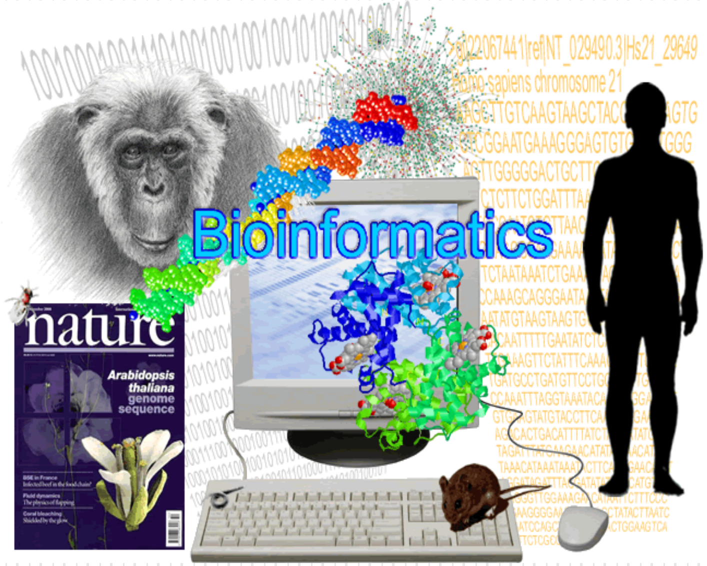
Bioinformatics, Syntenic Biology & Genome Editing


Bioinformatics
>gi|22067441|ref|NT_029490.3|Hs21_29649
Homo sapiens chromosome 21
nature
Arabidopsis thaliana genome sequence

薛佑玲 PhD

Institute of Biomedical Sciences

National Sun Yat-sen University

ylshiue@mail.nsysu.edu.tw

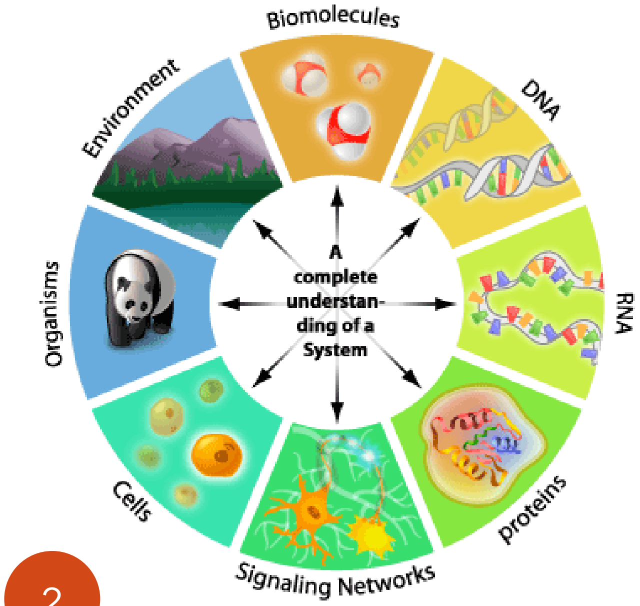
Outline

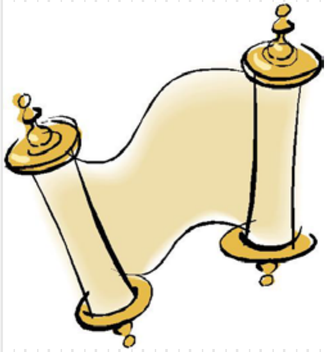
Introduction: a Short History About Bioinformatics

Bioinformatics Q & A

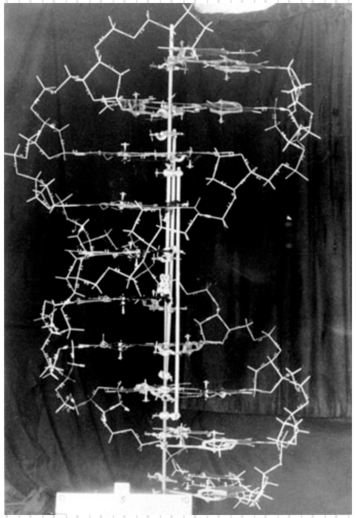
Synthetic biology

Genome Editing

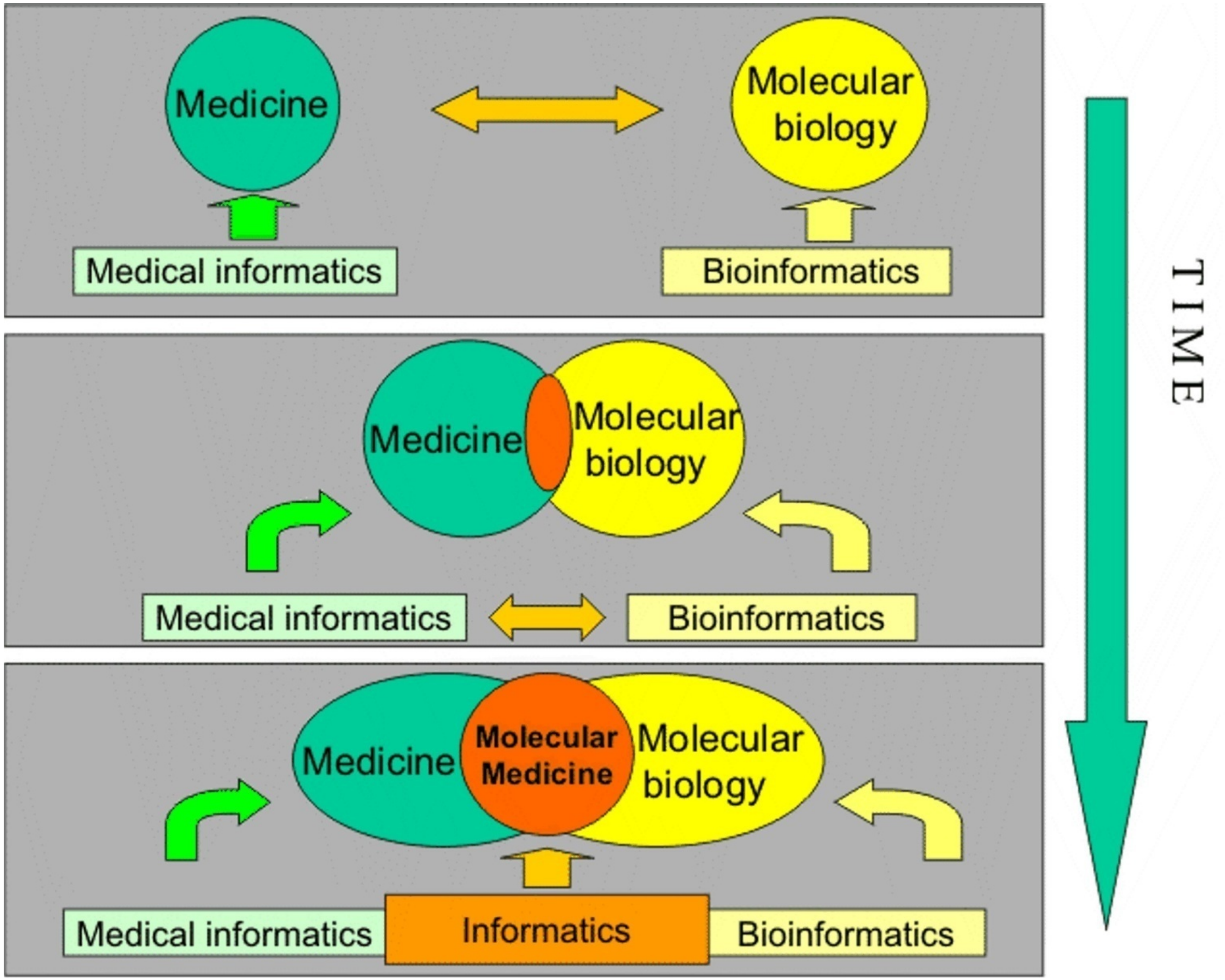




A Short History about Bioinformatics



The Convergence between MI & BI



Underlying molecules that response for **specific diseases**

Top Ten Medical Breakthroughs – since 1840

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

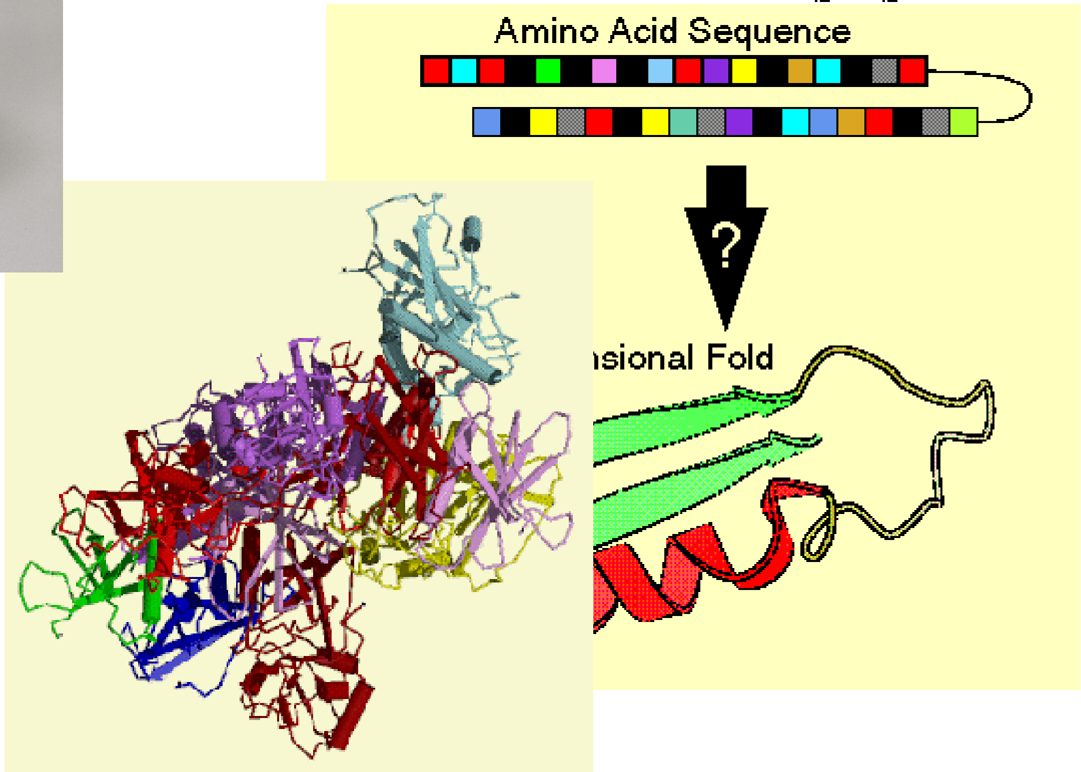
根據British Medical Journal 線上意見調查，自1840年創刊以來，最重要的醫學里程碑

Day 4: Computer Science and Medicine

CSedweek 11 部影片 訂閱

The Holy Grail of Bioinformatics

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN



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

A Short History Overview (I) - Wet

1953: Double helix of DNA (Waston & Crick)

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

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

1972: First DNA sequencing

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

1986: PCR (the photocopying machine of the biologist)

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

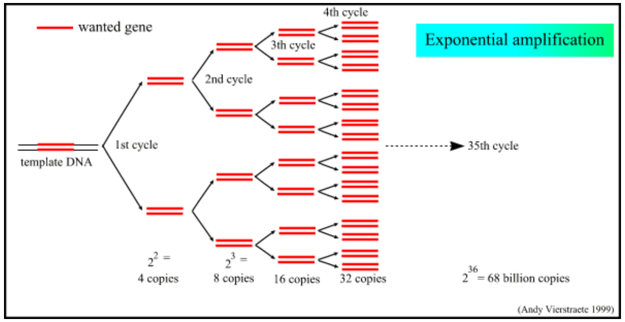
1995: Sequence of the genome of the bacteria: ***Haemophilus influenzae*** (2 *10^6 bp)

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

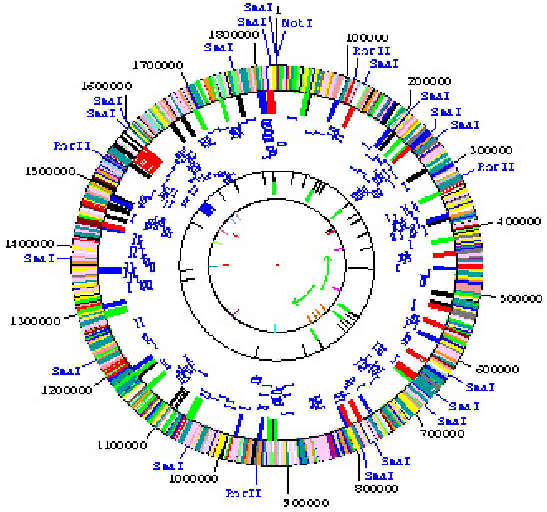
2000: Blue draft of the **human genome** (3*10^9 bp)

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

Recent: GOLD database



(Andy Vierstraete 1999)



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.)

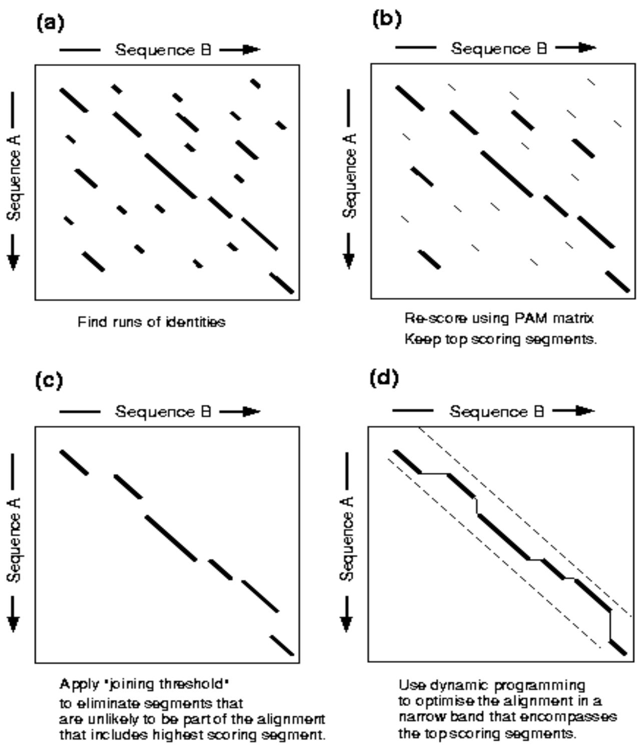
1986: Swiss-Prot (Protein Sequence Database)

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

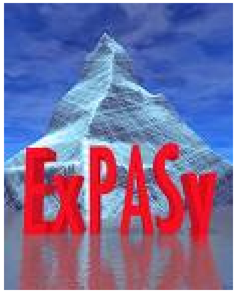


(a) Find runs of identities

(b) Re-score using PAM matrix
Keep top scoring segments.

(c) Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.

(d) Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.



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

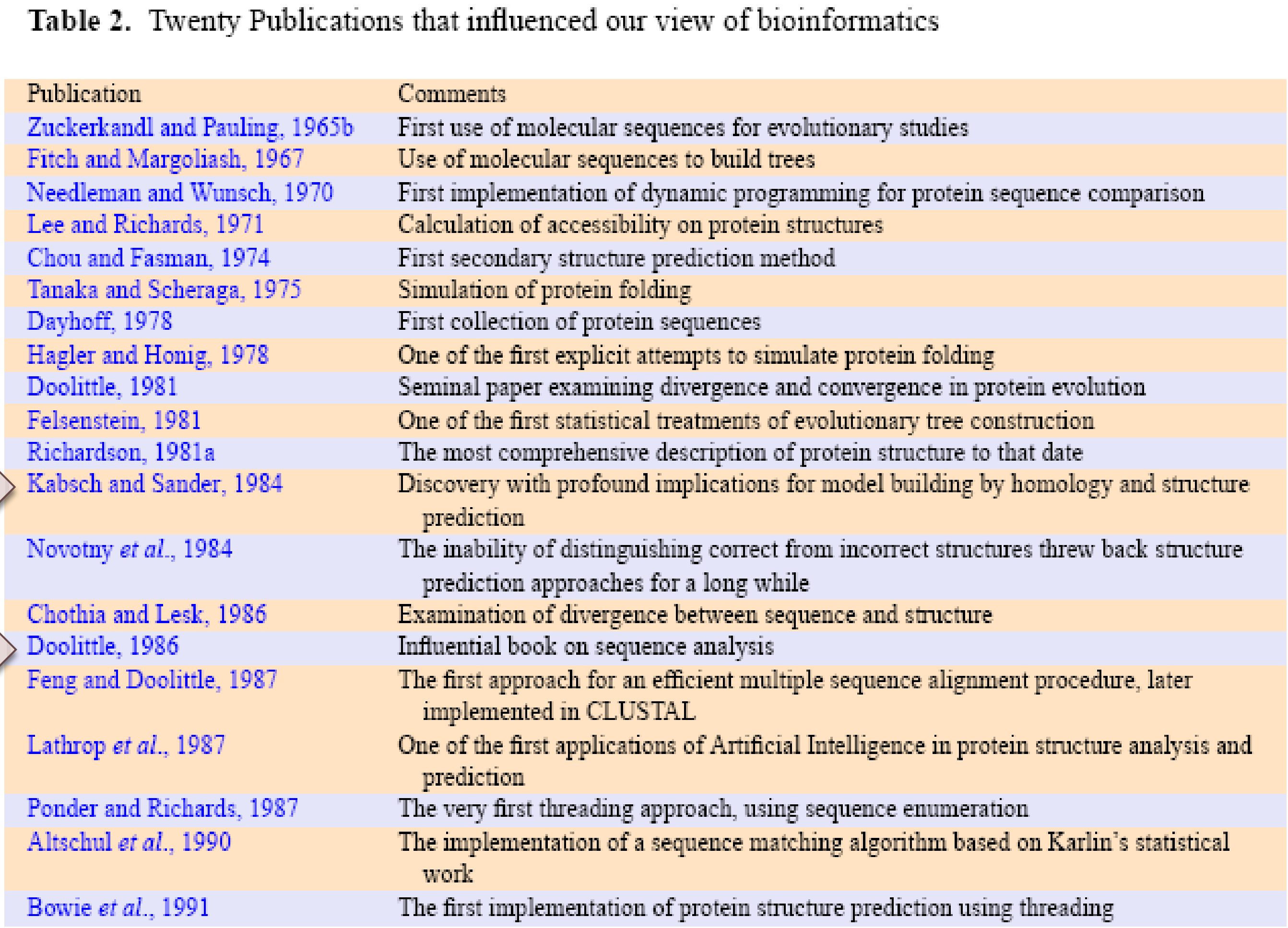


Table 2. Twenty Publications that influenced our view of bioinformatics

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

Bioinformatics: A Snapshot 10 Years Ago

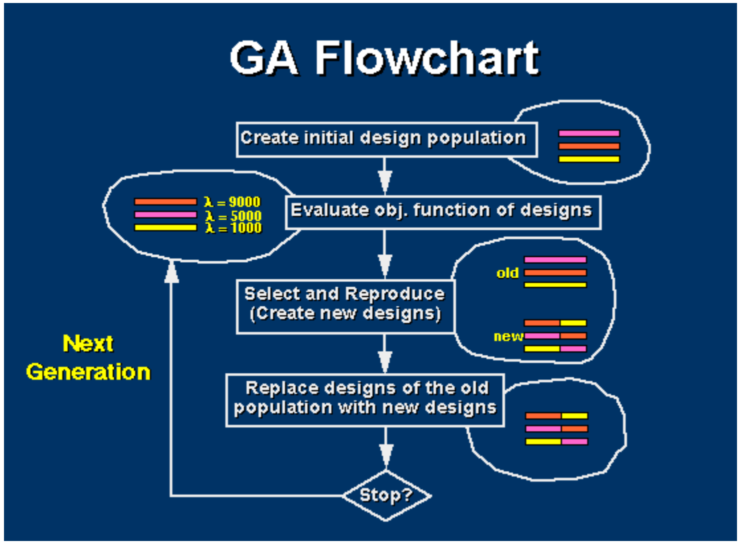
Pharmaceutical companies were **not interested**

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

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

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

DNA computers…



GA Flowchart

Create initial design population

Evaluate obj. function of designs

$\lambda = 9000$

$\lambda = 5000$

$\lambda = 1000$

Select and Reproduce (Create new designs)

old

new

Replace designs of the old population with new designs

Stop?

Next Generation

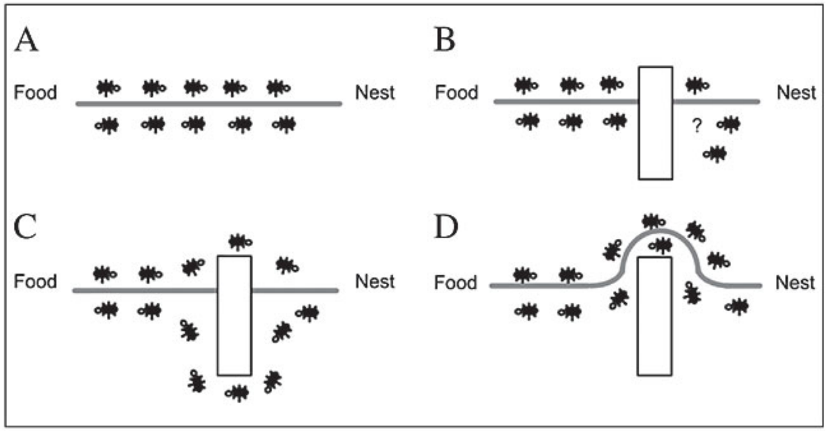


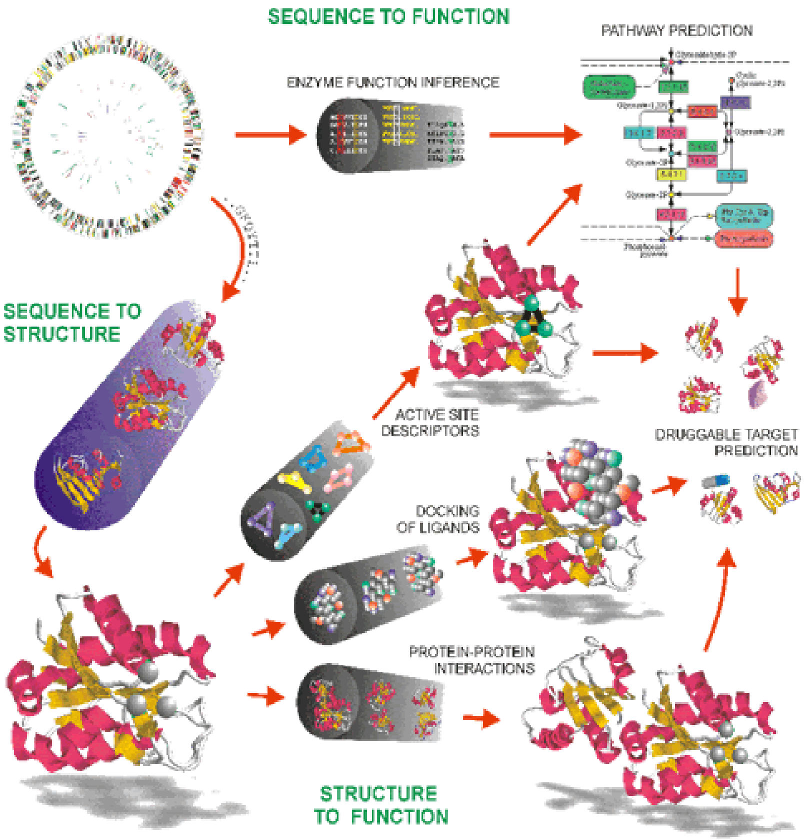
Figure 2. A. Ants in a pheromone trail between nest and food; **B.** an obstacle interrupts the trail; **C.** ants find two paths to go around the obstacle; **D.** a new pheromone trail is formed along the shorter path.

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

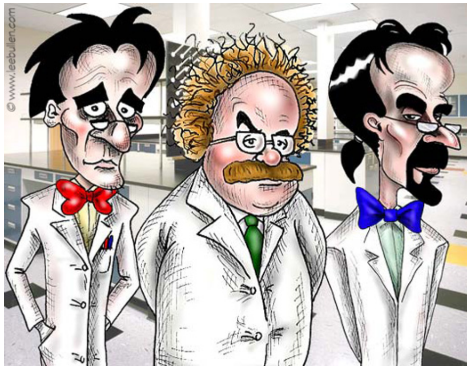


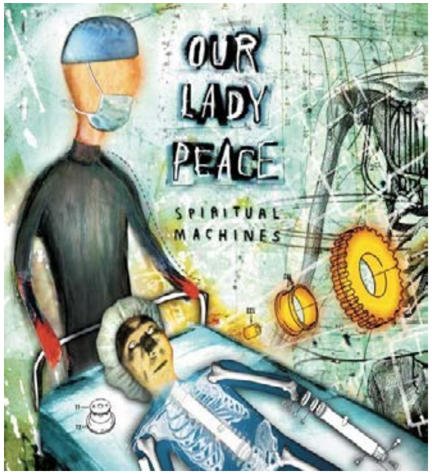
Bioinformatics in 2010

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**

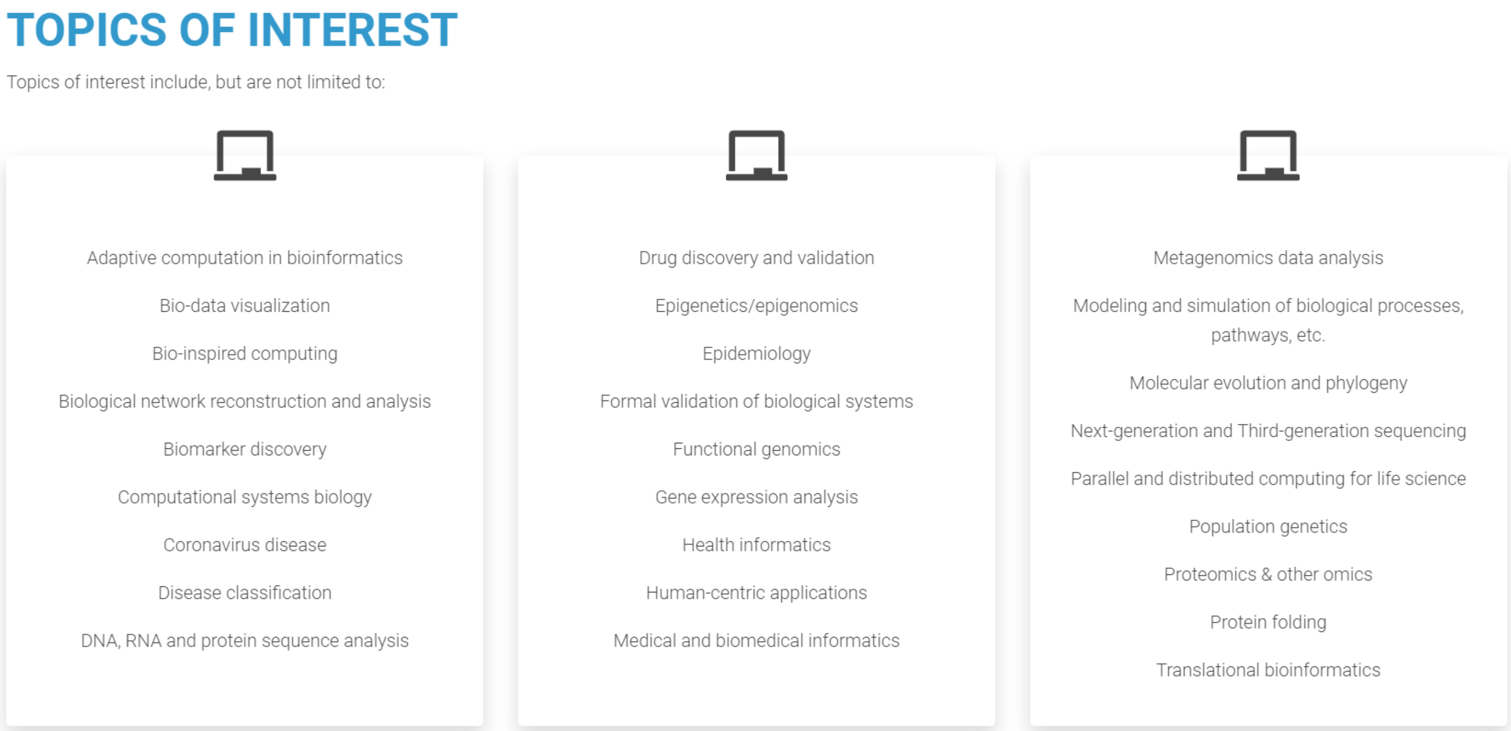
Computer scientists may have jumped on **another fancy subject:** Spiritual machines?





Bioinformatics in 2020

THEME: Innovation with AI and Cognitive Computing



TOPICS OF INTEREST

Topics of interest include, but are not limited to:

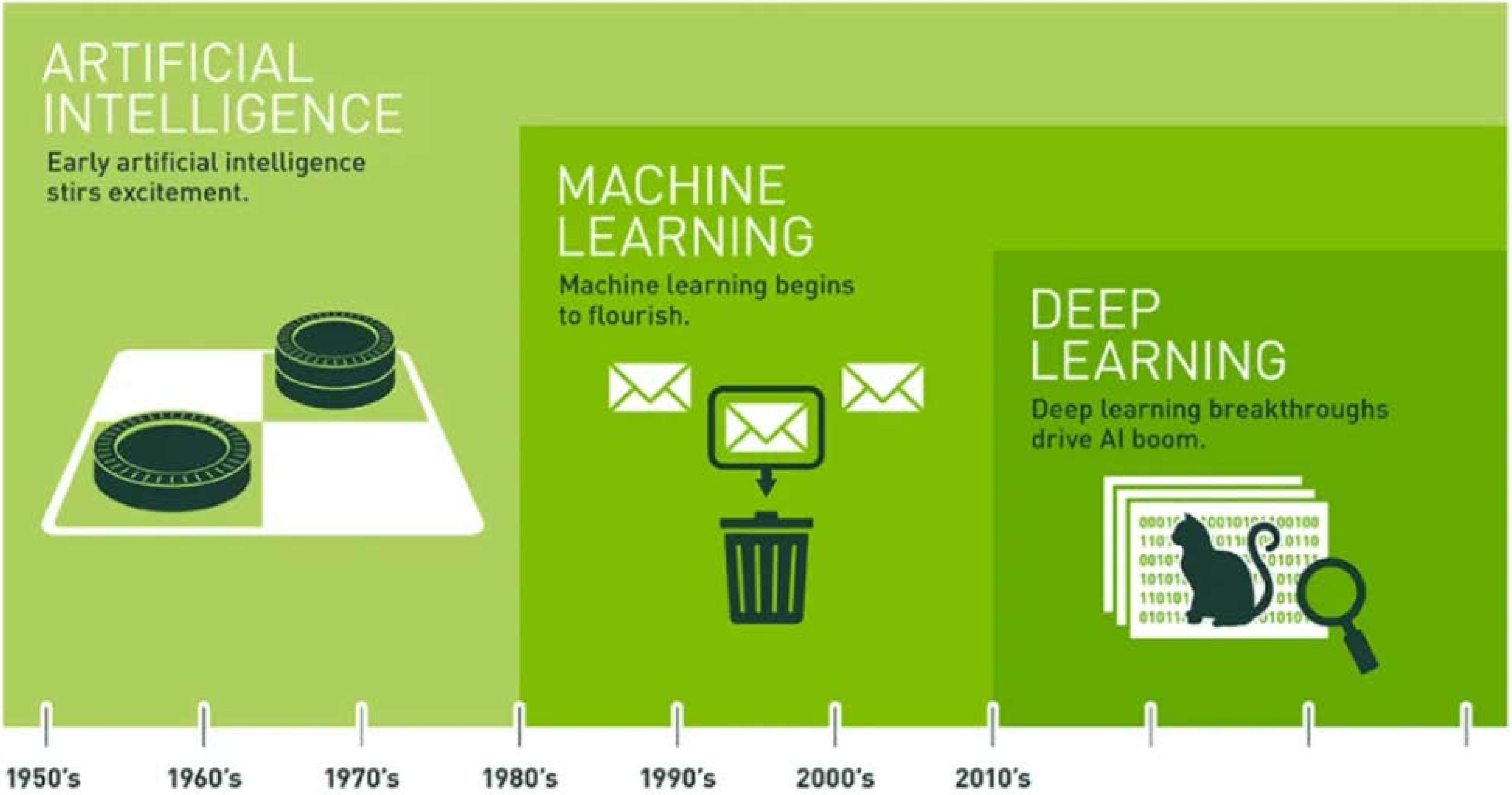
- Adaptive computation in bioinformatics
- Bio-data visualization
- Bio-inspired computing
- Biological network reconstruction and analysis
- Biomarker discovery
- Computational systems biology
- Coronavirus disease
- Disease classification
- DNA, RNA and protein sequence analysis

- Drug discovery and validation
- Epigenetics/epigenomics
- Epidemiology
- Formal validation of biological systems
- Functional genomics
- Gene expression analysis
- Health informatics
- Human-centric applications
- Medical and biomedical informatics

- Metagenomics data analysis
- Modeling and simulation of biological processes, pathways, etc.
- Molecular evolution and phylogeny
- Next-generation and Third-generation sequencing
- Parallel and distributed computing for life science
- Population genetics
- Proteomics & other omics
- Protein folding
- Translational bioinformatics

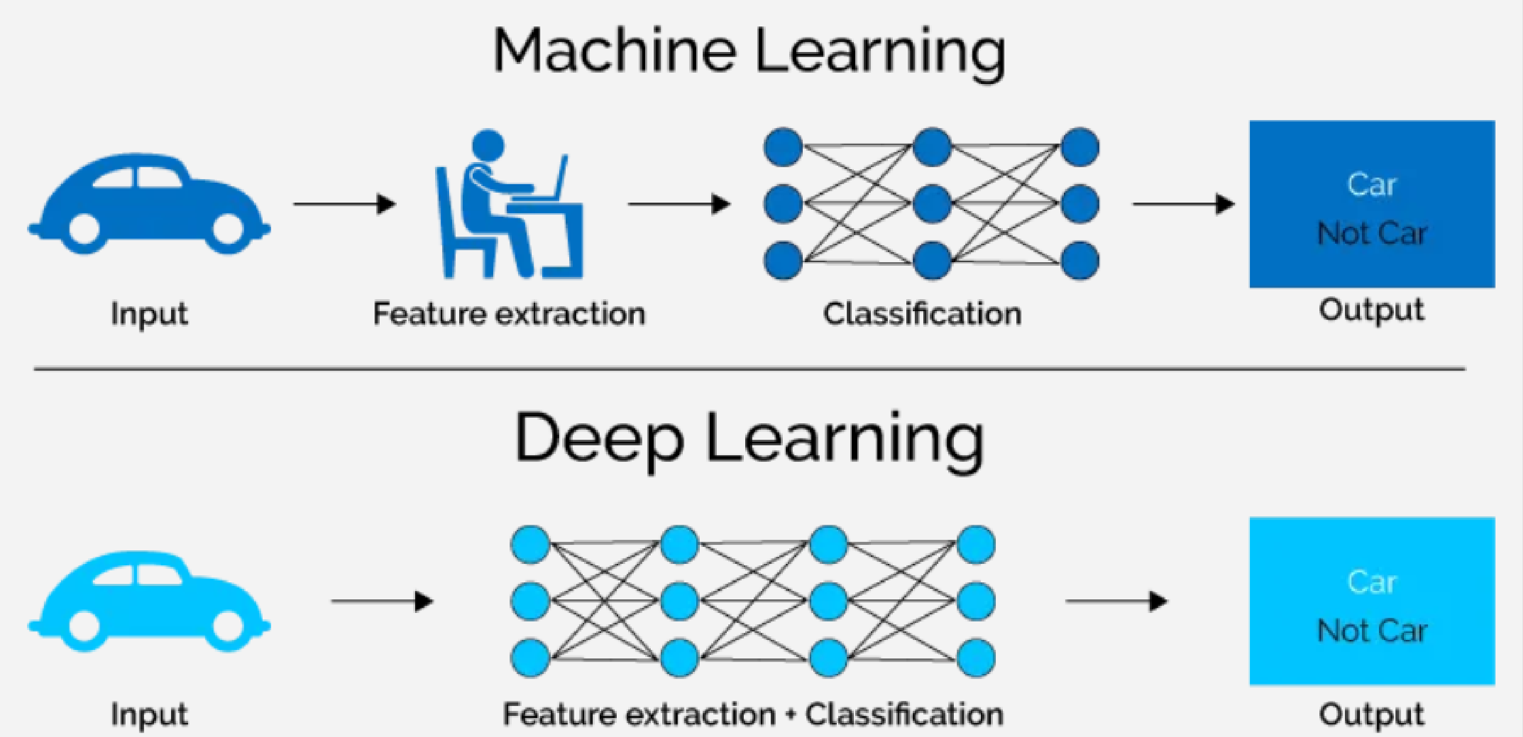
Artificial Intelligence

- 一般稱的 AI 其實是 Artificial Intelligence 的縮寫，而這個名字也清楚地表達了它的涵義。
  - 人工智慧的定義其實就是以「人工」編寫的電腦程式，去模擬出人類的「智慧」行為，其中包含模擬人類感官的「聽音辨讀、視覺辨識」、大腦的「推理決策、理解學習」、動作類的「移動、動作控制」等行為。

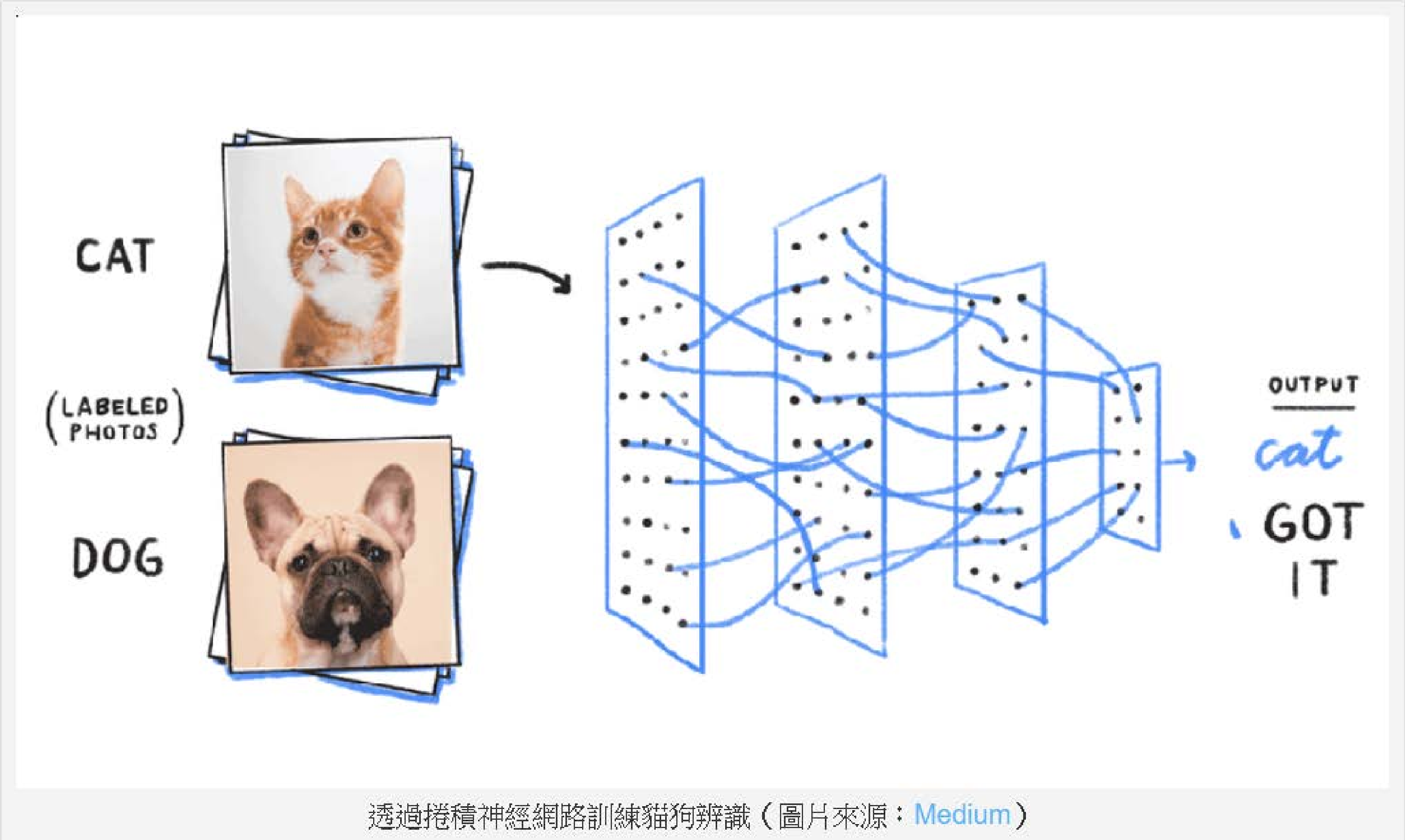

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

AI 演進（圖片來源：NVIDIA）

我們可以從上面這張圖清楚理解，AI、ML、DL 這三個名詞的關係就像洋蔥一樣層層遞進，機器學習（ML），是人工智慧（AI）底下的技術分支，而深度學習（DL）是近年才從機器學習衍伸出的領域，可以比喻為俄羅斯娃娃，一個子領域之中又有更深入的子領域。



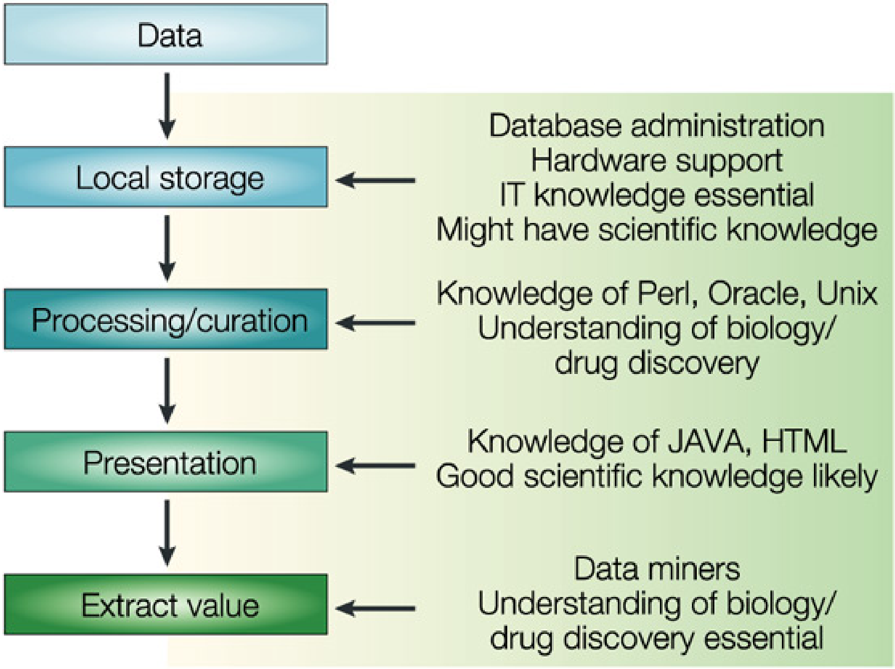
機器學習與深度學習的訓練比較：機器學習需要人為特徵分類（Feature extraction），而深度學習則是交給模型自己處理分類（圖片來源：XENONSTACK）



透過捲積神經網路訓練貓狗辨識（圖片來源：Medium）

Convolutional Neural Network, CNN

Resources: databases & software



Data

Local storage

Database administration
Hardware support
IT knowledge essential
Might have scientific knowledge

Processing/curation

Knowledge of Perl, Oracle, Unix
Understanding of biology/ drug discovery

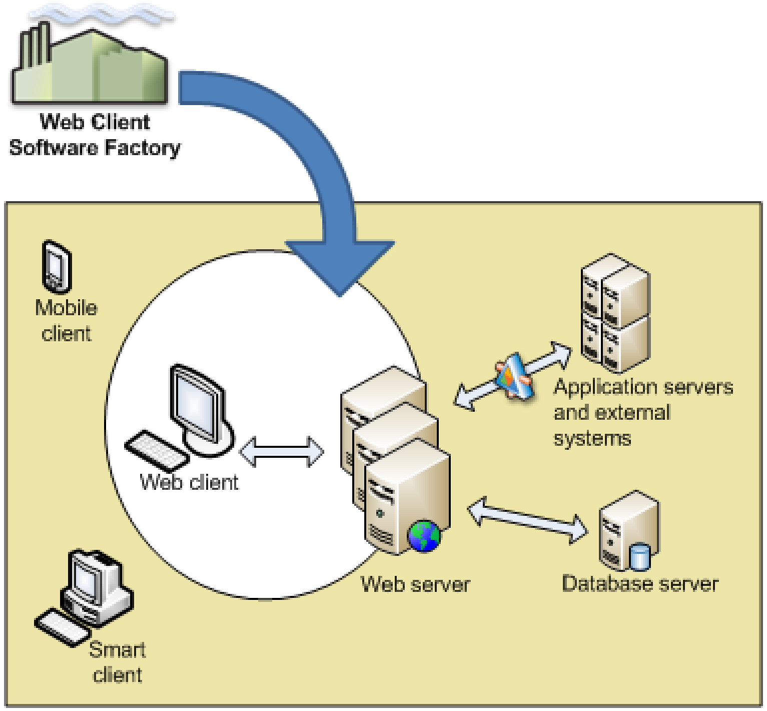
Presentation

Knowledge of JAVA, HTML
Good scientific knowledge likely

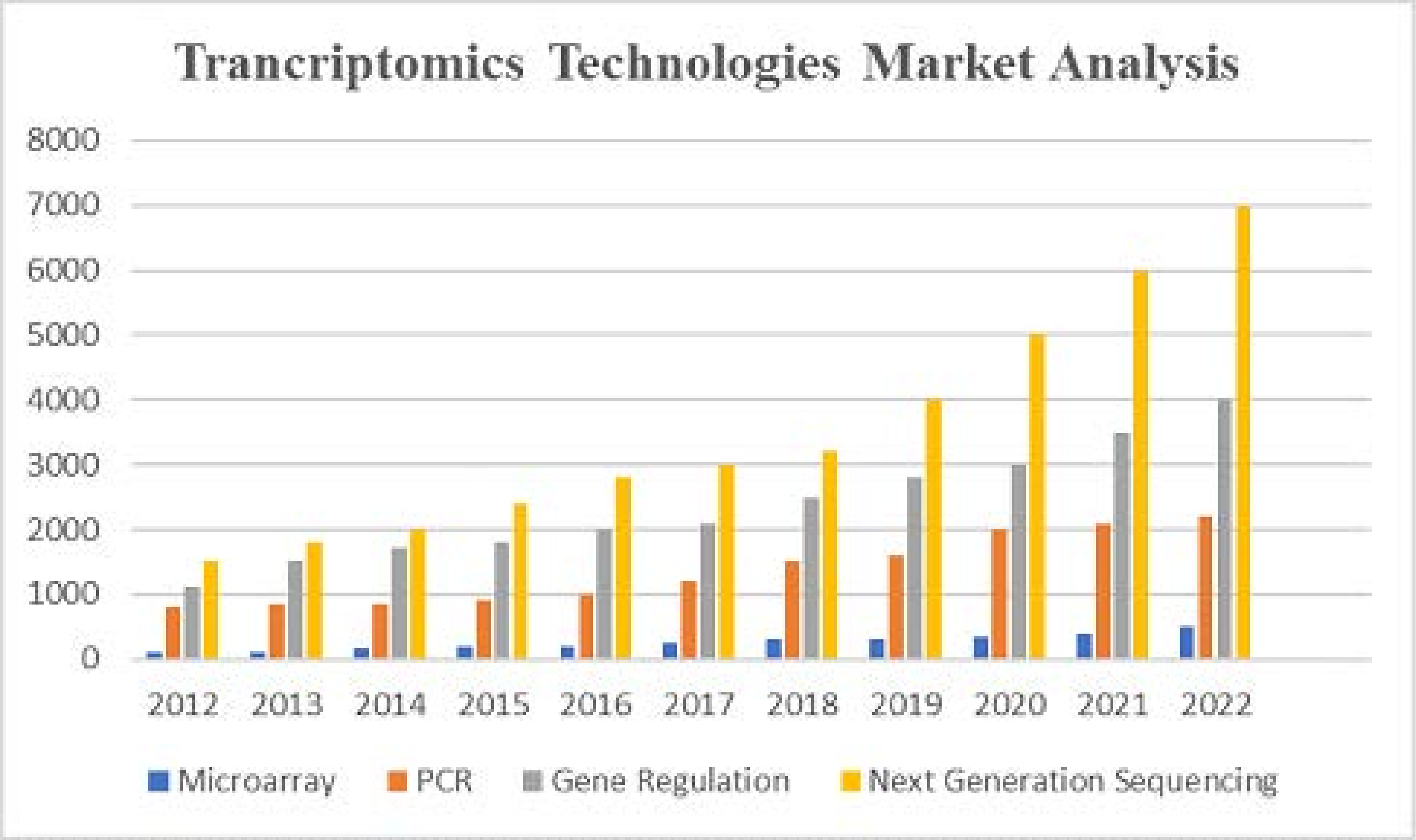
Extract value

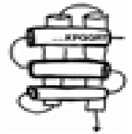
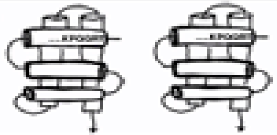
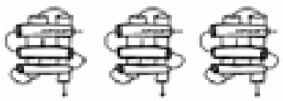
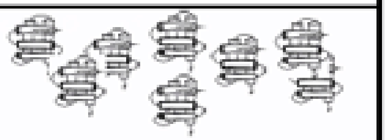
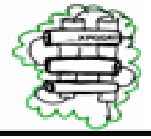
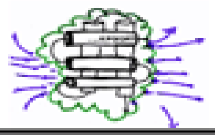
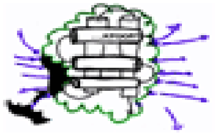
Data miners
Understanding of biology/ drug discovery essential

Nature Reviews | Drug Discovery



Nature Reviews Drug Discovery 3, 281 (2004)


Trancriptomics Technologies Market Analysis
8000
7000
6000
5000
4000
3000
2000
1000
0
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Microarray
PCR
Gene Regulation
Next Generation Sequencing

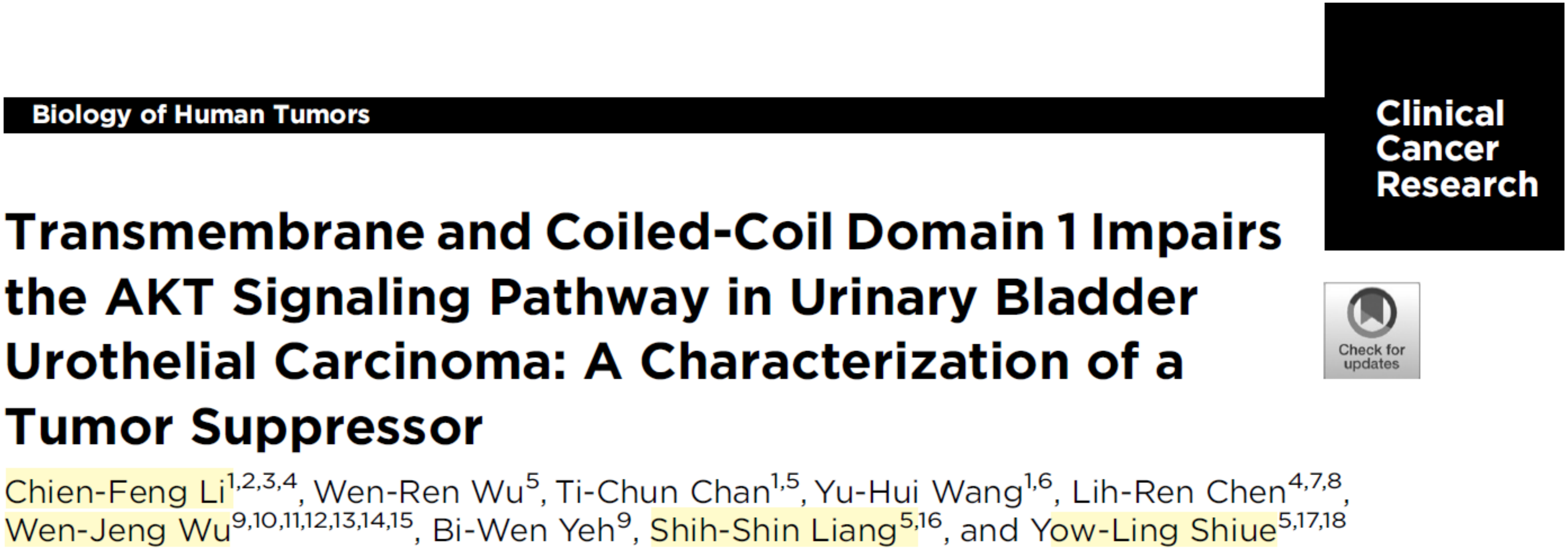
			Breadth: Homologs, Large-scale Surveys, Informatics—			
				pairwise comparison, sequence & structure alignment	multiple alignment, patterns, templates, trees	databases, scoring schemes, censuses
			1	2	3-100	100+
Depth: Rational Drug Design (physi		Genome Sequence	atcgatcgatatttgggatttggga	atcgatcgatatttgggatttggga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

COFFEE BREAK

"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



Biology of Human Tumors

Clinical Cancer Research

Transmembrane and Coiled-Coil Domain 1 Impairs the AKT Signaling Pathway in Urinary Bladder Urothelial Carcinoma: A Characterization of a Tumor Suppressor

Check for updates

Chien-Feng Li[1,2,3,4], Wen-Ren Wu[5], Ti-Chun Chan[1,5], Yu-Hui Wang[1,6], Lih-Ren Chen[4,7,8], Wen-Jeng Wu[9,10,11,12,13,14,15], Bi-Wen Yeh[9], Shih-Shin Liang[5,16], and Yow-Ling Shiue[5,17,18]

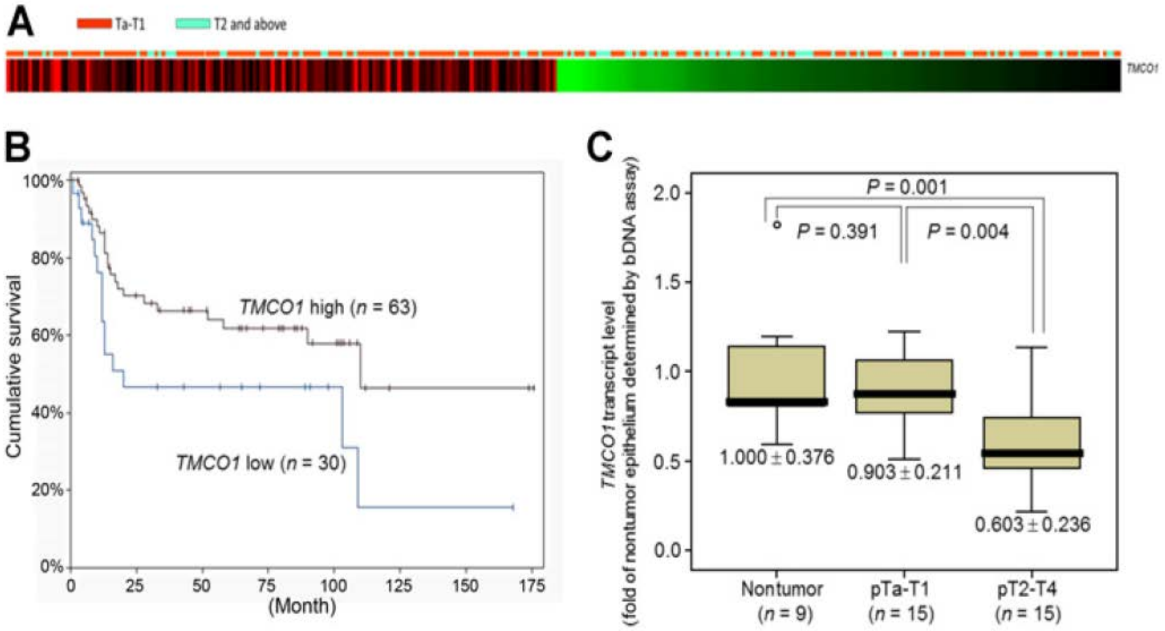
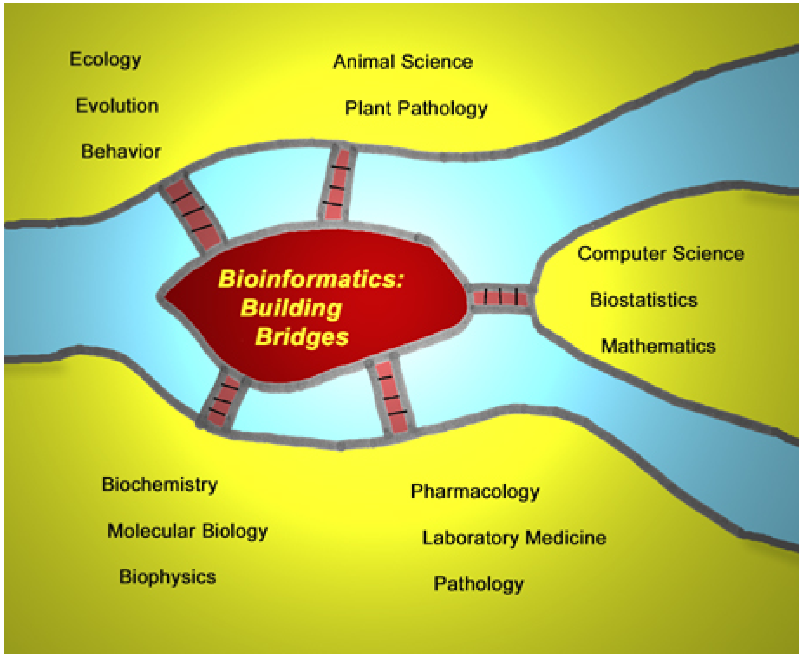


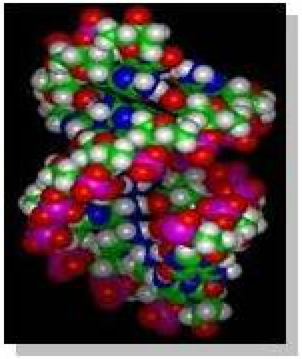
Figure 1.

Downregulation of the TMCO1 protein predicts poor disease-specific and metastasis-free survivals. **A,** A heatmap shows the data analysis from GSE32894 (GEO dataset), which identified that the *TMCO1* transcript is significantly downregulated ($P = 0.0009$) in muscle-invasive UBUC (blue bars). **B,** The downregulation of the *TMCO1* transcript was also predictive of poor overall survival in an independent dataset (GSE31684, GEO, NCBI; $P = 0.0425$). **C,** Quantitative RT-PCR

Q & A



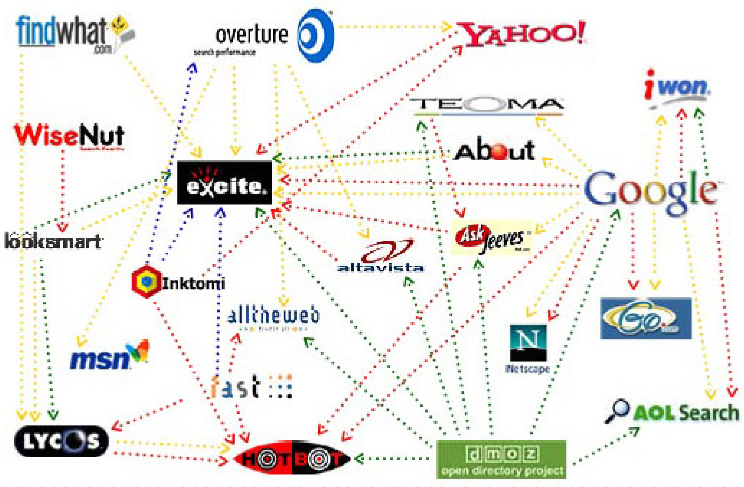
Computer systems

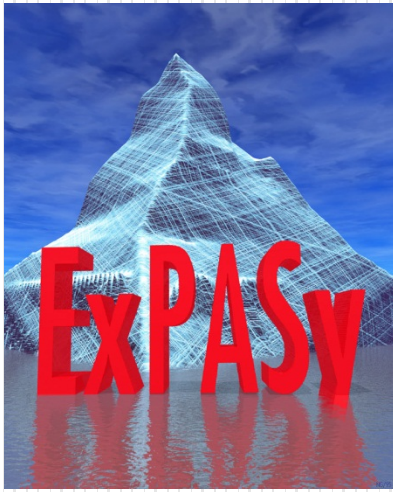


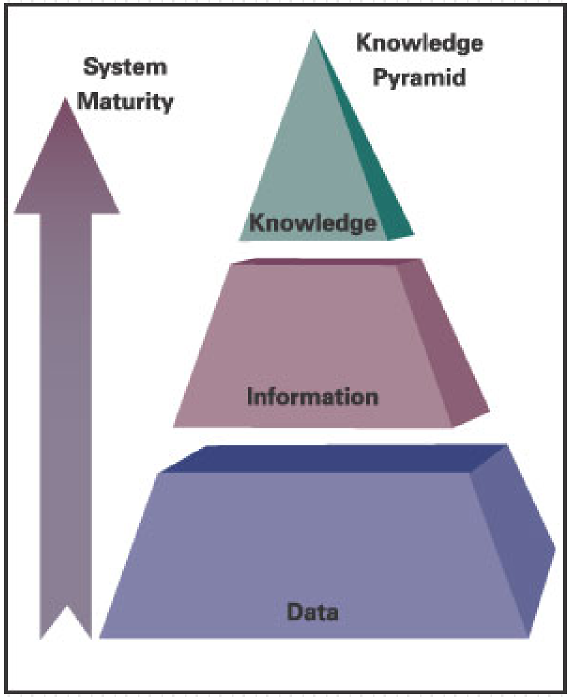
Biological systems

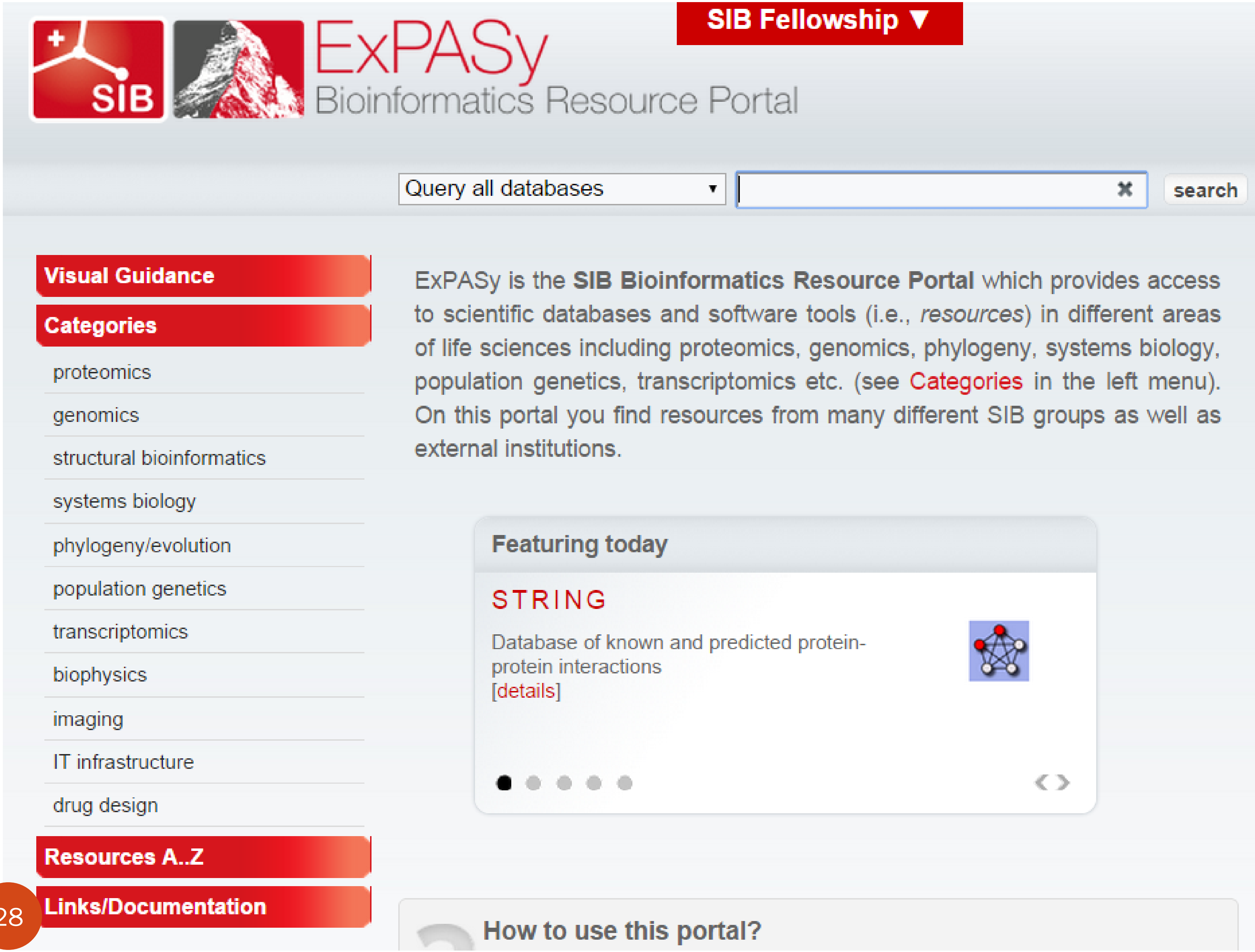
BIOINFORMATICS

Q: How to Find the Right Stuffs?









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]

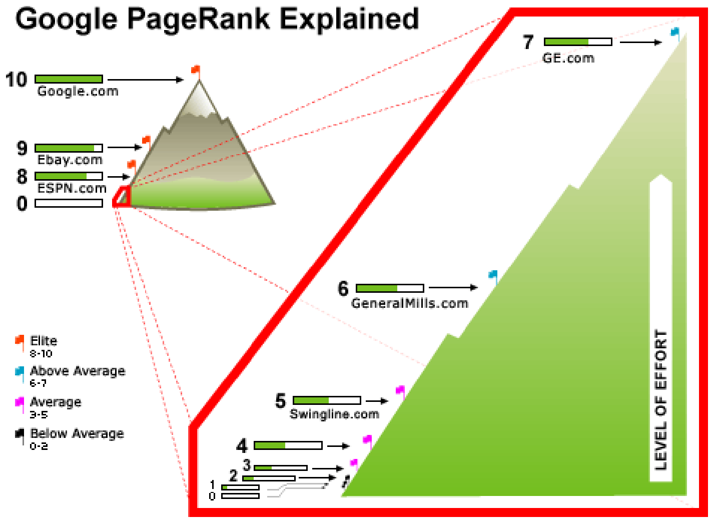
● ● ● ● ● ‹ ›

How to use this portal?

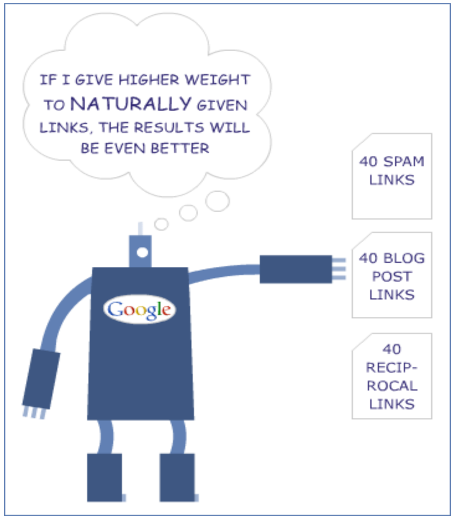
How to Find the Right Stuffs

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



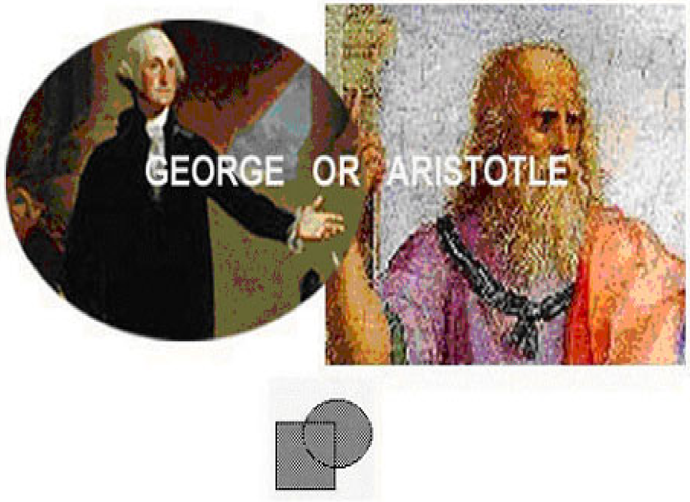
©2007 Elliance,Inc.

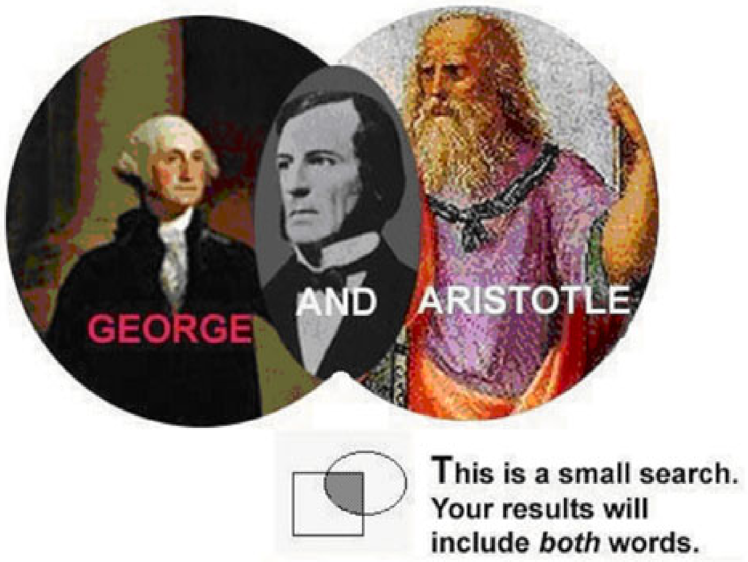


IF I GIVE HIGHER WEIGHT TO NATURALLY GIVEN LINKS, THE RESULTS WILL BE EVEN BETTER
40 SPAM LINKS
40 BLOG POST LINKS
40 RECIP-ROCAL LINKS
Google

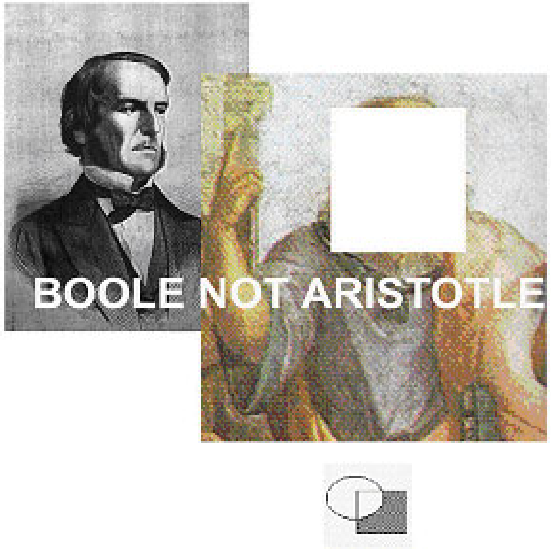
Search Efficiently

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

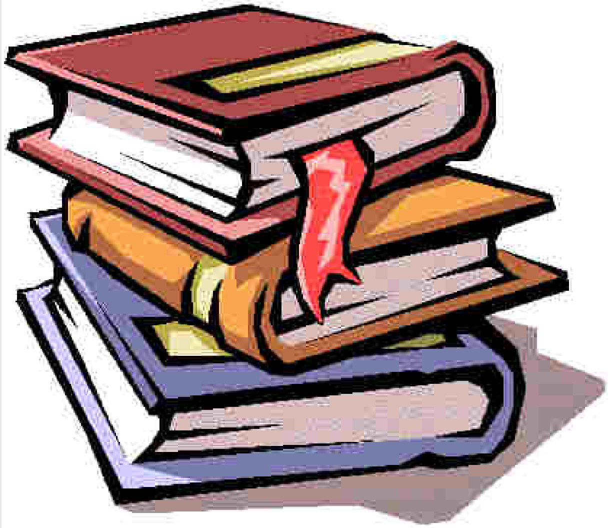




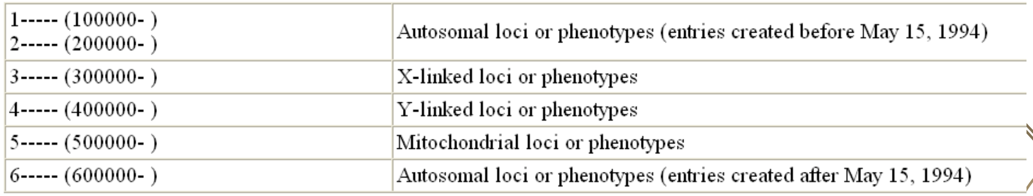
This is a small search. Your results will include *both* words.



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

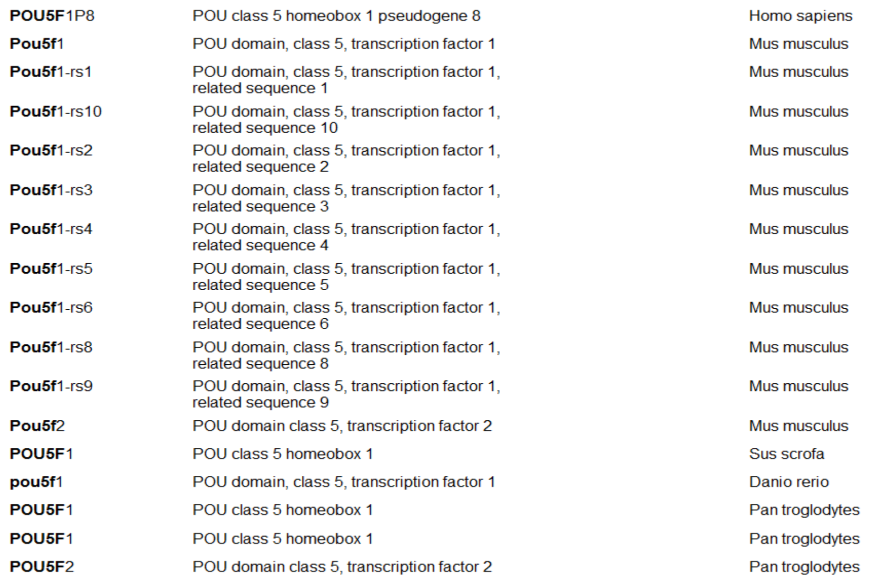


Gene or Disease – Official Symbol



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

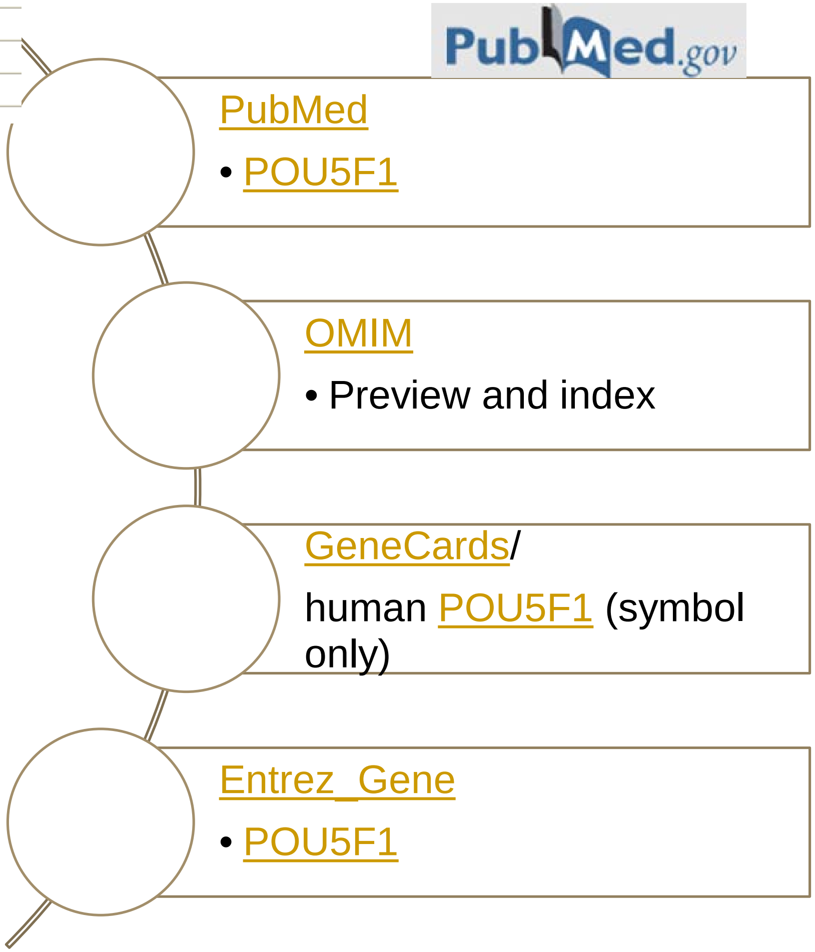
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University



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

GeneCards®
The Human Gene Compendium

PubMed.gov



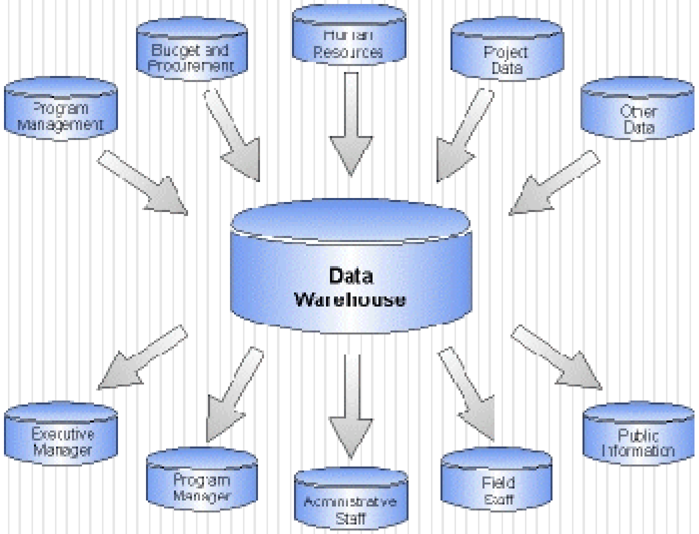
PubMed
- POU5F1

OMIM
- Preview and index

GeneCards/
human POU5F1 (symbol only)

Entrez_Gene
- POU5F1

Q: What is Derivative Databases?



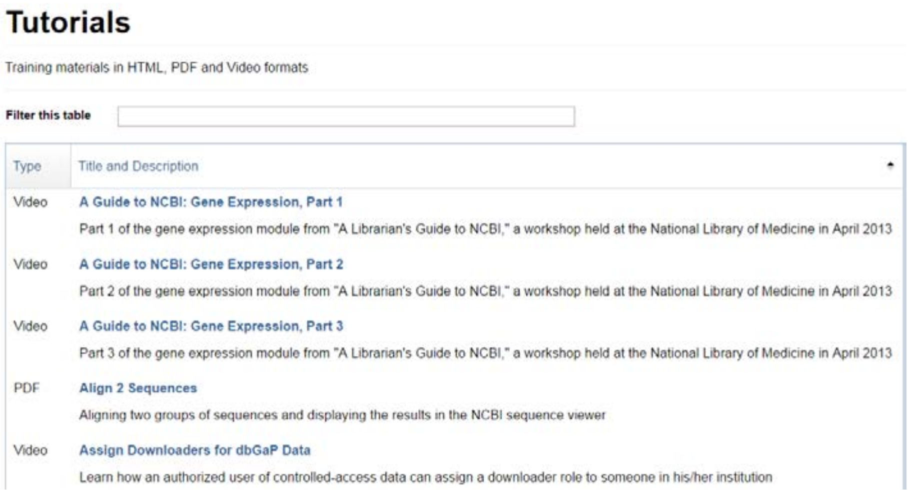
Leading Bioinformatic Centers

NCBI, USA

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

EBI, UK

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

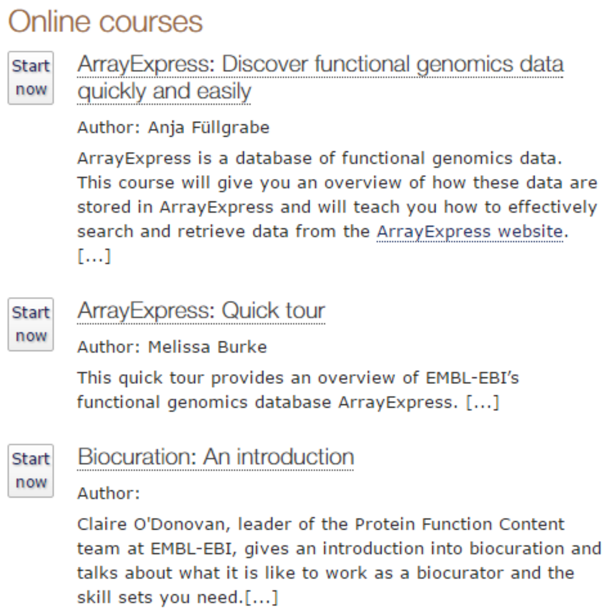


Tutorials

Training materials in HTML, PDF and Video formats

Filter this table

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



Online courses

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. [...]

ArrayExpress: Quick tour

Author: Melissa Burke

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

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**

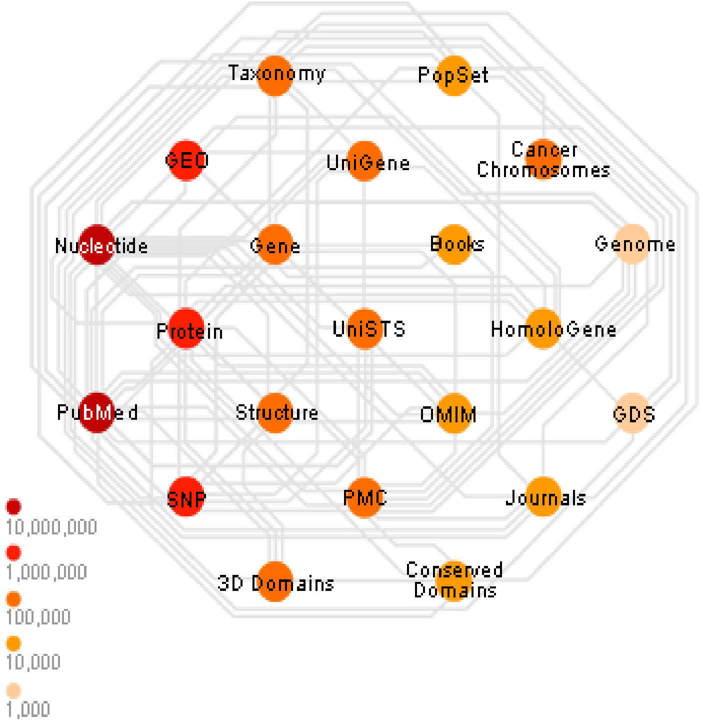
Contents

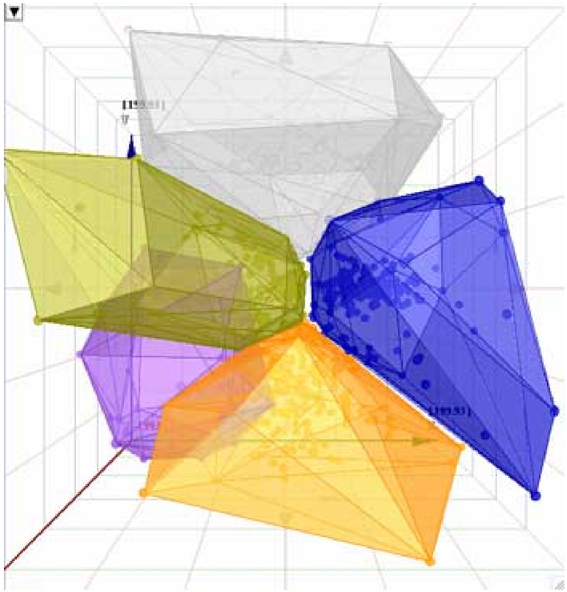
Databases

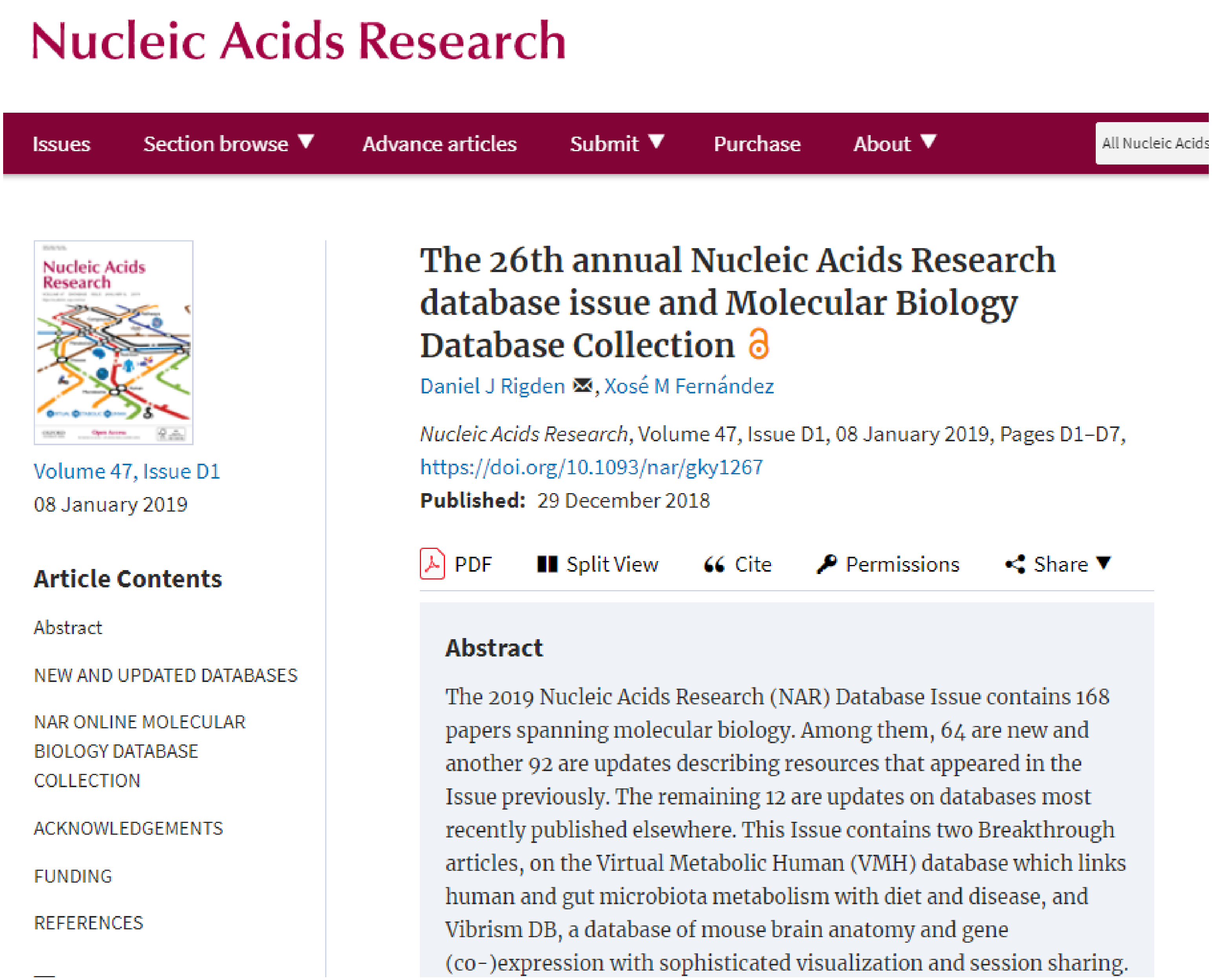
- **Primary** vs. **derivative** databases
  - **Value-added**

Methodologies (tools)

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







Volume 47, Issue D1
08 January 2019

Article Contents

The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection

Daniel J Rigden ✉, Xosé M Fernández

Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D1–D7, https://doi.org/10.1093/nar/gky1267

Published: 29 December 2018

Abstract

The 2019 Nucleic Acids Research (NAR) Database Issue contains 168 papers spanning molecular biology. Among them, 64 are new and another 92 are updates describing resources that appeared in the Issue previously. The remaining 12 are updates on databases most recently published elsewhere. This Issue contains two Breakthrough articles, on the Virtual Metabolic Human (VMH) database which links human and gut microbiota metabolism with diet and disease, and Vibrism DB, a database of mouse brain anatomy and gene (co-)expression with sophisticated visualization and session sharing.

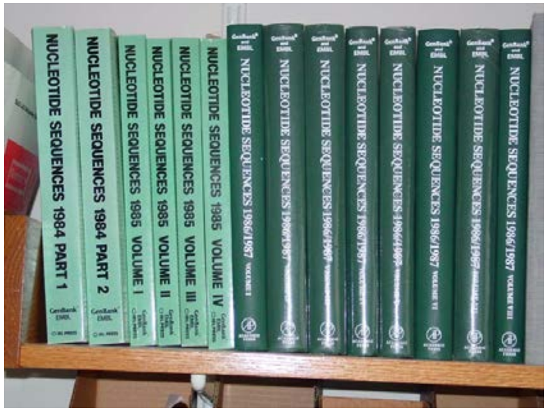
Primary vs. Derivative Databases - NCBI

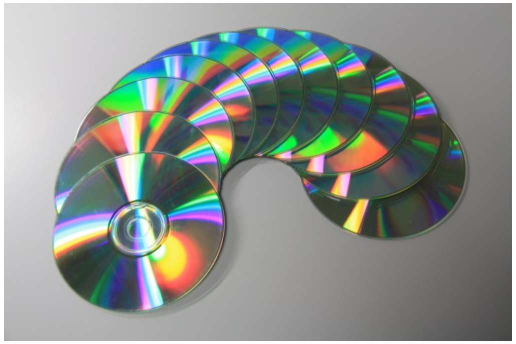
Primary databases

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

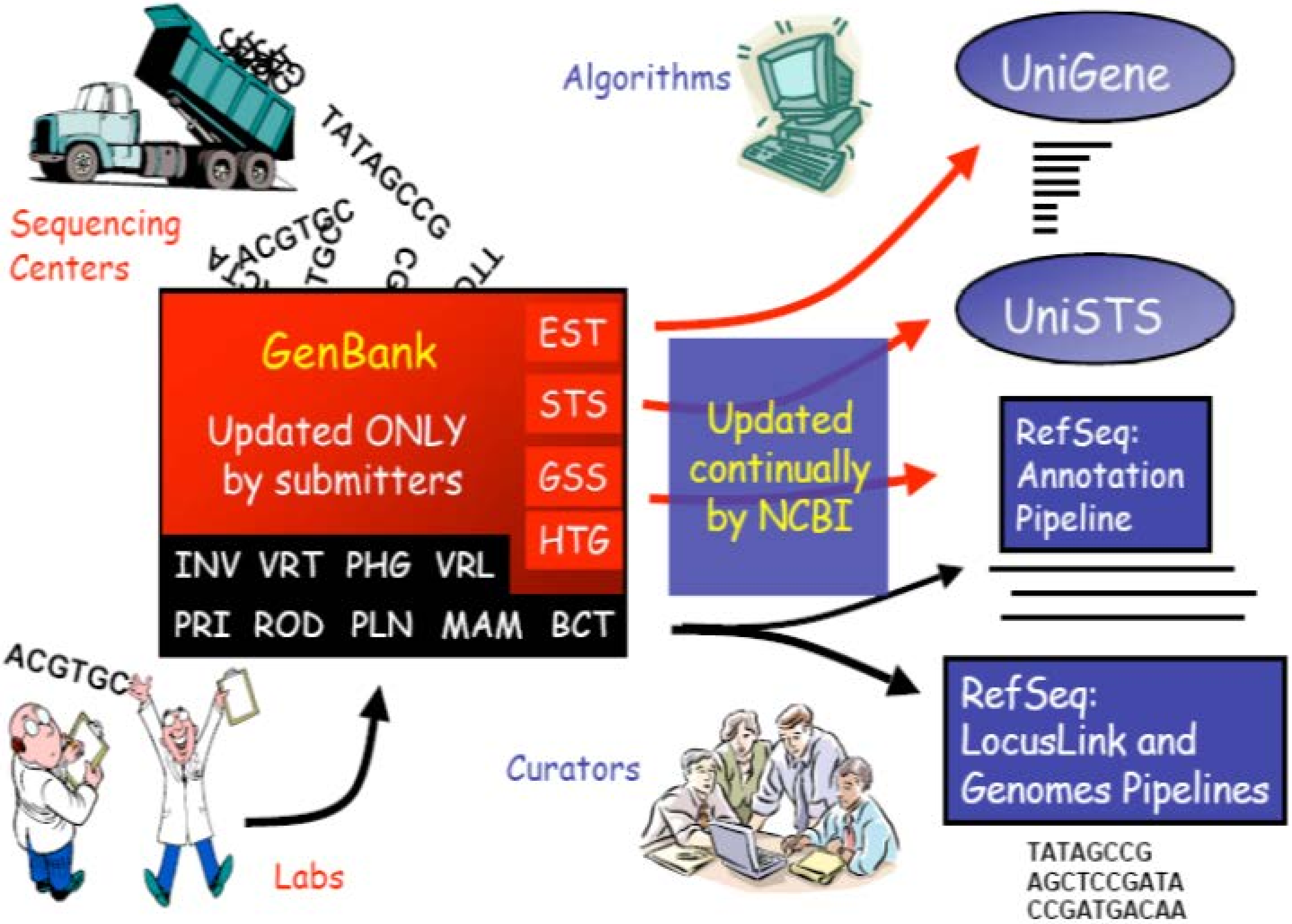
Derivative databases

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


NUCLEOTIDE SEQUENCES 1984 PART 1
NUCLEOTIDE SEQUENCES 1984 PART 2
NUCLEOTIDE SEQUENCES 1985 VOLUME I
NUCLEOTIDE SEQUENCES 1985 VOLUME II
NUCLEOTIDE SEQUENCES 1985 VOLUME III
NUCLEOTIDE SEQUENCES 1985 VOLUME IV
NUCLEOTIDE SEQUENCES 1986/1987



Primary vs. Derivative Databases - NCBI



Primary DNA Databases

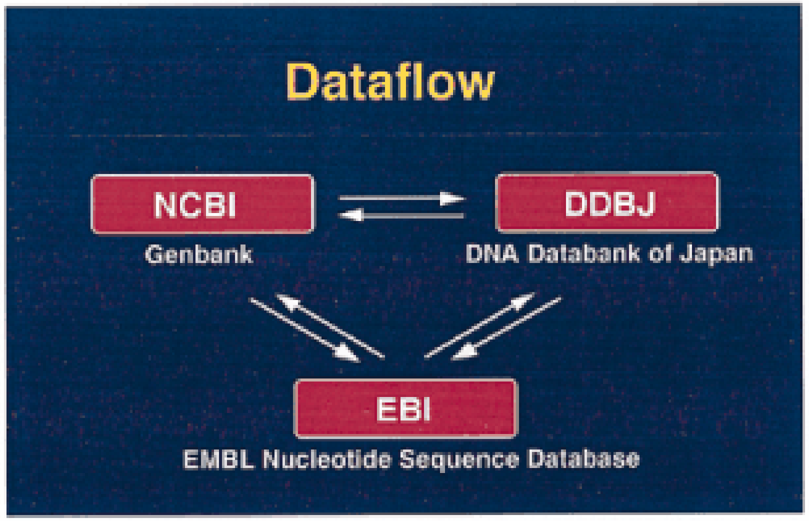
DDBJ Japan
INSDC IAM ICM
GenBank USA
EMBL Europe

NCBI

EMBL NUCLEOTIDE SEQUENCE DATABASE

DDBJ

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



Dataflow

NCBI
Genbank

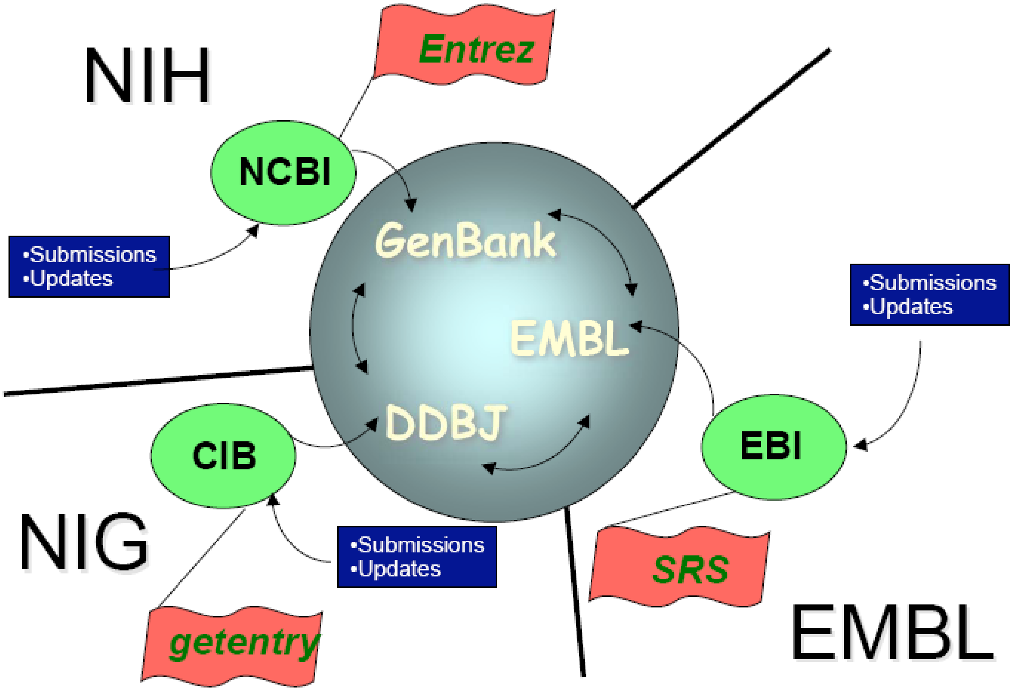
DDBJ
DNA Databank of Japan

EBI
EMBL Nucleotide Sequence Database

National Institute of Health (**NIH**)

National Center for Biotechnology (**NCBI**)

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



National Institute of Genetics (NIG)

Center for Information Biology (CIB)

Research Organization of Information and Systems
National Institute of Genetics

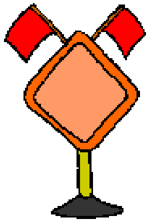
The European Bioinformatics Institute (**EBI**)

Sequence Retrieval System (**SRS**)

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

EMBL/GenBank/DDBJ Annotations

Warning!!!



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

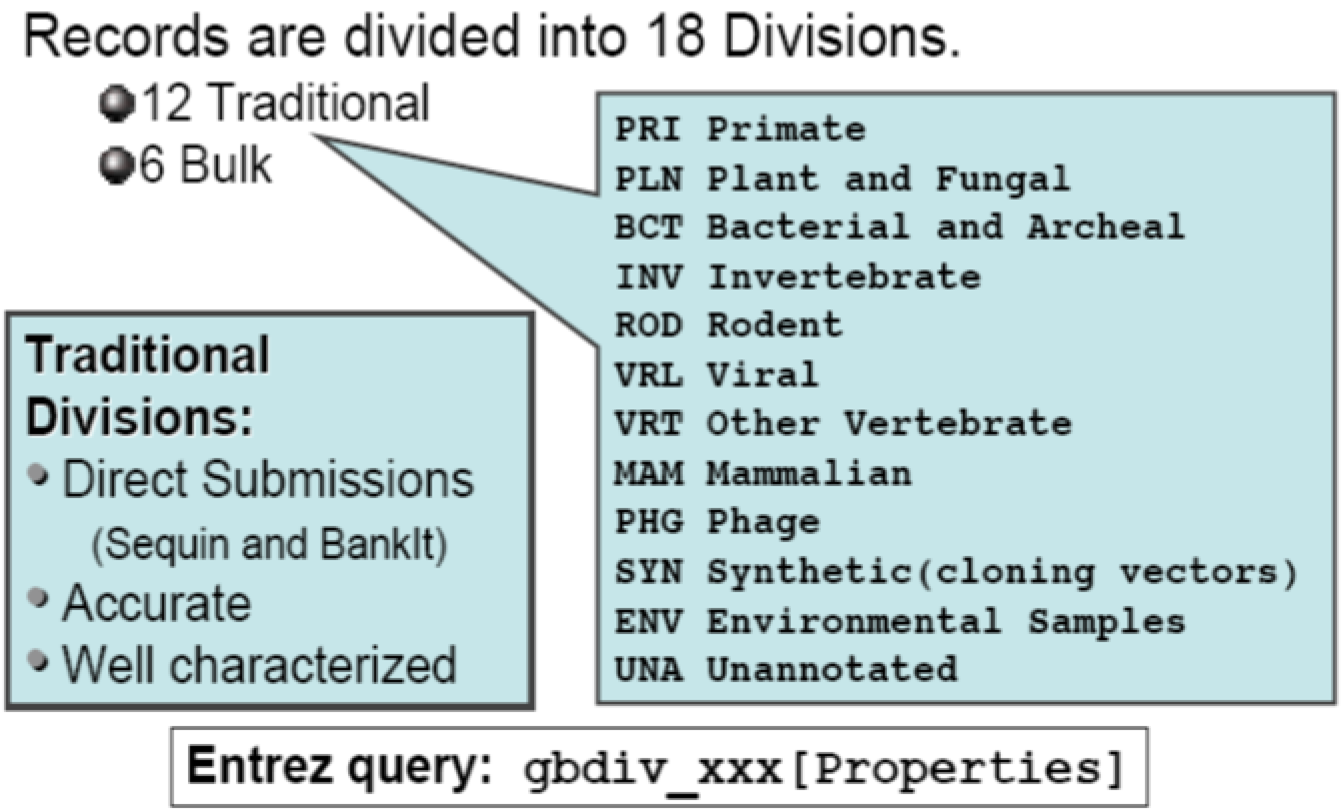
Some Interesting Sequence Annotation

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

Or:

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

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]`

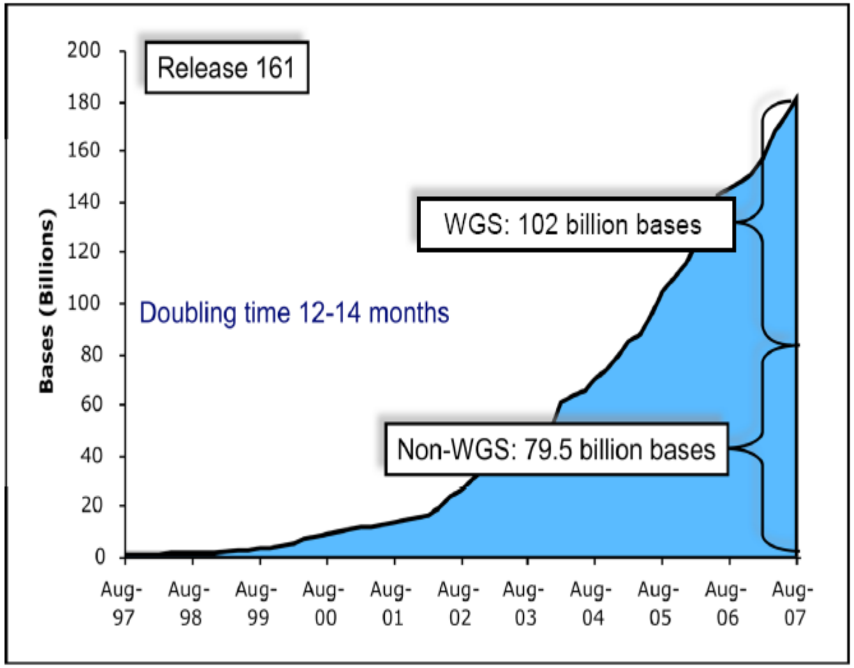
Bulk GenBank Divisions

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

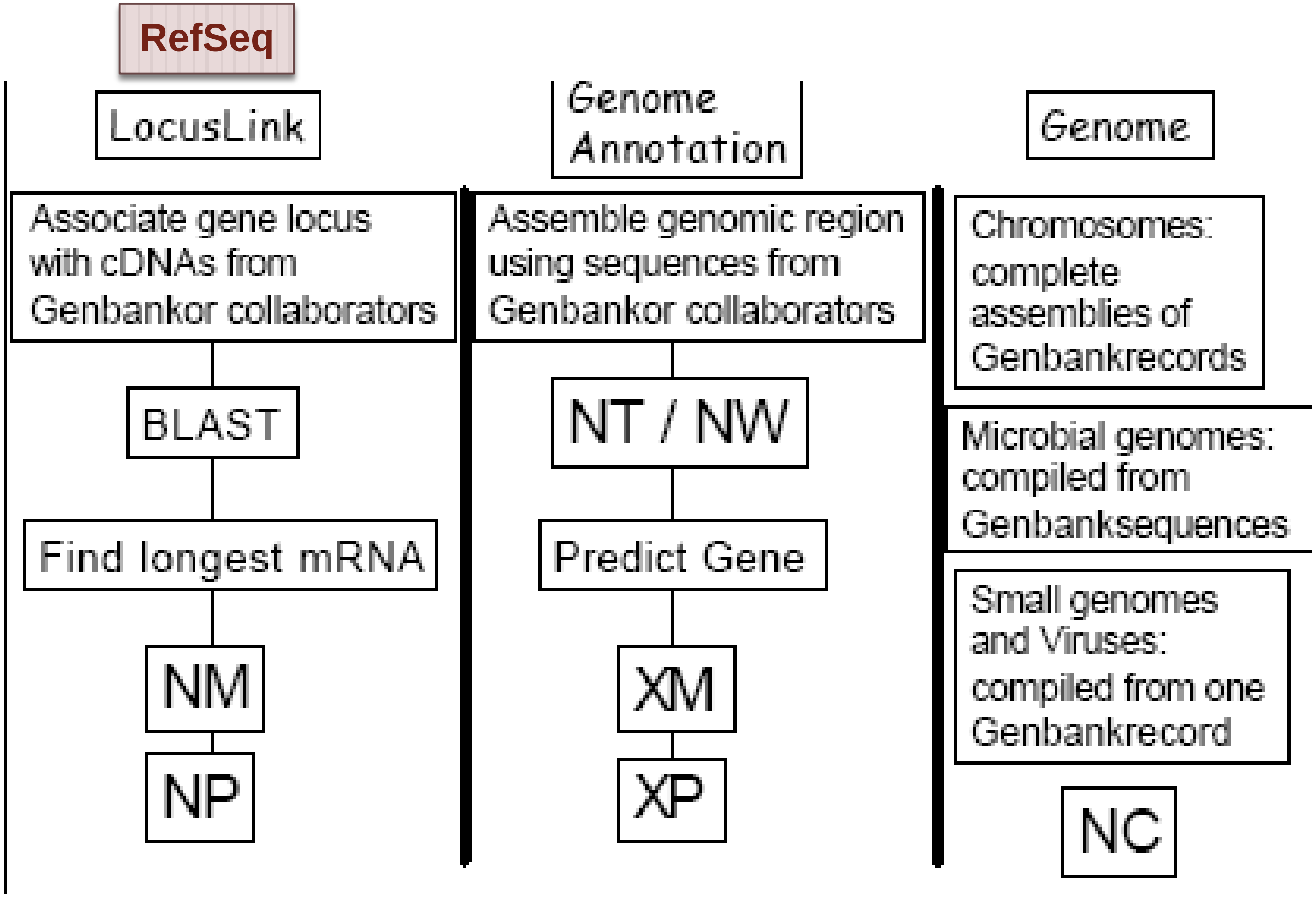
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTG	High Throughput Genomic
STS	Sequence Tagged Site
HTC	High Throughput cDNA
PAT	Patent

BULK Divisions:

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

Entrez query: `gbdiv_xxx[Properties]`

RefSeq Pipelines



RefSeq

LocusLink

Associate gene locus with cDNAs from Genbankor collaborators

BLAST

Find longest mRNA

NM

NP

Genome Annotation

Assemble genomic region using sequences from Genbankor collaborators

NT / NW

Predict Gene

XM

XP

Genome

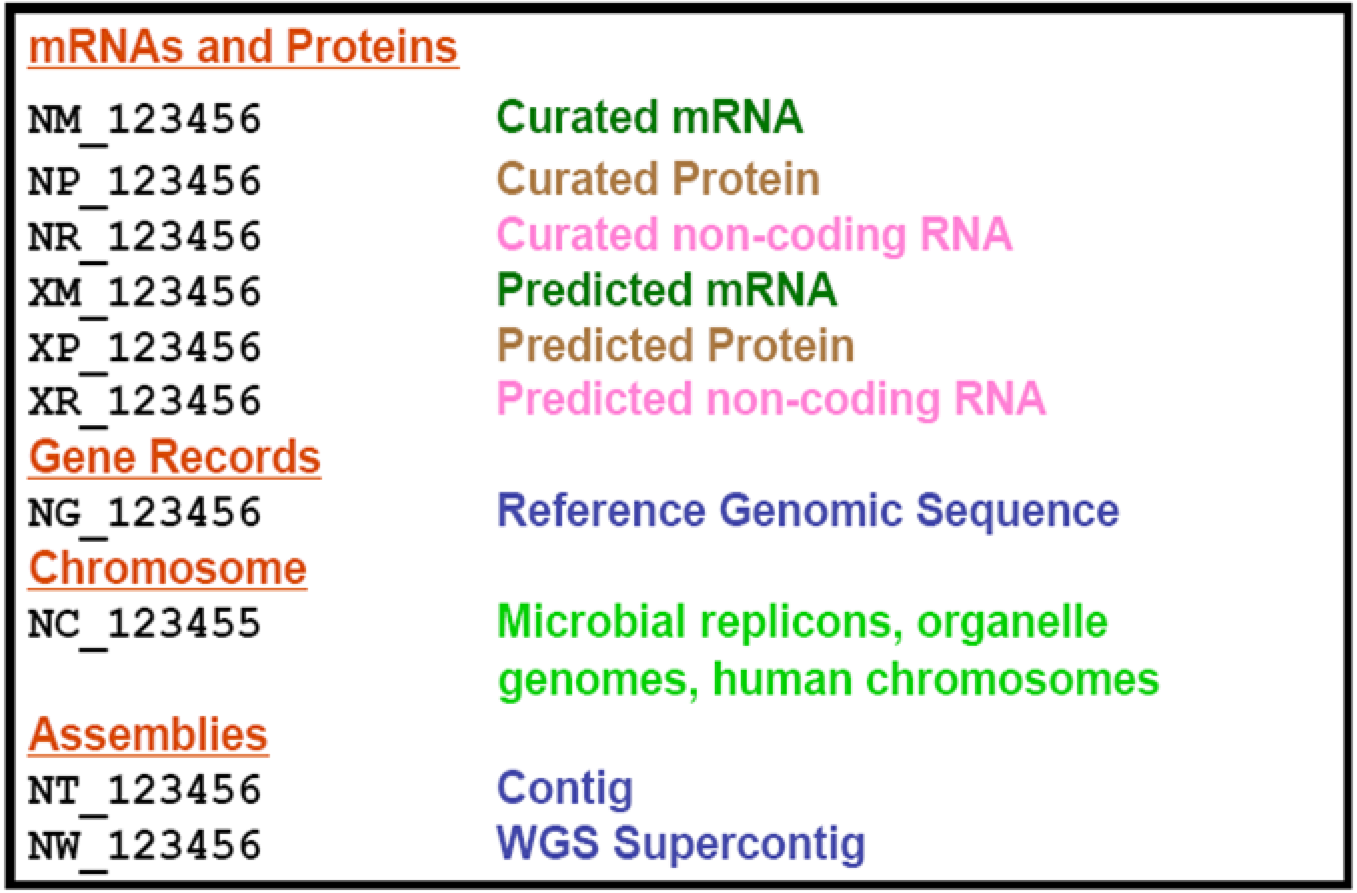
Chromosomes: complete assemblies of Genbankrecords

Microbial genomes: compiled from Genbanksequences

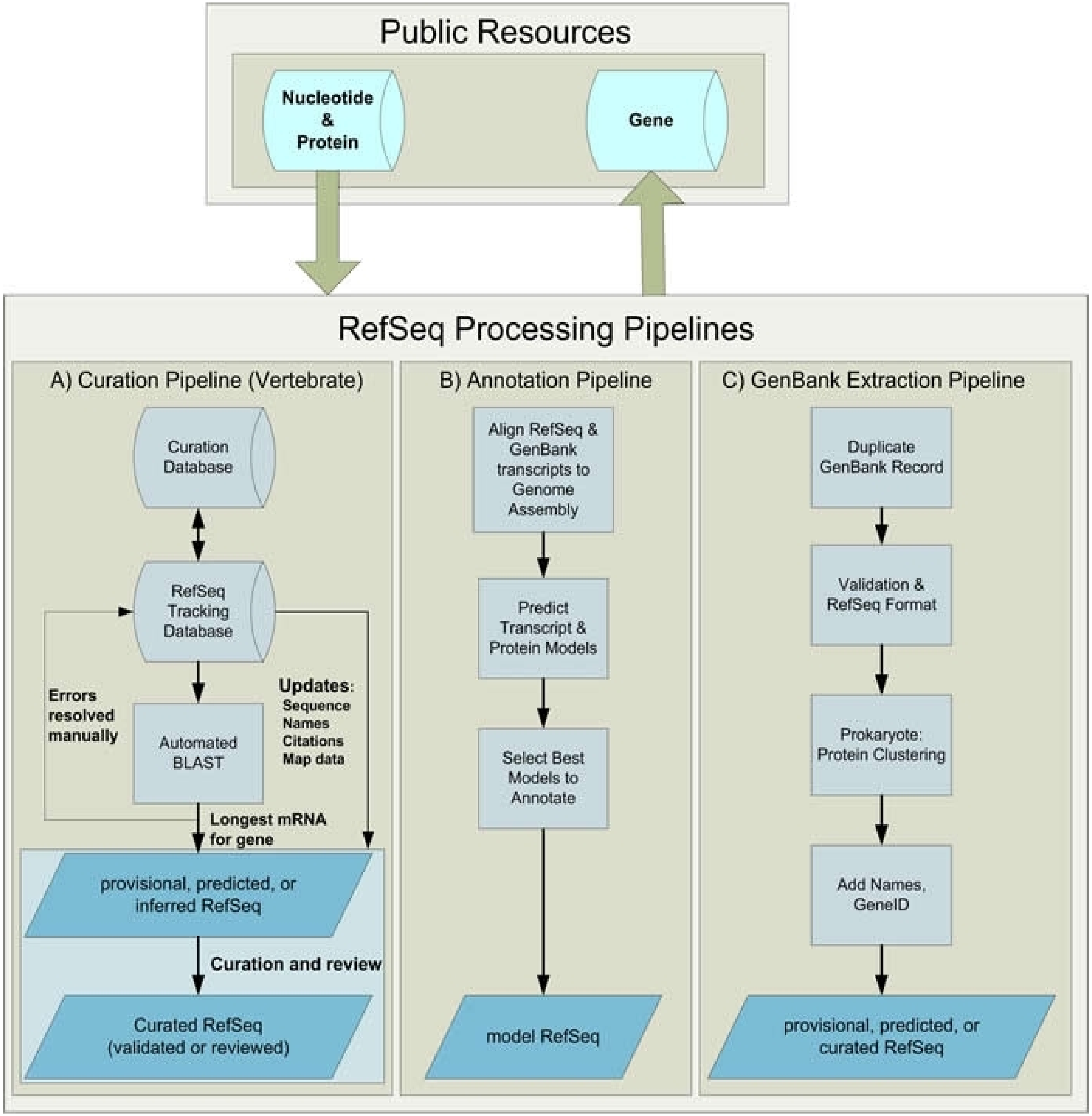
Small genomes and Viruses: compiled from one Genbankrecord

NC

Selected RefSeq Accession Number



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



Public Resources

Nucleotide & Protein

Gene

RefSeq Processing Pipelines

A) Curation Pipeline (Vertebrate)

Curation Database

RefSeq Tracking Database

Errors resolved manually

Automated BLAST

Updates: Sequence Names Citations Map data

Longest mRNA for gene

provisional, predicted, or inferred RefSeq

Curation and review

Curated RefSeq (validated or reviewed)

B) Annotation Pipeline

Align RefSeq & GenBank transcripts to Genome Assembly

Predict Transcript & Protein Models

Select Best Models to Annotate

model RefSeq

C) GenBank Extraction Pipeline

Duplicate GenBank Record

Validation & RefSeq Format

Prokaryote: Protein Clustering

Add Names, GeneID

provisional, predicted, or curated RefSeq

RefSeq: NCBI’s Derivative Sequence Database

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

```
srcdb_refseq[Properties]
```

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

RefSeq Benefits

Non-redundancy

Explicitly linked nucleotide & protein sequences

Updates to reflect current sequence data & biology

Data **validation**

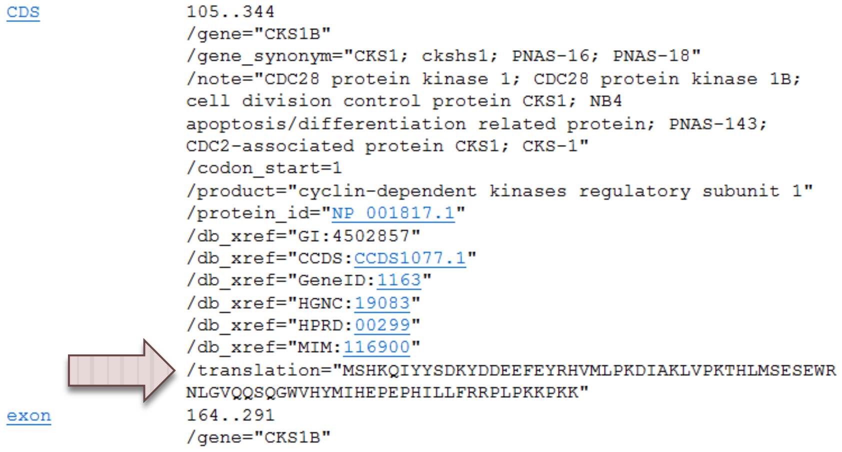
Format **consistency**

Distinct **accession** series

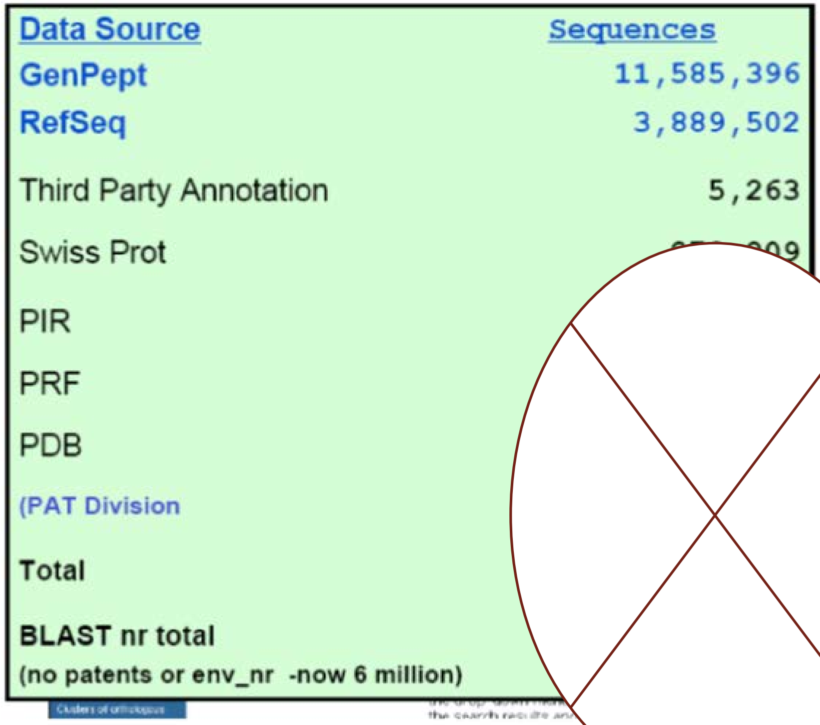
Stewardship by **NCBI staffs & collaborators**

Entrez Protein: Derivative Databases

Example: CKS1B



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

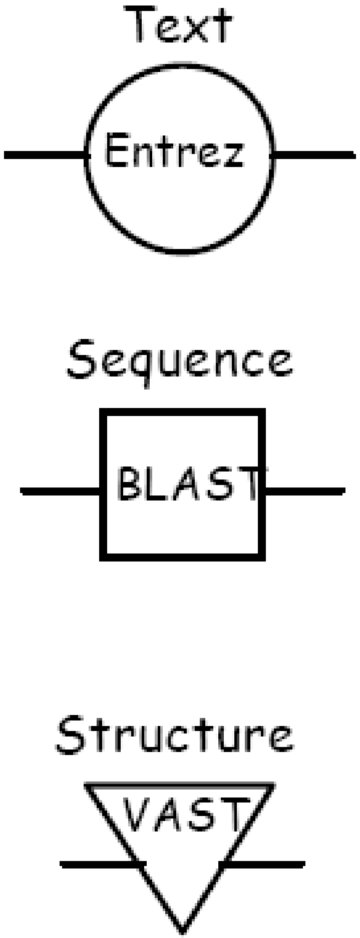


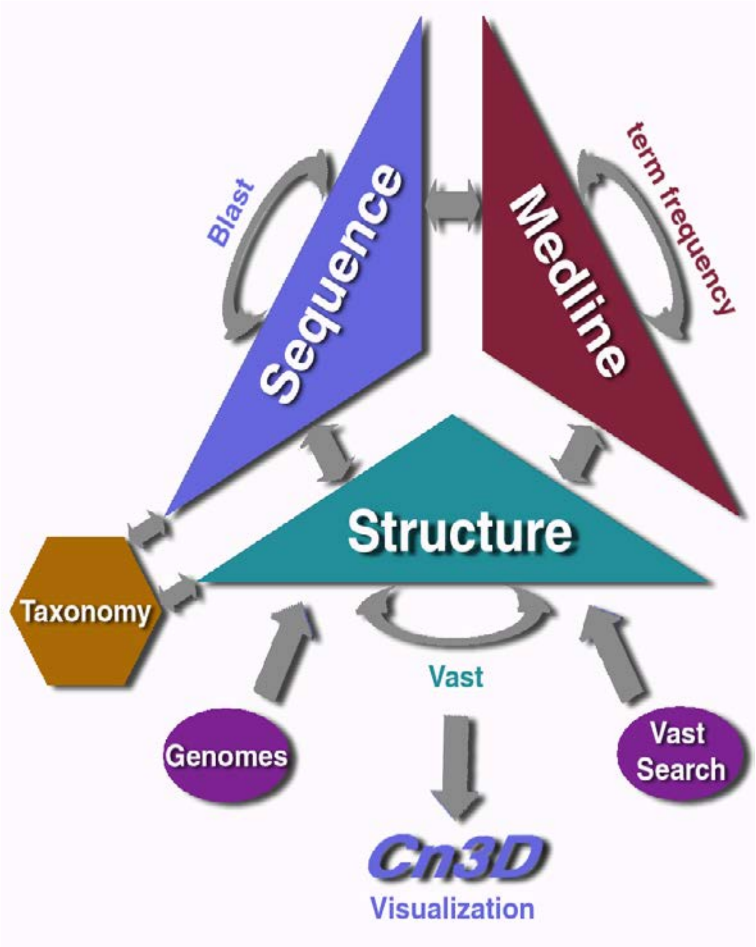
Data Source	Sequences
GenPept	11,585,396
RefSeq	3,889,502
Third Party Annotation	5,263
Swiss Prot	[illegible]
PIR	[illegible]
PRF	[illegible]
PDB	[illegible]
(PAT Division	[illegible]
Total	[illegible]
BLAST nr total	[illegible]
(no patents or env_nr -now 6 million)	

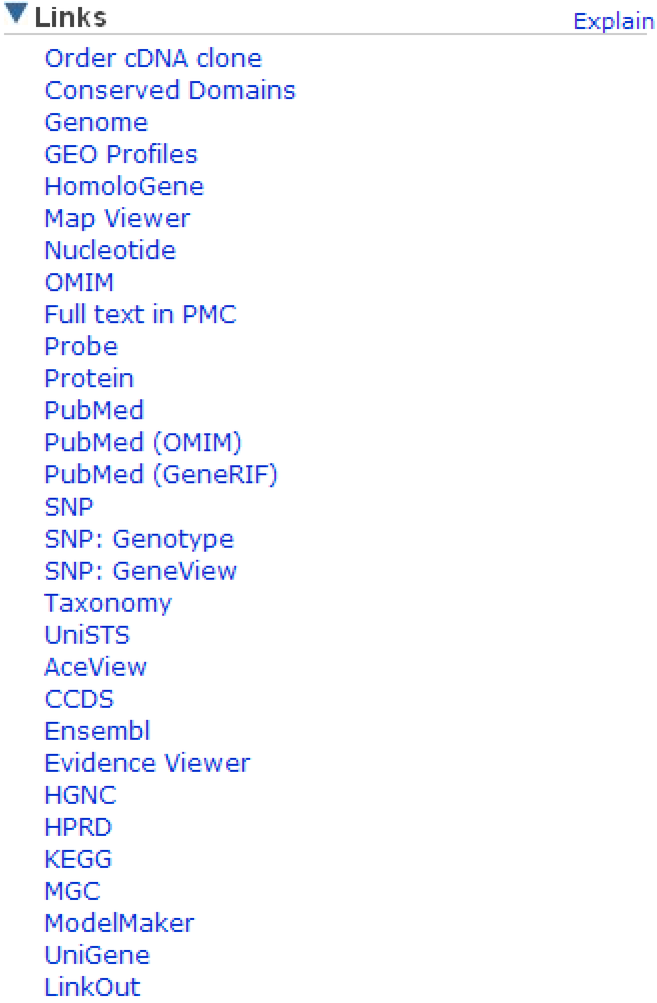
PAT: patent

Search in NCBI Databases

Searches	**Text**: e.g., *POU5F1* (Oct3/4);
	Sequence: e.g., *POU5F1*
	Structure: e.g., BRCA1



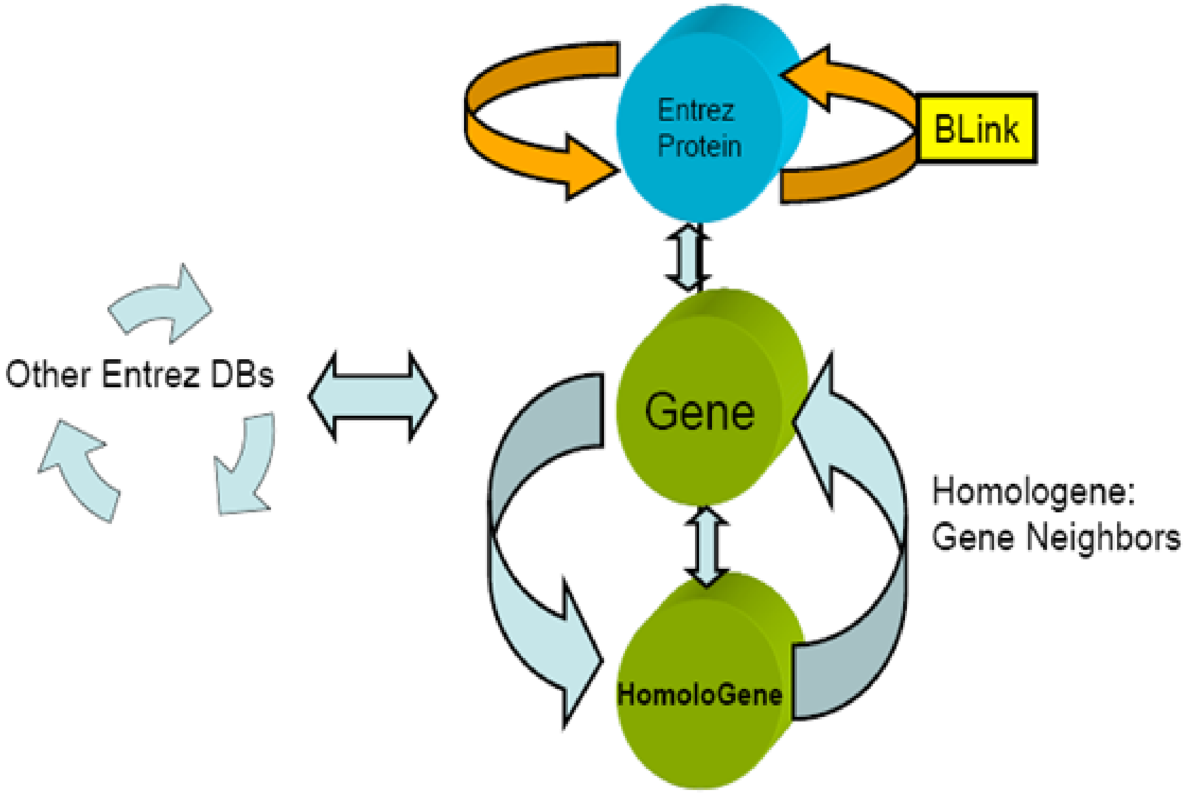



▼ Links

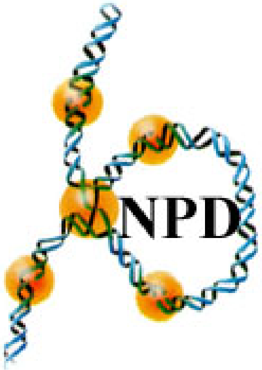
Order cDNA clone
Conserved Domains
Genome
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
OMIM
Full text in PMC
Probe
Protein
PubMed
PubMed (OMIM)
PubMed (GeneRIF)
SNP
SNP: Genotype
SNP: GeneView
Taxonomy
UniSTS
AceView
CCDS
Ensembl
Evidence Viewer
HGNC
HPRD
KEGG
MGC
ModelMaker
UniGene
LinkOut

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

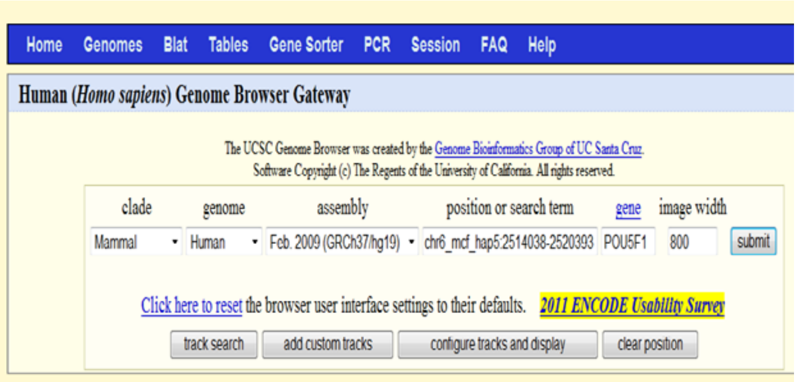
Entrez: Use Gene for everything



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



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



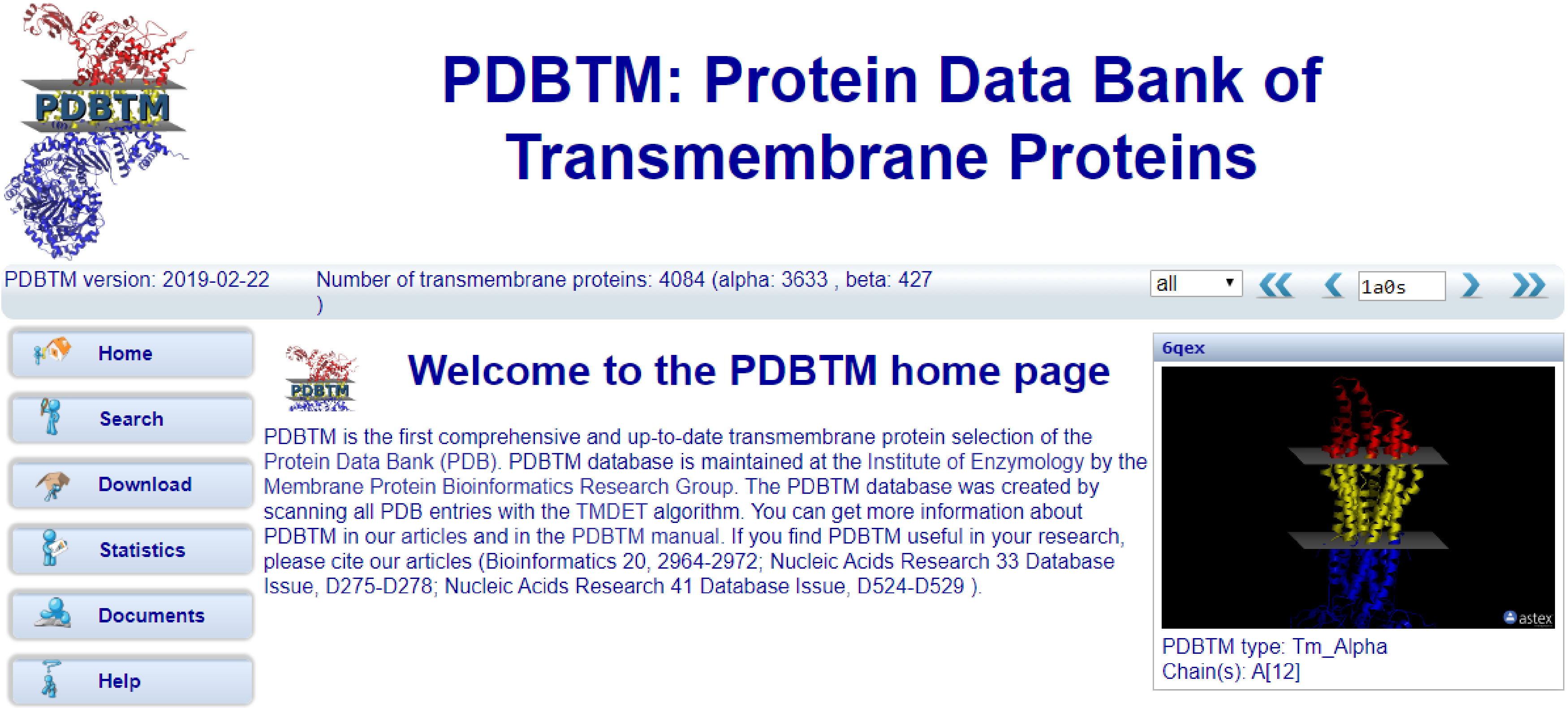
Human (*Homo sapiens*) Genome Browser Gateway

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	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

PDBTM version: 2019-02-22 Number of transmembrane proteins: 4084 (alpha: 3633 , beta: 427)

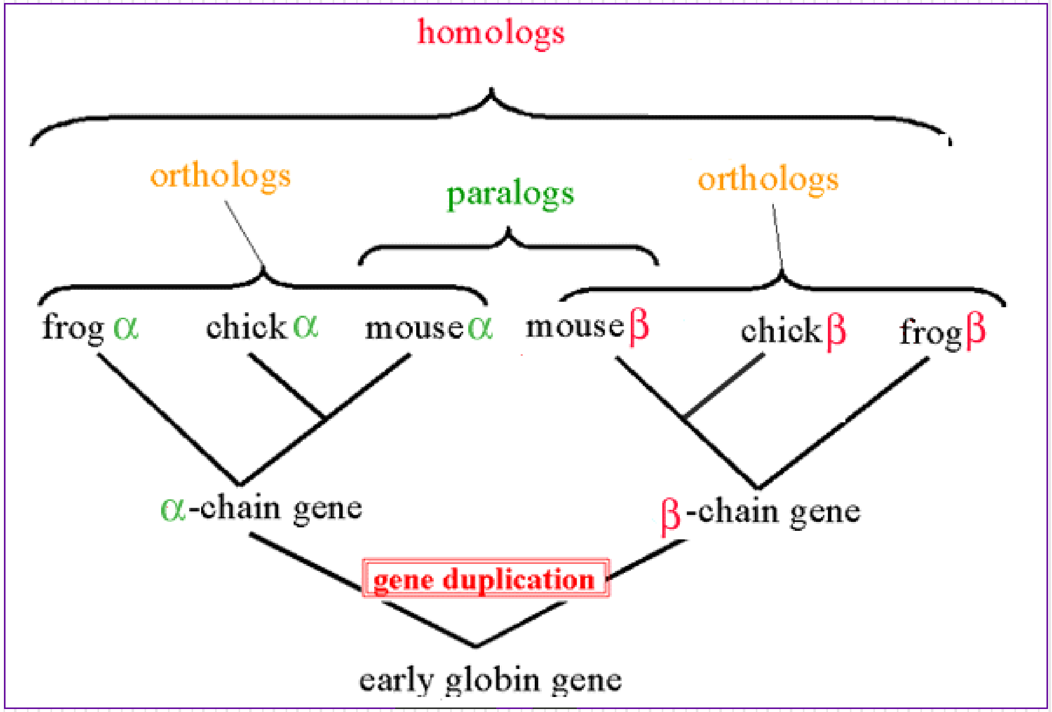
Welcome to the PDBTM home page

PDBTM is the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained at the Institute of Enzymology by the Membrane Protein Bioinformatics Research Group. The PDBTM database was created by scanning all PDB entries with the TMDET algorithm. You can get more information about PDBTM in our articles and in the 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-D278; Nucleic Acids Research 41 Database Issue, D524-D529).

6qex

PDBTM type: Tm_Alpha
Chain(s): A[12]

Q: How Do You Find the Orthologs from Other Species



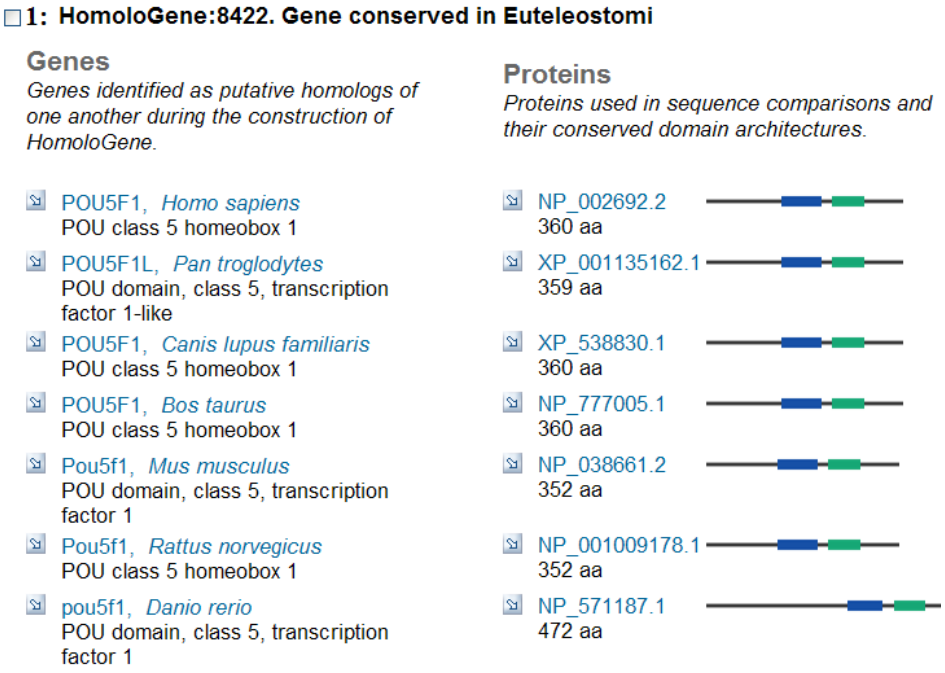
Homologs (1)

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



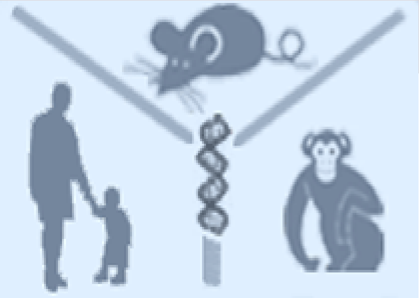
□ **1: HomoloGene:8422. Gene conserved in Euteleostomi**

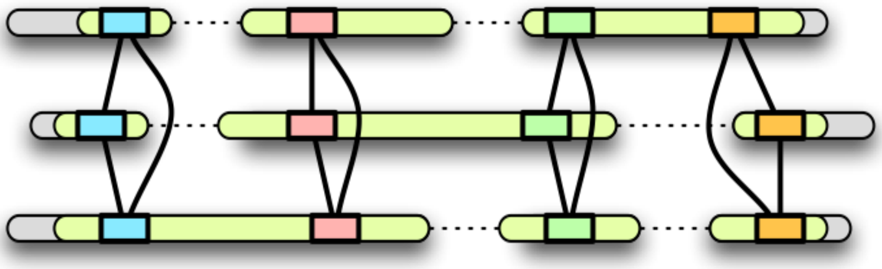
Genes *Genes identified as putative homologs of one another during the construction of HomoloGene.*	**Proteins** *Proteins used in sequence comparisons and their conserved domain architectures.*
POU5F1, *Homo sapiens* POU class 5 homeobox 1	NP_002692.2 360 aa
POU5F1L, *Pan troglodytes* POU domain, class 5, transcription factor 1-like	XP_001135162.1 359 aa
POU5F1, *Canis lupus familiaris* POU class 5 homeobox 1	XP_538830.1 360 aa
POU5F1, *Bos taurus* POU class 5 homeobox 1	NP_777005.1 360 aa
Pou5f1, *Mus musculus* POU domain, class 5, transcription factor 1	NP_038661.2 352 aa
Pou5f1, *Rattus norvegicus* POU class 5 homeobox 1	NP_001009178.1 352 aa
pou5f1, *Danio rerio* POU domain, class 5, transcription factor 1	NP_571187.1 472 aa

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)

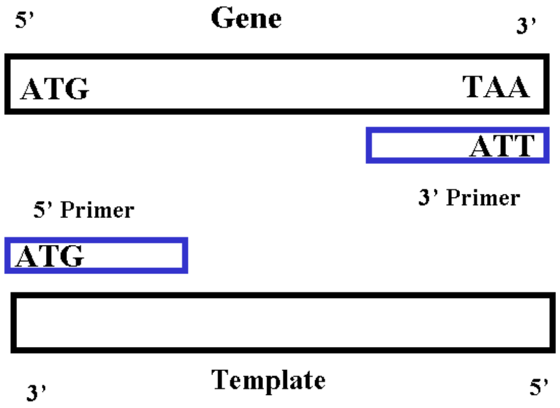




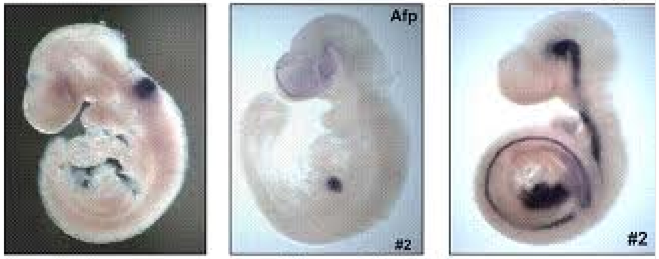
Mercator

Multiple Whole-Genome Orthology Map Construction

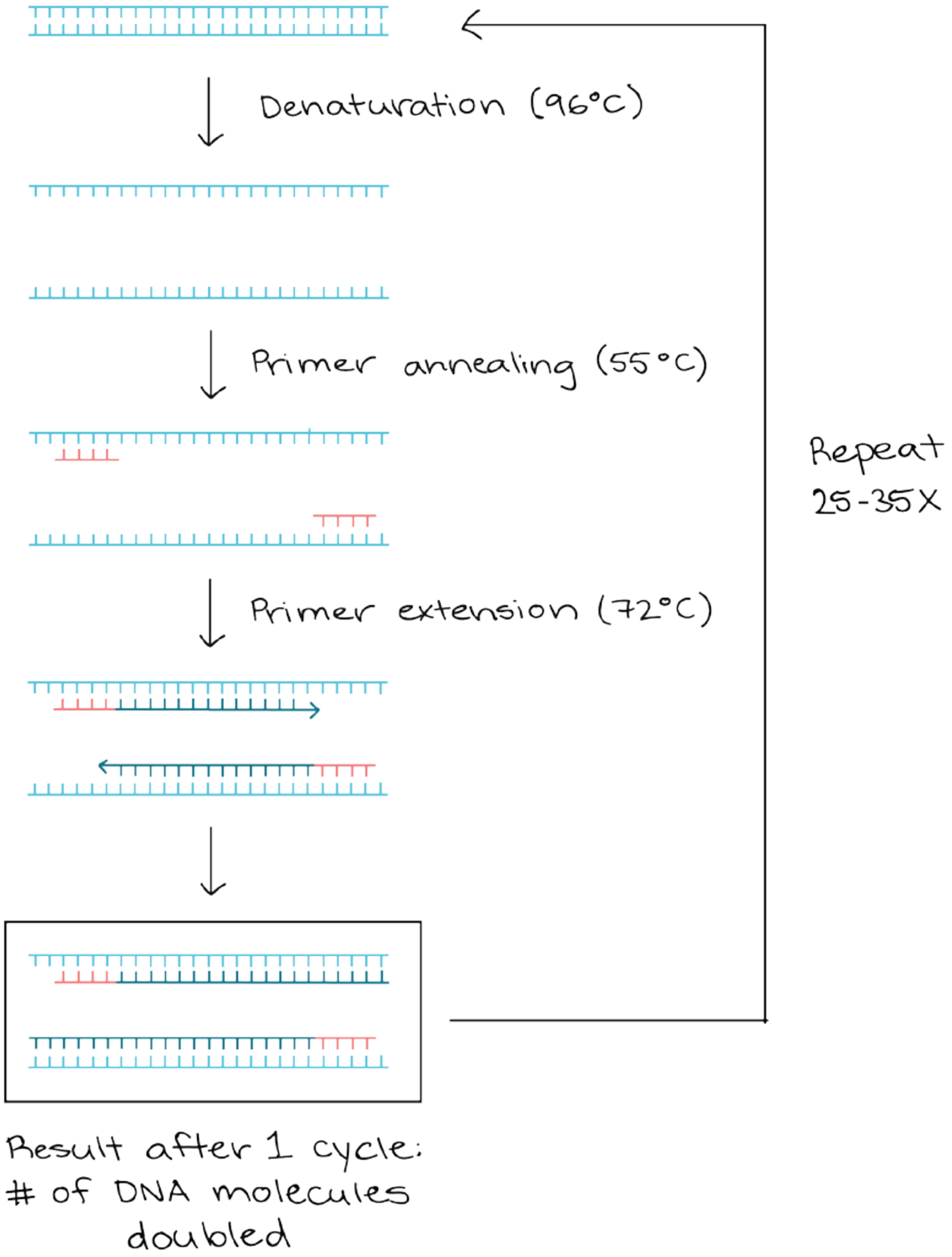
Q: How to Design Primers/Probes for PCR/qPCR/Cloning/in situ hybridization

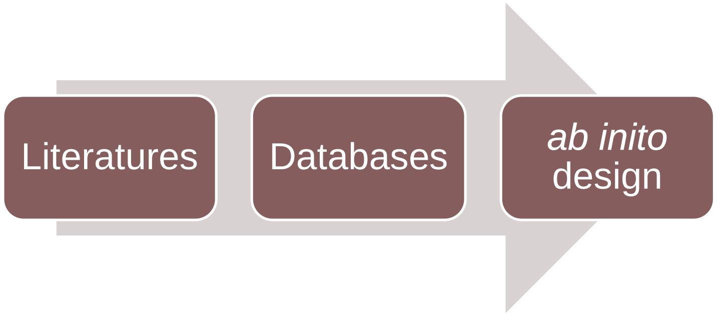


E9.5 embryos



Krox20 α-fetoprotein HNF3β

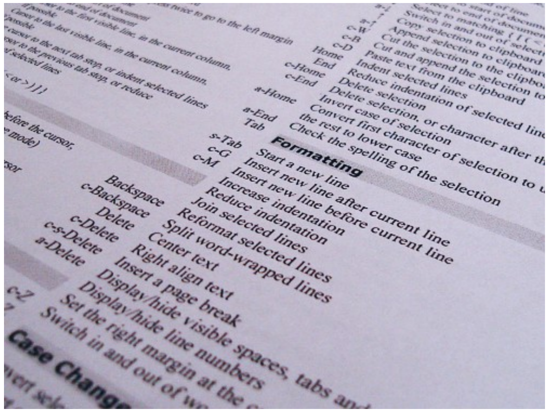

Denaturation (96°C)
Primer annealing (55°C)
Primer extension (72°C)
Repeat
25-35X
Result after 1 cycle:
# of DNA molecules
doubled



Literatures

Databases

ab inito design



↓

BLAST

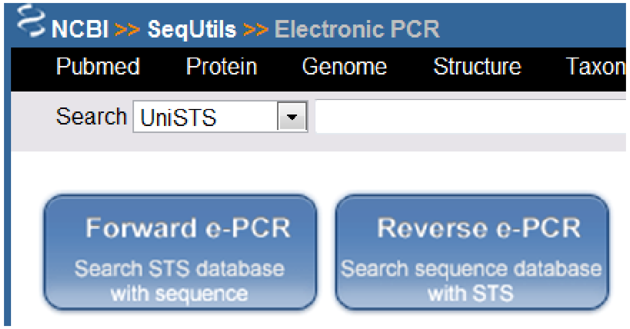
RT PRIMER DB

↓

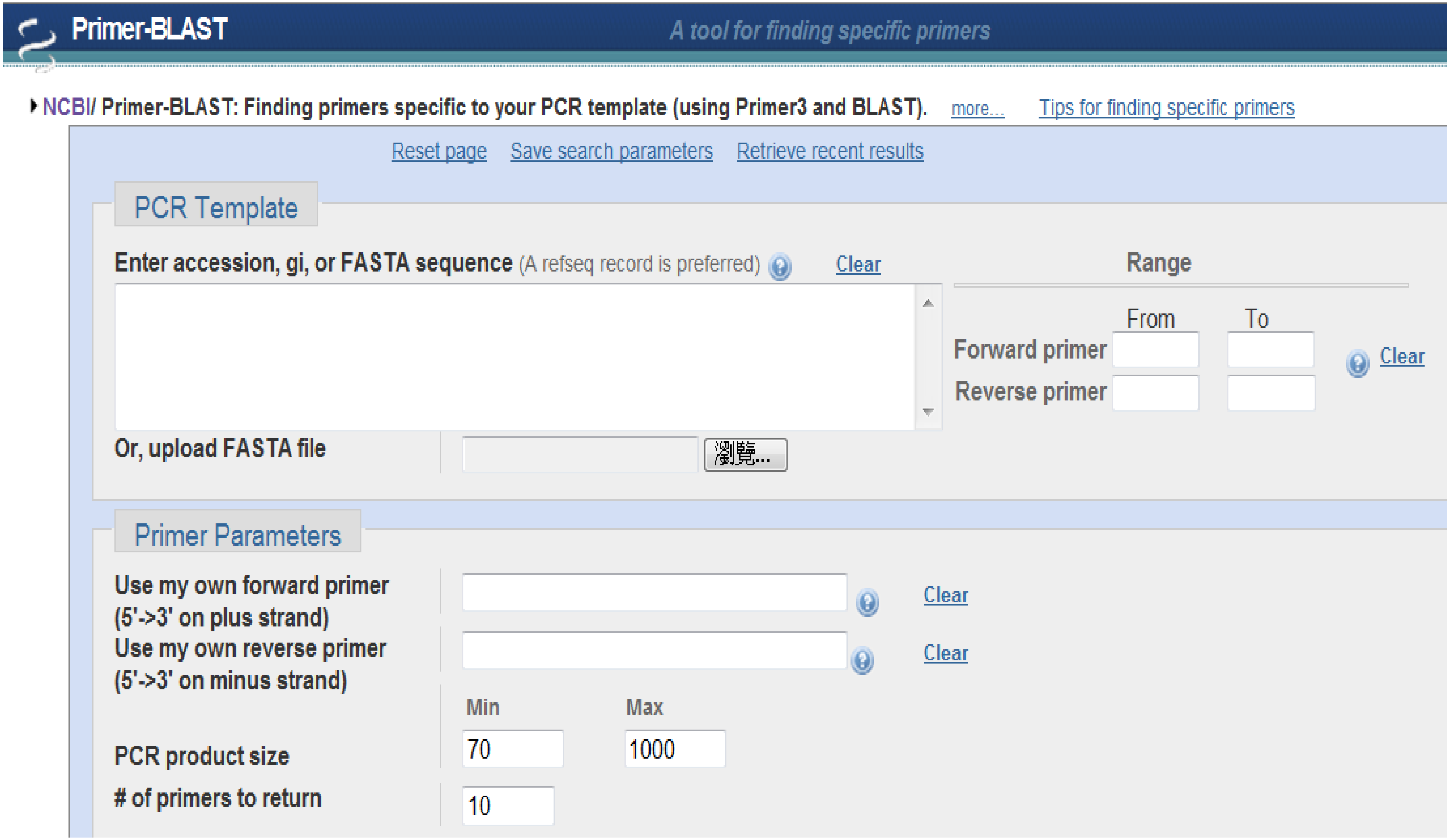
e.g., ACTB

↓

BLAST



NCBI >> SeqUtils >> Electronic PCR
Pubmed
Protein
Genome
Structure
Taxon
Search UniSTS
Forward e-PCR
Search STS database with sequence
Reverse e-PCR
Search sequence database with STS



Primer-BLAST

A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST). more... Tips for finding specific primers

Reset page Save search parameters Retrieve recent results

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

Range

	From	To
Forward primer		
Reverse primer		

Clear

Or, upload FASTA file 瀏覽...

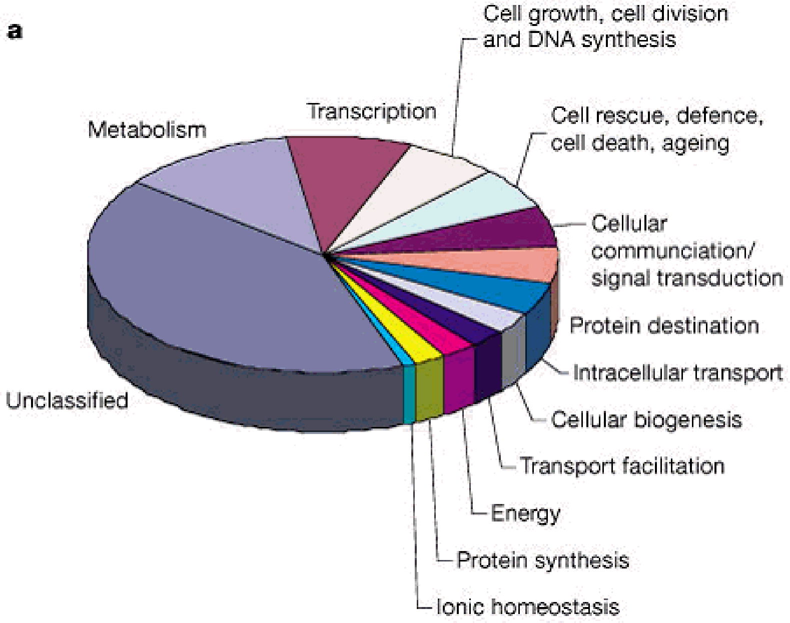
Primer Parameters

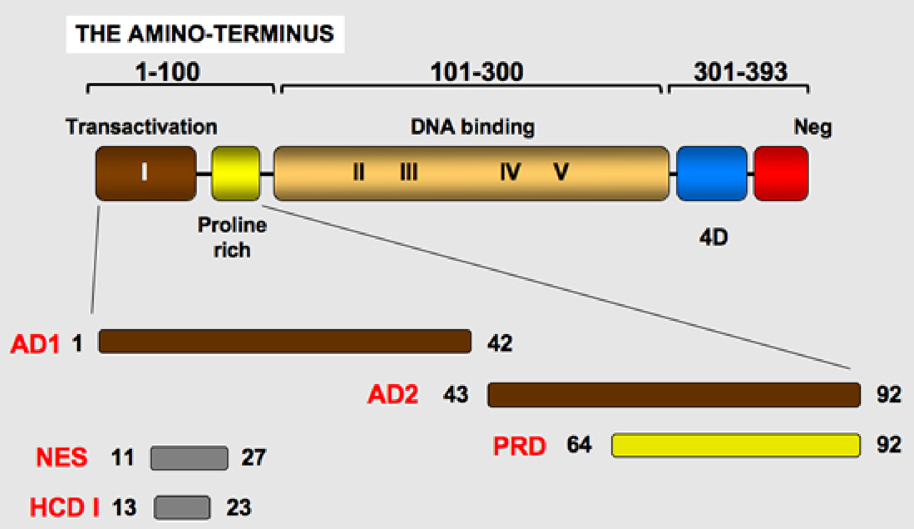
Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

	Min	Max
PCR product size	70	1000
# of primers to return	10	

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

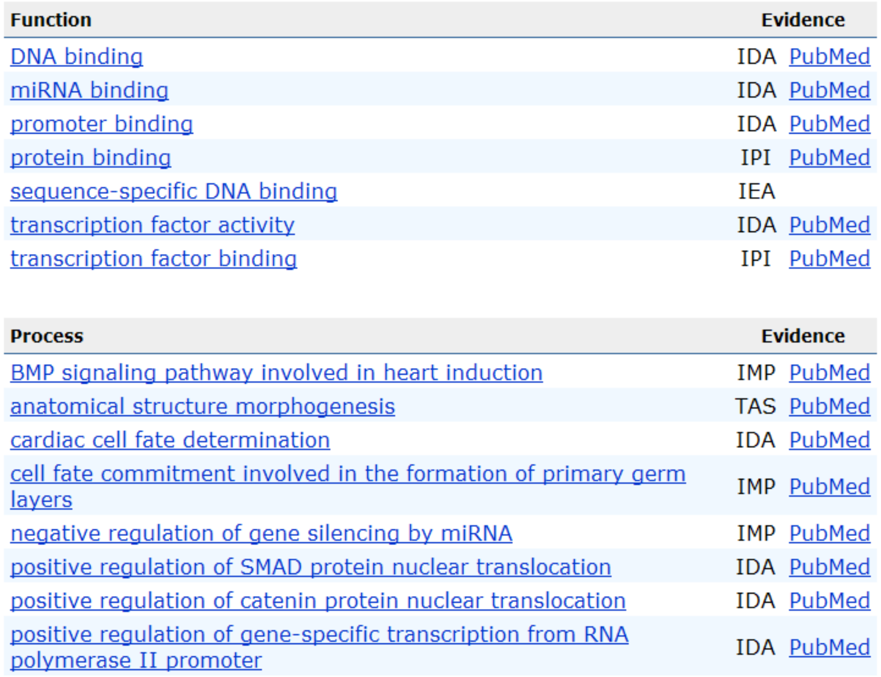




1. Gene Ontology

Through **integrated databases**

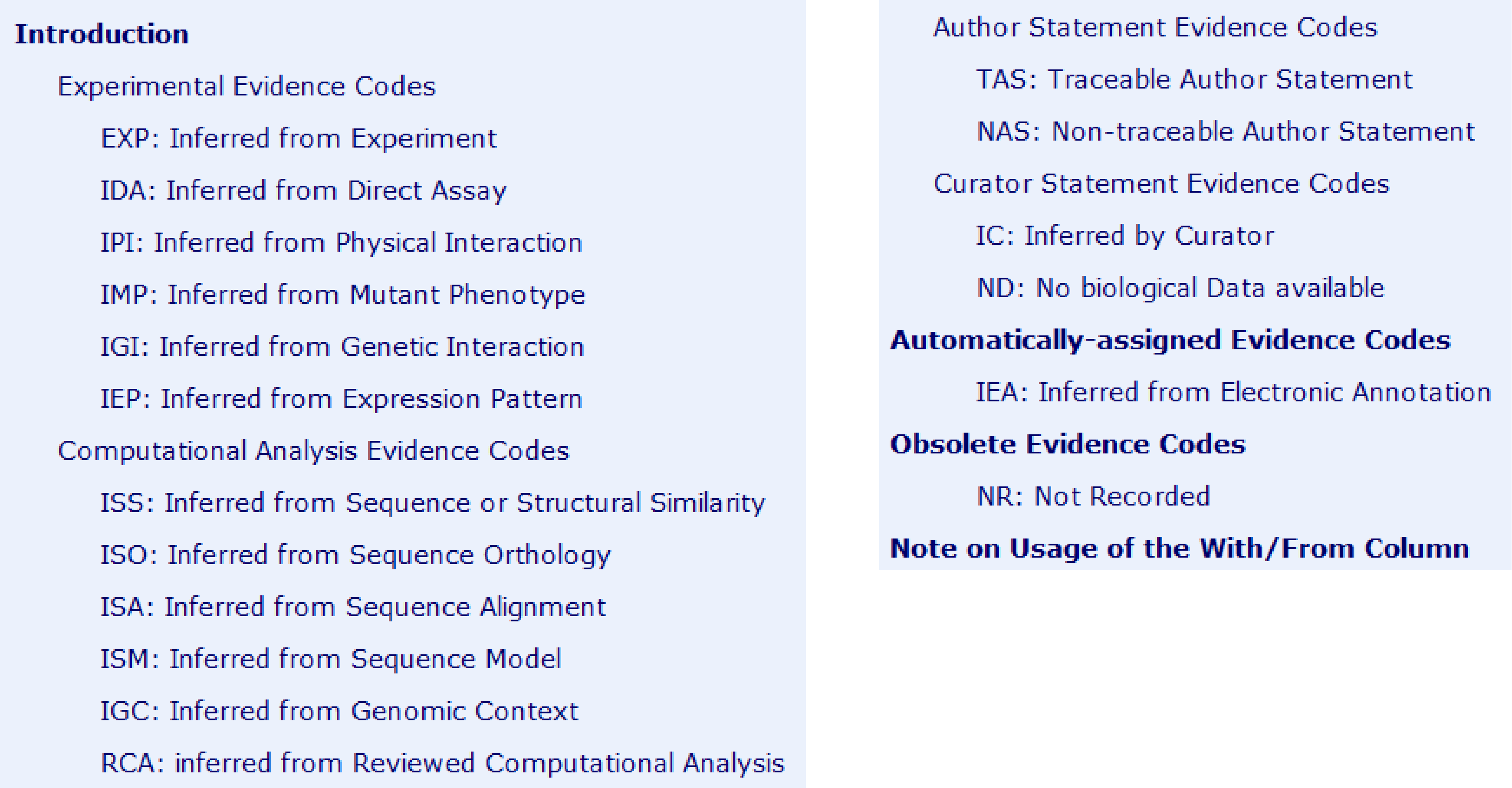
- Entrez_Gene
  - **GO terms**
- GeneCards
  - **GO terms**
- Uniprot/Swiss-Prot
  - POU5F1_Human
  - General annotation (comments)
  - Ontologies



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

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

GO Evidence Code



Introduction

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

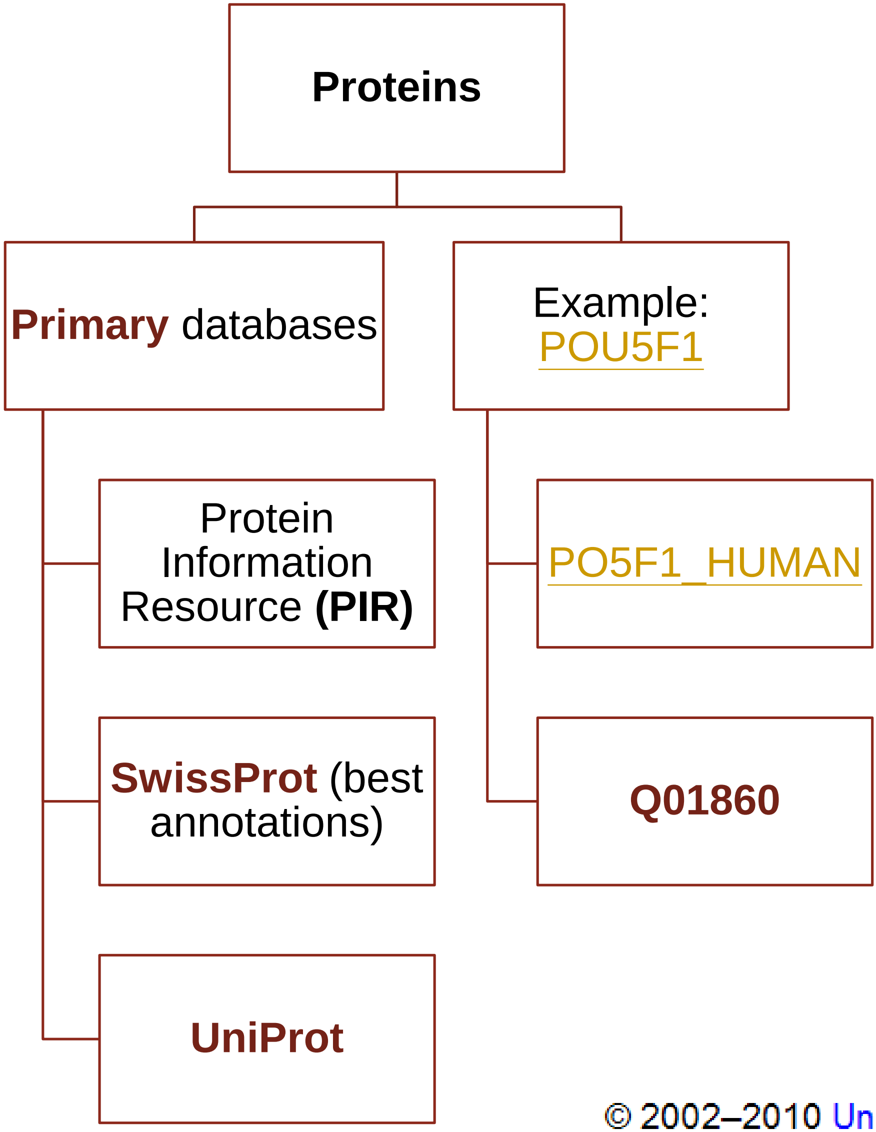
Automatically-assigned Evidence Codes

- IEA: Inferred from Electronic Annotation

Obsolete Evidence Codes

- NR: Not Recorded

Note on Usage of the With/From Column



Proteins

Primary databases

Example: POU5F1

Protein Information Resource **(PIR)**

PO5F1_HUMAN

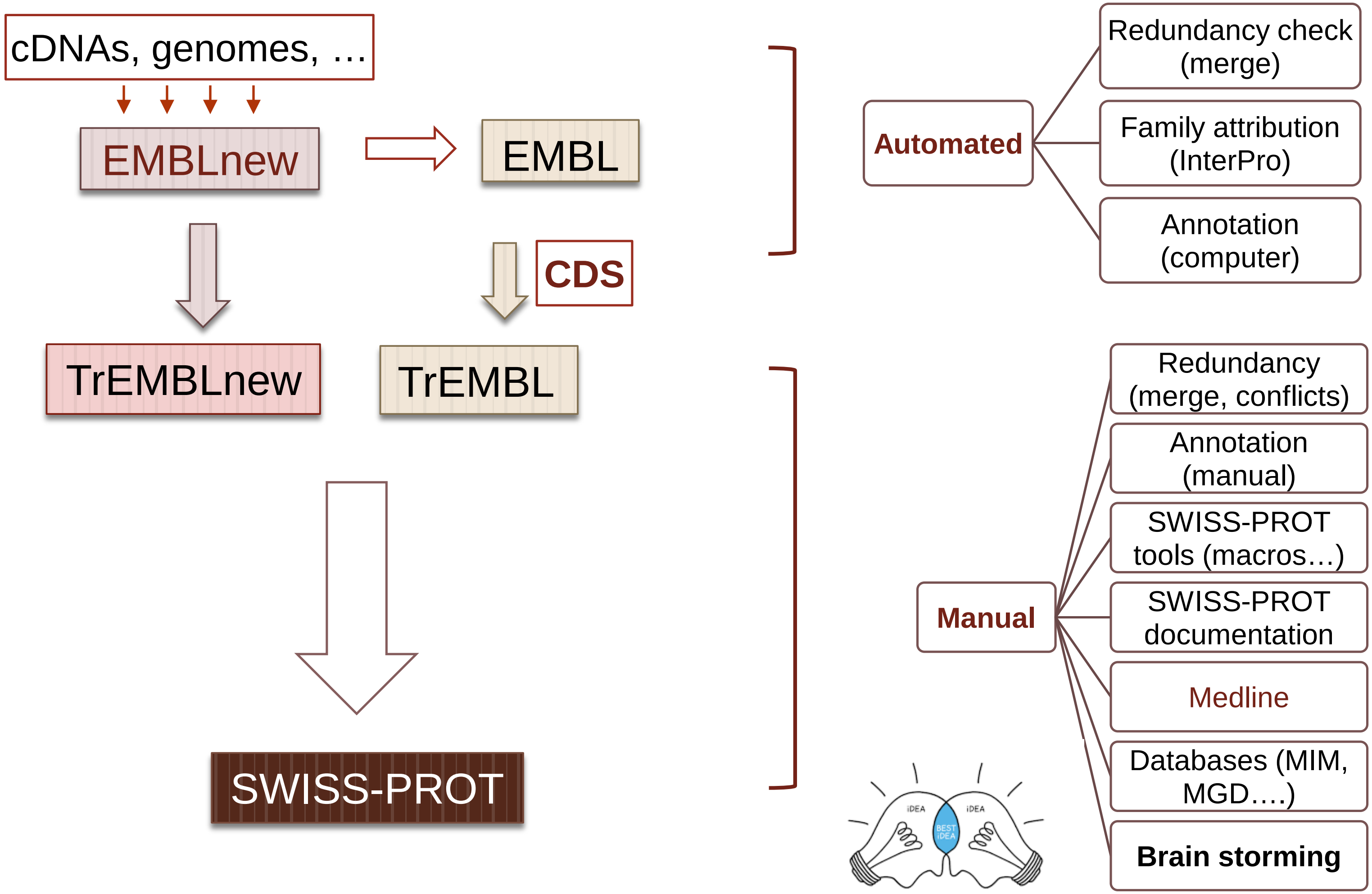
SwissProt (best annotations)

Q01860

UniProt

EMBL-EBI

The Simplified Story of a SWISS-PROT Entry



cDNAs, genomes, …

EMBLnew

EMBL

CDS

TrEMBLnew

TrEMBL

SWISS-PROT

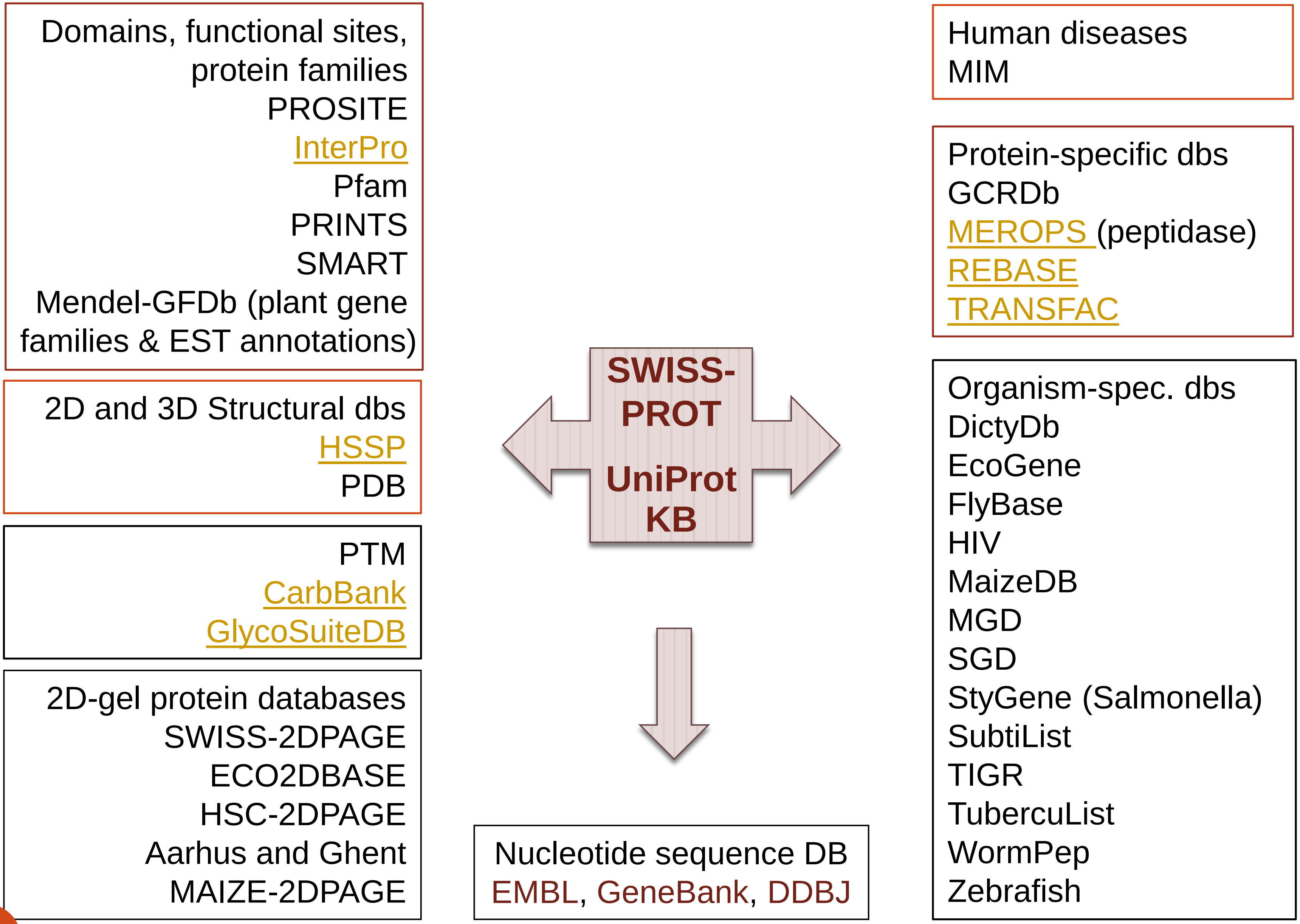
Automated

- Redundancy check (merge)
- Family attribution (InterPro)
- Annotation (computer)

Manual

- Redundancy (merge, conflicts)
- Annotation (manual)
- SWISS-PROT tools (macros…)
- SWISS-PROT documentation
- Medline
- Databases (MIM, MGD….)
- **Brain storming**

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



Domains, functional sites, protein families
PROSITE
InterPro
Pfam
PRINTS
SMART
Mendel-GFDb (plant gene families & EST annotations)

2D and 3D Structural dbs
HSSP
PDB

PTM
CarbBank
GlycoSuiteDB

2D-gel protein databases
SWISS-2DPAGE
ECO2DBASE
HSC-2DPAGE
Aarhus and Ghent
MAIZE-2DPAGE

SWISS-PROT

UniProt KB

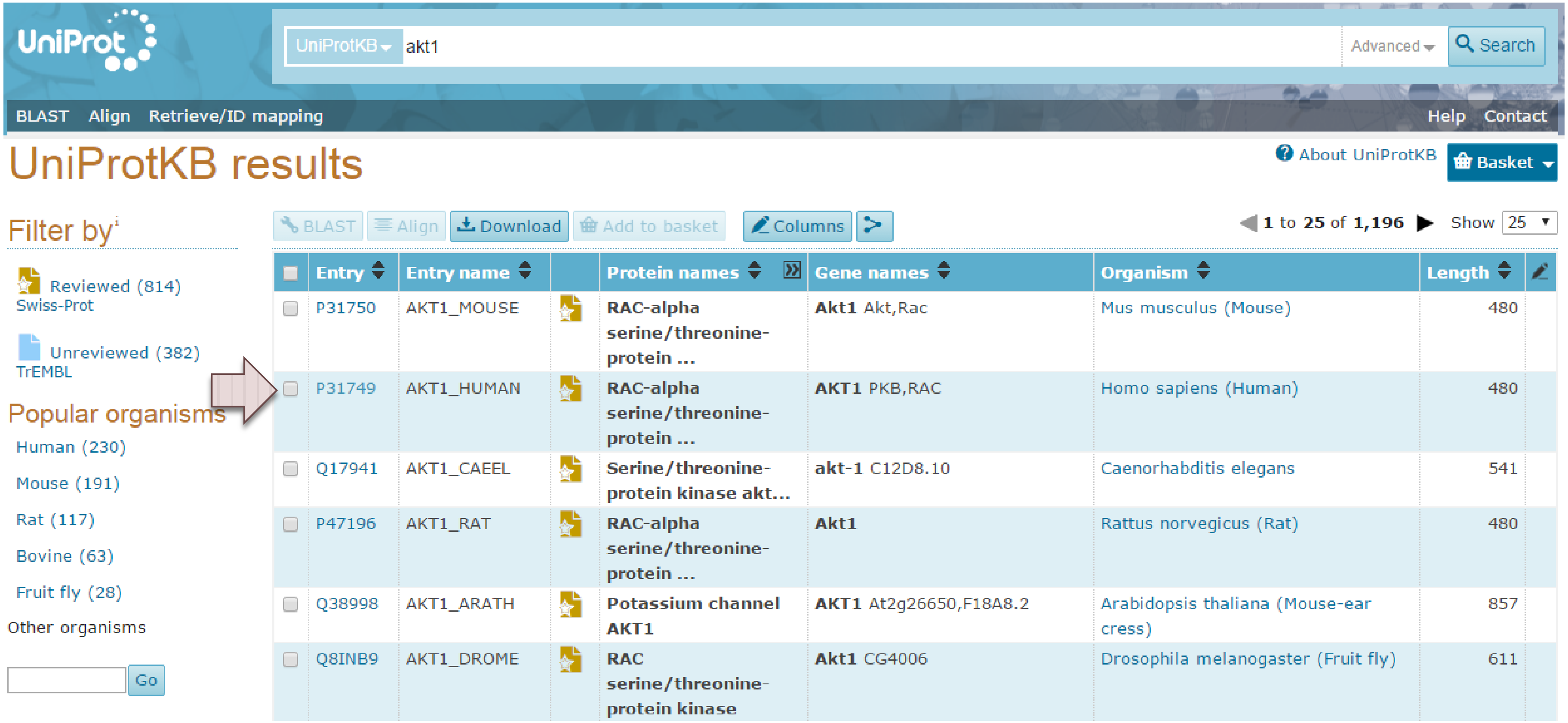
Nucleotide sequence DB
EMBL, GeneBank, DDBJ

Human diseases
MIM

Protein-specific dbs
GCRDb
MEROPS (peptidase)
REBASE
TRANSFAC

Organism-spec. dbs
DictyDb
EcoGene
FlyBase
HIV
MaizeDB
MGD
SGD
StyGene (Salmonella)
SubtiList
TIGR
TuberculList
WormPep
Zebrafish

2. UniProt/InterProt Annotations



Search: akt1

UniProtKB results

1 to 25 of 1,196

Filter by

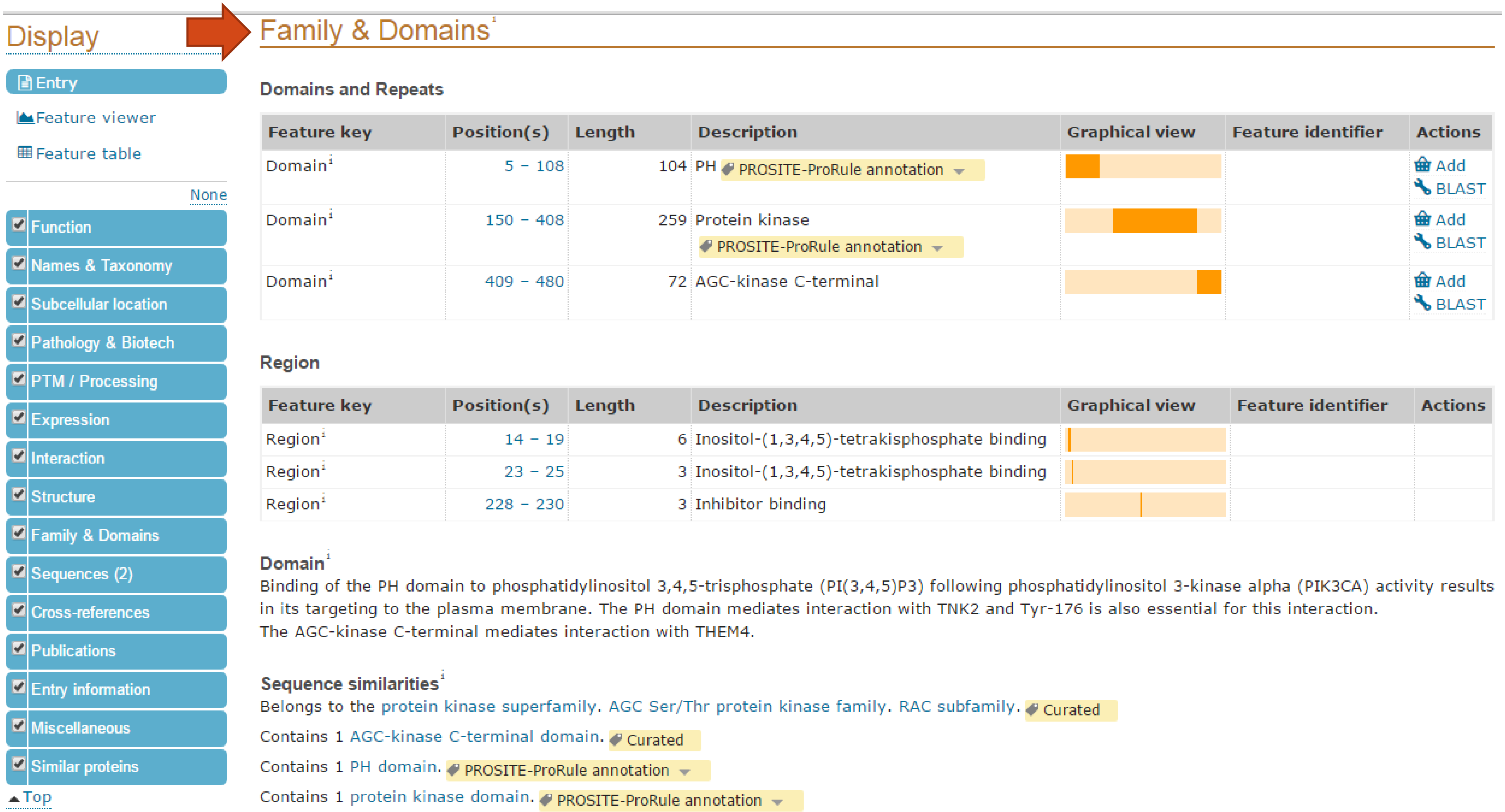
- Reviewed (814) Swiss-Prot
- Unreviewed (382) TrEMBL

Popular organisms

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

Other organisms

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



Family & Domains

Domains and Repeats

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

Region

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

Domain

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

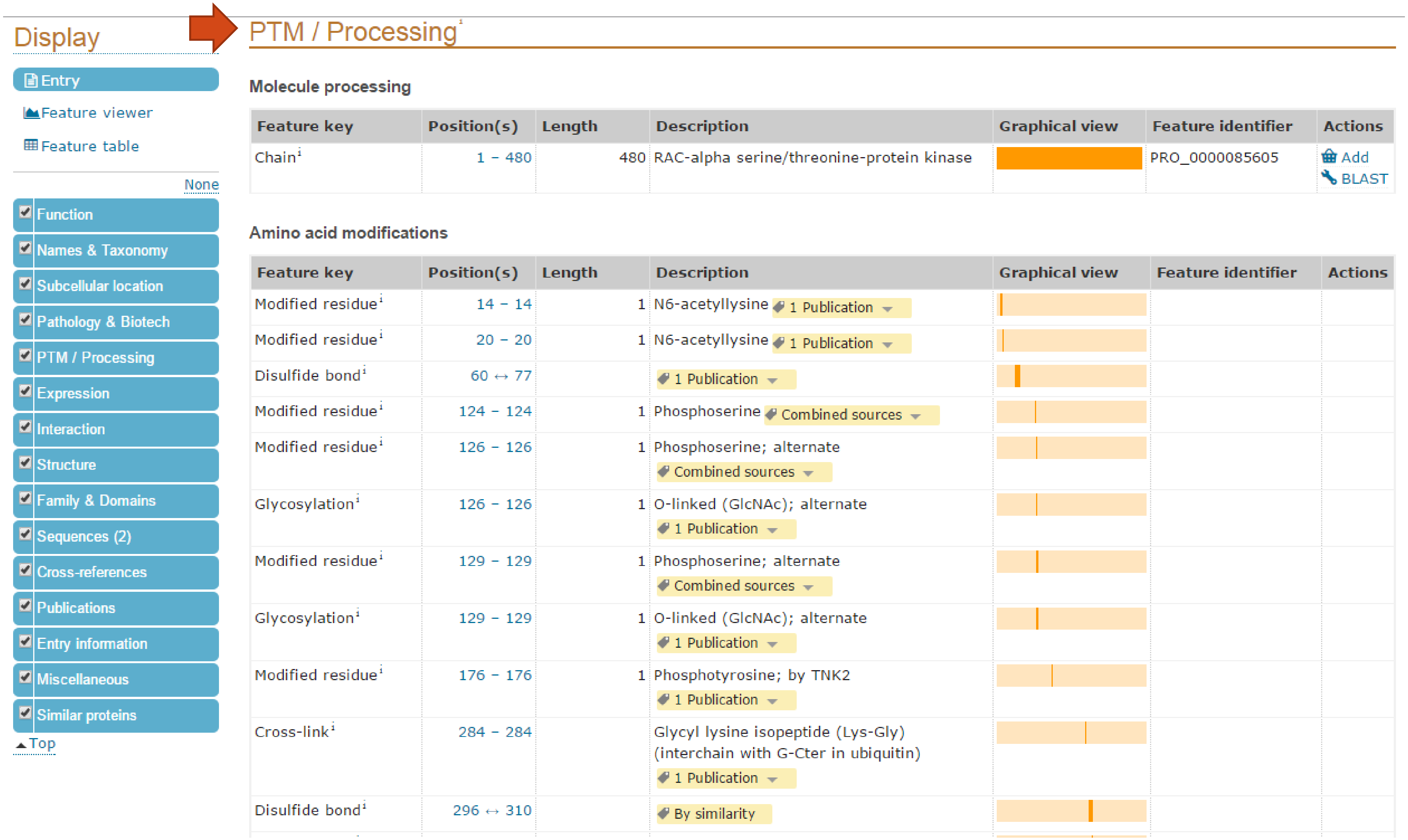
Sequence similarities

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

Contains 1 AGC-kinase C-terminal domain. Curated

Contains 1 PH domain. PROSITE-ProRule annotation

Contains 1 protein kinase domain. PROSITE-ProRule annotation



Display

- Entry
- Feature viewer
- Feature table

None

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

Top

PTM / Processing

Molecule processing

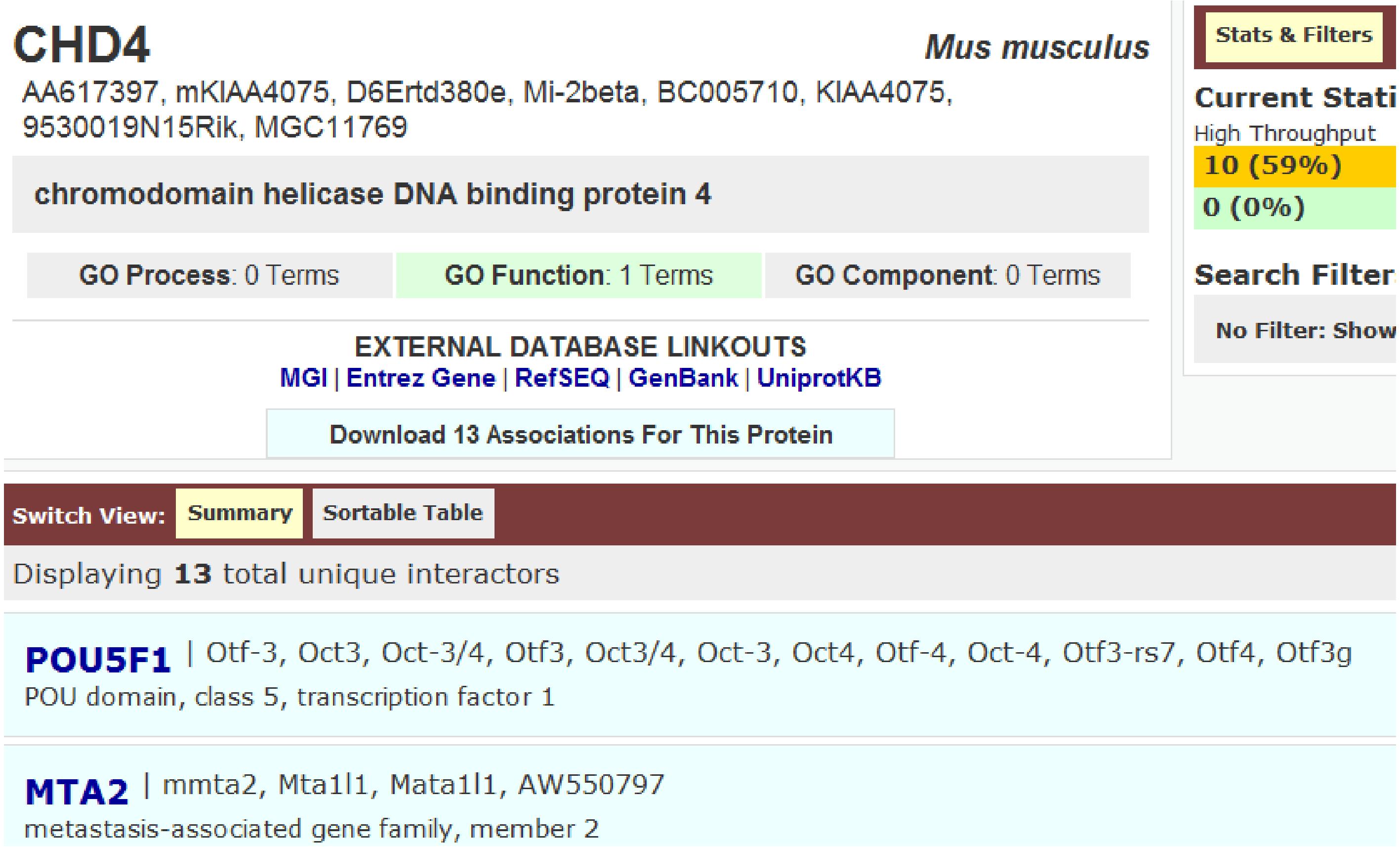
Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Chain	1 – 480	480	RAC-alpha serine/threonine-protein kinase		PRO_0000085605	Add BLAST

Amino acid modifications

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

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

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

Current Stati

High Throughput

10 (59%)

0 (0%)

Search Filter

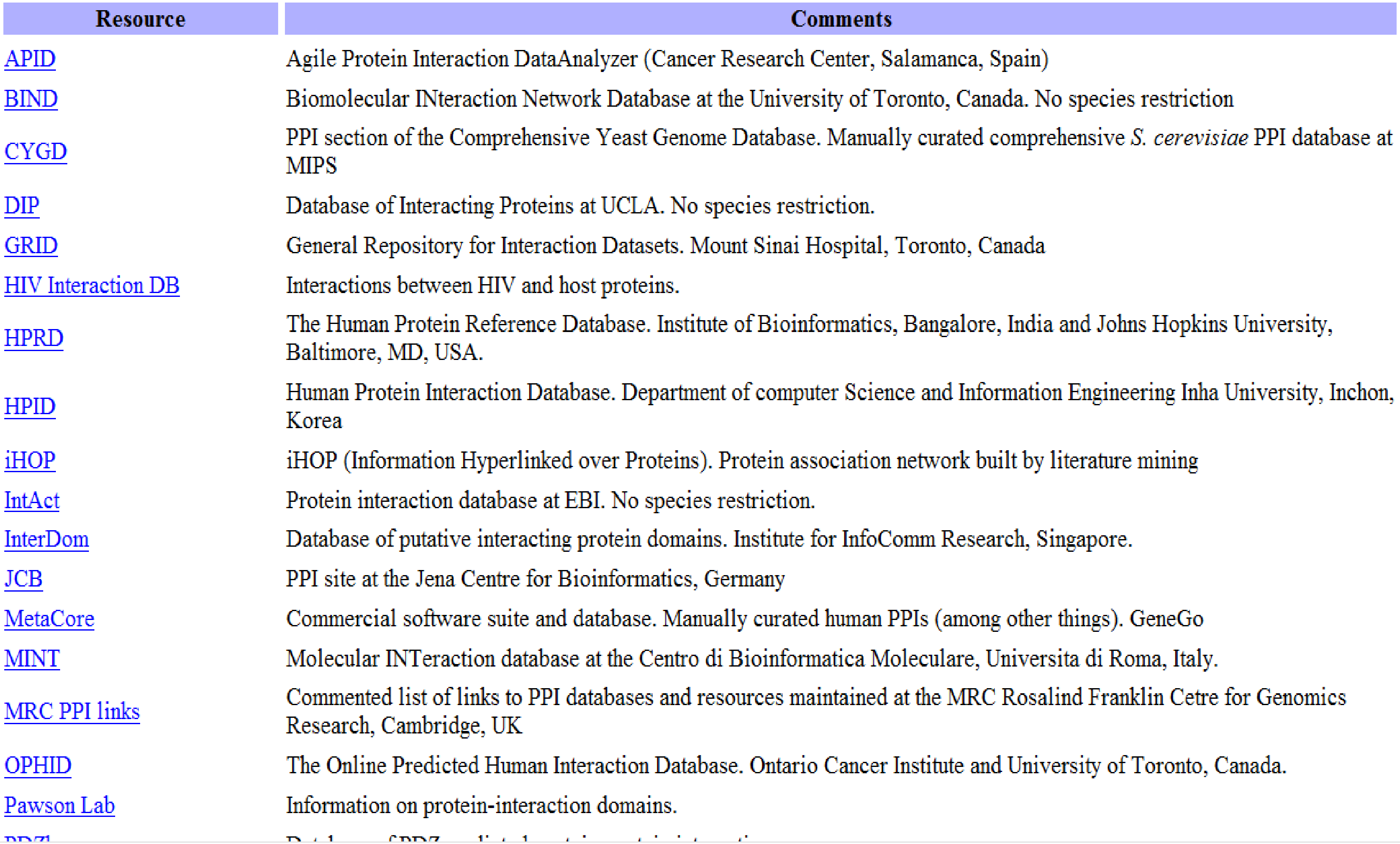
No Filter: Show

Displaying **13** total unique interactors

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

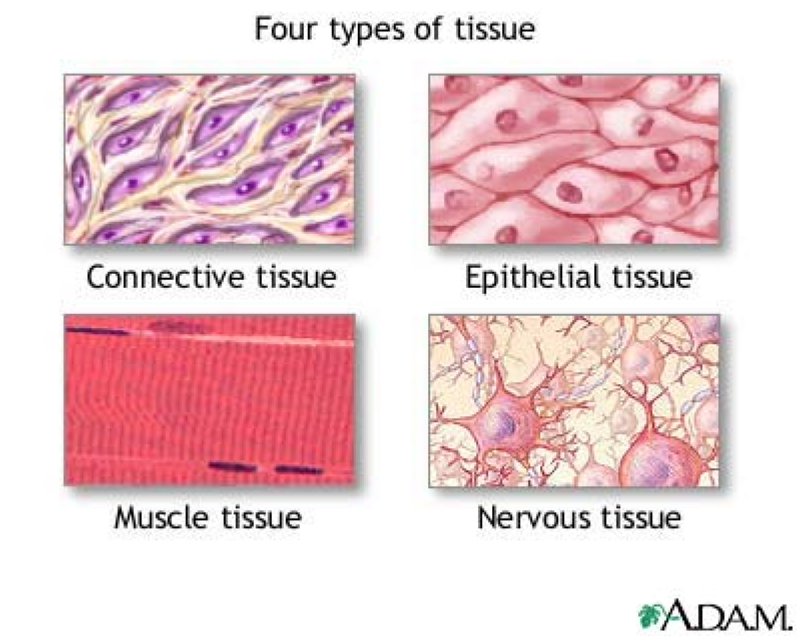
MTA2 | mmta2, Mta1l1, Mata1l1, AW550797
metastasis-associated gene family, member 2

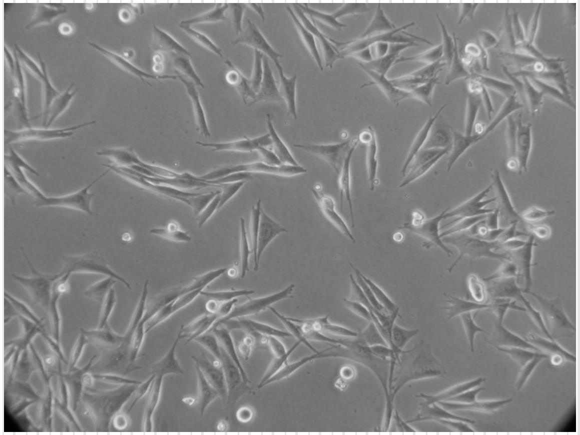
Databases for Protein – Protein Interaction

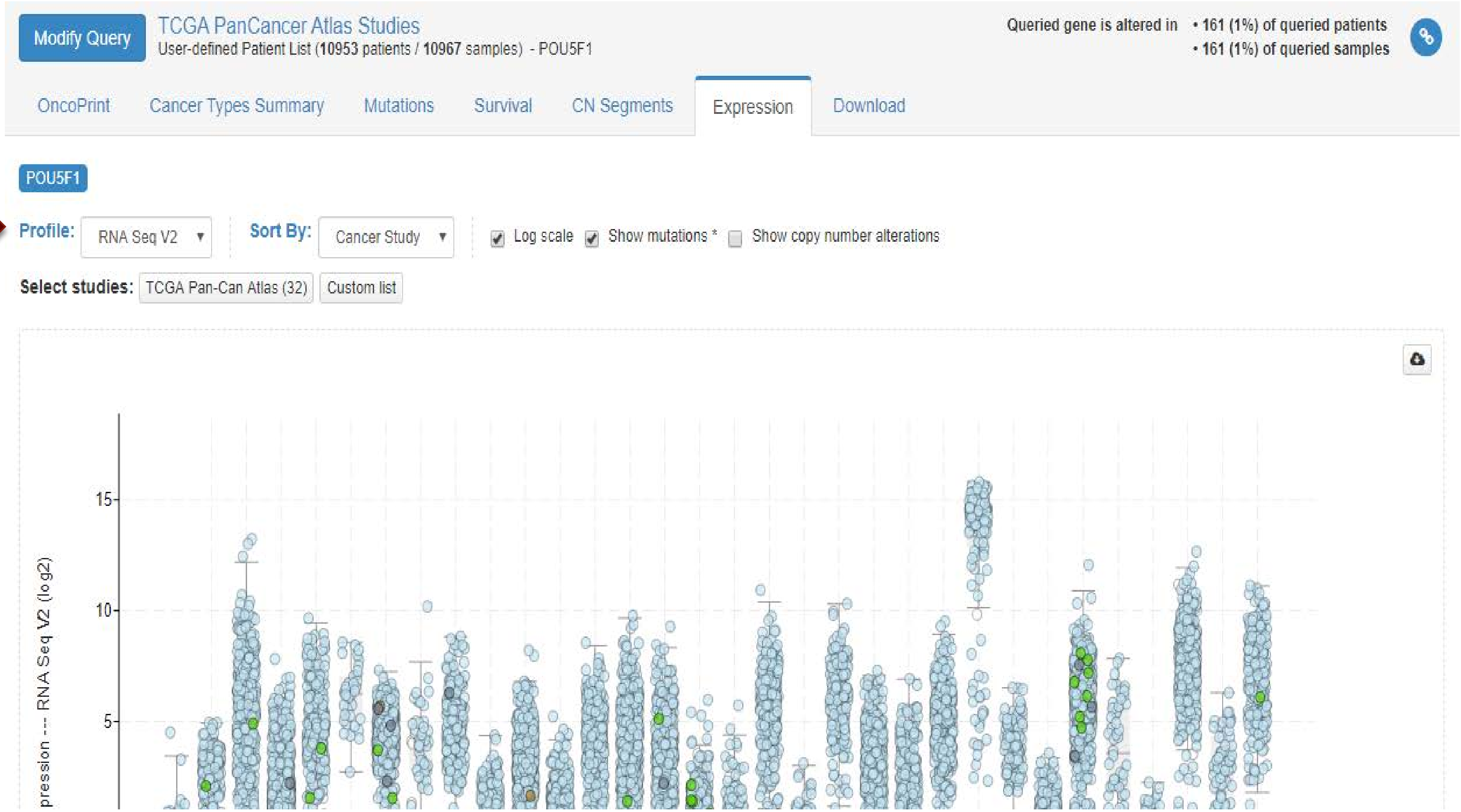


Resource	Comments
APID	Agile Protein Interaction DataAnalyzer (Cancer Research Center, Salamanca, Spain)
BIND	Biomolecular INteraction Network Database at the University of Toronto, Canada. No species restriction
CYGD	PPI section of the Comprehensive Yeast Genome Database. Manually curated comprehensive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

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







TCGA PanCancer Atlas Studies

User-defined Patient List (10953 patients / 10967 samples) - POU5F1

Queried gene is altered in
- 161 (1%) of queried patients
- 161 (1%) of queried samples

Profile: RNA Seq V2 ▾ | Sort By: Cancer Study ▾ | ☑ Log scale ☑ Show mutations * □ Show copy number alterations

Select studies: TCGA Pan-Can Atlas (32) | Custom list

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

Coffee Break

Tutorials for NCBI Tools

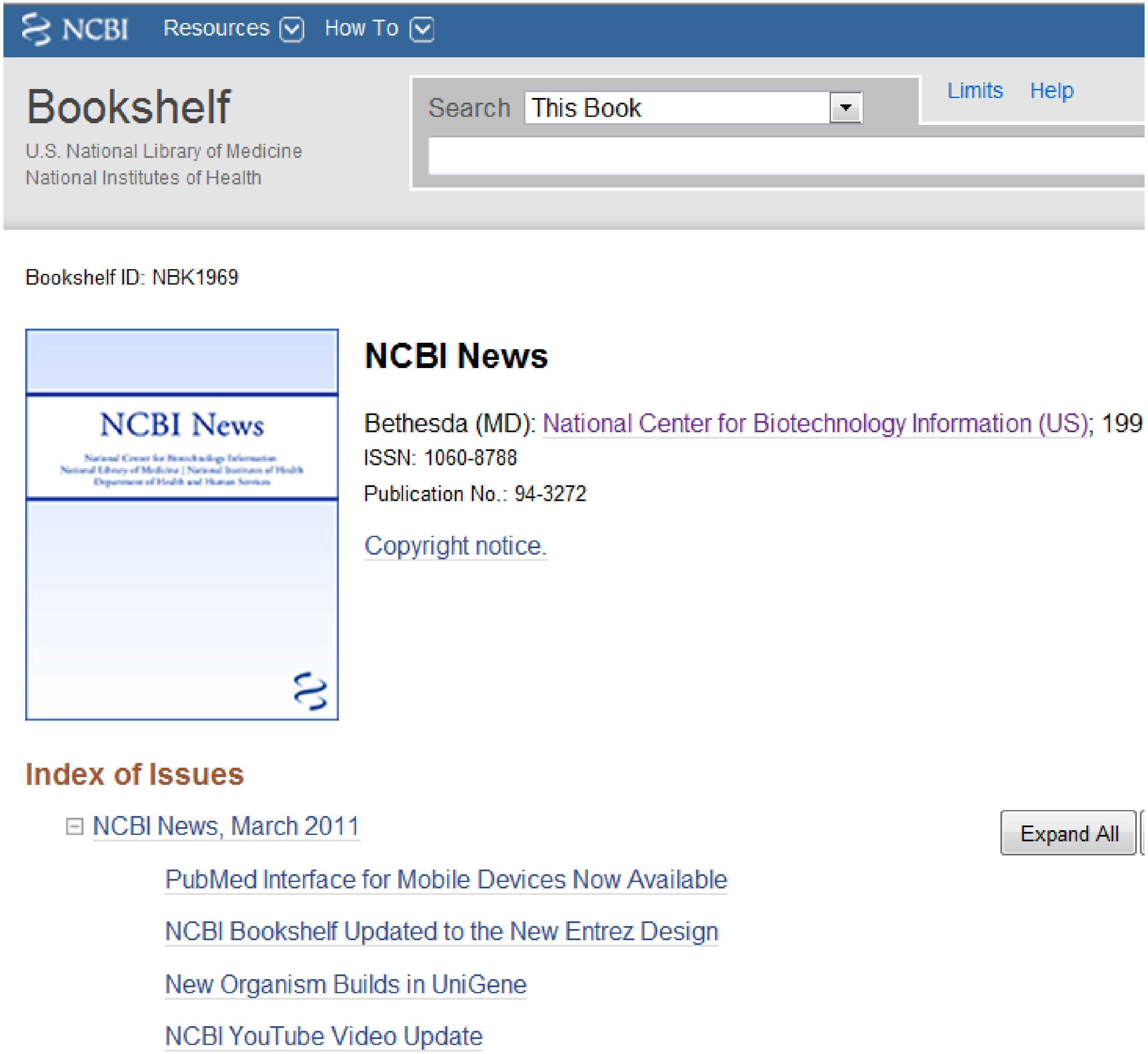
Edited by Laura Dean and Johanna McEntyre.

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.


NCBI
Resources
How To
Search
This Book
Limits
Help
NCBI News
National Center for Biotechnology Information
National Library of Medicine | National Institutes of Health
Department of Health and Human Services
Expand All

Bookshelf

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

Bookshelf ID: NBK1969

NCBI News

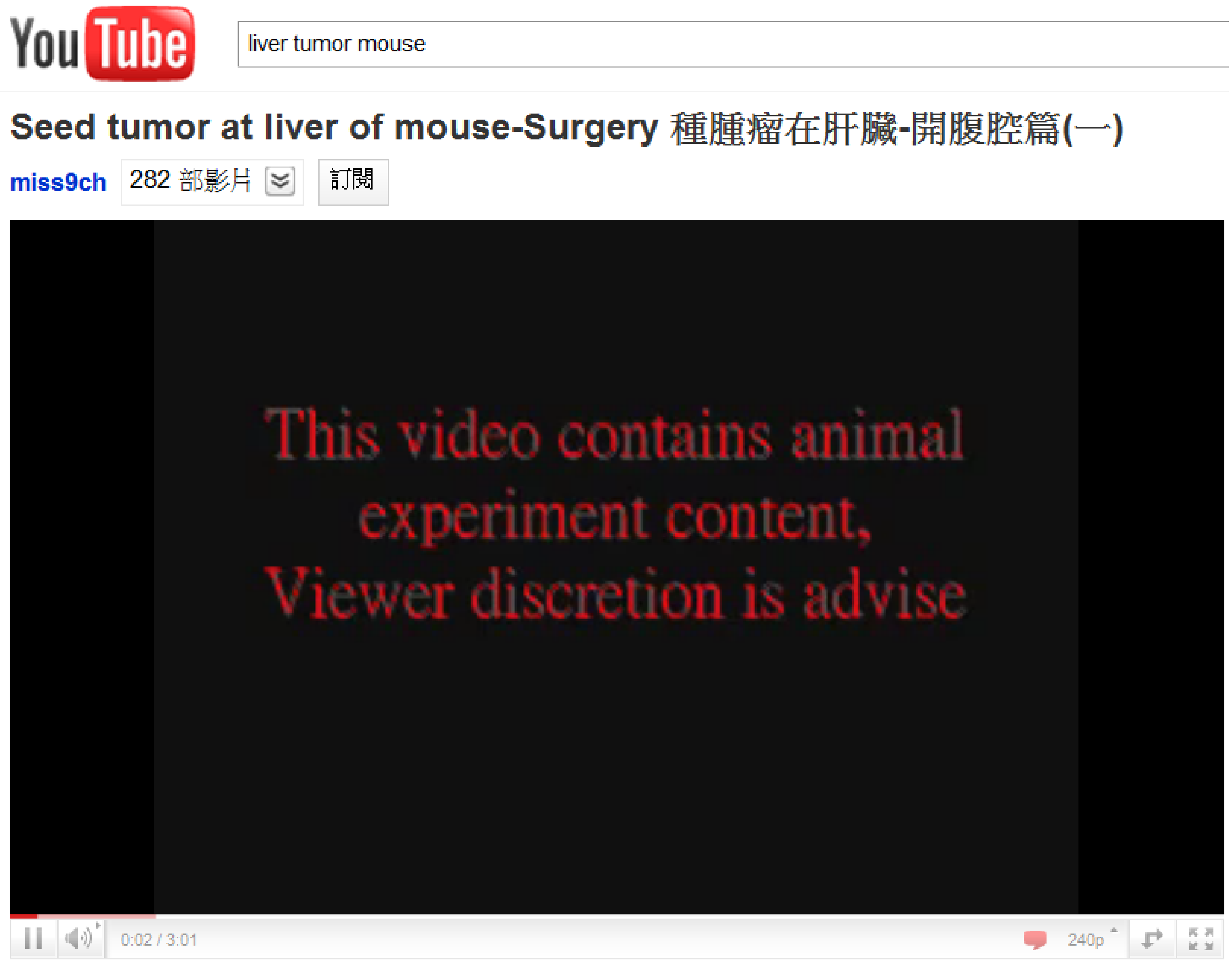
Bethesda (MD): National Center for Biotechnology Information (US); 199
ISSN: 1060-8788
Publication No.: 94-3272

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



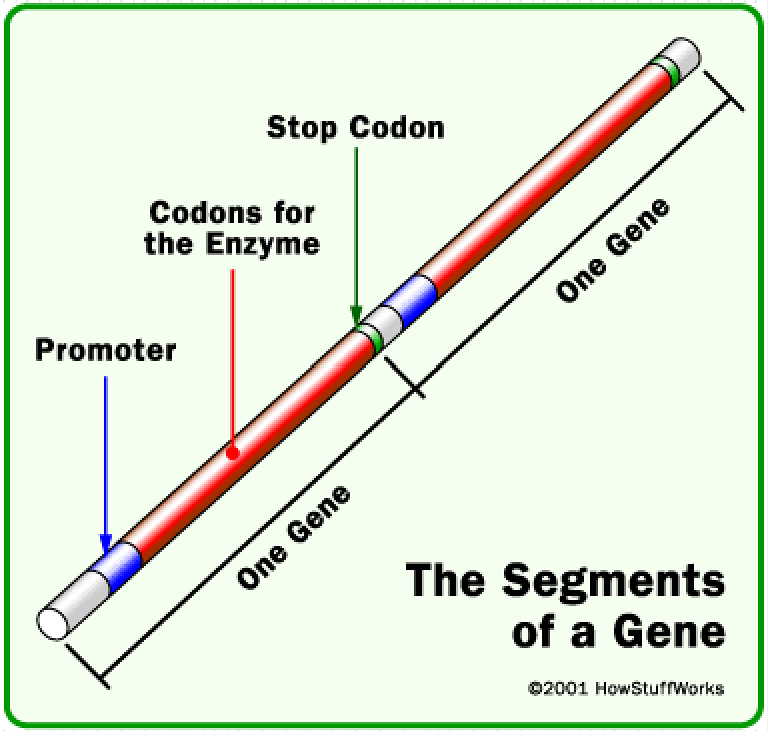
liver tumor mouse

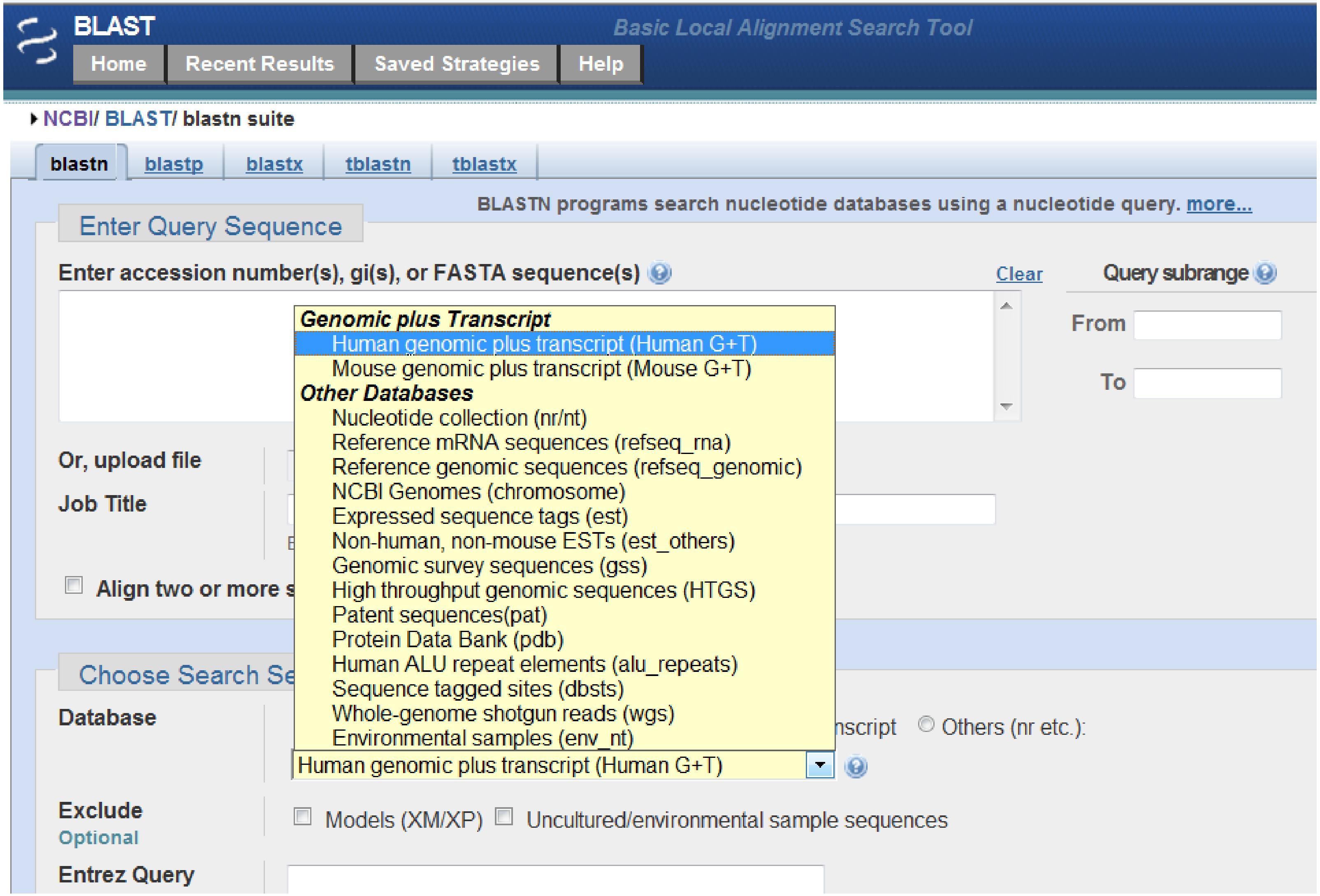
Seed tumor at liver of mouse-Surgery 種腫瘤在肝臟-開腹腔篇(一)

miss9ch

This video contains animal experiment content,
Viewer discretion is advise

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





BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file

Job Title

□ Align two or more s

Choose Search Se

Database ...nscript ○ Others (nr etc.):

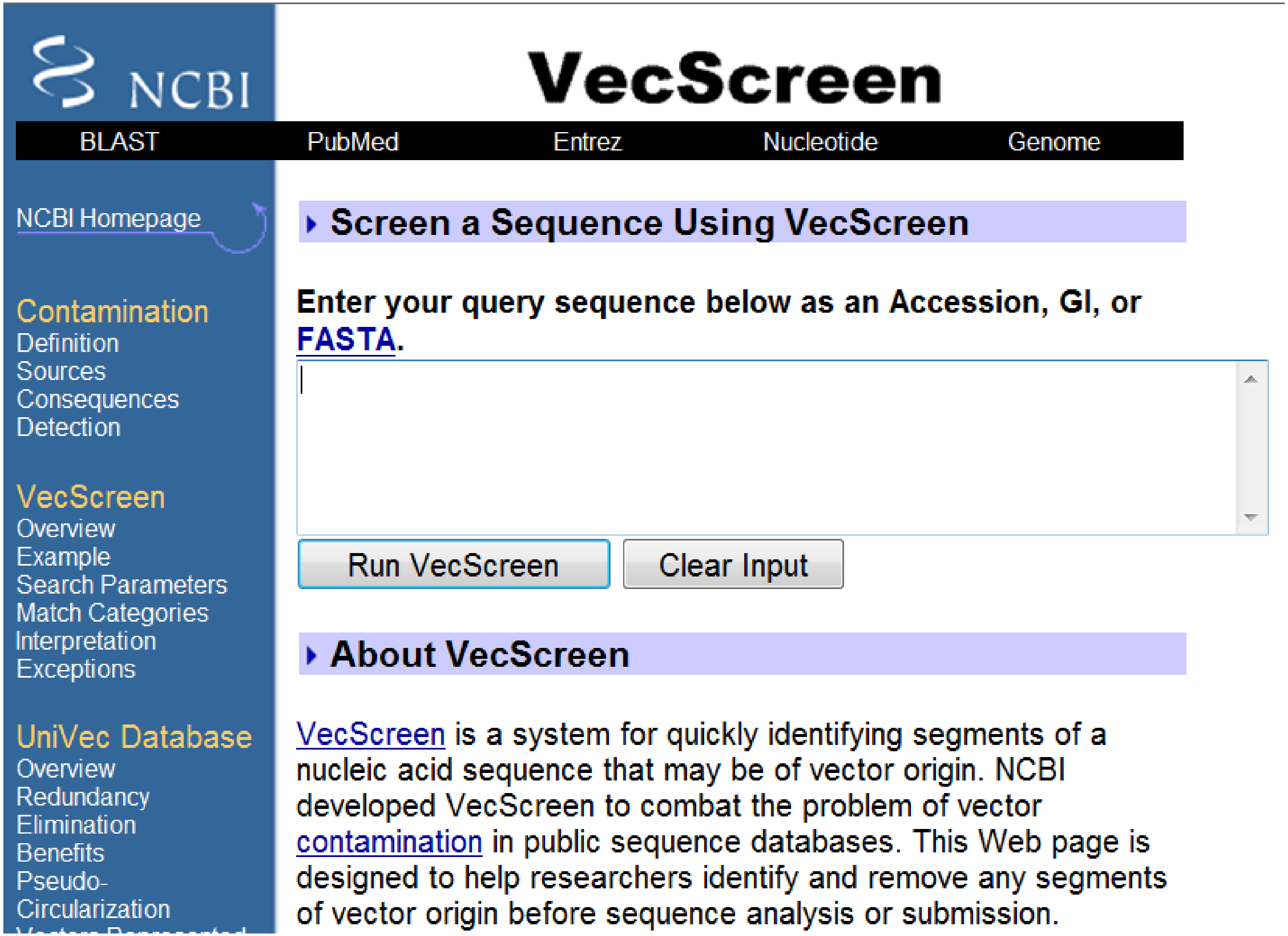
Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

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)

Exclude
Optional
□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query



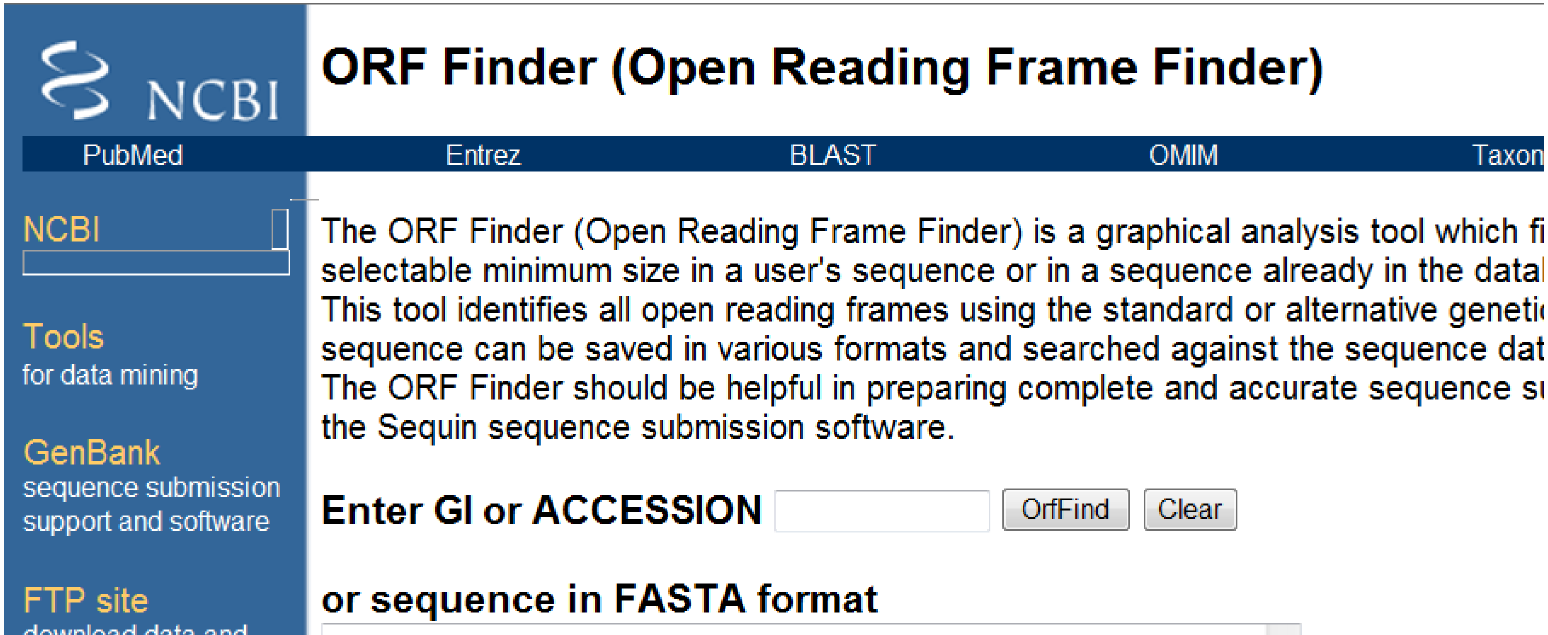
VecScreen

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.



NCBI

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxon

NCBI

Tools
for data mining

GenBank
sequence submission support and software

FTP site
download data and software

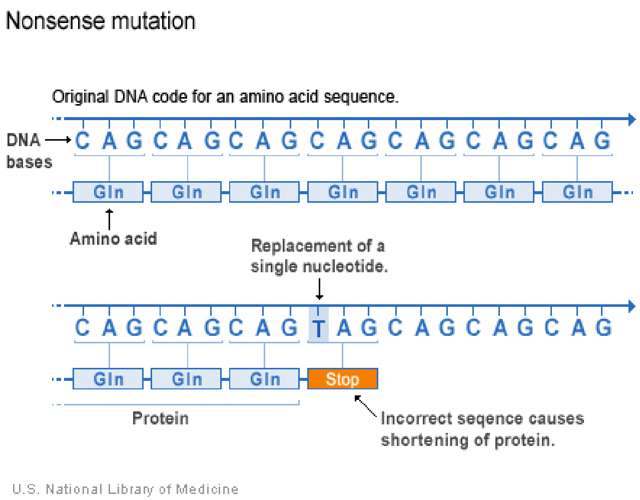
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which fi
selectable minimum size in a user's sequence or in a sequence already in the data
This tool identifies all open reading frames using the standard or alternative geneti
sequence can be saved in various formats and searched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

Enter GI or ACCESSION OrfFind Clear

or sequence in FASTA format

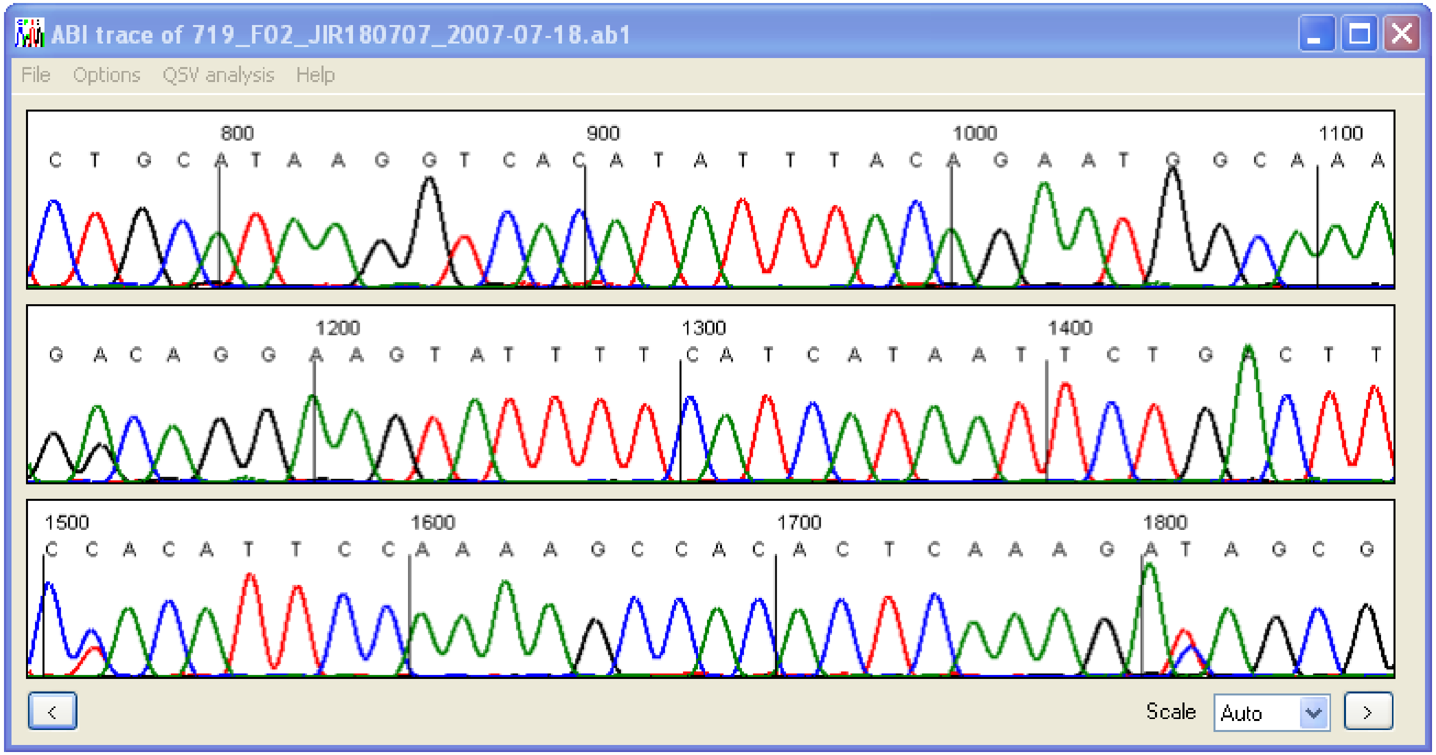
FROM: **TO:**

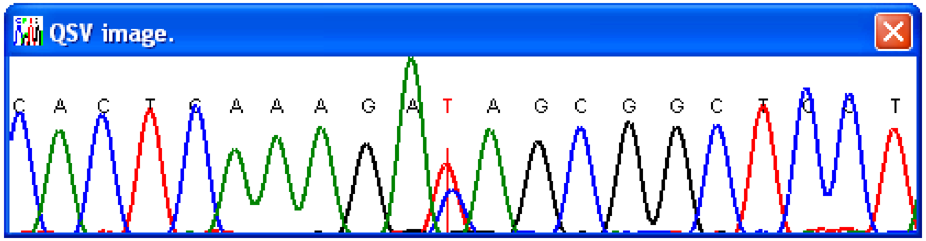
Genetic codes 1 Standard


Nonsense mutation

Original DNA code for an amino acid sequence.
DNA bases
C A G C A G C A G C A G C A G C A G C A G
Gln Gln Gln Gln Gln Gln Gln
Amino acid
Replacement of a single nucleotide.
C A G C A G C A G T A G C A G C A G C A G
Gln Gln Gln Stop
Protein
Incorrect seqence causes shortening of protein.

U.S. National Library of Medicine


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


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

When Cloned by Emails – get the map & confirmed

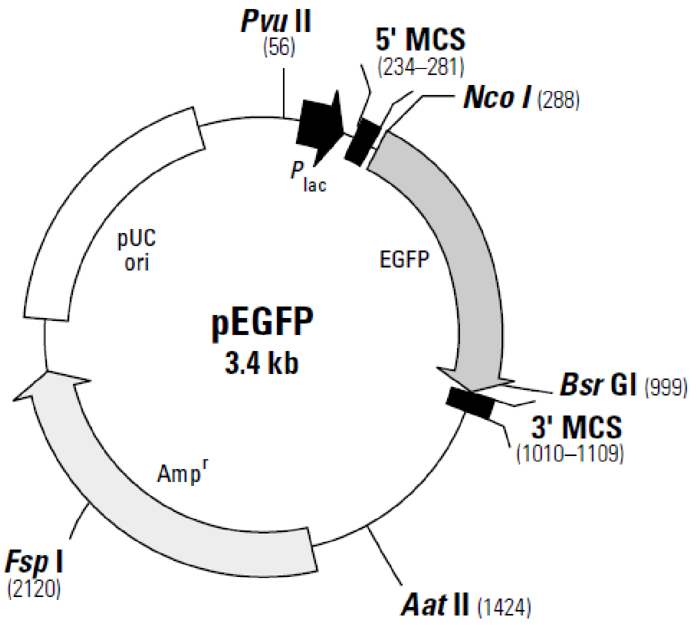
Specific EGFP Monoclonal Antibody for Westerns, IP and IC

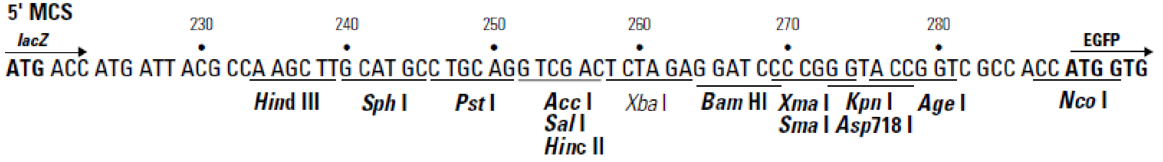
Visit our website for more details! click here...

pEGFP Vector Information

PT3078-5

Catalog #6077-1

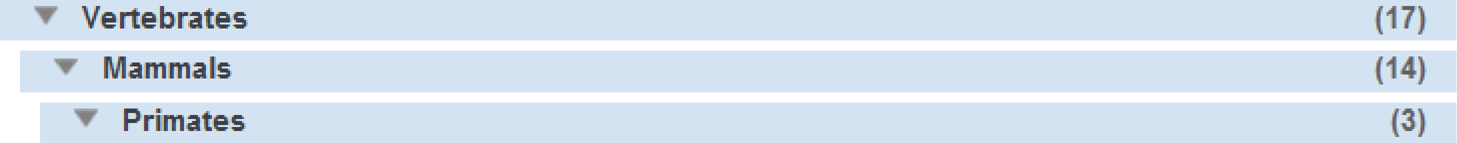




ATG ACC ATG ATT ACG CCA AGC TTG CAT GCC TGC AGG TCG ACT CTA GAG GAT CCC CGG GTA CCG GTC GCC ACC ATG GTG

Q: How to Get a Specific Sequence from Genome Databases

Genome Biology

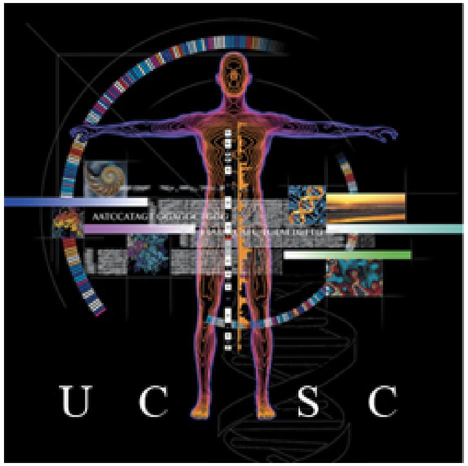


▼ Vertebrates	(17)
▼ Mammals	(14)
▼ Primates	(3)

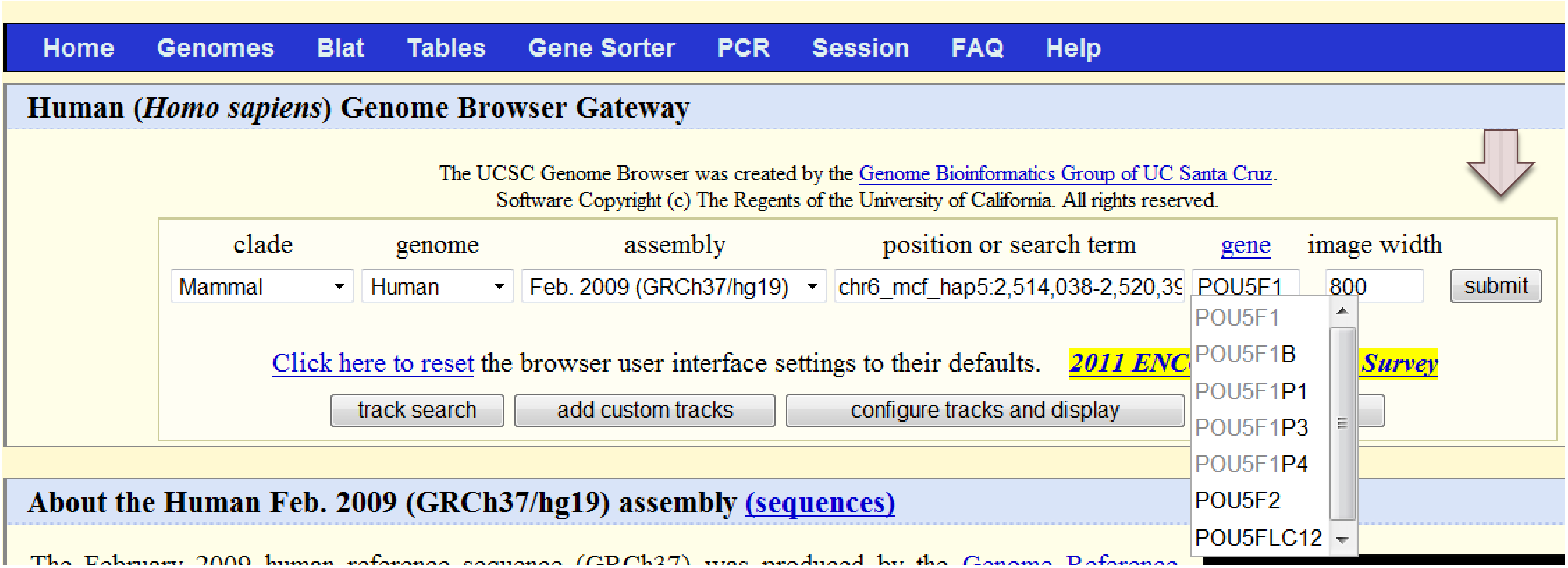
Map Viewer, NCBI

Genome Browser, UCSC

Ensembl Genome Browser, EBI



e!Ensembl



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

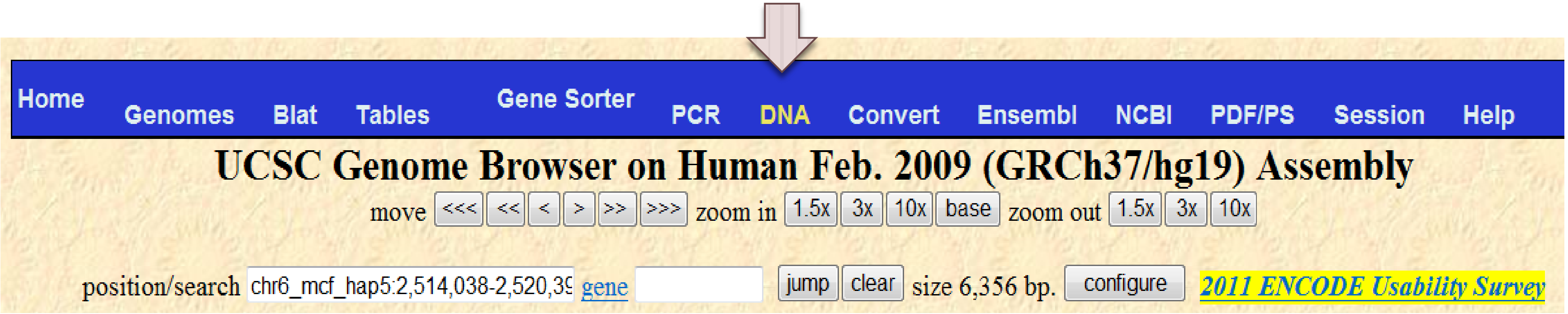
clade	genome	assembly	position or search term	gene	image width	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr6_mcf_hap5:2,514,038-2,520,39	POU5F1	800	submit

POU5F1
POU5F1B
POU5F1P1
POU5F1P3
POU5F1P4
POU5F2
POU5FLC12

Click here to reset the browser user interface settings to their defaults. 2011 ENC Survey

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

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

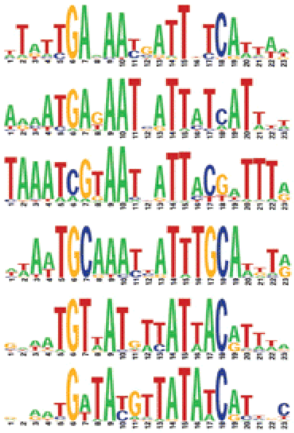


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

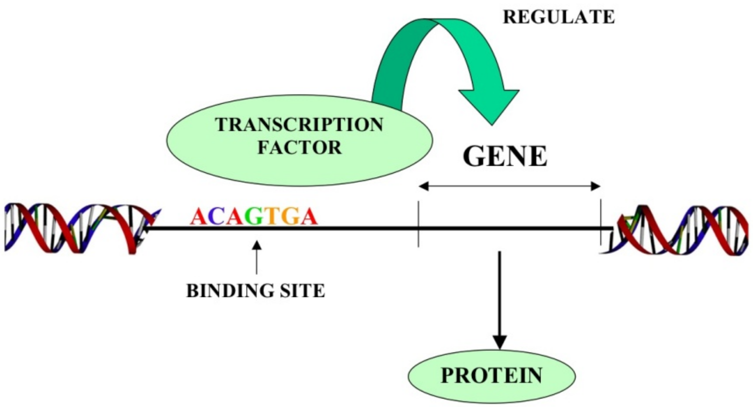
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

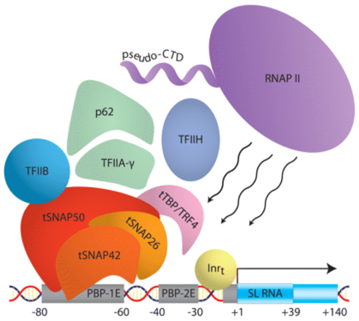
position/search chr6_mcf_hap5:2,514,038-2,520,39 gene jump clear size 6,356 bp. configure 2011 ENCODE Usability Survey

Q: How to Identify Potential Regulators?

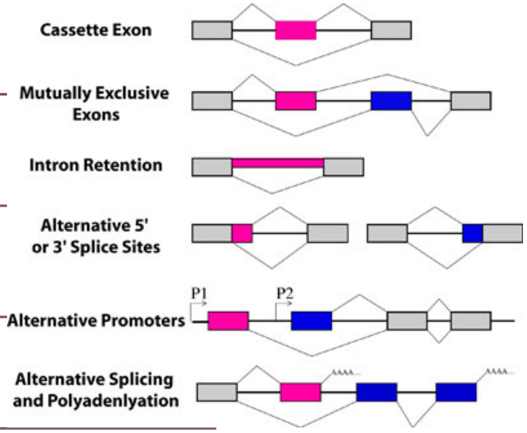


Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.

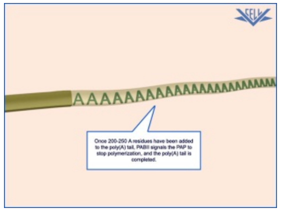


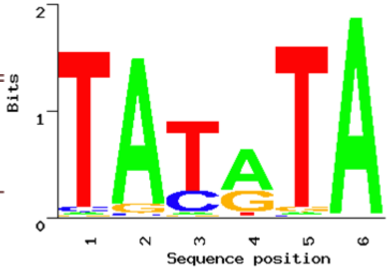


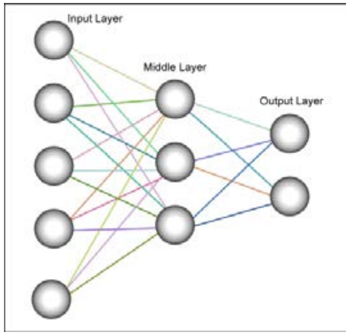
Feature-Based Methods

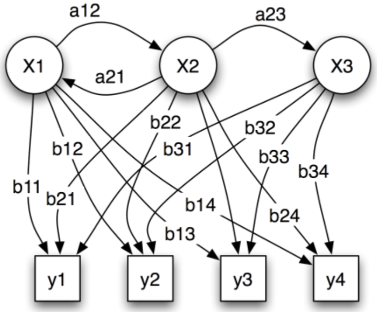


Based on identifying **gene signals**	**Promoter elements**
	Splice sites
	Start/stop codons
	PolyA sites…
Wide range of **methods**	Consensus sequences
	Weight matrices
	Neural networks (NNs) Decision trees
	Hidden Markov Models (HMMs)


Once 200-250 A residues have been added to the poly(A) tail, PABII signals the PAP to stop polymerization, and the poly(A) tail is completed.



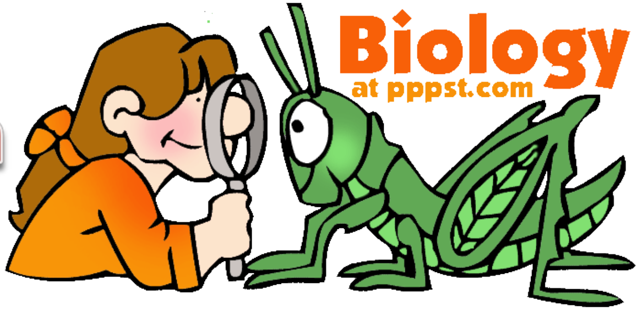




Promoter Databases and sites for analysis, prediction and search

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

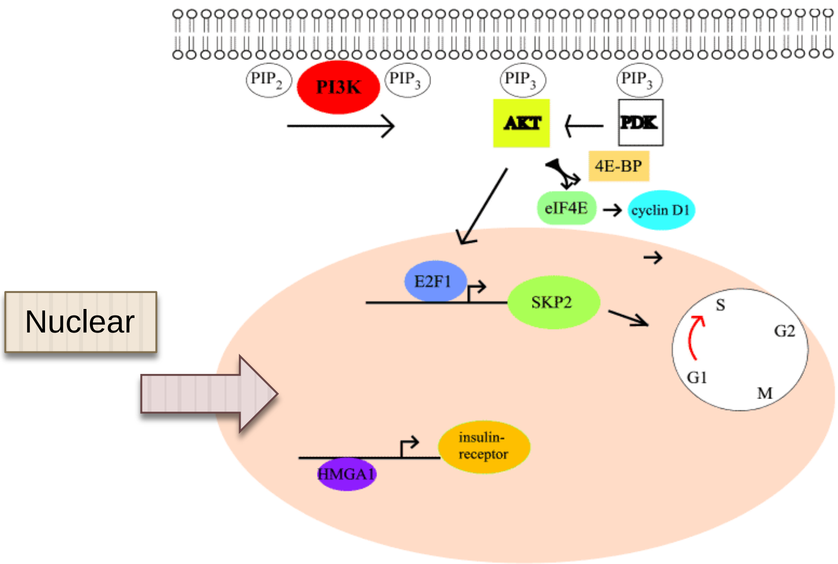
Pattern-driven



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

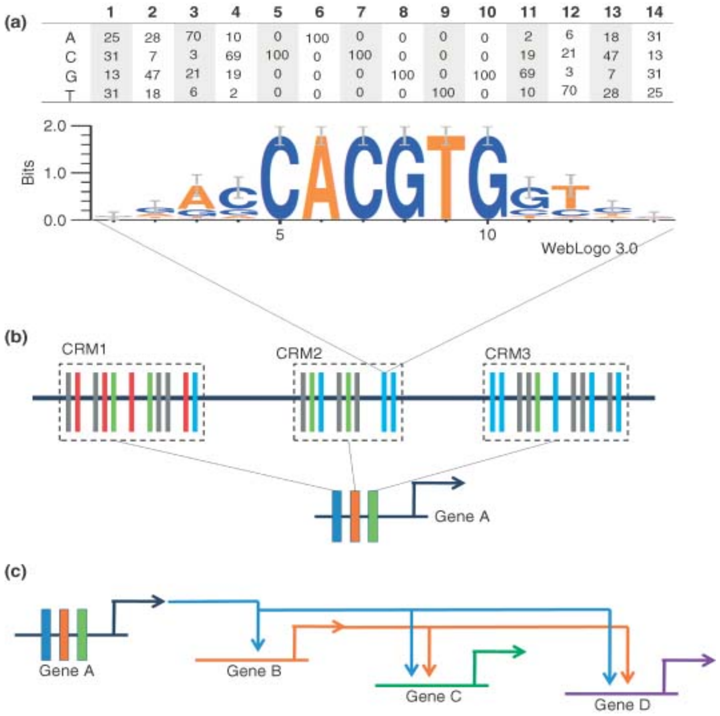
- Tend to produce huge numbers of **false-positive**

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



Nuclear

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

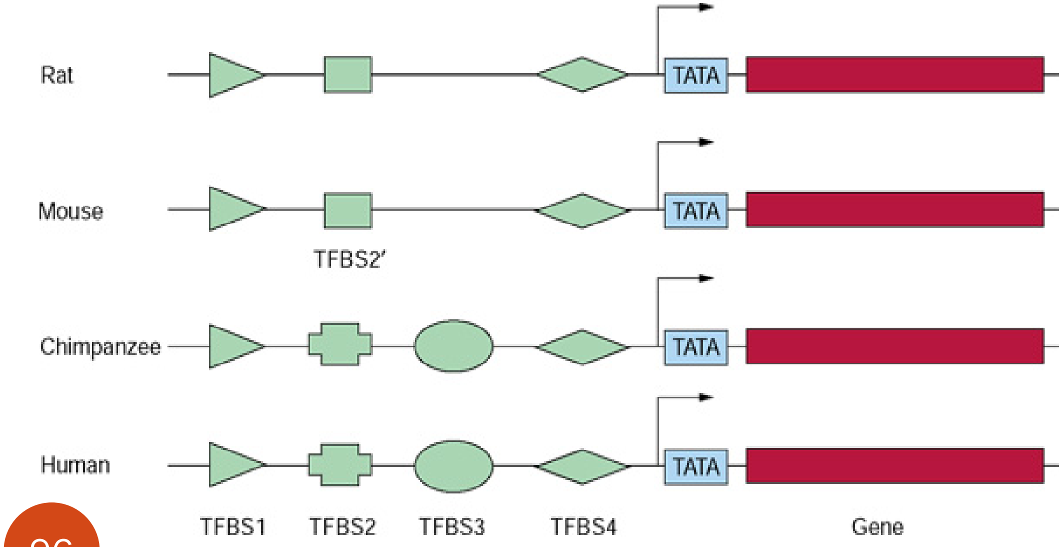
(Do the **wet lab** experiments!!!)

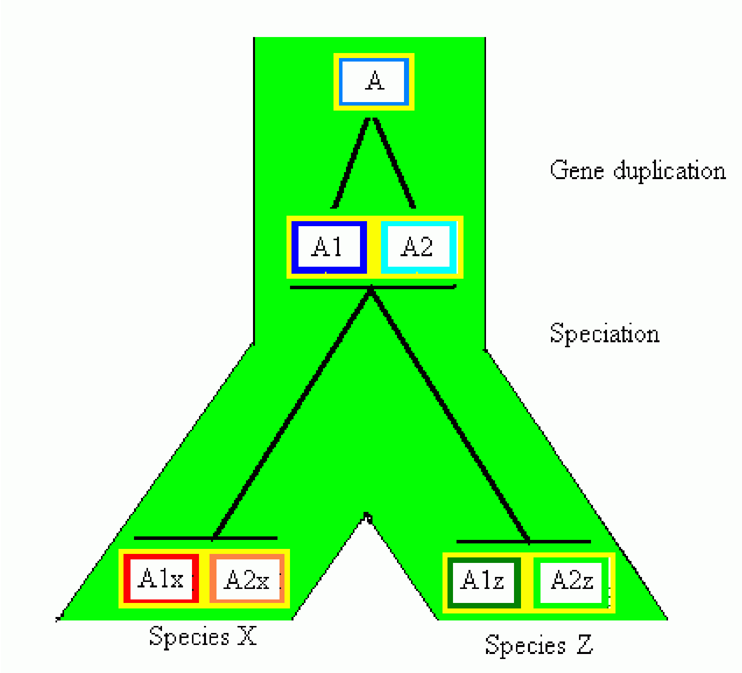
Eukaryotes: clusters of TFBSs are common	**Probability** of “real” binding site increases if annotated **transcription start site (TSS) nearby** • But **NOT** for enhancers • Only **a small fraction of TSSs** have been experimentally mapped	**Comparative** promoter mapping

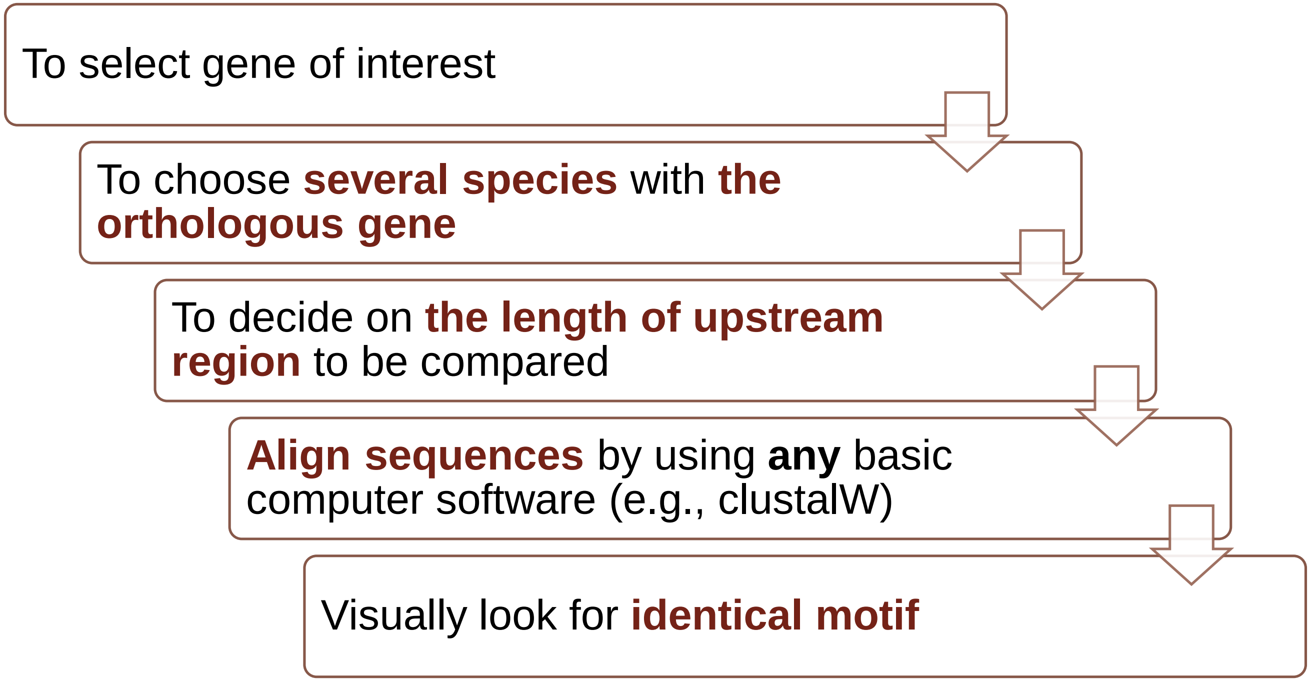
Phylogenetic Footprinting

Patterns of gene regulation are often conserved across species

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







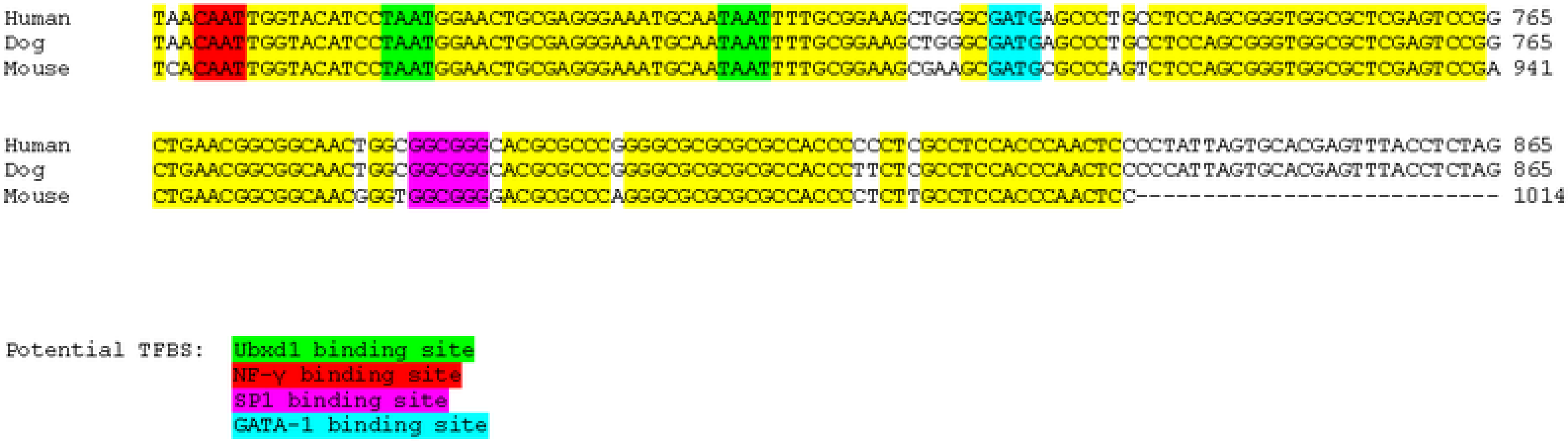
To select gene of interest

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

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

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

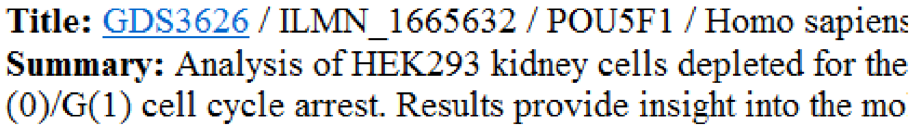
Visually look for **identical motif**



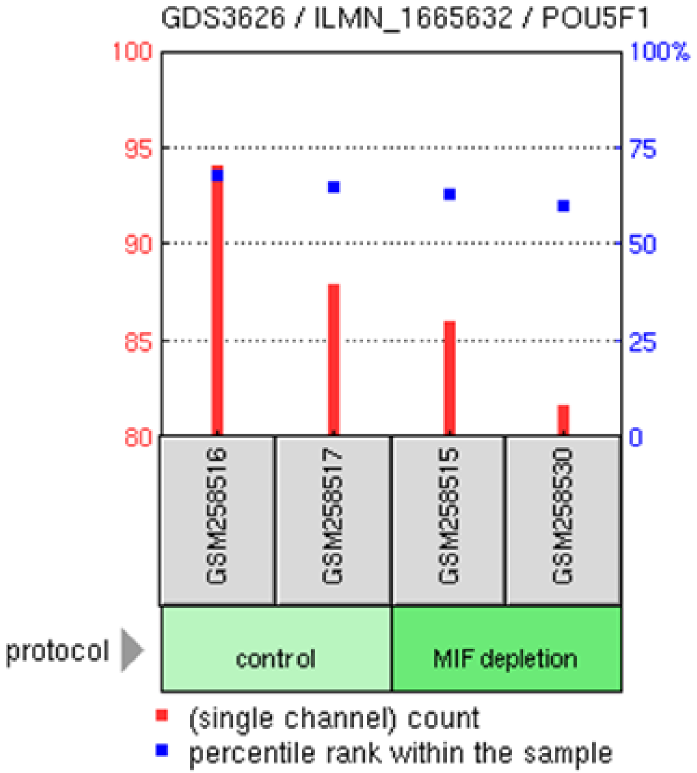
```
Human    TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Dog      TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Mouse    TCACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCGAAGCGATGCGCCCAGTCTCCAGCGGGTGGCGCTCGAGTCCGA 941

Human    CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCGCCACCCCCCCTCGCCTCCACCCAACTCCCCTATTAGTGCACGAGTTTACCTCTAG 865
Dog      CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCGCCACCCCTTCTCGCCTCCACCCAACTCCCCCATTAGTGCACGAGTTTACCTCTAG 865
Mouse    CTGAACGGCGGCAACGGGTGGCGGGGACGCGCCCAGGGCGCGCGCGCCACCCCCTCTTGCCTCCACCCAACTCC--------------------------- 1014


Potential TFBS:  Ubxd1 binding site
                 NF-Y binding site
                 SP1 binding site
                 GATA-1 binding site
```

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

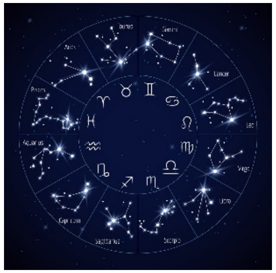
One More Trick - Coregulation

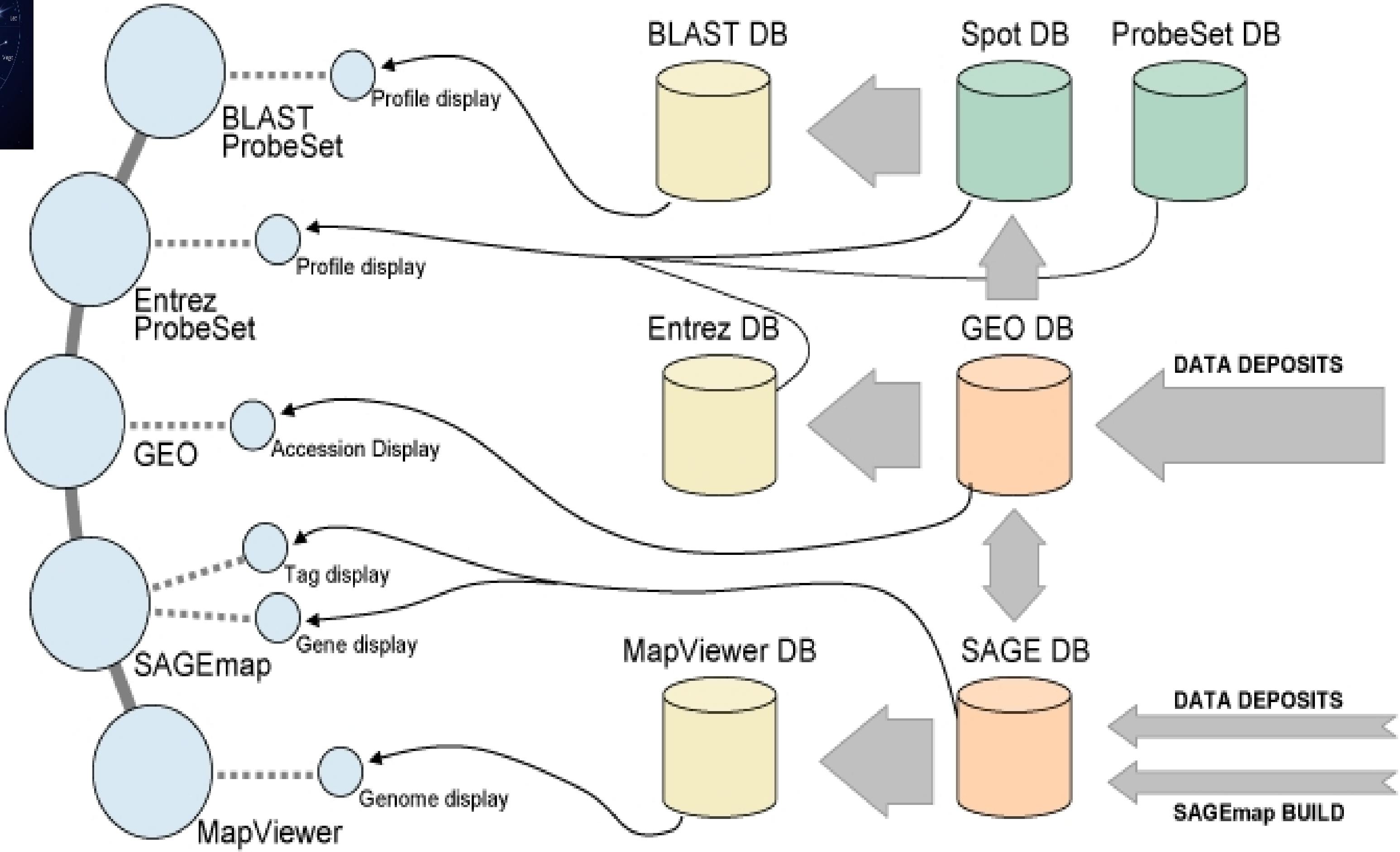


Title: GDS3626 / ILMN_1665632 / POU5F1 / Homo sapiens
Summary: Analysis of HEK293 kidney cells depleted for the
(0)/G(1) cell cycle arrest. Results provide insight into the mo

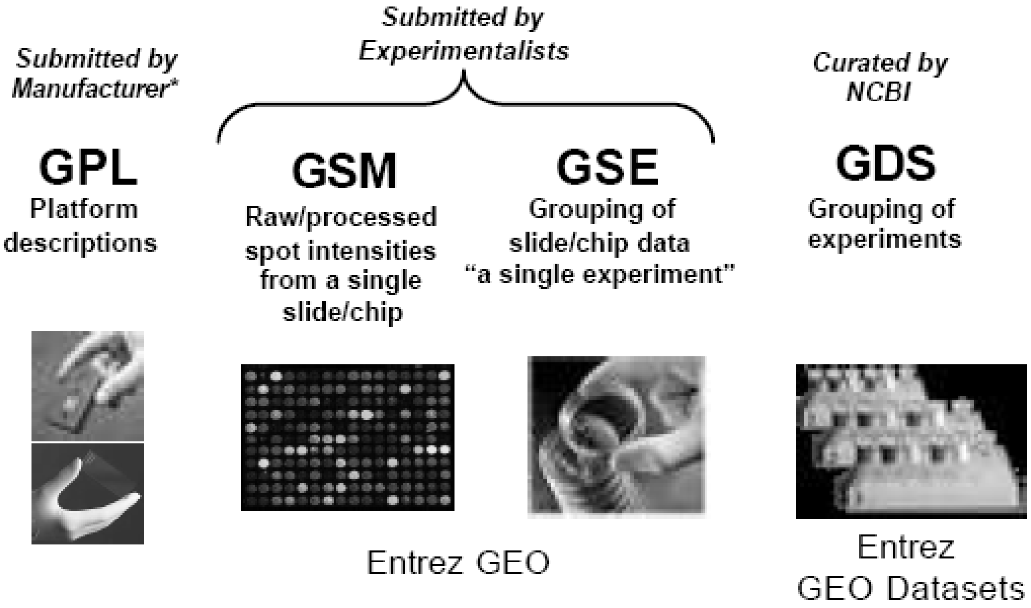


Constellation of NCBI Gene Expression Resources





Gene Expression Omnibus (GEO) (1)



*Submitted by Manufacturer**

GPL

Platform descriptions

Submitted by Experimentalists

GSM

Raw/processed spot intensities from a single slide/chip

GSE

Grouping of slide/chip data "a single experiment"

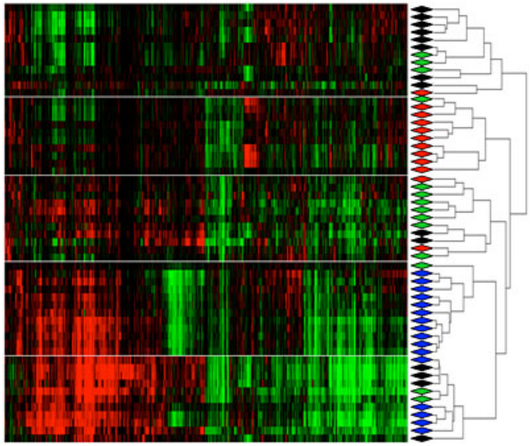
Entrez GEO

Curated by NCBI

GDS

Grouping of experiments

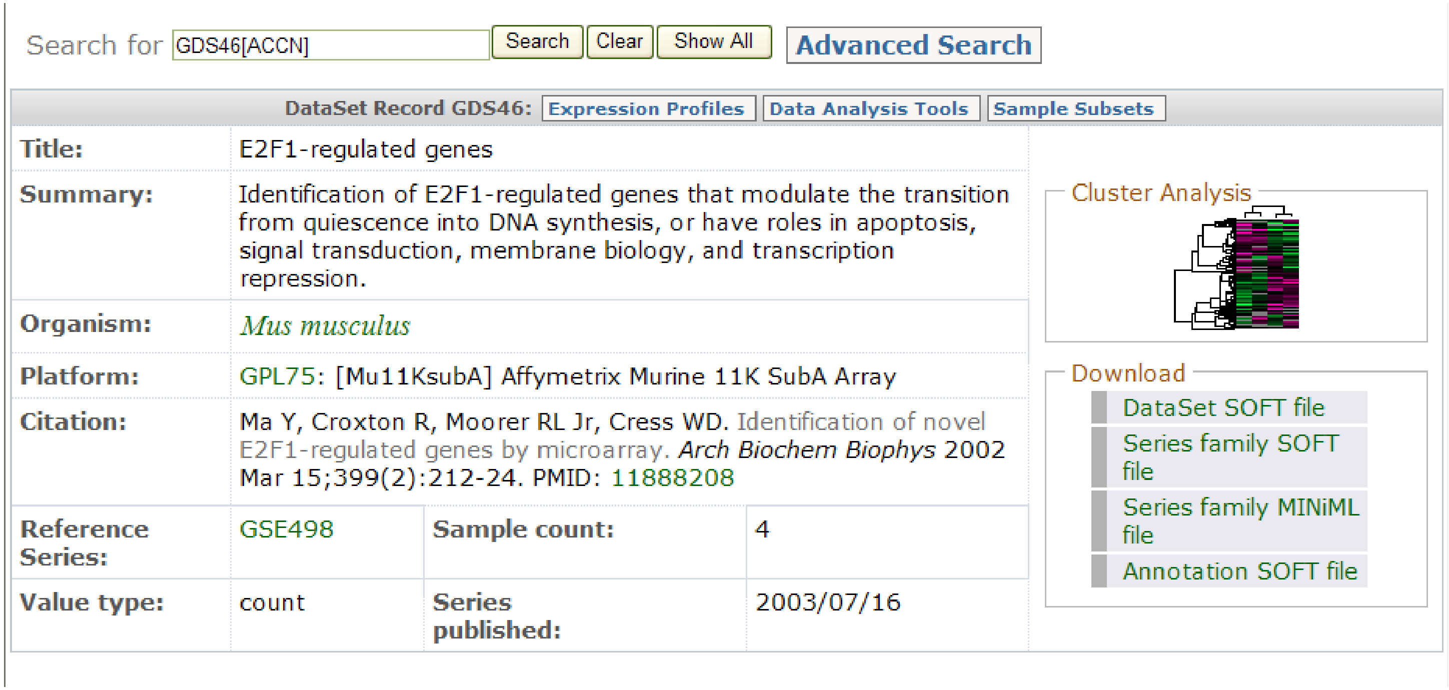
Entrez GEO Datasets



Gene Expression Omnibus (GEO) (2)

- Search GEO Profiles: POU5F1
  - Or **Limit**, **Preview / Index**

- GDS vs. GSE

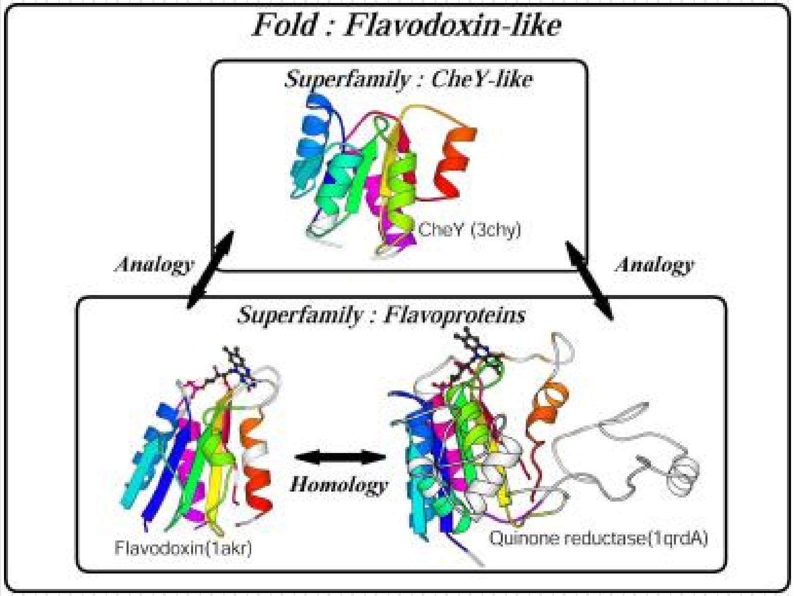


Search for GDS46[ACCN]

DataSet Record GDS46:

Title:	E2F1-regulated genes		
Summary:	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.		
Organism:	*Mus musculus*		
Platform:	GPL75: [Mu11KsubA] Affymetrix Murine 11K SubA Array		
Citation:	Ma Y, Croxton R, Moorer RL Jr, Cress WD. Identification of novel E2F1-regulated genes by microarray. *Arch Biochem Biophys* 2002 Mar 15;399(2):212-24. PMID: 11888208		
Reference Series:	GSE498	**Sample count:**	4
Value type:	count	**Series published:**	2003/07/16

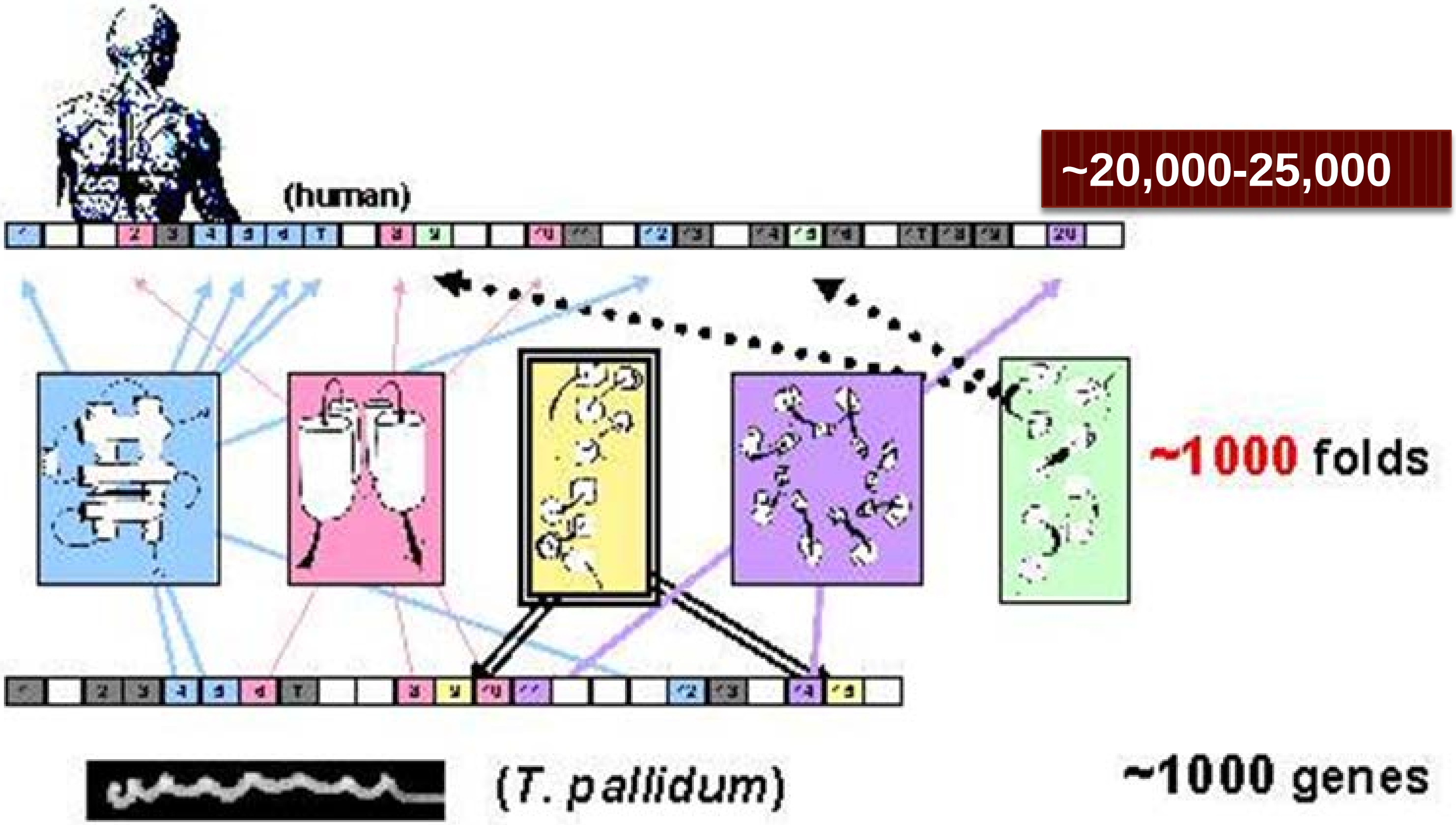
Q: Can You Speculate the Function of YFG from Structure Similarity?



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
		90	
		80	
		70	
Increasing divergence		60	
		50	Consensus Methods
		40	
	Twilight Zone	30	Profile Methods
		20	
Distant relationship	Midnight Zone	10	Structure Prediction
		0	

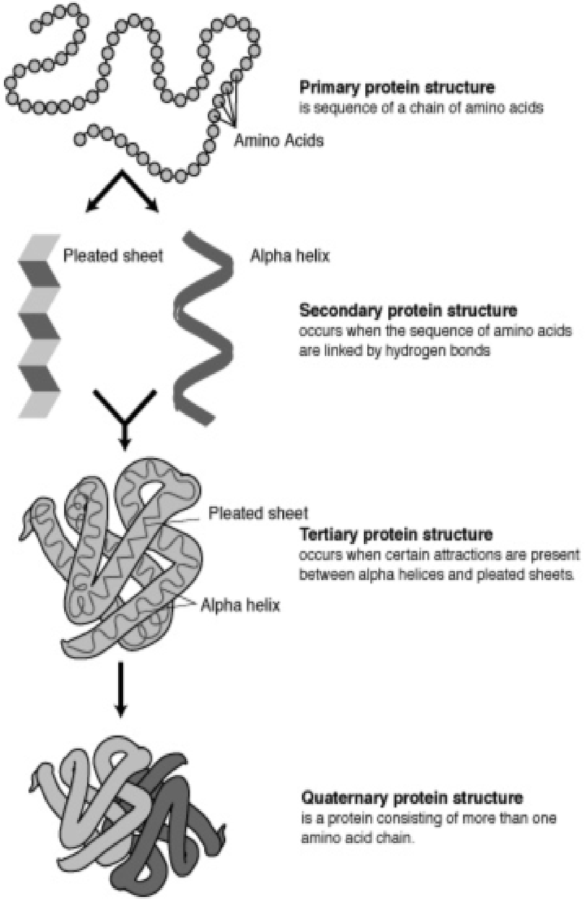
Simplifying Genomes with Folds, Pathways

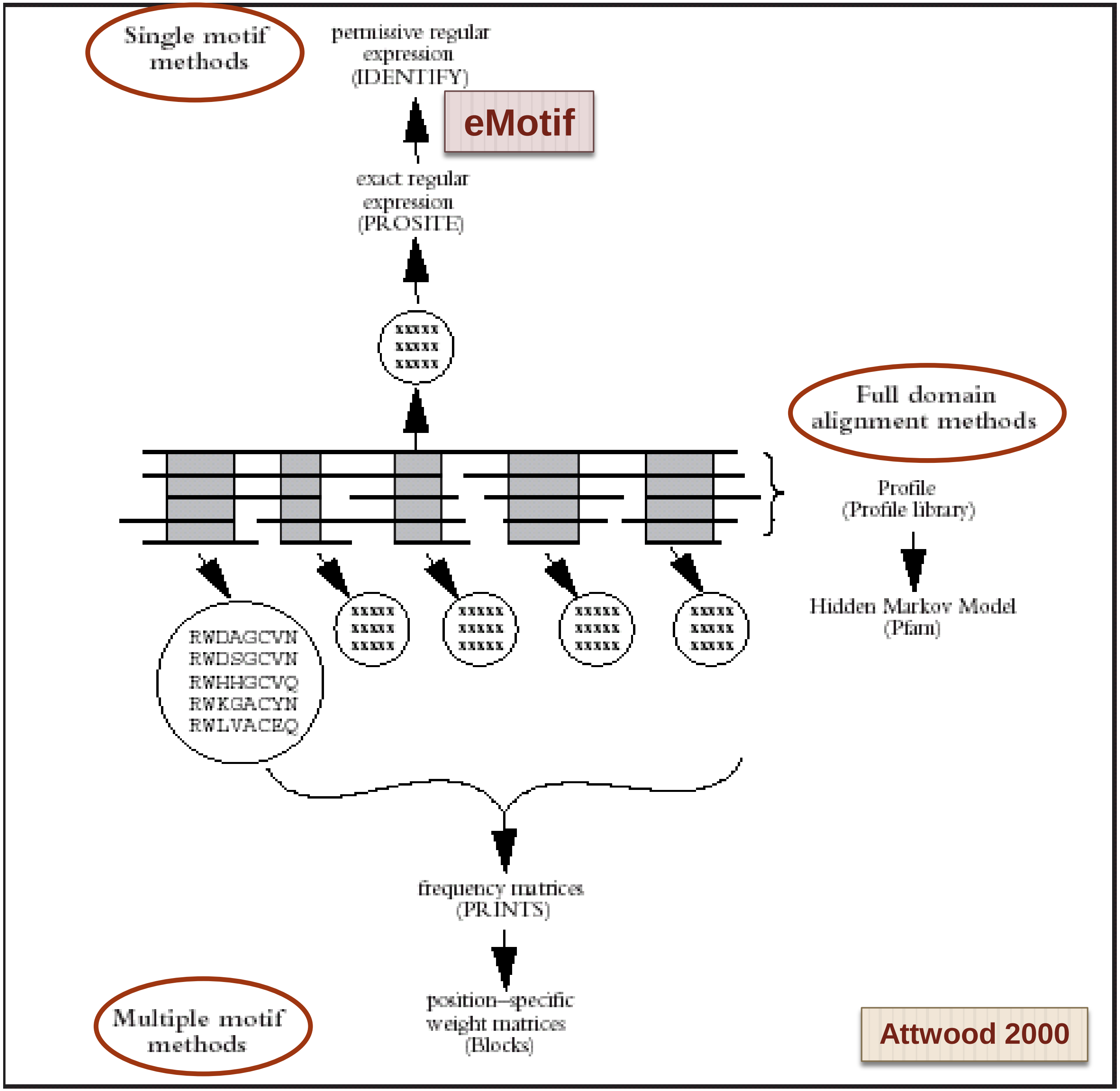


Significance: fold # << sequence ##

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 @, *, #


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

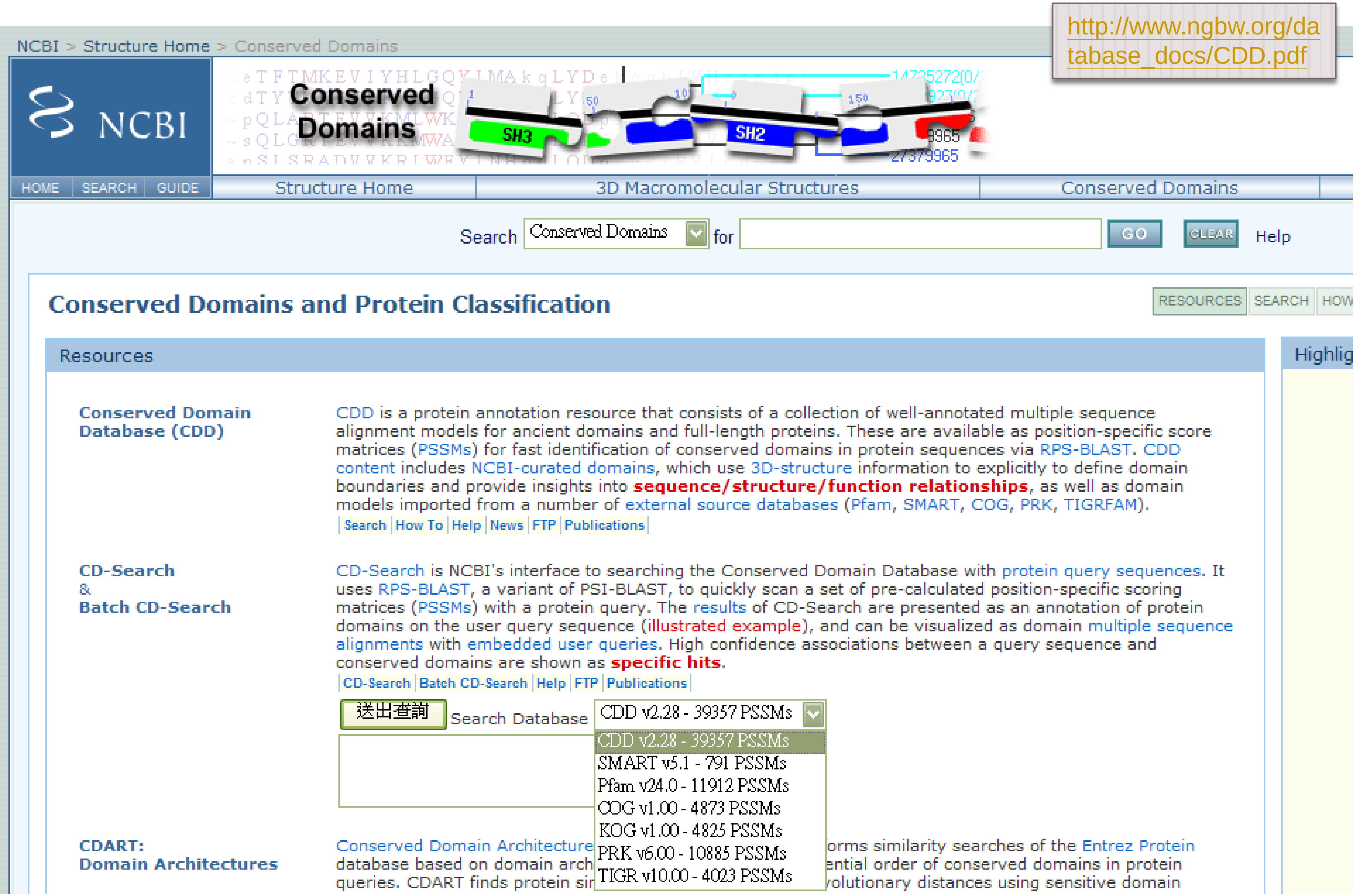

Single motif methods
permissive regular expression (IDENTIFY)
eMotif
exact regular expression (PROSITE)
xxxxx xxxxx xxxxx
Full domain alignment methods
Profile (Profile library)
Hidden Markov Model (Pfam)
RWDAGCVN
RWDSGCVN
RWHHGCVQ
RWKGACYN
RWLVACEQ
xxxxx xxxxx xxxxx
xxxxx xxxxx xxxxx
xxxxx xxxxx xxxxx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

Attwood 2000

Major Secondary "Pattern" Database

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

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



Conserved Domains and Protein Classification

Resources

Conserved Domain Database (CDD)

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

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

CD-Search & Batch CD-Search

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

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

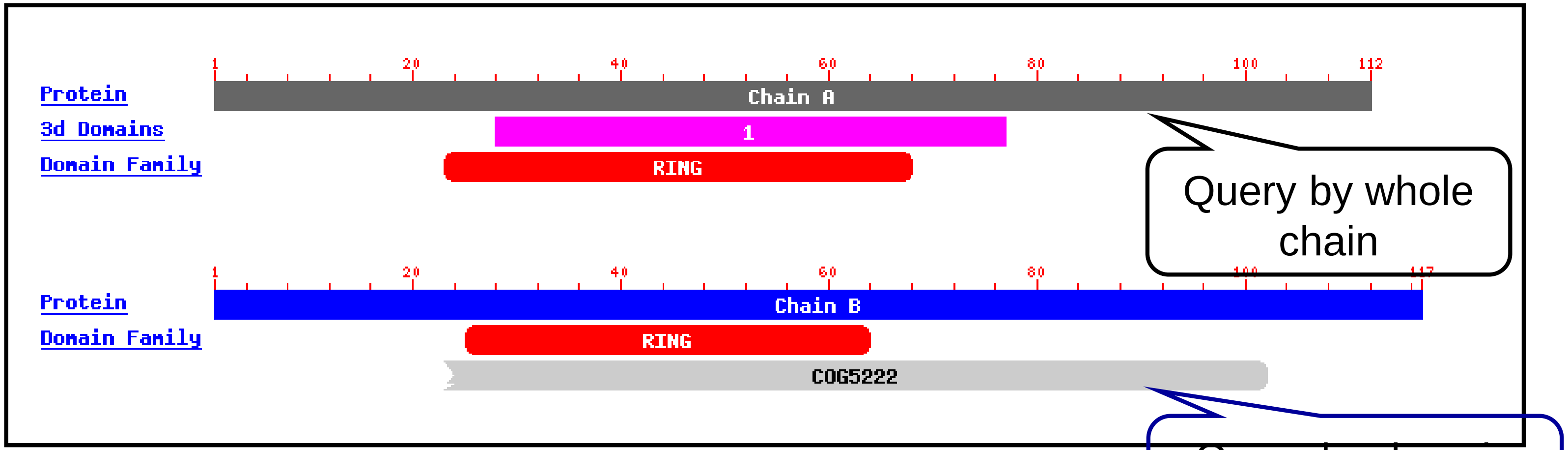
送出查詢 Search Database CDD v2.28 - 39357 PSSMs

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

CDART: Domain Architectures

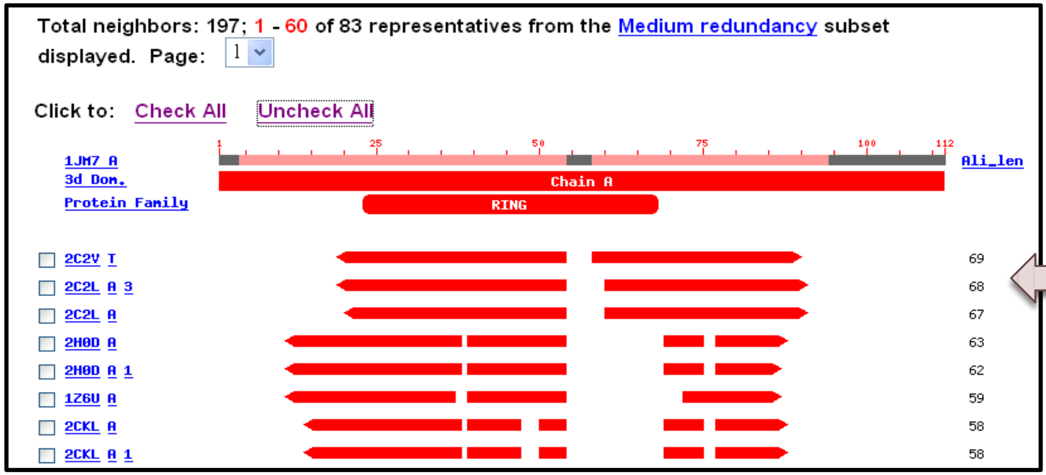
Conserved Domain Architecture … orms similarity searches of the Entrez Protein database based on domain arch… ential order of conserved domains in protein queries. CDART finds protein si… volutionary distances using sensitive domain

VAST: Query by Chain or 3D Domain



Query by whole chain

Query by domain COG5222



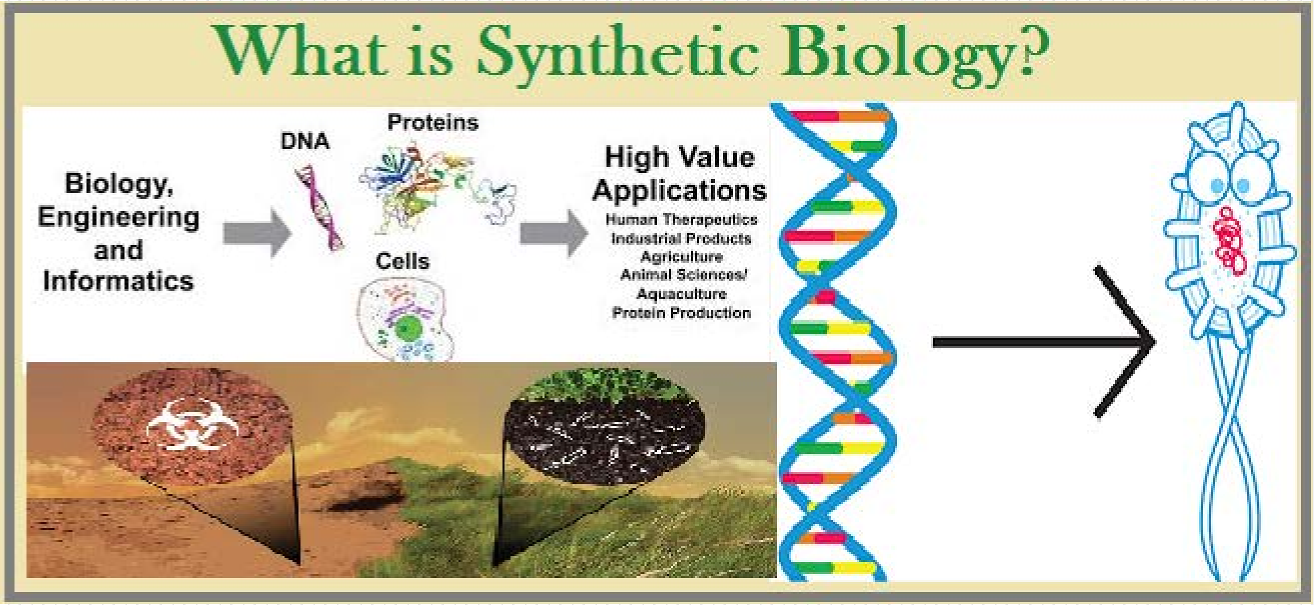
Total neighbors: 197; 1 - 60 of 83 representatives from the Medium redundancy subset displayed. Page: 1

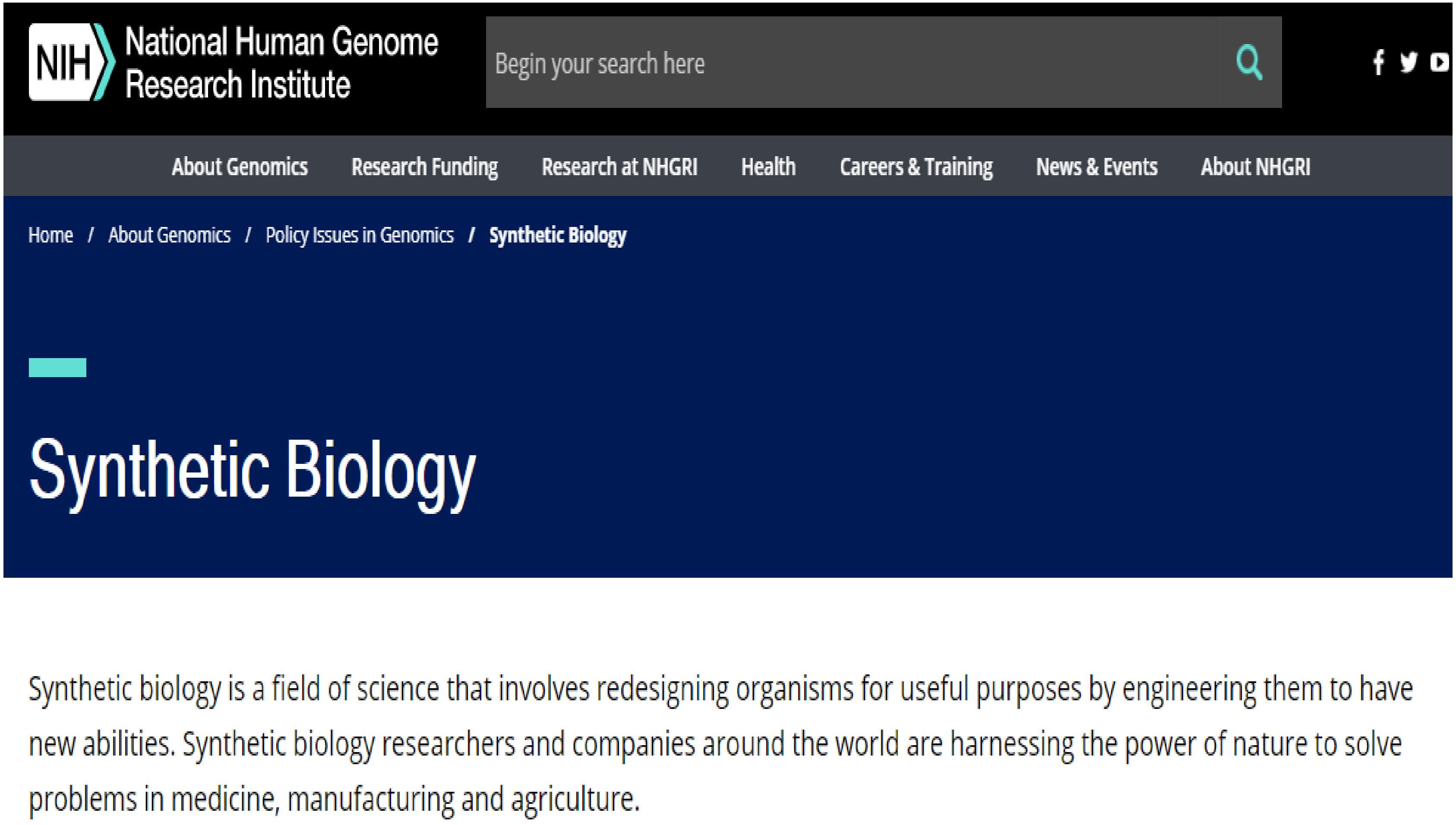
Click to: Check All Uncheck All

		Ali_len
□	2C2V T	69
□	2C2L A 3	68
□	2C2L A	67
□	2H0D A	63
□	2H0D A 1	62
□	1Z6U A	59
□	2CKL A	58
□	2CKL A 1	58

Not found with chain query

Synthetic Biology

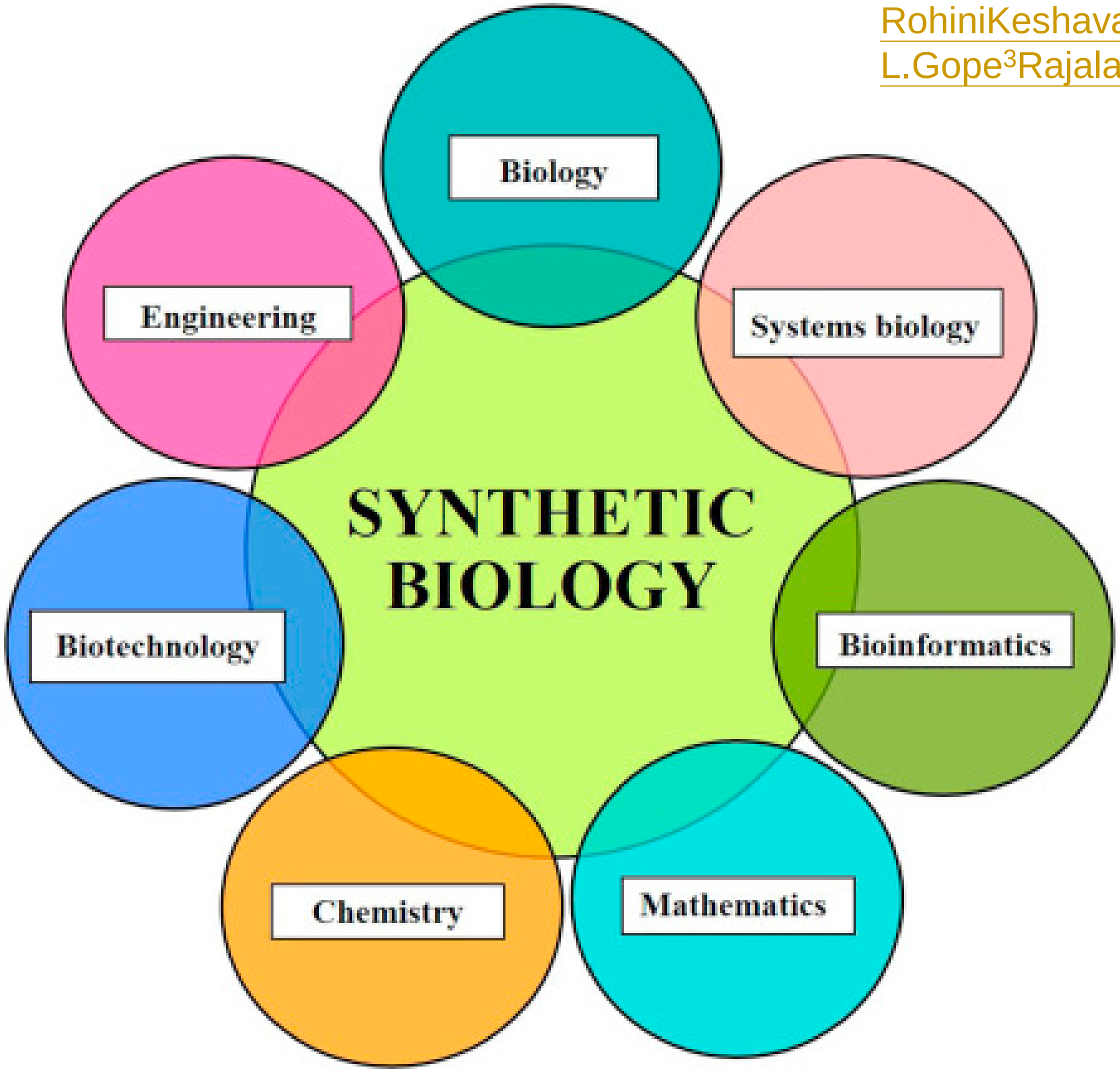




Synthetic Biology

Synthetic biology is a field of science that involves redesigning organisms for useful purposes by engineering them to have new abilities. Synthetic biology researchers and companies around the world are harnessing the power of nature to solve problems in medicine, manufacturing and agriculture.

Chapter 4 - Synthetic Biology: Overview and Applications Author links open overlay panel RohiniKeshava[1]RohanMitra[2]Mohan L.Gope[3]RajalakshmiGope[2]

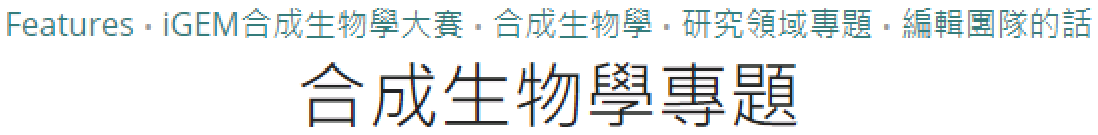


Omics Technologies and Bio-Engineering Towards Improving Quality of Life
2018, Pages 63-93

Example 1

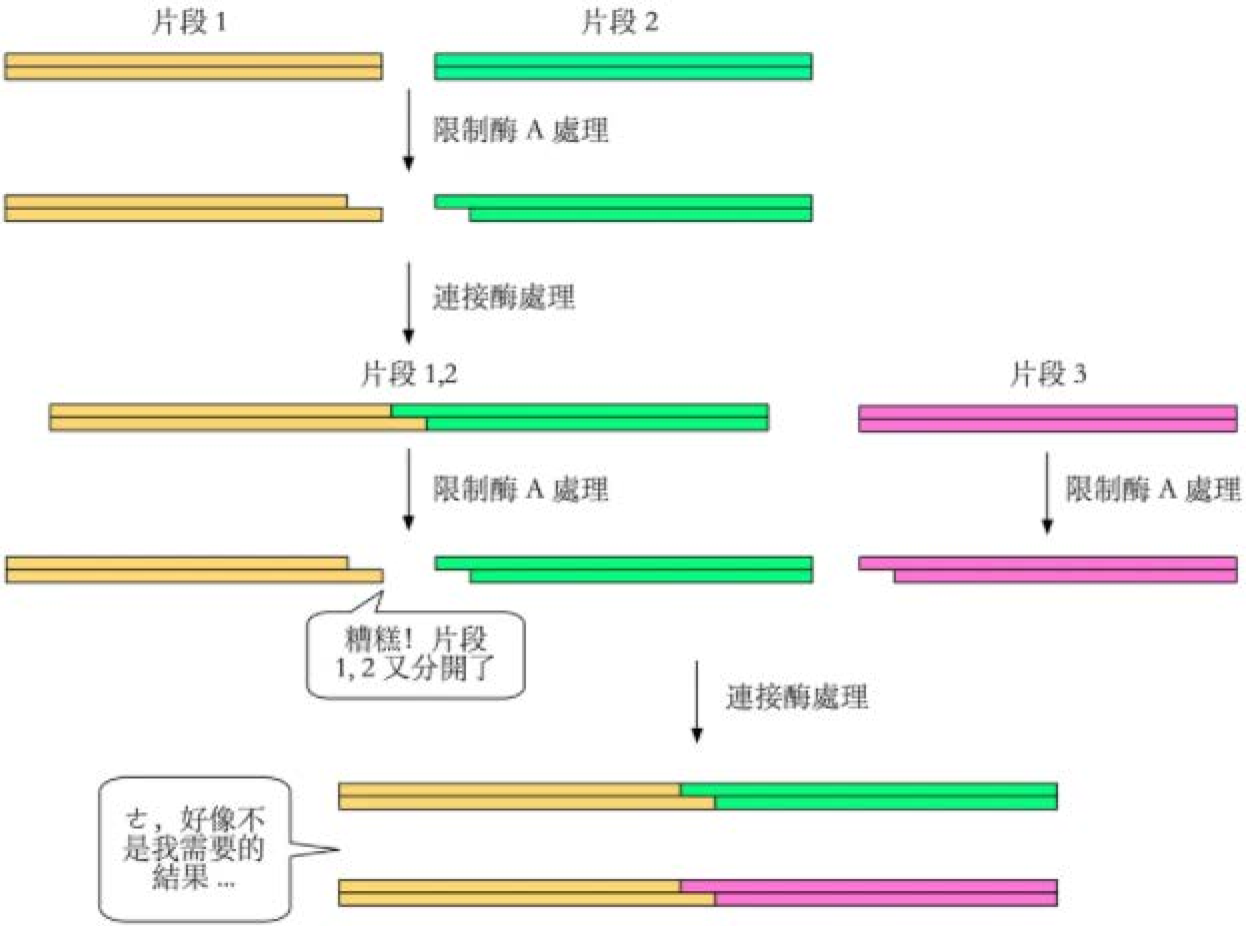
- 因為每合成一個鹼基對 (base pair, bp) DNA 的價格，三十年前要價數十至數百美元不等，而如今降低到只需要一美元或低於一美元，有人將這種現象比擬為生命科學研究上的摩爾定律。
  - DNA 合成技術的成熟，大大降低了DNA 合成的經濟門檻，也預告著大尺度基因體工程與合成生物學研究時代的來臨。

  - 2008 年，JCVI (J. Craig Venter Institute) 的研究人員用5000 ~ 7000 bp 大小的化學合成DNA 片段 (chemically synthesized DNA fragments)，以人工方式兩兩相連接組裝成一個 582,970 bp 的 *Mycoplasma genitalium* 細菌基因體。

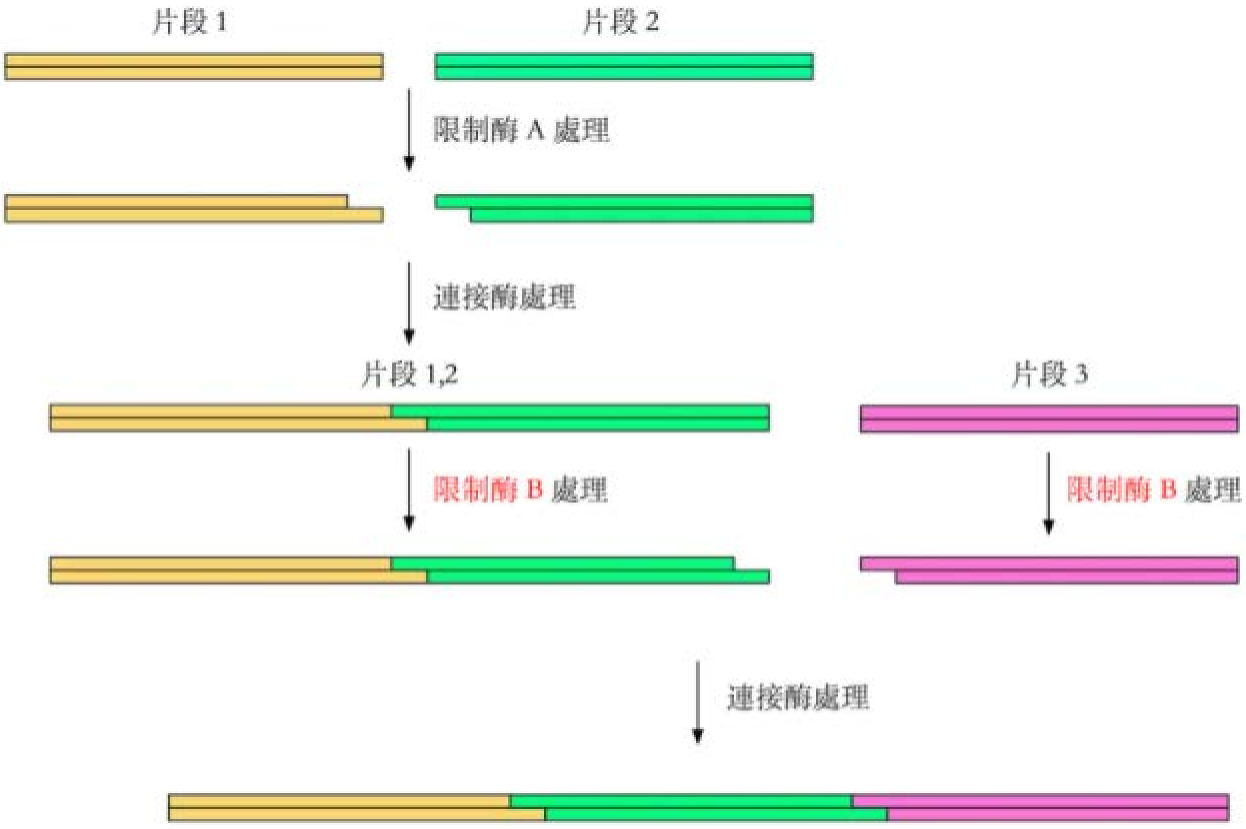


Features · iGEM合成生物學大賽 · 合成生物學 · 研究領域專題 · 編輯團隊的話

合成生物學專題


片段 1
片段 2
限制酶 A 處理
連接酶處理
片段 1,2
片段 3
限制酶 A 處理
限制酶 A 處理
糟糕！ 片段 1, 2 又分開了
連接酶處理
ㄜ，好像不是我需要的結果 ...

圖一 失敗的三段組裝



圖二 成功的三段組裝

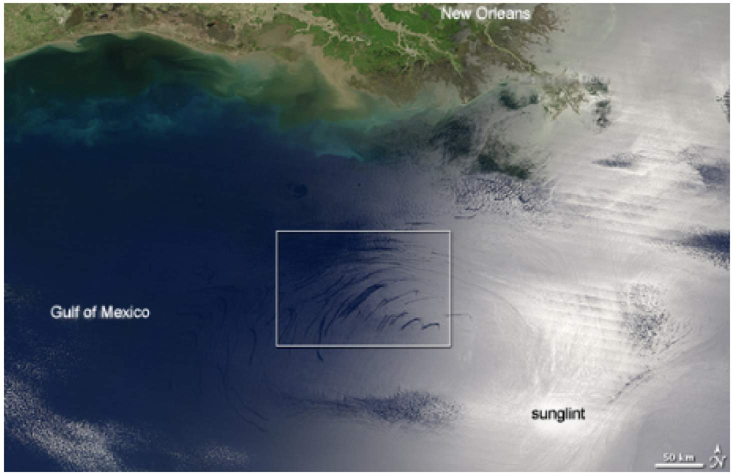
Example 2

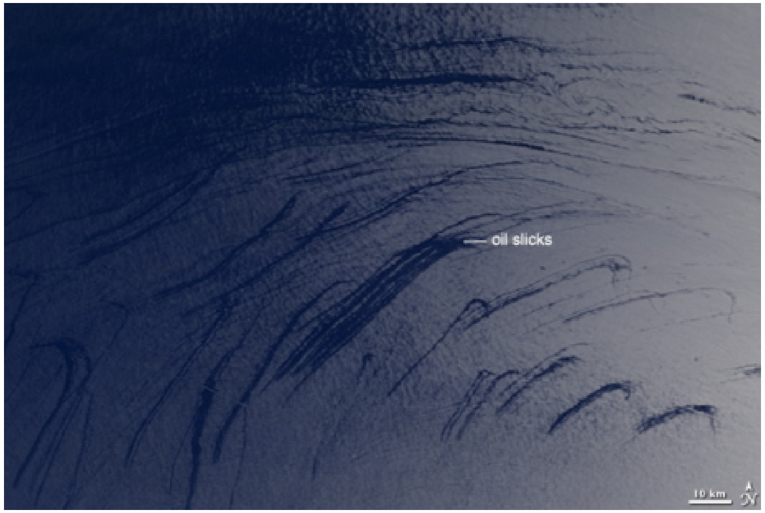
- Tom Knight 教授提出一種標準的 DNA 片段的組裝方式[9,10]，在每次的組裝可以使用相同的方式，不需要再費心選擇每次組裝使用的限制酶酵素。這樣的組裝方式，讓DNA 片段可以像積木一樣，一個片段一個片段一直連續組裝下去，生物零件 (biological Part) 的概念就因此誕生。
- 將生物 DNA 片段零件化，是工程思維應用在分子生物學的一個重大發明。因此，透過生物零件的定義與標準化的組裝方式，我們可以進一步組裝生物設備 (biological device)，或更進一步可以組裝一個生物系統 (biological system)，形成一個由生物零件為基礎的工程框架[11]。

Example 3: microorganisms harnessed for bioremediation

While invisible up close, microscopic oil **slicks**浮油from natural seeps滲透are visible from **space** because cohesion凝聚 between oil molecules flattens wave action to **form smooth areas** on the water (2010, BP)

©RB-DESKKART www.welt-atlas.de
Waco
Buffalo
Alexandria
Natchez
MISSISSIPPI
Hattiesburg
ALABAMA
Trifton
GEORGIA
Valdosta
Brunswick
AUSTIN
TEXAS
San Antonio
Port Lavaca
Beaumont
Houston
Galveston
BATON ROUGE
Lafayette
LOUISIANA
Morgan City
Gulfport
Biloxi
Mobile
Pensacola
New Orleans
Mississippi
TALLAHASSEE
Panama City
Lake City
Jacksonville
FLORIDA
USA
Daytona Beach
Orlando
Clearwater
St. Petersburg
Tampa
Sarasota
Fort Pierce
Fort Myers
Fort Lauderdale
Naples
Miami
Key West
Corpus Christi
Laredo
McAllen
Brownsville
GULF OF MEXICO
TROPIC OF CANCER
LA HABANA (HAVANNA)
Matanzas
Pinar del Río
CUBA
Tampico
MEXICO
Progreso
Río Lagartos
Mérida
Cancún
CARIBBEAN SEA
CIUDAD DE MEXICO (MEXICO CITY)
Campeche
Veracruz
Ciudad del Carmen
Puebla
Coatzacoalcos
500 km

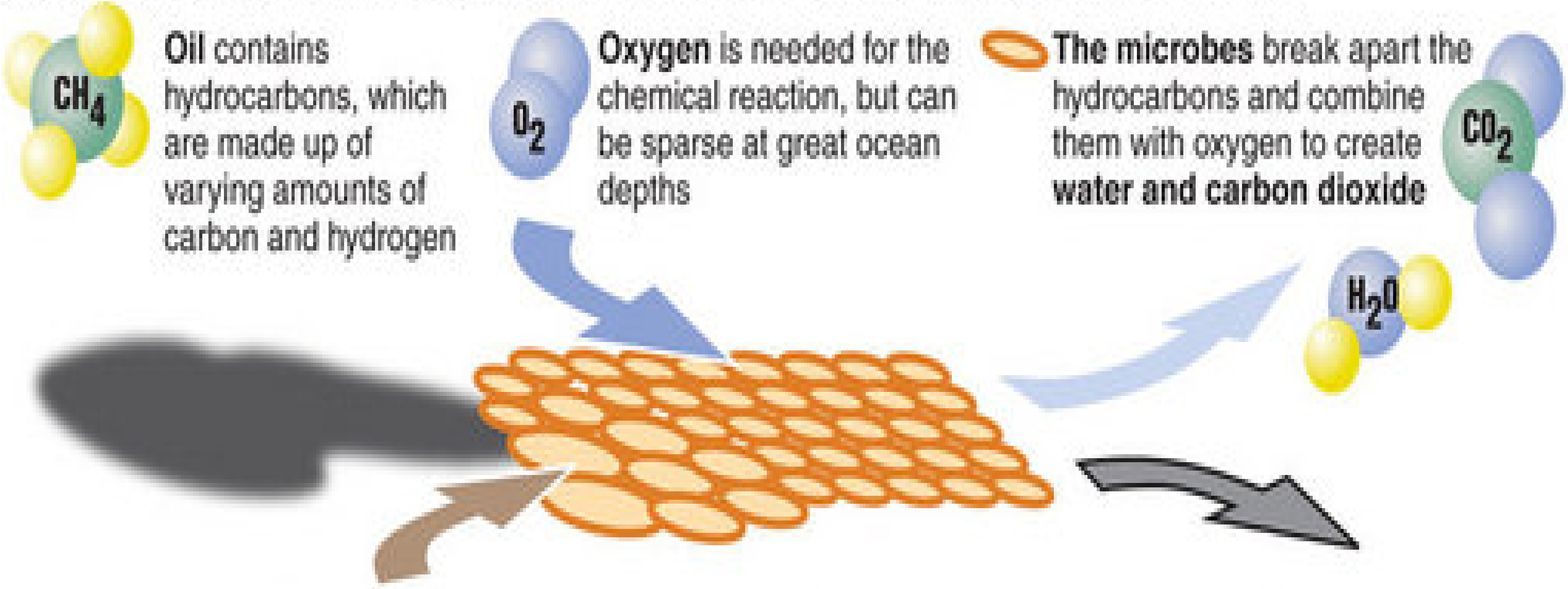




Petroleum-degrading microbes called *Oceanospirillales*

Oil-eating microbes

Naturally occurring microbes in the ocean feed on the hydrocarbons in oil. Scientists hope to speed up the process for the large spill in the Gulf of Mexico, where warm temperatures also aid the reaction.



Oil contains hydrocarbons, which are made up of varying amounts of carbon and hydrogen

Oxygen is needed for the chemical reaction, but can be sparse at great ocean depths

The microbes break apart the hydrocarbons and combine them with oxygen to create **water and carbon dioxide**

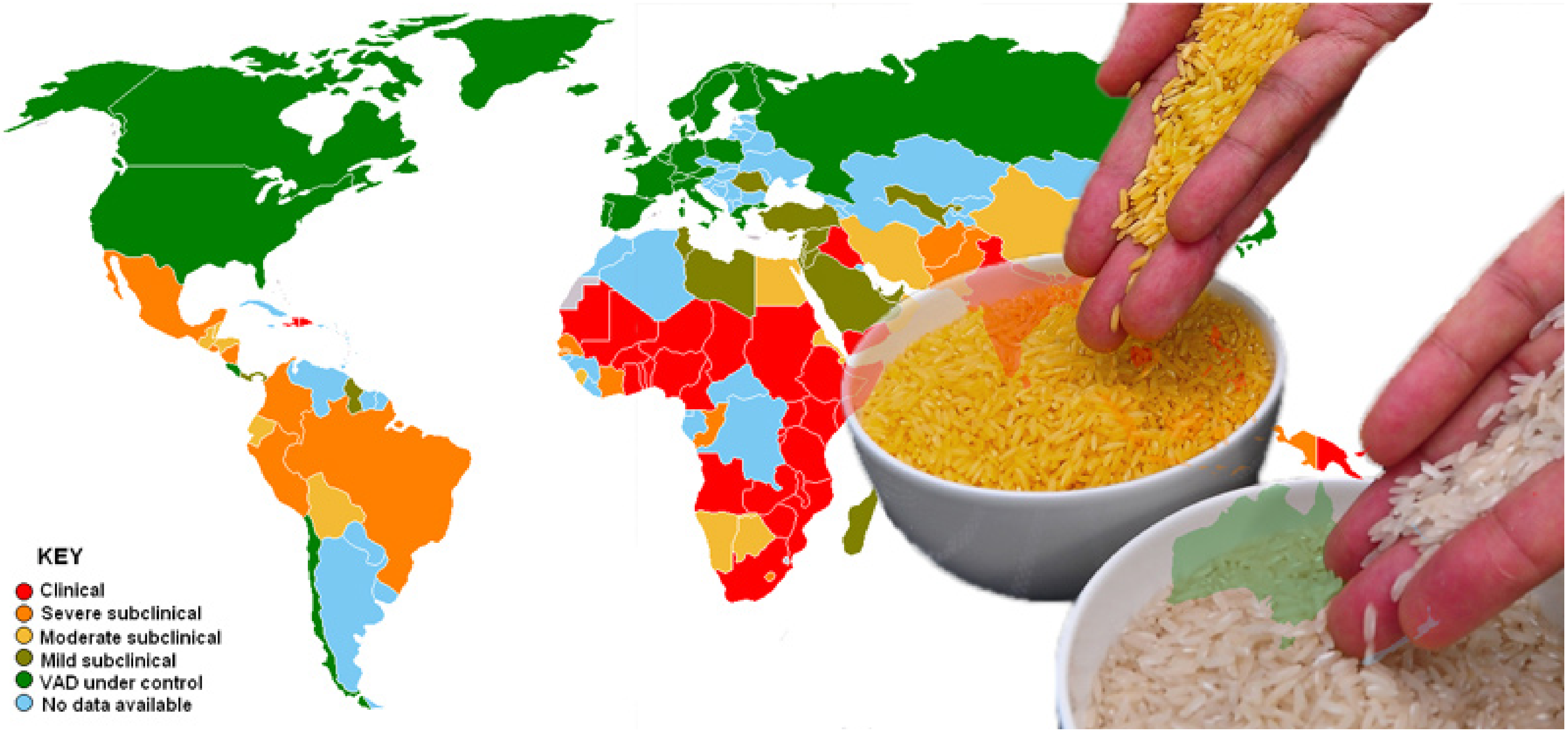
Adding fertilizer increases the size and number of the microbes so they can eat more oil; too much, however, can cause algae blooms, which starve the ecosystem of light and oxygen

Not all of the oil can be consumed, but what is left over is more easily dispersed by currents and wind

Source: Terry Hazen, Lawrence Berkeley National Lab
Graphic: Miami Herald

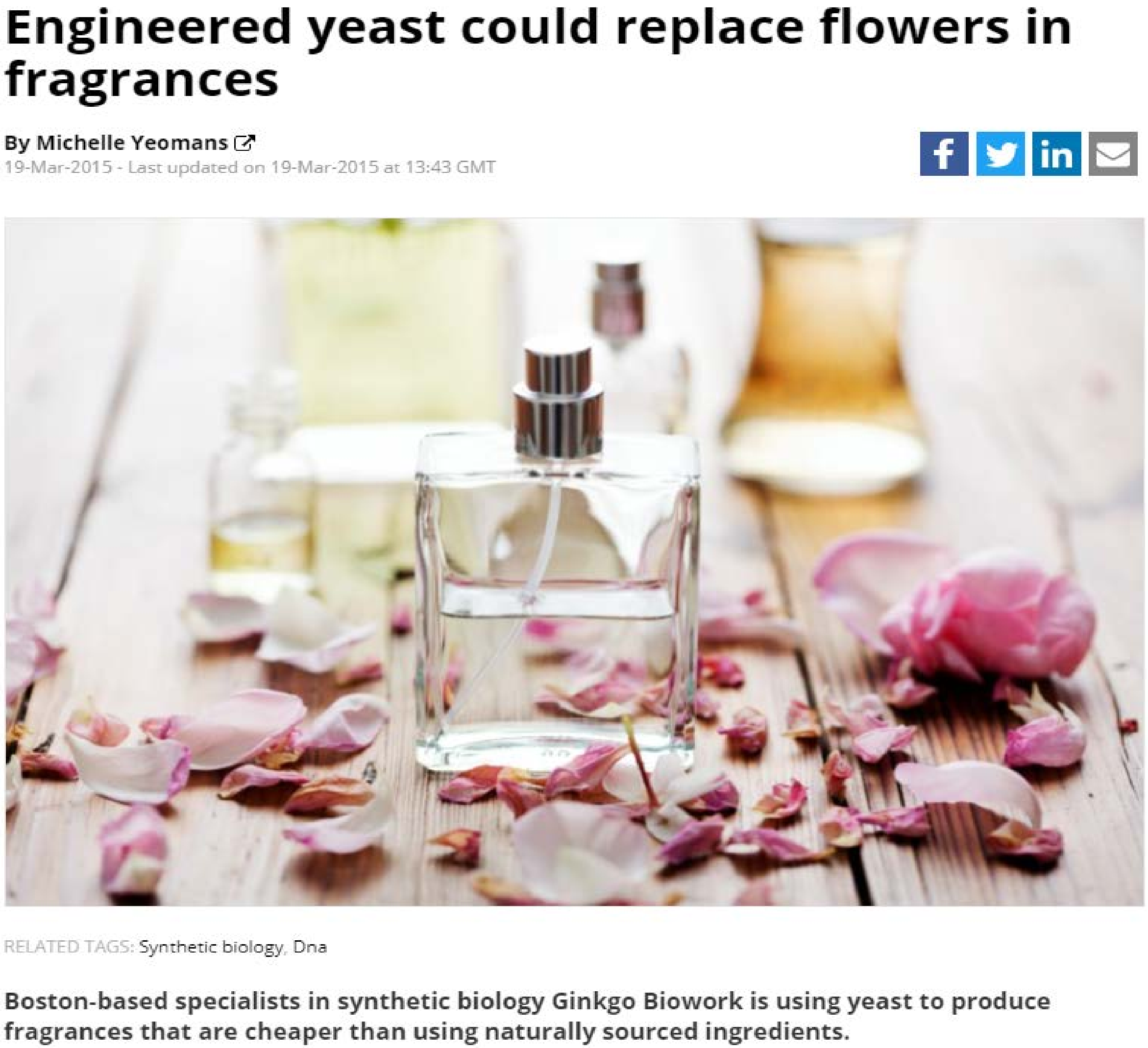
© 2010 MCT

Example 4: Rice modified to produce beta-carotene, a nutrient usually associated with carrots, that prevents vitamin A deficiency



Vitamin A deficiency causes blindness in 250,000 - 500,000 children every year and greatly increases a child's risk of death from infectious diseases.

Example 5: Yeast engineered to produce rose oil as an eco-friendly and sustainable substitute for real roses that perfumers use to make luxury scents.



Engineered yeast could replace flowers in fragrances

By Michelle Yeomans
19-Mar-2015 - Last updated on 19-Mar-2015 at 13:43 GMT

RELATED TAGS: Synthetic biology, Dna

Boston-based specialists in synthetic biology Ginkgo Biowork is using yeast to produce fragrances that are cheaper than using naturally sourced ingredients.

Genome Editing

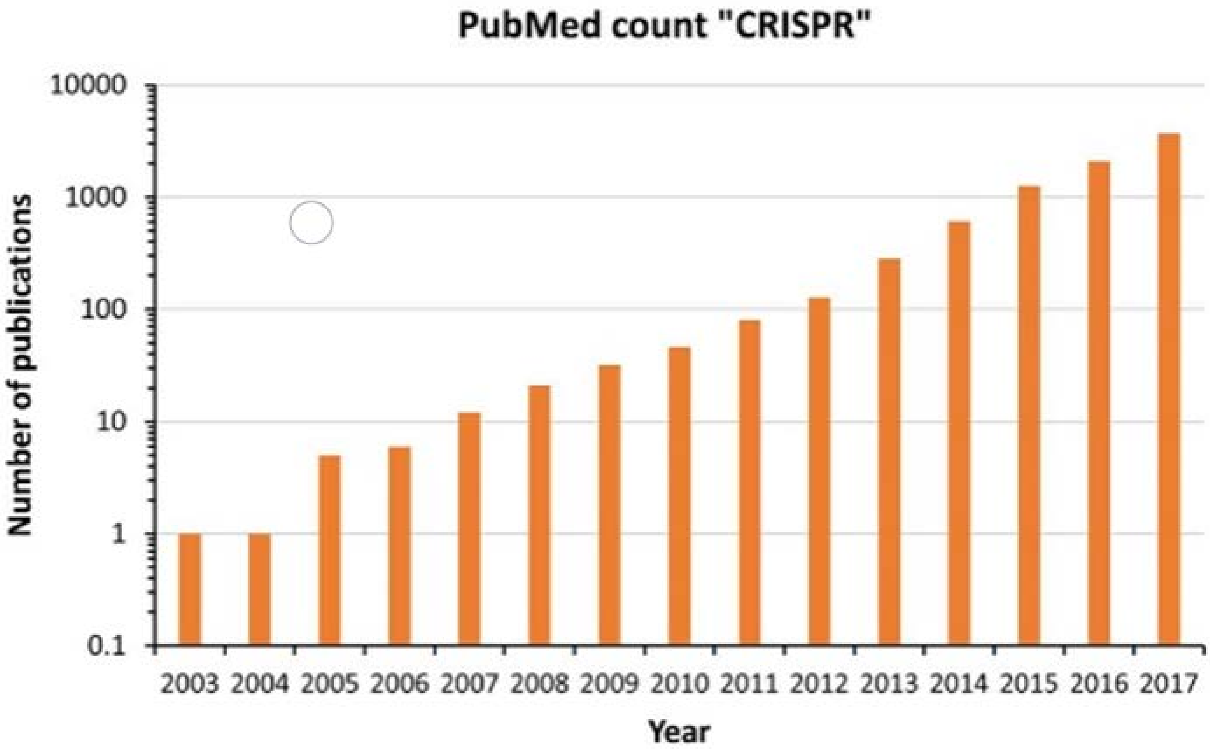
What is the difference between synthetic biology and genome editing?

- In some ways, synthetic biology is similar to another approach called "genome editing" because both involve changing an organism's genetic code; however, some people draw a distinction between these two approaches based on how that change is made.

- In synthetic biology, scientists typically stitch together long stretches of DNA and insert them into an organism's genome.
  - These synthesized pieces of DNA could be genes that are found in other organisms or they could be entirely novel.

- In genome editing, scientists typically use tools to make smaller changes to the organism's own DNA. Genome editing tools can also be used to delete or add small stretches of DNA in the genome.

CIRSPR/Cas9 Applications are Exploding and Revolutionize Molecular Biology



CRISPR-Cas Advanced Plant Breeding

Crop Insights by Jeffry Sander, Ph.D.[1] and Mark Jeschke, Ph.D.[2]

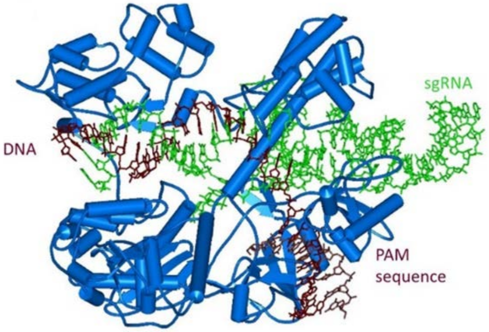
Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)

- A genome editing technique that
  - Targets a specific section of DNA
  - Make a precise cut/break at the target site

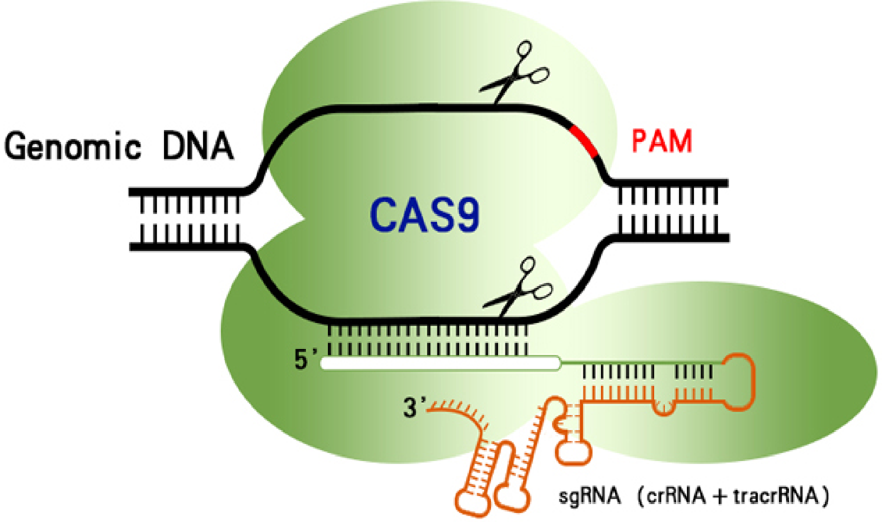
- Applications
  1. To make a gene nonfunctional (knockout)
  2. Replace on version of a gene with another
     - E.g., gene therapy
     - David Vetter was born without a functioning immune system and spent his life in a bubble that protected him from germs. He died at age 12 in 1984. Scientists are using gene therapy to treat the disorder so that children can live normally.

Adenosine Deaminase (ADA)

CRISPR/Cas9 Applications are Exploding and Revolutionize Molecular Biology



Structure of *staphylococcus aureus* Cas9 (blue) bound to single guide RNA (green) & targeted DNA (brown) (Nishimasu et al. 2015)



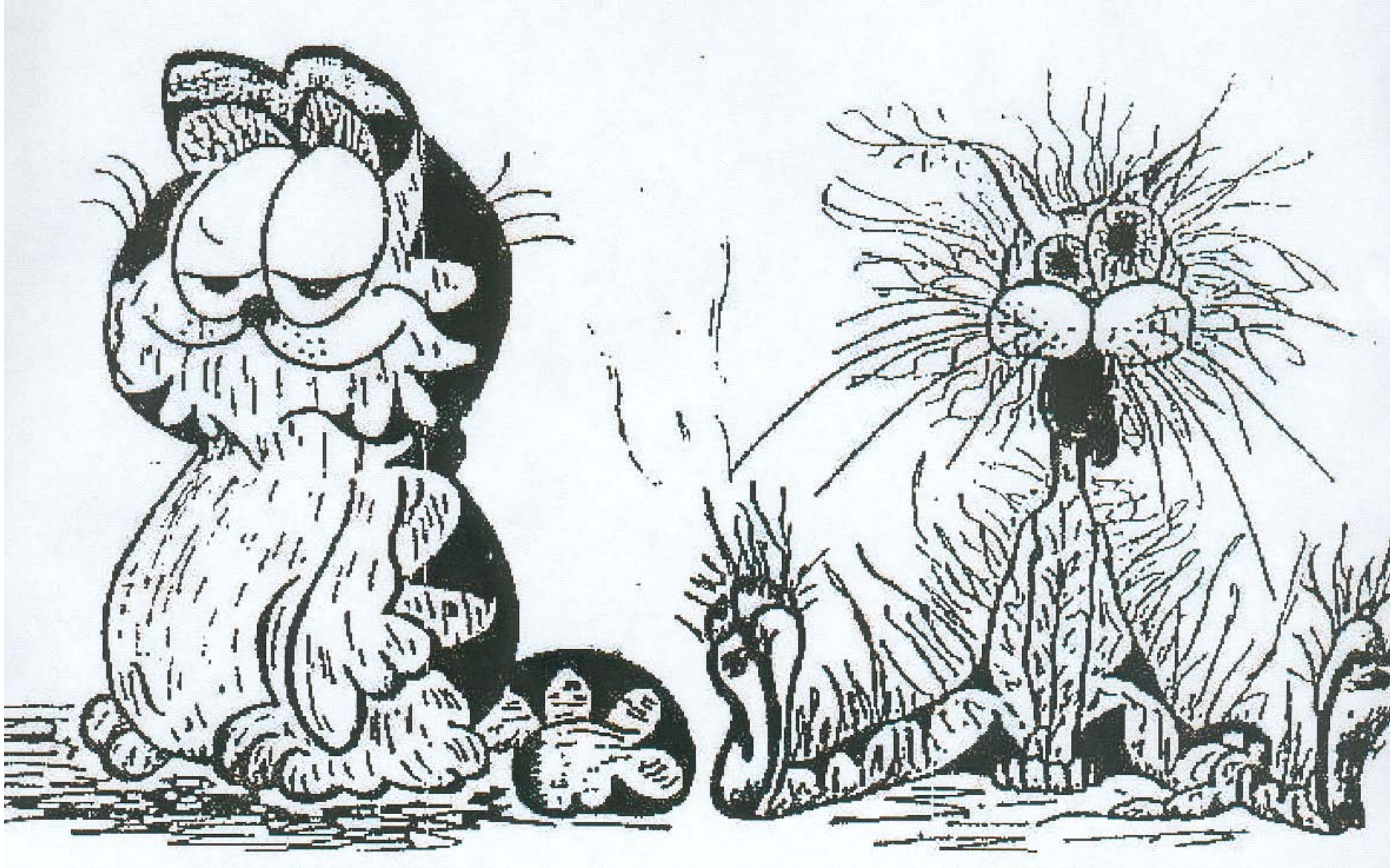
- Non-coding RNAs & Cas protein
- Protospacer adjacent motif (PAM) is a 2-6 base pair DNA sequence immediately following the DNA sequence targeted by the Cas9 nuclease in the CRISPR bacterial adaptive immune system
- sgRNA = single guide RNA = a targeting sequence (crRNA sequence) + (a Cas9 nuclease-recruiting sequence: tracrRNA)

CRISPR: Gene editing and beyond

s://www.youtube.com/watch?v=4YKFw2KZA5o

Genetic Engineering Will Change Everything Forever – CRISPR

Before…



After…