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?

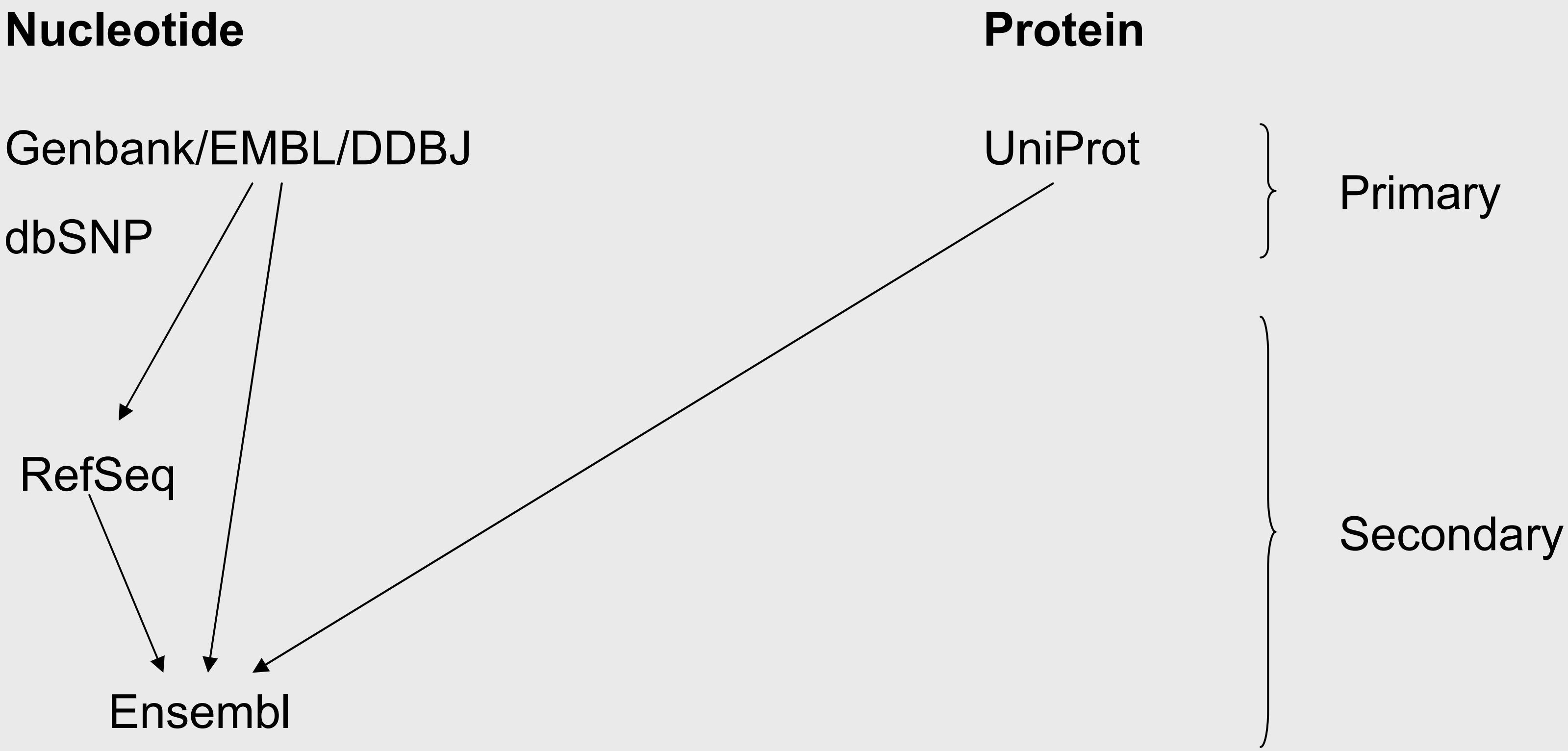
- **Quality**
  - Manual (slow)
  - No overlap between entries
  - Reliable
  - Some data might be missing

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

Main sequence databases

- **DNA:** **ACGGGCTATGTAGTGCTAGC**
    - EMBL / Genbank / DDBJ
    - RefSeq
- **Protein:** **YTCFSATFCFSAGDJSGAJGD**
    - UniProt / SWISS-PROT
    - RefSeq
- **Genomes:** **…**
    - Ensembl
    - UCSC Genome Browser

Hierarchy of databases - an illustrative example



Nucleotide

Protein

Genbank/EMBL/DDBJ

dbSNP

UniProt

Primary

RefSeq

Ensembl

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:

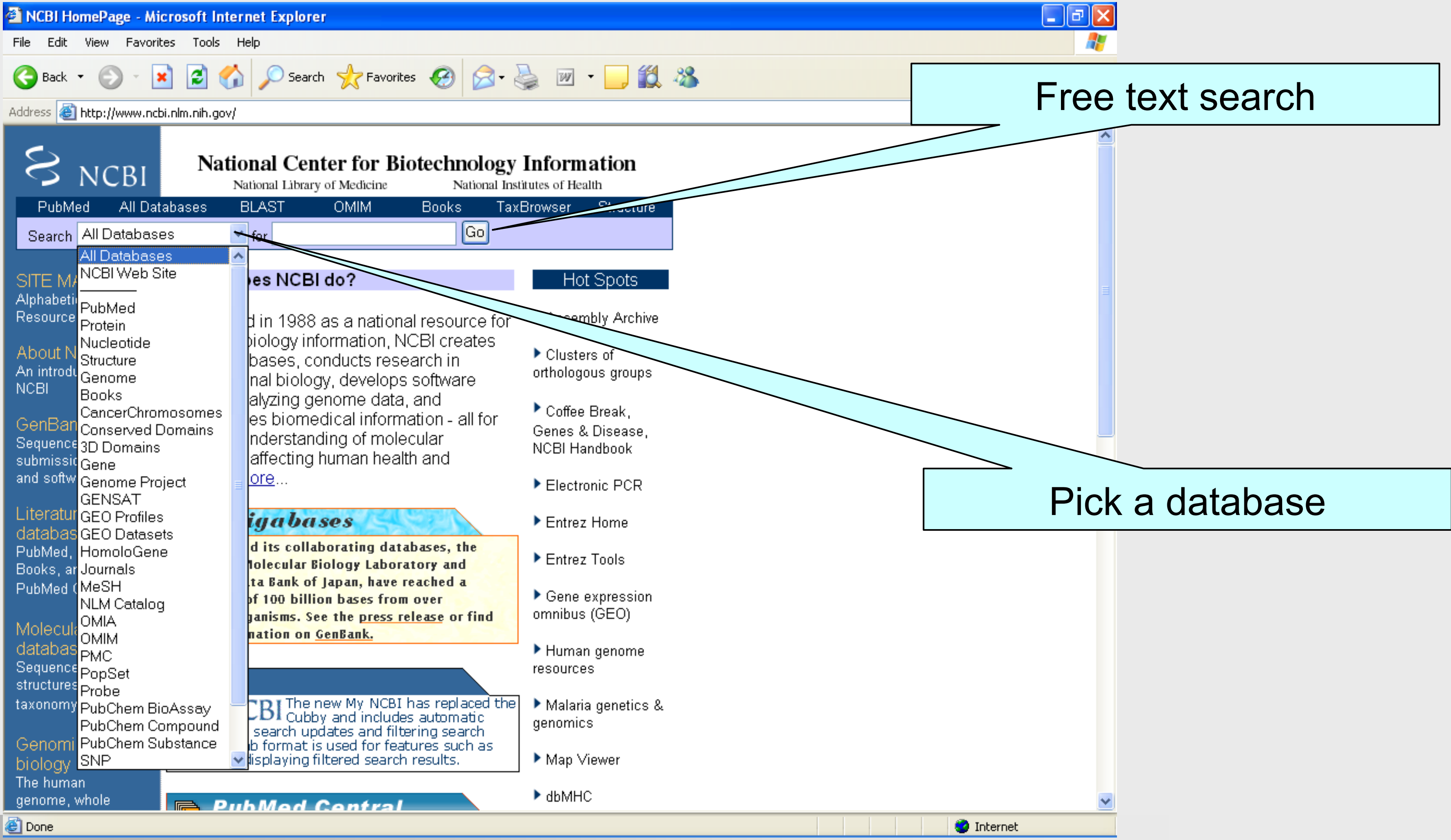
XRCC1	HUGO ID
M36089	EMBL accession number
P18887	UniProt accession number
NM_006297	RefSeq, nucleotide sequence
NP_006388	RefSeq, protein sequence
Hs.98493	UniGene ID
ENSG00000073050	Ensembl, gene sequence
ENSO00000262887	Ensembl, protein sequence
7515	Locuslink ID, Entrez Gene GeneID

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

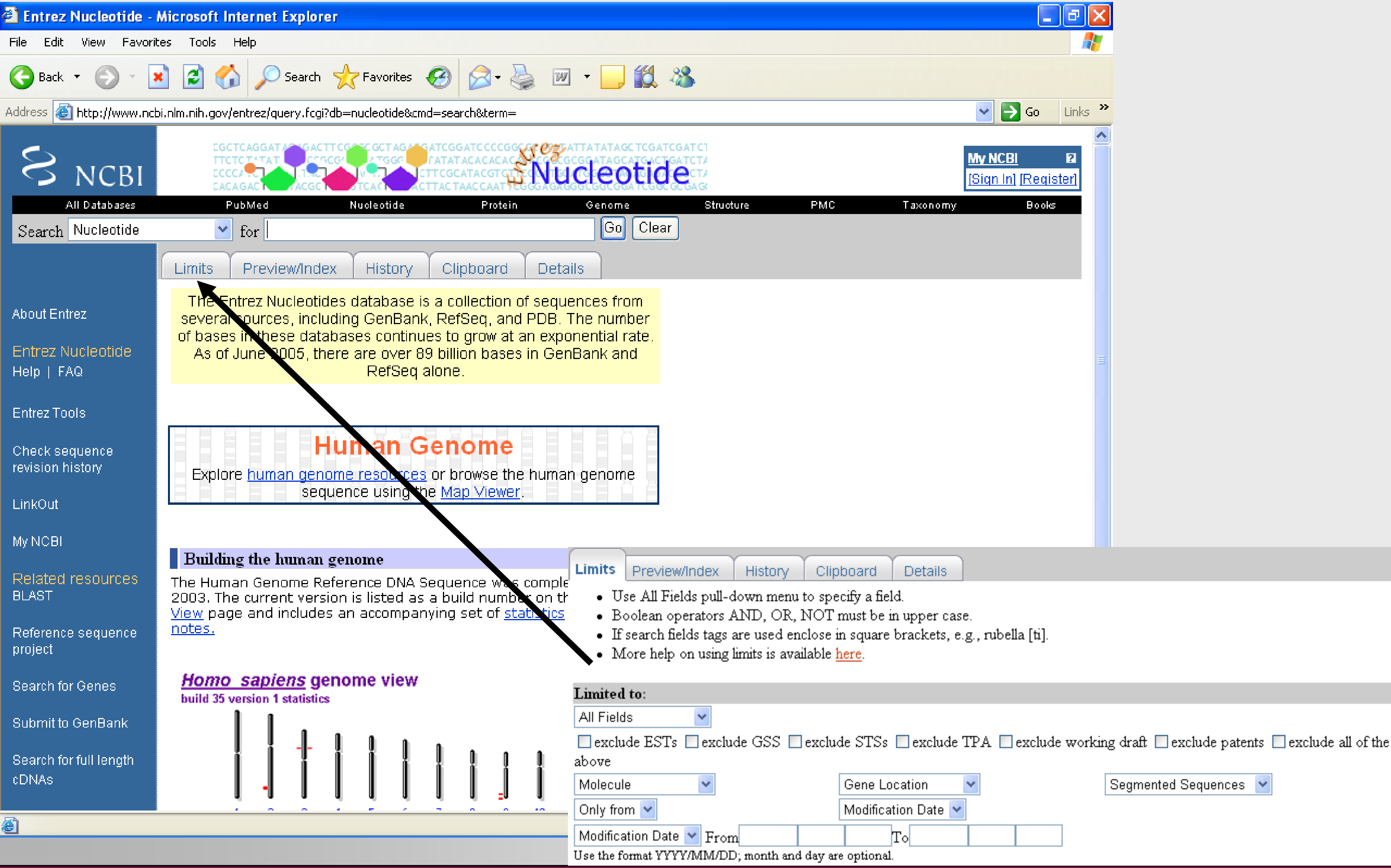
Entrez – main page



Free text search

Pick a database

Entrez – use limits for filtering



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

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 comple 2003. The current version is listed as a build number on th View page and includes an accompanying set of statistics notes.

***Homo sapiens* genome view**

build 35 version 1 statistics

- 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

☐ exclude ESTs ☐ exclude GSS ☐ exclude STSs ☐ exclude TPA ☐ exclude working draft ☐ exclude patents ☐ exclude all of the above

Molecule | Gene Location | Segmented Sequences

Only from | Modification Date

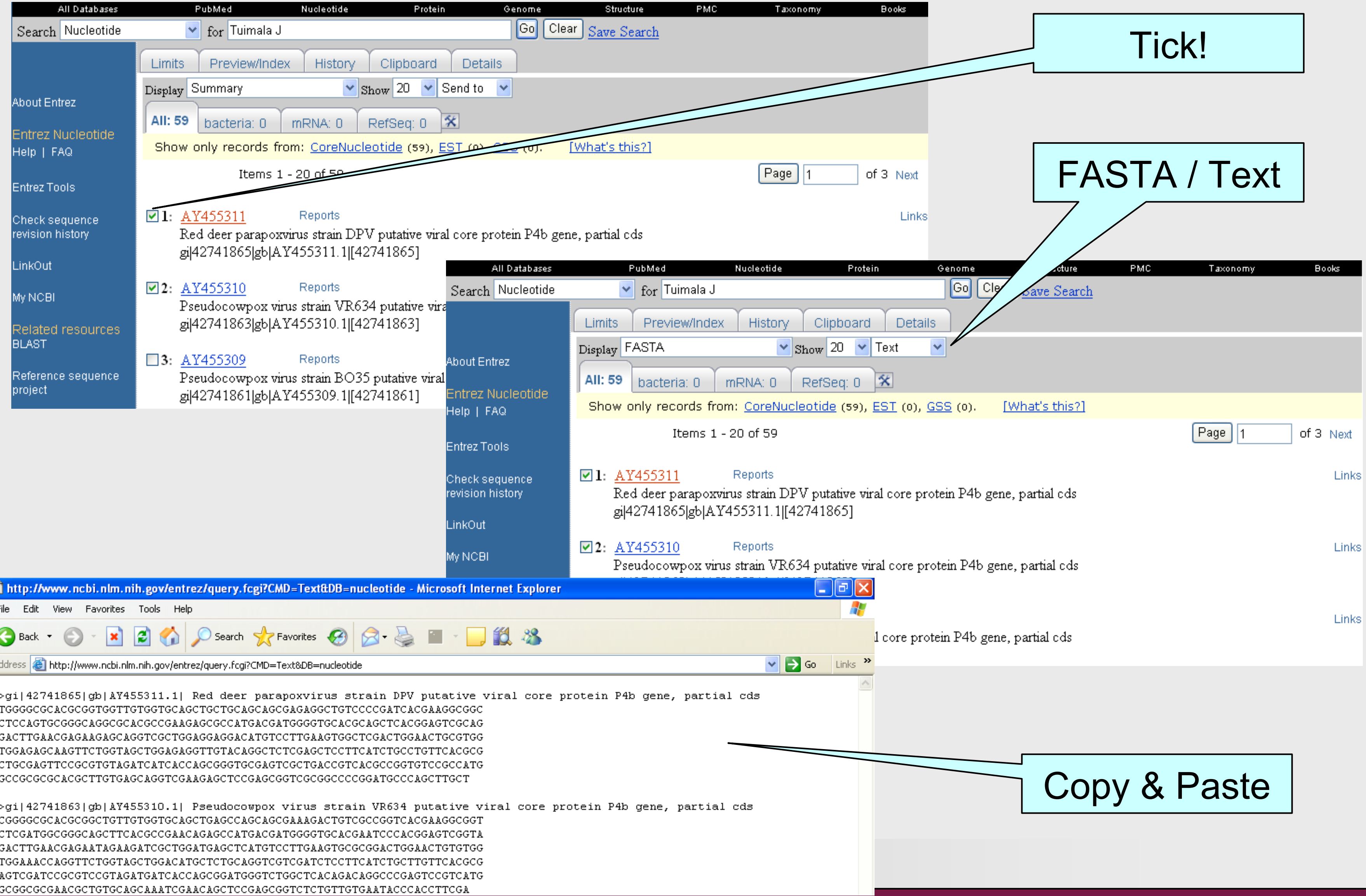
Modification Date From ___ To ___

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



Search Nucleotide for Tuimala J

Display Summary Show 20 Send to

All: 59 bacteria: 0 mRNA: 0 RefSeq: 0

Show only records from: CoreNucleotide (59), EST (0), GSS (0). [What's this?]

Items 1 - 20 of 59 Page 1 of 3

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

☑ 2: AY455310
Pseudocowpox virus strain VR634 putative viral core protein P4b gene, partial cds
gi|42741863|gb|AY455310.1|[42741863]

☐ 3: AY455309
Pseudocowpox virus strain BO35 putative viral
gi|42741861|gb|AY455309.1|[42741861]

Search Nucleotide for Tuimala J

Display FASTA Show 20 Text

All: 59 bacteria: 0 mRNA: 0 RefSeq: 0

Show only records from: CoreNucleotide (59), EST (0), GSS (0). [What's this?]

Items 1 - 20 of 59 Page 1 of 3

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

☑ 2: AY455310
Pseudocowpox virus strain VR634 putative viral core protein P4b gene, partial cds

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Text&DB=nucleotide - Microsoft Internet Explorer

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Text&DB=nucleotide

```
>gi|42741865|gb|AY455311.1| Red deer parapoxvirus strain DPV putative viral core protein P4b gene, partial cds
TGGGGCGCACGCGGTGGTTGTGGTGCAGCTGCTGCAGCAGCGAGAGGCTGTCCCCGATCACGAAGGCGGC
CTCCAGTGCGGGCAGGCGCACGCCGAAGAGCGCCATGACGATGGGGTGCACGCAGCTCACGGAGTCGCAG
GACTTGAACGAGAAGAGCAGGTCGCTGGAGGAGGACATGTCCTTGAAGTGGCTCGACTGGAACTGCGTGG
TGGAGAGCAAGTTCTGGTAGCTGGAGAGGTTGTACAGGCTCTCGAGCTCCTTCATCTGCCTGTTCACGCG
CTGCGAGTTCCGCGTGTAGATCATCACCAGCGGGTGCGAGTCGCTGACCGTCACGCCGGTGTCCGCCATG
GCCGCGCGCACGCTTGTGAGCAGGTCGAAGAGCTCCGAGCGGTCGCGGCCCCGGATGCCCAGCTTGCT

>gi|42741863|gb|AY455310.1| Pseudocowpox virus strain VR634 putative viral core protein P4b gene, partial cds
CGGGGCGCACGCGGCTGTTGTGGTGCAGCTGAGCCAGCAGCGAAAGACTGTCGCCGGTCACGAAGGCGGT
CTCGATGGCGGGCAGCTTCACGCCGAACAGAGCCATGACGATGGGGTGCACGAATCCCACGGAGTCGGTA
GACTTGAACGAGAATAGAAGATCGCTGGATGAGCTCATGTCCTTGAAGTGCGCGGACTGGAACTGTGTGG
TGGAAACCAGGTTCTGGTAGCTGGACATGCTCTGCAGGTCGTCGATCTCCTTCATCTGCTTGTTCACGCG
AGTCGATCCGCGTCCGTAGATGATCACCAGCGGATGGGTCTGGCTCACAGACAGGCCCGAGTCCGTCATG
GCGGCGCGAACGCTGTGCAGCAAATCGAACAGCTCCGAGCGGTCTCTGTTGTGAATACCCACCTTCGA
```

Tick!

FASTA / Text

Copy & Paste

FastA-format

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

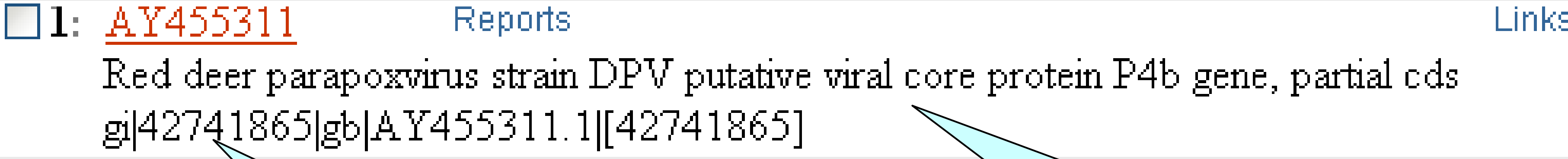
>RedDeer parapox DPV P4b AY455311
TGGGGCGCACGCGGTGGTTGTGGTGC

- **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]

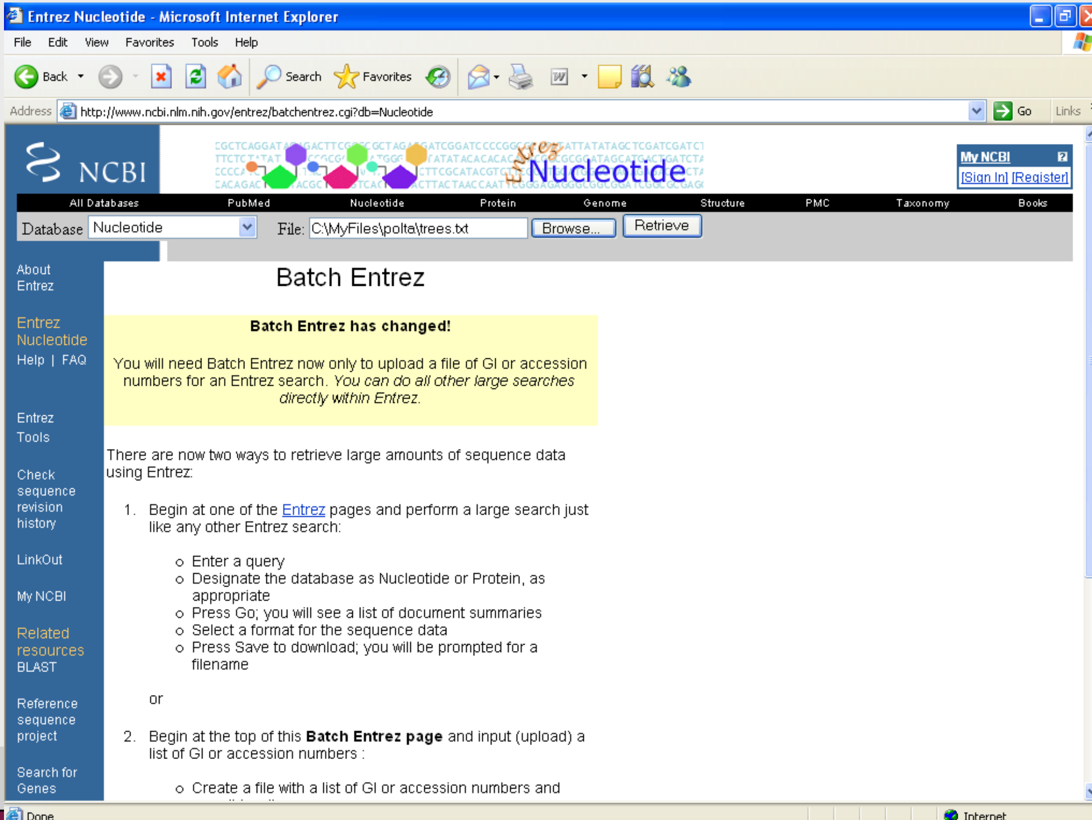
Sequence description

- 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



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

Database: Nucleotide

File: C:\MyFiles\polta\trees.txt

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:
   - Enter a query
   - Designate the database as Nucleotide or Protein, as appropriate
   - Press Go; you will see a list of document summaries
   - Select a format for the sequence data
   - 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 :
   - 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

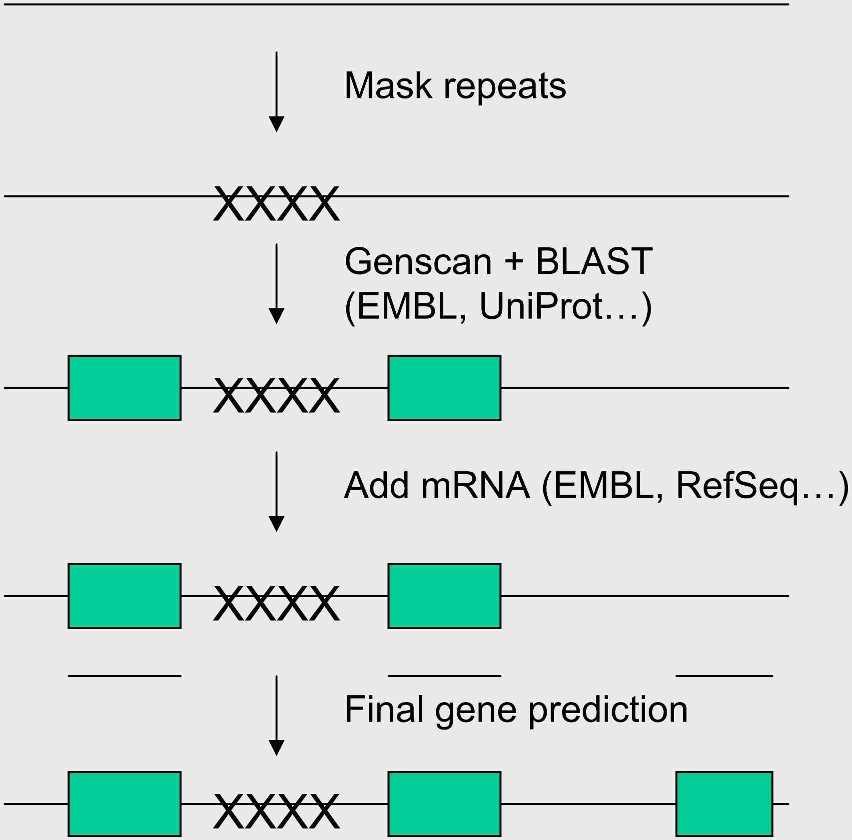
- **Ensembl**
  - http://www.ensembl.org
  - 19 species (Chordates!)
- **UCSC Genome Browser**
  - http://genome.ucsc.edu/
  - 28 species (Insects!)
- **NCBI MapViewer**
  - http://www.ncbi.nlm.nih.gov/mapview/
  - 38 species (Plants, Fungi!)

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



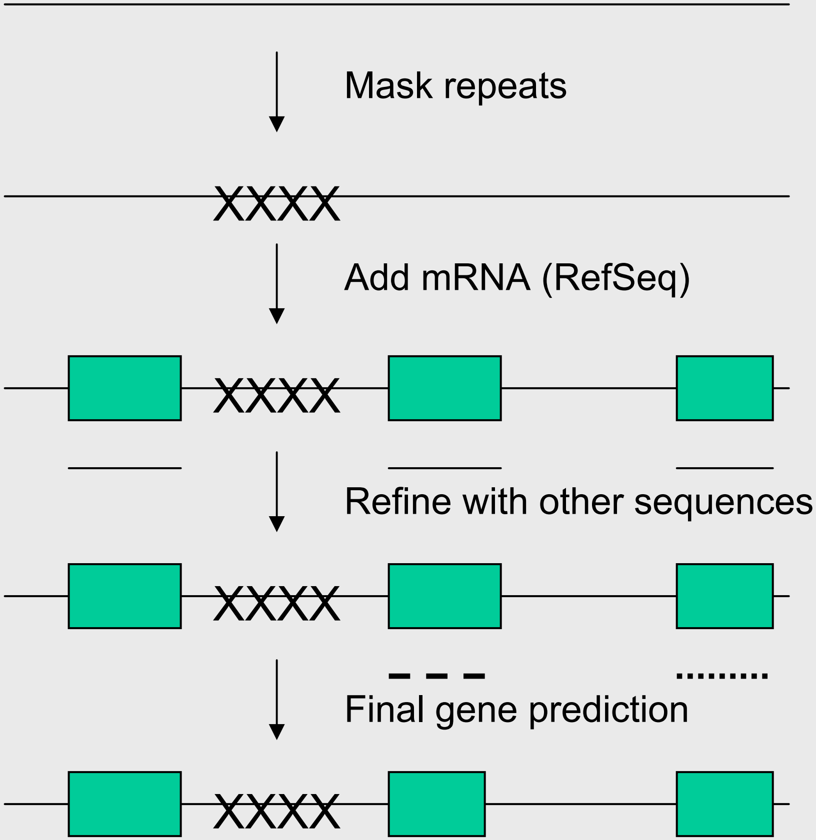
Mask repeats

Genscan + BLAST
(EMBL, UniProt…)

Add mRNA (EMBL, RefSeq…)

Final gene prediction

MapViewer



Mask repeats

Add mRNA (RefSeq)

Refine with other sequences

Final gene prediction

Other organisms

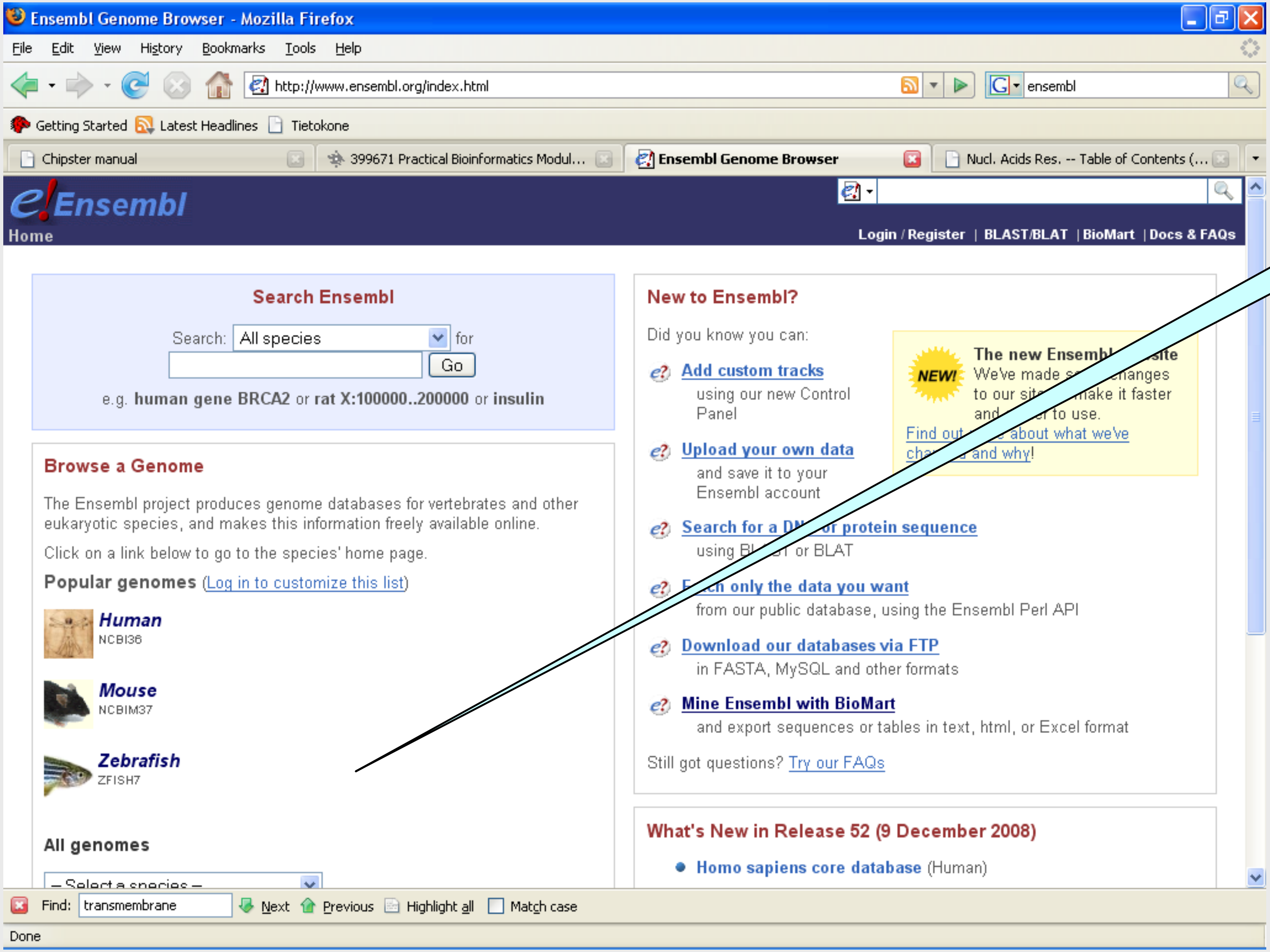
- **Yeast:**
  - http://www.yeastgenome.org/
- **Microbes:**
  - http://www.tigr.org/tdb/mdb/mdbcomplete.html
- **Parasites, single celled eukaryots…**
  - http://www.tigr.org/tdb/euk/
  - http://www.sanger.ac.uk/Projects/

Some considerations

- **Selection of the database**
  - Organism content
  - Speed (MapViewer 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



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 –

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 DN or protein sequence using Bl T or BLAT
- F ch 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

The new Ensembl site
We've made s changes to our sit make it faster and r to use.
Find out e about what we've cha and why!

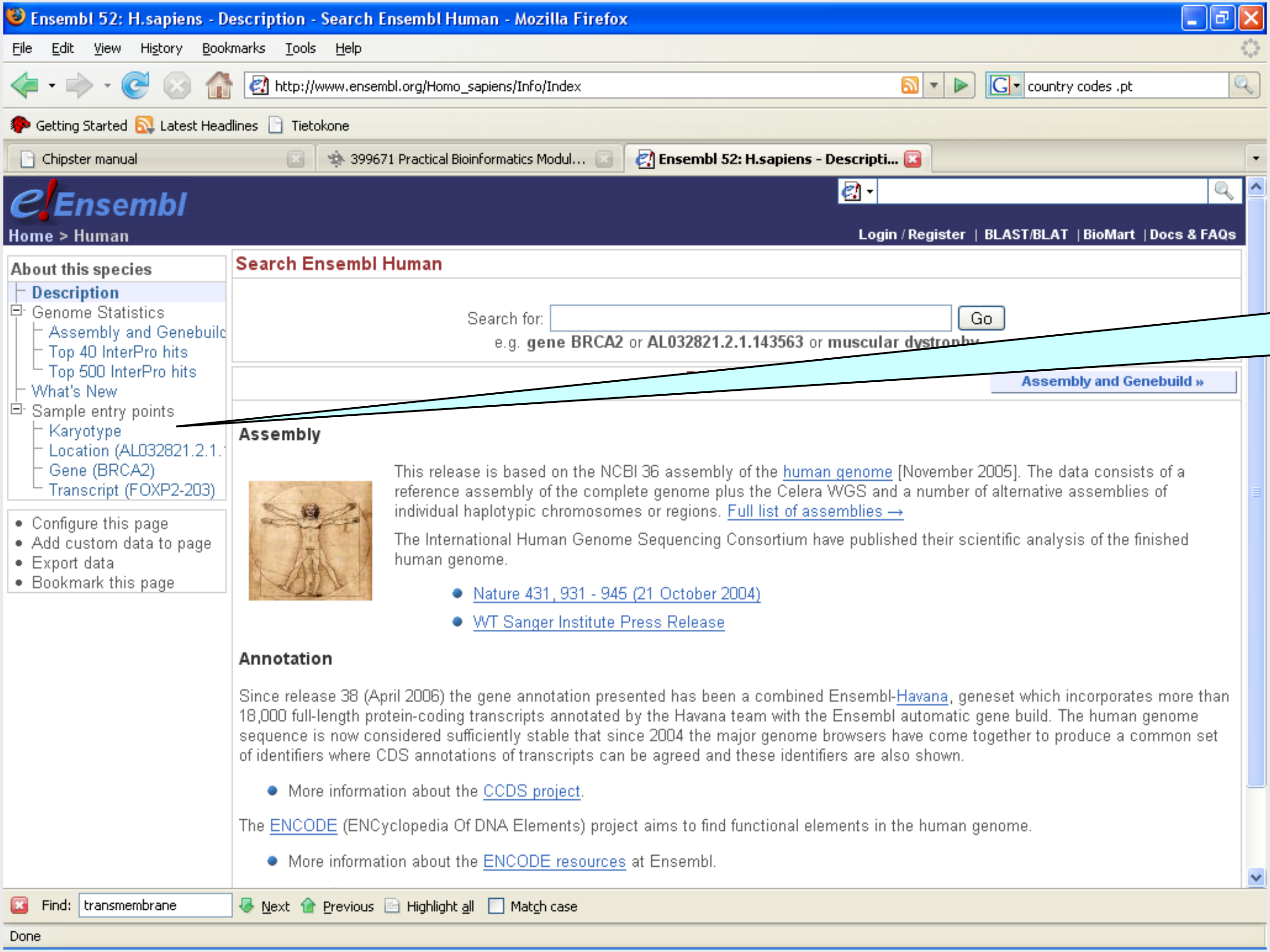
Still got questions? Try our FAQs

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

- **Homo sapiens core database** (Human)

Browse genome

Explore the genome



Search Ensembl Human

Search for:

e.g. **gene BRCA2** or **AL032821.2.1.143563** or **muscular dystrophy**

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.

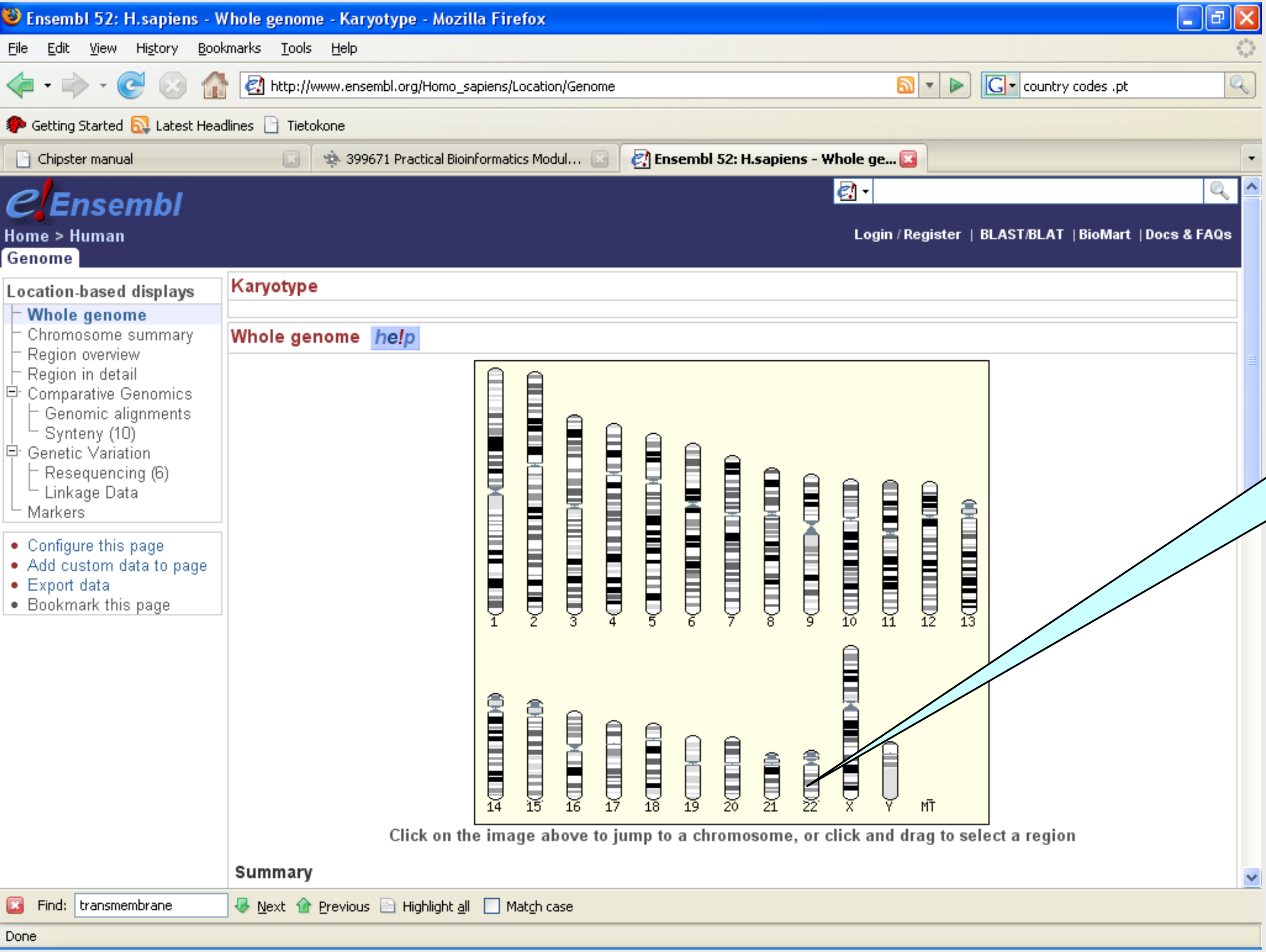
- More information about the CCDS project.

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

- More information about the ENCODE resources at Ensembl.

Karyotype (=chromosome view)

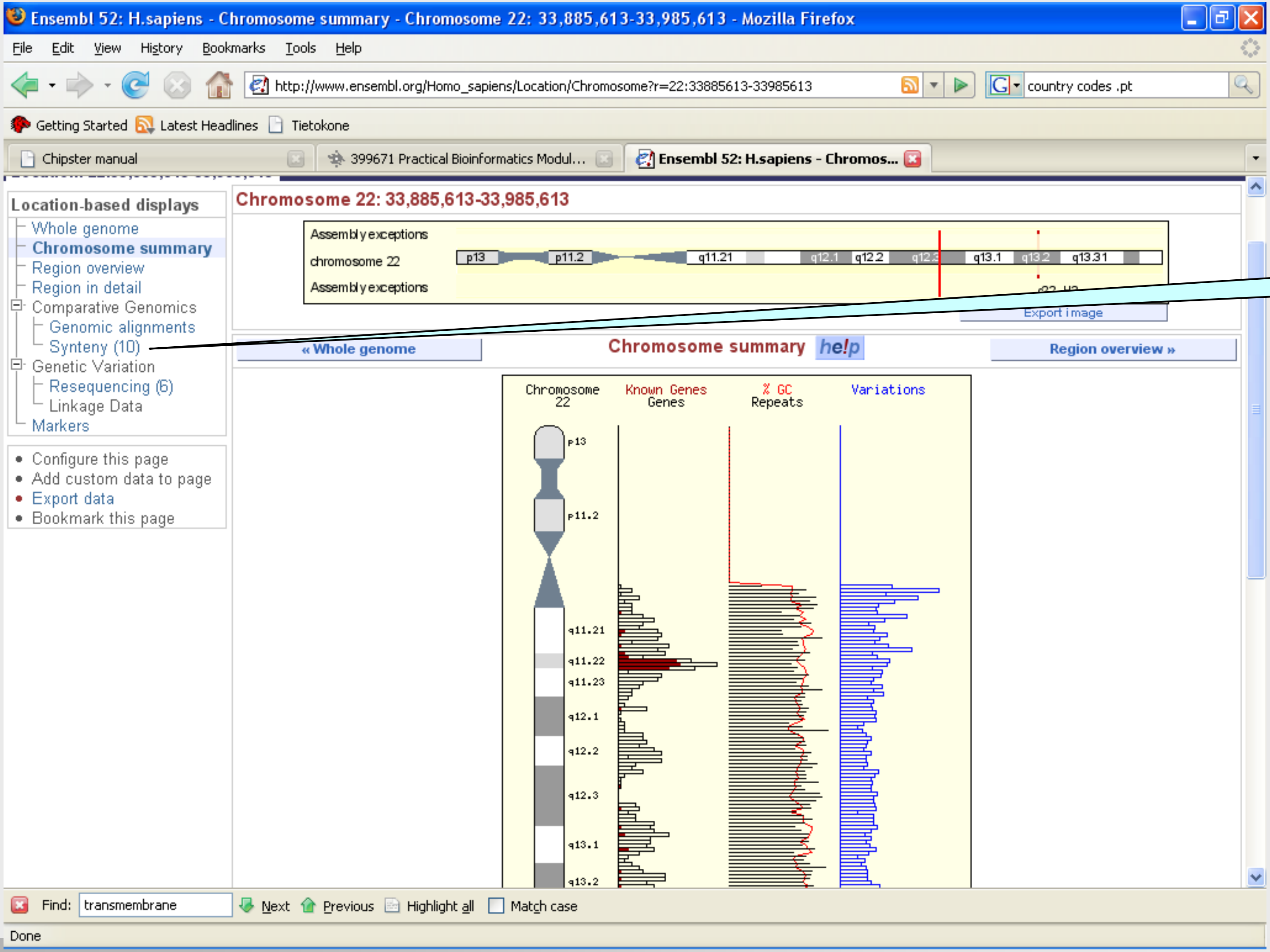
Explore chromosomes



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

Select a chromosome -> chromosome summary

Chromosome summary



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

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

country codes .pt

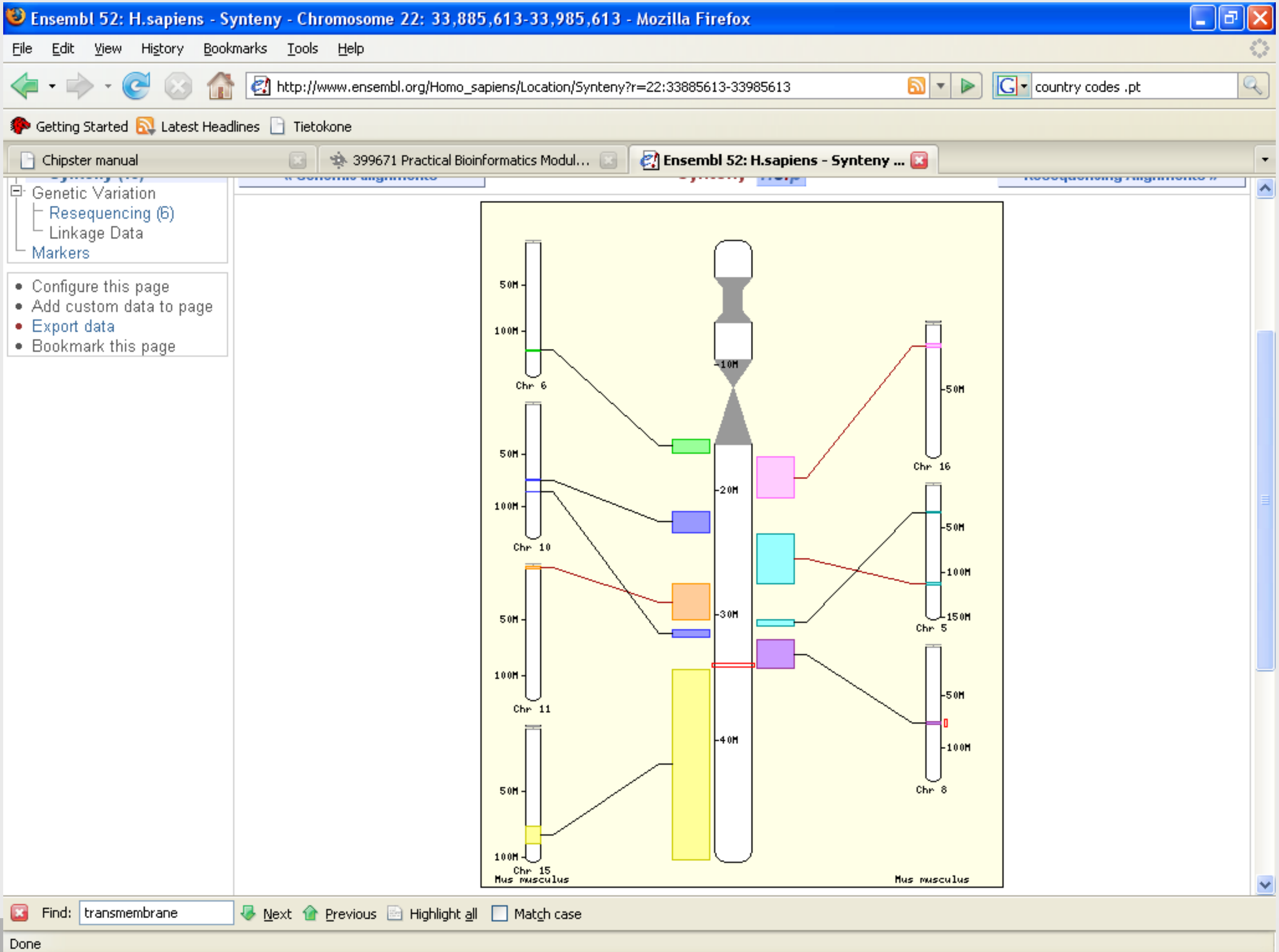
Chromosome 22: 33,885,613-33,985,613

Find: transmembrane

Done

Synteny

Synteny View



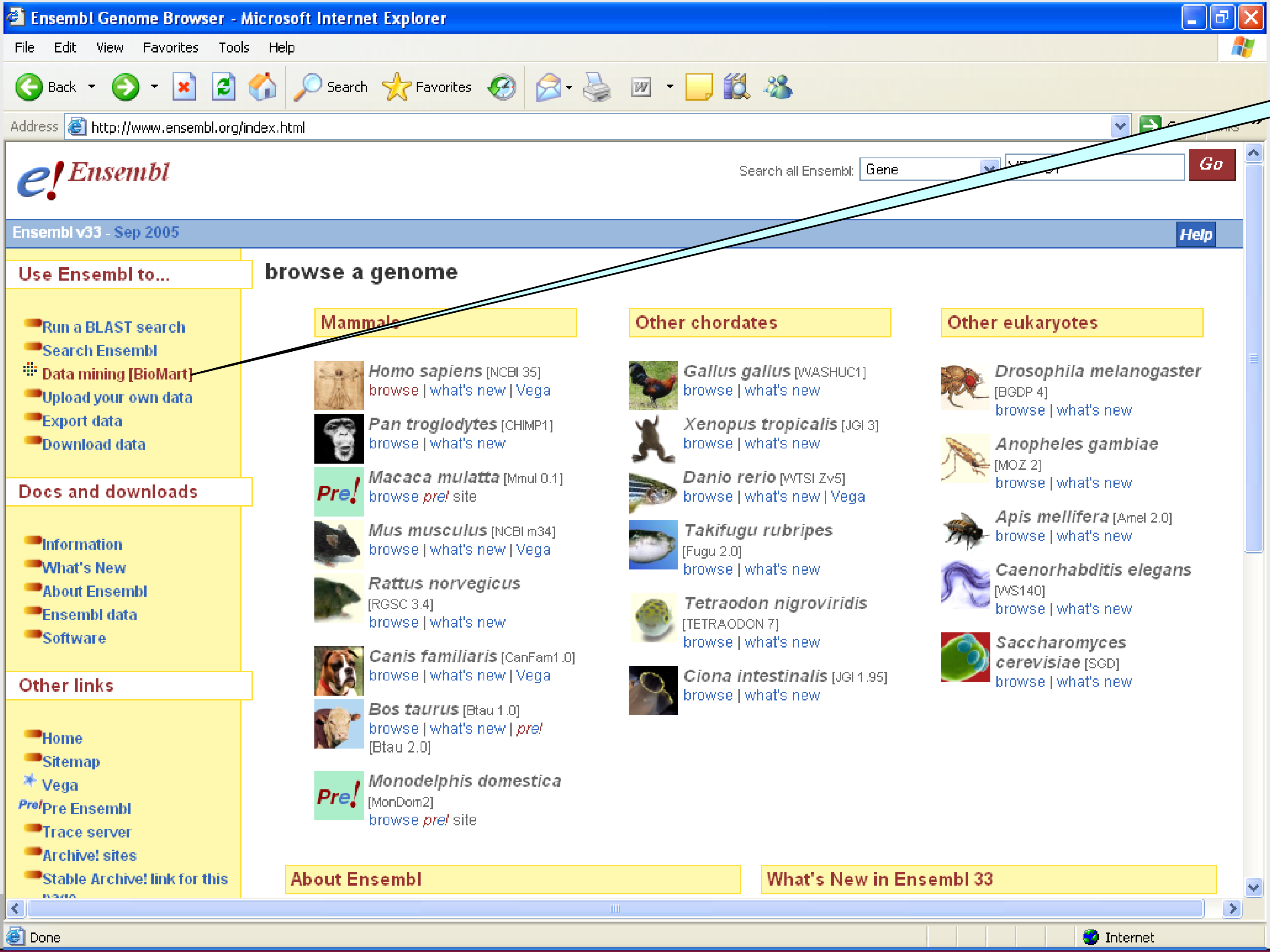
Ensembl 52: H.sapiens - Synteny - Chromosome 22: 33,885,613-33,985,613 - Mozilla Firefox

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

country codes .pt

Find: transmembrane

Ensembl front page



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

e! Ensembl

Search all Ensembl: Gene

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 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

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]

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 [WS140]
browse | what's new

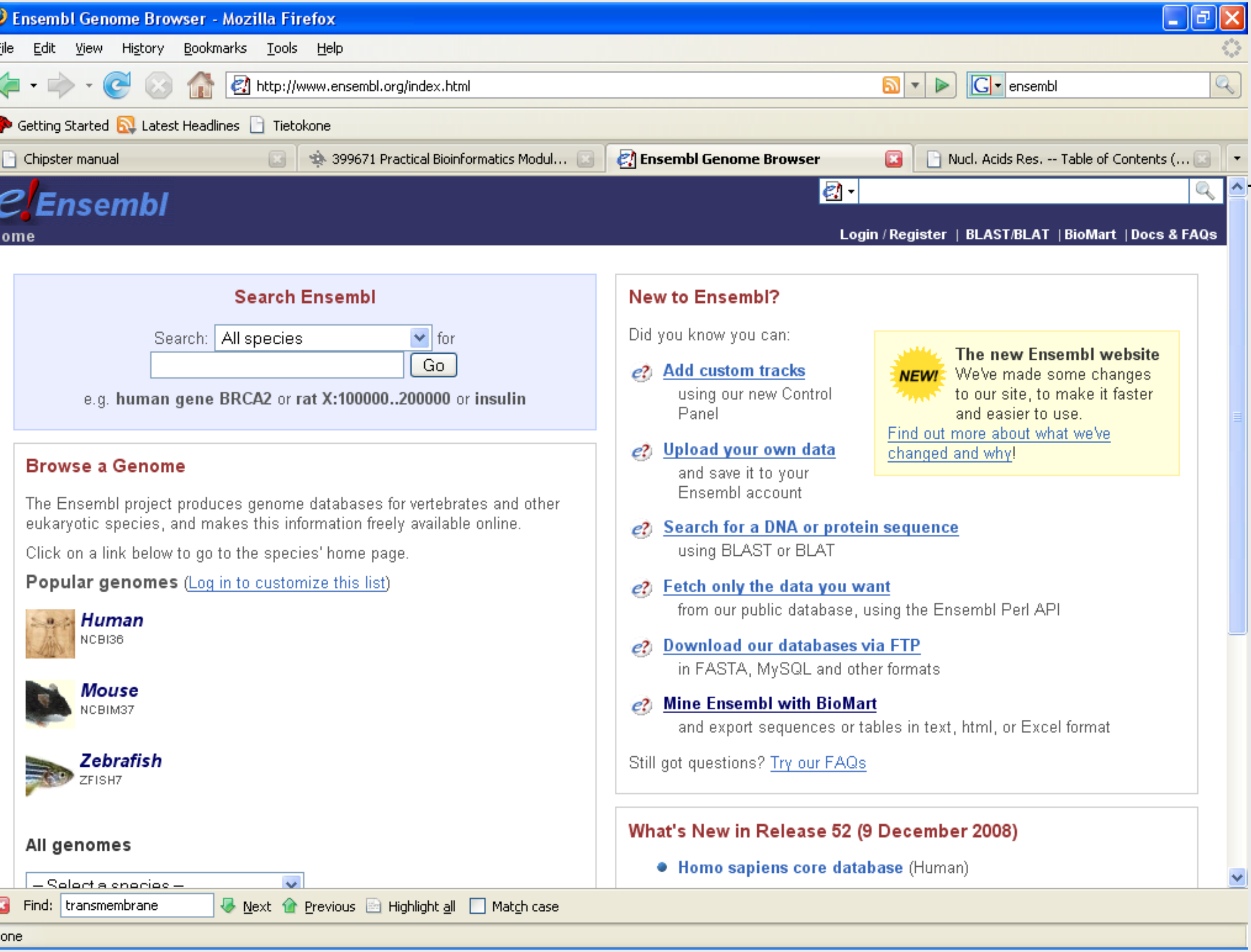
Saccharomyces cerevisiae [SGD]
browse | what's new

About Ensembl

What's New in Ensembl 33

BioMart

Ensembl front page



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

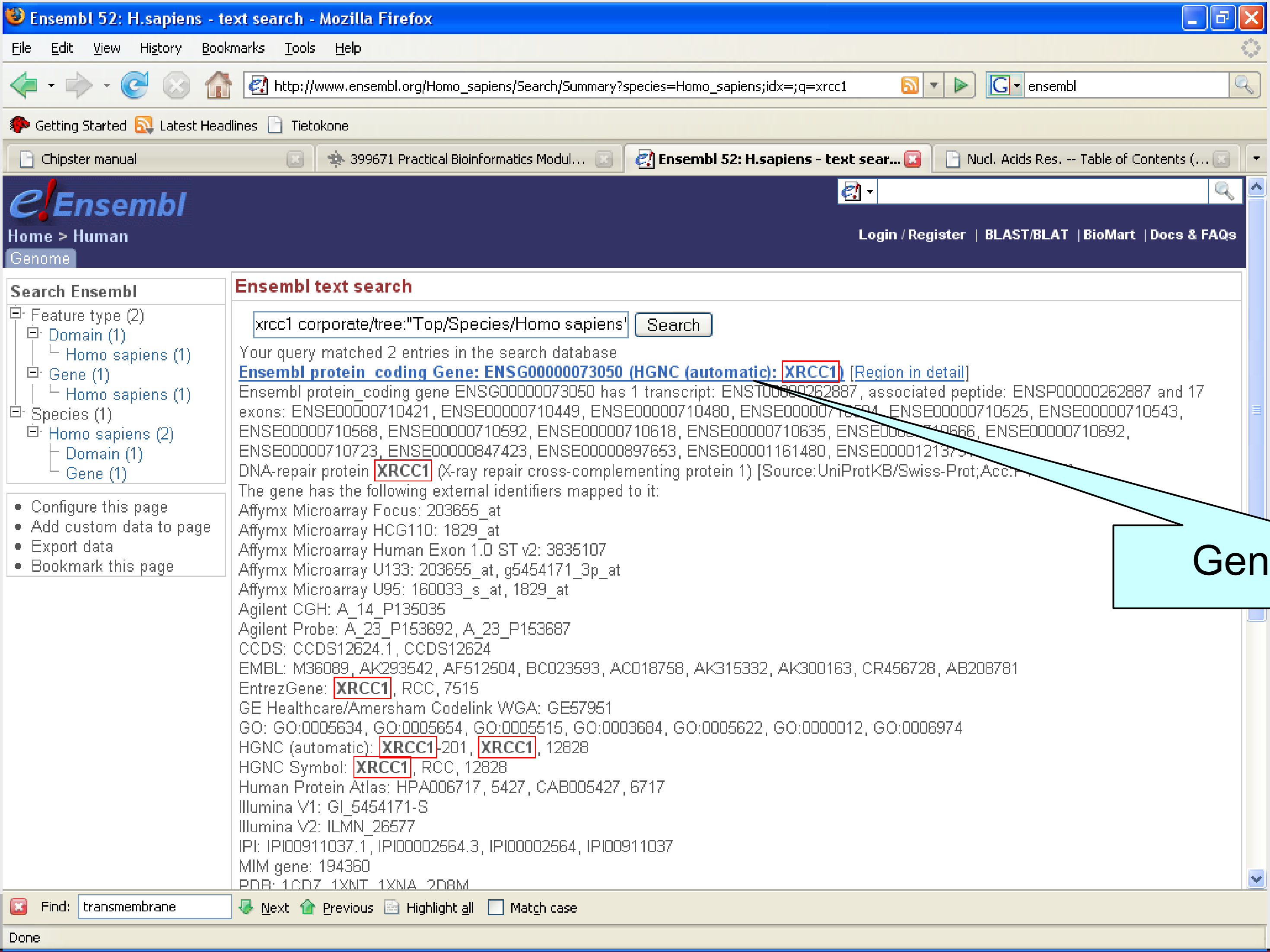
NEW! **The new Ensembl website** We've made some changes to our site, to make it faster and easier to use. Find out more about what we've changed and why!

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

- **Homo sapiens core database** (Human)

Quick search

Quick search results



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' Search

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, ENSE000007, ENSE00000710525, ENSE00000710543, ENSE00000710568, ENSE00000710592, ENSE00000710618, ENSE00000710635, ENSE000, ENSE00000710692, ENSE00000710723, ENSE00000847423, ENSE00000897653, ENSE00001161480, ENSE000012137

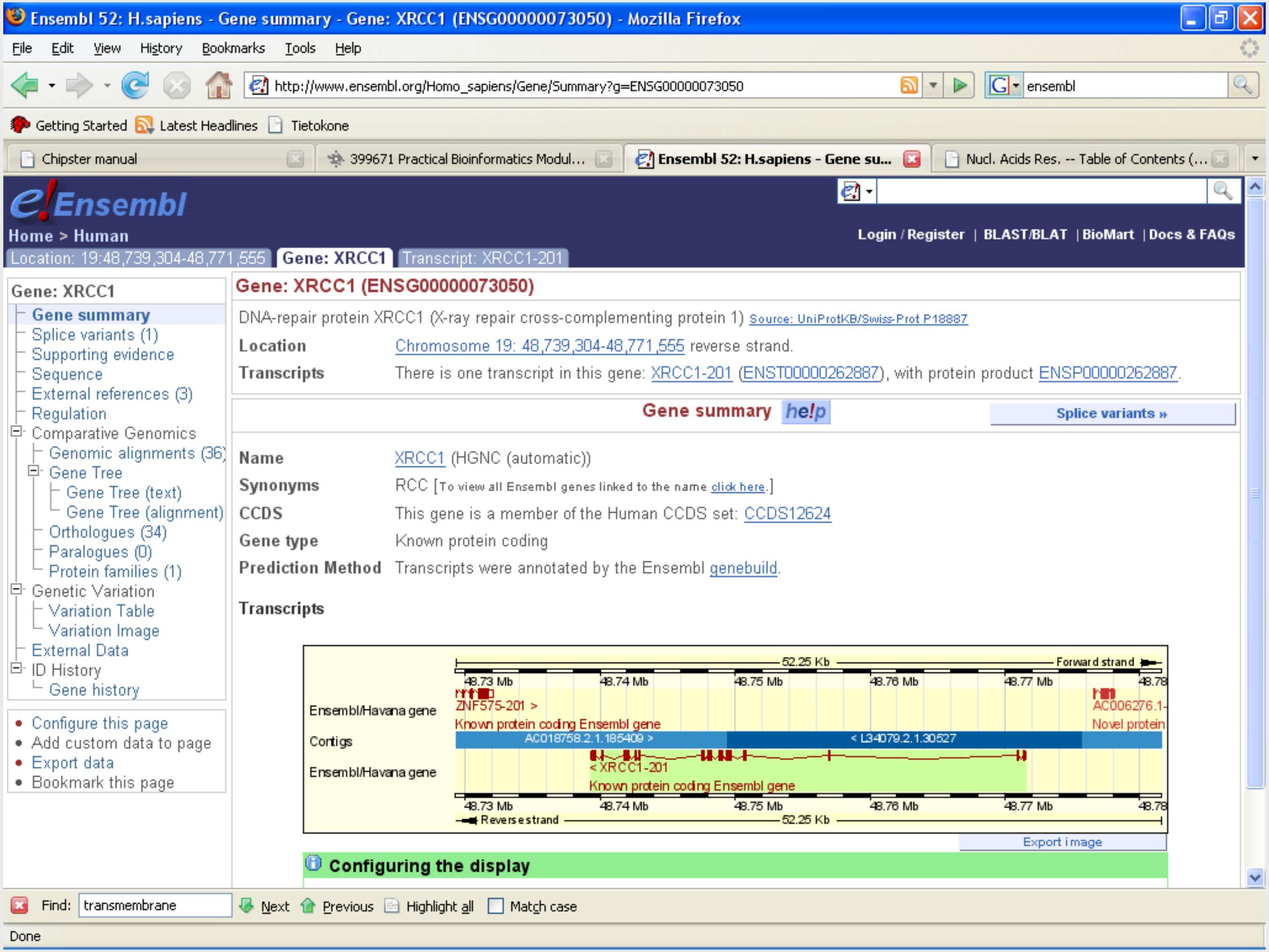
DNA-repair protein **XRCC1** (X-ray repair cross-complementing protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P

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: 1CD7, 1XNT, 1XNA, 2D8M

Geneview link

Gene View



Gene: XRCC1 (ENSG00000073050)

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

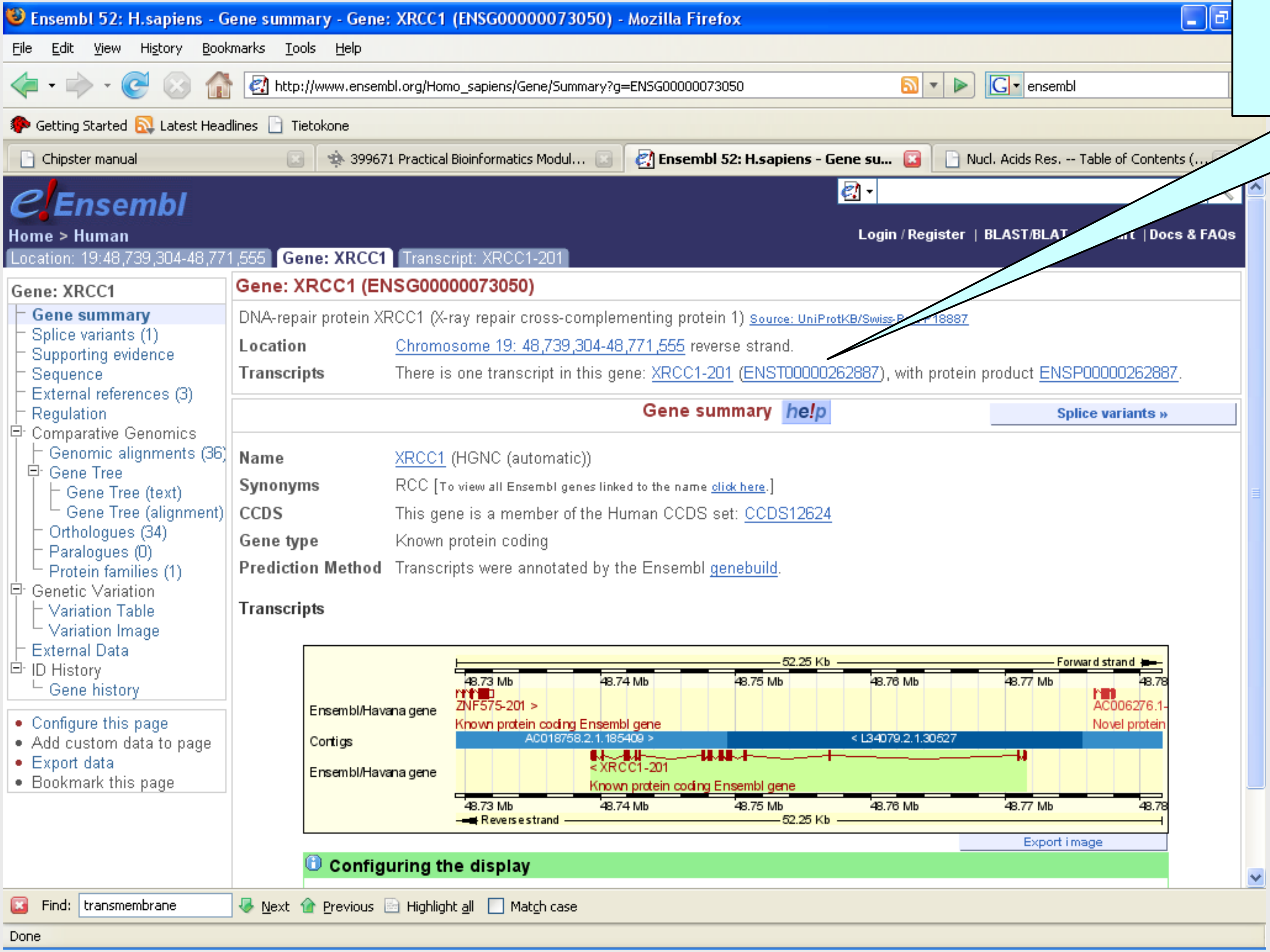
Location	Chromosome 19: 48,739,304-48,771,555 reverse strand.
Transcripts	There is one transcript in this gene: XRCC1-201 (ENST00000262887), with protein product ENSP00000262887.

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

Gene View

Information on transcripts and proteins.



Gene: XRCC1 (ENSG00000073050)

DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1) Source: UniProtKB/Swiss-P...18887

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.

Name XRCC1 (HGNC (automatic))

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

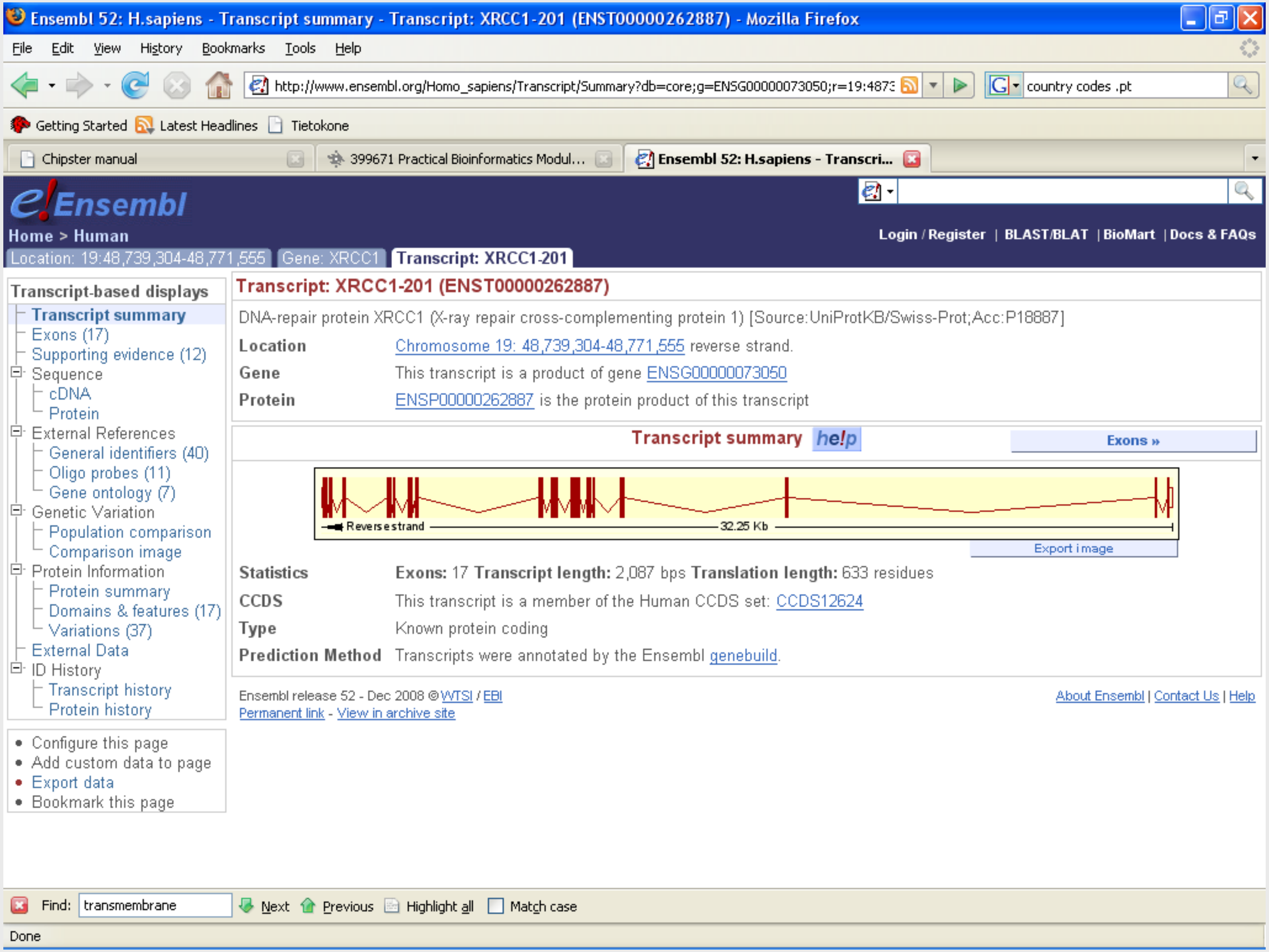
CCDS This gene is a member of the Human CCDS set: CCDS12624

Gene type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl genebuild.

Transcripts

Transcript info



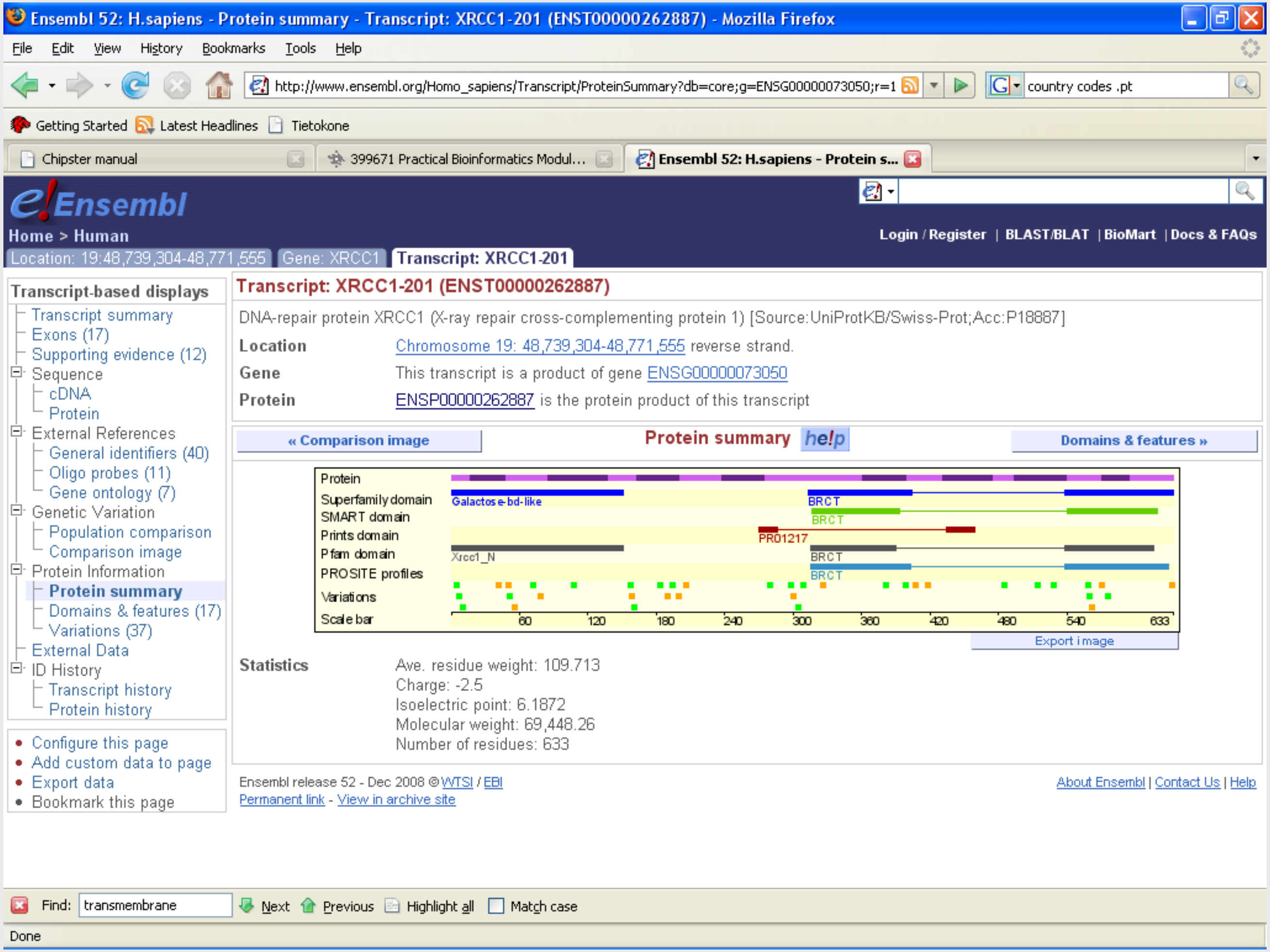
Transcript: XRCC1-201 (ENST00000262887)

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
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
Permanent link - View in archive site

Protein summary



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

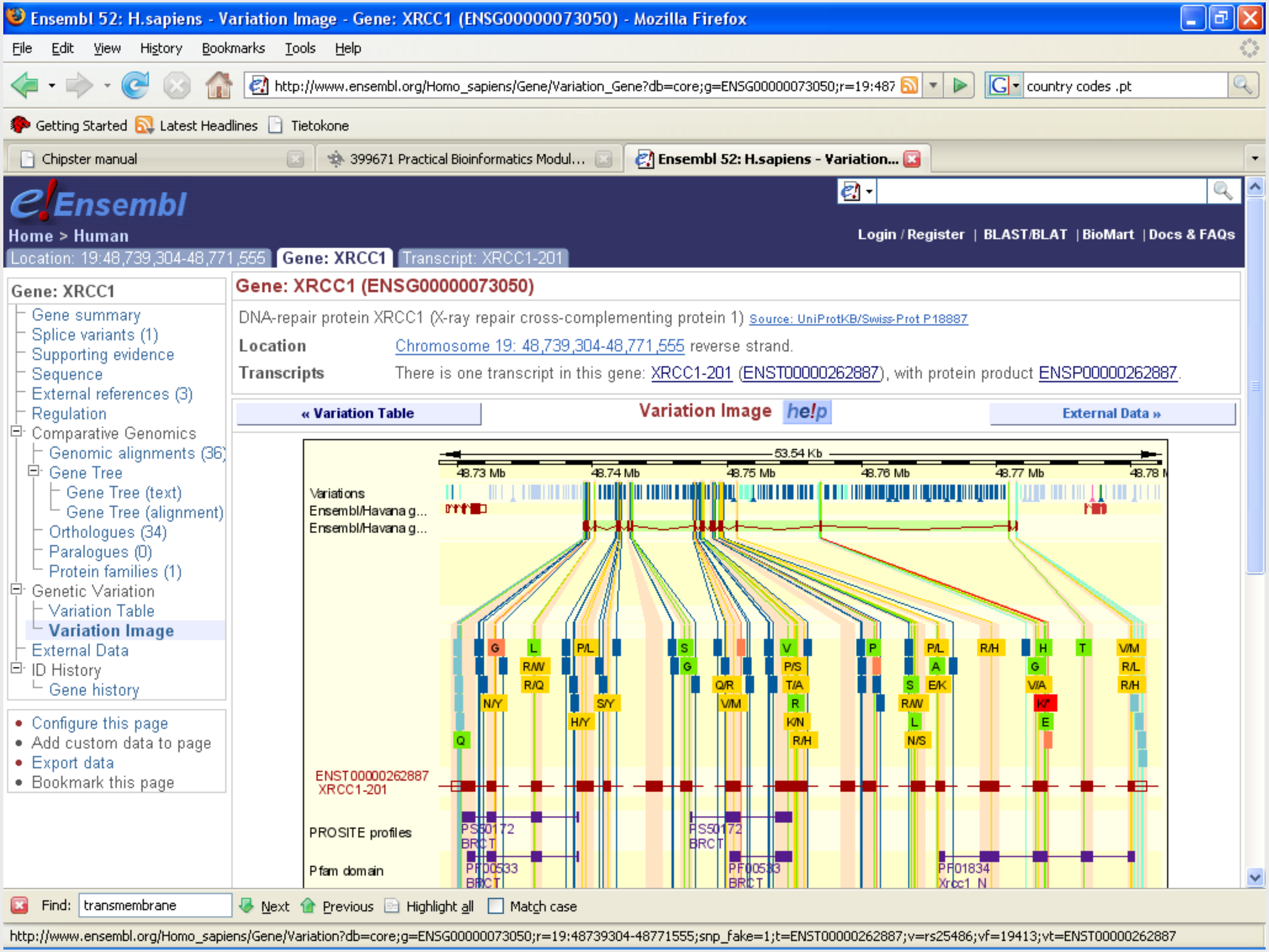
Statistics

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
Permanent link - View in archive site

About Ensembl | Contact Us | Help

Variation (SNP) view – from Gene tab



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

http://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene?db=core;g=ENSG00000073050;r=19:487

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.

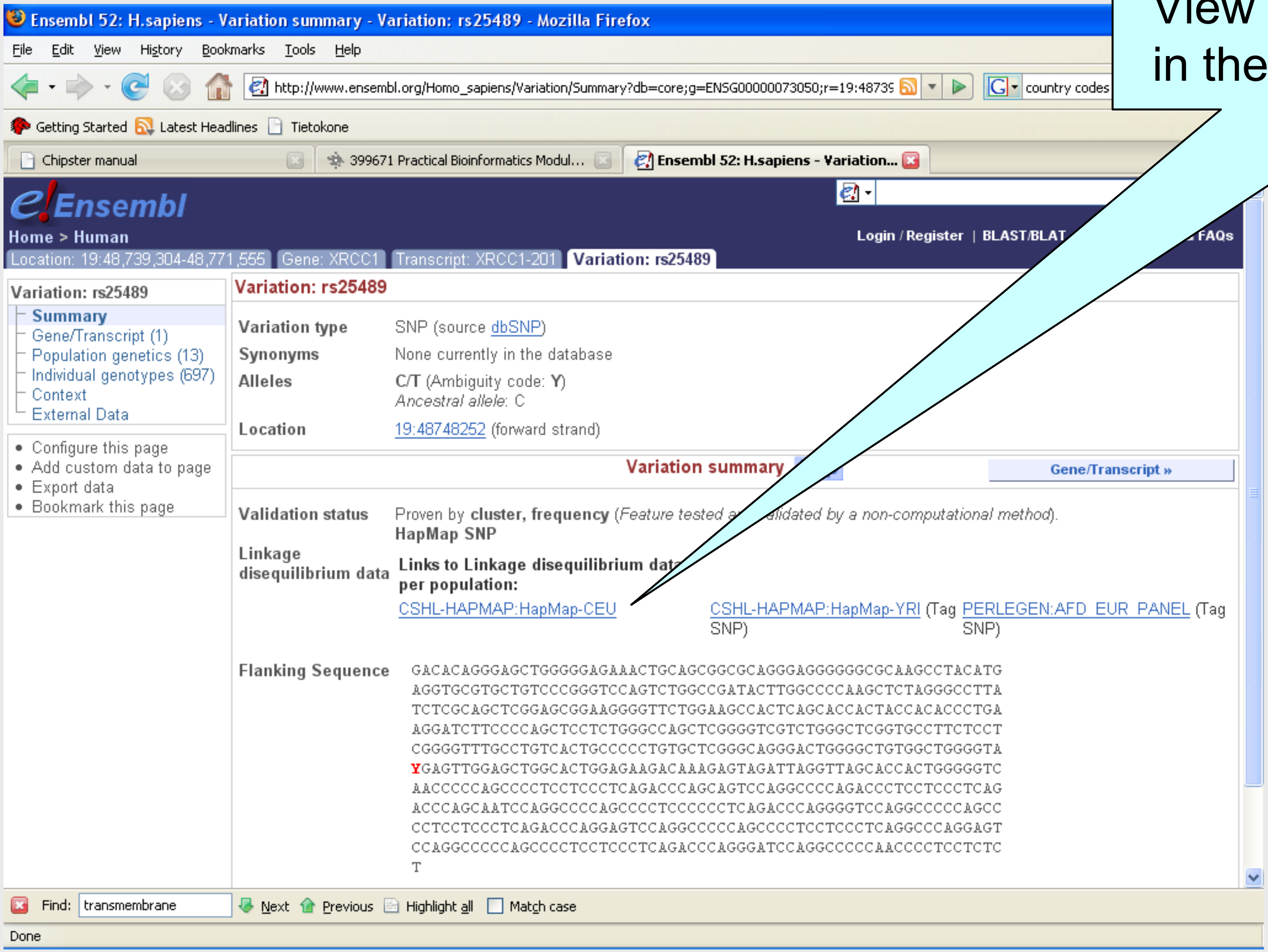
Transcripts There is one transcript in this gene: XRCC1-201 (ENST00000262887), with protein product ENSP00000262887.

Find: transmembrane

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).



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)

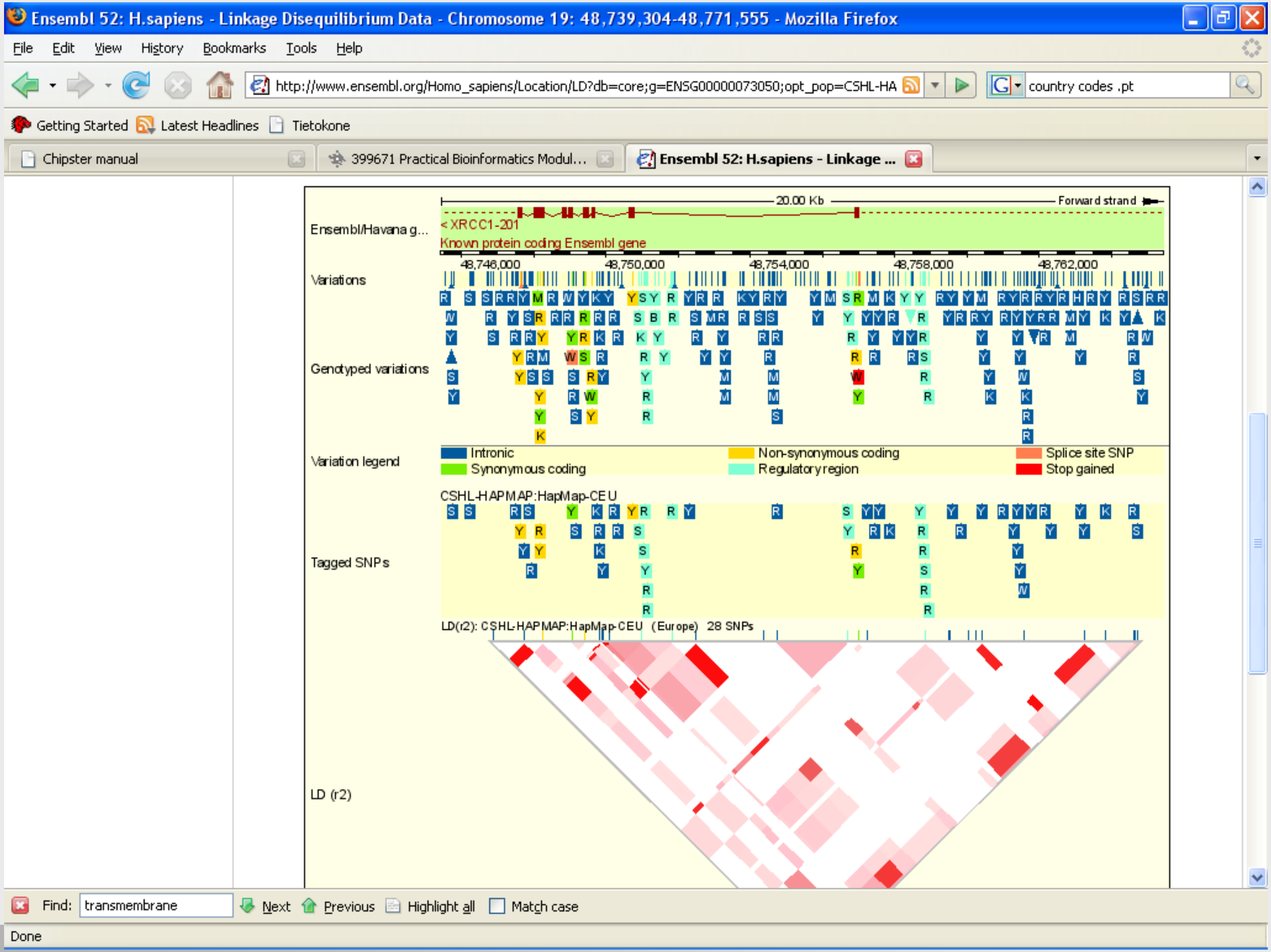
Variation summary

Validation status	Proven by **cluster, frequency** (*Feature tested a alidated by a non-computational method*). **HapMap SNP**
Linkage disequilibrium data	**Links to Linkage disequilibrium dat per population:** CSHL-HAPMAP:HapMap-CEU CSHL-HAPMAP:HapMap-YRI (Tag SNP) PERLEGEN:AFD_EUR_PANEL (Tag SNP)

Flanking Sequence

```
GACACAGGGAGCTGGGGGAGAAACTGCAGCGGCGCAGGGAGGGGGGCGCAAGCCTACATG
AGGTGCGTGCTGTCCCGGGTCCAGTCTGGCCGATACTTGGCCCCAAGCTCTAGGGCCTTA
TCTCGCAGCTCGGAGCGGAAGGGGTTCTGGAAGCCACTCAGCACCACTACCACACCCTGA
AGGATCTTCCCCAGCTCCTCTGGGCCAGCTCGGGGTCGTCTGGGCTCGGTGCCTTCTCCT
CGGGGTTTGCCTGTCACTGCCCCCTGTGCTCGGGCAGGGACTGGGGCTGTGGCTGGGGTA
YGAGTTGGAGCTGGCACTGGAGAAGACAAAGAGTAGATTAGGTTAGCACCACTGGGGGTC
AACCCCCAGCCCCTCCTCCCTCAGACCCAGCAGTCCAGGCCCCAGACCCTCCTCCCTCAG
ACCCAGCAATCCAGGCCCCAGCCCCTCCCCCCTCAGACCCAGGGGTCCAGGCCCCCAGCC
CCTCCTCCCTCAGACCCAGGAGTCCAGGCCCCCAGCCCCTCCTCCCTCAGGCCCAGGAGT
CCAGGCCCCCAGCCCCTCCTCCCTCAGACCCAGGGATCCAGGCCCCCAACCCCTCCTCTC
T
```

LDView

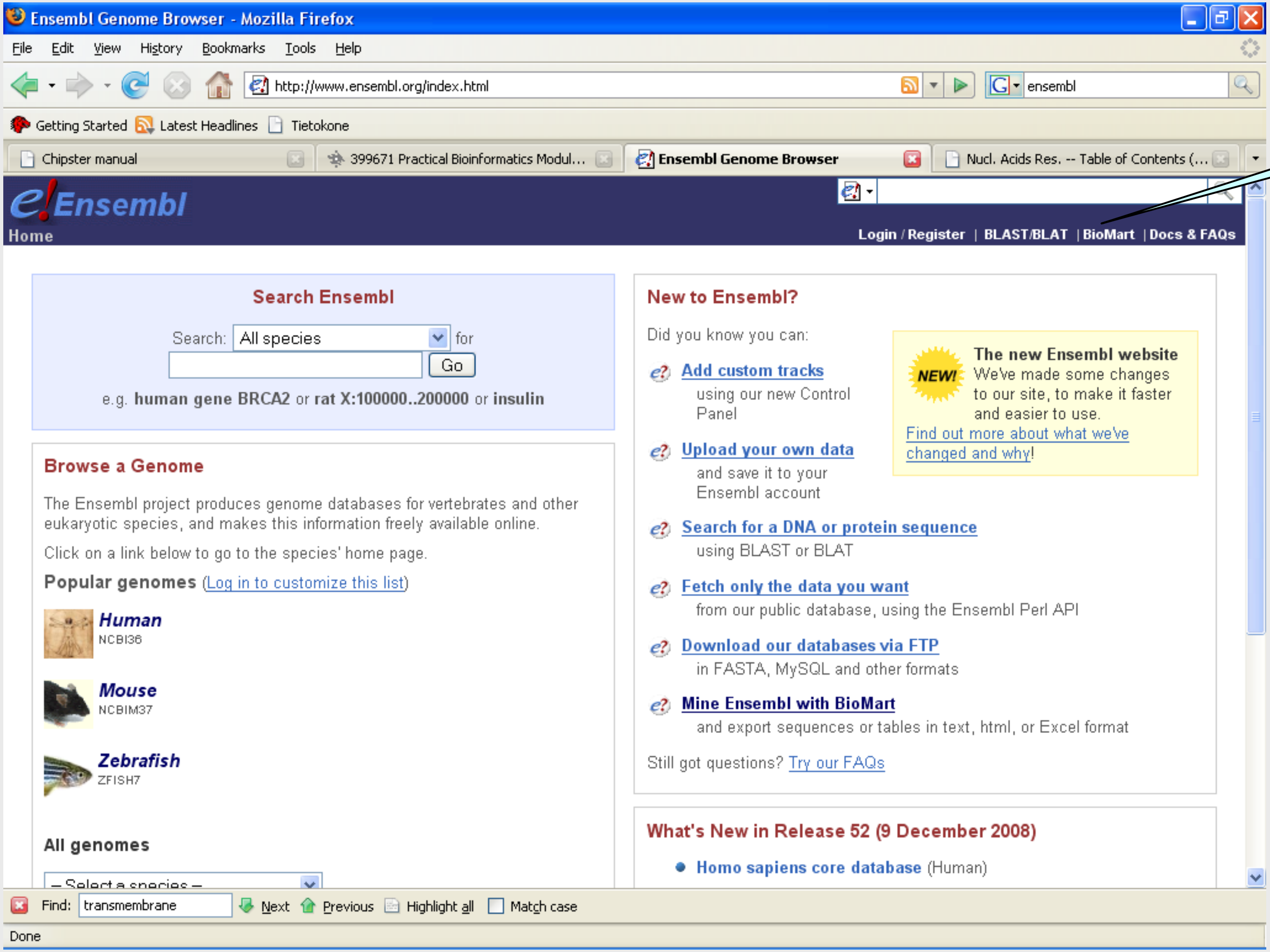


Ensembl 52: H.sapiens - Linkage Disequilibrium Data - Chromosome 19: 48,739,304-48,771,555 - Mozilla Firefox

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

Find: transmembrane

Ensembl front page



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

ensembl

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

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!

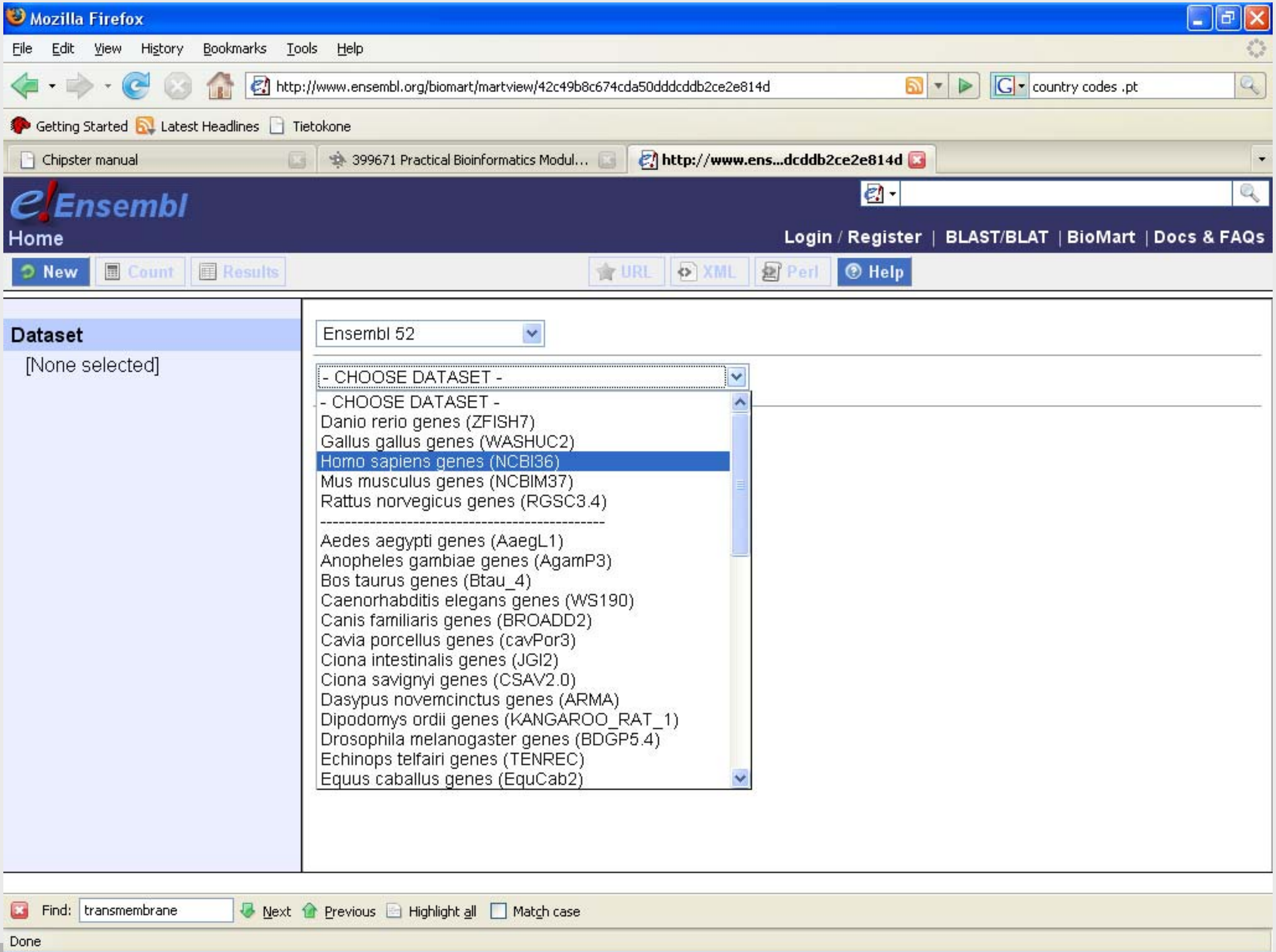
What's New in Release 52 (9 December 2008)

- **Homo sapiens core database** (Human)

Find: transmembrane

BioMart

MartView – select genome



Dataset
[None selected]

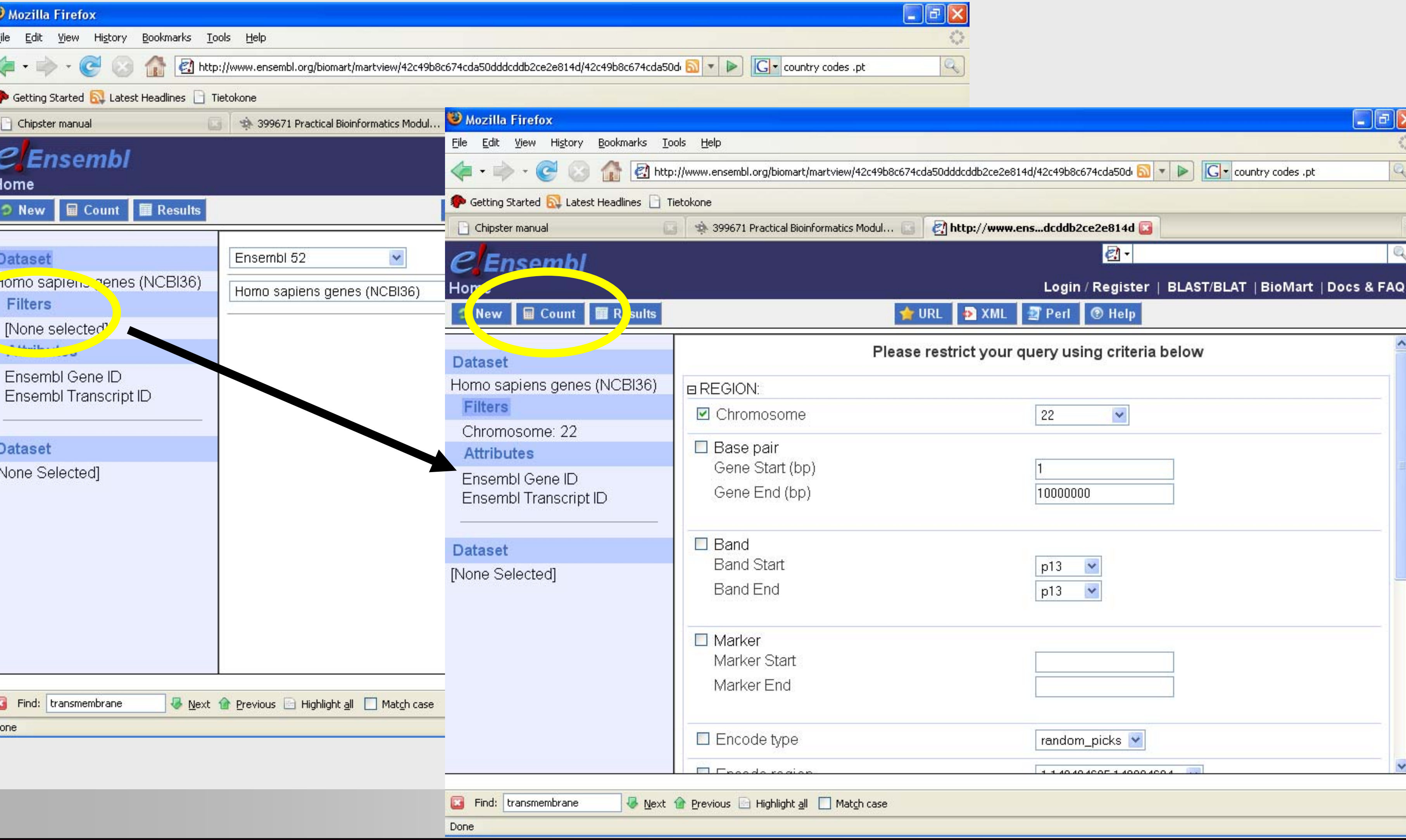
Ensembl 52

- 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)
Dasypus novemcinctus genes (ARMA)
Dipodomys ordii genes (KANGAROO_RAT_1)
Drosophila melanogaster genes (BDGP5.4)
Echinops telfairi genes (TENREC)
Equus caballus genes (EquCab2)

MartView - Filter



Left window:

Ensembl 52

Homo sapiens genes (NCBI36)

Dataset
Homo sapiens genes (NCBI36)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

Right window:

Dataset
Homo sapiens genes (NCBI36)
Filters
Chromosome: 22
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

Please restrict your query using criteria below

⊟ REGION:

☑ Chromosome: 22

☐ Base pair
Gene Start (bp): 1
Gene End (bp): 10000000

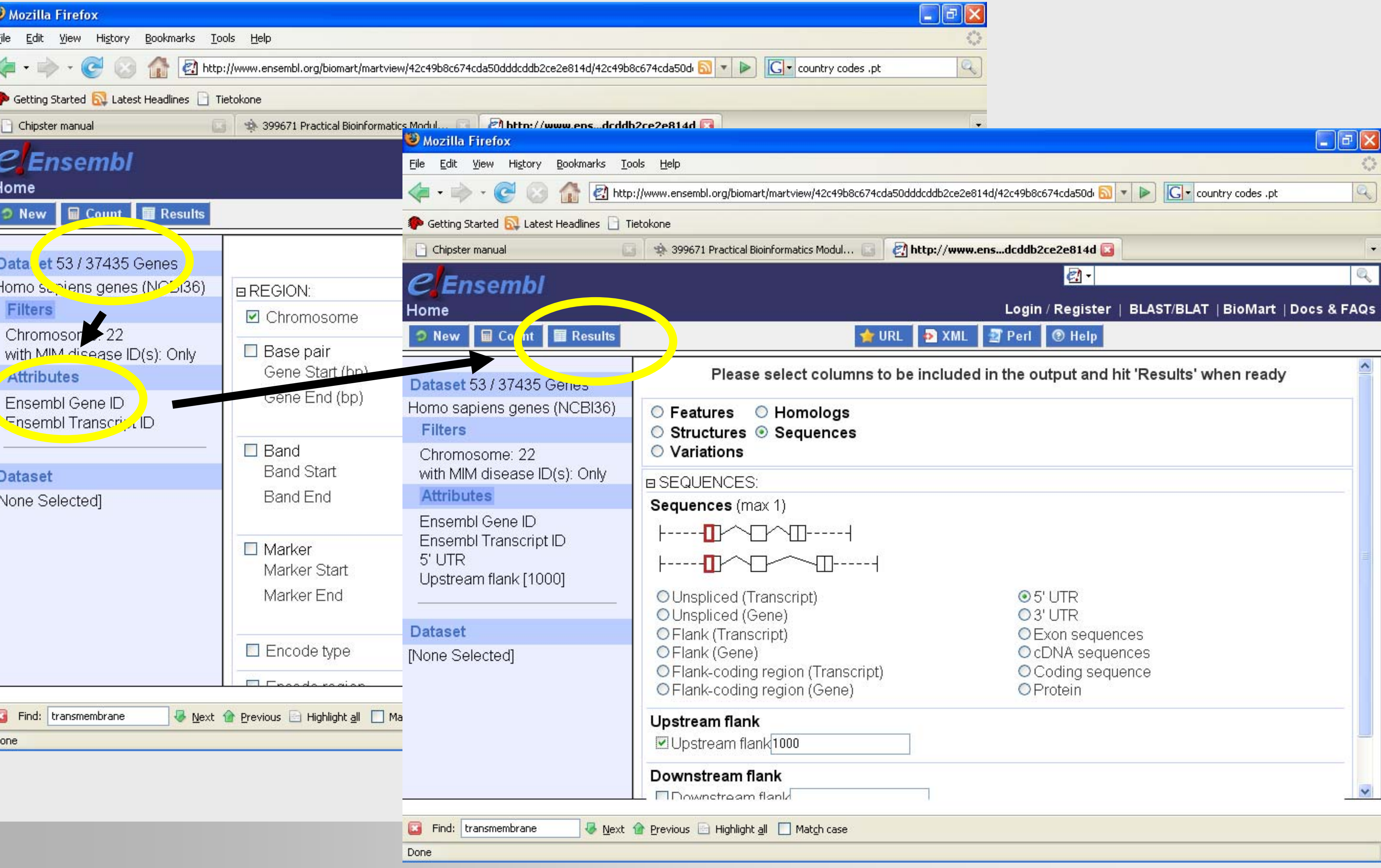
☐ Band
Band Start: p13
Band End: p13

☐ Marker
Marker Start:
Marker End:

☐ Encode type: random_picks

☐ Encode region

MartView - output



Dataset 53 / 37435 Genes
Homo sapiens genes (NCBI36)

Filters

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

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

⊟ REGION:

☑ Chromosome

☐ Base pair
Gene Start (bp)
Gene End (bp)

☐ Band
Band Start
Band End

☐ Marker
Marker Start
Marker End

☐ Encode type

☐ Encode region

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]

Please select columns to be included in the output and hit 'Results' when ready

○ Features ○ Homologs
○ Structures ● Sequences
○ Variations

⊟ SEQUENCES:

Sequences (max 1)

○ Unspliced (Transcript)
○ Unspliced (Gene)
○ Flank (Transcript)
○ Flank (Gene)
○ Flank-coding region (Transcript)
○ Flank-coding region (Gene)
● 5' UTR
○ 3' UTR
○ Exon sequences
○ cDNA sequences
○ Coding sequence
○ Protein

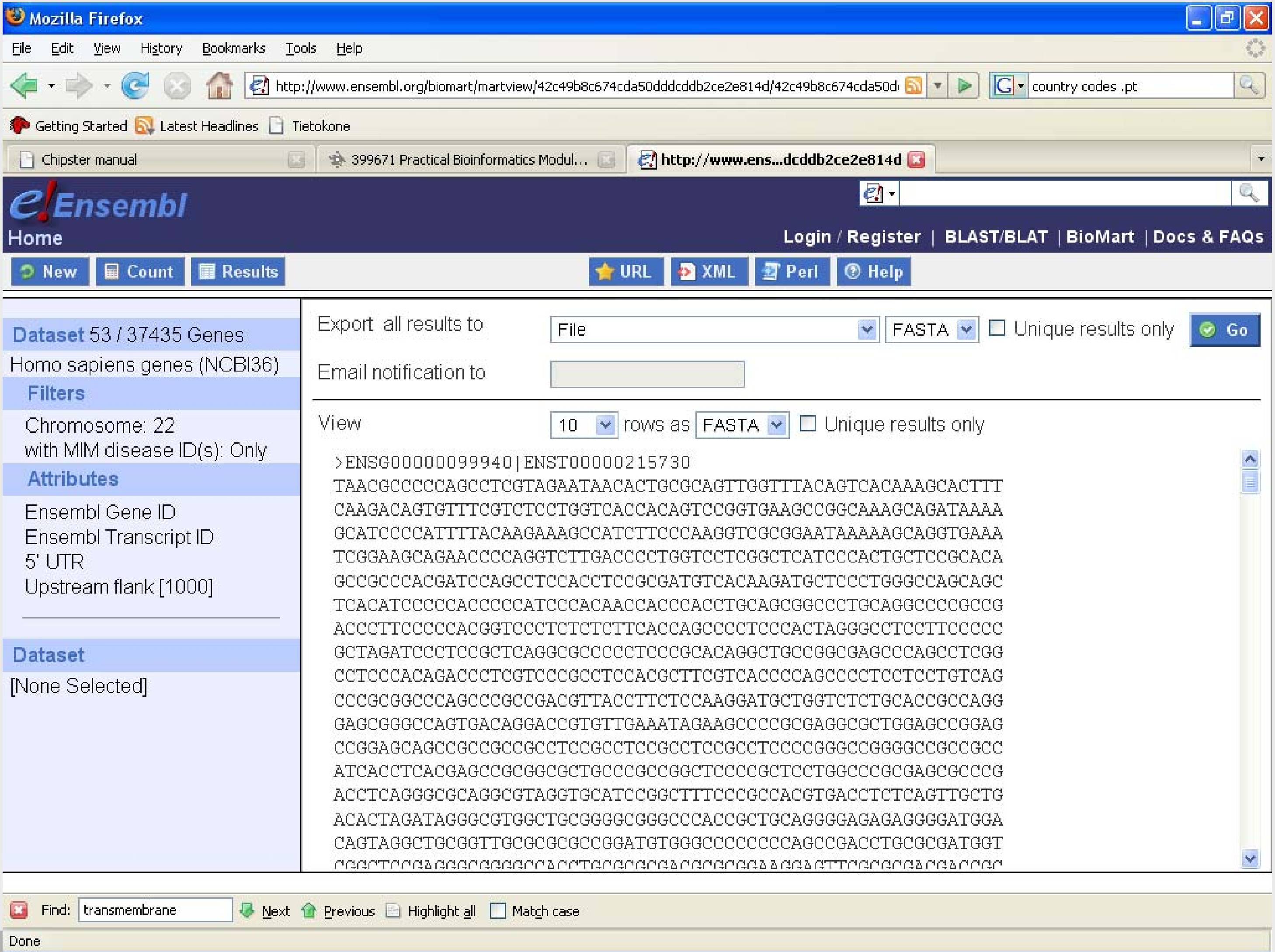
Upstream flank

☑ Upstream flank 1000

Downstream flank

☐ Downstream flank

MartView



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

Email notification to:

View 10 rows as FASTA ☐ Unique results only

```
>ENSG00000099940|ENST00000215730
TAACGCCCCCAGCCTCGTAGAATAACACTGCGCAGTTGGTTTACAGTCACAAAGCACTTT
CAAGACAGTGTTTCGTCTCCTGGTCACCACAGTCCGGTGAAGCCGGCAAAGCAGATAAAA
GCATCCCCATTTTACAAGAAAGCCATCTTCCCAAGGTCGCGGAATAAAAAGCAGGTGAAA
TCGGAAGCAGAACCCCAGGTCTTGACCCCTGGTCCTCGGCTCATCCCACTGCTCCGCACA
GCCGCCCACGATCCAGCCTCCACCTCCGCGATGTCACAAGATGCTCCCTGGGCCAGCAGC
TCACATCCCCCACCCCCATCCCACAACCACCCACCTGCAGCGGCCCTGCAGGCCCCGCCG
ACCCTTCCCCCACGGTCCCTCTCTCTTCACCAGCCCCTCCCACTAGGGCCTCCTTCCCCC
GCTAGATCCCTCCGCTCAGGCGCCCCCTCCCGCACAGGCTGCCGGCGAGCCCAGCCTCGG
CCTCCCACAGACCCTCGTCCCGCCTCCACGCTTCGTCACCCCAGCCCCTCCTCCTGTCAG
CCCGCGGCCCAGCCCGCCGACGTTACCTTCTCCAAGGATGCTGGTCTCTGCACCGCCAGG
GAGCGGGCCAGTGACAGGACCGTGTTGAAATAGAAGCCCCGCGAGGCGCTGGAGCCGGAG
CCGGAGCAGCCGCCGCCGCCTCCGCCTCCGCCTCCGCCTCCCCGGGCCGGGGCCGCCGCC
ATCACCTCACGAGCCGCGGCGCTGCCCGCCGGCTCCCCGCTCCTGGCCCGCGAGCGCCCG
ACCTCAGGGCGCAGGCGTAGGTGCATCCGGCTTTCCCGCCACGTGACCTCTCAGTTGCTG
ACACTAGATAGGGCGTGGCTGCGGGGCGGGCCCACCGCTGCAGGGGAGAGAGGGGATGGA
CAGTAGGCTGCGGTTGCGCGCGCCGGATGTGGGCCCCCCCCAGCCGACCTGCGCGATGGT
```

Find: transmembrane

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).**

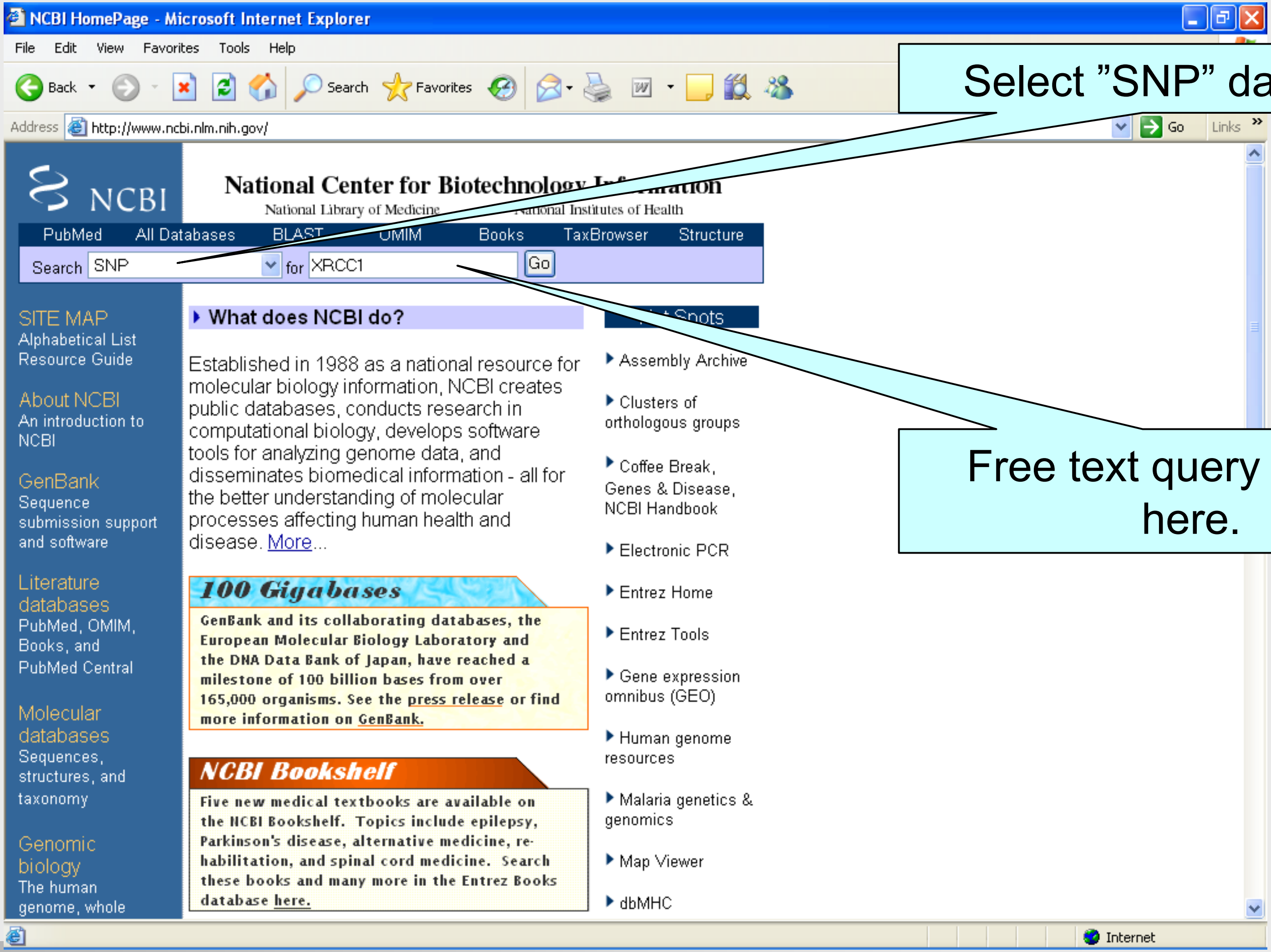
SNPs in databases I

- **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



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.

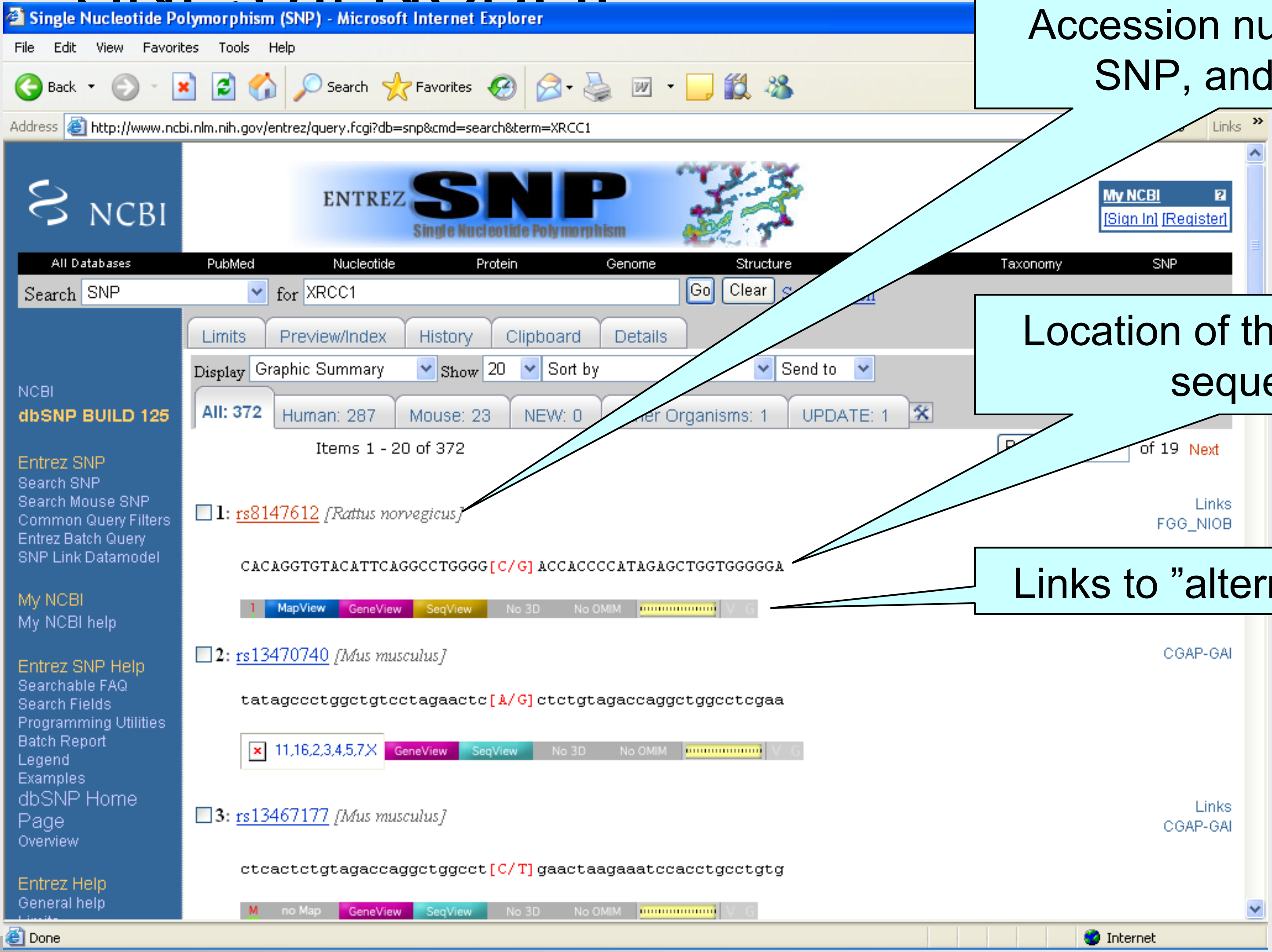
NCBI Bookshelf

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.

Select "SNP" database!

Free text query goes to here.

SNPs at NCBI II



http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp&cmd=search&term=XRCC1

Search SNP for XRCC1

Display Graphic Summary Show 20 Sort by

All: 372 | Human: 287 | Mouse: 23 | NEW: 0 | ...er Organisms: 1 | UPDATE: 1

Items 1 - 20 of 372 ... of 19

☐1: rs8147612 *[Rattus norvegicus]* Links FGG_NIOB

CACAGGTGTACATTCAGGCCTGGGG[C/G]ACCACCCCATAGAGCTGGTGGGGGA

☐2: rs13470740 *[Mus musculus]* CGAP-GAI

tatagccctggctgtcctagaactc[A/G]ctctgtagaccaggctggcctcgaa

11,16,2,3,4,5,7,X

☐3: rs13467177 *[Mus musculus]* Links CGAP-GAI

ctcactctgtagaccaggctggcct[C/T]gaactaagaaatccacctgcctgtg

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

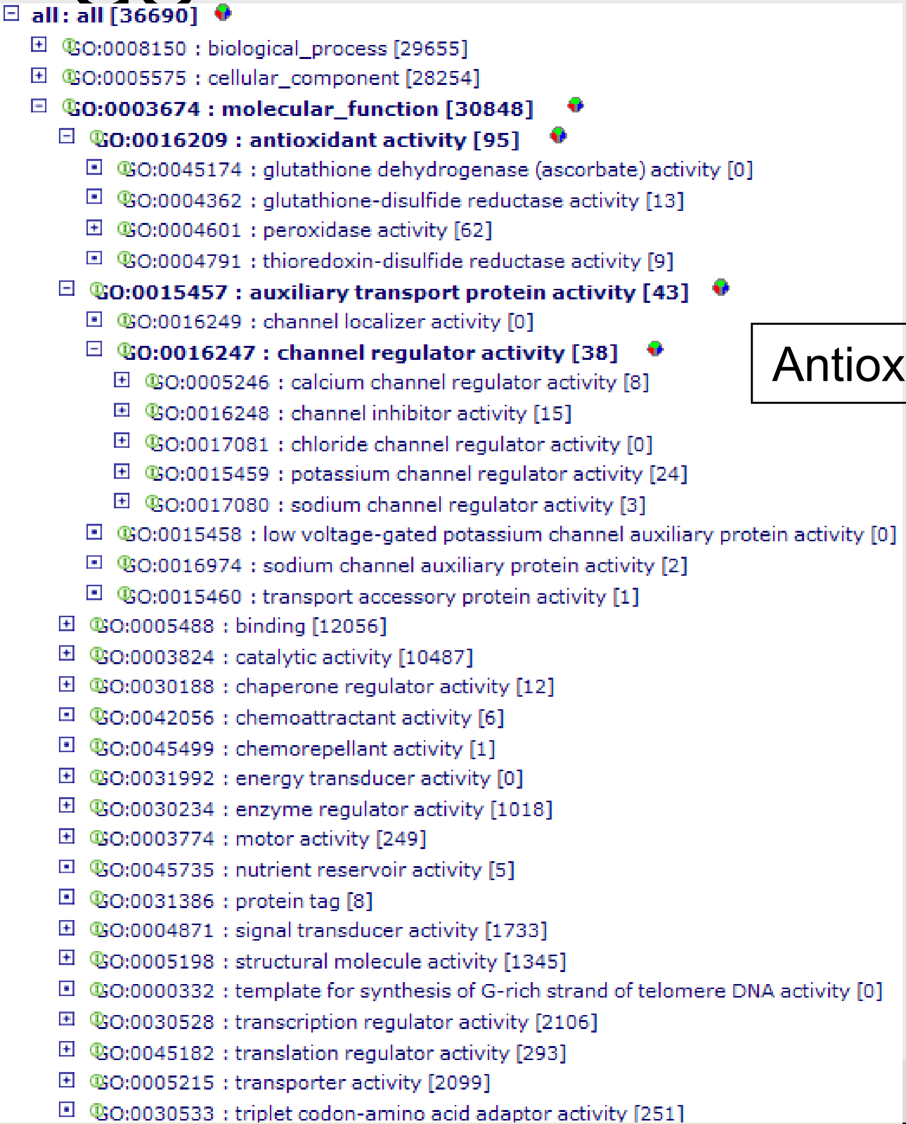
Gene Ontology (GO)

- **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]

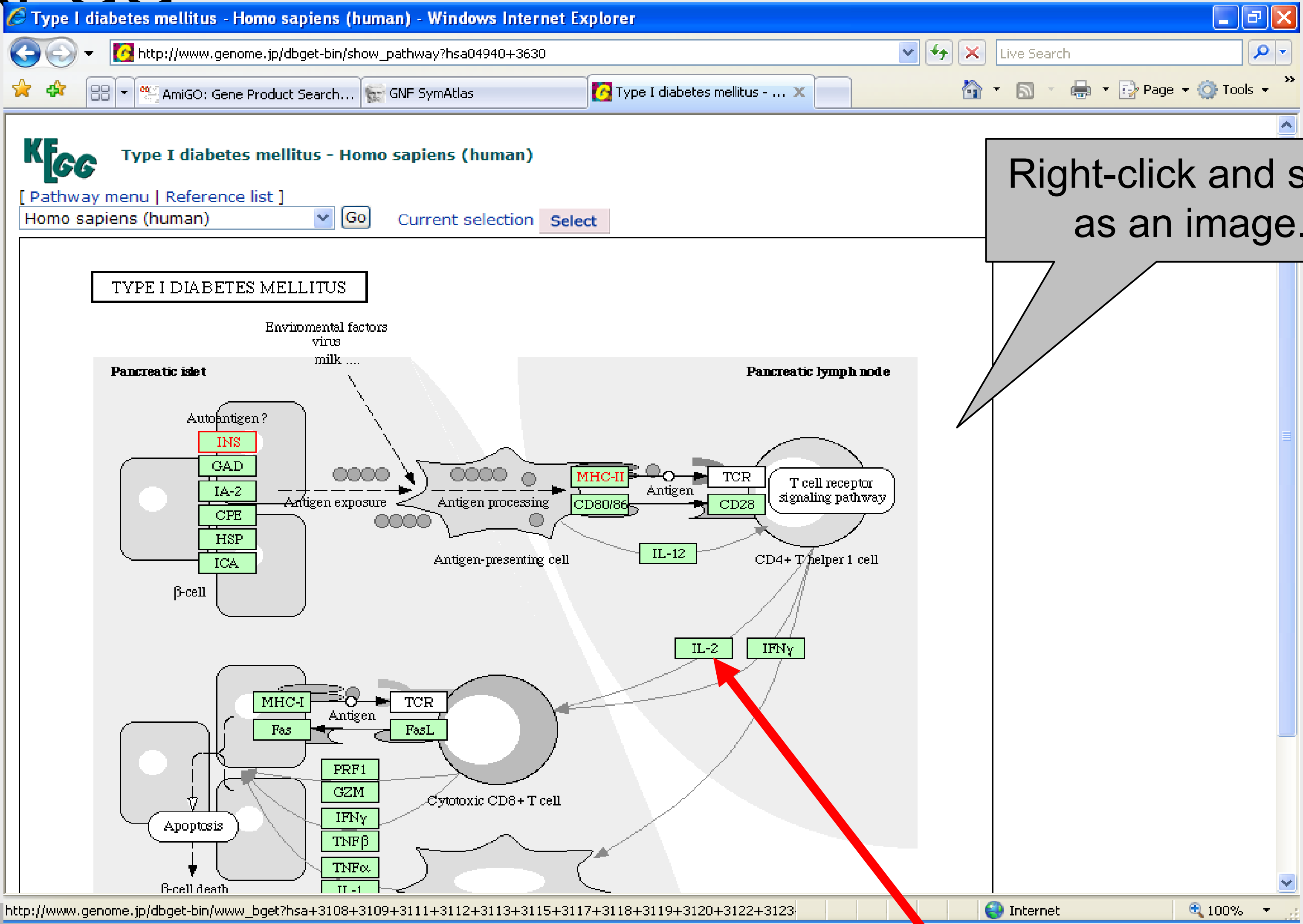
- Molecular function
  - Antioxidant activity
  - Auxiliary transport protein
    - Channel regulatory activity

KEGG

KEGG

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

KEGG



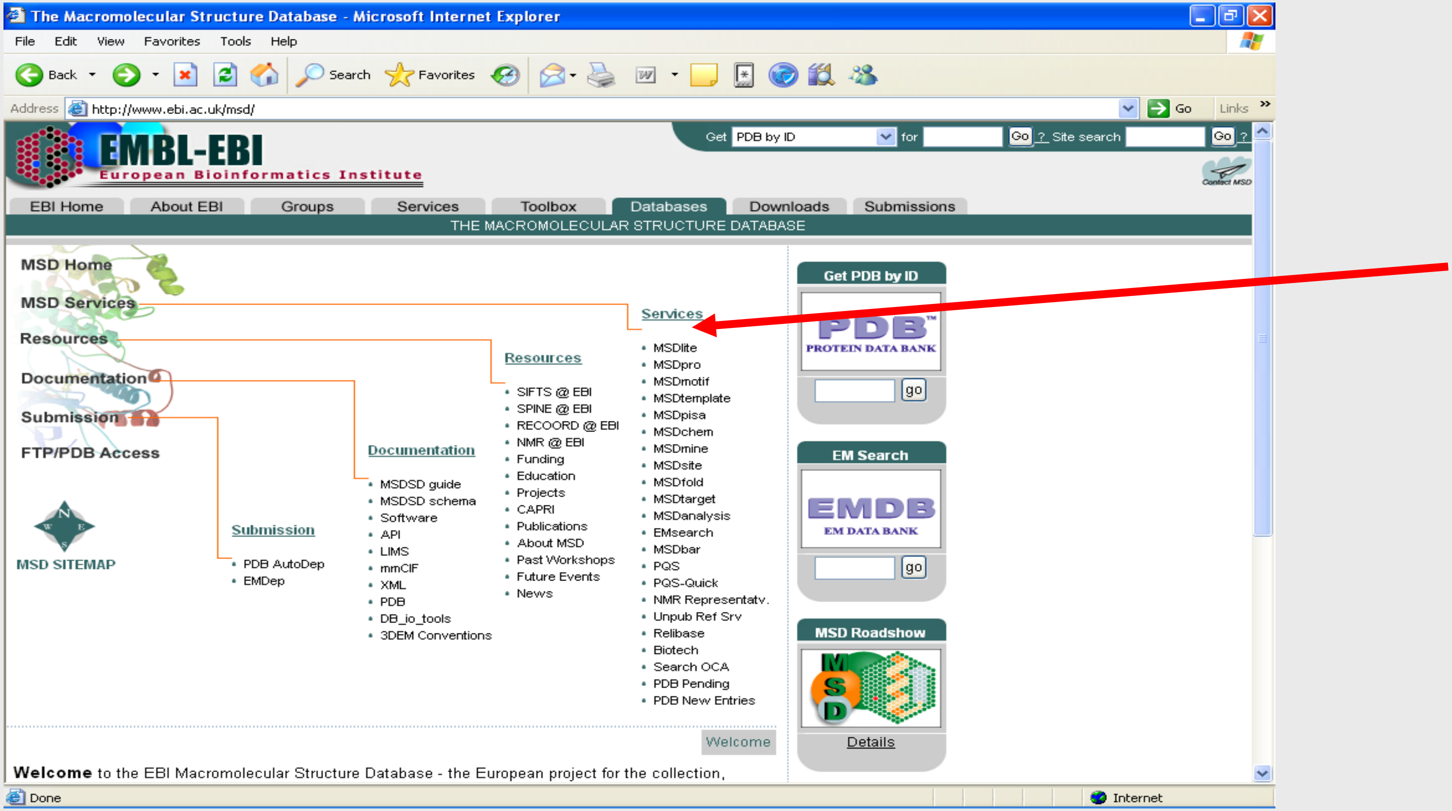
Right-click and save as an image.

Structural databases

PDB and MSD

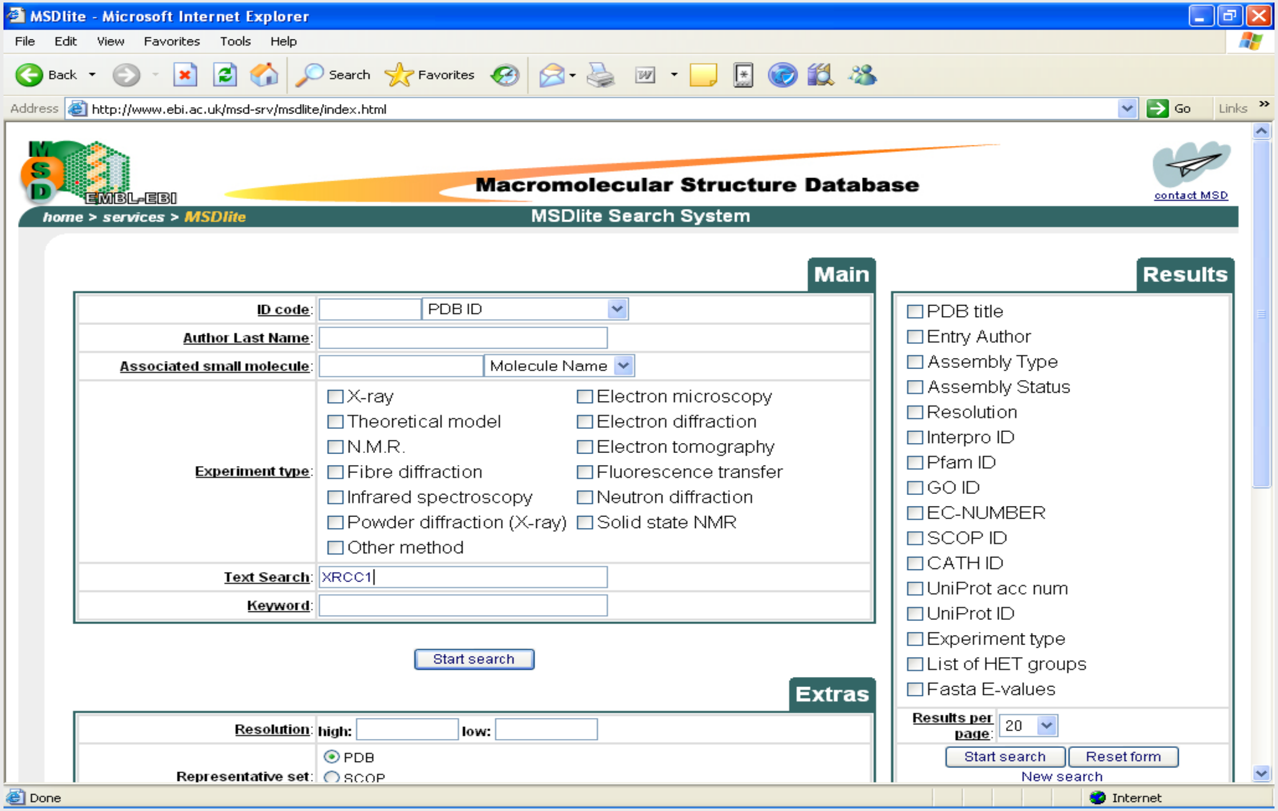
- **PDB contains structures of biological macromolecules.**
  - Mainly proteins, but also DNA and RNA structures
- **MSD is also a collection of biological structures, but it extends the PDB data format, and circumvents some problems.**

MSD 1/5



Welcome to the EBI Macromolecular Structure Database - the European project for the collection,

MSD 2/5



http://www.ebi.ac.uk/msd-srv/msdlite/index.html

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: XRCC1

Keyword:

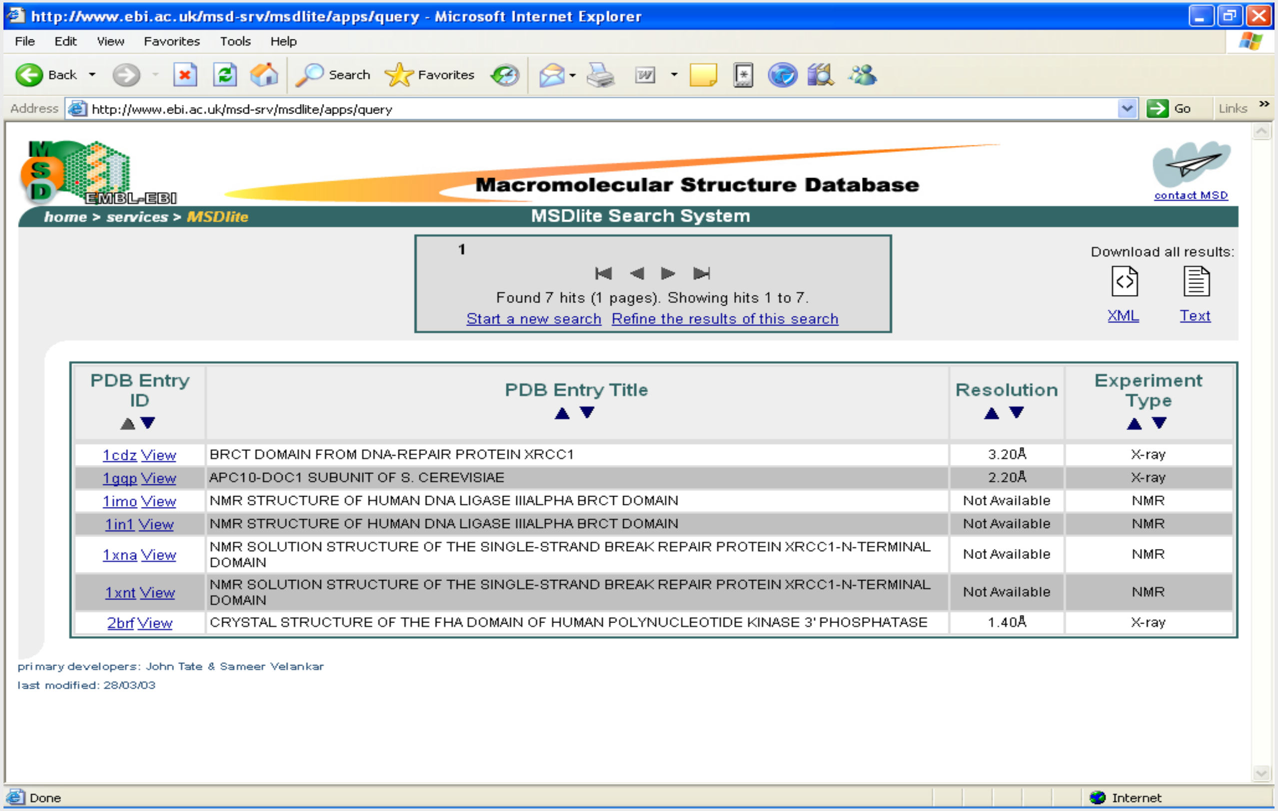
Resolution: high: low:

Representative set: ● PDB ○ SCOP

- ☐ 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: 20

MSD 3/5



Found 7 hits (1 pages). Showing hits 1 to 7.

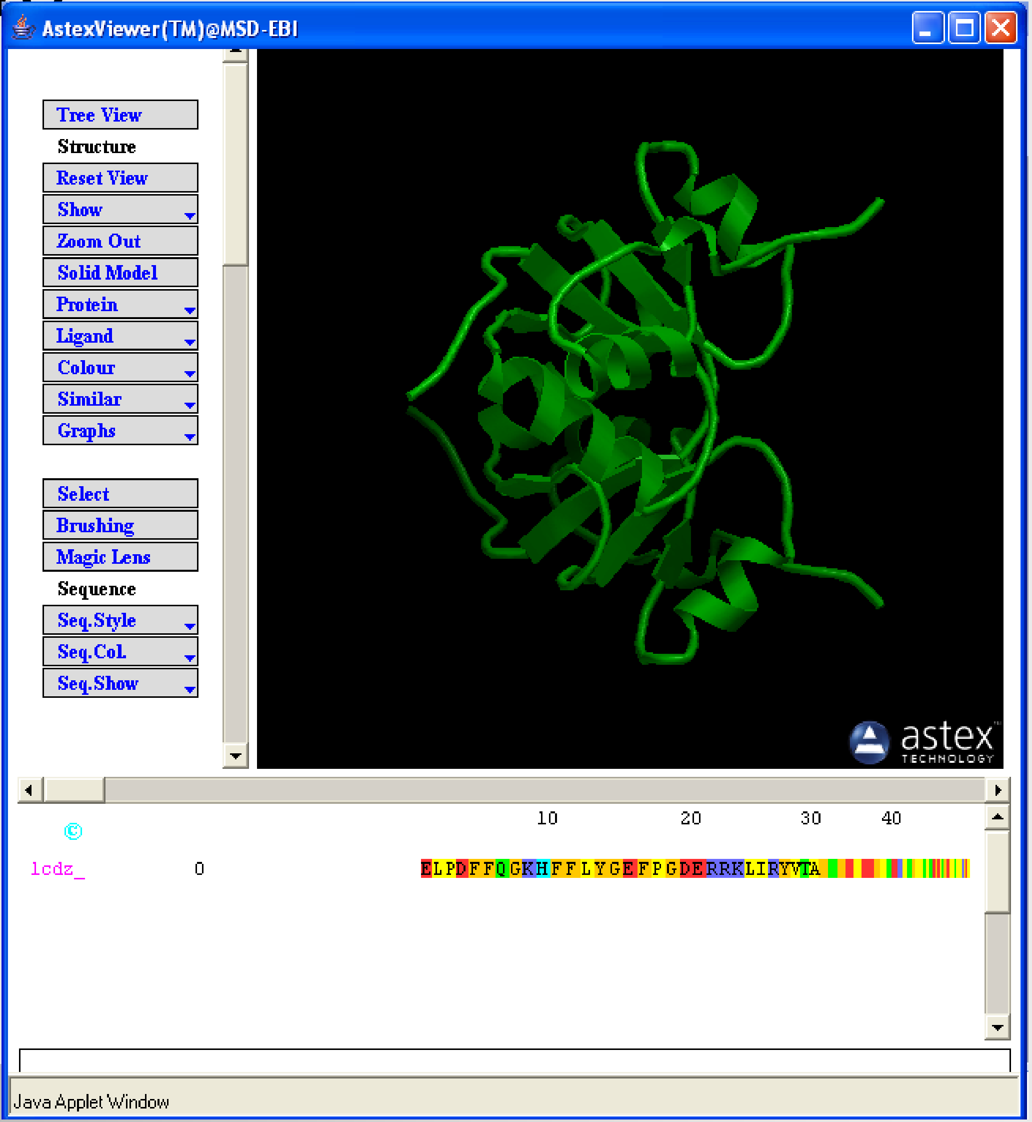
Start a new search Refine the results of this search

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

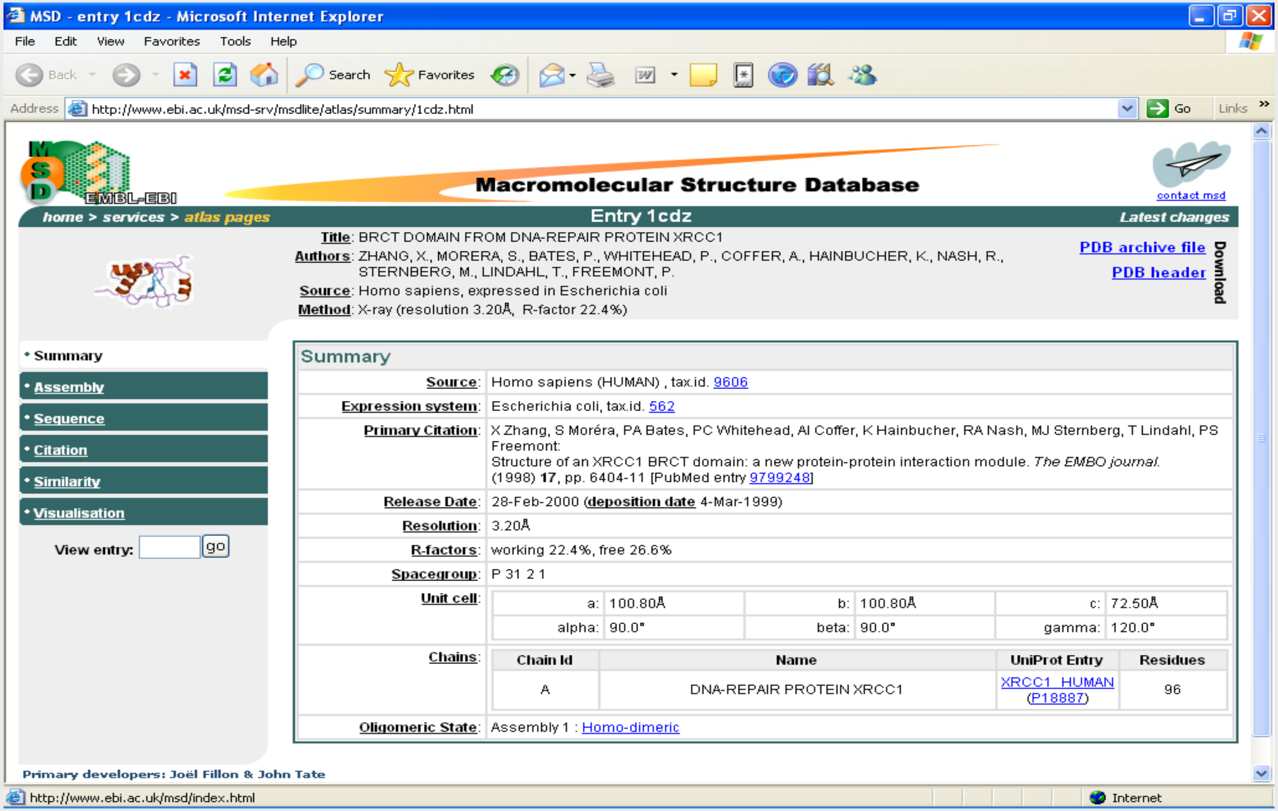
last modified: 28/03/03

MSD 4/5



```
1cdz_        0        ELPDFFQGKHFFLYGEFPGDERRKLIRYVTA
```

MSD 5/5



http://www.ebi.ac.uk/msd-srv/msdlite/atlas/summary/1cdz.html

MSD EMBL-EBI

Macromolecular Structure Database

contact msd

home > services > atlas pages

Entry 1cdz

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%)

Download
PDB archive file
PDB header

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

View entry: go

Summary

Source:	Homo sapiens (HUMAN) , tax.id. 9606
Expression system:	Escherichia coli, tax.id. 562
Primary Citation:	X Zhang, S Moréra, PA Bates, PC Whitehead, AI Coffer, 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%
Spacegroup:	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

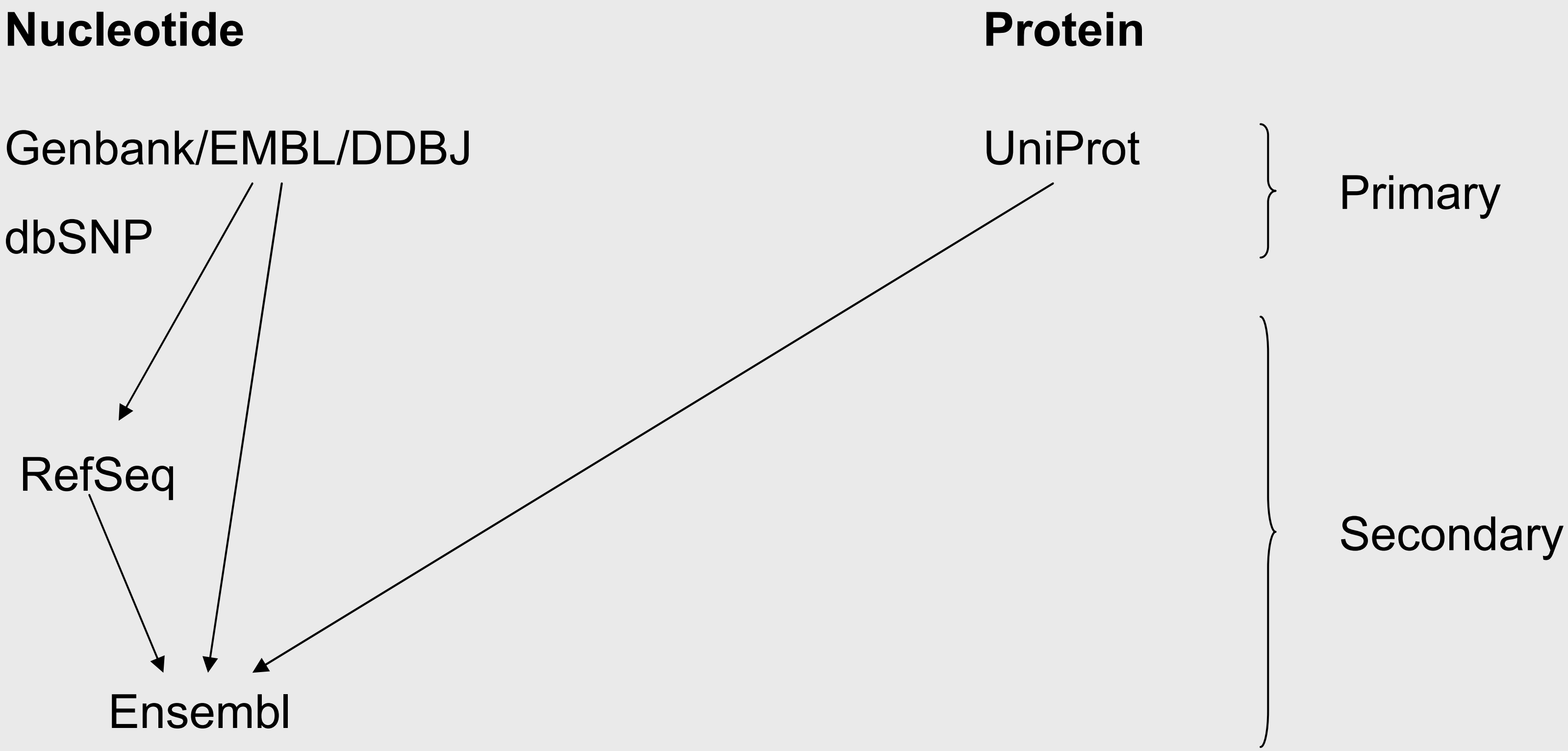
Integrating databases

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

UniProt

Primary

RefSeq

Ensembl

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

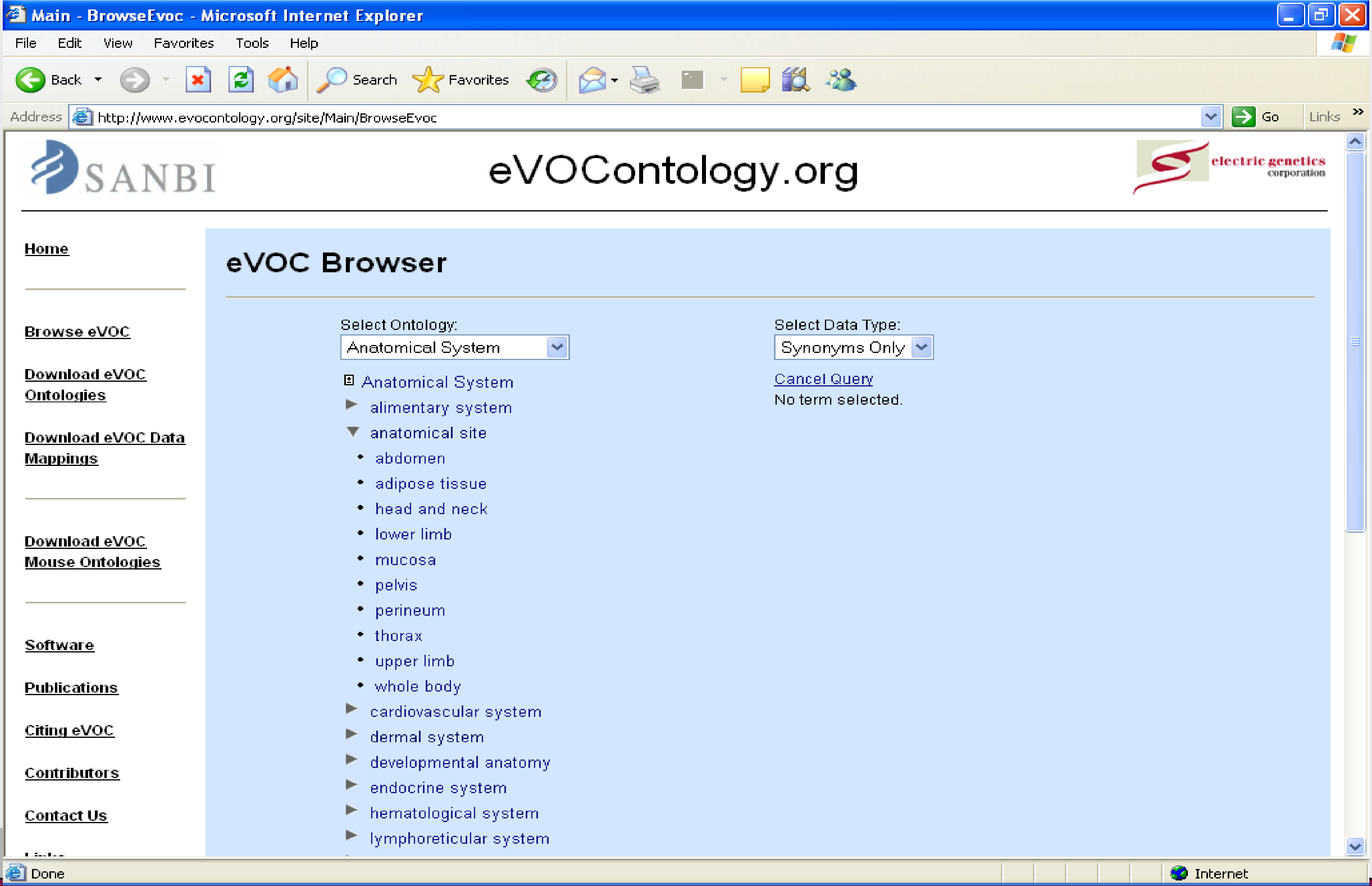
- **Differences in terminology**
  - Vector
    - A line with a direction (math.)
    - Carrier of an infectious agent (biol., med.)
    - Virus or DNA molecule used for transferring genetic material to or from cells (biol.)

    - 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



http://www.evocontology.org/site/Main/BrowseEvoc

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

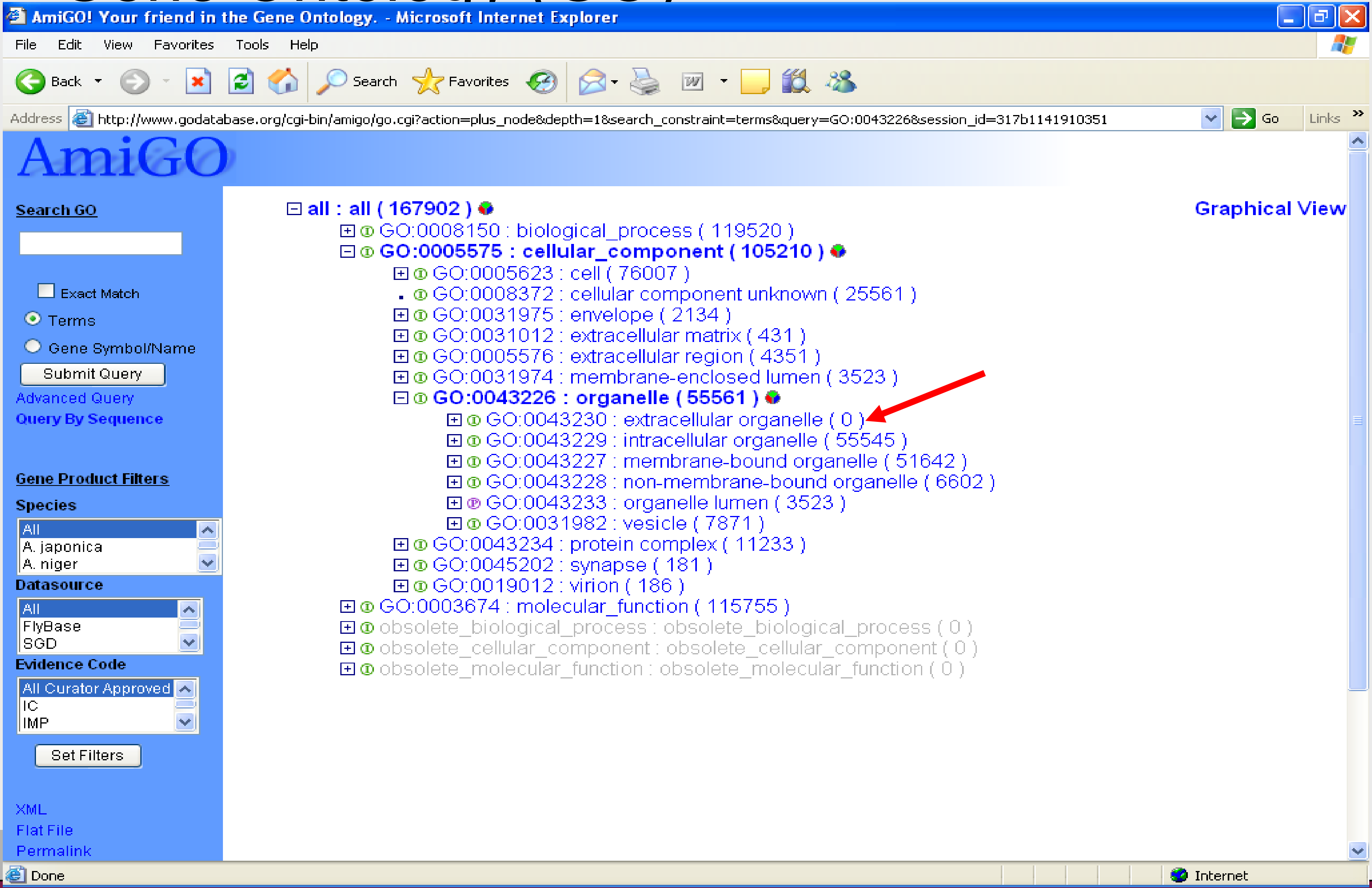
Select Data Type: Synonyms Only

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

Gene Ontology (GO)



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

☐ Exact Match
● Terms
○ Gene Symbol/Name

Species: All, A. japonica, A. niger

Datasource: All, FlyBase, SGD

Evidence Code: All Curator Approved, IC, IMP

- **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)