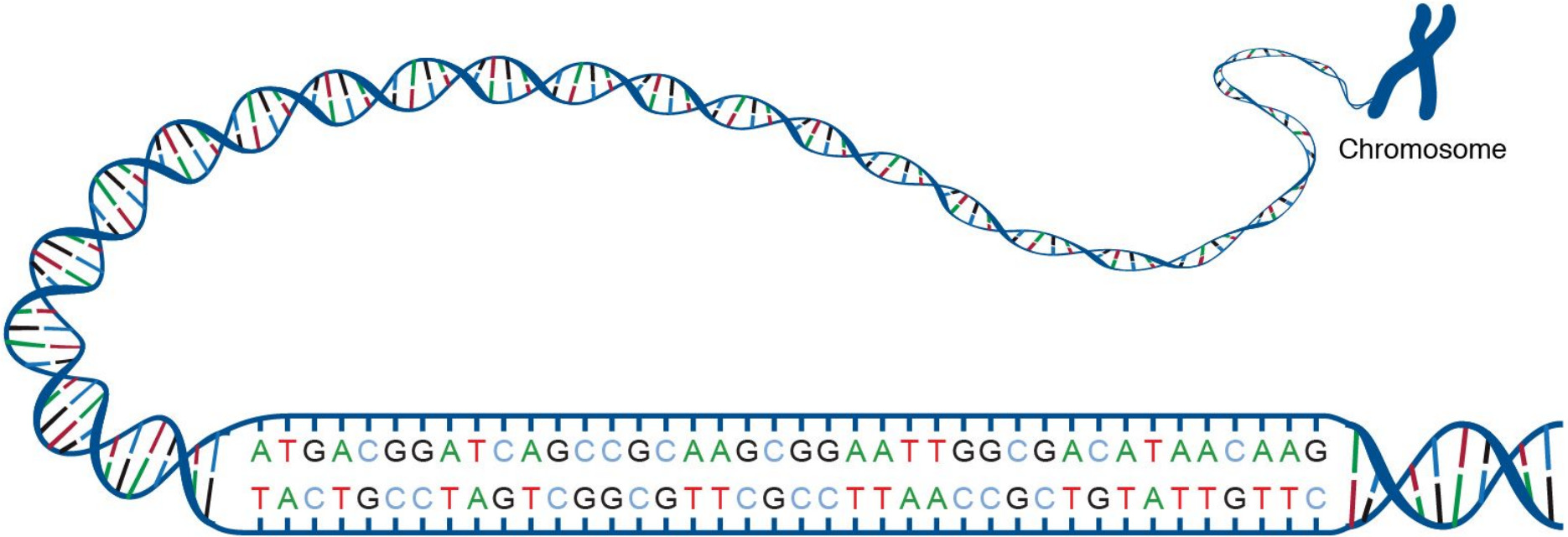
Basic Bioinformatics

PSB
Lect. Todsapol Techo

The Languages of Life (DNA)

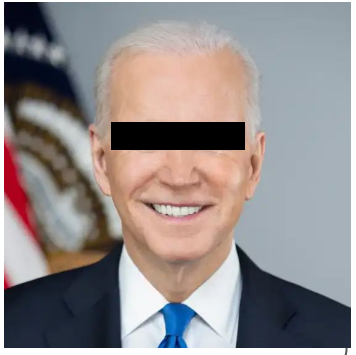
- Living information store in DNA sequence



~ 99.6% similarity

GCTCGAGAGATTTCTCTTTTACCTTTTTTTTACTATTTTTCACTCTCCCATAACCTCCTATATTGACTGAT
CTGTAATAACCACGATATTATTGGAATAAATAGGGGCTTGAAATTTGGAAAAAAAAAAAAAACTGAAATA
TTTTCGTGATAAGTGATAGTGATATTCTTCTTTTATTTGCTACTGTTACTAAGTCTCATGTACTAACATC
GATTGCTTCATTCTTTTTGTTGCTATATTATATGTTTAGAGGTTGCTGCTTTGGTTATTGATAACGGTTC
TGGTATGTGTAAAGCCGGTTTTGCCGGTGACGACGCTCCTCGTGCTGTCTTCCCATCTATCGTCGGTAGA
CCAAGACACCAAGGTATCATGGTCGGTATGGGTCAAAAAGACTCCTACGTTGGTGATGAAGCTCAATCCA
AGAGAGGTATCTTGACTTTACGTTACCCAATTGAACACGGTATTGTCACCAACTGGGACGATATGGAAAA
GATCTGGCATCATACCTTCTACAACGAATTGAGAGTTGCCCCAGAAGAACACCCTGTTCTTTTGACTGAA
GCTCCAATGAACCCTAAATCAAACAGAGAAAAGATGACTCAAATTATGTTTGAAACTTTCAACGTTCCAG
CCTTCTACGTTTCCATCCAAGCCGTTTTGTCCTTGTACTCTTCCGGTAGAACTACTGGTATTGTTTTGGA
TTCCGGTGATGGTGTTACTCACGTCGTTCCAATTTACGCTGGTTTCTCTCTACCTCACGCCATTTTGAGA
ATCGATTTGGCCGGTAGAGATTTGACTGACTACTTGATGAAGATCTTGAGTGAACGTGGTTACTCTTTCT
CCACCACTGCTGAAAGAGAAATTGTCCGTGACATCAAGGAAAAACTATGTTACGTCGCCTTGGACTTCGA
ACAAGAAATGCAAACCGCTGCTCAATCTTCTTCAATTGAAAAATCCTACGAACTTCCAGATGGTCAAGTC
ATCACTATTGGTAACGAAAGATTCAGAGCCCCAGAAGCTTTGTTCCATCCTTCTGTTTTGGGTTTGGAAT
CTGCCGGTATTGACCAAACTACTTACAACTCCATCATGAAGTGTGATGTCGATGTCCGTAAGGAATTATA
CGGTAACATCGTTATGTCCGGTGGTACCACCATGTTCCCAGGTATTGCCGAAAGAATGCAAAAGGAAATC





~ 98.8% similarity

GCTCGAGAGATTTCTCTTTTACCTTTTTTTTACTATTTTTCACTCTCCCATAACCTCCTATATTGACTGAT
CTGTAATAACCACGATATTATTGGAATAAATAGGGGCTTGAAATTTGGAAAAAAAAAAAAAACTGAAATA
TTTTCGTGATAAGTGATAGTGATATTCTTCTTTTATTTGCTACTGTTACTAAGTCTCATGTACTAACATC
GATTGCTTCATTCTTTTTGTTGCTATATTATATGTTTAGAGGTTGCTGCTTTGGTTATTGATAACGGTTC
TGGTATGTGTAAAGCCGGTTTTGCCGGTGACGACGCTCCTCGTGCTGTCTTCCCATCTATCGTCGGTAGA
CCAAGACACCAAGGTATCATGGTCGGTATGGGTCAAAAAGACTCCTACGTTGGTGATGAAGCTCAATCCA
AGAGAGGTATCTTGACTTTACGTTACCCAATTGAACACGGTATTGTCACCAACTGGGACGATATGGAAAA
GATCTGGCATCATACCTTCTACAACGAATTGAGAGTTGCCCCAGAAGAACACCCTGTTCTTTTGACTGAA
GCTCCAATGAACCCTAAATCAAACAGAGAAAAGATGACTCAAATTATGTTTGAAACTTTCAACGTTCCAG
CCTTCTACGTTTCCATCCAAGCCGTTTTGTCCTTGTACTCTTCCGGTAGAACTACTGGTATTGTTTTGGA
TTCCGGTGATGGTGTTACTCACGTCGTTCCAATTTACGCTGGTTTCTCTCTACCTCACGCCATTTTGAGA
ATCGATTTGGCCGGTAGAGATTTGACTGACTACTTGATGAAGATCTTGAGTGAACGTGGTTACTCTTTCT
CCACCACTGCTGAAAGAGAAATTGTCCGTGACATCAAGGAAAAACTATGTTACGTCGCCTTGGACTTCGA
ACAAGAAATGCAAACCGCTGCTCAATCTTCTTCAATTGAAAAATCCTACGAACTTCCAGATGGTCAAGTC
ATCACTATTGGTAACGAAAGATTCAGAGCCCCAGAAGCTTTGTTCCATCCTTCTGTTTTGGGTTTGGAAT
CTGCCGGTATTGACCAAACTACTTACAACTCCATCATGAAGTGTGATGTCGATGTCCGTAAGGAATTATA
CGGTAACATCGTTATGTCCGGTGGTACCACCATGTTCCCAGGTATTGCCGAAAGAATGCAAAAGGAAATC

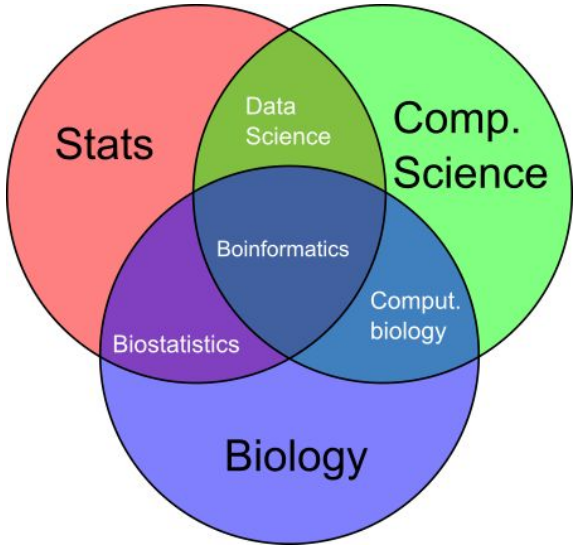
Similarity ?

What is Bioinformatics?

- **Interdisciplinary**
  - **Biology**
  - **Statistics/Mathematics**
  - **Computer Science**
- **Biological DATA**
  - **Nucleotide Sequence** (**DNA** and **RNA**)
  - Protein Sequence (Amino acid)
- **Big DATA**
  - Manipulate specific data from the database
  - Omic experiments

Bioinformatics

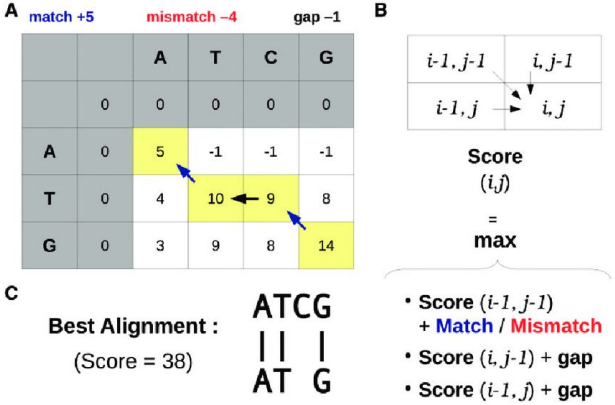
Find the ways to understand biological phenomenon from Biological Data.



Molecular Biology

Bioinformatics

Statistics/Mathematics

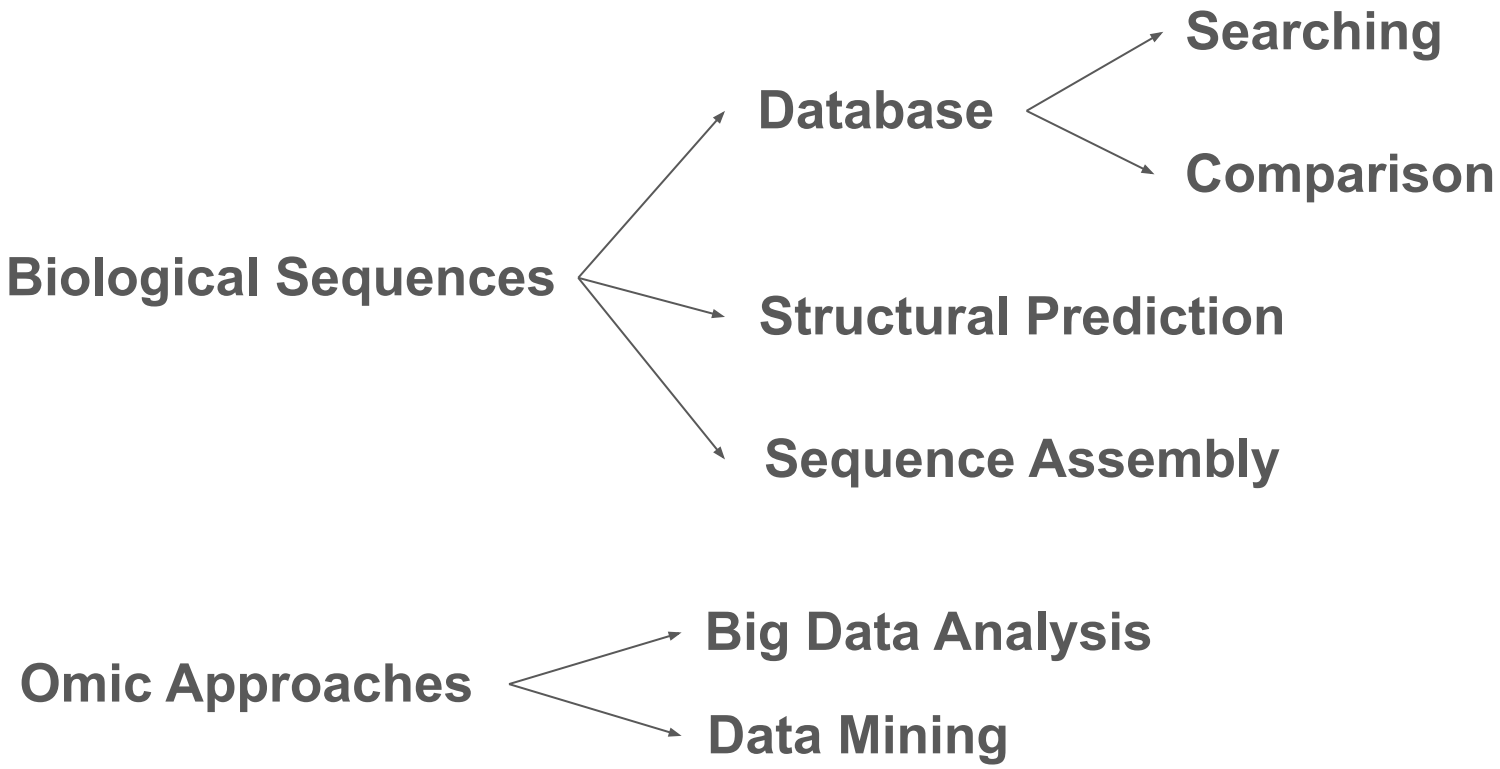


Computer Science

script.py

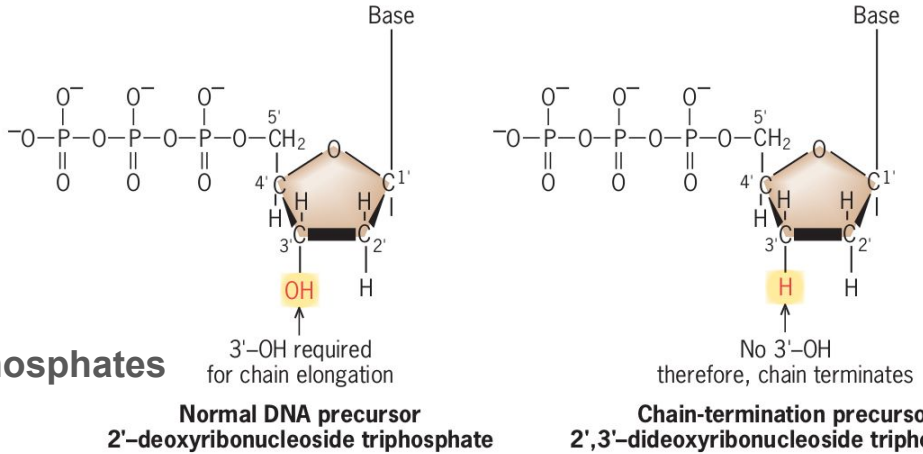
```
1 facebook = "Facebook's rating is"
2 fb_rating = 3.5
3
4 fb_rating_str = str(3.5)
5 fb = facebook + ' ' + fb_rating_str
6
7 print(fb)
```

Overview of Bioinformatics



DNA Sequencing

- **Sanger sequencing: Sequencing by Synthesis**
  - **An idea from DNA replication**
    - **DNA polymerase**
    - DNA template
      - PCR product
      - Recombinant plasmid
    - **single strand primer**
    - **dNTP**
    - **Dideoxyribonucleoside triphosphates**
      - **ddNTP**



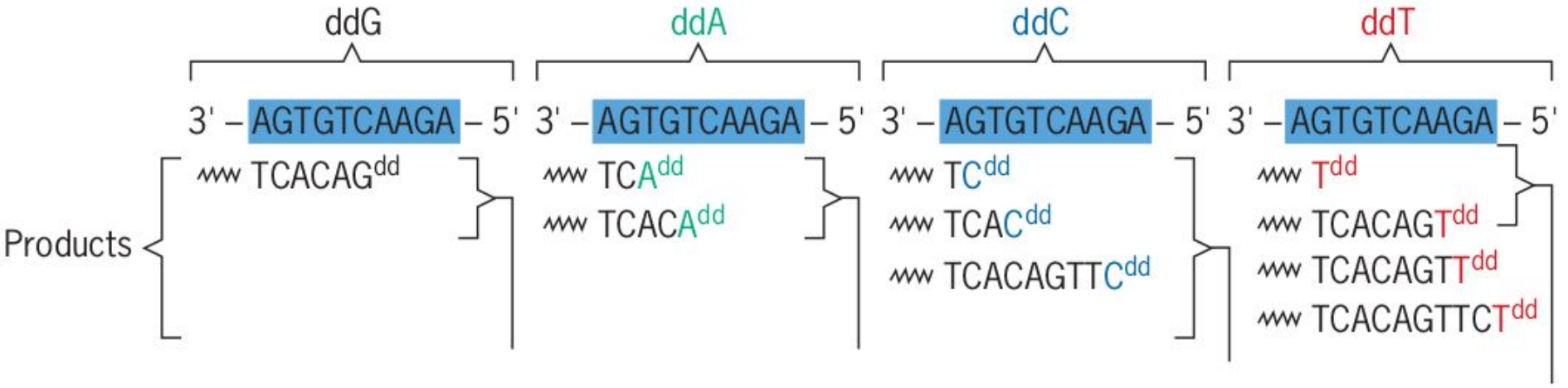
■ **FIGURE 14.16** Comparison of the structures of the normal DNA precursor 2′-deoxyribonucleoside triphosphate and the chain-terminator 2′,3′-dideoxyribonucleoside triphosphate used in DNA sequencing reactions.

Set up a DNA polymerization reaction containing the following:

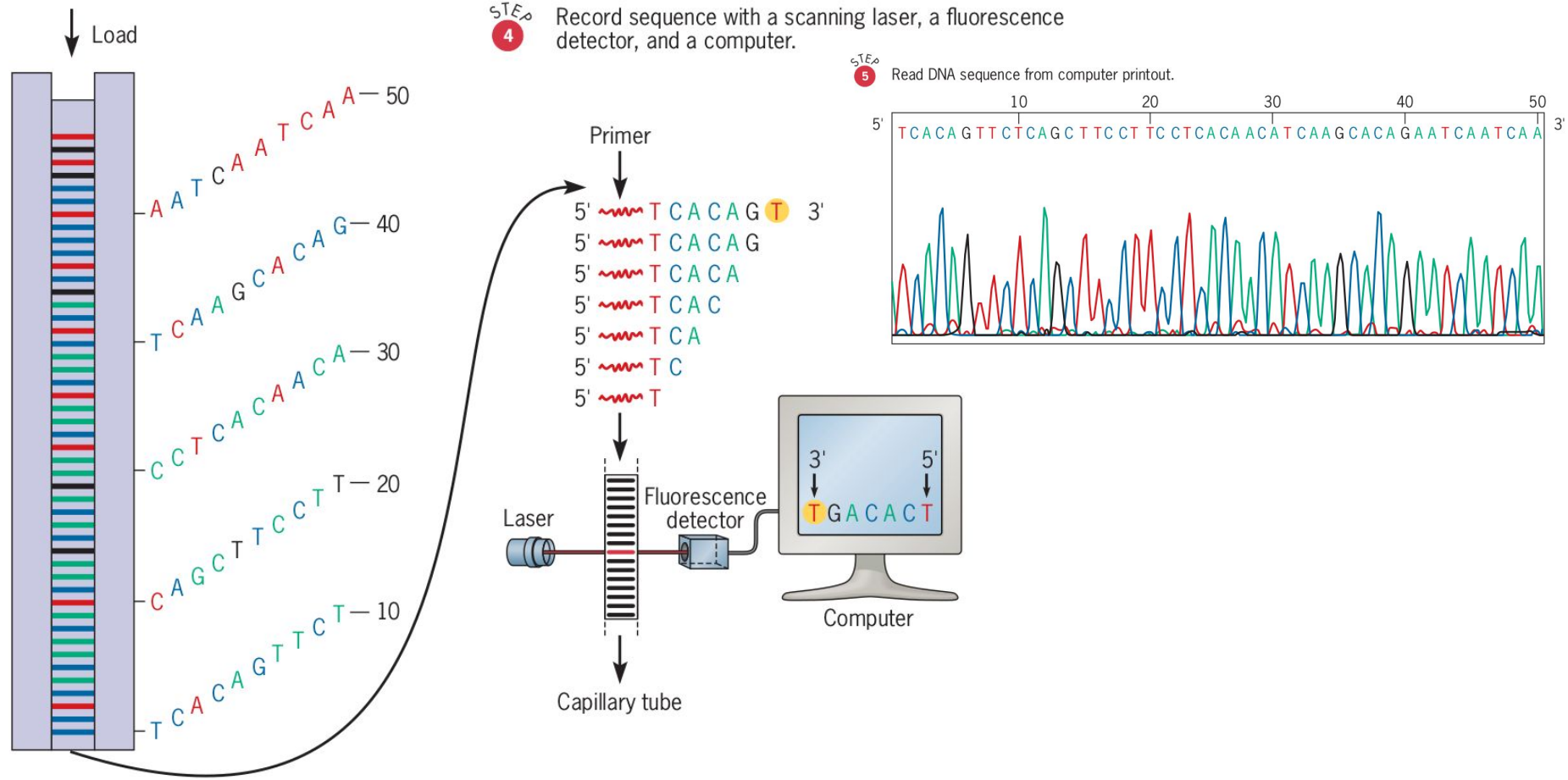
Template strand	3' – AGTGTCAAGA – 5'	DNA polymerase
Primer strand	5' ~~OH 3'	dGTP, dATP, dTTP, and dCTP

All four 2',3'–dideoxyribonucleoside triphosphate chain terminators, each labeled with a different fluorescent dye: ddGTP, ddATP, ddCTP, and ddTTP.

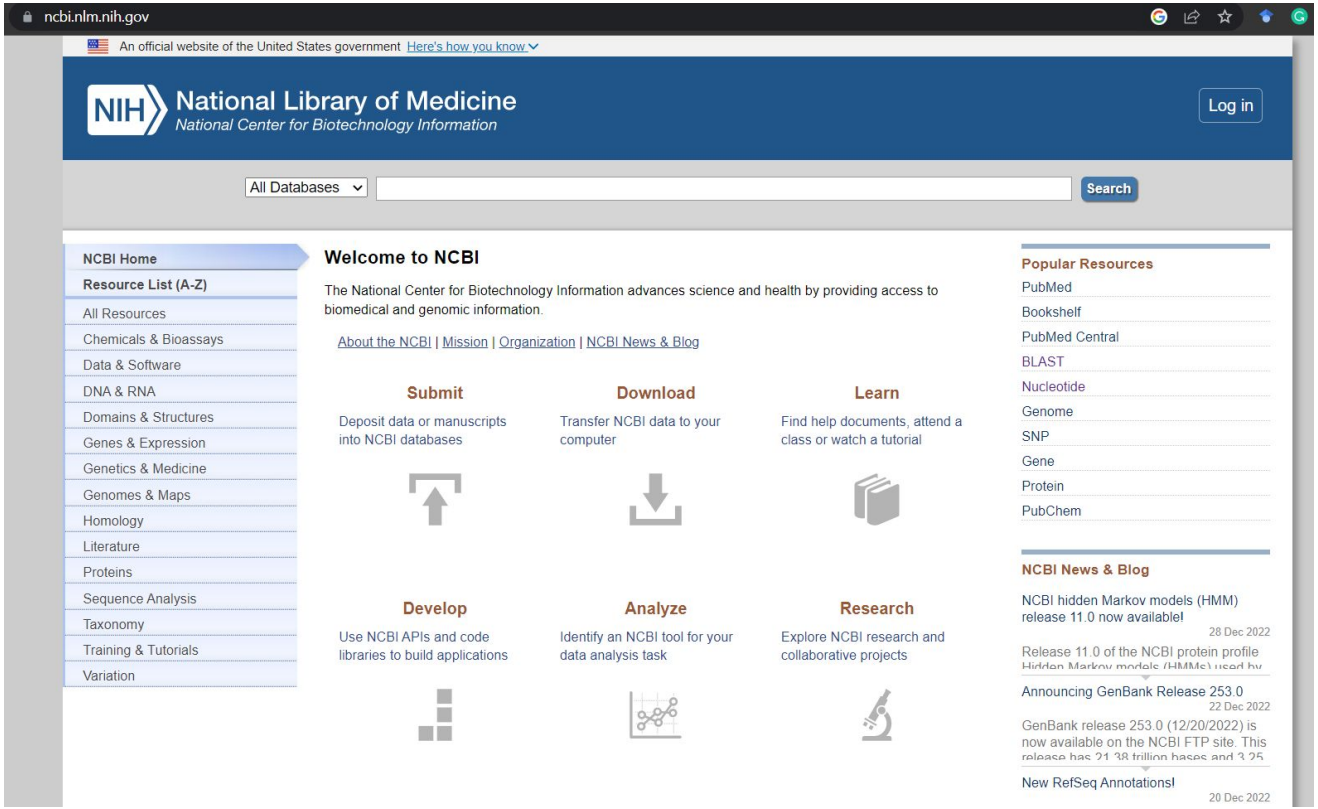
Incubate reaction mixture. Synthesized chains terminating with:



Denature products and separate by polyacrylamide capillary gel electrophoresis.

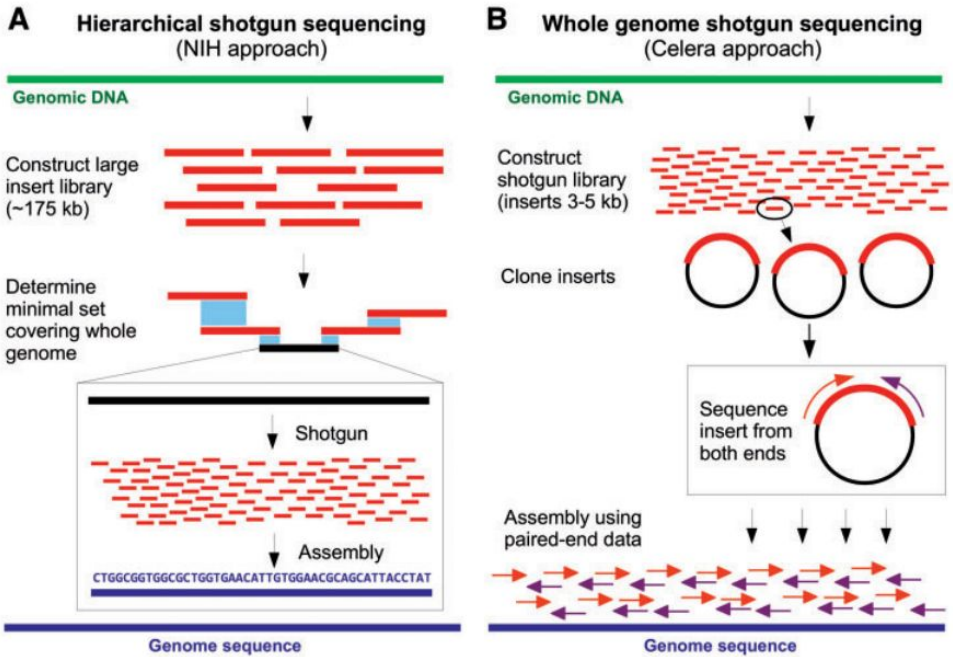


NCBI


ncbi.nlm.nih.gov
An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
NCBI hidden Markov models (HMM) release 11.0 now available!
28 Dec 2022
Release 11.0 of the NCBI protein profile
Announcing GenBank Release 253.0
22 Dec 2022
GenBank release 253.0 (12/20/2022) is now available on the NCBI FTP site. This
New RefSeq Annotations!
20 Dec 2022

Racing in Human Genome Sequencing

Public

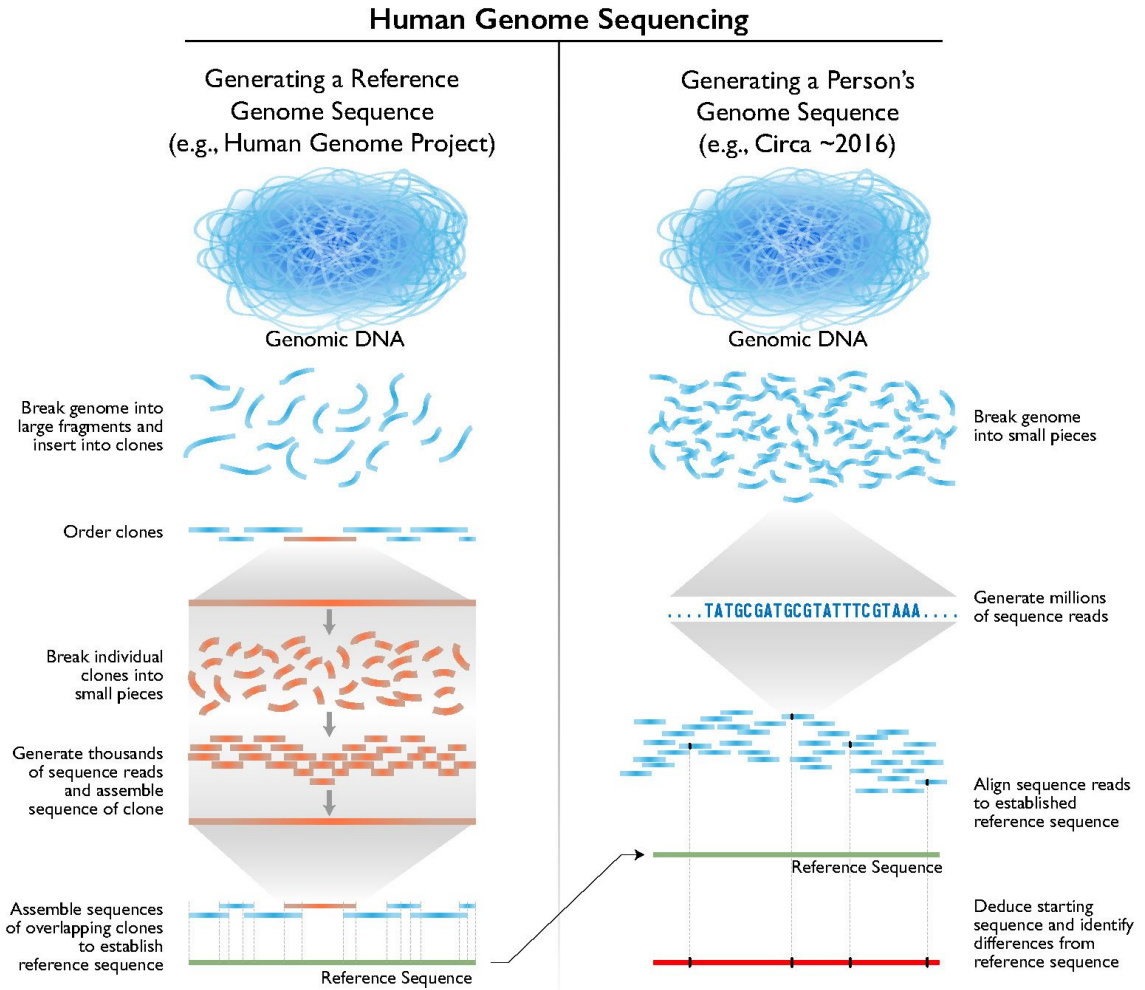


Private

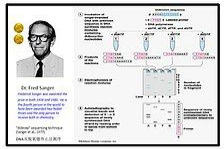
Have to use Computer to assemble the short DNA sequences

Figure 6. Hierarchical shotgun sequencing versus whole genome shotgun sequencing. Both approaches respectively exemplified the methodological rivalry between the public (NIH, **A**) and private (Celera, **B**) efforts to sequence the human genome. Whereas the NIH team believed that whole-genome shotgun sequencing (WGS) was technically unfeasible for gigabase-sized genomes, Venter's Celera team believed that not only this approach was feasible, but that it could also overcome the logistical burden of hierarchical shotgun sequencing, provided that efficient assembly algorithms and sufficient computational power are available. Because of the partial use of NIH data in Celera assemblies, the true feasibility of WGS sequencing for the human genome has been heavily debated by both sides [89, 90].

Recently Human Genome Sequencing

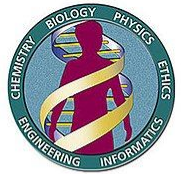


Next Generation Sequencing (NGS)



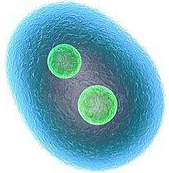
Sanger method

1977



Human Genome Project

1990



Complete eukaryotic genome

1996

Second generation sequencer：454 GS20

2005

Research Human Microbiome Project

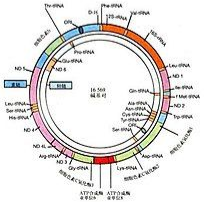
2008

Nanospace sequencing

2014

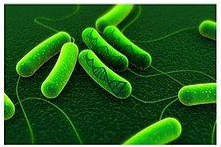
1981

Human mitochondrial genome sequence



1995

Complete cell genome



2001

Complete the Human Genome Project

2007

Second generation sequencer：Genetic Analyzer 2

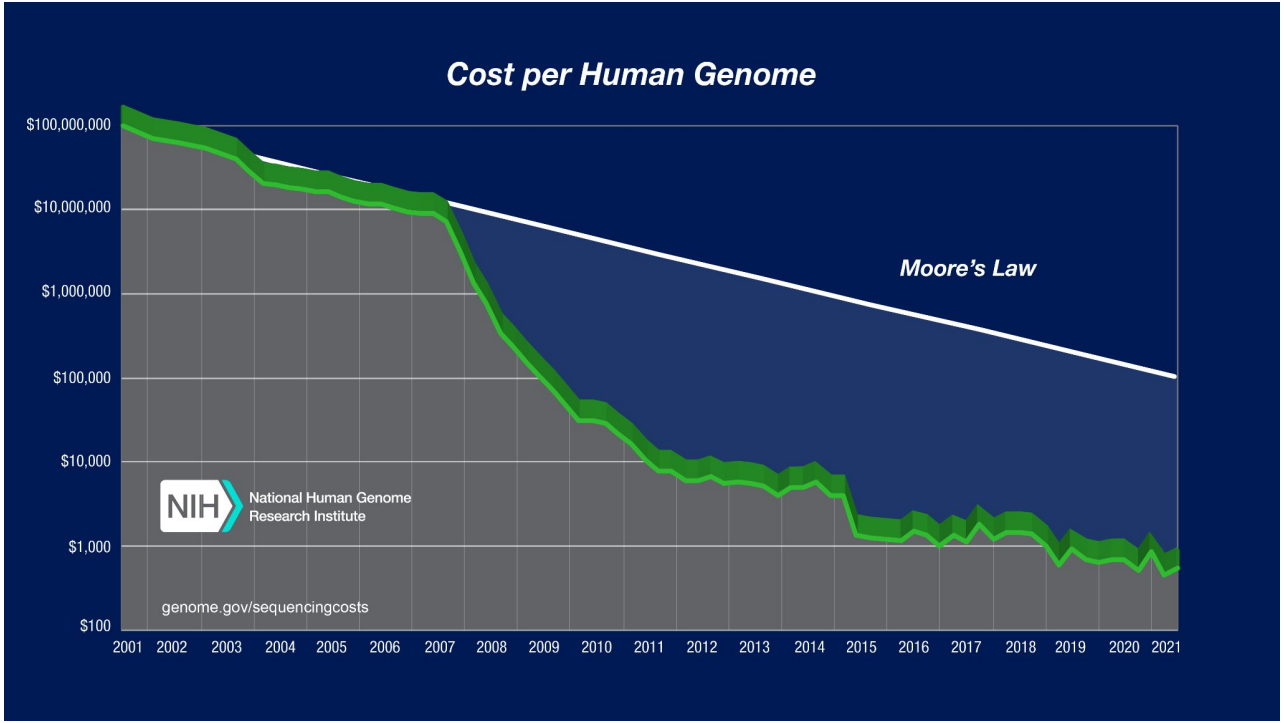
2011

Third generation sequencer：PacBio RS

2019

The third stage Human microbiome project

Cost of Genome Sequencing



Illumina Sequencer: 2023

- Flagship of Sequencing Company

iSeq 100

MiniSeq

MiSeq Series

NextSeq 550 Series

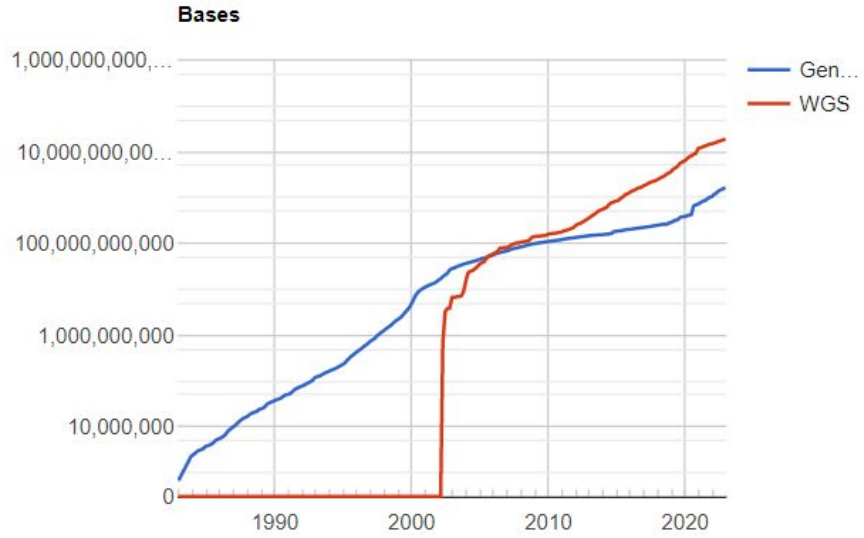
NextSeq 1000 & 2000

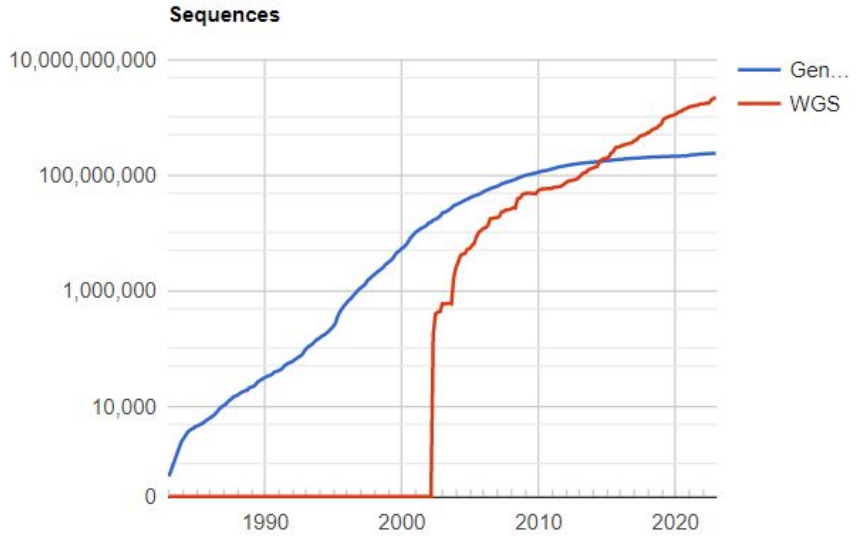
NovaSeq 6000 Series

NovaSeq X Series

	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000	NovaSeq 6000 Series	NovaSeq X Series
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours	~13–38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)	~13–21 hours (1.5B flow cells‡) ~18–24 hours (10B flow cells‡) ~48 hours (25B flow cells‡)
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb *	6000 Gb	16 Tb
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion *	20 billion	26 billion (single flow cells) 52 billion (dual flow cells)
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp	2 x 250 bp**	2 × 150 bp

Big Data Challenge: GenBank and WGS Statistics





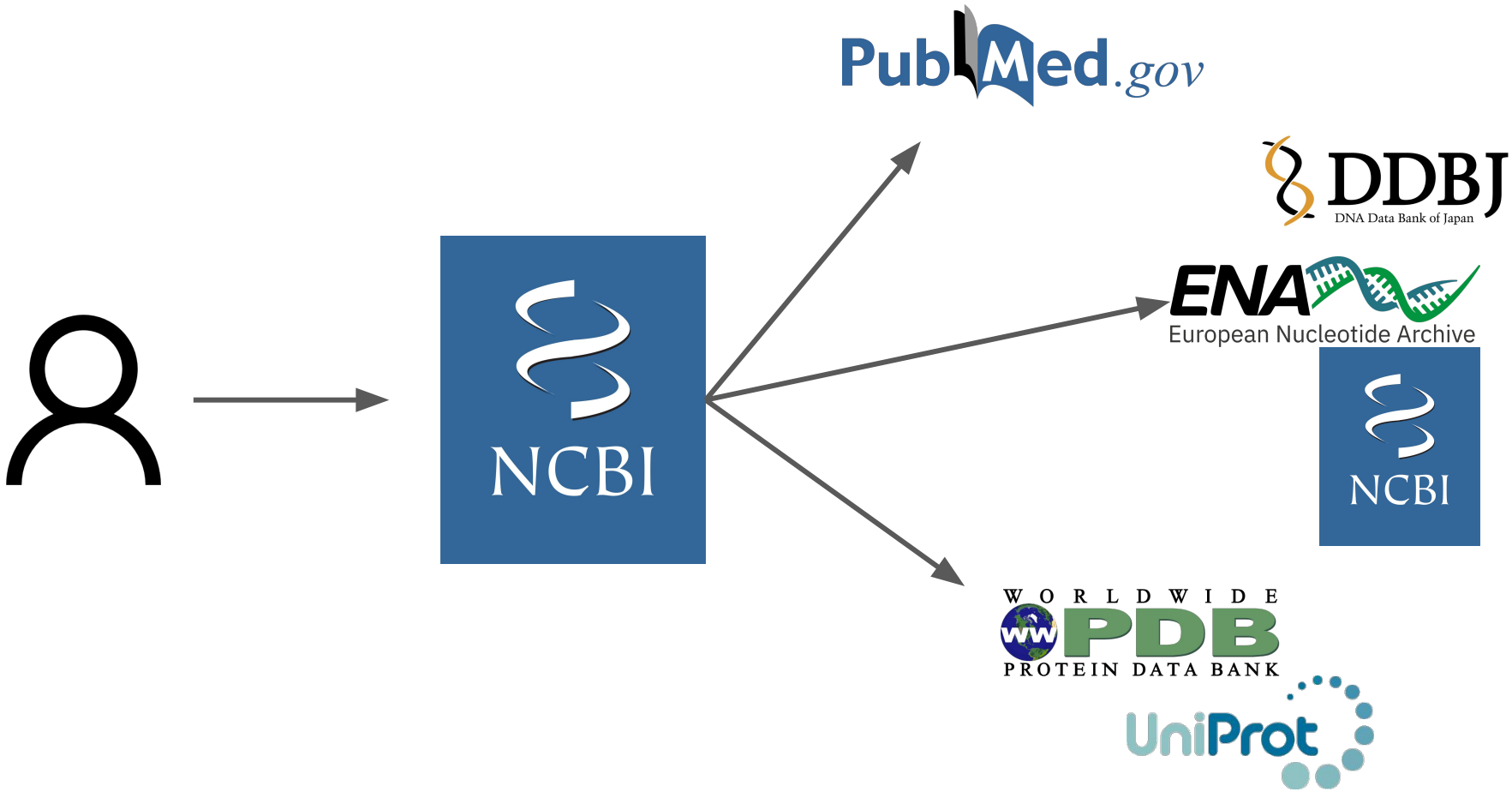
Date	Bases	Sequences	Bases	Sequences
Dec 1982	680338	606	-	-
Apr 2002	19072679701	16769983	692266338	172768
Dec 2022	1635594138493	241015745	19086596616569	2241439349

NCBI Database and Search tools

PSB
Lect. Todsapol Techo

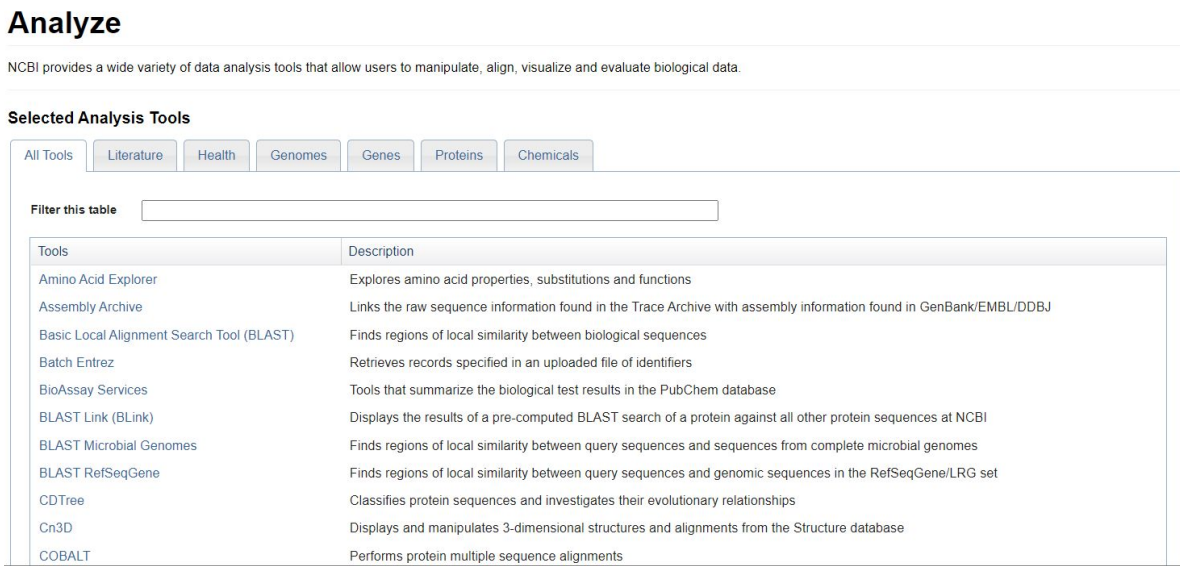
Introduction of Sequence Database

- **The National Center for Biotechnology Information (NCBI)**
  - **Famous Biological Sequence Database**
- Act like a portal to **access all biological database** including DNA, Protein, and etc.
  - **Nucleotide Sequence Database**
    - **Sanger Sequencing** (Old technology)
      - The International Nucleotide Sequence Database Collaboration (INSDC)
        - **GenBank (NCBI)**, **EMBL-EBI**, and **DDBJ**
    - **NGS** (New Technology)
      - **Sequence Read Archive (SRA)**
  - **Protein Sequence and Structure**
    - **Protein Data Bank (PDB)** and **Uniport**
  - **PubMed database**
    - **Abstract** and **Citation** of **scientific articles**
- Provide **web-based tools** for sequence analysis


PubMed.gov
DDBJ
DNA Data Bank of Japan
ENA
European Nucleotide Archive
NCBI
NCBI
WORLDWIDE
PDB
PROTEIN DATA BANK
UniProt

Web-based Tools

- BLAST (Basic Local Alignment Search Tool)
- COBALT (Constraint-based Multiple Alignment Tool)
- Primer-Blast (Primer Design Tools)



Analyze

NCBI provides a wide variety of data analysis tools that allow users to manipulate, align, visualize and evaluate biological data.

Selected Analysis Tools

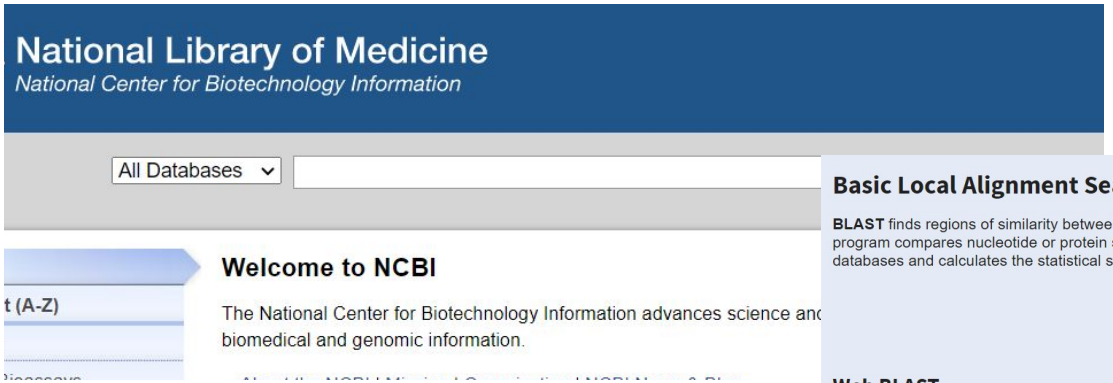
All Tools | Literature | Health | Genomes | Genes | Proteins | Chemicals

Filter this table

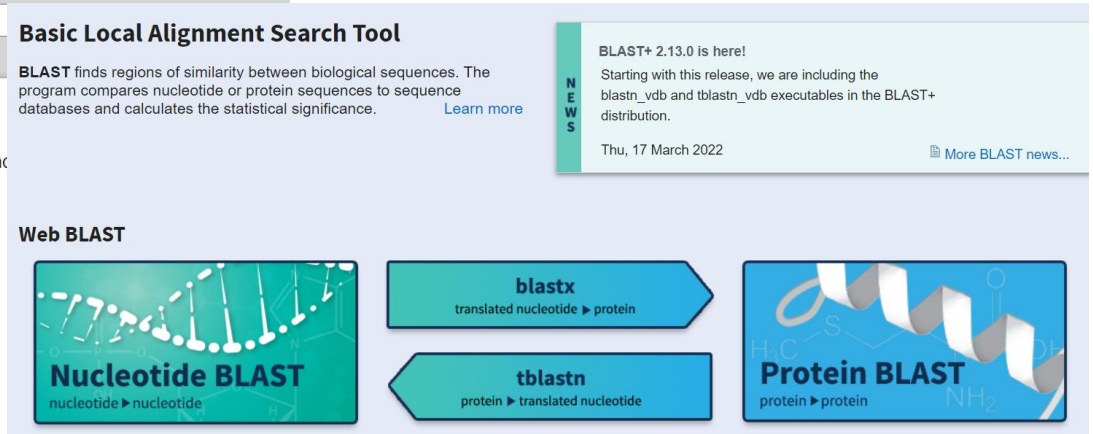
Tools	Description
Amino Acid Explorer	Explores amino acid properties, substitutions and functions
Assembly Archive	Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DDBJ
Basic Local Alignment Search Tool (BLAST)	Finds regions of local similarity between biological sequences
Batch Entrez	Retrieves records specified in an uploaded file of identifiers
BioAssay Services	Tools that summarize the biological test results in the PubChem database
BLAST Link (BLink)	Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI
BLAST Microbial Genomes	Finds regions of local similarity between query sequences and sequences from complete microbial genomes
BLAST RefSeqGene	Finds regions of local similarity between query sequences and genomic sequences in the RefSeqGene/LRG set
CDTree	Classifies protein sequences and investigates their evolutionary relationships
Cn3D	Displays and manipulates 3-dimensional structures and alignments from the Structure database
COBALT	Performs protein multiple sequence alignments

Search Systems in NCBI

- **To Find the interesting things** in NCBI-related database (Mostly Gene)
- **Two Main Search Systems on NCBI**
  - **Entrez** search systems (**Like Google**)
  - **B**asic **L**ocal **A**lignment **S**earch **T**ool (**BLAST**): sequence search



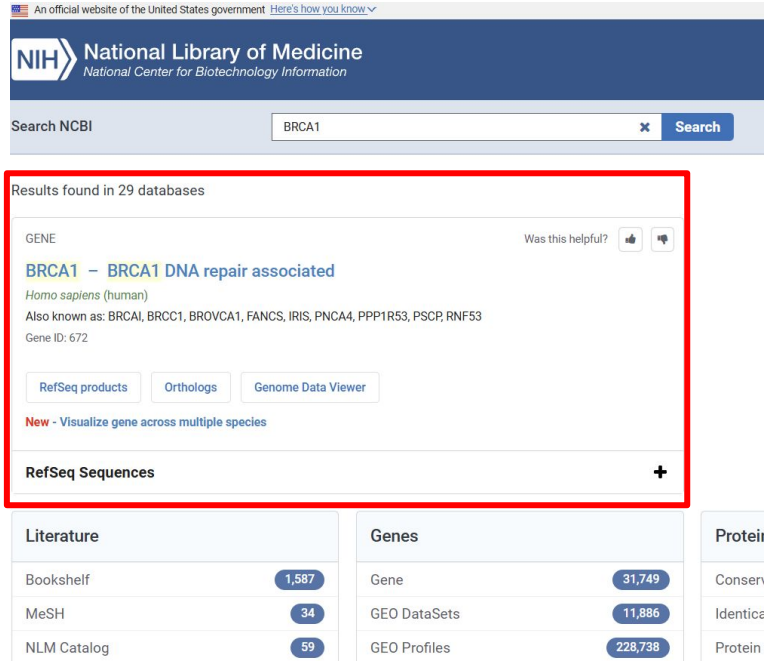
Entrez



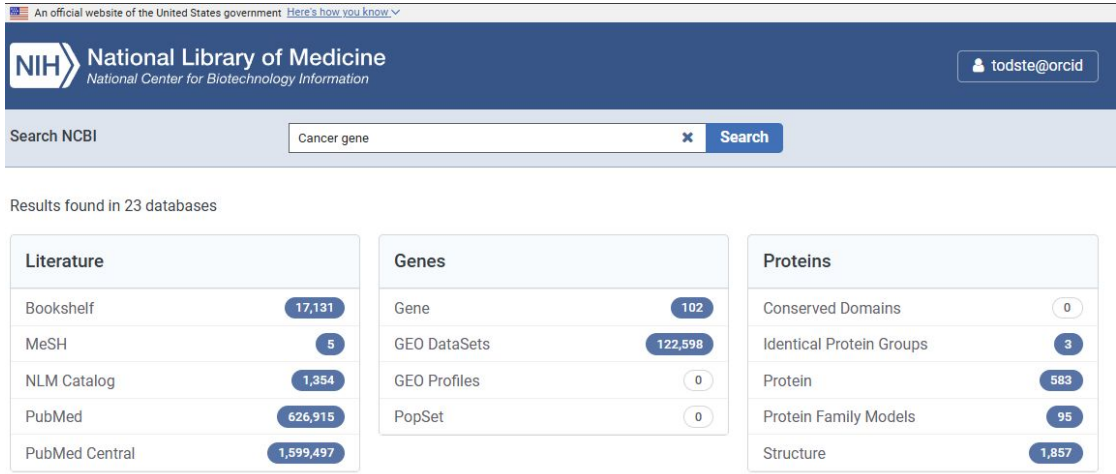
BLAST

Entrez Search System

- General Search to obtain Interesting Information from NCBI
- Narrow down your interesting information
- Usually start with the name of interesting gene, protein, or etc.



First Example: Search Well-Known Gene to obtain DNA, mRNA, and Protein sequence



Search Results in All Databases

Literature	
Bookshelf	1,587
MeSH	34
NLM Catalog	59
PubMed	23,465
PubMed Central	84,283

Genes	
Gene	31,749
GEO DataSets	11,886
GEO Profiles	228,738
PopSet	218

Proteins	
Conserved Domains	112
Identical Protein Groups	5,778
Protein	36,730
Protein Family Models	239
Structure	236

Genomes	
Assembly / Genome NCBI Datasets	
BioCollections	0
BioProject	650
BioSample	15,719
Nucleotide	49,584
SRA	10,840
Taxonomy	0

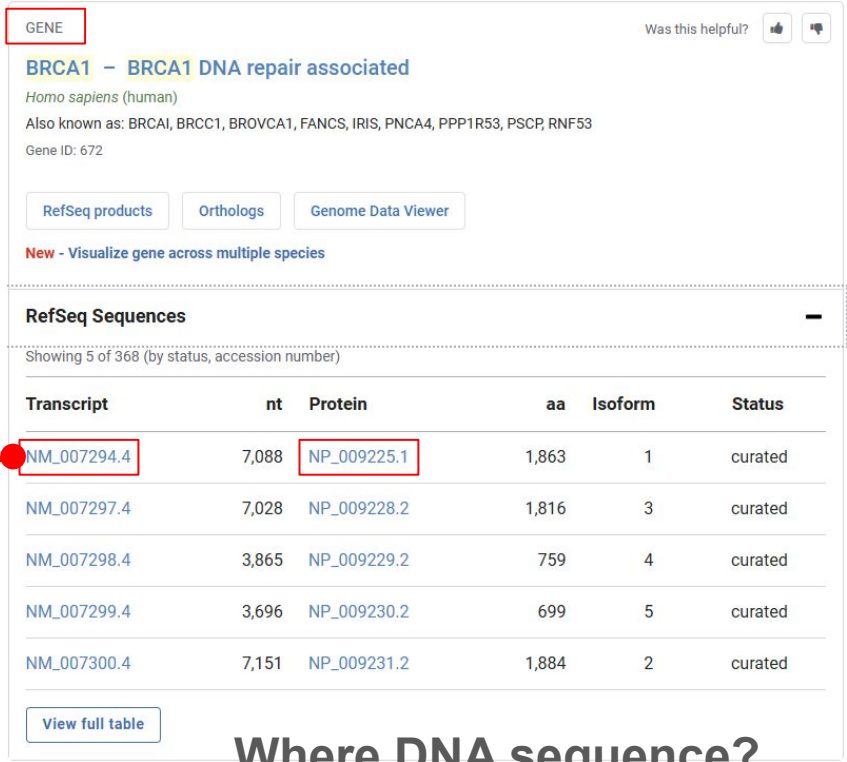
Clinical	
ClinicalTrials.gov	0
ClinVar	61,317
dbGaP	11
dbSNP	41,177
dbVar	7,744
GTR	657
MedGen	48
OMIM	194

PubChem	
BioAssays	1,255
Compounds	2
Pathways	3
Substances	152

NCBI Datasets All assembled genome data is now available through NCBI Datasets

Get mRNA and Protein sequences

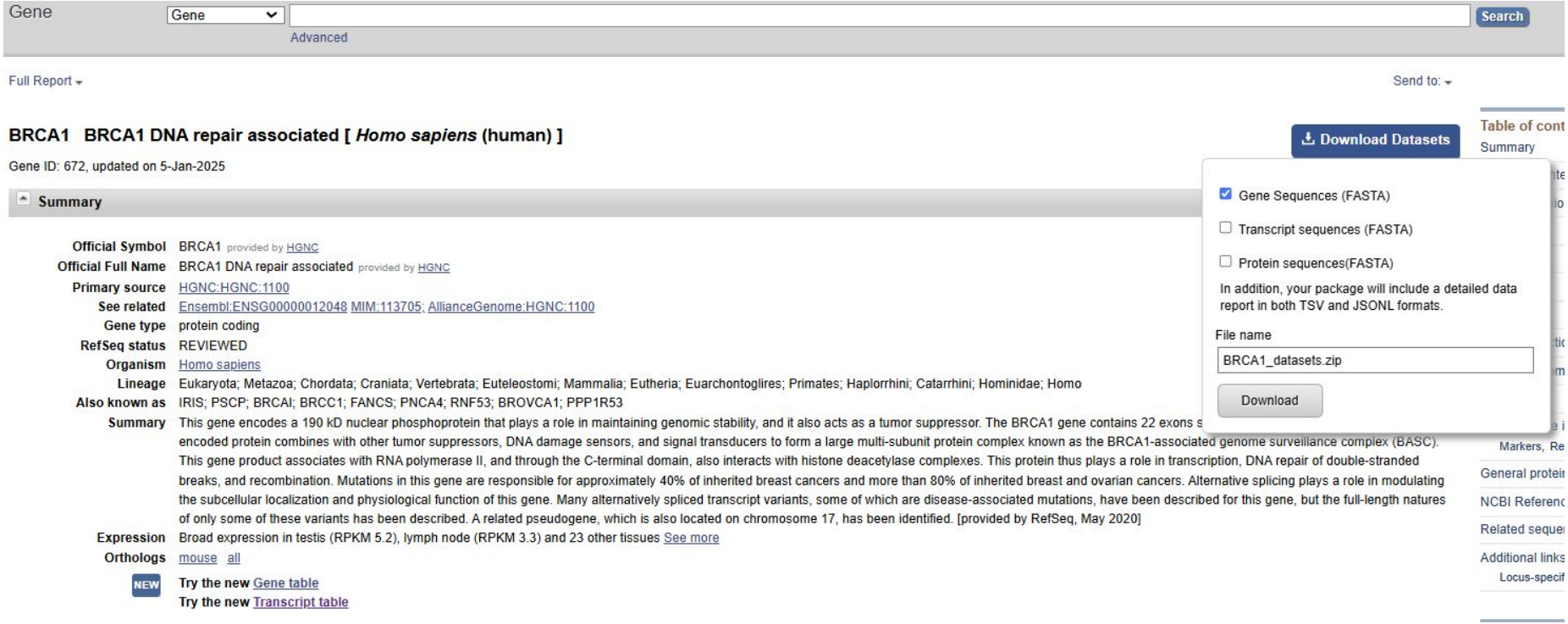
- **Get BRCA1 sequences**
  - **From Gene Database**
  - mRNA sequence
    - Mature mRNA sequence (**NM**)
  - Protein sequence
    - Related to mRNA sequence (**NP**)



Where DNA sequence?

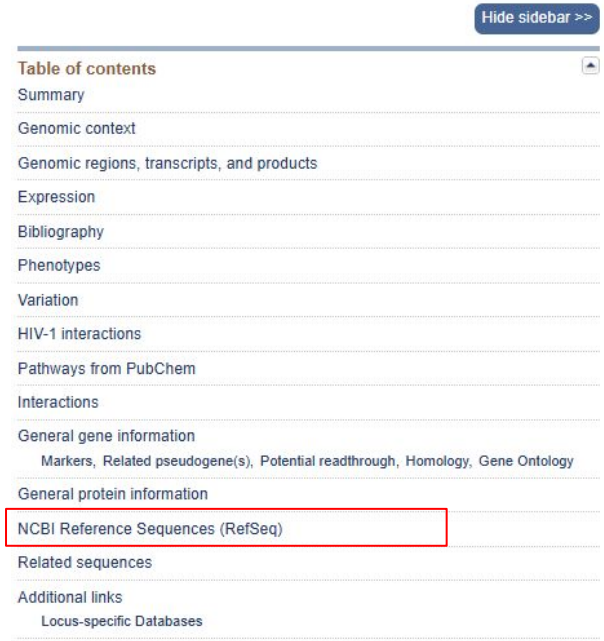
Get DNA sequence

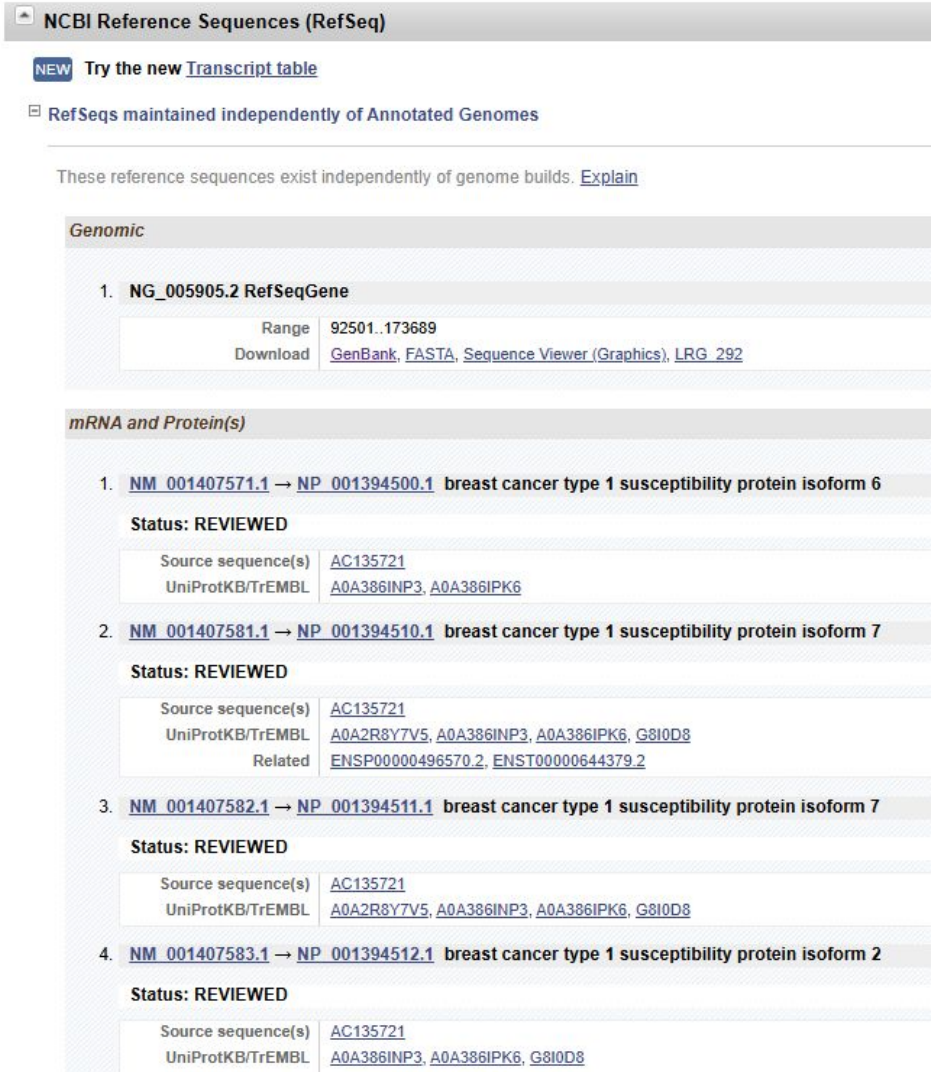
- From Gene Database



Get DNA sequence

- From Gene Database
  - To obtain the specific isoform or else





Reference Sequence Database (RefSeq Database)

- **Model RefSeq** : RNA and protein products that are generated by the **eukaryotic genome annotation pipeline (Prediction)**. These records use accession prefixes XM_, XR_, and XP_.
- **Known RefSeq** : RNA and protein products are mainly derived from GenBank cDNA and EST data and are supported by the RefSeq eukaryotic curation group **(From real experiment)**. These records use accession prefixes NM_, NR_, and NP_..
  - NM_001360016.2: **Transcript**
  - NR_: **Non coding RNA**
  - NP_001346945.1: **Protein**
  - NG_009015.2: **Genomic regions (Gene) or Segment in Chromosome**
  - NC_000023.11: **Chromosome (Assembled contig)**

Table 1.

RefSeq accession numbers and molecule types.

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NZ_[b]	Genomic	Complete genomes and unfinished WGS data
NM_	mRNA	Protein-coding transcripts (usually curated)
NR_	RNA	Non-protein-coding transcripts
XM_[c]	mRNA	Predicted model protein-coding transcript
XR_[c]	RNA	Predicted model non-protein-coding transcript
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	Annotated on genomic molecules without an instantiated transcript record
XP_[c]	Protein	Predicted model, associated with an XM_ accession
WP_	Protein	Non-redundant across multiple strains and species

[a] Whole Genome Shotgun sequence data.

[b] An ordered collection of WGS sequence for a genome.

[c] Computed.

FASTA

Homo sapiens glutathione synthetase (GSS), RefSeqGene (LRG_1168) on chromosome 20

NCBI Reference Sequence: NG_008848.2

GenBank Graphics

```
>NG_008848.2:5222-32596 Homo sapiens glutathione synthetase (GSS), RefSeqGene
(LRG_1168) on chromosome 20
GTCGCATCGAGGCCCCGCCCCCTGAGCCTGGGTAGCGGCGCGAGGGCCGGGAGAACCGTTCGCGGAGGAA
AGGCGAACTAGTAGGTTGGGGCGGCCACGGCGGCCGGCATGGGTCACGTTTCCTCGGGAGGAACGATGTG
AGGGAGGGGTCTGGCAAGAGATTGGAATTCCGGAGGCCGGGAGACCTTGTGGCTGAAACCCTTCGTAGGA
GCGGGGCAACTAGTGTCTAGTGAGGGGGTTGGGCTGGCGCGCACTGATCCCAGACTTTCCGGATCTTCTG
```

FASTA

glutathione synthetase [Homo sapiens]

NCBI Reference Sequence: NP_000169.1

GenPept Identical Proteins Graphics

```
>NP_000169.1 glutathione synthetase [Homo sapiens]
MATNWGSLLQDKQQLEELARQAVDRALAEGVLLRTSQEPTSSEVVSYAPFTLFPSLVPSALLEQAYAVQM
DFNLLVDAVSQNAAFLEQTLSSTIKQDDFTARLFDIHKQVLKEGIAQTVFLGLNRSDYMFQRSADGSPAL
KQIEINTISASFGGLASRTPAVHRHVLSVLSKTKEAGKILSNNPSKGLALGIAKAWELYGSPNALVLLIA
QEKERNIFDQRAIENELLARNIHVIRRTFEDISEKGSLDQDRRLFVDGQEIAVVYFRDGYMPRQYSLQNW
EARLLLERSHAAKCPDIATQLAGTKKVQQELSRPGMLEMLLPGQPEAVARLRATFAGLYSLDVGEEGDQA
IAEALAAPSRFVLKPQREGGGNNLYGEEMVQALKQLKDSEERASYILMEKIEPEPFENCLLRPGSPARVV
QCISELGIFGVYVRQEKTLVMNKHVGHLLRTKAIEHADGGVAAGVAVLDNPYPV
```

FASTA

Homo sapiens glutathione synthetase (GSS), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000178.4

GenBank Graphics

```
>NM_000178.4 Homo sapiens glutathione synthetase (GSS), transcript variant 1, mRNA
GTCGCATCGAGGCCCCGCCCCCTGAGCCTGGGTAGCGGCGCGAGGGCCGGGAGAACCGTTCGCGGAGGAA
AGGCGAACTAGTGTTGGGATGGCCACCAACTGGGGGAGCCTCTTGCAGGATAAACAGCAGCTAGAGGAGC
TGGCACGGCAGGCCGTGGACCGGGCCCTGGCTGAGGGAGTATTGCTGAGGACCTCACAGGAGCCCACTTC
CTCGGAGGTGGTGAGCTATGCCCCATTCACGCTCTTCCCCTCACTGGTCCCCAGTGCCCTGCTGGAGCAA
GCCTATGCTGTGCAGATGGACTTCAACCTGCTAGTGGATGCTGTCAGCCAGAACGCTGCCTTCCTGGAGC
AAACTCTTTCCAGCACCATCAAACAGGATGACTTTACCGCTCGTCTCTTTGACATCCACAAGCAAGTCCT
AAAAGAGGGCATTGCCCAGACTGTGTTCCTGGGCCTGAATCGCTCAGACTACATGTTCCAGCGCAGCGCA
GATGGCTCCCCAGCCCTGAAACAGATCGAAATCAACACCATCTCTGCCAGCTTTGGGGGCCTGGCCTCCC
GGACCCCAGCTGTGCACCGACATGTTCTCAGTGTCCTGAGTAAGACCAAAGAAGCTGGCAAGATCCTCTC
TAATAATCCCAGCAAGGGACTGGCCCTGGGAATTGCCAAAGCCTGGGAGCTCTACGGCTCACCCAATGCT
CTGGTGCTACTGATTGCTCAAGAGAAGGAAAGAAACATATTTGACCAGCGTGCCATAGAGAATGAGCTAC
TGGCCAGGAACATCCATGTGATCCGACGAACATTTGAAGATATCTCTGAAAAGGGGTCTCTGGACCAAGA
CCGAAGGCTGTTTGTGGATGGCCAGGAAATTGCTGTGGTTTACTTCCGGGATGGCTACATGCCTCGTCAG
TACAGTCTACAGAATTGGGAAGCACGTCTACTGCTGGAGAGGTCACATGCTGCCAAGTGCCCAGACATTG
CCACCCAGCTGGCTGGGACTAAGAAGGTGCAGCAGGAGCTAAGCAGGCCGGGCATGCTGGAGATGTTGCT
CCCTGGCCAGCCTGAGGCTGTGGCCCGCCTCCGCGCCACCTTTGCTGGCCTCTACTCACTGGATGTGGGT
GAAGAAGGGGACCAGGCCATCGCCGAGGCCCTTGCTGCCCCTAGCCGGTTTGTGCTAAAGCCCCAGAGAG
AGGGTGGAGGTAACAACCTATATGGGGAGGAAATGGTACAGGCCCTGAAACAGCTGAAGGACAGTGAGGA
GAGGGCCTCCTACATCCTCATGGAGAAGATCGAACCTGAGCCTTTTGAGAATTGCCTGCTACGGCCTGGC
AGCCCTGCCCGAGTGGTCCAGTGCATTTCAGAGCTGGGCATCTTTGGGGTCTATGTCAGGCAGGAAAAGA
CACTCGTGATGAACAAGCACGTGGGGCATCTACTTCGAACCAAAGCCATCGAGCATGCAGATGGTGGTGT
GGCAGCGGGAGTGGCAGTCCTGGACAACCCATACCCTGTGTGAGGGCACAACCAGGCCACGGGACCTTCT
ATCCTCTGTATTTGTCATTCCTCTCCTAGCCCTCCTGAGGGGTATCCTCCTAAAGACCTCCAAAGTTTTT
ATGGAAGGGTAAATACTGGTACCTTCCCCCAGCTTTCCATCTGAGGACCAGAAAAGTTGTGTCTCCCTTA
GATGAGATCTAGACGCCCCCAAATCCTTGAGATGTGGGTATAGCTCAGGGTAAGCTGCTCTGAGGTAAAG
GTCCATGAACCCTGCCCCACTCCTGTCAGCCCCTCATCAGCCTTTTCAGCAGGTTCCAGTGCCTGACTTG
GGATAGGACTGAGTGGTAGGAGGAGGGGGAGTGGAGGGGCATAGCCTTTCCCTAATTCTGCCTTAAATAA
AACTGCATTGCTGATTCAG
```

Entrez Search Options

- Using Boolean Operators
  - **AND**, **OR**, **NOT**
  - human **AND** mitochondrion
  - human **OR** mitochondrion
  - human **NOT** mitochondrion
- Indexed Fields
  - Human**[organism]**
  - Mitochondrion**[title]**
  - prolactin**[Protein Name]**
- Search builder

https://www.ncbi.nlm.nih.gov/books/NBK3837/

BLAST

Basic Local Alignment Search Tool

BLAST system

- Search tool for **interesting sequence**
- Require **sequences** as the input to search

```
GAACTCGGGAAGCCGGCGAGAAGTGTGAGGCCGCGG
TAGGGCCGCATCCCGCTCCGGAGAGAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCCGCGGAAGCTCGGTAA
TGATAAGCACGCCGGCCACTTTGCAGGGCGTCAC
CGCCTACACGCCCCCTCGTCTCTCGGACGGCGGCGT
CTAGCCTCGGGGCGCTCGGCCGCCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCCAATAGGGC
CGGCTTGACCCGCGAACAGGCGAGGGTTCCCGGG
GGAGTGGCGCGGCAGAAGGCCCCGCCCAGGAGCCGA
GGGACAGCCCAGAGGAGGCGTGGCCACGCTGCCG
GCGGAAGTGGAGCCCTCCGCGAGCGCGCGAGGCCGC
CGGGGCAGGCGGGGAAACCGGACAGTAGGGGCGG
GGCCGGGCCGGCGATGGGGATGCGGGAGCACTACGC
GGAGCTGCACCCGTGCCCGCCGGAATTGGGGATG
CAGAGCAGCGGCAGCGGGTATGGCAGGCAGCCGGCG
GGCCGGCCTCCAGCGCAGGTGCCCGAGAGGCAGG
GGCTGGCCTGGGATGCGCGCGCACCTGCCCTCGCCC
CGCCCCGCCCGCACGAGGGGTGGTGGCCGAGGCC
CCGCCCCGCACGCCTCGCCTGAGGCGGGTCCGCTC
```

```
MGRRGSAPGNGRTLRGCERGGRRRRSADSVMAEQ
VALSRTQVCGILREELFQGDAFHQSDTHIFIIMG
ASGDLAKKKIYPTIWWLFRDGLLPENTFIVGYAR
SRLTVADIRKQSEPFFKATPEEKLKLEDFFARNS
YVAGQYDDAASYQRLNSHMNALHLGSQANRLFYL
ALPPTVYEAVTKNIHESCMSQIGWNRIIVEKPFG
RDLQSSDRLSNHISSLFREDQIYRIDHYLGKEMV
QNLMVLRFANRIFGPIWNRDNIACVILTFKEPFG
TEGRGYFDEFGIIRDVMQNHLLQMLCLVAMEKPA
STNSDDVRDEKVKVLKCISEVQANNVVLGQYVGN
PDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVE
NERWDGVPFILRCGKALNERKAEVRLQFHDVAGD
IFHQQCKRNELVIRVQPNEAVYTKMMTKKPGMFF
NPEESELDLTYGNRYKNVKLPDAYERLILDVFCG
SQMHFVRSDELREAWRIFTPLLHQIELEKPKPIP
YIYGSRGPTEADELMKRVGFQYEGTYKWVNPHKL
```

```
>NC_000023.11:c154547569-154531390
Homo sapiens chromosome X,
GRCh38.p14 Primary Assembly
GAACTCGGGAAGCCGGCGAGAAGTGTGAGGCCGCGG
TAGGGCCGCATCCCGCTCCGGAGAGAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCCGCGGAAGCTCGGTAA
TGATAAGCACGCCGGCCACTTTGCAGGGCGTCAC
CGCCTACACGCCCCCTCGTCTCTCGGACGGCGGCGT
CTAGCCTCGGGGCGCTCGGCCGCCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCCAATAGGGC
CGGCTTGACCCGCGAACAGGCGAGGGTTCCCGGG
GGAGTGGCGCGGCAGAAGGCCCCGCCCAGGAGCCGA
GGGACAGCCCAGAGGAGGCGTGGCCACGCTGCCG
GCGGAAGTGGAGCCCTCCGCGAGCGCGCGAGGCCGC
CGGGGCAGGCGGGGAAACCGGACAGTAGGGGCGG
```

Fasta Format

- **Fasta:** Simple Formats for Sequences
  - filename.**fasta**, filename.**fa**, filename.**fna**, and etc.

Header

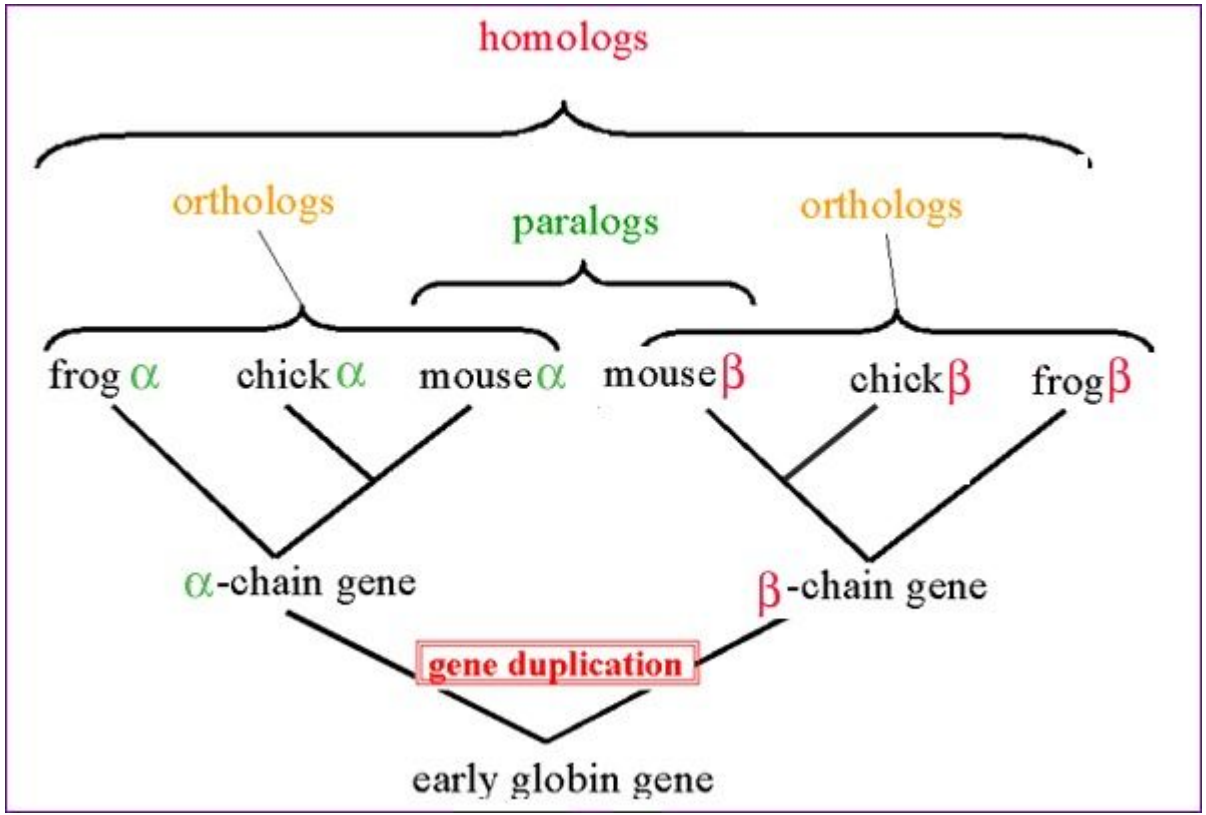
```
>NC_000023.11:c154547569-154531390 Homo sapiens chromosome X, GRCh38.p14 Primary Assembly
GAACTCGGGAAGCCGGCGAGAAGTGTGAGGCCGCGGTAGGGCCGCATCCCGCTCCGGAGAGAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCCGCGGAAGCTCGGTAATGATAAGCACGCCGGCCACTTTGCAGGGCGTCAC
CGCCTACACGCCCCCTCGTCTCTCGGACGGCGGCGTCTAGCCTCGGGGCGCTCGGCCGCCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCCAATAGGGCCGGCTTGACCCGCGAACAGGCGAGGGTTCCCGGG
GGAGTGGCGCGGCAGAAGGCCCCGCCCAGGAGCCGAGGGACAGCCCAGAGGAGGCGTGGCCACGCTGCCG
GCGGAAGTGGAGCCCTCCGCGAGCGCGCGAGGCCGCCGGGGCAGGCGGGGAAACCGGACAGTAGGGGCGG
```

Sequence

```
>Seq1 [organism=Carpodacus mexicanus] [clone=6b] actin (act) mRNA, partial cds
CCTTTATCTAATCTTTGGAGCATGAGCTGGCATAGTTGGAACCGCCCTCAGCCTCCTCATCCGTGCAGAATAATAATTTTCTTTATAGTAATACCAATCATGATCGGTGGT
TTCGGAAACTGACTAGTCCCACTCATAAT
>Seq2 [organism=uncultured bacillus sp.] [isolate=A2] corticotropin (CT) gene, complete cds
GGTAGGTACCGCCCTAAGNCTCCTAATCCGAGCAGAACTANGCCAACCCGGAGCCCTTCTGGGAGACGACTCAACACCACCTTCTTTGACCCAGCAGGAGGAGGAGACCCA
GTACTATACCAGCACCTATTCTGATTCTT
>Seq3 [organism=Phalaenopsis equestris var. leucaspis]
CCTATACCTAATTTTCGGCGCATGAGCCGGAATGGTGGGTACCGCTCTAAGCCTCCTCATTCGAGCAGAACTAGGCCAACCCGGAGCCCTTCTGGGAGACGACCAAGTCTA
CAACGTGGTTGTCACGGCCCATGCCTTCG
>Seq9 [organism=Petunia integrifolia subsp. inflata]
TAGTTGGAACAGCCCTCAGCCTACTCATCCGAGCAGAACTAGGCCAACCCGGAACCCTCCTGGGAGATGACCAAATCTACAATGTAATCGTCACTGCCCATGCCTTCGTAA
TAATCTTCTTCATAGTAATACCAGTCATA
```

Homology

- **Homologs**
  - Homologous biological components
- **Orthologs**
  - Same function different species arose from ancestor
- **Paralogs**
  - Same species arose from duplication in ancestor

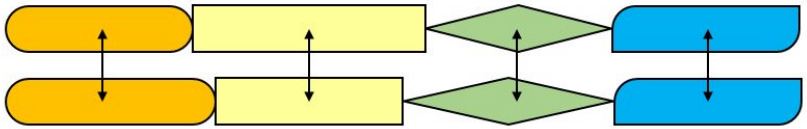


BLAST

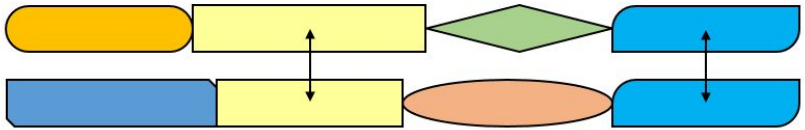
- BLAST tells you about **Non-chance similarities between biological sequences**
- **If similarity do not due to chance** then **they must be due to something else**
  - **Homology**: Similar sequences -> evolutionary relationship -> related function
  - **Identification** (Present or absent in database)
- When you have a sequence, the possible questions are
  - What is relate to?
    - **Homology**: **Orthologs** and **Paralogs**
    - **Function** of the sequences
  - Is it new or already in database?
    - Find the match sequences in database

Search Concept of BLAST

- **Compare Two sequences**
  - **Pairwise Alignment**
- **Looking for Similarity** between **two** sequences
  - **Query** (our interesting sequence)
  - **Subject** (sequences in database)
- **Two types of Pairwise Alignment**
  - **Global alignment** try to map all part of sequence from query to subject.
    - Needleman-Wunsch
    - **Most Similar Lengths**
  - **Local alignment** try to map some part of sequences in query to subject.
    - Smith-Waterman
    - **Different Lengths**
    - **BLAST**



Global Alignment



Local Alignment

Global Alignment and Local Alignment Examples

SEQUENCE 1 ATGACTTTACCAGGTGGTGTTATTTTAGTTTTCATTTTGGTCGGTTTGGCTTGTATTGCCATTATT

SEQUENCE 2 ATGATTTTGTCAGGTACCATTGCT

Global alignment

```
1  ATGACTTTACCAGGTGGTGTTATTTTAGTTTTCATTTTGGTCGGTTTGGCTTGTATTGCCATTATT  66
   |||| |||. . |||| |. .|| ||  |||
2  ATGATTTTGTCAGG--------------TACCATT-----------GCT---------------  24
```

Local alignment

```
1  ATGACTTTACCAGGT  15
   |||| |||. .|||||
1  ATGATTTTGTCAGGT  15
```

Simple Identical Scoring Systems

ATCG
ATTG
1+1+0+1=3

- **The system to quantify the similarity between sequence**

AT – C G
AT T - G
1+1+0+0+1=3

$$id(x, y) = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases}$$

Gap Penalties

- Open Gap = -2
- Gap = -1

$$id(x, y) = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases}$$

ATCG
ATTG

1+1+0+1=3

AT – C G
AT T - G

1+1-2-1-2-1+1=-3

ATC - - T A
ATT T T TA

1+1+0-2-1-1+1+1=0

AT - C - T A
AT T T T TA

1+1-2-1+0-2-1+1+1=-2

Proteins Scoring Systems

- **Problem with scoring systems in Protein sequence alignment**
- If **V** easy to change to **A** but not **W** to **A, should we still use the same score?**
- If **V** change to **A, then change back to V due to the evolution, should we still use the same score for unchanged V?**

VGK – GI… WGK – GI…

AGKVGL… AGKVGL

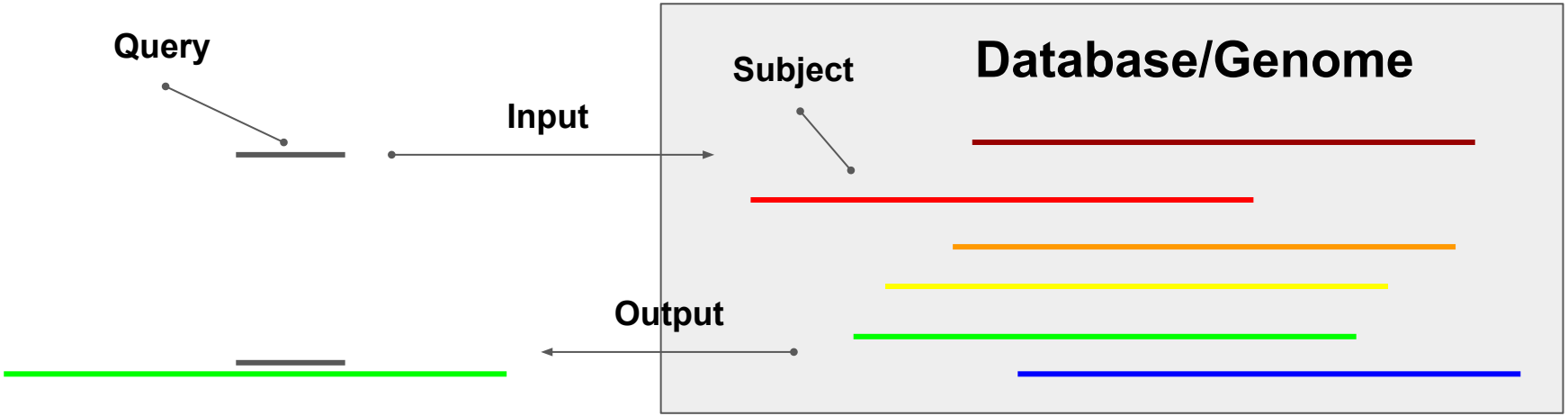
Protein Scoring Systems

- Scoring matrix
  - Substitution matrix
    - BLOSUM
    - PAM

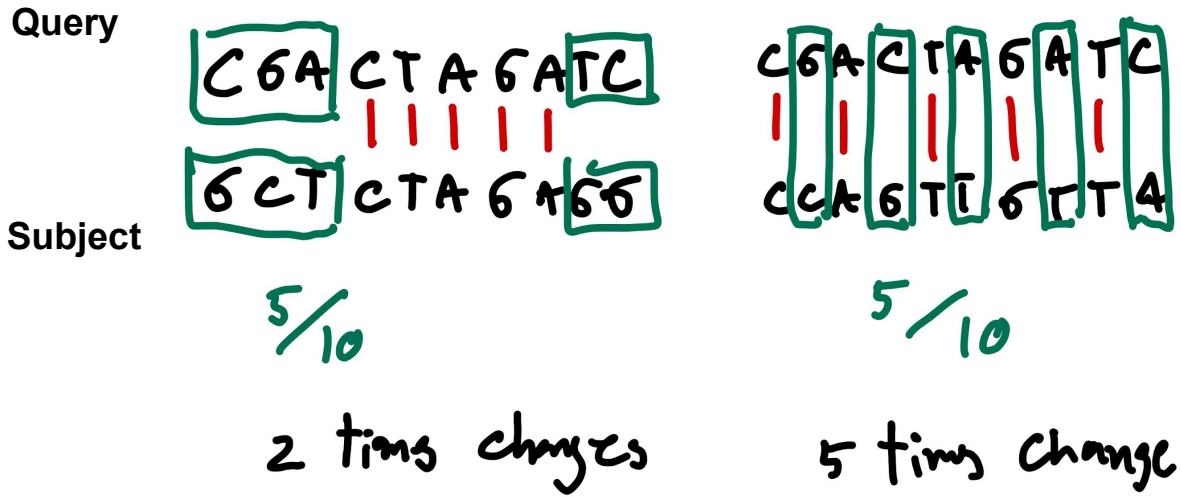
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

BLAST Algorithm

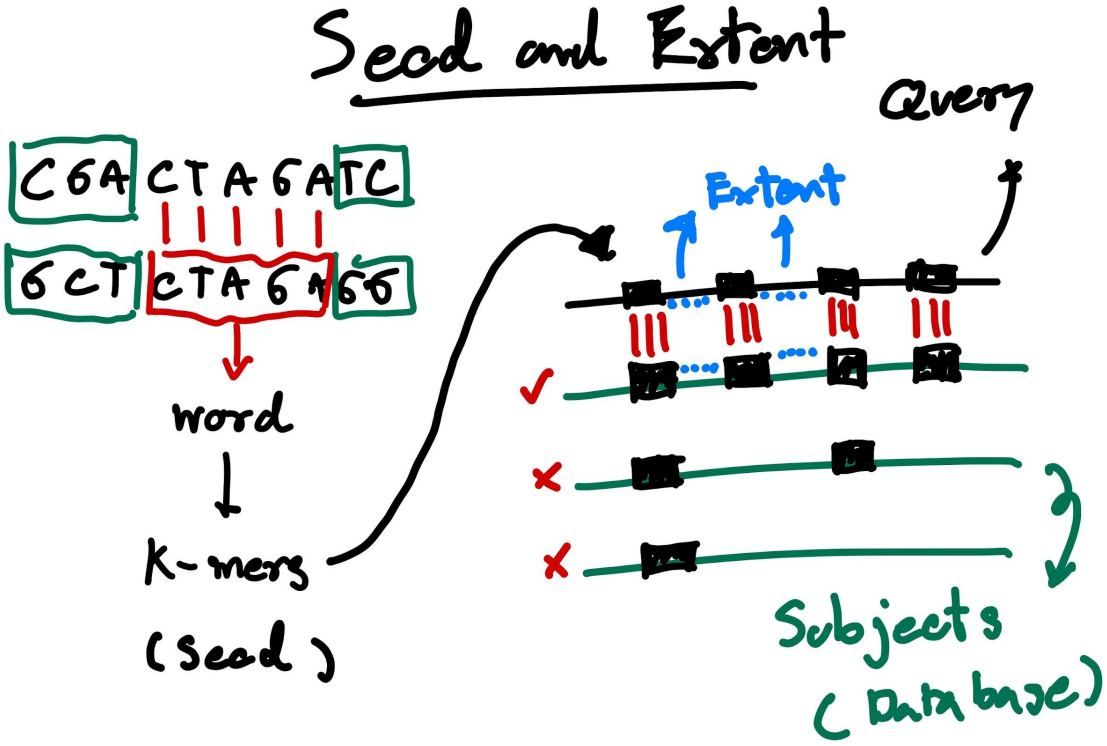
- **Generate words (K-mers) from sequence**
- **Seed and Extent:** Search words and extent scoring in Database
- **Local Alignment:** Smith–Waterman algorithm
- **Statistic Calculation:** Matching by **change** or **not**



Word Searching rather than Global Searching



Seed and Extent



BLAST Scoring Systems

- **Local Alignment** Based on **Smith–Waterman algorithm** for scoring systems
- **High score mean highly similar between 2 sequences**

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$

$W_1 = 2$

Statistical Calculation: Expectation Value (E-value)

- This value tell you about the possibility that **your query sequence similar to subject sequence by chance.**

10^{-16}

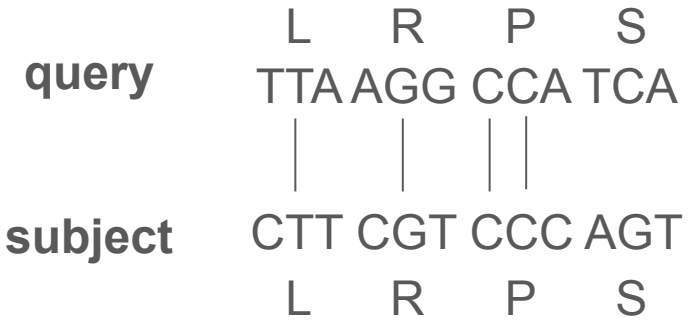
ATCG
ATTG

10^{-2}

AT – C G
AT T - G

Nucleotide and Protein Searching

- Study in evolutionary relationships

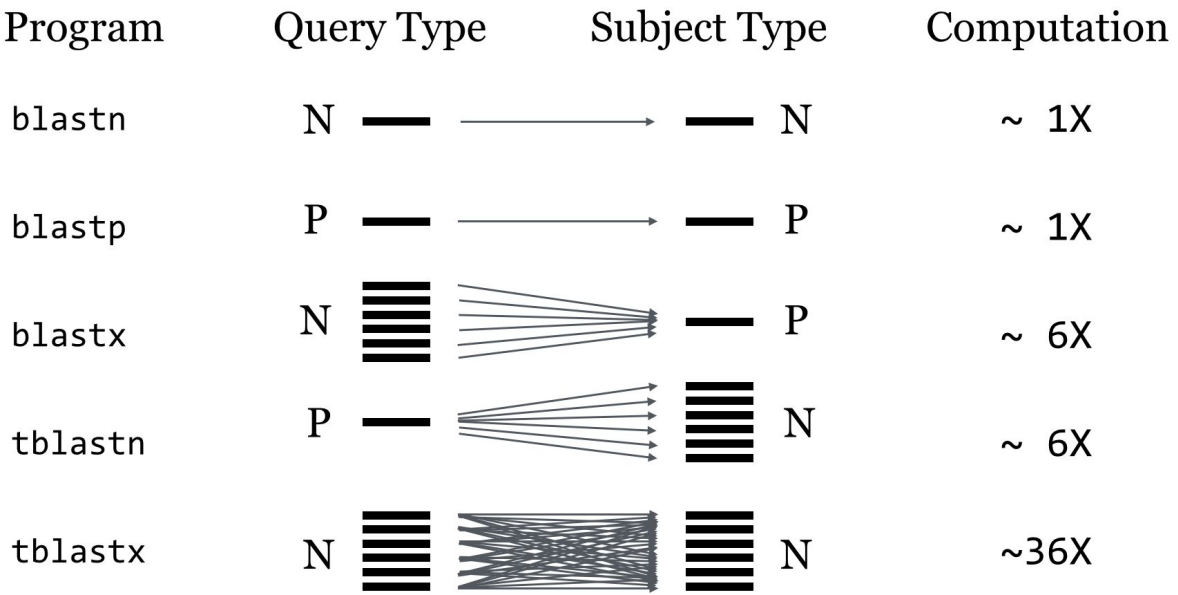


4 match in DNA

all match in Protein

BLAST Types

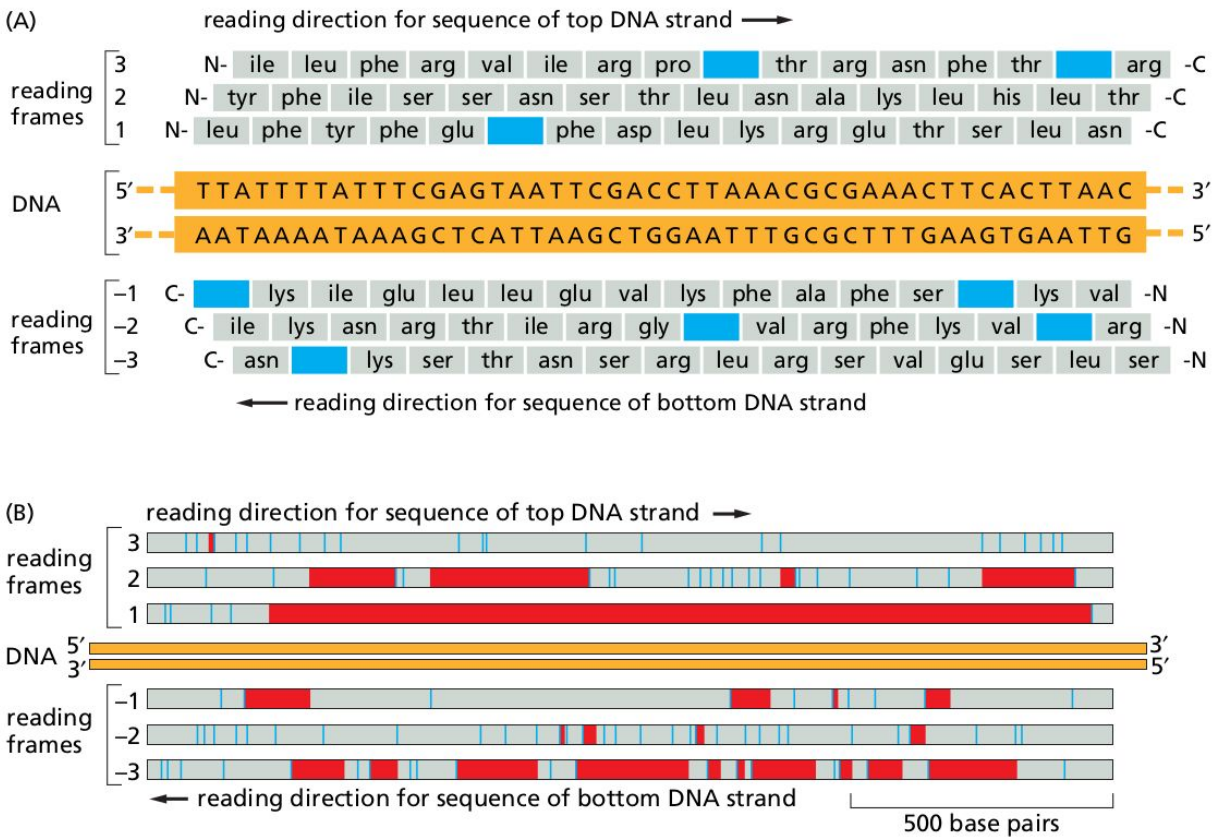
- **Basic Search**
  - blastn
  - blastp
- **Translated Search**
  - blastx
  - tblastn
  - tblastx
  - **6 Reading Frames**



(other BLAST types not listed: `psiblast`, `deltablast`, `rpsblast`)

Reading Frame

- 6 frames per sequences



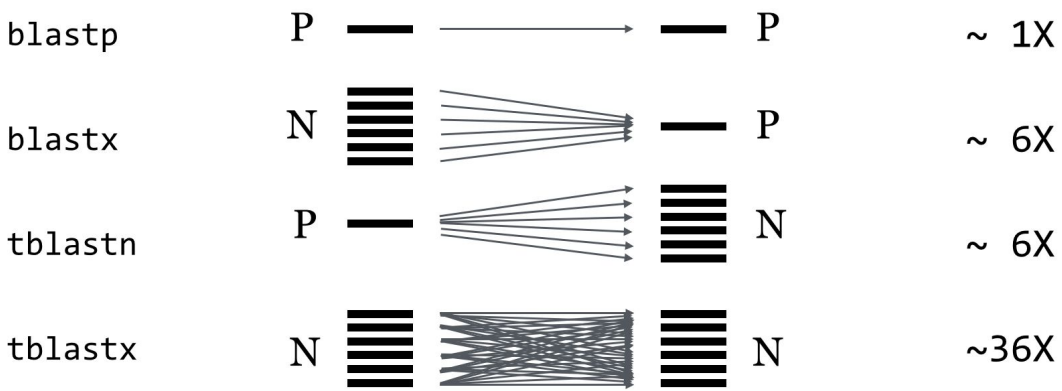
blastn

- **blastn**
  - **Small Word Size**
  - Traditional blast algorithm
- **megablast (general option)**
  - **Large word size**
  - Different gaping models
    - **Contiguous megablast**
      - Nearly identical sequences (same species)
    - **Discontiguous megablast**
      - Cross-species sequences (allow substitution of letter)

blastp

- **blastp**
  - **3-word**
- Translating search
  - blastx (annotate sequence)
  - tblastn
  - tblastx

G	Glycine	Gly	**P**	Proline	Pro
A	Alanine	Ala	**V**	Valine	Val
L	Leucine	Leu	**I**	Isoleucine	Ile
M	Methionine	Met	**C**	Cysteine	Cys
F	Phenylalanine	Phe	**Y**	Tyrosine	Tyr
W	Tryptophan	Trp	**H**	Histidine	His
K	Lysine	Lys	**R**	Arginine	Arg
Q	Glutamine	Gln	**N**	Asparagine	Asn
E	Glutamic Acid	Glu	**D**	Aspartic Acid	Asp
S	Serine	Ser	**T**	Threonine	Thr



National Library of Medicine
National Center for Biotechnology Information

todste@orcid

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 March 2022

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

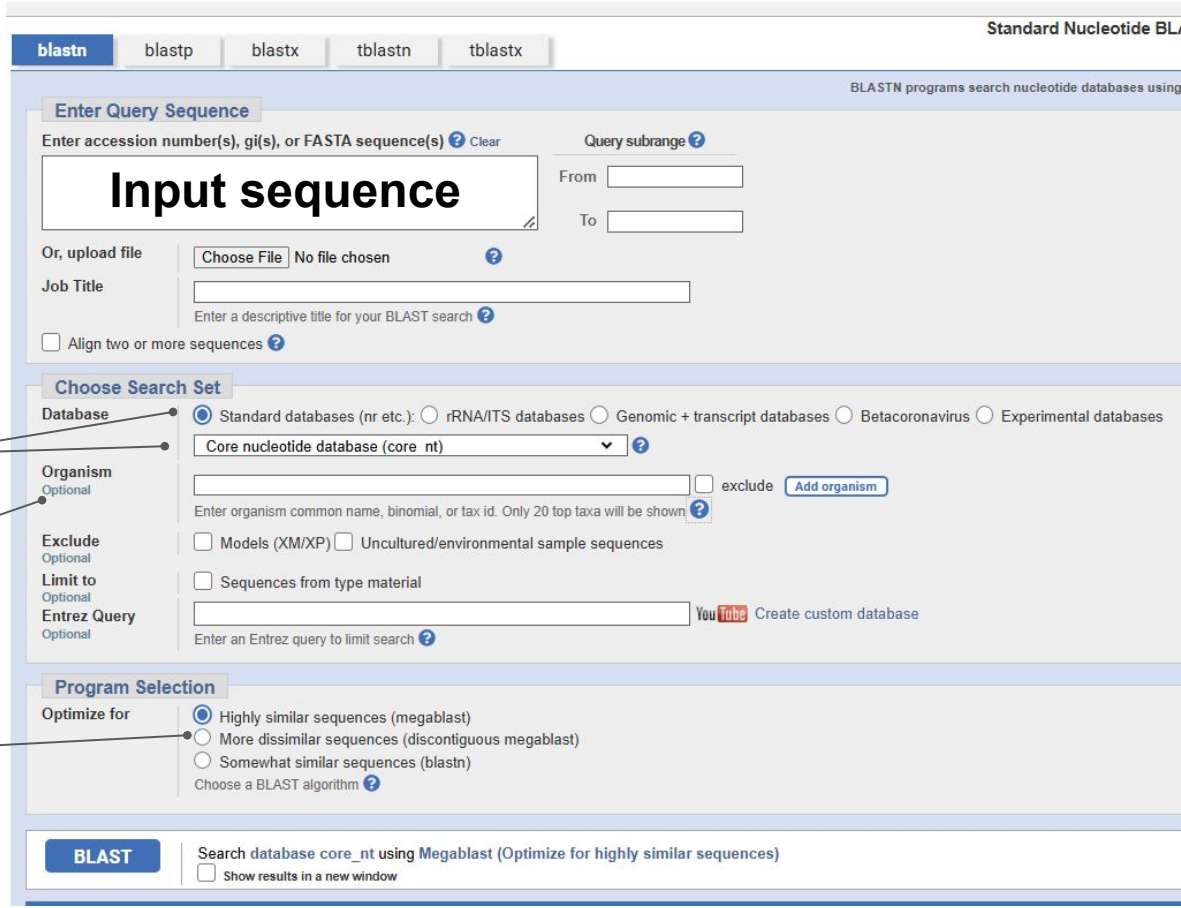
blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

Simple Searching of Nucleotide sequence

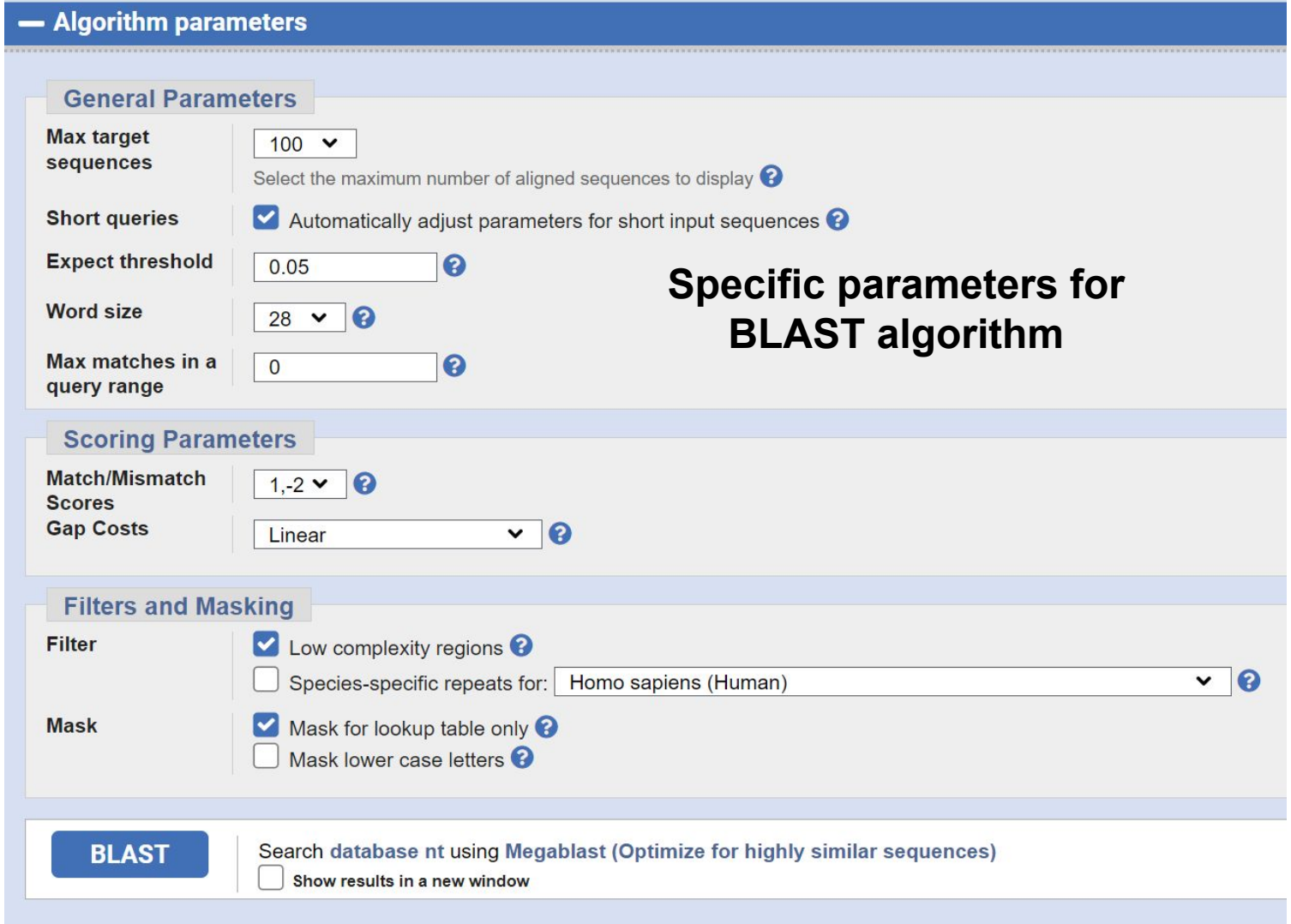
- blastn
  - Nucleotide searching



Select Database

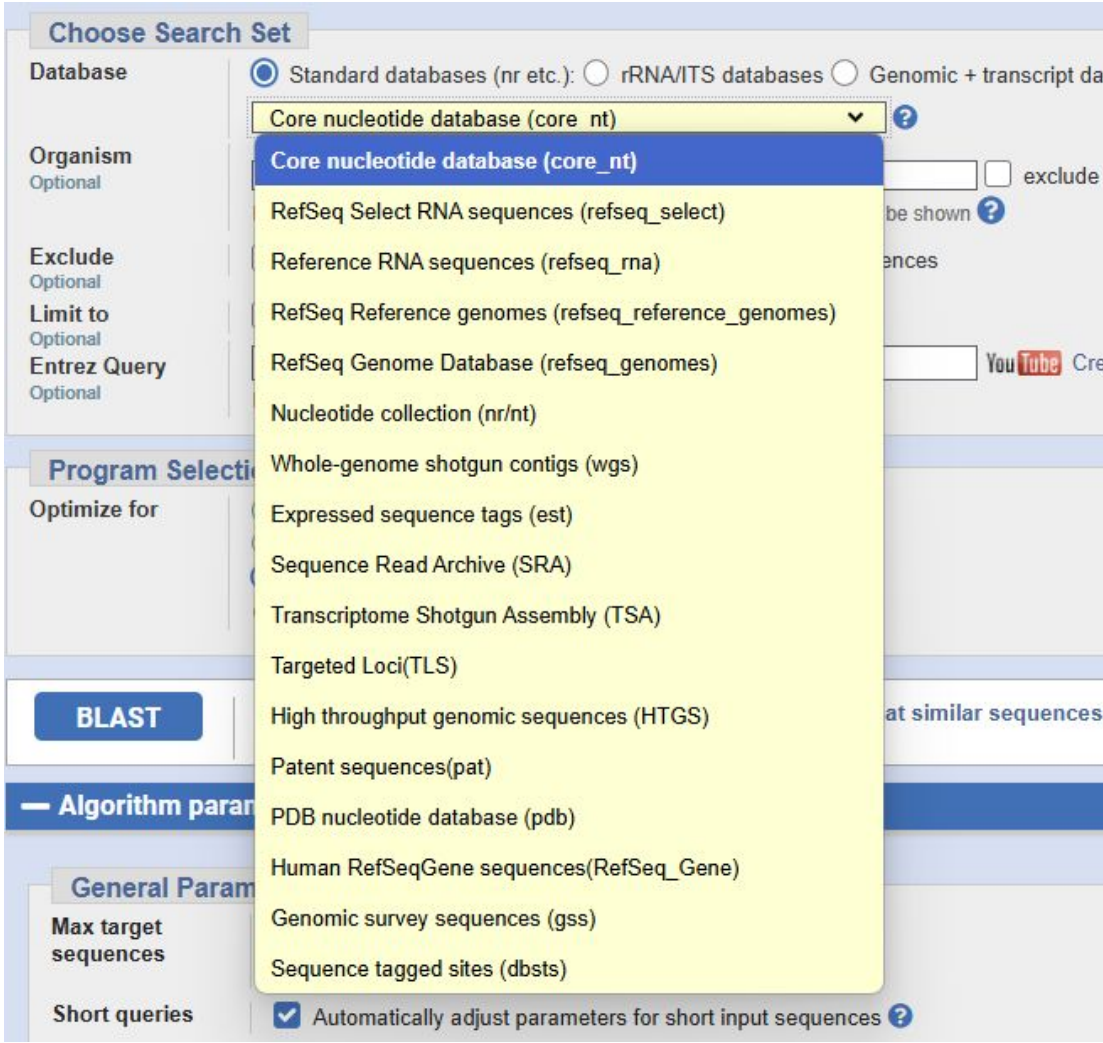
Specific organism

BLAST algorithm


Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
0.05
Word size
28
Max matches in a query range
0
Specific parameters for BLAST algorithm
Scoring Parameters
Match/Mismatch Scores
1,-2
Gap Costs
Linear
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Homo sapiens (Human)
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database nt using Megablast (Optimize for highly similar sequences)
Show results in a new window

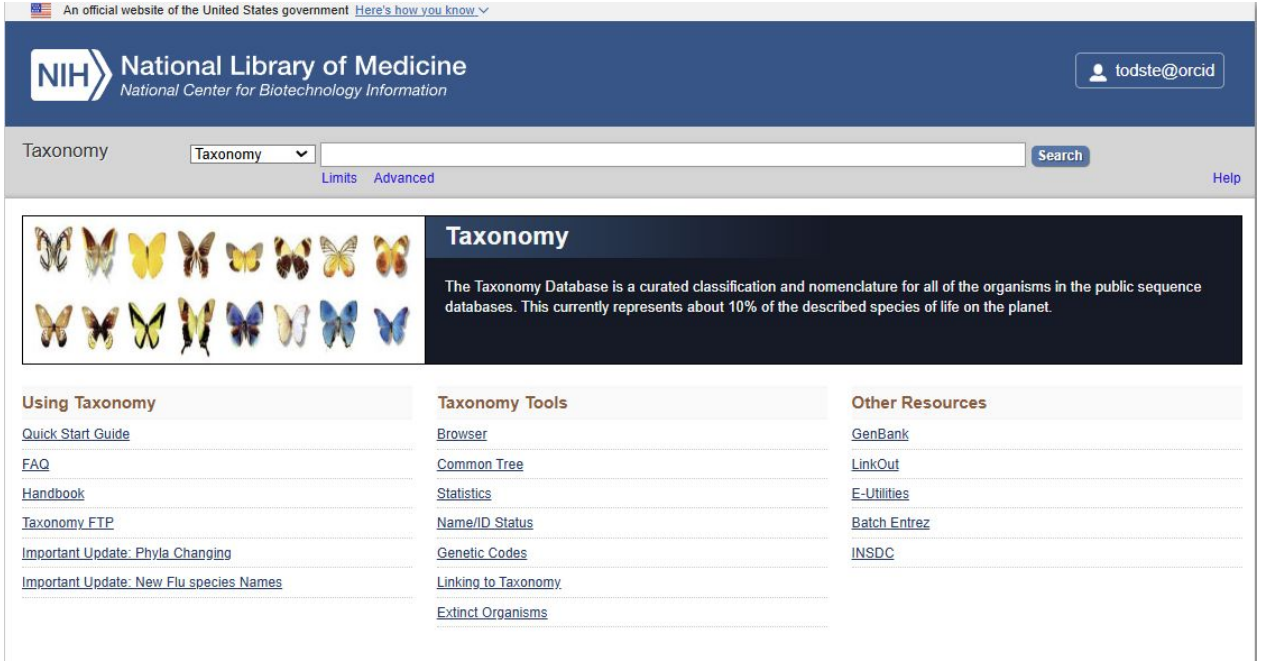
Nucleotide Database

- **core_nt**
- **nr/nt**
- refseq_rna



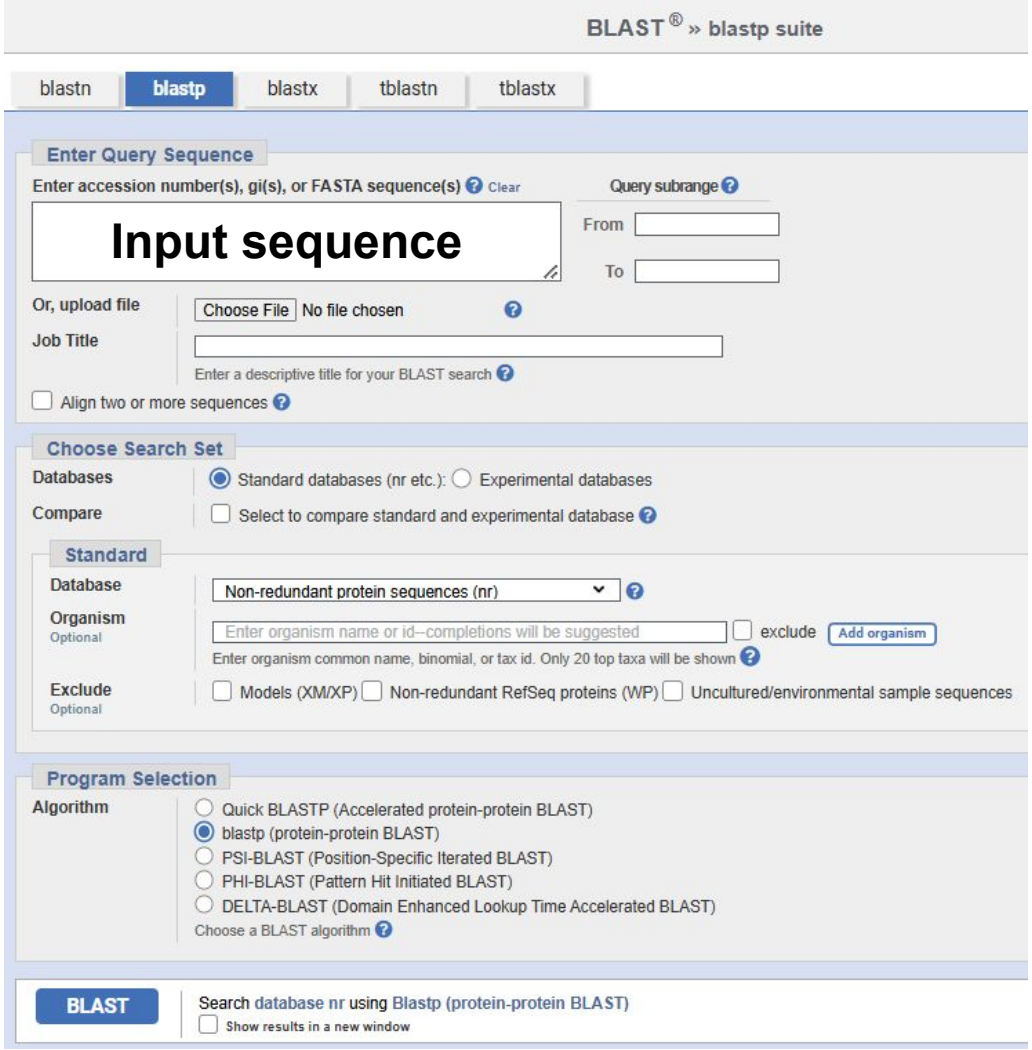
Taxonomy Browser

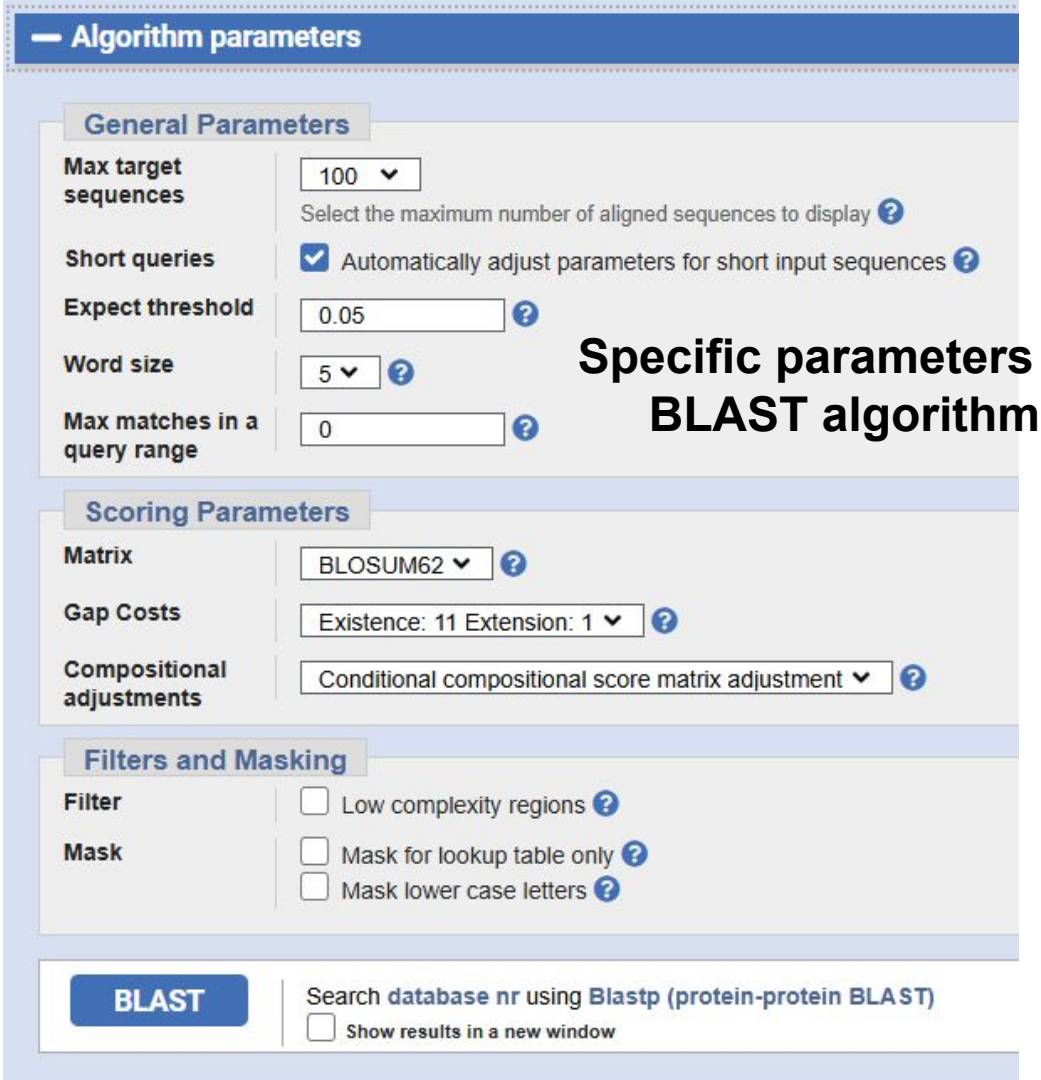
- To search the species id or names for limit the searching in BLAST



Searching of amino acid sequence

- blastp
  - Amino acid searching

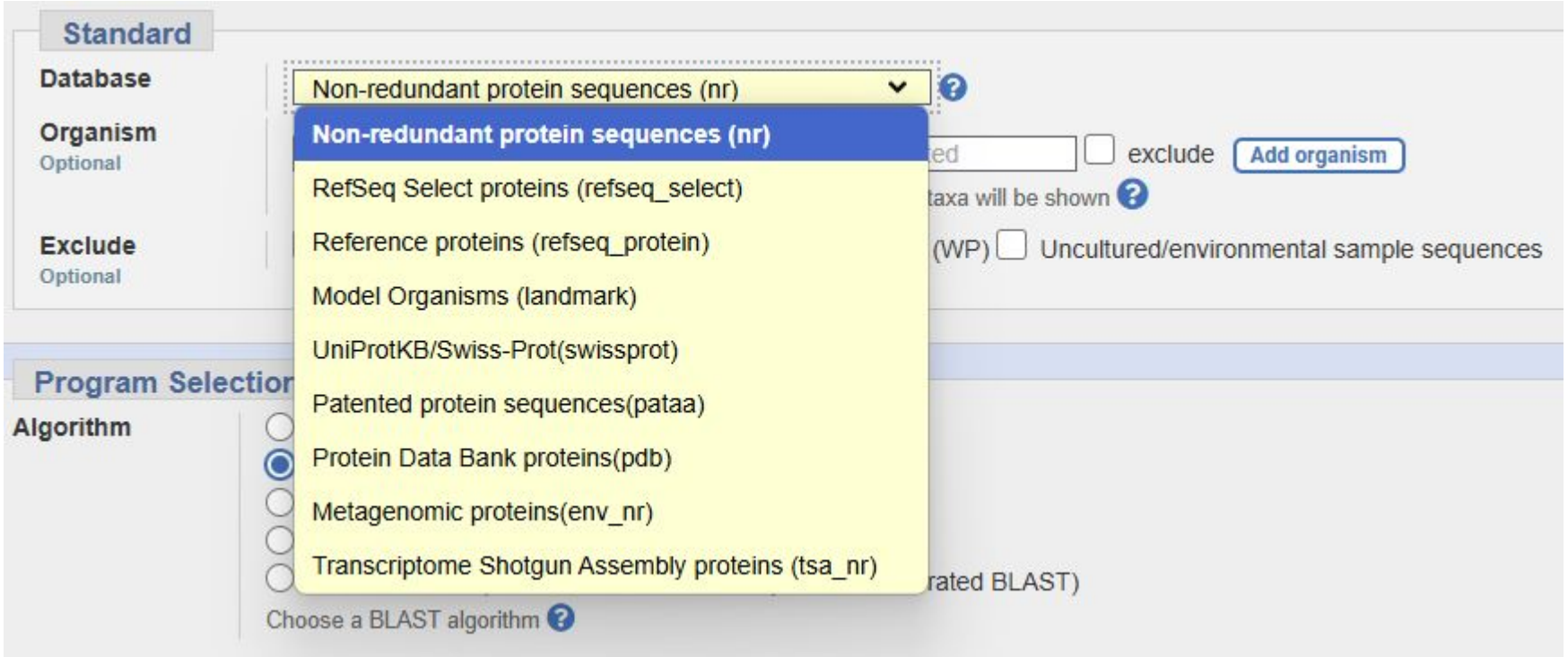




Specific parameters for BLAST algorithm

Protein databases

nr or **refseq_protein**



BLAST Results

RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain [Macaca mulatta]

Sequence ID: P02026.1 **Length:** 146 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 146 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
299 bits(765)	4e-102	Compositional matrix adjust.	146/146(100%)	146/146(100%)	0/146(0%)

```
Query  1    VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKV  60
            VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKV
Sbjct  1    VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKV  60

Query  61   KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK  120
            KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
Sbjct  61   KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK  120

Query  121  EFTPQVQAAYQKVVAGVANALAHKYH  146
            EFTPQVQAAYQKVVAGVANALAHKYH
Sbjct  121  EFTPQVQAAYQKVVAGVANALAHKYH  146
```

Related Information

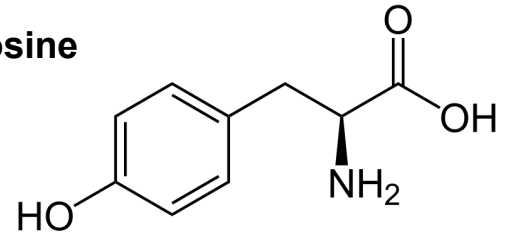
Gene - associated gene details

Identical Proteins - Identical proteins to P02026.1

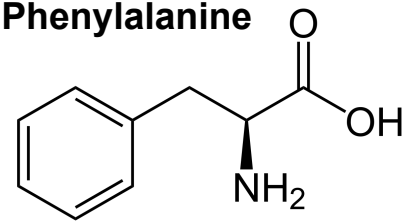
BLAST Results

- **Score**: Higher is better
  - Bit score is a normalized score in search space.
- **Expect** (**E-value**):**Lower is better**
  - Tell us about this match is occurred by chance or others
- **Identity**
  - Percentage of Matched Sequence
- **Positive**
  - Indicate **a conservative substitution** or substitutions that are **often observed in related proteins**.
- **Gap**
  - Percentage of GAP

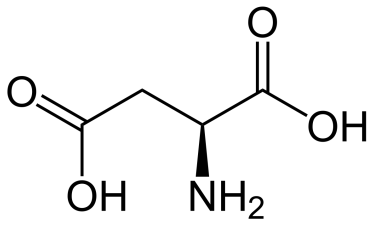
Tyrosine



Phenylalanine



Aspartic acid



Score	Expect	Method	Identities	Positives	Gaps
281 bits(718)	6e-95	Compositional matrix adjust.	137/146(94%)	141/146(96%)	0/146(0%)

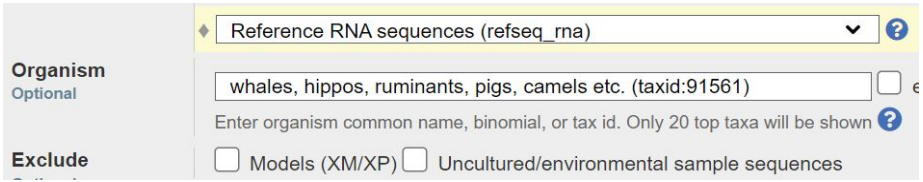
```
Query  1    VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVYPWTQRFFESFGDLSSPDAVMGNPKV  60
            VHLTPEEK+AVT LWGKVNVDEVGGEALGRLL+VYPWTQRFFESFGDLS+PDAVMGNPKV
Sbjct  1    VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  60

Query  61   KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK  120
            KAHGKKVLGAFSDGL HLDNLKGTFA LSELHCDKLHVDPENF+LLGNVLV VLAHHFGK
Sbjct  61   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVGVLAHHFGK  120

Query  121  EFTPQVQAAYQKVVAGVANALAHKYH  146
            EFTP VQAAYQKVVAGVANALAHKYH
Sbjct  121  EFTPPVQAAYQKVVAGVANALAHKYH  146
```

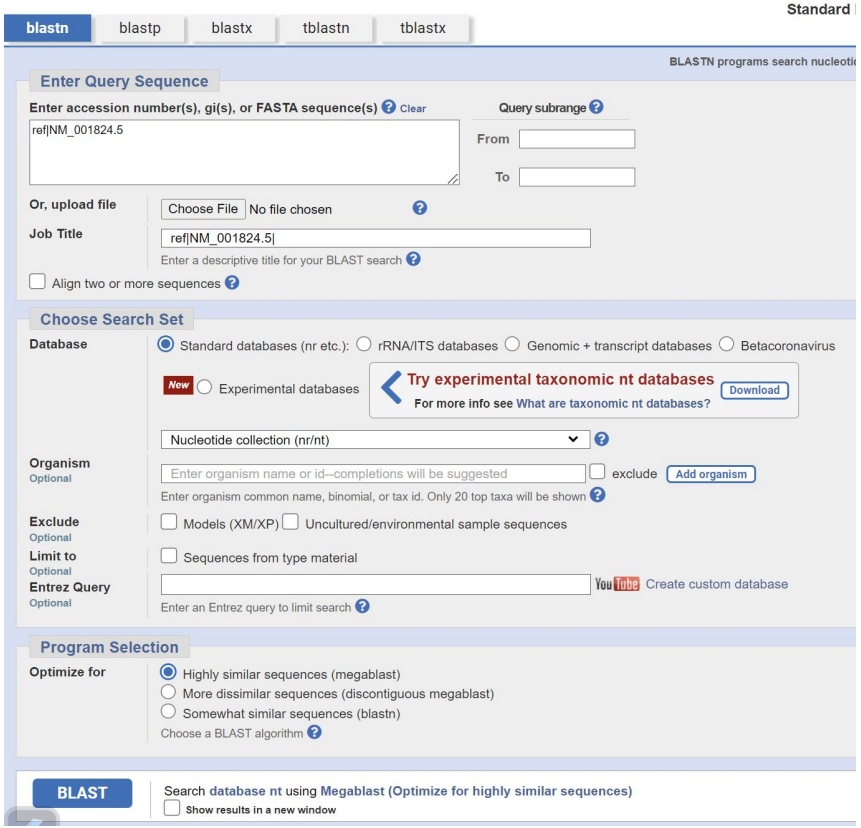
Practice

- https://github.com/todtec/SC153401

The Study of unknown3 homology

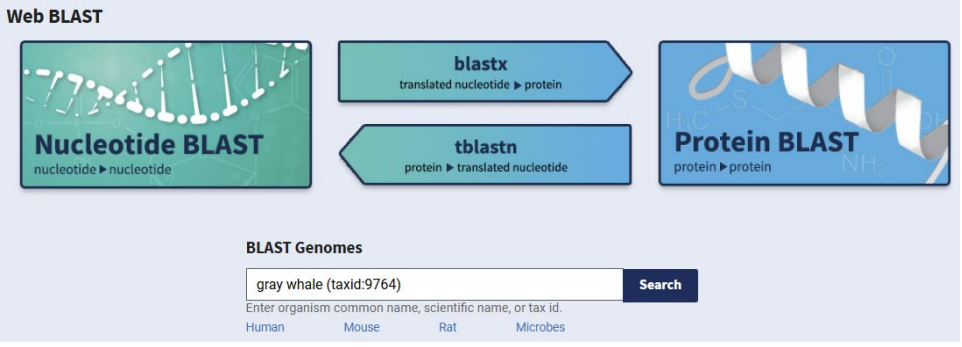
- Tool
  - blastn
- Database
  - core_nt (111500860)
  - nr/nt (111726244)
  - refseq_rna (66876125)
- Blastn options
  - Megablast
  - blastn

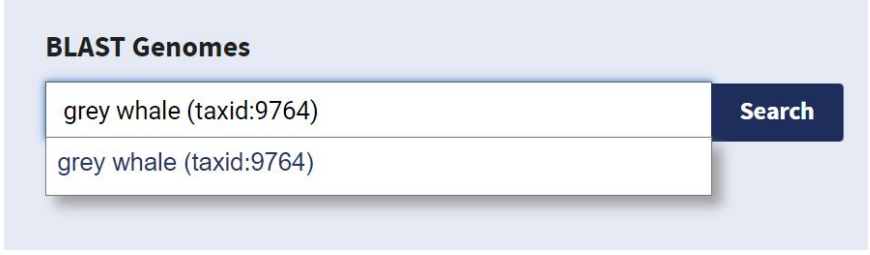


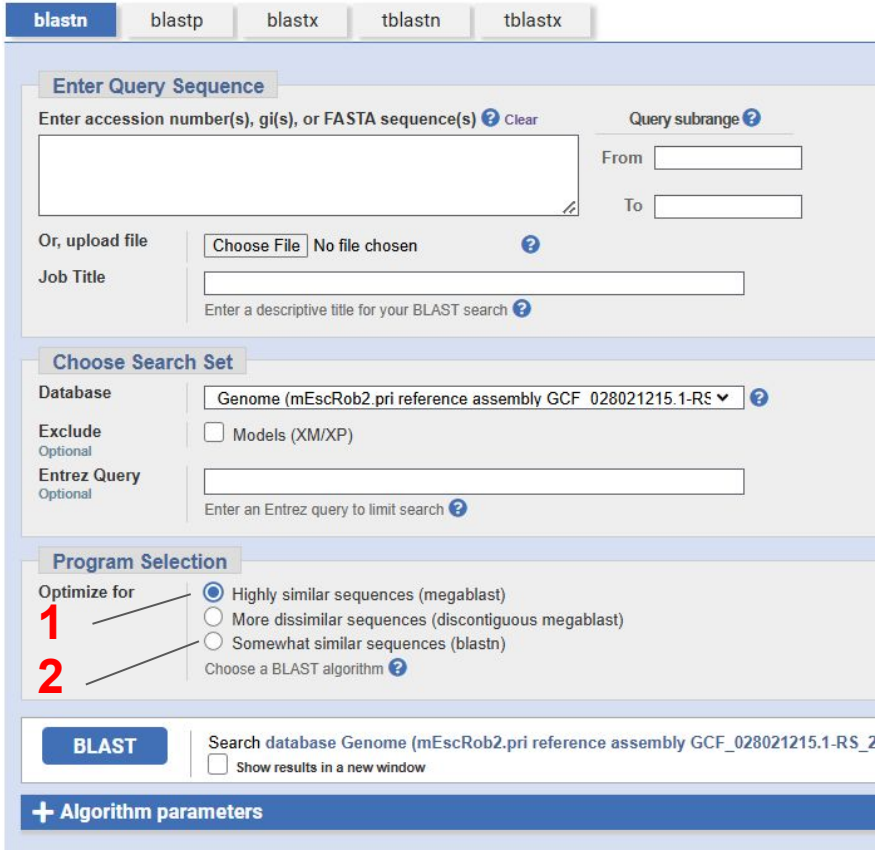


Annotating Gene by mRNA sequence

- BLAST Genomes
- NM_001824.5
- Grey whale



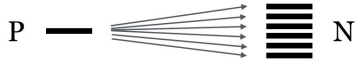




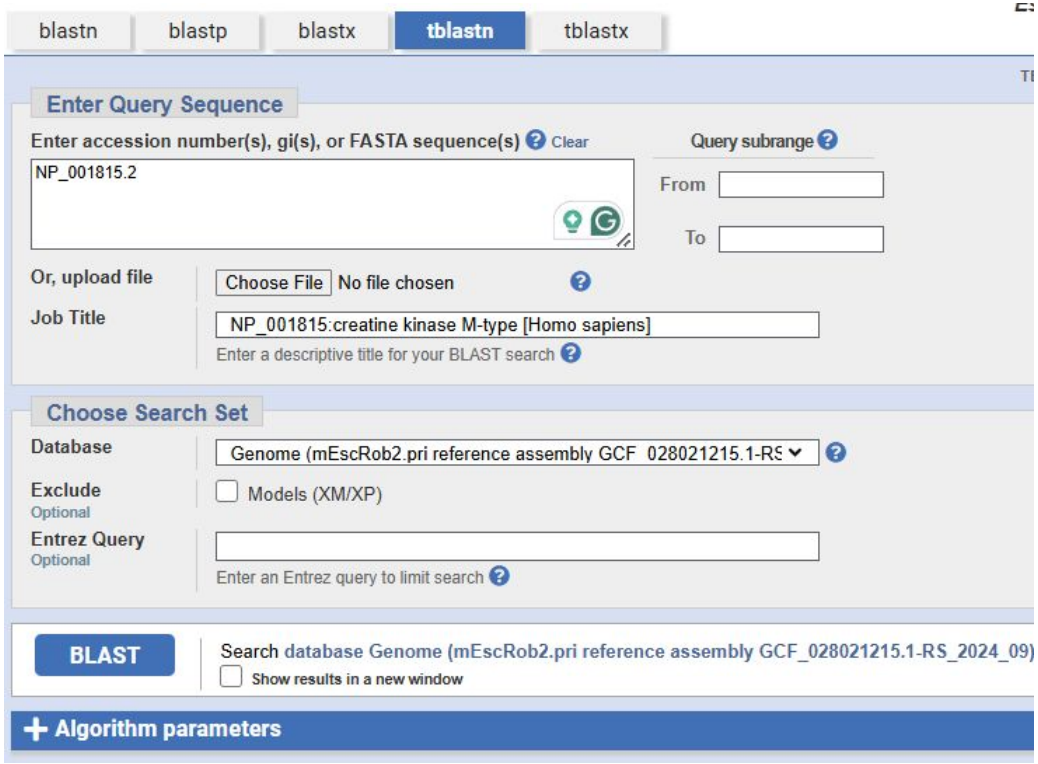
Annotating Gene by Protein Sequence

- NP_001815.2
- tblastn
  - Translate protein sequence to nucleotide before using BLAST

tblastn

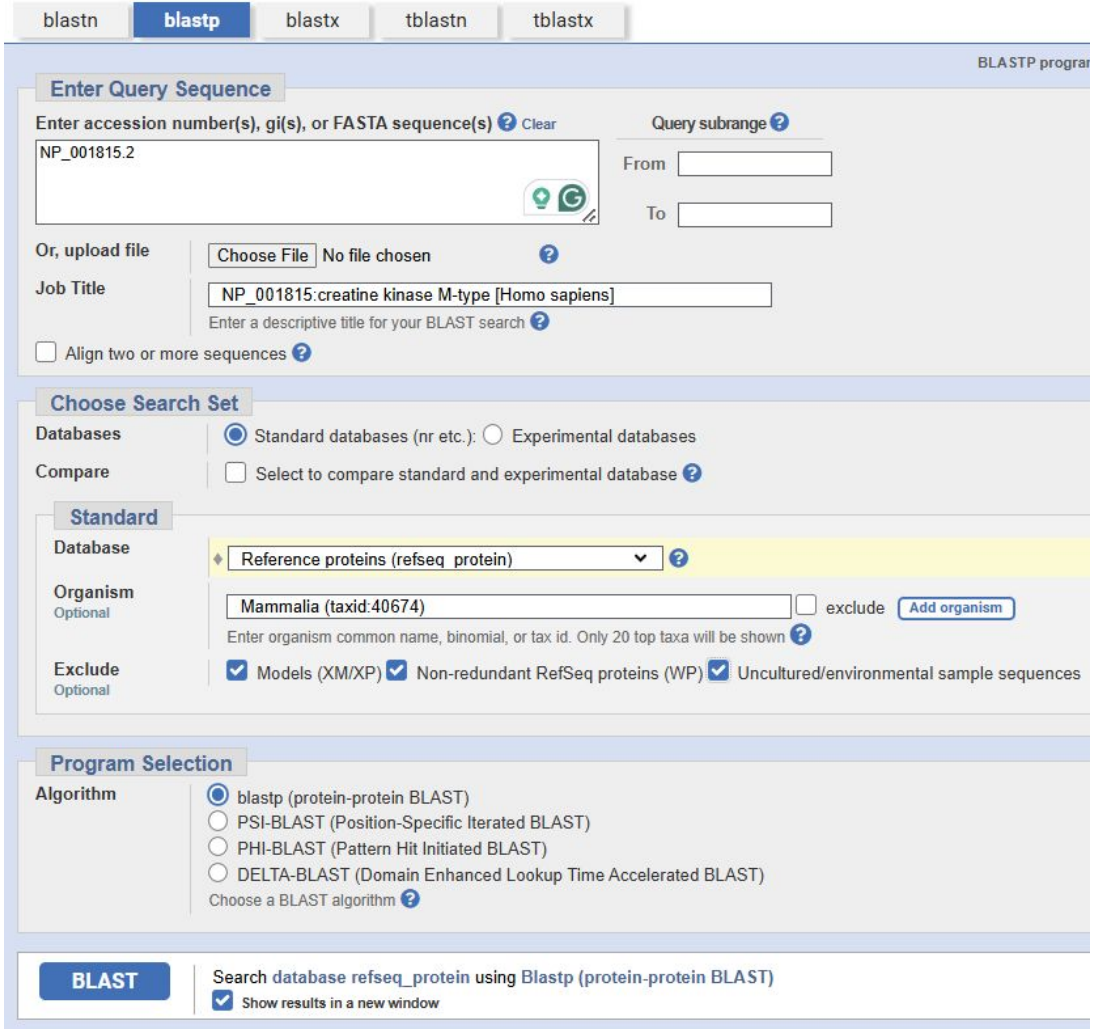


~ 6X

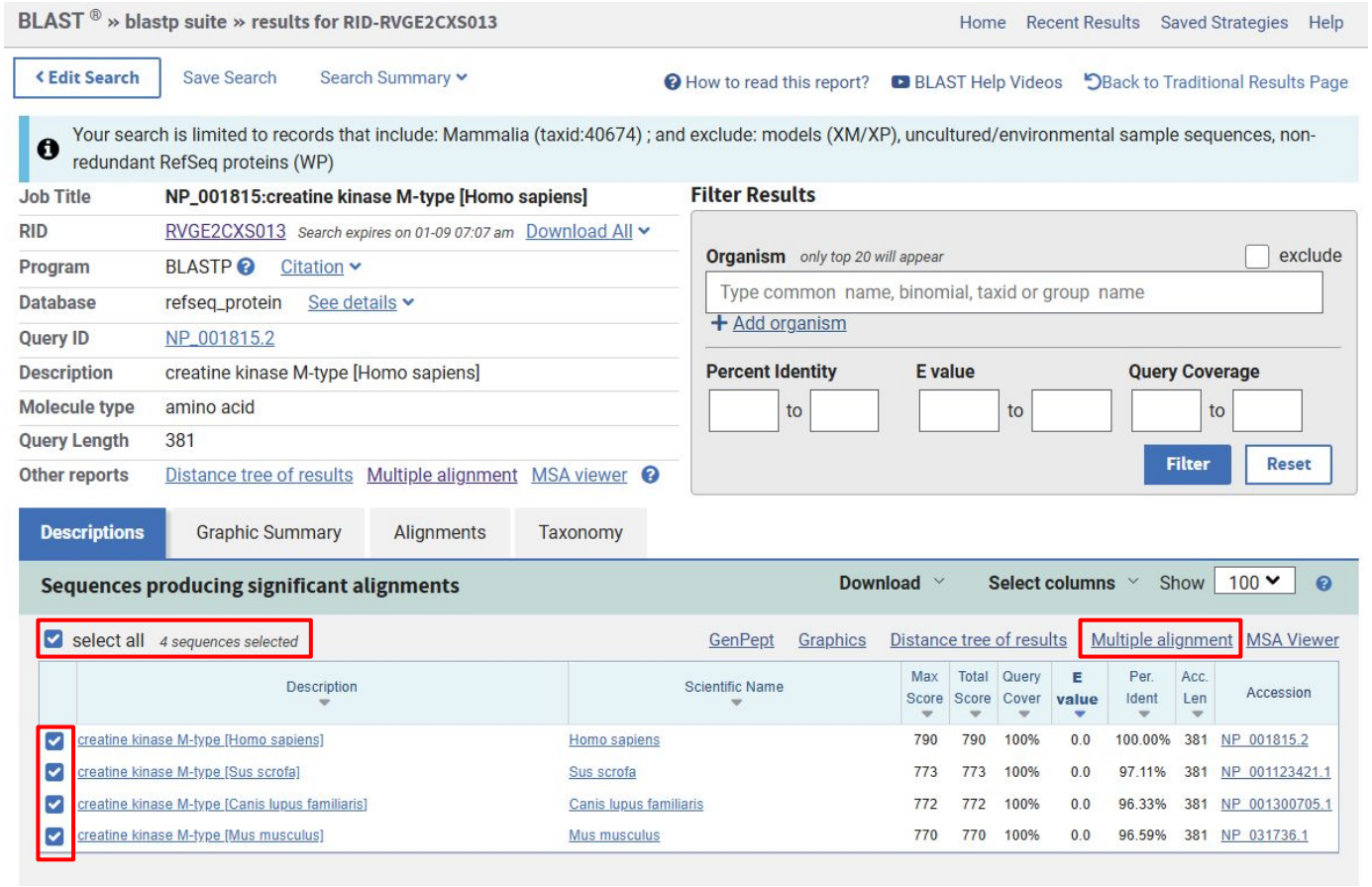


blastp

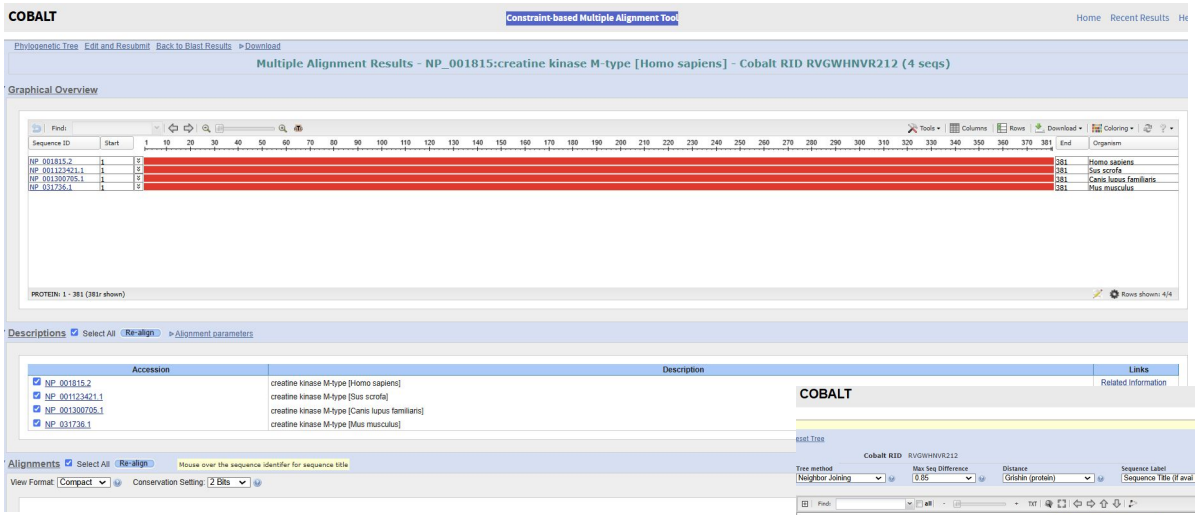
- **Find Orthologs** based on **BLAST**
- Use **multiple alignment** to identify **evolutionary relationship (COBALT)**
  - **Constraint-based Multiple Alignment Tool**
- Build Simply **Phylogenetic Tree**

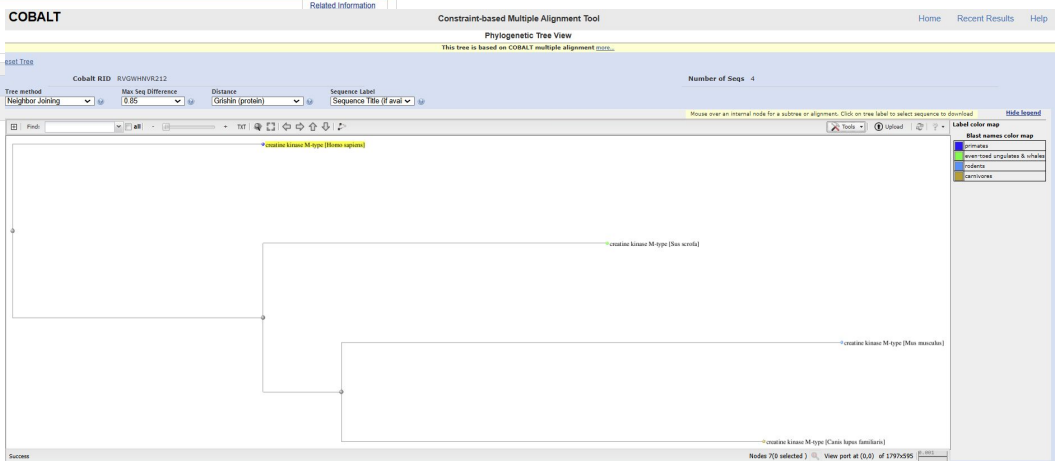


blastp result



COBALT


COBALT
Constraint-based Multiple Alignment Tool
Home
Recent Results
Phylogenetic Tree
Edit and Resubmit
Back to Blast Results
Download
Multiple Alignment Results - NP_001815:creatine kinase M-type [Homo sapiens] - Cobalt RID RVGWHNVR212 (4 seqs)
Graphical Overview
Sequence ID
Start
End
Organism
NP_001815.2
Homo sapiens
NP_001123421.1
Sus scrofa
NP_001300705.1
Canis lupus familiaris
NP_031736.1
Mus musculus
Descriptions
Select All
Re-align
Alignment parameters
Accession
Description
Links
creatine kinase M-type [Homo sapiens]
creatine kinase M-type [Sus scrofa]
creatine kinase M-type [Canis lupus familiaris]
creatine kinase M-type [Mus musculus]
Related Information
Alignments
Mouse over the sequence identifier for sequence title
View Format:
Compact
Conservation Setting:
2 Bits


COBALT
Constraint-based Multiple Alignment Tool
Home
Recent Results
Help
Phylogenetic Tree View
This tree is based on COBALT multiple alignment
Cobalt RID
RVGWHNVR212
Number of Seqs
4
Tree method
Neighbor Joining
Max Seq Difference
0.85
Distance
Grishin (protein)
Sequence Label
Sequence Title (if aval
Label color map
Blast names color map
primates
even-toed ungulates & whales
rodents
carnivores
creatine kinase M-type [Homo sapiens]
creatine kinase M-type [Sus scrofa]
creatine kinase M-type [Mus musculus]
creatine kinase M-type [Canis lupus familiaris]

END