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

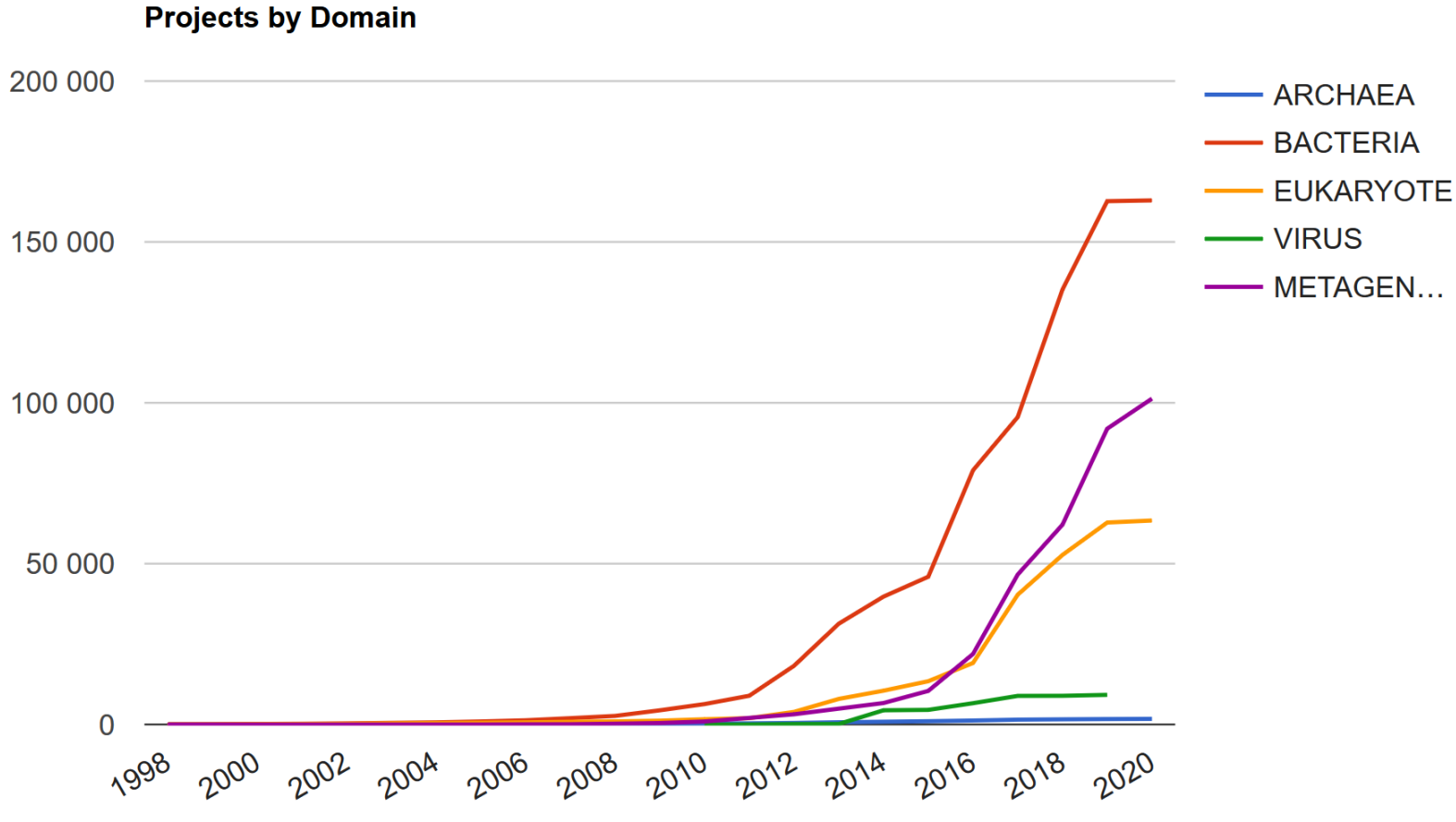
- 2002 Séquence préliminaire du génome de la souris (Waterston et al., 2002)
- 2004 ENCODE, Identification de tous les éléments fonctionnels du génome humain
- 2005 Roche 454: Séquenceur auto. haut-débit de 2ème génération par pyroséquençage : GS20
- 2007 Illumina/Solexa NGS de 2ème génération par synthèse 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

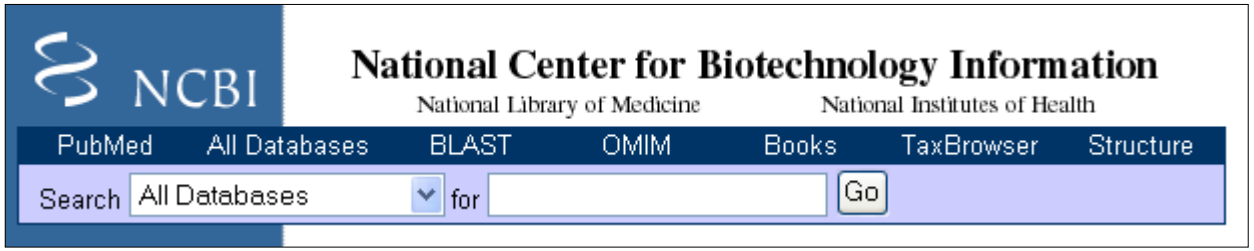


Centres de bioinformatique

- EBI (European Bioinformatics Institute)

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

- NCBI (National Center for Biotechnology Information)



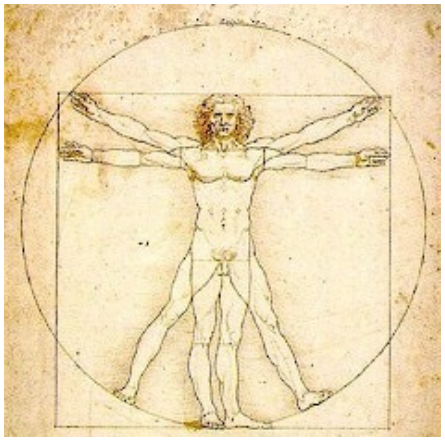
http://www.ncbi.nlm.nih.gov/

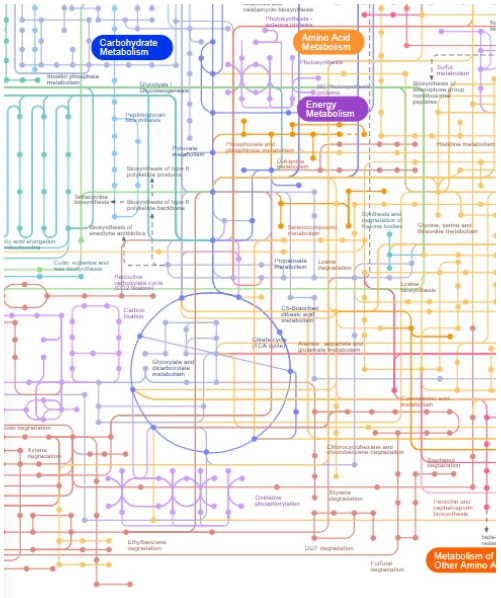
Banques de données en biologie moléculaire

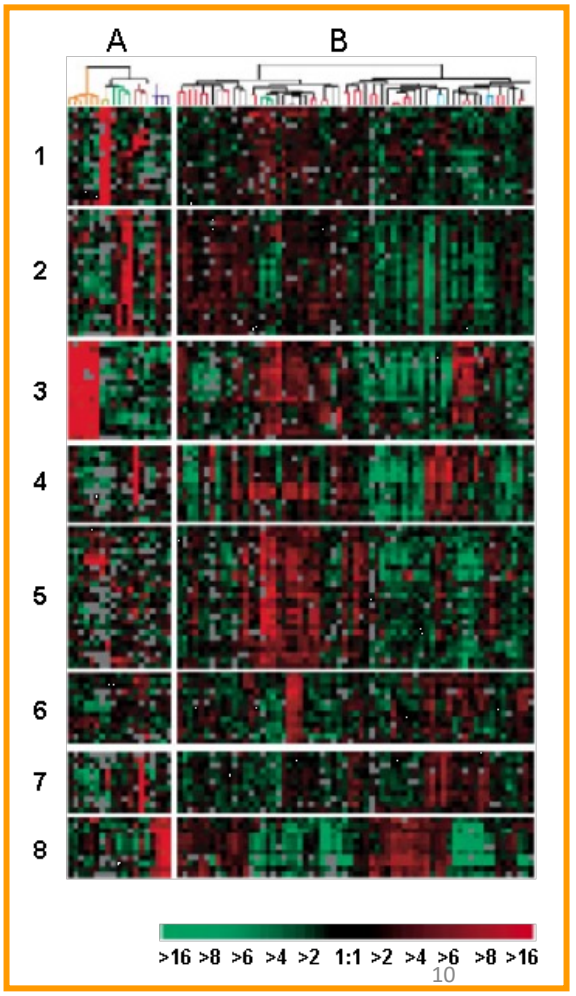
- Rôles des banques
  - Stockage
  - Diffusion (ftp, web…)
  - Organisation et standardisation des données
  - Connectivité avec autres banques
  - Actualisation

Multiplicité des banques

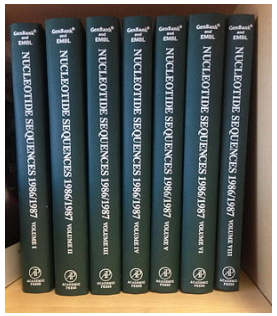
MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVGALELAGGPGA







Banques de séquences nucléiques généralistes



GenBank

NCBI

EMBL

ENA
European Nucleotide Archive

DNA databank of Japan

DDBJ

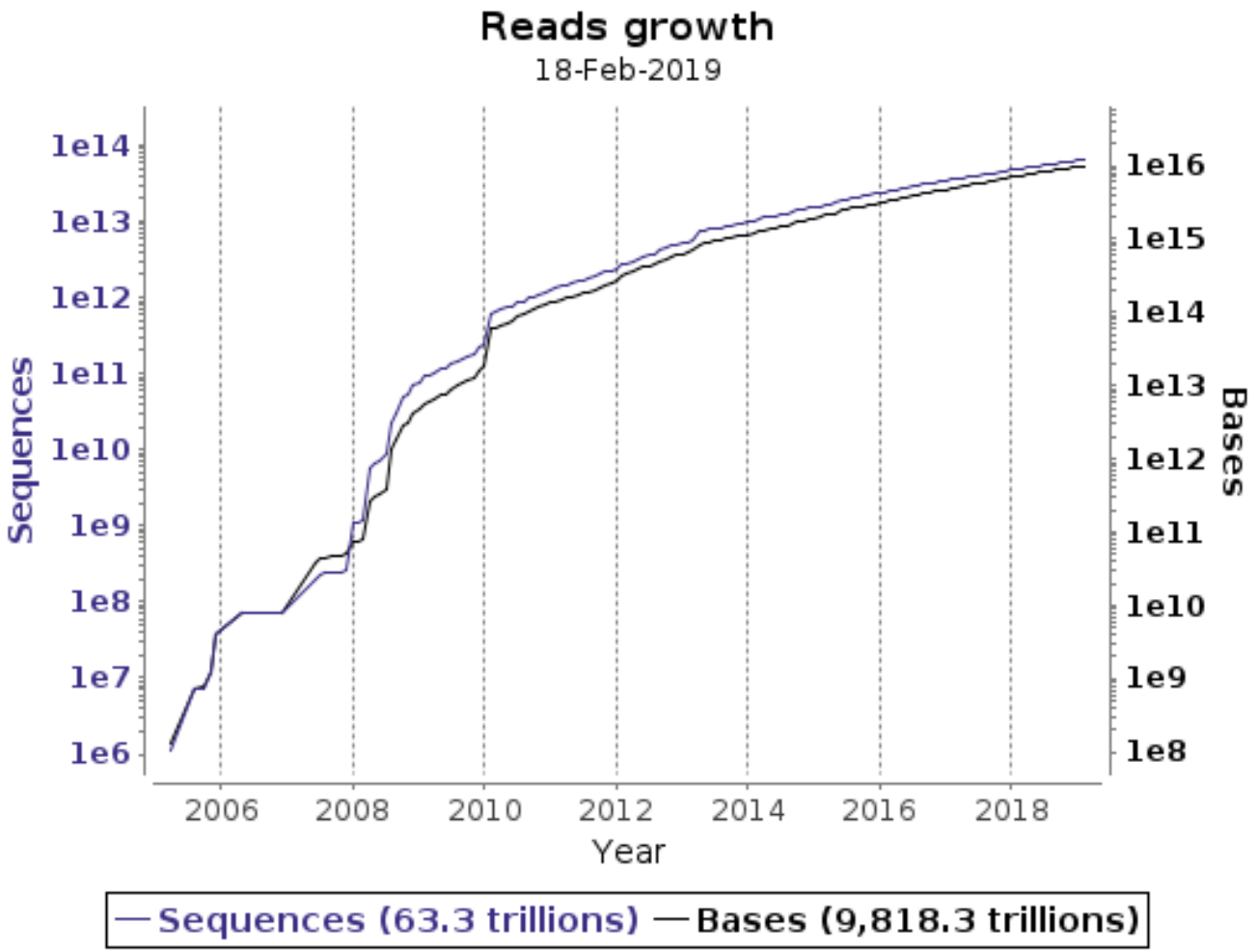
- 3 banques
- Échanges quotidiens des séquences collectées
- Effort d'unification=> format
  - accord entre GenBank et EMBL en 1986
  - accord entre GenBank/EMBL et DDBJ in 1987

INSDC International Nucleotide Sequence Database Collaboration

Banques de séquences

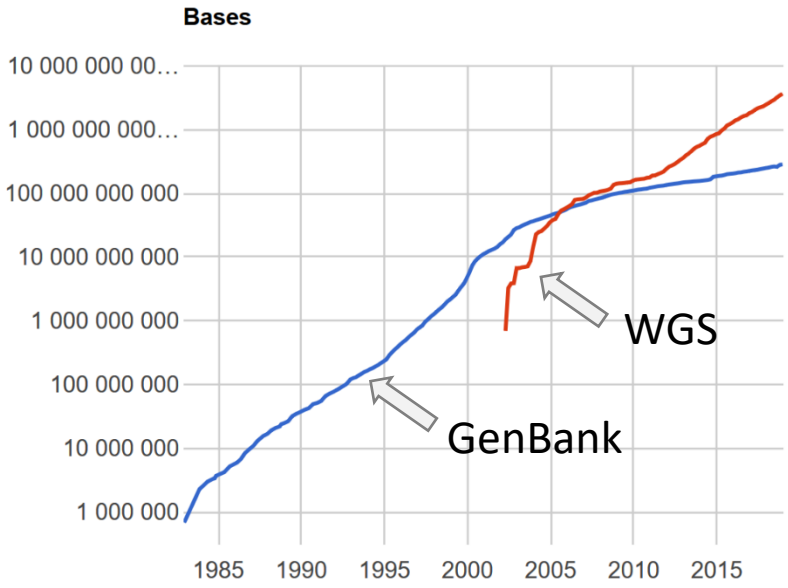
- Des banques incontournables :
  - dépôt obligatoire dans une des 3 banques avant publication
  - unique moyen d'accès aux séquences
- Alimentation :
  - soumission directe par la communauté scientifique (associée ou non à une publication)
  - dépôts de brevets
- Conséquences
  - banques exhaustives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

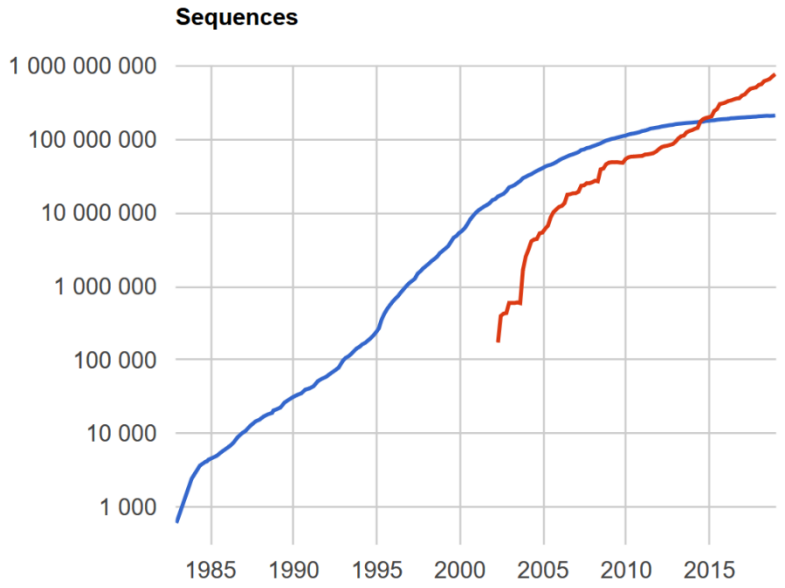
Evolution de la banque EMBL



http://www3.ebi.ac.uk/Services/DBStats/

Evolution de la banque GenBank





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

https://www.ncbi.nlm.nih.gov/genbank/statistics/

Banques de séquences protéiques généralistes

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

http://www.uniprot.org/

03/2018	**01/2019**	**02/2020**		10/2016	02/2018	**02/2020**
106,245,682	**130,366,644**	**167,278,920**	TrEMBL	68,493,254	109,414,541	**179,812,129**

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

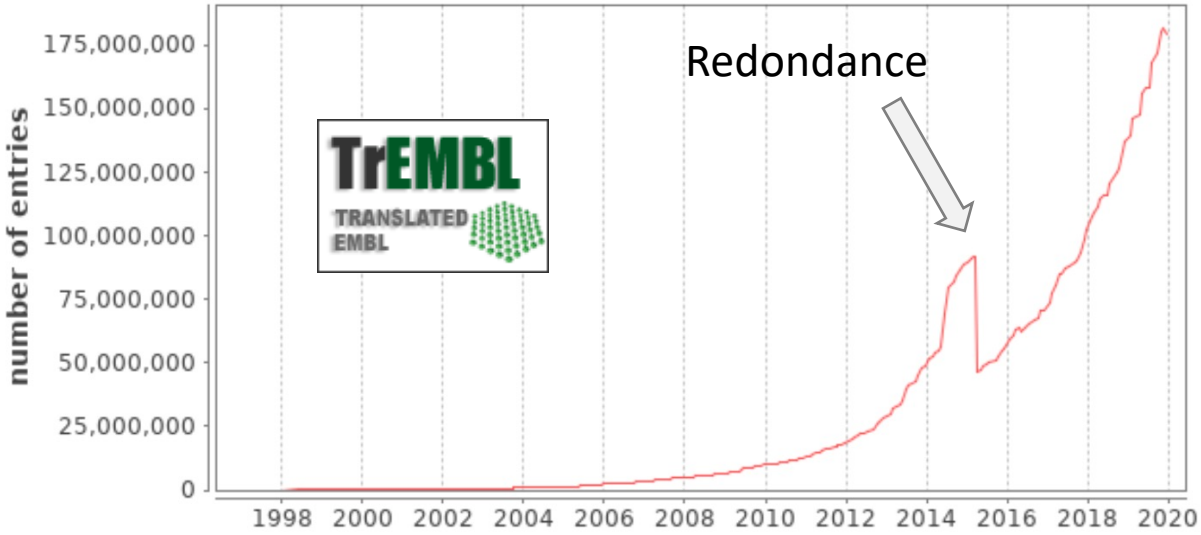
TrEMBL:
179,250,561 entrées

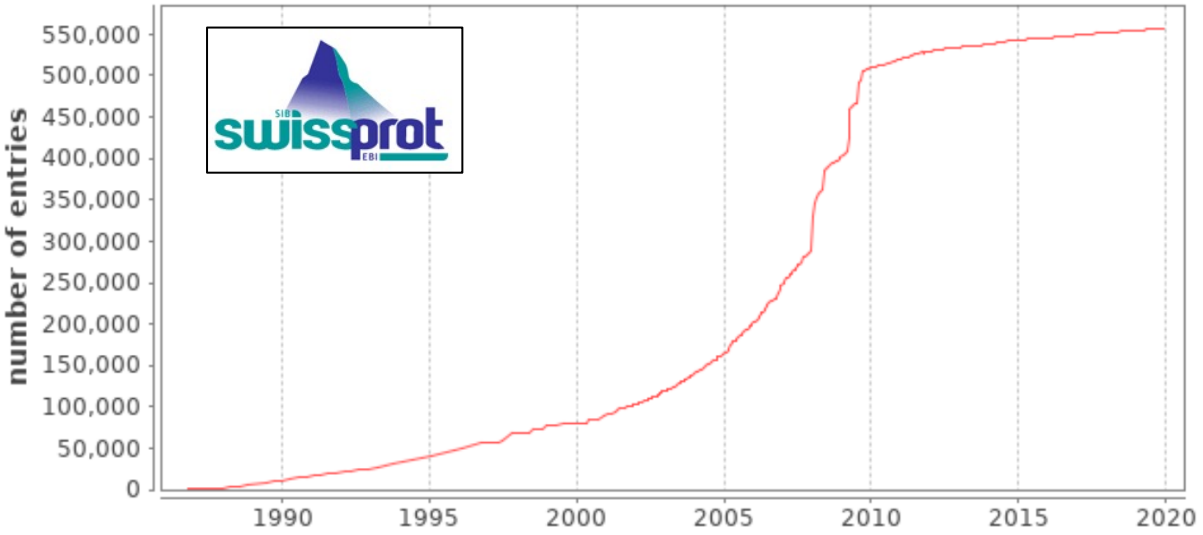
Swiss-Prot:
561,568 entrées

- 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 dont **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 silmilarity** » , « **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 »

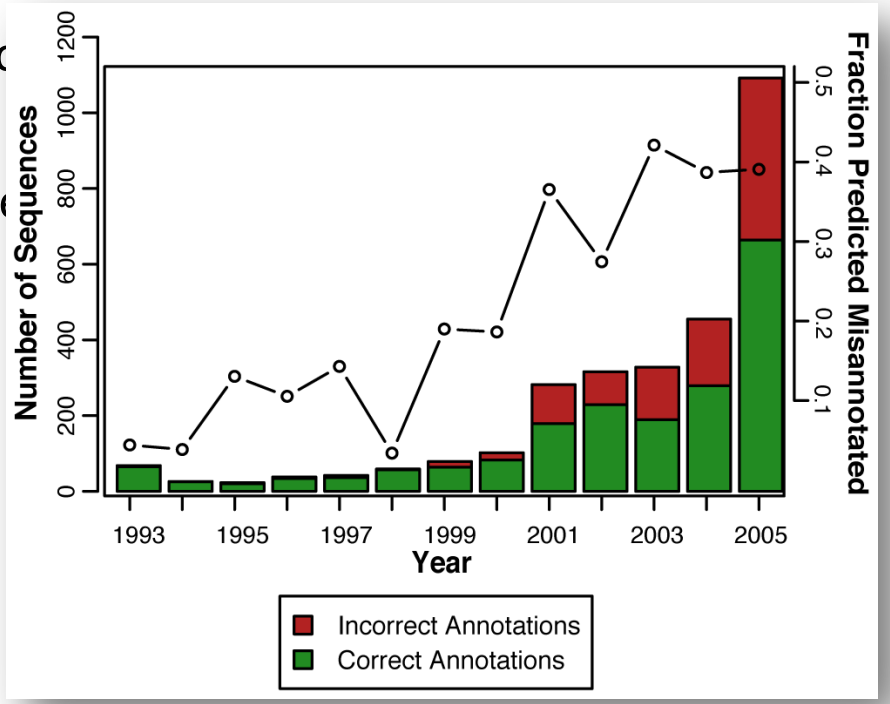
Exemple de l'importance de l'annotation

Exemple 1: DUF domain = **D**omain of **U**nknown **F**unction

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

Exemple 3: Analyse de 37 familles de pro

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



Schnoes *et al* PLOS Genetics 2009

Evolution des bases de données protéiques

Bases de données majeures

collecte des données individuelles et collectives

Attention à la qualité de ces données

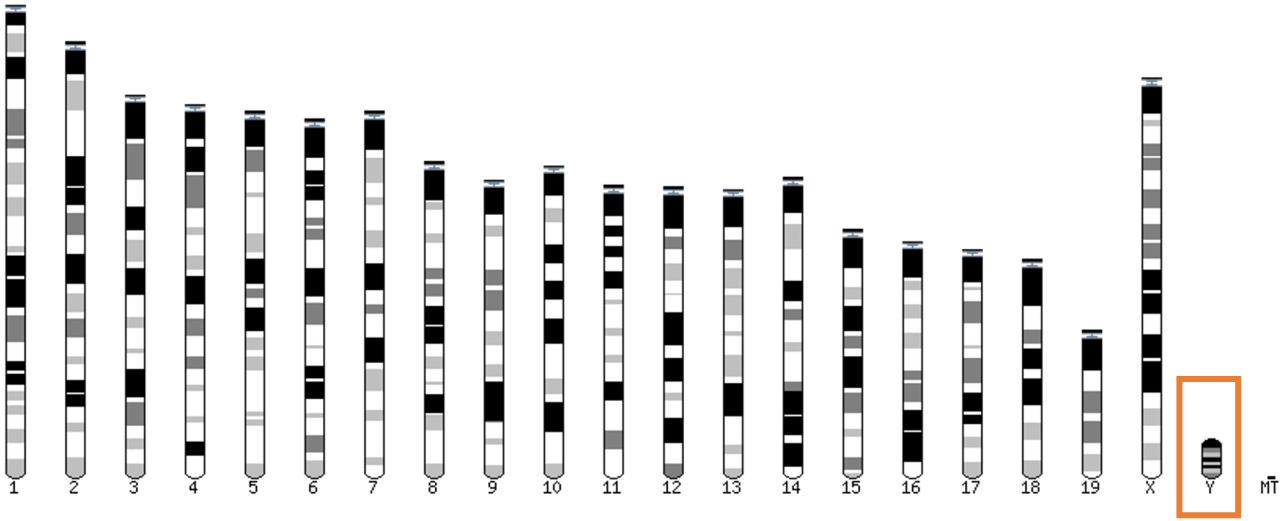
bases avec les Raw data vs Annotation

Ces données seront agrégées sur le génome humain

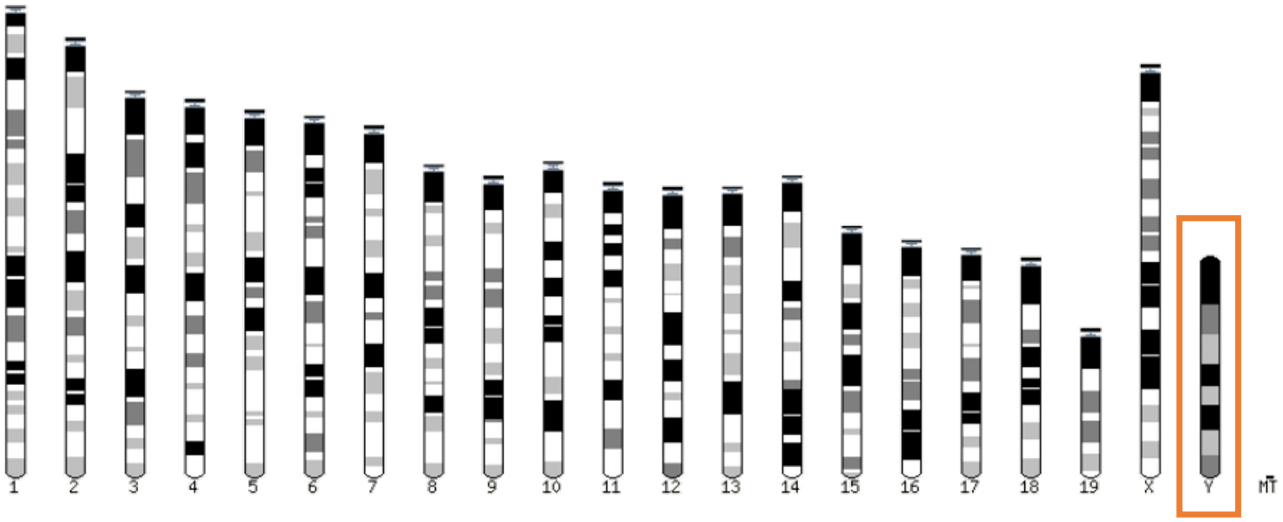
Genome browsers

Genome builds

mm9



mm10



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)

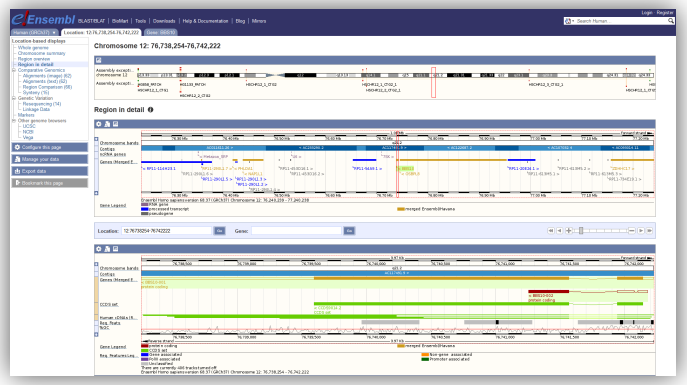
Source: https://genome.ucsc.edu/FAQ/FAQreleases.html

Genome Browsers – L'outil de référence

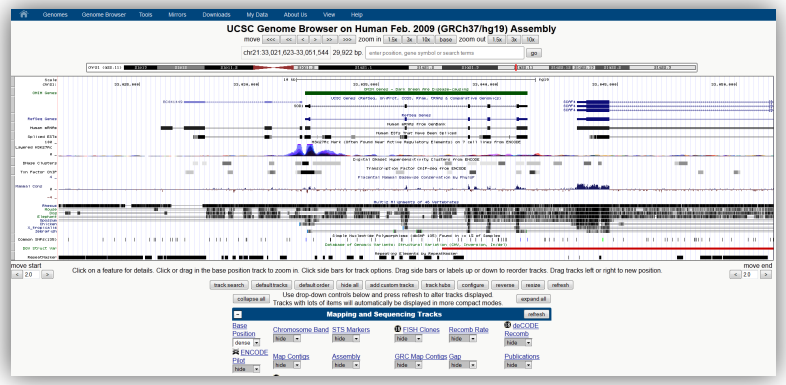
- Elément de référence absolue le **génome**
- Agrégateur et générateur d'informations/annotations
  - Prédictions de gènes
  - Protéines
  - Données d'expression
  - Variations
- Synthèse rapide et visuelle de données primordiales

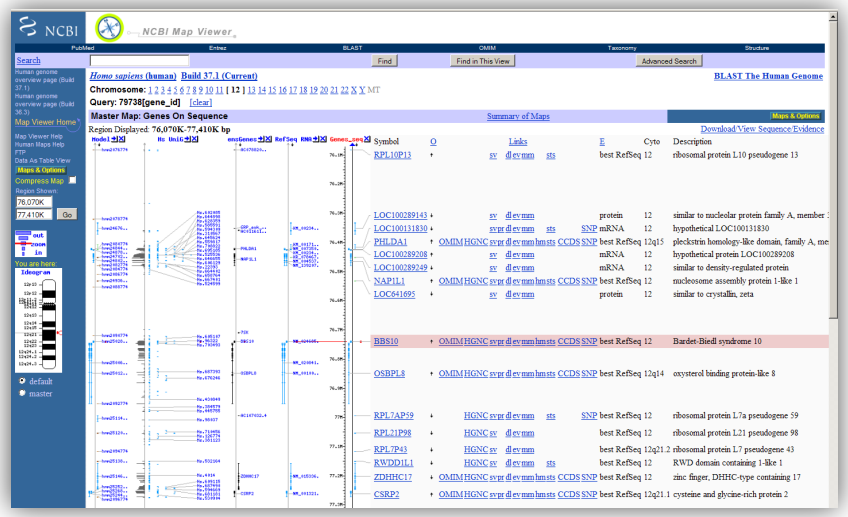
Il y a Genome Browsers…

EBI - Ensembl



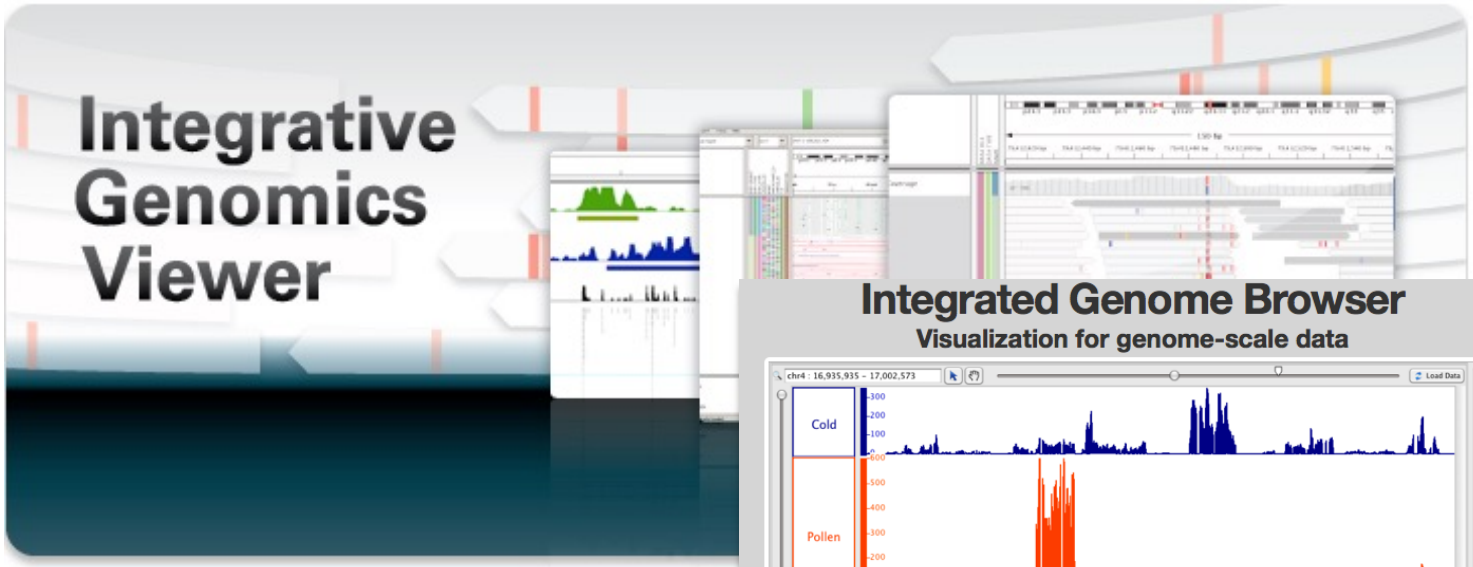
UCSC – Genome Browser





NCBI – Map Viewer

Et Genome browsers



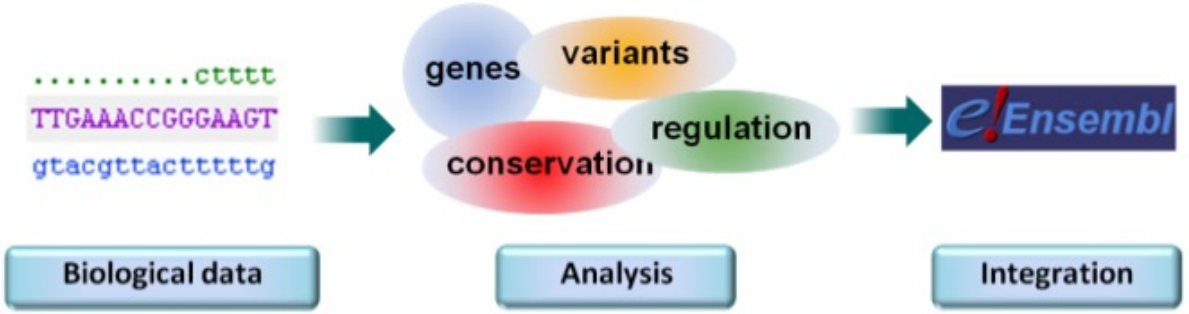
Ensembl

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

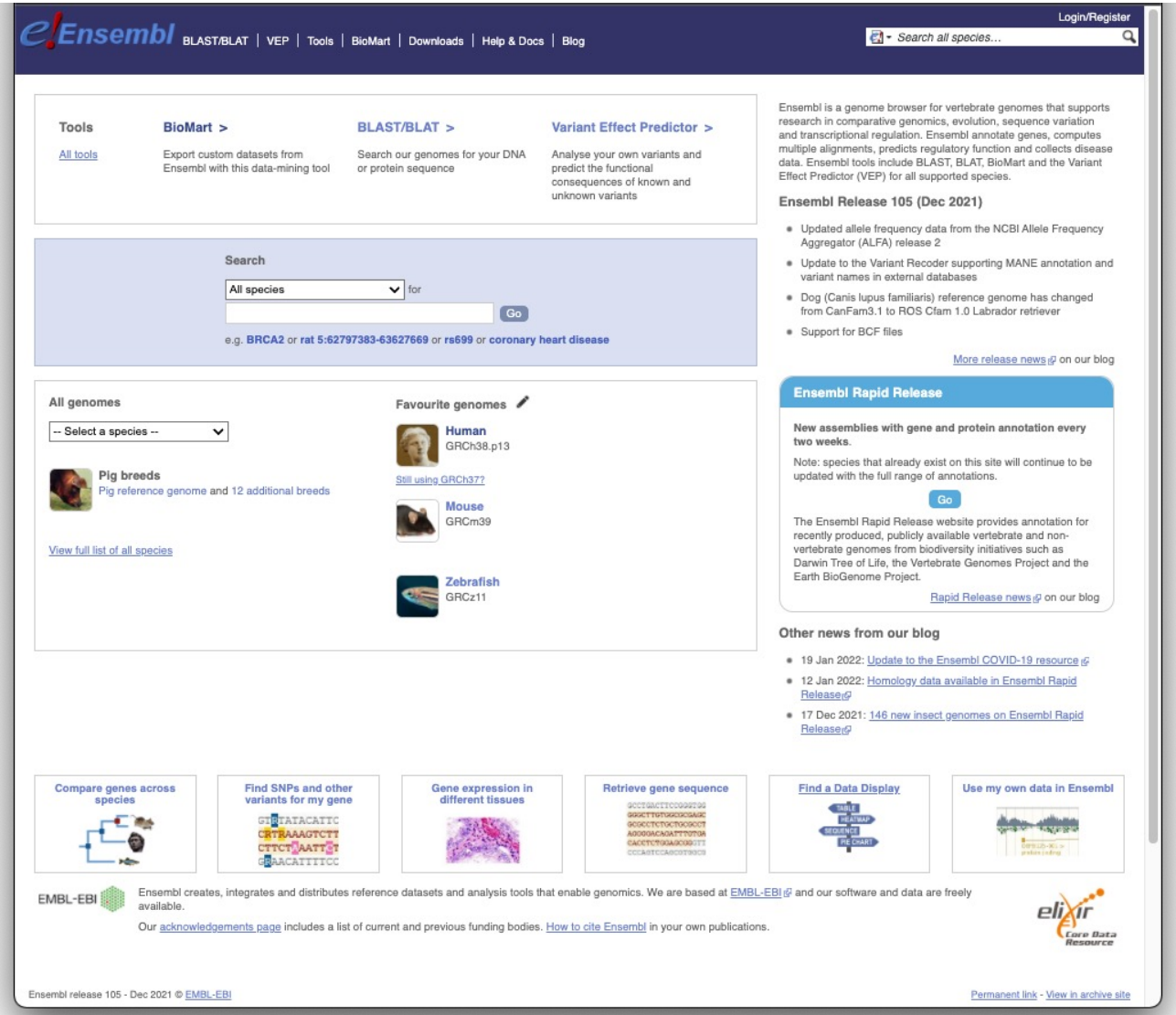


- Lancement du site web en juillet 2000 (au début il n'y avait que le génome humain)

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



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

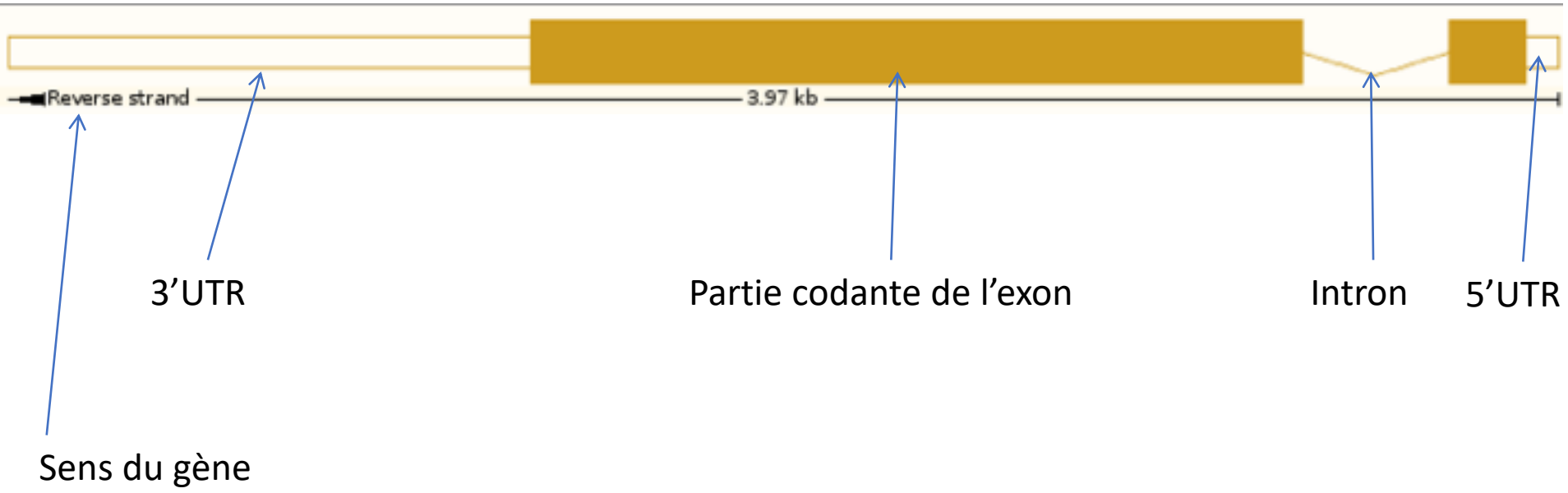
- Les transcrits d'Ensembl sont basés sur les bases de données suivantes :
  - Uniprot/Swiss-Prot (*curation* manuelle)
  - Uniprot/TrEMBL
  - NCBI refSeq (*curation* manuelle)

Protéine/ ARNm Assemblage de séquence Gène Ensembl

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



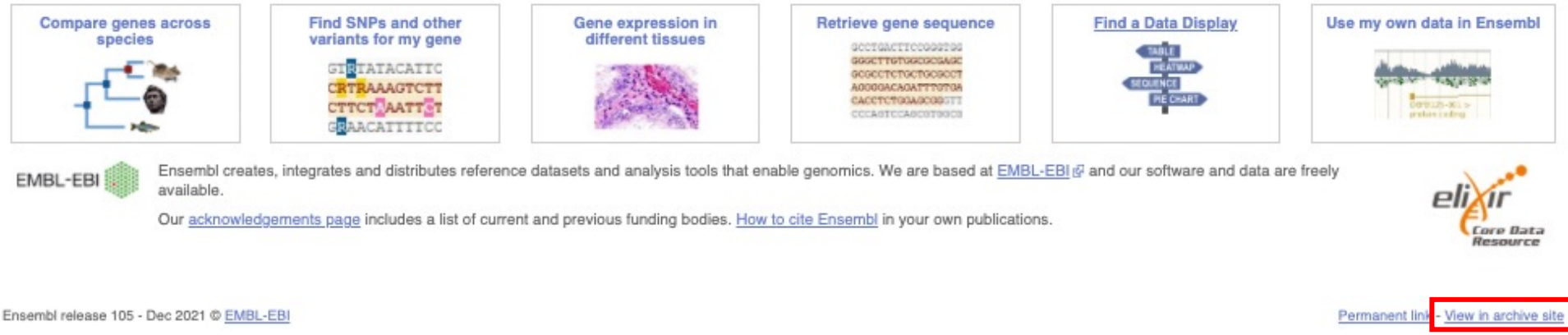
Identifiants Ensembl

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

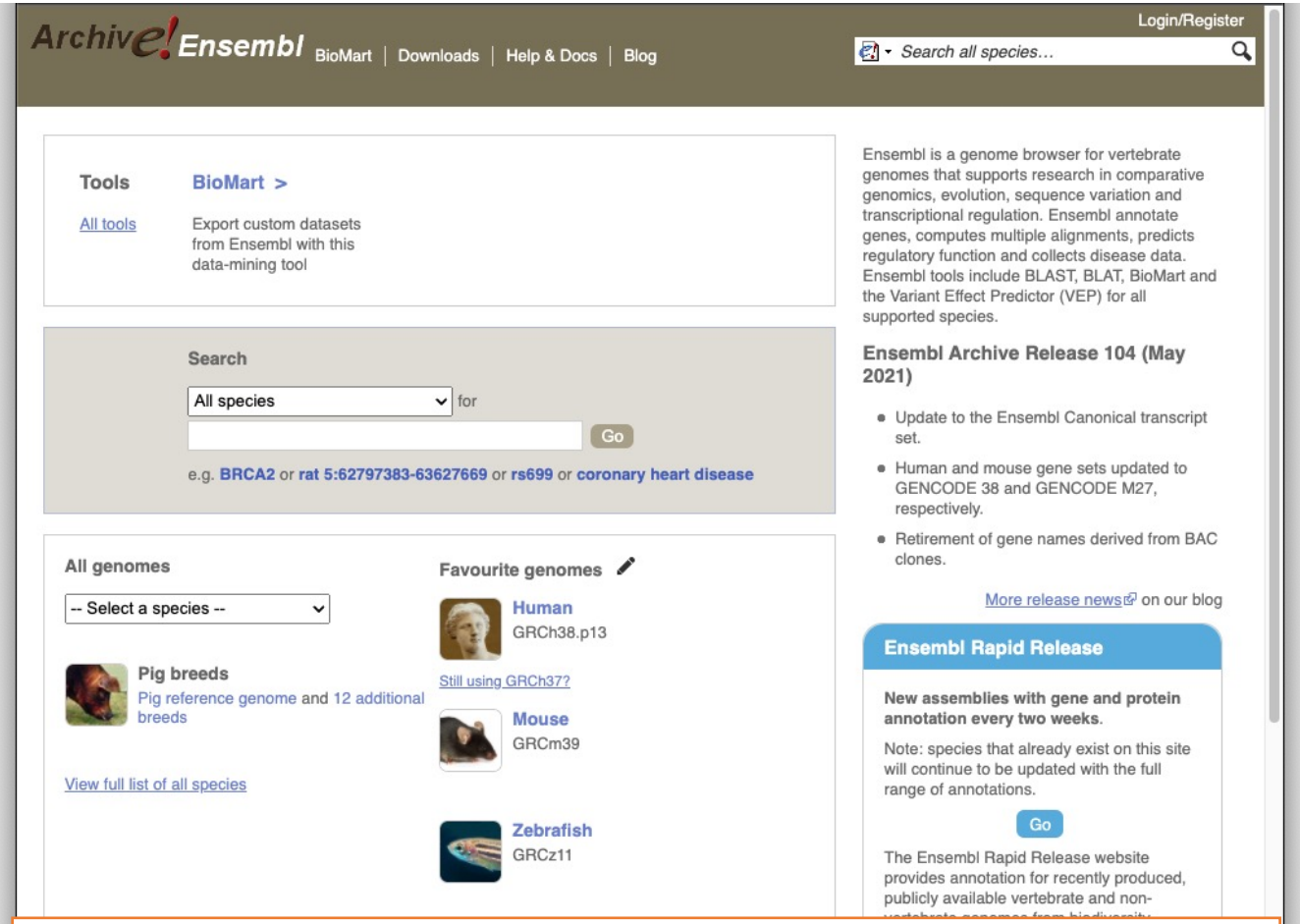


- Lien vers une version particulière d'Ensembl : http://Dec2021.archive.ensembl.org/index.html

Ensembl : Archives

http://www.ensembl.org/info/website/archives/index.html

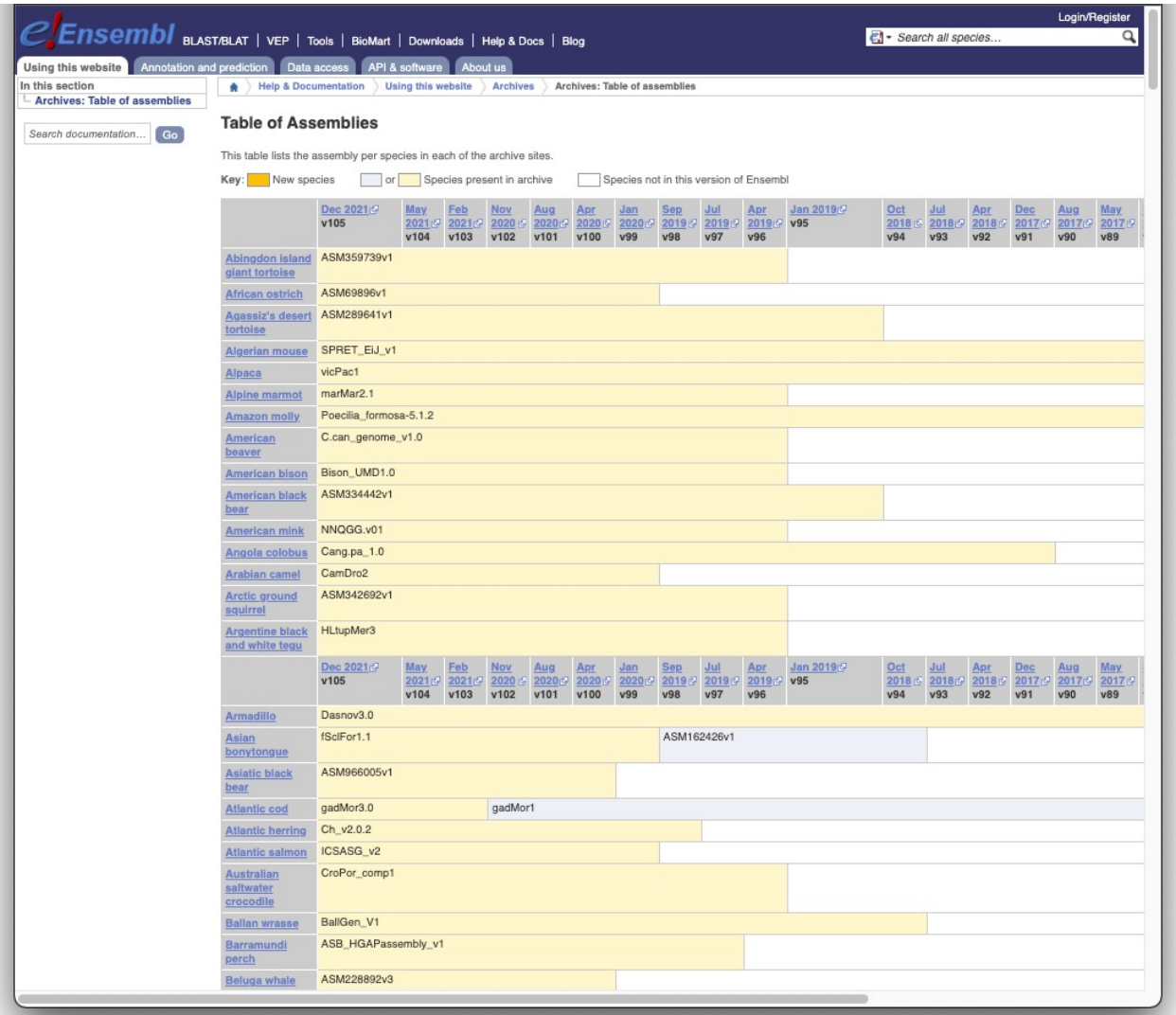
Ensembl : Archives



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

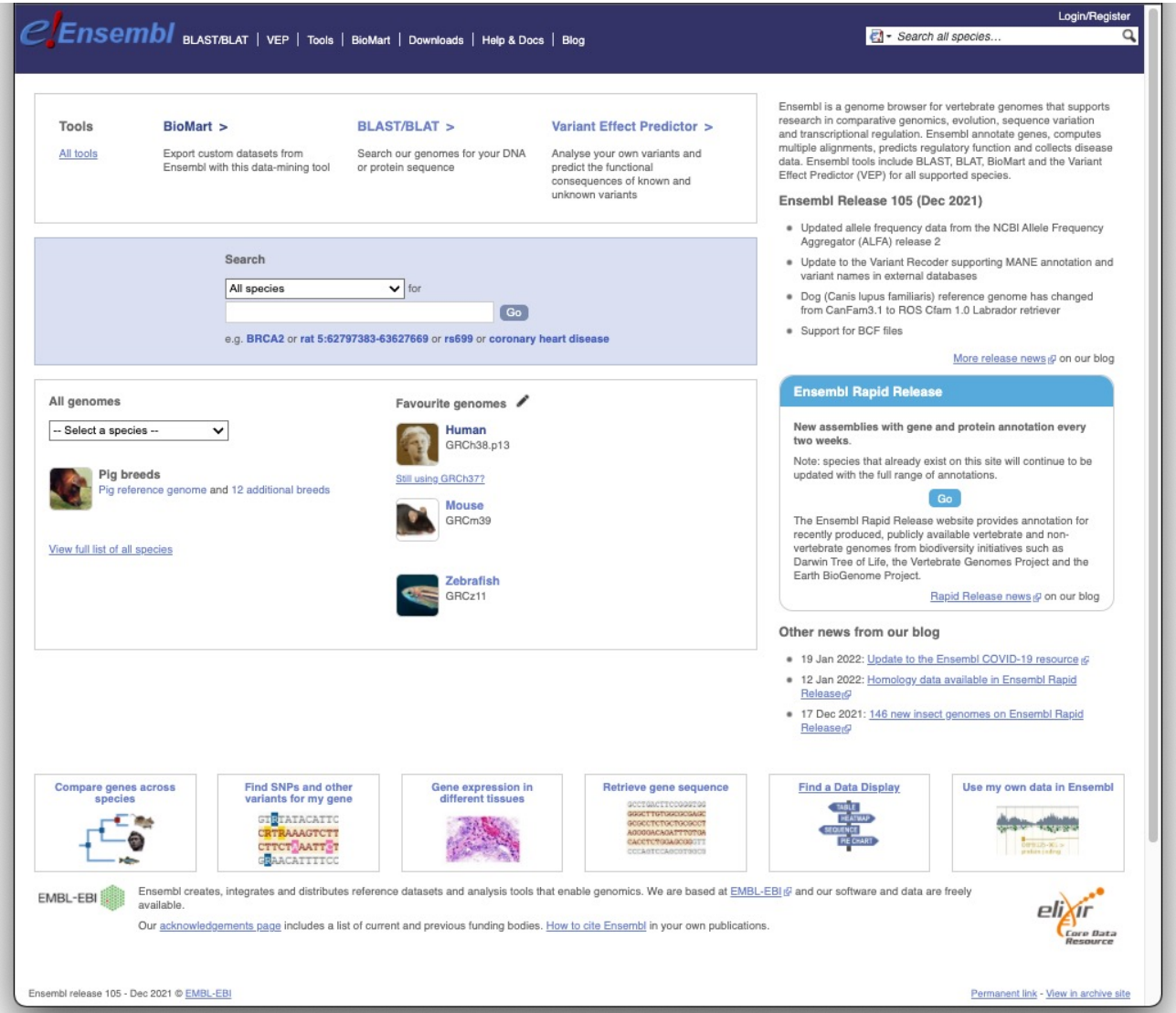


Aide et documentations

- 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)

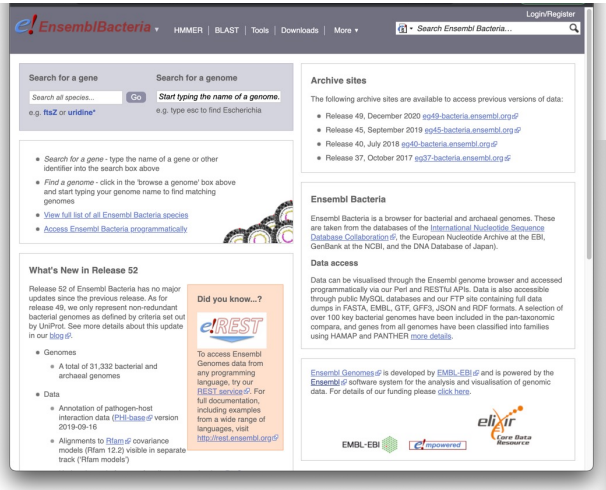
Naviguer dans ensembl

www.ensembl.org

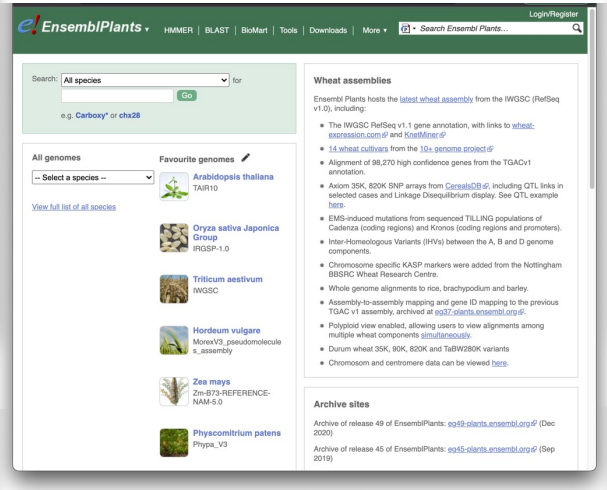


Ensembl Genomes

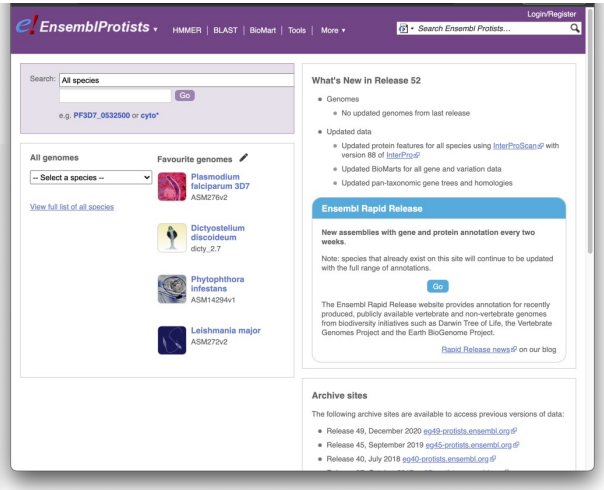
Bactéries



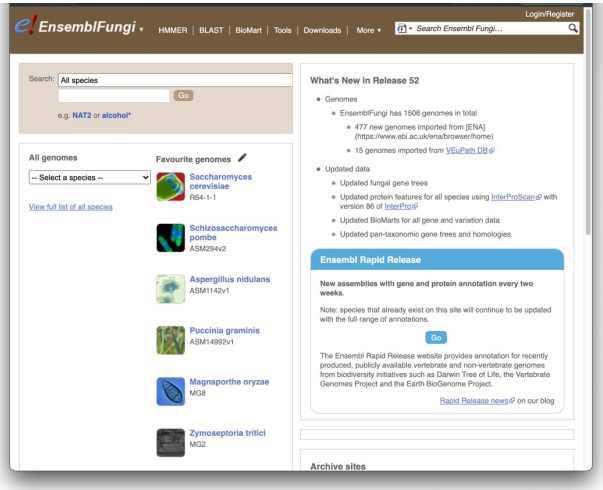
Plantes



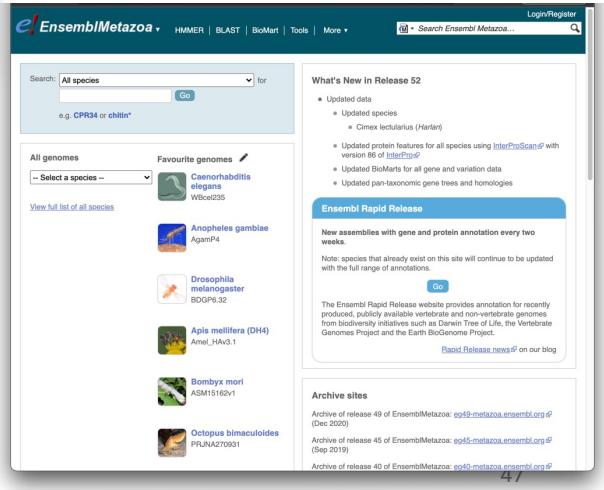
Protistes



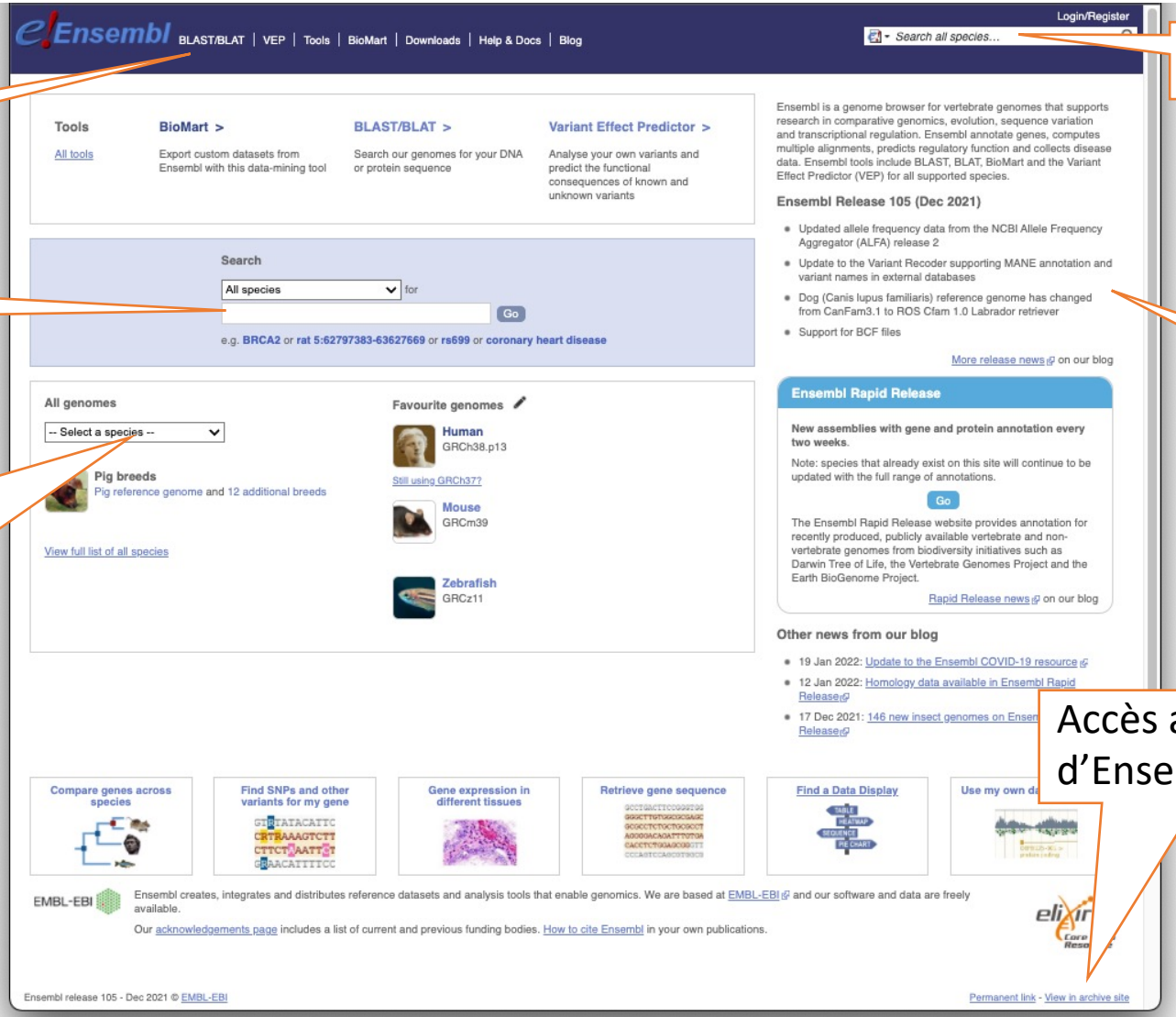
Fungi



Métazoaires



Le site web Ensembl: page d'accueil



Outils

Recherche

Recherche

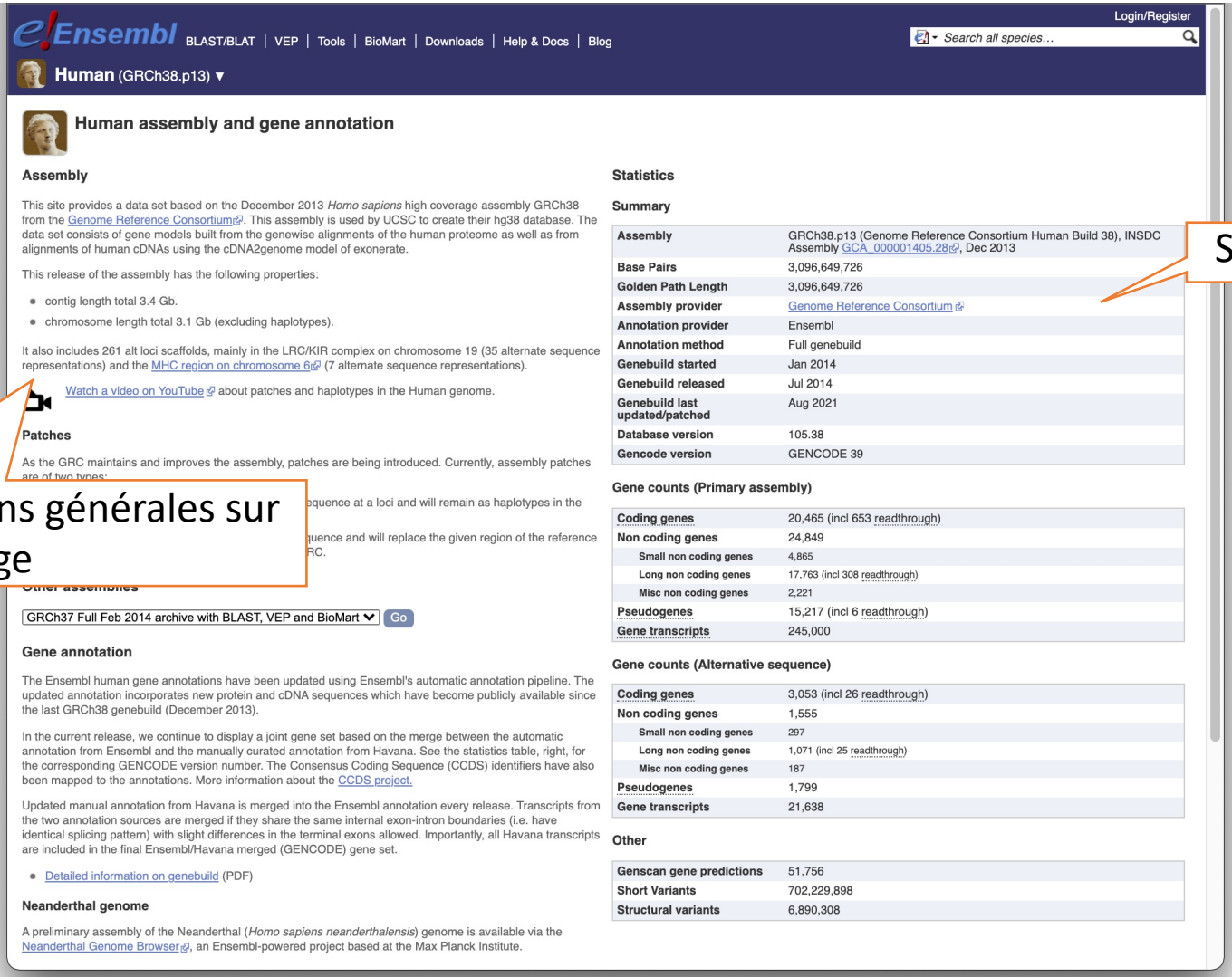
Liste déroulante Accès aux génomes

News

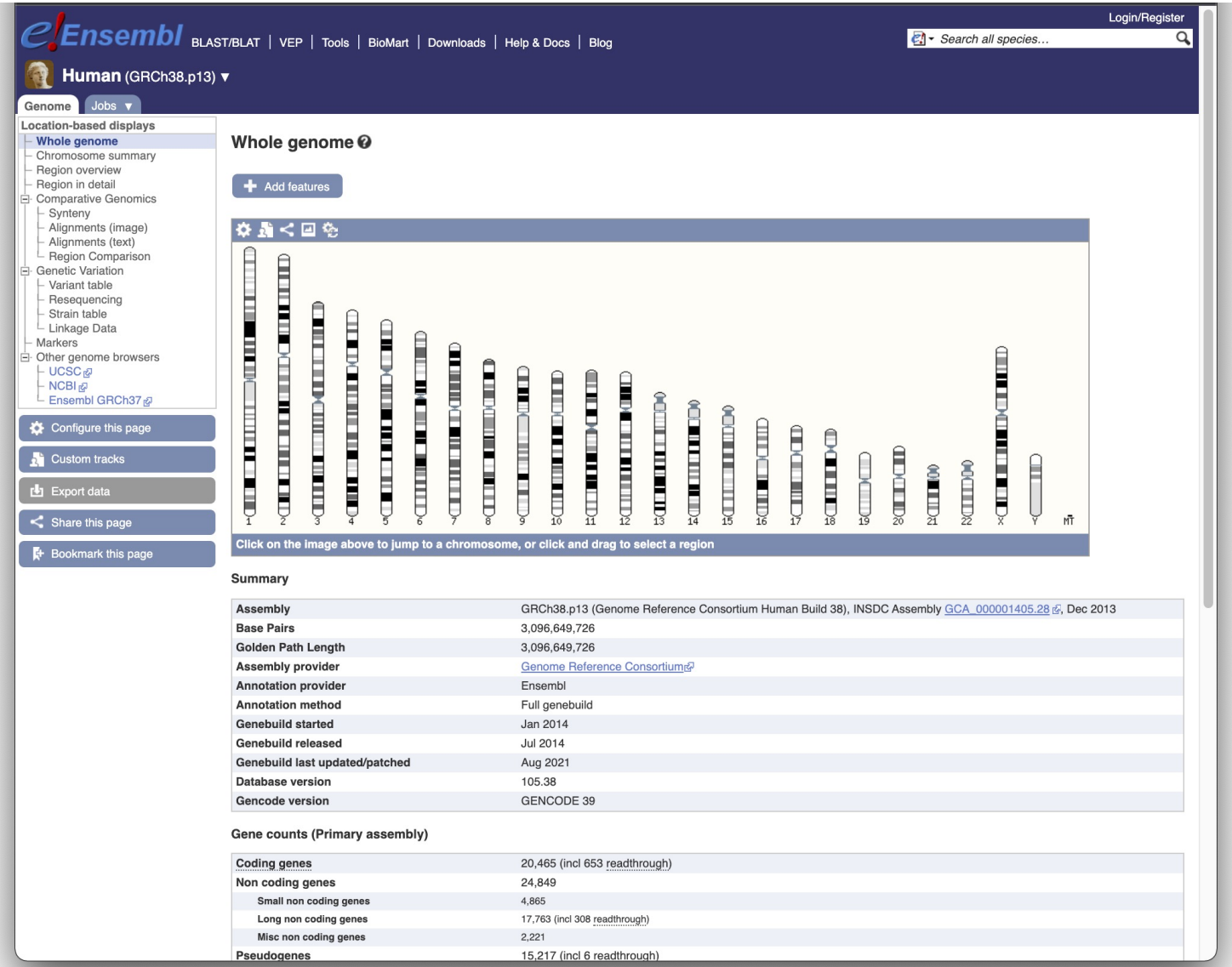
Accès aux archives d'Ensembl

Le site web Ensembl: les génomes

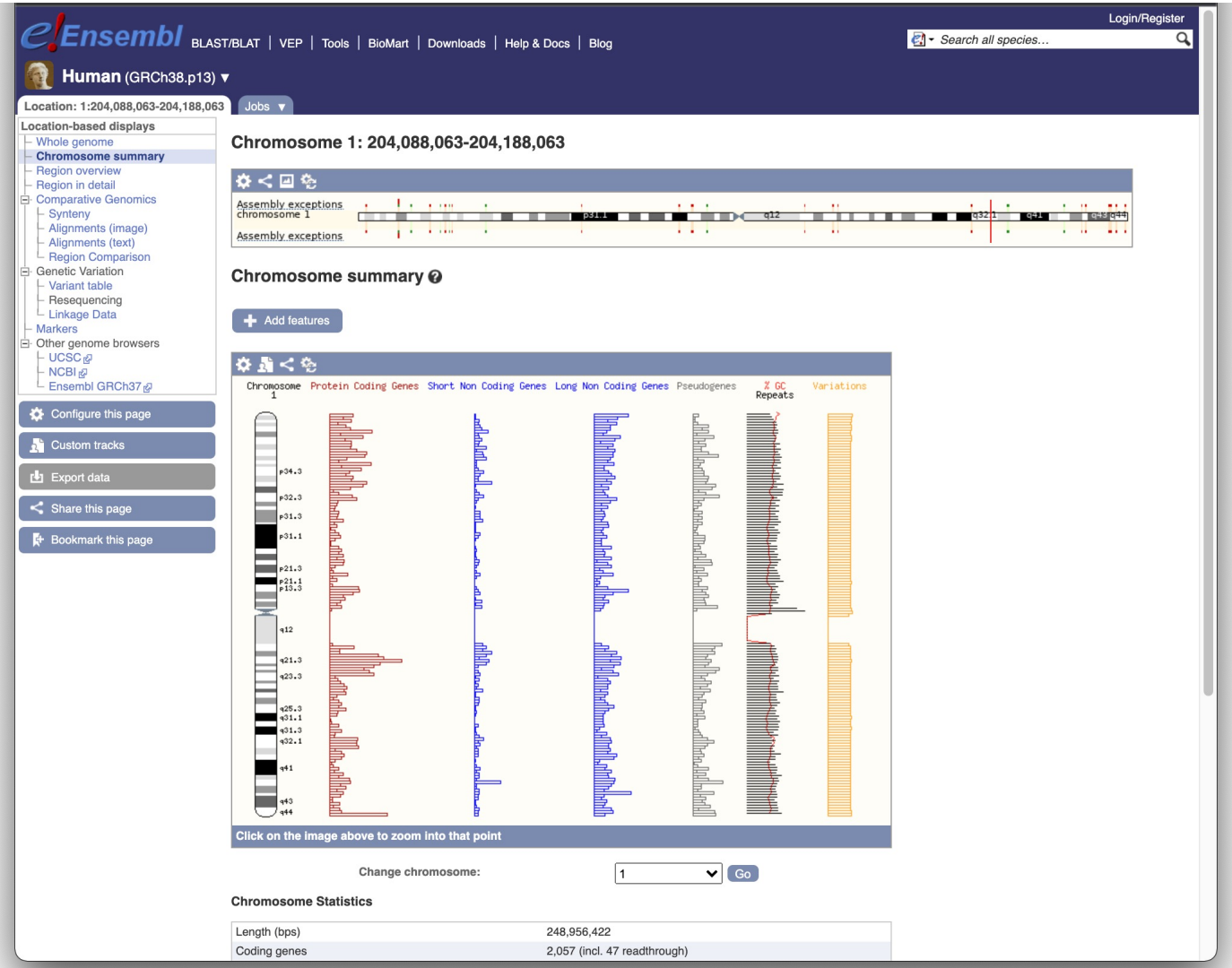
Le site web Ensembl: statistiques des génomes



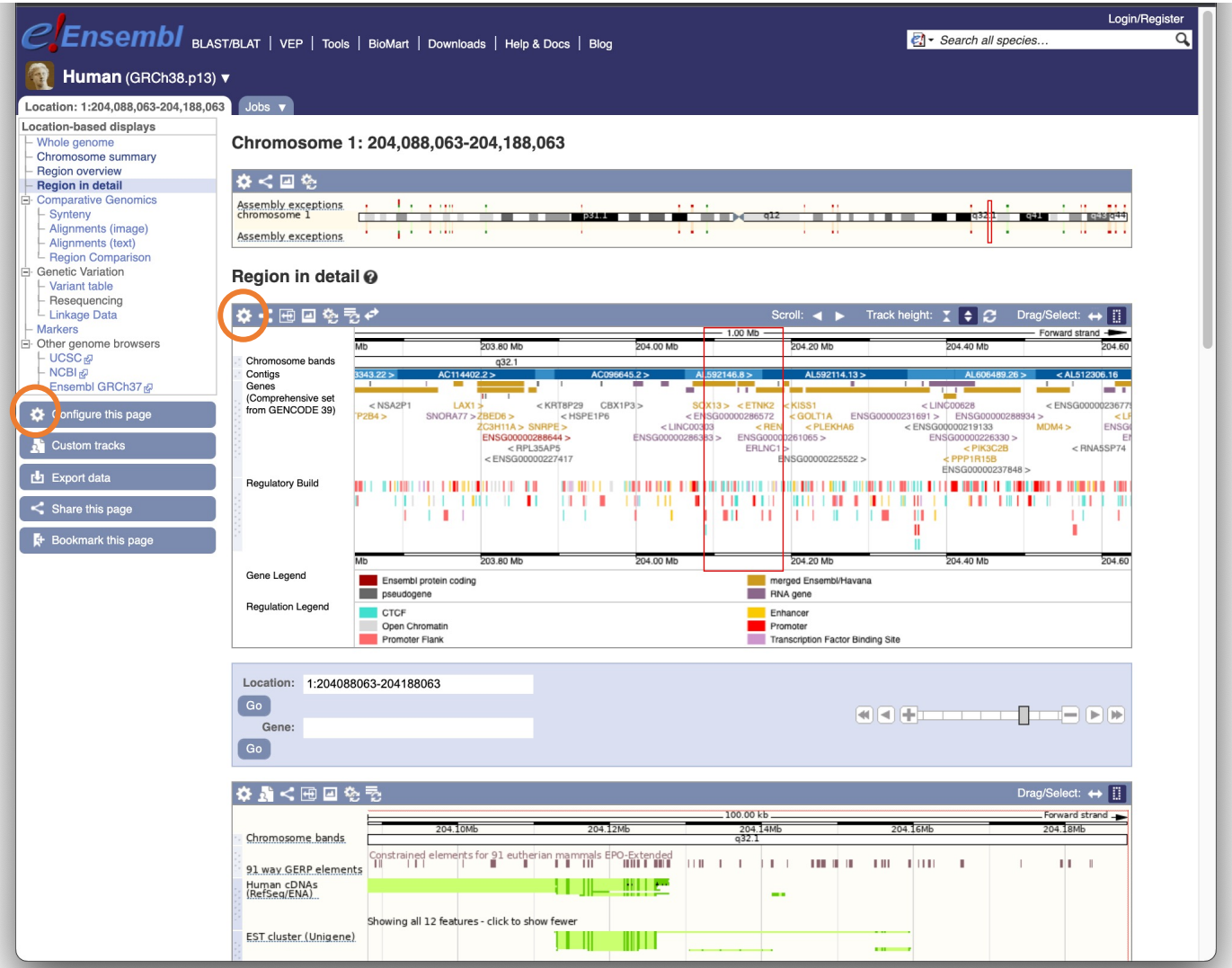
Le site web Ensembl: caryotype



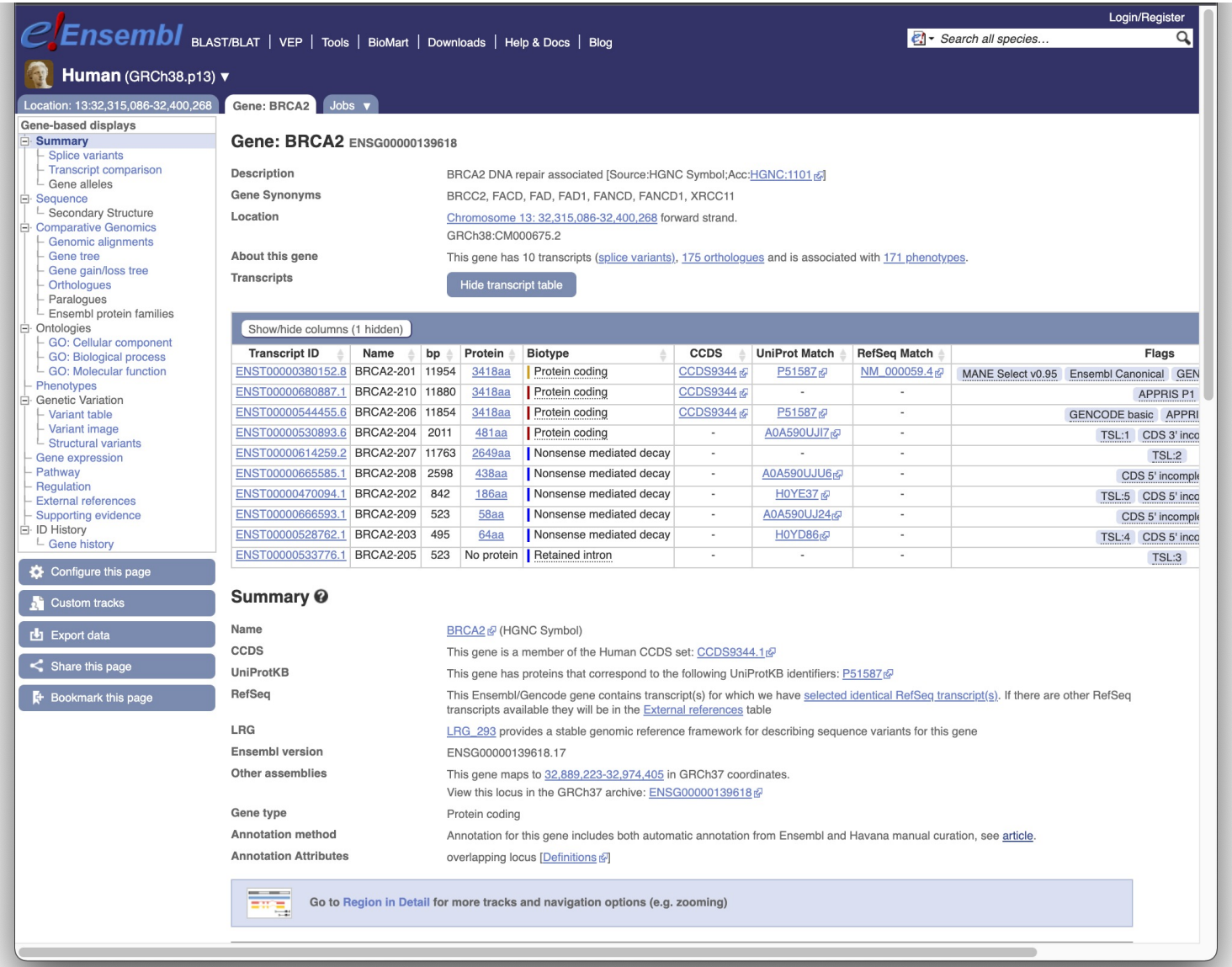
Le site web Ensembl : statistiques par chromosome



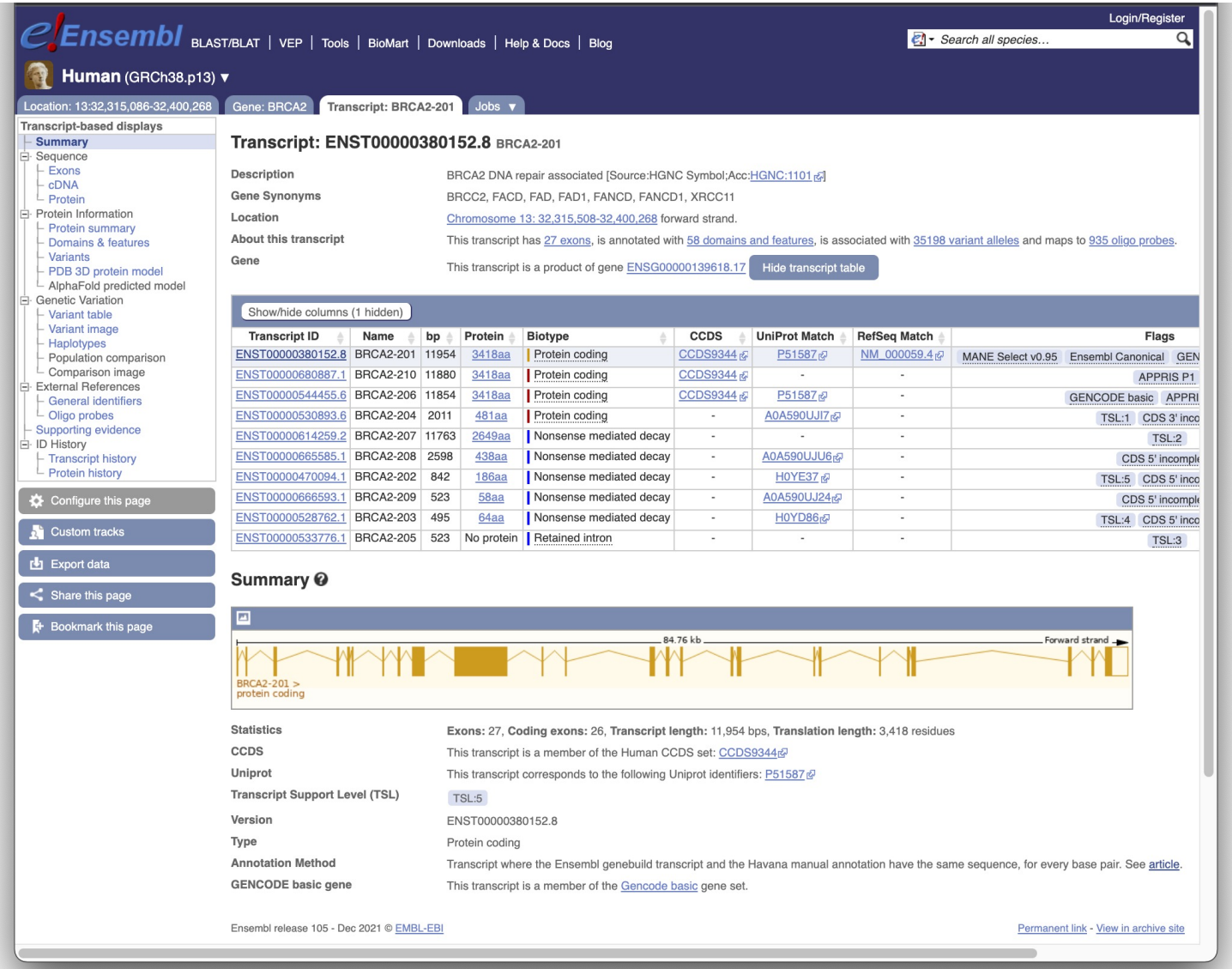
Le site web Ensembl : navigateur de génome



Le site web Ensembl : le gène



Le site web Ensembl : le transcrit



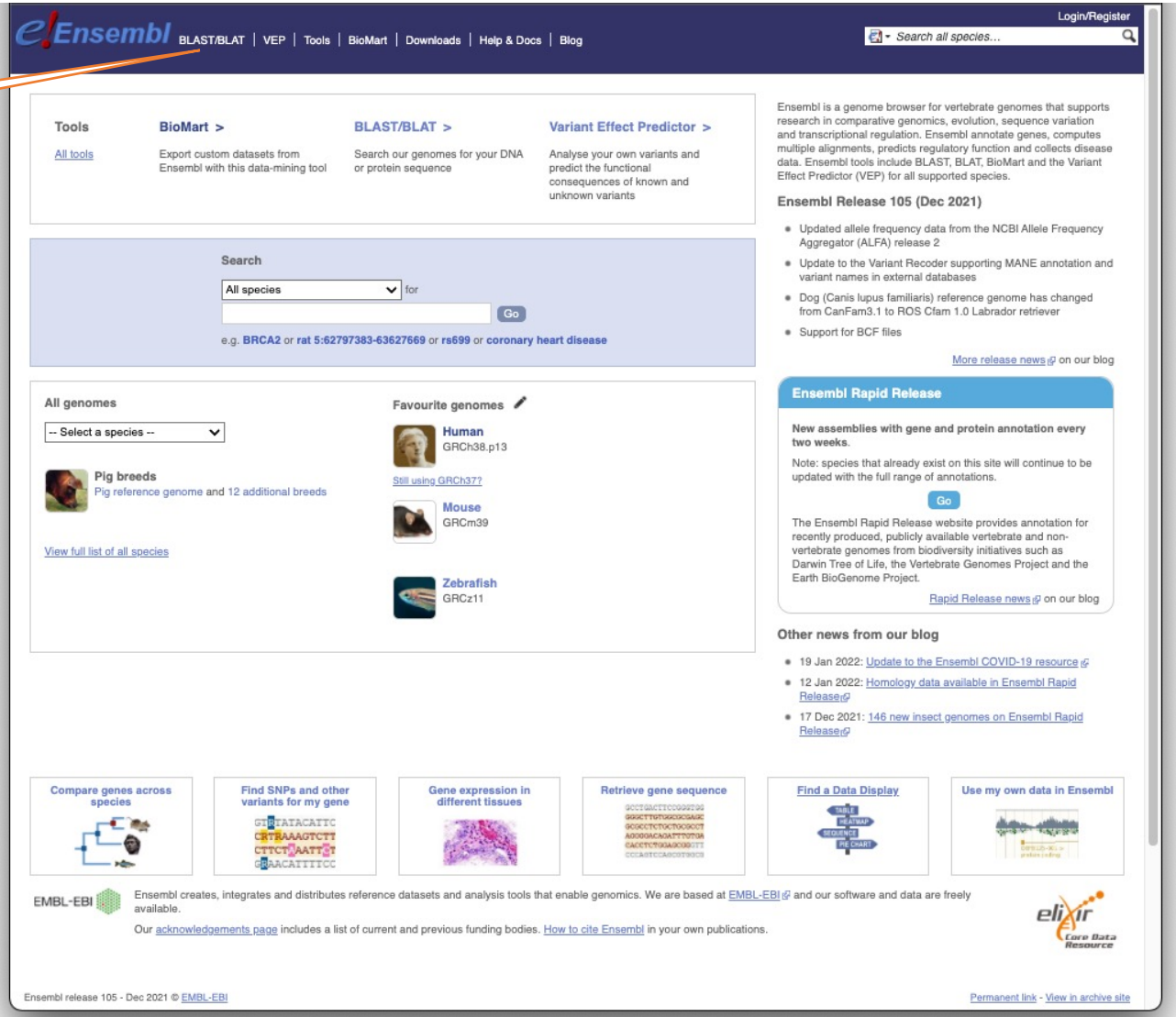
Naviguer dans Ensembl : Partie pratique

Visualiser ses propres données

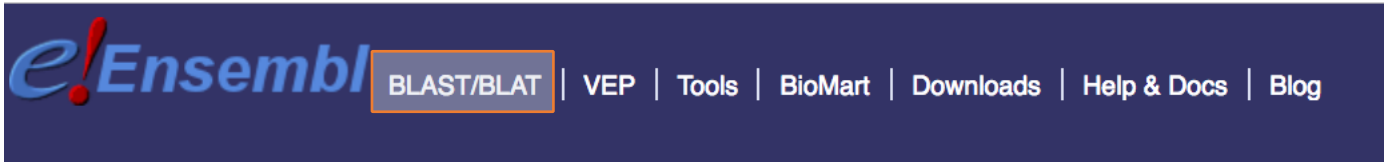
LES OUTILS

Les outils

BLAST/BLAT



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:
  - Détecter 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

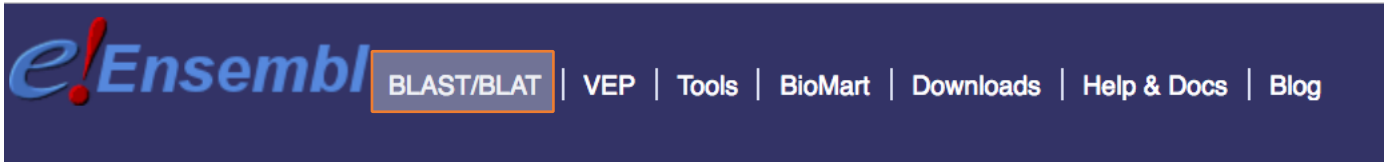
Les différentes comparaisons

BLAT (BLAST-Like Alignment Tool)

- An mRNA/DNA and cross-species protein sequence analysis tool to quickly find sequences of ≥ 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



MADTQYILPNDIGVSSLDCREAFRLLSPTERLYAYHLSRAAWYGGLAVLLQTSPEAPYIYALLSRLFRAQDP
DQLRQHALAEGLTEEEYQAFLVYAAGVYSNMGNYKSFGDTKFVPNLPKEKLERVILGSEAAQQHPEEVRGLW
QTCGELMFSLEPRLRHLGLGKEGITTYFSGNCTMEDAKLAQDFLDSQNLSAYNTRLFKEVDGEGKPYYEVRL
ASVLGSEPSLDSEVTSKLKSYEFRGSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIESFTQG
SIEAHKRGSRFWIQDKGPIVESYIGFIESYRDPFGSRGEFEGFVAVVNKAMSAKFERLVASAEQLLKELPWP
PTFEKDKFLTPDFTSLDVLTFAGSGIPAGINIPNYDDLRQTEGFKNVSLGNVLAVAYATQREKLTFLEEDDK
DLYILWKGPSFDVQVGLHELLGHGSGKLFVQDEKGAFNFDQETVINPETGEQIQSWYRSGETWDSKFSTIAS
SYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIYVNWLNMVRAGLLALEFYTPEAFNWRQAHMQARFVIL
RVLLEAGEGLVTITPTTGSDGRPDARVRLDRSKIRSVGKPALERFLRRLQVLKSTGDVAGGRALYEGYATVT
DAPPECFLTLRDTVLLRKESRKLIVQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEEILTQLAT
ADARFWKGPSEAPSGQA

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 1961 alignments, 23 hits [RawResult] view ▶

Summary

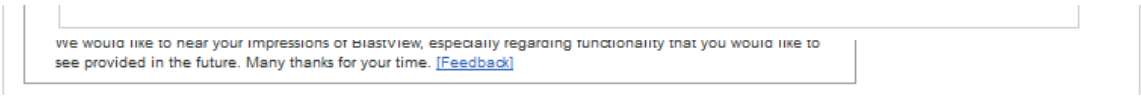
▶ **setup**
- Homo_sapiens
- Genomic sequence
- TBLASTN
- Low sensitivity

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

▶ **results**

▶ **display**

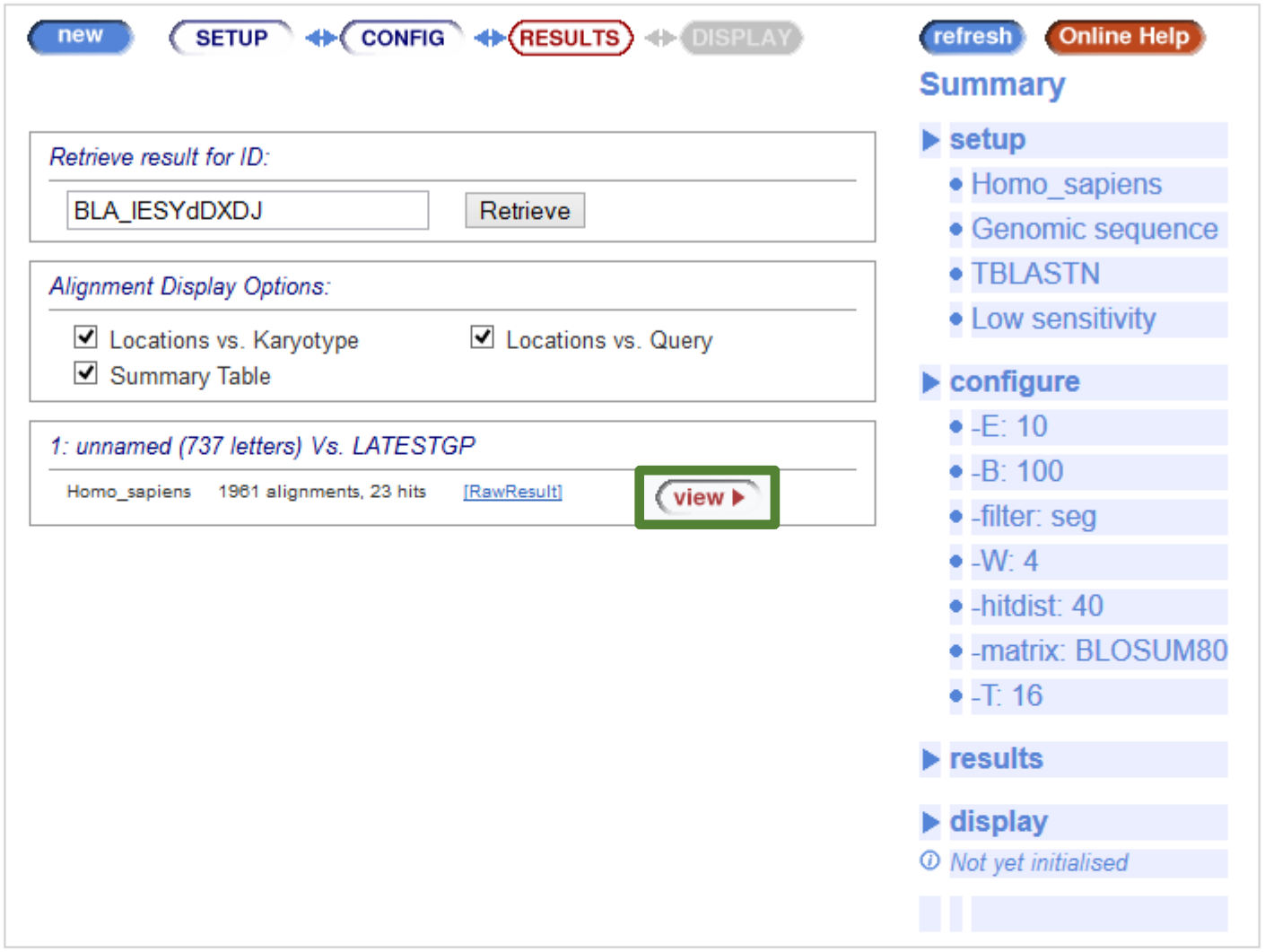
ⓘ *Not yet initialised*



```
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:	Reading Frame	High Score	Smallest Sum Probability 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	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	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	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	89	0.98	5
GL000202.1 dna:supercontig supercontig:GRCh37:GL000202.1:...	+1	111	0.995	3
GL000218.1 dna:supercontig supercontig:GRCh37:GL000218.1:...	-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	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 INPETGEQIQSWYRSGETWDSKFSTIASSYEECRAESVGLYLCLHPQVLEIFGFEGADAE 536
               INPE  EQIQSWYRS +TWDSKFSTI SSYEECRAESVGLYLCLHPQVLE FGFEGADAE
Sbjct: 76090065 INPEMREQIQSWYRSMKTWDSKFSTIVSSYEECRAESVGLYLCLHPQVLETFGFEGADAE 76089886

Query:      537 DVIYVNWLNMVRAGLLALEFYTPEAFNWRQAHMQARFVILRVLLEAGEGLVTITPTTGSD 596
               +VI VNWLNMV AGLLALEFYTPEA NW+QAH++AR VILRVL EAGEGL TITPT GSD
Sbjct: 76089885 EVISVNWLNMVGAGLLALEFYTPEASNWQQAHIRARIVILRVLPEAGEGLGTITPTAGSD 76089706

Query:      597 GRPDARVRLDRSKIRSVGKPALERFLRRLQVLKSTGDVAGGRALYEGYATVTDAPPECFL 656
               GRP+A+VRLDRSKI+SVG PALERFLRR      STGDVAGG  LYE YA V DAPPE FL
Sbjct: 76089705 GRPEAQVRLDRSKIQSVGNPALERFLRRCW---STGDVAGGWTLYERYAAVADAPPEGFL 76089535

Query:      657 TLRDTVLLRKESRKLIVQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEEILT 716
               TLRD VLLRKES KLIVQPN RLEGSDVQLLEYE SAAGLIRSFSE FPEDG ELE+ILT
Sbjct: 76089534 TLRDRVLLRKESWKLIVQPNIRLEGSDVQLLEYEVSAAGLIRSFSEHFPEDGLELEDILT 76089355

Query:      717 QLATADARFWKGPSEAPSGQA 737
               QLATADA+F KGPSEAPSGQA
Sbjct: 76089354 QLATADAQF*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 PSLDSEVTSKLKSYEFRGSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIE 283
               P L  +  SKLKS EFRGSPFQVT G+Y PILQKVVEQLEKAK YAANSHQ QMLAQYIE
Sbjct: 76090816 PGLRGD--SKLKS*EFRGSPFQVTWGNYMPILQKVVEQLEKAKTYAANSHQEQMLAQYIE 76090643

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

Query:      344 ERLVASAEQLLKELPWPPTFEKDKFLTPDFTSLDVLTFAGSGIPAGINIPNYDDLRQTEG 403
               E LV SAEQLLKELPW P FEKDKFLTPDFTS+DVLTFAGSGI AGINI NY+DL+QTEG
Sbjct: 76090462 EWLVVSAEQLLKELPWSPAFEKDKFLTPDFTSVDVLTFAGSGIAAGINISNYNDLKQTEG 76090283

Query:      404 FKNVSLGNVLAVAYATQREKLTFLEEDDKDLYILWKGPSFDVQVGLHELLGHGSGKLFVQ 463
               FKNVSLGNVLAV  ATQ EKLT LEE DKDLYI+  GPSFDVQVGLHELLG+GSGKL  Q
Sbjct: 76090282 FKNVSLGNVLAVV*ATQWEKLTVLEESDKDLYIVLMGPSFDVQVGLHELLGYGSGKLIEQ 76090103

Query:      464 DEKGAFNFDQET 475
               DEKGAFNFDQET
Sbjct: 76090102 DEKGAFNFDQET 76090067
```


new
SETUP
CONFIG
RESULTS
DISPLAY
refresh
Online Help
Summary
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 1961 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
Not yet initialised

new | SETUP | CONFIG | RESULTS | DISPLAY

refresh | Online Help

Summary

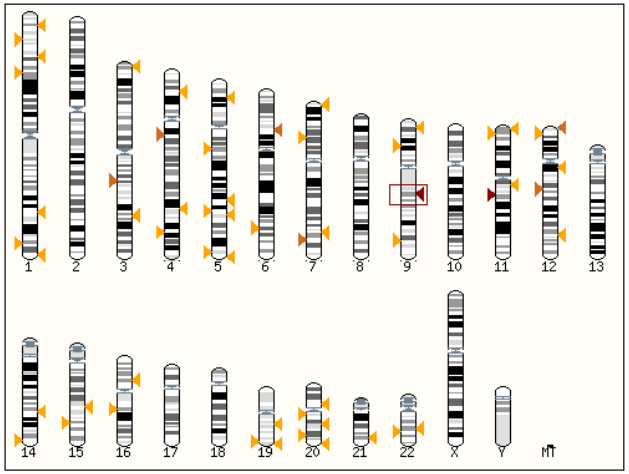
- **setup**
  - Homo_sapiens
  - Genomic sequence
  - TBLASTN
  - Low sensitivity
- **configure**
  - -E: 10
  - -B: 100
  - -filter: seg
  - -W: 4
  - -hitdist: 40
  - -matrix: BLOSUM80
  - -T: 16
- **results**
- **display**

Not yet initialised

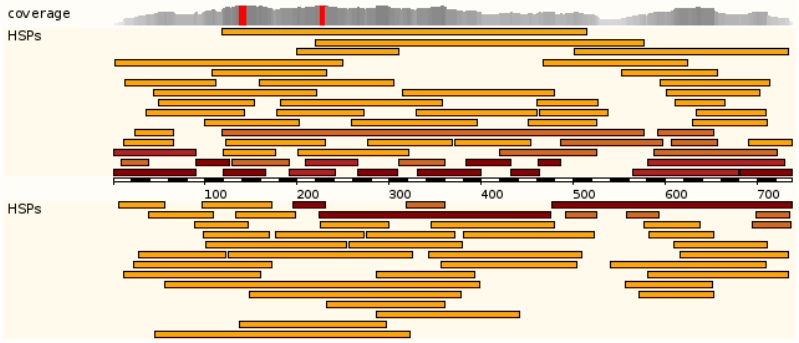
Displaying unnamed sequence alignments vs Homo_sapiens LATESTGP database

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)



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	<Score
Start	Start	Start	Start	Start	Start	Start	E-val	>Score

Links	Query Start	Query End	Ori	Chromosome Name	Chromosome Start	Chromosome End	Ori	Stats 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] [S] [G] [C]

[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
Percentage identity: 84.29

Query:      477 INPETGEQIQSWYRSGETWDSKFSTIASSYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIY 536
               INPE  EQIQSWYRS +TWDSKFSTI SSYEECRAESVGLYLCLHPQVLE FGFEGADAE
Sbjct: 76090065 INPEMREQIQSWYRSMKTWDSKFSTIVSSYEECRAESVGLYLCLHPQVLETFGFEGADAE 76089886

Query:      537 DVIYVNWLNMVRAGLLALEFYTPEAFNWRQAHMQARFVILRVLLEAGEGLVTITPTTGSD 596
               +VI VNWLNMV AGLLALEFYTPEA NW+QAH++AR VILRVL EAGEGL TITPT GSD
Sbjct: 76089885 EVISVNWLNMVGAGLLALEFYTPEASNWQQAHIRARIVILRVLPEAGEGLGTITPTAGSD 76089706

Query:      597 GRPDARVRLDRSKIRSVGKPALERFLRRLQVLKSTGDVAGGRALYEGYATVTDAPPECFL 656
               GRP+A+VRLDRSKI+SVG PALERFLRR     STGDVAGG  LYE  V V DAPPE FL
Sbjct: 76089705 GRPEAQVRLDRSKIQSVGNPALERFLRRCW---STGDVAGGWTLYERAAVADAPPEGFL 76089535

Query:      657 TLRDTVLLRKESRKLIVQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEEILT 716
               TLRD VLLRKES KLIVQPN RLEGSDVQLLEYE SAAGLIRSFSE FPEDG ELE+ILT
Sbjct: 76089534 TLRDRVLLRKESWKLIVQPNIRLEGSDVQLLEYEVSAAGLIRSFSEHFPEDGLELEDILT 76089355

Query:      717 QLATADARFWKGPSEAPSGQA 737
               QLATADA+F KGPSEAPSGQA
Sbjct: 76089354 QLATADAQF*KGPSEAPSGQA 76089292
```

[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

```
>unnamed
MADTQYILPNDIGVSSLDCREAFRLLSPTERLYAYHLSRAAWYGGLAVLLQTSPEAPYIY
ALLSRLFRAQDPDQLRQHALAEGLTEEEYQAFLVYAAGVYSNMGNYKSFGDTKFVPNLPK
EKLERVILGSEAAQQHPEEVRGLWQTCGELMFSLEPRLRHLGLGKEGITTYFSGNCTMED
AKLAQDFLDSQNLSAYNTRLFKEVDGEGKPYYEVRLASVLGSEPSLDSEVTSKLKSYEFR
GSPFQVTRGDYAPILQKVVEQLEKAKAYAANSHQGQMLAQYIESFTQGSIEAHKRGSRFW
IQDKGPIVESYIGFIESYRDPFGSRGEFEGFVAVVNKAMSAKFERLVASAEQLLKELPWP
PTFEKDKFLTPDFTSLDVLTFAGSGIPAGINIPNYDDLRQTEGFKNVSLGNVLAVAYATQ
REKLTFLEEDDKDLYILWKGPSFDVQVGLHELLGHGSGKLFVQDEKGAFNFDQETVINPE
TGEQIQSWYRSGETWDSKFSTIASSYEECRAESVGLYLCLHPQVLEIFGFEGADAEDVIY
VNWLNMVRAGLLALEFYTPEAFNWRQAHMQARFVILRVLLEAGEGLVTITPTTGSDGRPD
ARVRLDRSKIRSVGKPALERFLRRLQVLKSTGDVAGGRALYEGYATVTDAPPECFLTLRD
TVLLRKESRKLIVQPNTRLEGSDVQLLEYEASAAGLIRSFSERFPEDGPELEEILTQLAT
ADARFWKGPSEAPSGQA
```

[G]Sequence

```
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
```

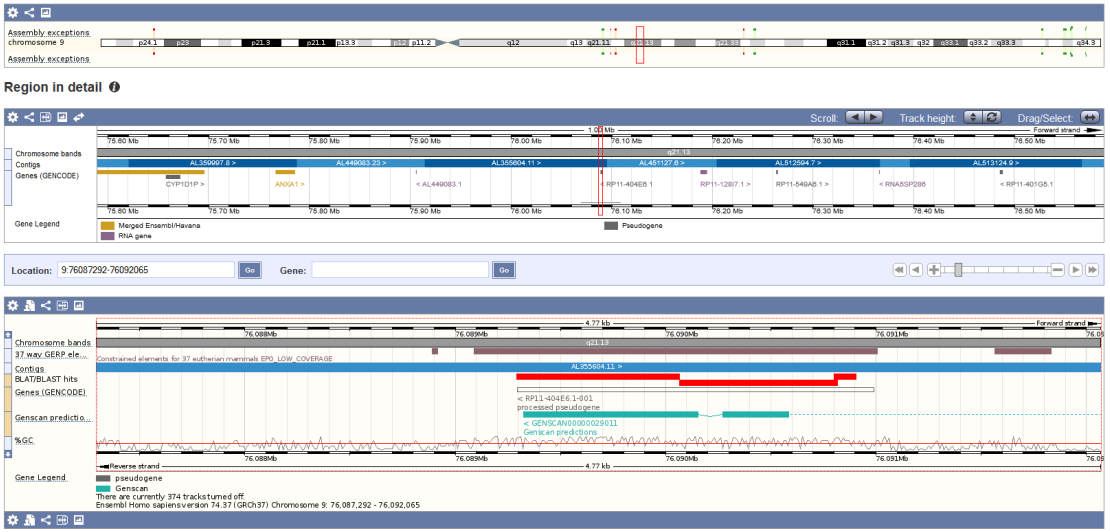
5' Flanking sequence 300 (bp)
3' Flanking sequence 300 (bp)
Coordinate system: Chromosome
Orientation: Forward relative to selected alignment
Alignment markup: All alignments | Both orientations
Feature markup: Ensembl exons | Both orientations
Line numbering: No numbers
update

THIS STYLE: Location of selected alignment
THIS STYLE: Location of other alignments
THIS STYLE: Location of Exons

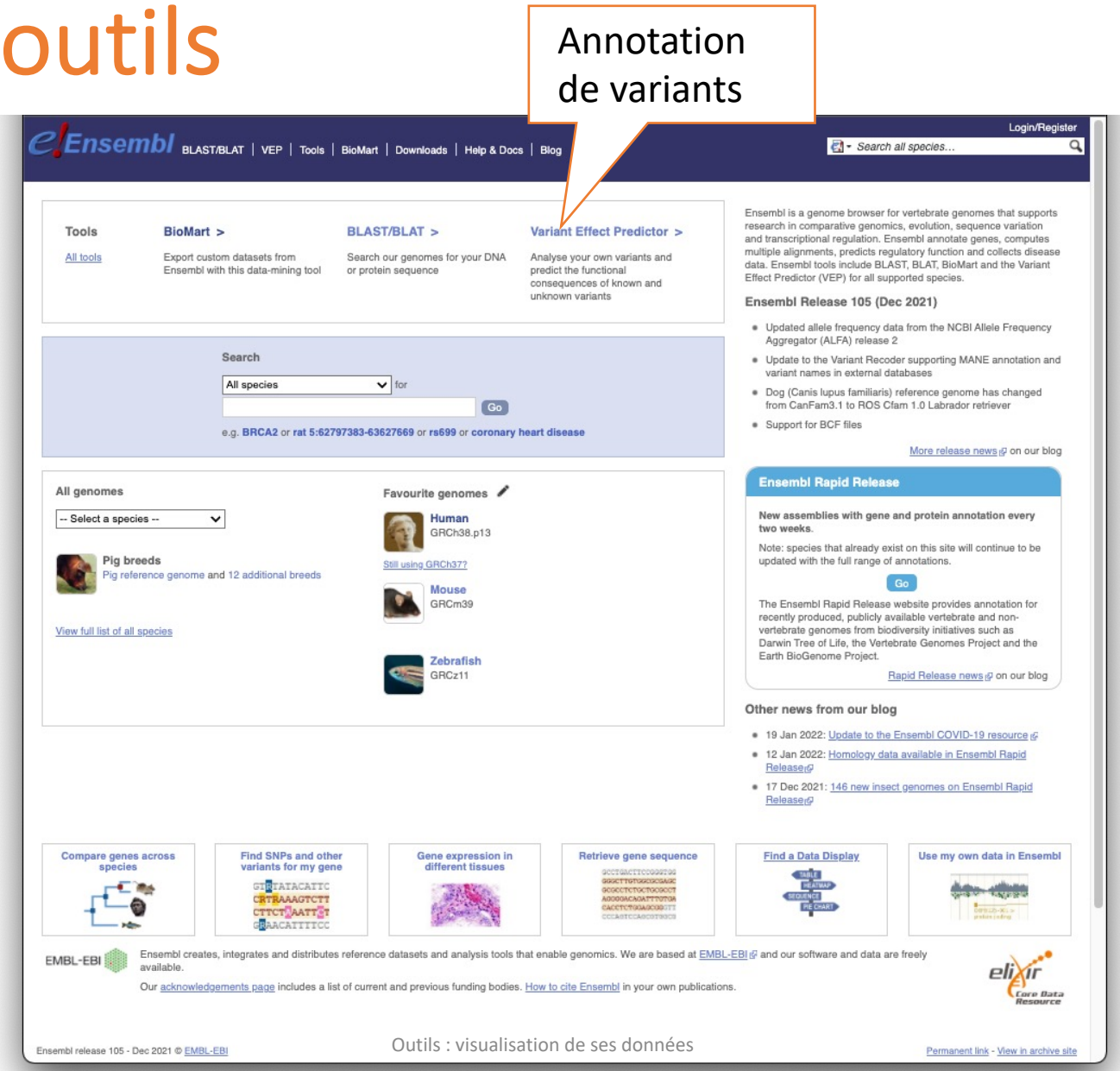
```
>chromosome:GRCh37:9:76088992:76090365:-1
TGGATGTTCTTACCTTCGCTGGCTCTGGCATCGCTGCCGGCATCAACATCTCCAACTACA
ATGACCTGAAGCAGACGGAAGGCTTTAAGAACGTGTCACTGGGGAACGTCCTGGCTGTGG
TCTAAGCCACGCAGTGGGAGAAGCTCACCGTTCTGGAGGAGAGTGACAAGGACTTGTACA
TCGTCCTGATGGGGCCCTCCTTCGATGTACAGGTGGGCCTGCACGAACTGCTGGGCTATG
GCAGCGGCAAGCTCATCGAGCAGGATGAAAAAGGAGCATTCAACTTTGACCAGGAGACCC
ATCAATCCAGAGATGAGAGAGCAGATTCAGAGCTGGTATCGGAGCATGAAGACCTGGGAC
AGCAAGTTCAGCACCATTGTCTCCAGCTACGAAGACTGCCGGGCTGAGAGCGTGGCCCTC
TACCTCTGTCTCCACCCGCAAGTGCTGGAGACCTTTGGCTTTGAGGGGGCTGATGCAGAG
GAGGTGATCTCCGTGAACTGGCTCAACATGGTTGGGGCTGGGCTGCTCGCTCTGGAGTTC
TACACACCTGAGGCCTCCAACTGGCAGCAGGCCCATATACGGGCCCGGATTGTGATCCTG
AGAGTCTTACCAGAGGCTGGCGAGGGACTCGGTACCATCACTCCCACCGCAGGCTCCGAT
GGGCGCCCGGAAGCCCAGGTCCGTCTCGACCGCAGCAAGATCCAGTCTGTGGGCAACCCT
GCCCTAGAGCGCTTCCTGCGGAGGTGCTGGTCCACAGGGGATGTGGCCGGAGGGTGGACC
CTGTACGAGAGGTATGCAGCAGTCGCTGATGCGCCCCCTGAGGGCTTTCTCACCCTCAGG
GACAGGGTGCTGCTGCGTAAGGAATCTTGGAAGCTCATCGTTCAGCCCAACATTCGCCTT
GAAGGCTCAGACGTGCAGCTTCTGGAATACGAGGTCTCAGCTGCTGGCCTCATCAGATCC
TTCTCTGAGCATTTCCCAGAGGATGGACTCGAGTTGGAGGACATCCTAACACAGCTGGCA
ACAGCCGATGCCCAATTCTGAAAGGGCCCCAGTGAGGCCCCATCTGGCCAGGCTTGTGGA
AGAAGTGTGTGGCCTCTCCCCCAACTGCATCAAACCAAGGCTGCAAGTGGCCCTCCATTC
GTGTTGTATTTAGGGGCTGGGGAGGGGGAGGGGGAGGGGCAGGAGCTTGGACCTTGGTAC
TACCTCAGCTGAGGGTGGTGACACTAACCCCTTCCATTTGTCAGCACCTTCCAGCGTGCC
AAGTGCTTCCCCTCTGTGAACTCATTTCATATGCACTGCCATATGTGGAGTGAGCAAGAC
CGTGCTCACCATCCCGTCTACCAGATGAGGAAATGGCAGTTCTGAGAAATCACG
```

[C]ontig view (?)

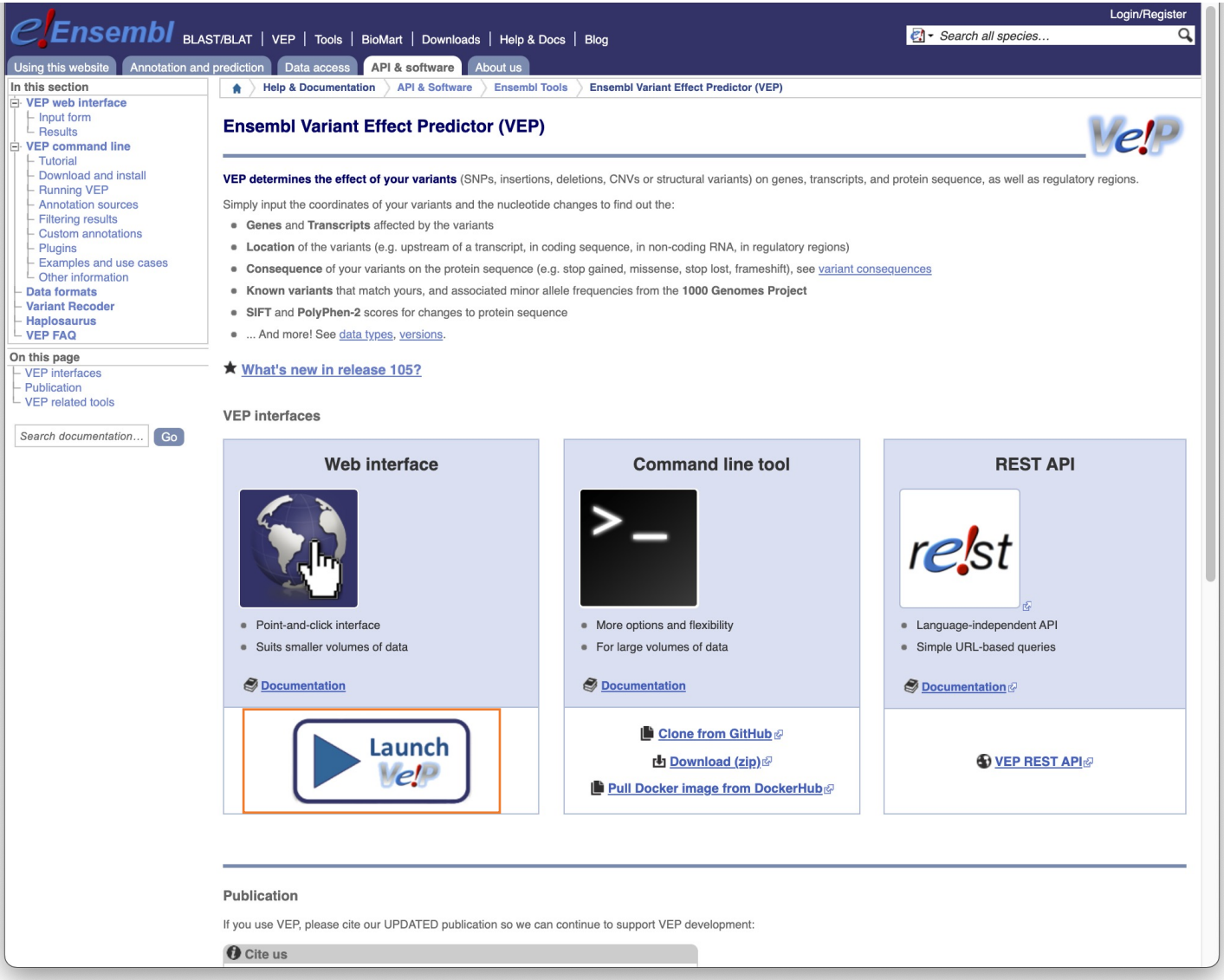
Chromosome 9: 76,087,292-76,092,065



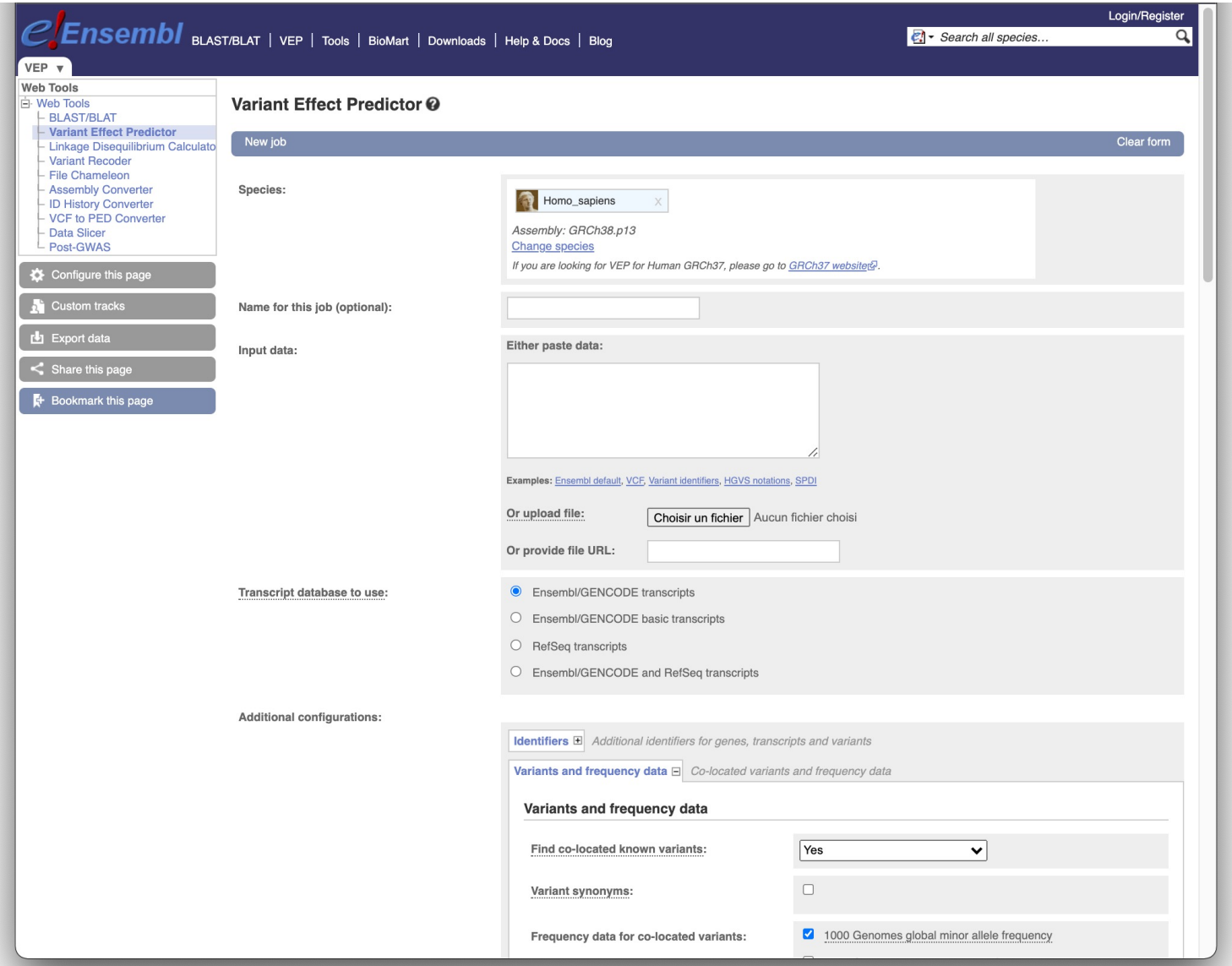
Les outils


Annotation de variants
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
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
Search
All species
for
Go
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease
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
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
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
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
Permanent link - View in archive site

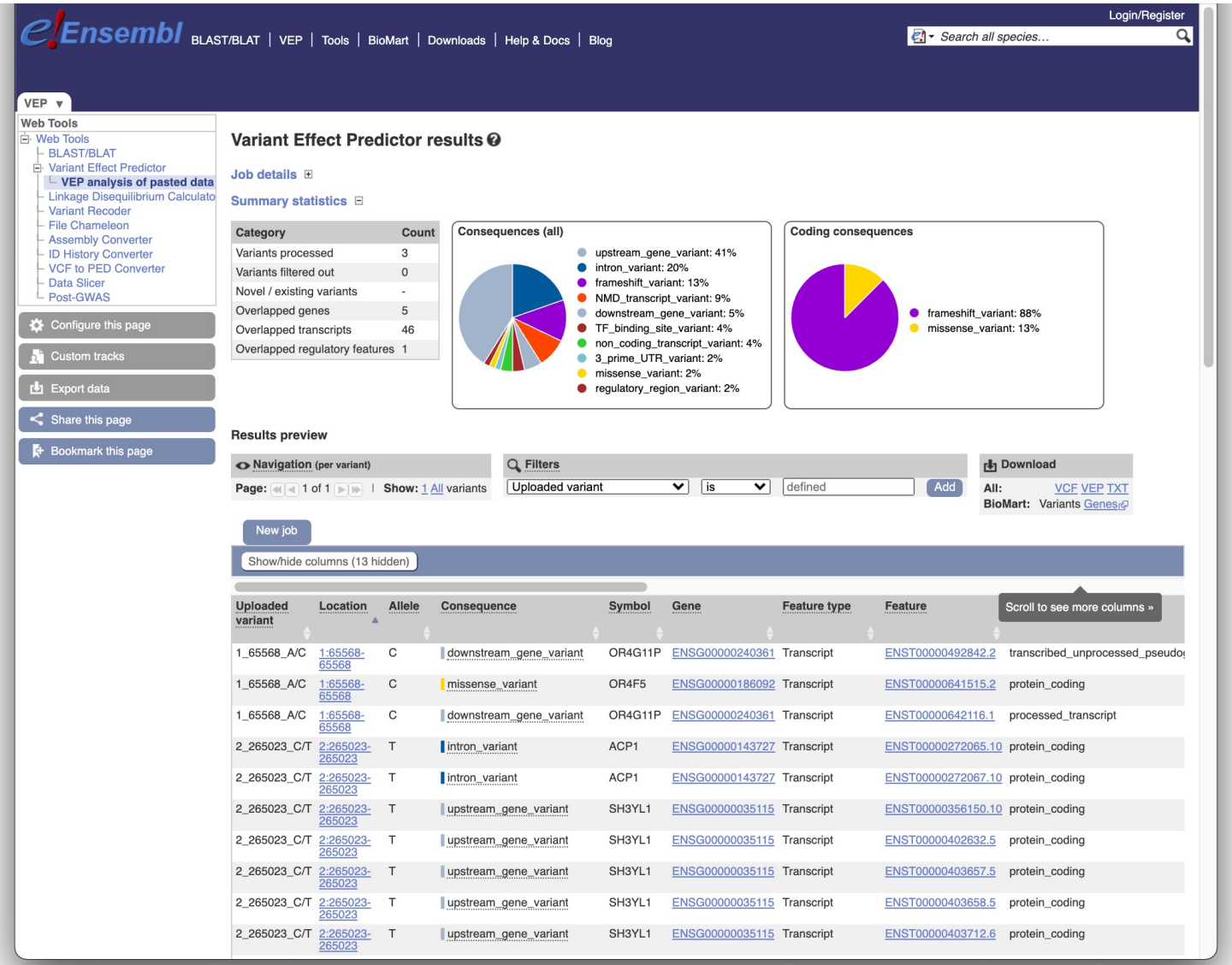
Variant Effect Predictor



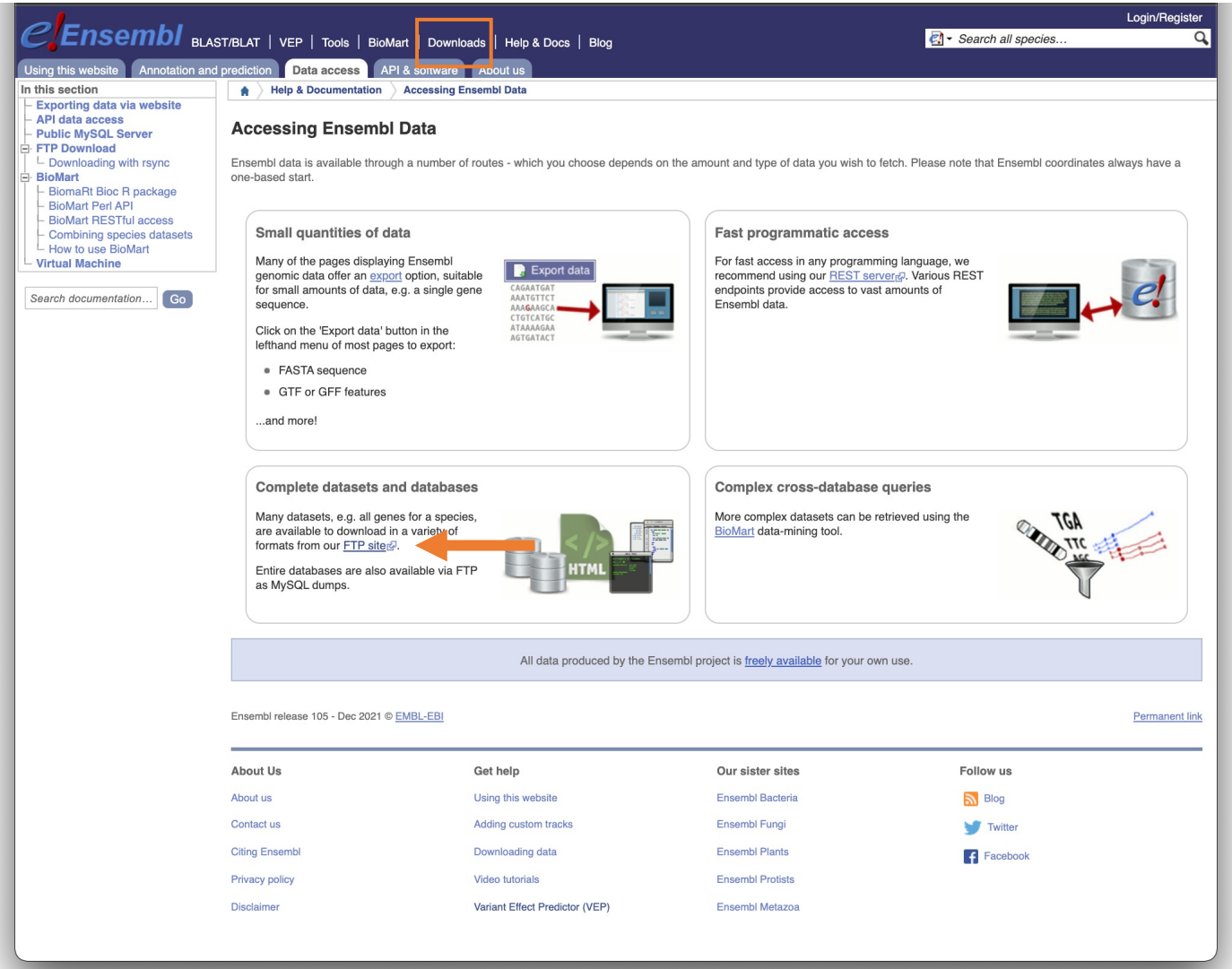
Variant Effect Predictor



Variant Effect Predictor



Outils de récupération de données

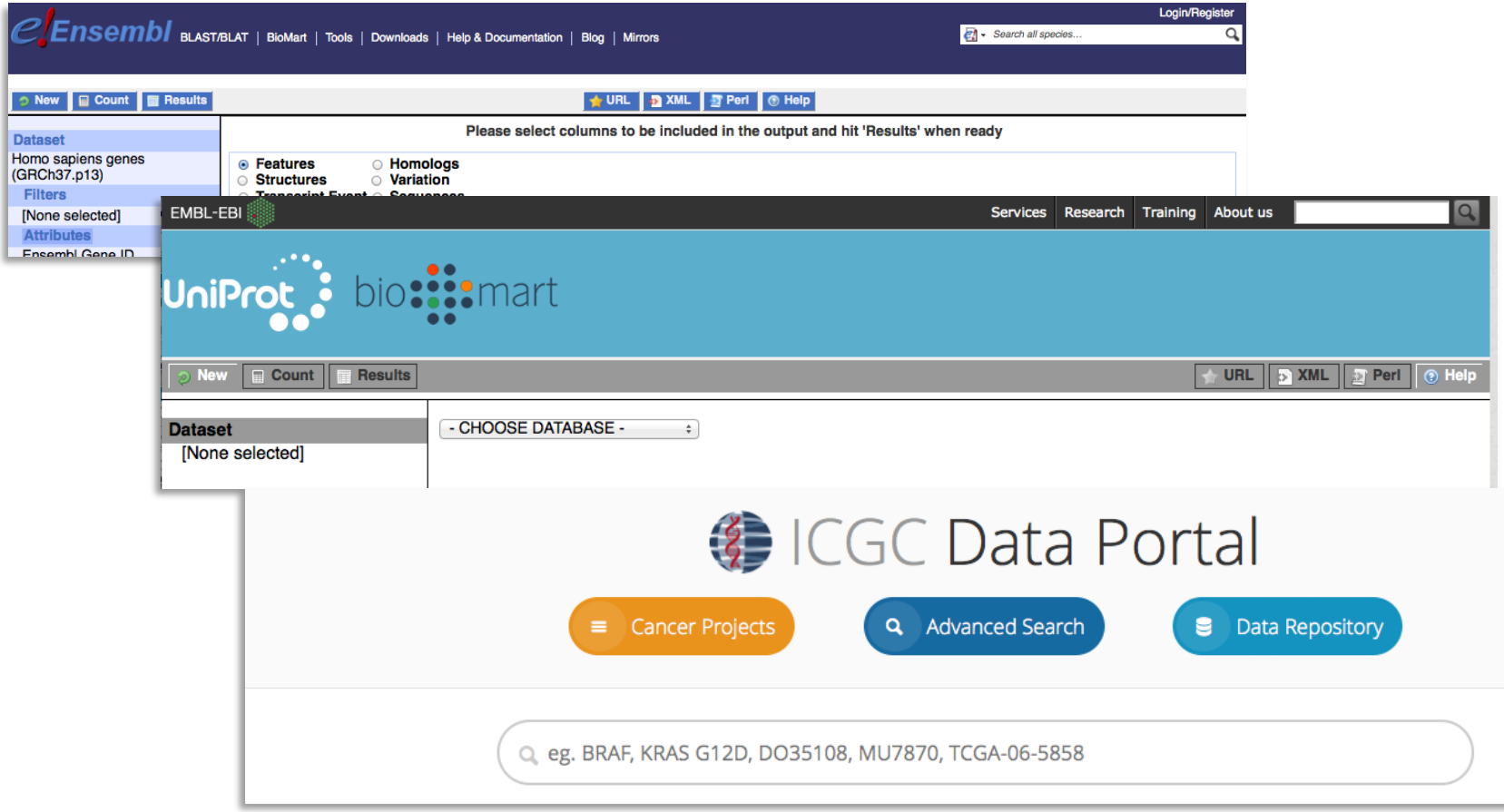


BioMart

Le projet BioMart

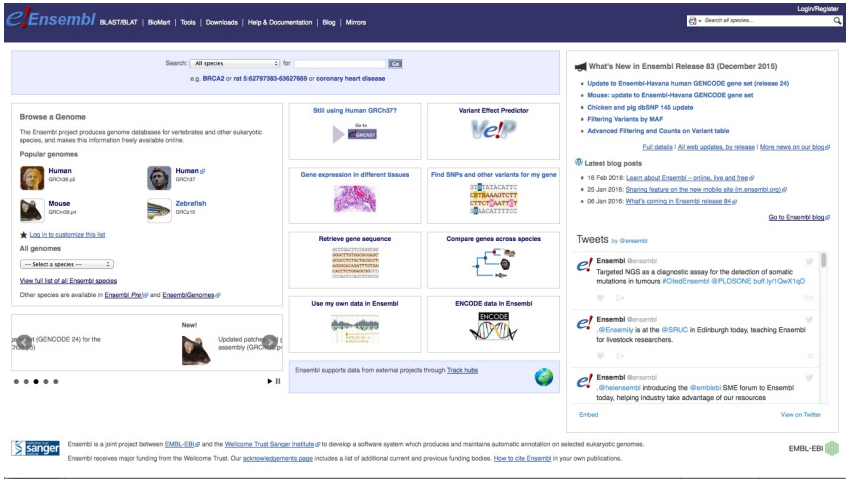
- http://www.biomart.org/
- Développé conjointement par :
  - EBI
  - Cold Spring Harbor Laboratory (CSHL)
- Arek Kasprzyk : « BioMart can access diverse databases from a single interface »
- Créer un système générique de stockage et de gestion de données
- « Data-agnostic » : manipulation de n'importe 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”


Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
New
Count
Results
URL
XML
Perl
Help
Dataset
Homo sapiens genes (GRCh37.p13)
Filters
[None selected]
Attributes
Please select columns to be included in the output and hit 'Results' when ready
Features
Structures
Homologs
Variation
EMBL-EBI
Services
Research
Training
About us
UniProt
biomart
New
Count
Results
URL
XML
Perl
Help
Dataset
[None selected]
- CHOOSE DATABASE -
ICGC Data Portal
Cancer Projects
Advanced Search
Data Repository
eg. BRAF, KRAS G12D, DO35108, MU7870, TCGA-06-5858

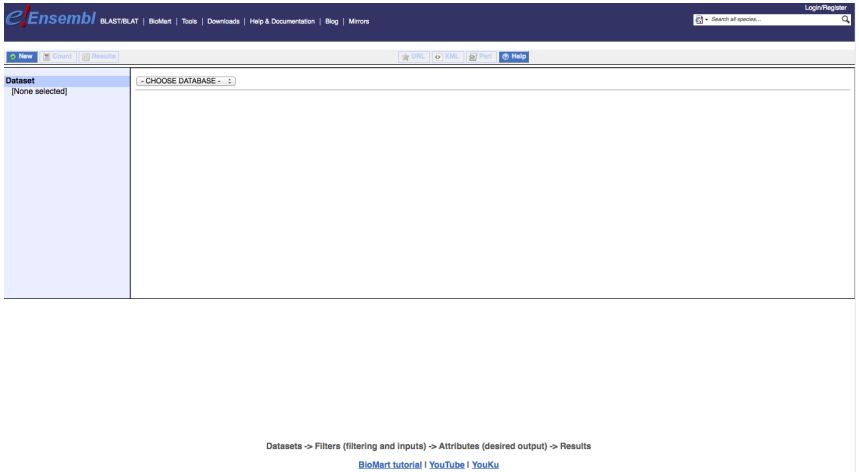
Accéder aux données d'Ensembl

Site web



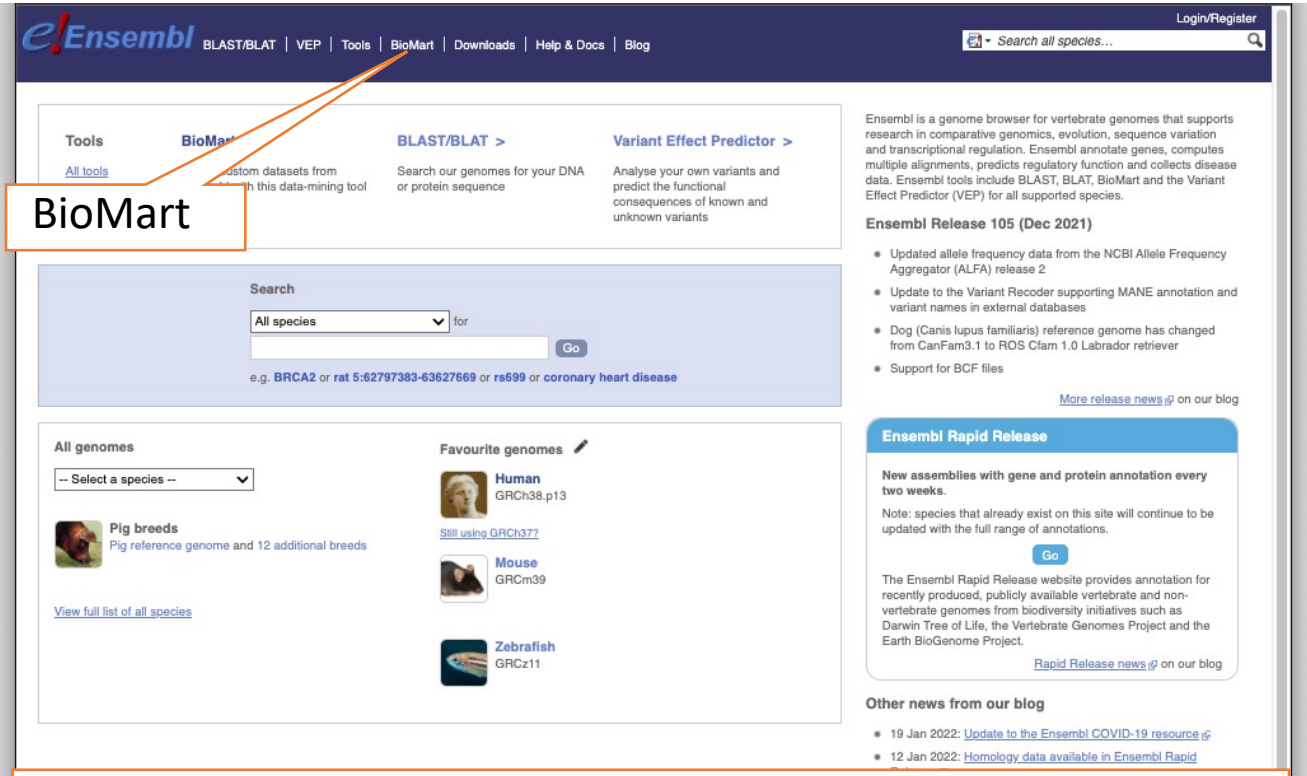
- Simple d'utilisation
- Facile à comprendre
- Une seule requête à la fois

Outil de fouille: BioMart



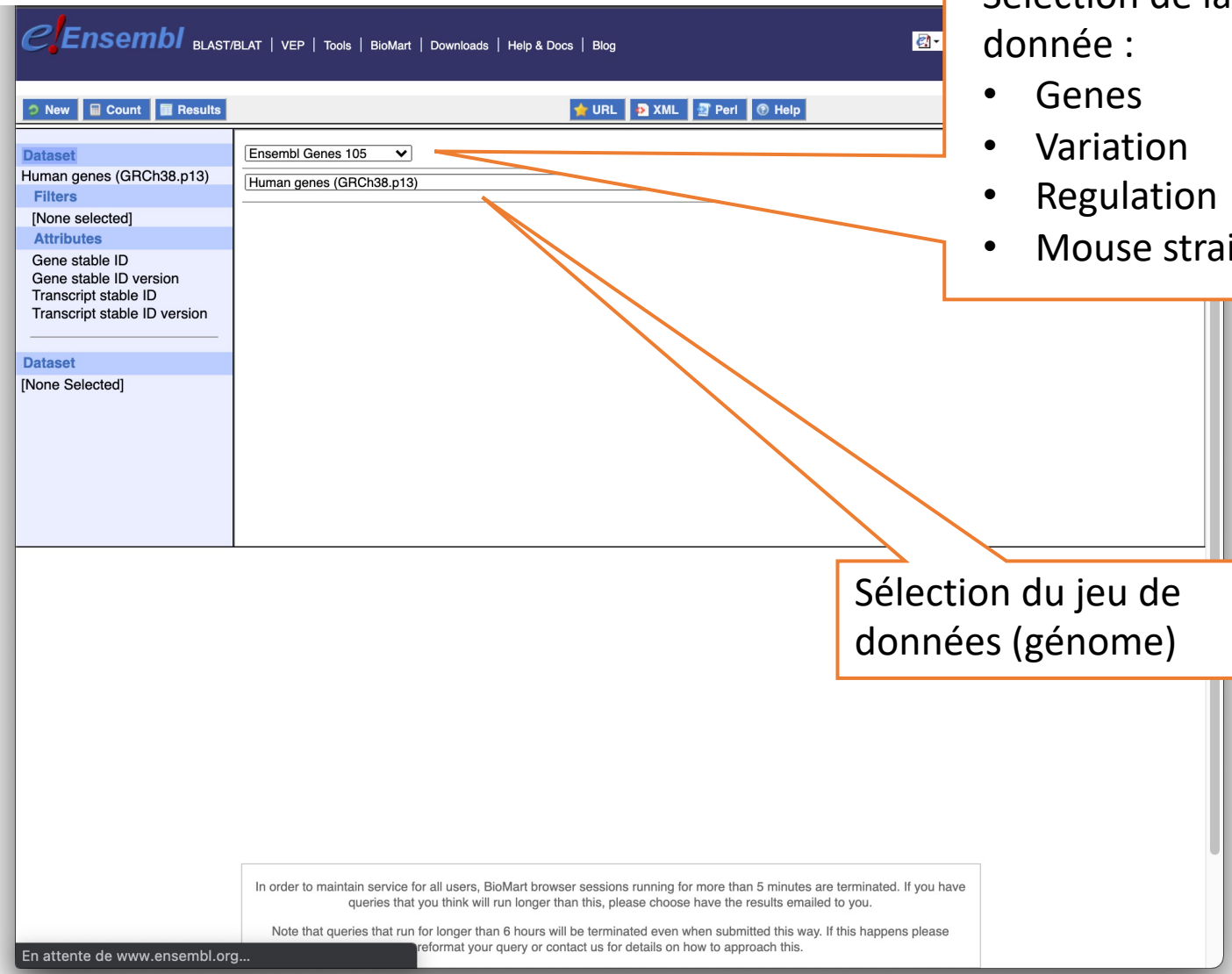
- Requête complexe
- Rapide
- Requiert une formation

BioMart/Ensembl



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

BioMart/Ensembl

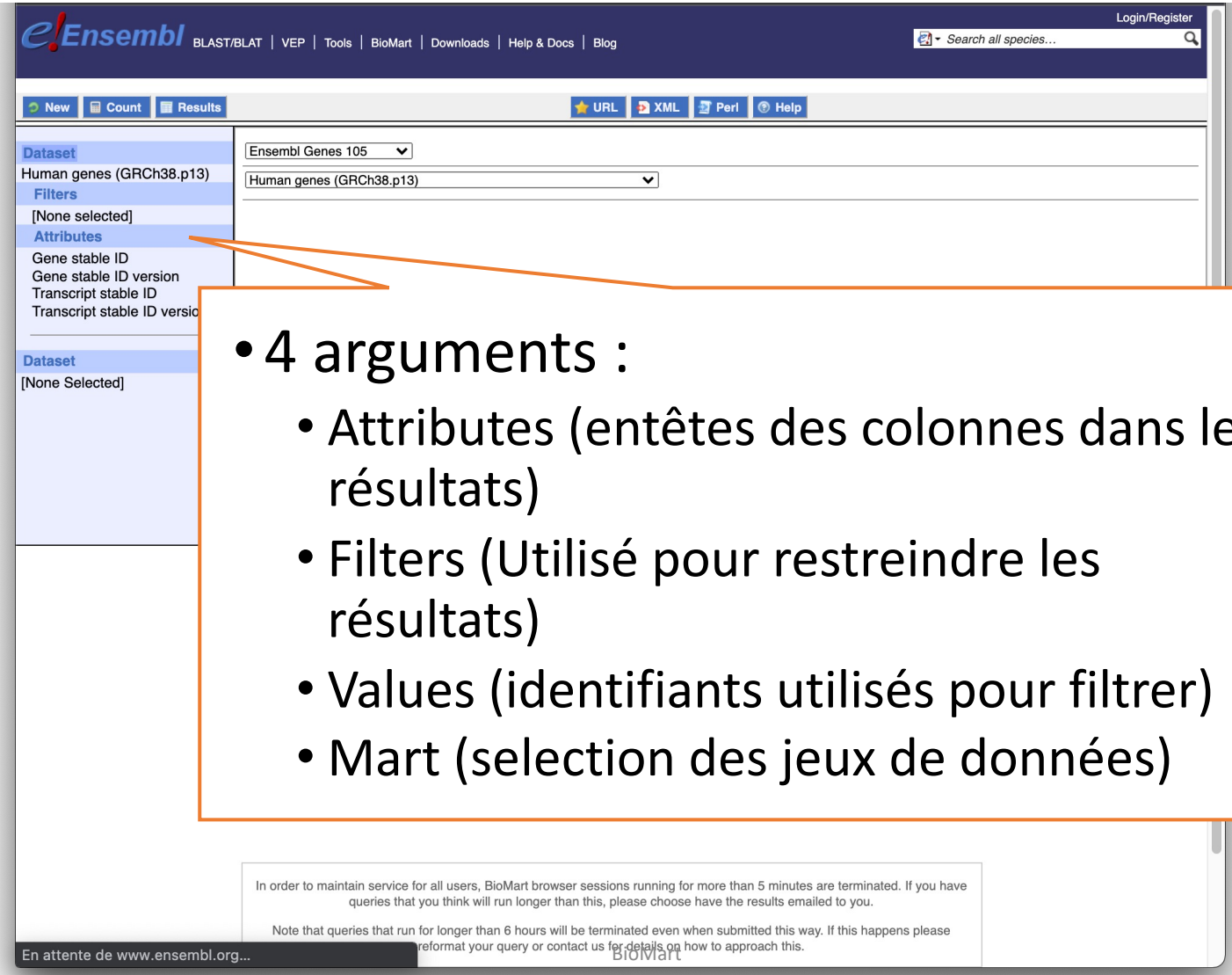


Selection de la Base de donnée :

- Genes
- Variation
- Regulation
- Mouse strain

Sélection du jeu de données (génome)

BioMart/Ensembl



- 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