Introduction to Bioinformatics
2. **DNA Sequence Retrieval and comparison**

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

- Retrieving a known DNA sequence
- Similarity searching with a DNA sequence
- BLAST

Retrieving a DNA sequence

You read a paper and..

- Is full-length clone of gene available?
- Is at least some of the DNA sequence available? (EST sequence)?

Finding Sequences in Databases

- The public DNA and protein sequence databases are huge.
- In order for these databases to be useful, the data must be readily accessible to researchers.

What Are You Looking For?

- **A gene?**
  - DNA or protein sequence?
- DNA sequences are essentially all in **GenBank**
  - Genomic, mRNA, cDNA, EST?
- Proteins are harder to pin down
  - **GenPept** (GenBank Peptides) is huge and poorly annotated - lots of junk
  - **SwissProt** is carefully annotated, but not fully comprehensive
  - **PIR** is somewhere in between

Large Databases

- Once upon a time, **GenBank** sent out sequence updates on CD-ROM disks a few times per year.
- Now **GenBank** is over 95 Gigabytes (28 **billion** bases)
- Most biocomputing sites update their copy of **GenBank** every day over the internet.
- Scientists access **GenBank** directly over the Web

You can search DNA sequence database

- Retrieve known sequences by
  - Keyword search
  - Accession numbers
- If you know some DNA sequence
  - Compare your DNA sequence with those in database
  - Basic Local Alignment Search Tool (BLAST) searches

Retrieve a DNA sequence

- ENTREZ
  - http://www.ncbi.nlm.nih.gov/Entrez/
- Click – Nucleotide
  - GenBank
  - OR

  http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
- Accession number
- Keyword search

Entrez is a Tool for Finding Sequences

- **GenBank** is managed by the **NCBI** (National Center for Biotechnology Information) which is a part of the US National Library of Medicine.
- NCBI has created a Web-based tool called **Entrez** for finding sequences in **GenBank**.

  http://www.ncbi.nlm.nih.gov
- Each sequence in **GenBank** has a unique "**accession number**".
- **Entrez** can also search for keywords such as gene names, protein names, and the names of orgainisms or biological functions

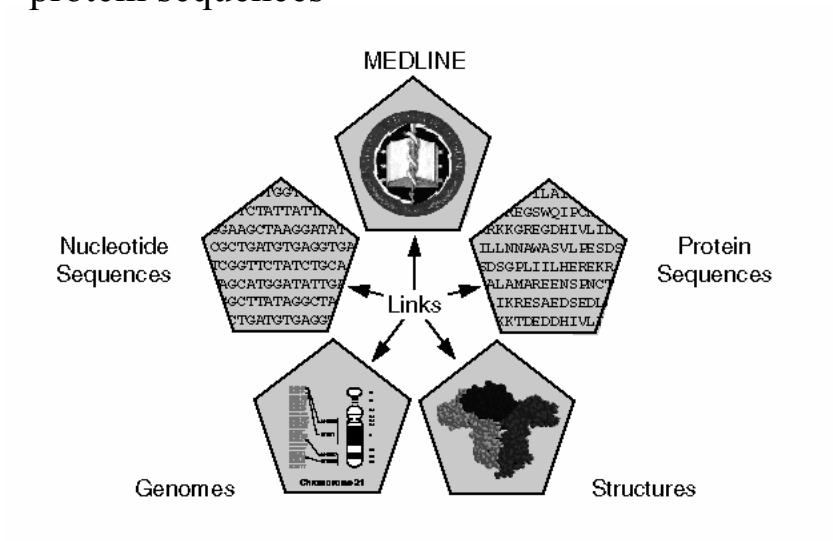
Entrez is a Database

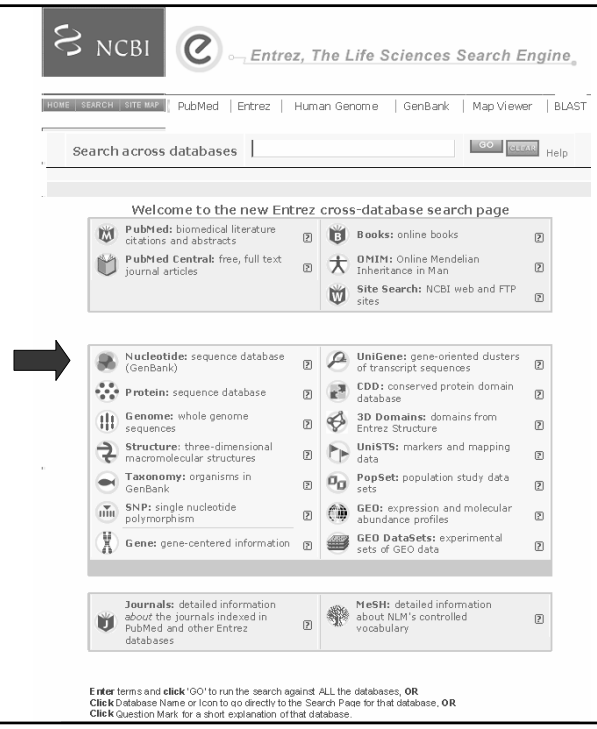
- The **Entrez** database contains all of the nucleotide and protein sequences in **GenBank** (updated daily) along with all of the literature in **MEDLINE** and the 3-D protein structures in **PDB** (**Protein Data Base**).
- **Entrez** is much more than a database, it is a both a powerful search engine and a pre-computed list of relationships amnog all of its data elements

Entrez is Internally Cross-linked

- DNA and protein sequences are linked to other similar sequences
- **Medline** citations are linked to other citations that contain similar keywords
- 3-D structures are linked to similar structures

Databases contain more than just DNA & protein sequences

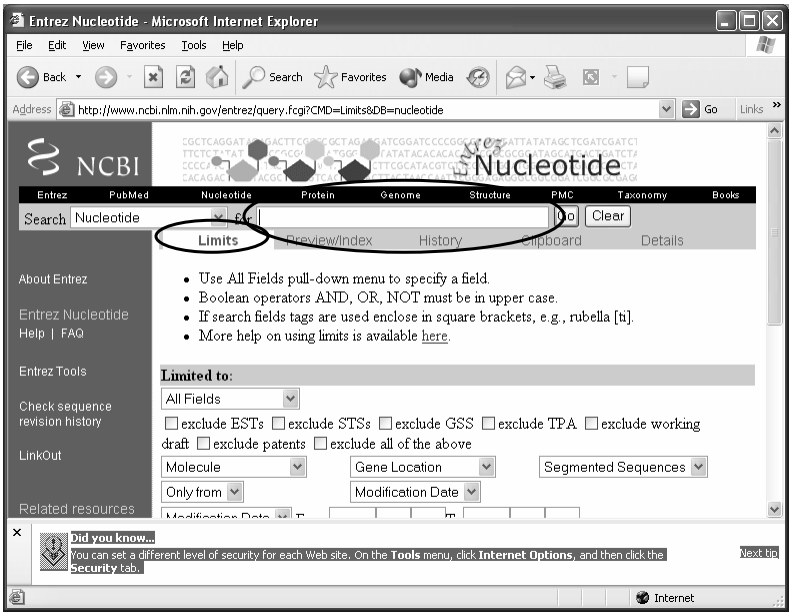




GenBank

- National Institute of Health, National Library of Medicine, National Center for Biotechnology Information
- http://www.ncbi.nlm.nih.gov/
- http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html

- Retrieve a sequence from GenBank
- Analyze raw sequence data
  - Base calling
  - Editing
  - Obtaining a consensus sequence
  - Translating
  - Restriction mapping
  - Similarity comparisons
  - Motif searches

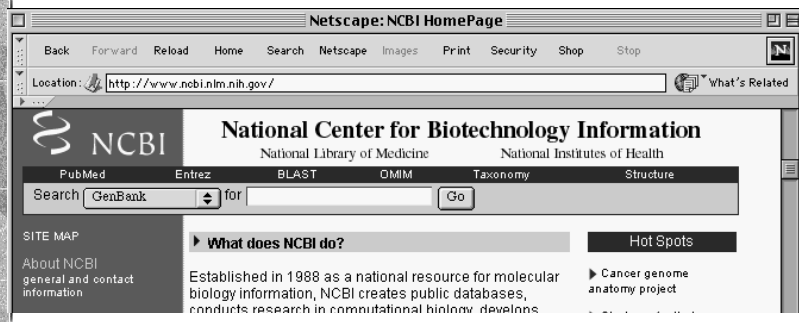


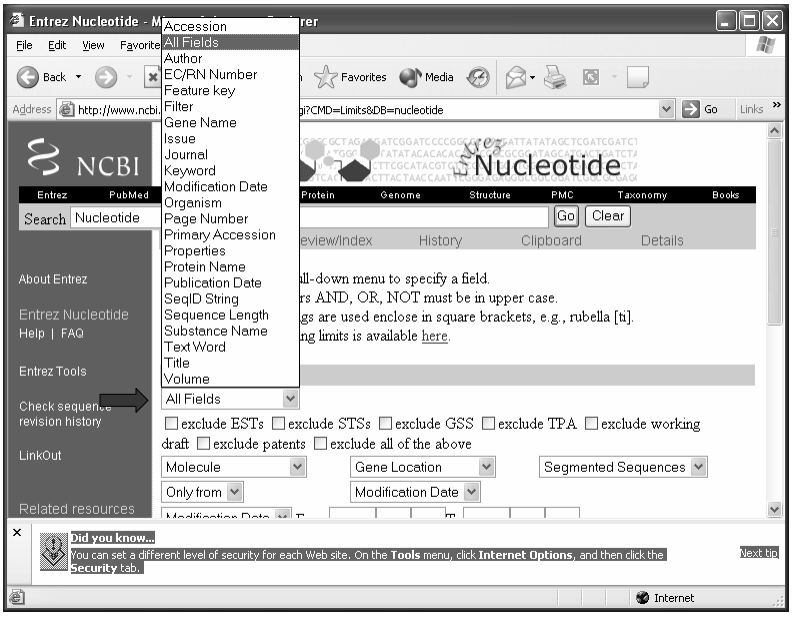
Accession Numbers!!

- Databases are designed to be searched by accession numbers (and locus IDs)
- These are guaranteed to be non-redundant, accurate, and not to change.
- Searching by gene names and keywords is inexact and retrieves more than one record usually

Type in a Query term

- Enter your search words in the query box and hit the "**Go**" button





Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for soybean AND chalcone AND synthase Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

All Fields

☐ exclude ESTs ☐ exclude STSs ☐ exclude GSS ☐ exclude TPA ☐ exclude working draft ☐ exclude patents ☐ exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Home Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for soybean AND chalcone AND synthase Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

All Fields

☑ exclude ESTs ☐ exclude STSs ☐ exclude GSS ☐ exclude TPA ☐ exclude working draft ☐ exclude patents ☐ exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide soybean AND chalcone synthase 6 Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

All Fields

☐ exclude ESTs ☐ exclude STSs ☐ exclude GSS ☐ exclude TPA ☐ exclude working draft ☐ exclude patents ☐ exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for soybean AND chalcone synthase 6 Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

Display Summary Show: 20 Send to Text

Items 1-9 of 9 One page.

☐ 1: BG725328 Links
sae35d12.y1 Gm-c1051 Glycine max cDNA clone GENOME SYSTEMS CLONE ID: Gm-c1051-7103 5' similar to SW:CHS6_SOYBN P30080 CHALCONE SYNTHASE 6 ;, mRNA sequence
gi|14008724|gb|BG725328.1|[14008724]

☐ 2: BE346944 Links
sp33b03.y1 Gm-c1043 Glycine max cDNA clone GENOME SYSTEMS CLONE ID: Gm-c1043-6 5' similar to SW:CHSA_IPOCO P48393 CHALCONE SYNTHASE A ;, mRNA sequence

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724 Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File

1: BG725328. sae35d12.y1 Gm-c1...[gi:14008724] Links

IDENTIFIERS

dbEST Id: **8490125**
EST name: sae35d12.y1
GenBank Acc: BG725328
GenBank gi: 14008724

CLONE INFO
Clone Id: GENOME SYSTEMS CLONE ID: Gm-c1051-7103 (5')
DNA type: cDNA

PRIMERS
PolyA Tail: Unknown

SEQUENCE

```
GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTGAGTCC
```

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.
Next tip

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724 Go Links

SEQUENCE

```
GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTGAGTCC
TACTTGGCGAAGGTGGCCATCAATAGCGCCTTCGCTGTCTGGAAGGATTGTTTGTGCAGT
CCACACAAGCTCAAACAAAGGCTTTTCAGCTGGCAGAGGATCTGATCCAACAATGACAGC
AGCTGCACCATCTCCAAACAAGGCTTGCCCCACAAGGCTGTCAAGATGTGTGTCACTCGG
GCCACGAAATGTGACTGCTGTGATCTCCGA
```

Quality: High quality sequence stops at base: 400

Entry Created: May 8 2001
Last Updated: Jul 22 2004

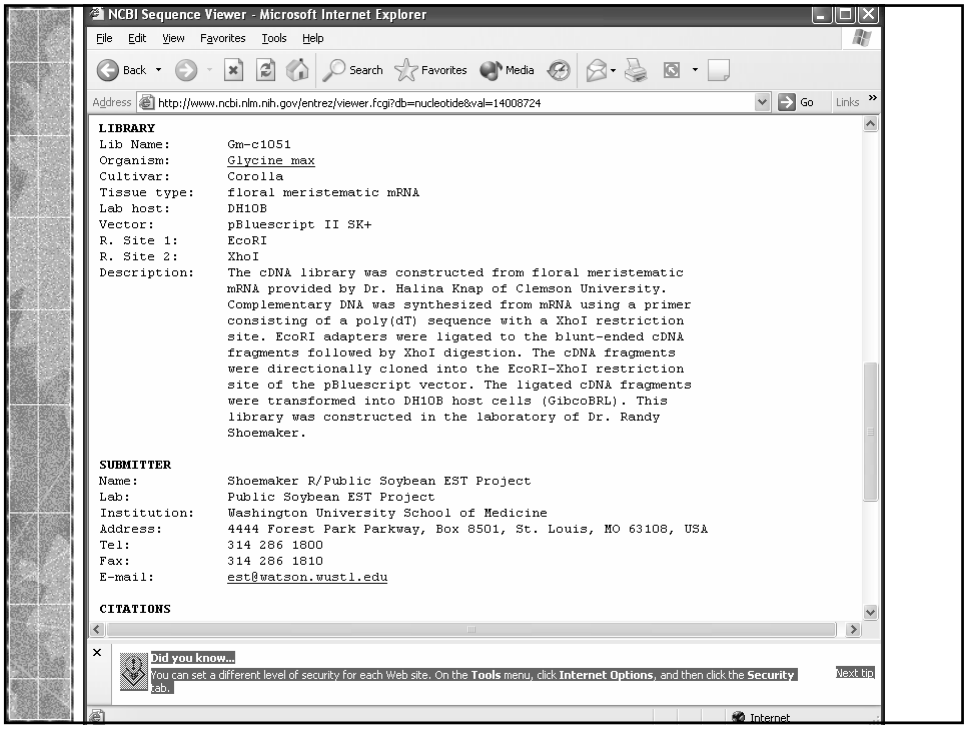
COMMENTS

When it has been determined, an EST from the other end of this clone is listed in the 'Other ESTs on clone' field. Possible reversed clone: similarity on wrong strand This clone is available through: Biogenetic Services, 801 32nd Ave. Brookings, SD 57006 USA (phone: 800 423 4163; email: info@biogeneticservices.com)

PUTATIVE ID Assigned by submitter
SW:CHS6_SOYBN P30080 CHALCONE SYNTHASE 6 ;

LIBRARY
Lib Name: Gm-c1051
Organism: Glycine max
Cultivar: Corolla
Tissue type: floral meristematic mRNA
Lib host: DH10B

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.
Next tip

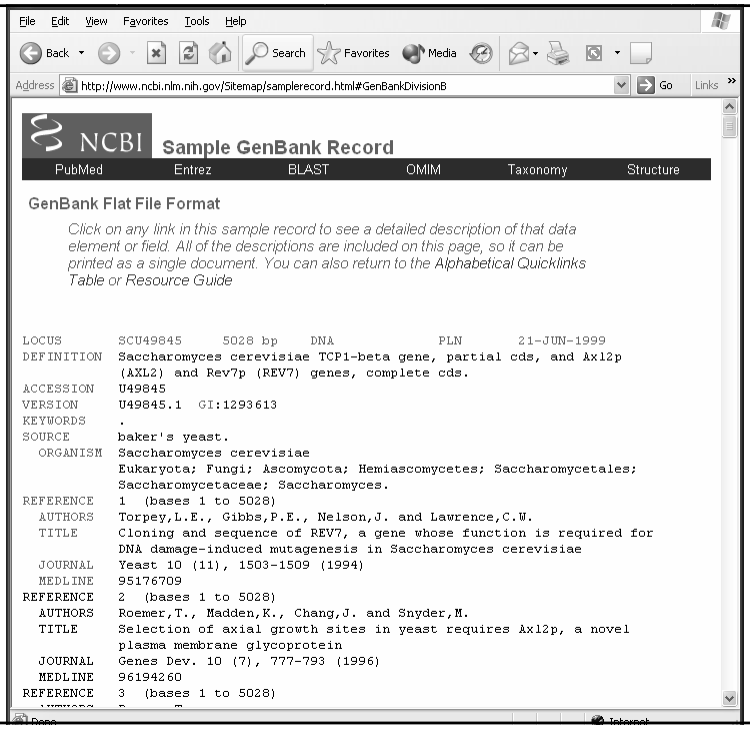


GenBank Records

- Databases are composed of records
- Flat File Format
- Provides information
- Standard, consistent organization of data

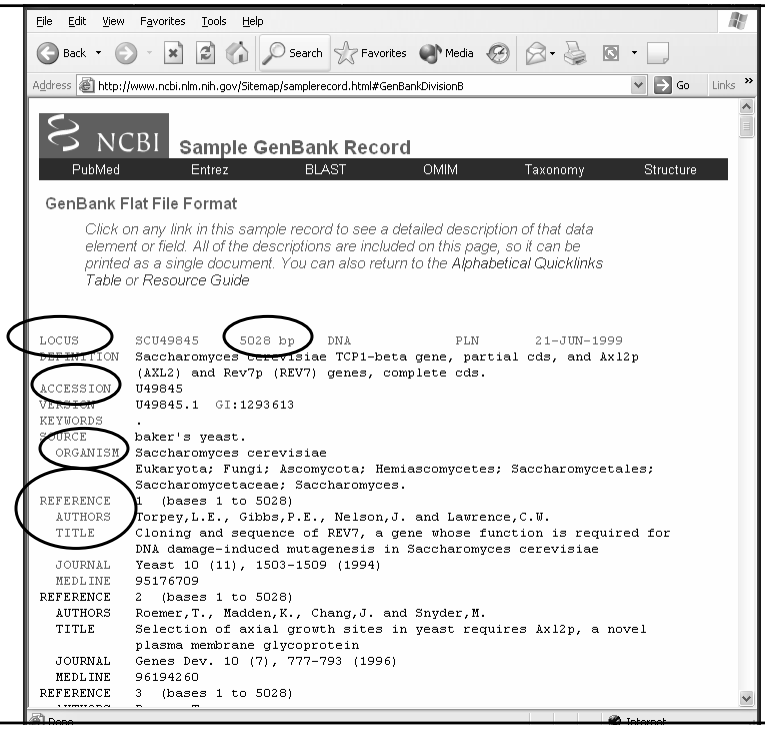
Flat file format

- Organized in a structured manner
- One big file
- Large body of information assembled and distributed in consistent format
- Lack support for procession transactions (inserts and updates)



Some Fields of GenBank Record

- Locus Name
- Sequence length
- Molecule type
- Definition
- GenBank accession number
- Version
- Keywords
- Source
- Organism
- Reference
- Reference
- Authors
- Title
- Journal
- Medline
- Other references
- features
- Amino acid translation
- Nucleotide sequence

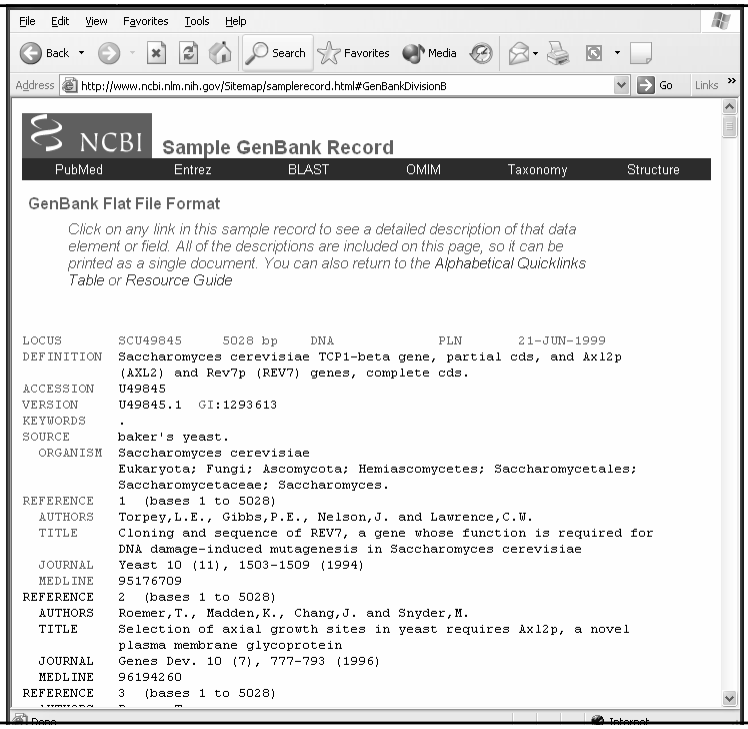


Locus Name

Unique

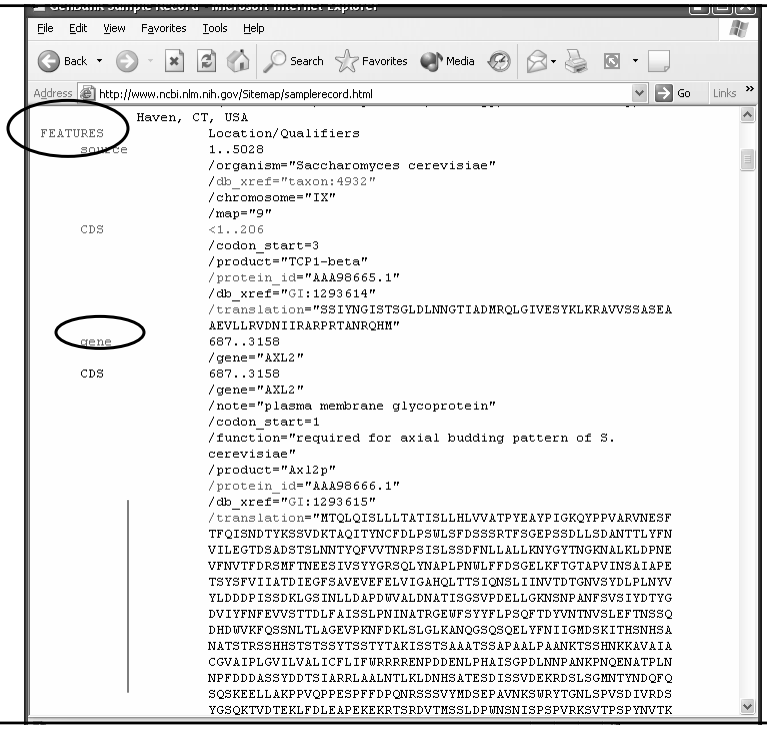
- Up to 10 characters
- 6 character
  - Genus species
- 8 characters
  - Just accession number

- Better to search for accession number than Locus Name



GenBank Accession Number

- Unique identifier for sequence record
- Usually a combination of letter(s) and numbers
- Do not change even if information changes
- Newer accession numbers to new submission using some of this data



Address http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html Go Links

```
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
gene                 complement(3300..4037)
                     /gene="REV7"
CDS                  complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
BASE COUNT     1510 a   1074 c    835 g   1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
```

Internet

Refine the Query

- Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query
- The “History” feature allows you to combine any of your past queries.
- The “Limits” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- [Many other features are designed to search for literature in MEDLINE]

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

All Fields

☐ exclude ESTs ☐ exclude STSs ☐ exclude GSS ☐ exclude TPA ☐ exclude working draft ☐ exclude patents ☐ exclude all of the above

Molecule | Gene Location | Segmented Sequences

Only from | Modification Date

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Refresh Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Index&DB=nucleotide Go Links

NCBI Entrez Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Preview Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

- Enter terms and click Preview to see only the number of search results.
- To combine searches use # before search number, e.g., (#2 OR #3) AND asthma. No history available

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

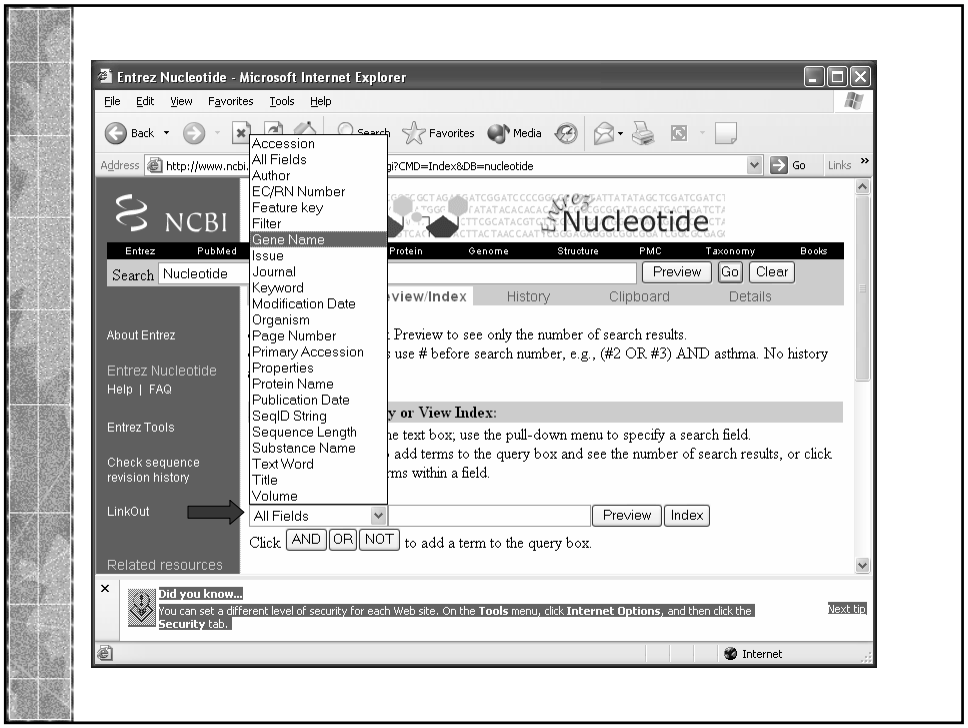
All Fields Preview Index

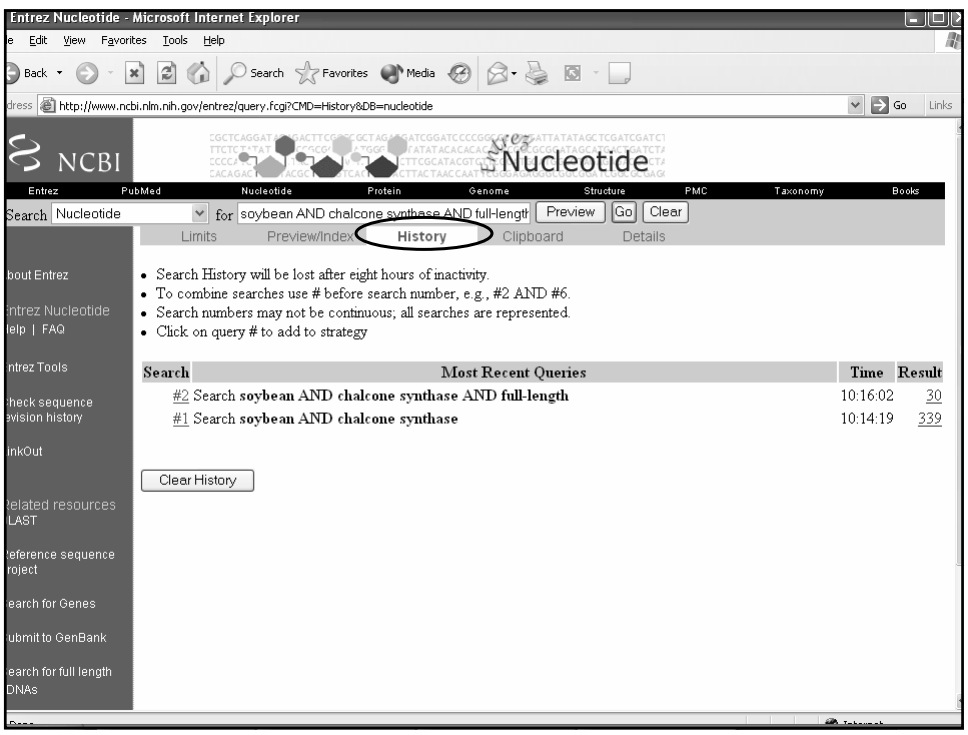
Click AND OR NOT to add a term to the query box.

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet


Entrez Nucleotide - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.
gi?CMD=Index&DB=nucleotide
Go Links
NCBI
Nucleotide
Entrez PubMed Protein Genome Structure PMC Taxonomy Books
Search Nucleotide
Preview Go Clear
view/Index History Clipboard Details
Accession
All Fields
Author
EC/RN Number
Feature key
Filter
Gene Name
Issue
Journal
Keyword
Modification Date
Organism
Page Number
Primary Accession
Properties
Protein Name
Publication Date
SeqID String
Sequence Length
Substance Name
Text Word
Title
Volume
Preview to see only the number of search results.
use # before search number, e.g., (#2 OR #3) AND asthma. No history
y or View Index:
he text box; use the pull-down menu to specify a search field.
add terms to the query box and see the number of search results, or click
rms within a field.
About Entrez
Entrez Nucleotide
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
Related resources
All Fields
Preview Index
Click AND OR NOT to add a term to the query box.
Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.
Next tip
Internet


Entrez Nucleotide - Microsoft Internet Explorer
Edit View Favorites Tools Help
Back Search Favorites Media
dress http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=History&DB=nucleotide
Go Links
NCBI
Nucleotide
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Nucleotide for soybean AND chalcone synthase AND full-length
Preview Go Clear
Limits Preview/Index History Clipboard Details
bout Entrez
ntrez Nucleotide
elp | FAQ
ntrez Tools
heck sequence
evision history
inkOut
elated resources
LAST
eference sequence
roject
earch for Genes
ubmit to GenBank
earch for full length
DNAs
Search History will be lost after eight hours of inactivity.
To combine searches use # before search number, e.g., #2 AND #6.
Search numbers may not be continuous; all searches are represented.
Click on query # to add to strategy
Search Most Recent Queries Time Result
#2 Search soybean AND chalcone synthase AND full-length 10:16:02 30
#1 Search soybean AND chalcone synthase 10:14:19 339
Clear History

DNA similarity search

Find related seqeunces

- Find ESTs
  - If not full-length, may allow assembly from ESTs
- Find other family members
  - Organization and function
- Find similar genes from other cultivars
  - SNP discovery
- Find similar genes from other organisms
  - Phylogenetic relationships

ESTs (Expressed Sequence Tags)

- partial cDNA sequences
- dbEST at NCBI
  - a comprehensive set of all public EST data
- UniGene at NCBI
  - clusters of ESTs and know genes from key species
    - does NOT have consensus sequences
    - has far too many clusters to be representative of real genes (129 K human clusters)

Find related DNA sequences

- Similarity Search (BLAST)
- NCBI GenBank database

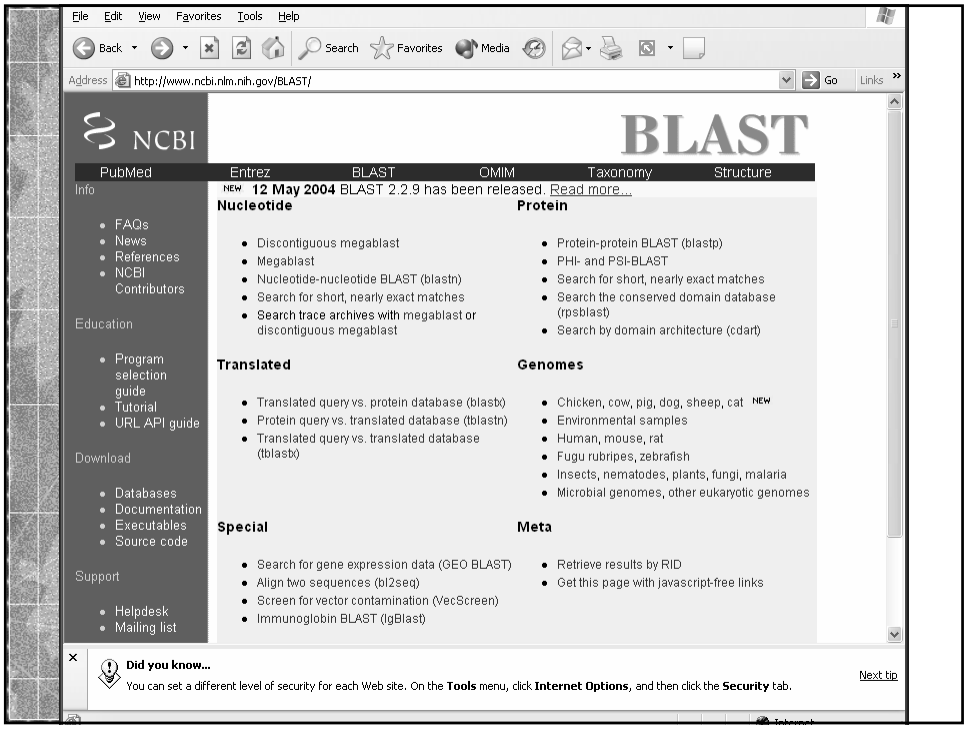
BLAST Searches

- Compare your sequence with database
- http://www.ncbi.nlm.nih.gov/BLAST/
- Nucleotide
- Protein
- Targeted to a genome

- BLAST
  - **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Local alignment
  - Tutorial at: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

BLAST

- Discontiguous Mega BLAST
  - Comparison of diverged sequences especially from different organisms
  - Alignments with low degree of identity
  - Looks for hits in “non-consecutive positions”
- Mega BLAST
  - Slight differences in similarity
  - Not effective at low degree of identity
  - Faster; handles longer sequences
- BLAST
  - Local alignment tool
  - http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html



File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ncbi.nlm.nih.gov/BLAST/ Go Links

- FAQs
- News
- References
- NCBI Contributors

Education

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

NCBI BLAST - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/ Go Links

NCBI

BLAST

PubMed Entrez BLAST OMIM Taxonomy Structure

NEW **12 May 2004** BLAST 2.2.9 has been released. Read more...

Info

- FAQs
- News
- References
- NCBI Contributors

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Databases
- Documentation
- Executables
- Source code

Support

- Helpdesk

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Chicken, cow, pig, dog, sheep, cat NEW
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

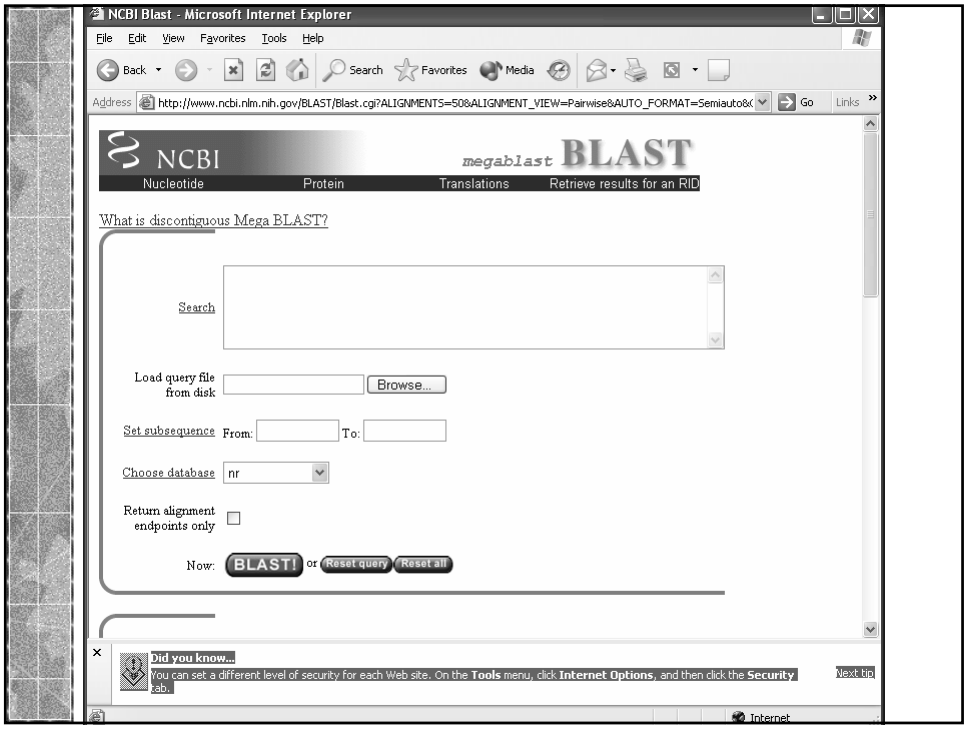
Next tip

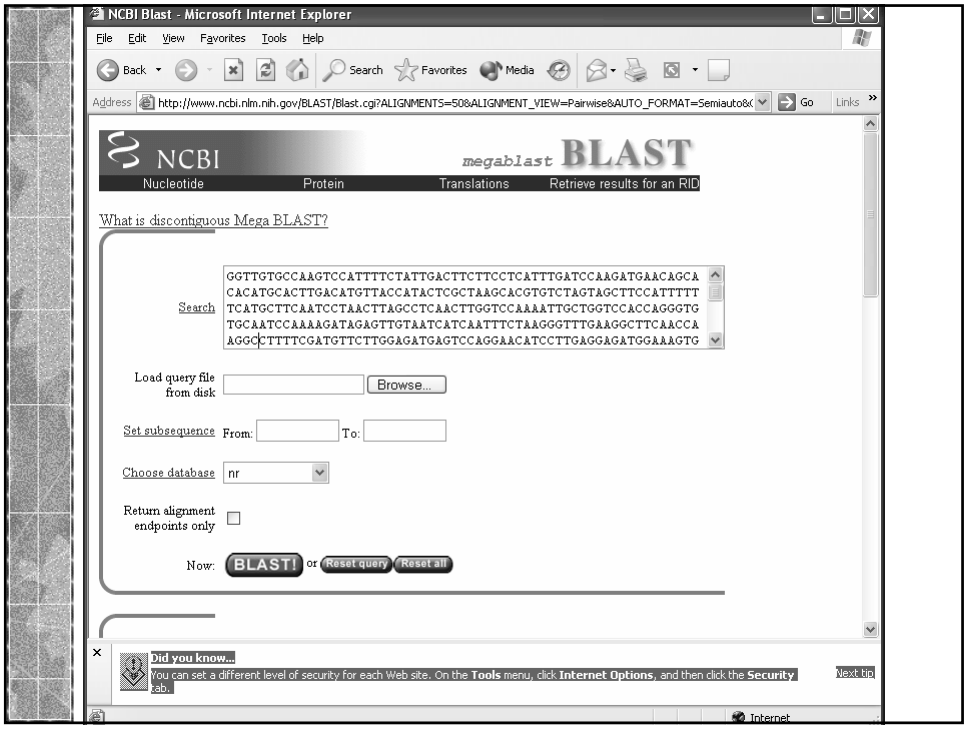
Internet

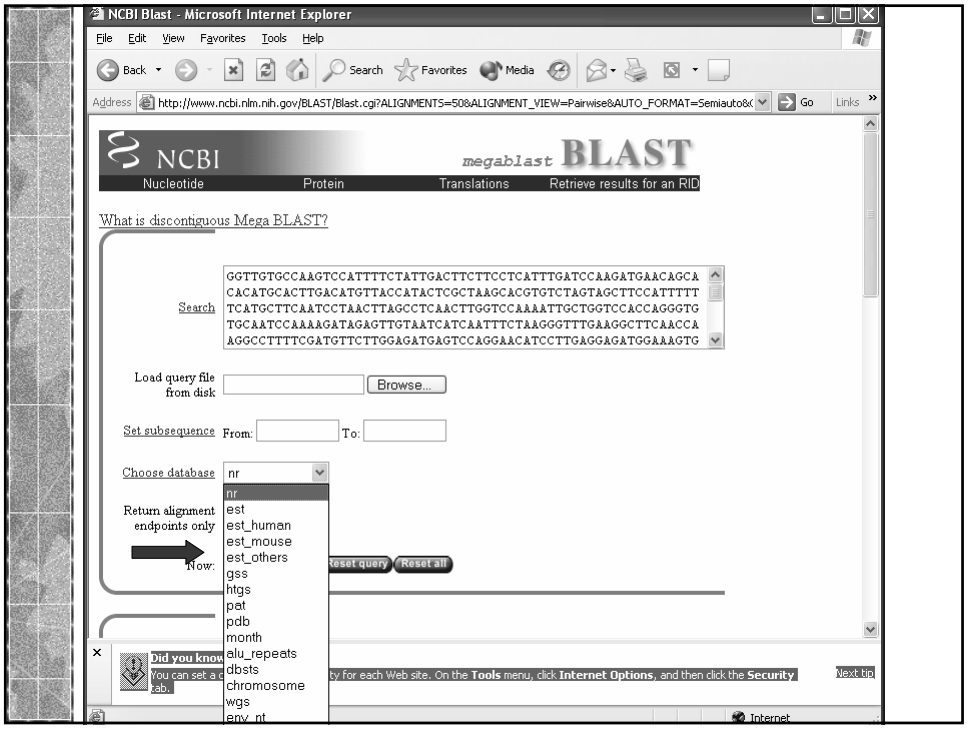
You have a sequence.
Does it have similarity to other known genes???
Copy DNA sequence from file

GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTGAGTCC
TACTTGGCGAAGGTGGCCATCAATAGCGCCTTCGCTGTCTGGAAGGATTGTTTGTGCAGT
CCACACAAGCTCAAACAAAGGCTTTTCAGCTGGCAGAGGATCTGATCCAACAATGACAGC
GCTGCACCATCTCCAAACAAGGCTTGCCCCACAAGGCTGTCAAGATGTGTGTCACTCGG

GCCACGAAATGTGACTGCTGTGATCTCCGA




NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&
Go Links
NCBI
megablast BLAST
Nucleotide Protein Translations Retrieve results for an RID
What is discontiguous Mega BLAST?
Search
GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCA
CACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTT
TCATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTG
TGCAATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCA
AGGCCTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTG
Load query file from disk
Browse...
Set subsequence
From:
To:
Choose database
nr
Return alignment endpoints only
Now:
BLAST!
or
Reset query
Reset all
Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.
Next tip
Internet


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&
Go Links
NCBI
megablast BLAST
Nucleotide Protein Translations Retrieve results for an RID
What is discontiguous Mega BLAST?
Search
GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCA
CACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTT
TCATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTG
TGCAATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCA
AGGCCTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTG
Load query file from disk
Browse...
Set subsequence
From:
To:
Choose database
nr
nr
est
est_human
est_mouse
est_others
gss
htgs
pat
pdb
month
alu_repeats
dbsts
chromosome
wgs
env_nt
Return alignment endpoints only
Now:
Reset query
Reset all
Did you know
You can set a c
ty for each Web site. On the Tools menu, click Internet Options, and then click the Security
Next tip
Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&(Go Links

History

from disk Browse...

Set subsequence From: To:

Choose database est_others

Return alignment endpoints only

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query Glycine max or select from: All organisms

Choose filter ☑ Low complexity ☐ Human repeats ☑ Mask for lookup table only ☐ Mask lower case

Expect 10

Word Size 11

Percent Identity, match, mismatch scores None, 1, -2

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi Go Links

NCBI *formatting* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = (509 letters)

Your search was limited by an Entrez query: Glycine max

The request ID is 1090952418-29109-8211948582.BLASTQ4

Format! or Reset all

The results are estimated to be ready in 23 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-zi Alignment in HTML format

Did you know...

You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

Internet

NCBI formatting BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.
Query = (1509 letters)
Your search was limited by an Entrez query: Glycine max

The request ID is 1089221208-15060-31262165038.BLASTQ4

Format! or Reset all

The results are estimated to be ready in 37 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi Go Links

NCBI results of BLAST

BLASTN 2.2.9 [May-01-2004]

RID: 1090952418-29109-8211948582.BLASTQ4

Database: GenBank non-mouse and non-human EST entries
12,989,815 sequences; 6,991,256,704 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

Query=
(509 letters)

Distribution of 101 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

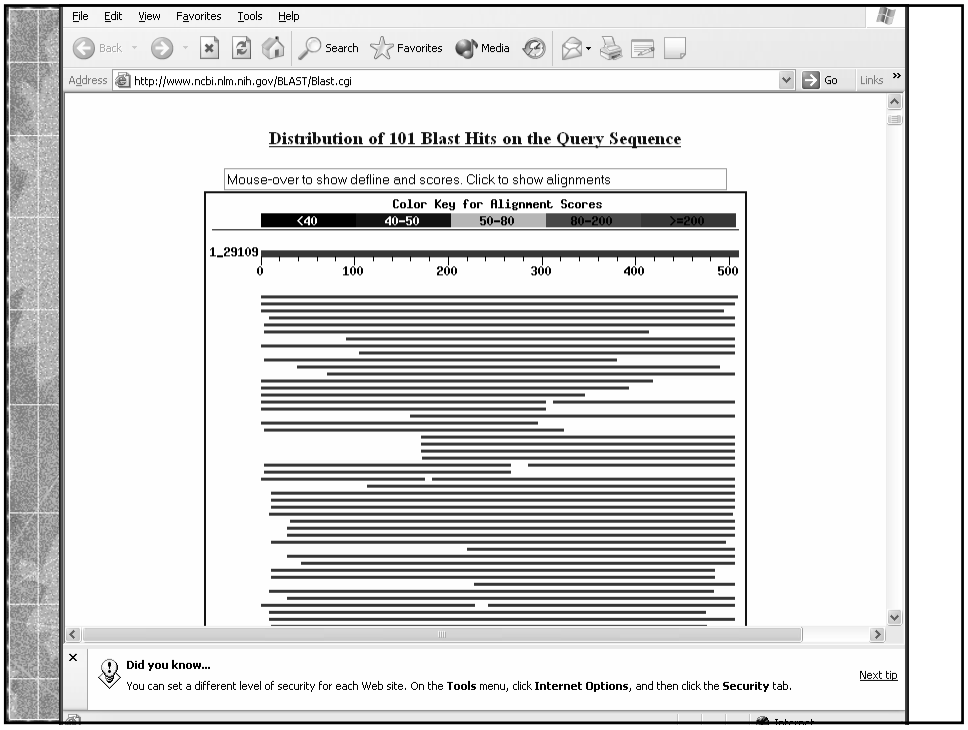
Color Key for Alignment Scores

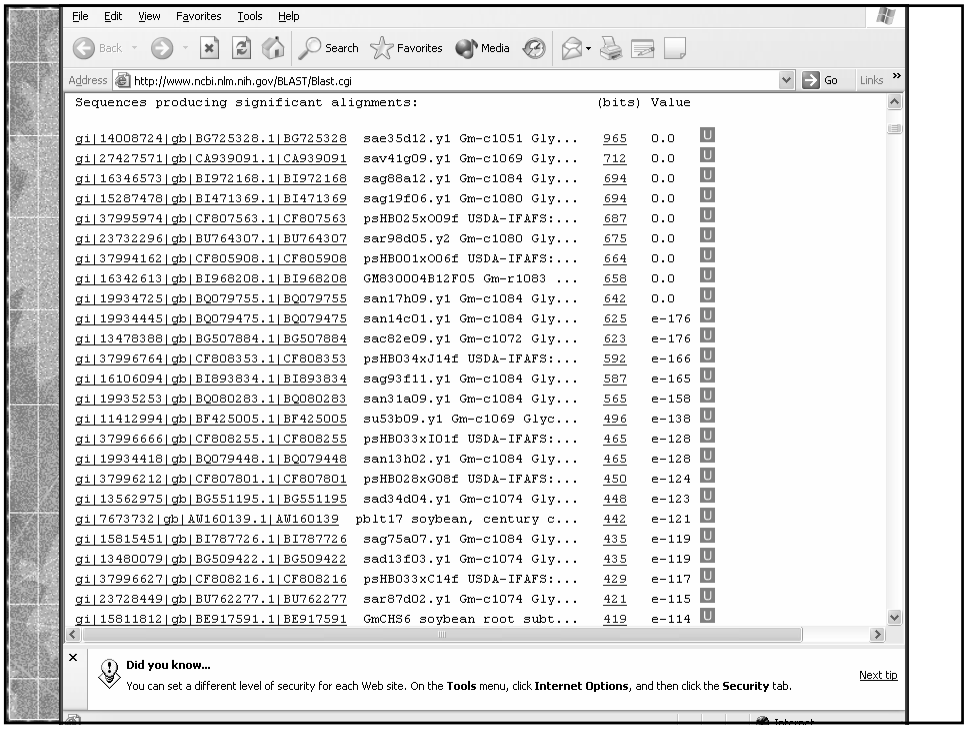
<40	40-50	50-80	80-200	>=200

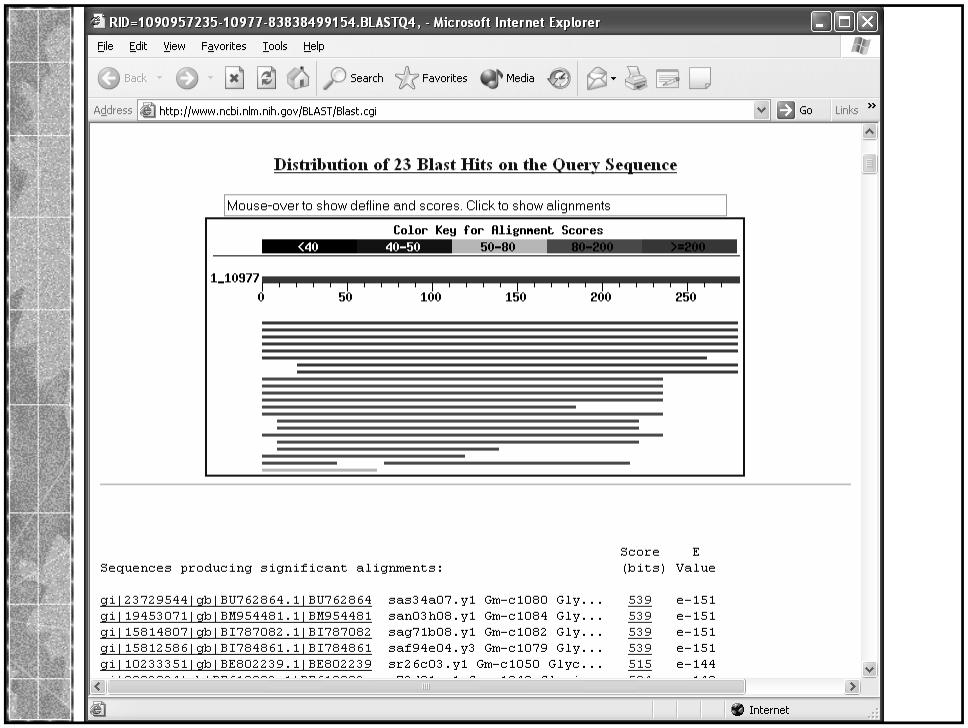
1_29109 0 100 200 300 400 500

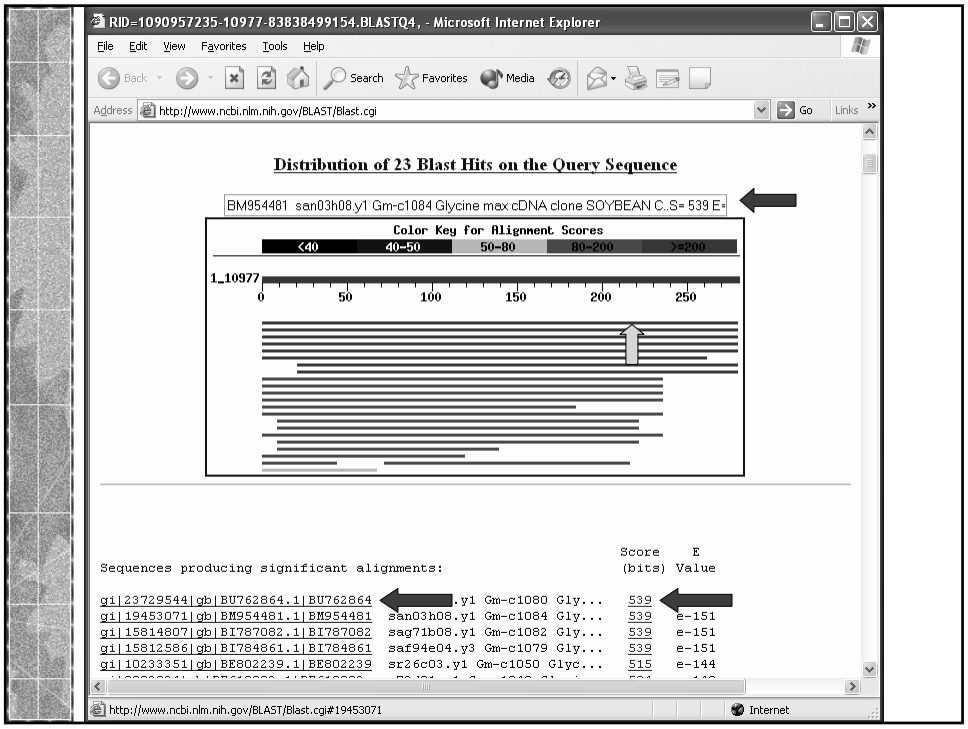
Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip


File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Distribution of 101 Blast Hits on the Query Sequence
Mouse-over to show define and scores. Click to show alignments
Color Key for Alignment Scores
<40 40-50 50-80 80-200 >=200
1_29109
0 100 200 300 400 500
Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.
Next tip


File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Sequences producing significant alignments: (bits) Value
gi|14008724|gb|BG725328.1|BG725328 sae35d12.y1 Gm-c1051 Gly... 965 0.0
gi|27427571|gb|CA939091.1|CA939091 sav41g09.y1 Gm-c1069 Gly... 712 0.0
gi|16346573|gb|BI972168.1|BI972168 sag88a12.y1 Gm-c1084 Gly... 694 0.0
gi|15287478|gb|BI471369.1|BI471369 sag19f06.y1 Gm-c1080 Gly... 694 0.0
gi|37995974|gb|CF807563.1|CF807563 psHB025xO09f USDA-IFAFS:... 687 0.0
gi|23732296|gb|BU764307.1|BU764307 sar98d05.y2 Gm-c1080 Gly... 675 0.0
gi|37994162|gb|CF805908.1|CF805908 psHB001xO06f USDA-IFAFS:... 664 0.0
gi|16342613|gb|BI968208.1|BI968208 GM830004B12F05 Gm-r1083 ... 658 0.0
gi|19934725|gb|BQ079755.1|BQ079755 san17h09.y1 Gm-c1084 Gly... 642 0.0
gi|19934445|gb|BQ079475.1|BQ079475 san14c01.y1 Gm-c1084 Gly... 625 e-176
gi|13478388|gb|BG507884.1|BG507884 sac82e09.y1 Gm-c1072 Gly... 623 e-176
gi|37996764|gb|CF808353.1|CF808353 psHB034xJ14f USDA-IFAFS:... 592 e-166
gi|16106094|gb|BI893834.1|BI893834 sag93f11.y1 Gm-c1084 Gly... 587 e-165
gi|19935253|gb|BQ080283.1|BQ080283 san31a09.y1 Gm-c1084 Gly... 565 e-158
gi|11412994|gb|BF425005.1|BF425005 su53b09.y1 Gm-c1069 Glyc... 496 e-138
gi|37996666|gb|CF808255.1|CF808255 psHB033xI01f USDA-IFAFS:... 465 e-128
gi|19934418|gb|BQ079448.1|BQ079448 san13h02.y1 Gm-c1084 Gly... 465 e-128
gi|37996212|gb|CF807801.1|CF807801 psHB028xG08f USDA-IFAFS:... 450 e-124
gi|13562975|gb|BG551195.1|BG551195 sad34d04.y1 Gm-c1074 Gly... 448 e-123
gi|7673732|gb|AW160139.1|AW160139 pblt17 soybean, century c... 442 e-121
gi|15815451|gb|BI787726.1|BI787726 sag75a07.y1 Gm-c1084 Gly... 435 e-119
gi|13480079|gb|BG509422.1|BG509422 sad13f03.y1 Gm-c1074 Gly... 435 e-119
gi|37996627|gb|CF808216.1|CF808216 psHB033xC14f USDA-IFAFS:... 429 e-117
gi|23728449|gb|BU762277.1|BU762277 sar87d02.y1 Gm-c1074 Gly... 421 e-115
gi|15811812|gb|BE917591.1|BE917591 GmCHS6 soybean root subt... 419 e-114
Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.
Next tip


RID=1090957235-10977-83838499154.BLASTQ4, - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Distribution of 23 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40 40-50 50-80 80-200 >=200
1_10977
0 50 100 150 200 250
Score E
Sequences producing significant alignments: (bits) Value
gi|23729544|gb|BU762864.1|BU762864 sas34a07.y1 Gm-c1080 Gly... 539 e-151
gi|19453071|gb|BM954481.1|BM954481 san03h08.y1 Gm-c1084 Gly... 539 e-151
gi|15814607|gb|BI787082.1|BI787082 sag71b08.y1 Gm-c1082 Gly... 539 e-151
gi|15812586|gb|BI784861.1|BI784861 saf94e04.y3 Gm-c1079 Gly... 539 e-151
gi|10233351|gb|BE802239.1|BE802239 sr26c03.y1 Gm-c1050 Glyc... 515 e-144
Internet


RID=1090957235-10977-83838499154.BLASTQ4, - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Distribution of 23 Blast Hits on the Query Sequence
BM954481 san03h08.y1 Gm-c1084 Glycine max cDNA clone SOYBEAN C..S= 539 E=
Color Key for Alignment Scores
<40 40-50 50-80 80-200 >=200
1_10977
0 50 100 150 200 250
Score E
Sequences producing significant alignments: (bits) Value
gi|23729544|gb|BU762864.1|BU762864 .y1 Gm-c1080 Gly... 539
gi|19453071|gb|BM954481.1|BM954481 san03h08.y1 Gm-c1084 Gly... 539 e-151
gi|15814607|gb|BI787082.1|BI787082 sag71b08.y1 Gm-c1082 Gly... 539 e-151
gi|15812586|gb|BI784861.1|BI784861 saf94e04.y3 Gm-c1079 Gly... 539 e-151
gi|10233351|gb|BE802239.1|BE802239 sr26c03.y1 Gm-c1050 Glyc... 515 e-144
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#19453071
Internet

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=23729544&dopt=GenBank Go Links

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Show: 20 Send to File Features

1: BU762864. sas34a07.y1 Gm-c1...[gi:23729544] Links

```
LOCUS       BU762864                 527 bp    mRNA    linear   EST 02-JUL-2004
DEFINITION  sas34a07.y1 Gm-c1080 Glycine max cDNA clone SOYBEAN CLONE ID:
            Gm-c1080-5414 5' similar to TR:Q9XEX5 Q9XEX5 CYTIDINE DEAMINASE ;,
            mRNA sequence.
ACCESSION   BU762864
VERSION     BU762864.1  GI:23729544
KEYWORDS    EST.
SOURCE      Glycine max (soybean)
  ORGANISM  Glycine max
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids
            ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae;
            Glycine.
REFERENCE   1  (bases 1 to 527)
  AUTHORS   Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna
            ,A., Bolla,B., Marra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C.,
            Wylie,T., Underwood,K., Steptoe,M., Theising,B., Allen,M., Bowers
            ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk
            ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann
            ,R., Waterston,R. and Wilson,R.
  TITLE     Public Soybean EST Project
  JOURNAL   Unpublished (1999)
COMMENT     Contact: Shoemaker R/Public Soybean EST Project
            Public Soybean EST Project
```

Internet

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=23729544&dopt=GenBank Go Links

```
                    200401) was  used to synthesize the cDNA. First-strand
                    synthesis was performed  with 5-methyl dCTP, hence the
                    ligated cDNA was hemimethylated. A  modification of
                    Stratagene's first-strand synthesis primer was used. An
                    'anchor' nucleotide (V=A, C, or G) was added to the 3' end
                    of the primer [GAGAGAGAGAGAGAGAGAACTAGTCTCGAG(T)18V] to
                    anchor the primer at the 5' end of the poly(A) tract.
                    After second-strand synthesis, the cDNA ends were filled
                    in with cloned Pfu DNA, ligated to EcoRI adapters and
                    subsequently phosphorylated. The cDNA was then
                    precipitated and redissolved in sterile, RNase-,
                    DNase-free water. The XhoI site within the first- strand
                    synthesis primer was then restricted by digestion with
                    XhoI from Promega (40U/ul); all XhoI sites in the cDNA
                    would be protected by their hemimethylated status. The
                    cDNA constructs were size-fractionated with  a 500bp
                    cutoff, using Sephacryl S-500 High Resolution (Pharmacia
                    Biotech) in a 2-mm diameter column and a bed volume of
                    approximately 1ml. The column eluent was precipitated,
                    redissolved, and ligated into Stratagene's pBluescript II
                    XR Predigested vector (pBluescript II SK(+) vector that
                    has been digested with EcoRI and XhoI, and phosphorylated
                    by Stratagene). This library was constucted in the
                    laboratory of Dr. Paul Keim and Dr. Virginia H. Coryell at
                    Northern Arizona University."
ORIGIN
        1 tacagcgcgt ccccctccgg cgtcgcgctt ctcgattcca aggggaatgt ttttaaaggc
       61 tcttacattg agtccgctgc ttataacccc agcttgggac cgcttcaggc cgccatcgtc
      121 gccttcatcg ccggcggcgg tggggattat gaagagattg ttgcggcggt gttggtggag
      181 aaggaagggg cggtcatcaa acaggatcac actgcaaggt tgctgctcca ttccatagcg
      241 ccacgctgcc acttcaacaa ttttcttgct tctcaatctc cttcaacttg aacattgttt
      301 ttttttctct ctaccctaaa ttaataatat atatatacta tttgagagtg gatttggatg
      361 cacccttacc tgacctcacc ctgtaataat gtgattgtca cggatagaca tggtgatggt
      421 gtataataat ccccttccta tttatgccta tgtggttcta atattttaag acaagctttt
      481 tctttttgta tattgttaca tctaatgaag aatattgtcc tgtcaaa
//
```

Internet

FASTA/BLAST **Statistics**

- E() value is equivalent to standard P value
- Significant if E() < 0.05 (smaller numbers are more significant)
  - The E-value represents the likelihood that the observed alignment is due to chance alone. A value of 1 indicates that an alignment this good would happen by chance with any random sequence searched against this database.
- The histogram should follow expectations (asterisks) except for hits

Interpretation of output

- very low E() values (e-100) are homologs or identical genes
- moderate E() values are related genes
- long list of gradually declining of E() values indicates a large gene family
- long regions of moderate similarity are more significant than short regions of high identity

What this does for you

- You identified what gene is encoded by your clone's sequence
- Perhaps you may have found the function of your gene
- You have more cDNA sequences to add together to build a consensus and perhaps a full-length cDNA

Biological Relevance

- It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the mismatches conservative ones?
- Are the matching regions important structural components of the genes or just introns and flanking regions?

Borderline similarity

- What to do with matches with E() values in the 0.5 -1.0 range?
- this is the **"Twilight Zone"**
- retest these sequences and look for related hits (not just your original query sequence)
- similarity is transitive:

  if **A~B** and **B~C**, then **A~C**

Advanced Similarity Techniques

Automated ways of using the results of one search to initiate multiple searches

- **INCA** (**I**terative **N**eighborhood **C**luster **A**nalysis)
  http://itsa.ucsf.edu/~gram/home/inca/
  - Takes results of one BLAST search, does new searches with each one, then combines all results into a single list
  - JAVA applet, compatibility problems on some computers
- **PSI BLAST**
  http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html
  - Creates a "position specific scoring matrix" from the results of one BLAST search
  - Uses this matrix to do another search
  - builds a family of related sequences
  - can't trust the resulting e-values

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank Go Links

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Show: 20 Send to File Features

1: BG725328. sae35d12.y1 Gm-c1...[gi:14008724] Links

```
LOCUS       BG725328                 510 bp    mRNA    linear   EST 22-JUL-2004
DEFINITION  sae35d12.y1 Gm-c1051 Glycine max cDNA clone GENOME SYSTEMS CLONE
            ID: Gm-c1051-7103 5' similar to SW:CHS6_SOYBN P30080 CHALCONE
            SYNTHASE 6 ;, mRNA sequence.
ACCESSION   BG725328
VERSION     BG725328.1  GI:14008724
KEYWORDS    EST.
SOURCE      Glycine max (soybean)
  ORGANISM  Glycine max
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids
            ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae;
            Glycine.
REFERENCE   1  (bases 1 to 510)
  AUTHORS   Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna
            ,A., Bolla,B., Marra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C.,
            Wylie,T., Underwood,K., Steptoe,M., Theising,B., Allen,M., Bowers
            ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk
            ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann
            ,R., Waterston,R. and Wilson,R.
  TITLE     Public Soybean EST Project
  JOURNAL   Unpublished (1999)
```

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank Go Links

```
                     /db_xref="taxon:3847"
                     /clone="GENOME SYSTEMS CLONE ID: Gm-c1051-7103"
                     /tissue_type="floral meristematic mRNA"
                     /lab_host="DH10B"
                     /clone_lib="Gm-c1051"
                     /note="Vector: pBluescript II SK+; Site_1: EcoRI; Site_2:
                     XhoI; The cDNA library was constructed from floral
                     meristematic mRNA provided by Dr. Halina Knap of Clemson
                     University. Complementary DNA was synthesized from mRNA
                     using a primer consisting of a poly(dT) sequence with a
                     XhoI restriction site. EcoRI adapters were ligated to the
                     blunt-ended cDNA fragments followed by XhoI digestion. The
                     cDNA fragments were directionally cloned into the
                     EcoRI-XhoI restriction site of the pBluescript vector. The
                     ligated cDNA fragments were transformed into DH10B host
                     cells (GibcoBRL). This library was constructed in the
                     laboratory of Dr. Randy Shoemaker."
ORIGIN
        1 ggttgtgcca agtccatttt ctattgactt cttcctcatt tgatccaaga tgaacagcac
       61 acatgcactt gacatgttac catactcgct aagcacgtgt ctagtagctt ccatttttc
      121 atgcttcaat cctaacttag cctcaacttg gtccaaaatt gctggtccac cagggtgtgc
      181 aatccaaaag atagagttgt aatcatcaat ttctaagggt ttgaaggctt caaccaaggc
      241 cttttcgatg ttcttggaga tgagtccagg aacatccttg aggagatgga aagtgagtcc
      301 tacttggcga aggtggccat caatagcgcc ttcgctgtct ggaaggattg tttgtgcagt
      361 ccacacaagc tcaaacaaag gcttttcagc tggcagagga tctgatccaa caatgacagc
      421 agctgcacca tctccaaaca aggcttgccc cacaaggctg tcaagatgtg tgtcactcgg
      481 gccacgaaat gtgactgctg tgatctccga
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

Next tip

FASTA format

- One of three formats used for sequences
- Begins with single-line description followed by sequence data
- Description line starts with “>”
- Example:
- >gi|532319|pir|TVFV2E|TVFV2E envelope protein
  ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
  QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
  HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
  MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
  TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
  APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
  LAAVEAQQQMLKLTIWGVK

The nucleic acid codes supported are:

A --> adenosine	M --> A C (amino)
C --> cytidine	S --> G C (strong)
G --> guanine	W --> A T (weak)
T --> thymidine	B --> G T C
U --> uridine	D --> G A T
R --> G A (purine)	H --> A C T
Y --> T C (pyrimidine)	V --> G C A
K --> G T (keto)	N --> A G C T (any)

- gap of indeterminate length

accepted amino acid codes are:

A alanine
B aspartate or asparagine
C cystine
D aspartate
E glutamate
F phenylalanine
G glycine
H histidine
I isoleucine
K lysine
L leucine
M methionine
N asparagine

P proline
Q glutamine
R arginine
S serine
T threonine
U selenocysteine
V valine
W tryptophan
Y tyrosine
Z glutamate or glutamine
X any
* translation stop

- gap of indeterminate length

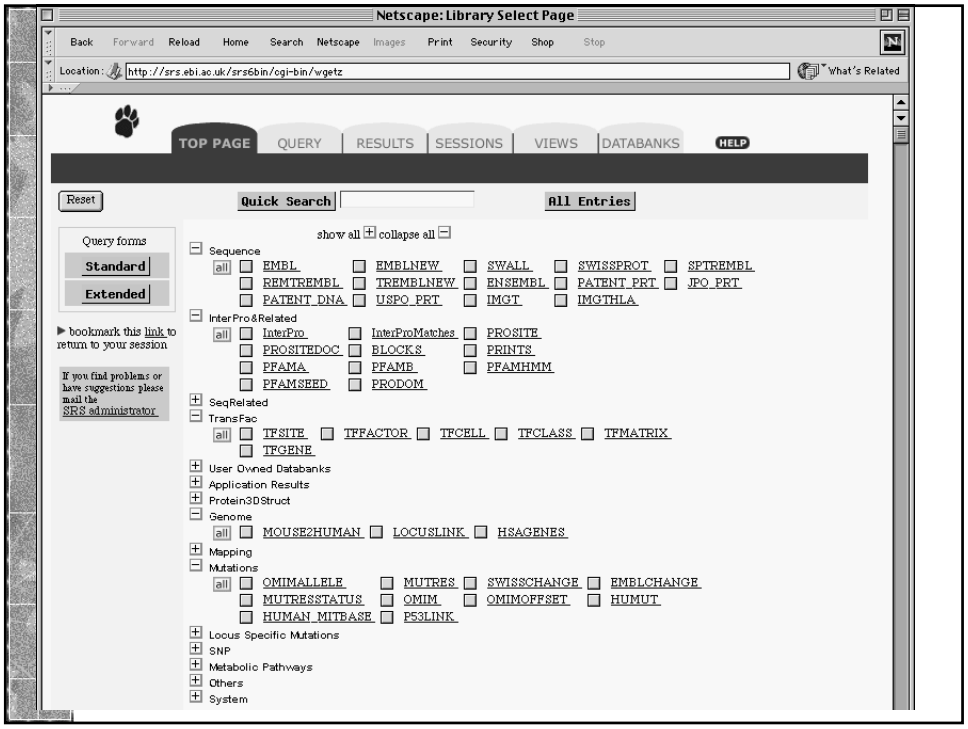
Types of data integrated in genome browsers

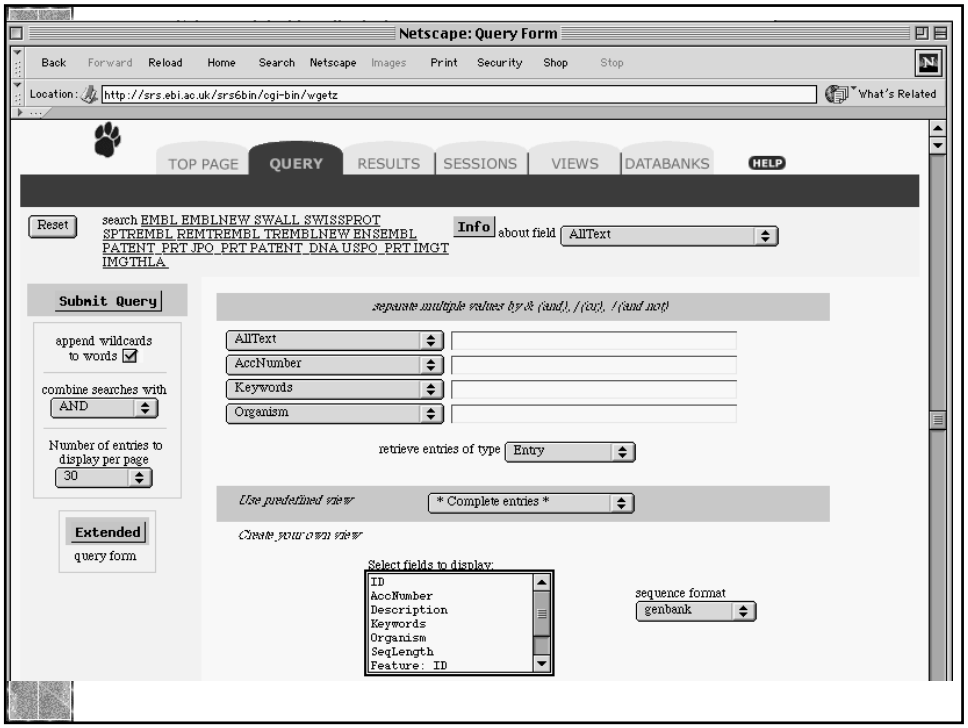
- Genomic sequence
- RefSeq mRNAs (non-redundant)
- GenBank mRNAs (redundant)
- ESTs
- Gene predictions
- SNPs
- Homologous sequences from other organisms
- STSs

Other Sequence Search Tools

- SRS (Sequence Retrival Service) was created by Dr. Thure Etzold: *[CABIOS 9(1); 49-57 (1993)]*
- It is a meta search engine for all types of biological data in hundreds of databases as well as about 20 sequence analysis programs
- SRS can be accessed over the WWW from many servers (mostly in Europe):

http://srs.ebi.ac.uk/
http://www.infobiogen.fr/srs6bin/cgi-bin/wgetz?-page+top
http://www.sanger.ac.uk/srs6bin/cgi-bin/wgetz?-page+top
http://iubio.bio.indiana.edu/srs6bin/cgi-bin/wgetz?-page+top



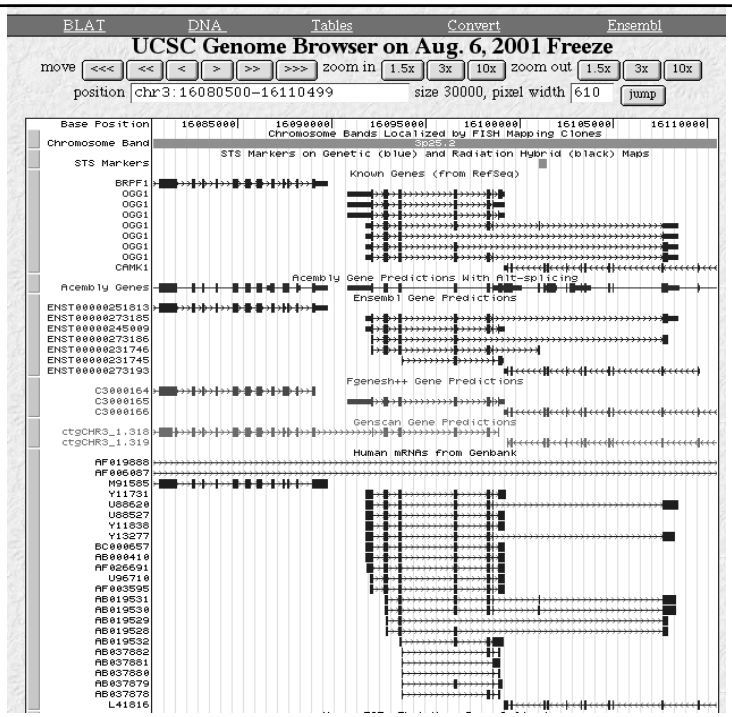


Why So Many Databases?

- If GenBank has all sequence data and Entrez is such a good query tool, then why are there so many other sequence databases?
  - Specialized data (single species, immunoglobulins, etc.)
  - Better annotation (i.e. SwissProt)
  - Sequences linked to other data (ACEDB)
  - Subbornness and local pride - EMBL, DDBJ
- Well designed databases are interlinked with others for supplemental data
- It is very hard to get all relevant information across all databases for any gene

Other Genetic Databases

- Genome Sequence - where does a gene fall on the genome
  - integrate multiple layers of information
    - > Sequence contigs, mRNAs, predicted exons, etc.
  - Single species?
- ESTs: dbEST @ NCBI
- SNPs: dbSNP @ NCBI,
  http://snp.cshl.org (SNP Consortium)
- Metabolism/Pathways
- Gene Function (Genome Ontology)
- Protein motifs/domains and protein families



Genome Databases

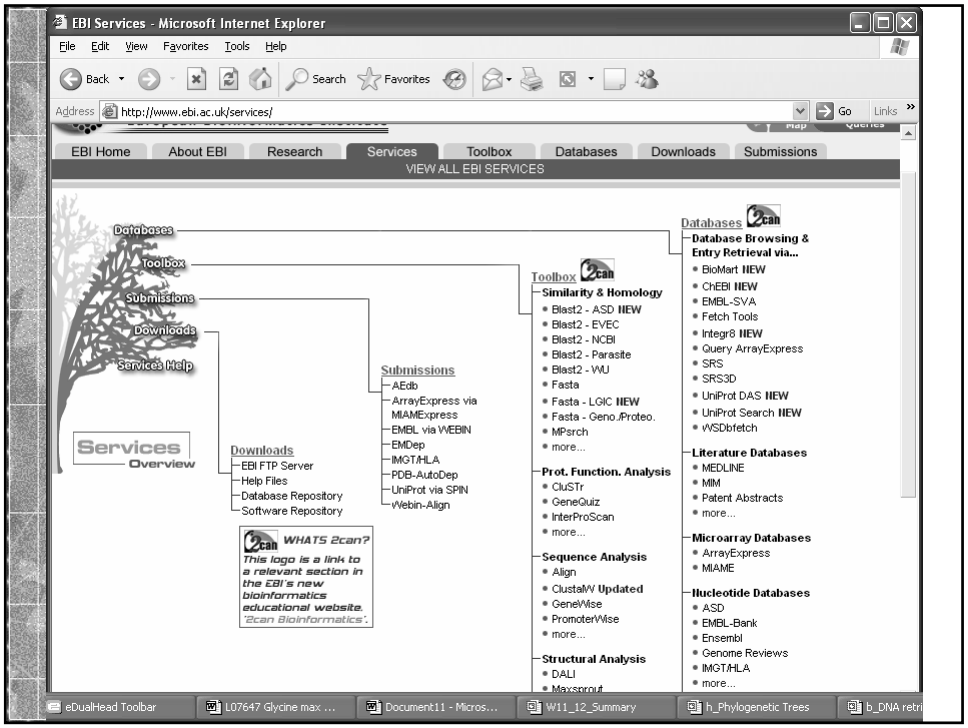
- New area - in desperate need of development

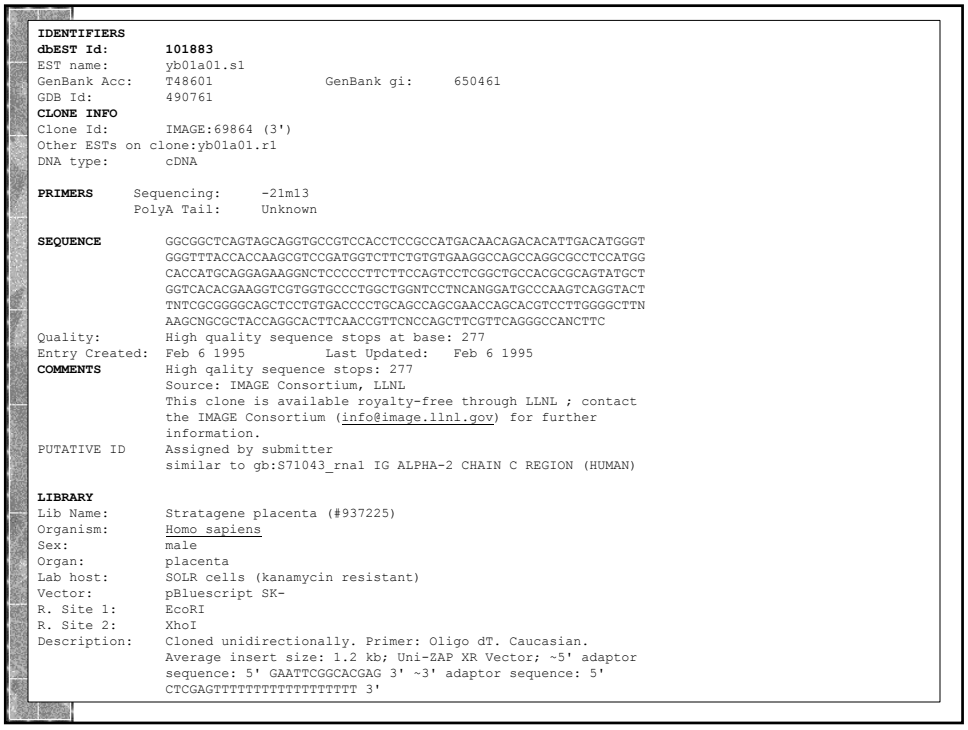
  Chromosomes::Sequence::Contigs::Clones::
  STS Markers::Genetic Markers::Genes::
  Features::Expression data::Phenotype
- No single database can hold it all
- UCSC is probably the best right now

  genome.ucsc.edu
- Need a data exchange and linkage infrastructure

European Bioinformatics Institute

- Products and services
- Databases
  - Literature
  - Microarray
  - nucleotide
- Toolbox with software
  - Similarity searches
  - Protein function
  - Sequence analysis
  - Structure
- http://www.ebi.ac.uk/





```
IDENTIFIERS
dbEST Id:        101883
EST name:        yb01a01.s1
GenBank Acc:     T48601              GenBank gi:     650461
GDB Id:          490761
CLONE INFO
Clone Id:        IMAGE:69864 (3')
Other ESTs on clone:yb01a01.r1
DNA type:        cDNA

PRIMERS      Sequencing:     -21m13
             PolyA Tail:     Unknown

SEQUENCE         GGCGGCTCAGTAGCAGGTGCCGTCCACCTCCGCCATGACAACAGACACATTGACATGGGT
                 GGGTTTACCACCAAGCGTCCGATGGTCTTCTGTGTGAAGGCCAGCCAGGCGCCTCCATGG
                 CACCATGCAGGAGAAGGNCTCCCCCTTCTTCCAGTCCTCGGCTGCCACGCGCAGTATGCT
                 GGTCACACGAAGGTCGTGGTGCCCTGGCTGGNTCCTNCANGGATGCCCAAGTCAGGTACT
                 TNTCGCGGGGCAGCTCCTGTGACCCCTGCAGCCAGCGAACCAGCACGTCCTTGGGGCTTN
                 AAGCNGCGCTACCAGGCACTTCAACCGTTCNCCAGCTTCGTTCAGGGCCANCTTC
Quality:         High quality sequence stops at base: 277
Entry Created:   Feb 6 1995          Last Updated:   Feb 6 1995
COMMENTS         High qality sequence stops: 277
                 Source: IMAGE Consortium, LLNL
                 This clone is available royalty-free through LLNL ; contact
                 the IMAGE Consortium (info@image.llnl.gov) for further
                 information.
PUTATIVE ID      Assigned by submitter
                 similar to gb:S71043_rna1 IG ALPHA-2 CHAIN C REGION (HUMAN)

LIBRARY
Lib Name:        Stratagene placenta (#937225)
Organism:        Homo sapiens
Sex:             male
Organ:           placenta
Lab host:        SOLR cells (kanamycin resistant)
Vector:          pBluescript SK-
R. Site 1:       EcoRI
R. Site 2:       XhoI
Description:     Cloned unidirectionally. Primer: Oligo dT. Caucasian.
                 Average insert size: 1.2 kb; Uni-ZAP XR Vector; ~5' adaptor
                 sequence: 5' GAATTCGGCACGAG 3' ~3' adaptor sequence: 5'
                 CTCGAGTTTTTTTTTTTTTTTTTT 3'
```

Database Search Strategies

- General search principles - not limited to sequence (or to biology)
- Use accession numbers whenever possible
- Start with broad keywords and narrow the search using more specific terms
- Try variants of spelling, numbers, etc.
- Search all relevant databases
- **Be persistent!!**

What we covered today

- Retrieving a known DNA sequence
- Similarity searching with a DNA sequence
- BLAST