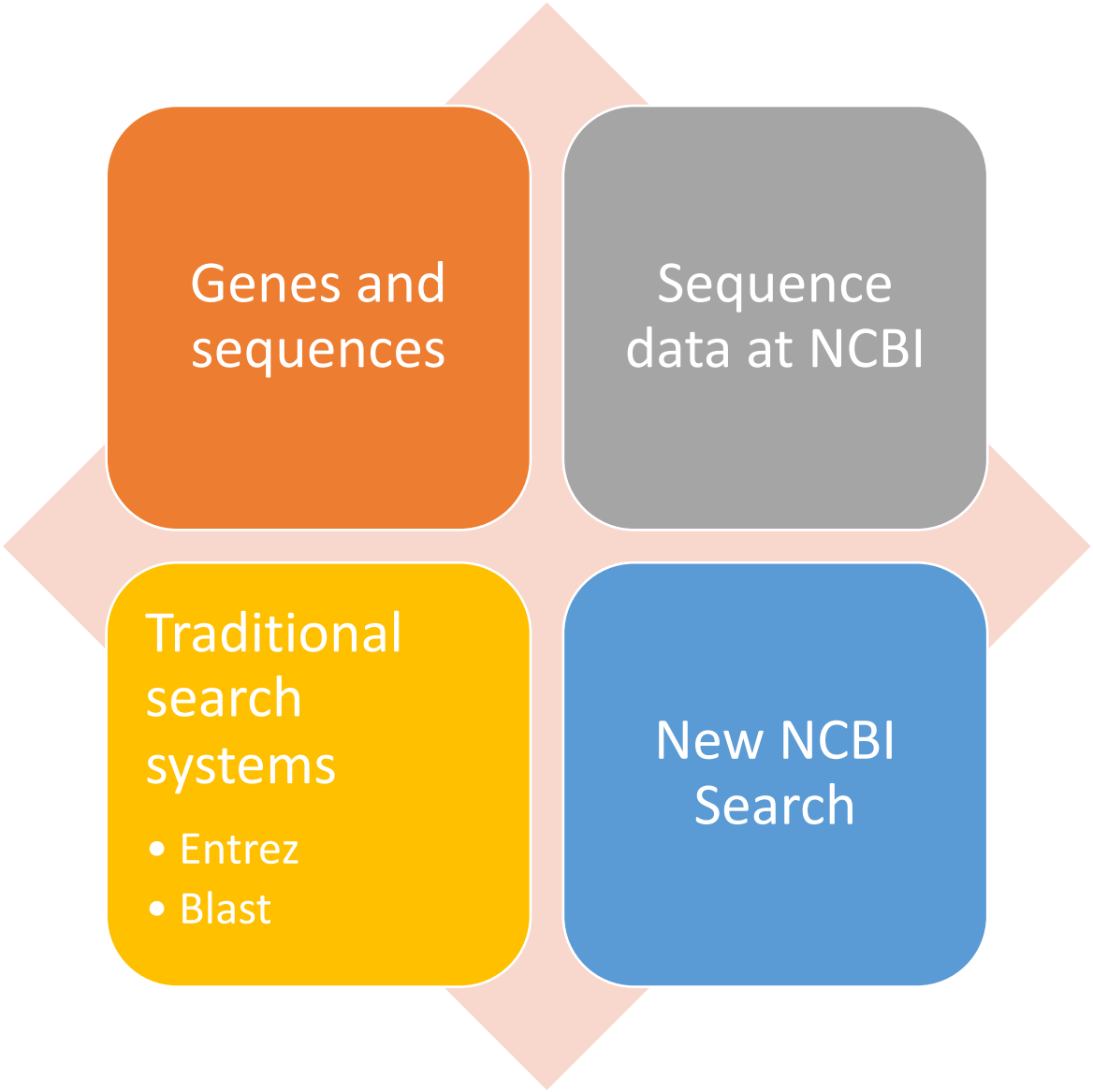
Introduction to NCBI

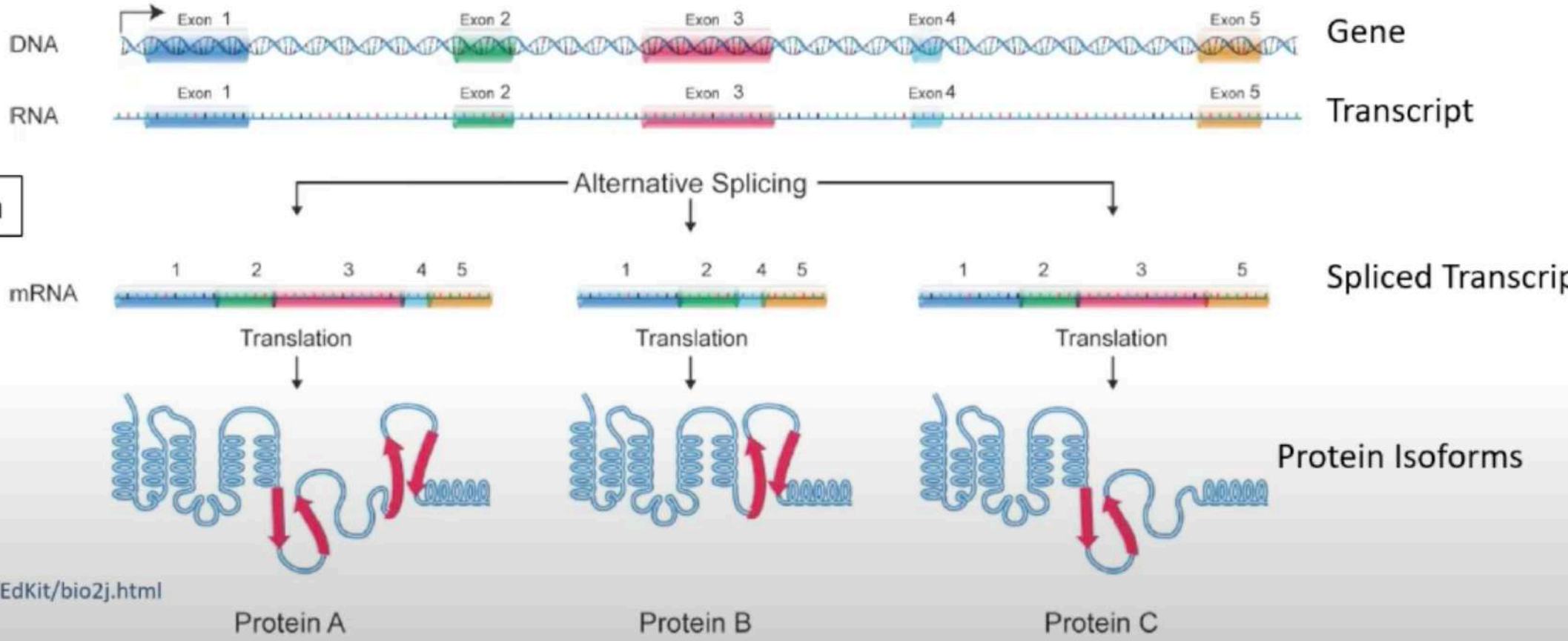
Assoc. Prof. Ilker BUYUK

Outline


Genes and sequences
Sequence data at NCBI
Traditional search systems
• Entrez
• Blast
New NCBI Search

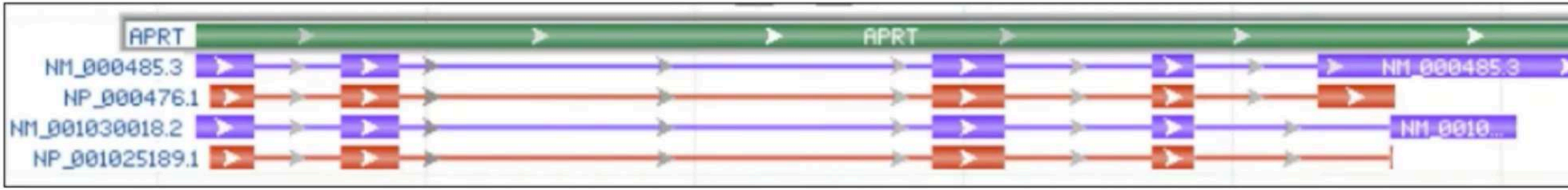
Transcription, splicing & translation

The Central Dogma



http://www.genome.gov/EdKit/bio2j.html

Gene and Transcripts at NCBI



Gene

chromosome 16: NC_000016.10 (88,809,339..88,811,928)

RefSeq Gene: NG_008013.1

Gene products

Two RefSeq Transcripts (splice forms) NM_000485.3, NM_001030018.2

Two RefSeq Proteins (isoforms) NP_000476.1, NP_001025189.1

Sources of sequence data at NCBI

- Submitted nucleotide sequences and corresponding proteins
  - International Sequence Database Colloration (INSDC)
    - Standard Sequences
      - **GenBank** -- US Sequence Database at NCBI
      - **European Nucleotide Archive** at EBI
      - **DNA Databank of Japan** at NIG
    - Next-Gen sequencing reads
      - Sequence Read Archive (SRA)
- High quality curated DNA and protein records
  - NCBI Reference Sequences
  - Swiss-Prot
  - Proteins from structures (PDB)

Example sequences for APRT

Human non-model RefSeqs (based on INSDC submisssions)

Gene products

Transcripts	Proteins
NM_000485.3 →	NP_000476.1
NM_001030018.2 →	NP_001025189.1

Genomic

Primary Assembly

NC_000016.10 (88,809,339..88,811,928)

RefSeqGene

NG_008013.1 (5,007..7,596)

Rhesus Model RefSeqs (based on gene prediction and next-gen read data)

Transcripts	Proteins
XM_028841000.1 →	XP_028696833.1
XM_001089867.4 →	XP_001089867.2
XM_015126795.2 →	XP_014982281.2

GenBank accessions

Transcripts

- BC107151.2
- BC106894.2
- BC106895.2

Genomic

- AY306126.1
- AC092384.5
- M16446.1
- U04709.1
- U09817.1

Other Protein

PDB	UniProtKB
1ORE_A	P07741.2
4X44_A	
6FCI_A	
4X45_A	
6FCH_A	
6FD4_A	

PDB sequence to structure links

adenine phosphoribosyltransferase isoform a [Homo sapiens]

NCBI Reference Sequence: NP_000476.1

Identical Proteins FASTA Graphics

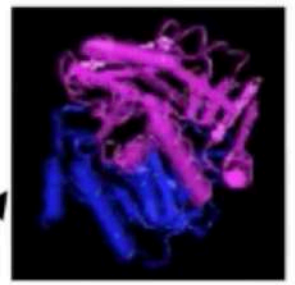
Go to:

```
LOCUS       NP_000476                180 aa            linear   PRI 22-AUG-2019
DEFINITION  adenine phosphoribosyltransferase isoform a [Homo sapiens].
ACCESSION   NP_000476
VERSION     NP_000476.1
DBSOURCE    REFSEQ: accession NM_000485.3
KEYWORDS    RefSeq; RefSeq Select.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

Customize view

Analyze this sequence

Protein 3D Structure



Crystal Stru
Human API
type in com
PDB: 6HG
Source: H
sapiens
Method: X

Diffraction
Resolution: 1.55 Å

NCBI Structure

RefSeq NP_000476 → 6HGS_A PDB protein chain

BLAST Similarity

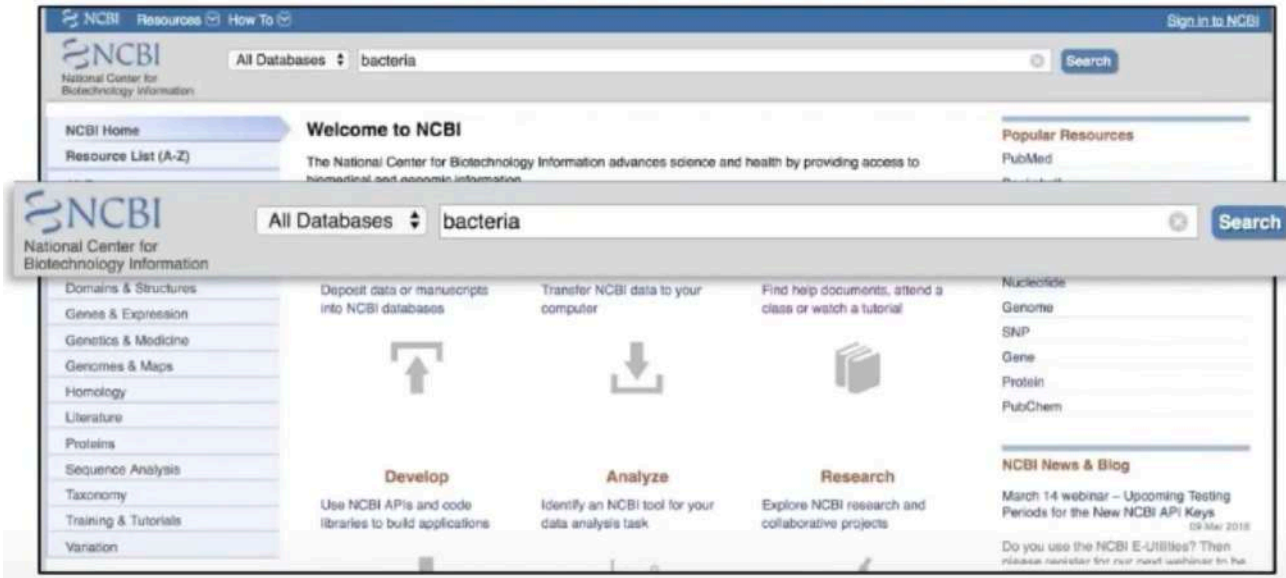
Web Sequence Databases

- Nucleotide www.ncbi.nlm.nih/nuccore/
  - RNA, Genomic DNA sequences
  - 409 million (non-bulk) records
    - 348 million submitted (INSDC),
    - 61 million NCBI Reference Sequences
- Protein www.ncbi.nlm.nih/protein/
  - 774 million records
    - 619 million INSDC coding region translations
    - 153 million NCBI Reference Sequences
    - 2 million imported from outside
      - Includes 440 thousand from 3-D structures (PDB)

Does not including next-gen data (SRA)

Searching NCBI data

NCBI Web Search Systems



Entrez text search system

40-plus integrated databases

Free text and database-specific fielded searches

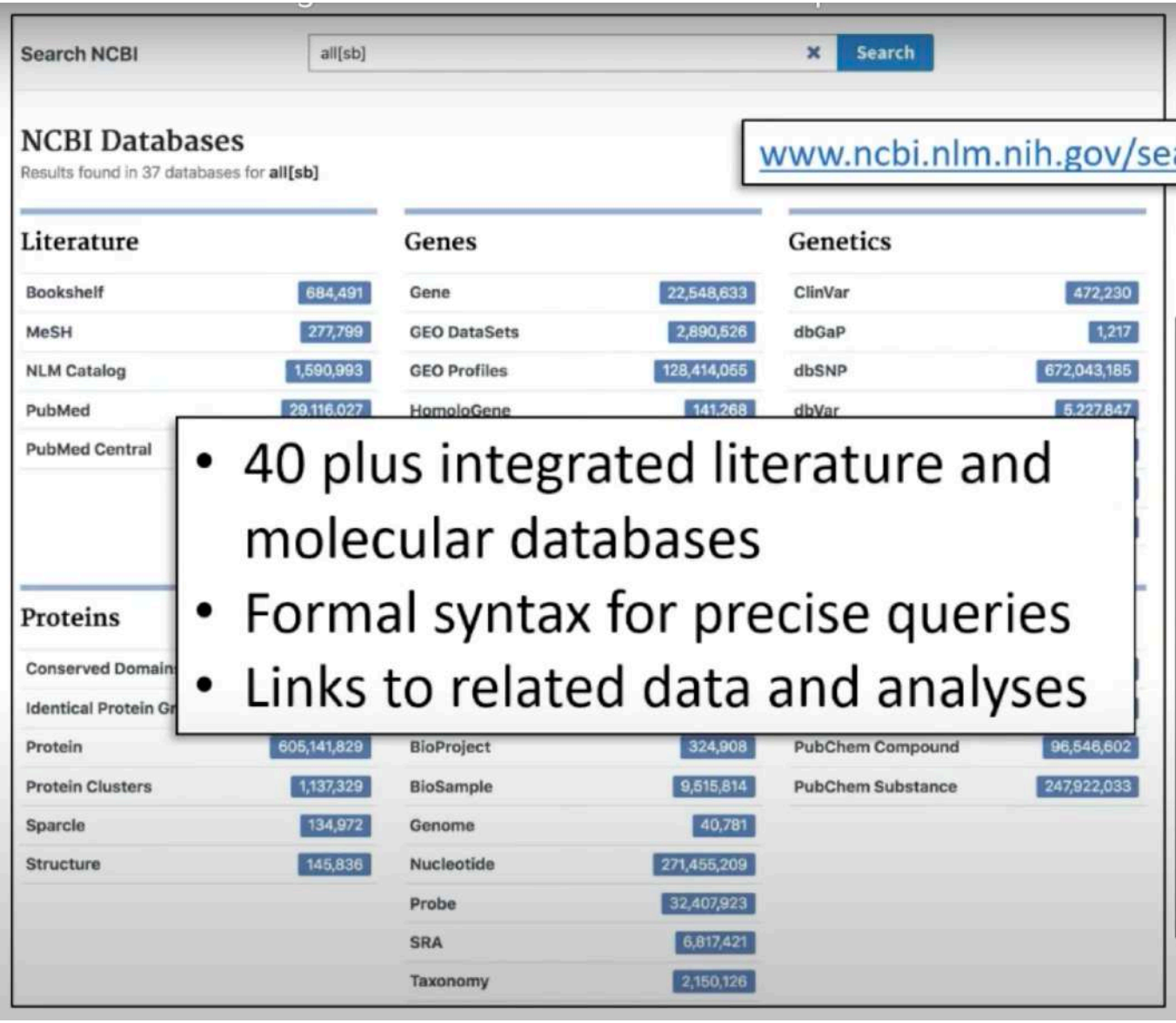
The PubMed search engine

BLAST — Basic Local Alignment Search Tool

- Sequence similarity search tools
- Finds related nucleotide and protein sequences
  - Designed to find homologs
  - Used for other sequence analysis tasks

The Entrez System

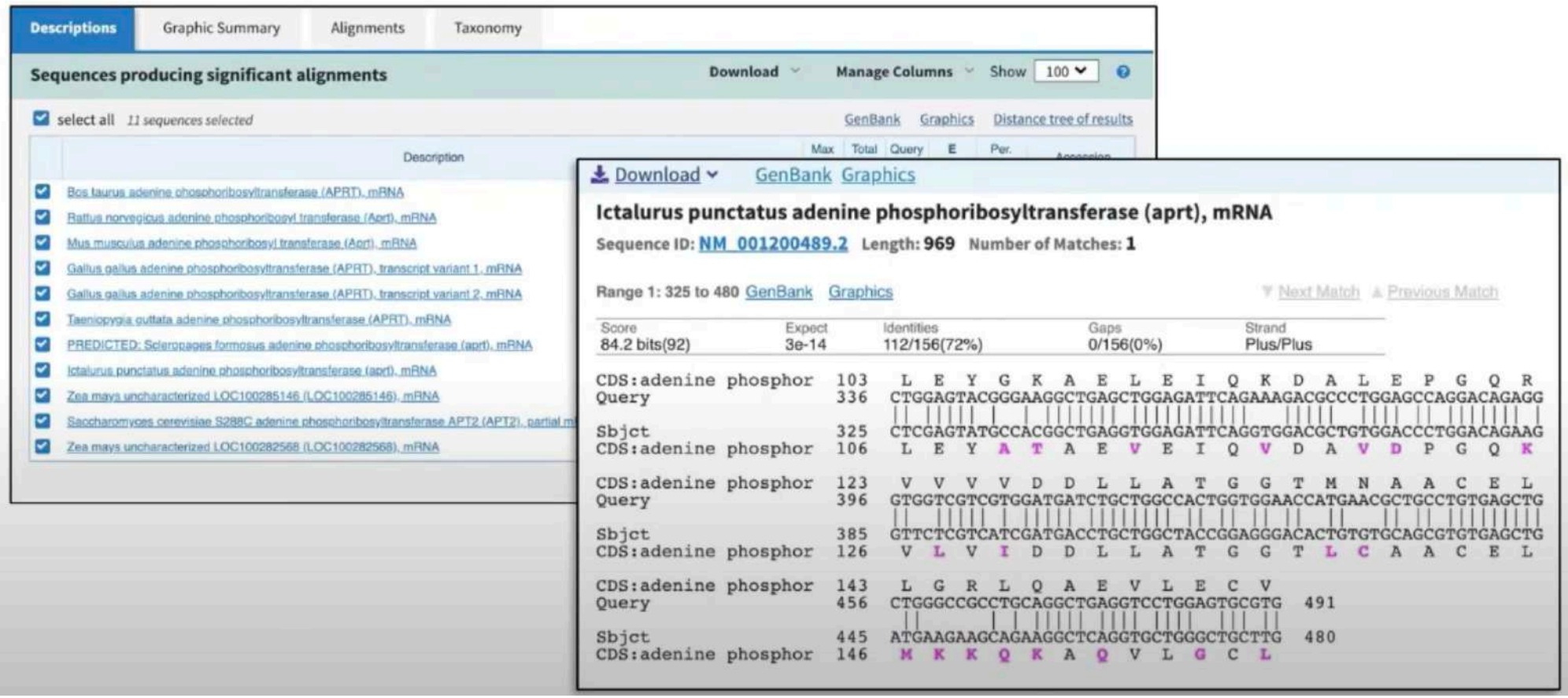
www.ncbi.nlm.nih.gov/search/



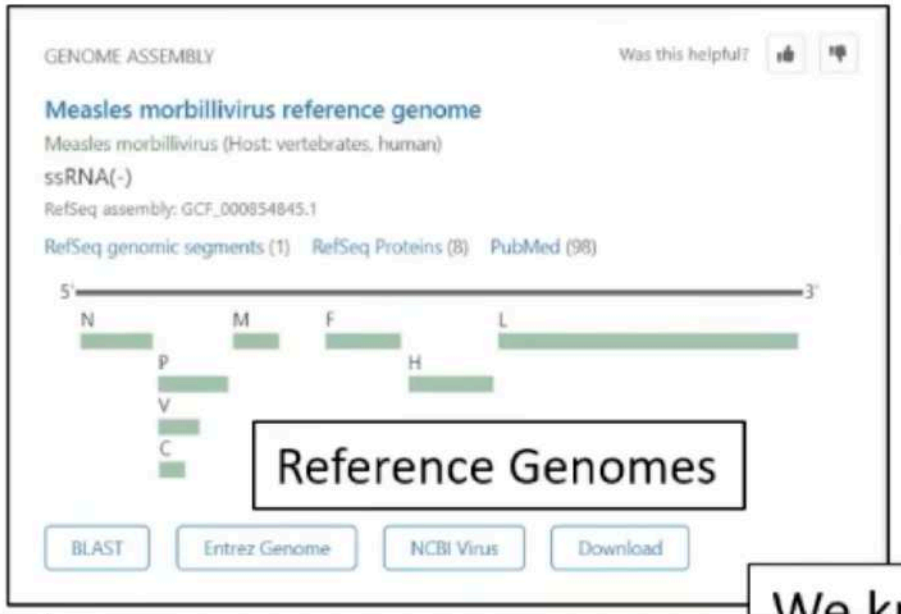
- 40 plus integrated literature and molecular databases
- Formal syntax for precise queries
- Links to related data and analyses

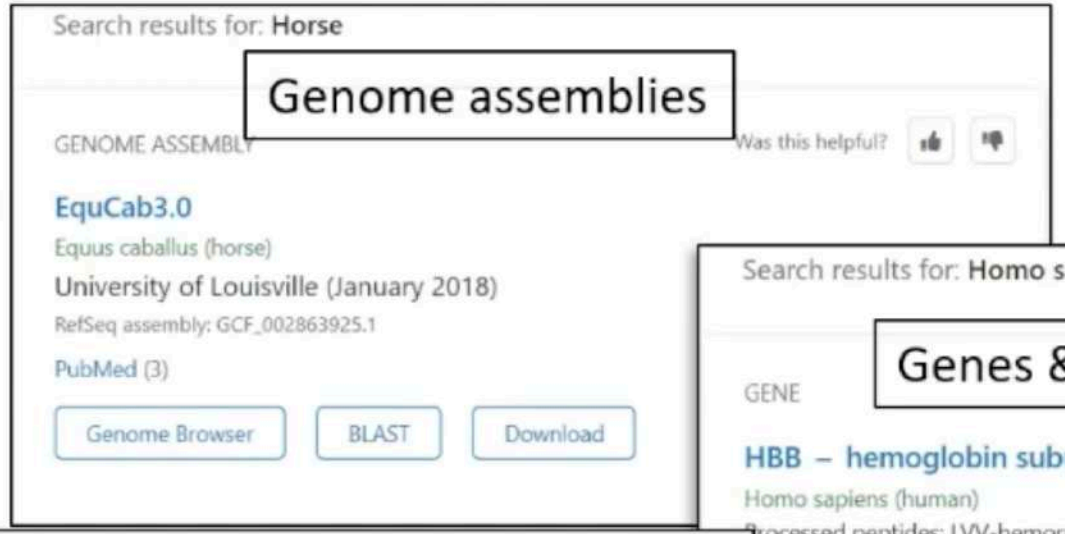
- Where to begin??
- Typically people use only a few for direct searching
  - Start in central hubs
    - Gene
    - Assembly
    - Taxonomy
    - BioProject

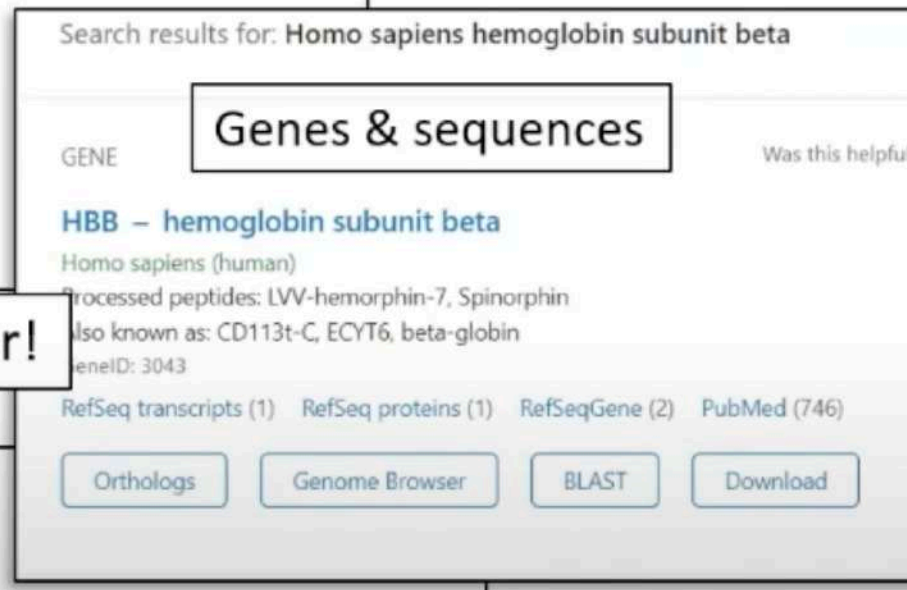
BLAST results: finding related sequences



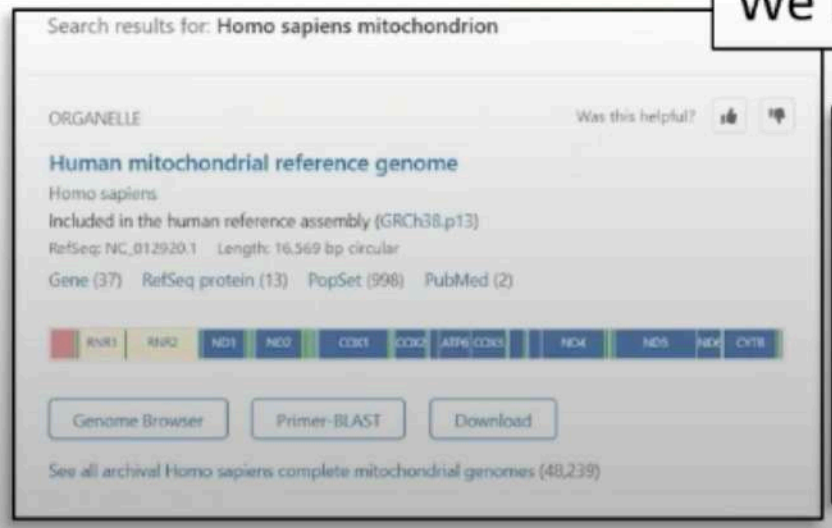
A new search experience ...

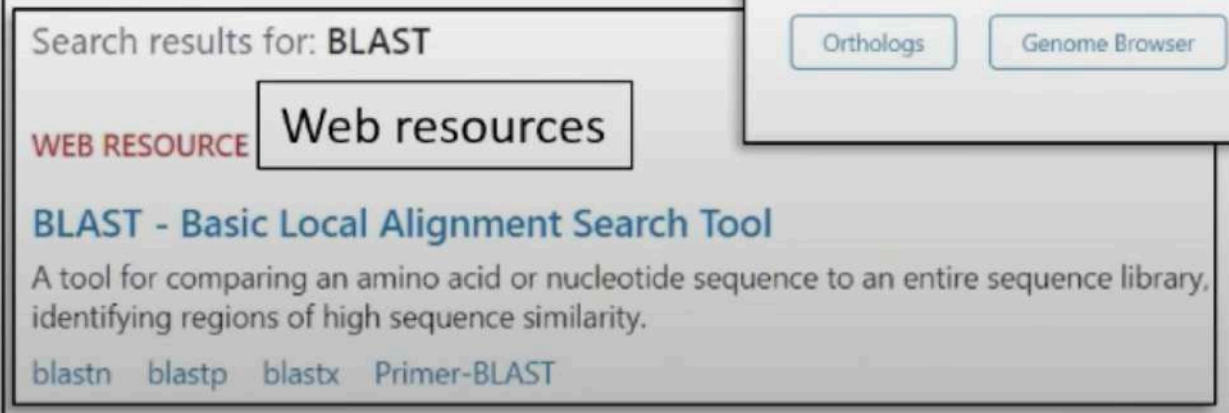




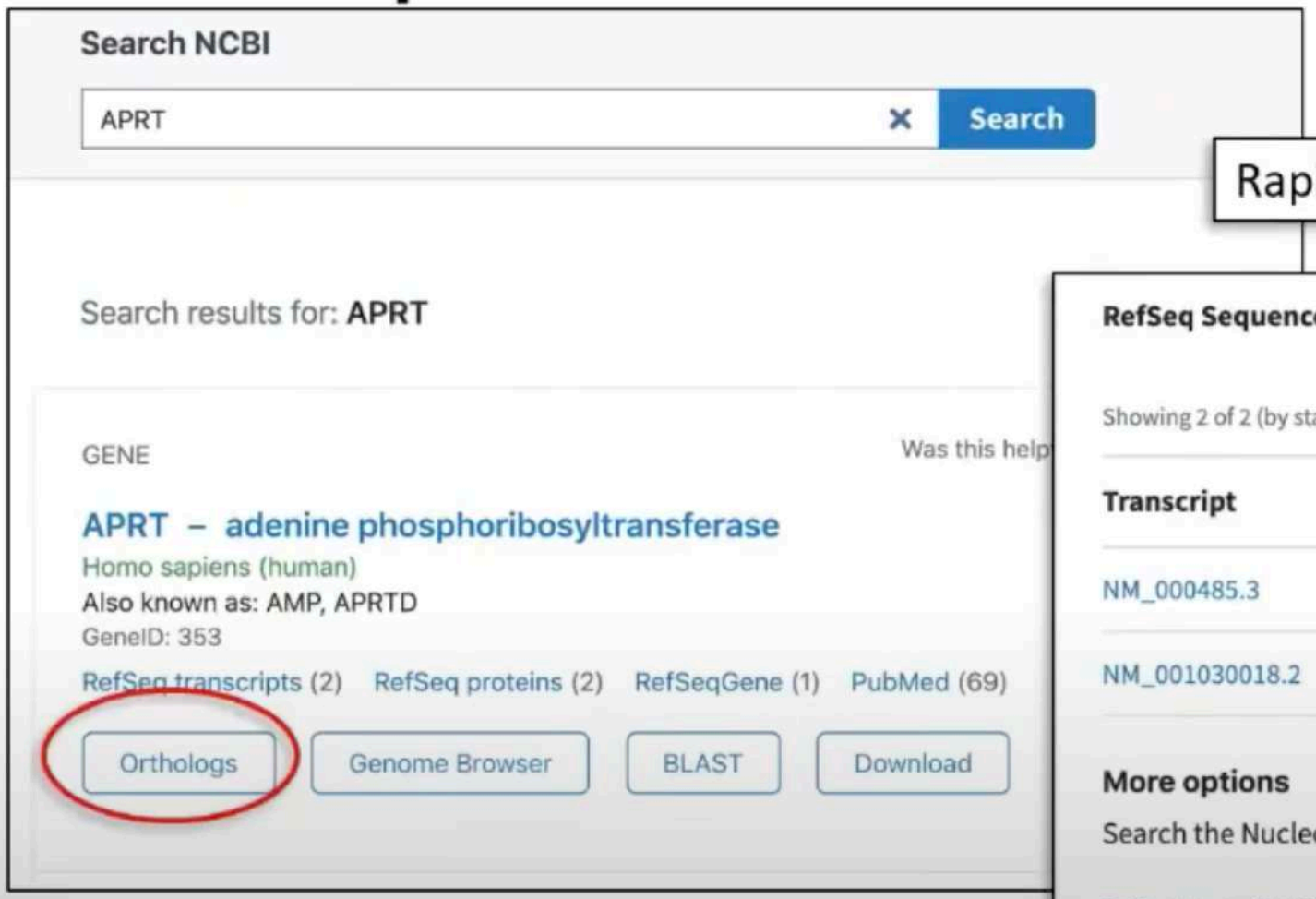


We know what you are looking for!

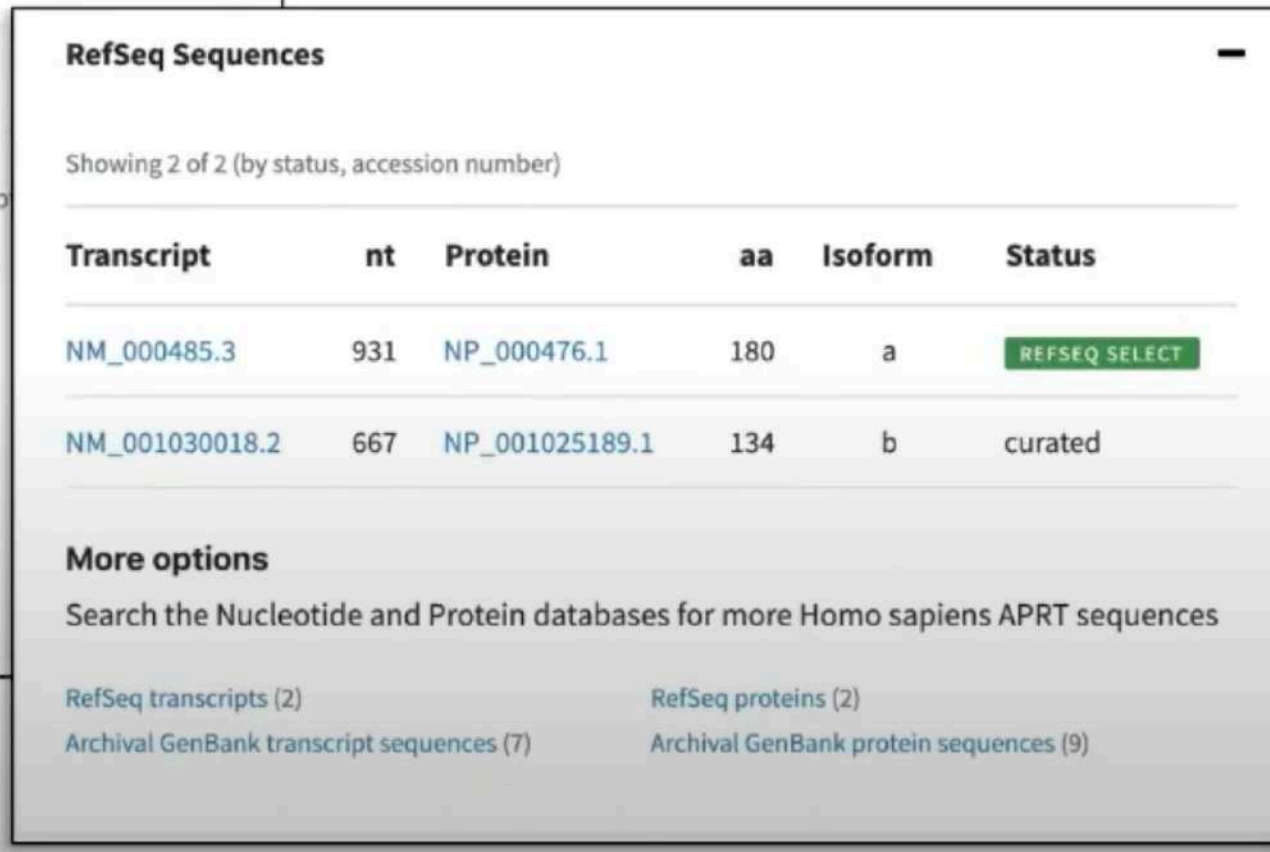




New NCBI Search makes finding genes and sequences easier

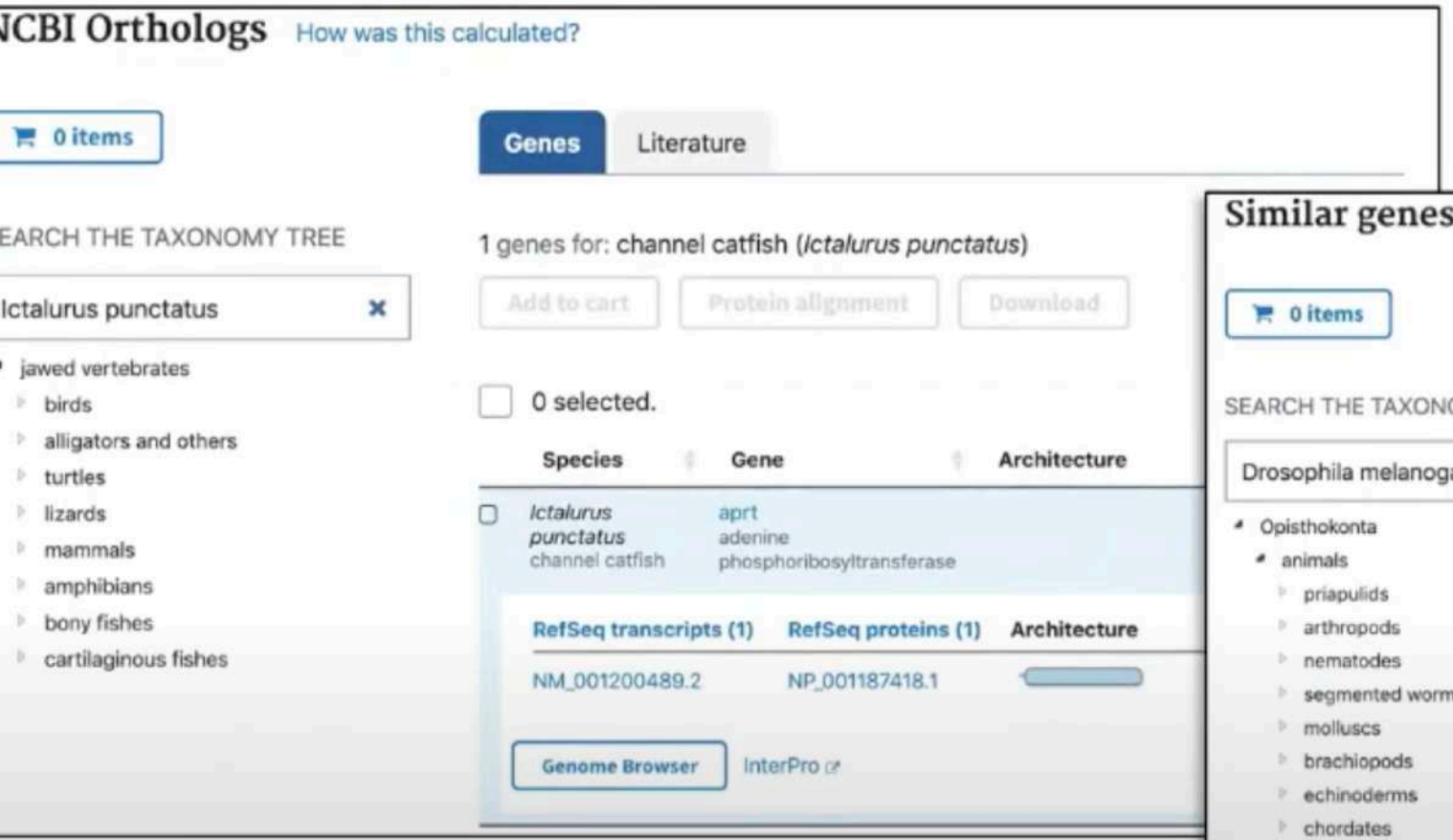


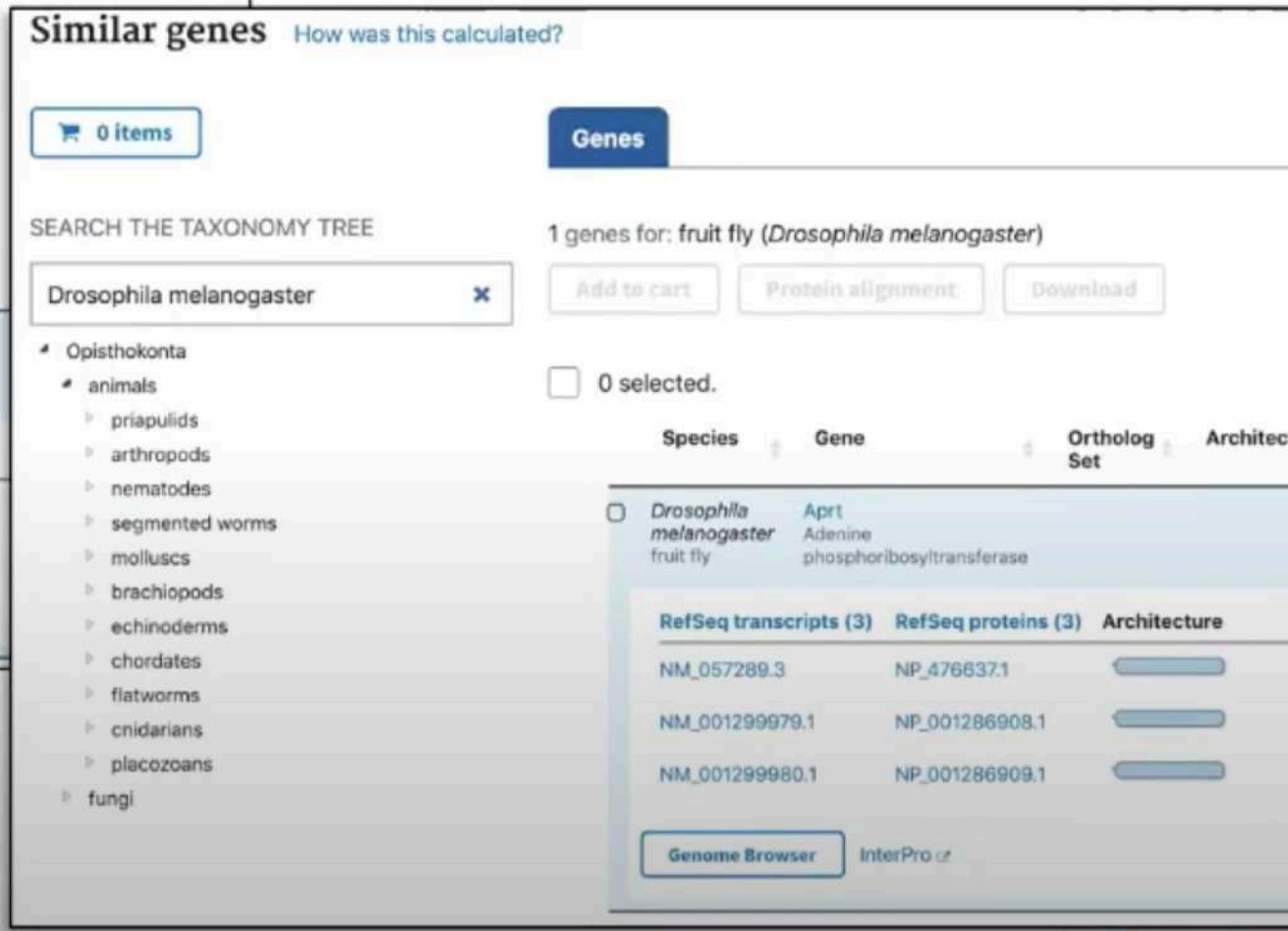
Rapid access to genes and sequences



Orthologs / Similar genes partially replaces BLAST

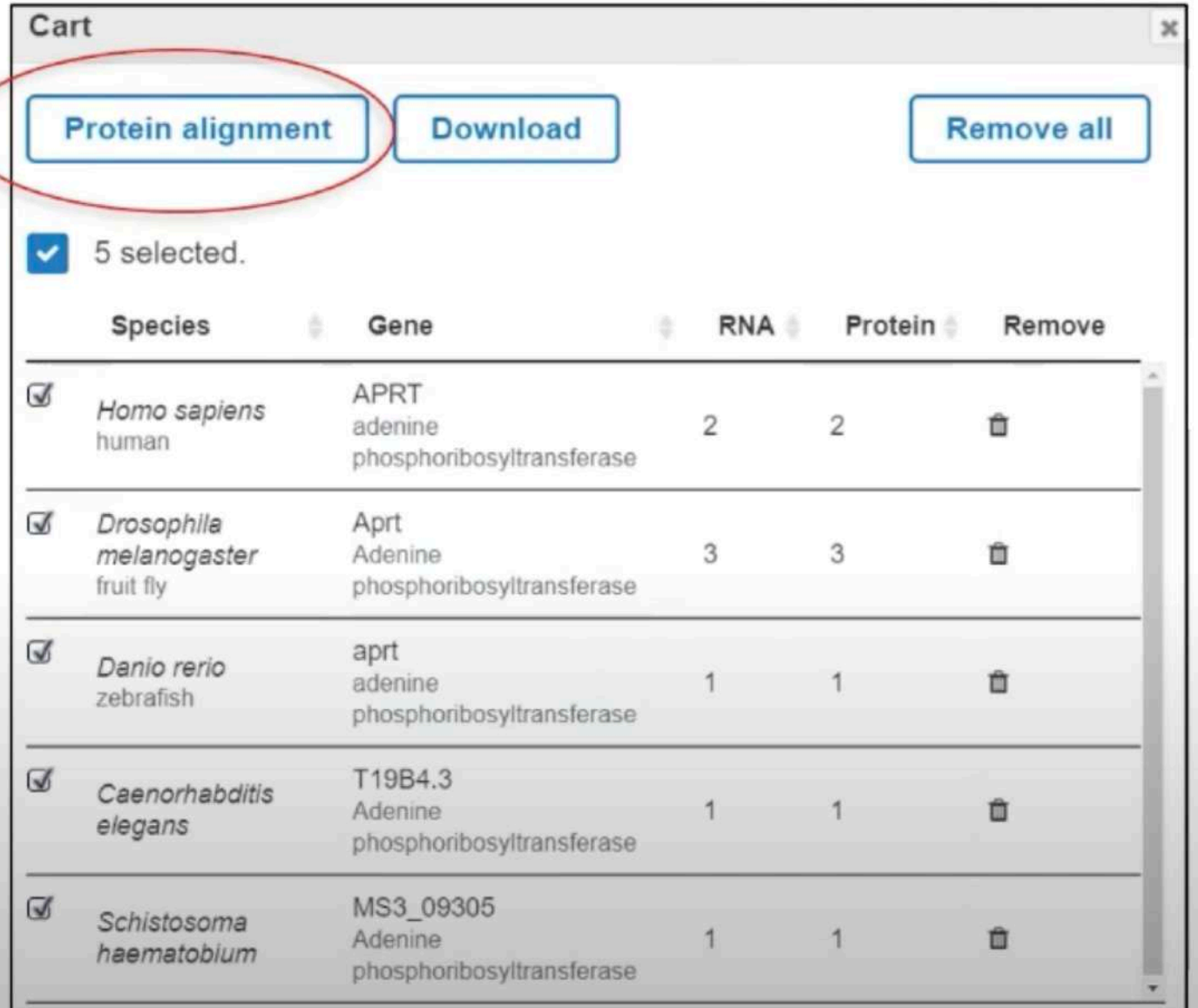
New Search makes finding related genes easier

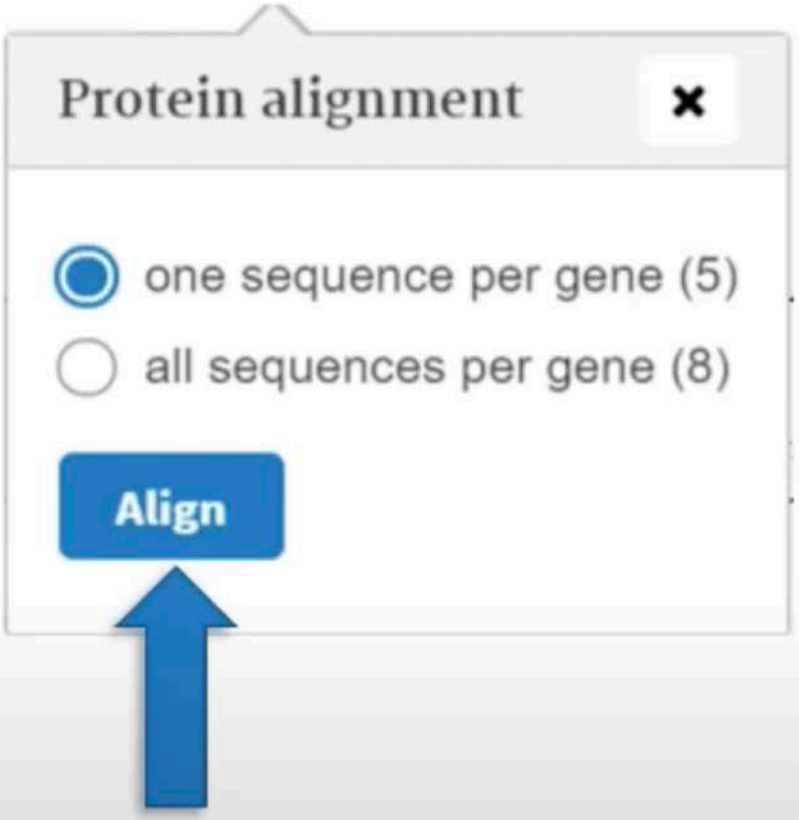




Orthologs of vertebrate genes
Similar genes for all eukaryotes

And comparing sequences easier





Live demonstrations

Sequences for adenine ribosyltransferase

The known item search for APRT

- Finding related sequences
  - Orthologs
  - Similar Genes
  - Accessing COBALT
  - Dowloading genes
- BLAST 2 Sequences
- Protein BLAST for bacterial adenine ribosyltransferase
- Structure of APRT protein in iCn3D

Thanks for listening