

Data mining in Ensembl with BioMart Worked Example

The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs? Do they have any functions predicted by Interpro?

What are their cDNA sequences?

STEP 1:
Go to the Ensembl
Main Page
(www.ensembl.org)

The screenshot shows the Ensembl homepage for release 43 (Feb 2007). A yellow callout box labeled "STEP 1:" points to the top navigation bar with the text "Go to the Ensembl Main Page (www.ensembl.org)". Another yellow callout box labeled "STEP 2:" points to the "Ensembl tools" section, specifically to the "Mine Ensembl with BioMart" link. The page layout includes a left sidebar with navigation links, a main content area with search and tool options, and a right sidebar with popular and more genomes. A footer at the bottom contains copyright information and a public use disclaimer.

Ensembl
Ensembl release 43 - Feb 2007

Search Ensembl
Search: for
e.g. mouse chromosome 2 or rat X:10000..20000 or human gene BRCA2

Ensembl tools

- Start a sequence search →
Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA
- Mine Ensembl with BioMart →
Cross-reference Ensembl datasets with BioMart, a powerful data-mining tool.
- Customise Your Ensembl →
Register with Ensembl to bookmark your favourite pages, customise your home page and much more!
- Fetch data with the Ensembl API →
Learn how to extract data from the public Ensembl database with this tutorial.

Ensembl 43 **Pre! species**

Popular genomes

- Homo sapiens**
NCBI 36 | Vega
- Mus musculus**
NCBI m36 | Vega
- Danio rerio**
Zf6 | Vega

More genomes

- ▶ **Aedes aegypti** AsegL1
- ▶ **Anopheles gambiae** AgamP3
- ▶ **Bos taurus** Blau_3.1 **UPDATED!**
- ▶ **Caenorhabditis elegans** WS160
- ▶ **Canis familiaris** Canfam 2.0
- ▶ **Cavia porcellus** cavPol2 **NEW!**
- ▶ **Ciona intestinalis** JGI12
- ▶ **Ciona savignyi** CSAV 2.0
- ▶ **Dasyatis novemcinctus** ARMA
- ▶ **Drosophila melanogaster** BDGP 4.3
- ▶ **Echinops telfairi** TENREC
- ▶ **Erinaceus europaeus** eEur1 **NEW!**
- ▶ **Felis catus** CAT **NEW!**
- ▶ **Gallus gallus** WASHUC2
- ▶ **Gasterosteus aculeatus** BROAD S1
- ▶ **Loxodonta africana** BROAD E1
- ▶ **Macaca mulatta** MMUL 1.0
- ▶ **Monodelphis domestica** MonDom 4.0
- ▶ **Ornithorhynchus anatinus** Oana5.0
- ▶ **Oryctolagus cuniculus** RABBIT
- ▶ **Oryzias latipes** HsR
- ▶ **Pan troglodytes** PanTro 2.1
- ▶ **Rattus norvegicus** RnOSC 3.4
- ▶ **Saccharomyces cerevisiae** SOD1.01
- ▶ **Takifugu rubripes** Fugu 4.0
- ▶ **Tetraodon nigroviridis** TETRAODON 7
- ▶ **Tupaia belangeri** tupBel1 **NEW!**
- ▶ **Xenopus tropicalis** JGI 4.1

Other pre-build species are available in [Ensembl Pre! →](#)
[Log in](#) to customise this list [Register](#)

Headlines: Release 43 (February 2006)

- Cow assembly and genebuild (*Bos taurus*)
- New species - Cat (*Felis catus*)
- New species - Tree shrew (*Tupaia belangeri*)
- New species - Hedgehog (*Erinaceus europaeus*)
- New species - Guinea Pig (*Cavia porcellus*)

More news...
[Log in](#) to see customised news - [Register](#)

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other Ensembl websites

- ▶ archive - past releases of Ensembl
- ▶ VEGA - Vertebrate Genome Annotation
- ▶ Ensembl Pre! - pre-release species
- ▶ EBI Genome Reviews database - mainly archaea and bacteria
- ▶ Trace server

Other sites using Ensembl software...

© 2007 [WTSI](#) / [EBI](#) Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

New

Help

Count

Results

» Dataset:

Homo sapiens genes (NCBI36)

» Attributes (Features)

Ensembl Gene ID

Ensembl Transcript ID

» Filters

[None selected]

» Dataset:

[None Selected]

Database: Ensembl 43

Dataset: Homo sapiens genes (NCBI36)

Choose a **Dataset** above, then use the left panel to navigate through the **Attributes** and **Filters** making your selections in this main panel. To preview the results click the **Results** button in the top panel.

[Mini Tutorial](#)

biomart version 0.5

STEP 3:

Select the primary datasets:
Ensembl genes (version 43)
and the species of interest
(*Homo sapiens*)

HOME · BLAST · BIOMART · SITEMAP · HELP

New

Help

Count

Results

» Dataset:

Homo sapiens genes (NCBI36)

» Attributes (Features)

Ensembl Gene ID

Ensembl Transcript ID

» Filters

[None selected]

» Dataset:

[None Selected]

Features

Homologs

Structures

Sequences

SNPs

REGION:

GENE:

EXPRESSION:

PROTEIN:

GENOMIC FEATURES:

biomart version 0.5

STEP 4:

Click on 'Attributes'
to select output
options (i.e. what we
would like to know
about our geneset).

HOME - E

New Help Count Results

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Features)
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description

» Filters
[None selected]

» Dataset:
[None Selected]

EXPRESSION:

PROTEIN:

Family Attributes
☐ Family Description ☐ Ensembl Family ID

Domain Attributes (max 1)
☐ Prosite ID ☐ PFAM ID
☐ PRINTS ID

Interpro Attributes
☒ Interpro Short Description ☐ Interpro ID
☐ Interpro Description

Transmembrane and Signal Domain Attributes
☐ Transmembrane domain ☐ Signal domain

GENOMIC FEATURES:

biomart version 0.5

STEP 5:
Expand the 'PROTEIN'
panel and
Select 'Interpro Short
Description'

New Help Count Results

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Features)
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description

» Filters
[None selected]

» Dataset:
[None Selected]

Features Homologs
Structures Sequences
SNPs

REGION:

GENE:

Ensembl Attributes
☒ Ensembl Gene ID ☐ Ensembl Peptide length
☒ Ensembl Transcript ID ☐ Transcript count
☐ Ensembl Peptide ID ☐ % GC content
☐ External Gene ID ☐ Description
☐ External Gene DB ☐ Biotype
☐ Ensembl CDS length ☐ Source
☐ Ensembl cDNA length ☐ Status

GO Attributes
☐ GO ID ☐ GO evidence code

biomart version 0.5

STEP 6:
Expand the 'GENE'
option in the
'Features' page.

STEP 7:
Select, along with the
default options,
'External Gene ID'.
Scroll down to select:
'EntrezGene ID' and
'Mim Gene
Accession' (this is
the OMIM ID).

HOME · BLAST · BIOMART · SITEMAP [HELP](#)

New **Help** **Count** **Results**

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

REGION:

GENE:

GENE ONTOLOGY:

EXPRESSION:

MULTI SPECIES COMPARISONS:

PROTEIN:

SNP:

biomart version 0.5

STEP 8:
View geneset options by clicking on 'Filters' on the left. Click on the '+' in front of 'REGION' to expand the choices.

New **Help** **Count** **Results**

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description

» **Filters**
Chromosome: X
Start : q28
End : q28

» **Dataset:**
[None Selected]

REGION:

☒ Chromosome X

☐ Base pair
Start 1
End 10000000

☒ Band
Start q28
End q28

☐ Marker
Start
End

☐ Encode type manual_picks

biomart version 0.5

STEP 9:
Select 'Chromosome X'

STEP 10:
Select 'Band Start q28' and 'End q28'

HOME · BLAST · BIOMART · SITEMAP · HELP

New Help **Count** Results

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes (Features)**
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description

» **Filters**
Chromosome: X
Start : q28
End : q28
with Disease association: Only

» **Dataset:**
[None Selected]

☐ Encode region 11:115962316:116462315

☐ In encode region ☒ Only ☐ Excluded

☒ GENE:
☒ ID LIST FILTERS: with Disease association ☒ Only ☐ Excluded

☐ ID list limit Ensembl Gene ID(s)

☐ Transcript count >=

☐ Entries with a 5' UTR ☒ Only ☐ Excluded

biomart version 0.5

STEP 11:
Expand the
'GENE' panel and
choose 'with
Disease
Association only'.

The filters have determined our gene set.
Click 'Count' (at the top) to see how many
genes have passed these filters.

STEP 12:
Click 'RESULTS' at the top
to preview the output.

HOME · BLAST · BIOMART · SITEMAP · HELP

New Help Count **Results**

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes (Features)**
Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description
EntrezGene ID
Mim Gene Accession

» **Filters**
Chromosome: X
Start : q28
End : q28
with Disease association: Only

Display maximum 10 rows as HTML

Export all results to File

Ensembl Gene ID	Ensembl Transcript ID	Interpro Short Description	EntrezGene ID	M
ENSG00000155966	ENST00000370460	AF-4	2334	3
ENSG00000155966	ENST00000286437	AF-4	2334	3
ENSG00000010404	ENST00000340855	Sulfatase	3423	3
ENSG00000013619	ENST00000370401	PRO_rich	0046	3
ENSG00000013619	ENST00000370401	PRO_rich	28030	3
ENSG00000013619	ENST00000370401	PRO_rich	0818	3
ENSG00000013619	ENST00000262858	PRO_rich	046	3
ENSG00000013619	ENST00000262858	PRO_rich	030	3
ENSG00000013619	ENST00000262858	PRO_rich	018	3
ENSG00000171100	ENST00000306167	TYR_phosphatase		3

Note the summary of
selected attributes and
filters (i.e. output
options and gene set
determinants)

STEP 13:
To view the complete table,
change the option to 'Export all
results to Browser' and click
'Go'.

Result Table 1

Ensembl Gene ID	Ensembl Transcript ID	Interpro Short Description	EntrezGene ID	Mim Gene Accession
ENSG00000155966	ENST00000370460	AF-4	2334	309548
ENSG00000155966	ENST00000286437	AF-4	2334	309548
ENSG00000010404	ENST00000340855	Sulfatase	3423	309900
ENSG00000013619	ENST00000370401	PRO_rich	10046	300120
ENSG00000013619	ENST00000370401	PRO_rich	728030	300120
ENSG00000013619	ENST00000370401	PRO_rich	730818	300120
ENSG00000013619	ENST00000262858	PRO_rich	10046	300120
ENSG00000013619	ENST00000262858	PRO_rich	728030	300120
ENSG00000013619	ENST00000262858	PRO_rich	730818	300120
ENSG000000171100	ENST00000306167	TYR_phosphatase	4534	300415
ENSG000000171100	ENST00000306167	GRAM	4534	300415
ENSG000000171100	ENST00000306167	Myotub-related	4534	300415
ENSG000000147383	ENST00000370274	Epimerase_Dh	50814	300275
ENSG000000147383	ENST00000370274	3Beta_HSD	50814	300275
ENSG000000147383	ENST00000370274	Polysac_CapD	50814	300275
ENSG000000147383	ENST00000370274	Male_sterile_C	50814	300275
ENSG000000130821	ENST00000330048	Na/ntran_symport	6535	300036
ENSG000000130821	ENST00000330048	Crt_transporter	6535	300036
ENSG000000130821	ENST00000253122	Na/ntran_symport	6535	300036
ENSG000000130821	ENST00000253122	Crt_transporter	6535	300036
ENSG000000185825	ENST00000345046	Macro_scav_rcpt	10134	300398
ENSG000000185825	ENST00000345046	Bap31	10134	300398
ENSG000000185825	ENST00000370133	Macro_scav_rcpt	10134	300398
ENSG000000185825	ENST00000370133	Bap31	10134	300398
ENSG000000101986	ENST00000218104	ABC_transp_like	215	300371
ENSG000000101986	ENST00000218104	ABC_Ald_N	215	300371
ENSG000000101986	ENST00000218104	ABC_transp_like	642762	300371
ENSG000000101986	ENST00000218104	ABC_Ald_N	642762	300371
ENSG000000198910	ENST00000370060	FnIII_subd	3897	308840
ENSG000000198910	ENST00000370060	VEGFR_N	3897	308840
ENSG000000198910	ENST00000370060	FN_III	3897	308840
ENSG000000198910	ENST00000370060	Ig_I-set	3897	308840
ENSG000000198910	ENST00000370060	Ig_V-set	3897	308840
ENSG000000198910	ENST00000370060	Ig	3897	308840
ENSG000000198910	ENST00000370060	Ig-like	3897	308840
ENSG000000198910	ENST00000361699	FnIII_subd	3897	308840
ENSG000000198910	ENST00000361699	VEGFR_N	3897	308840
ENSG000000198910	ENST00000361699	FN_III	3897	308840

STEP 14:

Click on
'Attributes'.

The screenshot shows the Biomart interface with the 'Attributes' panel selected in the left sidebar. The main content area displays various attribute categories: Features (selected), Homologs, Structures, Sequences, and SNPs. Under 'Features', there are sub-sections for REGION, GENE, Ensembl Attributes, and GO Attributes. The 'Ensembl Attributes' section is expanded, showing a list of attributes with checkboxes. The 'Dataset' section on the left shows 'Homo sapiens genes (NCBI36)' and 'Attributes (Features)' selected. The 'Filters' section shows 'Chromosome: X', 'Start: q28', 'End: q28', and 'with Disease association: Only'. The 'Dataset' section shows '[None Selected]'. The footer indicates 'biomart version 0.5' and a copyright notice for 2007 WTSI / EBI.

HOME · BLAST · BIOMART · SITEMAP · HELP

Count Results

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Features)

Ensembl Gene ID
Ensembl Transcript ID
Interpro Short Description
EntrezGene ID
Mim Gene Accession

» Filters

Chromosome: X
Start : q28
End : q28
with Disease association: Only

» Dataset:
[None Selected]

Features Homologs
Structures Sequences
SNPs

REGION:

GENE:

Ensembl Attributes

☒ Ensembl Gene ID ☐ Ensembl Peptide length
☒ Ensembl Transcript ID ☐ Transcript count
☐ Ensembl Peptide ID ☐ % GC content
☐ External Gene ID ☐ Description
☐ External Gene DB ☐ Biotype
☐ Ensembl CDS length ☐ Source
☐ Ensembl cDNA length ☐ Status

GO Attributes

☐ GO ID ☐ GO evidence code

biomart version 0.5

© 2007 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

STEP 15:
Select 'Sequences' and
expand the
'SEQUENCES' panel.

The screenshot shows the Biomart interface with the 'Sequences' panel selected in the left sidebar. The main content area displays various attribute categories: Features, Homologs, Structures, Sequences (selected), and SNPs. Under 'Sequences', there is a sub-section for SEQUENCES, which is expanded to show a list of sequence types with checkboxes. The 'Dataset' section on the left shows 'Homo sapiens genes (NCBI36)' and 'Attributes (Sequences)' selected. The 'Filters' section shows 'Chromosome: X', 'Start: q28', 'End: q28', and 'with Disease association: Only'. The 'Dataset' section shows '[None Selected]'. The footer indicates 'biomart version 0.5' and a copyright notice for 2007 WTSI / EBI.

HOME · BLAST · BIOMART · SITEMAP · HELP

New Help Count Results

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Sequences)

Chromosome
Ensembl Gene ID
Biotype
cDNA sequences

» Filters

Chromosome: X
Start : q28
End : q28
with Disease association: Only

» Dataset:
[None Selected]

Features Homologs
Structures Sequences
SNPs

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 (Transcript)
Exon sequences (Gene)
☒ cDNA sequences
Coding sequence
Peptide

biomart version 0.5

© 2007 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

STEP 16:
Select 'cDNA
sequences'

STEP 17:
Expand the 'Header Information' panel to view the default options. (Chromosome, Ensembl Gene ID and Biotype are selected).

STEP 18:
Click on 'Results'.

© 2007 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

STEP 19:
To view all the sequences, export to browser.

© 2007 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

cDNA 2