Bioinformatics: The Management and Analysis of Biological Data

Elliot J. Lefkowitz

'OMICS – MIC 741

January 29, 2016

Contact Information: Elliot Lefkowitz Professor, Microbiology

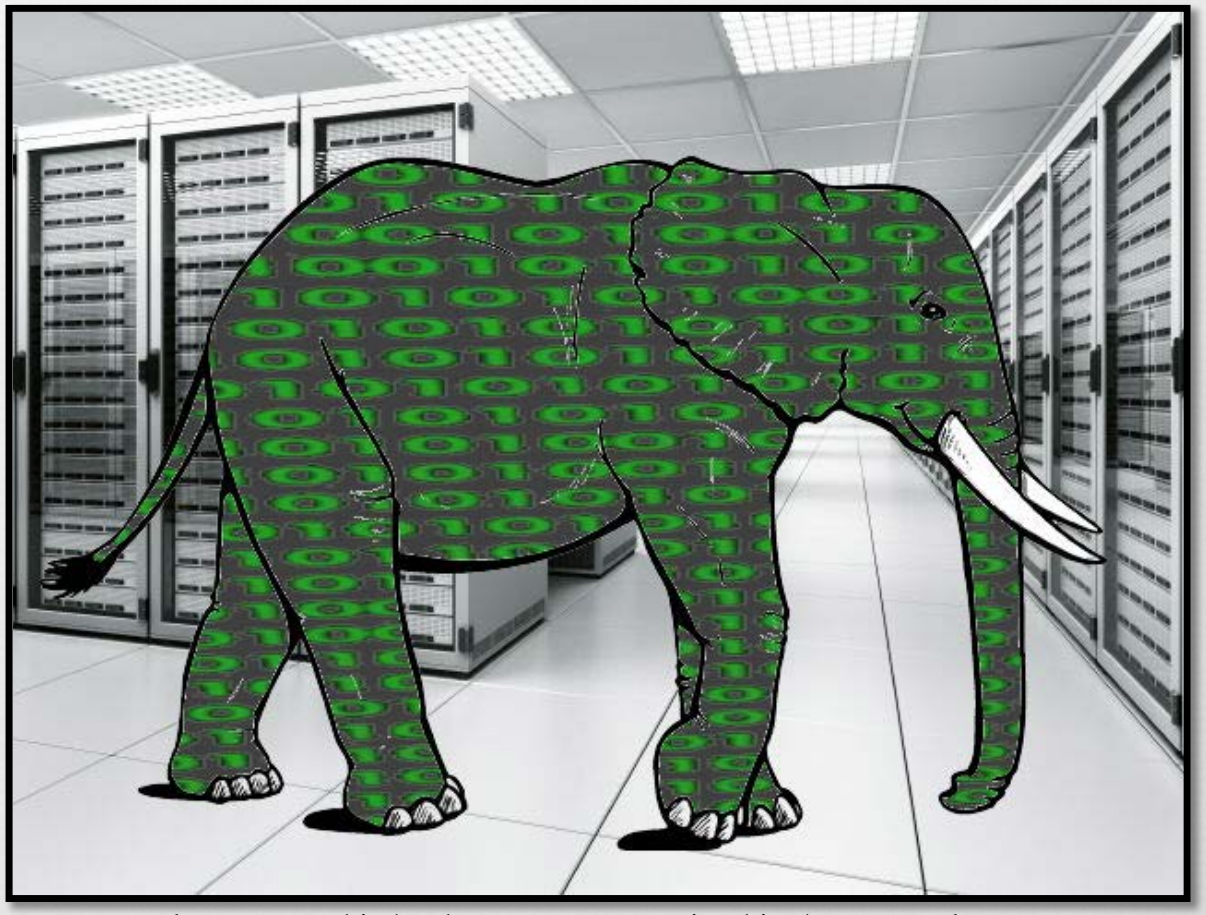
- Email
  - ElliotL@uab.edu
- Web Site
  - http://bioinformatics.uab.edu
- Office
  - BBRB 277A
- Phone
  - 934-1946

Objectives

- Defining and understanding the role of Bioinformatics in biological sciences
- Becoming familiar with the basic bioinformatic vocabulary
- Providing an overview of biomedical data and databases
- Providing an overview of biomedical analytical tools
- Learning how to discover, access, and utilize information resources

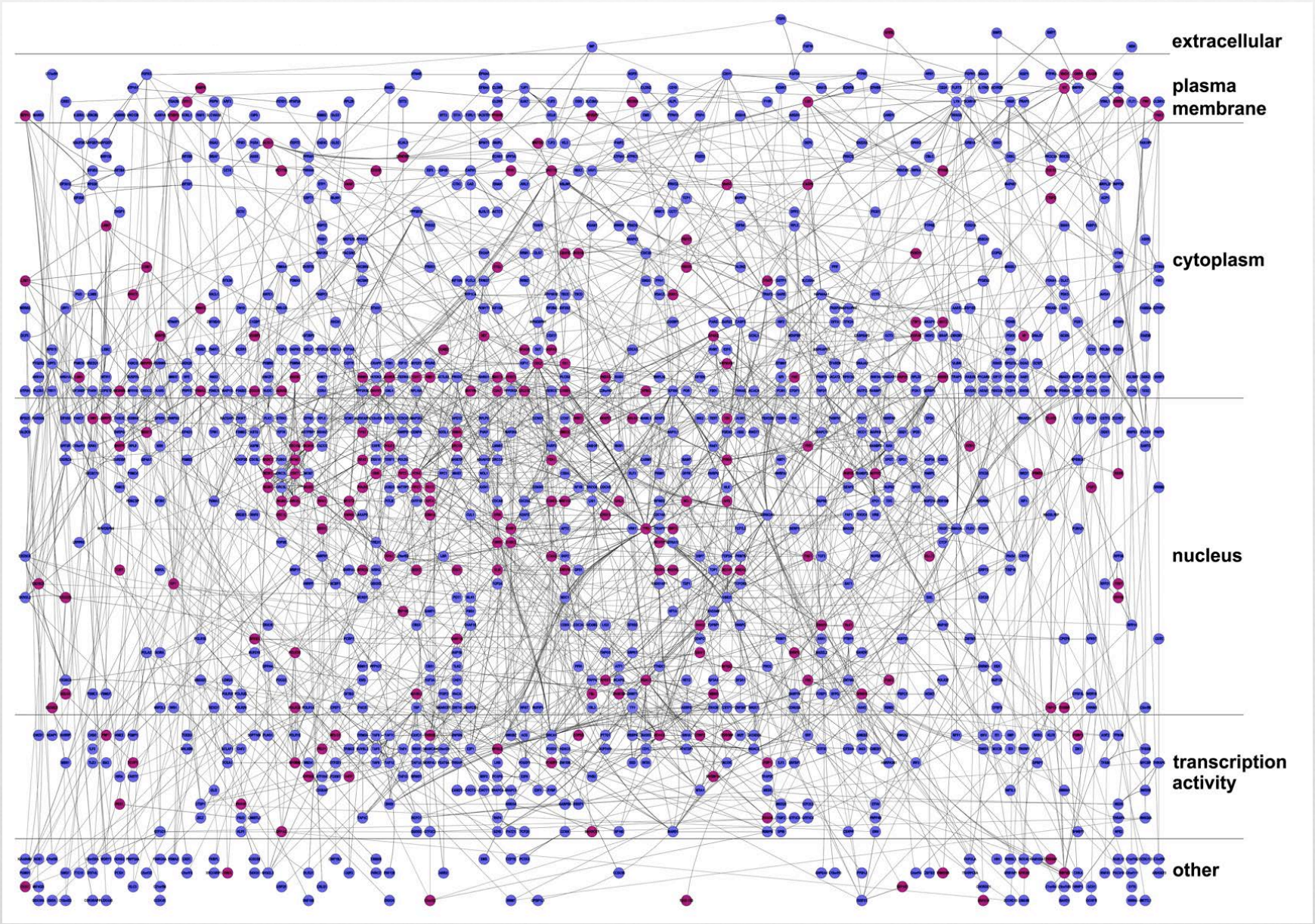
Biological Big Data

- Genomic
- Transcriptomic
- Epigenetic
- Proteomic
- Metabolomic
- Glycomic
- Lipidomic
- Imaging
- Health
- Integrated data sets



http://www.bigdatabytes.com/managing-big-data-starts-here/

Systems Biology



BMC Bioinformatics 2010, **11**:117

Bioinformatics

What is Bioinformatics?

- Computer-aided analysis of biological information

Bioinformatics = Pattern Discovery

- Discerning the characteristic (repeatable) patterns in biological information that help to explain the properties and interactions of biological systems.

Biological Patterns

- A pattern is a property of a biological system that can be defined such that the definition can be used to detect other occurrences of the same property
  - Sequence
  - Expression
  - Structure
  - Function
  - …

Bibliography: Web Sites

- Databases:
  - Nucleotide: www.ncbi.nlm.nih.gov
  - Protein: www.uniprot.org/
  - Protein structure: www.rcsb.org/pdb/
- Analysis Tools:
  - Protein analysis: us.expasy.org/
  - UAB Galaxy: www.uab.edu/galaxy
  - Penn State Galaxy: main.g2.bx.psu.edu/
  - Comparative genomics: ecrbrowser.dcode.org
- Genome Browsers:
  - NCBI: (www.ncbi.nlm.nih.gov/genomes)
  - Ensembl: (www.ensembl.org)
  - UCSC Genome Browser: (genome.ucsc.edu)

Bibliography: Journals

- Nucleic Acids Research
  - Database issue
    - Every January
  - Web Server issue
    - Every July
- Bioinformatics
- BMC Bioinformatics
- PLoS Computational Biology

Bioinformatics at UAB

- Center for Clinical and Translational Sciences (CCTS)
  - Biomedical Informatics
  - http://www.uab.edu/informatics
- Center for AIDS Research
  - Molecular and Genetic Bioinformatics Facility
  - http://www.genome.uab.edu/
- Heflin Center for Genetics
  - http://www.heflingenetics.uab.edu/
- Comprehensive Cancer Center's Bioinformatics Shared Facility
  - http://www.uab.edu/ccc/biostatistics/
- Section on Statistical Genetics
  - http://www.soph.uab.edu/ssg/
- UAB-IT Research Computing
  - http://www.uab.edu/cores/uab-it-research-computing-services/

Bioinformatics Services Provided by CCTS Informatics

- Provision of shared hardware supporting UAB's bioinformatics needs
- Access to software
  - Galaxy
  - Other supplemental software as needed
- Access to databases
  - Sequence, structure, motif, proteomic, expression…
- Education
  - Introductory lectures in core graduate courses
  - Bioinformatics module in upper-level courses
- Training
  - Assistance in the use of software and databases
  - Assistance in planning, organizing, and carrying out bioinformatics analyses
- Analytical assistance
- Research Collaborations

Major Information Resources

- NCBI
  - National Center for Biotechnology Information
  - Databases, tools, links
  - http://www.ncbi.nih.gov/
- EBI
  - European Bioinformatics Institute
  - Research Services Training
  - http://www.ebi.ac.uk/

National Centers for Biomedical Computing

- National Centers for Biomedical Computing
  - NIH program to enhance the computational infrastructure for biomedical computing
  - http://www.ncbcs.org/

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

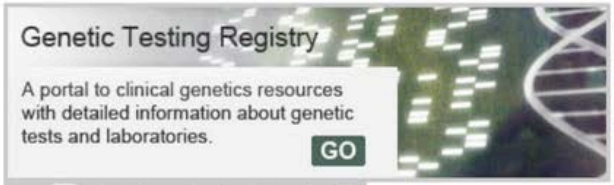
Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases



Popular Resources

NCBI Announcements

NCBI Eukaryotic Genome Annotation Pipeline breaks record; over 100 organisms annotated this year
Nov 20, 2014

NCBI BankIt webinar on December 17th
Nov 20, 2014
On December 17th, NCBI will have a webinar entitled "A Submitter's Guide to

NCBI E-Utilities webinar video now on YouTube
Nov 13, 2014
October's webinar, "An Introduction to

More...

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

USA.gov

Analytical Resources

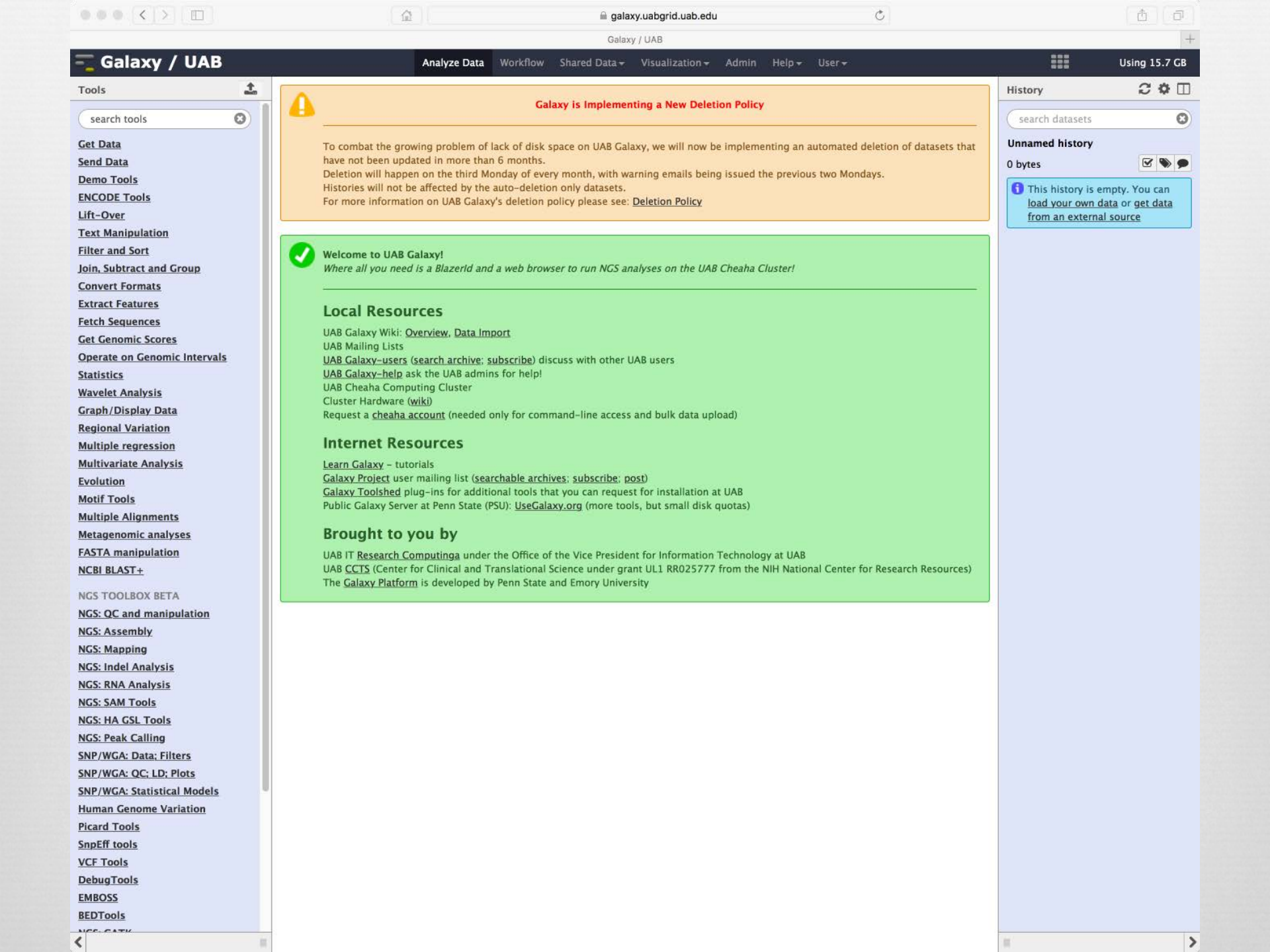
- Bioinformatics Packages
- Web-based tools
  - General Protein Analysis Tools
    - http://us.expasy.org/
  - Galaxy
    - https://usegalaxy.org/

Local Bioinformatics Packages

- Geneious
  - http://www.geneious.com
- MacVector
  - http://www.macvector.com
- Vector NTI
  - http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Cloning/vector-nti-software.html?CID=fl-vectornti
- DNAStar
  - http://www.dnastar.com
- Sequencher
  - http://genecodes.com

Galaxy

- Comprehensive set of bioinformatics tools
  - Basic sequence analysis (EMBOSS)
  - Next generation Sequencing (NGS) analysis
  - Statistical analysis
- http://www.uab.edu/galaxy
  - Requires BlazerID login
- http://main.g2.bx.psu.edu/


galaxy.uabgrid.uab.edu
Galaxy / UAB
Galaxy / UAB
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 15.7 GB
Tools
search tools
Get Data
Send Data
Demo Tools
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: HA GSL Tools
NGS: Peak Calling
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Human Genome Variation
Picard Tools
SnpEff tools
VCF Tools
DebugTools
EMBOSS
BEDTools
Galaxy is Implementing a New Deletion Policy
To combat the growing problem of lack of disk space on UAB Galaxy, we will now be implementing an automated deletion of datasets that have not been updated in more than 6 months.
Deletion will happen on the third Monday of every month, with warning emails being issued the previous two Mondays.
Histories will not be affected by the auto-deletion only datasets.
For more information on UAB Galaxy's deletion policy please see: Deletion Policy
Welcome to UAB Galaxy!
Where all you need is a BlazerId and a web browser to run NGS analyses on the UAB Cheaha Cluster!
Local Resources
UAB Galaxy Wiki: Overview, Data Import
UAB Mailing Lists
UAB Galaxy-users (search archive; subscribe) discuss with other UAB users
UAB Galaxy-help ask the UAB admins for help!
UAB Cheaha Computing Cluster
Cluster Hardware (wiki)
Request a cheaha account (needed only for command-line access and bulk data upload)
Internet Resources
Learn Galaxy - tutorials
Galaxy Project user mailing list (searchable archives; subscribe; post)
Galaxy Toolshed plug-ins for additional tools that you can request for installation at UAB
Public Galaxy Server at Penn State (PSU): UseGalaxy.org (more tools, but small disk quotas)
Brought to you by
UAB IT Research Computinga under the Office of the Vice President for Information Technology at UAB
UAB CCTS (Center for Clinical and Translational Science under grant UL1 RR025777 from the NIH National Center for Research Resources)
The Galaxy Platform is developed by Penn State and Emory University
History
search datasets
Unnamed history
0 bytes
This history is empty. You can load your own data or get data from an external source

ExPASy Bioinformatics Resource Portal

Visual Guidance

Categories

- proteomics
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

World-2DPAGE Repository

Public repository for gel-based proteomics data published in the literature
[details]

How to use this portal?

- Features and updates
- New to ExPASy
- Experienced ExPASy users: what is different

Popular resources

- UniProtKB
- SWISS-MODEL
- STRING
- PROSITE

Latest News

Protein Spotlight: The hidden things - 2015-01-08

Nature has its secret ways. During the course of the 19th century, the Augustinian friar Gregor Mendel worked out the basics of genetic inheritance as he crossbred pea plants. More

UniProt Knowledgebase release 2015_01 - 2015-01-07

Release notes
547,357 UniProtKB/Swiss-Prot entries (More.)
89,451,166 UniProtKB/TrEMBL entries (More).

[More news] [SIB news]

Bioinformatic Databases

Something to compare against

Major Sequence Databases

- International Nucleotide Sequence Database Collaboration
  - Genbank
  - EMBL
  - DDBJ
- Protein
  - UniProt
    - PIR
    - Swiss-Prot
    - Swiss-Prot TrEMBL
- NCBI

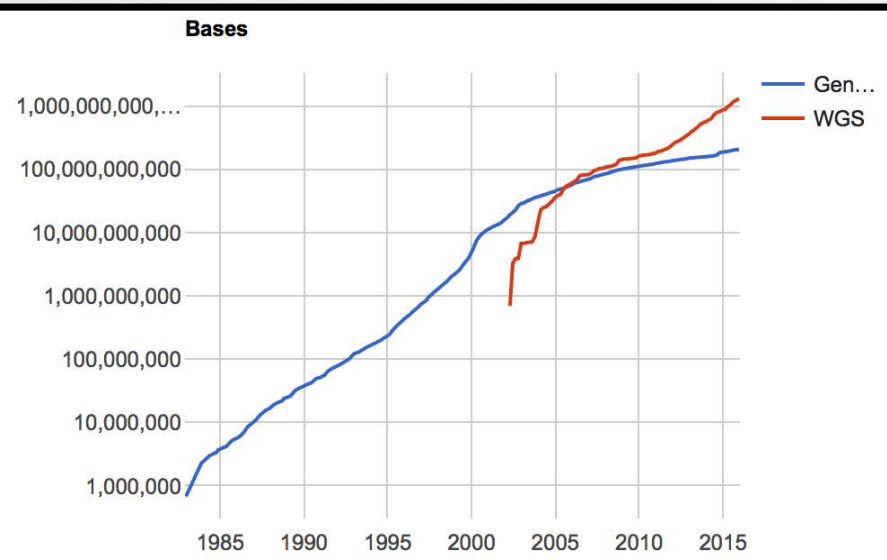
Other Databases

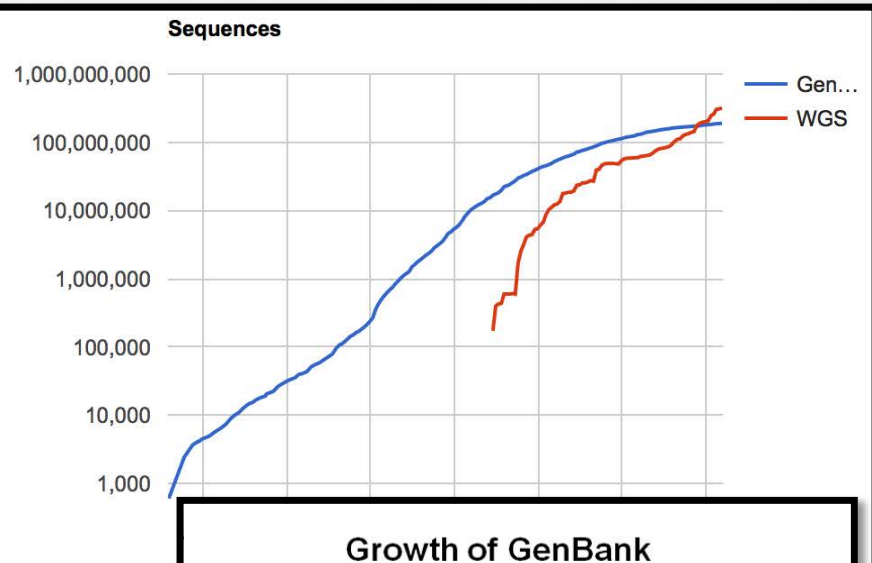
- Structural
  - Protein Data Bank (PDB): http://www.rcsb.org/pdb/
- Expression
  - Microarray Gene Expression Data Society (MGED): http://www.mged.org/
  - Gene Expression Omnibus (GEO – NCBI): http://www.ncbi.nlm.nih.gov/geo/
- Proteomic
  - Mascot: http://www.matrixscience.com/
- Metabolism
  - BioCyc: http://biocyc.org/
  - Reactome: http://reactome.org/
- Ontology
  - Gene Ontology (GO) Consortium: http://www.geneontology.org/
  - Controlled vocabulary for the description of biological processes

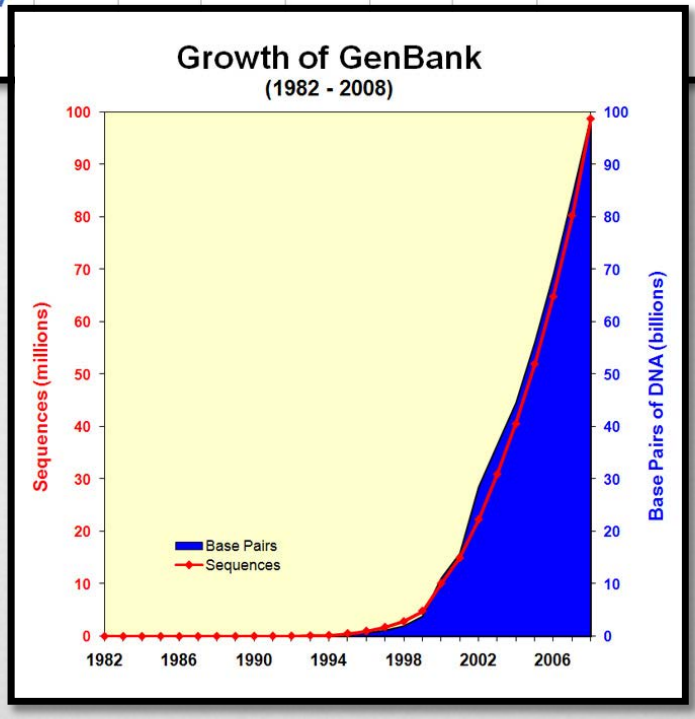
GenBank

- Primary nucleic acid sequence database
- Maintained by NCBI
  - National Center for Biotechnology Information
  - http://www.ncbi.nlm.nih.gov/genbank/
  - December, 2016; Release 211
  - 203,939,111,071 bases
  - 189,232,925 sequences
- WGS (Whole Genome Shotgun) Sequences
  - Separate from GenBank
    - 1,297,865,618,365 Bases
    - 317,122,157 Sequences

GenBank Growth







NCBI Databases for NGS Data

- Bioproject
  - Genomics, functional genomics, and genetics studies and links to their datasets
- Biosample
  - Descriptions of biological source materials used in experimental assays.
- Sequence Read Archive (SRA)
  - Raw sequence data
- Assembly archive

Other NCBI Databases

- RefSeq
- HomoloGene
- Genomic
- SNPs
- …

RefSeq

- NCBI Reference Sequence project
- Provides reference sequence standards for the naturally occurring molecules from chromosomes to mRNAs to proteins
- Stable reference point for:
  - mutation analysis
  - gene expression studies
  - polymorphism discovery
- Accession numbers have two letters, an underscore, and six numbers
  - NM_123456

HomoloGene

- Database of homologs among the annotated genes of several completely sequenced eukaryotic geneomes

Release 68 Statistics

Release date:2014-04-09
Total species:21
Total HomoloGene groups:44233

Download

Scientific name	Common name	Genome release	Annotation release	Input genes	Grouped genes	Groups
Homo sapiens	human	GRCh38	106	19806	19129	18732
Pan troglodytes	chimpanzee	Pan_troglodytes-2.1.4	102	22004	18730	18385
Macaca mulatta	Rhesus monkey	Mmul_051212	Build 1.2	22231	16843	16517
Canis lupus familiaris	dog	CanFam3.1	103	19872	18117	17434
Bos taurus	cow	Bos_taurus_UMD_3.1	103	21025	18798	17530
Mus musculus	mouse	GRCm38.p2	104	22627	21207	19033
Rattus norvegicus	rat	Rnor_5.0	104	22892	20616	18871
Gallus gallus	chicken	Gallus_gallus-4.0	102	17133	14600	13352
Xenopus tropicalis	western clawed frog	Xtropicalis_v7	101	22156	18447	14412
Danio rerio	zebrafish	Zv9	103	27270	20897	14559
Drosophila melanogaster	fruit fly	Release 5	Release 5.48	13944	8438	7016
Anopheles gambiae	malaria mosquito	AgamP3	AgamP3.3	12600	8428	6957
Caenorhabditis elegans	nematode	WS195	WS195	20519	7575	4308
Saccharomyces cerevisiae	budding yeast	R64-1-1	NA	5903	4579	4113
Kluyveromyces lactis	ascomycetes	ASM251v1	NA	5084	4283	4211
Eremothecium gossypii	ascomycetes	ASM9102v4	NA	4768	3874	3820
Schizosaccharomyces pombe	fission yeast	ASM294v2	NA	5132	3018	2741
Magnaporthe oryzae	rice blast fungus	MG8	NA	12673	6598	6061
Neurospora crassa	ascomycetes	ASM18292v1	NA	9821	5807	5701
Arabidopsis thaliana	thale cress	TAIR10	NA	27393	19143	10463
Oryza sativa	rice	Build 4.0	NA	28521	16112	9787

Multi-Genome Comparison

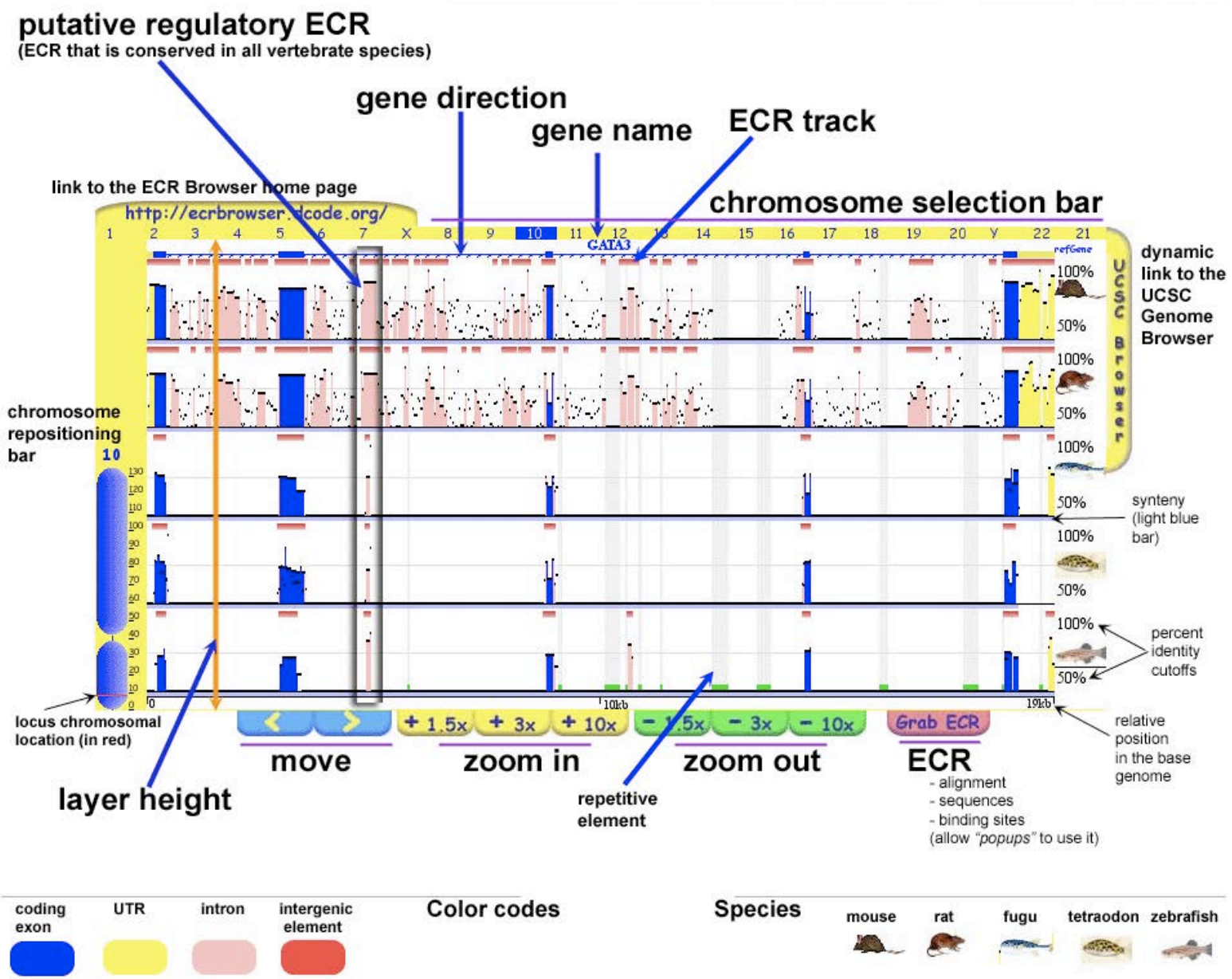
- Evolutionarily Conserved Regions
  - Coding Regions
  - Regulatory Regions
- ECR Browser
  - http://ecrbrowser.dcode.org

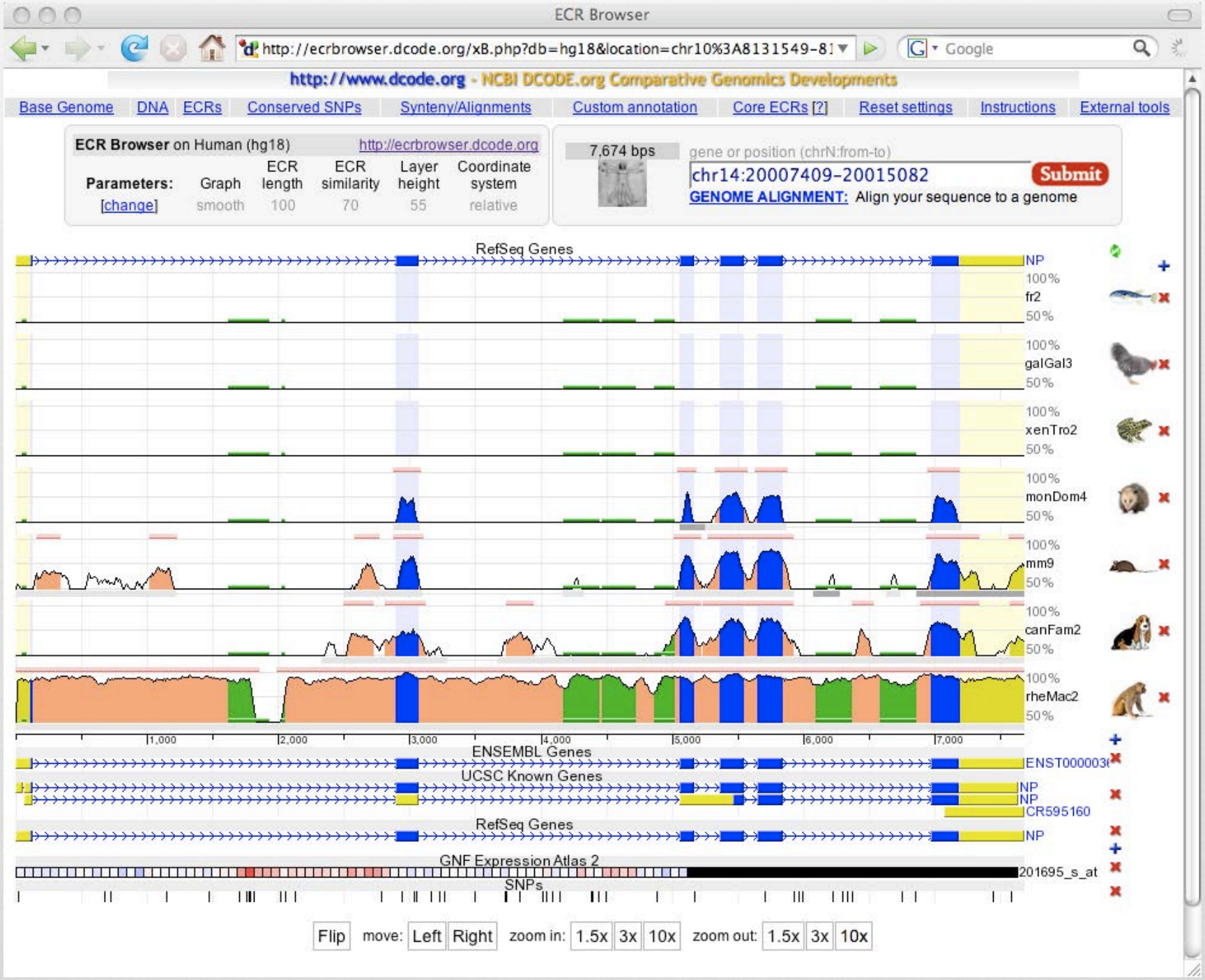
Mozilla Firefox

http://ecrbrowser.dcode.org/ecrLegend.html

Google

ECR Browser (http://ecrbrowser.dcode.org/) legend




ECR Browser
http://ecrbrowser.dcode.org/xB.php?db=hg18&location=chr10%3A8131549-81
Google
http://www.dcode.org - NCBI DCODE.org Comparative Genomics Developments
Base Genome
DNA
ECRs
Conserved SNPs
Synteny/Alignments
Custom annotation
Core ECRs [?]
Reset settings
Instructions
External tools
ECR Browser on Human (hg18)
http://ecrbrowser.dcode.org
Parameters:
Graph
ECR length
ECR similarity
Layer height
Coordinate system
[change]
smooth
100
70
55
relative
7,674 bps
gene or position (chrN:from-to)
chr14:20007409-20015082
Submit
GENOME ALIGNMENT: Align your sequence to a genome
RefSeq Genes
NP
100%
fr2
50%
galGal3
xenTro2
monDom4
mm9
canFam2
rheMac2
1,000
2,000
3,000
4,000
5,000
6,000
7,000
ENSEMBL Genes
ENST000003
UCSC Known Genes
NP
NP
CR595160
RefSeq Genes
NP
GNF Expression Atlas 2
201695_s_at
SNPs
Flip
move:
Left
Right
zoom in:
1.5x
3x
10x
zoom out:
1.5x
3x
10x

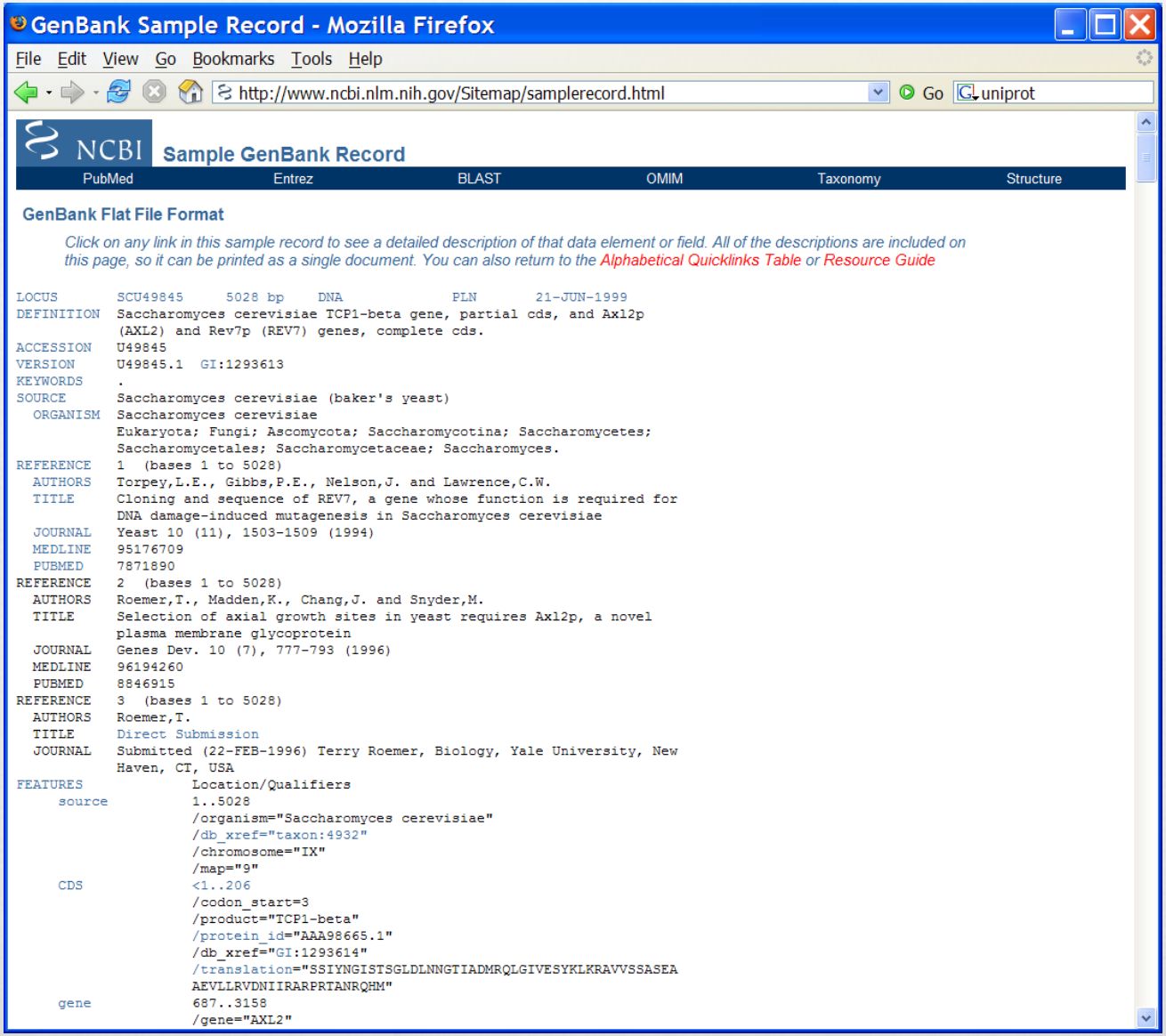
Sequences

The Sequence Record

- Different for each database
- Locus (Name)
- Accession Number
- Keywords
- Description
- Properties
- References
- The Sequence

GenBank Sample Record

- http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html



GenBank Record Human CFTR

```
LOCUS       HUMCFTRM                6129 bp    mRNA    linear   PRI 27-APR-1993
DEFINITION  Human cystic fibrosis mRNA, encoding a presumed transmembrane
            conductance regulator (CFTR).
ACCESSION   M28668
VERSION     M28668.1  GI:180331
KEYWORDS    cystic fibrosis; transmembrane conductance regulator.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 6129)
  AUTHORS   Riordan,J.R., Rommens,J.M., Kerem,B., Alon,N., Rozmahel,R.,
Grzelczak,Z., Zielenski,J., Lok,S., Plavsic,N., Chou,J.-L.,
Drumm,M.L., Iannuzzi,M.C., Collins,F.S. and Tsui,L.-C.
  TITLE     Identification of the cystic fibrosis gene: cloning and
            characterization of complementary DNA
  JOURNAL   Science 245 (4922), 1066-1073 (1989)
   PUBMED   2475911
```

Accession Numbers

- Each sequence submitted to a database is assigned a unique primary accession number
- Accession numbers do not change
- If a sequence is merged with another, a new accession number is assigned, and the original number becomes a secondary accession number
- Accession numbers may include version numbers
  - AO2428.2

GI numbers

- GenInfo Identifier
- Series of digits that are assigned consecutively to each sequence record processed by NCBI
- The GI number bears no resemblance to the Accession number of the sequence record
- Used to track sequence histories in GenBank and the other sequence databases
  - If a sequence is changed/updated, a new gi number is assigned

```
COMMENT     A three base-pair deletion spanning positions 1654-1656 is observed
            in cDNAs from cystic fibrosis patients.
FEATURES             Location/Qualifiers
     source          1. .6129
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     CDS             133. .4575
                     /note="cystic fibrosis transmembrane conductance
                     regulator"
                     /codon_start=1
                     /db_xref="PID:g180332"
                     /translation="MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVD
                     SADNLSEKLEREWDRELASKKNPKLINALRRCFFWRFMFYGIFLYLGEVTKAVQPLLL
                     LNRFSKDIAILDDLLPLTIFDFIQLLLIVIGAIAVVAVLQPYIFVATVPVIVAFIMLR
                     AYFLQTSQQLKQLESEGRSPIFTHLVTSLKGLWTLRAFGRQPYFETLFHKALNLHTAN
                     WFLYLSTLRWFQMRIEMIFVIFFIAVTFISILTTGEGEGRVGIILTLAMNIMSTLQWA
                     VNSSIDVDSLMRSVSRVFKFIDMPTEGKPTKSTKPYKNGQLSKVMIIENSHVKKDDIW
                     PSGGQMTVKDLTAKYTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLLN
                     TEGEIQIDGVSWDSITLQQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVAD
                     EVGLRSVIEQFPGKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDP
                     VTYQIIRRTLKQAFADCTVILCEHRIEAMLECQQFLVIEENKVRQYDSIQKLLNERSL
                     FRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL"
BASE COUNT     1886 a   1181 c   1330 g   1732 t
ORIGIN
```

```
HUMCFTRM  Length: 6129  April 13, 1998 13:00  Type: N  Check: 6781  ..

     1  AATTGGAAGC AAATGACATC ACAGCAGGTC AGAGAAAAAG GGTTGAGCGG

    51  CAGGCACCCA GAGTAGTAGG TCTTTGGCAT TAGGAGCTTG AGCCCAGACG

   101  GCCCTAGCAG GGACCCCAGC GCCCGAGAGA CCATGCAGAG GTCGCCTCTG

   151  GAAAAGGCCA GCGTTGTCTC CAAACTTTTT TTCAGCTGGA CCAGACCAAT

   201  TTTGAGGAAA GGATACAGAC AGCGCCTGGA ATTGTCAGAC ATATACCAAA

   251  TCCCTTCTGT TGATTCTGCT GACAATCTAT CTGAAAAATT GGAAAGAGAA

   301  TGGGATAGAG AGCTGGCTTC AAAGAAAAAT CCTAAACTCA TTAATGCCCT

   351  TCGGCGATGT TTTTTCTGGA GATTTATGTT CTATGGAATC TTTTTATATT

   401  TAGGGGAAGT CACCAAAGCA GTACAGCCTC TCTTACTGGG AAGAATCATA

   451  GCTTCCTATG ACCCGGATAA CAAGGAGGAA CGCTCTATCG CGATTTATCT
```

```
analyze% typedata -ref GB_PR:HUMIFNRF1A

!!NA_SEQUENCE 1.0
LOCUS       HUMIFNRF1A   7721 bp    DNA             PRI       10-NOV-1992
DEFINITION  Homo sapiens interferon regulatory factor 1 gene, complete
cds.
ACCESSION   L05072
NID         g184648
KEYWORDS    interferon regulatory factor 1.
SOURCE      Homo sapiens Placenta DNA.
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 7721)
  AUTHORS   Cha,Y., Sims,S.H., Romine,M.F., Kaufmann,M. and
Deisseroth,A.B.
  TITLE     Human interferon regulatory factor 1: intron/exon organization
  JOURNAL   DNA Cell Biol. 11, 605-611 (1992)
  MEDLINE   93000481
```

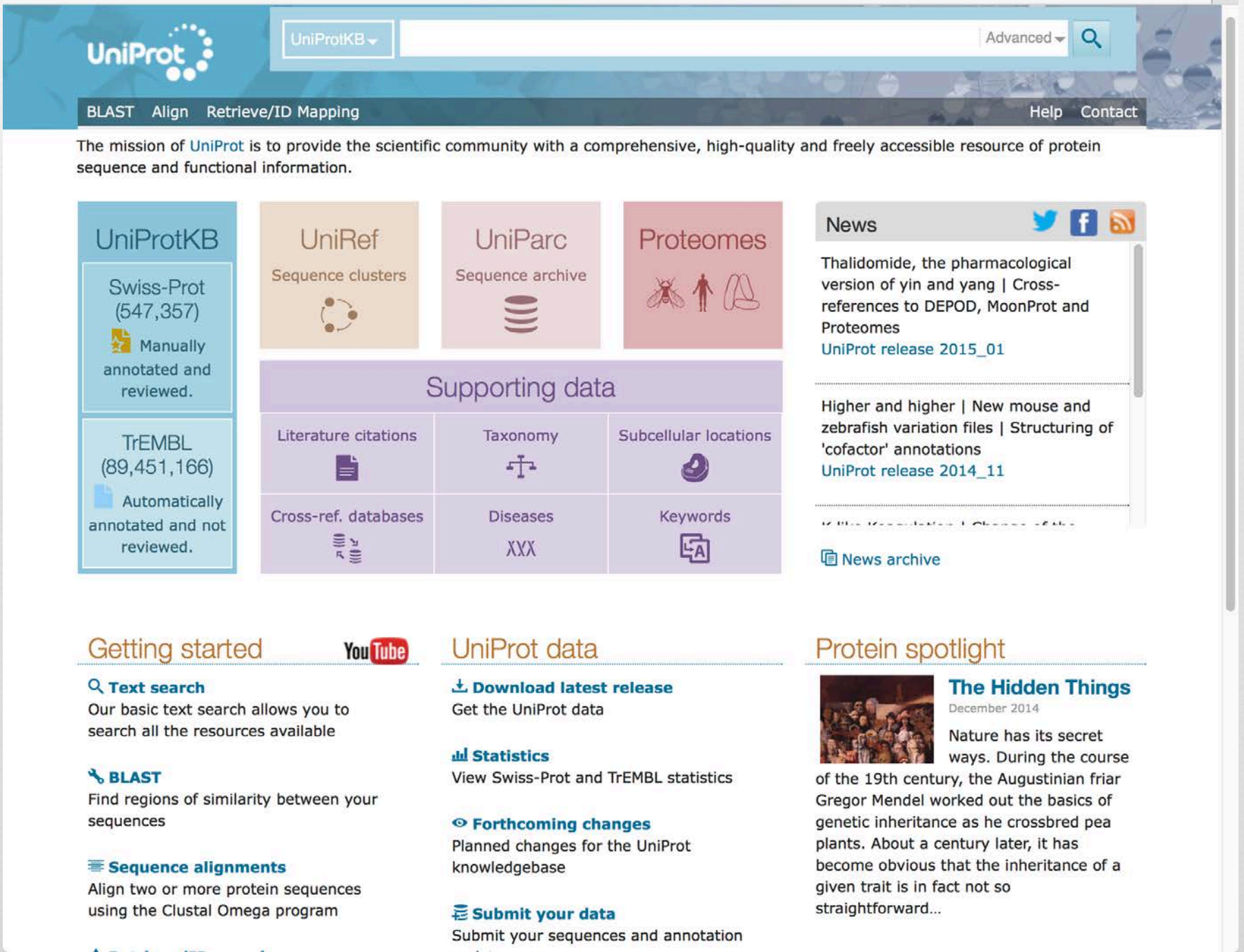
```
FEATURES             Location/Qualifiers
     source          1. .7721
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /tissue_type="Placenta"
                     /map="5q23-q31"
     exon            1. .219
                     /gene="IRF1"
                     /note="putative"
                     /number=1
     5'UTR           join(1. .219,1279. .1287)
                     /gene="IRF1"
     gene            join(1. .219,1279. .1287)
                     /gene="IRF1"
     intron          220. .1278
                     /gene="IRF1"
                     /number=1
     exon            1279. .1374
                     /gene="IRF1"
                     /number=2
     CDS             join(1288. .1374,2738. .2837,3630. .3806,3916. .3965,
                     4073. .4202,4386. .4508,5040. .5089,6248. .6383,6670.
                     .6794)
                     /gene="IRF1"
                     /codon_start=1
                     /product="interferon regulatory factor 1"
                     /db_xref="PID:g184649"
                     /translation="MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIFQIPWKHAAKH
                     GWDINKDACLFRSWAIHTGRYKAGEKEPDPKTWKANFRCAMNSLPDIEEVKDQSRNKG
                     SSAVRVYRMLPPLTKNQRKERKSKSSRDAKSKAKRKSCGDSSPDTFSDGLSSSTLPDD
                     HSSYTVPGYMQDLEVEQALTPALSPCAVSSTLPDWHIPVEVVPDSTSDLYNFQVSPMP
```

```
     intron          1375. .2737
                     /gene="IRF1"
                     /number=2
     exon            2738. .2837
                     /gene="IRF1"
                     /number=3
     intron          2838. .3629
                     /gene="IRF1"
                     /number=3
     exon            3630. .3806
                     /gene="IRF1"
                     /number=4
     intron          3807. .3915
                     /gene="IRF1"
                     /number=4
     exon            3916. .3965
                     /gene="IRF1"
                     /number=5
     intron          3966. .4072
                     /gene="IRF1"
                     /number=5

                      ...

     exon            5040. .5089
                     /gene="IRF1"
                     /number=8
     intron          5090. .6247
                     /gene="IRF1"
                     /number=8
     exon            6248. .6383
                     /gene="IRF1"
                     /number=9
     intron          6384. .6669
                     /gene="IRF1"
                     /number=9
     exon            6670. .7656
                     /gene="IRF1"
                     /number=10
     3'UTR           6795. .7656
BASE COUNT     1750 a   1946 c   2253 g   1772 t
ORIGIN
```

Protein Sequence Databases

UniProt

- Single site for access to protein sequences
- http://www.uniprot.org/
- Combines
  - Swiss-Prot
  - TrEMBL
  - PIR


www.uniprot.org
UniProt
UniProtKB
Advanced
BLAST Align Retrieve/ID Mapping Help Contact
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
UniProtKB
Swiss-Prot (547,357)
Manually annotated and reviewed.
TrEMBL (89,451,166)
Automatically annotated and not reviewed.
UniRef
Sequence clusters
UniParc
Sequence archive
Proteomes
Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords
News
Thalidomide, the pharmacological version of yin and yang | Cross-references to DEPOD, MoonProt and Proteomes
UniProt release 2015_01
Higher and higher | New mouse and zebrafish variation files | Structuring of 'cofactor' annotations
UniProt release 2014_11
News archive
Getting started
YouTube
Text search
Our basic text search allows you to search all the resources available
BLAST
Find regions of similarity between your sequences
Sequence alignments
Align two or more protein sequences using the Clustal Omega program
UniProt data
Download latest release
Get the UniProt data
Statistics
View Swiss-Prot and TrEMBL statistics
Forthcoming changes
Planned changes for the UniProt knowledgebase
Submit your data
Submit your sequences and annotation
Protein spotlight
The Hidden Things
December 2014
Nature has its secret ways. During the course of the 19th century, the Augustinian friar Gregor Mendel worked out the basics of genetic inheritance as he crossbred pea plants. About a century later, it has become obvious that the inheritance of a given trait is in fact not so straightforward...

Databases of Biological Knowledge

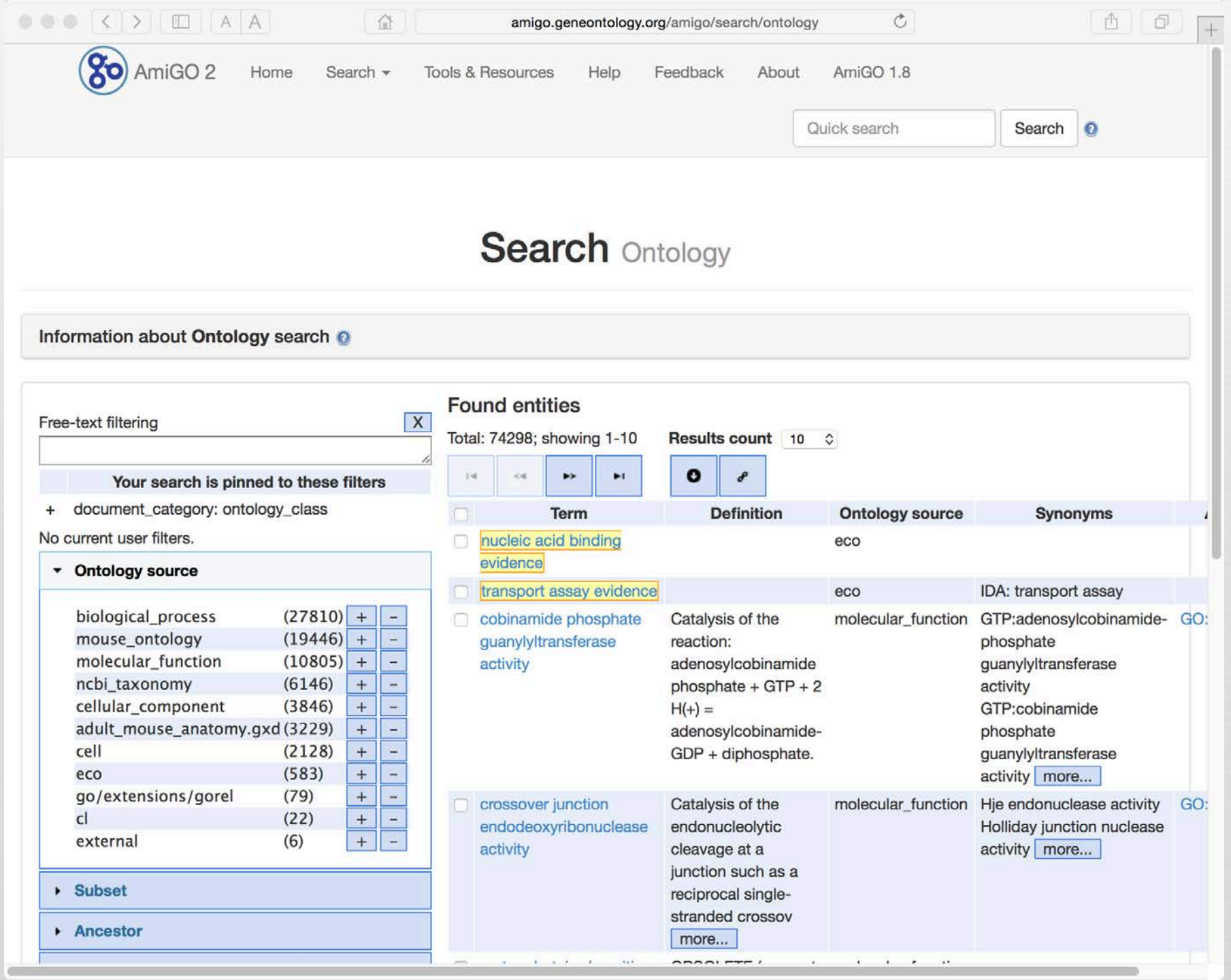
Categorizing, Describing, and Referencing Biological Information

Biological Ontologies

- Dictionary of terms describing biological processes
- Uses a "controlled vocabulary" to precisely define the terms to be used along with their definitions
  - Also provides a table of synonyms
- Provides the ability to construct reliable systems for computer-searchable databases and comparative analyses

A Few Ontologies

- Sequence Ontology (SO)
  - Ontology suitable for describing biological sequences
    - Sequence features with coordinates
      - exon, promoter, binding site…
    - Properties of features
  - http://song.sourceforge.net/
- Gene Ontology (GO)
  - Provides a controlled vocabulary to describe gene and gene product attributes in any organism
    - biological processes
    - cellular components
    - molecular functions
  - www.geneontology.org


amigo.geneontology.org/amigo/search/ontology
AmiGO 2
Home
Search
Tools & Resources
Help
Feedback
About
AmiGO 1.8
Quick search
Search
Search Ontology
Information about Ontology search
Free-text filtering
X
Your search is pinned to these filters
+ document_category: ontology_class
No current user filters.
Ontology source
biological_process (27810)
mouse_ontology (19446)
molecular_function (10805)
ncbi_taxonomy (6146)
cellular_component (3846)
adult_mouse_anatomy.gxd (3229)
cell (2128)
eco (583)
go/extensions/gorel (79)
cl (22)
external (6)
Subset
Ancestor
Found entities
Total: 74298; showing 1-10
Results count 10
Term
Definition
Ontology source
Synonyms
nucleic acid binding evidence
eco
transport assay evidence
eco
IDA: transport assay
cobinamide phosphate guanylyltransferase activity
Catalysis of the reaction: adenosylcobinamide phosphate + GTP + 2 H(+) = adenosylcobinamide-GDP + diphosphate.
molecular_function
GTP:adenosylcobinamide-phosphate guanylyltransferase activity GTP:cobinamide phosphate guanylyltransferase activity more...
crossover junction endodeoxyribonuclease activity
Catalysis of the endonucleolytic cleavage at a junction such as a reciprocal single-stranded crossov more...
molecular_function
Hje endonuclease activity Holliday junction nuclease activity more...

Searching for Information

Entrez Searching

- http://www.ncbi.nlm.nih.gov/entrez/
- Search via text patterns
- Cross-database search interface
  - Sequence
  - PubMed
  - OMIM
  - Linkage information
  - …

Search NCBI databases

Help

human cftr Search

Results found in 35 databases for "human cftr"

Literature

Books	141	books and reports
MeSH	14	ontology used for PubMed indexing
NLM Catalog	3	books, journals and more in the NLM Collections
PubMed	6,947	scientific & medical abstracts/citations
PubMed Central	11,510	full-text journal articles

Health

ClinVar	1,290	human variations of clinical significance
dbGaP	51	genotype/phenotype interaction studies
GTR	97	genetic testing registry
MedGen	1	medical genetics literature and links
OMIM	81	online mendelian inheritance in man
PubMed Health	15	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	31	biological projects providing data to NCBI
BioSample	124	descriptions of biological source materials
Clone	1,362	genomic and cDNA clones
dbVar	2,605	genome structural variation studies
Epigenomics	5,161	epigenomic studies and display tools
Genome	2	genome sequencing projects by organism
GSS	3	genome survey sequences
Nucleotide	6,572	DNA and RNA sequences
Probe	1,368	sequence-based probes and primers
SNP	10,723	short genetic variations
SRA	105	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	3	expressed sequence tag sequences
Gene	2,233	collected information about gene loci
GEO DataSets	186	functional genomics studies
GEO Profiles	234,414	gene expression and molecular abundance profiles
HomoloGene	13	homologous gene sets for selected organisms
PopSet	5	sequence sets from phylogenetic and population studies
UniGene	9	clusters of expressed transcripts

Proteins

Conserved Domains	6	conserved protein domains
Protein	1,075	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	64	experimentally-determined biomolecular structures

Chemicals

BioSystems	865	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	204	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	33	deposited substance and chemical information

NCBI Gene

- Provides a single query interface to curated sequence and descriptive information about genetic loci
  - Nomenclature
  - Aliases
  - Sequence accessions
  - Phenotypes
  - EC numbers
  - OMIM numbers
  - UniGene clusters
  - Homology
  - Map locations
  - Web sites

CFTR cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) [*Homo sapiens* (human)]

Gene ID: 1080, updated on 1-Feb-2015

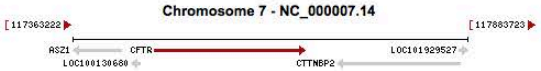
Summary

Official Symbol CFTR provided by HGNC
Official Full Name cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) provided by HGNC
Primary source HGNC:HGNC:1884
See related Ensembl:ENSG00000001626; HPRD:03883; MIM:602421; Vega:OTTHUMG00000023076
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as CF; MRP7; ABC35; ABCC7; CFTR/MRP; TNR-CFTR; dJ760C5.1
Summary This gene encodes a member of the ATP-binding cassette (ABC) transporter superfamily. ABC proteins transport various molecules across extra- and intra-cellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the MRP subfamily that is involved in multi-drug resistance. The encoded protein functions as a chloride channel and controls the regulation of other transport pathways. Mutations in this gene are associated with the autosomal recessive disorders cystic fibrosis and congenital bilateral aplasia of the vas deferens. Alternatively spliced transcript variants have been described, many of which result from mutations in this gene. [provided by RefSeq, Jul 2008]
Orthologs mouse all

Genomic context

Location: 7q31.2 See CFTR in MapViewer
Exon count: 31

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	7	NC_000007.14 (117470772..117668665)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	7	NC_000007.13 (117120017..117308719)

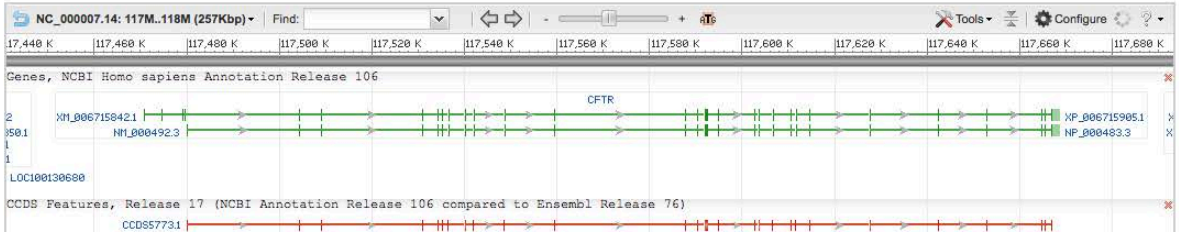


Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000007.14 chromosome 7 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank



Related information
Order cDNA clone
3D structures
BioAssay
BioAssay by Target (List)
BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Active
BioAssays, RNAi Target, Tested
BioProjects
BioSystems
Books
CCDS
ClinVar
Conserved Domains
dbVar
Full text in PMC
Full text in PMC_nucleotide
GAP
Gene neighbors
Genome
GEO Profiles
GTR
HomoloGene
Map Viewer
MedGen
Nucleotide
OMIM
Probe

OMIM

- Online Mendelian Inheritance in Man
- Database of gene-linked genetic disorders
- Maintained at Johns Hopkins University
  - Dr. Victor A. McKusick
- Provides link to GeneTests
  - Laboratories that provide testing for specific genetic disorders

*602421

CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR; CFTR

Alternative titles; symbols
ATP-BINDING CASSETTE, SUBFAMILY C, MEMBER 7; ABCC7

HGNC Approved Gene Symbol: CFTR

Cytogenetic location: 7q31.2 ***Genomic coordinates (GRCh37): 7:117,120,016-117,308,718*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Phenotype mapping key
7q31.2	Congenital bilateral absence of vas deferens	277180	3
	Cystic fibrosis	219700	3
	Sweat chloride elevation without CF		3
	{Bronchiectasis with or without elevated sweat chloride 1, modifier of}	211400	3
	{Hypertrypsinemia, neonatal}		3
	{Pancreatitis, idiopathic}	167800	3

TEXT

Description

The CFTR gene encodes an ATP-binding cassette (ABC) transporter that functions as a low conductance Cl(-)-selective channel gated by cycles of ATP binding and hydrolysis at its nucleotide-binding domains (NBDs) and regulated tightly by an intrinsically disordered protein segment distinguished by multiple consensus phosphorylation sites termed the regulatory domain (summary by Wang et al., 2014).

Cloning and Expression

Riordan et al. (1989) isolated overlapping cDNA clones from epithelial cell libraries with a genomic DNA segment containing a portion of the putative gene causing cystic fibrosis (CF; 219700). Transcripts approximately 6,500 nucleotides in size were detectable in the tissues affected in patients with CF. The predicted protein consists of 2 similar motifs, each with a domain having properties consistent with membrane-association, and a domain believed to be involved in ATP binding. In CF patients, a deleted phenylalanine residue occurs at the center of the putative first nucleotide-binding fold (NBF). The predicted protein has 1,480 amino acids with a molecular mass of 168,138 Da. The characteristics are remarkably similar to those of the mammalian multidrug resistant P-glycoprotein (171050), which also maps to 7q, and to a number of other membrane-associated proteins. To avoid confusion with the previously named CF antigen (123885), Riordan et al. (1989) referred to the protein as cystic fibrosis transmembrane conductance regulator (CFTR).

Cystic fibrosis represents the first genetic disorder elucidated strictly by the process of reverse genetics (later called positional cloning), i.e., on the basis of map location but without the availability of chromosomal rearrangements or deletions such as those that have greatly facilitated previous success in the cloning of human disease genes in Duchenne muscular dystrophy (310200), retinoblastoma (180200), and chronic granulomatous disease (306400), for example. By use of a combination of chromosome walking and jumping, Rommens et al. (1989) succeeded in covering the CF region on 7q. The jumping technique was particularly useful in bypassing 'unclonable' regions, which are estimated to constitute 5% of the human genome. (Yeast artificial chromosome (YAC) vectors represent an alternative strategy.) The identification of undermethylated CpG islands was 1 tip-off; another was screening of a cDNA library constructed from cultured sweat gland cells of a non-CF individual. The CF gene proved to be about 250,000 bp long, a surprising finding since the absence of apparent genomic rearrangements in CF chromosomes and the evidence of a limited number of CF mutations predicted a small mutational target.

Green and Olson (1990) described a general strategy for cloning and mapping large regions of human DNA with yeast artificial chromosomes (YACs). By analyzing 30 YAC clones from the region of chromosome 7 containing the CFTR gene, a contig map spanning more than 1.5 Mbp was assembled. Individual YACs as large as 790 kb

Comparative Analysis

Similarities and Differences

Comparative Analysis

- Identification of similarities
  - Primary sequence
  - Structure
  - Function
- Identification of differences
  - Gene complement
  - Genotypic differences resulting in phenotypic changes
- Phylogenetic inference
  - Predicting evolutionary history

Finding Sequences by Similarity

Homology vs. Similarity

- Similarity
  - Arises from:
    - Homology
    - Convergence
    - Gene Conversion
    - Chance
- Homology
  - Implies a common evolutionary origin
    - Homologs
    - Orthologs
    - Paralogs

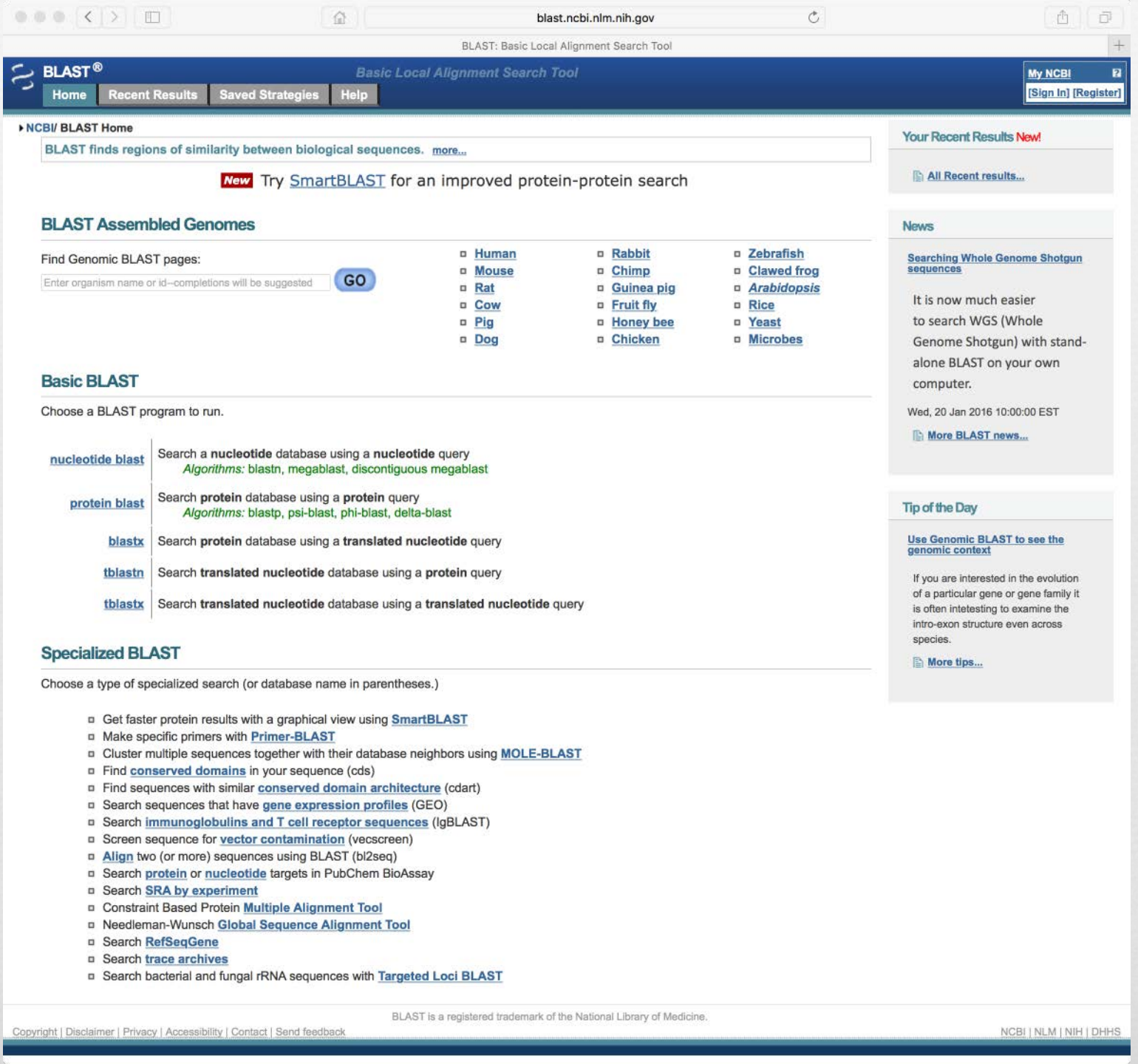
BLAST

- Basic Local Alignment Search Tool
- Search a sequence database for primary sequence similarities to some query sequence
- Provides a measure of the significance of the similarity
- Does not necessarily imply common evolutionary origin

BLAST

- All search combinations possible
- nt vs. nt database
  - blastn
- protein vs. protein database
  - blastp
- translated nt vs. protein database
  - blastx
- protein vs. translated nt database
  - tblastn
- translated nt vs. translated nt database
  - tblastx

BLAST Searching at NCBI


blast.ncbi.nlm.nih.gov
BLAST: Basic Local Alignment Search Tool
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New Try SmartBLAST for an improved protein-protein search
BLAST Assembled Genomes
Find Genomic BLAST pages:
Enter organism name or id--completions will be suggested
GO
Human
Mouse
Rat
Cow
Pig
Dog
Rabbit
Chimp
Guinea pig
Fruit fly
Honey bee
Chicken
Zebrafish
Clawed frog
Arabidopsis
Rice
Yeast
Microbes
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Get faster protein results with a graphical view using SmartBLAST
Make specific primers with Primer-BLAST
Cluster multiple sequences together with their database neighbors using MOLE-BLAST
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architecture (cdart)
Search sequences that have gene expression profiles (GEO)
Search immunoglobulins and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Needleman-Wunsch Global Sequence Alignment Tool
Search RefSeqGene
Search trace archives
Search bacterial and fungal rRNA sequences with Targeted Loci BLAST
Your Recent Results New!
All Recent results...
News
Searching Whole Genome Shotgun sequences
It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.
Wed, 20 Jan 2016 10:00:00 EST
More BLAST news...
Tip of the Day
Use Genomic BLAST to see the genomic context
If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species.
More tips...
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

BLAST® *Basic Local Alignment Search Tool*

Home | Recent Results | Saved Strategies | Help

My NCBI

Welcome elliotl. [Sign Out]

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

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

Reset page | Bookmark

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

Query subrange

From

To

Or, upload file: Choose File | no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

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

Program Selection

Algorithm:

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

Choose a BLAST algorithm

BLAST

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: 5000

Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

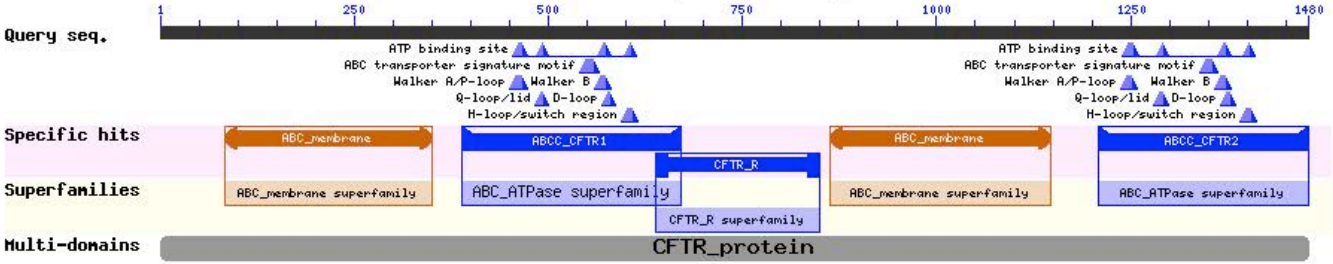
Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Job Title: gi|90421313|ref|NP_000483.3| cystic fibrosis...

Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	**CZARABWB014**
Status	Searching
Submitted at	Mon Feb 2 11:50:50 2015
Current time	Mon Feb 2 11:51:05 2015
Time since submission	00:00:15

This page will be automatically updated in **1** seconds

gi|90421313|ref|NP_000483.3| cystic fibrosis...

RID CZARABWB014 (Expires on 02-03 23:50 pm)

Query ID lcl|38978
Description gi|90421313|ref|NP_000483.3| cystic fibrosis transmembrane conductance regulator [Homo sapiens]
Molecule type amino acid
Query Length 1480

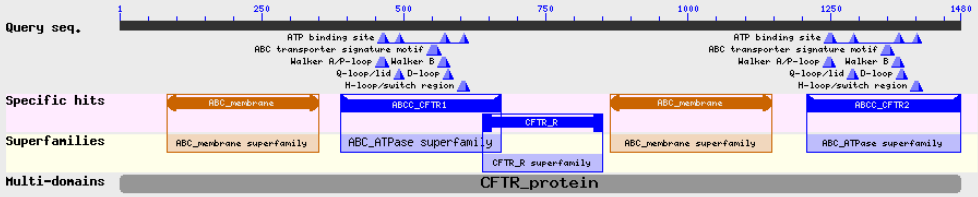
Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.2.30+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

Graphic Summary

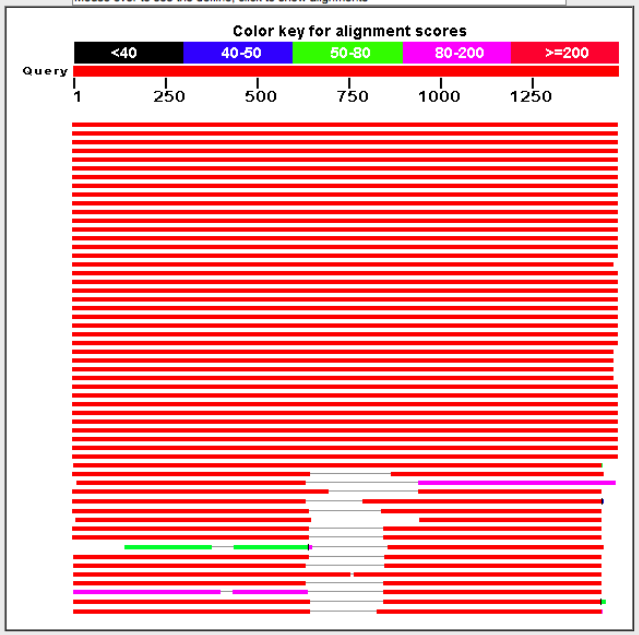
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 200 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	3046	3046	100%	0.0	100%	P13569.3
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	3034	3034	100%	0.0	99%	Q2IBF6.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	3033	3033	100%	0.0	99%	Q2QLE5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	3017	3017	100%	0.0	99%	Q2IBE4.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	3016	3016	100%	0.0	99%	Q07DX5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2994	2994	100%	0.0	98%	Q7JII8.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2994	2994	100%	0.0	98%	Q9TSP5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2992	2992	100%	0.0	98%	Q9TUQ2.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2991	2991	100%	0.0	98%	Q2IBA1.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2972	2972	100%	0.0	98%	Q07DY5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2956	2956	100%	0.0	97%	Q2QLF9.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2954	2954	100%	0.0	97%	Q07DV2.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2953	2953	100%	0.0	97%	Q2QLB4.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2947	2947	100%	0.0	97%	Q09YH0.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2945	2945	100%	0.0	96%	Q09YK5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2796	2796	100%	0.0	94%	Q2QL83.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2783	2783	99%	0.0	94%	Q2QLA3.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2762	2762	100%	0.0	92%	Q6PQZ2.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2760	2760	100%	0.0	92%	Q00554.4
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2734	2734	100%	0.0	93%	Q2QLH0.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2727	2727	100%	0.0	91%	Q07E42.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2703	2703	100%	0.0	92%	Q00PJ2.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2699	2699	100%	0.0	92%	Q07E16.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2694	2694	100%	0.0	90%	Q5U820.2
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2679	2679	100%	0.0	91%	Q108U0.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2674	2674	100%	0.0	91%	Q00552.2
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2669	2669	99%	0.0	91%	Q07DW5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2667	2667	99%	0.0	91%	Q09YJ4.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2652	2652	99%	0.0	91%	Q00555.2
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2652	2652	99%	0.0	91%	P35071.2
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2649	2649	100%	0.0	90%	Q2IBB3.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2615	2615	100%	0.0	89%	Q2QLC5.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2580	2580	100%	0.0	87%	Q2QL74.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2551	2551	100%	0.0	86%	Q5D1Z7.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2494	2494	100%	0.0	83%	Q07DZ6.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2340	2340	100%	0.0	78%	P26361.2
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2291	2291	100%	0.0	77%	P26363.1
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2281	2281	100%	0.0	78%	P34158.3
RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel	2129	2129	100%	0.0	71%	P26362.1
RecName: Full=Multidrug resistance-associated protein 1; AltName: Full=ATP-binding cassette sub-family C member 1; AltName: Full=Leukotriene C(4) transporter; Shor	488	565	96%	1e-144	26%	Q5F364.1
RecName: Full=Multidrug resistance-associated protein 4; AltName: Full=ATP-binding cassette sub-family C member 4; AltName: Full=MRP/cMOAT-related ABC transpor	421	971	82%	3e-122	34%	O15439.3
RecName: Full=Probable multidrug resistance-associated protein lethal(2)03659; AltName: Full=Wunen region A protein [Drosophila melanogaster]	347	843	78%	8e-98	32%	P91660.3
RecName: Full=ATP-binding cassette transporter abc2; Short=ABC transporter abc2; AltName: Full=ATP-energized glutathione S-conjugate pump abc2; AltName: Full=G	308	663	80%	2e-84	34%	Q10185.1

RecName: Full=Cystic fibrosis transmembrane conductance regulator; Short=CFTR; AltName: Full=ATP-binding cassette sub-family C member 7; AltName: Full=Channel conductance-controlling ATPase; AltName: Full=Dogfish transmembrane conductance regulator; AltName: Full=cAMP-dependent chloride channel [Squalus acanthias]

Sequence ID: sp|P26362.1|CFTR_SQUAC Length: 1492 Number of Matches: 1

Related Information

Range 1: 1 to 1492 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
2129 bits(5516)	0.0	Compositional matrix adjust.	1069/1497(71%)	1251/1497(83%)	22/1497(1%)

```
Query  1     MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLSEKLEREWDRE  60
             MQRSP+EKA+  SKLFF W RPIL+KGYRQ+LELSDIYQIPS DSAD LSE LEREWDRE
Sbjct  1     MQRSPIEKANAFSKLFFRWPRPILKKGYRQKLELSDIYQIPSSDSADELSEMLEREWDRE  60

Query  61    LA-SKKNPKLINALRRCFFWRFMFYGIFLYLGEVTKAVQPLLLGRIIASYDPDNKEERSI  119
             LA SKKNPKL+NALRRCFFWRF+FYGI LY  E TKAVQPL LGRIIASY+  N  ER I
Sbjct  61    LATSKKNPKLVNALRRCFFWRFLFYGILLYFVEFTKAVQPLCLGRIIASYNAKNTYEREI  120

Query  120   AIYLGIGLCLLFIVRTLLLHPAIFGLHHIGMQMRIAMFSLIYKKTLKLSSRVLDKISIGQ  179
             A YL +GLCLLF+VRTL LHPA+FGL H+GMQMRIA+FSLIYKK LK+SSRVLDKI  GQ
Sbjct  121   AYYLALGLCLLFVVRTLFLHPAVFGLQHLGMQMRIALFSLIYKKILKMSSRVLDKIDTGQ  180

Query  180   LVSLLSNNLNKFDEGLALAHFVWIAPLQVALLMGLIWELLQASAFCGLGFLIVLALFQAG  239
             LVSLLSNNLNKFDEG+A+AHFVWIAP+QV LLMGLIW  L    FCGLGFLI+LALFQA
Sbjct  181   LVSLLSNNLNKFDEGVAVAHFVWIAPVQVVLLMGLIWNELTEFVFCGLGFLIMLALFQAW  240

Query  240   LGRMMMKYRDQRAGKISERLVITSEMIENIQSVKAYCWEEAMEKMIENLRQTELKLTRKA  299
             LG+ MM+YRD+RAGKI+ERL ITSE+I+NIQSVK YCWE+AMEK+I+++RQ ELKLTRK
Sbjct  241   LGKKMMQYRDKRAGKINERLAITSEIIDNIQSVKVYCWEDAMEKIIDDIRQVELKLTRKV  300

Query  300   AYVRYFNSSAFFFSGFFVVFLSVLPYALIKGIILRKIFTTISFCIVLRMAVTRQFPWAVQ  359
             AY RYF+SSAFFFSGFFVVFLSV+PYA I  I LR+IFTTIS+ IVLRM VTRQFP A+Q
Sbjct  301   AYCRYFSSSAFFFSGFFVVFLSVVPYAFIHTIKLRRIFTTISYNIVLRMTVTRQFPSAIQ  360

Query  360   TWYDSLGAINKIQDFLQKEYKTLEYNLTTTEVVMENVTAFWEEGFGELFEKAKQNNNNR  419
             TWYDSLGAI KIQDFL K E+KT+EYNLTT EV M NVTA W+EG GELFEK KQN++ R
Sbjct  361   TWYDSLGAIRKIQDFLHKDEHKTVEYNLTTKEVEMVNVTASWDEGIGELFEKVKQNDSER  420

Query  420   KTSNGDDSLFFSNFSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMVIMGELEPSE  479
             K +NGDD LFFSNFSL  TPVLK+I+FK+E+G+LLA+AGSTG+GK+SLLM+IMGELEPS+
Sbjct  421   KMANGDDGLFFSNFSLHVTPVLKNISFKLEKGELLAIAGSTGSGKSSLLMMIMGELEPSD  480

Query  480   GKIKHSGRISFCSQFSWIMPGTIKENIIFGVSYDEYRYRSVIKACQLEEDISKFAEKDNI  539
             GKIKHSGRIS+  Q  WIMPGTIK+NIIFG+SYDEYRY SV+ ACQLEEDI+ F  KD
Sbjct  481   GKIKHSGRISYSPQVPWIMPGTIKDNIIFGLSYDEYRYTSVVNACQLEEDITVFPNKDKT  540

Query  540   VLGEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEKEIFESCVCKLMANKT  599
             VLG+GGITLSGGQRARISLARA+YKDADLYLLDSPF +LDV TEK+IFESC+CKLM NKT
Sbjct  541   VLGDGGITLSGGQRARISLARALYKDADLYLLDSPFSHLDVTTEKDIFESCLCKLMVNKT  600

Query  600   RILVTSKMEHLKKADKILILHEGSSYFYGTFSELQNLQPDFSSKLMGCDSFDQFSAERRN  659
             RILVTSK+EHLKKADKIL+LHEG  YFYGTFSELQ  +PDFSS+L+G   FD FSAERRN
Sbjct  601   RILVTSKLEHLKKADKILLLHEGHCYFYGTFSELQGEKPDFSSQLLGSVHFDSFSAERRN  660

Query  660   SILTETLHRFSL---EGDAPVSWTETKKQSFKQ-TGEFGEKRKNS-ILNPINSIRKFSIV  714
             SILTET  R S+   +G    S++ET+K SFKQ   EF EKRK+S I+NPI S +KFS+V
Sbjct  661   SILTETFRRCSVSSGDGAGLGSYSETRKASFKQPPPEFNEKRKSSLIVNPITSNKKFSLV  720

Query  715   QK--TPLQMNGIEEDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVLNL  772
             Q   +  Q NG+E+ + EP ER  SL+P++E GE    PR ++  +    QA RRQSVL L
Sbjct  721   QTAMSYPQTNGMEDATSEPGERHFSLIPENELGEPTKPRSNIFKSELPFQAHRRQSVLAL  780

Query  773   MTHSVNQGQNIHRKTTASTRKVSLAPQANL--TELDIYSRRLSQETGLEISEEINEEDLK  830
             MTHS      IH + +A  RK+S+  Q N   +E+DIYSRRLS++   EISEEINEEDLK
Sbjct  781   MTHSSTSPNKIHARRSA-VRKMSMLSQTNFASSEIDIYSRRLSEDGSFEISEEINEEDLK  839

Query  831   ECFFDDMESIPAVTTWNTYLRYITVHKSLIFVLIWCLVIFLAEVAASLVVLWLLG----N  886
             ECF D+ E     TTW+TYLRY+T  +++L+FVLI CLVIFLAEVAASL  LW++      N
Sbjct  840   ECFADEEEIQNVTTTWSTYLRYVTTNRNLVFVLILCLVIFLAEVAASLAGLWIISGLAIN  899

Query  887   TPLQDKGNSTH-SRNNSYAVIITSTSSYYVFYIYVGVADTLLAMGFFRGLPLVHTLITVS  945
             T  Q    ST  S  + ++  IT+ S YY+FYIYVG+AD+ LA+G  RGLPLVHTL+TVS
Sbjct  900   TGSQTNDTSTDLSHLSVFSKFITNGSHYYIFYIYVGLADSFLALGVIRGLPLVHTLVTVS  959

Query  946   KILHHKMLHSVLQAPMSTLNTLKAGGILNRFSKDIAILDDLLPLTIFDFIQLLLIVIGAI  1005
             K LH +MLHSVLQ PM+  N +KAG ILNRF KD AI+DD+LPLT+FDF+QL+LIV+GAI
Sbjct  960   KDLHKQMLHSVLQGPMTAFNKMKAGRILNRFIKDTAIIDDMLPLTVFDFVQLILIVVGAI  1019

Query  1006  AVVAVLQPYIFVATVPVIVAFIMLRAYFLQTSQQLKQLESEGRSPIFTHLVTSLKGLWTL  1065
              VV+VLQPY  +A +PV V FIMLRAYFL+TSQQLKQLESE RSPIF+HL+TSL+GLWT+
Sbjct  1020  CVVSVLQPYTLLAAIPVAVIFIMLRAYFLRTSQQLKQLESEARSPIFSHLITSLRGLWTV  1079

Query  1066  RAFGRQPYFETLFHKALNLHTANWFLYLSTLRWFQMRIEMIFVIFFIAVTFISILTTGEG  1125
             RAFGRQ YFETLFHKALNLHTANWFLYLSTLRWFQMRI+++FV+FFIAVTFI+I T   G
Sbjct  1080  RAFGRQSYFETLFHKALNLHTANWFLYLSTLRWFQMRIDIVFVLFFIAVTFIAIATHDVG  1139

Query  1126  EGRVGIILTLAMNIMSTLQWAVNSSIDVDSLMRSVSRVFKFIDMPTEGKPTKSTKPYKNG  1185
             EG+VGIILTLAMNI STLQWAVNSSIDVD LMRSVSRVFK+ID+P EG  TK+     N
Sbjct  1140  EGQVGIILTLAMNITSTLQWAVNSSIDVDGLMRSVSRVFKYIDIPPEGSETKNRHNANNP  1199

Query  1186  QLSKVMIIENSHVKKDDIWPSGGQMTVKDLTAKYTEGGNAILENISFSISPGQRVGLLGR  1245
               S V++IEN H+ K+  WPSGGQM V +LTAKYT  G A+L+++SFS++ GQRVGLLGR
Sbjct  1200  --SDVLVIENKHLTKE--WPSGGQMMVNNLTAKYTSDGRAVLQDLSFSVNAGQRVGLLGR  1255

Query  1246  TGSGKSTLLSAFLRLLNTEGEIQIDGVSWDSITLQQWRKAFGVIPQKVFIFSGTFRKNLD  1305
             TG+GKSTLLSA LRLL+TEGEIQIDG+SW+S++LQ+WRKAFGVIPQKVF+FSGTFRKNLD
Sbjct  1256  TGAGKSTLLSALLRLLSTEGEIQIDGISWNSVSLQKWRKAFGVIPQKVFVFSGTFRKNLD  1315

Query  1306  PYEQWSDQEIWKVADEVGLRSVIEQFPGKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAK  1365
             PYEQWSD+EIWKV +EVGL+S+IEQFP KL+FVLVDGG +LS+GHKQLMCLARS+LSKAK
Sbjct  1316  PYEQWSDEEIWKVTEEVGLKSMIEQFPDKLNFVLVDGGYILSNGHKQLMCLARSILSKAK  1375

Query  1366  ILLLDEPSAHLDPVTYQIIRRTLKQAFADCTVILCEHRIEAMLECQQFLVIEENKVRQYD  1425
             ILLLDEP +HLDPVT+QIIR+TLK  F++CTVIL EHR+EA+LECQQFLVIE   V+Q+D
Sbjct  1376  ILLLDEPFSHLDPVTFQIIRKTLKHSFSECTVILSEHRVEALLECQQFLVIEGSSVKQFD  1435

Query  1426  SIQKLLNERSLFRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL  1480
             SIQKLLNERSLFRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL
Sbjct  1436  SIQKLLNERSLFRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL  1492
```

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Cluster multiple sequences together with their database neighbors using MOLE-BLAST
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA by experiment
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search trace archives

Your Recent Results New!

All Recent results...

News

BLAST 2.2.30+ released

A new version of the stand-alone BLAST executables is now available.

Wed, 29 Oct 2014 15:00:00 EST

More BLAST news...

Tip of the Day

Use Genomic BLAST to see the genomic context

If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species.

More tips...

Pairwise Sequence Comparison

Detecting Similarity

- Is there a similarity between two sequences?
  - Identical symbols (nucleotides or amino acids)
  - Related symbols (amino acids)
- Do gaps/rearrangements allow for a higher degree of similarity?

Symbol Comparison Tables (Scoring Matrices)

- How do we compare one sequence to another
  - What is a match?
- Define match values for all possible symbol comparisons
  - Nucleotides
  - Amino acids

Nucleotide Tables

- Identical matches
  - A=A, T=T, C=C, G=G, U=U, T=U
- Matches with ambiguous bases
  - Based on ambiguity symbols
  - Y=T, Y=C
  - R=A, R=G
- Scores for matches or mismatches may differ depending on the table in use

Amino Acid Tables

- Measure of similarity between amino acids
- Not a simple match/mismatch relationship
- Values vary depending on degree of relatedness
- Based on evolution, chemistry, or structure

BLOSUM62 Table

- Default table for amino acid comparisons
- Many other similarity matrices are available

BLOSUM62 amino acid substitution matrix.

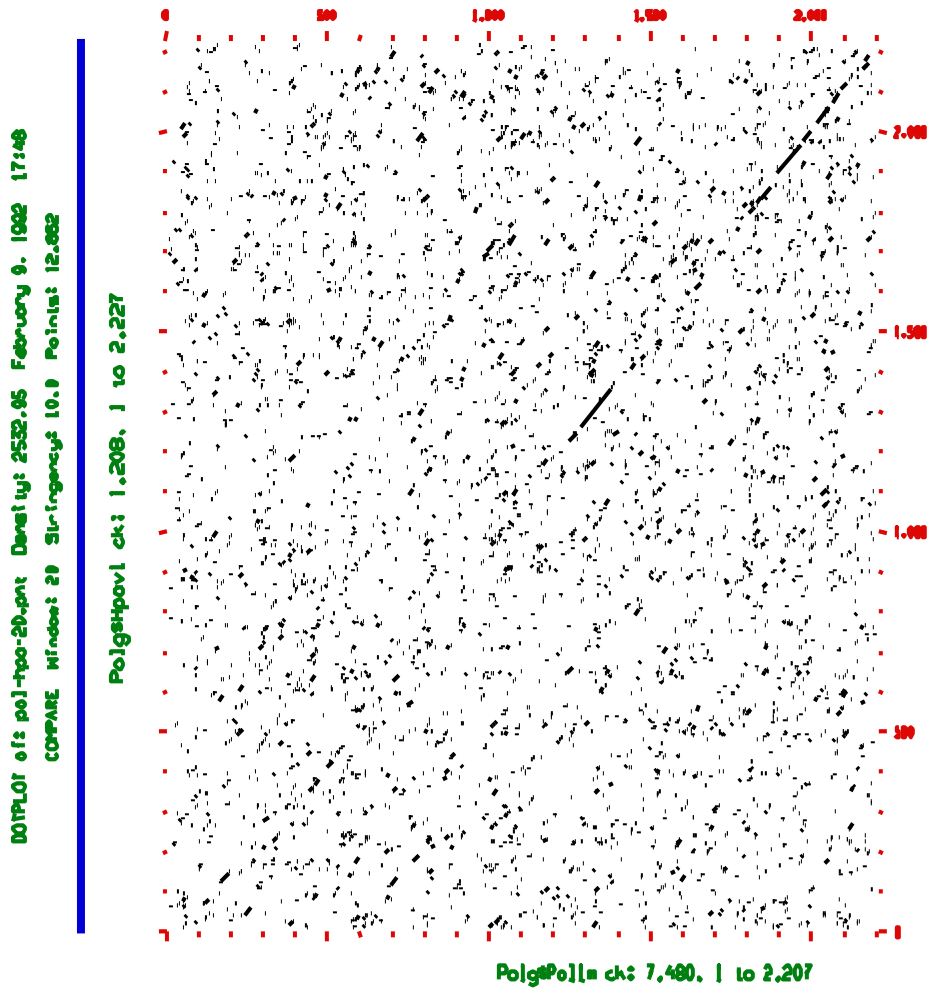
```
{
GAP_CREATE 12
GAP_EXTEND 4
}
```

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	2	0
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-3
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-2	1
L	-1	-4	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-3
M	-1	-3	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-2
N	-2	1	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

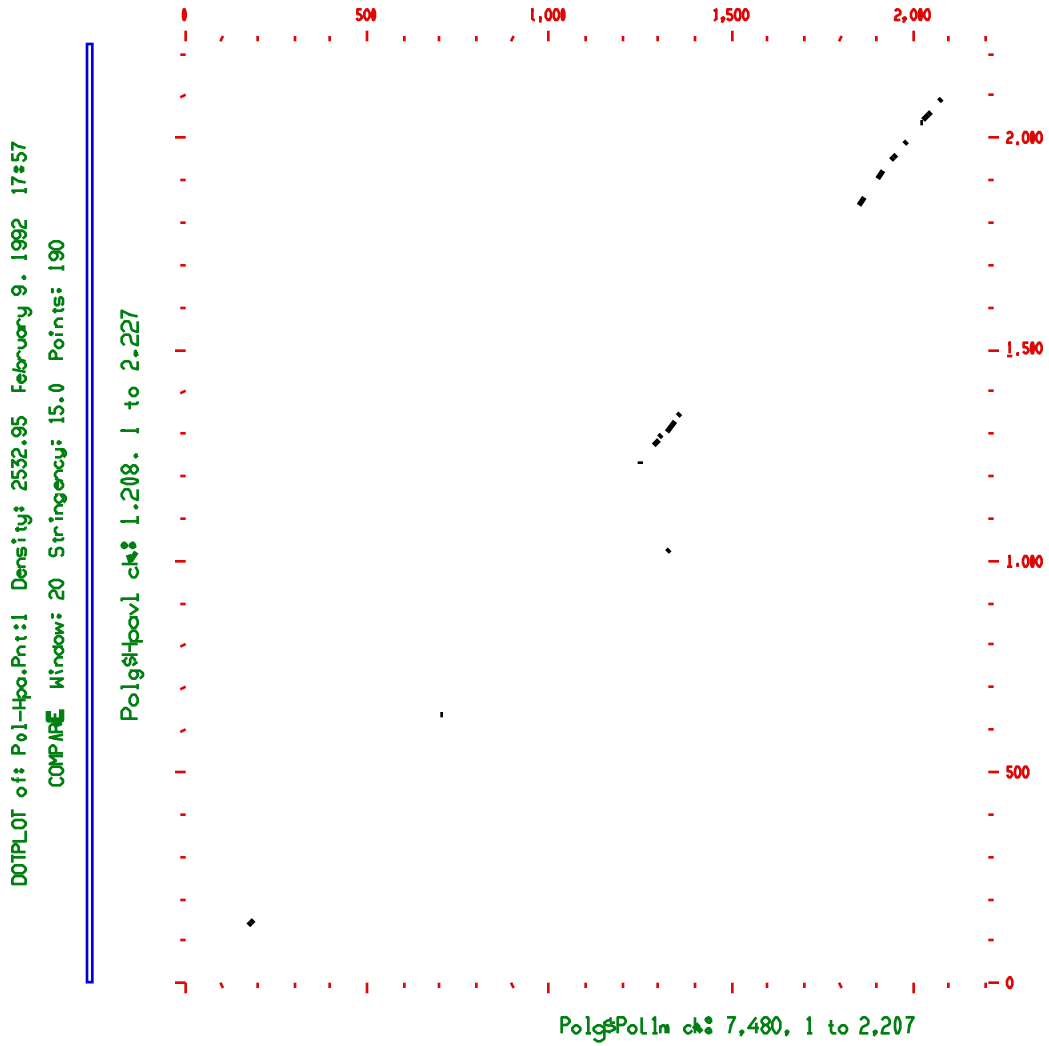
Dot Plots

- Allow comparison of two sequences in all registers
- Produces a graph (Dotplot) of sequence similarities
- The human brain interprets the results

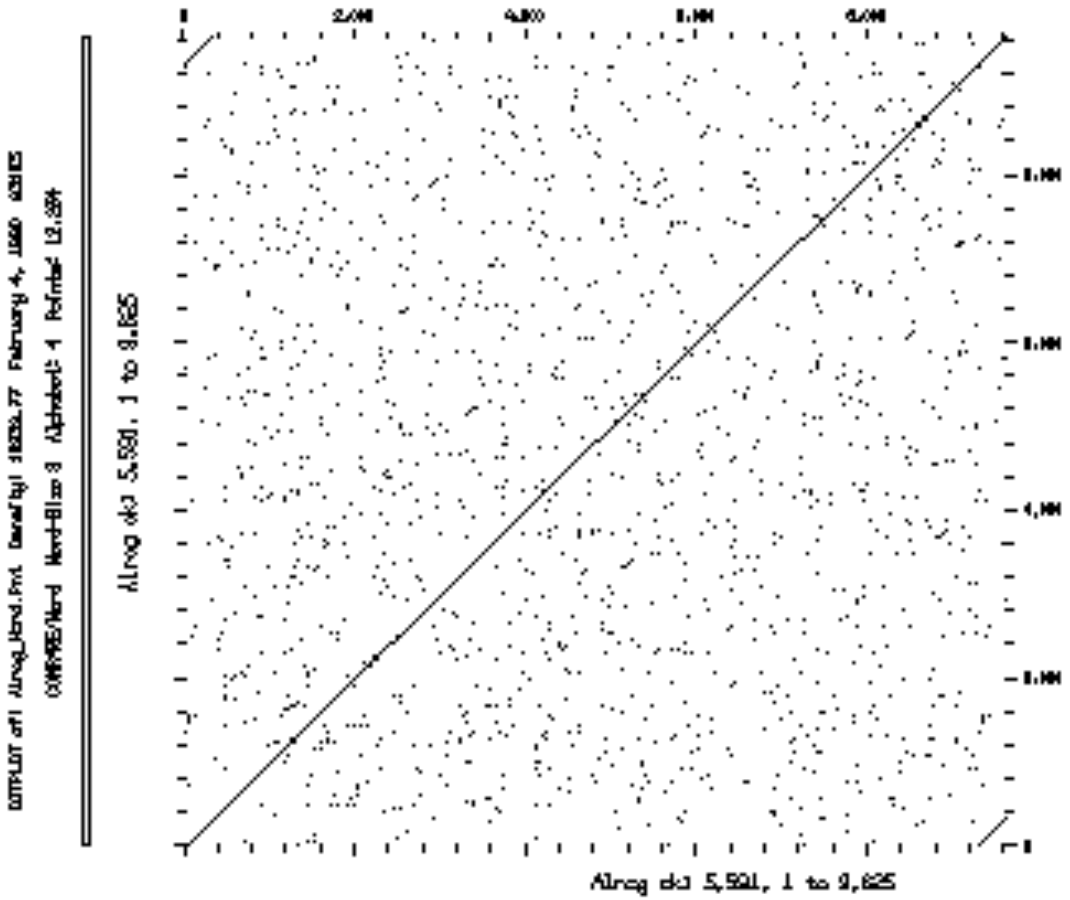
Polio:Hepatitis A 20/10 (Protein)



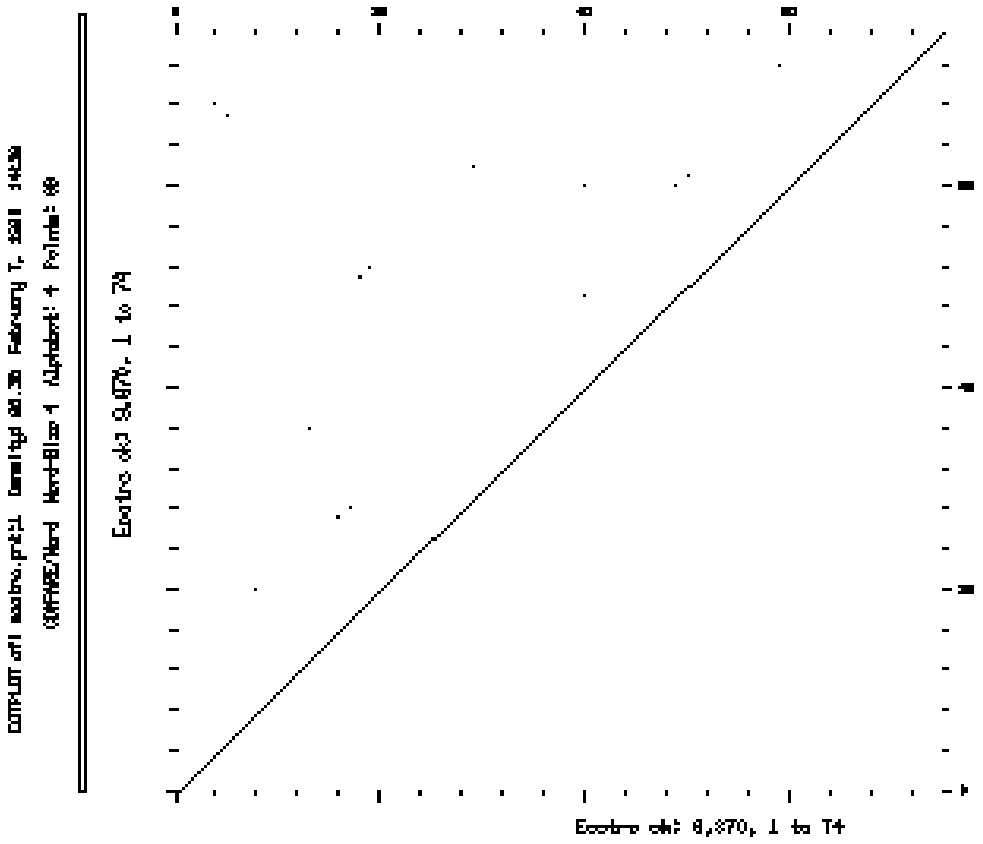
Polio:Hepatitis A 20/15 (Protein)



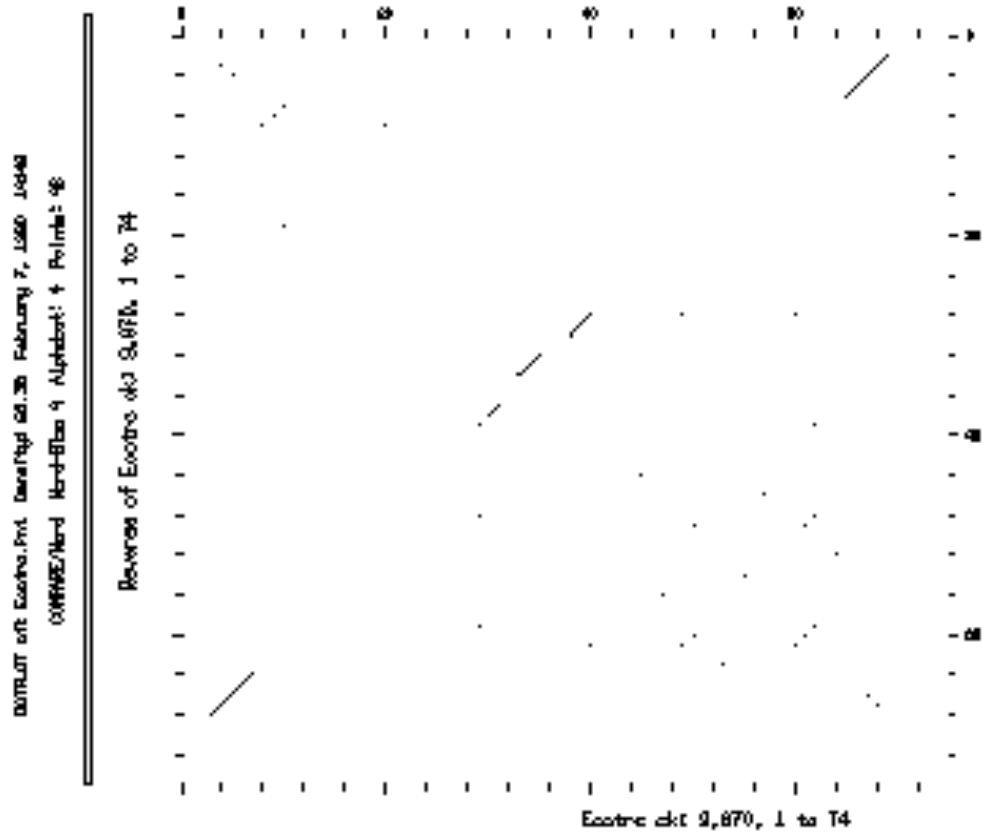
Rous Sarcoma Virus vs. Itself



E. coli tRNA-Cys vs. Itself

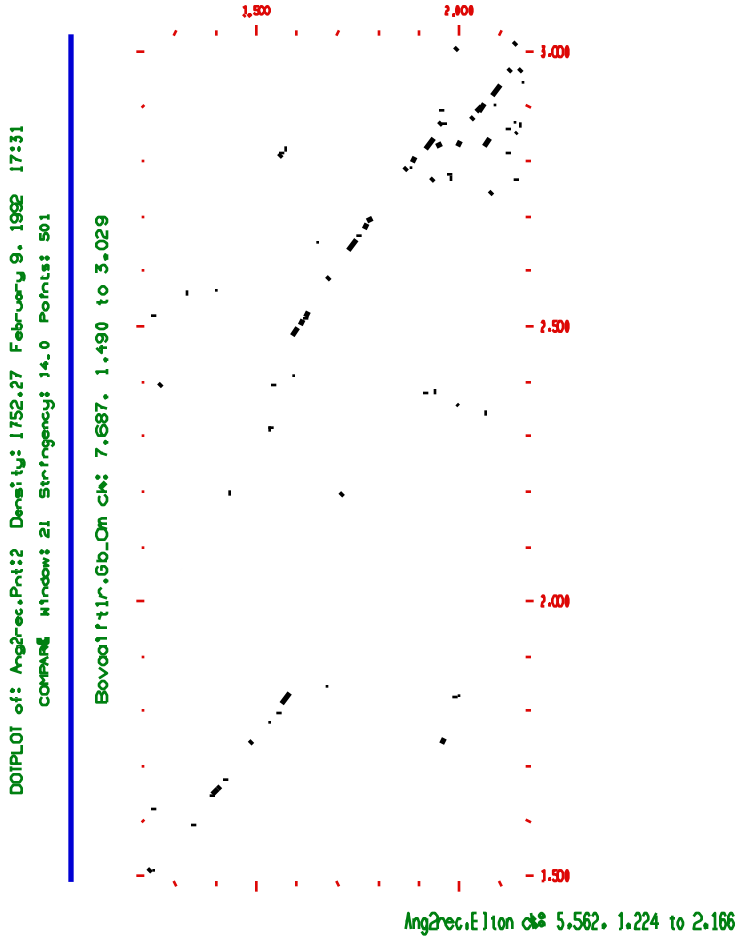


E. coli tRNA-Cys vs. Its Reverse-Complement



Angiotensin II Receptor mRNA

3' non-coding end; Bovine:Rat



Π σ Ρτ

Computer Generated Alignments

- Align two (or more) sequences by the introduction of gaps
  - Maximize sequence identity/similarity
  - Minimize gaps

Local vs. Global Alignments

- Local: What is the best region of similarity between two sequences?
  - Smith and Waterman
  - Not necessarily the whole sequence
- Global: What is the best possible alignment between two sequences?
  - Needleman–Wunsch
  - Always the whole sequence

Quality for Nucleotide Alignments

- Matches Rewarded
- Mismatches Penalized
- Gaps Penalized
  - GapWeight
    - Penalty for the introduction of a new gap
  - GapLengthWeight
    - Penalty for extension of an existing gap

Quality for Protein Alignments

- GapWeight and GapLengthWeight apply as for nucleotides
- A sum of comparison values between aligned amino acids is used in place of match and mismatch values

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

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

Reset page | Bookmark

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

NP_000483.3

Query subrange

From

To

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

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

NP_001038348.1

Subject subrange

From

To

Or, upload file Choose File no file selected

Program Selection

Algorithm

◉ blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST

Search protein sequence using Blastp (protein-protein BLAST)

☐ Show results in a new window

⊕ Algorithm parameters

Multiple sequence alignments

Get it right the first time!!

Why MSAs?

- Starting point for many types of comparative analyses
- Identification of functional motifs
  - Regulatory
  - Protein domains
- Evolutionary inference
  - Evolutionary history
  - Selection pressures

Conserved functional domains

- Sequences required for common function
- Conserved among different species

Variable domains

- Sequences under selection
  - Antibody escape mutants

Peptide motifs

- Active sites
- Binding motifs
- Protein modification motifs

The Alignment

- Critical!!!
- For coding regions first align the protein sequences and then align the nucleotide sequences to the protein sequence alignment
- Construct trees for both the protein and nucleotide alignments

Choose Your Program

- Simple nt or aa alignments
  - Clustal
- Alignment of many (hundreds) of sequences
- Large alignments (Whole genomes)
  - Sequence lengths from thousands to millions of residues

MSA Programs

- Clustal (W and X)
- DIALIGN
- MAFFT
- MAUVE
- MAVID
- MUSCLE
- T-Coffee
- …

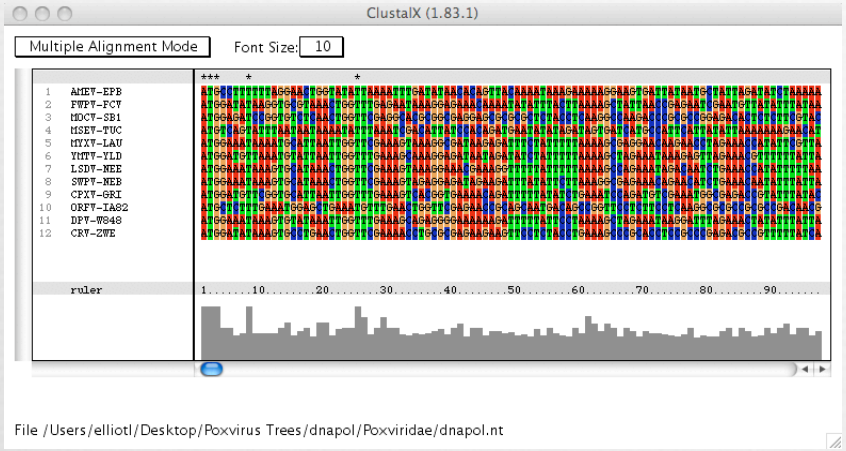
Choose your Parameters

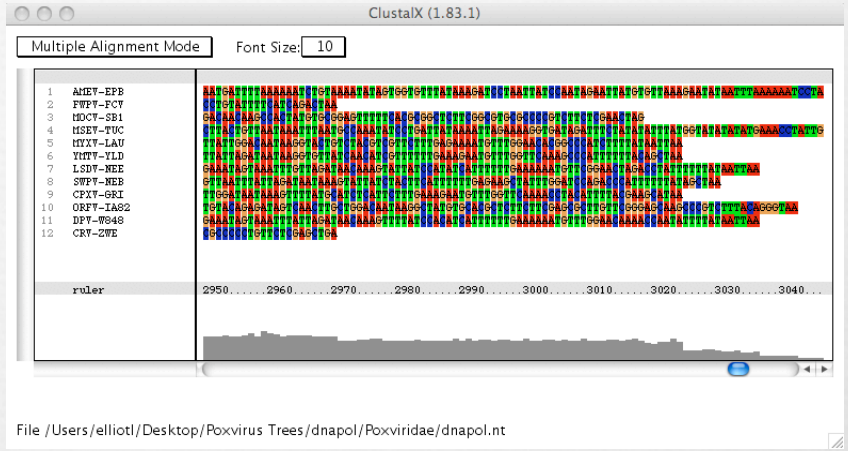
- Gap weight penalty
  - Penalty to introduce a new gap
- Gap length weight penalty
  - Penalty to extend a gap
- Substitution matrices
  - Evolutionary distance for various aa substitutions
- Start with the default settings
  - Inspect the alignment
  - Change parameters as needed

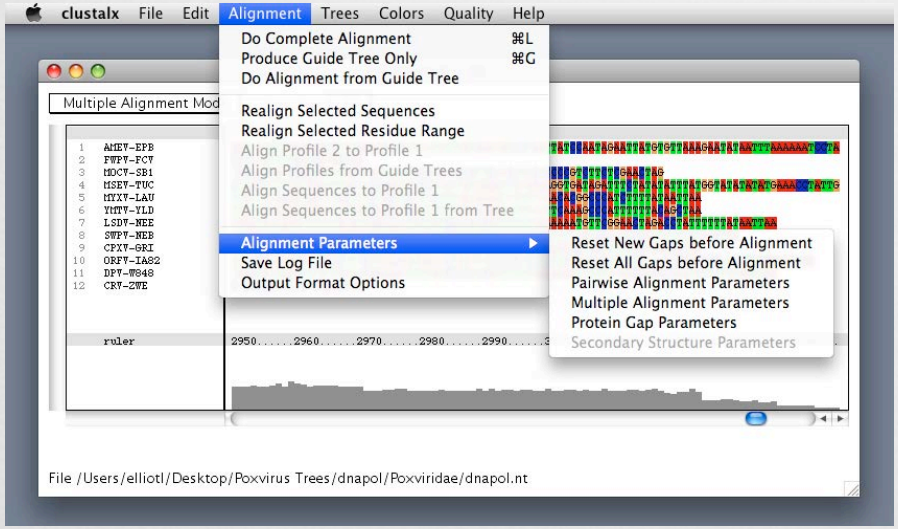
Inspect Your Results

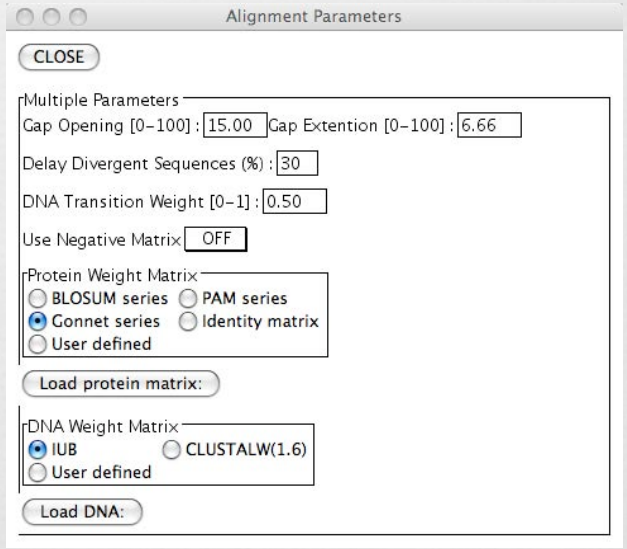
- Alignment Uncertainty and Genomic Analysis Science v319, p473, 1/25/2008
  - Learn to critically evaluate anything the computer tells you
  - Understand the limitations of any particular analysis tool
  - Understand the parameters that influence the behavior of any particular analysis tool

Using ClustalX

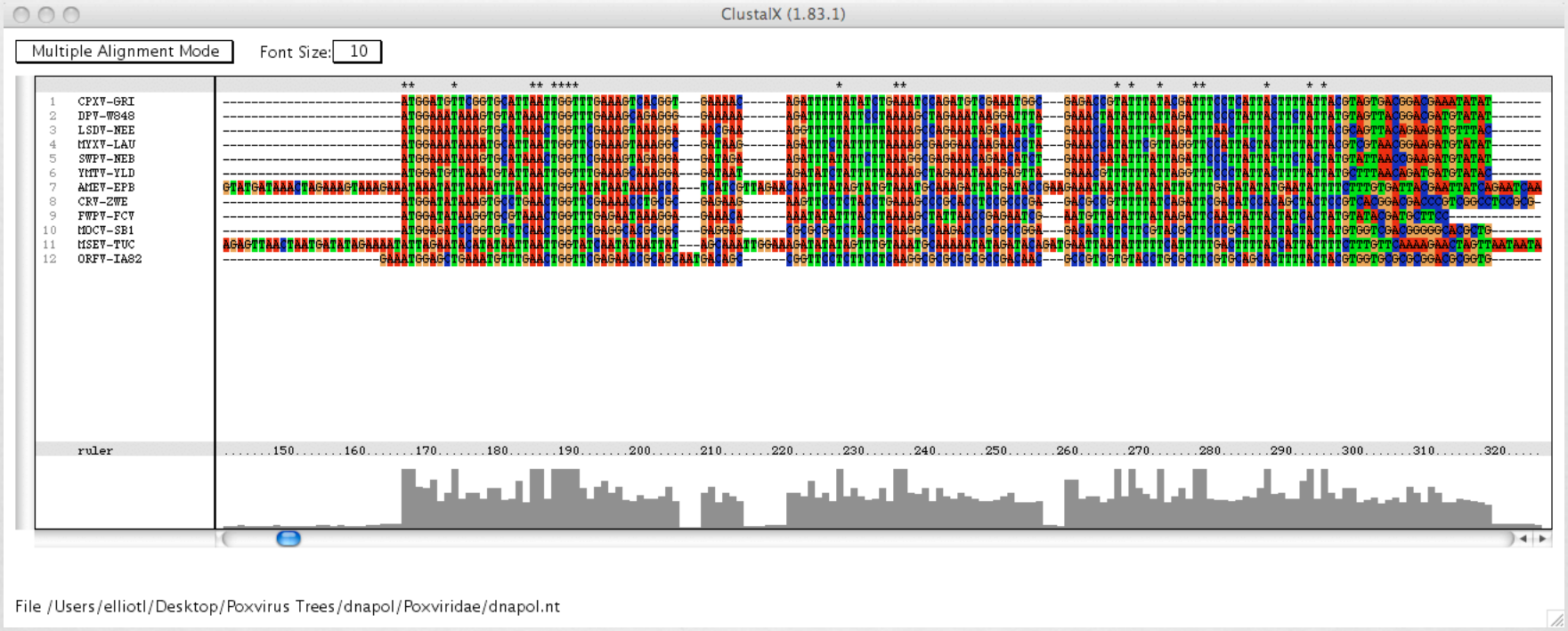




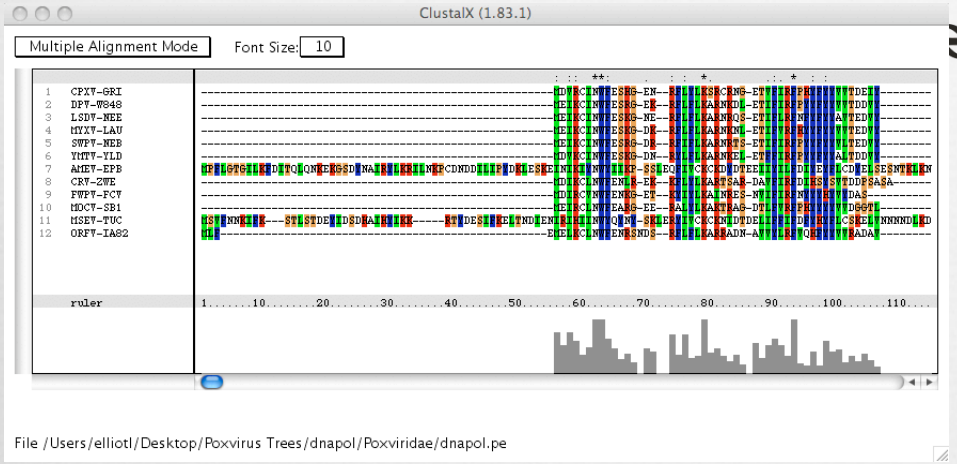


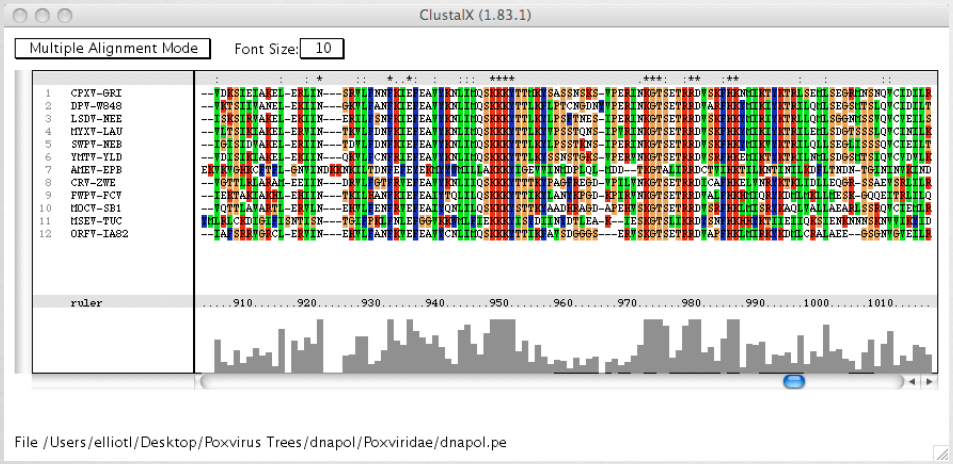


Clustal NT Alignment



Clustal AA Alignment





- Align Protein Seqs
- "Back translate" NT Seqs to AA alignment
  - MEGA
  - T-Coffee
- Provides codon-aligned NT MSA

Genomics

Genomic Resources

- NCBI Genome Resources
  - http://www.ncbi.nih.gov/Genomes/
- Ensembl
  - European Molecular Biology Laboratory (EMBL)
    - European Bioinformatics Institute
  - http://www.ensembl.org/
- UCSC Genome Browser
  - http://genome.ucsc.edu

Human (*Homo sapiens*) Genome Browser Gateway

group: Mammal | genome: Human | assembly: Feb. 2009 (GRCh37/hg19) | position: chr21:33,031,597-33,041,570 | search term: enter position, gene symbol or search terms | submit

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Homo sapiens
(Graphic courtesy of CBSE)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or
15q11;15q13	chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for
rs1042522;rs1800370	other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zahler	Lists mRNAs deposited by scientist named Zahler
Evans,J.E.	Lists mRNAs deposited by co-author J.E. Evans

Use this last format for author queries. Although GenBank requires the search format *Evans JE*, internally it uses the format *Evans,J.E.*.

Assembly Details

The GRCh37 build reference sequence is considered to be "finished", a technical term indicating that the sequence is highly accurate (with fewer than one error per 10,000 bases) and highly contiguous (with the only remaining gaps corresponding to regions whose sequence cannot be reliably resolved with current technology). Future work on the reference sequence will focus on improving accuracy and reducing gaps in the sequence. Statistics for the GRCh37 build assembly can be found on the NCBI Build 37.1 Statistics web page. For a glossary of assembly-related terms, please see the GRC Assembly Terminology page.

Note on chrM
Since the release of the UCSC hg19 assembly, the *Homo sapiens* mitochondrion sequence (represented as "chrM" in the Genome Browser) has been replaced in GenBank with the record NC_012920. We have not replaced the original sequence, NC_001807, in the hg19 Genome Browser. We plan to use the Revised Cambridge Reference Sequence (rCRS) in the next human assembly release.

Chromosome naming scheme
In addition to the "regular" chromosomes, the hg19 browser contains nine haplotype chromosomes and 59 unplaced contigs. If an unplaced contig is localized to a chromosome, the contig name is appended to the regular chromosome name, as in `chr1_gl000191_random`. If the chromosome is unknown, the contig is represented with the name "chrUn" followed by the contig identifier, as in `chrUn_gl000211`. Note that the chrUn contigs are no longer placed in a single, artificial chromosome as they have been in previous UCSC assemblies. See the sequences page for a complete list of hg19 chromosome names.

The nine haplotype chromosomes are:

name	accession	UCSC chr name
HSCHR6_MHC_APD_CTG1	GL000250.1	chr6_apd_hap1
HSCHR6_MHC_COX_CTG1	GL000251.1	chr6_cox_hap2
HSCHR6_MHC_DBB_CTG1	GL000252.1	chr6_dbb_hap3
HSCHR6_MHC_MANN_CTG1	GL000253.1	chr6_mann_hap4
HSCHR6_MHC_MCF_CTG1	GL000254.1	chr6_mcf_hap5
HSCHR6_MHC_QBL_CTG1	GL000255.1	chr6_qbl_hap6
HSCHR6_MHC_SSTO_CTG1	GL000256.1	chr6_ssto_hap7
HSCHR4_1_CTG9	GL000257.1	chr4_ctg9_hap1
HSCHR17_1_CTG5	GL000258.1	chr17_ctg5_hap1

See the Wellcome Trust Sanger Institute MHC Haplotype Project web site for additional information on the chr6 alternate haplotype assemblies.

The Y chromosome in this assembly contains two pseudoautosomal regions (PARs) that were taken from the corresponding regions in the X chromosome and are exact duplicates:

chrY:10001-2649520 and chrY:59034050-59363566
chrX:60001-2699520 and chrX:154931044-155260560

For further information on GRCh37 build see the NCBI GRCh37 release notes.

Bulk downloads of the sequence and annotation data are available via the Genome Browser FTP server or the Downloads page. We recommend that you use rsync for downloading large or multiple files.

The hg19 annotation tracks were generated by UCSC and collaborators worldwide. See the Credits page for a detailed list of the organizations and individuals who contributed to this release.

Genbank Pipeline Details

For the purposes of the Genbank alignment pipeline, this assembly is considered to be: **finished**.

One Final Word of Wisdom…

- "…although the computer is a wonderful helpmate for the sequence searcher and comparer, biochemists and molecular biologists must guard against the blind acceptance of any algorithmic output; given the choice, think like a biologist and not a statistician."
  - - Russell F. Doolittle, 1990