NCBI FieldGuide

Computational Genomics
CSHL
November 7–14, 2006

NCBI Genome Resources
Eric W. Sayers, PhD
National Center for Biotechnology Information

Genomic Biology
**Whole Genome Shotgun Projects**

- Traditional GenBank Divisions; Keyword: WGS
- 12-digit Accessions (eg. AABC00000000)
- 396 species
  - Viruses (1)
  - Bacteria (269)
  - Environmental sequences (1)
  - Archaea (3)
  - Eukaryotes (122) featuring:
    - Cow, Chicken, Rat, Mouse, Dog, Chimpanzee, Human
    - Pufferfish, Zebrafish
    - Honeybee, Anopheles, Fruit Flies, Silkworm
    - Nematode (C. briggsae)
    - Yeasts, Aspergillus
    - Rice

---

**From WGS to RefSeq**

1. AAE00000000.0 Passes
   Canis familiaris, whole genome shotgun sequencing project
   gi[6345788],gi[AAE00000000.0],gi[AAE00000000.0],gi[6345788]

   ![Image]

   ![Image]

   ![Image]
Chromosome 17 Assembly

WGS contigs (NW) separated by gaps

A WGS Contig (NW)

Let's Do a Search!

Search **Gene** for thyroid peroxidase (TPO)

\[\text{tpo} \text{ [sym]} \land \text{canis familiaris} \text{ [organism]}\]

Only 1 record!

Why not start all your NCBI searches this way?
Gene Annotation on a Chromosome (NC)

Thyroid peroxidase (TPO)

Two Views of the Assembly

Map Viewer

NC_006599

Links from Gene

Homologene
Beyond Gene

If your organism is not in Gene...

- WGS sequences in Entrez Nucleotide (wgs[prop])
  - remember the armadillo!
- UniGene: gene-based clusters of cDNAs and ESTs
- Trace Archive

What is UniGene?

A gene-oriented view of sequence entries

- 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. Rice
5. Zebrafish
6. Wheat
7. Pig
8. Frog (X. tropicalis)
9. Corn
10. Chicken

Finding UniGene Clusters

by link

human TPO

by Entrez search

UniGene Cluster for TPO

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Gene Symbol</th>
<th>Organism</th>
<th>Ensembl Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_001914</td>
<td>Human thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Homo sapiens</td>
<td>ENSG00000165569</td>
</tr>
<tr>
<td>NM_009946</td>
<td>Mouse thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Mus musculus</td>
<td>ENSMUSG00000054714</td>
</tr>
<tr>
<td>NM_010063</td>
<td>Rat thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Rattus norvegicus</td>
<td>ENSRNOG00000011576</td>
</tr>
<tr>
<td>NM_001281</td>
<td>Rice thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Oryza sativa</td>
<td>ENSOsaG00000007954</td>
</tr>
<tr>
<td>NM_001290</td>
<td>Zebrafish thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Danio rerio</td>
<td>ENSDARG00000001050</td>
</tr>
<tr>
<td>NM_001291</td>
<td>Wheat thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Triticum aestivum</td>
<td>ENSAatG00000000283</td>
</tr>
<tr>
<td>NM_001292</td>
<td>Pig thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Sus scrofa</td>
<td>ENSuscG00000001340</td>
</tr>
<tr>
<td>NM_001293</td>
<td>Frog (X. tropicalis) thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Xenopus tropicalis</td>
<td>ENSxtrG00000011475</td>
</tr>
<tr>
<td>NM_001294</td>
<td>Corn thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Zea mays</td>
<td>ENSzmgG00000001273</td>
</tr>
<tr>
<td>NM_001295</td>
<td>Chicken thyroid peroxidase (TPO)</td>
<td>TPO</td>
<td>Gallus gallus</td>
<td>ENSgagG00000011802</td>
</tr>
</tbody>
</table>
Trace Archive

Find the link under Hot Spots on the Home Page

News, Events and Notifications

New query builder!

New design of Trace Archive

Please be advised that this page is a new tool. We tested this tool with most existing user browsers, but if you are experiencing any problems with a new browser, we will work to fix it with the latest release. Please let us know if you have any feedback.

Last week Top 10 Articles (6/5/2020 - 6/11/2020)

<table>
<thead>
<tr>
<th>Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title1</td>
<td>Date1</td>
</tr>
<tr>
<td>Title2</td>
<td>Date2</td>
</tr>
<tr>
<td>Title3</td>
<td>Date3</td>
</tr>
<tr>
<td>Title4</td>
<td>Date4</td>
</tr>
<tr>
<td>Title5</td>
<td>Date5</td>
</tr>
<tr>
<td>Title6</td>
<td>Date6</td>
</tr>
<tr>
<td>Title7</td>
<td>Date7</td>
</tr>
<tr>
<td>Title8</td>
<td>Date8</td>
</tr>
<tr>
<td>Title9</td>
<td>Date9</td>
</tr>
<tr>
<td>Title10</td>
<td>Date10</td>
</tr>
</tbody>
</table>

Short-tailed opossum traces

Viewing Simple Genomes

• Full chromosomal sequences are provided
• Genes are annotated
• The annotation can be shown graphically and linked to sequence records
### NC_000913 in Entrez Genome

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Symbol</th>
<th>BLAST Hit ID</th>
<th>Link</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>mutL</td>
<td></td>
<td></td>
<td></td>
<td>Case sensitive!</td>
</tr>
</tbody>
</table>

- Gene ID: mutL
- Gene Symbol: mutL
- BLAST Hit ID: "NC_000913"
- Link: Entrez Gene and Protein
- Note: Case sensitive!

### NC_000913 in Entrez Genome

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Symbol</th>
<th>BLAST Hit ID</th>
<th>Link</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>mutL</td>
<td></td>
<td></td>
<td></td>
<td>Case sensitive!</td>
</tr>
</tbody>
</table>

- Gene ID: mutL
- Gene Symbol: mutL
- BLAST Hit ID: "NC_000913"
- Link: Entrez Gene and Protein
- Note: Case sensitive!

### 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 Summary

Map Content

- Sequence Maps
  - Core assembly
  - Annotation evidence
  - Clones & Markers
  - Polymorphisms
- Genetic Maps
  - Cytogenetic maps
  - Linkage maps
  - Radiation hybrid maps

View the Assembly near TPO
Curated CD Record

PSSM for CD00054

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network clients
NCBI Web BLAST

Megablast: NCBI's Genome Annotator

- Long alignments of very similar DNA sequences
- Default word size $W = 28$; best with $W > 16$
- Concatenation of query sequences
- Faster than blastn; less sensitive
NCBI FieldGuide

MegaBLAST

Templates for Discontiguous MegaBLAST

\[
\begin{array}{l}
W = 11, t = 16, \text{coding: } 1101101101101101 \\
W = 11, t = 16, \text{non-coding: } 1110010110110111 \\
W = 12, t = 16, \text{coding: } 1111101101101101 \\
W = 12, t = 16, \text{non-coding: } 1110110110110111 \\
W = 11, t = 18, \text{coding: } 101101101101101101 \\
W = 11, t = 18, \text{non-coding: } 111010110010110111 \\
W = 12, t = 18, \text{coding: } 101010101010101010 \\
W = 12, t = 18, \text{non-coding: } 11101001010010010111 \\
W = 11, t = 21, \text{coding: } 100111010011001101 \\
W = 11, t = 21, \text{non-coding: } 11101001010010010111 \\
W = 12, t = 21, \text{coding: } 100111010101010101 \\
W = 12, t = 21, \text{non-coding: } 11101101010101010111 \\
\end{array}
\]

Ma, B., Tromp, J., Li, M., "PatternHunter: faster and more sensitive homology search", Bioinformatics 2002
Mar:18(3):440-5

Selected BLAST Databases for Nucleotides

- **nr (nt)**
  - Traditional GenBank Divisions
    - NM_ and XM_ RefSeqs
- **refseq_rna**
  - NM_ and XM_ RefSeqs
- **refseq_genomic**
  - NC_ NC_ RefSeqs
- **wgs**
  - whole genome shotgun
- **month**
  - updated within 30 days

- **est**
  - EST Division
- **htgs**
  - HTG division
- **gss**
  - GSS division
**BLAST Options - Nucleotide**

---

**Example Entrez Queries for Custom Databases**

- `all[Filter] NOT mammalia[Organism]`
- `green plants[Organism]`
- `biomol mrna[Properties]`
- `nucleotide gene[filter]`

**Other Advanced Options**

- `-e 10000` expect value
- `-v 2000` descriptions
- `-b 2000` alignments

---

**Protein BLAST**

- standard protein BLAST
- constructs PSIBlast; pattern search
- W = 2; E limit = 20000; PAM30
- CD-Search
- CDART: protein architectures

---

**BLAST Databases for Proteins**

- `nr` (non-redundant protein sequences)
  - GenBank CDS translations
  - NP, RefSeqs
  - PIR, Swiss-Prot, PRF
  - PDB (sequences from structures)
- refseq – NP, XP, RefSeqs
- swissprot
- pat – patents
- pdb – sequences with 3D structures
- env_nr – environmental sequences
- month – sequences updated within 30 days
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

Neighbors: Precomputed BLAST

Entrez Related Sequences produces a list of sequences sorted by BLAST score, but with no alignment details.
Nucleotide vs. Protein BLAST

Comparing ADSS from H. sapiens and A. thaliana

<table>
<thead>
<tr>
<th>Human:</th>
<th>A. th.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRVT</td>
<td>SQVR</td>
</tr>
<tr>
<td>VLG</td>
<td>VLG</td>
</tr>
<tr>
<td>QWGD</td>
<td>QWGD</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.

Nucleotide Translated BLAST Protein

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>S</td>
<td>PPP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>tbllnx</td>
<td>L</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

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

BLAST Hit on the Genome

BLASTn Hit on the Map Viewer
### Results of Discontiguous Megablast

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_000538</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>NP_000538</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>

### TBLASTN Results Using NP_000538

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_000538</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>NP_000538</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>

### For More Information...

- The Entrez Gene
- The Reference Sequence Database
- The UniProt Knowledgebase
- The Protein Data Bank
- The National Library of Medicine
- The National Center for Biotechnology Information
- The Gene Expression Omnibus
- The Gene Ontology
- The Kyoto Encyclopedia of Genes and Genomes
- The Database of Genomic Variants
- The Mouse Genome Database
- The Rat Genome Database
- The Human Protein Reference Database
- The Human Protein Atlas
# NCBI Courses (all free!)

<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full day introductory course on NCBI web resources</td>
<td>Two-day intensive course on NCBI web resources</td>
</tr>
<tr>
<td>Medical Library Association Course</td>
<td>Eleven half-day courses, focusing on either a resource or problem</td>
</tr>
<tr>
<td>PubChem Course</td>
<td>Three-day course on NCBI resources designed for medical librarians providing bioinformatics support</td>
</tr>
<tr>
<td>PowerTools Technical Workshop Centre</td>
<td>Intense multi-day courses taught at NLM, including NCBI PowerScripting: 4 days on the E-utils</td>
</tr>
<tr>
<td>Full day course covering NCBI PubChem</td>
<td>NCBI 4-Pak: Four mini-courses in 2 days</td>
</tr>
<tr>
<td>Full day course covering NCBI 3D structure resources</td>
<td></td>
</tr>
</tbody>
</table>

## 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**