



Session 3

---

**Biomart**

**Lecture**

**- Michael P Schroeder -**

## **Where to find the gene info?**

This lecture is about how to retrieve basic genomic information like, gene names, locus, all the ids and more annotation

- ! Watch out to the exclamation marks throughout the presentation, they hint at hidden problems in every day work in genetics

## What is Biomart?



## A Database:

A solution to provide big data sets.

The data is accessible through a website..

Cosmic <http://www.sanger.ac.uk/genetics/CGP/cosmic/biomart/martview>

Ensembl <http://www.ensembl.org/biomart/martview>

Intogen <http://biomart.intogen.org/>

etc...

...and web services which allow programmatic access.

biomaRt (R packackage)

Gitools (Biomart importer)

## What is Biomart?



## A distributed system:

The data can be distributed on different servers (and locations) meanwhile they are connected to each other. All instances of biomart are connected to the central portal:

<http://central.biomart.org/>

# Biomart central

### IDENTIFIER SEARCH

  
  
Examples: KRAS, ENSG00000146648

### TOOLS

**Gene retrieval** | Variant retrieval | Sequence retrieval | ID converter

- Cancer genes
- Ensembl
- Ensembl Bacteria
- Ensembl Fungi
- Ensembl Metazoa
- Ensembl Plants
- Ensembl Protists
- Mouse Genome Informatics
- VEGA

### DATABASE SEARCH

Search by type | Search by organism | Search by database name (A-Z)

- ▶ Genome
- ▶ Gene annotation
- ▶ Protein sequence and structure
- ▶ Interaction and pathways
- ▶ Gene expression
- ▶ Cancer
- ▶ Model organism databases
- ▶ Other

### BIO MART CENTRAL PORTAL

Databases: 41

| Country        | Count |
|----------------|-------|
| Canada         | 1     |
| United States  | 8     |
| United Kingdom | 21    |
| Spain          | 1     |
| France         | 4     |
| Italy          | 1     |
| China          | 1     |
| South Korea    | 1     |
| Japan          | 1     |
| Peru           | 1     |
| Chile          | 1     |

Click on the map to view the list of databases

# Biomart central

## IDENTIFIER SEARCH

Examples: KRAS, ENSG00000146648

## TOOLS

- Gene retrieval
- Variant retrieval
- Sequence retrieval
- ID converter

- Cancer genes
- Ensembl ←
- Ensembl Bacteria
- Ensembl Fungi
- Ensembl Metazoa
- Ensembl Plants
- Ensembl Protists
- Mouse Genome Informatics
- VEGA

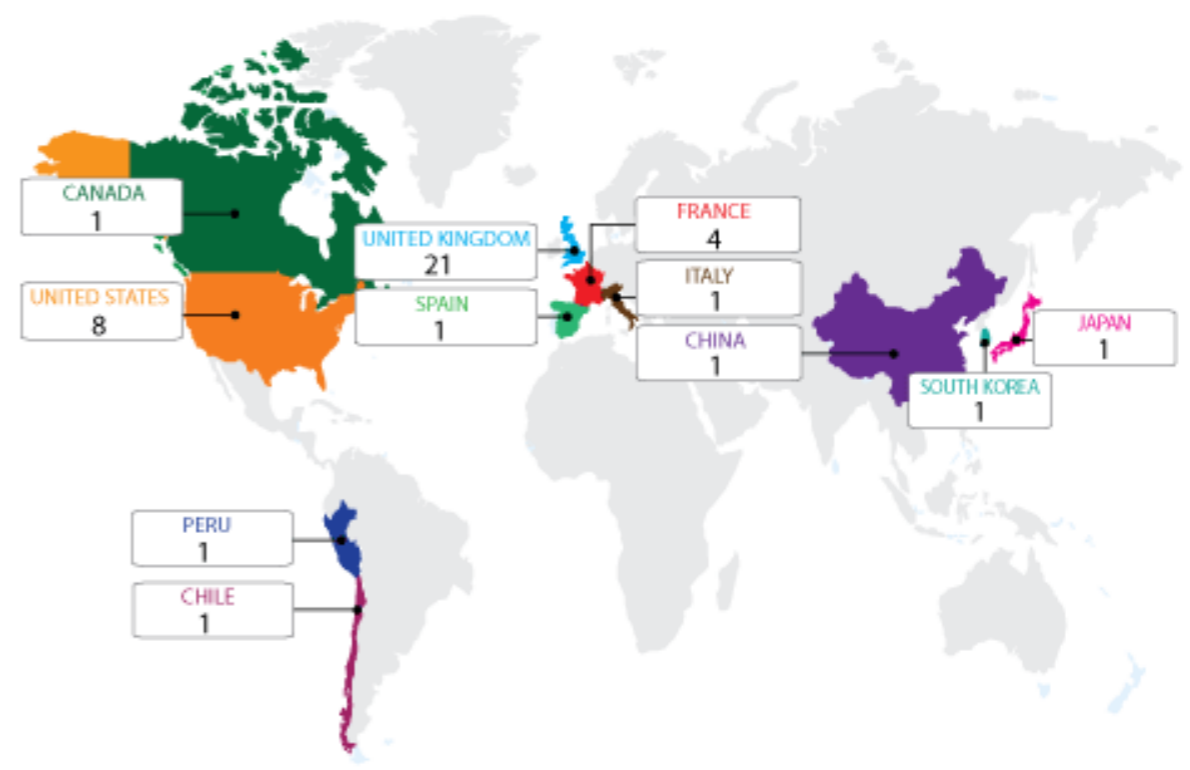
## DATABASE SEARCH

- Search by type
- Search by organism
- Search by database name (A-Z)

- ▶ Genome
- ▶ Gene annotation
- ▶ Protein sequence and structure
- ▶ Interaction and pathways
- ▶ Gene expression
- ▶ Cancer
- ▶ Model organism databases
- ▶ Other

## BIO MART CENTRAL PORTAL

Databases: 41



Click on the map to view the list of databases

## BioMart Central Portal

Home > Gene retrieval

VIEW:

### 1. SELECT DATASETS

- Homo sapiens genes (GRCh37.p3)
- Mus musculus genes (NCBIM37)
- Rattus norvegicus genes (RGSC3.4)
- Danio rerio genes (Zv9)
- Gallus gallus genes (WASHUC2)
- Drosophila melanogaster genes (BDGP5.25)
- Caenorhabditis elegans genes (WS220)
- Ailuropoda melanoleuca genes (ailMe1)
- Anolis carolinensis genes (AnoCar2.0)
- Bos taurus genes (Btau\_4.0)
- Callithrix jacchus genes (calJac3)
- Canis familiaris genes (CanFam\_2.0)
- Cavia porcellus genes (cavPor3)
- Choloepus hoffmanni genes (choHof1)
- Ciona intestinalis genes (JGI2)



### 2. RESTRICT SEARCH

Chromosome:

Gene Start (bp):

Gene End (bp):

Gene Biotype:

Gene Status:

Entries with IDs:

Go »

## BioMart Central Portal

[Home](#) > Gene retrieval

VIEW:

### 1. SELECT DATASETS

- Homo sapiens genes (GRCh37.p3)**
- Mus musculus genes (NCBIM37)
- Rattus norvegicus genes (RGSC3.4)
- Danio rerio genes (Zv9)
- Gallus gallus genes (WASHUC2)
- Drosophila melanogaster genes (BDGP5.25)
- Caenorhabditis elegans genes (WS220)
- Ailuropoda melanoleuca genes (ailMel1)
- Anolis carolinensis genes (AnoCar2.0)
- Bos taurus genes (Btau\_4.0)
- Callithrix jacchus genes (calJac3)
- Canis familiaris genes (CanFam\_2.0)
- Cavia porcellus genes (cavPor3)
- Choloepus hoffmanni genes (choHof1)
- Ciona intestinalis genes (JGI2)



### 2. RESTRICT SEARCH

Chromosome:

Gene Start (bp):

Gene End (bp):

Gene Biotype:

Gene Status:

Entries with IDs:

**Go »**



[Home](#) > [Gene retrieval](#)

Ensembl » [Homo Sapiens Genes \(GRCh37.P3\)](#) !

 [Bookmark](#)

 [REST / SOAP](#)

 [SPARQL](#)

 [Java](#)

 [Download data](#)

 [Back](#)

| Associated Gene Name ↕ | Ensembl Gene ID ↕               | Chromosome Name ↕ | Gene Start (bp) ↕ | Gene End (bp) ↕ | Strand ↕ | Band ↕ | Transcript count ↕ | Gene Biotype ↕ |
|------------------------|---------------------------------|-------------------|-------------------|-----------------|----------|--------|--------------------|----------------|
| <a href="#">CHL1</a>   | <a href="#">ENSG00000134121</a> | 3                 | 238279            | 451090          | 1        | p26.3  | 15                 | protein_coding |

Ensembl » Homo Sapiens Genes (GRCh37.P3) !

[Bookmark](#) | [REST / SOAP](#) | [SPARQL](#) | [Java](#) | [Download data](#) | [Back](#)

| Associated Gene Name | Ensembl Gene ID                 | Chromosome Name | Gene Start (bp) | Gene End (bp) | Strand | Band  | Transcript count | Gene Biotype   |
|----------------------|---------------------------------|-----------------|-----------------|---------------|--------|-------|------------------|----------------|
| <a href="#">CHL1</a> | <a href="#">ENSG00000134121</a> | 3               | 238279          | 451090        | 1      | p26.3 | 15               | protein_coding |

Powered by  bio:mart

GRCh37.P3 = **G**enome **R**eference **C**onsortium  
**h**uman, build **37**, patch **3**

Widely used:

CRCh37 a.k.a hg19

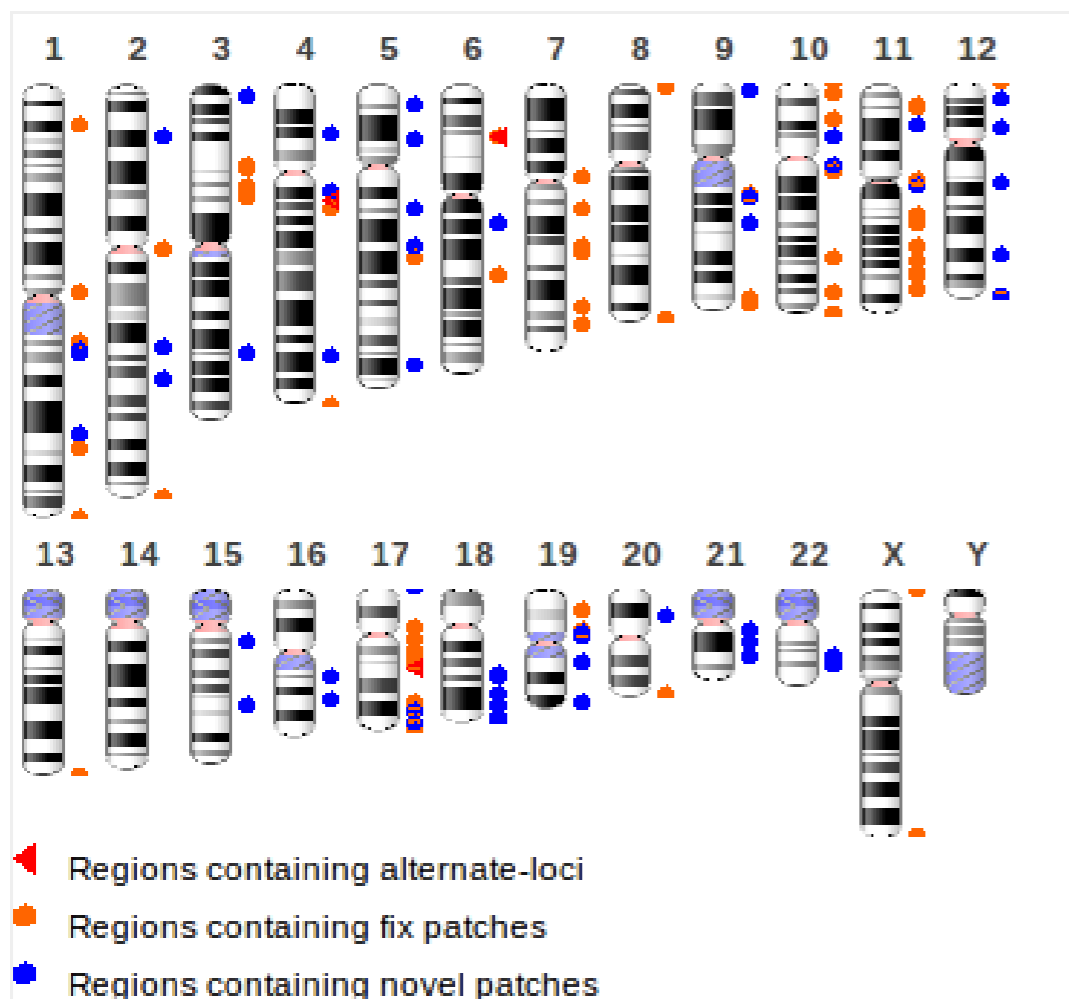
CRCh36 a.k.a hg18

# Genome Reference Consortium

[GRC Home](#)[Data](#)[Help](#)[Report an Issue](#)[Contact Us](#)[Credits](#)[Curators Only](#)[Human Overview](#)[Human Issues under Review](#)[Human Assembly Data](#)[Report a problem](#)

## Human Genome Overview

Information concerning the continuing improvement of the human genome.



An ideogram representation of the latest human assembly, GRCh37.p8 (not showing unplaced or unlocalized sequences).

The GRC is working hard to provide the best possible reference assembly for human. We do this by both generating multiple representations ( [alternate loci](#) ) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as [patches](#) . This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

### Getting Data

GRCh37 (Latest Major Release): [FTP](#)

GRCh37 patch release 8 (Latest Minor Release): [FTP](#)

Information on regions under review: [FTP](#)

We are planning to update the human reference assembly to GRCh38 in the summer of 2013. If you have questions or concerns about this [let us know](#) .

See our [blog](#) for more information on why we think this is important.

### Next assembly update

The next assembly update (patch release 9) will be a minor update (only patches) and will happen in Jul 2012

Ensembl » Homo Sapiens Genes (GRCh37.P3) !

[Bookmark](#) | [REST / SOAP](#) | [SPARQL](#) | [Java](#) | [Download data](#) | [Back](#)

| Associated Gene Name | Ensembl Gene ID | Chromosome Name | Gene Start (bp) | Gene End (bp) | Strand | Band  | Transcript count | Gene Biotype   |
|----------------------|-----------------|-----------------|-----------------|---------------|--------|-------|------------------|----------------|
| CHL1                 | ENSG00000134121 | 3               | 238279          | 451090        | 1      | p26.3 | 15               | protein_coding |

Powered by  bio:mart

Always use the id to work: ENSG00000xxxxxx

Gene names or gene symbols are very ambiguous!

Excel transforms gene symbols like MAR1 to dates: Mar-1



## Gene Symbol Report

### CHL1

|                            |  |
|----------------------------|--|
| Approved Symbol +          | CHL1   |
| Approved Name +            | cell adhesion molecule with homology to L1CAM (close homolog of L1)                                  |
| HGNC ID +                  | HGNC:1939  |
| Previous Symbols & Names + | "cell adhesion molecule with homology to L1CAM (close homologue of L1)"                              |
| Synonyms +                 | CALL, "cell adhesion molecule L1-like", FLJ44930, L1CAM2, MGC132578, "neural cell adhesion molecule" |
| Locus Type +               | gene with protein product  |
| Chromosomal Location +     | 3p26   |

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

## BioMart Central Portal

[Home](#) > [Gene retrieval](#)

Ensembl » [Homo Sapiens Genes \(GRCh37.P3\)](#) !

 [Bookmark](#)

 [REST / SOAP](#)

 [SPARQL](#)

 [Java](#)

 [Download data](#)

 [Back](#)

| Associated Gene Name | Ensembl Gene ID                 | Chromosome Name | Gene Start (bp) | Gene End (bp) | Strand | Band  | Transcript count | Gene Biotype   |
|----------------------|---------------------------------|-----------------|-----------------|---------------|--------|-------|------------------|----------------|
| <a href="#">CHL1</a> | <a href="#">ENSG00000134121</a> | 3               | 238279          | 451090        | 1      | p26.3 | 15               | protein_coding |

## Profile of gene CHL1

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors Login · Register

Human (GRCh37) Location: 3:238,279-451,090 Gene: CHL1

### Gene: CHL1 ENSG00000134121

**Description** cell adhesion molecule with homology to L1CAM (close homolog of L1) [Source:HGNC Symbol;Acc:1939]  
**Location** [Chromosome 3: 238,279-451,090](#) forward strand.  
**Transcripts** This gene has 15 transcripts

**Transcript and Gene level displays**  
In Ensembl we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domain annotation.
- Gene views which provide displays for data associated at the gene level such as orthologues, paralogues, regulatory regions and splice variants.

This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information you want using the menu at the left hand side of the page. To return to viewing gene level information click on the Gene tab in the menu bar at the top of the page.

### Gene summary [help](#)

**Name** [CHL1](#) (HGNC Symbol)  
**Synonyms** CALL, FLJ44930, L1CAM2, MGC132578 [To view all Ensembl genes linked to the name [click here](#).]  
**CCDS** This gene is a member of the Human CCDS set: [CCDS2556](#)  
**Gene type** Known protein coding  
**Prediction Method** Annotation for this gene includes both automatic annotation from Ensembl and [Havana](#) manual curation, see [article](#).  
**Alternative genes** This gene corresponds to the following database identifiers:  
**Havana gene:** [OTTHUMG00000090601](#) (version 6) [[view all locations](#)]

The genomic track displays the CHL1 gene structure on Chromosome 3. The x-axis represents genomic coordinates from 240.00 Kb to 440.00 Kb, with a total length of 232.81 Kb. The forward strand is indicated. Several transcripts are shown as horizontal bars with exons and introns: CHL1-001 (protein coding), CHL1-005 (processed transcript), CHL1-002 (protein coding), CHL1-006 (protein coding), and CHL1-007 (protein coding). The protein coding regions are highlighted in light green.

- Gene-based displays
    - Gene summary
    - Splice variants (15)
    - Supporting evidence
    - Sequence
    - External references
    - Regulation
  - Comparative Genomics
    - Genomic alignments
    - Gene Tree (image)
      - Gene Tree (text)
      - Gene Tree (alignment)
    - Orthologues (66)
    - Paralogues (13)
    - Protein families (4)
  - Phenotype
  - Genetic Variation
    - Variation Table
    - Variation Image
    - Structural Variation
  - External Data
    - Personal annotation
  - ID History
    - Gene history
- Configure this page  
Manage your data  
Export data  
Bookmark this page

# Ensembl

Search Ensembl, EBI or Sanger Institute

Jump from gene to location using tabs

The screenshot displays the Ensembl genome browser interface. At the top, the Ensembl logo is on the left, and navigation links for 'Home > Human', 'Login / Register', 'BLAST/BLAT', 'BioMart', and 'Docs & FAQs' are on the right. Below the header, the current location is shown as 'Location: 6:131,533,782-131,677,240' with tabs for 'Gene: AKAP7' and 'Transcript: AKAP7-001'. A search bar is located in the top right corner.

On the left side, there is a 'Location-based displays' menu with options: 'Whole genome', 'Chromosome summary', 'Region overview', 'Region in detail' (which is selected), 'Comparative Genomics' (with sub-options for 'Genomic alignments (35)', 'Multi-species comp. (39)', and 'Synteny (10)'), 'Genetic Variation' (with sub-option for 'Resequencing (6)'), 'Markers', and 'Export location data'. Below this menu are three options: 'Bookmark this page', 'Configure this page', and 'Add custom data to page'.

The main content area is titled 'Chromosome 6: 131,533,782-131,677,240'. It features a track for 'Assembly exceptions' for 'chromosome 6' and another for 'Assembly exceptions' for '6\_COX' and '6\_OBL'. Below these are navigation tabs: '« Region overview', 'Region in detail', and 'Genes & transcripts »'. The 'Region in detail' view shows a genomic map with 'Cortigs' (contigs) and 'Ensembl/Havana gene' tracks. A 1.00 Mb scale bar is shown above the contig track, with markers at 131.20 Mb, 131.50 Mb, and 131.80 Mb. The 'Forward strand' is indicated by an arrow. Genes shown include EPB41L2, AKAP7 (highlighted in green), RP1-2096S2, RP11-123H21.1, ARG1, CRSP3, and ENPP3.

Click and drag the mouse to recentre the display

Use the left-hand menus to navigate, export data and customise the page



# Ensembl

Search Ensembl, EBI or Sanger Institute

Jump from gene to location using tabs

The screenshot displays the Ensembl genome browser interface. At the top, the Ensembl logo is on the left, and navigation links for 'Home > Human', 'Login / Register', 'BLAST/BLAT', 'BioMart', and 'Docs & FAQs' are on the right. Below the header, the current location is shown as '6:131,533,782-131,677,240' with tabs for 'Gene: AKAP7' and 'Transcript: AKAP7-001'. A left-hand menu titled 'Location-based displays' includes options like 'Whole genome', 'Chromosome summary', 'Region overview', 'Region in detail', 'Comparative Genomics', 'Genetic Variation', 'Markers', and 'Export location data'. Below the menu are options to 'Bookmark this page', 'Configure this page', and 'Add custom data to page'. The main content area shows 'Chromosome 6: 131,533,782-131,677,240' with a track for 'Assembly exceptions' and 'chromosome 6'. Below this are tabs for 'Region overview' and 'Region in detail'. The bottom section shows a genomic track with 'Cortigs' and 'Ensembl/Havana gene' tracks. The 'AKAP7' gene is highlighted in green. A 1.00 Mb scale bar is shown above the track, and a 'Forward strand' arrow is on the right.

Click and drag the mouse to recentre the display

Use the left-hand menus to navigate, export data and customise the page

# Ensembl biomart

<http://www.ensembl.org/biomart/martview>

The screenshot shows the Ensembl BioMart interface. At the top, the Ensembl logo is on the left, and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors are on the right. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results' on the left, and 'URL', 'XML', 'Perl', and 'Help' on the right. The main content area is divided into two panels. The left panel shows the 'Dataset' as 'Homo sapiens genes (GRCh37.p6)' and 'Filters' as 'Gene type : protein\_coding'. The right panel displays a list of columns to be included in the output, with the instruction 'Please select columns to be included in the output and hit 'Results' when ready'. The columns are grouped into categories: Features (selected), Structures, Transcript Event, Homologs, Variation, and Sequences. Below these are expandable sections for GENE, EXTERNAL, EXPRESSION, and PROTEIN DOMAINS.

**Dataset**  
Homo sapiens genes (GRCh37.p6)

**Filters**  
Gene type : protein\_coding

**Attributes**  
Ensembl Gene ID  
Ensembl Transcript ID

**Dataset**  
[None Selected]

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

- Features**
- Structures**
- Transcript Event**
- Homologs**
- Variation**
- Sequences**

GENE:  
 EXTERNAL:  
 EXPRESSION:  
 PROTEIN DOMAINS:

!

Ensembl changes version every three months

# Biomart central

## IDENTIFIER SEARCH

Examples: KRAS, ENSG00000146648

## TOOLS

- Gene retrieval
- Variant retrieval
- Sequence retrieval
- ID converter

- Cancer genes
- Ensembl
- Ensembl Bacteria
- Ensembl Fungi
- Ensembl Metazoa
- Ensembl Plants
- Ensembl Protists
- Mouse Genome Informatics
- VEGA

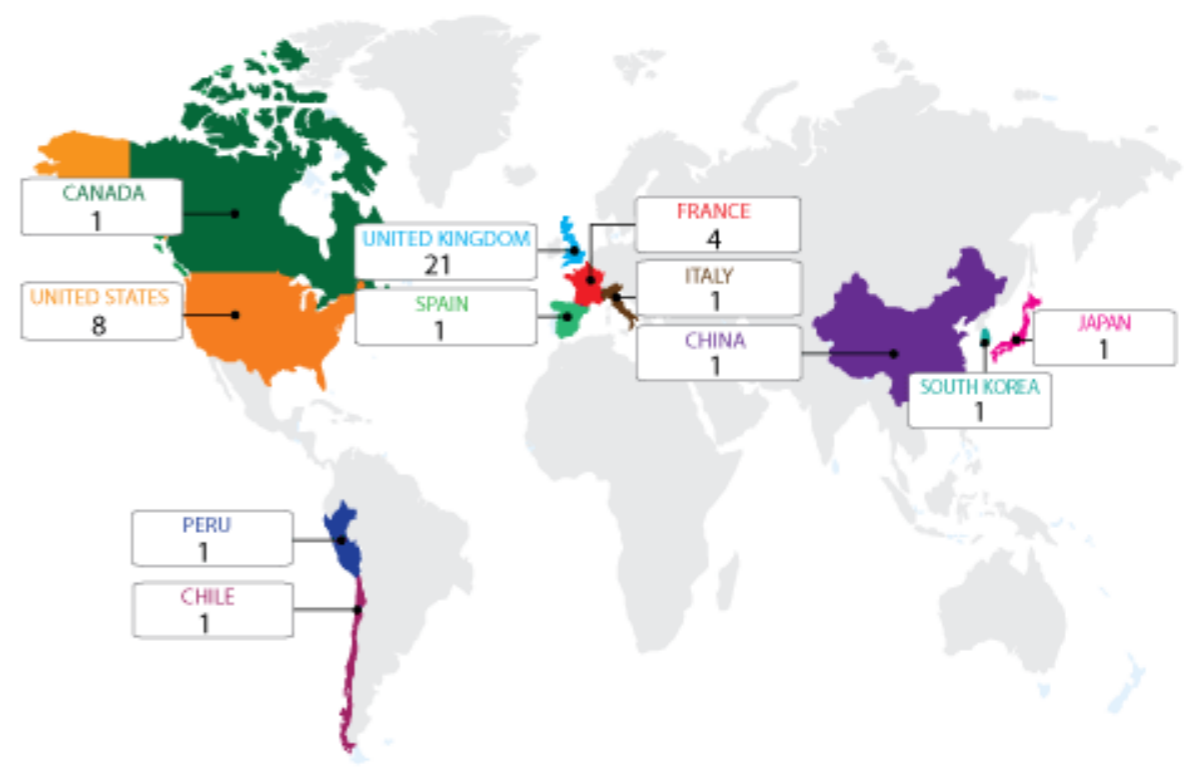
## DATABASE SEARCH

- Search by type
- Search by organism
- Search by database name (A-Z)

- ▶ Genome
- ▶ Gene annotation
- ▶ Protein sequence and structure
- ▶ Interaction and pathways
- ▶ Gene expression
- ▶ Cancer
- ▶ Model organism databases
- ▶ Other

## BIO MART CENTRAL PORTAL

Databases: 41



Click on the map to view the list of databases

## BioMart Central Portal

[Home](#) > [Converter](#) > ID converter

### ID CONVERTER

Dataset:

#### CONVERT

Entries with following IDs:


  

[upload file](#)

→

#### To

[Go »](#)

Powered by  biomart