Topic 9.3: Genomes

Genome Projects

Currently there are ~500,000 organisms with genomes sequenced or being sequenced.

Mostly – ~400,000 – bacteria

~20,000 viruses

~50,000 Eukarya

~200 animals

https://gold.jgi.doe.gov/index

Comparative Genomics

By comparing genomes of related animals, we may be able to see what sequences are conserved

5-10% of our genome is conserved cf. other animals

Remember: <1% of our genome is ”coding DNA”

(exons that code for protein)

→ a LOT of non-coding functional sequence information for chromosome packaging, segregation, and replication; non-coding RNA, and gene regulation

Viewing Genomes

Sites that allow us to view various genomes:

- NCBI: Genome Data Viewer
- UC Santa Cruz Genome Browser
- JBrowse http://jbrowse.org/
- Ensemble's Genomes

The free Integrative Genomics Viewer (IGV) program allows us to access the same data but runs as a standalone program.

Other web sites:

The gnomAD web site allows us to look at variations in 120,000 genomes.

The Galaxy web site has one of the largest collection of algorithms for analyzing genomes

Data in Genome Browsers

Besides the actual genomic sequences themselves, these genome browsers usually also have

- RefSeq genes (mRNA and protein)
- Ensembl genes
- Expressed sequence tags
- Gene predictions
- SNPs from a database like dbSNP
- Non-coding functional elements as found by projects like ENCoDE

Human genome data in the UCSC Genome Browser

The University of California Santa Cruz genome browser was set up to make the human genome accessible to the general public

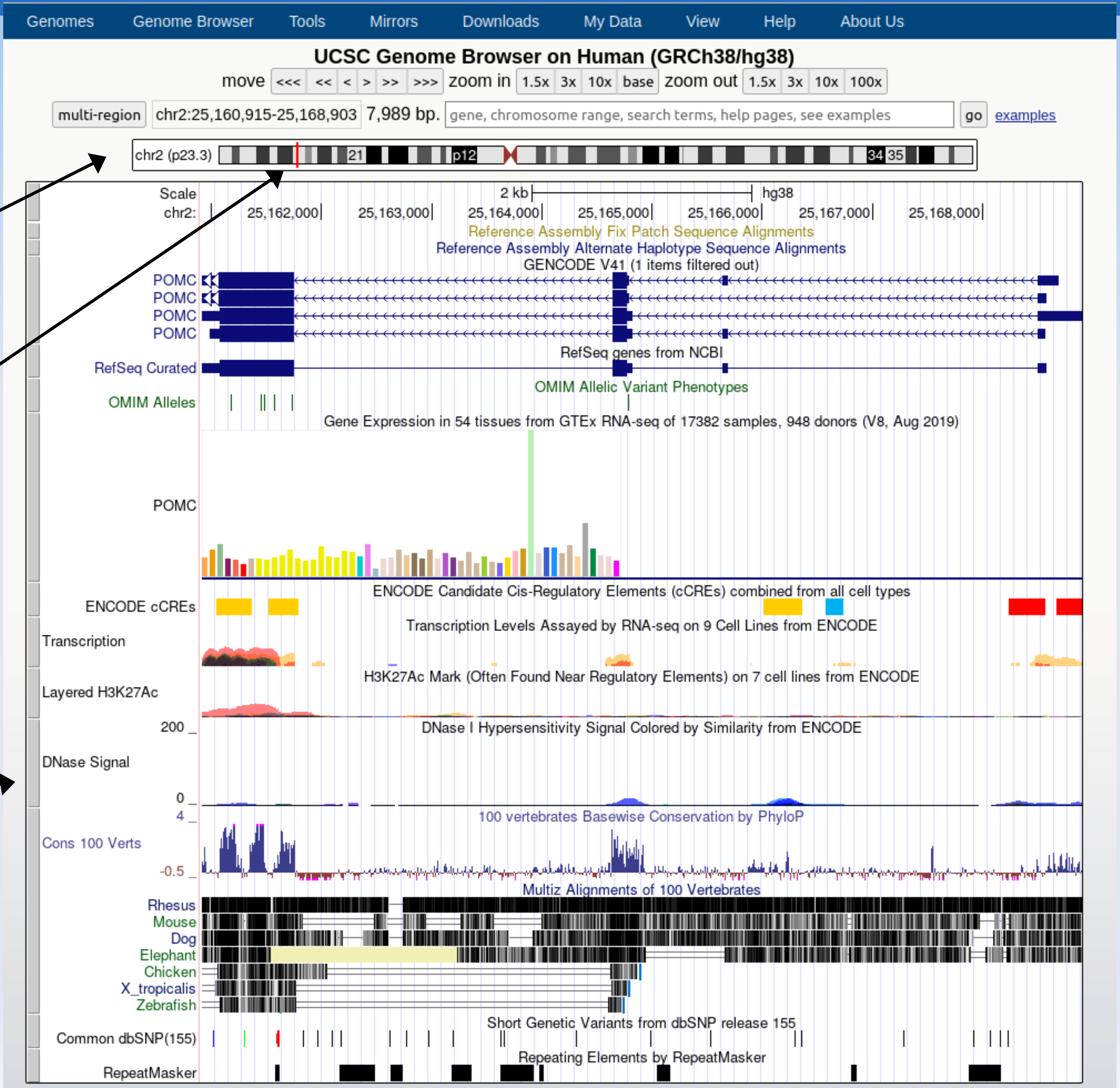
Go to: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38

UCSC Genome Browser

We will be shown data for a a particular chromosome and at a particular location:

We cannot see indivdual nucleotides because there are a lot of them in a chromosome.

Notice all the tracks of data



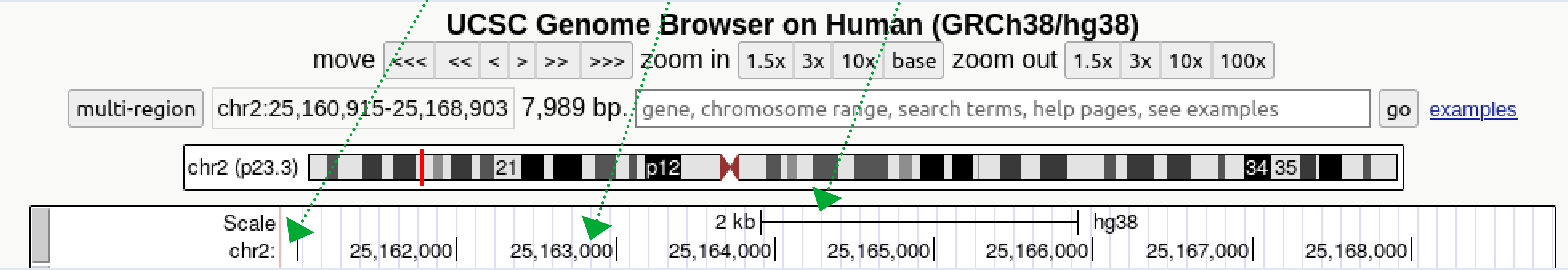
UCSC Genome Browser on Human (GRCh38/hg38)

chr2:25,160,915-25,168,903 7,989 bp.

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
gene, chromosome range, search terms, help pages, see examples
go
examples
chr2 (p23.3)
21
p12
34 35
Scale
chr2:
2 kb
hg38
25,162,000
25,163,000
25,164,000
25,165,000
25,166,000
25,167,000
25,168,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V41 (1 items filtered out)
POMC
POMC
POMC
POMC
RefSeq genes from NCBI
RefSeq Curated
OMIM Allelic Variant Phenotypes
OMIM Alleles
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
POMC
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Transcription
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
200
DNase I Hypersensitivity Signal Colored by Similarity from ENCODE
DNase Signal
0
4
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
-0.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Repeating Elements by RepeatMasker
RepeatMasker

UCSC Genome Browser Tracks

The first track displays the location (in bases) in a particular chromosome and a scale

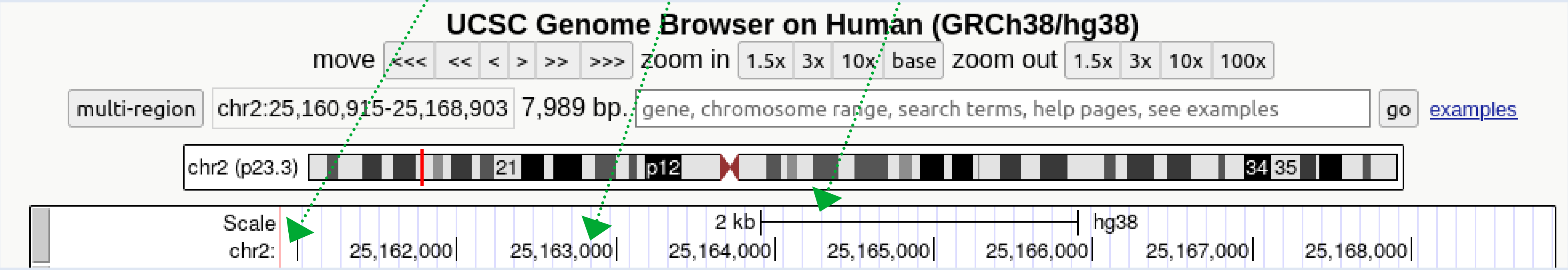


UCSC Genome Browser on Human (GRCh38/hg38)

chr2:25,160,915-25,168,903 7,989 bp.

UCSC Genome Browser Tracks

The first track displays the location (in bases) in a particular chromosome and a scale

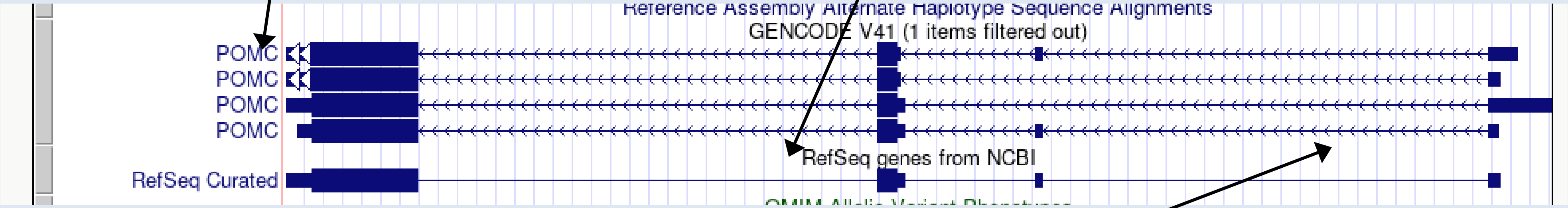


UCSC Genome Browser on Human (GRCh38/hg38)

chr2:25,160,915-25,168,903 7,989 bp.

UCSC Genome Browser Tracks

There are usually tracks displaying information about genes in this area of the chromosome, including, if available, curated RefSeq data

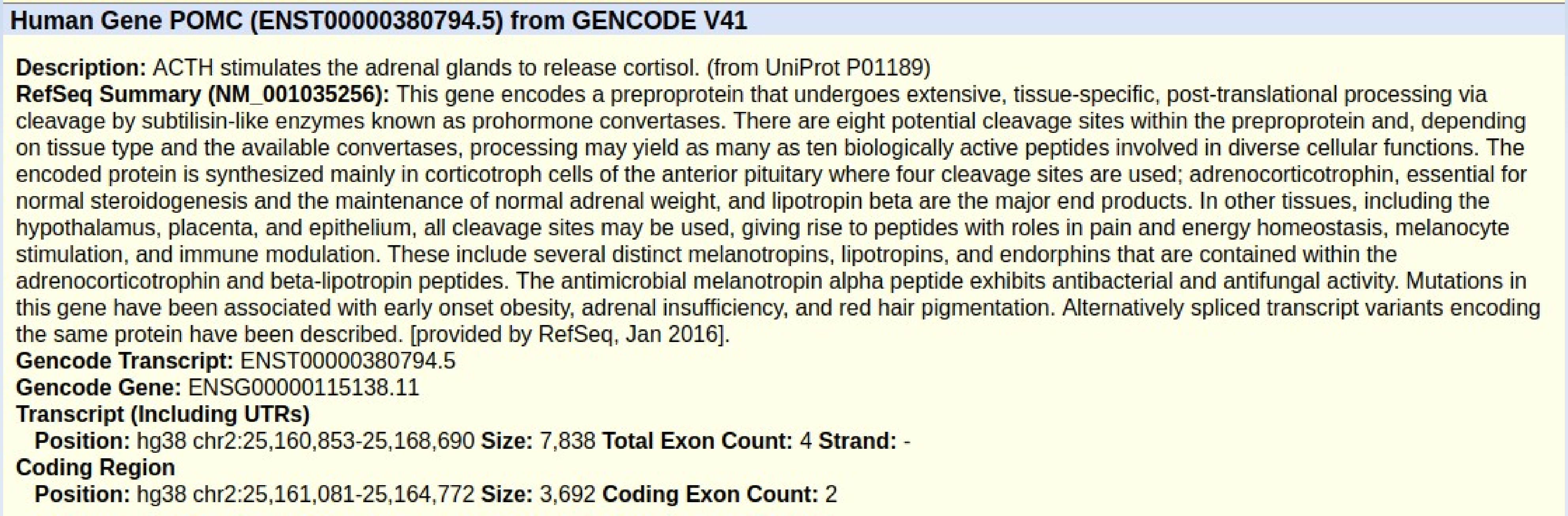

Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V41 (1 items filtered out)
POMC
POMC
POMC
POMC
RefSeq genes from NCBI
RefSeq Curated
OMIM Allelic Variant Phenotypes

In this case, we see that this particular gene is on the complementary strand

Clicking on the gene name will take us to a page displaying more details about the gene

UCSC Genome Browser Tracks

Clicking on the gene name will take us to a page displaying more details about the gene

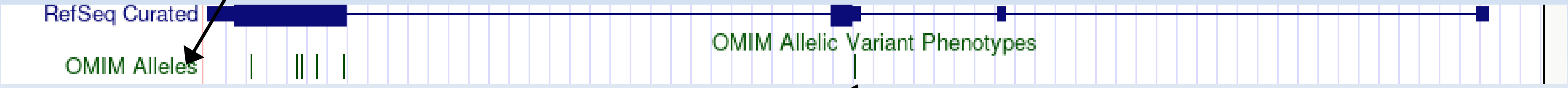


Human Gene POMC (ENST00000380794.5) from GENCODE V41

Description: ACTH stimulates the adrenal glands to release cortisol. (from UniProt P01189)
RefSeq Summary (NM_001035256): This gene encodes a preproprotein that undergoes extensive, tissue-specific, post-translational processing via cleavage by subtilisin-like enzymes known as prohormone convertases. There are eight potential cleavage sites within the preproprotein and, depending on tissue type and the available convertases, processing may yield as many as ten biologically active peptides involved in diverse cellular functions. The encoded protein is synthesized mainly in corticotroph cells of the anterior pituitary where four cleavage sites are used; adrenocorticotrophin, essential for normal steroidogenesis and the maintenance of normal adrenal weight, and lipotropin beta are the major end products. In other tissues, including the hypothalamus, placenta, and epithelium, all cleavage sites may be used, giving rise to peptides with roles in pain and energy homeostasis, melanocyte stimulation, and immune modulation. These include several distinct melanotropins, lipotropins, and endorphins that are contained within the adrenocorticotrophin and beta-lipotropin peptides. The antimicrobial melanotropin alpha peptide exhibits antibacterial and antifungal activity. Mutations in this gene have been associated with early onset obesity, adrenal insufficiency, and red hair pigmentation. Alternatively spliced transcript variants encoding the same protein have been described. [provided by RefSeq, Jan 2016].
Gencode Transcript: ENST00000380794.5
Gencode Gene: ENSG00000115138.11
Transcript (Including UTRs)
Position: hg38 chr2:25,160,853-25,168,690 **Size:** 7,838 **Total Exon Count:** 4 **Strand:** -
Coding Region
Position: hg38 chr2:25,161,081-25,164,772 **Size:** 3,692 **Coding Exon Count:** 2

UCSC Genome Browser Tracks

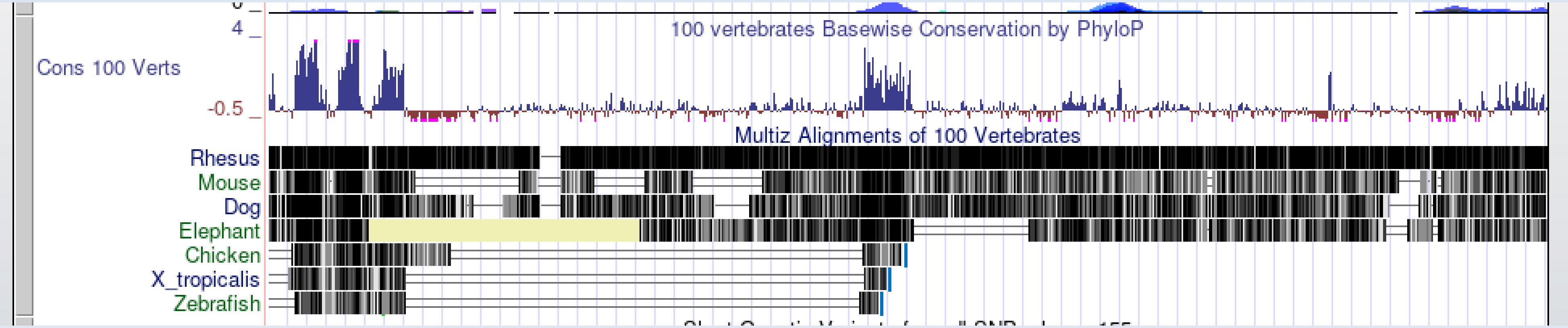
There may be an OMIM track displaying information about allelic variations



The bars represent variations that may be important in determining phenotypes or inherited diseases

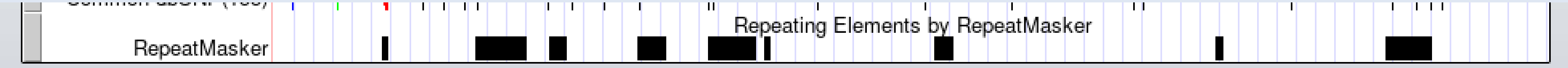
UCSC Genome Browser Tracks

A Conservation track displays best guesses for aligning this stretch of human genomic DNA with those of other vertebrates:


4
Cons 100 Verts
-0.5
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish

UCSC Genome Browser Tracks

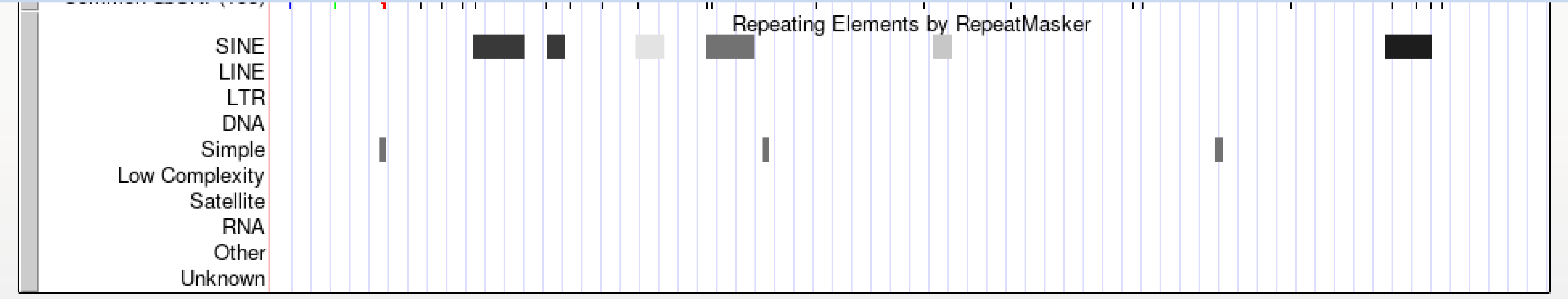
The last of the default tracks is often the RepeatMasker track showing any detected repeat sequences, which are mostly in non-coding parts of the genome.



Repeat sequences are genomic DNA with short and long interspersed nuclear elements (SINE and LINE), long terminal repeat elements (LTR), and other “low complexity” areas with short repeating sequences.

UCSC Genome Browser Tracks

RepeatMasker track: Click on the “RepeatMasker” link to reveal computationally detected repeat sequences.



Here we can get more details of repeat sequences - genomic DNA with

- short and long interspersed nuclear elements (SINE and LINE)
- long terminal repeat elements (LTR)
- other “low complexity” areas with short repeating sequences.

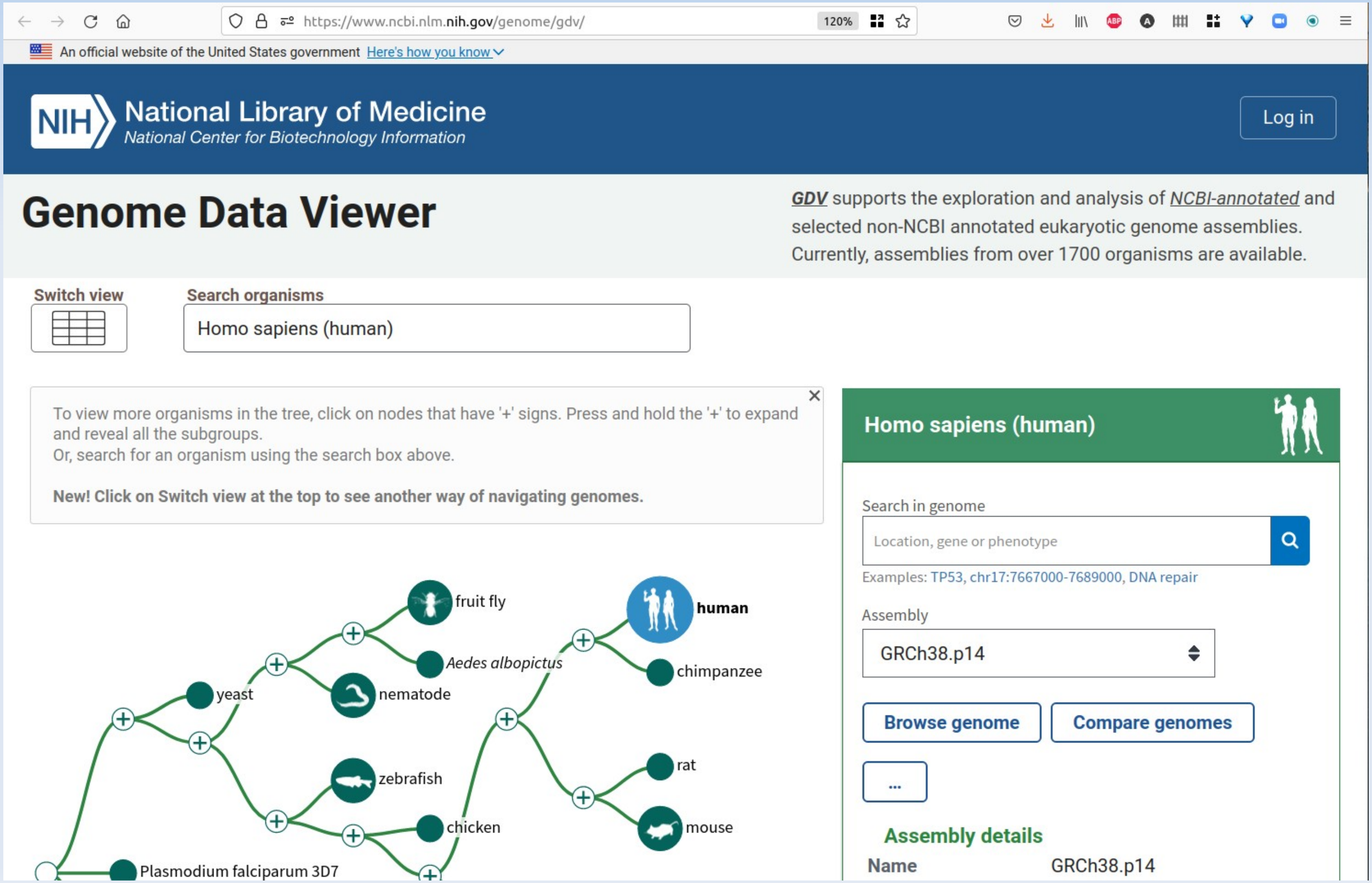
When performing sequence alignments, we can disable these repeat sequences from appearing in our alignments.

Zebra Mussel data in the NCBI Genome Data Viewer

Go to: https://www.ncbi.nlm.nih.gov/genome/gdv/

Zebra Mussel data in the NCBI Genome Data Viewer

Go to: https://www.ncbi.nlm.nih.gov/genome/gdv/



https://www.ncbi.nlm.nih.gov/genome/gdv/

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 1700 organisms are available.

Switch view

Search organisms: Homo sapiens (human)

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups.
Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

Homo sapiens (human)

Search in genome: Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, DNA repair

Assembly: GRCh38.p14

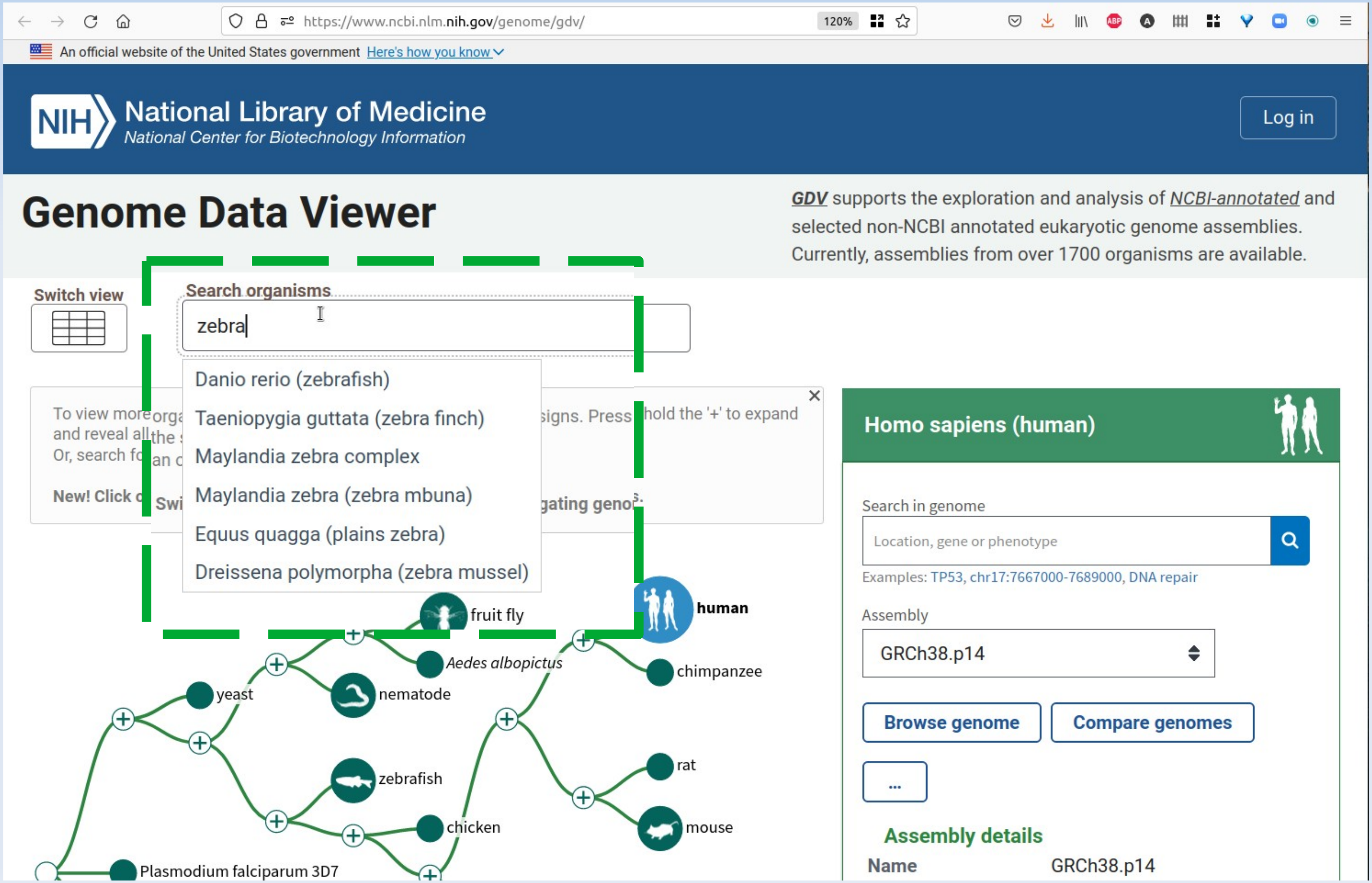
Browse genome | Compare genomes | ...

Assembly details

Name GRCh38.p14

Zebra Mussel data in the NCBI Genome Data Viewer

In the ”Search organisms” box, enter Zebra Mussel



https://www.ncbi.nlm.nih.gov/genome/gdv/

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Genome Data Viewer

GDV supports the exploration and analysis of NCBI-annotated and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 1700 organisms are available.

Search organisms: zebra

- Danio rerio (zebrafish)
- Taeniopygia guttata (zebra finch)
- Maylandia zebra complex
- Maylandia zebra (zebra mbuna)
- Equus quagga (plains zebra)
- Dreissena polymorpha (zebra mussel)

Homo sapiens (human)

Search in genome: Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, DNA repair

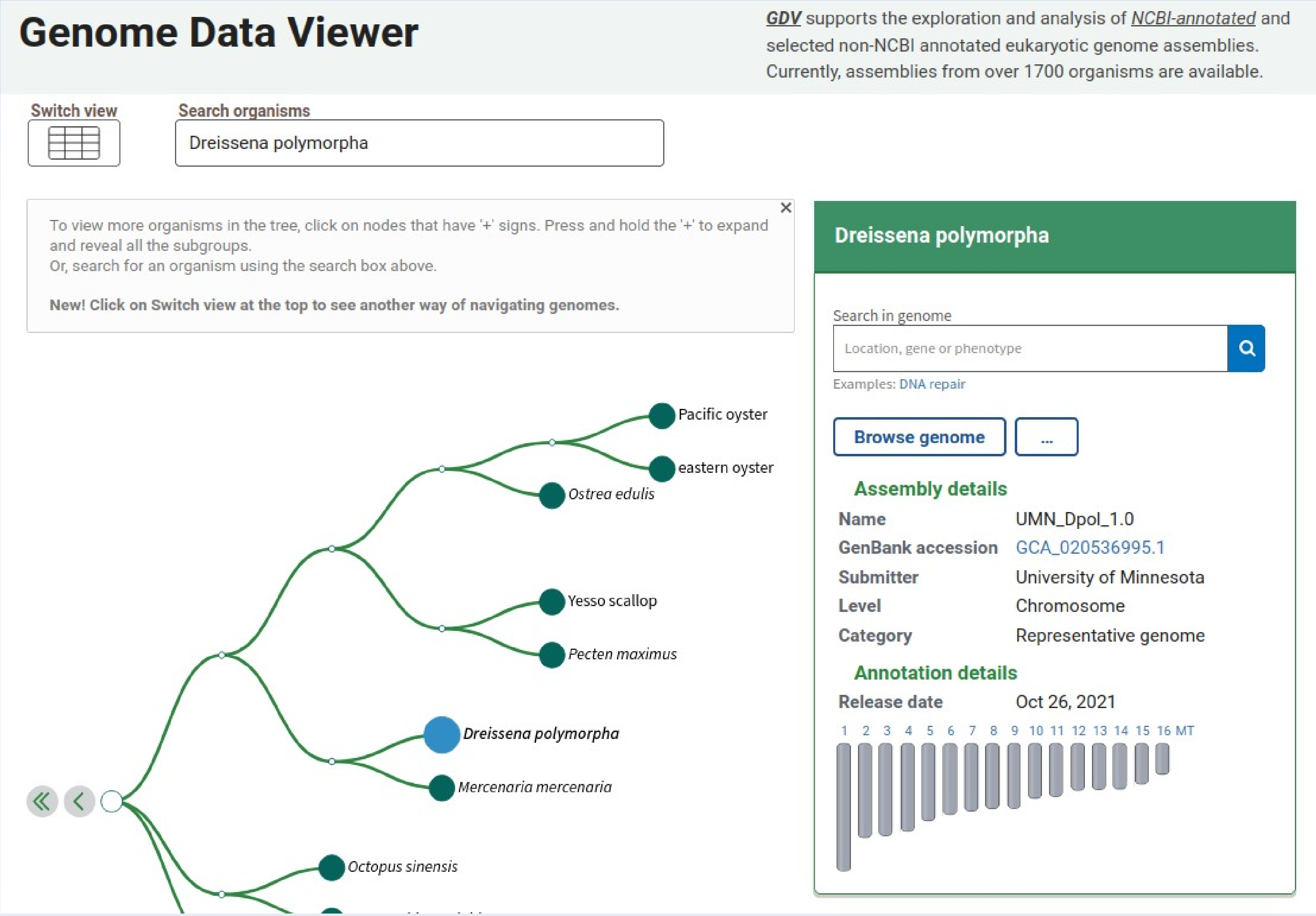
Assembly: GRCh38.p14

Assembly details

Name	GRCh38.p14

Zebra Mussel data in the NCBI Genome Data Viewer

It will show up as ”Dreissena polymorpha”



Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 1700 organisms are available.

Switch view

Search organisms: Dreissena polymorpha

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups.
Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

Dreissena polymorpha

Search in genome: Location, gene or phenotype

Examples: DNA repair

Browse genome | ...

Assembly details

Name	UMN_Dpol_1.0
GenBank accession	GCA_020536995.1
Submitter	University of Minnesota
Level	Chromosome
Category	Representative genome

Annotation details

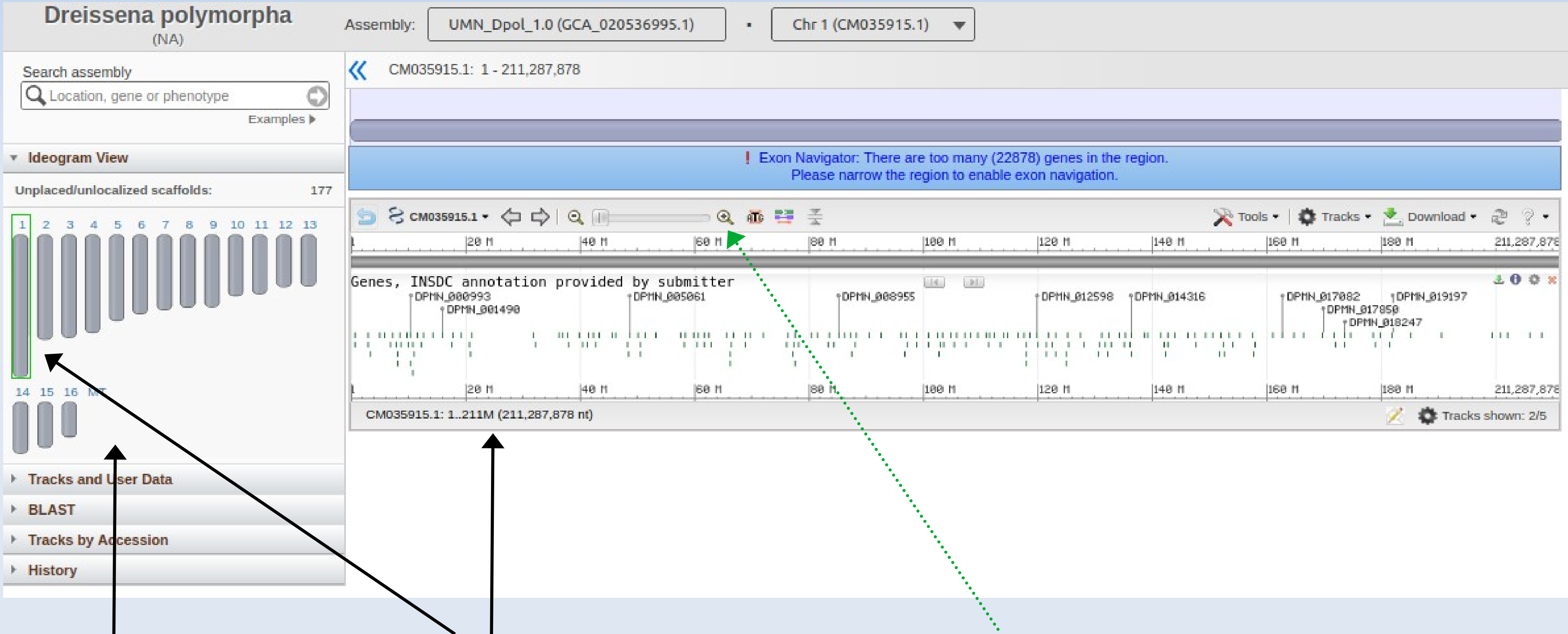
Release date	Oct 26, 2021

16 chromosomes plus a mitochondrial genome

Click Browse genome

Zebra Mussel data in the NCBI Genome Data Viewer

Genome Data Viewer will show us some of the details of Chromosome 1:



Dreissena polymorpha (NA)

Assembly: UMN_Dpol_1.0 (GCA_020536995.1) · Chr 1 (CM035915.1)

Search assembly: Location, gene or phenotype

CM035915.1: 1 - 211,287,878

! Exon Navigator: There are too many (22878) genes in the region. Please narrow the region to enable exon navigation.

Genes, INSDC annotation provided by submitter

CM035915.1: 1..211M (211,287,878 nt)

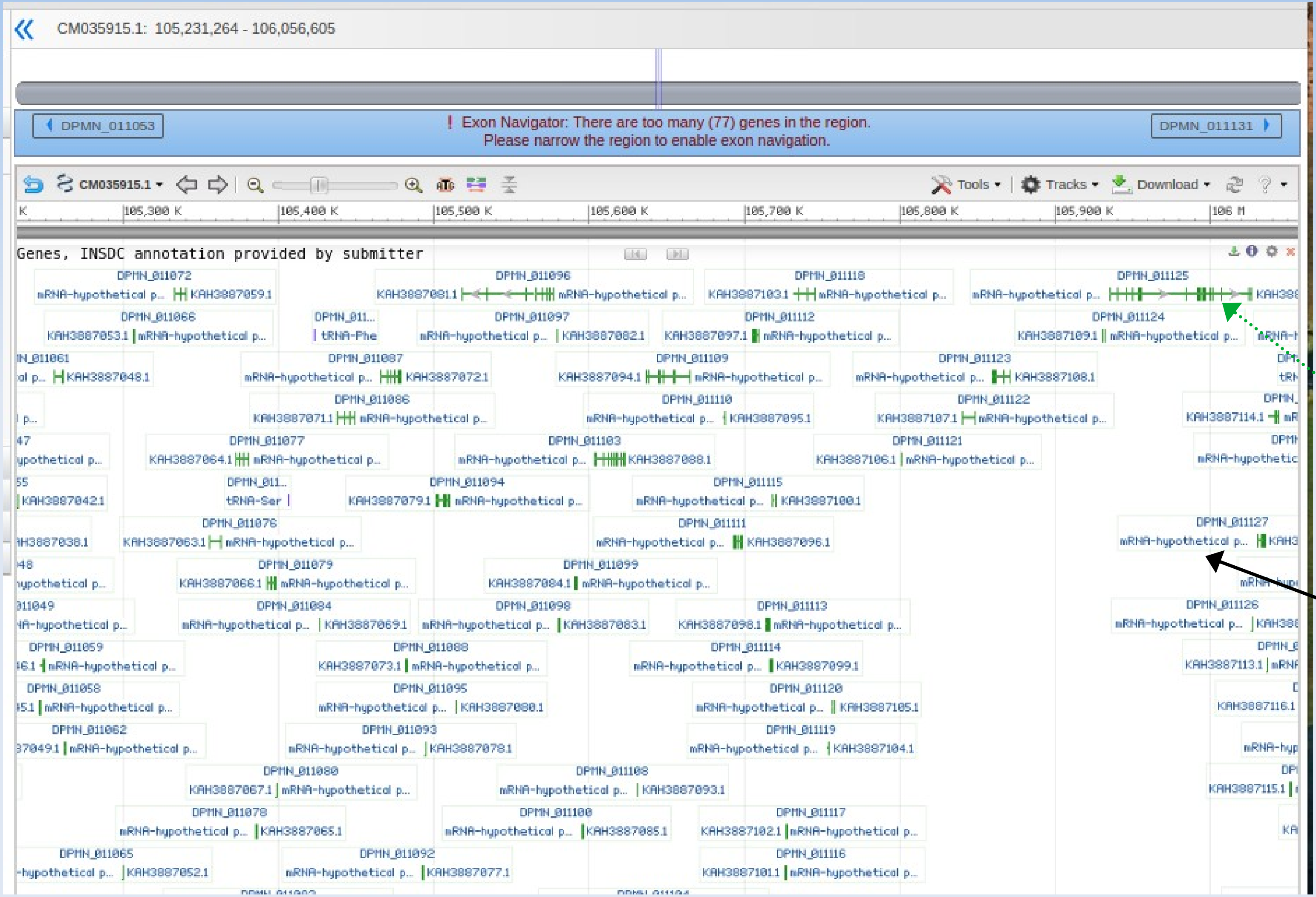
Tracks shown: 2/5

16 chromosomes plus a mitochondrial genome

Chromosome 1 is selected.
The view defaults to all 211 million bases.

Use the zoom controls to select less than about a million bases

Zebra Mussel data in the NCBI Genome Data Viewer



CM035915.1: 105,231,264 - 106,056,605

! Exon Navigator: There are too many (77) genes in the region. Please narrow the region to enable exon navigation.

Genes, INSDC annotation provided by submitter

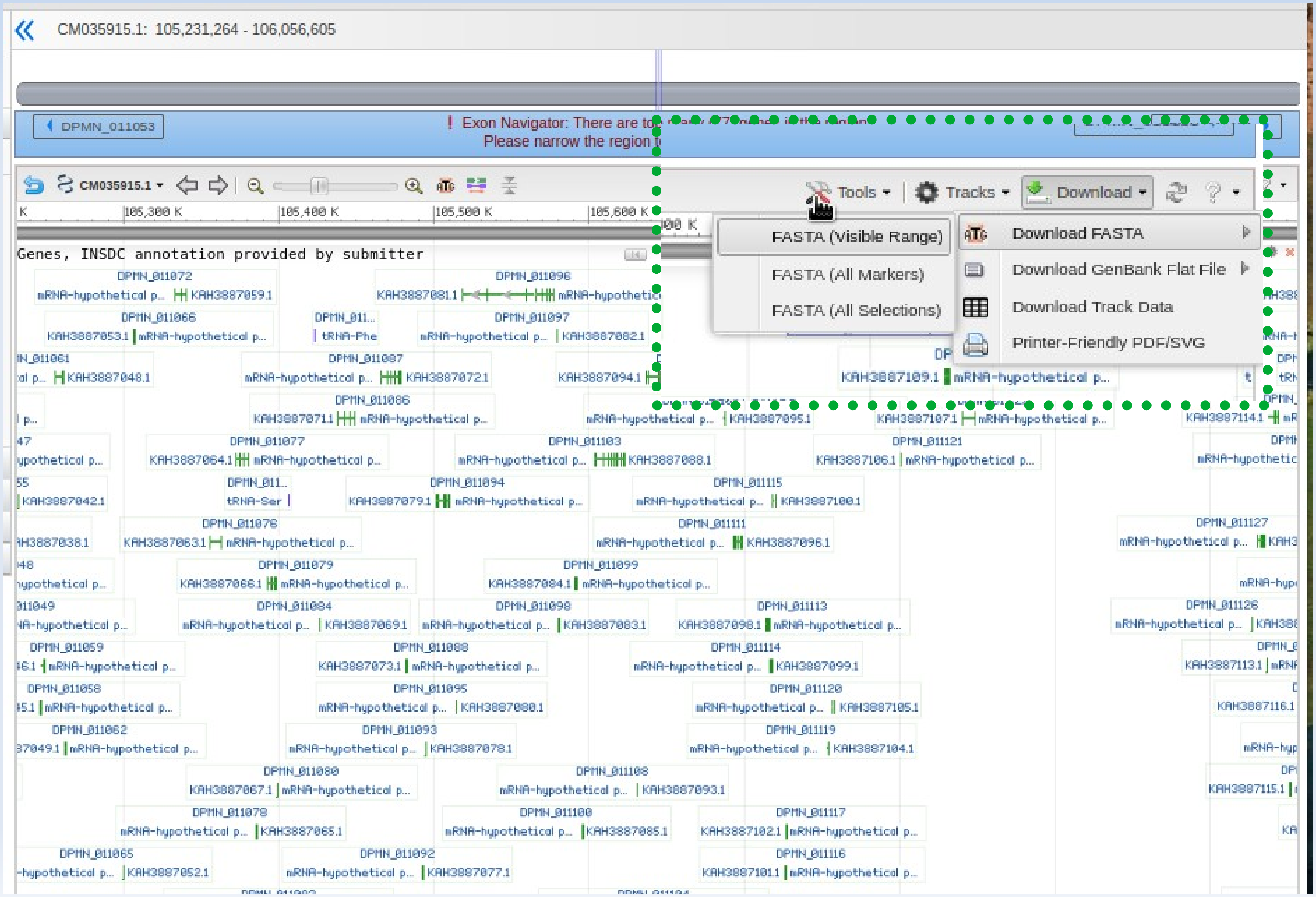
We can see exons and introns

A lot of the genes are marked as ”mRNA hypothetical” because they were detected using computational methods

At this resolution, we will see a lot more tracks

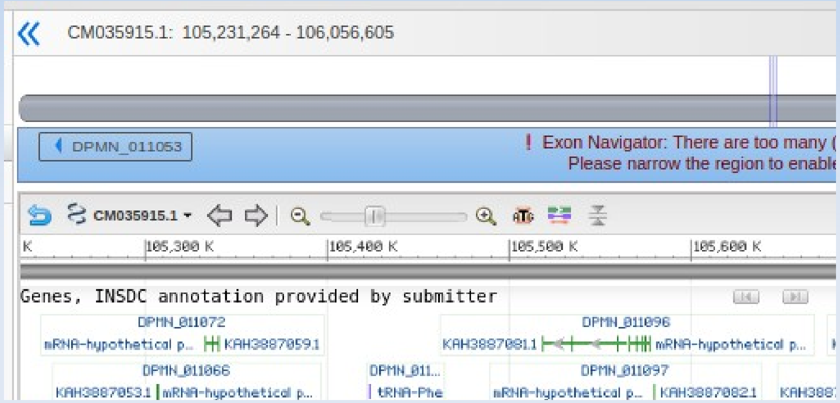
Zebra Mussel data in the NCBI Genome Data Viewer

We can download the nucleotides in this view as a Fasta file. Click on Download → Download Fasta → Fasta (Visible Range)



Zebra Mussel data in the NCBI Genome Data Viewer

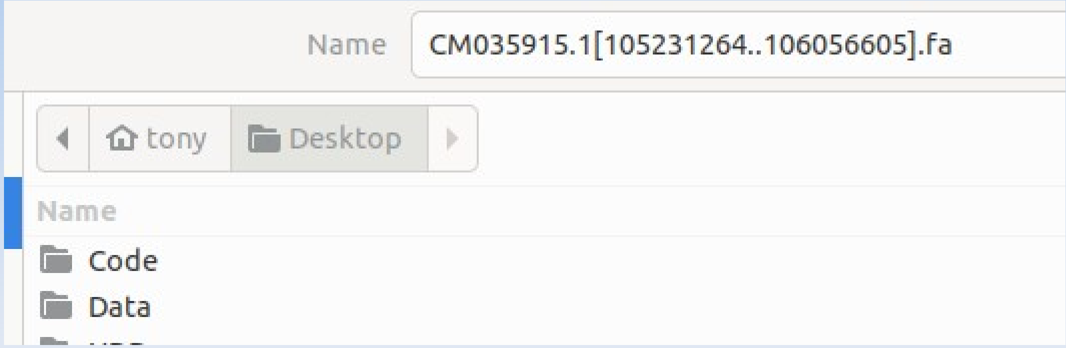
We will see a window asking us where to save a file with a name that contains the chromosome number and the range of nucleotides.



CM035915.1: 105,231,264 - 106,056,605

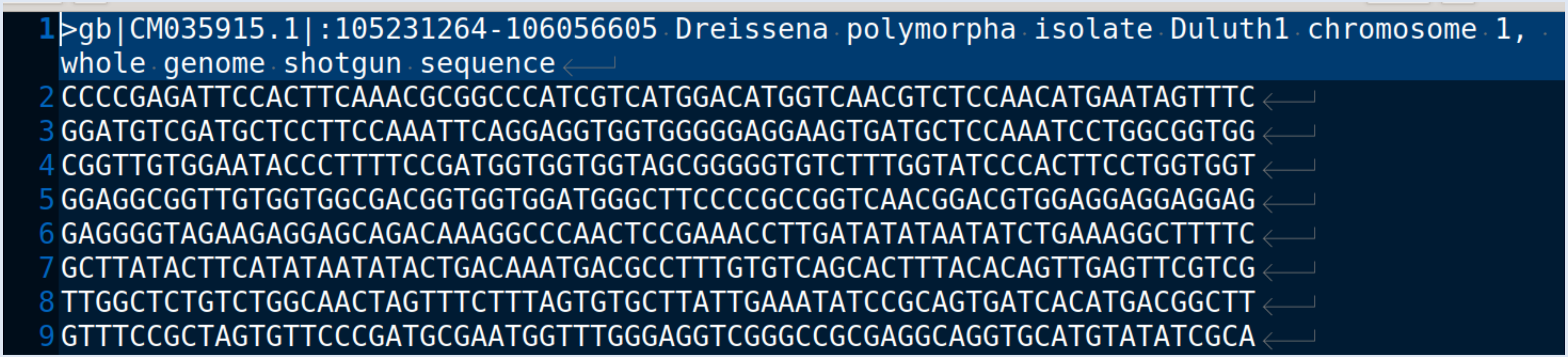
! Exon Navigator: There are too many (
Please narrow the region to enable

Genes, INSDC annotation provided by submitter



Name CM035915.1[105231264..106056605].fa

We can open up the file using a text editor like Notepad or Notepad++



```
>gb|CM035915.1|:105231264-106056605 Dreissena polymorpha isolate Duluth1 chromosome 1, whole genome shotgun sequence
CCCCGAGATTCCACTTCAAACGCGGCCCATCGTCATGGACATGGTCAACGTCTCCAACATGAATAGTTTC
GGATGTCGATGCTCCTTCCAAATTCAGGAGGTGGTGGGGGAGGAAGTGATGCTCCAAATCCTGGCGGTGG
CGGTTGTGGAATACCCTTTTCCGATGGTGGTGGTAGCGGGGGTGTCTTTGGTATCCCACTTCCTGGTGGT
GGAGGCGGTTGTGGTGGCGACGGTGGTGGATGGGCTTCCCCGCCGGTCAACGGACGTGGAGGAGGAGGAG
GAGGGGTAGAAGAGGAGCAGACAAAGGCCCAACTCCGAAACCTTGATATATAATATCTGAAAGGCTTTTC
GCTTATACTTCATATAATATACTGACAAATGACGCCTTTGTGTCAGCACTTTACACAGTTGAGTTCGTCG
TTGGCTCTGTCTGGCAACTAGTTTCTTTAGTGTGCTTATTGAAATATCCGCAGTGATCACATGACGGCTT
GTTTCCGCTAGTGTTCCCGATGCGAATGGTTTGGGAGGTCGGGCCGCGAGGCAGGTGCATGTATATCGCA
```

Viewing gene promoter data in the Ensemble site

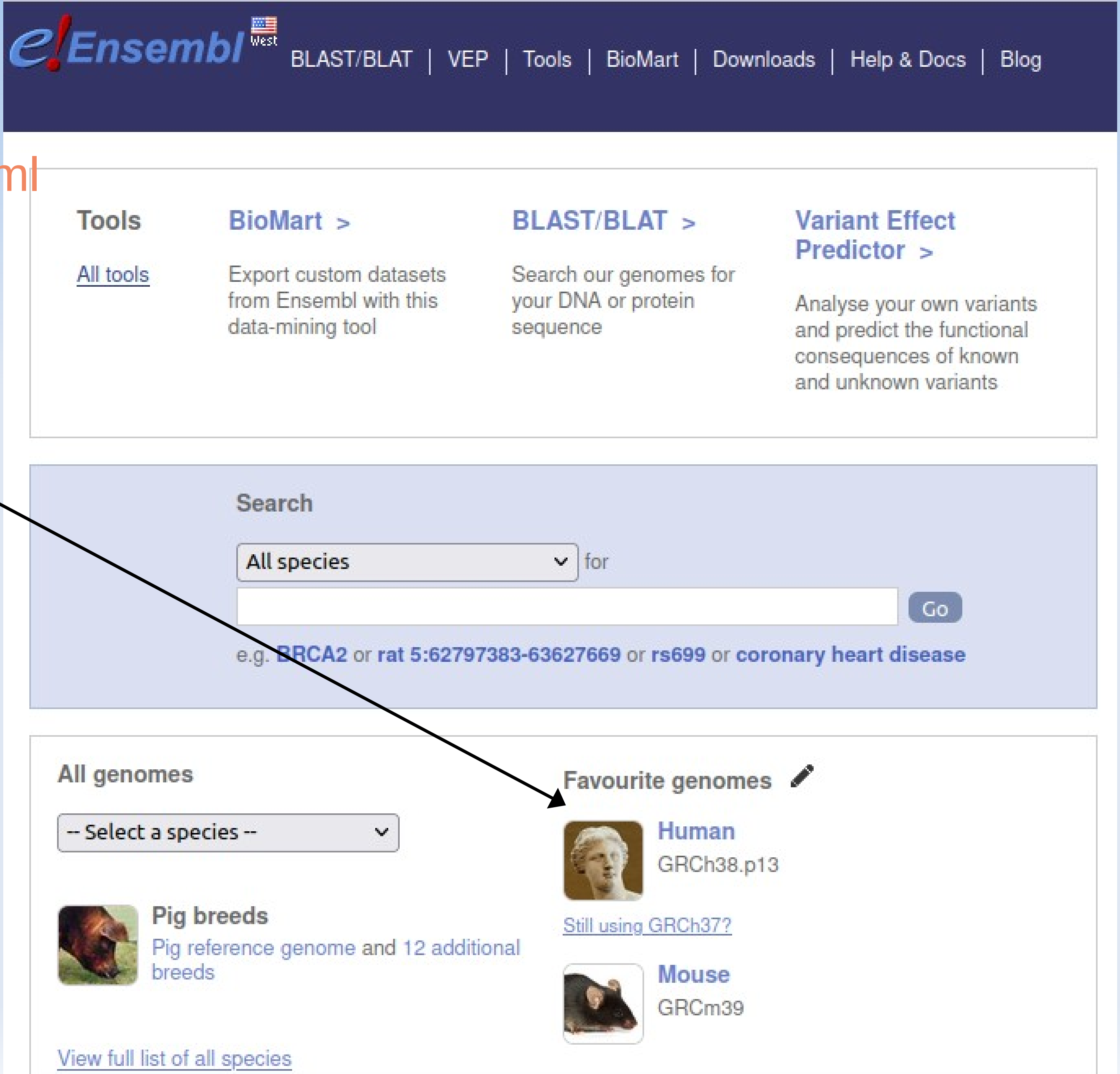
Go to: http://ensembl.org/index.html

Promoter data in Ensemble

Go to:

http://ensembl.org/index.html

Click ”Human”



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

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

Go

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

All genomes

-- Select a species --

Pig breeds

Pig reference genome and 12 additional breeds

View full list of all species

Favourite genomes

Human

GRCh38.p13

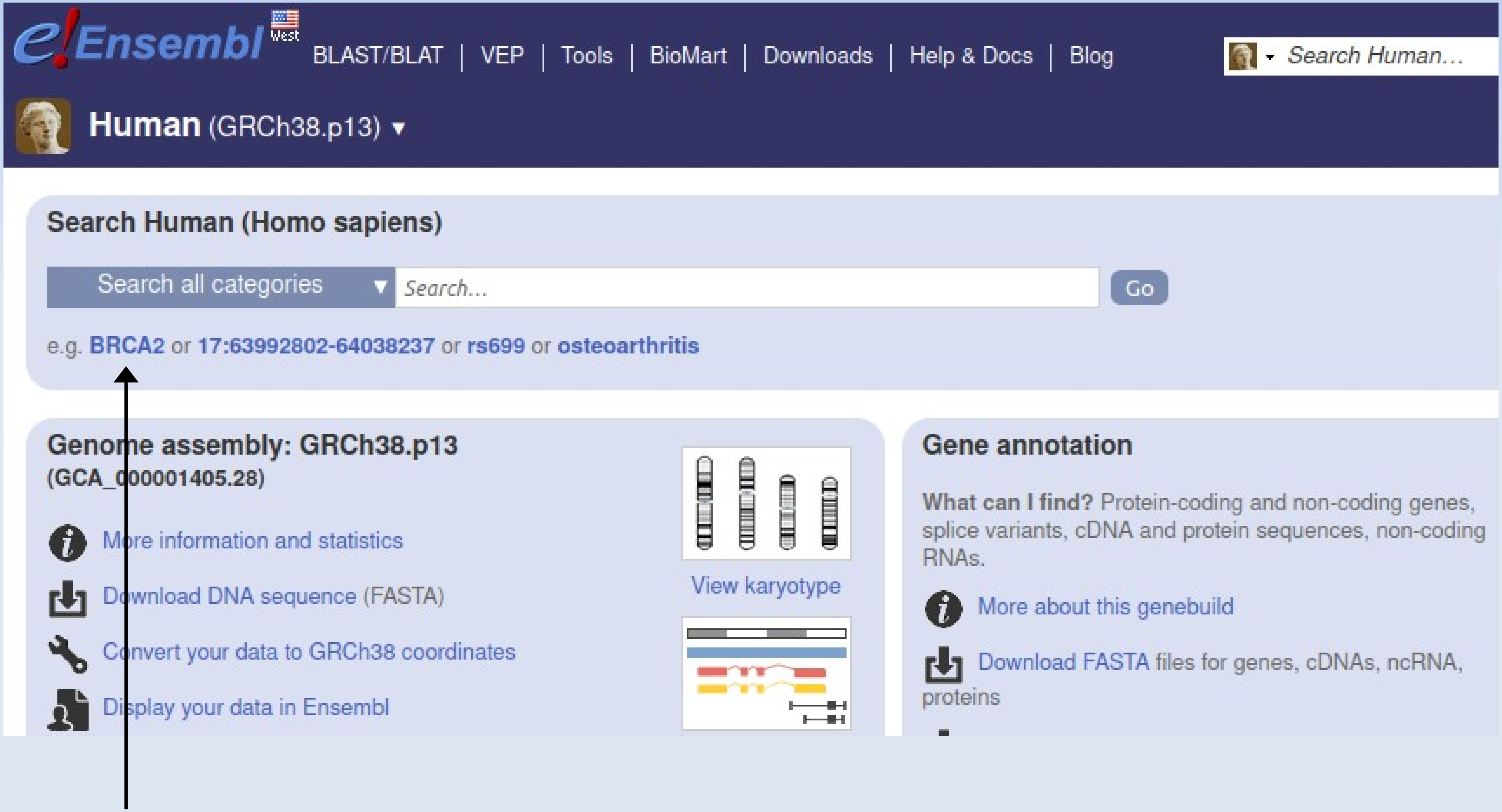
Still using GRCh37?

Mouse

GRCm39

Promoter data in Ensemble

A lot of options are presented



Ensembl West
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Search Human…
Human (GRCh38.p13) ▾

Search Human (Homo sapiens)

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

Genome assembly: GRCh38.p13 (GCA_000001405.28)

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

Gene annotation

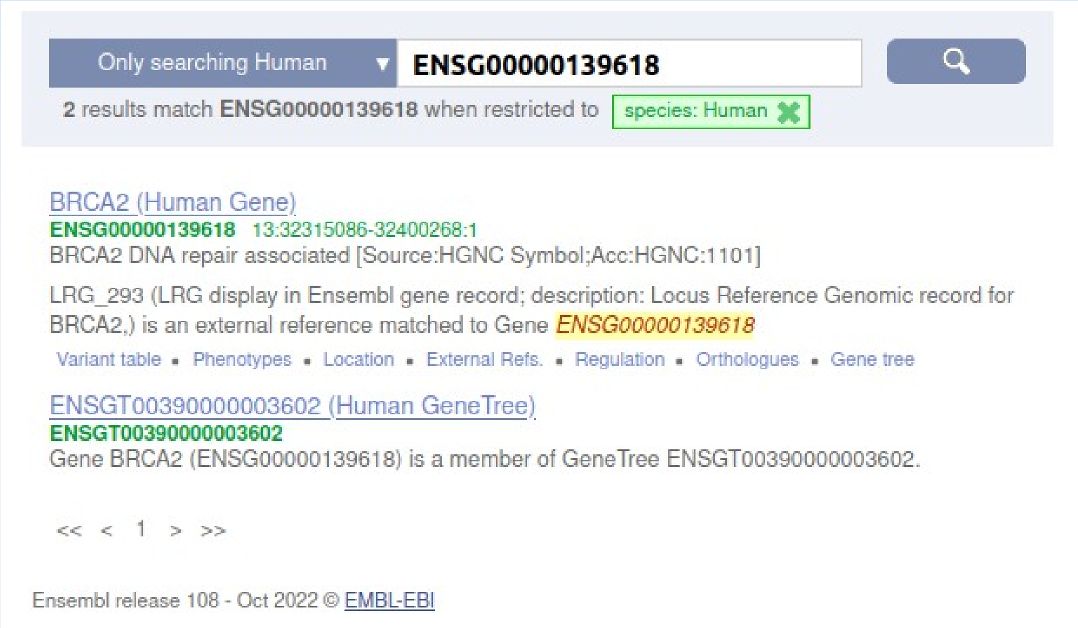
What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins

Click ”BRCA2” or search for a particular gene

Promoter data in Ensemble

Click the first search result



Only searching Human ▾ ENSG00000139618

2 results match **ENSG00000139618** when restricted to species: Human ✖

BRCA2 (Human Gene)
ENSG00000139618 13:32315086-32400268:1
BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

LRG_293 (LRG display in Ensembl gene record; description: Locus Reference Genomic record for BRCA2,) is an external reference matched to Gene *ENSG00000139618*

Variant table ▪ Phenotypes ▪ Location ▪ External Refs. ▪ Regulation ▪ Orthologues ▪ Gene tree

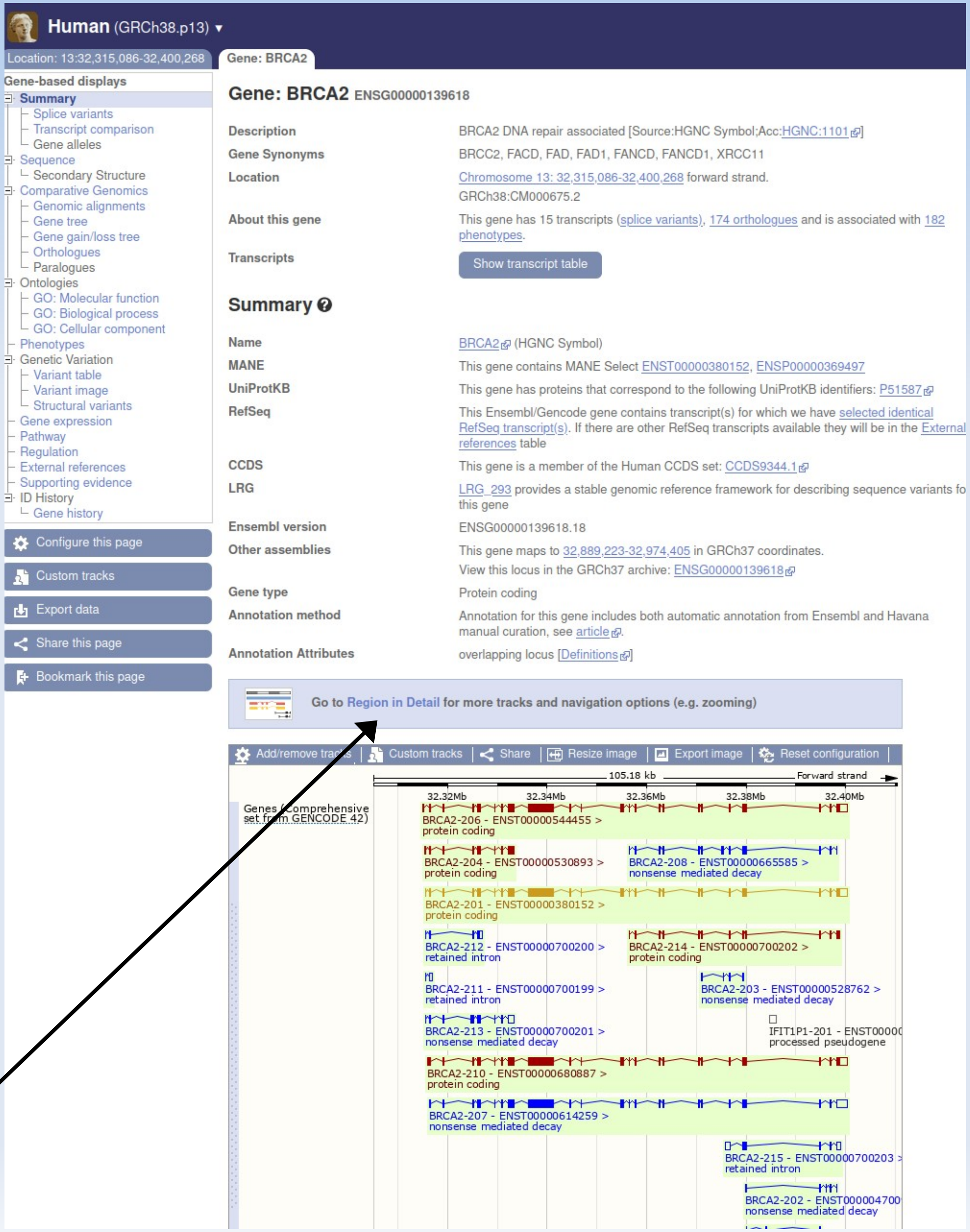
ENSGT00390000003602 (Human GeneTree)
ENSGT00390000003602
Gene BRCA2 (ENSG00000139618) is a member of GeneTree ENSGT00390000003602.

<< < 1 > >>

Ensembl release 108 - Oct 2022 © EMBL-EBI

We will get a page with

a lot of data:



Human (GRCh38.p13) ▾

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

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

Configure this page
Custom tracks
Export data
Share this page
Bookmark this page

Gene: BRCA2 ENSG00000139618

Description	BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
Gene Synonyms	BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11
Location	Chromosome 13: 32,315,086-32,400,268 forward strand. GRCh38:CM000675.2
About this gene	This gene has 15 transcripts (splice variants), 174 orthologues and is associated with 182 phenotypes.
Transcripts	Show transcript table

Summary ❓

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

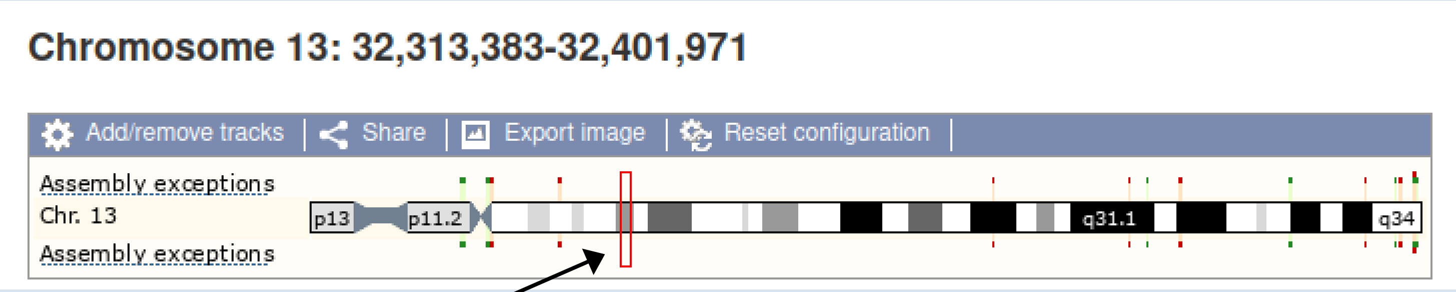
Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Add/remove tracks
Custom tracks
Share
Resize image
Export image
Reset configuration
105.18 kb
Forward strand
32.32Mb
32.34Mb
32.36Mb
32.38Mb
32.40Mb
Genes (Comprehensive set from GENCODE 42)
BRCA2-206 - ENST00000544455 >
protein coding
BRCA2-204 - ENST00000530893 >
protein coding
BRCA2-208 - ENST00000665585 >
nonsense mediated decay
BRCA2-201 - ENST00000380152 >
protein coding
BRCA2-212 - ENST00000700200 >
retained intron
BRCA2-214 - ENST00000700202 >
protein coding
BRCA2-211 - ENST00000700199 >
retained intron
BRCA2-203 - ENST00000528762 >
nonsense mediated decay
BRCA2-213 - ENST00000700201 >
nonsense mediated decay
IFIT1P1-201 - ENST0000C
processed pseudogene
BRCA2-210 - ENST00000680887 >
protein coding
BRCA2-207 - ENST00000614259 >
nonsense mediated decay
BRCA2-215 - ENST00000700203 >
retained intron
BRCA2-202 - ENST000004700
nonsense mediated decay

Click on ”Region in detail”

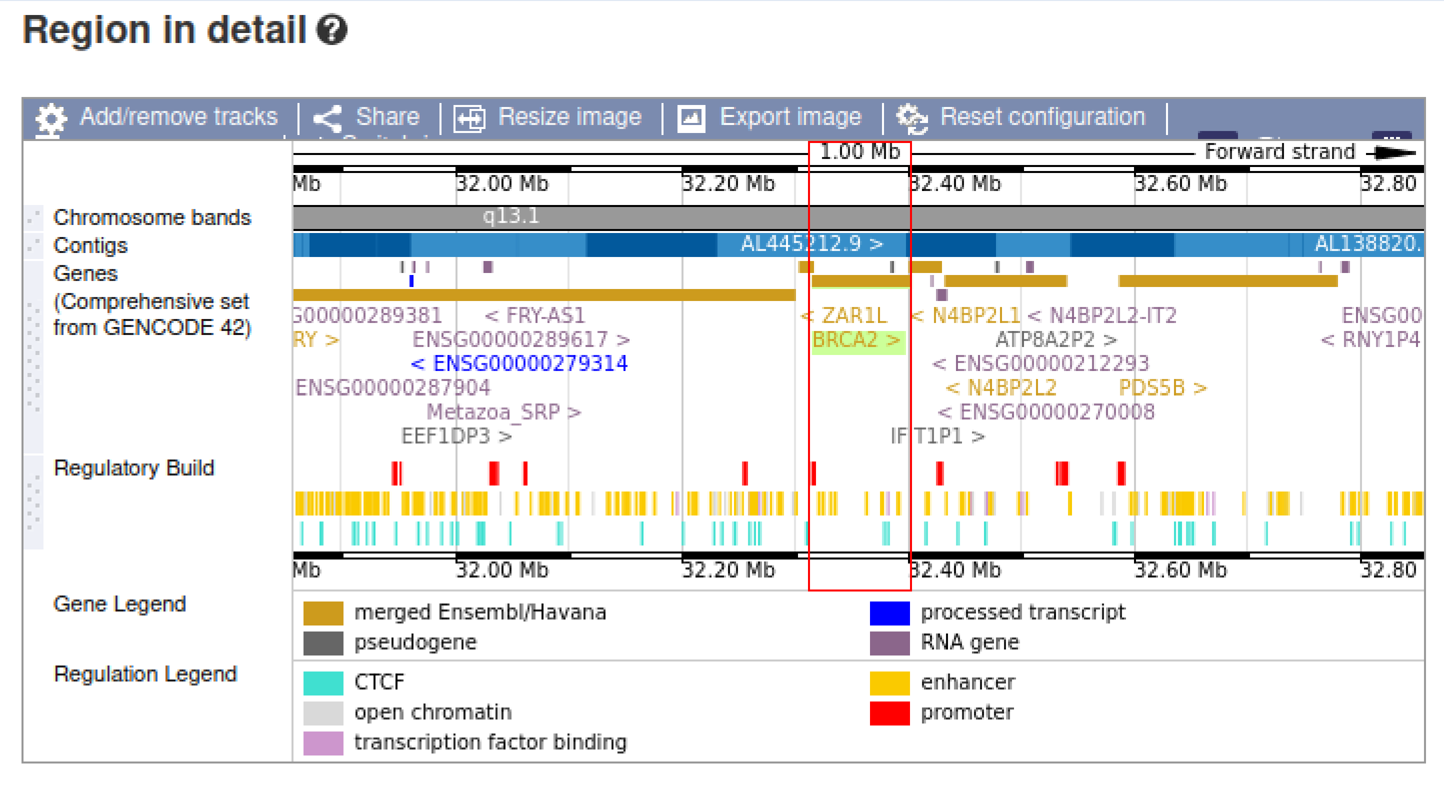
Promoter data in Ensemble

We will get a page with a view of the chromosome:



Chromosome 13: 32,313,383-32,401,971

A red box shows the part of the chromosome for which we will get a ”region in detail” view:

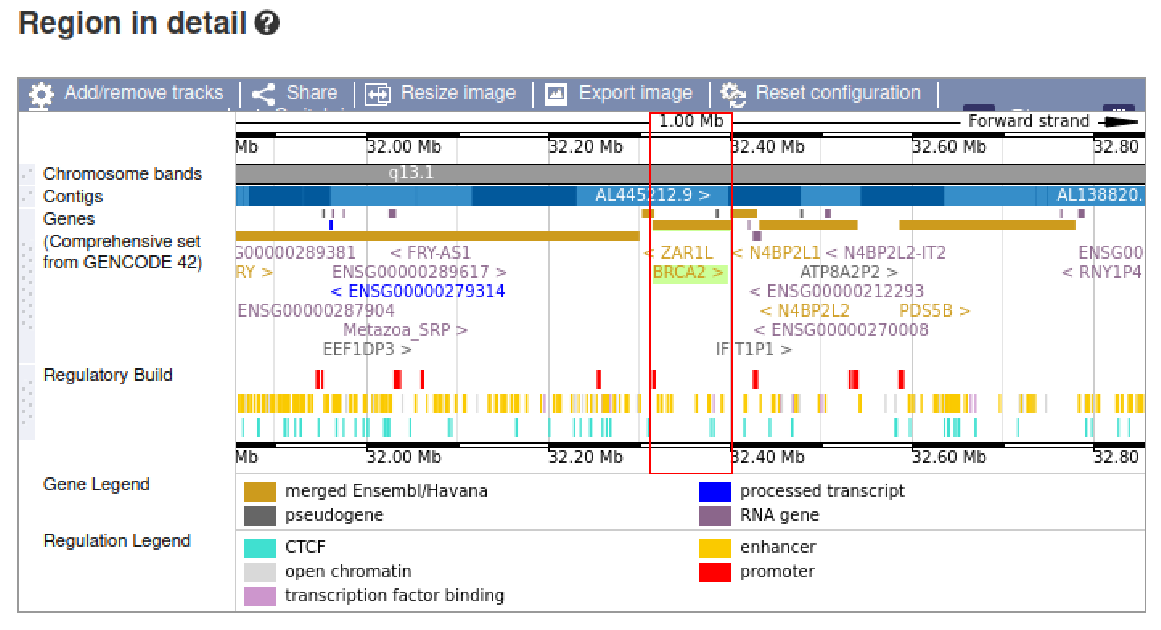


Region in detail

Promoter data in Ensemble

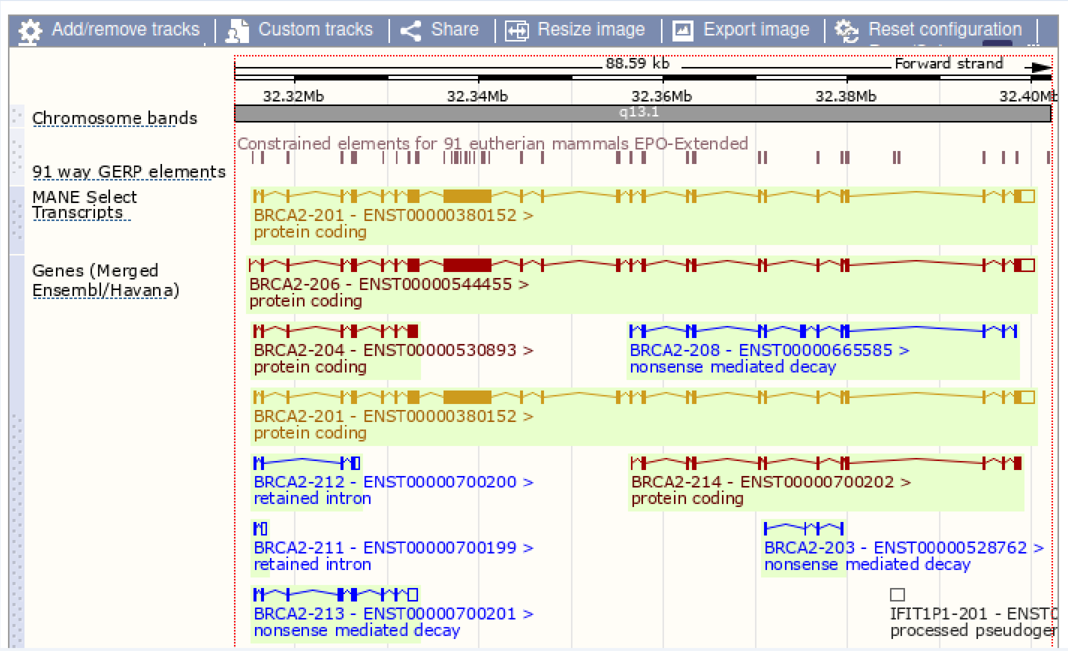
Below the Region

in detail view



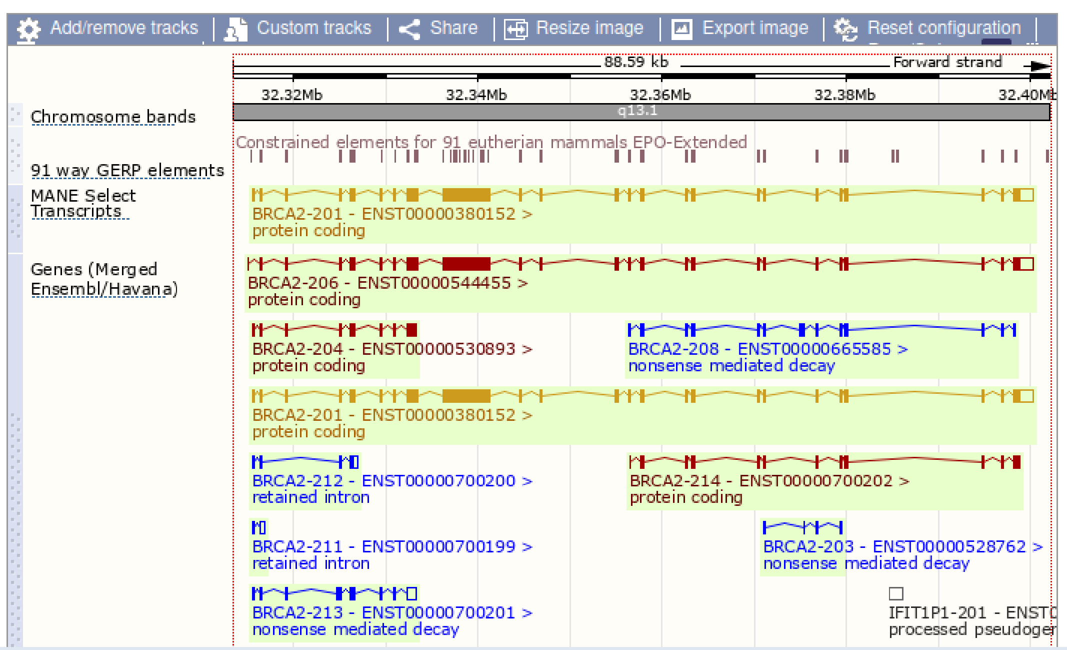
merged Ensembl/Havana	processed transcript
pseudogene	RNA gene
CTCF	enhancer
open chromatin	promoter
transcription factor binding	

we see more details of the gene:

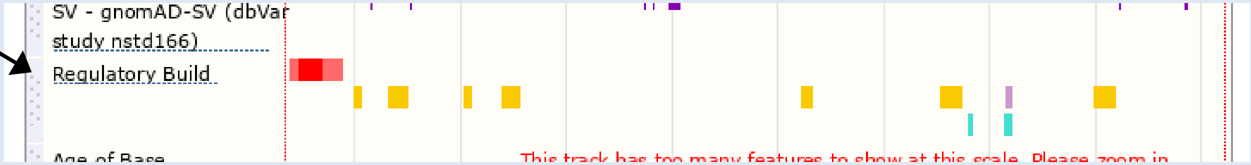


Promoter data in Ensemble

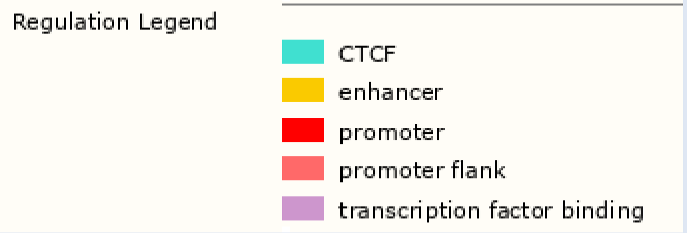
And below the gene view



In the Regulatory Build track, we see


SV - gnomAD-SV (dbVar study nstd166)
Regulatory Build
Age of Base
This track has too many features to show at this scale. Please zoom in

a promoter region and promoter flanks



Regulation Legend

🟦 CTCF

🟨 enhancer

🟥 promoter

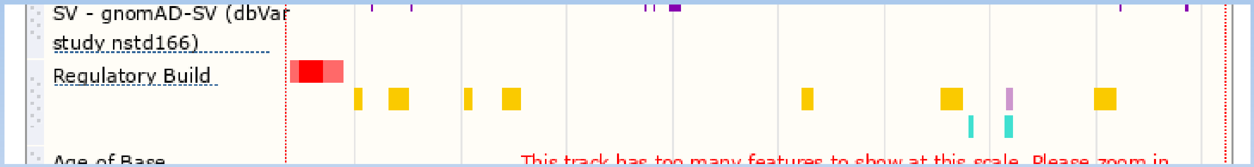
🟧 promoter flank

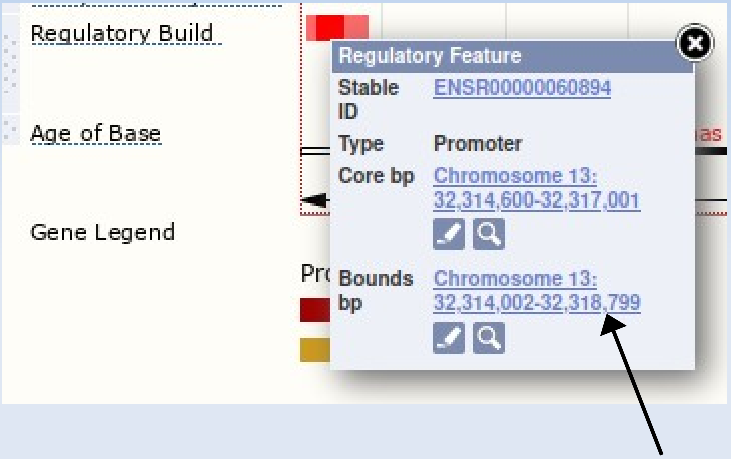
🟪 transcription factor binding

Promoter data in Ensemble

Click the promoter

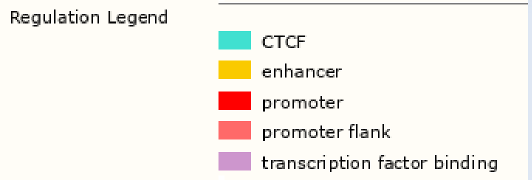
In the Regulatory Build track


SV - gnomAD-SV (dbVar
study nstd166)
Regulatory Build
Age of Base
This track has too many features to show at this scale. Please zoom in.


Regulatory Build
Age of Base
Gene Legend

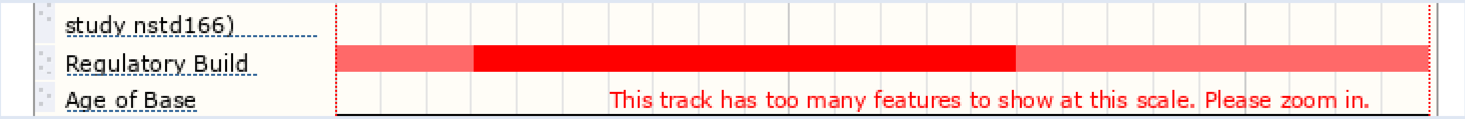
Regulatory Feature

Stable ID	ENSR00000060894
Type	Promoter
Core bp	Chromosome 13: 32,314,600-32,317,001
Bounds bp	Chromosome 13: 32,314,002-32,318,799

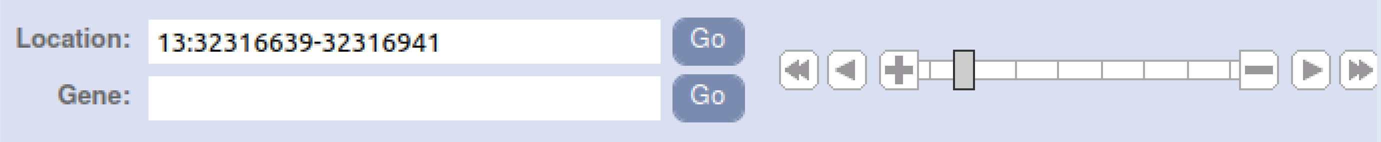

Regulation Legend

🟦	CTCF
🟨	enhancer
🟥	promoter
🟥	promoter flank
🟪	transcription factor binding

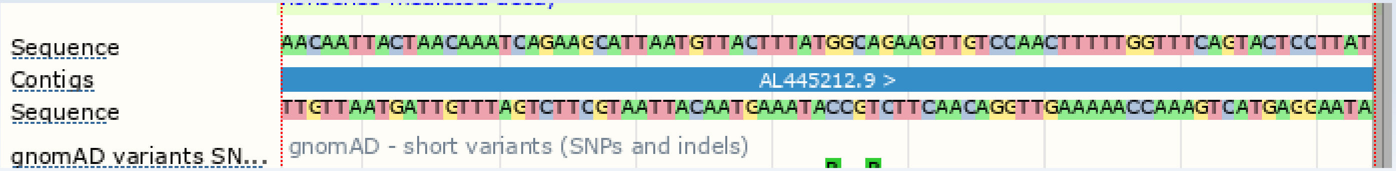
And in the pop-up window, choose the link for ”Bounds” for a closer view


study nstd166)
Regulatory Build
Age of Base
This track has too many features to show at this scale. Please zoom in.

Zoom in for a closer view


Location: 13:32316639-32316941 Go
Gene: Go

and eventually, we will be able to see nucleotides:


Sequence AACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGGCAGAAGTTGTCCAACTTTTTGGTTTCAGTACTCCTTAT
Contigs AL445212.9 >
Sequence TTGTTAATGATTGTTTAGTCTTCGTAATTACAATGAAATACCGTCTTCAACAGGTTGAAAAACCAAAGTCATGAGGAATA
gnomAD variants SN... gnomAD - short variants (SNPs and indels)