Introduction to Bioinformatics

8. Mining Genomic Sequence Data

Benjamin F. Matthews

United States Department of Agriculture

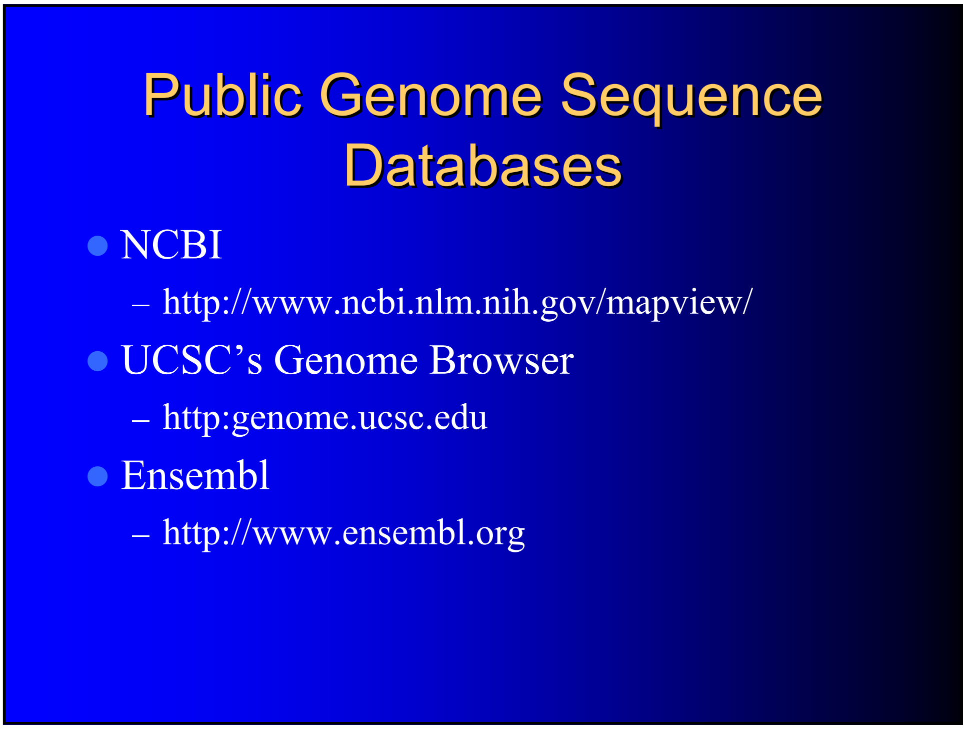
Soybean Genomics and Improvement Laboratory

Beltsville, MD 20708

matthewb@ba.ars.usda.gov

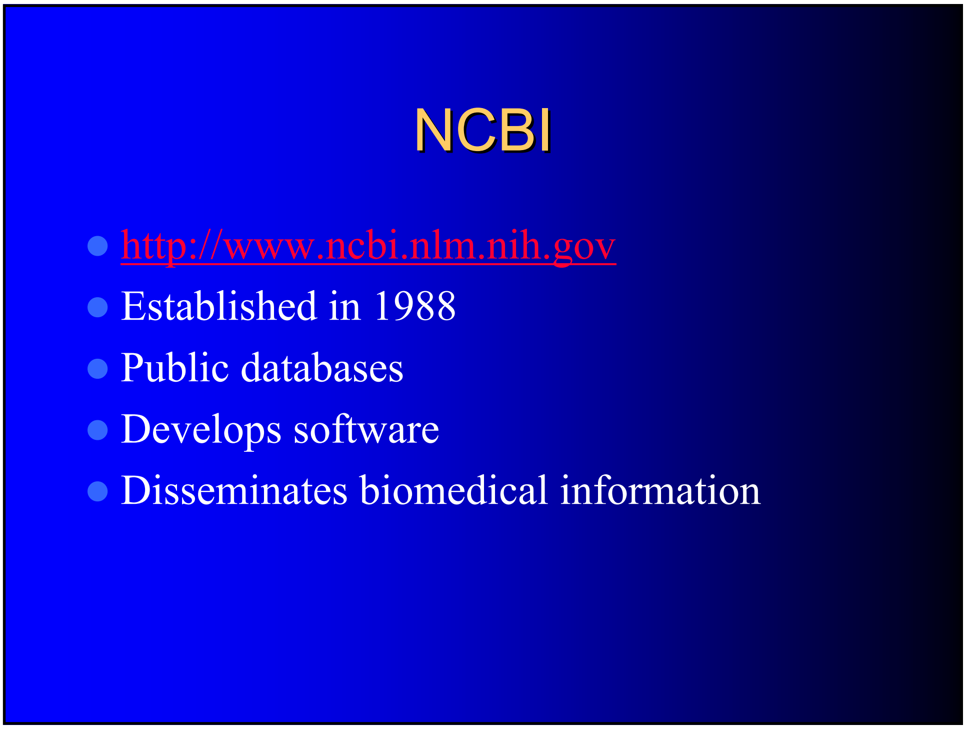
What we will cover today

- NCBI
- Genomic Databases
- UCSC
- Genomic DNA annotation



Public Genome Sequence Databases

- NCBI
  - http://www.ncbi.nlm.nih.gov/mapview/
- UCSC’s Genome Browser
  - http:genome.ucsc.edu
- Ensembl
  - http://www.ensembl.org

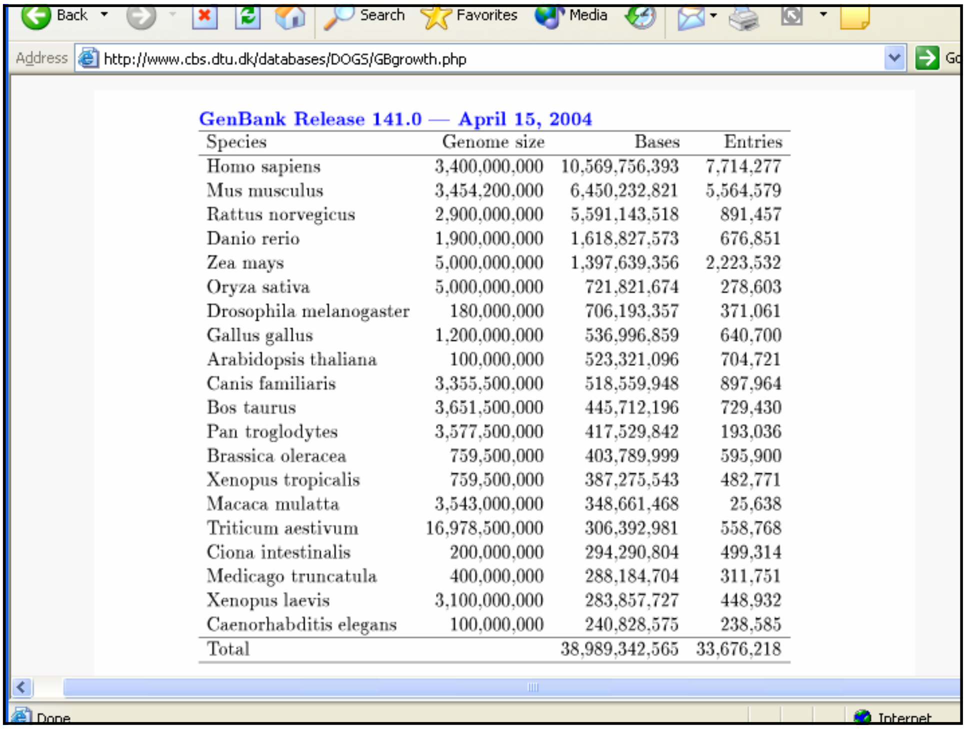


NCBI

- http://www.ncbi.nlm.nih.gov
- Established in 1988
- Public databases
- Develops software
- Disseminates biomedical information

Genomic Databases

- Sequencing of the whole genome of the organism
- Sequence must be annotated
  - Location of genes
  - Location of transcribed regions
  - Location of promoters
  - Function of motifs
  - Function of other DNA sequences



GenBank Release 141.0 — April 15, 2004

Species	Genome size	Bases	Entries
Homo sapiens	3,400,000,000	10,569,756,393	7,714,277
Mus musculus	3,454,200,000	6,450,232,821	5,564,579
Rattus norvegicus	2,900,000,000	5,591,143,518	891,457
Danio rerio	1,900,000,000	1,618,827,573	676,851
Zea mays	5,000,000,000	1,397,639,356	2,223,532
Oryza sativa	5,000,000,000	721,821,674	278,603
Drosophila melanogaster	180,000,000	706,193,357	371,061
Gallus gallus	1,200,000,000	536,996,859	640,700
Arabidopsis thaliana	100,000,000	523,321,096	704,721
Canis familiaris	3,355,500,000	518,559,948	897,964
Bos taurus	3,651,500,000	445,712,196	729,430
Pan troglodytes	3,577,500,000	417,529,842	193,036
Brassica oleracea	759,500,000	403,789,999	595,900
Xenopus tropicalis	759,500,000	387,275,543	482,771
Macaca mulatta	3,543,000,000	348,661,468	25,638
Triticum aestivum	16,978,500,000	306,392,981	558,768
Ciona intestinalis	200,000,000	294,290,804	499,314
Medicago truncatula	400,000,000	288,184,704	311,751
Xenopus laevis	3,100,000,000	283,857,727	448,932
Caenorhabditis elegans	100,000,000	240,828,575	238,585
Total		38,989,342,565	33,676,218

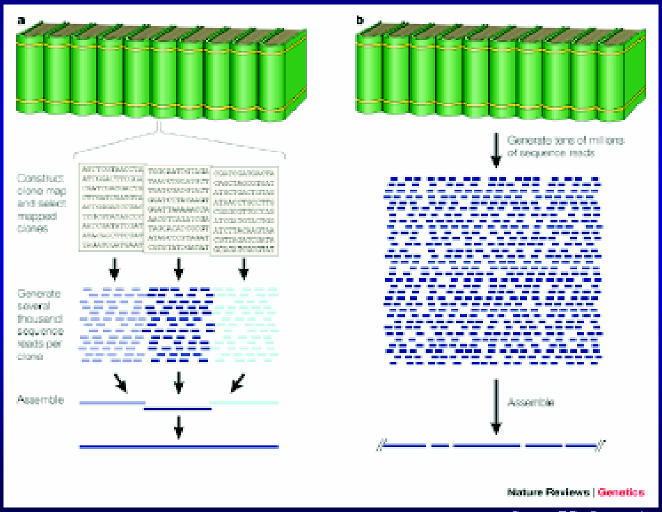
How was genomic sequence data generated?

- Clone-by-clone shotgun sequencing
- Whole-genome shotgun sequencing

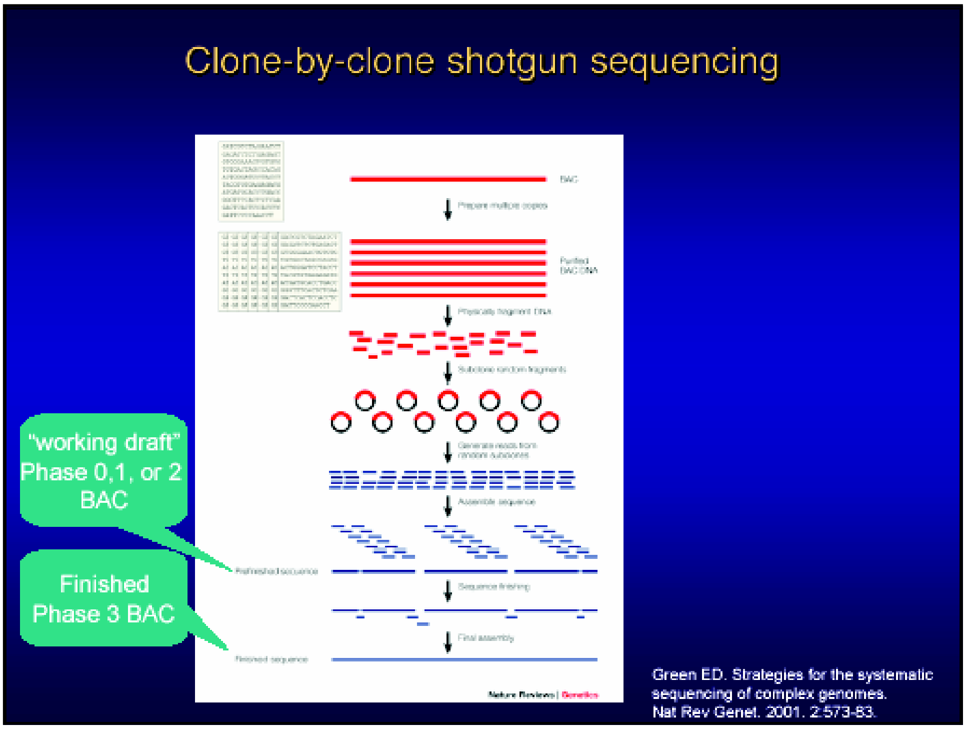
Overview of sequencing strategies

Clone-by-clone shotgun sequencing

Whole-genome shotgun sequencing


a
Construct clone map and select mapped clones
Generate several thousand sequence reads per clone
Assemble
b
Generate tens of millions of sequence reads
Assemble
Nature Reviews | Genetics

Green ED. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet. 2001. 2:573-83.

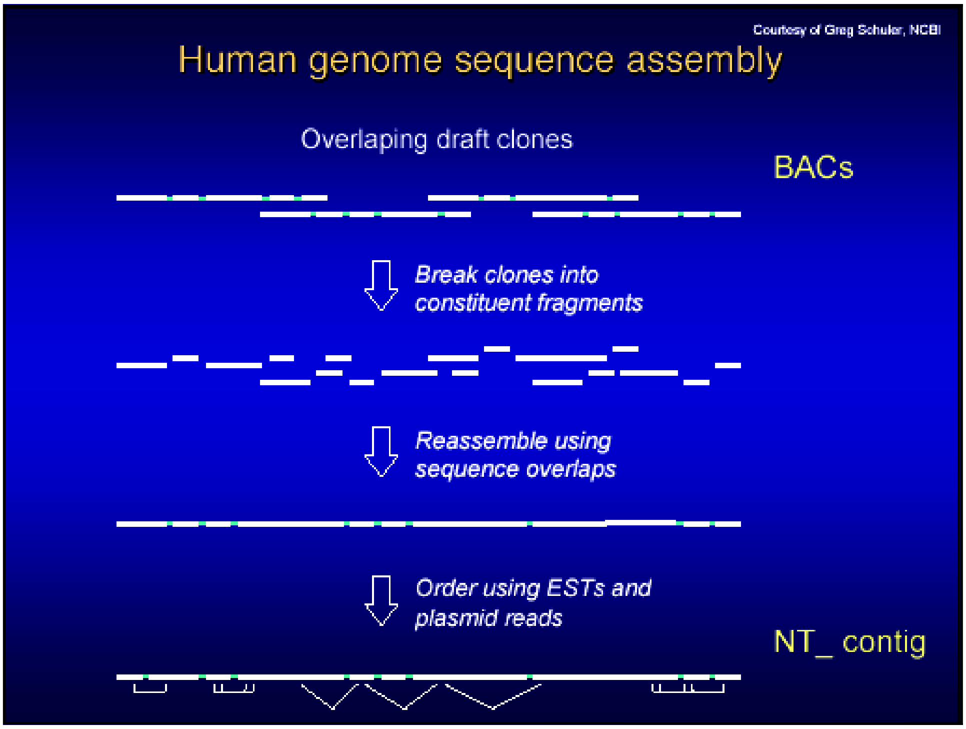


Clone-by-clone shotgun sequencing

"working draft" Phase 0,1, or 2 BAC

Finished Phase 3 BAC

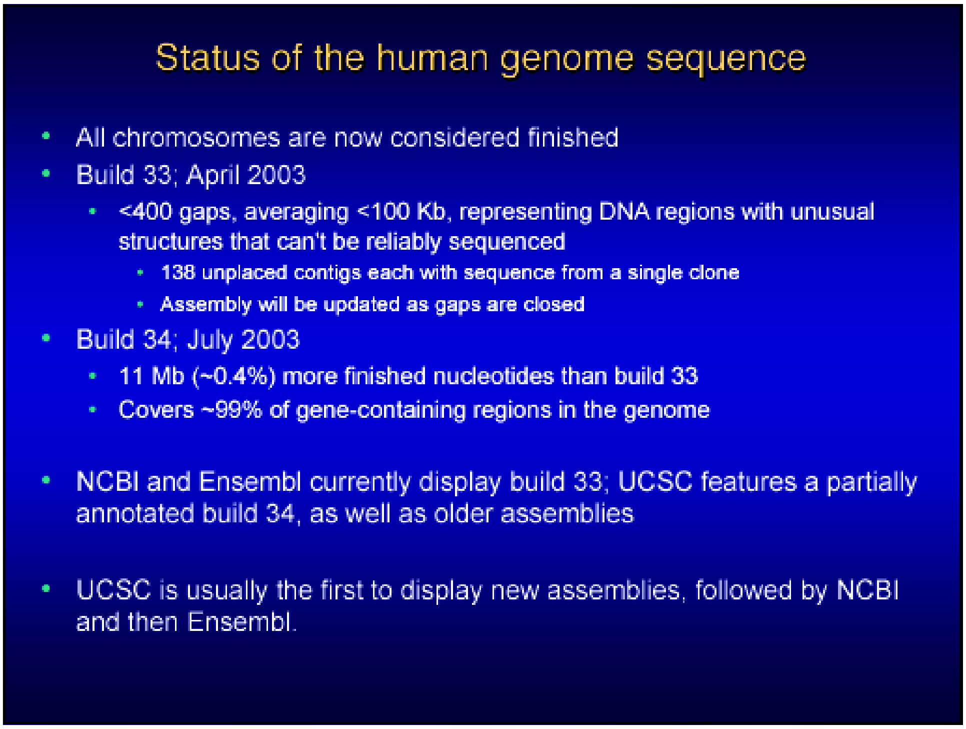
Green ED. Strategies for the systematic sequencing of complex genomes.
Nat Rev Genet. 2001. 2:573-83.



Courtesy of Greg Schuler, NCBI

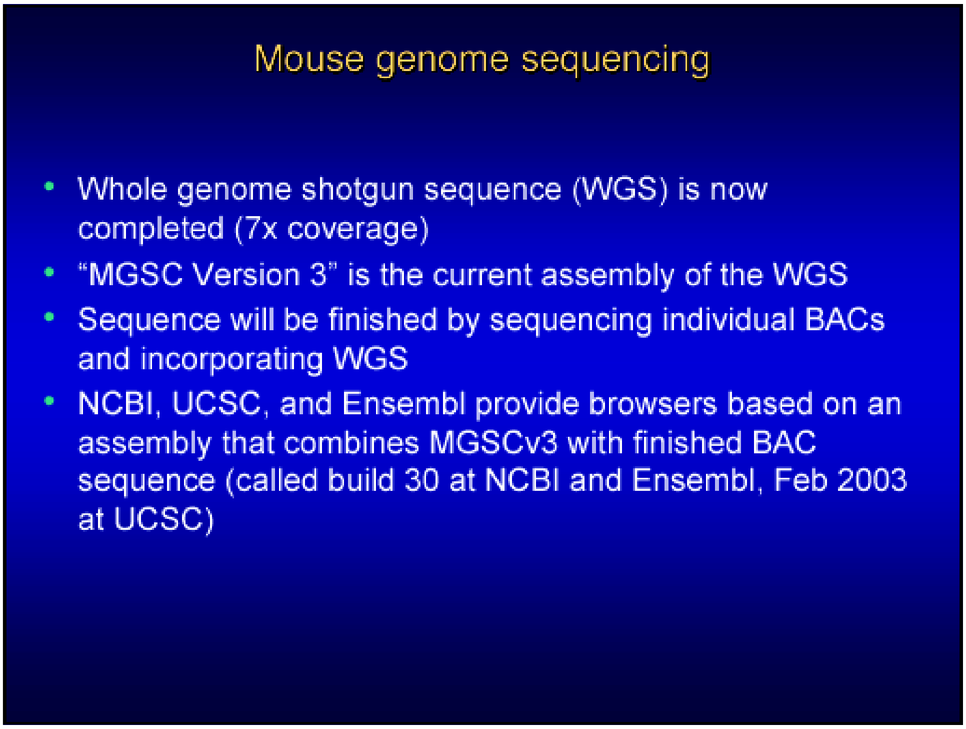
Human genome sequence assembly

Overlaping draft clones
BACs
Break clones into constituent fragments
Reassemble using sequence overlaps
Order using ESTs and plasmid reads
NT_ contig



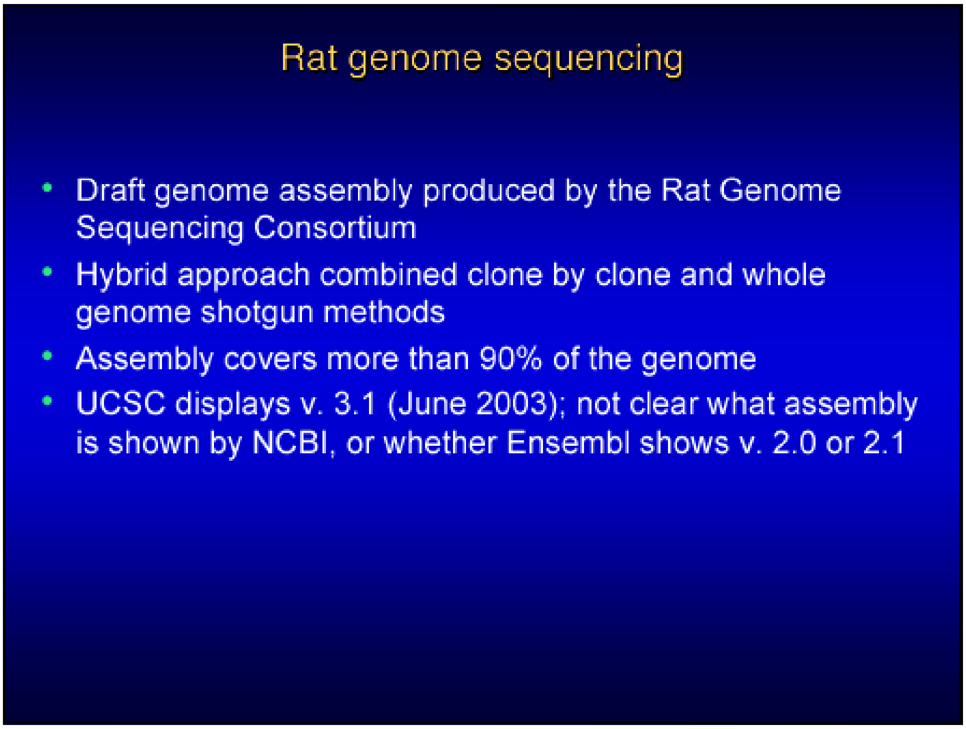
Status of the human genome sequence

- All chromosomes are now considered finished
- Build 33; April 2003
  - <400 gaps, averaging <100 Kb, representing DNA regions with unusual structures that can't be reliably sequenced
    - 138 unplaced contigs each with sequence from a single clone
    - Assembly will be updated as gaps are closed
- Build 34; July 2003
  - 11 Mb (~0.4%) more finished nucleotides than build 33
  - Covers ~99% of gene-containing regions in the genome
- NCBI and Ensembl currently display build 33; UCSC features a partially annotated build 34, as well as older assemblies
- UCSC is usually the first to display new assemblies, followed by NCBI and then Ensembl.



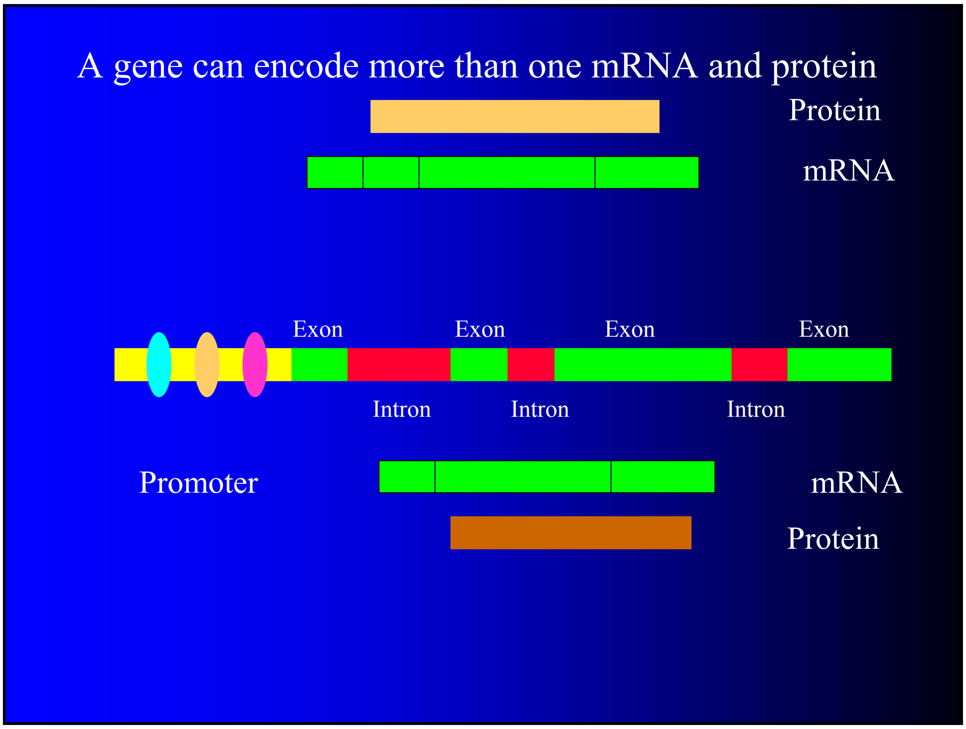
Mouse genome sequencing

- Whole genome shotgun sequence (WGS) is now completed (7x coverage)
- “MGSC Version 3” is the current assembly of the WGS
- Sequence will be finished by sequencing individual BACs and incorporating WGS
- NCBI, UCSC, and Ensembl provide browsers based on an assembly that combines MGSCv3 with finished BAC sequence (called build 30 at NCBI and Ensembl, Feb 2003 at UCSC)



Rat genome sequencing

- Draft genome assembly produced by the Rat Genome Sequencing Consortium
- Hybrid approach combined clone by clone and whole genome shotgun methods
- Assembly covers more than 90% of the genome
- UCSC displays v. 3.1 (June 2003); not clear what assembly is shown by NCBI, or whether Ensembl shows v. 2.0 or 2.1



A gene can encode more than one mRNA and protein

Protein

mRNA

Exon

Exon

Exon

Exon

Intron

Intron

Intron

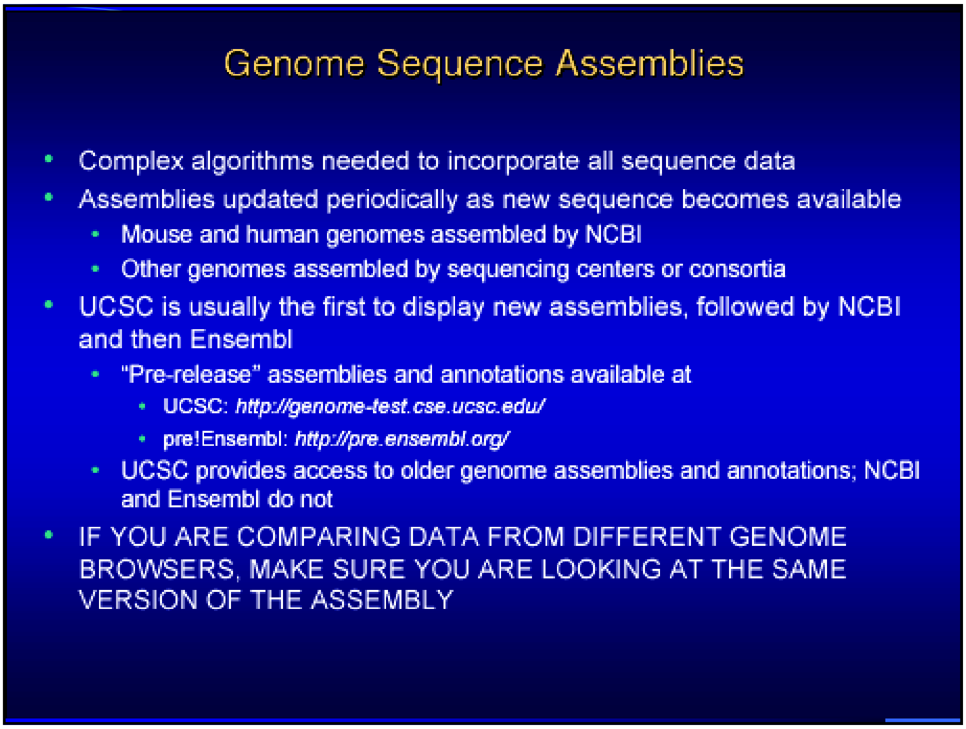
Promoter

mRNA

Protein

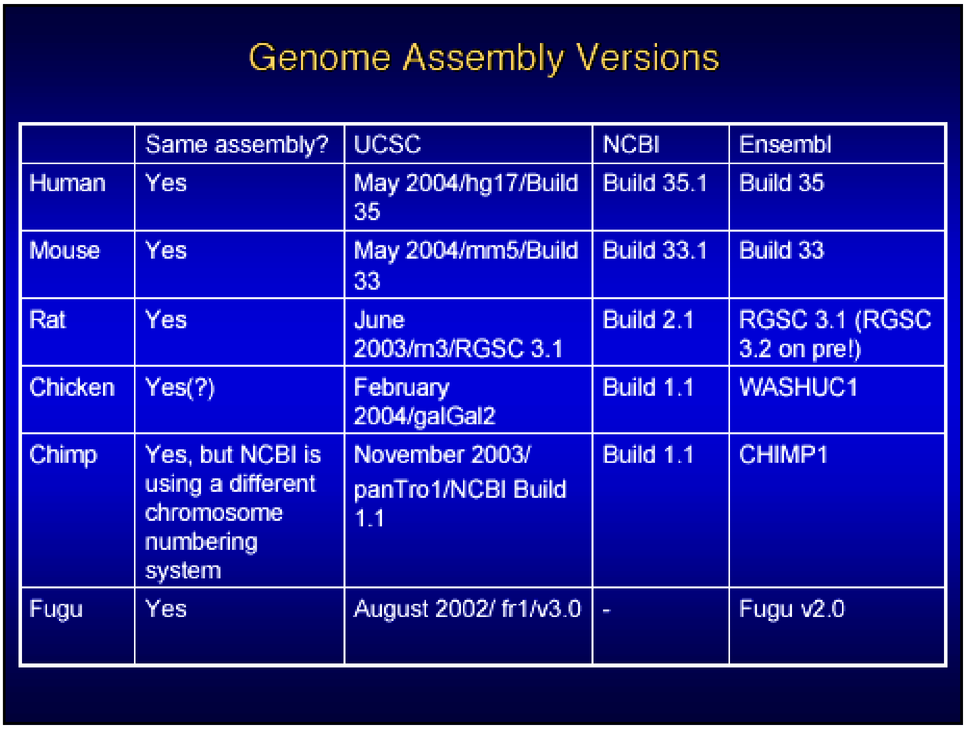
Specific Genome Databases

- Human
  - http://www.ncbi.nlm.nih.gov/genome/guide/human/
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
- Mouse
- Drosophila
- Nematode
- Arabidopsis
- Many others



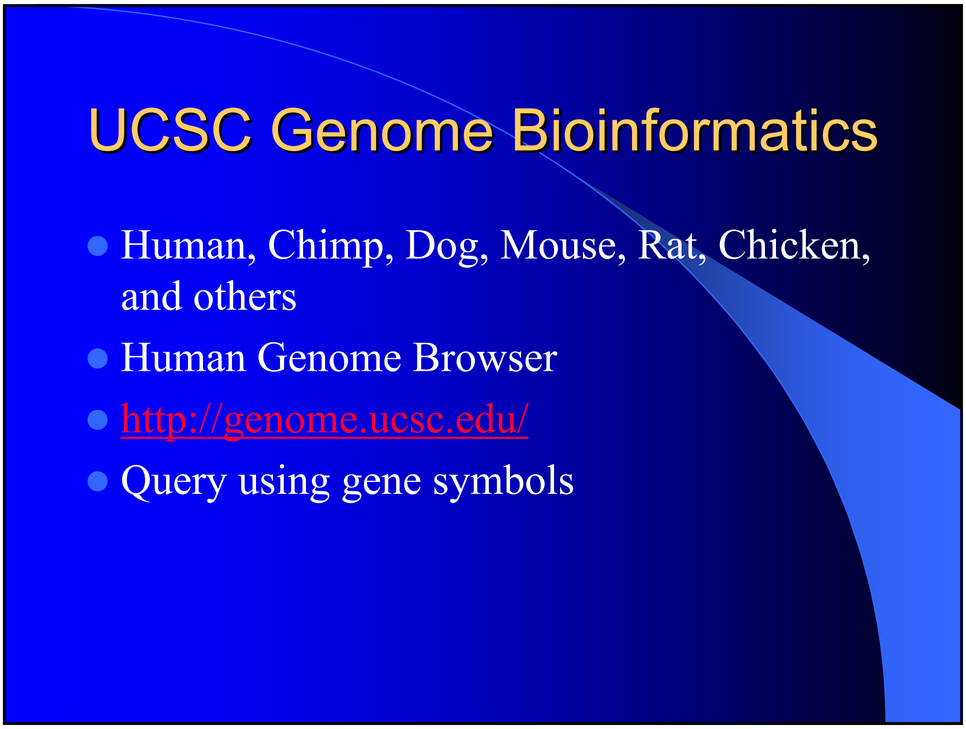
Genome Sequence Assemblies

- Complex algorithms needed to incorporate all sequence data
- Assemblies updated periodically as new sequence becomes available
  - Mouse and human genomes assembled by NCBI
  - Other genomes assembled by sequencing centers or consortia
- UCSC is usually the first to display new assemblies, followed by NCBI and then Ensembl
  - "Pre-release" assemblies and annotations available at
    - UCSC: *http://genome-test.cse.ucsc.edu/*
    - pre!Ensembl: *http://pre.ensembl.org/*
  - UCSC provides access to older genome assemblies and annotations; NCBI and Ensembl do not
- IF YOU ARE COMPARING DATA FROM DIFFERENT GENOME BROWSERS, MAKE SURE YOU ARE LOOKING AT THE SAME VERSION OF THE ASSEMBLY



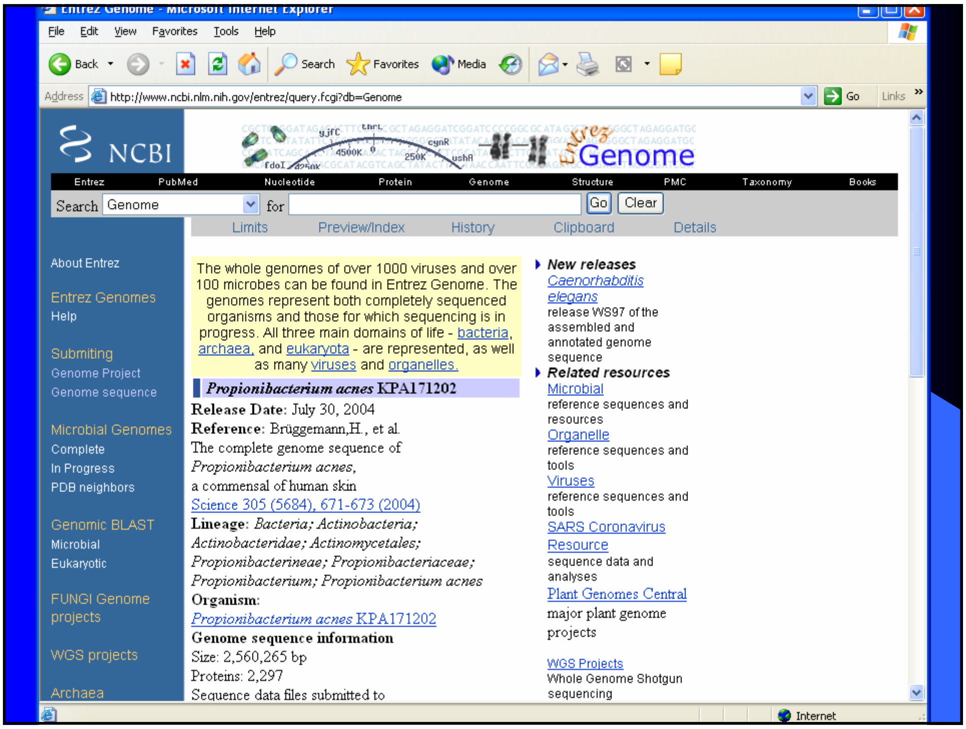
Genome Assembly Versions

	Same assembly?	UCSC	NCBI	Ensembl
Human	Yes	May 2004/hg17/Build 35	Build 35.1	Build 35
Mouse	Yes	May 2004/mm5/Build 33	Build 33.1	Build 33
Rat	Yes	June 2003/m3/RGSC 3.1	Build 2.1	RGSC 3.1 (RGSC 3.2 on pre!)
Chicken	Yes(?)	February 2004/galGal2	Build 1.1	WASHUC1
Chimp	Yes, but NCBI is using a different chromosome numbering system	November 2003/ panTro1/NCBI Build 1.1	Build 1.1	CHIMP1
Fugu	Yes	August 2002/ fr1/v3.0	-	Fugu v2.0



UCSC Genome Bioinformatics

- Human, Chimp, Dog, Mouse, Rat, Chicken, and others
- Human Genome Browser
- http://genome.ucsc.edu/
- Query using gene symbols


Entrez Genome - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Go Links
NCBI
Entrez Genome
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Genome for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Genomes
Help
Submitting
Genome Project
Genome sequence
Microbial Genomes
Complete
In Progress
PDB neighbors
Genomic BLAST
Microbial
Eukaryotic
FUNGI Genome projects
WGS projects
Archaea
Internet

The whole genomes of over 1000 viruses and over 100 microbes can be found in Entrez Genome. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses and organelles.

Propionibacterium acnes **KPA171202**

Release Date: July 30, 2004

Reference: Brüggemann,H., et al.

The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin

Science 305 (5684), 671-673 (2004)

Lineage: *Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Propionibacterineae; Propionibacteriaceae; Propionibacterium; Propionibacterium acnes*

Organism:

Propionibacterium acnes KPA171202

Genome sequence information

Size: 2,560,265 bp

Proteins: 2,297

Sequence data files submitted to

New releases

Caenorhabditis elegans release WS97 of the assembled and annotated genome sequence

Related resources

Microbial
reference sequences and resources

Organelle
reference sequences and tools

Viruses
reference sequences and tools

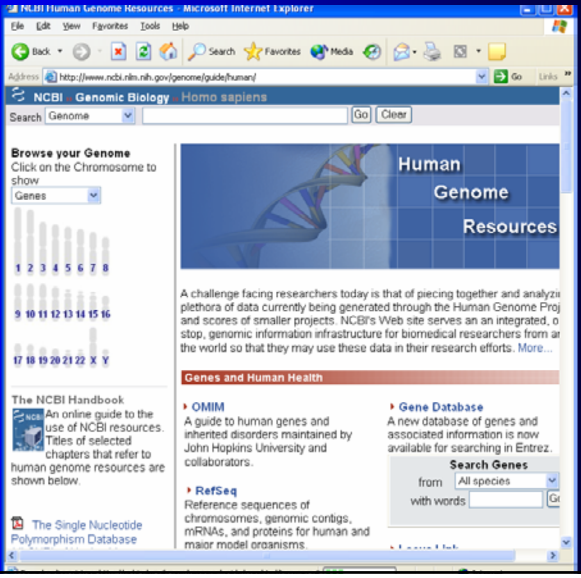
SARS Coronavirus Resource
sequence data and analyses

Plant Genomes Central
major plant genome projects

WGS Projects
Whole Genome Shotgun sequencing

Human

Humanhttp://www.ncbi.nlm.nih.gov/genome/guide/human/


NCBI Human Genome Resources - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/genome/guide/human/
Go Links
NCBI Genomic Biology Homo sapiens
Search Genome Go Clear
Genes
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 X Y
Human Genome Resources
Search Genes
from All species
with words

Browse your Genome
Click on the Chromosome to show

A challenge facing researchers today is that of piecing together and analyzi plethora of data currently being generated through the Human Genome Proj and scores of smaller projects. NCBI's Web site serves an an integrated, o stop, genomic information infrastructure for biomedical researchers from ar the world so that they may use these data in their research efforts. More...

The NCBI Handbook
An online guide to the use of NCBI resources. Titles of selected chapters that refer to human genome resources are shown below.

The Single Nucleotide Polymorphism Database

Genes and Human Health

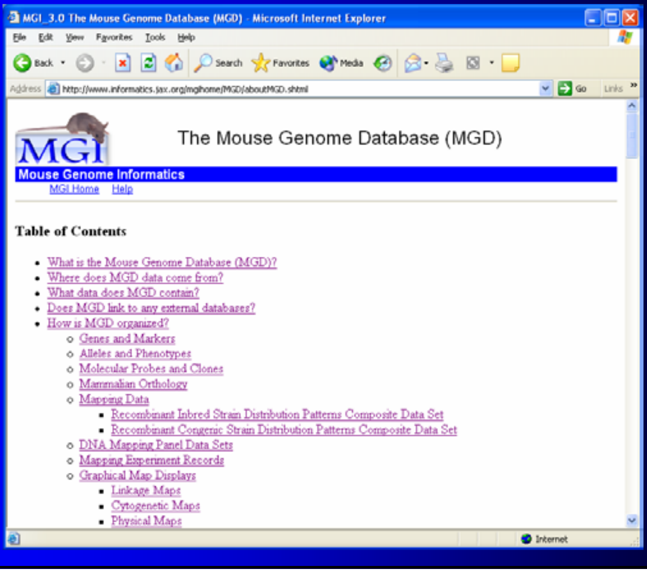
OMIM
A guide to human genes and inherited disorders maintained by John Hopkins University and collaborators.

RefSeq
Reference sequences of chromosomes, genomic contigs, mRNAs, and proteins for human and major model organisms.

Gene Database
A new database of genes and associated information is now available for searching in Entrez.

Mouse

http://www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml


MGI_3.0 The Mouse Genome Database (MGD) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Internet

Address: http://www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml

MGI

The Mouse Genome Database (MGD)

Mouse Genome Informatics

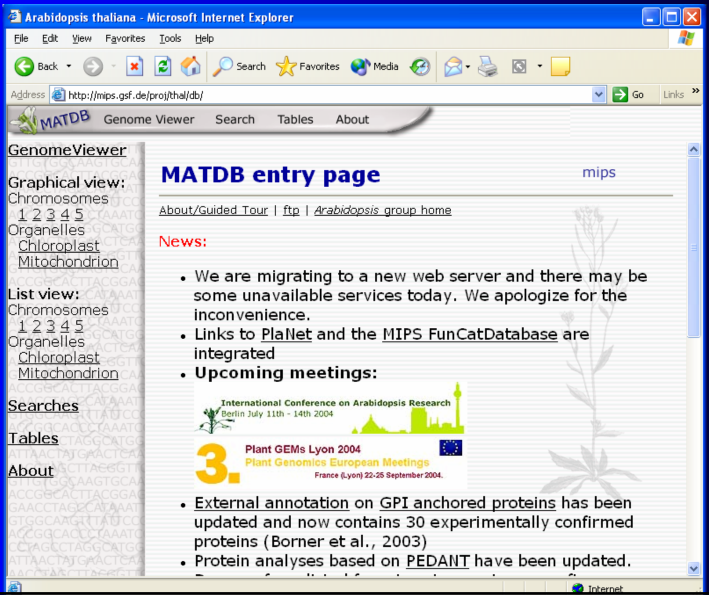
MGI Home Help

Table of Contents

- What is the Mouse Genome Database (MGD)?
- Where does MGD data come from?
- What data does MGD contain?
- Does MGD link to any external databases?
- How is MGD organized?
  - Genes and Markers
  - Alleles and Phenotypes
  - Molecular Probes and Clones
  - Mammalian Orthology
  - Mapping Data
    - Recombinant Inbred Strain Distribution Patterns Composite Data Set
    - Recombinant Congenic Strain Distribution Patterns Composite Data Set
  - DNA Mapping Panel Data Sets
  - Mapping Experiment Records
  - Graphical Map Displays
    - Linkage Maps
    - Cytogenetic Maps
    - Physical Maps

Arabidopsis

http://mips.gsf.de/proj/thal/db/


Arabidopsis thaliana - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
MATDB
Genome Viewer Search Tables About
Internet

Address: http://mips.gsf.de/proj/thal/db/

GenomeViewer

Graphical view:
Chromosomes
1 2 3 4 5
Organelles
Chloroplast
Mitochondrion

List view:
Chromosomes
1 2 3 4 5
Organelles
Chloroplast
Mitochondrion

Searches

Tables

About

MATDB entry page

mips

About/Guided Tour | ftp | *Arabidopsis* group home

News:

- We are migrating to a new web server and there may be some unavailable services today. We apologize for the inconvenience.
- Links to PlaNet and the MIPS FunCatDatabase are integrated
- **Upcoming meetings:**

International Conference on Arabidopsis Research
Berlin July 11th - 14th 2004

3. Plant GEMs Lyon 2004
Plant Genomics European Meetings
France (Lyon) 22-25 September 2004.

- External annotation on GPI anchored proteins has been updated and now contains 30 experimentally confirmed proteins (Borner et al., 2003)
- Protein analyses based on PEDANT have been updated.

Stanford Genomic Resources

- http://genome-www.stanford.edu/
- Saccharomyces
- Microarrays
- Arabidopsis
- Human, Mouse, Rat
- Candida
- Tetrahymena

UCSC

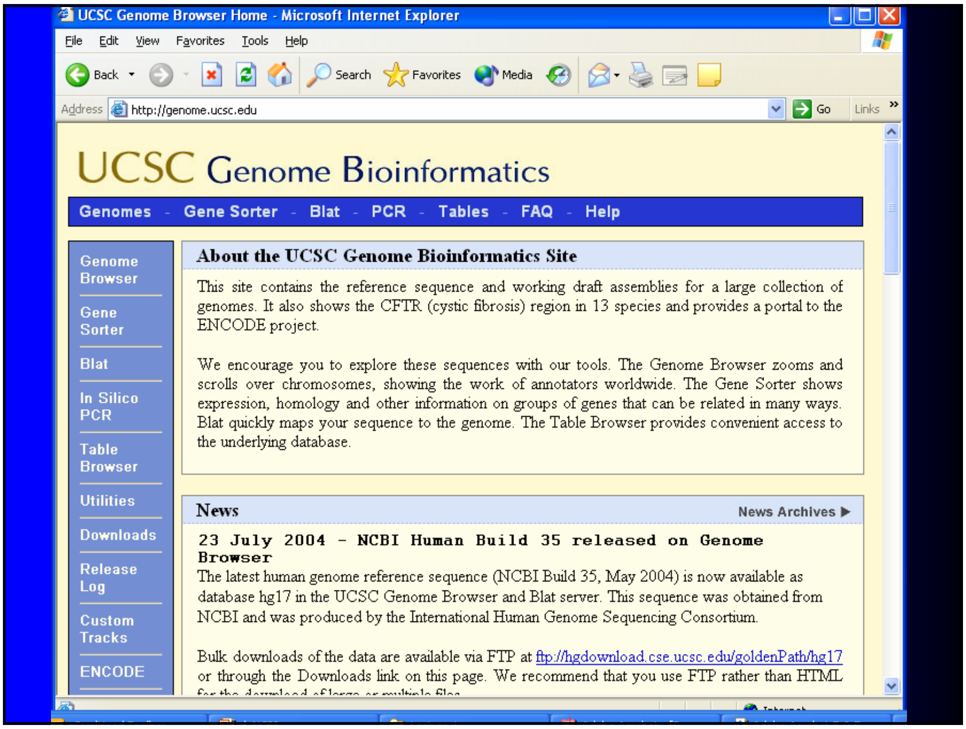
http://genome.ucsc.edu

Genomes available in database

- Human
- Chimp
- Dog
- Rat
- Chicken
- Drosophila
- C. elegans
- Yeast
- Others

UCSC Genome Bioinformatics

- Genomes
- Gene Sorter
  - Searches for related genes
- BLAT Search
  - Paste in query sequence to find it location in the genome
- In-Silico PCR
  - Searches sequence database with PCR primers
- Can download portions of database
- Encode
  - Information on function of DNA sequences



UCSC Genome Bioinformatics

Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Genome Browser
Gene Sorter
Blat
In Silico PCR
Table Browser
Utilities
Downloads
Release Log
Custom Tracks
ENCODE

About the UCSC Genome Bioinformatics Site

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also shows the CFTR (cystic fibrosis) region in 13 species and provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database.

News News Archives ▶

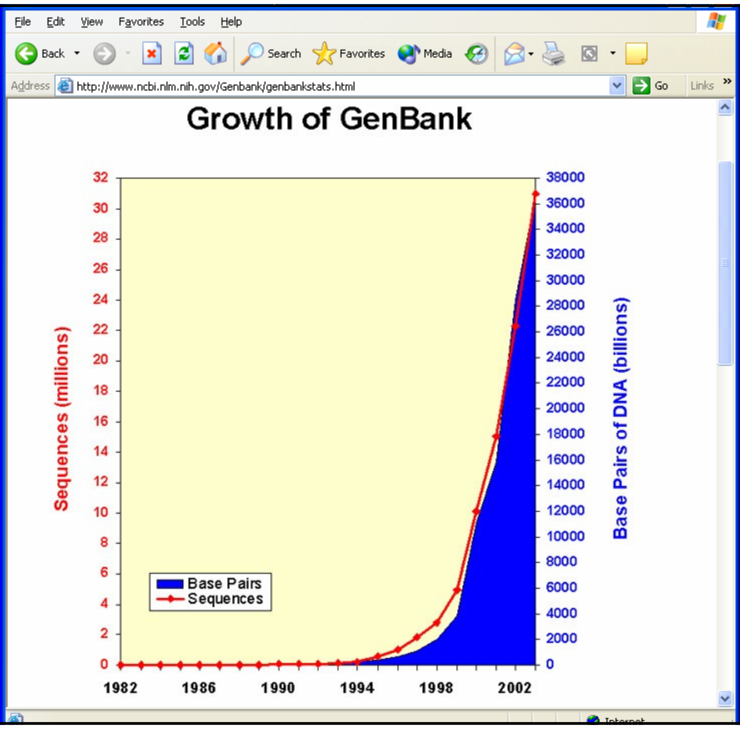
23 July 2004 - NCBI Human Build 35 released on Genome Browser

The latest human genome reference sequence (NCBI Build 35, May 2004) is now available as database hg17 in the UCSC Genome Browser and Blat server. This sequence was obtained from NCBI and was produced by the International Human Genome Sequencing Consortium.

Bulk downloads of the data are available via FTP at ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17 or through the Downloads link on this page. We recommend that you use FTP rather than HTML for the download of large or multiple files.

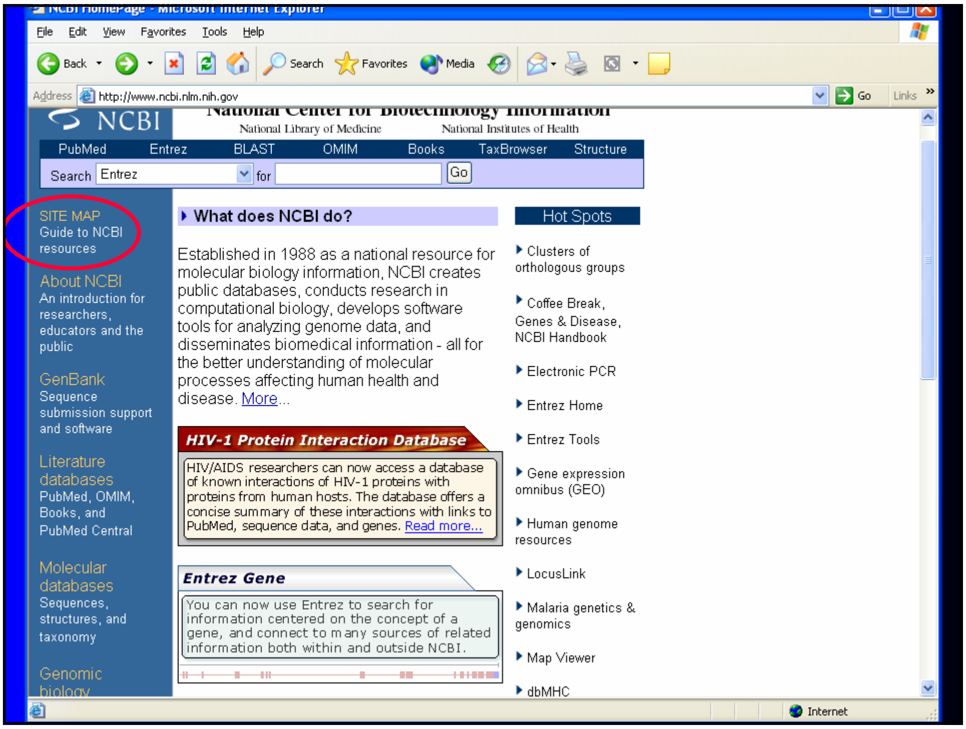
GenBank

- http://www.ncbi.nlm.nih.gov/Genbank/index.html
- Nucleotide sequences
- >130,000 organisms
- Annotated records with coding region features and amino acid translations



Address http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Growth of GenBank


NCBI HomePage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov
Go Links
Internet

NCBI National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search Entrez for Go

SITE MAP
Guide to NCBI resources

About NCBI
An introduction for researchers, educators and the public

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

HIV-1 Protein Interaction Database

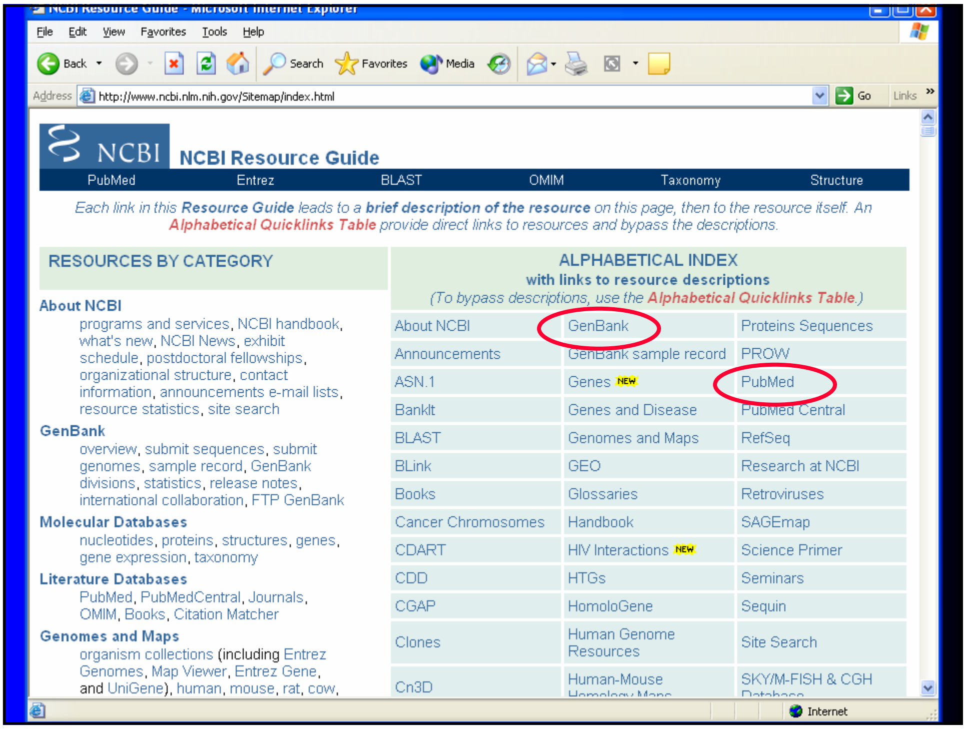
HIV/AIDS researchers can now access a database of known interactions of HIV-1 proteins with proteins from human hosts. The database offers a concise summary of these interactions with links to PubMed, sequence data, and genes. Read more...

Entrez Gene

You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

Hot Spots

- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- LocusLink
- Malaria genetics & genomics
- Map Viewer
- dbMHC


NCBI Resource Guide - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/Sitemap/index.html
Go Links
Internet

NCBI **NCBI Resource Guide**

PubMed Entrez BLAST OMIM Taxonomy Structure

Each link in this ***Resource Guide*** *leads to a* ***brief description of the resource*** *on this page, then to the resource itself. An* ***Alphabetical Quicklinks Table*** *provide direct links to resources and bypass the descriptions.*

RESOURCES BY CATEGORY

About NCBI
programs and services, NCBI handbook, what's new, NCBI News, exhibit schedule, postdoctoral fellowships, organizational structure, contact information, announcements e-mail lists, resource statistics, site search

GenBank
overview, submit sequences, submit genomes, sample record, GenBank divisions, statistics, release notes, international collaboration, FTP GenBank

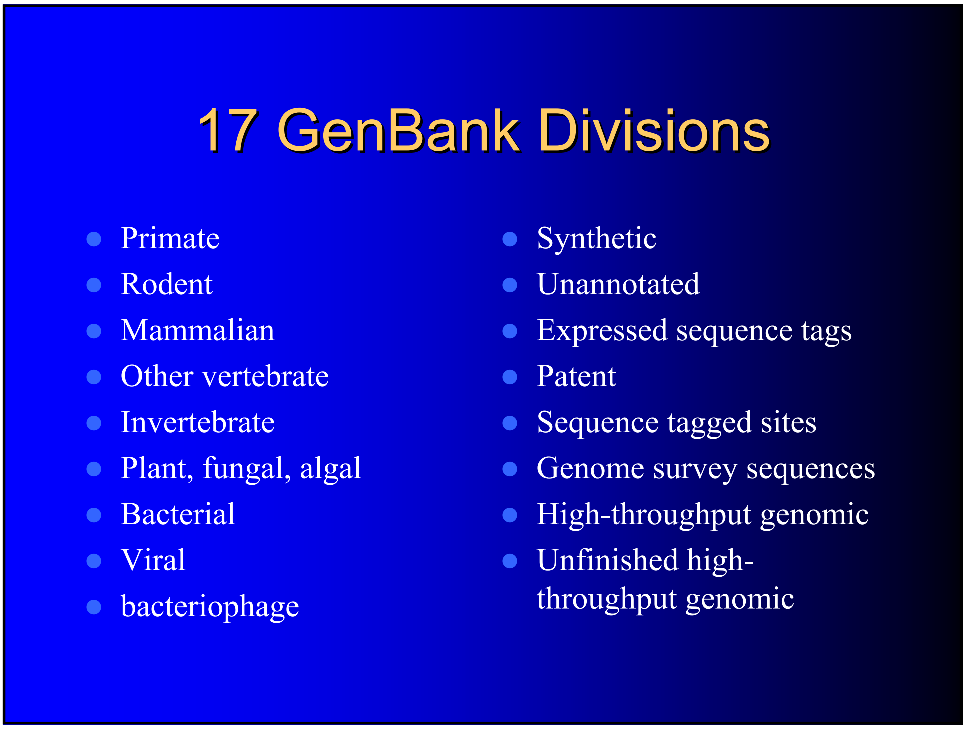
Molecular Databases
nucleotides, proteins, structures, genes, gene expression, taxonomy

Literature Databases
PubMed, PubMedCentral, Journals, OMIM, Books, Citation Matcher

Genomes and Maps
organism collections (including Entrez Genomes, Map Viewer, Entrez Gene, and UniGene), human, mouse, rat, cow,

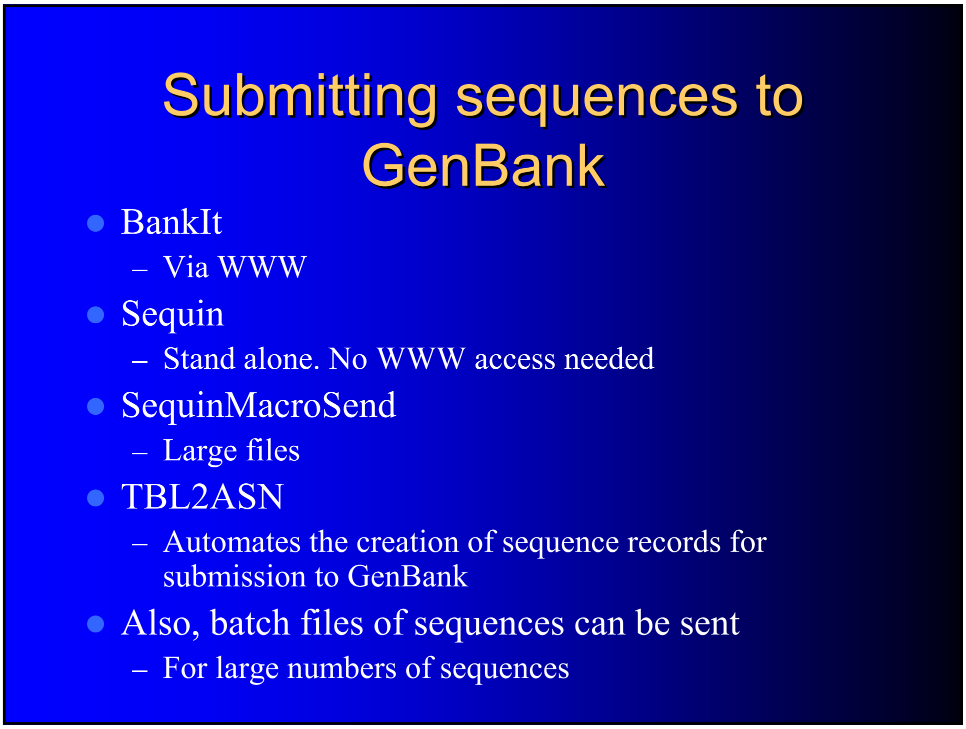
ALPHABETICAL INDEX
with links to resource descriptions
(To bypass descriptions, use the ***Alphabetical Quicklinks Table****.)*

About NCBI	GenBank	Proteins Sequences
Announcements	GenBank sample record	PROW
ASN.1	Genes NEW	PubMed
BankIt	Genes and Disease	PubMed Central
BLAST	Genomes and Maps	RefSeq
BLink	GEO	Research at NCBI
Books	Glossaries	Retroviruses
Cancer Chromosomes	Handbook	SAGEmap
CDART	HIV Interactions NEW	Science Primer
CDD	HTGs	Seminars
CGAP	HomoloGene	Sequin
Clones	Human Genome Resources	Site Search
Cn3D	Human-Mouse Homology Maps	SKY/M-FISH & CGH Database



17 GenBank Divisions

- Primate
- Rodent
- Mammalian
- Other vertebrate
- Invertebrate
- Plant, fungal, algal
- Bacterial
- Viral
- bacteriophage
- Synthetic
- Unannotated
- Expressed sequence tags
- Patent
- Sequence tagged sites
- Genome survey sequences
- High-throughput genomic
- Unfinished high-throughput genomic



Submitting sequences to GenBank

- BankIt
  - Via WWW
- Sequin
  - Stand alone. No WWW access needed
- SequinMacroSend
  - Large files
- TBL2ASN
  - Automates the creation of sequence records for submission to GenBank
- Also, batch files of sequences can be sent
  - For large numbers of sequences

Use BankIt if:

- you have one or a few sequence submissions
- you prefer to use a WWW-based submission tool
- your sequence annotation is not complicated
- you do not require sequence analysis tools to submit your sequence(s)

Use Sequin if:

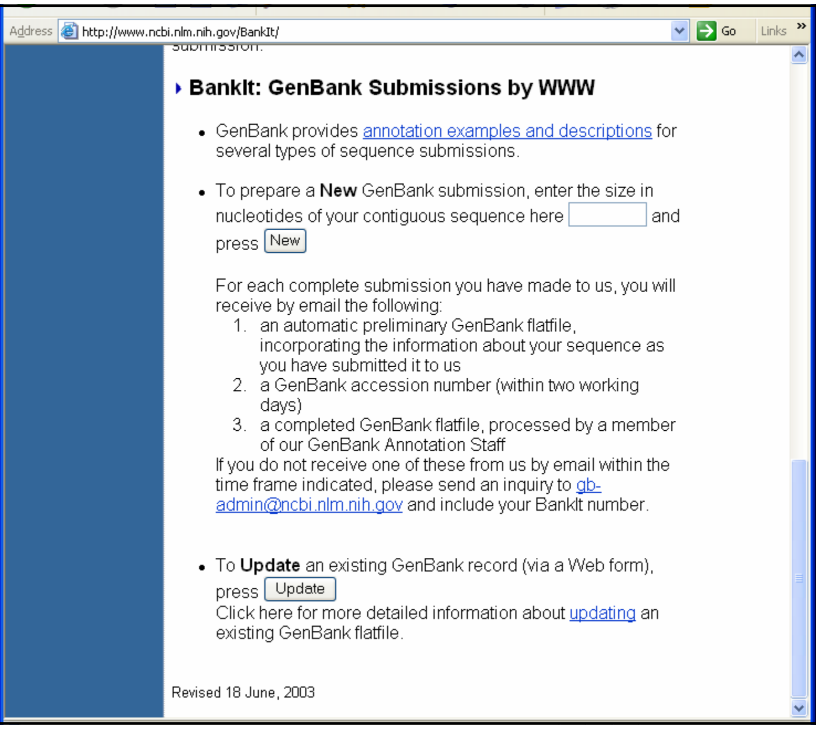
- you are submitting long or complex submissions
- you are submitting mutation, phylogenetic, population, environmental, or segmented sets
- you would like graphical viewing and editing options, including the alignment editor
- you would like network access to related analytical tools

At this time the following types of submissions are NOT acceptable:

- sequences of less than 50 bp in length
- a genomic sequence of multiple exons joined together without the sequence of the intervening introns
- primer only sequences
- protein only sequences
- non-biologically contiguous sequences containing internal unsequenced spacers
- sequences containing a mix of genomic and mRNA sequence represented as a single sequence
- Expressed Sequence Tag (EST) submissions (should be submitted through the dbEST system)
- Genome Survey Sequence (GSS) submissions (should be submitted through the dbGSS system)

BankIt

- http://www.ncbi.nlm.nih.gov/BankIt/
- Submit by WWW
- New submission
- Update an existing GenBank record



submission.

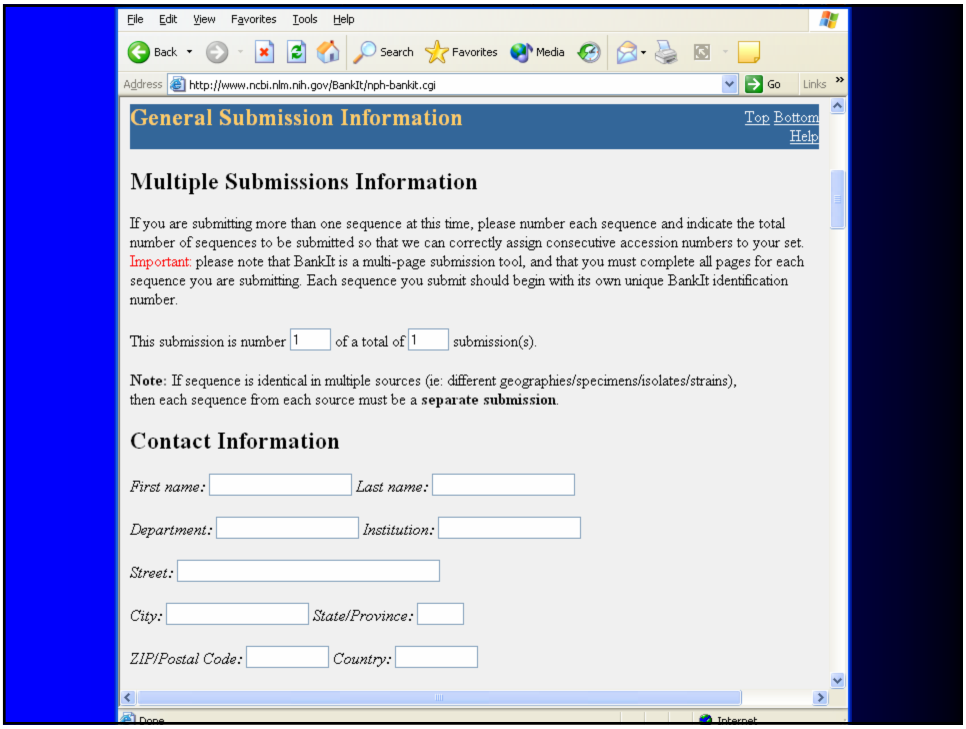
BankIt: GenBank Submissions by WWW

- GenBank provides annotation examples and descriptions for several types of sequence submissions.
- To prepare a **New** GenBank submission, enter the size in nucleotides of your contiguous sequence here ______ and press New

  For each complete submission you have made to us, you will receive by email the following:
  1. an automatic preliminary GenBank flatfile, incorporating the information about your sequence as you have submitted it to us
  2. a GenBank accession number (within two working days)
  3. a completed GenBank flatfile, processed by a member of our GenBank Annotation Staff

  If you do not receive one of these from us by email within the time frame indicated, please send an inquiry to gb-admin@ncbi.nlm.nih.gov and include your BankIt number.
- To **Update** an existing GenBank record (via a Web form), press Update
  Click here for more detailed information about updating an existing GenBank flatfile.

Revised 18 June, 2003


File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://www.ncbi.nlm.nih.gov/BankIt/nph-bankit.cgi
Go
Links
Top Bottom
Help
Done
Internet

General Submission Information

Multiple Submissions Information

If you are submitting more than one sequence at this time, please number each sequence and indicate the total number of sequences to be submitted so that we can correctly assign consecutive accession numbers to your set. Important: please note that BankIt is a multi-page submission tool, and that you must complete all pages for each sequence you are submitting. Each sequence you submit should begin with its own unique BankIt identification number.

This submission is number 1 of a total of 1 submission(s).

Note: If sequence is identical in multiple sources (ie: different geographies/specimens/isolates/strains), then each sequence from each source must be a **separate submission**.

Contact Information

First name: ______ *Last name:* ______

Department: ______ *Institution:* ______

Street: ______

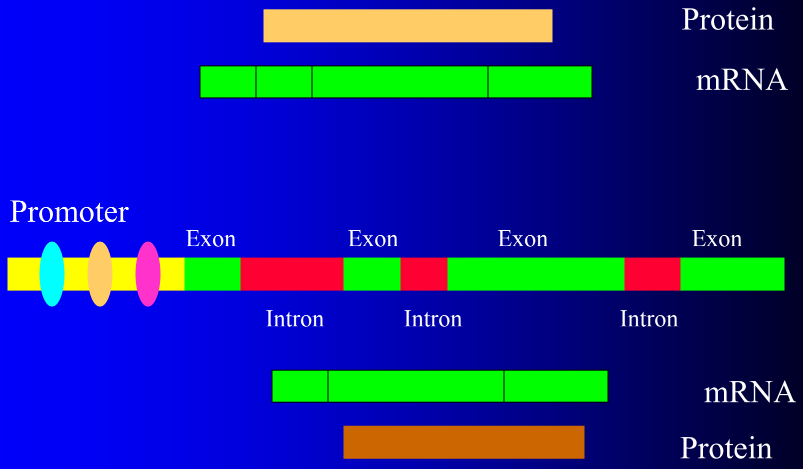
City: ______ *State/Province:* ______

ZIP/Postal Code: ______ *Country:* ______

Analysis of Genomic DNA sequences

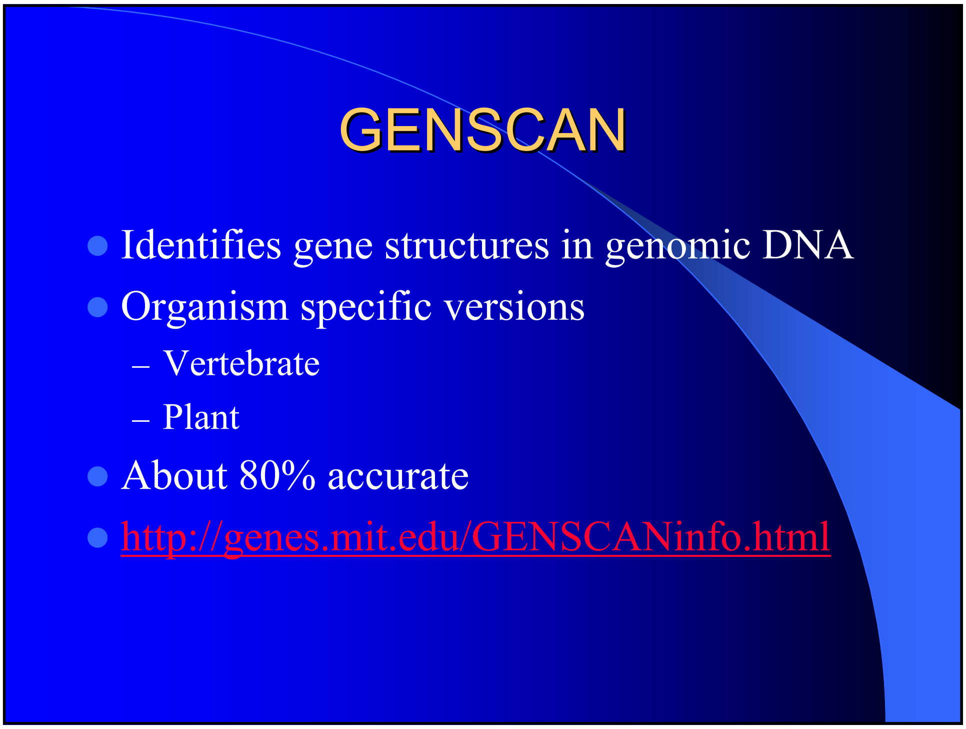
- You cloned a large piece of genomic DNA
- How will you annotate it
- Identify and describe introns, exons, promoters

A gene can encode more than one mRNA and protein



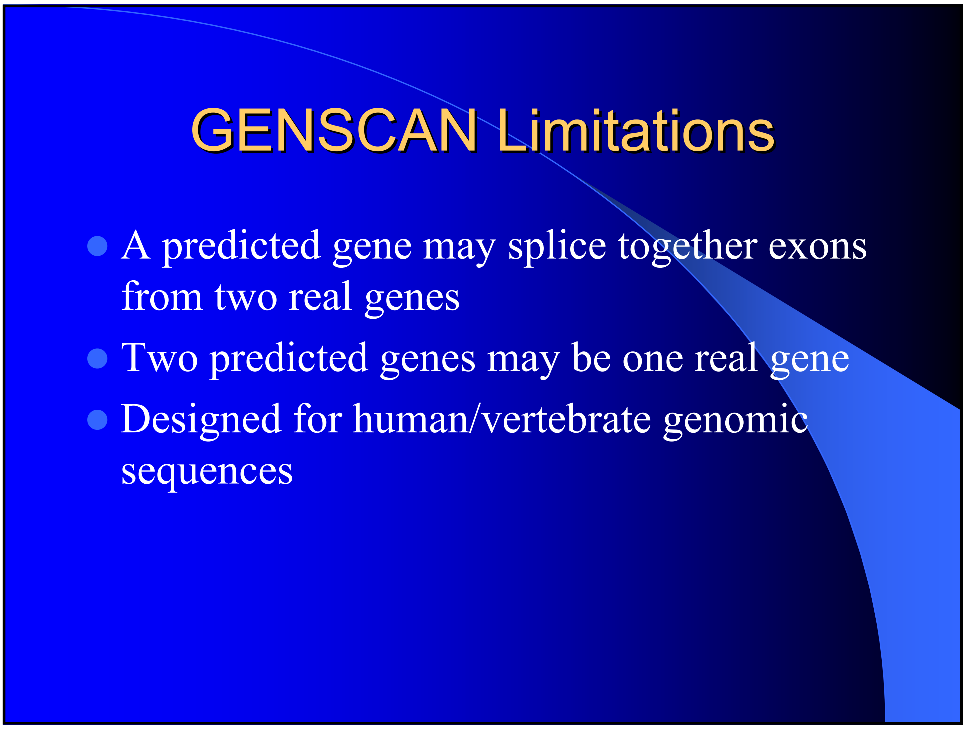
Software for genomic DNA analysis

- GeneScan
- GLIMMER
- GeneMark
- FGENE
- GRAIL
- FEX
- FGENESP



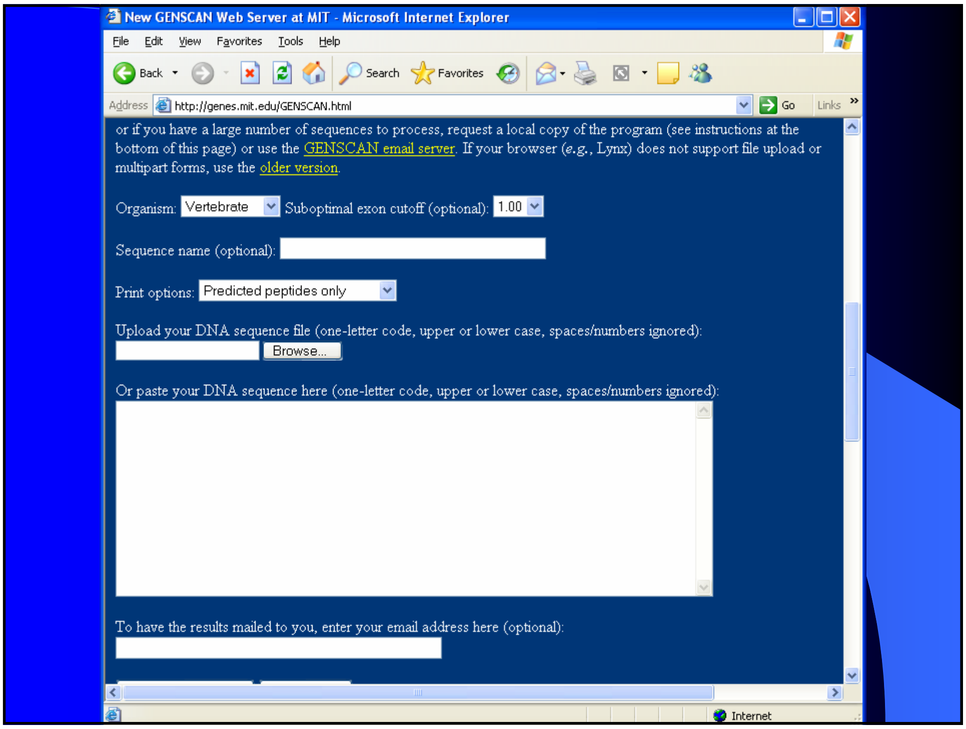
GENSCAN

- Identifies gene structures in genomic DNA
- Organism specific versions
  - Vertebrate
  - Plant
- About 80% accurate
- http://genes.mit.edu/GENSCANinfo.html



GENSCAN Limitations

- A predicted gene may splice together exons from two real genes
- Two predicted genes may be one real gene
- Designed for human/vertebrate genomic sequences


New GENSCAN Web Server at MIT - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://genes.mit.edu/GENSCAN.html
Go
Links
Internet

or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

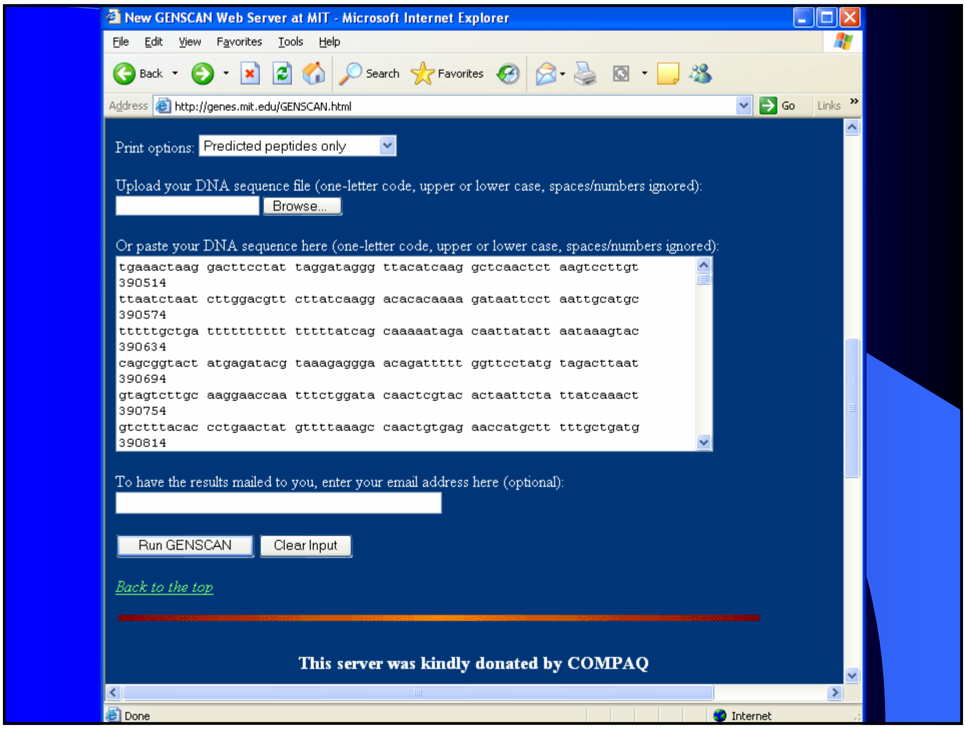
Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):
Browse...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

To have the results mailed to you, enter your email address here (optional):


New GENSCAN Web Server at MIT - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://genes.mit.edu/GENSCAN.html
Go
Links
Done
Internet

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):
Browse...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

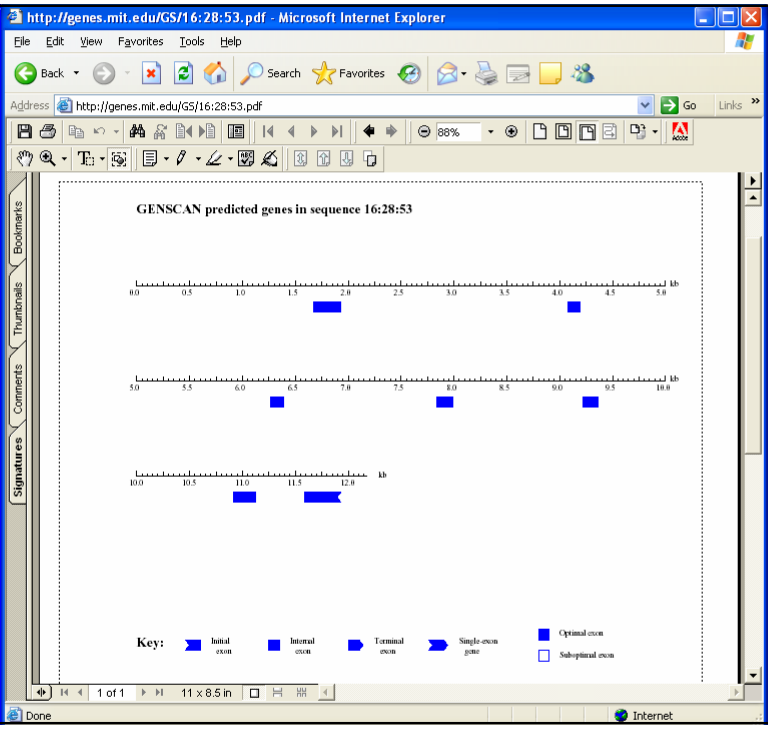
```
tgaaactaag gacttcctat taggataggg ttacatcaag gctcaactct aagtccttgt
390514
ttaatctaat cttggacgtt cttatcaagg acacacaaaa gataattcct aattgcatgc
390574
tttttgctga tttttttttt tttttatcag caaaaataga caattatatt aataaagtac
390634
cagcggtact atgagatacg taaagaggga acagattttt ggttcctatg tagacttaat
390694
gtagtcttgc aaggaaccaa tttctggata caactcgtac actaattcta ttatcaaact
390754
gtctttacac cctgaactat gttttaaagc caactgtgag aaccatgctt tttgctgatg
390814
```

To have the results mailed to you, enter your email address here (optional):

Run GENSCAN Clear Input

Back to the top

This server was kindly donated by COMPAQ

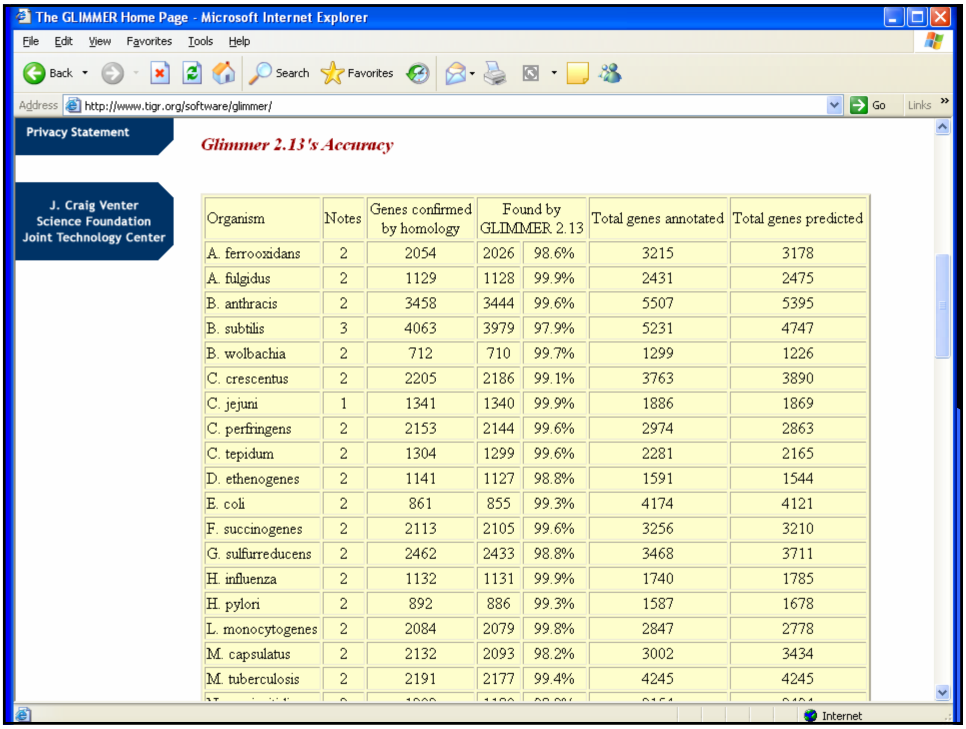


GENSCAN predicted genes in sequence 16:28:53

Key: Initial exon; Internal exon; Terminal exon; Single-exon gene; ■ Optimal exon; □ Suboptimal exon

GLIMMER

- Microbial DNA
- http://www.tigr.org/software/glimmer/


The GLIMMER Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.tigr.org/software/glimmer/
Go
Links
Privacy Statement
J. Craig Venter Science Foundation Joint Technology Center
Internet

Glimmer 2.13's Accuracy

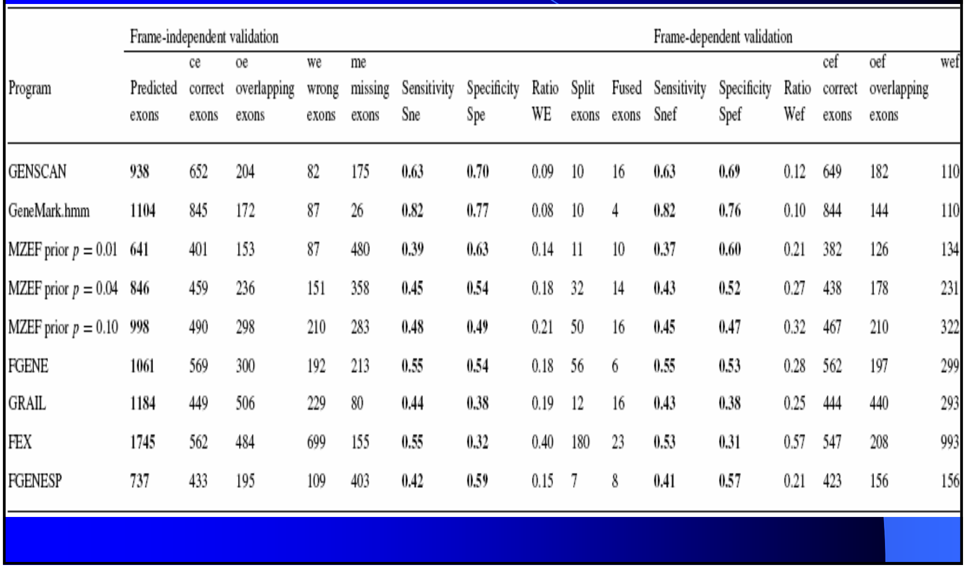
Organism	Notes	Genes confirmed by homology	Found by GLIMMER 2.13		Total genes annotated	Total genes predicted
A. ferrooxidans	2	2054	2026	98.6%	3215	3178
A. fulgidus	2	1129	1128	99.9%	2431	2475
B. anthracis	2	3458	3444	99.6%	5507	5395
B. subtilis	3	4063	3979	97.9%	5231	4747
B. wolbachia	2	712	710	99.7%	1299	1226
C. crescentus	2	2205	2186	99.1%	3763	3890
C. jejuni	1	1341	1340	99.9%	1886	1869
C. perfringens	2	2153	2144	99.6%	2974	2863
C. tepidum	2	1304	1299	99.6%	2281	2165
D. ethenogenes	2	1141	1127	98.8%	1591	1544
E. coli	2	861	855	99.3%	4174	4121
F. succinogenes	2	2113	2105	99.6%	3256	3210
G. sulfurreducens	2	2462	2433	98.8%	3468	3711
H. influenza	2	1132	1131	99.9%	1740	1785
H. pylori	2	892	886	99.3%	1587	1678
L. monocytogenes	2	2084	2079	99.8%	2847	2778
M. capsulatus	2	2132	2093	98.2%	3002	3434
M. tuberculosis	2	2191	2177	99.4%	4245	4245

GeneMark

- A family of gene prediction programs
- Bacteria
- Eukaryotes
- Viruses
- http://genes.mit.edu/GENSCANinfo.html

Eukaryotic GeneMark Accuracy

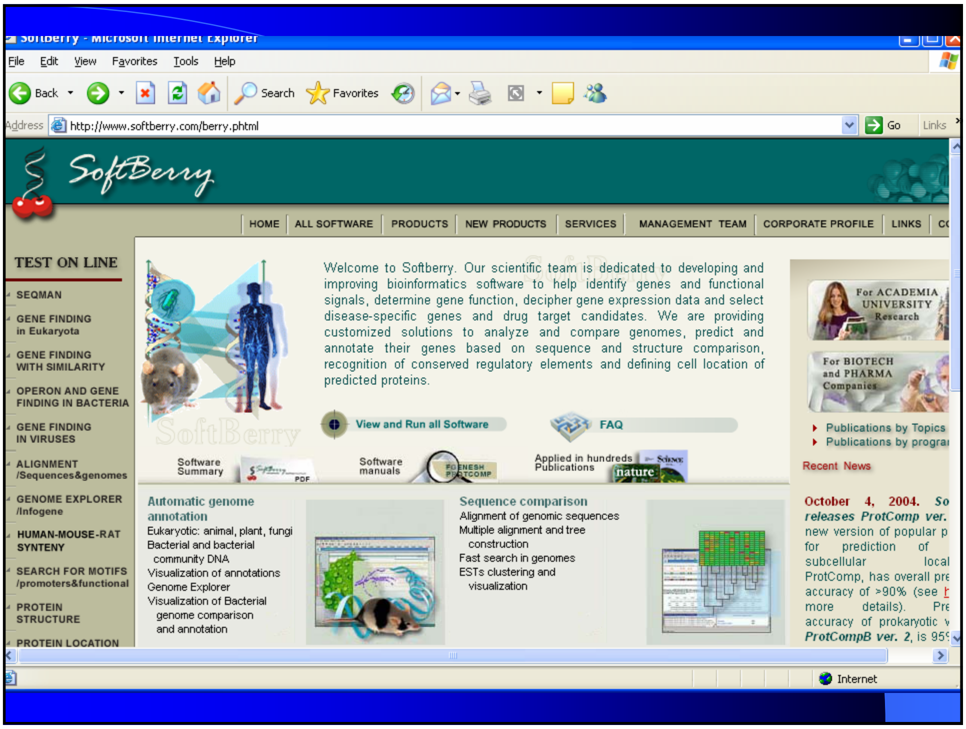
Arabidopsis thaliana Gene structure prediction



	Frame-independent validation										Frame-dependent validation					
		ce	oe	we	me									cef	oef	wef
Program	Predicted	correct	overlapping	wrong	missing	Sensitivity	Specificity	Ratio	Split	Fused	Sensitivity	Specificity	Ratio	correct	overlapping	
	exons	exons	exons	exons	exons	Sne	Spe	WE	exons	exons	Snef	Spef	Wef	exons	exons	
GENSCAN	938	652	204	82	175	0.63	0.70	0.09	10	16	0.63	0.69	0.12	649	182	110
GeneMark.hmm	1104	845	172	87	26	0.82	0.77	0.08	10	4	0.82	0.76	0.10	844	144	110
MZEF prior $p = 0.01$	641	401	153	87	480	0.39	0.63	0.14	11	10	0.37	0.60	0.21	382	126	134
MZEF prior $p = 0.04$	846	459	236	151	358	0.45	0.54	0.18	32	14	0.43	0.52	0.27	438	178	231
MZEF prior $p = 0.10$	998	490	298	210	283	0.48	0.49	0.21	50	16	0.45	0.47	0.32	467	210	322
FGENE	1061	569	300	192	213	0.55	0.54	0.18	56	6	0.55	0.53	0.28	562	197	299
GRAIL	1184	449	506	229	80	0.44	0.38	0.19	12	16	0.43	0.38	0.25	444	440	293
FEX	1745	562	484	699	155	0.55	0.32	0.40	180	23	0.53	0.31	0.57	547	208	993
FGENESP	737	433	195	109	403	0.42	0.59	0.15	7	8	0.41	0.57	0.21	423	156	156

Softberry

- Many software tools
- Some free trial versions on-line
- Some – pay for license
- http://www.softberry.com/berry.phtml



Address http://www.softberry.com/berry.phtml

TEST ON LINE

- SEQMAN
- GENE FINDING in Eukaryota
- GENE FINDING WITH SIMILARITY
- OPERON AND GENE FINDING IN BACTERIA
- GENE FINDING IN VIRUSES
- ALIGNMENT /Sequences&genomes
- GENOME EXPLORER /Infogene
- HUMAN-MOUSE-RAT SYNTENY
- SEARCH FOR MOTIFS /promoters&functional
- PROTEIN STRUCTURE
- PROTEIN LOCATION

Welcome to Softberry. Our scientific team is dedicated to developing and improving bioinformatics software to help identify genes and functional signals, determine gene function, decipher gene expression data and select disease-specific genes and drug target candidates. We are providing customized solutions to analyze and compare genomes, predict and annotate their genes based on sequence and structure comparison, recognition of conserved regulatory elements and defining cell location of predicted proteins.

View and Run all Software

FAQ

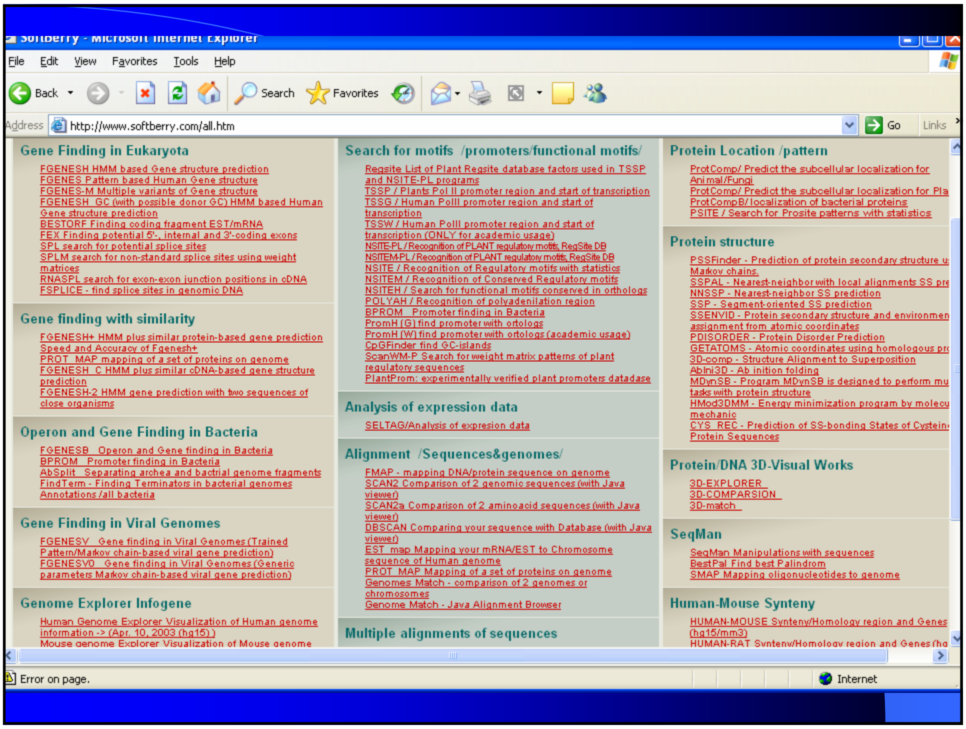
Automatic genome annotation
Eukaryotic: animal, plant, fungi
Bacterial and bacterial community DNA
Visualization of annotations
Genome Explorer
Visualization of Bacterial genome comparison and annotation

Sequence comparison
Alignment of genomic sequences
Multiple alignment and tree construction
Fast search in genomes
ESTs clustering and visualization

- Publications by Topics
- Publications by progra

Recent News

October 4, 2004. ***So releases ProtComp ver.*** new version of popular p for prediction of subcellular local ProtComp, has overall pre accuracy of >90% (see h more details). Pre accuracy of prokaryotic v ***ProtCompB ver. 2***, is 95



Address http://www.softberry.com/all.htm

Gene Finding in Eukaryota

FGENESH HMM based Gene structure prediction
FGENES Pattern based Human Gene structure
FGENES-M Multiple variants of Gene structure
FGENESH_GC (with possible donor GC) HMM based Human Gene structure prediction
BESTORF Finding coding fragment EST/mRNA
FEX Finding potential 5'-, internal and 3'-coding exons
SPL search for potential splice sites
SPLM search for non-standard splice sites using weight matrices
RNASPL search for exon-exon junction positions in cDNA
FSPLICE - find splice sites in genomic DNA

Gene finding with similarity

FGENESH+ HMM plus similar protein-based gene prediction
Speed and Accuracy of Fgenesh+
PROT_MAP mapping of a set of proteins on genome
FGENESH_C HMM plus similar cDNA-based gene structure prediction
FGENESH-2 HMM gene prediction with two sequences of close organisms

Operon and Gene Finding in Bacteria

FGENESB Operon and Gene finding in Bacteria
BPROM Promoter finding in Bacteria
AbSplit Separating archea and bactrial genome fragments
FindTerm - Finding Terminators in bacterial genomes
Annotations /all bacteria

Gene Finding in Viral Genomes

FGENESV Gene finding in Viral Genomes (Trained Pattern/Markov chain-based viral gene prediction)
FGENESV0 Gene finding in Viral Genomes (Generic parameters Markov chain-based viral gene prediction)

Genome Explorer Infogene

Human Genome Explorer Visualization of Human genome information -> (Apr. 10, 2003 (hg15))
Mouse genome Explorer Visualization of Mouse genome

Search for motifs /promoters/functional motifs/

Regsite List of Plant Regsite database factors used in TSSP and NSITE-PL programs
TSSP / Plants Pol II promoter region and start of transcription
TSSG / Human PolII promoter region and start of transcription
TSSW / Human PolII promoter region and start of transcription (ONLY for academic usage)
NSITE-PL / Recognition of PLANT regulatory motifs, RegSite DB
NSITEM-PL / Recognition of PLANT regulatory motifs, RegSite DB
NSITE / Recognition of Regulatory motifs with statistics
NSITEM / Recognition of Conserved Regulatory motifs
NSITEH / Search for functional motifs conserved in orthologs
POLYAH / Recognition of polyadenilation region
BPROM Promoter finding in Bacteria
PromH (G) find promoter with ortologs
PromH (W) find promoter with ortologs (academic usage)
CpGFinder find GC-islands
ScanWM-P Search for weight matrix patterns of plant regulatory sequences
PlantProm: experimentally verified plant promoters datadase

Analysis of expression data

SELTAG/Analysis of expresion data

Alignment /Sequences&genomes/

FMAP - mapping DNA/protein sequence on genome
SCAN2 Comparison of 2 genomic sequences (with Java viewer)
SCAN2a Comparison of 2 aminoacid sequences (with Java viewer)
DBSCAN Comparing your sequence with Database (with Java viewer)
EST_map Mapping your mRNA/EST to Chromosome sequence of Human genome
PROT_MAP Mapping of a set of proteins on genome
Genomes Match - comparison of 2 genomes or chromosomes
Genome Match - Java Alignment Browser

Multiple alignments of sequences

Protein Location /pattern

ProtComp/ Predict the subcellular localization for Animal/Fungi
ProtComp/ Predict the subcellular localization for Pla
ProtCompB/ localization of bacterial proteins
PSITE / Search for Prosite patterns with statistics

Protein structure

PSSFinder - Prediction of protein secondary structure u Markov chains.
SSPAL - Nearest-neighbor with local alignments SS pre
NNSSP - Nearest-neighbor SS prediction
SSP - Segment-oriented SS prediction
SSENVID - Protein secondary structure and environmen assignment from atomic coordinates
PDISORDER - Protein Disorder Prediction
GETATOMS - Atomic coordinates using homologous pr
3D-comp - Structure Alignment to Superposition
AbIni3D - Ab inition folding
MDynSB - Program MDynSB is designed to perform mu tasks with protein structure
HMod3DMM - Energy minimization program by molecu mechanic
CYS_REC - Prediction of SS-bonding States of Cystein Protein Sequences

Protein/DNA 3D-Visual Works

3D-EXPLORER
3D-COMPARSION
3D-match

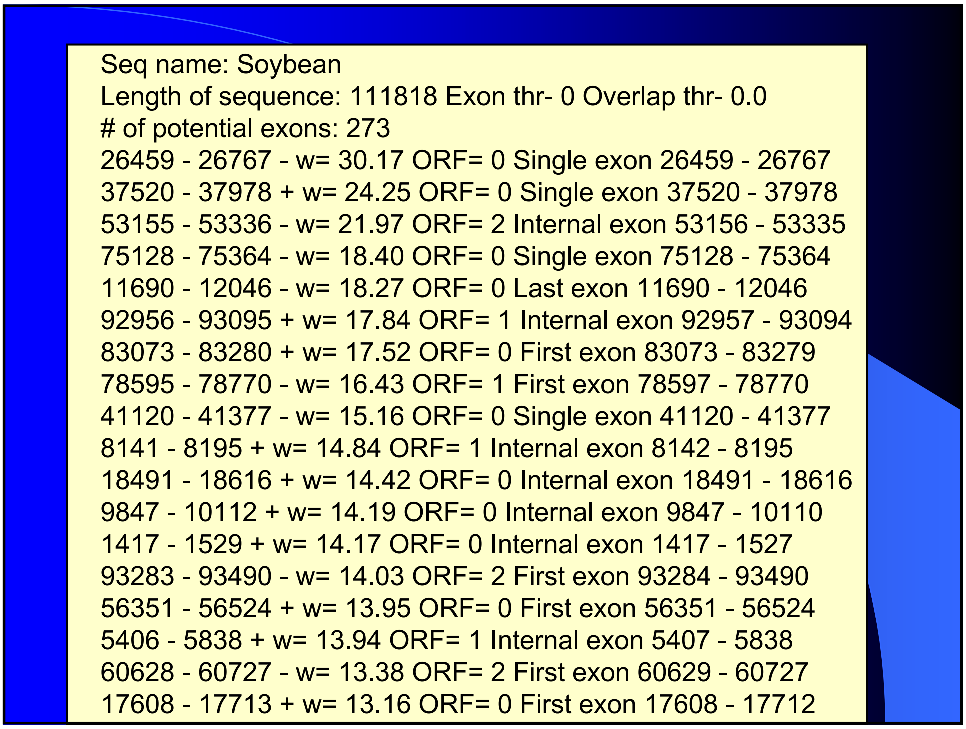
SeqMan

SeqMan Manipulations with sequences
BestPal Find best Palindrom
SMAP Mapping oligonucleotides to genome

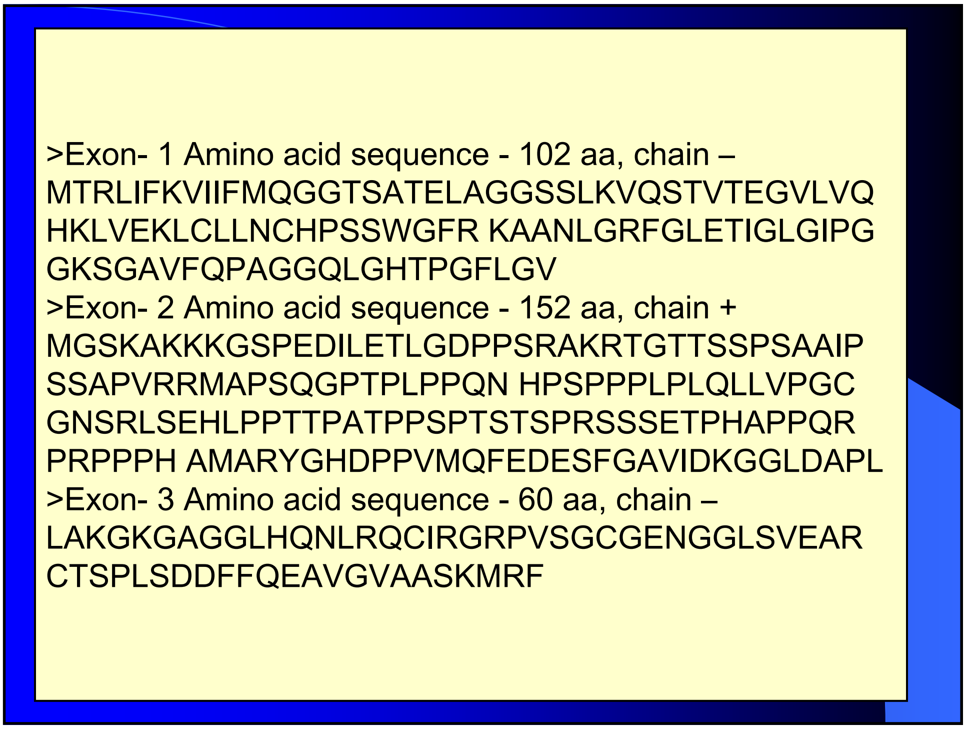
Human-Mouse Synteny

HUMAN-MOUSE Synteny/Homology region and Genes (hg15/mm3)
HUMAN-RAT Synteny/Homology region and Genes (hg

Error on page.



```
Seq name: Soybean
Length of sequence: 111818 Exon thr- 0 Overlap thr- 0.0
# of potential exons: 273
26459 - 26767 - w= 30.17 ORF= 0 Single exon 26459 - 26767
37520 - 37978 + w= 24.25 ORF= 0 Single exon 37520 - 37978
53155 - 53336 - w= 21.97 ORF= 2 Internal exon 53156 - 53335
75128 - 75364 - w= 18.40 ORF= 0 Single exon 75128 - 75364
11690 - 12046 - w= 18.27 ORF= 0 Last exon 11690 - 12046
92956 - 93095 + w= 17.84 ORF= 1 Internal exon 92957 - 93094
83073 - 83280 + w= 17.52 ORF= 0 First exon 83073 - 83279
78595 - 78770 - w= 16.43 ORF= 1 First exon 78597 - 78770
41120 - 41377 - w= 15.16 ORF= 0 Single exon 41120 - 41377
8141 - 8195 + w= 14.84 ORF= 1 Internal exon 8142 - 8195
18491 - 18616 + w= 14.42 ORF= 0 Internal exon 18491 - 18616
9847 - 10112 + w= 14.19 ORF= 0 Internal exon 9847 - 10110
1417 - 1529 + w= 14.17 ORF= 0 Internal exon 1417 - 1527
93283 - 93490 - w= 14.03 ORF= 2 First exon 93284 - 93490
56351 - 56524 + w= 13.95 ORF= 0 First exon 56351 - 56524
5406 - 5838 + w= 13.94 ORF= 1 Internal exon 5407 - 5838
60628 - 60727 - w= 13.38 ORF= 2 First exon 60629 - 60727
17608 - 17713 + w= 13.16 ORF= 0 First exon 17608 - 17712
```



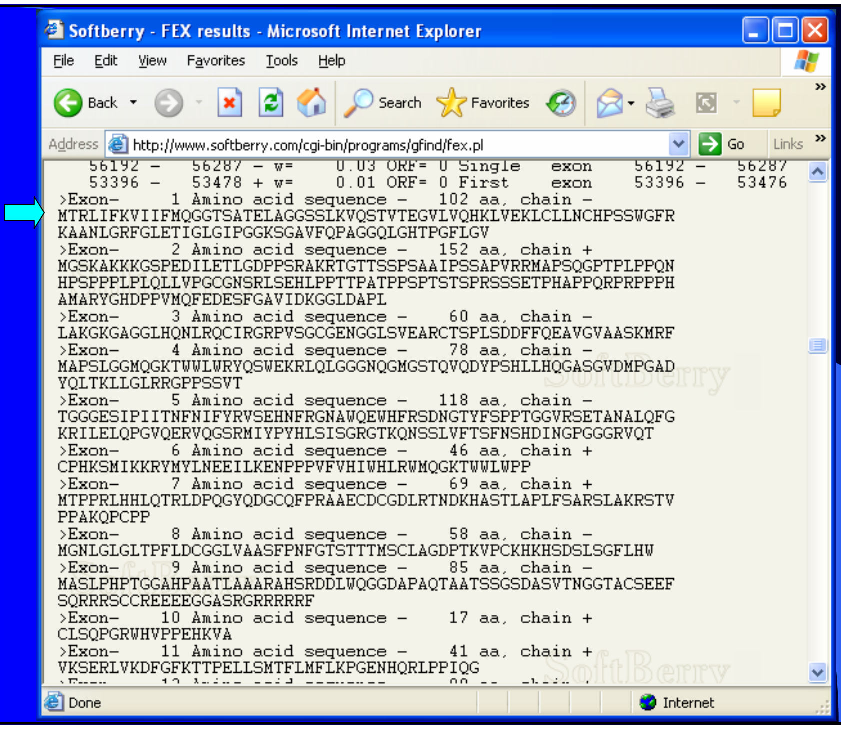
```
>Exon- 1 Amino acid sequence - 102 aa, chain –
MTRLIFKVIIFMQGGTSATELAGGSSLKVQSTVTEGVLVQ
HKLVEKLCLLNCHPSSWGFR KAANLGRFGLETIGLGIPG
GKSGAVFQPAGGQLGHTPGFLGV
>Exon- 2 Amino acid sequence - 152 aa, chain +
MGSKAKKKGSPEDILETLGDPPSRAKRTGTTSSPSAAIP
SSAPVRRMAPSQGPTPLPPQN HPSPPPLPLQLLVPGC
GNSRLSEHLPPTTPATPPSPTSTSPRSSSETPHAPPQR
PRPPPH AMARYGHDPPVMQFEDESFGAVIDKGGLDAPL
>Exon- 3 Amino acid sequence - 60 aa, chain –
LAKGKGAGGLHQNLRQCIRGRPVSGCGENGGLSVEAR
CTSPLSDDFFQEAVGVAASKMRF
```

Exon 1

BLASTP

Protein-Protein

No match

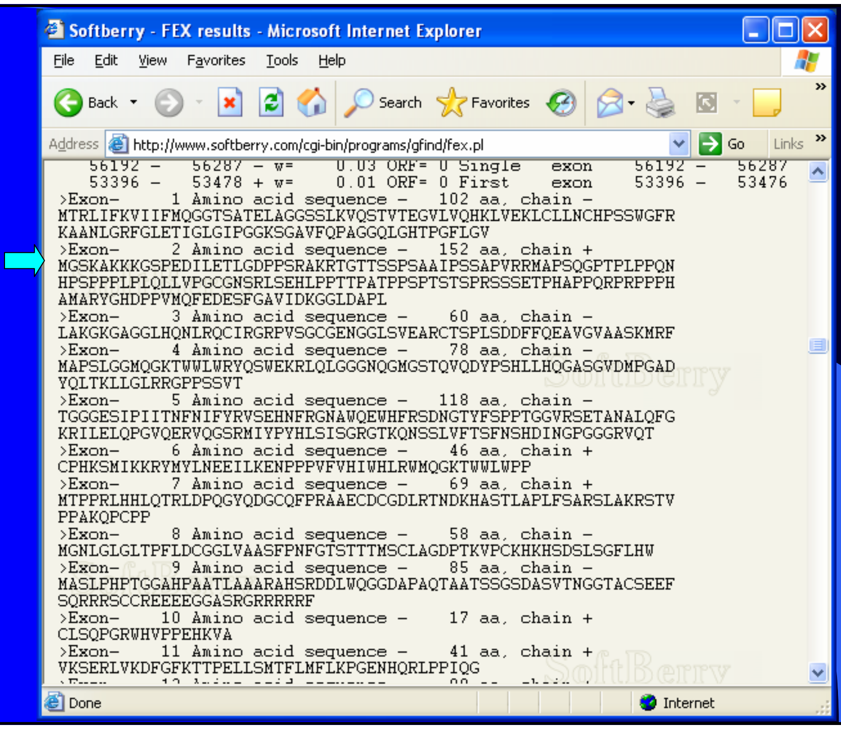

Softberry - FEX results - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://www.softberry.com/cgi-bin/programs/gfind/fex.pl
Go
Links
Done
Internet

```
    56192 -    56287 - w=    0.03 ORF= 0 Single   exon    56192 -    56287
    53396 -    53478 + w=    0.01 ORF= 0 First    exon    53396 -    53476
 >Exon-      1 Amino acid sequence -   102 aa, chain -
 MTRLIFKVIIFMQGGTSATELAGGSSLKVQSTVTEGVLVQHKLVEKLCLLNCHPSSWGFR
 KAANLGRFGLETIGLGIPGGKSGAVFQPAGGQLGHTPGFLGV
 >Exon-      2 Amino acid sequence -   152 aa, chain +
 MGSKAKKKGSPEDILETLGDPPSRAKRTGTTSSPSAAIPSSAPVRRMAPSQGPTPLPPQN
 HPSPPPLPLQLLVPGCGNSRLSEHLPPTTPATPPSPTSTSPRSSSETPHAPPQRPRPPPH
 AMARYGHDPPVMQFEDESFGAVIDKGGLDAPL
 >Exon-      3 Amino acid sequence -    60 aa, chain -
 LAKGKGAGGLHQNLRQCIRGRPVSGCGENGGLSVEARCTSPLSDDFFQEAVGVAASKMRF
 >Exon-      4 Amino acid sequence -    78 aa, chain -
 MAPSLGGMQGKTWWLWRYQSWEKRLQLGGGNQGMGSTQVQDYPSHLLHQGASGVDMPGAD
 YQLTKLLGLRRGPPSSVT
 >Exon-      5 Amino acid sequence -   118 aa, chain -
 TGGGESIPIITNFNIFYRVSEHNFRGNAWQEWHFRSDNGTYFSPPTGGVRSETANALQFG
 KRILELQPGVQERVQGSRMIYPYHLSISGRGTKQNSSLVFTSFNSHDINGPGGGRVQT
 >Exon-      6 Amino acid sequence -    46 aa, chain +
 CPHKSMIKKRYMYLNEEILKENPPPVFVHIWHLRWMQGKTWWLWPP
 >Exon-      7 Amino acid sequence -    69 aa, chain +
 MTPPRLHHLQTRLDPQGYQDGCQFPRAAECDCGDLRTNDKHASTLAPLFSARSLAKRSTV
 PPAKQPCPP
 >Exon-      8 Amino acid sequence -    58 aa, chain -
 MGNLGLGLTPFLDCGGLVAASFPNFGTSTTTMSCLAGDPTKVPCKHKHSDSLSGFLHW
 >Exon-      9 Amino acid sequence -    85 aa, chain -
 MASLPHPTGGAHPAATLAAARAHSRDDLWQGGDAPAQTAATSSGSDASVTNGGTACSEEF
 SQRRRSCCREEEEGGASRGRRRRRF
 >Exon-     10 Amino acid sequence -    17 aa, chain +
 CLSQPGRWHVPPEHKVA
 >Exon-     11 Amino acid sequence -    41 aa, chain +
 VKSERLVKDFGFKTTPELLSMTFLMFLKPGENHQRLPPIQG
```

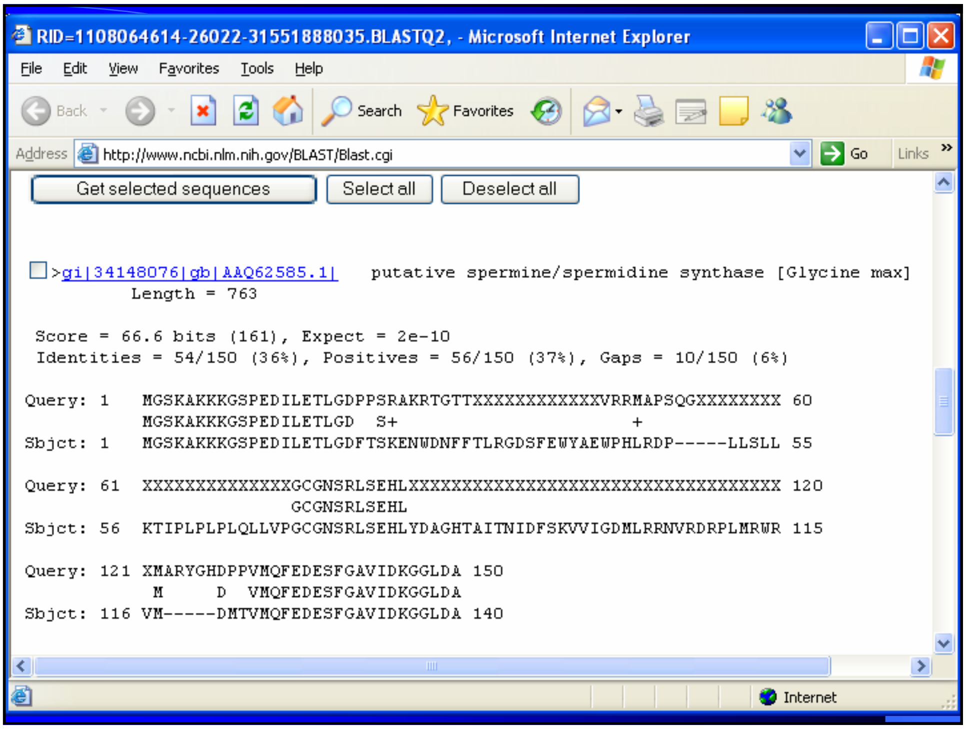
Exon 2

BLASTP

Protein-Protein


Softberry - FEX results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.softberry.com/cgi-bin/programs/gfind/fex.pl
Go
Links
Done
Internet

```
   56192 -    56287 - w=    0.03 ORF= 0 Single   exon    56192 -    56287
   53396 -    53478 + w=    0.01 ORF= 0 First    exon    53396 -    53476
>Exon-      1 Amino acid sequence -   102 aa, chain -
MTRLIFKVIIFMQGGTSATELAGGSSLKVQSTVTEGVLVQHKLVEKLCLLNCHPSSWGFR
KAANLGRFGLETIGLGIPGGKSGAVFQPAGGQLGHTPGFLGV
>Exon-      2 Amino acid sequence -   152 aa, chain +
MGSKAKKKGSPEDILETLGDPPSRAKRTGTTSSPSAAIPSSAPVRRMAPSQGPTPLPPQN
HPSPPPLPLQLLVPGCGNSRLSEHLPPTTPATPPSPTSTSPRSSSETPHAPPQRPRPPPH
AMARYGHDPPVMQFEDESFGAVIDKGGLDAPL
>Exon-      3 Amino acid sequence -    60 aa, chain -
LAKGKGAGGLHQNLRQCIRGRPVSGCGENGGLSVEARCTSPLSDDFFQEAVGVAASKMRF
>Exon-      4 Amino acid sequence -    78 aa, chain -
MAPSLGGMQGKTWWLWRYQSWEKRLQLGGGNQGMGSTQVQDYPSHLLHQGASGVDMPGAD
YQLTKLLGLRRGPPSSVT
>Exon-      5 Amino acid sequence -   118 aa, chain -
TGGGESIPIITNFNIFYRVSEHNFRGNAWQEWHFRSDNGTYFSPPTGGVRSETANALQFG
KRILELQPGVQERVQGSRMIYPYHLSISGRGTKQNSSLVFTSFNSHDINGPGGGRVQT
>Exon-      6 Amino acid sequence -    46 aa, chain +
CPHKSMIKKRYMYLNEEILKENPPPVFVHIWHLRWMQGKTWWLWPP
>Exon-      7 Amino acid sequence -    69 aa, chain +
MTPPRLHHLQTRLDPQGYQDGCQFPRAAECDCGDLRTNDKHASTLAPLFSARSLAKRSTV
PPAKQPCPP
>Exon-      8 Amino acid sequence -    58 aa, chain -
MGNLGLGLTPFLDCGGLVAASFPNFGTSTTTMSCLAGDPTKVPCKHKHSDSLSGFLHW
>Exon-      9 Amino acid sequence -    85 aa, chain -
MASLPHPTGGAHPAATLAAARAHSRDDLWQGGDAPAQTAATSSGSDASVTNGGTACSEEF
SQRRRSCCREEEEGGASRGRRRRRF
>Exon-     10 Amino acid sequence -    17 aa, chain +
CLSQPGRWHVPPEHKVA
>Exon-     11 Amino acid sequence -    41 aa, chain +
VKSERLVKDFGFKTTPELLSMTFLMFLKPGENHQRLPPIQG
```



```
□ >gi|34148076|gb|AAQ62585.1|   putative spermine/spermidine synthase [Glycine max]
          Length = 763

 Score = 66.6 bits (161), Expect = 2e-10
 Identities = 54/150 (36%), Positives = 56/150 (37%), Gaps = 10/150 (6%)

Query: 1   MGSKAKKKGSPEDILETLGDPPSRAKRTGTTXXXXXXXXXXXXVRRMAPSQGXXXXXXXX 60
            MGSKAKKKGSPEDILETLGD  S+                      +
Sbjct: 1   MGSKAKKKGSPEDILETLGDFTSKENWDNFFTLRGDSFEWYAEWPHLRDP-----LLSLL 55

Query: 61  XXXXXXXXXXXXXXGCGNSRLSEHLXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 120
                          GCGNSRLSEHL
Sbjct: 56  KTIPLPLPLQLLVPGCGNSRLSEHLYDAGHTAITNIDFSKVVIGDMLRRNVRDRPLMRWR 115

Query: 121 XMARYGHDPPVMQFEDESFGAVIDKGGLDA 150
             M     D  VMQFEDESFGAVIDKGGLDA
Sbjct: 116 VM-----DMTVMQFEDESFGAVIDKGGLDA 140
```

Gene Annotation Tips

- Use several different prediction software
  - Find Open Reading Frame (ORF)
  - Find Promoter
- Use software best suited for your organism
- Use BLAST and GenBank
- Use protein sequence and DNA coding sequence
- 5’ and 3’ ends are particularly difficult

What we covered today

- NCBI
- Genomic Databases
- UCSC
- Genomic DNA annotation