

NCBI Public Services

NCBI Overview

June 21, 2011



NCBI Public Services

Topics for Today

- About NCBI
- Overview of Molecular Databases
- Apolipoprotein E example
 - Using the Entrez system

NCBI Public Services

**The National Center for
Biotechnology Information**



***Created in 1988 as a part of the
National Library of Medicine at NIH***

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Aspects of Molecular Data

- Sequences
- Expression
- Genome Maps
- 3D Structures
- Protein Domains
- Homologous Genes, Proteins, Structures
- Pathways
- Genetic Variation

NCBI Public Services

Selected NCBI Databases

- Biomedical literature
 - PubMed [free Medline](#) 3-4 million searches per day
 - PubMed Central [full text online access](#)
 - NCBI Bookshelf [online biomedical textbooks](#)
- Biomolecular Databases
 - GenBank [primary nucleotide sequence database](#)
 - Protein [protein translations and outside sources](#)
 - RefSeq [curated NCBI reference sequences](#)
 - dbSNP [small scale genetic variations](#)
 - MMDB [NCBI's 3D structure database](#)
 - GEO [microarray expression data](#)

NCBI Public Services

Metadatabases: Information Hubs

- HomoloGene [homologs \(genes, sequences\) from selected eukaryotes](#)
- UniGene [sequence based gene catalog](#)
- Taxonomy [master browser for molecular data](#)
- Gene [molecular and literature related to genes](#)

NCBI Public Services

Selected NCBI Search Services and Tools

NCBI Public Services

- Entrez **integrated literature and molecular databases**
 - BLink **protein similarities**
 - Graphical Sequence Viewer **incipient genome browser**
- BLAST **highest volume sequence search service**
- VAST **structure similarity searches**
- Map Viewer **graphical genome map display**

Global Entrez Query: All NCBI Databases

NCBI Public Services

The screenshot shows the Global Entrez Query interface with a search bar at the top. Below the search bar, there is a grid of database search results. A central box highlights the text: "The Entrez system: 38 (and counting) integrated databases". The grid lists various databases such as PubMed, GenBank, and others, along with their respective record counts.

Sequence Databases at NCBI

NCBI Public Services

- **Primary**
 - GenBank: NCBI's primary sequence database
 - Entrez
 - Nucleotide
 - EST
 - GSS
 - Trace Archive: reads from capillary sequencers
 - Sequence Read Archive: next generation data
- **Derivative**
 - GenPept (GenBank translations)
 - Outside Protein (UniProt—Swiss-Prot, PDB)
 - NCBI Reference Sequences (RefSeq)

The Gene Database: Central Database

Entrez Gene
Genes and mapped phenotypes

Search: Gene [v] Limits Advanced search Help

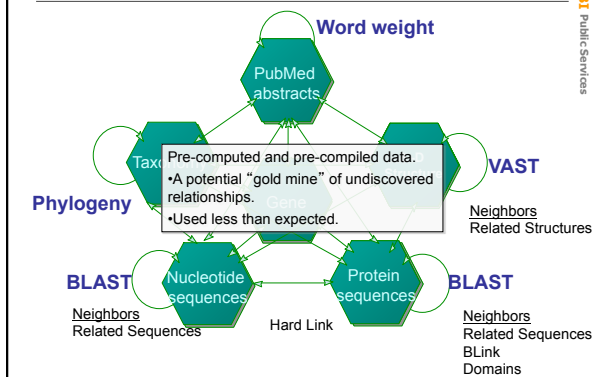
Welcome to Entrez Gene
Entrez Gene maintains information about genes from genomes of interest to the RefSeq group.

- Gene Centered Information
- Unifies NCBI-annotated and Submitted Genomes
- 6.7 million records for 6,740 taxa
- Central hub for accessing literature and molecular data

Entrez Gene News Spigen Protein Clusters

NCBI Public Services

Entrez: A Discovery System



NCBI Public Services

The Discovery Initiative

- Easier to use interfaces
- Promote higher quality resources
 - Gene
 - RefSeqs
- Expose the power of pre-computed similarities and pre-compiled links

NCBI Public Services

Discovery Components in Entrez

NCBI Public Services

- **Database Ads** – direct to related information in other database
- **Sensors** – point to other databases or special search tools where the query is more relevant
- **Analysis tools** – access to live analysis results

Apolipoprotein E (APOE)

NCBI Public Services

- Important serum lipid transport protein.
- Defects implicated in cardiovascular disease and late-onset Alzheimer disease (LOAD).
- Three common isoforms (alleles).

Isoform	Position 112 (130)	Position 158 (176)
e3	Cysteine (C)	Arginine (R)
e4	Arginine (R)	Arginine (R)
e2	Cysteine (C)	Cysteine (C)

The e4 isoform (allele) is associated with increased risk of LOAD

Human APOE Data Goals

NCBI Public Services

- Access to Gene Record
- Reference Sequences
 - transcript (mRNA)
 - protein
 - RefSeq Gene (gDNA)
- Sites of Expression
 - UniGene Expression profile
 - GEO profiles
- View genomic assemblies and Maps
 - Comparative Maps
- Disease gene polymorphisms
- Genotypes
 - Reference Genome
 - HuRef (JC Venter genome)
- Homologs in other species
 - HomoloGene
 - Blink to find Sea Lion homolog

NCBI Public Services

Live demonstrations

www.ncbi.nlm.nih.gov

NCBI Public Services

Human APOE Gene and RefSeqs

<ol style="list-style-type: none"> 1. Search All Databases with APOE from NCBI Homepage 2. Retrieve PubMed results 3. Follow Gene Ad to Gene ID 348 4. Follow link to “reference sequence details” 	<ul style="list-style-type: none"> • Transcript and protein <ul style="list-style-type: none"> – NM_000041.2 – NP_000032.1 • Genomic <ul style="list-style-type: none"> – RefSeq Gene NG_007084.2 • Genome Builds <ul style="list-style-type: none"> – Reference NC_000019.9 – HuRef
--	---

NCBI Public Services

Genome Maps

<p>APO Gene Cluster</p> <ol style="list-style-type: none"> 1. Gene links menu to Map Viewer 2. Maps and Options <ol style="list-style-type: none"> 1. Remove all but Gene 3. Zoom out 10X using graphic 4. Zoom our 2X by clicking on map 	<p>Comparative Maps</p> <ol style="list-style-type: none"> 1. Map Viewer Maps and Options <ol style="list-style-type: none"> 1. Add Chimp and Mouse Gene Maps
--	---

Allelic Variants and Disease

OMIM	dbSNP
<ul style="list-style-type: none"> Gene links menu to OMIM Retrieve APOLIPOPROTEIN E; APOE Click allelic variants Jump to .0016 ALZHEIMER DISEASE 2 [APOE, CYS112ARG] Recent activity back to APOE gene record 	<ul style="list-style-type: none"> Gene links menu to SNP: Gene View Check "Include clinically associated" and click "Refresh" Position 130 Cys->Arg Click on rs429358 in table Note HuRef has C allele coding for Arg

NCBI
Public Services

Expression Information

UniGene	GEO Profiles
<ol style="list-style-type: none"> 1. Gene links menu to UniGene 2. EST profile 3. Widespread expression 	<ol style="list-style-type: none"> 1. Gene links menu to GEO profiles 2. Use History to combine link from Gene AND GDS596 3. Predominant expression in liver, some in brain

NCBI
Public Services

Homologous Genes / Proteins

HomoloGene	Using BLAST
<ol style="list-style-type: none"> 1. Gene links menu to HomoloGene 2. Use "pairwise alignments" device to compare Human and Chimp proteins at position 130 3. Use UniGene portion to find rabbit (<i>Oryctolagus cuniculus</i>) homolog 	<ol style="list-style-type: none"> 1. Gene link to RefSeq protein NP_000032 2. "Run BLAST" from Discovery Column 3. Choose Reference Proteins database 4. Organism limit Giant Panda 5. BLAST button

NCBI
Public Services

Finding a Structure

Related Structures

1. Gene link to NP_000032
2. Links menu to Related Structures
3. Change to "All similar MMDB" and click "Go"
4. Link to sequence alignment from 1NFN_A and 1B68_A
5. Position 112 polymorphism

Structure and Cn3D

1. Related structures link to structure for 1B68
2. Click Structure View in Cn3D (Cn3D must be installed)
3. Manipulate structure
4. Add side chains using Style -> Edit Global Style
5. Highlight Arg 112

NCBI Public Services

Service Addresses

•**General Help** `info@ncbi.nlm.nih.gov`
 •**BLAST** `blast-help@ncbi.nlm.nih.gov`

Telephone support: 301- 496- 2475

NCBI Public Services
