

Gene-centered Information Resources

You will be introduced to the main information resources for genes and sequences. The tour includes *Entrez-Gene*, Gene-Cards, Gene Reports, Genes in Pathways and others.

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 on genes and their peculiarities that help you to interpret the 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?

- A **Accession number** – which database does the accession number belong to? Can I go from here?
- B **Gene symbol** – are you dealing with the official gene symbol or an alias?
- C **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?
- D **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 *AF117710* OR *NM_000518*

(3) **Ensembl Gene Reports** (<http://www.ensembl.org/>)

Scope: The project provides annotations of genes and genomes with a focus on vertebrate genomes.

- Search Ensembl (e.g. for hemoglobin – look up genes for *Homo sapiens*; look up the HGNC-category for official Gene-Symbol “HBB”)

(4) **GeneCards** (<http://www.genecards.org/>)

Scope: GeneCards® is a database of human genes, their products and their involvement in diseases. It offers concise information about the functions of all human genes that have HGNC-approved symbols, as well as other genes. Search GeneCards (e.g. for HBB or hemoglobin beta)

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

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

(6) **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”.

Gene information of genes involved in biochemical pathways

(7) **KEGG** (<http://www.genome.jp/kegg/>)- Kyoto Encyclopedia of Genes and Genomes

(8) **BioCarta** (<http://www.biocarta.com/>) – Dynamic graphical models of gene interactions
Browse/Search Pathways (e.g. by gene name; <http://www.biocarta.com/genes/index.asp>)

Related resources

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

(10) **HPRD Human Protein Reference Database** (<http://dev.hprd.org/>)

Scope: The Human Protein Reference Database represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome.

Access: Entrez-Gene/ Links

(11) **Nucleic Acids Research**

- Molecular Biology Database Collection (http://nar.oxfordjournals.org/content/vol36/suppl_1/index.dtl)
- Web Server issue (http://nar.oxfordjournals.org/content/vol35/suppl_2/)

MPG Bioinformatics Support Service

(<http://wwwex.biochem.mpg.de/iv/bioinfo.html>)

- Internal Resources, Tools & Groups (<http://wwwex.biochem.mpg.de/iv/mpsinternal.html>)
- Selected External Resources (<http://wwwex.biochem.mpg.de/iv/external.html>)

©N. Gaedeke

Last update: April 14th, 2008