Introduction to Biological Databases and Genome Browsers

Nuria Lopez-Bigas
Abel Gonzalez-Perez

BBGLab
IRB Barcelona
Biological Databases
Databank

1. An archive of information
2. Logical organization of the information (database schema)
3. Tools to access the information
Controlled vocabulary
A set of terms (such as gene names) that are unambiguous and non-redundant, and which are accepted as standards for the purpose of consistent database representation.

Ontology
• A systematically ordered representation of knowledge about a domain.
• An enumeration or formal description of the concepts used in a particular domain of knowledge, their definitions and relationships between them.
Bibliographic databanks

MEDLINE

• Repository of citations and abstracts of articles on the biomedical sciences

• Developed by the National Library of Medicine (NLM)

• Over 19 million unique records from 1950s
How to search on MEDLINE: PubMed

• Each article is indexed by title, abstract, authors, journal, year of publication, language and Medical Subject Heading (MeSH) terms

• MeSH is a **controlled vocabulary** that help to find articles that use alternative terms to refer to the same concept
A new nuclease member of the FAN club.

Huang M, D'Andrea AD.

Abstract
To cope with the life-threatening crisis of a DNA interstrand cross-link (ICL), human cells must invoke the Fanconi anemia (FA) DNA repair pathway. The FA pathway is a multistep repair process, requiring multiple nucleolytic incisions and translesion DNA synthesis. Recent work from four laboratories has identified a novel FA-associated nuclease, FAN1, that binds directly to monoubiquitinated FANCD2, resolving a decades-long puzzle regarding the function of this FANCD2 modification.

Publication Types, MeSH Terms, Substances, Grant Support
Publication Types:
- News
- Research Support, N.I.H., Extramural

MeSH Terms:
- Animals
- DNA Damage
- DNA Repair
- Exodeoxyribonucleases/metabolism
- Fanconi Anemia/enzymology
- Fanconi Anemia Complementation Group D2 Protein/metabolism
- Humans
- Proliferating Cell Nuclear Antigen/metabolism
- Ubiquitination
- Xenopus

Substances:
- Fanconi Anemia Complementation Group D2 Protein
- Proliferating Cell Nuclear Antigen
- Exodeoxyribonucleases
- FAN1 protein, human
The first databanks of biological information

First public molecular biology database was the Protein Data Bank (PDB) created in 1971.

In 1979, first DNA sequence database was established at Los Alamos National Lab (Gene Sequence Database), which derived into GenBank in 1982.

The European Molecular Biology Laboratory (EMBL) created its own electronic databank for DNA sequences in 1982, which contained 582 sequences.
Over time, an ecosystem of databases has developed
### Literature

<table>
<thead>
<tr>
<th>ID</th>
<th>Database</th>
<th>Description</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>7573</td>
<td>PubMed</td>
<td>scientific &amp; medical abstracts/citations</td>
<td>15</td>
</tr>
<tr>
<td>10414</td>
<td>PubMed Central</td>
<td>full-text journal articles</td>
<td>380</td>
</tr>
<tr>
<td>13</td>
<td>NLM Catalog</td>
<td>books, journals and more in the NLM Collections</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>MeSH</td>
<td>ontology used for PubMed indexing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Books</td>
<td>books and reports</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Site Search</td>
<td>NCBI web and FTP site index</td>
<td></td>
</tr>
</tbody>
</table>

### Health

<table>
<thead>
<tr>
<th>ID</th>
<th>Database</th>
<th>Description</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>PubMed Health</td>
<td>clinical effectiveness, disease and drug reports</td>
<td>100</td>
</tr>
<tr>
<td>26</td>
<td>MedGen</td>
<td>medical genetics literature and links</td>
<td>273</td>
</tr>
<tr>
<td>107</td>
<td>GTR</td>
<td>genetic testing registry</td>
<td>0</td>
</tr>
<tr>
<td>170</td>
<td>dbGaP</td>
<td>genotype/phenotype interaction studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ClinVar</td>
<td>human variations of clinical significance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OMIM</td>
<td>online mendelian inheritance in man</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OMIA</td>
<td>online mendelian inheritance in animals</td>
<td></td>
</tr>
</tbody>
</table>
**TP53**

<table>
<thead>
<tr>
<th>Nucleotide Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3361</strong> Nucleotide : DNA and RNA sequences</td>
</tr>
<tr>
<td><strong>6</strong> GSS : genome survey sequences</td>
</tr>
<tr>
<td><strong>1199</strong> EST : expressed sequence tag sequences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>60</strong> Genome : genome sequencing projects by organism</td>
</tr>
<tr>
<td><strong>0</strong> Assembly : genomic assembly information</td>
</tr>
<tr>
<td><strong>0</strong> Epigenomics : epigenomic studies and display tools</td>
</tr>
<tr>
<td><strong>126</strong> UniSTS : sequence-tagged sites for genome mapping</td>
</tr>
<tr>
<td><strong>1957</strong> SNP : short genetic variations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1340</strong> Gene : collected information about gene loci</td>
</tr>
<tr>
<td><strong>11</strong> HomoloGene : homologous gene sets for selected organisms</td>
</tr>
<tr>
<td><strong>141</strong> UniGene : clusters of expressed transcripts</td>
</tr>
</tbody>
</table>

| SRA : high-throughput DNA and RNA sequence read archive |
| PopSet : sequence sets from phylogenetic and population studies |
| Probe : sequence-based probes and primers |

| dbVar : genome structural variation studies |
| BioProject : biological projects providing data to NCBI |
| BioSample : descriptions of biological source materials |
| Clone : genomic and cDNA clones |

| GEO Profiles : gene expression and molecular abundance profiles |
| GEO DataSets : functional genomics studies |
When deciding which database to use...
1) The type of biological data

- DNA and protein sequences, including annotation
- Variations (e.g. SNPs, haplotypes, or disease associated mutations)
- Nucleic acid and protein structures, including annotation
- Databanks focused on organisms, including genome databases
- Databanks of protein expression patterns
- Databanks of metabolic, signalling and gene regulatory networks
- Databanks of molecular interactions (e.g. Protein-protein interactions)
2) The degree of “processing”

• Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the **submitter**
    • Examples: GenBank, Trace, SRA, SNP, GEO

• Derivative Databases
  - *Derived* from primary data
  - Content controlled by **third party**
    • Examples: NCBI Protein, Refseq, TPA, RefSNP, GEO

- Integrated HIC and Homology, Structures, Genomes...
3) The annotation and curation

• Plain
  - Experimental data with few (mostly automatic) annotations (Genomics Data Common)

• Enriched
  - Content (curation and annotation) carried out by specialists (Uniprot)
Genome browsers

Ensembl http://www.ensembl.org/  UCSC

http://www.genome.ucsc.edu/

Variation Databases


HapMap http://www.hapmap.org/
Gene Expression Databases

Array express [http://www.ebi.ac.uk/arrayexpress/] dbEST database

Tools for functional annotations of microarray experiments

Pathway databases and protein-protein interaction tools

Reactome http://www.reactome.org/  Biocarta

http://www.biocarta.com/


STRING http://string.embl.de/
Related with diseases, drugs


Disease cards http://www.diseasecard.org/


New databases:

Search for the database issue in the Nucleic Acids Research Journal

Database: the journal of biological databases and curation
http://database.oxfordjournals.org/
Exercises

(https://bbglab.irbbarcelona.org/courses/bco/practical_databases.php)

1. Searching PubMed

5. Retrieving protein-protein interaction data
Genome Browsers

...with a lot of help from ENSEMBL tutorials
Introduction to Ensembl

Why do we have genome browsers?

Why Ensembl?

Ensembl genes and genomes

Help and tutorials
Genome browsers provide a map

Figure adapted from the EnCODE project
www.nature.com/nature/focus/encode/
Genome Browsers

- Ensembl Genome browser
  http://www.ensembl.org

- NCBI Map Viewer

- UCSC Genome Browser
  http://genome.ucsc.edu
Ensembl Genome Browser

Protein/mRNA + Sequence Assembly → Ensembl Genes
UCSC Genome

UCSC Genome Browser on Human Mar. 2006 Assembly

Jump, clear, size 84,193 bp.

Tracks:
- Chromosome Band
- STS Markers
- FISH Clones
- BAC End Pairs

Mapping and Sequencing Tracks:
- Base Position
- Chromosome Band
- STS Markers
- FISH Clones
- BAC End Pairs

Move start and end:
- Click on a feature for details. Click on base position to zoom in around cunor. Click gray/blue bars on left for track options and descriptions.

Use drop down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.
Extending Ensembl across the taxonomic space


21 species
Drosophila (12)
Caenorhabditis (5)
Anopheles gambiae

EnsemblMetazoa

Chimpanzee
Pig
Platypus

Bacteria

Archaea

Eukaryota
Where does the data come from?

Proteins and cDNA/mRNA sequences from the research community found in:

- UniProt/Swiss-Prot *(manually curated)*
- UniProt/TrEMBL

[www.uniprot.org](http://www.uniprot.org)

- NCBI RefSeq *(manually curated)*

Combining genes and genomes

...tgccctgtag...

Exon

Untranslated+Coding

Exon

Coding

Exon

Untranslated
VEGA/Havana

• Automatic annotation pipeline: Gene building all at once (whole genome) Ensembl

• Manual curation: case-by-case basis VEGA: Vertebrate Genome Annotation Havana
Too many pieces...

Genome

Aligned cDNA and protein

Exon

Untranslated+Coding

Coding

Untranslated
Genes and Transcripts in Ensembl

• Ensembl known transcripts
• Ensembl novel transcripts
• Ensembl merged transcripts (Havana)

• EST clusters

• More manual curation (SGD, WormBase, FlyBase)
Ensembl/Havana

- Transcripts are labelled:
  - Ensembl
  - Havana
  - Ensembl/Havana merge

Known protein coding Ensembl/Havana merge transcript
Names in Ensembl

- **ENSG###** Ensembl Gene ID
- **ENST###** Ensembl Transcript ID
- **ENSP###** Ensembl Peptide ID
- **ENSE###** Ensembl Exon ID

For other species than human a suffix is added:

- MUS (*Mus musculus*) for mouse: ENSMUSG###
- DAR (*Danio rerio*) for zebrafish: ENSDARG###, etc.
Not just transcripts...

- Non-coding (nc)RNAs
- IDs in other databases
- microarray probes, clonesets, BAC maps
- Other features of the genome: repeats, CpG islands
- Comparative data: orthologues and paralogues, protein families, whole genome alignments, syntenic regions
- Variation data: SNPs, InDels
- Regulatory data (a first guess at promoter and enhancer elements)
- Data from external sources (DAS)
More information on ENSEMBL

http://www.ensembl.org/info/website/tutorials/index.html
Exercises

(https://bbglab.irbbarcelona.org/courses/bco/documents/Ensembl_genome_browser_handson.pdf)

Exercise on PTEN (1-3)