

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



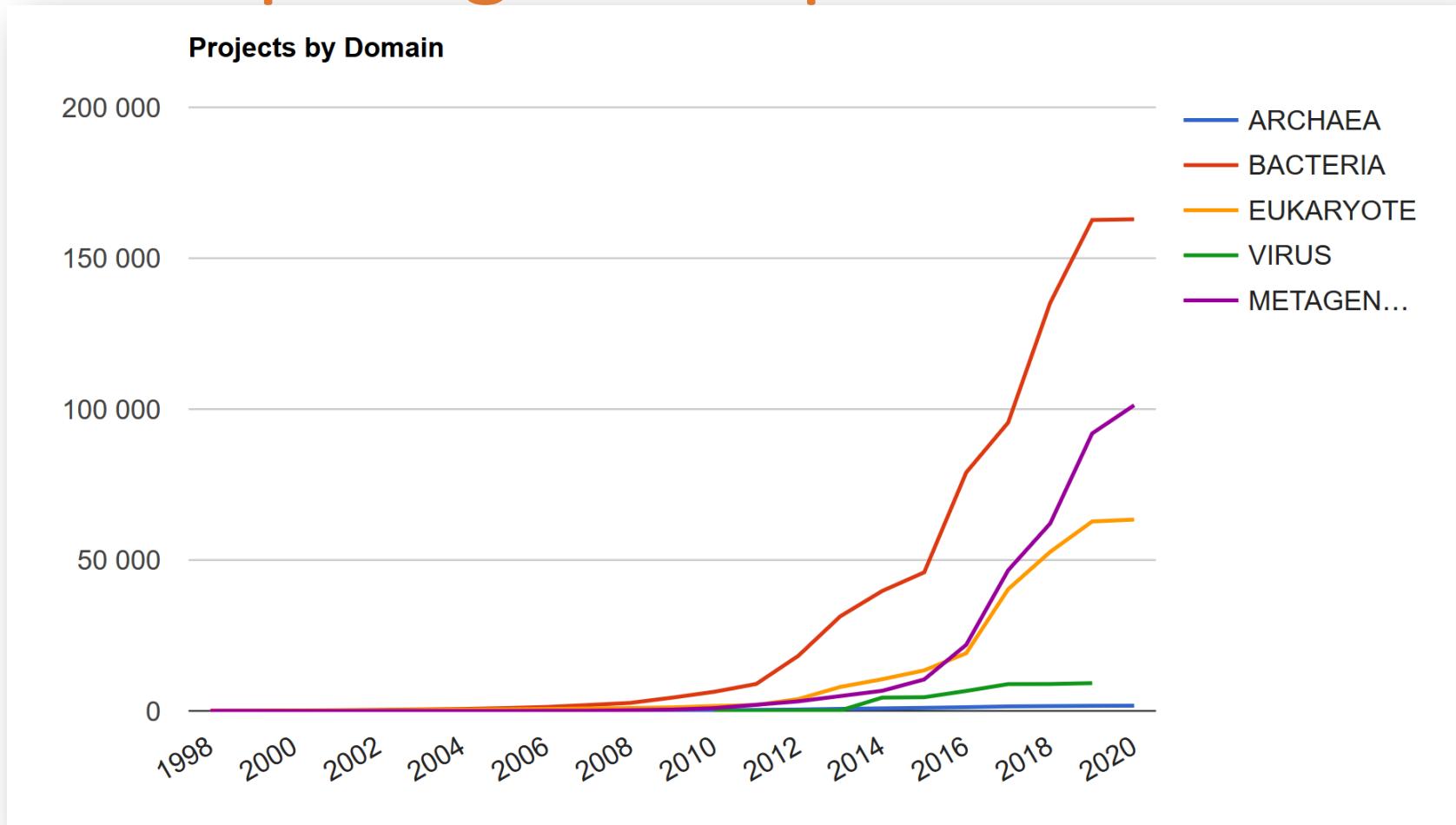
L'ère post-génomique

- 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 : GAIIx
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)
956 archées, 31736 bactéries et 9173 eukaryotes (www.genomesonline.org, 10/2016)

Exomes et génomes humains séquencés complètement (patients + pop. Générale)



L'ère post-génomique



Centres de bioinformatique

- EBI (European Bioinformatics Institute)



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

- NCBI (National Center for Biotechnology Information)

The screenshot shows the NCBI homepage with a blue header. On the left is the NCBI logo (a stylized 'S' icon) and the text 'NCBI'. To the right is the title 'National Center for Biotechnology Information' with smaller text 'National Library of Medicine' and 'National Institutes of Health' below it. A horizontal menu bar follows, with links: PubMed, All Databases, BLAST, OMIM, Books, TaxBrowser, and Structure. Below the menu is a search bar containing the text 'Search All Databases' with a dropdown arrow, a text input field, and a 'Go' button.

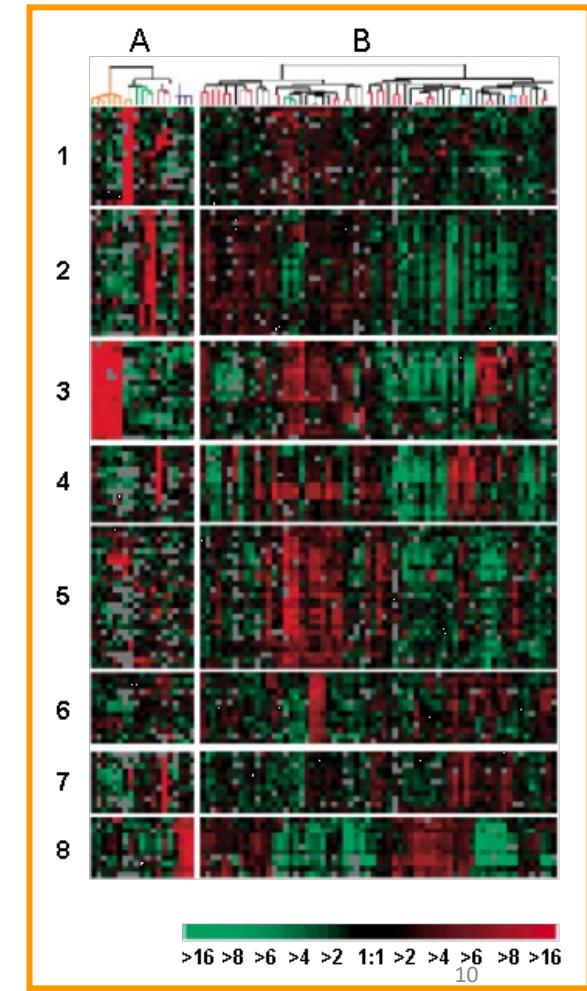
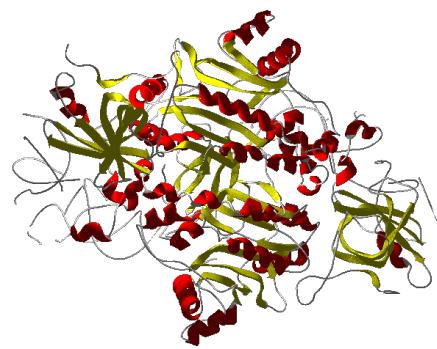
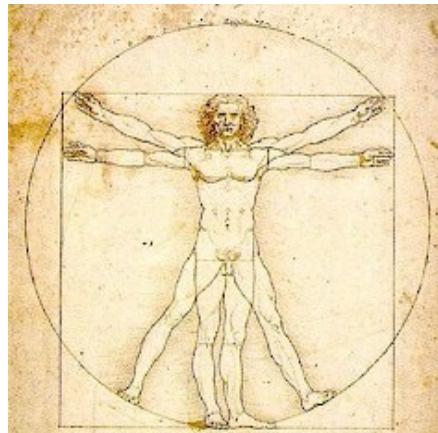
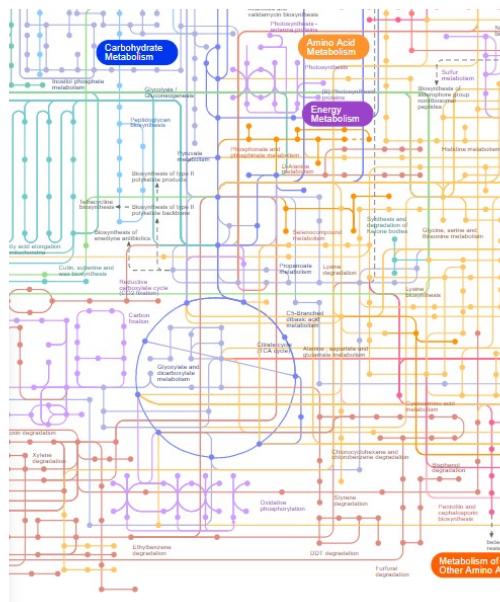
<http://www.ncbi.nlm.nih.gov/>

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

MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVGCALELAGGPGA



Banques de séquences nucléiques généralistes



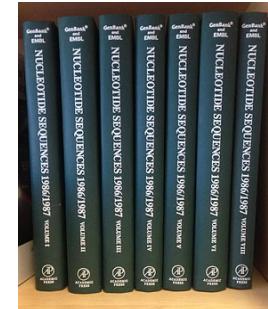
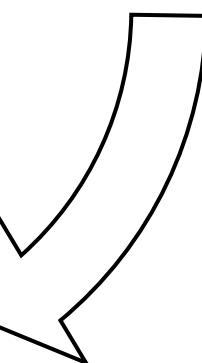
GenBank



EMBL



DNA
databank of
Japan



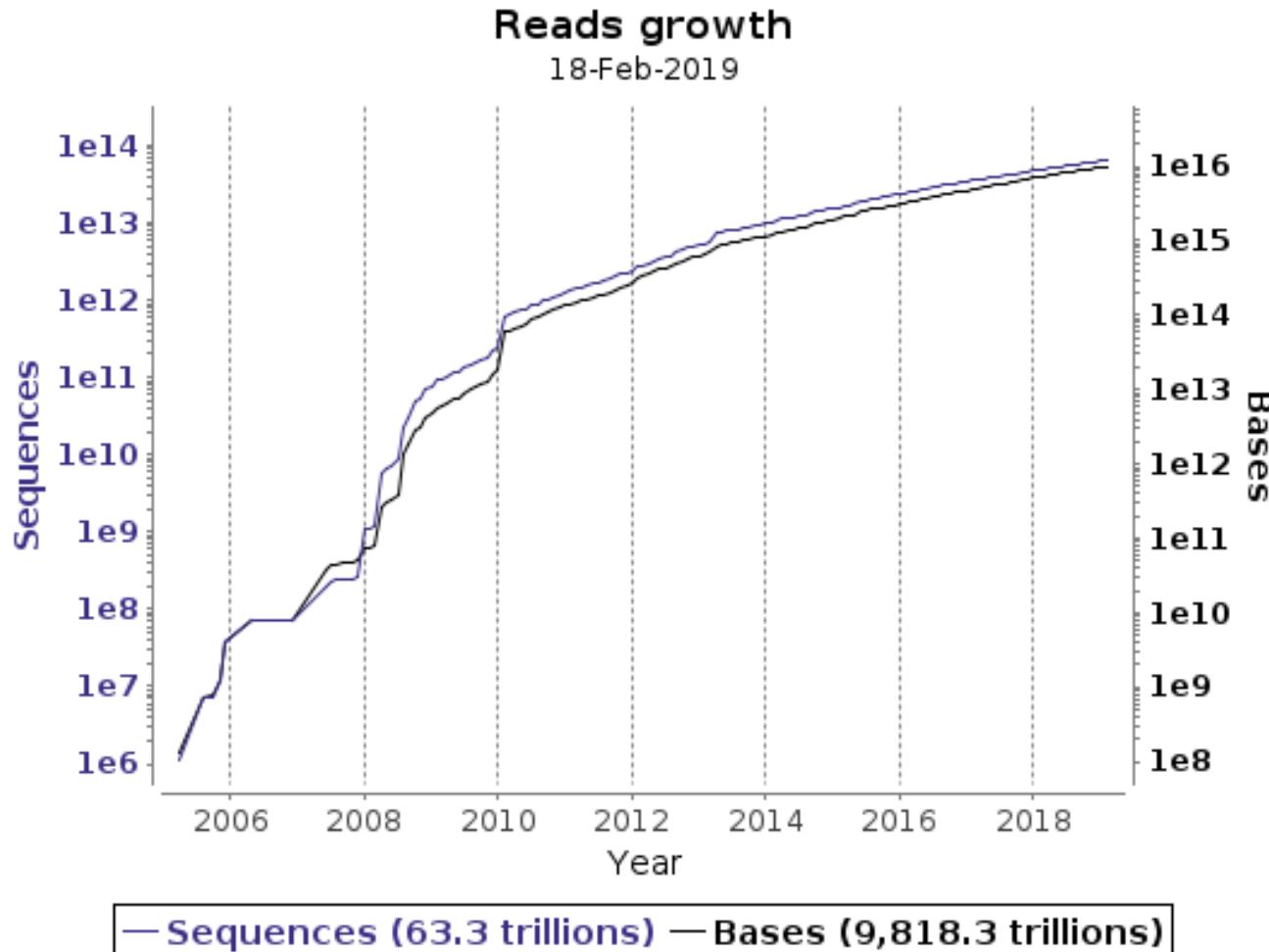
- 3 banques
- Échanges quotidiens des séquences collectées
- Effort d'unification=> format
 - accord entre GenBank et EMBL en 1986
 - accord entre GenBank/EMBL et DDBJ in 1987



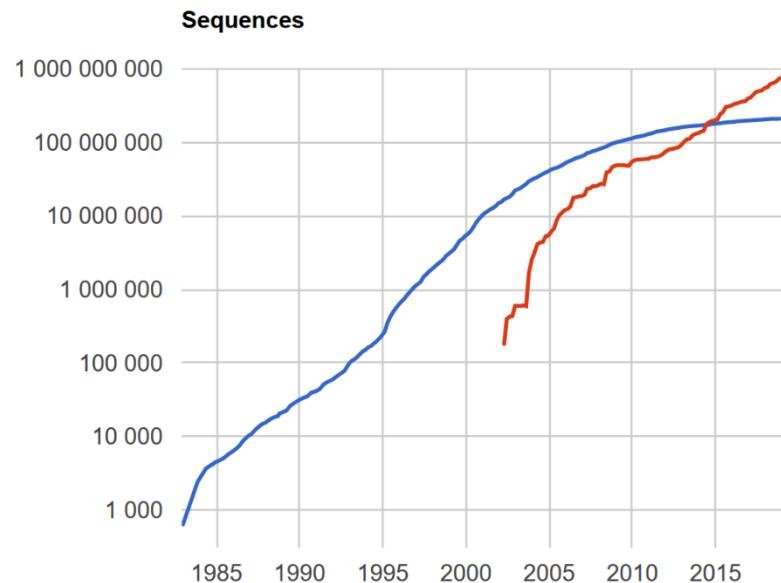
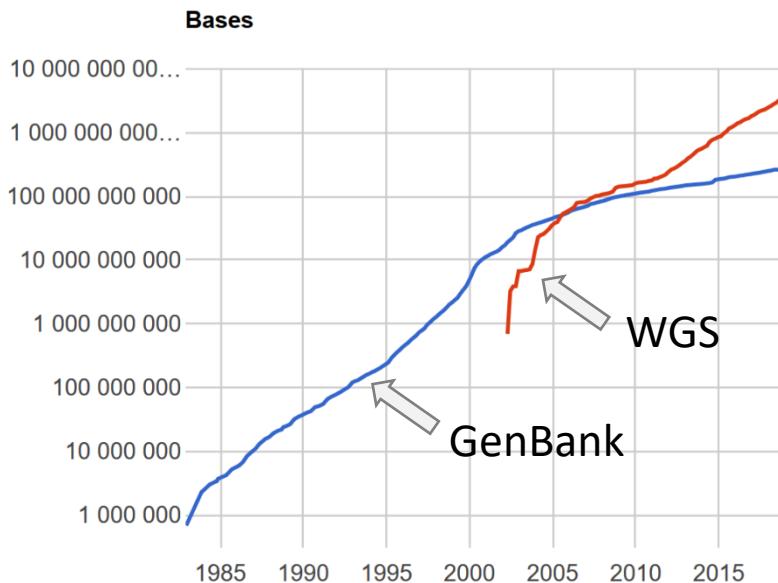
Banques de séquences

- 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

Evolution de la banque EMBL



Evolution de la banque GenBank



12/2018: 285 milliards de nucléotides, 211 millions d'entrées
Doublement tous les 18 mois

Banques de séquences protéiques généralistes



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

03/2018	01/2019	02/2020
106,245,682	130,366,644	167,278,920

Transcrits: 29,869,155
Organismes: 99,842



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

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



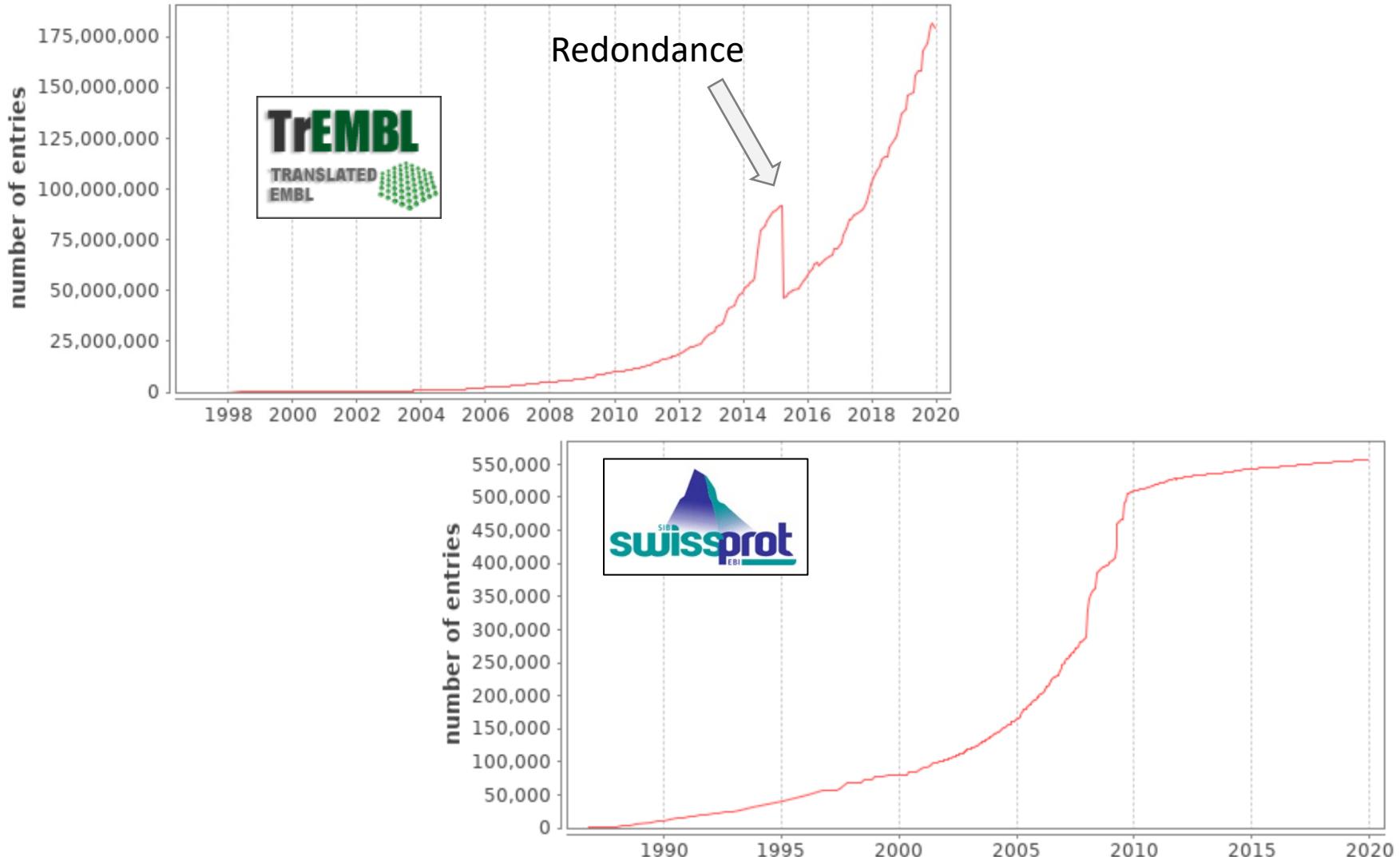
TrEMBL:
179,250,561 entrées

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	387,632	70,5%	11,091,443	21,1%
Predicted	11,465	2,1%	40,603,140	76,9%
Uncertain	1,955	0,4%	0	0%

Evolution des bases de données protéiques



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** »

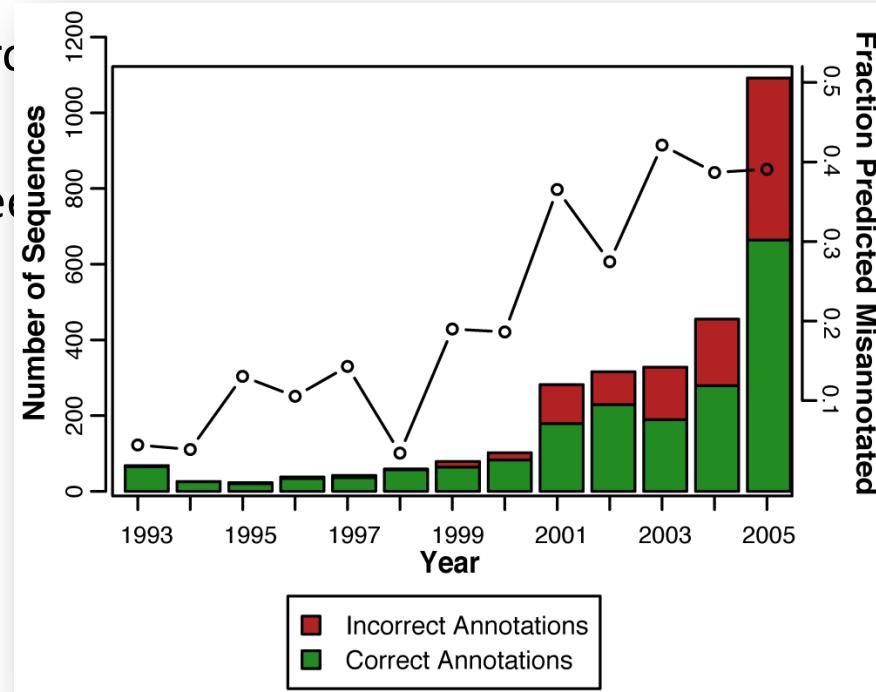
Exemple de l'importance de l'annotation

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

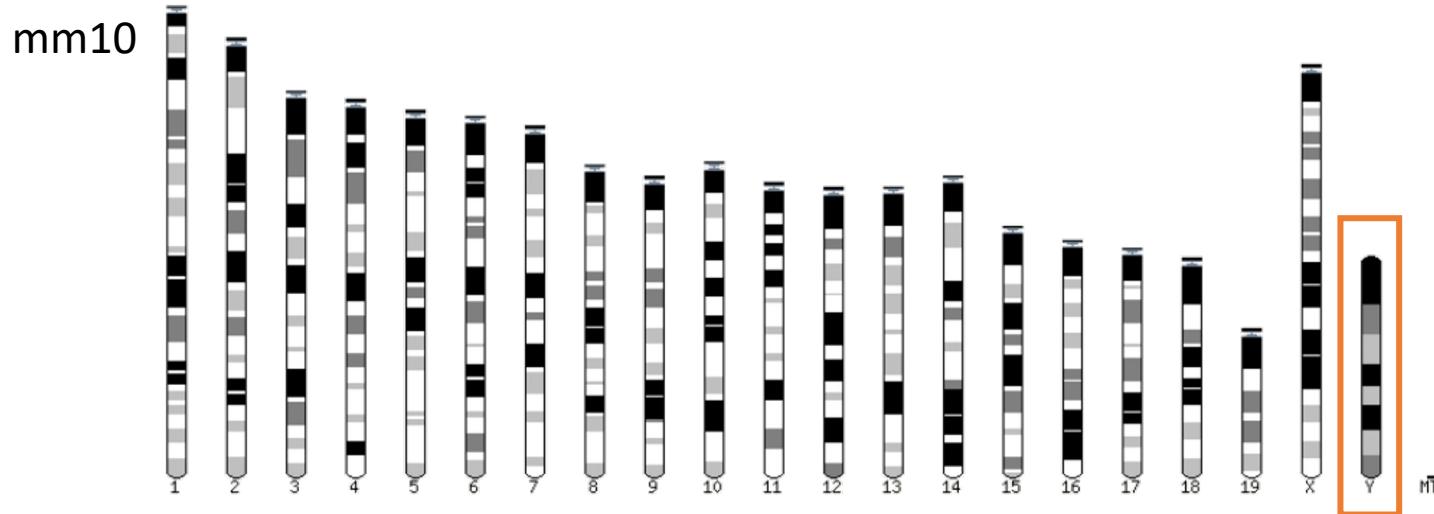
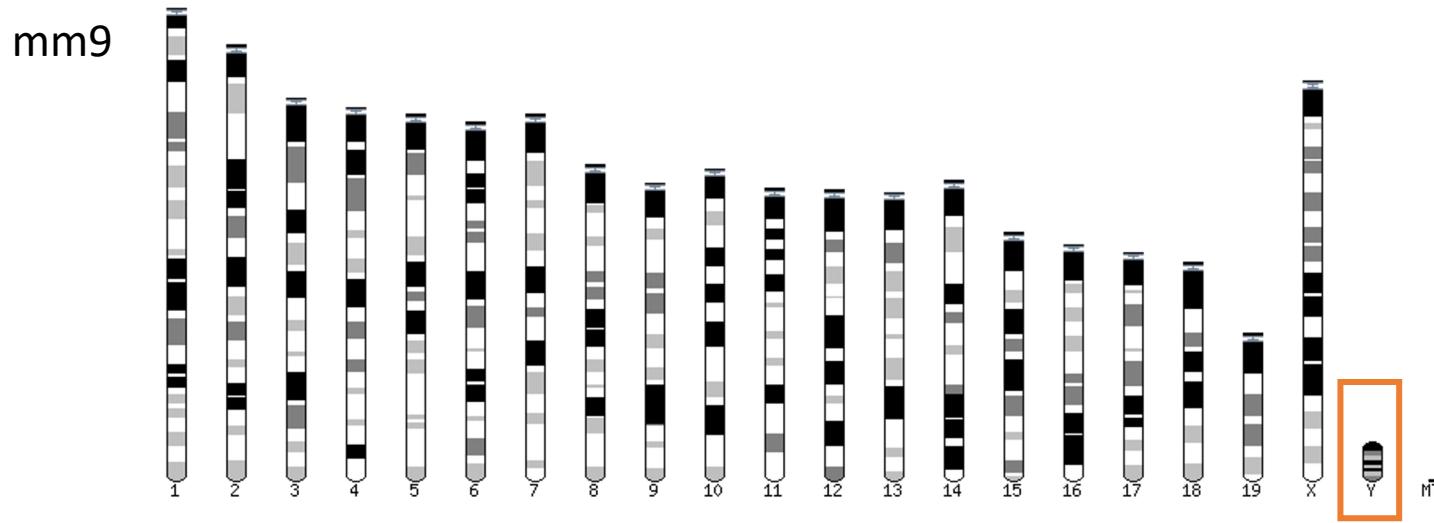
Bases de données majeures
collecte des données individuelles et collectives

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
MAMMALS				
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

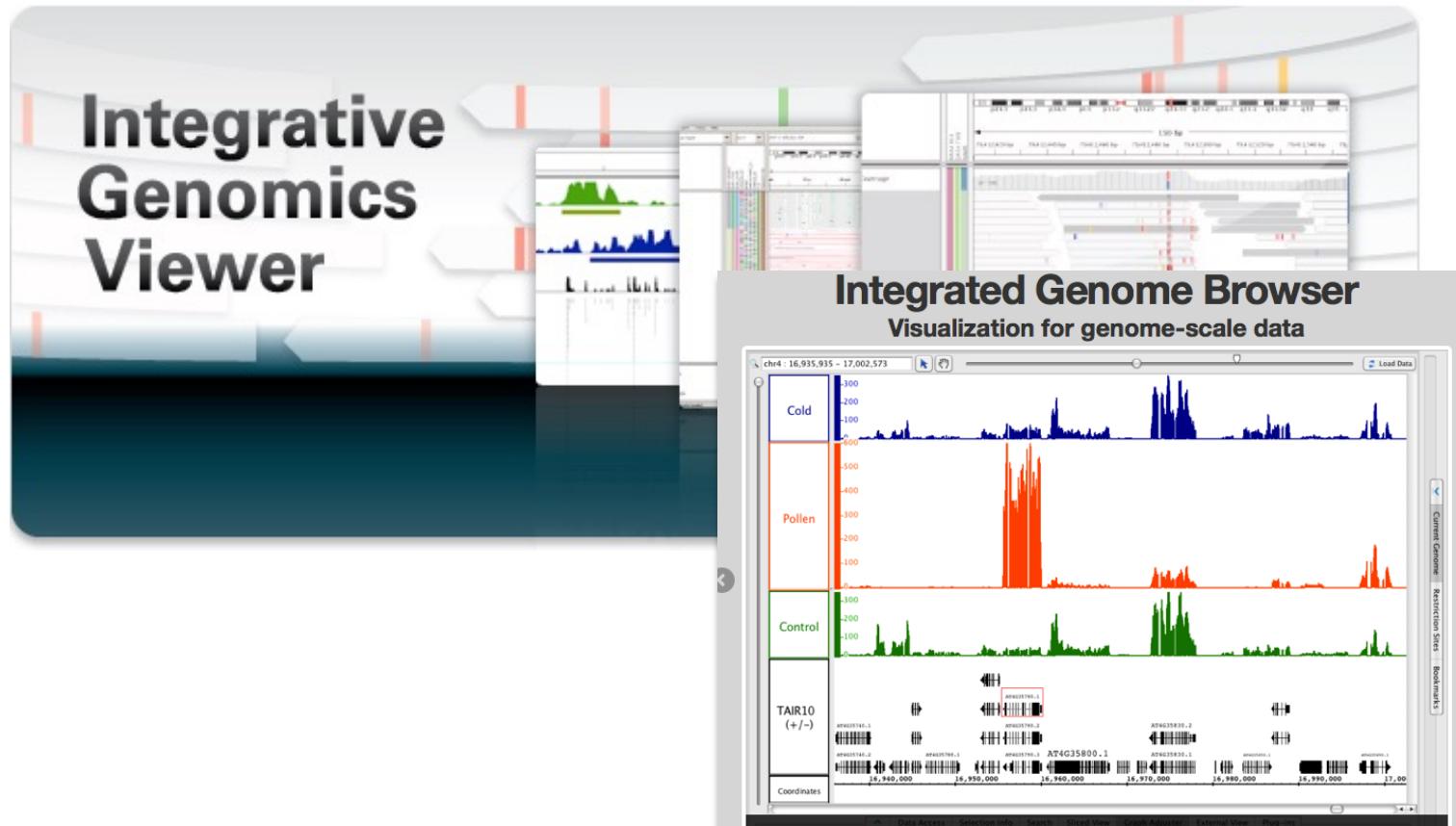
Il y a Genome Browsers...

EBI - Ensembl

UCSC – Genome Browser

NCBI – Map Viewer

Et Genome browsers



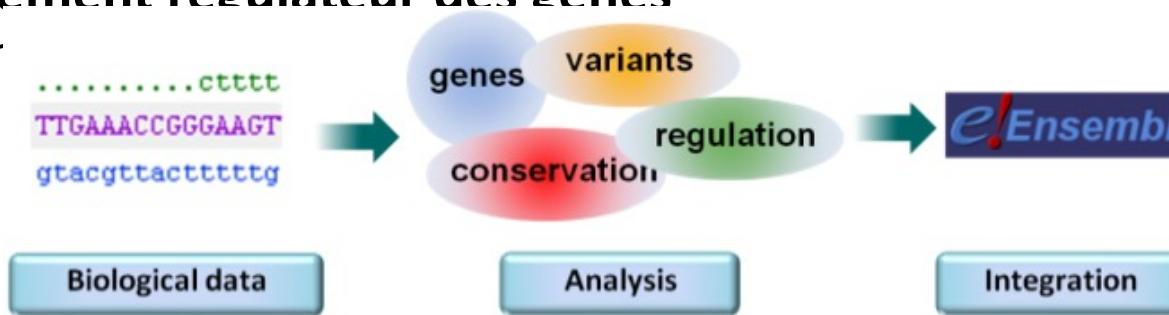
Ensembl

Le projet Ensembl

- Initié en 1999 (avant la première version du génome humain)
- Projet en collaboration entre l'European Bioinformatics Intitute (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
 - Ar



- 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

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

Login/Register Search all species...

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 12 additional breeds

[View full list of all species](#)

Favourite genomes

Human GRCh38.p13
Still using GRCh37?

Mouse GRCm39

Zebrafish GRCz11

Ensembl Release 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (*Canis lupus familiaris*) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

[More release news](#) on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

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

Go

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

[Rapid Release news](#) on our blog

Other news from our blog

- 19 Jan 2022: [Update to the Ensembl COVID-19 resource](#)
- 12 Jan 2022: [Homology data available in Ensembl Rapid Release](#)
- 17 Dec 2021: [146 new insect genomes on Ensembl Rapid Release](#)

Compare genes across species

Find SNPs and other variants for my gene

GTATACATTC
CTTAAAGCTT
CTTCATTG
GAACATTTCC

Gene expression in different tissues

Retrieve gene sequence

GGCTTGCTTCGGCGTTC
GGGGCTTGTGCGCGCGAC
GGGGCTCTCTGCGCGCGCT
AAGGGCGCGCGCGCGCGCG
GGGGCTCTGGCGCGCGCGCG
GGGGCTCTGGCGCGCGCGCG

Find a Data Display

TABLE
MAP
REGION
PCP
CHART

Use my own data in Ensembl

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available.

Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

elixir Core Data Resource

Ensembl release 105 - Dec 2021 © EMBL-EBI

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

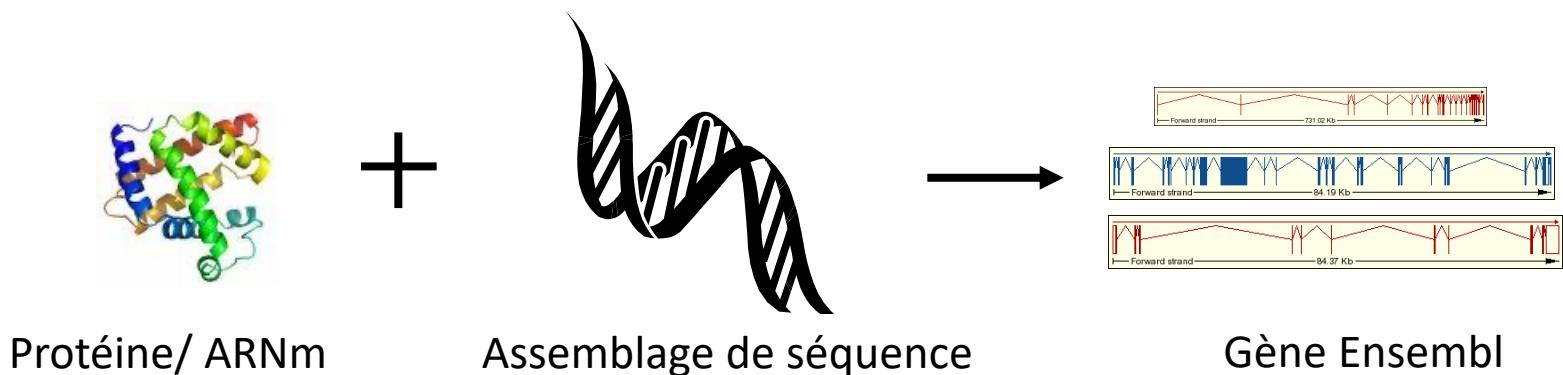
Comprendre ENSEMBL

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

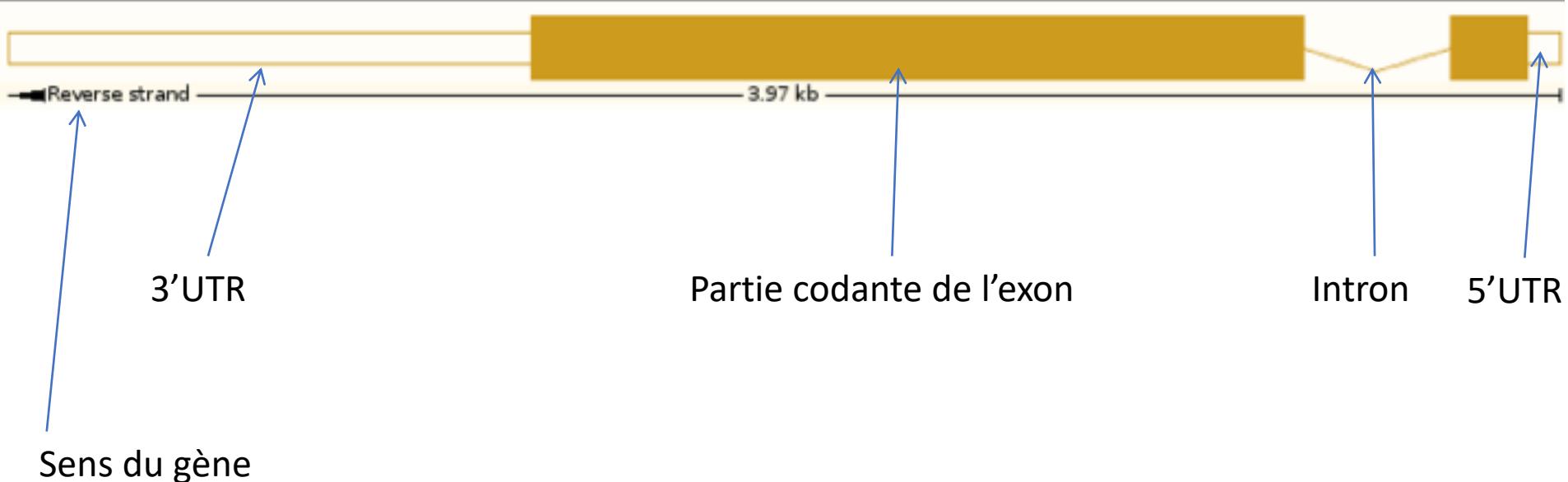
- Les transcrits d'Ensembl sont basés sur les bases de données suivantes :
 - Uniprot/Swiss-Prot (*curation manuelle*)
 - Uniprot/TrEMBL
 - NCBI refSeq (*curation manuelle*)



Les annotations

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

Transcrits Ensembl

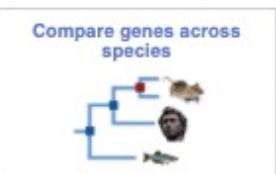


Identifiants Ensembl

- ENS**G**### Ensembl Gene ID
- ENST**T**### Ensembl Transcript ID
- ENSP**P**### Ensembl Peptide ID
- ENSE**E**### Ensembl Exon ID
- Ajout d'un suffix pour les autres espèces
 - MUS (*Mus musculus*) pour la souris: ENS**MUS**G###
 - DAR (*Danio rerio*) pour le zebrafish: ENS**DAR**G###
 - etc.

Version (Release)

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



Compare genes across species



Find SNPs and other variants for my gene



Gene expression in different tissues



Retrieve gene sequence

```
GCCTGACTTCGGGTGG  
GGGCTTGCGGGGAGC  
GGGGCTCTCTGGGGCT  
AAGGGCAAGATTGGGA  
CACCTCTGAGACGGTT  
CCGATCCGGCTGGCG
```



Find a Data Display



Use my own data in Ensembl



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.



- Lien vers une version particulière d'Ensembl : <http://Dec2021.archive.ensembl.org/index.html>

Ensembl : Archives

[Login/Register](#)

 BLAST/BLAT | VEP | Tools | BioMart | Downloads | More ▾

Search all species... 

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

In this section  Help & Documentation > Using this website > Archives

Archives: Table of assemblies

Search documentation 

Ensembl Archives

About Archive Ensembl

The main Ensembl site (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 NCBI37 assembly** is available at our [Ensembl 67 archive](#) site.
- As from August 2014 we are supporting the **human GRCh37 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

List of currently available archives

- [Ensembl GRCh37](#): Full Feb 2014 archive with BLAST, VEP and BioMart
- [Ensembl 105: Dec 2021](#) - this site
- [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 90: Aug 2017](#)
- [Ensembl 89: May 2017](#)
- [Ensembl 88: Mar 2017](#)
- [Ensembl 87: Dec 2016](#)
- [Ensembl 86: Oct 2016](#)
- [Ensembl 80: May 2015](#)
- [Ensembl 77: Oct 2014](#)
- [Ensembl 75: Feb 2014](#)
- [Ensembl 54: May 2008](#)

[Table of archives showing assemblies present in each one.](#)

<http://www.ensembl.org/info/website/archives/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 

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

All genomes  **Favourite genomes** 

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

 **Mouse**
GRCm39

 **Zebrafish**
GRCz11

[View full list of all species](#)

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.



The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-

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 'Archives: Table of assemblies' page on the Ensembl website. The page displays a grid where each row represents a species and its assembly version, and each column represents a month and year from December 2021 to May 2017. The background color of each cell indicates the status of the assembly for that specific species and period: yellow for new species, grey for species present in the archive, and white for species not in this version of Ensembl.

	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
Abingdon island giant tortoise	ASM359739v1																
African ostrich	ASM69896v1																
Agassiz's desert tortoise	ASM289641v1																
Algerian mouse	SPRET_EIJ_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	NNQGG.v01																
Angola colobus	Cang.pa_1.0																
Arabian camel	CamDro2																
Arctic ground squirrel	ASM342692v1																
Argentine black and white tegu	HLtupMer3																
	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
Armadillo	Dasnov3.0																
Asian bonytongue	fSciFor1.1																
Asiatic black bear	ASM966005v1																
Atlantic cod	gadMor3.0																
Atlantic herring	Ch_v2.0.2																
Atlantic salmon	ICSASG_v2																
Australian saltwater crocodile	CroPor_compl																
Ballan wrasse	BallGen_V1																
Barramundi perch	ASB_HGAPassembly_v1																
Beluga whale	ASM228892v3																

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

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

Login/Register

Search all species...

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 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (*Canis lupus familiaris*) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

[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.

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

[Rapid Release news](#) on our blog

Other news from our blog

- 1 Jan 2022: [Update to the Ensembl COVID-19 resource](#)
- 12 Jan 2022: [Homology data available in Ensembl Rapid Release](#)
- 17 Dec 2021: [146 new insect genomes on 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.

elixir Core Data Resource

Ensembl Genomes

Bactéries

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

Search for a gene | Search for a genome
e.g. [hsZ or uridine](#) | Start typing the name of a genome...
e.g. type egs to find Escherichia

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

- Release 49, December 2020 [eg9-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](#)

Search for a gene - type the name of a gene or other identifier into the search box above.
 Find a genome - click in the browse a genome box above and start typing your genome name to find matching genomes.
 View full list of all Ensembl Bacteria species
 Access Ensembl Bacteria programmatically

What's New in Release 52
Release 52 of Ensembl Bacteria has no major updates since the previous release. As for releases 49-45, we are defining bacterial genomes as defined by criteria set out by UniProt. See more details about this update in our [blog post](#).

• Genomes
 • A total of 31,332 bacterial and archaeal genomes

• Data
 • Annotation of pathogen-host interaction data ([PhI-base](#)) version 2019-09-16
 • Alignments to Rfam covariance alignments (Rfam 12.2) visible in separate track (Rfam module)

Did you know?
 To access Ensembl Genomes data from any programming language, try our [REST API](#). For full documentation, including examples and a range of genomic data, visit [http://rest.ensembl.org](#).

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  

Fungi

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

Search: All species | Go
e.g. [NAT2 or alcohol](#)

All genomes **Favourite genomes**  
Select a species -  
[View full list of all species](#)

What's New in Release 52
 • Genomes
 • EnsemblFungi has 1506 genomes in total
 • 477 new genomes imported from ENA ([https://www.ebi.ac.uk/ena/browser/home](#))
 • 15 genomes imported from VEuPath DB

• Updated data
 • Updated fungal gene trees
 • Updated protein features for all species using InterProScan with version 86 of InterPro

• Updated BioMarts for all gene and variation data
 • Updated pan-taxonomic gene trees and homologies

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

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

Rapid Release news on our blog

Archive sites

Plantes

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

Search: All species | Go
e.g. Carboxy* or chx28

All genomes **Favourite genomes**  TAIR10
Select a species - 
[View full list of 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 [wheat-expression.com](#) and [KoTeMapper](#)
- 14 wheat cultivars from the [10x genome project](#)
- Alignment of 98,270 high confidence genes from the TGACv1 annotation
- Axon 53K, 800K SNP arrays from [CerealiDB](#), including QTL links in selected cases and Linkage Disequilibrium display. See QTL example [here](#).
- EMS-induced mutations from sequenced TILLING populations of *Cassava* (coding regions) and *Kinnow* (coding regions and promoters).
- Inter-*homologous Variants* (IHVs) between the A, B and D genome components
- 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 *TAO4* assembly are available at [https://www.ensembl.org](#).
- Phylogenetic analysis, allowing users to view alignments among multiple wheat components simultaneously.
- Dunum wheat 35K, 90K, 200K and TaBW200K 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 | More | Search Ensembl Protists...

Search: All species | Go
e.g. PF3D7_0523500 or cyto*

All genomes **Favourite genomes**  WBC07v2
Select a species - 
[View full list of 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.

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

Rapid Release news on our blog

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

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

Métazoaires

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

Search: All species | Go
e.g. CP934 or chitin*

All genomes **Favourite genomes**  WBC07v35
Select a species - 
[View full list of all species](#)

What's New in Release 52
 • Updated data
 • Updated species
 • *Cimex lectularius* (Hirudin)

• Updated protein features for all species using InterProScan with version 86 of InterPro

• Updated BioMarts for all gene and variation data

• Updated pan-taxonomic gene trees and homologies

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

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

Rapid Release news on our blog

Archive sites
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](#) (47

Le site web Ensembl: page d'accueil

Outils

Recherche

Liste déroulante
Accès aux génomes

Recherche

News

Accès aux archives d'Ensembl

The screenshot shows the Ensembl homepage for release 105 (December 2021). The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar at the top right allows searching across all species. The main content area features several sections:

- Tools:** Links to BioMart, BLAST/BLAT, and Variant Effect Predictor.
- Search:** A search interface with a dropdown for species selection (set to "All species") and a text input for search terms (e.g., BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease), followed by a "Go" button.
- All genomes:** A section for selecting a species, currently set to "Pig breeds". It also lists "Human" (GRCh38.p13), "Mouse" (GRCm39), and "Zebrafish" (GRCz11).
- Favourite genomes:** Shows the selected genomes: Human, Mouse, and Zebrafish.
- Ensembl Rapid Release:** Information about new assemblies and gene/protein annotation releases every two weeks. It mentions the rapid release website for recently produced genomes from initiatives like Darwin Tree of Life and the Vertebrate Genomes Project.
- Other news from our blog:** A list of recent posts, including:
 - 1 Jan 2022: Update to the Ensembl COVID-19 resource
 - 12 Jan 2022: Homology data available in Ensembl Rapid Release
 - 17 Dec 2021: 146 new insect genomes on Ensembl Rapid Release
- Footer:** Logos for EMBL-EBI and ELIXIR Core Reso, and a note that Ensembl creates, integrates, and distributes reference datasets and analysis tools.

Le site web Ensembl: les génomes

Informations, statistiques

Recherche

Lien vers des exemples

The screenshot shows the Ensembl Human genome assembly page (GRCh38.p13). Key sections include:

- Search Human (Homo sapiens)**: Includes a search bar and a note about searching categories.
- Genome assembly: GRCh38.p13 (GCA_000001405.28)**: Provides links to more information and statistics, download DNA sequence (FASTA), convert data to GRCh38 coordinates, and display data in Ensembl.
- Comparative genomics**: Discusses homologues, gene trees, and whole genome alignments across species.
- Regulation**: Details DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.
- Gene annotation**: Describes protein-coding and non-coding RNA. Includes links to more about this genebuild, download FASTA files, download GTF or GFF3 files, and update old Ensembl IDs.
- Variation**: Details short sequence variants and longer structural variants, disease, and other phenotypes. Includes links to more about variation in Ensembl, download all variants (GVF), and Variant Effect Predictor (VeP).
- Example gene**: Shows the Pax6 gene with its genomic region, transcript, and protein domains.
- Example transcript**: Shows the Pax6 transcript with its exons and introns.
- Example variant**: Shows a SNP example with the sequence ATCGAGCT, ATCCAGCT, ATCGAGAT.
- Example phenotype**: Shows eye color as an example phenotype.
- Example structural variant**: Shows a structural variant with a diagram of DNA strands.

At the bottom, there are links to Ensembl release 105 (Dec 2021), a permanent link, and sister sites like Ensembl Bacteria and Ensembl Plants. There are also "Follow us" links for social media and a blog.

Le site web Ensembl: statistiques des génomes

Informations générales sur l'assemblage

Patches
As the GRC maintains and improves the assembly, patches are being introduced. Currently, assembly patches are of two types:

sequence at a loci and will remain as haplotypes in the sequence and will replace the given region of the reference RC.

Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ | Go

Gene annotation
The Ensembl human gene annotations have been updated using Ensembl's automatic annotation pipeline. The updated annotation incorporates new protein and cDNA sequences which have become publicly available since the last GRCh38 genebuild (December 2013).
In the current release, we continue to display a joint gene set based on the merge between the automatic annotation from Ensembl and the manually curated annotation from Havana. See the statistics table, right, for the corresponding GENCODE version number. The Consensus Coding Sequence (CCDS) identifiers have also been mapped to the annotations. More information about the [CCDS project](#).
Updated manual annotation from Havana is merged into the Ensembl annotation every release. Transcripts from the two annotation sources are merged if they share the same internal exon-intron boundaries (i.e. have identical splicing pattern) with slight differences in the terminal exons allowed. Importantly, all Havana transcripts are included in the final Ensembl/Havana merged (GENCODE) gene set.

• [Detailed information on genebuild \(PDF\)](#)

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

Statistics

Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28 , Dec 2013
Base Pairs	3,096,649,726
Golden Path Length	3,096,649,726
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Aug 2021
Database version	105.38
Gencode version	GENCODE 39

Gene counts (Primary assembly)

Coding genes	20,465 (incl 653 readthrough)
Non coding genes	24,849
Small non coding genes	4,865
Long non coding genes	17,763 (incl 308 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,217 (incl 6 readthrough)
Gene transcripts	245,000

Gene counts (Alternative sequence)

Coding genes	3,053 (incl 26 readthrough)
Non coding genes	1,555
Small non coding genes	297
Long non coding genes	1,071 (incl 25 readthrough)
Misc non coding genes	187
Pseudogenes	1,799
Gene transcripts	21,638

Other

Genscan gene predictions	51,756
Short Variants	702,229,898
Structural variants	6,890,308

Statistiques

Le site web Ensembl: caryotype

Ensembl

Human (GRCh38.p13) ▾

Genome Jobs

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

Whole genome

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

Summary

Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28 , Dec 2013
Base Pairs	3,096,649,726
Golden Path Length	3,096,649,726
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Aug 2021
Database version	105.38
Gencode version	GENCODE 39

Gene counts (Primary assembly)

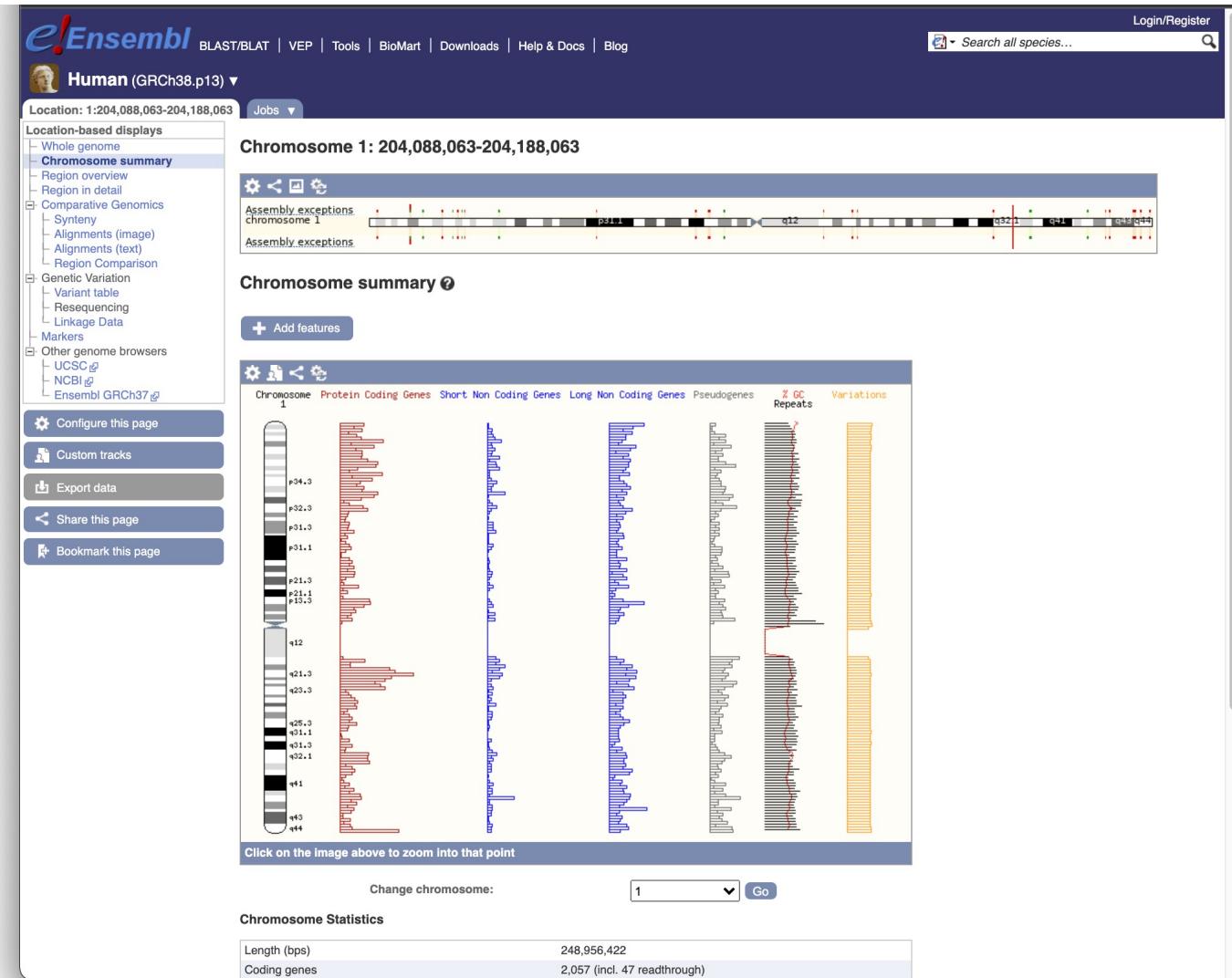
Coding genes	20,465 (incl 653 readthrough)
Non coding genes	24,849
Small non coding genes	4,865
Long non coding genes	17,763 (incl 308 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,217 (incl 6 readthrough)

Login/Register

Search all species...

51

Le site web Ensembl : statistiques par chromosome



Le site web Ensembl : navigateur de génome

The screenshot displays the Ensembl genome browser interface for the Human genome (GRCh38.p13). The main content area shows the "Region in detail" view for Chromosome 1, spanning from 204,088,063 to 204,188,063. The interface includes a navigation bar at the top with links to BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and a Blog. A search bar is also present.

The left sidebar contains a navigation tree under "Region in detail" and a "Configure this page" button, which is highlighted with a red circle. Other options in the sidebar include Custom tracks, Export data, Share this page, and Bookmark this page.

The central panel features three horizontal tracks:

- Chromosome bands:** Shows the physical map of the chromosome.
- Genes:** Displays gene models with their names and Ensembl IDs (e.g., AC114402.2, AC096645.2, AL392146.8, AL592114.13, AL606489.26, AL512306.16) and locations in Mb (e.g., 203.80 Mb to 204.40 Mb).
- Regulatory Build:** Shows various regulatory elements including CTCF, Open Chromatin, and Promoter Flank regions.

Below these tracks, there are legends for Gene Legend (Ensembl protein coding, pseudogene) and Regulation Legend (CTCF, Open Chromatin, Promoter Flank).

The bottom section of the screenshot shows a zoomed-in view of a smaller genomic region from 204,088,063 to 204,188,063, with a focus on the "91 way GERP elements" track, which displays conservation scores across 91 eutherian mammals.

Le site web Ensembl : le gène

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

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

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
 - Ensembl protein families
- Ontologies
 - GO: Cellular component
 - GO: Biological process
 - GO: Molecular function
- Phenotypes
- Genetic Variation
 - Variant table
 - Variant image
 - Structural variants
- Gene expression
- Pathway
- 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, FAD, FAD1, FANCD, FANCD1, XRCC11

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

About this gene This gene has 10 transcripts ([splice variants](#)), [175 orthologues](#) and is associated with [171 phenotypes](#).

Transcripts Hide transcript table

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000380152.8	BRCA2-201	11954	3418aa	Protein coding	CCDS9344	P51587	NM_000059.4	MANE Select v0.95 Ensembl Canonical GEN APPRI P1
ENST00000680887.1	BRCA2-210	11880	3418aa	Protein coding	CCDS9344	-	-	GENCODE basic APPRI
ENST00000544455.6	BRCA2-206	11854	3418aa	Protein coding	CCDS9344	P51587	-	TSL:1 CDS 3' inco
ENST00000530893.6	BRCA2-204	2011	481aa	Protein coding	-	A0A590UJ17	-	TSL:2
ENST00000614259.2	BRCA2-207	11763	2649aa	Nonsense mediated decay	-	-	-	CDS 5' incompl
ENST00000665585.1	BRCA2-208	2598	438aa	Nonsense mediated decay	-	A0A590UJU6	-	TSL:5 CDS 5' inco
ENST00000470094.1	BRCA2-202	842	186aa	Nonsense mediated decay	-	H0YE37	-	CDS 5' incompl
ENST00000666593.1	BRCA2-209	523	58aa	Nonsense mediated decay	-	A0A590UJ24	-	TSL:4 CDS 5' inco
ENST00000528762.1	BRCA2-203	495	64aa	Nonsense mediated decay	-	H0YD86	-	TSL:3
ENST00000533776.1	BRCA2-205	523	No protein	Retained intron	-	-	-	

Summary [↗](#)

Name [BRCA2](#) (HGNC Symbol)

This gene is a member of the Human CCDS set: [CCDS9344.1](#) [↗](#)

CCDS This gene has proteins that correspond to the following UniProtKB identifiers: [P51587](#) [↗](#)

UniProtKB 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.

RefSeq [LRG_293](#) provides a stable genomic reference framework for describing sequence variants for this gene

LRG ENSG00000139618.17

Ensembl version This gene maps to [32,889,223-32,974,405](#) in GRCh37 coordinates.

Other assemblies View this locus in the GRCh37 archive: [ENSG00000139618](#) [↗](#)

Gene type Protein coding

Annotation method Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see [article](#).

Annotation Attributes overlapping locus [Definitions](#) [↗](#)

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Le site web Ensembl : le transcript

Ensembl Human (GRCh38.p13) ▾

Location: 13:32,315,086-32,400,268 Gene: BRCA2 Transcript: BRCA2-201 Jobs ▾

Transcript: ENST00000380152.8 BRCA2-201

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

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

Location: Chromosome 13: 32,315,508-32,400,268 forward strand.

About this transcript: This transcript has 27 exons, is annotated with 58 domains and features, is associated with 35198 variant alleles and maps to 935 oligo probes.

Gene: This transcript is a product of gene ENSG00000139618.17 Hide transcript table

Show/hide columns (1 hidden)

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000380152.8	BRCA2-201	11954	3418aa	Protein coding	CCDS9344	P51587	NM_000059.4	MANE Select v0.95 Ensembl Canonical GEN
ENST00000680887.1	BRCA2-210	11880	3418aa	Protein coding	CCDS9344	-	-	APPRIS P1
ENST00000544456.5	BRCA2-206	11854	3418aa	Protein coding	CCDS9344	P51587	-	GENCODE basic APPRI
ENST00000530893.6	BRCA2-204	2011	481aa	Protein coding	-	A0A590UJ17	-	TSL:1 CDS 3' inco
ENST00000614259.2	BRCA2-207	11763	2649aa	Nonsense mediated decay	-	-	-	TSL:2
ENST00000665585.1	BRCA2-208	2598	438aa	Nonsense mediated decay	-	A0A590UJU6	-	CDS 5' incompl
ENST00000470094.1	BRCA2-202	842	186aa	Nonsense mediated decay	-	H0YE37	-	TSL:5 CDS 5' inco
ENST00000666593.1	BRCA2-209	523	58aa	Nonsense mediated decay	-	A0A590UJ24	-	CDS 5' incompl
ENST00000528762.1	BRCA2-203	495	64aa	Nonsense mediated decay	-	H0YD86	-	TSL:4 CDS 5' inco
ENST00000533776.1	BRCA2-205	523	No protein	Retained intron	-	-	-	TSL:3

Summary

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

Ensembl release 105 - Dec 2021 © EMBL-EBI Permanent link - View in archive site

55

Naviguer dans Ensembl : Partie pratique

Visualiser ses propres données

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 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (*Canis lupus familiaris*) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

[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.

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

[Rapid Release news](#) on our blog

Other news from our blog

- 19 Jan 2022: [Update to the Ensembl COVID-19 Release](#)
- 12 Jan 2022: [Homology data available in Ensembl](#)
- 17 Dec 2021: [146 new insect genomes on Ensembl](#)

Visualiser ses propres données

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.

Navigation dans Ensembl

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

elixir Core Data Resource

LES OUTILS

Les outils

BLAST/BLAT

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 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (*Canis lupus familiaris*) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

[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.

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

[Rapid Release news](#) on our blog

Other news from our blog

- 1 Jan 2022: [Update to the Ensembl COVID-19 resource](#)
- 12 Jan 2022: [Homology data available in Ensembl Rapid Release](#)
- 17 Dec 2021: [146 new insect genomes on Ensembl Rapid Release](#)

Tools

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

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

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

Search

All species for

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

All genomes

-- Select a species --

Pig breeds Pig reference genome and 12 additional breeds

Favourite genomes

Human GRCh38.p13
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

Find a Data Display

Use my own data in Ensembl

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.

ELIXIR Core Data Resource

Ensembl release 105 - Dec 2021 © EMBL-EBI

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

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 ≥ 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



MADTQYILPNDIGVSSLDCREAFLLSPTERLYAYHLSRAAWYGLAVLLQTSPPEAPYIYALLSRLFRAQDP
DQLRQHALAEGLTEEEYQAFLVYAAGVYSNMGNYKSFGDTKFVPNLPKEKLERVILGSEAAQQHPEEVRG
QTCGELMFSLEPRLRHLGLGKEGITTYFSGNCTMEDAKLAQDFLDSQNLSAYNTRLFKEVDGEKGPKYYEVRL
ASVLGSEPSLDSEVTSKLKSYEFRGSPFQVTRGDYAPILOQVVEQLEKAKAYAANSHQGQMLAQYIESFTQG
SIEAHKRGSRFWIQDKGPIVESYIGFIESYRDPFGSRGEFEFGFVAVVNKAMSAKFERLVASAELQLLKELPWP
PTFEKDKEFLPDFTSLDVLTFAGSGIPAGINIPNYDDLQRTEGFKNVSLGNVLAVAYATQREKLTFLEEDDK
DLYILWKGPSFDVQVGLHELLGHGSGKLFVQDEKGAFNFQETVINPETGEQIQSWYRSGETWDSKFSTIAS
SYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIYVNWLNMVRAGLLALEFYTPPEAFNWRQAHMQARFVIL
RVLLEAGEGLVTITPTTGSDGRPDARVRLDRSKIRSVGKPALERFLRRLQVLKSTGDVAGGRALYEGYATVT
DAPPECFLTLRDTVLLRKESRKLIIVQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEELTQLAT
ADARFWKGPSAEPSGOA

new SETUP CONFIG RESULTS DISPLAY refresh Online Help

new SETUP CONFIG RESULTS DISPLAY refresh Online Help

new SETUP CONFIG RESULTS DISPLAY refresh Online Help

Retrieve result for ID:
BLA_IESYdDXDJ 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 ►

► 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

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....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:	Reading	High	Probability	Smallest Sum	
				Frame	Score

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		
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:...:1	+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:...:1	-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:...:1	-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:...:1	-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:...:1	-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 INPETGEOIOSWYRSGETWDWSKFSTIASSYEECRAESVGLYLCILHPOVLEIFGFEGADAE 536
Sbjct: 76090065 INPE EQIQSWYRS +TWDSKFSTI SSYEECRAESVGLYLCILHPOVLE IFGFEGADAE 76089886

Query: 537 DVIVVNWLNMVRAGLIALEFYTPAFAFNWRQAHMQRARVILRLVLEAGEGLVITPTTGSD 596
Sbjct: 76089885 EVISVNWLNMVGAGLLADEFYTPAFAWNWQQAHIRARIVRLVLPAGEGLGTITPTAGSD 76089706

Query: 597 GRPDARVRLDRSKIRSVGKPALERFLRLRLOVLKSTGDVAGGRALYEGYATVDAPPCEL 656
GRPA+A+VRLDRSKI+SVG PALERFLRR STGDVAGG LYE YA V DAPPE FL
Sbjct: 76089705 GRPEAQVRLDRSKIQSVDGNPALERFLRLRW---STGDVAGGWTLERYAAVADAPPGEFL 76089535

Query: 657 TLRDTVLLRKESRKLIQPNTRLEGSDVOLLEYEASAAGLIRSFSERFPEDGPELEEILT 716
TLRD VLLRKES KLIVQPN RLEGSDVOLLEYE SAAGLIRSFS FPEDG ELE+ILT
Sbjct: 76089534 TLRDRVLLRKESWKLIQPNIRLEGSDVOLLEYEVSAAGLIRSFS EHFPEDGLEDEILD 76089355

Query: 717 QLATADAFWKGPKSEAPSGQA 737
QLATADA+F KGPSEAPSGQA
Sbjct: 76089354 QLATADAOF*KGPSEAPSGQA 76089292

Score = 1700 (557.6 bits), Expect = 0., Sum P(6) = 0.
Identities = 212/252 (84%), Positives = 221/252 (87%), Frame = -2

Query: 224 PSLDSEVTSKLKSYSYERGSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIE 283
P L + SKLKS EFRGSFPQVT G+Y PILQKVVEQLEKAK YAANSHQ QMLAQYIE
Sbjct: 76090816 PGLRGD--SKLKS*EFRGSFPQVITWGNYMPILQKVVEQLEKAKTYAANSHQEQLAQYIE 76090643

Query: 284 SFTQGSIEAHKRGSRFWIQDKGPIVESYIGFIESYRDPFGSRGEFEGFVAVVNKAMSAKF 343
SFTQGS EAHK-GSRFWI DKGPIVESYI FI+SYRD FGSRG EGFVAVVNKAMSAKF
Sbjct: 76090642 SFTQGSTEAHKKGSRFWI*DKGPIVESYIEFIQSYRDSFGSRGVCEGFVAVVNKAMSAKF 76090463

Query: 344 ERLVASAEQLLKELPWPPTFEKDKFLTPDFTSLDVLT FAGSGIPAGINIPNYDDLRTQTEG 403
E LV SAEQLLKELPW P FEKDKFLTPDFTS+DVLTFAGSGI AGINI NY+DL+QTEG
Sbjct: 76090462 EWLVVSAEQLLKELPWSPAFEKDKFLTPDFTSVDVLT FAGSGIAAGINISNYNDLKQTEG 76090283

Query: 404 FKNVSLGNVLAVAYATQREKLT FLEEDDKDLYILWKGPSFDVQVGLHELLGHGSGKLFVQ 463
FKNVSLGNVLAV ATQ EKLT LEE DKDLYI+ GPSFDVQVGLHELLG+GSGKL Q
Sbjct: 76090282 FKNVSLGNVLAVV*ATQWEKLTVEESDKDLYIVLMGPSFDVQVGLHELLGYGSGKLFIEQ 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

ⓘ 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 ►**

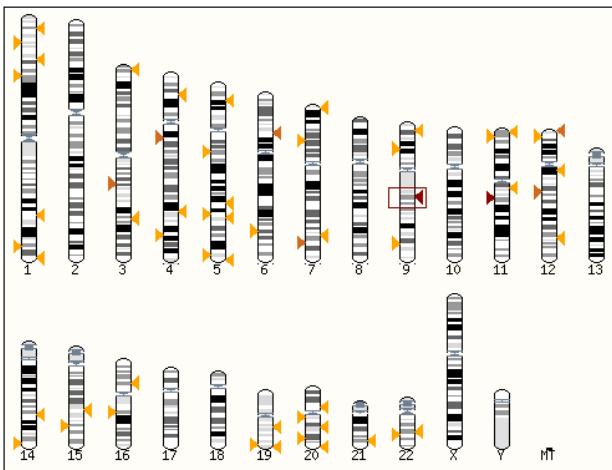
new SETUP CONFIG RESULTS DISPLAY

Displaying unnamed sequence alignments vs Homo_sapiens LATESTGP database

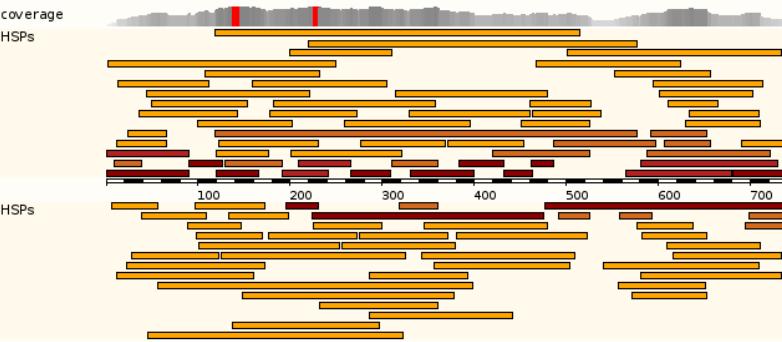
Showing top 100 alignments of 1961, sorted by Raw Score

refresh

Alignment Locations vs. Karyotype (click arrow to hide)



Alignment Locations vs. Query (click arrow to hide)



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

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	_off	_off	_off	_off	_off	_off	_off	_Lrg
Name	Name	Name	Name	Name	Name	Name	Score	<Lrg
Start	Start	Start	Start	Start	Start	Start	E-val	<Score

Links	Query	Chromosome	Stats											
	Start	End	Ori	Name	Start	End	Ori	Score	E-val	%ID	Length			
[A]	[S]	[G]	[C]	477	737	+	Chr:9	76089292	76090065	-	1765	0.	84.29	261
[A]	[S]	[G]	[C]	224	475	+	Chr:9	76090067	76090816	-	1700	0.	84.13	252
[A]	[S]	[G]	[C]	119	577	+	Chr:4	65296878	65298248	+	1542	5.5e-250	49.70	497
[A]	[S]	[G]	[C]	581	729	+	Chr:4	65298493	65298930	+	854	5.5e-250	74.83	151
[A]	[S]	[G]	[C]	1	90	+	Chr:11	66249672	66249941	+	763	3.2e-292	100.00	90
[A]	[S]	[G]	[C]	330	399	+	Chr:11	66260186	66260395	+	552	3.2e-292	95.71	70
[A]	[S]	[G]	[C]	565	679	+	Chr:11	66264763	66265104	+	531	3.2e-292	63.71	124
[A]	[S]	[G]	[C]	1	90	+	Chr:4	65296627	65296899	+	529	5.5e-250	76.09	92
[A]	[S]	[G]	[C]	588	721	+	Chr:11	66271972	66272364	+	487	1.7e-276	55.63	142
[A]	[S]	[G]	[C]	681	737	+	Chr:11	66276549	66276719	+	477	3.2e-292	100.00	57
[A]	[S]	[G]	[C]	120	166	+	Chr:11	66254008	66254148	+	391	1.8e-273	97.87	47
[A]	[S]	[G]	[C]	420	526	+	Chr:11	66262674	66262961	+	384	3.2e-292	53.57	112
[A]	[S]	[G]	[C]	486	597	+	Chr:11	66263006	66263296	+	377	1.7e-276	51.72	116
[A]	[S]	[G]	[C]	266	309	+	Chr:11	66258962	66259093	+	375	3.2e-292	97.73	44
[A]	[S]	[G]	[C]	209	266	+	Chr:11	66258657	66258854	+	370	3.2e-292	75.76	66
[A]	[S]	[G]	[C]	384	432	+	Chr:11	66260513	66260650	+	310	5.1e-263	83.67	49
[A]	[S]	[G]	[C]	90	126	+	Chr:11	66252641	66252751	+	272	3.2e-292	89.19	37
[A]	[S]	[G]	[C]	432	463	+	Chr:11	66261009	66261104	+	270	1.7e-276	96.88	32
[A]	[S]	[G]	[C]	192	242	+	Chr:11	66255385	66255576	+	268	1.3e-266	64.06	64
[A]	[S]	[G]	[C]	196	230	+	Chr:9	76090801	76090905	-	257	0.	88.57	35
[A]	[S]	[G]	[C]	129	191	+	Chr:11	66254628	66254813	+	248	3.2e-292	56.06	66

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

[A]lign

```

Query location      : unnamed    477 to   737 (+)
Database location   : 9          76089292 to 76090065 (-)
Genomic location    : 9          76089292 to 76090065 (-)

Alignment score     : 1765
E-value             : 0
Alignment length    : 261
Expect              : 8.6e-02

```

Query: 477 INFE7GEIQ1QSWR8GETDHSK8FSTI3ASVECRAE3VGLYLCLHQRULEIIFGEGADE 656
INPE EIQ1QSWR3 -T+DHSK8FSTI 3SVECRAE3VGLYLCLHQRULE FGFEGADE
Sbjct: 76090065 INFEMREQPIQ5WYR3MSKWTDHSK8FSTI19VECRAE3VGLYLCLHQRULETFGEGADE 76089886

Query: 537 DIV1YNWHLIMRAGLILFPEFTPEAFNWNQRAQM0ARAFV1LRLVLEAGEGLVIIITGSD 596
VI VNLWHLIMRAGLILFPEFTPEA N+QH++AR L1RLVLEAGEGLVIIITGSD
Sbjct: 76089885 DIV1YNWHLIMRAGLILFPEFTPEASWNRAR1ARIVLRLVLEAGEGLVIIITGSD 76089706

Query: 597 GRPDARVRDLRSKIRSVGKPALERFLRRLQVLKSTGVDVAGRALYEGVA VTDAPPECFL 656
GRP+A+VLRDLRSK1-SVG PALERFLR STGDVAGG LY Y V DAPPE FL
Sbjct: 76089705 GRPEAQVVRDLRSK1Q-SVGPNPALERFLRRCW--- STGDVAGGWTLYER AAVADAPPECFL 76089535

Query: 657 TLRDVLLRKESRK1LVIQPNTRLEGSDVQLLVEYE3AAGLIRLSF---RFEDGFLEELIT 716
TLRD VLLRKES LK1QPN RLEGSDVQLLVEYE SAAGLIRS FEDG ELE+ILT
Sbjct: 76089534 TLRDVLLRKESRK1LVIQPNRILEGSDVQLLVEYE3AAGLIRLS FSEHF DFLGDELEELIT 76089355

Query: 717 QLATADARFWKGPSAEAPSQQA 737
QLATADA+F KGSPAEAPSQQA
Sbjct: 76089354 QLATADAF+KGSPAEAPSQQA 76089282

[S]equence

Query location : unnamed 477 to 737 (+)
Database location : 9 76089292 to 76090065 (-)
Genomic location : 9 76089292 to 76090065 (-)

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

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

3

MADTYQILPNIDGVSSLDCREAFLRLSPTERLYAHLRSAWYGGLAVLLQTSPPEAFYIY
ALLSRLSFRAQPPDQLRQHALAEGLTIEEYQAFLLVVAAGYVSNMNGNYSKSGDFTKFPVNF
EKLERVLGSEAAQHQPVEFVRGLTGCTGELMSLFEPRPLRLHGKGKEGTEFVSGNCNTMED
AKLAQDFLDSQNLSSAYNTRLFKEVDEGGPKPYEVRLASVLGSEPSLDSEVTSLKLSYEFFF
GSPFPTVRGDYAPILQVKVEQLEKAKAYAANSHQMLAQYIESFTQGSAEIAHKRSLPWF
IDQKPGPIEVSYIHFIESYDGFRRGFRGEFFGVAVWQNKAMRERLVAQAEKLKRPWL
PTFEKKDKFLTBDFTSLDVLTFAGSGIFAGINIPNYNDLDRQTEGFKNVSLGNVLAVAYATQ
REKLTLFEEEDKDLYILWKGSFSDVQVGLHELLHGGSGLKFVQDCEKGFANFDQETV1NP
TGEQIQISWYRSGETWDKSFTIASSYEECAESVGLYLCLHPQVLEI1FGEGADADEV
VNWLNMVRAGLLEAFYTFEAFNRFQRQAHMQARFVILVLEAGEGVL1TPTGSDGRPF
ARVRDLRSKRISKVGPALERFLRLQVLKSTGDVAGGRALYEYGVATVTDAPPECFLITR
TVLRLKESRKLI1VQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEEILTQLAT
ADARFWKGPSSEAPSQGA

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

[G]Sequence

[C]ontig view (?)

```

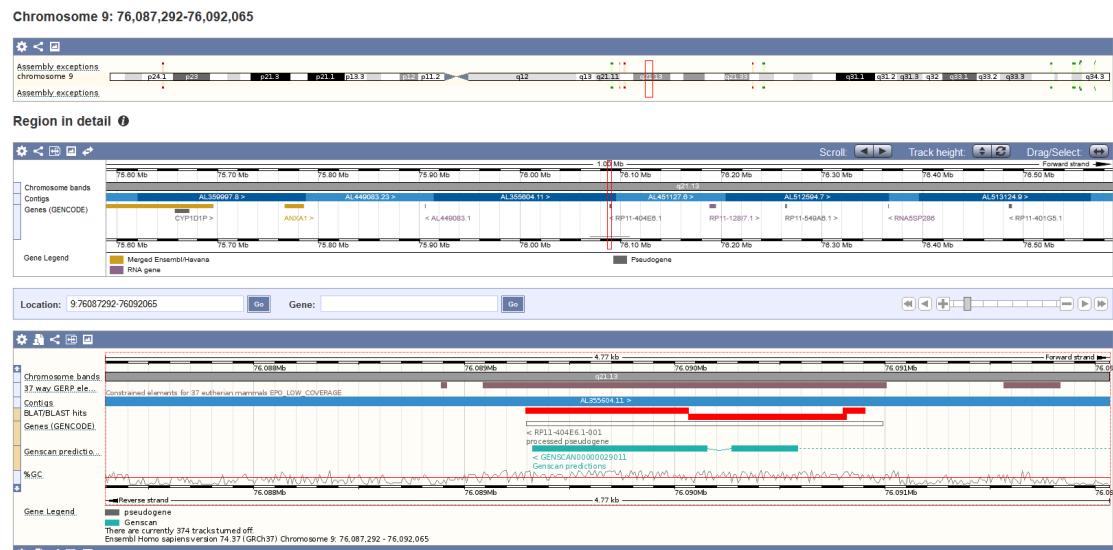
Query location      : unnamed        477 to      737 (+)
Database location   : 9      76059292 to 76090065 (-)
Genomic location    : 9      76059292 to 76090065 (-)

Alignment score     : 1265

```

Alignment score : 1768
E-value : 0.
Alignment length : 261
Percentage identity: 84.2

5' Flanking sequence	<input type="text" value="300"/> (bp)
3' Flanking sequence	<input type="text" value="300"/> (bp)
Coordinate system	Chromosome <input type="button" value="▼"/>
Orientation	Forward relative to selected alignment <input type="button" value="▼"/>
Alignment markup	All alignments <input type="button" value="▼"/> Both orientations <input type="button" value="▼"/>
Feature markup	Ensembl exons <input type="button" value="▼"/> Both orientations <input type="button" value="▼"/>
Line numbering	No numbers <input type="button" value="▼"/>



Les outils

Annotation de variants

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

Search all species...

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 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (*Canis lupus familiaris*) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

[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.

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

[Rapid Release news](#) on our blog

Other news from our blog

- 1 Jan 2022: [Update to the Ensembl COVID-19 resource](#)
- 12 Jan 2022: [Homology data available in Ensembl Rapid Release](#)
- 17 Dec 2021: [146 new insect genomes on Ensembl Rapid Release](#)

Tools

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

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

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

Search

All species for Go

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

All genomes

-- Select a species --

Pig breeds Pig reference genome and 12 additional breeds

Favourite genomes

Human GRC38.p13
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

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.

elixir Core Data Resource

Ensembl release 105 - Dec 2021 © EMBL-EBI

Outils : visualisation de ses données

Permanent link - View in archive site

Variant Effect Predictor

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

Login/Register Search all species...

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

In this section

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

On this page

- VEP interfaces
- Publication
- VEP related tools

Search documentation... Go

Ensembl Variant Effect Predictor (VEP)

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

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

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

★ [What's new in release 105?](#)

VEP interfaces

Web interface



- Point-and-click interface
- Suits smaller volumes of data

[Documentation](#)



Command line tool



- More options and flexibility
- For large volumes of data

[Documentation](#)

[Clone from GitHub](#) [Download \(zip\)](#) [Pull Docker image from DockerHub](#)

REST API



- Language-independent API
- Simple URL-based queries

[Documentation](#)

[VEP REST API](#)

Publication

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

[Cite us](#)

Variant Effect Predictor

The screenshot shows the Variant Effect Predictor (VEP) tool on the Ensembl website. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar for species and a login/register link are also present. The left sidebar has a 'Web Tools' section with 'Variant Effect Predictor' selected, along with other options like BLAST/BLAT, Linkage Disequilibrium Calculator, Variant Recoder, File Chameleon, Assembly Converter, ID History Converter, VCF to PED Converter, Data Slicer, and Post-GWAS. Below this are buttons for 'Configure this page', 'Custom tracks', 'Export data', 'Share this page', and 'Bookmark this page'. The main content area is titled 'Variant Effect Predictor' and contains fields for 'Species' (set to Homo_sapiens), 'Name for this job (optional)', and a large text area for 'Input data' with a placeholder 'Either paste data:'. Below this are sections for 'Transcript database to use:' (radio buttons for Ensembl/GENCODE transcripts, Ensembl/GENCODE basic transcripts, RefSeq transcripts, and Ensembl/GENCODE and RefSeq transcripts) and 'Additional configurations:' (checkboxes for 'Identifiers', 'Variants and frequency data', and 'Find co-located known variants: Yes'). At the bottom, there are checkboxes for 'Variant synonyms' and 'Frequency data for co-located variants: 1000 Genomes global minor allele frequency'.

Variant Effect Predictor

New job Clear form

Species: Homo_sapiens

Name for this job (optional):

Input data:

Either paste data:

Examples: Ensembl default, VCF, Variant identifiers, HGVS notations, SPDI

Or upload file: Choisir un fichier Aucun fichier choisi

Or provide file URL:

Transcript database to use:

Additional configurations:

Identifiers Additional identifiers for genes, transcripts and variants

Variants and frequency data Co-located variants and frequency data

Find co-located known variants: Yes

Variant synonyms:

Frequency data for co-located variants: 1000 Genomes global minor allele frequency

Variant Effect Predictor

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

Login/Register

Search all species...

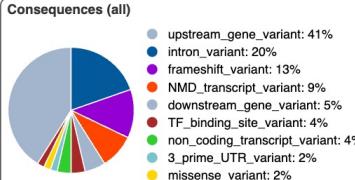
Variant Effect Predictor results 

Job details 

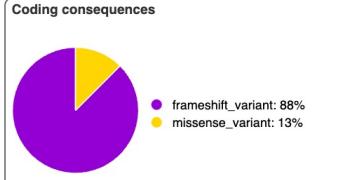
Summary statistics 

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

Consequences (all)



Coding consequences



Results preview

Navigation (per variant)  Filters  Download 

Page:  1 of 1  | Show: 1 All variants  is  defined 

All: [VCF VEP TXT](#)
BioMart: Variants [Genes](#)

New job 

Show/hide columns (13 hidden) 

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Scroll to see more columns »
1_65568_A/C 	1_65568_65568	C	downstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000492842.2	transcribed_unprocessed_pseudo
1_65568_A/C 	1_65568_65568	C	missense_variant	OR4F5	ENSG00000186092	Transcript	ENST00000641515.2	protein_coding
1_65568_A/C 	1_65568_65568	C	downstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000642116.1	processed_transcript
2_265023_C/T 	2_265023_265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272065.10	protein_coding
2_265023_C/T 	2_265023_265023	T	intron_variant	ACP1	ENSG00000143727	Transcript	ENST00000272067.10	protein_coding
2_265023_C/T 	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000356150.10	protein_coding
2_265023_C/T 	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000402632.5	protein_coding
2_265023_C/T 	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403657.5	protein_coding
2_265023_C/T 	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403658.5	protein_coding
2_265023_C/T 	2_265023_265023	T	upstream_gene_variant	SH3YL1	ENSG00000035115	Transcript	ENST00000403712.6	protein_coding

Outils de récupération de données

Screenshot of the Ensembl website (<https://www.ensembl.org>) showing data download options.

The page title is "Accessing Ensembl Data". The "Downloads" menu item is highlighted with a red box.

Small quantities of data: Many pages offer an "Export" option for small amounts of data, e.g. a single gene sequence. An arrow points from the "Export data" button to a computer monitor displaying a sequence: CAGATCAT AAATGTTT AAAGAGCA CTGTCATGC ATAAAAGAA AGTGATACT.

Fast programmatic access: For fast access in any programming language, use the REST server. An arrow points from a computer monitor to a database icon.

Complete datasets and databases: All datasets, e.g. all genes for a species, are available to download in various formats from the [FTP site](#). An orange arrow points from the "FTP site" link to a database icon.

Complex cross-database queries: More complex datasets can be retrieved using the [BioMart](#) data-mining tool. An arrow points from the BioMart link to a funnel icon.

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

Ensembl release 105 - Dec 2021 © EMBL-EBI Permanent link

About Us

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

Get help

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

Our sister sites

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

Follow us

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

BioMart

Le projet BioMart

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

Les “Marts”

The image displays three separate web interfaces for biological data marts:

- Ensembl BioMart:** A screenshot showing the interface for selecting columns from a dataset (Homo sapiens genes, GRCh37.p13). It includes a sidebar for filters and attributes, and a top navigation bar with links like BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors.
- UniProt BioMart:** A screenshot showing the interface for choosing a database. It features the UniProt logo and a search bar at the top, with a sidebar for dataset selection.
- ICGC Data Portal:** A screenshot of the main portal interface. It features a logo with a red and blue circular emblem, the text "ICGC Data Portal", and three buttons: "Cancer Projects" (orange), "Advanced Search" (blue), and "Data Repository" (blue). Below these is a search bar containing the placeholder text "eg. BRAF, KRAS G12D, DO35108, MU7870, TCGA-06-5858".

Accéder aux données d'Ensembl

Site web

The screenshot shows the Ensembl homepage with a search bar at the top. Below it, there's a section for browsing genomes, a 'What's New' section for release 83, and a 'Tweets' section from the Ensembl Twitter account (@ensembl). The page also features links to BLAST, BioMart, and other tools.

Outil de fouille: BioMart

The screenshot shows the BioMart interface. At the top, there's a search bar and a 'Dataset' dropdown menu 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 side of the header is a 'Login/Register' button and a search bar labeled 'Search all species...'. Below the header, there are three main sections: 'Tools' (with a 'BioMart' link), 'BLAST/BLAT >' (with a 'Variant Effect Predictor >' link), and 'Variant Effect Predictor >'. The 'BioMart' link in the 'Tools' section is highlighted with a red box. The 'BioMart' section in the main content area is also highlighted with a red box. It contains a search bar with dropdown menus for 'All species' and 'for', and a 'Go' button. Below the search bar is a note: 'e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease'. The main content area also includes sections for 'All genomes' (with a dropdown menu 'Select a species') and 'Favourite genomes' (listing Human (GRCh38.p13), Mouse (GRCm39), and Zebrafish (GRCz11)). To the right, there's a 'Ensembl Rapid Release' section with news about gene and protein annotation updates every two weeks, a note about existing species, and a 'Go' button. At the bottom, there's a 'Other news from our blog' section with links to COVID-19 resource updates and homology data availability.

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

BioMart/Ensembl

The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p13)), Filters (None selected), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version). The main area shows a dropdown menu for 'Dataset' set to 'Ensembl Genes 105'. Below it, a list shows 'Human genes (GRCh38.p13)' as the selected dataset. A large orange bracket on the right side of the interface spans from the 'Dataset' dropdown to the list of datasets, indicating the process of selecting the database. Another orange bracket at the bottom right points to the 'Human genes (GRCh38.p13)' entry, specifically highlighting the genome selection. A note at the bottom states: 'In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.' and 'Note that queries that run for longer than 6 hours will be terminated even when submitted this way. If this happens please reformat your query or contact us for details on how to approach this.'

Selection de la Base de donnée :

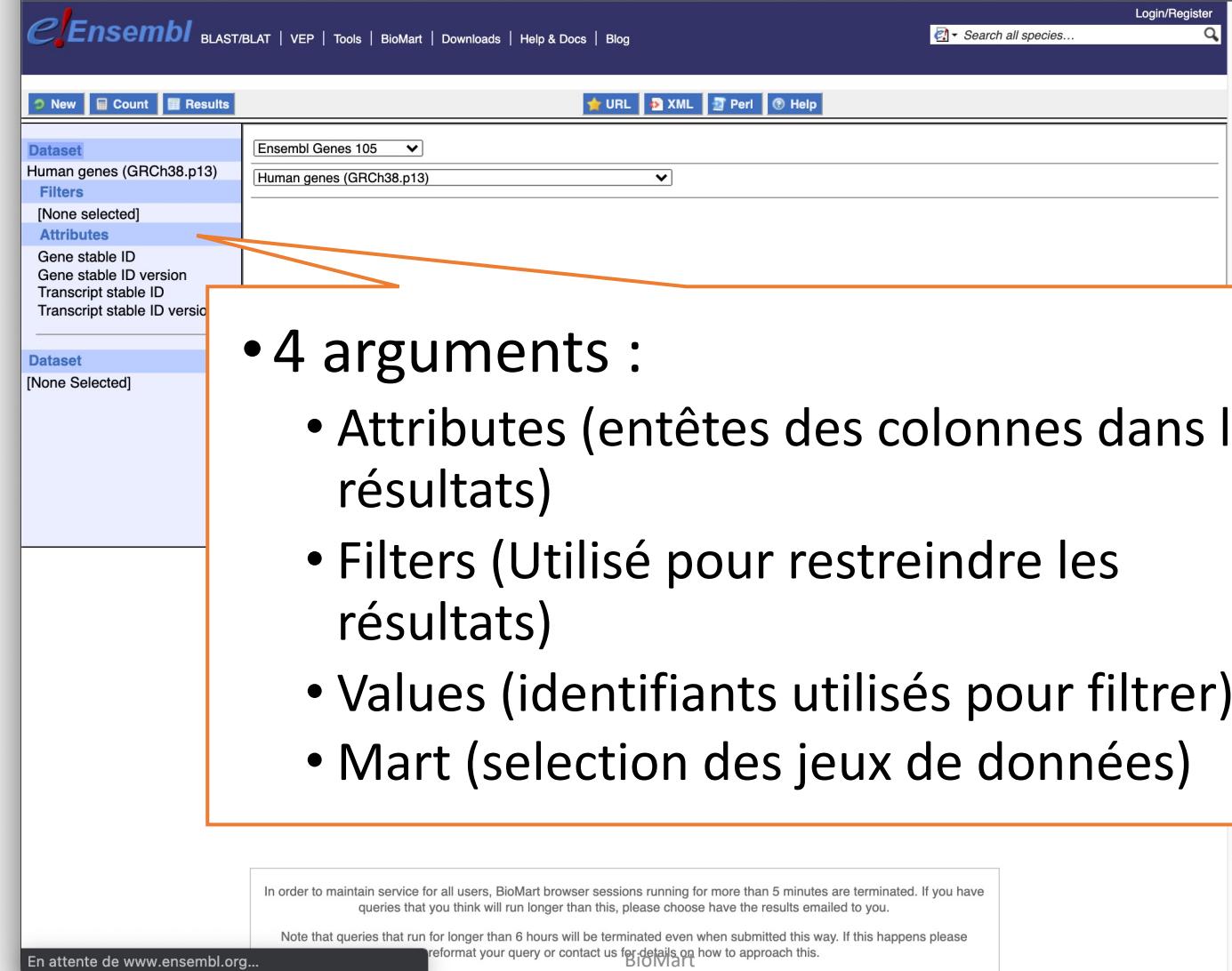
- Genes
- Variation
- Regulation
- Mouse strain

Sélection du jeu de données (génome)

In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.
Note that queries that run for longer than 6 hours will be terminated even when submitted this way. If this happens please reformat your query or contact us for details on how to approach this.

En attente de www.ensembl.org...

BioMart/Ensembl



The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p13)), Filters ([None selected]), and Attributes (Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version). The main area shows a dropdown for 'Dataset' set to 'Ensembl Genes 105' and another dropdown for 'Human genes (GRCh38.p13)'. A large orange box highlights the 'Attributes' section in the sidebar and the 'Values' dropdown in the main area. Below the sidebar, a message states: 'In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.' At the bottom, it says 'En attente de www.ensembl.org...' and 'BioMart'.

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

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