

NCBI-Resources for Genes & Genomes

This workshop will introduce you to the main information hubs regarding genomic biology. The focus of the workshop are the NCBI-Databases Gene, RefSeq, Genomes, Genome-Projects, and Taxonomy. Database searches and database contents will be compared.

Starting points for your search depend on the information you already have. If you have an accession number your search is more directed than a search with keywords. In order to interpret the data you need to know how the resource was created and what information is being provided. Databases can roughly be categorized into archival databases that provide raw data and duplicate records (as for example GenBank/EMBL/DDBJ) and curated databases that provide non-redundant, processed and often annotated data (as for example Swiss-Prot and RefSeq). This class discusses the most important databases for information of molecular data.

Types of Databases - archival vs. curated

- Preliminary: - web sites of sequencing centers
- Archival: - raw data, duplicate records, e.g., GenBank/EMBL/DDBJ
- Curated: - non-redundant, processed data, often annotated, e.g., Swiss-Prot, RefSeq
- Peer Reviewed: - UniGene, COGs

1. What kind of information do you start from?

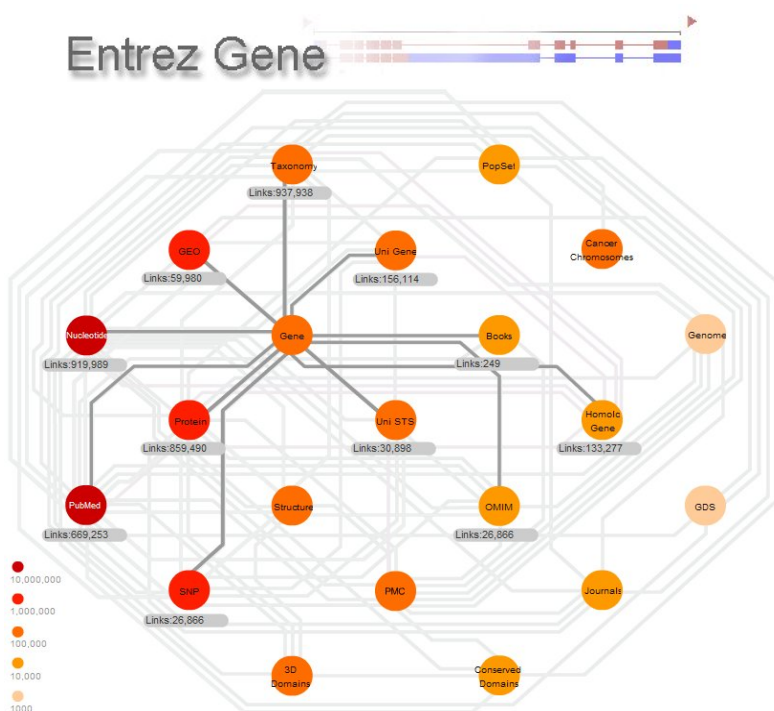
- **Accession number** – which database does the accession number belong to? Can I go from here?
- **Gene symbol** – are you dealing with the official gene symbol or an alias?
- **Gene description** – can you find descriptions in the categories of **gene ontology** or did you make up your own descriptions? Which words do I need in order to find what I am looking for?
- **Organism** – are you interested in taxonomy, the genome, a web-portal of the organism etc.?
- **Raw sequence (genomic/cDNA)** – in order to find more information on the sequence you will have to perform sequence similarity searches as for example with BLAST.

2. What information is actually provided in the resources discussed below?

(1) **Entrez-Gene** (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>)

Scope: *Entrez-Gene* is the NCBI database for gene-specific information. It does not include all known or predicted genes; instead

- *Entrez-Gene* focuses on the genomes that have been completely sequenced,
- that have an active research community to contribute gene-specific information, or
- that are scheduled for intense sequence analysis.
- The content of *Entrez-Gene* represents the result of curation and automated integration of data from NCBI's Reference Sequence project (RefSeq), from collaborating model organism databases, and from many other databases available from NCBI.



(2) **RefSeq** (<http://www.ncbi.nlm.nih.gov/RefSeq/>)

Scope: The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

- Explanations/Accessions (<http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions>)
- Examples *M28668 OR NM_000492*

(2a) **RefSeqGene** - RefSeqGene, a subset of NCBI's Reference Sequence (RefSeq) project, defines genomic sequences to be used as reference standards for well-characterized genes. These sequences, labeled with the keyword RefSeqGene, serve as a stable foundation for reporting mutations, for establishing conventions for numbering exons and introns, and for defining the coordinates of other variations.

- The RefSeqGene project is an active member of the **Locus Reference Genomic (LRG)** project. LRG sequences provide a stable genomic DNA framework for reporting mutations with a permanent ID and core content that never changes.

(3) **Taxonomy Database/Browser**
(<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>)

Scope: “The NCBI Taxonomy Database contains the names and lineages of ~247,000 organisms, both living and extinct, that are represented in the genetic databases with at least one nucleotide or protein sequence“.

Search examples: “dog”, HIV

Genome HUBS – the resources Entrez-Genome and BioProjects

The new Genome database shares a close relationship with the recently redesigned BioProject database (formerly Genome Project). Primary information about genome sequencing projects in the new Genome database is stored in the BioProject database. BioProject records of type "Organism Overview" have become Genome records with a Genome ID that maps uniquely to a BioProject ID. The new Genome database also includes all "genome sequencing" records in BioProject.

(3) **Entrez-Genome** (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>)

Scope: This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

The new Genome resource uses a new data model where a single record provides information about the organism (usually a species), its genome structure, available assemblies and annotations, and related genome-scale projects such as transcriptome sequencing, epigenetic studies and variation analysis. As before, the Genome resource represents genomes from all major taxonomic groups: Archaea, Bacteria, Eukaryote, and Viruses. The old Genome database represented only Refseq genomes, while the new resource extends this scope to all genomes either provided by primary submitters (INSDC genomes) or curated by NCBI staff (RefSeq genomes).

(4) **Entrez-BioProject** (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=bioproject>)

Scope: “A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.”

(5) **Genetics & Medicine**

- OMM
- OMIA
- Genes & Disease
-and more

Related resources

(a) **MapViewer** (<http://www.ncbi.nlm.nih.gov/mapview/>)

Scope: The MapViewer is a software component of *Entrez-Genome* that provides special browsing capabilities for a subset of organisms. It allows you to view and search an organism's complete genome, display chromosome maps, and zoom into progressively greater levels of detail, down to the sequence data for a region of interest.

Example: Browse Chromosomes and Maps, Look at the MapViewer Help pages.

Other Genome Browser:

- **ENSEMBL Genome Browser** (<http://www.ensembl.org/index.html>)
- **UCSC Genome Bioinformatics** (<http://genome.ucsc.edu/>)

- (b) **HUGO Gene Nomenclature Committee** (<http://www.genenames.org/>)

Scope: “Giving unique and meaningful names to every **human** gene”.

- (c) **Gene Ontology Home** (<http://www.geneontology.org/>)

Scope: The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism.

- Search the Gene Ontology Database for “oxygen transport activity”.

- (d) Life on Earth from the “**Tree of Life**” Web Project (<http://tolweb.org/>, *Root of the tree*)

- (e) **UniProt** (<http://www.uniprot.org/>) – The Universal Protein Resource (UniProtKB – UniProt Knowledge Base)

- (f) **Nucleic Acids Research**

- Molecular Biology Database Collection
<http://nar.oxfordjournals.org/content/40/D1.toc>
- Web Server issue
http://nar.oxfordjournals.org/content/39/suppl_2

- (g) **NCBI Training Tutorials**

<http://www.ncbi.nlm.nih.gov/guide/training-tutorials/>

MPG Bioinformatics Support Service

<http://www.biochem.mpg.de/en/facilities/ivs/SupportTraining/BioInfoSup/index.html>

Wiki on bioinformatics tools developed in the Max Planck Society

<http://www.bioinfowiki.mpg.de/>