

IBT 432 Aplikasi Bioinformatika

Database sekuen dan analisis genomika

Riza Arief Putranto

Rencana Perkuliahan

- ~~1. Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. Database sekuen dan analisis genomika
3. Anotasi sekuen ke genom - Praktik
4. Analisis komparasi genomika I
5. Analisis komparasi genomika II
6. Analisis komparasi genomika III
7. Analisis komparasi genomika – Praktik
8. Protein modelling I
9. Protein modelling II
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

A. Genomic analysis

BLAST

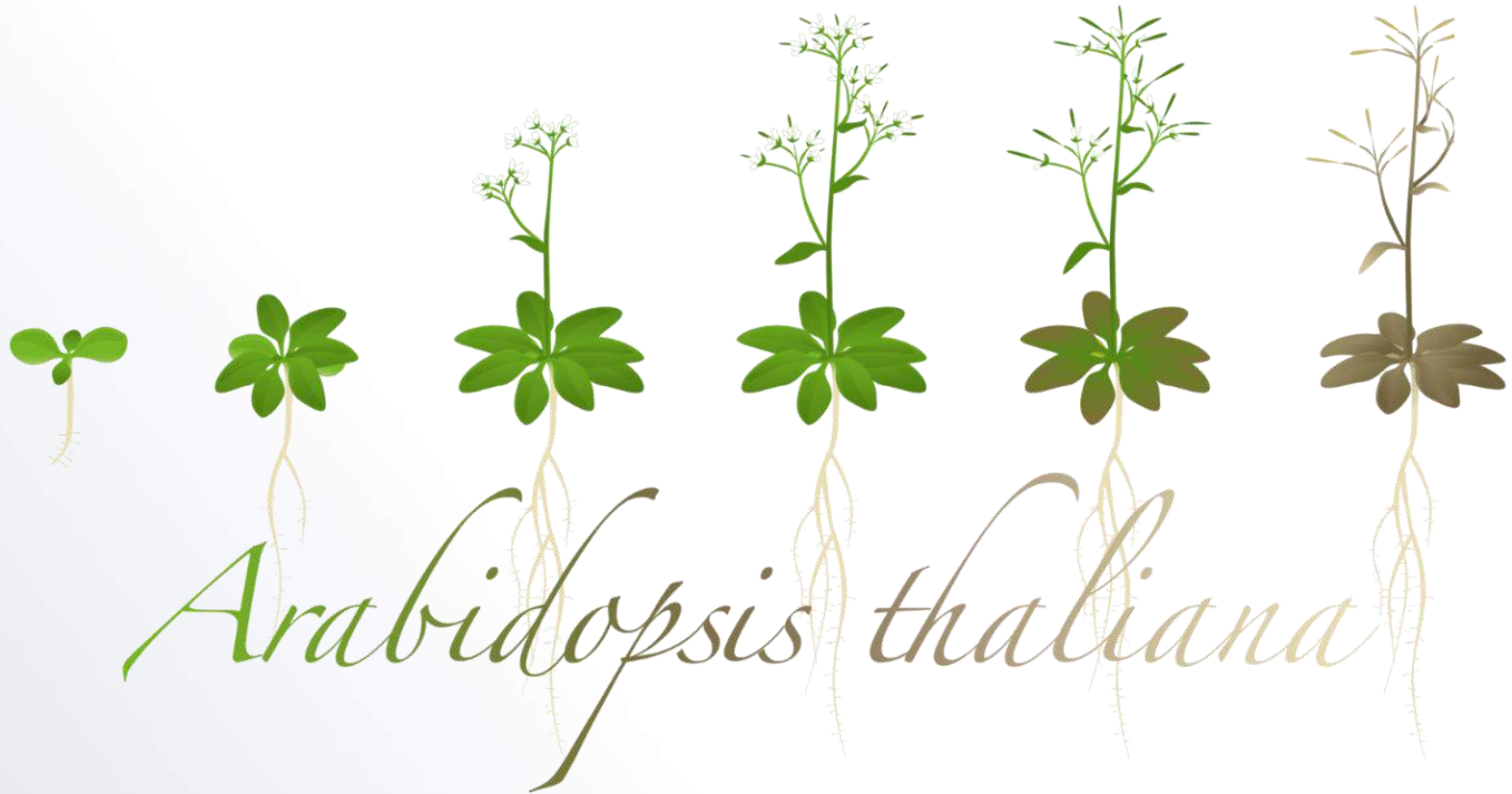
Multiple Sequence Alignment

ATGGAATTACAGAGAGGATTCCTCTGCTATAACCAGCAATCTAAAGCTCTATTGAAGAAGAACCTGTCGCTTTCATGGAGGAACAA
GAGCGCCTCATTCTCCAGTTGTTTTATCCTTGTTCTTCATTTTCTCCTCTTCTGCATTGAAAAAGCCATGAACGCTGCCAACTCC
GACACCACGGCGTATAAGTCGGTACTCGATCCTCAGCCTCTGGTTTCGCCGCCGATCCCTCCCTGCGAGGACAAGTTCTACATTAA
GAAGCCTTGCTTCGACTTTGTATGGAGCGGAAACTATAGCGATAGACTGAATCGTATTGTCAGTTCAATCATGGCAAATAATCCTG
GCAGGGTGATTACGTCCAATAAGGTTAAATCATTCAGAACAAACAGACGATGTAGACGTTTGGCTTTTGAATAATCCTATGCATTGC
CCTGGAGCTCTGCATTTCAAAGATATAAATGCCACTGTTATCAGCTATGGCGTACAGACTAATTCTACCCGAGTTATGAATCGGGGG
TATTCTGAAGATCCTACATTCAAATTTCAAATCCCACTTCAGATTGCAGCAGAGCGTGAAATTGCTCGGTCTATCATTGGAGATCCA
AACTTTAGCTGGGTAGTTGGACTTAAAGAATTTGCACACCCAGCAAAAAACAATTTCTCTGTGTTGGCTTCAATTGGACCAACTTT
CTTTCTTGCGTTTGGCATGTTGGGTTTGTGATGCAAATTGGATCTTTGGTGGTAGAGAAGGAACTCAAACCTTCGCCAGGCAATG
AGTATGACAGGTCTTATGAATCTGCTTATTGGTTCTCTTGATCACATGGGAAGGAATACTTCACTCGTATCATCACTTCTCCTTG
TTCTTTTTGGAATGATCTTTGACTTTGAGATCTTCAAGAAAAGAAATTTTCACTTCTCTTTCTTCTCTTCTATCTTTTTTCAGCTCAA
TATGGTTGGTTTTGCC
TTTTTTACACAGATTG
AATCTTCTTGGTATAG
GTGTGTACGAGATGA
TTTGTTTTGGCAATCT
GACTGGAAAAGGTG
AGAAGATGAGGATG
AGGTACGTGGACTTGCAAAGGTATATGCTGGGACTACGAAGATTGGTTGTTGTAATGCAAGAAAACCTCACCTTACCATGCTCTC
AAGGGCTTATGGATGAACTTTGCAAAGGATCAGTTATTTTGTCTCCTTGGACCAAATGGCGCTGGAAAAACTACCGCAATCAATT
GTTTGACAGGCTTAACACCTGTGACCAGTGGAGATGCTTTGATTTATGGATATTCCATTCGGAGCCCTGTTGGCATGTCCAACATTC
GAAGAATCATAGGAGTTTGTCCCCAGTTTGCATCCTTTGGGATGCATTATCTGGTGCAGAGCATCTCCATCTCTTTGCTAGCATT
AAGGCCTACCCCCAGATTCAATAAATTTGGTTGCTGAGGAATCATTAGCAGAGGTAAGACTCACTGAGGCAGCTAAAGTGAGAA
CCAGGAGTTACAGTGGAGGAATGAGACGCCGGCTCAGTGTTGCAATAGCACTTATTGGGAACCCAAAGTTGGTCATTCTAGACG
AACCGACTACTGGTATGGATCCAATATCGAGAAGACATGTCTGGGATATAATACAGAATGCAAAGAAAGGTCGTTCCATTGTCCTG
ACAACACATTCAATGGAAGAAGCTGACATTCTAAGTGATCGCATAGGAATTATGGCCAAGGGTAGGCTCCGATGCATCGGAACAT
CAATCAGTTGAAGTCGAGATTCGGTACCGTTTTCACTAATGTGAGCTTTATTGAAAGCAATGCTATGAGACGCCGGCTCAGT
GTTGCAATAGCACTTATTGGGAACCCAAAGTTGGTCATTCTAGACGAACCGACTACTGGTATGGATCCAATATCGAGAAGACATGT
CTGGGAATGAGACGCCGGCTCAGTGTTGCAATAGCACTTATTGGGAACCCAAAGTTGGTCATTCTAGACGAACCGACTACTGGTA
TGGATCCAATATCGAGAAGACATGTCTGGGATCCAATATCGAGAAGACATGTCTGGGATCCAATATCGAGAAGACATGTCTGGGA

A nucleotide sequence...

What it means?

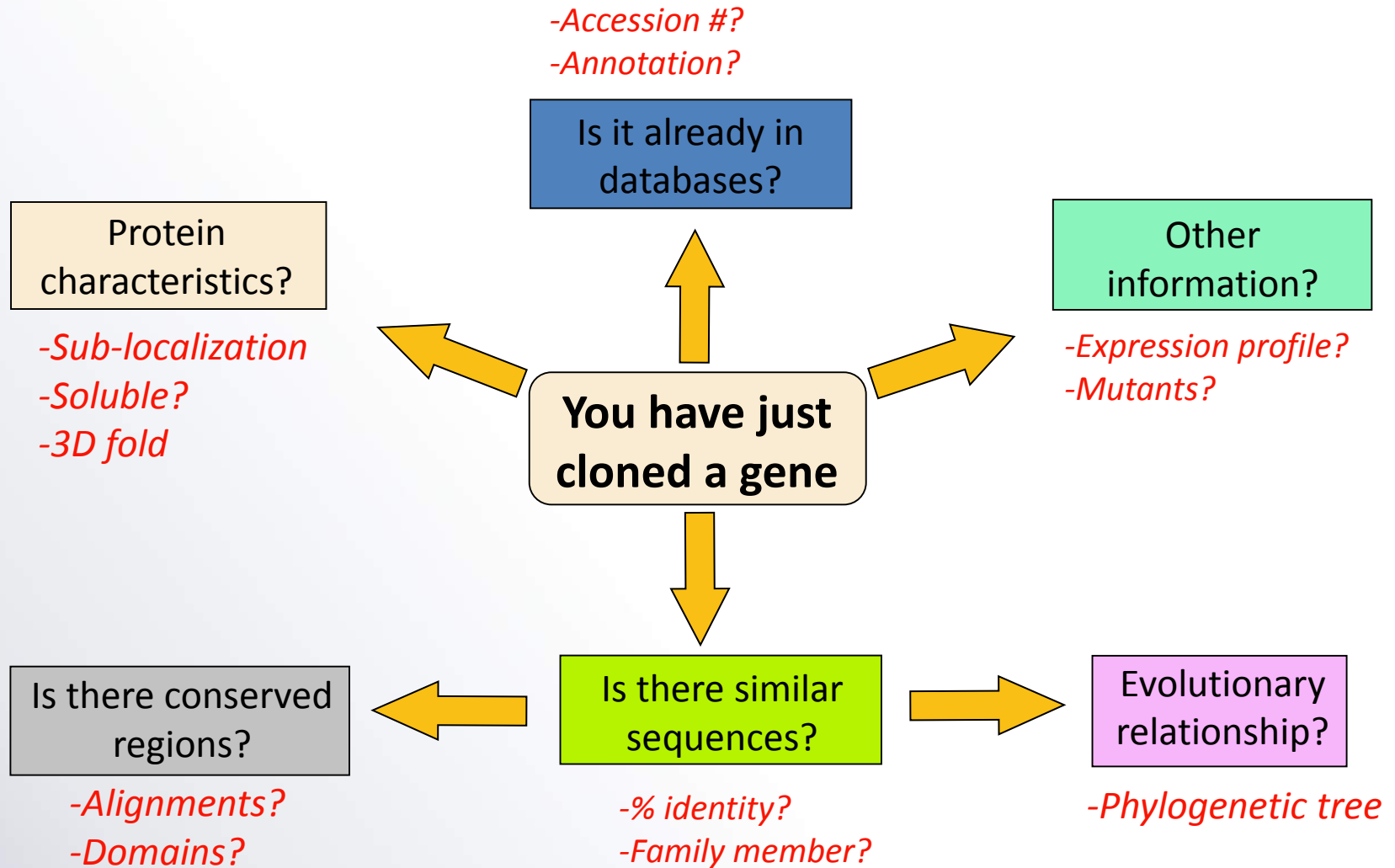
Do you feel dizzy?



Arabidopsis thaliana

A phenotype...
A gene that controls development
Do you still feel dizzy?

Analysis of genomics data



The steps of genomic data analysis

-Accession #?

-Annotation?

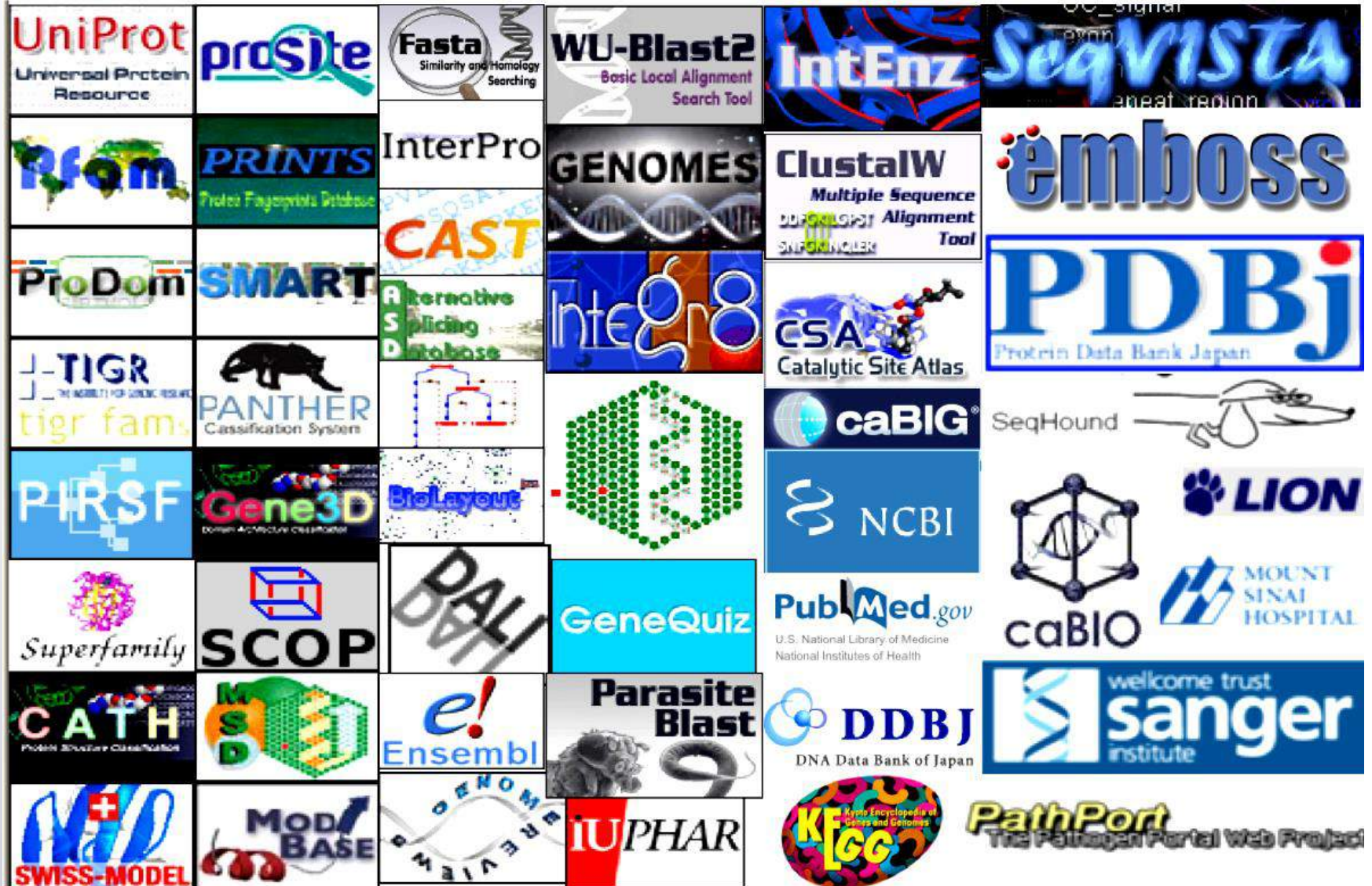
Is it already in
databases?



**You have just
cloned a gene**

How many Bioinformatics Resources?

The 2016 issue has a list of about 180 **databases**



DNA (nucleotide sequences) databases

- They are big databases and searching either one should produce similar results because they exchange information routinely.



GenBank (NCBI): <http://www.ncbi.nlm.nih.gov>



DDBJ (DNA DataBase of Japan): <http://www.ddbj.nig.ac.jp>



TIGR: <http://tigr.org/tdb/tgi>



Yeast: <http://yeastgenome.org>



Microbes: <https://img.jgi.doe.gov/>

- **Specialized databases: tissues, species...**

ESTs (Expressed Sequence Tags)

~at NCBI <http://www.ncbi.nlm.nih.gov/dbEST>

~at TIGR <http://tigr.org/tdb/tgi>

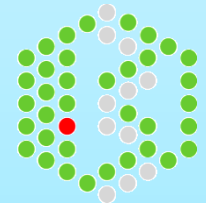
...many more

INSDC: International Nucleotide Sequence Database Collaboration

<http://www.ddbj.nig.ac.jp/>



INSDC
IAC
ICM



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

<http://www.ebi.ac.uk/ena>

Protein (amino acid) databases

➤ They are big databases too:



Swiss-Prot (very high level of annotation)

https://web.expasy.org/docs/swiss-prot_guideline.html



UniProt (protein identification resource) the world's most comprehensive catalog of information on proteins

<http://www.uniprot.org/>

➤ Translated databases:



TREMBL (translated EMBL): includes entries that have not been annotated yet into Swiss-Prot.

<http://www.ebi.ac.uk/trembl/access.html>



GenPept (translation of coding regions in GenBank)



PDB (sequences derived from the 3D structure Brookhaven PDB) <http://www.rcsb.org/pdb/>

Analysis of genomics data

-Accession #?

-Annotation?

Is it already in
databases?



**You have just
cloned a gene**



Is there similar
sequences?

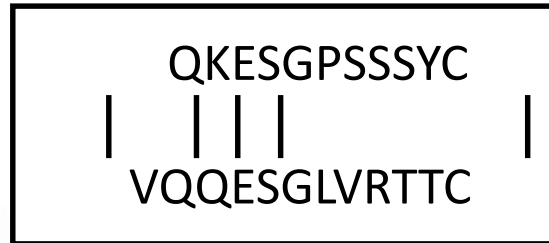
-% identity?

-Family member?

Database search method: Alignment

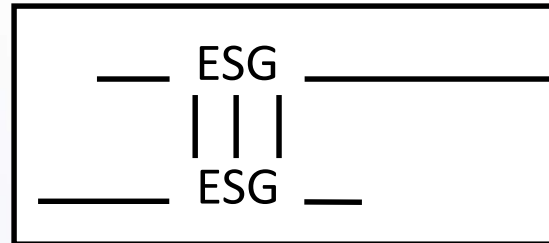
➤ Two broad classes of sequence alignments

Global alignment:



each residue, not sensitive

Local alignment:



group of region, faster

➤ The most widely used **local similarity algorithm**

- ✓ Smith-Waterman <http://www.ebi.ac.uk/MPsrch/>
- ✓ **Basic Local Alignment Search Tool (BLAST)**
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- ✓ Fast Alignment (FASTA) <http://fasta.genome.jp>; <http://www.ebi.ac.uk/fasta33/>;
<http://www.arabidopsis.org/cgi-bin/fasta/nph-TAIRfasta.pl>

Basic Local Alignment Search Tool (BLAST)

- Tool to **compare homologous sequences between species**
- **e-values** - **the probability of finding a random sequence** in the database. The lower e-values the more trustable (statistically) the result

The screenshot shows the BLAST website interface. At the top, there are logos for NIH (U.S. National Library of Medicine) and NCBI (National Center for Biotechnology Information), along with a "Sign in to NCBI" link. The main header includes the "BLAST" logo and navigation links for "Home", "Recent Results", "Saved Strategies", and "Help".

The main content area features a section titled "Basic Local Alignment Search Tool" with a brief description: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." A "Learn more" link is provided.

To the right of this section is a "NEWS" alert box titled "IgBLAST 1.8.0 released" with the text: "A new version of IgBLAST is now available. Wed, 15 Nov 2017 16:00:00 EST" and a link to "More BLAST news...".

Below the main section is a "Web BLAST" section with three options:

- Nucleotide BLAST**: nucleotide ► nucleotide (represented by a DNA double helix icon)
- blastx**: translated nucleotide ► protein (represented by a blue arrow pointing right)
- tblastn**: protein ► translated nucleotide (represented by a blue arrow pointing left)
- Protein BLAST**: protein ► protein (represented by a protein ribbon structure icon)

Analysis of genomics data

-Accession #?

-Annotation?

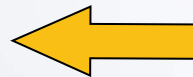
Is it already in
databases?



**You have just
cloned a gene**



Is there similar
sequences?



Is there conserved
regions?

-Alignments?

-Domains?

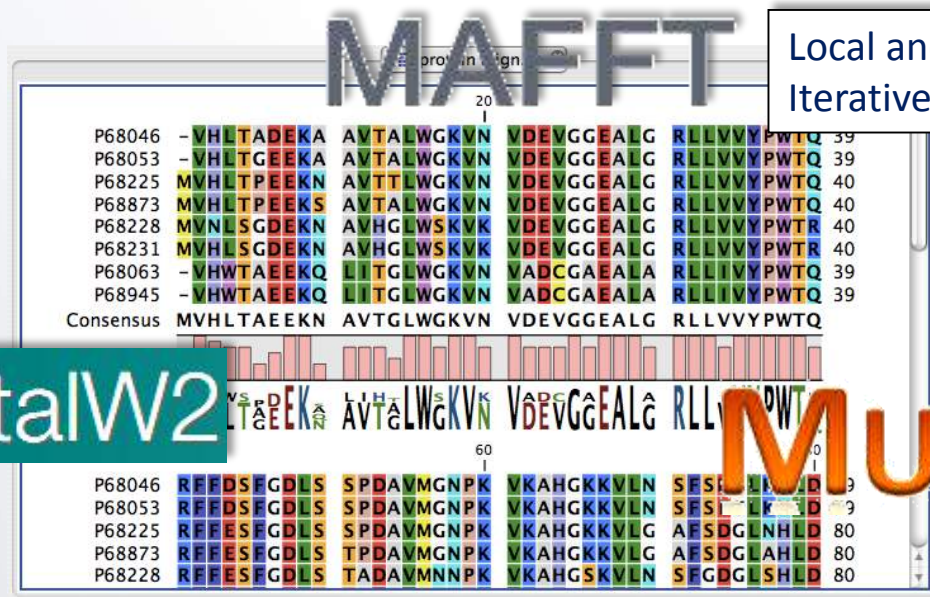
-% identity?

-Family member?

Multiple Sequence Alignment (MSA)

- Residues with **evolutionarily equivalent positions** across all sequences **matched**
- Indicates **relationship** between **residues** of different sequences
- Reveals **similarity/disimilarity**

Global alignment
Gaps



Local and global alignment
Iterative algorithm Fast-Fourier

Local and global alignment
Iterative algorithm

ClustalW2

MUSCLE

Other tools in NCBI: CD-search, MA


Sécurisé | <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

- rizaputra 70+ Gmail - Boîte de réce Webmail - Main Google Agenda Facebook Twitter / Accueil Welcome! | LinkedIn Riza Arief Putranto Riza-Arief Putranto - Google M


Specialized searches

SmartBLAST Find proteins highly similar to your query	Primer-BLAST Design primers specific to your PCR template	Global Align Compare two sequences across their entire span (Needleman-Wunsch)	CD-search Find conserved domains in your sequence
GEO Find matches to gene expression profiles	IgBLAST Search immunoglobulins and T cell receptor sequences	VecScreen Search sequences for vector contamination	CDART Find sequences with similar conserved domain architecture
Targeted Loci Search markers for phylogenetic analysis	Multiple Alignment Align sequences using domain and protein constraints	BioAssay Search protein or nucleotide targets in PubChem BioAssay	MOLE-BLAST Establish taxonomy for uncultured or enviromental sequences

Domain search of a primary protein



Conserved Domains



HOME SEARCH GUIDE NewSearch Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Conserved domains on [lcl|seqsig_MASQC_a27bead58672cb84221db2596c2506e2] View Concise Results ?

vibranium

Protein Classification ?

potato_inhibit domain-containing protein (domain architecture ID 10447212)
potato_inhibit domain-containing protein

Graphical summary Zoom to residue level show extra options >

Query seq. M A S Q C P V K N S W P E L V G T N G D I A A G I I Q T E N A N V K A I V V K E G L P I T Q D L N F N R V R V F V D E N R V V T Q V P A I G

Specific hits potato_inhibit

Superfamilies potato_inhibit superfamily

Search for similar domain architectures ? Refine search ?

List of domain hits

Name	Accession	Description	Interval	E-value
[-] potato_inhibit	pfam00280	Potato inhibitor I family; Potato inhibitor I family;	8-70	5.08e-22

Pssm-ID: 306734 Cd Length: 64 Bit Score: 79.83 E-value: 5.08e-22

seqsig_MASQC_a27bead58672cb84221db2596c2506e2 8 KNSWPELVGTNGDIAAGIIQTENANV-KAIVVKEGLPITQDLNFRVRFVDENRVVTQVPAIG 70
Cdd:pfam00280 1 KTSWPELVGKPAEEAKEIILKDRPDVtIVEVLPVGSPTDFRCNRRVRFVDGNGIVVQTPVVG 64

Protein domain

E-value

Residues

Analysis of genomics data

-Accession #?
-Annotation?

Is it already in
databases?

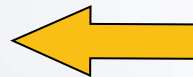


**You have just
cloned a gene**



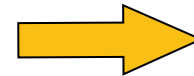
Is there conserved
regions?

-Alignments?
-Domains?



Is there similar
sequences?

-% identity?
-Family member?

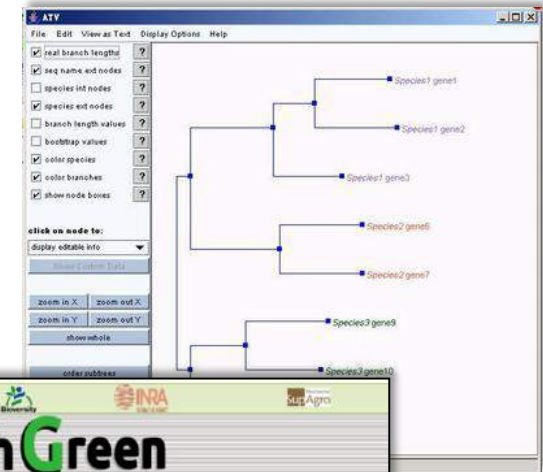
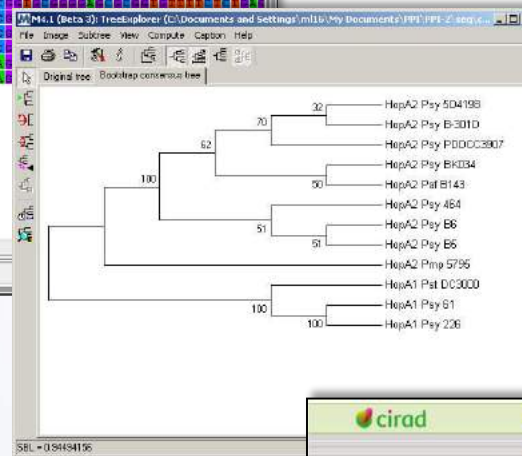
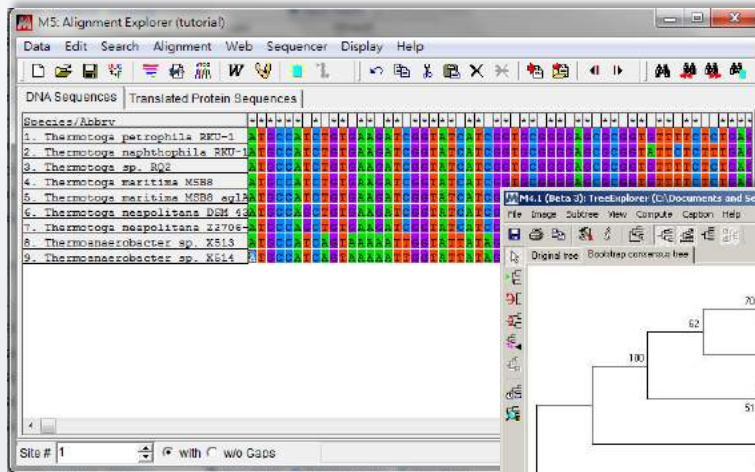


Evolutionary
relationship?

-Phylogenetic tree

Many softwares and algorithms for phylo

Phylo analysis uses different algorithms in softwares. The idea is to understand how to use them.



South Green bioinformatics platform website. The page features the logo for South Green, which includes the text 'South Green bioinformatics platform'. Below the logo, it says 'Welcome to GALAXY' and 'Our pre-configured and validated workflows'. The main content area is divided into several sections: 'NGS analyses' (with sub-sections for SNP calling and SNP analysis), 'SNP analysis' (describing the SNIPlay3 workflow), and 'GWAS' (with sub-sections for Structural variations, Chrom. reconstruction, and Metagenomics). The page also includes a navigation menu on the right side with links to 'Gene families' and 'Access workflow'.

- ✓ Neighbor joining
- ✓ Maximum likelihood
- ✓ Maximum parsimony
- ✓ Bayesian inference

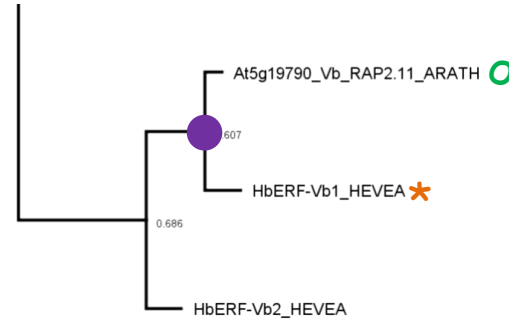
Different phylogenetic algorithms

- **Neighbor joining and UPGMA** are clustering algorithms that can make **quick trees** but are **not the most reliable**, especially when dealing with deeper divergence times. These methods are good to give you an idea about your data, but are **almost never acceptable for publication**.
- **Maximum parsimony and minimum evolution** are methods that try to **minimize branch lengths by either minimizing distance (minimum evolution) or minimizing the number of mutations (maximum parsimony)**. The major problem with these methods is that **they fail to take into account many factors of sequence evolution** (e.g. reversals and convergence homoplasies). Thus, the deeper the divergence times, more likely these methods will lead to erroneous or poorly supported groupings.
- **Maximum likelihood and Bayesian methods** can apply a **model of sequence evolution** and are **ideal for building a phylogeny using sequence data**. These methods are the two methods that are **most often used in publications** and many reviewers prefer these methods. The main downside of these methods is that they are computationally expensive. However, with today's computers this is not too much of a problem.

Phylo tree for evolutionary relationship

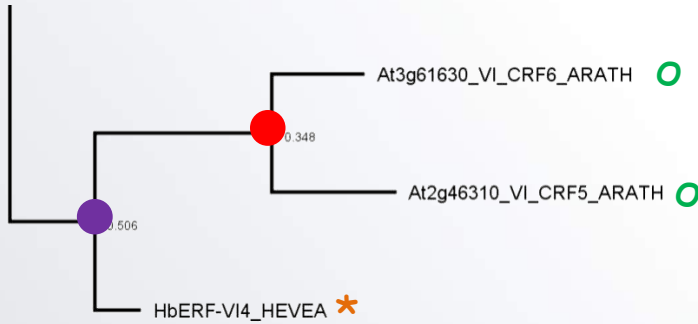


HbERF-II

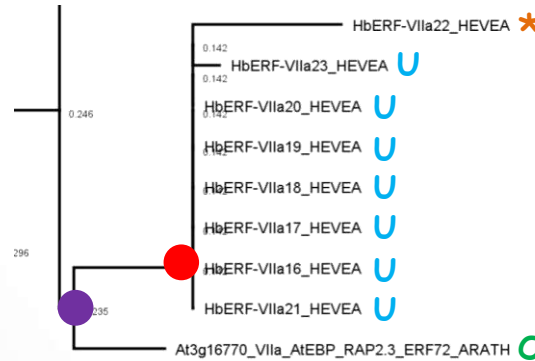


HbERF-V

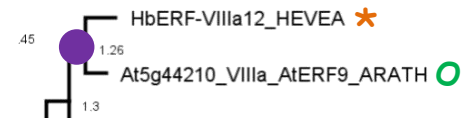
Identifikasi ortolog dan paralog dari famili gen Ethylene Response Factors
Hevea brasiliensis vs *Arabidopsis thaliana*



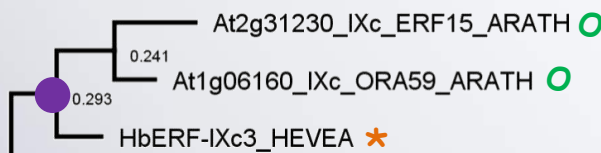
HbERF-VI



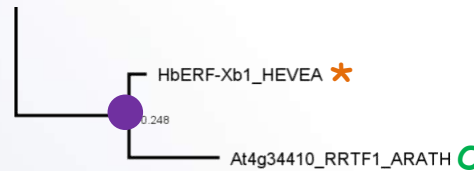
HbERF-VII



HbERF-VIII



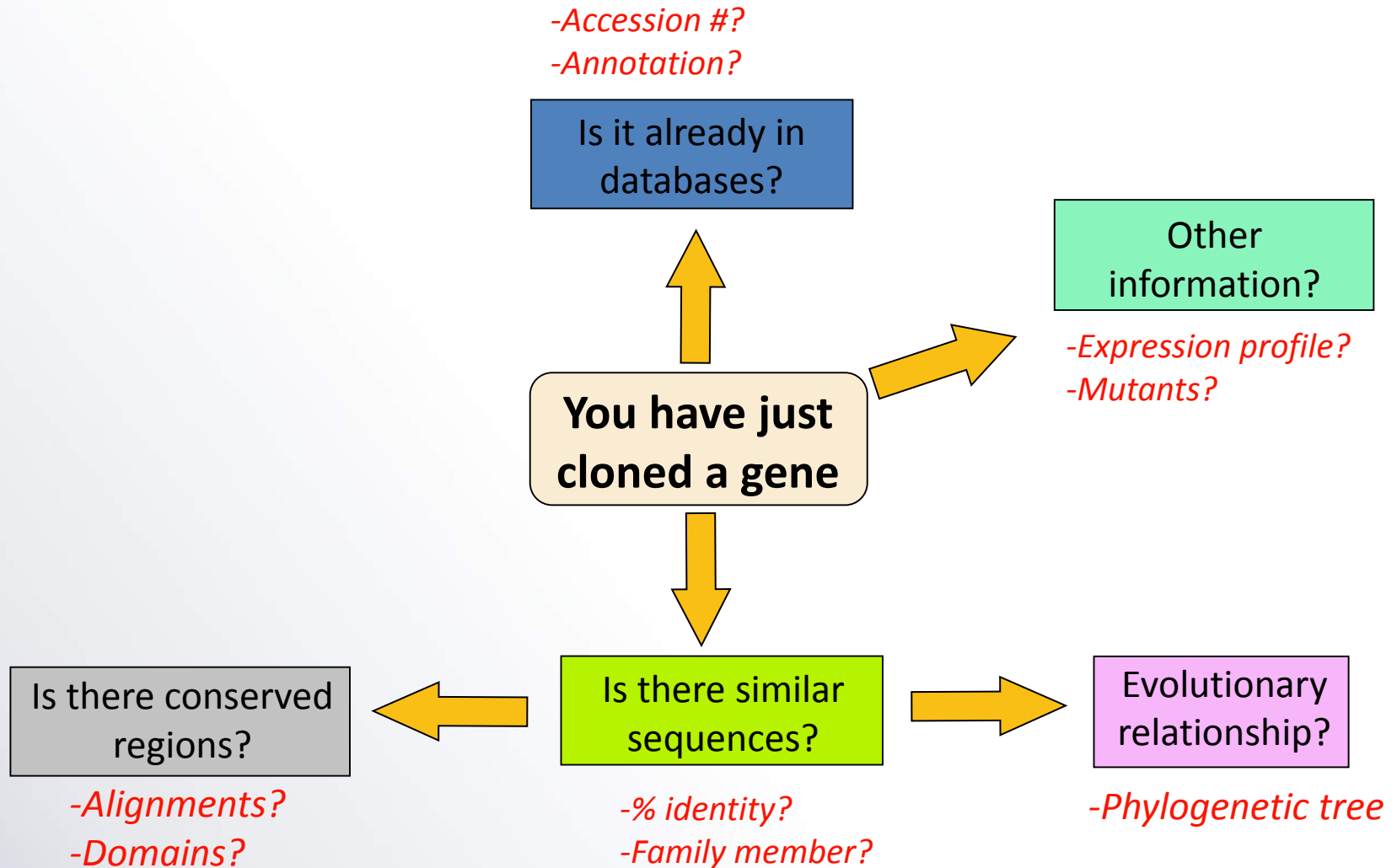
HbERF-IX



HbERF-X

- Speciation
- Duplication
- U Ultra paralog to query
- * Query
- Ortholog to query

Analysis of genomics data



Expression data on open source databases

The **Human Protein Atlas** is a Swedish-based program initiated in 2003 with the aim to **map all the human proteins in cells, tissues and organs** using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

THE HUMAN PROTEIN ATLAS

SEARCH [] Fields »

≡ MENU HELP NEWS

PCK2

TISSUE CELL PATHOLOGY

TISSUE ATLAS

PRIMARY DATA

GENE/PROTEIN

Antibody validation

Dictionary

GENERAL INFORMATION

Gene name ⁱ	PCK2
Gene description ⁱ	Phosphoenolpyruvate carboxykinase 2, mitochondrial
Protein class ⁱ	Citric acid cycle related proteins Disease related genes Enzymes Plasma proteins Potential drug targets Predicted intracellular proteins
Predicted localization ⁱ	Intracellular
Number of transcripts ⁱ	9

[SHOW MORE](#)

HUMAN PROTEIN ATLAS INFORMATION

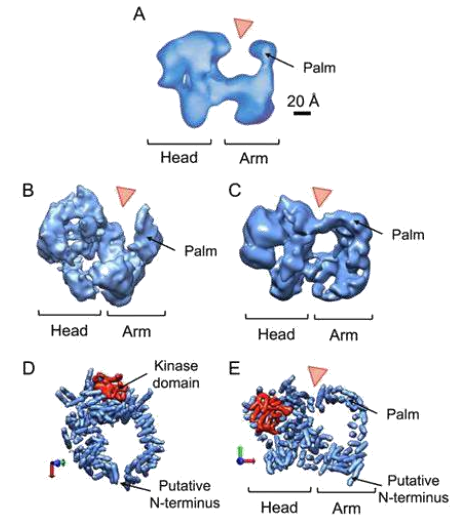
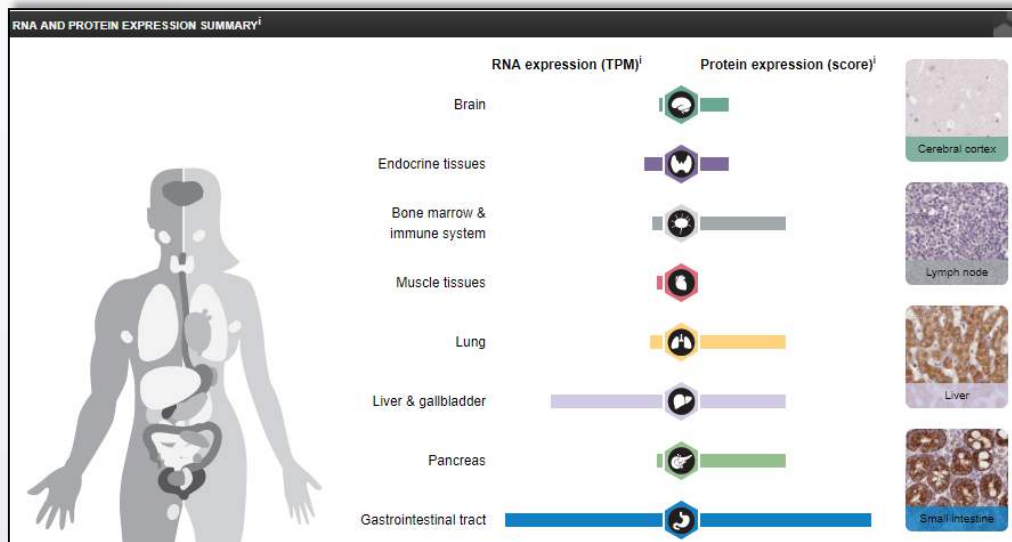
RNA tissue category ^j	HPA: Group enriched (duodenum, kidney, liver, small intestine) GTEx: Expressed in all FANTOM5: Group enriched (colon, liver, small intestine)
Protein evidence ⁱ	Evidence at protein level
Protein expression ⁱ	Distinct granular cytoplasmic expression, mainly in the small intestine, hepatocytes and renal tubules.

IMMUNOHISTOCHEMISTRY DATA RELIABILITY

Data reliability description ⁱ	Antibody staining mainly consistent with RNA expression data.
Reliability score ⁱ	Enhanced
Antibodies ⁱ	HPA051162 , HPA053502 , CAB018734

[SHOW MORE](#)

Example of PKC2 human protein



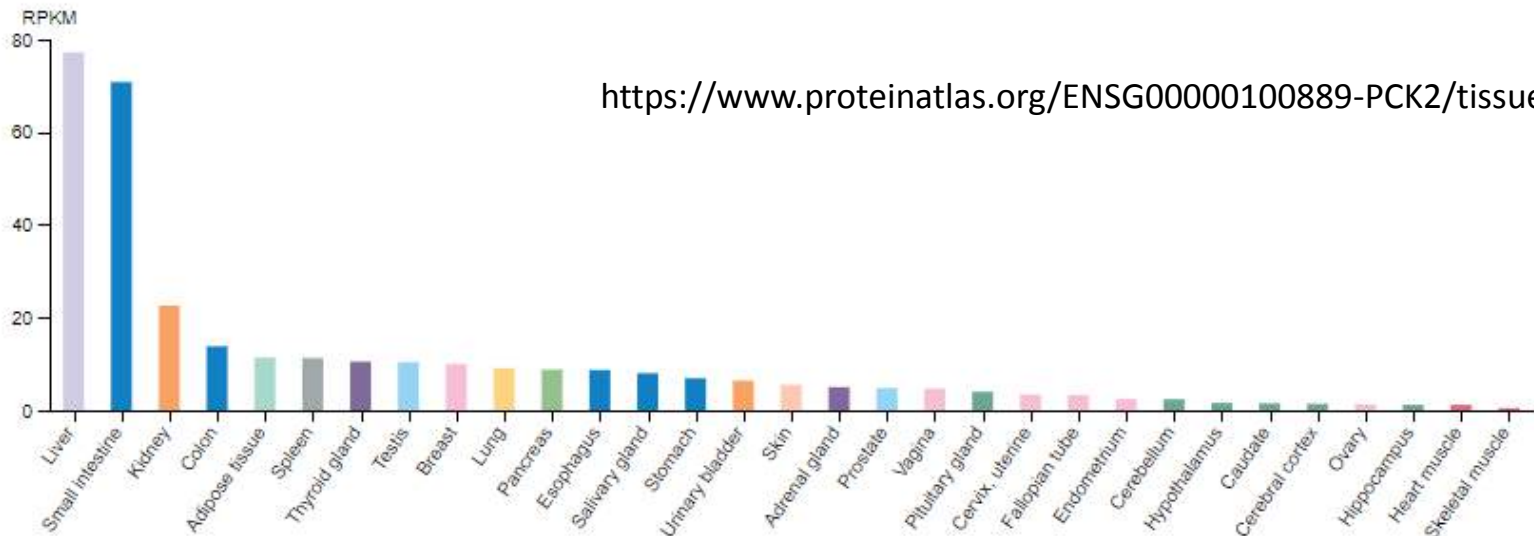
GTEX dataset¹

RNA tissue category: Expressed in all

Organ

Expression

Alphabetical



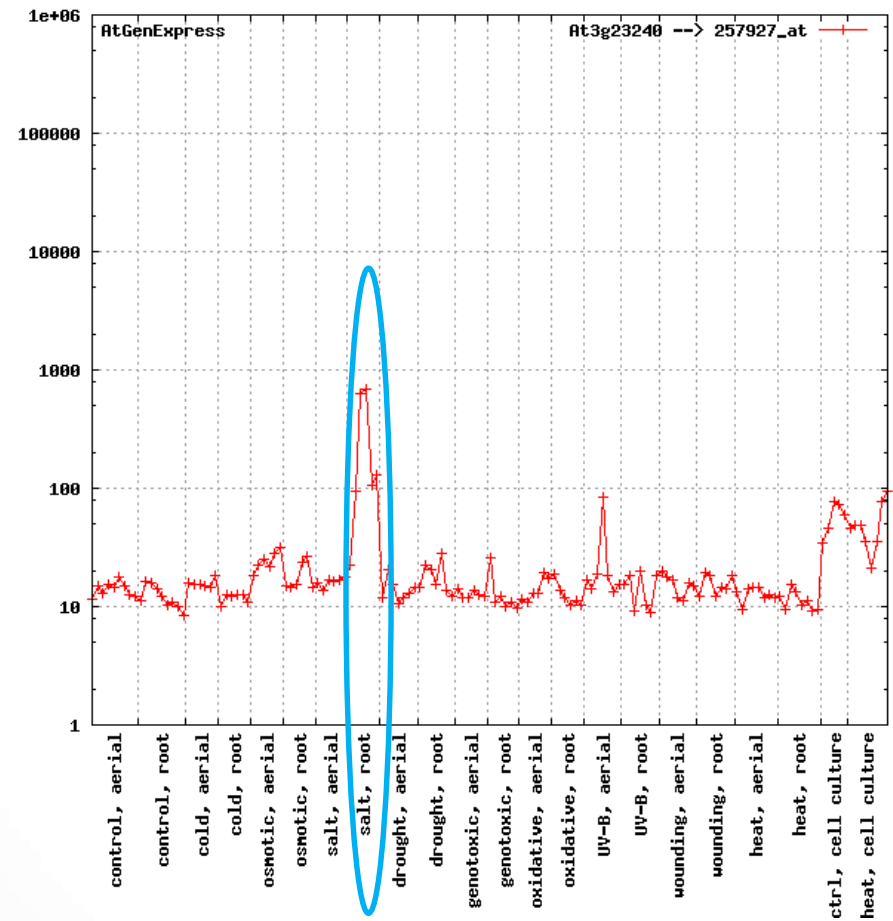
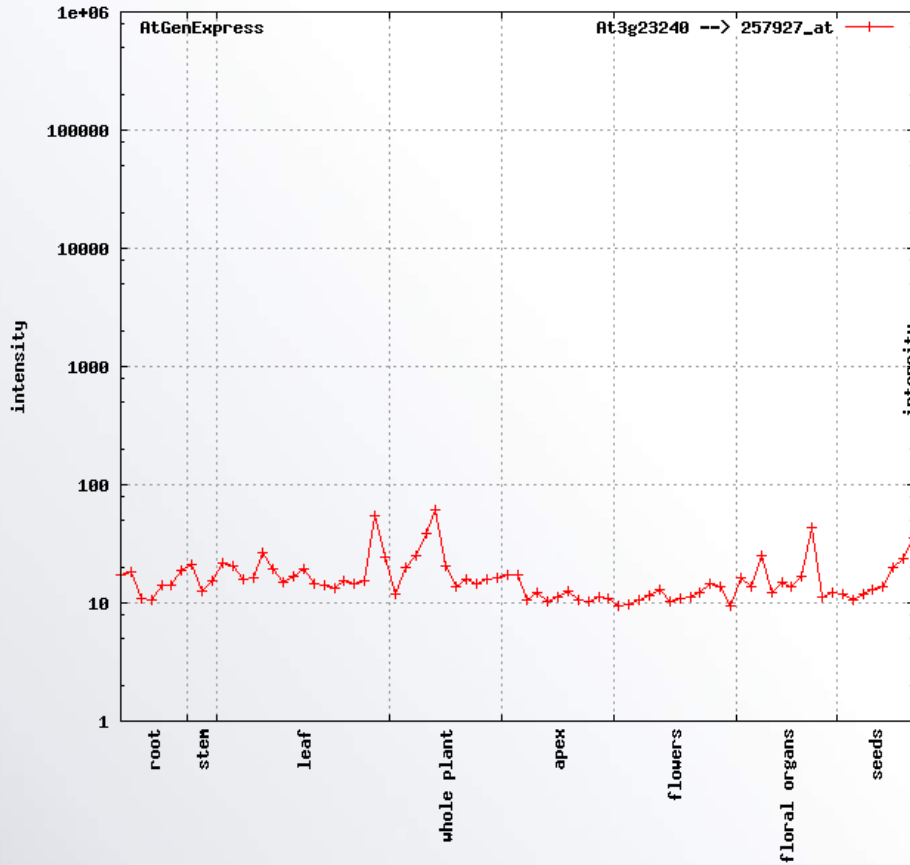
<https://www.proteinatlas.org/ENSG00000100889-PCK2/tissue>

Expression data on open source databases

The **AtGenExpress** project a comprehensive *Arabidopsis thaliana* genome transcript expression study was performed using the Affymetrix ATH1 microarray in order to understand regulatory networks in detail. We subjected, in a high-resolution kinetic series, *Arabidopsis* plants, of **identical genotype** grown under identical conditions, to **different environmental stresses** (heat, cold, drought, salt, high osmolarity, UV-B light and wounding).



Example of At3g23240.1 gene

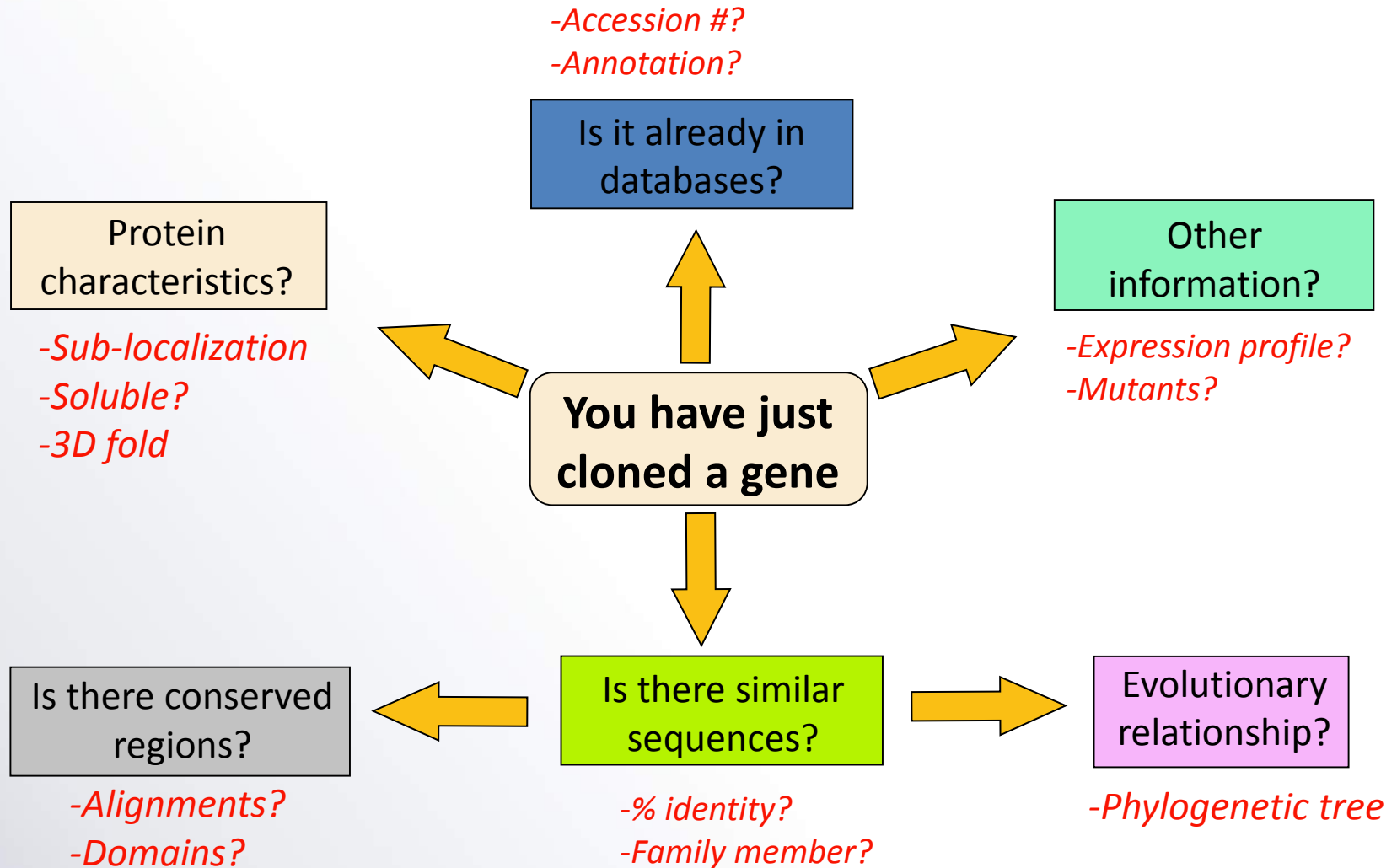


At3g23240 or **AtERF1** encodes a member of the ERF (**ethylene response factor**) subfamily B-3 of ERF/AP2 transcription factor family (ERF1). The protein contains one AP2 domain.



<http://jsp.weigelworld.org/expviz>

Analysis of genomics data

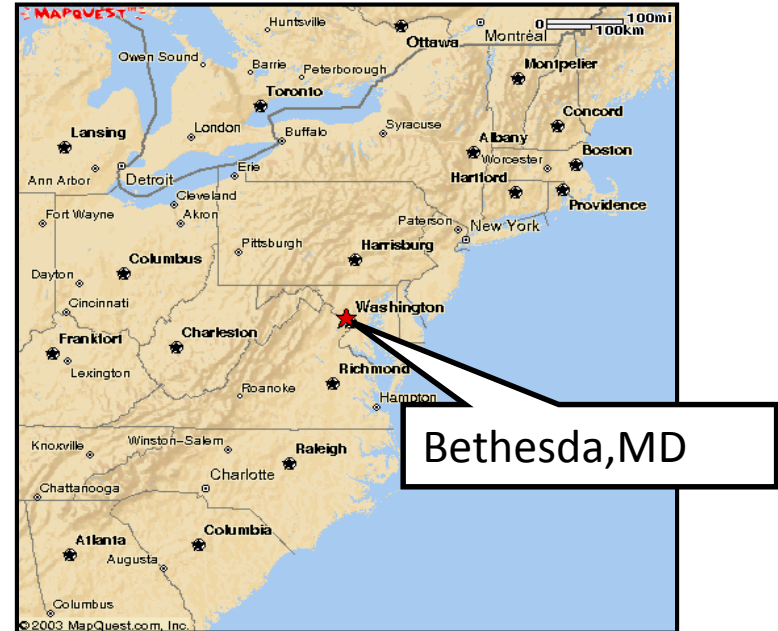


A critic for current bioinformatics:
the **lack of a single software package** that can perform all of these functions.

B. Sequence database

NCBI databases
Genome Data Viewer (GDV)

The National Center for Biotechnology Information



**Created in 1988 as a part of
the National Library of Medicine at NIH**

- Establish **public databases**
- Research in **computational biology**
- Develop **software tools for sequence analysis**
- Disseminate **biomedical** information

Web access: www.ncbi.nlm.nih.gov

The image shows a screenshot of the NCBI website homepage. At the top, there is a navigation bar with the NCBI logo, 'Resources' and 'How To' dropdown menus, and 'My NCBI | Sign In' links. Below this is a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. A yellow box highlights the text 'New pages!' in the top right corner. On the left side, there is a vertical 'Resources' menu with items like 'NCBI Home', 'All Resources (A-Z)', 'Literature', 'DNA & RNA', 'Proteins', 'Sequence Analysis', 'Genes & Expression', 'Genomes', 'Maps & Markers', 'Domains & Structure', 'Genetics & Medicine', 'Taxonomy', 'Data & Software', 'Training & Tutorials', 'Homology', 'Small Molecules', and 'Variation'. The main content area features a 'Welcome to NCBI' section with a description of the center's mission and a 'Genome' section with a sub-header '1000 prokaryotic genomes are now completed and available in the Genome database' and an image of yellow and blue microbes. To the right, there is a 'Popular Resources' section with a list of links including PubMed, PubMed Central, Bookshelf, BLAST, Gene, Nucleotide, Protein, GEO, Conserved Domains, Structure, and PubChem. At the bottom, there is a 'Common footer' section with a grid of links under the headings 'GETTING STARTED', 'RESOURCES', 'POPULAR', 'FEATURED', and 'NCBI INFORMATION'. A 'Help Desk' link is located in the top right corner of the footer area.

NCBI Resources How To My NCBI | Sign In

National Center for Biotechnology Information Search All Databases for Search

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structure
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

You are here: NCBI

GETTING STARTED

- Site Map
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES

- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

POPULAR

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

FEATURED

- GenBank
- Reference Sequences
- Map Viewer
- Genome Projects
- Human Genome
- Mouse Genome
- Influenza Virus
- Primer-BLAST
- Short Read Archive

NCBI INFORMATION

- About NCBI
- Research at NCBI
- NCBI Newsletter
- NCBI FTP Site
- Contact Us

Help Desk

Common footer

What are databases?

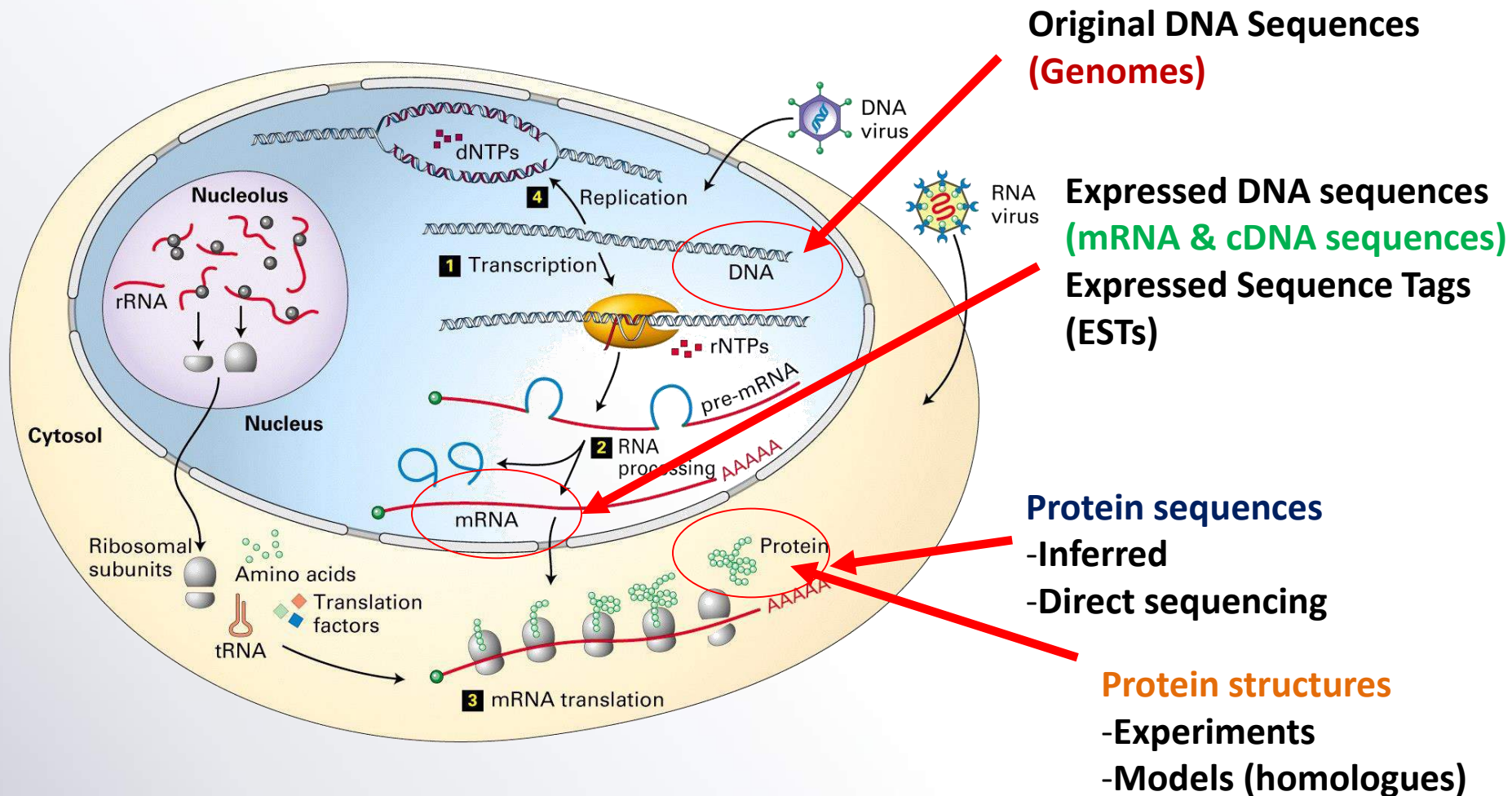
- ❑ **Structured** collection of information.
- ❑ Consists of basic units called **records** or **entries**.
- ❑ Each record consists of **fields**, which hold **pre-defined** data related to the record.
- ❑ For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence)

The “Perfect” database

- ❑ Comprehensive, but **easy to search**
- ❑ **Annotated**, but not “too annotated”
- ❑ A **simple**, easy to understand **structure**
- ❑ **Cross-referenced**
- ❑ Minimum **redundancy**
- ❑ **Easy retrieval** of data



The molecular biology dogma and biological data



NCBI databases and services

- ❑ **GenBank** primary sequence database
- ❑ **Free public access** to biomedical literature
 - PubMed free Medline (3 million searches per day)
 - PubMed Central full text online access
- ❑ **Entrez** integrated molecular and literature databases



Type of molecular databases

□ **Primary** Databases

- ✓ **Original submissions** by experimentalists
- ✓ Content controlled by the **submitter**

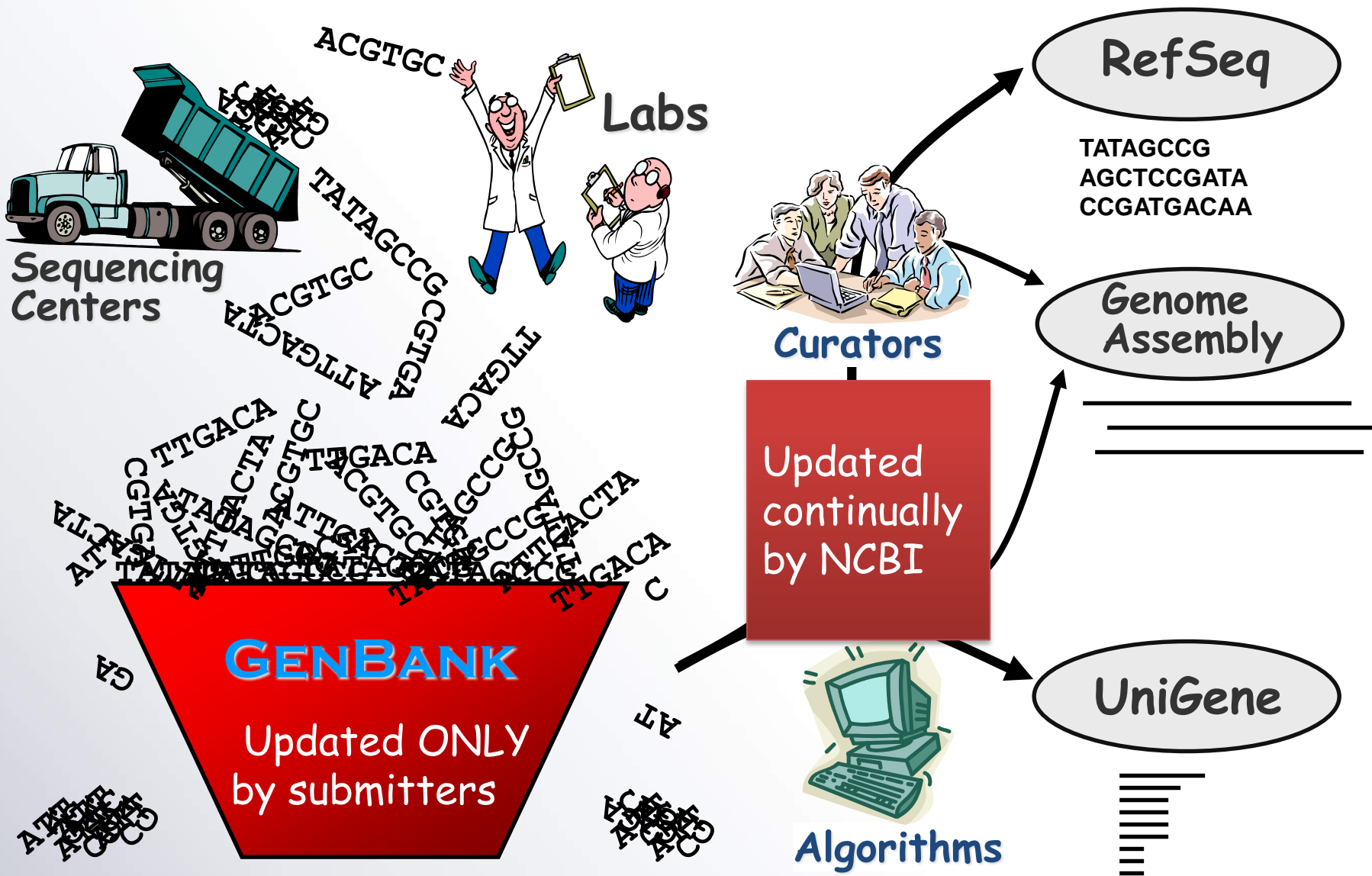
Examples: GenBank, Trace, SRA, SNP, GEO

□ **Derivative** Databases

- ✓ **Derived from primary data**
- ✓ Content controlled by **third party** (NCBI)

Examples: NCBI Protein, Refseq, TPA, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain

Primary sequence submission



Primary sequence submission

The screenshot shows the GenBank Submission Portal. At the top left is the NIH logo and the text "U.S. National Library of Medicine National Center for Biotechnology Information". At the top right is a user profile icon labeled "putranto". Below the header is a navigation bar with "Home", "My submissions" (which is underlined), "Templates", and "My profile". The main content area features the "GenBank" logo and a blue "New submission" button. A light blue informational box contains a note: "Note: Submit only ribosomal RNA (rRNA), rRNA-ITS or Influenza sequences here. All other submission types should use one of the alternate submission tools (e.g. BankIt, Sequin, tbl2asn, etc.)." Below this, a section titled "Prokaryotic rRNA submissions" lists requirements: all sequences must be prokaryotic; FASTA files must contain sequences from 16S, 23S, or 16S-23S ribosomal RNA; samples must be from uncultured environmental sources or pure cultured strains; and next-generation sequencing data must be assembled or processed into OTUs, bins, or phylotypes.

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

putranto

Submission Portal

Home My submissions Templates My profile

GenBank [New submission](#)

Note: Submit only **ribosomal RNA (rRNA)**, **rRNA-ITS** or **Influenza** sequences here.
All other submission types should use one of the alternate [submission tools](#) (e.g. [BankIt](#), [Sequin](#), [tbl2asn](#), etc.)

Prokaryotic rRNA submissions must meet the following requirements:

- All sequences are prokaryotic
- All sequences in the FASTA file contain sequences from one of the following types: 16S ribosomal RNA, 23S ribosomal RNA, or 16S-23S ribosomal RNA intergenic spacer region
- Sampled from an uncultured, environmental source or from pure cultured strains
- Sequences from 454, Illumina or next generation sequencing technologies are accepted only if they are assembled (each sequence was assembled from two or more overlapping sequence reads) or processed into OTUs, bins, or individual phylotypes.

Follow the steps of submission – easy to handle but prepare all the information and annotation as complete as possible

Sequence databases at NCBI

□ Primary

- ✓ **GenBank**: NCBI's primary sequence database
- ✓ **Trace Archive**: reads from capillary sequencers
- ✓ **Sequence Read Archive**: next generation data

□ Derivative

- ✓ **GenPept** (GenBank translations)
- ✓ Outside Protein (**UniProt—Swiss-Prot, PDB**)
- ✓ NCBI Reference Sequences (**RefSeq**)

GenBank – Primary sequence database

- ❑ **Nucleotide only** sequence database

- ❑ **Archival** in nature
 - ✓ Historical
 - ✓ Reflective of submitter point of view (subjective)
 - ✓ **Redundant**

- ❑ **Data**
 - ✓ **Direct submissions** (traditional records)
 - ✓ Batch submissions
 - ✓ FTP accounts (genome data)

Principal GenBank record

LOCUS HSHMLHI 2503 bp mRNA linear PRI 31-MAR-1994
DEFINITION Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION U07418
VERSION U07418.1 GI:466461
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 2503)

Accession

- Stable
- Reportable
- Universal

ACCESSION U07418

VERSION U07418.1 GI:466461

Version

Tracks changes in sequence

GI number

NCBI internal use

ei Wei, Molecular
Medical Center Drive

20850, USA

Principal GenBank record

LOCUS	HSHMLHI	2503 bp	mRNA	linear	PRI 31-MAR-1994
DEFINITION	Human DNA mismatch repair protein 1	FEATURES			
ACCESSION	U07418	Location/Qualifiers			
VERSION	U07418.1 GI:46111	source 1..2503			
KEYWORDS	.	/organism="Homo sapiens"			
SOURCE	Homo sapiens (human)	/db_xref="taxon:9606"			
ORGANISM	<u>Homo sapiens</u>	/chromosome=1			
	Eukaryota; Metazoa; Mammalia; Eutheria	/map=1			
REFERENCE	1 (bases 1 to 2503)	/tissue="colon" /dev="col" /gene="HSHMLHI" /feature="CDS" /function="DNA mismatch repair" /note="SwissProt: P12111" /codon="standard" /protein="HSHMLHI" /db_xref="PubMed:8128251" /translation="TSIQVLSALASISRRKALVFGNAVETVYAF LGSNSSLDAFLQLTKGTS LQBEINDFANFYFSLERFYSIRFLQLANL" //			
		BASE COUNT 723 a 539 c 599 g 642 t			
		ORIGIN			
		1 gttgaacatc tagacgtttc cttggctcct ctggcgccaa aatgtcgttc gtggcagggg			
		61 ttattcggcg gctggacgag acagtgggga accgcatcgc ggcgggggaa gttatccagc			
		121 ggccagctaa tgctatcaaa gagatgattg agaactgttt agatgcaaaa tccacaagta			
		181 ttcaagtgtat tgtaaaagag ggaggcctga agttgattca gatccaagac aatggcaccg			
		241 ggatcaggaa agaagatctg gatattgtat gtgaaagggt cactactagt aaactcgagt			
		301 cctttgagga tttagccagt atttctacct atggctttcg aggtgaggct ttggccagca			
		361 taagccatgt ggctcatgtt actattacaa cgaaaacagc tgatggaaaag tgtgcataca			
		421 gagcaagtta ctcagatgga aaactgaaag cccctcctaa accatgtgct ggcaatcaag			
		481 ggaccagat cacgggtggag gacctttttt acaacatagc cacgaggaga aaagcttaa			
		541 aaaatccaag tgaagaatat gggaaaattt tggaaagtgt tggcaggtat tcagtacaca			
		601 atgcaggcat tagtttctca gttaaaaaac aaggagagac agtagctgat gttaggacac			
		661 tacccaatgc ctcaaccgtg gacaatattc gctccgtcct tggaaatgct gttagtcgag			
		721 aactgataga aattggatgt gaggataaaa ccctagcctt caaaatgaat ggttacatat			
		781 ccaatgcaaa ctactcagtg aagaagtgca tcttcttact cttcatcaac catcgtctgg			
		841 tagaatcaac ttcttgaga aaagccatag aaacagtgtat tgcagcctat ttgccaaaa			
		901 acacacacc attcctgtac ctcagtttag aaatcagtc ccagaatgtg gatgtaaatg			
		961 tgcaccccac aaagcatgaa gttcacttcc tgcacgagga gacatcctg ggacgggtgc			
		1021 agcagccatc cgagagcaag ctctgggctt ccaattctct caggatgtac ttcaccagaa			
		1081 ctttgctacc aggacttgct ggcccctctg gggagatggt taaatccaca acaagtctga			
		1141 cctcgtcttc tacttctgga agtagtgata agtctatgc ccaccagatg gttcgtacag			
		1201 attcccggga acagaagctt gatgcatctt tgcagcctct gagcaaaccc ctgtccagtc			
		1261 agccccaggc cattgtcaca gaggataaga cagatatttc tagtggcagg gctaggcagg			
		1321 aagatgagga gatgcttgaa ctcccagccc ctgctgaagt ggctgcaaaa aatcagagct			
		1381 tggaggggga tacaacaag gggacttcag aaatgcaga gaagagagga cctacttcca			
		1441 gcaaccccag aaagagacat cgggaagatt ctgatgtgga aatggtggaag gatgattccc			
		1501 gaaaggaaat gactgcagct gtaccctccc ggagaagatt tttaccctcc cattaacctc			
		1561 tgagtctcca ggaagaaatt aatgagcagg gacatgaggt tctccgggag atggttgcata			
		1621 accactcctt cgtgggctgt gtgaatcctc agtgggcctt ggcacagcat caaaccaagt			
		1681 tataccttct caacaccacc aagcttagtg aagaactggt ctaccagata ctcatattg			
		1741 attttgccaa ttttgggttt ctcaggttat cggagccagg accgctcttt gaccttgcca			
		1801 tgcttgctt agatagttcca gagagtggtt ggacagagga agatggtccc aaagaaggac			
		1861 ttgctgaata cattgttgat ttctgaaga agaaggctga gatgcttga gactatttct			
		1921 ctttggaaat tgatgaggaa gggaaacctga ttgattacc ccttctgatt gacaaactatg			
		1981 tgcccccttt ggagggactg cctatcttca ttctcgact agcccactgag gtgaattggg			
		2041 acgaagaaaa ggaatgttt gaaagcctca gtaaagaatg cgctatgttc tattccatcc			
		2101 ggaagcagta catatctgag gagtgcagcc tctcaggcca gcagagtgaat gtcctggct			
		2161 ccattccaaa ctctggaag tggactgtgg aacacattgt ctataaagcc ttgctgctcac			
		2221 acattctgct tcctaaacct ttcacagaag atggaaatat cctgcagctt gctaacctgt			
		2281 ctgactata caaagctctt gagaggtggt aaatctggtt ttgctgcaat gttggcagtg			
		2341 gttctcttt ctctgattc cgatacaag ttgtgtatca aagtgtgata tacaaagttg			
		2401 accaacataa gtgtgtgag cacttaagac ttatacttgc cttctgatag tattcctta			
		2461 tacacagtgg attgattata aataaataga tgtgtcttaa cat			

ACCESSION

VERSION

Version
Tracks changes in sequence


Well annotated

The sequence is the data

GenPept: GenBank CDS Translations

```
FEATURES             Location/Qualifiers
     source            1..2484
                        /organism="Homo sapiens"
                        /mol_type="mRNA"
                        /db_xref="taxon:9606"
                        /chromosome="3"
                        /map="3p22-p23"
     gene              1..2484
                        /gene="M
CDS             22..2292
                        /gene="M
                        /note="homolog of S. cerevisiae PMS1 (Swiss-Prot Accession
                        Number P14242), S. cerevisiae MLH1 (GenBank Accession
                        Number U07187), E. coli MUTL (Swiss-Prot Accession Number
                        P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession
                        Number P14161) and Streptococcus pneumoniae (Swiss-Prot
                        Accession Number P14160) "
                        /codon_start=1
                        /product="DNA mismatch repair protein homolog"
                        /protein_id="AAC50285.1"
                        /db_xref="GI:463989"
                        /translation="MSFVAGVIRRLDET VVNRIAAGEVIQR PANAIKEMIENCLDAK
                        SSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTT SKLQSFEDLASISTYGF
                        RGEALASISHVAHVTTITTKTADGKCA YRASYS DGK LKAPPKPCAGNQTQIT
                        VEDLFYNIATRRKALKNPSEEY GKILEVVG RYSVHNAGISF SVKKQGETVADV
                        R TLPNASTVDNIRS
```

>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDET VVNRIAAGEVIQR PANAIKEMIENCLDAKSTS IQVIV...
EDLDIVCERFTT SKLQSFEDLASISTYGF RGEALASISHVAHVTTITTKTAD...

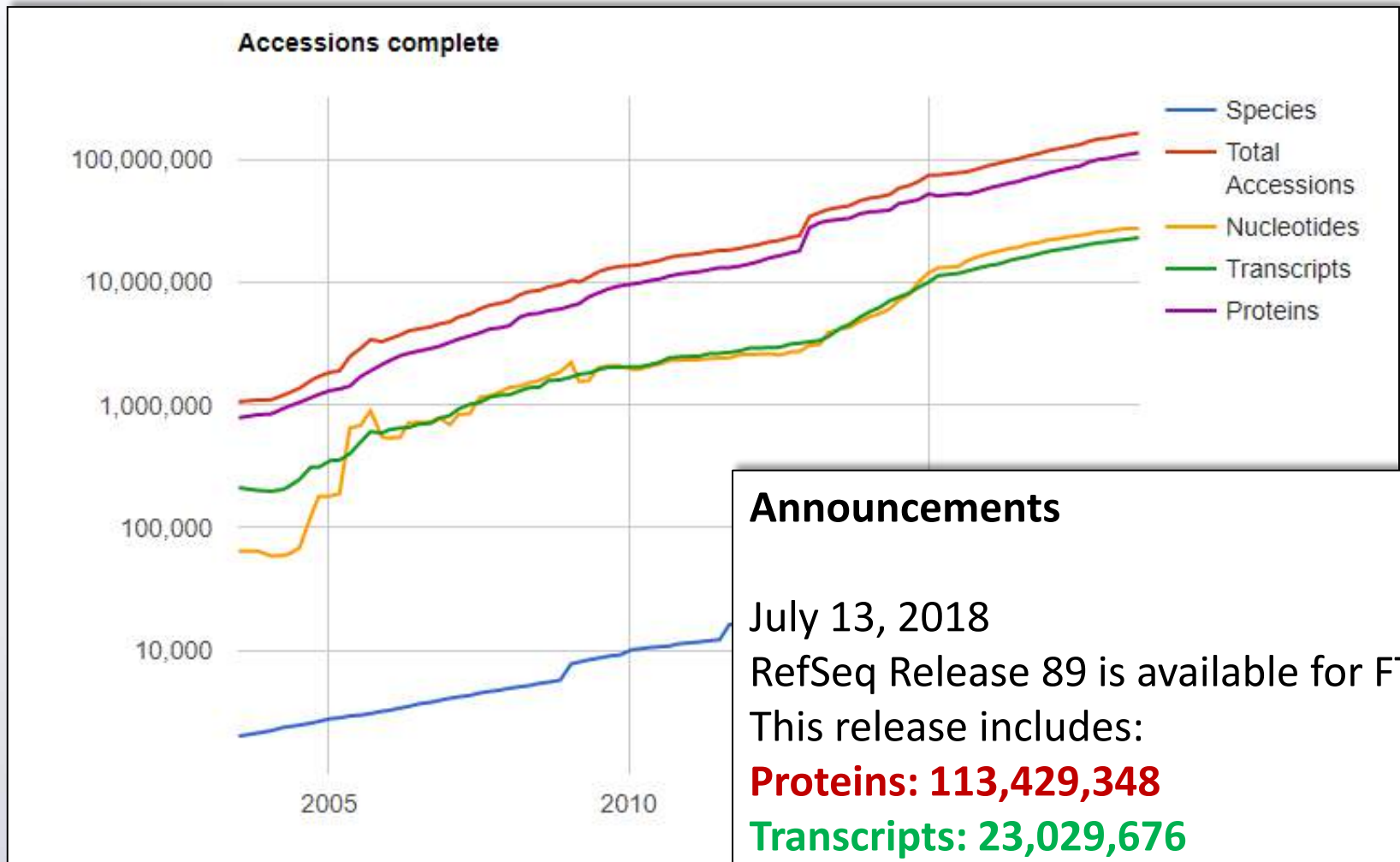


RefSeq: Derivative sequence database

- ❑ **Curated** transcripts and proteins
- ❑ **Model** transcripts and proteins
- ❑ **Assembled Genomic Regions**
- ❑ **Chromosome** records
 - Human genome
 - Microbial
 - Organelle

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

RefSeq: Derivative sequence database



Announcements

July 13, 2018

RefSeq Release 89 is available for FTP

This release includes:

Proteins: 113,429,348

Transcripts: 23,029,676

Organisms: 81,345

From GenBank to RefSeq

[Human apolipoprotein E \(epsilon-4 allele\) gene, complete cds](#)
1. 5,515 bp linear DNA
M10065.1 GI:178852

[Human mRNA fragment for apolipoprotein E \(apo E\)](#)
2. 528 bp linear mRNA
X00199.1 GI:28808

[H.sapiens mRNA](#)
3. 275 bp linear m
Z70760.1 GI:1263

[Homo sapiens apolipoprotein E \(APOE\), mRNA](#)
1,223 bp linear mRNA

[Homo sapiens c](#)
4. 1,023 bp linear
AK314898.1 GI:16

NM_000041.2 GI:48762938

[Human apolipoprotein E mRNA, complete cds](#)
5. 1,157 bp linear mRNA
M12529.1 GI:178848

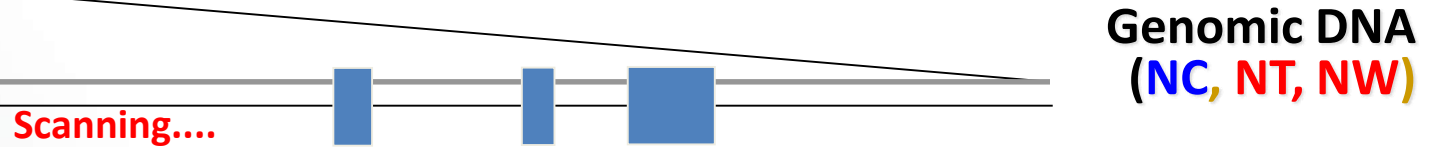
Curation number


RefSeq

[Homo sapiens preapolipoprotein E \(APOE\) mRNA, complete cds](#)
6. 1,156 bp linear mRNA
K00396.1 GI:178850

[Homo sapiens apolipoprotein E, mRNA \(cDNA clone MGC:1571 IMAGE:3355712\), complete cds](#)
7. 1,186 bp linear mRNA
BC003557.1 GI:13097698

RefSeq: The finalized version of GenBank



Model mRNA (XM)
(XR)  → Model protein (XP)

Curated mRNA (NM)
(NR)  → Curated Protein (NP)

RefSeq

GenBank
Sequences



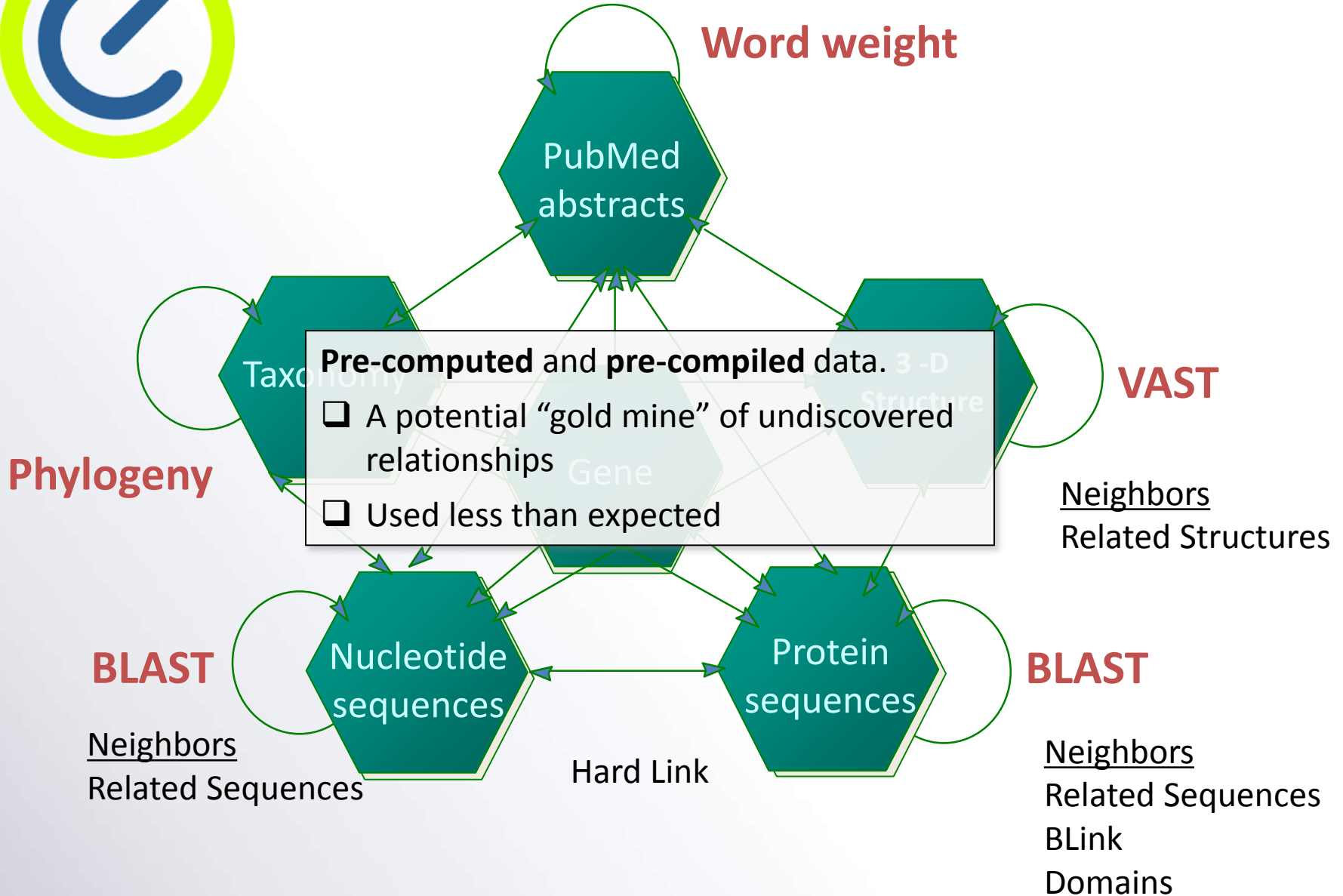
RefSeq benefits

- ❑ **Non-redundancy**
- ❑ **Updates** to reflect current sequence data and biology
- ❑ Data **validation**
- ❑ Format **consistency**
- ❑ **Distinct accession series**
- ❑ **Stewardship by NCBI staff and collaborators**



The Entrez interface

Word weight



Entrez: All access to database in NCBI

Search NCBI databases

apolipoprotein E



Search

Results found in 32 databases for **apolipoprotein E**

Literature

Bookshelf	486	Books and reports
MeSH	29	Ontology used for PubMed indexing
NLM Catalog	23	Books, journals and more in the NLM Collections
PubMed	23,329	Scientific and medical abstracts/citations
PubMed Central	34,431	Full-text journal articles
PubMed Health	37	Clinical effectiveness, disease and drug reports

Genes

EST	3,386	Expressed sequence tag sequences
Gene	5,653	Collected information about gene loci
GEO DataSets	3,600	Functional genomics studies
GEO Profiles	277,988	Gene expression and molecular abundance profiles
HomoloGene	13	Homologous gene sets for selected organisms
PopSet	33	Sequence sets from phylogenetic and population studies
UniGene	33	Clusters of expressed transcripts

The Entrez system: 41 (and counting) **integrated** databases

When in GenBank you see sequence individually, in GDV you see the whole information in the genome...

What is NCBI Genome Browser?

- ❑ **NCBI's main genome (assembly) browser** for eukaryotic organisms -- replacing Map Viewer
- ❑ Familiar Sequence Viewer panel with functions
 - ✓ **Navigating** and **searching** across the assembly
 - ✓ Uploading **mapped data**
 - ✓ Accessing **analysis tools**
- ❑ Over 500 eukaryotic genome assemblies available

The Genome Data Viewer (GDV): NCBI's Genome Browser

Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 500 eukaryotic RefSeq genome assemblies. ⓘ

Select organism

Homo sapiens (human)

Homo sapiens (human) genome

Search in genome

Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

[Browse genome](#) [BLAST genome](#)

Assembly details

Name	GRCh38.p11
RefSeq accession	GCF_000001405.37
GenBank accession	GCA_000001405.26
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome

Annotation details

Annotation Release	108
Release date	2016-06-06

HHS NIH NATIONAL LIBRARY OF MEDICINE NCBI

www.ncbi.nlm.nih.gov/genome/gdv

Browsing the species

Select organism

Homo sapiens (human)

A phylogenetic tree with a root on the left and two main branches. The top branch includes zebrafish and Japanese medaka. The bottom branch includes Xenopus tropicalis, chicken, rat, dog, and pig. The human node is highlighted with a blue circle and a human icon.

Select organism

Danio rerio (zebrafish)

Danio rerio (zebrafish) genome

Search in genome

Location, gene or phenotype

800, DNA repair

GRCz11

[Browse genome](#) [BLAST genome](#)

Assembly details

Name	GRCz11
RefSeq accession	GCF_000002035.6
GenBank accession	GCA_000002035.4
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome

Annotation details

Annotation Release	106
Release date	2017-06-26

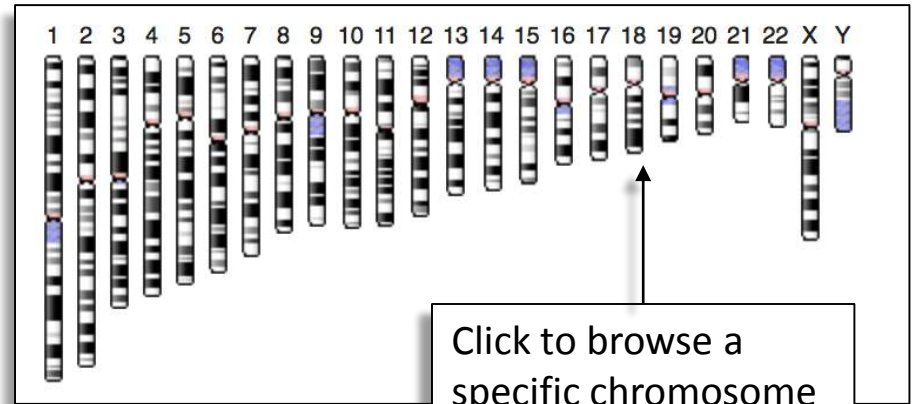
An ideogram of the 25 chromosomes of the zebrafish genome, numbered 1 through 25. The chromosomes are represented as vertical bars of varying lengths and are arranged in a roughly descending order of size.

Phylogenetic tree provides a convenient way to **browse by organism**

Browsing the human genome

Search with

- ✓ gene names
- ✓ SNP ids
- ✓ RefSeq Accessions
- ✓ chromosome positions



Homo sapiens (human) genome

Search in genome

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

Assembly

GRCh38.p11

[Browse genome](#) [BLAST genome](#)

Pick Assembly

- ✓ GRCh38.p11
- ✓ GRCh37.p13
- ✓ HuRef (Celera)
- ✓ CHM1_1.1 (hydatid molar cell line)

Enter browser at chromosome 1

BLAST search the assembled sequences

Main browser interface

The screenshot displays the Genome Data Viewer interface for the region NC_000001.11: 109,687,201 - 109,694,340 on Chromosome 1. The interface includes several key components:

- Current Chromosome Ideogram:** Located at the top, showing a chromosome ideogram with the current region highlighted in blue.
- Gene & Exon navigator:** A navigation bar below the ideogram showing the selected gene (GSTM1) and transcript (NM_000561.3), along with exon navigation controls.
- Search results:** A sidebar on the left showing search results for 'GSTM1', listing genes and their locations on various chromosomes.
- Sequence viewer:** The main central area displaying multiple tracks including gene annotations, dbSNP variants, RNA-seq exon coverage, and RNA-seq intron features.
- Additional functions:** A sidebar at the bottom left lists options like 'Your Data', 'BLAST', 'Add Tracks', 'Assembly Region Details', and 'History'.

Name	Location
GSTM1	Chr1: 109,687.8K - 109,693.7K
TP53	Chr17: 7,668,402 - 7,687,550
MTHFR	Chr1: 11,785,730 - 11,806,103
GSTT1	NT_187633.1: 270.3K - 278.5K
VEGFA	Chr6: 43,770,209 - 43,786,487
GSTP1	Chr11: 67,583,595 - 67,586,653
IL6	Chr7: 22,725,869 - 22,732,002
ACE	Chr17: 63,477,061 - 63,498,380

Play with NCBI

1. Buka website NCBI: <https://www.ncbi.nlm.nih.gov/>
2. Ketik kata kunci “apolipoprotein E” di box serch dan amati hasilnya
3. Ada berapa gen yang muncul? Ada berapa protein?
4. Klik “Gene” dan pilihlah gen “APOE”
5. Bagaimana anda tahu jika gen “APOE” sudah versi *curated version*? Silahkan cek <https://www.ncbi.nlm.nih.gov/gene/348>
6. Kembali ke Entrez
7. Carilah struktur protein dari “apolipoprotein E”

Play with GDV

1. Buka website GDV:
<https://www.ncbi.nlm.nih.gov/genome/gdv/>
2. Pilihlah “Human genome”
3. Carilah gen BRCA1 (gen penyebab kanker payudara)
4. Surfing pada Genome Browser
5. Di kromosom berapa gen tersebut ditemukan?
6. Ada berapa exon (sekuen penyandi protein) dari gen tersebut?

It was **still the second course,**
don't get dizzy yet

