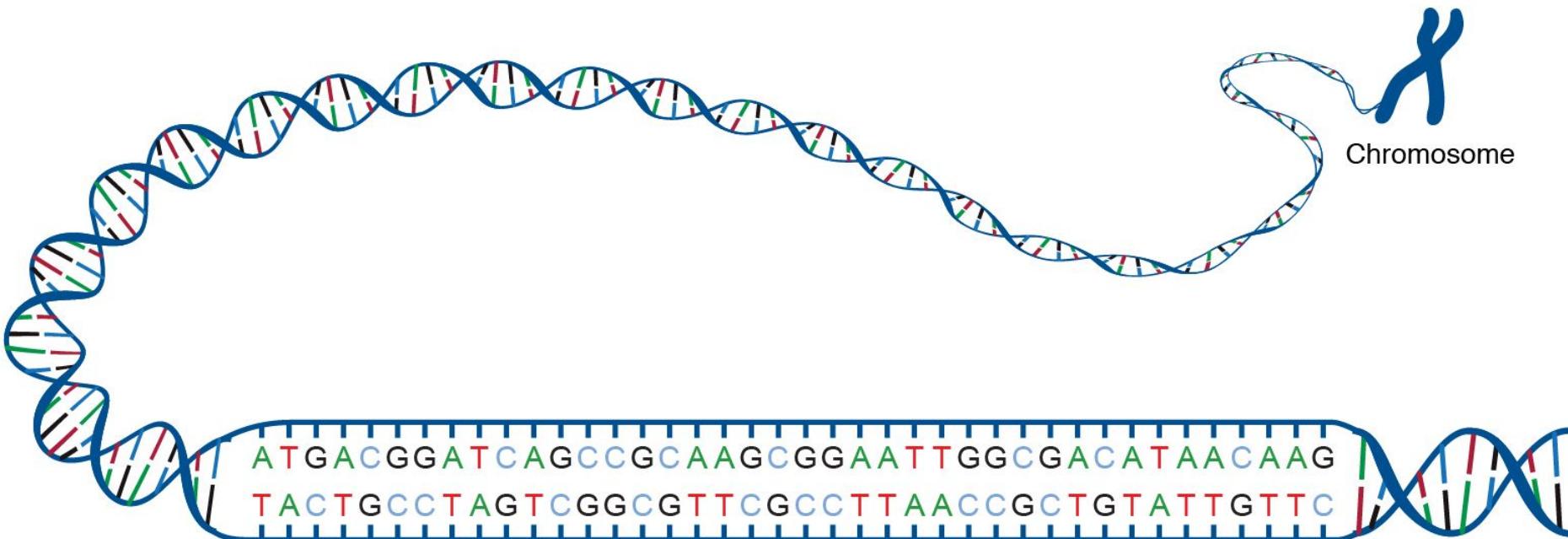


Basic Bioinformatics

PSB
Lect. Todsapol Techo

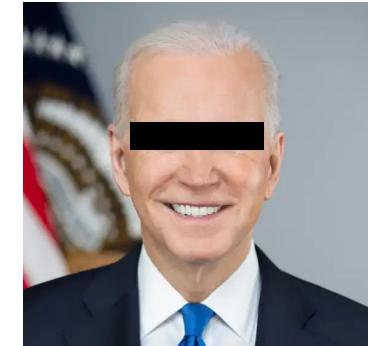
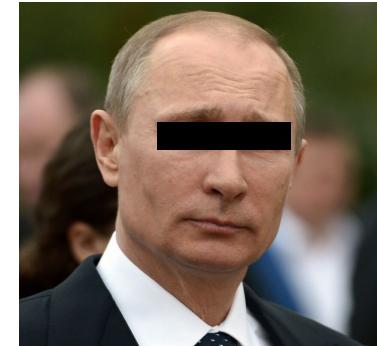
The Languages of Life (DNA)

- Living information store in DNA sequence



1 gaactcggga agccggcgag aagtgtgagg ccgcggtagg gcccacatccc gctccggaga
61 gaagtctgag tccgcccagc tctgcaggcc cgccggaaagct cggtaatgtat aagcaccccg
121 gccactttgc agggcgtcac cgcctacacg ccccctcgtc tctcgacgg cggcgtctag
181 cctcggggcg ctccggcccc ccgcctctc cgggggagga atcaagaaga gactgccccaa
241 tagggccggc ttgaccgcg aacaggcgag gttcccgccc ggatggcgc ggcagaaggc
301 cccgcccagg agccgagggc cagcccagag gaggcgtggc cacgctgcg gcggaaagtgg
361 agccctccgc gagcgcgcga ggccgcgggg gcaggcgggg aaacccgaca gttagggcgg
421 ggccggggcg gcgtatgggg tgcgggagca ctacgcggag ctgcacccgt gcccggcgg
481 attggggatg cagagcagcg gcagcgggta tggcaggcag ccggcgggccc ggcctccagg
541 gcaggtgccc gagaggcagg ggctggcctg ggatgcgcgc gcacccgtccc tcgccccccc
601 cccgcccac gaggggttgtt ggccgaggcc ccgcctcgct gaggcgggtt
661 cgctcagccc aggccccccccc gccgattaaa tggccggcg gggctcagcc
721 cccggaaacg gtcgtacact tcggggctgc gagcgcggag ggcgacgacg acgaagcgc
781 ggttaaccggc cggggggcggc ccgcgcaggc ggaggagcgt actgtcccg cgtgcggc
841 gcggcggtaa aatacagct gtttttgtt ctgagaacc gagcagaatc gagagggtct
901 taaccaatcc ctttataccccc cgcacccctt ctcttgagcc cctgagaccc cgagagcga
961 ggggacttgc cgaccgggtt caccagctt ggcaggggga gggctggagc tgaactccag
1021 catctgcacc atctcccatg ctccaggta ttgtggagtt cccgctacag tcgggaatga
1081 gatggtcctg ggcacgcagt tccatgcccc acaaggattt tactcggtt tccagaattt
1141 atgctgttagt cggaaatacac caatgtttt agtaattttt taatgtacac cctgaaatgaa
1201 ggctgcctag gagagagtgg ctggagccca gagccagcag tttctaaccct atcaaccact
1261 ccccaatgcc cagccgttca caaggagtga ttgggcaat cagggttac cctgggtgtga
1321 gaccccgag gaactctcaa gaaaggggct aacttctcaa tgctctctt ttcttctgccc
1381 ttgttaacga gccttcttc caccagacag cgtcatggca gagcagggtt ccctgagcc
1441 gaccccgagt tgccggatcc tgccggaaaga gctttccag ggcgtgcct tccatcagtc
1501 ggatacacac atattcatca tcatgggtgc atcggttagt atctcccaagg ccccaatctt
1561 aaaagccagg aagtgcctgc tccatgcctc agcttttcca actaattttt gcaggggcccc
1621 acaggctctg ctaagtttagt ctccctctcc cgcctctgtc agtgcggcag gatcatcccc
1681 gctctacccgt gtgtgcataat gcacttgcac gactttttt tttggagacg
1741 gagttttgtt ctggcgtccc aggatggggt gcagtggcgc catcttgc cactgtcaacc
1801 tcggccctctt gggttcaagc gatttctctt ctcagcccc ccaagtagct gagattacag
1861 gcatgcacca ccacaccggc ctaattttt tatttttagt agagacggag ttccacatgt
1921 ttgaccaggc ttgtctcaaa ctccgtacact caagtgtatcc gcccgcctcg gcctccaaa
1981 gtgctggat gacaggcatg agccacccacg cccggctgcga ggacgtttt gaaacctgt
2041 cctagccgtc gaatgttagcc cttaaatcttgc gatctctggc tcttactaa
2101 tggacttctt gggaaaaca gggaaatgtca ttcatctgt agtttttca gacctgttca
2161 tgccttaaca aatgagttcc aaaggcgaga tctgaggcct gatctctgccc tcttactaa
2221 cacttcaagt tccgtcgata caaacactta gctttctt gaacactgtc tctttgcctc

~ 99.6% similarity



~ 98.8% similarity

GCTCGAGAGATTCTTTACCTTTTACTATTTCACTCTCCATAACCTCCTATATTGACTGAT
CTGTAATAACCAAGATATTGGAAATAATAGGGGCTGAAATTGGAAAAAAAACGTGAAATA
TTTCGTTGATAAGTGATAGTGATATTCTCTTTATTGACTGTTACTAAGTCTCATGACTAACATC
GATTGCTTCAATTCTTTGTTGCTATATTATGTTAGAGGTTGCTGCTTGGTTATTGATAACGGTT
TGGTATGTTGAAAGCCGGTTTGCCTGCGTGACAGCCTCTGTCGTTCCCATCTCATGCGTAGA
CCAAGACACCAAGGTTATGGTGTGAAAGACTCTACGTTGGTGTGAACTCAATC
AGAGAGGTATCTTGACTTTACGTTACCAATTGAAACACGGTTATTGTCACCAACTGGGACGATATGGAAA
GATCTGGCATACCTTACAAAGAATTGGAGGTTGACTGACTTGAGGTTGAAAGCTGGTTACTCTTCT
CCACCACTGCTGAAAGAGAAATTGTCGTGACATCAAGGAAAACATGTTACGTCGCTTGGACTTCGA
ACAAGAAATGCAAACCGCTGCTCAATTCTCTCAATTGAAACATCTACGAATTCCAGATGGTCAAGTC
ATCACTATTGTAACGAAAGATTGAGGCCAGAAGCTTGTTCATCCTCTGTTTGGTTGGAAAT
CTGCCGGTATTGACCAAAACTTACAACCTCATGAGATGTGATGTCGATGCCGTAAGGAATTATA
CGGTAACATCGTTATGCCGGTGTACCCACATGTTCCAGGTATTGCCGAAAGAATGCCAAAGGAATC

Similarity ?

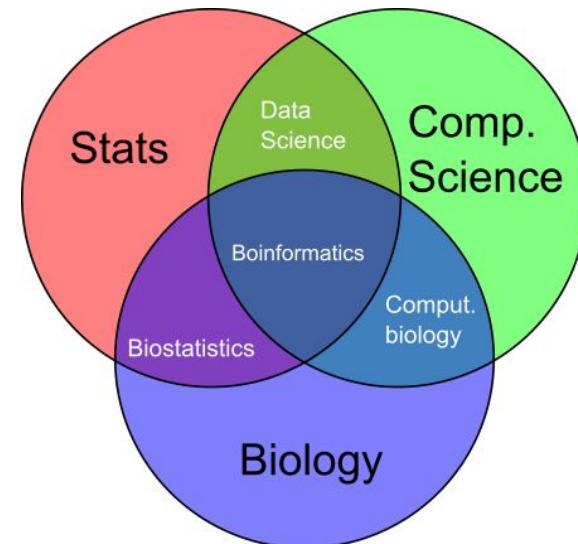


What is Bioinformatics?

- 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.



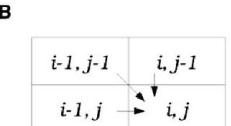


Molecular Biology

Bioinformatics

Statistics/Mathematics

		A	T	C	G
0	0	0	0	0	0
A	0	5	-1	-1	-1
T	0	4	10	9	8
G	0	3	9	8	14



Score
(i,j)
=
 \max

C

Best Alignment : ATCG

(Score = 38)

|| |

AT G

- Score ($i-1, j-1$) + Match / Mismatch
- Score ($i, j-1$) + gap
- Score ($i-1, j$) + gap

Computer Science

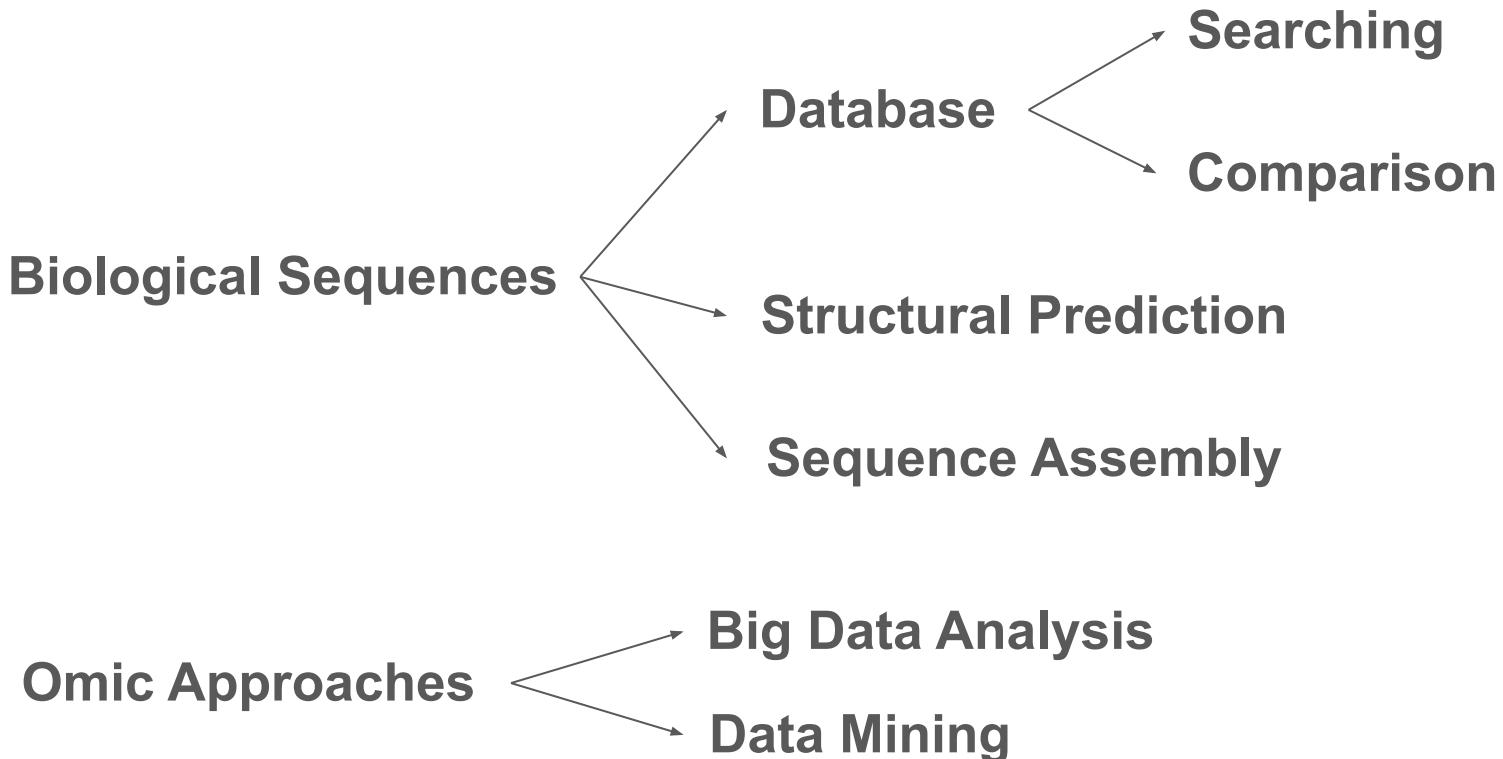
script.py

```

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)
8

```

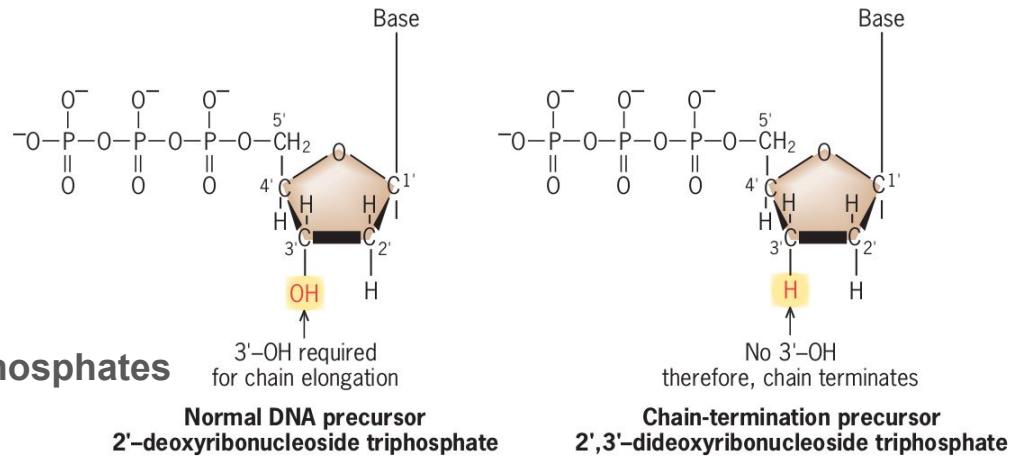
Overview of Bioinformatics



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'

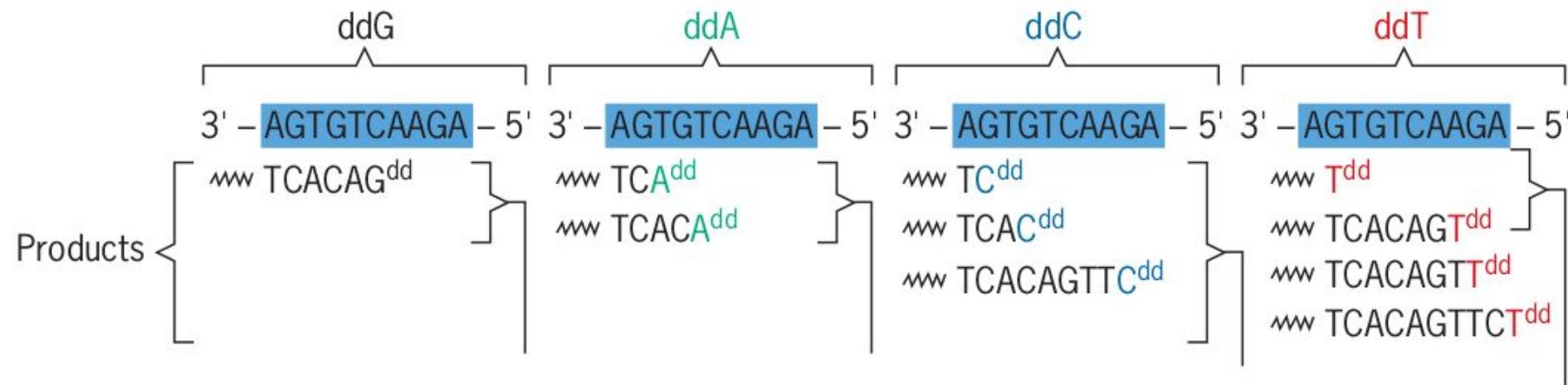
DNA polymerase

Primer strand 5' ~OH 3'

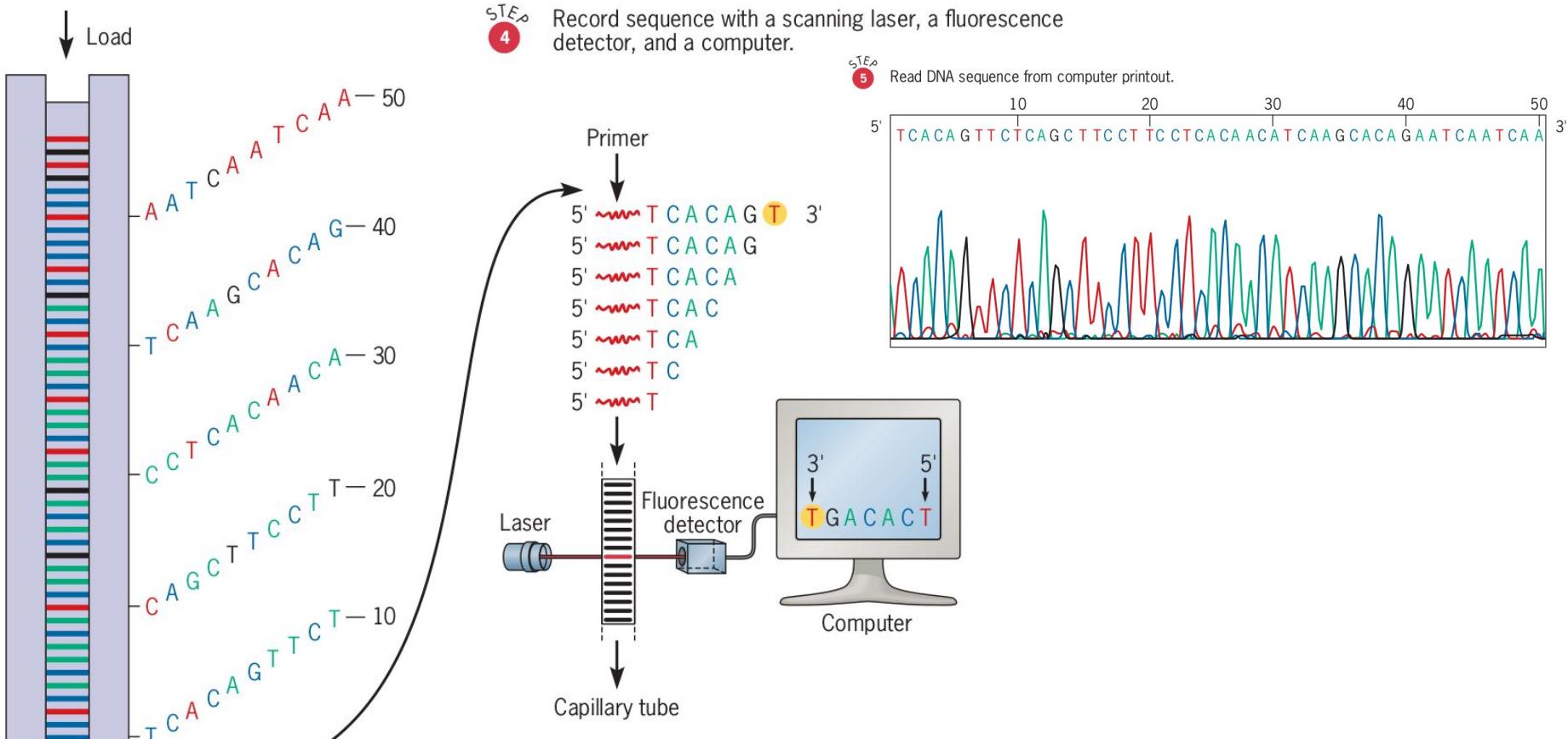
dGTP, dATP, dTTP, and dCTP

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.nlm.nih.gov

An official website of the United States government [Here's how you know](#)

Log in

National Library of Medicine National Center for Biotechnology Information

All Databases

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 *Hidden Markov models (HMMs) search*

Announcing GenBank Release 253.0 22 Dec 2022

GenBank release 253.0 (12/20/2022) is now available on the NCBI FTP site. This release has 21.3B bases and 3.25

New RefSeq Annotations! 20 Dec 2022

Racing in Human Genome Sequencing

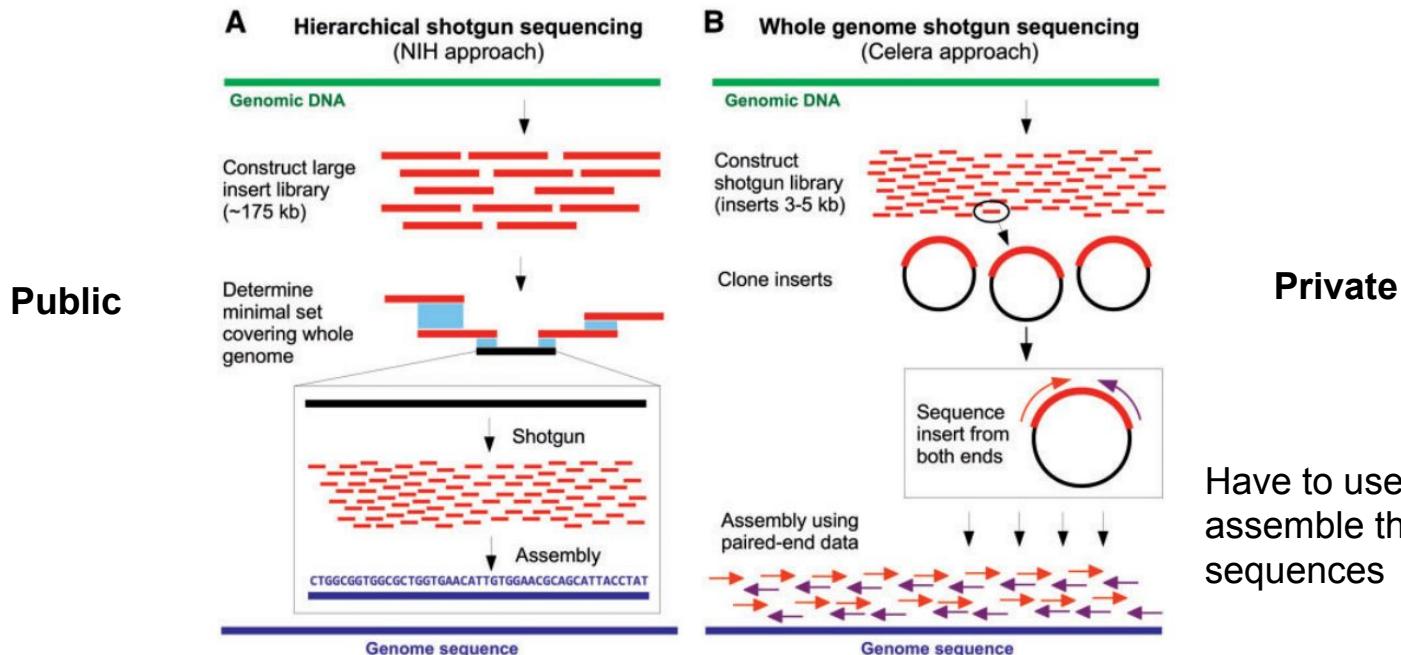
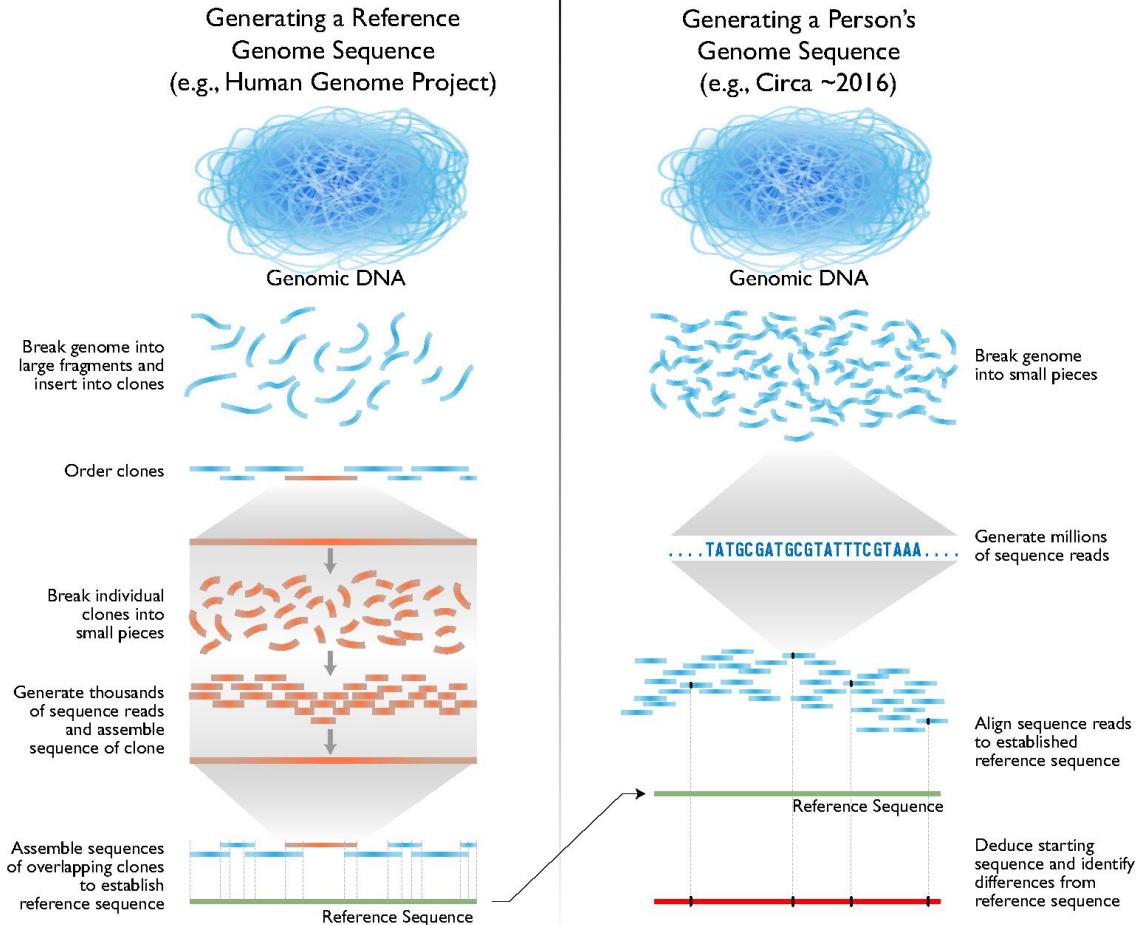


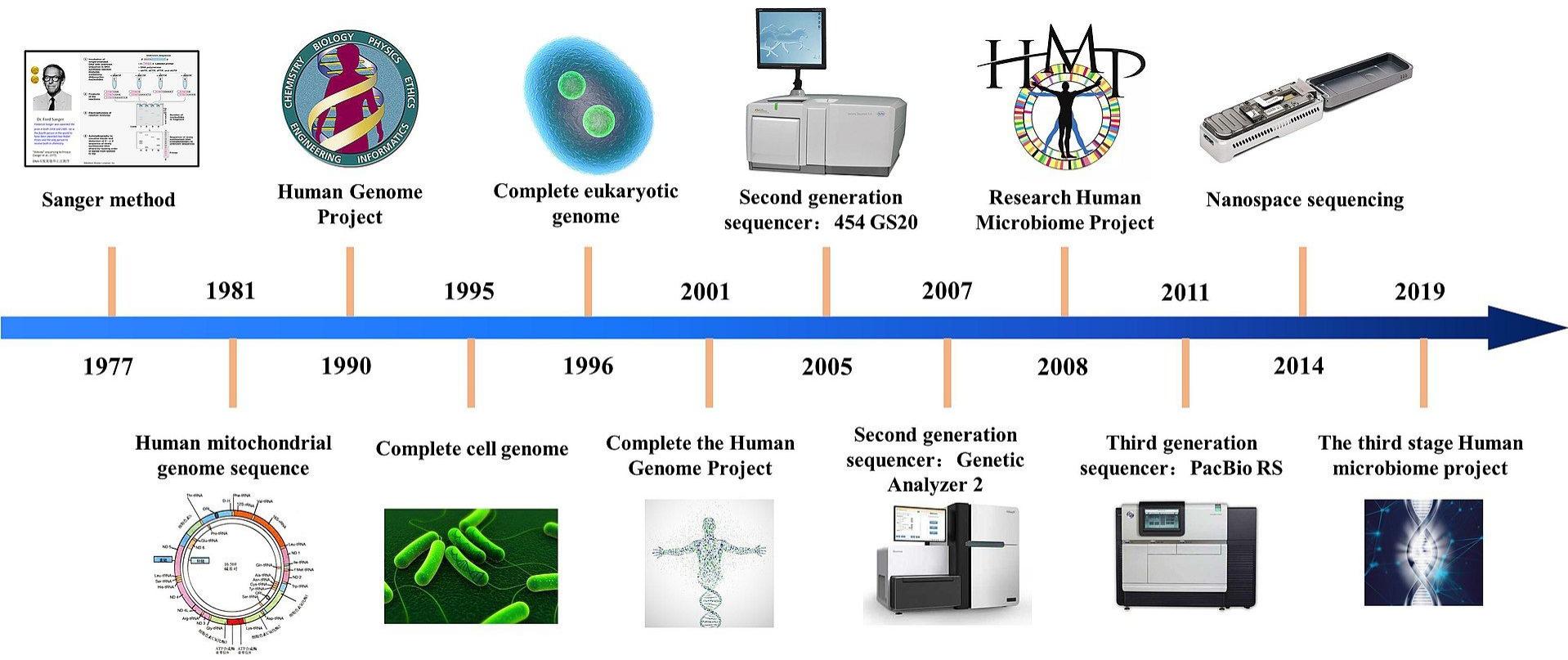
Figure 6. Hierarchical shotgun sequencing versus whole genome shotgun sequencing. Both approaches respectively exemplified the methodological rivalry between the public (NIH, A) and private (Celera, B) efforts to sequence the human genome. Whereas the NIH team believed that whole-genome shotgun sequencing (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

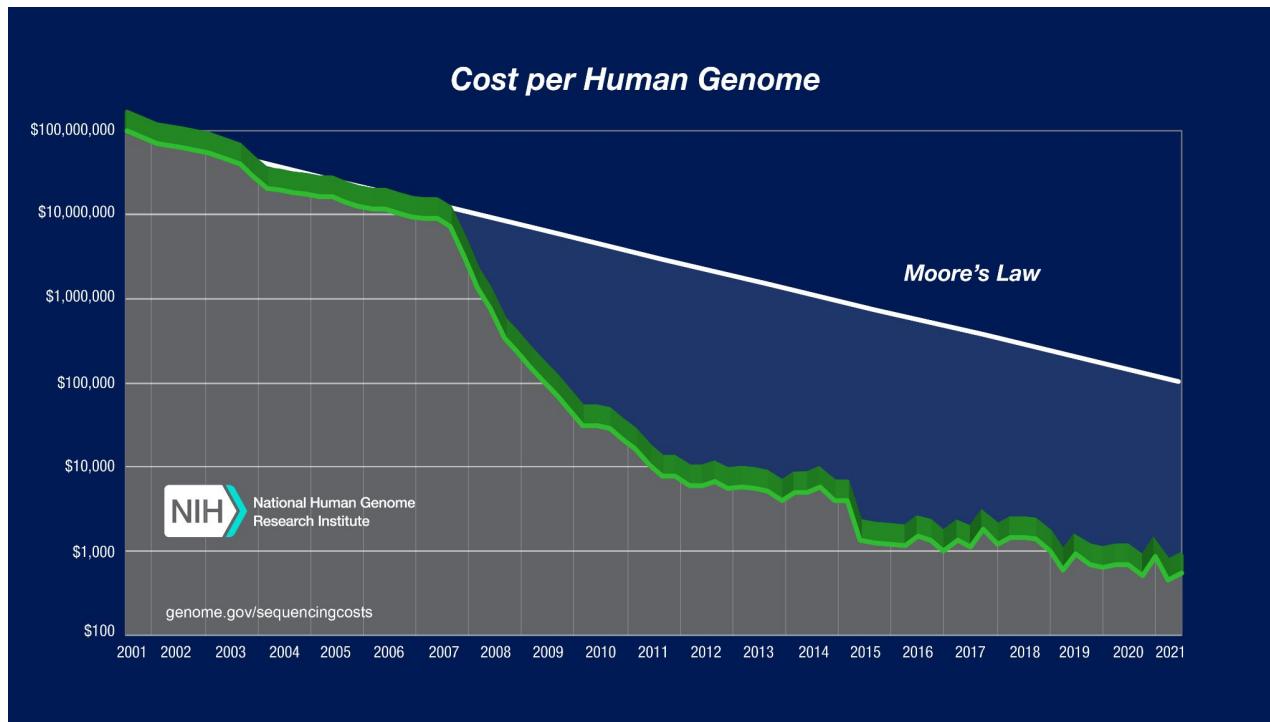
Human Genome Sequencing



Next Generation Sequencing (NGS)



Cost of Genome Sequencing

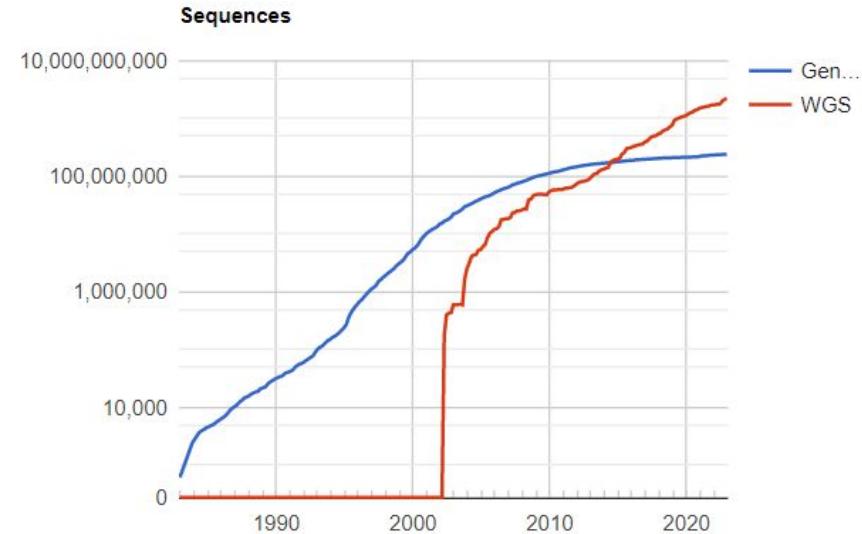
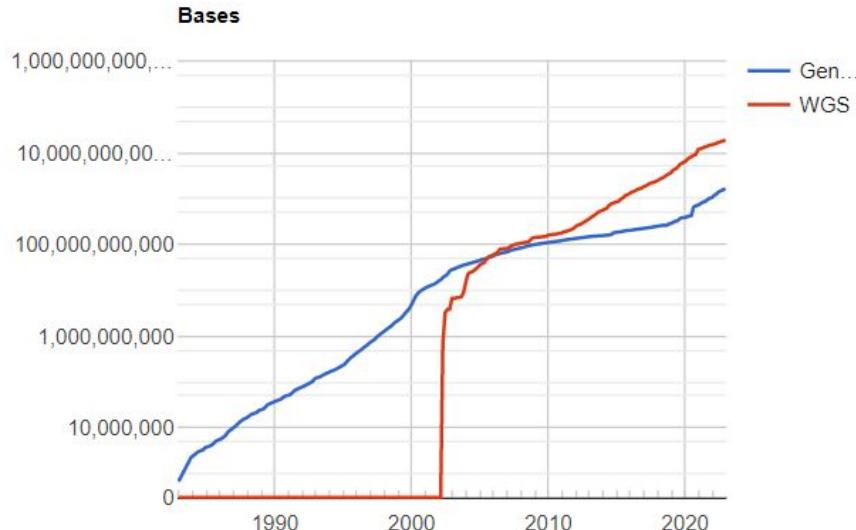


Illumina Sequencer: 2023

- Flagship of Sequencing Company

	iSeq 100	MiniSeq	MiSeq Series 	NextSeq 550 Series 	NextSeq 1000 & 2000	NovaSeq 6000 Series 	NovaSeq X Series
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 × 250 bp**	2 × 150 bp

Big Data Challenge: GenBank and WGS Statistics



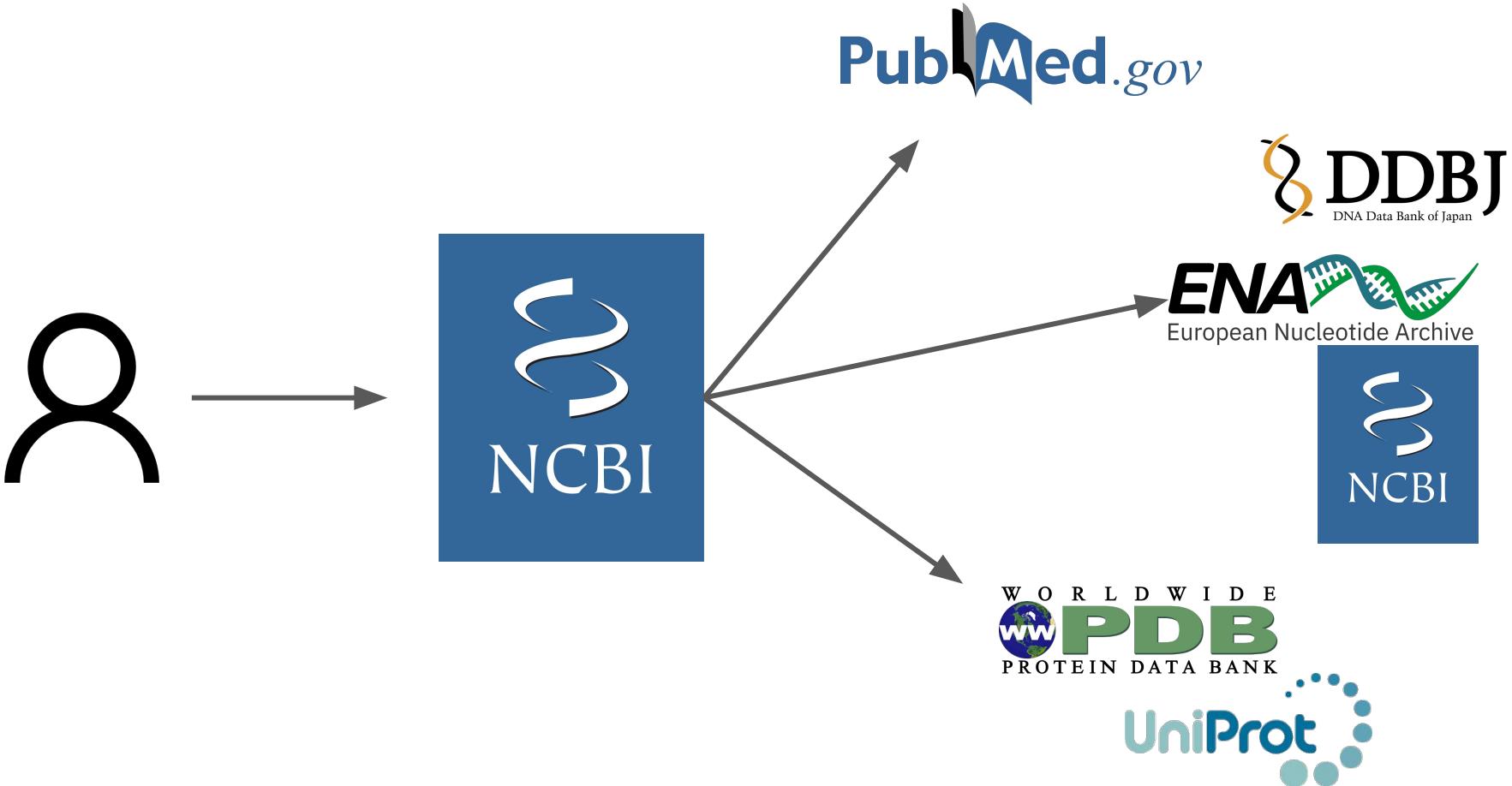
Date	Bases	Sequences	Bases	Sequences
Dec 1982	680338	606	-	-
Apr 2002	19072679701	16769983	692266338	172768
Dec 2022	1635594138493	241015745	19086596616569	2241439349

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



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 <input type="text"/>																													
<table border="1"><thead><tr><th>Tools</th><th>Description</th></tr></thead><tbody><tr><td>Amino Acid Explorer</td><td>Explores amino acid properties, substitutions and functions</td></tr><tr><td>Assembly Archive</td><td>Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DDBJ</td></tr><tr><td>Basic Local Alignment Search Tool (BLAST)</td><td>Finds regions of local similarity between biological sequences</td></tr><tr><td>Batch Entrez</td><td>Retrieves records specified in an uploaded file of identifiers</td></tr><tr><td>BioAssay Services</td><td>Tools that summarize the biological test results in the PubChem database</td></tr><tr><td>BLAST Link (BLink)</td><td>Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI</td></tr><tr><td>BLAST Microbial Genomes</td><td>Finds regions of local similarity between query sequences and sequences from complete microbial genomes</td></tr><tr><td>BLAST RefSeqGene</td><td>Finds regions of local similarity between query sequences and genomic sequences in the RefSeqGene/LRG set</td></tr><tr><td>CDTree</td><td>Classifies protein sequences and investigates their evolutionary relationships</td></tr><tr><td>Cn3D</td><td>Displays and manipulates 3-dimensional structures and alignments from the Structure database</td></tr><tr><td>COBALT</td><td>Performs protein multiple sequence alignments</td></tr></tbody></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
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)
 - Basic Local Alignment Search Tool (BLAST): sequence search

National Library of Medicine
National Center for Biotechnology Information

All Databases ▾

Welcome to NCBI

The National Center for Biotechnology Information advances science and biomedical and genomic information.

t (A-Z)

BLAST

Entrez

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

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.

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Search NCBI

Results found in 29 databases

GENE Was this helpful?

BRCA1 – BRCA1 DNA repair associated

Homo sapiens (human)
Also known as: BRCA1, BRCC1, BROVCA1, FANCS, IRIS, PNCA4, PPP1R53, PSCP, RNF53
Gene ID: 672

[RefSeq products](#) [Orthologs](#) [Genome Data Viewer](#)

New - Visualize gene across multiple species

RefSeq Sequences

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

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Search NCBI

Results found in 23 databases

Literature	Genes	Proteins
Bookshelf 17,131	Gene 102	Conserved Domains 0
MeSH 5	GEO DataSets 122,598	Identical Protein Groups 3
NLM Catalog 1,354	GEO Profiles 0	Protein 583
PubMed 626,915	PopSet 0	Protein Family Models 95
PubMed Central 1,599,497		Structure 1,857

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**)

Nucleotide Nucleotide Help

GenBank

Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript variant 1, mRNA
NCBI Reference Sequence: NM_007294.4
[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_007294 7088 bp mRNA linear PRI 05-OCT-2024
DEFINITION Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript variant 1, mRNA.
ACCESSION NM_007294
VERSION NM_007294.4
KEYWORDS RefSeq; MAMM Select.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Change region shown
Analyze this sequence
Pick Primers
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer

Articles about the BRCA1 gene
BRCA1 and 53BP1 regulate reprogramming

GENE Was this helpful?

BRCA1 – BRCA1 DNA repair associated
Homo sapiens (human)
Also known as: BRCA1, BRCC1, BROVCA1, FANCS, IRIS, PNCA4, PPP1R53, PSCP, RNF53
Gene ID: 672

New - Visualize gene across multiple species

RefSeq Sequences

Showing 5 of 368 (by status, accession number)

Transcript	nt	Protein	aa	Isoform	Status
NM_007294.4	7,088	NP_009225.1	1,863	1	curated
NM_007297.4	7,028	NP_009228.2	1,816	3	curated
NM_007298.4	3,865	NP_009229.2	759	4	curated
NM_007299.4	3,696	NP_009230.2	699	5	curated
NM_007300.4	7,151	NP_009231.2	1,884	2	curated

Where DNA sequence?

Get DNA sequence

- From Gene Database

Gene Advanced

Full Report

BRCA1 BRCA1 DNA repair associated [*Homo sapiens* (human)]

Gene ID: 672, updated on 5-Jan-2025

[Download Datasets](#)

Gene Sequences (FASTA)
 Transcript sequences (FASTA)
 Protein sequences(FASTA)

In addition, your package will include a detailed data report in both TSV and JSONL formats.

Table of contents

Summary

Official Symbol BRCA1 provided by HGNC
Official Full Name BRCA1 DNA repair associated provided by HGNC
Primary source HGNC:HGNC:1100
See related Ensembl:ENSG00000012048 MIM:113705; AllianceGenome:HGNC:1100
Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as IRIS; PSCP; BRCA1; BRCC1; FANCS; PNCA4; RNF53; BROVCA1; PPP1R53
Summary This gene encodes a 190 kD nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. The BRCA1 gene contains 22 exons and encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC). This gene product associates with RNA polymerase II, and through the C-terminal domain, also interacts with histone deacetylase complexes. This protein thus plays a role in transcription, DNA repair of double-stranded breaks, and recombination. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers. Alternative splicing plays a role in modulating the subcellular localization and physiological function of this gene. Many alternatively spliced transcript variants, some of which are disease-associated mutations, have been described for this gene, but the full-length nature of only some of these variants has been described. A related pseudogene, which is also located on chromosome 17, has been identified. [provided by RefSeq, May 2020]
Expression Broad expression in testis (RPKM 5.2), lymph node (RPKM 3.3) and 23 other tissues [See more](#)
Orthologs mouse all
NEW Try the new [Gene table](#)
Try the new [Transcript table](#)

Get DNA sequence

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

[Hide sidebar >>](#)

[Table of contents](#)

[Summary](#)

[Genomic context](#)

[Genomic regions, transcripts, and products](#)

[Expression](#)

[Bibliography](#)

[Phenotypes](#)

[Variation](#)

[HIV-1 interactions](#)

[Pathways from PubChem](#)

[Interactions](#)

[General gene information](#)

Markers, Related pseudogene(s), Potential readthrough, Homology, Gene Ontology

[General protein information](#)

[NCBI Reference Sequences \(RefSeq\)](#)

[NCBI Reference Sequences \(RefSeq\)](#)

[Related sequences](#)

[Additional links](#)

[Locus-specific Databases](#)

NCBI Reference Sequences (RefSeq)

NEW Try the new [Transcript table](#)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. [Explain](#)

Genomic

1. [NG_005905.2 RefSeqGene](#)

Range	92501..173689
Download	GenBank , FASTA , Sequence Viewer (Graphics) , LRG 292

mRNA and Protein(s)

1. [NM_001407571.1 → NP_001394500.1 breast cancer type 1 susceptibility protein isoform 6](#)

Status: REVIEWED

Source sequence(s)	AC135721
UniProtKB/TrEMBL	A0A386INP3 , A0A386IPK6

2. [NM_001407581.1 → NP_001394510.1 breast cancer type 1 susceptibility protein isoform 7](#)

Status: REVIEWED

Source sequence(s)	AC135721
UniProtKB/TrEMBL	A0A2R8Y7V5 , A0A386INP3 , A0A386IPK6 , G8I0D8
Related	ENSP00000496570.2 , ENST00000644379.2

3. [NM_001407582.1 → NP_001394511.1 breast cancer type 1 susceptibility protein isoform 7](#)

Status: REVIEWED

Source sequence(s)	AC135721
UniProtKB/TrEMBL	A0A2R8Y7V5 , A0A386INP3 , A0A386IPK6 , G8I0D8

4. [NM_001407583.1 → NP_001394512.1 breast cancer type 1 susceptibility protein isoform 2](#)

Status: REVIEWED

Source sequence(s)	AC135721
UniProtKB/TrEMBL	A0A386INP3 , A0A386IPK6 , G8I0D8

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_	<u>mRNA</u>	Protein-coding transcripts (usually curated)
NR_	<u>RNA</u>	Non-protein-coding transcripts
XM_ ^c	<u>mRNA</u>	Predicted model protein-coding transcript
XR_ ^c	<u>RNA</u>	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

```
GTCGCATCGAGGCCCGCCCCCTGAGCCTGGTAGCGGCGGAGGGCCGGAGAACGTTCGGGAGGA  
AGCGAACTAGTAGTTGGGGCGGCCACGGCGCCGATGGTCAGCTTCCCTCGGGAGGAACGATGTG  
AGGGAGGGCTGGCAAGAGATTGGAATTCCGGAGGCCGGAGCTTGCTGAAACCCCTGCTAGGA  
GCGGGCAACTAGTGTCTAGTGAGGGGTTGGCTGGCGCACTGATCCCAGACTTCCGATCTCTG
```

FASTA ▾

glutathione synthetase [Homo sapiens]

NCBI Reference Sequence: NP_000169.1

[GenPept](#) [Identical Proteins](#) [Graphics](#)

>NP_000169.1 glutathione synthetase [Homo sapiens]

```
MATNWGSLLDKQQLEELARQAVDRALAEGVLLRTSQEPTSSEVSYAPFTLFPSSLVPSALLEQAYAVQM  
DFNLLVDASQNAFLEQTLSSTIKQDDFTARLFDIHKQVLKEGIAQTVFLGLNRSDYMFQRSADGSPAL  
KQIEINTISASFGLASRTPAVHRHVLSKTEAKGILSNNSPKGLALGIKAWELEYGSPNALVLLIA  
QEKENRNFQRAIENELLARNIHIWRRTFEDISEKGSLDQDRRLFVDGQEIAVVYFRDGYMPRQYSLNQW  
EARLLEERSHAAKCPDIATQLAGTKVQQEELSRRGMLLEMPLPGQPEAVARLRTFAGLYSLDVGEEGDQA  
IAEALAAPSRFVLPKPREGGGNLNYGEEMVQALKQLDSEERASYILMEKIEPEPFENCLLRPGSPARV  
QCISELGIFGVYVRQEKTLMVNKHVGHLRTKAEHADGVAAGVAVLNDNPYPV
```

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

```
GTCGCATCGAGGCCCGCCCCCTGAGCCTGGTAGCGGCGGAGGGCCGGAGAACGTTCGGGAGGA  
AGCGAACTAGTAGTGGGAGGCCACACTGGGGAGCCTGCTGAGGATAAACAGCAGCTAGAGGAGC  
TGGCACGGCAGGGCTGGACCGGGCCCTGGCTGAGGGAGTATTGCTGAGGACCTCACAGGAGCCACTTC  
CTCGGAGGGTGTGAGCTATGCCCATTCACGCTCTCCCTCACTGGTCCCAGTGGCTGCTGGACAA  
GCCATGCTGAGATGACTTCAACCTGCTAGTGGATGCTGTCAGCCAGACGCTGCTTCTGGAGAC  
AAACTCTTCCAGCACCATAACAGGATGACTTTACCGCTGCTCTTGGACATCCAAAGCAAGTCTC  
AAAAGAGGGCATTGCCAGCACTGTGTTCTGGGCTGAATGCTCAGACTACATGTTCCAGCGAGCGA  
GATGGCTCCCAGCCCTGAAACAGATCGAAATCACACCATCTGCGAGCTTGGGGCTGGCCCTCC  
GGACCCCAGCTGTGACCCGACATGTTCTCAGTGTCTGAGTAAGACCAAAGAACGCTGGCAAGATCCTC  
TAATAATCCCAGCAAGGACTGGCCCTGGGAATTGCCAAAGCCTACGGCTCACCCAAATGCT  
CTGGTGTACTGATTGCTCAAGAGAAAGGAAACATATTGACCAGCTGCTCATAGAGAATGAGCTAC  
TGGCAGGAAACATCCATGATCGACGAATTGGAATGATCTCTGAAAAGGGTCTGGACCAAGA  
CCGAAGGGTGTGGATGGCAGGAATTGCTGTTTACTCCGGATGGTACATGCTCGTCAG  
TACAGTCAGAATGGGAAGACGTCAGTGTGGAGAGGTCACATGCTGCGCAAGTGCCAGACATTG  
CCACCCAGCTGGCTGACGGCAATAGAAGGTGACAGGAGCTAACAGGCCGGCATGCTGGAGATGTTGCT  
CCCTGGCCAGCTGGGCTGTCGGCCCTCGGCCACCTTCTGCTGGCTTACTCACTGGATGTTGCT  
GAAGAAGGGGACCGGCCATCGCCAGGCCCTGCTGCCCTAGCCGGTTGCTAAAGCCCCAGAGAG  
AGGGTGAGGTAAACAACCTATATGGGAGGAATGGTACAGGCCCTGAAACAGCTGAAGGACAGTGGAGA  
GAGGGCTCTACATCTCATGGAGAAGATCGAACCTGAGCCTTTGAGAATTGCTGCTACGGCTGGC  
AGCCCTGGCAGTGGTCAGTCAGTCAGGCTGGGATCTTGGGCTTATGTCAGGCAGGAAAAGA  
CACTCGTGTGAACAAGCAGCTGGGGCATCTACTTCGAACCAAAGCCTACATGAGCTGAGATGGTGT  
GGCAGCGGGAGTGGCAGTCTGGACAACCCATACCTGTTGAGGGCACAACAGGCCACGGGACCTTCT  
ATCCTCTGTTATGCTCTCTCTAGGACCTCTGAGGGTATCTCTAAAGACCTCAAAGTTTTT  
ATGGAAGGGTAATACTGGTACCTCCCCACGCTTCCATCTGAGGACAGAAAAGTTGTTGCTCCCTTA  
GATGAGATCTAGACGCCCAATCTTGGAGATGGGATATGCTCAGGGTAAGCTGCTGAGGTAAG  
GTCATGAACCCCTGCCCACTCTGTCAGCCCCCTCATCAGCCTTTCAGCAGGTTCCAGTGCCTGACTT  
GGATAGGACTGAGTGGTAGGAGGGGGAGTGGAGGGCATAGCCTTCCCTAATTGCTGCTTAAATAA  
AACTGCATTGCTGATTCA
```

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

Nucleotide Advanced Search Builder

Use the builder below to create your search

[Edit](#) [Clear](#)

Builder

All Fields AND All Fields

or [Add to history](#)



<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

GAACTCGGGAAGCCGGCAGAAGTGTGAGGCCGG
TAGGGCCGCATCCCGCTCCGGAGAGAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCGGAAAGCTCGGTAA
TGATAAGCACGCCGCCACTTGCAGGGCGTCAC
CGCCTACAGCCCCCTCGTCTCGGACGGCGCGT
CTAGCCTGGGGCGCTGGCCCGCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCAATAGGC
CGGCTTGACCCGCAACAGGGAGGGTTCCCGGG
GGAGTGGCGCGCAGAAGGCCCGCCAGGAGGCCA
GGGACAGCCCAGAGGAGGCGTGGCACGCTGCCG
GCGGAAGTGGAGCCCTCCGCAGCGCGAGGCCGC
CGGGCAGGCGGGAAACCGGACAGTAGGGCGG
GGCCGGGCCGGCGATGGGGATGCGGGAGACTACGC
GGAGCTGCACCCGTGCCGCCGAATTGGGGATG
CAGAGCAGCGGCAGGGGTATGGCAGGCCAGCGCG
GGCCGGCCCTCAGCGCAGGTGCCGAGAGGCAGG
GGCTGGCCTGGGATGCGCGCACCTGCCCTCGCC
CGCCCCGCCGCACGAGGGGTGGTGGCCGAGGCC
CGCCCCCGCACGCCCTGCCGTAGGCAGGTCCGCTC

MGRRGSAPGNRTRLGCERGGRRRSADSVMAEQ
VALSRTQVCGILREELFQGDAFHQSDTHIFIIMG
ASGDLAKKIIYPTIWWLFRDGLLPENTFIVGYAR
SRLTVADIRKQSEPFFKATPEEKLKLEDFFFARN
YVAGQYDDAASYQRLNSHMNALHLGSQANRLFYL
ALPPTVYEAVTKNIHESCMSQIGWNRIIVEKPF
RDLQSSDRLSNHISSLFREDQIYRIDHYLGKEMV
QNLMVLRFANRIFGPIWNRDNIACVILTFKEPFG
TEGRGYFDEFGIIRDVMQNLLQMLCLVAMEKPA
STNSDDRDEKVVLKCISEVQANNVVLQYVGN
PDGEGEATKGYLDPTVPRGSTTATFAAVVLYVE
NERWDGVFILRCGKALNERKAEVRLQFHVDVAGD
IFHQQCKRNELVIRQPNEAVYTMMTKKPGMFF
NPEESELDTYGNRYKNVLPAYERLILIDVFCG
SQMHFVRDELREWIFTPLLHQIELEPKPKIP
YIYGSRGPTEADELMKRVGFQYEGTYKWVNPHKL

>NC_000023.11:c154547569-154531390
Homo sapiens chromosome X,
GRCh38.p14 Primary Assembly
GAACTCGGGAAGCCGGCGAGAAGTGTGAGGCCGG
TAGGGCCGCATCCCGCTCCGGAGAGAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCGCCGGAAAGCTCGGTAA
TGATAAGCACGCCGCCACTTGCAGGGCGTCAC
CGCCTACAGCCCCCTCGTCTCGGACGGCGCGT
CTAGCCTGGGGCGCTGGCCGCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCAATAGGGC
CGGCTTGACCCCGAACAGGGAGGGTTCCCGGG
GGAGTGGCGCGCAGAAGGCCGCCAGGAGGCCA
GGGACAGCCCAGAGGAGGCGTGGCCACGCTGCCG
GCGGAAGTGGAGCCCTCCGCAGCGCGAGGCCGC
CGGGCAGGCGGGAAACCGGACAGTAGGGCGG

Fasta Format

- **Fasta:** Simple Formats for Sequences

Header

- filename.**fasta**, filename.**fa**, filename.**fna**, and etc.

The diagram illustrates the Fasta format with a yellow box highlighting the header line. Two black arrows point from the left and right towards the sequence line, which starts with '>NC_000023.11:c154547569-154531390'. The sequence line contains a long string of DNA bases.

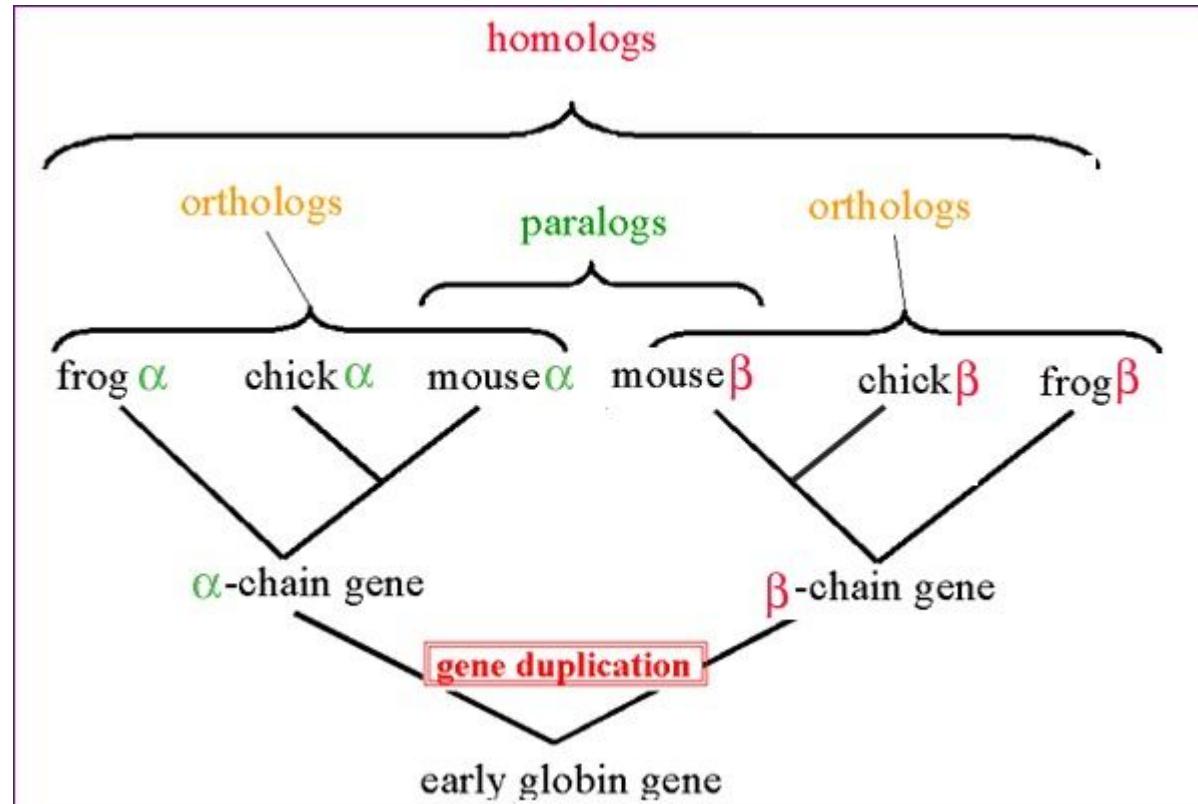
>NC_000023.11:c154547569-154531390 Homo sapiens chromosome X, GRCh38.p14 Primary Assembly
GAACTCGGGAAGCCGGCGAGAAAGTGTGAGGCCGCGTAGGGCCGCATCCCGCTCCGGAGAGAAAGTCTGAG
TCCGCCAGGCTCTGCAGGCCCGCGGAAGCTCGTAATGATAAGCACGCCGCCACTTGCAGGGCGTCAC
CGCCTACACGCCCTCGTCTCGGACGGCGCGTCTAGCCTGGGGCGCTGGCCGCCCGCCCTCTC
CGGGGGAGGAATCAAGAAGAGACTGCCAATAGGGCCGGCTTGACCCGCGAACAGGGCAGGGTTCCCGGG
GGAGTGCGCGGCAGAAGGCCGCCAGGAGCCGAGGGACAGCCCAGAGGAGGCGTGGCACGCTGCCG
GCGGAAGTGGAGGCCCTCCGCGAGCGCGAGGGCGCCGGGCAGGCGGGAAACCGGACAGTAGGGCGG

Sequence

```
>Seq1 [organism=Carpodacus mexicanus] [clone=6b] actin (act) mRNA, partial cds
CCTTATCTAATTTGGAGCATGAGCTGGCATAGTTGAAACGCCCTCAGCCTCCTCATCCGTGCAGAATAATAATTCTTATAGTAATACCAATCATGATCGGTGGT
TTCGGAAACTGACTAGTCCCCTACTCATAAT
>Seq2 [organism=uncultured bacillus sp.] [isolate=A2] corticotropin (CT) gene, complete cds
GGTAGGTACCGCCCTAAGNCCTTAATCCGAGCAGAACTANGCCAACCCGGAGCCCTCTGGGAGACGACTCAACACCACCTCTTGACCCAGCAGGAGGAGACCCA
GTACTATAACCAGCACCTATTCTGATTCTT
>Seq3 [organism=Phalaenopsis equestris var. leucaspis]
CCTATACTTAATTTGGCGCATGAGCCGAATGGTGGGTACCGCTCTAACGCCCTCATTGAGCAGAACTAGGCCAACCCGGAGCCCTCTGGGAGACGACCAAGTCTA
CAACGTGGTTGTACGGCCCATGCCTCG
>Seq9 [organism=Petunia integrifolia subsp. inflata]
TAGTTGAAACGCCCTCAGCTACTCATCCGAGCAGAACTAGGCCAACCCGGAAACCCCTCTGGGAGATGACCAAATCTACAATGTAATCGTCACTGCCCATGCCTCGAA
TAATCTTCTTATAGTAATACCAAGTCATA
```

Homology

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

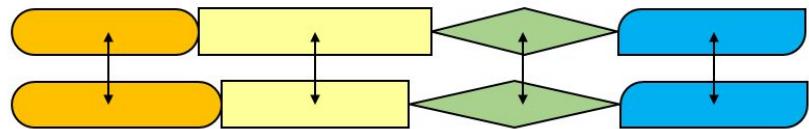


BLAST

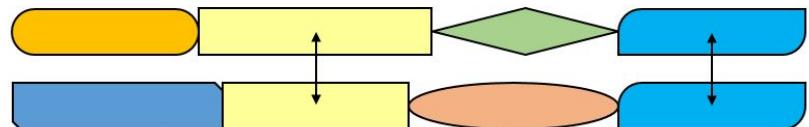
- BLAST tells you about **Non-chance similarities between biological sequences**
- **If similarity do not due to chance then they must be due to something else**
 - **Homology:** Similar sequences -> evolutionary relationship -> related function
 - **Identification** (Present or absent in database)
- 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 ATGACTTTACCAGGTGGTGTATTAGTTCTAGTTGGTCGGTTGGCTTGATTGCCATTATT
SEQUENCE 2 ATGATTTGTCAGGTACCAATTGCT

Global alignment

1	ATGACTTTACCAGGTGGTGTATTAGTTCTAGTTGGTCGGTTGGCTTGATTGCCATTATT	66
2 ATGATTTGTCAGG-----TACCAATT-----GCT-----	24

Local alignment

1	ATGACTTTACCAGGT	15
	. ..	
1	ATGATTTGTCAGGT	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

$$1+1+0+0+1=3$$

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

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
1+1+0+1=3
ATTG

AT - C G
AT T - G

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

ATC - - TA
ATT T T TA

AT - C - TA
AT T T T TA

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

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...

AGKVGL...

WGK – GI...

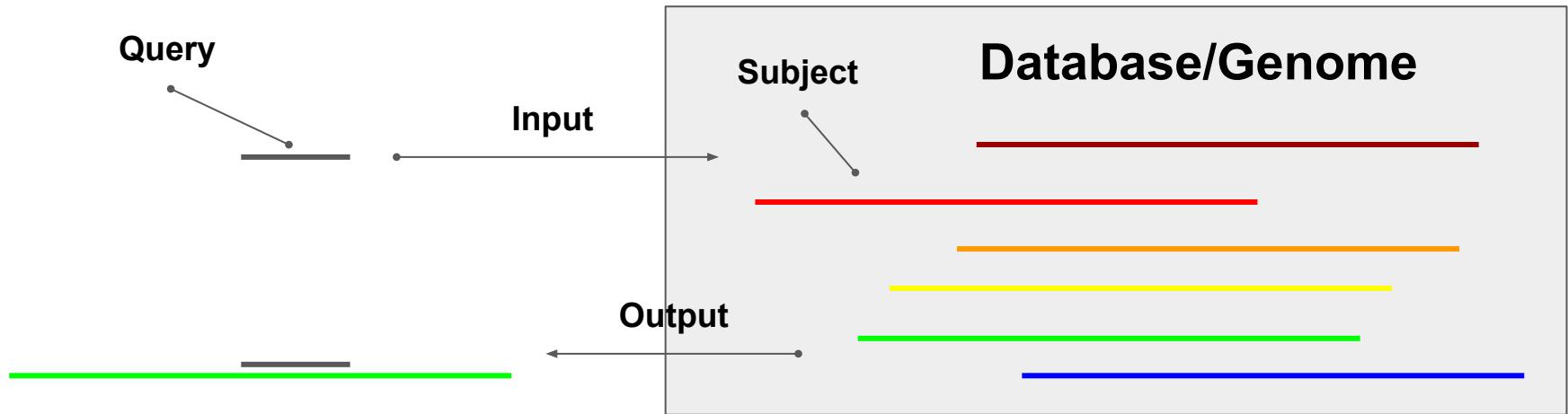
AGKVGL

Protein Scoring Systems

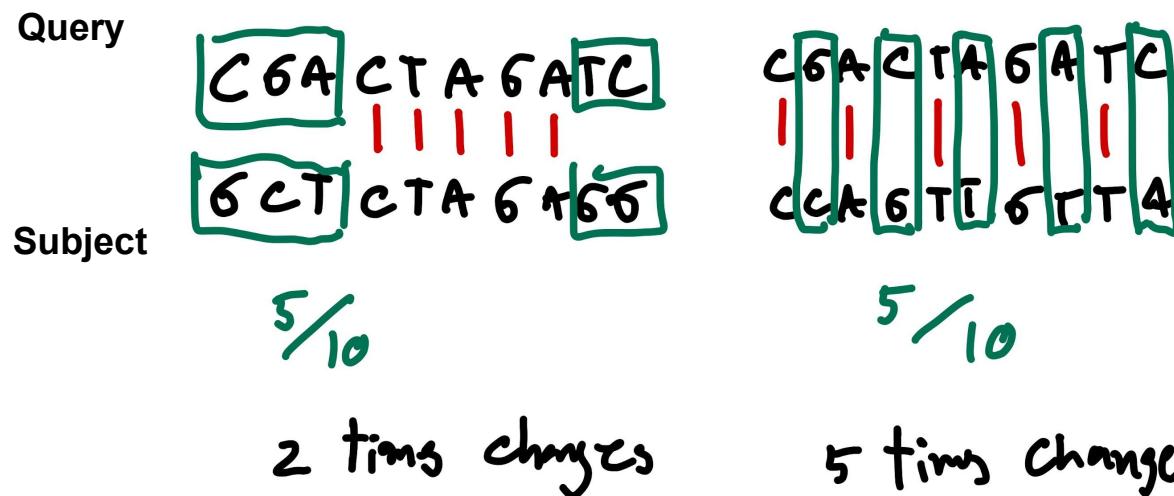
- Scoring matrix
 - Substitution matrix
 - BLOSUM
 - PAM

BLAST Algorithm

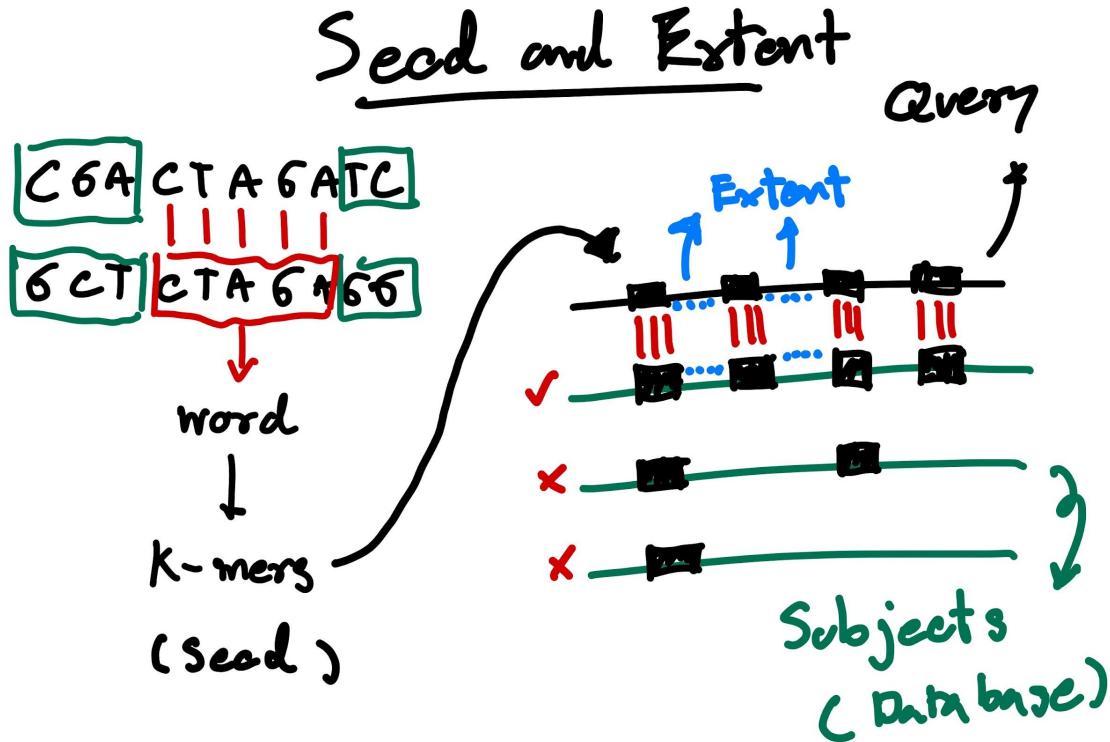
- Generate words (K-mers) from sequence
- Seed and Extent: Search words and extent scoring in Database
- Local Alignment: Smith–Waterman algorithm
- Statistic Calculation: Matching by change or not



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
T	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

	L	R	P	S
query	TTA	AGG	CCA	TCA
subject	CTT	CGT	CCC	AGT
	L	R	P	S

4 match in DNA

all match in Protein

BLAST Types

- **Basic Search**

- blastn
- blastp

Program	Query Type	Subject Type	Computation
blastn	N ——————>	— N	~ 1X

- **Translated Search**

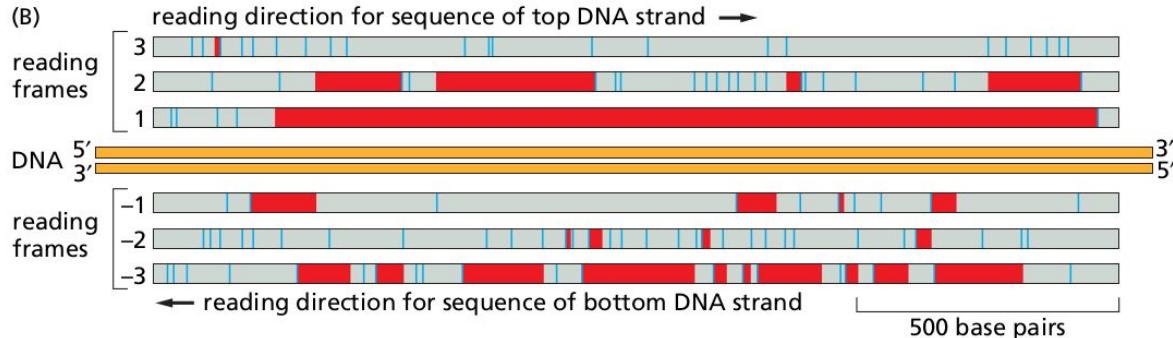
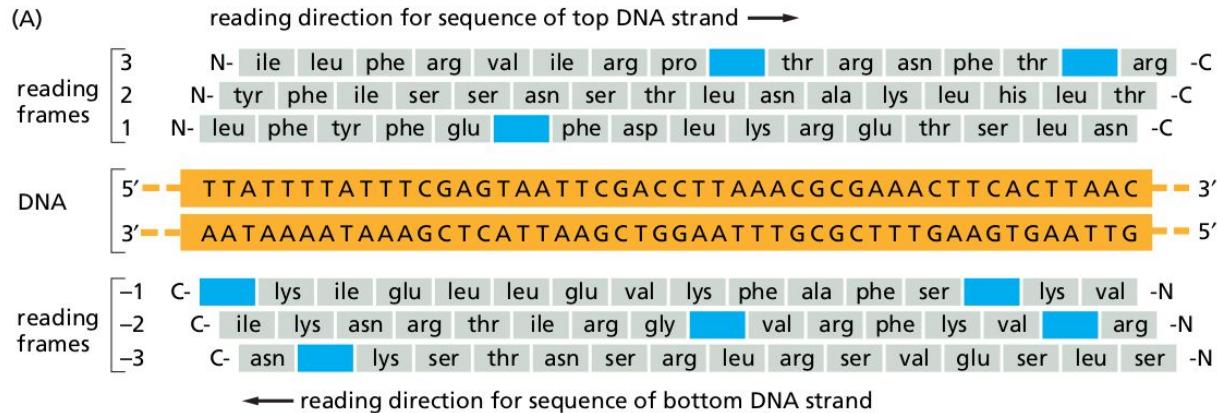
- blastx
- tblastn
- tblastx
- **6 Reading Frames**

blastp	P ——————>	— P	~ 1X
blastx	N —————>	— P	~ 6X
tblastn	P ——————>	N	~ 6X
tblastx	N —————>	N	~36X

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

Reading Frame

- 6 frames per sequences



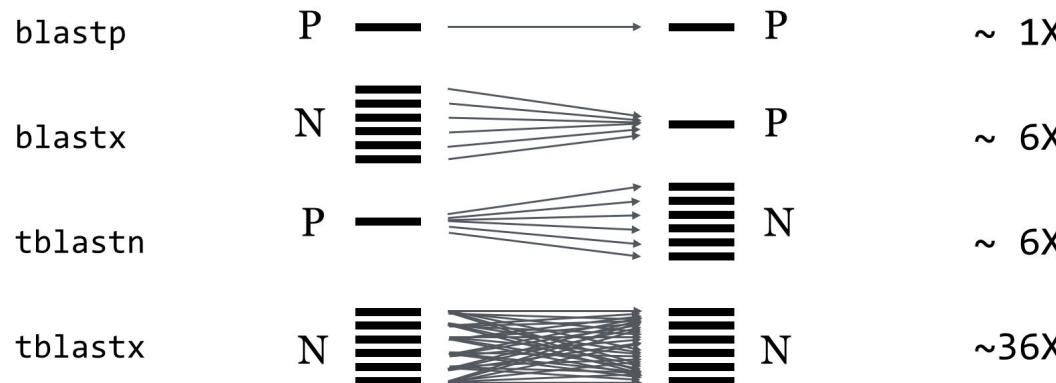
blastn

- **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





 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

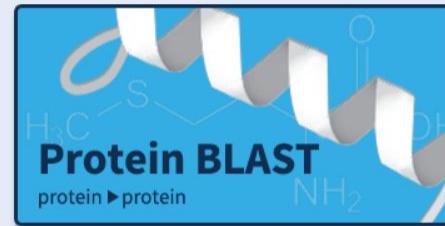
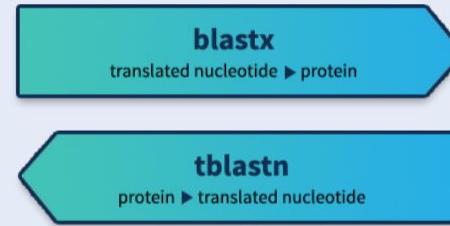
BLAST+ 2.13.0 is here!

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



Simple Searching of Nucleotide sequence

- blastn
 - Nucleotide searching

Select Database

Specific organism

BLAST algorithm

Standard Nucleotide BL...

BLASTN programs search nucleotide databases using

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Query subrange

Input sequence

Or, upload file No file chosen

Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus Experimental databases
Core nucleotide database (core_nt)

Organism exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Models (XM/XP) Uncultured/environmental sample sequences

Limit to Sequences from type material

Entrez Query Create custom database
Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST Search database core_nt using Megablast (Optimize for highly similar sequences)
 Show results in a new window

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: [?](#)

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

Choose Search Set

Database	<input checked="" type="radio"/> Standard databases (nr etc.) <input type="radio"/> rRNA/ITS databases <input type="radio"/> Genomic + transcript da
Organism Optional	Core nucleotide database (core_nt) <input type="checkbox"/> ? Core nucleotide database (core_nt) <input checked="" type="checkbox"/> RefSeq Select RNA sequences (refseq_select) Reference RNA sequences (refseq_rna) RefSeq Reference genomes (refseq_reference_genomes) RefSeq Genome Database (refseq_genomes) Nucleotide collection (nr/nt) Whole-genome shotgun contigs (wgs) Expressed sequence tags (est) Sequence Read Archive (SRA) Transcriptome Shotgun Assembly (TSA) Targeted Loci(TLS) High throughput genomic sequences (HTGS) Patent sequences(pat) PDB nucleotide database (pdb) Human RefSeqGene sequences(RefSeq_Gene) Genomic survey sequences (gss) Sequence tagged sites (dbsts)
Exclude Optional	<input type="checkbox"/> exclude be shown <input type="checkbox"/> ?
Limit to Optional	
Entrez Query Optional	

Program Selection

Optimize for

BLAST

Algorithm parameters

General Parameters

Max target sequences

Short queries Automatically adjust parameters for short input sequences

Taxonomy Browser

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

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

todste@orcid

Taxonomy [Limits](#) [Advanced](#) [Help](#)



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

[Quick Start Guide](#)
[FAQ](#)
[Handbook](#)
[Taxonomy FTP](#)
[Important Update: Phyla Changing](#)
[Important Update: New Flu species Names](#)

Taxonomy Tools

[Browser](#)
[Common Tree](#)
[Statistics](#)
[Name/ID Status](#)
[Genetic Codes](#)
[Linking to Taxonomy](#)
[Extinct Organisms](#)

Other Resources

[GenBank](#)
[LinkOut](#)
[E-Utilities](#)
[Batch Entrez](#)
[INSDC](#)

[blastn](#) **blastp** [blastx](#) [tblastn](#) [tblastx](#)

Searching of amino acid sequence

- **blastp**
 - Amino acid searching

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)
Input sequence
From _____
To _____
Or, upload file [Choose File](#) No file chosen [?](#)
Job Title _____
Enter a descriptive title for your BLAST search [?](#)
 Align two or more sequences [?](#)

Choose Search Set
Databases Standard databases (nr etc.) Experimental databases
Compare Select to compare standard and experimental database [?](#)

Standard
Database Non-redundant protein sequences (nr) [?](#)
Organism Optional Enter organism name or id—completions will be suggested exclude [Add organism](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)
Exclude Optional Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection
Algorithm Quick BLASTP (Accelerated protein-protein BLAST)
 blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm [?](#)

BLAST [Search database nr using Blastp \(protein-protein BLAST\)](#)
 Show results in a new window

Algorithm parameters

General Parameters

Max target sequences	100	?
Select the maximum number of aligned sequences to display ?		
Short queries	<input checked="" type="checkbox"/>	Automatically adjust parameters for short input sequences ?
Expect threshold	0.05	?
Word size	5	?
Max matches in a query range	0	?

Scoring Parameters

Matrix	BLOSUM62	?
Gap Costs	Existence: 11 Extension: 1	?
Compositional adjustments	Conditional compositional score matrix adjustment	?

Filters and Masking

Filter	<input type="checkbox"/> Low complexity regions	?
Mask	<input type="checkbox"/> Mask for lookup table only	?
	<input type="checkbox"/> Mask lower case letters	?

Specific parameters for BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Protein databases

nr or refseq_protein

Standard

Database

- Non-redundant protein sequences (nr)
- RefSeq Select proteins (refseq_select)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot(swissprot)
- Patented protein sequences(pataa)
- Protein Data Bank proteins(pdb)
- Metagenomic proteins(env_nn)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Choose a BLAST algorithm

Program Selection

Algorithm

selected exclude Add organism

taxa will be shown

(WP) Uncultured/environmental sample sequences

rated BLAST)

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](#)

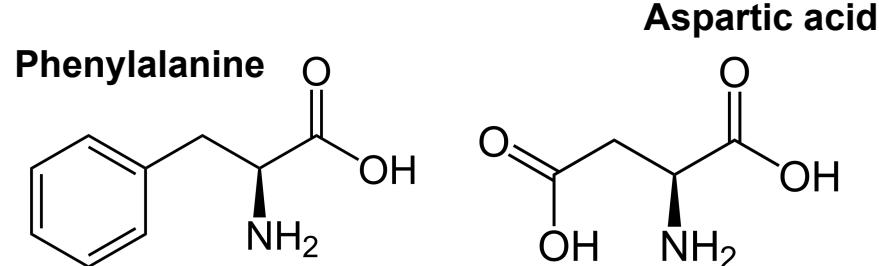
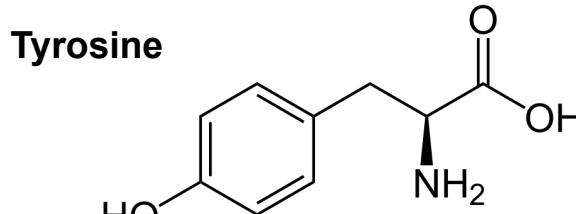
Score	Expect	Method	Identities	Positives	Gaps
299 bits(765)	4e-102	Compositional matrix adjust.	146/146(100%)	146/146(100%)	0/146(0%)
Query 1	VHLTPEEKNAVTLWGKVNDEVGGEALGRLLL	VYPWTQRFFESFGDLSSPDAVMGNPKV	60		
Sbjct 1	VHLTPEEKNAVTLWGKVNDEVGGEALGRLLL	VYPWTQRFFESFGDLSSPDAVMGNPKV	60		
Query 61	KAHGKKVLGAFSDGLNHLDNLKGTF	AQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK	120		
Sbjct 61	KAHGKKVLGAFSDGLNHLDNLKGTF	AQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK	120		
Query 121	EFTPQVQAAYQKV	VAGVANALAHKYH	146		
Sbjct 121	EFTPQVQAAYQKV	VAGVANALAHKYH	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



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	VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLL VYPWTQRFFESFGDLSSPDAVMGNPKV				60
Sbjct 1	VHLTPEEK+AVT LWGKVNVDEVGGEALGRLLL+VYPWTQRFFESFGDLS+PDAMGNPKV				60
Query 61	KAHGKKVVLGAFSDGLNHDNLKGTFQAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK				120
Sbjct 61	KAHGKKVVLGAFSDGL HLDNLKGTF LSELHCDKLHVDPENF+LLGNVLV VLAAHFGK				120
Query 121	EFTPQVQAAQKVVAGVANALAHKYH	146			
Sbjct 121	EFTP VQAAQKVVAGVANALAHKYH	146			
	EFTPPVQAAQKVVAGVANALAHKYH	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

Organism Optional	<input type="text" value="Reference RNA sequences (refseq_rna)"/> <input type="button" value="▼"/>
Exclude	<input type="text" value="whales, hippos, ruminants, pigs, camels etc. (taxid:91561)"/> <input type="checkbox"/>
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown <input type="button" value="?"/>	
<input type="checkbox"/> Models (XM/XP) <input type="checkbox"/> Uncultured/environmental sample sequences	

The screenshot shows the NCBI BLASTN search interface. The 'Standard' tab is selected at the top. The main search area has the accession number 'ref|NM_001824.5' entered into the 'Enter Query Sequence' field. Below it, there are fields for 'Or, upload file' (Choose File, No file chosen), 'Job Title' (ref|NM_001824.5|), and a checkbox for 'Align two or more sequences'. The 'Choose Search Set' section includes a 'Database' dropdown (set to Standard databases (nr etc.)), an 'Organism' dropdown (set to Nucleotide collection (nr/nt)), and an 'Exclude' section with checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'. The 'Program Selection' section shows 'Optimize for' set to 'Highly similar sequences (megablast)'.

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ref|NM_001824.5

Query subrange From To

Or, upload file No file chosen

Job Title ref|NM_001824.5| Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

New Experimental databases For more info see What are taxonomic nt databases?

Organism Nucleotide collection (nr/nt)

Enter organism name or id—completions will be suggested exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Models (XM/XP) Uncultured/environmental sample sequences

Limit to Sequences from type material

Entrez Query Enter an Entrez query to limit search

YouTube Create custom database

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database nt using Megablast (Optimize for highly similar sequences) Show results in a new window

Annotating Gene by mRNA sequence

- BLAST Genomes
- NM_001824.5
- Grey whale

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

gray whale (taxid:9764)

Enter organism common name, scientific name, or tax id.
Human Mouse Rat Microbes

BLAST Genomes

grey whale (taxid:9764)

grey whale (taxid:9764)

blastn blastp blastx tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Query subrange ?
From _____ To _____

Or, upload file No file chosen ?

Job Title
Enter a descriptive title for your BLAST search ?

Choose Search Set
Database
Exclude Models (XM/XP)
Optional
Entrez Query
Optional
Enter an Entrez query to limit search ?

Program Selection
Optimize for
1 Highly similar sequences (megablast)
2 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm ?

Search database Genome (mEscRob2.pri reference assembly GCF_028021215.1-RS_2
 Show results in a new window

Annotating Gene by Protein Sequence

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

tblastn P —————→ N ~ 6X

The screenshot shows the NCBI BLAST search interface. The top navigation bar has tabs for blastn, blastp, blastx, **tblastn**, and blastx. The main search area is titled "Enter Query Sequence" and contains a text input field with "NP_001815.2". Below it is a file upload section with "Choose File" and "No file chosen". A "Job Title" field is filled with "NP_001815:creatine kinase M-type [Homo sapiens]". The "Choose Search Set" section includes a "Database" dropdown set to "Genome (mEscRob2.pri reference assembly GCF_028021215.1-RS)", an "Exclude Optional" checkbox for "Models (XM/XP)", and an "Entrez Query Optional" input field. At the bottom, there is a large blue "BLAST" button, a search database dropdown set to "Search database Genome (mEscRob2.pri reference assembly GCF_028021215.1-RS_2024_09)", and a checkbox for "Show results in a new window". A "Algorithm parameters" section is at the very bottom.

blastp

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

blastn blastp blastx tblastn tblastx BLASTP program

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

From
To

Or, upload file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Databases Standard databases (nr etc.) Experimental databases

Compare Select to compare standard and experimental database

Standard

Database

Organism exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Optional

Program Selection

Algorithm blastp (protein-protein BLAST) PSI-BLAST (Position-Specific Iterated BLAST) PHI-BLAST (Pattern Hit Initiated BLAST) DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database refseq_protein using Blastp (protein-protein BLAST) Show results in a new window

blastp result

BLAST® » blastp suite » results for RID-RVGE2CXS013

Home Recent Results Saved Strategies Help

[Edit Search](#) Save Search Search Summary [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Your search is limited to records that include: Mammalia (taxid:40674) ; and exclude: models (XM/XP), uncultured/environmental sample sequences, non-redundant RefSeq proteins (WP)

Job Title NP_001815:creatine kinase M-type [Homo sapiens]
RID RVGE2CXS013 Search expires on 01-09 07:07 am [Download All](#)
Program BLASTP [Citation](#)
Database refseq_protein [See details](#)
Query ID NP_001815.2
Description creatine kinase M-type [Homo sapiens]
Molecule type amino acid
Query Length 381
Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name
[+ Add organism](#)

Percent Identity E value Query Coverage

to to to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Select columns Show 100 [?](#)

select all 4 sequences selected GenPept Graphics Distance tree of results [Multiple alignment](#) MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	creatine kinase M-type [Homo sapiens]	Homo sapiens	790	790	100%	0.0	100.00%	381	NP_001815.2
<input checked="" type="checkbox"/>	creatine kinase M-type [Sus scrofa]	Sus scrofa	773	773	100%	0.0	97.11%	381	NP_001123421.1
<input checked="" type="checkbox"/>	creatine kinase M-type [Canis lupus familiaris]	Canis lupus familiaris	772	772	100%	0.0	96.33%	381	NP_001300705.1
<input checked="" type="checkbox"/>	creatine kinase M-type [Mus musculus]	Mus musculus	770	770	100%	0.0	96.59%	381	NP_031736.1

COBALT

COBALT Constraint-based Multiple Alignment Tool

Home Recent Results Help

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

Find: Tools Columns Rows Download Coloring

Sequence ID	Start	End	Organism
NP_001815.2	1	381	Homo sapiens
NP_00112421.1	1	381	Sus scrofa
NP_001303796.1	1	381	Canis lupus familiaris
NP_031736.1	1	381	Ovis canadensis

PROTEIN: 1 - 381 (381 shown) Rows shown: 4/4

Descriptions Select All Realign > Alignment parameters

Accession	Description	Links
NP_001815.2	creatine kinase M-type [Homo sapiens]	Related Information
NP_00112421.1	creatine kinase M-type [Sus scrofa]	
NP_001303796.1	creatine kinase M-type [Canis lupus familiaris]	
NP_031736.1	creatine kinase M-type [Ovis canadensis]	

COBALT Constraint-based Multiple Alignment Tool Phylogenetic Tree View This tree is based on COBALT multiple alignment [here](#).

Alignments Select All Realign Mouse over the sequence identifier for sequence title

View Format Compact Conservation Setting 2 Bits

Tree method Neighbor Joining Max Seq Difference 0.05 Distance Origin (protein) Sequence Label Sequence Title (if avail.) Number of Seqs: 4

Mouse over an internal node for a subtree or alignment. Click on tree label to select sequence to download Hide legend

Label color map Best names color map

inverted triangles & white nodes

dark nodes

Success

creatine kinase M-type [Homo sapiens]

creatine kinase M-type [Sus scrofa]

creatine kinase M-type [Canis lupus familiaris]

creatine kinase M-type [Ovis canadensis]

Notes 7/0 infected View port at (0.0) of 1797x595

END