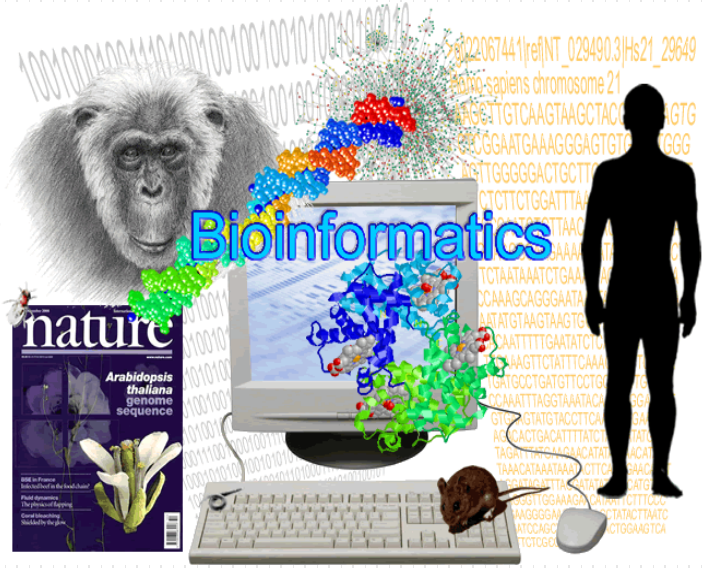
Bioinformatics, Syntenic Biology & Genome Editing



薛 佑 玲 PhD

Institute of Biomedical Sciences

National Sun Yat-sen University

ylshiue@mail.nsysu.edu.tw

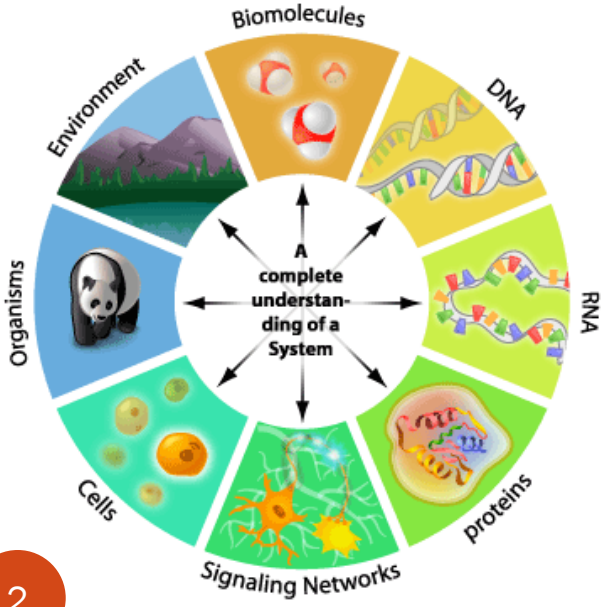
Outline

Introduction: a Short History About Bioinformatics

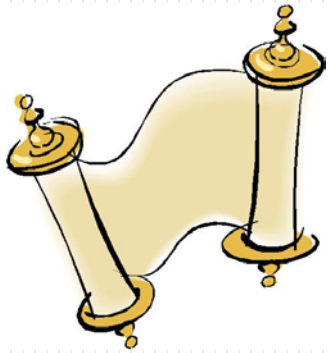
Bioinformatics Q & A

Synthetic biology

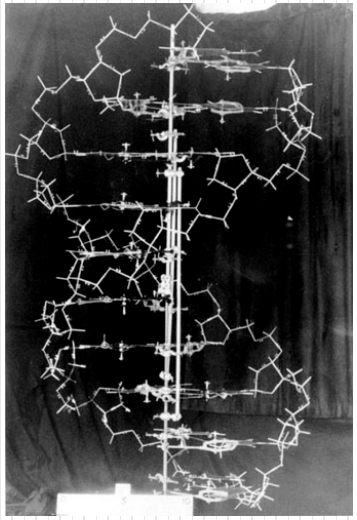
Genome Editing



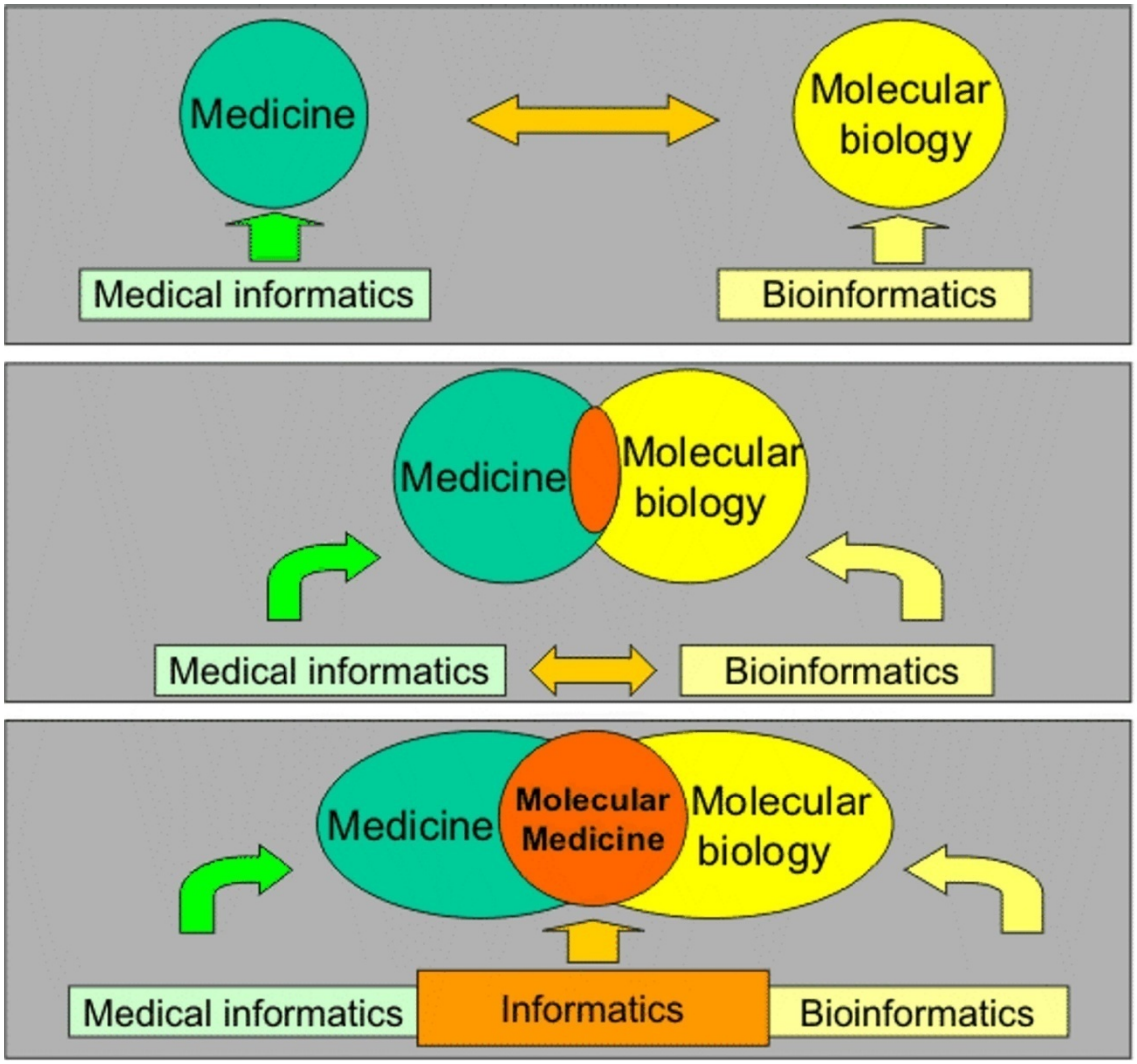
b101nf0rmat1cs



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**

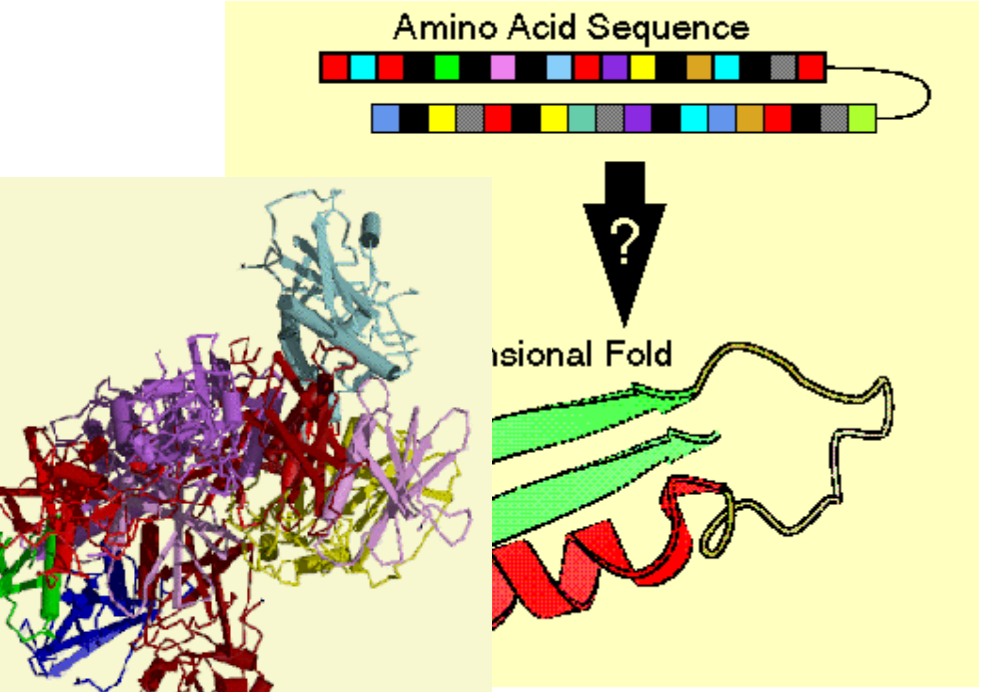
根據British Medical Journal 線上意見調查，
自1840年創刊以來，最重要的醫學里程碑

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)

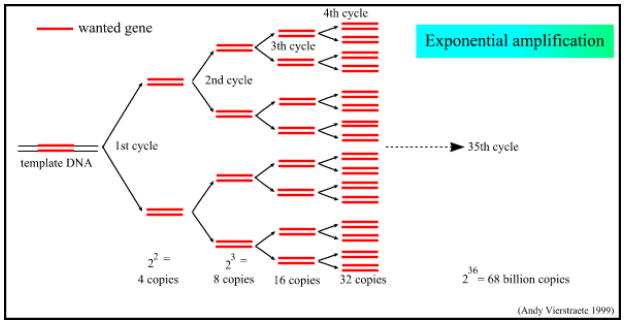
1995: Sequence of the genome of the bacteria: ***Haemophilus influenzae*** ($2*10^6$ bp)

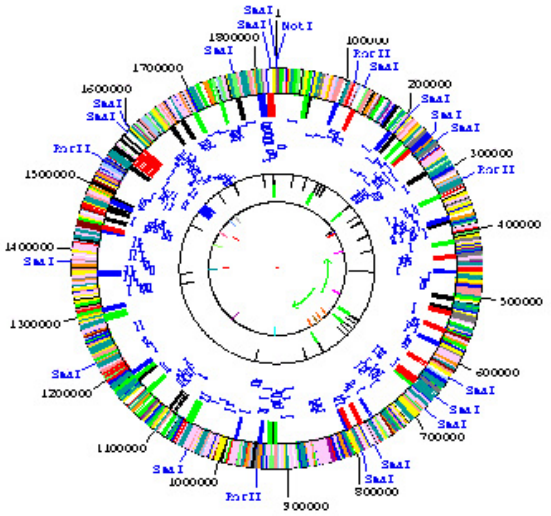
1999: Sequence of the genome of **a multi-cellular organism**: *Caenorhabditis elegans* (10^8 bp)

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

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

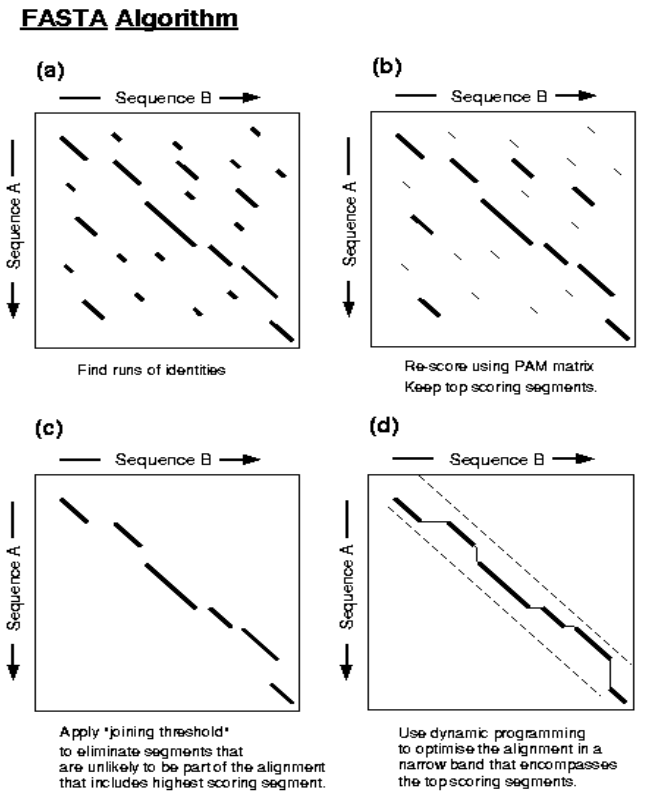
Recent: GOLD database

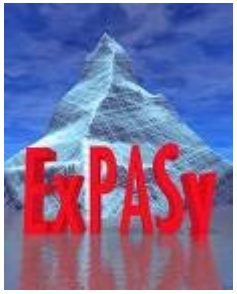




A Short History Overview (I) - Dry

- **1965**: «Atlas of protein sequence and structure» (**Dayhoff**)
- **1967**: Fitch WM (Phylogenetic trees)
- **1970**: **Needleman/Wunsch (1st similarity search algorithm)**
- **1971**: PDB (3D structure database)
- **1977**: **Staden (1st sequence analysis software suite)**
- **1980**: **EMBL Heidelberg**
- **1980: Smith/Waterman algorithm**
- **1982**: EMBL Nucleotide Sequence Database and GenBank
- **1985: CABIOS (1st scientific journal for bioinformatics)**
- **1985**: FASTP (ancestor of **FASTA**, Blast, etc.)
- **1986**: Swiss-Prot (Protein Sequence Database)
- **1988: Creation of the NCBI in the USA**
- **1992**: EBI founded as EMBL outstation in **Hinxton** (Wellcome Trust Campus)
- **1993**: **ExPASy** (1st WWW server for the life sciences)…





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

Table 2. Twenty Publications that influenced our view of bioinformatics

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

Bioinformatics: A Snapshot 10 Years Ago

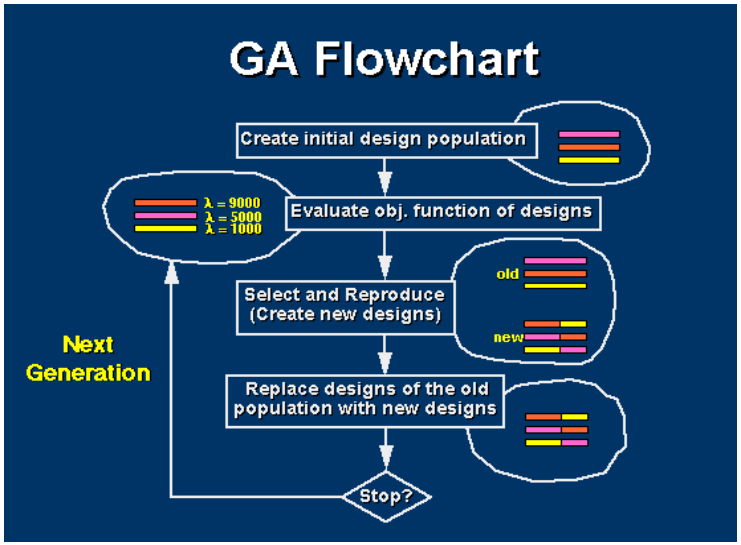
Pharmaceutical companies were **not interested**

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

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

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

DNA computers…



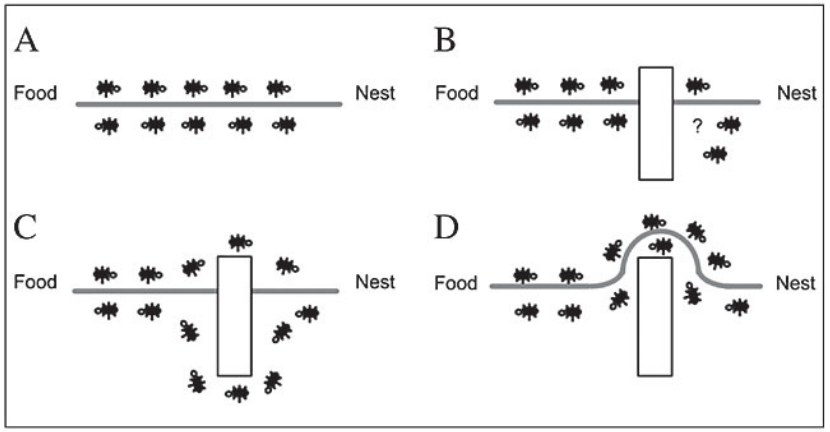


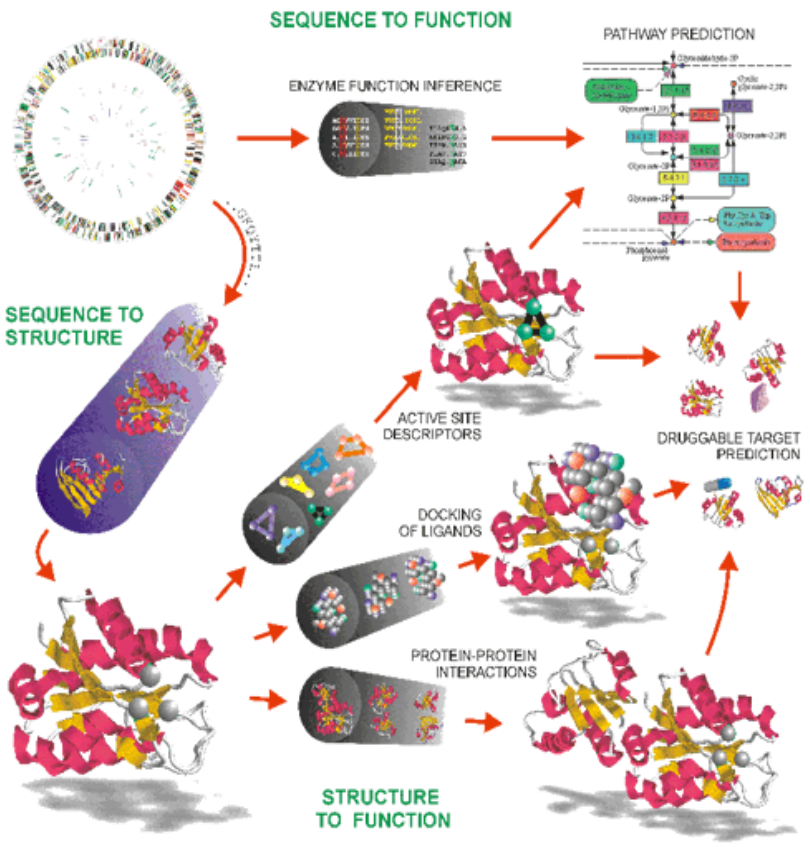
Figure 2. A. Ants in a pheromone trail between nest and food; **B.** an obstacle interrupts the trail; **C.** ants find two paths to go around the obstacle; **D.** a new pheromone trail is formed along the shorter path.

Bioinformatics in 2003

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

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

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

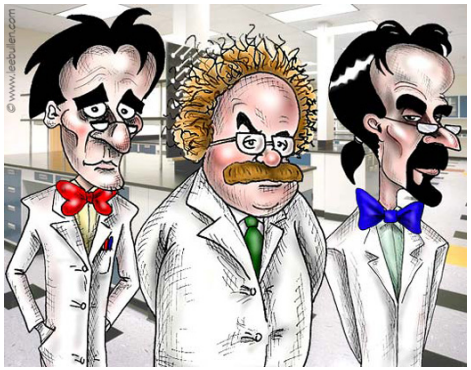


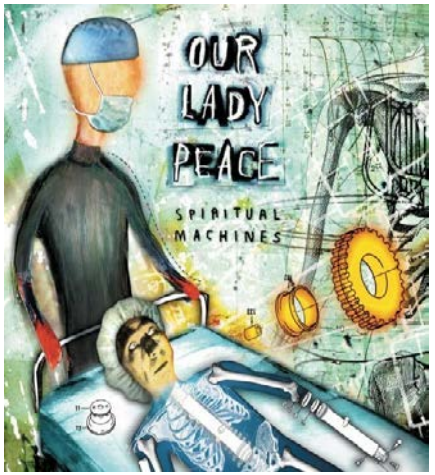
Bioinformatics in 2010

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

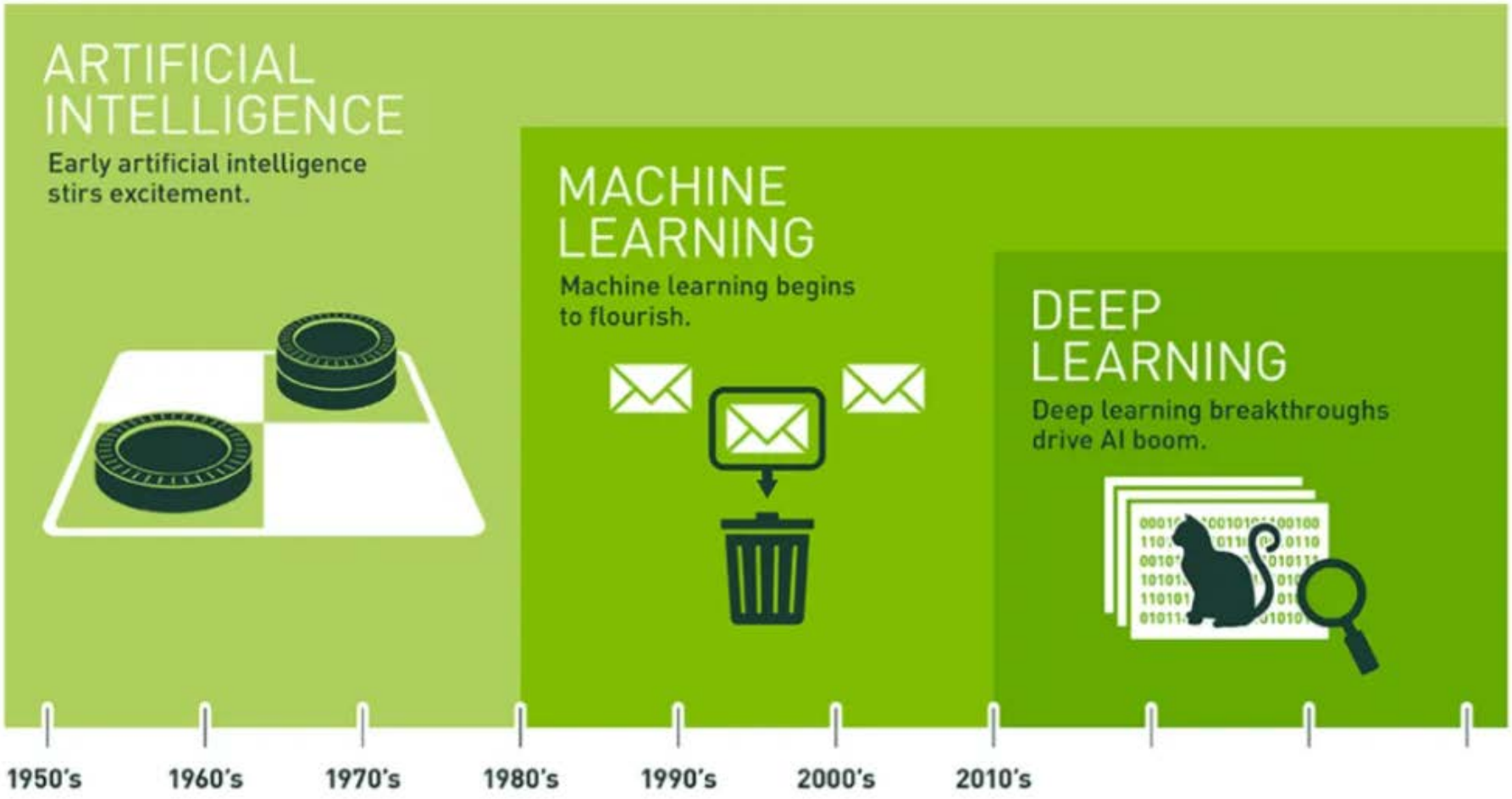
TOPICS OF INTEREST

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

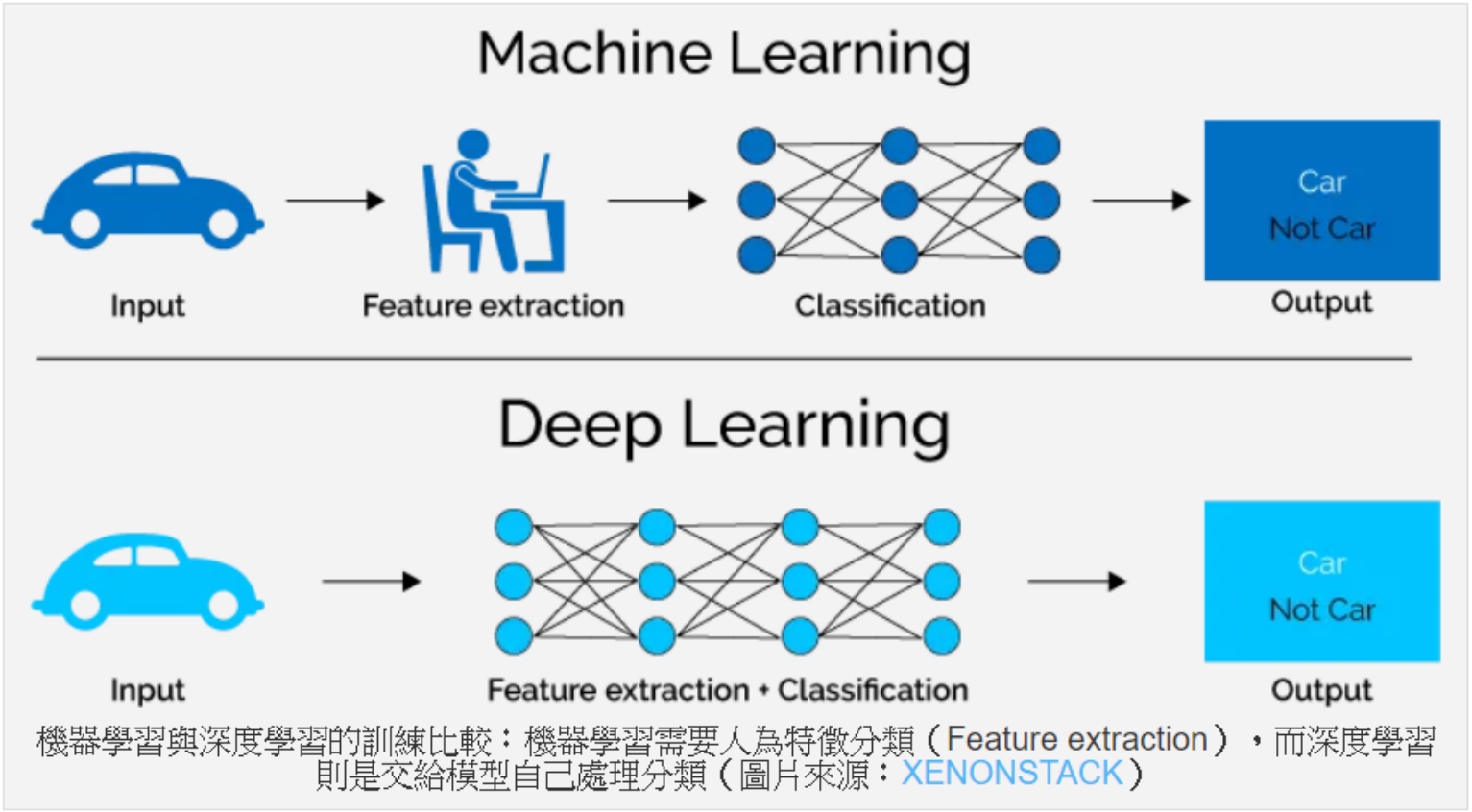
Artificial Intelligence

- 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。

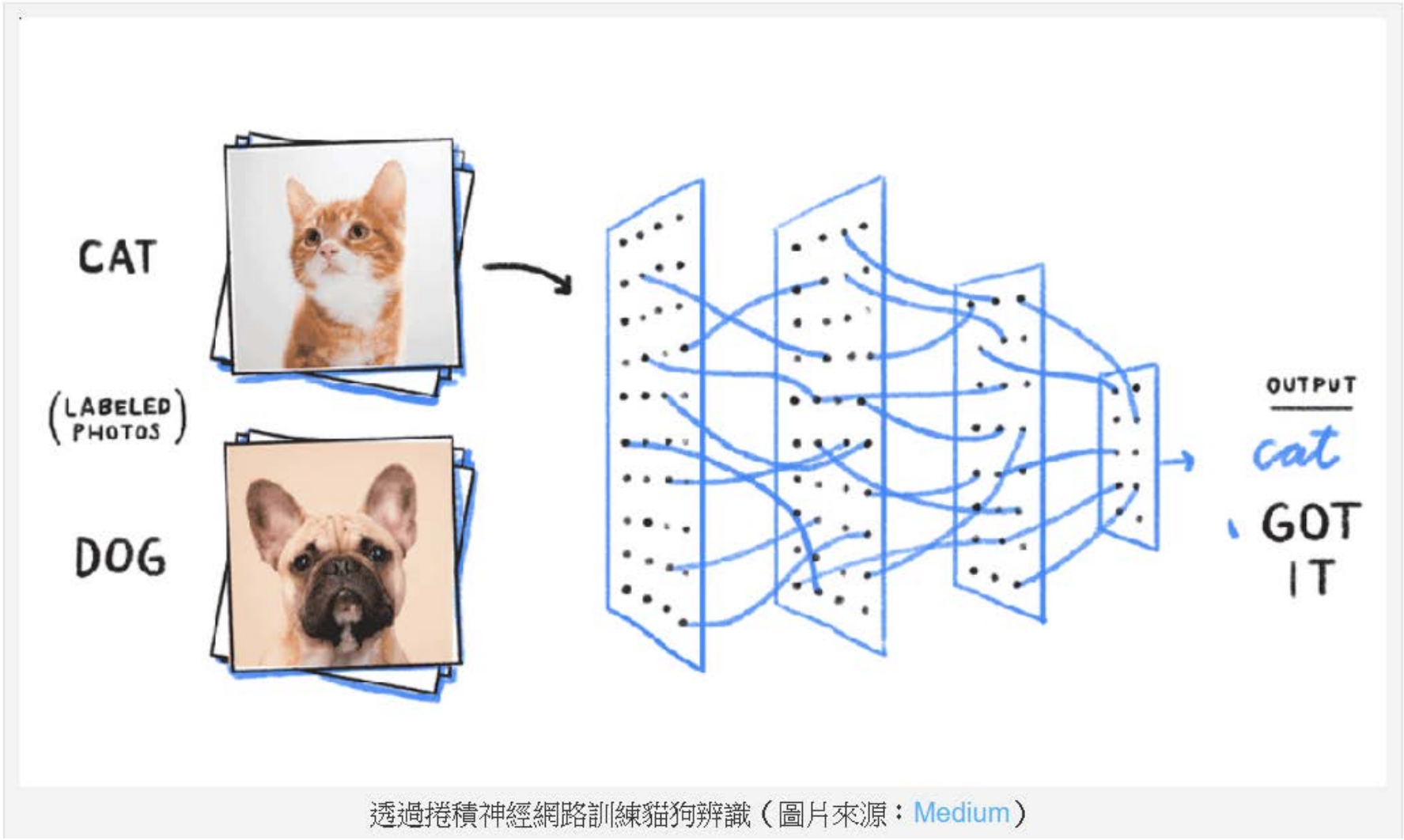


AI 演進（圖片來源：NVIDIA）

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。



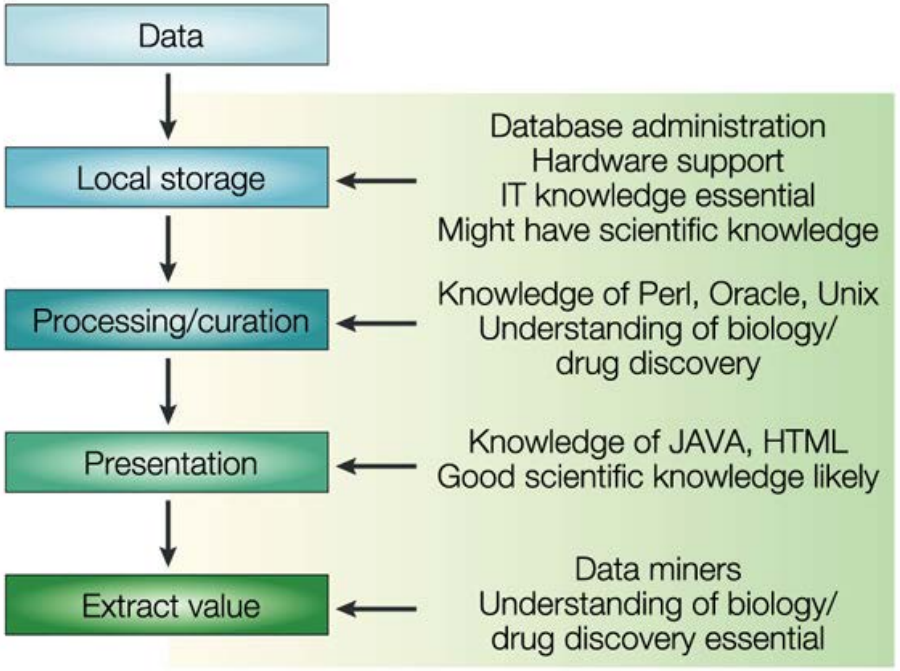
機器學習與深度學習的訓練比較：機器學習需要人為特徵分類（Feature extraction），而深度學習則是交給模型自己處理分類（圖片來源：XENONSTACK）

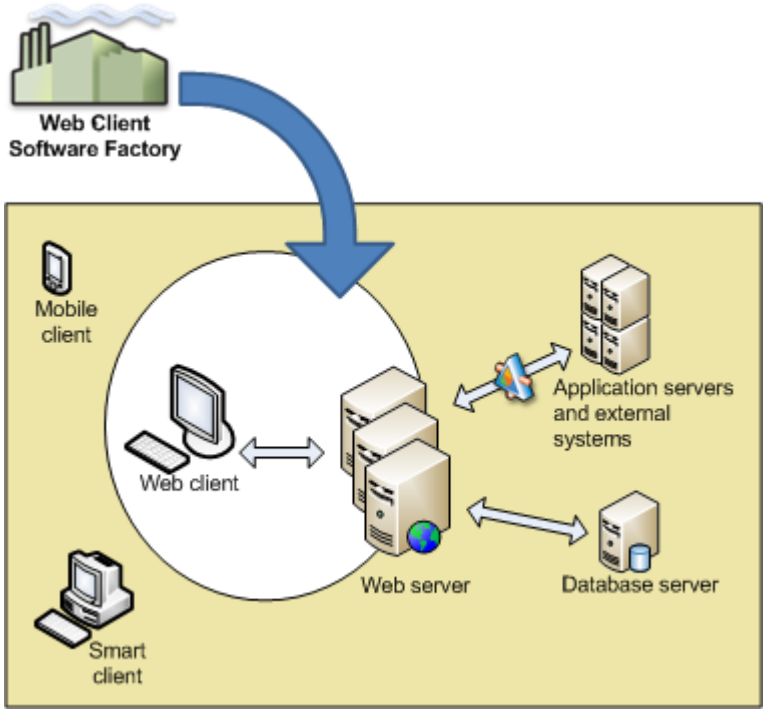


透過捲積神經網路訓練貓狗辨識（圖片來源：Medium）

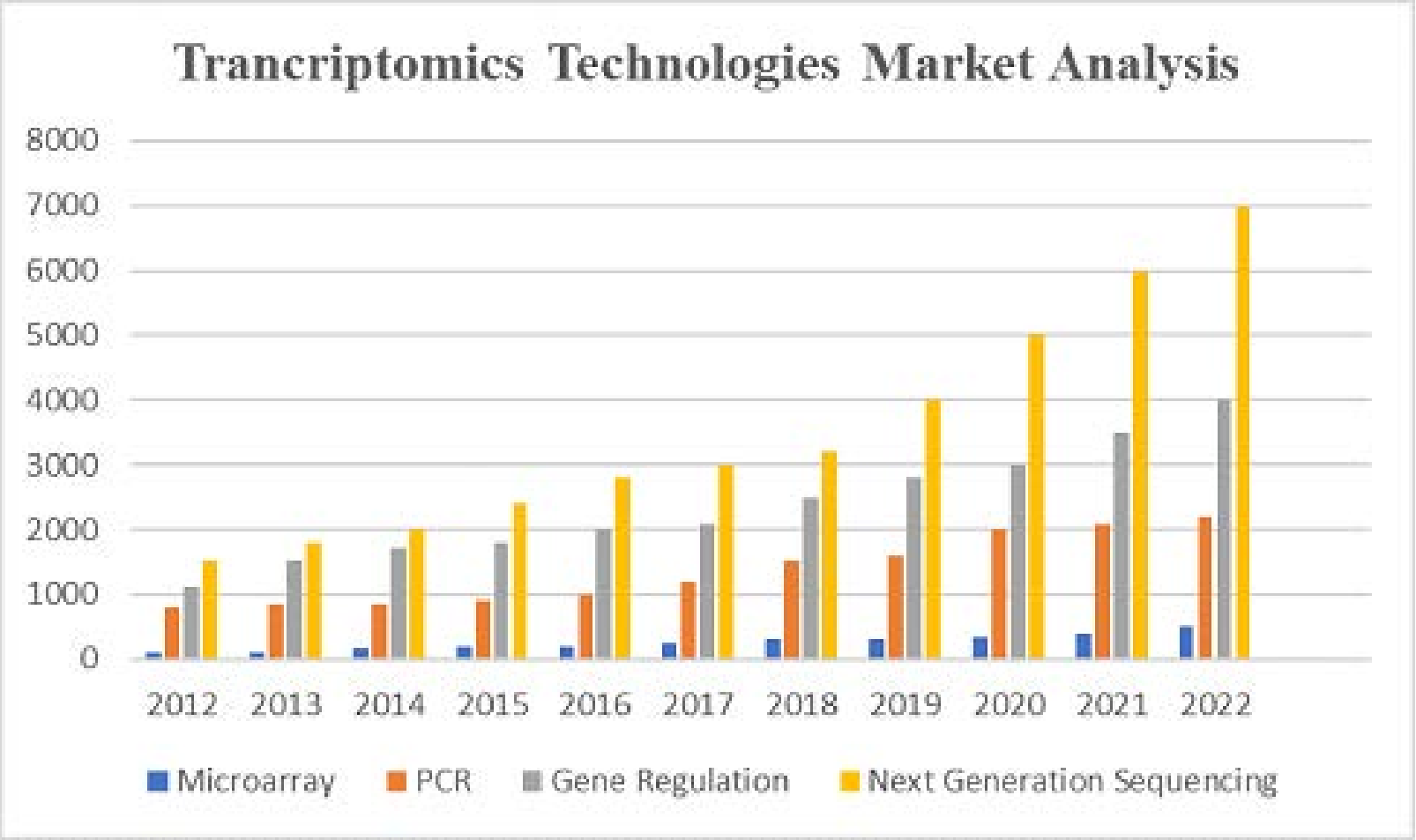
Convolutional Neural Network, CNN

Resources: databases & software



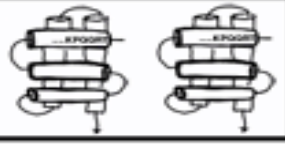
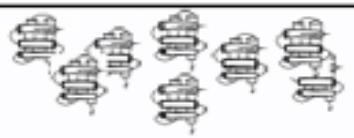


Nature Reviews Drug Discovery 3, 281 (2004)


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

Depth: Rational Drug Design (physi

Breadth: Homologs, Large-scale Surveys, Informatics–

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

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

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

Case Study

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

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]

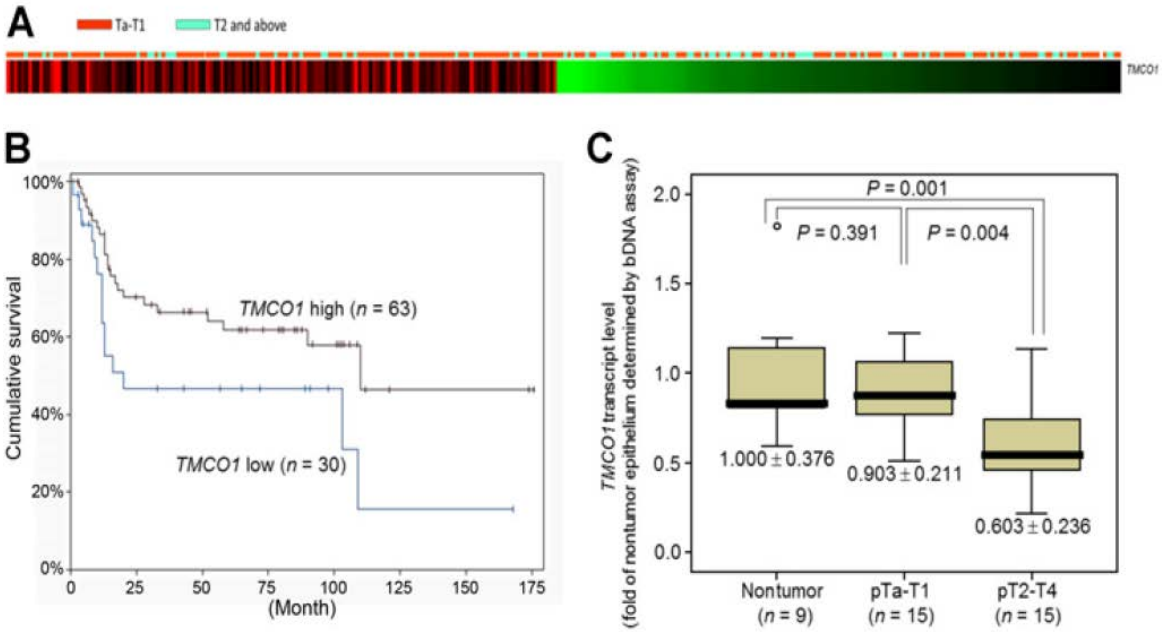


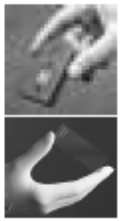
Figure 1.

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

Gene Expression Omnibus (GEO)

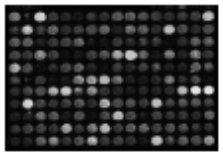
*Submitted by Manufacturer**

GPL
Platform descriptions



Submitted by Experimentalists

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



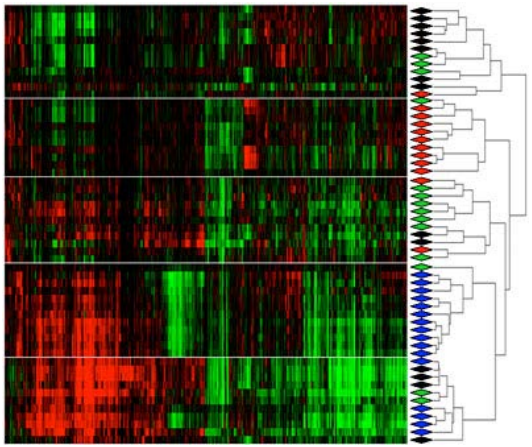
GSE
Grouping of slide/chip data "a single experiment"

Entrez GEO

Curated by NCBI

GDS
Grouping of experiments

Entrez GEO Datasets

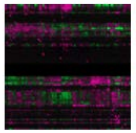


NCBI

DataSet CURATED BROWSER

Search for GDS4456[ACCN] Search Clear Show All Advanced Search

Cluster Analysis



Download

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

DataSet Record GDS4456: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	High-risk bladder cancer		
Summary:	Analysis of bladder cancer specimens from a high-risk population of patients who underwent radical cystectomy (Memorial Sloan-Kettering Cancer Center cohort). Results provide insight into the prediction of survival in high-risk urothelial carcinoma of the urinary bladder.		
Organism:	*Homo sapiens*		
Platform:	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array		
Citations:	Riester M, Taylor JM, Feifer A, Koppie T et al. Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin Cancer Res* 2012 Mar 1;18(5):1323-33. PMID: 22228636 Riester M, Werner L, Bellmunt J, Selvarajah S et al. Integrative analysis of 1q23.3 copy-number gain in metastatic urothelial carcinoma. *Clin Cancer Res* 2014 Apr 1;20(7):1873-83. PMID: 24486590		
Reference Series:	GSE31684	**Sample count:**	93
Value type:	transformed count	**Series published:**	2012/01/23

Data Analysis Tools

Find genes

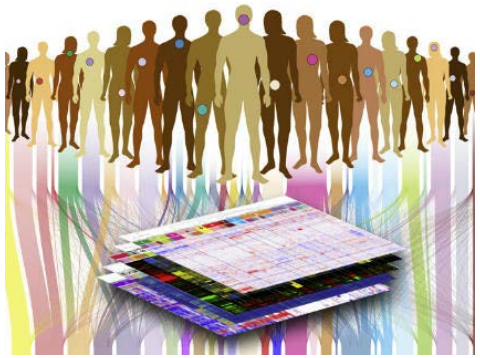
Compare 2 sets of samples

Cluster heatmaps

Find gene name or symbol: Go

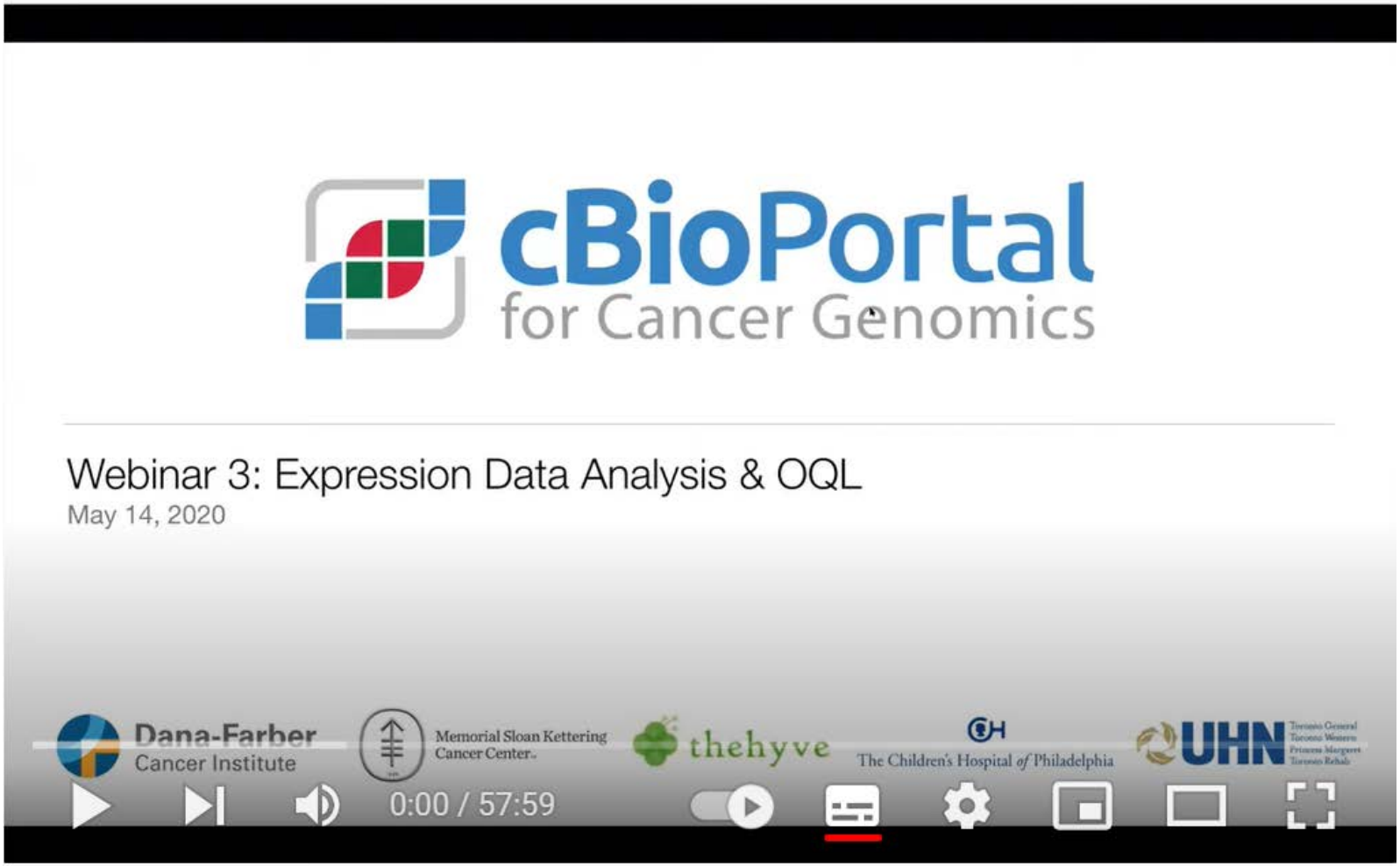
Find genes that are up/down for this condition(s): ☑ specimen ☑ disease state Go

TCGA: The Cancer Genome Atlas Program



- The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over **20,000 primary cancer** and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions

- Over the next dozen years, TCGA generated over 2.5 petabytes (2^{50}) of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.



cBioPortal Webinar 3: Expression Data Analysis

Cell

Volume 173, Issue 2, 5 April 2018, Page 530

SnapShot

SnapShot: TCGA-Analyzed Tumors

Amy Blum, Peggy Wang, Jean C. Zenklusen

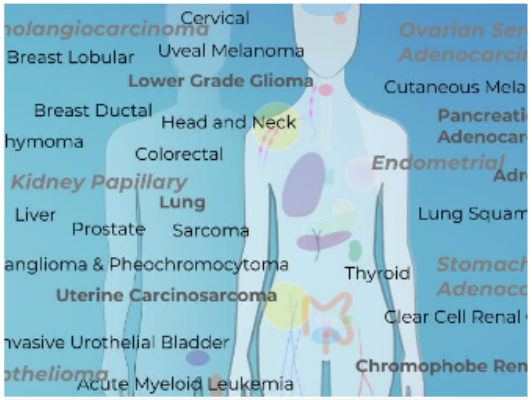
Show more

Cancer type	Prevalence	TCGA cases assessed	Key findings
Breast lobular carcinoma	3,327,552	203	FOXA1 elevated in lobular carcinoma, GATA3 in ductal carcinoma; lobular enriched for PTEN loss and Akt activation
Breast ductal carcinoma		784	Four distinct subtypes: basal, Her2, luminal A, and luminal B; most common driver mutations: TP53, PIK3CA, GATA3; basal subtype similar to serous ovarian cancer
Prostate cancer	3,085,209	333	Highly heterogeneous with 26% driven by unknown alterations; ETS gene fusions or mutations in SPOP, FOXA1, or IDH1 define seven subtypes; actionable lesions in PI3K, MAPK, and DNA repair pathways
Colorectal adenocarcinoma	1,317,247	276	Colon and rectal cancers have similar genomic profiles; hypermutated subtype associated with favorable prognosis; new potential drivers: ARID1A, SOX9, FAM123B/WTX
Cutaneous melanoma	1,169,351	331	Established four subtypes: BRAF mutant, RAS mutant, NF1 mutant, and triple wild-type based on driver mutations; higher levels of immune lymphocyte infiltration correlated with better survival
Thyroid carcinoma	726,646	496	Majority driven by RAS or BRAFV600E mutations
Endometrial carcinoma	710,228	373	Classified endometrial cancers into four categories: POLE ultramutated, MSI (microsatellite instability) hypermutated, copy-number low, copy-number high
Uterine carcinosarcoma		57	Strong and varied degree of epithelial-mesenchymal transition; TP53 mutations in 91% of samples; PI3K alterations in half

TCGA Cancers Selected for Study

The Cancer Genome Atlas (TCGA) selected the following cancers for study based on specific criteria that include:

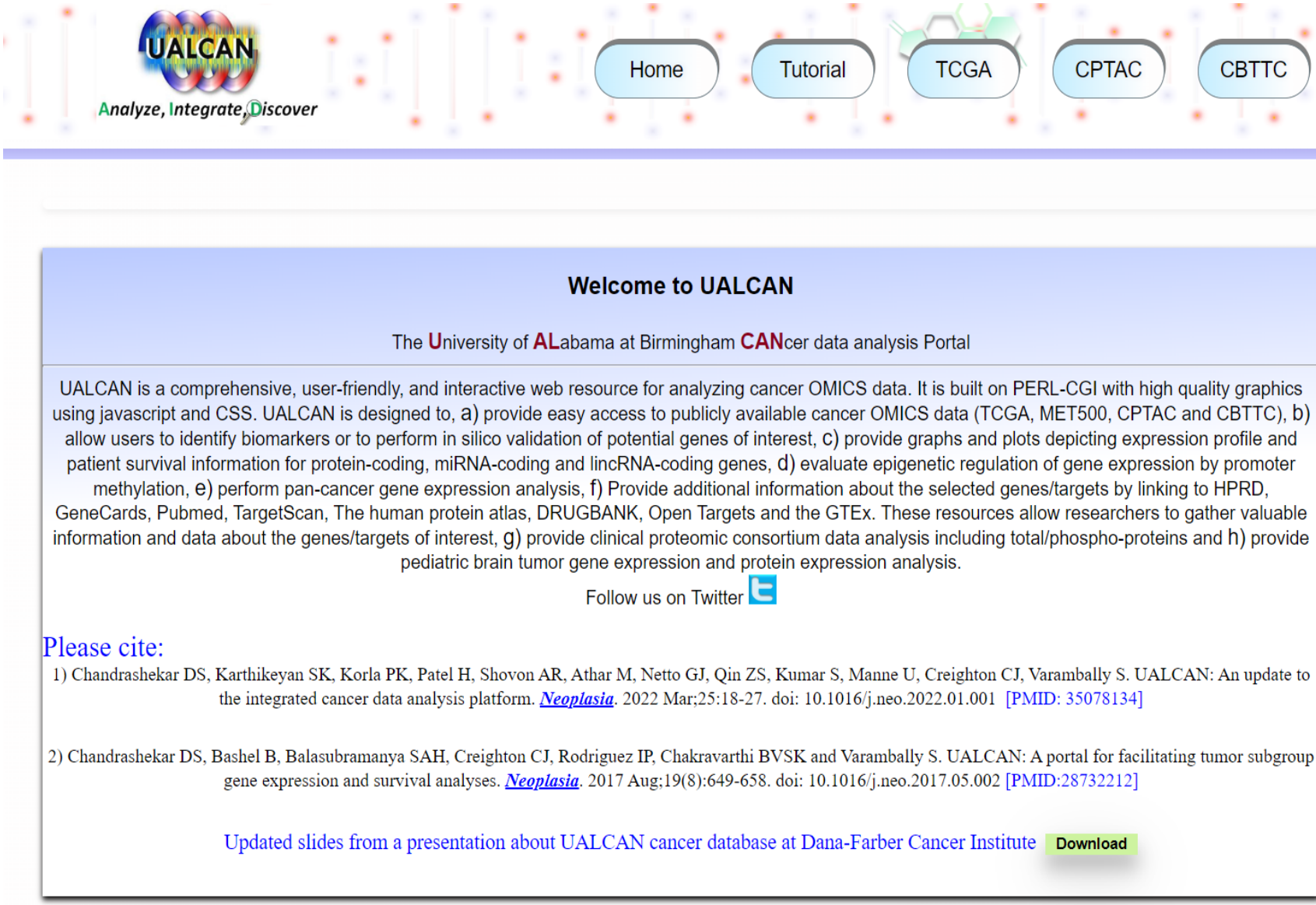
- Poor prognosis
- Overall public health impact
- Availability of samples meeting standards for patient consent
- Availability of samples meeting standards for quality and quantity that include:
  - Primary, untreated tumor with a source of matched normal tissue or blood sample
  - Frozen, sufficiently sized, resection samples
  - Samples composed of at least 80% tumor nuclei (threshold later lowered to 60% with improved sequencing technology and computational methods)



TCGA cancers selected for study
Credit: National Cancer Institute

TCGA's Cancers Selected for Study

Cancer Type Studied	# Cases Characterized (# Cases in Marker Paper)	Publication
Acute Myeloid Leukemia	200 (200)	NEJM 2013
Adrenocortical Carcinoma	92 (91)	Cancer Cell 2016
Bladder Urothelial Carcinoma	412 (131)	Nature 2014, Cell 2017
Breast Ductal Carcinoma	778 (430)	Nature 2012
Breast Lobular Carcinoma	201 (127)	Cell 2015
Cervical Carcinoma	307 (228)	Nature 2017
Cholangiocarcinoma	51 (38)	Cell Reports 2017
Colorectal Adenocarcinoma	633 (276)	Nature 2012
Esophageal Carcinoma	185 (164)	Nature 2017
Gastric Adenocarcinoma	443 (295)	Nature 2014
Glioblastoma Multiforme	617 (206)	Nature 2008, Cell 2013
Head and Neck Squamous Cell Carcinoma	528 (279)	Nature 2015


UALCAN
Analyze, Integrate, Discover
Home
Tutorial
TCGA
CPTAC
CBTTC
Welcome to UALCAN
The University of ALabama at Birmingham CANcer data analysis Portal
UALCAN is a comprehensive, user-friendly, and interactive web resource for analyzing cancer OMICS data. It is built on PERL-CGI with high quality graphics using javascript and CSS. UALCAN is designed to, a) provide easy access to publicly available cancer OMICS data (TCGA, MET500, CPTAC and CBTTC), b) allow users to identify biomarkers or to perform in silico validation of potential genes of interest, c) provide graphs and plots depicting expression profile and patient survival information for protein-coding, miRNA-coding and lincRNA-coding genes, d) evaluate epigenetic regulation of gene expression by promoter methylation, e) perform pan-cancer gene expression analysis, f) Provide additional information about the selected genes/targets by linking to HPRD, GeneCards, Pubmed, TargetScan, The human protein atlas, DRUGBANK, Open Targets and the GTEx. These resources allow researchers to gather valuable information and data about the genes/targets of interest, g) provide clinical proteomic consortium data analysis including total/phospho-proteins and h) provide pediatric brain tumor gene expression and protein expression analysis.
Follow us on Twitter
Please cite:
1) Chandrashekar DS, Karthikeyan SK, Korla PK, Patel H, Shovon AR, Athar M, Netto GJ, Qin ZS, Kumar S, Manne U, Creighton CJ, Varambally S. UALCAN: An update to the integrated cancer data analysis platform. Neoplasia. 2022 Mar;25:18-27. doi: 10.1016/j.neo.2022.01.001 [PMID: 35078134]
2) Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Rodriguez IP, Chakravarthi BVSK and Varambally S. UALCAN: A portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017 Aug;19(8):649-658. doi: 10.1016/j.neo.2017.05.002 [PMID:28732212]
Updated slides from a presentation about UALCAN cancer database at Dana-Farber Cancer Institute
Download

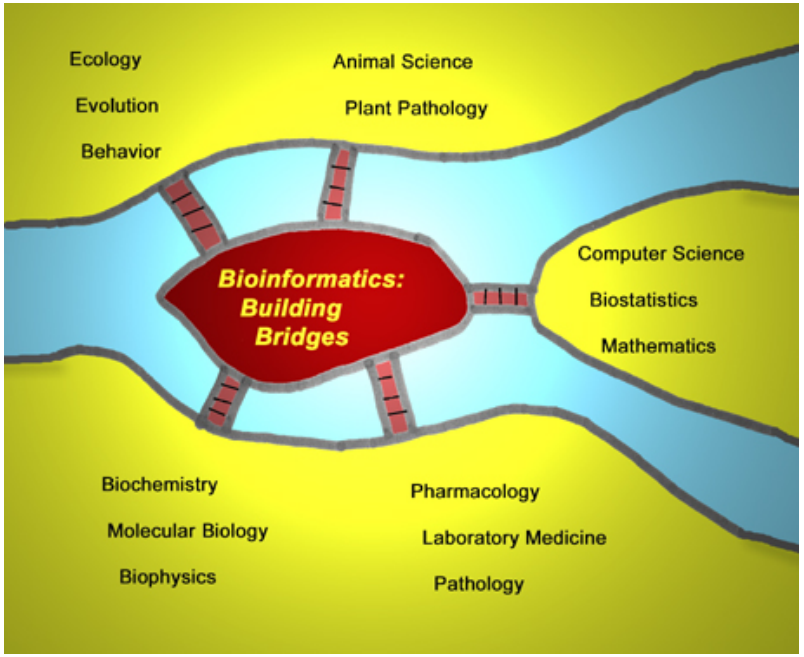
NEOPLASIA
www.neoplasia.com
Volume 19 Number 8 August 2017 pp. 649–658 649

UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses[1]

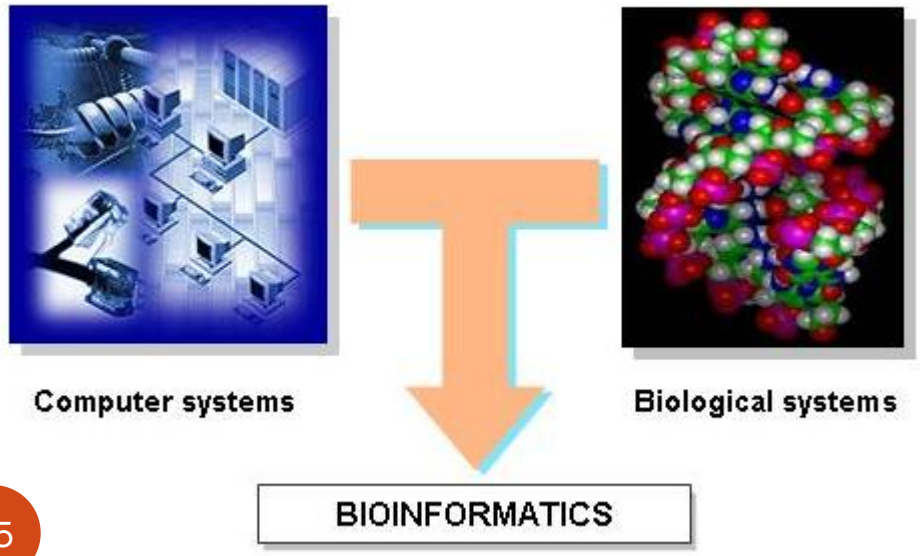
Darshan S. Chandrashekar[*,†], Bhuwan Bashel[*], Sai Akshaya Hodigere Balasubramanya[*,†], Chad J. Creighton[‡], Israel Ponce-Rodriguez[*], Balabhadrapatruni V.S.K. Chakravarthi[*,†] and Sooryanarayana Varambally[*,†]

[*]Molecular and Cellular Pathology, Department of Pathology, University of Alabama at Birmingham; [†]Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35233, USA; [‡]Department of Medicine, Dan L. Duncan Comprehensive Cancer Center, and Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA

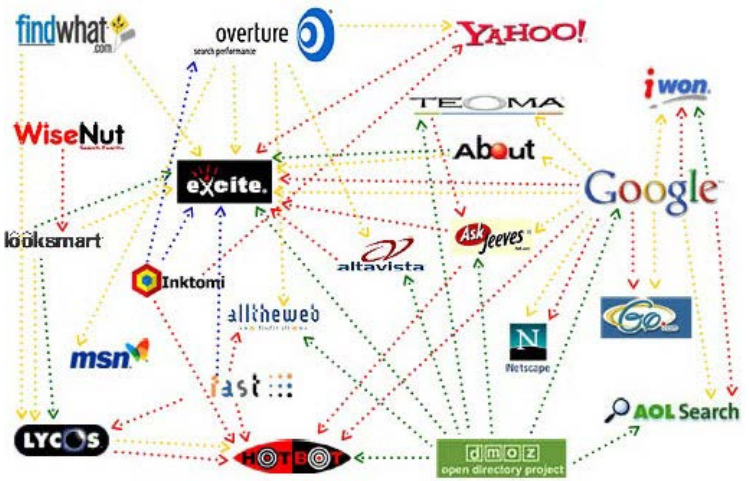
•Genomics data from The Cancer Genome Atlas (TCGA) project has led to the comprehensive molecular characterization of multiple cancer types. The large sample numbers in TCGA offer an excellent opportunity to address questions associated with tumor heterogeneity. Exploration of the data by cancer researchers and clinicians is imperative to unearth novel therapeutic/diagnostic biomarkers. Various computational tools have been developed to aid researchers in carrying out specific TCGA data analyses; however there is need for resources to facilitate the study of gene expression variations and survival associations across tumors. Here, we report UALCAN, an easy to use, interactive web-portal to perform to in-depth analyses of TCGA gene expression data. UALCAN uses TCGA level 3 RNA-seq and clinical data from 31 cancer types. The portal's user-friendly features allow to perform: 1) analyze relative expression of a query gene(s) across tumor and normal samples, as well as in various tumor sub-groups based on individual cancer stages, tumor grade, race, body weight or other clinicopathologic features, 2) estimate the effect of gene expression level and clinicopathologic features on patient survival; and 3) identify the top over- and under-expressed (up and down-regulated) genes in individual cancer types. This resource serves as a platform for in silico validation of target genes and for identifying tumor sub-group specific candidate biomarkers. Thus, UALCAN web-portal could be extremely helpful in accelerating cancer research. UALCAN is publicly available at http://ualcan.path.uab.edu.

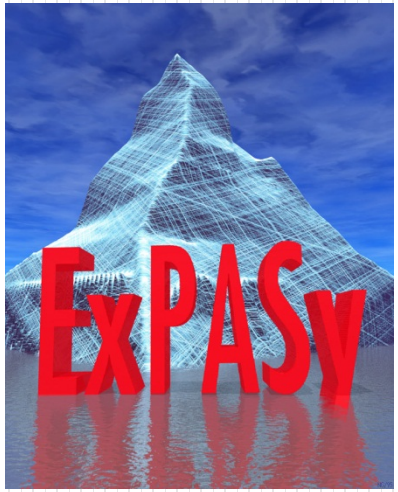

Ecology
Evolution
Behavior
Animal Science
Plant Pathology
Bioinformatics: Building Bridges
Computer Science
Biostatistics
Mathematics
Biochemistry
Molecular Biology
Biophysics
Pharmacology
Laboratory Medicine
Pathology

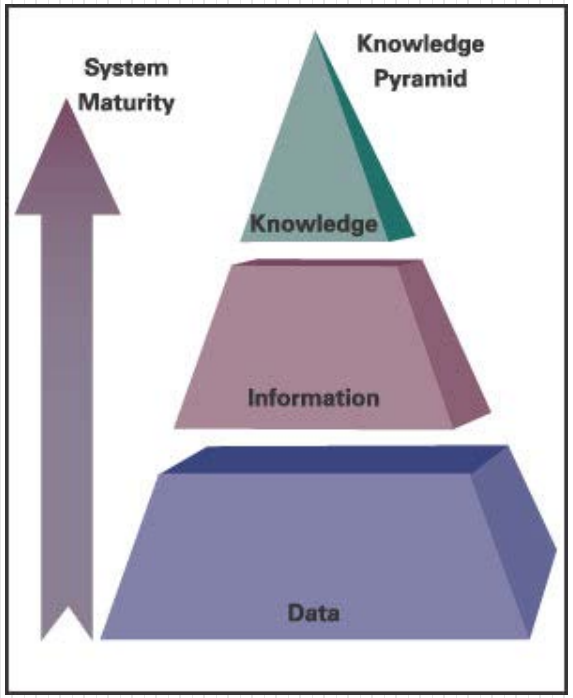
Q & A


Computer systems
Biological systems
BIOINFORMATICS

Q: How to Find the Right Stuffs?





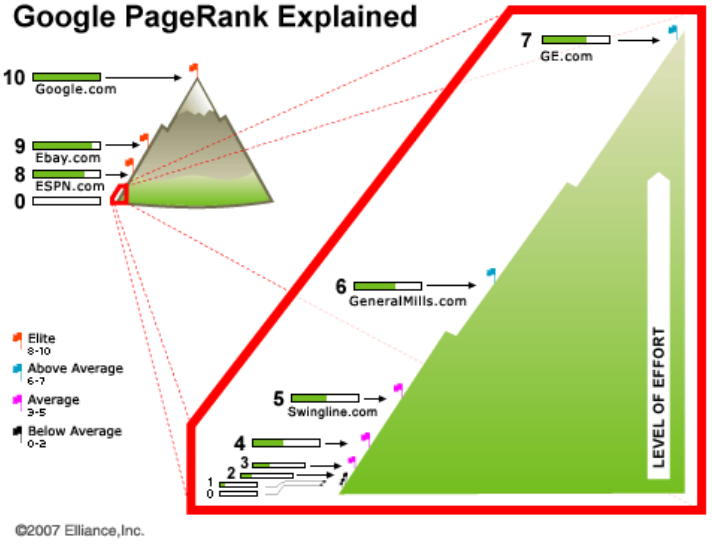


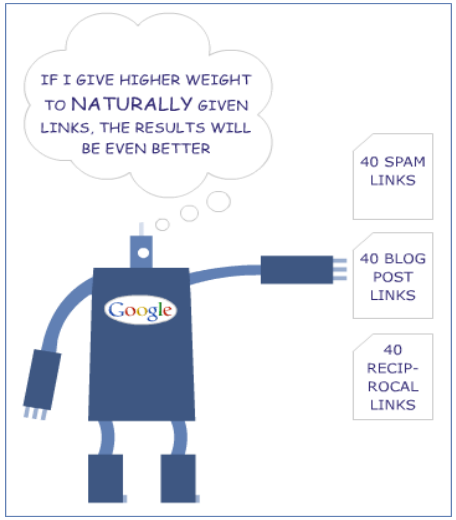

SIB Fellowship ▼
ExPASy
Bioinformatics Resource Portal
Query all databases
search
Visual Guidance
Categories
proteomics
genomics
structural bioinformatics
systems biology
phylogeny/evolution
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
drug design
Resources A..Z
Links/Documentation
ExPASy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.
Featuring today
STRING
Database of known and predicted protein-protein interactions
[details]
How to use this portal?

How to Find the Right Stuffs

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

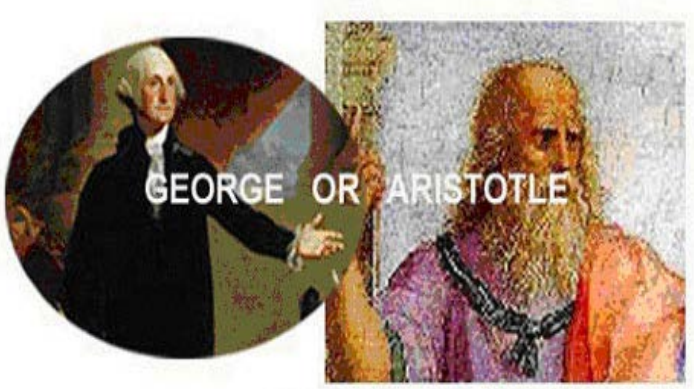
Fidelity: edu > gov > org > com



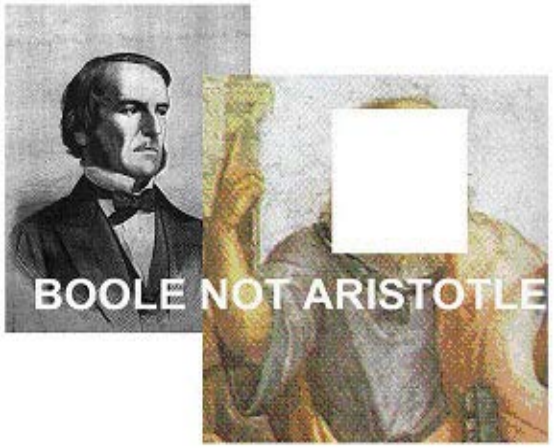


Search Efficiently

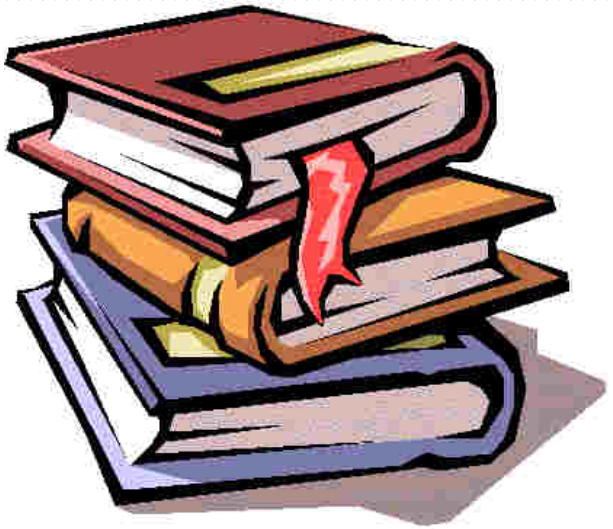
Quick Tours	Search PubMed by Authors
	My NCBI…
Boolean operators	**AND** OR **NOT**



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



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



Gene or Disease – Official Symbol

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

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

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

GeneCards® The Human Gene Compendium

PubMed.gov

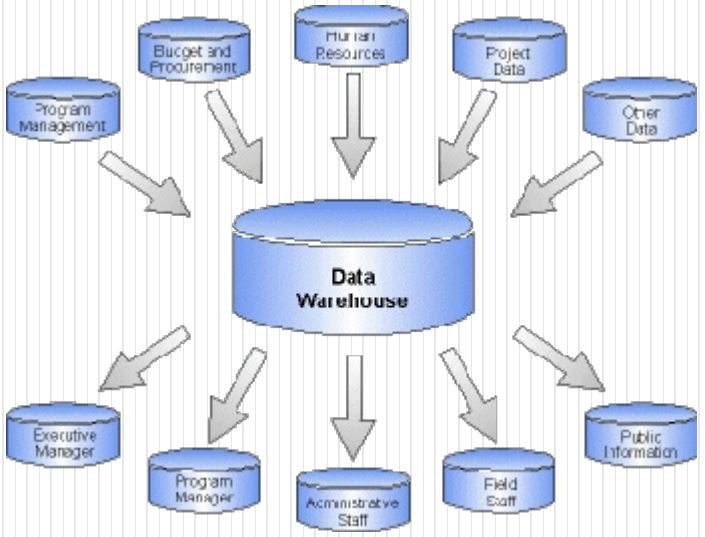
PubMed
- POU5F1

OMIM
- Preview and index

GeneCards/
human POU5F1 (symbol only)

Entrez_Gene
- POU5F1

Q: What is Derivative Databases?



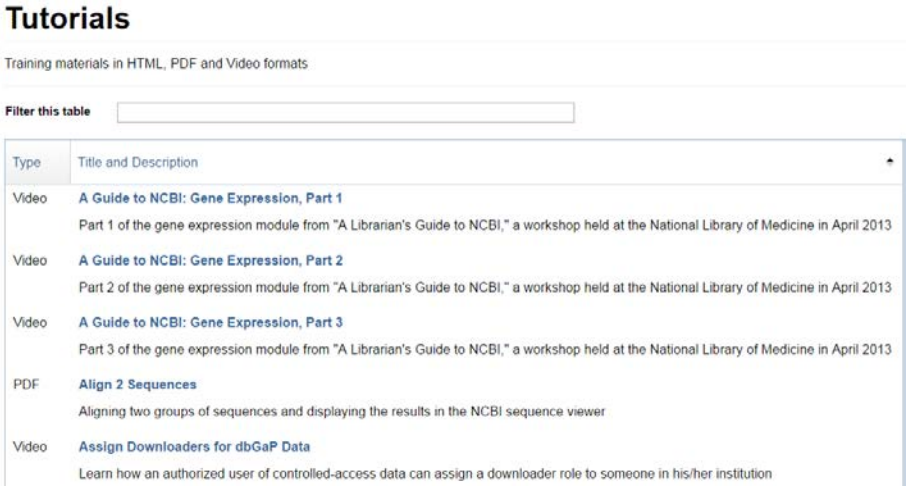
Leading Bioinformatic Centers

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**

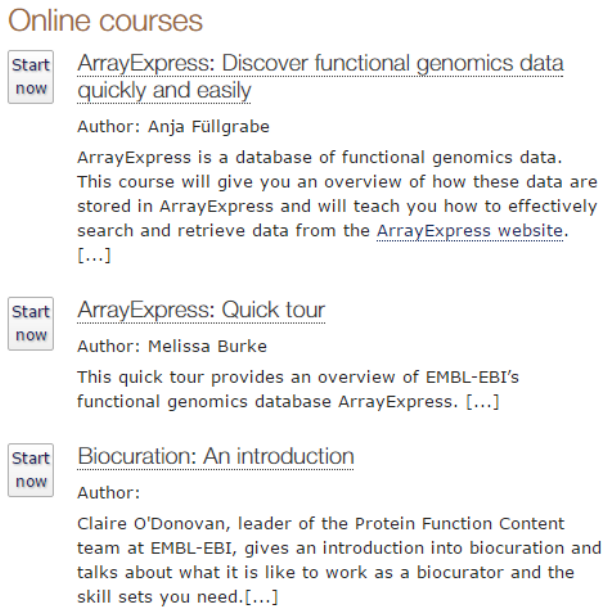


Tutorials

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

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

Author: Anja Füllgrabe

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

Start now — ArrayExpress: Quick tour

Author: Melissa Burke

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

Start now — Biocuration: An introduction

Author:

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

The National Center for Biotechnology Information (NCBI)

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

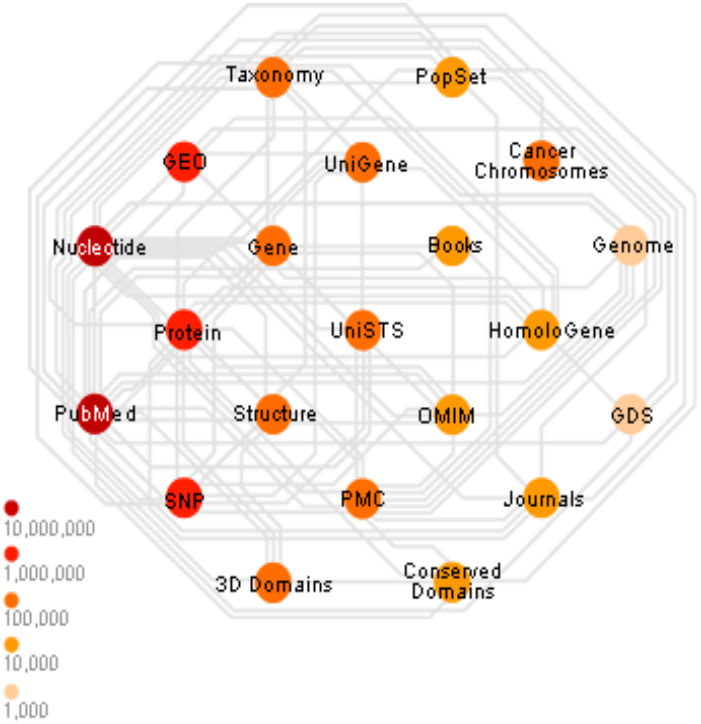
Contents

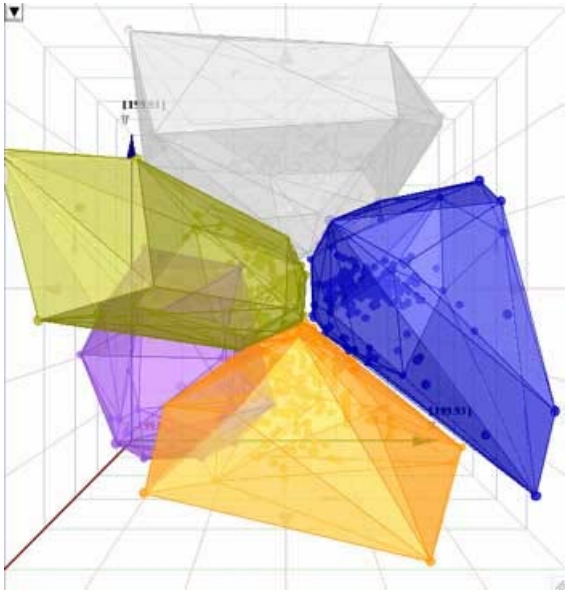
Databases

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

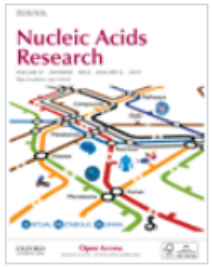
Methodologies (tools)

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





Nucleic Acids Research



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

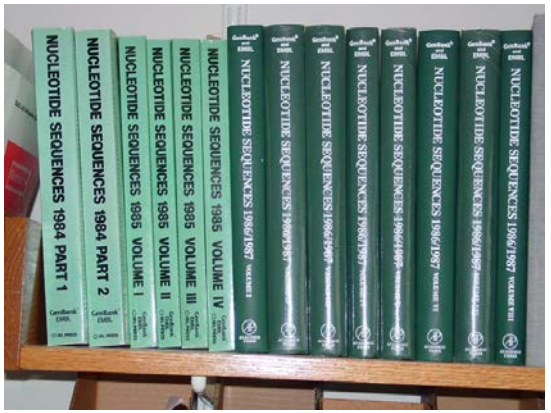
PDF | Split View | Cite | Permissions | Share ▼

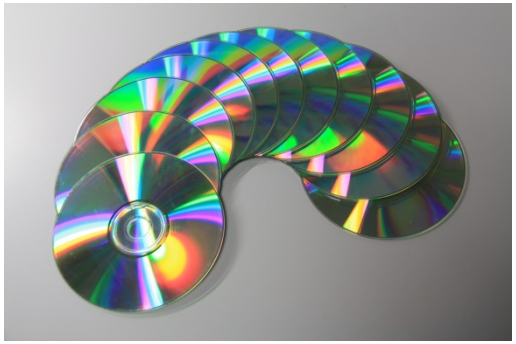
Abstract

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.

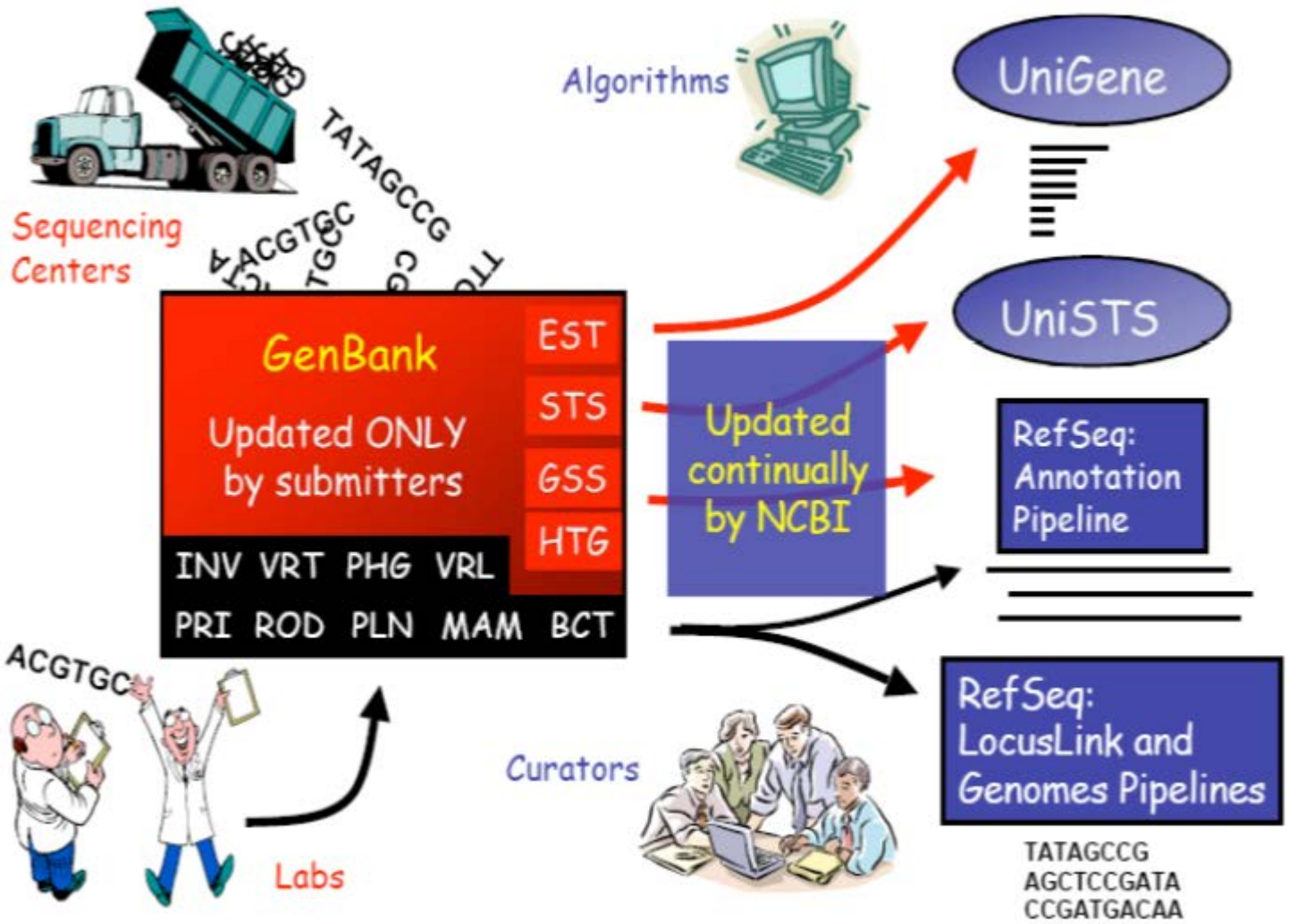
Primary vs. Derivative Databases - NCBI

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





Primary vs. Derivative Databases - NCBI



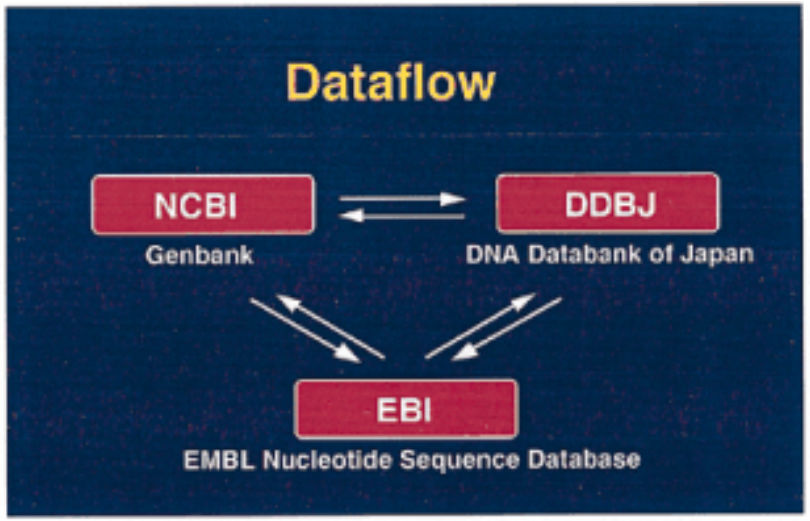
Primary DNA Databases

NCBI

EMBL NUCLEOTIDE SEQUENCE DATABASE

DDBJ

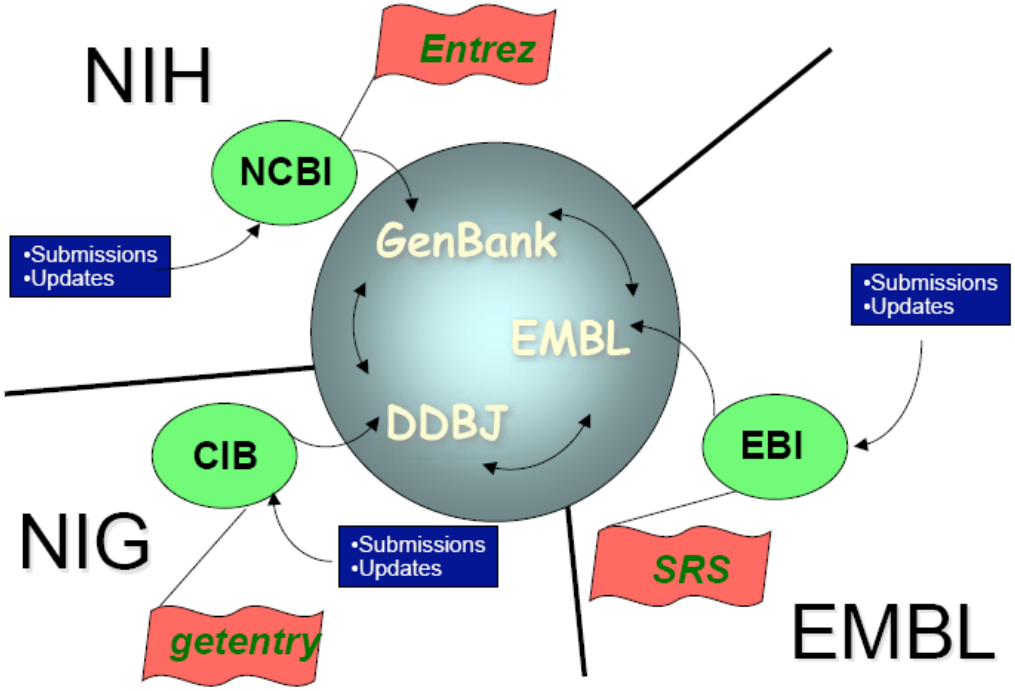
- GenBank (USA)
- EMBL (Europe)
- DDBJ (Japan)



National Institute of Health (**NIH**)

National Center for Biotechnology (**NCBI**)

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



National Institute of Genetics (NIG)

Center for Information Biology (CIB)

Research Organization of Information and Systems
National Institute of Genetics

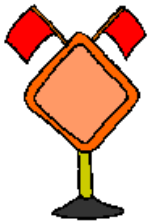
The European Bioinformatics Institute (**EBI**)

Sequence Retrieval System (**SRS**)

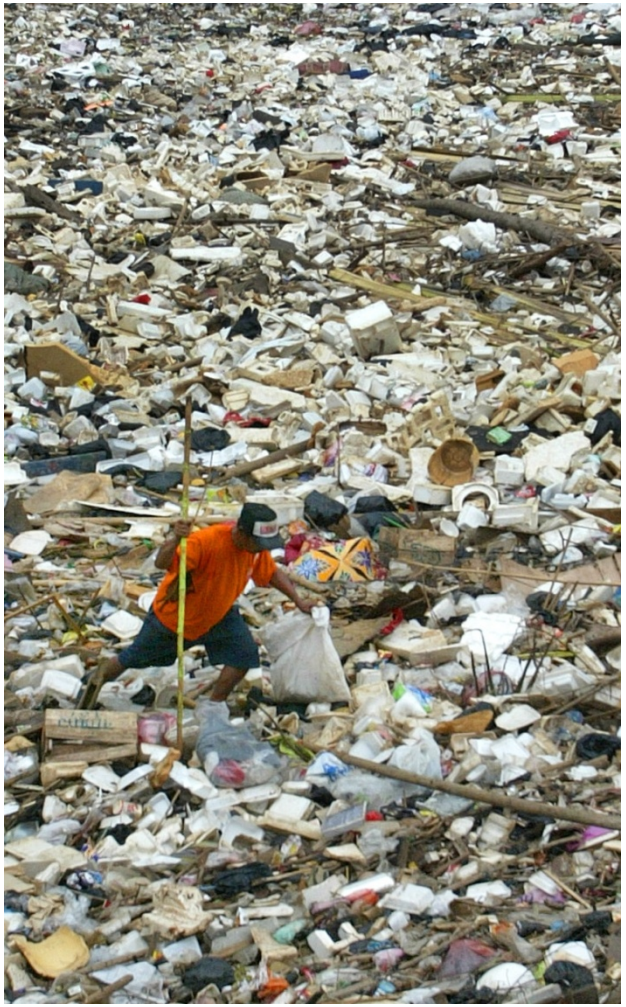
The *European Molecular Biology Laboratory* (***EMBL***)

EMBL/GenBank/DDBJ Annotations

Warning!!!



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



Some Interesting Sequence Annotation

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

Or:

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

???

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]`

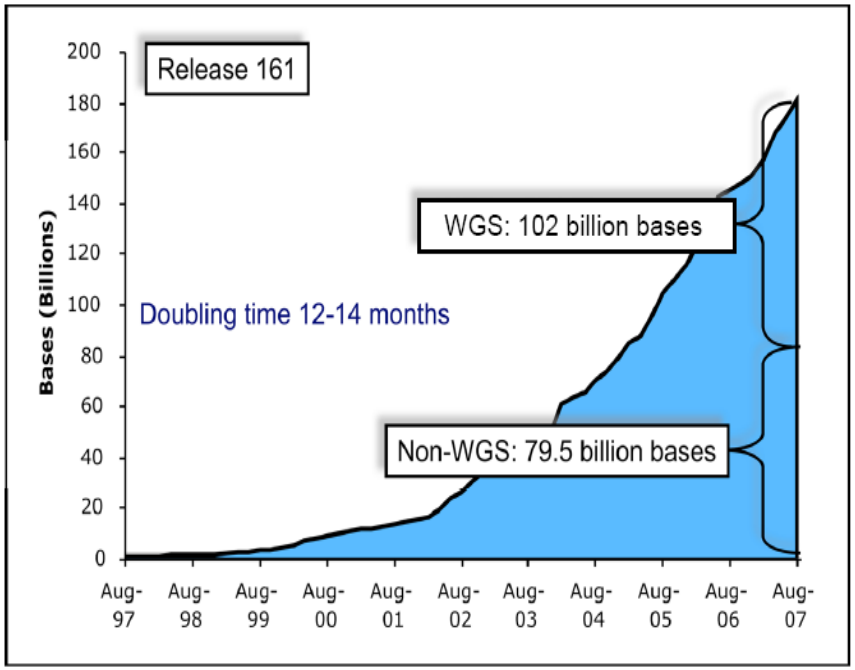
Bulk GenBank Divisions

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

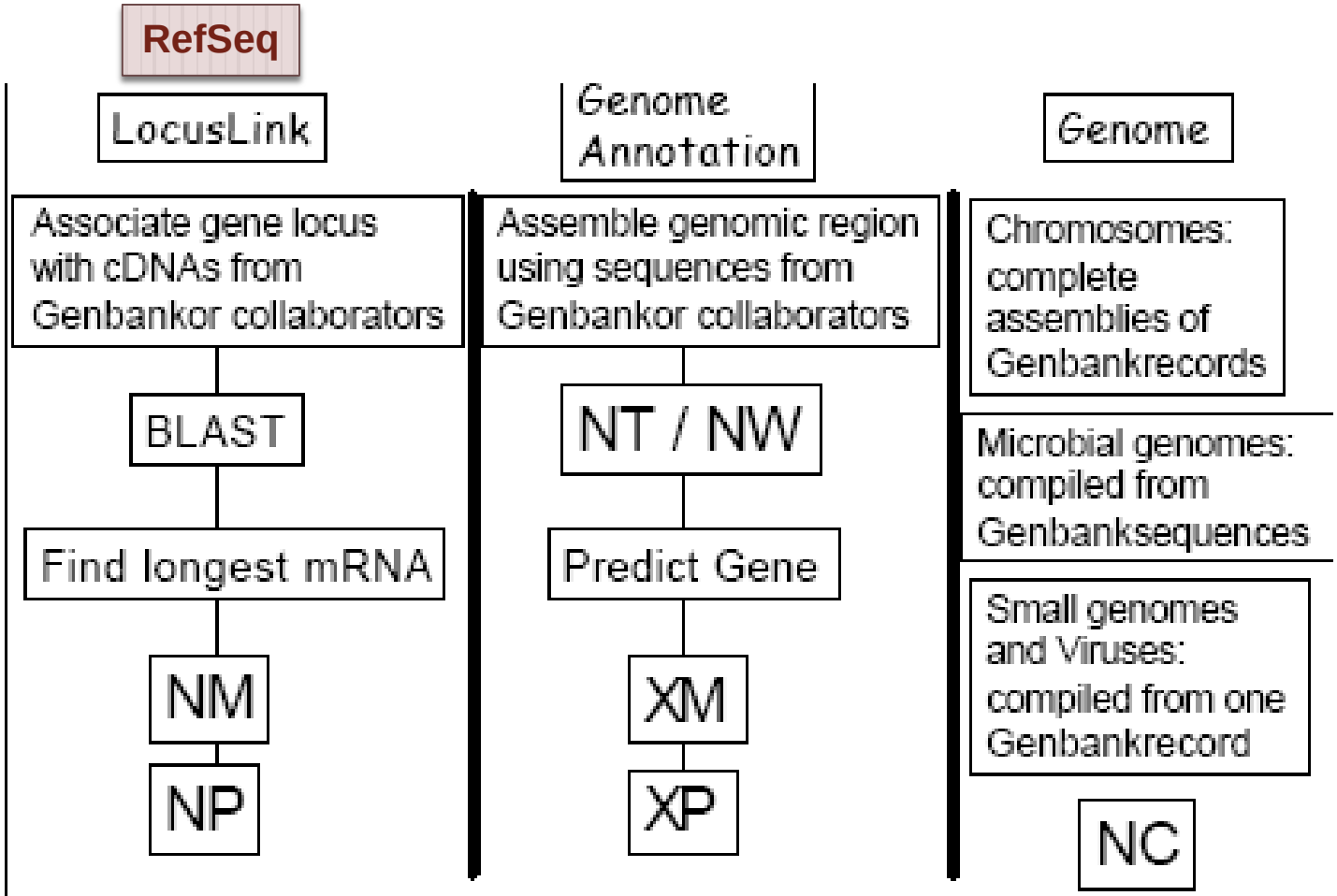
```
EST Expressed Sequence Tag
GSS Genome Survey Sequence
HTG High Throughput Genomic
STS Sequence Tagged Site
HTC High Throughput cDNA
PAT Patent
```

BULK Divisions:

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

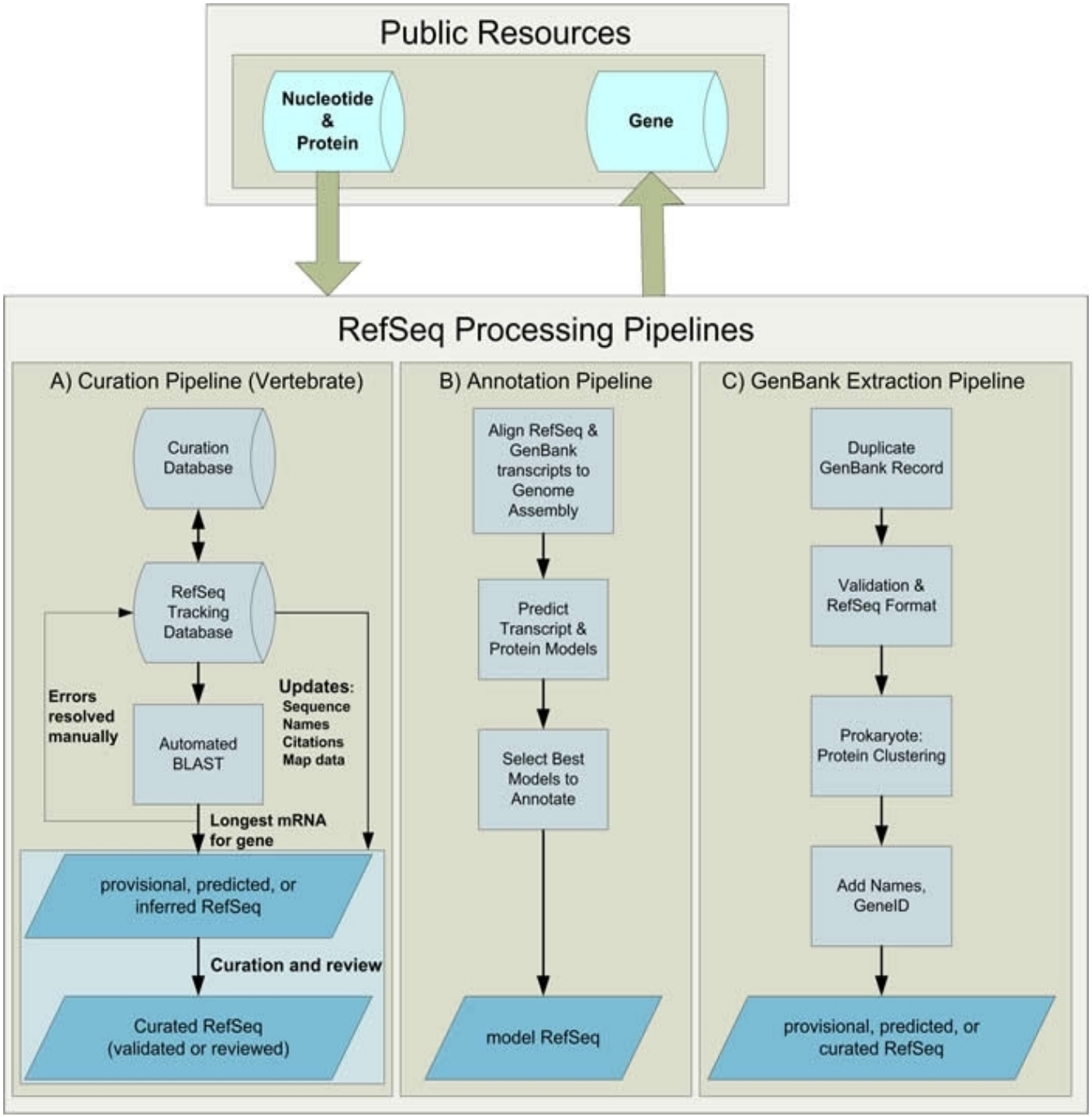
Entrez query: `gbdiv_xxx[Properties]`

RefSeq Pipelines



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

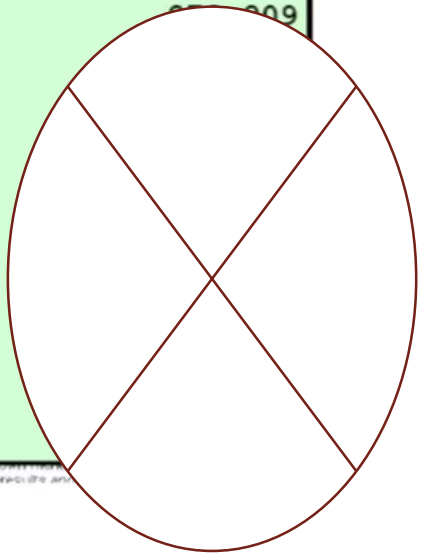
Stewardship by **NCBI staffs** **& collaborators**

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"
```

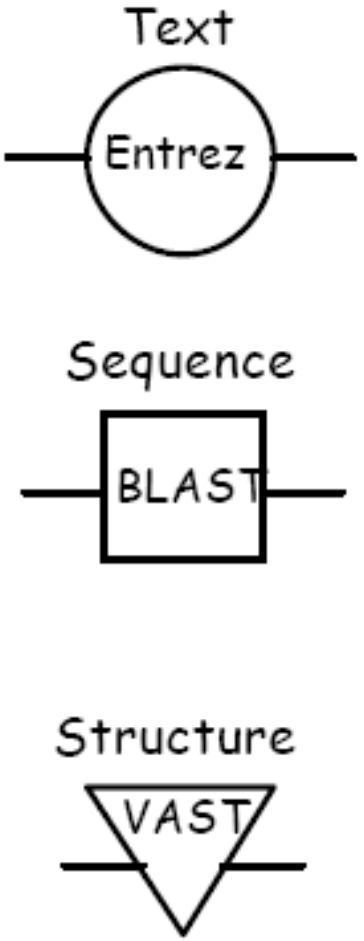
Data Source	Sequences
GenPept	11,585,396
RefSeq	3,889,502
Third Party Annotation	5,263
Swiss Prot	
PIR	
PRF	
PDB	
(PAT Division	
Total	
BLAST nr total (no patents or env_nr -now 6 million)	

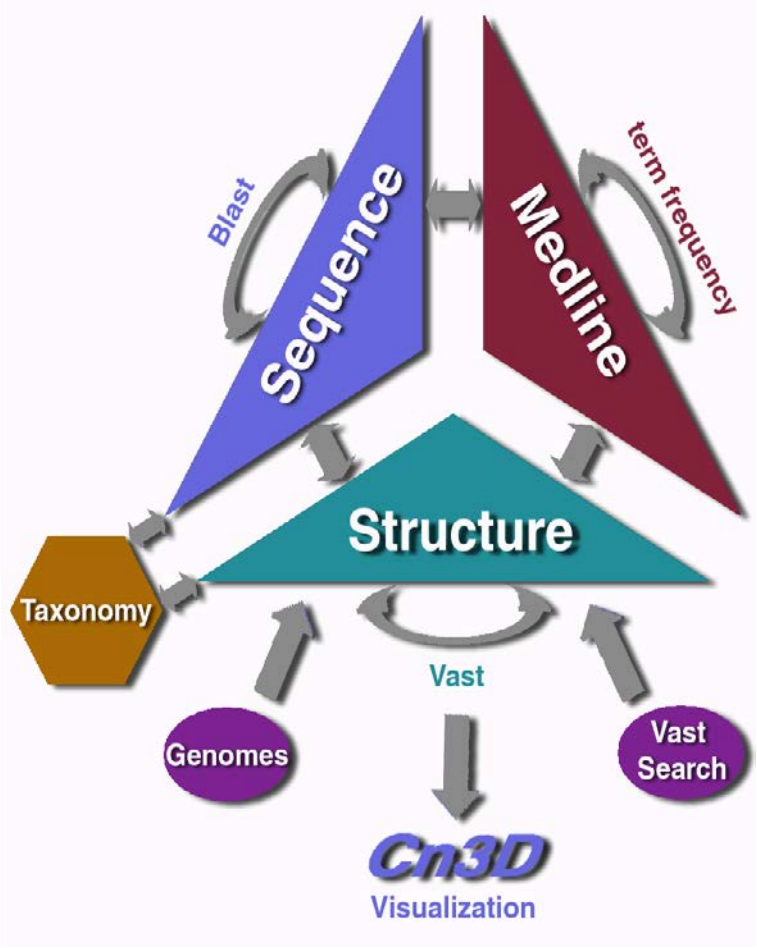


PAT: patent

Search in NCBI Databases

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



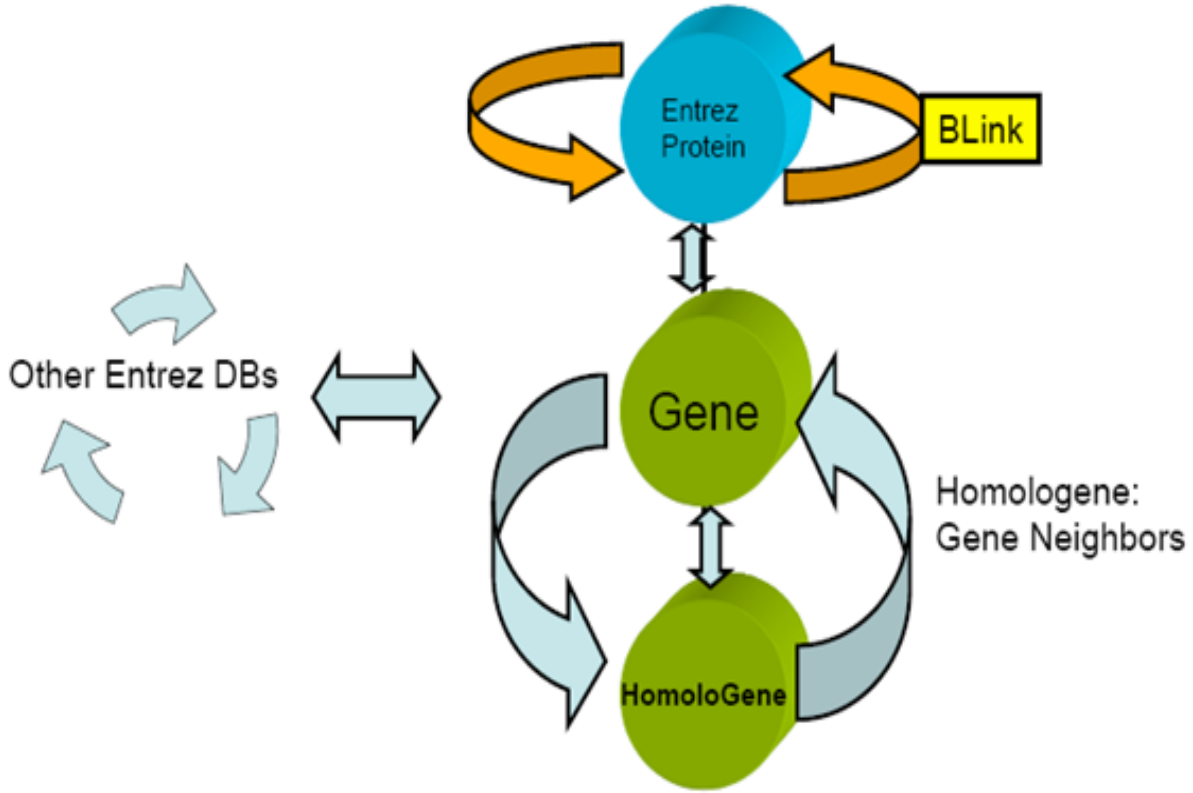


▼ Links Explain

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

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

Entrez: Use Gene for everything

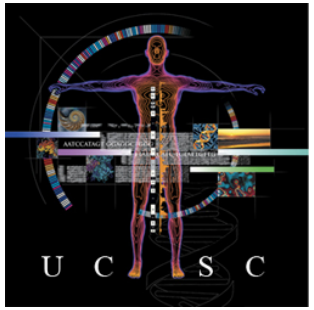


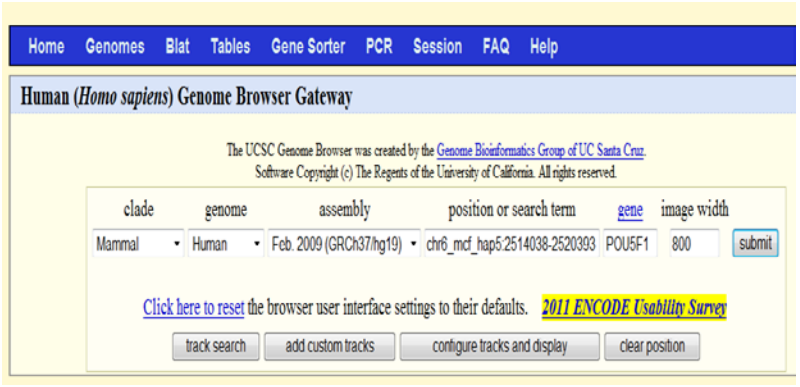
Examples in Other Databases: Using the Official Symbol All the Time (except for protein structure)

STRING 8.3

NPD

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





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

PDBTM: Protein Data Bank of Transmembrane Proteins

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

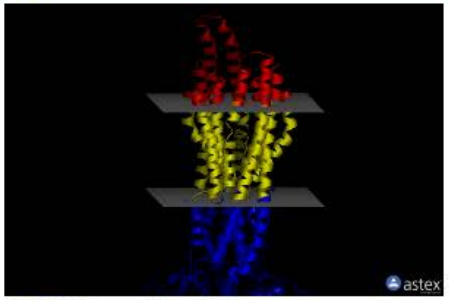
all 1a0s

- Home
- Search
- Download
- Statistics
- Documents
- Help

Welcome to the PDBTM home page

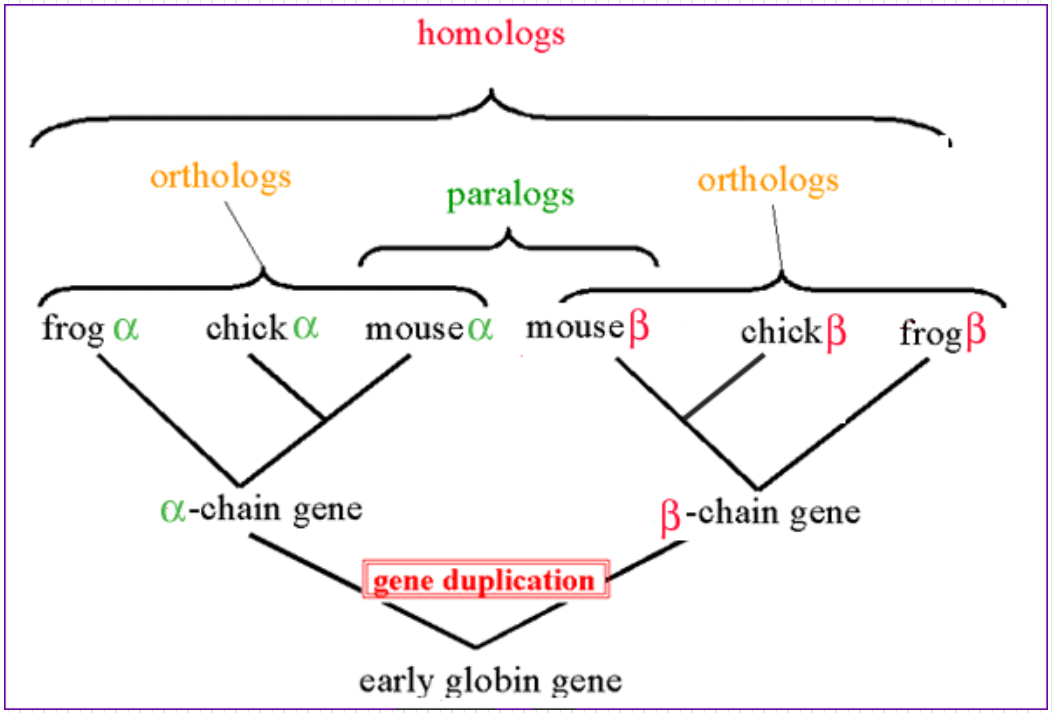
PDBTM is the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained at the Institute of Enzymology by the Membrane Protein Bioinformatics Research Group. The PDBTM database was created by scanning all PDB entries with the TMDET 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



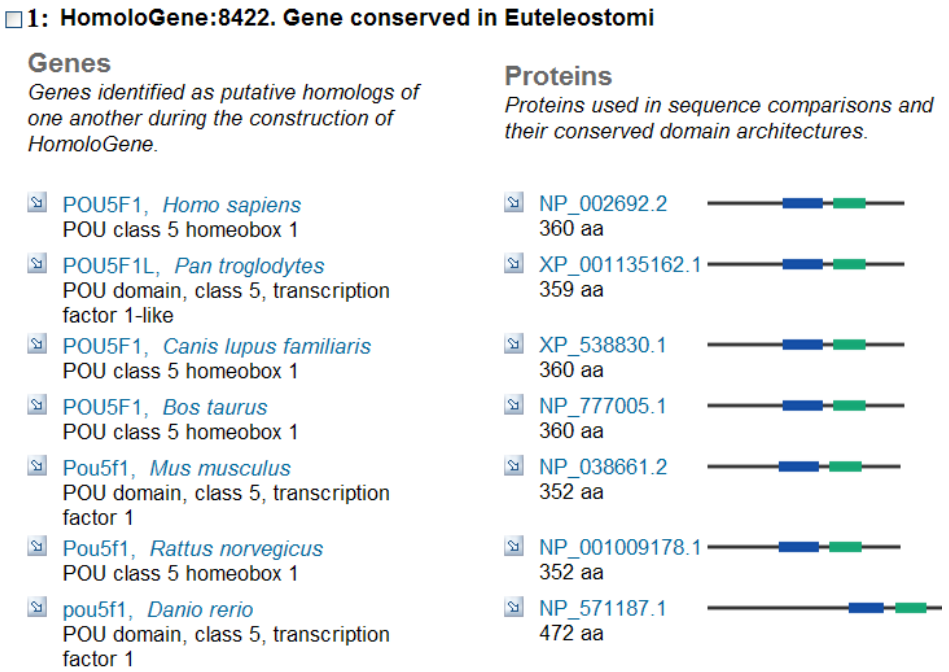
Homologs (1)

NCBI_Homologene (links)

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

Example

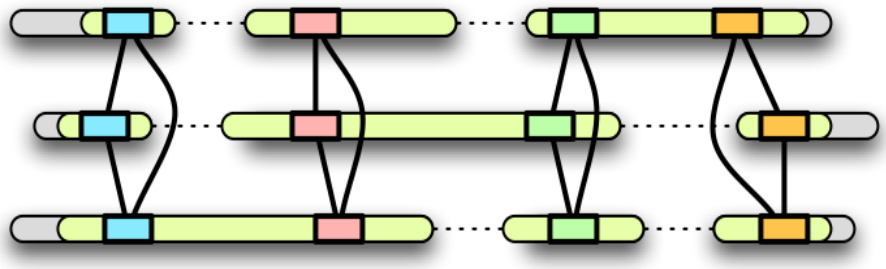
- ***POU5F1*** (via ENTREZ_GENE) **Links** to the "Homologene"
  - Protein: multiple alignment
  - Conserved domains
  - PubMed (references)
  - Protein → All links from this record → BLink



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

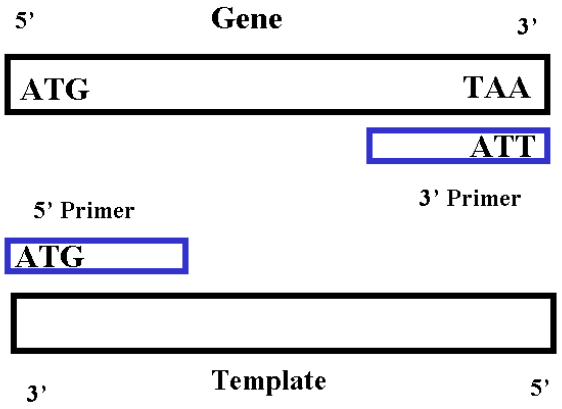
- Mouse Genome Informatics
  - Mm: *Pou5f1* (chr. 17; 19.23 cM)

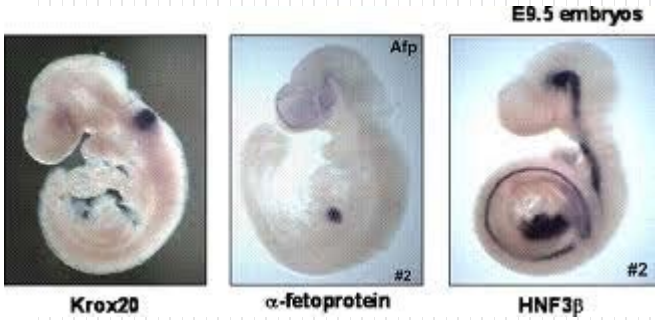


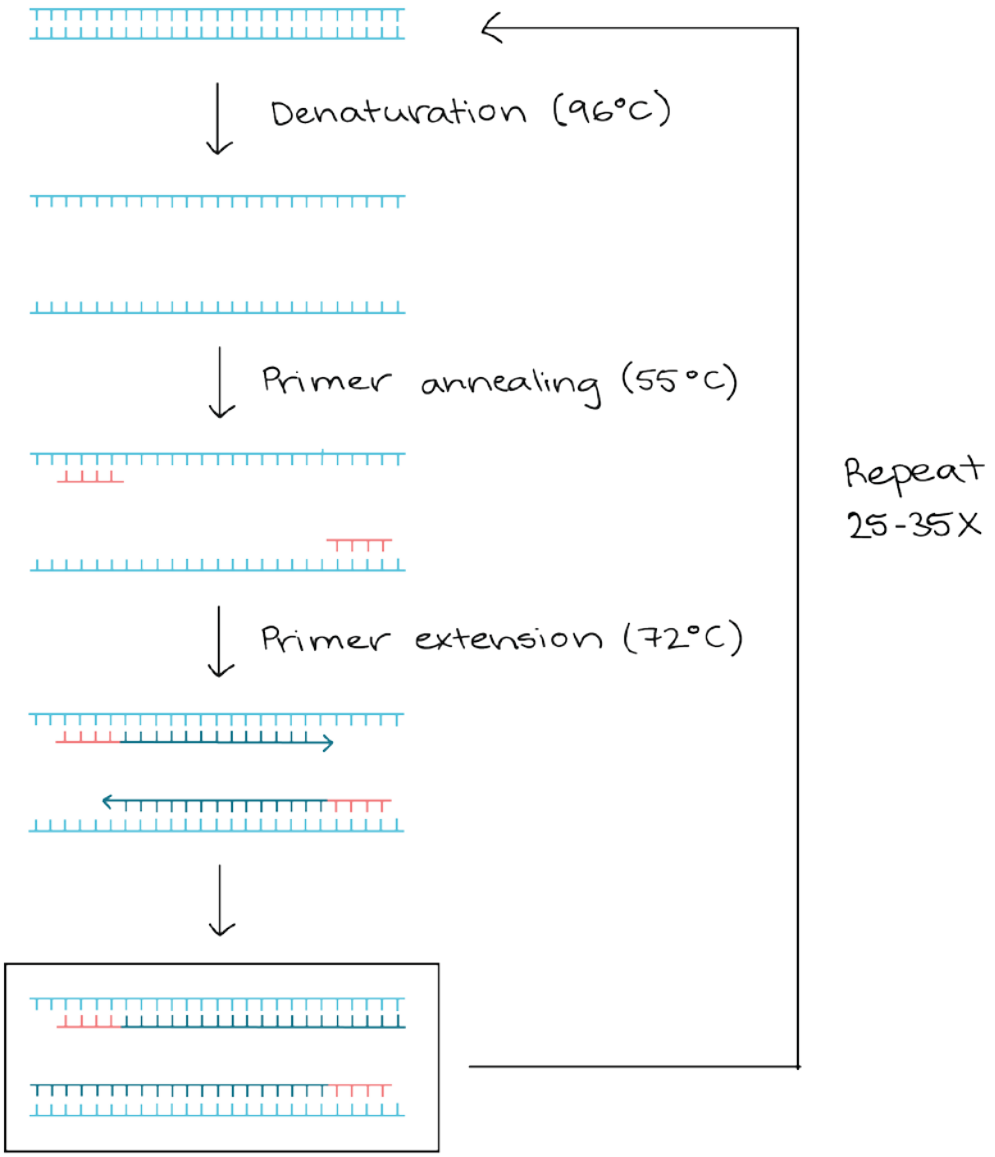
Mercator

Multiple Whole-Genome Orthology Map Construction

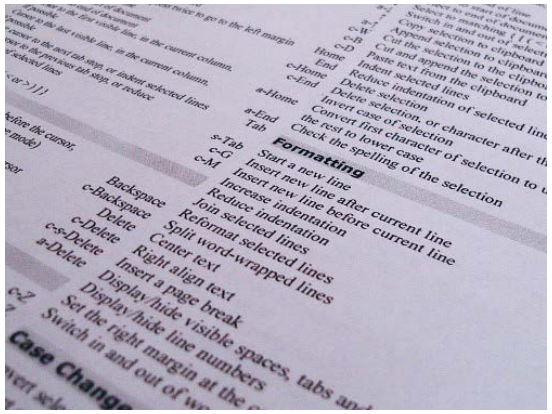
Q: How to Design Primers/Probes for PCR/qPCR/Cloning/in situ hybridization







Result after 1 cycle:
# of DNA molecules
doubled



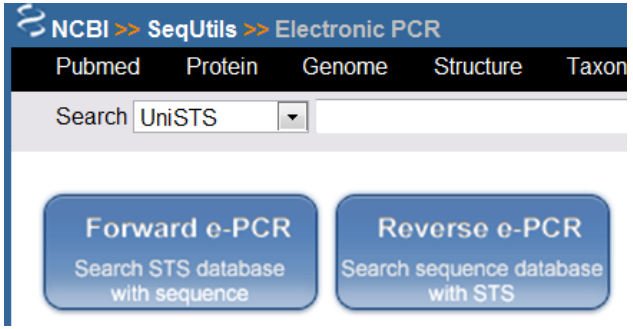
RT PRIMER DB

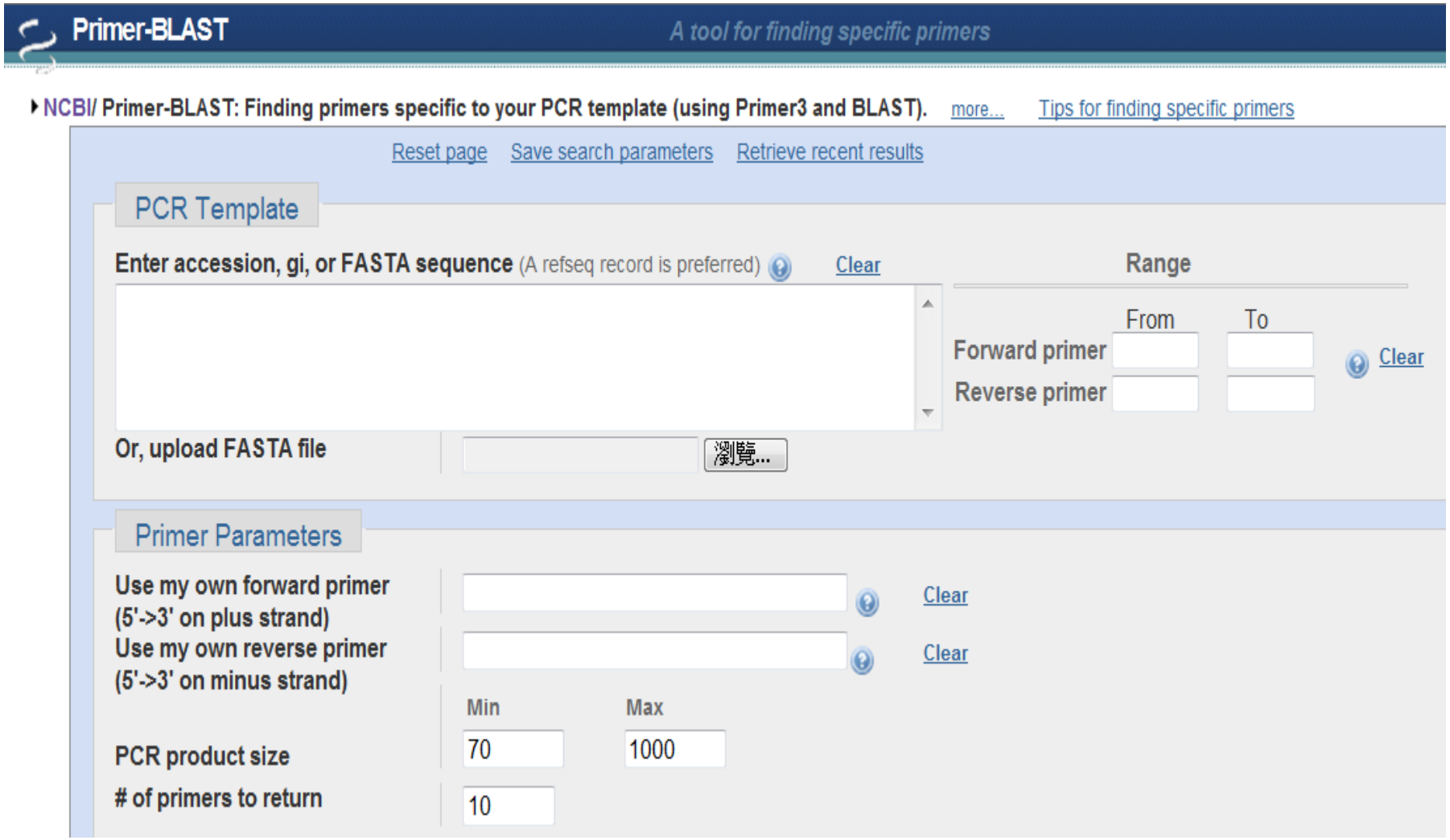
e.g., ACTB

BLAST

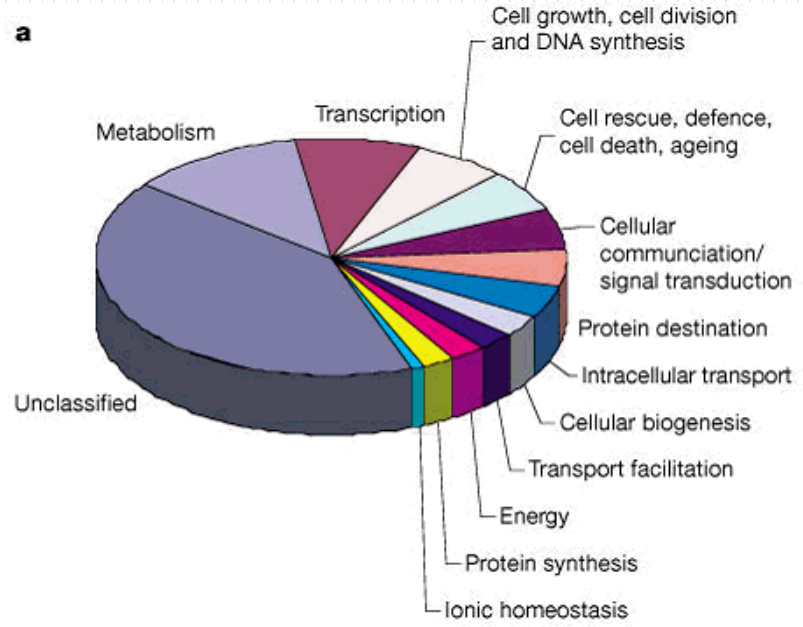
BLAST

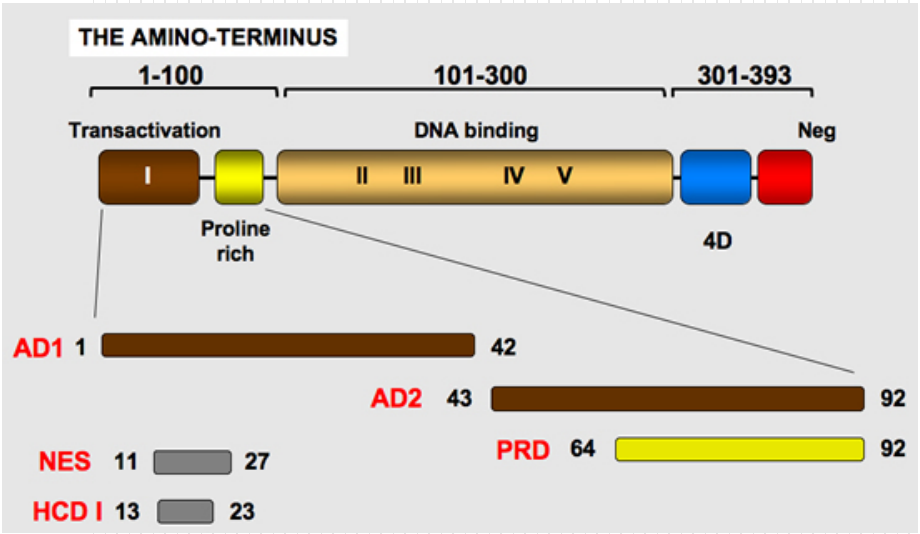
Literatures → Databases → *ab inito* design




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

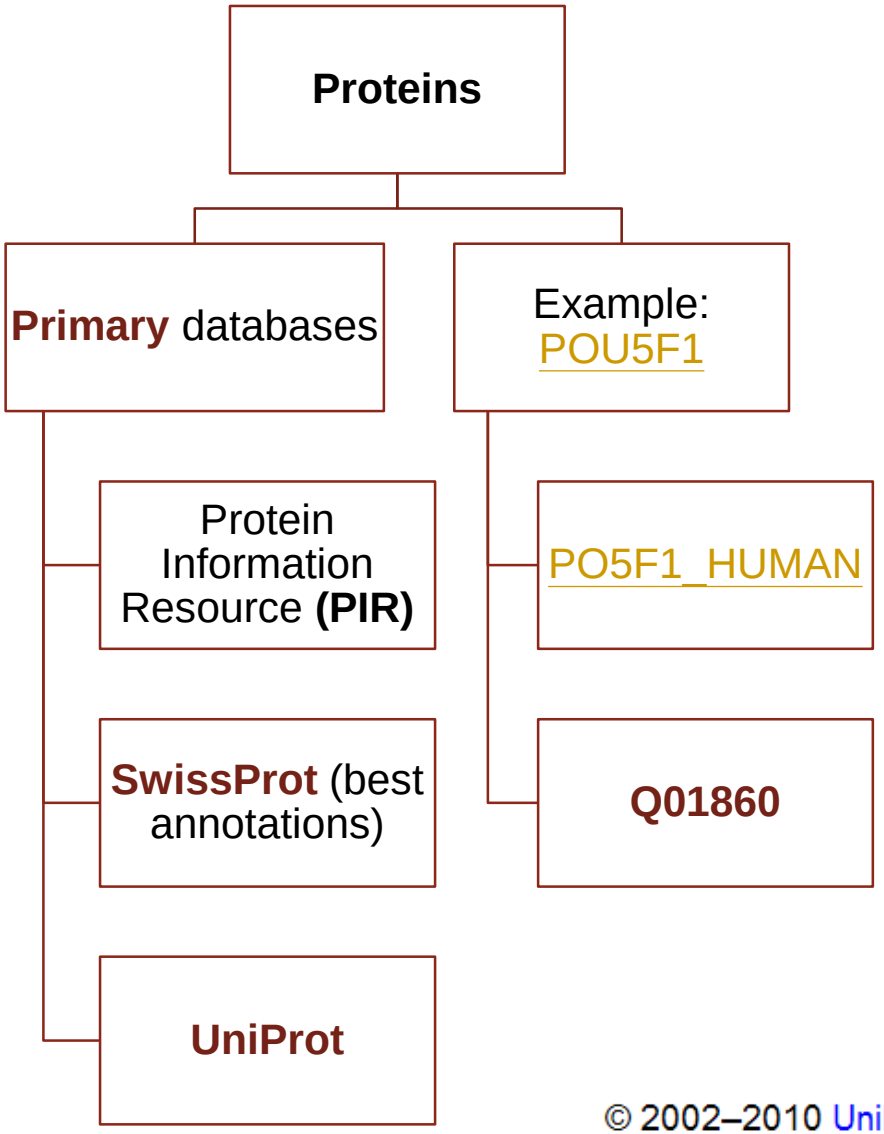
Automatically-assigned Evidence Codes

- IEA: Inferred from Electronic Annotation

Obsolete Evidence Codes

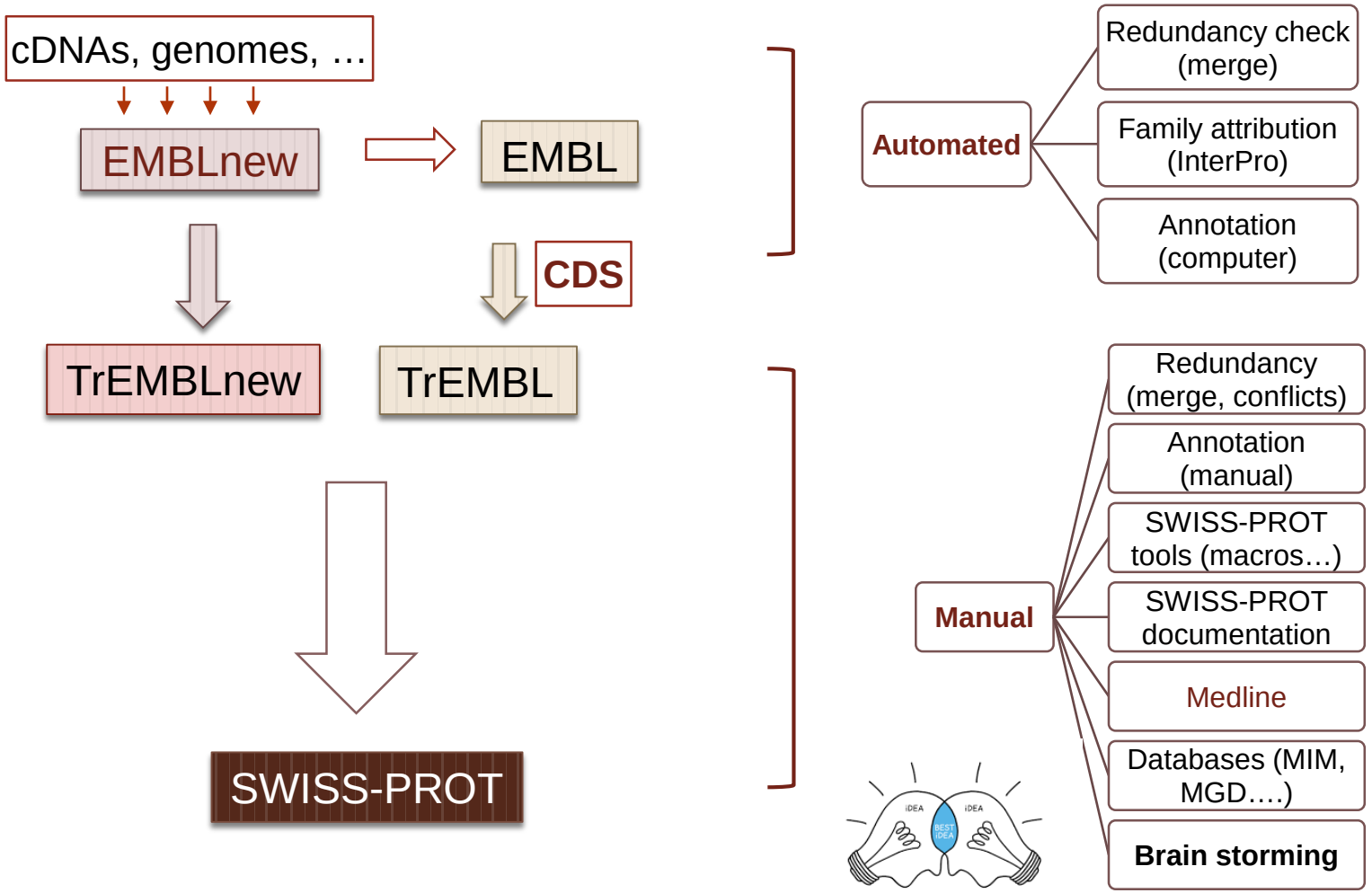
- NR: Not Recorded

Note on Usage of the With/From Column



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

The Simplified Story of a SWISS-PROT Entry



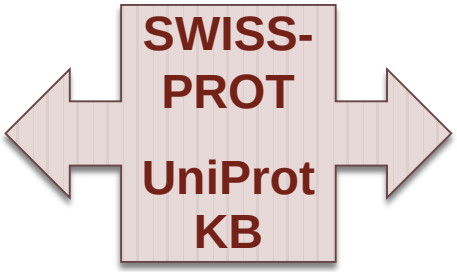
Once in SWISS-PROT, the entry is no more in TrEMBL, **but still in EMBL (archive)**

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

2D and 3D Structural dbs
HSSP
PDB

PTM
CarbBank
GlycoSuiteDB

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



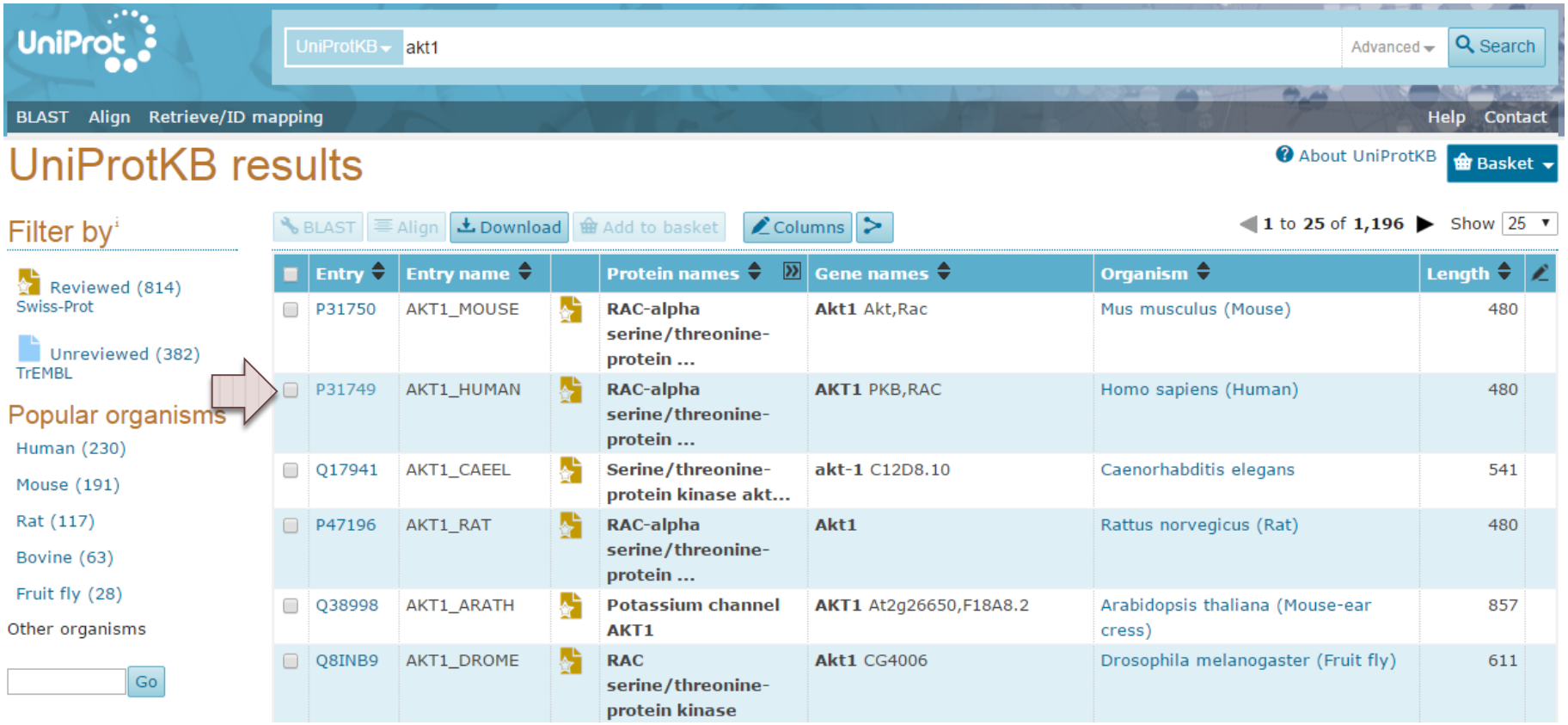
Nucleotide sequence DB
EMBL, GeneBank, DDBJ

Human diseases
MIM

Protein-specific dbs
GCRDb
MEROPS (peptidase)
REBASE
TRANSFAC

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

2. UniProt/InterProt Annotations



Display

- Entry
- Feature viewer
- Feature table

None

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

Top

Family & Domains

Domains and Repeats

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

Region

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

Domain

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

Sequence similarities

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

Contains 1 AGC-kinase C-terminal domain. Curated

Contains 1 PH domain. PROSITE-ProRule annotation

Contains 1 protein kinase domain. PROSITE-ProRule annotation

Display

- Entry
- Feature viewer
- Feature table

None

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

Top

PTM / Processing

Molecule processing

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

Amino acid modifications

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

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

Stats & Filters

Current Stati

High Throughput

10 (59%)

0 (0%)

Search Filter

No Filter: Show

Switch View: Summary | Sortable Table

Displaying **13** total unique interactors

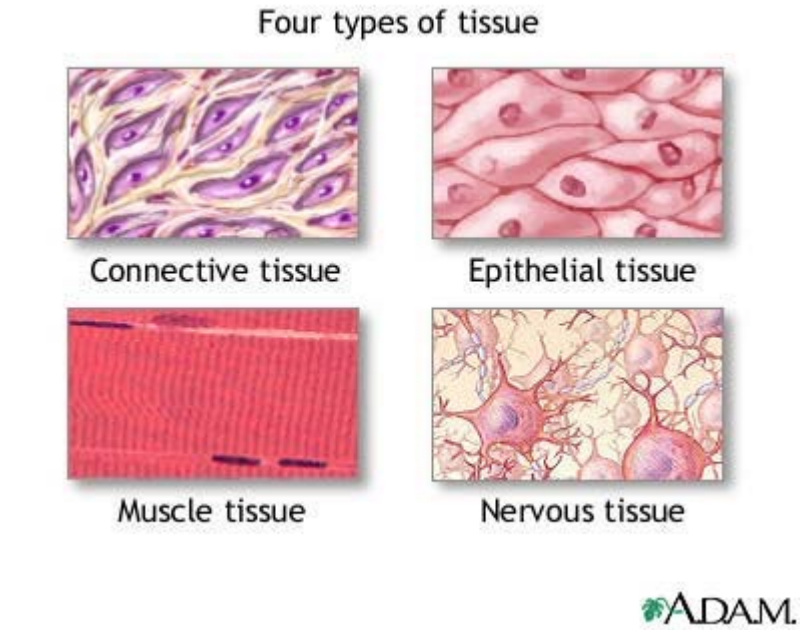
POU5F1 | Otf-3, Oct3, Oct-3/4, Otf3, Oct3/4, Oct-3, Oct4, Otf-4, Oct-4, Otf3-rs7, Otf4, Otf3g
POU domain, class 5, transcription factor 1

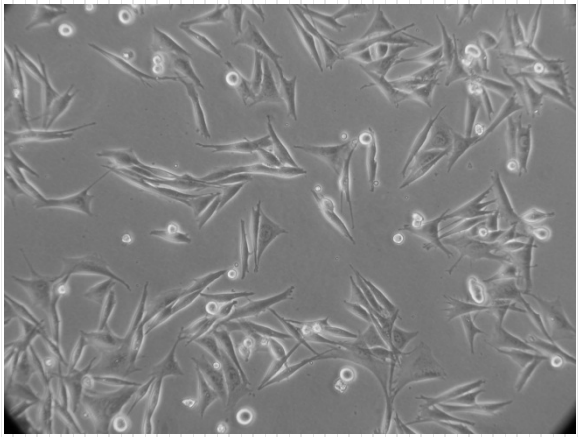
MTA2 | mmta2, Mta1l1, Mata1l1, AW550797
metastasis-associated gene family, member 2

Databases for Protein – Protein Interaction

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

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






Modify Query
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
OncoPrint
Cancer Types Summary
Mutations
Survival
CN Segments
Expression
Download
POU5F1
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
15
10
5
pression --- RNA Seq V2 (log2)

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

Bookshelf

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

Search This Book Limits Help

Bookshelf ID: NBK1969

NCBI News

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

Copyright notice.

Index of Issues

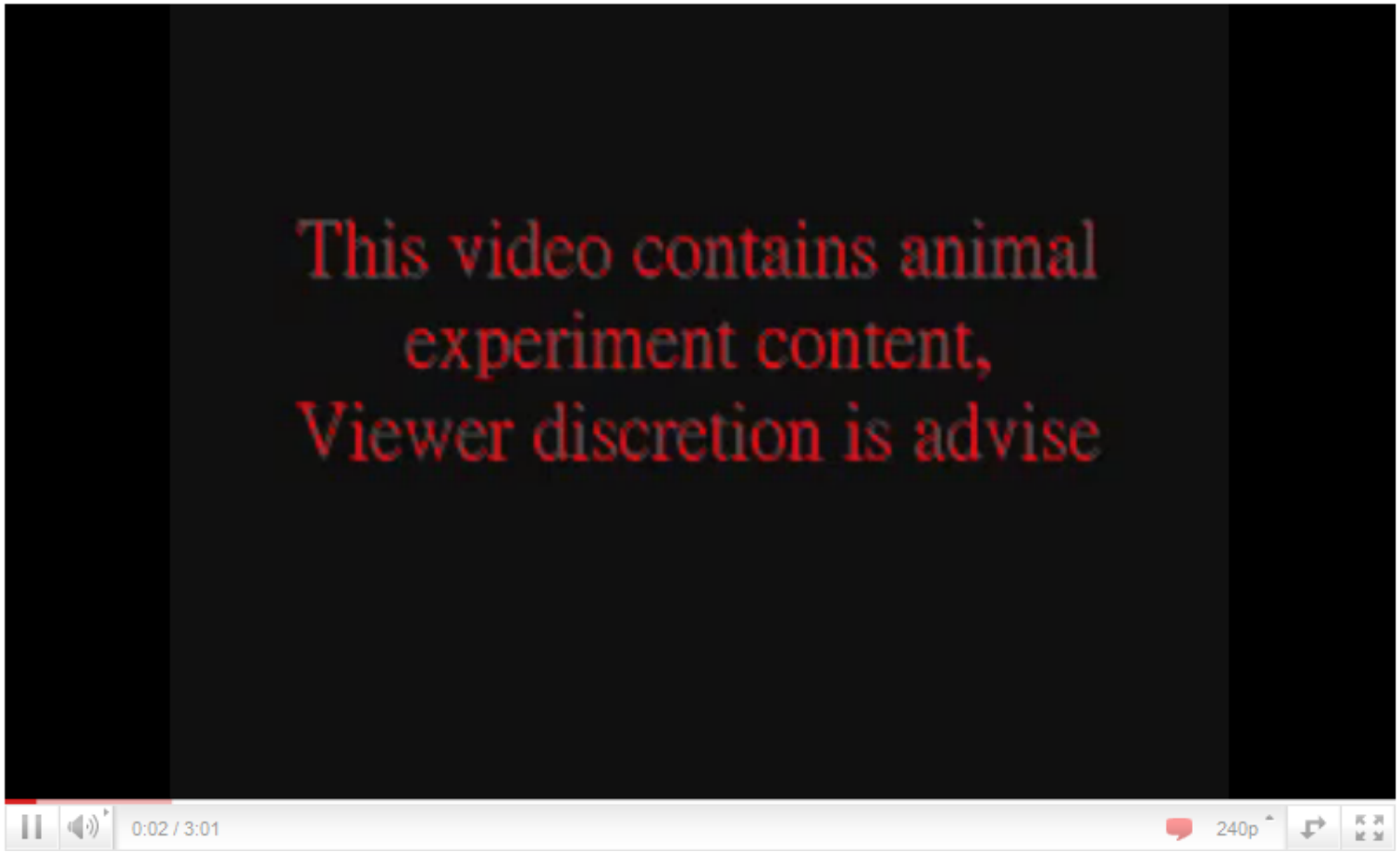
Expand All

- 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

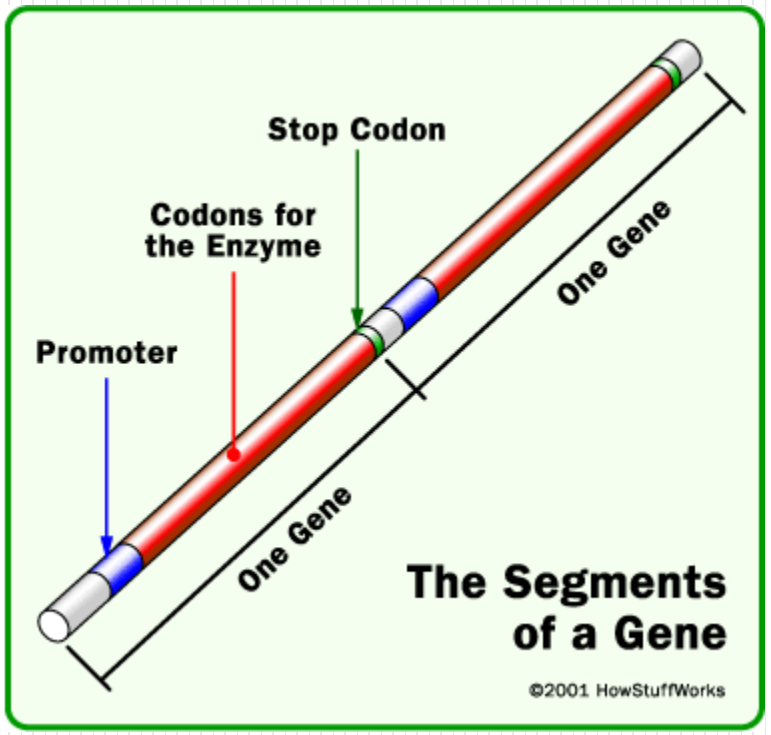
YouTube liver tumor mouse

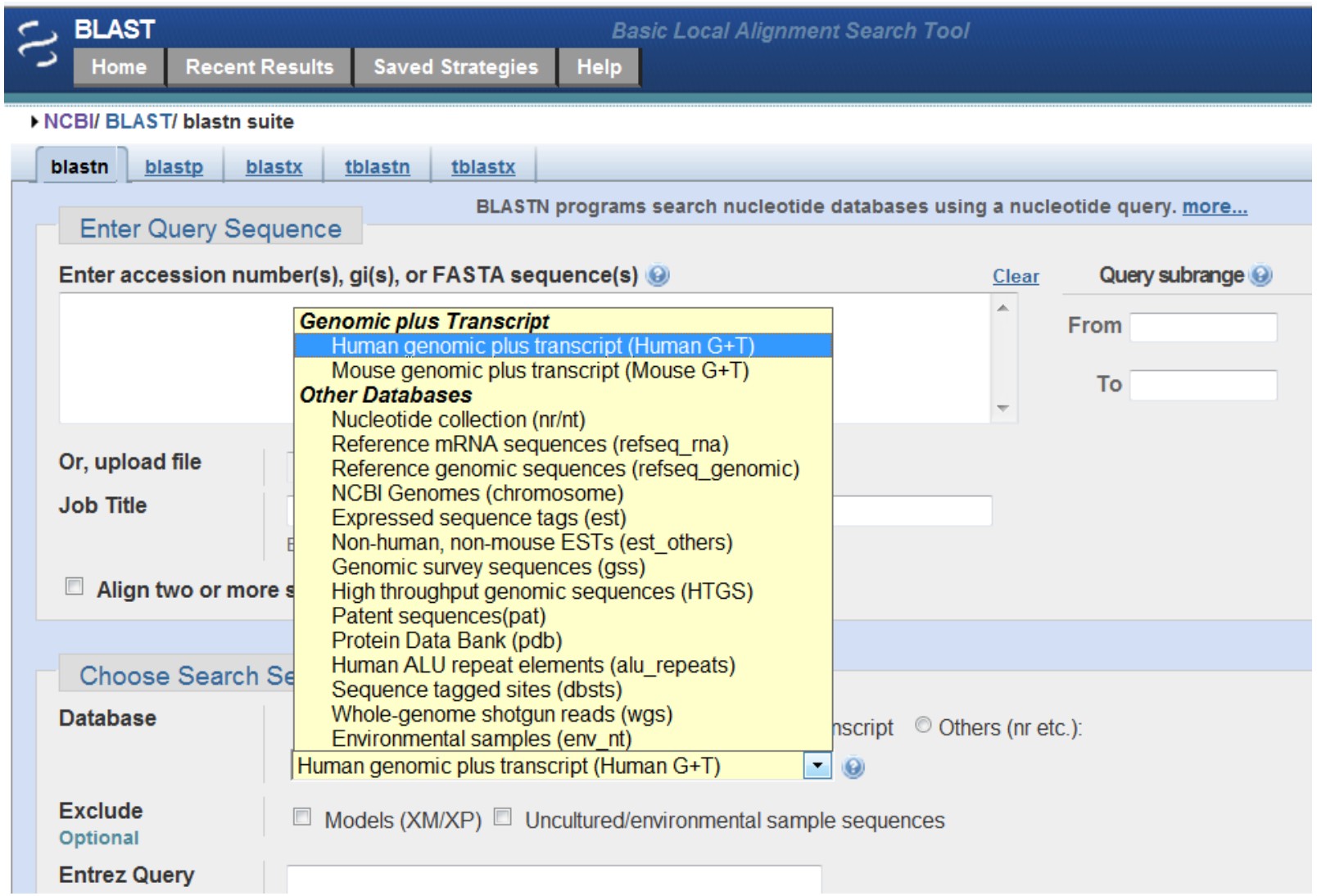
Seed tumor at liver of mouse-Surgery 種腫瘤在肝臟-開腹腔篇(一)

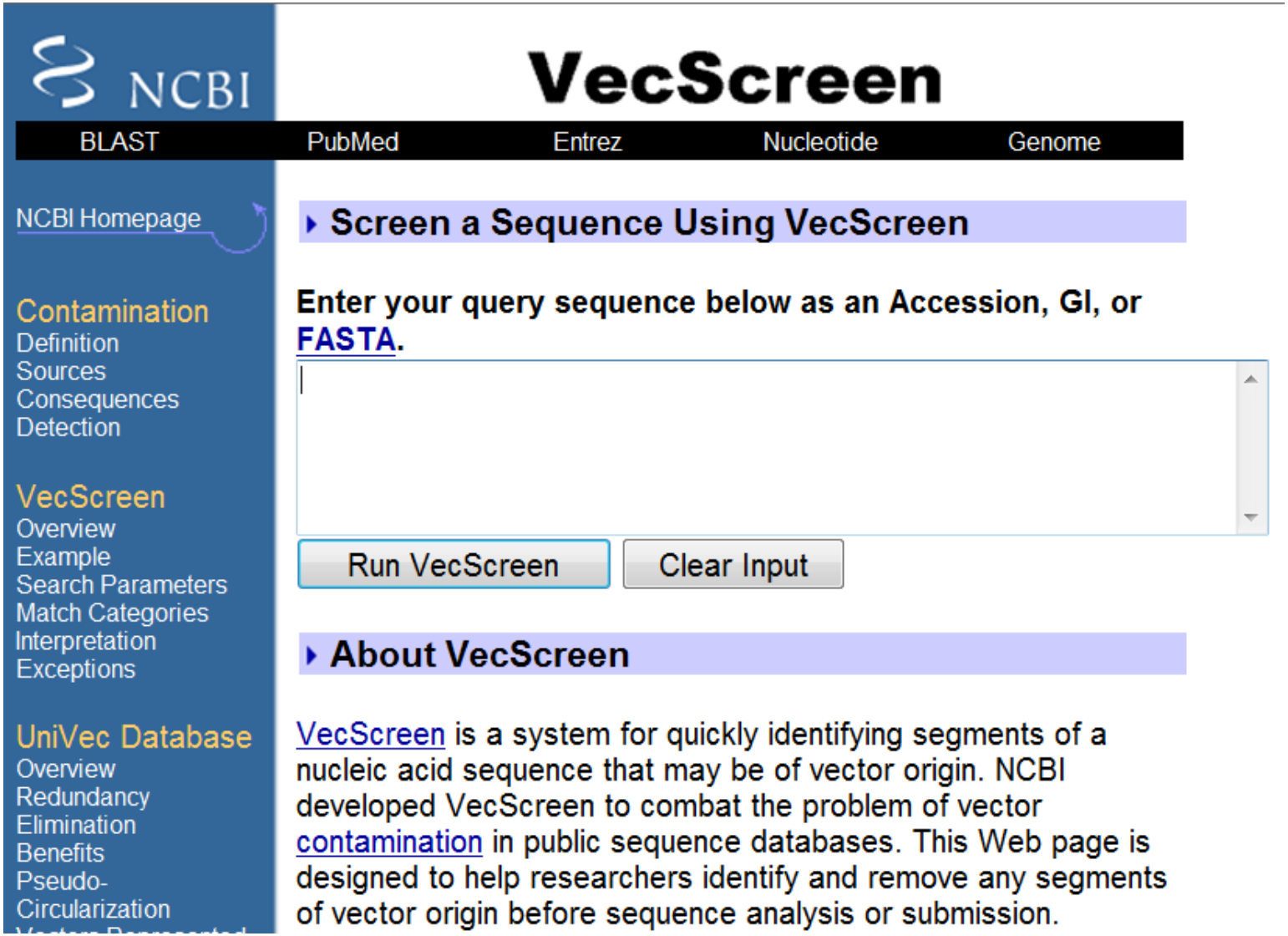
miss9ch 282 部影片 訂閱



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




BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
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)
Or, upload file
Job Title
Align two or more s
Choose Search Se
Database
nscript
Others (nr etc.):
Human genomic plus transcript (Human G+T)
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query


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

ORF Finder (Open Reading Frame Finder)

NCBI

PubMed Entrez BLAST OMIM Taxon

NCBI

Tools
for data mining

GenBank
sequence submission
support and software

FTP site
download data and
software

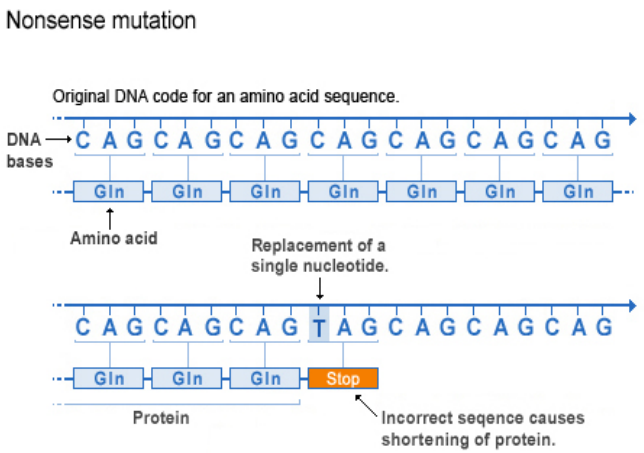
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which fi
selectable minimum size in a user's sequence or in a sequence already in the datal
This tool identifies all open reading frames using the standard or alternative genetic
sequence can be saved in various formats and searched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

Enter GI or ACCESSION OrfFind Clear

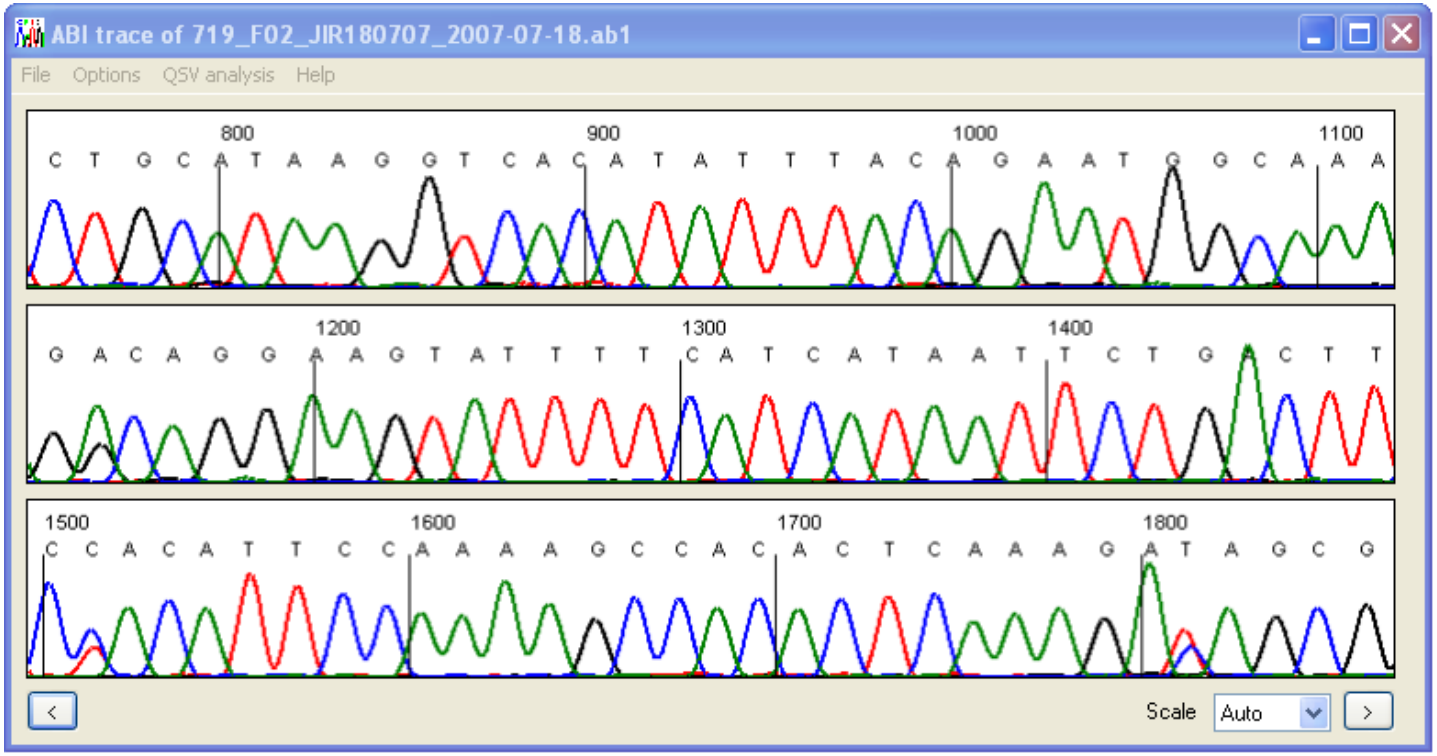
or sequence in FASTA format

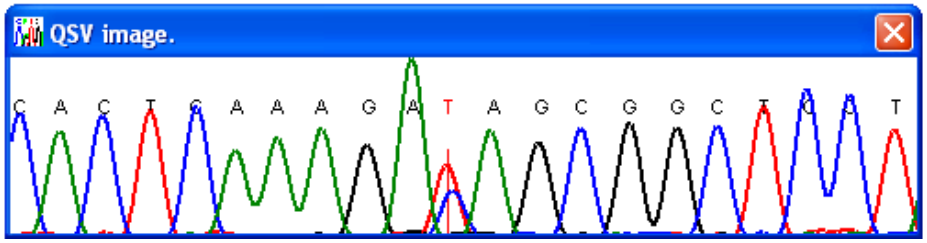
FROM: **TO:**

Genetic codes 1 Standard



U.S. National Library of Medicine


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


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

When Cloned by Emails – get the map & confirmed

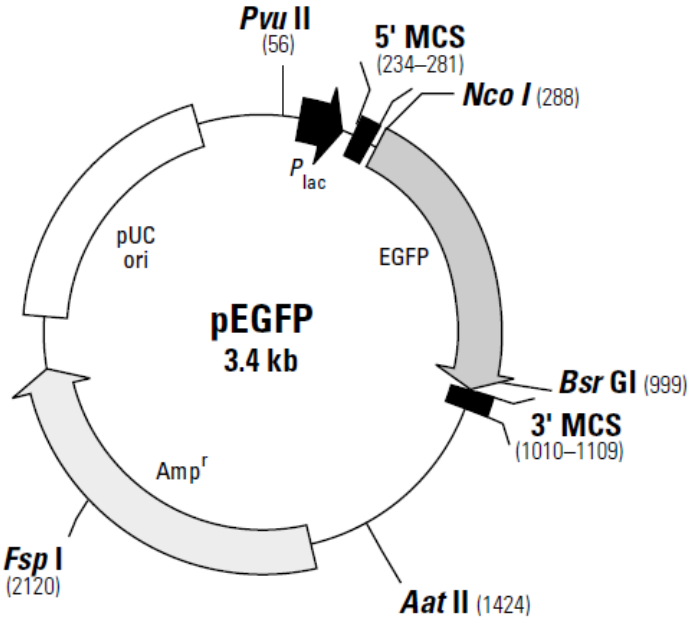
Specific EGFP Monoclonal Antibody for Westerns, IP and IC

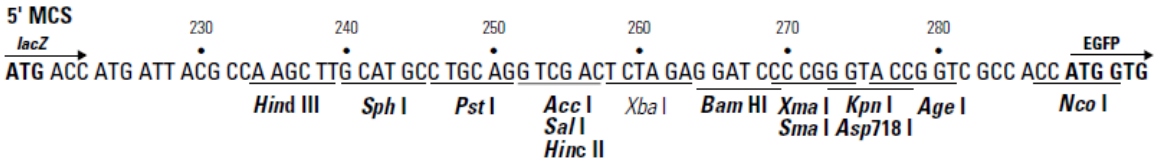
Visit our website for more details! click here...

pEGFP Vector Information

PT3078-5

Catalog #6077-1





Q: How to Get a Specific Sequence from Genome Databases

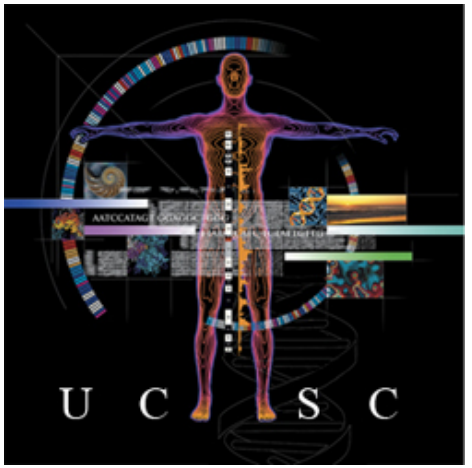
Genome Biology

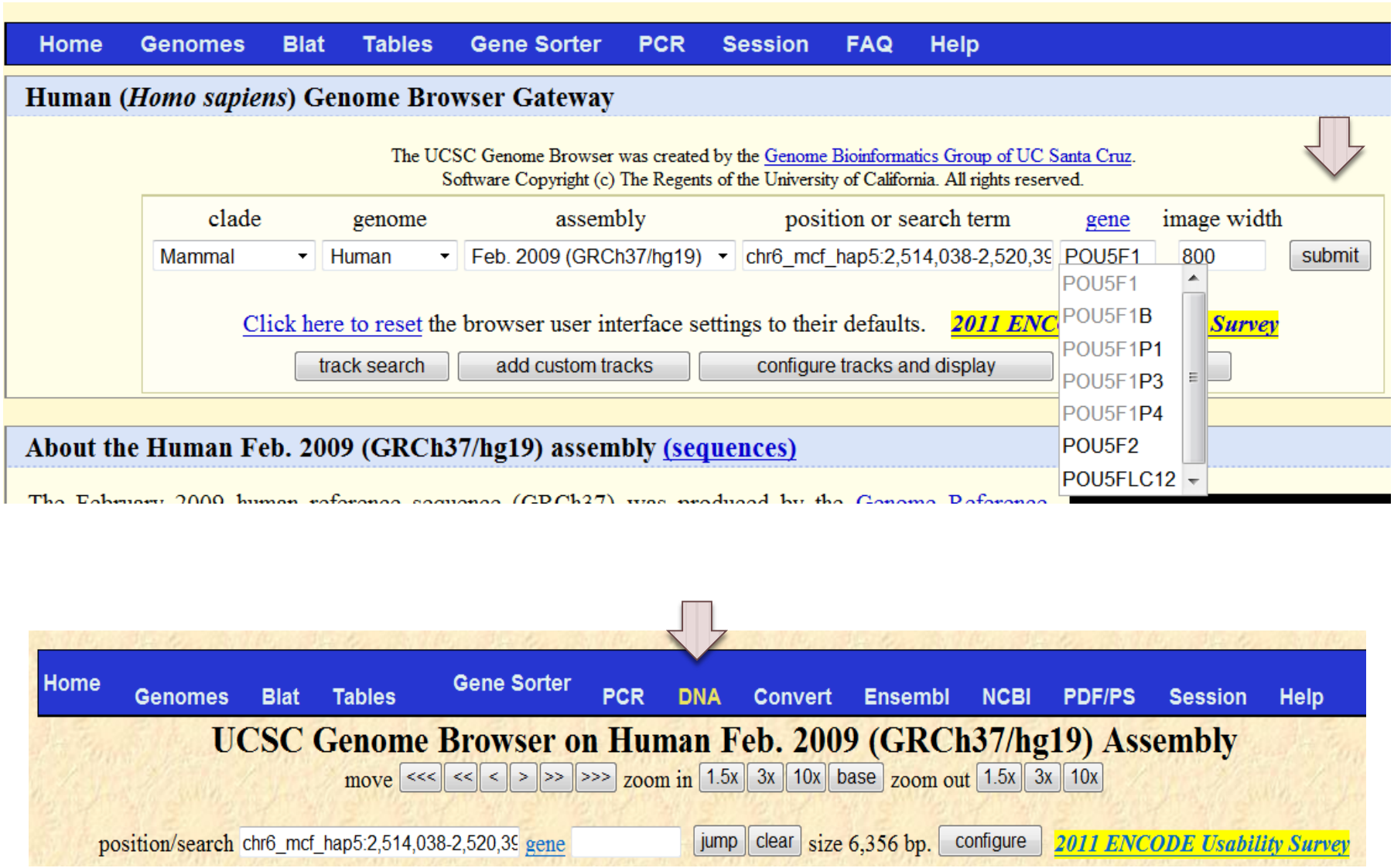
- Vertebrates (17)
  - Mammals (14)
    - Primates (3)

Map Viewer, NCBI

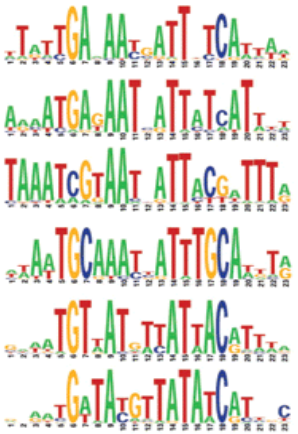
Genome Browser, UCSC

Ensembl Genome Browser, EBI

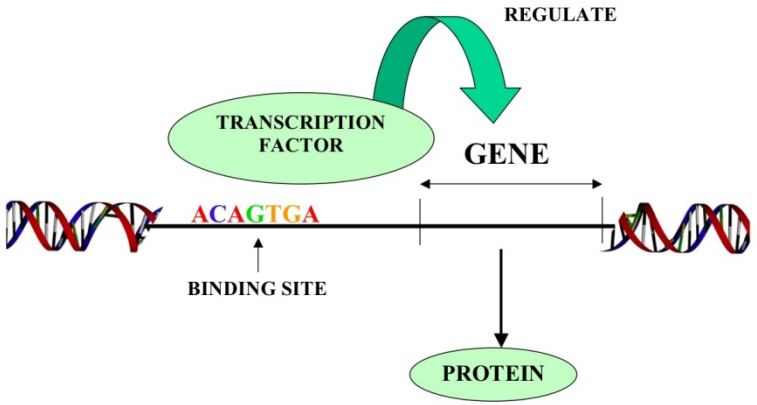


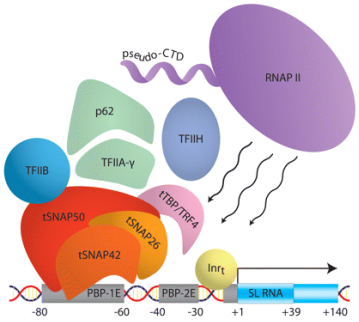

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help
Human (Homo sapiens) Genome Browser Gateway
The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.
clade genome assembly position or search term gene image width
Mammal Human Feb. 2009 (GRCh37/hg19) chr6_mcf_hap5:2,514,038-2,520,39 POU5F1 800 submit
POU5F1
POU5F1B
POU5F1P1
POU5F1P3
POU5F1P4
POU5F2
POU5FLC12
Click here to reset the browser user interface settings to their defaults.
2011 ENC Survey
track search
add custom tracks
configure tracks and display
About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr6_mcf_hap5:2,514,038-2,520,39 gene jump clear size 6,356 bp. configure
2011 ENCODE Usability Survey

Q: How to Identify Potential Regulators?



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

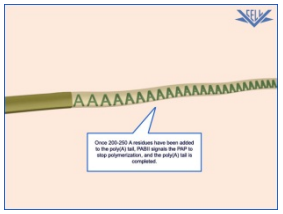


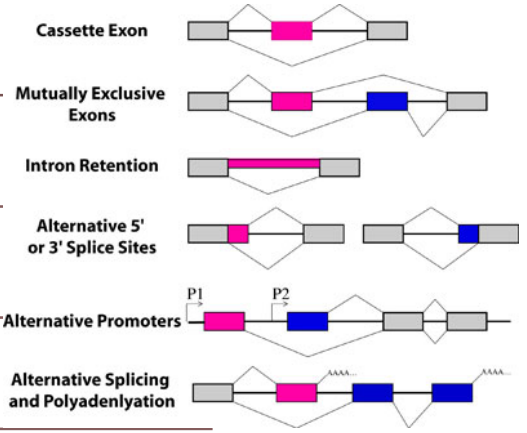


Feature-Based Methods

Based on identifying **gene signals**

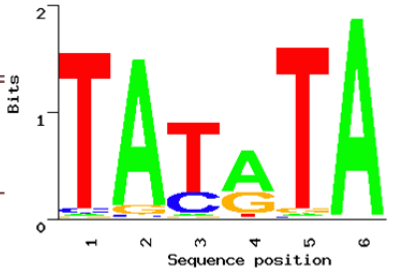
- **Promoter elements**
- Splice sites
- **Start/stop** codons
- PolyA sites…

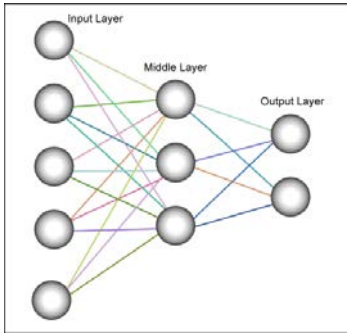


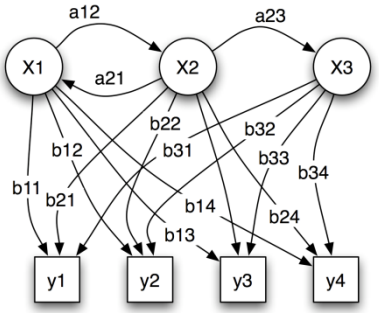


Wide range of **methods**

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



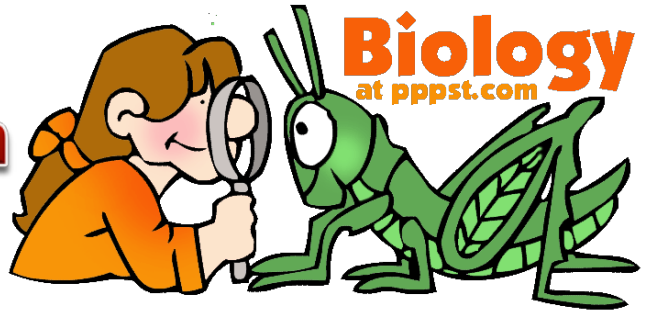




Promoter Databases and sites for analysis, prediction and search

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

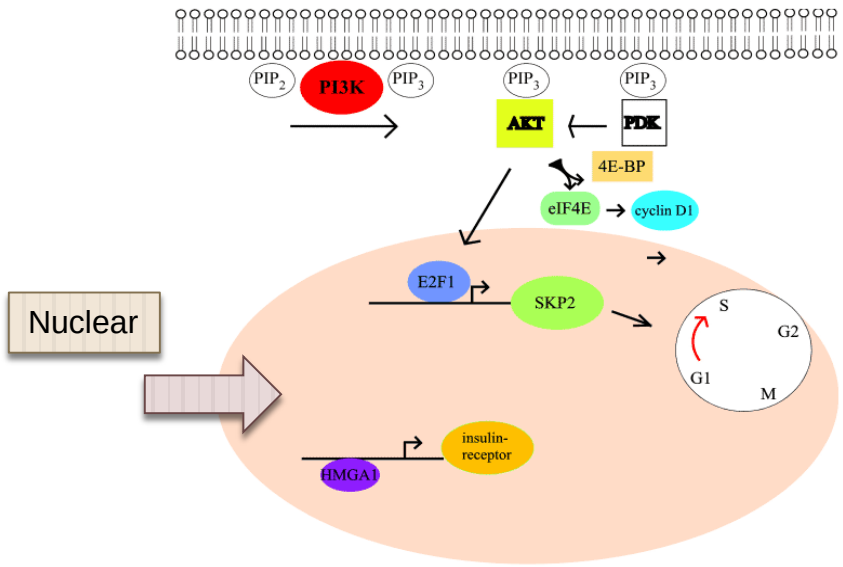
Pattern-driven



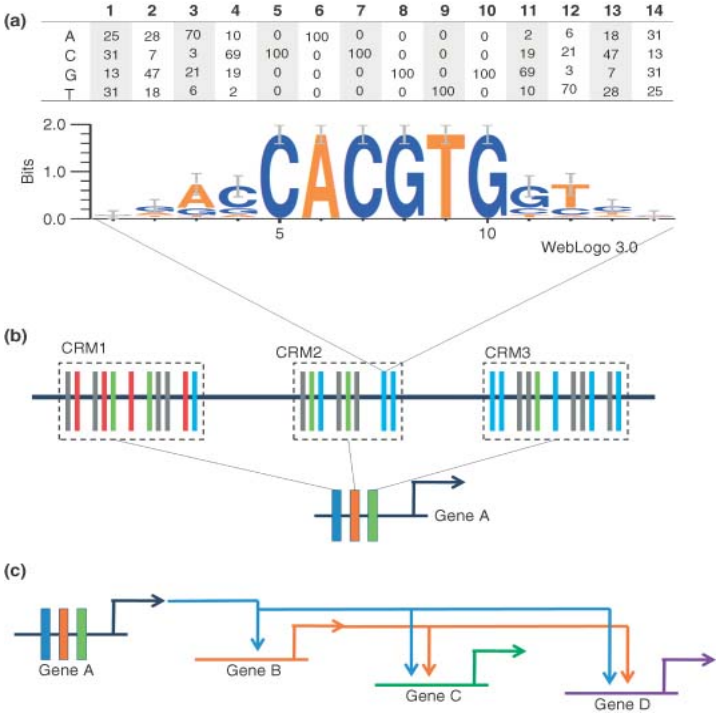
Success depends on **available of collections** of **annotated binding sites**

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

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



PI3K/AKT signaling in pancreatic cancer cells

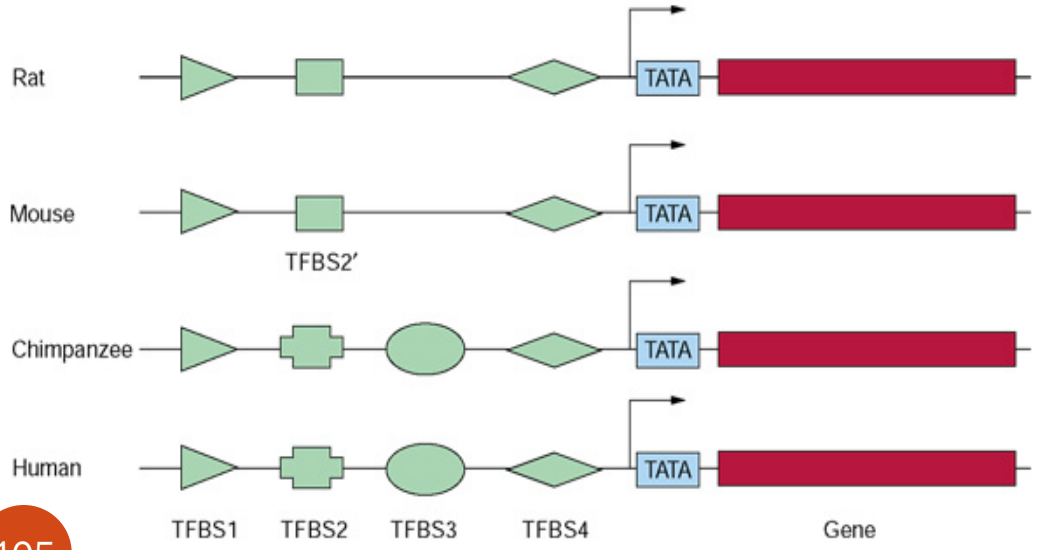


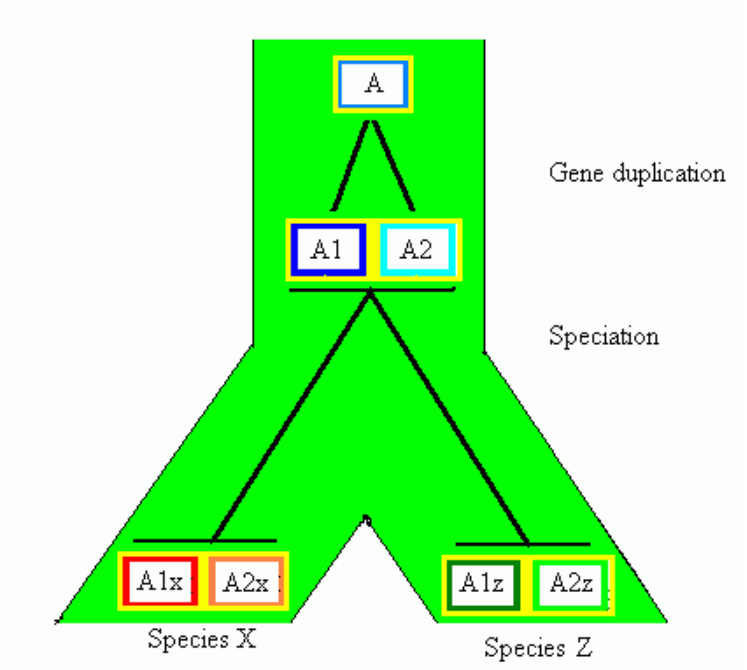
Taking **sequence context/biology** into account (Do the **wet lab** experiments!!!)		
Eukaryotes: clusters of TFBSs are common	**Probability** of "real" binding site increases if annotated **transcription start site (TSS) nearby** • But **NOT** for enhancers • Only **a small fraction of TSSs** have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

Phylogenetic Footprinting

Patterns of gene regulation are often conserved across species

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





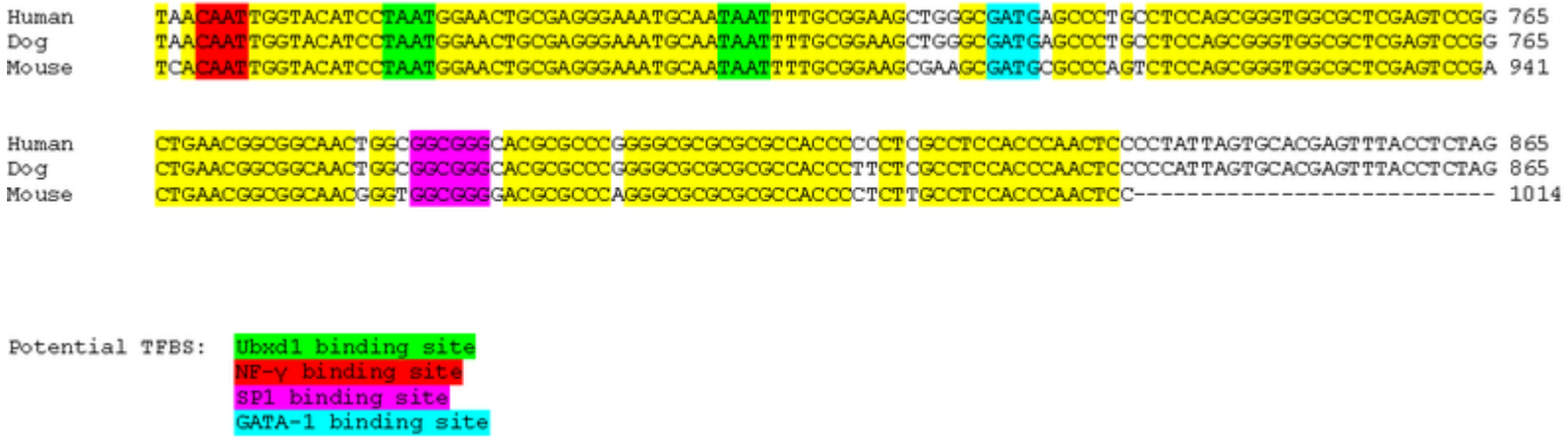
To select gene of interest

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

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

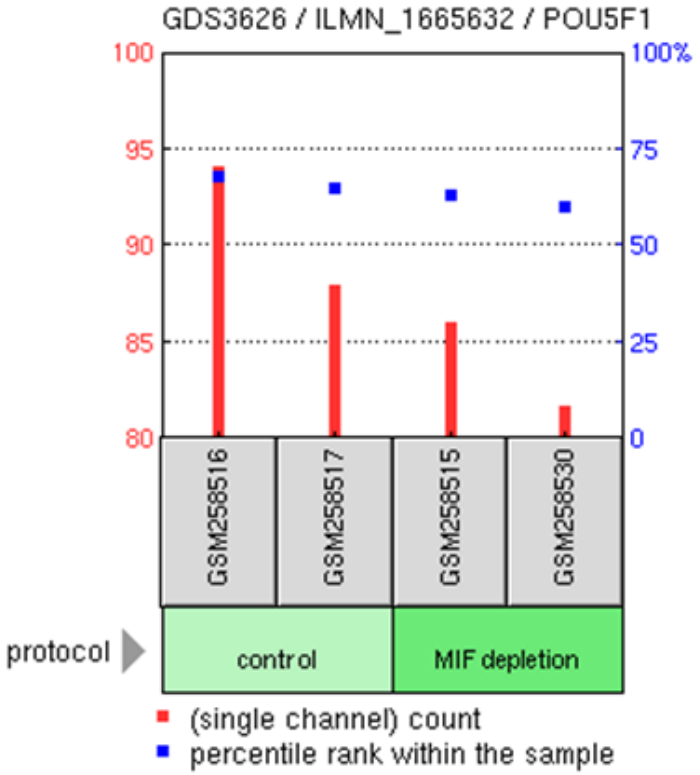
Align sequences by using **any** basic computer software (e.g., clustalW)

Visually look for **identical motif**

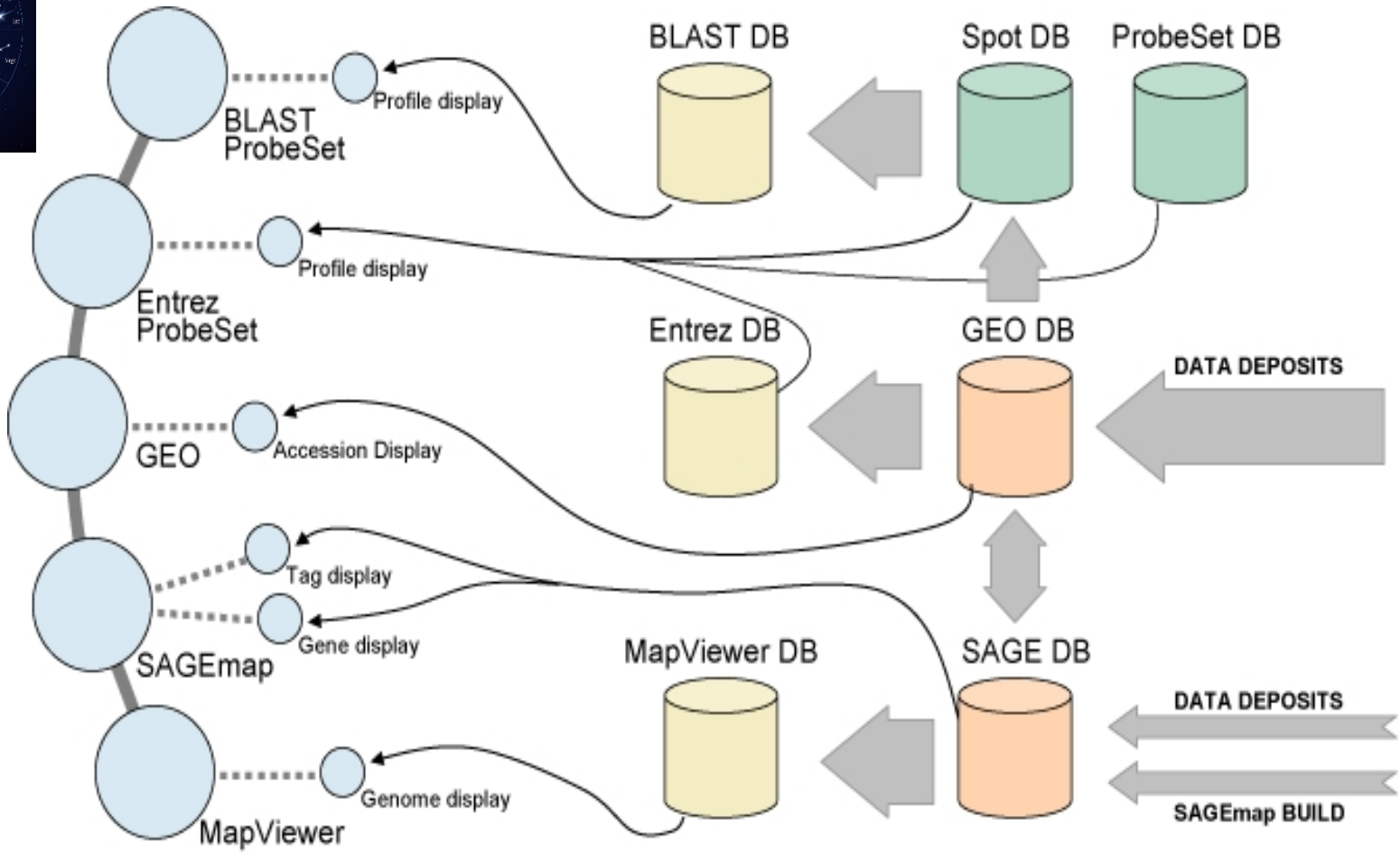


One More Trick - Coregulation

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



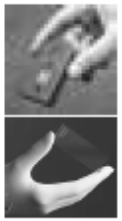
Constellation of NCBI Gene Expression Resources



Gene Expression Omnibus (GEO) (1)

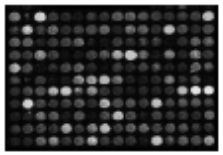
*Submitted by Manufacturer**

GPL
Platform descriptions



Submitted by Experimentalists

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



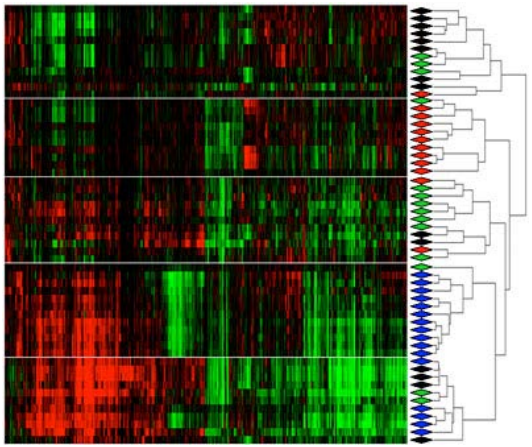
GSE
Grouping of slide/chip data
"a single experiment"

Entrez GEO

Curated by NCBI

GDS
Grouping of experiments

Entrez GEO Datasets



Gene Expression Omnibus (GEO) (2)

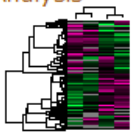
- Search GEO Profiles: POU5F1
  - Or **Limit**, **Preview / Index**
- GDS vs. GSE

Search for GDS46[ACCN] Search Clear Show All **Advanced Search**

DataSet Record GDS46: Expression Profiles | Data Analysis Tools | Sample Subsets

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

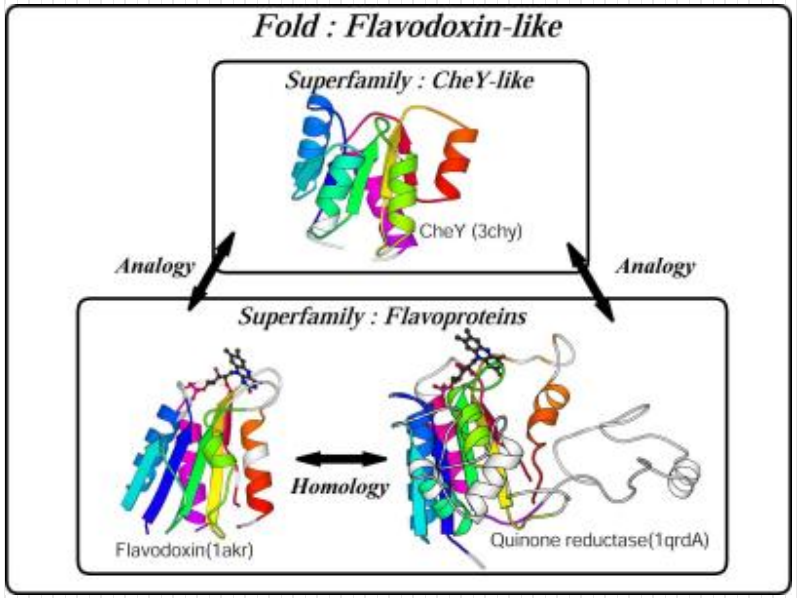
Cluster Analysis



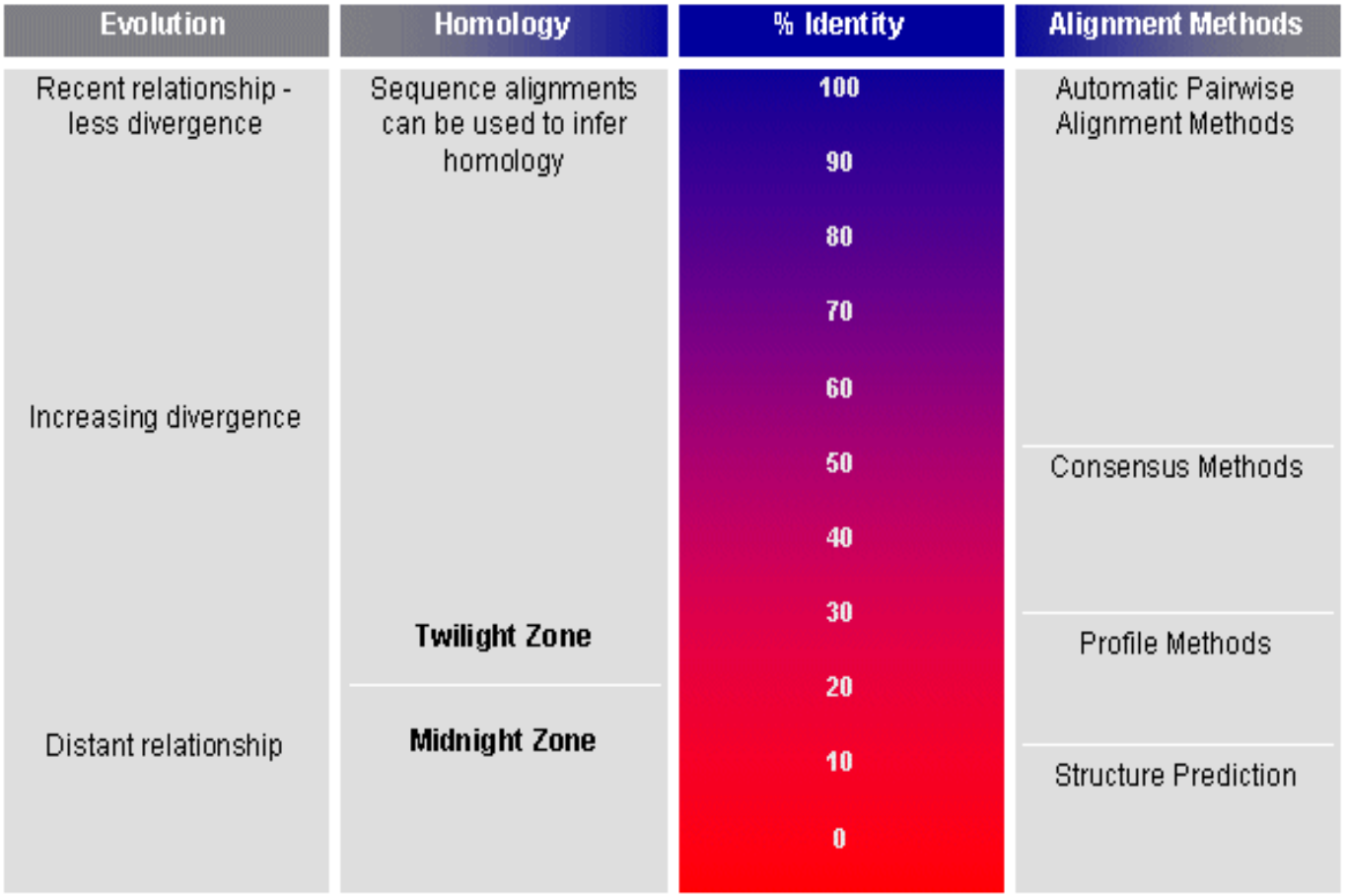
Download

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

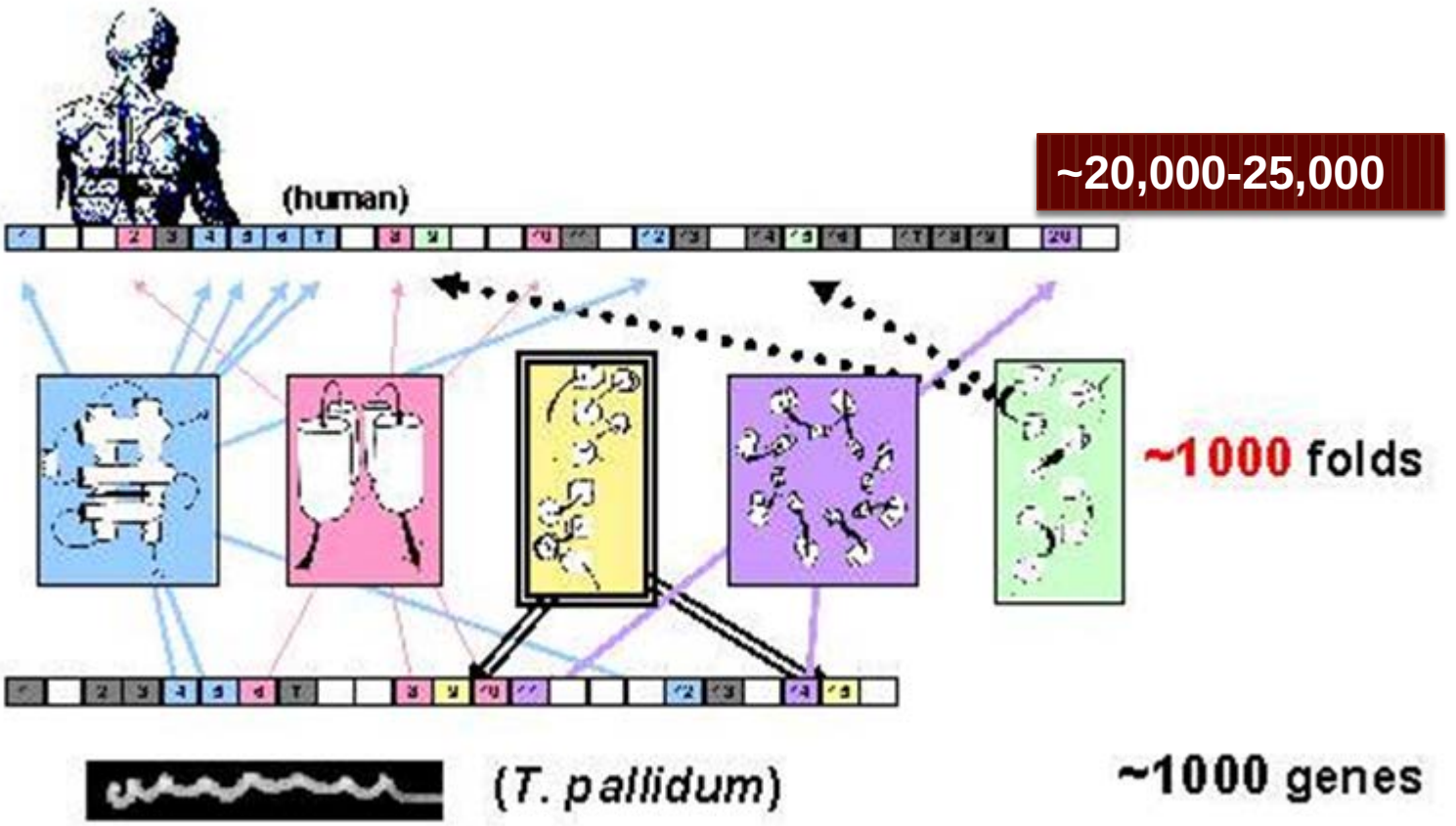
Q: Can You Speculate the Function of YFG from Structure Similarity?



Structures are More Conserved Than Sequences



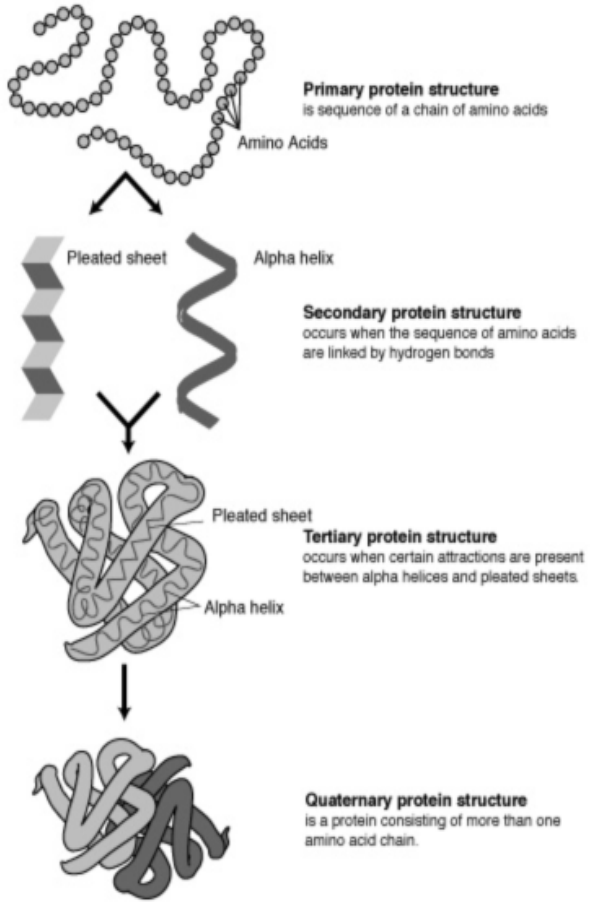
Simplifying Genomes with Folds, Pathways

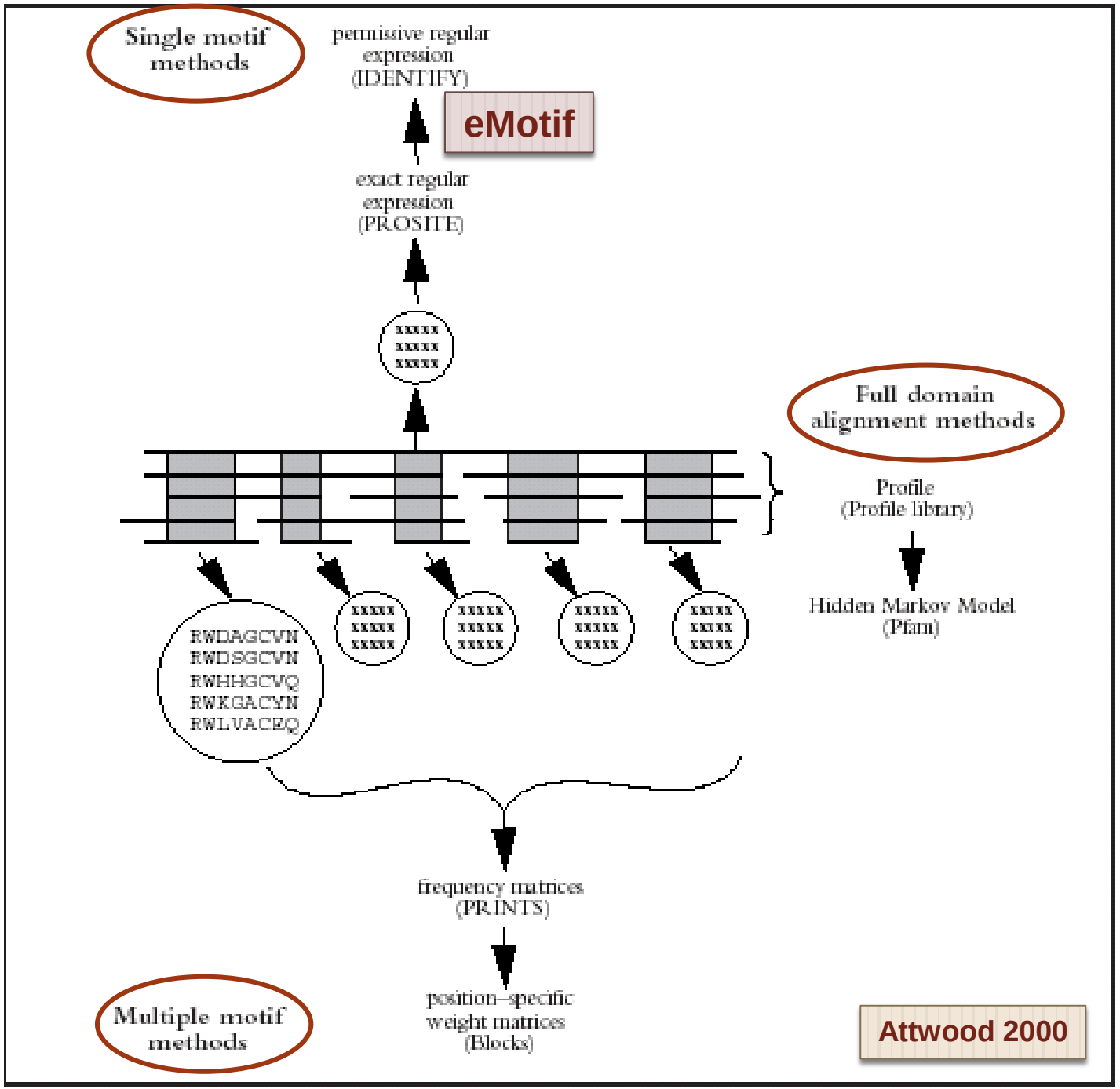


Significance: fold # << sequence ##

Levels of Protein Sequence & Structure Organization

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




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

Major Secondary "Pattern" Database

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

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

NCBI > Structure Home > Conserved Domains

NCBI

Conserved Domains

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

Search Conserved Domains for GO CLEAR Help

Conserved Domains and Protein Classification

RESOURCES | SEARCH | HOW

Resources

Conserved Domain Database (CDD)

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

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

CD-Search & Batch CD-Search

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

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

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

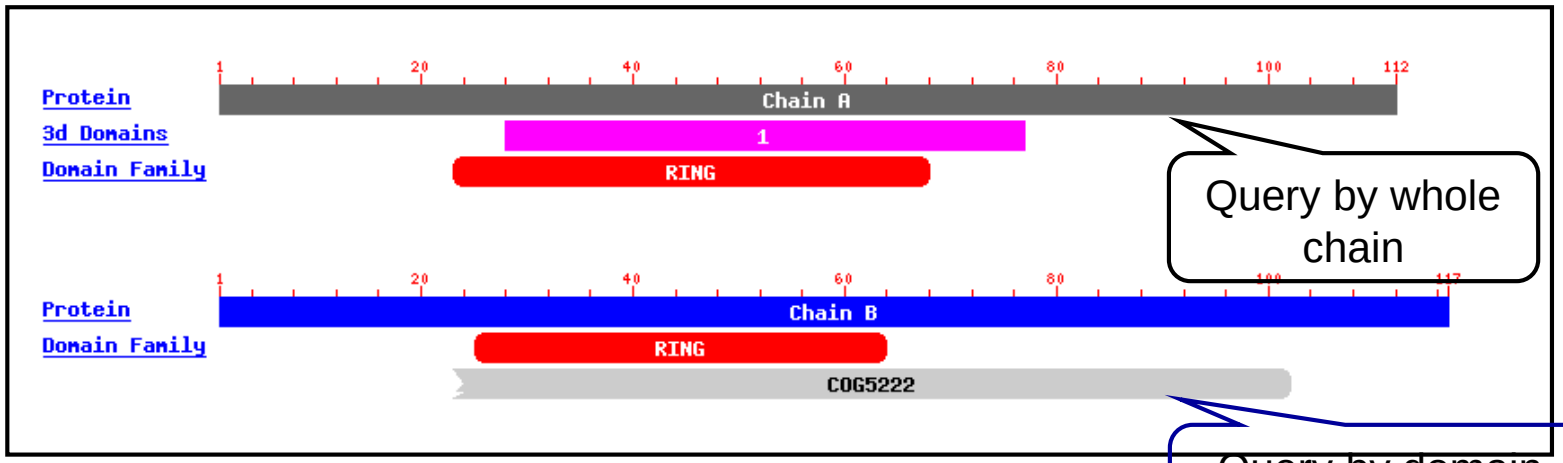
- CDD v2.28 - 39357 PSSMs
- SMART v5.1 - 791 PSSMs
- Pfam v24.0 - 11912 PSSMs
- COG v1.00 - 4873 PSSMs
- KOG v1.00 - 4825 PSSMs
- PRK v6.00 - 10885 PSSMs
- TIGR v10.00 - 4023 PSSMs

CDART: Domain Architectures

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

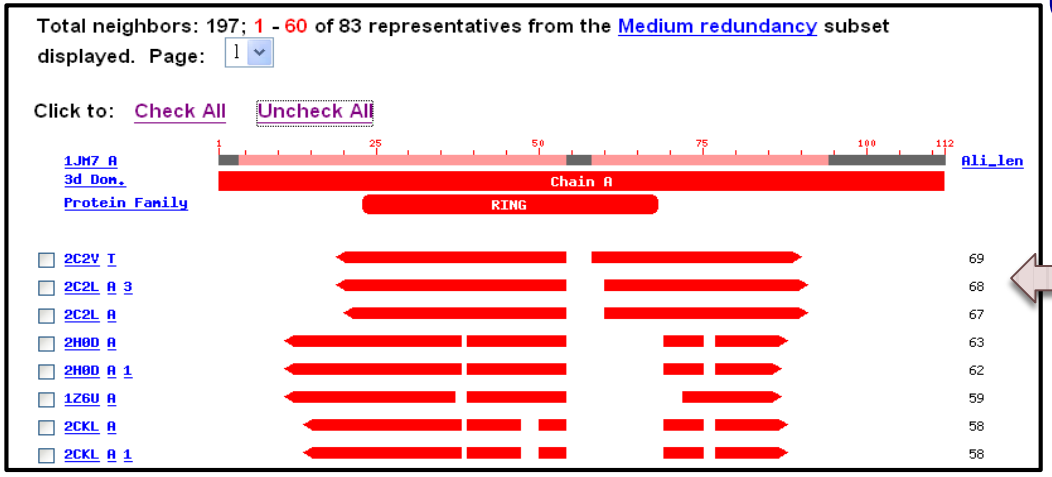
Highlig

VAST: Query by Chain or 3D Domain



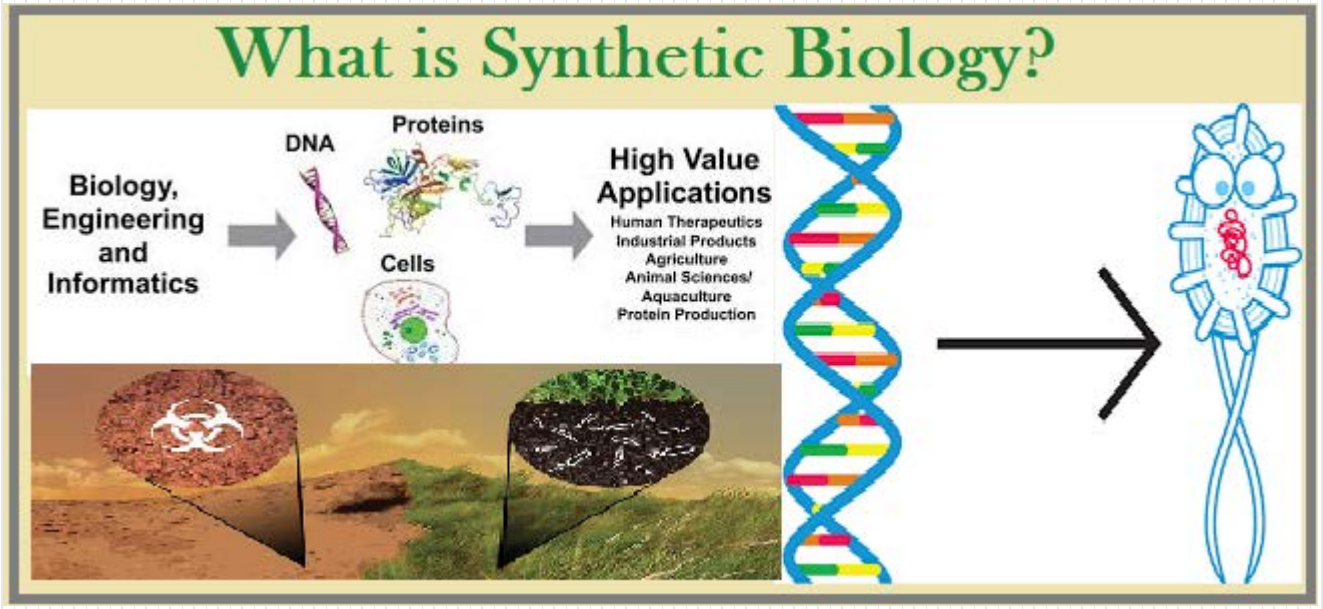
Query by whole chain

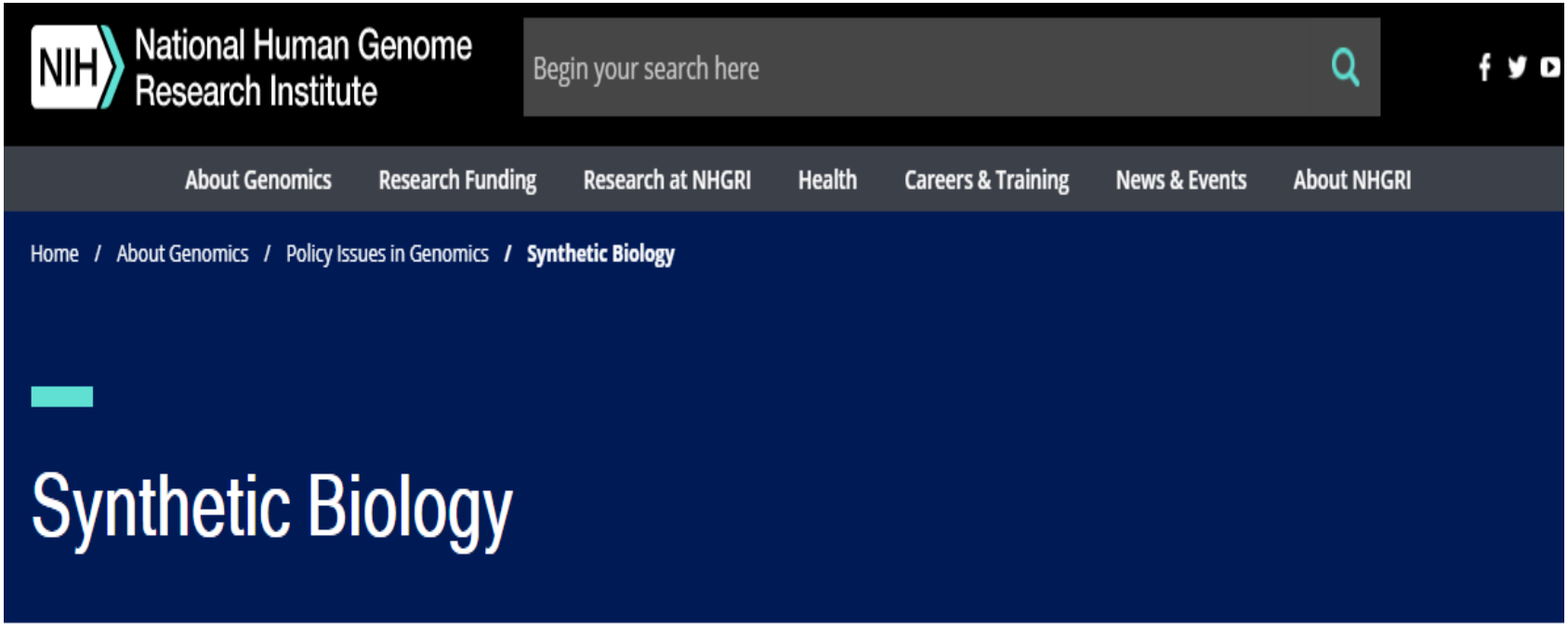
Query by domain COG5222



Not found with chain query

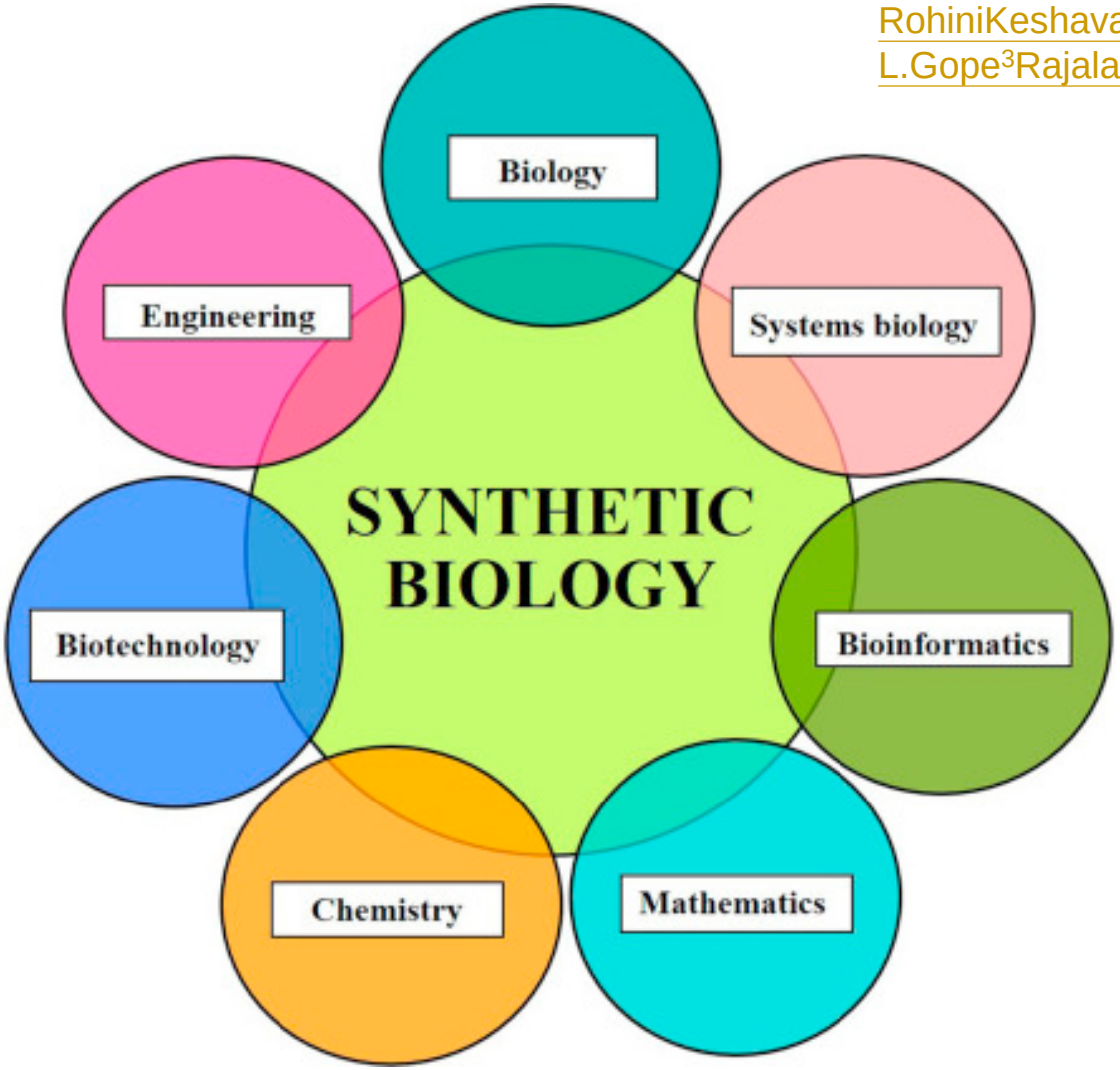
Synthetic Biology





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.

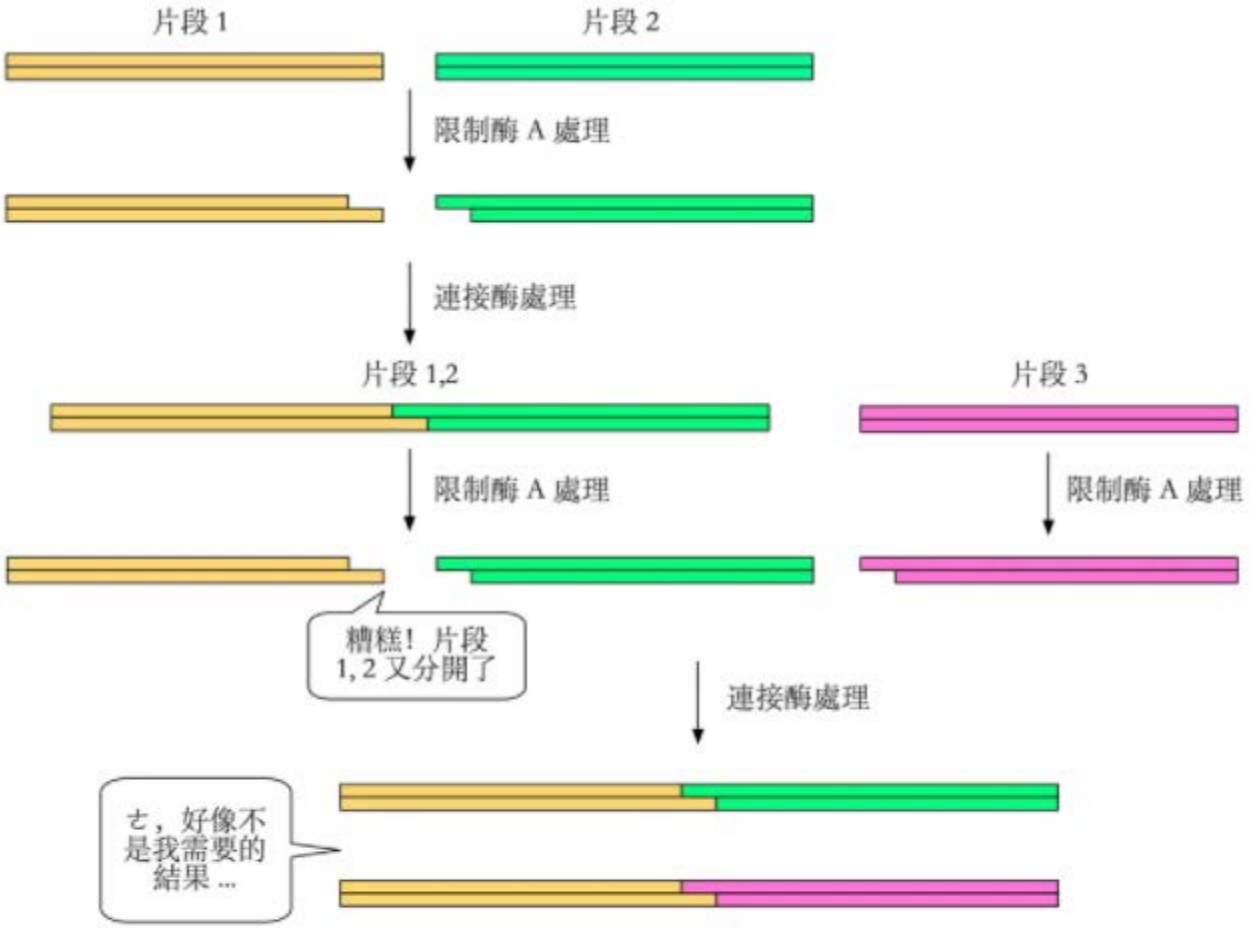
Chapter 4 - Synthetic Biology: Overview and Applications Author links open overlay panel RohiniKeshava[1]RohanMitra[2]Mohan L.Gope[3]RajalakshmiGope[2]



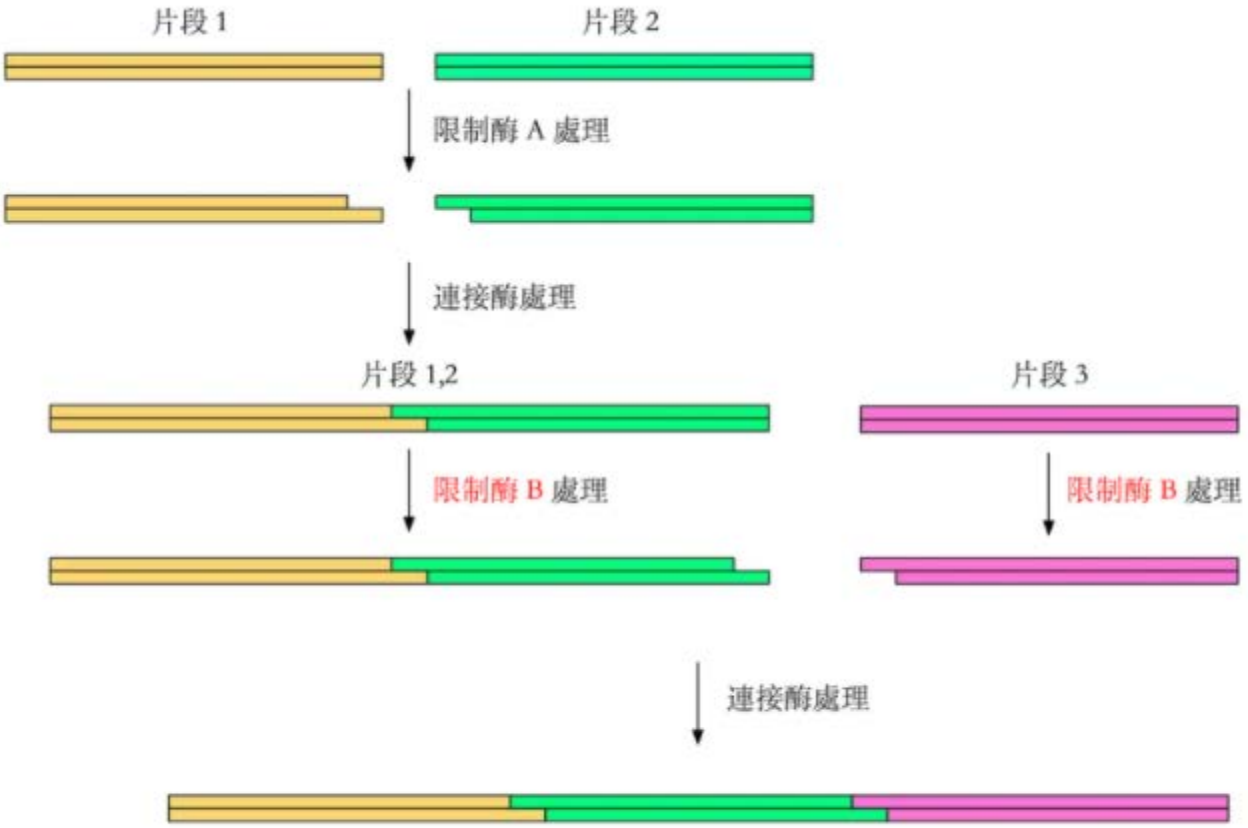
Omics Technologies and Bio-Engineering Towards Improving Quality of Life
2018, Pages 63-93

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。


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



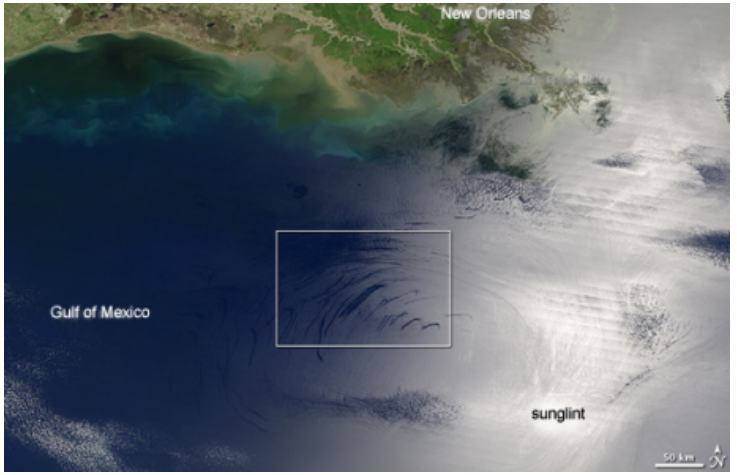
圖二 成功的三段組裝

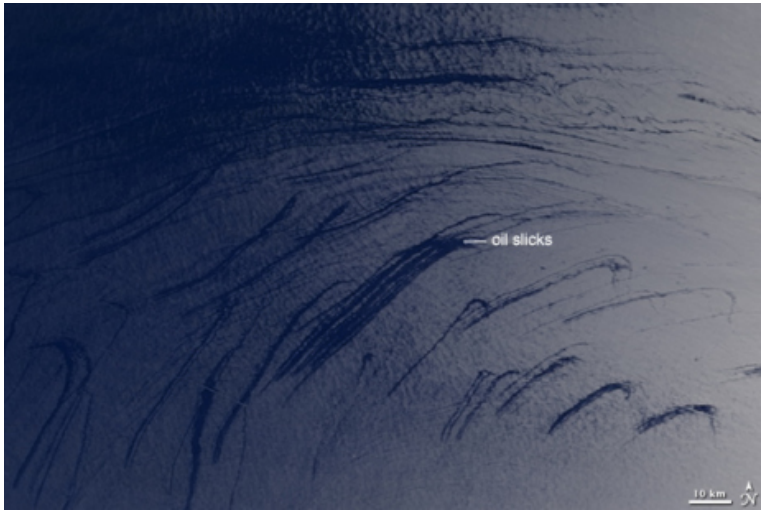
Example 2

- 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)

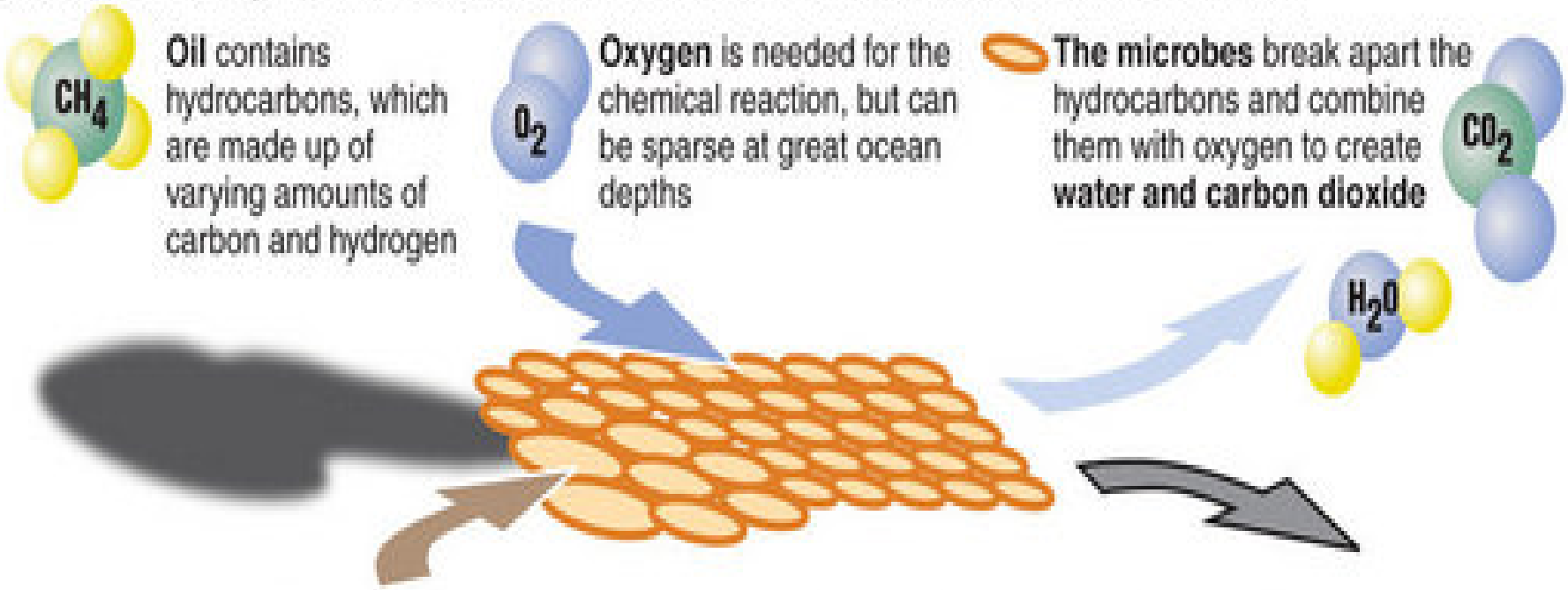




Petroleum-degrading microbes called *Oceanospirillales*

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.



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

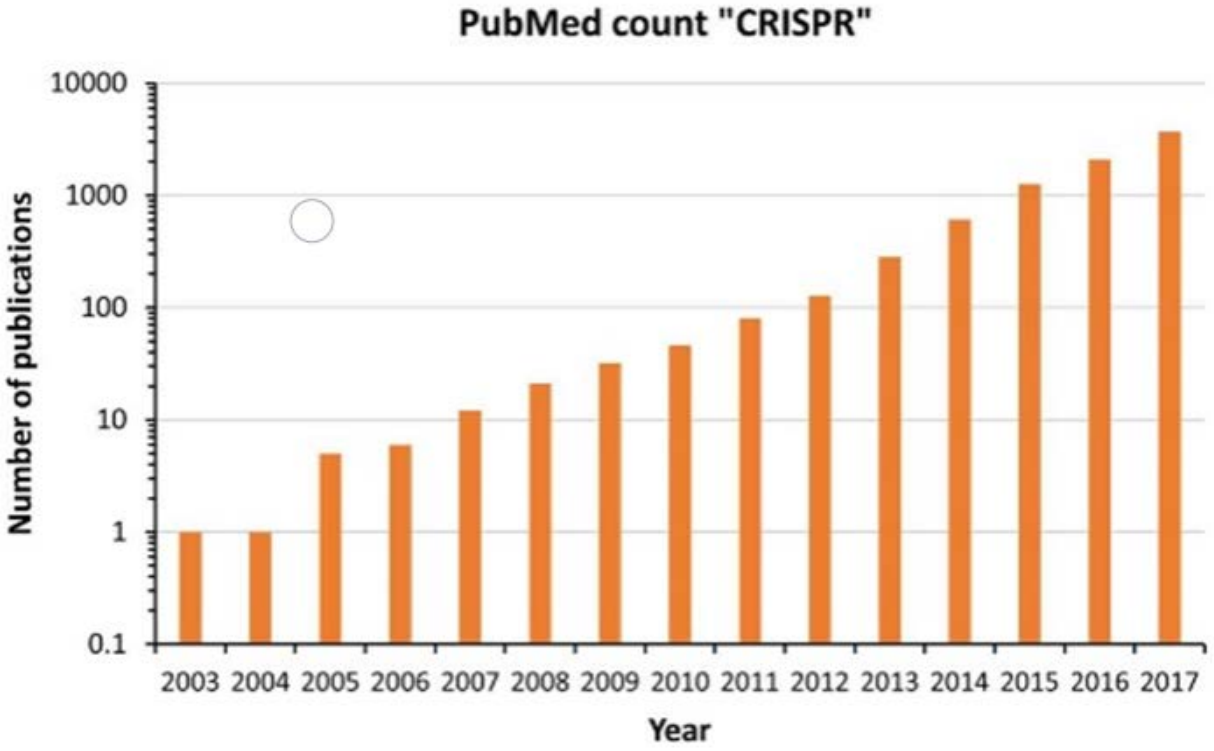
What is the difference between synthetic biology and genome editing?

- 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]

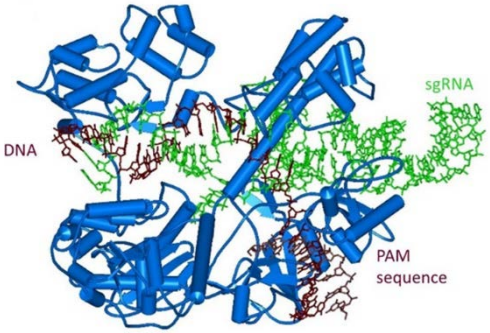
Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)

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

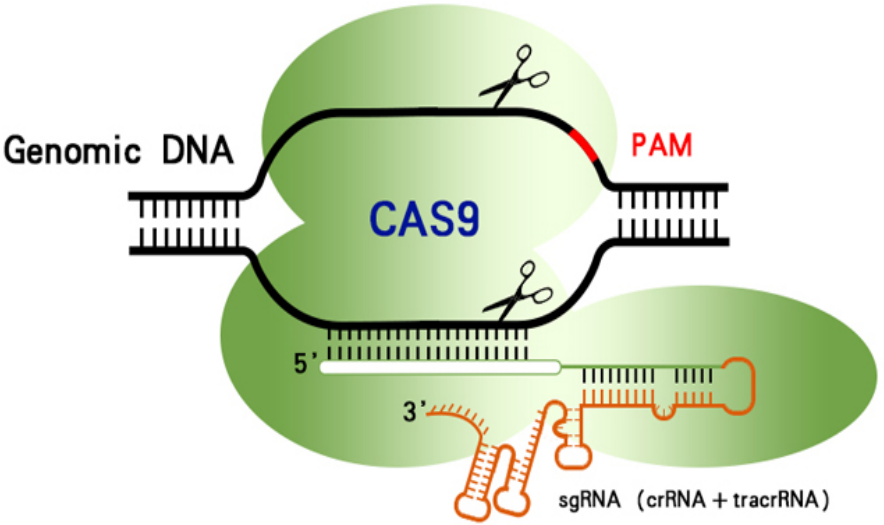
- Applications
  1. To make a gene nonfunctional (knockout)
  2. Replace on version of a gene with another
     - E.g., gene therapy
     - David Vetter was born without a functioning immune system and spent his life in a bubble that protected him from germs. He died at age 12 in 1984. Scientists are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



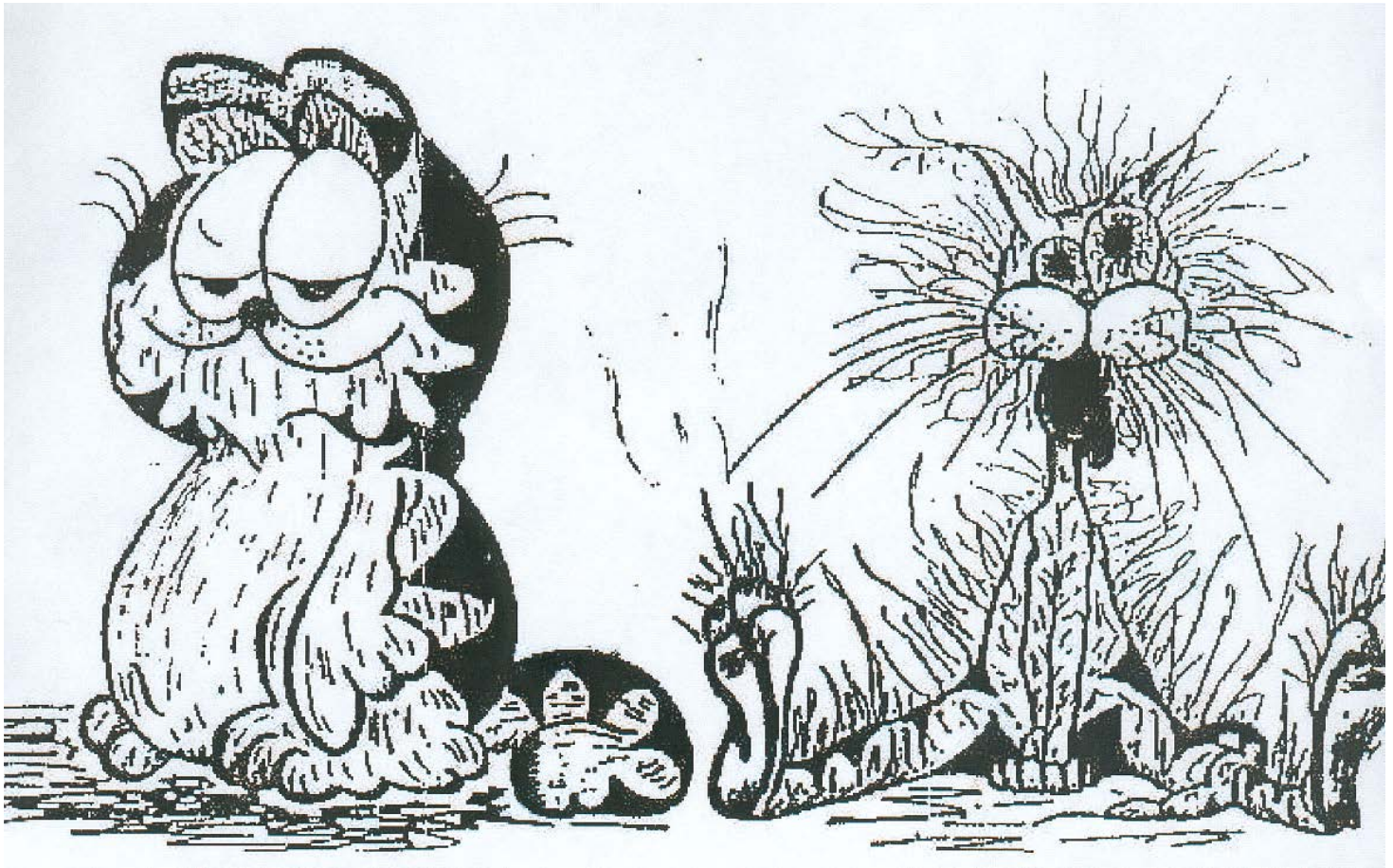
- Non-coding RNAs & Cas protein
- Protospacer adjacent motif (PAM) is a 2-6 base pair DNA sequence immediately following the DNA sequence targeted by the Cas9 nuclease in the CRISPR bacterial adaptive immune system
- sgRNA = single guide RNA = a targeting sequence (crRNA sequence) + (a Cas9 nuclease-recruiting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

CRISPR: Gene editing and beyond

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

Genetic Engineering Will Change Everything Forever – CRISPR

Before…



After…