Introduction to Sequence Databases

1. DNA & RNA
2. Proteins
What are Databases?

- A database is a structured collection of information.
- A database consists of basic units called records or entries.
- Each record consists of fields, which hold pre-defined data related to the record.
- For example, a protein database would have protein sequences as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence, …)
A database can be thought of as a large table, where the rows represent records and the columns represent fields.

<table>
<thead>
<tr>
<th>Field Record</th>
<th>Name</th>
<th>Length</th>
<th>Sequence</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA001</td>
<td>MTGA</td>
<td>243</td>
<td>MYQWI…</td>
<td>yes</td>
</tr>
<tr>
<td>QA002</td>
<td>Ribosomal protein L9</td>
<td>267</td>
<td>MAAPV…</td>
<td>no</td>
</tr>
<tr>
<td>QA003</td>
<td>Flagellin</td>
<td>374</td>
<td>GSSIL…</td>
<td>no</td>
</tr>
<tr>
<td>QA004</td>
<td>GDPMH</td>
<td>157</td>
<td>MFLRQ…</td>
<td>yes</td>
</tr>
</tbody>
</table>

**Accession Numbers**: Unique identifiers of the database records.
Ideal minimal content of an entry in a sequence database

- Sequence
- Accession number (AC)
- Taxonomic data
- References
- Annotation/Curation
- Keywords
- Cross-references
- Documentation

Sources of data:
- research groups (direct submission)
- literature supplementary information
- genome sequencing institutes
- patents
Within a database, the format needs to be kept consistent.

A SwissProt entry, in **Fasta** format:

```
>sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens (Human).
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRPLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAWSGLRSLTLRLRALGAQKEAISPDD
AASAAAPLRRTITADTFRKLFRVYSNFLRGKLLKLYTGEACRTGDR
```
Why Databases?

- The purpose of databases is not merely to collect and organize data, but to allow intelligent data retrieval.
- A **query** is a method to retrieve information from the database.
- The organization of each record into predetermined fields, allows us to use **queries on fields**.
Databases on the Internet

- Biological databases often have web interfaces, which allow users to send queries to the databases.
- Some databases can be accessed by different web servers, each offering a different interface.
Database download

• Nearly all biological databases are available for download as simple text (flat) files.
• A local version of the database allows one greater freedom in processing the data.
• Processing data in files requires some computer-programming skills. PERL is an easy programming language that can be used for extraction and analysis of data from files.
There are approximately 286,730,369,256 sequence records in the traditional GenBank divisions as of 2011.
### Table 1. Growth of GenBank Divisions (nucleotide base-pairs)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TSA</td>
<td>Transcriptome shotgun data</td>
<td>39 829 979</td>
<td>398 676 845</td>
<td>900.9</td>
</tr>
<tr>
<td>ENV</td>
<td>Environmental samples</td>
<td>1 091 072 890</td>
<td>1 723 286 428</td>
<td>57.9</td>
</tr>
<tr>
<td>PAT</td>
<td>Patented sequences</td>
<td>5 592 927 651</td>
<td>8 519 294 473</td>
<td>52.3</td>
</tr>
<tr>
<td>BCT</td>
<td>Bacteria</td>
<td>4 107 328 206</td>
<td>5 333 010 385</td>
<td>29.8</td>
</tr>
<tr>
<td>VRL</td>
<td>Viruses</td>
<td>779 481 462</td>
<td>970 125 245</td>
<td>24.5</td>
</tr>
<tr>
<td>PHG</td>
<td>Phages</td>
<td>36 100 172</td>
<td>43 456 808</td>
<td>20.4</td>
</tr>
<tr>
<td>MAM</td>
<td>Other mammals</td>
<td>576 977 646</td>
<td>679 274 390</td>
<td>17.7</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrates</td>
<td>1 734 996 371</td>
<td>2 036 240 836</td>
<td>17.4</td>
</tr>
<tr>
<td>WGS</td>
<td>WGS data</td>
<td>148 165 117 763</td>
<td>169 253 846 128</td>
<td>14.2</td>
</tr>
<tr>
<td>GSS</td>
<td>Genome survey sequences</td>
<td>16 738 219 857</td>
<td>18 442 479 673</td>
<td>10.2</td>
</tr>
<tr>
<td>PLN</td>
<td>Plants</td>
<td>3 695 552 256</td>
<td>4 038 424 961</td>
<td>9.3</td>
</tr>
<tr>
<td>SYN</td>
<td>Synthetic</td>
<td>131 361 806</td>
<td>142 548 355</td>
<td>8.5</td>
</tr>
<tr>
<td>VRT</td>
<td>Other vertebrates</td>
<td>2 366 300 257</td>
<td>2 533 789 261</td>
<td>7.1</td>
</tr>
<tr>
<td>EST</td>
<td>ESTs</td>
<td>34 522 977 161</td>
<td>36 803 930 321</td>
<td>6.6</td>
</tr>
<tr>
<td>HTC</td>
<td>High-throughput cDNA</td>
<td>636 472 189</td>
<td>659 355 057</td>
<td>3.6</td>
</tr>
<tr>
<td>PRI</td>
<td>Primates</td>
<td>5 751 413 009</td>
<td>5 943 029 356</td>
<td>3.3</td>
</tr>
<tr>
<td>ROD</td>
<td>Rodents</td>
<td>4 206 718 960</td>
<td>4 298 354 944</td>
<td>2.2</td>
</tr>
<tr>
<td>HTG</td>
<td>High-throughput genomic</td>
<td>23 895 733 886</td>
<td>24 276 862 305</td>
<td>1.6</td>
</tr>
<tr>
<td>UNA</td>
<td>Unannotated</td>
<td>119 348</td>
<td>120 289</td>
<td>0.8</td>
</tr>
<tr>
<td>STS</td>
<td>Sequence tagged sites</td>
<td>629 573 650</td>
<td>634 263 196</td>
<td>0.7</td>
</tr>
<tr>
<td>TOTAL</td>
<td>All GenBank sequences</td>
<td>254 698 274 519</td>
<td>286 730 369 256</td>
<td>12.6</td>
</tr>
</tbody>
</table>

(Benson et al. (2011) Nucleic Acids Res D32:7)
<table>
<thead>
<tr>
<th>Organism</th>
<th>Non-WGS base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>14 792 487 417</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>8 859 010 528</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>6 443 768 086</td>
</tr>
<tr>
<td><em>Bos taurus</em></td>
<td>5 361 712 195</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>5 037 629 354</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>4 783 381 701</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>3 137 945 523</td>
</tr>
<tr>
<td><em>Strongylocentrotus purpuratus</em></td>
<td>1 352 920 226</td>
</tr>
<tr>
<td><em>Oryza sativa Japonica Group</em></td>
<td>1 197 245 122</td>
</tr>
<tr>
<td><em>Nicotiana tabacum</em></td>
<td>1 187 388 273</td>
</tr>
<tr>
<td><em>Xenopus (Silurana) tropicalis</em></td>
<td>1 147 132 278</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>1 047 707 620</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>1 001 926 471</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>1 001 073 627</td>
</tr>
<tr>
<td><em>Canis lupus familiaris</em></td>
<td>943 043 649</td>
</tr>
<tr>
<td><em>Vitis vinifera</em></td>
<td>913 911 649</td>
</tr>
<tr>
<td><em>Gallus gallus</em></td>
<td>891 463 513</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>886 103 518</td>
</tr>
<tr>
<td><em>Macaca mulatta</em></td>
<td>821 393 285</td>
</tr>
<tr>
<td><em>Ciona intestinalis</em></td>
<td>748 350 657</td>
</tr>
</tbody>
</table>
EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 108, June 2011), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in *Nucleic Acids Research* 2009 37: D19-D25, provides further information and details.

The EMBL nucleotide sequence database forms part of the European Nucleotide Archive, an EBI project led by Guy Cochrane as part of the The Protein and Nucleotide Database Group (PANDA) under Ewan Birney.

### Link

<table>
<thead>
<tr>
<th>Link</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography.</td>
</tr>
<tr>
<td>Submission</td>
<td>Primary sequence submissions, third party annotation, updates.</td>
</tr>
<tr>
<td>Publications</td>
<td>Group publications</td>
</tr>
<tr>
<td>People</td>
<td>Group members</td>
</tr>
<tr>
<td>Contact</td>
<td>How to contact the EMBL Nucleotide Sequence Database</td>
</tr>
<tr>
<td>News</td>
<td>List of recent changes on this site</td>
</tr>
</tbody>
</table>

### Contact

For information, comments and/or suggestions, please use the EBI Support Form page [http://www.ebi.ac.uk/support/](http://www.ebi.ac.uk/support/)
DDBJ: DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in USA.

Hot Topics

- Jul. 1, 2011 Release of TSA and EST data of Botryococcus braunii
- Feb. 22, 2011 DDBJ will continue Sequence Raw Data Archiving

Maintenance

- Jun. 23, 2011 (Jul. 1) ARSA database search temporary unavailable

Sequence Data Submission

- Submit my sequences
  Orientation for the data submission
- Update my entries
  Guidance for the update of the entry

FTP/Web API

- FTP (ftp.ddbj.nig.ac.jp)
  Download data files
- Web API
  Programmatic interfaces of DDBJ Web services

DNA Data Bank of Japan (DDBJ)

Center for Information Biology and DNA Data Bank of Japan (CIB–DDBJ)
National Institute of Genetics (NIQ)
SOKEI DAI
Department of Genetics
Research Organization of Information and Systems

International Nucleotide Sequence Database Collaboration

DDBJ exchanges data via the SINET3 computer network.
The “perfect” database

1. Comprehensive, but easy to search.
2. Annotated, but not “too annotated”.
3. A simple, easy to understand structure.
4. Cross-referenced.
5. Minimum redundancy.
6. Easy retrieval of data.
Problems with General Sequence Databases

• Databases that strive for encyclopedic completeness are now so huge as to be close to unmanageable.

1. Redundancy (nothing ever goes out).
2. Inadequate sequences.
   – old sequences
   – partially annotated sequences
   – inconsistent & outdated annotations (submitter annotation)
   – error sequences, low-quality sequences
   – contaminations
   – anonymous sequence
Release 57.5 of 07-Jul-09 of UniProtKB/Swiss-Prot contains 471,472 sequence entries, comprising 167,326,533 amino acids abstracted from 181,042 references.
NCBI Reference Sequences

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq is a foundation for medical, functional, and diversity studies; they provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses. [more...]

Scope

NCBI provides RefSeqs for taxonomically diverse organisms including eukaryotes, bacteria, and viruses. Additional records are added to the collection as data become publicly available.

May 12, 2011: RefSeq Release 47 available for FTP

This release includes:

- Proteins: 12,625,466
- Organisms: 12,000

To receive announcements of future RefSeq releases and incremental large updates please subscribe to NCBI's refseq-announce mail list: refseq-announce
### The RefSeq Accession number format and molecule types

<table>
<thead>
<tr>
<th>Accession</th>
<th>Molecule type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_xxxxxx</td>
<td>Complete genomic molecule</td>
</tr>
<tr>
<td>NG_xxxxxx</td>
<td>Genomic region</td>
</tr>
<tr>
<td>NM_xxxxxx</td>
<td>mRNA</td>
</tr>
<tr>
<td>NP_xxxxxx</td>
<td>Protein</td>
</tr>
<tr>
<td>NR_xxxxxx</td>
<td>RNA</td>
</tr>
<tr>
<td>NT_xxxxxx</td>
<td>computed Genomic contig</td>
</tr>
<tr>
<td>XM_xxxxxx</td>
<td>computed mRNA</td>
</tr>
<tr>
<td>XP_xxxxxx</td>
<td>computed Protein</td>
</tr>
</tbody>
</table>
Using Biological Databases

- What databases should I use?
- What kind of information I expect to find in this database?
- Is the data in database of interest to me?
- How reliable is it?
Practical Session: Outline

• Integrated systems: e.g., NCBI (Protein, Nucleotide, Gene, OMIM, etc.)
• Protein Databases: e.g., ExPASy (SwissProt + TrEMBL)
• Protein structures: e.g., PDB and PDBsum
• Pathway databases: e.g., KEGG (Kyoto Encyclopedia of Genes and Genomes)
Tips for the Practical Session

• We will go over several databases in a very short time. Don’t expect to remember all the small details. They are not important. All respectable databases have a “HELP” component.

• Try to:
  – Learn the common features of biological databases.
  – Understand the main features of every database.
  – Learn how to use the online HELP.
  – Judge and compare databases.
EBI/NCBI/DDBJ

• These 3 databases contain mainly the same information within 2-3 days (few differences in format and syntax)
• Serve as **archives** containing all sequences (single genes, ESTs, complete genomes, etc.) derived from:
  – Genome projects
  – Sequencing centers
  – Individual scientists
  – Literature
  – Patent offices
• Non-confidential data exchanged daily
• The database triples approximately every 12 months.
• Heterogeneous: sequence length, genomes, variants, fragments, …
• Minimum sequence size: 10 bp
• **Archive**: nothing goes out -> highly redundant!
• full of errors: in sequences, in annotations, in CDS attribution…. 
• no consistency of annotations; most annotations are done by the submitters; heterogeneity of the quality and the completion and updating of the information
• Unexpected information you can find:
• ACCESSION Z71230

<table>
<thead>
<tr>
<th>FT</th>
<th>source</th>
<th>1..124</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:4097&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/organelle=&quot;plastid:chloroplast&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Nicotiana tabacum&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/isolate=&quot;Cuban Cahibo cigar, gift from President Fidel Castro&quot;</td>
<td></td>
</tr>
</tbody>
</table>

• ACCESSION NC_001610

<table>
<thead>
<tr>
<th>FT</th>
<th>source</th>
<th>1..17084</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>/chromosome=&quot;complete mitochondrial genome&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:9267&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/organelle=&quot;mitochondrion&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Didelphis virginiana&quot;</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/dev_stage=&quot;adult&quot;</td>
<td></td>
</tr>
</tbody>
</table>
• There are 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011 (different today 2013, find out!).

• Most biocomputing sites update their copy of GenBank every day over the internet.

• Scientists access GenBank directly over the Web.
These billions of Gs, As, Ts, and Cs would be useless without the "annotation" in each sequence record.
The **LOCUS** field consists of five different subfields:

1a **Locus Name** (HSFHE) - The locus name is a tag for grouping similar sequences. The first two or three letters usually designate the organism. In this case **HS** stands for *Homo sapiens*. The last several characters are associated with another group designation, such as gene product. In this example, the last three digits represent the gene symbol, HFE. Currently, the only requirement for assigning a locus name to a record is that it is unique.

1b **Sequence Length** (12146 bp) - The total number of nucleotide base pairs (or amino acid residues) in the sequence record.
2 DEFINITION - Brief description of the sequence. The description may include source organism name, gene or protein name, or designation as untranscribed or untranslated sequences (e.g., a promoter region). For sequences containing a coding region (CDS), the definition field may also contain a “completeness” qualifier such as "complete CDS" or "exon 1."
3 ACCESSION (Z92910) - Unique identifier assigned to a complete sequence record. This number never changes, even if the record is modified. An accession number is a combination of letters and numbers that are usually in the format of one letter followed by five digits (e.g., M12345) or two letters followed by six digits (e.g., AC123456).
4 VERSION (Z92910.1) - Identification number assigned to a single, specific sequence in the database. This number is in the format “accession.version.” If any changes are made to the sequence data, the version part of the number will increase by one. For example U12345.1 becomes U12345.2. A version number of Z92910.1 for this HFE sequence indicates that the sequence data has not been altered since its original submission.
GI (1890179) - Also a sequence identification number. Whenever a sequence is changed, the version number is increased and a new GI is assigned. If a nucleotide sequence record contains a protein translation of the sequence, the translation will have its own GI number.
**KEYWORDS** (haemochromatosis; HFE gene) - A keyword can be any word or phrase used to describe the sequence. Keywords are not taken from a controlled vocabulary. Notice that in this record the keyword, "haemochromatosis," employs British spelling, rather than the American "hemochromatosis." Many records have no keywords. A period is placed in this field for records without keywords.
7 **SOURCE** (human) - Usually contains an abbreviated or common name of the source organism.

8 **ORGANISM** (*Homo sapiens*) - The scientific name (usually genus and species) and phylogenetic lineage. See the NCBI Taxonomy Homepage for more information about the classification scheme used to construct taxonomic lineages.
**9 REFERENCE** - Citations of publications by sequence authors that support information presented in the sequence record. Several references may be included in one record. References are automatically sorted from the oldest to the newest. Cited publications are searchable by author, article or publication title, journal title, or MEDLINE unique identifier (UID). The UID links the sequence record to the MEDLINE record.
1c Molecule Type (DNA) - Type of molecule that was sequenced. All sequence data in an entry must be of the same type.

1d GenBank Division (PRI) - There are different GenBank divisions. In this example, PRI stands for primate sequences. Some other divisions include ROD (rodent sequences), MAM (other mammal sequences), PLN (plant, fungal, and algal sequences), and BCT (bacterial sequences).

1e Modification Date (23-July-1999) - Date of most recent modification made to the record. The date of first public release is not available in the sequence record. This information can be obtained only by contacting NCBI at info@ncbi.nlm.nih.gov.
If the REFERENCE TITLE contains the words "Direct Submission," contact information for the submitter(s) is provided.
The FEATURES table

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Location/Qualifiers</th>
</tr>
</thead>
</table>
| source   | 1..12146
          | /organism="Homo sapiens"
          | /mol_type="genomic DNA"
          | /db_xref="taxon:9606"
          | /chromosome="5"
          | /map="5p"
          | /clone="ICRFy901D1223"
          | /clone_lib="ICRF YAC-library"
| gene     | 1028..10637
          | /gene="HFE"
| exon     | 1028..1324
          | /gene="HFE"
          | /number=1 |
| CDS      | join(1249..1324,4652..4915,5125..5400,6494..6769,
          | 6928..7041,7995..8035)
          | /gene="HFE"
          | /function="iron metabolism"
          | /note="haemochromatosis candidate gene"
          | /codon_start=1
          | /protein_id="CAB07442.1"
          | /db_xref="GI:1890180"
          | /db_xref="GO:Q030201"
          | /db_xref="UniProt/Swiss-Prot:Q030201"
          | /translation="MGPRARPALLLMLLTAVLQGRLLRSRSHSLHYLFMGASEQDLGL
          | SLFEALGYVDQLQVFYDHESSRVEPRTPWSSRISQMULQLSQSLKGWDHMFTVDF
          | UTIMENHNSKESHTLQVILGCMEQEDNSTEGYKYGYDQGDHLEFCPDRTLDUAAEP
          | RAUPTKLEVERHKIRARQNRAYLERDCPAQLQOLLELRGVLDDQVPPLVKVTTHVTS
          | SVTTLRICALNYYQPNNITKWLKDRQPMDAKEFEPKDVLPNGDGTYGQWITLAVPPGE
          | EQRYTCCQVEHPLQDLQPLIVIWEPSPSGTLVIGVISGIAVFFVVFILFIGILFIILRKRQG
          | SRGAMGHYVAERE"
| intron   | 1325..4651
          | /gene="HFE"
          | /number=1 |
| polyA_signal | 10617..10622 |
          | /gene="HFE" |
A feature is simply an annotation that describes a portion of the sequence.

Each feature includes a location (sequence location or interval) and one or several qualifiers.

Clicking on the feature name will open a record for the sequence interval identified in the feature location.

A list of features can be found in http://www.ncbi.nlm.nih.gov/collab/FT/
source - An obligatory feature. The source gives the length of the entire sequence, the scientific name of the source organism, and the Taxon ID number.

Other types of information that the submitter may include in this field are chromosome number, map location, clone, and strain identification.
**gene** - Sequence portion that delineates the beginning and end of a gene.
exon - Sequence segment that contains an exon. Exons may contain portions of 5' and 3' UTRs (untranslated regions). The name of the gene to which the exon belongs and exon number are provided.

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Location/Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td></td>
</tr>
<tr>
<td>/organism=&quot;Homo sapiens&quot;</td>
<td></td>
</tr>
<tr>
<td>/mol_type=&quot;genomic DNA&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;taxon:9606&quot;</td>
<td></td>
</tr>
<tr>
<td>/chromosome=&quot;6&quot;</td>
<td></td>
</tr>
<tr>
<td>/map=&quot;6p&quot;</td>
<td></td>
</tr>
<tr>
<td>/clone=&quot;ICRFy9012223&quot;</td>
<td></td>
</tr>
<tr>
<td>/clone_lib=&quot;ICRF YAC-library&quot;</td>
<td></td>
</tr>
<tr>
<td>gene</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/number=1</td>
<td></td>
</tr>
<tr>
<td>exon</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/number=1</td>
<td></td>
</tr>
<tr>
<td>CBS</td>
<td></td>
</tr>
<tr>
<td>/join=&quot;1249..1324,4652..4915,5125..5400,6494..6769,6928..7041,7995..8035&quot;</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/function=&quot;iron metabolism&quot;</td>
<td></td>
</tr>
<tr>
<td>/note=&quot;haemochromatosis candidate gene&quot;</td>
<td></td>
</tr>
<tr>
<td>/codon_start=1</td>
<td></td>
</tr>
<tr>
<td>/protein_id=&quot;CA807442.1&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;GI:1890180&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;GOA:Q00201&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;UniProt/Swiss-Prot:Q00201&quot;</td>
<td></td>
</tr>
<tr>
<td>/translation=&quot;MGPRARPAALLMLMLQTAALLQLAGRLRSLASLKLHFMLGAQEQQLQLGQLSLFELA0TVQGQLVYDFHDEKEEREPFRPVWSSRSSQWWLQWLSQSLQGLKGWDMFTVDFWTIMEPHHDSKESHTQVILGCQEMQCGSTEGYKDYQDGQBDLFCPDFTLDDQAEYFRAWPTKLEW3G3H334G4ARONRAYLQDCPQVLQQLKLLRGLGVLPQVQPLVWKV3334TSVTLTCLR33NYPN1THKWLKDQHPD3AKFEPKDVLHNPQGTYGQWTALVPPGEEEYTCQVERPGDQLQLIVWEPSSPGTGLVQIVGIAVFWVFLIGFILJERKQQQSR611GHTVLAERE&quot;</td>
<td></td>
</tr>
<tr>
<td>intron</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
<tr>
<td>/number=1</td>
<td></td>
</tr>
<tr>
<td>polyA_signal</td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;HFE&quot;</td>
<td></td>
</tr>
</tbody>
</table>

"
**CDS** - Sequence of nucleotides that code for amino acids of the protein product (coding sequence).

The CDS begins with the first nucleotide of the start codon and ends with the third nucleotide of the stop codon.

This feature includes the translation into amino acids and may also contain gene name, gene product function, link to protein sequence record, and cross-references to other database entries.
intron - Transcribed but spliced-out parts. Intron number is shown.
**polyA_signal** - Identifies the sequence portion required for endonuclease cleavage of an mRNA transcript. Consensus sequence for the polyA signal is AATAAA.
BASE COUNT & ORIGIN

BASE COUNT - Base Count gives the total number of adenine (A), cytosine (C), guanine (G), and thymine (T) bases in the sequence.

ORIGIN - Origin contains the sequence data, which begins on the line immediately below the field title.
Molecule-specific and topic-specific databases

AsDb - Aberrant Splicing db
ACUTS - Ancient conserved untranslated DNA sequences db
Codon Usage Db
EPD - Eukaryotic Promoter db
HOVERGEN - Homologous Vertebrate Genes db
IMGT - ImMunoGeneTics db [Mirror at EBI]
ISIS - Intron Sequence and Information System
RDP - Ribosomal db Project
gRNAs db - Guide RNA db
PLACE - Plant cis-acting regulatory DNA elements db
PlantCARE - Plant cis-acting regulatory DNA elements db
sRNA db - Small RNA db
ssu rRNA - Small ribosomal subunit db
lsu rRNA - Large ribosomal subunit db
5S rRNA - 5S ribosomal RNA db
tmRNA Website
tmRDB - tmRNA db
tRNA - tRNA compilation from the University of Bayreuth
uRNADB - uRNA db
RNA editing - RNA editing site
RNAmmod db - RNA modification db
SOS-DGBD - Db of Drosophila DNA sequences annotated with regulatory binding sites
TelDB - Multimedia Telomere Resource
TRADAT - TRAnscription Databases and Analysis Tools
Subviral RNA db - Small circular RNAs db (viroid and viroid-like)
UPDB - Molecular probe db
OPD - Oligonucleotide probe db
VectorDB - Vector sequence db
Organism specific databases:

FlyBase (Drosophila)
SGD (yeast)
MaizeDB (maize)
SubtiList (B. subtilis).
The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive links from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of Arabidopsis thaliana and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation with additional support from TAIR sponsors.

Updates on TAIR funding are available here.

New: GBrowse for 8 plant species at TAIR
(beta version)

Arabidopsis thaliana & lyra
Oryza sativa indica & japonica
Populus trichocarpa
Sorghum bicolor
Vitis vinifera
Zea mays
Physcomitrella patens
Brachypodium distachyon

GBrowse now available for eight plant species at TAIR
[May 19, 2011]
GBrowse instances for the following plants have been added to TAIR: Arabidopsis lyrata, Brachypodium distachyon, Oryza sativa japonica, Oryza sativa indica, Populus trichocarpa, Physcomitrella patens, Sorghum bicolor, Vitis vinifera, Zea mays.

AraCyc 8.0 available at the PMN [April 14, 2011]
The newest AraCyc release has 446 pathways and superpathways plus over 5500 enzymes. Come check out the new data on plant metabolism at the PMN!

ABRC and NASC fee increase [March 25, 2011]
The fees for individual stocks will increase to $7 each for academic users and to $56 each for commercial users.

New Education Resources available at ABRC [March 23, 2011]
Seven Education Kits containing seeds, DNA and written materials are now available from ABRC.
The search and retrieval system that integrates information from the National Center for Biotechnology (NCBI) databases.

These databases include nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and MEDLINE, through PubMed.
Input your search keywords or the Boolean expression

Search across databases: hiv - 2

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>biomedical literature citations and abstracts</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>free, full text journal articles</td>
</tr>
<tr>
<td>Books</td>
<td>online books</td>
</tr>
<tr>
<td>OMIM</td>
<td>online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>Site Search</td>
<td>NCBI web and FTP sites</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>sequence database (GenBank)</td>
</tr>
<tr>
<td>Protein</td>
<td>sequence database</td>
</tr>
<tr>
<td>Genome</td>
<td>whole genome sequences</td>
</tr>
<tr>
<td>Structure</td>
<td>three-dimensional macromolecular structures</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>organisms in GenBank</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>Gene</td>
<td>gene-centered information</td>
</tr>
<tr>
<td>HomoloGene</td>
<td>eukaryotic homology groups</td>
</tr>
<tr>
<td>PubChem Compound</td>
<td>small molecule chemical structures</td>
</tr>
<tr>
<td>PubChem Substance</td>
<td>chemical substances screened for bioactivity</td>
</tr>
<tr>
<td>Genome Project</td>
<td>genome project information</td>
</tr>
<tr>
<td>UniGene</td>
<td>gene-oriented clusters of transcript sequences</td>
</tr>
<tr>
<td>CDD</td>
<td>conserved protein domain database</td>
</tr>
<tr>
<td>3D Domains</td>
<td>domains from Entrez Structure</td>
</tr>
<tr>
<td>UniSTS</td>
<td>markers and mapping data</td>
</tr>
<tr>
<td>PopSet</td>
<td>population study data sets</td>
</tr>
<tr>
<td>GEO Profiles</td>
<td>expression and molecular abundance profiles</td>
</tr>
<tr>
<td>GEO DataSets</td>
<td>experimental sets of GEO data</td>
</tr>
<tr>
<td>Cancer Chromosomes</td>
<td>cytogenetic databases</td>
</tr>
<tr>
<td>PubChem BioAssay</td>
<td>bioactivity screens of chemical substances</td>
</tr>
<tr>
<td>GENSAT</td>
<td>gene expression atlas of mouse central nervous system</td>
</tr>
</tbody>
</table>
Databases: protein sequences


- **TrEMBL**: created in 1996; complement to SWISS- PROT; derived from EMBL CDS translations (« proteomic » version of EMBL)

- **PIR-PSD**: Protein Information Resources [http://pir.georgetown.edu/](http://pir.georgetown.edu/)

- **Genpept**: « proteomic » version of GenBank

- Many specialized protein databases for specific families or groups of proteins.
  - Examples: **AMSDb** (antibacterial peptides), **GPCRDB** (7 TM
<table>
<thead>
<tr>
<th>SWISS-PROT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collaboration between the SIB (CH) and EMBL/EBI (UK)</strong></td>
</tr>
<tr>
<td><strong>Manually annotated: non-redundant, cross-referenced, fully documented.</strong></td>
</tr>
<tr>
<td><strong>Weekly releases; available from about 50 servers across the world, the main source being ExPASy in Geneva</strong></td>
</tr>
</tbody>
</table>
- 495,880 sequences
- 174,780,353 amino acid residues
- 11,891 species
- 2,000 journals
- 276,903 authors
• 531,473 sequences
• 188,463,640 amino acid residues
• 12,564 species
• 2,154 journals
• 306,144 authors
TrEMBL (Translation of EMBL)

- It is impossible to cope with the *quantity* of newly generated data AND to maintain the high *quality* of SWISS-PROT -> TrEMBL, created in 1996.

- TrEMBL is automatically generated (from annotated EMBL coding sequences (CDS)) and annotated using software tools.

- Contains all that is *not* in SWISS-PROT. SWISS-PROT + TrEMBL = all known protein sequences.
The simplified story of a SWISS-PROT entry

cDNAs, genomes, ...

EMBLnew → EMBL

TrEMBLnew → TrEMBL

SWISS-PROT

Some data are not submitted to the public databases!!
(delayed or cancelled...)

« Automated »
- Redundancy check (merge)
- Family attribution (InterPro)
- Annotation (computer)

« Manual »
- Redundancy (merge, conflicts)
- Annotation (manual)
- SWISS-PROT tools (macros...)
- SWISS-PROT documentation
- Medline
- Databases (MIM, MGD....)

Once in SWISS-PROT, the entry is no more in TrEMBL, but still in EMBL (archive)

CDS: proposed and submitted at EMBL by authors or by genome projects (can be experimentally proven or derived from gene prediction programs). TrEMBL neither translates DNA sequences, nor uses gene prediction programs; only takes CDS proposed by the submitting authors in the EMBL entry.
NCBI - RefSeq

Main features of the RefSeq collection include:

1. Non-redundancy.
2. Explicitly linked nucleotide and protein sequences
3. Data validation and format consistency
5. Ongoing curation by NCBI staff and collaborators, with review status indicated on each record
Text based searching

- **Terminology**: query, hit, fields, logical/Boolean operator.
- **General principles**:
  1. All main databases provide a convenient tool for text base searching.
  2. We can search for query words in specific fields.
  3. We can search more than one database at a time.
  4. We can Pose additional limits, such as modification date.
Schizosaccharomyces pombe strain 972h- complete genome
Entrez cross-database search

Search across databases "homo sapiens" hemoglobin alpha HBA

- PubMed: biomedical literature citations and abstracts
- PubMed Central: free, full text journal articles
- Site Search: NCBI web and FTP sites
- Nucleotide: sequence database (includes GenBank)
- Protein: sequence database
- Genome: whole genome sequences
- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank
- SNP: single nucleotide polymorphism
- Gene: gene-centered information
- HomoloGene: eukaryotic homology groups
- PubChem Compound: unique small molecule chemical structures
- PubChem Substance: deposited chemical substance records
- UniGene: gene-oriented clusters of transcript sequences
- CDD: conserved protein domain database
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- PopSet: population study data sets
- GEO Profiles: expression and molecular abundance profiles
- GEO DataSets: experimental sets of GEO data
- Cancer Chromosomes: cytogenetic databases
- PubChem BioAssay: bioactivity screens of chemical substances
- GENSAT: gene expression atlas of mouse central nervous system
1: NM_000518
   Homo sapiens hemoglobin, beta (HBB), mRNA
gi:28302128[refNM_000518.4][28302128]

2: NM_000517
   Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA
gi:14043068[refNM_000517.3][14043068]

3: NM_007912
   Mus musculus epidermal growth factor receptor (Egfr), transcript variant 2, mRNA
gi:90403618[refNM_007912.4][90403618]

4: NM_207655
   Mus musculus epidermal growth factor receptor (Egfr), transcript variant 1, mRNA
gi:90403617[refNM_207655.2][90403617]

5: NG_000007
   Homo sapiens beta globin region (HBB) on chromosome 11
gi:28380635[refNG_000007.3][28380635]

6: NG_000006
   Homo sapiens beta globin region (HBB) on chromosome 11
gi:28380635[refNG_000006.3][28380635]
Sequence formats: FASTA format

>sequence name

[sequence]...
A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month:

**Glucansucrase**

We brush our teeth twice a day with fluoride toothpaste, use mouthwash, limit sugars in our diet... and we still get cavities. Cavities are caused by bacteria that consume some of the sugar in our diet, ferment it, and then release acids. These acids eat away at the hard minerals in our teeth. It seems like it would be easy to brush these bacteria away, and get rid of them once and for all. However, they have a trick to avoid this.

**Full Article...**

Protein Structure Initiative Featured Molecule:

**RNA Chaperone NMB1681**

MCSC researchers have solved the structure of an RNA chaperone, revealing a form-fitting binding site and a flexible tail.

**Full Article | PSI Featured Molecule Archive | PSI Structural Biology Knowledgebase**
Insulin

Insulin Therapy

Diabetes mellitus may be treated by manually replacing the insulin that is missing in the blood. Of course, we need a plentiful source of insulin for use in these treatments. Fortunately, insulin from pigs (at left, PDB entry 4ins) differs from human insulin (at right, PDB entry 2hlu) by only one amino acid:
UCSC Genome Bioinformatics

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Table quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGen lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBS) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

16 June 2011 - Re-engineered OMIM Tracks Released

We announce today the release of our newly re-engineered OMIM (Online Mendelian Inheritance in Man) tracks for both hg18 and hg19. With the kind assistance of Ada Hamosh (director), Joanna Amberger and Francois Schiefele of the OMIM project, we have divided the OMIM records into three separate tracks:

OMIM Allelic Variant SNPs
Variants in the OMIM database that have associated dbSNP identifiers.

OMIM Genes
The genomic positions of gene entries in the OMIM database. The coloring indicates the associated OMIM phenotype class.

OMIM Phenotypes - Gene Unknown
Regions known to be associated with a phenotype, but for which no specific gene is known to be causal. This track also includes known multi-gene syndromes.

The OMIM tracks are searchable by OMIM number. In most cases, simply typing the 6-digit OMIM number into the position/search box on the Browser will take you to the record.

The OMIM data are the property of Johns Hopkins University and will not be available for download from UCSC. Please contact the OMIM project at omim.org for download information.

UCSC thanks engineers Fan Hsu, Brooke Rhead and Robert Kahn for this release.

9 June 2011 - UCSC Preview Browser Available

Early access to ENCODE and other UCSC browser data tracks under construction is now available from the new UCSC Preview Browser site: http://genome-preview.ucsc.edu. Read more.

7 June 2011 - Updated Lizard Browser Available: We have released a Genome Browser for the May 2010 genome assembly of the green anole lizard, Anolis carolinensis (Broad version AnoCar2.0, UCSC version anoCar2). Read more.
Summary

• What is the best db for sequence analysis?
• Which does contain the highest quality data?
• Which is the more comprehensive?
• Which is the more up-to-date?
• Which is the less redundant?
• Which is the more indexed (allows complex queries)?
• Which Web server does respond most quickly?
Presents new databases and updates of existing databases
Thanks to:

- Laurent Falquet