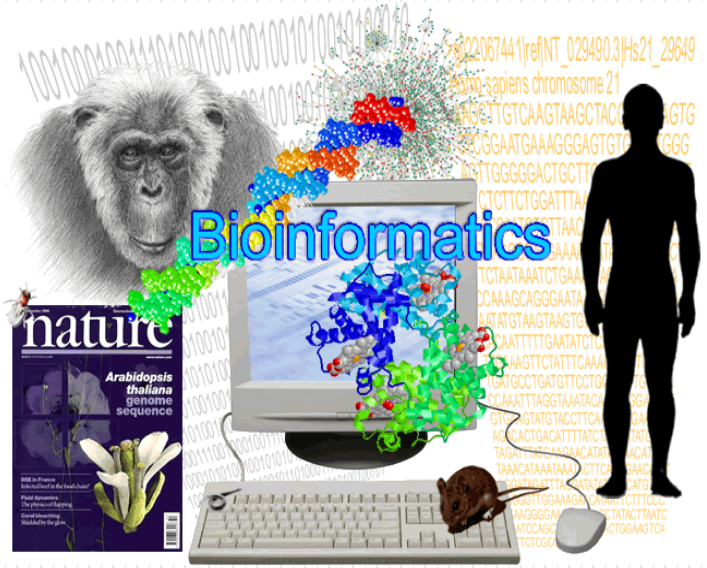
Bioinformatics & Functional Genomics



薛 佑 玲 PhD

Institute of Biomedical Sciences

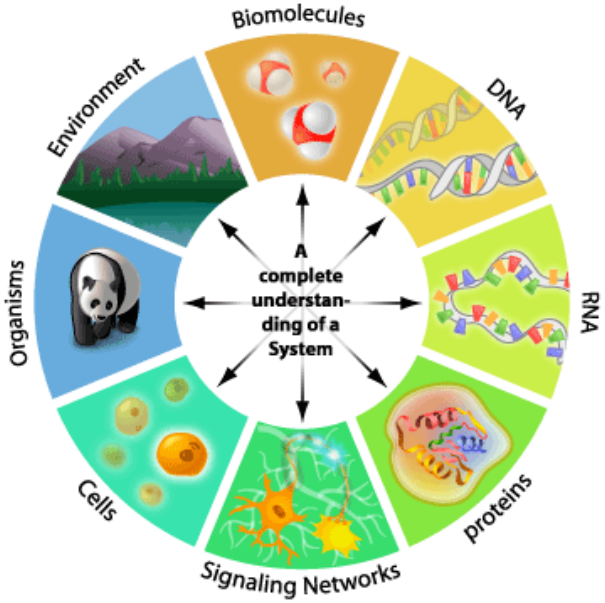
National Sun Yat-sen University

ylshiue@mail.nsysu.edu.tw

Outline

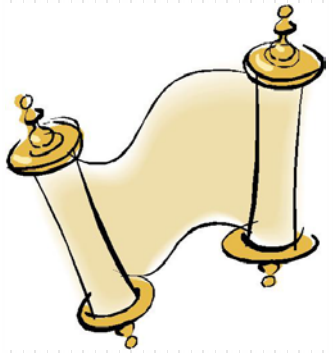
Introduction: a Short History About Bioinformatics

Bioinformatics Q & A

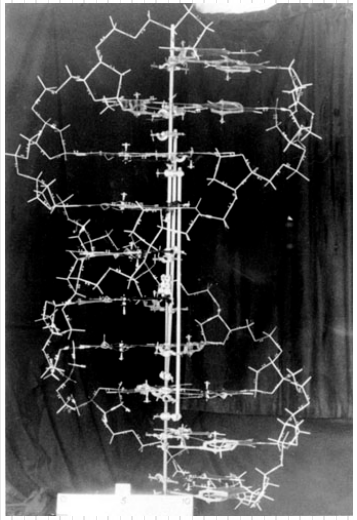


b101nf0rmat1cs

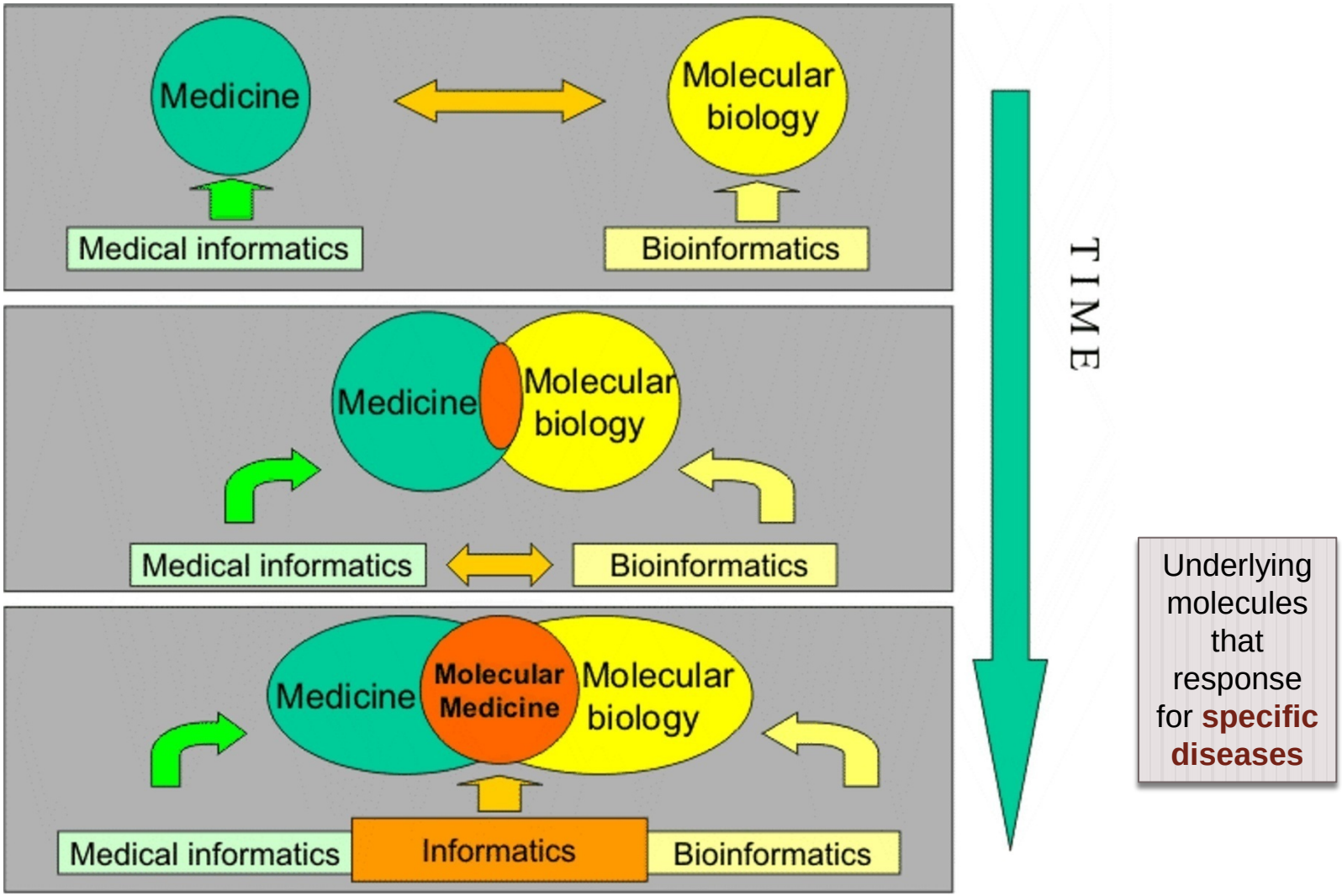
ACCATGGATTACATA011011011000110101O
GATTCCATTATAAGGA011001110000000100
TGCCGGCAATAGGCAA011101010001101O1
CAATAAGCATTCCACO01010101101011011



A Short History about Bioinformatics



The Convergence between MI & BI



Top Ten Medical Breakthroughs – since 1840

- Hygiene equipment
- Antibiotics
- Anesthetic
- Vaccine
- Discovery of DNA structure
- Microbiology theory
- 'The Pill': the combined oral contraceptive pill
- Evidence-based Medicine
- Medical imagining（e.g., X-ray, MRI…)
- **Computer**
- **Stem cell therapy**

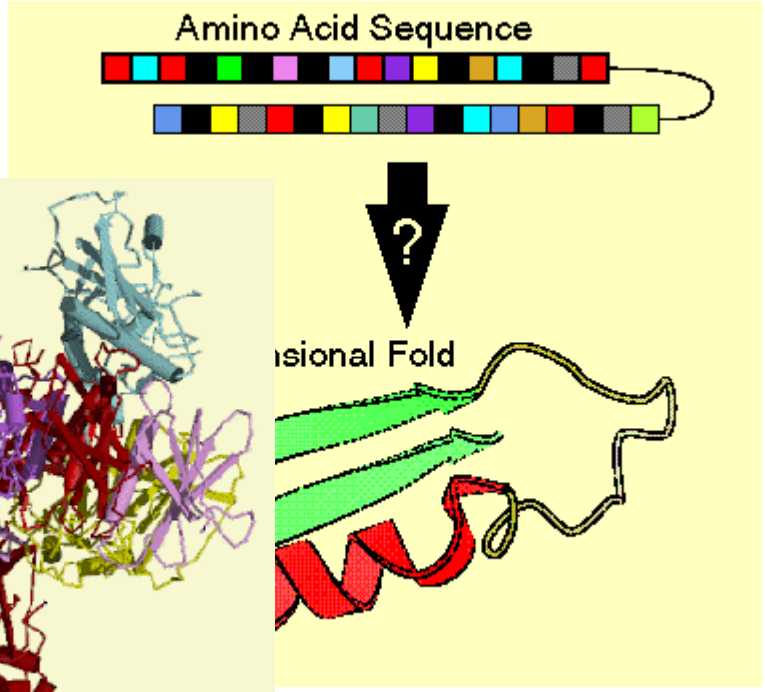
根據British Medical Journal 線上意見調查，
自1840年創刊以來，最重要的醫學里程碑

Day 4: Computer Science and Medicine

CSedweek 11 部影片 訂閱

The Holy Grail of Bioinformatics

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN



…to be able to understand **the words in a sequence sentence** that form a particular protein **structure** (from Attwood & Parry-Smith 1999)

A Short History Overview (I) - Wet

1953: Double helix of DNA (Waston & Crick)

1954: First protein sequence (**insulin** by **Sanger**)

1958: First X-ray 3D structure of a protein (**myoglobin** by Kendrew)

1972: First DNA sequencing

1977: Rapid **sequencing** techniques (**Gilbert & Sanger**)

1986: PCR (the photocopying machine of the biologist)

1992: Sequence of **yeast** chromosome III ($3*10^5$ bp)

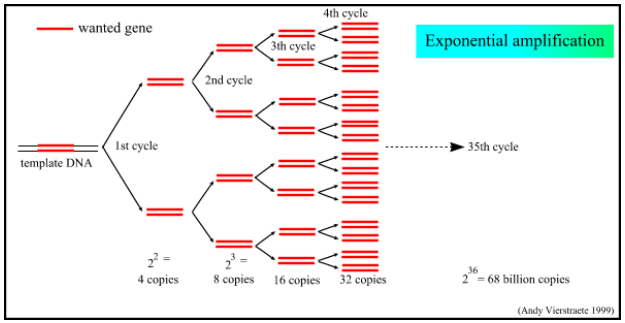
1995: Sequence of the genome of the bacteria: ***Haemophilus influenzae*** ($2*10^6$ bp)

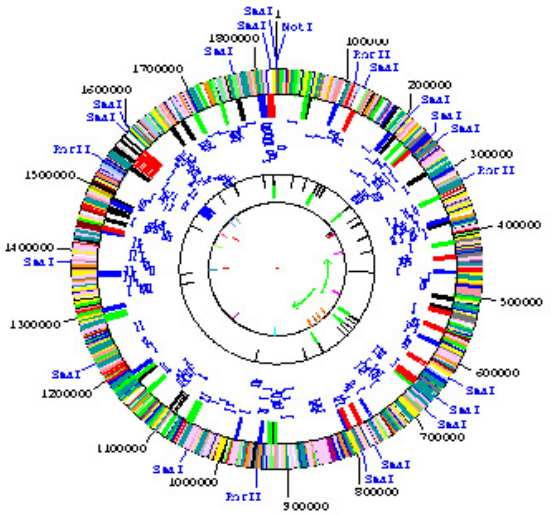
1999: Sequence of the genome of **a multi-cellular organism**: *Caenorhabditis elegans* (10^8 bp)

2000: Blue draft of the **human genome** ($3*10^9$ bp)

2002: Genome of *Ashbya gossypii* ***(Saccharomycetes)***

Recent: GOLD database





A Short History Overview (I) - Dry

1965: «Atlas of protein sequence and structure» (**Dayhoff**)

1967: Fitch WM (Phylogenetic trees)

1970: **Needleman/Wunsch (1st similarity search algorithm)**

1971: PDB (3D structure database)

1977: **Staden (1st sequence analysis software suite)**

1980: **EMBL Heidelberg**

1980: Smith/Waterman algorithm

1982: EMBL Nucleotide Sequence Database and GenBank

1985: CABIOS (1st scientific journal for bioinformatics)

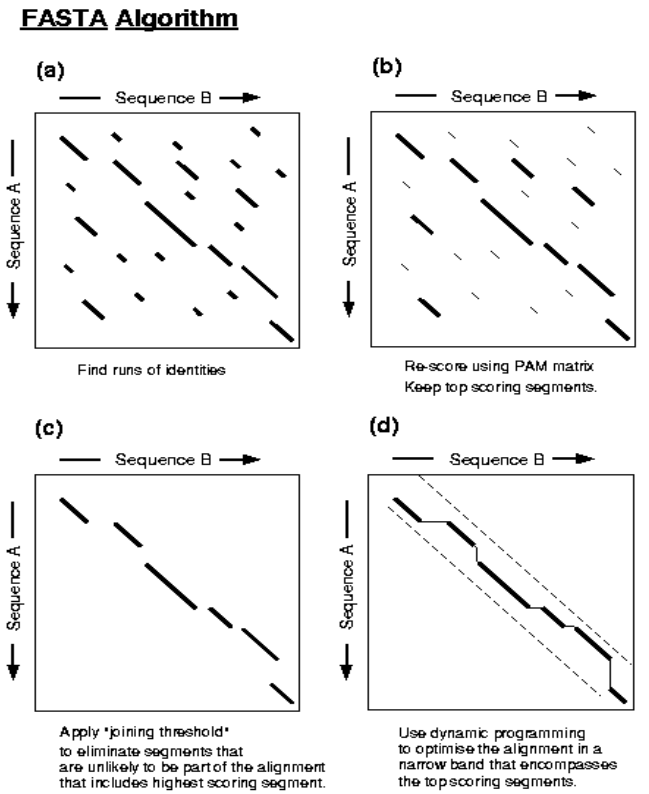
1985: FASTP (ancestor of **FASTA**, Blast, etc.)

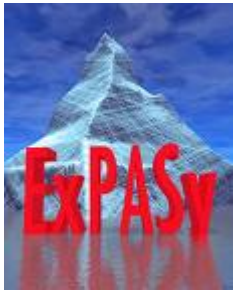
1986: Swiss-Prot (Protein Sequence Database)

1988: Creation of the NCBI in the USA

1992: EBI founded as EMBL outstation in **Hinxton** (Wellcome Trust Campus)

1993: **ExPASy** (1st WWW server for the life sciences)…





Early Bioinformatics: the birth of a discipline – Quzounis CA & Valencia A (2003)

Table 2. Twenty Publications that influenced our view of bioinformatics

Publication	Comments
Zuckerkandl and Pauling, 1965b	First use of molecular sequences for evolutionary studies
Fitch and Margoliash, 1967	Use of molecular sequences to build trees
Needleman and Wunsch, 1970	First implementation of dynamic programming for protein sequence comparison
Lee and Richards, 1971	Calculation of accessibility on protein structures
Chou and Fasman, 1974	First secondary structure prediction method
Tanaka and Scheraga, 1975	Simulation of protein folding
Dayhoff, 1978	First collection of protein sequences
Hagler and Honig, 1978	One of the first explicit attempts to simulate protein folding
Doolittle, 1981	Seminal paper examining divergence and convergence in protein evolution
Felsenstein, 1981	One of the first statistical treatments of evolutionary tree construction
Richardson, 1981a	The most comprehensive description of protein structure to that date
Kabsch and Sander, 1984	Discovery with profound implications for model building by homology and structure prediction
Novotny *et al.*, 1984	The inability of distinguishing correct from incorrect structures threw back structure prediction approaches for a long while
Chothia and Lesk, 1986	Examination of divergence between sequence and structure
Doolittle, 1986	Influential book on sequence analysis
Feng and Doolittle, 1987	The first approach for an efficient multiple sequence alignment procedure, later implemented in CLUSTAL
Lathrop *et al.*, 1987	One of the first applications of Artificial Intelligence in protein structure analysis and prediction
Ponder and Richards, 1987	The very first threading approach, using sequence enumeration
Altschul *et al.*, 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie *et al.*, 1991	The first implementation of protein structure prediction using threading

Bioinformatics: A Snapshot 10 Years Ago

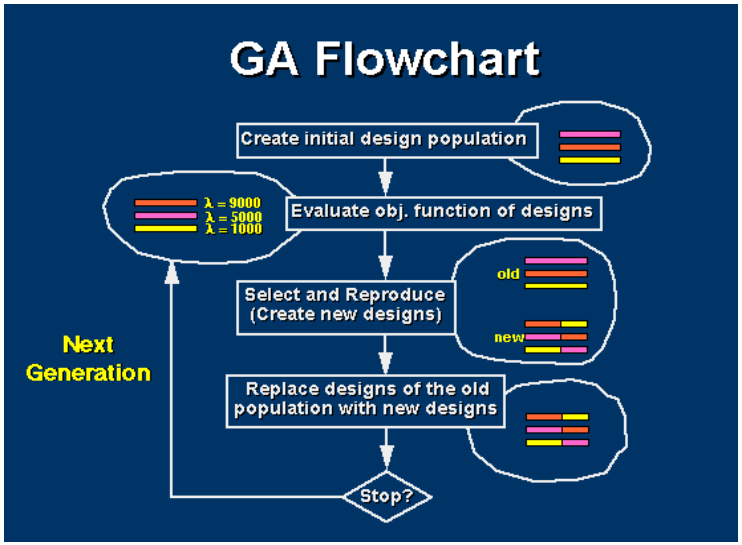
Pharmaceutical companies were **not interested**

Life scientists believed that it was an **outlet** for **failed biologists** that want to play around with computers

Computer scientists did not even consider it important, they confused it with **bio-inspired** **"computer sciences"**

E.g., **genetic algorithm**, artificial life, **ant algorithm**, neural network

DNA computers…



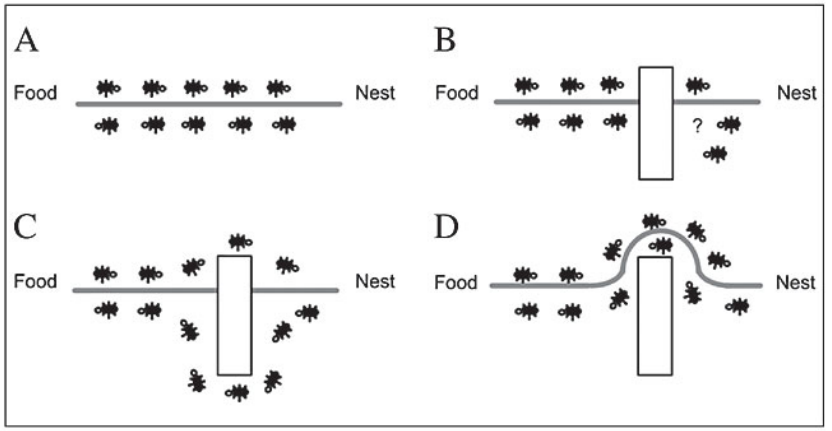


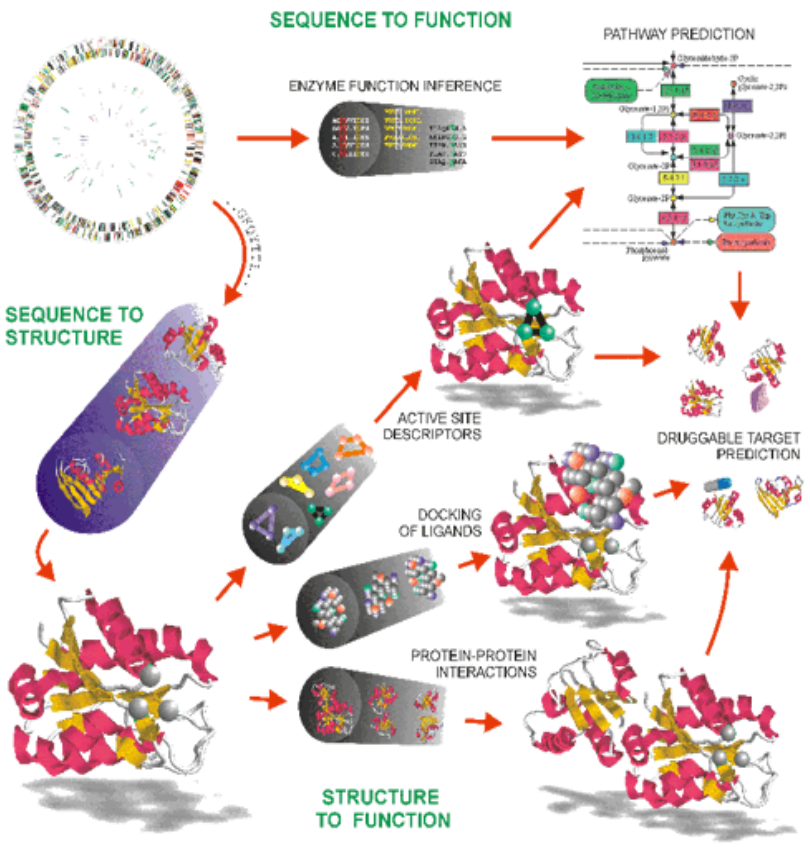
Figure 2. A. Ants in a pheromone trail between nest and food; **B.** an obstacle interrupts the trail; **C.** ants find two paths to go around the obstacle; **D.** a new pheromone trail is formed along the shorter path.

Bioinformatics in 2003

Pharmaceutical companies believe that it is **the most efficient way** to streamline the process of **drug discovery**

Some life scientists believe it is **the solution to all problems in life sciences** and that it will allow them **to avoid** doing **some experiments**

Computer scientists are very interested: **the scope and complexity** of the domain makes it the ideal field of application of **new software techniques** and specialized hardware developments

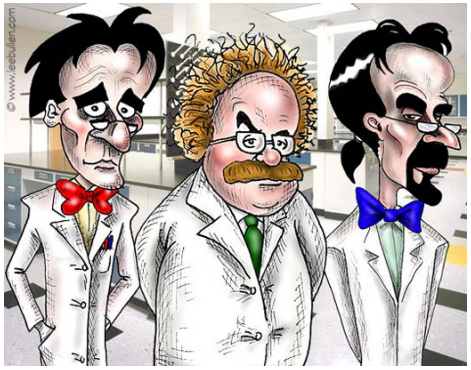


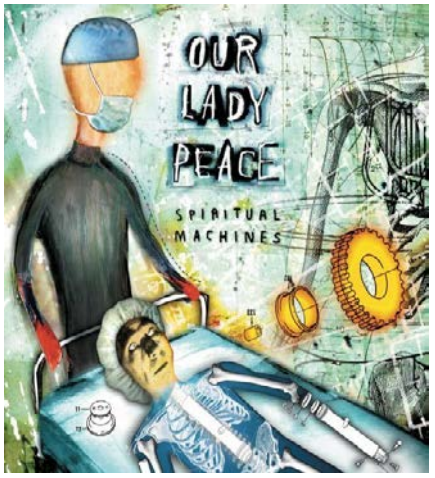
Bioinformatics after 2010

Pharmaceutical companies use it **routinely**, but have realized that it **complements** rather than **replaces** experimental work

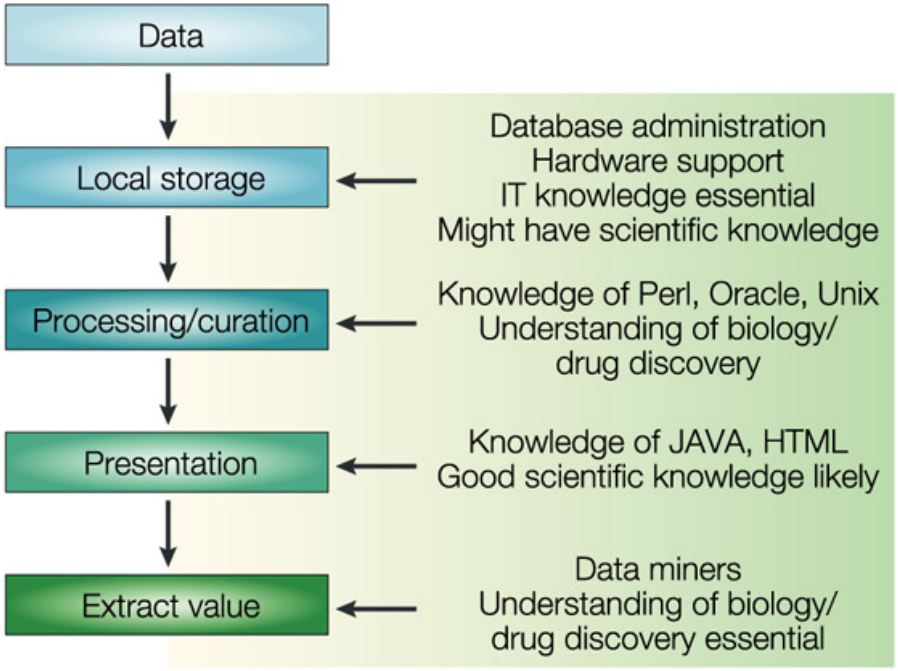
Life scientists use it **efficiently every day** and therefore **forget that it exists**

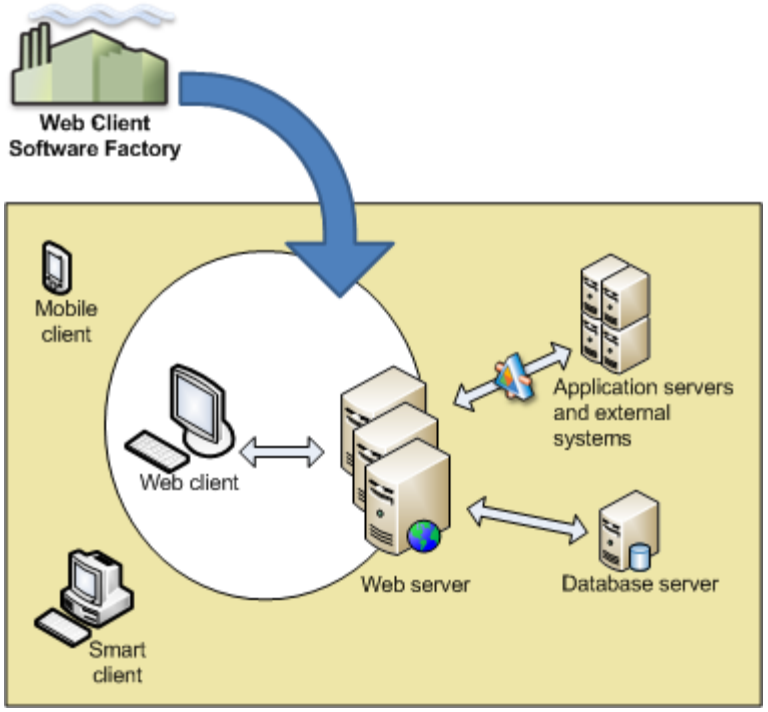
Computer scientists may have jumped on **another fancy subject:** Spiritual machines?





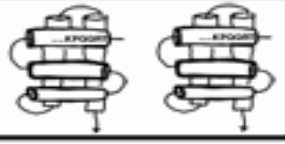
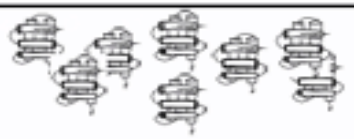
Resources: databases & software





Nature Reviews Drug Discovery 3, 281 (2004)

Depth: Rational Drug Design (physics)→

		Breadth: Homologs, Large-scale Surveys, Informatics–			
			pairwise comparison, sequence & structure alignment	multiple alignment, patterns, templates, trees	databases, scoring schemes, censuses
		1	2	3-100	100+
	Genome Sequence	atcgatcgatatttgggatttgggga	atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga	atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga	atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga atcgatcgatatttgggatttgggga
gene finding	↓				
	Protein Sequence	ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT
structure prediction	↓				
	Protein Structure				
geometry calculation	↓				
	Protein Surface				
molecular simulation	↓				
	Force Field				
structure docking	↓				
	Ligand Complex				

"Don't just sit there! If you've processed all the data there is, go out and find more data!"

Reproduced in R.L. Weber, "*A random walk in science*", IOP Publishing, 1973

Case Study

Case Study

1: Cell. 2003 May 30;113(5):631-42.

The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.

Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S.

Laboratory of Animal Molecular Technology, Research and Education Center for Genetic Information, Nara Institute of Science and Technology, Nara 630-0192, Japan.

Embryonic stem (ES) cells derived from the inner cell mass (ICM) of blastocysts grow infinitely while maintaining pluripotency. Leukemia inhibitory factor (LIF) can maintain self-renewal of mouse ES cells through activation of Stat3. However, LIF/Stat3 is dispensable for maintenance of ICM and human ES cells, suggesting that the pathway is not fundamental for pluripotency. In search of a critical factor(s) that underlies pluripotency in both ICM and ES cells, we performed in silico differential display and identified several genes specifically expressed in mouse ES cells and preimplantation embryos. We found that one of them, encoding the homeoprotein Nanog, was capable of maintaining ES cell self-renewal independently of LIF/Stat3. nanog-deficient ICM failed to generate epiblast and only produced parietal endoderm-like cells. nanog-deficient ES cells lost pluripotency and differentiated into extraembryonic endoderm lineage. These data demonstrate that Nanog is a critical factor underlying pluripotency in both ICM and ES cells.

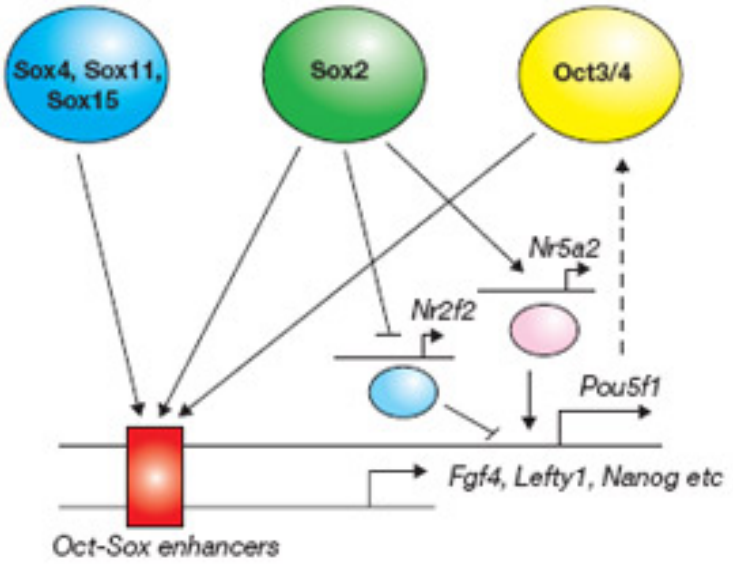
PMID: 12787504 [PubMed - indexed for MEDLINE]

Introduction

LIF/gp130/Stat3 are **not** fundamental for **pluripotency** & predict the existence of **a novel pathway(s)** that maintains pluripotency in both ICM & ES cells

Objective: To identify the **LIF/Stat3-independent factor(s)** that underlies **pluripotency** in both ICM & ES cells

To this end, **DDD** identified genes **expression** in ES cells as **specifically** as ***oct3/4***

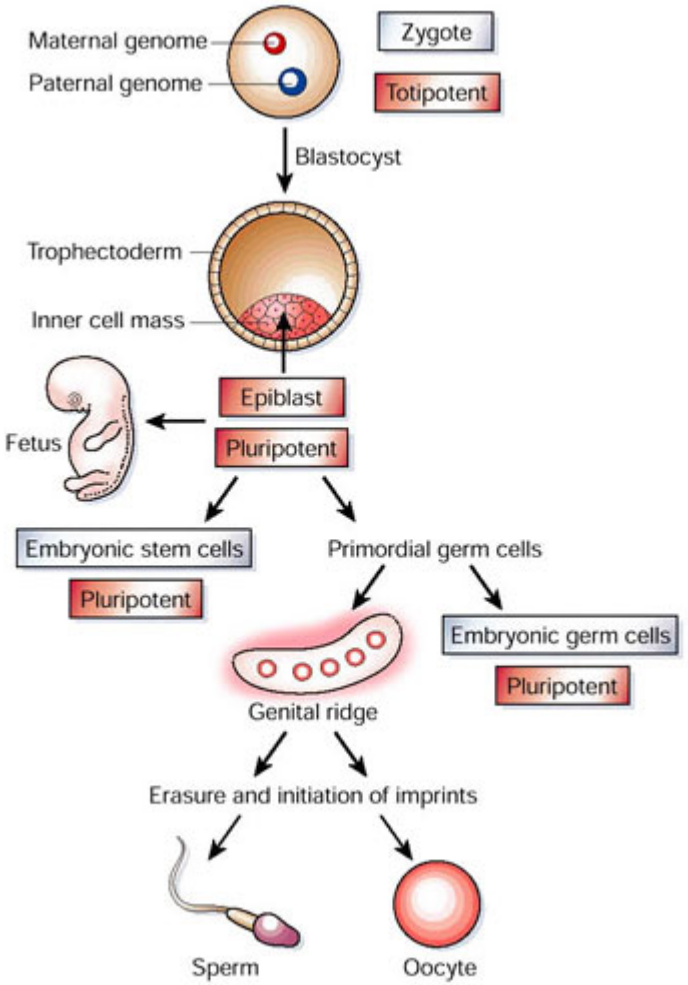


Results

Identification of ***ecat*** by **DDD**

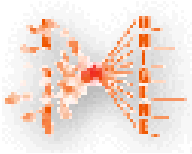
To identify candidate of the LIF/Stat3-independent factor(s) essential for **pluripotent cells**, DDD was performed to **compare expressed sequence tag** (EST) libraries from mouse ES cells & those from **various somatic tissues**

A number of genes were found **overrepresented in ES cell-derived libraries** (Table, next slide)



Data-Mining

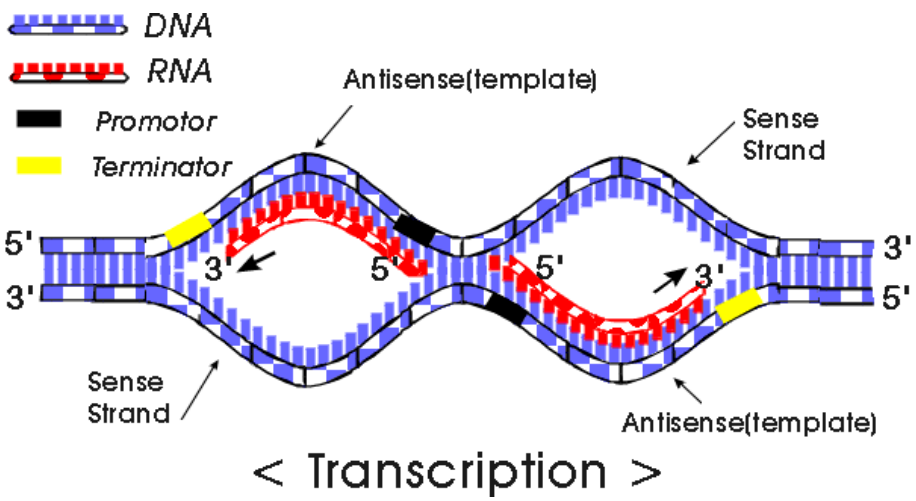
The Cancer Genome Anatomy Project (CGAP) - aims to decipher the molecular anatomy of cancer cells. CGAP develops profiles of cancer cells by comparing gene expression in normal, precancerous, and malignant cells from a wide variety of tissues.

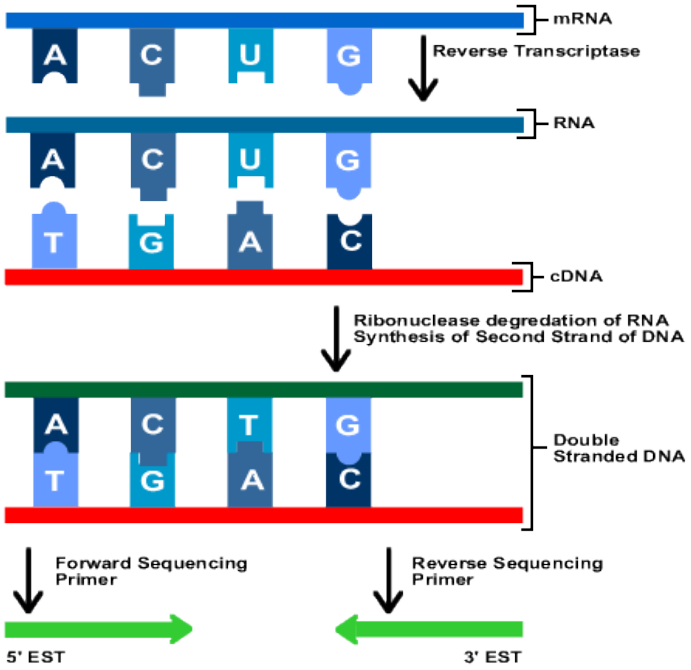


UniGene DDD - Digital Differential Display - an online tool to compare computed gene expression profiles between selected cDNA libraries. Using a statistical test, genes whose expression levels differ significantly from one tissue to the next are identified and shown to the user. Additional information about UniGene is above, including a list of organisms represented.

http://www.ncbi.nlm.nih.gov/Tools/

DDD Basics (1)


DNA
RNA
Promotor
Terminator
Antisense(template)
Sense Strand
5'
3'
3'
5'
5'
3'
3'
5'
Sense Strand
Antisense(template)
< Transcription >


mRNA
A C U G
Reverse Transcriptase
RNA
A C U G
T G A C
cDNA
Ribonuclease degredation of RNA
Synthesis of Second Strand of DNA
A C T G
T G A C
Double Stranded DNA
Forward Sequencing Primer
Reverse Sequencing Primer
5' EST
3' EST

DDD Basics (2)

The **foundation** of DDD

UniGene

A tool for **comparing** **EST-based expression profile** among the various libraries, or pools of libraries, represented in **UniGene**

Between %

DDD employs a **statistical** **method of comparison**

The Fisher Exact Test (Conditional chi-square test); **N < 40 or theoretical value <1**

Output restriction

Statistically significant differences (***P* ≤ 0.05**)

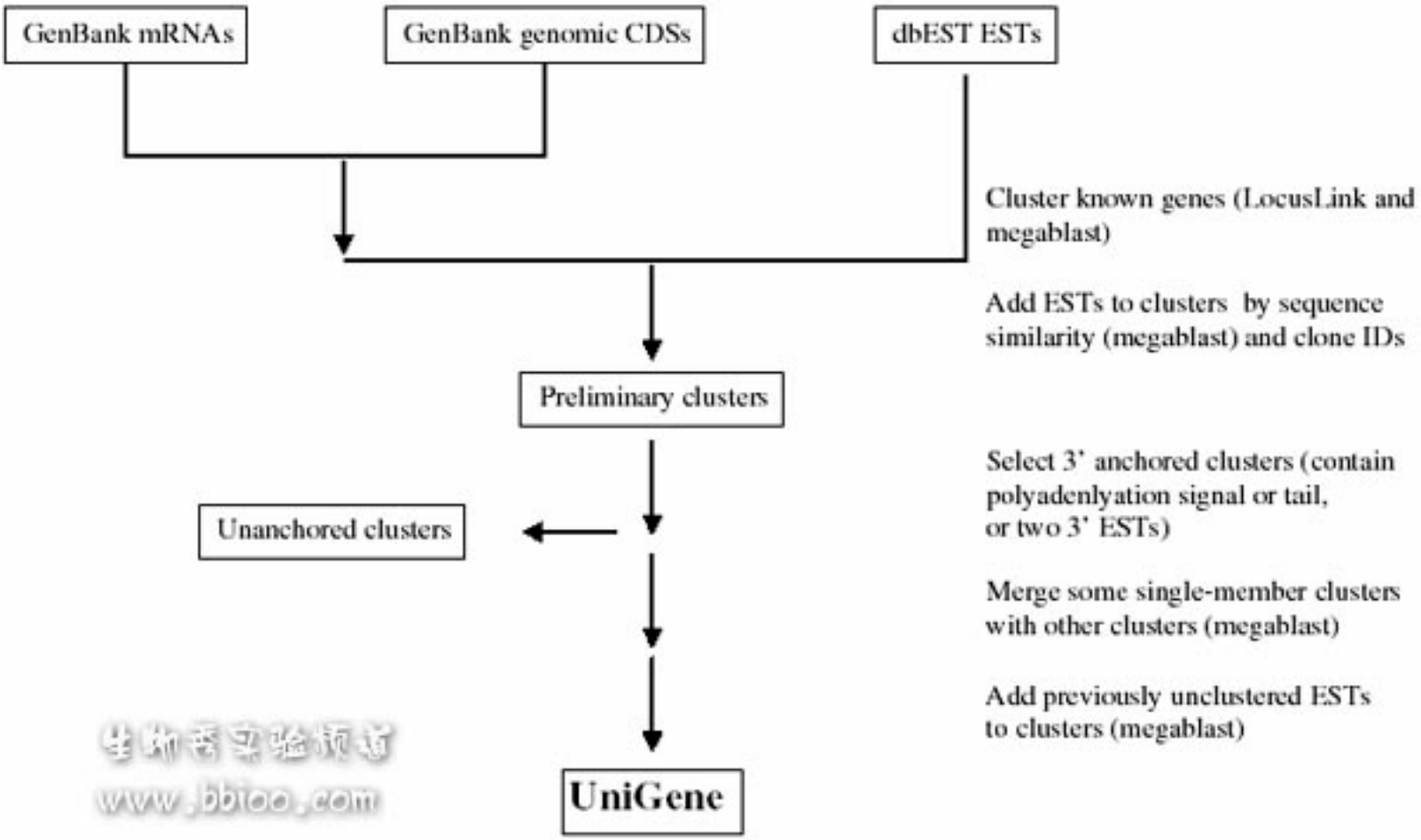
Libraries restriction

Only those with **over 1,000 sequences** in UniGene are included in DDD

References

Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: **The NCBI Handbook**. Bethesda (MD): National Center for Biotechnology Information; 2003.

UniGene



For any given of **pool size (N, M)** and **gene counts (c and C),** the **probability** of the table being generated by chance is calculated where

$$p = [N!M!c!C!]/[(N+M)!a!b!A!B!)$$

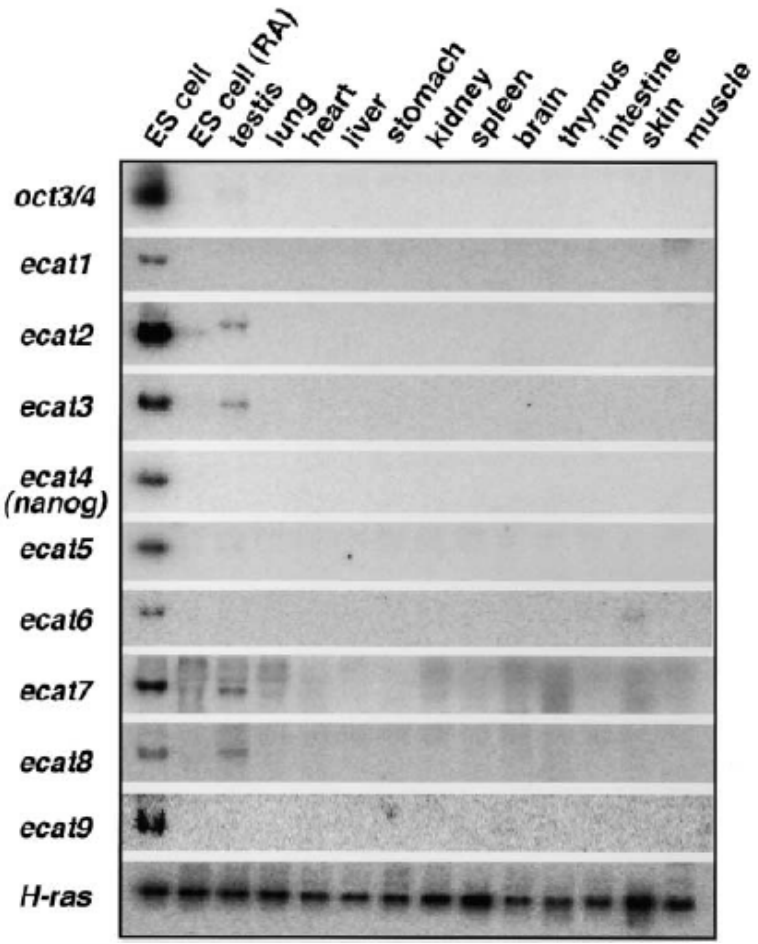
The Fisher Exact Test

In the context of DDD, the relevant 2 X 2 tables are of the form:

	Gene1	All Other Genes	Total
PoolA	a = # sequences in poolA assigned to Gene1	A = # sequences in poolA NOT assigned to Gene1	N = a + A total number of sequences in PoolA
PoolB	b = # sequences in poolB assigned to Gene1	B = # sequences in poolB NOT assigned to Gene1	M = b + B total number of sequences in PoolB
Total	c = a + b	C = A + B	N + M = c + C

Northern Blot Analysis

Unigene	Frequency ES/EC cell	Frequency Others	Symbol	Symbol NEW
Mm.361968	0.00295	0	ecat2	Dppa5
Mm.10205	0.0028	0	utf1	Utf1
Mm.5090	0.00079	0	cripto	Tdgf1
Mm.157658	0.00076	0	ecat1	2410004A20Rik
Mm.128134	0.00076	0	hnRNB-G	1700012H05Rik
Mm.5180	0.00073	0	Nr0b1/dax-1	Nr0b1
Mm.6047	0.00067	0	ecat4/nanog	Nanog
Mm.28369	0.00049	0	ecat3/fbx15	Fbxo15
Mm.17031	0.0004	0	oct3/4	Pou5f1
Mm.285848	0.00033	0	Zfp42/rex1	Zfp42
Mm.45676	0.00033	0	EST	LOC433110
Mm.258773	0.00033	0	ecat5/ERas	Zmynd11
Mm.23310	0.0003	0	zfp296	Zfp296
Mm.18154	0.00027	0	tcl1	Tcl1
Mm.13433	0.00027	0	ecat7	Dnmt3l
Mm.47904	0.00027	0	ecat8	2410004F06Rik
Mm.299742	0.00024	0	ecat9	Gdf3
Mm.913	0.00024	0	brachyury (T)	T
Mm.256916	0.00024	0	tex20	Sall4
Mm.158190	0.00021	0	ecat6	2410039E07Rik



ecat, for ES cell associated transcripts

	Pool	Lib ID(s)	Clustered ESTs
Edit...	A.ES cells	2512	15303
Edit...	B. Adult Pool	2581, 5369, 5354, 2518, 2602, 2509, 5393, 2591, 2607, 2606, 2590, 9946, 5430, 2513, 5360, 9974, 7215, 9742, 7216, 2570, 2571, 7222, 5429, 2562, 5390, 7218, 5361, 7219, 2551, 4140, 12264, 5352, 5357, 9952	223254

New...

Statistically Significant Differences

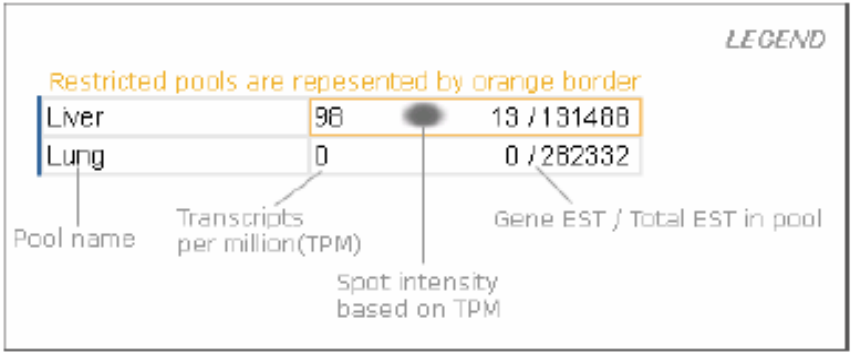
A.	B.	Gene index	Gene description
	990	Mm.99395	Seminal vesicle protein, secretion 2 (Svs2)
0.00000	0.00443		
A<B	B>A		
55	170	Mm.336743	Heat shock protein 8 (Hspa8)
0.00359	0.00076		
A>B	B<A		
41		Mm.10205	Undifferentiated embryonic cell transcription factor 1 (Utf1)
0.00268	0.00000		
A>B	B<A		
39	87	Mm.28222	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39 (Ddx39)
0.00255	0.00039		
A>B	B<A		
32		Mm.361968	Developmental pluripotency associated 5 (Dppa5)
0.00209	0.00000		
A>B	B<A		

Breakdown by Tissue

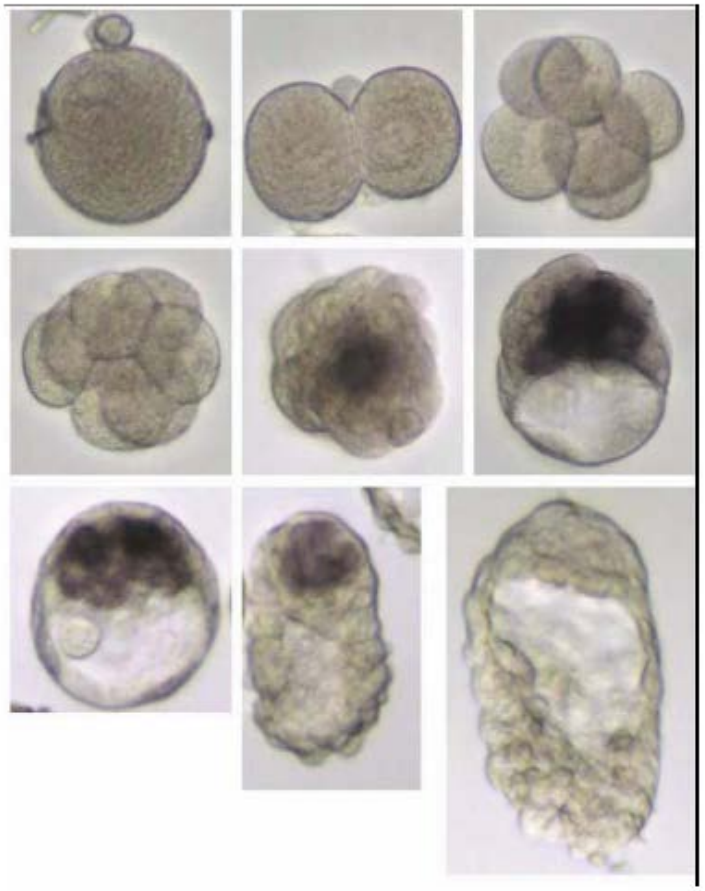
bone	0		0/38928
bone marrow	0		0/37245
brain	0		0/486112
colon	0		0/52004
eye	0		0/169091
heart	0		0/53082
kidney	0		0/116533
liver	0		0/104225
lung	0		0/43564
lymph node	0		0/25515
mammary gland	0		0/348007
muscle	0		0/19349
ovary	0		0/14904
pancreas	0		0/80182
placenta	0		0/32674
pituitary g...	0		0/43777
skin	0		0/83521
spleen	0		0/69103
stomach	0		0/31299
testis	19		2/102659
thymus	0		0/99645
uterus	0		0/6588

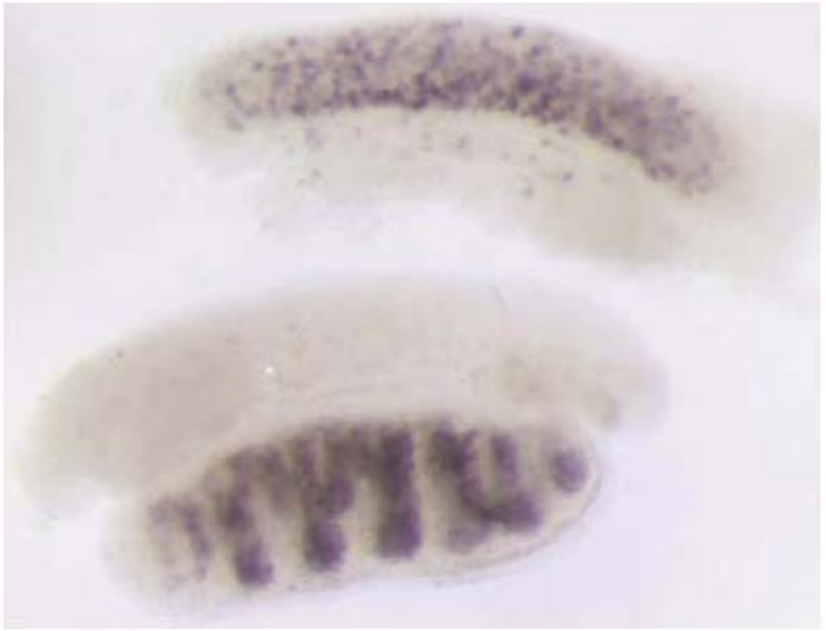
Breakdown by Developmental Stage

egg	0		0/23603
pre-implant...	295		46/155645
post-implan...	16		1/60985
mid-gestati...	35		15/422752
late-gestat...	0		0/218646
neonate	0		0/57179
post natal	0		0/68404
adult	0		0/848043



Expression of *Nanog* in Vivo





E11.5 **genital ridges** from female (top) & **male** (bottom)

Preimplantation embryos. Top: embryos of **1, 2, and 6 cells**. Middle: 8-cell embryo, late morula & early blastocyst, **bottom**: blastocysts at expanded, hatched & **implanting** stages

Pou5f1 (oct3/4) Expression Profile

Breakdown by tissue

bone	0		0/38928
bone marrow	0		0/37245
brain	0		0/486112
colon	0		0/52004
eye	5		1/169091
heart	0		0/53082
kidney	0		0/116533
liver	0		0/104225
lung	0		0/43564
lymph node	0		0/25515
mammary gland	0		0/348007
muscle	0		0/19349
ovary	0		0/14904
pancreas	0		0/80182
placenta	0		0/32674
pituitary g...	0		0/43777
skin	23		2/83521
spleen	0		0/69103
stomach	0		0/31299
testis	29		3/102659
thymus	0		0/99645
uterus	0		0/6588

Breakdown by developmental stage

egg	0		0/23603
pre-implant...	205		32/155645
post-implan...	196		12/60985
mid-gestati...	2		1/422752
late-gestat...	0		0/218646
neonate	0		0/57179
post natal	0		0/68404
adult	3		3/848043

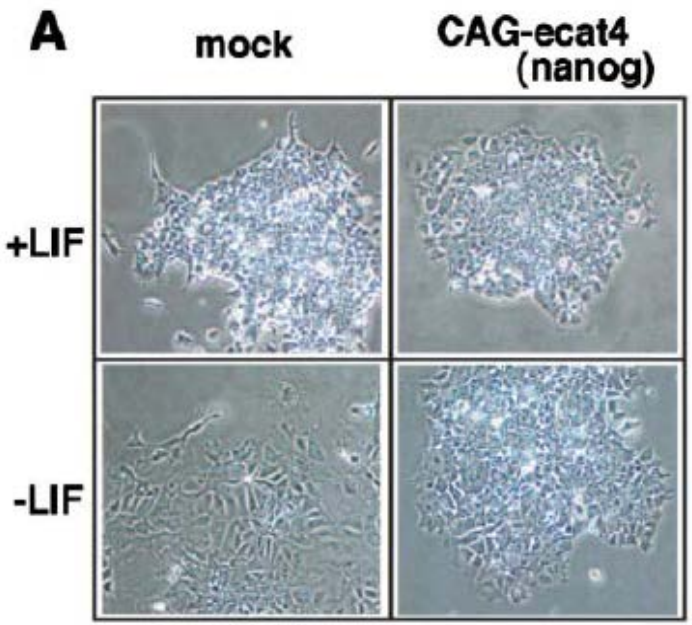
Ecat9 & ***Sox2*** induce massive **cell death**

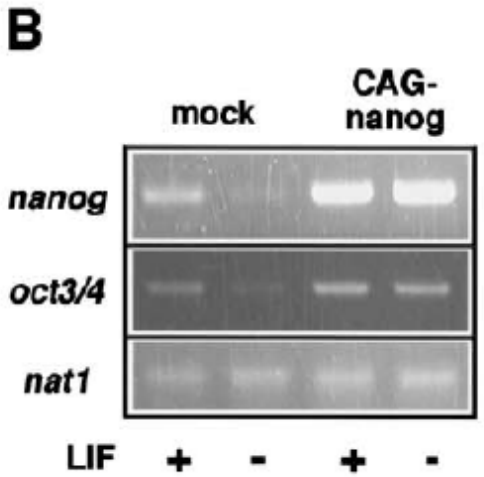
When cultured **with LIF**, all of them showed **normal morphology**

When cultured **without LIF**, all but one **differentiated** normally as judged by **flattened morphology** & **reduced *oct3/4* expression**

Cells constitutively expressing ***ecat4*** did not show such a morphological change even after **prolonged culture (> 1 month) without LIF**

Expression of ***oct3/4*** also remained normal





CAG: the CMV early enhancer/chicken beta actin (CAG) promoter:
Nanog **from Tir Na Nog (land of the ever young)**

2005

THE JOURNAL OF BIOLOGICAL CHEMISTRY

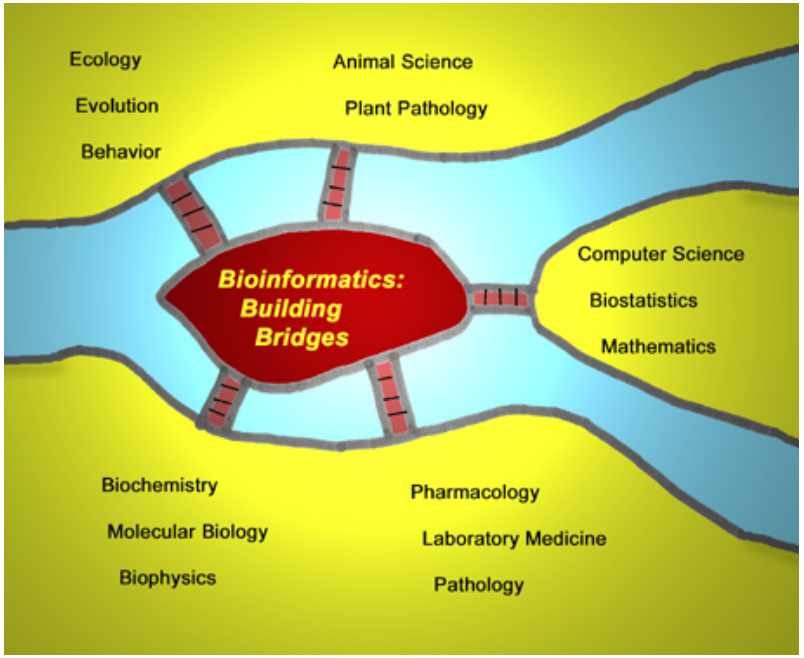
Vol. 280, No. 26, Issue of July 1, pp. 24371–24379, 2005
Printed in U.S.A.

Differential Roles for Sox15 and Sox2 in Transcriptional Control in Mouse Embryonic Stem Cells*[S]

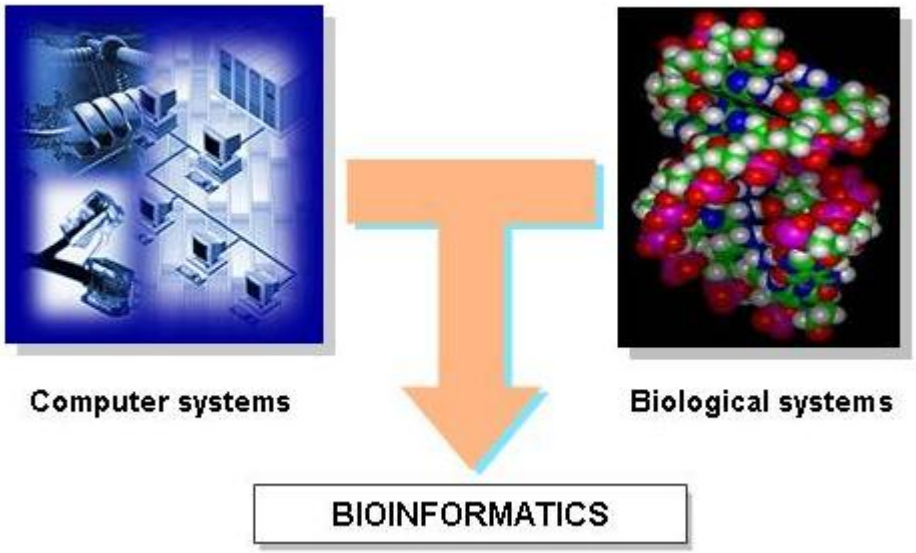
Received for publication, February 7, 2005, and in revised form, April 5, 2005
Published, JBC Papers in Press, April 29, 2005, DOI 10.1074/jbc.M501423200

Masayoshi Maruyama, Tomoko Ichisaka, Masato Nakagawa, and Shinya Yamanaka‡

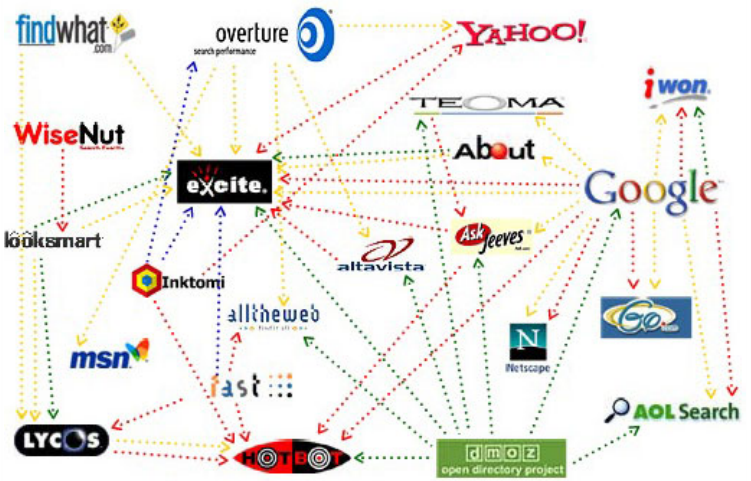
From the Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan and CREST, Japan Science and Technology Agency, Kyoto 606-8507, Japan

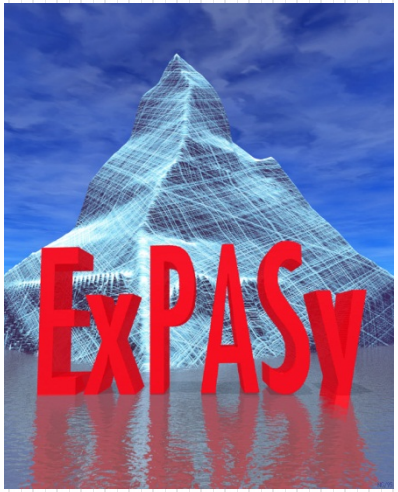

Ecology
Evolution
Behavior
Animal Science
Plant Pathology
Bioinformatics: Building Bridges
Computer Science
Biostatistics
Mathematics
Biochemistry
Molecular Biology
Biophysics
Pharmacology
Laboratory Medicine
Pathology

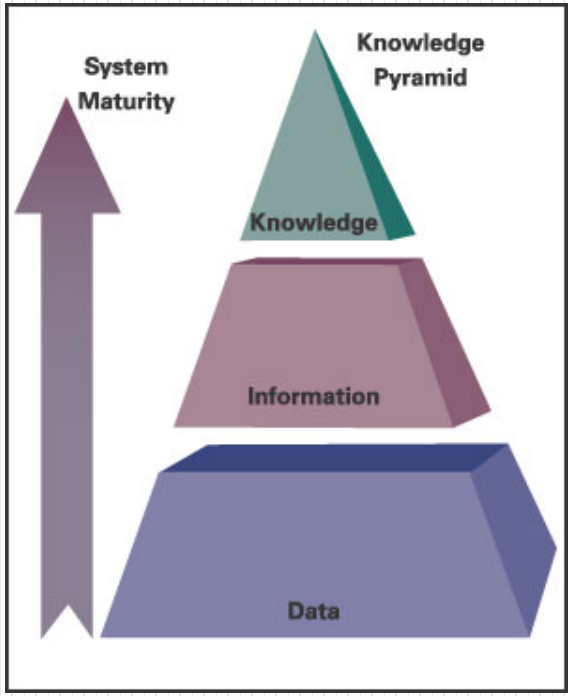
Q & A


Computer systems
Biological systems
BIOINFORMATICS

Q: How to Find the Right Stuffs?







SIB Fellowship ▼

ExPASy
Bioinformatics Resource Portal

Query all databases ▾ search

- Visual Guidance
- Categories
  - proteomics
  - genomics
  - structural bioinformatics
  - systems biology
  - phylogeny/evolution
  - population genetics
  - transcriptomics
  - biophysics
  - imaging
  - IT infrastructure
  - drug design
- Resources A..Z
- Links/Documentation

ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

STRING

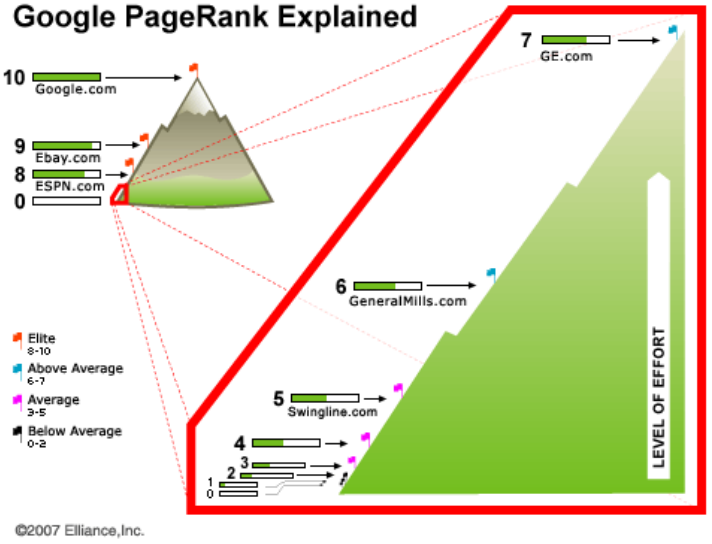
Database of known and predicted protein-protein interactions
[details]

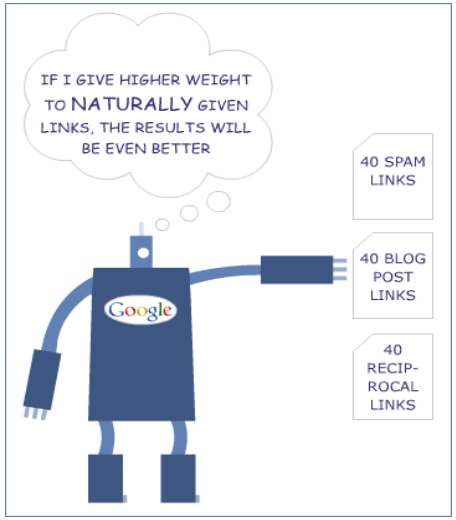
How to use this portal?

How to Find the Right Stuffs

Google	Algorithm: **PageRank™**
	PDF, 庫存頁面…
Askcom	**ExpertRank** algorithm
	Subject-specific popularity
Use **the right key words**	PubMed: MeSH
	OMIM: index
	Gene name: HUGO

Fidelity: edu > gov > org > com

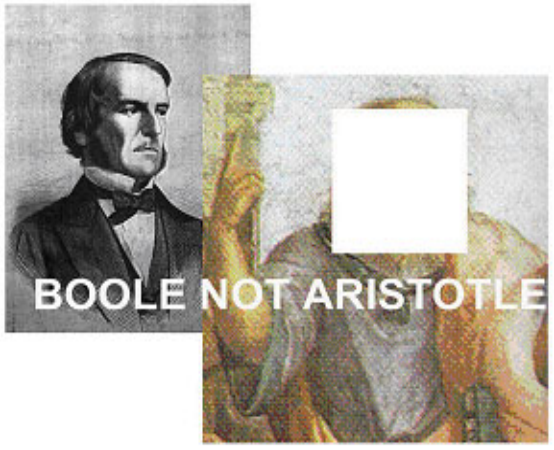




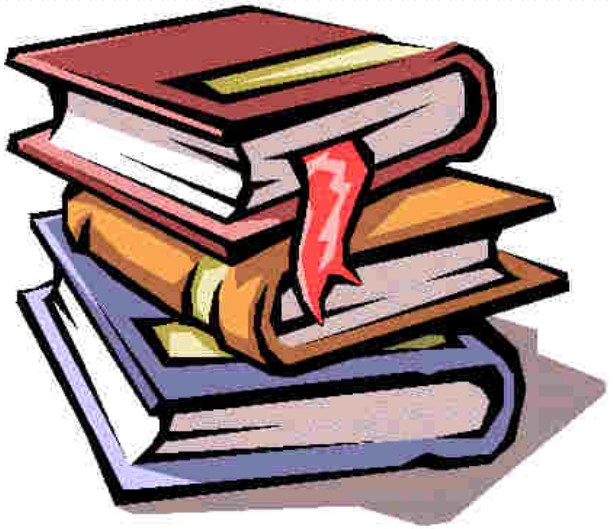
Search Efficiently

Quick Tours	Search PubMed by Authors
	My NCBI…
Boolean operators	**AND** OR **NOT**

This is a small search.
Your results will
include *both* words.



Q: How to Find References Related to Your Favorite Gene (YFG)



Gene or Disease – Official Symbol

1----- (100000-) 2----- (200000-)	Autosomal loci or phenotypes (entries created before May 15, 1994)
3----- (300000-)	X-linked loci or phenotypes
4----- (400000-)	Y-linked loci or phenotypes
5----- (500000-)	Mitochondrial loci or phenotypes
6----- (600000-)	Autosomal loci or phenotypes (entries created after May 15, 1994)

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

POU5F1P8	POU class 5 homeobox 1 pseudogene 8	Homo sapiens
Pou5f1	POU domain, class 5, transcription factor 1	Mus musculus
Pou5f1-rs1	POU domain, class 5, transcription factor 1, related sequence 1	Mus musculus
Pou5f1-rs10	POU domain, class 5, transcription factor 1, related sequence 10	Mus musculus
Pou5f1-rs2	POU domain, class 5, transcription factor 1, related sequence 2	Mus musculus
Pou5f1-rs3	POU domain, class 5, transcription factor 1, related sequence 3	Mus musculus
Pou5f1-rs4	POU domain, class 5, transcription factor 1, related sequence 4	Mus musculus
Pou5f1-rs5	POU domain, class 5, transcription factor 1, related sequence 5	Mus musculus
Pou5f1-rs6	POU domain, class 5, transcription factor 1, related sequence 6	Mus musculus
Pou5f1-rs8	POU domain, class 5, transcription factor 1, related sequence 8	Mus musculus
Pou5f1-rs9	POU domain, class 5, transcription factor 1, related sequence 9	Mus musculus
Pou5f2	POU domain class 5, transcription factor 2	Mus musculus
POU5F1	POU class 5 homeobox 1	Sus scrofa
pou5f1	POU domain, class 5, transcription factor 1	Danio rerio
POU5F1	POU class 5 homeobox 1	Pan troglodytes
POU5F1	POU class 5 homeobox 1	Pan troglodytes
POU5F2	POU domain class 5, transcription factor 2	Pan troglodytes

GeneCards® The Human Gene Compendium

PubMed
- POU5F1

PubMed.gov

OMIM
- Preview and index

GeneCards/
human
- POU5F1 (symbol only)

Entrez_Gene
- POU5F1

iHOP
- POU5F1

iHOP information hyperlinked over proteins

iHOP: Model building

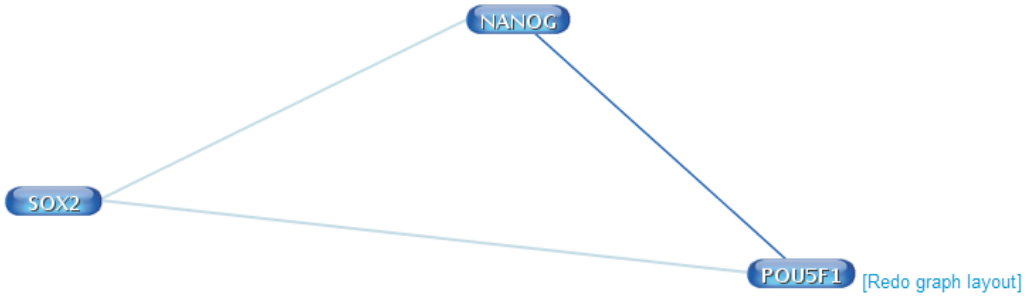
WikiGenes	edit this page new
UniProt	Q01860, A6NLL8, Q15167
IntAct	Q01860
OMIM	164177
NCBI Gene	5460
NCBI RefSeq	NP_976034, NP_002692
NCBI RefSeq	NM_002701, NM_203289
NCBI UniGene	5460
NCBI Accession	BAB63311, CAA79974

Homologues of POU5F1 ...

Definitions for POU5F1 ...

Most recent information for POU5F1 ...

Enhanced PubMed/Google query ...

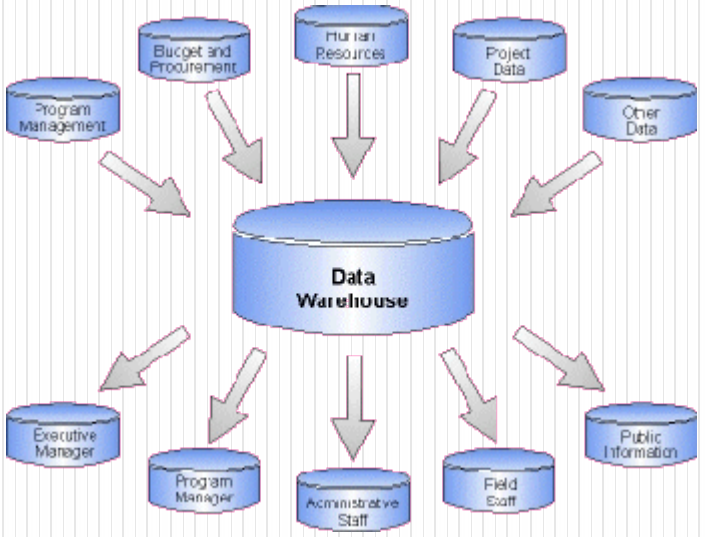


[Redo graph layout]

We found that Oct4 [POU5F1] was transiently **activated** at the 2-cell stage (P-value <0.05) while Nanog [NANOG] and Sox2 [SOX2] were activated at the 4-cell stage (P-value <0.05) in in vitro embryos.

Please cite the use of iHOP as "Hoffmann, R., Valencia, A. A gene network for navigating the literature. Nature Genetics 36, 664 (2004)" and as "iHOP - http://www.ihop-net.org/".

Q: What is Derivative Databases?



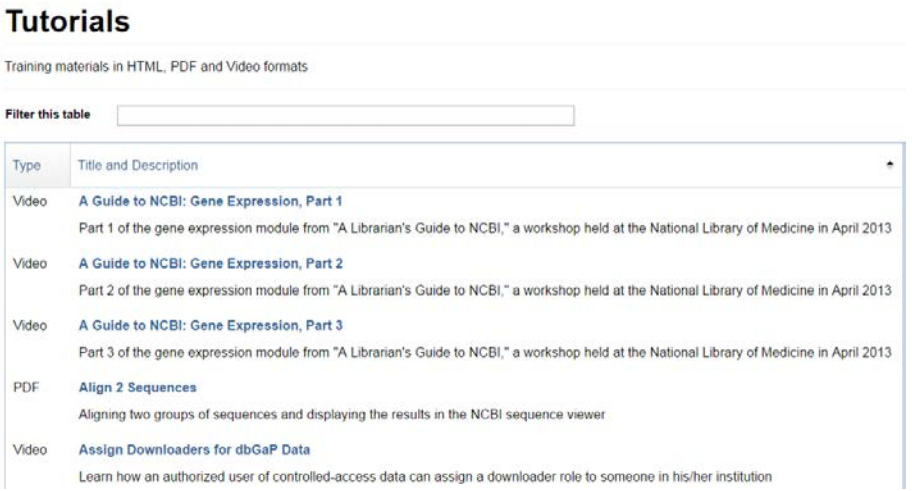
Leading Bioinformatic Centers

NCBI, USA

- To develop **new methods** for integrative, **computer-based data analysis** to **mine** massive and complex **data sets**

EBI, UK

- The EBI is a centre for **research** and **services** in **bioinformatics**
- The Institute manages **databases** of **biological data** including **nucleic acid, protein sequences & macromolecular structures**

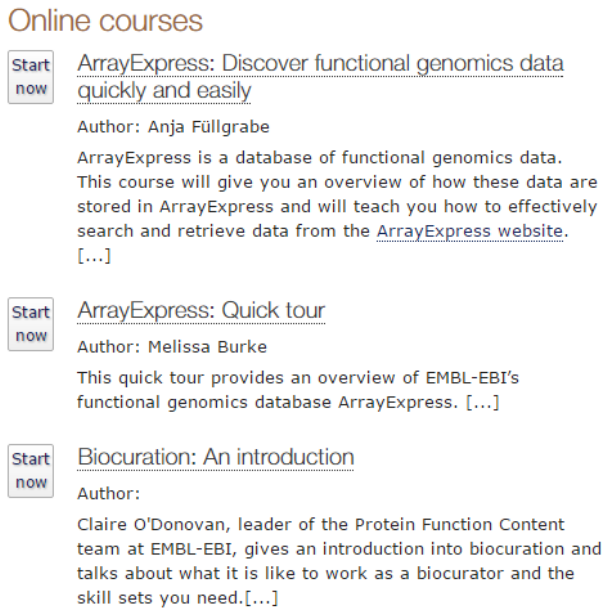


Tutorials

Training materials in HTML, PDF and Video formats

Filter this table

Type	Title and Description
Video	**A Guide to NCBI: Gene Expression, Part 1** Part 1 of the gene expression module from "A Librarian's Guide to NCBI," a workshop held at the National Library of Medicine in April 2013
Video	**A Guide to NCBI: Gene Expression, Part 2** Part 2 of the gene expression module from "A Librarian's Guide to NCBI," a workshop held at the National Library of Medicine in April 2013
Video	**A Guide to NCBI: Gene Expression, Part 3** Part 3 of the gene expression module from "A Librarian's Guide to NCBI," a workshop held at the National Library of Medicine in April 2013
PDF	**Align 2 Sequences** Aligning two groups of sequences and displaying the results in the NCBI sequence viewer
Video	**Assign Downloaders for dbGaP Data** Learn how an authorized user of controlled-access data can assign a downloader role to someone in his/her institution



Online courses

Start now — ArrayExpress: Discover functional genomics data quickly and easily

Author: Anja Füllgrabe

ArrayExpress is a database of functional genomics data. This course will give you an overview of how these data are stored in ArrayExpress and will teach you how to effectively search and retrieve data from the ArrayExpress website. [...]

Start now — ArrayExpress: Quick tour

Author: Melissa Burke

This quick tour provides an overview of EMBL-EBI's functional genomics database ArrayExpress. [...]

Start now — Biocuration: An introduction

Author:

Claire O'Donovan, leader of the Protein Function Content team at EMBL-EBI, gives an introduction into biocuration and talks about what it is like to work as a biocurator and the skill sets you need.[...]

The National Center for Biotechnology Information (NCBI)

Founded	**1988**
NCBI	**The leading American information provider;** a division of the National Library of Medicine (NLM), NIH (Bethesda, USA)
Roles	To develop **new information technologies** to aid our understanding of the **molecular** and **genetic processes** that underlie **health and disease**

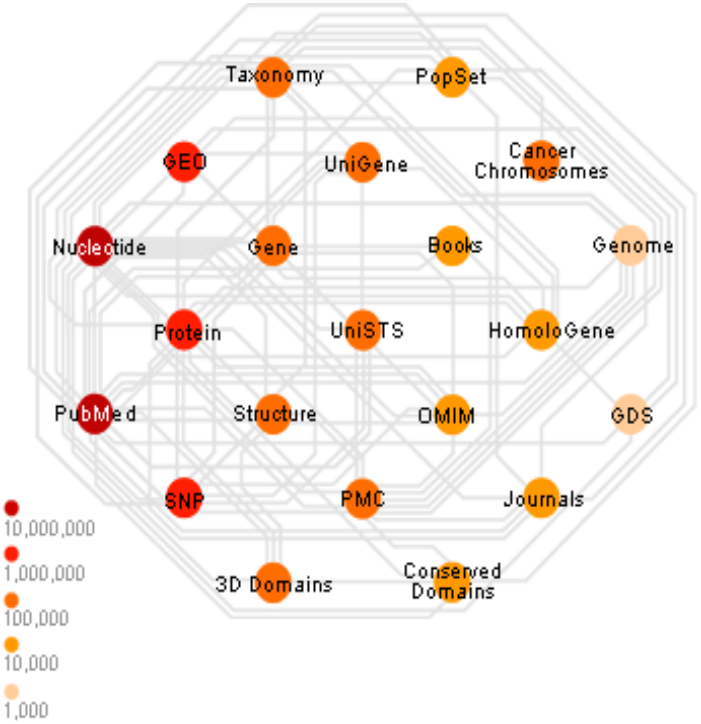
Contents

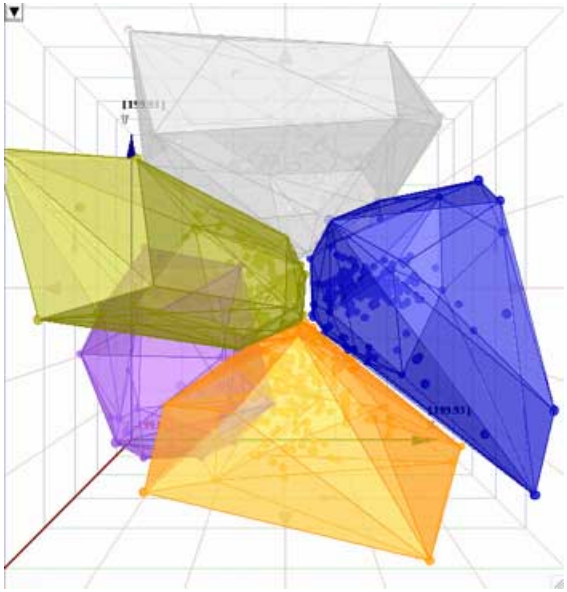
Databases

- **Primary** vs. **derivative** databases
  - **Value-added**

Methodologies (tools)

- Tools: e.g., BLAST, NCBI
- **Algorithms**
  - Neural network (NN)
    - Self-organizing map (SOM)
  - **Hidden Markov Model** (HMM)
  - K-means clustering





Nucleic Acids Research

The 2016 database issue of *Nucleic Acids Research* and an updated molecular biology database collection

Daniel J. Rigden[1,*], Xosé M. Fernández-Suárez[2] and Michael Y. Galperin[3,*]

+ Author Affiliations

*To whom correspondence should be addressed. Tel: +1 301 435 5910; Fax: +1 301 435 7793; Email: nardatabase@gmail.com

Correspondence may also be addressed to Daniel J. Rigden. Tel: +44 151 795 4467; Fax: +44 151 795 4406; Email: drigden@liv.ac.uk

Received November 22, 2015.
Accepted November 23, 2015.

Abstract

« Previous | Next Article »
Table of Contents

This Article

Nucl. Acids Res. (04 January 2016) 44 (D1): D1-D6. doi: 10.1093/nar/gkv1356

This article appears in: Database issue

» Abstract *Free*
Full Text (HTML) *Free*
Full Text (PDF) *Free*
Database Summaries

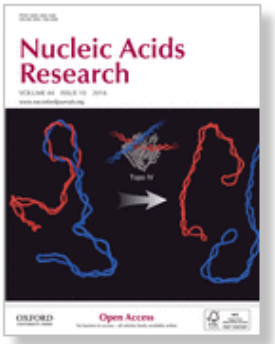
- Classifications

Database Issue

Search this journal:
GO
Advanced »

Current Issue

02 June 2016 44 (10)



The 2016 database issue of *Nucleic Acids Research* and an updated molecular biology database collection

The NAR online Molecular Biology Database Collection in 2016

Daniel J. Rigden, Xos&#65533; M. Fern&#65533;ndez-Su&#65533;rez, and Michael Y. Galperin

This year's update of the NAR online Molecular Biology Database Collection (which is freely available at http://www.oxfordjournals.org/nar/database/c/) involved inclusion of 62 new databases (Table 1) and 15 databases that have been previously described elsewhere and were not part of this Collection (Table 2). In addition, the Collection has been expanded by including such databases as Integrative Cancer Genomics (IntOGen) and Disease Variant Store (DIVAS) (52,53). Our curation checks revealed 121 non-responsive databases, of which 23 obsolete entries have been removed from the Collection and the rest marked for potential removal next year. In addition, 26 entries in the Collection have been updated with respect to their URLs, descriptions, and/or author contact information.

We welcome suggestions for inclusion in the Collection of additional databases that have been published in other journals. Such suggestions should be addressed to XMFS at xose.m.fernandez@gmail.com and should include database summaries in plain text, organized in accordance with the http://www.oxfordjournals.org/nar/database/summary/1 template.

The category and database order generally follows that in the compilation paper. However, many databases appear in more than one category.

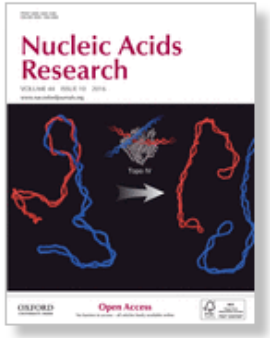
This Article

doi: 10.1093/nar/gkv1356
Nucleic Acids Res 04 January 2016 vol. 44 no. D1 D1-D6

Abstract ***Free***
Full Text (HTML) ***Free***
Full Text (PDF)
» Database Summaries

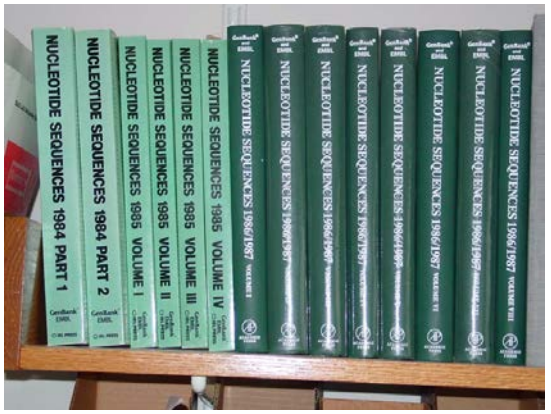
Current Issue

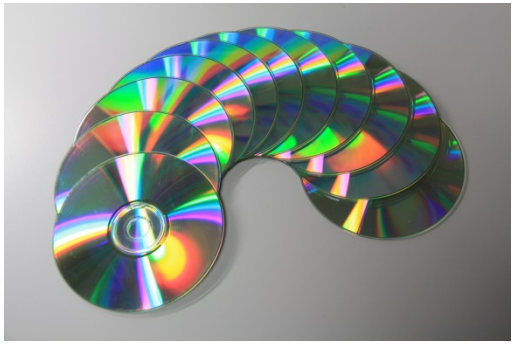
02 June 2016 44 (10)



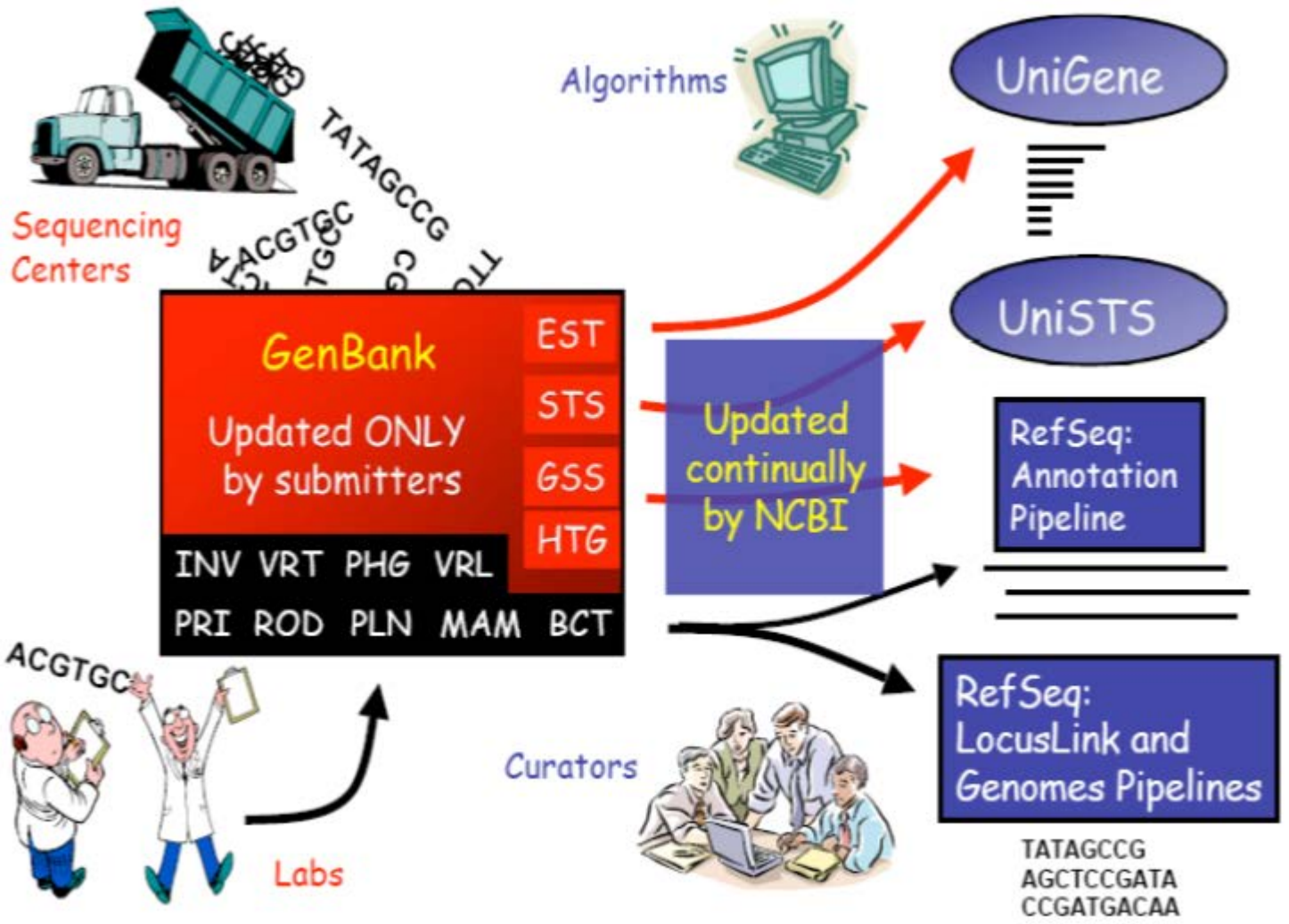
Primary vs. Derivative Databases - NCBI

Primary databases	Derivative databases
• **Original** submissions by **experimenta lists**	• **Curated** by NCBI stuffs
• **Submitters** retain editorial control of records	• **NCBI** retains **editorial control** of records
• Archival in nature	• Record content is **updated continually**



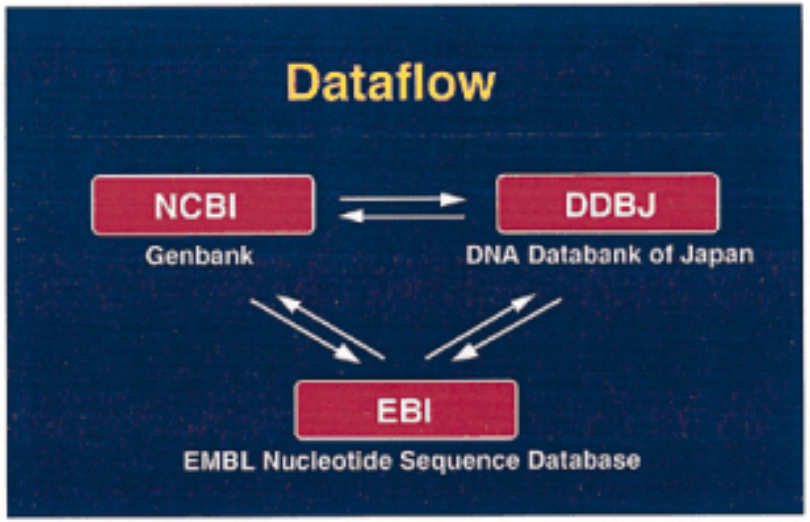


Primary vs. Derivative Databases - NCBI



Primary DNA Databases

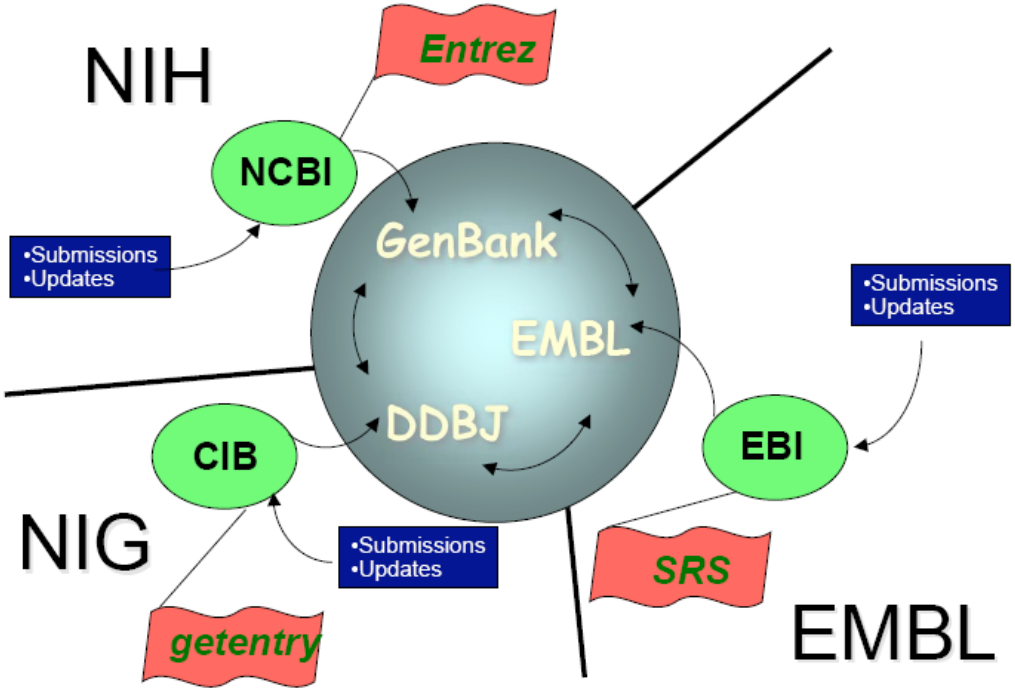
- GenBank (USA)
- EMBL (Europe)
- DDBJ (Japan)



National Institute of Health (**NIH**)

National Center for Biotechnology (**NCBI**)

Retrieval System Across all Databases in NCBI (**ENTREZ**)



National Institute of Genetics (NIG)

Center for Information Biology (CIB)

Research Organization of Information and Systems
National Institute of Genetics

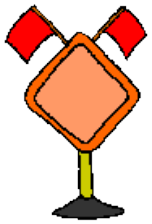
The European Bioinformatics Institute (**EBI**)

Sequence Retrieval System (**SRS**)

The *European Molecular Biology Laboratory* (***EMBL***)

EMBL/GenBank/DDBJ Annotations

Warning!!!



DNA data base annotations are **full of errors**	In sequences, in annotations, in CDs attribution…
	No consistency of annotations
	Most annotations are done by the **submitters**
	Heterogeneity of quality and updating

Some Interesting Sequence Annotation

```
FT   source          1..124
FT                   /db_xref="taxon:4097"
FT                   /organelle="plastid:chloroplast"
FT                   /organism="Nicotiana tabacum"
FT                   /isolate="Cuban cahibo cigar, gift from President Fidel
FT                   Castro"
```

Or:

```
FT   source          1..17084
FT                   /chromosome="complete mitochondrial genome"
FT                   /db_xref="taxon:9267"
FT                   /organelle="mitochondrion"
FT                   /organism="Didelphis virginiana"
FT                   /dev_stage="adult"
FT                   /isolate="fresh road killed individual"
FT                   /tissue_type="liver"
```

???

UniProt › Taxonomy

Search in: Taxonomy | Query | Search | Clear | Advanced Search »

SPECIES **Didelphis marsupialis virginiana (North American opossum)**

UniProtKB (151) | Taxonomy help

Mnemonic	DIDMA
Taxon identifier	9267
Scientific name	Didelphis marsupialis virginiana
Common name	North American opossum
Synonym	-
Other names	› Didelphis virginiana › Virginia opossum
Rank	SPECIES
Lineage	› cellular organisms › Eukaryota › Fungi/Metazoa group › Metazoa › Eumetazoa › Bilateria › Coelomata › Deuterostomia › Chordata › Craniata › Vertebrata › Gnathostomata › Teleostomi › Euteleostomi

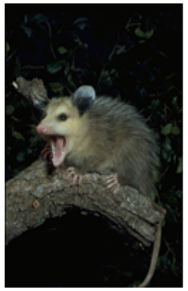
Taxonomy navigation

- ⇧ › Didelphis
- ⇩ Terminal (leaf) node.

Images may be subject to copyright.

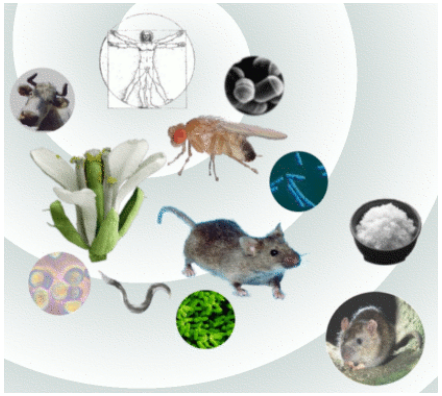
calphotos.berkeley.edu

calphotos.berkeley.edu



calphotos.berkeley.edu

[Taxonomy Browser @ EBI](#) vs. [NCBI Taxonomy](#)



Organization of GenBank: Traditional Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

```
PRI Primate
PLN Plant and Fungal
BCT Bacterial and Archeal
INV Invertebrate
ROD Rodent
VRL Viral
VRT Other Vertebrate
MAM Mammalian
PHG Phage
SYN Synthetic(cloning vectors)
ENV Environmental Samples
UNA Unannotated
```

Traditional Divisions:

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

Entrez query: `gbdiv_xxx[Properties]`

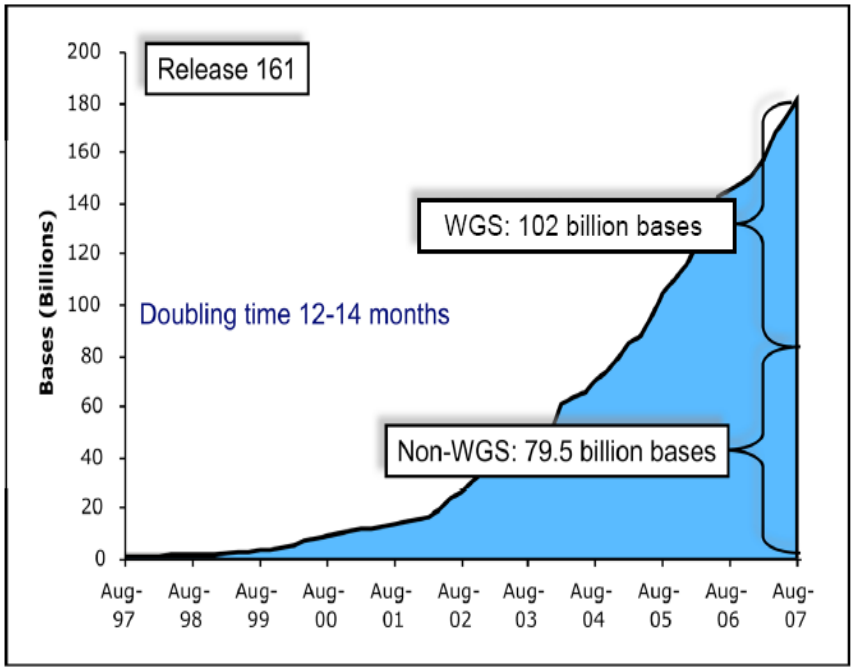
Bulk GenBank Divisions

Batch submission & htg (email & ftp)

Inaccurate & poorly characterized

- **EST:** Expressed Sequence Tag
- **GSS:** Genome Survey Sequence
- **HTG**: High Throughput Genome
- **HTC:** High Throughput cDNA
- **STS:** Sequence Tagged Site

The Growth of GenBank



Organization of GenBank: Bulk Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

```
EST Expressed Sequence Tag
GSS Genome Survey Sequence
HTG High Throughput Genomic
STS Sequence Tagged Site
HTC High Throughput cDNA
PAT Patent
```

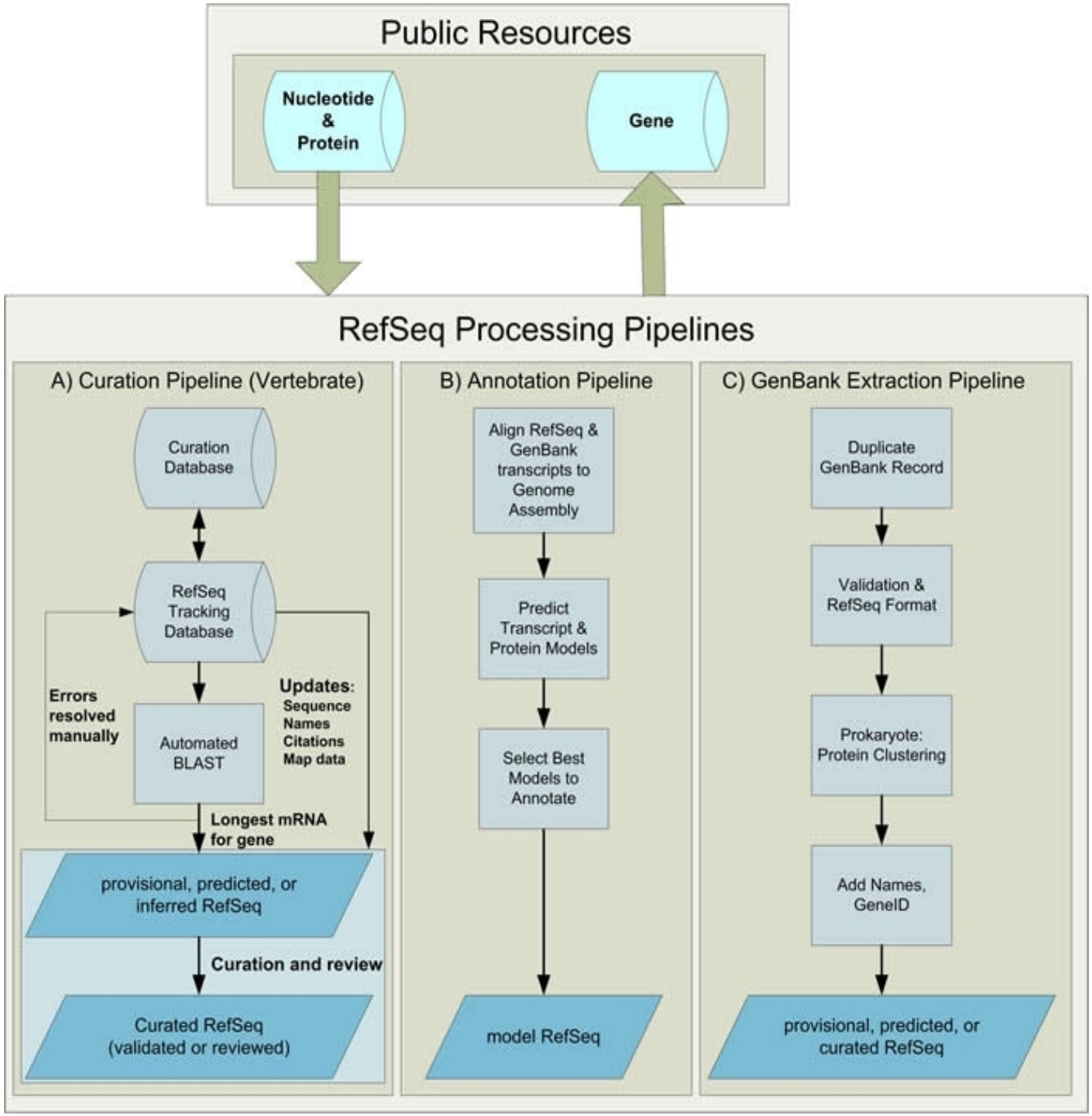
BULK Divisions:

- Batch Submission
  (Email and FTP)
- Inaccurate
- Poorly characterized

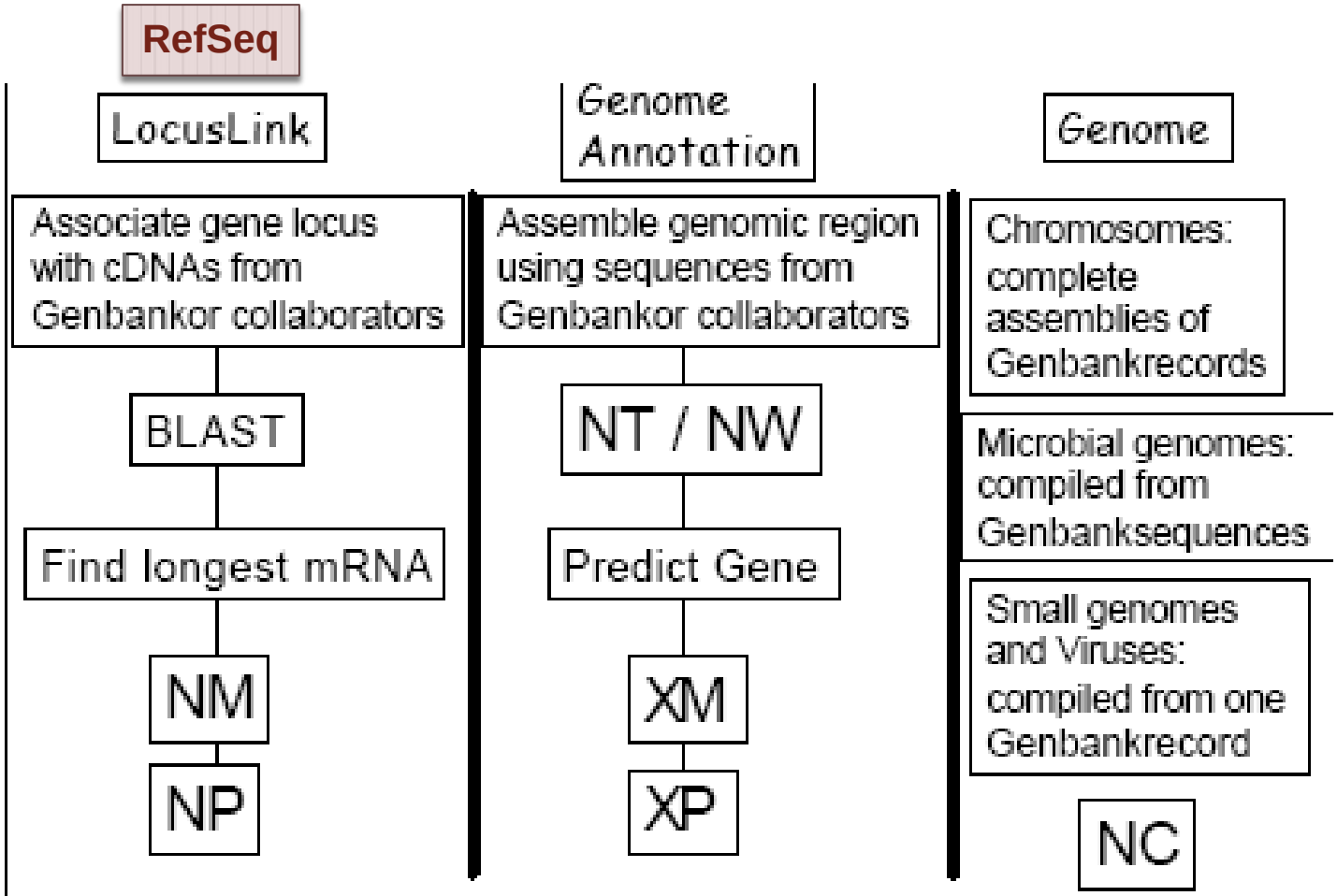
Entrez query: `gbdiv_xxx[Properties]`

Selected RefSeq Accession Number

mRNAs and Proteins	
NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
Gene Records	
NG_123456	Reference Genomic Sequence
Chromosome	
NC_123455	Microbial replicons, organelle genomes, human chromosomes
Assemblies	
NT_123456	Contig
NW_123456	WGS Supercontig


Public Resources
Nucleotide & Protein
Gene
RefSeq Processing Pipelines
A) Curation Pipeline (Vertebrate)
Curation Database
RefSeq Tracking Database
Errors resolved manually
Automated BLAST
Updates:
Sequence
Names
Citations
Map data
Longest mRNA for gene
provisional, predicted, or inferred RefSeq
Curation and review
Curated RefSeq (validated or reviewed)
B) Annotation Pipeline
Align RefSeq & GenBank transcripts to Genome Assembly
Predict Transcript & Protein Models
Select Best Models to Annotate
model RefSeq
C) GenBank Extraction Pipeline
Duplicate GenBank Record
Validation & RefSeq Format
Prokaryote: Protein Clustering
Add Names, GeneID
provisional, predicted, or curated RefSeq

RefSeq Pipelines



RefSeq: NCBI's Derivative Sequence Database

- **Curated transcripts and proteins**
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis microbial genomes (proteins), and more
- **Model transcripts and proteins**
- **Assembled Genomic Regions (contigs)**
  - human genome
  - mouse genome
  - rat genome
  - chicken
  - honeybee
  - sea urchin
- **Chromosome records**
  - Human genome
  - microbial
  - organelle

`srcdb_refseq[Properties]`

ftp://ftp.ncbi.nih.gov/refseq/release/

RefSeq Benefits

Non-redundancy

Explicitly linked nucleotide & protein sequences

Updates to reflect current sequence data & biology

Data **validation**

Format **consistency**

Distinct **accession** series

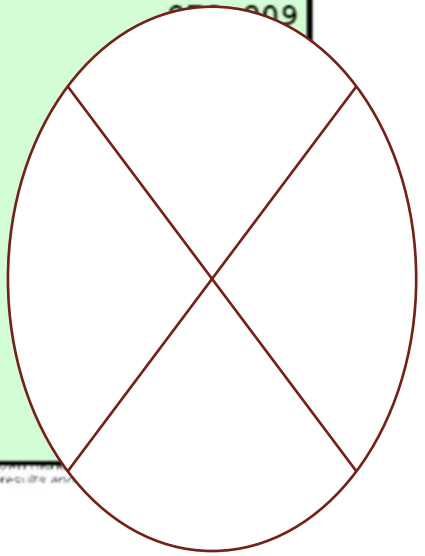
Stewardship by **NCBI staffs** **& collaborators**

Entrez Protein: Derivative Databases

Example: CKS1B

```
CDS             105..344
                /gene="CKS1B"
                /gene_synonym="CKS1; ckshs1; PNAS-16; PNAS-18"
                /note="CDC28 protein kinase 1; CDC28 protein kinase 1B;
                cell division control protein CKS1; NB4
                apoptosis/differentiation related protein; PNAS-143;
                CDC2-associated protein CKS1; CKS-1"
                /codon_start=1
                /product="cyclin-dependent kinases regulatory subunit 1"
                /protein_id="NP_001817.1"
                /db_xref="GI:4502857"
                /db_xref="CCDS:CCDS1077.1"
                /db_xref="GeneID:1163"
                /db_xref="HGNC:19083"
                /db_xref="HPRD:00299"
                /db_xref="MIM:116900"
                /translation="MSHKQIYYSDKYDDEEFEYRHVMLPKDIAKLVPKTHLMSESEWR
                NLGVQQSQGWVHYMIHEPEPHILLFRRPLPKKPKK"
exon            164..291
                /gene="CKS1B"
```

Data Source	Sequences
GenPept	11,585,396
RefSeq	3,889,502
Third Party Annotation	5,263
Swiss Prot	
PIR	
PRF	
PDB	
(PAT Division	
Total	
BLAST nr total	
(no patents or env_nr -now 6 million)	



PAT: patent

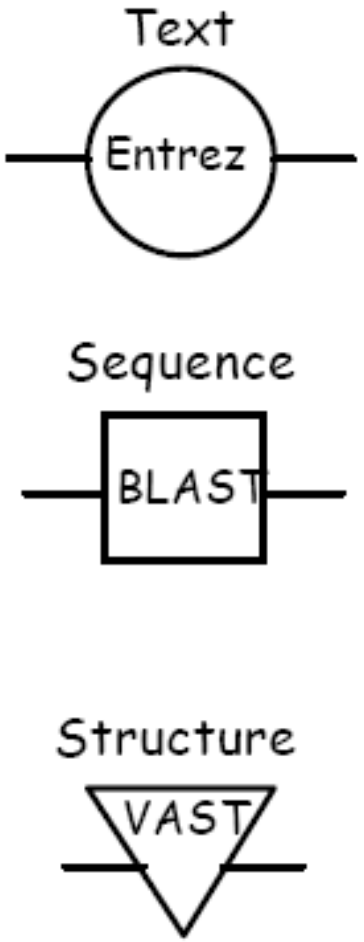
Search in NCBI Databases

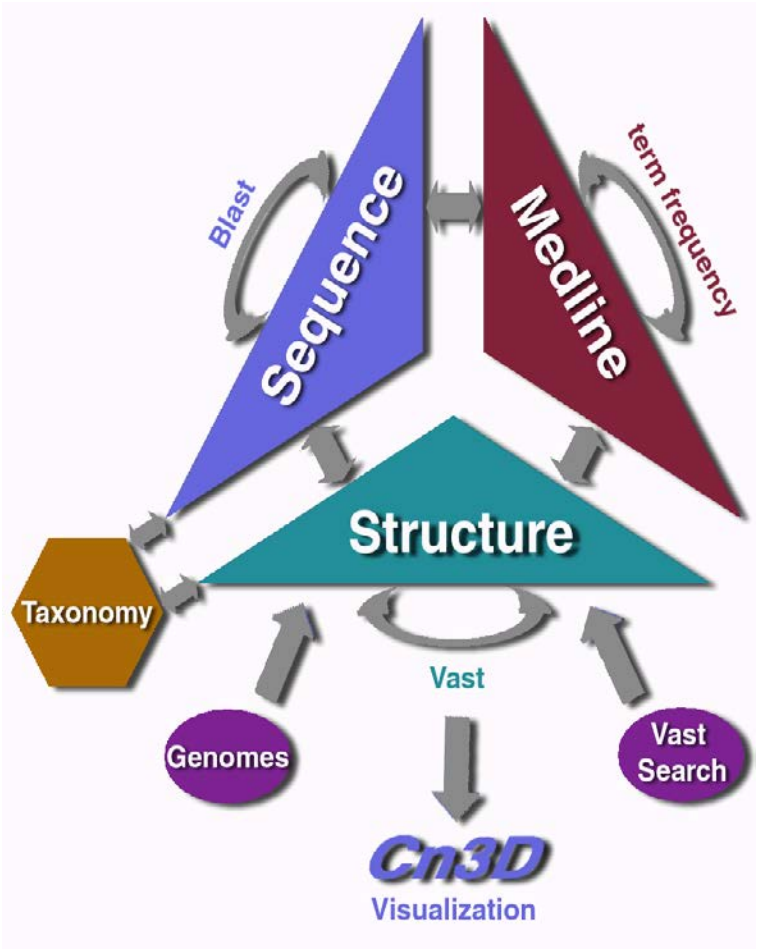
Searches

Text: e.g., *POU5F1* (Oct3/4);

Sequence: e.g., *POU5F1*

Structure: e.g., BRCA1



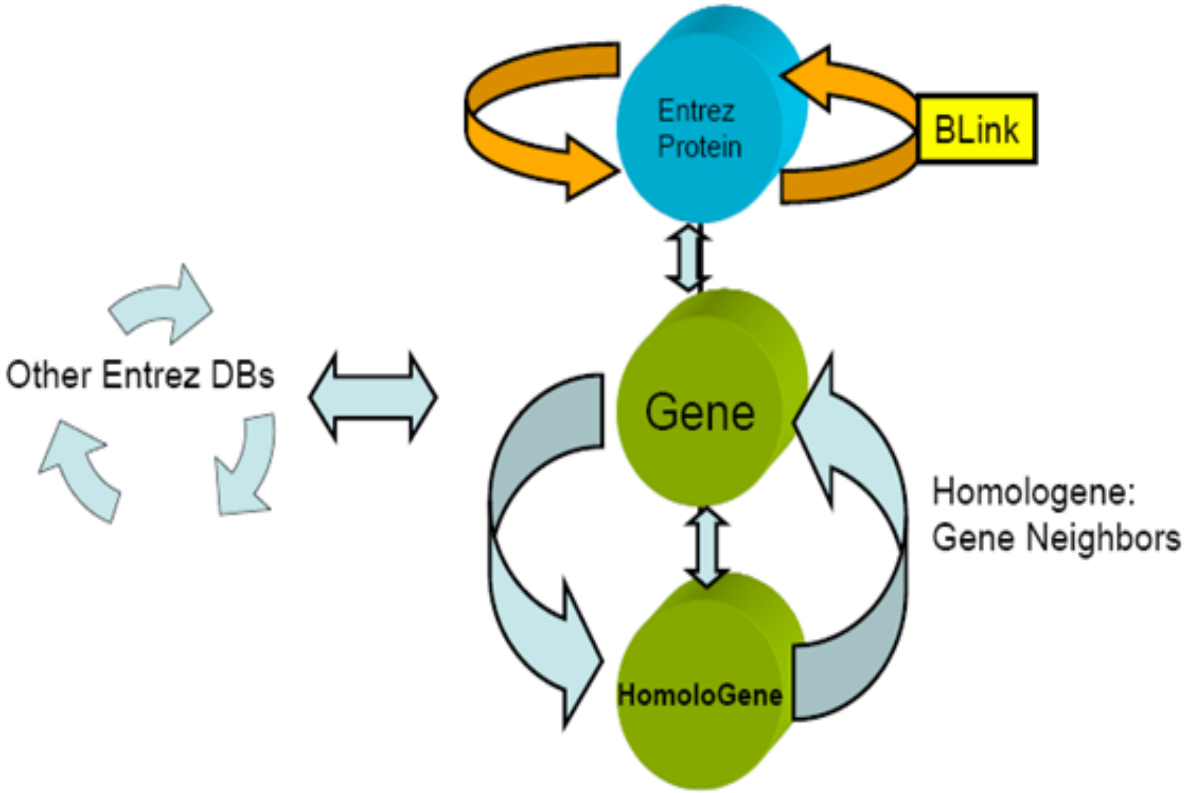


Links Explain

- Order cDNA clone
- Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (OMIM)
- PubMed (GeneRIF)
- SNP
- SNP: Genotype
- SNP: GeneView
- Taxonomy
- UniSTS
- AceView
- CCDS
- Ensembl
- Evidence Viewer
- HGNC
- HPRD
- KEGG
- MGC
- ModelMaker
- UniGene
- LinkOut

© HENRY STEWART PUBLICATIONS 1467-5463. **BRIEFINGS IN BIOINFORMATICS.** VOL 5. NO 2. 204–206. JUNE 2004

Entrez: Use Gene for everything

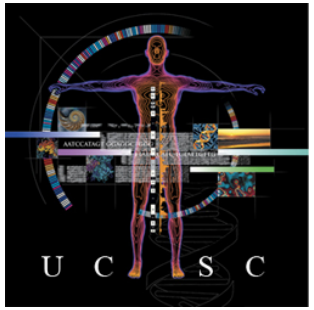


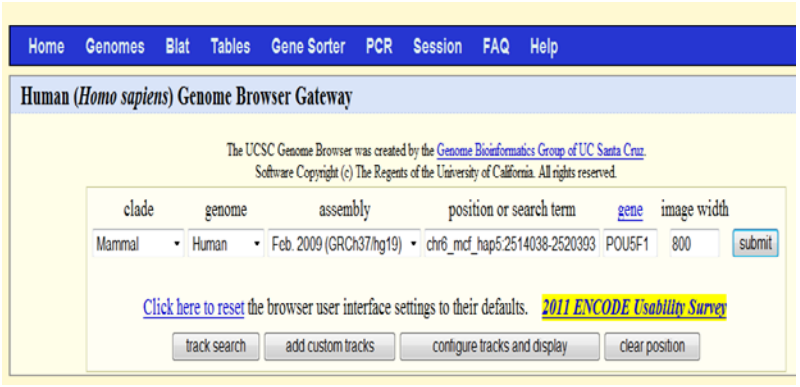
Examples in Other Databases: Using the Official Symbol All the Time (except for protein structure)

STRING 8.3

NPD

The Nuclear Protein Database (e.g., TP53)





Examples in Other Databases: Using the Official Symbol All the Time (except for protein structure)

PDBTM

Protein Data Bank of Transmembrane Proteins

Sun 01 May, 2011 1403 TM structures | Version 2.3 | 65441 visitors.

Menu

Home
Search
Download
Statistics
Documents
News
TMDET
Comment

PDBTM home

Welcome to the PDBTM, the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained in the Institute of Enzymology by the Protein Structure Research Group. PDBTM database was created by scanning all PDB entries with TMDET algorithm. You can read more about PDBTM in our articles and in PDBTM manual. If you find PDBTM useful in your research, please cite our articles (Bioinformatics 20, 2964-2972; Nucleic Acids Research 33 Database Issue, D275-8).

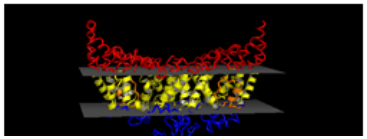
Current holdings

- **73492** structures,
- **1403** transmembrane structures,
  - 1204 alpha helical,
  - 198 beta barrel.

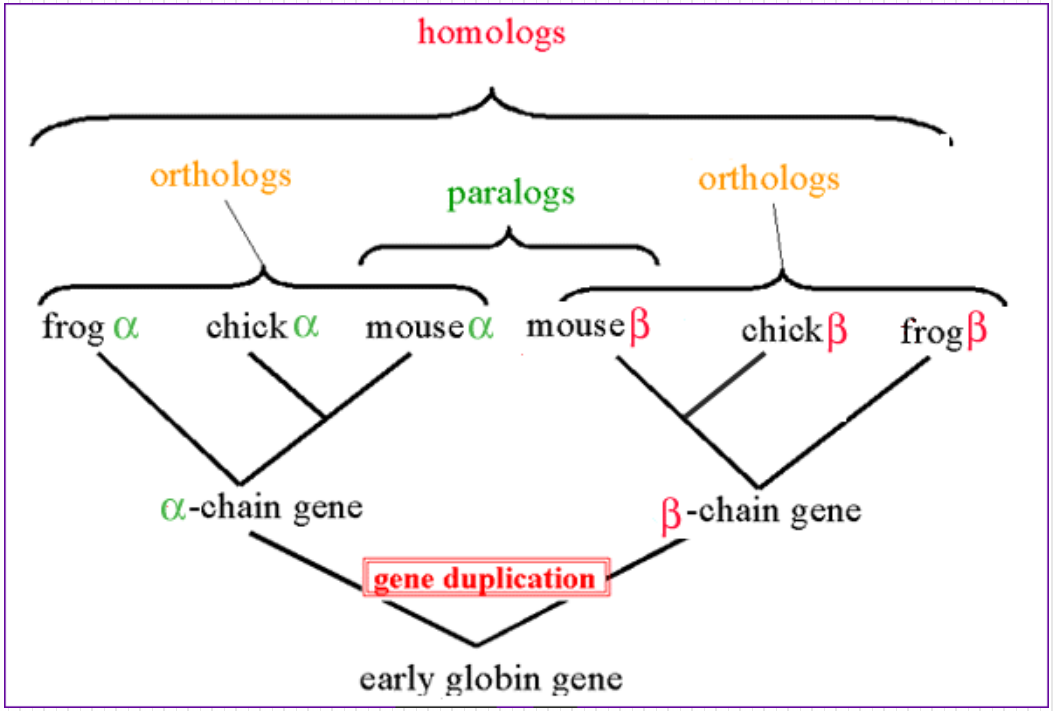
more

Molecule of the month

Phosphorylation-coupled saccharide transporter



Q: How Do You Find the Orthologs from Other Species



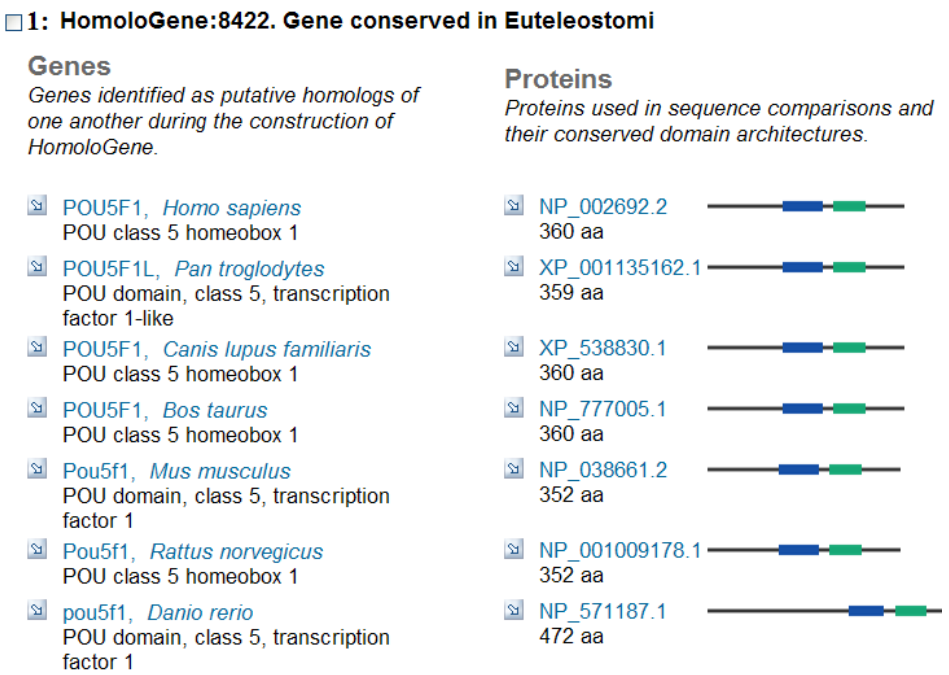
Homologs (1)

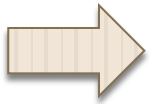
NCBI_Homologene (links)

- A set of maps that shown **chromosomal regions** homologous between mouse, human & other species

Example

- ***POU5F1*** (via ENTREZ_GENE) **Links** to the "Homologene"
  - Protein: multiple alignment
  - Conserved domains
  - PubMed (references)
  - Protein → All links from this record → BLink





All links from this record

BLink

Related Sequences

Identical Proteins

BioSystems

CDD Search Results

Conserved Domains (Concise)

Conserved Domains (Full)

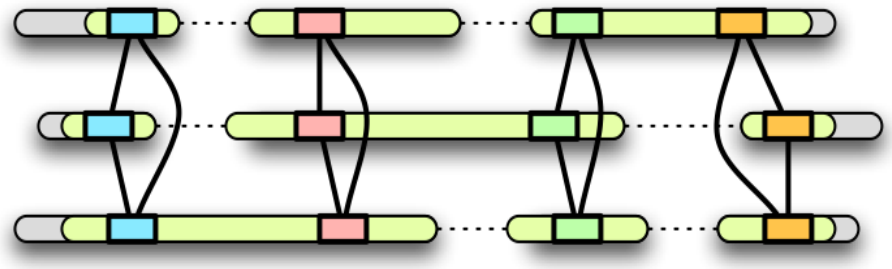
Domain Relatives

Encoding mRNA

Homologs (2)

Hs and **Mm** links adjacent to each map name show **the mouse-human homology map** with the master chromosome as human or mouse

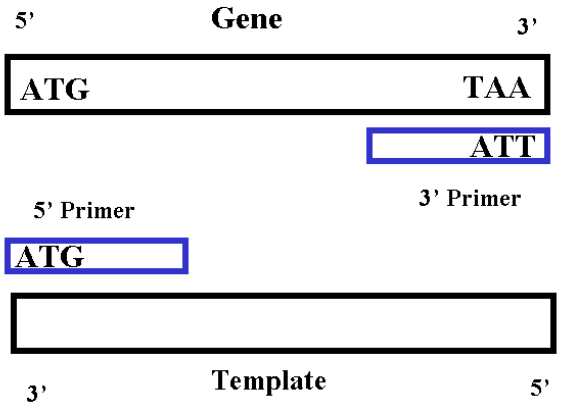
- Mouse Genome Informatics
  - Mm: *Pou5f1* (chr. 17; 19.23 cM)

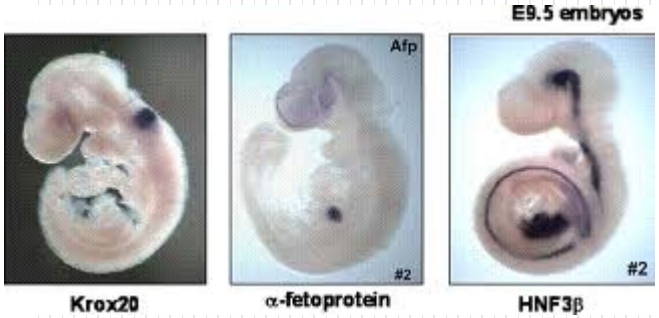


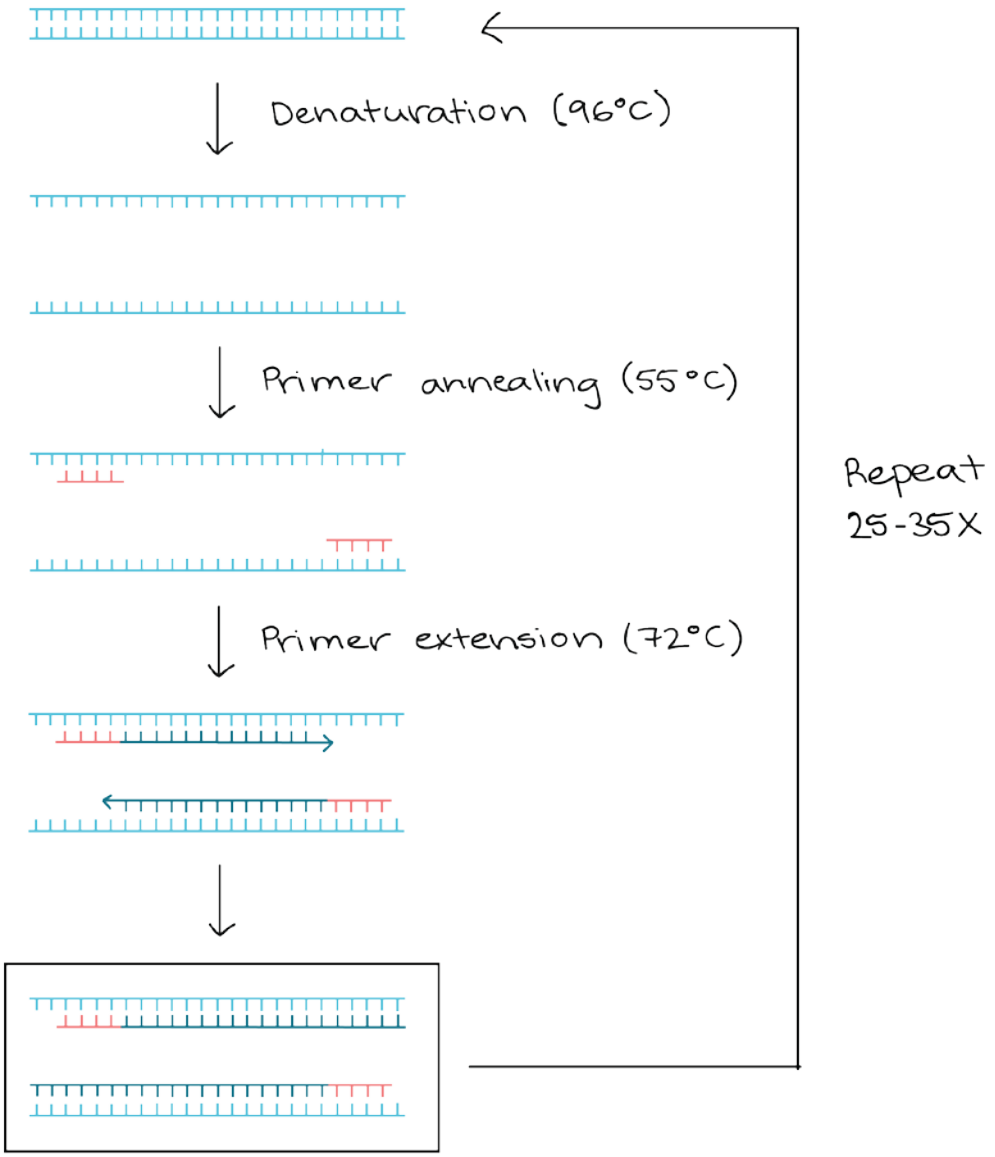
Mercator

Multiple Whole-Genome Orthology Map Construction

Q: How to Design Primers/Probes for PCR/qPCR/Cloning/in situ hybridization



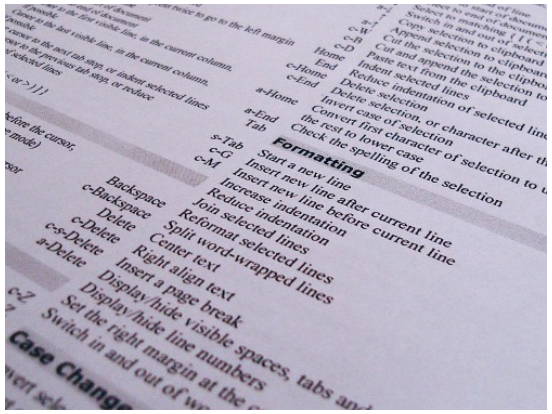



Denaturation (96°C)
Primer annealing (55°C)
Primer extension (72°C)
Repeat
25-35X
Result after 1 cycle:
# of DNA molecules
doubled

Literatures

Databases

ab inito design

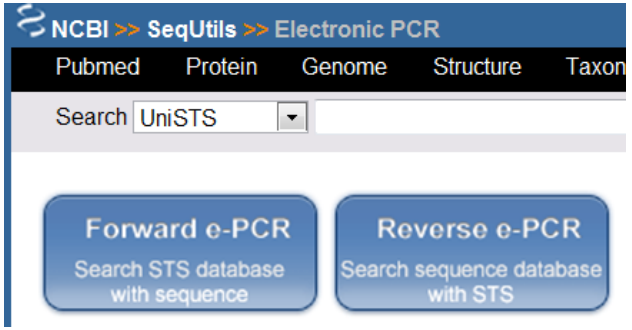


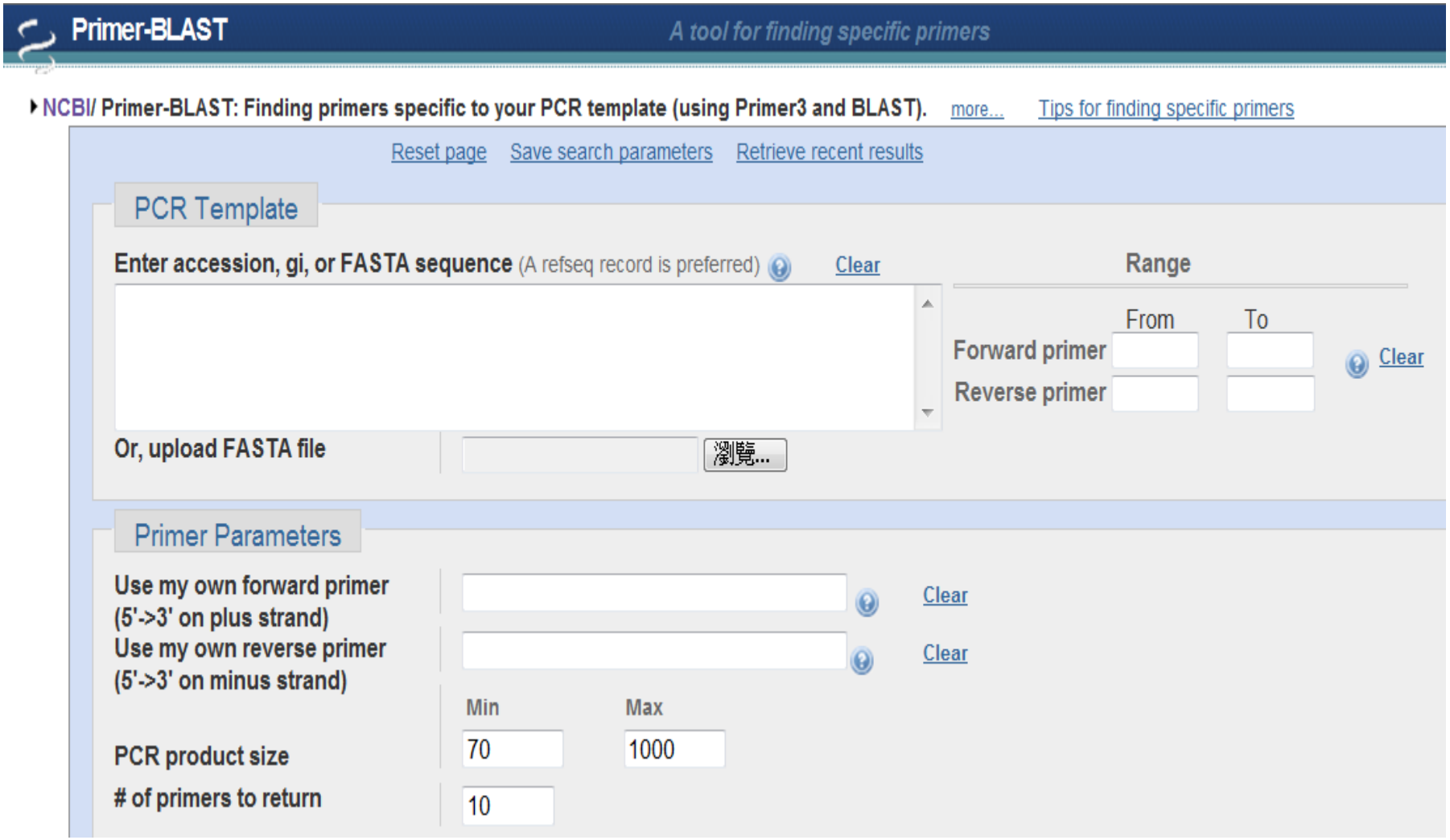
RT PRIMER DB

e.g., ACTB

BLAST

BLAST




Primer-BLAST
A tool for finding specific primers
NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST).
more...
Tips for finding specific primers
Reset page
Save search parameters
Retrieve recent results
PCR Template
Enter accession, gi, or FASTA sequence (A refseq record is preferred)
Clear
Range
From
To
Forward primer
Reverse primer
Clear
Or, upload FASTA file
瀏覽...
Primer Parameters
Use my own forward primer (5'->3' on plus strand)
Clear
Use my own reverse primer (5'->3' on minus strand)
Clear
Min
Max
PCR product size
70
1000
# of primers to return
10

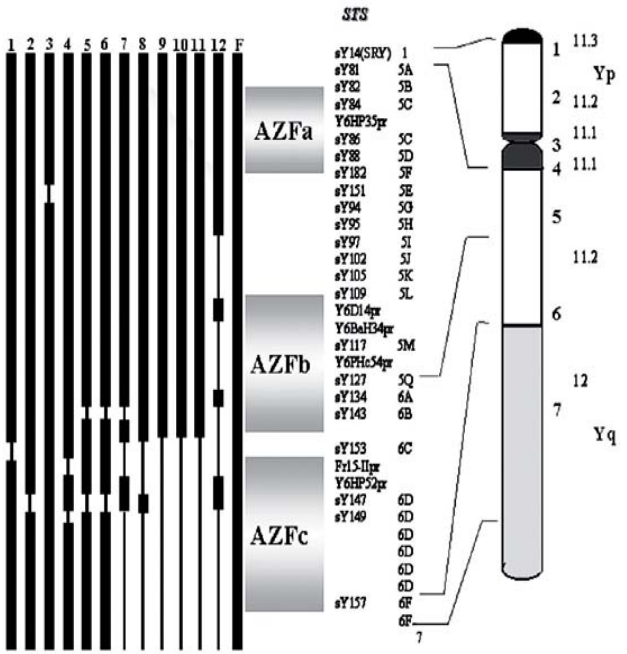
Sequence Tagged Sites (STSs)

The NCBI's **electronic PCR (e-PCR) tool**

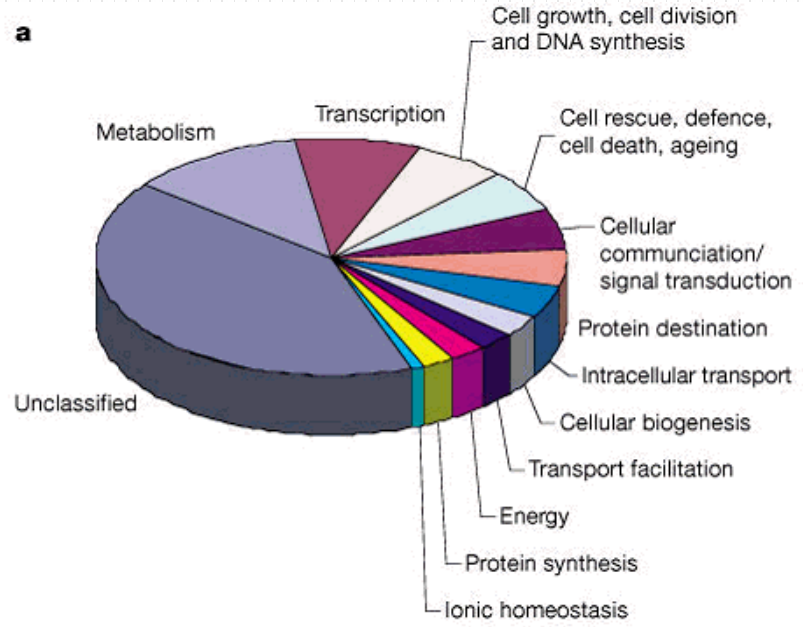
- A part of the UniSTS resource, can be used to find **STS markers within a DNA fragment of interest**

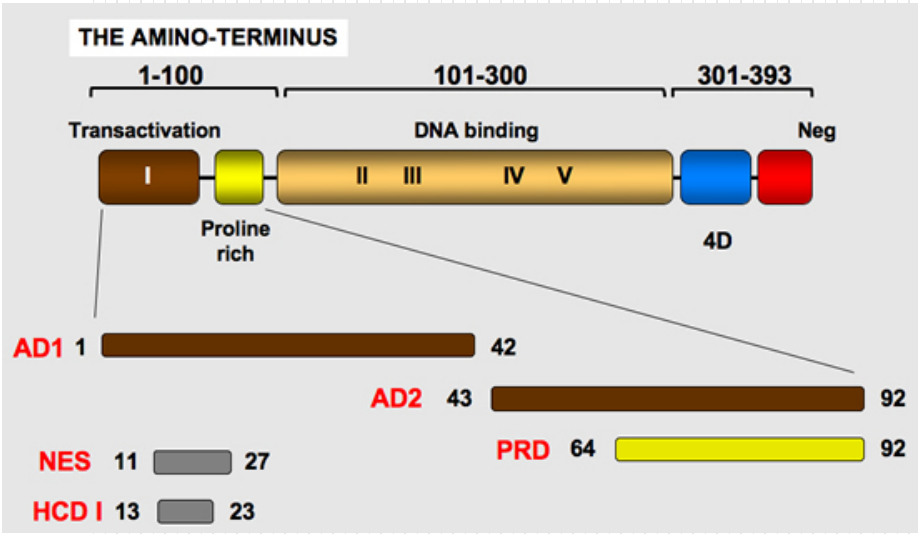
UniSTS contains all the available data on **STS markers (through electronic PCR)**

- **Primer sequences**
- Product (**amplicon**) size
- **Mapping** information
- **Cross references** (Links)



Q: How to Find the Function and/or Structure of YFG





1. Gene Ontology

Through **integrated databases**

- Entrez_Gene
  - **GO terms**
- GeneCards
  - **GO terms**
- Uniprot/Swiss-Prot
  - POU5F1_Human
  - General annotation (comments)
  - Ontologies

Function	Evidence
DNA binding	IDA PubMed
miRNA binding	IDA PubMed
promoter binding	IDA PubMed
protein binding	IPI PubMed
sequence-specific DNA binding	IEA
transcription factor activity	IDA PubMed
transcription factor binding	IPI PubMed

Process	Evidence
BMP signaling pathway involved in heart induction	IMP PubMed
anatomical structure morphogenesis	TAS PubMed
cardiac cell fate determination	IDA PubMed
cell fate commitment involved in the formation of primary germ layers	IMP PubMed
negative regulation of gene silencing by miRNA	IMP PubMed
positive regulation of SMAD protein nuclear translocation	IDA PubMed
positive regulation of catenin protein nuclear translocation	IDA PubMed
positive regulation of gene-specific transcription from RNA polymerase II promoter	IDA PubMed

GO Evidence Code

Introduction

- Experimental Evidence Codes
  - EXP: Inferred from Experiment
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern
- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis
- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available

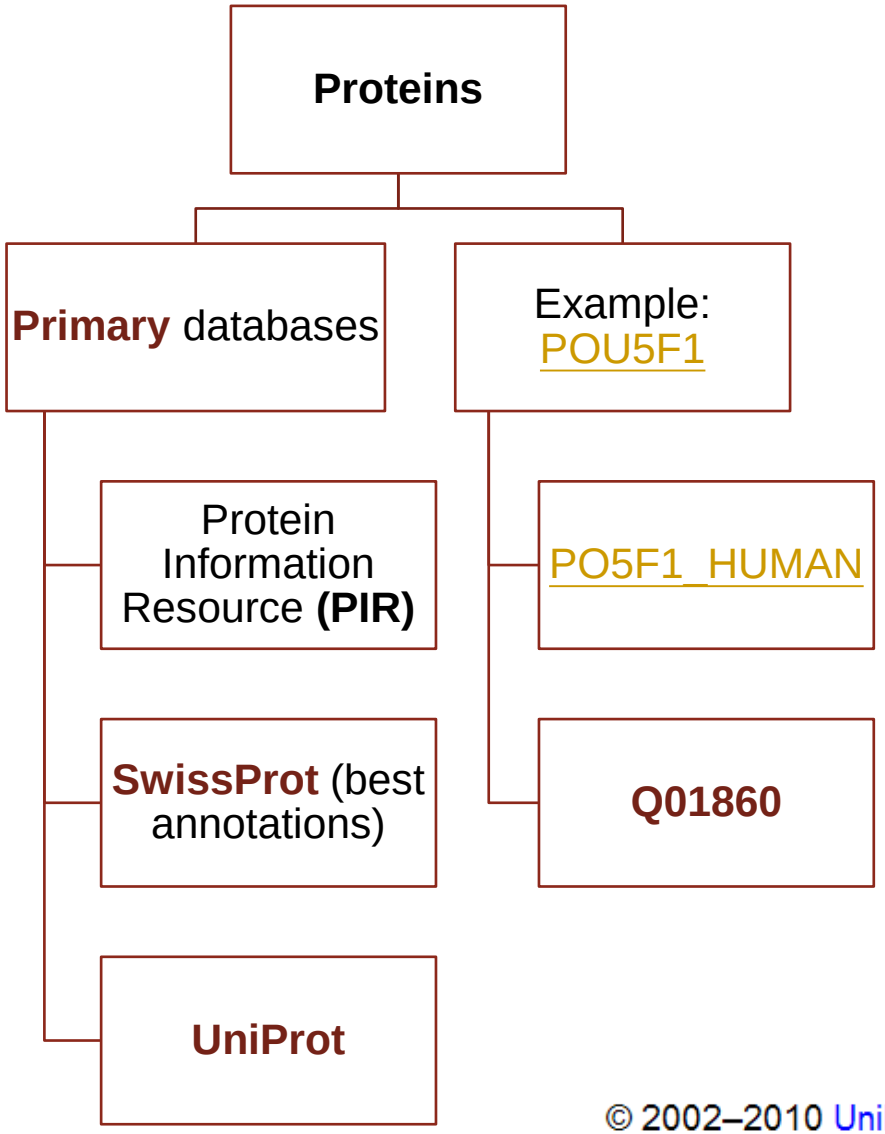
Automatically-assigned Evidence Codes

- IEA: Inferred from Electronic Annotation

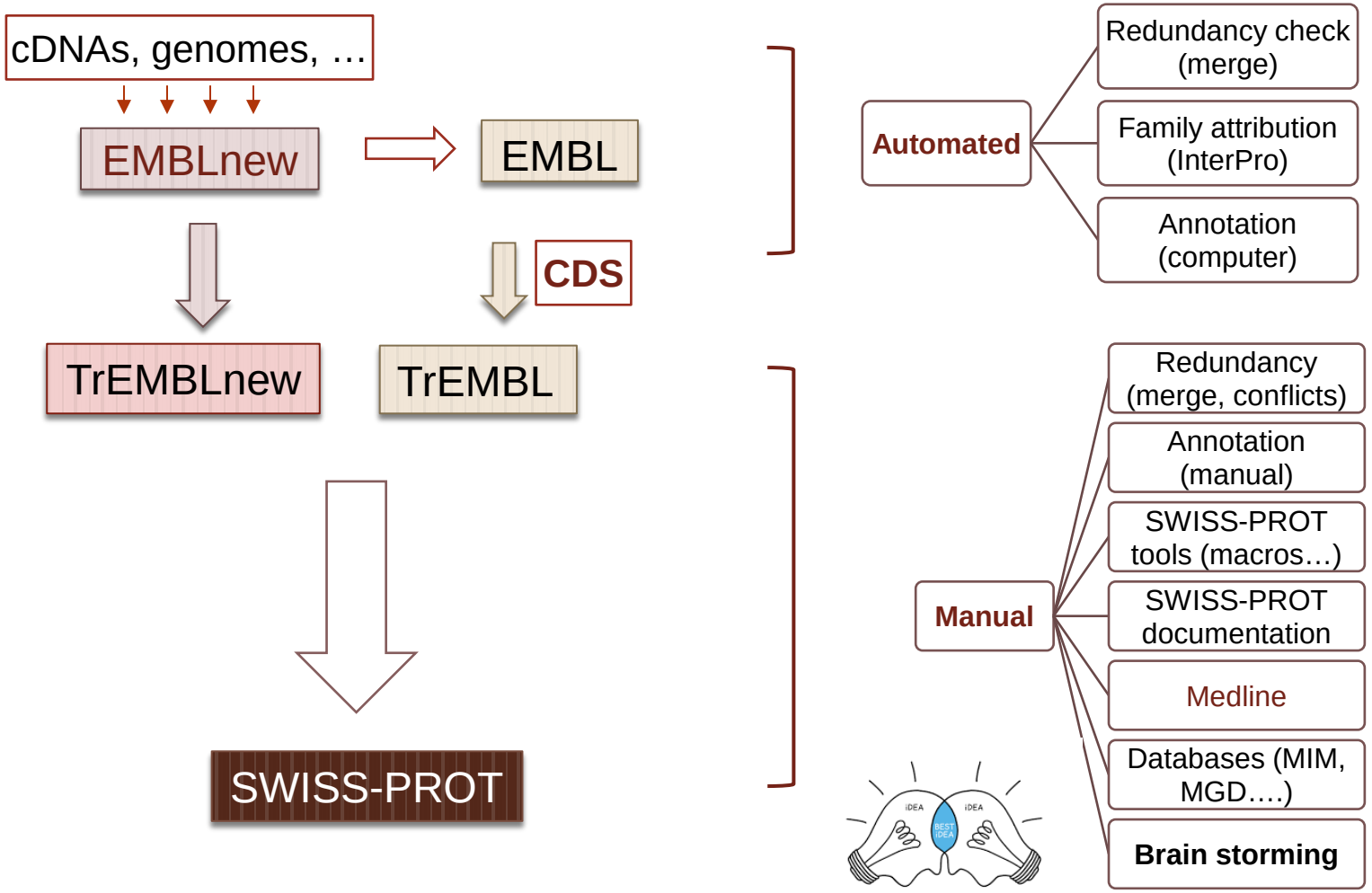
Obsolete Evidence Codes

- NR: Not Recorded

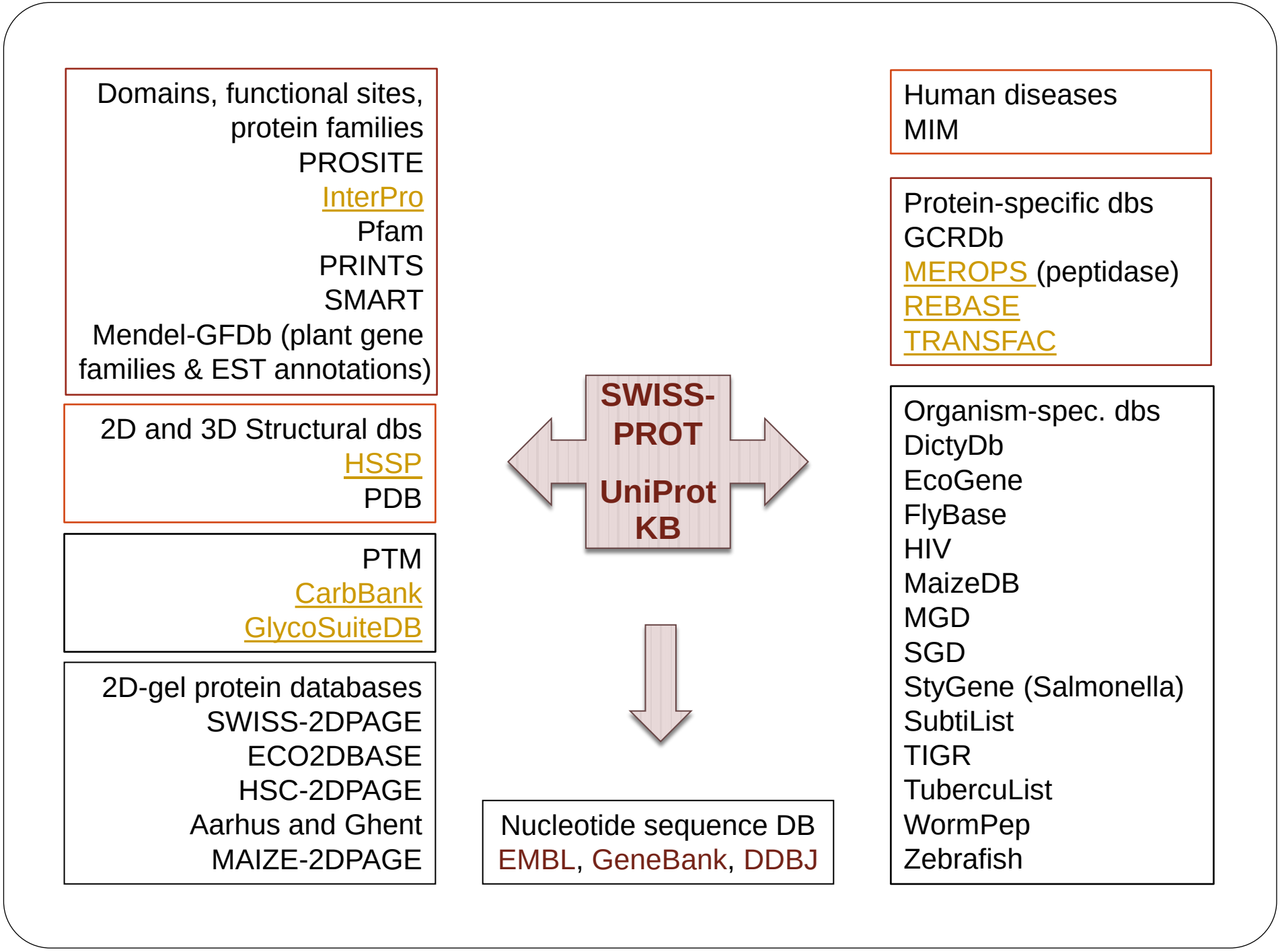
Note on Usage of the With/From Column



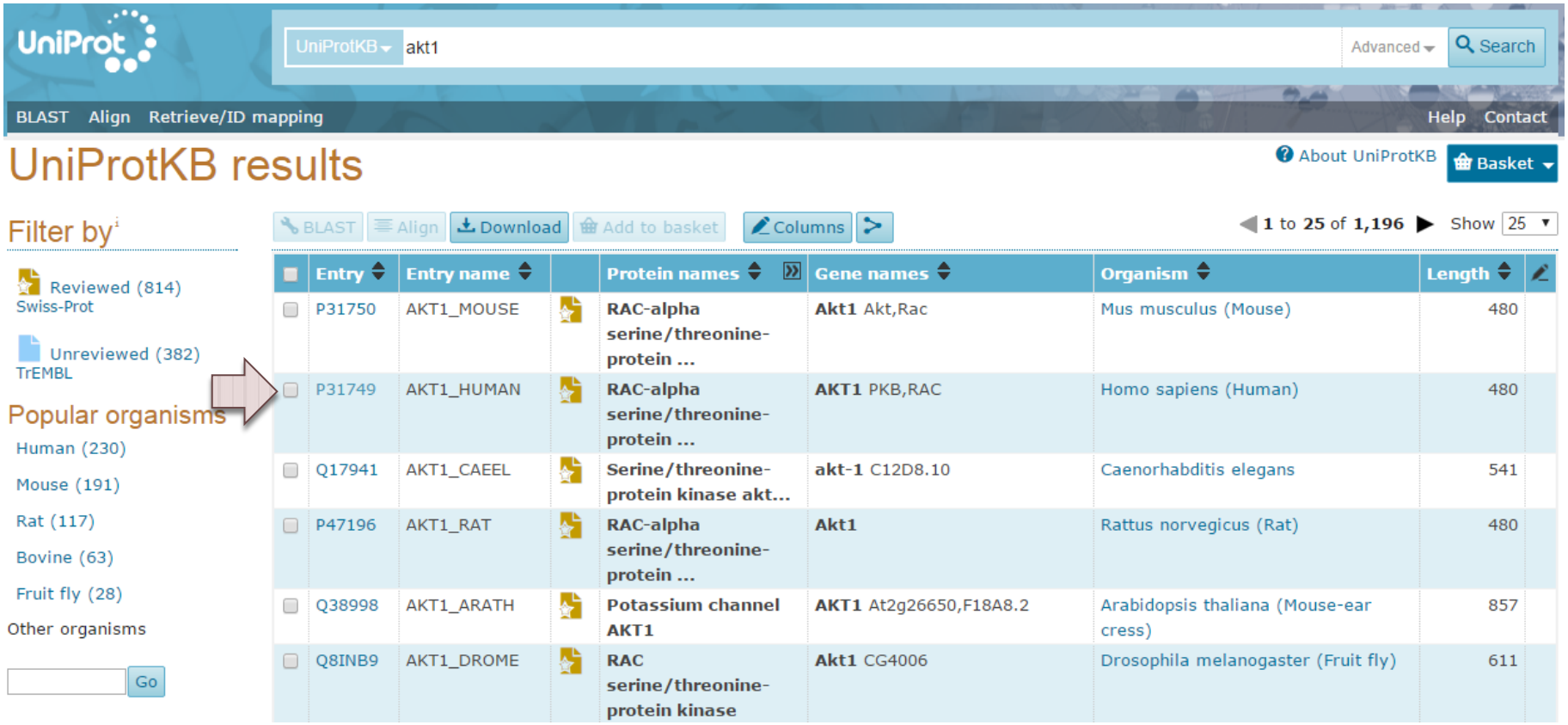
The simplified story of a SWISS-PROT entry



Once in SWISS-PROT, the entry is no more in TrEMBL, **but still in EMBL (archive)**


Domains, functional sites, protein families
PROSITE
InterPro
Pfam
PRINTS
SMART
Mendel-GFDb (plant gene families & EST annotations)
2D and 3D Structural dbs
HSSP
PDB
PTM
CarbBank
GlycoSuiteDB
2D-gel protein databases
SWISS-2DPAGE
ECO2DBASE
HSC-2DPAGE
Aarhus and Ghent
MAIZE-2DPAGE
SWISS-PROT
UniProt KB
Nucleotide sequence DB
EMBL, GeneBank, DDBJ
Human diseases
MIM
Protein-specific dbs
GCRDb
MEROPS (peptidase)
REBASE
TRANSFAC
Organism-spec. dbs
DictyDb
EcoGene
FlyBase
HIV
MaizeDB
MGD
SGD
StyGene (Salmonella)
SubtiList
TIGR
TubercuList
WormPep
Zebrafish

2. UniProt/InterProt Annotations



Display

- Entry
- Feature viewer
- Feature table

None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences (2)
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similar proteins

Top

Family & Domains

Domains and Repeats

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Domain	5 – 108	104	PH PROSITE-ProRule annotation			Add BLAST
Domain	150 – 408	259	Protein kinase PROSITE-ProRule annotation			Add BLAST
Domain	409 – 480	72	AGC-kinase C-terminal			Add BLAST

Region

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Region	14 – 19	6	Inositol-(1,3,4,5)-tetrakisphosphate binding			
Region	23 – 25	3	Inositol-(1,3,4,5)-tetrakisphosphate binding			
Region	228 – 230	3	Inhibitor binding			

Domain

Binding of the PH domain to phosphatidylinositol 3,4,5-trisphosphate (PI(3,4,5)P3) following phosphatidylinositol 3-kinase alpha (PIK3CA) activity results in its targeting to the plasma membrane. The PH domain mediates interaction with TNK2 and Tyr-176 is also essential for this interaction.
The AGC-kinase C-terminal mediates interaction with THEM4.

Sequence similarities

Belongs to the protein kinase superfamily. AGC Ser/Thr protein kinase family. RAC subfamily. Curated

Contains 1 AGC-kinase C-terminal domain. Curated

Contains 1 PH domain. PROSITE-ProRule annotation

Contains 1 protein kinase domain. PROSITE-ProRule annotation

Display

- Entry
- Feature viewer
- Feature table

None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences (2)
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similar proteins

Top

PTM / Processing

Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Chain	1 – 480	480	RAC-alpha serine/threonine-protein kinase		PRO_0000085605	Add BLAST

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Modified residue	14 – 14	1	N6-acetyllysine 1 Publication			
Modified residue	20 – 20	1	N6-acetyllysine 1 Publication			
Disulfide bond	60 ↔ 77		1 Publication			
Modified residue	124 – 124	1	Phosphoserine Combined sources			
Modified residue	126 – 126	1	Phosphoserine; alternate Combined sources			
Glycosylation	126 – 126	1	O-linked (GlcNAc); alternate 1 Publication			
Modified residue	129 – 129	1	Phosphoserine; alternate Combined sources			
Glycosylation	129 – 129	1	O-linked (GlcNAc); alternate 1 Publication			
Modified residue	176 – 176	1	Phosphotyrosine; by TNK2 1 Publication			
Cross-link	284 – 284		Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin) 1 Publication			
Disulfide bond	296 ↔ 310		By similarity			

3. If YFG Involves in Specific Function/Pathway? - through its interacted protein

BioGRID 3.1

CHD4

Mus musculus

AA617397, mKIAA4075, D6Ertd380e, Mi-2beta, BC005710, KIAA4075, 9530019N15Rik, MGC11769

chromodomain helicase DNA binding protein 4

GO Process: 0 Terms | **GO Function**: 1 Terms | **GO Component**: 0 Terms

EXTERNAL DATABASE LINKOUTS

MGI | Entrez Gene | RefSEQ | GenBank | UniprotKB

Download 13 Associations For This Protein

Stats & Filters

Current Stati

High Throughput

10 (59%)

0 (0%)

Search Filter

No Filter: Show

Switch View: Summary | Sortable Table

Displaying **13** total unique interactors

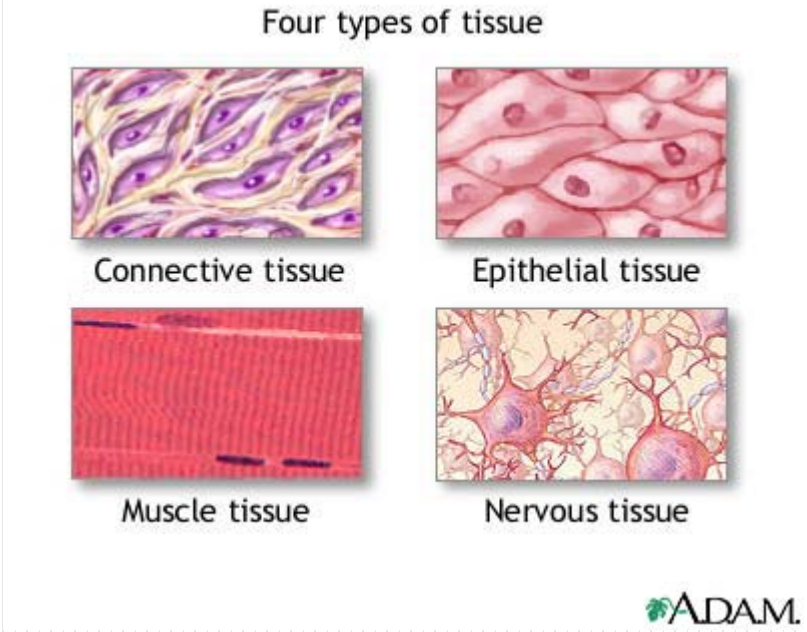
POU5F1 | Otf-3, Oct3, Oct-3/4, Otf3, Oct3/4, Oct-3, Oct4, Otf-4, Oct-4, Otf3-rs7, Otf4, Otf3g
POU domain, class 5, transcription factor 1

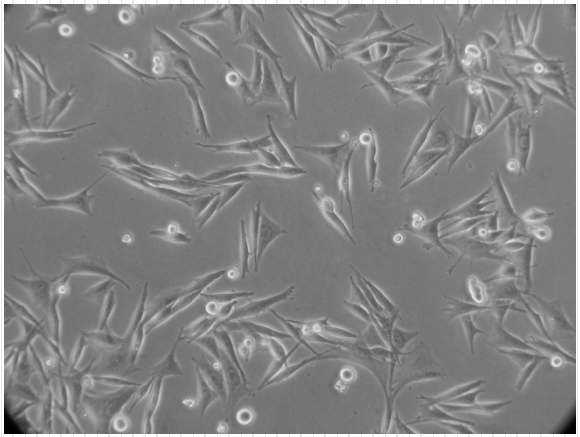
MTA2 | mmta2, Mta1l1, Mata1l1, AW550797
metastasis-associated gene family, member 2

Databases for Protein – Protein Interaction

Resource	Comments
APID	Agile Protein Interaction DataAnalyzer (Cancer Research Center, Salamanca, Spain)
BIND	Biomolecular INteraction Network Database at the University of Toronto, Canada. No species restriction
CYGD	PPI section of the Comprehensive Yeast Genome Database. Manually curated comprehensive *S. cerevisiae* PPI database at MIPS
DIP	Database of Interacting Proteins at UCLA. No species restriction.
GRID	General Repository for Interaction Datasets. Mount Sinai Hospital, Toronto, Canada
HIV Interaction DB	Interactions between HIV and host proteins.
HPRD	The Human Protein Reference Database. Institute of Bioinformatics, Bangalore, India and Johns Hopkins University, Baltimore, MD, USA.
HPID	Human Protein Interaction Database. Department of computer Science and Information Engineering Inha University, Inchon, Korea
iHOP	iHOP (Information Hyperlinked over Proteins). Protein association network built by literature mining
IntAct	Protein interaction database at EBI. No species restriction.
InterDom	Database of putative interacting protein domains. Institute for InfoComm Research, Singapore.
JCB	PPI site at the Jena Centre for Bioinformatics, Germany
MetaCore	Commercial software suite and database. Manually curated human PPIs (among other things). GeneGo
MINT	Molecular INTeraction database at the Centro di Bioinformatica Moleculare, Universita di Roma, Italy.
MRC PPI links	Commented list of links to PPI databases and resources maintained at the MRC Rosalind Franklin Cetre for Genomics Research, Cambridge, UK
OPHID	The Online Predicted Human Interaction Database. Ontario Cancer Institute and University of Toronto, Canada.
Pawson Lab	Information on protein-interaction domains.

Q: What kind of Cell Lines or Tissues I Should Use for PCR-based Cloning YFG?





NCBI

UniGene

ORGANIZED VIEW

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM

Search UniGene for POU5F1 Go Clear

Limits | Preview/Index | History | Clipboard | Details

UniGene Homepage

UniGene: An Organized View of the Transcriptome.

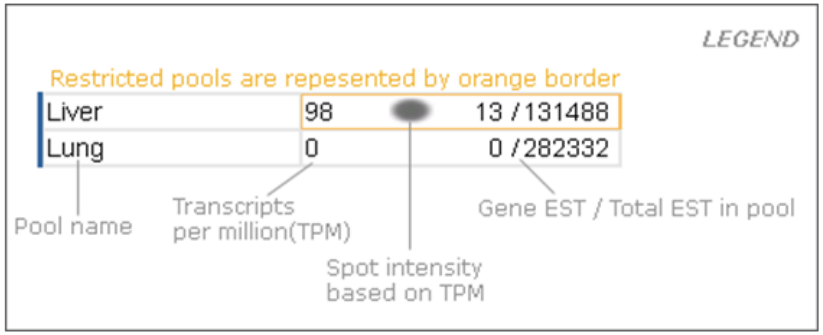
ascites	0		0 /40015
bladder	0		0 /29757
blood	0		0 /123478
bone	0		0 /71655
bone marrow	0		0 /48801
brain	1		2 /1100989
cervix	0		0 /48171
connective tissue	0		0 /149255
ear	0		0 /16212
embryonic tissue	407		88 /215722
esophagus	0		0 /20209
eye	0		0 /211054
heart	0		0 /89626
intestine	0		0 /234472
kidney	0		0 /211777
larynx	0		0 /24145
liver	4		1 /207743
lung	8		3 /336974
lymph	0		0 /44270
lymph node	0		0 /91610
mammary gland	6		1 /153271
mouth	0		0 /67052
muscle	0		0 /107715
nerve	0		0 /15768
ovary	58		6 /102051
pancreas	0		0 /214812
parathyroid	0		0 /20539
pharynx	0		0 /41328
pituitary gland	0		0 /16585
placenta	0		0 /280825

Hs.249184

embryoid body	14		1 /70761
blastocyst	754		47 /62319
fetus	0		0 /564012
neonate	0		0 /31097
infant	0		0 /23620
juvenile	0		0 /55556
adult	8		17 /1939121

- Hs.249184 representation biased toward **blastocyst** [more like this]

EST profiles show **approximate** gene expression patterns as inferred from EST counts ar
normalized, subtracted, or otherwise biased have been removed, but for a variety of reaso



Q: What Would I Do When I am having Breakfast or having a Coffee Break?

Coffee Break

Tutorials for NCBI Tools

Edited by Laura Dean and Johanna McEntyre.

National Center for Biotechnology Information

Bethesda (MD): National Center for Biotechnology Information (US); 1999-.

Coffee Break is a resource at NCBI that combines reports on recent biomedical discoveries with use of NCBI tools. The result is an interactive tutorial that tells a biological story. Each report is based on a discovery reported in one or more articles from the recently published peer-reviewed literature. After a brief introduction that sets the work described into a broader context, the report focuses on how a molecular understanding can provide explanations of observed biology and lead to therapies for diseases.

Bookshelf ID: NBK1969

NCBI News

Bethesda (MD): National Center for Biotechnology Information (US); 199
ISSN: 1060-8788
Publication No.: 94-3272

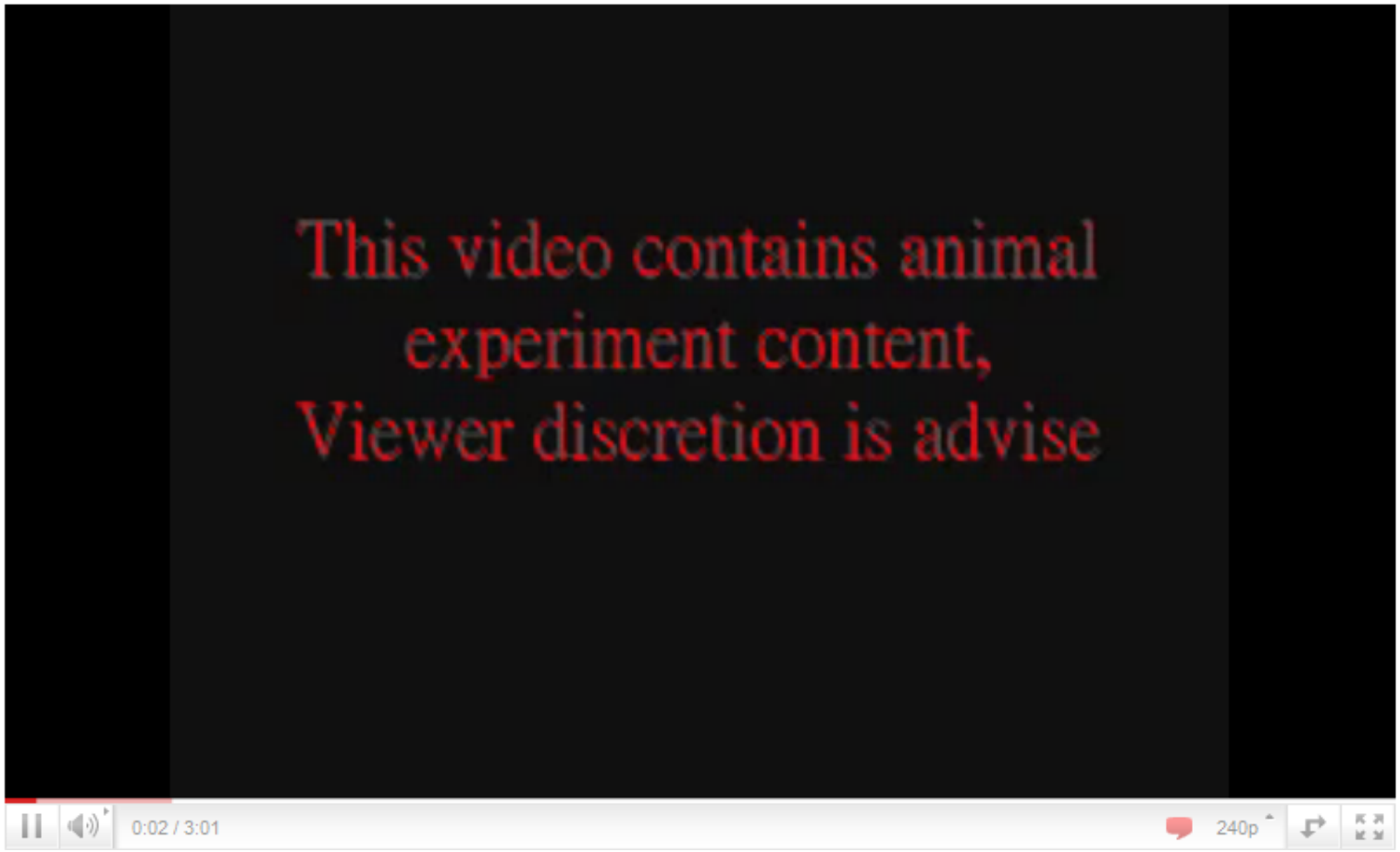
Index of Issues

Expand All

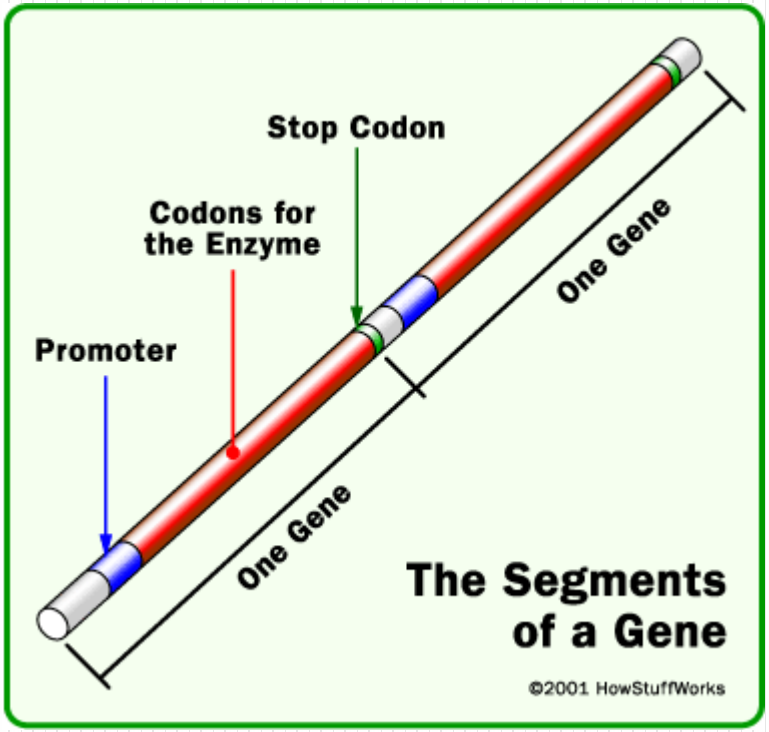
liver tumor mouse

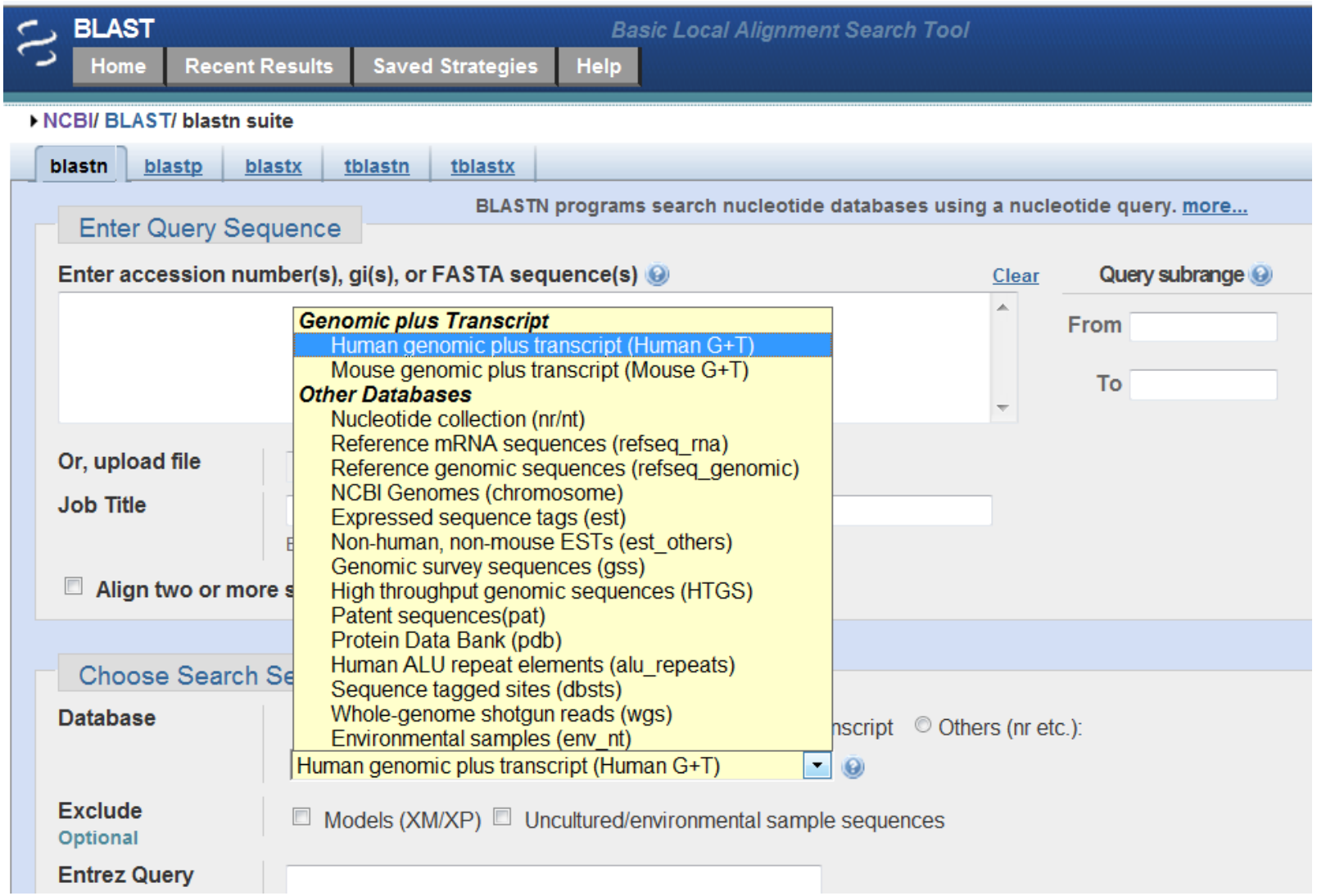
Seed tumor at liver of mouse-Surgery 種腫瘤在肝臟-開腹腔篇(一)

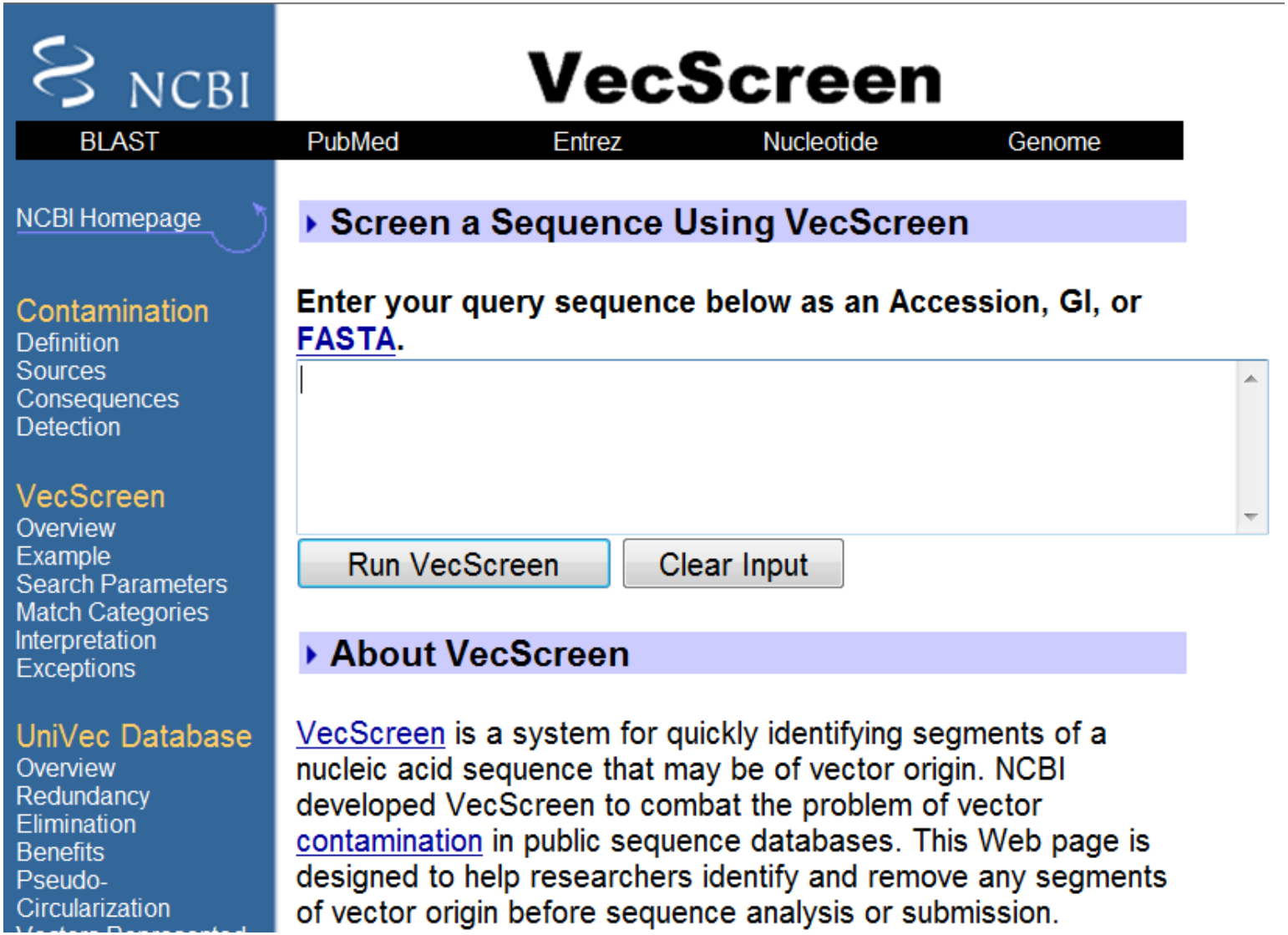
miss9ch 282 部影片 訂閱



Q: How do You Know You've Cloned the Correct YGF? (wild type vs. mutant?)




BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Job Title
Align two or more s
Choose Search Se
Database
Genomic plus Transcript
Human genomic plus transcript (Human G+T)
Mouse genomic plus transcript (Mouse G+T)
Other Databases
Nucleotide collection (nr/nt)
Reference mRNA sequences (refseq_rna)
Reference genomic sequences (refseq_genomic)
NCBI Genomes (chromosome)
Expressed sequence tags (est)
Non-human, non-mouse ESTs (est_others)
Genomic survey sequences (gss)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
Protein Data Bank (pdb)
Human ALU repeat elements (alu_repeats)
Sequence tagged sites (dbsts)
Whole-genome shotgun reads (wgs)
Environmental samples (env_nt)
nscript
Others (nr etc.):
Human genomic plus transcript (Human G+T)
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query


NCBI
VecScreen
BLAST
PubMed
Entrez
Nucleotide
Genome
NCBI Homepage
Contamination
Definition
Sources
Consequences
Detection
VecScreen
Overview
Example
Search Parameters
Match Categories
Interpretation
Exceptions
UniVec Database
Overview
Redundancy
Elimination
Benefits
Pseudo-
Circularization
Screen a Sequence Using VecScreen
Enter your query sequence below as an Accession, GI, or FASTA.
Run VecScreen
Clear Input
About VecScreen
VecScreen is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin. NCBI developed VecScreen to combat the problem of vector contamination in public sequence databases. This Web page is designed to help researchers identify and remove any segments of vector origin before sequence analysis or submission.

NCBI

ORF Finder (Open Reading Frame Finder)

PubMed | Entrez | BLAST | OMIM | Taxon

NCBI

Tools
for data mining

GenBank
sequence submission support and software

FTP site
download data and software

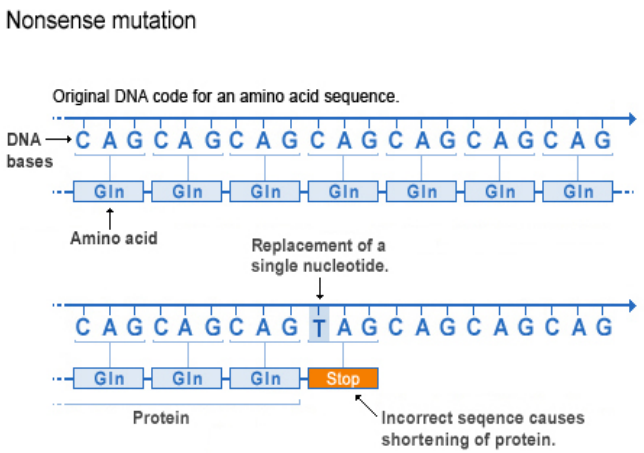
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which fi
selectable minimum size in a user's sequence or in a sequence already in the datal
This tool identifies all open reading frames using the standard or alternative genetic
sequence can be saved in various formats and searched against the sequence dat
The ORF Finder should be helpful in preparing complete and accurate sequence su
the Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

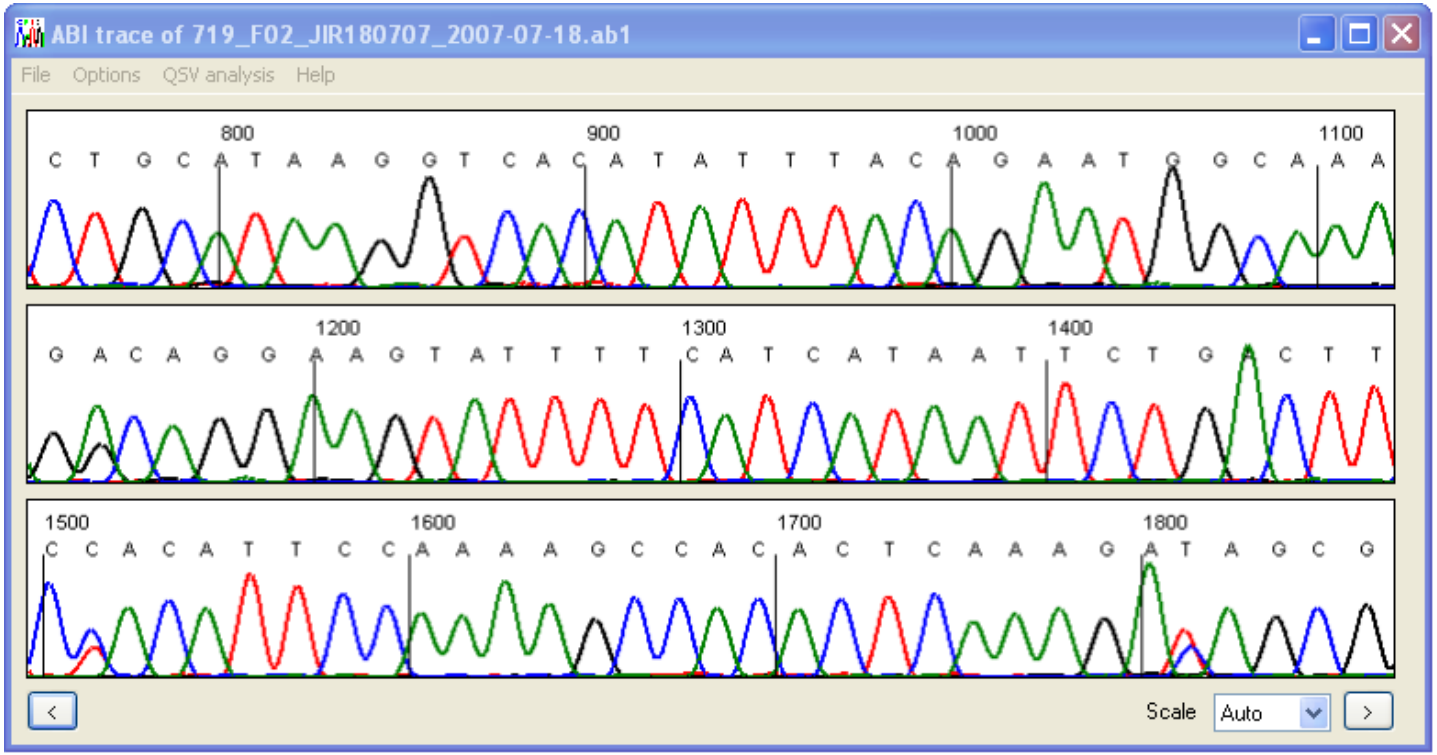
or sequence in FASTA format

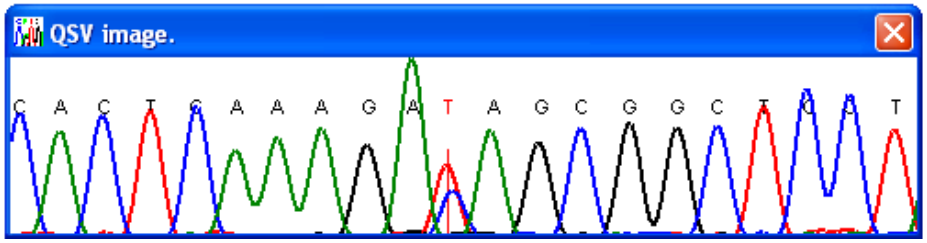
FROM: **TO:**

Genetic codes 1 Standard



U.S. National Library of Medicine


ABI trace of 719_F02_JIR180707_2007-07-18.ab1
File
Options
QSV analysis
Help
800
900
1000
1100
C T G C A T A A G G T C A C A T A T T T A C A G A A T G G C A A A
1200
1300
1400
G A C A G G A A G T A T T T T C A T C A T A A T T C T G C T T
1500
1600
1700
1800
C C A C A T T C C A A A A G C C A C A C T C A A A G A T A G C G
Scale
Auto


QSV image.
C A C T G A A A G A T A G C G G C T C A T

When Cloned by Emails – get the map & confirmed

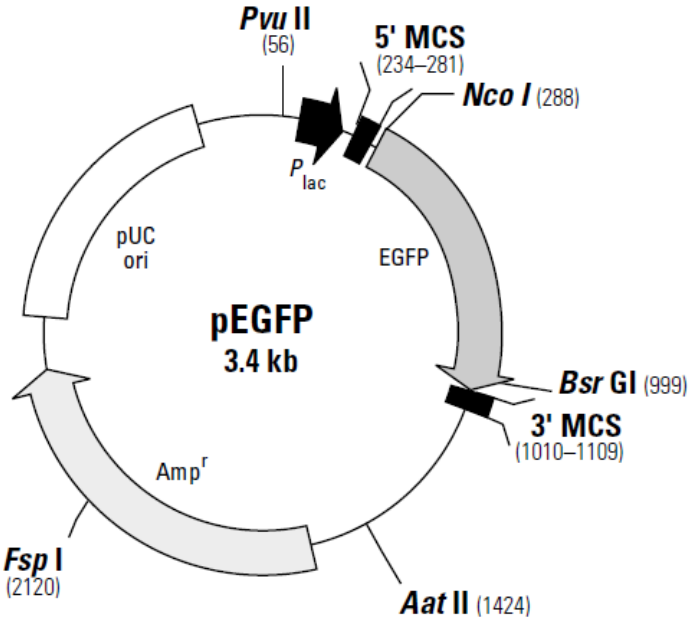
Specific EGFP Monoclonal Antibody for Westerns, IP and IC

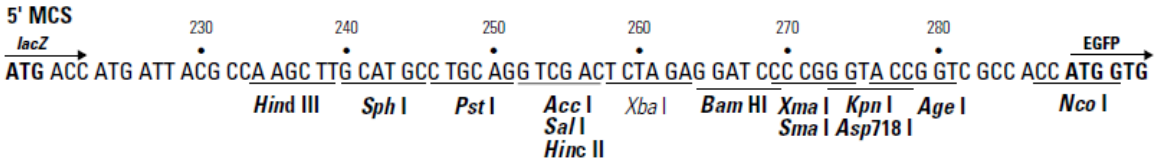
Visit our website for more details! click here...

pEGFP Vector Information

PT3078-5

Catalog #6077-1





Q: How to Get a Specific Sequence from Genome Databases

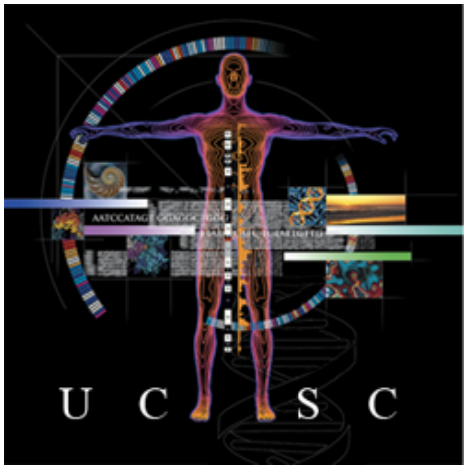
Genome Biology

- Vertebrates (17)
  - Mammals (14)
    - Primates (3)

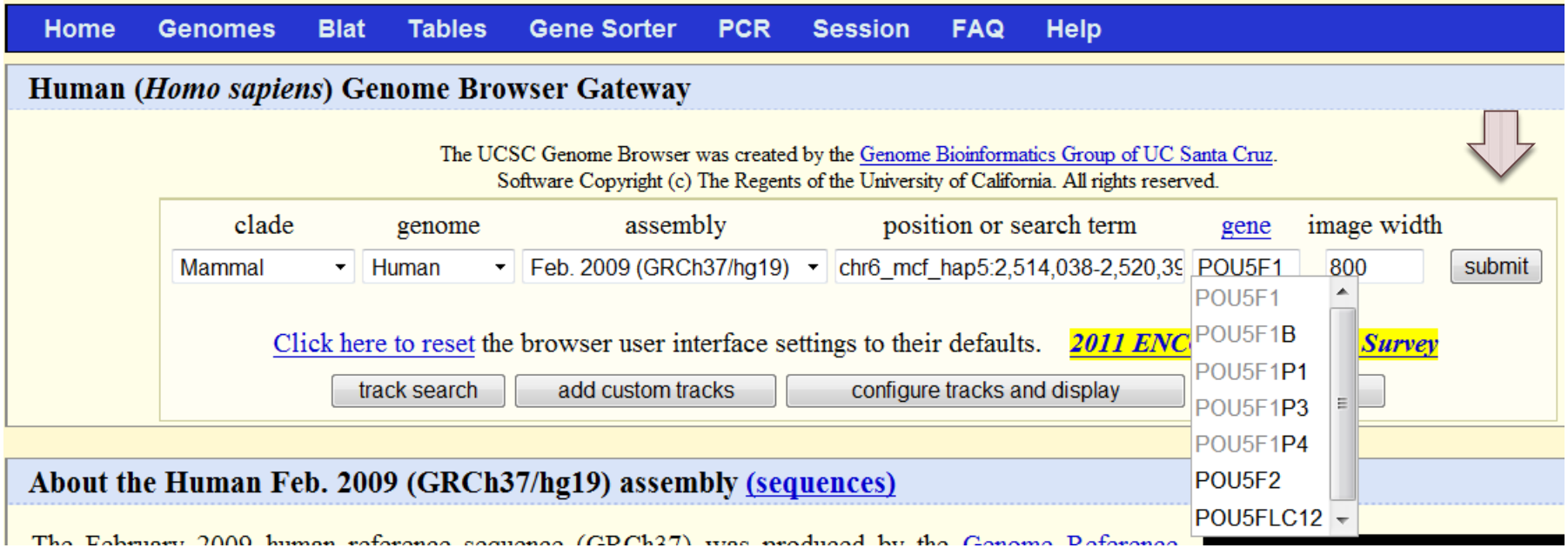
Map Viewer, NCBI

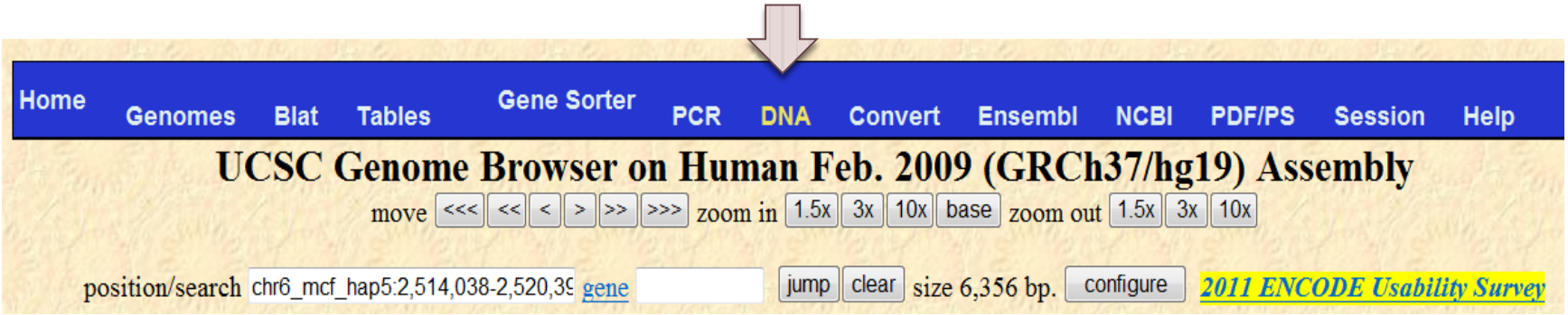
Genome Browser, UCSC

Ensembl Genome Browser, EBI



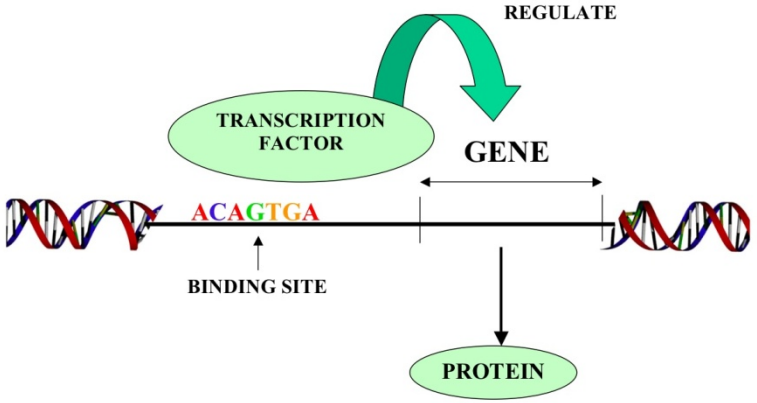
e!Ensembl

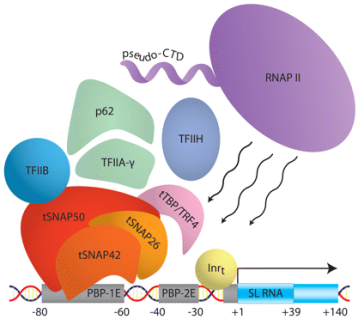

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help
Human (Homo sapiens) Genome Browser Gateway
The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.
clade genome assembly position or search term gene image width
Mammal Human Feb. 2009 (GRCh37/hg19) chr6_mcf_hap5:2,514,038-2,520,39 POU5F1 800 submit
POU5F1
POU5F1B
POU5F1P1
POU5F1P3
POU5F1P4
POU5F2
POU5FLC12
Click here to reset the browser user interface settings to their defaults.
2011 ENC Survey
track search
add custom tracks
configure tracks and display
About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr6_mcf_hap5:2,514,038-2,520,39 gene jump clear size 6,356 bp. configure
2011 ENCODE Usability Survey

Q: How to Identify Potential Regulators?

Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.

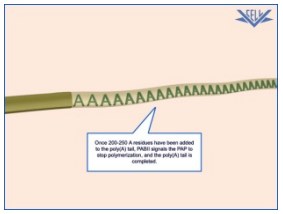


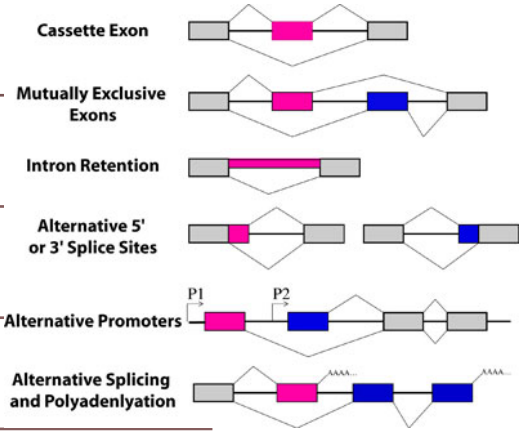


Feature-Based Methods

Based on identifying **gene signals**

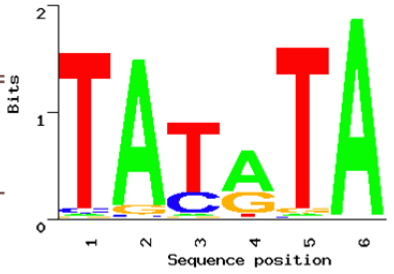
- **Promoter elements**
- Splice sites
- **Start/stop** codons
- PolyA sites…

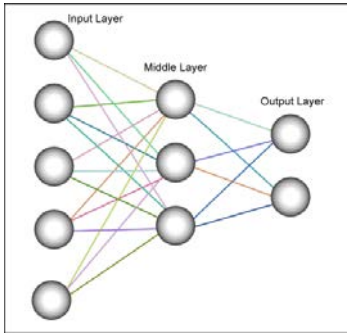


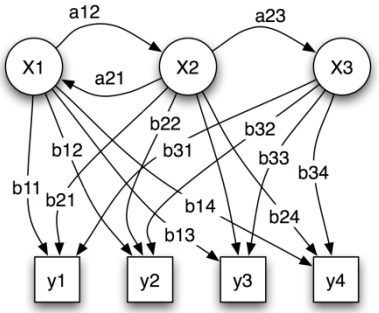


Wide range of **methods**

- Consensus sequences
- Weight matrices
- **Neural networks (NNs)** Decision trees
- **Hidden Markov Models (HMMs)**



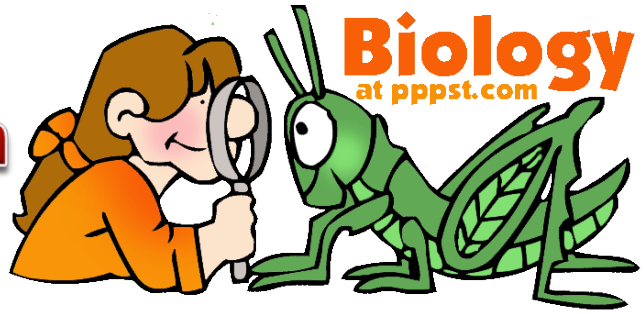




Promoter Databases and sites for analysis, prediction and search

AlignACE	motif-finding algorithm.
Promoter Binding Element Database	Arabidopsis thaliana promoter binding element database
CpG promoter	promoter mapping using CpG islands
Core promoter	to predict putative Transcriptional Start Site (TSS)
dbtss	Database of Transcriptional Start Sites
Dragon Promoter Finder	an advanced system for promoter recognition in vertebrates
EPD	an annotated non-redundant collection of eukaryotic POL II promoters
FirstEF	a 5' terminal exon and promoter prediction program
Human Promoter Database	Search for transcriptional start site
Mcpromoter	A statistical tool for the prediction of transcription start sites
Motif Explorer	Motif & promoter visualization
Neural Network Promoter Prediction	Neural Network Promoter Prediction

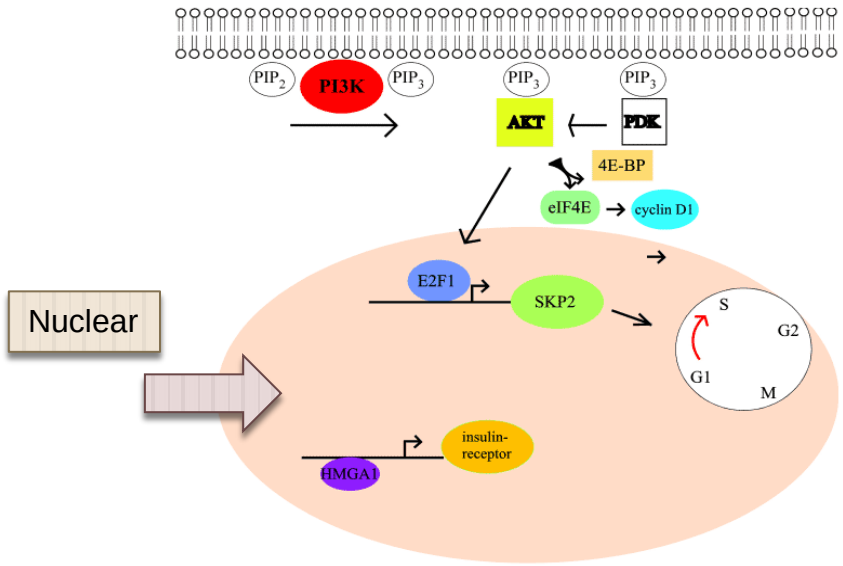
Pattern-driven



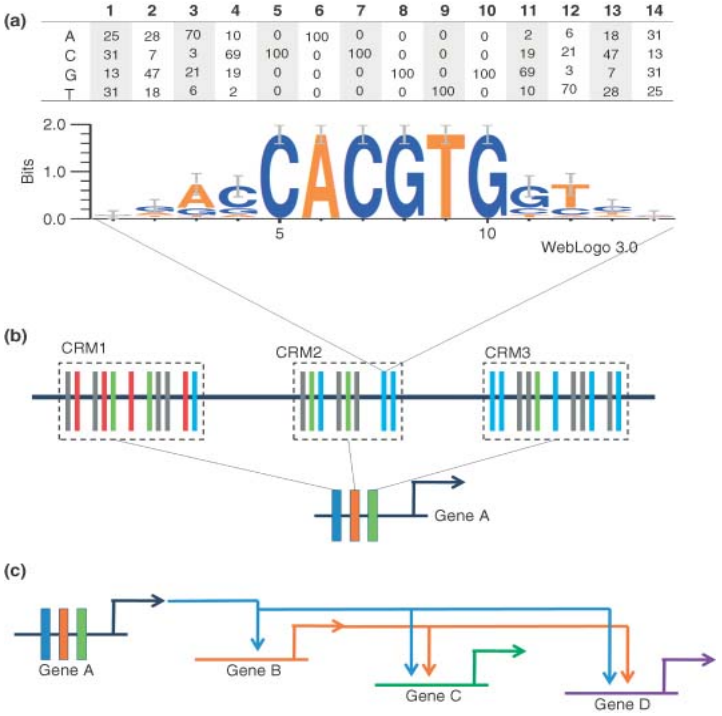
Success depends on **available of collections** of **annotated binding sites**

- Tend to produce huge numbers of **false-positive**

- **Reasons**
  - Binding sites (BS) for specific TFs often **variable**
  - Binding sites are short (typically **5-15 bp**)
  - **Interactions** between TFs (& other proteins) influence **affinity** & **specificity** of TF binding
  - One binding site often recognized by **multiple TFs**
  - **Biology is complex**: promoters often specific to **organism/cell/stage/environmental** condition



PI3K/AKT signaling in pancreatic cancer cells

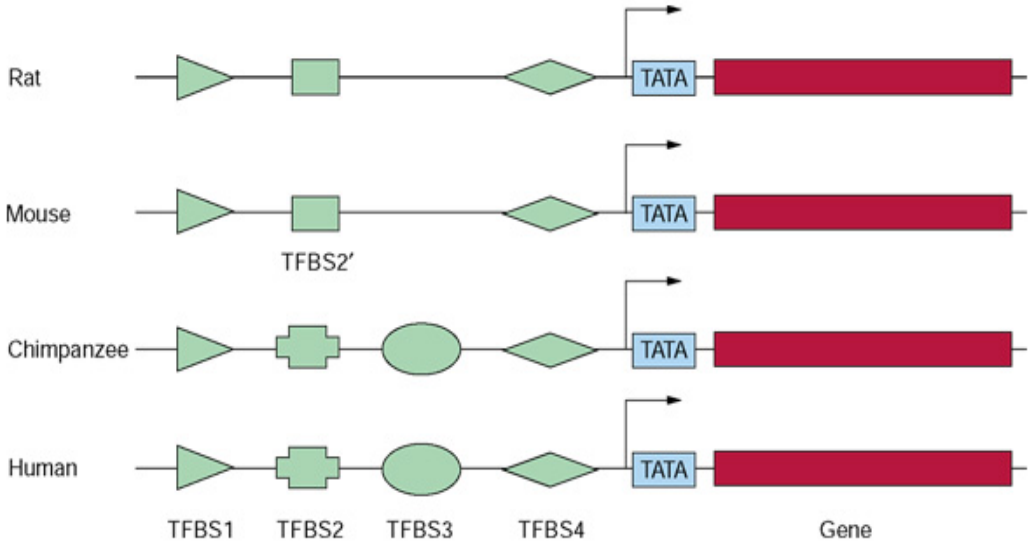


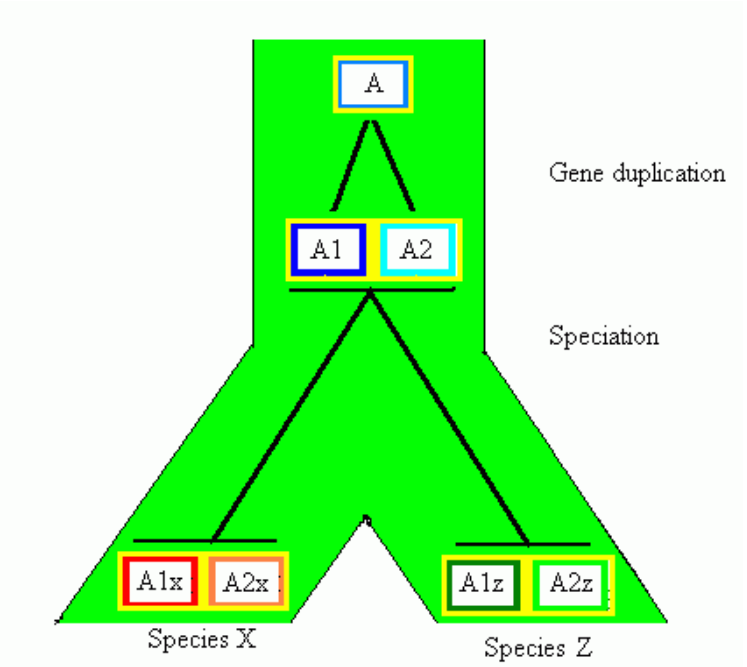
Taking **sequence context/biology** into account (Do the **wet lab** experiments!!!)		
Eukaryotes: clusters of TFBSs are common	**Probability** of "real" binding site increases if annotated **transcription start site (TSS) nearby** • But **NOT** for enhancers • Only **a small fraction of TSSs** have been experimentally mapped	**Comparative** promoter mapping

Phylogenetic Footprinting

Patterns of gene regulation are often conserved across species

- Interspecies comparisons ⇨ to identify **common regulatory sequences** (Wasserman et al. 2000)
  - The selection of appropriate species, critical





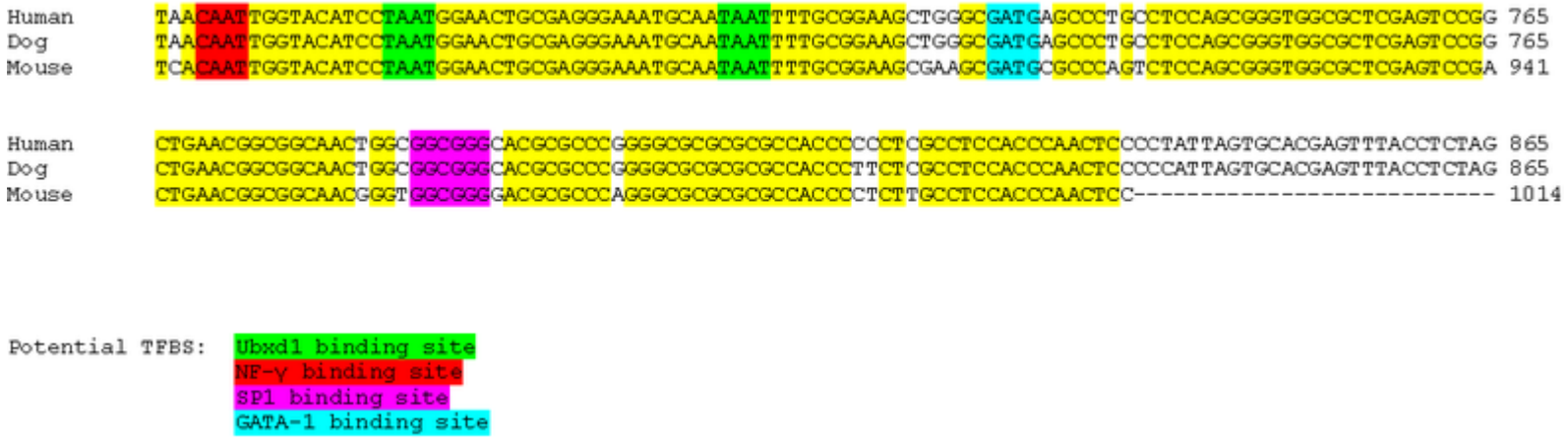
To select gene of interest

To choose **several species** with **the orthologous gene**

To decide on **the length of upstream region** to be compared

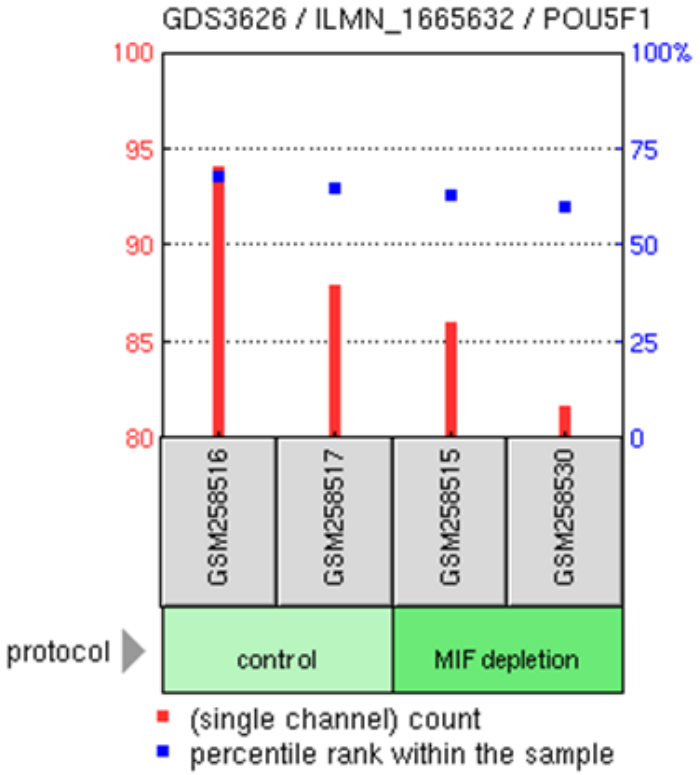
Align sequences by using **any** basic computer software (e.g., clustalW)

Visually look for **identical motif**

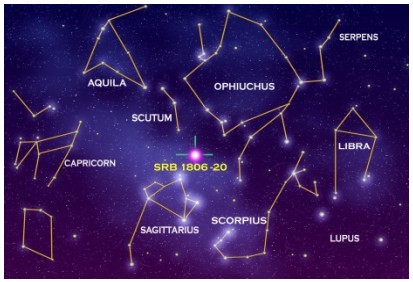


One More Trick - Coregulation

Title: GDS3626 / ILMN_1665632 / POU5F1 / Homo sapiens
Summary: Analysis of HEK293 kidney cells depleted for the
(0)/G(1) cell cycle arrest. Results provide insight into the mo



Constellation of NCBI Gene Expression Resources

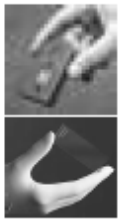


BLAST DB
Spot DB
ProbeSet DB
Profile display
BLAST ProbeSet
Profile display
Entrez ProbeSet
Entrez DB
GEO DB
DATA DEPOSITS
Accession Display
GEO
Tag display
Gene display
SAGEmap
MapViewer DB
SAGE DB
DATA DEPOSITS
SAGEmap BUILD
Genome display
MapViewer

Gene Expression Omnibus (GEO) (1)

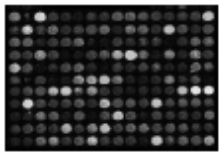
*Submitted by Manufacturer**

GPL
Platform descriptions



Submitted by Experimentalists

GSM
Raw/processed spot intensities from a single slide/chip



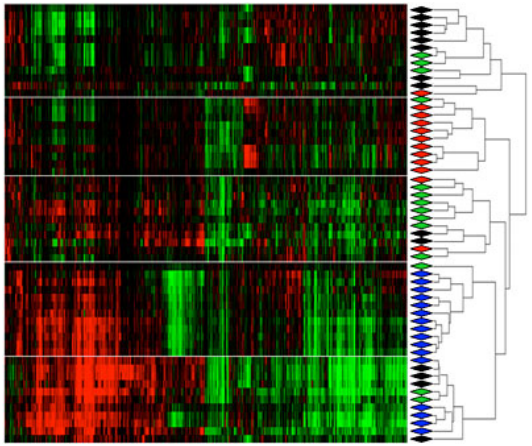
GSE
Grouping of slide/chip data
"a single experiment"

Entrez GEO

Curated by NCBI

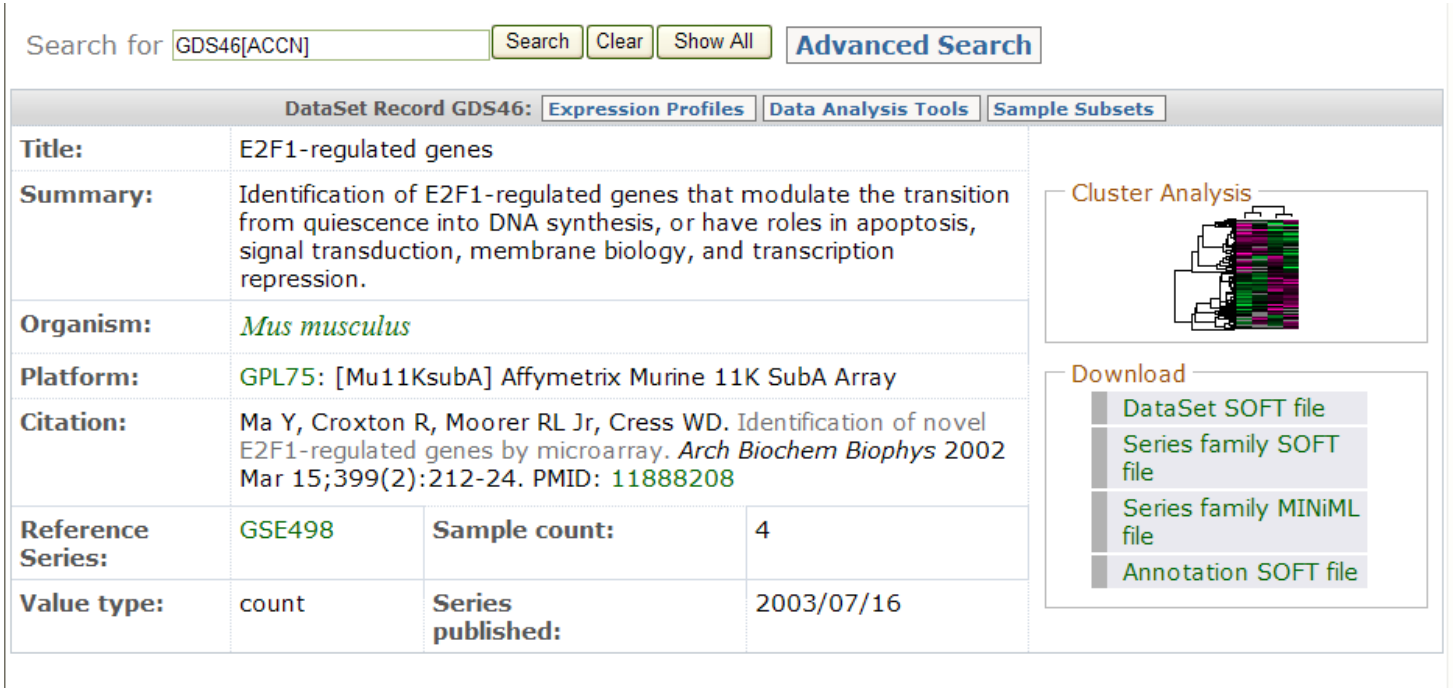
GDS
Grouping of experiments

Entrez GEO Datasets

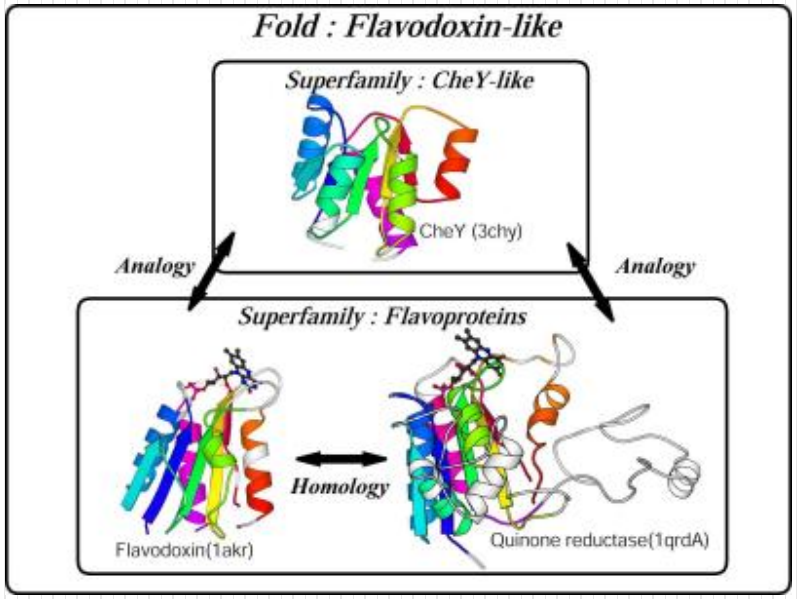


Gene Expression Omnibus (GEO) (2)

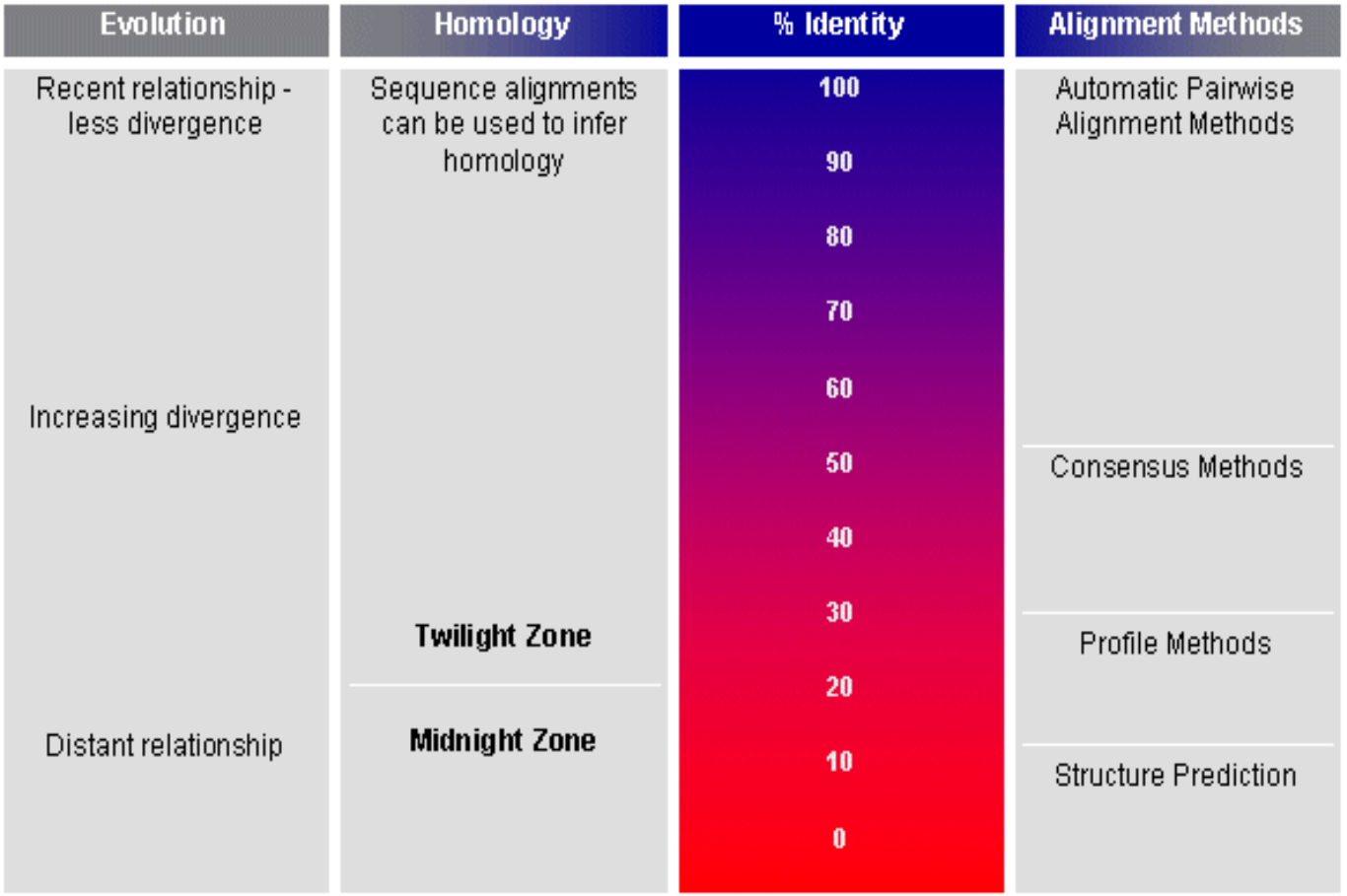
- Search GEO Profiles: POU5F1
  - Or **Limit**, **Preview / Index**
- GDS vs. GSE



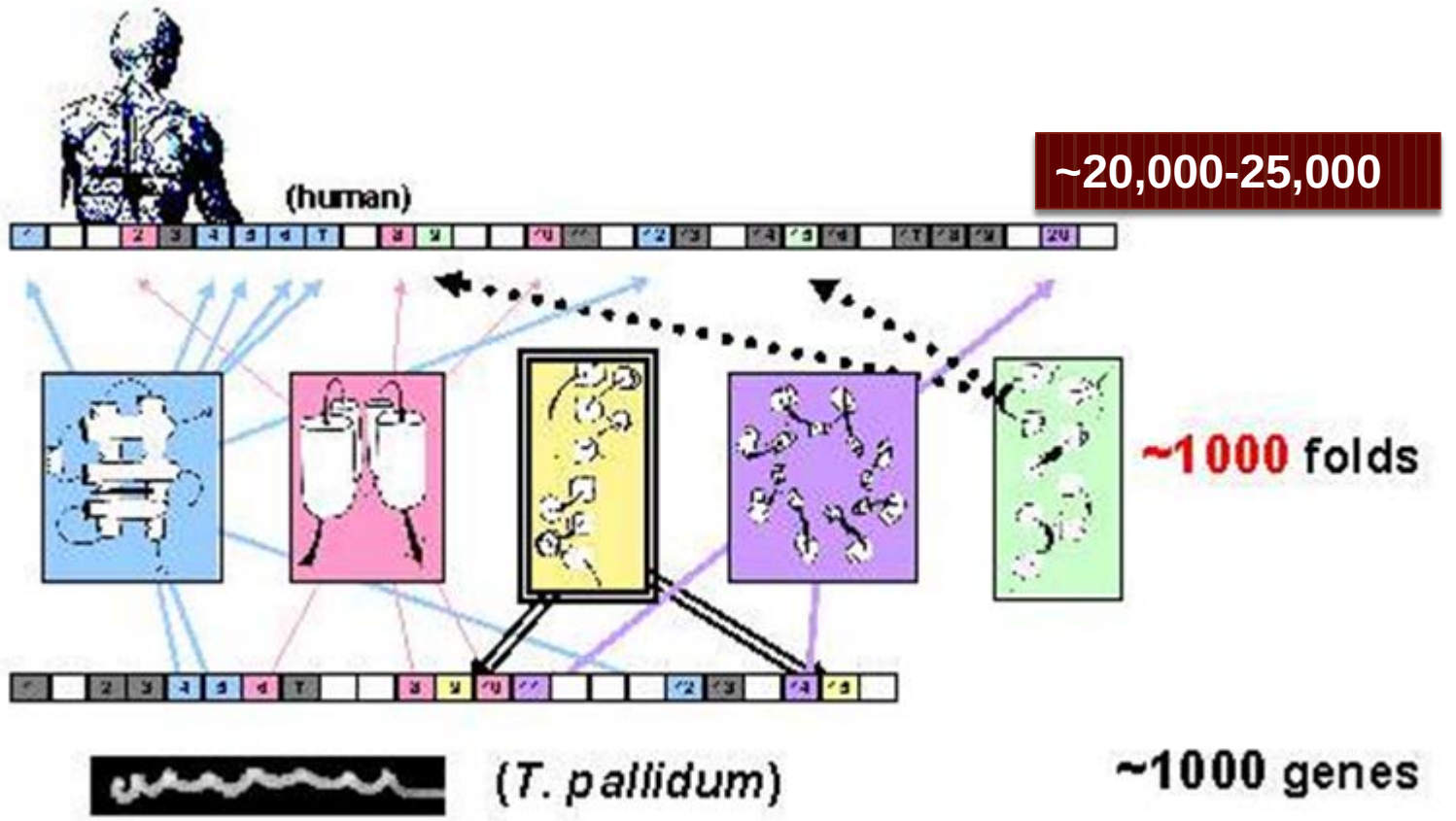
Q: Can You Speculate the Function of YFG from Structure Similarity?



Structures are More Conserved Than Sequences



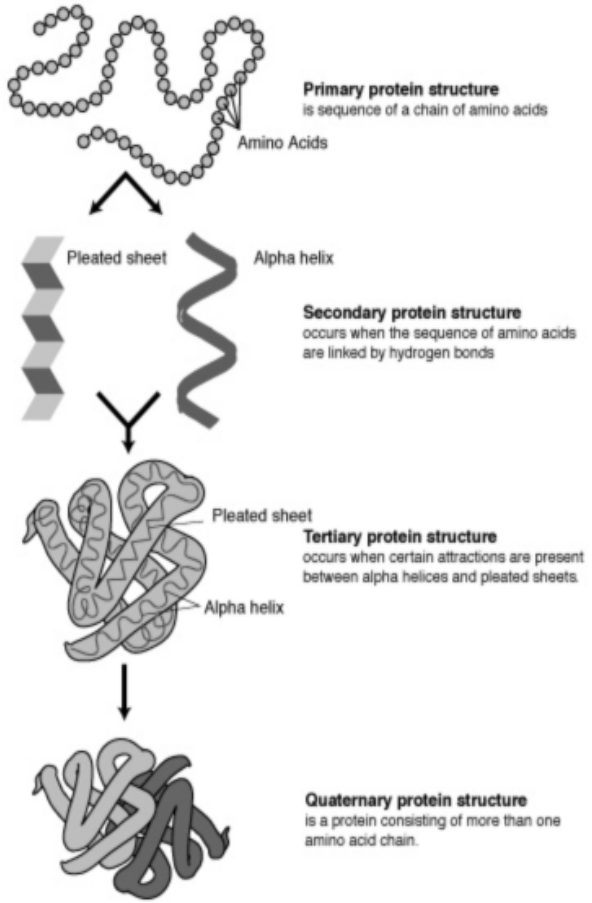
Simplifying Genomes with Folds, Pathways

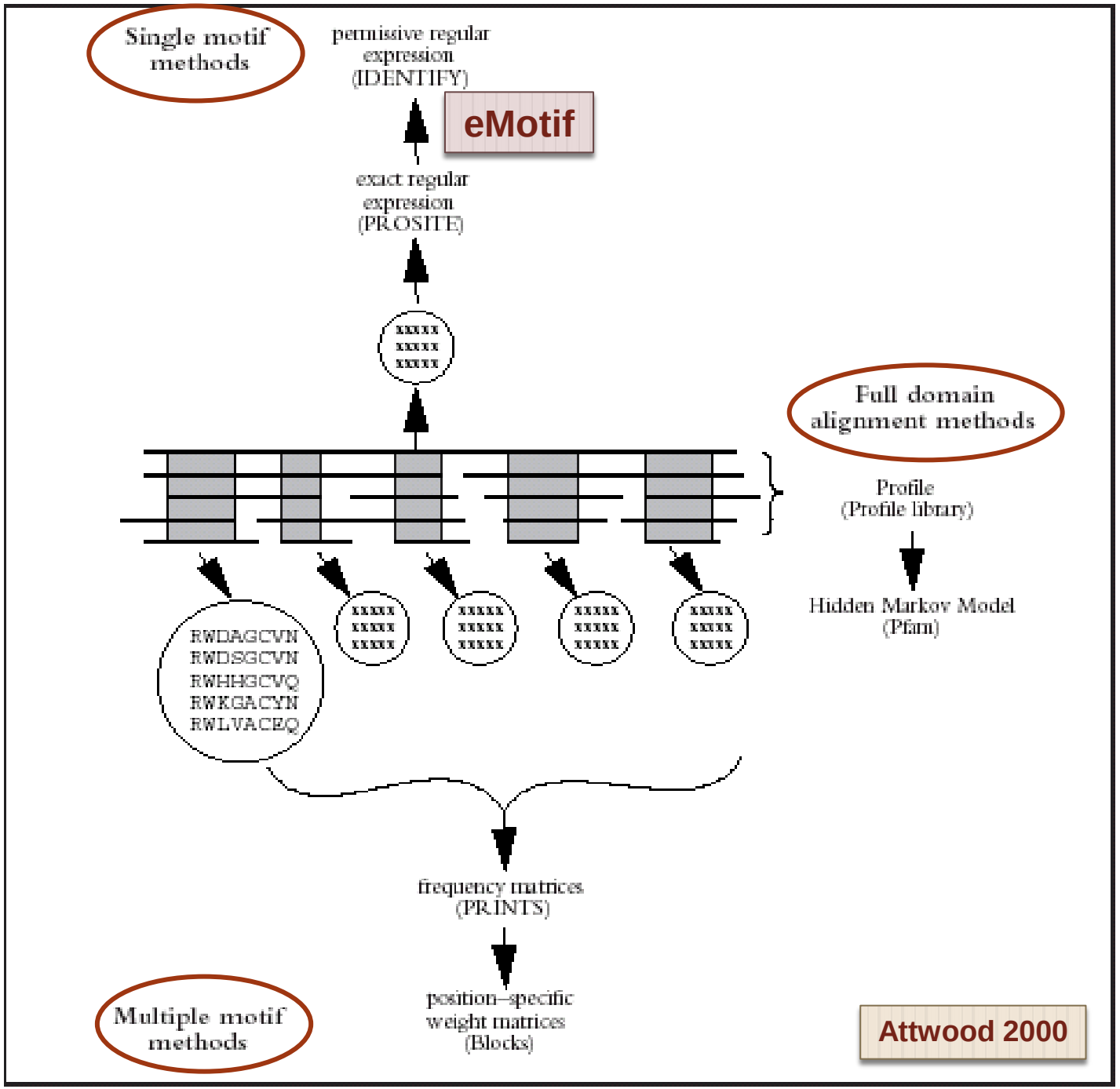


Significance: fold # << sequence ##

Levels of Protein Sequence & Structure Organization

Level/ Database	Content	Example
Primary	Sequence	“AVILDRYFH”
Secondary	Motif	[AS]-[IL]2-X[DE]-R-[FYW]2-H
Tertiary	Domain/ module	a,b,c or @, *, #




Single motif methods
permissive regular expression (IDENTIFY)
eMotif
exact regular expression (PROSITE)
XXXXX
XXXXX
XXXXX
Full domain alignment methods
Profile
(Profile library)
Hidden Markov Model
(Pfam)
RWDAGCVN
RWDSGCVN
RWHHGCVQ
RWKGACYN
RWLVACEQ
frequency matrices
(PRINTS)
position–specific weight matrices
(Blocks)
Multiple motif methods
Attwood 2000

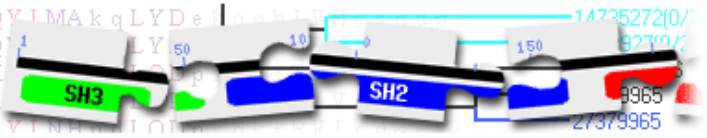
Major Secondary "Pattern" Database

2nd Database	Primary Source	Stored Information
PROSITE	SWISS-PROT	Regular expression (**pattern**)
PROSITE	BLOCKS+/Prints	**Fuzzy** expression (**pattern**)
PRINTS	SWISS-PROT/ TrEMBL	Aligned motifs - fingerprints
Profiles (**Prosite**)	SWISS-PROT	Weighted matrices (**profiles**)
Pfam/**SMART**	SWISS-PROT	Hidden Markov Models (**HMMs**)
Conserved Domain Database (**CDD**)	NCBI	Position-specific scoring matrices (**PSSMs**)

http://www.ngbw.org/database_docs/CDD.pdf

NCBI > Structure Home > Conserved Domains

NCBI



HOME | SEARCH | GUIDE | Structure Home | 3D Macromolecular Structures | Conserved Domains

Search Conserved Domains for GO CLEAR Help

Conserved Domains and Protein Classification

RESOURCES | SEARCH | HOW

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly to define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

Search | How To | Help | News | FTP | Publications

CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**.

CD-Search | Batch CD-Search | Help | FTP | Publications

送出查詢 Search Database CDD v2.28 - 39357 PSSMs

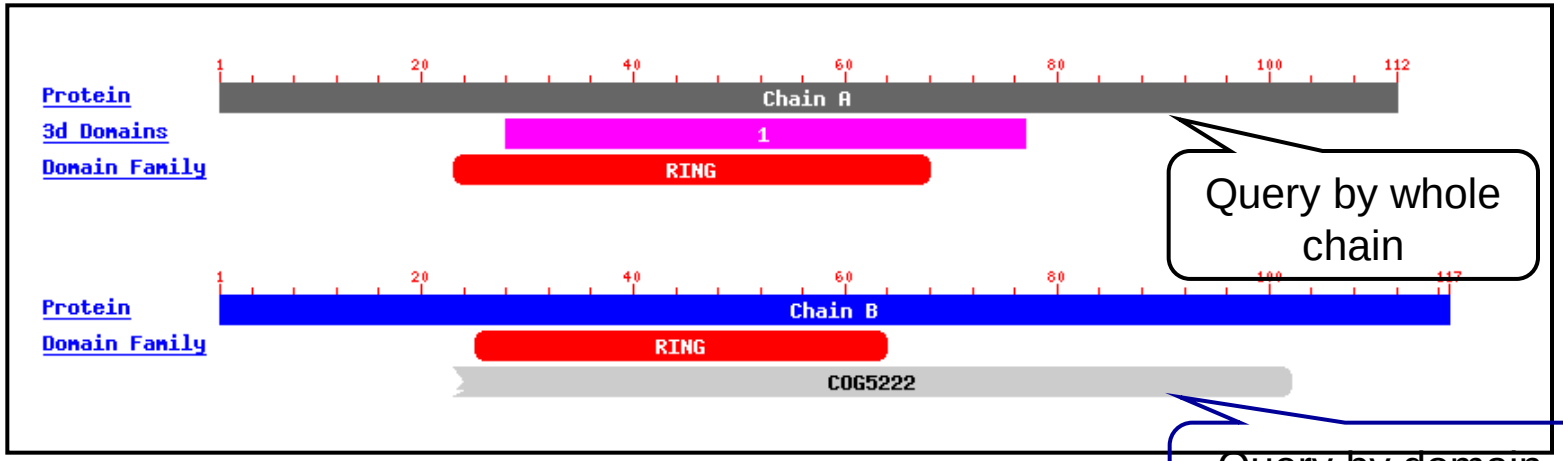
- CDD v2.28 - 39357 PSSMs
- SMART v5.1 - 791 PSSMs
- Pfam v24.0 - 11912 PSSMs
- COG v1.00 - 4873 PSSMs
- KOG v1.00 - 4825 PSSMs
- PRK v6.00 - 10885 PSSMs
- TIGR v10.00 - 4023 PSSMs

CDART: Domain Architectures

Conserved Domain Architecture ...orms similarity searches of the Entrez Protein database based on domain arch ...ential order of conserved domains in protein queries. CDART finds protein sir ...volutionary distances using sensitive domain

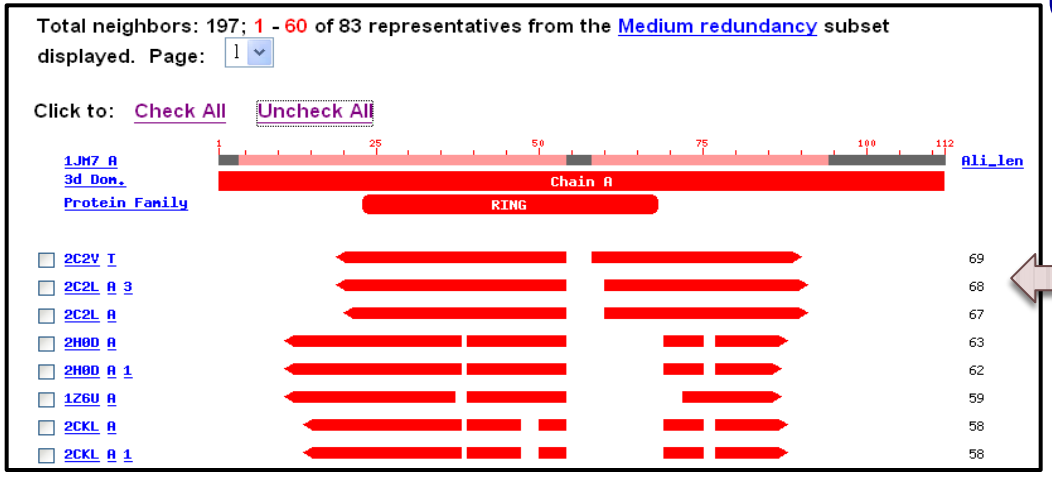
Highlig

VAST: Query by Chain or 3D Domain



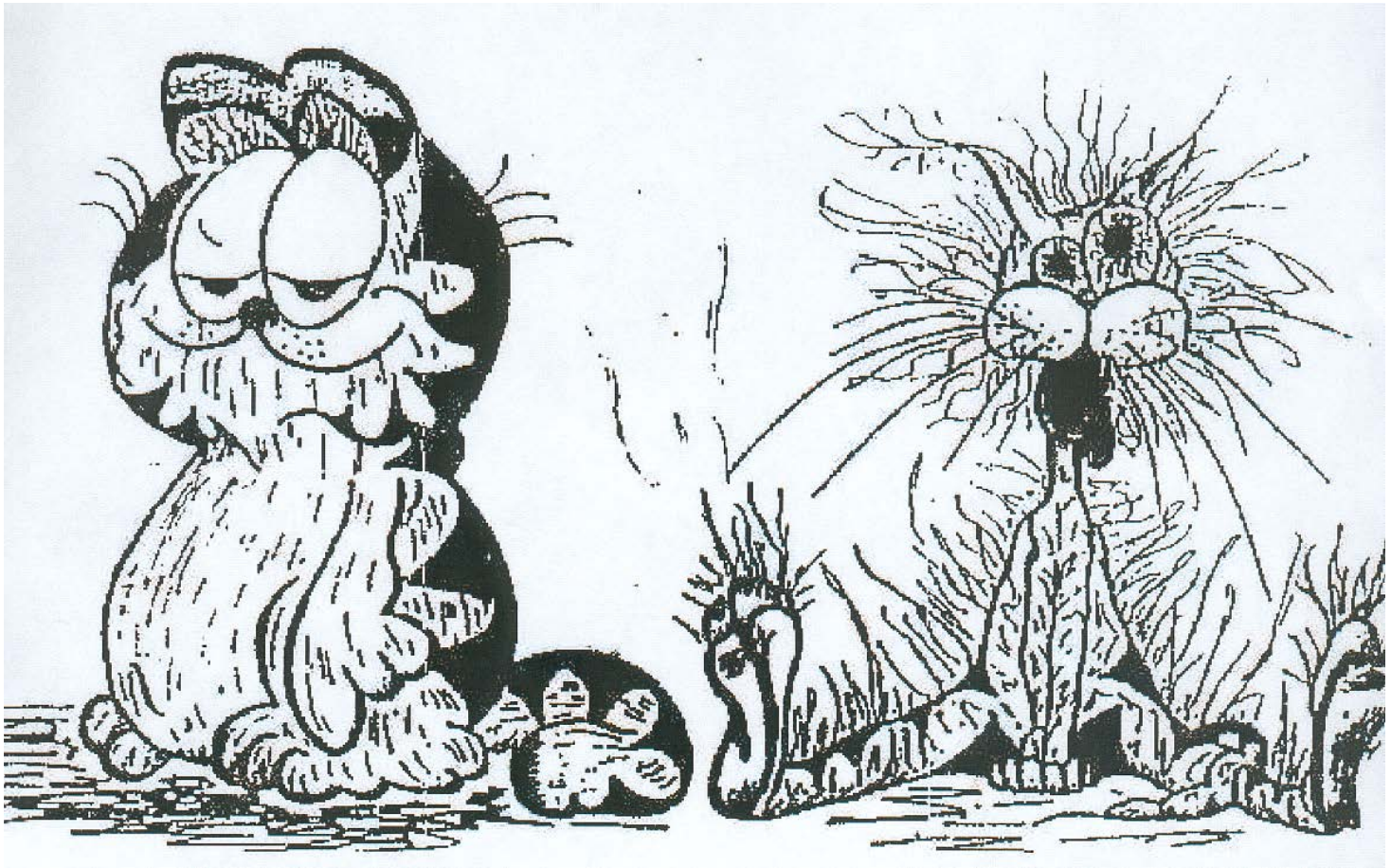
Query by whole chain

Query by domain COG5222



Not found with chain query

Before…



After…