Universitas Esa Unggul

Smart, Creative and Entrepreneurial

IBT 432 Aplikasi Bioinformatika

Database sekuen dan analisis genomika

Riza Arief Putranto

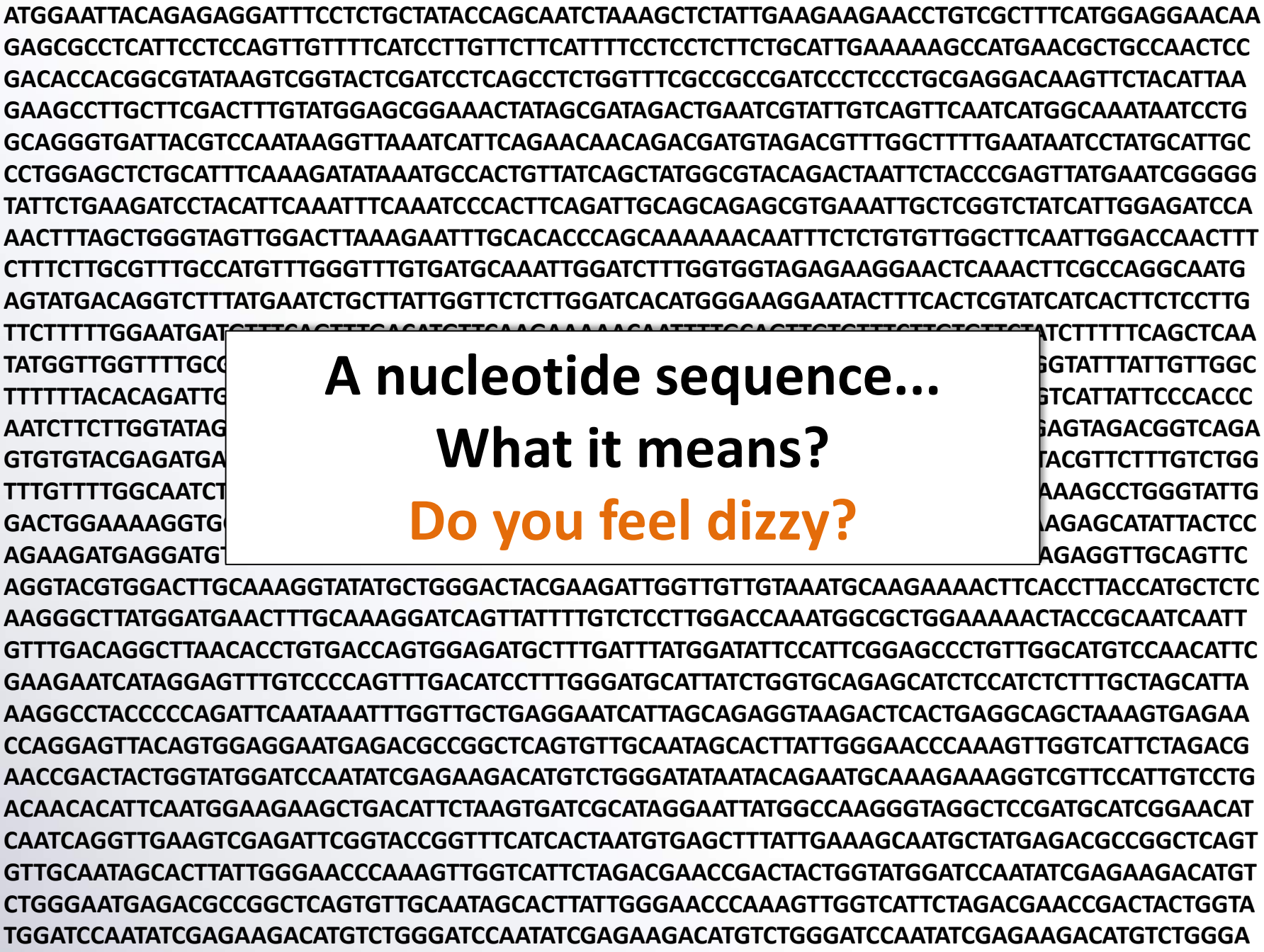
Rencana Perkuliahan

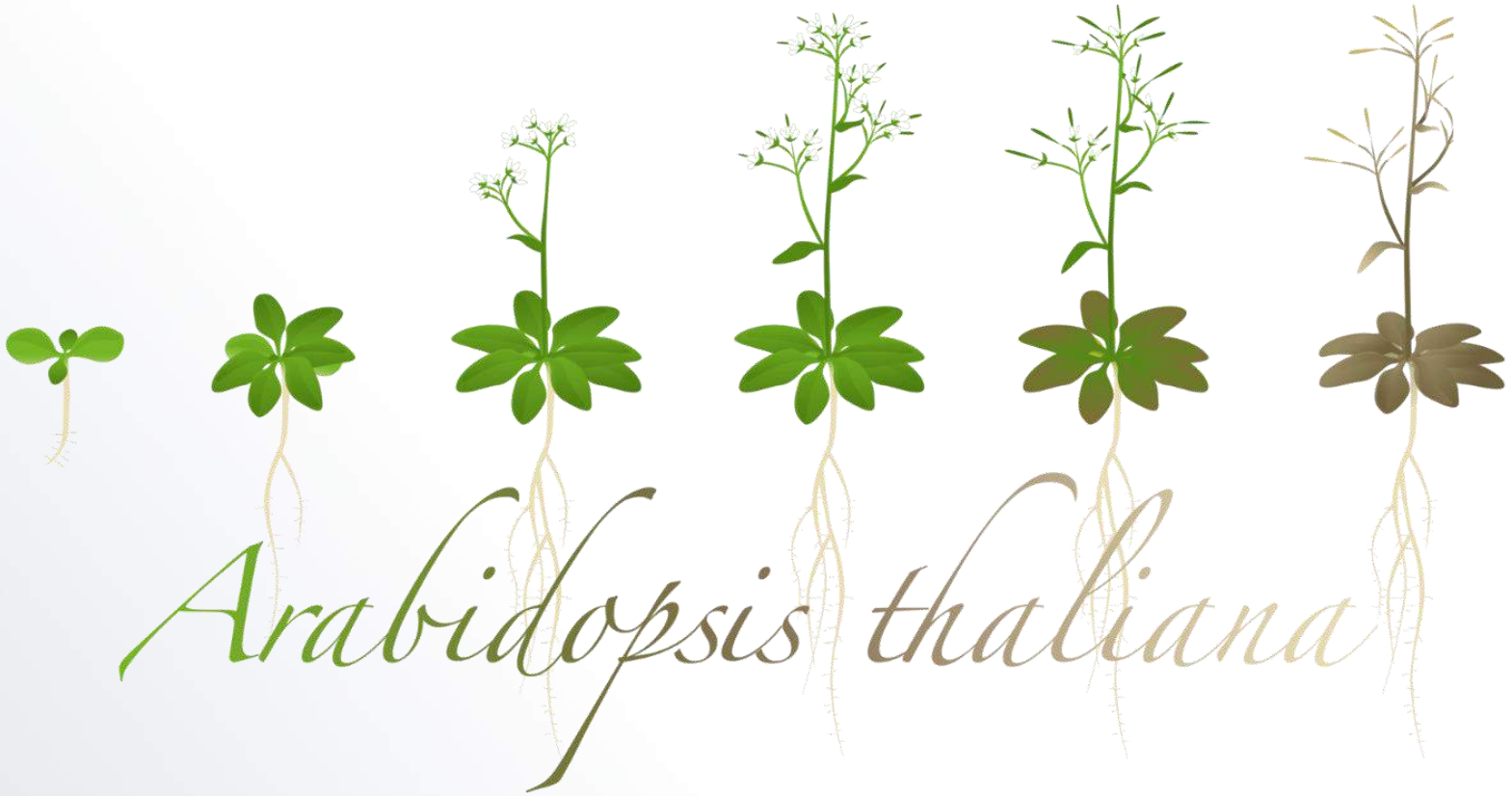
1. ~~Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. Database sekuen dan analisis genomika
3. Anotasi sekuen ke genom - Praktik
4. Analisis komparasi genomika I
5. Analisis komparasi genomika II
6. Analisis komparasi genomika III
7. Analisis komparasi genomika – Praktik
8. Protein modelling 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


A nucleotide sequence...
What it means?
Do you feel dizzy?



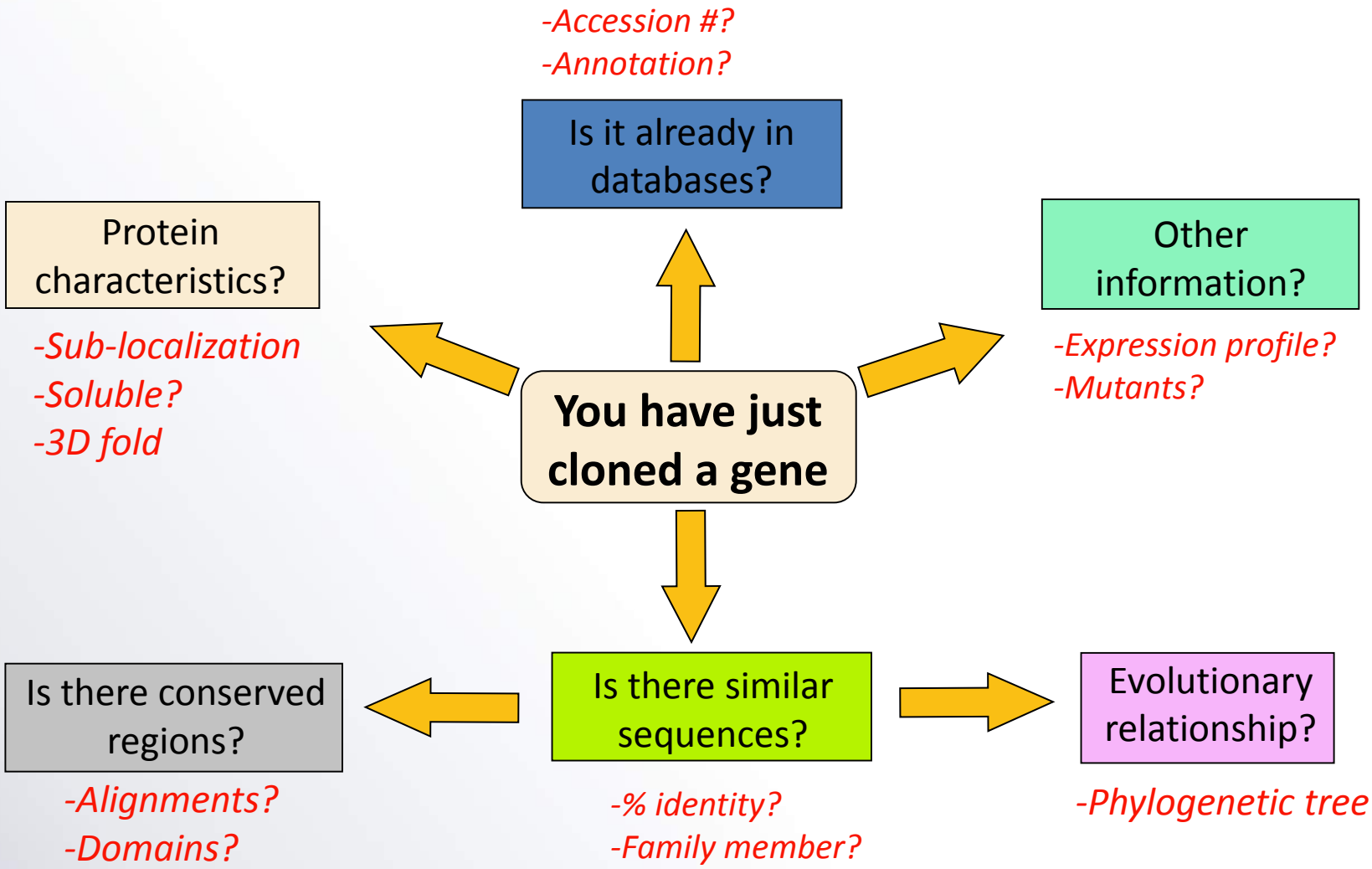
A phenotype...

A gene that controls development

Do you still feel dizzy?

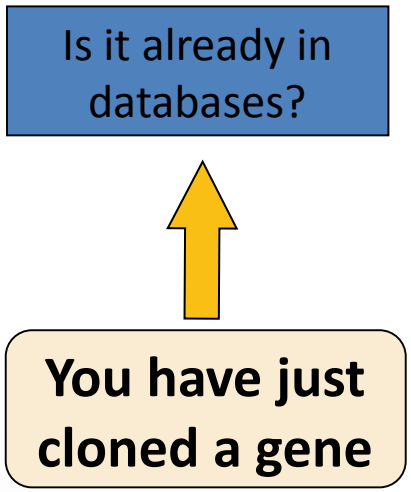
vexedmuddler.artstation.com

Analysis of genomics data



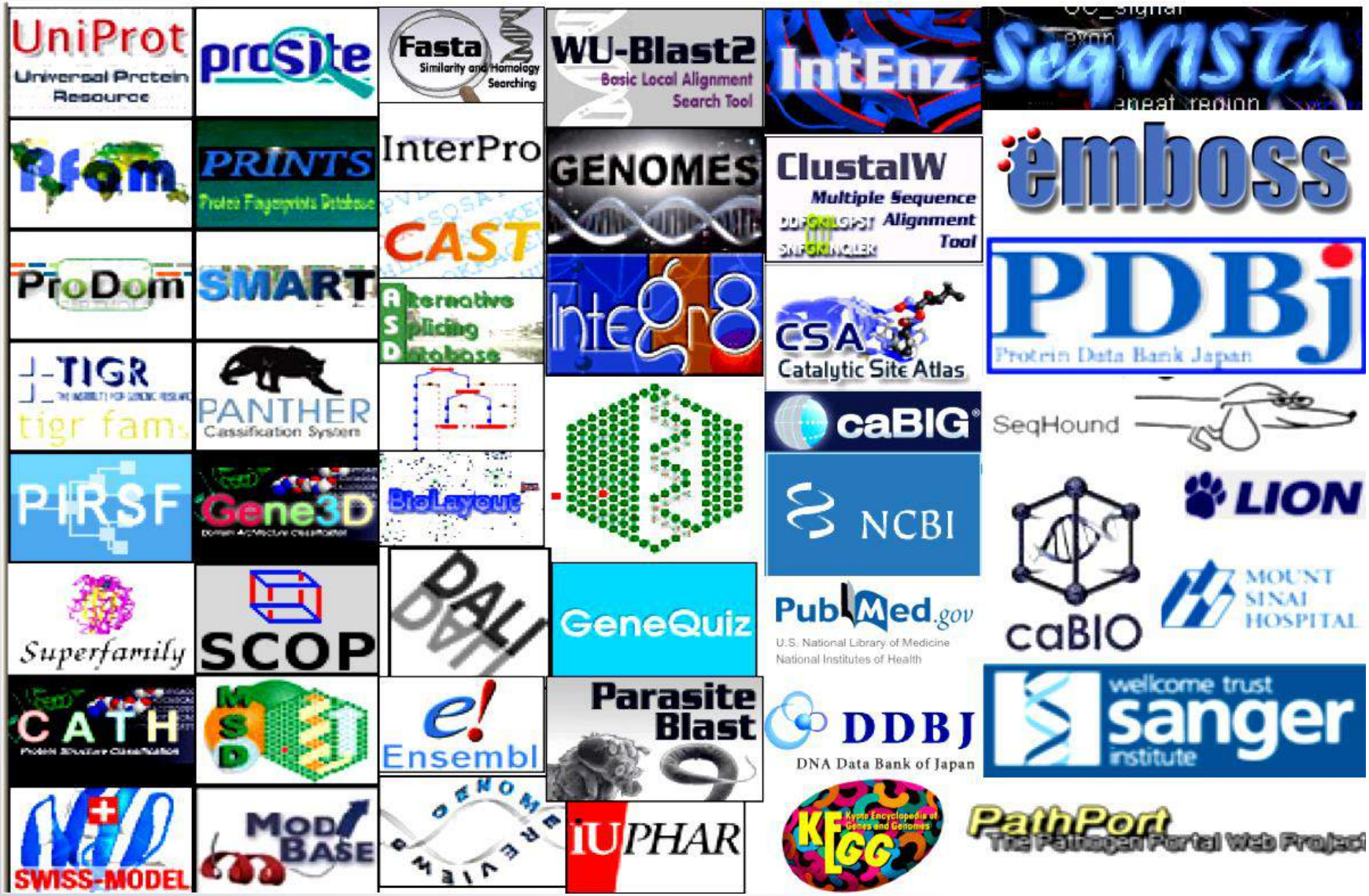
The steps of genomic data analysis

-Accession #?
-Annotation?



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…**

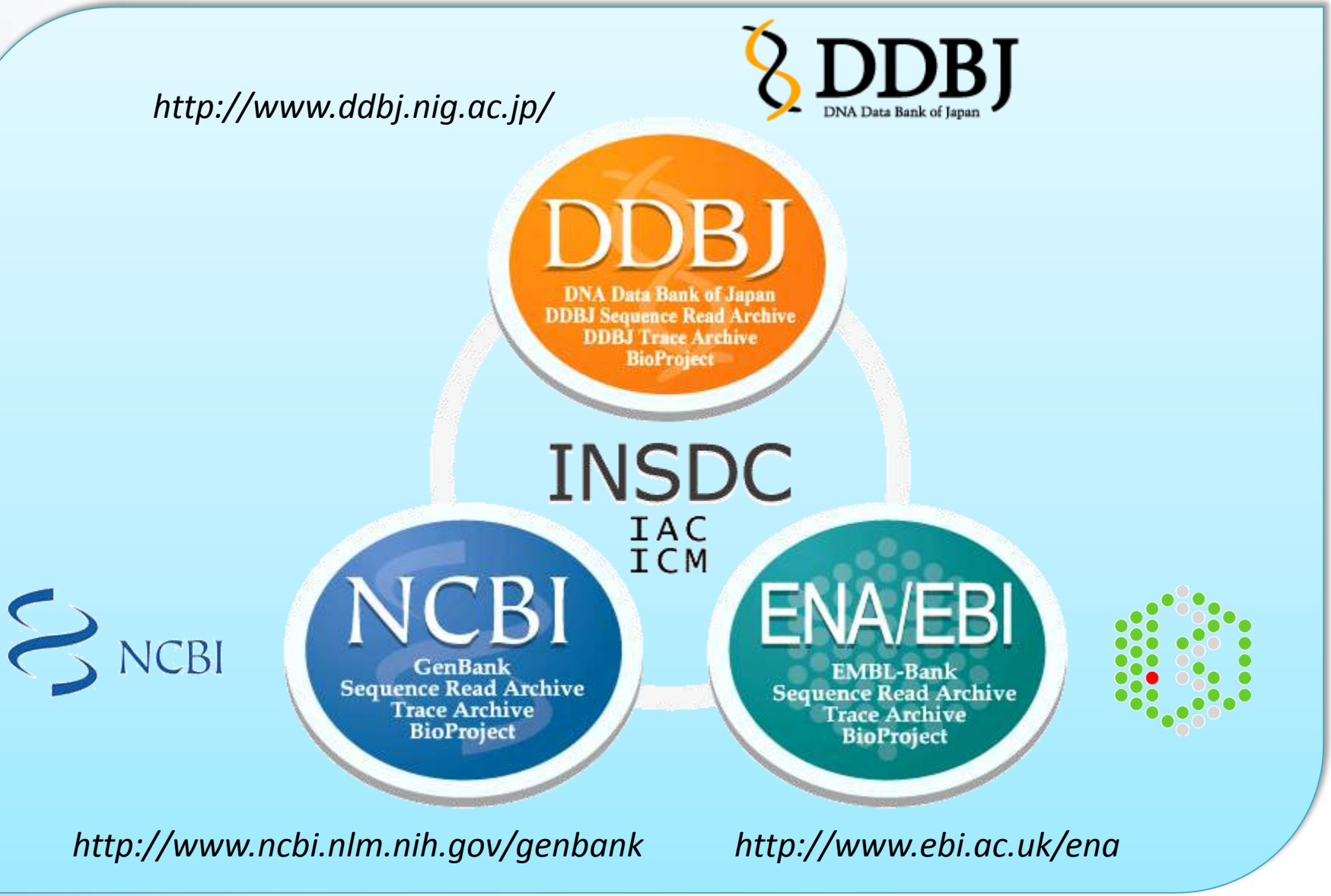
ESTs (Expressed Sequence Tags)

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

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

...many more

INSDC: International Nucleotide Sequence Database Collaboration



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:**

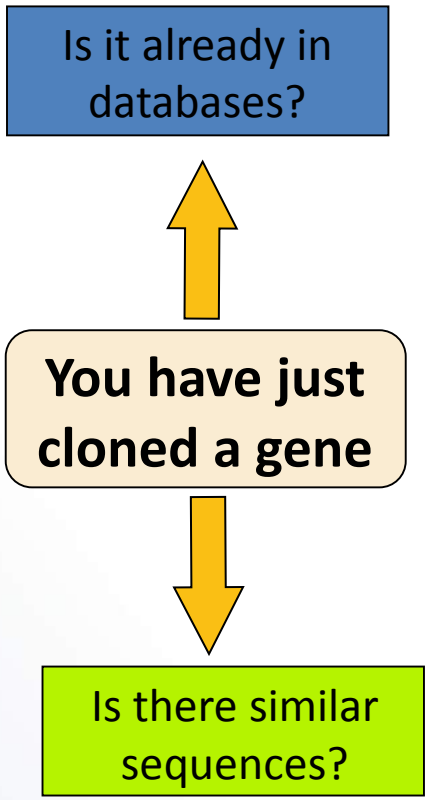
TREMBL (translated EMBL): includes entries that have not been annotated yet into Swiss-Prot.
http://www.ebi.ac.uk/trembl/access.html

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?

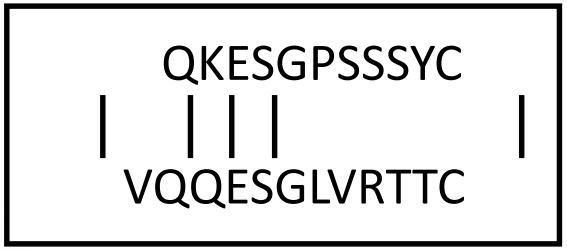


-% identity?
-Family member?

Database search method: Alignment

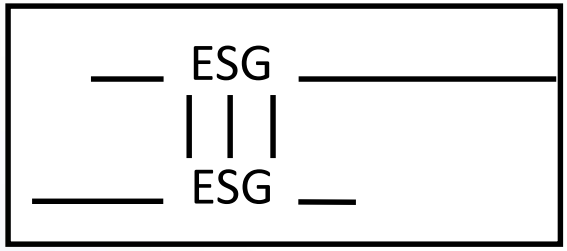
- Two broad classes of sequence alignments

Global alignment:



each residue, not sensitive

Local alignment:

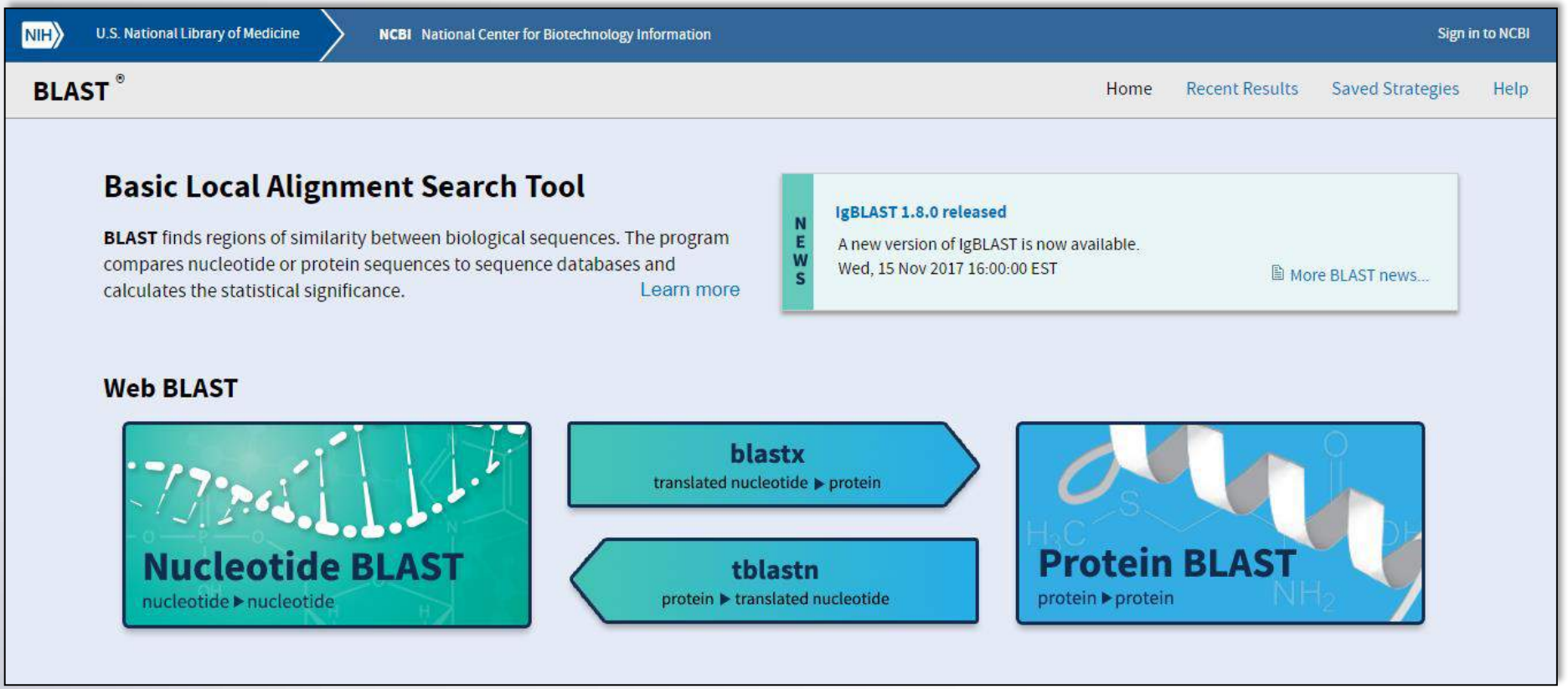


group of region, faster

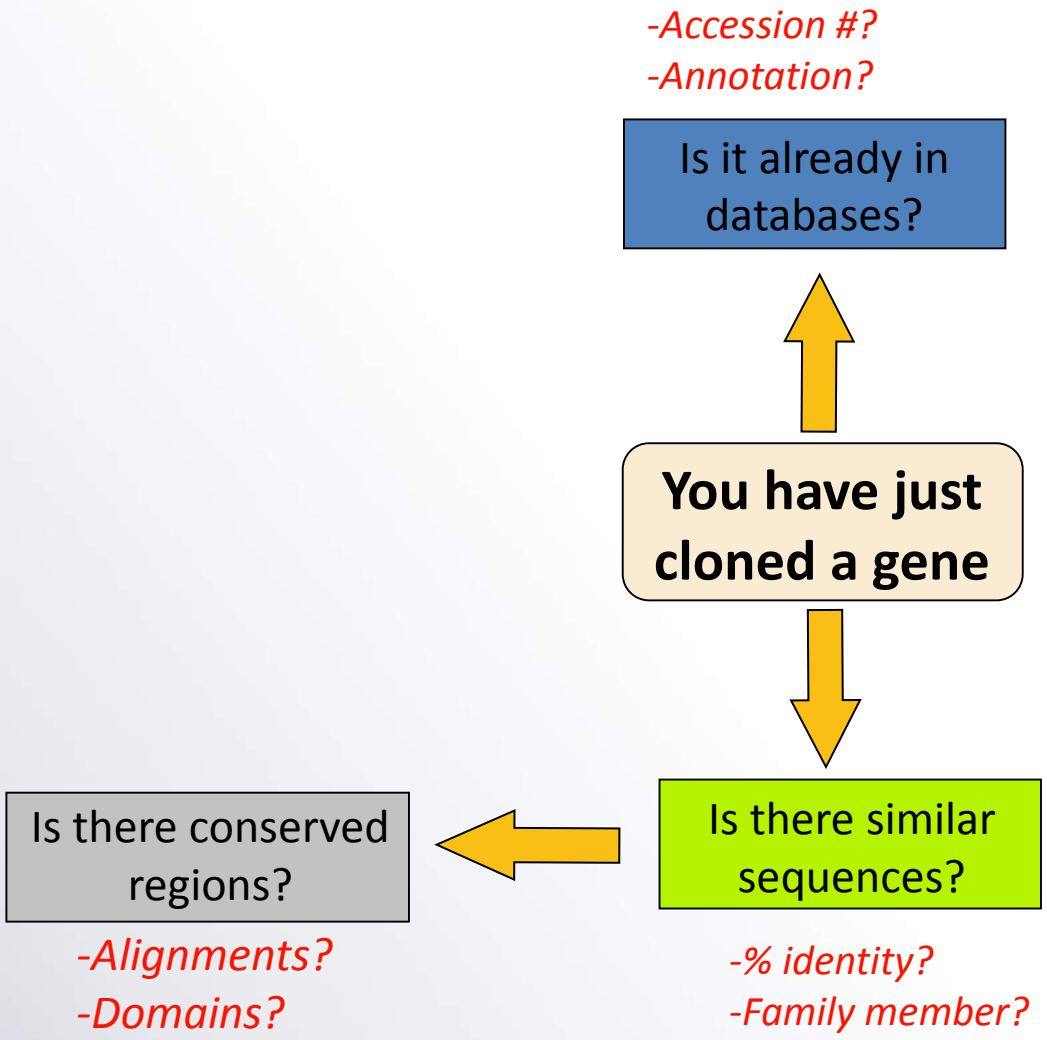
- The most widely used **local similarity algorithm**
  - ✓ Smith-Waterman http://www.ebi.ac.uk/MPsrch/
  - ✓ **Basic Local Alignment Search Tool (BLAST)** https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - ✓ Fast Alignment (FASTA) http://fasta.genome.jp; http://www.ebi.ac.uk/fasta33/; http://www.arabidopsis.org/cgi-bin/fasta/nph-TAIRfasta.pl

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



Analysis of genomics data



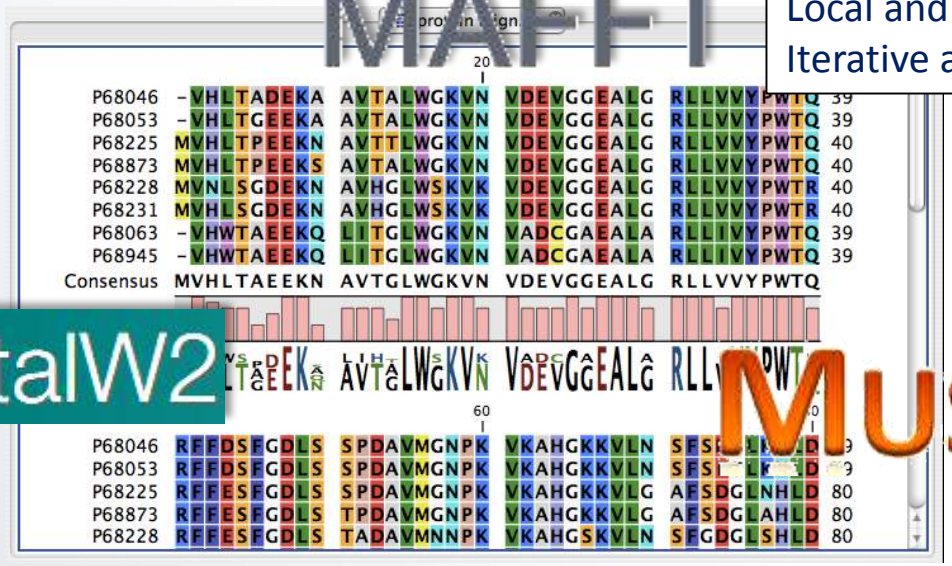
Multiple Sequence Alignment (MSA)

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

MAFFT

Local and global alignment
Iterative algorithm Fast-Fourier

Global alignment
Gaps

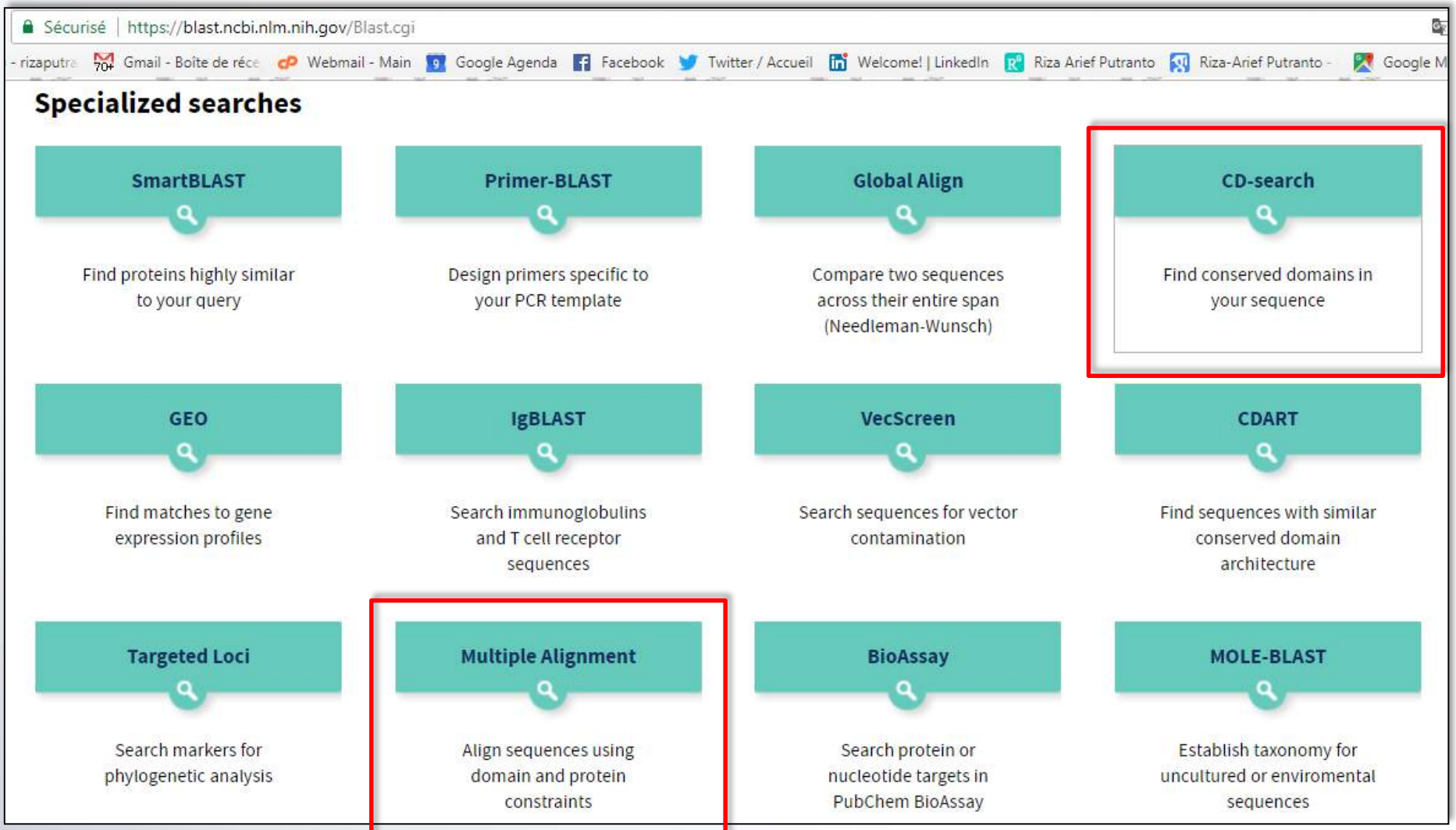


ClustalW2

MUSCLE

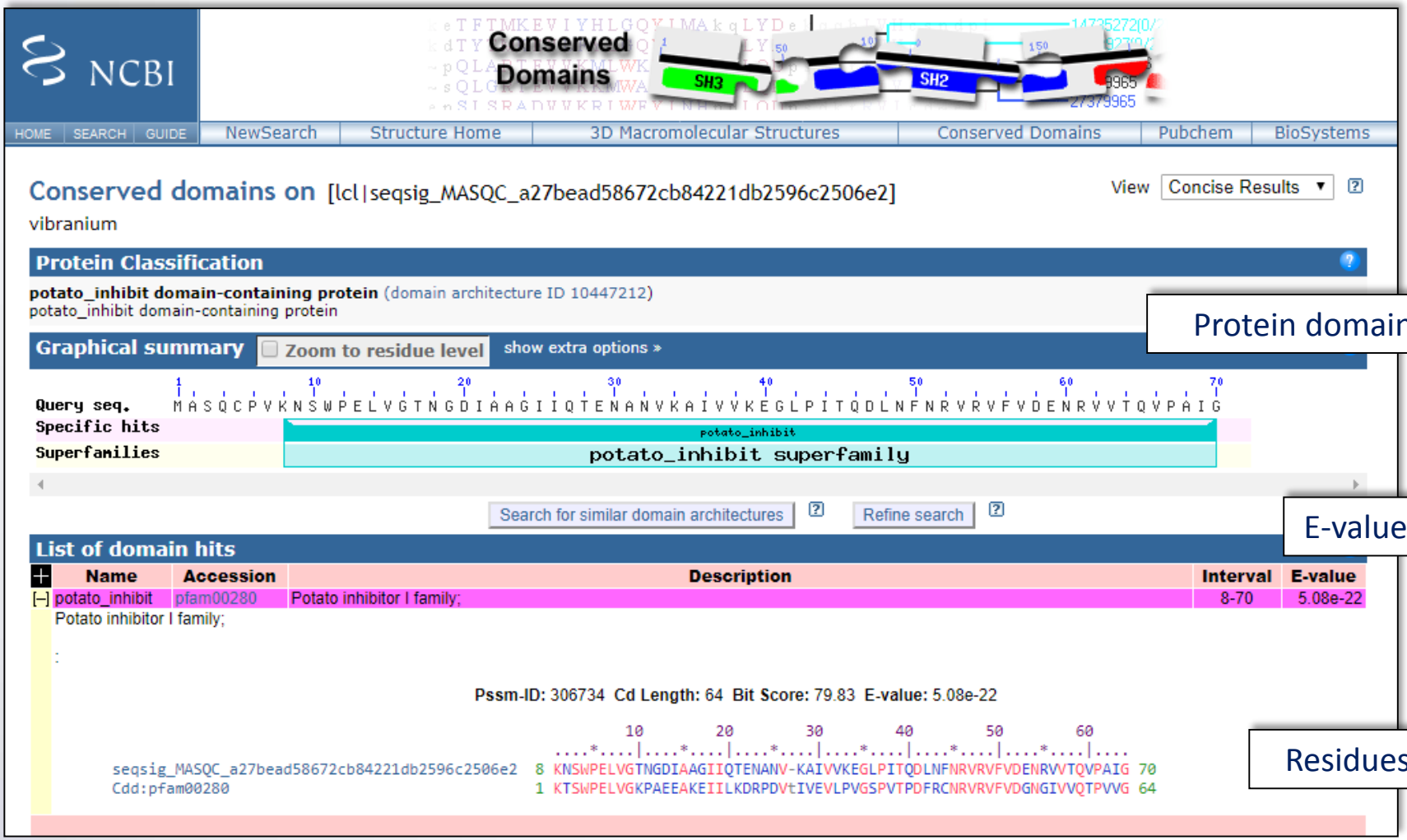
Local and global alignment
Iterative algorithm

Other tools in NCBI: CD-search, MA



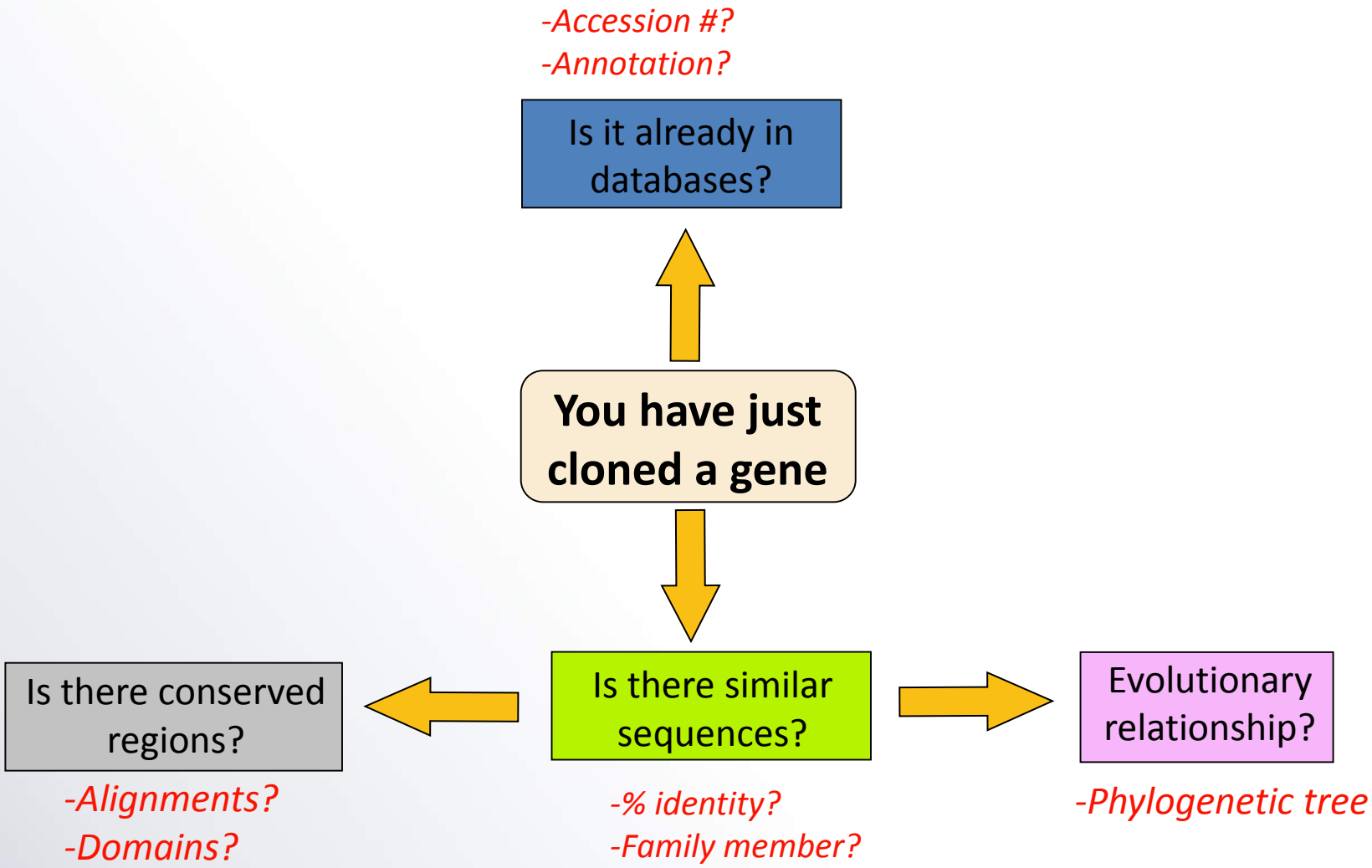
https://blast.ncbi.nlm.nih.gov/Blast.cgi

Domain search of a primary protein



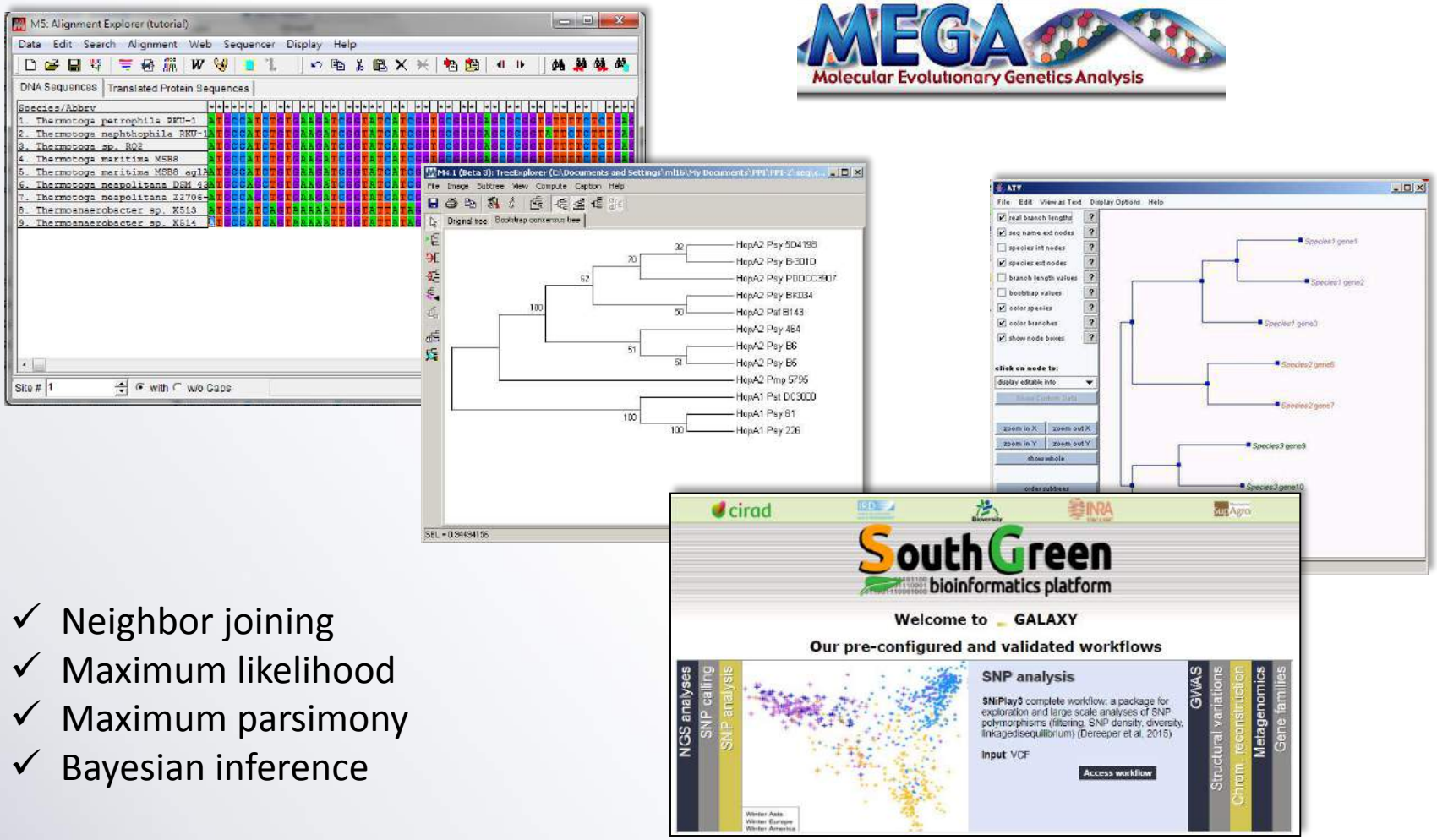
https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Analysis of genomics data



Many softwares and algorithms for phylo

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

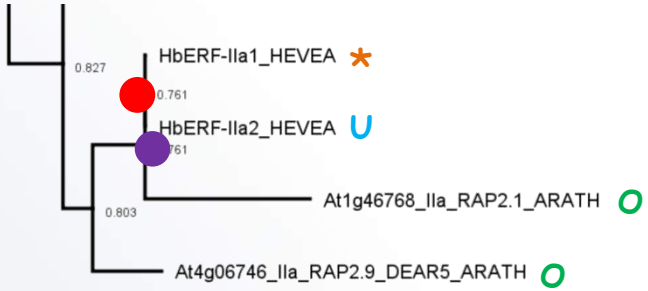


- ✓ Neighbor joining
- ✓ Maximum likelihood
- ✓ Maximum parsimony
- ✓ Bayesian inference

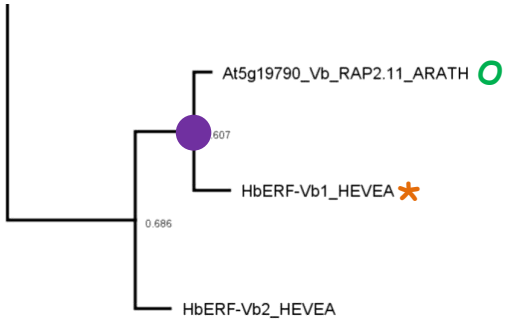
Different phylogenetic algorithms

- **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 method 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 **the 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 computational expensive. However, with the today's computers this is not too much of a problem.

Phylo tree for evolutionary relationship



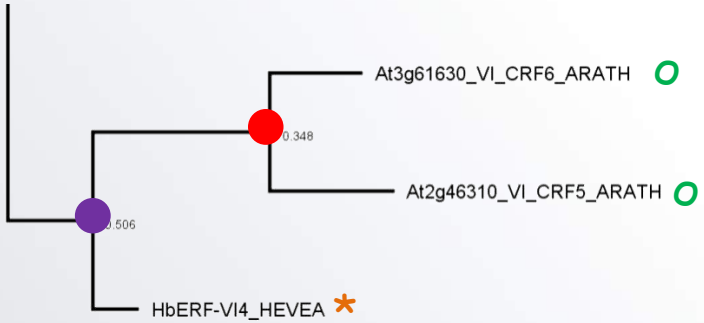
HbERF-II



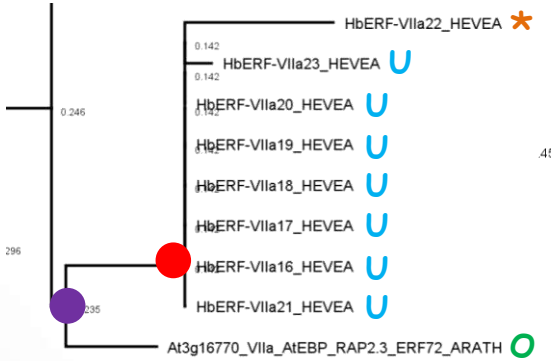
HbERF-V

Identifikasi ortolog dan paralog dari famili gen Ethylene Response Factors

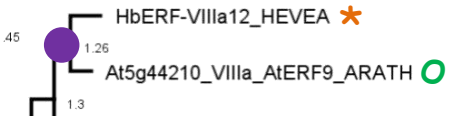
Hevea brasiliensis vs Arabidopsis thaliana



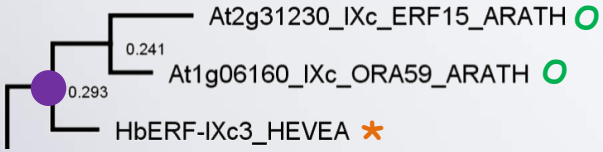
HbERF-VI



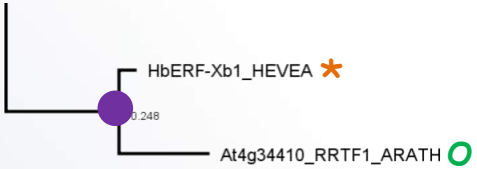
HbERF-VII



HbERF-VIII



HbERF-IX

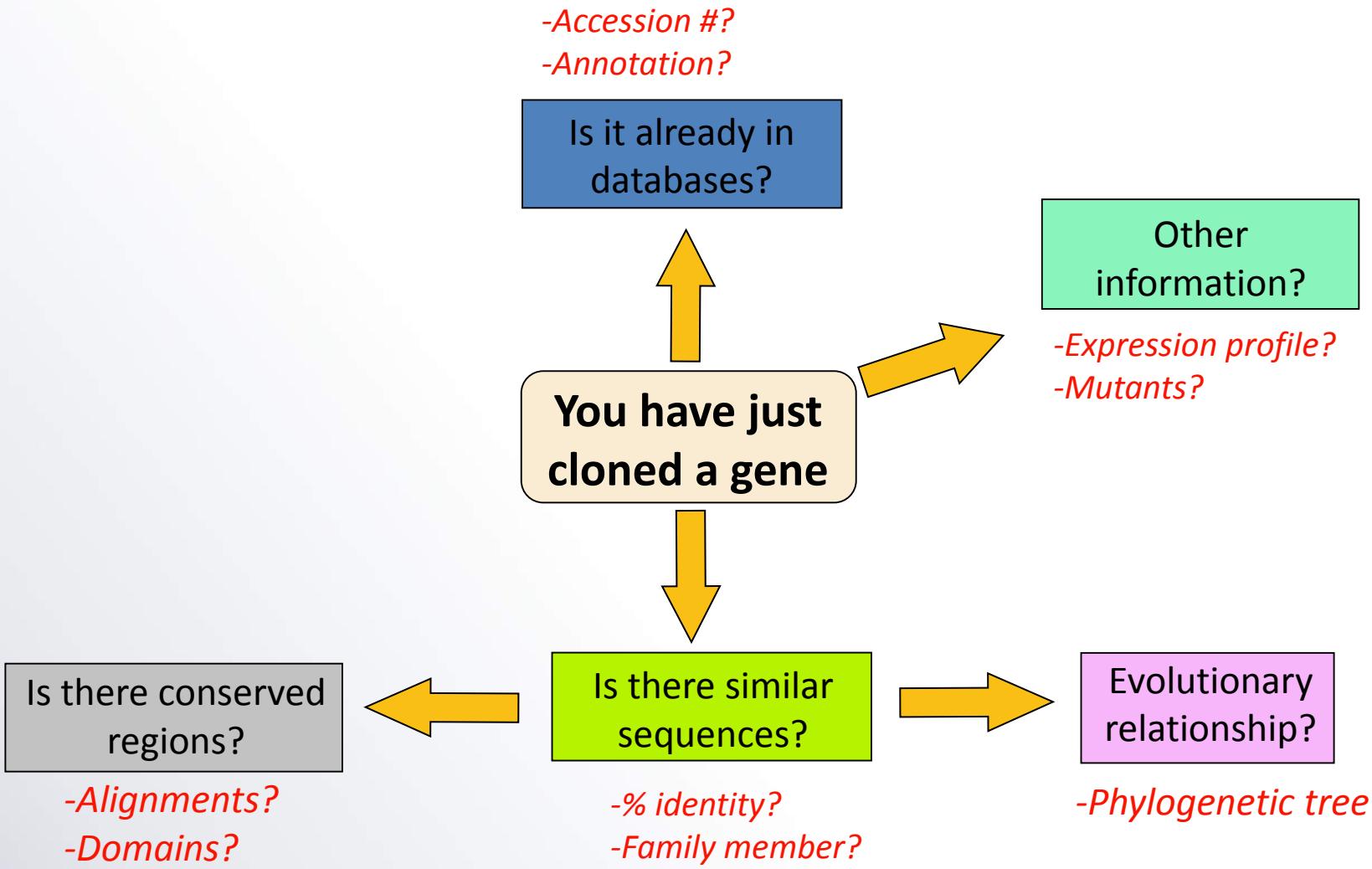


HbERF-X

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

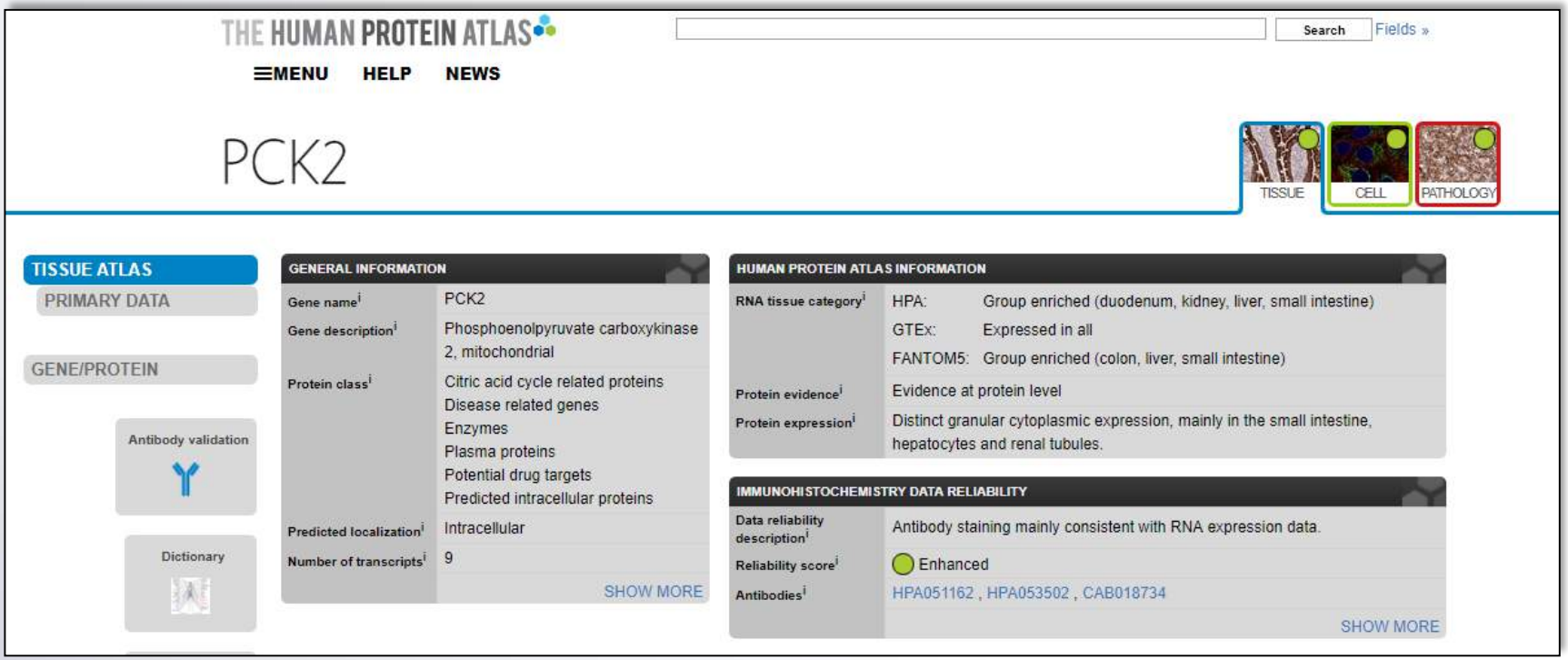
Putranto *et al* 2015 *PLoS ONE,* 10, e0123618.

Analysis of genomics data



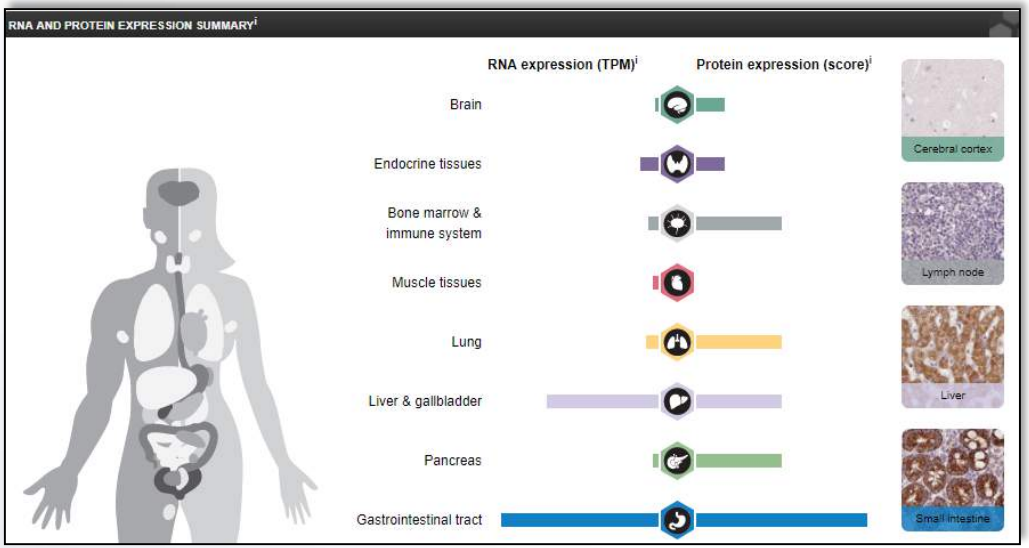
Expression data on open source databases

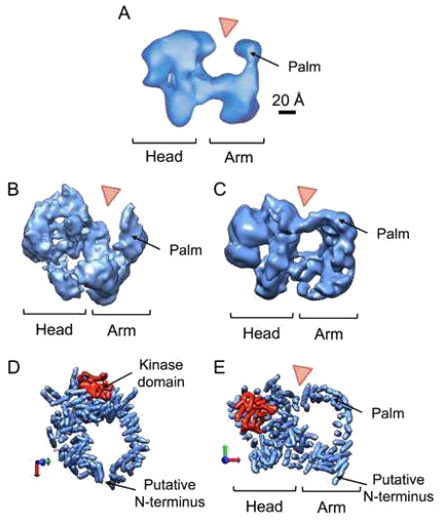
The Human Protein Atlas is a Swedish-based program initiated in 2003 with the aim to **map all the human proteins in cells, tissues and organs** using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

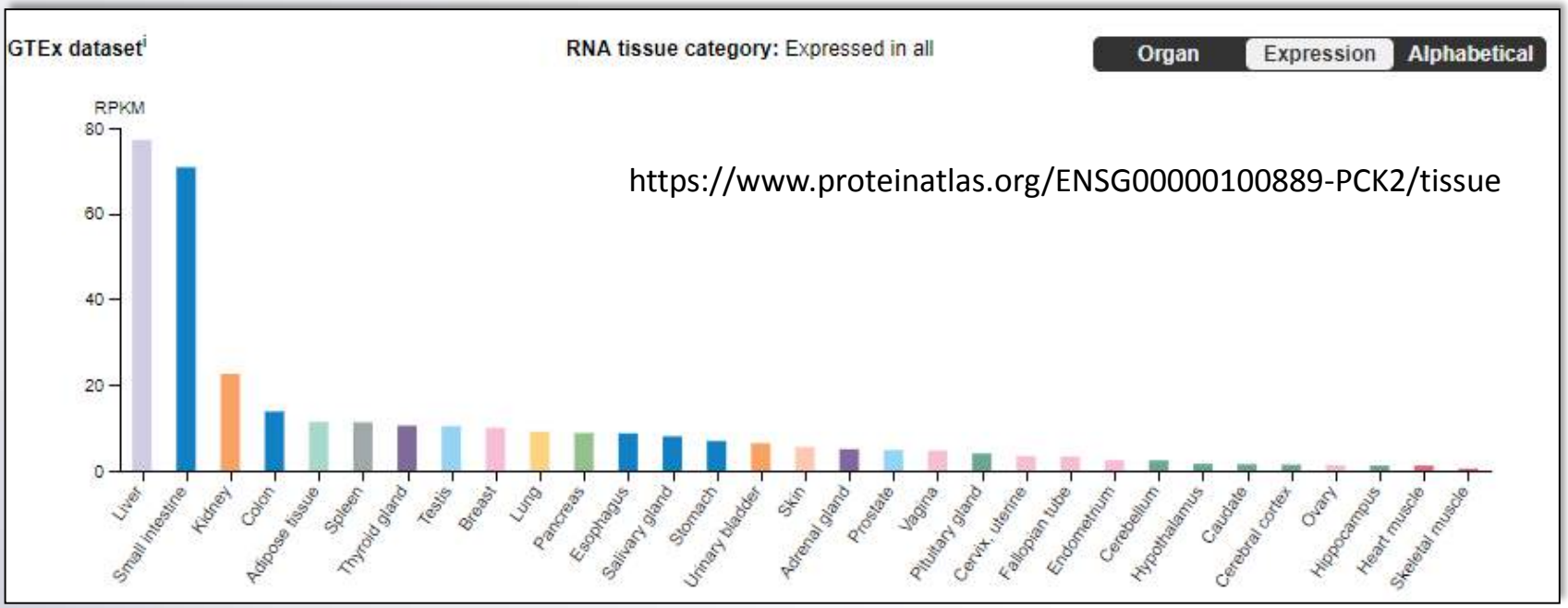


https://www.proteinatlas.org/about

Example of PKC2 human protein



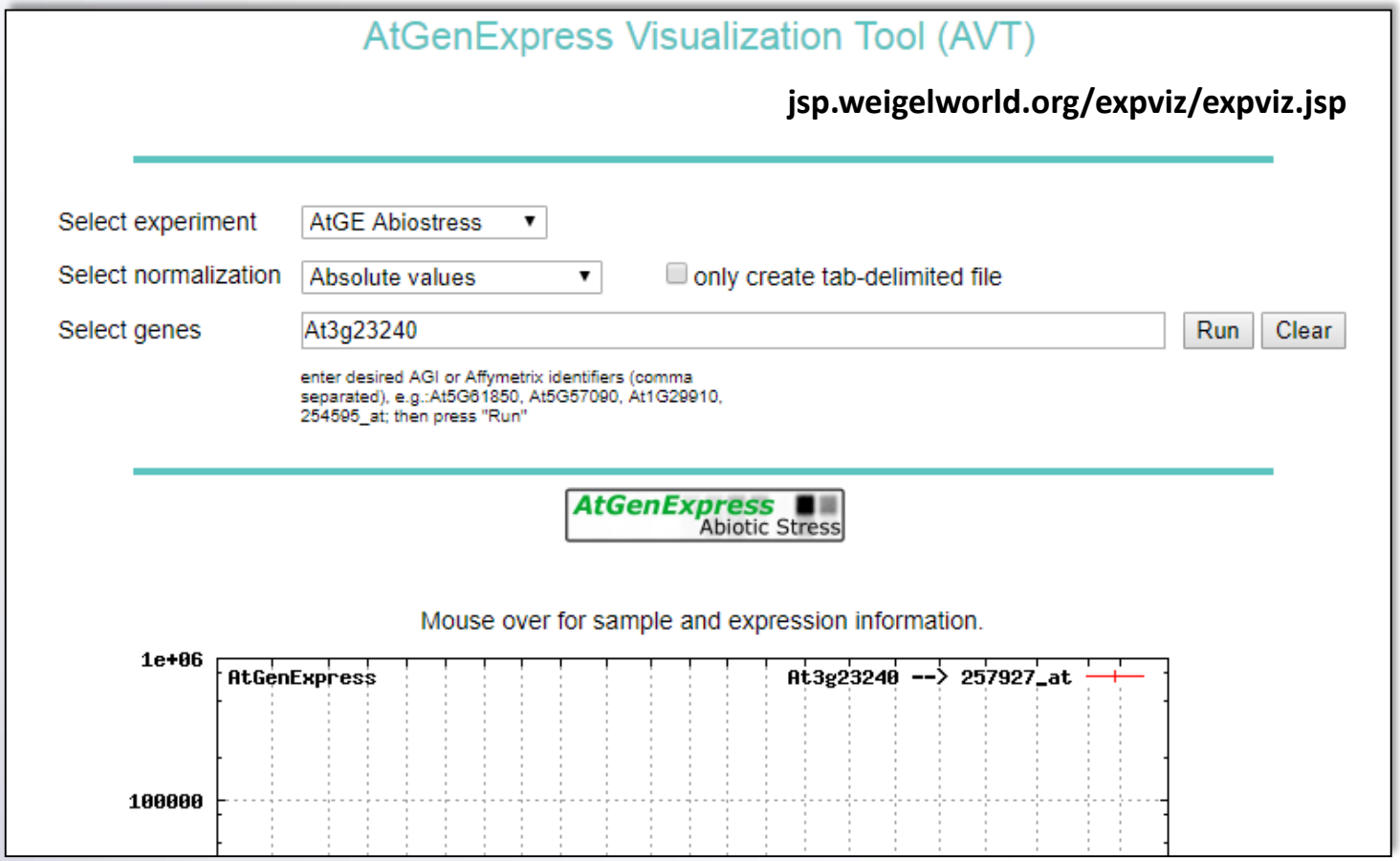




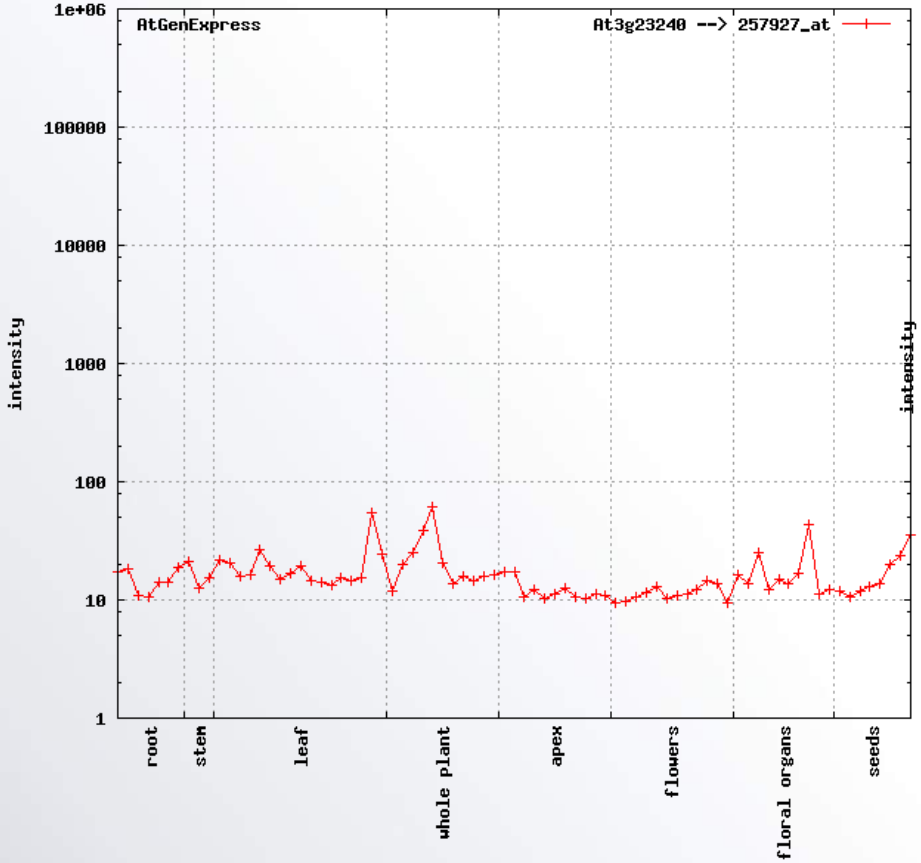
https://www.proteinatlas.org/ENSG00000100889-PCK2/tissue

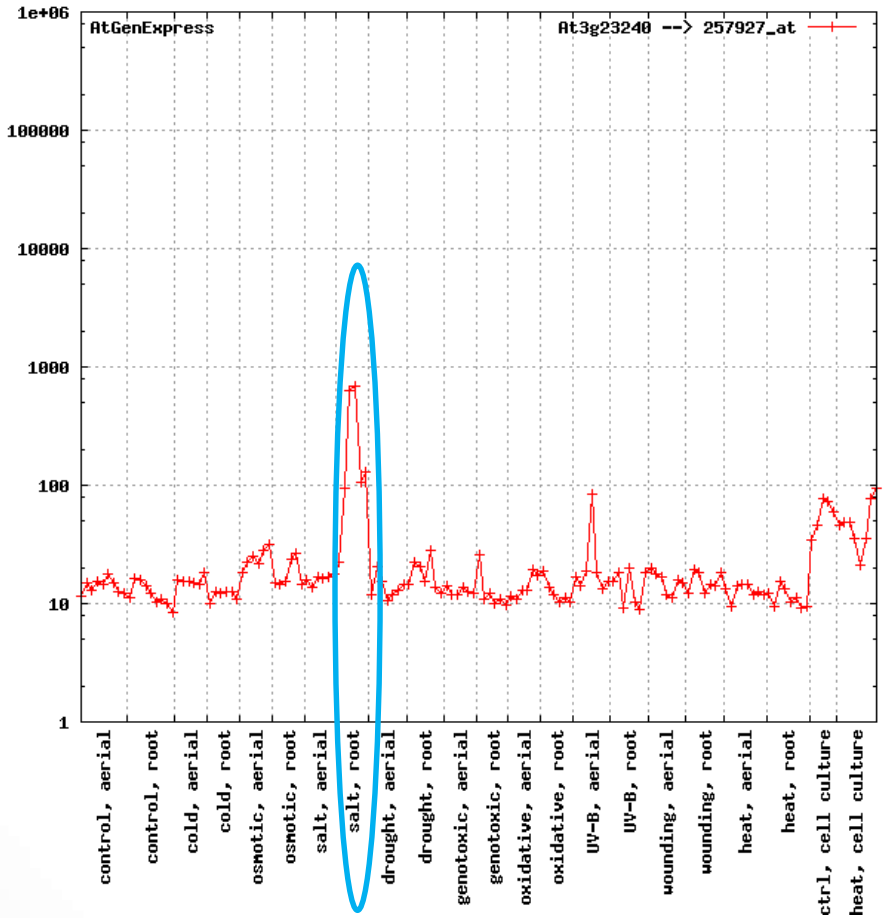
Expression data on open source databases

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



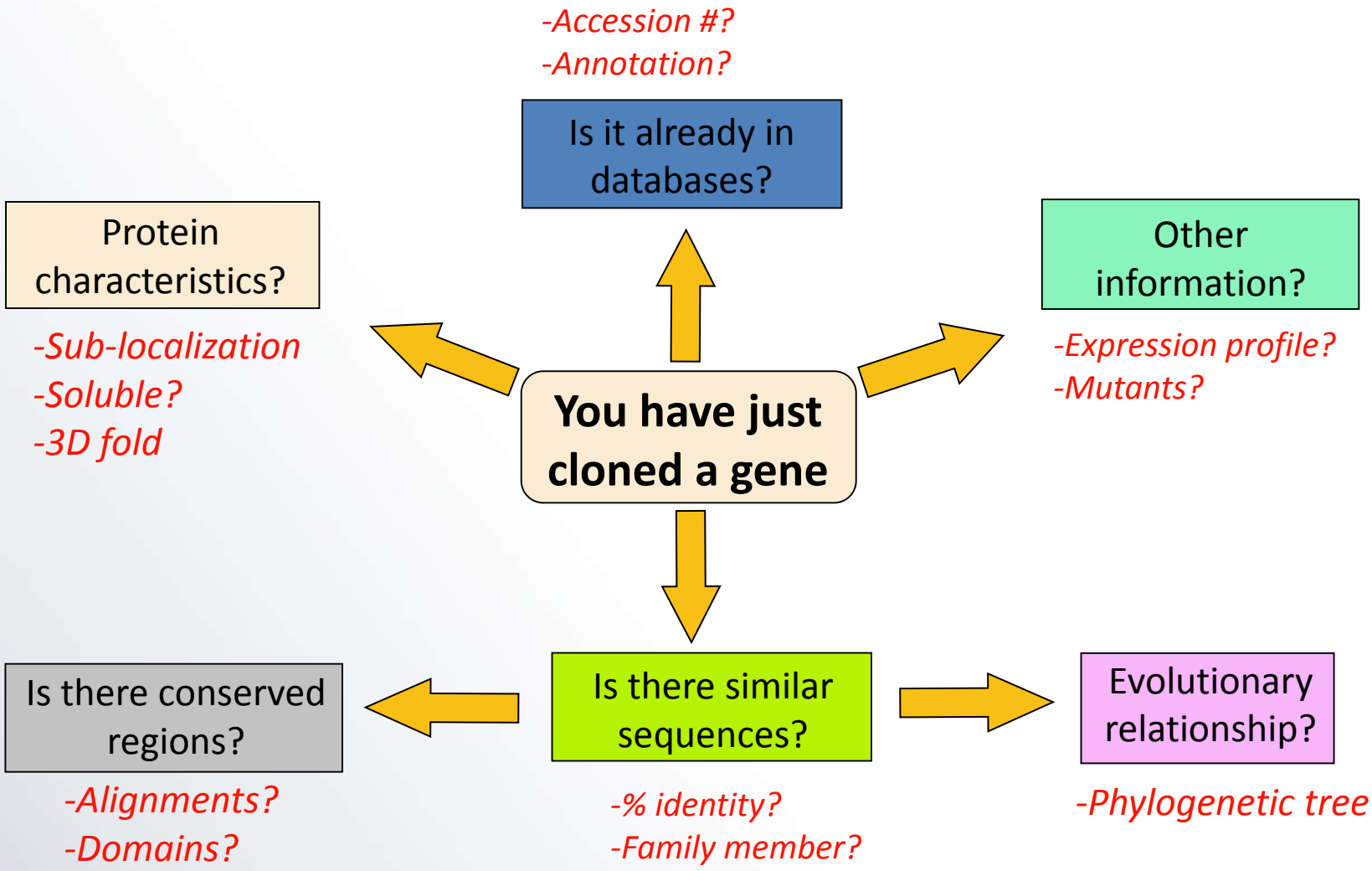


At3g23240 or **AtERF1** encodes a member of the ERF (**ethylene response factor**) subfamily B-3 of ERF/AP2 transcription factor family (ERF1). The protein contains one AP2 domain.

AtGenExpress Abiotic Stress

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

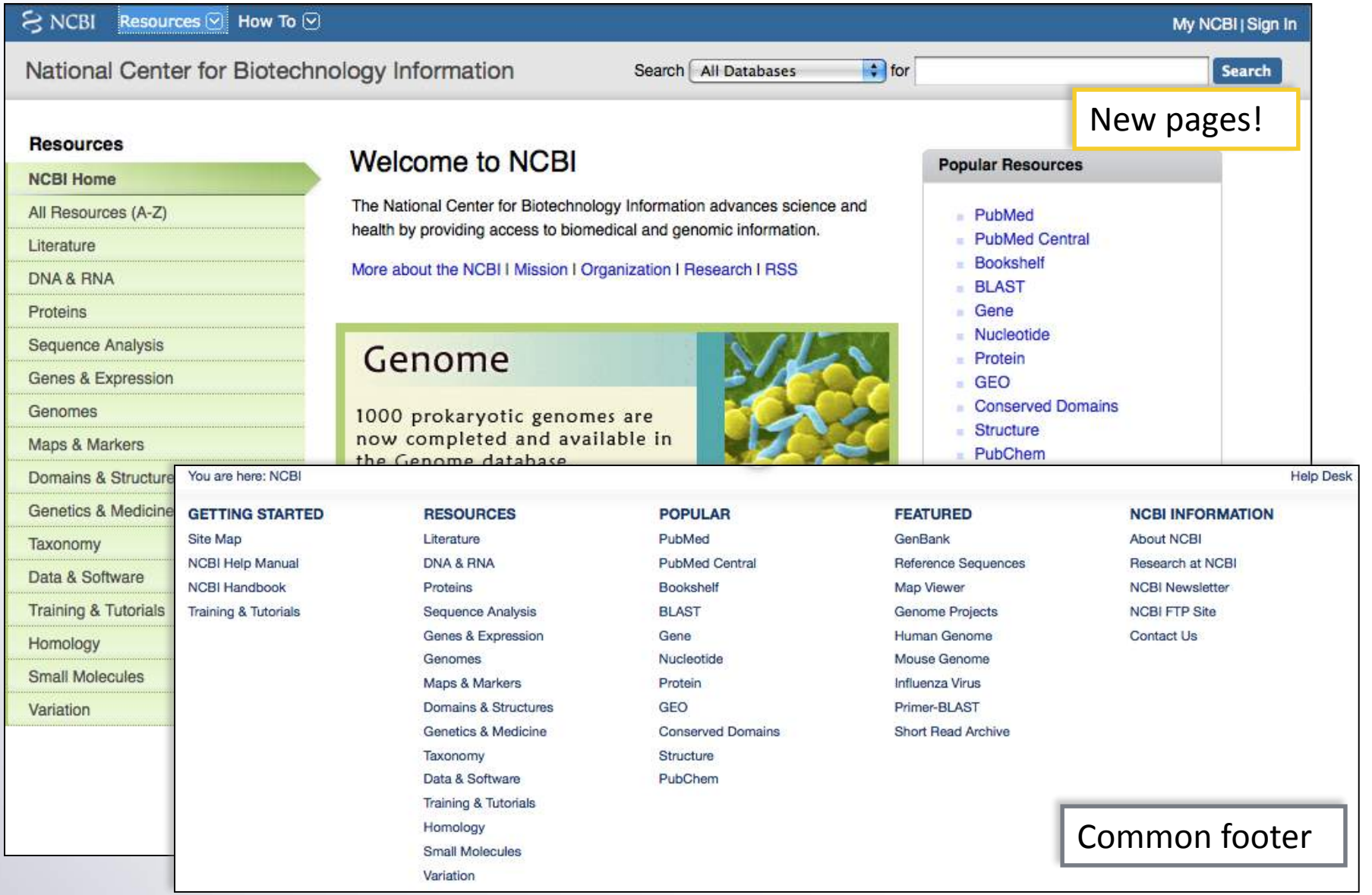
Genome Data Viewer (GDV)

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



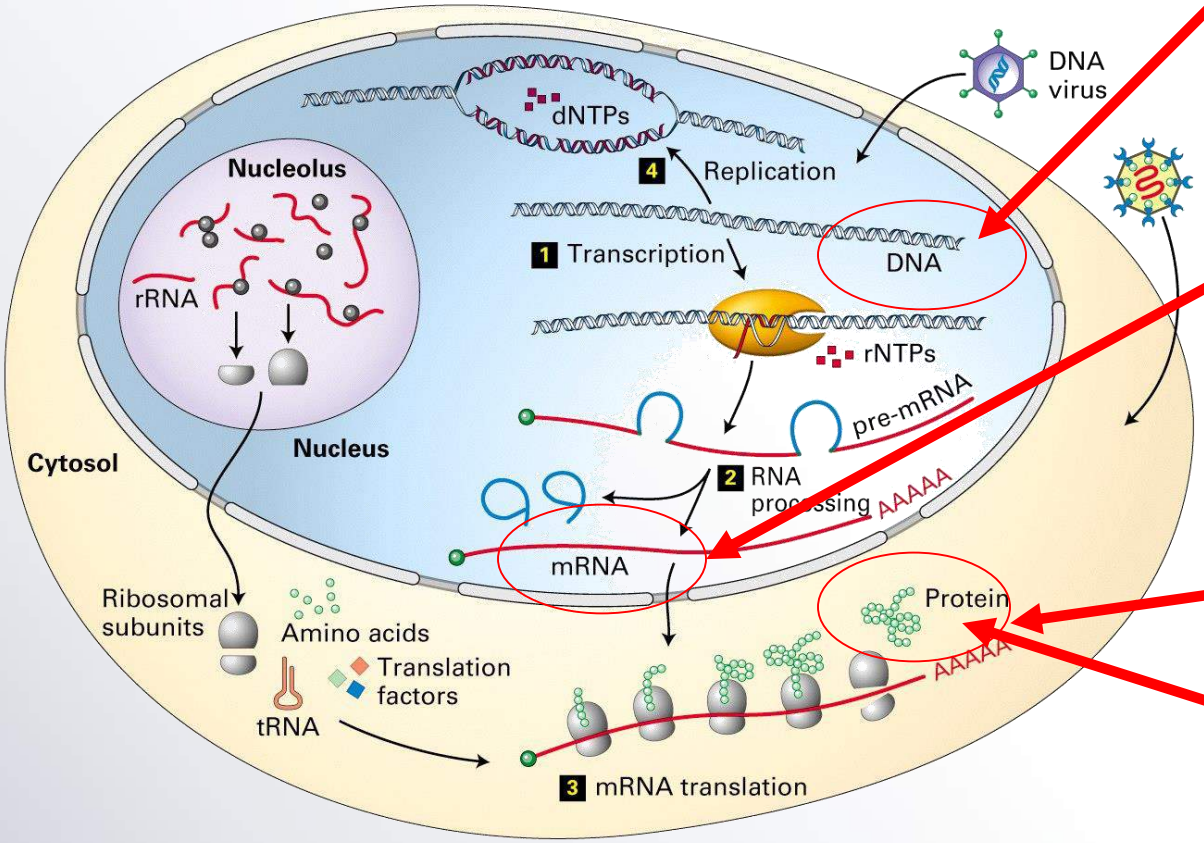
What are databases?

- **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



Original DNA Sequences (Genomes)

Expressed DNA sequences (mRNA & cDNA sequences)
Expressed Sequence Tags (ESTs)

Protein sequences
-Inferred
-Direct sequencing

Protein structures
-Experiments
-Models (homologues)

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

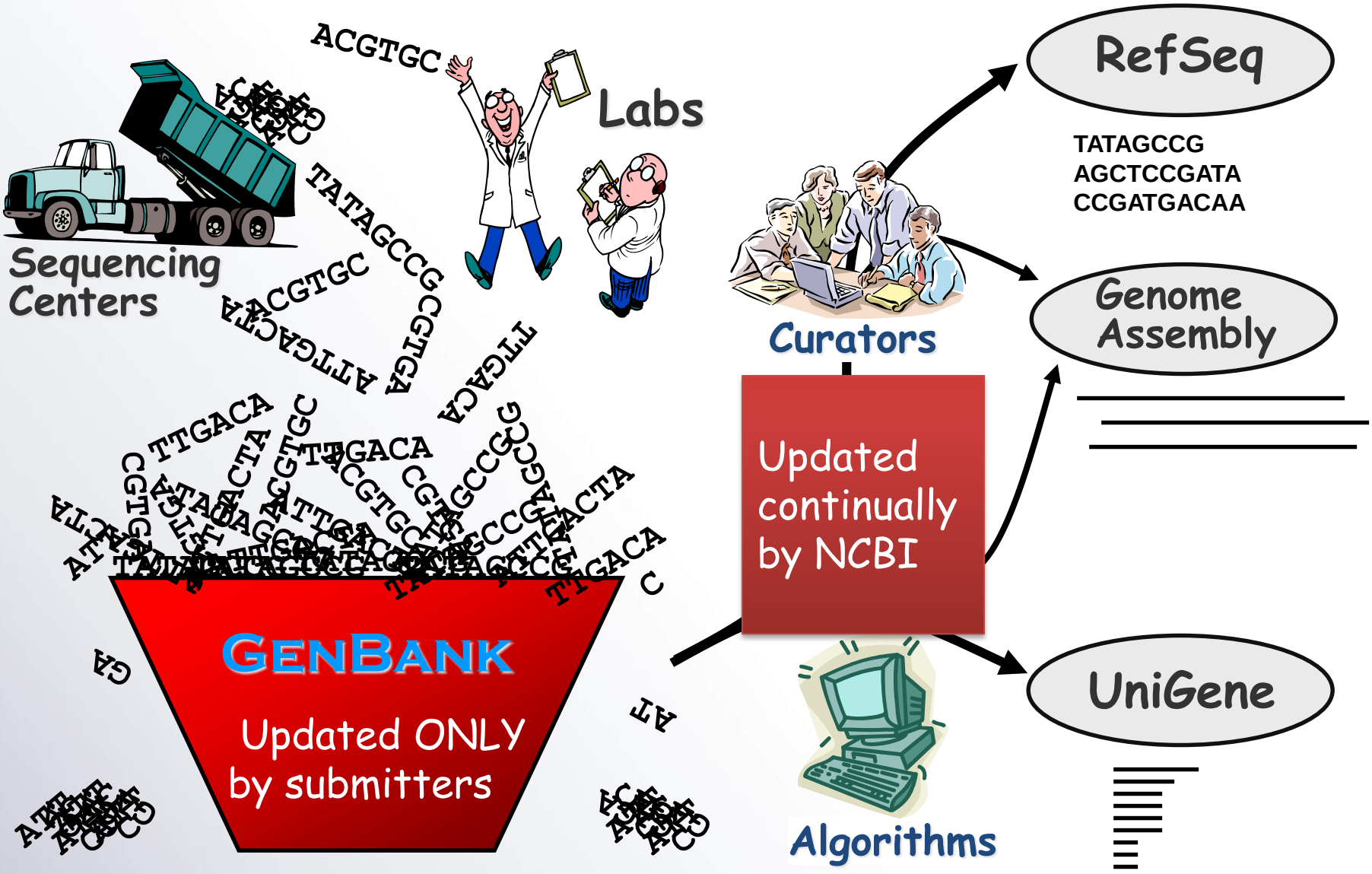
- **Primary** Databases
  - ✓ **Original submissions** by experimentalists
  - ✓ Content controlled by the **submitter**

    Examples: GenBank, Trace, SRA, SNP, GEO

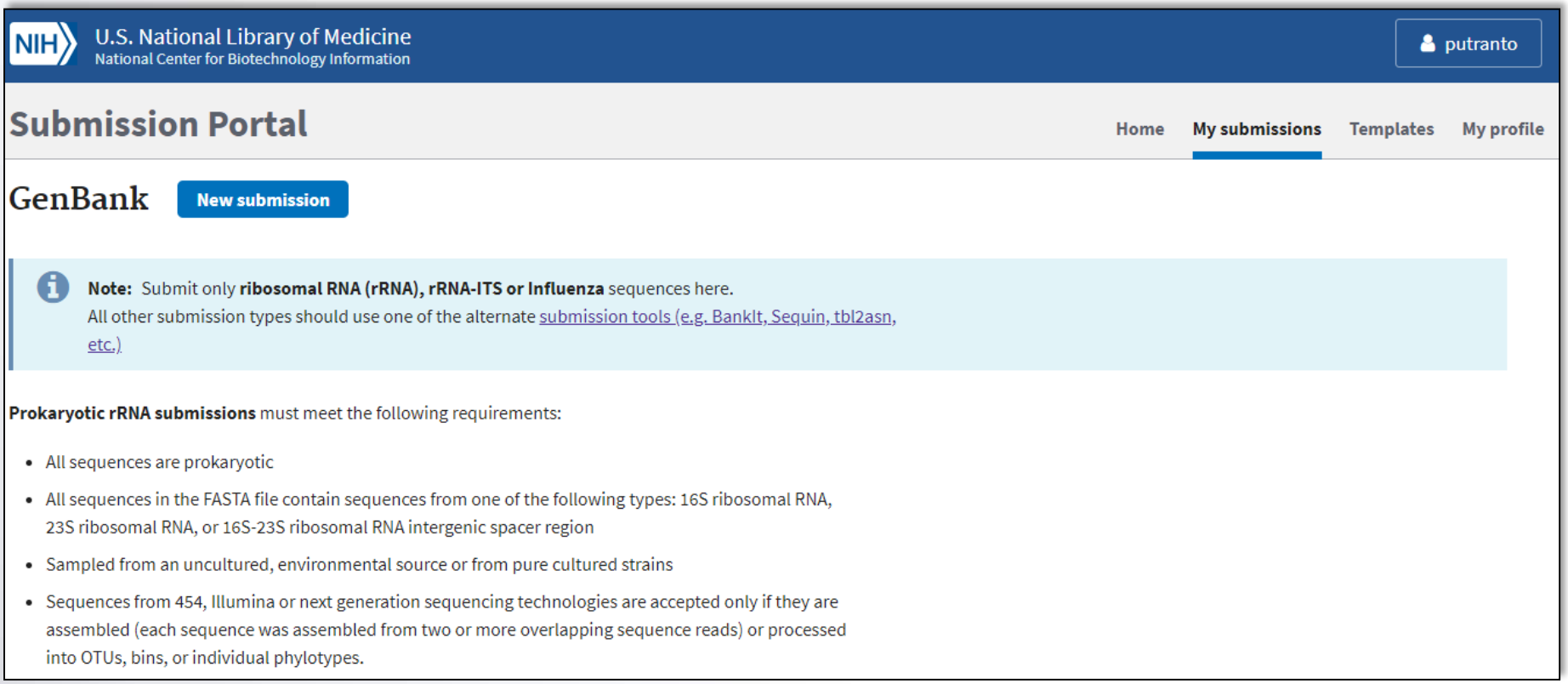
- **Derivative** Databases
  - ✓ **Derived from primary data**
  - ✓ Content controlled by **third party** (NCBI)

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

Primary sequence submission



Primary sequence submission



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

Sequence databases at NCBI

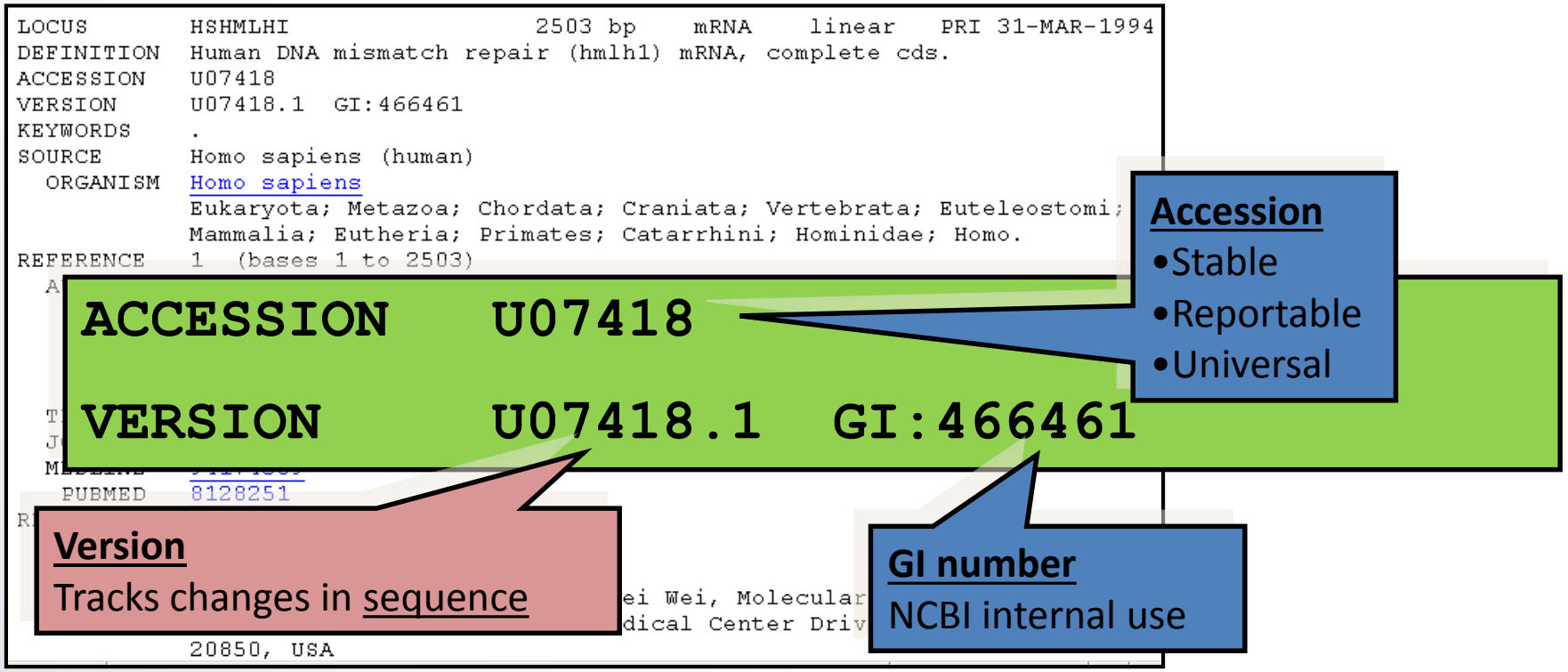
- **Primary**
  - ✓ **GenBank**: NCBI's primary sequence database
  - ✓ **Trace Archive**: reads from capillary sequencers
  - ✓ **Sequence Read Archive**: next generation data

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

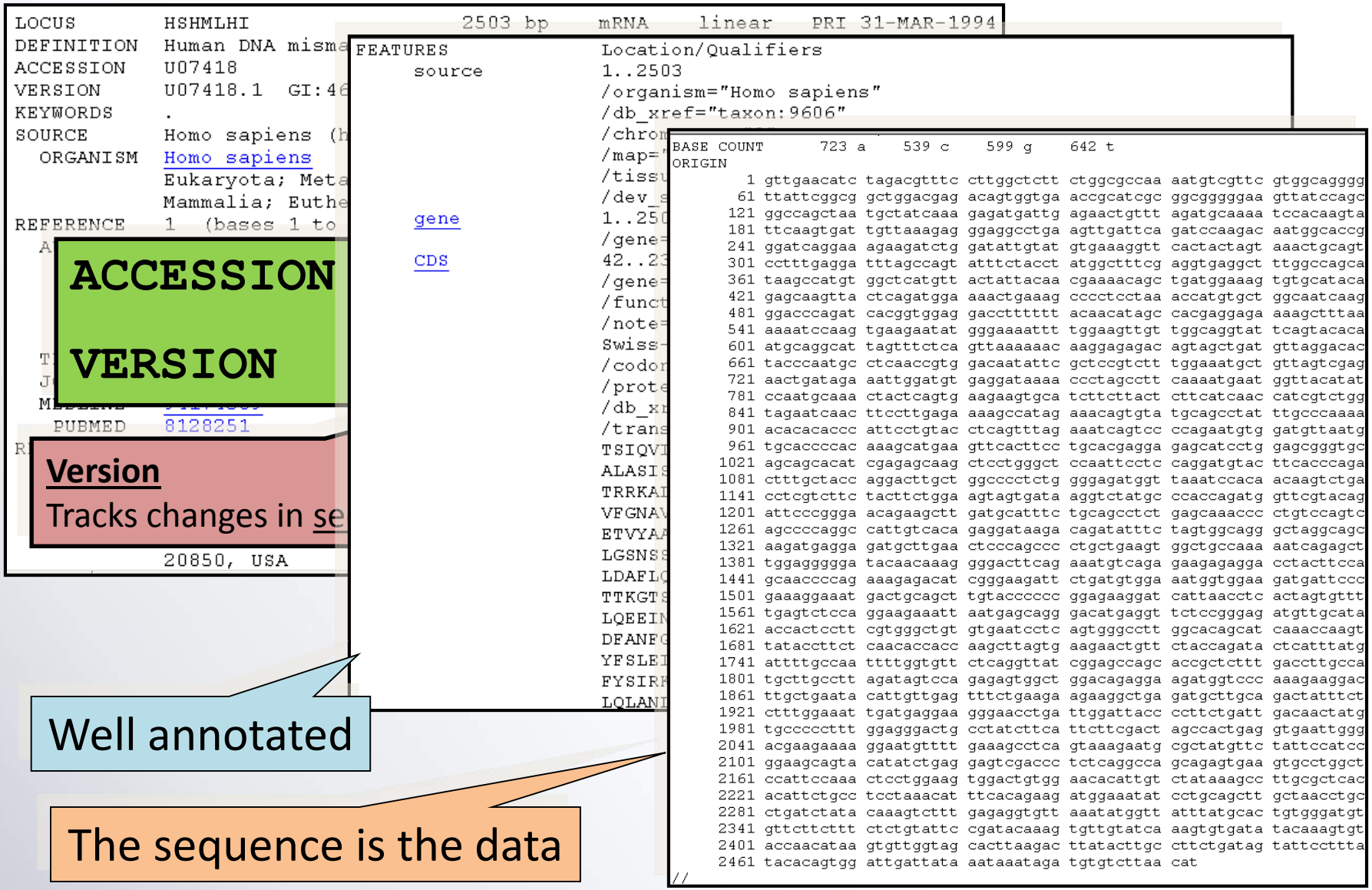
GenBank – Primary sequence database

- **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



Principal GenBank record



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="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
```

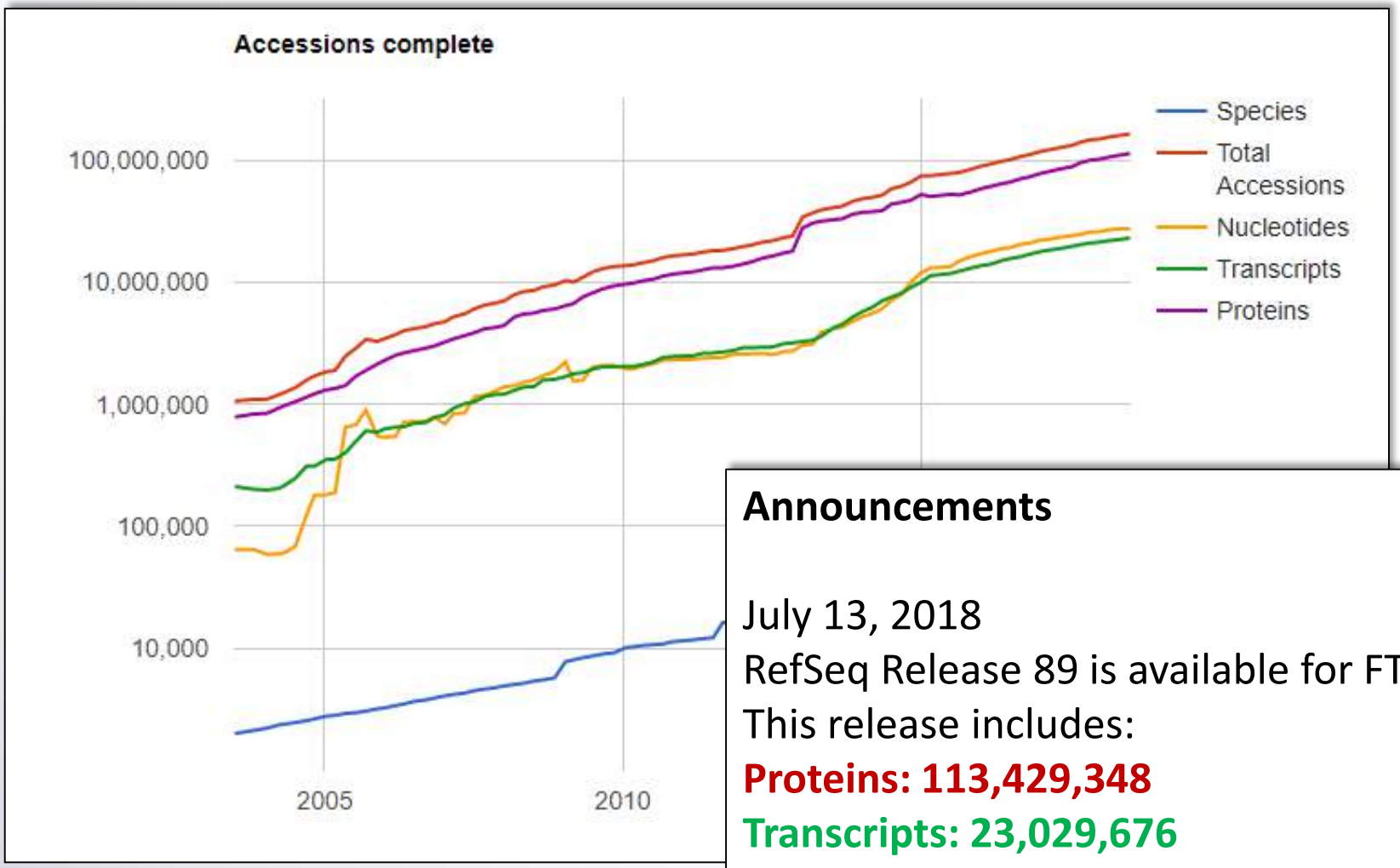
```
>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

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/

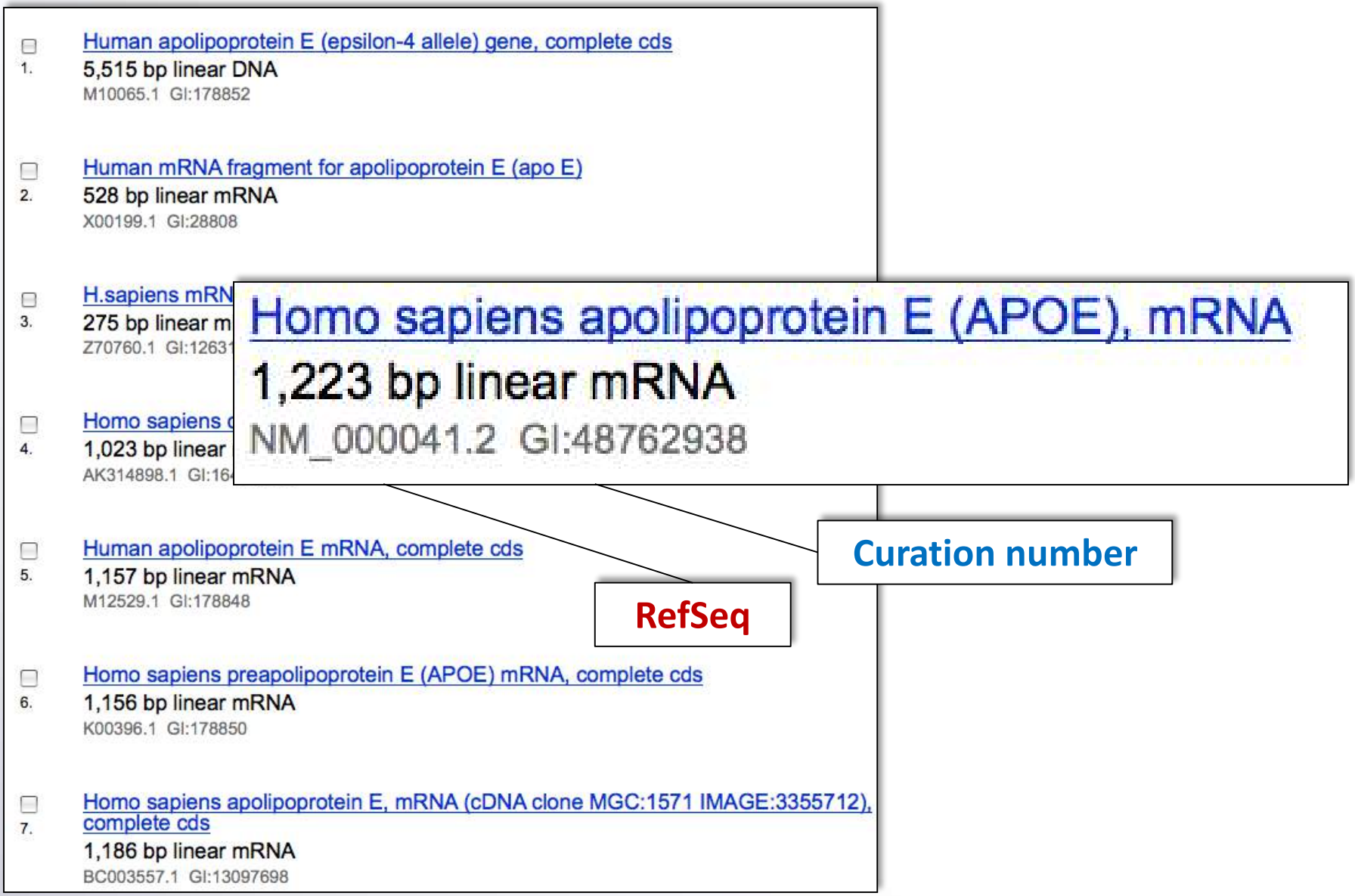
RefSeq: Derivative sequence database



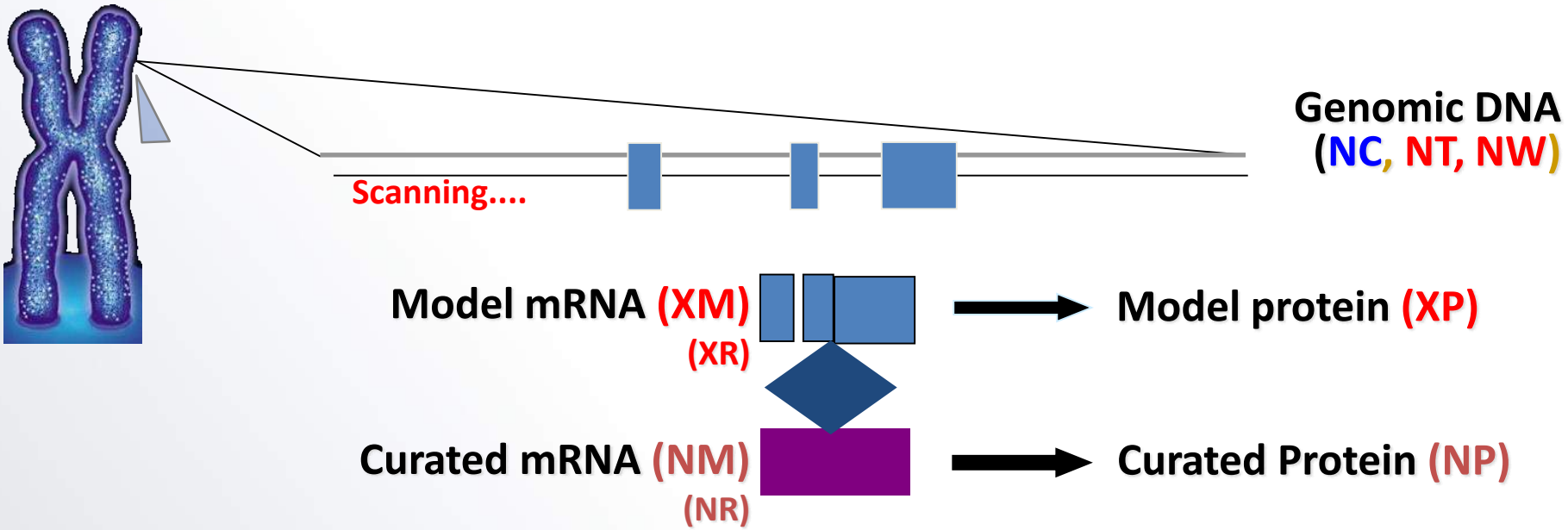
Announcements

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

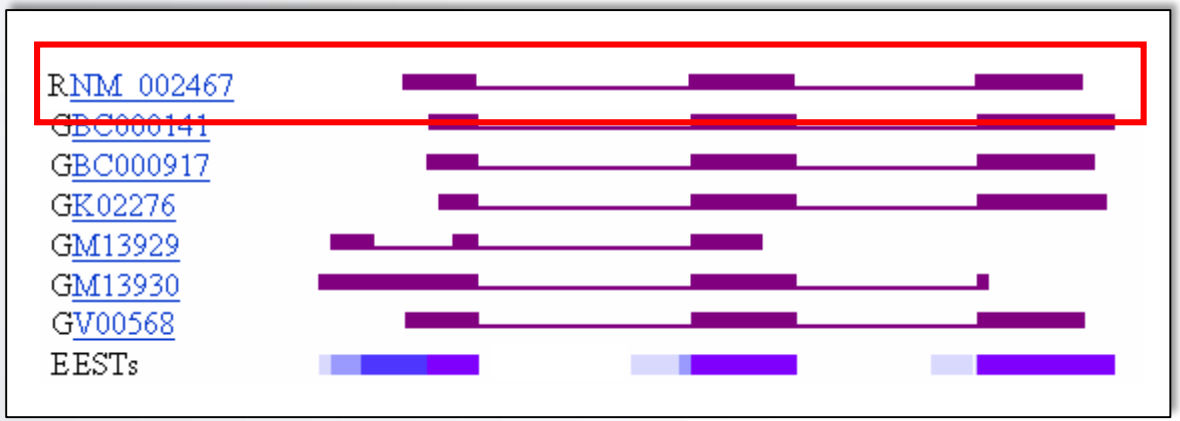


RefSeq: The finalized version of GenBank



RefSeq

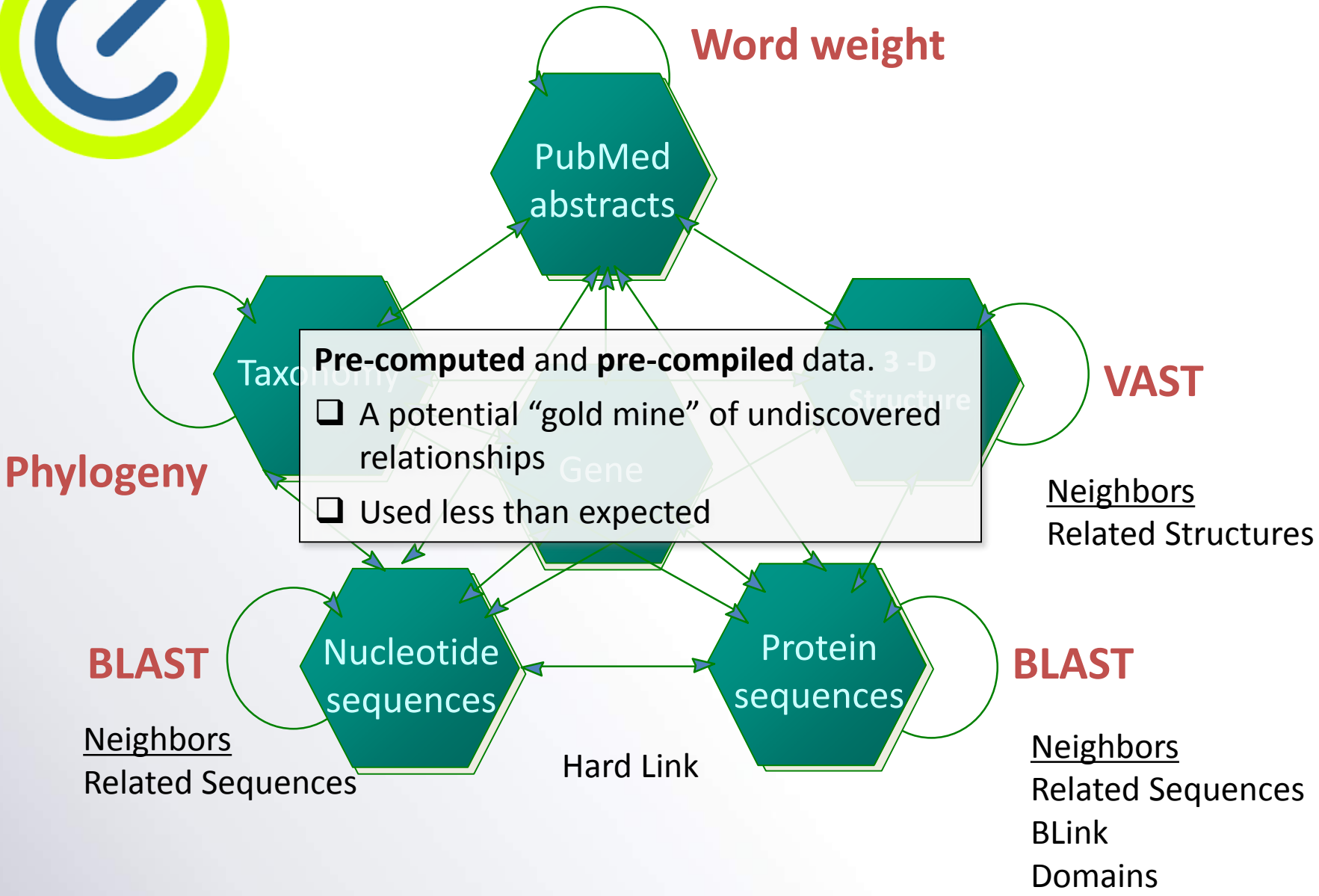
GenBank Sequences



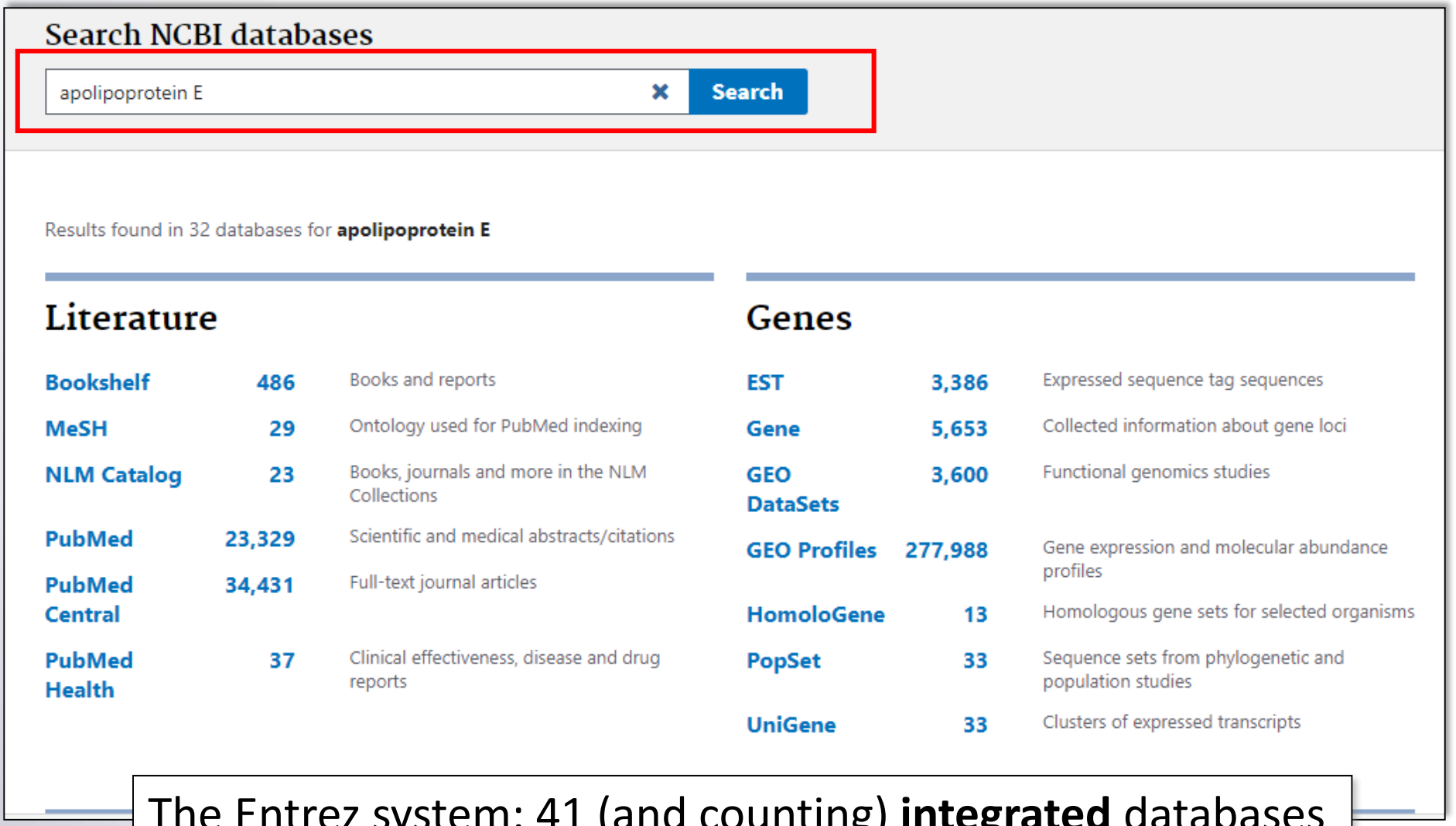
RefSeq benefits

- 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



Entrez: All access to database in NCBI



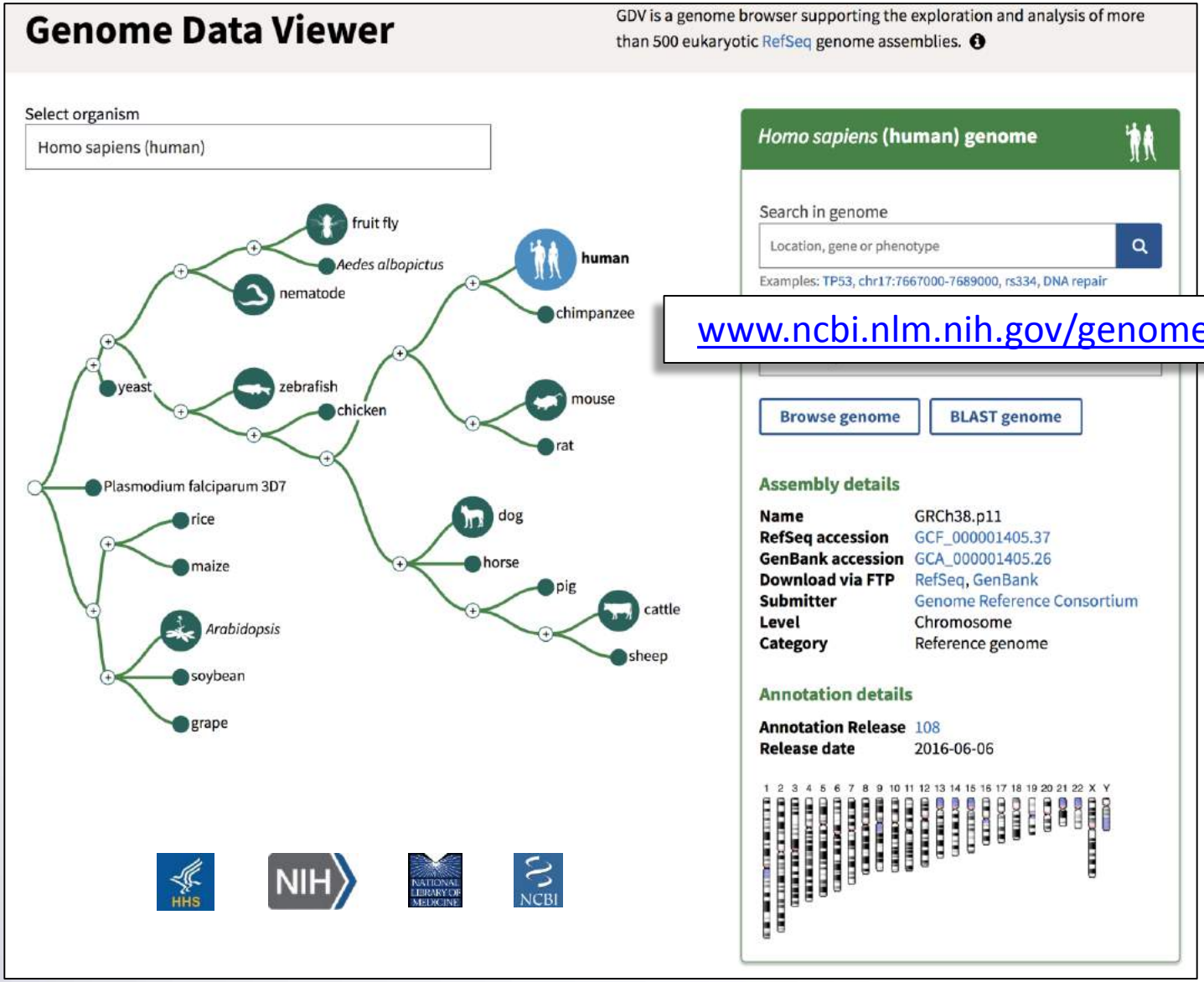
The Entrez system: 41 (and counting) **integrated** databases

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

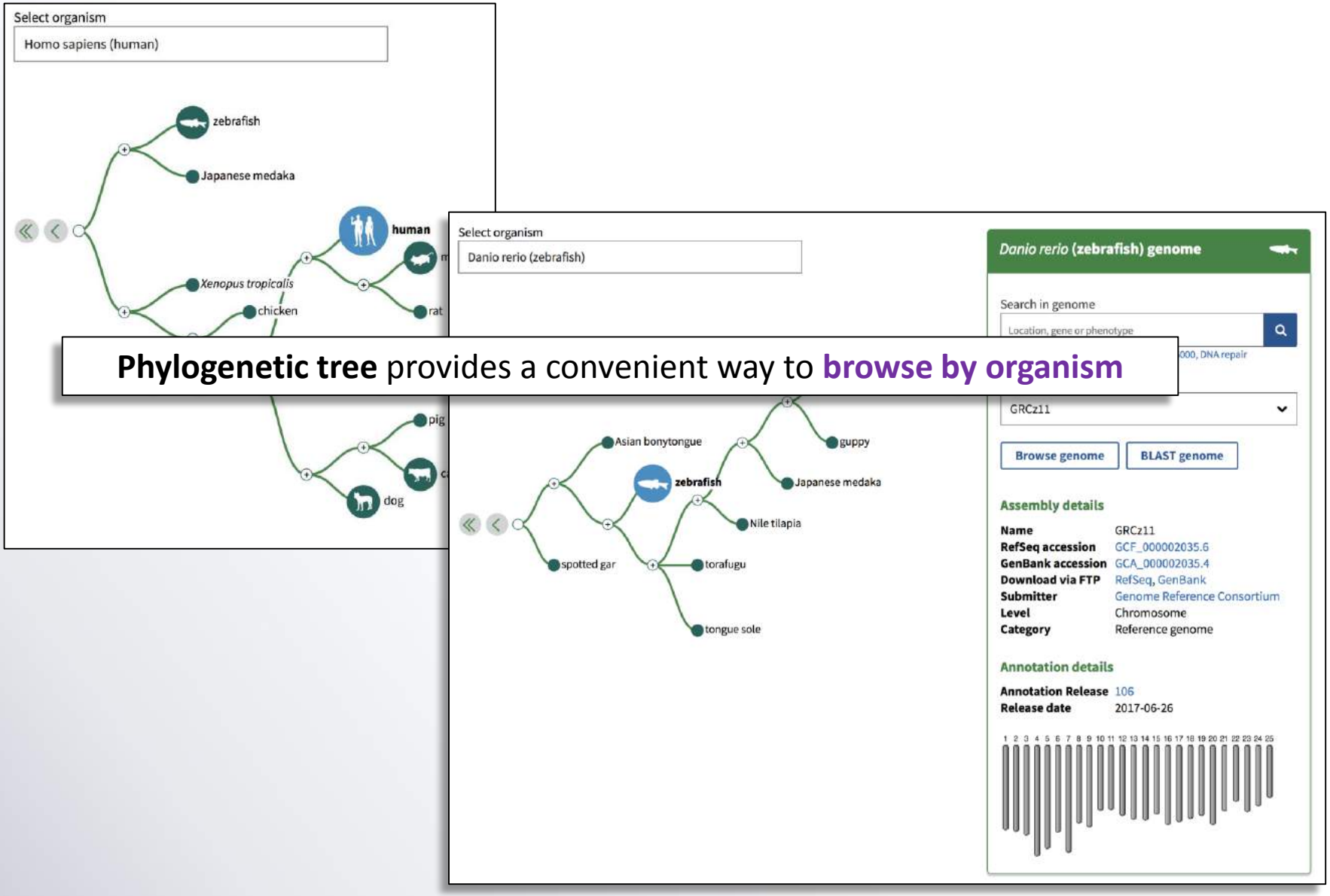
What is NCBI Genome Browser?

- **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



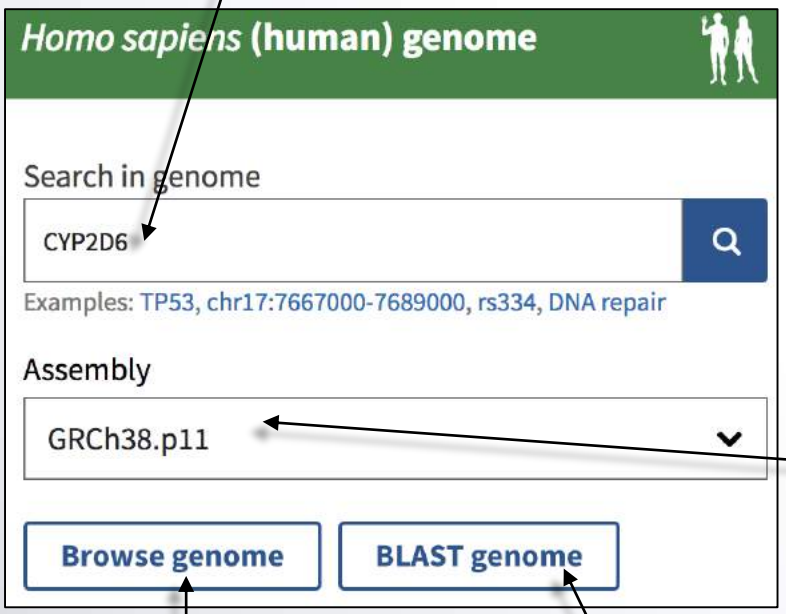
Browsing the species

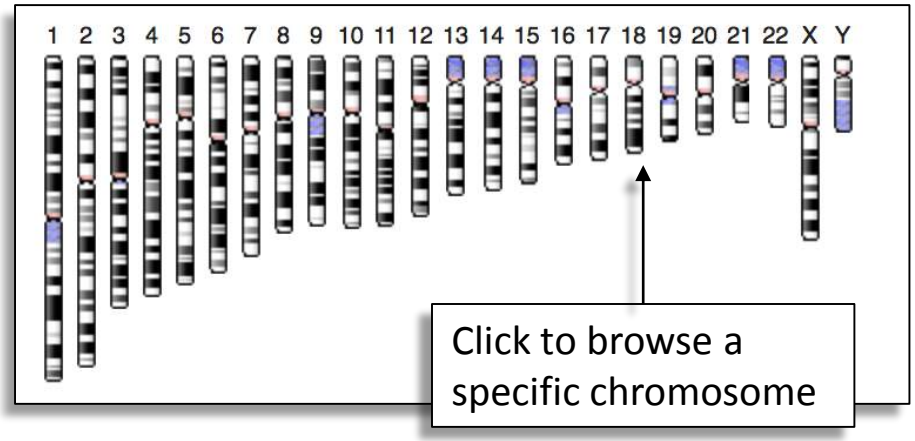


Phylogenetic tree provides a convenient way to **browse by organism**

Browsing the human genome

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





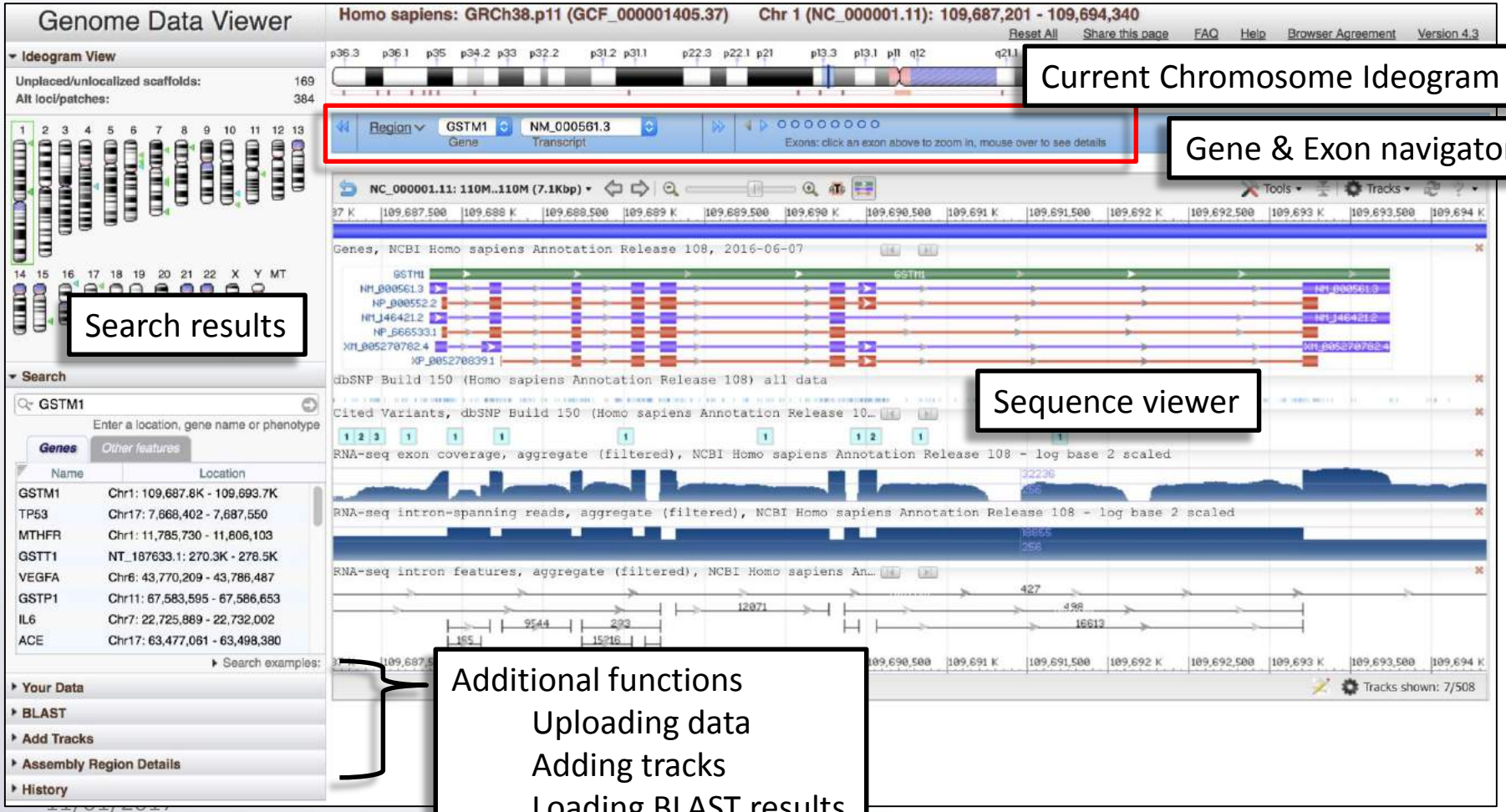
Click to browse a specific chromosome

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



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