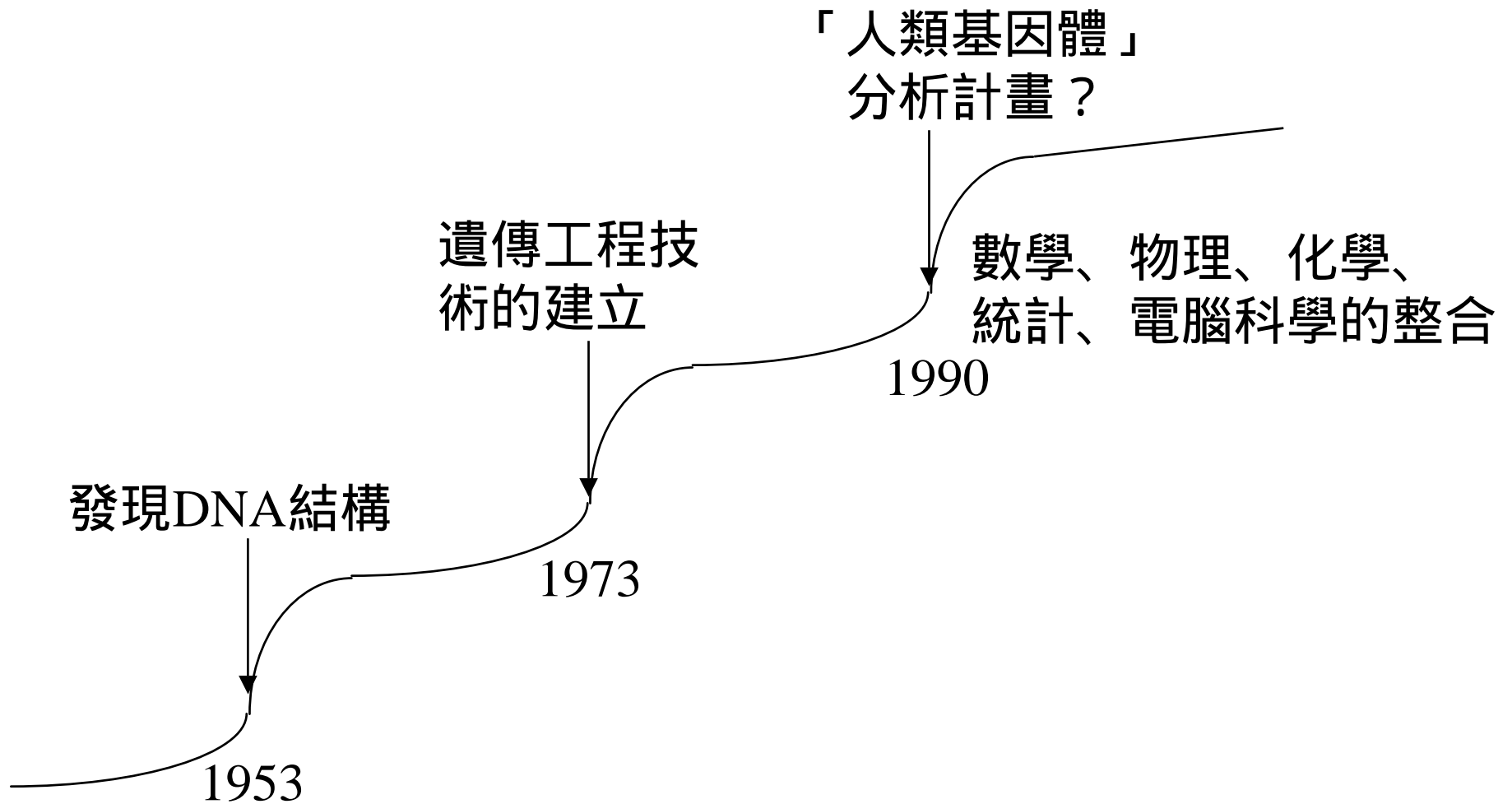


人類基因體分析計畫 與生物資訊學

陽明大學 生科系, 生化所,
生物資訊學程
楊永正

觀察歷史有助於了解時代趨勢



基因體分析計畫帶動了 生物科技的發展

- 基因圖譜有助於致病基因的定位
- 序列有助於找尋同源的基因
- 基因體全序列有助於解釋生物體的功能 (細胞分化, 細胞交互作用等)

定序的原理

...GTCGACTGCAAT...

ddA

...GTCGACTGCAA

...GTCGACTGCA

...GTCGA

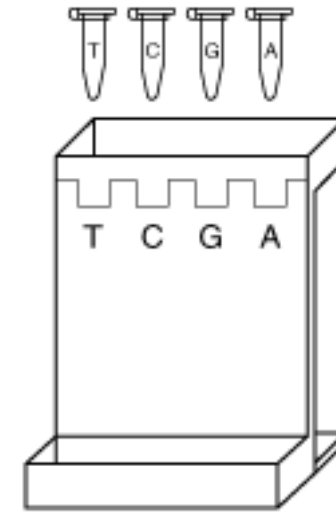
ddC

...GTCGACTGC

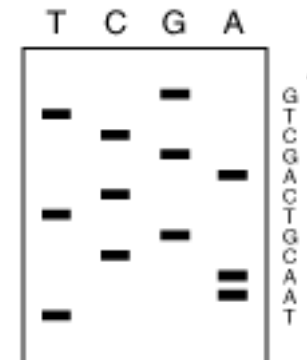
...GTCGAC

...GTC

1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation



2. Sequence read (bottom to top) from gel autoradiogram



基因體分析計畫發展史

- 決定cDNA或基因體序列?
- 由下而上或由上而下的策略
 - 由下而上: 霰彈槍法(shotgun method)
 - 由上而下: 先建基因圖譜(map)
- 霰彈槍法決定全基因體序列

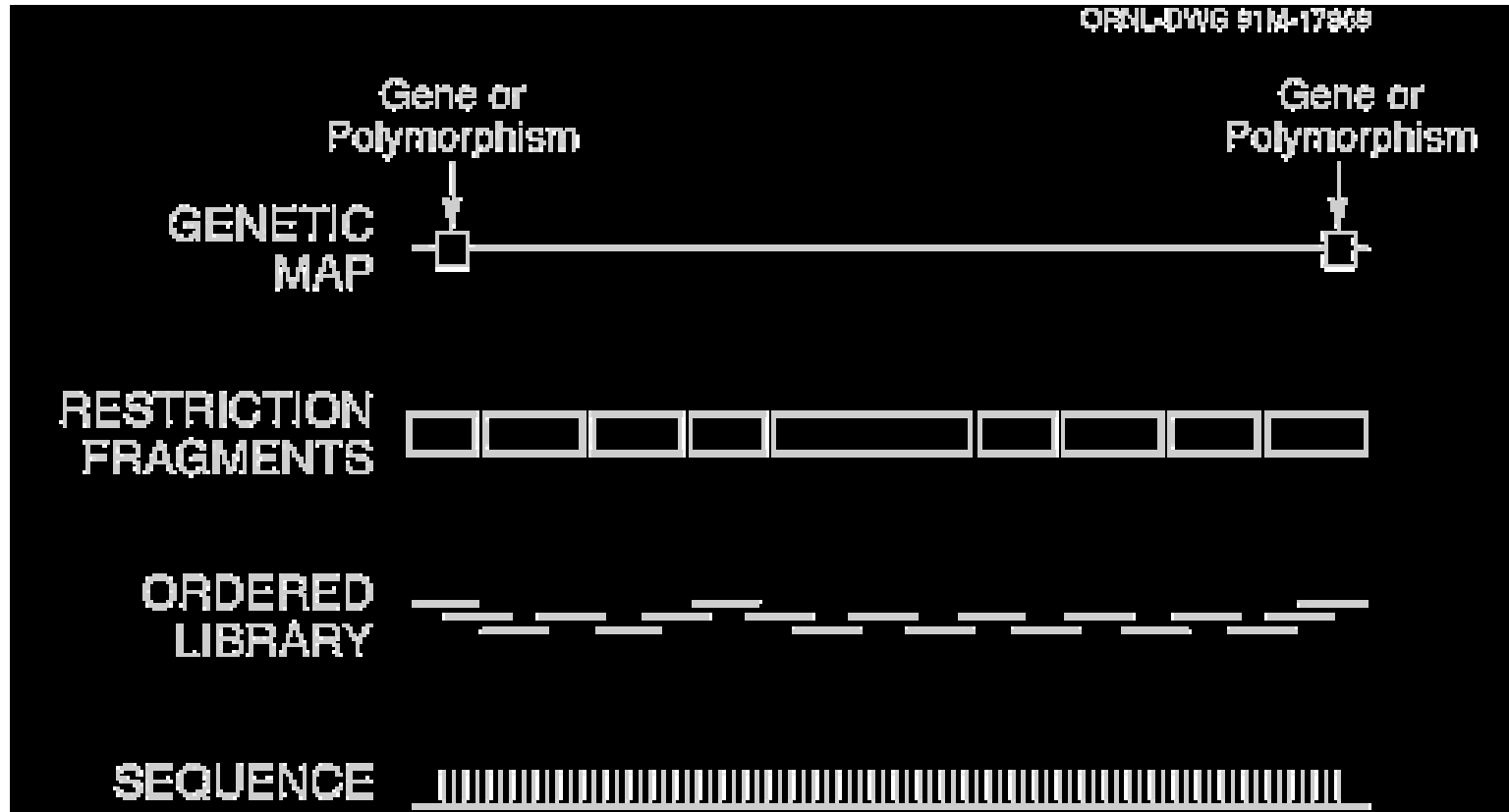
決定cDNA或基因體序列?

- 決定cDNA序列
 - different cells express different set of genes
 - different mRNAs have different abundance
 - don't know when to stop; less efficient
- 決定基因體序列
 - complete set of genetic information
 - less redundancy than cDNA approach
 - goal is clear; more efficient

霰彈槍法(shotgun method)

定序初期效率高,後期越來越困難
重複序列會增加組合序列的困難度

由上而下的策略

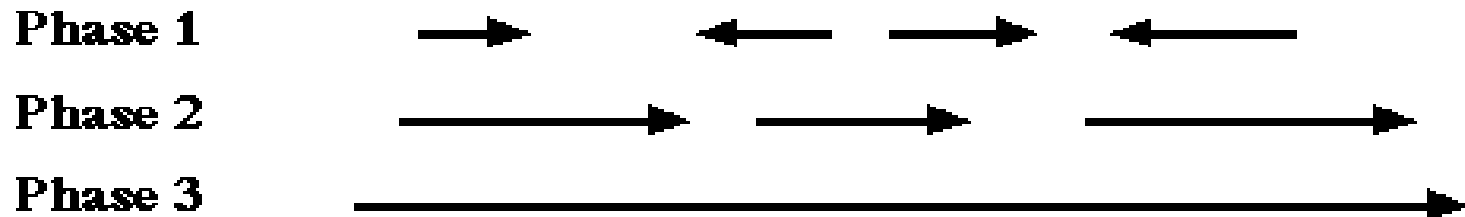


百慕達協定

“Unfinished” sequence data should be released as soon as it is “usable” for homology searching and other types of sequence analysis.

人類基因體分析計畫

第一期：方向, 順序未定
第二期：不連續
第三期：已連續



GenBank存放序列的方法

- Organismal Divisions:
 - BCT: Bacterial seq.
 - PRI: Primate seq., include Human Phase 3
 - ROD: Rodent seq.
 - MAM: Other mammalian seq.
 - VRT: Other vertebrate seq.
 - INV: Invertebrate seq., include Drosophila, C. elegans Phase 3
 - PLN: Plant and Fungal seq., include Arabidopsis Phase 3
 - VRL: Viral seq.
 - PHG: Phage seq.
 - RNA: Structural RNA seq.
 - SYN: Synthetic and chimeric seq.
 - UNA: Unannotated seq.
- Functional Divisions:
 - EST: Expressed Seq. Tags
 - STS: Sequence Tagged Sites
 - GSS: Genome Survey seq.
 - HTG: High Throughput Genomic seq., include Phase 1 and 2 from all organisms

霰彈槍法決定全基因體序列

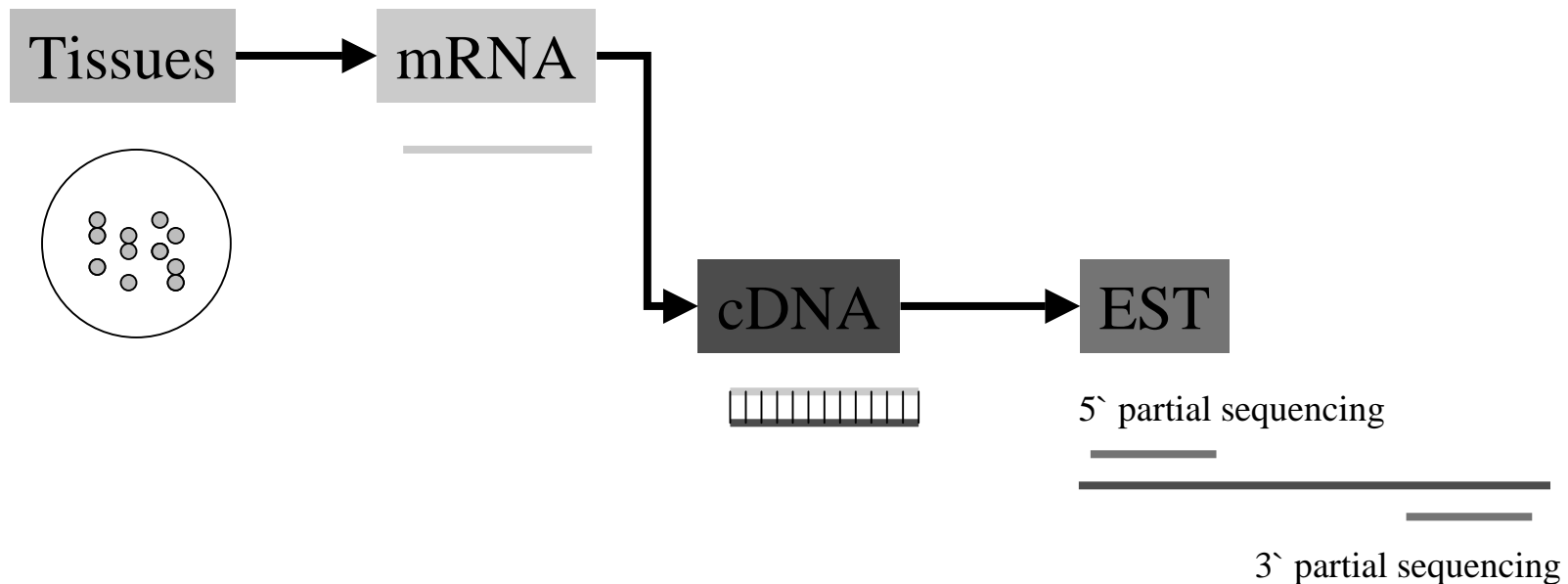
- 由上而下的策略
 - 需做基因庫(library)與定序兩種人才
 - 決定圖譜(mapping)是的速率決定步驟
- 霰彈槍法決定全基因體序列
 - 做完基因庫後,只需定序人才
 - 增加定序儀,即可增加定序的速率
 - ➡ 單一步驟,可平行處理
- 攪亂了群雄割據的佈局

具爭議性的Craig Venter

- 表現序列標幟(expressed sequence tag, EST)
- 以霰彈槍法決定微生物的全基因體序列
- 以霰彈槍法決定果蠅與人的全基因體序列

Expressed Sequence Tag (EST)

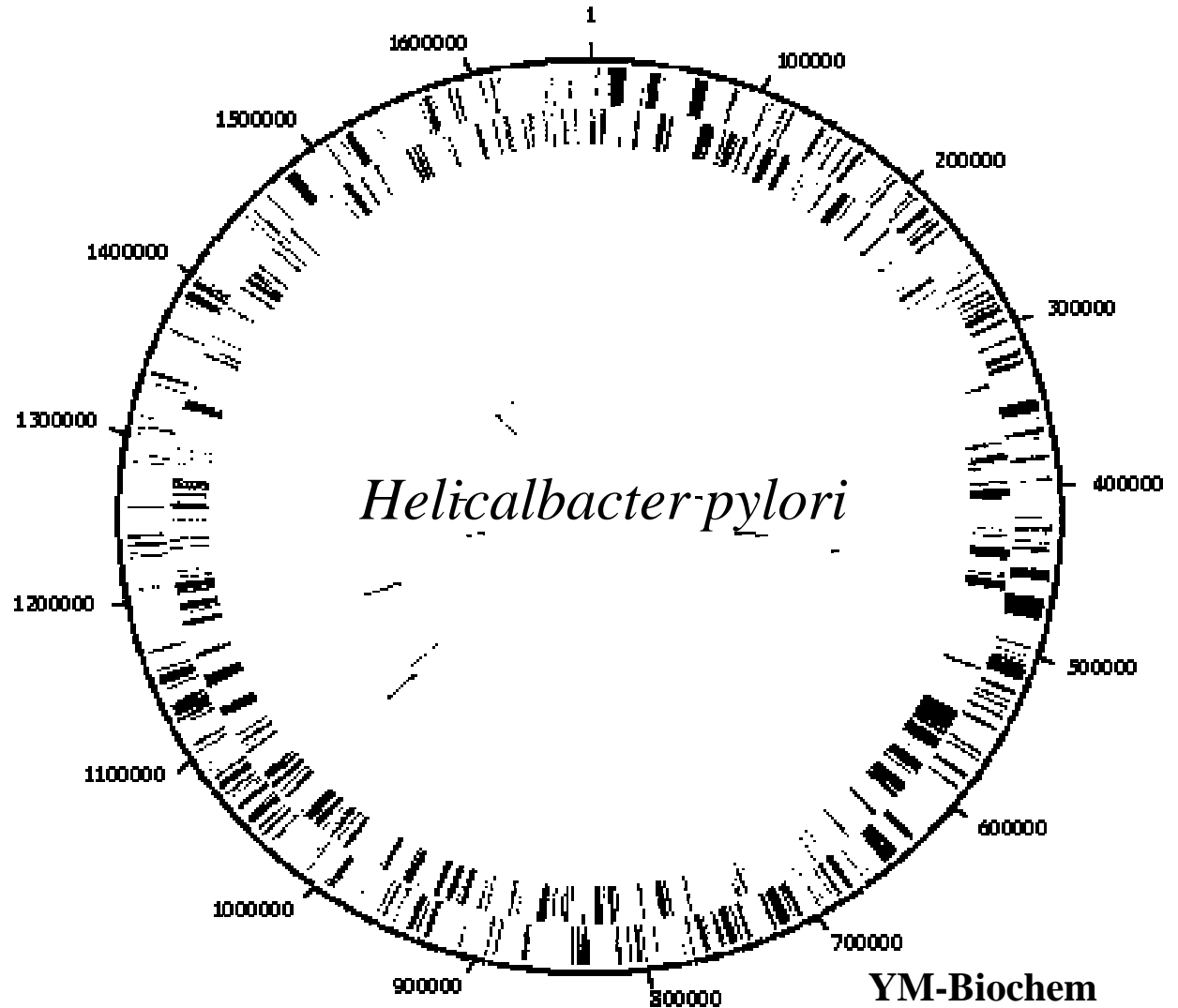
- Partial cDNA sequences of genes expressed in different tissues



有全基因體序列的好處

You will know not only what genes are there, but also what genes are missing.

You will have not only the information about components, but also the interactions among components.



定序技術的改進

滎陽千萬鹼基定序計畫

一、硬體設施：

ABI 377

Acrylamide gel
(lane tracking
is needed)



x 5

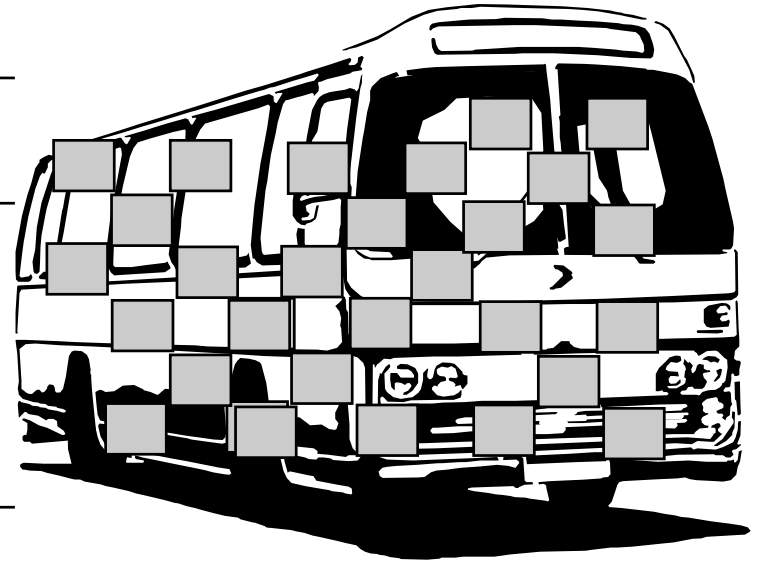
ABI 3700
Capillary
electrophoresis



x 1

人類基因體分析計畫

	初稿	完稿
結束日期	June 26, 2000	2003
序列覆蓋率	90%	>99%
錯誤率	1%	0.01%



Phase 1



Phase 2



Phase 3



Celera公司釋出序列資訊的方式

Not to reproduce, redistribute, re-package, adapt or prepare derivative works of Celera data ... for third party, in any form whatsoever, for any purpose.

Science期刊對Celera公司 開特例的後果

生物資訊中心不能釋出註解Celera
序列的加值資訊

=> 須購買資訊或有能力自行分析

榮陽千萬鹼基計畫的意義

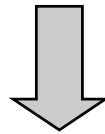
國際意義

火車頭的帶動角色

對生物醫學與生技產業的影響

後基因體分析世代

所有基因體序列資訊



瞭解DNA中蘊涵的資訊
生物科技的應用

DNA 語言為何難懂?

5 10 15 20 25 30 35
GARBA GEYOU THROW OUTJU NKYOU KEPT HEREF
40 45 50 55 60 65 70
ORE90 %OFTH EGENO MEISJ UNKSB RENNE R

**Garbage you throw out, junk you keep. Therefore, 90%
of the genome is junk. - S. Brenner.**

語言的三大要素： 字,詞與文法

**Garbage you throw out, junk you keep. Therefore,
90% of the genome is junk. - S. Brenner.**



A->B, B->C, C->D,
D->E, E->F, --- *etc.*

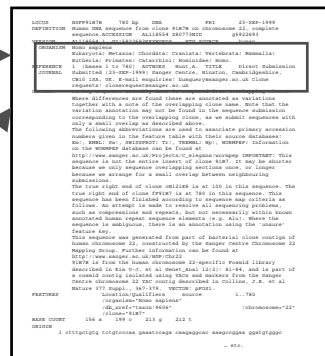
Hbscbhf zpv uispv pvu, kvol zpv lffq. Uifsfgpsf,
01^ fg uif hfopnf jt kvol. =T. Csfoofs.

語言學的類比 - 字與詞

Junk

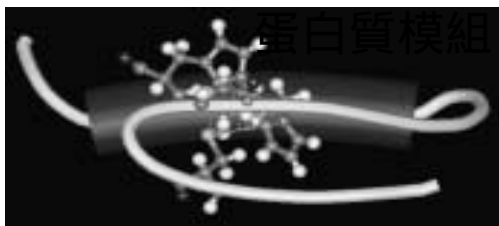
Garbage you throw out, junk you keep. Therefore, 90% of the genome is junk.

- S. Brenner.

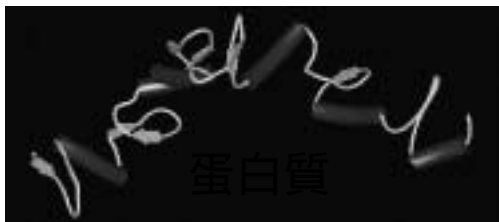


生命之書

字

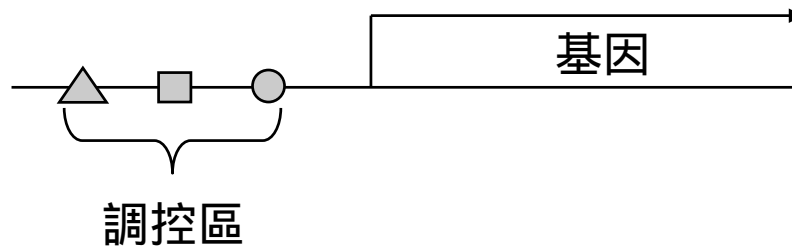


詞



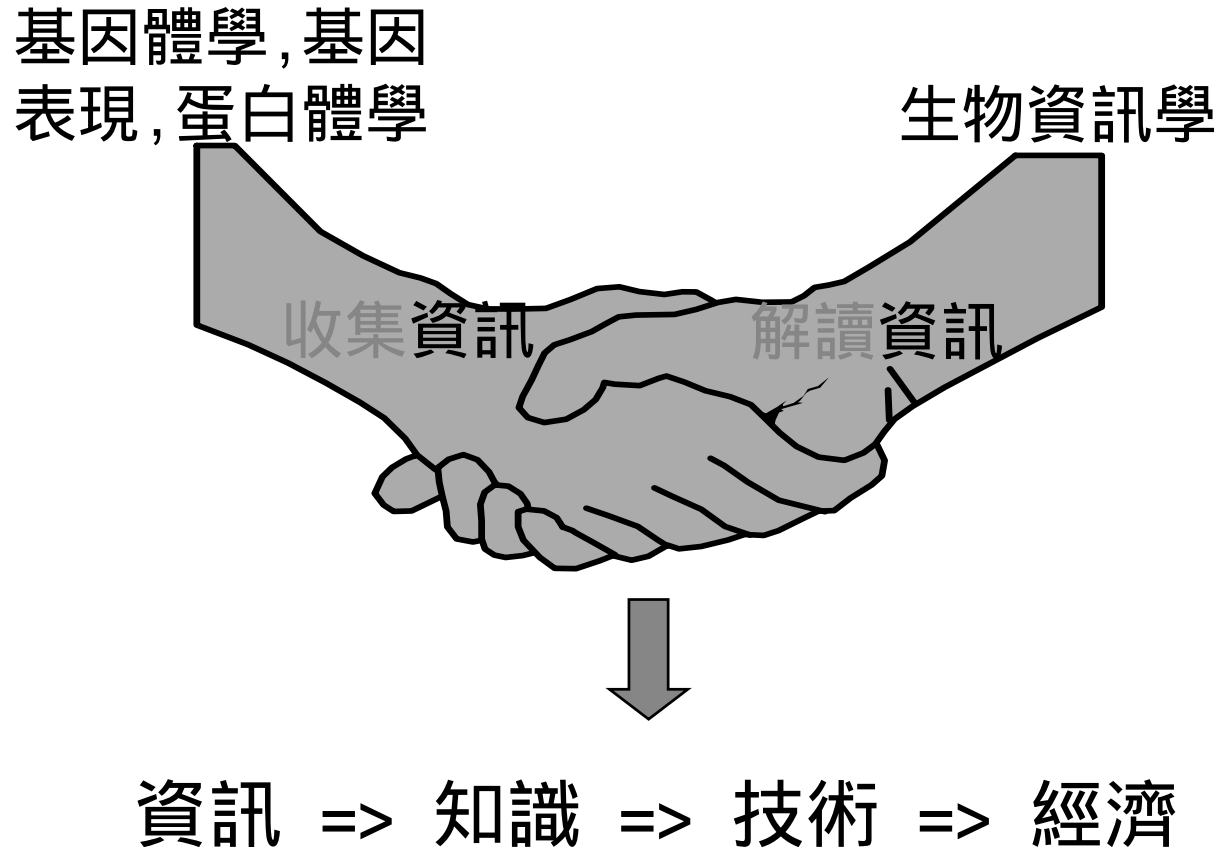
獨立折疊單元

轉錄因子接合位置

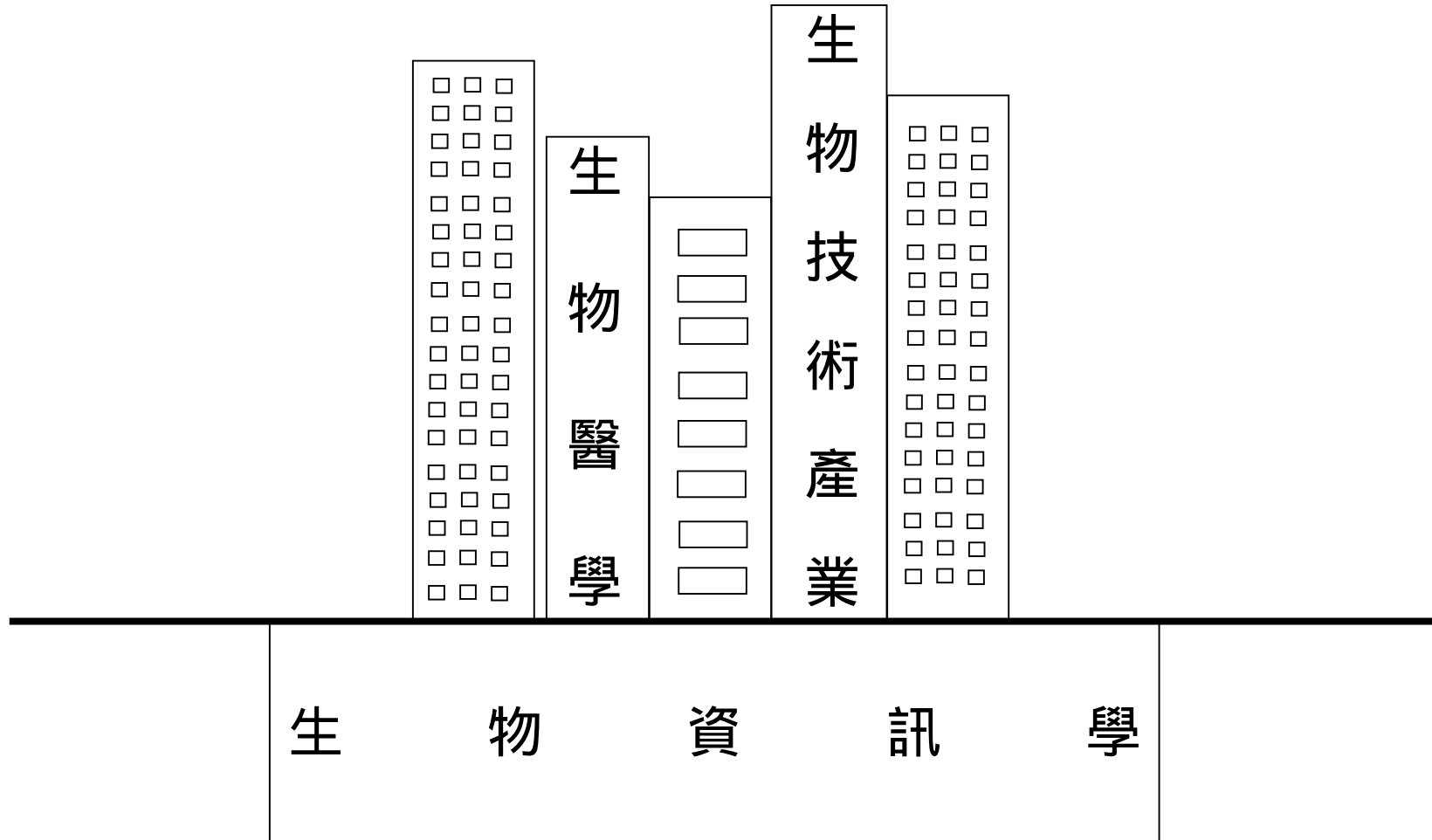


序列決定因子

生物資訊學是發展生物 科技的催化劑

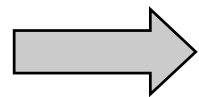


生物資訊學的重要性



分子醫學

- 疾病診斷方法的改進
- 遺傳疾病的早期診斷
- 循理化藥物設計
- 建立疾病的動物模型以測試藥物
- 基因治療 (gene therapy)
- 個人化之疾病治療



尋找致病基因或路徑

如何運用基因體序列資訊?

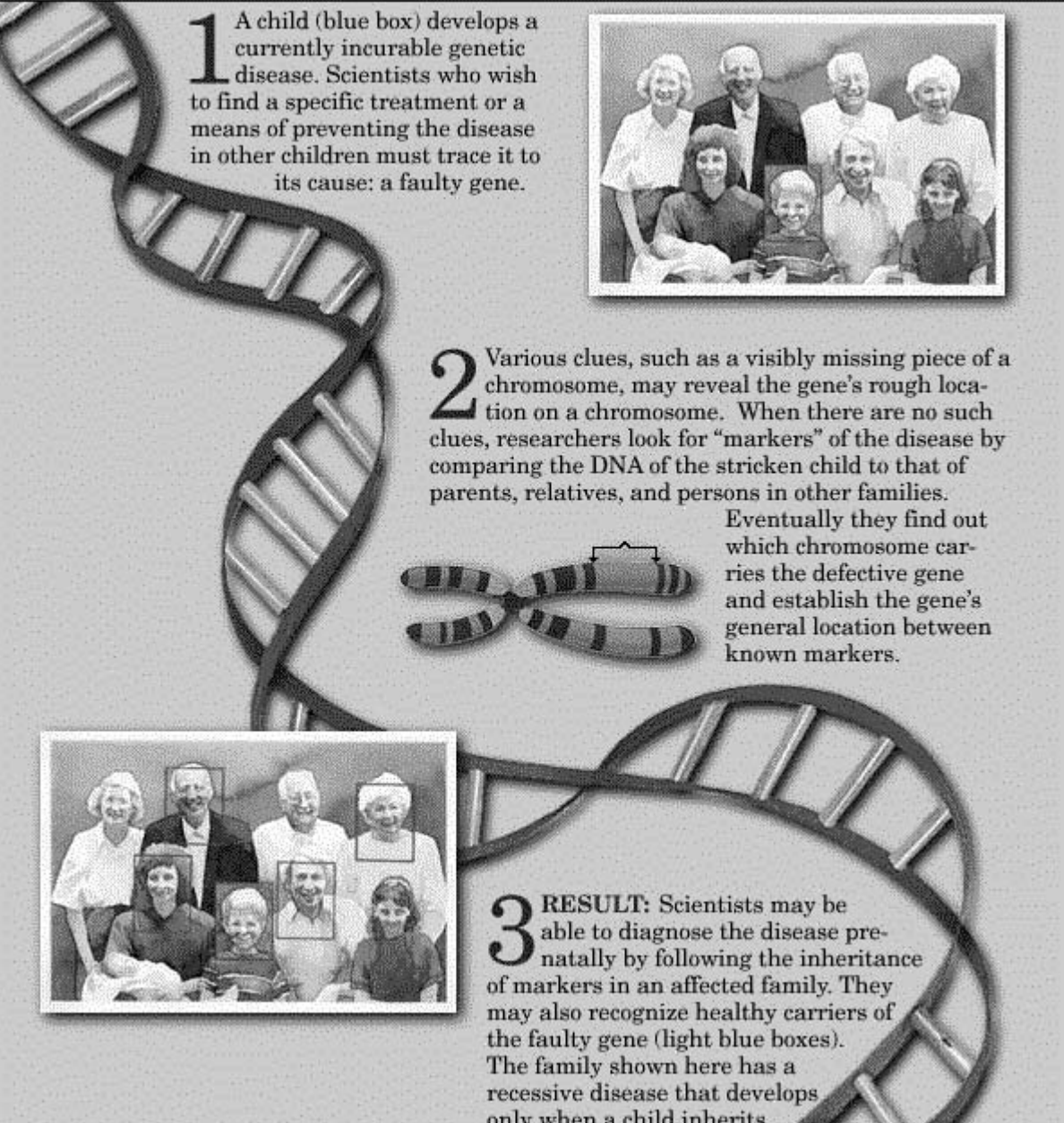
尋找新基因的功能
驗證預測基因的存在
尋找造成疾病的基因

尋找致病基因


Important databases

- OMIM
- Gene variation databases
 - Cystic fibrosis mutation database
 - Huntington disease mutation database
 - *etc.*

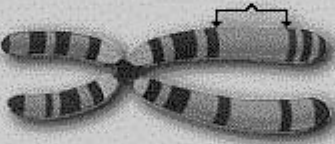
<http://www.hhmi.org/GeneticTrail/gatefold/gate2.htm>




1 A child (blue box) develops a currently incurable genetic disease. Scientists who wish to find a specific treatment or a means of preventing the disease in other children must trace it to its cause: a faulty gene.



2 Various clues, such as a visibly missing piece of a chromosome, may reveal the gene's rough location on a chromosome. When there are no such clues, researchers look for "markers" of the disease by comparing the DNA of the stricken child to that of parents, relatives, and persons in other families. Eventually they find out which chromosome carries the defective gene and establish the gene's general location between known markers.

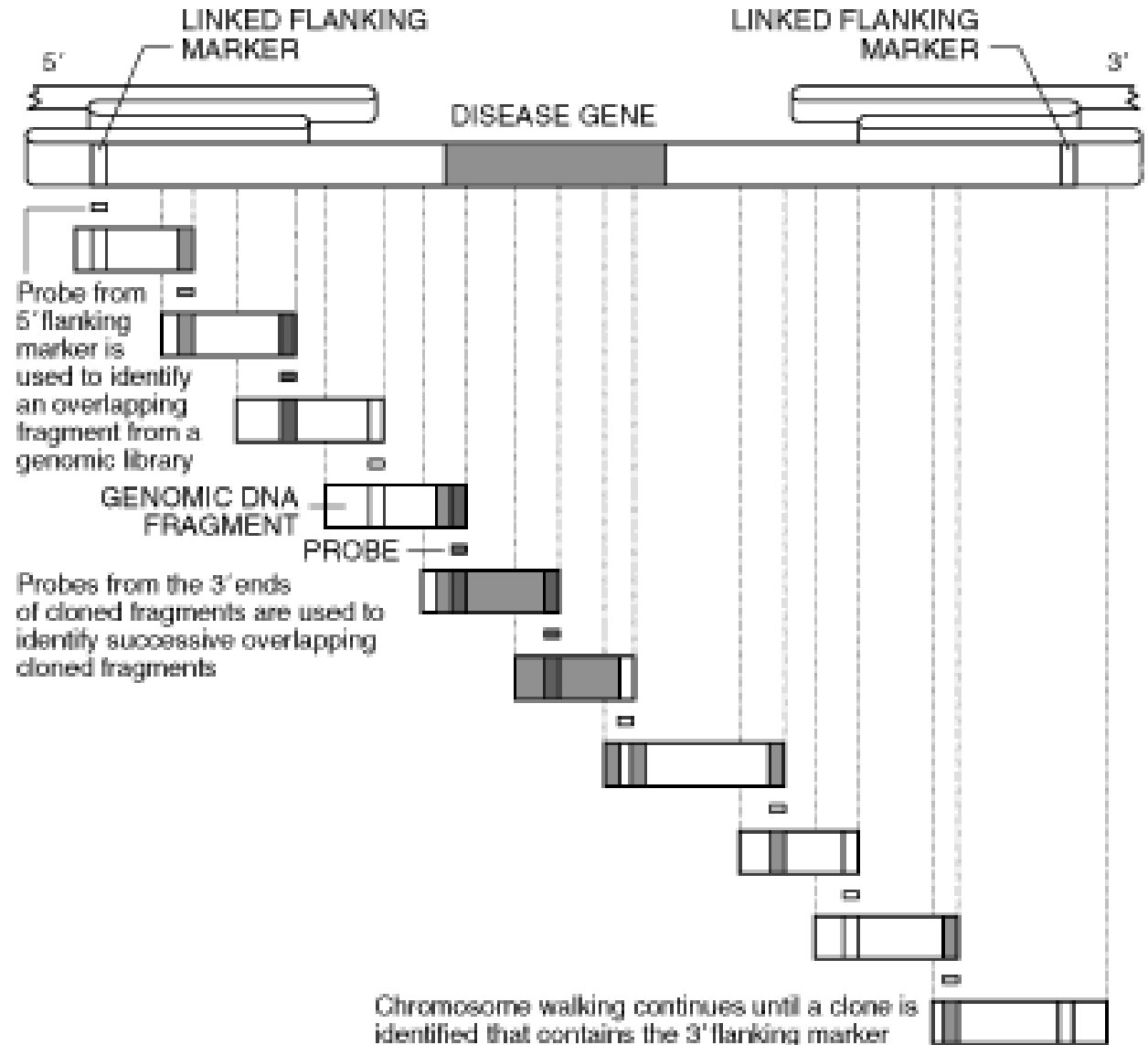


3 RESULT: Scientists may be able to diagnose the disease prenatally by following the inheritance of markers in an affected family. They may also recognize healthy carriers of the faulty gene (light blue boxes). The family shown here has a recessive disease that develops only when a child inherits



利用定位選殖法 尋找致病基因

Chromosome walking
& positional cloning



尋找新基因的方法

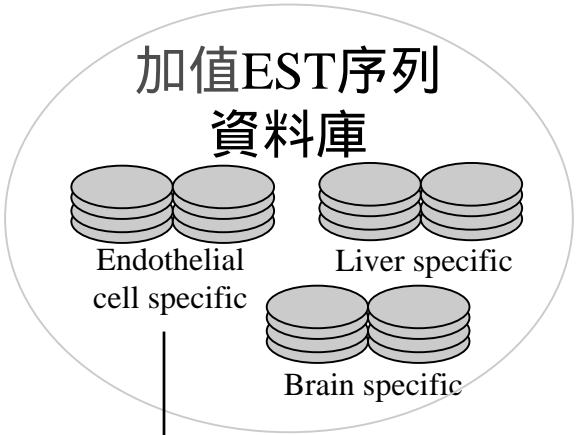
利用基因註解

利用EST資料庫

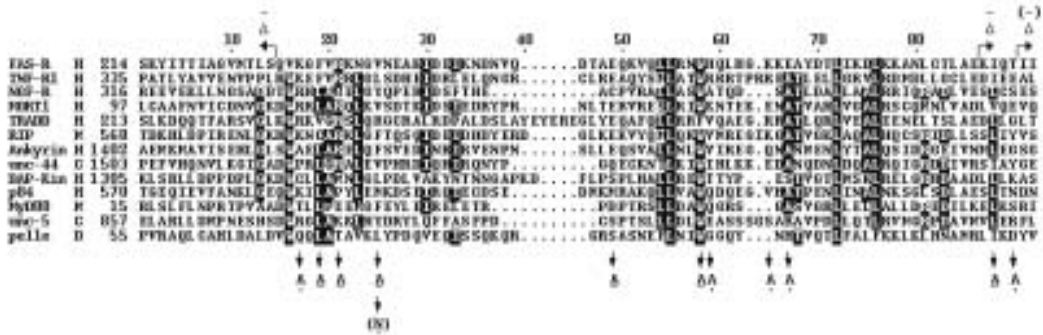
利用比較基因體學

利用基因表現的差異

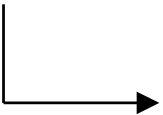
利用EST資料庫 尋找新基因



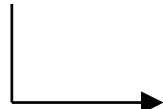
Search for known
protein family →



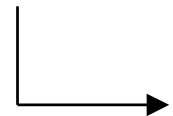
e.g. death domain consensus



EST records



Look for gene that
encode this EST

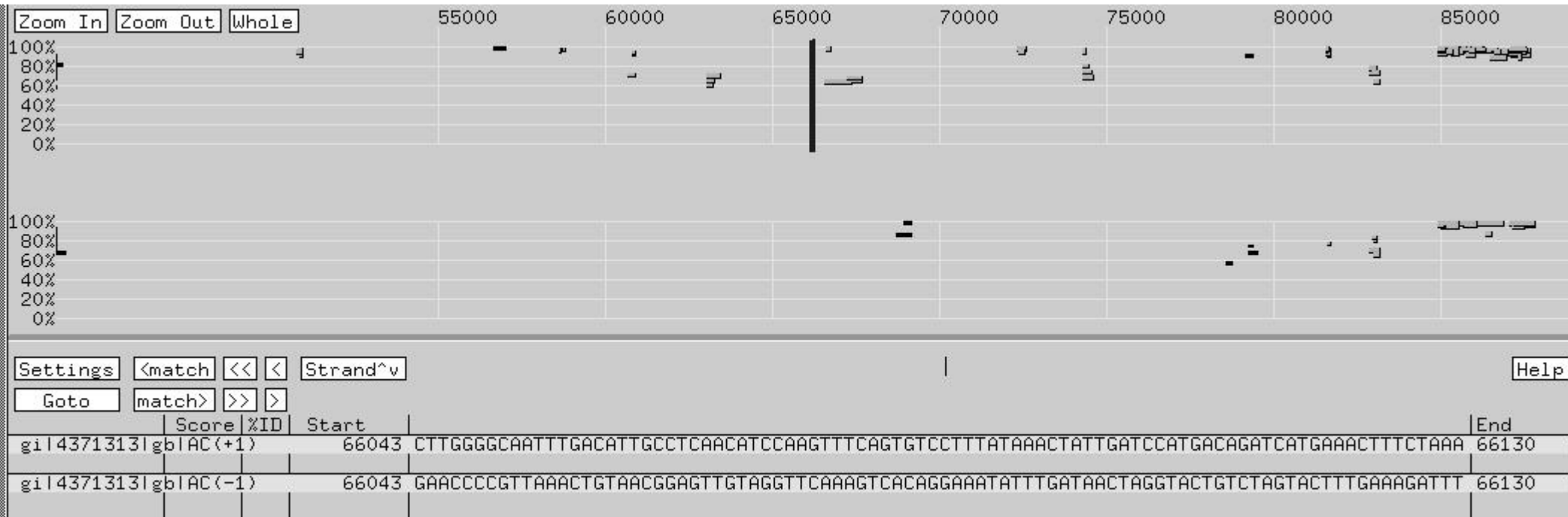


Gene
identification

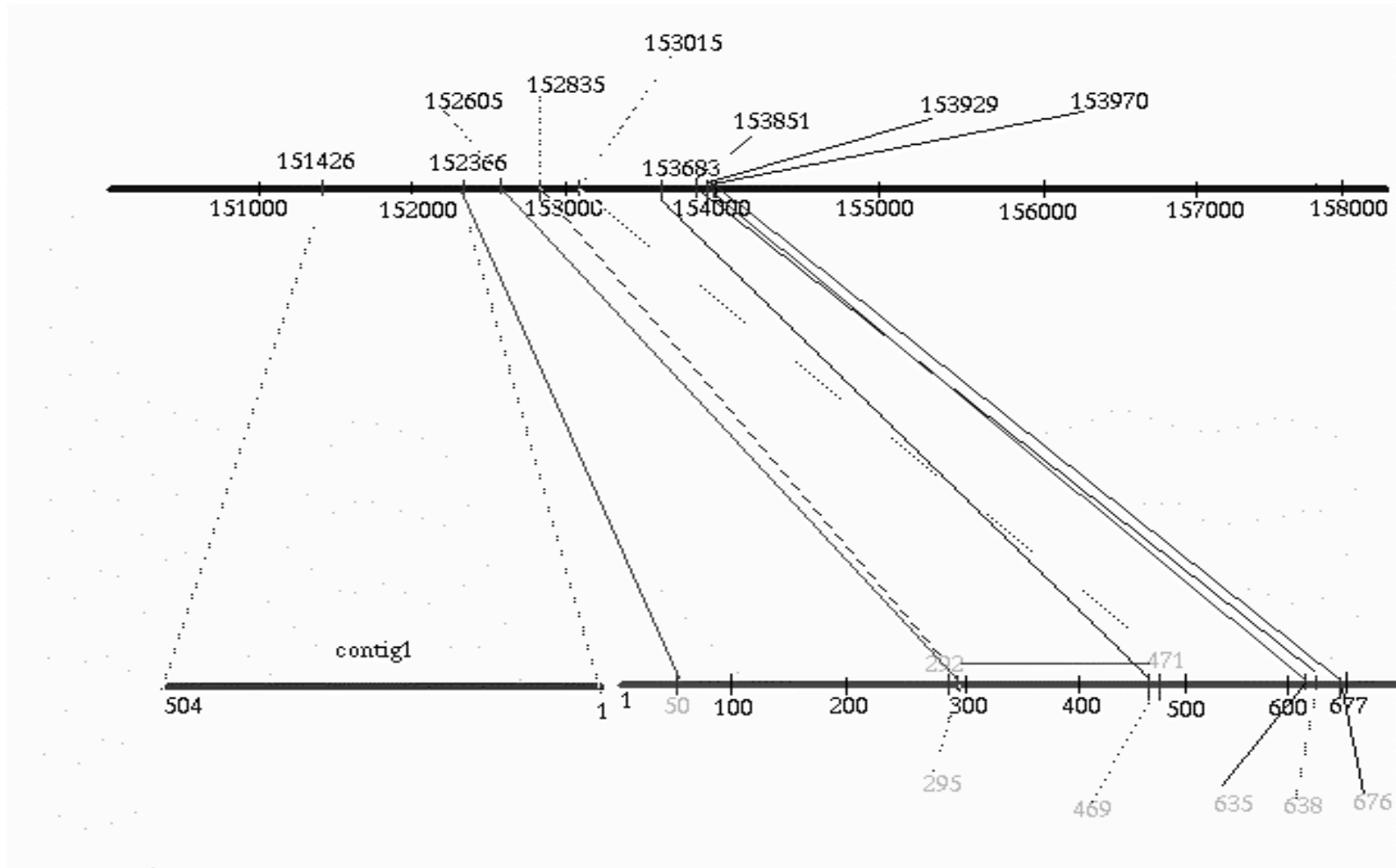
Zhai *et al* (1999) VEGI, a novel cytokine of the tumor necrosis factor family, is an angiogenesis inhibitor that suppresses the growth of colon carcinomas *in vivo*. **FASEB** 13, 181-189.

利用基因體序列尋找新基因的功能

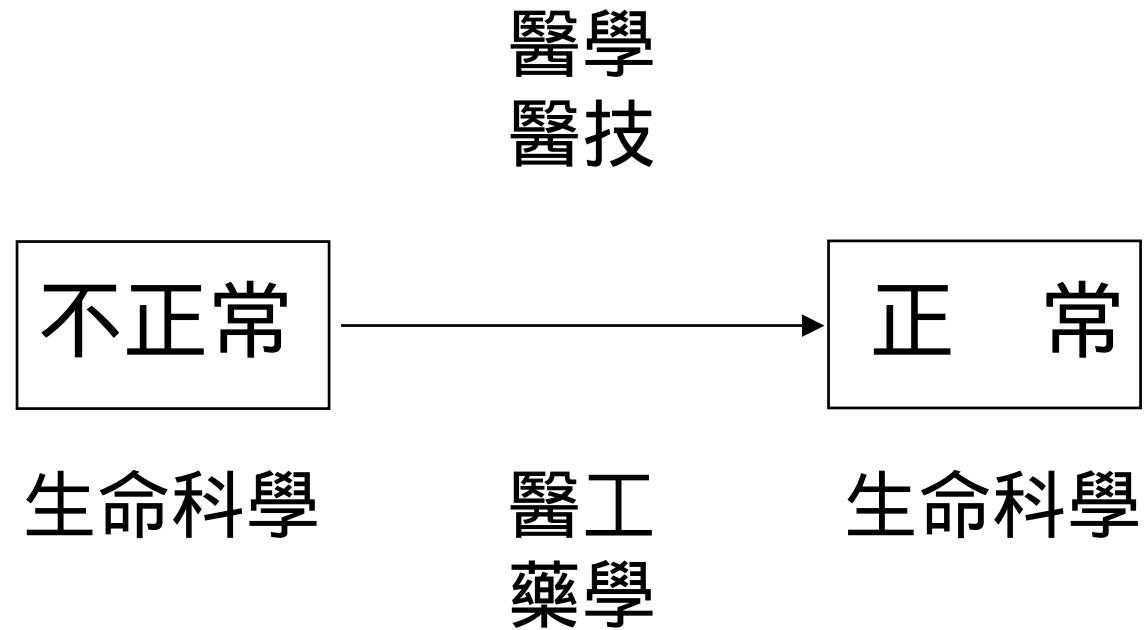
EST => HTGS => annotation => Reveal possible function



利用基因體序列驗證預測的基因



醫學與生命科學的關係



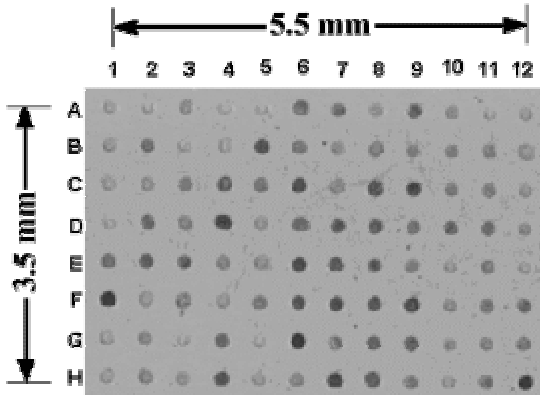
常用的研究策略

尋找差異

判定因果關係

找到主控因子/步驟

微陣列(microarray)分析



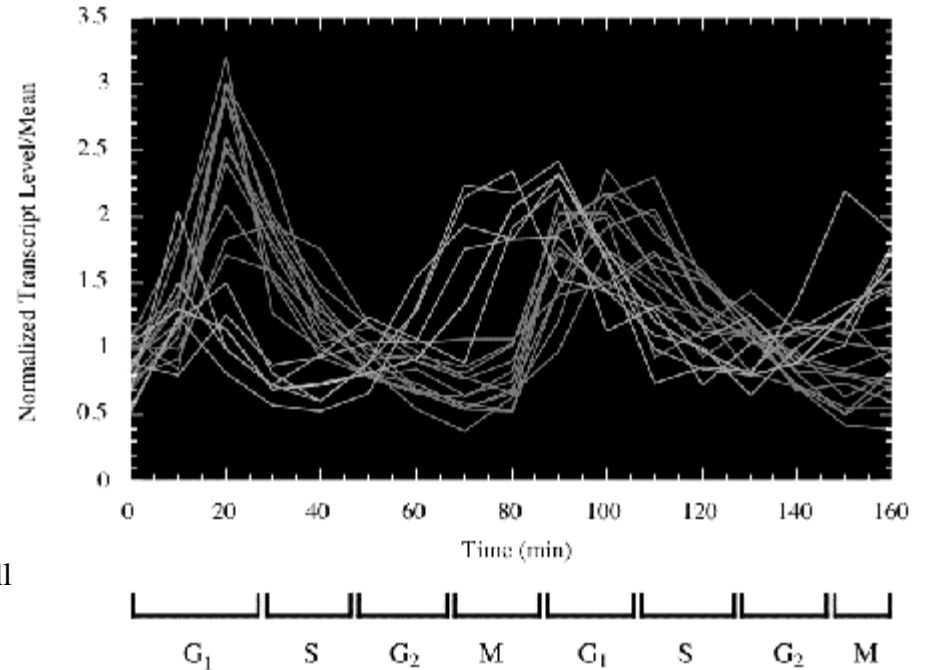
Picture taken from
http://www.ibms.sinica.edu.tw/~peck/chinese/marray_4.htm

Explanation: Below is the full data for every gene in yeast. Data between timepoints have been normalized with respect to each other. There are some bacterial geneon each of the 4 chips. We recommend that this document be downloaded and opened in Excel. The 17 data after each gene are the normalized fluorescence between 0 and 160 minutes after cell cycle reinitiation from START. Have fun.

Gene Name zero ten twenty thirty forty fifty sixty seventy eighty ninety
 hundred one-ten one-twenty one-thirty one-fourty one-fifty one-sixty
 18srRnaa 22 38 41 43 23 29 25 20 17 98 46 27 23 38 27 28 287
 18srRnab 5 9 -13 -9 -14 -13 -11 -18 -1 -18 9 -8 -15 -6 -19 -35 150
 18srRnac 3 -2 13 5 6 5 -3 -1 -6 37 8 -3 -3 7 7 0 182

...
 (data from http://genomics.stanford.edu/yeast/full_data.html)

Transcriptional Regulation of DNA Replication Genes



orange=pre-replication complex genes: mcm2,
 mcm3, cdc46, cdc47, cdc54, cdc6
 blue=replication genes involved in DNA synthesis

Picture taken from
http://genomics.stanford.edu/yeast/additional_figures_link.html

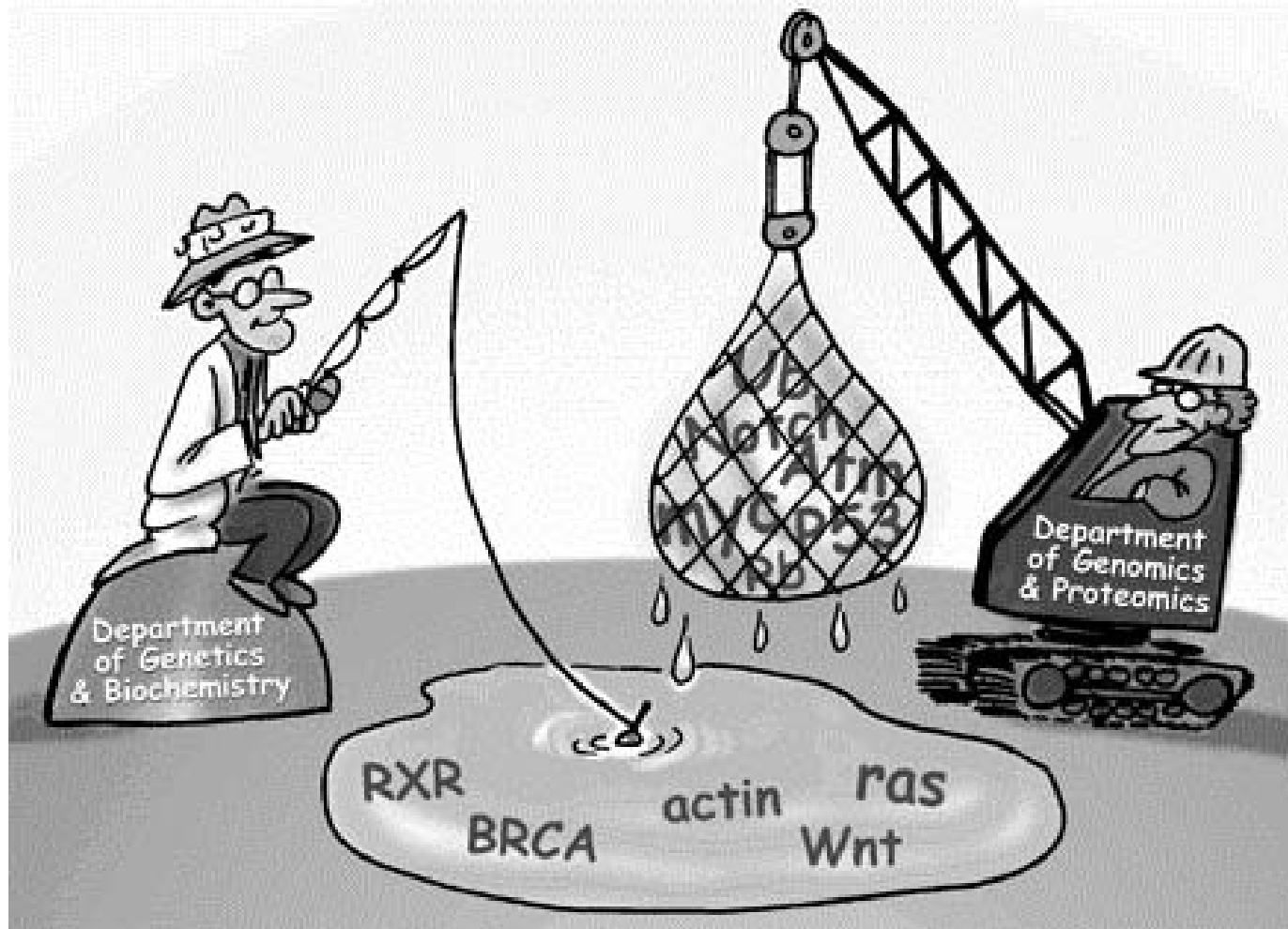
基因體分析的精神

宏觀分析 (global analysis)

=> 得到所有的資訊

(與傳統上“由假說設計實驗”的方法不同)

Traditional analysis vs high throughput analysis



Descriptive vs predictive sciences

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Period																			
1	1 H																	2 He	
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne	
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar	
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr	
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe	
6	55 Cs	56 Ba	*	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Uun	111 Uuu	112 Uub	113 Uut	114 Uuq	115 Uup	116 Uuh	117 Uus	118 Uuo
*Lanthanoids	*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb				
**Actinoids	**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No				

* Picture made from screenshot of <http://www.shef.ac.uk/~chem/web-elements/>
YM-Biochem

資訊驅動之生物醫學研究

好的生物資訊研究不應只為生物醫學研究服務,而要為生物醫學研究開拓新的領域與研究方法,如UniGene 資料庫的建立

另一個例子

Putative alternative splicing site (PALS) database
<http://pals.ym.edu.tw/>

InterPro search Results.

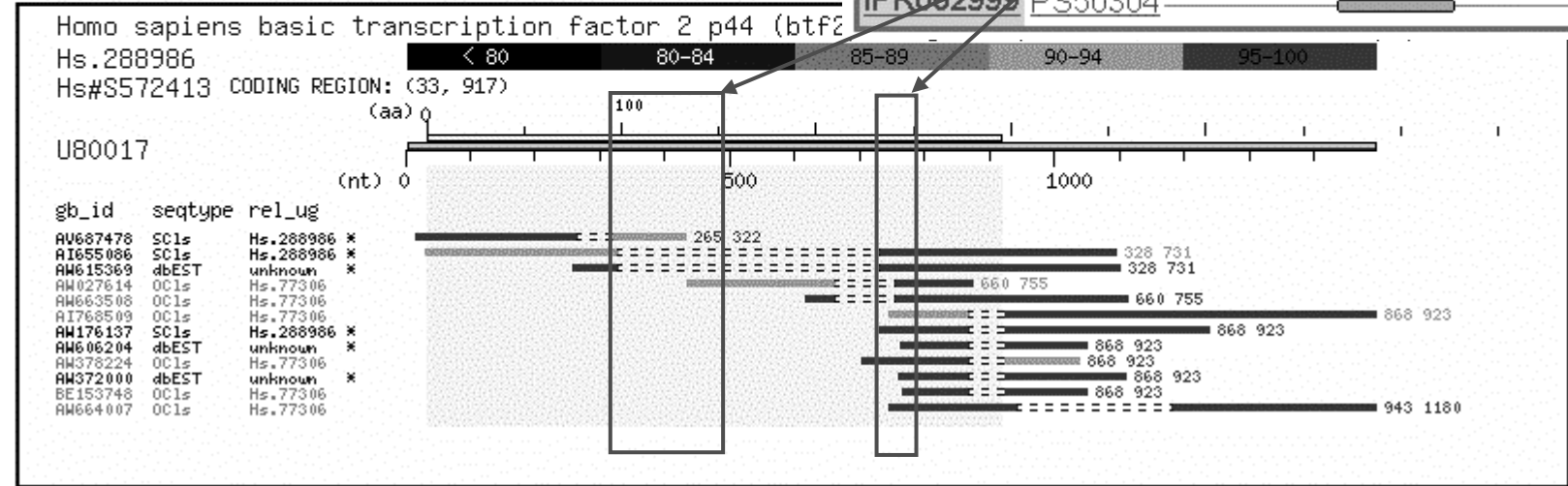
1 Query Sequence submitted Length 294 aa.

InterPro	Results of PPsearch against PROSITE	Results of PFScan against PROSITE	Results of Finger against PRINTS
IPR000694		PS50099 [185-251]	
IPR002999		PS50304 [91-151]	

Welcome to the PASS db Ver. 0.

Created by Huang, Y.H., Chen, Y.T., Lai, J.J. and Yang, U-C

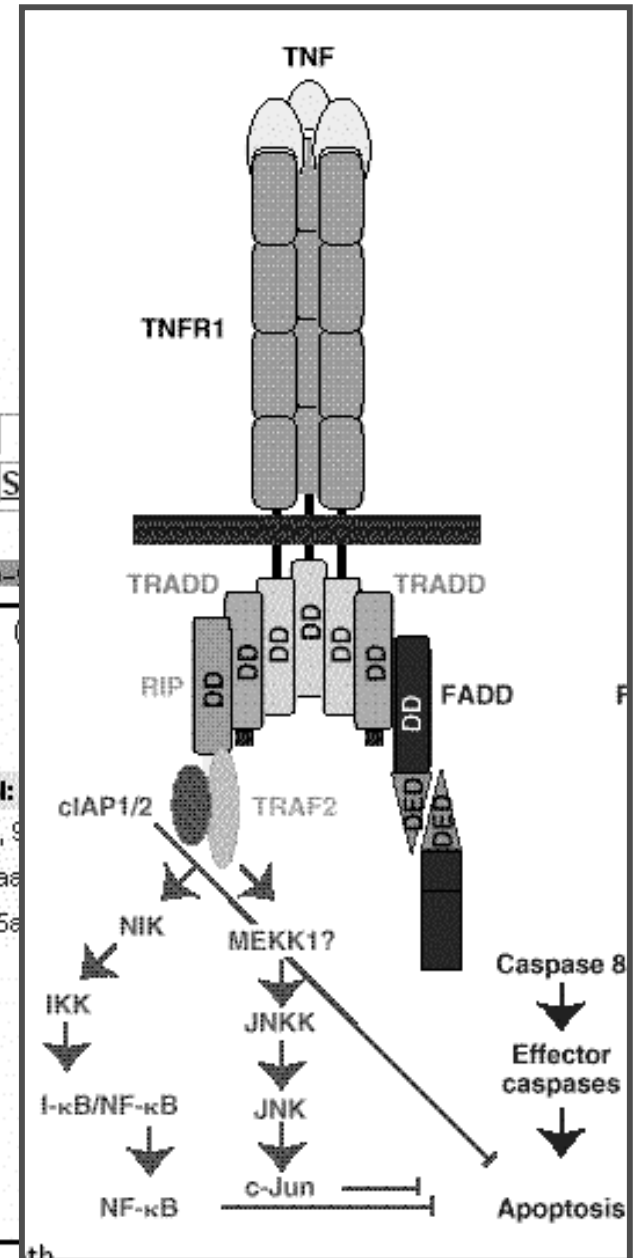
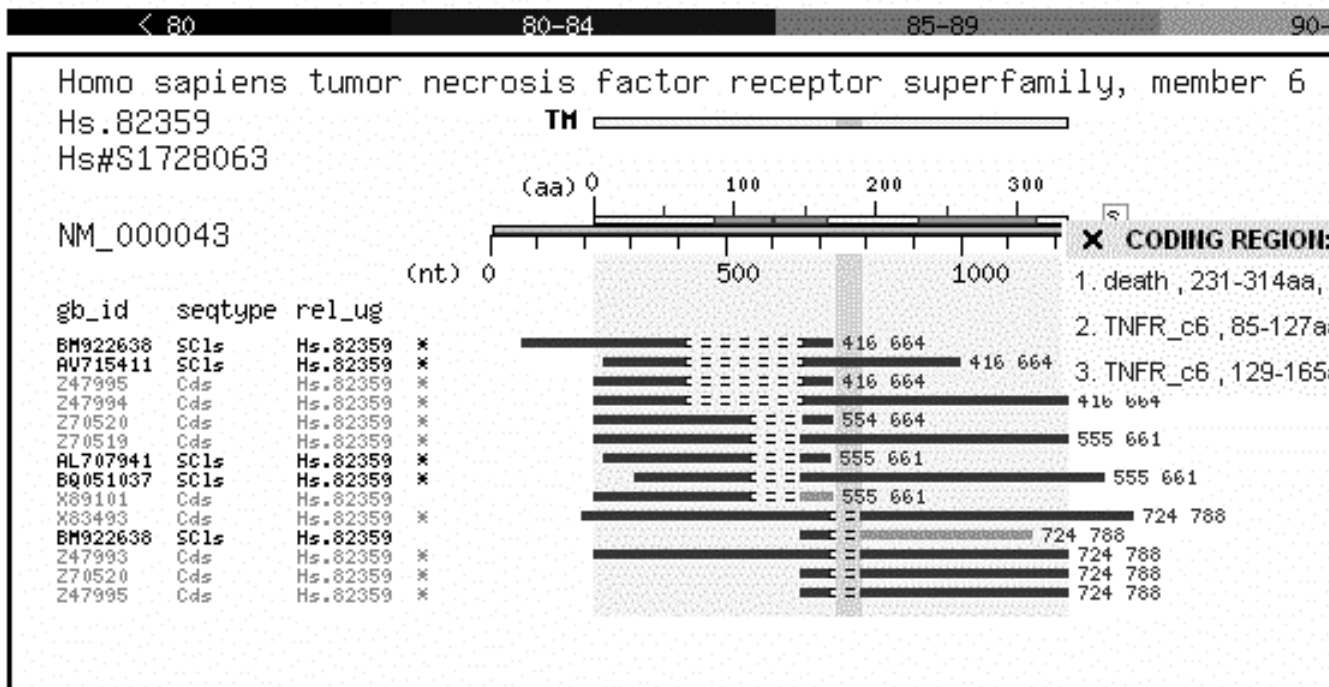
Summary Text **InterPro** Dart CDD dbSNP SAGE Ensembl



“Information integration” may make new discovery

Release 3, Web Interface Ver. 0.9.5.1

[Summary](#)
[Text](#)
[OMIM](#)
[HUGO](#)
[dbSNP](#)
[SAGE](#)
[GeneCards](#)
[Ensembl](#)
[Search again!](#)
[iProClass](#)
[InterPro](#)
[Dart](#)
[CDD](#)
[PSORT](#)
[TMHMM](#)
[NetOGlyc](#)
[ma_mRNA](#)
[Homologs](#)

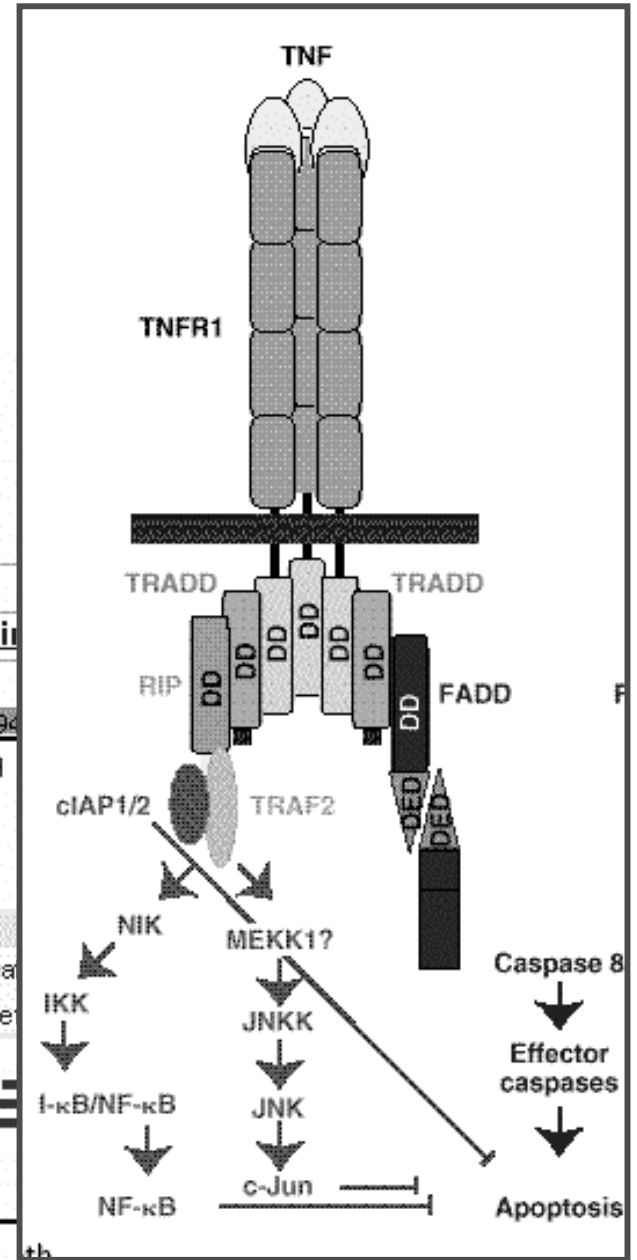
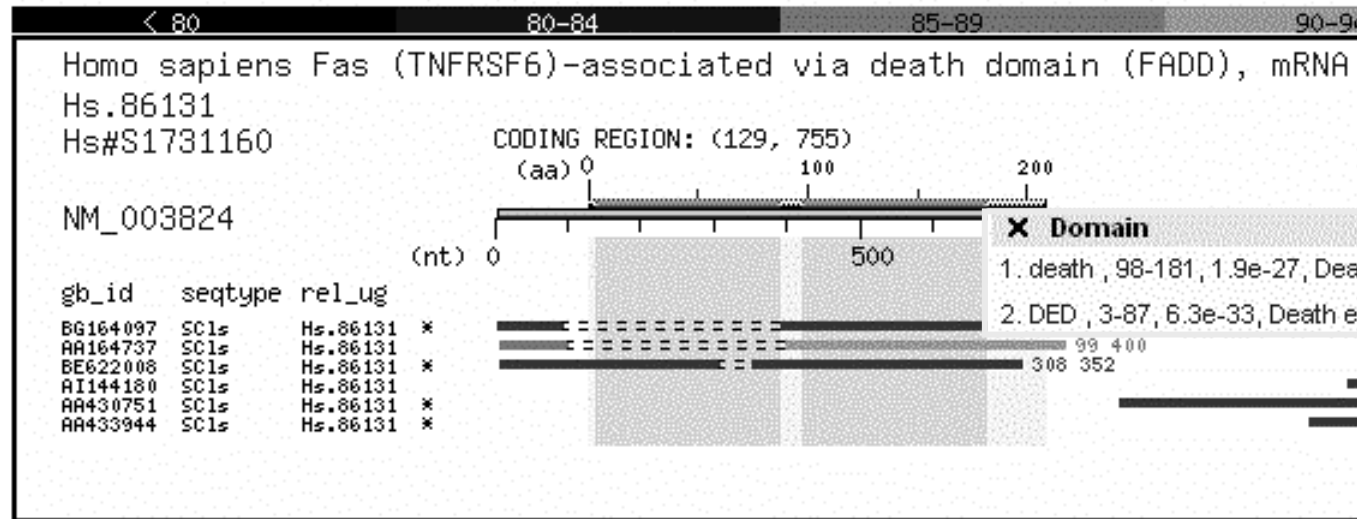


Hypothesis can be made by comparing value-added information

Putative Alternative Splicing Database (PALS db)

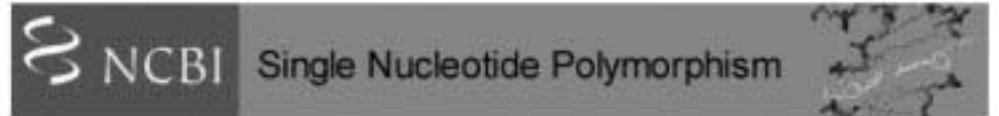
Release 3, Web Interface Ver. 0.9.5.1

Summary	Text	OMIM	HUGO	dbSNP	SAGE	GeneCards	Ensembl	Search again!
iProClass	InterPro	Dart	CDD	PSORT	TMHMM	NetOGlyc	ma_mRNA	Homologs



另一個例子

Putative alternative splicing site (PALS) database
<http://pals.ym.edu.tw/>



SNP List

General

- [SNP Home](#)
- [@SNP Summary](#)
- [How To Submit](#)
- [Genome SNP RFA](#)
- [FAQ](#)

Search Result:

Click on either a local SNP ID or a NCBI Assay ID (ss#) to view the SNP record
 Click "Download" to save the complete query result list (not just this page).
 The default filename is download_list.cgi.
 Overwrite it with any name of your choice, ex.

Welcome to the PASS db Ver. 0.9.4.127

Created by [Huang, Y.H.](#), [Chen, Y.T.](#), [Lai, J.J.](#) and [Yang, U-C](#)

- [Summary](#)
- [Text](#)
- [InterPro](#)
- [Dart](#)
- [CDD](#)
- [dbSNP](#)
- [SAGE](#)
- [Ensembl](#)
- [GeneCards](#)
- [OMIM](#)
- [Help](#)
- [Search again!](#)

ter_snp_id

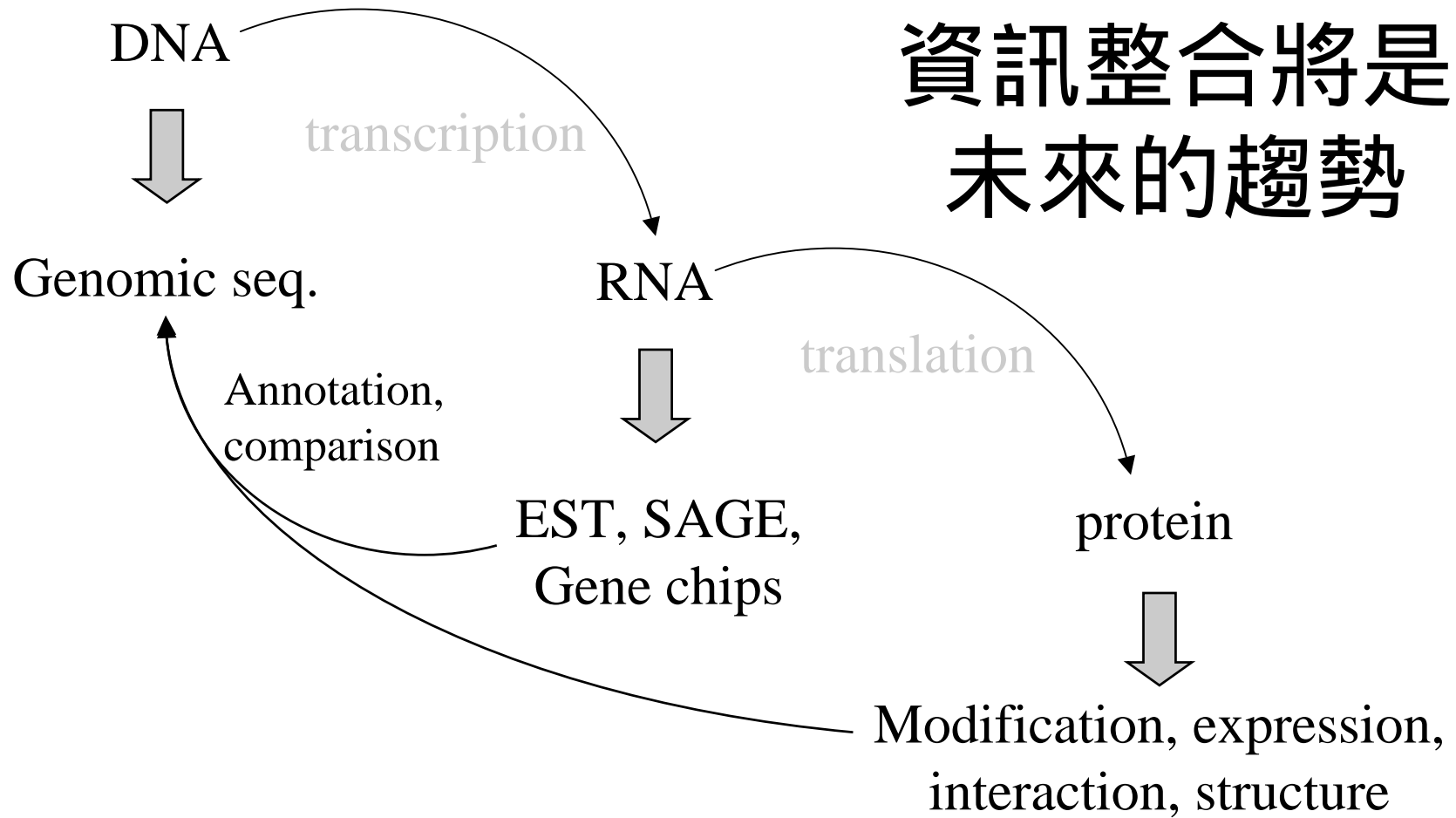
ID	RefSNP Cluster ID(rs#)
9	rs8581
4	rs15545
55	rs363965
15	rs150443
64	rs211479
09	rs152153
30	rs152159
19	rs363910
82	rs377575
60	rs150491

Homo sapiens basic transcription factor 2 p44 (btf2p44) gene, partial cds, neuronal apoptosis inh
 Hs.288986
 Hs#S572413 CODING REGION: (33, 917)

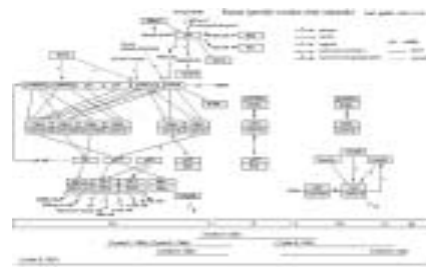
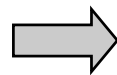
U80017

gb_id	seqtype	rel Ug
AV687478	SC1s	Hs.288986 *
AI655086	SC1s	Hs.288986 *
AW615369	dbEST	unknown *
AW027614	OC1s	Hs.77306
AW663508	OC1s	Hs.77306
AI768509	OC1s	Hs.77306
AW176137	SC1s	Hs.288986 *
AW606204	dbEST	unknown *
AW378224	OC1s	Hs.77306
AW372000	dbEST	unknown *
BE153748	OC1s	Hs.77306
AW664007	OC1s	Hs.77306

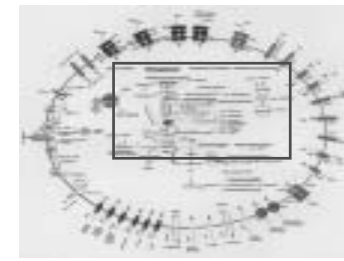
資訊整合將是 未來的趨勢



Components

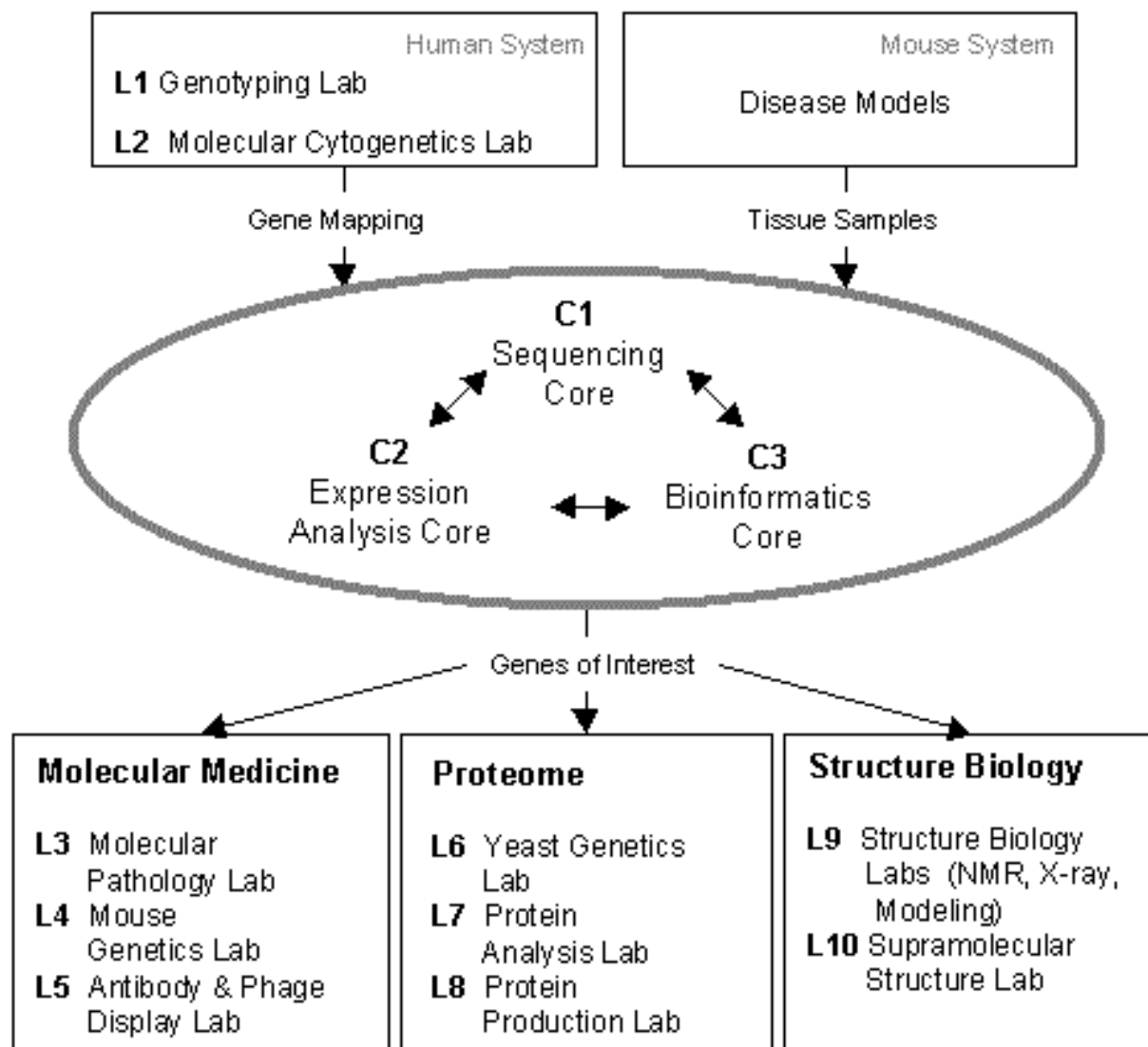


Pathways and regulatory circuits

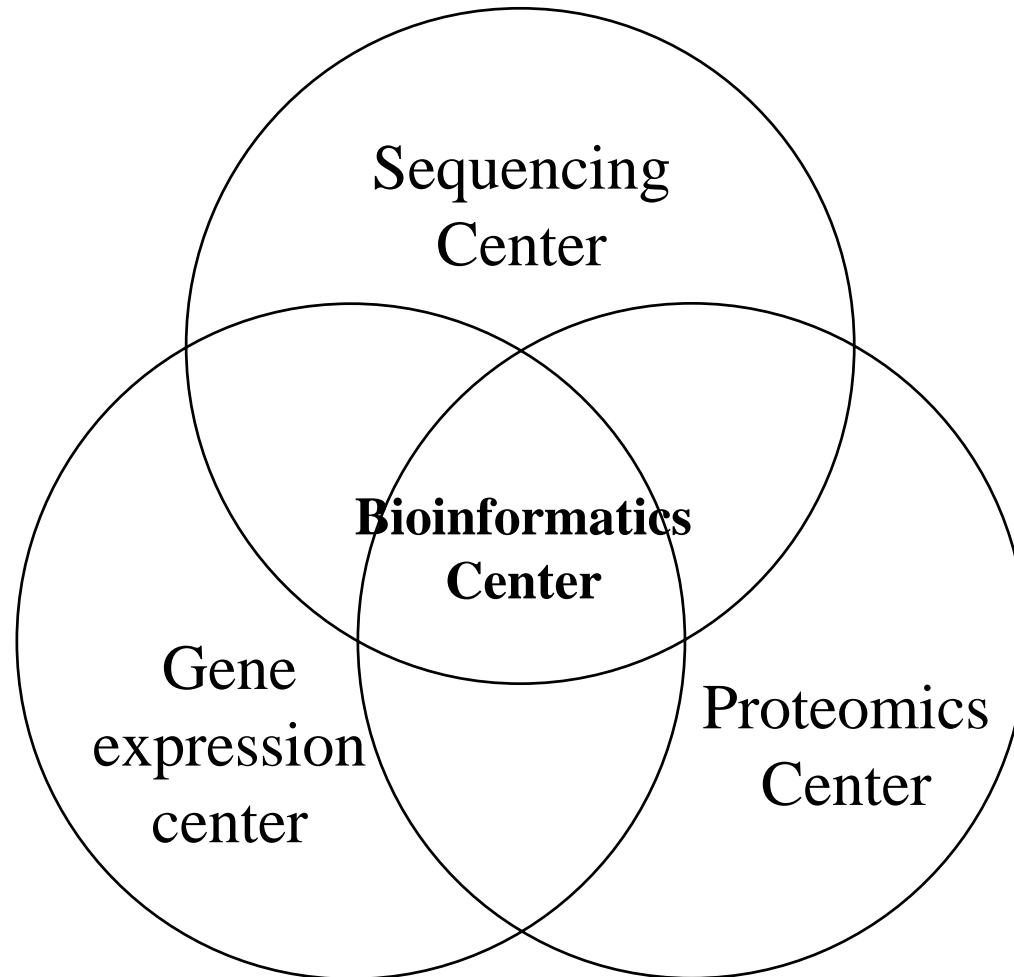


Hypothetical cell

基因體研究的 基礎架構是成 敗的關鍵



生物資訊學須與資訊提供者 結合才能發揮力量



http://binfo.ym.edu.tw/idg/

BioInfoFab



Ymuyang
個人訊息
會員服務
點離會員

全文搜尋
搜尋
即時新聞

站務公告
煩請各位讀者點選上頭 [會員服務] 以了解本站之基本服務項目
若有任何建議請mailto:binfo@ym.edu.tw 聯絡管理者 [陽明生資]
感謝SourceFab所提供的專案計劃- FabulousWeb [先進媒

最新消息



Leroy Hood博士認為五年後 新藥開發成本將劇降

責任編輯: 陽明生資 截稿時間: 2001/03/20 @09:43AM 點閱率: 192
美國系統生物科學研究中心 [Systems Biology] 的總裁Leroy Hood博士認為, 未來生技專家最艱難的工作, 就是要從大量的基因資訊中, 瞭解這些資訊對於人體組織的影響...
[詳細內容... \(討論篇數: 00\)](#)

何大一: 目前世界各國都積極發展生物科技, 台灣現階段應該急起直追

責任編輯: 陽明生資 截稿時間: 2001/03/08 @09:34AM 點閱率: 1096
去年受總統之邀擔任國內生物科技最高顧問的何大一指出, 生物科技的發表對台灣相當重要, 但需要政府相關政策配合...
[詳細內容... \(討論篇數: 05\)](#)

Abbott實驗室製藥部基因體學主任談基因資訊產業

責任編輯: 陽明生資 截稿時間: 2001/03/08 @08:28AM 點閱率: 426
美國亞培 (Abbott) 實驗室製藥部基因體學主任 (哈伯Donald N. Halbert) 指出, 生物資訊學已經成為科學家在基因體序列中, 發掘其價值所在的關鍵之鑰...
[詳細內容... \(討論篇數: 00\)](#)

- [00] 晶基生物晶片技術授權給摩托羅拉代工量產
- [01] 台灣的生物資訊競賽
- [04] 台灣創投對台灣的生技產業均持著有信心的樂觀態度

投票

你有聽過SRS嗎?

- 聽過, 而且用過
- 聽過但是沒用過
- SRS?是一個新的影片分級等級嗎??

投下 [目前結果 | 辦個投票]

目前票數: 38
討論篇數: 1

常駐論壇

- [127] 生物資訊在哪裡?
- [054] 資訊學園地
- [053] 生資名詞知多少?
- [046] GCG套裝軟體
- [043] 生物技術交流
- [033] 生資工具使用經驗交流
- [023] 一千零一個為什麼?
- [022] 演講及研討會