NCBI Genome Resources

Eric W. Sayers, PhD
National Center for Biotechnology Information

Computational Genomics
CSHL
November 7–13, 2007

NCBI Resources

• Part 1: Biological Models in Entrez
  – The NCBI Entrez System
  – Sequences of a Model
  – Protein Annotations
    ** Intermission **
• Part 2: Relationships between Models
  – Assemblies: Genomes and Homologene
  – DNA and RNA Annotations
  – NCBI Web BLAST
Assemblies: Entrez Genome Project

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Release</th>
<th>Depth</th>
<th>Status</th>
<th>Method</th>
<th>Assembly Status</th>
<th>Center/Institution</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anolis carolinensis</td>
<td>1.04Gbp</td>
<td>1500</td>
<td>In Progress</td>
<td>Chromosome-based</td>
<td>NCBI-NIH Integrated Sequencing Center</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Betta splendens</td>
<td>2.09Gbp</td>
<td>1000</td>
<td>In Progress</td>
<td>WGS</td>
<td>NCBI-NIH Integrated Sequencing Center</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Canis lupus</td>
<td>2.70Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Washington University, OFB, ND</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Canis lupus</td>
<td>2.70Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Washington University, OFB, ND</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>1.14Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Broad Institute</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Canis lupus</td>
<td>2.70Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Broad Institute</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>1.14Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Broad Institute</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Canis lupus</td>
<td>2.70Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Broad Institute</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>1.14Gbp</td>
<td>500</td>
<td>In Progress</td>
<td>WGS</td>
<td>Broad Institute</td>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>

Links to Genomic Sequences

1. AAC000000000 [Report]
   Canis familiaris, whole genome shotgun sequencing project
   [g63145786][g63145786][g63145786][g63145786][g63145786]

2. NC_006532 [Report]
   Canis familiaris chromosome 7, whole genome shotgun sequence
   [g74032736][g74032736][g74032736][g74032736][g74032736]

3. NC_006530 [Report]
   Canis familiaris chromosome 6, whole genome shotgun sequence
   [g74032736][g74032736][g74032736][g74032736][g74032736]

4. NC_006617 [Report]
   Canis familiaris chromosome 15, whole genome shotgun sequence
   [g74032736][g74032736][g74032736][g74032736][g74032736]

5. NC_006541 [Report]
   Canis familiaris chromosome 35, whole genome shotgun sequence
   [g74032736][g74032736][g74032736][g74032736][g74032736]

Armadillo WGS project

Dog WGS project

Dog chromosomes
**Canis familiaris** Genome Project

*Genome Project = Canis familiaris (dogs)*

Model organism that is notable for its extensive genetic diversity and morphological variation.

**Lineage:** Eukaryota, Metazoa, Chordata, Gnathostomata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Caniformia, Feliformia, Canidae, Canis, Canis familiaris

**Search Map:** Viewer for  
- Available maps:
  - Sequence Maps
  - RH maps

**Clones**

- *Whole Genome Shotgun (WGS)*
  - Example query: wgs[prop] AND gbdiv pri[prop]
  - 12-digit Accessions (e.g. AABC00000000)
  - >1000 projects (> 30 million sequences)
    - Viruses (3), Bacteria (822), Archaea (8)
    - Environmental sequences (13)
    - Eukaryotes (214)

- *Finished sequences:*
  - biomol genomic[prop] AND gbdiv xxx[prop]
    - > 37 million sequences
  - Draft sequences: gbdiv htg[prop] (>100,000 sequences)
From Clones/WGS to Chromosomes

1: AAXX0000000

Reports

Canis familiaris, whole genome shotgun sequencing project
g[63145786]gb|AAXX000000002|AAXX02000000[63145786]

GenBank

NCBI FieldGuide

Chromosome 17 Assembly

36: NC_006599

Reports

Canis familiaris chromosome 17, whole genome shotgun sequence
g[74031862]ref|NC_006599.2|NC_006599[74031862]

RefSeq

NC – RefSeq chromosome

WGS contigs (NW) separated by gaps

FEATURES

<table>
<thead>
<tr>
<th>Location/Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>source 1..57347517</td>
</tr>
<tr>
<td>/organism=&quot;Canis familiaris&quot;</td>
</tr>
<tr>
<td>/mol_type=&quot;genomic DNA&quot;</td>
</tr>
<tr>
<td>/db_xref=&quot;canon:9615&quot;</td>
</tr>
<tr>
<td>/chromosome=&quot;17&quot;</td>
</tr>
<tr>
<td>/breed=&quot;beaver&quot;</td>
</tr>
</tbody>
</table>

CONTIG

join|gap(3000000)NU 876263.111.51591990, gap(1000), NU 876284.111.22731627 |

NW – RefSeq WGS contig
A WGS Contig (NW)  

**Continued:**

**Features:**
- Source: NCBI  
- **Location/Qualifiers:**
  - chromosome: "17"  
- **Contig:**
  - join: constant, A000019551:1..6159 (1), NW  
  - complement: A000019547:11..26309 (1), NW  
  - complement: A000019548:11..26309 (1), NW  
  - complement: A000019549:11..26309 (1), NW  

---

A Clone-based Contig (NT)

**Continued:**

**Features:**
- Source: NCBI  
- **Location/Qualifiers:**
  - chromosome: "2"  
- **Contig:**
  - join: constant, A000023149:11..18440, NT  
  - NT  
  - NT – RefSeq clone-based contig

---
Let’s Do a Search!

Search Gene for thyroid peroxidase (TPO)

tpo [sym] AND canis familiaris [organism]

[Gene/protein name] (if [sym] doesn’t work)

| All: 1 | Current: 1 | Genes: Genomes 1 | SNP: GeneView: 0 |

TPO

thyroid peroxidase [Canis familiaris]
Chromosome: 17
GeneID: 403521

Only 1 record!
Why not start all your NCBI searches this way?

Gene Annotation on a Chromosome (NC)

thyroid peroxidase (TPO)

exons
mRNA
CDS
protein
Genome Annotation in Entrez Nucleotide

NCBI FieldGuide

GenBank Components (clones, WGS) → NT/NW Contigs → NC

Assembly

Components

Genome Components

Master

NM/XM

mRNA

Components

Genome Annotation Links

curated mRNA

human chromosome 2

the 18 contigs of the chromosome 2 assembly
Beyond GenBank: The Trace Archive

Find the link under Hot Spots on the Home Page

News, Events and Notifications

Enter a query string [Enter Query Builder] on 11 members

New query builder!

Last week Top 10 Arrivals (06/04/2006 - 06/10/2006)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETRODUS VAMPYRUS</td>
<td>1,452,491</td>
</tr>
<tr>
<td>CHLOROPUS HUMANNI</td>
<td>1,159,836</td>
</tr>
<tr>
<td>ELYSIUS DANAE</td>
<td>742,934</td>
</tr>
<tr>
<td>PHYTHODORMUS INFESTANS</td>
<td>742,744</td>
</tr>
<tr>
<td>Hyperopus POLIARIS</td>
<td>491,174</td>
</tr>
<tr>
<td>MUS MUSCULUS</td>
<td>351,068</td>
</tr>
<tr>
<td>SCROHAUMBUS COLOR</td>
<td>169,314</td>
</tr>
</tbody>
</table>

New design of NCBI Trace Archive

Please be advised that our pages got a new look. We tested the pages with most existing web browser but if you are experiencing any problem with a page, please try to reload it with the 'SHIFT' button pressed. If that does not help please send a message with the exact problem (see the link at the bottom of the page) and provide us with your e-mail address, so we would be able to solve the problem with you directly.

Please check out our new query builder with the suggestion box. Simply start typing a field's value, and the query builder will help you to choose the right one by showing all available values if any.

Short-tailed opossum traces
Annotations on a Model

DNA → RNA → Protein

- SNP
- OMIM
- Structure
- CDD

- Genome
- CoreNucleotide
- Protein
- Gene

Sequence Polymorphisms

- SNP
- OMIM

General Polymorphisms
- Primary database of submitted SNPs
- Curated database of reference SNPs
- Contains more than just SNPs:
  - True SNPs
  - MNP (multiple nucleotide)
  - Insertions
  - Deletions
  - Microsatellites
  - Mixed
  - No variation (constant)

Human Phenotypes
- Clinical literature database
- Curated at Johns Hopkins Univ
- Links human genes and genetic disorders to human disease
- Lists allelic variants that have clinical consequences

Variations in SNP are not necessarily in OMIM, and vice versa!
Linking to SNP

NCBI FieldGuide

Entrez Gene - TPO

Links to SNP are also available from CoreNucleotide and Protein

Entrez SNP

NCBI FieldGuide

SNP: rs11888851 (Homo sapiens)

CAGGGCGCTGCCAAAACCGATGCC [C/G] TTTCTTTTCAAAGGCCCCTCGTGCC

Reference SNP (rsSNP) Cluster Report: rs11888851

SNP UID: rs#

SNP Details are categorized in the following sections:

Submission, Field, Reference, GeneView, Map, Variation, Validation

Submitter records for this rsSNP Cluster

The submission s17642223 has been submitted by a member of the Entrez Gene team and was used to initiate sequence for rs11888851 during BLAST build.

primary data: ss#
### GeneView in dbSNP

| rs14176261 | N.D. | synonymous | G | Ile | 3 | 4 |
| rs17222496 | N.D. | synonymous | A | Pro | 3 | 204 |
| rs183846 | 14218813 | N.D. | synonymous | T | Ser | 3 | 267 |
| rs320137 | N.D. | synonymous | T | Ser | 3 | 273 |
| rs3125377 | N.D. | synonymous | C | Thr | 2 | 298 |
| rs1184258 | N.D. | synonymous | A | Met | 1 | 218 |
| rs132609 | N.D. | synonymous | T | Pro | 3 | 115 |

### SNP Detail – Genotype Data

**refSNP ID:** rs28991293  
**Organism:** human (Homo sapiens)  
**Molecule Type:** Gene  
**Created Updated in build:** 125/125  
**Map to Genome Build:** 36.1

- **Allele**:
  - **Allele Name**: SNP
  - **SNP Type**: single nucleotide polymorphism
  - **Allele Code**: C/T
  - **Ancestral Allele**: T

### Population Diversity

<table>
<thead>
<tr>
<th>Sample Asserternable</th>
<th>Genotypes</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>#94</td>
<td>CT: 1.000</td>
<td>T: 1.000</td>
</tr>
<tr>
<td>#44</td>
<td>CT: 1.000</td>
<td>T: 1.000</td>
</tr>
<tr>
<td>#21</td>
<td>CT: 0.929</td>
<td>T: 1.000</td>
</tr>
<tr>
<td>#15</td>
<td>CT: 1.000</td>
<td>T: 1.000</td>
</tr>
<tr>
<td>#23</td>
<td>0.374: 0.536: 0.035: 0.082: 0.193</td>
<td></td>
</tr>
</tbody>
</table>

**Total Samples:** 109  
Genotype Detail
dbGaP – Whole Genome Associations

Data is either open access or controlled access

Annotations on a Model
What is UniGene?

A gene-oriented view of expressed sequences

- MegaBlast based automated sequence clustering
- Now informed by genome hits
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes known genes and uncharacterized ESTs
- Useful for gene discovery and selection of mapping reagents

Organisms in UniGene

**Top Ten**
1. Human
2. Mouse
3. Rat
4. Corn
5. Zebrafish
6. Pig
7. Cow
8. Frog (X. tropicalis)
9. Rice
10. Wheat

**Mammalia**
- *Homo sapiens*
- *Mus musculus*
- *Rattus norvegicus*

**Aves**
- *Gallus gallus*
- *Xenopus laevis*

**Arthropoda**
- *Drosophila melanogaster*

**Phyla**
- *Trematoda*
- *Schistosoma mansoni*
- *Schistosoma japonicum*

**Cnidaria**
- *Hydra vulgaris*
- *Hydra magnipapillata*

**Bryozoa**
- *Physoplax pataens*

**Contralesca**
- *Pisum sativum*

**Eukaryotes**
- *Arabidopsis thaliana*
- *Citrus sinensis*
- *Glycine max*
- *Helianthus annuus*
- *Lactuca sativa*
- *Lotus corniculatus*
- *Lygodium japonicum*
- *Malus domestica*
- *Medicago truncatula*
- *Peploides trimera x Peploides trimera var. solanum*
- *Visna sativa*

**Myxozoa**
- *Hordoneum vulgare*
- *Oryza sativa*
- *Saccharum officinarum*
- *Sorghum bicolor*
- *Triticum aestivum*
- *Zea mays*

**Chlorophyta**
- *Chlorella vulgaris*
- *Chlamydomonas reinhardii*

**Chromadorea**
- *Chromadorea elegans*

**Eukaryotes**
- *Dicytostelium discoideum*
- *Arachnida*
- *Drosophila melanogaster*

**Fungi**
- *Magnaporthe grisea*
- *Neurospora crassa*

**Protista**
- *Glycera elongata*

**Bacteria**
- *Phytophthora infestans*
Finding UniGene Clusters

Universal Gene Cluster for TPO

UniGene Hs:467554  Homo sapiens  TPO

Thyroid peroxidase (TPO)

Selected Protein Similarities

- Comparison of sequences in UniGene with proteins of a genome. The alignments can suggest function of a gene.

<table>
<thead>
<tr>
<th>Species</th>
<th>Protein Name</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. elegans</td>
<td>C. elegans</td>
<td>100%</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>D. melanogaster</td>
<td>100%</td>
</tr>
<tr>
<td>M. musculus</td>
<td>M. musculus</td>
<td>100%</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>R. norvegicus</td>
<td>100%</td>
</tr>
</tbody>
</table>

Gene Expression

- Tissues and development stages from the gene's sequence expression. Links to other WGBS expression resources.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Expression Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>View expression levels using UniGene Profileviewer</td>
</tr>
</tbody>
</table>

Note: Highly represented (1.9% of 895,421) in PubMol, GCAP, Th3, GEO.

EST Sequences (32 of 32) [Show all sequences]

<table>
<thead>
<tr>
<th>EST ID</th>
<th>Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000193.2</td>
<td>Clone</td>
<td>100%</td>
</tr>
<tr>
<td>C000194.2</td>
<td>Clone</td>
<td>100%</td>
</tr>
<tr>
<td>C000195.2</td>
<td>Clone</td>
<td>100%</td>
</tr>
<tr>
<td>C000196.2</td>
<td>Clone</td>
<td>100%</td>
</tr>
</tbody>
</table>

NCBI FieldGuide
GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile

Submitted by Experimentalists

Curated by NCBI FieldGuide

GPL
Platform descriptions

GSM
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data “a single experiment”

GDS
Grouping of experiments

Entrez GEO Profiles

Entrez GEO Datasets

GEO Profile

GEO Profile
GEO Dataset

**Accession:** GDS1732
**Title:** Papillary thyroid cancer

**Summary:**
Expression profiling of 7 papillary thyroid carcinoma (PTC) samples. PTC is the most common type of thyroid cancer, representing up to 80% of all malignant thyroid tumors. Results provide insight into potential molecular markers for PTC.

**Platform:** GPL570, Affymetrix GeneChip Human Genome U133 Plus 2.0 Array

**Sample organism:** Homo sapiens

**Feature count:** 5403

**Series:** GSE6879

**Series published:** 01/01/2006

**Last GDS update:** 05/03/2006

---

Viewing Simple Genomes

All are RefSeq NC records in Entrez Genome

- Full chromosomal sequences are provided
- Genes are annotated
- The annotation can be shown graphically and linked to sequence records
Viewing Complex Genomes

NCBI Map Viewer

• Map Viewer Home Page
  - Shows all supported organisms
  - Provides links to genomic BLAST

• Genome Overview Page
  - Provides links to individual chromosomes
  - Shows hits on a genome graphically

• Chromosome Viewing Page
  - Allows interactive views of annotation details
  - Provides numerous maps unique to each genome

Map Viewer Home Page
Map Summary

NCBI FieldGuide

Build 36 vs Build 35

TPO's contig!

View the Assembly near TPO
Assembly of Chr. 2

Assembly of Chromosome 2
From Gene to the Map Viewer

Ensembl gene

UniGene

NCBI gene

Viewing Syntenic Regions

Choose these maps from Maps & Options!
### NCBI Web BLAST

#### BLAST Home
BLAST finds regions of similarity between biological sequences. Learn more about how to use the new BLAST design.

#### BLAST Assembled Genomes
Choose a species genome to search, or list all generic BLAST databases.

- **Human**
- **Mouse**
- **Bat**
- **Anobius rhabdax**
- Genus *mycota*
- Genus *reves*
- Genus *melanogaster*
- Gallus *galus*
- Pan troglodytes
- *Micrurus*
- *Aph melifera*

#### Basic BLAST
Choose a BLAST program to run.

- **Nucleotide blast**
  - Search a nucleotide database using a nucleotide query
    - Algorithms: blastn, megablast, distrigu pipes megablast
- **Protein blast**
  - Search protein database using a protein query
    - Algorithms: blastp, dban, phiblast
- **Nucleotide dban**
  - Search translated nucleotide database using a nucleotide query
- **Protein dban**
  - Search translated protein database using a protein query
- **Nucleotide dban**
  - Search translated nucleotide database using a translated-nucleotide query

---

### Nucleotide BLAST

#### Enter accession number, gi, or FASTA sequence

Or, upload file

Job Title

Enter a descriptive title for your BLAST search.

#### Choose Search Set

- **Database**
  - Human genomic + transcript
  - Mouse genomic + transcript
  - Other (e.g., Human genomic plus transcript [Human G+T])

Enter Query

Optional

Enter an entropy query to limit search.

#### Program Selection

- **Optimize for**
  - Highly similar sequences (megablast)
  - More dissimilar sequences (distrigu pipes megablast)
  - Somewhat similar sequences (blat)

Choose a BLAST algorithm.
Nucleotide Algorithms

**MegaBLAST**
- Useful for long alignments of very similar DNA sequences
- Default word size (W) = 28; best with W ≥ 16
- Faster than blastn, but less sensitive

**Discontiguous MegaBLAST**

| W = 11, t = 16, coding: | 1101101101101101 |
| W = 11, t = 16, non-coding: | 1110010110110111 |
| W = 12, t = 16, coding: | 1111101101101101 |
| W = 12, t = 16, non-coding: | 1110110110110111 |

- **W** = word size; # matches in template
- **t** = template length (window size within which the word match is evaluated)


**blastn**
- "standard" nucleotide BLAST
- Default word size (W) = 11; minimum W = 7

Nucleotide Databases: Human and Mouse

**Megablast, blastn**

- Human genomic and transcript is now the default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences
Sortable Results

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Max score</th>
<th>Fold score</th>
<th>Query coverage</th>
<th>E value</th>
<th>Max ident</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_001255.1</td>
<td>Homo sapiens CDC20 cell division cycle 20 homolog (Cdc20) human</td>
<td>2876</td>
<td>2876</td>
<td>95%</td>
<td>0.0</td>
<td>97%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_000367.17</td>
<td>Homo sapiens chromosome 9</td>
<td>2629</td>
<td>2629</td>
<td>94%</td>
<td>0.0</td>
<td>95%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_909881.1</td>
<td>Homo sapiens chromosome 9</td>
<td>2561</td>
<td>2561</td>
<td>94%</td>
<td>0.0</td>
<td>95%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_909881.1</td>
<td>Homo sapiens chromosome 1</td>
<td>428</td>
<td>428</td>
<td>95%</td>
<td>9e-117</td>
<td>100%</td>
<td>UGO</td>
</tr>
</tbody>
</table>

Separate Sections for Transcript and Genome

Pseudogene on Chromosome 9

Functional Gene on Chromosome 1

Total Score: All Segments

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Max score</th>
<th>Fold score</th>
<th>Query coverage</th>
<th>E value</th>
<th>Max ident</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_001255.1</td>
<td>Homo sapiens CDC20 cell division cycle 20 homolog (Cdc20) human</td>
<td>2876</td>
<td>2876</td>
<td>95%</td>
<td>0.0</td>
<td>97%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_000367.17</td>
<td>Homo sapiens chromosome 9</td>
<td>2629</td>
<td>2629</td>
<td>94%</td>
<td>0.0</td>
<td>95%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_909881.1</td>
<td>Homo sapiens chromosome 9</td>
<td>2561</td>
<td>2561</td>
<td>94%</td>
<td>0.0</td>
<td>95%</td>
<td>UGO</td>
</tr>
<tr>
<td>NM_909881.1</td>
<td>Homo sapiens chromosome 1</td>
<td>428</td>
<td>428</td>
<td>95%</td>
<td>9e-117</td>
<td>100%</td>
<td>UGO</td>
</tr>
</tbody>
</table>

Functional Gene Now First
Nucleotide Databases: Traditional

Selected BLAST Databases for Nucleotides

- **nr (nt)**
  - Traditional GenBank Divisions
  - NM_ and XM_ RefSeqs
- **refseq_rna**
  - NM_ and XM_ RefSeqs
- **refseq_genomic**
  - NC_, NG_ RefSeqs
- **wgs**
  - whole genome shotgun

- **est**
  - EST Division
- **htgs**
  - HTG division
- **gss**
  - GSS division
Limiting the Database Using Entrez

Organism autocomplete

all[filter] NOT mammals[organism]
gene_in_mitochondrion[Properties]
2006-2007 [Modification Date]
biomol_mrna[Properties]
biomol_genomic[Properties]

Algorithm parameters: Nucleotide

• Masks LC sequence (simple repeats)
  • Prevents starting alignment in masked region
  • Allows extensions through masked regions

• Masks species-specific interspersed repeats
  • Essential for genomic query sequences
**Nucleotide vs. Protein BLAST**

Comparing ADSS from *H. sapiens* and *A. thaliana*

<table>
<thead>
<tr>
<th>Human:</th>
<th>NRVTVLGAQWGDEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A.th.:</td>
<td>SQVSLGCQWGDEG</td>
</tr>
</tbody>
</table>

BLASTp finds three matching words
BLASTn finds no match, because there are no 7 bp words

Protein searches are generally more sensitive than nucleotide searches.

**NCBI FieldGuide**

**Translated BLAST**

Particularly useful for nucleotide sequences without protein annotations, such as ESTs or genomic DNA

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastx</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ddd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tblastn</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ddd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tblastx</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ddd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Protein BLAST

**Enter accession number, gi, or FASTA sequence**

**Query subrange**

**Or, upload file**

**Job title**

**Choose Search Set**

- Database: Reference proteins
- Organism: Optional
- Enter organism common name, binomial, or taxid. Only 20 top taxa will be shown.

**Program Selection**

- Algorithm:
  - blastp (protein-protein BLAST)
  - PSI-BLAST (Position-Specific Iterated BLAST)
  - PHI-BLAST (Pattern Hit Initiated BLAST)

### Protein Algorithms

**Position independent matrices**

- **blastp**: protein → protein
- **blastx**: translated nucleotide → protein
- **tblastn**: protein → translated nucleotide
- **tblastx**: translated nucleotide → translated nucleotide

**PHI-BLAST**: Hits must contain an input Prosite pattern

**PSSMs**

**PSI-BLAST**: Generates PSSM from protein hits

**RPS-BLAST**: Searches database of PSSMs (CD-Search)
BLAST Databases for Proteins

**nr** (non-redundant protein sequences)
- GenBank CDS translations
- NP_, XP_ RefSeqs
- PIR, Swiss-Prot, PRF
- PDB (sequences from structures)

**refseq** – NP_, XP_ RefSeqs

**swissprot** – Swiss-Prot

**pat** – patents

**pdb** – sequences with 3D structures

**env_nr** – environmental sequences

Algorithm parameters: Protein

- **Expand**
- **May limit results**
- **Adjust to set stringency**
- **Default statistics adjustment for compositional bias**
- **Off now by default. Conflicts with comp-based stats**
Automatic Short Sequence Adjustment

NCBI FieldGuide

e-value 20000
Word Size 2
Matrix PAM30
Comp Stats Off
Low Comp Filter Off

Formatting Results

NCBI FieldGuide
Sample Alignment Options

Query: NP_000538
Program: blastp
Database: Refseq, excluding primates

Pairwise

flat query-anchored with identities

BLAST TreeView

Query: AAA61217
Program: blastp
Database: Refseq, excluding status=model
Pre-computed BLAST Services

Find in or near the Links menu…

**Related Sequences**
- Nucleotide: very restrictive blastn
- Protein: similar to default blastp (e < 1e-6)

**Conserved Domains**
- RPS-BLAST (CD-Search)

**BLink**
- Graphical view of top 200 proteins

**Related Structure**
- blastp against pdb

**UniGene**
- blastn clustering of ESTs and cDNAs
- blastx of ESTs to reference proteins

**Homologene**
- blastp and blastn between related gene models

**Map Viewer**
- sequence maps
- alignments of mRNAs, GenBank clones, and ESTs to a genomic build

Recent Results and Saved Strategies

Previous BLAST results are saved for ~2 days
**BLAST strategies** can be saved indefinitely in MyNCBI

Login to My NCBI to save search strategies
Genomic BLAST

- These pages provide customized nucleotide and protein databases for each genome
- If a Map Viewer is available, the BLAST hits can be viewed on the maps

BLAST the Chicken Genome

Program

Accession for human TPO mRNA
NCBI Courses (all free!)

NCBI Courses
- Field Guide
- Field Guide Plus
- MiniCourses
- Medical Library Association Course
- PubChem Course
- PowerTools Technical Workshop Series
- Structure Course

Full day introductory course on NCBI web resources
Two-day intensive course on NCBI web resources
Eleven half-day courses, focusing on either a resource or problem
Three-day course on NCBI resources designed for medical librarians providing bioinformatics support
Full day course covering NCBI PubChem
Intense multi-day courses taught at NLM, including NCBI PowerScripting: 4 days on the E-utilities NCBI 4-Pak: Four mini-courses in 2 days
Full day course covering NCBI 3D structure resources

For More Information...

E-mail addresses
- Eric Sayers sayers@ncbi.nlm.nih.gov
- General Help info@ncbi.nlm.nih.gov
- BLAST blast-help@ncbi.nlm.nih.gov

The (free!) NCBI Newsletter
http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook
Follow the link from the NCBI Home Page

The NCBI Education Page