

Practical Course in Biodatabases

Practical Bioinformatics Module III –Biodatabases

Jarno Tuimala

CSC

What is a database?

- **A database is a collection of information stored in a computer in a systematic way, such that a computer program can consult it to answer questions. The software used to manage and query a database is known as a database management system (DBMS). The properties of database systems are studied in information science.**

Database types

- **Flat files (semi-structured text files)**
 - Traditionally used for sequence databases
 - large indexes needed
- **XML database**
 - Typically extensions of flat files
- **Relational databases**
 - Used for gene expression and genome databases
 - Data stored in tables (that are cross-referenced)

What makes a good database?

➤ **Quality**

- Manual (slow)
- No overlap between entries
- Reliable
- Some data might be missing

➤ **Coverage**

- Automatic (fast)
- Overlapping entries
- Errors, biases
- Up-to-date

Main sequence databases

➤ **DNA:**

- EMBL / Genbank / DDBJ
- RefSeq

ACGGGCTATGTAGTGCTAGC

➤ **Protein:**

- UniProt / SWISS-PROT
- RefSeq

YTCFSATFCFSAGDJSGAJGD

➤ **Genomes:**

- Ensembl
- UCSC Genome Browser

...

Hierarchy of databases - an illustrative example

Nucleotide

Protein

Genbank/EMBL/DDBJ

dbSNP

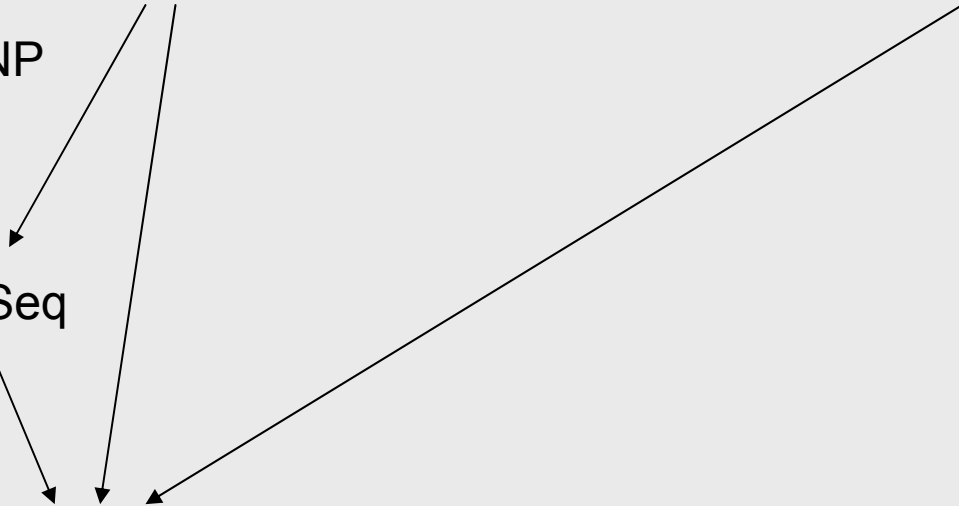
RefSeq

Ensembl

UniProt

} Primary

} Secondary



DNA sequences

➤ **EMBL / Genbank**

- Primary DNA sequence databases!
- Release v. update
- Divisions (hum, mus, est, ...)

➤ **RefSeq**

- Curated
- Less redundancy and errors

Protein sequences

➤ UniProt

- Minimal redundancy, specialist annotation, extensive cross-referencing
- Basically, contains three parts:
 - SWISS-PROT
 - TrEMBL ("translated EMBL")
 - PIR

NRDB

- **NRDB (non-redundant database) contains information combined from several sources**
 - Nucleotide: Genbank, RefSeq...
 - No ESTs, STSs (sequence tagged sites), GSSs (genome survey sequences) or HTGSs (high throughput genomic sequences)
 - Protein: translated Genbank, SWISS-PROT, RefSeq
- **Non-redundancy doesn't hold anymore!**

Genomes

➤ **Ensembl**

- European effort
- Contains only eukaryots

➤ **UCSC**

- University of California effort
- Insects! (hard to find elsewhere)

Others

- **dbSNP**
 - Database for single nucleotide polymorphisms (SNPs)
- **dbEST**
 - Database for expressed sequence tags (ESTs)
- **UniGene**
 - ESTs clustered to represent "genes"

Nuc. Acids. Res (2009) vol. 37 suppl. 1

http://nar.oxfordjournals.org/content/vol37/suppl_1/index.dtl

Others, cont.

➤ **OMIM / OMIA**

- Online Mendelian Inheritance in Man / Animals
- Used to be published as a book

➤ **Pubmed**

- Public Medline
- Contains abstract and links to articles

➤ **GO ontology**

- A controlled vocabulary of functionality etc. terms for, e.g., gene annotation

About accession numbers

- Every sequence entry is individually labeled with an accession number. E.g., from Genbank you can always retrieve the same sequence, if you know the accession number.
- Accession number: alpha-numeric code
- ID: human readable sequence name
- Some examples:

XRCC1	HUGO ID
M36089	EMBL accession number
P18887	UniProt accession number
NM_006297	RefSeq, nucleotide sequence
NP_006388	RefSeq, protein sequence
Hs.98493	UniGene ID
ENSG00000073050	Ensembl, gene sequence
ENSO00000262887	Ensembl, protein sequence
7515	Locuslink ID, Entrez Gene GeneID

Downloading the DBs

- **Most of the sequence databases can be downloaded and installed locally (on your own computer)**
 - This will make, e.g., blast searches much faster, but it takes much disk space
- **Some links**
 - <http://hgdownload.cse.ucsc.edu/downloads.html>
 - <http://www.ncbi.nlm.nih.gov/Ftp/>
 - <http://www.ebi.ac.uk/uniprot/database/download.html>
- **Remember that these are copyrighted!**

Queries

Entrez – main page

NCBI HomePage - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/>

NCBI
National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for Go

SITE MAP
Alphabetical Resource
About NCBI
GenBank
Literature databases
Molecular databases
Genomic biology

What NCBI do?
Established in 1988 as a national resource for biology information, NCBI creates databases, conducts research in molecular biology, develops software analyzing genome data, and disseminates biomedical information - all for understanding of molecular biology affecting human health and disease.

Hot Spots
► Clusters of orthologous groups
► Coffee Break, Genes & Disease, NCBI Handbook
► Electronic PCR
► Entrez Home
► Entrez Tools
► Gene expression omnibus (GEO)
► Human genome resources
► Malaria genetics & genomics
► Map Viewer
► dbMHC

Sequencing databases
► GenBank
► EMBL
► DDBJ

GenBank
GenBank has reached a milestone of 100 billion bases from over 100,000 organisms. See the [press release](#) or find out more [information on GenBank](#).

My NCBI
The new My NCBI has replaced the Cubby and includes automatic search updates and filtering search results. The new format is used for features such as displaying filtered search results.

Free text search

Pick a database

Entrez – use limits for filtering

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&cmd=search&term=> Go Links >>

NCBI

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for [] Go Clear

Limits Preview/Index History Clipboard Details

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of June 2005, there are over 89 billion bases in GenBank and RefSeq alone.

Human Genome
Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

Building the human genome
The Human Genome Reference DNA Sequence was complete in 2003. The current version is listed as a build number on the [View](#) page and includes an accompanying set of [statistics notes](#).

Homo sapiens genome view
build 35 version 1 statistics

Limits Preview/Index History Clipboard Details

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available [here](#).

Limited to:

All Fields [v]
 exclude ESTs exclude GSS exclude STSs exclude TPA exclude working draft exclude patents exclude all of the above

Molecule [v] Gene Location [v] Segmented Sequences [v]

Only from [v] Modification Date [v]

Modification Date [v] From [] [] [] To [] [] []
Use the format YYYY/MM/DD; month and day are optional.

Entrez – use qualifiers as filters

- **Search all the sequences containing term XRCC**
 - P4b*
- **Search using the author's name**
 - Tuimala J or Tuimala J[AUTH]
- **Search using the journal's name**
 - J. Gen. Virol.[JOUR]
- **Search using an organism name**
 - Human[ORGN]
- **Search using the sequence length**
 - 400[SLEN]
 - 400:500[SLEN]
- **Date limits**
 - 2005/12/01:2005/12/31[PDAT] (YYYY/MM/DD)
- **Search by accession numbers**
 - M18838[ACCN] or M18838:M18848[ACCN]
- **Terms can be combined, e.g.**
 - P4b* Tuimala J[AUTH] 400:500[SLEN]

Getting the sequences

The screenshot shows the NCBI Entrez search interface. The search term is 'Tuimala J'. The results are displayed in a list format. The first three results are:

- 1: [AY455311](#) Reports
Red deer parapoxvirus strain DPV putative viral core protein P4b gene, partial cds
[gi|42741865|gb|AY455311.1|](#)[42741865]
- 2: [AY455310](#) Reports
Pseudocowpox virus strain VR634 putative viral core protein P4b gene, partial cds
[gi|42741863|gb|AY455310.1|](#)[42741863]
- 3: [AY455309](#) Reports
Pseudocowpox virus strain BO35 putative viral core protein P4b gene, partial cds
[gi|42741861|gb|AY455309.1|](#)[42741861]

The 'Display' dropdown is set to 'Summary'. The 'Show' dropdown is set to '20'. The 'Text' dropdown is set to 'Text'. The 'FASTA' format is selected in the 'Display' dropdown. The 'Clipboard' button is highlighted. The 'Tick!' callout points to the 'Clipboard' button. The 'FASTA / Text' callout points to the 'Text' dropdown. The 'Copy & Paste' callout points to the 'Clipboard' button.

Tick!

FASTA / Text

Copy & Paste

```
>gi|42741865|gb|AY455311.1| Red deer parapoxvirus strain DPV putative viral core protein P4b gene, partial cds
TGGGGCCGACGCCGGTGGTTGTTGGTGCAGCTGCTGCAGCAGCGAGAGGGCTGTCCCGGATCACGAAGGCCGG
CTCCAGCTGCCGGCAGCGCCACGCCGAAAGAGCGCCATGACATGGGGTGCACGCAAGTCCAGGAGTCGCAG
GACTTGAACGAGAAGAGCAGGTCGCTGGAGGAGGACATGTCCCTTGAAGTGGCTCGACTGGAACCTGCGTGG
TGGAGAGCAAGTTCTGGTAGCTGGAGAGGTTGTACAGGCTCTCCAGCTCCTTACATCTGCCCTGTTACAGCG
CTGCGAGTTCGGCGTGTAGATCATCACAGCGGGTGGCAGTGCCTGACCGTCACGCGCGGTGTCGGCCATG
GCCGCGCGCAGCTTGTGAGCAGGTGCAAGAGCTCCGAGCGGTGCGGGCCCCGGATGCCACGCTTGTCT
```

```
>gi|42741863|gb|AY455310.1| Pseudocowpox virus strain VR634 putative viral core protein P4b gene, partial cds
CGGGCCGACGCCGGTGGTTGTTGGTGCAGCTGAGCCAGCAGCGAAGAGCTGTCCCGGTCACGAAGGCCGGT
CTCGATGGCGGGCAGCTTCCAGCGCAACAGAGCCATGACATGGGGTGCACGCAAGTCCAGGAGTCGGTA
GACTTGAACGAGAATGAAAGATCGCTGGATGAGCTCATGTCTTGAAGTGGCGGACTGGAACCTGTGTGG
TGGAAACAGGTTCTGGTAGCTGGAGCATGCTCTGCAGGTCGCTGATCTCCTTACATCTGCTTGTTCACGCG
AGTGCATCCCGTCCGATGATGATCATCACAGCGGATGGGCTCGGCTCACAGCAGGCGCGAGTCCGCTCATG
CGGGCCGCAAGCGCTGTGCAGCAATCGAACAGCTCCGAGCGGTCTCTGTTGAATACCCACCTTCGA
```

FastA-format

```
>gi|42741865|gb|AY455311.1| Red deer...  
TGGGGCGCACGCGGTGGTTGTGGTGC
```

```
>RedDeer parapox DPV P4b AY455311  
TGGGGCGCACGCGGTGGTTGTGGTGC
```

- You can modify the title to your liking, but always retain the accession number!

A sequence record

Genbank accession number

Links to other data sources (get protein seqs, etc.)

1: [AY455311](#)

Reports

Links

Red deer parapoxvirus strain DPV putative viral core protein P4b gene, partial cds
gi|42741865|gb|AY455311.1|[42741865]

Sequence description

- Genbank GI-number (identifies the sequence, also).
- Note that the accession number (AY455311.1) has a version number in the end. If you search other databases (like EMBL) for the same sequence, you need to omit the version number (.1).

What if you have the acc. numbers?

- Under Entrez tool, there a Batch Entrez that let's you retrieve several sequences at the same time, if you know the accession numbers.

- Make a list of the numbers in Notepad (one per line)
- Save the list as text file
- Retrieve the sequences

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Go Print Mail New Tab

Address http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide

NCBI Nucleotide

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Database Nucleotide File: C:\MyFiles\polta\trees.txt Browse... Retrieve

Batch Entrez

Batch Entrez has changed!

You will need Batch Entrez now only to upload a file of GI or accession numbers for an Entrez search. You can do all other large searches directly within Entrez.

There are now two ways to retrieve large amounts of sequence data using Entrez:

1. Begin at one of the [Entrez](#) pages and perform a large search just like any other Entrez search:
 - o Enter a query
 - o Designate the database as Nucleotide or Protein, as appropriate
 - o Press Go, you will see a list of document summaries
 - o Select a format for the sequence data
 - o Press Save to download, you will be prompted for a filename

or

2. Begin at the top of this **Batch Entrez page** and input (upload) a list of GI or accession numbers :
 - o Create a file with a list of GI or accession numbers and

Sequence submission

Genbank / EMBL policy

- **Some journals need you to submit the sequences before the publication is accepted.**
- **You can select a date when your sequences will be published, i.e., sequences can be submitted to the database, but kept secret for a few month before the publication appears in the journal.**
- **Only the submitter can later change the record.**

Genbank / EMBL submission

- **One or a few sequences**
 - BankIt (Genbank), Webin (EMBL)
- **Several sequences**
 - Sequin (Genbank), personal contact (EMBL)
- **You need to fill in the www-forms with all the data that appears in the record, including the description the coding regions, etc.**

Genbank does not allow

- **Shorter than 50bp of sequence**
- **Only primers**
- **Only protein sequence**
- **Multiple exons without the intron sequences**
- **Mix of genomic and mRNA sequence**

Genome databases:
Ensembl, UCSC, MapViewer



What are genome databases?

- **Genome databases contain, well, genomic information collected from many sources.**
 - Genome assembly
 - Gene predictions
 - Known genes, mRNA, ESTs, proteins
 - Genetic maps, markers and polymorphisms
 - Gene expression and phenotypes
 - Annotations
 - Interspecies homologues

Why genome databases?

- **Genome structure**
- **Gene identification**
- **Complete catalog or blueprint**
- **Rapid identification of proteins**
- **Genetic, transcriptome, proteome analysis**
- **Comparative genomics**

Databases to be introduced

➤ **Ensembl**

- <http://www.ensembl.org>
- 19 species (Chordates!)

➤ **UCSC Genome Browser**

- <http://genome.ucsc.edu/>
- 28 species (Insects!)

➤ **NCBI MapViewer**

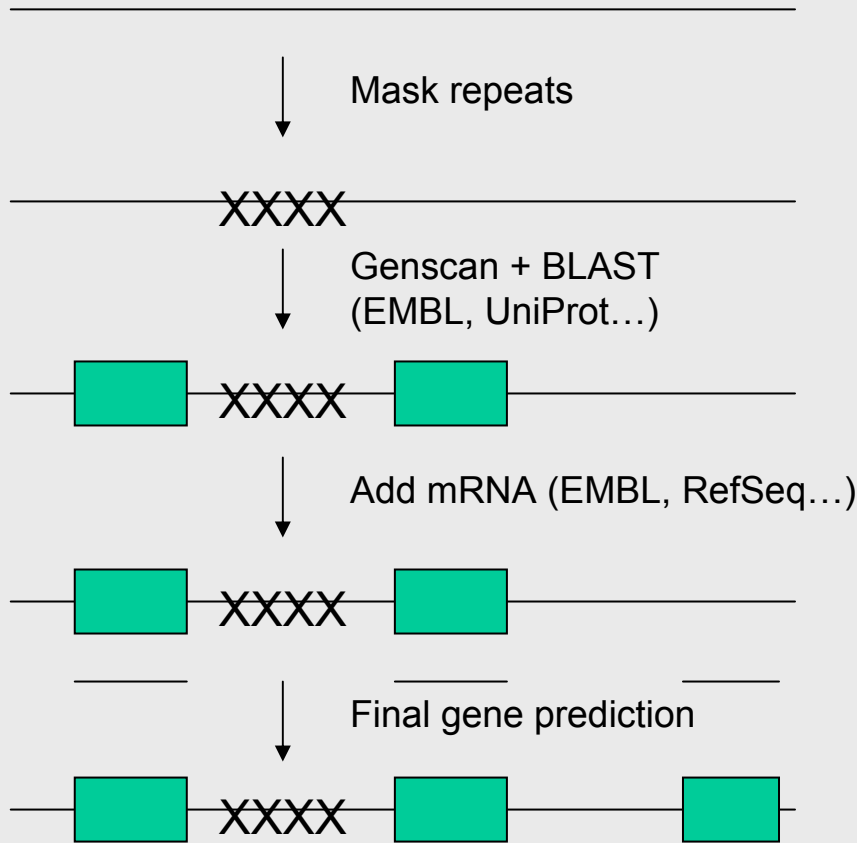
- <http://www.ncbi.nlm.nih.gov/mapview/>
- 38 species (Plants, Fungi!)

There's no single truth

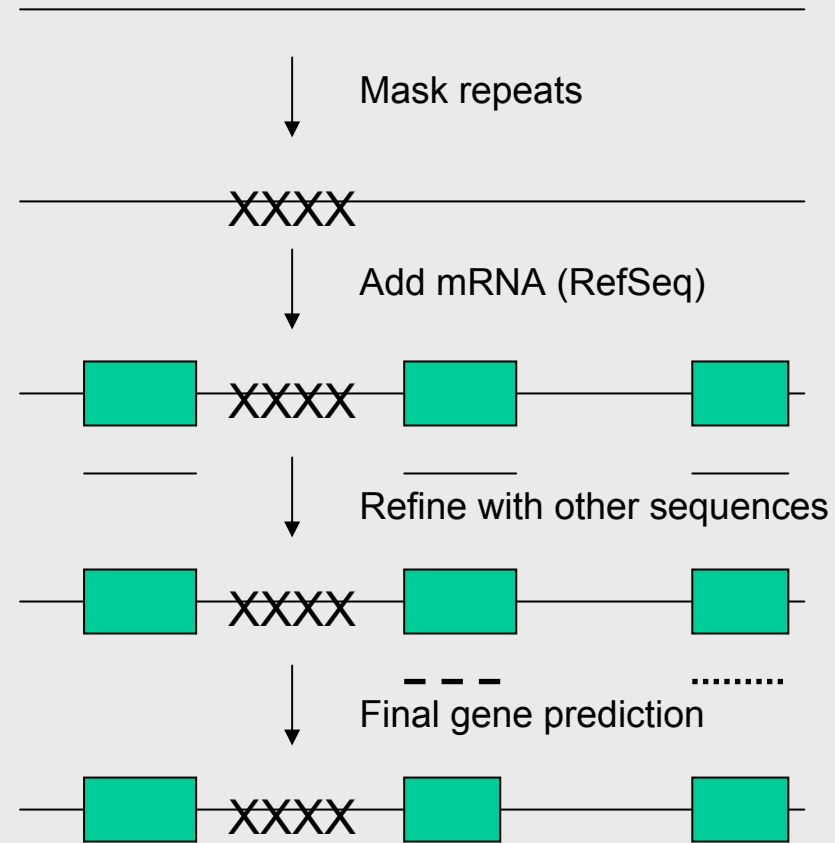
- **Number of human genes:**
 - 24 194 (Ensembl)
 - 23 951 (UCSC)
 - 26 626 (MapViewer)
 - 24 625 (RefSeq mRNAs)
- **And all use (almost) the same genomic assembly from 2004!**
- **So where is the difference?**

Gathering data

Ensembl



MapViewer



Other organisms

➤ **Yeast:**

- <http://www.yeastgenome.org/>

➤ **Microbes:**

- <http://www.tigr.org/tdb/mdb/mdbcomplete.html>

➤ **Parasites, single celled eukaryots...**

- <http://www.tigr.org/tdb/euk/>
- <http://www.sanger.ac.uk/Projects/>

Some considerations

- **Selection of the database**
 - Organism content
 - Speed (MapView can be slow)
- **Organism specific databases can be more up-to-date than general databases**
- **Genome databases are not a one stop shop for all information, other databases like EMBL and UniProt are still needed**

Queries to Ensembl

Ensembl front page

Ensembl Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/index.html

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl Genome Browser Nucl. Acids Res. -- Table of Contents (...)

e!Ensembl
Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl

Search: All species for
 Go

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

- Human**
NCBI36
- Mouse**
NCBIM37
- Zebrafish**
ZFISH7

All genomes

Select a species

New to Ensembl?

Did you know you can:

- Add custom tracks**
using our new Control Panel
- Upload your own data**
and save it to your Ensembl account
- Search for a DNA or protein sequence**
using BLAST or BLAT
- Fetch only the data you want**
from our public database, using the Ensembl Perl API
- Download our databases via FTP**
in FASTA, MySQL and other formats
- Mine Ensembl with BioMart**
and export sequences or tables in text, html, or Excel format

Still got questions? [Try our FAQs](#)

What's New in Release 52 (9 December 2008)

- Homo sapiens core database** (Human)

Find: transmembrane Next Previous Highlight all Match case

Done

Browse genome

Explore the genome

Ensembl 52: H.sapiens - Description - Search Ensembl Human - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Info/Index

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Descripti...

e!Ensembl
Home > Human Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

About this species

- Description
 - Genome Statistics
 - Assembly and Genebuild
 - Top 40 InterPro hits
 - Top 500 InterPro hits
 - What's New
- Sample entry points
 - Karyotype**
 - Location (AL032821.2.1.)
 - Gene (BRCA2)
 - Transcript (FOXP2-203)

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Search Ensembl Human

Search for: Go
e.g. [gene BRCA2](#) or [AL032821.2.1.143563](#) or [muscular dystrophy](#)

Assembly and Genebuild »

Assembly

This release is based on the NCBI 36 assembly of the [human genome](#) [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS and a number of alternative assemblies of individual haplotypic chromosomes or regions. [Full list of assemblies →](#)

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- [Nature 431, 931 - 945 \(21 October 2004\)](#)
- [WT Sanger Institute Press Release](#)

Annotation

Since release 38 (April 2006) the gene annotation presented has been a combined Ensembl-[Havana](#), geneset which incorporates more than 18,000 full-length protein-coding transcripts annotated by the Havana team with the Ensembl automatic gene build. The human genome sequence is now considered sufficiently stable that since 2004 the major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed and these identifiers are also shown.

- More information about the [CCDS project](#).

The [ENCODE](#) (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

- More information about the [ENCODE resources](#) at Ensembl.

Find: Next Previous Highlight all Match case

Done

Karyotype
(=chromosome
view)

Explore chromosomes

Ensembl 52: H.sapiens - Whole genome - Karyotype - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Location/Genome

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Whole ge...

Ensembl

Home > Human

Genome

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

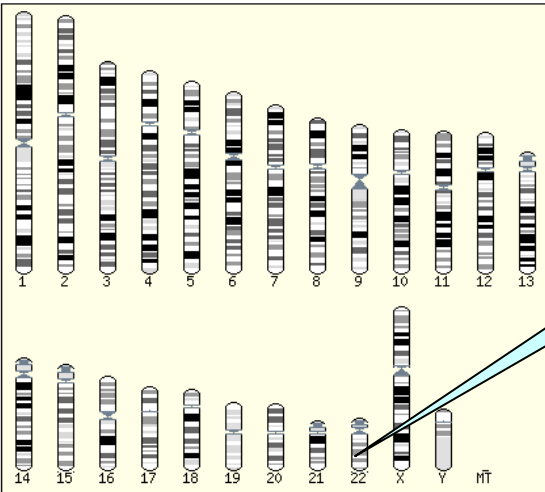
Location-based displays

- Whole genome**
 - Chromosome summary
 - Region overview
 - Region in detail
- Comparative Genomics
 - Genomic alignments
 - Synteny (10)
- Genetic Variation
 - Resequencing (6)
 - Linkage Data
 - Markers

• Configure this page
• Add custom data to page
• Export data
• Bookmark this page

Karyotype

Whole genome [help](#)



1 2 3 4 5 6 7 8 9 10 11 12 13
14 15 16 17 18 19 20 21 22 X Y MT

Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

Find: transmembrane

Next Previous Highlight all Match case

Done

Select a chromosome -> chromosome summary

Chromosome summary

Ensembl 52: H.sapiens - Chromosome summary - Chromosome 22: 33,885,613-33,985,613 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Location/Chromosome?r=22:33885613-33985613

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Chromos...

Location-based displays

- Whole genome
- Chromosome summary**
- Region overview
- Region in detail
- Comparative Genomics
 - Genomic alignments
 - Synteny (10)
- Genetic Variation
 - Resequencing (6)
 - Linkage Data
 - Markers

● Configure this page
● Add custom data to page
● Export data
● Bookmark this page

Chromosome 22: 33,885,613-33,985,613

Assembly exceptions
chromosome 22 p13 p11.2 q11.21 q12.1 q12.2 q12.3 q13.1 q13.2 q13.31
Assembly exceptions

Export Image

« Whole genome **Chromosome summary help** Region overview »

Chromosome 22	Known Genes	% GC Repeats	Variations
p13			
p11.2			
q11.21			
q11.22			
q11.23			
q12.1			
q12.2			
q12.3			
q13.1			
q13.2			

Find: transmembrane Next Previous Highlight all Match case

Done

Synteny

Synteny View

Ensembl 52: H.sapiens - Synteny - Chromosome 22: 33,885,613-33,985,613 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Location/Synteny?r=22:33885613-33985613

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Synteny ...

Genetic Variation
Resequencing (6)
Linkage Data
Markers

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

50M
100M
Chr 6

50M
100M
Chr 10

50M
100M
Chr 11

50M
100M
Chr 15
Mus musculus

10M
20M
30M
40M

50M
100M
Chr 16

50M
100M
Chr 5

50M
100M
Chr 8
Mus musculus

Find: transmembrane Next Previous Highlight all Match case

Done

Ensembl front page

BioMart

Ensembl Genome Browser - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Stop

Address <http://www.ensembl.org/index.html>

e! Ensembl Search all Ensembl:

Ensembl v33 - Sep 2005

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre! Pre Ensembl
- Trace server
- Archive! sites
- Stable Archive! link for this page

browse a genome

Mammals

- Homo sapiens** [NCBI 35]
[browse](#) | [what's new](#) | [Vega](#)
- Pan troglodytes** [CHIMP1]
[browse](#) | [what's new](#)
- Pre!** **Macaca mulatta** [Mmul 0.1]
[browse](#) [pre!](#) site
- Mus musculus** [NCBI m34]
[browse](#) | [what's new](#) | [Vega](#)
- Rattus norvegicus**
[RGSC 3.4]
[browse](#) | [what's new](#)
- Canis familiaris** [CanFam1.0]
[browse](#) | [what's new](#) | [Vega](#)
- Bos taurus** [Btau 1.0]
[browse](#) | [what's new](#) | [pre!](#)
[Btau 2.0]
- Pre!** **Monodelphis domestica**
[MonDom2]
[browse](#) [pre!](#) site

Other chordates

- Gallus gallus** [WASHUC1]
[browse](#) | [what's new](#)
- Xenopus tropicalis** [JGI 3]
[browse](#) | [what's new](#)
- Danio rerio** [WTSI Zv5]
[browse](#) | [what's new](#) | [Vega](#)
- Takifugu rubripes**
[Fugu 2.0]
[browse](#) | [what's new](#)
- Tetraodon nigroviridis**
[TETRAODON 7]
[browse](#) | [what's new](#)
- Ciona intestinalis** [JGI 1.95]
[browse](#) | [what's new](#)

Other eukaryotes

- Drosophila melanogaster**
[BGDP 4]
[browse](#) | [what's new](#)
- Anopheles gambiae**
[MOZ 2]
[browse](#) | [what's new](#)
- Apis mellifera** [Amel 2.0]
[browse](#) | [what's new](#)
- Caenorhabditis elegans**
[WVS140]
[browse](#) | [what's new](#)
- Saccharomyces cerevisiae** [SGD]
[browse](#) | [what's new](#)

About Ensembl **What's New in Ensembl 33**

Done Internet

Ensembl front page

Ensembl Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/index.html

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl Genome Browser Nucl. Acids Res. -- Table of Contents (...)

Ensembl
ome

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl

Search: All species for

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

- Human**
NCBI36
- Mouse**
NCBIM37
- Zebrafish**
ZFISH7

All genomes

Select a species -

Find: Highlight all Match case

Quick
search

Quick search results

Ensembl 52: H.sapiens - text search - Mozilla Firefox

http://www.ensembl.org/Homo_sapiens/Search/Summary?species=Homo_sapiens;idx=q;qx=rcc1

Ensembl
Home > Human
Genome

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl

- Feature type (2)
 - Domain (1)
 - Homo sapiens (1)
- Gene (1)
 - Homo sapiens (1)
- Species (1)
 - Homo sapiens (2)
 - Domain (1)
 - Gene (1)

[Configure this page](#)

[Add custom data to page](#)

[Export data](#)

[Bookmark this page](#)

Ensembl text search

xrcc1 corporate/tree:"Top/Species/Homo sapiens"

Your query matched 2 entries in the search database

[Ensembl protein_coding Gene: ENSG00000073050 \(HGNC \(automatic\): XRCC1\) \[Region in detail\]](#)

Ensembl protein_coding gene ENSG00000073050 has 1 transcript: ENST00000262887, associated peptide: ENSP00000262887 and 17 exons: ENSE00000710421, ENSE00000710449, ENSE00000710480, ENSE00000710504, ENSE00000710525, ENSE00000710543, ENSE00000710568, ENSE00000710592, ENSE00000710618, ENSE00000710635, ENSE00000710666, ENSE00000710692, ENSE00000710723, ENSE00000847423, ENSE00000897653, ENSE00001161480, ENSE00001213781, ENSE00001213782

DNA-repair protein [XRCC1](#) (X-ray repair cross-complementing protein 1) [Source:UniProtKB/Swiss-Prot;Acc:Q99566]

The gene has the following external identifiers mapped to it:

Affymx Microarray Focus: 203655_at

Affymx Microarray HCG110: 1829_at

Affymx Microarray Human Exon 1.0 ST v2: 3835107

Affymx Microarray U133: 203655_at, g5454171_3p_at

Affymx Microarray U95: 160033_s_at, 1829_at

Agilent CGH: A_14_P135035

Agilent Probe: A_23_P153692, A_23_P153687

CCDS: CCDS12624.1, CCDS12624

EMBL: M36089, AK293542, AF512504, BC023593, AC018758, AK315332, AK300163, CR456728, AB208781

EntrezGene: [XRCC1](#), RCC, 7515

GE Healthcare/Amersham Codelink WGA: GE57951

GO: GO:0005634, GO:0005654, GO:0005515, GO:0003684, GO:0005622, GO:0000012, GO:0006974

HGNC (automatic): [XRCC1](#)-201, [XRCC1](#), 12828

HGNC Symbol: [XRCC1](#), RCC, 12828

Human Protein Atlas: HPA006717, 5427, CAB005427, 6717

Illumina V1: GI_5454171-S

Illumina V2: ILMN_26577

IPI: IPI00911037.1, IPI00002564.3, IPI00002564, IPI00911037

MIM gene: 194360

PDB: 1CDZ 1XNT 1XNA 2D8M

Find: transmembrane Match case

Done

Geneview link

Gene View

Ensembl 52: H.sapiens - Gene summary - Gene: XRCC1 (ENSG00000073050) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000073050

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Gene su... Nucl. Acids Res. -- Table of Contents (...)

e!Ensembl
Home > Human
Location: 19:48,739,304-48,771,555 **Gene: XRCC1** Transcript: XRCC1-201
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Gene: XRCC1

- Gene summary
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Comparative Genomics
 - Genomic alignments (36)
 - Gene Tree
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (34)
 - Paralogues (0)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
- ID History
 - Gene history

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Gene: XRCC1 (ENSG00000073050)

DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1) [Source: UniProtKB/Swiss-Prot P18887](#)

Location [Chromosome 19: 48,739,304-48,771,555](#) reverse strand.

Transcripts There is one transcript in this gene: [XRCC1-201 \(ENST00000262887\)](#), with protein product [ENSP00000262887](#).

Gene summary [help](#) [Splice variants »](#)

Name [XRCC1](#) (HGNC (automatic))

Synonyms RCC [To view all Ensembl genes linked to the name [click here](#).]

CCDS This gene is a member of the Human CCDS set: [CCDS12624](#)

Gene type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl [genebuild](#).

Transcripts

Configuring the display [Export image](#)

Find: transmembrane [Next](#) [Previous](#) Highlight all Match case

Done

Gene View

Information on transcripts and proteins.

Ensembl 52: H.sapiens - Gene summary - Gene: XRCC1 (ENSG00000073050) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000073050

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Gene su... Nucl. Acids Res. -- Table of Contents (...)

e!Ensembl
Home > Human
Login / Register | BLAST/BLAT | Docs & FAQs

Location: 19:48,739,304-48,771,555 **Gene: XRCC1** Transcript: XRCC1-201

Gene: XRCC1

- Gene summary
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Comparative Genomics
 - Genomic alignments (36)
 - Gene Tree
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (34)
 - Paralogues (0)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
- ID History
 - Gene history

• Configure this page
• Add custom data to page
• Export data
• Bookmark this page

Gene: XRCC1 (ENSG00000073050)
DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1) Source: UniProtKB/Swiss-Prot: P18887

Location [Chromosome 19: 48,739,304-48,771,555](#) reverse strand.

Transcripts There is one transcript in this gene: [XRCC1-201 \(ENST00000262887\)](#), with protein product [ENSP00000262887](#).

Gene summary [help](#) [Splice variants »](#)

Name [XRCC1](#) (HGNC (automatic))

Synonyms RCC [To view all Ensembl genes linked to the name [click here](#).]

CCDS This gene is a member of the Human CCDS set: [CCDS12624](#)

Gene type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl [genebuild](#).

Transcripts

Ensembl/Havana gene
Known protein coding Ensembl gene
Cortigs
Ensembl/Havana gene
Known protein coding Ensembl gene

48.73 Mb 48.74 Mb 48.75 Mb 48.76 Mb 48.77 Mb 48.78 Mb

52.25 Kb Forward strand

Reverse strand

[Export image](#)

Configuring the display

Find: [Next](#) [Previous](#) Highlight all Match case

Done

Transcript info

Ensembl 52: H.sapiens - Transcript summary - Transcript: XRCC1-201 (ENST00000262887) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000073050;r=19:487c country codes .pt

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Transcri...

e!Ensembl

Home > Human Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Location: 19,48,739,304-48,771,555 Gene: XRCC1 Transcript: XRCC1-201

Transcript-based displays

- Transcript summary
- Exons (17)
- Supporting evidence (12)
- Sequence
 - cDNA
 - Protein
- External References
 - General identifiers (40)
 - Oligo probes (11)
 - Gene ontology (7)
- Genetic Variation
 - Population comparison
 - Comparison image
- Protein Information
 - Protein summary
 - Domains & features (17)
 - Variations (37)
- External Data
- ID History
 - Transcript history
 - Protein history

• Configure this page
• Add custom data to page
• Export data
• Bookmark this page

Transcript: XRCC1-201 (ENST00000262887)

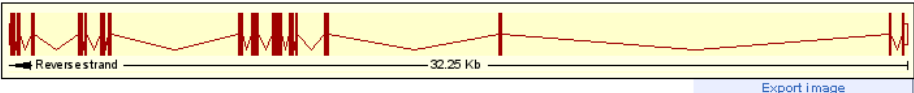
DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P18887]

Location [Chromosome 19: 48,739,304-48,771,555](#) reverse strand.

Gene This transcript is a product of gene [ENSG00000073050](#)

Protein [ENSP00000262887](#) is the protein product of this transcript

Transcript summary [help](#) [Exons »](#)



Statistics Exons: 17 Transcript length: 2,087 bps Translation length: 633 residues

CCDS This transcript is a member of the Human CCDS set: [CCDS12624](#)

Type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl [genebuild](#).

Ensembl release 52 - Dec 2008 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)
[Permanent link](#) - [View in archive site](#)

Find: [Next](#) [Previous](#) Highlight all Match case

Done

Protein summary

Ensembl 52: H.sapiens - Protein summary - Transcript: XRCC1-201 (ENST00000262887) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Transcript/ProteinSummary?db=core;g=ENSG00000073050;r=1 country codes .pt

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Protein s...

Ensembl
Home > Human
Location: 19:48,739,304-48,771,555 Gene: XRCC1 Transcript: XRCC1-201
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Transcript-based displays

- Transcript summary
- Exons (17)
- Supporting evidence (12)
- Sequence
 - cDNA
 - Protein
- External References
 - General identifiers (40)
 - Oligo probes (11)
 - Gene ontology (7)
- Genetic Variation
 - Population comparison
 - Comparison image
- Protein Information
 - Protein summary**
 - Domains & features (17)
 - Variations (37)
- External Data
- ID History
 - Transcript history
 - Protein history

• Configure this page
• Add custom data to page
• Export data
• Bookmark this page

Transcript: XRCC1-201 (ENST00000262887)

DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P18887]

Location [Chromosome 19: 48,739,304-48,771,555](#) reverse strand.

Gene This transcript is a product of gene [ENSG00000073050](#)

Protein [ENSP00000262887](#) is the protein product of this transcript

« Comparison image **Protein summary help** Domains & features »

Statistics

- Ave. residue weight: 109.713
- Charge: -2.5
- Isoelectric point: 6.1872
- Molecular weight: 69,448.26
- Number of residues: 633

Ensembl release 52 - Dec 2008 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)
[Permanent link](#) - [View in archive site](#)

Find: transmembrane Next Previous Highlight all Match case

Done

Variation (SNP) view – from Gene tab

The screenshot displays the Ensembl genome browser interface for the gene **XRCC1** (ENSG00000073050) on chromosome 19. The main view is the **Variation Image**, which shows a genomic region from approximately 48.73 Mb to 48.78 Mb. The variation image consists of several tracks: Variations (Ensembl/Havana g...), ENST00000262887 (XRCC1-201) transcript, PROSITE profiles (P550172 BRCT, PFD0533 BRCT, P550172 BRCT, PFD0533 BRCT, PFD1834 Xrcc1_N), and Pfam domain (PFD1834 Xrcc1_N). The variation image shows numerous SNPs (Single Nucleotide Polymorphisms) across the gene, with various amino acid substitutions indicated by colored boxes (e.g., G, L, PL, S, V, P, P/L, R/H, H, T, VM, R/W, R/Q, SY, Q/R, V/M, T/A, R, KN, R/H, N/Y, H/Y, R, N/S, V/A, E). The left sidebar contains navigation options for the gene, including Gene summary, Splice variants (1), Supporting evidence, Sequence, External references (3), Regulation, Comparative Genomics, Genomic alignments (36), Gene Tree, Orthologues (34), Paralogues (0), Protein families (1), Genetic Variation, Variation Table, Variation Image, External Data, ID History, and Gene history. The bottom of the browser shows the search bar with the text "transmembrane" and the URL: http://www.ensembl.org/Homo_sapiens/Gene/Variation?db=core;g=ENSG00000073050;r=19:48739304-48771555;snp_fake=1;t=ENST00000262887;v=rs25486;vf=19413;vt=ENST00000262887

SNPView (link from protein summary)

View linkage disequilibrium in the population (LDView).

Ensembl 52: H.sapiens - Variation summary - Variation: rs25489 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Variation/Summary?db=core;g=ENSG00000073050;r=19:48736 country codes

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Variation...

Home > Human Login / Register | BLAST/BLAT | FAQs

Location: 19:48,739,304-48,771,555 Gene: XRCC1 Transcript: XRCC1-201 Variation: rs25489

Variation: rs25489

- Summary
- Gene/Transcript (1)
- Population genetics (13)
- Individual genotypes (697)
- Context
- External Data

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Variation: rs25489

Variation type SNP (source [dbSNP](#))

Synonyms None currently in the database

Alleles C/T (Ambiguity code: Y)
Ancestral allele: C

Location [19:48748252](#) (forward strand)

Variation summary Gene/Transcript »

Validation status Proven by **cluster, frequency** (Feature tested & validated by a non-computational method).
HapMap SNP

Linkage disequilibrium data **Links to Linkage disequilibrium data per population:**
[CSHL-HAPMAP:HapMap-CEU](#) [CSHL-HAPMAP:HapMap-YRI](#) (Tag SNP) [PERLEGEN:AFD_EUR_PANEL](#) (Tag SNP)

Flanking Sequence

```
GACACAGGGAGCTGGGGAGAAAAGCTGCAGCGGCCAGGGAGGGGGCGCAAGCCTACATG
AGGTGCGTGTGCCCGGTCCAGTCTGGCCGAT&CTGGCCCCAAGCTCTAGGGCCTTA
TCTCGCAGCTCGGAGCGGAAGGGGTTCTGGAAGCCACTCAGCACCCTACCACACCCCTGA
AGGATCTTCCCCAGCTCCTCTGGGCCAGCTCGGGTCTGGGCTCGGTGCCTTCTCCT
CGGGTTTGCCCTGTC&CTGCCCTGTGCTCGGGCAGGGACTGGGGCTGGGCTGGGGTA
Y&GATTGGAGCTGGC&CTGGAGAAGACA&AG&GT&GATTAGGTTAGCACC&CTGGGGGTC
AACCCCCAGCCCTCCTCCTCAGACCCAGCAGTCCAGGCCCCAGACCCCTCCTCCTCAG
ACCCAGCAATCC&AGGGCC&AGCCCTC&AG&AGGGTCC&AGGGCC&AGCC
CCTCCTCCTCAG&AGCC&AG&GTCC&AGGGCC&AGCCCTCCTCCTCAGGGCC&AG&GT
CC&AGGGCC&AGCCCTCCTCCTCAG&AGCC&AG&GTCC&AGGGCC&AGCCCTCCTCCTC
T
```

Find: transmembrane Next Previous Highlight all Match case

Done

LDView

Ensembl 52: H.sapiens - Linkage Disequilibrium Data - Chromosome 19: 48,739,304-48,771,555 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/Homo_sapiens/Location/LD?db=core;g=ENSG00000073050;opt_pop=CSHL-HA country codes .pt

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl 52: H.sapiens - Linkage ...

Ensembl/Havana g... <XRCC1-201

Known protein coding Ensembl gene

Variations

Genotyped variations

Variation legend

Tagged SNPs

LD (r2)

CSHL-HAPMAP:HapMap-CEU (Europe) 28 SNPs

Find: transmembrane Next Previous Highlight all Match case

Done

Ensembl front page

The screenshot shows the Ensembl Genome Browser front page in a Mozilla Firefox browser window. The browser's address bar displays the URL <http://www.ensembl.org/index.html>. The page features a search bar with the text "Search: All species" and a "Go" button. Below the search bar, there are links for "Browse a Genome" and "Popular genomes" (Human, Mouse, Zebrafish). A "New to Ensembl?" section lists various features like "Add custom tracks", "Upload your own data", "Search for a DNA or protein sequence", "Fetch only the data you want", "Download our databases via FTP", and "Mine Ensembl with BioMart". A "What's New in Release 52 (9 December 2008)" section highlights the "Homo sapiens core database (Human)". A callout box labeled "BioMart" points to the "BioMart" link in the top navigation bar.

Ensembl Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

<http://www.ensembl.org/index.html> ensembl

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... Ensembl Genome Browser Nucl. Acids Res. -- Table of Contents (...)

Ensembl
Home Login / Register | BLAST/BLAT | **BioMart** | Docs & FAQs

Search Ensembl

Search: All species for

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

- Human**
NCBI36
- Mouse**
NCBIM37
- Zebrafish**
ZFISH7

All genomes

Select a species -

Find: Highlight all Match case

New to Ensembl?

Did you know you can:

- [Add custom tracks](#)
using our new Control Panel
- [Upload your own data](#)
and save it to your Ensembl account
- [Search for a DNA or protein sequence](#)
using BLAST or BLAT
- [Fetch only the data you want](#)
from our public database, using the Ensembl Perl API
- [Download our databases via FTP](#)
in FASTA, MySQL and other formats
- [Mine Ensembl with BioMart](#)
and export sequences or tables in text, html, or Excel format

Still got questions? [Try our FAQs](#)

NEW! **The new Ensembl website**
We've made some changes to our site, to make it faster and easier to use.
[Find out more about what we've changed and why!](#)

What's New in Release 52 (9 December 2008)

- Homo sapiens core database** (Human)

BioMart

MartView – select genome

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/biomart/martview/42c49b8c674cda50dddcd2ce2e814d

country codes .pt

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... http://www.ens...dcddb2ce2e814d

e!Ensembl

Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset

[None selected]

Ensembl 52

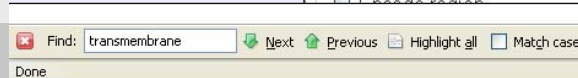
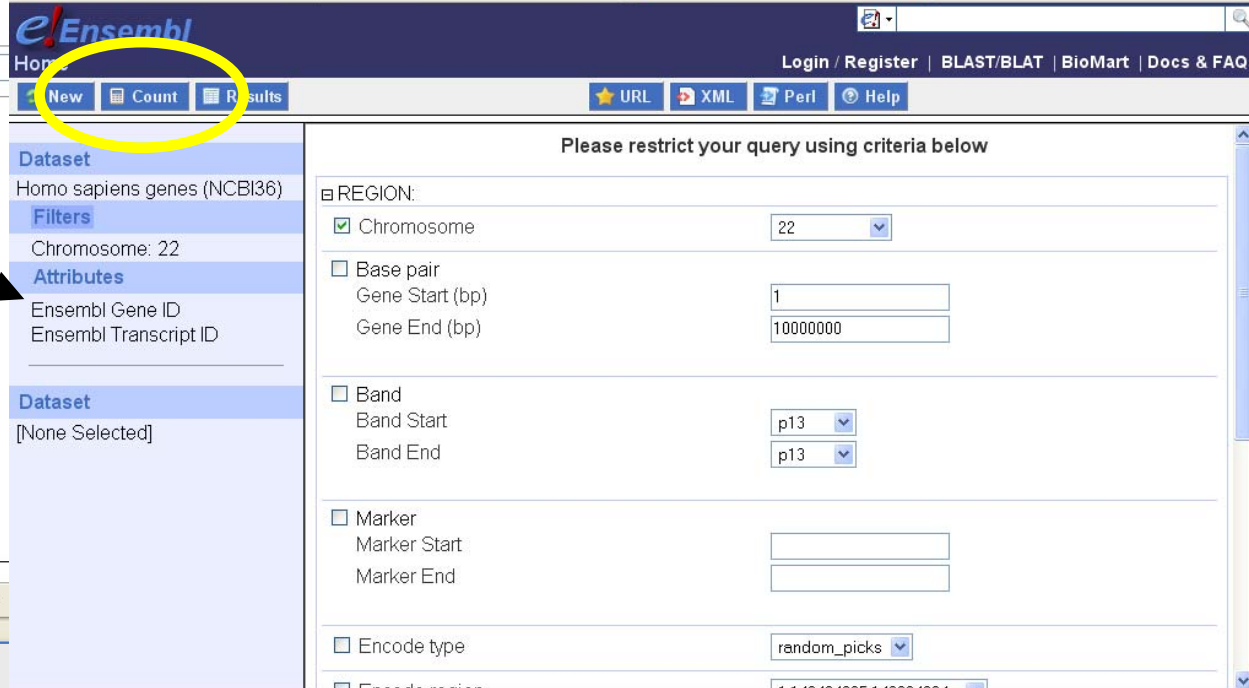
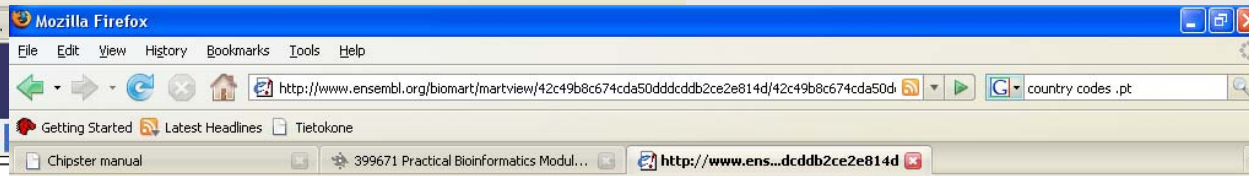
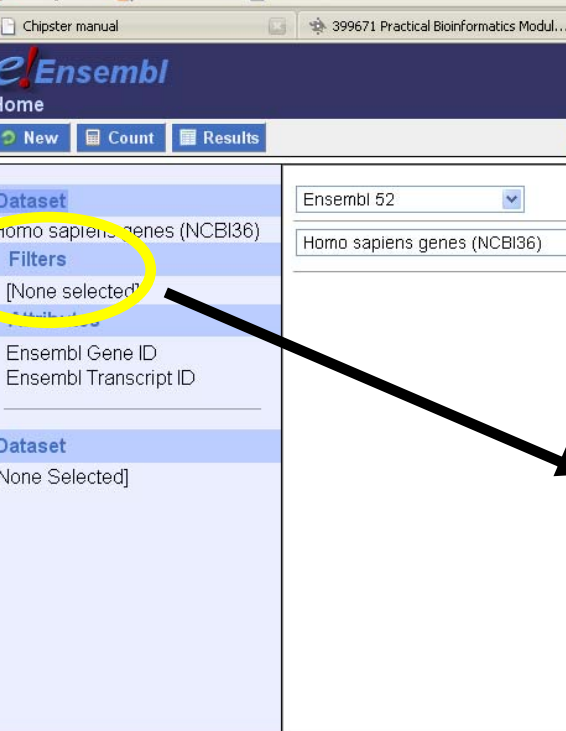
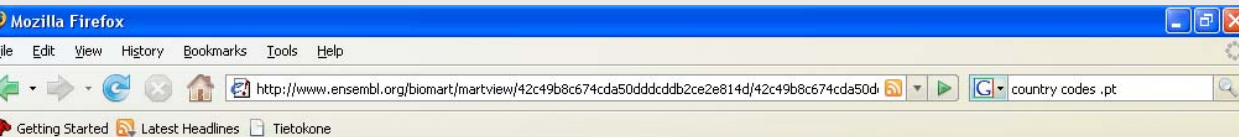
- CHOOSE DATASET -

- CHOOSE DATASET -
- Danio rerio genes (ZFISH7)
- Gallus gallus genes (WASHUC2)
- Homo sapiens genes (NCBI36)**
- Mus musculus genes (NCBIM37)
- Rattus norvegicus genes (RGSC3.4)
-
- Aedes aegypti genes (AaegL1)
- Anopheles gambiae genes (AgamP3)
- Bos taurus genes (Btau_4)
- Caenorhabditis elegans genes (WS190)
- Canis familiaris genes (BROADD2)
- Cavia porcellus genes (cavPor3)
- Ciona intestinalis genes (JGI2)
- Ciona savignyi genes (CSAV2.0)
- Dasyus novemcinctus genes (ARMA)
- Dipodomys ordii genes (KANGAROO_RAT_1)
- Drosophila melanogaster genes (BDGP5.4)
- Echinops telfairi genes (TENREC)
- Equus caballus genes (EquCab2)

Find: transmembrane Next Previous Highlight all Match case

Done

MartView - Filter



MartView - output

The image displays two overlapping screenshots of the Ensembl MartView interface in a Mozilla Firefox browser. The top screenshot shows the initial query configuration page, and the bottom screenshot shows the column selection page.

Top Screenshot (Query Configuration):

- Dataset:** 53 / 37435 Genes
- Filters:** Chromosome: 22, with MIM disease ID(s): Only
- Attributes:** Ensembl Gene ID, Ensembl Transcript ID
- REGION:** Chromosome, Base pair, Gene Start (bp), Gene End (bp)
- Band:** Band, Band Start, Band End
- Marker:** Marker, Marker Start, Marker End
- Encode type:** Encode region
- Buttons:** New, Count, Results

Bottom Screenshot (Column Selection):

- Buttons:** New, Count, Results
- Message:** Please select columns to be included in the output and hit 'Results' when ready
- Feature Selection:** Features, Homologs, Structures, Sequences, Variations
- SEQUENCES:** Sequences (max 1)
- Diagram:** Two gene models showing exons and introns with UTRs.
- Feature List:** 5' UTR, 3' UTR, Exon sequences, cDNA sequences, Coding sequence, Protein
- Upstream flank:** Upstream flank [1000]
- Downstream flank:** Downstream flank

Yellow circles highlight the 'New', 'Count', and 'Results' buttons in both screenshots. Arrows point from the 'Attributes' section in the top screenshot to the 'Attributes' section in the bottom screenshot.

MartView

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ensembl.org/biomart/martview/42c49b8c674cda50ddcdcb2ce2e814d/42c49b8c674cda50d country codes .pt

Getting Started Latest Headlines Tietokone

Chipster manual 399671 Practical Bioinformatics Modul... http://www.ens...dcddb2ce2e814d

Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset 53 / 37435 Genes
Homo sapiens genes (NCBI36)

Filters
Chromosome: 22
with MIM disease ID(s): Only

Attributes
Ensembl Gene ID
Ensembl Transcript ID
5' UTR
Upstream flank [1000]

Dataset
[None Selected]

Export all results to File FASTA Unique results only Go

Email notification to

View 10 rows as FASTA Unique results only

```
> ENSG00000099940 | ENST00000215730  
TAACGCCCCAGCCTCGTAGAATAACACTGCGCAGTTGGTITACAGTCACAAAGCACTTT  
CAAGACAGTGTTCCTCTCCTGGTACCACAGTCCGGTGAAGCCGGCAAAGCAGATAAAA  
GCATCCCCATTTTACAAGAAAGCCATCTTCCCAAGGTCGCGGAATAAAAAGCAGGTGAAA  
TCGGAAAGCAGAACCCAGGTCTTGACCCTGGTCCTCGGCTCATCCACTGCTCCGCACA  
GCCGCCCAGATCCAGCCTCCACCTCCGCGATGTACAAGATGCTCCCTGGGCCAGCAGC  
TCACATCCCCACCCCATCCACAACCACCACCTGCAGCGGCCCTGCAGGCCCCGCCG  
ACCCTTCCCCCACGGTCCCTCCTCTTCAACCAGCCCTCCCCTAGGGCTCCTTCCCC  
GCTAGATCCCTCCGCTCAGGCGCCCTCCCGCACAGGCTGCCGGCAGCCAGCCTCGG  
CCTCCACAGACCCTCGTCCCGCTCCACGCTTCGTACCCAGCCCTCCTCTGTACG  
CCCGCGCCAGCCCGCCGACGTTACCTTCTCCAAGGATGCTGTTCTGTGACCCGCCAGG  
GAGCGGGCCAGTGACAGGACCCTGTGAAATAGAAGCCCGGAGGCGCTGGAGCCGGAG  
CCGGAGCAGCCCGCCGCCCTCCGCCCTCCGCTCCCGGGCCGGGGCCCGCCGC  
ATCACCCTACGAGCCCGCGCTGCCCGCGGCTCCCCTCCCTGCCCCTGCAGCGCCG  
ACCTCAGGGCCGAGGCGTAGTGCATCCGGCTTCCCGCACAGTGCACCTCTCAGTTGCTG  
ACACTAGATAGGGCGTGCTGCCGGCGGGCCACCCTGCAGGGGAGAGAGGGGATGGA  
CAGTAGGCTGCCGTTGCCGCGCCGGATGTGGGCCCCCCCAGCCGACCTGCCGGATGGT  
CGCTCGCAGACCGGCGCGCCAGCTGCGCGCGAGAGGCGCGCAGAGGCGCGCAGCGCGC
```

Find: transmembrane Next Previous Highlight all Match case

Done

SNP databases

SNPs and disease research

- **One might be interested in studying how certain SNPs are associated with, say, length of the nose.**
- **In a typical setting we collect a number of individuals, make an interview and collect background information (age, sex, parity, etc.), and genotype the individuals for (certain) SNPs, and look for correlations (used here as in common language) between SNPs, background data and length of the nose.**
- **These data are typically best preserved in a database (inside the lab).**

SNPs in databases I

➤ **dbSNP**

- Contains SNPs, microsatellites (ACACAC), and other small polymorphisms
- >10 million SNPs for human, 4.5 million validated
- Mouse, chicken, dog, maize, chimp are other dominant species in the DB
- Most of the data is not in Genbank, but is cross-referenced to it.

SNPs in databases II

➤ **HGVbase**

- Last update in 2003, but contains good quality data

➤ **HapMap data in Ensembl**

- Validated SNP data from three populations (269/270 individuals)

➤ **SNP500CANCER**

- Validated SNP data from three populations (102 individuals)

SNPs at NCBI I

The image shows a screenshot of the NCBI homepage in Microsoft Internet Explorer. The browser's address bar shows the URL <http://www.ncbi.nlm.nih.gov/>. The page features a navigation menu with links to PubMed, All Databases, BLAST, OMIM, Books, TaxBrowser, and Structure. A search bar is present with the text "SNP" entered and a dropdown menu set to "for XRCCT". A "Go" button is next to the search bar. Two callout boxes are overlaid on the page: one pointing to the search bar with the text "Select 'SNP' database!" and another pointing to the search input field with the text "Free text query goes to here." The main content area includes a "What does NCBI do?" section, a "100 Gigabases" announcement, and an "NCBI Bookshelf" section. A sidebar on the left contains links for "SITE MAP", "About NCBI", "GenBank", "Literature databases", "Molecular databases", and "Genomic biology".

NCBI HomePage - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/>

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search SNP for XRCCT Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

100 Gigabases

GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the [press release](#) or find more information on [GenBank](#).

NCBI Bookshelf

Five new medical textbooks are available on the NCBI Bookshelf. Topics include epilepsy, Parkinson's disease, alternative medicine, rehabilitation, and spinal cord medicine. Search these books and many more in the Entrez Books database [here](#).

Assembly Archive

Clusters of orthologous groups

Coffee Break, Genes & Disease, NCBI Handbook

Electronic PCR

Entrez Home

Entrez Tools

Gene expression omnibus (GEO)

Human genome resources

Malaria genetics & genomics

Map Viewer

dbMHC

Select "SNP" database!

Free text query goes to here.

Internet

SNPs at NCBI II

The screenshot shows the NCBI Entrez SNP search results for the query 'XRCC1'. The browser title is 'Single Nucleotide Polymorphism (SNP) - Microsoft Internet Explorer'. The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp&cmd=search&term=XRCC1>. The search results are displayed in a table with columns for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'Taxonomy', and 'SNP'. The search criteria are 'SNP for XRCC1'. The results are sorted by 'Graphic Summary' and show 372 items. The first three items are highlighted:

- 1: [rs8147612](#) [*Rattus norvegicus*]
CACAGGTGTACATTCAGGCCTGGGG[C/G]ACCACCCCATAGAGCTGGTGGGGGA
MapView GeneView SeqView No 3D No OMM
- 2: [rs13470740](#) [*Mus musculus*]
tatagccctggctgtcctctagaactc[A/G]ctctgtagaccaggctggcctcgaa
CGAP-GAI
- 3: [rs13467177](#) [*Mus musculus*]
ctcactctgtagaccaggctggcct[C/T]gaactaagaaatccacctgcctgtg
CGAP-GAI

Accession number for the SNP, and species.

Location of the SNP in the sequence.

Links to "alternative" views.

Biological pathways

Pathway databases

➤ **Reactome**

- Curated
- Pathways and reactions

➤ **KEGG**

- Curated
- Manually drawn pathway maps for molecular interactions and reactions
- Used extensively

➤ **Both contain data for several species**

Pathway databases

- **cMAP**
 - Resembles KEGG
- **GO**
 - Gene ontologies
- **MINT**
 - Protein-protein interactions

GO

Gene Ontology (GO)

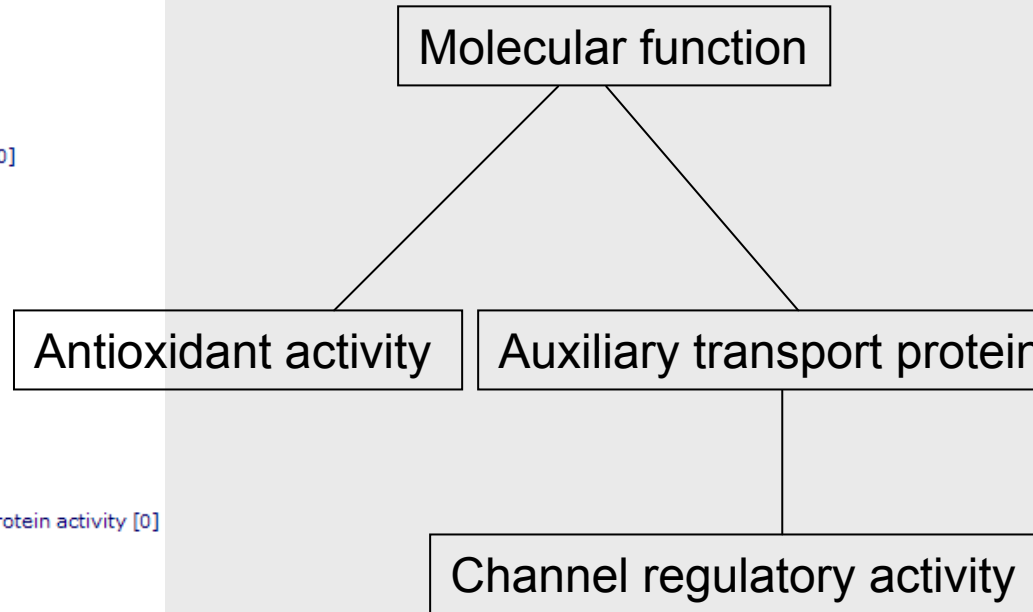
- **A controlled vocabulary for describing gene product**
- **The same term always describes the same entity**
 - V. cell (battery, prison, part of a table, ...)
- **Go annotation describe activities and localizations of gene products**
 - Evidence codes!

GO

- **GO is a hierarchy (directed acyclic graph)**
- **AmiGO**
- **Three related ontologies**
 - Biological process
 - Cellular component
 - Molecular function

GO

- all: all [36690]
- GO:0008150 : biological_process [29655]
- GO:0005575 : cellular_component [28254]
- GO:0003674 : **molecular_function** [30848]
- GO:0016209 : **antioxidant activity** [95]
- GO:0045174 : glutathione dehydrogenase (ascorbate) activity [0]
- GO:0004362 : glutathione-disulfide reductase activity [13]
- GO:0004601 : peroxidase activity [62]
- GO:0004791 : thioredoxin-disulfide reductase activity [9]
- GO:0015457 : **auxiliary transport protein activity** [43]
- GO:0016249 : channel localizer activity [0]
- GO:0016247 : **channel regulator activity** [38]
- GO:0005246 : calcium channel regulator activity [8]
- GO:0016248 : channel inhibitor activity [15]
- GO:0017081 : chloride channel regulator activity [0]
- GO:0015459 : potassium channel regulator activity [24]
- GO:0017080 : sodium channel regulator activity [3]
- GO:0015458 : low voltage-gated potassium channel auxiliary protein activity [0]
- GO:0016974 : sodium channel auxiliary protein activity [2]
- GO:0015460 : transport accessory protein activity [1]
- GO:0005488 : binding [12056]
- GO:0003824 : catalytic activity [10487]
- GO:0030188 : chaperone regulator activity [12]
- GO:0042056 : chemoattractant activity [6]
- GO:0045499 : chemorepellant activity [1]
- GO:0031992 : energy transducer activity [0]
- GO:0030234 : enzyme regulator activity [1018]
- GO:0003774 : motor activity [249]
- GO:0045735 : nutrient reservoir activity [5]
- GO:0031386 : protein tag [8]
- GO:0004871 : signal transducer activity [1733]
- GO:0005198 : structural molecule activity [1345]
- GO:0000332 : template for synthesis of G-rich strand of telomere DNA activity [0]
- GO:0030528 : transcription regulator activity [2106]
- GO:0045182 : translation regulator activity [293]
- GO:0005215 : transporter activity [2099]
- GO:0030533 : triplet codon-amino acid adaptor activity [251]



KEGG

KEGG

- **Kyoto encyclopedia of genes and genomes.**
- **Established in 1995**
- **Curated**
- **Pathways and reactions (in the pathway database) – enzymes!**

KEGG

Type I diabetes mellitus - Homo sapiens (human) - Windows Internet Explorer

http://www.genome.jp/dbget-bin/show_pathway?hsa04940+3630

AmiGO: Gene Product Search... GNF SymAtlas Type I diabetes mellitus - ...

KEGG Type I diabetes mellitus - Homo sapiens (human)

[Pathway menu | Reference list]

Homo sapiens (human) Go Current selection Select

TYPE I DIABETES MELLITUS

Environmental factors
virus
milk

Pancreatic islet

Autoantigen?
INS
GAD
IA-2
CPE
HSP
ICA

β -cell

Antigen exposure

Pancreatic lymph node

Antigen processing

Antigen-presenting cell

MHC-II
CD80/86

Antigen

TCR
CD28

T cell receptor signaling pathway

IL-12

CD4+ T helper 1 cell

IL-2

IFN γ

MHC-I
TCR
FasL

Antigen

Fas

Cytotoxic CD8+ T cell

PRF1
GZM
IFN γ
TNF β
TNF α
IL-1

Apoptosis

β -cell death

Right-click and save as an image.

Structural databases

PDB and MSD

- **PDB contains structures of biological macromolecules.**
 - Mainly proteins, but also DNA and RNA structures
- **MSD is also a collection of biological structures, but it extends the PDB data format, and circumvents some problems.**

MSD 1/5

The screenshot shows the EMBL-EBI Macromolecular Structure Database website. The browser window title is "The Macromolecular Structure Database - Microsoft Internet Explorer". The address bar shows "http://www.ebi.ac.uk/msd/". The page header includes the EMBL-EBI logo and navigation tabs: EBI Home, About EBI, Groups, Services, Toolbox, Databases, Downloads, and Submissions. The main content area is titled "THE MACROMOLECULAR STRUCTURE DATABASE" and contains several sections:

- MSD Home**
- MSD Services**
- Resources**
- Documentation**
- Submission**
- FTP/PDB Access**
- MSD SITEMAP**

The **Services** section is highlighted with a red arrow and contains the following links:

- MSDlite
- MSDpro
- MSDmotif
- MSDtemplate
- MSDpisa
- MSDchem
- MSDmine
- MSDsite
- MSDfold
- MSDtarget
- MSDanalysis
- EMsearch
- MSDbar
- PQS
- PQS-Quick
- NMR Representatv.
- Unpub Ref Srv
- Relibase
- Biotech
- Search OCA
- PDB Pending
- PDB New Entries

The **Resources** section contains:

- SIFTS @ EBI
- SPINE @ EBI
- RECOORD @ EBI
- NMR @ EBI
- Funding
- Education
- Projects
- CAPRI
- Publications
- About MSD
- Past Workshops
- Future Events
- News

The **Documentation** section contains:

- MSDSD guide
- MSDSD schema
- Software
- API
- LIMS
- mmcIF
- XML
- PDB
- DB_io_tools
- 3DEM Conventions

The **Submission** section contains:

- PDB AutoDep
- EMDep

The right sidebar features three widgets:

- Get PDB by ID**: A search box for the PDB Protein Data Bank with a "go" button.
- EM Search**: A search box for the EM Data Bank with a "go" button.
- MSD Roadshow**: A graphic with the letters M, S, and D, and a "Details" link below it.

At the bottom of the page, there is a "Welcome" message: "Welcome to the EBI Macromolecular Structure Database - the European project for the collection," and a "Done" status bar.

MSD 2/5

MSDlite - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ebi.ac.uk/msd-srv/msd-lite/index.html>

home > services > MSDlite

Macromolecular Structure Database

MSDlite Search System

contact MSD

Main

ID code: PDB ID

Author Last Name:

Associated small molecule: Molecule Name

Experiment type:

- X-ray
- Theoretical model
- N.M.R.
- Fibre diffraction
- Infrared spectroscopy
- Powder diffraction (X-ray)
- Other method
- Electron microscopy
- Electron diffraction
- Electron tomography
- Fluorescence transfer
- Neutron diffraction
- Solid state NMR

Text Search:

Keyword:

Results

- PDB title
- Entry Author
- Assembly Type
- Assembly Status
- Resolution
- Interpro ID
- Pfam ID
- GO ID
- EC-NUMBER
- SCOP ID
- CATH ID
- UniProt acc num
- UniProt ID
- Experiment type
- List of HET groups
- Fasta E-values

Results per page:

New search

Extras

Resolution: high: low:

Representative set: PDB SCOP

Done Internet

MSD 3/5

The screenshot shows a web browser window with the address bar displaying <http://www.ebi.ac.uk/msd-srv/msd-lite/apps/query>. The page title is "Macromolecular Structure Database" and the URL path is "home > services > MSDlite". The main heading is "MSDlite Search System".

Navigation and search options include "Back", "Search", "Favorites", and "Go". A search bar is present with the text "Found 7 hits (1 pages). Showing hits 1 to 7." Below this, there are links for "Start a new search" and "Refine the results of this search".

On the right side, there is a "Download all results:" section with icons for "XML" and "Text".

PDB Entry ID	PDB Entry Title	Resolution	Experiment Type
1cdz View	BRCT DOMAIN FROM DNA-REPAIR PROTEIN XRCC1	3.20Å	X-ray
1gqp View	APC10-DOC1 SUBUNIT OF S. CEREVISIAE	2.20Å	X-ray
1imo View	NMR STRUCTURE OF HUMAN DNA LIGASE IIIALPHA BRCT DOMAIN	Not Available	NMR
1in1 View	NMR STRUCTURE OF HUMAN DNA LIGASE IIIALPHA BRCT DOMAIN	Not Available	NMR
1xna View	NMR SOLUTION STRUCTURE OF THE SINGLE-STRAND BREAK REPAIR PROTEIN XRCC1-N-TERMINAL DOMAIN	Not Available	NMR
1xnt View	NMR SOLUTION STRUCTURE OF THE SINGLE-STRAND BREAK REPAIR PROTEIN XRCC1-N-TERMINAL DOMAIN	Not Available	NMR
2brf View	CRYSTAL STRUCTURE OF THE FHA DOMAIN OF HUMAN POLYNUCLEOTIDE KINASE 3' PHOSPHATASE	1.40Å	X-ray

primary developers: John Tate & Sameer Velankar
last modified: 28/03/03

Done Internet

MSD 4/5

AstexViewer(TM)@MSD-EBI

Tree View

Structure

Reset View

Show

Zoom Out

Solid Model

Protein

Ligand

Colour

Similar

Graphs

Select

Brushing


Magic Lens

Sequence

Seq.Style

Seq.Col

Seq.Show



astex TECHNOLOGY

10 20 30 40

lcdz_ 0 ELPDFFQCKHFFLYGEFPGDERRKLRVIA


Java Applet Window

MSD 5/5

MSD - entry 1cdz - Microsoft Internet Explorer


File Edit View Favorites Tools Help

Address <http://www.ebi.ac.uk/msd-srv/msd-lite/atlas/summary/1cdz.html>




Macromolecular Structure Database

Entry 1cdz



home > services > atlas pages Latest changes



Title: BRCT DOMAIN FROM DNA-REPAIR PROTEIN XRCC1

Authors: ZHANG, X., MORERA, S., BATES, P., WHITEHEAD, P., COFFER, A., HAINBUCHER, K., NASH, R., STERNBERG, M., LINDAHL, T., FREEMONT, P.

Source: Homo sapiens, expressed in Escherichia coli

Method: X-ray (resolution 3.20Å, R-factor 22.4%)

[PDB archive file](#)

[PDB header](#)

Download

- Summary
- Assembly
- Sequence
- Citation
- Similarity
- Visualisation

View entry:

Summary

Source: Homo sapiens (HUMAN) , tax.id. [9606](#)

Expression system: Escherichia coli, tax.id. [562](#)

Primary Citation: X Zhang, S Morera, PA Bates, PC Whitehead, AI Coffe, K Hainbucher, RA Nash, MJ Sternberg, T Lindahl, PS Freemont: Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. *The EMBO journal*. (1998) 17, pp. 6404-11 [PubMed entry [9799248](#)]

Release Date: 28-Feb-2000 (**deposition date** 4-Mar-1999)

Resolution: 3.20Å

R-factors: working 22.4%, free 26.6%

Space group: P 31 2 1

Unit cell:		
a: 100.80Å	b: 100.80Å	c: 72.50Å
alpha: 90.0°	beta: 90.0°	gamma: 120.0°

Chains:	Chain Id	Name	UniProt Entry	Residues
	A	DNA-REPAIR PROTEIN XRCC1	XRCC1_HUMAN (P18887)	96

Oligomeric State: Assembly 1 : [Homo-dimeric](#)

Primary developers: Joël Fillon & John Tate

<http://www.ebi.ac.uk/msd/index.html> Internet

Integrating databases

Why integration?

- **Data is distributed to several sources**
 - That can prevent efficient access to data
- **Genomics**
 - Study of whole genomes, knowledge of gene content, expression etc. needed
- **To get a better view to cells**
 - Systems biology
 - Reductionism doesn't work by itself anymore, we need integration of knowledge
One PhD student, one gene ;(
 - Add protein studies, metabolomics, etc.

Hierarchy of databases - an illustrative example

Nucleotide

Protein

Genbank/EMBL/DDBJ

dbSNP

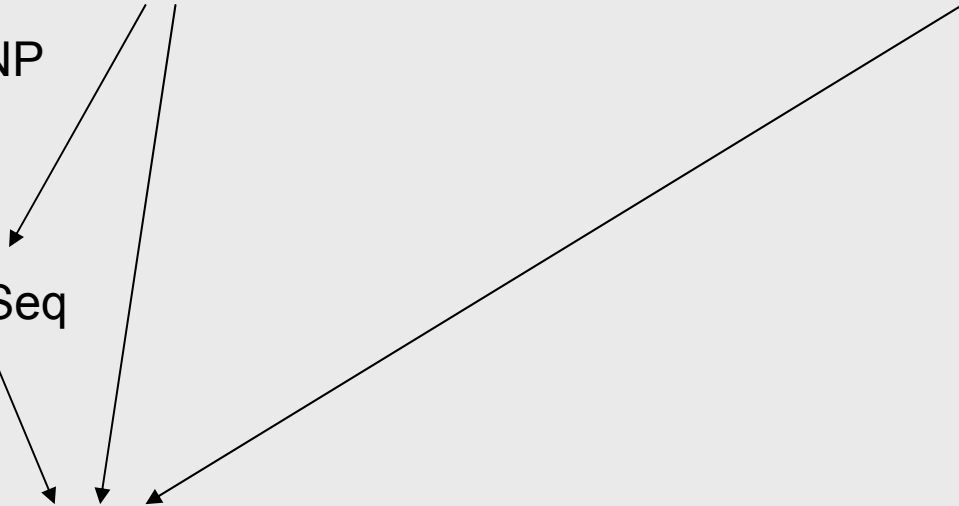
RefSeq

Ensembl

UniProt

} Primary

} Secondary



About accession numbers

- Every sequence entry is individually labeled with an accession number. E.g., from Genbank you can always retrieve the same sequence, if you know the accession number.
- Accession number: alpha-numeric code
- ID: human readable sequence name
- Some examples:

XRCC1	HUGO ID
M36089	EMBL accession number
P18887	UniProt accession number
NM_006297	RefSeq, nucleotide sequence
NP_006388	RefSeq, protein sequence
Hs.98493	UniGene ID
ENSG00000073050	Ensembl, gene sequence
ENSO00000262887	Ensembl, protein sequence
7515	Locuslink ID, Entrez Gene GeneID

Problems in integration

➤ **Integration can't be based on accession numbers**

- Every database uses a different system

➤ **Integration can't be based on sequences**

- Sequence is not necessarily unique

ACGT is a substring of ACGTACGTA and ACGTGGTATTGCTAG, so
which gene does it actually represent?

➤ **What about common terms (you wish!)**

Problems in semantic integration

➤ **Differences in terminology**

- Vector

A line with a direction (math.)

Carrier of an infectious agent (biol., med.)

Virus or DNA molecule used for transferring genetic material to or from cells (biol.)

Breakfast cereal manufactured by Kellogg (food)

A rock band (music)

Ghost town (Final Fantasy VI)

Solutions to terminology

➤ **Controlled vocabularies**

- A set list of terms that are used to describe certain elements
- GO ontology: hierarchical ontology of gene functions, cellular localizations, etc.
- eVOC ontology: describe elements of humans

➤ **Ontologies**

- Knowledge representation systems
- Use richer semantic terms to describe relationships between elements



eVOC

Main - BrowseEvoc - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://www.evocontology.org/site/Main/BrowseEvoc> Go Links

 **eVOContology.org** 

Home

Browse eVOC

Download eVOC Ontologies

Download eVOC Data Mappings

Download eVOC Mouse Ontologies

Software

Publications

Citing eVOC

Contributors

Contact Us

eVOC Browser

Select Ontology: Anatomical System

Select Data Type: Synonyms Only

[Cancel Query](#)
No term selected.

- Anatomical System
 - alimentary system
 - anatomical site
 - abdomen
 - adipose tissue
 - head and neck
 - lower limb
 - mucosa
 - pelvis
 - perineum
 - thorax
 - upper limb
 - whole body
 - cardiovascular system
 - dermal system
 - developmental anatomy
 - endocrine system
 - hematological system
 - lymphoreticular system

Done Internet

Gene Ontology (GO)

AmiGO! Your friend in the Gene Ontology. - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Stop

Address http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=plus_node&depth=1&search_constraint=terms&query=GO:0043226&session_id=317b1141910351 Go Links

AmiGO

Search GO

Exact Match
 Terms
 Gene Symbol/Name

[Advanced Query](#)
[Query By Sequence](#)

Gene Product Filters

Species

All
A. japonica
A. niger

Datasource

All
FlyBase
SGD

Evidence Code

All Curator Approved
IC
IMP

[XML](#)
[Flat File](#)
[Permalink](#)

Graphical View

- all : all (167902)
 - GO:0008150 : biological_process (119520)
 - GO:0005575 : cellular_component (105210)**
 - GO:0005623 : cell (76007)
 - GO:0008372 : cellular component unknown (25561)
 - GO:0031975 : envelope (2134)
 - GO:0031012 : extracellular matrix (431)
 - GO:0005576 : extracellular region (4351)
 - GO:0031974 : membrane-enclosed lumen (3523)
 - GO:0043226 : organelle (55561)**
 - GO:0043230 : extracellular organelle (0)
 - GO:0043229 : intracellular organelle (55545)
 - GO:0043227 : membrane-bound organelle (51642)
 - GO:0043228 : non-membrane-bound organelle (6602)
 - GO:0043233 : organelle lumen (3523)
 - GO:0031982 : vesicle (7871)
 - GO:0043234 : protein complex (11233)
 - GO:0045202 : synapse (181)
 - GO:0019012 : virion (186)
 - GO:0003674 : molecular_function (115755)
 - obsolete_biological_process : obsolete_biological_process (0)
 - obsolete_cellular_component : obsolete_cellular_component (0)
 - obsolete_molecular_function : obsolete_molecular_function (0)

