

# **Introduction to Bioinformatics**

**SC113603 Molecular Biology**  
Lect. Todsapol Techo

# How human understand Hieroglyphs?

Egyptian hieroglyphs



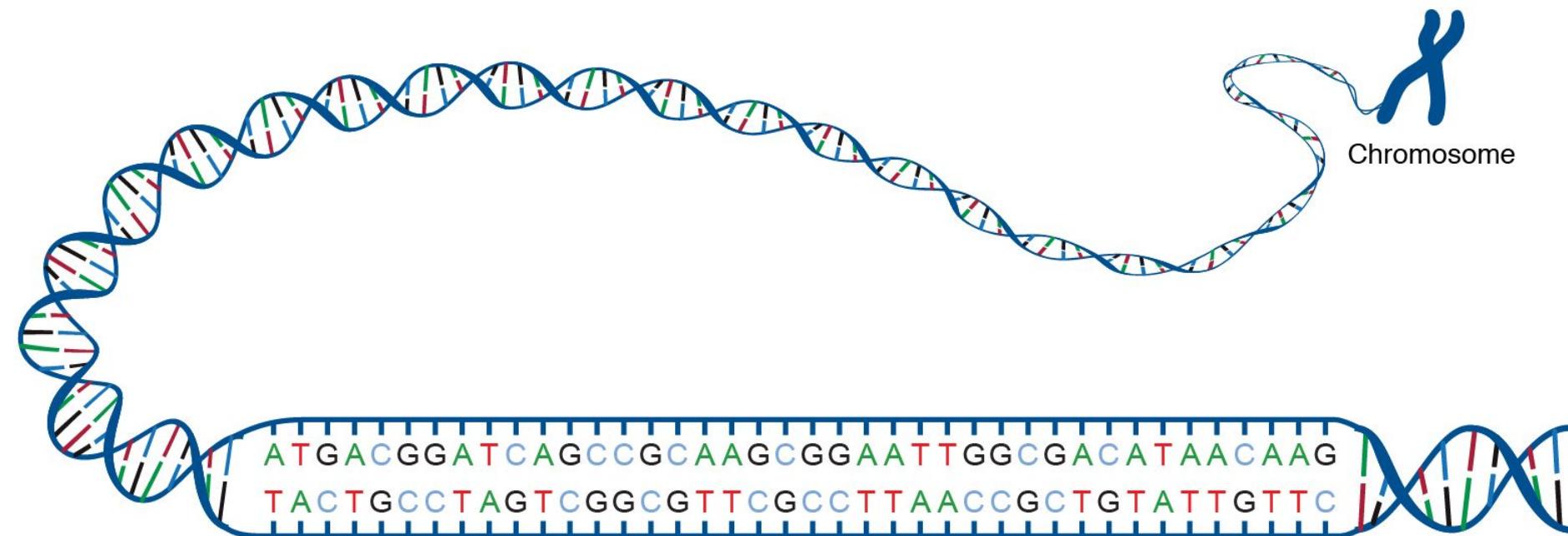
Demonic Text



- Let we are understand Ancient Egyptian life and culture

Greek Text

# The Languages of Life



GCTCGAGAGATTCTTTACCTTTACTATTTCACTCTCCATAACCTCCTATATTGACTGAT  
CTGTAATAACCACGATATTATTGGAATAAAATAGGGGCTTGAAATTGGAAAAAAA  
TTTCGTGATAAGTGTAGTGTATTCTTCTTTATTGCTACTGTTACTAAGTCTCATGTA  
GATTGCTTCATTCTTTGTTGCTATATTATGTTAGAGGTTGCTGCTTGGTTATTGATA  
TGGTATGTGAAAGCCGGTTTGCCGGTGACGACGCTCCTCGTGCTGTCTCCC  
CCAAGACACCAAGGTATCATGGTCGGTATGGGTCAAAAAGACTCCTACGTTGGT  
AGAGAGGGTATCTGACTTACGTTACCAATTGAACACGGTATTGTCACCAACT  
GATCTGGCATCACCTTCTACAACGAATTGAGAGTTGCCCAAGAAGAACACCC  
GCTCCAATGAACCCCTAAATCAAACAGAGAAAAGATGACTCAAATT  
CCTTCTACGTTCCATCCAAGCCGGTTGTCCTGTACTCTTCCGGTAGAA  
TTCCGGTGATGGTGTACTCACGTCGTTCAATTACGCTGGTTCTCTACCT  
ATCGATTGGCCGGTAGAGAGATTGACTGACTACTTGATGAAGATCTTGAGT  
CCACCACTGCTGAAAGAGAAATTGTCCTGACATCAAGGAAAAACTATGTT  
ACAAGAAAATGCAAACCGCTGCTCAATCTTCAATTGAAAAATCCTAC  
ATCACTATTGGTAACGAAAGATTGAGAGCCCCAGAAGCTTGTCC  
CTGCCGGTATTGACCAAAACTACTTACAACCTCCATGAAAGTGT  
CGGTAACATCGTTATGTCCGGTGGTACCAACCATGTTCCCAGGT  
CTGCTGTTGGGTTGGAATTATA

GCTGAGAGATTTCTTTACCTTTTACTATTTCACTCTCCATAACCTCTTATATTGACTGAT  
CTGAATAACACGATATTGGAAATAAATAGGGCCTGAAATTGGAAAAAAAATGAAATA  
TTTCGTATAAGTGATAGTGATATTCTTCTTTATTGCTACTGTTACTAAGTCTCATGTAACATC  
GATTGCTTCACTTTGGTCTATTATATGTTAGAGGTTGCTGCTTGGTTATTGATAACGGTTC  
TGGTATGTTGAAAGCGGTTGCGGTGACGCTCTCGTGTCTTCCCATATCGTCGGTAGA  
CCAAGACACCAAGGTATCGTGTGGTGGTCAAAAAGACTCTACGTTGGTAGAAGCTAATCCA  
AGAGAGGTACTTACGTTACCTTACGGTACCGGATATGGACAGCTGGGAGATATGGAAAAA  
GATCTGGATCATACCTTCTACAACGAATTGGAGGTTGCCAGAAGAACACCCGTTCTTGGACTGAA  
GCTCCAATGAAACCTTAATCAAACAGAGAAAAGATGACTCAATTATGTTGAAACTTCAACGTTCAAG  
CCTTCTAGTTCCATCCAAGCGTTGCTCTACTCTCCGGTAGAAACTACTGGTATTGTTGG  
TTCCGGTAGTGGTACTACGCTGCTTCAATTACGCTGCTCTACCTCACGCACTTTGGAGA  
ATCGATTGGCGGTAGAGATTGACTGACTCTGATGAAGATCTTGAGTAGCTGGTTACTCTT  
CCACCACTGCTGAAAGAGAAATTGCTGTGACATCAAGGAAAATATGTTACGTCGCTTGGACTTGA  
ACAAGAAATGCAAACCGCTCAATCTTCAATTGAAAATCTCACGCAACTTCCAGATGGTCAAGTC  
ATCACTATTGGTAACGAAAGATTAGGCCCCAGAAGCTTCCATCTCTGGTTGGGGTTGGAGA  
CTGCCGTATTGACCAAACACTTACAACTCCATCATGAAGTGTGATGTCGATGCCGTAAAGGAATTATA  
CGGTAACATCGTATGCCGGTACCCATGTCAGGTATTGCCGAAAGATGCAAAGGAATTAC



~ 98.8% similarity

GCTGAGAGATTTCTTTACCTTTTACTATTTCACTCTCCATAACCTCTTATATTGACTGAT  
CTGAATAACACGATATTGGAAATAAATAGGGCCTGAAATTGGAAAAAAAATGAAATA  
TTTCGTATAAGTGATAGTGATATTCTTCTTTATTGCTACTGTTACTAAGTCTCATGTAACATC  
GATTGCTTCACTTTGGTCTATTATATGTTAGAGGTTGCTGCTTGGTTATTGATAACGGTTC  
TGGTATGTTGAAAGCGGTTGCGGTGACGCTCTCGTGTCTTCCCATATCGTCGGTAGA  
CCAAGACACCAAGGTATCGTGTGGTGGTCAAAAAGACTCTACGTTGGTAGAAGCTAATCCA  
AGAGAGGTACTTACGTTACCTTACGGTACCGAATTGAAACAGGTATTGTCACCAACTGGGAGATATGGAAAAA  
GATCTGGATCATACCTTCTACAACGAATTGGAGGTTGCCAGAAGAACACCCGTTCTTGGACTGAA  
GCTCCAATGAAACCTTAATCAAACAGAGAAAAGATGACTCAATTATGTTGAAACTTCAACGTTCAAG  
CCTTCTAGTTCCATCAACGGCTTGTCTTGTACTCTCCGGTAGAAACTCTGTTGGGGTTGGAGA  
TTCCGGTAGTGGTACTACGTCGTTCAATTACGCTGGTAGTCTTCTACCTCACGCCATTGGAGA  
ATCGATTGGCGGTAGAGATTGACTGACTCTGATGAAGATCTTGAGTAGCTGGTTACTCTT  
CCACCACTGCTGAAAGAGAAATTGCTGTGACATCAAGGAAAATATGTTACGTCGCTTGGACTTGA  
ACAAGAAATGCAAACCGCTCAATCTTCAATTGAAAATCTCACGAAACTTCCAGATGGTCAAGTC  
ATCACTATTGGTAACGAAAGATTAGGCCCCAGAAGCTTGGTTCATCTCTGGGGTTGGAGA  
CTGCCGTATTGACCAAACACTTACAACTCCATCATGAAGTGTGATGTCGATGCCGTAAAGGAATTATA  
CGGTAACATCGTATGCCGGTACCCATGTCAGGTATTGCCGAAAGATGCAAAGGAATTAC

Similarity ?

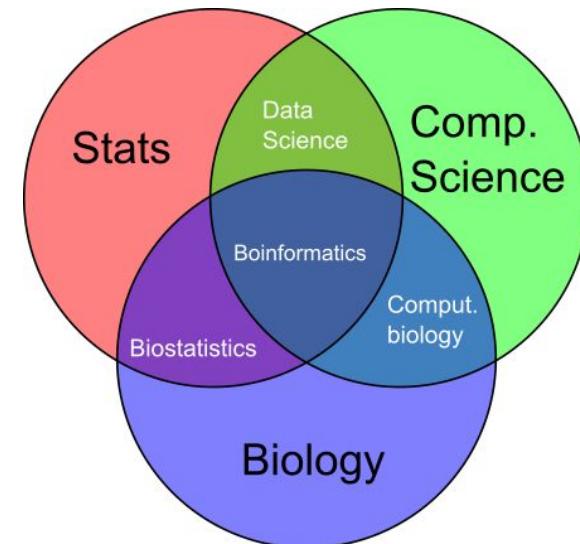


# What is Bioinformatics?

- 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

# Bioinformatics

Find the ways to understand biological phenomenon from Biological Data.



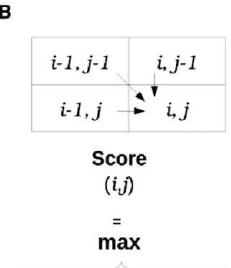


## Molecular Biology

# 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



**C**

Best Alignment : ATCG  
 (Score = 38)

```

  ATCG
  ||| |
  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
  
```

# Bioinformatics and Computational Biology

- Bioinformatics: Create tools for solving biological problems
  - Develop or Build tools (algorithm, program, pipeline or database) for solving biological problems using biological data.
- Computational Biology: Study Biology by using Computational Techniques
  - Learn biological knowledge about living systems
- The levels of Bioinformatician and Comp. Biologist
  - **Level 1 (Entry): Use public tools to generate experimental hypotheses**
  - Level 2 (Bio.Info.): Develop algorithms and database for analysis
  - Level 3 (Com.Bio.): Make biological discoveries from public data integration and modeling

# History of Bioinformatics

## Protein Wave

1955: Protein sequencing

1970: Needleman-Wunsch Algorithm  
(Global Sequence Alignment)

1973: Protein Data Bank (PDB)

1990: BLAST

1994-: Critical Assessment of  
Structure Prediction (CASP)

2020: AlphaFold2

Learning the rule from real DATA

## Nucleotide Wave

1953: DNA structure

1972: Recombinant DNA

1977: Sanger sequencing

1985: PCR

1988: NCBI

1990: BLAST

1995-2003: Human  
Genome sequencing

2005-: NGS

## Expression Wave

1995: Microarray

2006: RNA-seq  
(RNA sequencing)

2008: scRNA-seq  
(Single Cell RNA-seq)

# History of Bioinformatics: Protein Wave

- 1955: Sanger Sequence Bovine Insulin (**Protein sequence**)
- 1970: Needleman-Wunsch Algorithm (**Pairwise Sequence Alignment**)
- 1973: **Protein Data Bank (PDB)**
- 1990: **BLAST (Search protein sequence in database)**
- 1994-: Critical Assessment of Structure Prediction (**CASP**) (Community for Protein Structure Prediction) (**Structural Modeling**)
- 2020: AlphaFold2 (Deep learning): Learning the rule from real DATA (**Structural Modeling**)

# *de novo* assembly of protein sequence

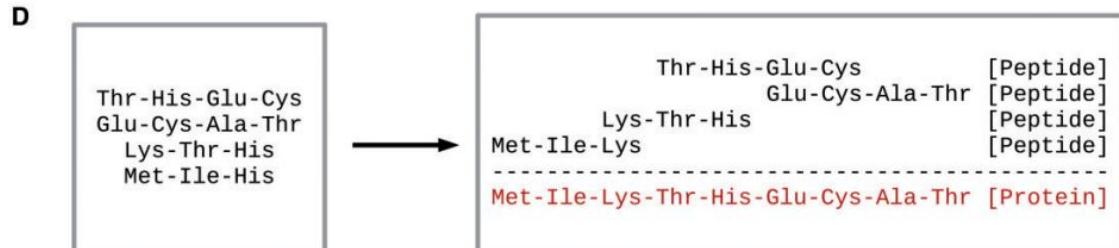


Margaret Oakley Dayhoff

**COMPROTEIN**  
*a de novo* sequence assembler

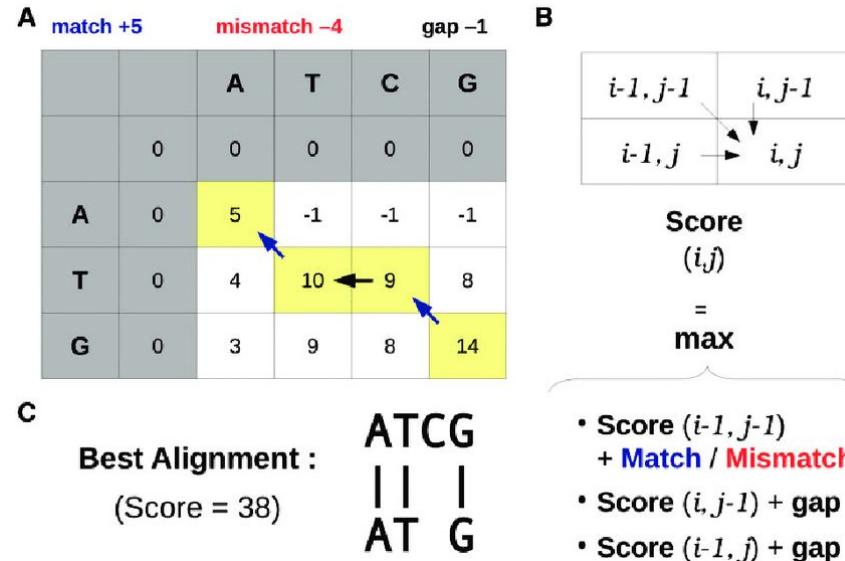
IBM 7090

FORTRAN code



# Needleman-Wunsch Algorithm (Global alignment)

- Algorithm: a process or set of rules to be followed in calculations or other problem-solving operations
- First algorithm for compare two amino acid sequences



# Protein Database

- Database for Protein structure and Protein sequences

The screenshot shows the homepage of the RCSB PDB (Protein Data Bank) website. The header includes the RCSB PDB logo, navigation links for Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, and Careers, and options for MyPDB and Contact us. The main content area features a search bar with '199,755 Structures from the PDB' and '1,000,361 Computed Structure Models (CSM)'. A banner at the top right promotes 'NEW! Computed Structure Models (CSM)'. The left sidebar has links for Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. The central column contains text about the database's mission and highlights for COVID-19 resources and seasons greetings. On the right, there is a 'December Molecule of the Month' section featuring a 'Click Chemistry' molecule.

rcsb.org

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers

MyPDB Contact us

**PDB** 199,755 Structures from the PDB  
1,000,361 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Advanced Search | Browse Annotations Help

PDB-101 www.PDB EMDDataResource NUCLEIC ACID DATABASE wwPDB Foundation

NEW! Computed Structure Models (CSM) Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

**COVID-19 CORONAVIRUS Resources**

Seasons Greetings

December Molecule of the Month

Click Chemistry

# BLAST (Local alignment)

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

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

**National Library of Medicine**  
National Center for Biotechnology Information

**Log in**

**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

# Protein Structure Prediction

- Make the rule to predict Protein Structure
- Learn the rule for predicting Protein Structure

## Article

### Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

Accepted: 12 July 2021

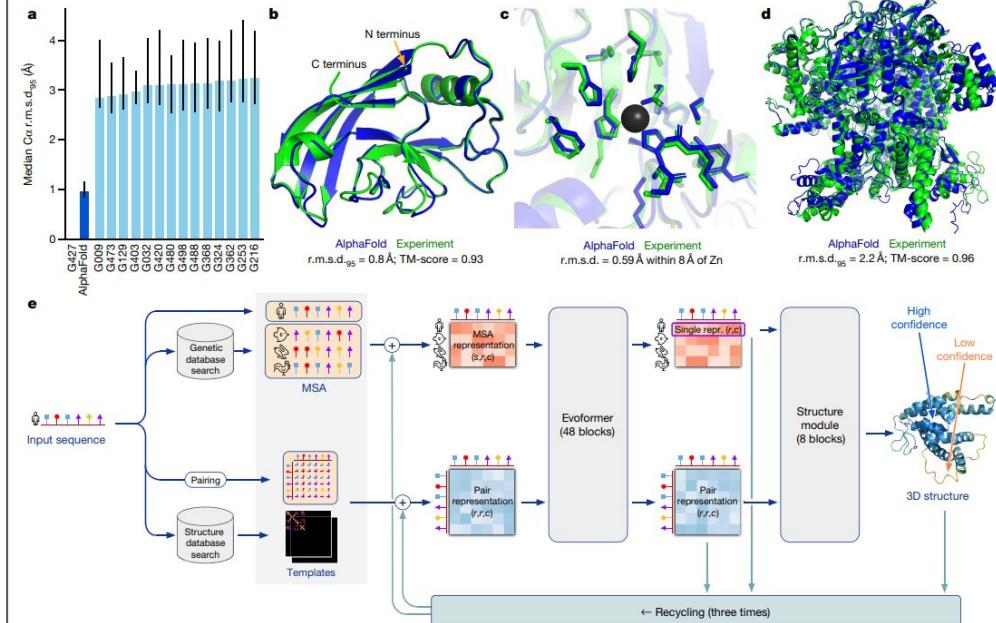
Published online: 15 July 2021

Open access

 Check for updates

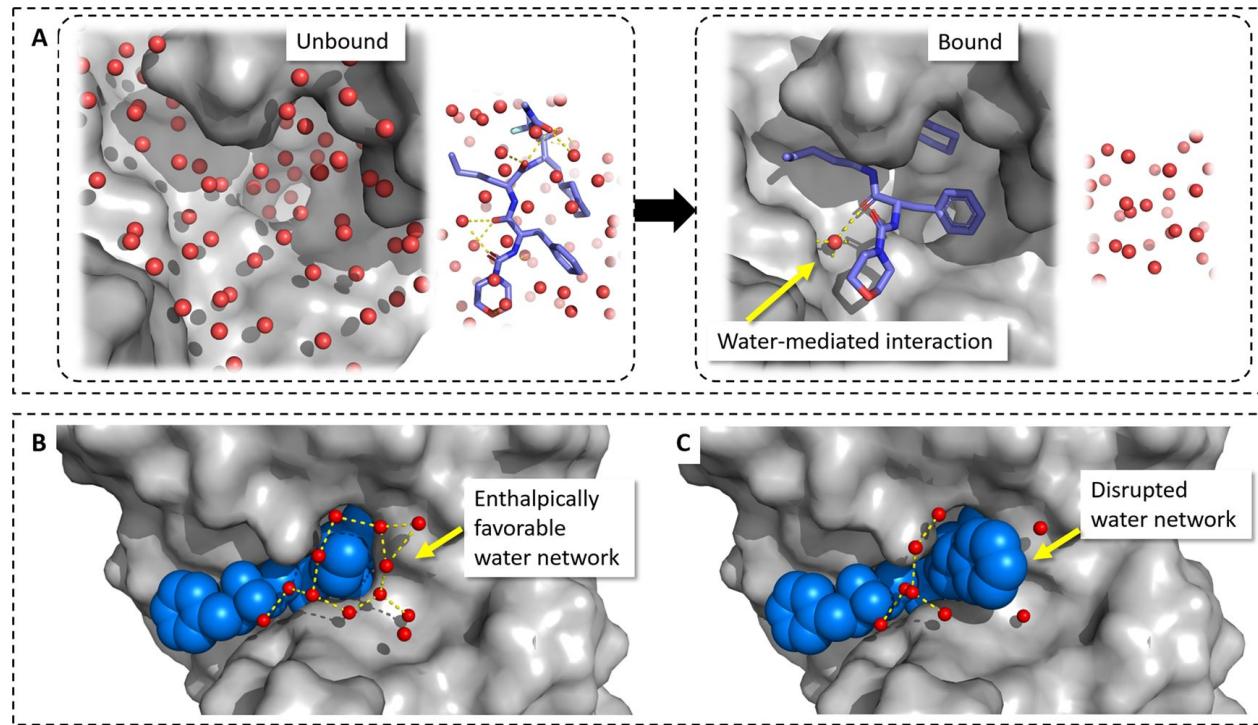
John Jumper<sup>1,4</sup>, Richard Evans<sup>1,4</sup>, Alexander Pritzel<sup>1,4</sup>, Tim Green<sup>1,4</sup>, Michael Figurnov<sup>1,4</sup>, Olaf Ronneberger<sup>1,4</sup>, Kathryn Tunyasuvunakool<sup>1,4</sup>, Russ Bates<sup>1,4</sup>, Augustin Žídek<sup>1,4</sup>, Anna Potapenko<sup>1,4</sup>, Alex Bridgland<sup>1,4</sup>, Clemens Meyer<sup>1,4</sup>, Simon A. A. Kohl<sup>1,4</sup>, Andrew J. Ballard<sup>1,4</sup>, Andrew Cowie<sup>1,4</sup>, Bernardino Romera-Paredes<sup>1,4</sup>, Stanislav Nikolov<sup>1,4</sup>, Rishabh Jain<sup>1,4</sup>, Jonas Adler<sup>1</sup>, Trevor Back<sup>1</sup>, Stig Petersen<sup>1</sup>, David Reiman<sup>1</sup>, Ellen Clancy<sup>1</sup>, Michał Zieliński<sup>1</sup>, Martin Steinegger<sup>2,3</sup>, Michałina Pacholska<sup>1</sup>, Tamas Berghammer<sup>1</sup>, Sebastian Bodenstein<sup>1</sup>, David Silver<sup>1</sup>, Oriol Vinyals<sup>1</sup>, Andrew W. Senior<sup>1</sup>, Koray Kavukcuoglu<sup>1</sup>, Pushmeet Kohli<sup>1</sup> & Demis Hassabis<sup>1,4</sup>

## Article

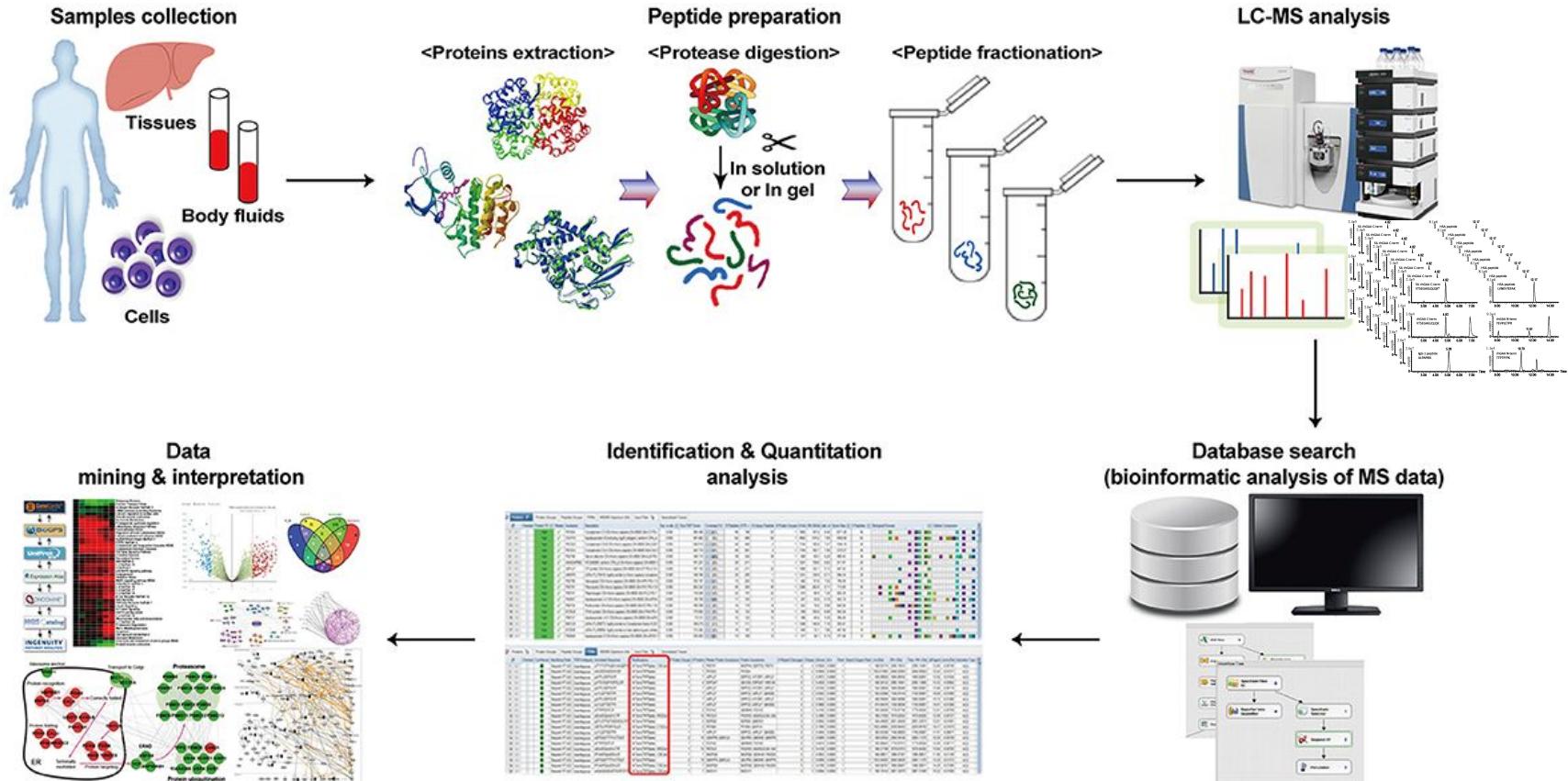


# Protein-Ligand Interaction

- Drug discovery



# Proteomics



# History of Bioinformatics

## Protein Wave

1955: Protein sequencing

1970: Needleman-Wunsch Algorithm  
(Global Sequence Alignment)

1973: Protein Data Bank (PDB)

1990: BLAST

1994-: Critical Assessment of  
Structure Prediction (CASP)

2020: AlphaFold2

Learning the rule from real DATA

## Nucleotide Wave

1953: DNA structure

1972: Recombinant DNA

1977: Sanger sequencing

1985: PCR

1988: NCBI

1990: BLAST

1995-2003: Human  
Genome sequencing

2005-: NGS

## Expression Wave

1995: Microarray

2006: RNA-seq  
(RNA sequencing)

2008: scRNA-seq  
(Single Cell RNA-seq)

# History of Bioinformatics: Nucleotide Wave

- 1953: DNA structure
- 1972: Recombinant DNA
- **1977: Sanger sequencing**
- 1985: PCR
- 1988: NCBI
- 1990: BLAST
- **1995-2003: Human Genome sequencing**
  - Public Project
  - Private Project
- **2005-: Next Generation Sequencing (Pyrosequencing)**

ncbi.nlm.nih.gov

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

## 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

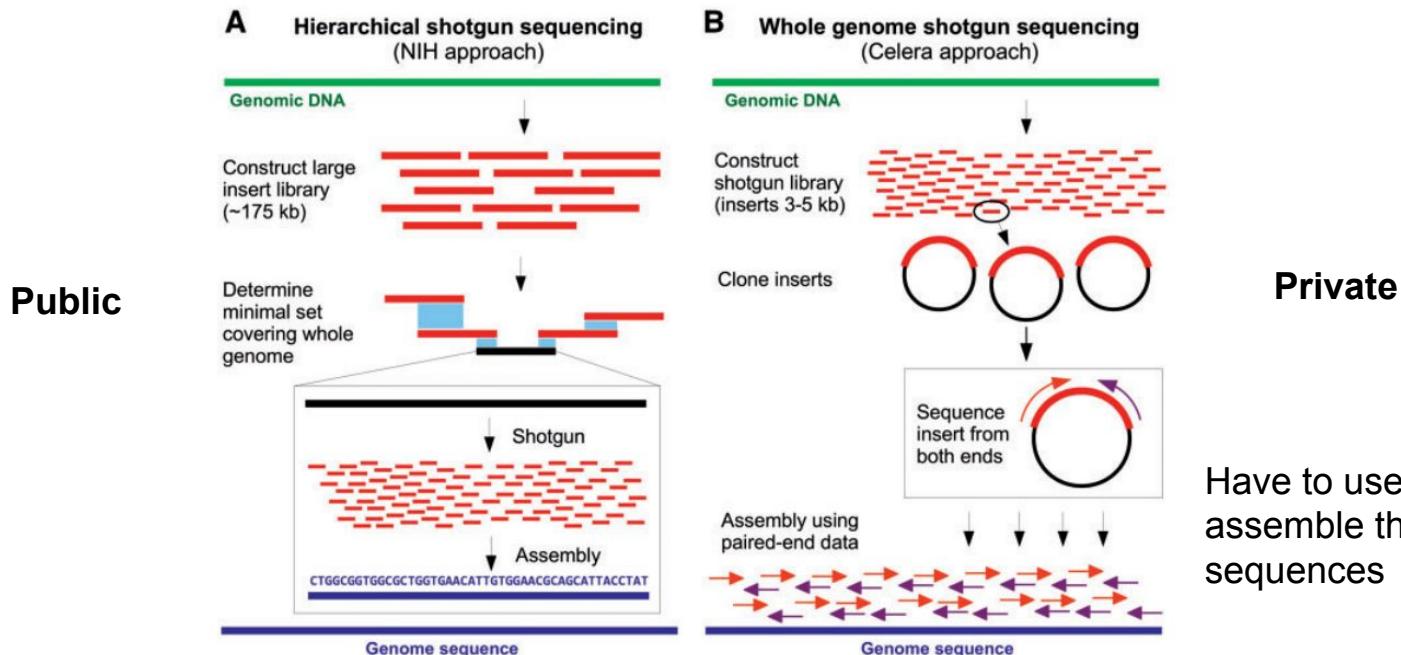
Release 11.0 of the NCBI protein profile Hidden Markov models (HMMs) [read more](#)

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 21.38 trillion bases and 3.25 [read more](#)

New RefSeq Annotations! 20 Dec 2022

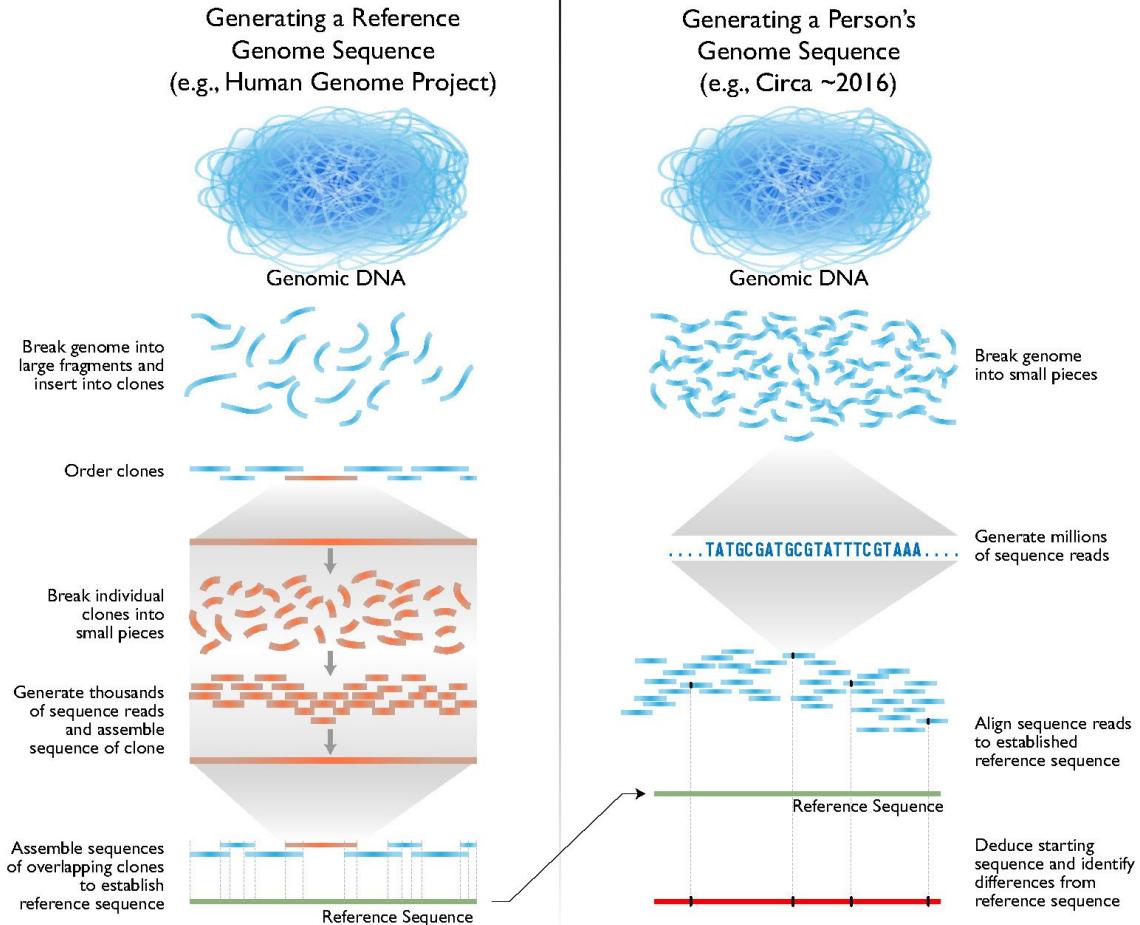
# Racing in Human Genome Sequencing



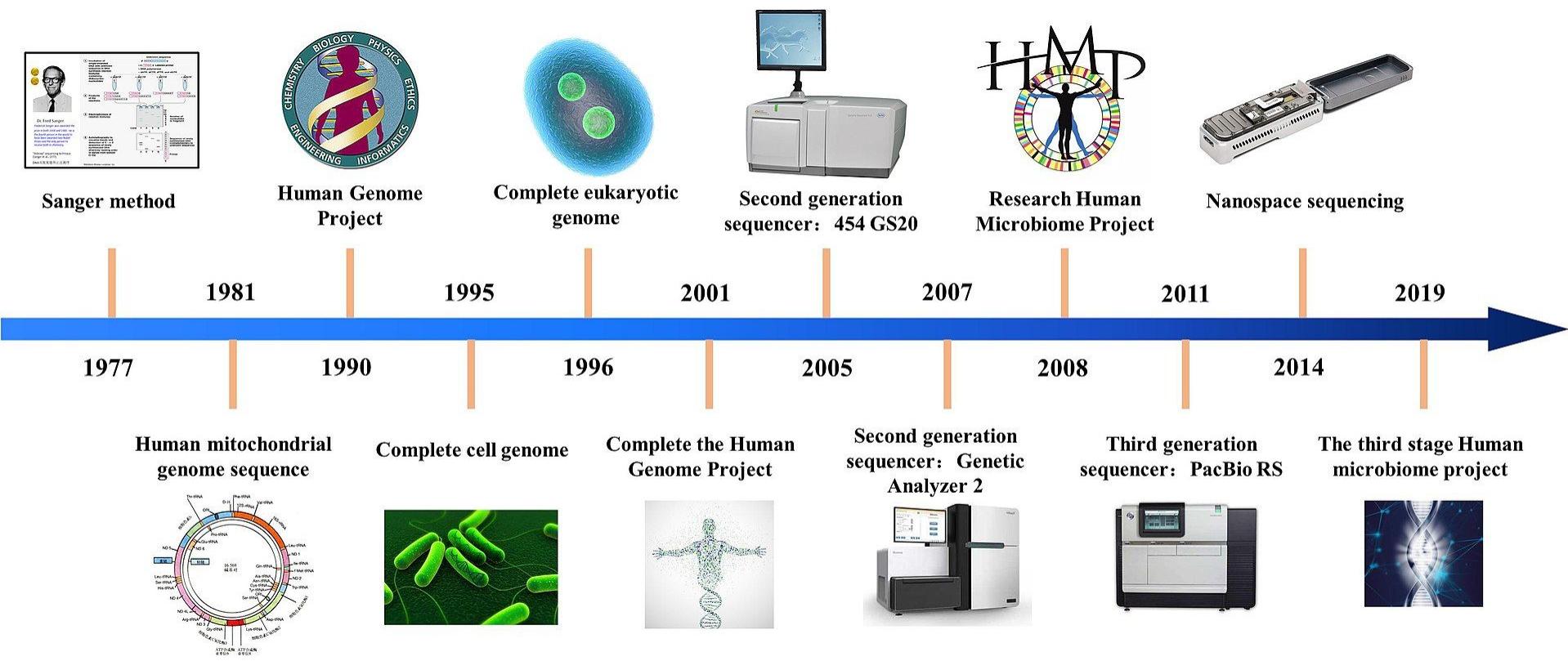
**Figure 6.** Hierarchical shotgun sequencing versus whole genome shotgun sequencing. Both approaches respectively exemplified the methodological rivalry between the public (NIH, A) and private (Celera, B) efforts to sequence the human genome. Whereas the NIH team believed that whole-genome shotgun sequencing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

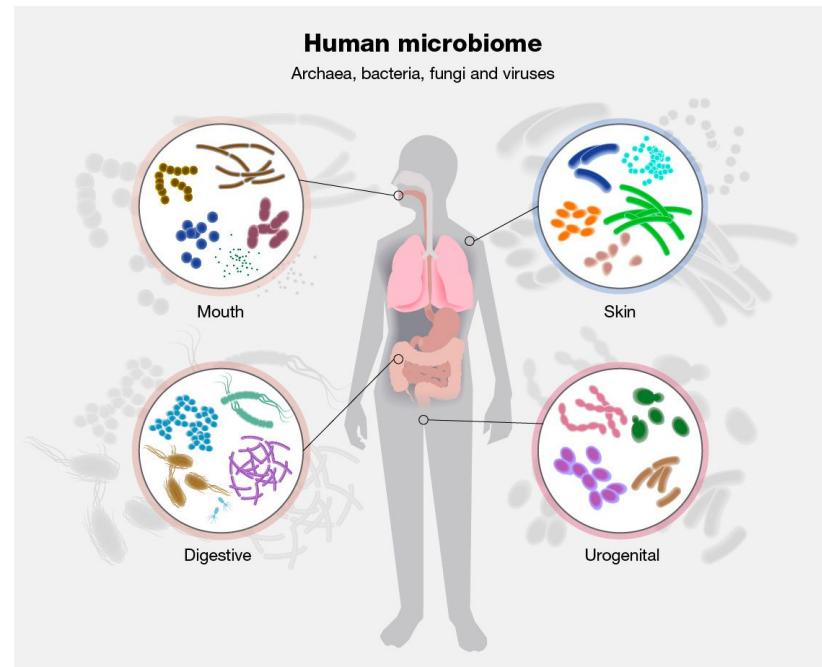
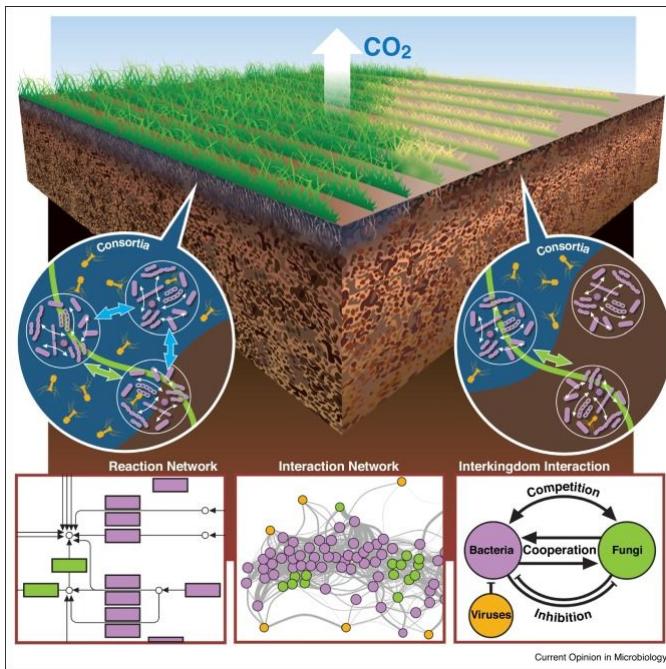


# Next Generation Sequencing (NGS)



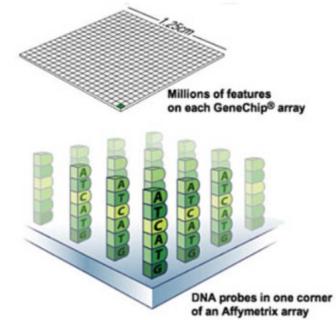
# Microbiome: Based on DNA sequencing

- **Metagenomics:** The study of the structure and function of entire nucleotide sequences isolated and analyzed from all the organisms (typically microbes) in a bulk sample.
- Study the microbial population in environment or microenvironment (inside other organisms)



# History of Bioinformatics: Expression Wave

- 1995 Microarray
- 2006 RNA-seq
- 2008 scRNA-seq



<http://www.affymetrix.com/>

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

	Condition 1	Condition 2	Condition N	
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

# Huge Data per Sample/Run

- Microarray and NGS sequencing data

10 M sequences

Tabular Microarray Viewer											Sequence Retriever	Information Panel
Marker	Scatter Plot										Sequence Retriever	Information Panel
	CB 2-23	CB 3-10	CB 3-30	CB 3-7	CB 6-8	CC 3-28	CC 4-14	CC 4-6	CC 7-25	CC 7-7		
37429_g_a...	7.36	7.26	7.37	7.30	7.17	7.07	7.43	7.16	7.80	7.36		
37430_at (...)	2.98	4.00	3.51	3.36	3.51	3.17	3.47	3.26	3.17	2.91		
37431_at (...)	7.36	7.18	7.14	7.61	7.51	7.76	6.88	7.25	7.33	7.15		
37432_g_a...	7.61	7.64	7.69	7.34	7.84	7.07	7.09	7.30	8.09	7.32		
37433_at (...)	5.94	5.44	5.77	5.51	6.09	5.93	5.61	6.06	5.87	5.71		
37434_at (...)	7.08	7.59	7.32	8.19	8.07	7.72	7.72	7.60	7.48	7.37		
37435_s_a...	5.67	6.15	6.21	5.60	5.75	5.48	5.17	5.20	5.48	5.72		
37436_at (...)	5.98	4.52	5.40	4.33	4.38	5.55	5.19	5.46	5.33	4.19		
37437_at (...)	4.78	4.48	6.52	1.26	5.27	5.52	4.97	4.94	5.62	4.60		
37438_at (...)	6.09	6.42	6.70	6.99	6.85	6.82	6.63	6.93	6.67	6.05		
37439_at (...)	2.63	2.56	4.14	2.87	4.33	2.94	2.96	1.14	2.56	2.26		
37440_at (...)	5.62	6.11	6.40	6.02	6.09	6.13	5.81	4.85	5.33	6.15		
37441_at (...)	4.25	4.78	4.75	5.13	4.53	4.87	3.79	5.48	4.86	4.15		
37442_at (...)	9.42	8.84	8.84	9.60	9.23	9.55	9.49	9.51	9.17	9.57		
37443_at (...)	7.14	6.88	6.61	6.76	6.83	7.15	6.86	6.66	6.11	6.90		
37444_at (...)	7.39	7.77	7.98	7.57	7.49	7.52	7.58	7.83	8.07	7.58		
37445_at (...)	7.32	7.58	7.71	8.08	7.82	8.03	7.71	7.57	7.89	7.73		
37446_at (...)	3.62	5.38	5.88	3.85	5.00	5.03	5.01	5.15	5.54	5.44		
37447_at (...)	5.24	6.87	7.63	6.35	6.25	6.76	6.15	5.49	6.21	6.40		

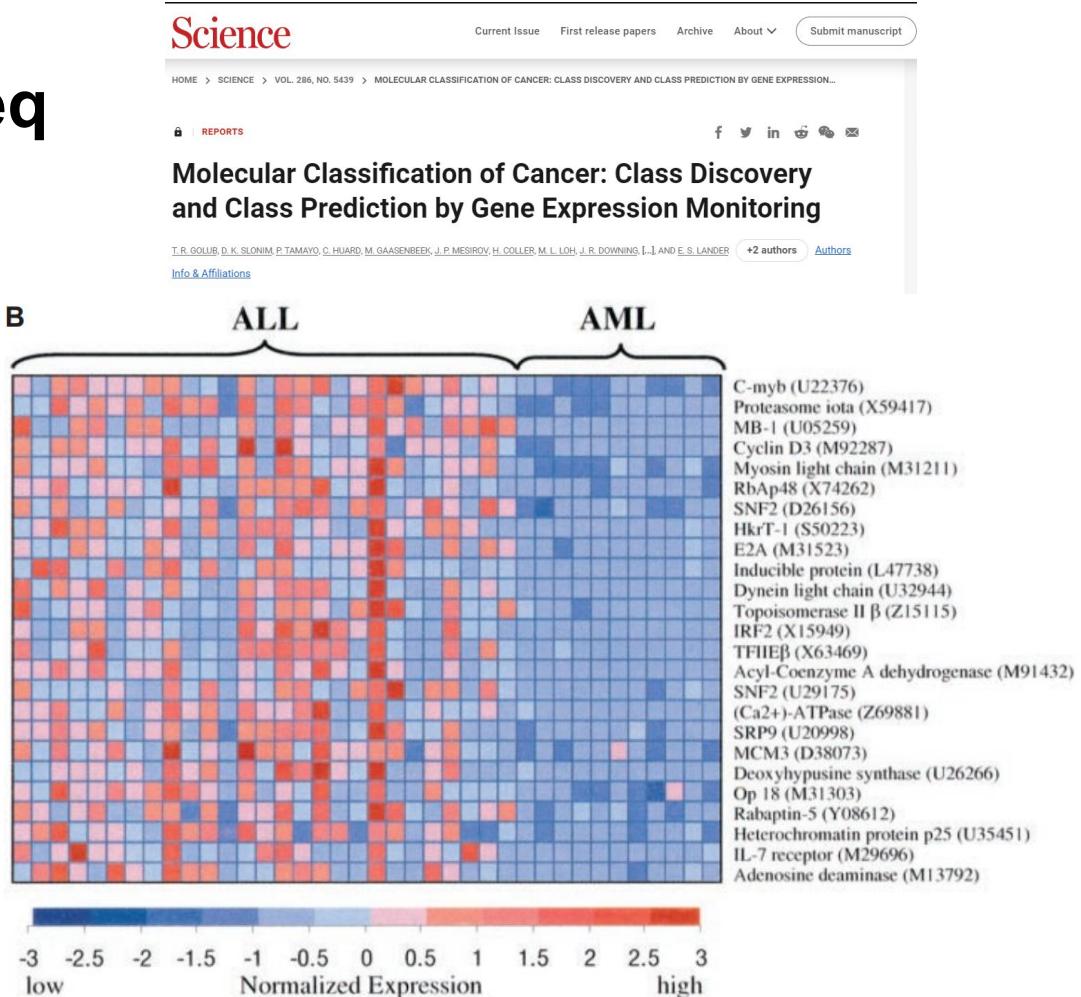
Format:  Number  Scientific Decimal places: 2 Export

~20k probe

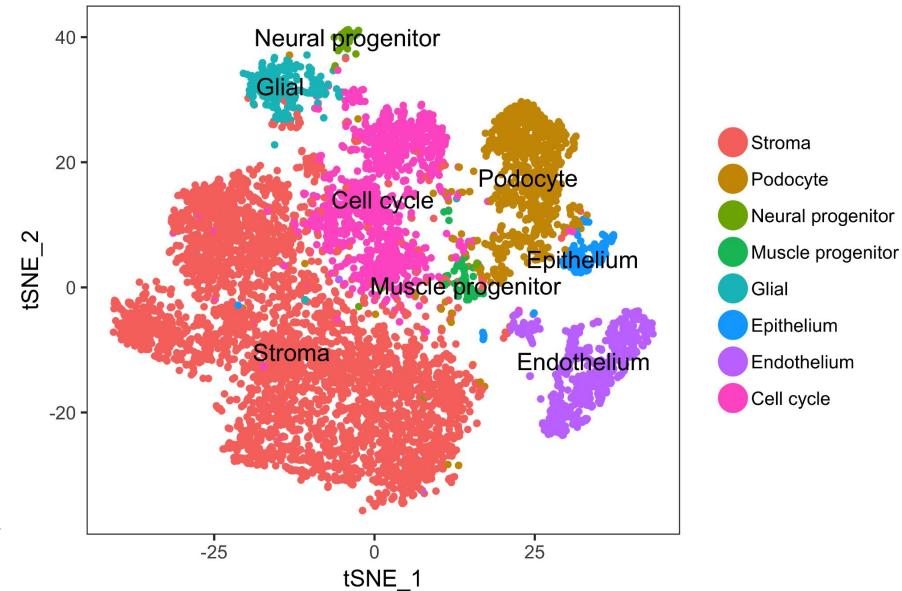
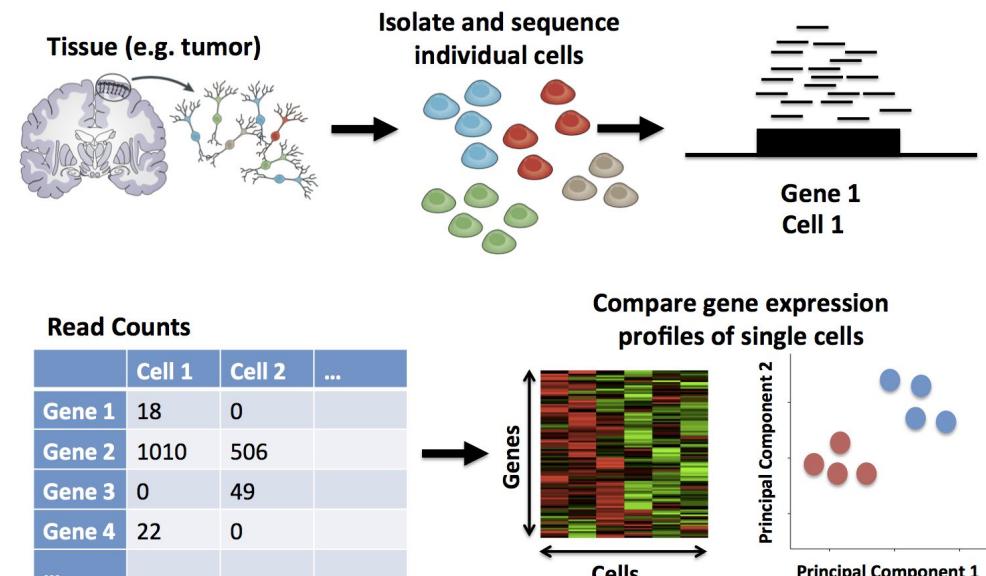
AAAAAA6AEEEEAEEE/EEEEEEAEEE/EEEEAEEE<EEEEAEEE  
@SRR13253046.14 14/1  
ACCAGACTATTTGTATACTAACGTAACAAATTAGGTTGTTGAACAACTTC  
+  
/AAAAAEE  
@SRR13253046.15 15/1  
AAGACTATTTGTATACTAACGTAACAAATTAGGTTGTTGAACAACTTC  
+  
AAAAAEE  
@SRR13253046.16 16/1  
AATTGTATACTAACGTAACAAATTAGGTTGTTGAACAACTTCGATCAT  
+  
AAAAAEEEEEEEEEEEEEEEEEEEEEEEE6EEEEEE6EEEE  
@SRR13253046.17 17/1  
GCGAAAGCTTATGAGAAAATACATGAATGACAGGTAAAATATTGGCTCG  
+  
AAAAAEEEEEEEEEEEEEEEEAAAAAEEEEE/EEEEEEEEEEEE  
@SRR13253046.18 18/1  
GCGAAAGCTTATGAGAAAATACATGAATGACAGGTAAAATATTGGCTCG  
+  
AAAAGEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.19 19/1  
GTAAAAATTTGGCTCGAAAAGAGGACAAAAGAGAAATCATAATGAGT  
+  
6AAAAAEEEEEEEEEEEEAAEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.20 20/1  
GCGAAAAGAGGACAAAAGAGAAATCATAATGAGTAAACCACTTGCTG  
+  
6/AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EEEEEEEE  
@SRR13253046.21 21/1  
CCAGGTACGAACTGCTAAGAAGACTCATAAAACTTCAGGTTAA  
+  
6/AAAEEAEEEEEEEEAEEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.22 22/1  
TGTACGAACTGCTAAGAAGACTCATAAAACTTCAGGTTAA  
+  
/AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.23 23/1  
ACTAGTGTCCGGAAAGCTACTTTATGATGTTCAAGGCTGAAGTTGAA  
+  
/AAAAAEEEEEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.24 24/1  
GCTGAAGTTGAATTTATGATGTTCAACATCAAATGTGTCTATTTG  
+  
/AAAAAEEEEEEEEEEEEEEEEEEEEEEEE  
@SRR13253046.25 25/1  
ACTGAAGTTGAATTTATGATGTTCAACATCAAATGTGTCTATTTG  
+  
/AAAAAEEEEEEEEEE<EEEEEEEEEEAEEEEEEE  
@SRR13253046.26 26/1

# Microarray and RNA-seq

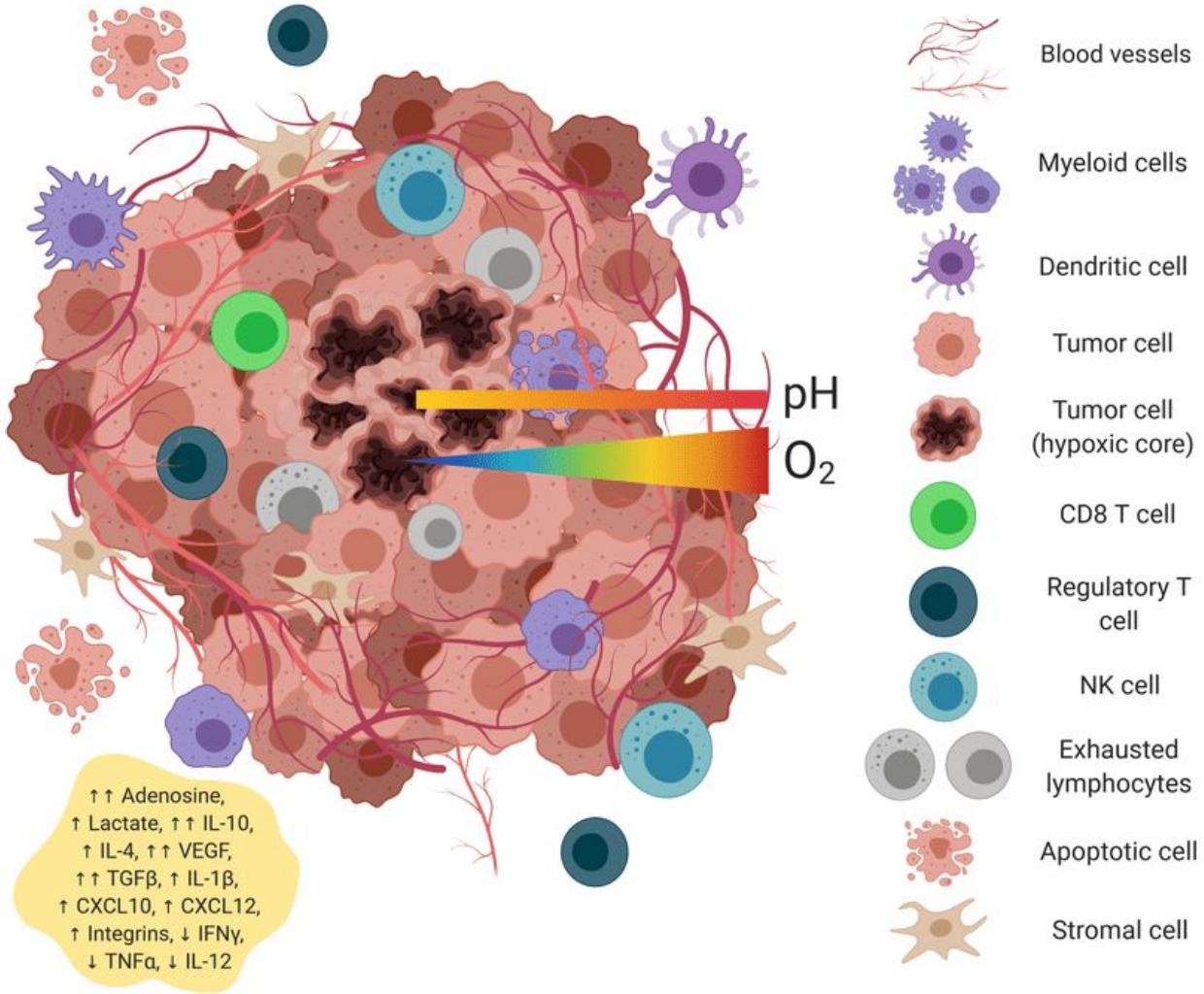
- Acute Lymphoblastic Leukemia (ALL)
- Acute Myeloid Leukemia (AML)
- Find the difference between closely related leukemia
- Precision Medicine



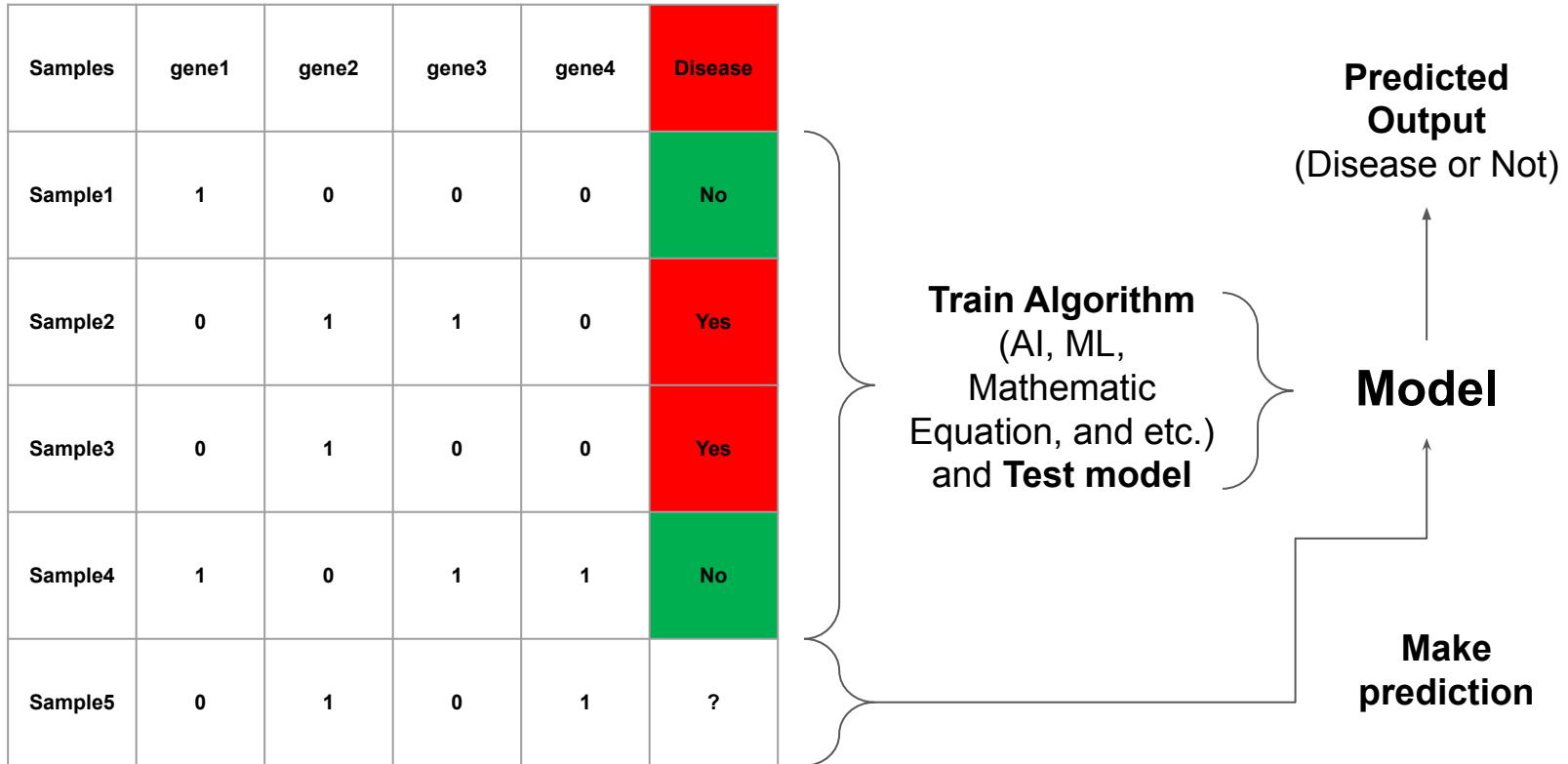
# Single Cell RNA Sequencing



# Cancer Microenvironment



# Machine Learning in Biology (Concepts)



# History of Bioinformatics

## Protein Wave

1955: Protein sequencing

1970: Needleman-Wunsch Algorithm  
(Global Sequence Alignment)

1973: Protein Data Bank (PDB)

1990: BLAST

1994-: Critical Assessment of  
Structure Prediction (CASP)

2020: AlphaFold2

Learning the rule from real DATA

## Nucleotide Wave

1953: DNA structure

1972: Recombinant DNA

1977: Sanger sequencing

1985: PCR

1988: NCBI

1990: BLAST

1995-2003: Human  
Genome sequencing

2005-: NGS

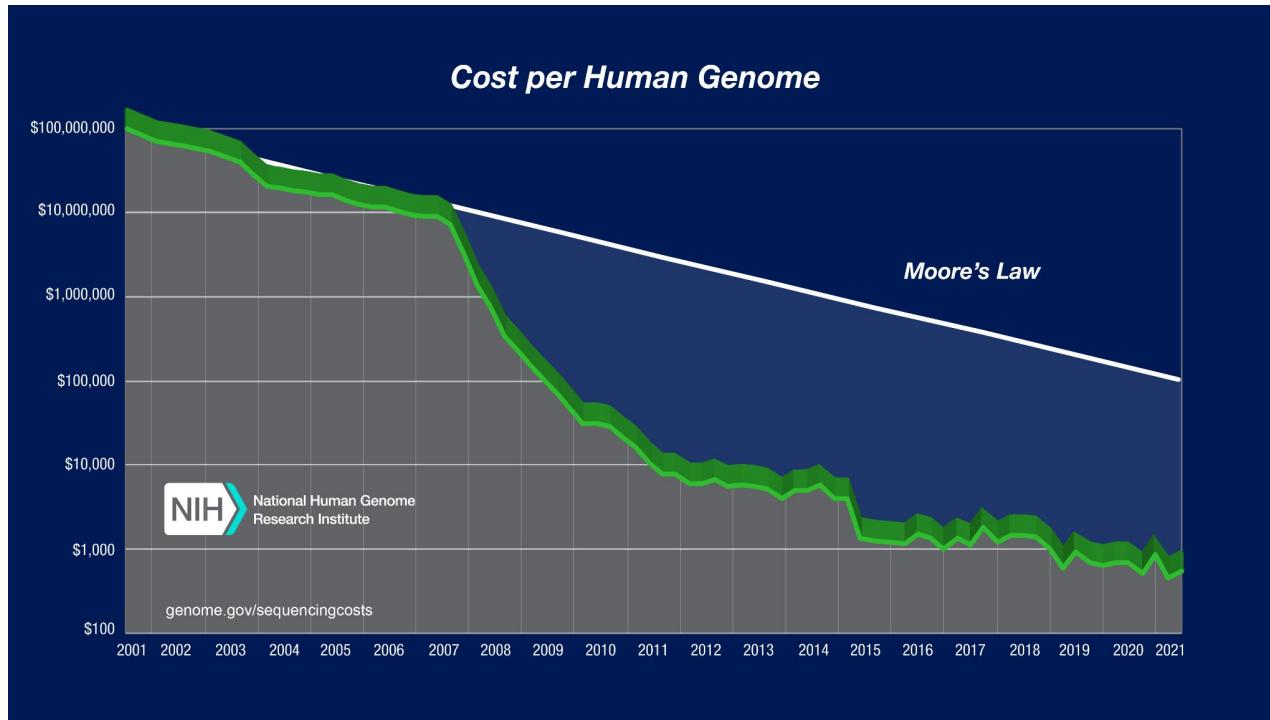
## Expression Wave

1995: Microarray

2006: RNA-seq  
(RNA sequencing)

2008: scRNA-seq  
(Single Cell RNA-seq)

# Cost of Genome Sequencing



# Illumina Sequencer: 2023

- Flagship of Sequencing Company



iSeq 100



MiniSeq



MiSeq Series



NextSeq 550 Series



NextSeq 1000 &amp; 2000



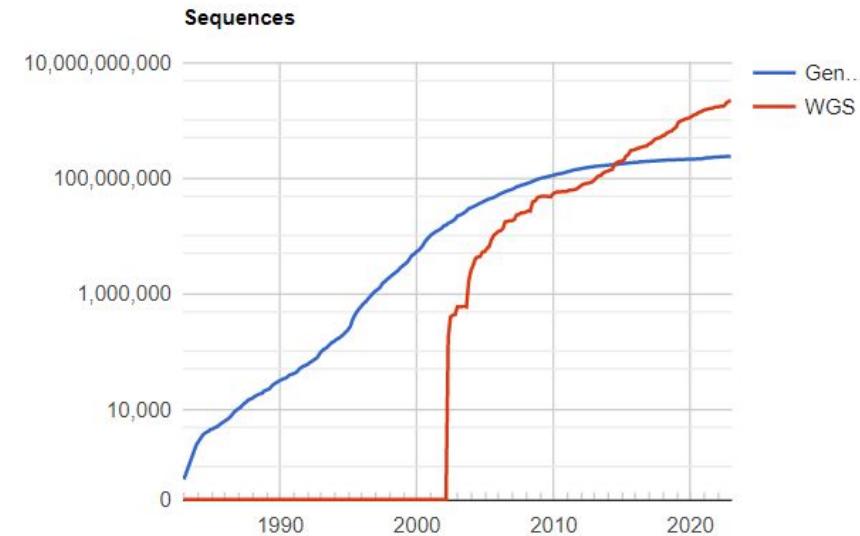
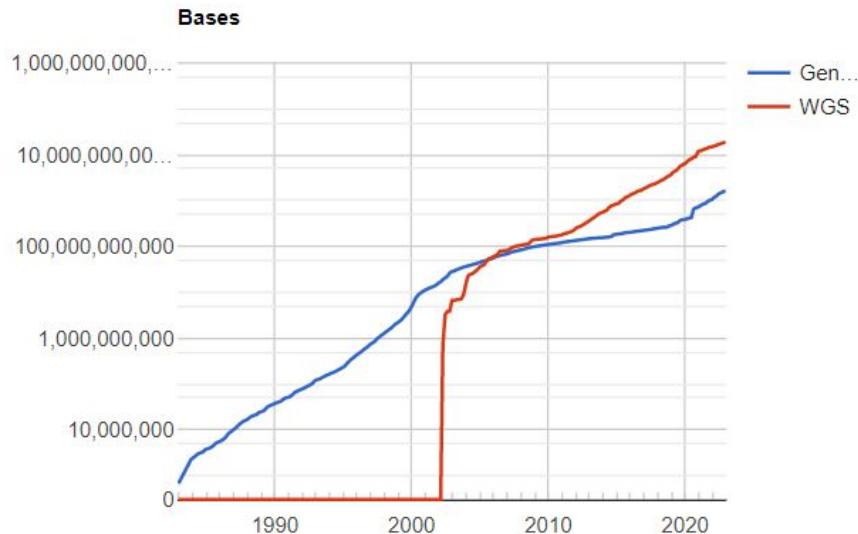
NovaSeq 6000 Series



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 <sup>‡</sup> ) ~48 hours (25B flow cells <sup>‡</sup> )
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 <sup>†</sup>	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



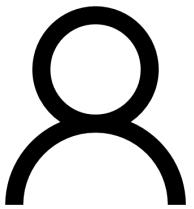
Date	Bases	Sequences	Bases	Sequences
Dec 1982	680338	606	-	-
Apr 2002	19072679701	16769983	692266338	172768
Dec 2022	1635594138493	241015745	19086596616569	2241439349

# **The National Center for Biotechnology Information (NCBI)**

**SC113603 Molecular Biology**  
Lect. Todsapol Techo

# Introduction of NCBI

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



Pub**Med**.gov

DDBJ  
DNA Data Bank of Japan

**ENA**  
European Nucleotide Archive

WORLDWIDE  
**PDB**  
PROTEIN DATA BANK

UniProt

# Web-based Tools

- BLAST
- COBALT
- etc.

## 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
- **Two Main Search Systems on NCBI**
  - Entrez search systems (Like Google)
  - Basic Local Alignment Search Tool (BLAST)

National Library of Medicine  
National Center for Biotechnology Information

All Databases ▾

Welcome to NCBI

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

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.

Results found in 30 databases

GENE

G6PD – glucose-6-phosphate dehydrogenase

*Homo sapiens* (human)

Also known as: G6PD1

Gene ID: 2539

RefSeq transcripts (3) RefSeq proteins (3) RefSeqGene (1) PubMed (477)

Orthologs Genome Data Viewer BLAST Download

RefSeq Sequences +

Literature

Bookshelf	450
MeSH	17
NLM Catalog	9
PubMed	6,510
PubMed Central	14,676

Genes

Gene	1,905
GEO DataSets	25
GEO Profiles	6,579
HomoloGene	2
PopSet	107

Proteins

Conserved Domains	1
Identical Protein Groups	255
Protein	52,112
Protein Family Models	12
Structure	43

Cancer

Clinical

PubChem

## Search Box

Results found in 30 databases

### Number of database matched search text

GENE

**G6PD – glucose-6-phosphate dehydrogenase***Homo sapiens* (human)

Also known as: G6PD1

Gene ID: 2539

RefSeq transcripts (3) RefSeq proteins (3) RefSeqGene (1) PubMed (477)

[Orthologs](#)[Genome Data Viewer](#)[BLAST](#)[Download](#)

Was this helpful?



RefSeq Sequences

**Get transcript, protein, or gene sequences**

RefSeq transcripts

G6PD – 3 of 3 transcripts

Transcript	Isoform	Len (nt)
NM_000402.4	a	2,406
NM_001042351.3	b	2,267
NM_001360016.2	b	2,223

[View full table](#)

NCBI Datasets

### The information of transcripts

Literature

Bookshelf  
MeSH  
NLM Catalog  
PubMed  
PubMed Central

Genes

Gene  
GEO DataSets  
GEO Profiles  
HomoloGene  
PopSet

Proteins

Conserved Domains	100
Identical Protein Groups	4,646
Protein	117,634
Protein Family Models	319
Structure	175

Genomes

Assembly  
BioCollections  
BioProject  
BioSample  
Genome  
Nucleotide  
SRA  
Taxonomy

Clinical

ClinicalTrials.gov  
ClinVar  
dbGaP  
dbSNP  
dbVar  
GTR  
MedGen  
OMIM

PubChem

BioAssays	915
Compounds	2
Pathways	3
Substances	134

**Results from other database**

GENE

Was this helpful?



## G6PD – glucose-6-phosphate dehydrogenase

*Homo sapiens* (human)

Also known as: G6PD1

Gene ID: 2539

[RefSeq transcripts](#) (3) [RefSeq proteins](#) (3) [RefSeqGene](#) (1) [PubMed](#) (477)

[Orthologs](#)

[Genome Data Viewer](#)

[BLAST](#)

[Download](#)

### RefSeq Sequences

Showing 3 of 3 (by status, accession number)

Transcript	nt	Protein	aa	Isoform	Status
<a href="#">NM_001360016.2</a>	2,223	<a href="#">NP_001346945.1</a>	515	b	<a href="#">MANE SELECT</a>
<a href="#">NM_000402.4</a>	2,406	<a href="#">NP_000393.4</a>	545	a	curated
<a href="#">NM_001042351.3</a>	2,267	<a href="#">NP_001035810.1</a>	515	b	curated

[View full table](#)

### More options

Search the Nucleotide and Protein databases for more sequences

[RefSeq transcripts](#) (3)

[Archival GenBank transcript sequences](#) (13)

[RefSeq proteins](#) (3)

[Archival GenBank protein sequences](#) (148)

### RefSeq transcripts

[G6PD](#) – 3 of 3 transcripts

Transcript	Isoform	Len (nt)
<a href="#">NM_000402.4</a>	a	2,406
<a href="#">NM_001042351.3</a>	b	2,267
<a href="#">NM_001360016.2</a>	b	2,223

[View full table](#)

NCBI Datasets

- **RefSeq Database**
  - Reference sequence for each gene
  
- **Example sequences**
  - NM\_001360016.2: Transcript
  - NP\_001346945.1: Protein
  - NG\_009015.2: Gene
  - NC\_000023.11: Chromosome (Assembly)

# Genome Data Viewer

Home Share this page Reset All More Info...

## Homo sapiens (human)

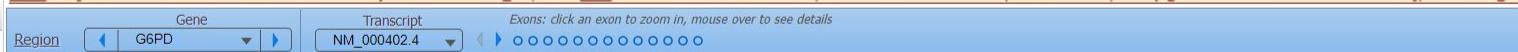
Search assembly

Location, gene or phenotype

Examples ►

Assembly: GRCh38.p14 (GCF\_000001405.40) • Chr X (NC\_000023.11)

NC\_000023.11: 154,529,769 - 154,549,205



## Assemblies

Assembly

Switch organism

GCF\_000001405.40 (GRCh38.p14)

Select an assembly to change view

## Locations for Gene G6PD

Sequence Location

NC\_000023.11 154,531,390 - 154,547,569

## Ideogram View

Unplaced/unlocalized scaffolds: 166 Alt loci/patches: 514

1 2 3 4 5 6 7 8 9 10 11 12 13

14 15 16 17 18 19 20 21 22 X Y MT

GenBank ▾

Send to: ▾

Change region shown

## Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), RefSeqGene on chromosome X

NCBI Reference Sequence: NG\_009015.2

[FASTA](#) [Graphics](#)Go to: 

LOCUS NG\_009015 23182 bp DNA linear PRI 15-DEC-2022  
 DEFINITION Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), RefSeqGene  
 on chromosome X.

### RefSeqGene Accession

VERSION NG\_009015.2  
 KEYWORDS RefSeq; RefSeqGene.  
 SOURCE Homo sapiens (human)  
 ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Homo.

COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff in  
 collaboration with Sujatha Mohan. The reference sequence was  
 derived from [AC244090.3](#).  
 This sequence is a reference standard in the [RefSeqGene](#) project.

On Jan 17, 2014 this sequence version replaced [NG\\_009015.1](#).

**Summary:** This gene encodes glucose-6-phosphate dehydrogenase. This protein is a cytosolic enzyme encoded by a housekeeping X-linked gene whose main function is to produce NADPH, a key electron donor in the defense against oxidizing agents and in reductive biosynthetic reactions. G6PD is remarkable for its genetic diversity. Many variants of G6PD, mostly produced from missense mutations, have been described with wide ranging levels of enzyme activity and associated clinical symptoms. G6PD deficiency may cause neonatal jaundice, acute hemolysis, or severe chronic

Customize view

Analyze this sequence

Run BLAST

Pick Primers

[Highlight Sequence Features](#)

Find in this Sequence

### Articles about the G6PD gene

Extracellular vesicles-transferred SBSN drives glioma aggressiveness by activ [Oncogene. 2022]

Glucose-6-phosphate Dehydrogenase (G6PD) A-Variant Frequency an [Am J Trop Med Hyg. 2022]

Suppression of ACE2 SUMOylation protects against SARS-CoV-2 infectio [Nat Commun. 2022]

See all...

### Reference sequence information

RefSeq alternative splicing

See 3 reference mRNA sequence splice variants for the G6PD gene.

### More about the G6PD gene

This gene encodes glucose 6 phosphate

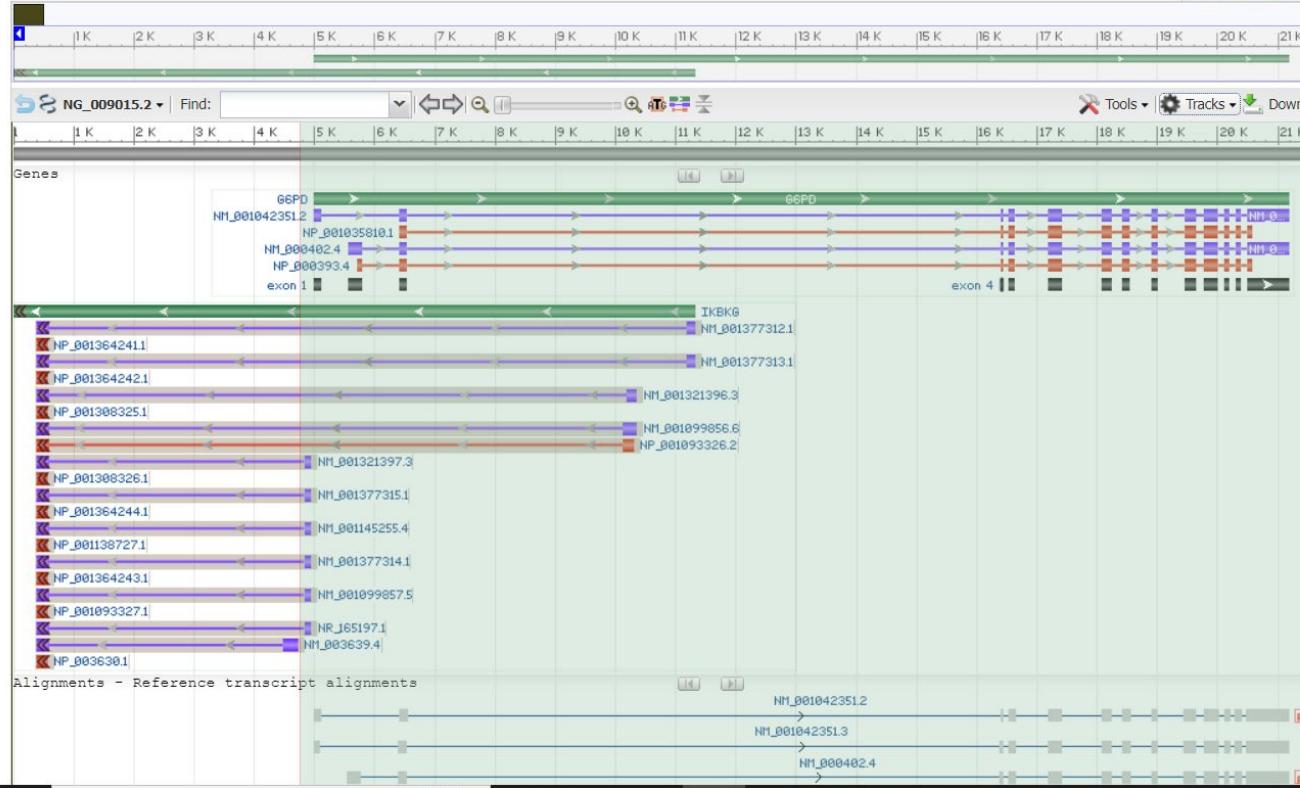
Graphics ▾

Send to: ▾

Analyze this sequence

# Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), RefSeqGene on chromosome X

NCBI Reference Sequence: NG\_009015.2

[GenBank](#) [FASTA](#)[Link To This View](#)

- [Range: 4752..21642](#)
- [16,891 nt](#)
- [Copy Sequence \(Selection\)](#)
- [Zoom On Range](#)
- [Zoom To Sequence](#)
- [Set New Marker On Range](#)
- [Modify Range](#)
- [Add New Panel On Range](#)
- [BLAST Search \(Selection\)](#)
- [Primer BLAST \(Selection\)](#)
- [Download FASTA \(Selection\)](#)
- [Download GenBank Flat File \(Selection\)](#)

## Reference sequence information

RefSeq alternative splicing  
See 3 reference mRNA sequence s gene.

## More about the G6PD gene

This gene encodes glucose-6-phosphate dehydrogenase, a cytosolic enzyme encoded by a linked gene whose main f...  
Also Known As: G6PD1

## Homologs of the G6PD gene

The G6PD gene is conserved in Rhesus monkey, mouse, rat, zebrafish, fruit fly, mosquito, *S.cerevisiae*, *K.lactis*, *E.gossypii*, *S.Neapolitana*, *A.thaliana*, rice, and frog.

## Related information

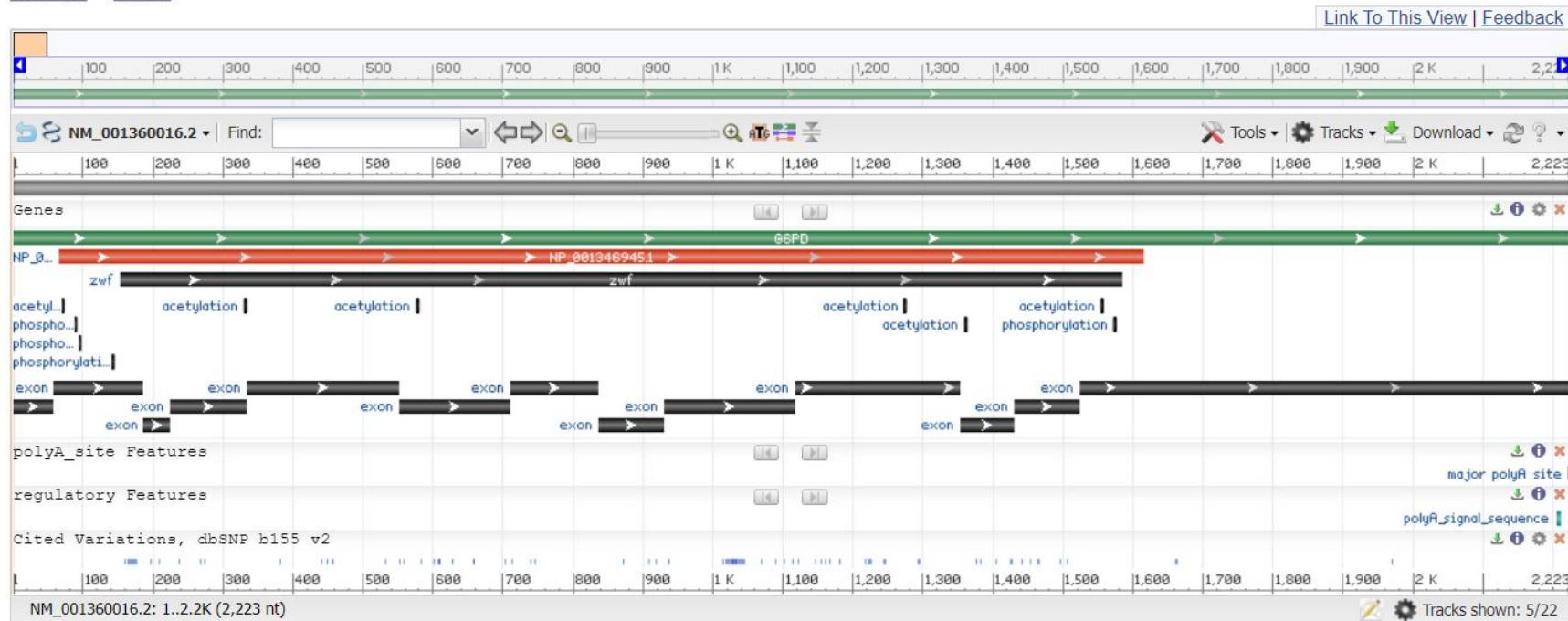
Graphics ▾

Send to: ▾

## Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), transcript variant 3, mRNA

NCBI Reference Sequence: NM\_001360016.2

[GenBank](#) [FASTA](#)



Graphics ▾

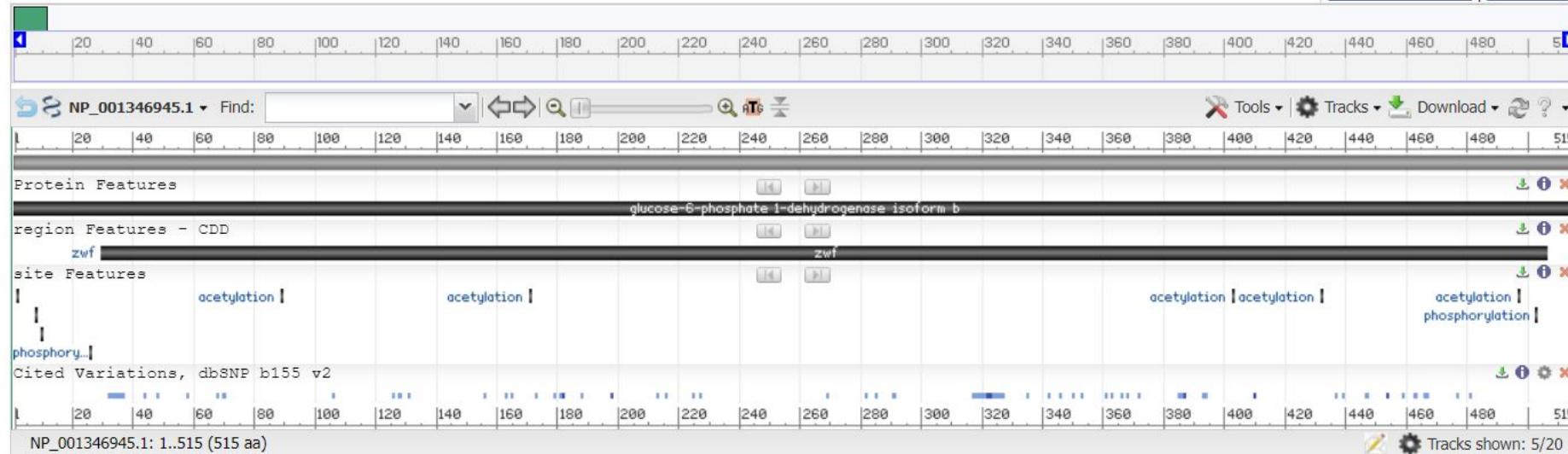
Send to: ▾

## glucose-6-phosphate 1-dehydrogenase isoform b [Homo sapiens]

NCBI Reference Sequence: NP\_001346945.1

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

[Link To This View](#) | [Feedback](#)



## Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), RefSeqGene on chromosome X

NCBI Reference Sequence: NG\_009015.2

[GenBank](#) [Graphics](#)

>NG\_009015.2 Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), RefSeqGene on chromosome X

```
TGCCCTGTCCAGGATGGACAGCACTTCATGAGTGCTCAGTGCAGGGACACAGAGGCCCTGGACATGTTGG
TCCAACCTGGAAGGTTCAATTCCAGCTCAGAGCTCCCTGTAAGATCCAGCAAGGCTGTCCTGGAGCCTAC
CCGGCAGCGTCTTCCTCCCTGCCCCCTGGCTTCTCCCCCTGTCCTCCTTCCCTCCAGTACTG
GTCCTGAGAGCATGCTAAACTTCCTCCCCGCTAATCTGCATCTCAGAGTCTACTTCTGAAACCAAGAGG
GAGGGCAGCACATCCCCACCCAGGCCAGGAAACGCCAGGGTGCAGGTGGGAGAACAGGCCCTCCCT
CAGCAGGGCTGGATCCCCTGACTCTTCCTCACCTGGAGCTCTGATTCTCCTCCAGGCAGCGCTGGAG
GGTCTCAGGAGCGCCCTGTTCTGAAGGCAGGTGCAAGCATGGCTGGCTCCCCAGAGGAGACTTCGCC
AGTACGTCCTGATCTGCTGCCGGGACCACACTGGGCTGACCACATCACAAGTTGGCTCTCCAGAGGT
GCCTATTCTACCAACAGGCAAGGGCTGAAAGAGAAAGGCCAGCAGGGAGTCACAGGAGACCCACATCT
TACCCAGCAGAAAAGCCTATGAGTTCTGGGGCTACATTTAAACAGAATGCTGATAATTACACCTAGTG
ATCAGGATGCTAAAGTGGGTTTCAGTCTAAGGGTAGCAGATACTCCATTCAAAGGGAGGTGCTGGA
CCACCCATGTGGGACTTGAAGGAGCTAGCTCCAGGCTGAAATAGAAGAATTCAACCCAGAGATGGAGA
GTCGGGACAAGACTGTTGTGGAGGGAGATGTCAGTATGCTGGAAATCTGATCCCTACACACCTCAA
GCAGGGAGAAGTGGAGGGTACCGGCCCTCATCCAGTGTAGTGGGCTGAATGGTAGCCCCCGAAAAGATA
TGCCCATGTTCAACCCCTGAAACATATGTGATTGGGAAAGGGCTTTGCAAGATGTAGGATCTCAAGA
TGAGATCATTCTGGGTGACGTGGGTGGGCCCTAAATCAAACACGTGCTTTGAGAGACAGAAGAGAG
AAGCCAGAGGGAGACGCAAGGGAGGGCATGTCGGACTGCCCTGGAGACTGGAGTGAGGCAGCCAC
AAGCCAGGAATAATGGGGCAGAAGCTGGGAGAGGCAAGGGAGGATTGCCCTTGAAGCCTTGAGG
GTGTGTGCTTCAACCTTGGTTTGACTCTAGTCTCAGAAGTGTGAGATTAAATGTCTGGCG
GGCGCAGTGGCTCACACCTGTAATCCAGCACTTGGGAGGCCAGGCGGGCAATCATGAGGTAGGAG
ATCGAGACCAGCCTGGCTAACACGGTAAACCTGCTACTAAAAAATGGCGGTCA
CGGTGGCTACGCCGCTGTAATCCAGCACTTGGGAGGCCAGCAGGGTGGATCATGAGGTAGGAGATTG
AGACCATCCTGGCTAACACAGTGAACACCTGCTACTAAAAAATAGCCAGGCATGGTAG
CGGGCGCCTGTAAGCCAGCTACCCAGGAGGCTGAGGCAAGGAGATGGCGTAACCCGGGAGGCAGGAGA
ATGGCGTGAACCCGGGATCGCGCACTGCACTCCAGCCTGGCAACAGAGCGAGACTCATCTCAAAAAA
AAAAAAATACAAAAAATTAGCAGAGTGTGGTGACACGCGCCTGAGTCCAGCTACTCGGGAGGCCGAAGC
AAGAGAATCGCTGAACCCAGGAGGCCAGTGGCAGTGAAGTGTGCCACTGCACTCCACCTT
```

# Gene Orthologs

Search NCBI

G6PD

Search

## G6PD - glucose-6-phosphate dehydrogenase

This gene encodes glucose-6-phosphate dehydrogenase. This protein is a cytosolic enzyme encoded by a housekeeping X-linked gene whose main function is to produce NADPH, a key electron donor in the defense against oxidizing agents and in reductive biosynthetic reactions. G6PD is remarkable for its genetic diversity. Many variants of G6PD, mostly produced from missense mutations, have been described with wide ranging levels of enzyme activity and associated clinical symptoms. G6PD deficiency may cause neonatal jaundice, acute hemolysis, or severe chronic non-spherocytic hemolytic anemia. Two transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]

Genes similar to G6PD

NCBI Orthologs How was this calculated?

 0 items

SEARCH THE TAXONOMY TREE

Enter taxonomic name

- Craniata
  - ▷ vertebrates

Protein alignment

one sequence per gene (3)  
 all sequences per gene (24)

Align

328 genes for: ver

Add to cart  Protein alignment  Download

3 selected

Previous  Next

Species

Gene

Architecture

aa

 Feedback

## COBALT

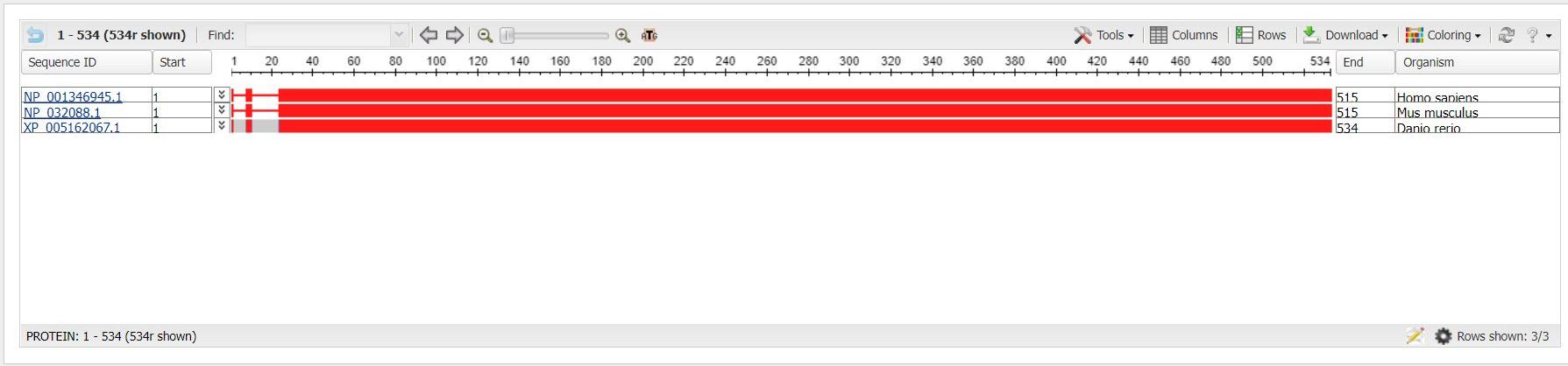
## Constraint-based Multiple Alignment Tool

Home | Recent Results | Help

Phylogenetic Tree | Edit and Resubmit | Download

## - Cobalt RID V5WTPGA6212 (3 seqs)

## ▼ Graphical Overview

▼ Descriptions  Select All | Re-align | Alignment parameters

Accession	Description	Links
<input checked="" type="checkbox"/> NP_001346945.1	glucose-6-phosphate 1-dehydrogenase isoform b [Homo sapiens]	<a href="#">Related Information</a>
<input checked="" type="checkbox"/> NP_032088.1	glucose-6-phosphate 1-dehydrogenase X [Mus musculus]	<a href="#">Related Information</a>
<input checked="" type="checkbox"/> XP_005162067.1	glucose-6-phosphate 1-dehydrogenase isoform X1 [Danio rerio]	<a href="#">Related Information</a>

## COBALT

## Constraint-based Multiple Alignment Tool

Home Recent Results Help

## Phylogenetic Tree View

This tree is based on COBALT multiple alignment more...

[Reset Tree](#)

Cobalt RID V5WTPGA6212

Number of Seqs 3

Tree method

Fast Minimum Evolution

Max Seq Difference

0.85

Distance

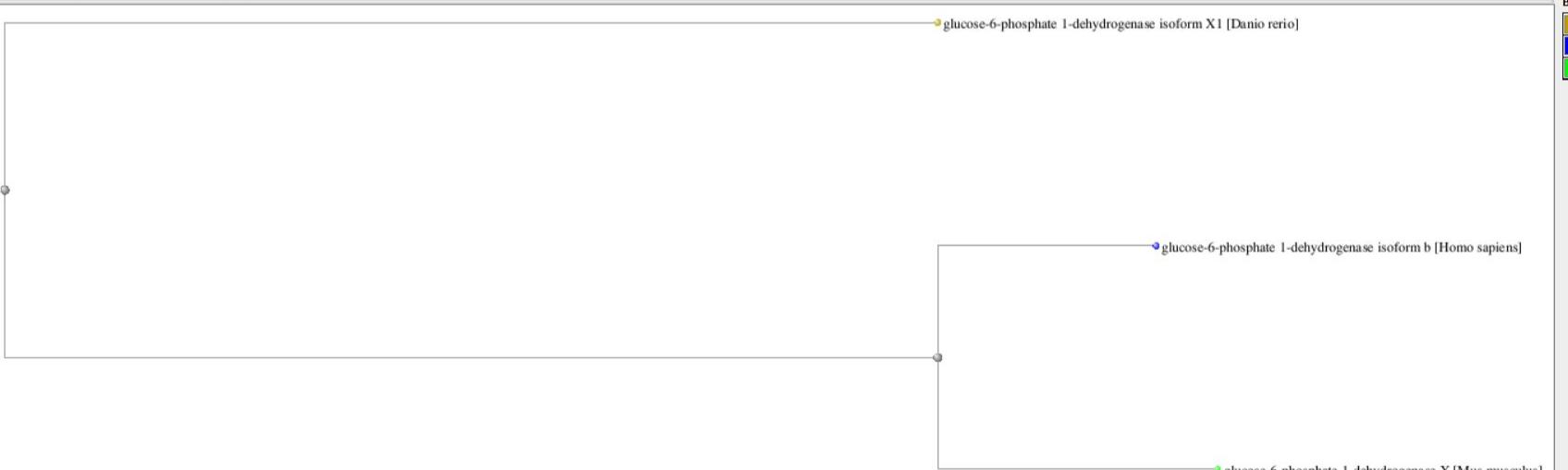
Grishin (protein)

Sequence Label

Sequence Title (if avail.)

[Hide legend](#) Find:   all              Label color map  
Blast names color map

<input checked="" type="checkbox"/>	bony fishes
<input checked="" type="checkbox"/>	primates
<input checked="" type="checkbox"/>	rodents

Success Nodes 5(0 selected)  0.01