

An Introduction to Bioinformatics Strategies

Apratim Mitra, Ph.D.

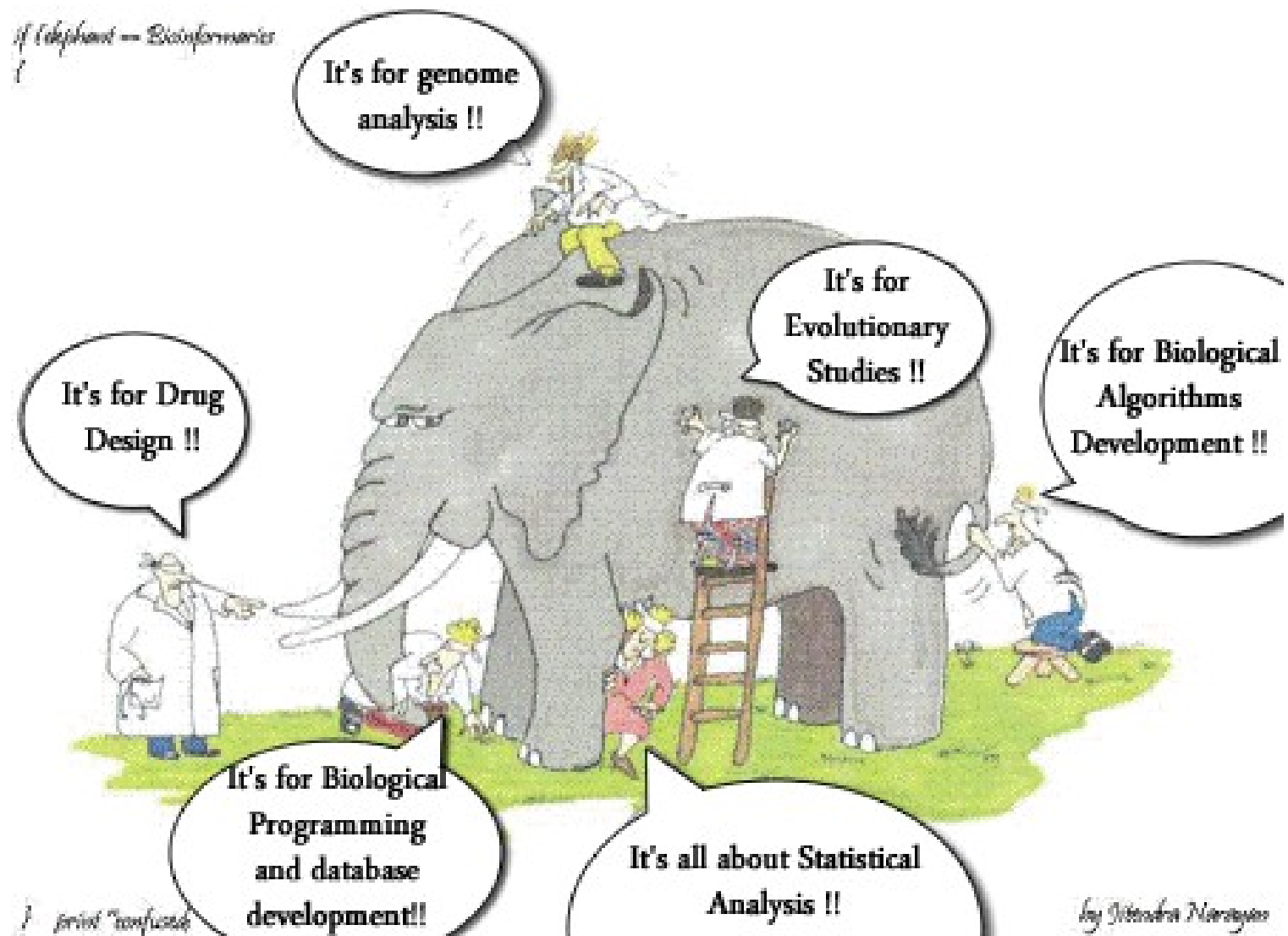
NICHD/NIH

5/17/2018

Outline

- What is bioinformatics?
 - NCBI portal – GenBank, PubMed
- Sequence analysis
 - NCBI BLAST
- Multiple sequence alignment
 - Clustal Omega
- Next generation sequencing
 - UCSC genome browser

What is bioinformatics?



What is bioinformatics?

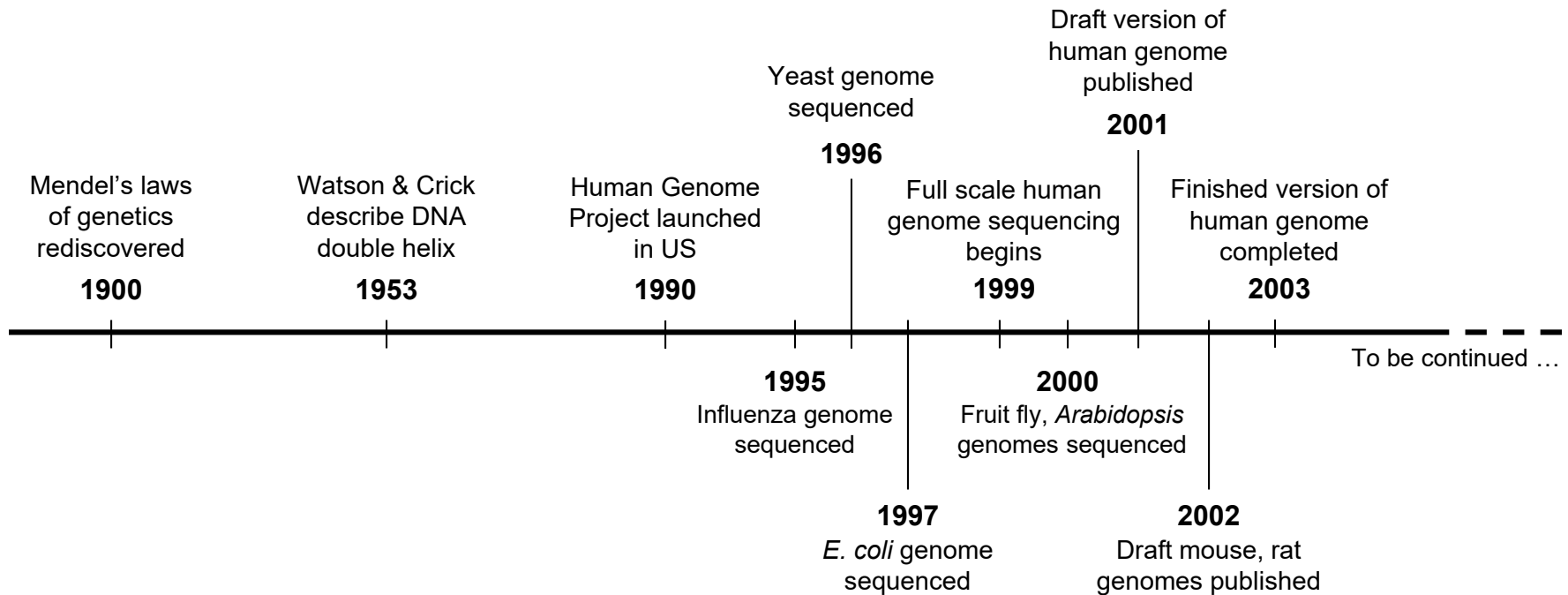
- **DEFINITION**

- Inter-disciplinary field that combines biology, computer science, mathematics, statistics and engineering

- **GOALS**

- Understand biological data
 - DNA, RNA, protein sequences and/or structures
- Solve biological problems
 - Human disease, physiological traits

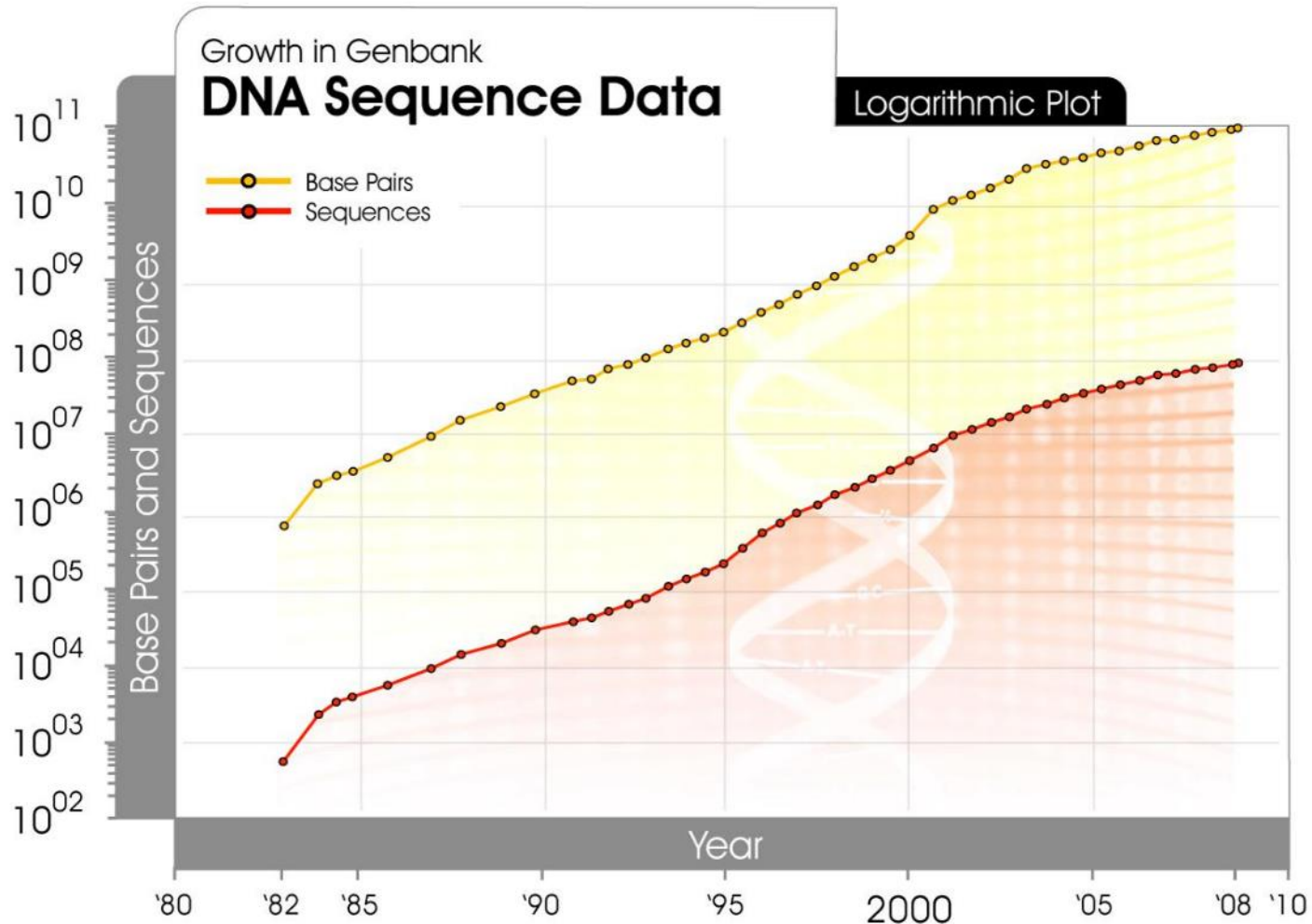
Bioinformatics: A brief history



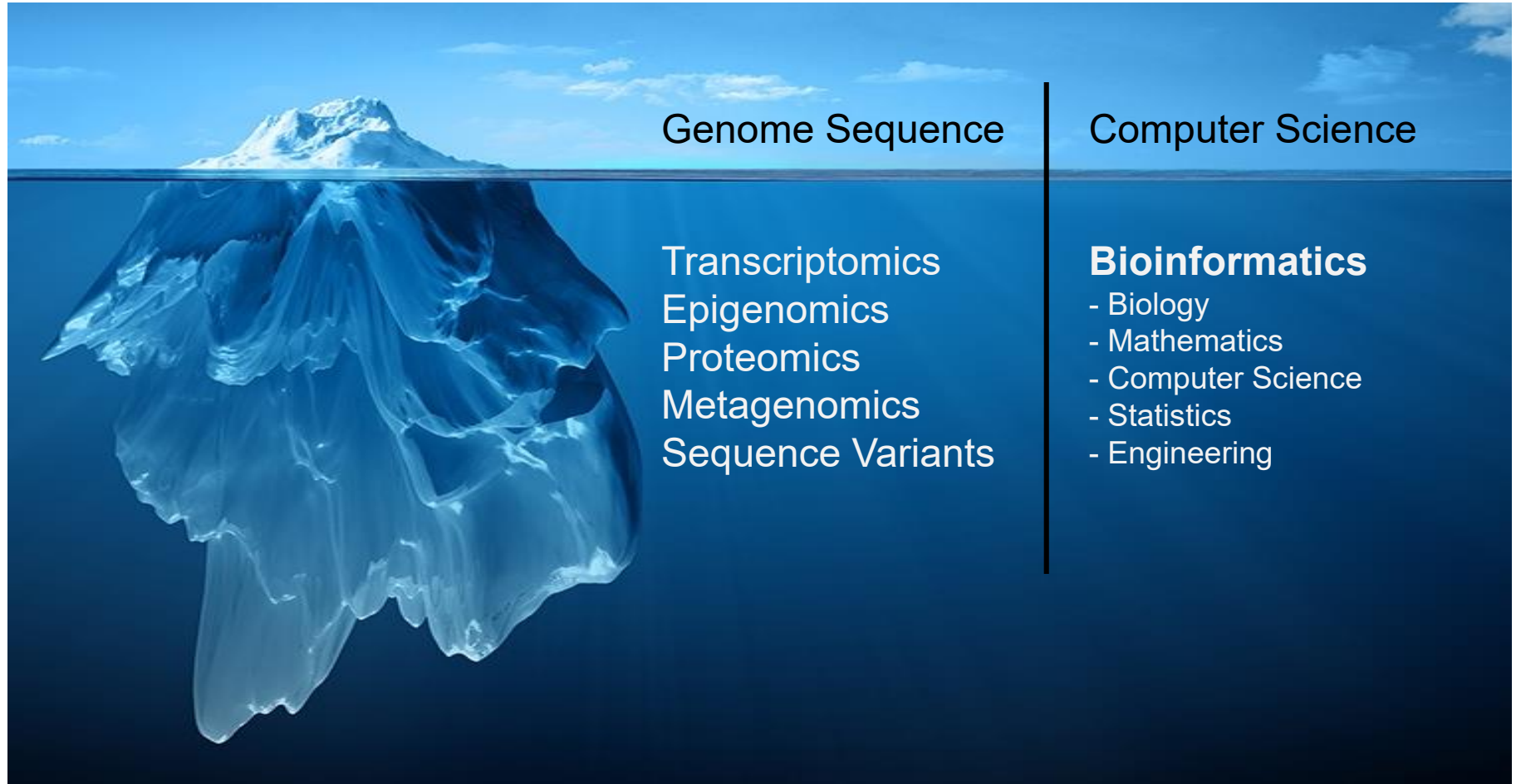
Bioinformatics: Early perceptions



Explosion in sequence data



Things aren't quite so simple ...







National Center for Biotechnology Information (NCBI)


- Portal to extensive database of resources
 - Analysis tools – BLAST
 - Sequence data – GenBank, GEO
 - Scientific literature – PubMed
 - Protein structures, Variation, etc

NCBI website

- <http://ncbi.nlm.nih.gov>

 NCBI Resources  How To Sign in to NCBI

 National Center for Biotechnology Information

All Databases 

Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation


Welcome to NCBI

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

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)


Submit

Deposit data or manuscripts into NCBI databases




Download

Transfer NCBI data to your computer




Learn

Find help documents, attend a class or watch a tutorial




Develop

Use NCBI APIs and code libraries to build applications




Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog


New releases from NCBI: Multiple Sequence Alignment Viewer 1.6, Tree Viewer 1.6.0, and Genome Workbench 2.12.0

14 Jul 2017

New on YouTube: "NCBI Minute: Tailor Your PubMed Search Experience with My NCBI"

13 Jul 2017

GenBank

 NCBI Resources ▾ How To ▾

Sign in to NCBI

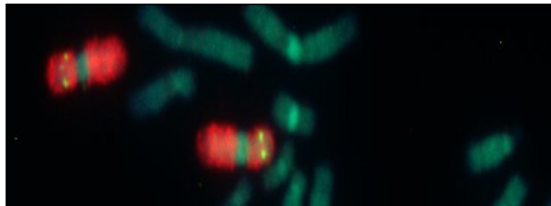
Gene

Gene ▾

Search

Advanced

Help



Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene

[Gene Quick Start](#)[FAQ](#)[Download/FTP](#)[RefSeq Mailing List](#)[Gene News](#) [Factsheet](#)

Gene Tools

[Submit GeneRIFs](#)[Submit Correction](#)[Statistics](#)[BLAST](#)[Genome Workbench](#)[Splign](#)

Other Resources

[HomoloGene](#)[OMIM](#)[RefSeq](#)[RefSeqGene](#)[UniGene](#)[Protein Clusters](#)

Representative queries

Find genes by...

Search text

free text

[human muscular dystrophy](#)

chromosome and symbol

[11\[chr\] OR 2\[chr\]\) AND adh*\[sym\]](#)

partial name and multiple species

[alive\[prop\] AND transporter\[title\] AND \("Drosophila melanogaster"\[orgn\] OR "Mus musculus"\[orgn\]\)](#)

associated sequence accession

[M11313\[accn\]](#)

gene name (symbol)

[BRCA1\[sym\]](#)

publication (PubMed ID)

[11331580\[PMID\]](#)

Demo

- Search 'gapdh' in GenBank

The screenshot shows the NCBI Gene search interface. At the top, the search bar contains 'gapdh' and the 'Search' button is visible. Below the search bar, there are links for 'Create RSS', 'Create alert', and 'Advanced'. The left sidebar lists various categories and sources, with 'Current' status selected. The main content area displays search results for 'gapdh'. A summary box at the top of the results indicates that there are 265 gene records for 'gapdh' across various species, including Homo sapiens, Mus musculus, and Rattus norvegicus. Below this, the 'Search results' section shows 'Items: 1 to 20 of 1122'. A table lists the first two results: GAPDH (human) and Gapdh (house mouse). The table columns are Name/Gene ID, Description, Location, Aliases, and MIM. To the right of the table, there are filters for 'Results by taxon' and 'Find related data'.

NCBI Resources How To Sign in to NCBI

Gene Gene gapdh Search

Create RSS Create alert Advanced Help

Gene sources
Genomic
Plasmids

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
✓ Current

Chromosome locations

Tabular 20 per page Sort by Relevance Send to

See GAPDH [glyceraldehyde-3-phosphate dehydrogenase](#)
[gapdh](#) in [Homo sapiens](#) [Mus musculus](#) [Rattus norvegicus](#) [All 265 Gene records](#)

Search results
Items: 1 to 20 of 1122
See also [1280 discontinued or replaced items](#).

Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> GAPDH ID: 2597	glyceraldehyde-3-phosphate dehydrogenase [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (6534405..6538375)	G3PD, GAPD, HEL-S-162eP	138400
<input type="checkbox"/> Gapdh ID: 14433	glyceraldehyde-3-phosphate dehydrogenase [<i>Mus musculus</i> (house mouse)]	Chromosome 6, NC_000072.6 (125161721..125166467,	Gapd	

clear

Filters: [Manage Filters](#)

Results by taxon

Top Organisms [\[Tree\]](#)
[Homo sapiens \(280\)](#)
[Rattus norvegicus \(95\)](#)
[Mus musculus \(66\)](#)
[Plasmodium falciparum 3D7 \(12\)](#)
[Arabidopsis thaliana \(11\)](#)
[All other taxa \(658\)](#)
[More...](#)

Find related data

Database: [Select](#)

[Find items](#)

- Available information
 - Sequence, genomic context, tissue expression, linked articles, etc.

PubMed

NCBI Resources How To

[Sign in to NCBI](#)

PubMed.gov

US National Library of Medicine
National Institutes of Health

PubMed

[Advanced](#)

Search

[Help](#)

PubMed

PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed

[PubMed Quick Start Guide](#)

[Full Text Articles](#)

[PubMed FAQs](#)

[PubMed Tutorials](#)

[New and Noteworthy](#)

PubMed Tools

[PubMed Mobile](#)

[Single Citation Matcher](#)

[Batch Citation Matcher](#)

[Clinical Queries](#)

[Topic-Specific Queries](#)

More Resources

[MeSH Database](#)

[Journals in NCBI Databases](#)

[Clinical Trials](#)

[E-Utilities \(API\)](#)

[LinkOut](#)

Latest Literature

New articles from highly accessed journals

Blood (15)

Cell (15)

Cochrane Database Syst Rev (3)

J Biol Chem (5)

J Clin Oncol (1)

Trending Articles

PubMed records with recent increases in activity

CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria.
Nature. 2017.

The effects of moderate- versus high-load resistance training on muscle growth, body composition, and performance in collegiate women.

J Strength Cond Res. 2017.

PubMed Commons

Featured comments

Modeling absence epilepsy in rats? @DepaulisAntoine et al critique analysis; author D Barth replies. [bit.ly/2u1YNqk](#)
Jul 14

Bias in biobank analyses—@MarcusMunafo et al discuss potential impact of low response rates on association estimates [bit.ly/2spEqrT](#)
Jul 13

Sequence Search

- **Problem**

- Search for an unknown (query) sequence from a large database of genomes

- **Tool**

BLAST

Basic Local Alignment Search Tool

BLAST workflow

- Choose appropriate BLAST program
- Enter query sequence
- Select database to search
- Run BLAST search
- Analyze output

NCBI BLAST



U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

Sign in to NCBI

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

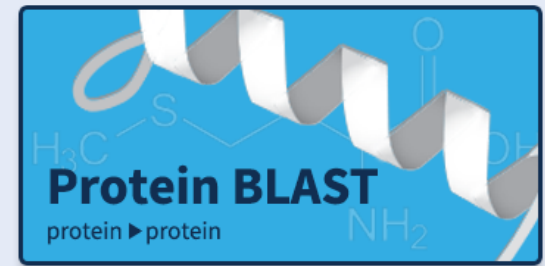
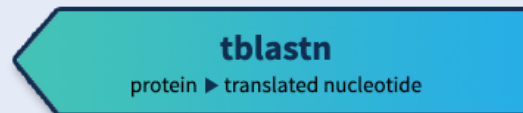
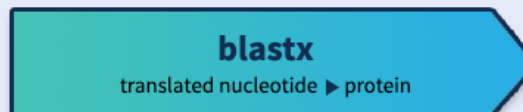
QuickBLASTP

Try [QuickBLASTP](#) for a fast protein search of nr.

Tue, 23 May 2017 13:00:00 EST

[More BLAST news...](#)

Web BLAST



BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

[Human](#)

[Mouse](#)

[Rat](#)

[Microbes](#)

Nucleotide BLAST (BLASTN)

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® » blastn suite Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#) Query subrange [?](#)

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

Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

Nucleotide collection (nr/nt) [?](#)

Organism [?](#)

Optional ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [?](#)

Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to [?](#)

Optional ☐ Sequences from type material

Entrez Query [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 [?](#)

Query Sequence

Select Database

Protein BLAST (BLASTP)

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® >> blastp suite Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#) Query subrange [?](#)

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

Database [?](#)

Organism Optional ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

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

Entrez Query Optional [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Algorithm

- ☐ Quick BLASTP (Accelerated protein-protein BLAST) **New**
- ☒ blastp (protein-protein BLAST)
- ☐ PSI-BLAST (Position-Specific Iterated BLAST)
- ☐ PHI-BLAST (Pattern Hit Initiated BLAST)
- ☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

BLAST output

- Job summary

BLAST® » **blastn suite** » RID-PUCHBH2N014

HomeRecent ResultsSaved StrategiesHelp

BLAST Results

[Edit and Resubmit](#) [Save Search Strategies](#) ▶ [Formatting options](#) ▶ [Download](#) [YouTube](#) [How to read this page](#) [Blast report description](#)

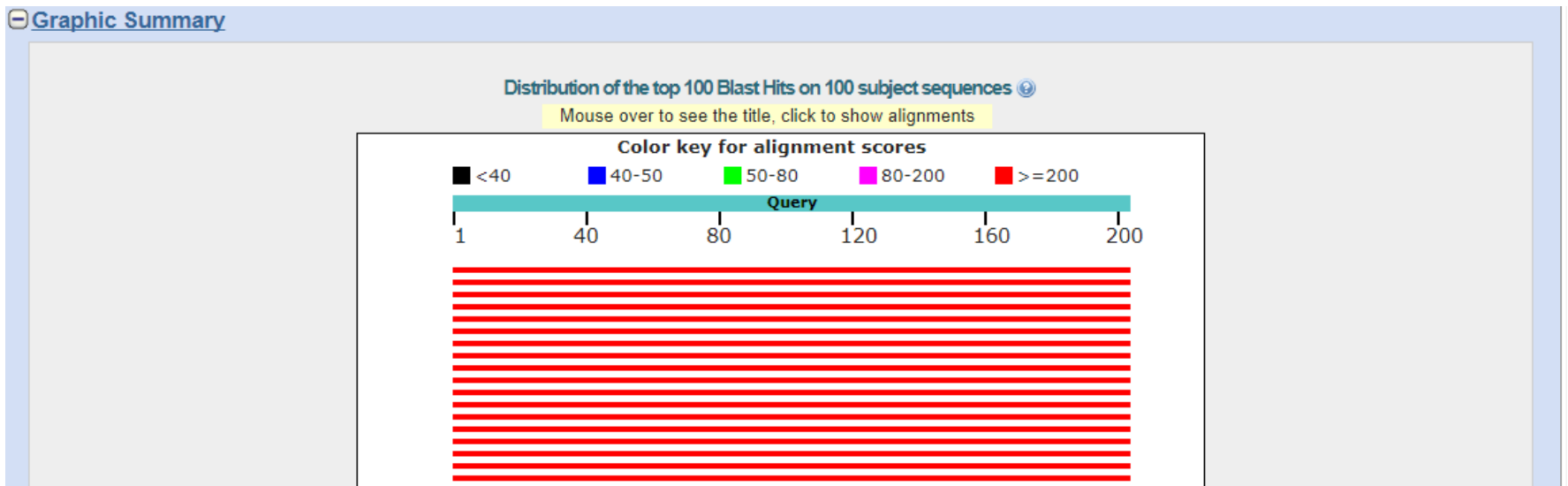
Job title: Nucleotide Sequence (204 letters)

RID	PUCHBH2N014 (Expires on 07-19 05:54 am)	
Query ID	Id Query_108271	Database Name nr
Description	None	Description Nucleotide collection (nt)
Molecule type	nucleic acid	Program BLASTN 2.6.1+ ▶ Citation
Query Length	204	

Other reports: ▶ [Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[MSA viewer\]](#)

BLAST output


- Graphic summary



- Color-coded based on scores
- Mouse-over for more information







BLAST output

- Ranked alignments with scores

 **Descriptions**

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

 [Alignments](#)  [Download](#)  [GenBank](#)  [Graphics](#)  [Distance tree of results](#) 

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Homo sapiens isolate GRC-TYR111292 tyrosinase precursor (TYR) gene, complete cds	369	369	100%	2e-98	100%	KJ528593.1
<input type="checkbox"/>	Homo sapiens haplotype HGDP00944_b tyrosinase (TYR) gene, complete cds	369	369	100%	2e-98	100%	KC201588.1
<input type="checkbox"/>	Homo sapiens haplotype HGDP00944_a tyrosinase (TYR) gene, complete cds	369	369	100%	2e-98	100%	KC201587.1
<input type="checkbox"/>	Homo sapiens haplotype H1_a tyrosinase (TYR) gene, complete cds	369	369	100%	2e-98	100%	KC201583.1
<input type="checkbox"/>	Homo sapiens haplotype HGDP00923_b tyrosinase (TYR) gene, complete cds	369	369	100%	2e-98	100%	KC201582.1
<input type="checkbox"/>	Homo sapiens haplotype HGDP00923_a tyrosinase (TYR) gene, complete cds	369	369	100%	2e-98	100%	KC201581.1

- Higher scores = better alignment
- Lower E-values = lower probability that alignment was random chance

BLAST output

- Detailed alignments

Alignments

Download ▾ [GenBank](#) [Graphics](#) [Next](#) [Previous](#) [Descriptions](#)

Homo sapiens isolate GRC-TYR111292 tyrosinase precursor (TYR) gene, complete cds
Sequence ID: [KJ528593.1](#) Length: 4651 Number of Matches: 1

Range 1: 4235 to 4438 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
369 bits(408)	2e-98	204/204(100%)	0/204(0%)	Plus/Plus
Query 1	TCAGCCCTTTTAACATTTTCCCTAAGCCCATATGTCTAAGGAAAGGATGCTATTTGGTA	60		
Sbjct 4235	TCAGCCCTTTTAACATTTTCCCTAAGCCCATATGTCTAAGGAAAGGATGCTATTTGGTA	4294		
Query 61	ATGAGGAACTGTTATTTGTATGTGAATTAAGTGCTCTTATTTAAAAAATTGAAATAAT	120		
Sbjct 4295	ATGAGGAACTGTTATTTGTATGTGAATTAAGTGCTCTTATTTAAAAAATTGAAATAAT	4354		
Query 121	TTTGATTTTGCCTTCTGATTATTTAAAGATCTATATGTTTTATTGGCCCTTCTTTA	180		
Sbjct 4355	TTTGATTTTGCCTTCTGATTATTTAAAGATCTATATGTTTTATTGGCCCTTCTTTA	4414		
Query 181	TTTTAATAAAACAGTGAGAAATCT	204		
Sbjct 4415	TTTTAATAAAACAGTGAGAAATCT	4438		

Related Information
[Gene](#) - associated gene details
[Map Viewer](#) - aligned genomic context

- Visual representation of alignment
- Can include gaps or mismatches

Demo

- <https://digitalworldbiology.com/BLAST/sequences>
- Copy and paste sequences into BLAST
 - Examine results

Multiple sequence alignment

- Alignment of three or more sequences to infer sequence homology and study evolutionary relationships
 - Protein, DNA or RNA
- **Tool**

CLUSTAL

Cluster Analysis of Pairwise Alignments

CLUSTAL workflow

- Input sequences or upload file
- Specify parameters
- Run CLUSTAL
- Analyze results

Clustal Omega

- <http://www.ebi.ac.uk/Tools/msa/clustalo/>

The screenshot shows the Clustal Omega web interface. At the top is a navigation bar with links: EMBL-EBI, Services, Research, Training, Industry, About us, and a search icon. On the right of the navigation bar is 'EMBL-EBI Hinxton'. Below the navigation bar is a teal header with 'Clustal Omega' in white. Under the header are tabs: 'Input form' (selected), 'Web services', and 'Help & Documentation'. On the right of the header are links for 'Feedback' and 'Share'. Below the header is a breadcrumb trail: 'Tools > Multiple Sequence Alignment > Clustal Omega'. The main heading is 'Multiple Sequence Alignment'. Below this is a paragraph: 'Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).' Below this is an 'Important note': 'This tool can align up to 4000 sequences or a maximum file size of 4 MB.' The main form is titled 'STEP 1 - Enter your input sequences'. It has a dropdown menu labeled 'Enter or paste a set of' with 'PROTEIN' selected. Below this is a text area labeled 'sequences in any supported format:'. At the bottom of the form is a section 'Or, upload a file:' with a 'Choose File' button and 'No file chosen' text. Below the main form is 'STEP 2 - Set your parameters'. Three arrows point to the form: one to the dropdown menu labeled 'Type of sequence', one to the text area labeled 'Query Sequence or file', and one to the 'Choose File' button.

STEP 1 - Enter your input sequences

Enter or paste a set of
PROTEIN

sequences in any supported format:

Or, upload a file: No file chosen

STEP 2 - Set your parameters

Type of sequence

Query Sequence or file

Clustal Omega Output

- Alignment

```
Borrelia      YKIVEIVSDGDYSIDEQIAVIEDDSGMRHNITMSFHWPVKVPITNYKERLIPSEPMLTQT
Candida       GTITSIAEAGSYNVEEPVLEVE-FDGKKHKYSMMHTWPVRVPRP-VAEKLTA DHPLLTGQ
Saccharomyces GTITWIAPAGEYTLDEKILEVE-FDGKKSDFTLYHTWPVRVPRP-VTEKLSADYPLLTGQ
Neurospora    GTITRIAEKGEYTVEEKILEVE-FDGKKTEYPMMQTWPVRVPRP-AAEKHSANQPFLVGQ
Trypanosoma   GRVTSIVPSGNYTLQDDIIELE-YNGTVKSLKLMHRWPVRTPRP-VASKESGNHPLLTGQ
Drosophila    GTVRYIAPSGNYKVDDVVLETE-FDGEITKHTMLQVWPVRHHAP-VTEKLPANHPLLTGQ
Acetabularia GTVTYIAAPGNYTINEKIIEVE-FQGAKYEYSMKQSWPVRSPRP-VVEKLLADTPLL TGQ
Daucus        GKITYVAPAGQYSLKDTVLELE-FQGVKKQFTMLQTWPVRTPRP-VASKLAADTPLL TGQ
Sulfolobus    GTLKE LAREGDYTVEDVVAVVD-MNGDEIPVKMYQKWPVRIPRP-YKEKLEPVEPLL TI
Thermococcus  GEIVEIAEEGDYTVEEVIVKVKKPDGTIEELKMYHRWPVRVKRP-YKQKLPPEVPLITGQ
               :  :.  *. * . . . :  .  . *      :   ***:      .:   * : .
```

Symbol	Meaning
*	Fully conserved residue
:	Strongly related residues
.	Weakly related residues

Clustal Omega Output

- Alignment

```

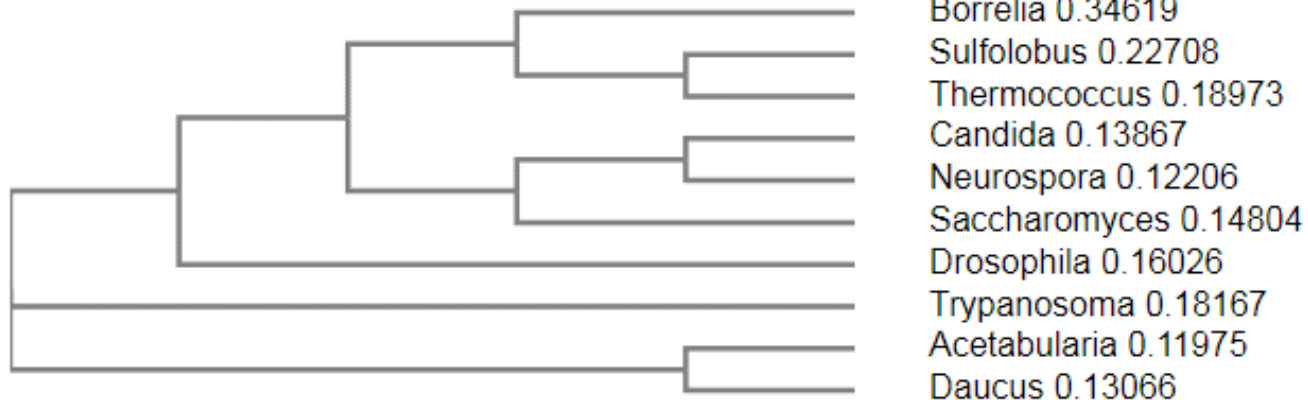
Borrelia      YKIVEIVSDGDYSIDEQIAVIEDDSGMRHNITMSFHWPVKVPITNYKERLIPSEPMLTQT
Candida       GTITSIAEAGSYNVEEPVLEVE-FDGKKHKYSMMHTWPVRVPRP-VAEKLTAADHPLLGTGQ
Saccharomyces GTITWIAPAGEYTLDEKILEVE-FDGKKSDFTLYHTWPVRVPRP-VTEKLSADYPLLGTGQ
Neurospora    GTITRIAEKGEYVEEKILEVE-FDGKKTEYPMMQTWPVRVPRP-AAEKHSANQPFLVGQ
Trypanosoma   GRVTSIVPSGNYTLQDDIIELE-YNGTVKSLKLMHRWPVRTPRP-VASKESGNHPLLGTGQ
Drosophila    GTVRYIAPSGNYKVDDVVLETE-FDGEITKHTMLQVWPVRHHAP-VTEKLPANHPLLGTGQ
Acetabularia GTVTYIAAPGNYTINEKIIEVE-FQGAKYEYSMKQSWPVRSPRP-VVEKLLADTPLLGTGQ
Daucus        GKITYVAPAGQYSLKDTVLELE-FQGVKKQFTMLQTWPVRTPRP-VASKLAADTPLLGTGQ
Sulfolobus    GTLKEELAREGDYTVEDVAVVD-MNGDEIPVKMYQKWPVRIPRP-YKEKLEPVEPLLGTI
Thermococcus  GEIVEIAEEGDYTVEEVIVKVKKPDGTIEELKMYHRWPVRVKRP-YKQKLPPEVPLITGQ
      :  :.  *.*.:.: :  .  .*      :  ***:      .:  *:..

```

Color	Residues	Properties
Red	AVFPMILW	Small (small+ hydrophobic (incl. aromatic -Y))
Blue	DE	Acidic
Magenta	RK	Basic
Green	STYHCNGQ	Hydroxyl + sulfhydryl + amine + G
Grey	Others	Unusual amino acids

- Phylogenetic tree

Branch length: ☒ Cladogram ☐ Real



Demo

- Download file:
 - http://www.bioinformaticsworld.com/clustal_seq.txt
- Go to Clustal Omega website
 - Run on downloaded file

Next-generation sequencing (NGS)

- Quantum leap of DNA sequencing technology
- Ability to sequence whole genomes fast and cheap
- Large repository of public data
 - ENCODE project
- Tools to browse genomes and view data
 - UCSC genome browser

ENCODE project

- Encyclopedia of DNA elements
 - <https://www.encodeproject.org/>
- Launched in 2003 to identify all functional elements in human genome
 - Extended to mouse, worm, fly
- Massive consortium led by NHGRI

ENCODE project: Datasets

HUMAN MOUSE WORM FLY

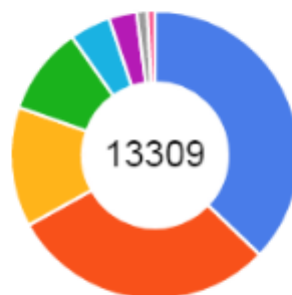
[View Assay Matrix](#)

Project



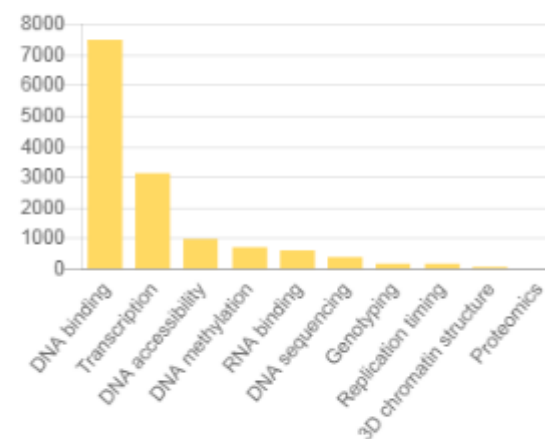
ENCODE
Roadmap
modENCODE
modERN
GGR

Biosample Type



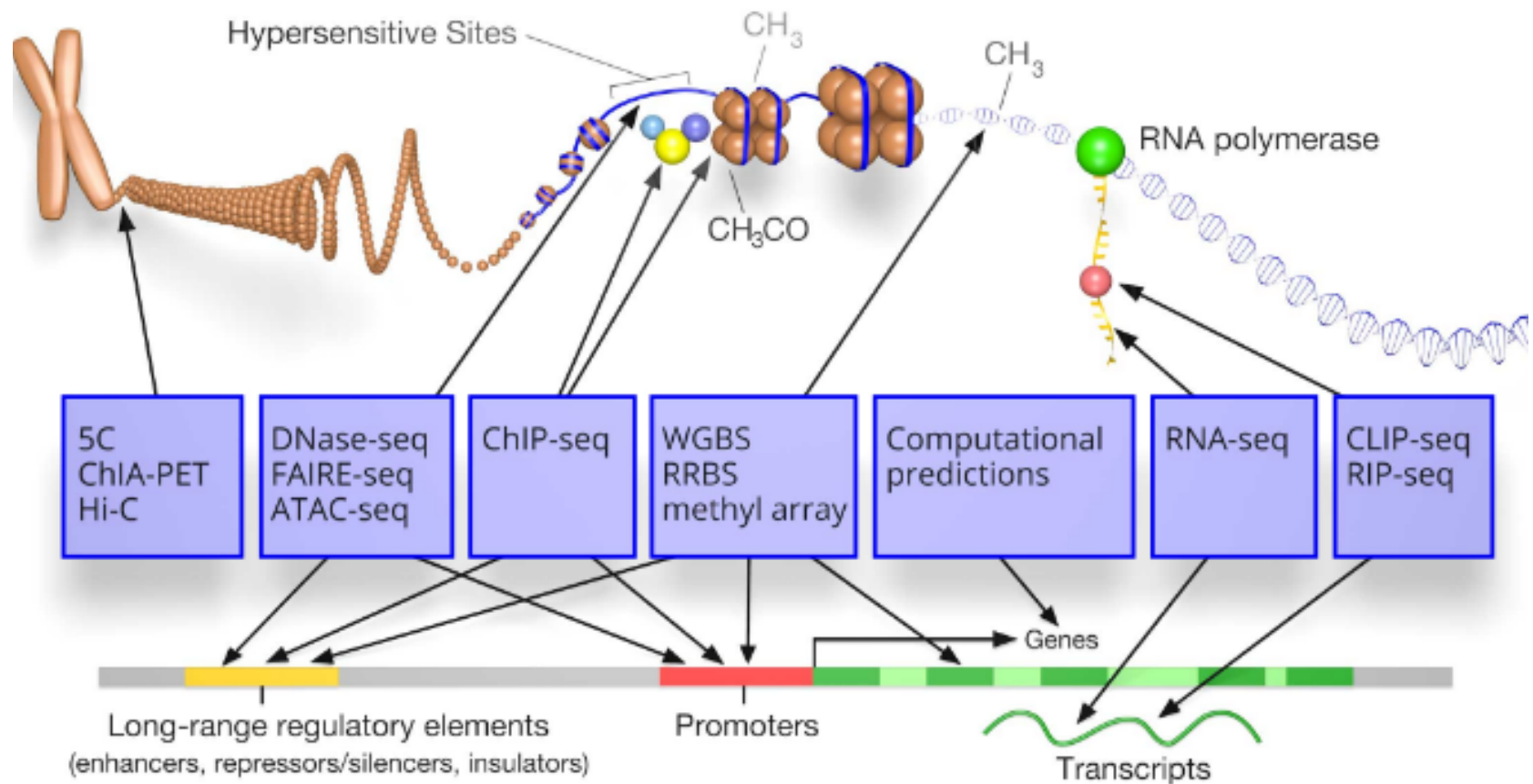
immortalized cell line
tissue
primary cell
whole organisms
in vitro differentiated cells
stem cell
in vitro sample
induced pluripotent stem cell line

Assay Categories

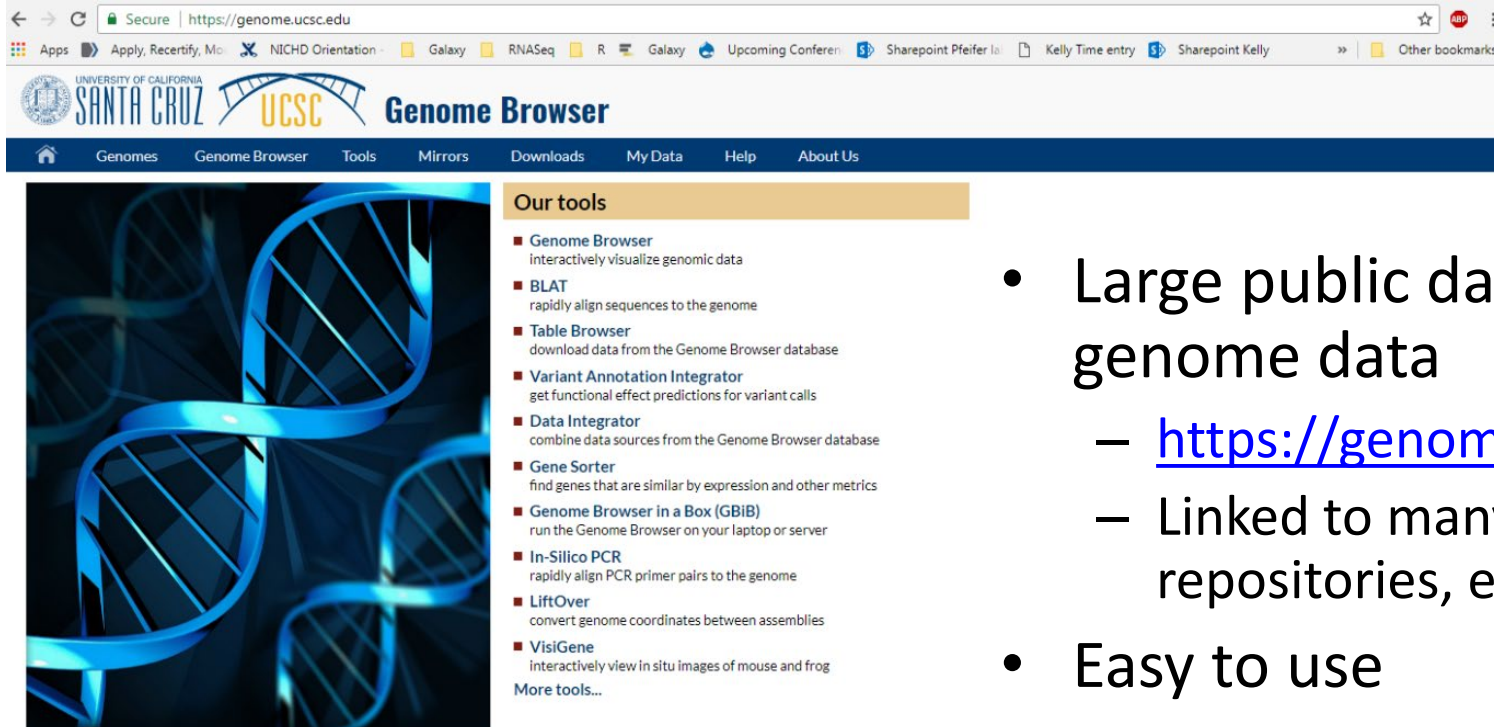


<https://www.encodeproject.org/>

ENCODE project: Assays



UCSC Genome Browser



The screenshot shows the UCSC Genome Browser homepage. At the top is a browser window with the URL <https://genome.ucsc.edu>. Below the browser window is the UCSC logo and the text "Genome Browser". A navigation bar contains links: Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The main content area features a large image of a DNA double helix on the left. To the right of the image is a section titled "Our tools" with a list of tools and their descriptions:

- **Genome Browser**
interactively visualize genomic data
- **BLAT**
rapidly align sequences to the genome
- **Table Browser**
download data from the Genome Browser database
- **Variant Annotation Integrator**
get functional effect predictions for variant calls
- **Data Integrator**
combine data sources from the Genome Browser database
- **Gene Sorter**
find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**
run the Genome Browser on your laptop or server
- **In-Silico PCR**
rapidly align PCR primer pairs to the genome
- **LiftOver**
convert genome coordinates between assemblies
- **VisiGene**
interactively view in situ images of mouse and frog

Below the list is a link "More tools...".

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly

What's new

May 03, 2018 - Updated GENCODE gene tracks for human assemblies, hg19 and hg38

Apr. 04, 2018 - New NCBI RefSeq tracks for model organisms

- Large public database of genome data
 - <https://genome.ucsc.edu/>
 - Linked to many public repositories, e.g. ENCODE
- Easy to use
- Customizable view
- Can upload your own data ('custom tracks')

UCSC Genome Browser

Mouse genome, mm9

Genomic location → chr19:10,810,307-10,861,564 51,258 bp

Search box → enter position, gene symbol or search terms

Annotation → UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics), Ensemble Gene Predictions - archive 65 - dec2011, Human Proteins Mapped by Chained tBLASTn, Non-Mouse RefSeq Genes, RefSeq Genes, Publications: Sequences in Scientific Articles, Mouse ESTs That Have Been Spliced, Placental Mammal Basewise Conservation by PhyloP

Conservation → Mammal Cons, Multiz Alignments of 39 Vertebrates

Repeat elements → RepeatMasker

Additional options → Mapping and Sequencing (Base Position, STS Markers, Assembly, BAC End Pairs, Chromosome Band, Gap, GC Percent, GRC Incident, Map Contigs, Mappability, MGI QTL, Restr Enzymes, Short Match, Wiki Track)

Annotation tracks

The screenshot displays the UCSC Genome Browser interface for Mouse mm9 chr19:10810. The browser window shows the URL <https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr19%3...>. The main track area shows multiple tracks including "Multiz Alignments of 38 Vertebrates", "Simple Nucleotide Polymorphisms (dbSNP build 128)", and "Repeats (RepeatMasker)". Below the tracks, there are controls for "move start" and "move end" with a range of 2.0. A toolbar includes buttons for "track search", "default tracks", "default order", "hide all", "add custom tracks", "track hubs", "configure", "multi-region", "reverse", "resize", "refresh", "collapse all", and "expand all".

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing refresh

Base Position dense ▾ GC Percent hide ▾ Short Match hide ▾	STS Markers dense ▾ GRC Incident hide ▾ Wiki Track hide ▾	Assembly hide ▾ Map Contigs hide ▾	BAC End Pairs hide ▾ Mappability hide ▾	Chromosome Band hide ▾ MGI QTL hide ▾	Gap hide ▾ Restr Enzymes hide ▾
---	--	---	--	--	--

Genes and Gene Predictions refresh

UCSC Genes pack ▾ AUGUSTUS hide ▾ Genscan Genes hide ▾ Old UCSC Genes hide ▾ UCSC Alt Events hide ▾	Ensembl Genes pack ▾ CCDS hide ▾ IKMC Genes hide ▾ ORFeome Clones hide ▾ Vega Genes hide ▾	Human Proteins pack ▾ CRISPR hide ▾ MGC Genes hide ▾ Pfam in UCSC Gene hide ▾ Yale Pseudo60 hide ▾	Other RefSeq hide ▾ Exoniphy hide ▾ miRNA hide ▾ SGP Genes hide ▾	RefSeq Genes dense ▾ Gene Trap hide ▾ N-SCAN hide ▾ Transcriptome hide ▾	AceView Genes hide ▾ Geneid Genes hide ▾ NIA Gene Index hide ▾ tRNA Genes hide ▾
--	---	---	--	---	---

Literature refresh

Publications dense ▾	Web Sequences hide ▾
---	---

mRNA and EST refresh

Spliced ESTs dense ▾	Mouse ESTs hide ▾	Mouse mRNAs hide ▾	Other mRNAs hide ▾	PolyA-Seq hide ▾	SIB Alt-Splicing hide ▾
---	--------------------------------------	---------------------------------------	---------------------------------------	-------------------------------------	--

ENCODE @ UCSC

The screenshot displays the UCSC Genome Browser interface for Mouse mm9 chr19:10810. The browser window shows the URL <https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr19%3...>. The interface includes various tracks for genomic data, with the 'Expression and Regulation' section highlighted by a red box and an arrow pointing to the 'LICR TFBS' track.

The tracks shown include:

- UCSC Alt Events
- Vega Genes
- Yale Pseudo60
- Literature
- mRNA and EST
- Phenotype and Allele
- Expression and Regulation
- Comparative Genomics

The 'Expression and Regulation' section contains the following tracks:

- Affy Exon...
- Affy GNF-1M
- Affy MOE430
- Affy U74
- Allen Brain
- Caltech Histone
- Caltech RNA-seq
- Caltech TFBS
- CpG Islands...
- CSHL Long RNA-seq
- FaceBase 24STypes
- FSU Repli-chip
- GNF Atlas 2
- GNF U74A
- GNF U74B
- GNF U74C
- LICR Histone
- LICR RNA-seq
- LICR TFBS
- NHGRI BiP
- NKI Nuc Lamina...
- ORegAnno
- PSU DNaseI HS
- PSU Histone
- PSU RNA-seq
- PSU TFBS
- REST
- Stan/Yale Histone
- Stan/Yale RNA-seq
- Stan/Yale TFBS
- TS miRNA sites
- UW DNaseI DGE
- UW DNaseI HS
- UW RNA-seq

The 'Comparative Genomics' section contains the following tracks:

- Conservation
- GERP
- Rat Chain/Net
- Guinea pig Chain/Net
- Rabbit Chain/Net
- Marmoset Chain/Net
- Rhesus Chain/Net
- Orangutan Chain/Net
- Chimp Chain/Net
- Human Chain/Net
- Panda Chain/Net
- Dog Chain/Net
- Cat Chain/Net
- Horse Chain/Net
- Sheep Chain/Net
- bosTau6 Chain/Net
- Pig Chain/Net
- Elephant Chain/Net
- Opossum Chain/Net
- Platypus Chain/Net
- Lizard Chain/Net
- Turkey Chain/Net
- Chicken Chain/Net
- X_tropicalis Chain/Net

ENCODE @ UCSC

LICR TFBS Track Settings

Transcription Factor Binding Sites by ChIP-seq from ENCODE/LICR (All Expression and Regulation tracks)

Maximum display mode: [Reset to defaults](#)

Select views (help):
Peaks

Select subtracks by cell line and factor:

Cell Line	Factor	CTCF	p300	Pol2	Input
Bone Marrow	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bone Marrow	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bone Marrow	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Bone Marrow	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Derived Macrophage	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Derived Macrophage	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Derived Macrophage	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Derived Macrophage	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Cerebellum	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cerebellum	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cerebellum	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Cerebellum	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ES-Bruc4	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ES-Bruc4	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ES-Bruc4	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
ES-Bruc4	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Heart	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Heart	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Heart	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Heart	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Kidney	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Kidney	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Kidney	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Kidney	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Limb	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Limb	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Limb	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Limb	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Liver	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Liver	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Liver	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Liver	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lung	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lung	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lung	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Lung	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
MEF	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MEF	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MEF	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
MEF	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Olfactory Bulb	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Olfactory Bulb	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Olfactory Bulb	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Olfactory Bulb	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Small Intestine	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Small Intestine	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Small Intestine	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Small Intestine	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Spleen	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spleen	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spleen	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Spleen	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Testis	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Testis	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Testis	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Testis	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Thymus	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Thymus	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Thymus	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Thymus	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Whole Brain	CTCF	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Whole Brain	p300	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Whole Brain	Pol2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Whole Brain	Input	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Select subtracks further by: (select multiple categories and items - help)

Age:
Adult 8 weeks
Immortal cells

List subtracks: ☒ only selected/visible ☐ all (18 of 96 selected)

Cell Line	Factor	Views	Age	Track Name	Restricted Until
Cerebellum	CTCF	Peaks	Adult 8 weeks	Cerebellum Adult 8 weeks CTCF TFBS ChIP-seq Peaks from ENCODE/LICR	2011-10-19
Cerebellum	CTCF	Signal	Adult 8 weeks	Cerebellum Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR	2011-10-19
Cerebellum	Pol2	Peaks	Adult 8 weeks	Cerebellum Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR	2011-12-07

ENCODE @ UCSC

- Separate website with ENCODE data also available

<https://genome.ucsc.edu/ENCODE/>



Encyclopedia of DNA Elements at UCSC 2003 - 2012

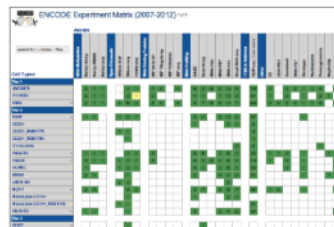
About

The [Encyclopedia of DNA Elements](#) (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

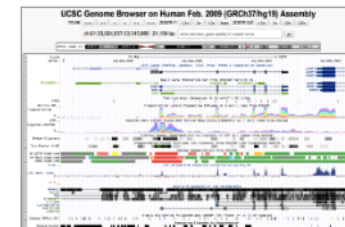
ENCODE results from 2007 and later are available from the ENCODE Project Portal, encodeproject.org. This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. **All ENCODE data at UCSC are freely available for download and analysis.**

Explore ENCODE data (2003 - 2012) at UCSC



View ENCODE data (2003 - 2012) in the UCSC Genome Browser



Search for data (current) at the ENCODE Portal

Search for ENCODE tracks (2003 - 2012) in the UCSC Browser

Search
human data

Search
mouse data

ENCODE @ UCSC

- Search full database for tracks to display

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Help](#) [About Us](#)

Search for Tracks in the Human Feb. 2009 (GRCh37/hg19) Assembly

Search

Advanced

Track Name: contains

and Description: contains

and Group: is

and Data Format: is

ENCODE terms

and is among [Cell, tissue or DNA sample](#)

and is among [Antibody or target protein](#)

About Track Search

Search for terms in track names, descriptions, groups, and ENCODE metadata. If multiple terms are entered, only tracks with all terms will be part of the results. [more help](#)

Recap

- What is bioinformatics?
 - NCBI portal – GenBank, Pubmed
- Sequence analysis
 - NCBI BLAST
- Multiple sequence alignment
 - Clustal Omega
- Next-generation sequencing (NGS)
 - UCSC Genome Browser
 - ENCODE project

Questions?

Demo files

- BLAST
 - <https://digitalworldbiology.com/BLAST/sequences>
- CLUSTAL
 - [http://www.bioinformaticsworld.com/clustal seq.
txt](http://www.bioinformaticsworld.com/clustal_seq.txt)