The NCBI Genomic Discovery System

Using Integrated Data from Multiple Genomes at NCBI to Accelerate Biological Discovery

Eric W. Sayers, Ph.D.
National Center for Biotechnology Information
Bethesda, MD USA

HGM2004
5 April 2004
17.00 – 19.00

This Evening’s Menu…

Gene of the Day: thyroid peroxidase

• Entrez Gene and RefSeq
  – Exploring evidence for gene annotations
• Beyond RefSeq
  – What if my species isn’t in RefSeq?
• Sequence Polymorphisms
  – Finding SNPs and phenotypes
• NCBI Map Viewer
  – View and retrieve annotations and associated data
• Finding Gene Homologs
  – Find homologs using Entrez, Map Viewer, and BLAST
• Example Problems
  – Searching for markers and exploring disease alleles

The National Institutes of Health
The National Center for Biotechnology Information

- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

Entrez Gene and RefSeq

- Entrez Gene is a new database of gene loci
- Entrez Gene includes all organisms in LocusLink plus all other organisms in RefSeq
- Entrez Gene is an Entrez database with full searching and linking capabilities
- Entrez Gene will ultimately replace LocusLink

Gene Record for Human TPO
Primary vs. Derivative Databases

RefSeq Benefits

- Non-redundant
- Explicitly linked nucleotide and protein sequences
- Updates to reflect current sequence data and biology
- Data validation
- Format consistency
- Distinct accession series
- Stewardship by NCBI staff and collaborators

RefSeq Accession Numbers

<table>
<thead>
<tr>
<th>Curated Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_123455</td>
</tr>
<tr>
<td>NG_123456</td>
</tr>
<tr>
<td>NM_123456</td>
</tr>
<tr>
<td>NR_123456</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_123456</td>
</tr>
<tr>
<td>NW_123456</td>
</tr>
<tr>
<td>NZ_ABCD12345678</td>
</tr>
<tr>
<td>XM_123456</td>
</tr>
<tr>
<td>XP_123456</td>
</tr>
<tr>
<td>ZP_123456</td>
</tr>
<tr>
<td>XR_123456</td>
</tr>
</tbody>
</table>
RefSeq Pipelines

- **LocusLink Pipeline**
  - Associate gene locus with cDNAs from Genbank collaborators
  - Find longest mRNA
  - BLAST
  - NM
  - NP

- **Genome Annotation**
  - Assemble genomic region using sequences from Genbank collaborators
  - Predict Gene
  - NT / NW
  - NM
  - NP

- **Genome**
  - Predict Gene
  - XM
  - XP

- **Chromosomes:**
  - Complete assemblies of Genbank records

- **Microbial genomes:**
  - Compiled from Genbank sequences

- **Small genomes and Viruses:**
  - Compiled from one Genbank record

- **Curated Eukaryotic RefSeqs**

<table>
<thead>
<tr>
<th>RefSeq</th>
<th>NM</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mus musculus</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Anopheles gambiae str. PEST</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Magnaporthe grisea 70-15</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Zea mays (japonica cultivar-group)</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Eremothecium gossypii</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Plasmodium falciparum 3D7</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Trypanosoma brucei</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Xenopus tropicalis</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>NC</td>
<td>NC</td>
</tr>
</tbody>
</table>
Finding Primary Sequences

• Search Entrez Nucleotide
  – 99.3% GenBank (primary data)
  – 0.7% RefSeq (curated data)

• Typical starting query:
  – thyroid peroxidase AND human[orgn]

• Useful search terms in [Properties]:
  – srcdb : Source database (srcdb genbank)
  – gbdiv : GenBank division (gbdiv est)
  – biomol : Biomolecule type (biomol mrna)

Global Entrez Search

Results

291 records
NM_xxxxxx : RefSeqs!
Not TPO!
Limit by Title and Database

Entrez Nucleotide
99.3% GenBank (primary data)
srcdb ddbj/embl/genbank[properties]
0.7% RefSeq (curated data)
srcdb refseq[properties]

#1: thyroid peroxidase AND human[orgn] 291
#2: thyroid peroxidase[title] AND human[orgn] 166
#3: #2 AND srcdb refseq[properties] 5
#4: #2 AND srcdb ddbj/embl/genbank[properties] 161

Entrez Trick
primary data

GenBank Divisions

PRI (27) Primate
ROD (11) Rodent
PHL (10) Plant and Fungal
BCT (8) Bacterial/Archaeal
INV (6) Invertebrate
VRL (3) Viral
VRT (4) Other Vertebrate
MAM (1) Mammalian
PHG (1) Phage
SYN (1) Synthetic
UNA (1) Unannotated

EST (288) Expressed Sequence Tag
GSS (98) Genome Survey Sequence
HTG (61) High Throughput Genomic
HTC (3) High Throughput cDNA
STS (3) Sequence Tagged Site

Traditional
• Direct Submissions (Sequin/BankIt)
• Accurate (~1 error per 10,000 bp)
• Well characterized
• Organized by taxonomy

Bulk
• For sequencing projects
• Batch submissions (ftp/email)
• Inaccurate
• Poorly characterized
• Organized by sequence type

Limit by GenBank Division

EST Division gbdiv est[prop]
Primate Division gbdiv pri[prop]

#1: thyroid peroxidase AND human[orgn] 291
#2: thyroid peroxidase[title] AND human[orgn] 166
#3: #2 AND gbdiv refseq[prop] 5
#4: #2 AND gbdiv ddbj/embl/genbank[prop] 161
#5: #2 AND gbdiv est [prop] 20
#6: #2 AND gbdiv pri [prop] 141

traditional GenBank records
Limit by Biomolecule Type

Genomic DNA: bi omrl genomic[prop]
cDNA: bi omrl mrna[prop]

- #1: thyroid peroxidase AND human[orgn] 291
- #2: thyroid peroxidase[title] AND human[orgn] 166
- #3: #2 AND srcdb refseq[properties] 5
- #4: #2 AND srcdb ddbj/embl/genbank[properties] 161
- #5: #2 AND gbdiv est[prop] 20
- #6: #2 AND gbdiv pri[prop] 141
- #7: #6 AND bi omrl genomic[prop] 25
- #8: #6 AND bi omrl mrna[prop] 116

MmA: cDNA

Genomic Records

Segmented Set

cDNA Records

Human IgK

Human IgH
Use NOT

Limit by Protein Name

116 records (title) – 4 records (protein name)

Search by Sequence
Related Sequences

- #19 AND srcdb ddjb/embl/genbank[prop] AND human[orgn]
  - Most similar
  - Least similar

Search  Most Recent Quarters  Time  Result
---  ------------------------  ----  ----
20  #20 AND srcdb ddjb/embl/genbank[prop] AND human[orgn]  10:01:09  8 sequences not found by a text search

Similar Human GenBank Sequences

8 sequences not found by a text search

Sorting order returns to "last in, first out."

2 thyroperoxidases

How did we do?

Entrez Gene Related Sequences

<table>
<thead>
<tr>
<th>GeneBank</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>A04439</td>
<td>A4ET71</td>
</tr>
<tr>
<td>A04439</td>
<td>A4ET71</td>
</tr>
</tbody>
</table>

Entrez Search

SWISS-PROT, no nucleotide links!
Creating NM Records

Genome annotation

NM/NP Records in Entrez

NM_000547: variant 1

COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff.
The reference sequence was derived from M17755.2 and AW874082.1.
On Feb 25, 2003 this sequence version replaced gi:21361188.

NM_175719: variant 2

COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff.
The reference sequence was derived from J02970.1, AW874082.1 and M17755.2.

EST that completes 3' end

NM/NP Records in Entrez Gene
Entrez Gene RefSeq Graphics

Annotation Pipeline

Model Eukaryotic RefSeqs
Creating Transcript Models (XMs)

Genomic DNA (NC, NT, NW) → Model mRNA (XM) → Model protein (XP)

Curated mRNA (NM) → Curated Protein (NP)

RetSeq

Genbank Sequences

So how about TPO?

**Thyroid peroxidase**

GeneID: 7173  Location: RefSeq: NC_001149.12

Transcripts and products: RefSeq below


Spidey Results

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>mRNA coordinates</th>
<th>Length</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-945</td>
<td>945</td>
<td>100.0%</td>
</tr>
<tr>
<td>2</td>
<td>866-1560</td>
<td>694</td>
<td>100.0%</td>
</tr>
<tr>
<td>3</td>
<td>753-1667</td>
<td>914</td>
<td>100.0%</td>
</tr>
<tr>
<td>4</td>
<td>8448-8685</td>
<td>237</td>
<td>100.0%</td>
</tr>
<tr>
<td>5</td>
<td>2796-2852</td>
<td>56</td>
<td>100.0%</td>
</tr>
<tr>
<td>X</td>
<td>451-565</td>
<td>115</td>
<td>100.0%</td>
</tr>
<tr>
<td>Y</td>
<td>12779-13065</td>
<td>2866</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
What about LOC400938?

BLAST Results for XM_171013

Is there any GenBank support for this mRNA?

The Perils of the XM
Genome Annotation in Entrez

GenBank Components (clones, WGS) → Assembly → NT/NW Contigs → Genome Components → NM/XM Master mRNA

Genome Components: NM/XM (curated mRNA)

Genome Annotation Links:
- Download all clones used in the assembly of human chromosome 2
- The 23 contigs of the chromosome 2 assembly
- Human chromosome 2

Entrez Trick:
- The 1983 clones of the chromosome 2 assembly
- The genomic contig on human chromosome 2 containing NM_000547
- Human chromosome 2
Getting an Annotation Summary

Getting the Annotation Details

These flat files contain all annotations in the gene and the full, explicit sequence.
Getting the Assembly Details

1. **LocusLink**: Associate gene locus with cDNAs from Genbank or collaborators.
2. **BLAST**: Find longest mRNA.
3. **NM**: Assemble genomic region using sequences from Genbank or collaborators.

Genome Pipeline

- **LocusLink**: Associate gene locus with cDNAs from Genbank or collaborators.
- **Genome Annotation**: Assemble genomic region using sequences from Genbank or collaborators.
- **Genome**: Chromosomes: complete assemblies of Genbank records.
- **NC**: Microbial genomes: compiled from Genbank sequences.
- **XM**: Small genomes and Viruses: compiled from one Genbank record.
- **XP**: Predict Gene.

Entrez Gene and Genome

Entrez Gene now includes all organisms in Entrez Genome.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Organism</th>
<th>Chromosome</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>adenylosuccinate synthetase</td>
<td>Salmonella enterica serovar Typhimurium</td>
<td>12.0</td>
<td>SM2053</td>
</tr>
<tr>
<td>adenylosuccinate synthetase</td>
<td>Salmonella enterica serovar Typhimurium</td>
<td>12.0</td>
<td>SM2053</td>
</tr>
<tr>
<td>adenylosuccinate synthetase</td>
<td>Shigella flexneri 2a</td>
<td>12.0</td>
<td>101962</td>
</tr>
<tr>
<td>adenylosuccinate synthetase</td>
<td>Shigella flexneri 2a</td>
<td>12.0</td>
<td>101962</td>
</tr>
<tr>
<td>adenylosuccinate synthetase</td>
<td>Shigella flexneri 2a</td>
<td>12.0</td>
<td>101962</td>
</tr>
</tbody>
</table>
Which Species?

works only for sets of less than 500 records

Finding Genes by Text

Query: thyroid peroxidase

<table>
<thead>
<tr>
<th>Gene symbol:</th>
<th>human thyroid peroxidase (TPO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tpo [syn] AND human [organism]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein name:</th>
<th>topoisomerase genes from Archaea</th>
</tr>
</thead>
<tbody>
<tr>
<td>topoisomerase[gene/protein name] AND archaea [organism]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chromosome and Links:</th>
<th>genes on human chromosome 2 with OMIM links</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>RefSeq status and variants:</th>
<th>Reviewed RefSeqs with transcript variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>srcdb refseq reviewed[prop] AND has transcript variants[prop]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disease and Gene Ontology:</th>
<th>Membrane proteins linked to cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>integral to plasma membrane[gene ontology] AND cancer [dis]</td>
<td></td>
</tr>
</tbody>
</table>
Scripting Gene Searches

Conduct Entrez searching by posting URLs

Find human genes with disease alleles that have been mapped uniquely

**ESearch**(gene)  human [organism] AND gene omim[filter]
- #1: Gene UIDs

**ELink**(gene, OMIM)
- #2: OMIM UIDs
- #3: OMIM UIDs
- #4: Gene UIDs

**ESearch**(OMIM)
- #5: OMIM UIDs

**ESummary**
- #6: Gene UIDs
- Summaries

Gene, LocusLink, and the Future

- Gene will eventually replace LocusLink
- Many LocusLink functions are now in the Entrez Links menu
- As much as possible, LocusLink IDs are equal to the corresponding Gene UID
- Currently there is no FTP site for Entrez Gene
- Therefore, we will maintain the LL_tmpl file and associated files until Entrez Gene has FTP files
Beyond RefSeq

If your organism does not have RefSeqs...

- UniGene: gene-based clusters of cDNAs and ESTs
- UniSTS: STS markers
- WGS sequences in Entrez Nucleotide (wgs[prop])
- Trace Archive

UniGene: Clustering Expressed Sequences

- Records are clusters of mRNAs and ESTs that ideally represent single genes
- Records are created automatically by a modified BLAST algorithm
- UniGene provides a means to identify an EST or unannotated mRNA

Organisms in UniGene

<table>
<thead>
<tr>
<th>Species</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>94,884</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>3,155</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>26,842</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>10,543</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>4,950</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>24,164</td>
</tr>
<tr>
<td>Homo troglodytes troglodytes</td>
<td>1,265</td>
</tr>
<tr>
<td>Canis lupus</td>
<td>3,823</td>
</tr>
<tr>
<td>Pongo abelii, pygmaeus</td>
<td>3,959</td>
</tr>
<tr>
<td>Chimpanzee paniscus</td>
<td>1,966</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>9,088</td>
</tr>
<tr>
<td>Apithecus blacki</td>
<td>10,986</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>10,176</td>
</tr>
<tr>
<td>Cavia porcellus</td>
<td>1,402</td>
</tr>
<tr>
<td>Microtus arvalis</td>
<td>2,085</td>
</tr>
<tr>
<td>Mus musculus carolinensis</td>
<td>1,855</td>
</tr>
<tr>
<td>Mus musculus molossinus</td>
<td>1,441</td>
</tr>
<tr>
<td>Mus musculus philgenii</td>
<td>1,181</td>
</tr>
<tr>
<td>Mus musculus spretus</td>
<td>1,384</td>
</tr>
<tr>
<td>Pan troglodytes troglodytes</td>
<td>5,167</td>
</tr>
<tr>
<td>Pan troglodytes schweinfurthi</td>
<td>2,389</td>
</tr>
<tr>
<td>Pan troglodytes troglodytes</td>
<td>1,542</td>
</tr>
<tr>
<td>Cricetus auratus</td>
<td>1,055</td>
</tr>
<tr>
<td>Peromyscus leucopus</td>
<td>1,210</td>
</tr>
<tr>
<td>Peromyscus maniculatus</td>
<td>1,055</td>
</tr>
<tr>
<td>Peromyscus leucopus</td>
<td>1,210</td>
</tr>
<tr>
<td>Peromyscus leucopus</td>
<td>1,055</td>
</tr>
<tr>
<td>Peromyscus leucopus</td>
<td>1,210</td>
</tr>
</tbody>
</table>
Finding UniGene Clusters

Searching Entrez UniGene

Nucleotide: Non-genomic sequences from the PLN division of Genbank

gbdiv_pln [properties] NOT biomol_genomic [properties]

mRNA: rat clusters that have at least one mRNA

rat [organism] NOT 0 [mRNA count]

Chromosome: e

Genethon [Map Name] AND human [organism] AND 12 [chromosome]

UnistS: Markers on the Genethon map of human chromosome 12

Structure: Structures of bacterial kinases with resolutions below 2 Å

Bacteria [organism] AND kinase AND 000.00:002.00 [resolution]
UniGene Links in Entrez

UniSTS

UniSTS Record

UniSTS: A unified set of markers
Linking Beyond Genes

- UniGene
- Nucleotide
- SNP
- UniSTS
- GEO
- Protein
- OMIM
- PubMed
- Homologene
- Taxonomy
- Books
- Genome
- 3D Domains
- Structure
- Domains
- PMC
- Journals
- PubMed
- GEO Datasets

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
Find Non-synonymous SNPs

Non-synonymous TPO SNPs

Locus View of SNPs
Links to OMIM

OMIM Record

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

<table>
<thead>
<tr>
<th>View Summary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1 diagram</td>
<td>Task 2 diagram</td>
</tr>
<tr>
<td>Region</td>
<td>Identifiers</td>
</tr>
<tr>
<td>Total Unique Contigs: 6</td>
<td>Total Unique Contigs: 6</td>
</tr>
<tr>
<td>Total Unique Clones: 6</td>
<td>Total Unique Clones: 6</td>
</tr>
<tr>
<td>Total Unique Features: 6</td>
<td>Total Unique Features: 6</td>
</tr>
</tbody>
</table>

### Map Content

- Sequence Maps
  - Core assembly
  - Annotation evidence
  - Clones & Markers
  - Polymorphisms
  - Links & Features

- Genetic Maps
  - Cytogenetic maps
  - Linkage maps
  - Radiation hybrid maps

---

View the Assembly near TPO
Map Content

Map content varies greatly by species!

• Sequence Maps
  • Core assembly
  • Annotation evidence
  • Clones & Markers
  • Polymorphisms
  • Links & Features

• Genetic Maps
  • Cytogenetic maps
  • Linkage maps
  • Radiation hybrid maps

Ab initio (model)
GenBank DNA
EST
UniGene
Gene

Annotation Evidence

GenBank records not used in assembly

Unigene Clusters

Ab initio models

Searching the Map Viewer
Viewing Results on the Maps

Linking to the Map Viewer

Linking from Gene
Gene Homologs in Entrez

ProtEST blast of UniGene cluster sequences against curated proteins from 8 organisms:
- H. sapiens
- M. musculus
- R. norvegicus
- D. melanogaster
- C. elegans
- S. cerevisiae
- A. thaliana
- E. coli

Gene Homologs in Entrez

Entrez Homologene: Homologs by protein BLAST

Pairwise Scores

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene</th>
<th>e-value</th>
<th>Bit Score</th>
<th>e-value</th>
<th>Bit Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Rhesus</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Mouse</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Elephant</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Giraffe</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Elephant</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Giraffe</td>
<td>TPO</td>
<td>1.2</td>
<td>74.9</td>
<td>0.39</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Gene Homologs in Map Viewer

Rat – Mouse - Human

A Closer Look…
Genomic BLAST

Results Summary

Results on Genome Overview
BLAST Hits on Map Viewer

Comparing the Dog cDNA

AAM26727

dog protein

Mouse

Human

<table>
<thead>
<tr>
<th>BLAST</th>
<th>Score</th>
<th>E Value</th>
<th>Hotels</th>
<th>Percent Identity</th>
<th>Query Range</th>
<th>Target Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>100%</td>
<td>0.0001</td>
<td>100</td>
<td>100</td>
<td>200-1000</td>
<td>200-1000</td>
</tr>
<tr>
<td>Mouse</td>
<td>99%</td>
<td>0.0002</td>
<td>95</td>
<td>99</td>
<td>300-1500</td>
<td>300-1500</td>
</tr>
</tbody>
</table>

Human

<table>
<thead>
<tr>
<th>BLAST</th>
<th>Score</th>
<th>E Value</th>
<th>Hotels</th>
<th>Percent Identity</th>
<th>Query Range</th>
<th>Target Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>100%</td>
<td>0.0001</td>
<td>100</td>
<td>100</td>
<td>200-1000</td>
<td>200-1000</td>
</tr>
<tr>
<td>Mouse</td>
<td>99%</td>
<td>0.0002</td>
<td>95</td>
<td>99</td>
<td>300-1500</td>
<td>300-1500</td>
</tr>
</tbody>
</table>

Mouse

<table>
<thead>
<tr>
<th>BLAST</th>
<th>Score</th>
<th>E Value</th>
<th>Hotels</th>
<th>Percent Identity</th>
<th>Query Range</th>
<th>Target Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>100%</td>
<td>0.0001</td>
<td>100</td>
<td>100</td>
<td>200-1000</td>
<td>200-1000</td>
</tr>
<tr>
<td>Mouse</td>
<td>99%</td>
<td>0.0002</td>
<td>95</td>
<td>99</td>
<td>300-1500</td>
<td>300-1500</td>
</tr>
</tbody>
</table>
Find a Locus Between Markers

Markers in Map Viewer

Targeting a Region
Exon 15 is the tenth exon of NM_000547

Create a Model with Exons 1-9

Desired model = 2-3-5-6-7-9-10-11-13

欲建模型 = 2-3-5-6-7-9-10-11-13-15-16-17-19-20-22-23-25

Compare Conserved Domains

NM_000547 → NP_000538
Explore a Disease SNP

Problem

Curated CD Record

Curated CDs (cd12345) are based on sequence and structure alignments.

Annotated features

Structural evidence

Curated CD Record

Feature 1: CAV1 binding site
For More Information…

• 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

Come See Us!

Booth #12

Poster #233