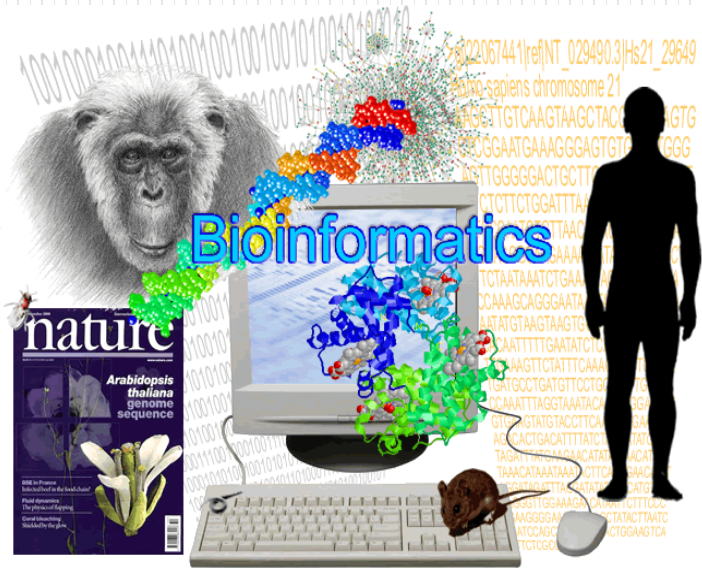


# Bioinformatics & Functional Genomics



薛佑玲 PhD

Institute of Biomedical Sciences

National Sun Yat-sen University

[ylshiu@mail.nsysu.edu.tw](mailto:ylshiu@mail.nsysu.edu.tw)

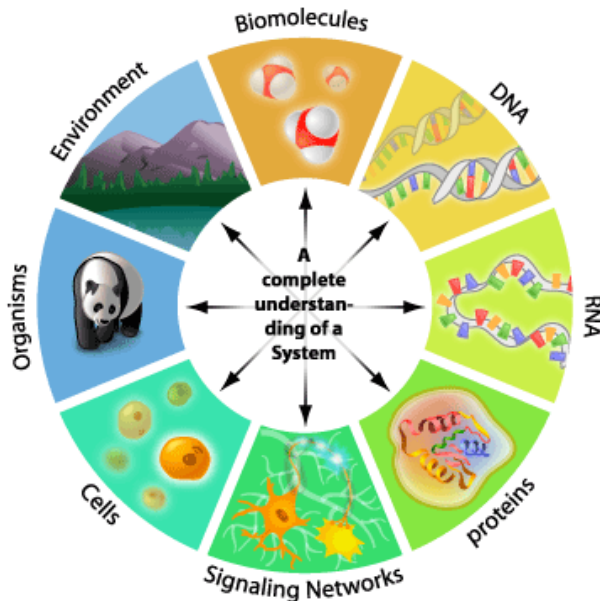
# Outline

---

## Introduction: a Short History About Bioinformatics

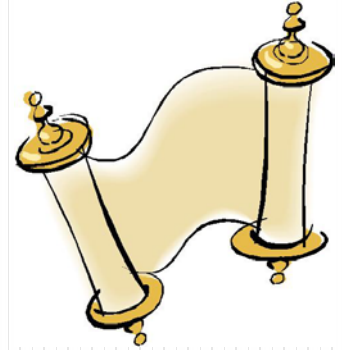
---

## Bioinformatics Q & A

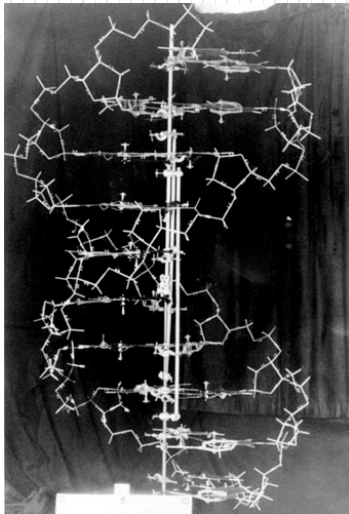


b101nf0rmat1cs

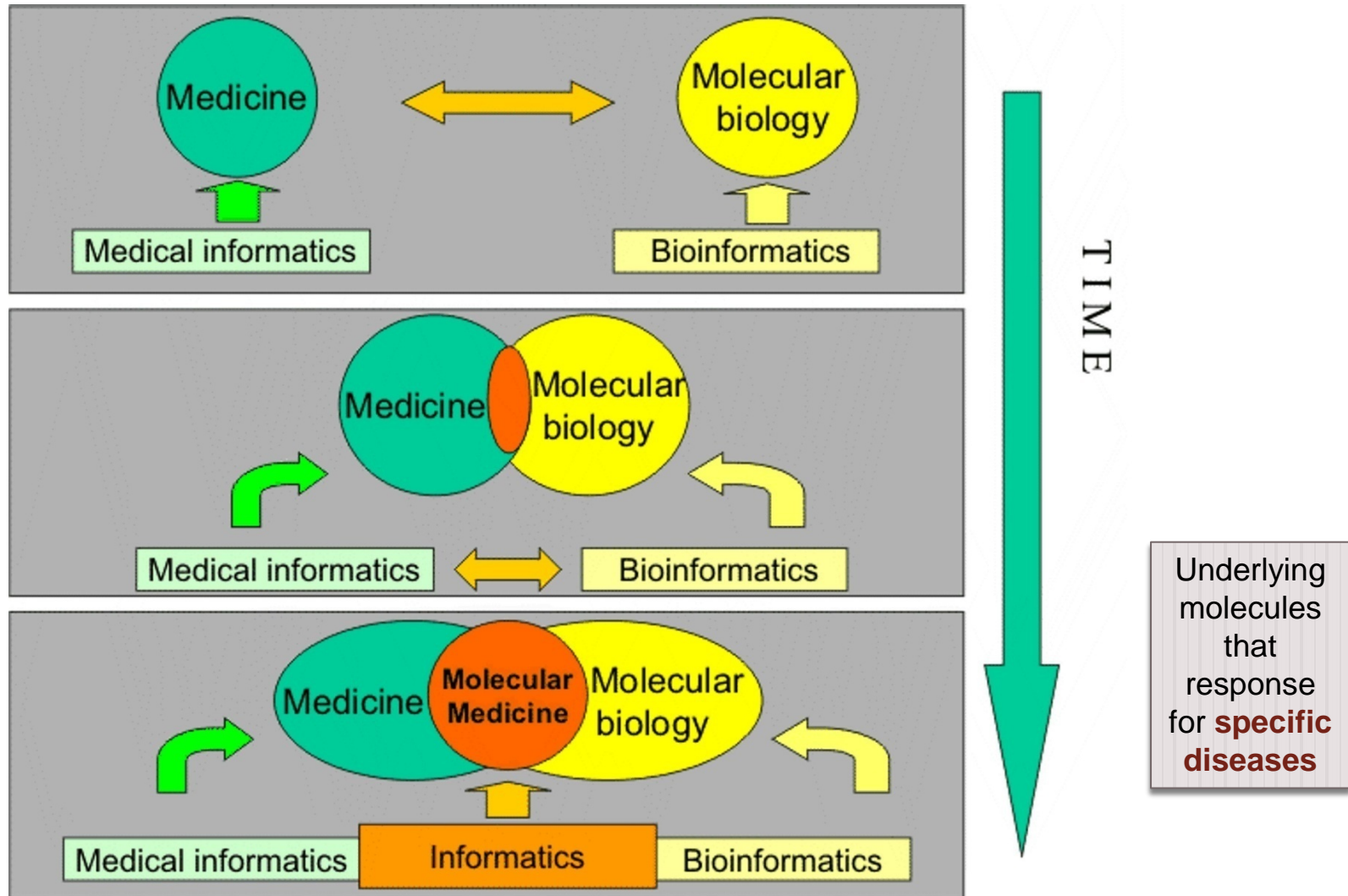
```
ACCATGGATTACATA000110110001101010  
GATTCATTATAAGGA01100111000000100  
TGCCGGCAATAGGCA001110101000110101  
CAATAAGCATTCCAC001010101101011011
```



# 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 imaging ( 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



0:04 / 2:03

480p



# The Holy Grail of Bioinformatics

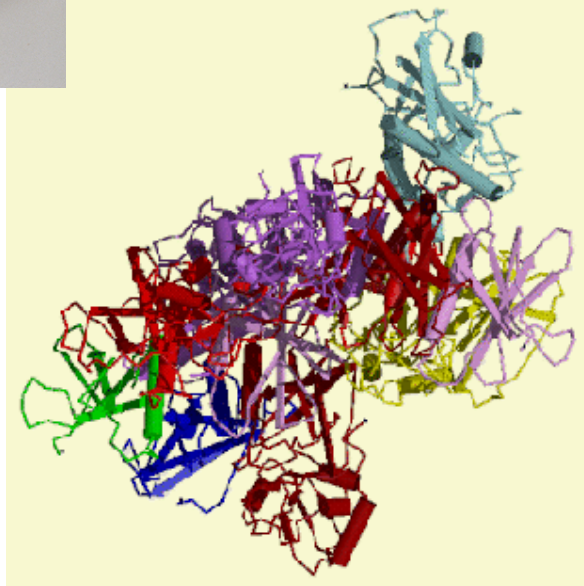
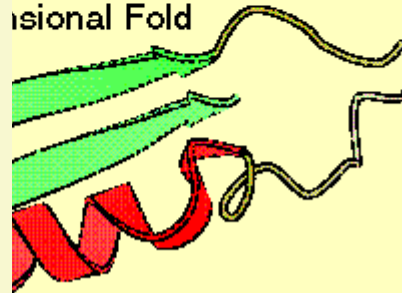


```
MNGTEGPNFYVPFSNKTGVVRS PF EAPQYYLAEPWQFSMLAAYMFL L I V L  
GFPINFLTLYVTVQHKKLRTP L N Y I L L N L A V A D L F M V F G G F T T T L Y T S L H  
G Y F V F G P T G C N L E G F F A T L G G E I A L W S L V V L A I E R Y V V V C K P M S N F R F G E  
N H A I M G V A F T W V M A L A C A A P P L V G W S R Y I P Q G M Q C S C G A L Y F T L K P E I N N
```

Amino Acid Sequence



3D Structural Fold



...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 ( $3 \cdot 10^5$  bp)

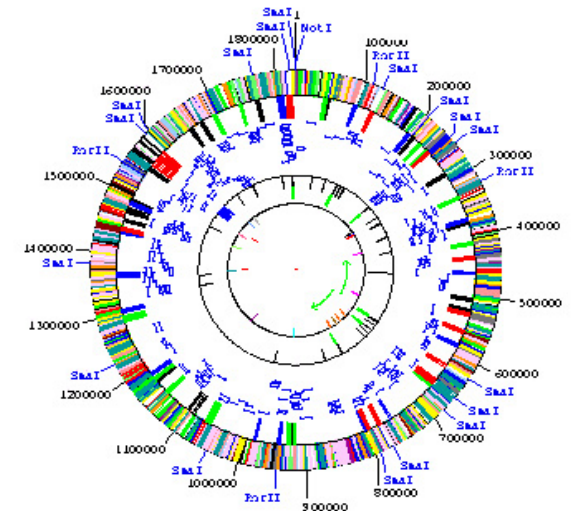
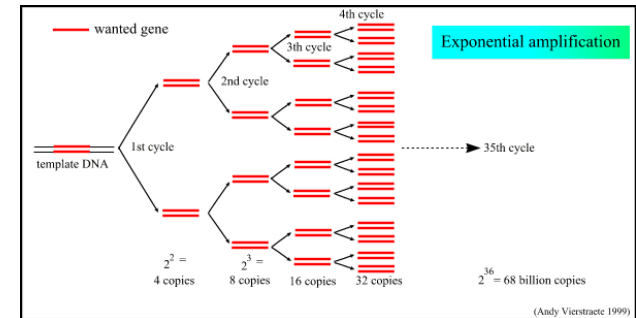
**1995:** Sequence of the genome of the bacteria: **Haemophilus influenzae** ( $2 \cdot 10^6$  bp)

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

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

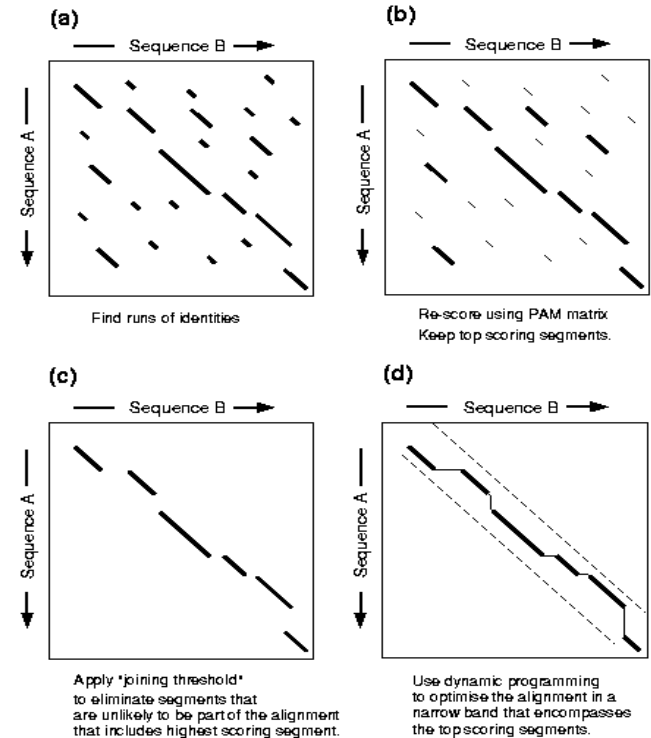
**1986:** Swiss-Prot (Protein Sequence Database)

**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)...

## **FASTA Algorithm**





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

**Table 2.** Twenty Publications that influenced our view of bioinformatics

Publication	Comments
Zuckerlandl 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 <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie <i>et al.</i> , 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...

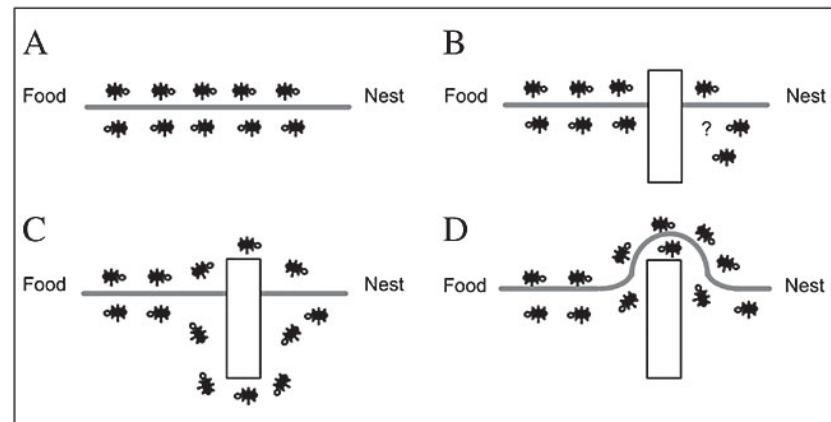
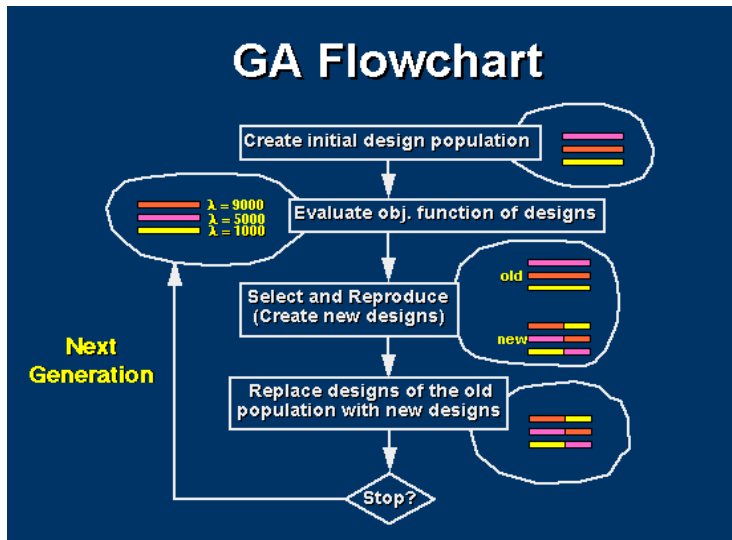


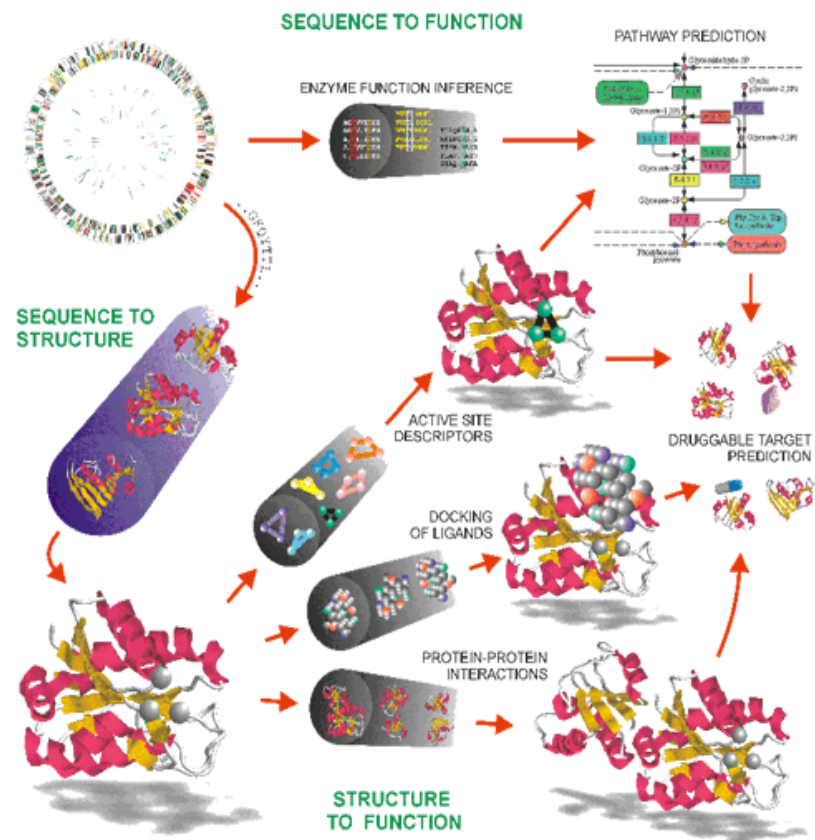
Figure 2. A. Ants in a pheromone trail between nest and food; B. an obstacle interrupts the trail; C. ants find two paths to go around the obstacle; D. a new pheromone trail is formed along the shorter path.

# 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



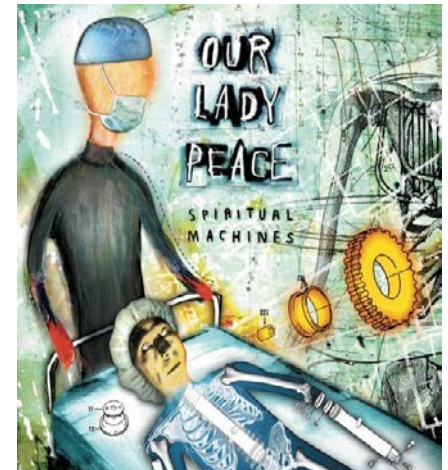
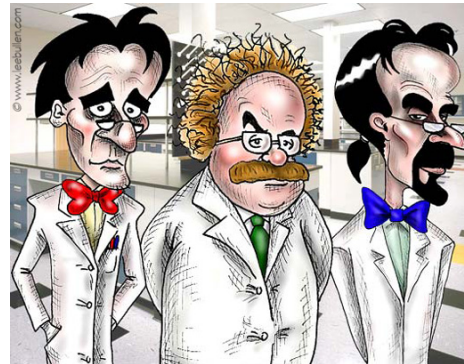
# Bioinformatics after 2010



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

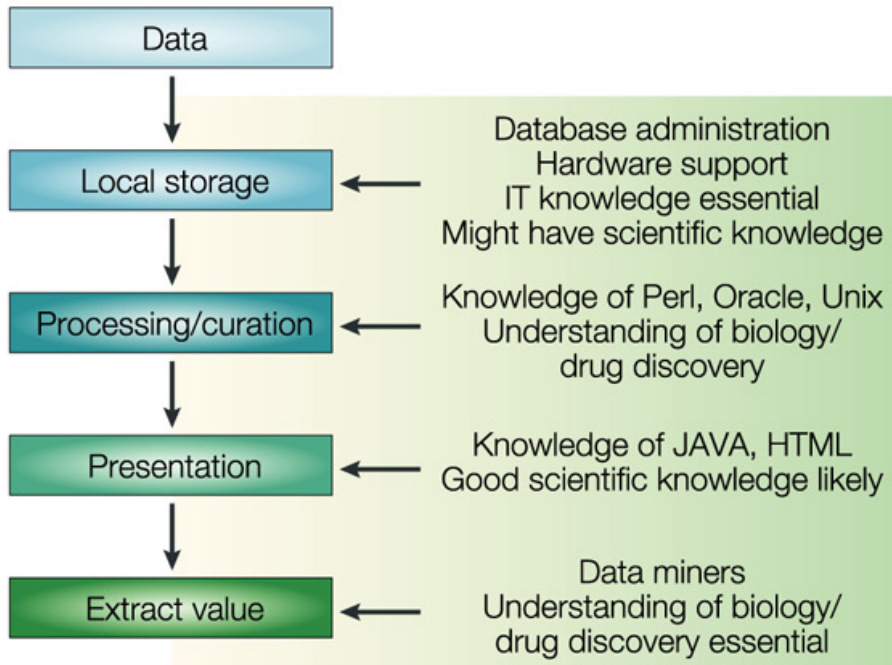
Life scientists use it **efficiently every day** and therefore **forget** that it **exists**

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



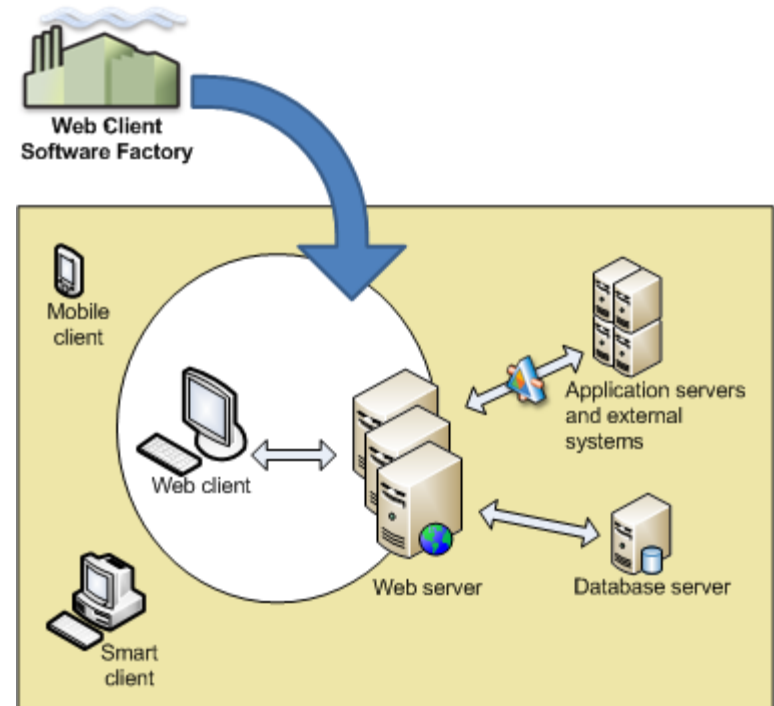


# Resources: databases & software



Nature Reviews | Drug Discovery








Nature Reviews Drug Discovery 3, 281 (2004)





# Breadth: Homologs, Large-scale Surveys, Informatics—

	pairwise comparison, sequence & structure alignment	multiple alignment, patterns, templates, trees	databases, scoring schemes, censuses
<b>1</b>	<b>2</b>	<b>3-100</b>	<b>100+</b>

	<b>Genome Sequence</b>	atc gatc gatatttgggatttgggga	atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga	atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga	atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga atc gatc gatatttgggatttgggga
gene finding	↓				
	<b>Protein Sequence</b>	ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT
structure prediction	↓				
	<b>Protein Structure</b>				
geometry calculation	↓				
	<b>Protein Surface</b>				
molecular simulation	↓				
	<b>Force Field</b>				
structure docking	↓				
	<b>Ligand Complex</b>				

Depth: Rational Drug Design (physics) →





*"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.

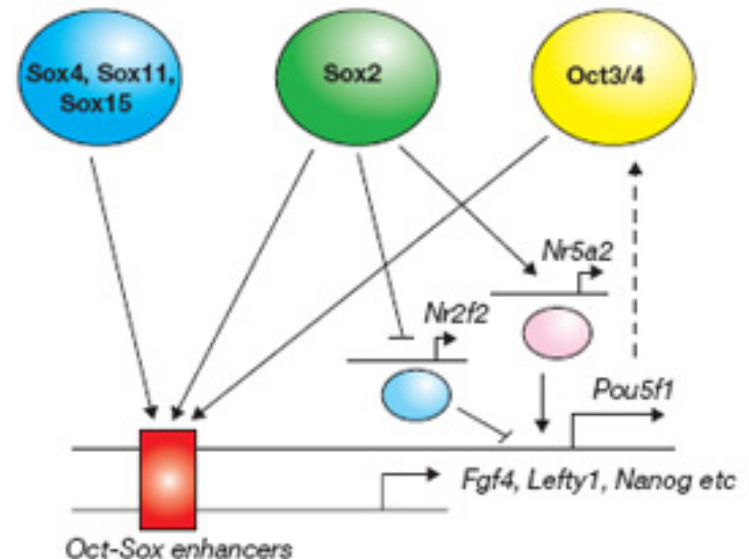
PMID: 12787504 [PubMed - indexed for MEDLINE]

# 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**



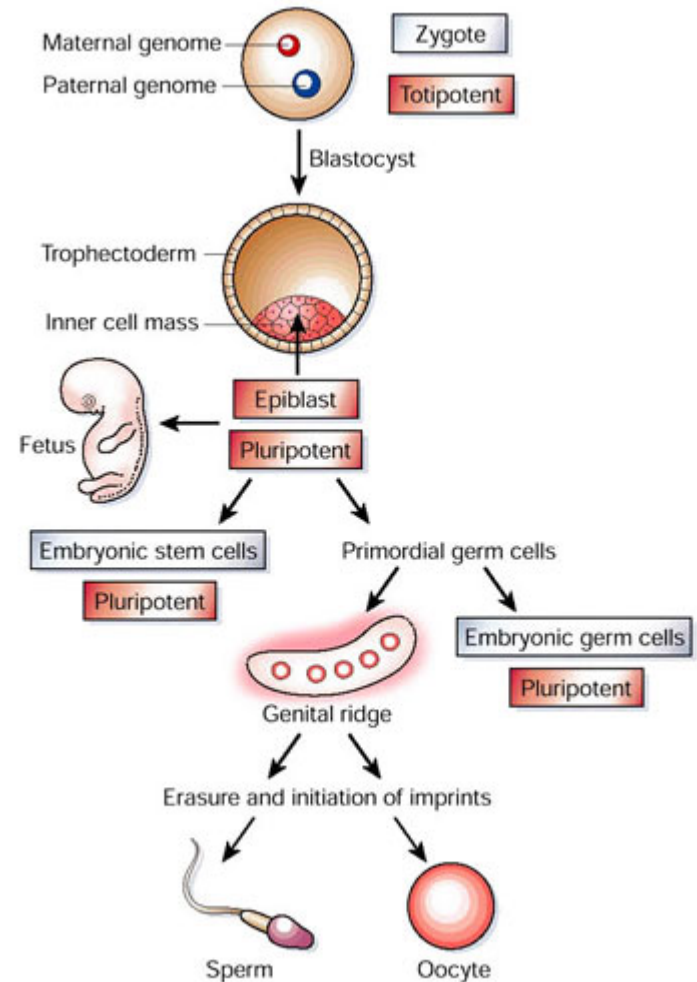


# Results

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**

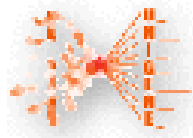
A number of genes were found **overrepresented in ES cell-derived libraries** (Table, next slide)



# 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.

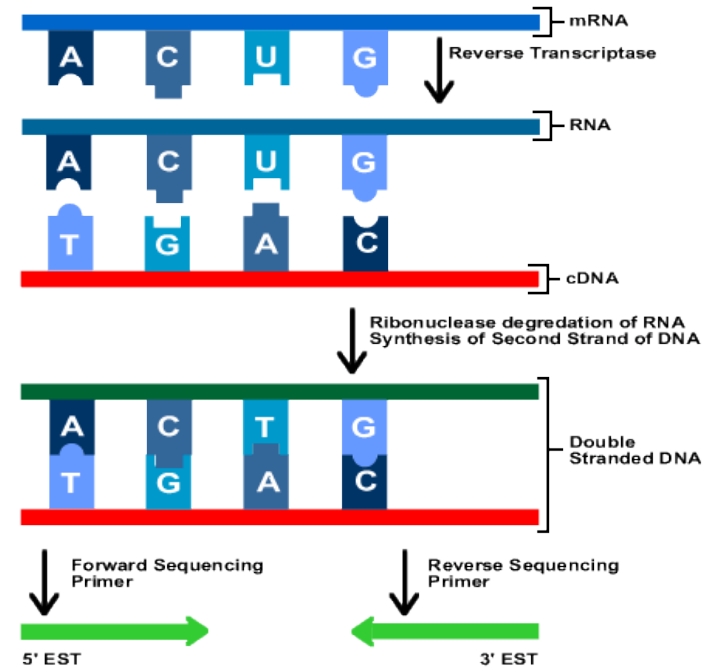
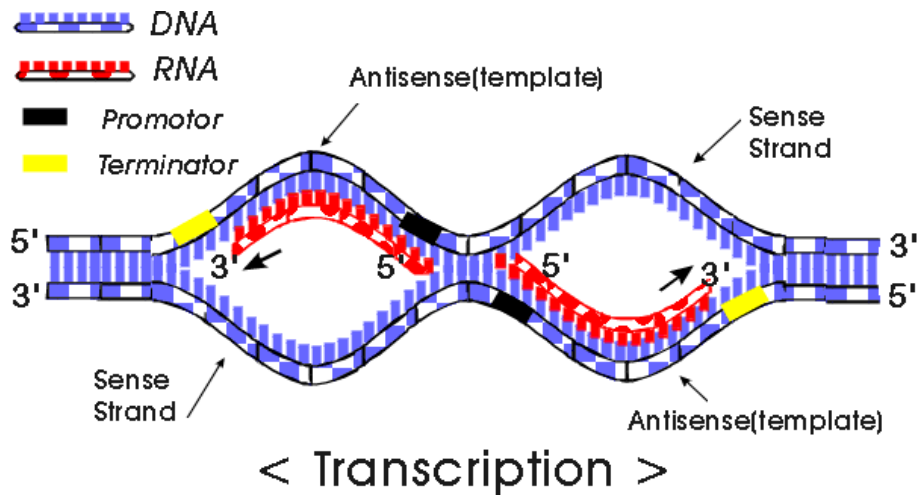


**UniGene DDD** - Digital Differential Display - an online tool to compare computed gene expression profiles between selected cDNA libraries. Using a statistical test, genes whose expression levels differ significantly from one tissue to the next are identified and shown to the user. Additional information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

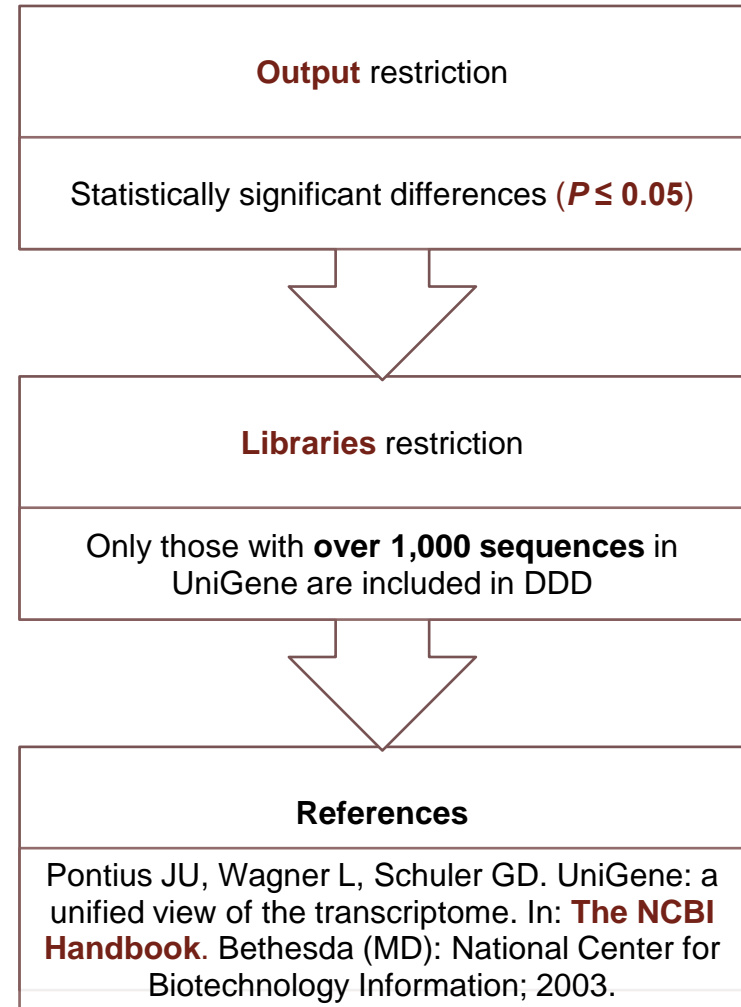
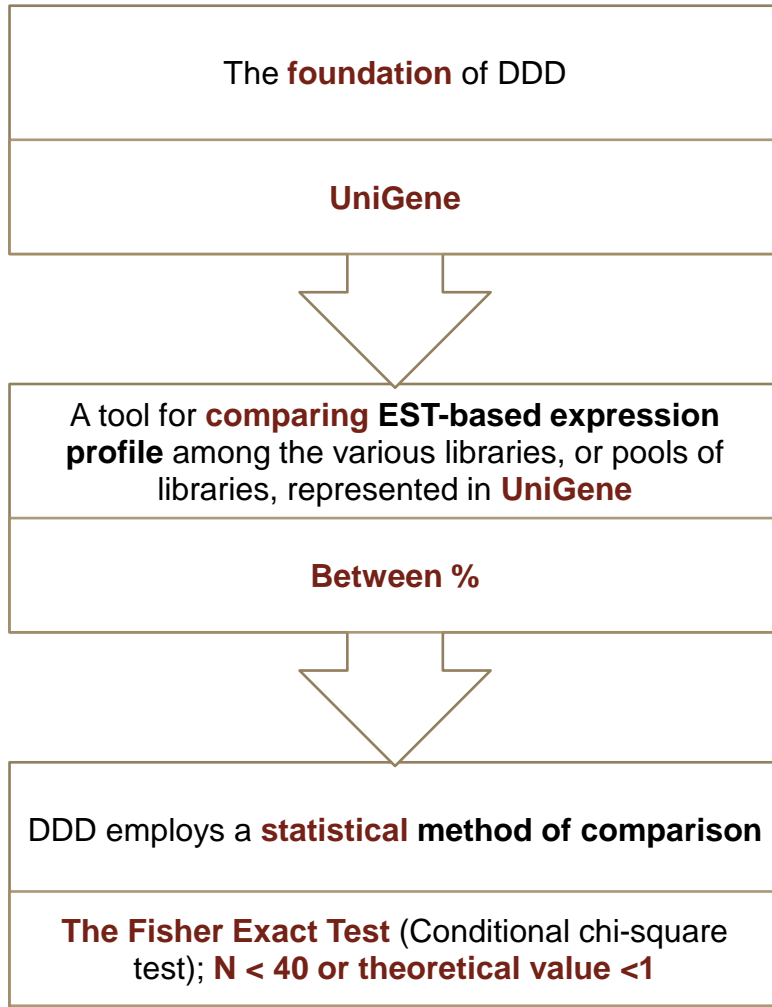
<http://www.ncbi.nlm.nih.gov/Tools/>

s/

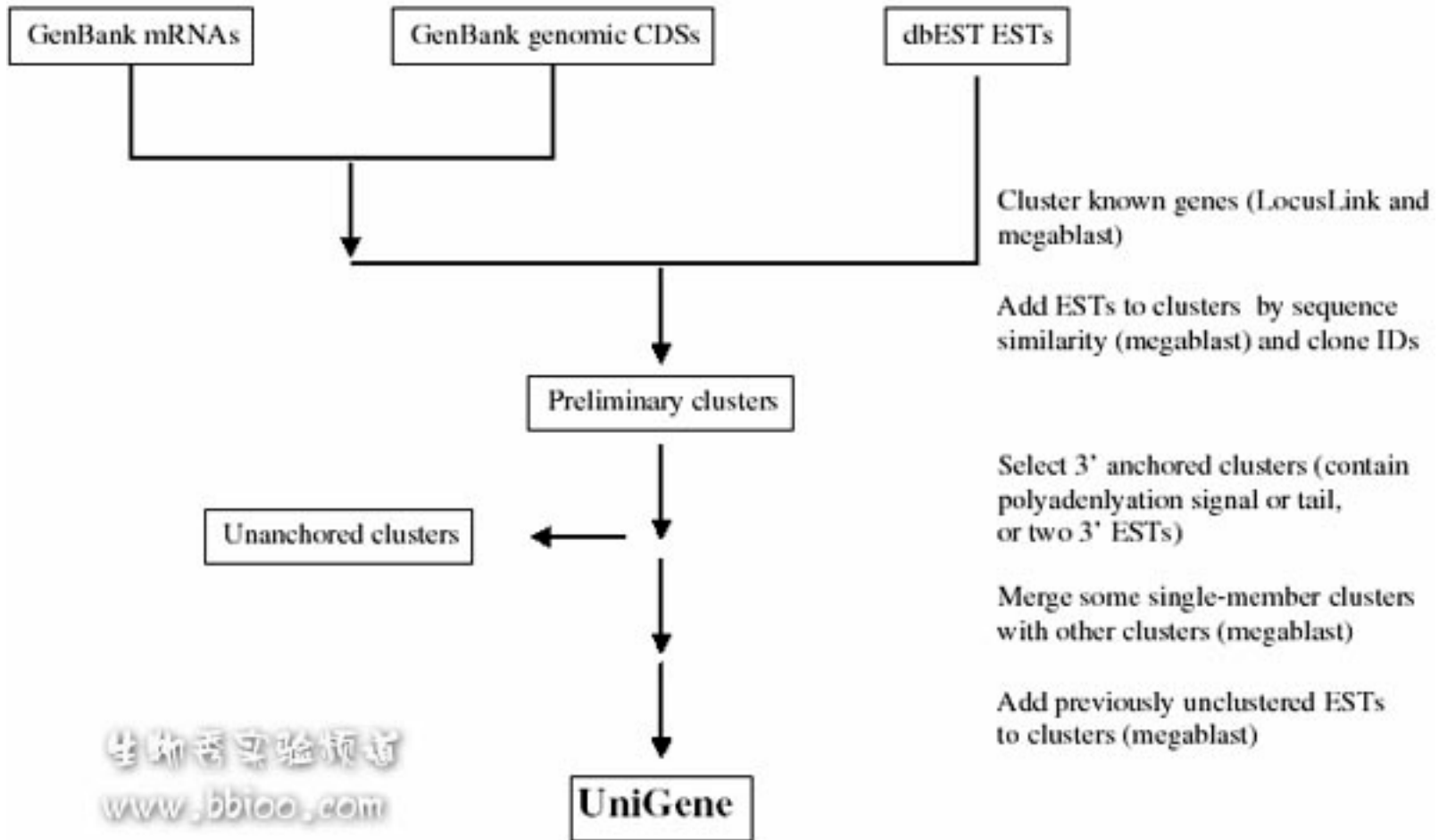
# DDD Basics (1)



# DDD Basics (2)



# 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

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

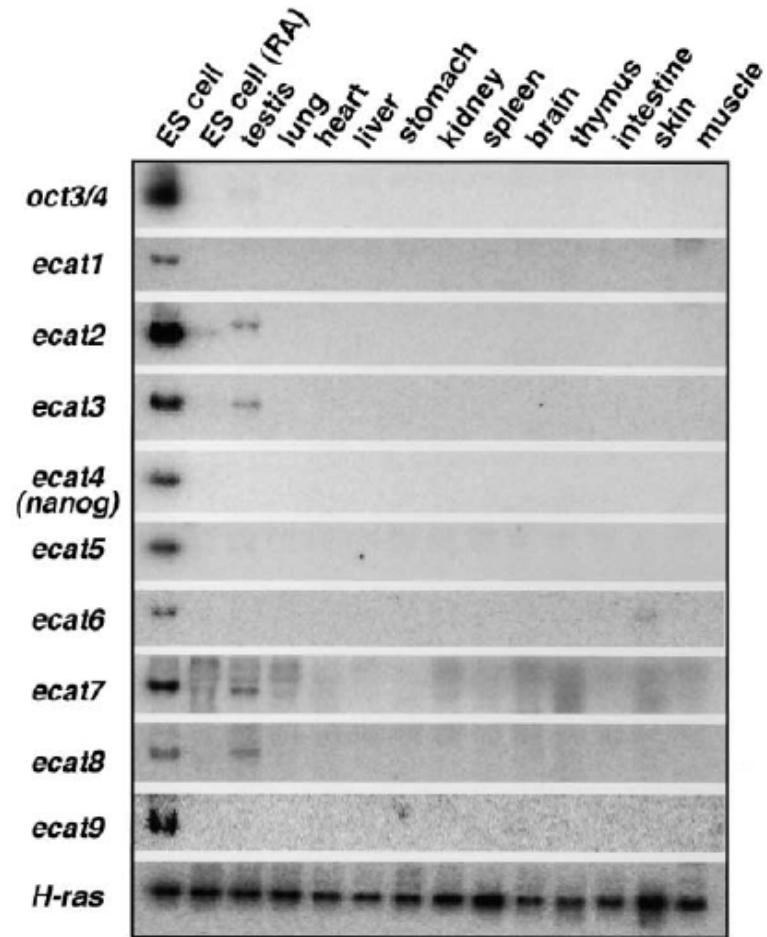
### The Fisher Exact Test

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	<a href="#">Dppa5</a>
Mm.10205	0.0028	0	utf1	<a href="#">Utf1</a>
Mm.5090	0.00079	0	cripto	<a href="#">Tdgf1</a>
Mm.157658	0.00076	0	ecat1	<a href="#">2410004A20Rik</a>
Mm.128134	0.00076	0	hnRNH-G	<a href="#">1700012H05Rik</a>
Mm.5180	0.00073	0	Nr0b1/dax-1	<a href="#">Nr0b1</a>
Mm.6047	0.00067	0	ecat4/nanog	<a href="#">Nanog</a>
Mm.28369	0.00049	0	ecat3/fbx15	<a href="#">Fbxo15</a>
Mm.17031	0.0004	0	oct3/4	<a href="#">Pou5f1</a>
Mm.285848	0.00033	0	Zfp42/rex1	<a href="#">Zfp42</a>
Mm.45676	0.00033	0	EST	<a href="#">LOC433110</a>
Mm.258773	0.00033	0	ecat5/ERas	<a href="#">Zmynd11</a>
Mm.23310	0.0003	0	zfp296	<a href="#">Zfp296</a>
Mm.18154	0.00027	0	tcl1	<a href="#">Tcl1</a>
Mm.13433	0.00027	0	ecat7	<a href="#">Dnmt3l</a>
Mm.47904	0.00027	0	ecat8	<a href="#">2410004F06Rik</a>
Mm.299742	0.00024	0	ecat9	<a href="#">Gdf3</a>
Mm.913	0.00024	0	brachyury (T )	<a href="#">Irf1</a>
Mm.256916	0.00024	0	tex20	<a href="#">Sall4</a>
Mm.158190	0.00021	0	ecat6	<a href="#">2410039E07Rik</a>



*ecat*, for ES cell associated transcripts

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

[New...](#)

### Statistically Significant Differences

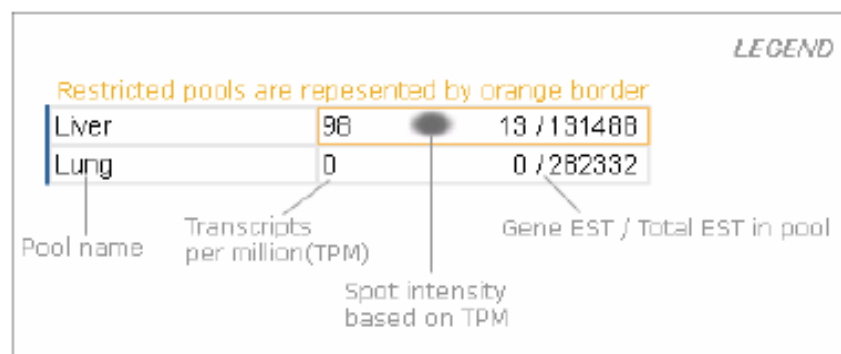
A.	B.	Gene index	Gene description
990	<a href="#">Mm.99395</a>		Seminal vesicle protein, secretion 2 (Svs2)
0.00000	0.00443		
A<B	B>A		
55	170	<a href="#">Mm.336743</a>	Heat shock protein 8 (Hspa8)
0.00359	0.00076		
A>B	B<A		
41	<a href="#">Mm.10205</a>		Undifferentiated embryonic cell transcription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

## Breakdown by Tissue

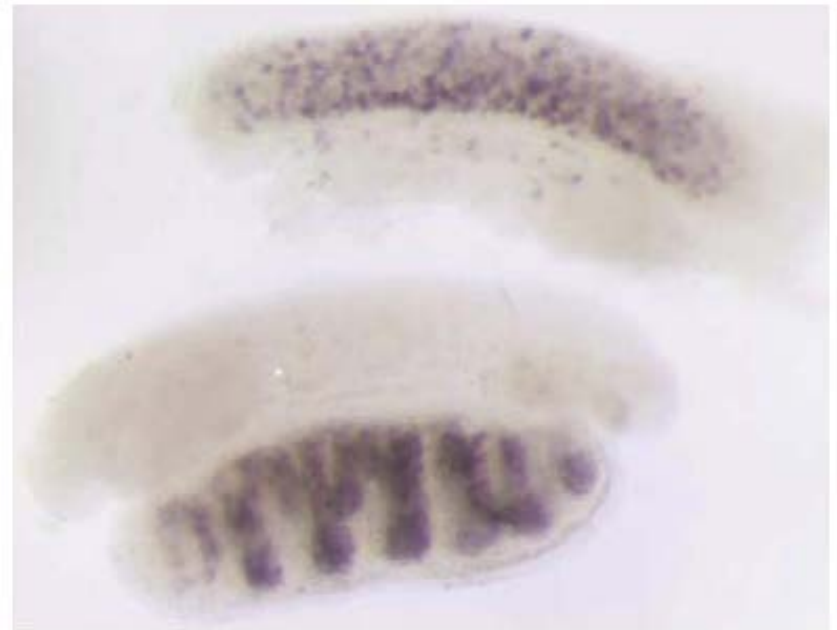
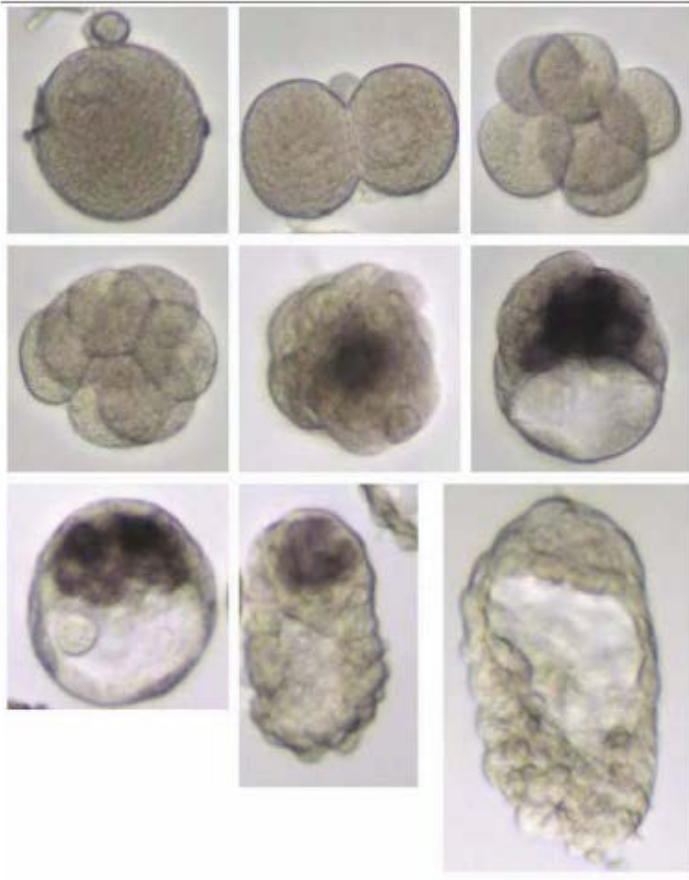
bone	0		0/38928
bone marrow	0		0/37245
brain	0		0/486112
colon	0		0/52004
eye	0		0/169091
heart	0		0/53082
kidney	0		0/116533
liver	0		0/104225
lung	0		0/43564
lymph node	0		0/25515
mammary gland	0		0/348007
muscle	0		0/19349
ovary	0		0/14904
pancreas	0		0/80182
placenta	0		0/32674
pituitary g...	0		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**

---

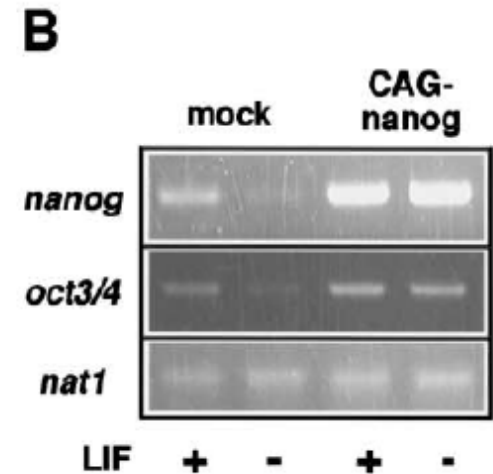
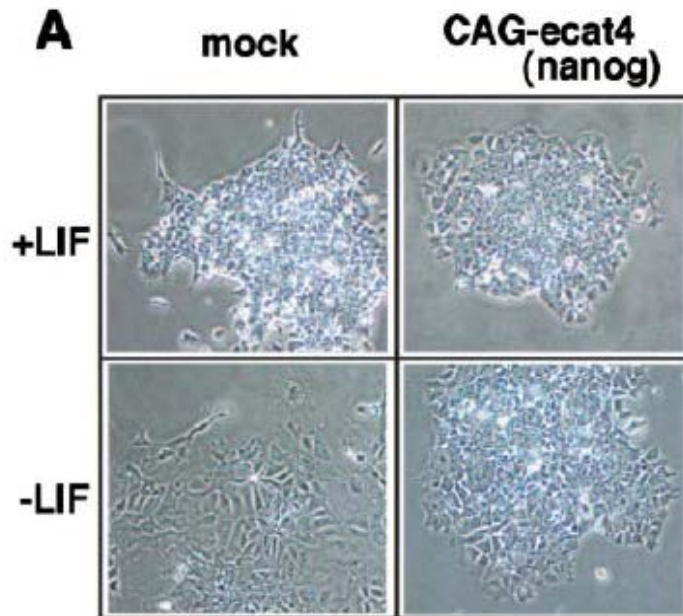
When cultured **without LIF**, all but one **differentiated** normally as judged by **flattened morphology & reduced *oct3/4* expression**

---

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  
© 2005 by The American Society for Biochemistry and Molecular Biology, Inc.

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\*<sup>§</sup>

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

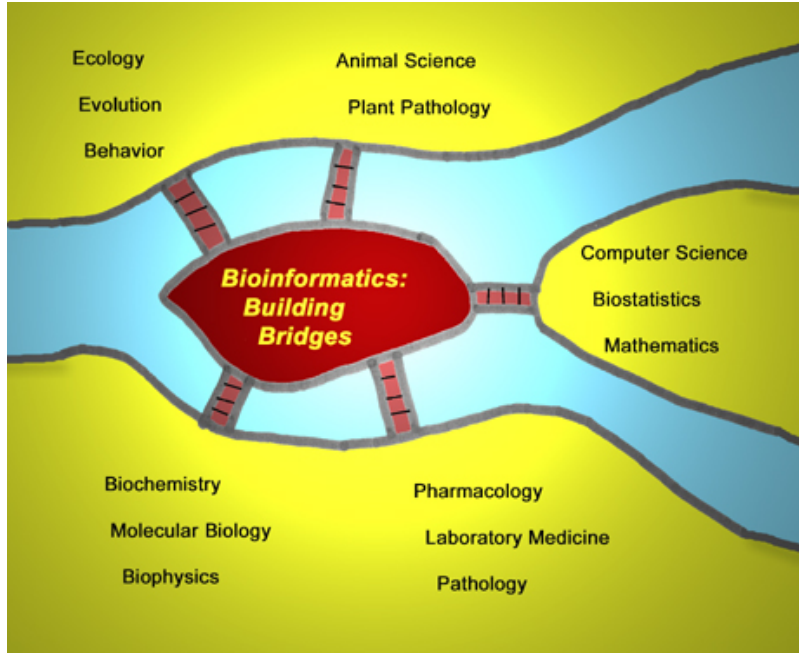
Masayoshi Maruyama, Tomoko Ichisaka, Masato Nakagawa, and Shinya Yamanaka<sup>‡</sup>

*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*

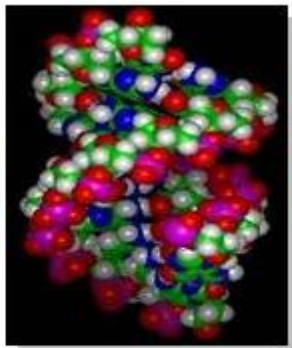
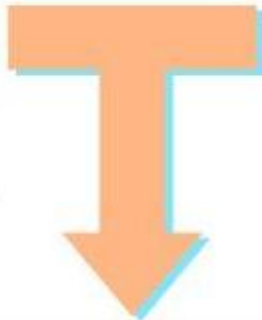




# Q & A



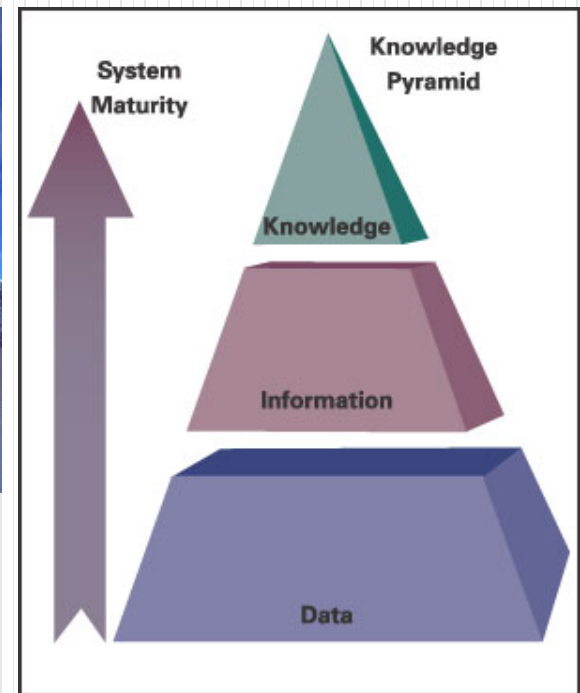
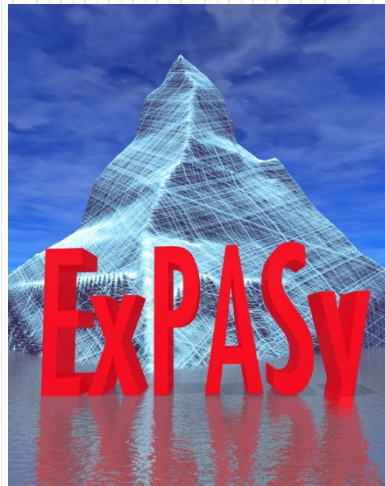
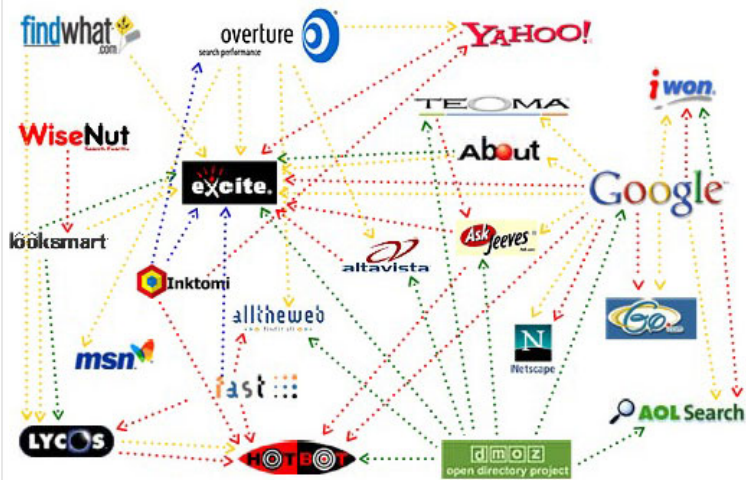
Computer systems



Biological systems

**BIOINFORMATICS**

# Q: How to Find the Right Stuffs?





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

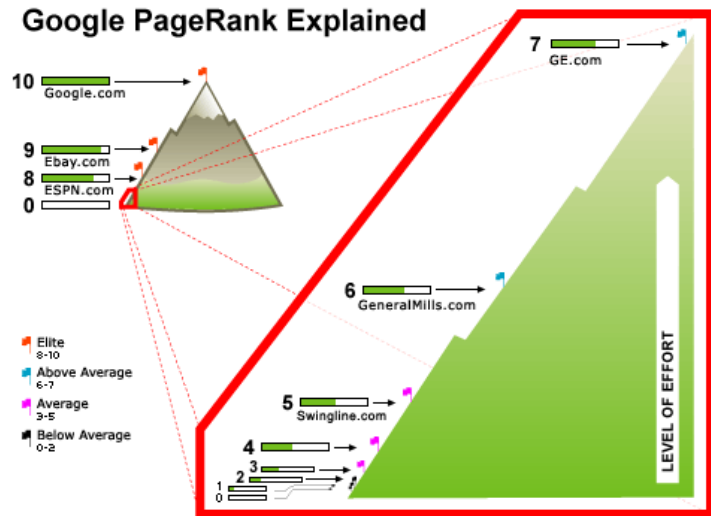
Subject-specific popularity

Use **the right key words** PubMed: **MeSH**

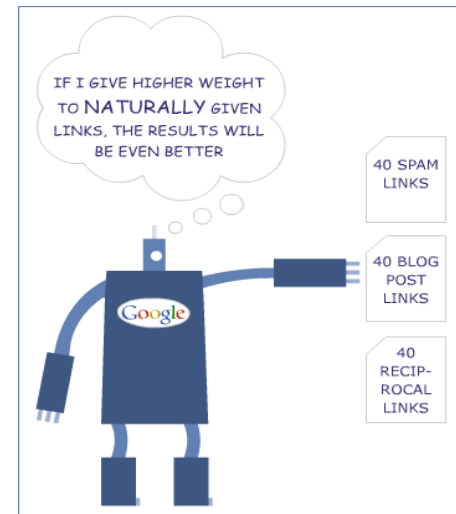
**OMIM**: index

Gene name: **HUGO**

Fidelity: edu > gov > org > com



©2007 Elliance, Inc.



# Search Efficiently

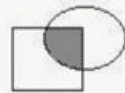
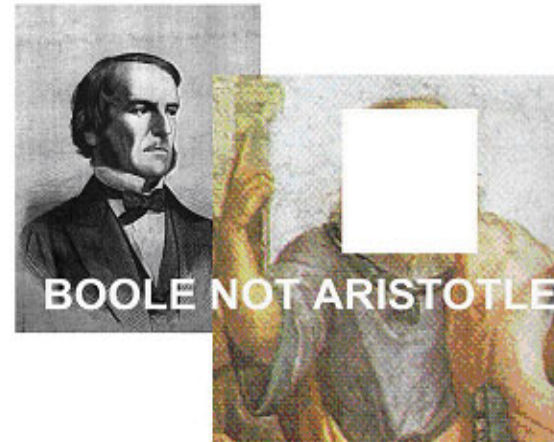
[Quick  
Tours](#)

[Search PubMed by Authors](#)

[My NCBI...](#)

[Boolean  
operators](#)

**AND OR NOT**

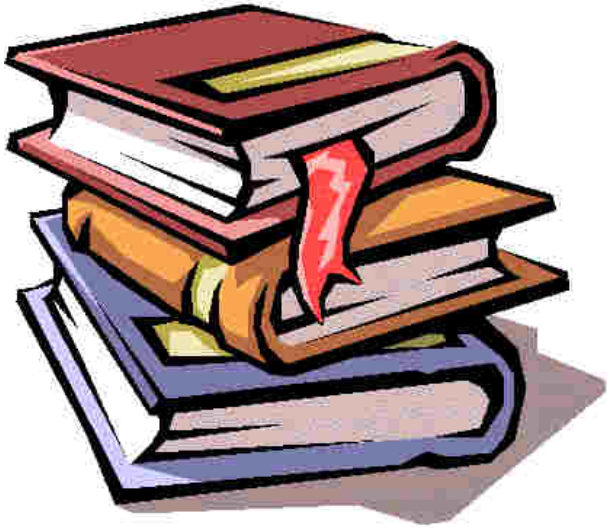


This is a small search.  
Your results will  
include *both* words.





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





# Gene or Disease – Official Symbol

1----- (100000- )	Autosomal loci or phenotypes (entries created before May 15, 1994)
2----- (200000- )	
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)



**PubMed**  
 • [POU5F1](#)

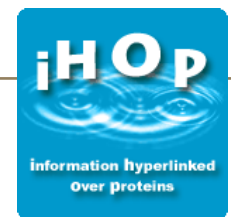
**OMIM**  
 • Preview and index

**GeneCards/**  
 human

• [POU5F1](#) (symbol only)

**Entrez Gene**  
 • [POU5F1](#)

**iHOP**  
 • [POU5F1](#)



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

# iHOP: Model building

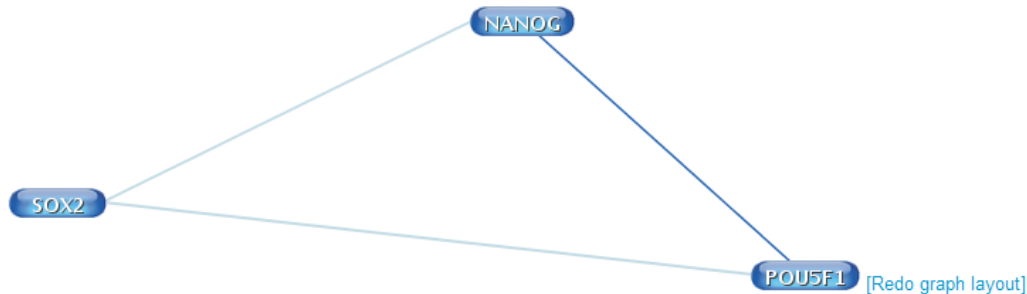
WikiGenes [edit this page](#) **new**  
UniProt [Q01860, A6NLL8, Q15167](#)  
IntAct [Q01860](#)  
OMIM [164177](#)  
NCBI Gene [5460](#)  
NCBI RefSeq [NP\\_976034, NP\\_002692](#)  
NCBI RefSeq [NM\\_002701, NM\\_203289](#)  
NCBI UniGene [5460](#)  
NCBI Accession [BAB63311, CAA79974](#)






[Homologues of POU5F1 ...](#)

[Definitions for POU5F1 !\[\]\(99f58673407353e96a019fbca558fd72\_img.jpg\) ...](#)

[Most recent information for POU5F1 !\[\]\(0f848bbd71cef6b345273b16f905912a\_img.jpg\) ...](#)

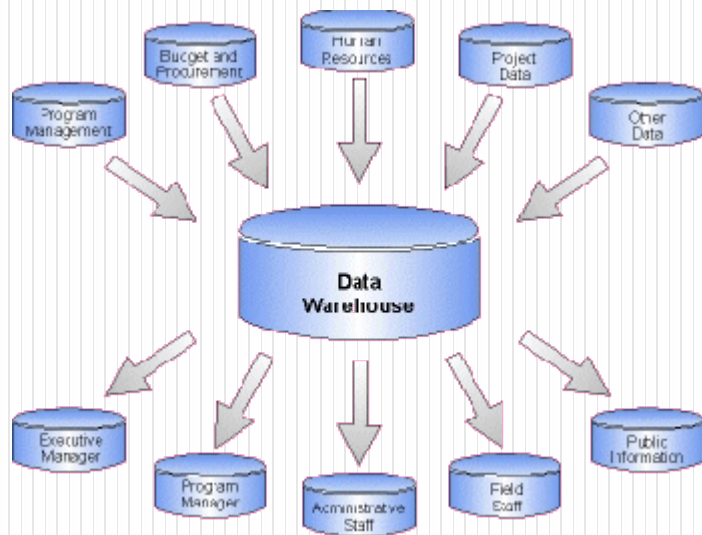
[Enhanced PubMed/Google query ...](#)



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	<b>A Guide to NCBI: Gene Expression, Part 1</b> 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	<b>A Guide to NCBI: Gene Expression, Part 2</b> 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	<b>A Guide to NCBI: Gene Expression, Part 3</b> 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	<b>Align 2 Sequences</b> Aligning two groups of sequences and displaying the results in the NCBI sequence viewer
Video	<b>Assign Downloaders for dbGaP Data</b> 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**



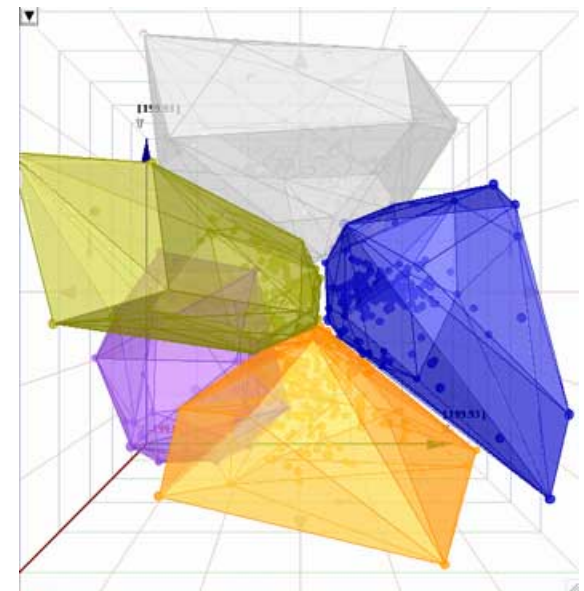
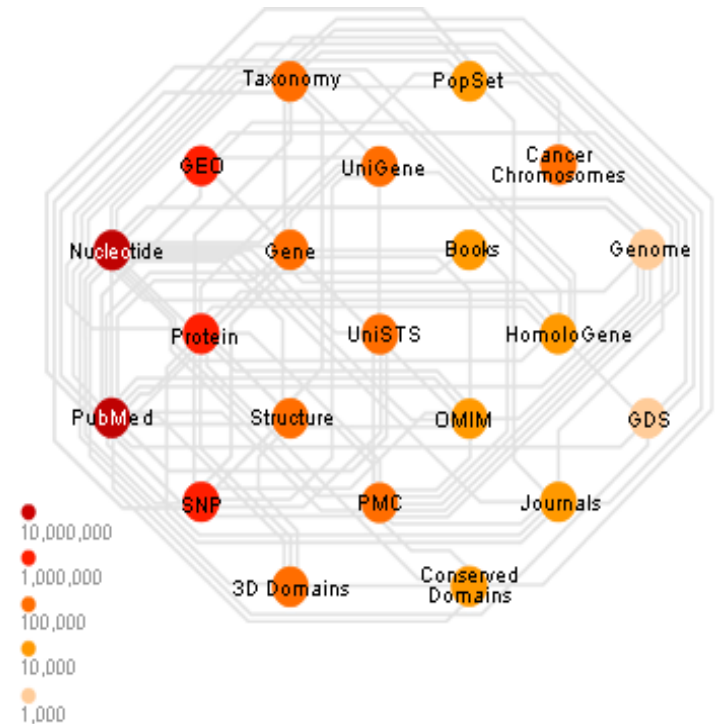
# Contents

## 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





**Institution:** NATIONAL SUN YAT-SEN UNIVERSITY [Sign In as Personal Subscriber](#)

[Oxford Journals](#) > [Science & Mathematics](#) > [Nucleic Acids Research](#) > [Volume 44, Issue D1](#) > Pp. D1-D6.



**BIOLOGY**  
Methods & Protocols

A NEW OPEN ACCESS JOURNAL  
Now accepting manuscripts, click here to submit

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

Daniel J. Rigden<sup>1,\*</sup>, Xosé M. Fernández-Suárez<sup>2</sup> and Michael Y. Galperin<sup>3,\*</sup>

+ Author Affiliations

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



[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, Xosé M. Fernández-Suárez, 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](mailto: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.

- Category List
- Summary Paper List
- Complete Category/Summary Paper List
- Search Summary Papers



### 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

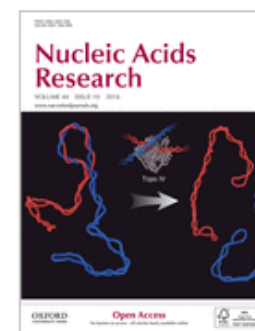
Search this journal:

GO

[Advanced »](#)

### Current Issue

02 June 2016 44 (10)



[Alert me to new issues](#)



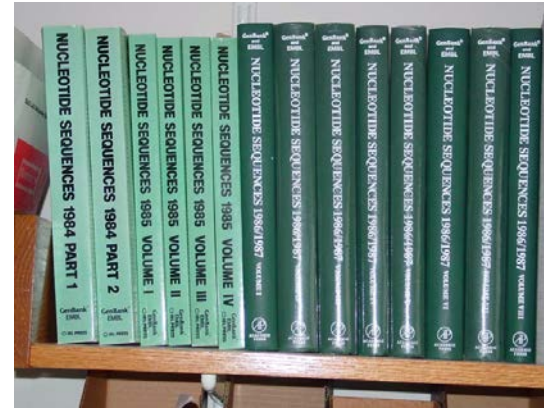
# Primary vs. Derivative Databases - NCBI

## Primary databases

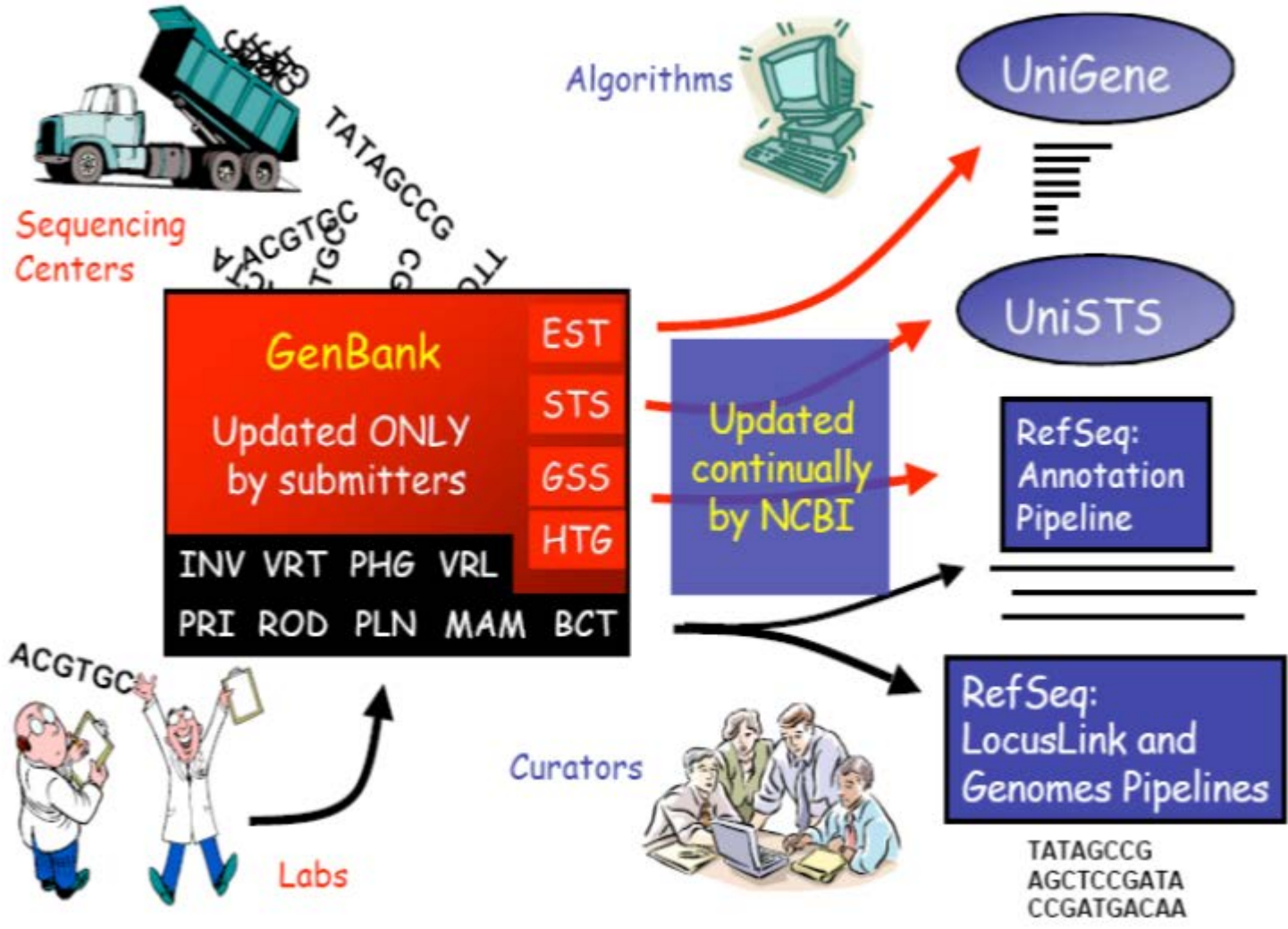
- **Original** submissions by **experimentalists**
- **Submitters** retain editorial control of records
- Archival in nature

## Derivative databases

- **Curated** by NCBI staffs
- **NCBI** retains **editorial control** of records
- Record content is **updated continually**

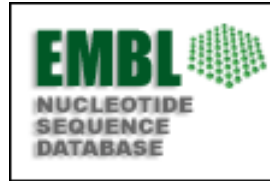


# Primary vs. Derivative Databases - NCBI





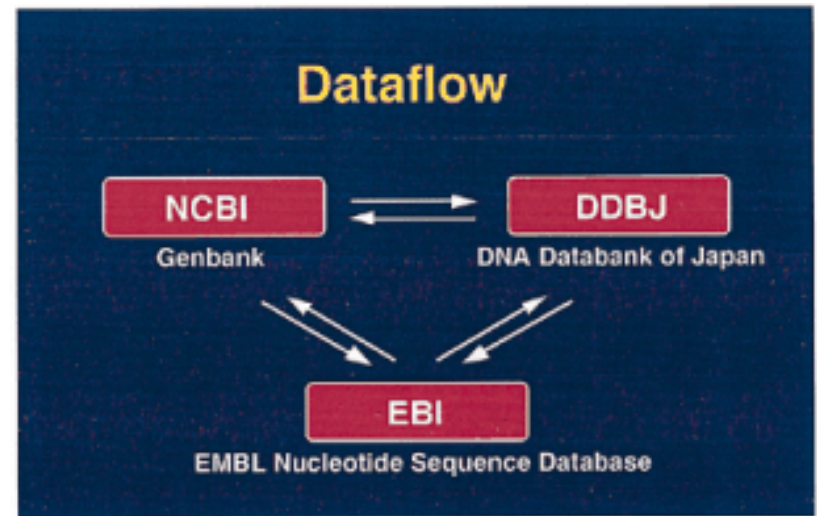
# Primary DNA Databases



GenBank (USA)

EMBL (Europe)

DDBJ (Japan)



National Institute of Health (**NIH**)

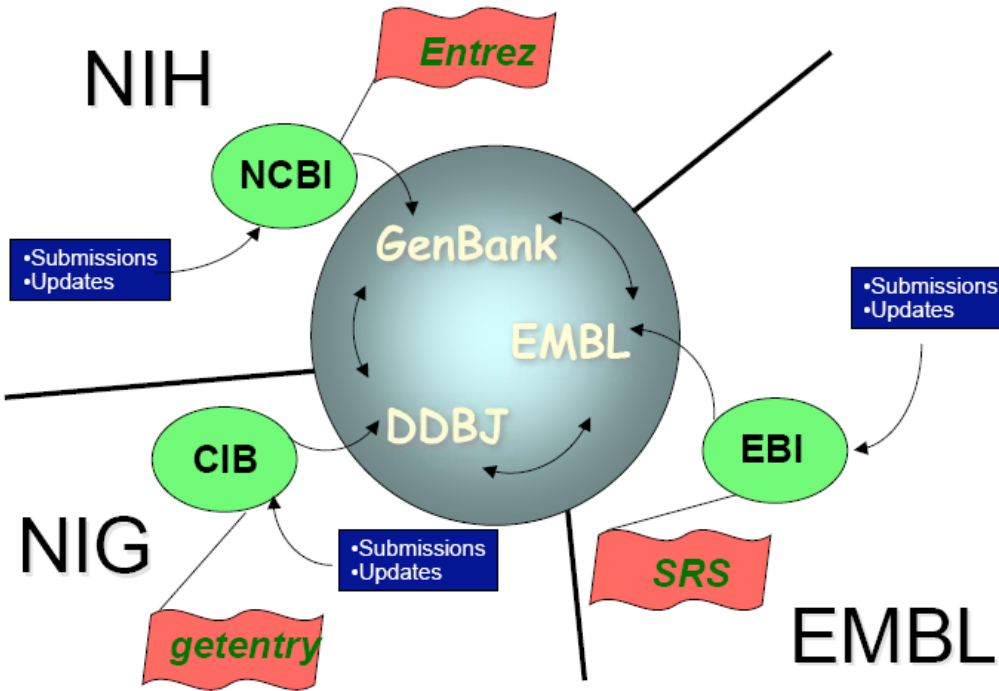
National Center for Biotechnology (**NCBI**)

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



National Institute of Genetics  
(NIG)

Center for Information Biology  
(CIB)



The European Bioinformatics  
Institute (**EBI**)

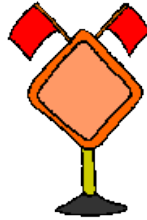
Sequence Retrieval System  
(**SRS**)

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



# EMBL/GenBank/DDBJ Annotations

Warning!!!



---

DNA  
data  
base  
annot  
ations  
are  
**full of  
errors**

In sequences, in annotations, in  
CDs attribution...

---

**No consistency** of annotations

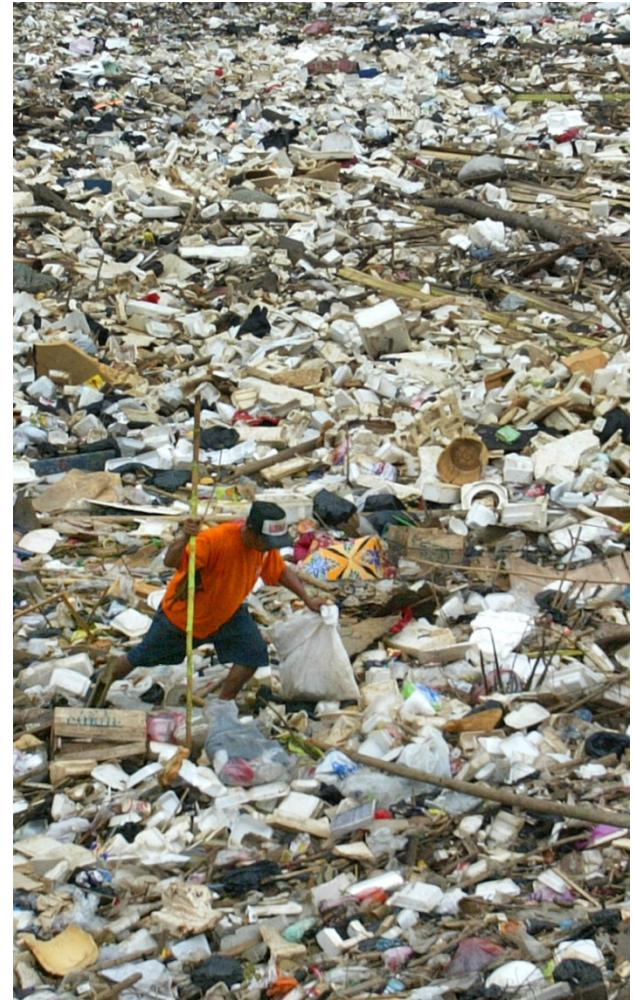
---

Most annotations are done by the  
**submitters**

---

**Heterogeneity** of quality and  
updating

---



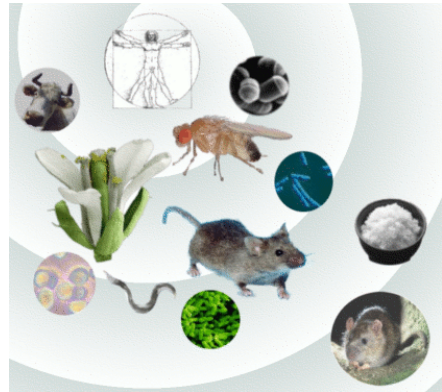
# Some Interesting Sequence Annotation

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"



Taxonomy  
Browser  
@ EBI vs. NCBI  
Taxonomy



Search in Taxonomy Query

[Advanced Search »](#)

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

 [UniProtKB \(151\)](#) |  [Taxonomy help](#)

<b>Mnemonic</b>	DIDMA
<b>Taxon identifier</b>	9267
<b>Scientific name</b>	<i>Didelphis marsupialis virginiana</i>
<b>Common name</b>	North American opossum
<b>Synonym</b>	-
<b>Other names</b>	<ul style="list-style-type: none"> <li>› <i>Didelphis virginiana</i></li> <li>› Virginia opossum</li> </ul>
<b>Rank</b>	SPECIES
<b>Lineage</b>	<ul style="list-style-type: none"> <li>› cellular organisms</li> <li>› <a href="#">Eukaryota</a></li> <li>› Fungi/Metazoa group</li> <li>› <a href="#">Metazoa</a></li> <li>› Eumetazoa</li> <li>› Bilateria</li> <li>› Coelomata</li> <li>› Deuterostomia</li> <li>› <a href="#">Chordata</a></li> <li>› <a href="#">Craniata</a></li> <li>› <a href="#">Vertebrata</a></li> <li>› Gnathostomata</li> <li>› Teleostomi</li> <li>› <a href="#">Euteleostomi</a></li> </ul>

**Taxonomy navigation**

↑ [› Didelphis](#)

↓ Terminal (leaf) node.

Images may be subject to copyright.



[calphotos.berkeley.edu](http://calphotos.berkeley.edu)



[calphotos.berkeley.edu](http://calphotos.berkeley.edu)



[calphotos.berkeley.edu](http://calphotos.berkeley.edu)

# Organization of GenBank: Traditional Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

## Traditional Divisions:

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

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

Entrez query: `gbdiv_xxx[Properties]`

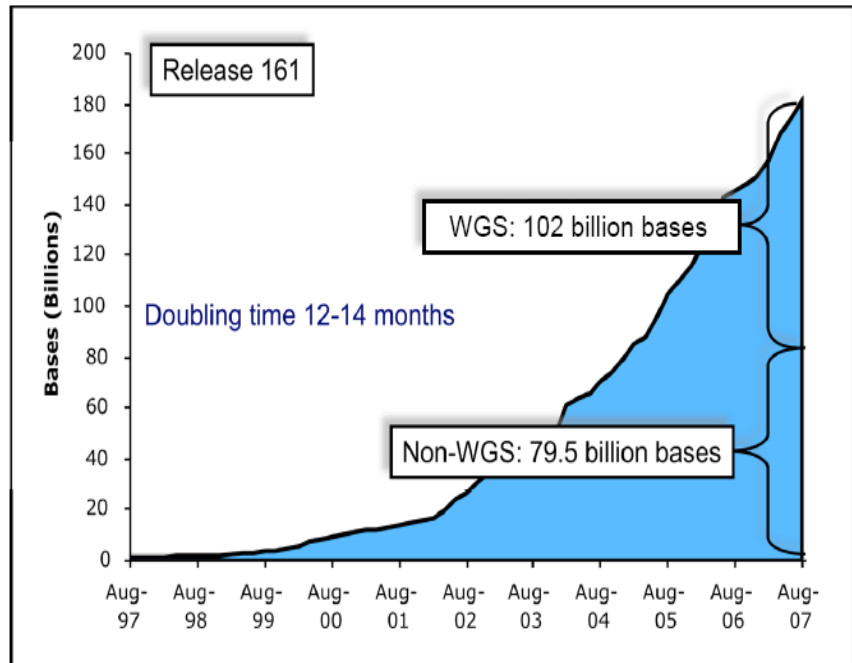
# Bulk GenBank Divisions

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

## **BULK Divisions:**

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

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

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

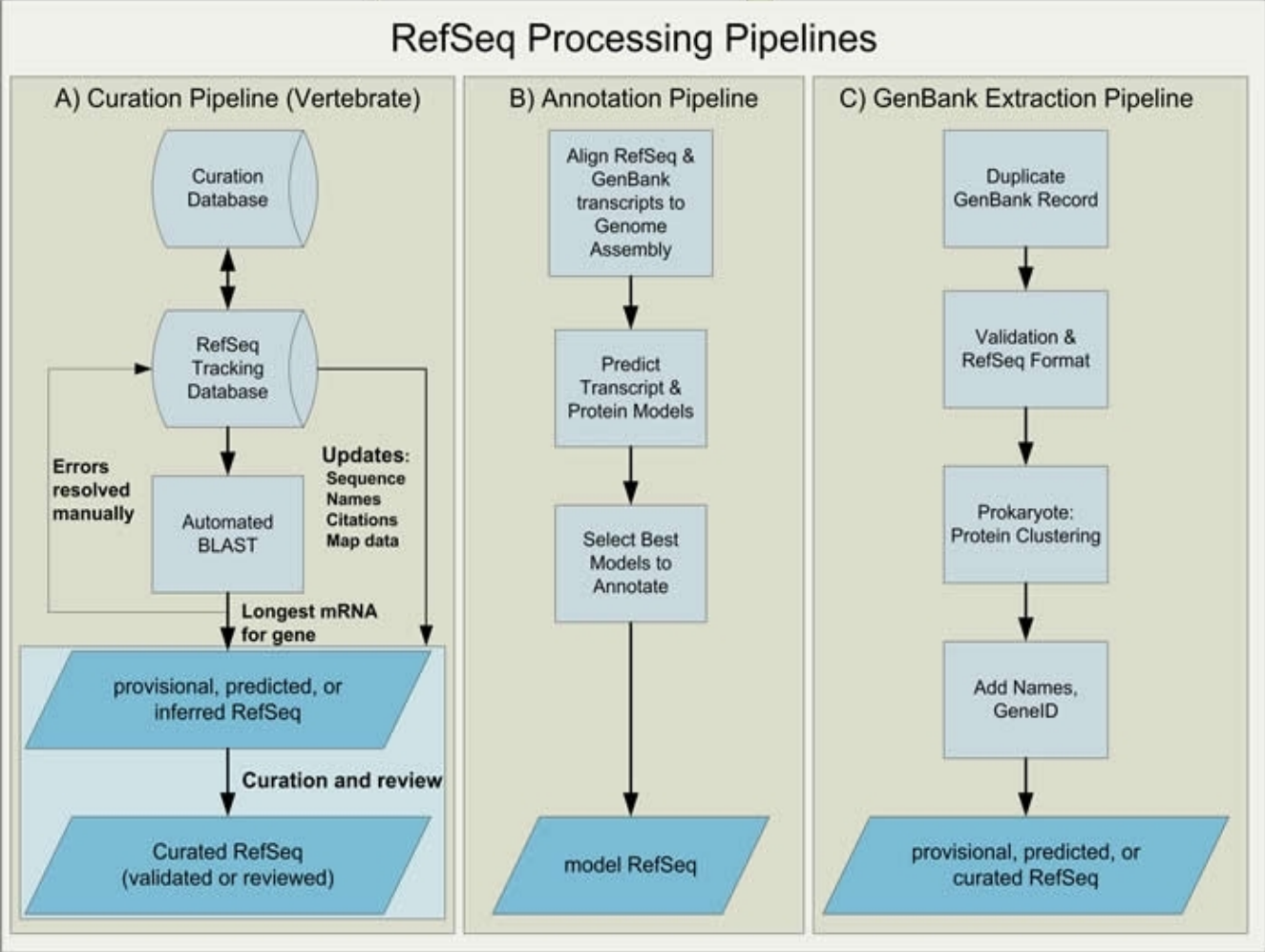
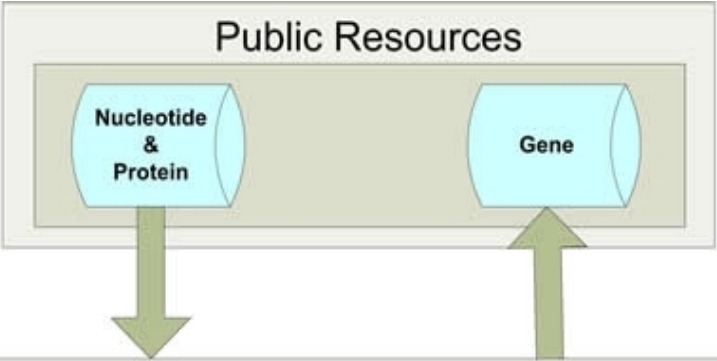
## Assemblies

NT\_123456

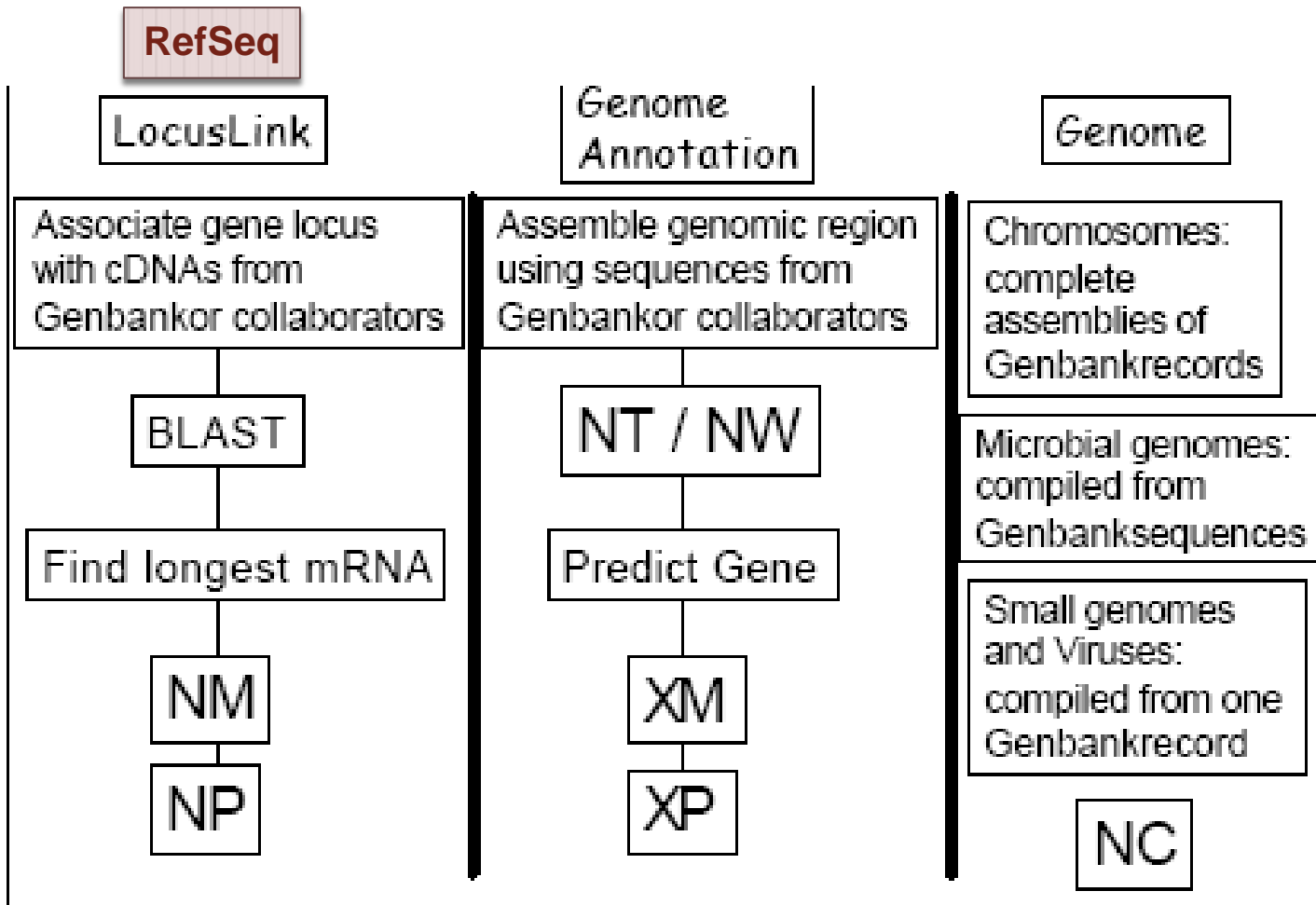
Contig

NW\_123456

WGS Supercontig



# 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      – chicken
  - mouse genome      – honeybee
  - rat genome          – 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**



interact with consistency

# Entrez Protein: Derivative Databases

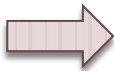
## Example: CKS1B

CDS

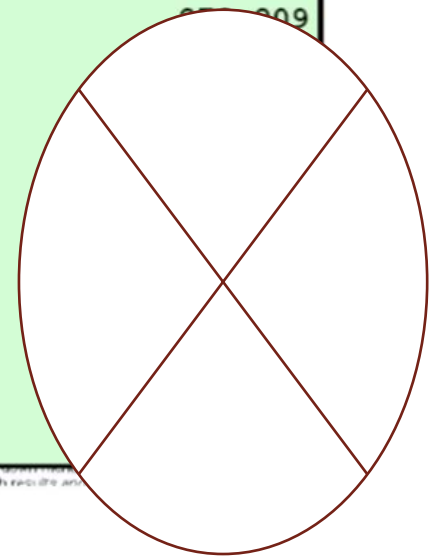
```

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="MSHKQIYYSDKYDDEEFYRHMVLPKDI AKLVPKTHLMSESEWR
NLGVQQSQGWVHYMIHEPEPHILLFRRLPKPKPKK"
164..291
/gene="CKS1B"
    
```

exon



Data Source	Sequences
GenPept	11,585,396
RefSeq	3,889,502
Third Party Annotation	5,263
Swiss Prot	2,000
PIR	
PRF	
PDB	
(PAT Division)	
<b>Total</b>	
<b>BLAST nr total</b>	
(no patents or env_nr -now 6 million)	



**PAT: patent**



# Search in NCBI Databases

---

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

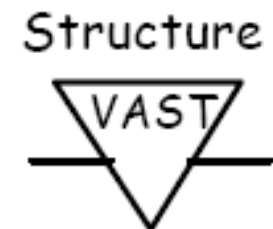
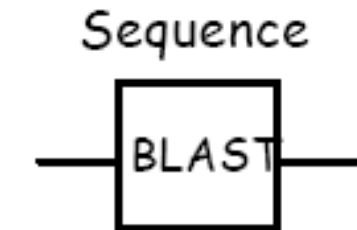
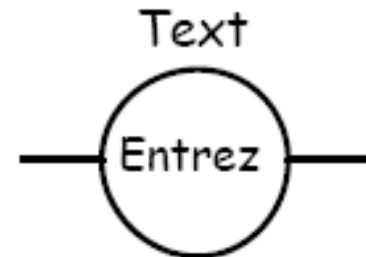
---

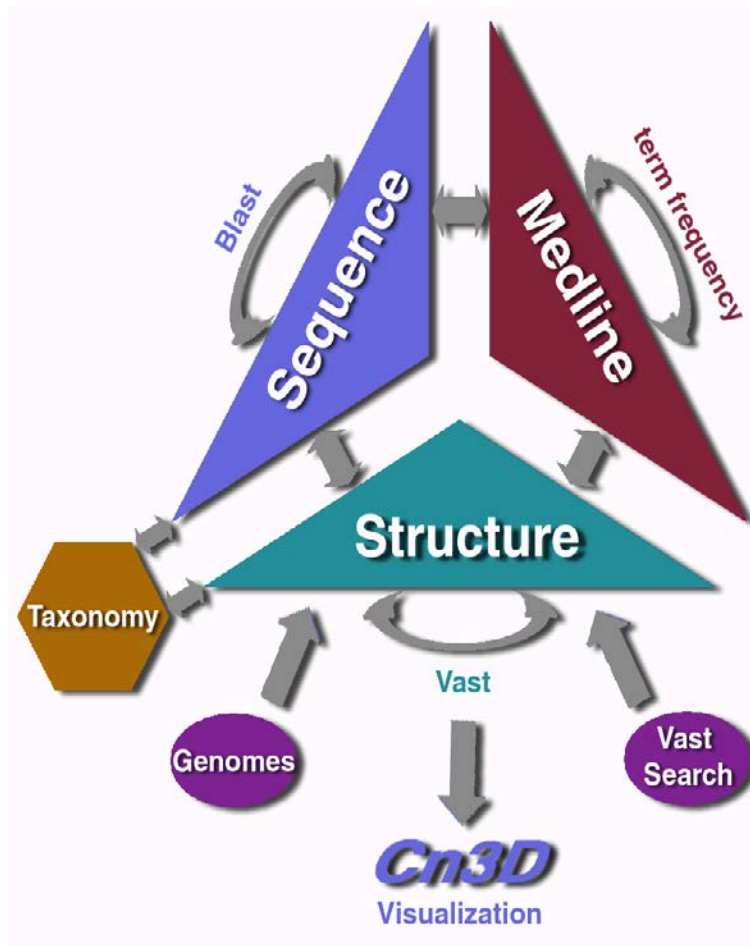
**Sequence:** e.g., [POU5F1](#)

---

**Structure:** e.g., [BRCA1](#)

---



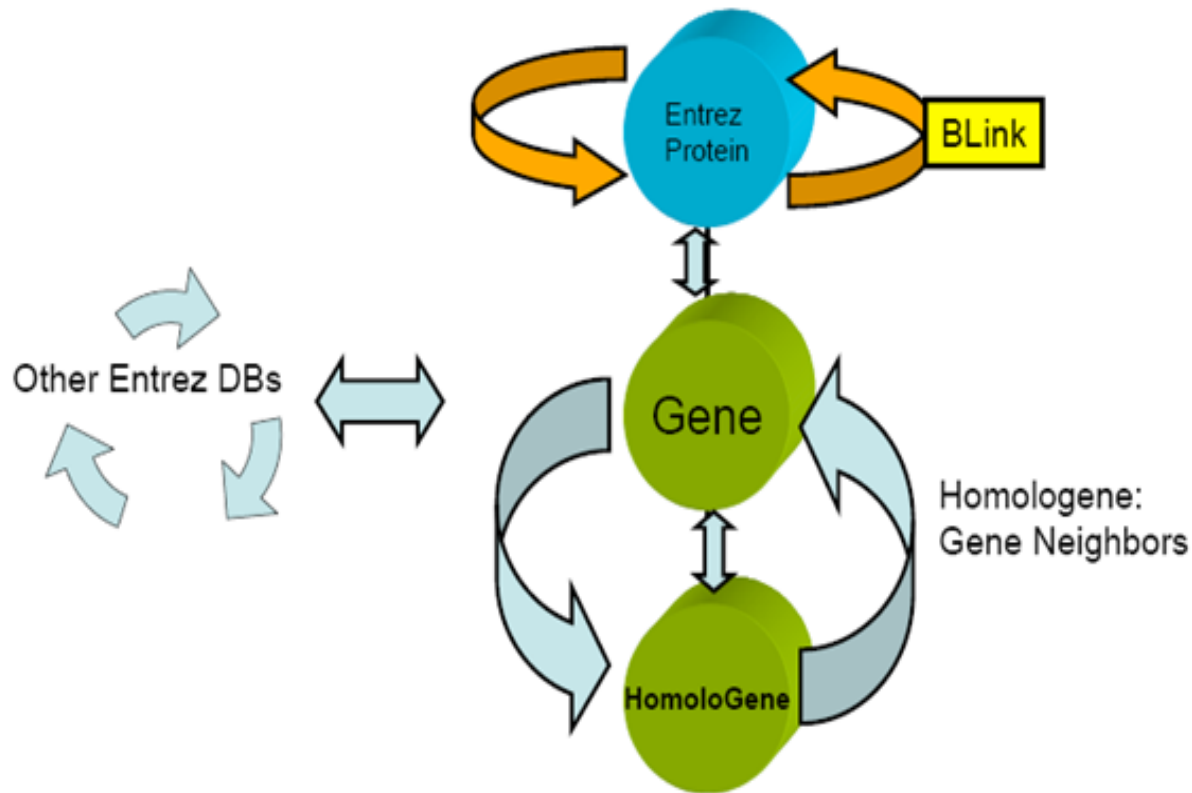


#### ▼ Links

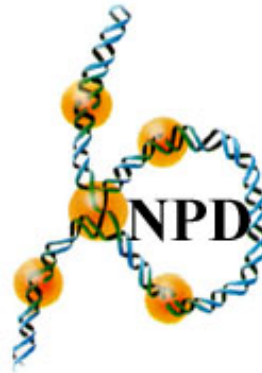
[Explain](#)

[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](#)

# Entrez: Use Gene for everything



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



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



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_hap5:2514038-2520393	POU5F1	800	<input type="button" value="submit"/>

[Click here to reset](#) the browser user interface settings to their defaults. [2011 ENCODE Usability Survey](#)

# 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](#)).

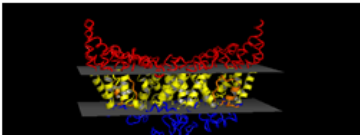
#### Current holdings

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

[more](#)

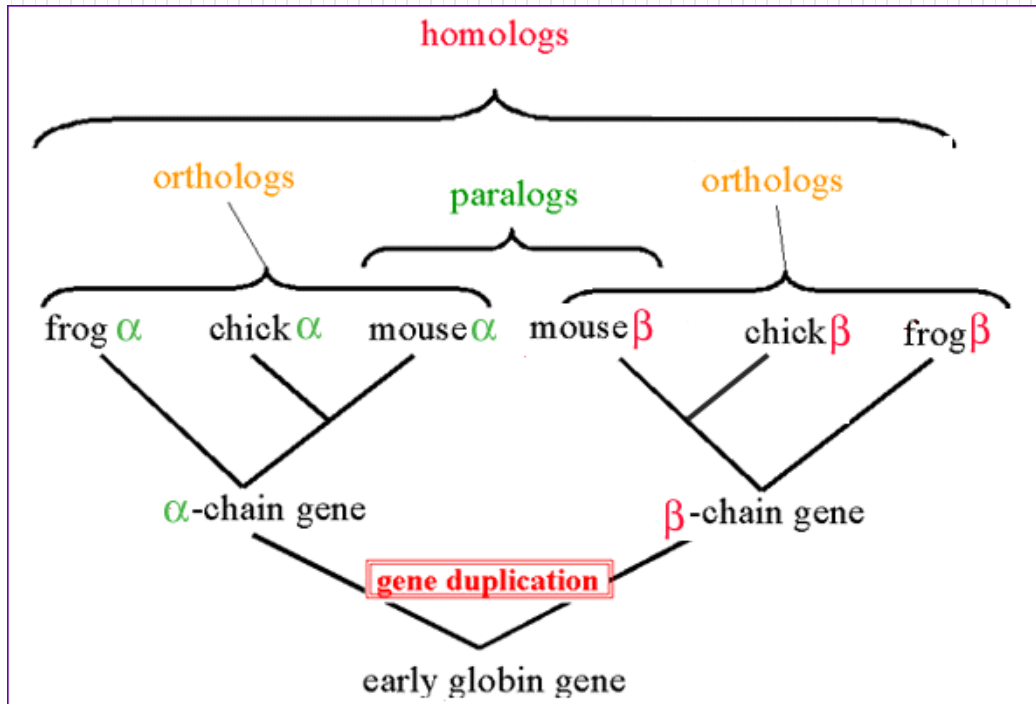
#### Molecule of the month

##### 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

### 1: HomoloGene:8422. Gene conserved in Euteleostomi








#### Genes

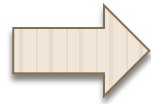
Genes identified as putative homologs of one another during the construction of HomoloGene.

- POU5F1, *Homo sapiens*  
POU class 5 homeobox 1
- POU5F1L, *Pan troglodytes*  
POU domain, class 5, transcription factor 1-like
- POU5F1, *Canis lupus familiaris*  
POU class 5 homeobox 1
- POU5F1, *Bos taurus*  
POU class 5 homeobox 1
- Pou5f1, *Mus musculus*  
POU domain, class 5, transcription factor 1
- Pou5f1, *Rattus norvegicus*  
POU class 5 homeobox 1
- pou5f1, *Danio rerio*  
POU domain, class 5, transcription factor 1

#### Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

- NP\_002692.2 360 aa 
- XP\_001135162.1 359 aa 
- XP\_538830.1 360 aa 
- NP\_777005.1 360 aa 
- NP\_038661.2 352 aa 
- NP\_001009178.1 352 aa 
- NP\_571187.1 472 aa 



---

### All links from this record

[BLink](#)

[Related Sequences](#)

[Identical Proteins](#)

[BioSystems](#)

[CDD Search Results](#)

[Conserved Domains \(Concise\)](#)

[Conserved Domains \(Full\)](#)

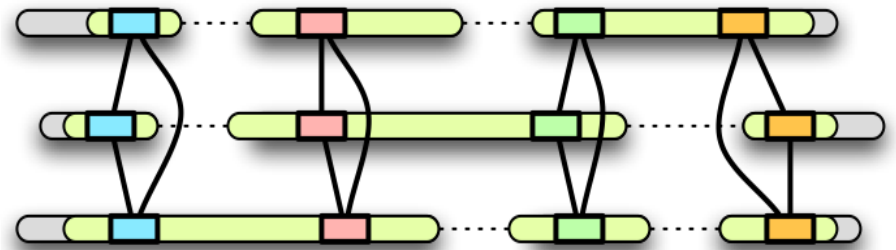
[Domain Relatives](#)

[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

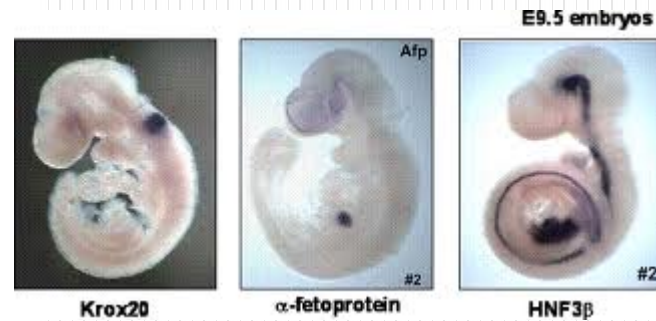
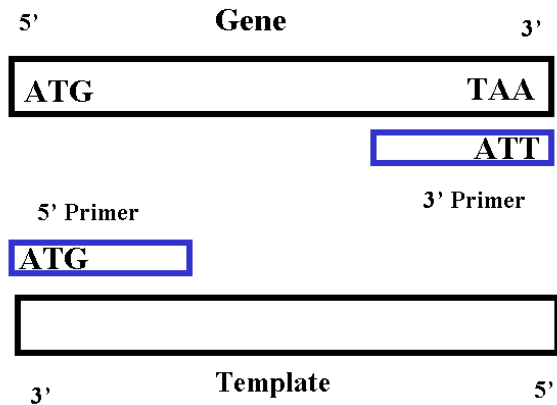
- [Mouse Genome Informatics](#)
- [Mm](#): *Pou5f1* (chr. 17; 19.23 cM)



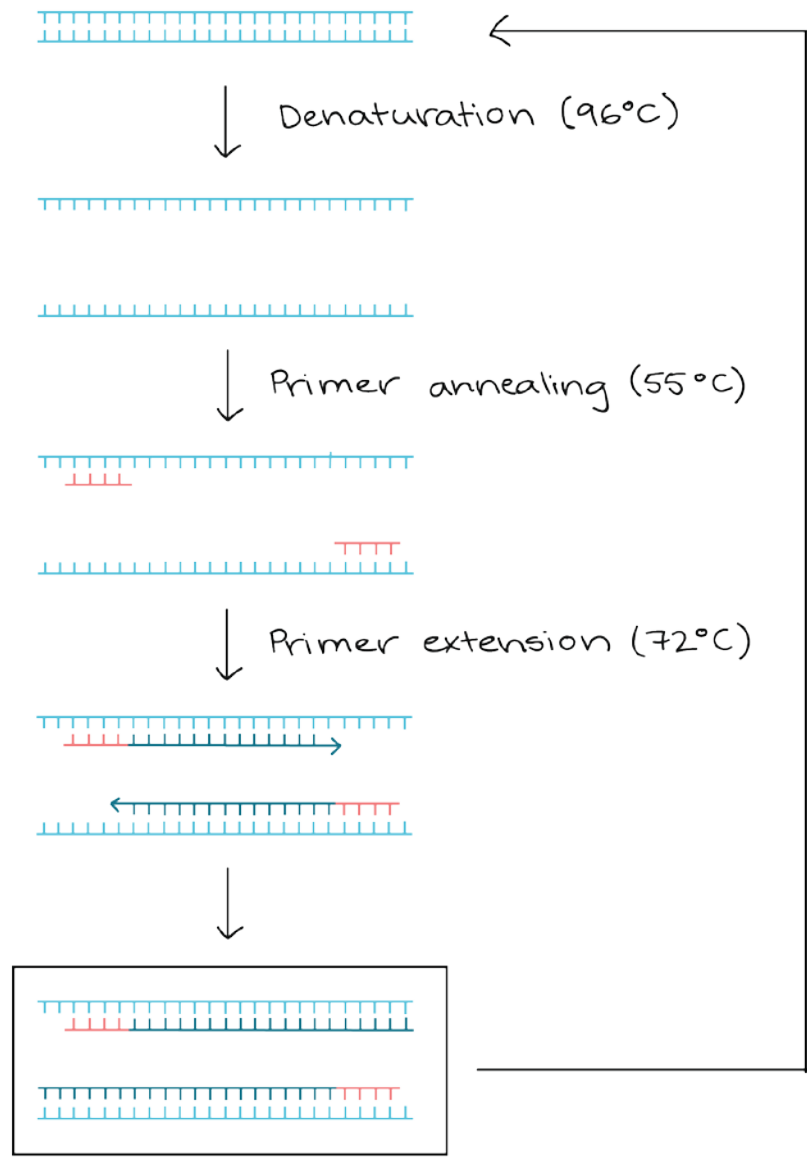
**Mercator**

**Multiple Whole-Genome Orthology Map Construction**

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

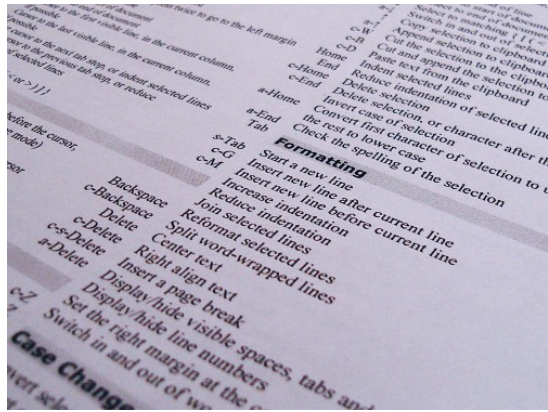
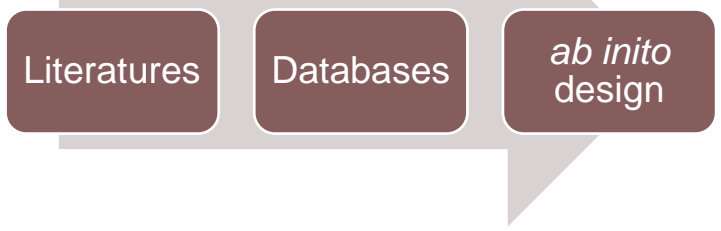






Repeat  
25-35X

Result after 1 cycle:  
# of DNA molecules  
doubled

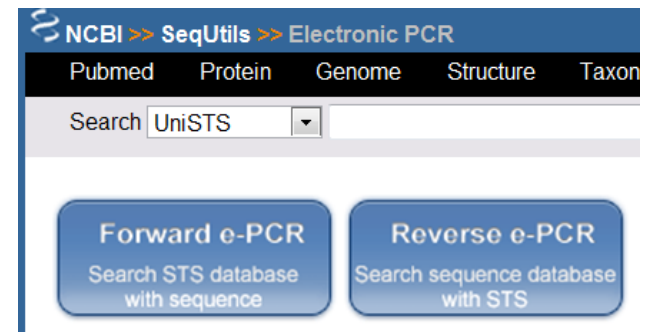


**e.g., ACTB**



**BLAST**

**BLAST**





► [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	<input type="text"/>	<input type="text"/>	<a href="#">Clear</a>
Reverse primer	<input type="text"/>	<input type="text"/>	

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
<input type="text" value="70"/>	<input type="text" value="1000"/>

PCR product size

# of primers to return

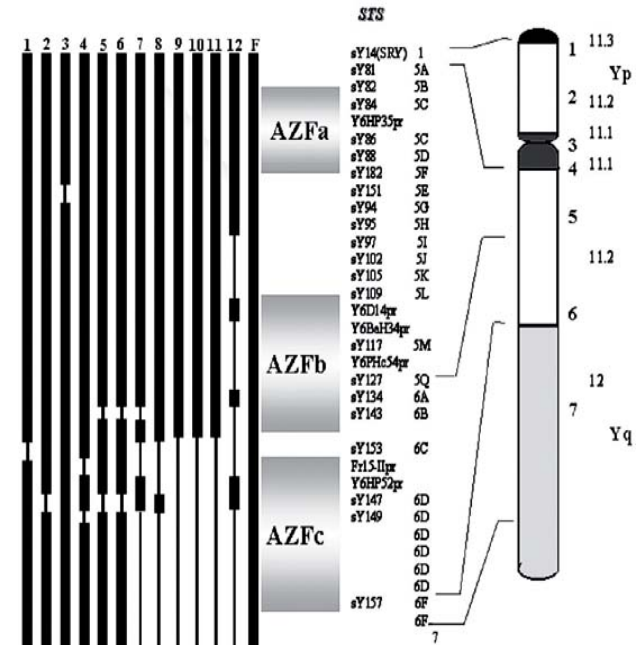
# Sequence Tagged Sites (ST<sub>S</sub>s)

## 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**

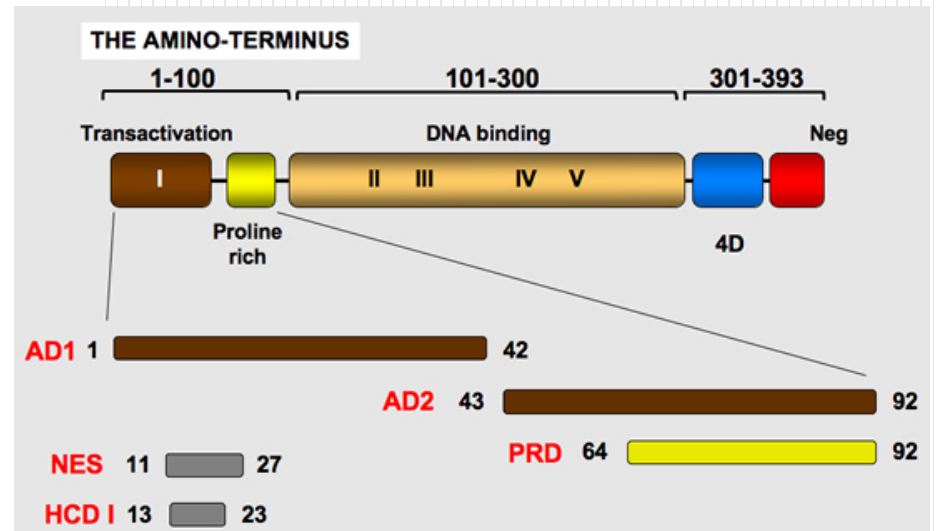
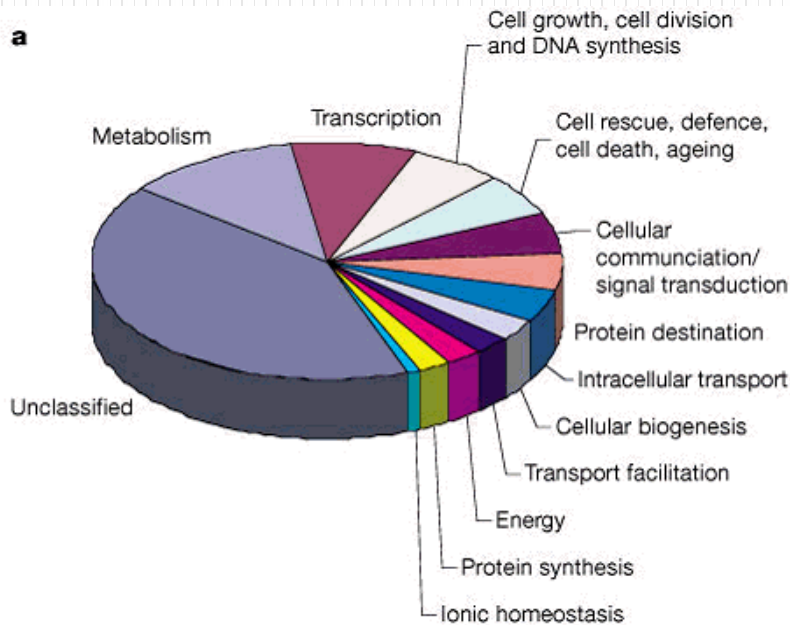
UniSTS contains all the available data on **STS markers (through electronic PCR)**

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



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

**a**



# 1. Gene Ontology

## Through integrated databases

- Entrez\_Gene
  - **GO terms**
- GeneCards
  - **GO terms**
- Uniprot/Swiss-Prot
  - POU5F1\_Human
  - General annotation (comments)
- Ontologies

Function	Evidence
<a href="#">DNA binding</a>	IDA <a href="#">PubMed</a>
<a href="#">miRNA binding</a>	IDA <a href="#">PubMed</a>
<a href="#">promoter binding</a>	IDA <a href="#">PubMed</a>
<a href="#">protein binding</a>	IPI <a href="#">PubMed</a>
<a href="#">sequence-specific DNA binding</a>	IEA
<a href="#">transcription factor activity</a>	IDA <a href="#">PubMed</a>
<a href="#">transcription factor binding</a>	IPI <a href="#">PubMed</a>

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



# 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

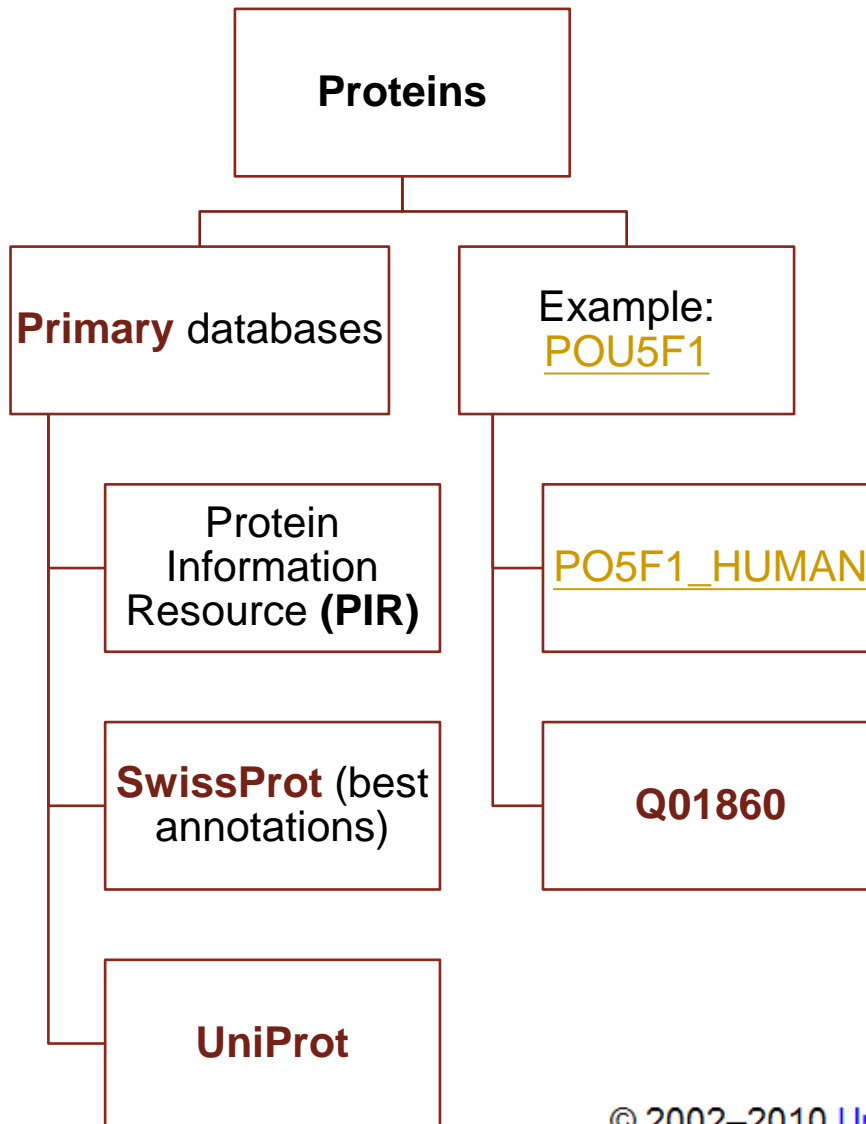
### **Automatically-assigned Evidence Codes**

IEA: Inferred from Electronic Annotation

### **Obsolete Evidence Codes**

NR: Not Recorded

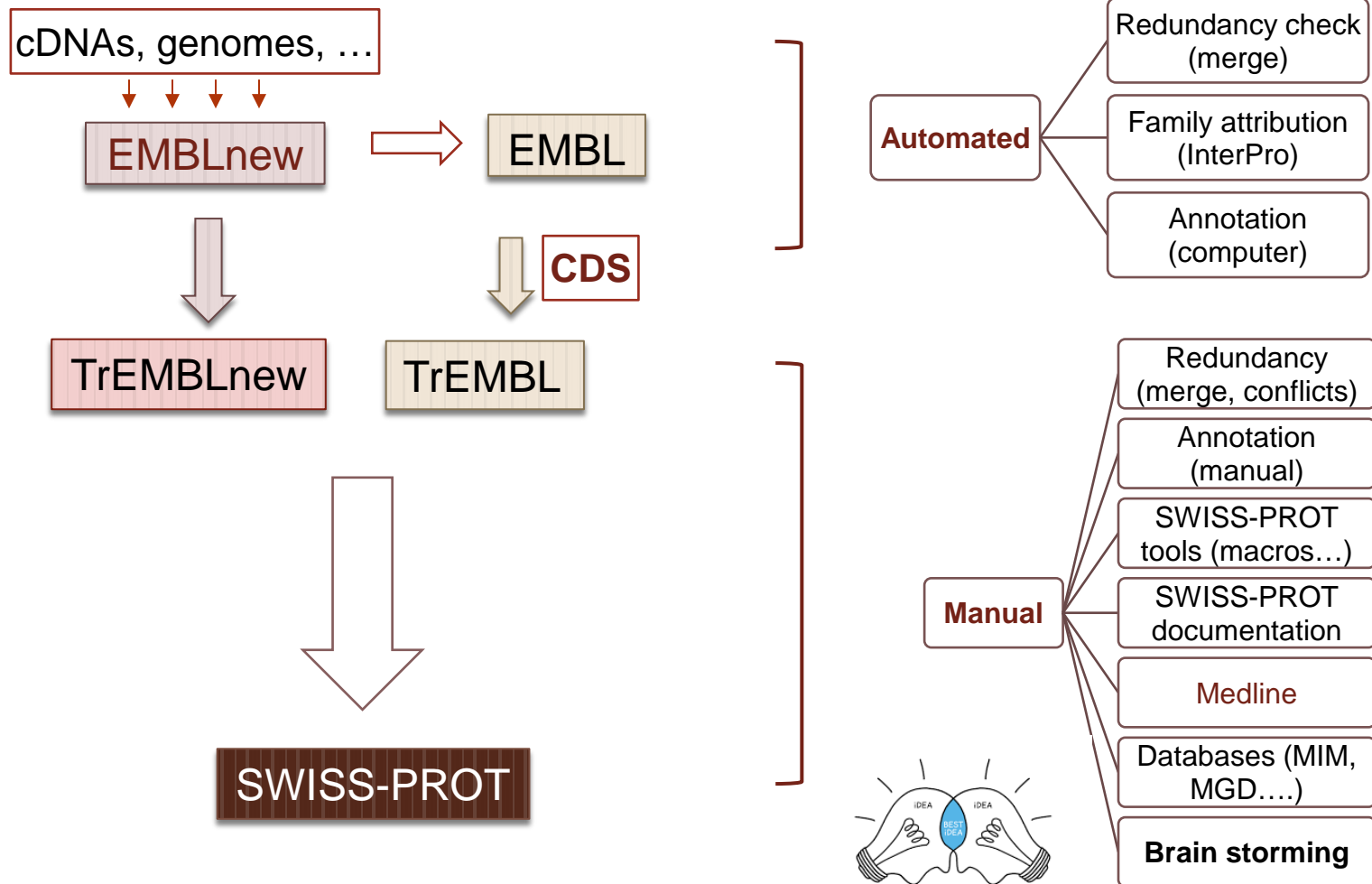
### **Note on Usage of the With/From Column**



© 2002–2010 UniProt Consortium | License & Disclaimer | Contact



# The simplified story of a SWISS-PROT entry



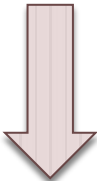
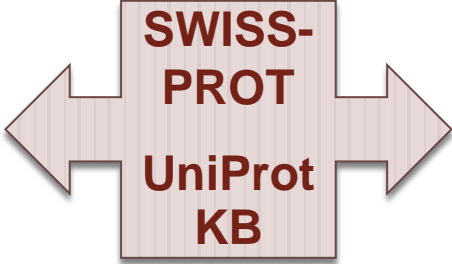
Once in SWISS-PROT, the entry is no more in TrEMBL, **but still in EMBL (archive)**

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



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

UniProtKB  Advanced


BLAST Align Retrieve/ID mapping Help Contact

## UniProtKB results

[About UniProtKB](#) [Basket](#)

Filter by <sup>i</sup> BLAST Align Download Add to basket Columns  1 to 25 of 1,196 Show 25

Reviewed (814) Swiss-Prot  
 Unreviewed (382) TrEMBL

Popular organisms 

- Human (230)
- Mouse (191)
- Rat (117)
- Bovine (63)
- Fruit fly (28)
- Other organisms

<input type="checkbox"/>	Entry	Entry name		Protein names	Gene names	Organism	Length	
<input type="checkbox"/>	P31750	AKT1_MOUSE		RAC-alpha serine/threonine-protein ...	Akt1 Akt,Rac	Mus musculus (Mouse)	480	
<input type="checkbox"/>	P31749	AKT1_HUMAN		RAC-alpha serine/threonine-protein ...	AKT1 PKB,RAC	Homo sapiens (Human)	480	
<input type="checkbox"/>	Q17941	AKT1_CAEEL		Serine/threonine-protein kinase akt...	akt-1 C12D8.10	Caenorhabditis elegans	541	
<input type="checkbox"/>	P47196	AKT1_RAT		RAC-alpha serine/threonine-protein ...	Akt1	Rattus norvegicus (Rat)	480	
<input type="checkbox"/>	Q38998	AKT1_ARATH		Potassium channel AKT1	AKT1 At2g26650,F18A8.2	Arabidopsis thaliana (Mouse-ear cress)	857	
<input type="checkbox"/>	Q8INB9	AKT1_DROME		RAC serine/threonine-protein kinase	Akt1 CG4006	Drosophila melanogaster (Fruit fly)	611	

## Display

## Family & Domains<sup>i</sup>

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

### Domains and Repeats

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

### Region

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

### Domain<sup>i</sup>

Binding of the PH domain to phosphatidylinositol 3,4,5-trisphosphate (PI(3,4,5)P<sub>3</sub>) 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<sup>i</sup>

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



## PTM / Processing<sup>1</sup>

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](#)

### Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Chain <sup>1</sup>	1 – 480	480	RAC-alpha serine/threonine-protein kinase		PRO_0000085605	<a href="#">Add</a> <a href="#">BLAST</a>

### Amino acid modifications

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

# 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

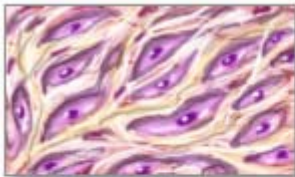
metastasis-associated gene family, member 2

# Databases for Protein – Protein Interaction

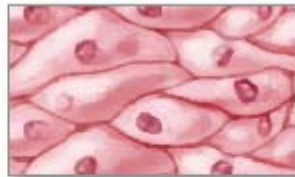
Resource	Comments
<a href="#">APID</a>	Agile Protein Interaction DataAnalyzer (Cancer Research Center, Salamanca, Spain)
<a href="#">BIND</a>	Biomolecular INteraction Network Database at the University of Toronto, Canada. No species restriction
<a href="#">CYGD</a>	PPI section of the Comprehensive Yeast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

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

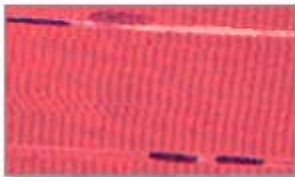
Four types of tissue



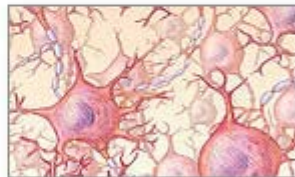
Connective tissue



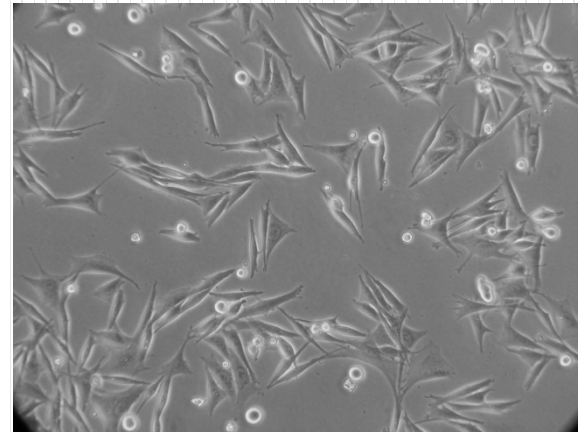
Epithelial tissue



Muscle tissue



Nervous tissue



All Databases PubMed Nucleotide Protein Genome Structure OMIM

Search UniGene for POU5F1

UniGene  
Homepage

UniGene: An Organized View of the Transcriptome.

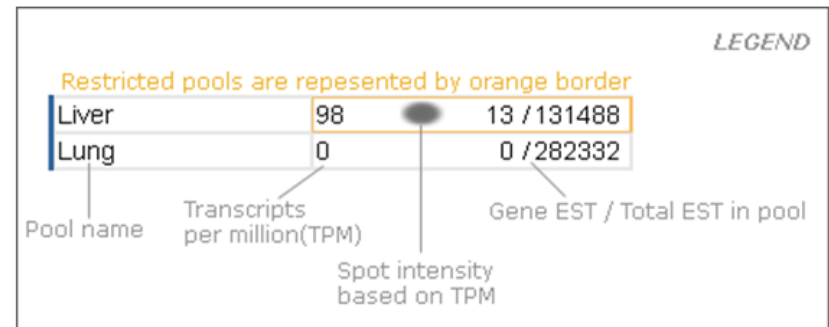
ascites	0	0 / 40015
bladder	0	0 / 29757
blood	0	0 / 123478
bone	0	0 / 71655
bone marrow	0	0 / 48801
brain	1	2 / 1100989
cervix	0	0 / 48171
connective tissue	0	0 / 149255
ear	0	0 / 16212
embryonic tissue	407	88 / 215722
esophagus	0	0 / 20209
eye	0	0 / 211054
heart	0	0 / 89626
intestine	0	0 / 234472
kidney	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.249104

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-.

[Copyright notice.](#)



*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

U.S. National Library of Medicine  
National Institutes of Health

Search  ▾

[Limits](#) [Help](#)

Bookshelf ID: NBK1969



## NCBI News

Bethesda (MD): [National Center for Biotechnology Information \(US\)](#); 199

ISSN: 1060-8788

Publication No.: 94-3272

[Copyright notice.](#)

## Index of Issues

☐ [NCBI News, March 2011](#)

[PubMed Interface for Mobile Devices Now Available](#)

[NCBI Bookshelf Updated to the New Entrez Design](#)

[New Organism Builds in UniGene](#)

[NCBI YouTube Video Update](#)

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

This video contains animal  
experiment content,  
Viewer discretion is advise



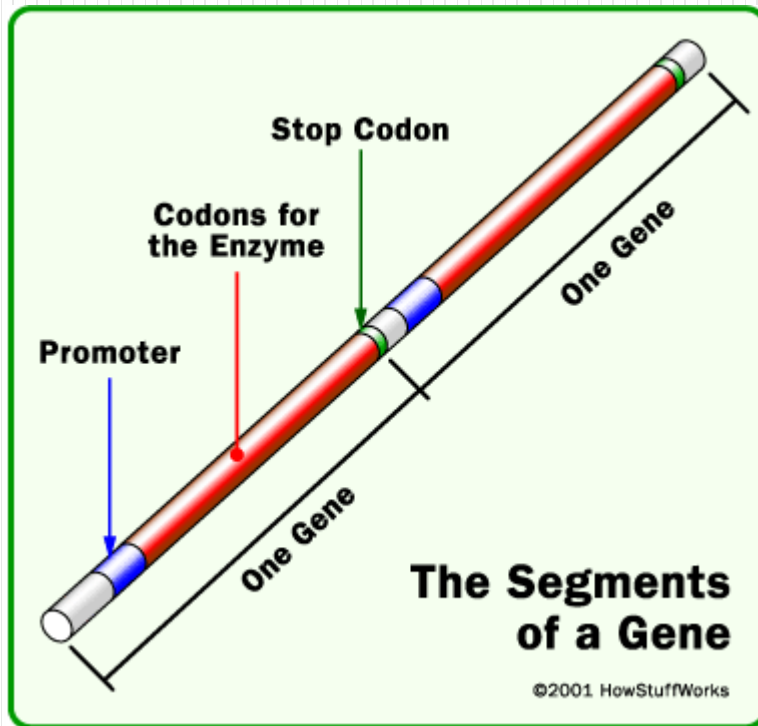
0:02 / 3:01



240p




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



[▶ 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 **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)

Human genomic plus transcript (Human G+T) 

Or, upload file

Job Title

 Align two or more s

## Choose Search Se

Database

Exclude

Optional

Entrez Query

 Models (XM/XP)  Uncultured/environmental sample sequencestranscript  Others (nr etc.):

[NCBI Homepage](#)

## Contamination

[Definition](#)  
[Sources](#)  
[Consequences](#)  
[Detection](#)

## VecScreen

[Overview](#)  
[Example](#)  
[Search Parameters](#)  
[Match Categories](#)  
[Interpretation](#)  
[Exceptions](#)

## UniVec Database

[Overview](#)  
[Redundancy](#)  
[Elimination](#)  
[Benefits](#)  
[Pseudo-](#)  
[Circularization](#)  
[Vector Representation](#)

## ▶ Screen a Sequence Using VecScreen

Enter your query sequence below as an Accession, GI, or **FASTA**.

## ▶ 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.





# 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 finds selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic code. The sequence can be saved in various formats and searched against the sequence database. The ORF Finder should be helpful in preparing complete and accurate sequence submissions to the Sequin sequence submission software.

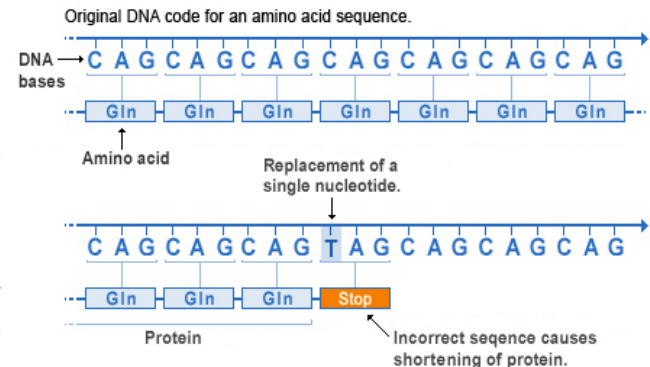
Enter GI or ACCESSION  OrfFind Clear

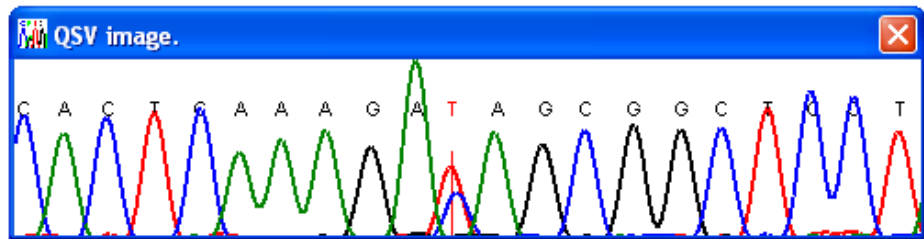
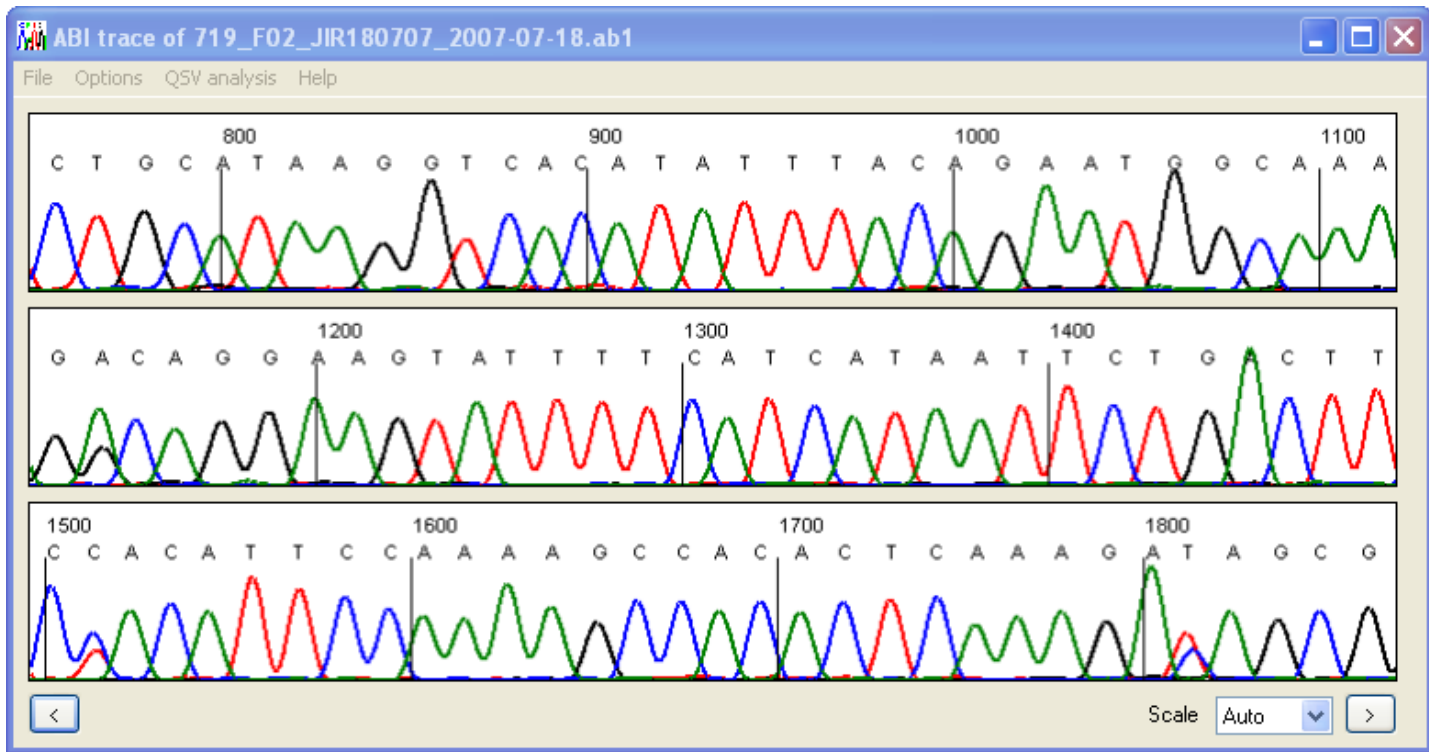
or sequence in FASTA format

FROM:  TO:

Genetic codes 1 Standard

Nonsense mutation





# When Cloned by Emails – get the map & confirmed

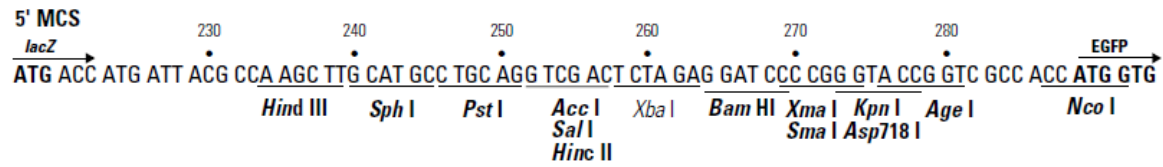
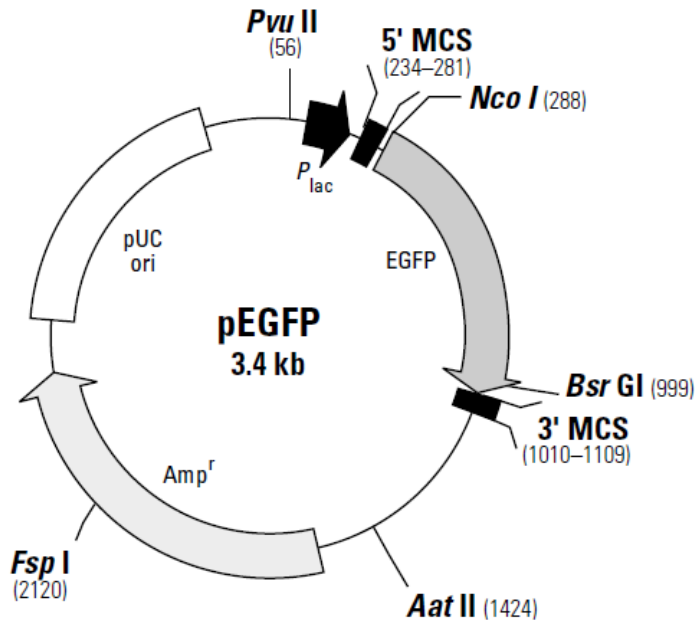
Specific EGFP Monoclonal Antibody for Westerns, IP and IC

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



## 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

[Click here to reset](#) the browser user interface settings to their defaults. **2011 ENC**

- POU5F1
- POU5F1B
- POU5F1P1
- POU5F1P3
- POU5F1P4
- POU5F2
- POU5FLC12

**Survey**

## About the Human Feb. 2009 (GRCh37/hg19) assembly ([sequences](#))

The February 2009 human reference sequence (GRCh37) was produced by the [Genome Reference](#)

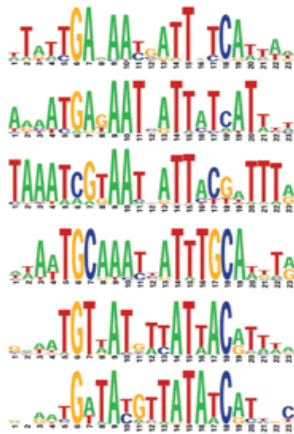
## UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

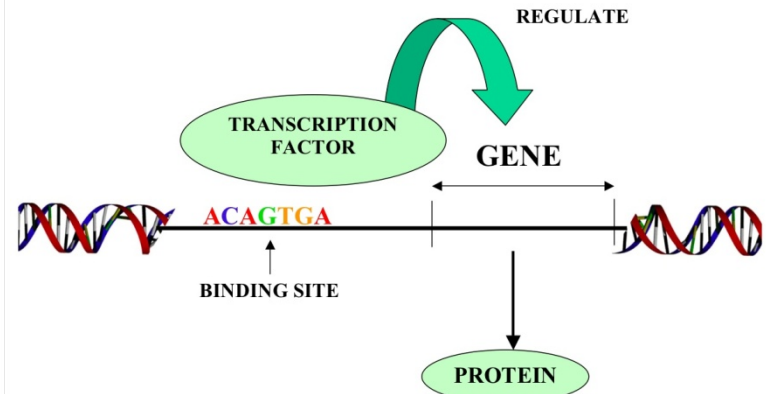
position/search chr6\_mcf\_hap5:2,514,038-2,520,39 [gene](#)    size 6,356 bp.  **2011 ENCODE Usability Survey**

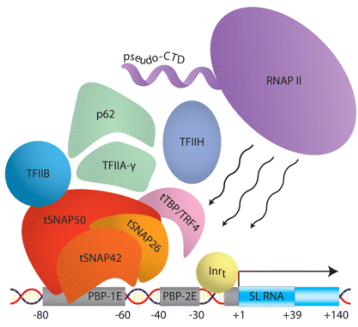


# Q: How to Identify Potential Regulators?



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

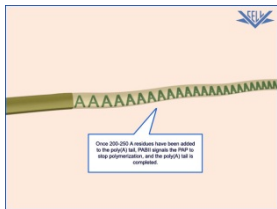
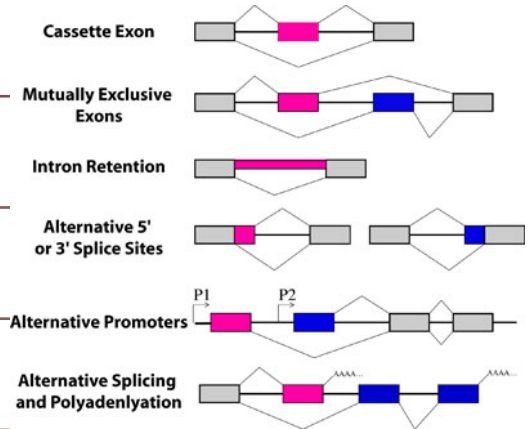
PolyA sites...

Consensus sequences

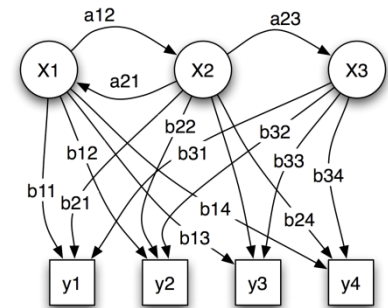
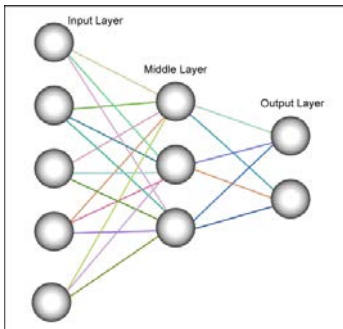
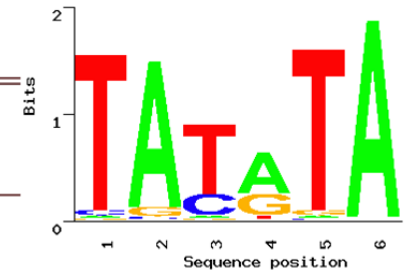
Weight matrices

**Neural networks (NNs)** Decision trees

**Hidden Markov Models (HMMs)**



Wide range of **methods**



# Promoter Databases and sites for analysis, prediction and search

[AlignACE](#)

motif-finding algorithm.

[Promoter Binding Element Database](#)  
[CpG promoter](#)

Arabidopsis thaliana promoter binding element database  
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

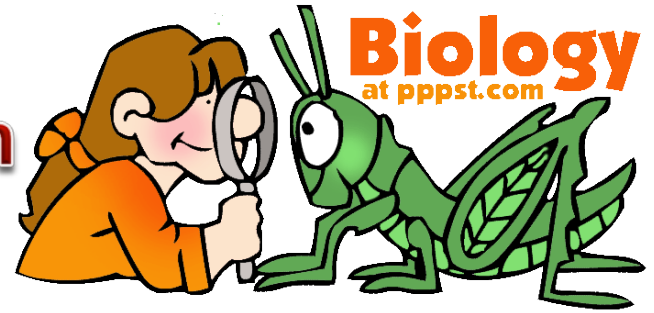
[Motif Explorer](#)

Motif & promoter visualization

[Neural Network Promoter Prediction](#)

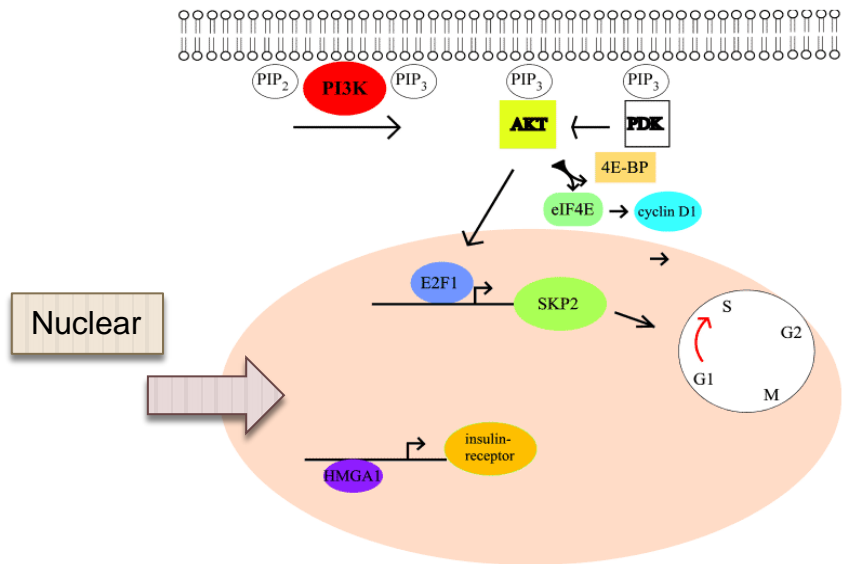
Neural Network Promoter Prediction

# Pattern-driven



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

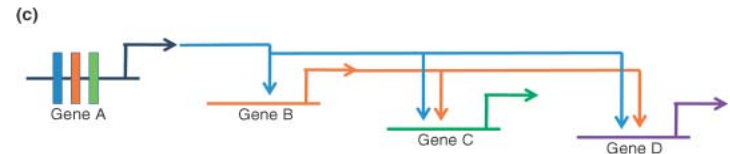
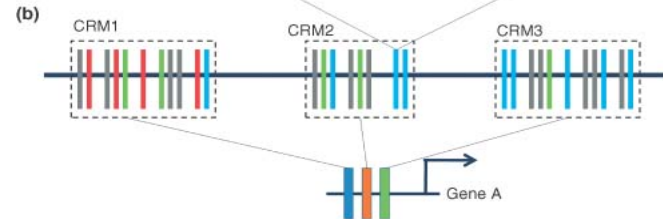
- Tend to produce huge numbers of **false-positive**
- **Reasons**
  - Binding sites (BS) for specific TFs often **variable**
  - Binding sites are short (typically **5-15 bp**)
  - **Interactions** between TFs (& other proteins) influence **affinity** & **specificity** of TF binding
  - One binding 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

(a)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	25	28	70	10	0	100	0	0	0	0	2	6	18	31
C	31	7	3	69	100	0	100	0	0	0	19	21	47	13
G	13	47	21	19	0	0	0	100	0	100	69	3	7	31
T	31	18	6	2	0	0	0	0	100	0	10	70	28	25



## 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**

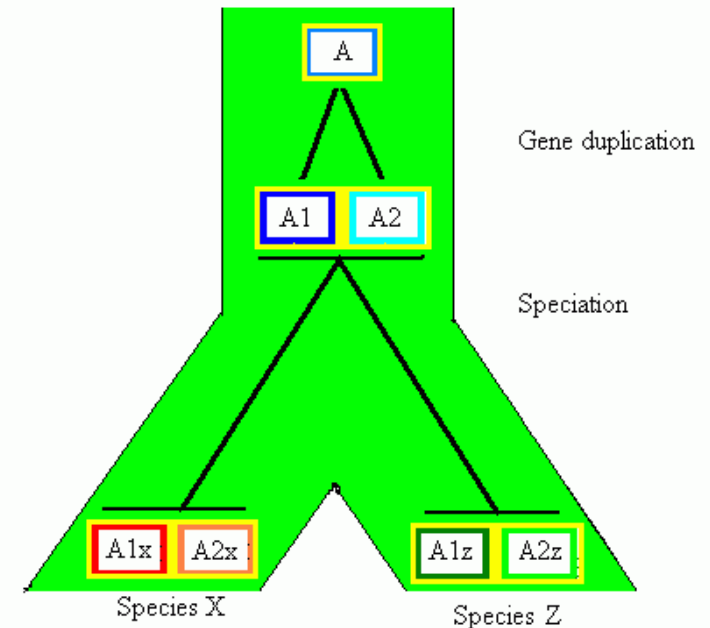
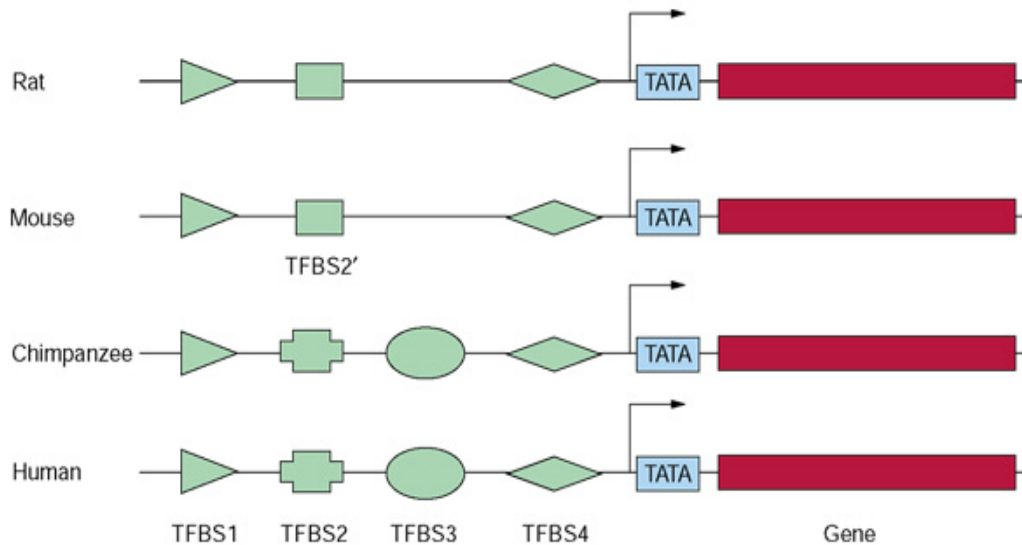
- But **NOT** for enhancers
- Only a **small fraction of TSSs** have been experimentally mapped

**Comparative** promoter mapping

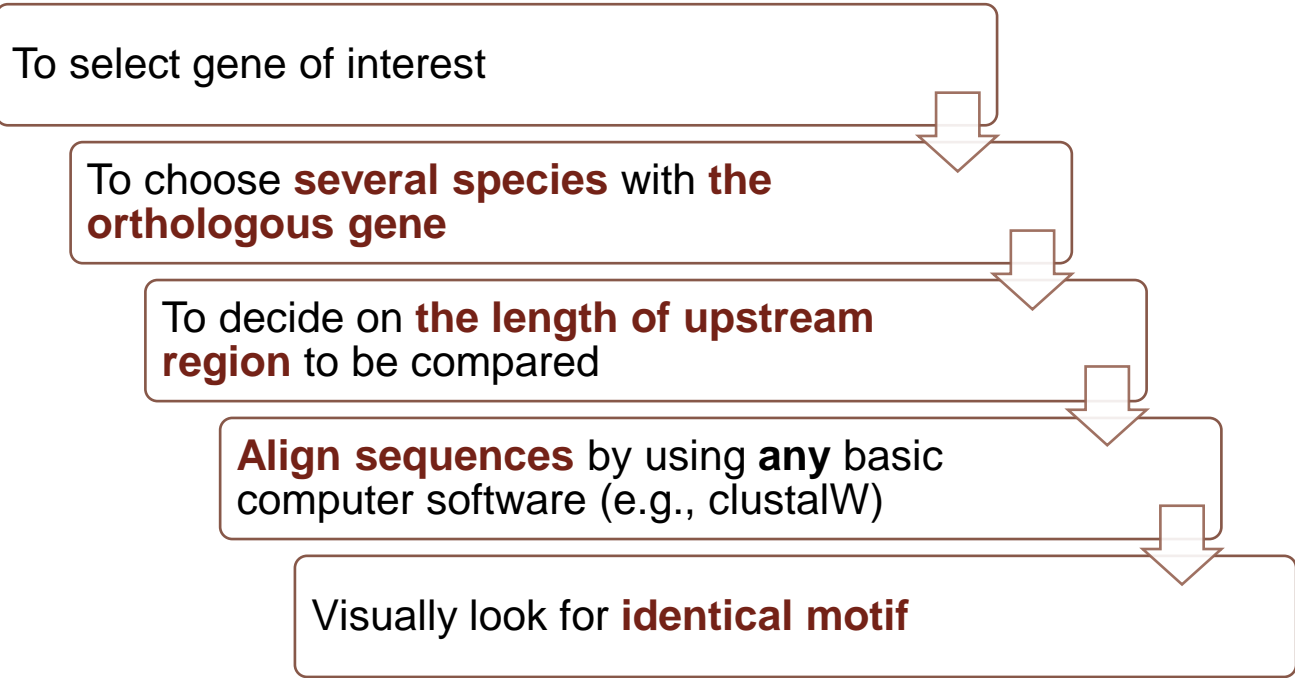
# Phylogenetic Footprinting

Patterns of gene regulation are often conserved across species

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







```

Human  TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Dog     TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Mouse  TCACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTGCGGAAGCGAAGCGATGCGCCAGTCTCCAGCGGGTGGCGCTCGAGTCCGA 941

Human  CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCCACCCCTTCGCCTCCACCCAACTCCCTATTAGTGCACGAGTTTACCTCTAG 865
Dog     CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCCACCCCTTCTCGCCTCCACCCAACTCCCCATTAGTGCACGAGTTTACCTCTAG 865
Mouse  CTGAACGGCGGCAACGGTGGCGGGGACGCGCCAGGGGCGCGCGCCACCCCTCTGCCTCCACCCAACTC----- 1014
  
```

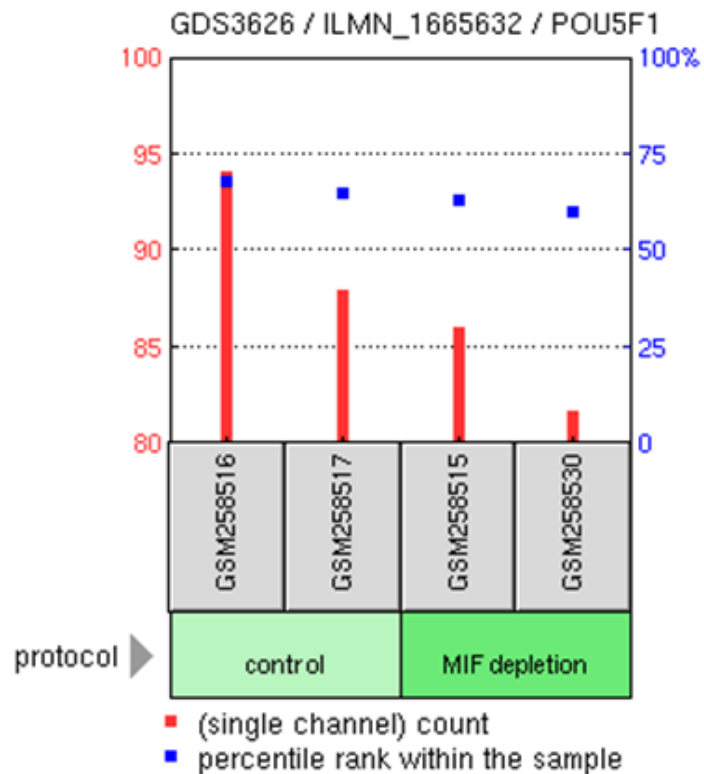
Potential TFBS: **Ubx1 binding site**  
**NP-γ binding site**  
**SP1 binding site**  
**GATA-1 binding site**

\*All TF names are from human with orthologous TFs present in both dog and mouse.

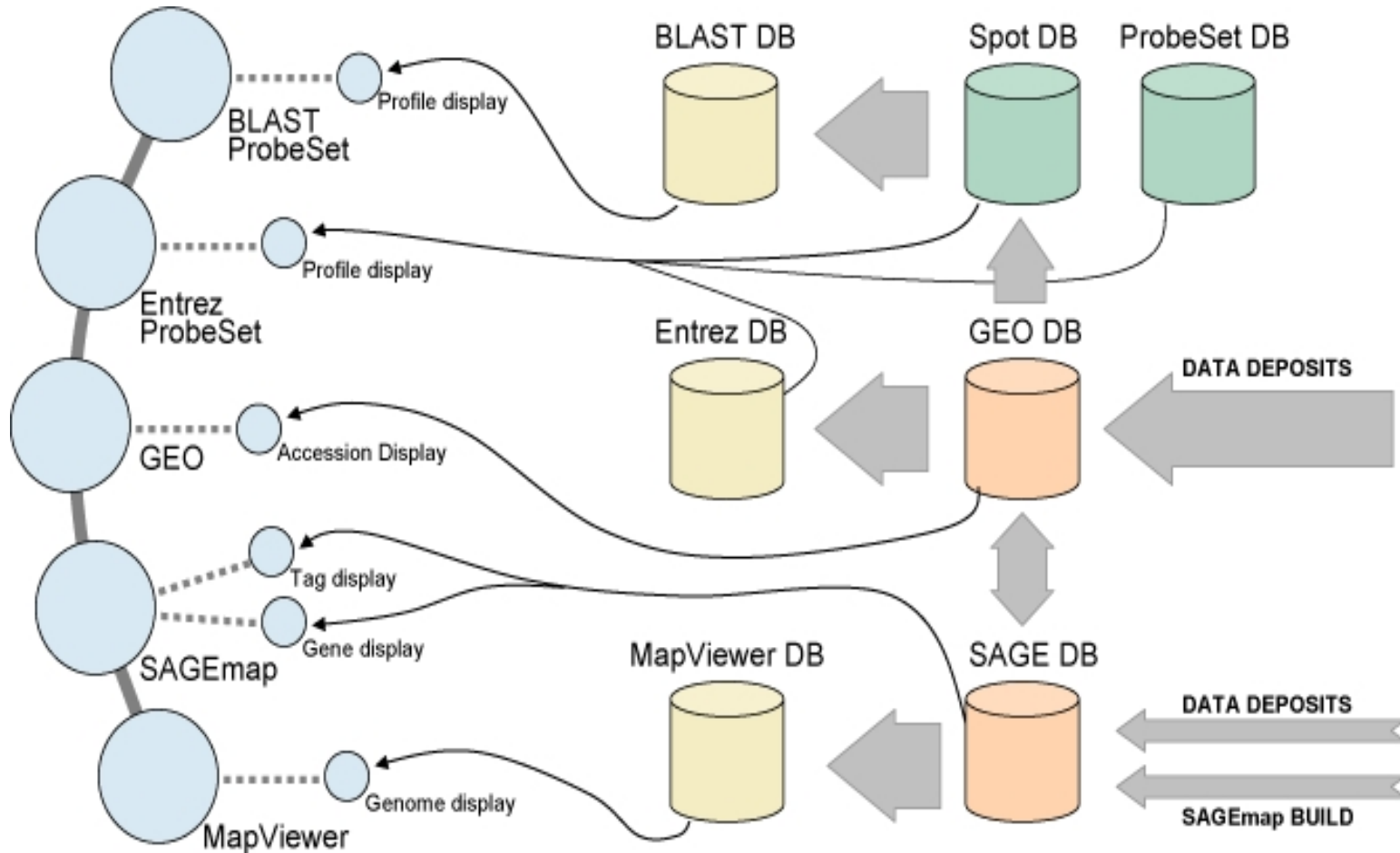
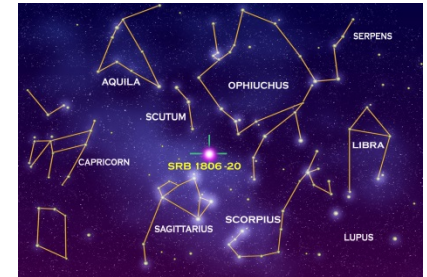
# One More Trick - Coregulation

**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



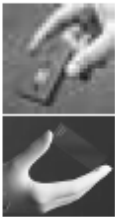
# Constellation of NCBI Gene Expression Resources



# Gene Expression Omnibus (GEO) (1)

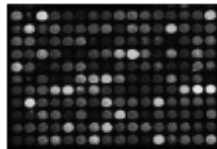
Submitted by  
Manufacturer\*

**GPL**  
Platform  
descriptions



Submitted by  
Experimentalists

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



Entrez GEO

**GSE**  
Grouping of  
slide/chip data  
“a single experiment”

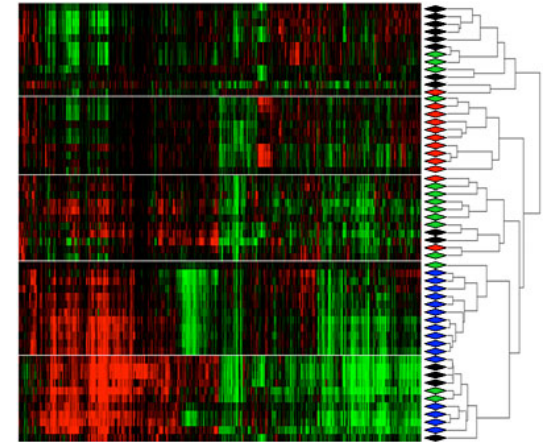


Curated by  
NCBI

**GDS**  
Grouping of  
experiments



Entrez  
GEO Datasets



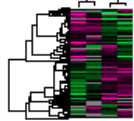
# Gene Expression Omnibus (GEO) (2)

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

Search for     [Advanced Search](#)

DataSet Record GDS46: <a href="#">Expression Profiles</a> <a href="#">Data Analysis Tools</a> <a href="#">Sample Subsets</a>			
<b>Title:</b>	E2F1-regulated genes		
<b>Summary:</b>	Identification of E2F1-regulated genes that modulate the transition from quiescence into DNA synthesis, or have roles in apoptosis, signal transduction, membrane biology, and transcription repression.		
<b>Organism:</b>	<i>Mus musculus</i>		
<b>Platform:</b>	GPL75: [Mu11KsubA] Affymetrix Murine 11K SubA Array		
<b>Citation:</b>	Ma Y, Croxton R, Moorer RL Jr, Cress WD. Identification of novel E2F1-regulated genes by microarray. <i>Arch Biochem Biophys</i> 2002 Mar 15;399(2):212-24. PMID: 11888208		
<b>Reference Series:</b>	<a href="#">GSE498</a>	<b>Sample count:</b>	4
<b>Value type:</b>	count	<b>Series published:</b>	2003/07/16

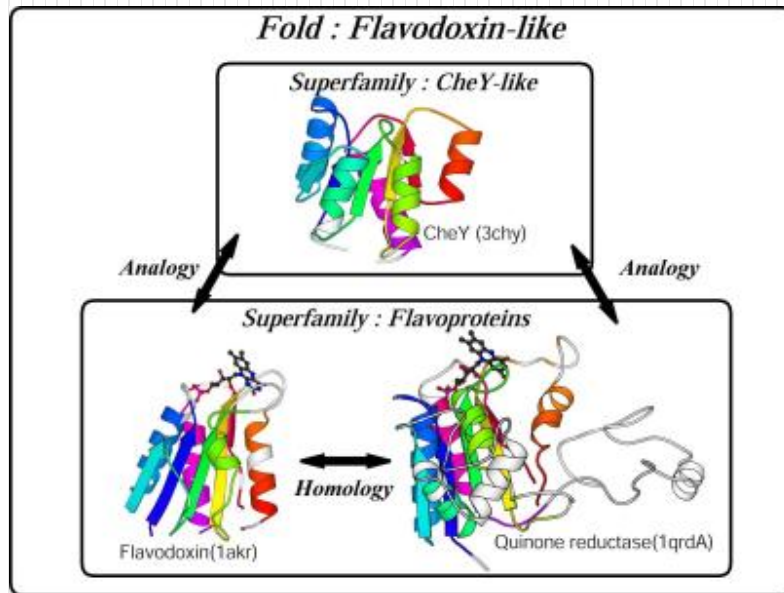
**Cluster Analysis**



**Download**

- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

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

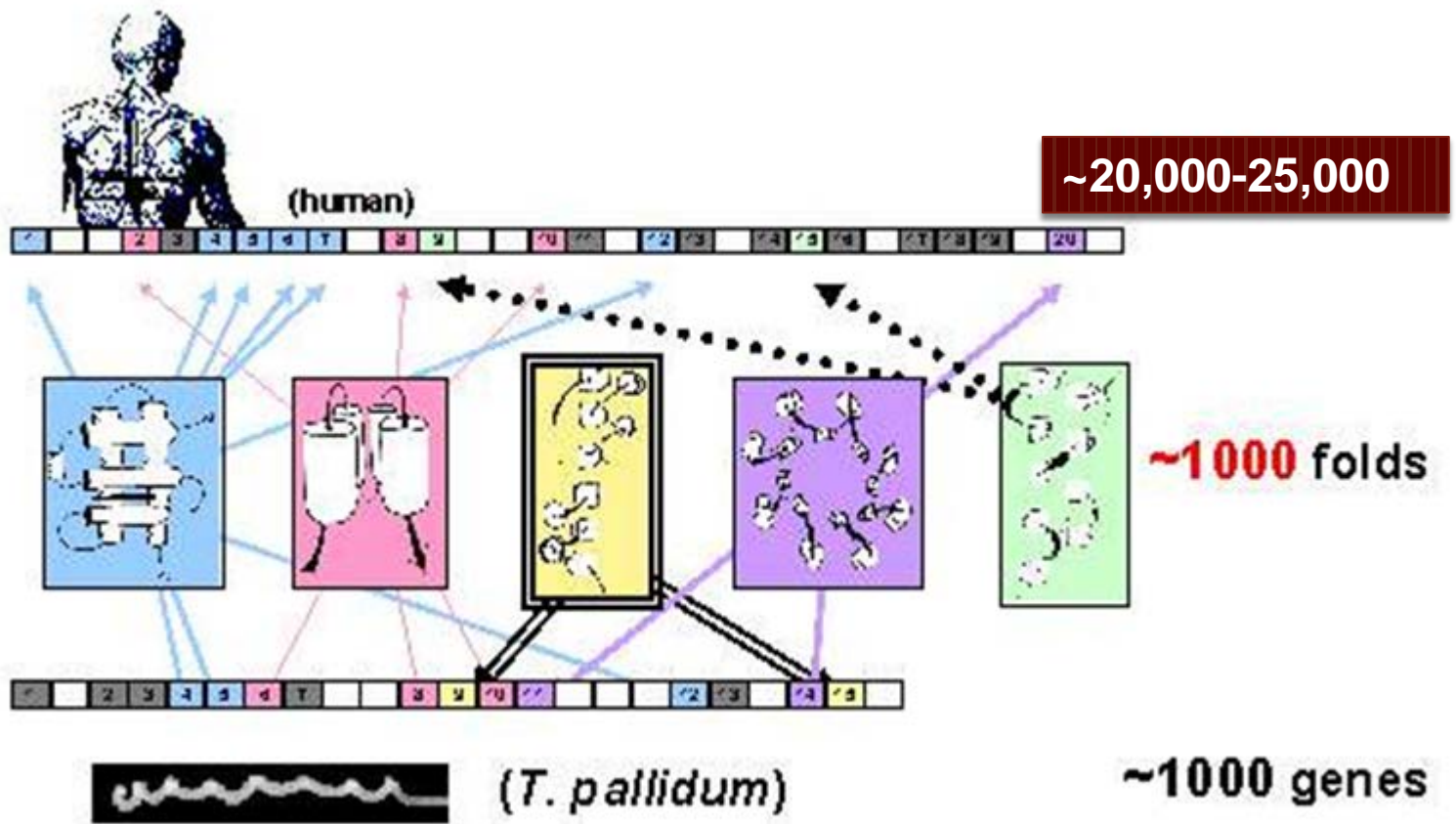


# Structures are More Conserved Than Sequences

Evolution	Homology	% Identity	Alignment Methods
Recent relationship - less divergence	Sequence alignments can be used to infer homology	100	Automatic Pairwise Alignment Methods
Increasing divergence		90	
		80	
		70	
		60	
		50	
Distant relationship		40	Consensus Methods
		30	Profile Methods
		20	Structure Prediction
		10	
		0	
	<b>Twilight Zone</b>		
	<b>Midnight Zone</b>		



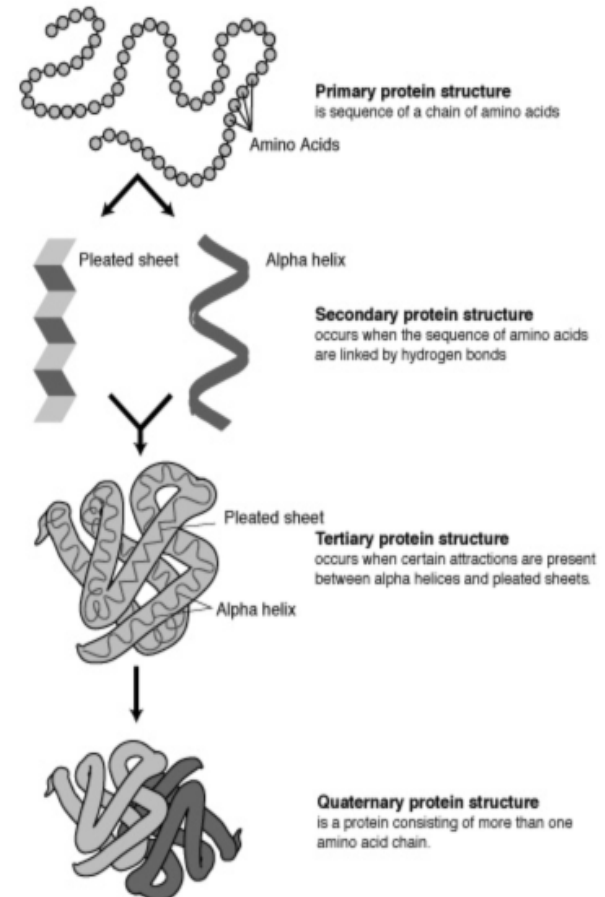
# Simplifying Genomes with Folds, Pathways



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 @, *, #



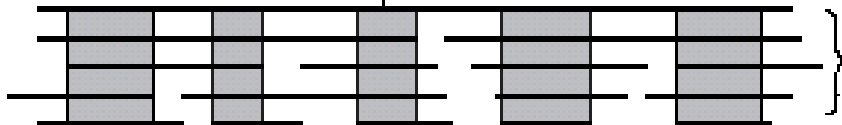
Single motif methods

permissive regular expression (IDENTIFY)

eMotif

exact regular expression (PROSITE)

XXXX  
XXXX  
XXXX



Full domain alignment methods

Profile (Profile library)

Hidden Markov Model (Pfam)

RWDAGCVN  
RWDSGCVN  
RWHHGCVQ  
RWKGACYN  
RWLVACEQ

XXXX  
XXXX  
XXXX

XXXX  
XXXX  
XXXX

XXXX  
XXXX  
XXXX

XXXX  
XXXX  
XXXX

frequency matrices (PRINTS)

position-specific weight matrices (Blocks)

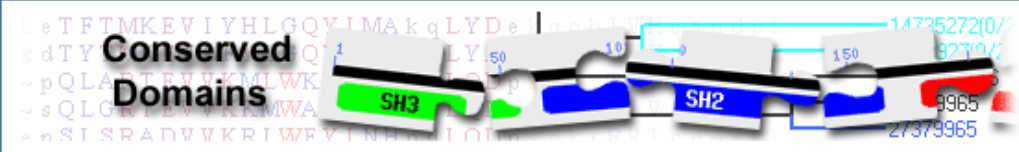
Multiple motif methods

Attwood 2000

# Major Secondary “Pattern” Database

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

[http://www.ncbi.nlm.nih.gov/Structure/structure\\_docs/CDD.pdf](http://www.ncbi.nlm.nih.gov/Structure/structure_docs/CDD.pdf)



Search  for    Help

## Conserved Domains and Protein Classification

[RESOURCES](#) [SEARCH](#) [HOW](#)

### Resources Highlig

#### 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](#)

Search Database

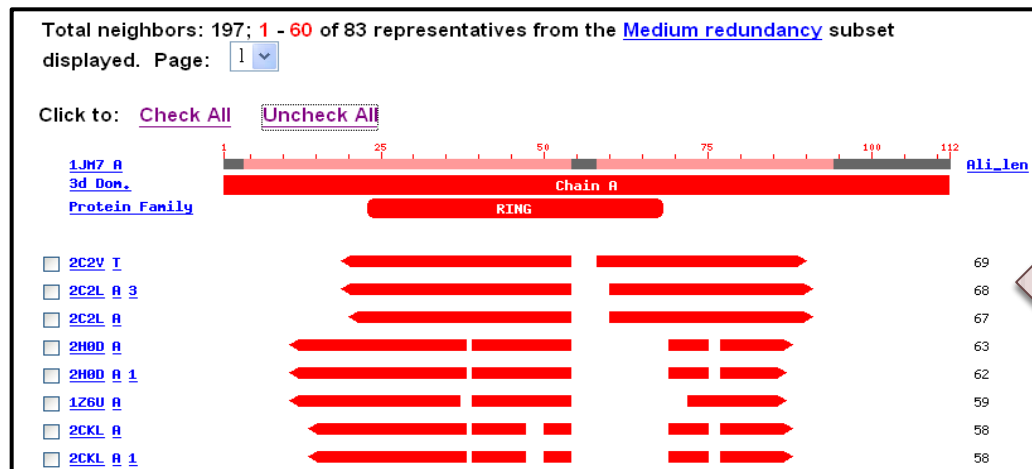
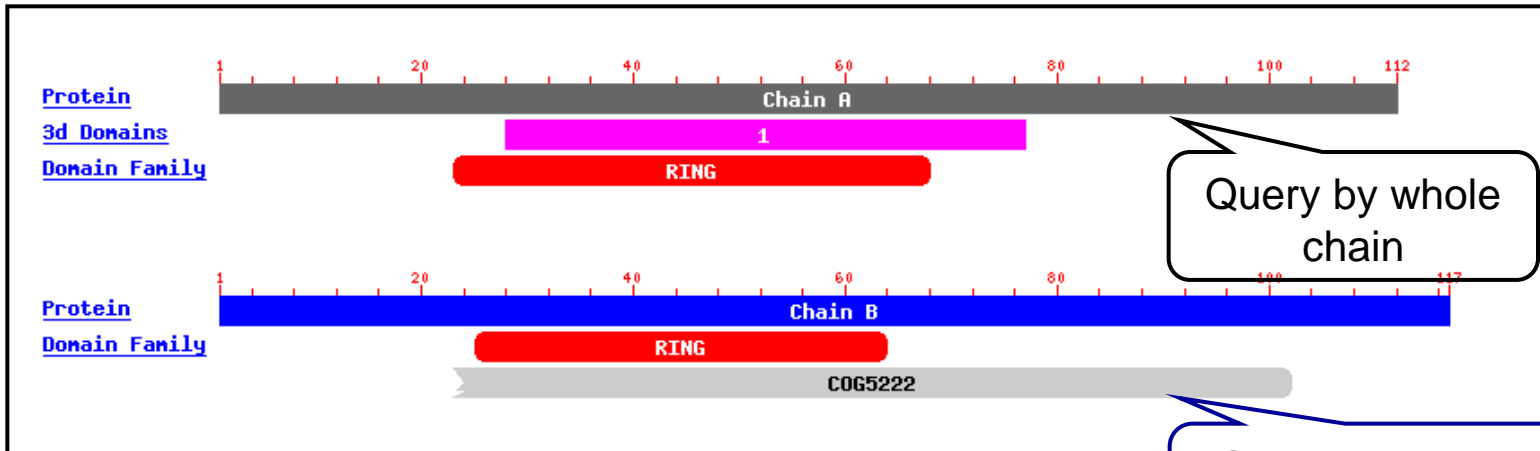
- CDD v2.28 - 39357 PSSMs
- SMART v5.1 - 791 PSSMs
- Pfam v24.0 - 11912 PSSMs
- COG v1.00 - 4873 PSSMs
- KOG v1.00 - 4825 PSSMs
- PRK v6.00 - 10885 PSSMs
- TIGR v10.00 - 4023 PSSMs

#### CDART: Conserved Domain Architectures

Conserved Domain Architecture database based on domain architecture queries. CDART finds protein sir

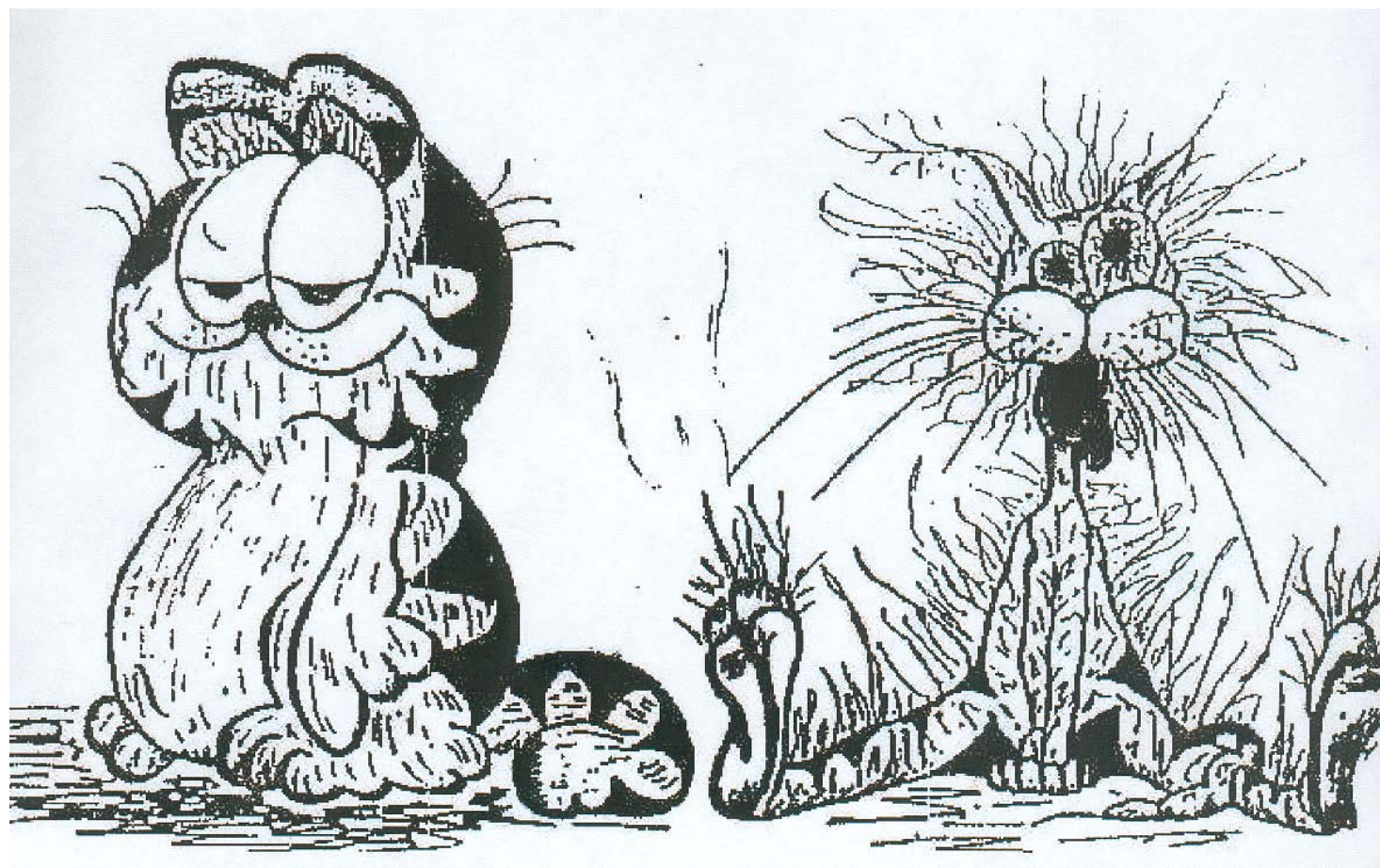
forms similarity searches of the Entrez Protein database in the natural order of conserved domains in protein sequences. CDART finds evolutionary distances using sensitive domain

# VAST: Query by Chain or 3D Domain





**Before...**



**After...**