

# 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



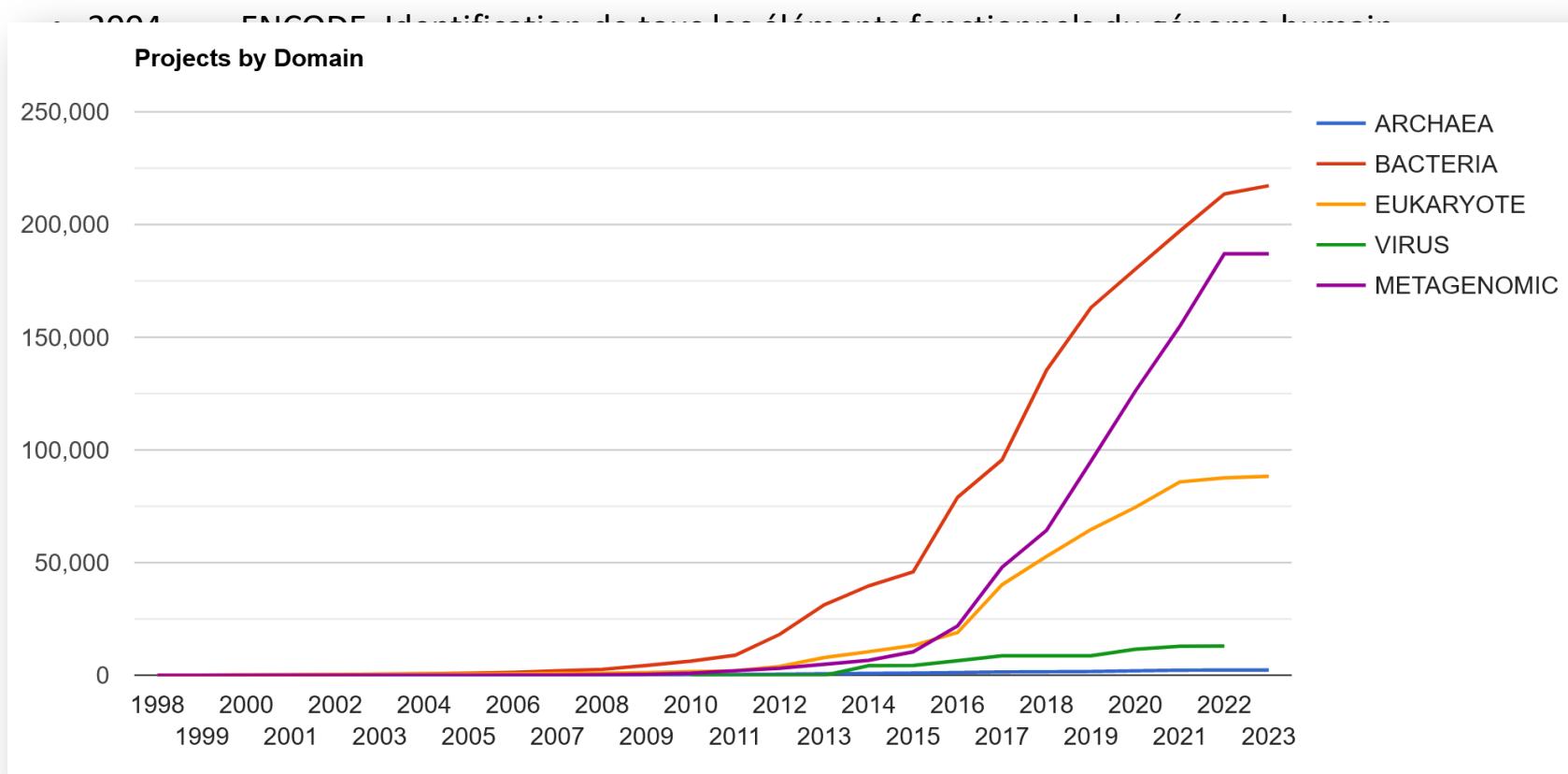
# L'ère post-génomique

- 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)

The screenshot shows the NCBI homepage. At the top left is the NCBI logo (a stylized 'S' icon followed by the letters 'NCBI'). To its right is the text 'National Center for Biotechnology Information'. Below this, it says 'National Library of Medicine' and 'National Institutes of Health'. A horizontal menu bar follows, containing links for 'PubMed', 'All Databases', 'BLAST', 'OMIM', 'Books', 'TaxBrowser', and 'Structure'. Below the menu is a search bar with the placeholder 'Search All Databases' and a dropdown arrow, followed by a text input field and a 'Go' button.

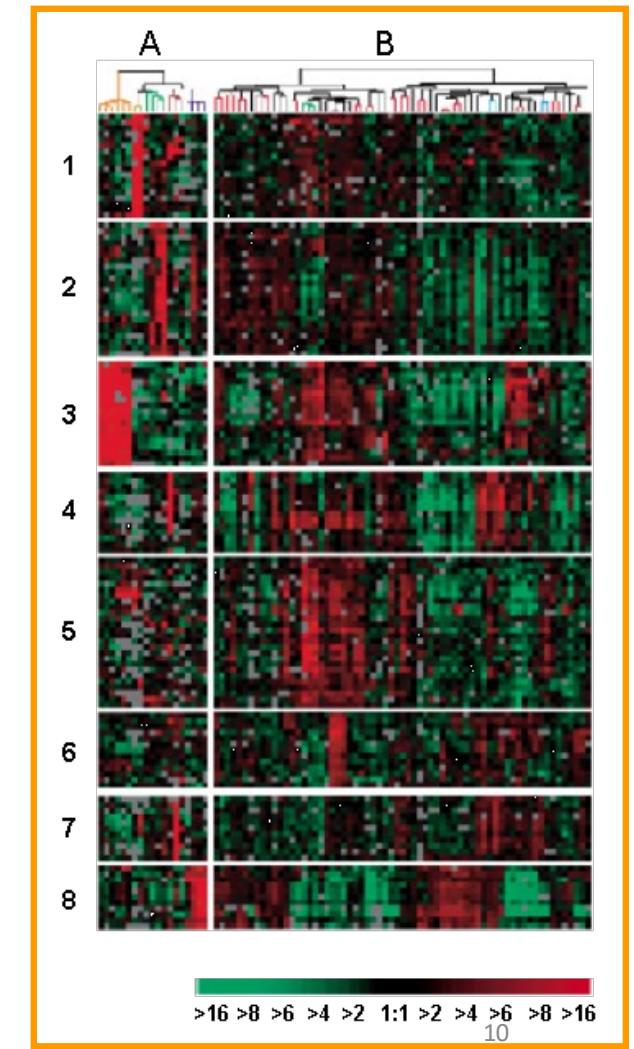
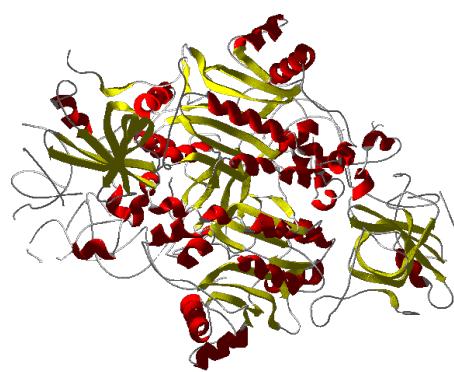
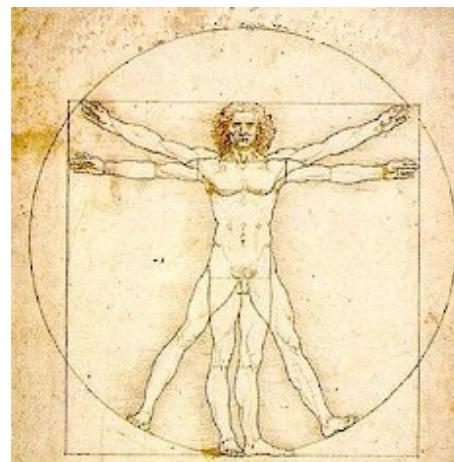
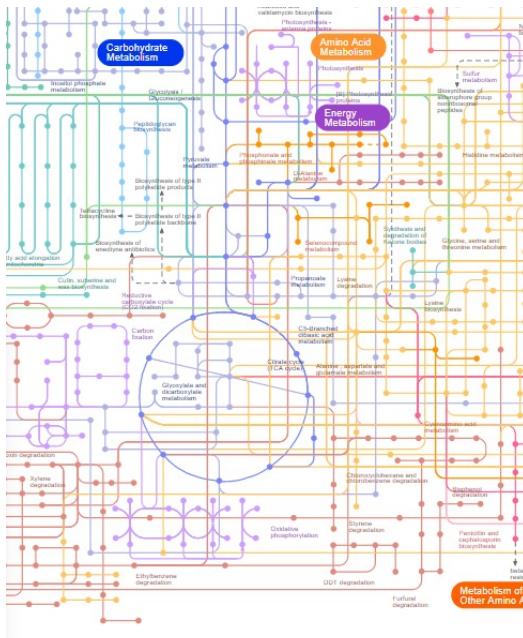
<http://www.ncbi.nlm.nih.gov/>

# 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

**MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFYTPKARREVEGPQVGALELAGGPGAA**



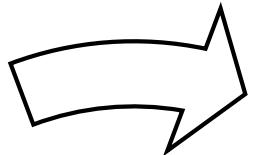
# Banques de séquences nucléiques généralistes

- 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

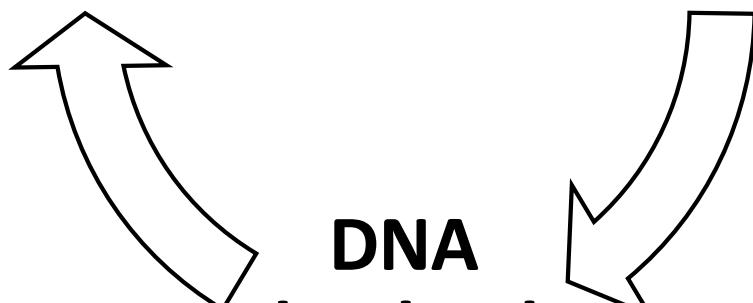
# Banques de séquences nucléiques généralistes



# 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

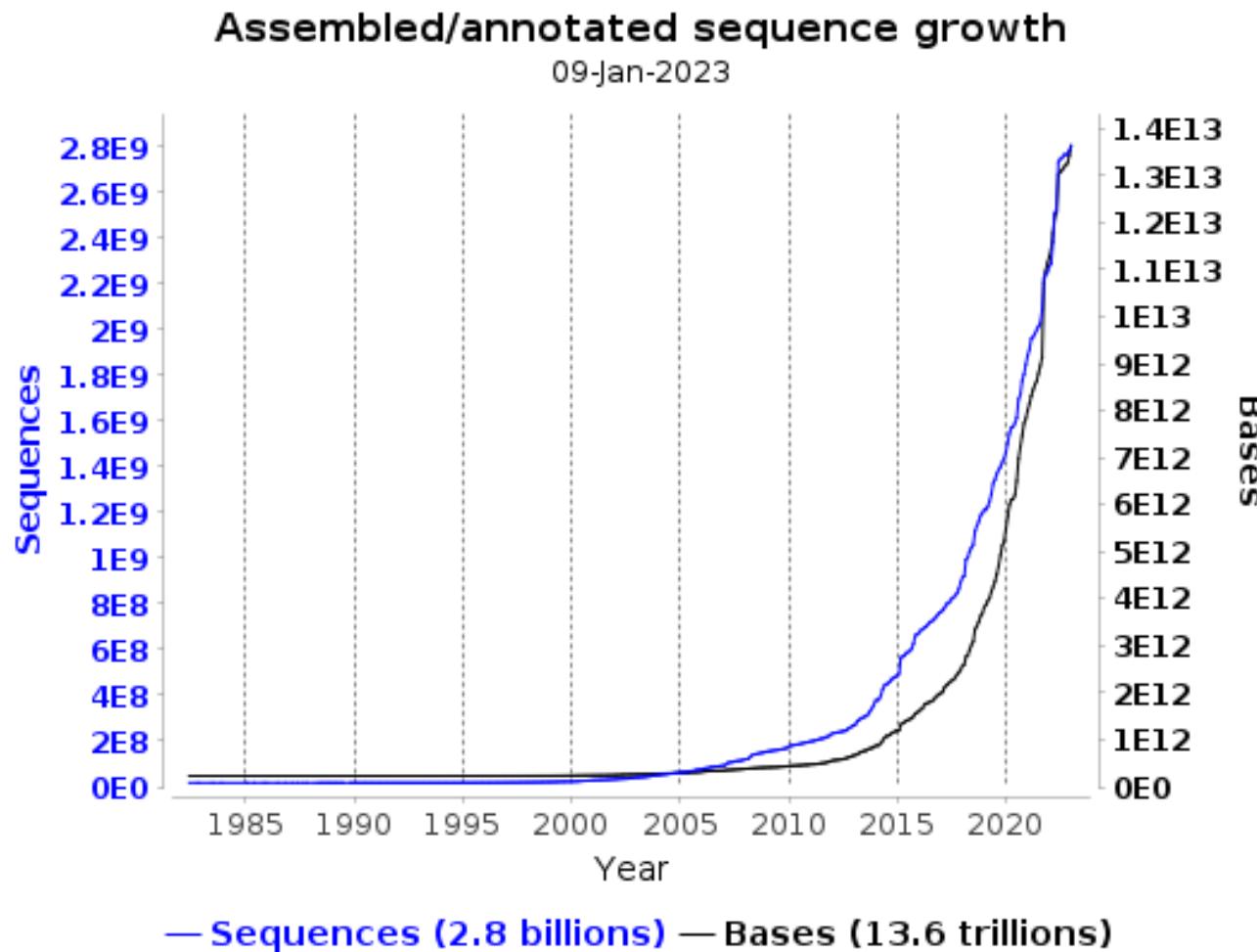


**NUCLEOTIDE  
SEQUENCES  
1986/1987**

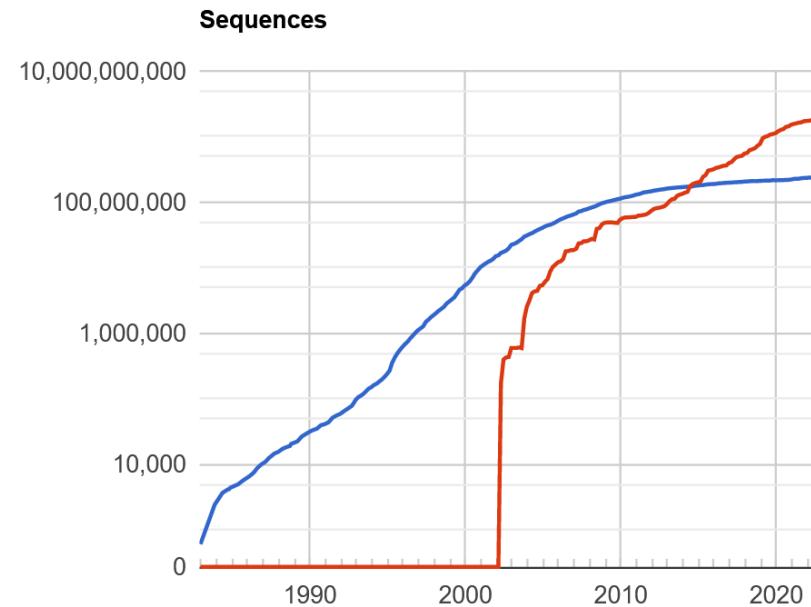
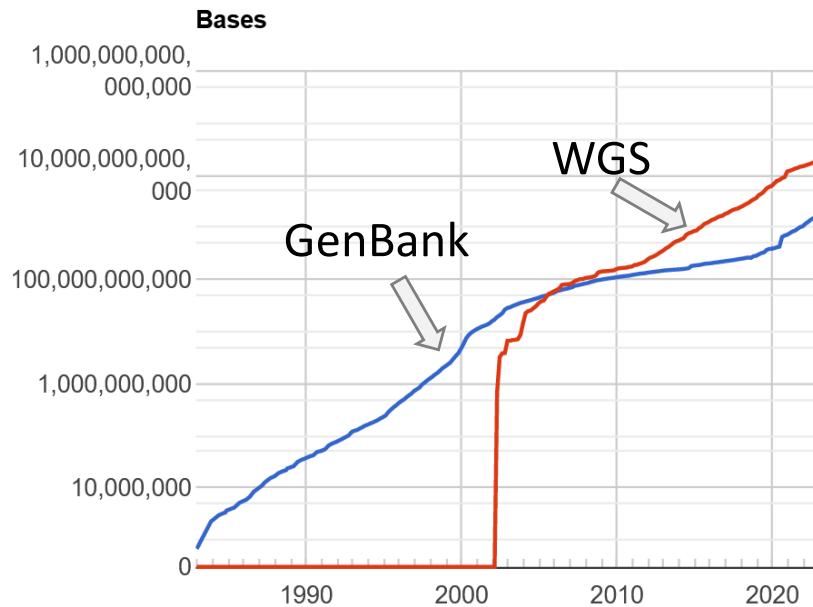
---

**VOLUME I**  
**PRIMATES**

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

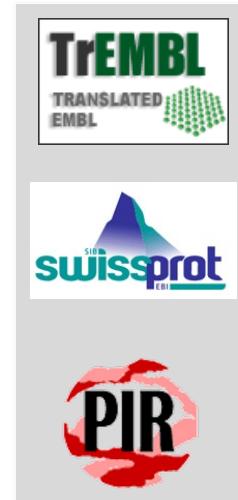
Transcrits: **29,869,155**

Organismes: **99,842**



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

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



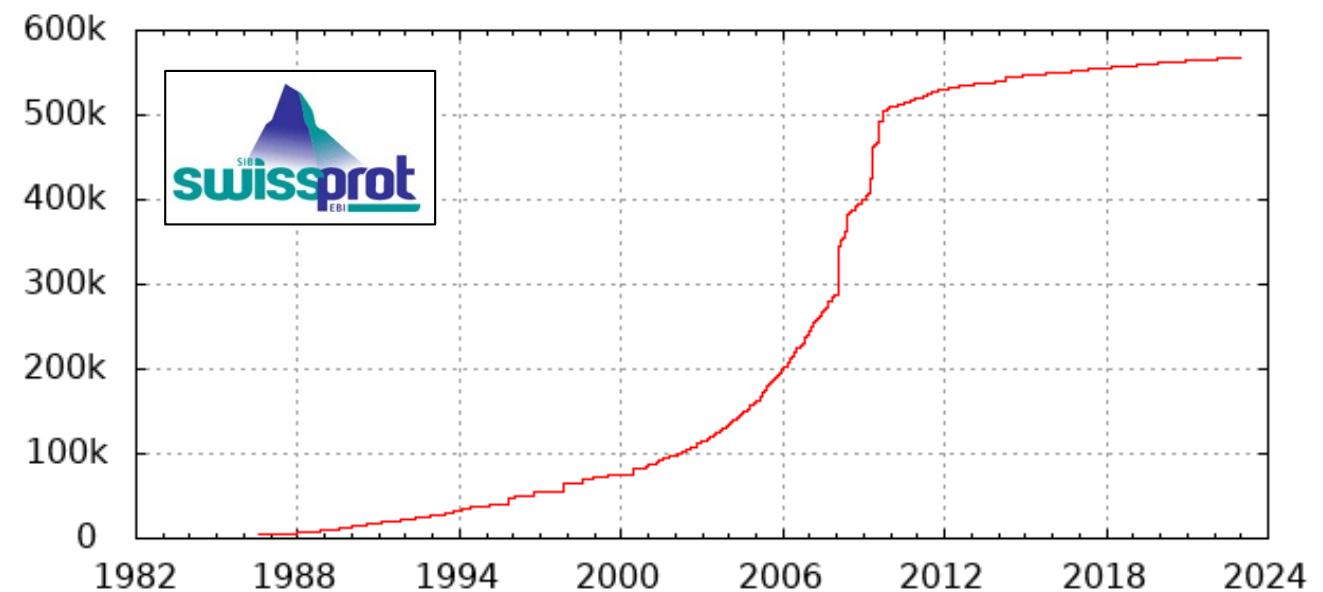
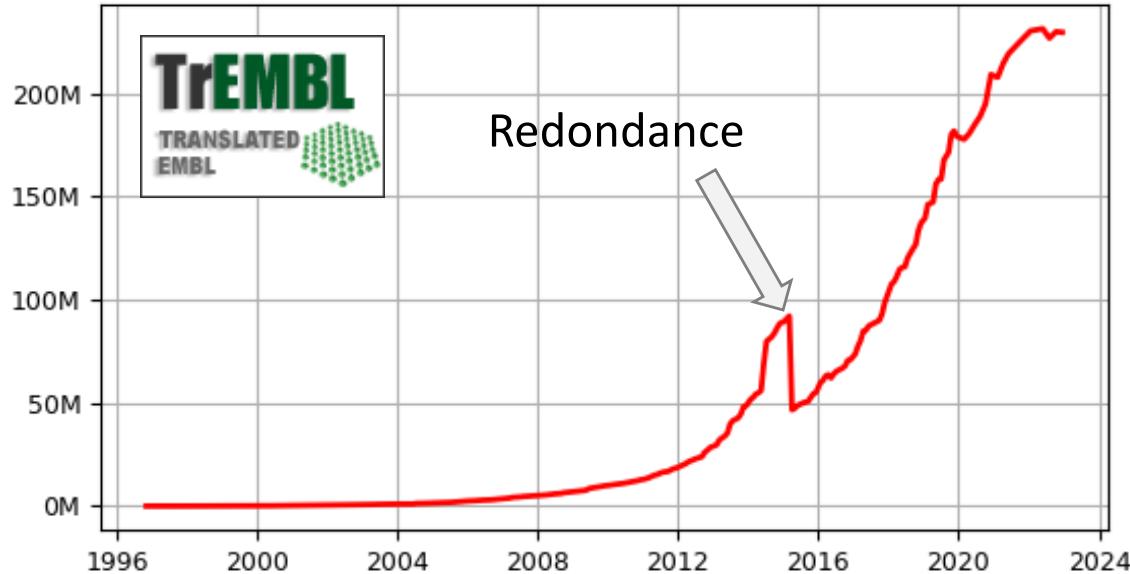
TrEMBL:  
**179,250,561** entrées

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

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

Entry Feature viewer Publications External links History

BLAST Download Add Add a publication Entry feedback

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 :

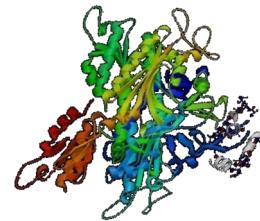
- regrouper les informations d'un même type
- faciliter l'accès à l'information

## 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

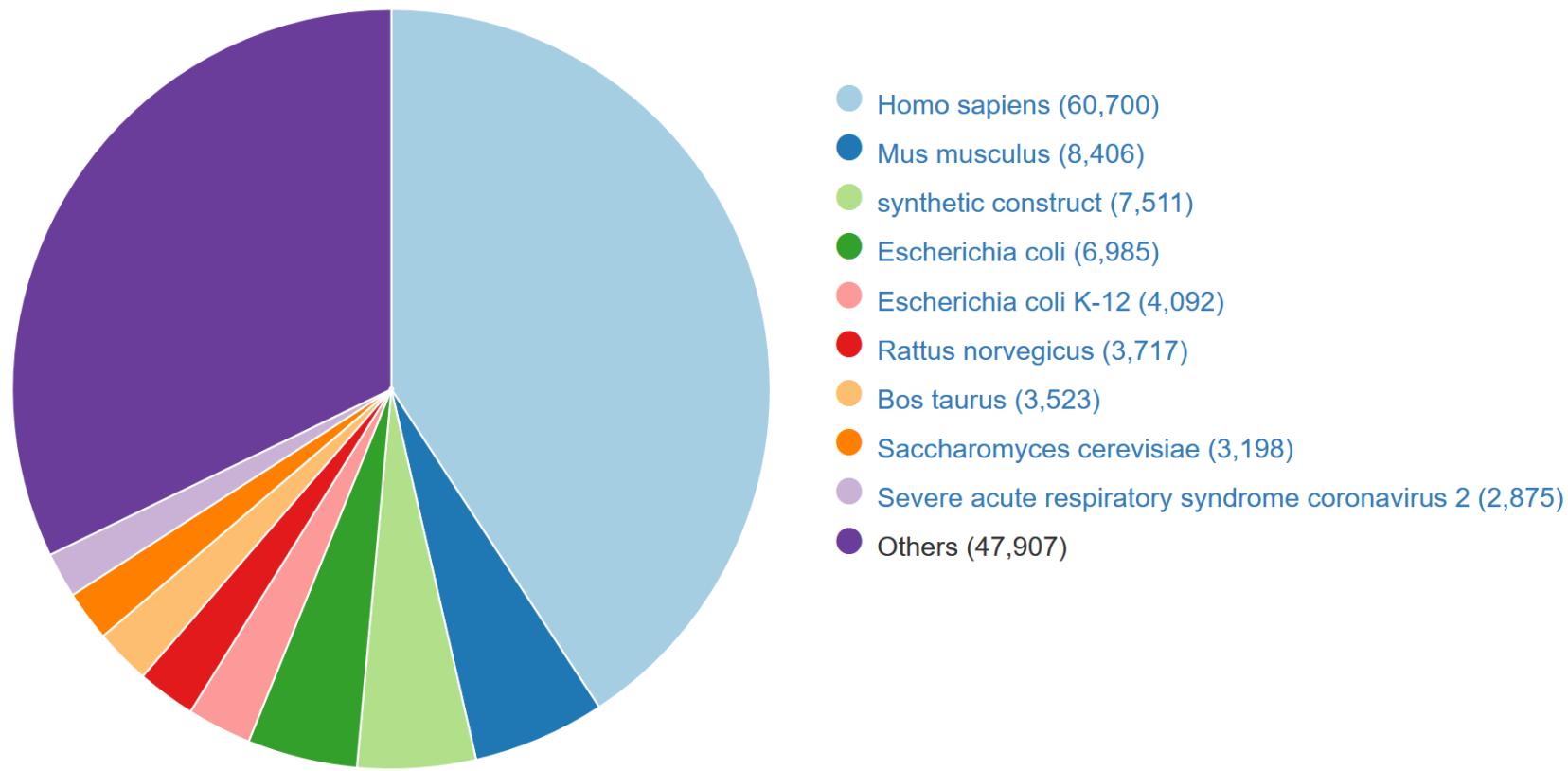
January Molecule of the Month

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** »

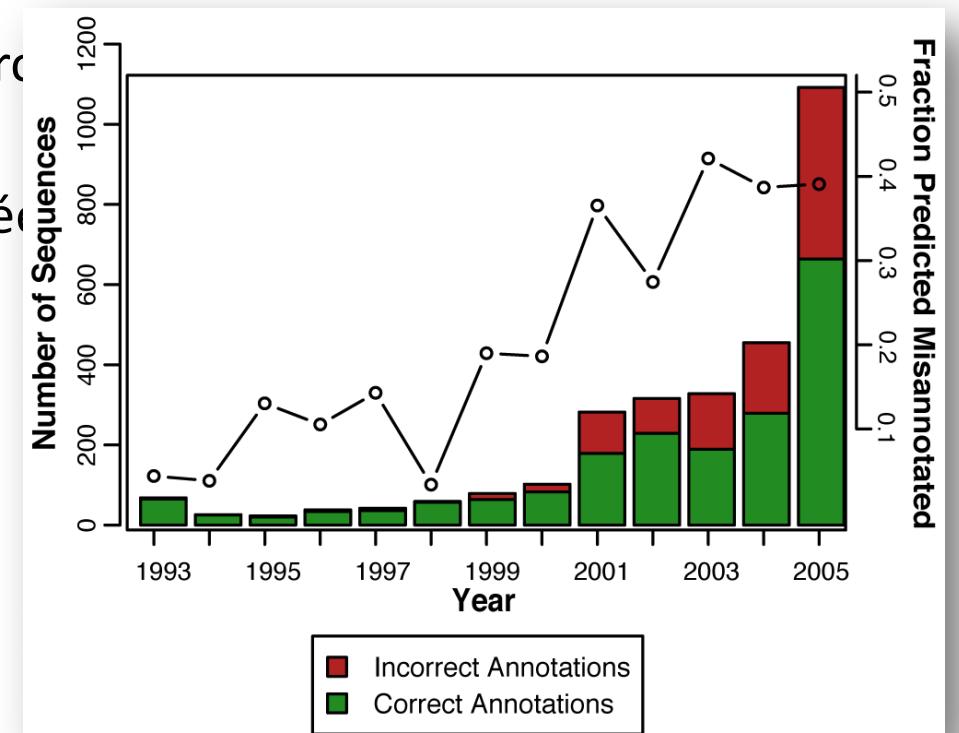
# Exemple de l'importance de l'annotation

**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**

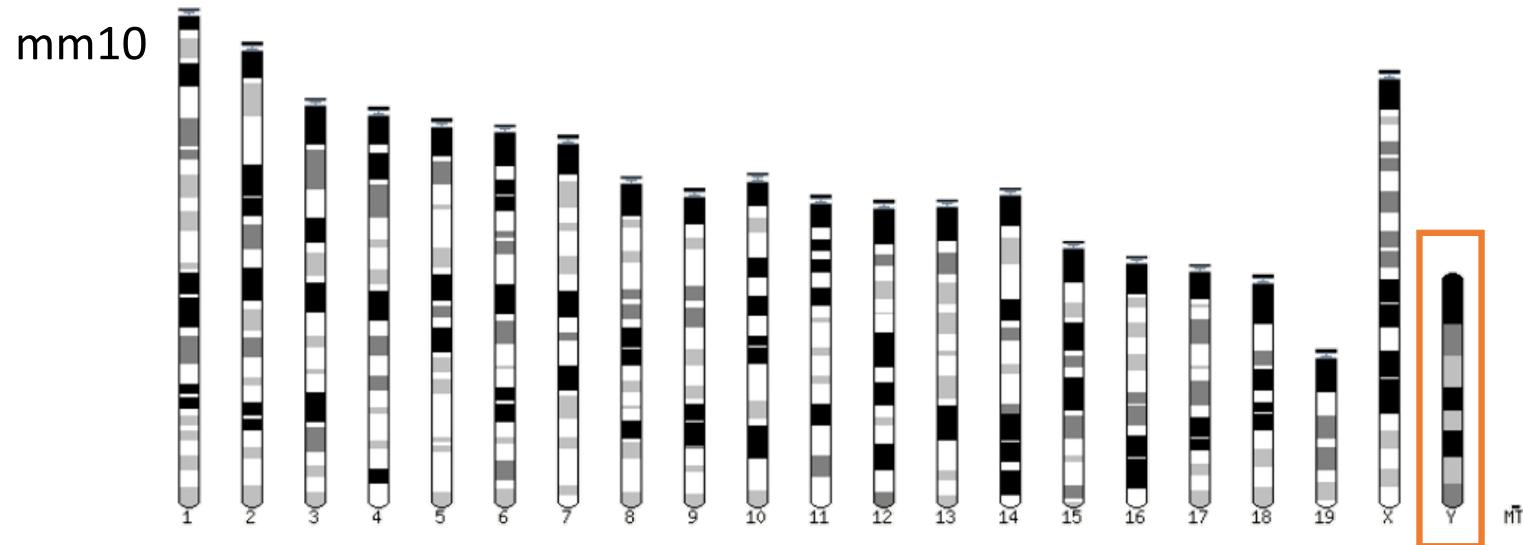
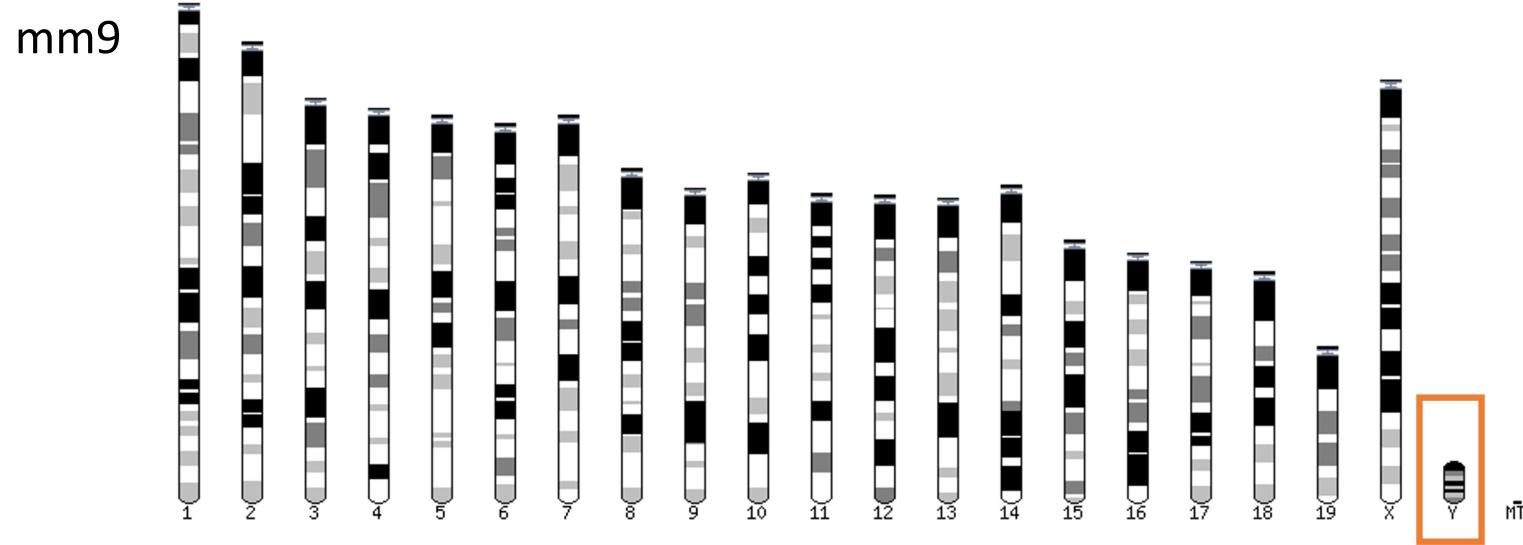
Bases de données majeures  
collecte des données individuelles et collectives

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

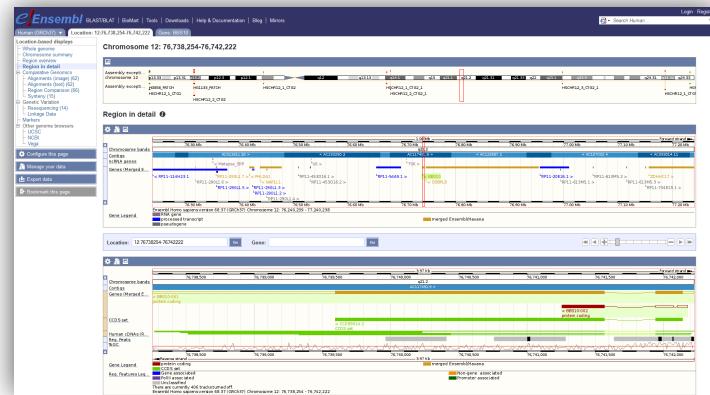
SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
<b>MAMMALS</b>				
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI 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

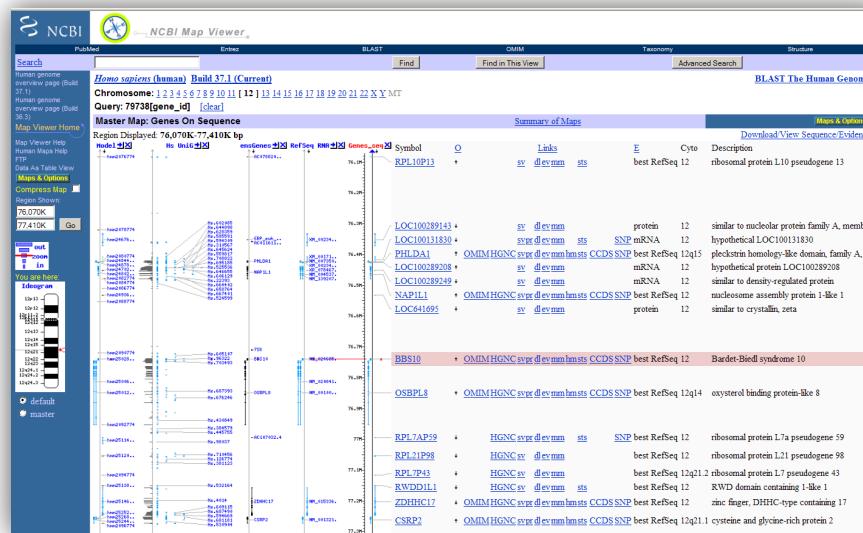
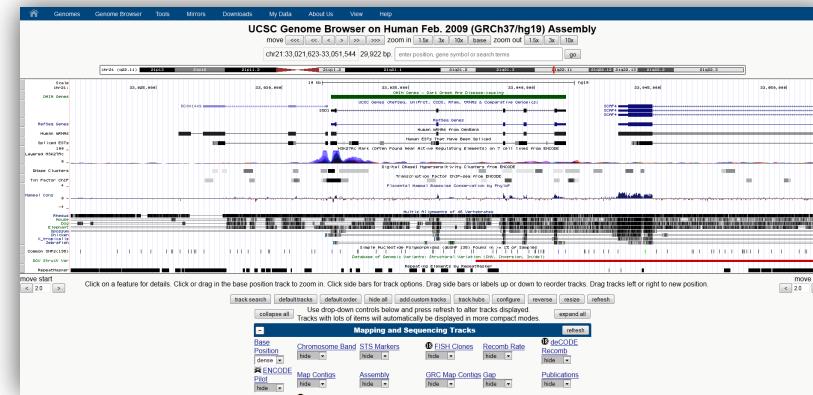
- Elément de référence absolue le **génome**
- Agrégateur et générateur d'informations/annotations
  - Prédictions de gènes
  - Protéines
  - Données d'expression
  - Variations
- Synthèse rapide et visuelle de données primordiales

# Il y a Genome Browsers...

EBI - Ensembl

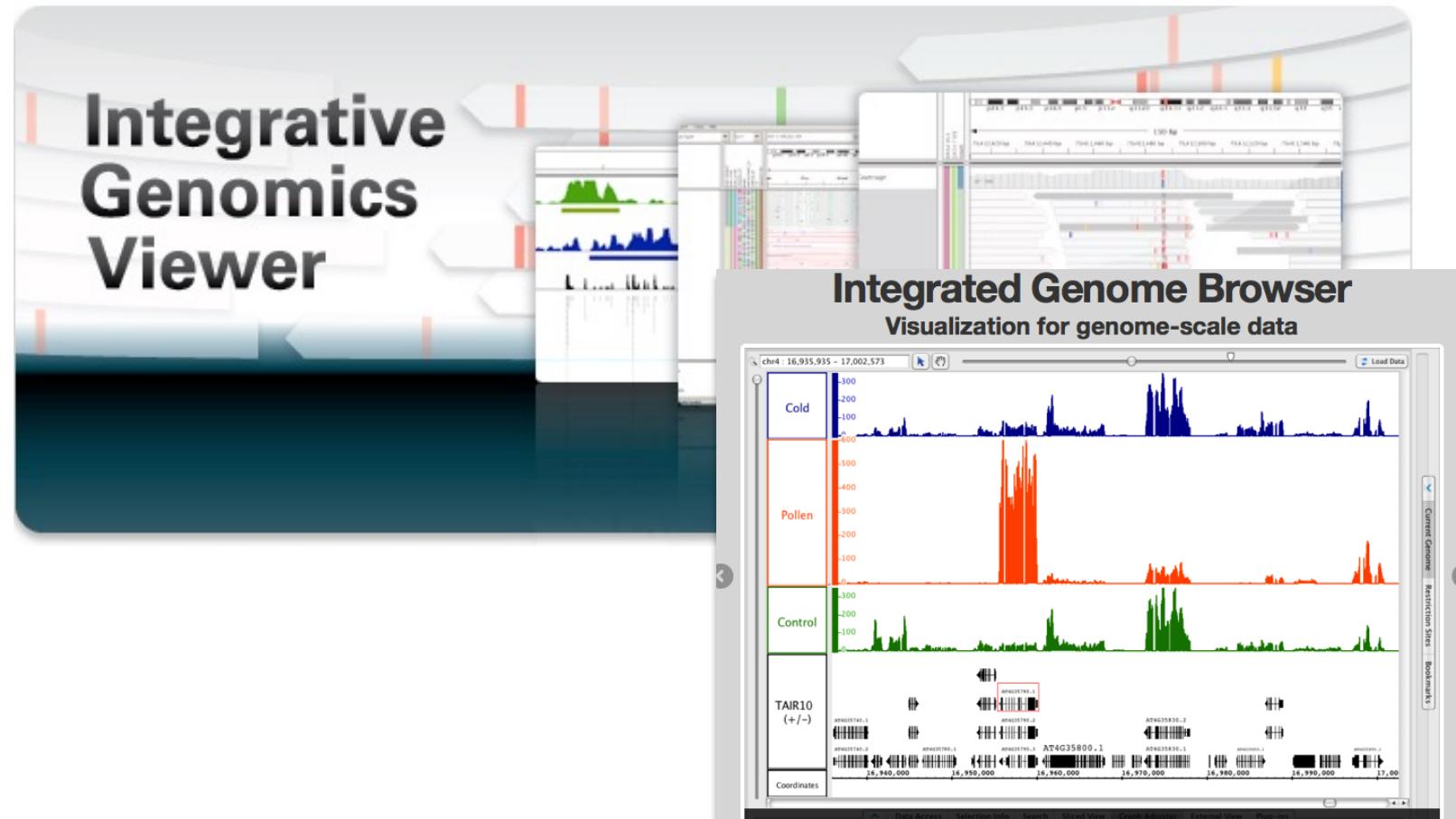


UCSC – Genome Browser



NCBI – Map Viewer

# Et Genome browsers



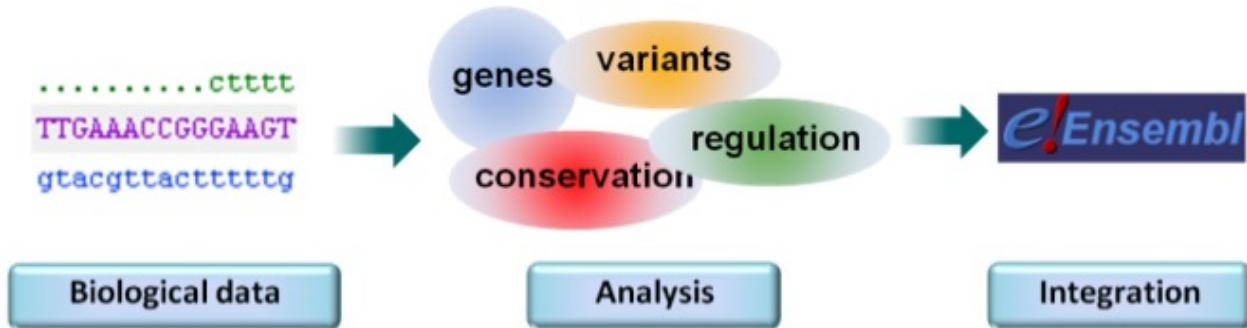
# Ensembl

# 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 [Login/Register](#)   

**Tools** [BioMart >](#) [BLAST/BLAT >](#) [Variant Effect Predictor >](#)

[All tools](#) Export custom datasets from Ensembl with this data-mining tool Search our genomes for your DNA or protein sequence 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 20 additional breeds **Favourite genomes** Human GRCh38.p14 Still using GRCh37? Mouse GRCm39 Zebrafish GRCz11

[View full list of all species](#)

**Ensembl Release 113 (October 2024)**

- Integration of lncRNA transcripts from the Capture Long-read Sequencing (CLS) project
- Additional breeds available for *Capra hircus* (Goat), *Ovis aries* (Sheep), and *Sus scrofa* (Pig)
- Ensembl VEP now supports the GENCODE Primary transcript set
- Regulatory annotation updates for *Homo sapiens* (Human) and *Mus musculus* (Mouse)

[More release news](#) on our blog

**Ensembl Rapid Release**

New genome assemblies are now being released to the [Ensembl Beta site](#). All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site. The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

[Find out more on our blog](#)

**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.

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

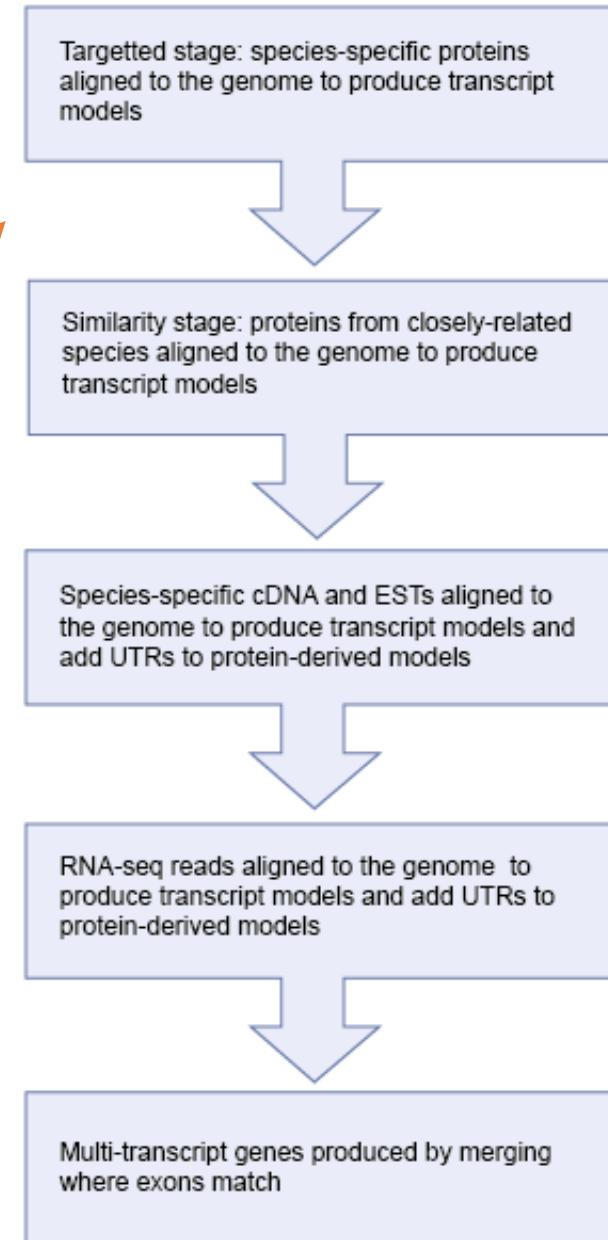
# Comprendre ENSEMBL

# 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

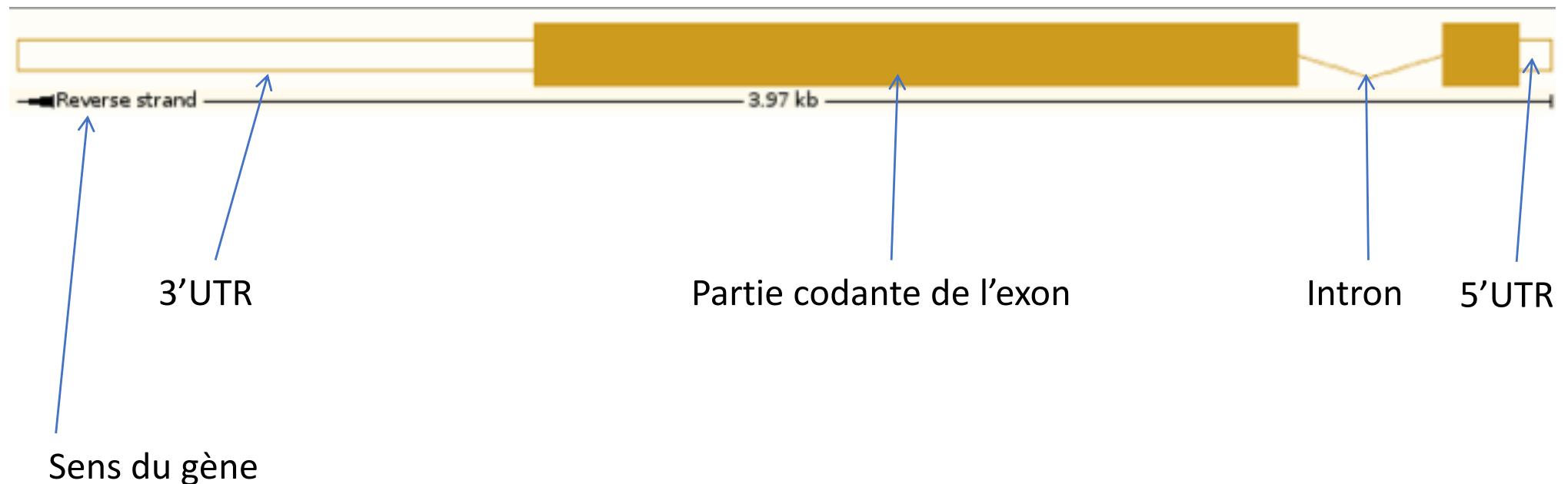
- Les annotations automatiques des gènes codants sont basés sur :
  - European Nucleotide Archive (ENA)
  - UniProtKB (curation manuelle)
  - NCBI refseq (curation manuelle)
- Les annotations automatiques des gènes non-codants sont basés sur :
  - RFAM (base de données de familles d'ARNs)
  - miRbase
  - tRNAscan-SE (algorithme de prediction des tRNA)



# 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, NCBI, HGNC, MGI)
  - <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

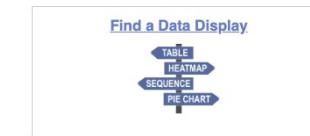
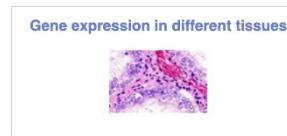
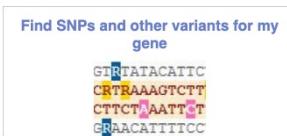
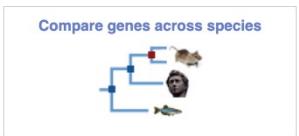


# Identifiants Ensembl

- ENSG### 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: ENSMUSG###
  - DAR (*Danio rerio*) pour le zebrafish: ENSDARG###
  - etc.

# Version (Release)

- ~ tous les 3 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 | Help & Documentation | Using this website | Archives

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 the 'Table of Assemblies' page from the Ensembl Archives. The page has a dark blue header with the Ensembl logo and navigation links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar 'Search all species...' and a 'Login/Register' link are also present. The main content area has a light gray background and displays a grid of species names in the left column and assembly version numbers in the right column. Each cell in the grid contains a small colored square indicating the status of the assembly: yellow for new species, white for species present in the archive, and gray for species not in this version of Ensembl. The grid spans from October 2022 at the top to September 2015 at the bottom.

	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	ASM34442v1																											
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

**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 113 (October 2024)**

- Integration of lncRNA transcripts from the Capture Long-read Sequencing (CLS) project
- Additional breeds available for *Capra hircus* (Goat), *Ovis aries* (Sheep), and *Sus scrofa* (Pig)
- Ensembl VEP now supports the GENCODE Primary transcript set
- Regulatory annotation updates for *Homo sapiens* (Human) and *Mus musculus* (Mouse)

[More release news](#) on our blog

**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 20 additional breeds

[View full list of all species](#)

**Favourite genomes**

 Human GRCh38.p14  
[Still using GRCh37?](#)

 Mouse GRCm39

 Zebrafish GRCz11

**Ensembl Rapid Release**

New genome assemblies are now being released to the [Ensembl Beta site](#). All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site. The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

Find out more on our [blog](#)

**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.

GLOBAL CORE BIO DATA RESOURCE 

# Ensembl Genomes

## Bactéries

**EnsemblBacteria** | HMMER | BLAST | Tools | Downloads | More | Search Ensembl Bacteria...

**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 EMBL-EBI's BioProject database 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...**

Ensembl Genomes is developed by EMBL-EBI and is powered by the Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please [click here](#).

**EMBL-EBI** | **Powered by**

## Fungi

**EnsemblFungi** | HMMER | BLAST | BioMart | Tools | Downloads | More | Search Ensembl Fungi...

**Search: All species**

**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.

[Rapid Release news](#) on our blog

**Archive sites**

## Plantes

**EnsemblPlants** | HMMER | BLAST | BioMart | Tools | Downloads | More | Search Ensembl Plants...

**Search: All species**

**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 [expression.com](#) and [Krauthein](#)
- Alignments from the 10+ genome project
- Alignment of 98,270 high confidence genes from the TIGCv1 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 TGAo 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** | HMMER | BLAST | BioMart | Tools | Downloads | More | Search Ensembl Protists...

**Search: All species**

**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](#)

**Archive sites**

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

## Métazoaires

**EnsemblMetazoa** | HMMER | BLAST | BioMart | Tools | More | Search Ensembl Metazoa...

**Search: All species**

**What's New in Release 52**

- Updated data
  - Updated species
    - Cimex lectularius (Herrler)
  - 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](#)

**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

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

Recherche

Tools

BioMart >

BLAST/BLAT >

Variant Effect Predictor >

[All tools](#)

Export custom datasets from Ensembl with this data-mining tool

Search our genomes for your DNA or protein sequence

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 113 (October 2024)

- Integration of lncRNA transcripts from the Capture Long-read Sequencing (CLS) project
- Additional breeds available for *Capra hircus* (Goat), *Ovis aries* (Sheep), and *Sus scrofa* (Pig)
- Ensembl VEP now supports the GENCODE Primary transcript
- Regulatory annotation updates for *Homo sapiens* (Human) and *Mus musculus* (Mouse)

[More release news](#) on our blog

Recherche

Search

All species for  Go

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

News

All genomes

-- Select a species --

Pig breeds  
Pig reference genome and 20 additional breeds

[View full list of species](#)

Favourite genomes

Human  
GRCh38.p14

Mouse  
GRCm39

Zebrafish  
GRCz11

Ensembl Rapid Release

New genome assemblies are now being released to the [Ensembl Beta site](#).

All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site.

The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

[Find out more on our blog](#)

Liste déroulante Accès aux génomes

species

Find SNPs and other variants for my gene

Gene expression in different tissues

Retrieve gene sequence

Find a Data Display

Use your own data in Ensembl

Accès aux archives d'Ensembl

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

Ensembl release 113 - October 2024 © EMBL-EBI

53

# Le site web Ensembl: les génomes

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

Human (GRCh38.p14) ▾

Search Human (Homo sapiens)

Search all categories ▾ Search...

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p14 (GCA\_000001405.29)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014

Comparative genome

What can I find? Homologous genes, orthologous genes, phylogenetic trees, and whole genome alignments across multiple species.

Informations, statistiques

More about regulatory features like enhancers and promoters, and regulatory activity including ATAC-seq and ChIP-seq tracks.

- More about the Ensembl regulatory annotation
- Experimental data sources
- Download all regulatory features (GFF)

Recherche

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein transcripts, and other genomic features.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Lien vers des exemples

Pax6 INS FGF2 BRCA2 DMD ssh

Example gene

Example transcript

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

VeP

ATCGAGCT ATCCAGCT ATCGAGAT

Example variant

Example phenotype

Example structural variant

# Le site web Ensembl: statistiques des génomes

**Informations générales sur l'assemblage**

**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).

**Patches**

As the genome reference assembly is updated, it improves the assembly, patches are being introduced. Currently, assembly patches are of two types: sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC in the region of the reference assembly at the next major assembly

**Statistics**

**Summary**

Assembly	GRCh38.p14 (Genome Reference Consortium Human Build 38) GCA_000001405.29, Dec 2013
Base Pairs	3,099,750,718
Golden Path Length	3,099,750,718
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 2024
Database version	113.38
Gencode version	GENCODE 47

**Gene counts (Primary assembly)**

Coding genes	19,868 (excl 659 readthrough)
Non coding genes	42,160
Small non coding genes	4,867
Long non coding genes	35,076 (excl 300 readthrough)
Misc non coding genes	2,217
Pseudogenes	15,206 (excl 1 readthrough)
Gene transcripts	387,944

**Gene counts (Alternative sequence)**

Coding genes	3,303 (excl 34 readthrough)
Non coding genes	2,008
Small non coding genes	349
Long non coding genes	1,457 (excl 33 readthrough)
Misc non coding genes	202
Pseudogenes	2,060 (excl 1 readthrough)
Gene transcripts	24,090

**Other**

Genscan gene predictions	50,174
Short Variants	1,111,915,564
Structural variants	7,862,163

# Le site web Ensembl: caryotype

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

Human (GRCh38.p14) ▾

Login/Register

Search all species...

Genome

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail

Comparative Genomics

- Synteny
- Alignments (image)
- Alignments (text)
- Region Comparison

Genetic Variation

- Variant table
- Resequencing
- Strain table
- Linkage Data
- Markers

Other genome browsers

- UCSC
- NCBI
- Ensembl GRCh37

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

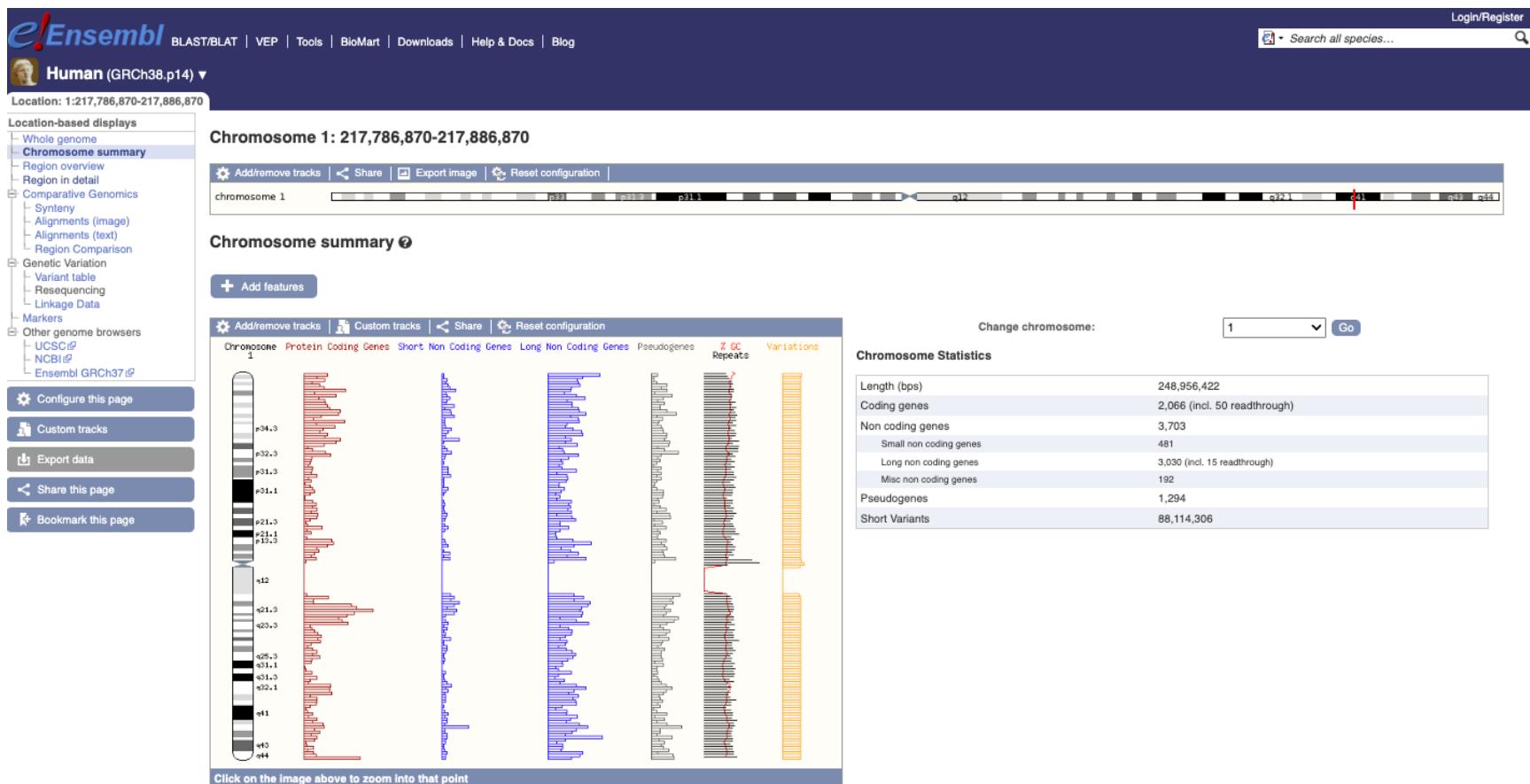
Assembly	GRCh38.p14 (Genome Reference Consortium Human Build 38), INSDC Assembly <a href="#">GCA_000001405.29</a> , Dec 2013
Base Pairs	3,099,750,718
Golden Path Length	3,099,750,718
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jul 2024
Database version	113.38
Gencode version	GENCODE 47

Gene counts (Primary assembly)

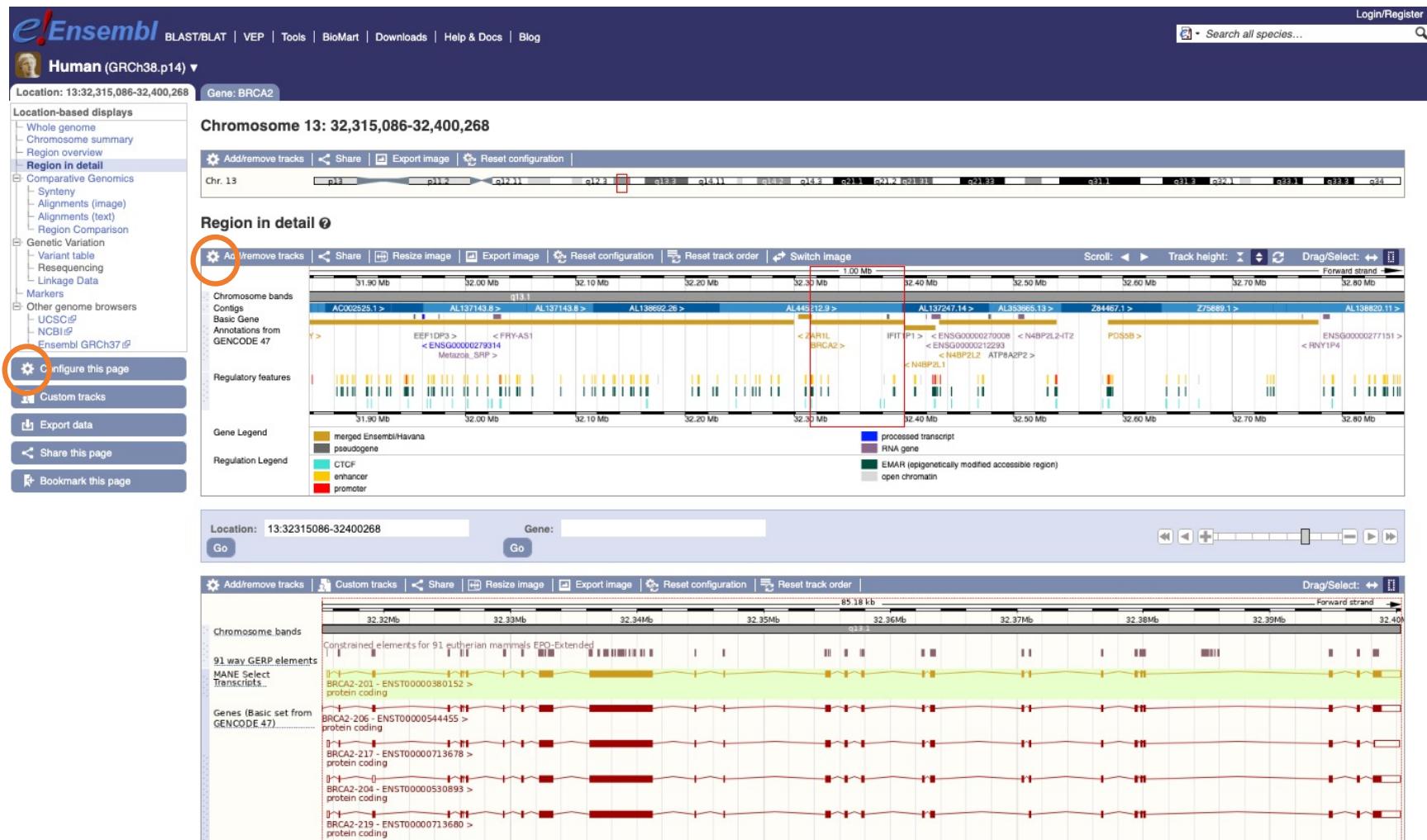
Coding genes	19,868 (excl 659 readthrough)
Non coding genes	42,160
Small non coding genes	4,867
Long non coding genes	35,076 (excl 300 readthrough)

<https://www.ensembl.org>

# Le site web Ensembl : statistiques par chromosome



# Le site web Ensembl : navigateur de génome



# Le site web Ensembl : le gène

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

Human (GRCh38.p14) ▾

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: Anatomical entity
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
  - Gene expression
  - Pathway
  - Molecular interactions
  - Regulation
  - External references
  - Supporting evidence
- ID History
  - Gene history

**Gene: BRCA2 ENSG00000139618**

Description: BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

Gene Synonyms: BRCC2, FACD, FAD1, FANCD, FANC1, XRCC11

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

About this gene: This gene has 19 transcripts (splice variants), 173 orthologues and is associated with 194 phenotypes.

Transcripts: Hide transcript table

Transcript ID	Name	bp	Protein	Translation ID	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000380152.8	BRCA2-201	11954	3418aa	ENSP00000369497.3	Protein coding	CCDS9344	P51587	NM_000059.4	MANE Select Ensembl Canonical GENCODE Primary GENCODE Basic APPRIS ALT2 TSL:1
ENST00000530893.7	BRCA2-204	11953	3295aa	ENSP00000499438.2	Protein coding		A0A590UJ17	-	GENCODE Primary GENCODE Basic APPRIS P4 TSL:1
ENST00000713678.1	BRCA2-217	11900	3232aa	ENSP00000518981.1	Protein coding		-	-	GENCODE Primary GENCODE Basic APPRIS ALT2
ENST00000680887.1	BRCA2-210	11880	3418aa	ENSP00000505050.1	Protein coding	CCDS9344	A0A7P0T9D7	P51587	APPRIS ALT2
ENST00000544455.6	BRCA2-206	11854	3418aa	ENSP00000439902.1	Protein coding	CCDS9344	P51587	-	GENCODE Basic APPRIS ALT2 TSL:1
ENST00000713680.1	BRCA2-219	11798	3366aa	ENSP00000518983.1	Protein coding		-	-	GENCODE Basic
ENST00000700202.2	BRCA2-214	10553	3401aa	ENSP00000514856.2	Protein coding		A0ABV8TPZ2	-	GENCODE Primary GENCODE Basic APPRIS ALT2
ENST00000470094.2	BRCA2-202	12077	3200aa	ENSP00000434898.2	Nonsense mediated decay		H0YE37	-	TSL:5
ENST00000713677.1	BRCA2-216	11958	118aa	ENSP00000518980.1	Nonsense mediated decay		-	-	-
ENST00000614259.2	BRCA2-207	11763	2649aa	ENSP00000506251.1	Nonsense mediated decay		A0A7P0TAP7	-	TSL:2
ENST00000713679.1	BRCA2-218	11428	2644aa	ENSP00000518982.1	Nonsense mediated decay		-	-	-
ENST00000665585.2	BRCA2-208	10917	2916aa	ENSP00000499570.2	Nonsense mediated decay		A0A590UJU6	-	-
ENST00000528762.2	BRCA2-203	10668	2898aa	ENSP00000433168.2	Nonsense mediated decay		H0YD86	-	TSL:4
ENST00000666593.2	BRCA2-209	9839	3097aa	ENSP00000449926.2	Nonsense mediated decay		A0A590UJ24	-	-
ENST00000700201.1	BRCA2-213	2103	129aa	ENSP00000514855.1	Nonsense mediated decay		A0ABV8TQQ4	-	-
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 External references table

CCDS: This gene is a member of the Human CCDS set: CCDS9344.1

# Le site web Ensembl : le transcript

Ensembl Human (GRCh38.p14) ▾

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

**Configure this page**

**Custom tracks**

**Export data**

**Share this page**

**Bookmark this page**

**Transcript: ENST00000380152.8 BRCA2-201**

Description: BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC-1101]

Gene Synonyms: BRCC2, FACD, FAD, FAD1, FANCN, FANCD1, XRCC11

Location: Chromosome 13: 32,315,086-32,400,268 forward strand.

About this transcript: This transcript has 27 exons, is annotated with 98 domains and features, is associated with 65674 variant alleles and maps to 930 oligo probes.

Gene: This transcript is a product of gene ENSG00000139618.19 Hide transcript table

Show/hide columns Filter

Transcript ID	Name	bp	Protein	Translation ID	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000380152.8	BRCA2-201	11954	3418aa	ENSP00000369497.3	Protein coding	CCDS9344.0	P51587.0	NM_000059.4.0	MANE Select Ensembl Canonical GENCODE Primary GENCODE Basic APPRIS ALT2 TSL-5
ENST00000530893.7	BRCA2-204	11953	3295aa	ENSP00000499438.2	Protein coding	-	A0A590UJ17.0	-	GENCODE Primary GENCODE Basic APPRIS P4 TSL-1
ENST00000713671.8	BRCA2-217	11903	3232aa	ENSP00000518861.1	Protein coding	-	-	-	GENCODE Primary GENCODE Basic APPRIS ALT2
ENST000006680887.1	BRCA2-210	11880	3418aa	ENSP00000595508.2	Protein coding	CCDS9344.0	A0A7P0T9D7.0	P51587.0	APPRIS ALT2
ENST00000544456.6	BRCA2-206	11854	3418aa	ENSP00000439902.1	Protein coding	CCDS9344.0	P51587.0	-	GENCODE Basic APPRIS ALT2 TSL-1
ENST00000713680.1	BRCA2-219	11798	3366aa	ENSP00000518983.1	Protein coding	-	-	-	GENCODE Basic
ENST00000700202.2	BRCA2-214	10553	3401aa	ENSP00000514856.2	Protein coding	-	A0ABV8TP22.0	-	GENCODE Primary GENCODE Basic APPRIS ALT2
ENST00000470094.2	BRCA2-202	12077	3200aa	ENSP00000434898.2	Nonsense mediated decay	-	H0YE37.0	-	TSL-5
ENST00000713671.7	BRCA2-216	11954	118aa	ENSP00000518860.1	Nonsense mediated decay	-	-	-	-
ENST00000614259.2	BRCA2-207	11763	2649aa	ENSP00000506251.1	Nonsense mediated decay	-	A0A7P0TAP7.0	-	TSL-2
ENST00000713671.9	BRCA2-218	11426	2644aa	ENSP00000518862.1	Nonsense mediated decay	-	-	-	-
ENST00000665585.2	BRCA2-208	10917	2916aa	ENSP00000499570.2	Nonsense mediated decay	-	A0A590UUJ6.0	-	-
ENST00000526762.2	BRCA2-203	10668	2898aa	ENSP00000431168.2	Nonsense mediated decay	-	H0YD86.0	-	TSL-4
ENST00000666593.2	BRCA2-209	9839	3097aa	ENSP00000499256.2	Nonsense mediated decay	-	A0A590UJ24.0	-	-
ENST00000700201.1	BRCA2-213	2103	129aa	ENSP00000514855.1	Nonsense mediated decay	-	A0ABV8TQ04.0	-	-
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

BRCA2-201 - ENST00000380152 > protein coding

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.

Annotation Attributes: CAGE supported TSS [Definitions]

GENCODE basic gene: This transcript is a member of the Gencode basic gene set.

Ensembl release 113 - October 2024 © EMBL-EBI Permanent link - View in archive site

# Naviguer dans Ensembl : Partie pratique

# Visualiser ses propres données

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

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

**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 113 (October 2024)**

- Integration of lncRNA transcripts from the Capture Long-read Sequencing (CLS) project
- Additional breeds available for *Capra hircus* (Goat), *Ovis aries* (Sheep), and *Sus scrofa* (Pig)
- Ensembl VEP now supports the GENCODE Primary transcript set
- Regulatory annotation updates for *Homo sapiens* (Human) and *Mus musculus* (Mouse)

[More release news](#) on our blog

**All genomes**  
-- Select a species --

**Pig breeds**  
Pig reference genome and 20 additional breeds

[View full list of all species](#)

**Favourite genomes**

- Human** GRCh38.p14
- Mouse** GRCm39
- Zebrafish** GRCz11

**Still using GRCh37?**

**New genome assemblies are now being released to the [Ensembl Beta site](#).**  
All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site.  
The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

[Find out more on our blog](#)

**Ensembl Rapid Release**

**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.

 **GLOBAL CORE BIODATA RESOURCE**

 **elixir Core Data Resource**

# LES OUTILS

# Les outils

**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

**Search**  
 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 20 additional breeds  
[View full list of all species](#)

**Favourite genomes**  
 Human GRCh38.p14  
[Still using GRCh37?](#)  
 Mouse GRCm39  
 Zebrafish GRCz11

**Compare genes across species**  


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


**Gene expression in different tissues**  


**Retrieve gene sequence**  
  
GCTTACATTCCTCCATTG  
GCGCTTGCGGGCGCGCG  
GGGCTCTCCTGCGCGCTG  
AAGGGACAGAATTGTTGCG  
CACCTCTGGAGCGCGCTG  
CCAGCTGCCGCGCTGCG

**Find a Data Display**  


**Use my own data in Ensembl**  


**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 113 (October 2024)**

- Integration of lncRNA transcripts from the Capture Long-read Sequencing (CLS) project
- Additional breeds available for *Capra hircus* (Goat), *Ovis aries* (Sheep), and *Sus scrofa* (Pig)
- Ensembl VEP now supports the GENCODE Primary transcript set
- Regulatory annotation updates for *Homo sapiens* (Human) and *Mus musculus* (Mouse)

[More release news](#) on our blog

**Ensembl Rapid Release**

New genome assemblies are now being released to the [Ensembl Beta site](#). All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site. The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated. Find out more on our [blog](#)

**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 BISCUIT RESOURCE** **elixir** Core Data

# 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  
LW QTCGELMFSLEPRLRHLGLGKEGITYFSGNCTMEDAKLAQDFLDSQNLSAYNTRLFKEVDGEKPYYE  
VRL ASVLGSEPSLDSEVTSKLKSYEFRGSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIESFTQ  
G SIEAHKRGSRFWIQDKGPIVESYIGFIESYRDPFGSRGEFEVFVAVVNKAMSAKFERLVASAEQLLKE  
LPWP PTFEKDKFLTPDFTSLDVLTFAKGSGIPAGINIPNYDDLQRTEGFKNVSLGNVLAVAYATQRE  
KLTLEEDDK DLYILWKGPSFDVQVGLHELLGHGSGKLFVQDEKGAFNFDQETVINPETGEQIQSWYRS  
GETWDSKFSTIAS SYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIYVNWLNMVRAGLLALEFY  
TPEAFNWRQAHMQARFVIL RVLLEAGEGLVTITPTTGSDGRPDARVRLDRSKIRSVGKP  
ALERFLRRIQVLKSTGDVAGGRALYEGYATVT DAPPECFLTLRDTVLLRKESRK  
LIVQPNT  
RLEGSDVQLLEYASAAGLIRS  
FSERFPEDGPELEEILTQLAT ADARFWKGPSEAPSGQA

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:  
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  
+VI VNWLNMV AGILLEAFYTFPEA NW+QAH++AR VILRVL EAGEGL TITPT GSD  
Sbjct: 76089885 EVISVNLNMVGAGLLALEFYTFPEASNWQOAHIRARIVLRLPEAGEGLGTITPTAGSD 76089706  
  
Query: 597 GRPDARVRLDRSKIRSVGKPALERFLRRLOVLKSTGDVAGGRALYEGYATVTDAPPECFL 656  
GRP+A+VRLDRSKI+SVG PALERFLRR STGDVAGG LYE YA V DAPPE FL  
Sbjct: 76089705 GRPEAQVRLDRSKIOSVGNPALERFLRRCW---STGDVAGGWTLYERYAAVADAPPEGFL 76089535  
  
Query: 657 TLRDTVLLRKESRKLIIVQPNTRLLEGSDVOLLEYEASAAGLIRSFSEFPEDGPELEEILT 716  
TLRD VLLRKES KLIIVQPNTLLEGSDVOLLEYE SAAGLIRSFSE FPEDG ELE+ILT  
Sbjct: 76089534 TLDRVLLRKESWKLIIVQPNTLLEGSDVOLLEYEASAAGLIRSFSEHFPEDGLEDILT 76089355  
  
Query: 717 QLATADARFWKGSEAPSGOA 737  
QLATADA+F KGPSEAPSGQA  
Sbjct: 76089354 QLATADAOF\*KGPSEAPSGOA 76089292  
  
Score = 1700 (557.6 bits), Expect = 0., Sum P(6) = 0.  
Identities = 212/252 (84%), Positives = 221/252 (87%), Frame = -2  
  
Query: 224 PSLDSEVTSKLKSYEFRGSPFQVTTRGDYAPILKQVVEQLEKAKAYAANSHQGQMQLAQYIE 283  
P L + SKLKS EFRGSPFQVT G+Y PILQKVVQLEKAK YAANSHQ QMLAQYIE  
Sbjct: 76090816 PGLRGD--SKLKS\*EFRGSPFQVTWGNYMPILQKVVQLEKAKTYAANSHQEQMQLAQYIE 76090643  
  
Query: 284 SFTQGSIEAHKRGSRFWIQDKGPIVESYIYGIESYRDPFGSRGEFEGFVAVVNKAMSAKF 343  
SFTQGS EAHK+GSRFWI DKGPIVESYI FI+SYRD FGSRG EGFVAVVNKAMSAKF  
Sbjct: 76090642 SFTQGSTEAHKKGSRFWI\*DKGPIVESYIEFIQSYRDSFGSRGVCEGFVAVVNKAMSAKF 76090463  
  
Query: 344 ERLVASAEEQLLKELPWPPTFEKDKFLTPDFTSLDVLTFAGSGIPAGINIPNYDDLRLQTEG 403  
E LV SAEQLLKELPW P FEKDKFLTPDFTS+DVLTFAGSGI AGINI NY+DL+QTEG  
Sbjct: 76090462 EWLVVAEQLLKELPWSFAFEKDKEFLTPDFTSVDLTFAGSGIAAGINISNYNDLKQTEG 76090283  
  
Query: 404 FKNVSLGNVLAVAYATQREKLTFLLEDDKDLYILWKGPSFDVQVGLHELLGHGSGKLFVQ 463  
FKNVSLGNVLAV ATQ EKLT LEE DKDLYI+ GPSFDVQVGLHELLG+GSGKL Q  
Sbjct: 76090282 FKNVSLGNVLAVV\*ATQWEKLTVLEESDKDLYIVLMGPSFDVQVGLHELLGYGSGKLFIEQ 76090103  
  
Query: 464 DEKGAFNFDQET 475  
DEKGAFNFDQET  
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:

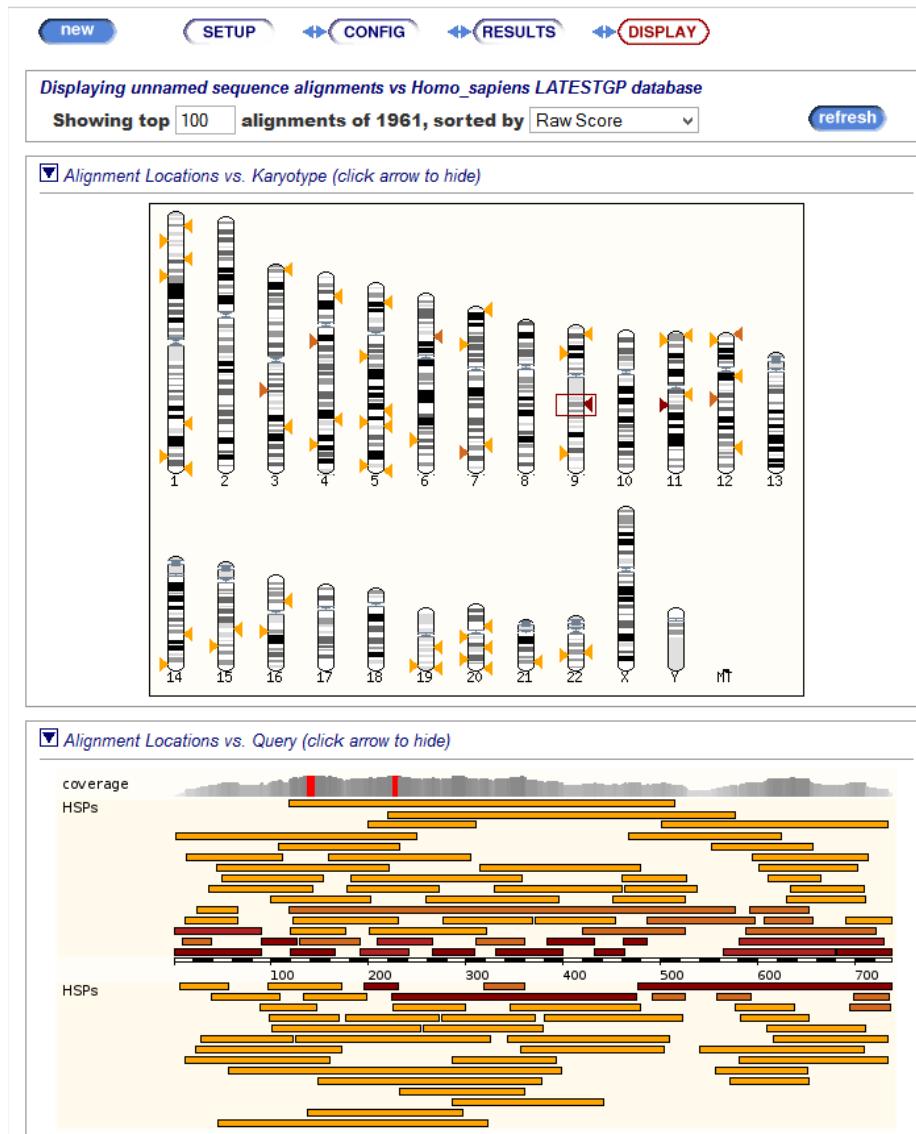
BLA\_IESTYdDXDJ    Retrieve

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_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Score E-val	>Lrg <Score >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

Alignments length: 24  
Percentage identity: 88.29

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 INPEEQIQSWYRSGETWDKFSTIASSYEECRAESVGLYLCLHNPQVLIEFGFEGADDE 536
         INPE EQIQSWYRS +IWDSKFSTI SSYEECRAESVGLYLCLHNPQVLIEFGFEGADDE
Sbjct: 76090065 INPEEQIQSWYRSKMTWDKFSTIASSYEECRAESVGLYLCLHNPQVLIEFGFEGADDE 76089886

Query:   537 DIVLYSNWLMLMVAGGILALEYYTPEAANWQAHM+QARFVILVLLVEAGLGLVIIITPFGSD 596
         +VI VNHLMLMV AGGILALEYYTPEA NW+QAH+ +AR VILVLF EAGEL GLVIIITPFGSD
Sbjct: 76089885 EVISUNWLMLMVAGGILALEYYTPEAUNWQAHIRARIVILVLFPEAGEGLGVIIITPFGSD 76089706

Query:   597 GRFDARVRLDRSK1RSVKGPKALERFLRQLVLR3TGWAGGRALYEYHNTDAPPECFL 656
         GRFA+A+VRLDRSK1+SVC PALERFLRQLVLR3TGWAGGRALYEYHNTDAPPECFL
Sbjct: 76089705 GRFEPAVRLDRSK1QVGUNALERFLRQRCW--STGUVAGWTNLKREANADAPPEGFL 76089593

Query:   657 TLRDTVLLRRESKLRLVQPNTRLEGSVDQVLLEYEASAAGLIRS+ERFPEDGELEIILT 716
         TLRD VLLRKES KLIVQPN RLEGSVDQVLLEYEASAAGLIRS SE FPDG ELE+ILT
Sbjct: 76089594 TLRD VLLRKESKLVLVQPNTRLEGSVDQVLLEYEASAAGLIRS FSEMFHDGLELEDIT 76089355

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

```

## [S]equence

Alignment score : 1765  
E-value : 0.  
Alignment length : 261  
Residue identity : 84.00

```

percentage identity: 84.29

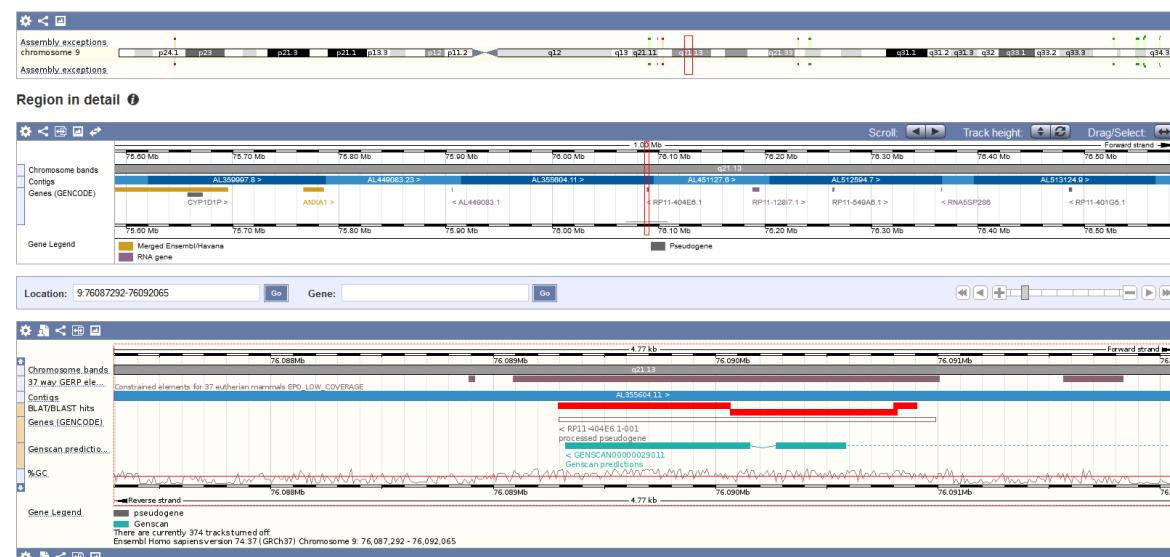
THIS STYLE: Matching bases for selected HSP
THIS STYLE: Matching bases for other HSPs in selected hit

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

```

## [C]ontig view (?)

Chromosome 9: 76.087.292-76.092.065



# Les outils

## Annotation de variants

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

Login/Register

Search all species...

**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  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 20 additional breeds

[View full list of all species](#)

**Favourite genomes**

Human  
GRCh38.p14  
[Still using GRCh37?](#)

Mouse  
GRCm39

Zebrafish  
GRCz11

**Ensembl Rapid Release**

New genome assemblies are now being released to the [Ensembl Beta site](#).  
All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site.  
The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

[Find out more on our blog](#)

**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 113 - October 2024 © EMBL-EBI

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

74

# Variant Effect Predictor

Screenshot of the Ensembl Variant Effect Predictor (VeP) website:

The page title is "Ensembl Variant Effect Predictor (VEP)". The main content area includes:

- A brief description: "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."
- A list of features:
  - 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](#).
- A section titled "What's new in release 113!"
- A "VEP interfaces" section with three options:
  - Web interface**: Point-and-click interface, suits smaller volumes of data. Includes a "Launch VeP" button.
  - Command line tool**: More options and flexibility, for large volumes of data. Includes links to "Clone from GitHub", "Download (zip)", and "Pull Docker image from DockerHub".
  - REST API**: Language-independent API, simple URL-based queries. Includes a "VeP REST API" link.
- A "Publication" section with citation information:

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, Fllicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biology* Jun 6;17(1):122. (2016) doi:10.1186/s13059-016-0974-4. [GR](#)

# 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...". Below the navigation bar, the "Variant Effect Predictor" logo is displayed, followed by a "New job" button and a "Clear form" button.

The main form area has several sections:

- Species:** Set to "Homo\_sapiens".  
Assembly: GRCh38.p14  
Change species
- Name for this job (optional):** An empty input field.
- Input data:** A large text area for pasting data, with examples listed below it: Ensembl default, VCF, Variant identifiers, HGVS notations, SPDI. There are also options to upload a file or provide a file URL.
- Transcript database to use:** A group of radio buttons:
  - Ensembl/GENCODE transcripts
  - Ensembl/GENCODE basic transcripts
  - Ensembl/GENCODE primary transcripts
  - RefSeq transcripts
  - Ensembl/GENCODE and RefSeq transcripts
- Additional configurations:** A section with several buttons:
  - Identifiers**: Additional identifiers for genes, transcripts and variants
  - Variants and frequency data**: Co-located variants and frequency data
  - Additional annotations**: Additional transcript, protein and regulatory annotations
  - Predictions**: Variant deleteriousness predictions, e.g. SIFT, PolyPhen
  - Filtering options**: Pre-filter results by frequency or consequence type
  - Advanced options**: Additional enhancements

A large green "Run >" button is located at the bottom of the configuration section.

# Variant Effect Predictor

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

Login/Register

Search all species...

VEP ▾

**Variant Effect Predictor results**

**Job details**

**Summary statistics**

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	6
Overlapped transcripts	47
Overlapped regulatory features	1

**Consequences (all)**

**Coding consequences**

**Results preview**

**Navigation (per variant)**

**Filters**

Page: 1 of 1 | Show: All variants

Uploaded variant is defined Add

All: VCF VEP TXT BioMart: Variants Genes

**Show/hide columns (18 hidden)**

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Score	AAE	MLE
1_65568_A/C	1_65568_65568	C	downstream_gene_variant	OR4G1P	ENSG00000240361	Transcript	ENST00000492842.2	transcribed_unprocessed_pseudogene	-	-	-	-	-	-	A	A/C
1_65568_A/C	1_65568_65568	C	missense_variant	OR4F5	ENSG00000166092	Transcript	ENST00000641515.2	protein_coding	2/3	64	4	2	K/Q	AAG/CAG	A	A/C
1_65568_A/C	1_65568_65568	C	downstream_gene_variant	-	ENSG00000290826	Transcript	ENST00000642116.1	lncRNA	-	-	-	-	-	-	A	A/C
1_65568_A/C	1_65568_65568	C	downstream_gene_variant	-	ENSG00000290826	Transcript	ENST00000832531.1	lncRNA	-	-	-	-	-	-	A	A/C
2_265023_C/T	2_265023_265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272065.10	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272067.11	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000356150.10	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000402632.5	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403657.5	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403658.5	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403712.6	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000405233.5	protein_coding	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	intron_variant, NMD_transcript_variant	ACP1	ENSG00000143727	Transcript	ENST00000405364.2	nonsense-mediated_decay	-	-	-	-	-	-	C	C/T
2_265023_C/T	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000405430.5	protein_coding	-	-	-	-	-	-	C	C/T

# Outils de récupération de données

Screenshot of the Ensembl website showing data retrieval tools.

The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads (highlighted with a red box), Help & Docs, and Blog. A search bar at the top right says "Search all species..." with a magnifying glass icon. The left sidebar has sections for "Using this website", "Annotation and prediction", "Data access" (also highlighted with a red box), "API & software", and "About us". A "Help & Documentation" link leads to the current page.

## Accessing Ensembl Data

Ensembl data is available through a number of routes - which you choose depends on the amount and type of data you wish to fetch. Please note that Ensembl coordinates always have a one-based start.

### Small quantities of data

Many of the pages displaying Ensembl genomic data offer an [export](#) option, suitable for small amounts of data, e.g. a single gene sequence.

Click on the 'Export data' button in the lefthand menu of most pages to export:

- FASTA sequence
- GTF or GFF features

...and more!



### Fast programmatic access

For fast access in any programming language, we recommend using our [REST server](#). Various REST endpoints provide access to vast amounts of Ensembl data.



### Complete datasets and databases

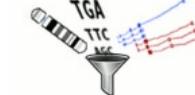
Many datasets, e.g. all genes for a species, are available to download in a variety of formats from our [FTP site](#).

Entire databases are also available via FTP as MySQL dumps.



### Complex cross-database queries

More complex datasets can be retrieved using the [BioMart](#) data-mining tool.



All data produced by the Ensembl project is [freely available](#) for your own use.

Ensembl release 113 - October 2024 © EMBL-EBI

[Permanent link](#)

#### About Us

- [About us](#)
- [Contact us](#)
- [Citing Ensembl](#)
- [Privacy policy](#)
- [Long-term data preservation](#)
- [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](#)

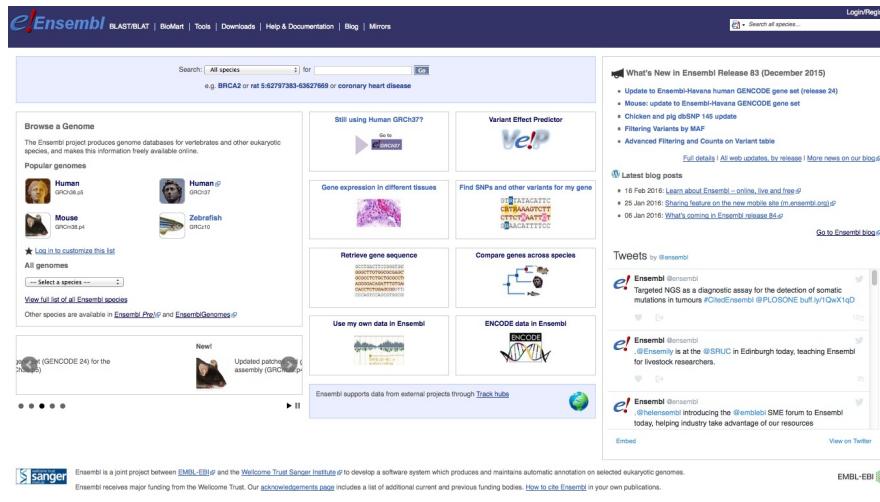
# BioMart

# Le projet BioMart

- 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

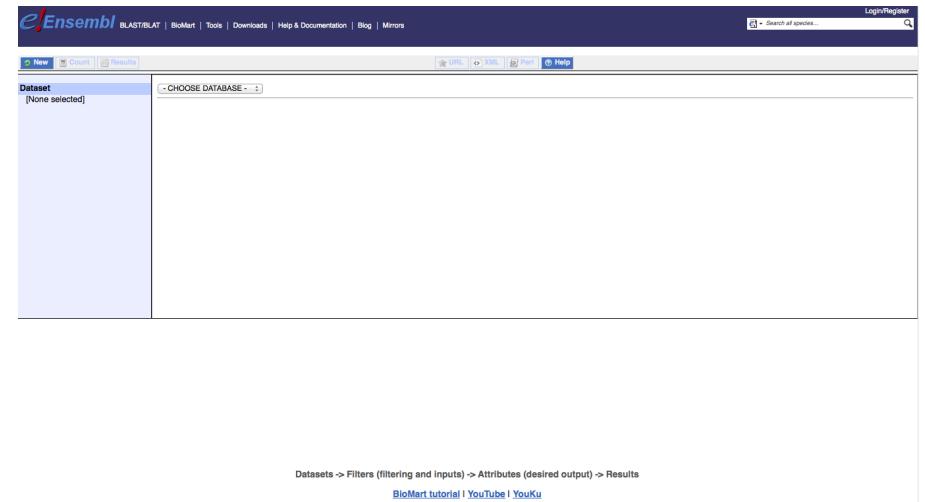
# Accéder aux données d'Ensembl

## Site web



The screenshot shows the main Ensembl website interface. At the top, there's a search bar with placeholder text "Search: All species for" and a dropdown menu showing "e.g. BRCA2 or rat 5:62797383-63627869 or coronary heart disease". Below the search bar, there's a "Browse a Genome" section with links for Human (GRCh37), Mouse (GRCh38), Zebrafish (zv9), and others. There are also sections for "Variant Effect Predictor", "Gene expression in different tissues", "Find SNPs and other variants for my gene", "Retrieve gene sequence", "Compare genes across species", "Use my own data in Ensembl", and "ENCODE data in Ensembl". A sidebar on the left lists "Popular genomes" like Human, Mouse, Zebrafish, and others. A "What's New in Ensembl Release 83 (December 2015)" section highlights updates for Human GRCh37, Mouse, Chicken, and Pig. A "Latest blog posts" section shows tweets from the Ensembl account. At the bottom, there's a footer with the Sanger logo and EMBL-EBI links.

## Outil de fouille: BioMart



The screenshot shows the BioMart interface. At the top, there's a search bar with placeholder text "Search for species..." and a "Login/Register" button. Below the search bar, there are tabs for "New", "Count", and "Results". A "Dataset" dropdown menu is set to "[None selected]". The main area is currently empty, indicating no specific dataset has been chosen yet.

-  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. The main content area has several sections: "Tools" (with a "All tools" link), "BioMart" (with a "BioMart" link highlighted by a red box), "BLAST/BLAT >" (with a "Search our genomes for your DNA or protein sequence" link), and "Variant Effect Predictor >" (with a "Analyse your own variants and predict the functional consequences of known and unknown variants" link). Below these is a "Search" section with a dropdown menu set to "All species" and a "Go" button. A note below says "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease". To the left, there's a "All genomes" section with a dropdown menu and a "Pig breeds" section showing "Pig reference genome and 20 additional breeds". To the right, there's a "Favourite genomes" section with entries for "Human" (GRCh38.p14) and "Mouse" (GRCm39), both with edit icons. A "Ensembl Rapid Release" box notes new genome assemblies available at the Ensembl Beta site, mentioning release 65 and the end of the Rapid Release website. At the bottom, there's a "Compare genes across species" section, EMBL-EBI acknowledgements, and logos for "own data in Ensembl", "GLOBAL CORE BIODATA RESOURCE", and "elixir Core Data Resource".

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

# BioMart/Ensembl

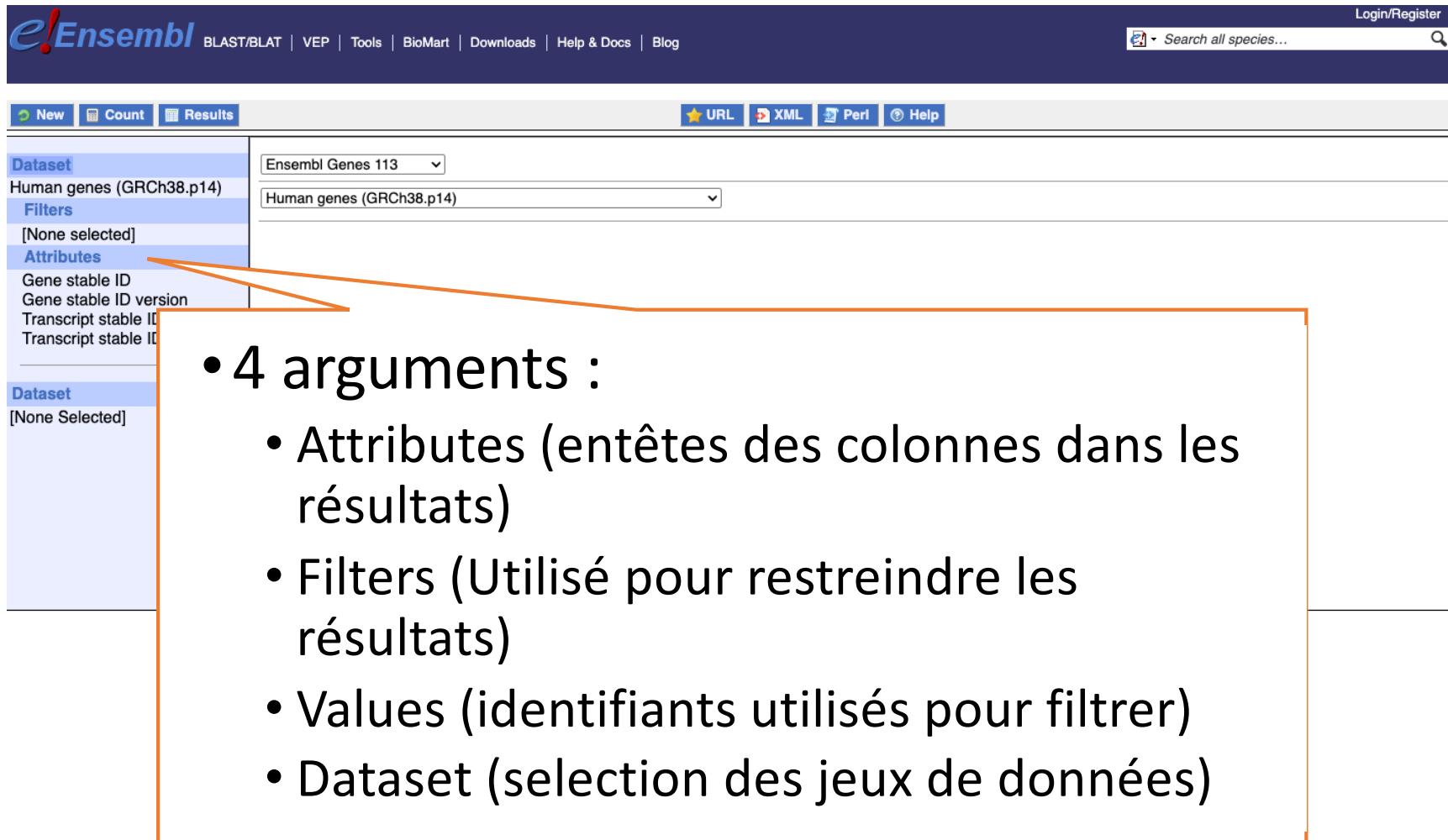
The screenshot shows the Ensembl BioMart interface. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. Below the navigation bar, there are three tabs: New, Count, and Results. The 'New' tab is selected. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p14)), Filters (None selected), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version). The main area is titled 'Dataset' and shows '[None Selected]'. Above the main area, there are two dropdown menus: one for 'Dataset' set to 'Ensembl Genes 113' and another for 'Species' set to 'Human genes (GRCh38.p14)'. Orange arrows point from the text 'Sélection de la Base de données :' to the 'Dataset' dropdown and from 'Sélection du jeu de données (génome)' to the 'Species' dropdown.

Sélection de la Base de données :

- Genes
- Variation
- Regulation
- Mouse strain

Sélection du jeu de données (génome)

# BioMart/Ensembl



The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p14)), Filters ([None selected]), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID). The main area shows a dropdown for 'Dataset' set to 'Ensembl Genes 113' and another dropdown for 'Human genes (GRCh38.p14)'. At the top, there are tabs for New, Count, Results, and links for URL, XML, Perl, and Help. A search bar at the top right says 'Search all species...'.

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

# 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