

# Introduction à Ensembl/Biomart

Stéphanie Le Gras

Jean Muller

# Objectifs

- Révision sur les banques/bases de données biologiques
- Connaitre l’existence et l’utilité des principaux “Genome browser”
- Comprendre comment fonctionne le “Genome browser : Ensembl”
- S’initier à
  - la navigation dans Ensembl
  - l’utilisation des outils d’Ensembl
  - l’utilisation de Biomart

# Plan

- Introduction
  - Les banques/bases de données biologiques
  - Les “genome browsers”
- Le projet Ensembl
- Comprendre Ensembl
- Navigation dans le “genome browser” Ensembl
- Les outils intégrés à Ensembl
- Utilisation de Biomart

# Les banques/Bases de données biologiques

# De l'artisanat au haut débit...

- 1951 première séquence protéique
- 1967 construction d'arbres phylogénétiques
- 1970 algorithme de Needleman & Wunsch
- 1977 séquençage de l'ADN (Méthode Sanger)
  - premier package bioinformatique (Staden)
- 1978 bases de données Pir, EMBL, Genbank
- 1981 algorithme d'alignement local (Smith & Waterman)
- 1990 programme Blast
- 1991 étiquettes d'ADNc « EST »
- 1995 séquençage du génome complet d'une bactérie
- 1996 séquençage complet du génome de la levure
- 2001 première version du génome humain

=> Début de l'ère post-génomique



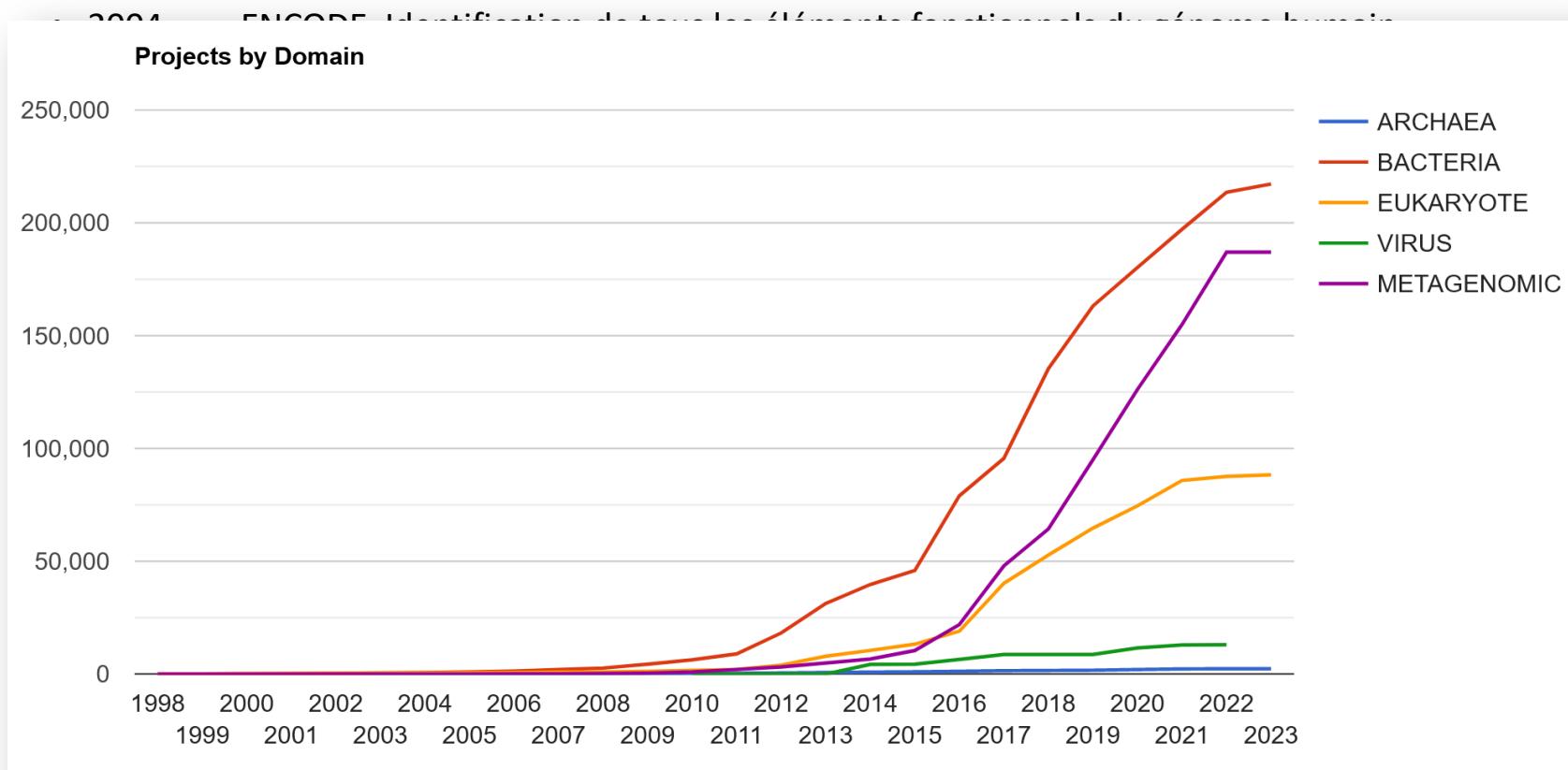
# L'ère post-génomique

- 2002 Séquence préliminaire du génome de la souris (Waterston et al., 2002)
  - 2004 ENCODE, Identification de tous les éléments fonctionnels du génome humain
  - 2005 Roche 454: Séquenceur auto. haut-débit de 2ème génération par pyroséquençage : GS20
  - 2007 Illumina/Solexa NGS de 2ème génération par synthèse microfluidique : GAIx  
Applied Biosystems NGS de 2ème génération par ligation : système SOLiD
  - 2008 Helicos Séquenceur auto. de 2ème génération par synthèse sans pré-amplification
  - 2012 ENCODE Encyclopédie des éléments fonctionnels du génome humain
  - 2014 Génome à 1000\$ 2 annonces Illumina et Life Technologies
  - 2016->40 000 génomes complets publiés (3 domaines du vivant)  
4989 archées, 409995 bactéries, 47196 eukaryotes et 18327 virus  
([www.genomesonline.org](http://www.genomesonline.org), 01/2023)
- Exomes et génomes humains séquencés complètement (patients + pop. Générale)



# L'ère post-génomique

- 2002 Séquence préliminaire du génome de la souris (Waterston et al., 2002)



# Centres de bioinformatique

- EBI (European Bioinformatics Institute)



<http://www.ebi.ac.uk/>

- NCBI (National Center for Biotechnology Information)



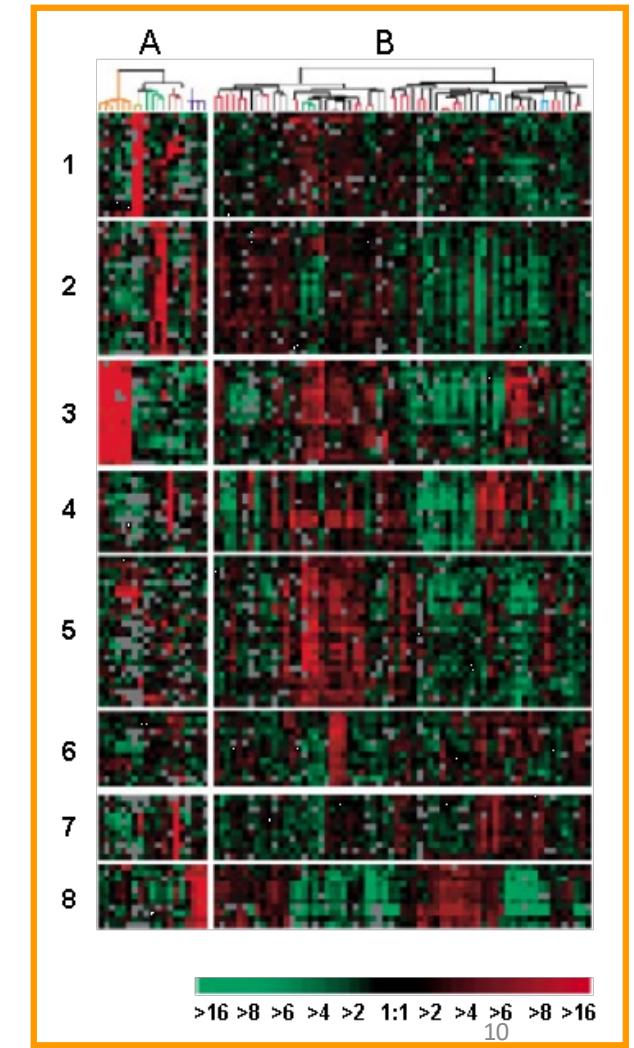
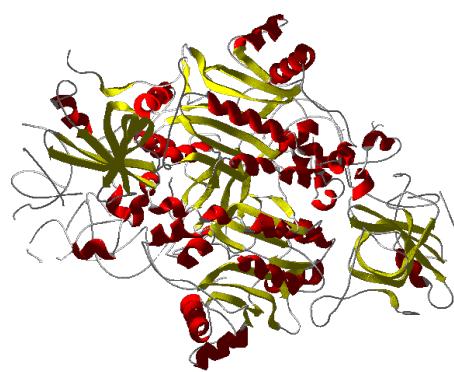
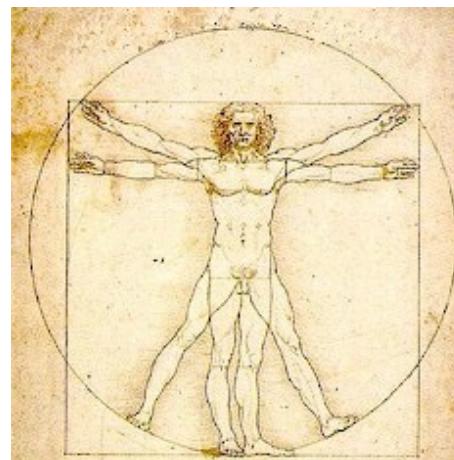
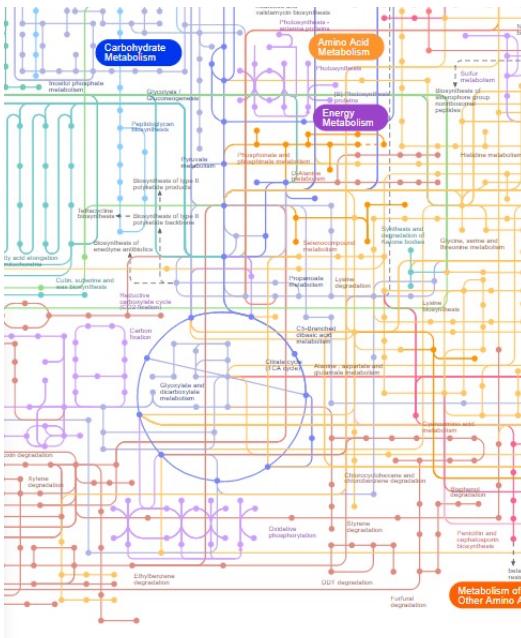
<http://www.ncbi.nlm.nih.gov/>

# Banques de données en biologie moléculaire

- Rôles des banques
  - Stockage
  - Diffusion (ftp, web...)
  - Organisation et standardisation des données
  - Connectivité avec autres banques
  - Actualisation

# Multiplicité des banques

**MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFYTPKARREVEGPQVGCALELAGGPGAA**



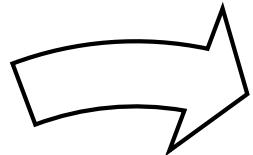
# Banques de séquences nucléiques généralistes

- Des banques incontournables :
  - dépôt obligatoire dans une des 3 banques avant publication
  - unique moyen d'accès aux séquences
- Alimentation :
  - soumission directe par la communauté scientifique  
(associée ou non à une publication)
  - dépôts de brevets
- Conséquences
  - banques exhaustives
  - banques extrêmement redondantes
  - contiennent des erreurs

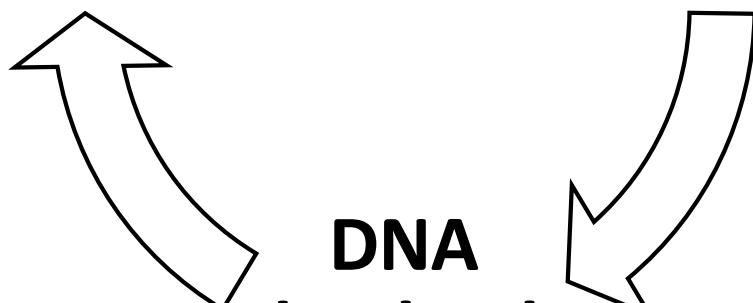
# Banques de séquences nucléiques généralistes



**GenBank**

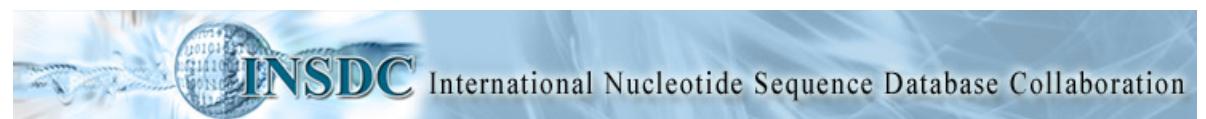


**EMBL**

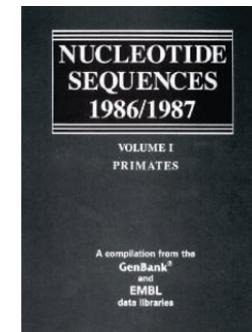


**DNA  
databank  
of Japan**

- 3 banques
- Échanges quotidiens des séquences collectées
- Effort d'unification=> format
  - 1986: accord entre GenBank/EMBL
  - 1987: accord entre GenBank/EMBL/DDBJ

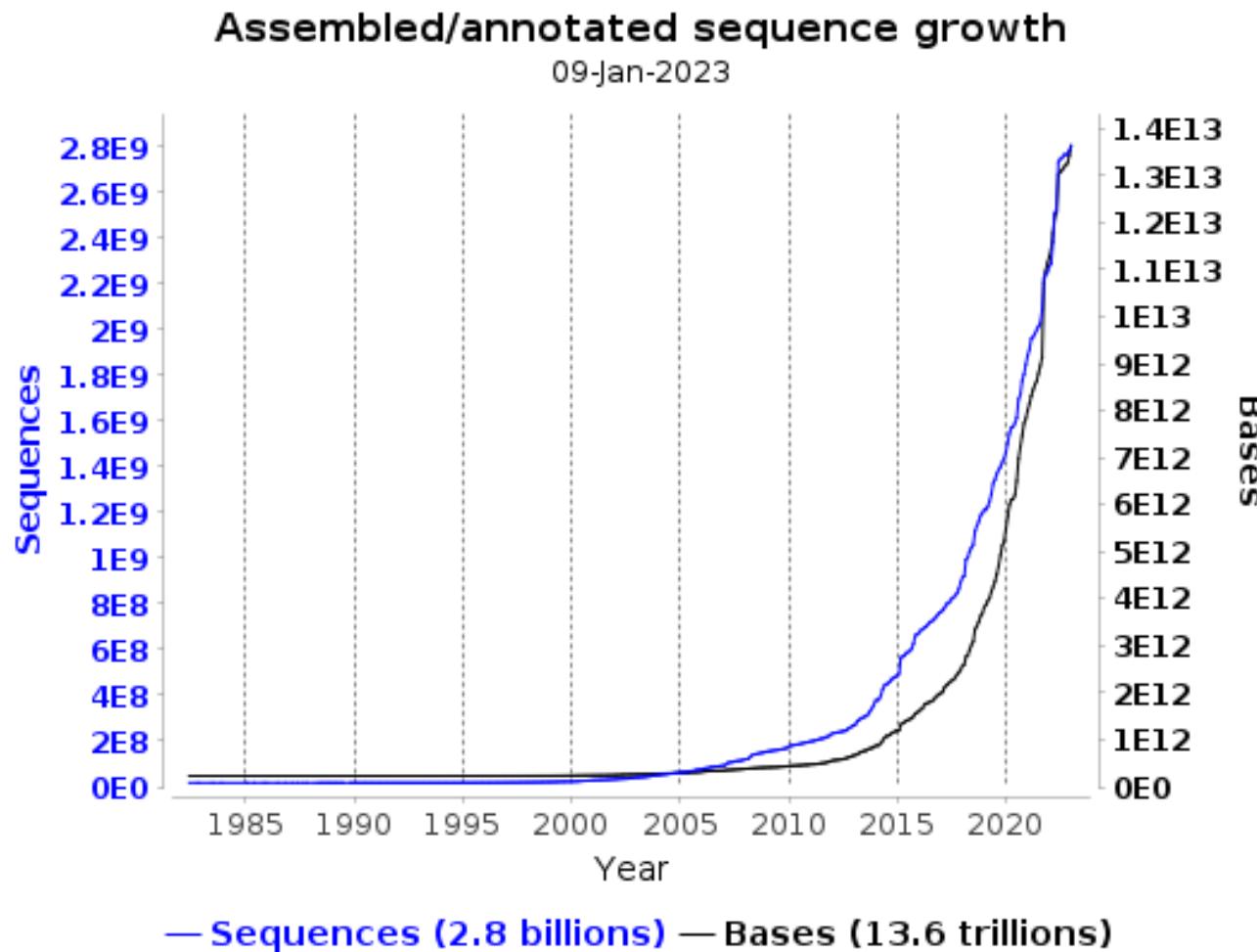


wikipedia

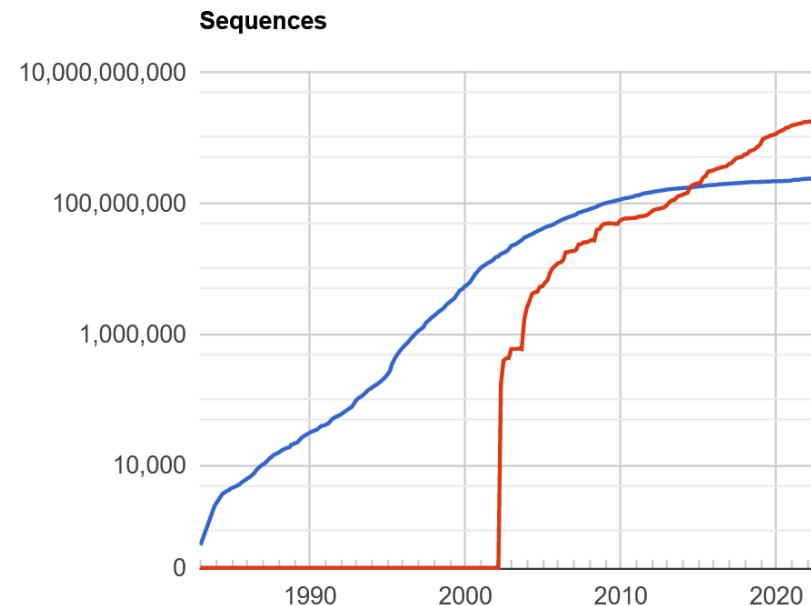
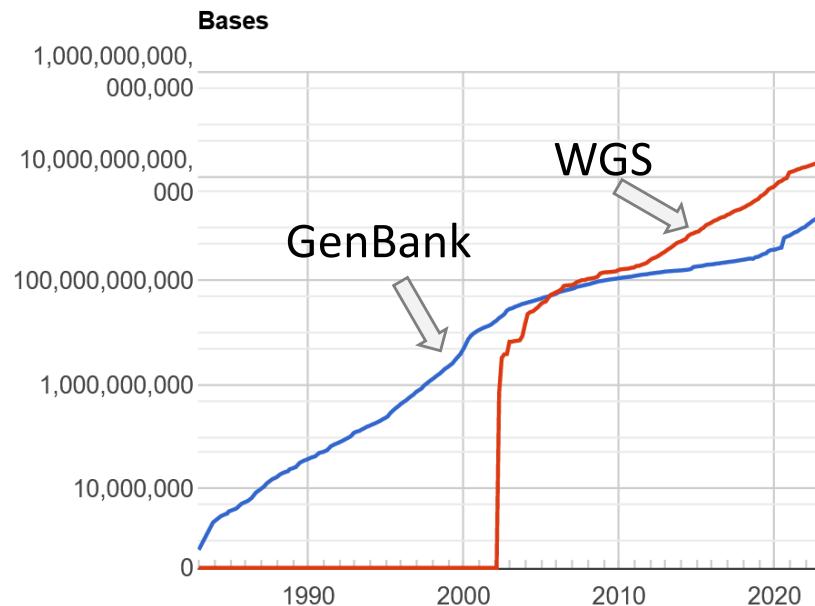


A compilation from the  
GenBank®  
and  
EMBL  
data libraries

# Evolution de la banque EMBL



# Evolution de la banque GenBank



**01/2023:** 1635 milliards de nucléotides, 241 millions d'entrées  
Doublement tous les 18 mois

# Banques de séquences protéiques généralistes



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

|                       |                        |                               |                               |
|-----------------------|------------------------|-------------------------------|-------------------------------|
| 09/2016<br>70 427 238 | 03/2018<br>106,245,682 | 01/2019<br><b>130,366,644</b> | 02/2020<br><b>167,278,920</b> |
|-----------------------|------------------------|-------------------------------|-------------------------------|

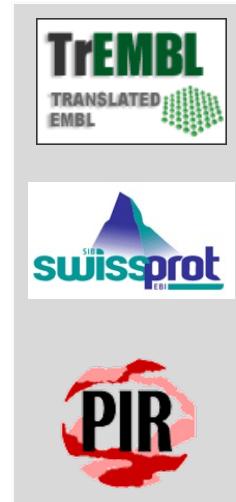
Transcrits: **29,869,155**

Organismes: **99,842**



<http://www.uniprot.org/>

|                       |                        |                               |
|-----------------------|------------------------|-------------------------------|
| 10/2016<br>68,493,254 | 02/2018<br>109,414,541 | 02/2020<br><b>179,812,129</b> |
|-----------------------|------------------------|-------------------------------|



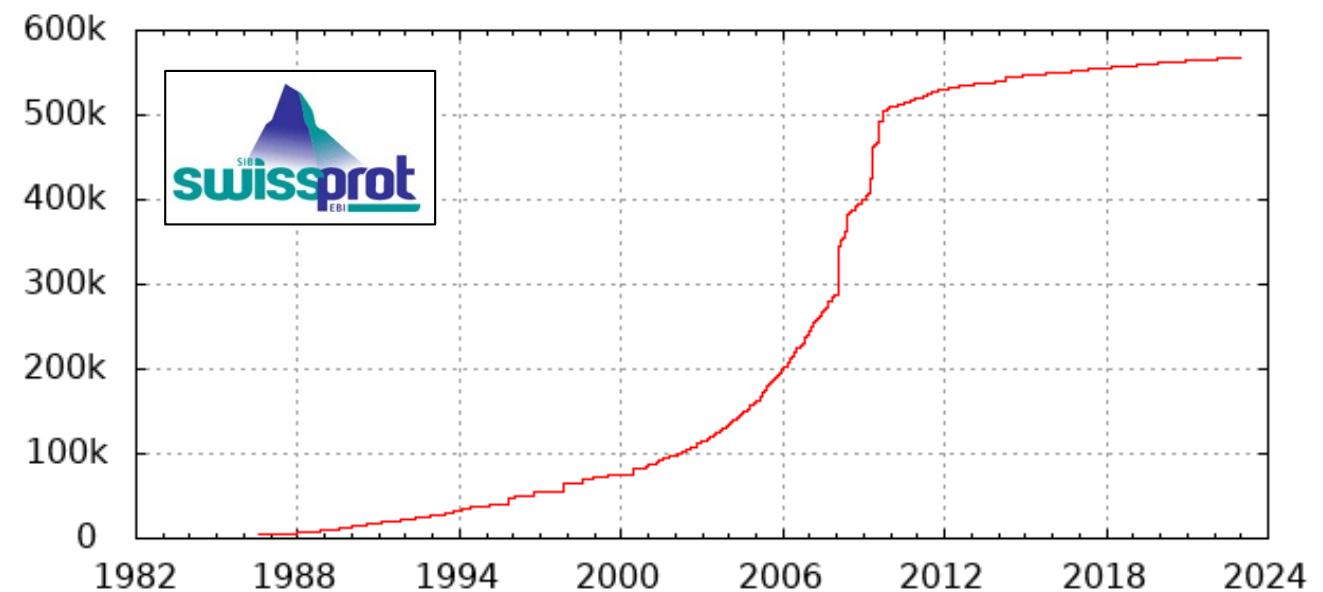
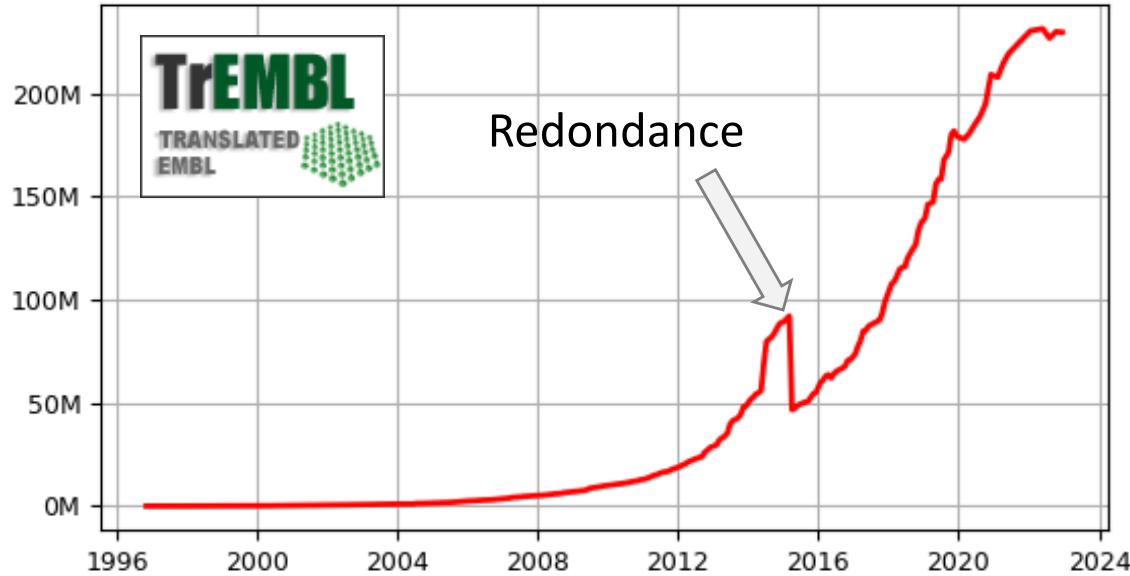
TrEMBL:  
**179,250,561** entrées

Swiss-Prot:  
**561,568** entrées

- 2 banques majeures
- Qualité variable/stabilisée
- Exhaustivité / Annotation

| Annotation                   | UniProt        |              | TrEMBL            |              |
|------------------------------|----------------|--------------|-------------------|--------------|
| Evidence at protein level    | 90,921         | 16,5%        | 118,013           | 0,2%         |
| Evidence at transcript level | 57,673         | 10,5%        | 971,005           | 1,8%         |
| Inferred from homology       | <b>387,632</b> | <b>70,5%</b> | <b>11,091,443</b> | <b>21,1%</b> |
| Predicted                    | <b>11,465</b>  | <b>2,1%</b>  | <b>40,603,140</b> | <b>76,9%</b> |
| Uncertain                    | 1,955          | 0,4%         | 0                 | 0%           |

## Evolution des bases de données protéiques



UniProt BLAST Align Peptide search ID mapping SPARQL Release 2022\_05 | Statistics 📦 🗑️ 📧 Help

## Find your protein

UniProtKB ▾ Examples: Insulin, APP, Human, P05067, organism\_id:9606 Advanced | List Search

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)

**Proteins**  
UniProt Knowledgebase

Reviewed (Swiss-Prot) 568,744  
Unreviewed (TrEMBL) 229,580,745

**Species Proteomes**

Protein sets for species with sequenced genomes from across the tree of life

**Protein Clusters**  
UniRef

Clusters of protein sequences at 100%, 90% & 50% identity

**Sequence Archive**  
UniParc

Non-redundant archive of publicly available protein sequences seen across different databases

Feedback Help

# Une entrée Swiss-Prot

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

Function Q8TAM1 · BBS10\_HUMAN

|                      |                       |                                  |                                |                           |
|----------------------|-----------------------|----------------------------------|--------------------------------|---------------------------|
| Names & Taxonomy     | Protein <sup>i</sup>  | Bardet-Biedl syndrome 10 protein | Amino acids                    | 723                       |
| Subcellular Location | Gene <sup>i</sup>     | BBS10                            | Protein existence <sup>i</sup> | Evidence at protein level |
| Disease & Variants   | Status <sup>i</sup>   | UniProtKB reviewed (Swiss-Prot)  | Annotation score <sup>i</sup>  | 5/5                       |
| PTM/Processing       | Organism <sup>i</sup> | Homo sapiens (Human)             |                                |                           |

Expression      Entry Feature viewer Publications External links History

Interaction      BLAST Download Add Add a publication Entry feedback

Structure

Family & Domains

Sequence

Similar Proteins

Feedback

Help

**Function<sup>i</sup>**

Probable molecular chaperone that assists the folding of proteins upon ATP hydrolysis (PubMed:[20080638](#)).  
Plays a role in the assembly of BBSome, a complex involved in ciliogenesis regulating transports vesicles to the cilia (PubMed:[20080638](#)).  
Involved in adipogenic differentiation (PubMed:[19190184](#)).  2 Publications

# Une entrée Swiss-Prot

## Un enregistrement (entrée) :

- les informations liées à la séquence
- la séquence elle-même
- indicateur de fin d'enregistrement

## Les champs :

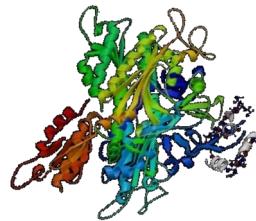
- regrouper les informations d'un même type
- faciliter l'accès à l'information

## Format général (flat file) :

- enregistrements organisés séquentiellement
- fichier texte (ASCII)
- fichiers disponibles en XML

ID BBS10\_HUMAN Reviewed; 723 AA.  
AC Q8TAM1; Q96CW2; Q9H5D2;  
DT 16-MAY-2006, integrated into UniProtKB/Swiss-Prot.  
DT 16-MAY-2006, sequence version 2.  
DT 14-OCT-2015, entry version 117.  
DE RecName: Full=Bardet-Biedl syndrome 10 protein;  
DR CCDS; CCDS9014.2; -.  
DR RefSeq; NP\_078961.3; NM\_024685.3.  
DR STRING; 9606.ENSP00000376946; -.  
DR Ensembl; ENST00000393262; ENSP00000376946; ENSG00000179941.  
DR GeneID; 79738; -.  
DR KEGG; hsa:79738; -.  
DR GO; GO:0005929; C:cilium; IEA:UniProtKB-SubCell.  
DR GO; GO:0005524; F:ATP binding; IEA:UniProtKB-KW.  
DR GO; GO:0001103; F:RNA polymerase II repressing transcription factor binding; IPI:MG1.  
DR GO; GO:0051131; P:chaperone-mediated protein complex assembly; IMP:MG1.  
DR GO; GO:0035058; P:nonmotile primary cilium assembly; IMP:BHF-UCL.  
DR GO; GO:0045494; P:photoreceptor cell maintenance; IMP:BHF-UCL.  
DR InterPro; IPR002423; Cpn60/TCP-1.  
DR InterPro; IPR027413; GROEL-like\_equatorial.  
DR Pfam; PF00118; Cpn60\_TCP1; 2. major  
PE 1: Evidence at protein level;  
KW ATP-binding; Bardet-Biedl syndrome; Cell projection; Chaperone;  
KW Ciliopathy; Complete proteome; Disease mutation; Mental retardation;  
KW Nucleotide-binding; Obesity; Polymorphism; Reference proteome;  
KW Sensory transduction; Vision.  
FT VARIANT 715 715 H -> R. {ECO:0000269|PubMed:21344540}.  
FT /FTId=VAR\_066261.  
SQ SEQUENCE 723 AA; 80838 MW; 558143FFA5F191DD CRC64;  
MLSSMAAAGS VKAAALQVAEV LEAIVSCCVG PEGRQVLCTK PTGEVLLSRN GGRLEALHL  
EHPIARMIVD CVSSHKKTG DGAKTIIIFL CHLLRGLHAI TDREKDPLMC ENIQTHGRHW  
KNCSRWFIS QALLTFQTQI LDGIMDQYLS RHFLSIFSSA KERTLCRSSL ELLLEAYFCG  
RVGRNNHKFI SQLMCDYFFF CMTCKGIGV FELVDDHFVE LNVGVTGLPV SDSRIIAGLV  
LQKDFSVYRP ADGDMRMVIV TETIQPLFST SGSEFILNSE AQFQTSQFWI MEKTKAIMKH  
LHSQNVKLLI SSVKQPDLVs YYAGVNNGISV VECLSSEEVS LIRRIIGLSP FVPPQAFSQC  
EIPNTALVKF CKPLILRSKR YVHLGLISTC AFIPHISIVLC GPVHGLIEQH EDALHGALKM  
LRQLFKDLDL NYMTQTNDQN GTSSLFIYKN SGESYQAPDP GNGSIQRPYQ DTVAENKDAL  
EKTQTYLKVN SNLVIPDVEL ETYIPYSTPT LTPTDTFQTV ETLTCLSLER NRLTDYYEPL  
LKNNSTAYST RGNRIEISYE NLQVTNITRK GSMLPVSKL PNMGTSQSYL SSSMPAGCVL  
PVGGNFEILL HYYLLNYAKK CHQSEETMVS MIANALLGI PKVLYKSKTG KYSFPHTYIR  
AVHALQTNQP LVSSQTGLES VMGKYQLLTS VLQCLTKILT IDMVITVKRH PQKVHNQDSE  
DEL  
//

# Les banques de structures



- La Protein Data Bank (PDB)

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB Contact us

**RCSB PDB PROTEIN DATA BANK** 200,069 Structures from the PDB 1,000,357 Computed Structure Models (CSM) ▾ 3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Help Advanced Search | Browse Annotations

PDB-101 wwPDB EMDDataResource Nucleic Acid Database Foundation

**NEW! Computed Structure Models (CSM)** Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

COVID-19 CORONAVIRUS Resources

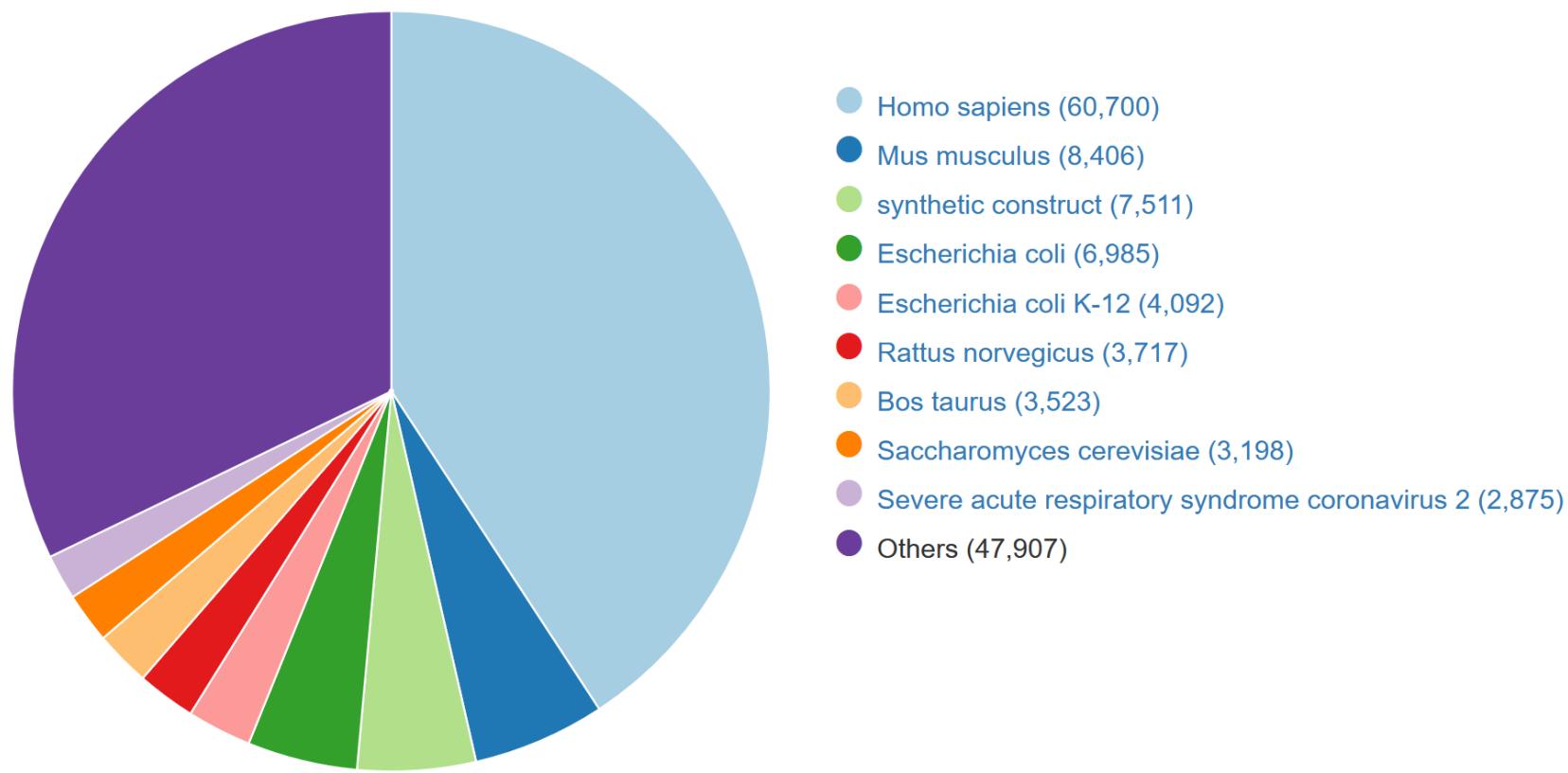
January Molecule of the Month

200000

<http://www.rcsb.org/pdb/>

# Les banques de structures

Distribution en fonction de l'organisme d'origine



# Hétérogénéité de la qualité en fonction de leur origine

La séquence des protéines est prédite!



La qualité des séquences de protéines dépend de la source et est donc très hétérogène

cDNA clonés et séquencés individuellement => protéine  
(complets, séquençage multiple, vérification)



HTC (High-Throughput cDNA) => protéine  
(full-length mais séquence brute, indels, multiple codons initiateur)



Structure 3D => protéine  
(attention au substitutions ponctuelles/délétions)



Séquence génomique procaryote => protéine prédite  
(prédiction réalisée par outils bioinformatiques, erreurs de codon initiateur de traduction fréquents, indels en Nter)



Séquence génomique eucaryote => protéine prédite  
(prédiction réalisée par outils bioinformatiques, erreurs de prédictions de sites d'épissage fréquents, frameshifts, indels)



# Hétérogénéité de la qualité en fonction de leur origine

## 1) Annotations manuelles



Réalisées par des experts, les entrées sont traitées une par une (UniProt/SwissProt)

## 2) Annotations automatiques



Réalisées par des outils bioinformatiques de prédiction de domaines, de fonctions...

« **by similarity** », « **homologous to** », « **related to** », « **-like** », « **putative** », « **potential** »

Sont produites en haut-débit (ex: annotation de génomes)

Elles sont légions dans les banques ... et en attente d'une validation

## 3) Absence d'annotations



« **hypothetical protein** »

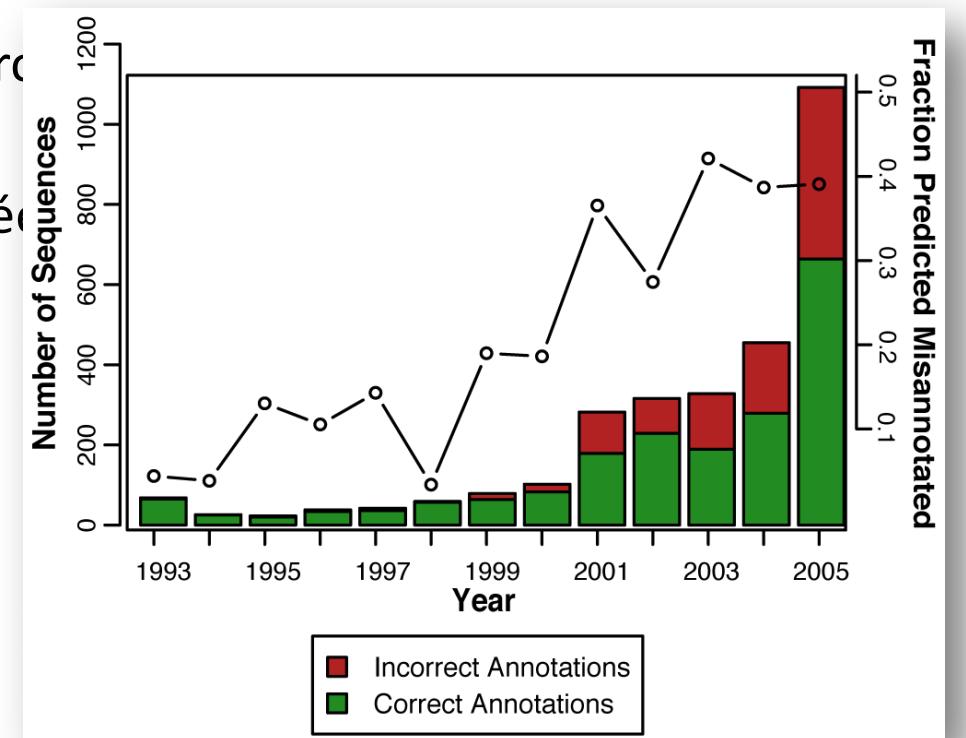
# Exemple de l'importance de l'annotation

**Exemple 1:** DUF domain = Domain of Unknown Function

**Exemple 2:** FAM20C = Family with sequence similarity 20, member C

**Exemple 3:** Analyse de 37 familles de protéines

L'augmentation de la **quantité** de données ne signifie pas une augmentation de la **qualité** de ces données.



# **Evolution des bases de données protéiques**

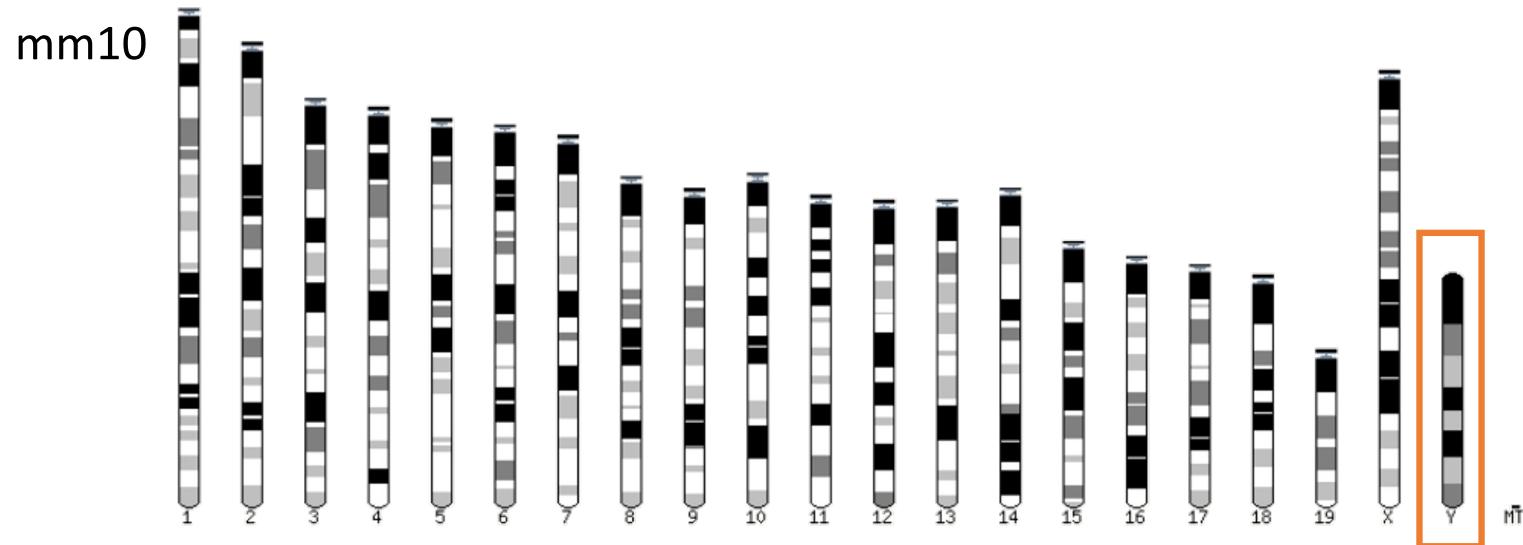
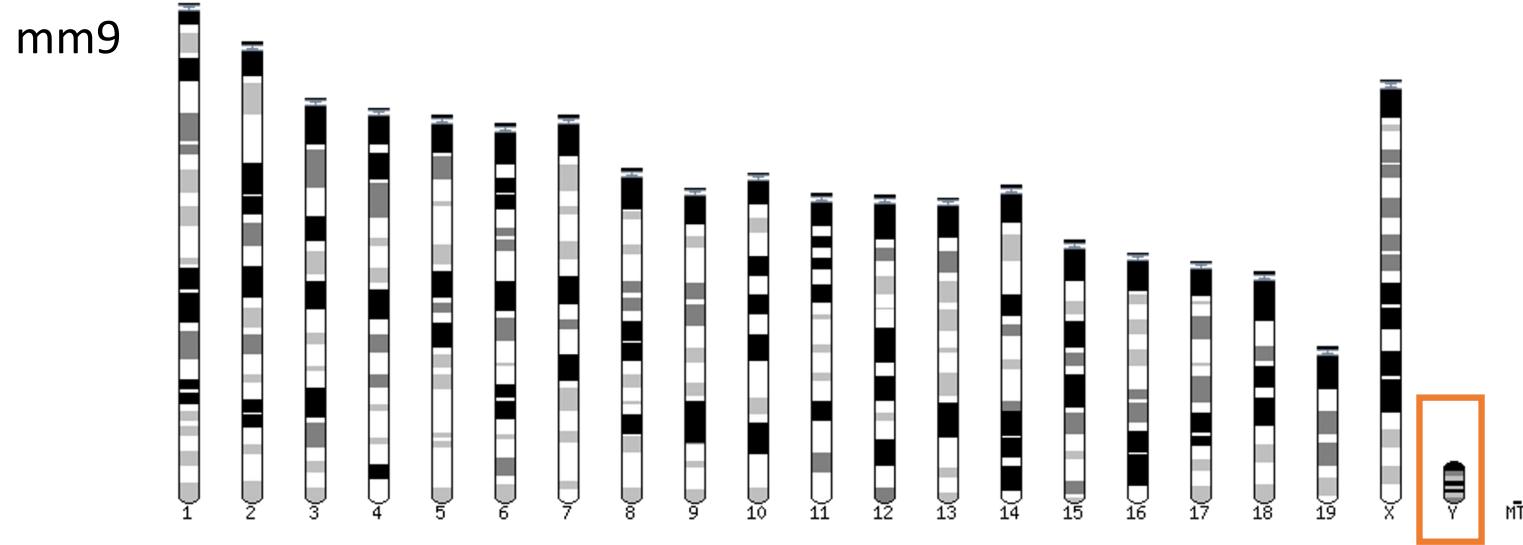
Bases de données majeures  
collecte des données individuelles et collectives

Attention à la qualité de ces données  
bases avec les Raw data vs Annotation

Ces données seront agrégées sur le génome humain

# Genome browsers

# Genome builds



# Human Genome Builds

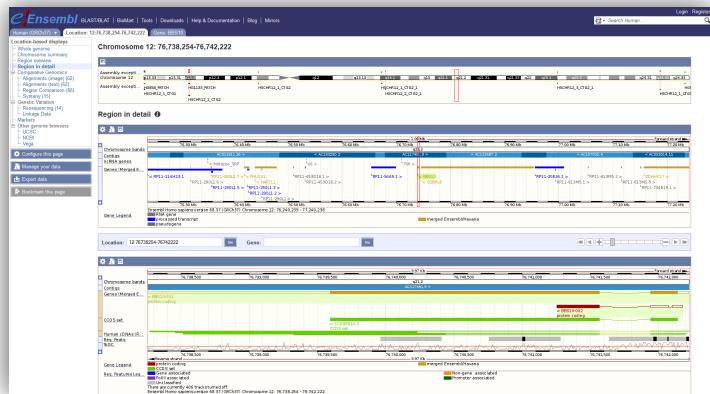
| SPECIES        | UCSC VERSION | RELEASE DATE | RELEASE NAME                       | STATUS               |
|----------------|--------------|--------------|------------------------------------|----------------------|
| <b>MAMMALS</b> |              |              |                                    |                      |
| Human          | hg38         | Dec. 2013    | Genome Reference Consortium GRCh38 | Available            |
|                | hg19         | Feb. 2009    | Genome Reference Consortium GRCh37 | Available            |
|                | hg18         | Mar. 2006    | NCBI Build 36.1                    | Available            |
|                | hg17         | May 2004     | NCBI Build 35                      | Available            |
|                | hg16         | Jul. 2003    | NCBI Build 34                      | Available            |
|                | hg15         | Apr. 2003    | NCBI Build 33                      | Archived             |
|                | hg13         | Nov. 2002    | NCBI Build 31                      | Archived             |
|                | hg12         | Jun. 2002    | NCBI Build 30                      | Archived             |
|                | hg11         | Apr. 2002    | NCBI Build 29                      | Archived (data only) |
|                | hg10         | Dec. 2001    | NCBI Build 28                      | Archived (data only) |
|                | hg8          | Aug. 2001    | UCSC-assembled                     | Archived (data only) |
|                | hg7          | Apr. 2001    | UCSC-assembled                     | Archived (data only) |
|                | hg6          | Dec. 2000    | UCSC-assembled                     | Archived (data only) |
|                | hg5          | Oct. 2000    | UCSC-assembled                     | Archived (data only) |
|                | hg4          | Sep. 2000    | UCSC-assembled                     | Archived (data only) |
|                | hg3          | Jul. 2000    | UCSC-assembled                     | Archived (data only) |
|                | hg2          | Jun. 2000    | UCSC-assembled                     | Archived (data only) |
|                | hg1          | May 2000     | UCSC-assembled                     | Archived (data only) |

# Genome Browsers – L'outil de référence

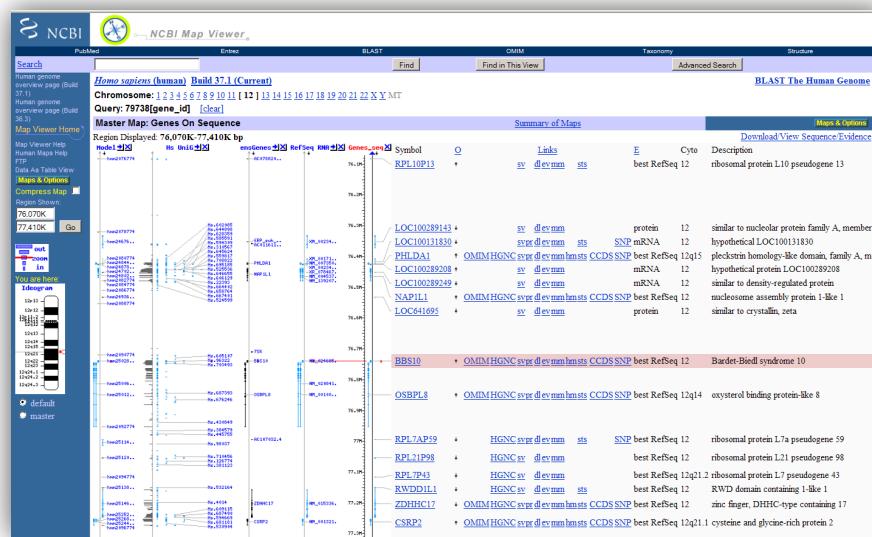
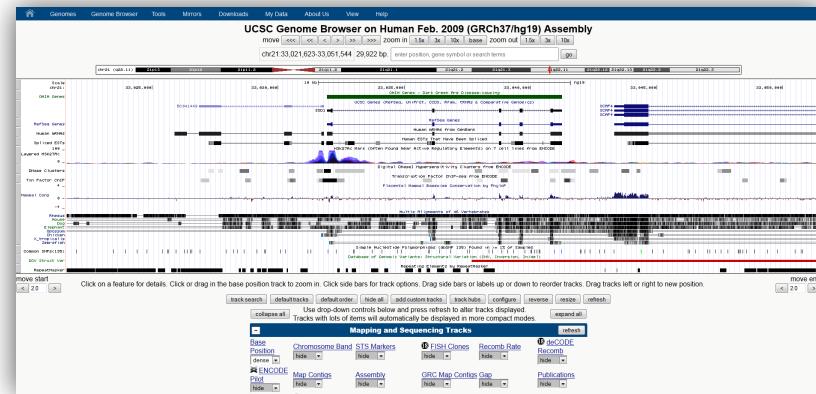
- Elément de référence absolue le **génome**
- Agrégateur et générateur d'informations/annotations
  - Prédictions de gènes
  - Protéines
  - Données d'expression
  - Variations
- Synthèse rapide et visuelle de données primordiales

# Il y a Genome Browsers...

EBI - Ensembl

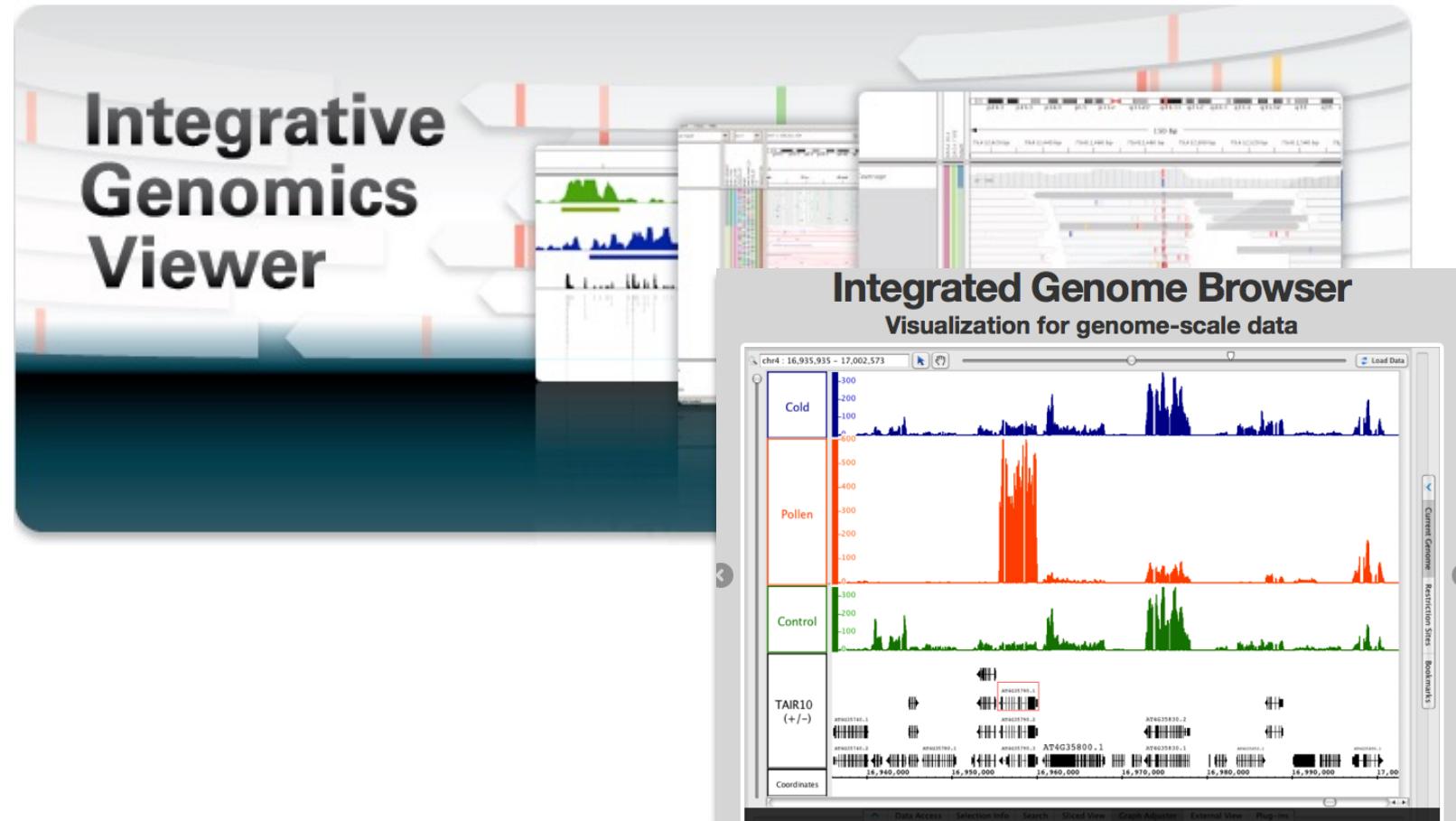


UCSC – Genome Browser



NCBI – Map Viewer

# Et Genome browsers



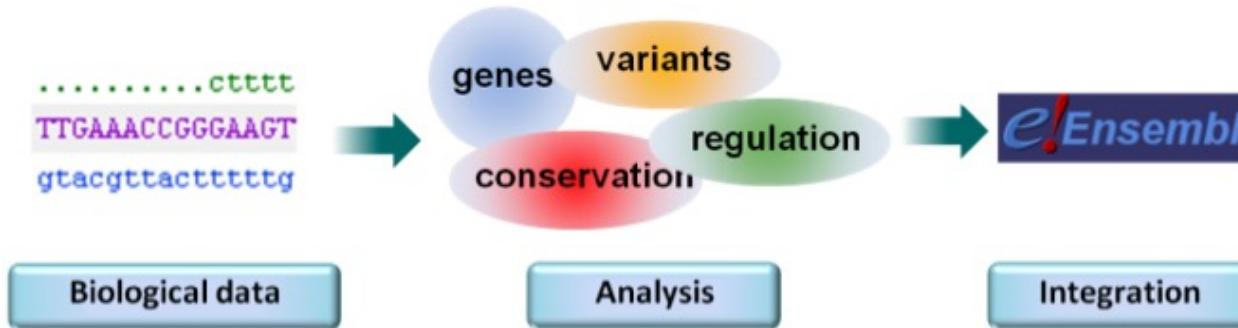
# Ensembl

# Le projet Ensembl

- Initié en 1999 (avant la première version du génome humain)
- Projet en collaboration entre l'European Bioinformatics Institute (EBI) et le Wellcome Trust Sanger Institute (WTSI)
- Objectif :
  - Annoter automatiquement les génomes
  - Ajouter des données biologiques aux annotations
  - Rendre publique les annotations sur le web
- Ensembl ne produit pas ses propres données d'assemblage de génome!

# Le projet Ensembl

- Données disponibles :
  - Génomes
  - Données de génomique comparative
  - Variations
  - Elément régulateur des gènes
  - Annotations externes



- Lancement du site web en juillet 2000 (au début il n'y avait que le génome humain)

# Les génomes d'Ensembl

- Espèces de vertébrés dans <http://ensembl.org>
- EnsemblGenomes (avril 2009) :  
<https://ensemblgenomes.org/>
  - Métazoaires : <http://metazoa.ensembl.org>
  - Bactéries : <http://bacteria.ensembl.org>
  - Plantes : <http://plants.ensembl.org>
  - Fungi : <http://fungi.ensembl.org>
  - Protistes : <http://protists.ensembl.org>

# L'interface web

**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search  All species for  e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Tools [All tools](#)

**BioMart >** Export custom datasets from Ensembl with this data-mining tool

**BLAST/BLAT >** Search our genomes for your DNA or protein sequence

**Variant Effect Predictor >** Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

**Ensembl Release 108 (Oct 2022)**

- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWiCH consortium for chicken
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail and collared flycatcher
- Retirement of postGAP tool

[More release news](#) on our blog

All genomes  Pig breeds Pig reference genome and 12 additional breeds [View full list of all species](#)

Favourite genomes  Human GRCh38.p13 [Still using GRCh37?](#) Mouse GRCm39 Zebrafish GRCz11

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.  
Note: species that already exist on this site will continue to be updated with the full range of annotations.

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

Other news from our blog

- 02 Dec 2022: Job: Senior Full Stack Developer
- 24 Nov 2022: The first invertebrate-themed Ensembl Rapid Release is out!
- 18 Nov 2022: Geek for a Week : Georgia Argiroou

Compare genes across species Find SNPs and other variants for my gene Gene expression in different tissues Retrieve gene sequence Find a Data Display Use my own data in Ensembl

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available. Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Ensembl release 108 - Oct 2022 © EMBL-EBI Permanent link - View in archive site

GLOBAL CORE BIODATA RESOURCE elixir Core Data Resource

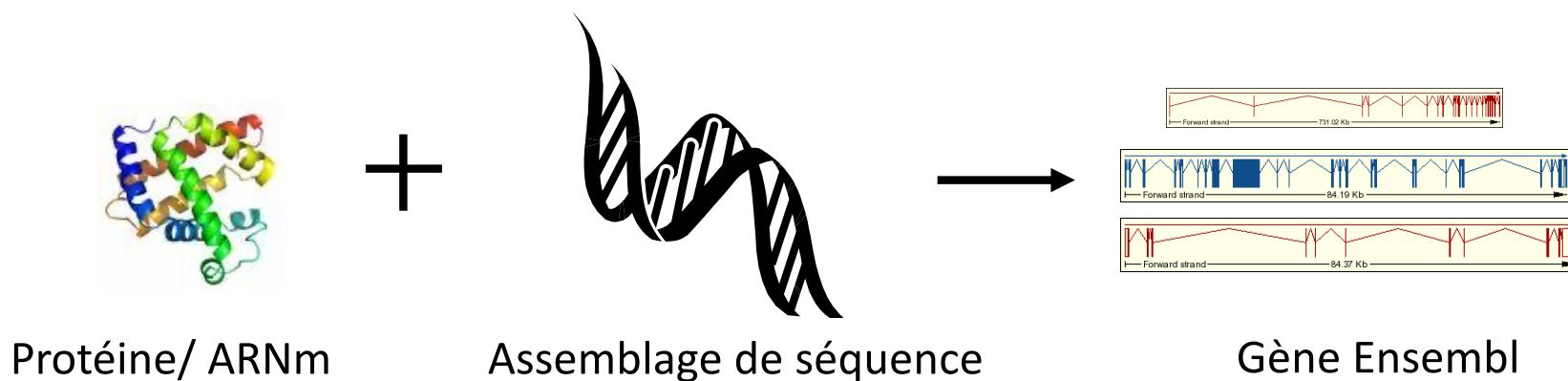
# Comprendre ENSEMBL

# Les annotations

- 3 à 6 mois
- Annotation par Ensembl
  - Annotation automatique (Ensembl Genebuild) :
    - Détermination des transcrits dans le génome entier
    - Basées sur des séquences d'ARNm et protéiques extraites des banques de données publiques
  - *Curation* manuelle : au cas par cas. Ex: l'humain, la souris, le rat, le zebrafish + autres vertébrés (produit par le groupe HAVANA du WTSI)
  - Fusion des annotations automatiques et manuelles (Gold)
- + Annotations importées depuis flyBase, WormBase, SGD

# Les annotations

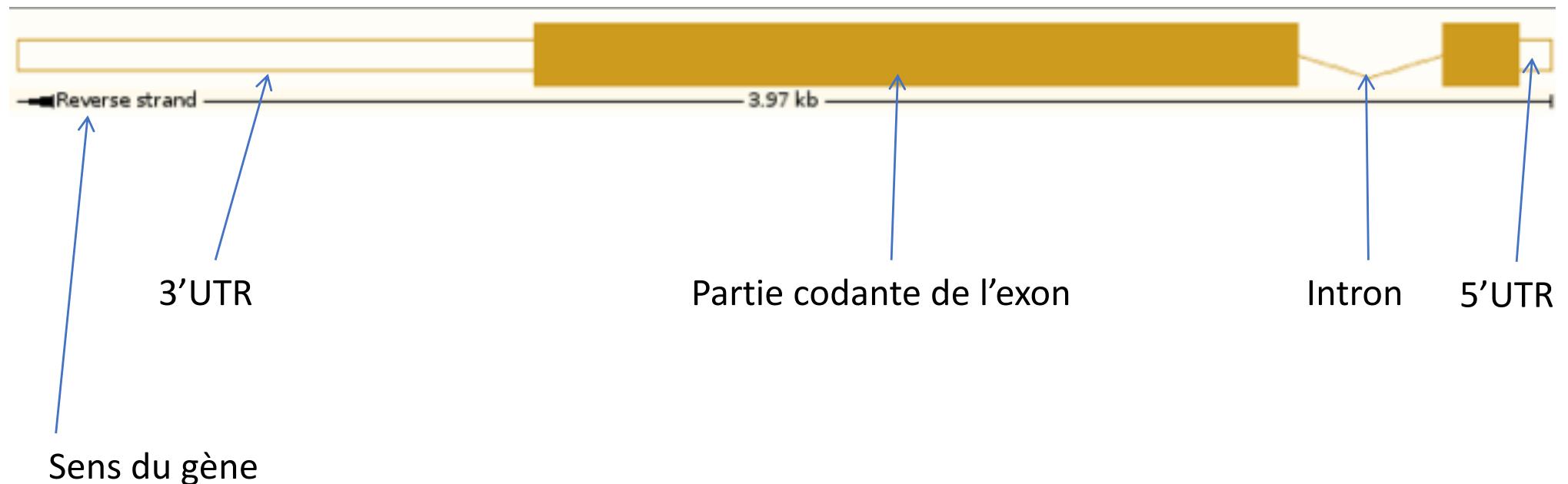
- Les transcrits d'Ensembl sont basés sur les bases de données suivantes :
  - Uniprot/Swiss-Prot (*curation manuelle*)
  - Uniprot/TrEMBL
  - NCBI refSeq (*curation manuelle*)



# Les annotations

- Les annotations des gènes peuvent varier entre les différents genome browsers (Ensembl, UCSC, NCBI)
- CCDS (Consensus CDS) est un jeu de données de gènes codants validés par tous les membres du consortium (EBI, HGNC, MGI, NCBI, WTSI)
  - <http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi>
  - Il faut que l'assemblage du génome soit suffisamment stable pour identifier les gènes dont les positions sont identiques entre les différentes sources (chez humain et souris)

# Transcrits Ensembl

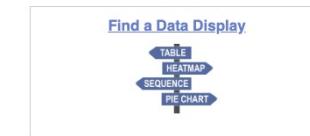
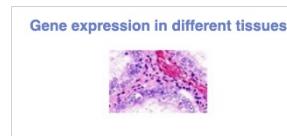
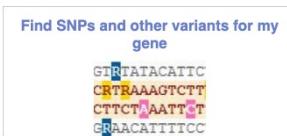
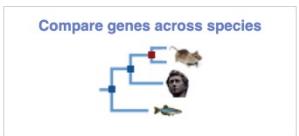


# Identifiants Ensembl

- ENS**G**### Ensembl **Gene** ID
- ENST### Ensembl **Transcript** ID
- ENSP### Ensembl **Peptide** ID
- ENSE### Ensembl **Exon** ID
- Ajout d'un suffix pour les autres espèces
  - MUS (*Mus musculus*) pour la souris: ENS**MUS**G###
  - DAR (*Danio rerio*) pour le zebrafish: ENS**DAR**G###
  - etc.

# Version (Release)

- ~ tous les 3-4 mois
- Lien vers la dernière version d'Ensembl est toujours : <http://www.ensembl.org>



Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at [EMBL-EBI](#) and our software and data are freely available. Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.



Ensembl release 108 - Oct 2022 © EMBL-EBI

[Permanent link - View in archive site](#)

- Lien vers une version particulière d'Ensembl : <http://Oct2022.archive.ensembl.org/index.html>

# Ensembl : Archives

Using this website | Annotation and prediction | Data access | API & software | About us

In this section | Archives: Table of assemblies

Search documentation | Go

## Ensembl Archives

### About Archive Ensembl

The main Ensembl site ([www.ensembl.org](http://www.ensembl.org)) and the mirror sites are updated with the latest data approximately every three months. We maintain the Ensembl Archive sites so that there are stable links to data from a particular release. As of December 2016 these will be available for [five years](#), together with the following longer term archives:

- Annotation on the human NCBI36 assembly is available at our [Ensembl 54 archive](#) site.
- Annotation on the mouse NCBIm37 assembly is available at our [Ensembl 67 archive](#) site.
- As from August 2014 we are supporting the human GRCm37 assembly at our dedicated [GRCh37 human](#) site. Unlike the other Ensembl archive sites, this will be updated to the latest web interface every Ensembl release and there may be occasional data updates to human.

Archived databases are also maintained for at least 10 years. Currently all databases are available from 2004. More information is available from our [MySQL database documentation](#). We also maintain data archives from 2004 available from our [FTP site](#).

For all enquiries, please [contact the Ensembl HelpDesk](#).

### Notes

- Ensembl aims to maintain stable identifiers for genes (ENSG), transcripts (ENST), proteins (ENSP) and exons (ENSE) as long as possible. Changes within the genome sequence assembly or an updated genome annotation may dramatically change a gene model. In these cases, the old set of stable IDs is retired and a new one assigned. Gene and transcript pages both have an ID History view which maps changes in the ID from the earliest version in Ensembl.
- Protein family identifiers (fam), Ensembl EST gene identifiers (ENSESTG) and Genscan identifiers (GENSCAN) are currently not stable.
- With the exception of the GRCh37 human site **BLAST**, **BLAT** and **other tools** are not available from the archive sites.
- Accounts are shared between the current site and almost all archives. The exceptions are the older human NCBI36 and the mouse GRCm37 sites where changes in architecture and code make sharing logins impractical.

### Linking to the Archive Ensembl sites

The Archive Ensembl sites have the format: <http://<three-letter-month><year>.archive.ensembl.org> for example <http://nov2008.archive.ensembl.org>

In the footer of each current Ensembl page, there is a link called 'Permanent link', which links to the corresponding page in the Ensembl Archive. A similar link on each archive page links back to the current site (i.e. [www.ensembl.org](http://www.ensembl.org)).

For example if you are looking at the Alternative Splicing view for human gene BRCA2 on the [main Ensembl site](#) in August 2015, when Ensembl 80 was the current version, the URL would be:  
[http://www.ensembl.org/Homo\\_sapiens/Gene/Splice?db=core:g=ENSG00000139618;r=13;31787617-31871809;t=ENST00000380152](http://www.ensembl.org/Homo_sapiens/Gene/Splice?db=core:g=ENSG00000139618;r=13;31787617-31871809;t=ENST00000380152)

and the equivalent archived page URL would be:  
[http://jul2015.archive.ensembl.org/Homo\\_sapiens/Gene/Splice?db=core:g=ENSG00000139618;r=13;31787617-31871809;t=ENST00000380152](http://jul2015.archive.ensembl.org/Homo_sapiens/Gene/Splice?db=core:g=ENSG00000139618;r=13;31787617-31871809;t=ENST00000380152)

Unfortunately, owing to the change in site organisation between releases it is not always possible to map pages one-to-one between the current Ensembl site and the older archives. If the link does not take you to the data you expected, trying using the search facility to locate the information.

Ensembl release 108 - Oct 2022 © EMBL-EBI

Permanent link

### List of currently available archives

- [Ensembl GRCm37](#): Full Feb 2014 archive with BLAST, VEP and BioMart
- [Ensembl 108: Oct 2022](#) - this site
- [Ensembl 107: Jul 2022](#)
- [Ensembl 106: Apr 2022](#)
- [Ensembl 105: Dec 2021](#)
- [Ensembl 104: May 2021](#)
- [Ensembl 103: Feb 2021](#)
- [Ensembl 102: Nov 2020](#)
- [Ensembl 101: Aug 2020](#)
- [Ensembl 100: Apr 2020](#)
- [Ensembl 99: Jan 2020](#)
- [Ensembl 98: Sep 2019](#)
- [Ensembl 97: Jul 2019](#)
- [Ensembl 96: Apr 2019](#)
- [Ensembl 95: Jan 2019](#)
- [Ensembl 94: Oct 2018](#)
- [Ensembl 93: Jul 2018](#)
- [Ensembl 92: Apr 2018](#)
- [Ensembl 91: Dec 2017](#)
- [Ensembl 80: May 2015](#)
- [Ensembl 77: Oct 2014](#)
- [Ensembl 75: Feb 2014](#)
- [Ensembl 54: May 2009](#)

Table of archives showing assemblies present in each one.

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

# Ensembl : Archives

**Archive! Ensembl** BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species... 

**Tools** **BioMart >**

[All tools](#)

Export custom datasets from Ensembl with this data-mining tool

**Search**

All species for  **Go**

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

**All genomes** **Favourite genomes** 

-- Select a species --

**Pig breeds**  
Pig reference genome and 12 additional breeds  
  
[View full list of all species](#)

**Human**  
GRCh38.p13  
  
[Still using GRCh37?](#)

**Mouse**  
GRCm39  


**Zebrafish**  
GRCz11  


**Ensembl Archive Release 104 (May 2021)**

- Update to the Ensembl Canonical transcript set.
- Human and mouse gene sets updated to GENCODE 38 and GENCODE M27, respectively.
- Retirement of gene names derived from BAC clones.

[More release news](#)  on our blog

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.  
Note: species that already exist on this site will continue to be updated with the full range of annotations.

**Go**

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomic data from individual sites.

Les anciennes version d'Ensembl sont conservées pendant 5 ans sauf si elles contiennent la dernière version de l'annotation d'un génome.

# Ensembl : Archives

- <http://www.ensembl.org/info/website/archives/assembly.html>

The screenshot shows a table titled "Table of Assemblies" on the Ensembl Archives website. The table lists species names in the first column and assembly versions for specific months and years in the subsequent columns. A legend at the top indicates three types of entries: "New species" (yellow), "Species present in archive" (light blue), and "Species not in this version of Ensembl" (white).

|                                | Oct 2022 v108            | Jul 2022 v107 | Apr 2022 v106 | Dec 2021 v105 | May 2021 v104 | Feb 2021 v103 | Nov 2020 v102 | Aug 2020 v101 | Apr 2020 v100 | Jan 2020 v99 | Sep 2019 v98 | Jul 2019 v97 | Apr 2019 v96 | Jan 2019 v95 | Oct 2018 v94 | Jul 2018 v93 | Apr 2018 v92 | Dec 2017 v91 | Aug 2017 v90 | May 2017 v89 | Mar 2017 v88 | Dec 2016 v87 | Oct 2016 v86 | Jul 2016 v85 | Mar 2016 v84 | Dec 2015 v83 | Sep 2015 v82 |
|--------------------------------|--------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Abingdon island giant tortoise | ASM359739v1              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| African ostrich                | ASM69896v1               |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Agassiz's desert tortoise      | ASM289641v1              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Algerian mouse                 | SPRET_EiU_v1             |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Alpaca                         | vicPac1                  |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Alpine marmot                  | marMar2.1                |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Amazon molly                   | Poecilia_formosa-5.1.2   |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| American beaver                | C.can_genome_v1.0        |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| American bison                 | Bison_UMD1.0             |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| American black bear            | ASM33442v1               |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| American mink                  | NNGG.v01                 |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Angola colobus                 | Cang.pa_1.0              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Arabian camel                  | CamDro2                  |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Arctic ground squirrel         | ASM342692v1              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Argentine black and white tegu | HLtpMer3                 |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
|                                | Oct 2022 v108            | Jul 2022 v107 | Apr 2022 v106 | Dec 2021 v105 | May 2021 v104 | Feb 2021 v103 | Nov 2020 v102 | Aug 2020 v101 | Apr 2020 v100 | Jan 2020 v99 | Sep 2019 v98 | Jul 2019 v97 | Apr 2019 v96 | Jan 2019 v95 | Oct 2018 v94 | Jul 2018 v93 | Apr 2018 v92 | Dec 2017 v91 | Aug 2017 v90 | May 2017 v89 | Mar 2017 v88 | Dec 2016 v87 | Oct 2016 v86 | Jul 2016 v85 | Mar 2016 v84 | Dec 2015 v83 | Sep 2015 v82 |
| Armadillo                      | Dasnov3.0                |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Asian bonytongue               | fSciFor1.1               |               |               |               |               |               |               |               |               |              |              |              |              |              | ASM162426v1  |              |              |              |              |              |              |              |              |              |              |              |              |
| Asiatic black bear             | ASM966005v1              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Atlantic cod                   | gadMor3.0                |               |               |               |               |               |               |               |               |              |              |              |              | gadMor1      |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Atlantic herring               | Ch_v2.0.2                |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Atlantic salmon                | Ssal_v3.1                |               |               |               |               |               |               |               |               |              |              |              |              | ICSASG_v2    |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Australian saltwater crocodile | CroPor_comp1             |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Ballan wrasse                  | BallGen_V1               |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Barramundi perch               | ASB_HGAPassembly_v1      |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Beluga whale                   | ASM228892v3              |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Bengalese finch                | LonStrDom1               |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Bicolor damselfish             | Stegastes_partitus-1.0.2 |               |               |               |               |               |               |               |               |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |

# Aide et documentations

- Vidéo Youtube (workshop...)
- FAQ
- Exercices
- Cours en ligne
- Publications :
  - Flicek, P. et al. **Ensembl 2013**. Nucleic Acids Res. Advanced Access (Database Issue).  
<http://www.ncbi.nlm.nih.gov/pubmed/23203987>
  - Xosé M. Fernández-Suárez and Michael K. Schuster. **Using the Ensembl Genome Server to Browse Genomic Sequence Data.** UNIT 1.15 in Current Protocols in Bioinformatics, Jun 2010
  - Giulietta M Spudich and Xosé M Fernández Suárez. **Touring Ensembl: A practical guide to genome browsing.** BMC Genomics 2010, 11:295 (11 May 2010)

# Naviguer dans ensembl

# www.ensembl.org

**e|Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register  

**Tools**

**BioMart >** Export custom datasets from Ensembl with this data-mining tool

**BLAST/BLAT >** Search our genomes for your DNA or protein sequence

**Variant Effect Predictor >** Analyse your own variants and predict the functional consequences of known and unknown variants

**Search**

All species for  Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

**All genomes**

-- Select a species --

**Pig breeds** Pig reference genome and 12 additional breeds

[View full list of all species](#)

**Favourite genomes**

**Human** GRCh38.p13  
Still using GRCh37?

**Mouse** GRCm39

**Zebrafish** GRCz11

**Ensembl Release 108 (Oct 2022)**

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWiCH consortium for chicken
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail and collared flycatcher
- Retirement of postGAP tool

[More release news](#) on our blog

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

**Other news from our blog**

- 02 Dec 2022: [Job: Senior Full Stack Developer](#)
- 24 Nov 2022: [The first invertebrate-themed Ensembl Rapid Release is out!](#)
- 18 Nov 2022: [Geek for a Week - Georgie Argirou](#)

**Compare genes across species**

**Find SNPs and other variants for my gene**

**Gene expression in different tissues**

**Retrieve gene sequence**

```
GCCTGACTTCCTGGGTGCG  
GTTTATACATTC  
CTTAAGCTT  
CTCTAATT  
GAACTTCC
```

**Find a Data Display**

TABLE  
HEATMAP  
SEQUENCE  
PIE CHART

**Use my own data in Ensembl**

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available. Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

GLOBAL CORE BIODATA RESOURCE

elixir Core Data Resource

Ensembl release 108 - Oct 2022 © EMBL-EBI

Permanent link - [View in archive site](#)

# Ensembl Genomes

## Bactéries

**EnsemblBacteria**

Search for a gene | Search for a genome

Archive sites

The following archive sites are available to access previous versions of data:

- Release 49, December 2020 [eg49-bacteria.ensembl.org](#)
- Release 45, September 2019 [eg45-bacteria.ensembl.org](#)
- Release 40, July 2018 [eg40-bacteria.ensembl.org](#)
- Release 37, October 2017 [eg37-bacteria.ensembl.org](#)

**Ensembl Bacteria**

Ensembl Bacteria is a browser for bacterial and archaeal genomes. These are taken from the databases of the International Nucleotide Sequence Database Collaboration, the European Nucleotide Archive at the EBI, GenBank at the NCBI, and the DNA Database of Japan.

**Data access**

Data can be visualised through the Ensembl genome browser and accessed programmatically via our Perl and REST APIs. Data is also accessible through public databases such as NCBI's BioProject, BioSample and BioAssay dumps in FASTA, EMBL, GTF, GFF3, JSON and RDF formats. A selection of over 100 key bacterial genomes have been included in the pan-taxonomic compara, and genes from all genomes have been classified into families using HAMAP and PANTHER more details.

**What's New in Release 52**

Release 52 of Ensembl Bacteria has no major updates from the previous release. As for release 49, we only represent non-redundant bacterial genomes as defined by criteria set out by UniProt. See more details about this update in our [blog post](#).

**Did you know...**

**e!REST**  
To access Ensembl data programmatically using any programming language, try our REST service or. For detailed examples, including examples from a wide range of languages, visit [http://rest.ensembl.org](#).

**EMBL-EBI**  

## Fungi

**EnsemblFungi**

Search: All species | Go

**What's New in Release 52**

- EnsemblFungi has 1506 genomes in total
  - 477 new genomes imported from ENA (<https://www.ebi.ac.uk/ena/browser/home>)
  - 15 genomes imported from VFPPathDB
- Updated data
  - Updated fungal gene trees
  - Updated protein features for all species using InterProScan with version 86 of InterPro
  - Updated BioMarts for all gene and variation data
  - Updated pan-taxonomic gene trees and homologies

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

**Archive sites**

## Plantes

**EnsemblPlants**

Search: All species | Go

**Wheat assemblies**

Ensembl Plants hosts the [latest wheat assembly](#) from the IWGSC (RefSeq v1.0), including:

- The IWGSC RefSeq v1.1 gene annotation, with links to [sheath-expression.com](#) and [Krauthein](#)
- Alignments from the 10+ genome project
- Alignment of 98,270 high confidence genes from the IWGSC v1 annotation
- Axion 35K, 200K SNP arrays from [CerealsDB](#), including QTL links in selected cases and Linkage Disequilibrium display. See QTL example here
- EMS-induced mutations from sequenced TILLING populations of Cadenza (coding regions) and Kronos (coding regions and promoters)
- Inter-Homologous Variants (IHVs) between the A, B and D genome alignments
- Chromosome specific KASP markers were added from the Nottingham BBSRC Wheat Research Centre
- Whole genome alignments to rice, brachypodium and barley
- Assembly-to-assembly mapping and gene ID mapping to the previous ZGA v1 assembly, archived at [eg37-plants.ensembl.org](#)
- Polyloid view enabled, allowing users to view alignments among multiple wheat components simultaneously
- Durum wheat 35K, 80K, 200K and TaIW280K variants
- Chromosome and centromere data can be viewed here

**Archive sites**

Archive of release 49 of EnsemblPlants: [eg49-plants.ensembl.org](#) (Dec 2020)

Archive of release 45 of EnsemblPlants: [eg45-plants.ensembl.org](#) (Sep 2019)

Navigation dans Ensembl

## Protistes

**EnsemblProtists**

Search: All species | Go

**What's New in Release 52**

- Genomes
  - No updated genomes from last release
- Updated data
  - Updated protein features for all species using InterProScan with version 86 of InterPro
  - Updated BioMarts for all gene and variation data
  - Updated pan-taxonomic gene trees and homologies

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

[Go](#)

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

**Archive sites**

The following archive sites are available to access previous versions of data:

- Release 49, December 2020 [eg49-protists.ensembl.org](#)
- Release 45, September 2019 [eg45-protists.ensembl.org](#)
- Release 40, July 2018 [eg40-protists.ensembl.org](#)

## Métazoaires

**EnsemblMetazoa**

Search: All species | Go

**What's New in Release 52**

- Updated data
  - Updated species
    - Cimex lectularius (Hem)
  - Updated protein features for all species using InterProScan with version 86 of InterPro
  - Updated BioMarts for all gene and variation data
  - Updated pan-taxonomic gene trees and homologies

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

[Go](#)

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

**Archive sites**

Archive of release 49 of EnsemblMetazoa: [eg49-metazoa.ensembl.org](#) (Dec 2020)

Archive of release 45 of EnsemblMetazoa: [eg45-metazoa.ensembl.org](#) (Sep 2019)

Archive of release 40 of EnsemblMetazoa: [eg40-metazoa.ensembl.org](#)

# Le site web Ensembl: page d'accueil

Outils



Recherche

All genomes

– Select a species –

Pig breeds

Pig reference genome and 12 additional breeds

View full list of species

BioMart >  
Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >  
Search our genomes for your DNA or protein sequence

Variant Effect Predictor >  
Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

## Ensembl Release 108 (Oct 2022)

- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWiCH consortium for chick, zebrafish, flycatcher
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail
- Retirement of postGAP tool

More release news on our blog

News

## Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

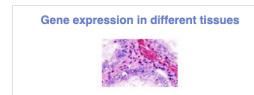
Rapid Release news on our blog

## Other news from our blog

- 02 Dec 2022: Job: Senior Full Stack Developer
- 24 Nov 2022: The first invertebrate-themed Ensembl Rapid Release is out!
- 18 Nov 2022: Geek for a Week : Georgia Argirou

Liste déroulante  
Accès aux génomes

Ensembl release 108 - Oct 2022 © EMBL-EBI



Accès aux archives d'Ensembl



# Le site web Ensembl: les génomes

**Recherche**

**Lien vers des exemples**

**Informations, statistiques**

**Gene annotation**

**Variation**

**Example gene**

**Example transcript**

**Example variant**

**Example phenotype**

**Example structural variant**

Ensembl release 108 - Oct 2022 © EMBL-EBI

[Permanent link](#) - [View in archive site](#)

## About Us

- [About us](#)
- [Contact us](#)
- [Citing Ensembl](#)
- [Privacy policy](#)

## Get help

- [Using this website](#)
- [Adding custom tracks](#)
- [Downloading data](#)
- [Variant tutorial](#)

## Our sister sites

- [Ensembl Bacteria](#)
- [Ensembl Fungi](#)
- [Ensembl Plants](#)
- [Ensembl Protists](#)

## Follow us

- [Blog](#)
- [Twitter](#)
- [Facebook](#)

# Le site web Ensembl: statistiques des génomes

**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

**Human assembly and gene annotation**

**Assembly**

This site provides a data set based on the December 2013 *Homo sapiens* high coverage assembly GRCh38 from the [Genome Reference Consortium](#). This assembly is used by UCSC to create their hg38 database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb.
- chromosome length total 3.1 Gb (excluding haplotypes).

It also includes 261 alt loci scaffolds, mainly in the LRC/KIR complex on chromosome 19 (35 alternate sequence representations) and the [MHC region on chromosome 6](#) (7 alternate sequence representations).

[Watch a video on YouTube](#) about patches and haplotypes in the Human genome.

**Patches**

As the GRC maintains a patching process for the assembly, patches are being introduced. Currently, assembly patches are of two types:

- Novel patch: adds new sequence at a locus that will remain as a haplotype in the next major assembly release by GRC
- Fix patch: adds sequence that corrects the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC

**Informations générales sur l'assemblage**

The Ensembl human genome annotations have been updated using Ensembl's automatic annotation pipeline. The updated annotation incorporates new protein and cDNA sequences which have become publicly available since the last GRCh38 genebuild (December 2013).

In the current release, we continue to display a joint gene set based on the merge between the automatic annotation from Ensembl and the manually curated annotation from Havana. See the statistics table, right, for the corresponding GENCODE version number. The Consensus Coding Sequence (CCDS) identifiers have also been mapped to the annotations. More information about the [CCDS project](#).

Updated manual annotation from Havana is merged into the Ensembl annotation every release. Transcripts from the two annotation sources are merged if they share the same internal exon-intron boundaries (i.e. have identical splicing pattern) with slight differences in the terminal exons allowed. Importantly, all Havana transcripts are included in the final Ensembl/Havana merged (GENCODE) gene set.

- [Detailed information on genebuild](#) (PDF)

**Neanderthal genome**

A preliminary assembly of the Neanderthal (*Homo sapiens neanderthalensis*) genome is available via the [Neanderthal Genome Browser](#), an Ensembl-powered project based at the Max Planck Institute.

**More information**

General information about this species can be found in [Wikipedia](#).

**Statistics**

**Summary**

|                                |   |
|--------------------------------|---|
| Assembly                       | GRCh38.p13 (Genome Reference Consortium Human Build 38). INSDC Assembly <a href="#">GCA_000001405.28</a> . Dec 2013 |
| Base Pairs                     | 3,096,649,726   |
| Golden Path Length             | 3,096,649,726   |
| Assembly provider              | <a href="#">Genome Reference Consortium</a>   |
| Annotation provider            | Ensembl   |
| Annotation method              | Full genebuild  |
| Genebuild started              | Jan 2014  |
| Genebuild released             | Jul 2014  |
| Genebuild last updated/patched | Jul 2022  |
| Database version               | 108.38  |
| Gencode version                | GENCODE 42  |

**Statistiques**

**Gene counts (Primary assembly)**

|                        |                               |
|------------------------|-------------------------------|
| Coding genes           | 19,813 (excl 651 readthrough) |
| Non coding genes       | 25,972                        |
| Small non coding genes | 4,864                         |
| Long non coding genes  | 18,887                        |
| Misc non coding genes  | 2,221                         |
| Pseudogenes            | 15,241                        |
| Gene transcripts       | 252,477                       |

**Gene counts (Alternative sequence)**

|                        |                             |
|------------------------|-----------------------------|
| Coding genes           | 3,028 (excl 26 readthrough) |
| Non coding genes       | 1,682                       |
| Small non coding genes | 297                         |
| Long non coding genes  | 1,198                       |
| Misc non coding genes  | 187                         |
| Pseudogenes            | 1,796                       |
| Gene transcripts       | 21,630                      |

**Other**

|                          |             |
|--------------------------|-------------|
| Genscan gene predictions | 51,756      |
| Short Variants           | 715,081,156 |
| Structural variants      | 7,097,115   |

# Le site web Ensembl: caryotype

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Login/Register

Search all species...

**Whole genome**

+ Add features

Add/remove tracks | Custom tracks | Share | Export image | Reset configuration

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

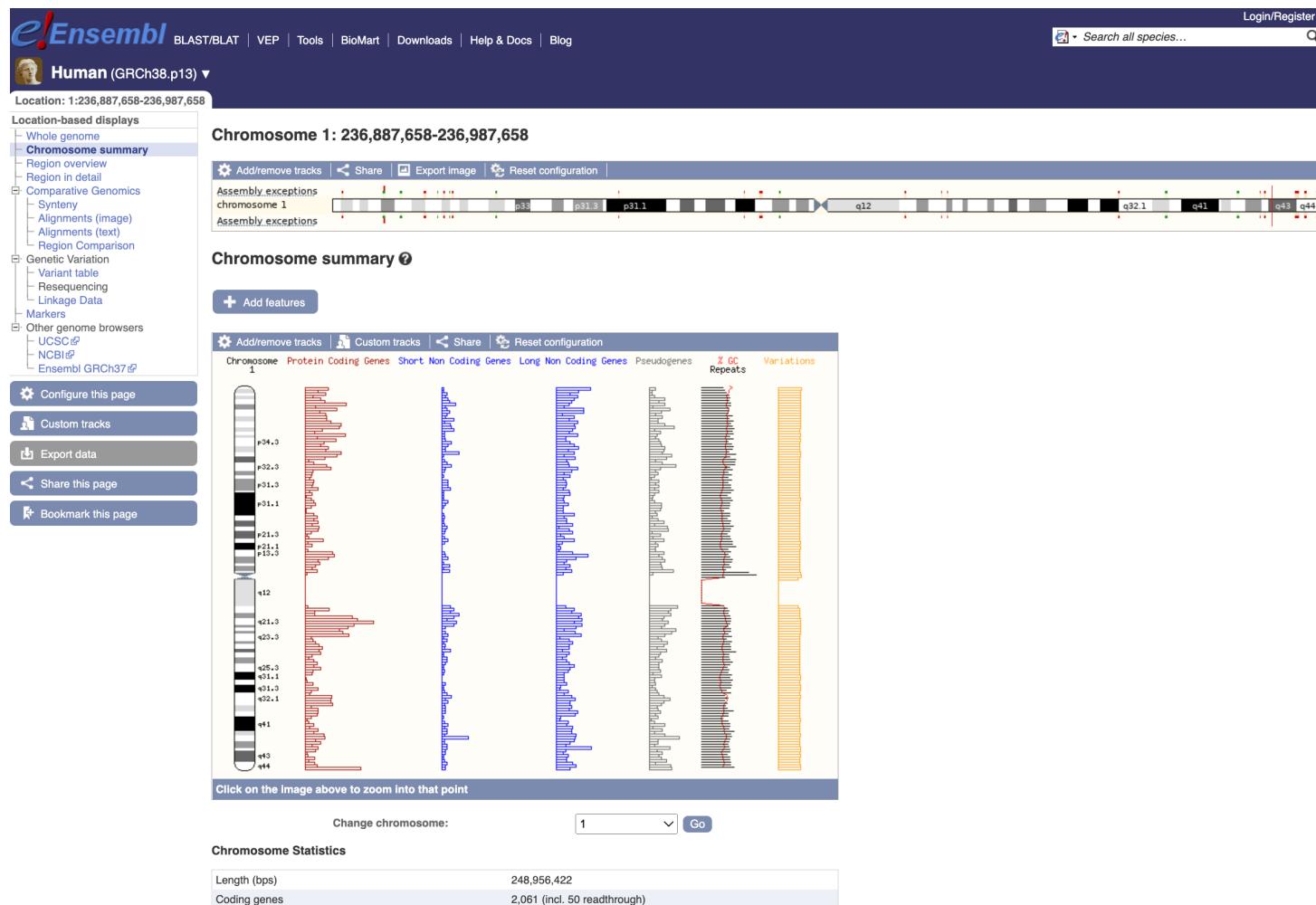
**Summary**

|                                |   |
|--------------------------------|---|
| Assembly                       | GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly <a href="#">GCA_000001405.28</a> , Dec 2013 |
| Base Pairs                     | 3,096,649,726   |
| Golden Path Length             | 3,096,649,726   |
| Assembly provider              | Genome Reference Consortium   |
| Annotation provider            | Ensembl   |
| Annotation method              | Full genebuild  |
| Genebuild started              | Jan 2014  |
| Genebuild released             | Jul 2014  |
| Genebuild last updated/patched | Jul 2022  |
| Database version               | 108.38  |
| Gencode version                | GENCODE 42  |

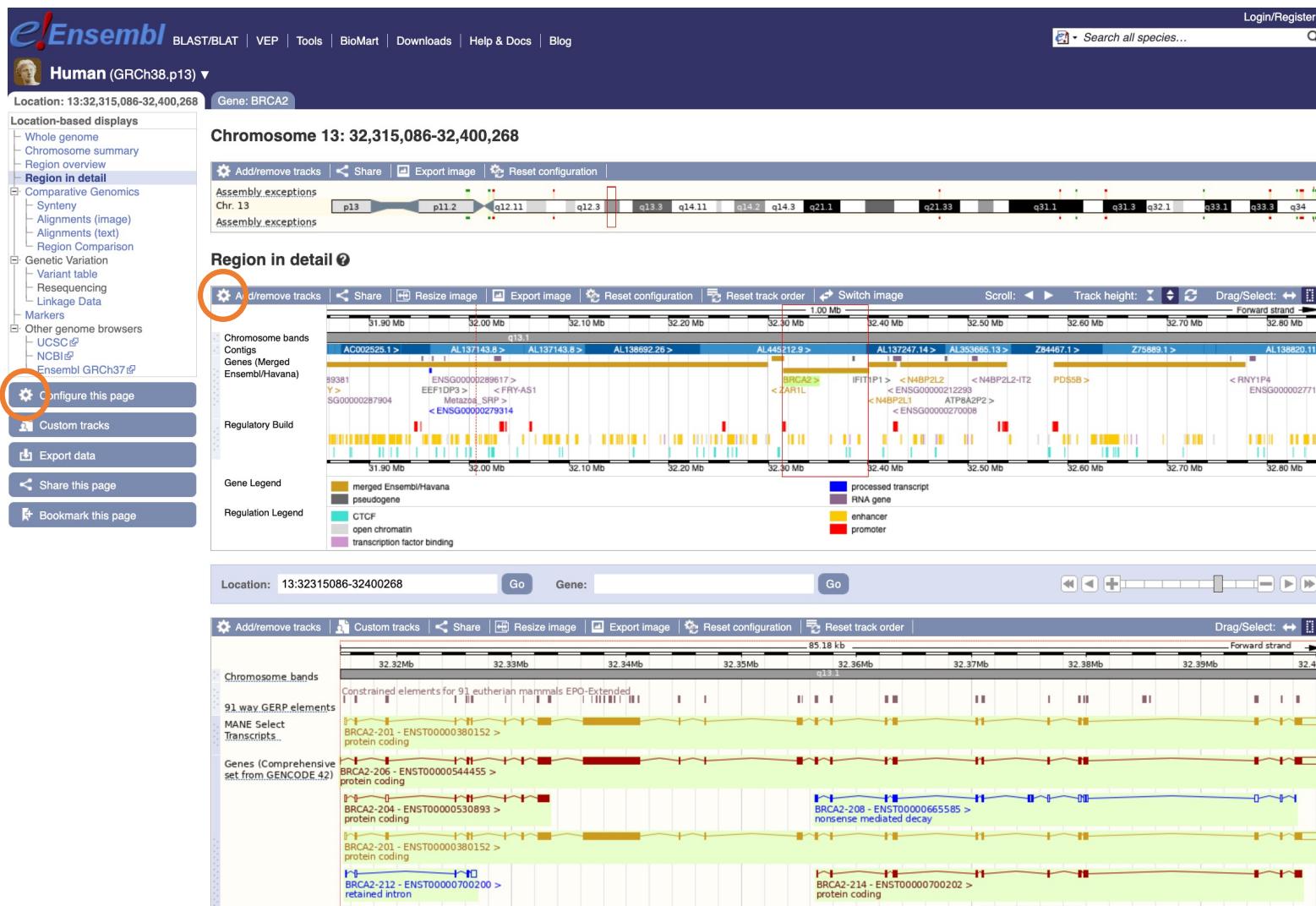
**Gene counts (Primary assembly)**

|                        |                               |
|------------------------|-------------------------------|
| Coding genes           | 19,813 (excl 651 readthrough) |
| Non coding genes       | 25,972                        |
| Small non coding genes | 4,864                         |
| Long non coding genes  | 18,887                        |
| Misc non coding genes  | 2,221                         |

# Le site web Ensembl : statistiques par chromosome



# Le site web Ensembl : navigateur de génome



# Le site web Ensembl : le gène

**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Location: 13:32,315,086-32,400,268 Gene: BRCA2

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Ontologies
  - GO: Molecular function
  - GO: Biological process
  - GO: Cellular component
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
  - Gene expression
  - Pathway
  - Regulation
  - External references
  - Supporting evidence
- ID History
- Gene history

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

**Gene: BRCA2 ENSG00000139618**

**Description** BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:[HGNC:1101](#)]

**Gene Synonyms** BRCC2, FACD, FAD, FAD1, FANCD1, XRCC11

**Location** Chromosome 13: 32,315,086-32,400,268 forward strand.  
GRCh38:CM000675.2

**About this gene** This gene has 15 transcripts ([splice variants](#)), [174 orthologues](#) and is associated with [182 phenotypes](#).

**Transcripts** Hide transcript table

| Show/hide columns (1 hidden) |           |       |            |                         |          |               |              |             |                   |                   | Filter            |       |
|------------------------------|-----------|-------|------------|-------------------------|----------|---------------|--------------|-------------|-------------------|-------------------|-------------------|-------|
| Transcript ID                | Name      | bp    | Protein    | Biotype                 | CCDS     | UniProt Match | RefSeq Match | Flags       |                   |                   |                   |       |
| ENST00000380152.8            | BRCA2-201 | 11954 | 3418aa     | Protein coding          | CCDS9344 | P51587        | NM_000059.4  | MANE Select | Ensembl Canonical | Gencode basic     | APPRIS P1         | TSL:5 |
| ENST00000680887.1            | BRCA2-210 | 11880 | 3418aa     | Protein coding          | CCDS9344 | A0A7P0T9D7    | -            |             |                   | APPRIS P1         |                   |       |
| ENST00000544455.6            | BRCA2-206 | 11854 | 3418aa     | Protein coding          | CCDS9344 | P51587        | -            |             |                   | Gencode basic     | APPRIS P1         | TSL:1 |
| ENST0000070202.1             | BRCA2-214 | 2673  | 890aa      | Protein coding          |          | -             | -            |             |                   | CDS 5' incomplete |                   |       |
| ENST00000530893.6            | BRCA2-204 | 2011  | 481aa      | Protein coding          |          | A0A590UJ1Z    | -            |             |                   | TSL:1             | CDS 3' incomplete |       |
| ENST00000614259.2            | BRCA2-207 | 11763 | 2649aa     | Nonsense mediated decay |          | A0A7P0TAP7    | -            |             |                   | TSL:2             |                   |       |
| ENST00000665585.1            | BRCA2-208 | 2598  | 438aa      | Nonsense mediated decay |          | A0A590UJU6    | -            |             |                   | CDS 5' incomplete |                   |       |
| ENST00000700201.1            | BRCA2-213 | 2103  | 129aa      | Nonsense mediated decay |          | -             | -            |             |                   | -                 |                   |       |
| ENST00000470094.1            | BRCA2-202 | 842   | 186aa      | Nonsense mediated decay |          | H0YE37        | -            |             |                   | TSL:5             | CDS 5' incomplete |       |
| ENST00000666593.1            | BRCA2-209 | 523   | 58aa       | Nonsense mediated decay |          | A0A590UJ24    | -            |             |                   | CDS 5' incomplete |                   |       |
| ENST00000528762.1            | BRCA2-203 | 495   | 64aa       | Nonsense mediated decay |          | H0YD86        | -            |             |                   | TSL:4             | CDS 5' incomplete |       |
| ENST00000700203.1            | BRCA2-215 | 2532  | No protein | Retained intron         |          | -             | -            |             |                   | -                 |                   |       |
| ENST00000700200.1            | BRCA2-212 | 860   | No protein | Retained intron         |          | -             | -            |             |                   | -                 |                   |       |
| ENST00000700199.1            | BRCA2-211 | 553   | No protein | Retained intron         |          | -             | -            |             |                   | -                 |                   |       |
| ENST00000533776.1            | BRCA2-205 | 523   | No protein | Retained intron         |          | -             | -            |             |                   | TSL:3             |                   |       |

**Summary**

|                  |   |
|------------------|---|
| Name             | BRCA2 (HGNC Symbol)   |
| MANE             | This gene contains MANE Select ENST00000380152, ENSP00000369497   |
| UniProtKB        | This gene has proteins that correspond to the following UniProtKB identifiers: P51587   |
| RefSeq           | This Ensembl/Gencode gene contains transcript(s) for which we have selected identical RefSeq transcript(s). If there are other RefSeq transcripts available they will be in the <a href="#">External references</a> table |
| CCDS             | This gene is a member of the Human CCDS set: CCDS9344.1   |
| LRG              | LRG_293 provides a stable genomic reference framework for describing sequence variants for this gene  |
| Ensembl version  | ENSG00000139618.18  |
| Other assemblies | This gene maps to 32,889,223-32,974,405 in GRCh37 coordinates.<br>View this locus in the GRCh37 archive: ENSG00000139618  |
| Gene type        | Protein coding  |

# Le site web Ensembl : le transcript

Screenshot of the Ensembl transcript page for BRCA2-201.

Header: Human (GRCh38.p13) ▾ Location: 13:32,315,086-32,400,268 Gene: BRCA2 Transcript: BRCA2-201

Transcript-based displays:

- Summary
- Sequence
  - Exons
  - cDNA
  - Protein
- Protein Information
  - Protein summary
  - Domains & features
  - Variants
  - PDB 3D protein model
  - AlphaFold predicted model
- Genetic Variation
  - Variant table
  - Variant image
  - Haplotypes
  - Population comparison
  - Comparison image
- External References
  - General identifiers
  - Oligo probes
  - Supporting evidence
- ID History
  - Transcript history
  - Protein history

Description: BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:[HGNC:1101](#)] Gene Synonyms: BRCC2, FACD, FAD1, FANCD, FANCD1, XRCC11 Location: Chromosome 13: 32,315,508-32,400,268 forward strand. About this transcript: This transcript has 27 exons, is annotated with 68 domains and features, is associated with 35622 variant alleles and maps to 958 oligo probes. Gene: This transcript is a product of gene [ENSG00000139618.18](#) Hide transcript table

Show/hide columns (1 hidden)

| Transcript ID     | Name      | bp    | Protein    | Biotype                 | CCDS     | UniProt Match | RefSeq Match | Flags   |
|-------------------|-----------|-------|------------|-------------------------|----------|---------------|--------------|---|
| ENST00000380152.8 | BRCA2-201 | 11954 | 3418aa     | Protein coding          | CCDS9344 | P51587        | NM_000059.4  | MANE Select Ensembl Canonical GENCODE basic APPRIS P1 TSL:5 |
| ENST00000680887.1 | BRCA2-210 | 11880 | 3418aa     | Protein coding          | CCDS9344 | A0A7P0T9D7    | -            | APPRIS P1   |
| ENST00000544455.6 | BRCA2-206 | 11854 | 3418aa     | Protein coding          | CCDS9344 | P51587        | -            | GENCODE basic APPRIS P1 TSL:1                               |
| ENST00000700202.1 | BRCA2-214 | 2673  | 890aa      | Protein coding          | -        | -             | -            | CDS 5' incomplete   |
| ENST00000530893.6 | BRCA2-204 | 2011  | 481aa      | Protein coding          | -        | A0A590UJ17    | -            | TSL:1 CDS 3' incomplete                                     |
| ENST00000614259.2 | BRCA2-207 | 11763 | 2649aa     | Nonsense mediated decay | -        | A0A7P0TAP7    | -            | TSL:2   |
| ENST00000665585.1 | BRCA2-208 | 2598  | 438aa      | Nonsense mediated decay | -        | A0A590UJU6    | -            | CDS 5' incomplete   |
| ENST00000700201.1 | BRCA2-213 | 129aa | 129aa      | Nonsense mediated decay | -        | -             | -            | -   |
| ENST00000470094.1 | BRCA2-202 | 842   | 186aa      | Nonsense mediated decay | -        | HOYE37        | -            | TSL:5 CDS 5' incomplete                                     |
| ENST00000666593.1 | BRCA2-209 | 523   | 58aa       | Nonsense mediated decay | -        | A0A590UJ24    | -            | CDS 5' incomplete   |
| ENST00000528762.1 | BRCA2-203 | 495   | 64aa       | Nonsense mediated decay | -        | HOYD86        | -            | TSL:4 CDS 5' incomplete                                     |
| ENST00000700203.1 | BRCA2-215 | 2532  | No protein | Retained intron         | -        | -             | -            | -   |
| ENST00000700200.1 | BRCA2-212 | 860   | No protein | Retained intron         | -        | -             | -            | -   |
| ENST00000700199.1 | BRCA2-211 | 553   | No protein | Retained intron         | -        | -             | -            | -   |
| ENST00000533776.1 | BRCA2-205 | 523   | No protein | Retained intron         | -        | -             | -            | TSL:3   |

Summary:

Export image

Statistics: Exons: 27, Coding exons: 26, Transcript length: 11,954 bps, Translation length: 3,418 residues  
MANE: This MANE Select transcript contains ENSP00000369497 and matches to NM\_000059.4 and NP\_000050.3.  
Uniprot: This transcript corresponds to the following Uniprot identifiers: P51587.  
CCDS: This transcript is a member of the Human CCDS set: CCDS9344.  
Transcript Support Level (TSL): TSL:5  
Version: ENST00000380152.8  
Type: Protein coding  
Annotation Method: Transcript where the Ensembl genebuild transcript and the Havana manual annotation have the same sequence, for every base pair. See [article](#).  
GENCODE basic gene: This transcript is a member of the Gencode basic gene set.

# Naviguer dans Ensembl : Partie pratique

# Visualiser ses propres données

**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog [Login/Register](#)

**Tools** [All tools](#)

**BioMart >** Export custom datasets from Ensembl with this data-mining tool

**BLAST/BLAT >** Search our genomes for your DNA or protein sequence

**Variant Effect Predictor >** Analyse your own variants and predict the functional consequences of known and unknown variants

**Search**  
All species for  
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

**All genomes**  
Select a species --

**Pig breeds** Pig reference genome and 12 additional breeds

[View full list of all species](#)

**Favourite genomes**

- Human** GRCh38.p13
- Still using GRCh37?
- Mouse** GRCm39
- Zebrafish** GRCz11

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.  
Note: species that already exist on this site will continue to be updated with the full range of annotations.

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate such as Darwin Tree of Life, the Vertebrate Genome Project.

**Other news from our blog**

- 02 Dec 2022: Job: Senior Full Stack Developer
- 24 Nov 2022: The first invertebrate-themed Ensembl Rapid Release
- 18 Nov 2022: Geek for a Week : Georgia Argirou

**Visualiser ses propres données**

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available. Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Permanent link - View in archive site

Ensembl release 108 - Oct 2022 © EMBL-EBI



GLOBAL CORE BIODATA RESOURCE



About Us  
About us  
Contact us

Get help  
Using this website  
Adding custom tracks

Our sister sites  
Ensembl Bacteria  
Ensembl Fungi

Follow us  
Blog  
Twitter

Navigation dans Ensembl

62

# LES OUTILS

# Les outils

[Login/Register](#)

**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog



**BLAST/BLAT**

**BioMart >** Export custom datasets from Ensembl with this data-mining tool

**BLAST/BLAT >** Search our genomes for your DNA or protein sequence

**Variant Effect Predictor >** Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

**Ensembl Release 108 (Oct 2022)**

- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWiCH consortium for chicken
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail and collared flycatcher
- Retirement of postGAP tool

[More release news](#) on our blog

**Search**

All species  for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

**All genomes**

-- Select a species --

 **Pig breeds**  
Pig reference genome and 12 additional breeds

[View full list of all species](#)

**Favourite genomes**

 **Human**  
GRCh38.p13  
[Still using GRCh37?](#)

 **Mouse**  
GRCm39

 **Zebrafish**  
GRCz11

**Ensembl Rapid Release**

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

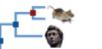
The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

**Other news from our blog**

- 02 Dec 2022: [Job: Senior Full Stack Developer](#)
- 24 Nov 2022: [The first invertebrate-themed Ensembl Rapid Release is out!](#)
- 18 Nov 2022: [Geek for a Week : Georgia Argirogi](#)

**Compare genes across species**



**Find SNPs and other variants for my gene**

  
GATATACATTC  
CTTAAAGTCCTT  
CTTCTAATTCT  
GAACATTTCC

**Gene expression in different tissues**



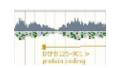
**Retrieve gene sequence**

```
GGCTGAGCTTCCGCGTGCG  
GGCGCTTGTGCGCGCGCGCG  
GGCGCTCTGCGCGCGCGCG  
AGGGGAGAGATTGTGTG  
CACCTCTGGAAACCGGGTTT  
GCCAGTCGCGCGCGCGCG
```

**Find a Data Display**



**Use my own data in Ensembl**



EMBL-EBI  Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at [EMBL-EBI](#) and our software and data are freely available. Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

GLOBAL CORE BIODATA RESOURCE 

Ensembl release 108 - Oct 2022 © EMBL-EBI

[Permanent link](#) - [View in archive site](#)

## About Us

About us

Contact us

Get help

Using this website

### [Adding custom tracks](#)

Our sister sites

Ensembl Bacteria

Ensembl Fungi

Follow us



# Blast



- Recherche de similarité
  - 1 séquence (**Query**) comparée à des milliers ou des millions de séquences (**base de données**) par comparaison 2 à 2.
- But:
  - Déetecter des séquences proches
  - Annotation simple (domaines protéiques, localisation génomique, nombre d'exons)

# Les différentes comparaisons

## BLAST : Basic Local Alignment Search Tool

Altschul *et al.* Basic local alignment search tool. *J. Mol. Biol.* 1990

Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997

| Programmes | Requête                         | Banque                          | Comparaison | Exemples d'utilisation  |
|------------|---------------------------------|---------------------------------|-------------|---|
| Blastn     | ADN                             | ADN                             | nucléique   | Recherche d'ARN structuraux, d'éléments régulateurs                               |
| Blastp     | Protéine                        | protéines                       | protéique   | Recherche de protéines homologues   |
| Tblastn    | Protéine                        | ADN (traduit dans les 6 cadres) | protéique   | Recherche de similarités entre une protéine et une séquence génomique mal annotée |
| Blastx     | ADN (traduit dans les 6 cadres) | protéines                       | protéique   | Recherche des phases de lecture dans une séquence codante                         |
| Tblastx    | ADN (traduit dans les 6 cadres) | ADN (traduit dans les 6 cadres) | protéique   | Avantages de tblastn et blastx mais très long                                     |

# Les différentes comparaisons

## BLAT (BLAST-Like Alignment Tool)

- An mRNA/DNA and cross-species protein sequence analysis tool to quickly find sequences of  $\geq 95\%$  similarity of length  $\geq 40$  bases.
- was developed by Jim Kent at the University of California Santa Cruz (UCSC) in the early 2000s to assist in the assembly and annotation of the human genome.
- The target database of BLAT is not a set of GenBank sequences, but instead an index derived from the assembly of the entire genome. **Blat works by keeping an index of an entire genome in memory.**
- By default, the index consists of all non-overlapping 11-mers for DNA and 4-mers for protein.
- Kent, W.J.. BLAT -- The BLAST-Like Alignment Tool. *Genome Research* 2002

# Blast



MADTQYILPNDIGVSSLDCREAFRLSPTERLYAYHLSRAAWYGGLAVLLQTSPEAPYIYALLSRLFRAQDP  
DQLRQHALAEGLTEEEYQAFLVVAAGVYSNMGNYSFGDTKFVNPNLKEKLERVILGSEAAQQHPEEVRG  
QTCGELMFSLEPRLRHLGLGKEGITYFSGNCTMEDAKLAQDFLDSQNLSAYNTRLFKEVDGEGKPYYEV  
ASVLGSEPSLDSEVTSKLKSYEFRGSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIESFT  
SIEAHKRGSRFWIQDKGPIVESYIGFIESYRDPFGSRGEFEVFVAVVNKAMSAKFERLVASA  
EQLLKELPWP  
PTFEKDKFLTPDFTSLDVLTFAFGSGIPAGINIPNYDDLQRTEGFKNVSLGNVL  
AVAYATQREKLT  
FLEEDDK  
DLYILWKGPSFDVQVGLHELLGHGSGKLFVQDEKGAFNFDQETVINPETGEQIQSWYRS  
GETWD  
SKFSTIAS  
SYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIYVNWL  
NMVRAGLLALEFY  
TPEAFNWRQAHMQARF  
VIL  
RVILLEAGEGLVTITPTTGSDGRPDARVRLDRSKIRSVGKPA  
LERFLRRIQVL  
KSTGDVAGGRAL  
YEGYAT  
VT  
DAPPECFLTLRDTVLLRKESRK  
LIVQP  
NTRLEGSDVQL  
LEYEASAAGL  
IRSF  
SERFP  
EDGPE  
LEEILT  
QLAT  
ADARFWKG  
PSEAPSGQA

new SETUP CONFIG RESULTS DISPLAY refresh Online Help

Summary

► setup

- Homo\_sapiens
- Genomic sequence
- TBLASTN
- Low sensitivity

► configure

- -E: 10
- -B: 100
- -filter: seg
- -W: 4
- -hitdist: 40
- -matrix: BLOSUM80
- -T: 16

► results

► display

① Not yet initialised

Retrieval result for ID:  
BLAIESYdDXDJ      Retrieve

Alignment Display Options:

Locations vs. Karyotype     Locations vs. Query  
 Summary Table

1: unnamed (737 letters) Vs. LATESTGP  
Homo\_sapiens 1981 alignments, 23 hits      [\[RawResult\]](#)      [view ▶](#)

We would like to hear your impressions of blastview, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback](#)

Content-type: text/plain

TBLASTN 2.0MP-WashU [04-May-2006] [linux26-x64-I32LPF64 2006-05-10T17:22:28]

Copyright (C) 1996-2006 Washington University, Saint Louis, Missouri USA.  
All Rights Reserved.

Reference: Gish, W. (1996-2006) <http://blast.wustl.edu>

Query= unnamed  
(737 letters)

WARNING: Precomputed values for Lambda, K and H are unavailable for the BLOSUM80 scoring matrix, when used with gap penalties +9 and +2. Unless overridden on the command line, the values computed for ungapped alignments will be used instead, but the reported E-values and P-values may be much too low.

Database: Homo\_sapiens.GRCh37.dna.toplevel.fa  
297 sequences; 32,036,512,383 total letters.

WARNING: Use of the hspsepSmax parameter should be considered with long database sequences, to improve the biological relevance of the HSP groups that are assembled and to improve the statistical discrimination of these groups from random background.

Searching....10....20....30....40....50....60....70....80....90....100% done

WARNING: hspmax=1000 was exceeded by 37 of the database sequences, causing the associated cutoff score, S2, to be transiently set as high as 73.

| Sequences producing High-scoring Segment Pairs:                | Smallest Sum |       |             |    |
|--|--------------|-------|-------------|----|
|  | Reading      | High  | Probability |    |
|  | Frame        | Score | P(N)        | N  |
| 9 dna:chromosome chromosome:GRCh37:9:1:141213431:1 REF         | -3           | 1765  | 0.          | 6  |
| 11 dna:chromosome chromosome:GRCh37:11:1:135006516:1 REF       | +3           | 763   | 3.2e-292    | 9  |
| 4 dna:chromosome chromosome:GRCh37:4:1:191154276:1 REF         | +3           | 1542  | 5.5e-250    | 4  |
| 20 dna:chromosome chromosome:GRCh37:20:1:63025520:1 REF        | -1           | 131   | 0.0035      | 9  |
| 16 dna:chromosome chromosome:GRCh37:16:1:90354753:1 REF        | +1           | 120   | 0.014       | 10 |
| 12 dna:chromosome chromosome:GRCh37:12:1:133851895:1 REF       | -2           | 126   | 0.060       | 11 |
| 19 dna:chromosome chromosome:GRCh37:19:1:59128983:1 REF        | -1           | 128   | 0.069       | 9  |
| 22 dna:chromosome chromosome:GRCh37:22:1:51304566:1 REF        | +1           | 130   | 0.10        | 10 |
| GL000199.1 dna:supercontig supercontig:GRCh37:GL000199.1:...+3 | +3           | 149   | 0.11        | 2  |
| 14 dna:chromosome chromosome:GRCh37:14:1:107349540:1 REF       | +2           | 167   | 0.21        | 8  |
| 1 dna:chromosome chromosome:GRCh37:1:1:249250621:1 REF         | -1           | 134   | 0.25        | 8  |
| GL000220.1 dna:supercontig supercontig:GRCh37:GL000220.1:...-3 | -3           | 124   | 0.26        | 4  |
| 5 dna:chromosome chromosome:GRCh37:5:1:180915260:1 REF         | +1           | 127   | 0.33        | 9  |
| GL000224.1 dna:supercontig supercontig:GRCh37:GL000224.1:...-2 | -2           | 126   | 0.49        | 2  |
| 7 dna:chromosome chromosome:GRCh37:7:1:159138663:1 REF         | -3           | 129   | 0.88        | 9  |
| 21 dna:chromosome chromosome:GRCh37:21:1:48129895:1 REF        | -2           | 131   | 0.98        | 9  |
| GL000237.1 dna:supercontig supercontig:GRCh37:GL000237.1:...-2 | -2           | 89    | 0.98        | 5  |
| GL000202.1 dna:supercontig supercontig:GRCh37:GL000202.1:...+1 | +1           | 111   | 0.995       | 3  |
| GL000218.1 dna:supercontig supercontig:GRCh37:GL000218.1:...-1 | -1           | 145   | 0.996       | 5  |
| 15 dna:chromosome chromosome:GRCh37:15:1:102531392:1 REF       | +2           | 134   | 0.999       | 12 |
| 6 dna:chromosome chromosome:GRCh37:6:1:171115067:1 REF         | -2           | 118   | 0.9991      | 13 |
| 3 dna:chromosome chromosome:GRCh37:3:1:198022430:1 REF         | -3           | 118   | 0.9998      | 11 |
| GL000206.1 dna:supercontig supercontig:GRCh37:GL000206.1:...-3 | -3           | 92    | 0.99992     | 6  |

>9 dna:chromosome chromosome:GRCh37:9:1:141213431:1 REF  
Length = 141,213,431

Score = 1765 (578.9 bits), Expect = 0., Sum P(6) = 0.  
Identities = 220/261 (84%), Positives = 230/261 (88%), Frame = -3

|  |  |          |
|--|--|----------|
| Query:   | 477 INPETGEOIQOSWYRSGETWDSKFTIASSYECRAESVGLYCLHPOVLEIFGFEGADAE           | 536      |
| Sbjct:   | 76090065 INPE EQIQSWYRS +TWDSKFSTI SSYECCRASVGLYCLHPOVLE FGFEAGADE       | 76089886 |
| Query:   | 537 DVIVVNWLNMVRAGLLALEFYTFPEAFNWRQAHMOARFVILRVLLEAGEGLVTITPTTGS         | 596      |
| Sbjct:   | 76089885 +VI VNWLNMV AGILLEAFYTFPEA NW+QAH++AR VILRVL EAGEGL TITPT GSD   | 76089706 |
| Query:   | 597 GRPDARVRLDRSKIRSVGKPALERFLRRLOVLKSTGDVAGGRALYEGYATVTDAPPECFL         | 656      |
| Sbjct:   | 76089705 GRP+A+VRLDRSKI+SVG PALERFLRR STGDVAGG LYE YA V DAPPE FL         | 76089535 |
| Query:   | 657 TLRDTVLLRKESRKLIIVQPNTRLLEGSDVOLLEYEASAAGLIRSFSEFPEDGPELEEILT        | 716      |
| Sbjct:   | 76089534 TLRD VLLRKES KLIIVQPNTLLEGSDVOLLEYEASAAGLIRSFSEHFPEDGLEDILT     | 76089355 |
| Query:   | 717 QLATADARFWKGSEAPSGOA 737   |          |
| Sbjct:   | 76089354 QLATADAOF*KGPSEAPSGOA 76089292                                  |          |
| Score = 1700 (557.6 bits), Expect = 0., Sum P(6) = 0.<br>Identities = 212/252 (84%), Positives = 221/252 (87%), Frame = -2 |  |          |
| Query:   | 224 PSLDSEVTSKLKSYEFRGSPFQVTTRGDYAPILKQVVEQLEKAKAYAANSHQGQMQLAQYIE       | 283      |
| Sbjct:   | 76090816 P L + SKLKS EFRGSPFQVT G+Y PILQKVVQELEKAK YAANSHQ QMLAQYIE      | 76090643 |
| Query:   | 284 SFTQGSIEAHKRGSRFWI QDKGPIVESYI F+SYRD FGSRG EGFVAVVNKAMSAKF          | 343      |
| Sbjct:   | 76090642 SFTQGSTEAHKKGSRFWI*DKGPIVESYIEFIQS YRDSFGSRGVCEGFVAVVNKAMSAKF   | 76090463 |
| Query:   | 344 ERLVASAEEQLLKELPWPPTFEKDFKFLTPDFTSLDVLTFA GSGIPAGINIPNYDDL RQTEG     | 403      |
| Sbjct:   | 76090462 E WL VSAEQLLKELPWP FEKDFKFLTPDFTS+DVL TFA GSGI AGINI NY+DL+QTEG | 76090283 |
| Query:   | 404 FKNVSLGNVLAVAYATQREKLTFL EDDKDLYILWKGPSFDVQVGLHELLGHGSGKLFVQ         | 463      |
| Sbjct:   | 76090282 FKNVSLGNVLAV ATQ EKL T LEE DKDLYI+ GPSFDVQVGLHELLG+GSGKL Q      | 76090103 |
| Query:   | 464 DEKGAFNFDQET 475   |          |
| Sbjct:   | 76090102 DEKGAFNFDQET 76090067   |          |

new    SETUP    CONFIG    RESULTS    DISPLAY

refresh    Online Help

### Summary

► setup

- Homo\_sapiens
- Genomic sequence
- TBLASTN
- Low sensitivity

► configure

- -E: 10
- -B: 100
- -filter: seg
- -W: 4
- -hitdist: 40
- -matrix: BLOSUM80
- -T: 16

► results

► display

① Not yet initialised

Retrieve result for ID:

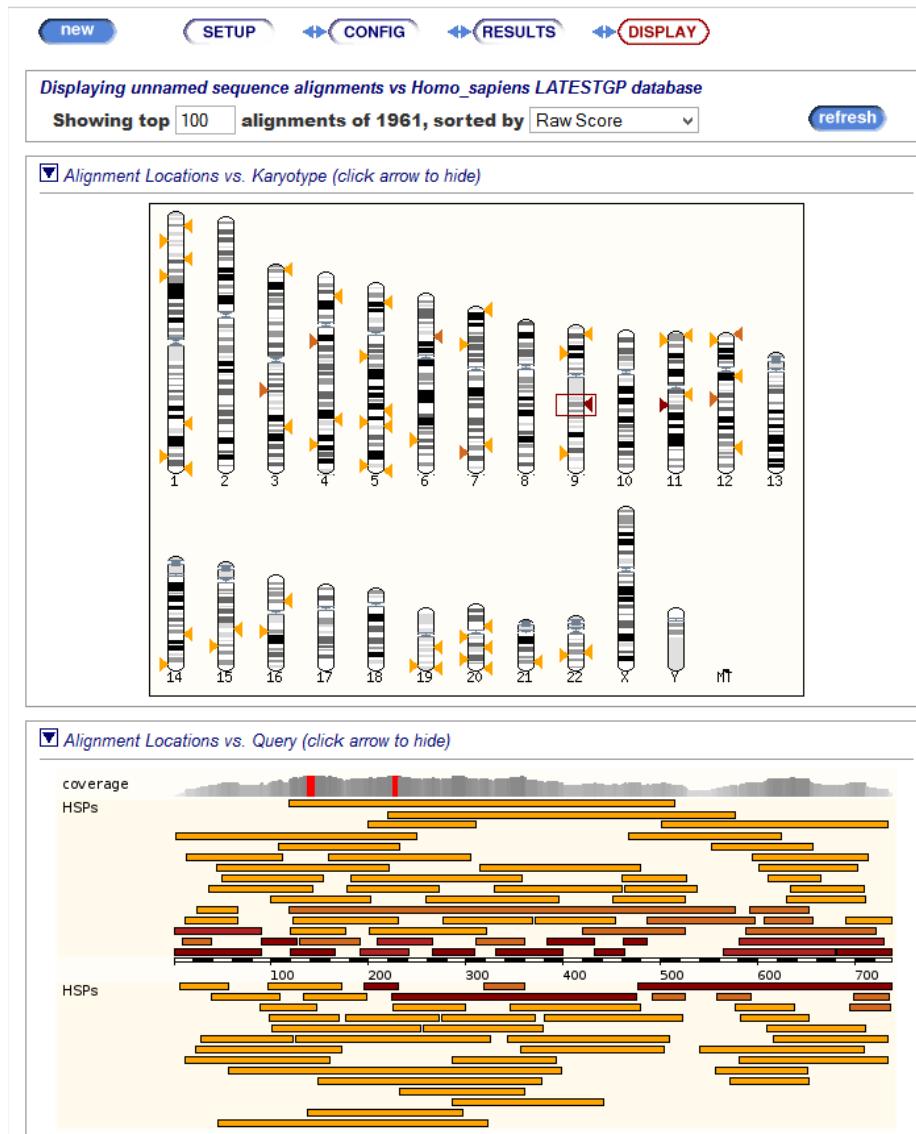
BLA\_IESTYdDXDJ    Retrieve

Alignment Display Options:

Locations vs. Karyotype     Locations vs. Query  
 Summary Table

1: unnamed (737 letters) Vs. LATESTGP

Homo\_sapiens 1961 alignments, 23 hits    [RawResult]    **view ►**



refresh Online Help

### Summary

- setup
  - *Homo\_sapiens*
  - Genomic sequence
  - TBLASTN
  - Low sensitivity
- configure
  - -E: 10
  - -B: 100
  - -filter: seg
  - -W: 4
  - -hdist: 40
  - -matrix: BLOSUM80
  - -T: 16
- results
- display
  - ① Not yet initialised

Alignment Summary (click arrow to hide)

Select rows to include in table, and type of sort  
(Use the 'ctrl' key to select multiples)

refresh

| Query   | Subject                | Chromosome             | Supercontig            | Clone                  | Contig                 | Lrg                    | Stats                   | Sort By                  |
|---|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-------------------------|--------------------------|
| _off_<br>Name<br>Start  | _off_<br>Name<br>Start | _off_<br>Name<br>Start | _off_<br>Name<br>Start | _off_<br>Name<br>Start | _off_<br>Name<br>Start | _off_<br>Name<br>Start | _off_<br>Score<br>E-val | >Lrg<br><Score<br>>Score |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 477 737 +              | <a href="#">Chr:9</a>  | 76089292               | 76090065 -             | 1765 0.                | 84.29                  | 261                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 224 475 +              | <a href="#">Chr:9</a>  | 76090067               | 76090816 -             | 1700 0.                | 84.13                  | 252                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 119 577 +              | <a href="#">Chr:4</a>  | 65296878               | 65298248 +             | 1542 5.5e-250          | 49.70                  | 497                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 581 729 +              | <a href="#">Chr:4</a>  | 65298493               | 65298530 +             | 854 5.5e-250           | 74.83                  | 151                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 1 90 +                 | <a href="#">Chr:11</a> | 66249672               | 66249941 +             | 763 3.2e-292           | 100.00                 | 90                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 330 399 +              | <a href="#">Chr:11</a> | 66260186               | 66260395 +             | 552 3.2e-292           | 95.71                  | 70                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 565 679 +              | <a href="#">Chr:11</a> | 66264763               | 66265104 +             | 531 3.2e-292           | 63.71                  | 124                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 1 90 +                 | <a href="#">Chr:4</a>  | 65296627               | 65296899 +             | 529 5.5e-250           | 76.09                  | 92                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 588 721 +              | <a href="#">Chr:11</a> | 66271972               | 66272364 +             | 487 1.7e-276           | 55.63                  | 142                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 681 737 +              | <a href="#">Chr:11</a> | 66276549               | 66276719 +             | 477 3.2e-292           | 100.00                 | 57                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 120 166 +              | <a href="#">Chr:11</a> | 66254008               | 66254148 +             | 391 1.8e-273           | 97.87                  | 47                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 420 526 +              | <a href="#">Chr:11</a> | 66262674               | 66262961 +             | 384 3.2e-292           | 53.57                  | 112                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 486 597 +              | <a href="#">Chr:11</a> | 66263006               | 66263296 +             | 377 1.7e-276           | 51.72                  | 116                     |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 266 309 +              | <a href="#">Chr:11</a> | 66258962               | 66259093 +             | 375 3.2e-292           | 97.73                  | 44                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 209 266 +              | <a href="#">Chr:11</a> | 66258657               | 66258854 +             | 370 3.2e-292           | 75.76                  | 66                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 384 432 +              | <a href="#">Chr:11</a> | 66260513               | 66260650 +             | 310 5.1e-263           | 83.67                  | 49                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 90 126 +               | <a href="#">Chr:11</a> | 66252641               | 66252751 +             | 272 3.2e-292           | 89.19                  | 37                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 432 463 +              | <a href="#">Chr:11</a> | 66261009               | 66261104 +             | 270 1.7e-276           | 96.88                  | 32                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 192 242 +              | <a href="#">Chr:11</a> | 66255385               | 66255576 +             | 268 1.3e-266           | 64.06                  | 64                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 196 230 +              | <a href="#">Chr:9</a>  | 76090801               | 76090905 -             | 257 0.                 | 88.57                  | 35                      |                          |
| <a href="#">[A]</a> <a href="#">[S]</a> <a href="#">[G]</a> <a href="#">[C]</a> | 129 191 +              | <a href="#">Chr:11</a> | 66254628               | 66254813 +             | 248 3.2e-292           | 56.06                  | 66                      |                          |

[A] [S] [G] [C] 477 737 + Chr:9 76089292 76090065 - 1765 0. 84.29 261

[A] [S] [G] [C]

[G]Sequence

5' Flanking sequence 300 (bp)

3' Flanking sequence 300 (bp)

Coordinate system Chromosome

Orientation Forward relative to selected alignment

Alignment markup All alignments Both orientations

Feature markup Ensembl exons Both orientations

Line numbering No numbers

[A]lign

```

Alignment score : 1765
E-value : 0.
Alignment length : 261
Percentage identity: 84.29

Query:   477 INPEEQI5WVRSGETWDKFSTIASSYEECRAESVGLYLCLHQPQVLIEFGFEGADDE 536
         INPE EQIQ5WVRS +IWDSKFSTI SSYEECRAESVGLYLCLHQPQVLIEFGFEGADDE
Sbjct: 76090065 INPEMREQI5WVRSKMTWDKFSTI5VSSEYEECRAESVGLYLCLHQPQVLIEFGFEGADDE 76089886

Query:   537 DIVLYNWLMLMVYIPEAFNNRNQAHM5QAREFVILVLLVEAGLGLVII+TITGSD 596
         +VI VNHLMLMV AGILLALEYYTFPEA NW+QAH+ =AR VLRLV PEAGEGL TITFT PL
Sbjct: 76089885 EVISUNWLMLMVAGILLALEYYTFPEASWNQAHIRARIVILRVLVPEAGEGLVITTPAGSD 76089706

Query:   597 GRFDARVLRDLRSK1RSV5GKPALERFLRRLQVL3TGWAGGRALYEYH+WTDAPECFL 656
         GRFA+A+VLRDLRSK1+SVC PALERFLRRLQVL3TGWAGGRALYEYH+WTDAPECFL
Sbjct: 76089705 GRFEPAWVLRDLRSK1QVGVNALALERFLRRCW--STGDVAGGWTLEIYANADAPPEGFL 76089593

Query:   657 TLRDTVLLRRESKRLLIVQPNTRLEGSDVQLLEYEASAAGLIRS+ERFPEDGELEIILT 716
         TLRD VLLRKES KLIQVNP RLEGSDVQLLEYEASAAGLIRS SE FPDG ELE+ILT
Sbjct: 76089534 TLRD VLLRKESKWLIVQPNTRLEGSDVQLLEYEASAAGLIRS FSEMFHDGLELEDIT 76089355

Query:   717 QLATADARWVGPSEAPSQGA 737
         QLATADAF+ KGSEAPSQGA
Sbjct: 76089354 QLATADAF+ KGSEAPSQGA 76089292

```

[S]equence 

|                   |   |         |          |    |         |     |
|-------------------|---|---------|----------|----|---------|-----|
| Query location    | : | unnamed | 477      | to | 737     | (+) |
| Database location | : | 9       | 76089292 | to | 7609065 | (-) |
| Genomic location  | : | 9       | 76089292 | to | 7609065 | (-) |

```
Alignment score      : 1765
E-value              : 0.
Alignment length     : 261
Expect              : 24.26
```

```

percentage identity: 84.29

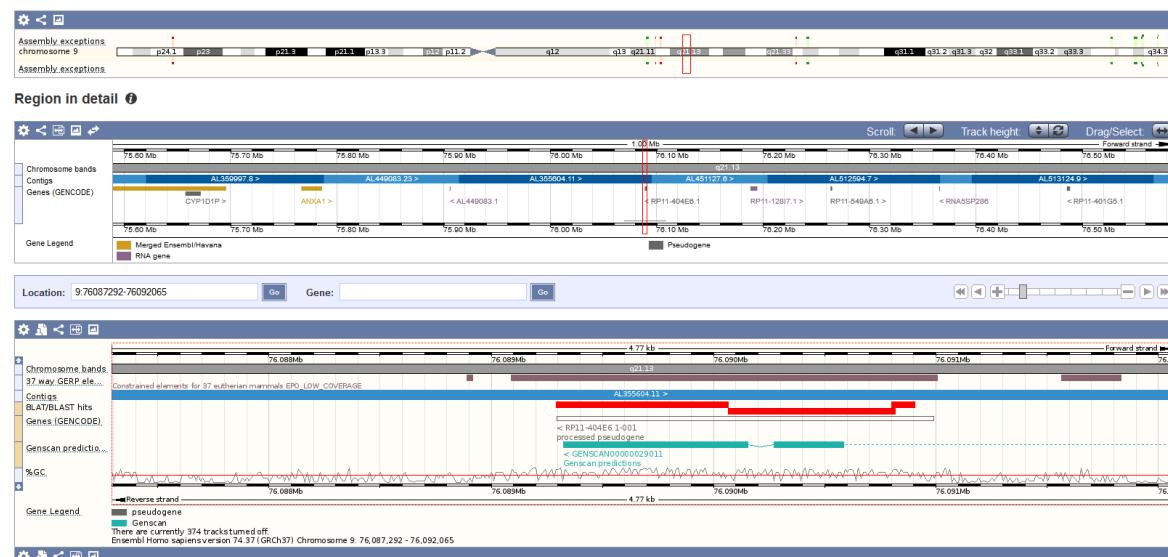
THIS STYLE: Matching bases for selected HSP
THIS STYLE: Matching bases for other HSPs in selected hit

>unnamed
MADTQYILPNDIGVSSLDCREAFRLLSPTERLYAHHLRSAAWYGGGLAVLQLQTSPPEAPYI
ALLSRFLRAQDPDQLRQHALAEGLTEEEYQAFLVYAAVGVSNMGNYKSFDTKFVNPNLPI
EKLERVLGSEAAQHPEEVEGLWLWTCGELMFSLEPRRLRHGLGEKITTGFSGNCTMEL
AKLAQDFLDSQNLSAYNTRLFKEVDGEKGPKYYEVRLASVGLGEPSLSDSEVTSKLKSYEFT
GSFPVTRGDYAPILOKVVQELEKAKAYAANSHQGQMLAQYIESFTQGSIIEAHKGRSGL
IQQDKGPIVESYQFIESYQFIESYQFIESYQFIESYQFIESYQFIESYQFIESYQFIESYQF
PTFEKDCKFLTPDFTSLDVLTFAGSGIAPAGINIPNYDDLRQTEGFKNVSLGNVLAVAYAT
REKLTFLLEEDDKDLYIWKGPSPFDVGVGLHELLGHGSCKLFVQDEKGAFNFDQETVNP
TGEQIQISWYRSGETWDSKFTIASSYECRAESVGLYLCLHCPVQLEIFGEGADADEVIP
VNWLNMVRAGLLALEFTYPEAFNWQRQAHMQARFVILVRLEAGEGLVTITPTGSDGRP
ARVRLDRSKRISVKGPALERFLRLQVLKSTGDVAGGRALYEGYATVTDAPPECFLTLR
TVLRLRSRKLWYQPNTRLEGSDVQYLLEYEASAAGLIRLSFSERFPEDGPELEEILTQLA
ADARFWKGPSSEAPSGOA

```

## [C]ontig view (?)

Chromosome 9: 76 087 292-76 092 065



# Les outils

# Annotation de variants



BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search all species...

Login/Register

**Tools**

[BioMart >](#)  
Export custom datasets from Ensembl with this data-mining tool

[All tools](#)

**BLAST/BLAT >**  
Search our genomes for your DNA or protein sequence

**Variant Effect Predictor >**  
Analyse your own variants and predict the functional consequences of known and unknown variants

**Search**  
 for   
e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

**All genomes**

**Pig breeds**  
Pig reference genome and 12 additional breeds  
  
[View full list of all species](#)

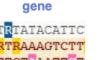
**Favourite genomes** 

 **Human**  
GRCh38.p13  
[Still using GRCh37??](#)

 **Mouse**  
GRCm39

 **Zebrafish**  
GRCz11

**Compare genes across species**  


**Find SNPs and other variants for my gene**  
  
GTTATACATT  
CCTAAAGTCTT  
CTCTTAATT  
GACATTTC

**Gene expression in different tissues**  


**Retrieve gene sequence**  
  
GCTCTGCTTCGCGATGG  
GCGCTTGTCTGGCGACGC  
GGCCCTCTGCTGGCGCTT  
AAGGCGACATTTGTGAG  
CACCTCTGCGCGCGTT  
CCCCATCGCACTGGCGCG

**Find a Data Display**  


**Use my own data in Ensembl**  
  
BAM (9125-90) >  
protein coding

EMBL-EBI  Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at [EMBL-EBI](#) and our software and data are freely available.  
Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

GLOBAL CORE BIODATA RESOURCE  elixir Core Data 

Document ID: Mgmt-00100000000000000000000000000000

# Variant Effect Predictor

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species...

Using this website Annotation and prediction Data access API & software About us

Help & Documentation API & Software Ensembl Tools Ensembl Variant Effect Predictor (VEP)

Ve!P

In this section

- VEP web interface
  - Input form
  - Results
- VEP command line
  - Tutorial
  - Download and install
  - Running VEP
  - Annotation sources
  - Filtering results
  - Custom annotations
  - Plugins
  - Examples and use cases
  - Other information
- Data formats
- Variant Recoder
- HaploSaurus
- VEP FAQ

On this page

- VEP interfaces
- Publication
- VEP related tools

Search documentation Go

### Ensembl Variant Effect Predictor (VEP)

**VEP determines the effect of your variants** (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- Genes and Transcripts affected by the variants
- Location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- Consequence of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift), see [variant consequences](#)
- Known variants that match yours, and associated minor allele frequencies from the [1000 Genomes Project](#)
- SIFT and PolyPhen-2 scores for changes to protein sequence
- ... And more! See [data types, versions](#).

★ [What's new in release 108?](#)

#### VEP interfaces

**Web interface**  
Point-and-click interface  
Suits smaller volumes of data  
[Documentation](#)

**Command line tool**  
More options and flexibility  
For large volumes of data  
[Documentation](#)

**REST API**  
Language-independent API  
Simple URL-based queries  
[Documentation](#)

[Launch Ve!P](#)

**Publication**

If you use VEP, please cite our UPDATED publication so we can continue to support VEP development:

Cite us

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biology* Jun 6;17(1):122. (2016) doi:10.1186/s13059-016-0974-4 CR

VEP related tools

# Variant Effect Predictor

The screenshot shows the Ensembl Variant Effect Predictor (VEP) interface. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there are buttons for Login/Register and a search bar labeled "Search all species...". The main content area has a sidebar titled "Web Tools" containing links for BLAST/BLAT, Variant Effect Predictor (which is selected and highlighted in blue), Linkage Disequilibrium Calculator, Variant Recoder, File Chameleon, Assembly Converter, ID History Converter, VCF to PED Converter, and Data Slicer. Below this are buttons for "Configure this page", "Custom tracks", "Export data", "Share this page", and "Bookmark this page". The main form is titled "Variant Effect Predictor" and includes fields for "Species" (set to Homo\_sapiens), "Name for this job (optional)", "Input data" (with a text area for pasting data and options to upload a file or provide a URL), "Transcript database to use" (radio buttons for Ensembl/Gencode transcripts, Ensembl/Gencode basic transcripts, RefSeq transcripts, and Ensembl/Gencode and RefSeq transcripts, with the first option selected), and "Additional configurations" (checkboxes for Identifiers, Variants and frequency data, Additional annotations, Predictions, Filtering options, and Advanced options). A large green "Run" button is at the bottom.

## Recent jobs

You have no jobs currently running or recently completed.

Ensembl release 108 - Oct 2022 © EMBL-EBI

# Variant Effect Predictor

**Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species...

**VEP** ▾

Web Tools

- Web Tools
- BLAST/BLAT
- Variant Effect Predictor
- VEP analysis of pasted data
- Linkage Disequilibrium Calculator
- Variant Recoder
- File Chameleon
- Assembly Converter
- ID History Converter
- VCF to PED Converter
- Data Slicer
- Post-GWAS

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

### Variant Effect Predictor results ⓘ

Job details ⓘ

Summary statistics ⓘ

| Category                       | Count |
|--------------------------------|-------|
| Variants processed             | 3     |
| Variants filtered out          | 0     |
| Novel / existing variants      | -     |
| Overlapped genes               | 5     |
| Overlapped transcripts         | 46    |
| Overlapped regulatory features | 1     |

**Consequences (all)**

**Coding consequences**

**Results preview**

Navigation (per variant) | Filters | Download

Page: 1 of 1 | Show: All variants | Uploaded variant is defined Add All: VCF VEP TXT BioMart: Variants Genes

New job

Show/hide columns (13 hidden)

| Uploaded variant | Location        | Allele | Consequence             | Symbol  | Gene            | Feature type | Feature            | Scroll to see more columns >       |
|------------------|-----------------|--------|-------------------------|---------|-----------------|--------------|--------------------|------------------------------------|
| 1_65568_A/C      | 1:65568-65568   | C      | downstream_gene_variant | OR4G11P | ENSG00000240361 | Transcript   | ENST00000492842.2  | transcribed_unprocessed_pseudogene |
| 1_65568_A/C      | 1:65568-65568   | C      | missense_variant        | OR4F5   | ENSG00000186092 | Transcript   | ENST00000641515.2  | protein_coding                     |
| 1_65568_A/C      | 1:65568-65568   | C      | downstream_gene_variant | OR4G11P | ENSG00000240361 | Transcript   | ENST00000642116.1  | processed_transcript               |
| 2_265023_C/T     | 2:265023-265023 | T      | intron_variant          | ACP1    | ENSG00000143727 | Transcript   | ENST00000272065.10 | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | intron_variant          | ACP1    | ENSG00000143727 | Transcript   | ENST00000272067.10 | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | upstream_gene_variant   | SH3YL1  | ENSG00000035115 | Transcript   | ENST00000356150.10 | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | upstream_gene_variant   | SH3YL1  | ENSG00000035115 | Transcript   | ENST00000402632.5  | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | upstream_gene_variant   | SH3YL1  | ENSG00000035115 | Transcript   | ENST00000403657.5  | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | upstream_gene_variant   | SH3YL1  | ENSG00000035115 | Transcript   | ENST00000403658.5  | protein_coding                     |
| 2_265023_C/T     | 2:265023-265023 | T      | upstream_gene_variant   | SH3YL1  | ENSG00000035115 | Transcript   | ENST00000403712.6  | protein_coding                     |

# Outils de récupération de données

The screenshot shows the Ensembl website's "Data access" section. A red box highlights the "Downloads" menu item. The page content includes:

- Accessing Ensembl Data**: Describes routes for fetching data, noting one-based starts.
- Small quantities of data**: Shows an "Export data" button and a FASTA sequence example: CAGATGAT AAATGTTCT AAAGAAGCA CTGGATGTC ATAAAGAAA AGTGATACT.
- Fast programmatic access**: Recommends using the REST server for fast access in programming languages.
- Complete datasets and databases**: Shows a database icon and mentions MySQL dumps via FTP, with an orange arrow pointing to the "FTP site" link.
- Complex cross-database queries**: Illustrates a funnel merging multiple data streams into a single output.

At the bottom, a note states: "All data produced by the Ensembl project is [freely available](#) for your own use."

Ensembl release 108 - Oct 2022 © EMBL-EBI

[Permanent link](#)

## About Us

- [About us](#)
- [Contact us](#)
- [Citing Ensembl](#)
- [Privacy policy](#)
- [Disclaimer](#)

## Get help

- [Using this website](#)
- [Adding custom tracks](#)
- [Downloading data](#)
- [Video tutorials](#)
- [Variant Effect Predictor \(VEP\)](#)

## Our sister sites

- [Ensembl Bacteria](#)
- [Ensembl Fungi](#)
- [Ensembl Plants](#)
- [Ensembl Protists](#)
- [Ensembl Metazoa](#)

## Follow us

- [Blog](#)
- [Twitter](#)
- [Facebook](#)

# BioMart

# Le projet BioMart

- <http://www.biomart.org/>
- Développé conjointement par :
  - EBI
  - Cold Spring Harbor Laboratory (CSHL)
- Arek Kasprzyk : « BioMart can access diverse databases from a single interface »
- Créer un système générique de stockage et de gestion de données
- « Data-agnostic » : manipulation de n'importe quel type de donnée avec le même software
- Applicable à
  - Tout type de données descriptives (y compris des données biologiques)
  - de grands volumes de données

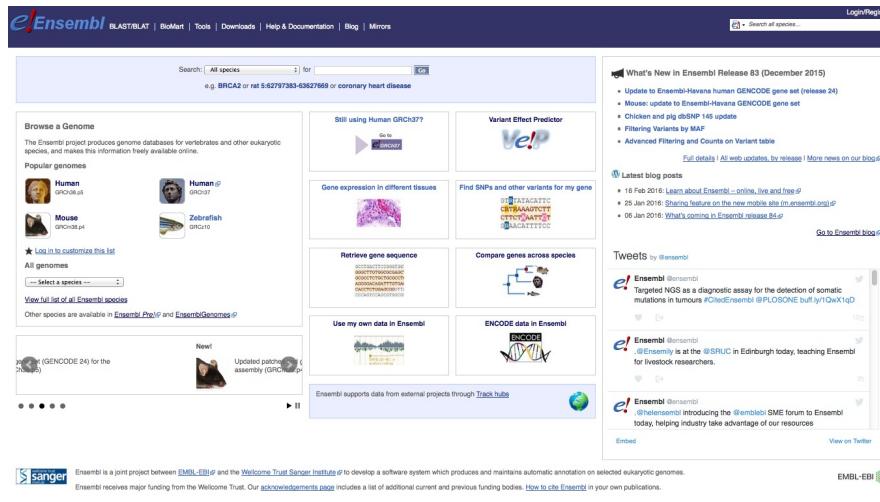
# Les “Marts”

The image displays three separate web interfaces side-by-side:

- Ensembl BioMart:** A dark blue header with the Ensembl logo and links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar at the top right says "Search all species...". Below the header, there are tabs for New, Count, and Results. A "Dataset" dropdown is set to "Homo sapiens genes (GRCh37.p13)". A "Filters" section shows "[None selected]". An "Attributes" section includes "Ensembl Gene ID". On the right, a list of output columns is shown with "Features" selected. A footer navigation bar includes Services, Research, Training, and About us.
- UniProt BioMart:** A light blue header with the UniProt logo and "bioMart". Below it is a similar interface to the Ensembl one, with tabs for New, Count, and Results, and a "Dataset" dropdown set to "[None selected]". A "Choose Database" dropdown menu is open. The footer has the same navigation bar as the Ensembl interface.
- ICGC Data Portal:** A white header with the ICGC logo and "Data Portal". Below the header are three buttons: "Cancer Projects" (orange), "Advanced Search" (blue), and "Data Repository" (teal). At the bottom is a search bar with placeholder text "eg. BRAF, KRAS G12D, DO35108, MU7870, TCGA-06-5858".

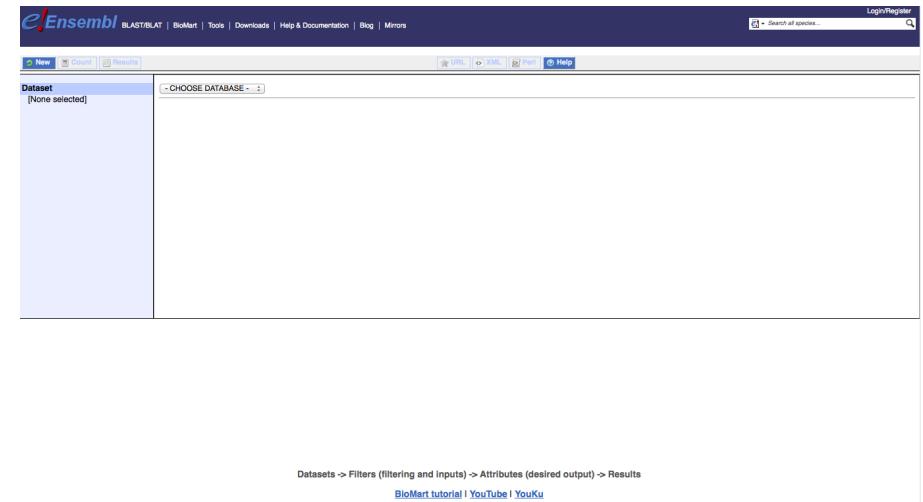
# Accéder aux données d'Ensembl

## Site web



The screenshot shows the Ensembl homepage with a search bar at the top. Below it, there's a section for browsing genomes (Human, Mouse, Zebrafish) and a "What's New" section for Ensembl Release 83 (December 2015). The page also features links to BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors.

## Outil de fouille: BioMart



The screenshot shows the BioMart interface with a search bar at the top. Below it, there's a "Dataset" dropdown menu set to "[None selected]" and a "CHOOSE DATABASE" button. The page also includes links to New, Count, and Results tabs, as well as a URL input field and a Help button.

-  Simple d'utilisation
-  Facile à comprendre
-  Une seule requête à la fois

-  Requête complexe
-  Rapide
-  Requiert une formation

# BioMart/Ensembl

The screenshot shows the Ensembl BioMart homepage. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there's a search bar labeled "Search all species..." and a "Login/Register" button. Below the navigation, there are three main sections: "Tools" (with a "All tools" link), "BioMart >" (with a link to "Export custom datasets from this site"), and "BLAST/BLAT >" (with a link to "Search our genomes for your DNA or protein sequence"). To the right of these is the "Variant Effect Predictor >" section. The central part of the page features a search bar with dropdown menus for "All species" and "for", and a "Go" button. Below the search bar is a text input field containing "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease". The main content area is divided into two columns: "All genomes" (with a dropdown menu for "Select a species" and a "Pig breeds" section) and "Favourite genomes" (listing Human, Mouse, and Zebrafish). To the right, there's a "Ensembl Rapid Release" section with a "Go" button and a "Rapid Release news" link. Below this are "Other news from our blog" links.

- Accès à :
- Annotation génomique (gènes, SNPs)
- Annotation fonctionnelle
- Expression

EMBL-EBI Ensembl created Our acknowledgements  
Ensembl release 108 - Oct 2022 © EMBL-EBI  
About Us About us Contact us

Data in Ensembl  
elixir Core Data Resource  
Link - View in archive site

# BioMart/Ensembl

The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p13)), Filters ([None selected]), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version). Below that is another Dataset section with [None Selected]. At the top right, there are buttons for New, Count, Results, URL, XML, Perl, and Help. A dropdown menu labeled "Ensembl Genes 108" is open, showing "Human genes (GRCh38.p13)" as the selected option. Orange arrows point from the text "Selection de la Base de donnée :" to the "Ensembl Genes 108" dropdown and from "Sélection du jeu de données (génome)" to the "Human genes (GRCh38.p13)" option.

- Selection de la Base de donnée :
- Genes
  - Variation
  - Regulation
  - Mouse strain

Sélection du jeu de données (génome)

In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose to have the results emailed to you.

Note that queries that run for longer than 6 hours will be terminated even when submitted this way. If this happens please reformat your query or contact us for details on how to approach this.

# BioMart/Ensembl

The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with 'Dataset' set to 'Human genes (GRCh38.p13)', 'Filters' set to '[None selected]', and 'Attributes' expanded to show options like 'Gene stable ID', 'Gene stable ID version', etc. The main area has a dropdown for 'Ensembl Genes 108' and another for 'Human genes (GRCh38.p13)'. At the top right are links for 'Login/Register', a search bar, and download options ('URL', 'XML', 'Perl', 'Help'). A large orange box highlights the 'Attributes' section in the sidebar, and an orange arrow points from this box to a bulleted list of four arguments.

- 4 arguments :
- Attributes (entêtes des colonnes dans les résultats)
- Filters (Utilisé pour restreindre les résultats)
- Values (identifiants utilisés pour filtrer)
- Mart (selection des jeux de données)

Note that queries that run for longer than 6 hours will be terminated even when submitted this way. If this happens please reformat your query or contact us for details on how to approach this.

# Biomart : Partie pratique

# Comparaison des browsers

- Différences majeures entre Ensembl vs UCSC/NCBI
  - NCBI vs ensembl (UCSC?) – à l'origine de l'assemblage
  - Utilisation d'un pipeline automatique pour la création des jeux de données
  - Utilisation:
    - Visuel: ensembl/UCSC vs NCBI
    - Web: ensembl vs UCSC/NCBI
    - Rapidité/confort: UCSC vs ensembl/NBI
    - Organisation: ensembl/UCSC? Vs NCBI