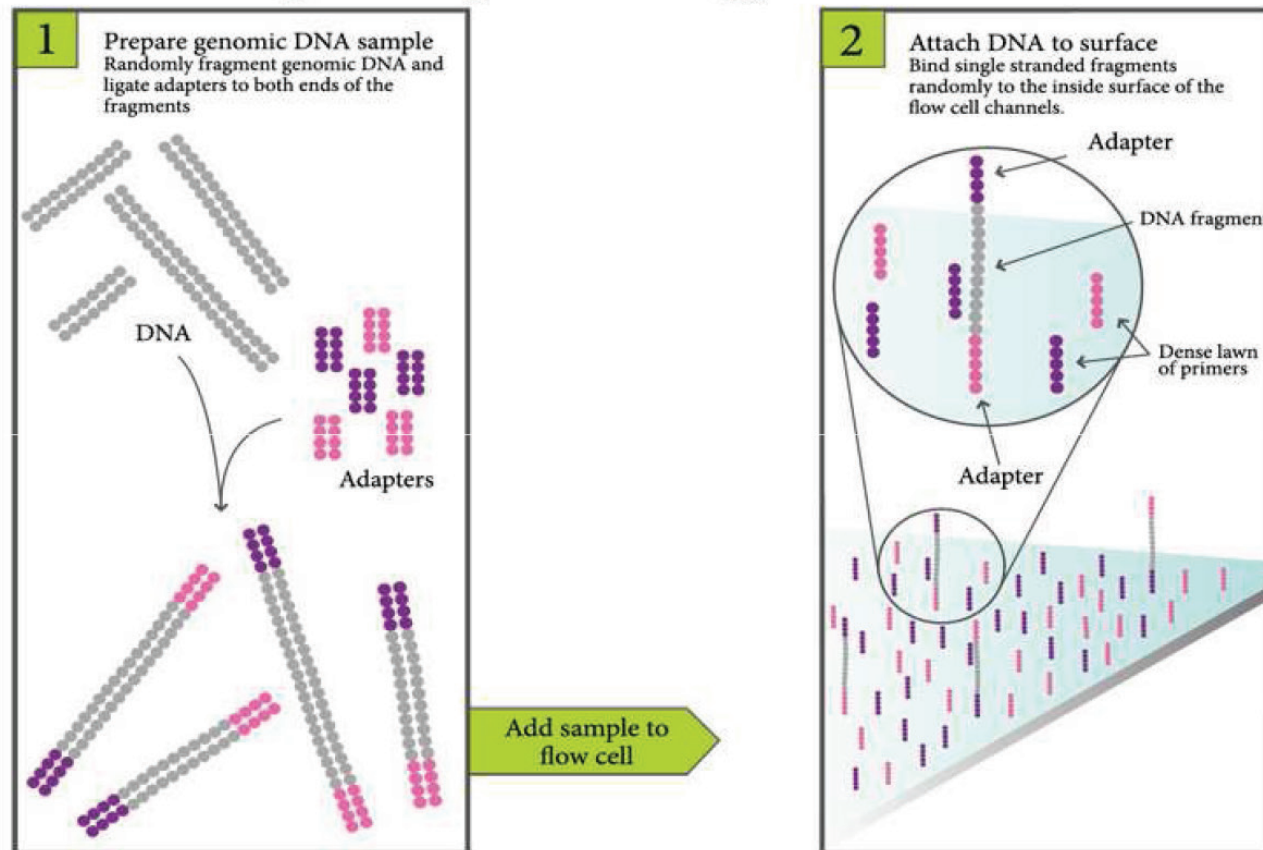
- 傳統定序法 Sanger sequencing
  - 一次定序長度 ~ 500bp - 1 kb
  - 每次上機 96 sample
  - 1977年發明
- Next generation sequencing (次世代定序, NGS)
  - 又稱**大規模並行測序** (Massive parallel sequencing)
  - ~ 100 - 500 bp /read (依定序技術不同)
  - 每次上機可獲得 > 100 Gb資料量
  - 不同公司技術不同，目前以illumina開發的SBS技術為主流  
(2006年發明)
  - 可用於基因體、轉錄體、基因甲基化及環境微生物定序



# SBS

- 將Genomic DNA打斷為約 ~500 bp, 並接上adapter
- 將DNA變性並固定在含有引子的定序盤
- 引子的序列為根據adapter設計

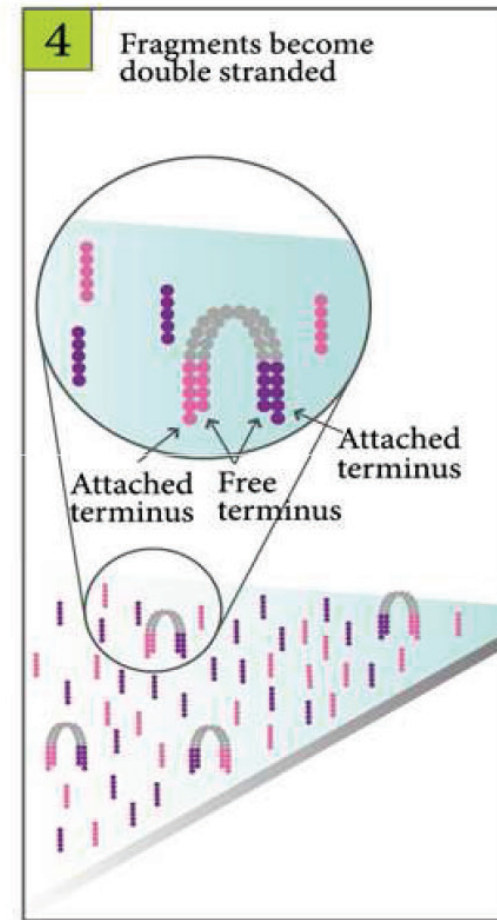
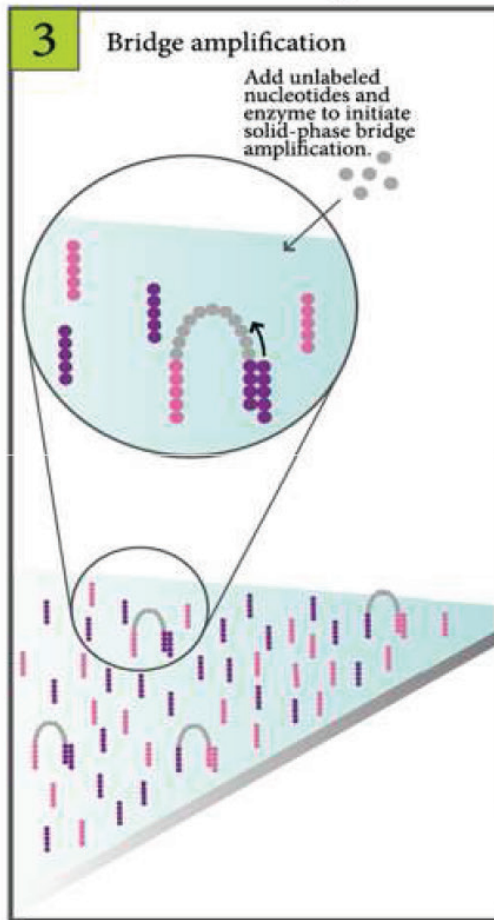
## Sequencing Technology Overview



# SBS

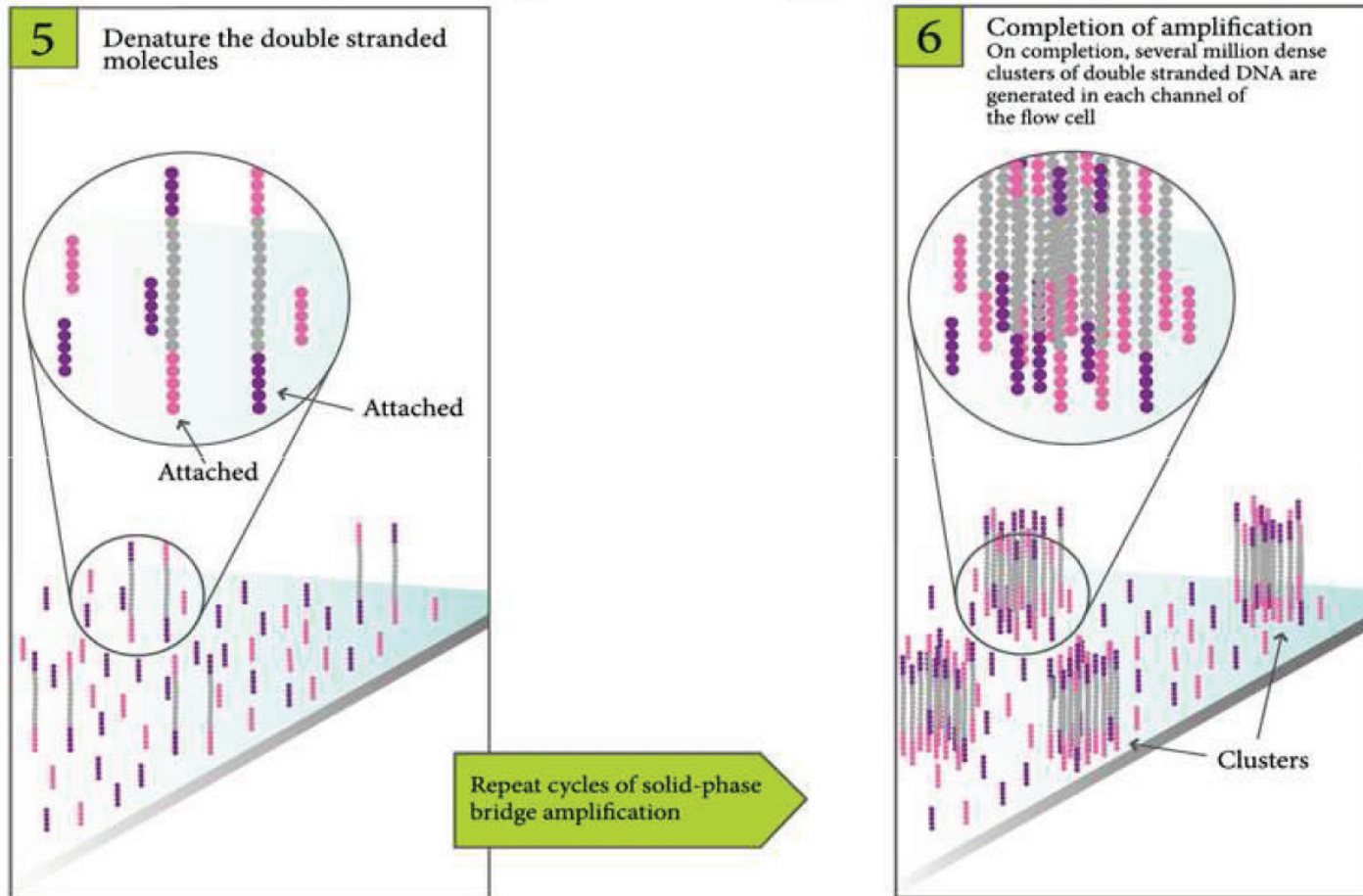
- 利用引子進行PCR，將單股的標的序列複製為雙股

## Sequencing Technology Overview



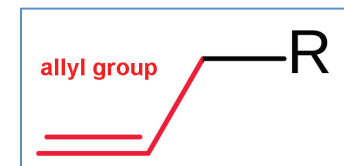
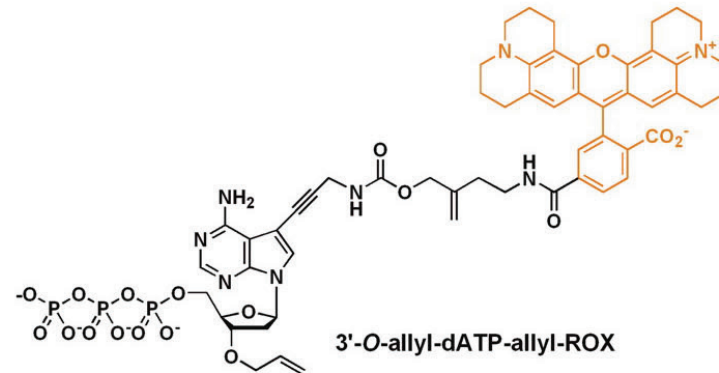
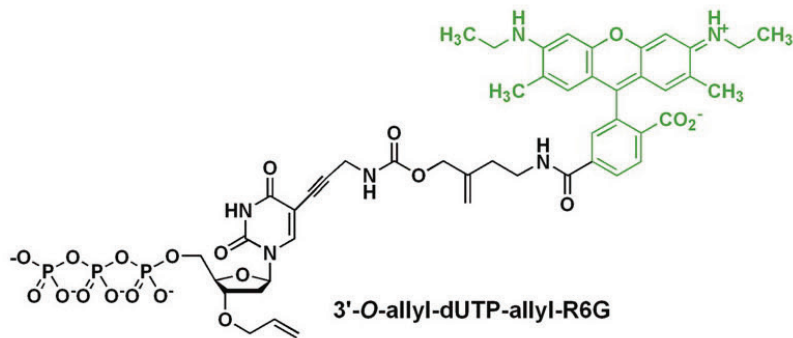
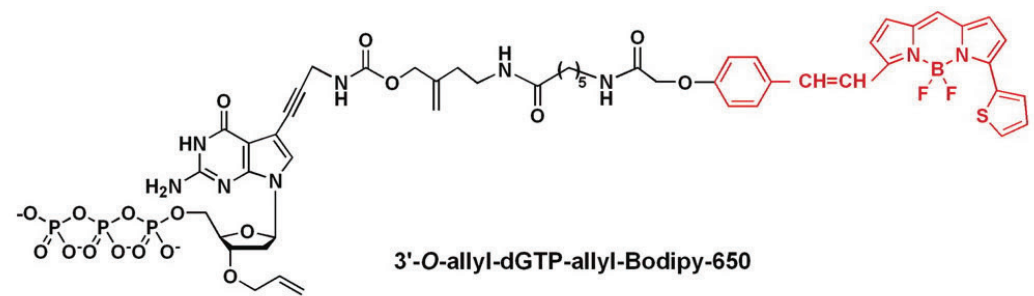
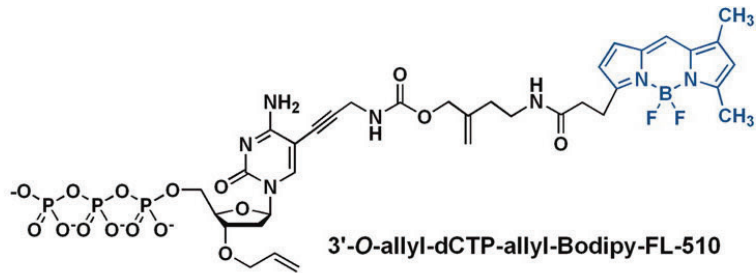
- 將複製的雙股變性成單股，並重覆PCR步驟，序列數目將以2次方成長
- 複製的目的為擴大定序訊號

## Sequencing Technology Overview

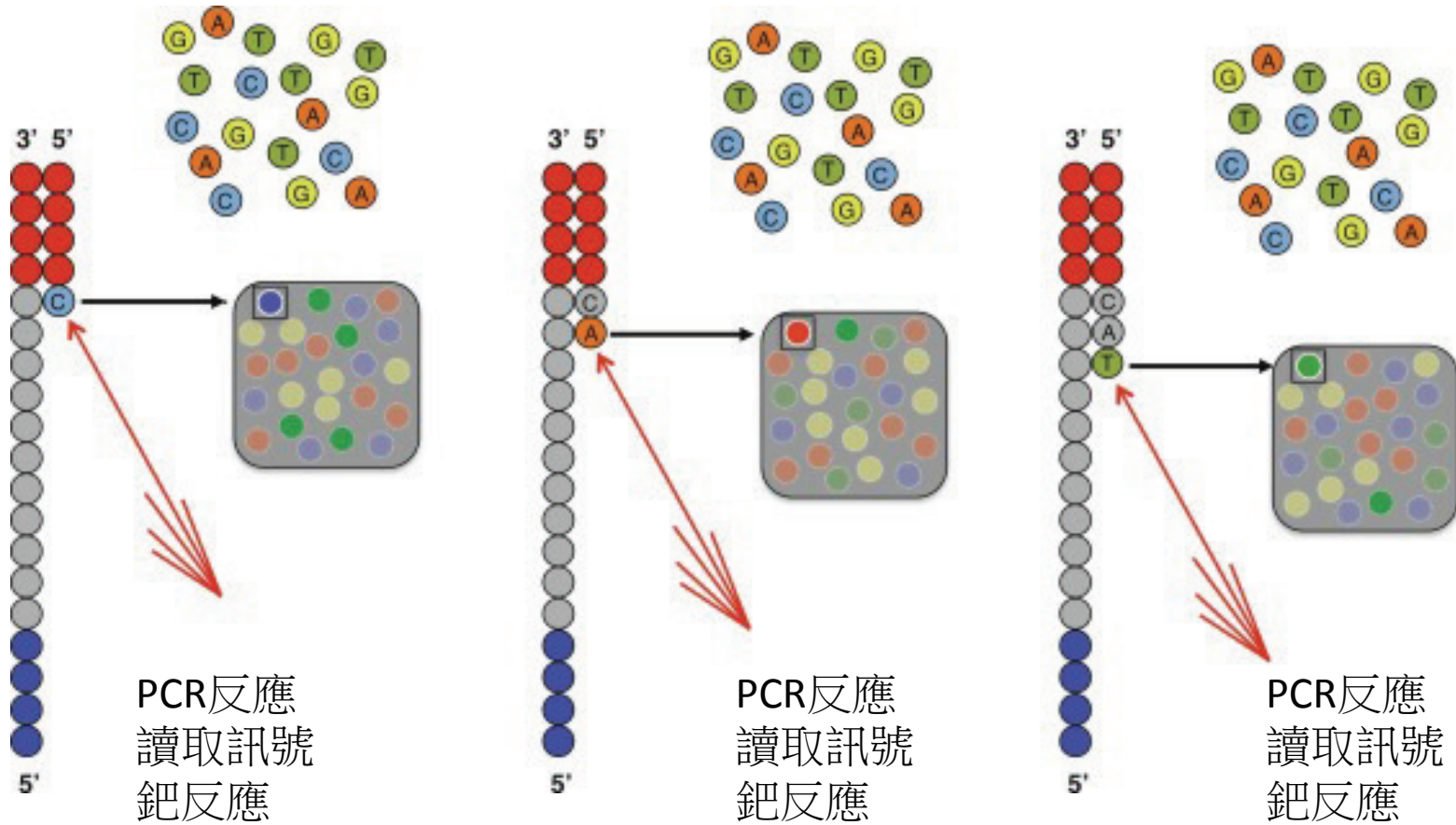


# SBS — 反應用特殊核苷酸

- SBS所使用的核苷酸為特殊鹼基，上頭帶有發色基，可在激光後發出不同顏色
- 核苷酸五碳糖3端上帶有烯丙基(allyl group), 可使PCR反應無法進行
- 發色基及五碳糖3端的烯丙基可以鈦(pd)反應將其移除，使PCR反應繼續

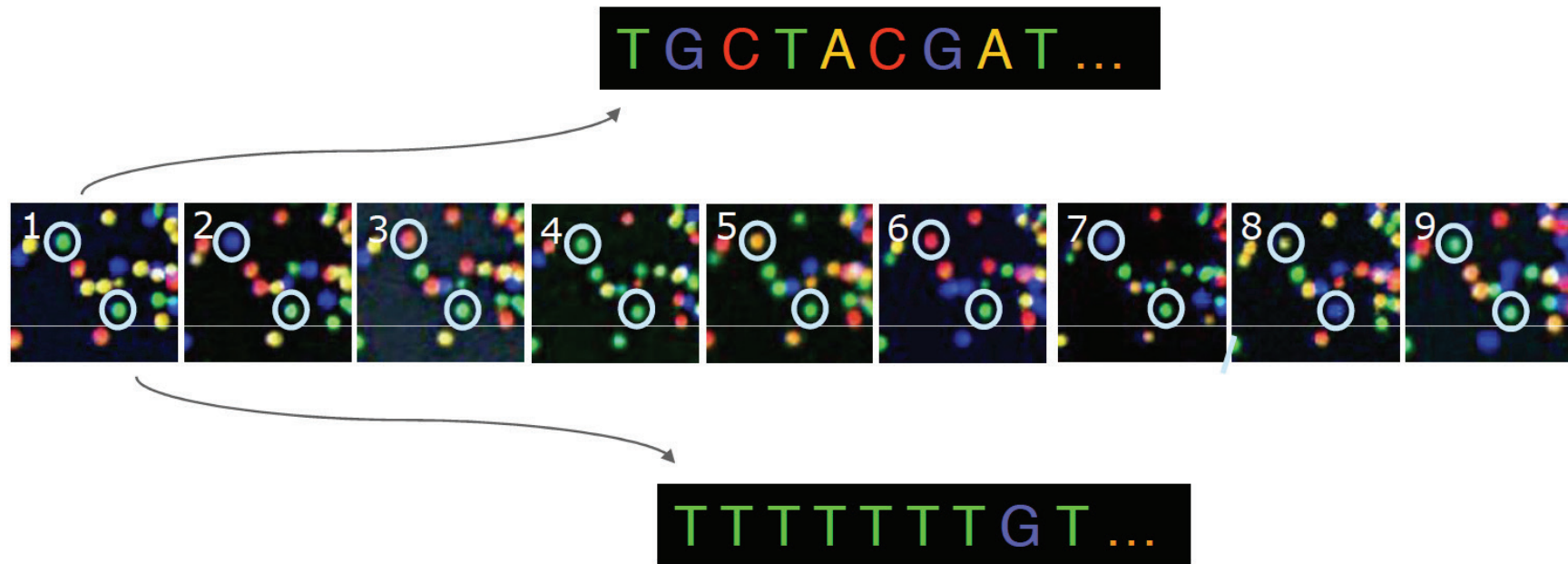


# SBS – PCR 反應及訊號讀取





# SBS—訊號解讀 (base calling)



The identity of each base of a cluster is read off from sequential images.



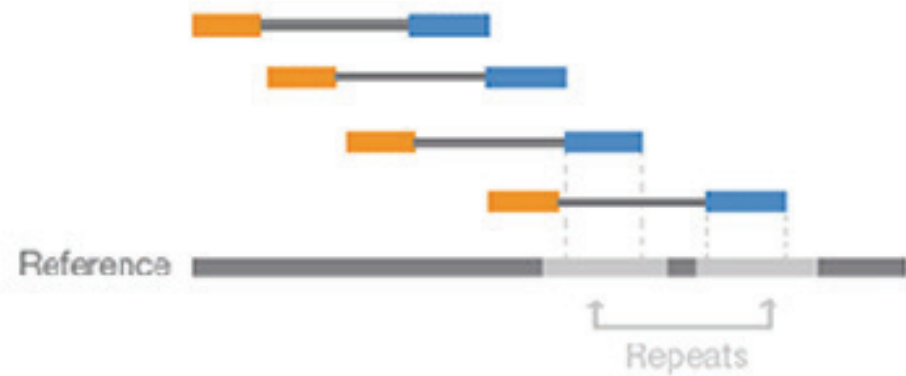
# Genome assembly (基因組序列組裝)

- 定序時目標序列為 ~500 bp，每次由其兩端讀取 150 bp，這些定序小片段稱之為read
- 序列的中間部份為未被定序區域
- 同一目標序列上會有二條read，可幫助序列組裝正確性
- 組裝原理為具有重覆序列的片段即可能位在基因體同一位置

Paired-End Reads

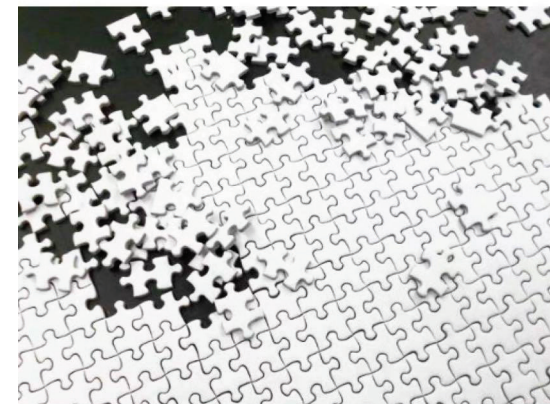
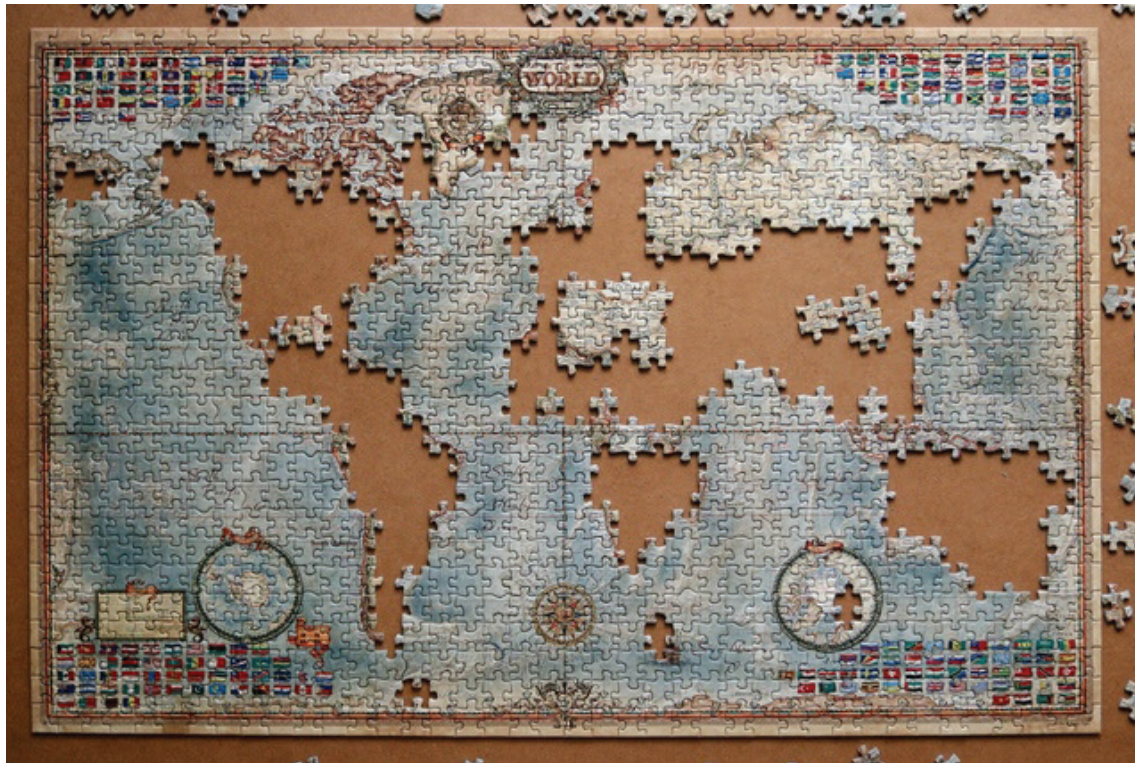


Alignment to the Reference Sequence



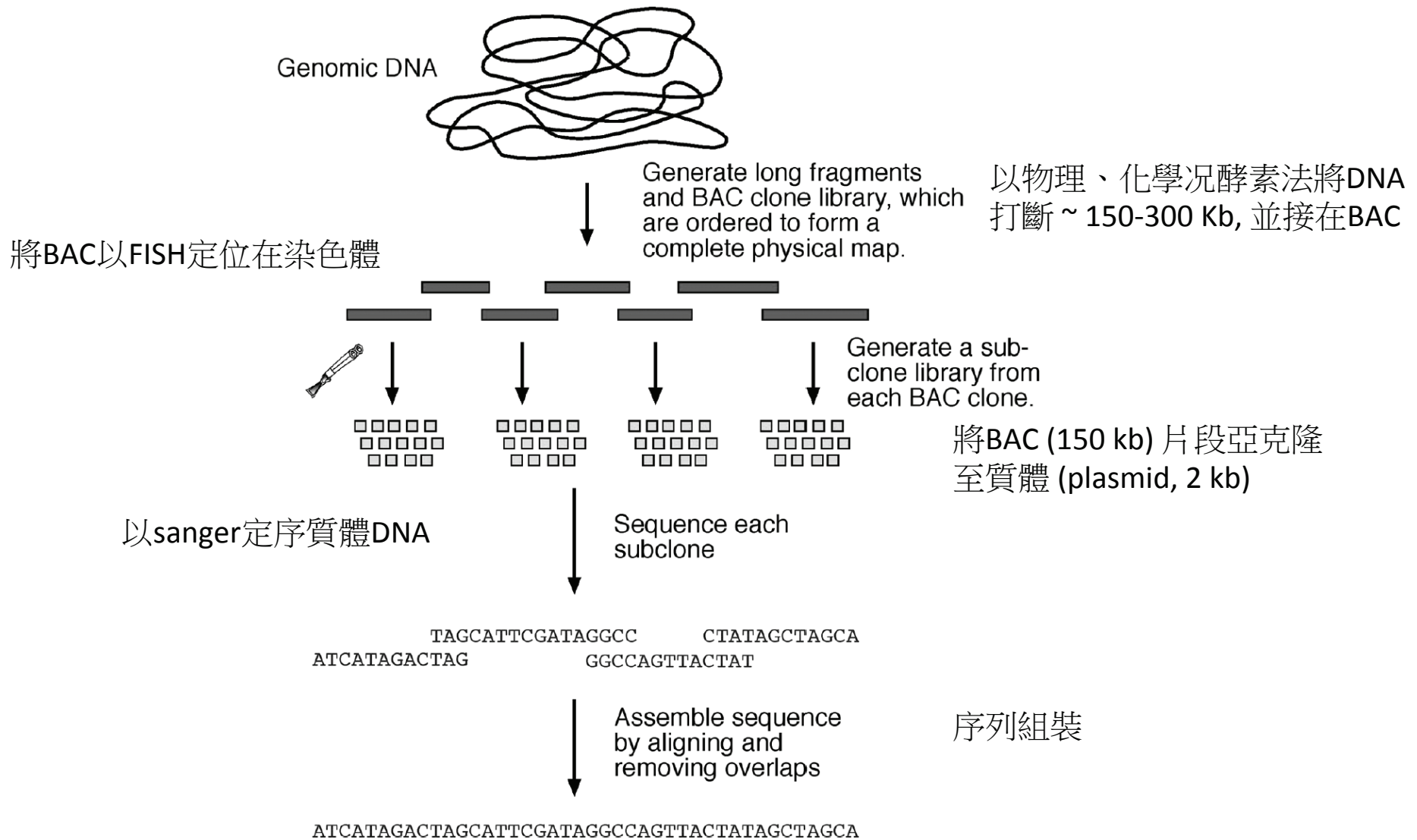
# 定序組裝(assembly)的難題

- 人類基因組為3G(30億個鹼基)，假設定序時每次只能讀 150 bp，那麼至少要有2百萬條reads，若考慮重疊性，那至少要六百萬條reads才能完整涵蓋基因組。
- 每條序列都如同一塊小拼圖，如何把六百萬條序列完整地對到染色體？尤其是染色體不同位置但序列卻即為相似或序列多次重覆之位置。  
--> 將染色體分為大片段，並得知每大片段是由染色體那一區域而來



# Hierarchical sequencing approach (階層式定序法)

## Hierarchical Sequencing Approach



# DNA 載體

載體	承載量	宿主細胞
人類人造染體 (HAC)	6000 - 10000 Kb	human cell
酵母人造染色體 (YAC)	100 - 3000 kb	Yeast
細菌人造染色體(BAC)	150 ~ 350 kb	E. Coli
噬菌體載體 (PAC)	100- 300 kb	E. Coli
黏質體 (Cosmid, 噬菌體載體/質體之複合體)	35-45 kb	E. Coli
質體 (plasmid)	<= 15kb	E. Coli

如果以一倍的覆蓋率計算，人類基因組 (3,200,000 kb) 需量

320 HAC

1,066 YAC

9,142 BAC

10,666 PAC

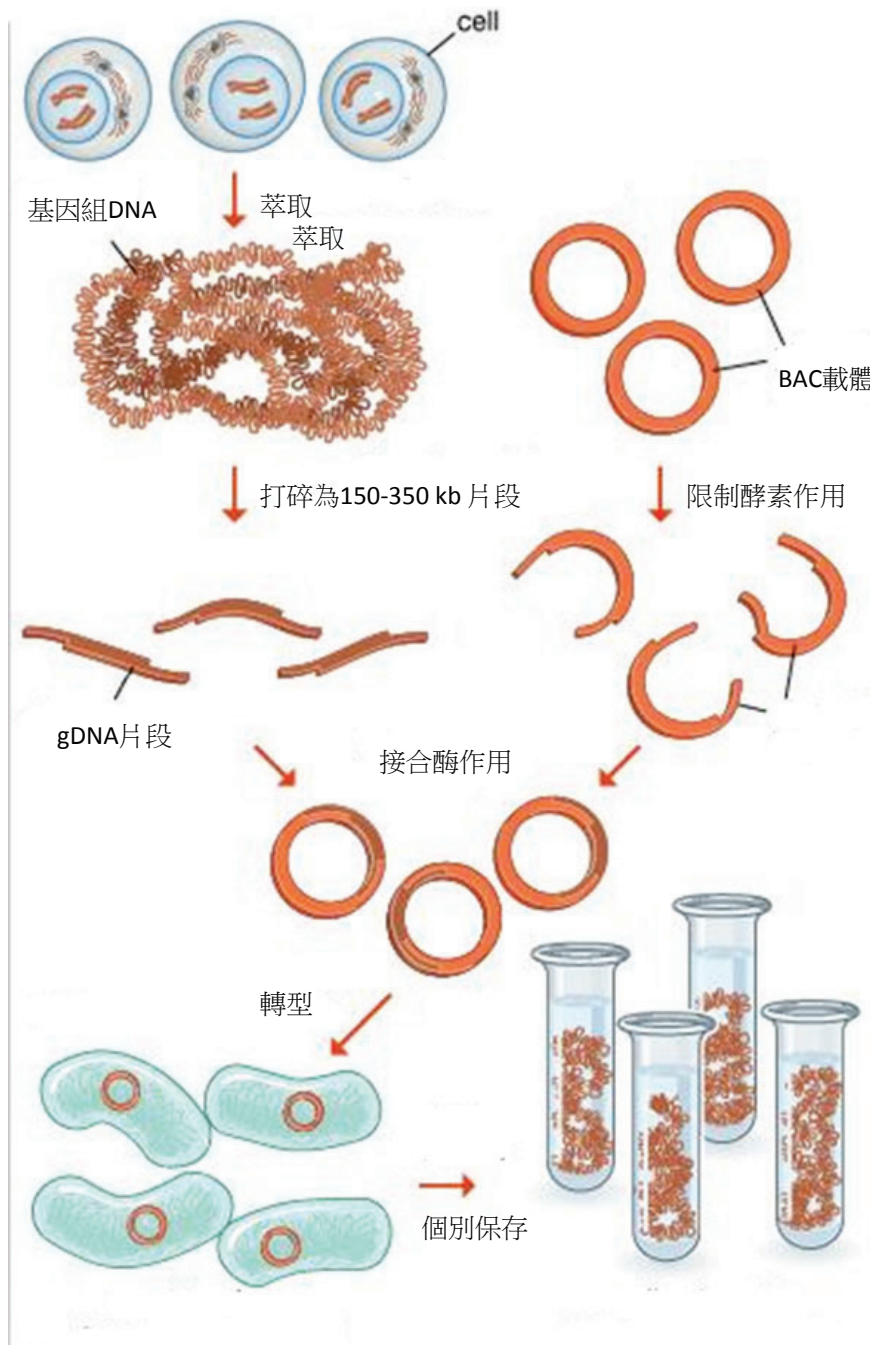
71,111 Cosmid

213,333 plasmid

\* 通常一個基因組庫的要求為6倍覆蓋率以上

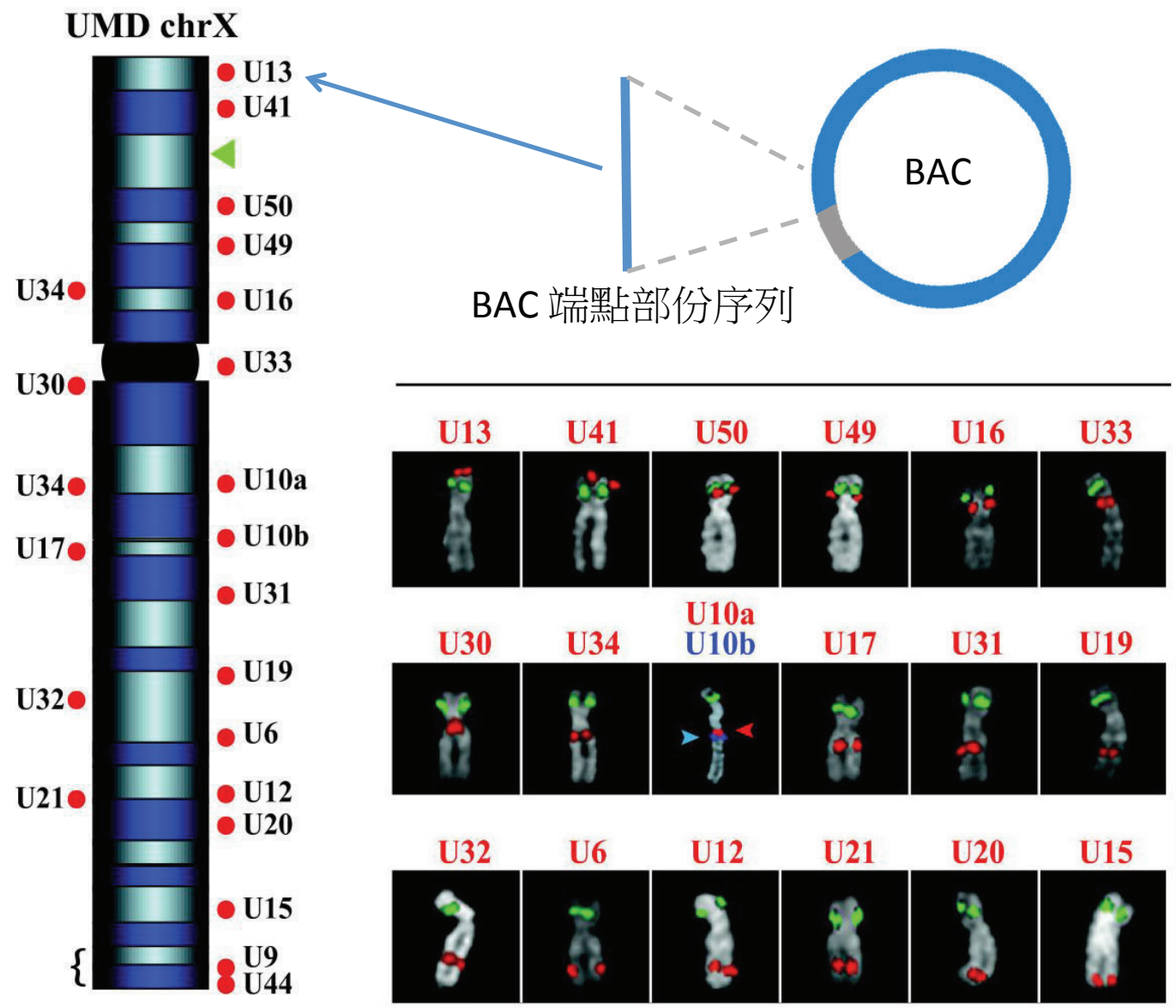


# 基因組庫建構 (Genomic Library Construction)



如果一個BAC可承載150kb,對於人類基因組3.2 Gb而言，9,142個BAC為一倍的覆蓋率，而若要達到90%以上的覆蓋率，至少要6倍的BAC數目(54,852個)

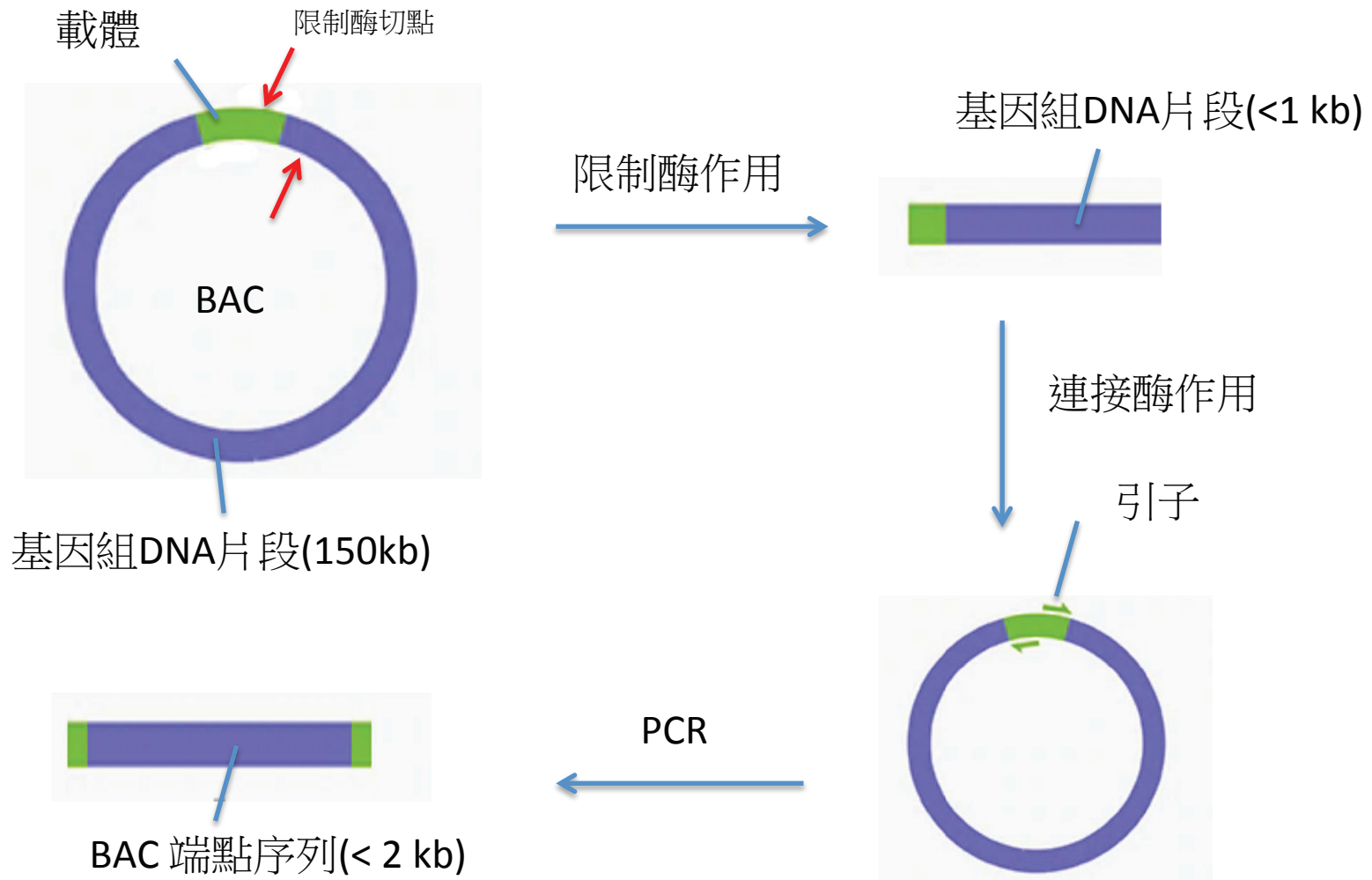
# 以螢光原位雜合法(FISH)將BAC定位在染色體上定位



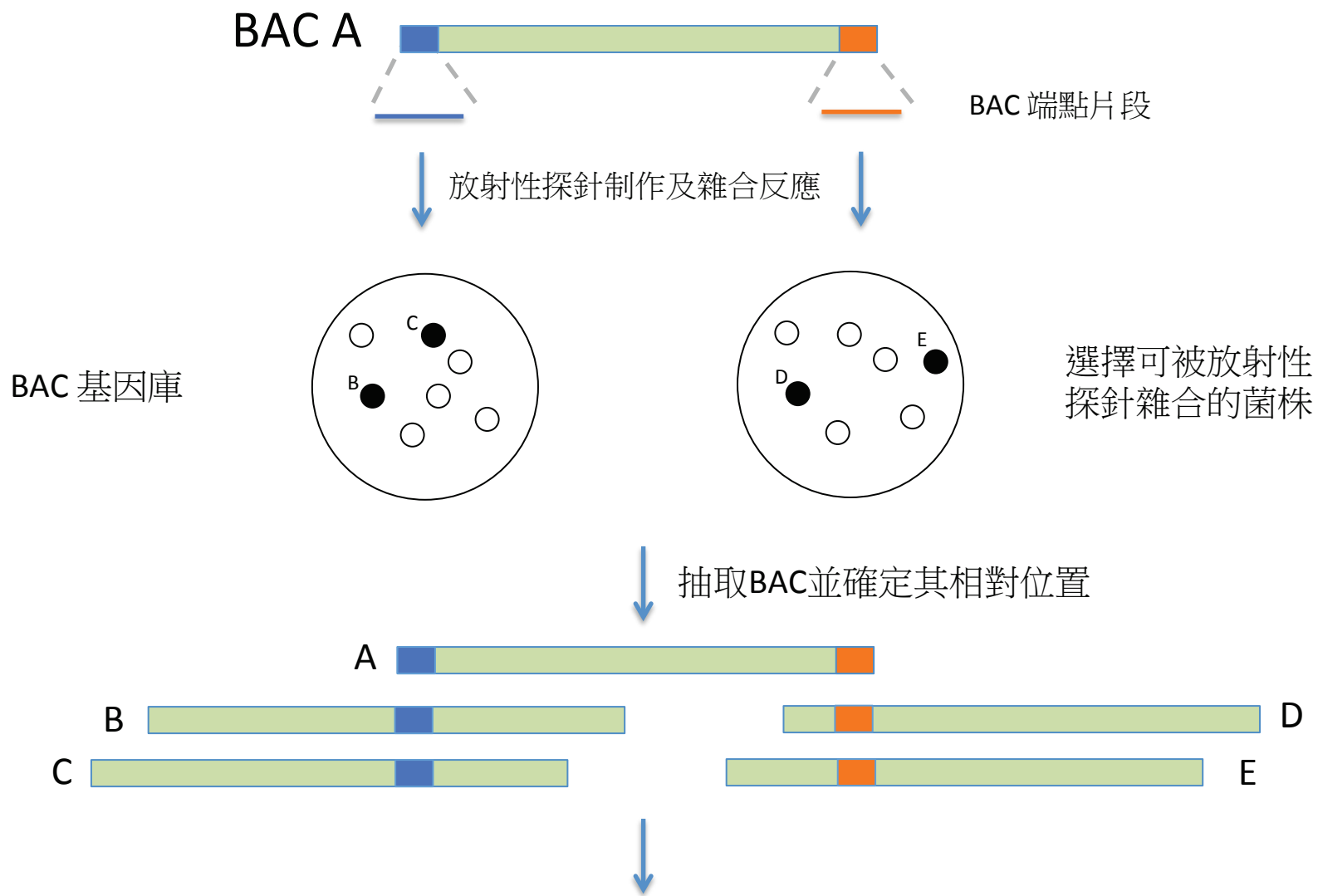


# Inverse PCR (反向PCR)

- 最好聚合酶最長可做到30kb，但很貴，成功率也很低。大部份聚合酶僅能做到 ~ 2 kb
- 基因組DNA片段為未知，無法設計引子。引子只能設計在載體序列
- 目的為擴增BAC端序列，以進行FISH或BAC library screening

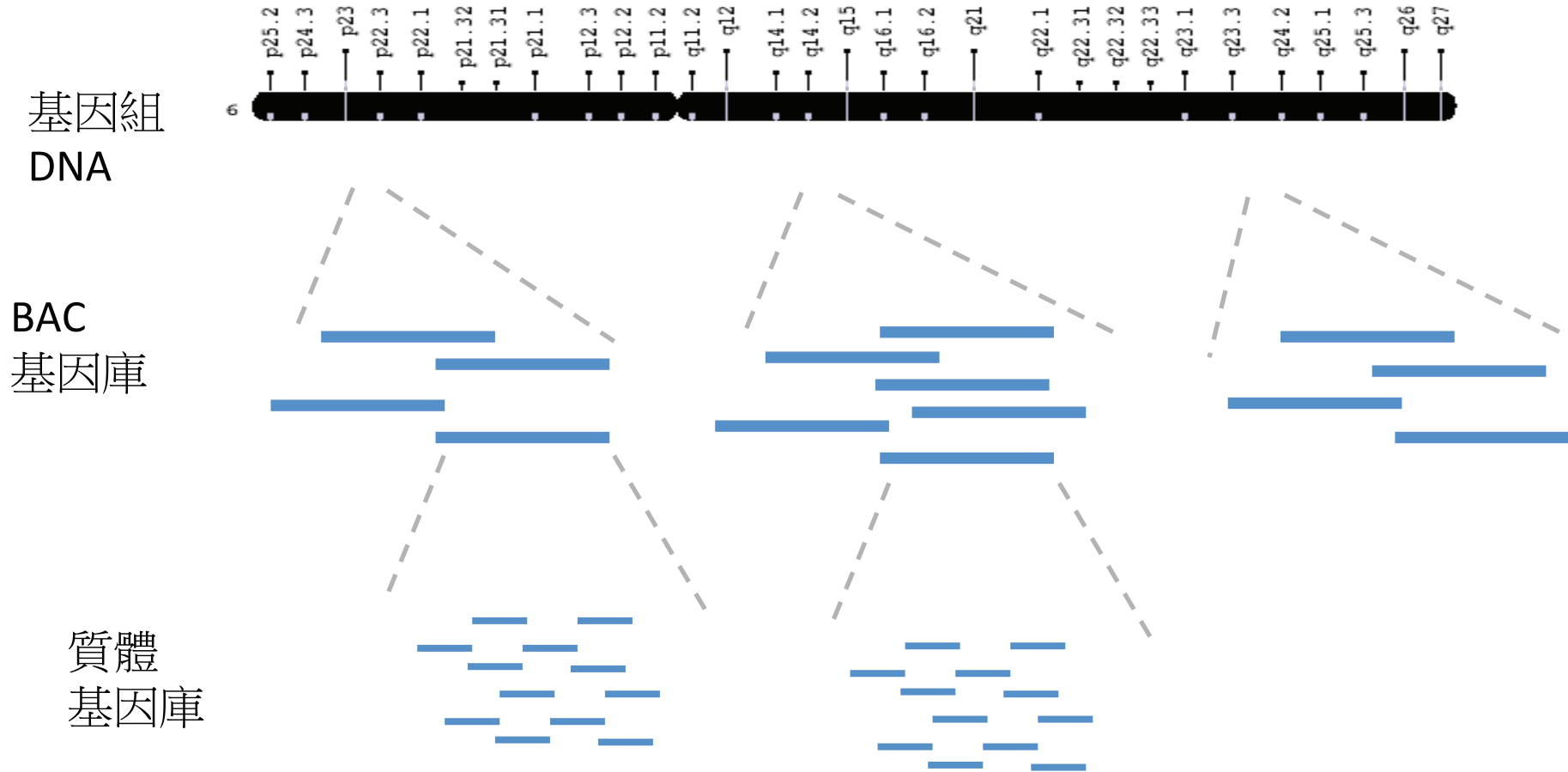


# BAC基因庫篩選



以BAC B, C, D, E端點片段重覆以上實驗以找到上下游BAC

# 階層式定序

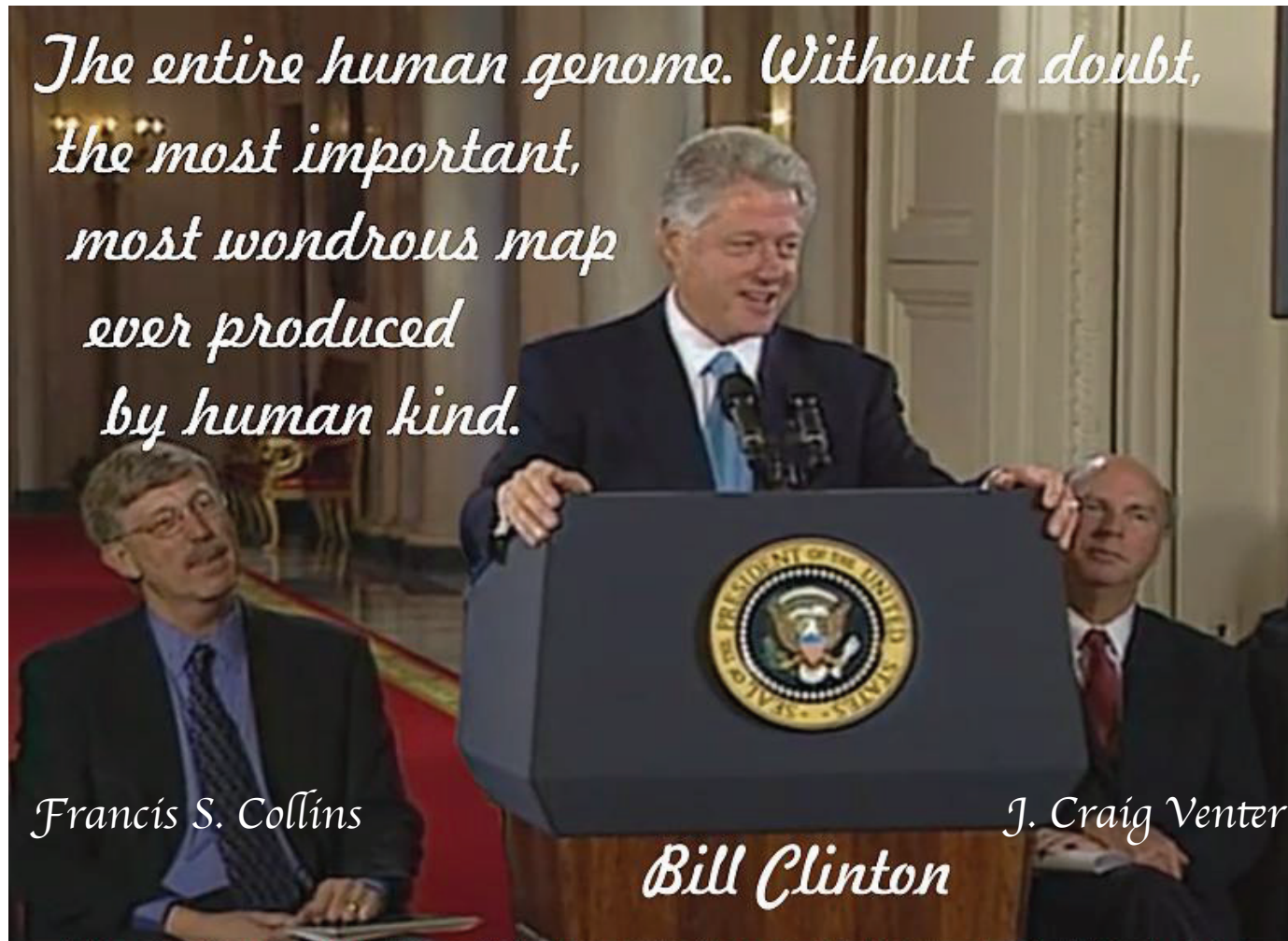


## 霰彈式定序及階層式定序比較

	霰彈式定序	階層式定序
時間	短	長
經費	少	多
人力	少	多
空缺區域(gap)	大	小

- 當第一個物種以階層式定序出來後，可作為其它物種基因組參考圖譜，因此其它物種即使使用霰彈式定序法，也可以得到較精確的基因組圖譜。
- 目前已有公司號稱只要**1,000**美元即可完成個人定序

人類基因體計劃共識－所有基因為人類共享，不可註冊



*The entire human genome. Without a doubt,  
the most important,  
most wondrous map  
ever produced  
by human kind.*

*Francis S. Collins*

*J. Craig Venter*

*Bill Clinton*

More science quotes at Today in Science History [todayinsci.com](http://todayinsci.com)



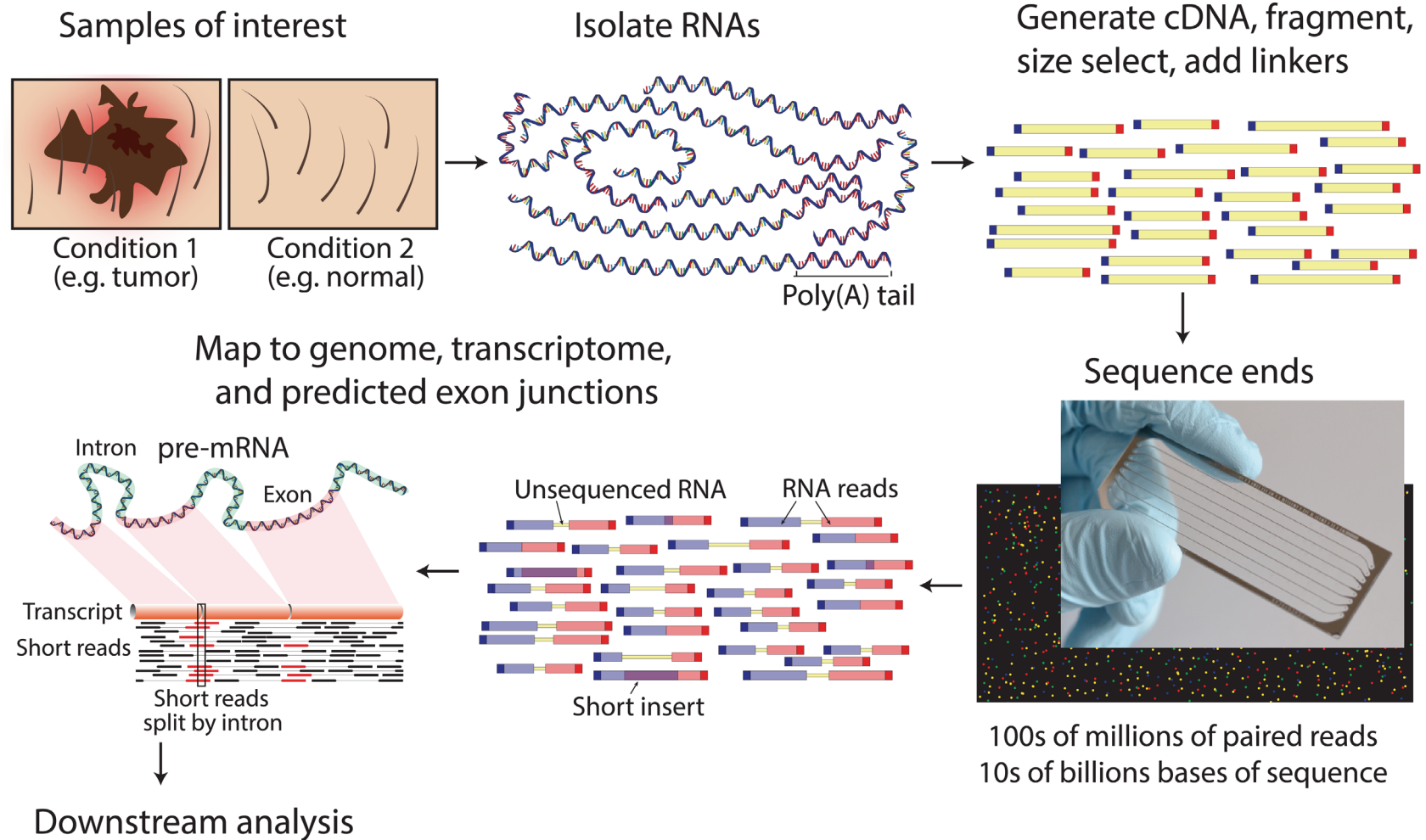
# 定序後的挑戰－基因註解

**Q:**人類基因組約**3.2G**，含有約**20,000-25,000**個基因，如果以每個基因平均長度**10 kb**計算，那麼這些基因約佔基因組**8% (250,000 kb)**。那麼我們要如何知道基因的位置？

1. 軟體預測：由基因的特質預測，如果有參考基因庫會較準確
2. 轉錄體定序：定序**RNA**的序列，並將其對回至基因組相對位置，該方法較準確，但部份表現量較低的基因不一定會被定序

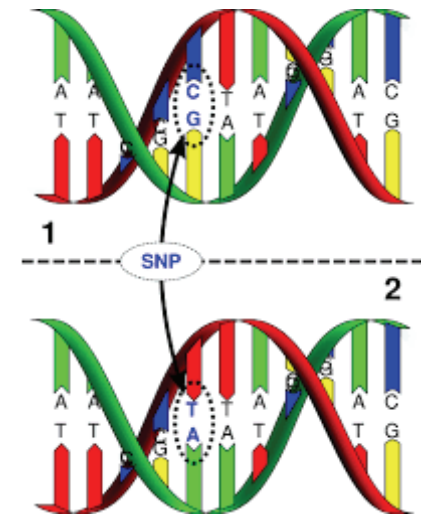
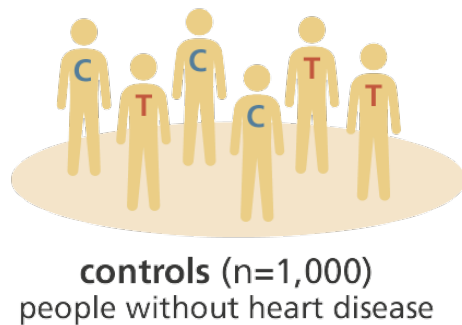
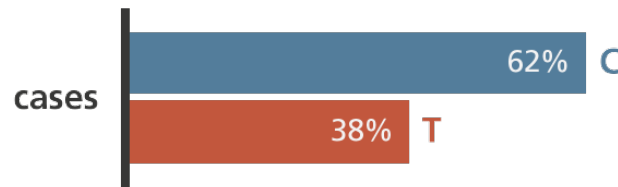
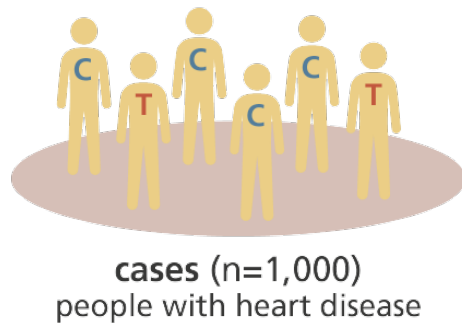
# 轉錄體定序 (RNA-seq)

抽取RNA → 轉成雙股DNA並切成小片段 → 霰彈式定序 → 對應到基因組相對位置




# 人類基因體的應用

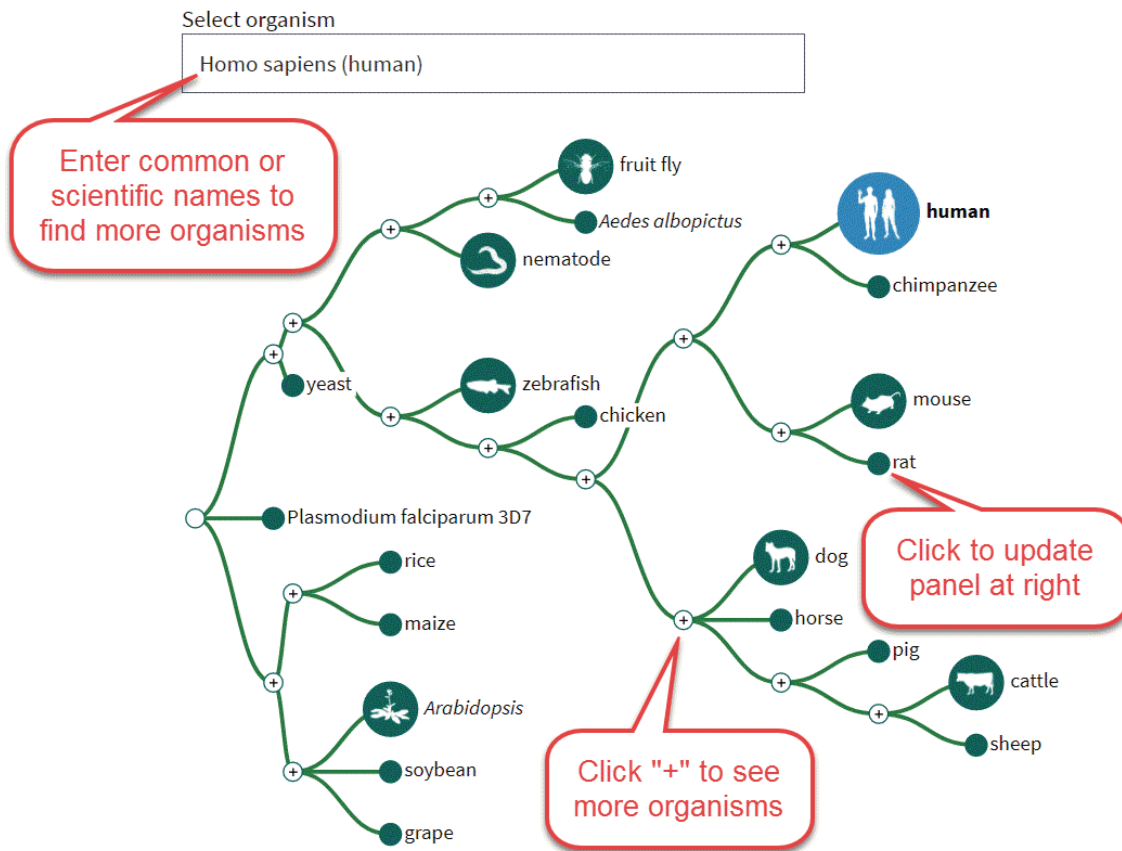
1. 基因功能分析
2. 醫葯開發 - 精準醫學
3. 基因尋找 – GWAS
4. Metagenomics – 微生物菌相分析
5. Epigenomics – DNA 甲基化分析
6. 其它



# NCBI Genome Data viewer

## Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 540 eukaryotic [RefSeq](#) genome assemblies. 



### Homo sapiens (human) genome

Search within selected assembly

Search in genome

Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

Assembly

GRCh38.p11

[Browse genome](#) [BLAST genome](#)

Select assembly version

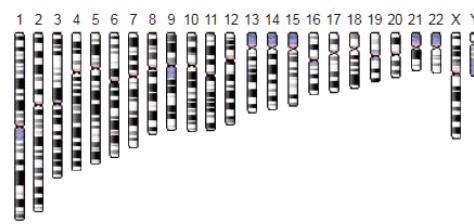
#### Assembly details

Name	GRCh38.p11
RefSeq accession	GCF_000001405.37
GenBank accession	GCA_000001405.26
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome

#### Annotation details

Annotation Release 108

Release date



# Homo sapiens (human) genome



Search in genome

aquaporin



Genes

Other

Name	Location
AQP4	Chr18: 26.85M - 26.87M
AQP1	Chr7: 30.91M - 30.93M
AQP2	Chr12: 49.95M - 49.96M
AQP3	Chr9: 33.44M - 33.45M
AQP5	Chr12: 49.96M - 49.97M
AQP9	Chr15: 58.14M - 58.19M
AQP8	Chr16: 25.22M - 25.23M

Examples: [TP53](#), [chr17:7667000-7689000](#), [rs334](#), [DNA repair](#)

Assembly

GRCh38.p12

Search "Aquaporin"

result



Chromosome 7

location: 7p14.3  
(cytological map)

Transcript ID

Current position  
(physical map)

### Genome Data Viewer

Homo sapiens: GRCh38.p12 (GCF\_000001405.38) Chr 7 (NC\_000007.14): 30,910,312 - 30,926,899

Reset All Share this page FAQ Help Browser Agreement Version 4.5

Region: AQP1 (Gene) NM\_001329872.1 (Transcript)  
Exon cds range: 30,911,910 - 30,912,293, range: 30,911,694 - 30,912,293

NC\_000007.14

Genes, NCBI Homo sapiens Annotation Release 109, 2018...  
AQP1 [4+]

Genes, Ensembl release 93  
000240583 ...

dbSNP Build 151 (Homo sapiens Annotation Release 108) all data

Cited Variants, dbSNP Build 150 (Homo sapiens Annotat...)

RNA-seq exon coverage, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log base 2 scaled

RNA-seq intron-spanning reads, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log

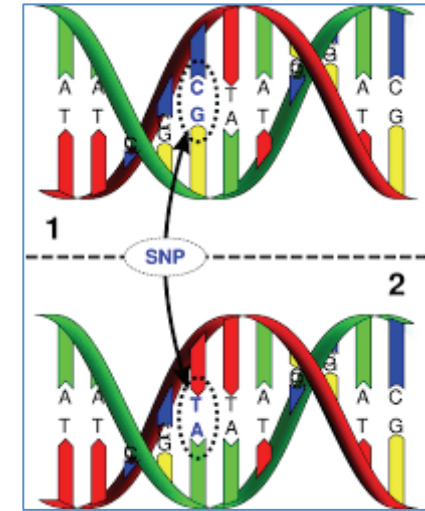
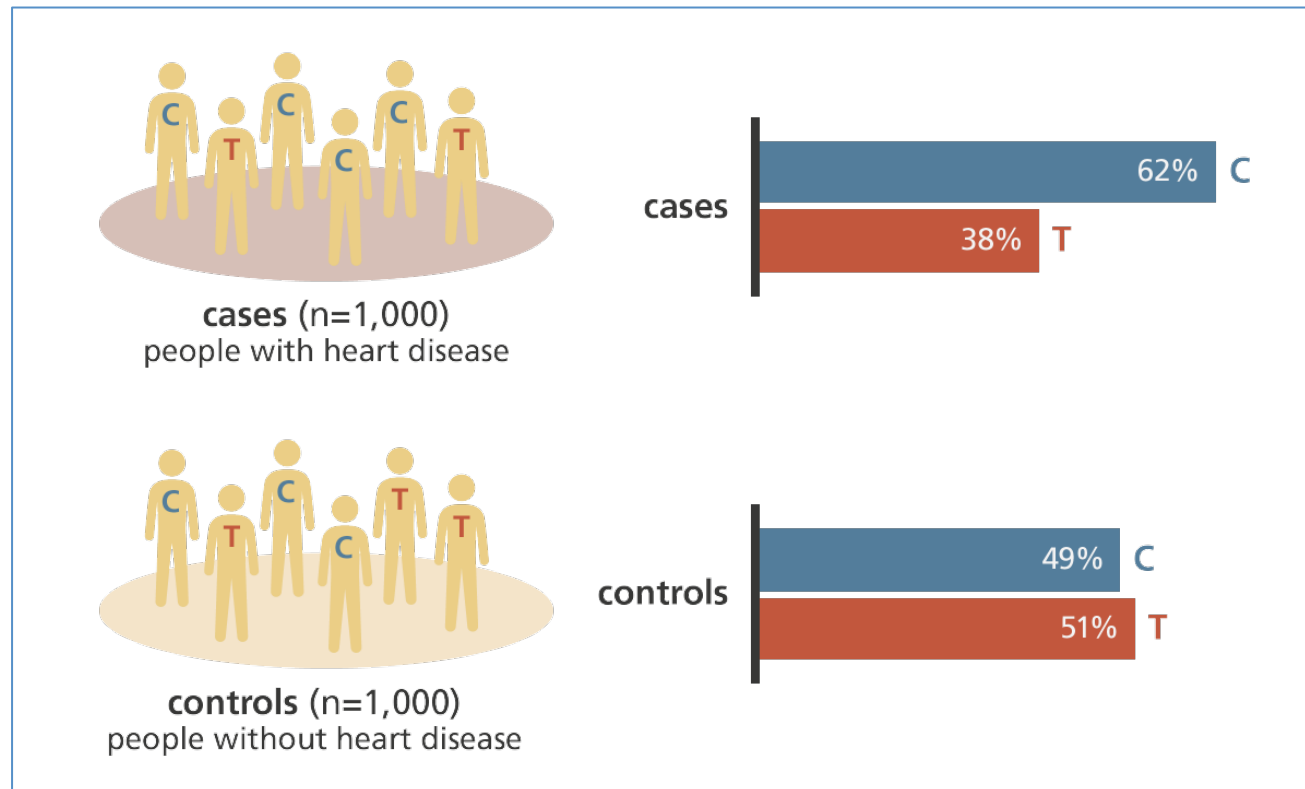
Feedback

The screenshot displays the Genome Data Viewer interface. On the left, an ideogram view shows chromosome 7 highlighted in green. Below it, a search table lists aquaporin genes across various chromosomes. The main panel shows a detailed view of the AQP1 gene on chromosome 7, including its structure with exons and introns, and associated genomic data like dbSNPs and RNA-seq coverage. Red arrows from external text boxes point to specific features: 'Chromosome 7' points to the ideogram, 'location: 7p14.3 (cytological map)' points to the cytogenetic band on the ideogram, 'Transcript ID' points to the transcript ID in the region dropdown, 'Current position (physical map)' points to the genomic coordinates at the top, and 'Gene structure' points to the gene model tracks. A 'Feedback' button is visible in the bottom right corner.

Search result

# 人類基因體的應用

1. 基因功能分析
2. 醫葯開發 - 精準醫學
3. 基因尋找 – GWAS
4. Metagenomics – 微生物菌相分析
5. Epigenomics – DNA 甲基化分析
6. 其它



SNP (單一核苷酸多型性)

GWAS分析

# GWAS (全基因組關聯分析)

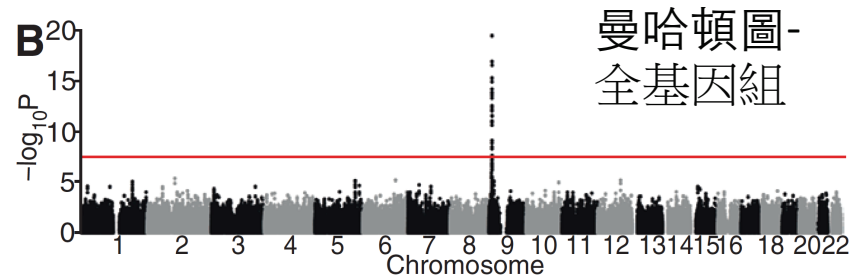
- Genome-wide association study
- 由人類全基因組中找出某性狀關聯基因
- 方法為比較在具有不同性狀的族群之間的個體基因組差異，如金髮及黑髮或具有遺傳疾病及正常人
- 兩群個體其它性狀不能差異太大
- 全基因組定序 → GWAS分析 → SNP candidate → 基因功能分析



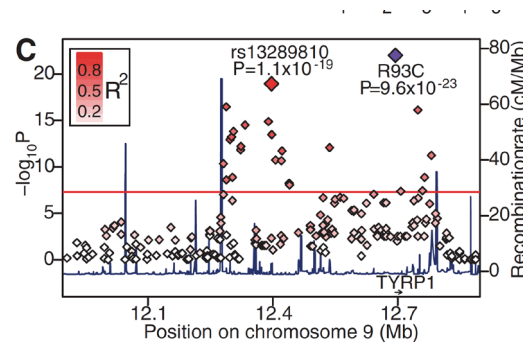
42位



43位



曼哈頓圖-  
全基因組



曼哈頓圖-  
基因TYRP1

**Thanks for your attention**