An Introduction to Bioinformatics Strategies

Apratim Mitra, Ph.D.

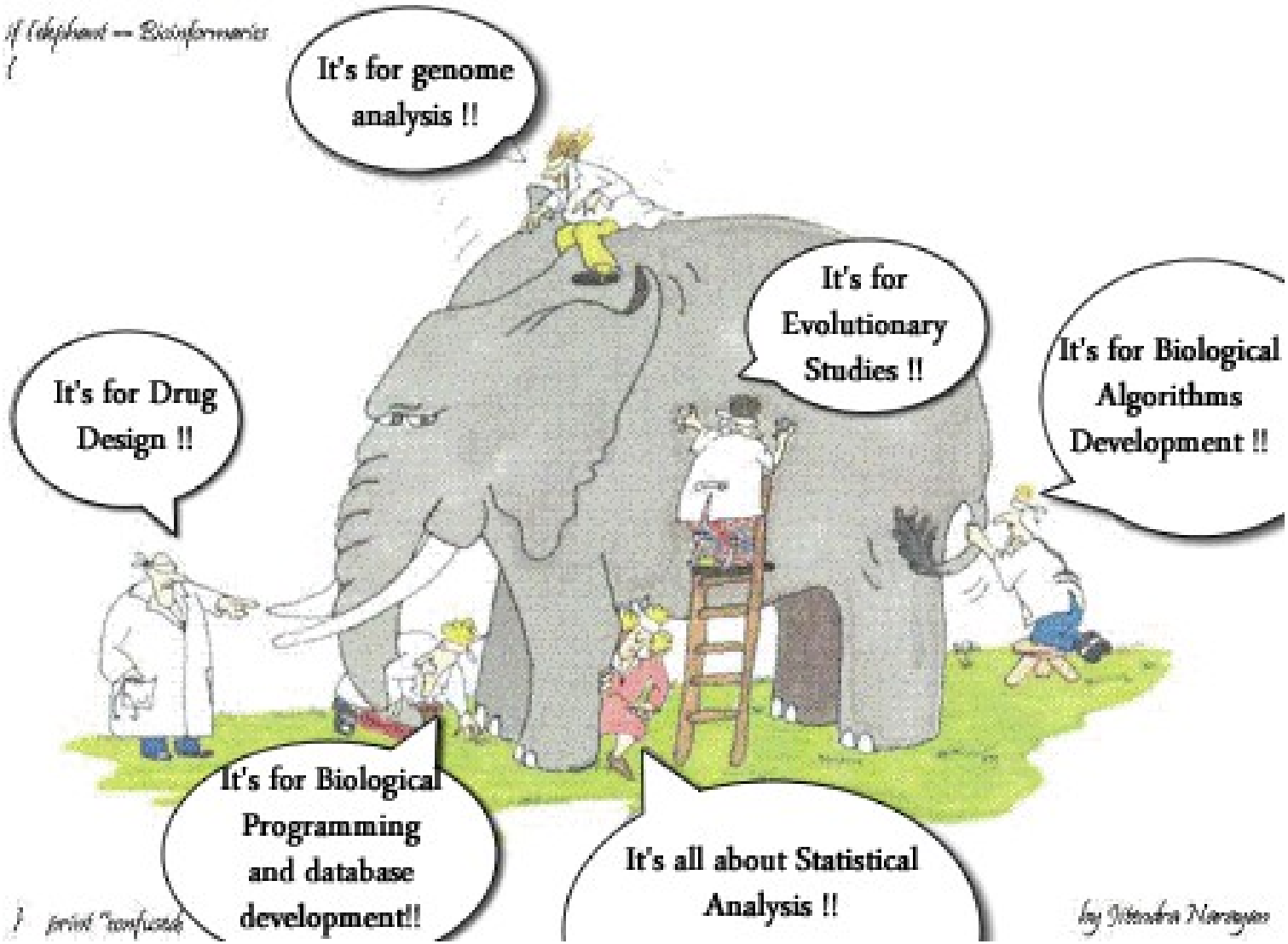
NICHD/NIH

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?

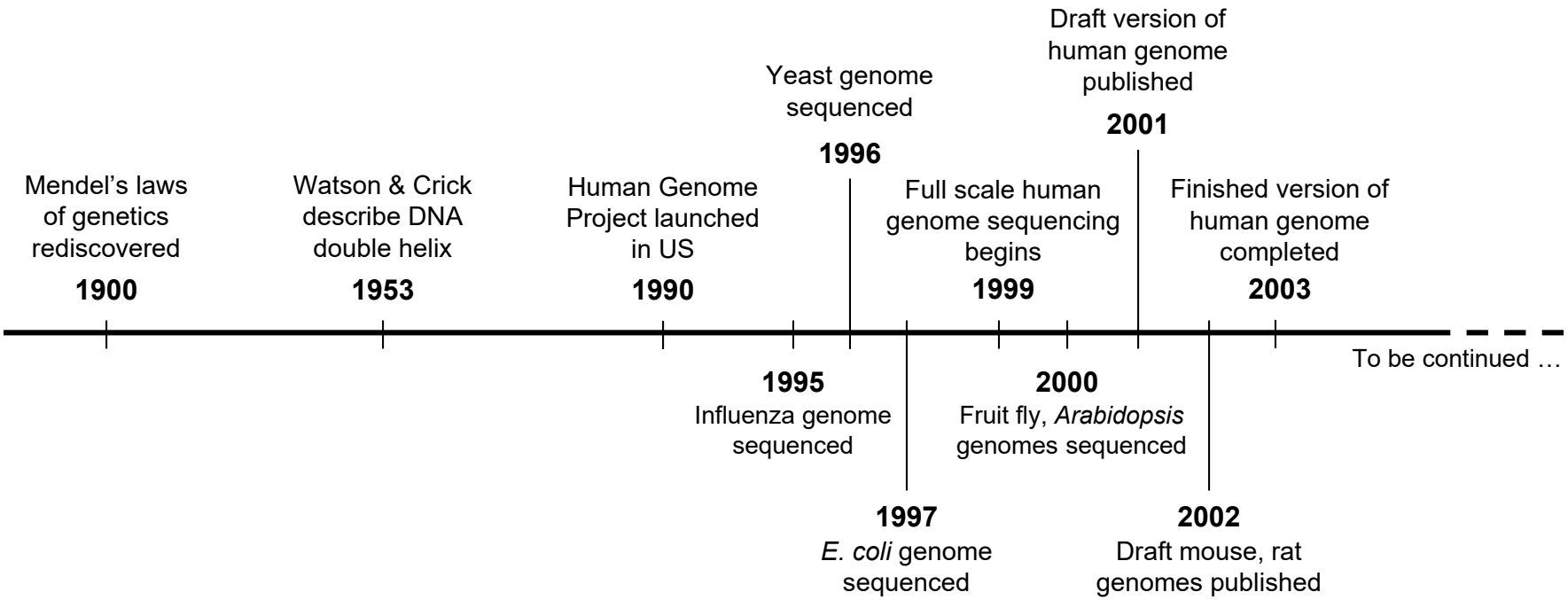


http://bioinformaticsonline.com/file/download/4482

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

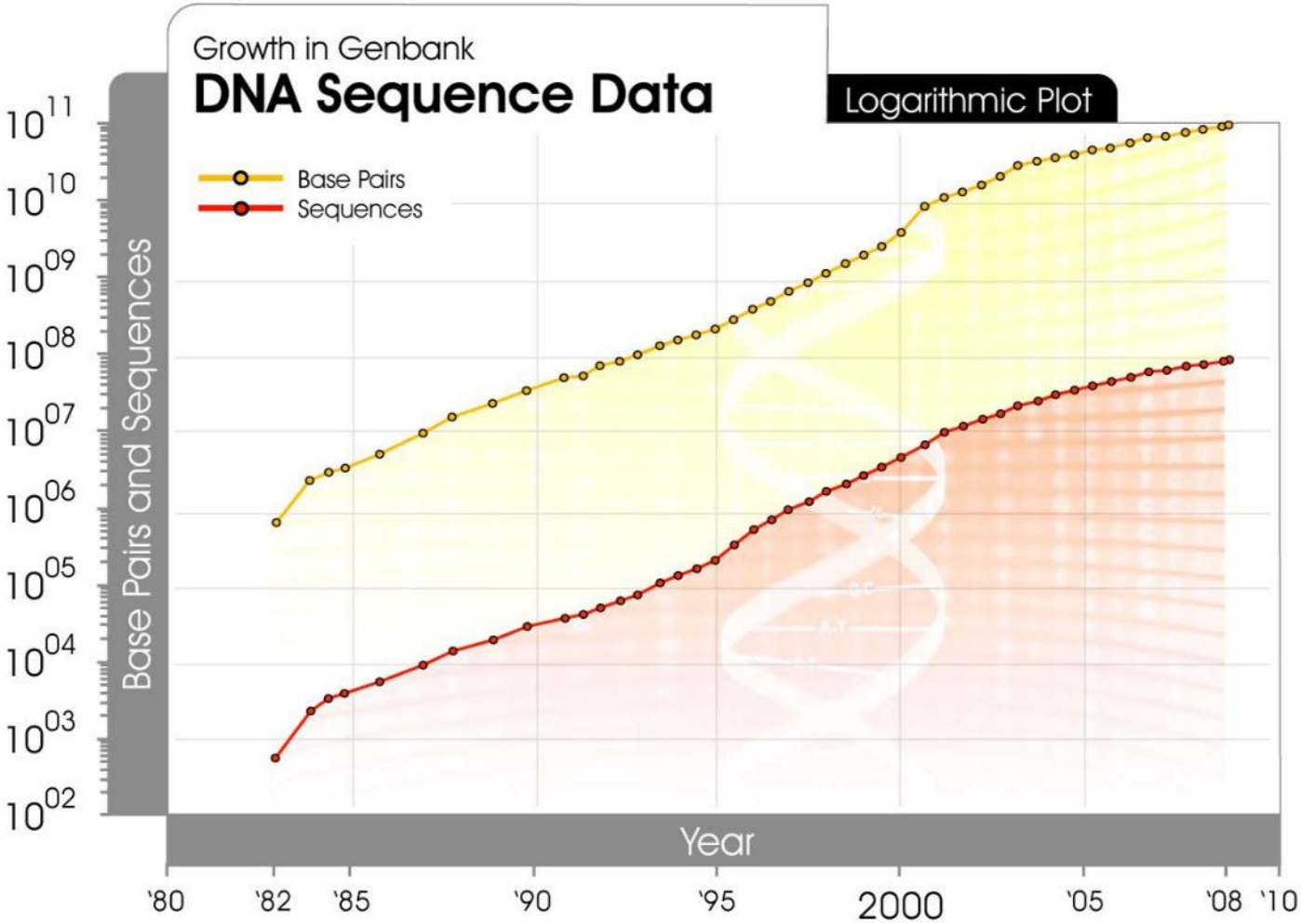
Bioinformatics: A brief history



Bioinformatics: Early perceptions

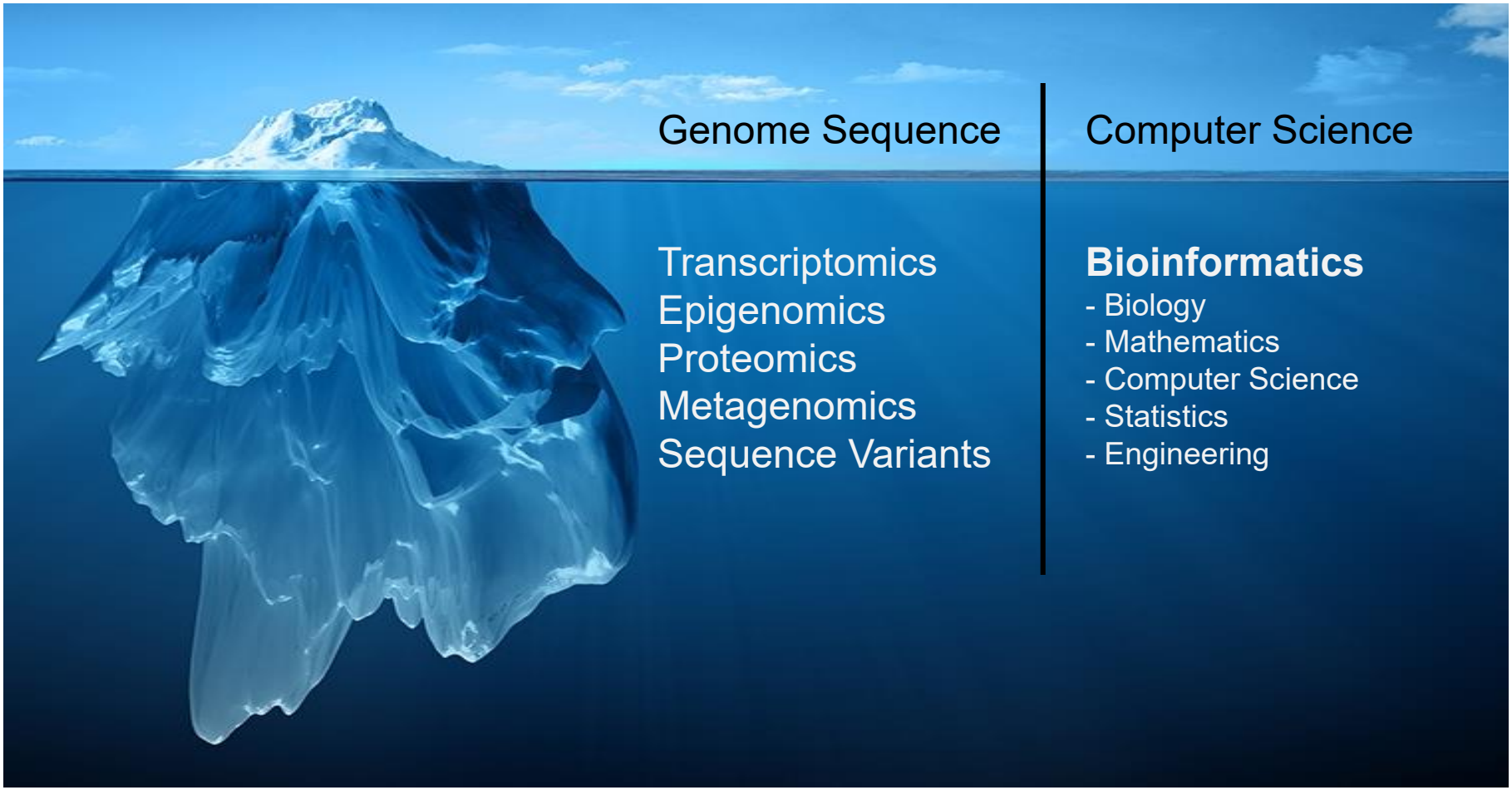
http://www.philcallaway.ab.ca/images/Cartoons/privacy%20cartoon2.jpg

Explosion in sequence data



http://www.kurzweilai.net/dna_sequencing_data.jpg

Things aren’t quite so simple …

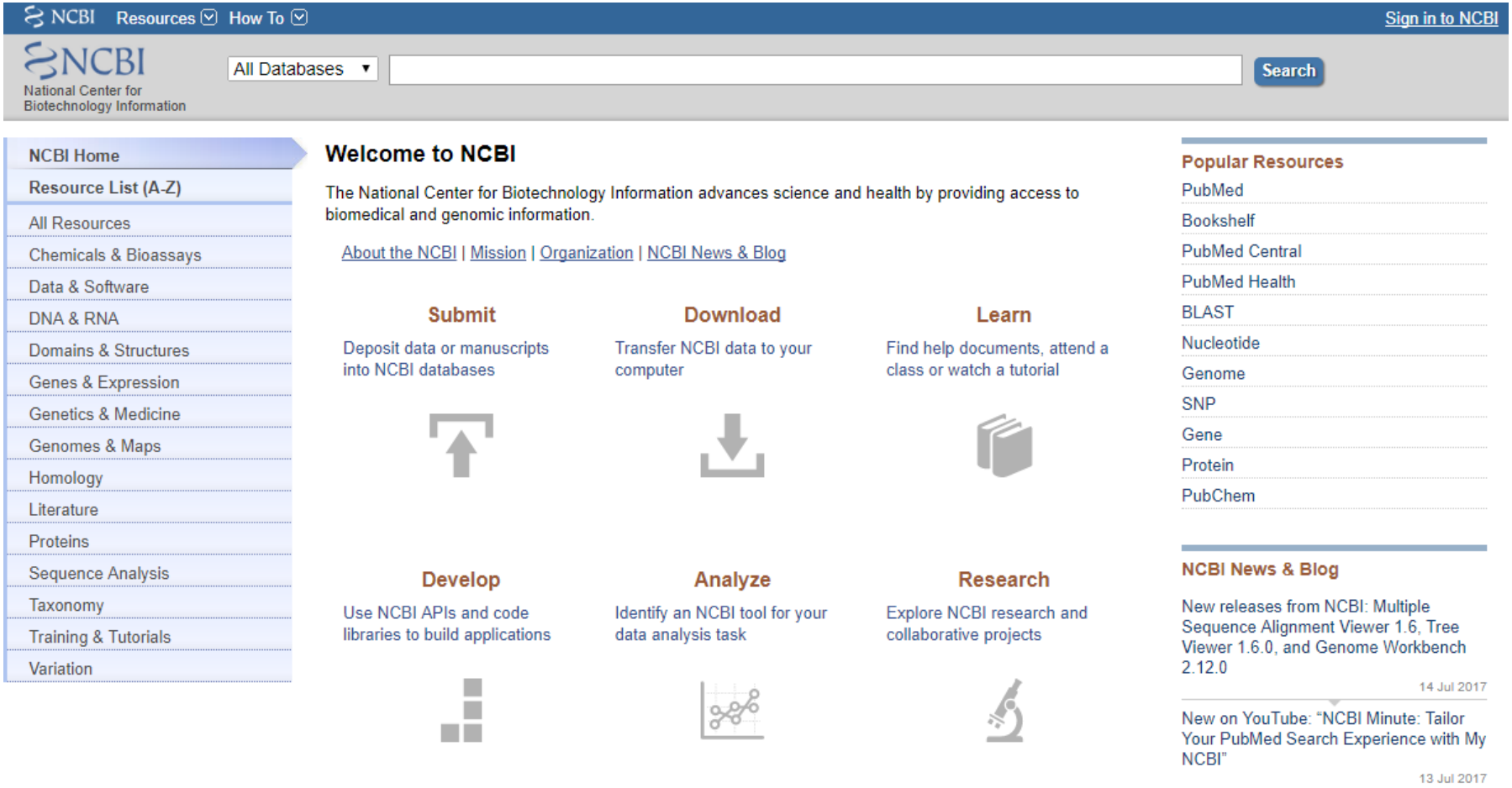


National Center for Biotechnology Information (NCBI)

- 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

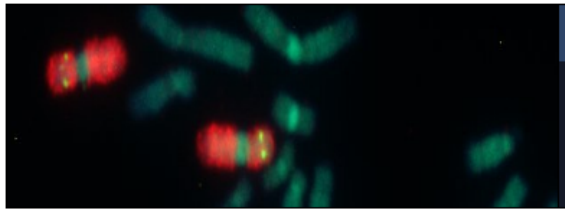


GenBank

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	(II[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

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 clear
✓ 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

<< First < Prev Page 1 of 57 Next > Last >>

See also 1280 discontinued or replaced items.

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

Hide sidebar >>

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 Search

Advanced 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/2ulYNqk
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**

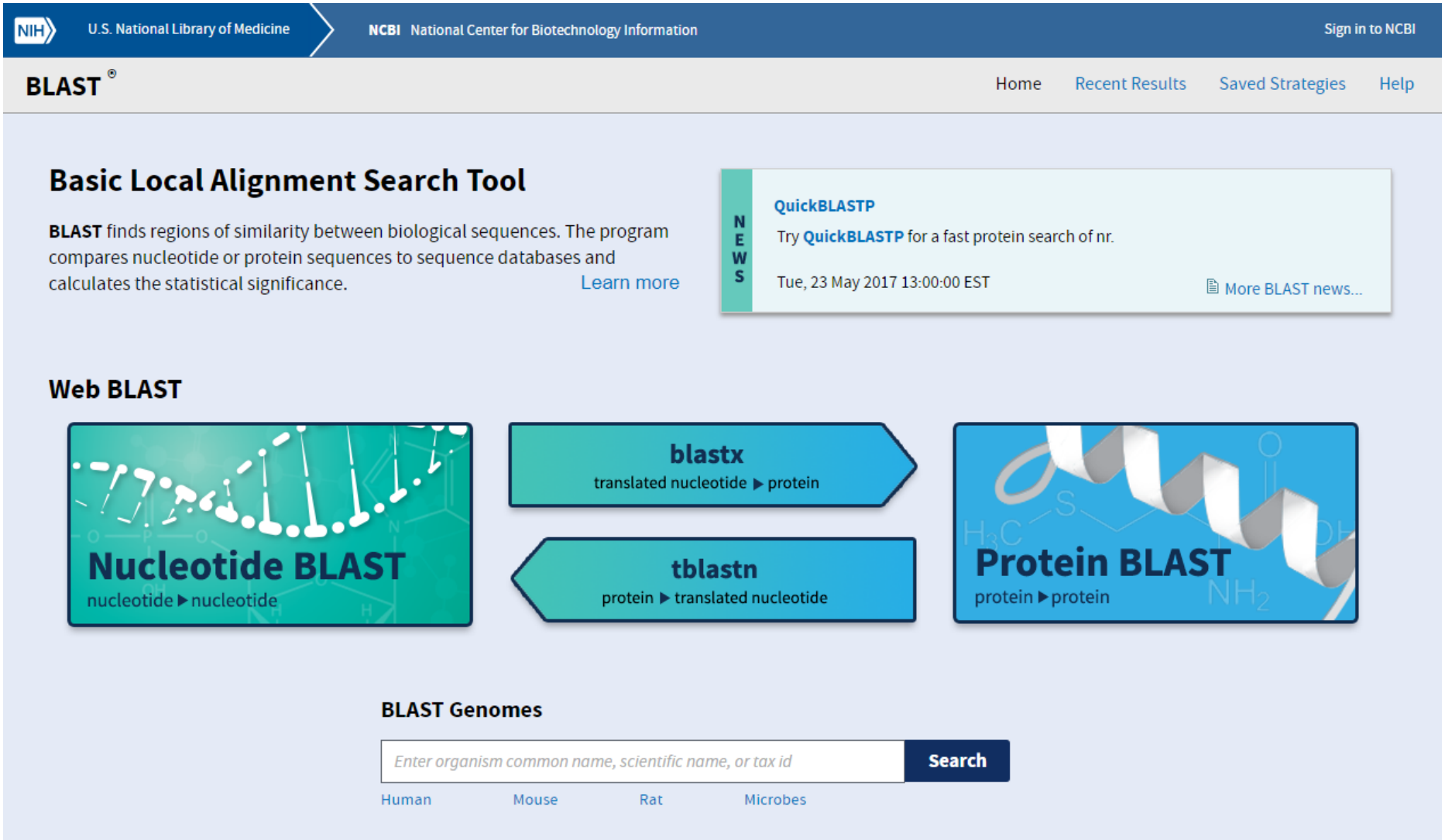
BLAST

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


NIH
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
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
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

Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Query Sequence

Choose Search Set

Database

Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Select Database

Organism
Optional

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
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Limit to
Optional

Sequences from type material

Entrez Query
Optional

YouTube 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)

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

Reset page | Bookmark

Enter Query Sequence

BLASTP programs search protein databases using a protein query. more...

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

Query Sequence

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): 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 (Optional): Models (XM/XP) | Uncultured/environmental sample sequences

Entrez Query (Optional): YouTube Create custom database

Enter an Entrez query to limit search

Select Database

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

Home Recent Results Saved Strategies Help

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 lcl|Query_108271
Description None
Molecule type nucleic acid
Query Length 204

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.6.1+ ▷Citation

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

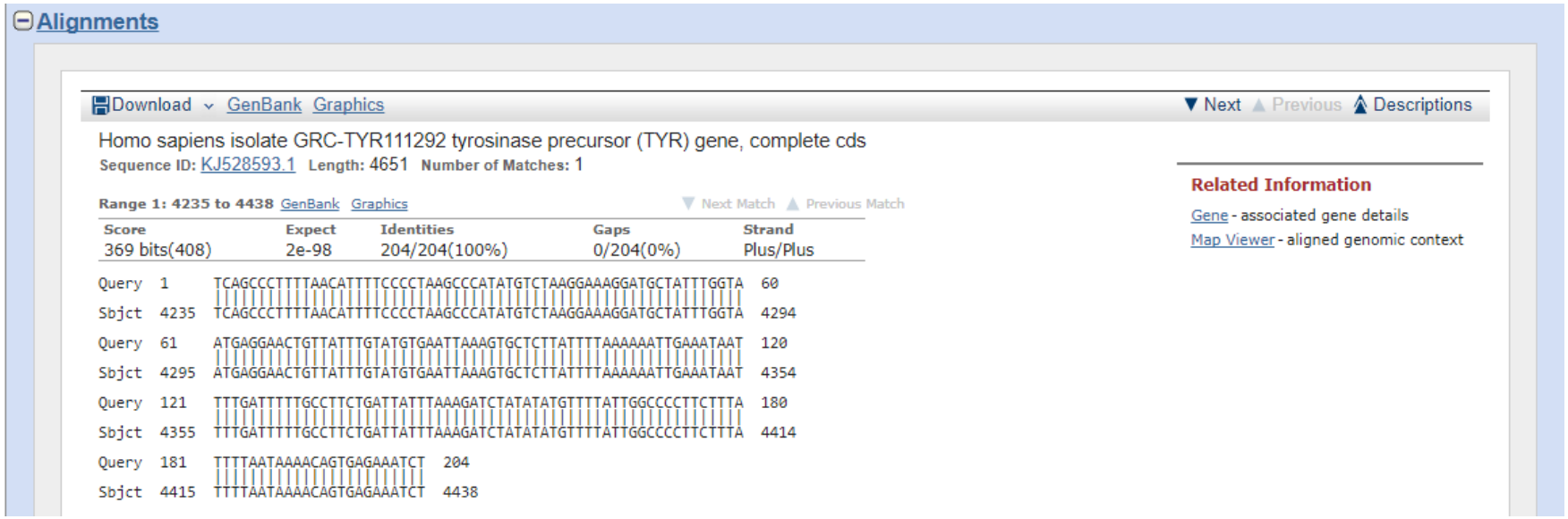
Alignments Download GenBank Graphics Distance tree of results

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



  - 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

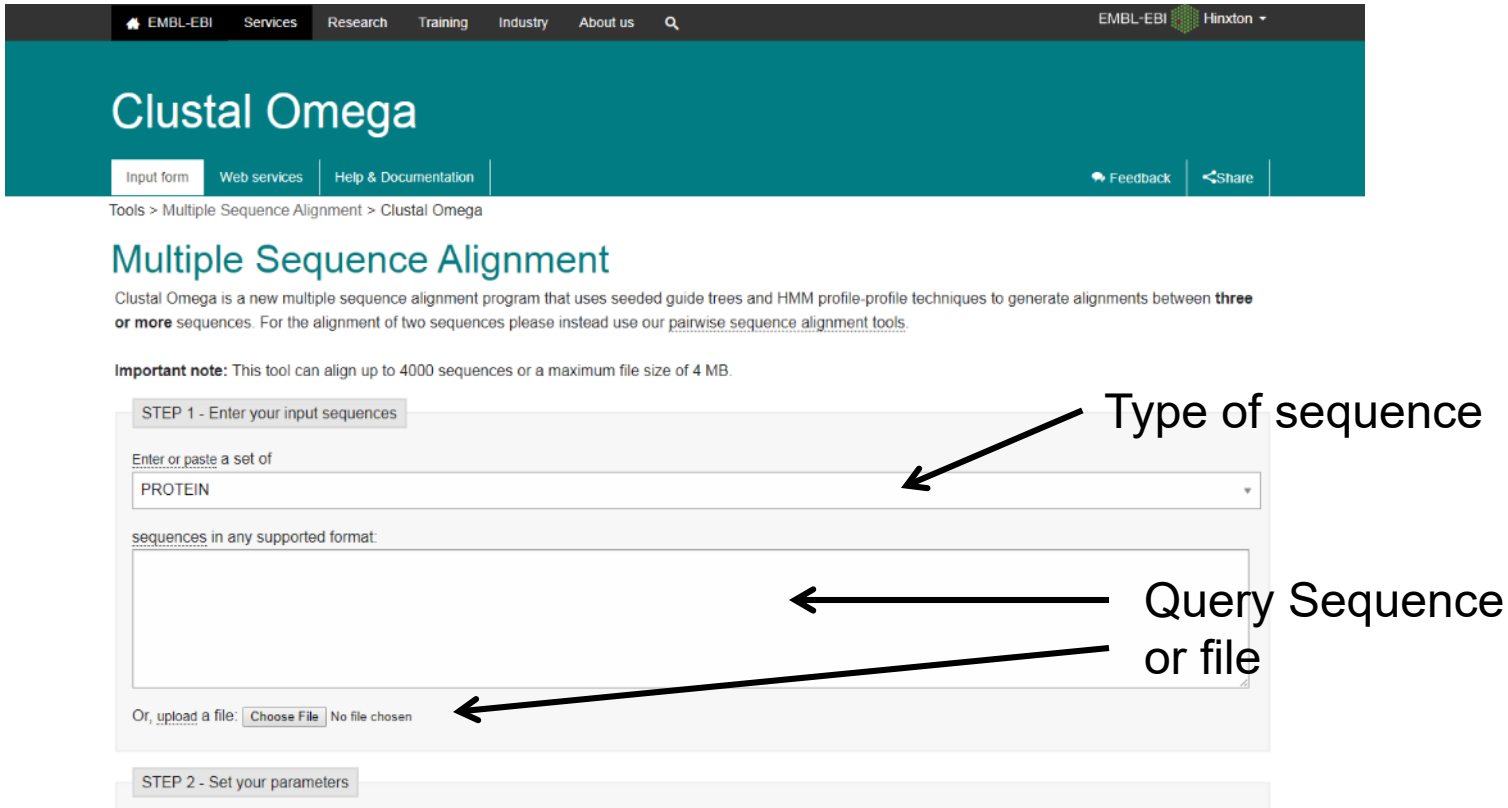
Cluster Analysis of Pairwise Alignments

CLUSTAL workflow

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

Clustal Omega

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



Clustal Omega Output

- Alignment

```
Borrelia        YKIVEIVSDGDYSIDEQIAVIEDDSGMRHNITMSFHWPVKVPITNYKERLIPSEPMLTQT
Candida         GTITSIAEAGSYNVEEPVLEVE-FDGKKHKYSMMHTWPVRVPRP-VAEKLTADHPLLTGQ
Saccharomyces   GTITWIAPAGEYTLDEKILEVE-FDGKKSDFTLYHTWPVRVPRP-VTEKLSADYPLLTGQ
Neurospora      GTITRIAEKGEYTVEEKILEVE-FDGKKTEYPMMQTWPVRVPRP-AAEKHSANQPFLVGQ
Trypanosoma     GRVTSIVPSGNYTLQDDIIELE-YNGTVKSLKLMHRWPVRTPRP-VASKESGNHPLLTGQ
Drosophila      GTVRYIAPSGNYKVDDVVLETE-FDGEITKHTMLQVWPVRHHAP-VTEKLPANHPLLTGQ
Acetabularia    GTVTYIAAPGNYTINEKIIEVE-FQGAKYEYSMKQSWPVRSPRP-VVEKLLADTPLLTGQ
Daucus          GKITYVAPAGQYSLKDTVLELE-FQGVKKQFTMLQTWPVRTPRP-VASKLAADTPLLTGQ
Sulfolobus      GTLKELAREGDYTVEDVVAVVD-MNGDEIPVKMYQKWPVRIPRP-YKEKLEPVEPLLTGI
Thermococcus    GEIVEIAEEGDYTVEEVIVKVKKPDGTIEELKMYHRWPVRVKRP-YKQKLPPEVPLITGQ
                  :  :.  *.*.:.: :   .  .*      :   ***:       .:     *::.
```

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

Clustal Omega Output

- Alignment

```
Borrelia        YKIVEIVSDGDYSIDEQIAVIEDDSGMRHNITMSFHWPVKVPITNYKERLIPSEPMLTQT
Candida         GTITSIAEAGSYNVEEPVLEVE-FDGKKHKYSMMHTWPVRVPRP-VAEKLTADHPLLTGQ
Saccharomyces   GTITWIAPAGEYTLDEKILEVE-FDGKKSDFTLYHTWPVRVPRP-VTEKLSADYPLLTGQ
Neurospora      GTITRIAEKGEYTVEEKILEVE-FDGKKTEYPMMQTWPVRVPRP-AAEKHSANQPFLVGQ
Trypanosoma     GRVTSIVPSGNYTLQDDIIELE-YNGTVKSLKLMHRWPVRTPRP-VASKESGNHPLLTGQ
Drosophila      GTVRYIAPSGNYKVDDVVLETE-FDGEITKHTMLQVWPVRHHAP-VTEKLPANHPLLTGQ
Acetabularia    GTVTYIAAPGNYTINEKIIEVE-FQGAKYEYSMKQSWPVRSPRP-VVEKLLADTPLLTGQ
Daucus          GKITYVAPAGQYSLKDTVLELE-FQGVKKQFTMLQTWPVRTPRP-VASKLAADTPLLTGQ
Sulfolobus      GTLKELAREGDYTVEDVVAVVD-MNGDEIPVKMYQKWPVRIPRP-YKEKLEPVEPLLTGI
Thermococcus    GEIVEIAEEGDYTVEEVIVKVKKPDGTIEELKMYHRWPVRVKRP-YKQKLPPEVPLITGQ
                  :  :.  *.*.:.: :   .  .*      :   ***:       .:     *::.
```

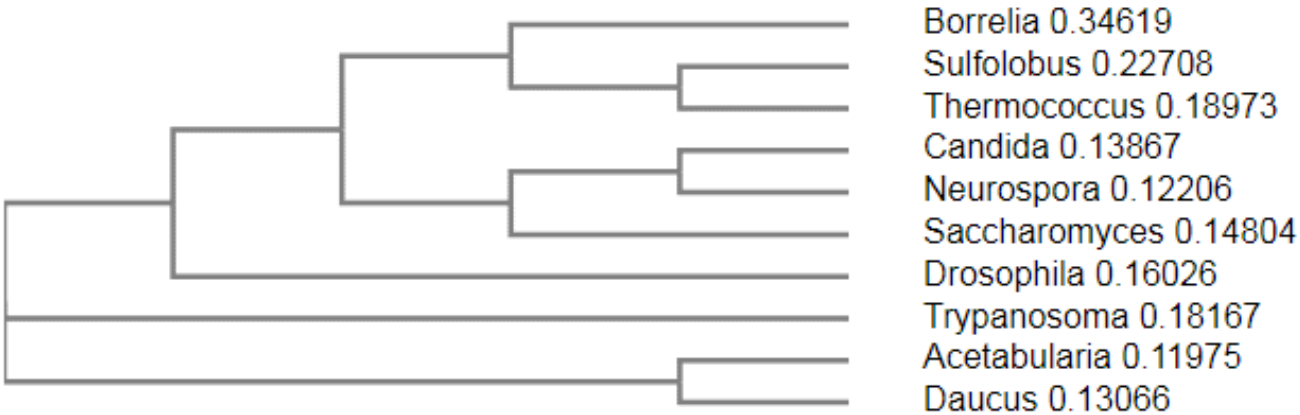
Color	Residues	Properties
Red	AVFPMILW	Small (small+ hydrophobic (incl. aromatic -Y))
Blue	DE	Acidic
Magenta	RK	Basic
Green	STYHCNGQ	Hydroxyl + sulfhydryl + amine + G
Grey	Others	Unusual amino acids

Clustal Omega Output

- Phylogenetic tree

Download Phylogenetic Tree Data

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)

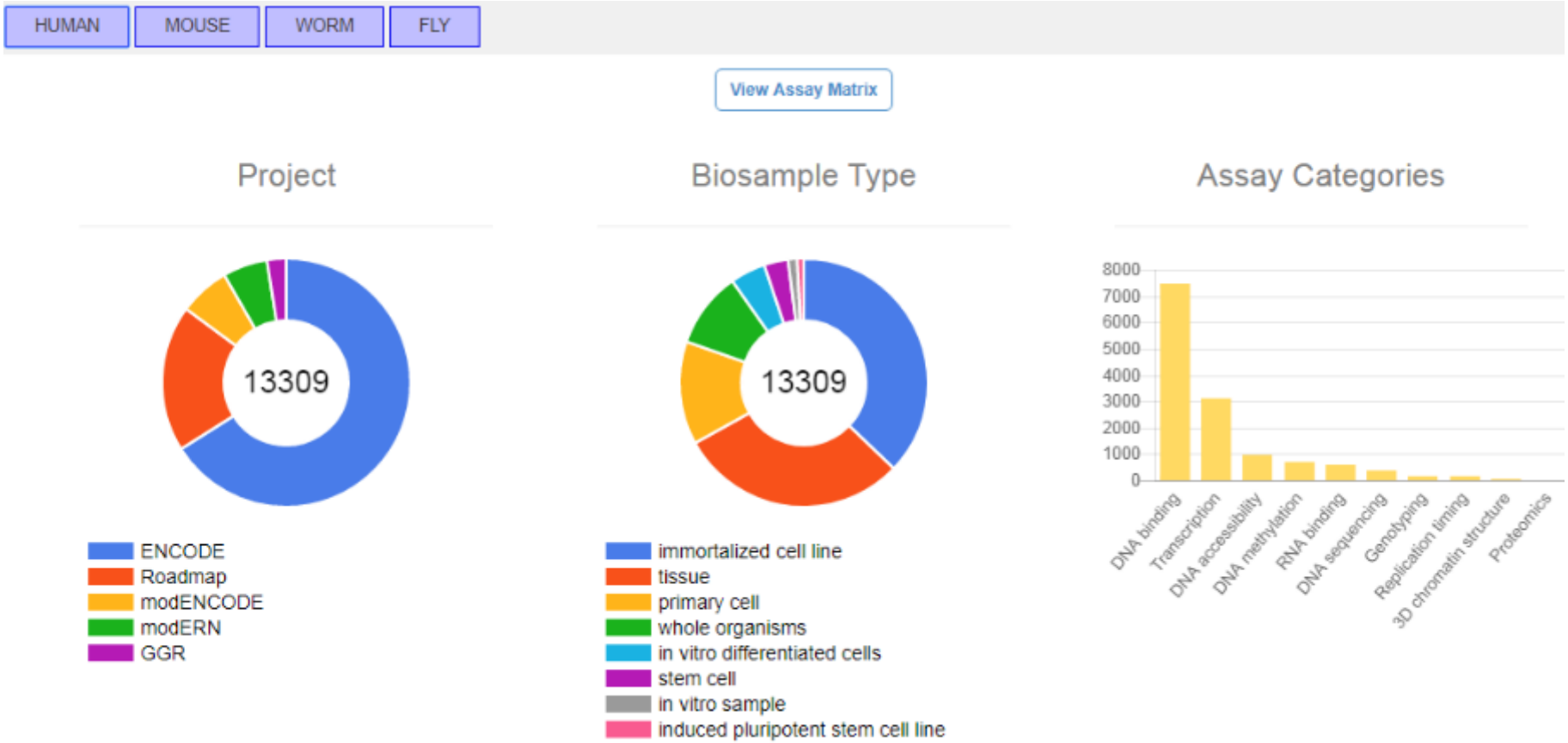
- Quantum leap of DNA sequencing technology
- Ability to sequence whole genomes fast and cheap
- Large repository of public data
  - ENCODE project
- Tools to browse genomes and view data
  - UCSC genome browser

ENCODE project

- Encyclopedia of DNA elements
  - https://www.encodeproject.org/

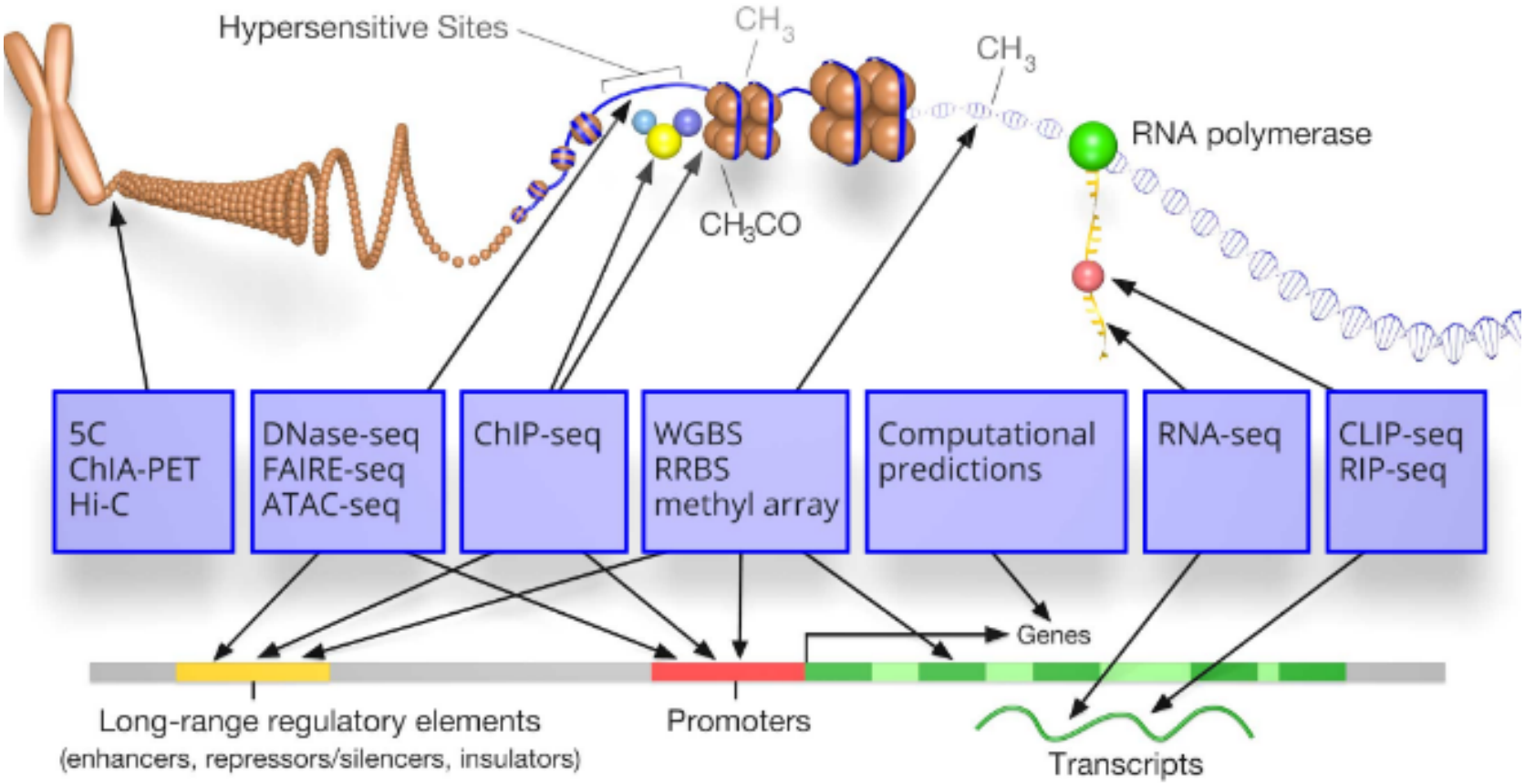
- Launched in 2003 to identify all functional elements in human genome
  - Extended to mouse, worm, fly
- Massive consortium led by NHGRI

ENCODE project: Datasets



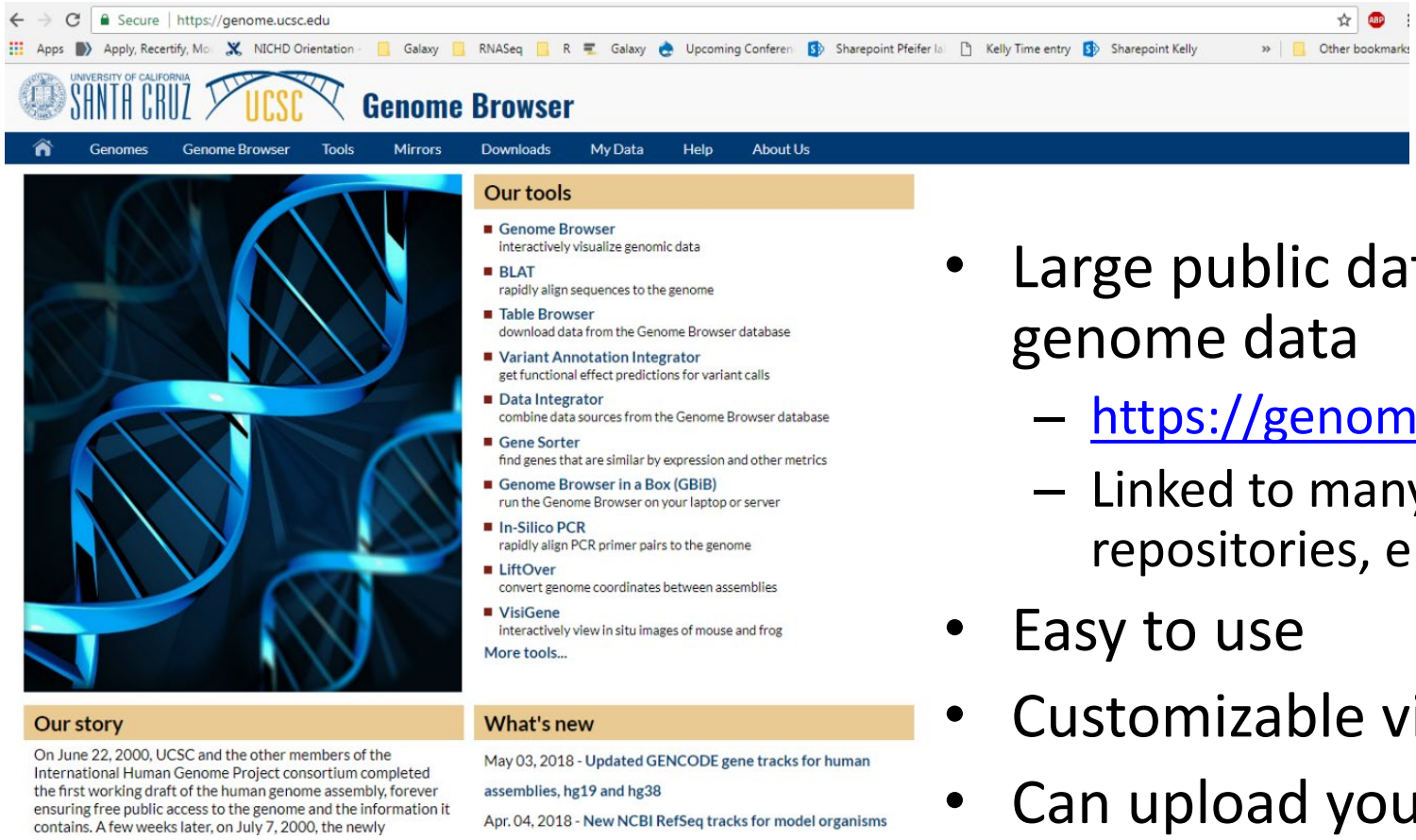
https://www.encodeproject.org/

ENCODE project: Assays



https://www.encodeproject.org/

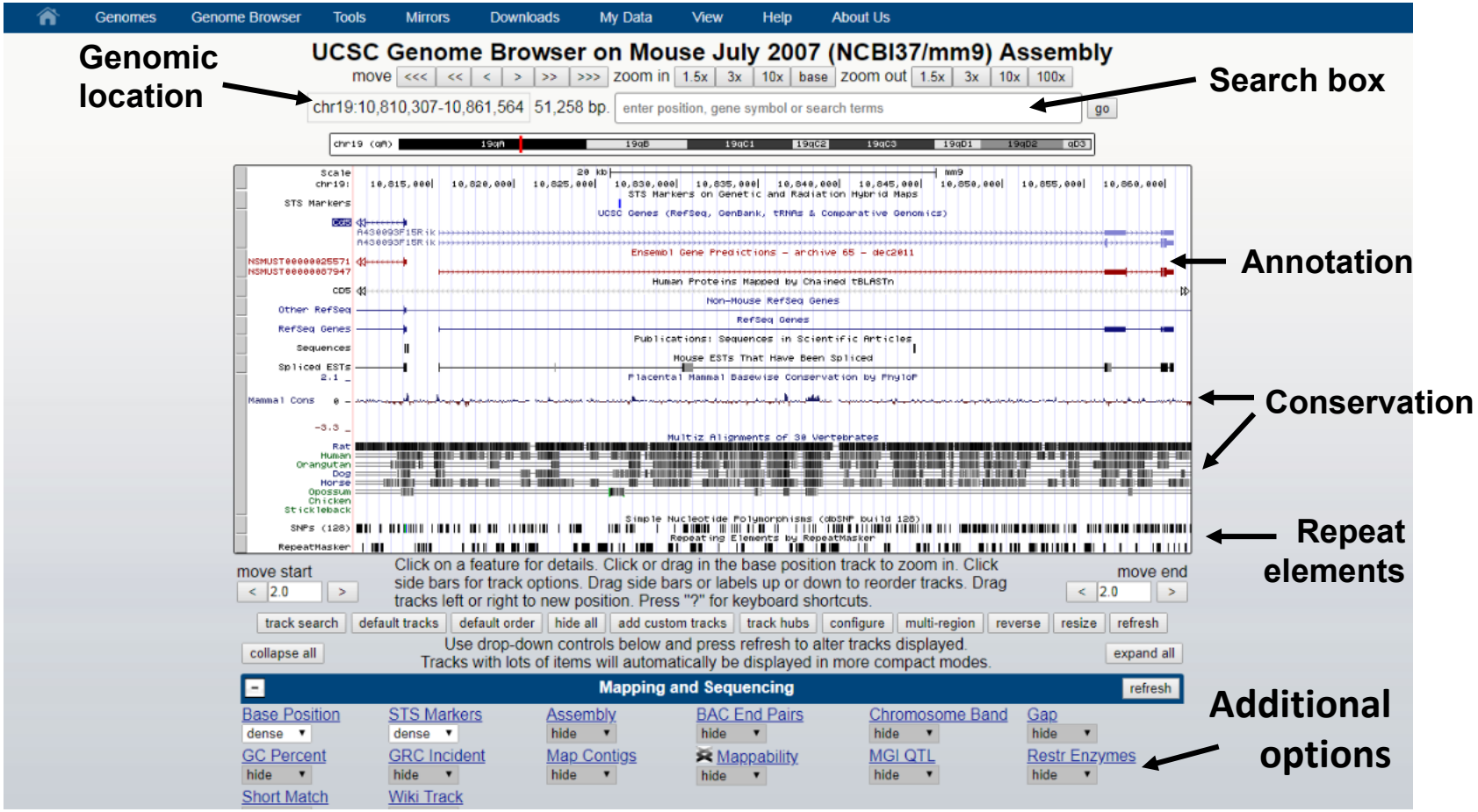
UCSC Genome Browser



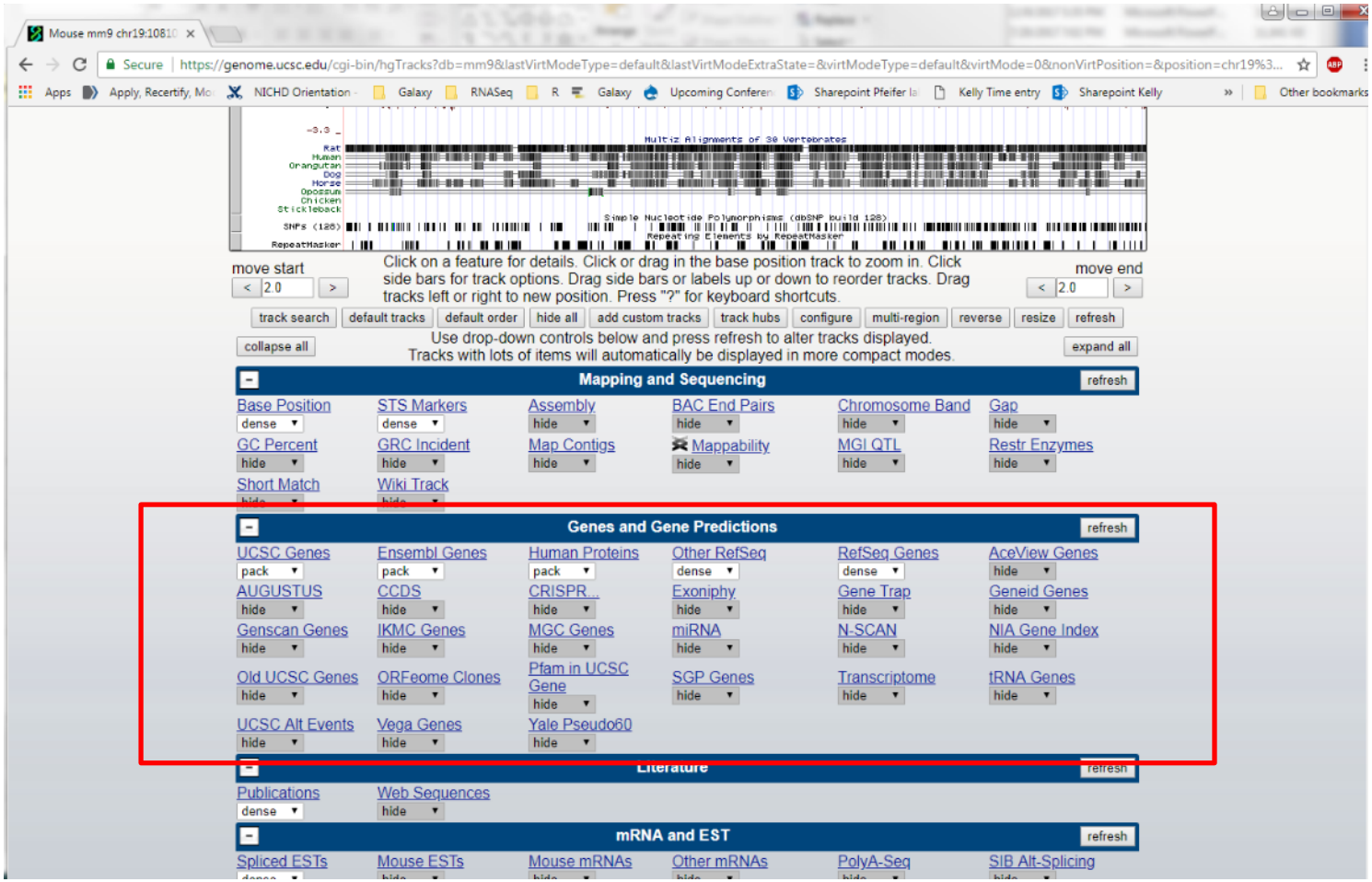
- Large public database of genome data
  - https://genome.ucsc.edu/
  - Linked to many public repositories, e.g. ENCODE
- Easy to use
- Customizable view
- Can upload your own data ('custom tracks')

UCSC Genome Browser

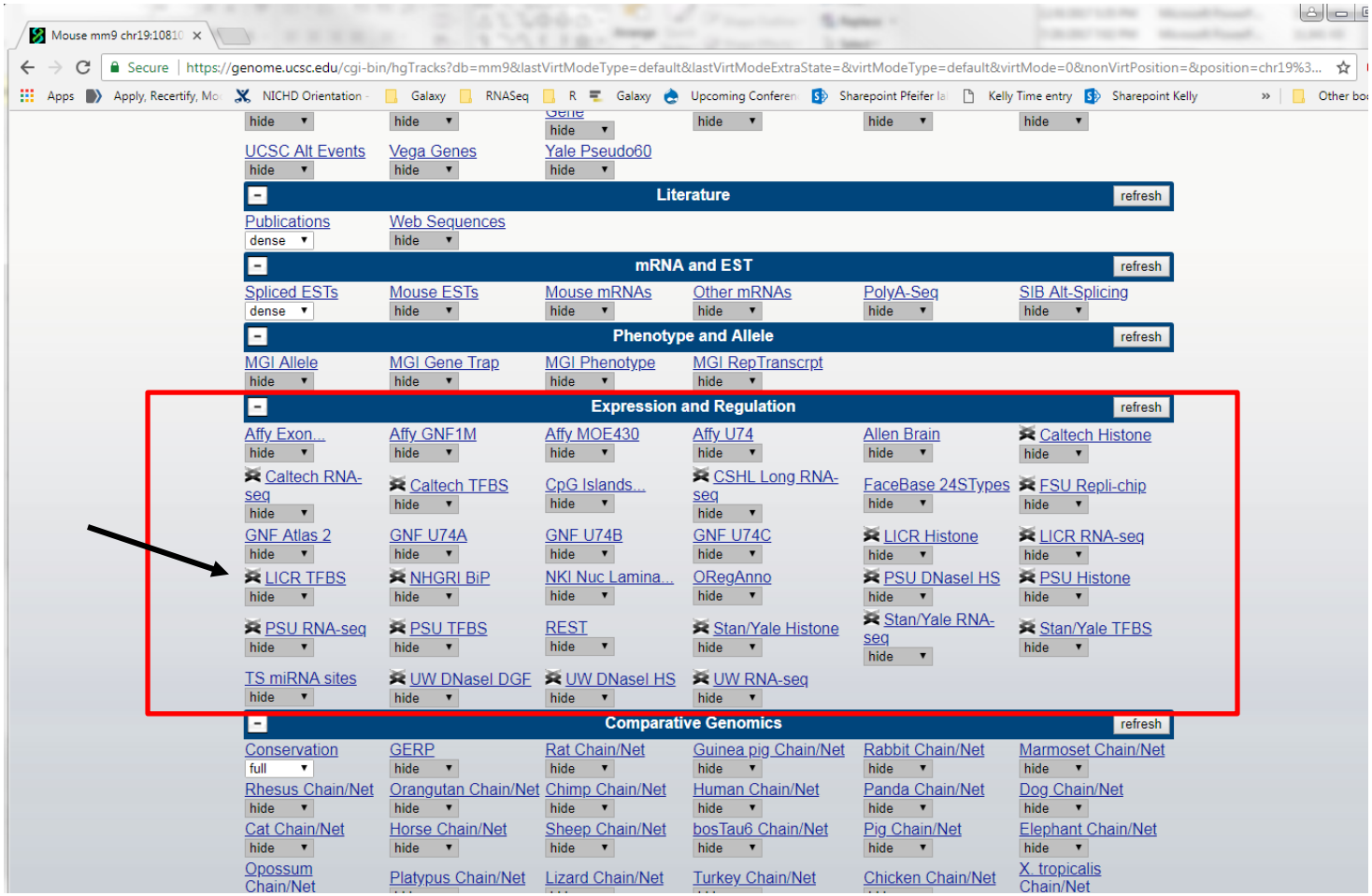
Mouse genome, mm9



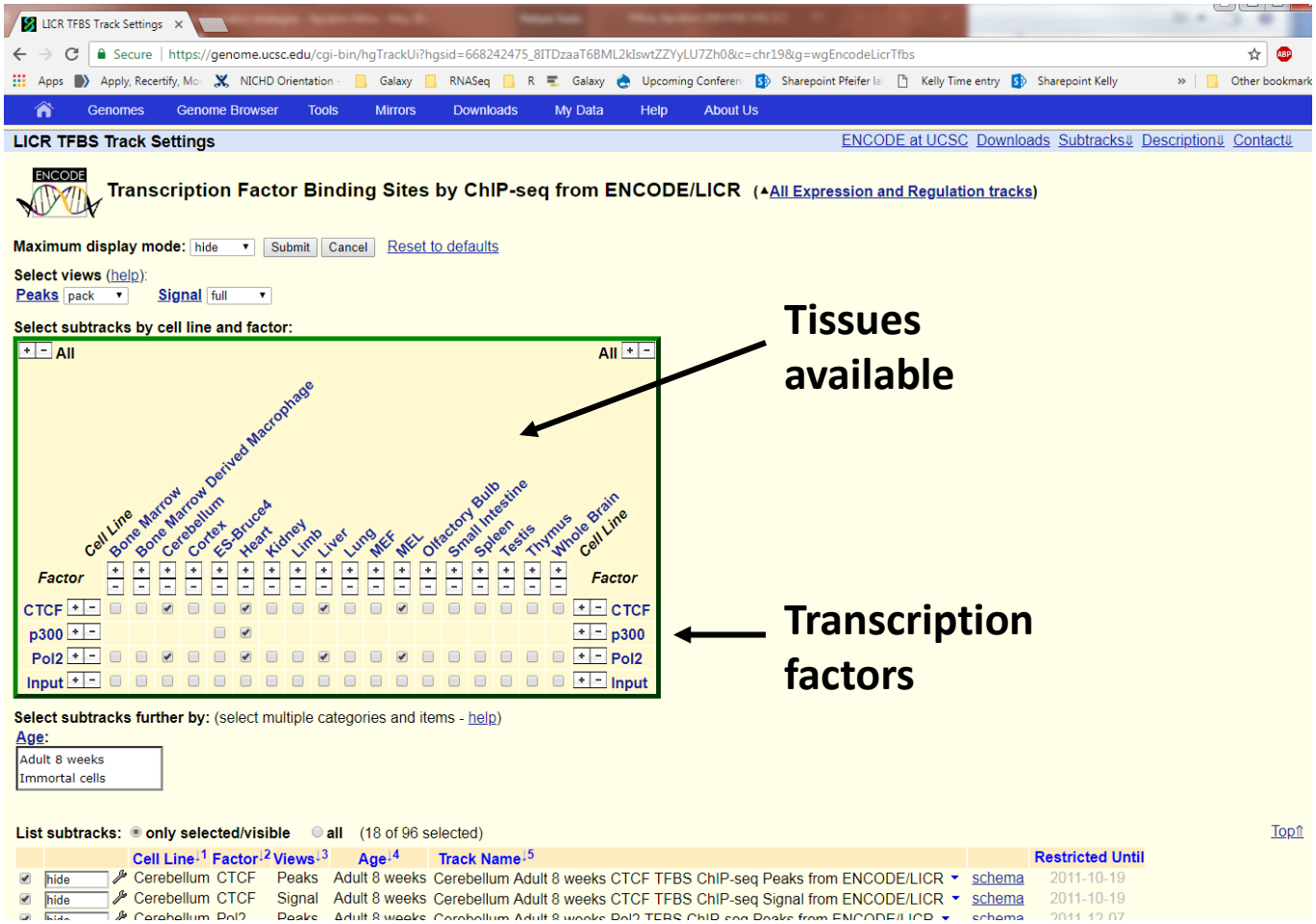
Annotation tracks



ENCODE @ UCSC

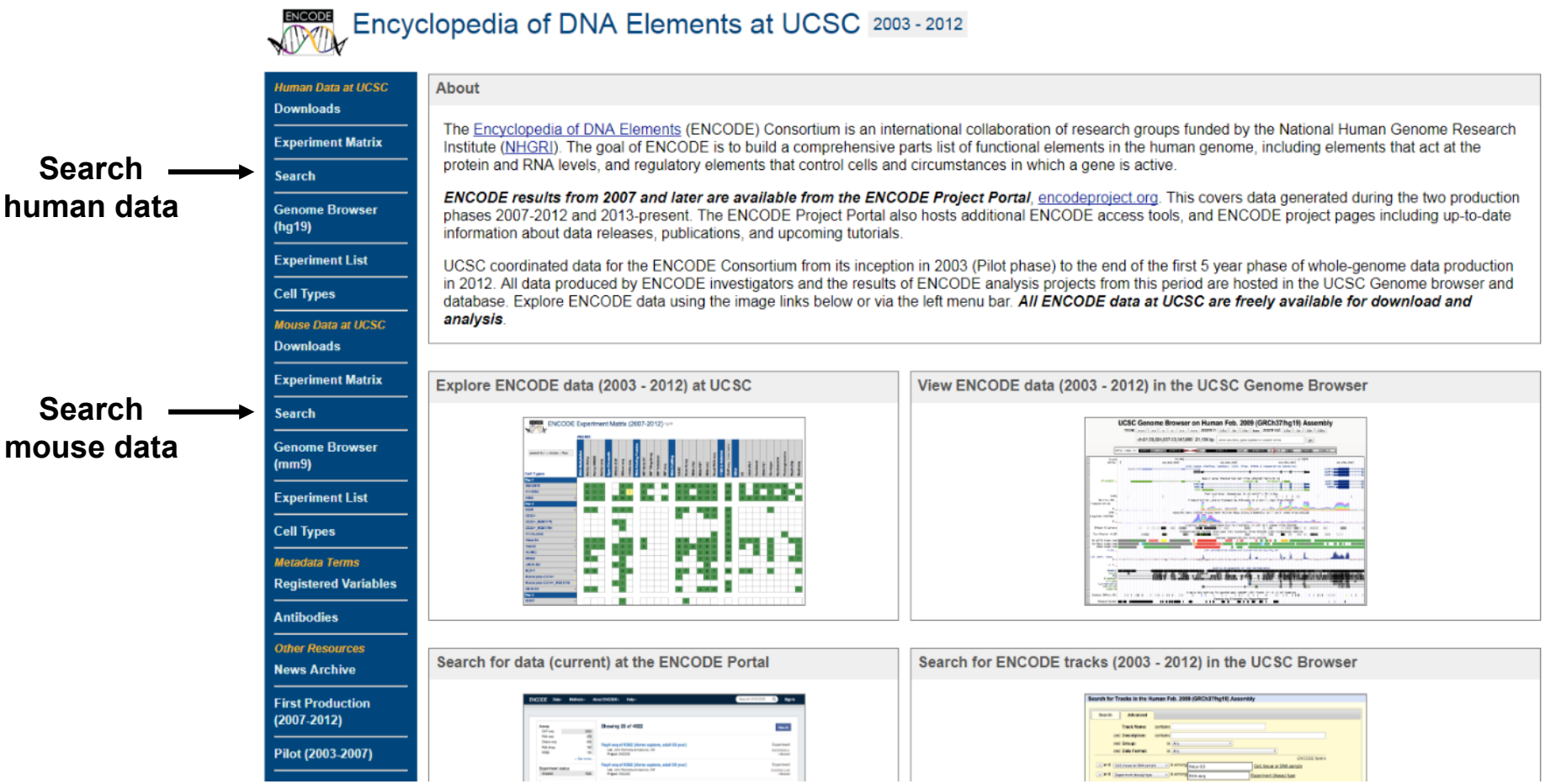

Mouse mm9 chr19:10810
Secure | https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr19%3...
Apps
Apply, Recertify, Mo
NICHD Orientation
Galaxy
RNASeq
R
Galaxy
Upcoming Conferen
Sharepoint Pfeifer la
Kelly Time entry
Sharepoint Kelly
Other bo
UCSC Alt Events
Vega Genes
Yale Pseudo60
Literature
refresh
Publications
Web Sequences
mRNA and EST
refresh
Spliced ESTs
Mouse ESTs
Mouse mRNAs
Other mRNAs
PolyA-Seq
SIB Alt-Splicing
Phenotype and Allele
refresh
MGI Allele
MGI Gene Trap
MGI Phenotype
MGI RepTranscrpt
Expression and Regulation
refresh
Affy Exon...
Affy GNF1M
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 DGF
UW DNaseI HS
UW RNA-seq
Comparative Genomics
refresh
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
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us
ENCODE at UCSC
Downloads
Subtracks
Description
Contact
Transcription Factor Binding Sites by ChIP-seq from ENCODE/LICR
All Expression and Regulation tracks
Maximum display mode:
Submit
Cancel
Reset to defaults
Select views (help):
Peaks
Signal
Select subtracks by cell line and factor:
Cell Line
Bone Marrow
Bone Marrow Derived Macrophage
Cerebellum
Cortex
ES-Bruce4
Heart
Kidney
Limb
Liver
Lung
MEF
MEL
Olfactory Bulb
Small Intestine
Spleen
Testis
Thymus
Whole Brain
Factor
CTCF
p300
Pol2
Input
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
Cerebellum CTCF Signal Adult 8 weeks Cerebellum Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
schema
2011-10-19
Tissues available
Transcription factors

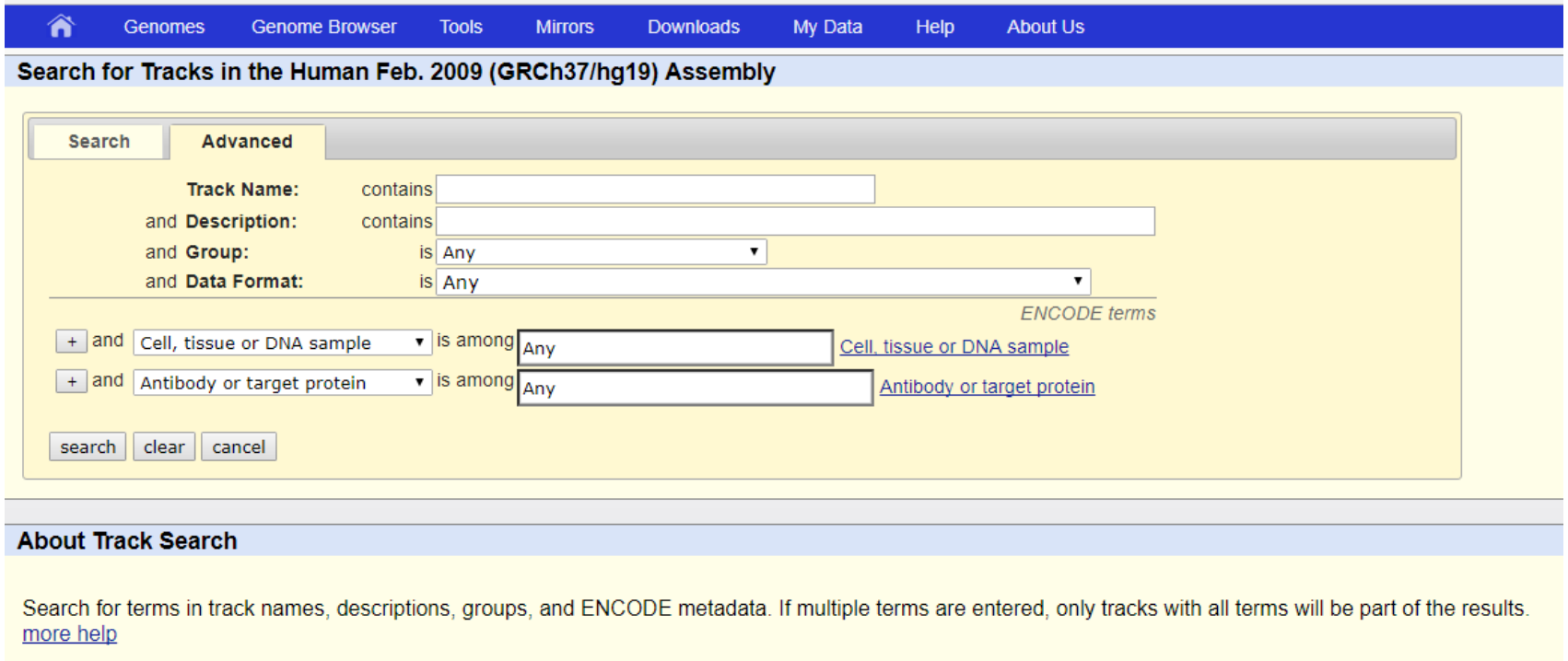
ENCODE @ UCSC

- Separate website with ENCODE data also available
  https://genome.ucsc.edu/ENCODE/



ENCODE @ UCSC

- Search full database for tracks to display



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