Genomic Computing

Politecnico di Milano 13-17 March, 2017

POLITECNICO DI MILANO

Dipartimento di Elettronica, Informazione e Bioingegneria

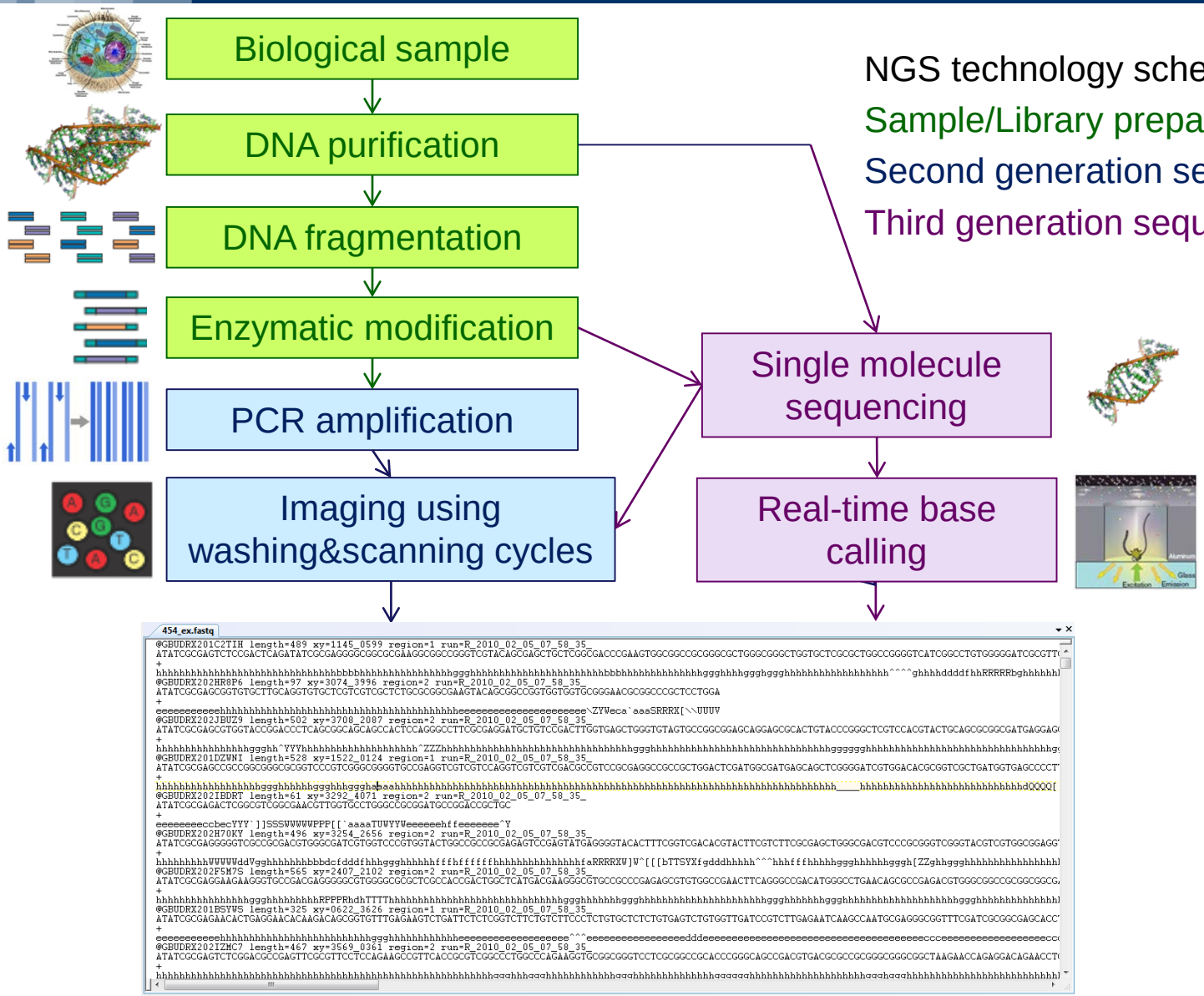
DNA viewing and qualitative analysis using Genome Browsers

Fernando Palluzzi & Marco Masseroli *masseroli@elet.polimi.it*

Motivations

- Next generation sequencing have given the possibility to sequence **entire genomes from a variety of different organisms** at low costs.
- The huge amount of sequencing data must be stored and organized to be easily used.
- Working with entire genomes is **computationally intensive**, requiring huge hardware facilities.
- Genome browsers offer the possibility to **browse entire genomes**, **viewing and downloading** genomic sequences together with all the available information contained into the main annotation databases, such as NCBI, OMIM, dbSNP, DGV, UniProt, etc.
- A good genome browser should provide also instruments for **database querying**, **sequence alignment** and other common bioinformatics tasks, together with direct links to databases and other web services.
- In this session we will see how the **UCSC Genome Browser** works.

From the biological molecule to the sequence



NGS technology schema:

Sample/Library preparation

Second generation sequencing

Third generation sequencing

Read only

Sequencing technologies and sources of error

- **Low sample quality**

  e.g. The presence of exogenous DNA molecules or other contaminants

- **Sample origin**

  e.g. Different species have diverse base composition, affecting sequencing efficiency

  Ancient DNA may come with compositional biases

- **Library type** (single-end, paired-ends, mate-pairs)
- **Coverage biases** (some regions could be over-represented)
- **Intrinsic technology errors**

  > **DNA-polymerase efficiency** decreases with sequence length

  > **Dephasing** (DNA-polymerases become asynchronous with time)

  > **Homopolymers** (e.g. AAAAAAAAAAA…)

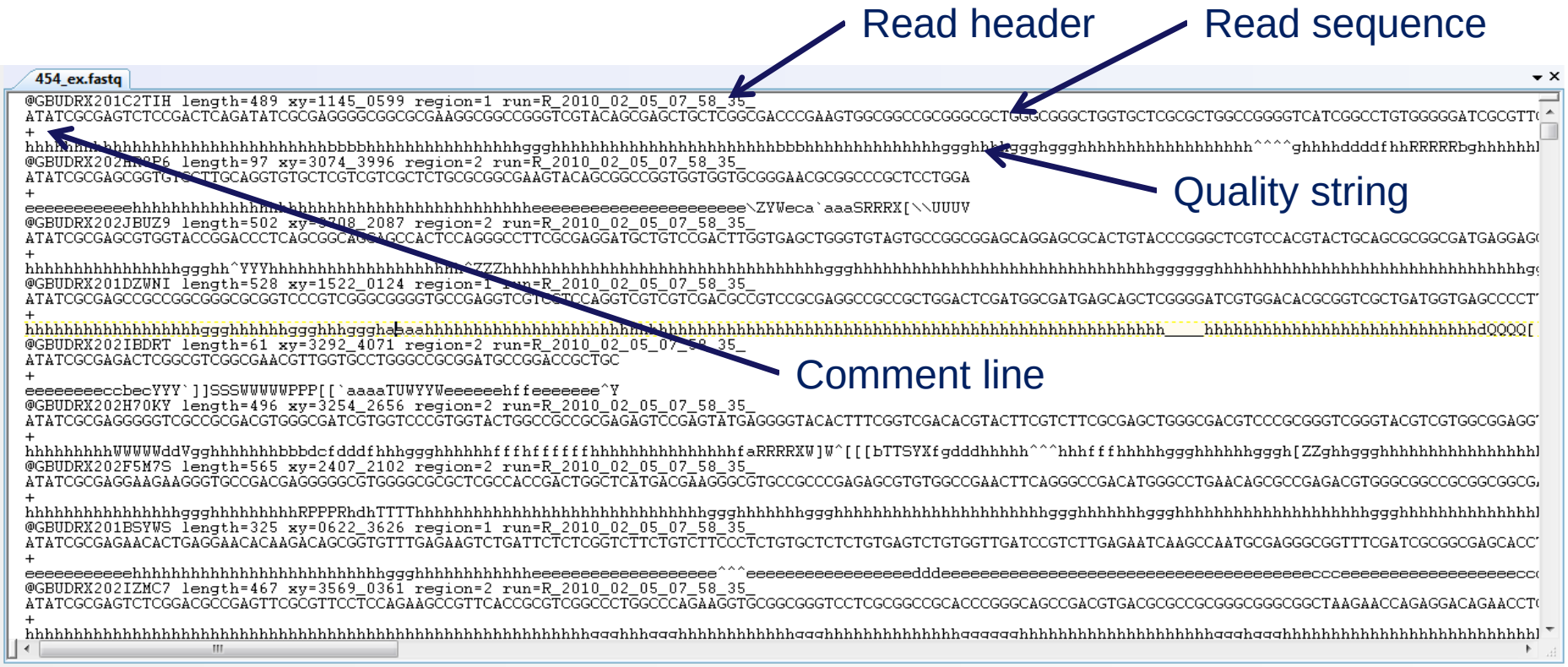
  > **Amplification (PCR) errors**

  > **Signal weakness** (if the technique does not require amplification)

The sequence file can represent the biological sequence as long as the sequencing quality is reasonably high.

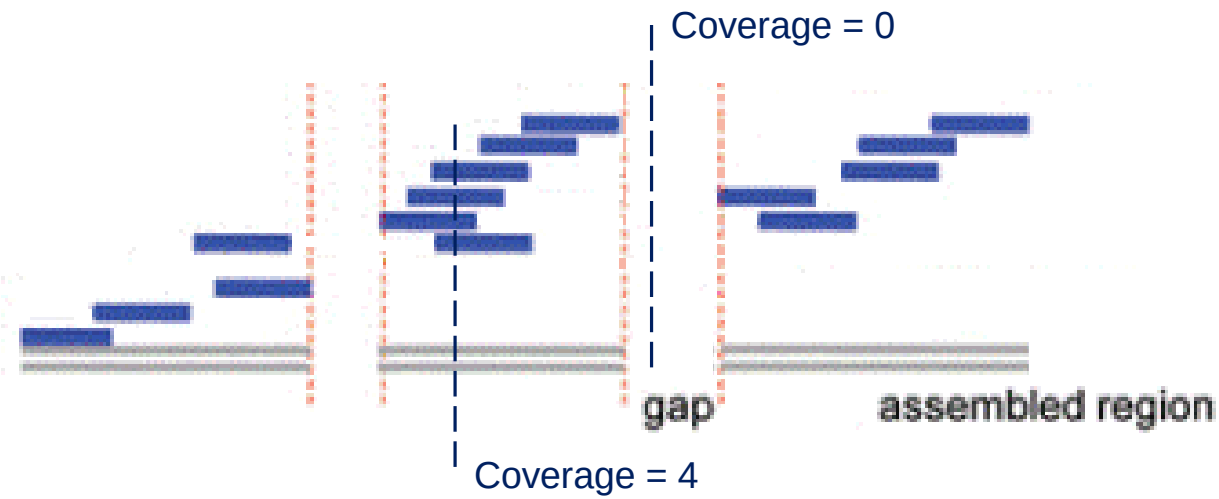
Assemblies and reference genomes

- **DNA samples are fragmented** because of a reduced DNA-polymerase efficiency with an increasing sequence length.
- Our output file (**FASTQ** format) contains our sample represented by a multiple sequence file. Each fragment in the file is called a **read**.



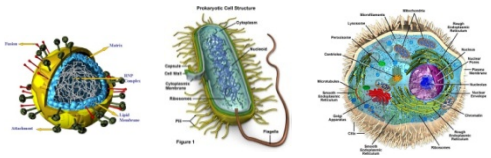
Assemblies and reference genomes

- Reads collected in FASTQ files are **overlapping** so they can be aligned.
- Genomes are obtained assembling up to millions of reads per file.
- For this reason a genome sequence can also be called **assembly**.
- The same genome may have different assemblies (e.g., hg19, GRCh38).
- A **reference genome** is an haploid assembly of fragments from different individuals, often sequenced with different technologies.



- Reference genomes are useful for **reference-based alignments** and for **identifying biologically functional genomic sequences**.

Genomes as dynamic entities



- A **genome** represents the **inheritable part of an organism variation** (i.e. its **hereditary information**)
- DNA and RNA are the biological molecules used to **store**, **encode** and **transmit** this information
- Genome information is **always changing** …

  due to either endogenous factors:

  > **Spontaneous mutation** (e.g. deamination)

  > **Duplication errors**

  > **Repair errors**

  > **Hypermutation** (a mechanism to induce immune variation)

  > **Transposable elements** (“copy&paste” or “cut&paste”)

  or induced mutations:

  > **Radiation** (e.g. UV or X-rays)

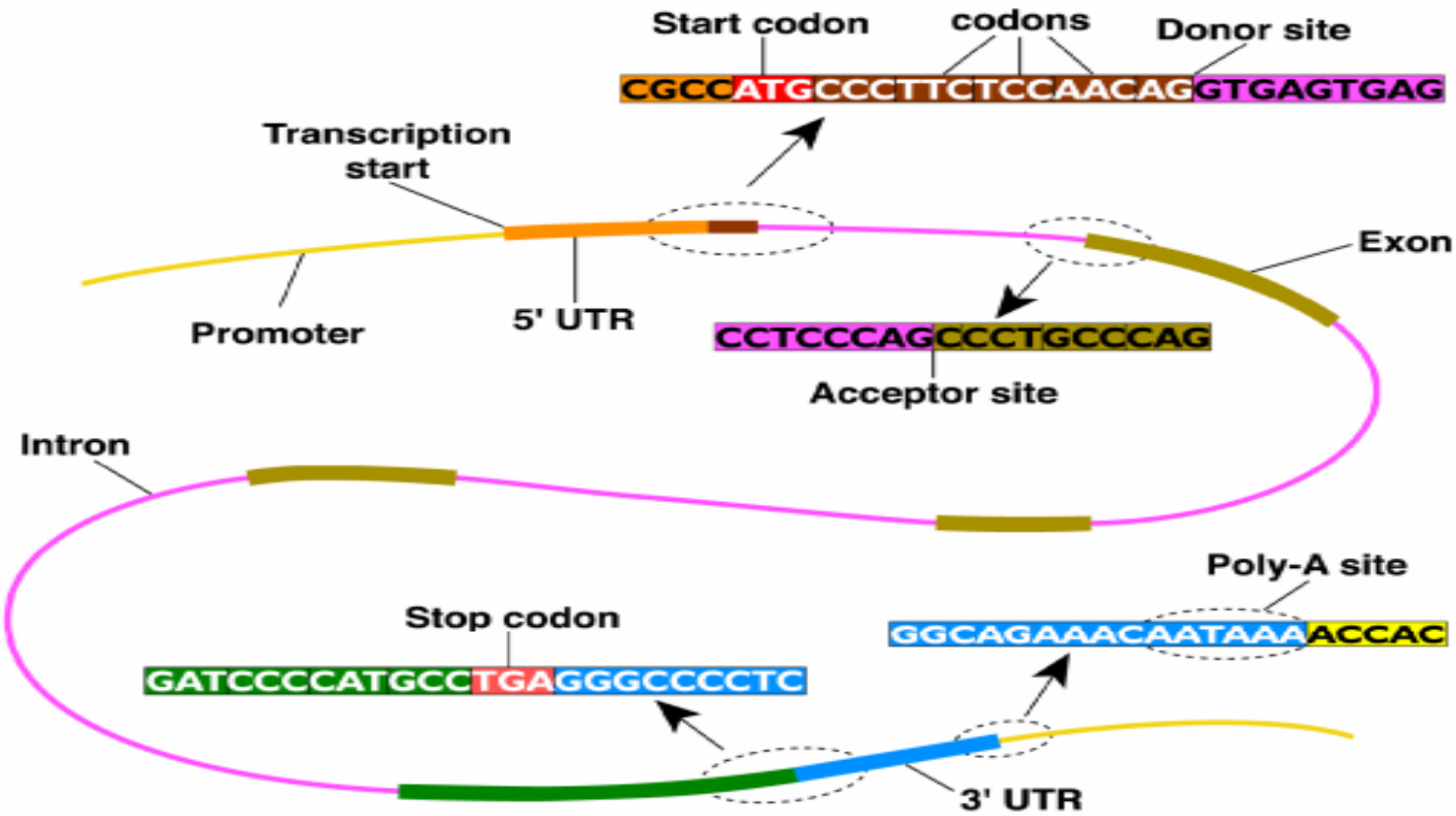
  > **Chemicals** (e.g. intercalating agents, radicals, etc.)

  > **Biological factors** (Viruses)

Genes

The entire nucleic acid sequence that is necessary for the synthesis of a functional polypeptide or RNA.

So a gene contains none, one or more **Coding DNA Sequences (CDSs)**, plus different **regulatory regions**.



Eukaryotic gene elements

PROMOTER Promoter region contains **Transcription Factor Binding Sites** (TFBSs) that may bind TFs required to initiate transcription. They can also co-locate with histone modifications such as **H3K27Ac**.

EXONs Part of the DNA sequence of a gene that is retained in the mature **mRNA**, after **splicing** (see slide 12). One gene usually have multiple exons that can be alternatively excluded, generating different **mRNA**s or **non-coding RNA**s from a single gene.

INTRONs Introns are part of the gene that are removed during splicing. Some introns can codify for proteins or can be processed after splicing, producing non-coding RNA. Introns can also be mobile genetic elements.

Eukaryotic gene elements

UTR

Eukaryotic genes have one Untranslated Terminal Region for each terminus (5’ and 3’). **5’UTR** contains regulatory regions, used to tune gene expression and the **Ribosome Binding Site** (RBS). The **3’UTR** starts just after the stop codon of the CDS. This UTR contains regulatory sequences used to regulate mature mRNA lifespan and expression: poly-A tail, protein binding sites and micro RNAs (miRNAs) binding sites.

Poly-A site

The **poly-A tail** is an Adenosine tail that is added to a mature mRNA 3’ end. This tail is involved in regulating mRNA lifespan. The poly-A site is the point in which the poly-A tail is added and it is placed at the end of 3’UTR.

TSS

Transcription Start Site. The first base transcribed into RNA. Genes may have more than one TSS.

Other regulatory elements

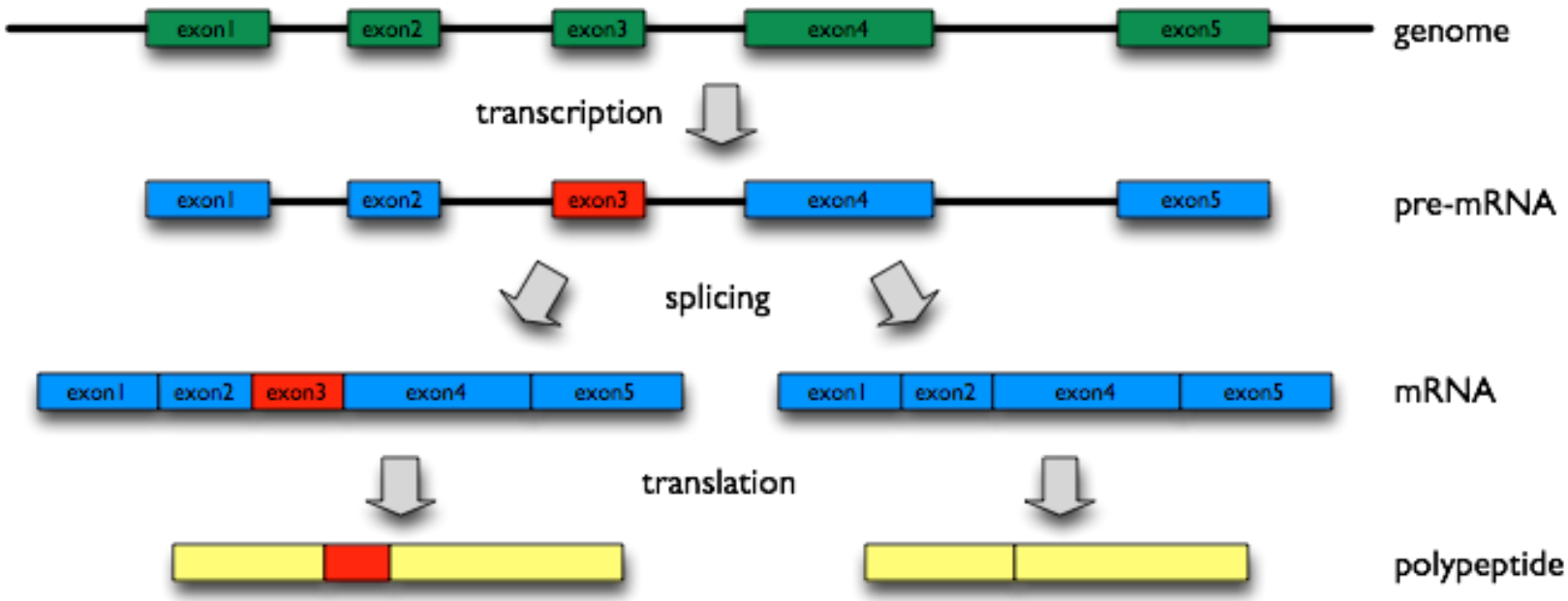
TFBS **Transcription Factor Binding Site**. TFBSs are binding regions for factors regulating gene expression (i.e. Transcription Factors). Their position is important to understand how genes are co-regulated (genes regulated by the same or interacting TFs are co-regulated)

ENHANCER Regulatory elements, bound by TFs, that enhance transcription rate of a set of genes that are implied in the same regulatory process (e.g. the production of a protein or a set of co-regulated proteins). They can also co-locate with histone modifications such as **H3K27Ac**.

CpG ISLAND Genomic regions rich in Cs and Gs, with an observed-to-expected CG content higher than 60%, on the same strand (bound by a phospho(p)-diester bound, thus the name “CpG”). In mammalian genomes they are 0,3 to 3k bp. About 70% of the **human promoters** have a high CpG content. **Methylation of GpC islands at promoters** and their flanking regions (CpG shores) is correlated with **gene repression**.

One gene, more transcripts

A gene may give rise to **more than one transcript**, and so **more than one polypeptide**. This is possible because of **splicing**. During splicing, **introns** are removed from the pre-mRNA molecule, so (usually) only **exons** are present in the mature mRNA molecule. The alternative exclusion of one or more exons, leads to the synthesis of different mRNAs and so different polypeptides.



Sense and antisense strands

- Usually we visualize single strand sequences, but biological DNA sequences have two strands: the **sense strand** (or **forward**, or **+**) and the **antisense strand** (or **reverse**, or **-**).
- In biology, a sense strand can generate a RNA version of the same sequence, which is translated into proteins.
  e.g. `5′ ATGTTCCTGTAG 3′` is a + strand if
  `5′ AUGUUCCUGUAG 3′` (the equivalent mRNA) is translated
- The antisense is the **reverse complement** of the + strand.
  e.g. `3′ TACAAGGACATC 5′` is the reverse complement of
  `5′ ATGTTCCTGTAG 3′`
- So, in a double stranded DNA (dsDNA), the + strand is equal to the RNA (with U instead of T), while the – strand is its complement.
- Conventionally, **5' is reported on the left side and 3' on the right**.
- <u>But be careful!</u> Because genes may map the reference genome in the reverse strand, you could visualize them reversed!

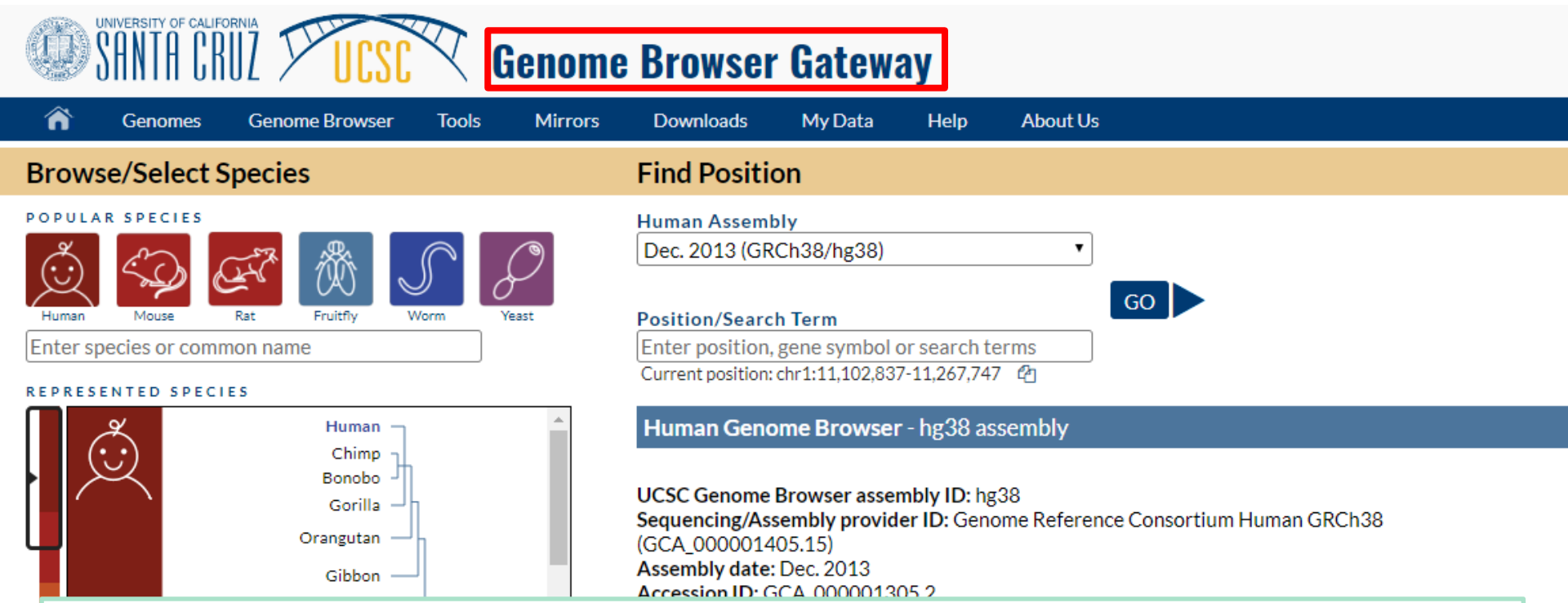
UCSC Genome Browser

The UCSC Genome Browser is an on-line browser hosted by the University of California, Santa Cruz. It offers access to different vertebrate and invertebrate genomes, providing a large database of integrated annotations, aligned with each genomic region. The browser is a graphical viewer is an OPEN-SOURCE, web-based tool suite built on top a MySQL database.

The interface allow the user to explore the database using different tools, like the Genome Browser, the Table Browser, BLAT, Gene Sorter, etc.

The Genome Browser offers also hyperlinks to other databases, like NCBI, UniProt or OMIM, and other web services, like Galaxy.

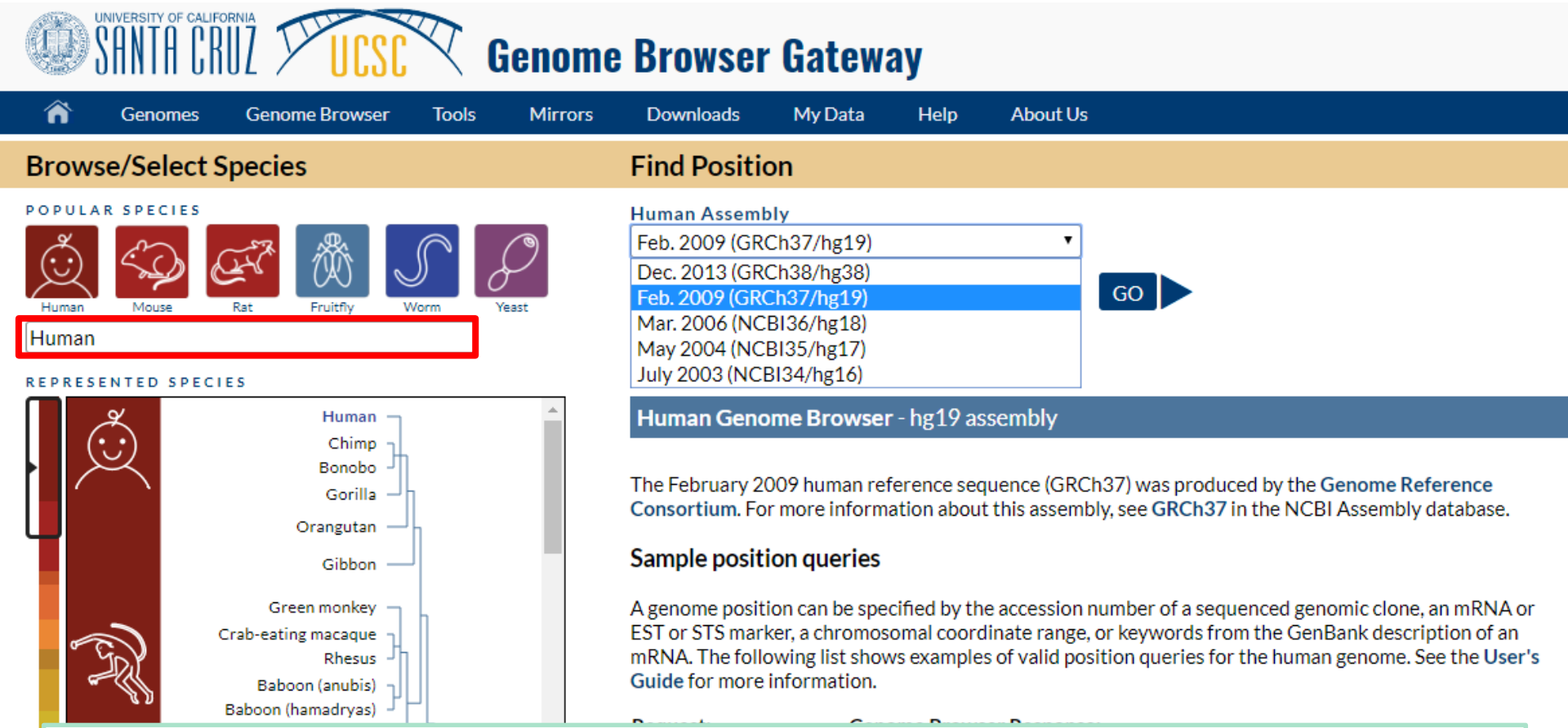
UCSC Genome Browser Gateway



The Gateway is the access point to the Genome Browser. The required fields are **species or common name** (of the species whose genome we want to browse), **assembly** (the assembly of the genome of interest) and **position** (the genomic region we want to visualize), **gene symbol** or **search term** (a gene name or description of the gene, transcript or protein we want to search for).

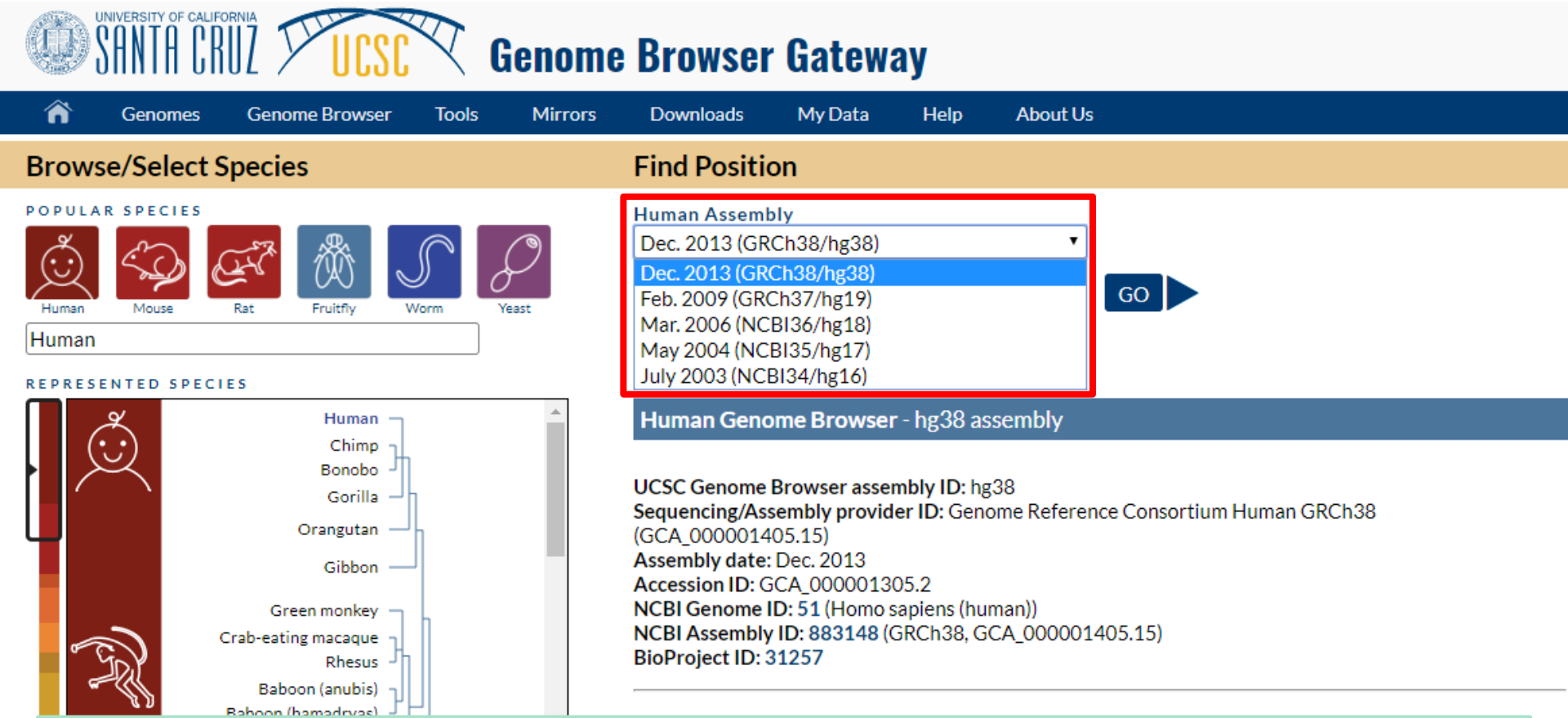
Position/search term are exclusive search parameters, while the others are compulsory settings.

Example: Human genomes …



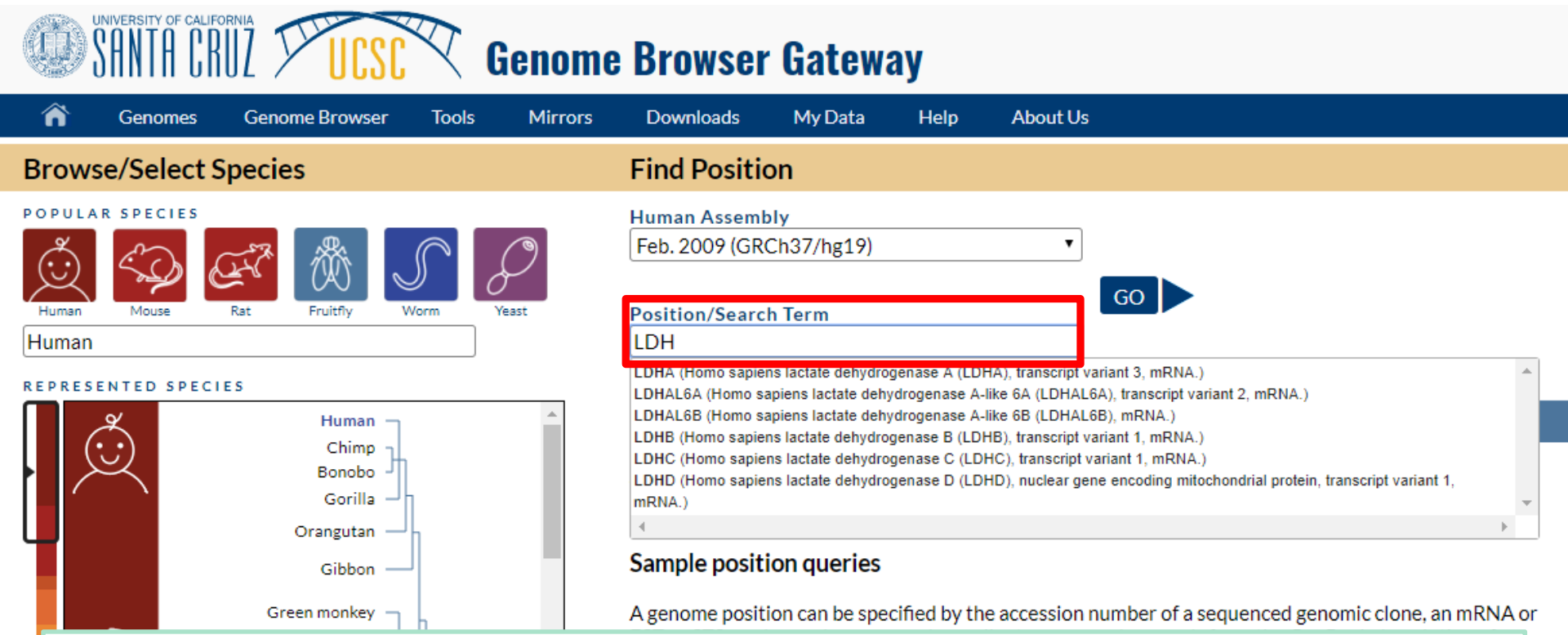
Select Human genome.

Example: the most recent assemblies of the human genome …



Select Human genome >>> Assembly GRCh38/hg38.

We can either search by a specific term…



Let's search for "LDHB" gene.

The search bar uses the autocomplete option, but it is not showing all the possible options.

So type "LDHB" and then press ENTER.

‘LDHB’ Search results

Known Genes

```
LDHB (uc001rfe.4) at chr12:21635342-21657842 - Homo sapiens lactate dehydrogenase B (LDHB), transcript variant 1, mRNA. (from RefSeq NM_002300)
LDHB (uc058lvt.1) at chr12:21654657-21757831 - lactate dehydrogenase B (from HGNC LDHB)
LDHB (uc058lvs.1) at chr12:21654355-21657971 - lactate dehydrogenase B (from HGNC LDHB)
LDHB (uc058lvr.1) at chr12:21644050-21757857 - Belongs to the LDH/MDH superfamily. (from UniProt F5H793)
LDHB (uc058lvq.1) at chr12:21643946-21657839 - Belongs to the LDH/MDH superfamily. (from UniProt C9J7H8)
LDHB (uc058lvp.1) at chr12:21643545-21654570 - lactate dehydrogenase B (from HGNC LDHB)
LDHB (uc058lvo.1) at chr12:21638368-21657362 - (S)-lactate + NAD(+) = pyruvate + NADH. (from UniProt A8MW50)
LDHB (uc058lvn.1) at chr12:21635342-21638538 - lactate dehydrogenase B (from HGNC LDHB)
LDHB (uc058lvm.1) at chr12:21635342-21637374 - lactate dehydrogenase B (from HGNC LDHB)
LDHB (uc001rfd.4) at chr12:21635342-21657794 - Homo sapiens lactate dehydrogenase B (LDHB), transcript variant
LDHA (uc021qep.2) at chr11:18394562-18408218 - Homo sapiens lactate dehydrogenase A (LDHA), transcript variant
LDHA (uc010rdd.2) at chr11:18396266-18408218 - Homo sapiens lactate dehydrogenase A (LDHA), transcript variant
LDHA (uc010rdc.2) at chr11:18394562-18407667 - Homo sapiens lactate dehydrogenase A (LDHA), transcript variant
LDHA (uc001mol.4) at chr11:18394586-18408218 - Homo sapiens lactate dehydrogenase A (LDHA), transcript variant
PEX5 (uc009zfu.3) at chr12:7189163-7218571 - Homo sapiens peroxisomal biogenesis factor 5 (PEX5), transcript va
PEX5 (uc010sgc.3) at chr12:7189687-7210431 - Homo sapiens peroxisomal biogenesis factor 5 (PEX5), transcript va
PEX5 (uc001qsw.4) at chr12:7189715-7210262 - Homo sapiens peroxisomal biogenesis factor 5 (PEX5), transcript va
PEX5 (uc001qsv.4) at chr12:7190361-7211480 - Homo sapiens peroxisomal biogenesis factor 5 (PEX5), transcript va
PEX5 (uc001qsu.4) at chr12:7189584-7211483 - Homo sapiens peroxisomal biogenesis factor 5 (PEX5), transcript va
LRRK2 (uc001rmg.5) at chr12:40225074-40369281 - Homo sapiens leucine-rich repeat kinase 2 (LRRK2), mRNA. (from RefSeq NM_198578)
TERF1 (uc003xzd.3) at chr8:73008864-73046471 - Homo sapiens telomeric repeat binding factor 1 (TERF1), transcript variant 1, mRNA. (from RefSeq NM_017489)
TERF1 (uc003xze.3) at chr8:73008864-73048122 - Homo sapiens telomeric repeat binding factor 1 (TERF1), transcript variant 2, mRNA. (from RefSeq NM_003218)
```

Known Genes reports the list of genes as annotated in the Genome Browser, with their UCSC ID in round brackets.

NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*)

```
NM_001174097.2 at chr12:21635342-21657794
NM_001315537.1 at chr12:21635342-21657855
NM_002300.7 at chr12:21635342-21657855
```

RefSeq Genes reports the list of **reference sequences** as reported in the National Center for Biotechnology Information (**NCBI**). Many sequence databases contain **multiple sequencing versions** for the same biological sequence. For example, this may be the consequence of improvements in sequencing conditions, that allow to create a more reliable sequence version. Many databases contain each of these versions, generating a high redundancy. Instead, the RefSeq repository contains **non-redundant** (only one sequencing version per each gene, transcript or protein, the most accurate one) and **well-annotated** (i.e. manually checked and annotated) sequences.

‘LDHB’ Search results

*The **Encyclopedia of DNA Elements** (**ENCODE**) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.*

*The aim of **GENCODE** as a sub-project of the ENCODE scale-up project is to annotate all evidence-based gene features in the entire human genome at a high accuracy.*

Links:

ENCODE@NHGRI	http://www.genome.gov/10005107
ENCODE@USCS	http://genome.ucsc.edu/ENCODE/
GENCODE	http://www.gencodegenes.org/

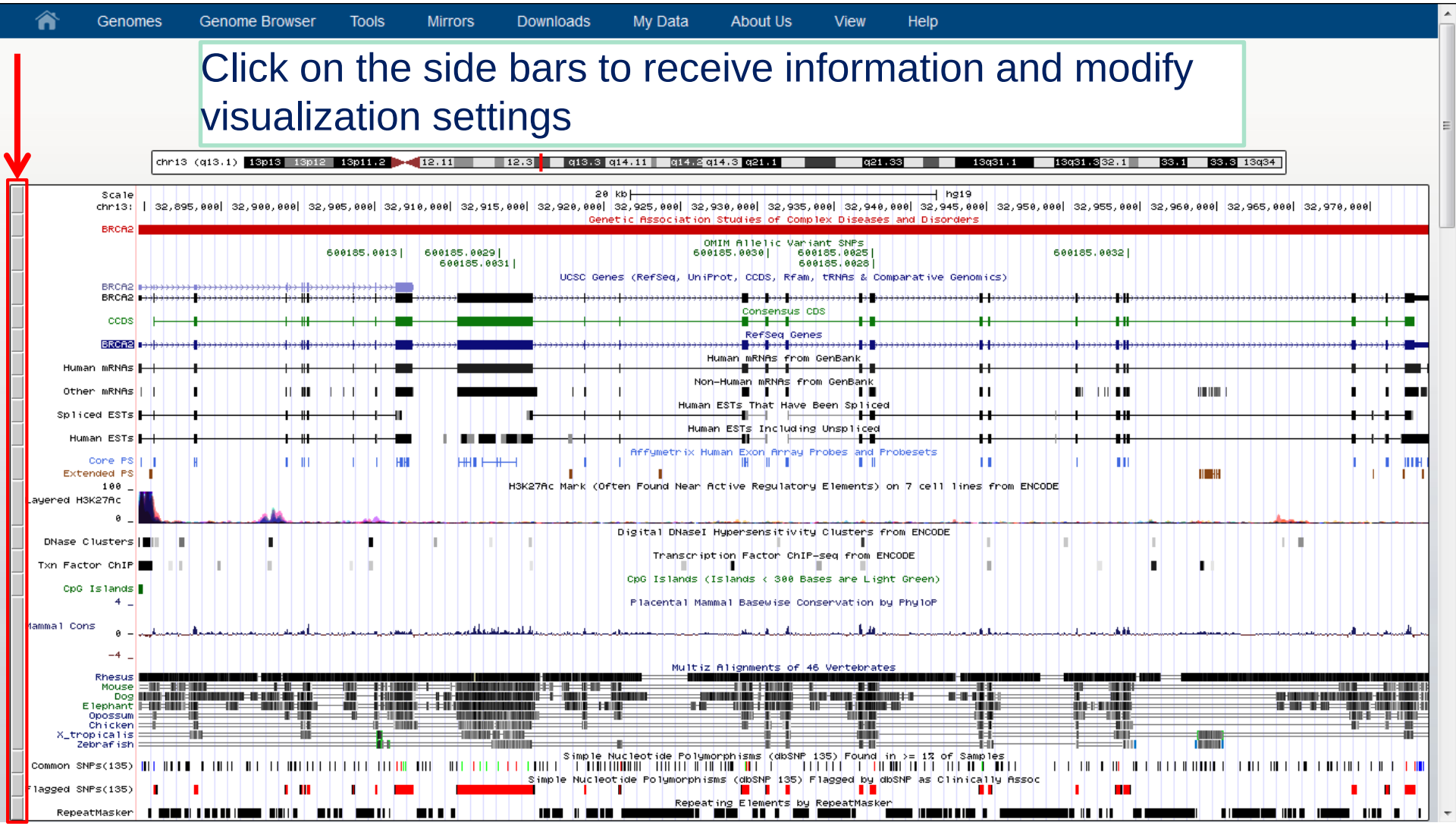
LDHB at chr12:21635353-21657851 - (NM_001316338) L-lactate dehydrogenase B chain isoform LDHBx
LDHB at chr12:21635353-21657851 - (NM_174100) L-lactate dehydrogenase B chain isoform LDHB
LDHB at chr12:21635342-21657363 - (NM_001320089) L-lactate dehydrogenase B chain isoform LDHB
LDHB at chr12:21635342-21657853 - (NM_001320086) L-lactate dehydrogenase B chain isoform LDHBx
LDHB at chr12:21635342-21657853 - (NM_001195809) L-lactate dehydrogenase B chain isoform LDHB

Basic Gene Annotation Set from GENCODE Version 28 (Ensembl 92)

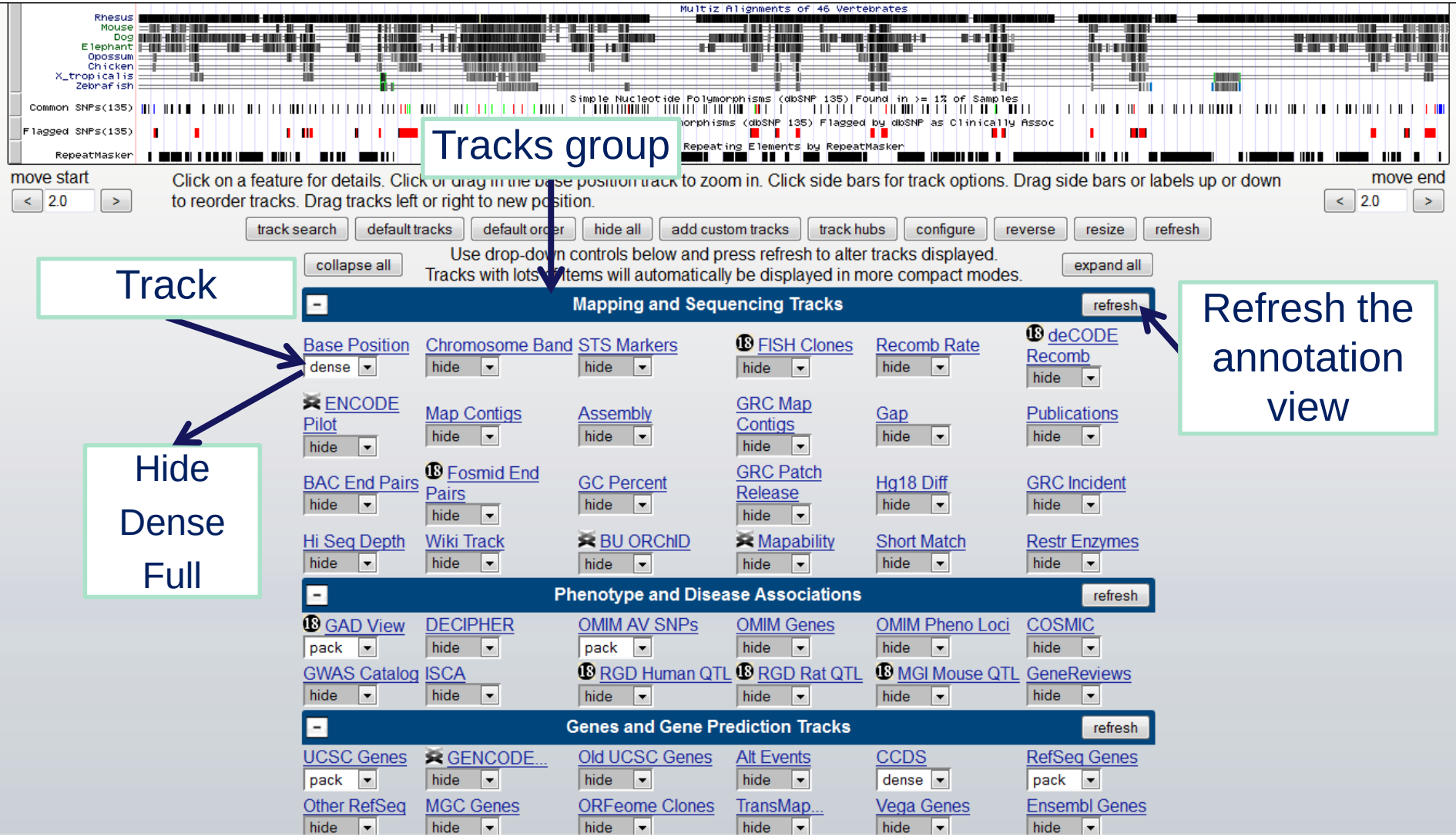
LDHB at chr12:21635342-21657794
LDHB at chr12:21635342-21657842

Annotation tracks, a general view

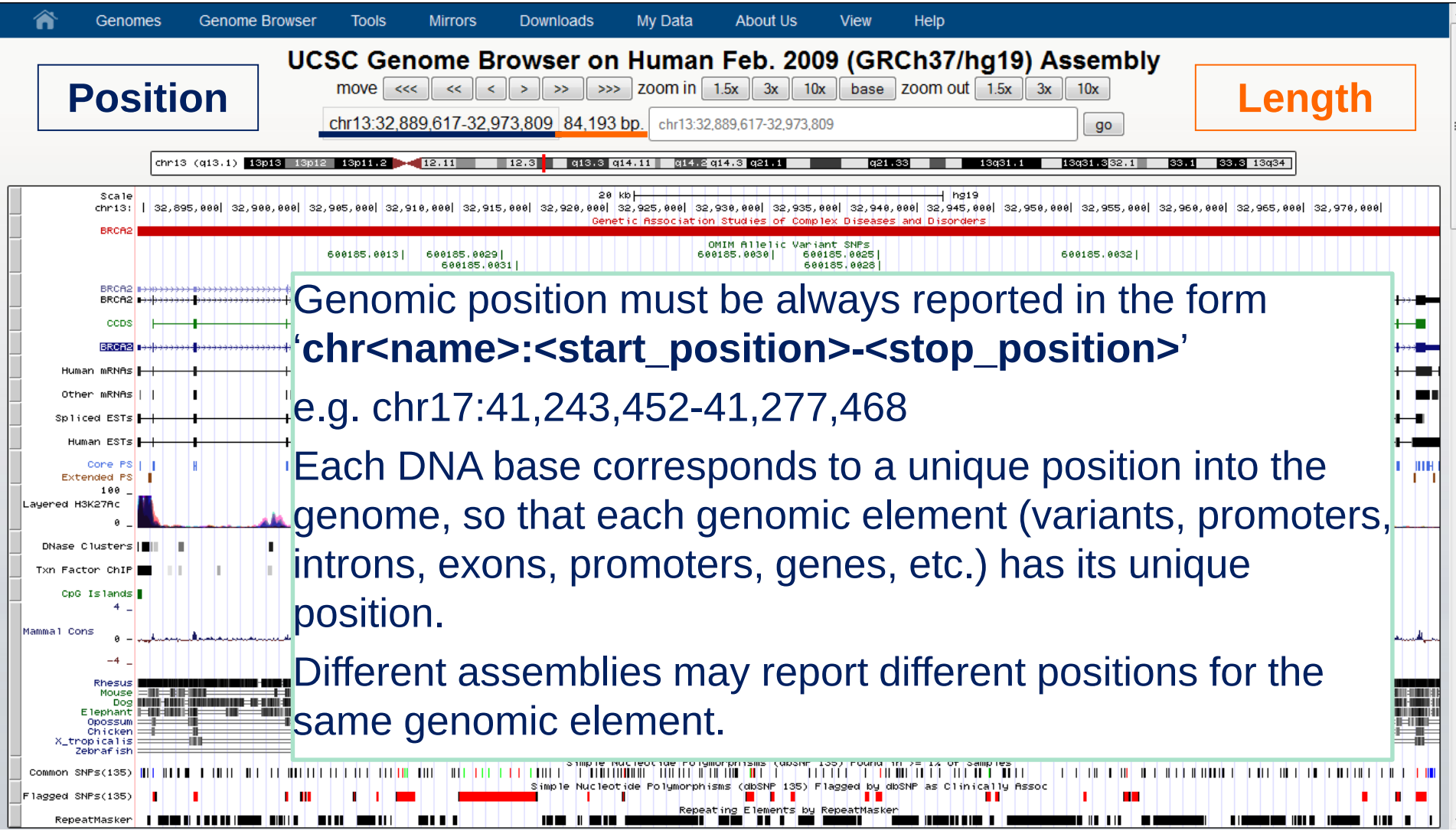
Click on the side bars to receive information and modify visualization settings



Customizing annotation tracks



Annotation tracks



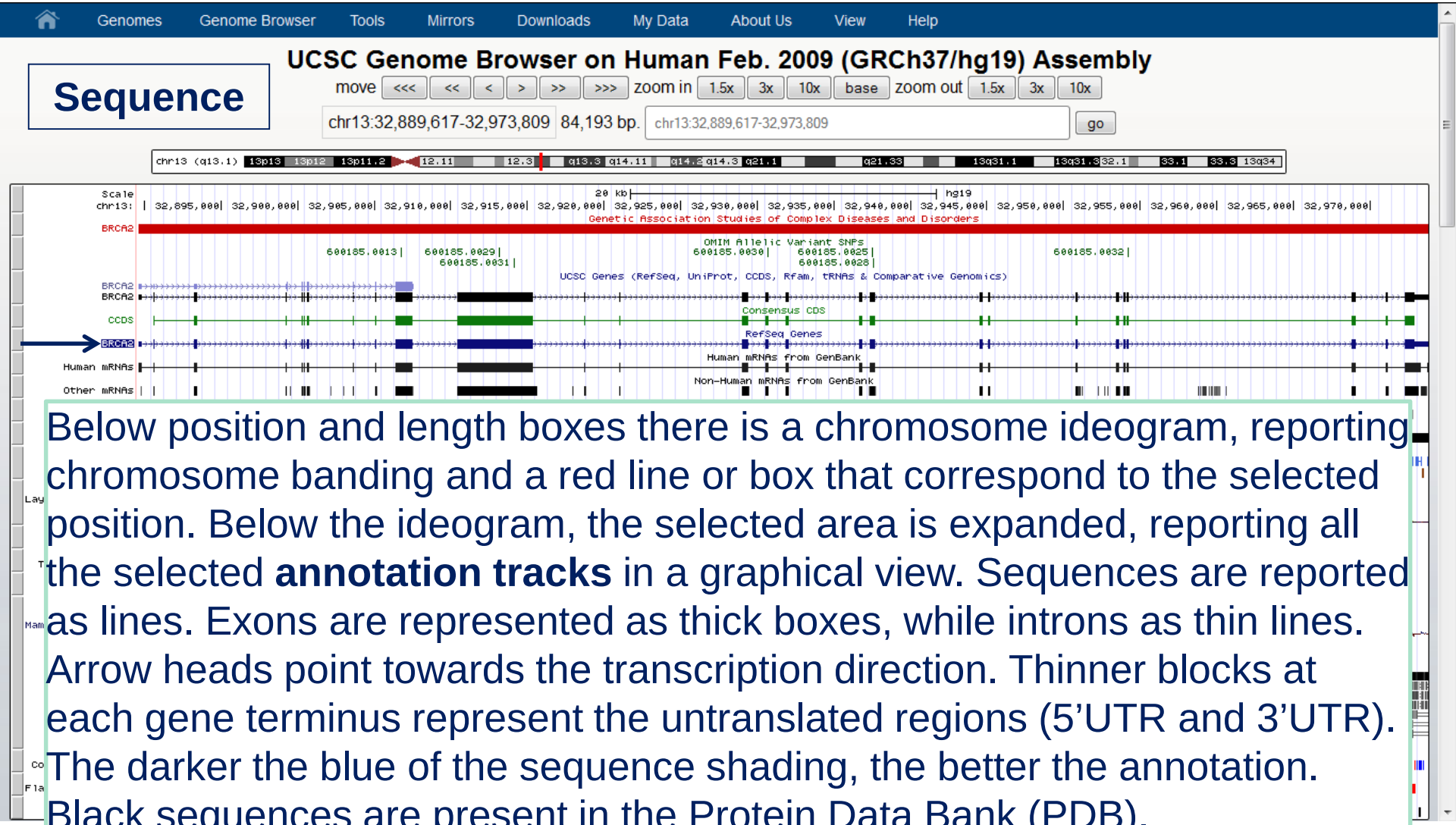
Genomic position must be always reported in the form ‘**chr<name>:<start_position>-<stop_position>**’

e.g. chr17:41,243,452-41,277,468

Each DNA base corresponds to a unique position into the genome, so that each genomic element (variants, promoters, introns, exons, promoters, genes, etc.) has its unique position.

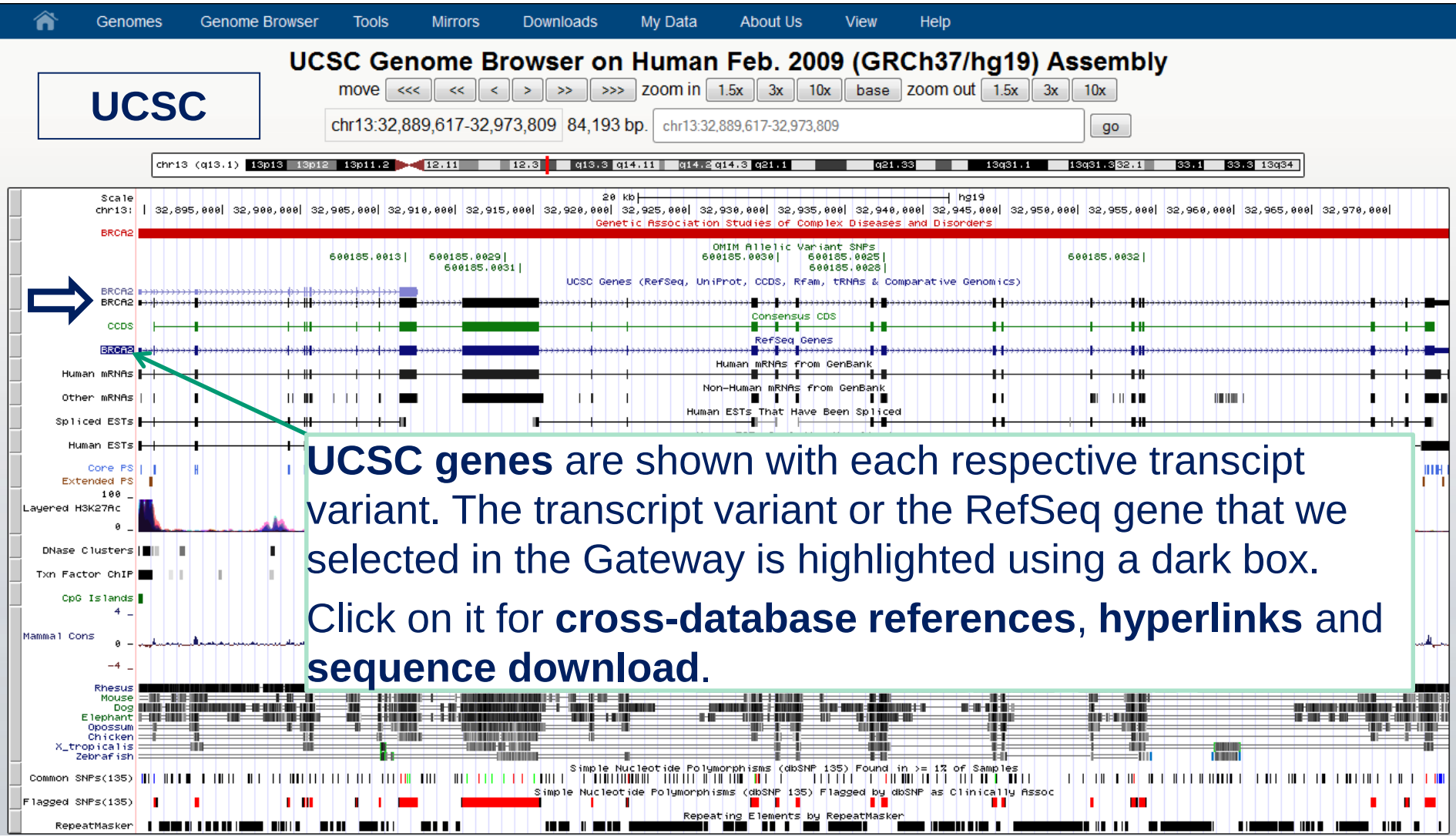
Different assemblies may report different positions for the same genomic element.

Annotation tracks



Below position and length boxes there is a chromosome ideogram, reporting chromosome banding and a red line or box that correspond to the selected position. Below the ideogram, the selected area is expanded, reporting all the selected **annotation tracks** in a graphical view. Sequences are reported as lines. Exons are represented as thick boxes, while introns as thin lines. Arrow heads point towards the transcription direction. Thinner blocks at each gene terminus represent the untranslated regions (5'UTR and 3'UTR). The darker the blue of the sequence shading, the better the annotation. Black sequences are present in the Protein Data Bank (PDB).

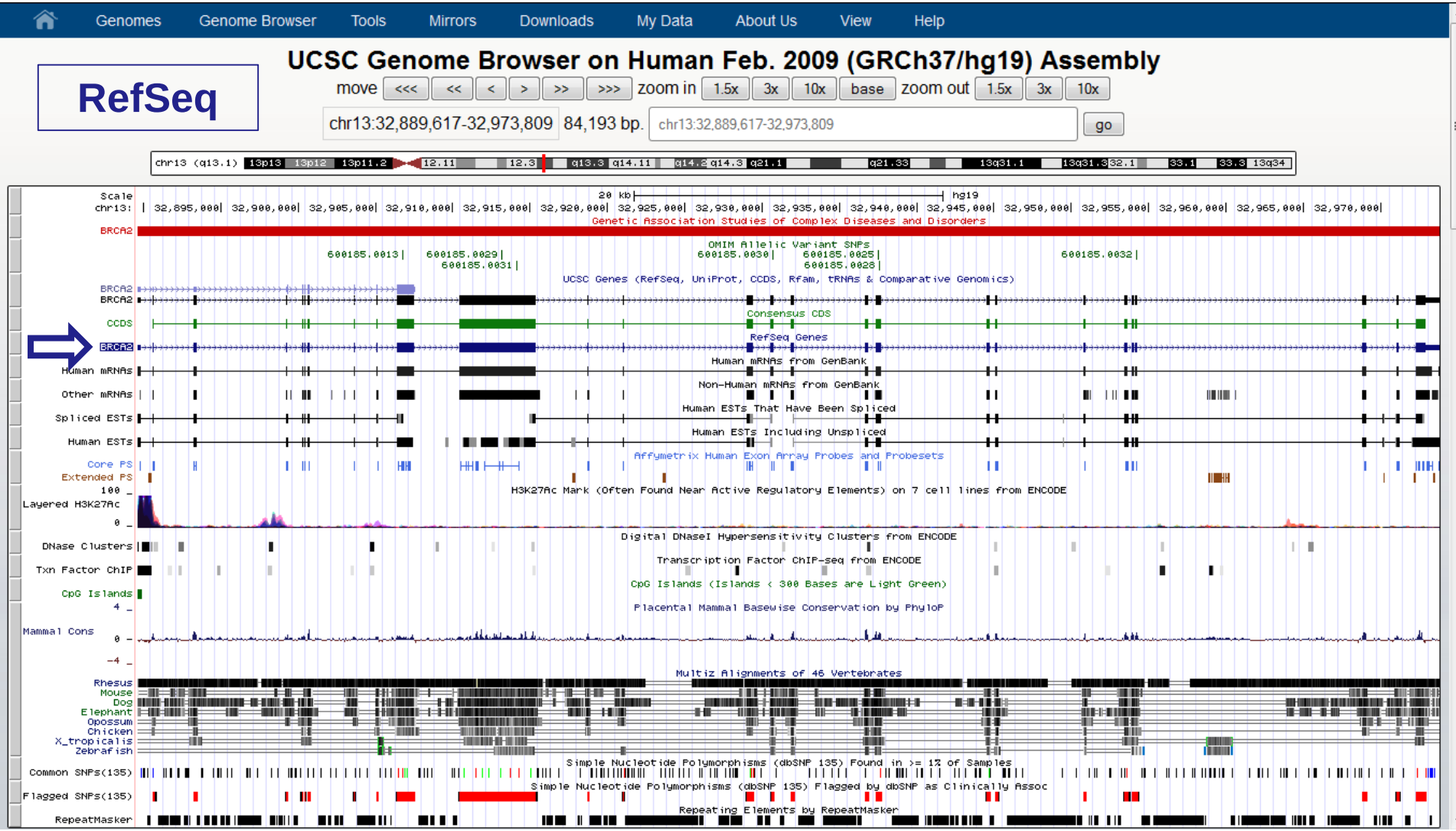
Annotation tracks



UCSC genes are shown with each respective transcipt variant. The transcript variant or the RefSeq gene that we selected in the Gateway is highlighted using a dark box.

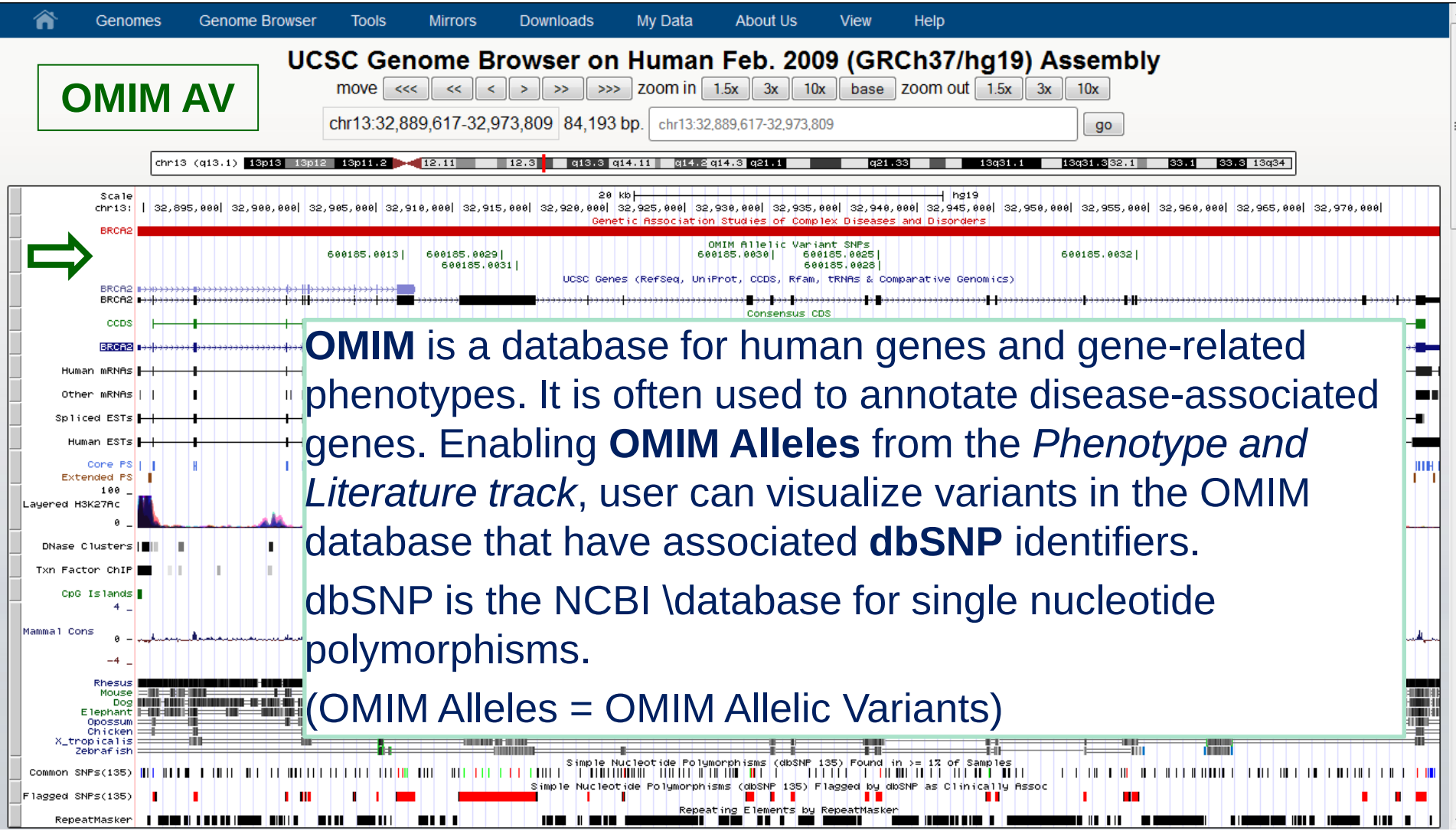
Click on it for **cross-database references**, **hyperlinks** and **sequence download**.

Annotation tracks



Annotation tracks

OMIM AV



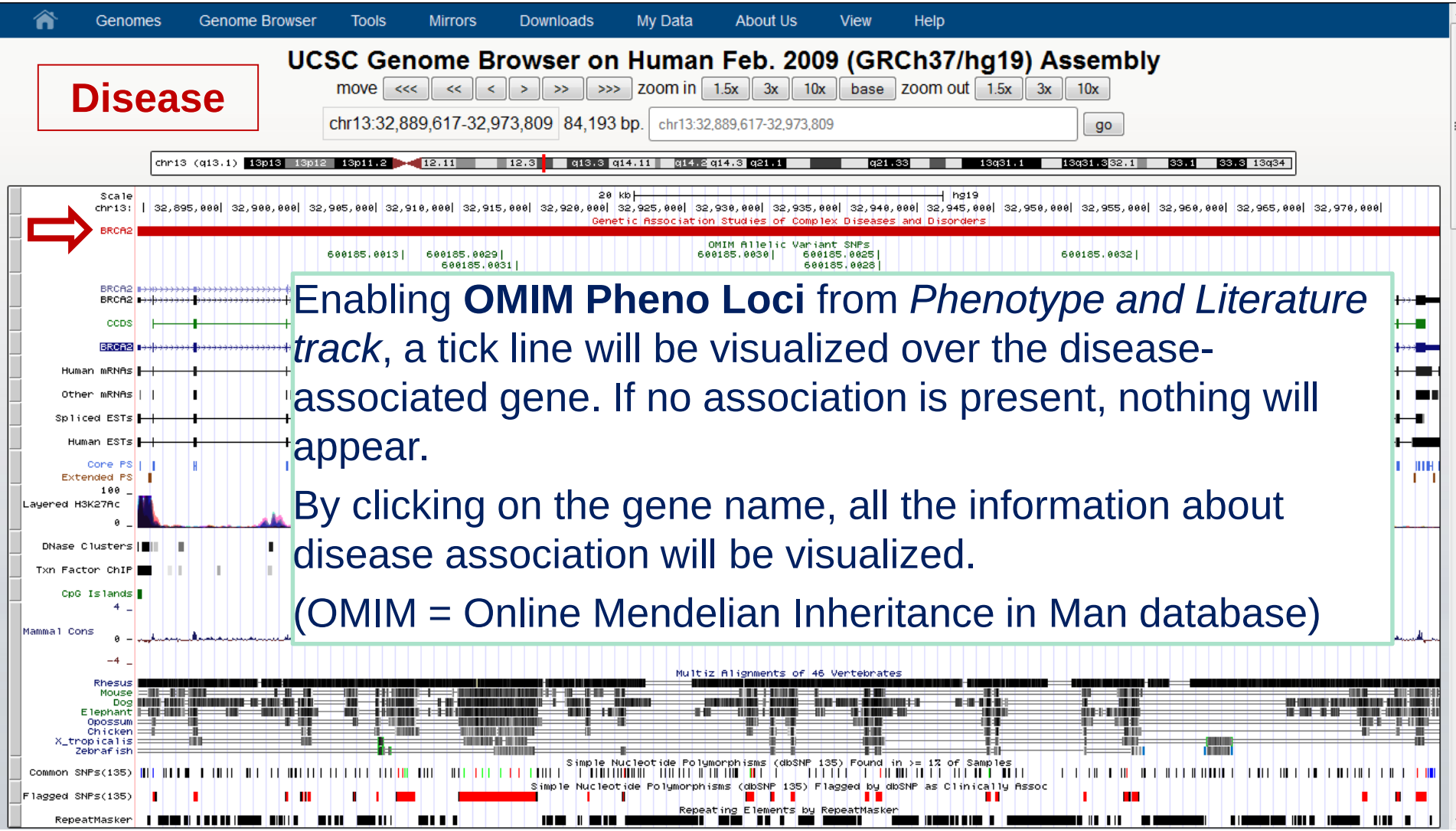
OMIM is a database for human genes and gene-related phenotypes. It is often used to annotate disease-associated genes. Enabling **OMIM Alleles** from the *Phenotype and Literature track*, user can visualize variants in the OMIM database that have associated **dbSNP** identifiers.

dbSNP is the NCBI \database for single nucleotide polymorphisms.

(OMIM Alleles = OMIM Allelic Variants)

Annotation tracks

Disease

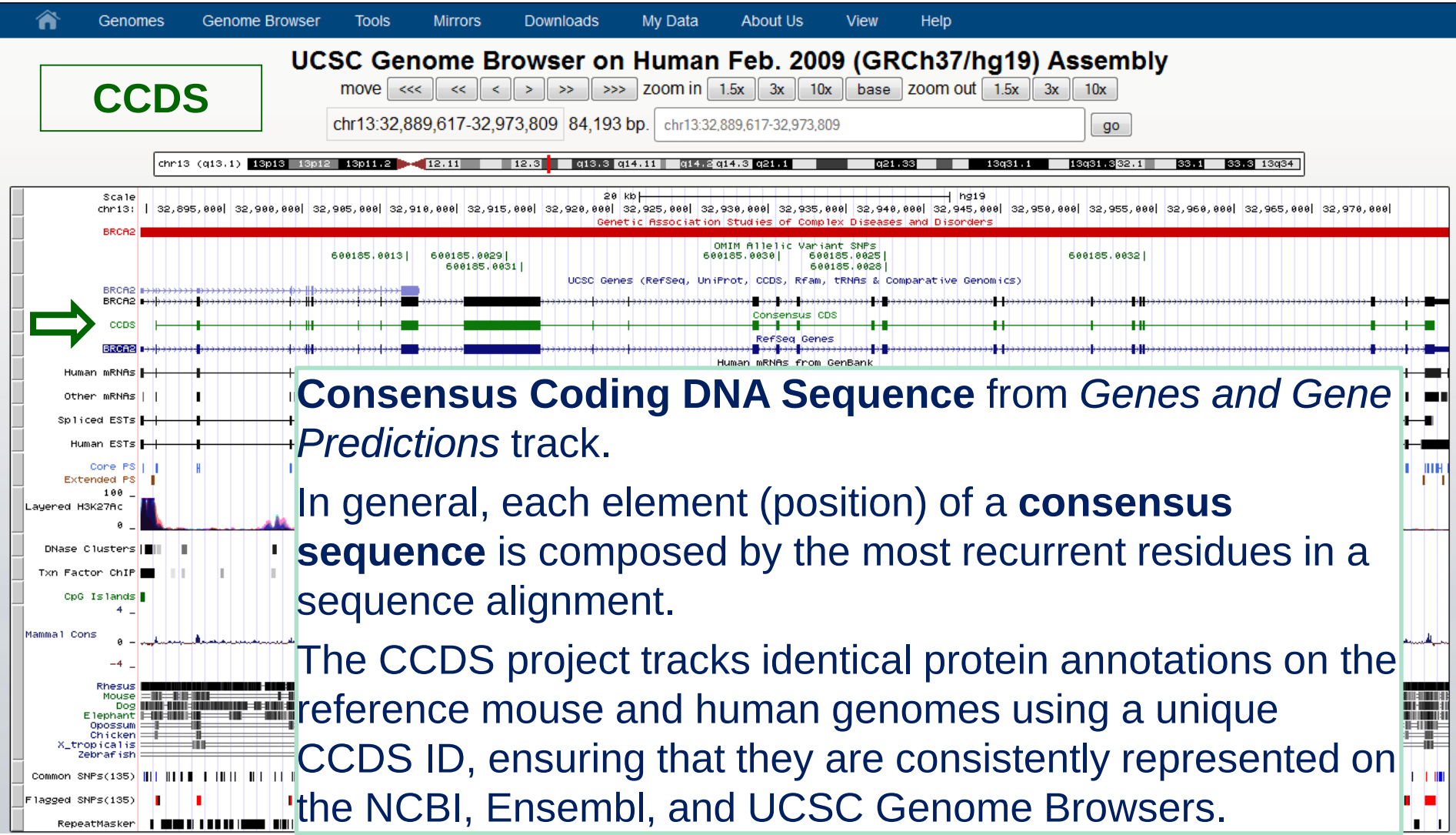


Enabling **OMIM Pheno Loci** from *Phenotype and Literature track*, a tick line will be visualized over the disease-associated gene. If no association is present, nothing will appear.

By clicking on the gene name, all the information about disease association will be visualized.

(OMIM = Online Mendelian Inheritance in Man database)

Annotation tracks



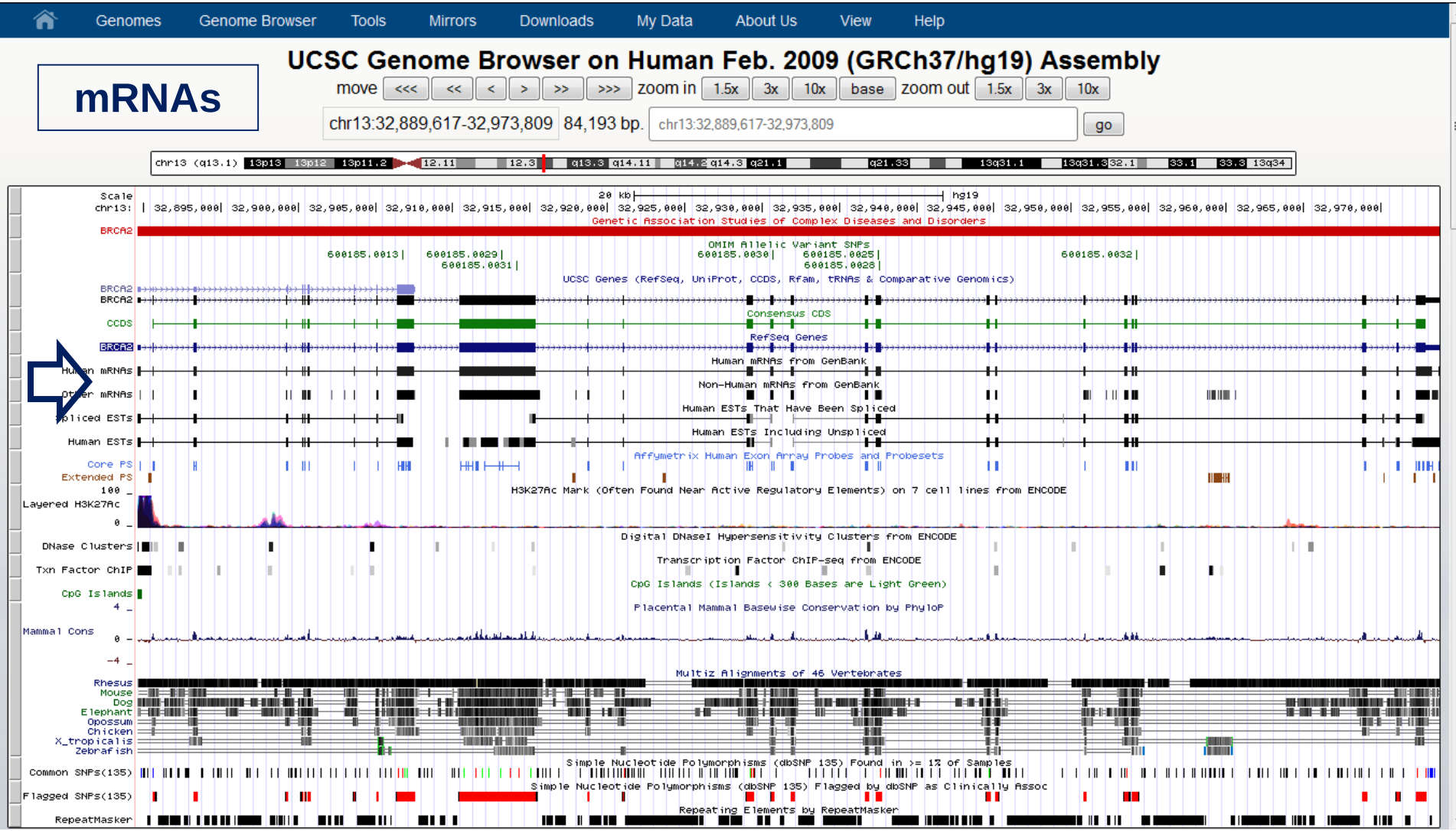
CCDS

Consensus Coding DNA Sequence from *Genes and Gene Predictions* track.

In general, each element (position) of a **consensus sequence** is composed by the most recurrent residues in a sequence alignment.

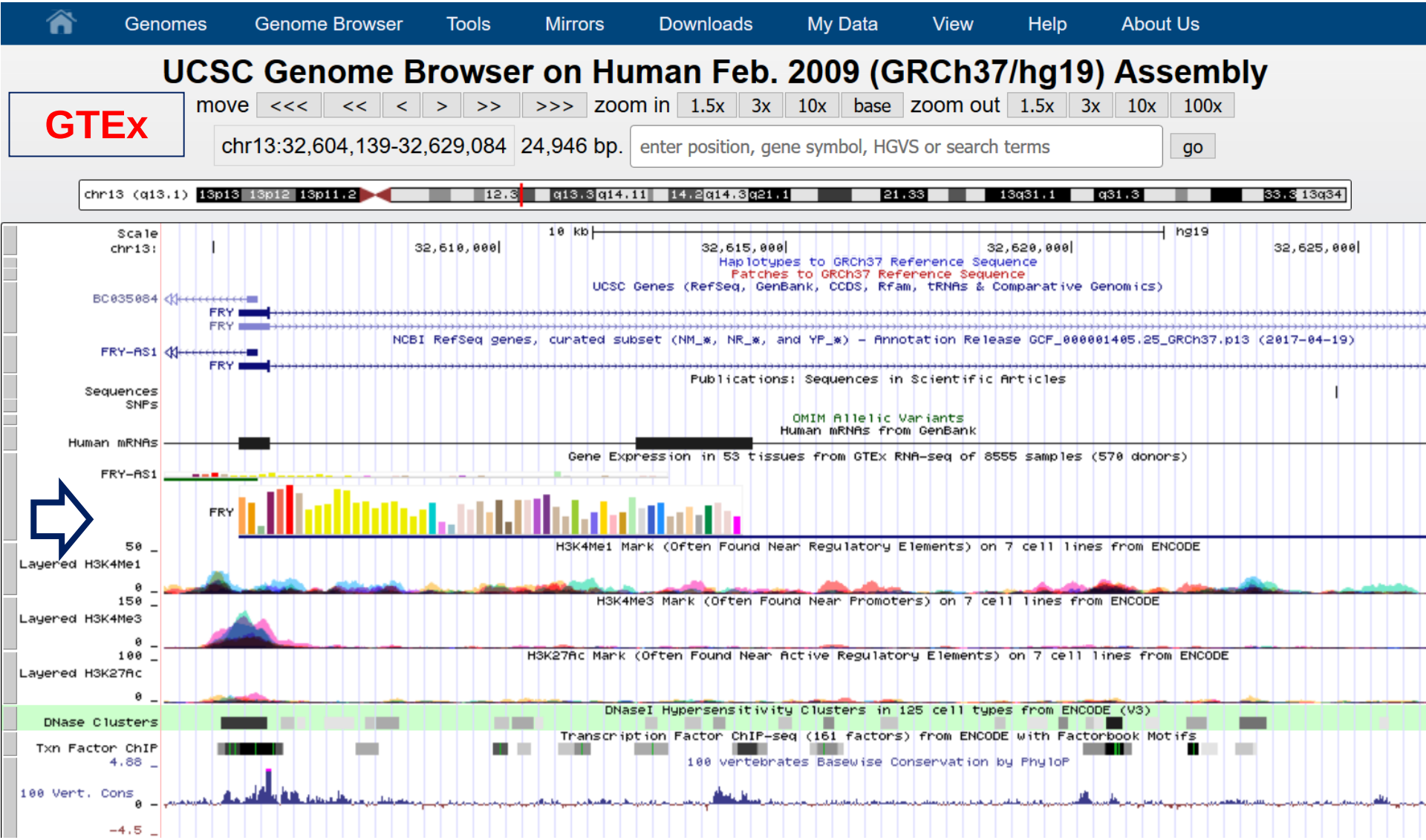
The CCDS project tracks identical protein annotations on the reference mouse and human genomes using a unique CCDS ID, ensuring that they are consistently represented on the NCBI, Ensembl, and UCSC Genome Browsers.

Annotation tracks

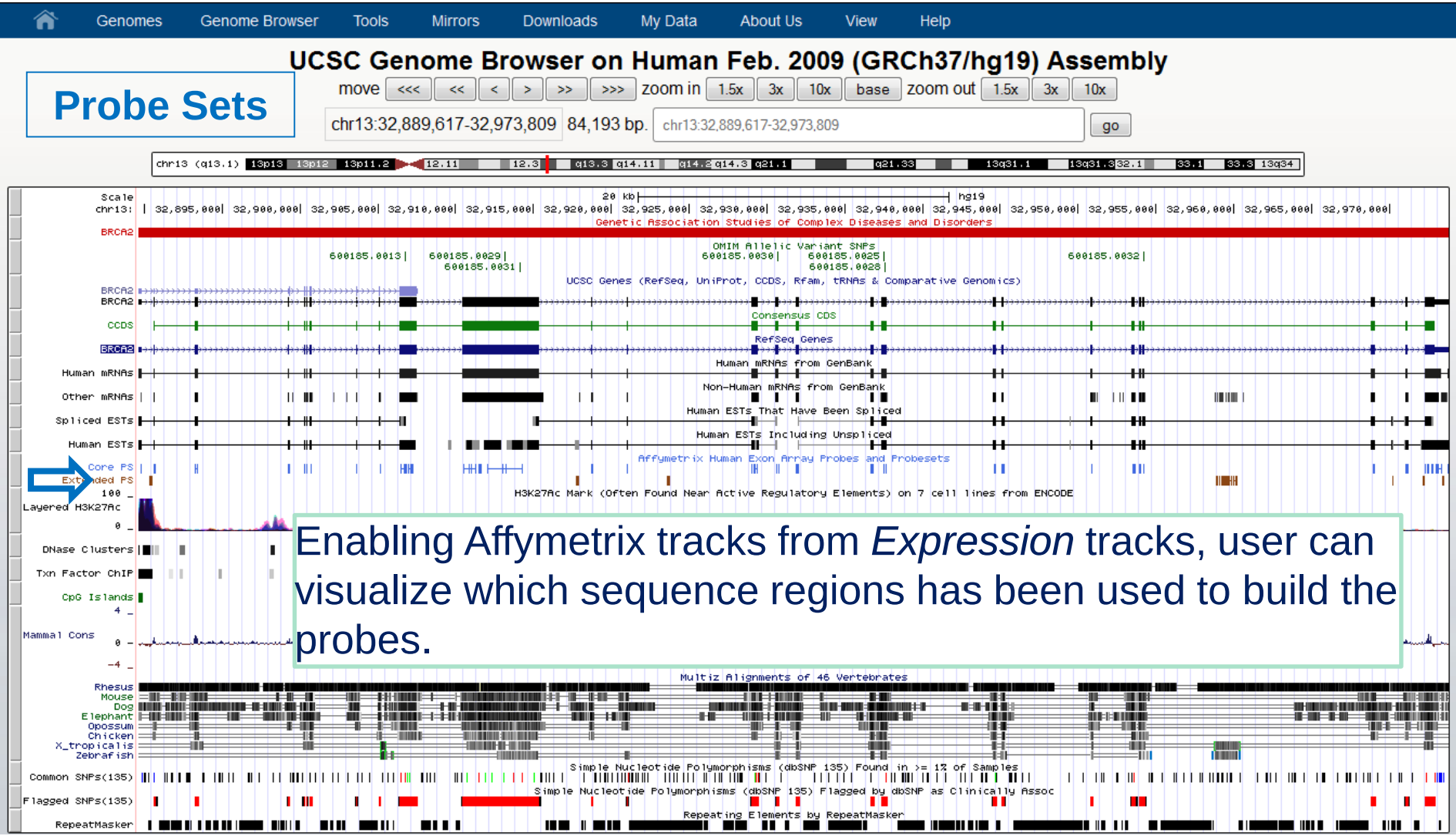


mRNAs

Annotation tracks



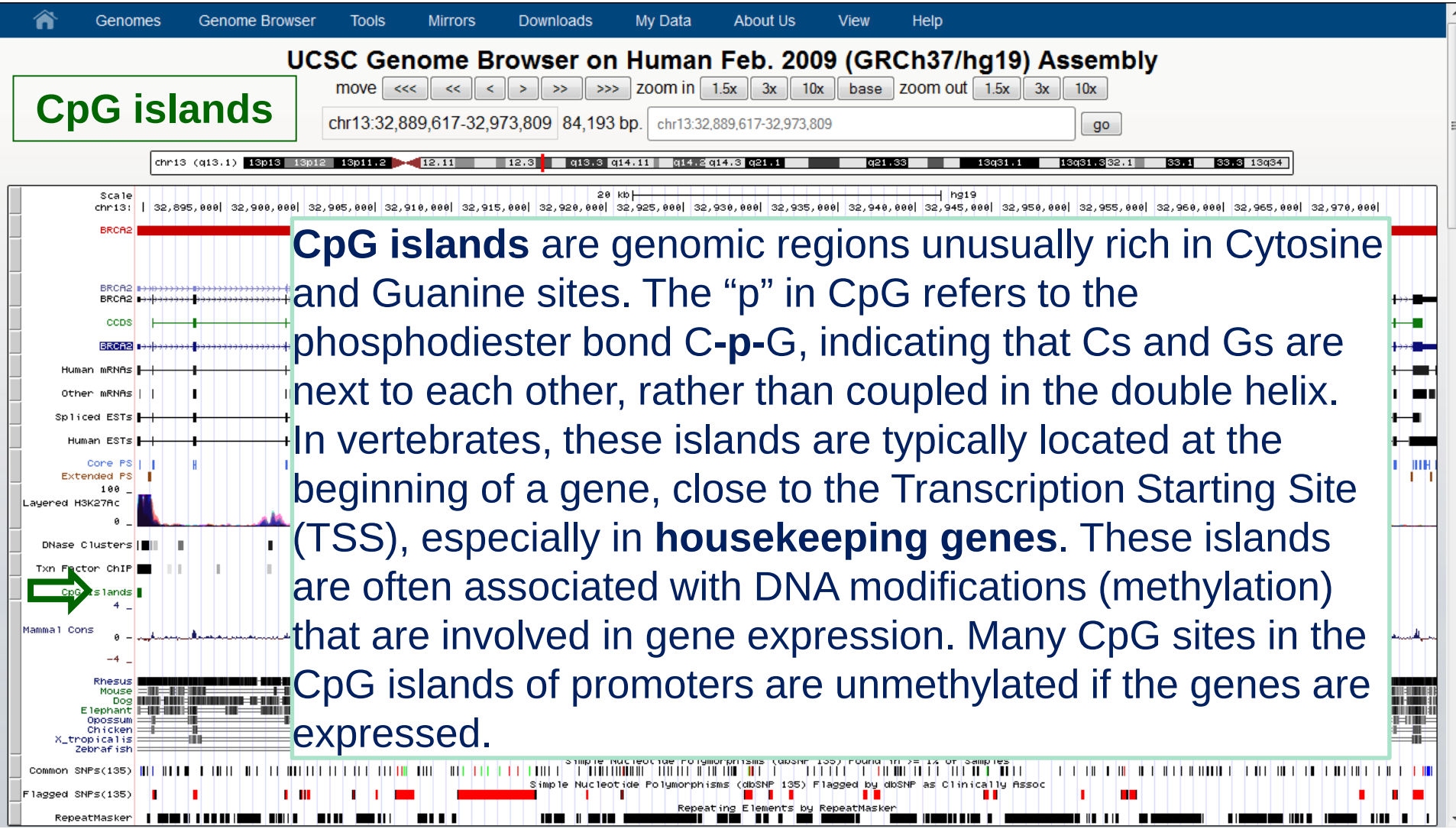
Annotation tracks



Enabling Affymetrix tracks from *Expression* tracks, user can visualize which sequence regions has been used to build the probes.

Annotation tracks

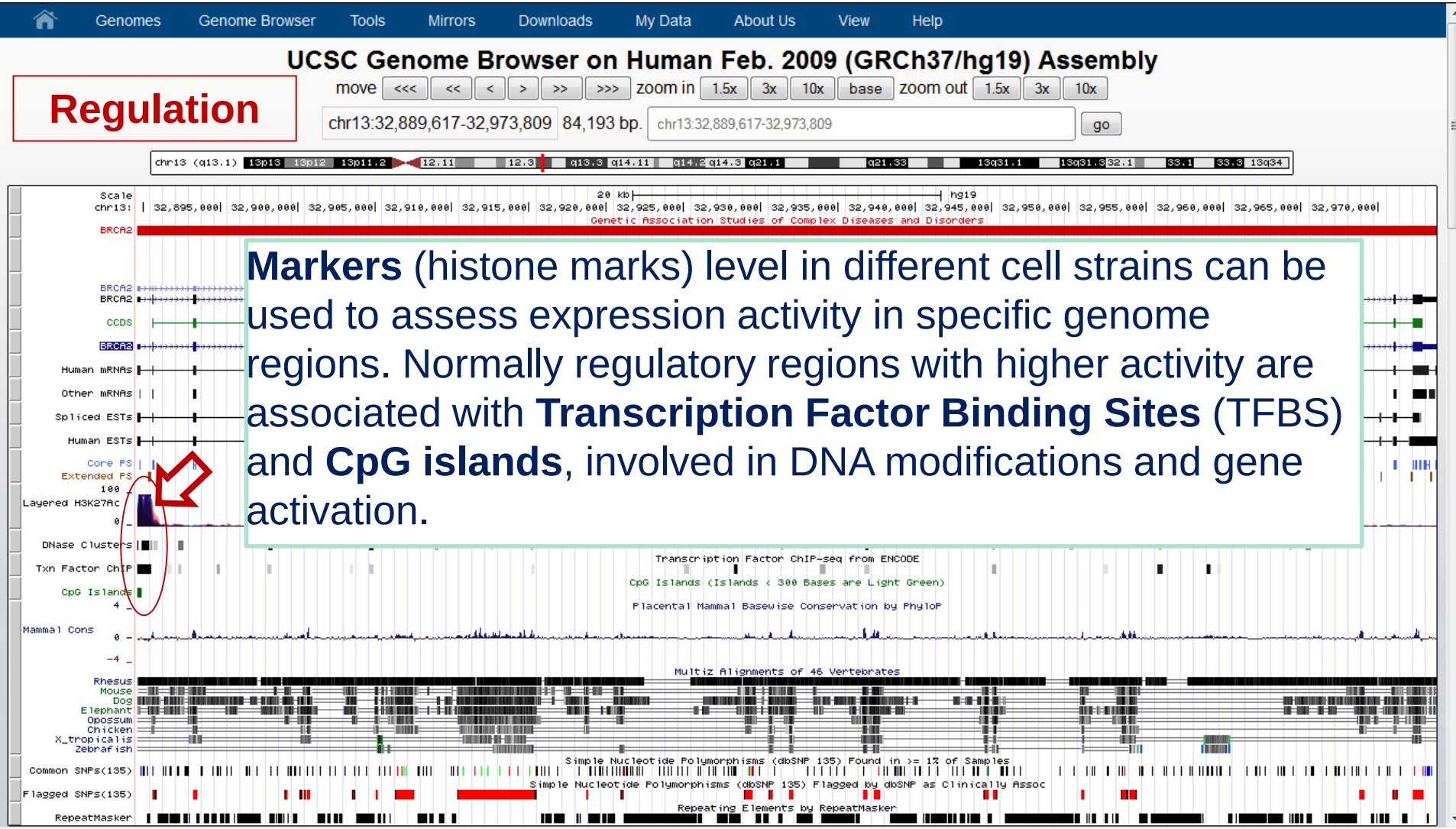
CpG islands



CpG islands are genomic regions unusually rich in Cytosine and Guanine sites. The “p” in CpG refers to the phosphodiester bond C-**p**-G, indicating that Cs and Gs are next to each other, rather than coupled in the double helix. In vertebrates, these islands are typically located at the beginning of a gene, close to the Transcription Starting Site (TSS), especially in **housekeeping genes**. These islands are often associated with DNA modifications (methylation) that are involved in gene expression. Many CpG sites in the CpG islands of promoters are unmethylated if the genes are expressed.

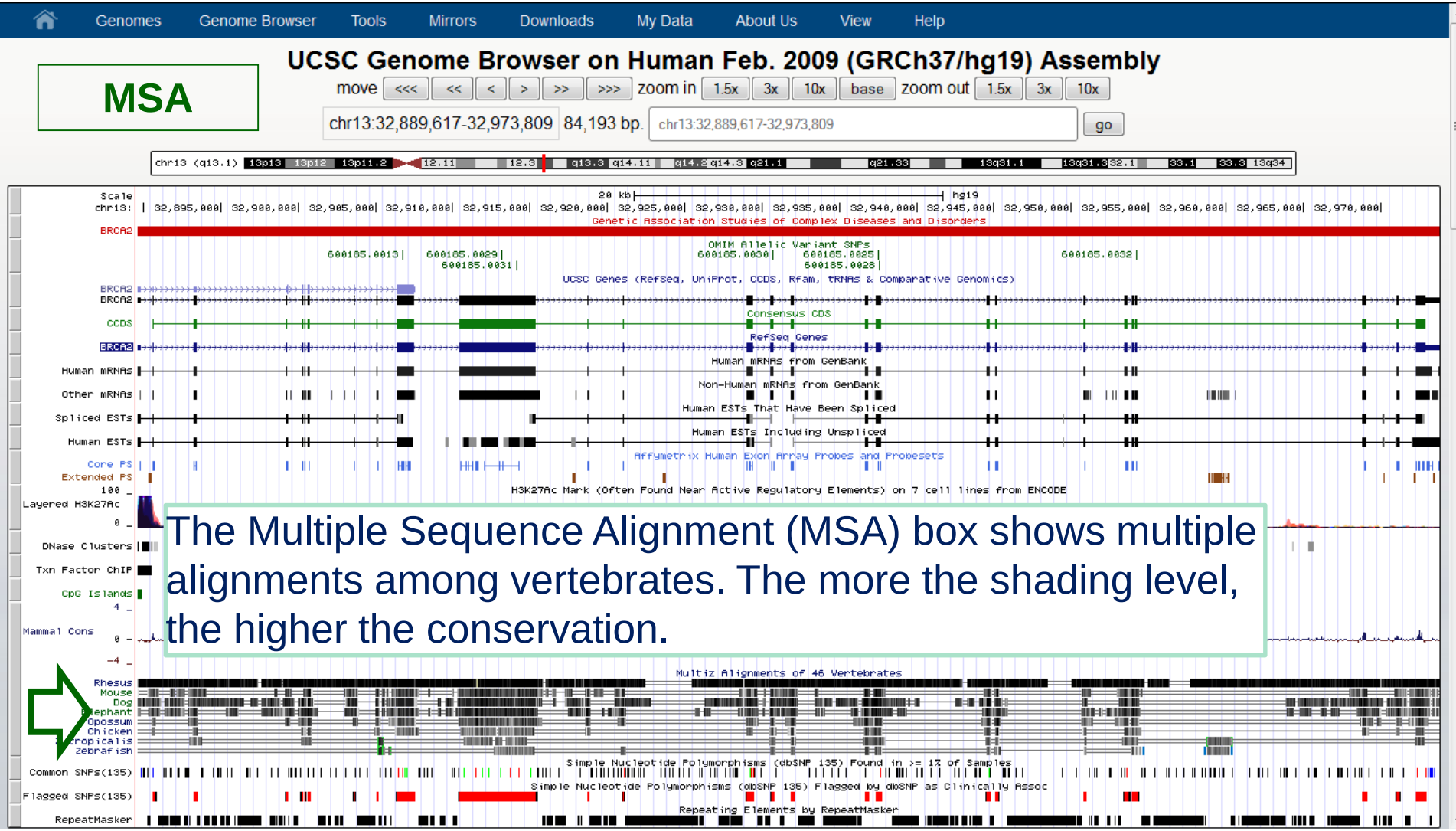
Annotation tracks

Regulation



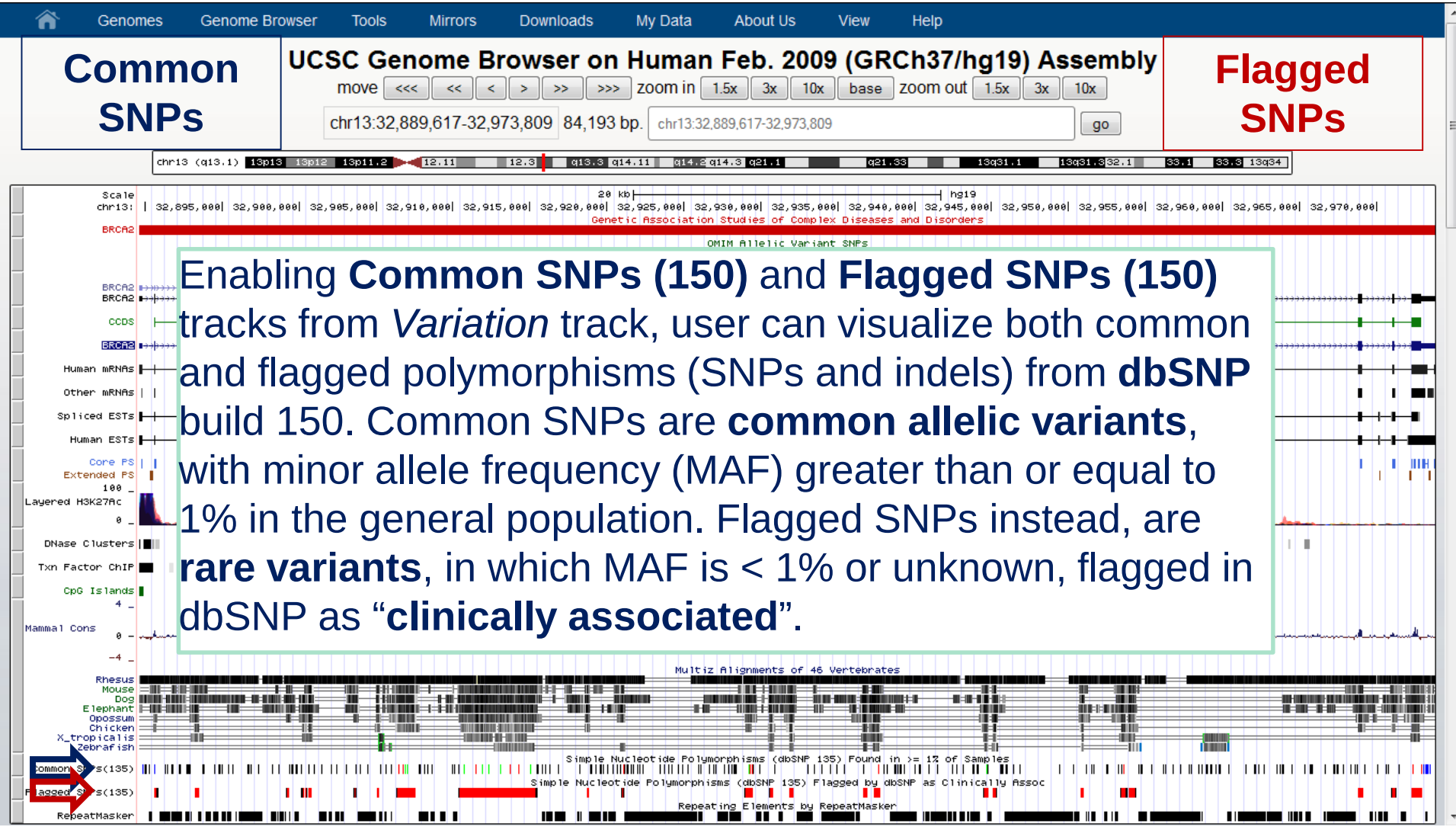
Markers (histone marks) level in different cell strains can be used to assess expression activity in specific genome regions. Normally regulatory regions with higher activity are associated with **Transcription Factor Binding Sites** (TFBS) and **CpG islands**, involved in DNA modifications and gene activation.

Annotation tracks



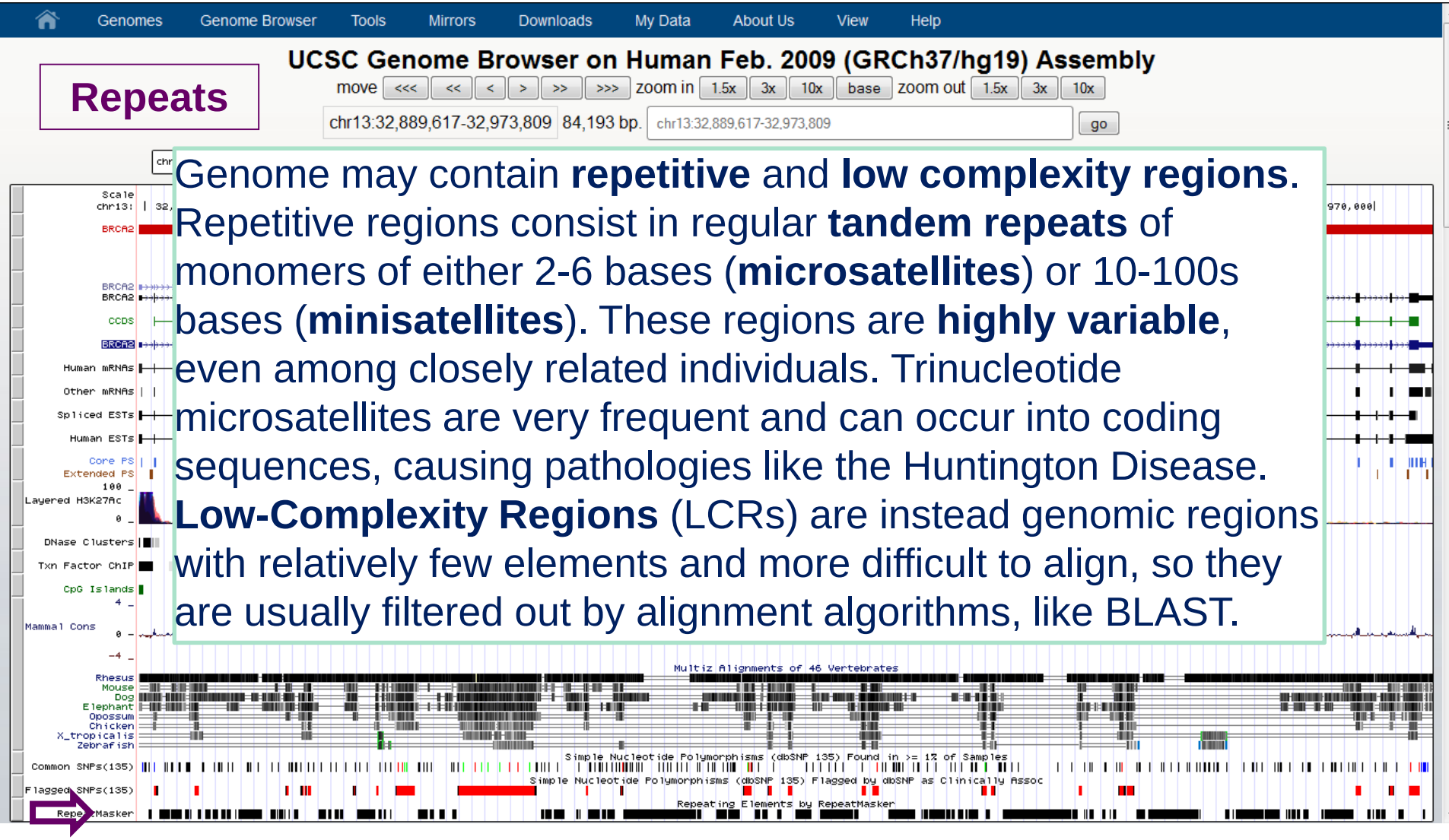
The Multiple Sequence Alignment (MSA) box shows multiple alignments among vertebrates. The more the shading level, the higher the conservation.

Annotation tracks



Enabling **Common SNPs (150)** and **Flagged SNPs (150)** tracks from *Variation* track, user can visualize both common and flagged polymorphisms (SNPs and indels) from **dbSNP** build 150. Common SNPs are **common allelic variants**, with minor allele frequency (MAF) greater than or equal to 1% in the general population. Flagged SNPs instead, are **rare variants**, in which MAF is < 1% or unknown, flagged in dbSNP as "**clinically associated**".

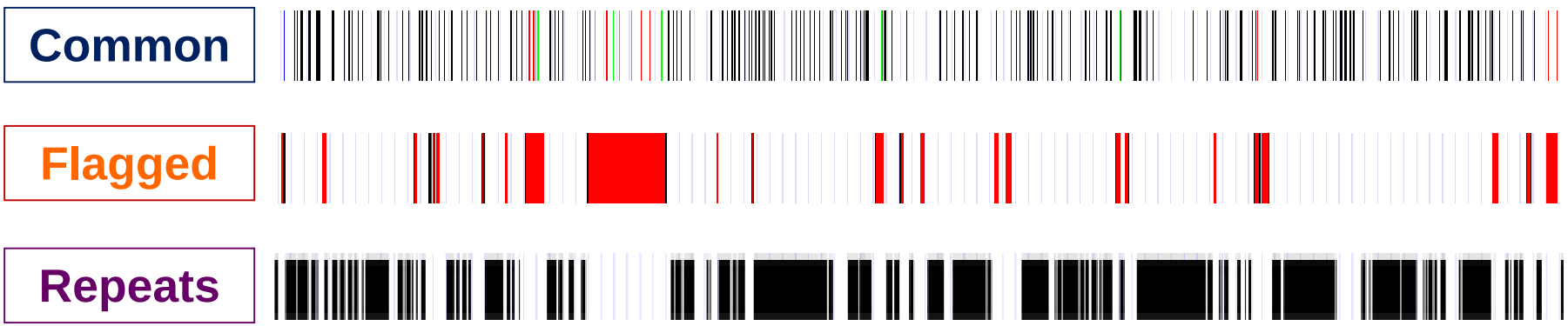
Annotation tracks



Genome may contain **repetitive** and **low complexity regions**. Repetitive regions consist in regular **tandem repeats** of monomers of either 2-6 bases (**microsatellites**) or 10-100s bases (**minisatellites**). These regions are **highly variable**, even among closely related individuals. Trinucleotide microsatellites are very frequent and can occur into coding sequences, causing pathologies like the Huntington Disease. **Low-Complexity Regions** (LCRs) are instead genomic regions with relatively few elements and more difficult to align, so they are usually filtered out by alignment algorithms, like BLAST.

Repeats, common variants and flagged SNPs … Do you see something odd?

Answer in the next slide



Synonymous SNPs

Non-synonymous SNPs

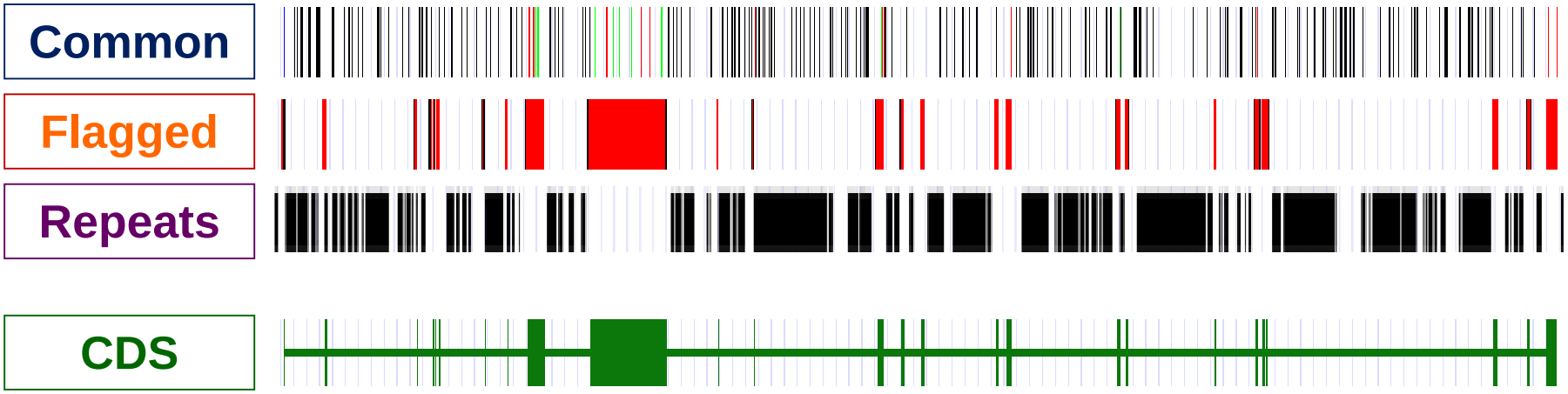
Untranslated SNPs

The higher the repeat mask shading the higher the presence of DNA sequences for interspersed repeats and low complexity regions.

Hint: look at the SNPs distribution.

Repeats, common variants and flagged SNPs … Do you see something odd?

Answer

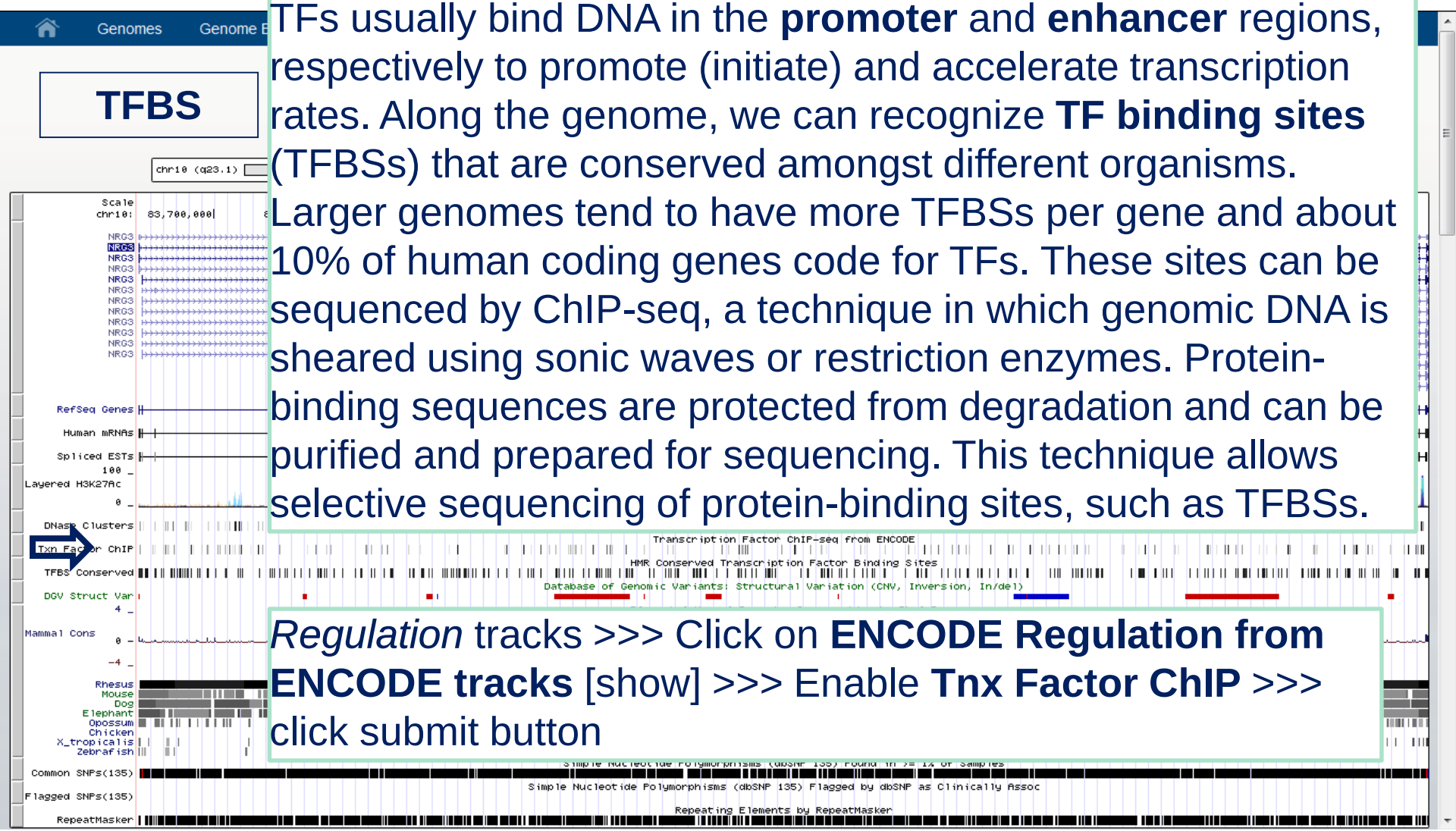


Common variants are distributed more uniformly than clinically-related (flagged) variants. Flagged variants seems to be distributed mainly in non-repetitive regions. Amongst common variants, non-synonymous changes are more often found in non-repetitive regions. This because the Coding DNA Sequence of a gene is a non-repetitive region and a change in this DNA region may be the cause of a structural and functional change in the gene product.

Annotation tracks

TFBS

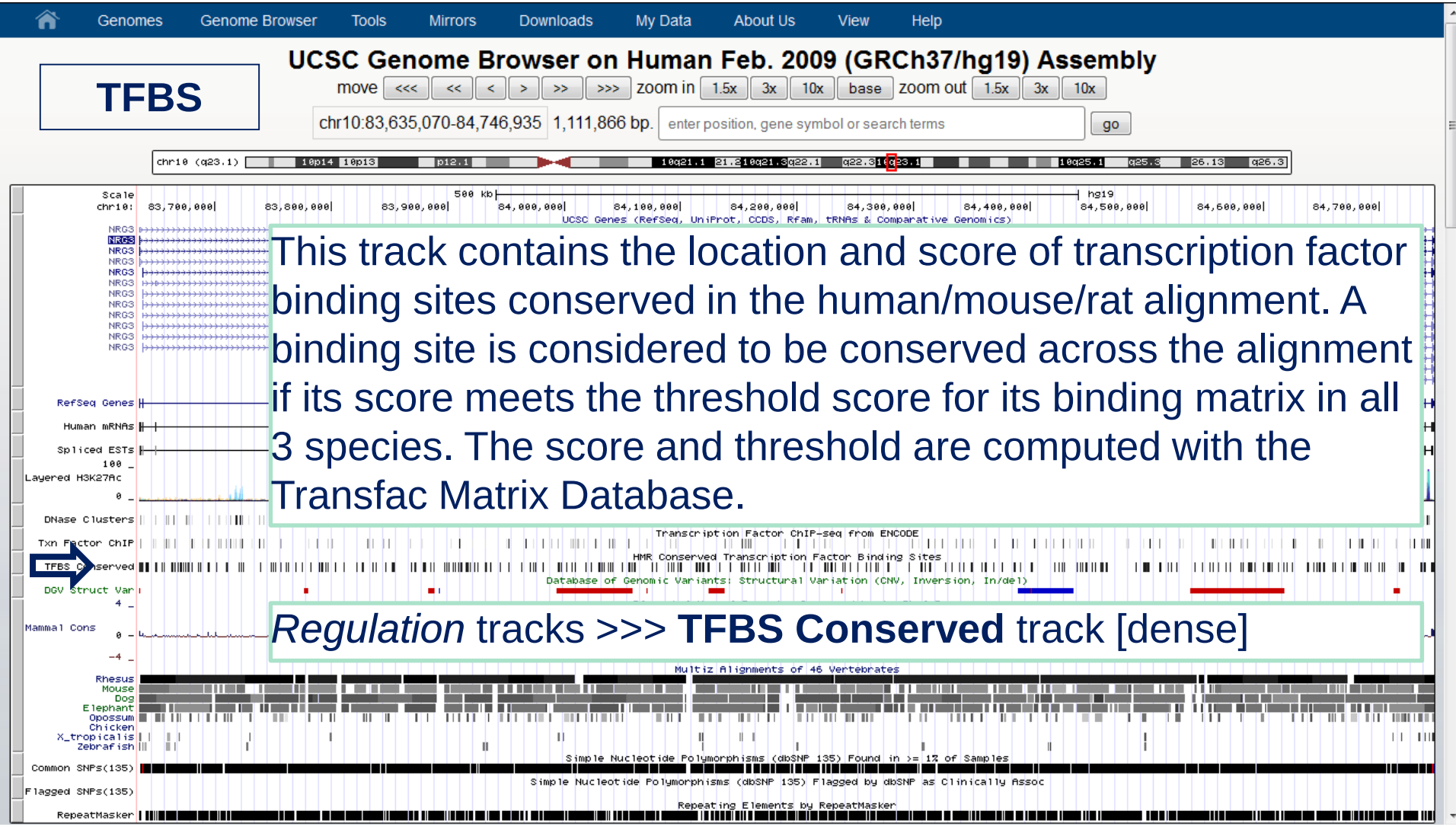
TFs usually bind DNA in the **promoter** and **enhancer** regions, respectively to promote (initiate) and accelerate transcription rates. Along the genome, we can recognize **TF binding sites** (TFBSs) that are conserved amongst different organisms. Larger genomes tend to have more TFBSs per gene and about 10% of human coding genes code for TFs. These sites can be sequenced by ChIP-seq, a technique in which genomic DNA is sheared using sonic waves or restriction enzymes. Protein-binding sequences are protected from degradation and can be purified and prepared for sequencing. This technique allows selective sequencing of protein-binding sites, such as TFBSs.



Regulation tracks >>> Click on **ENCODE Regulation from ENCODE tracks** [show] >>> Enable **Tnx Factor ChIP** >>> click submit button

Annotation tracks

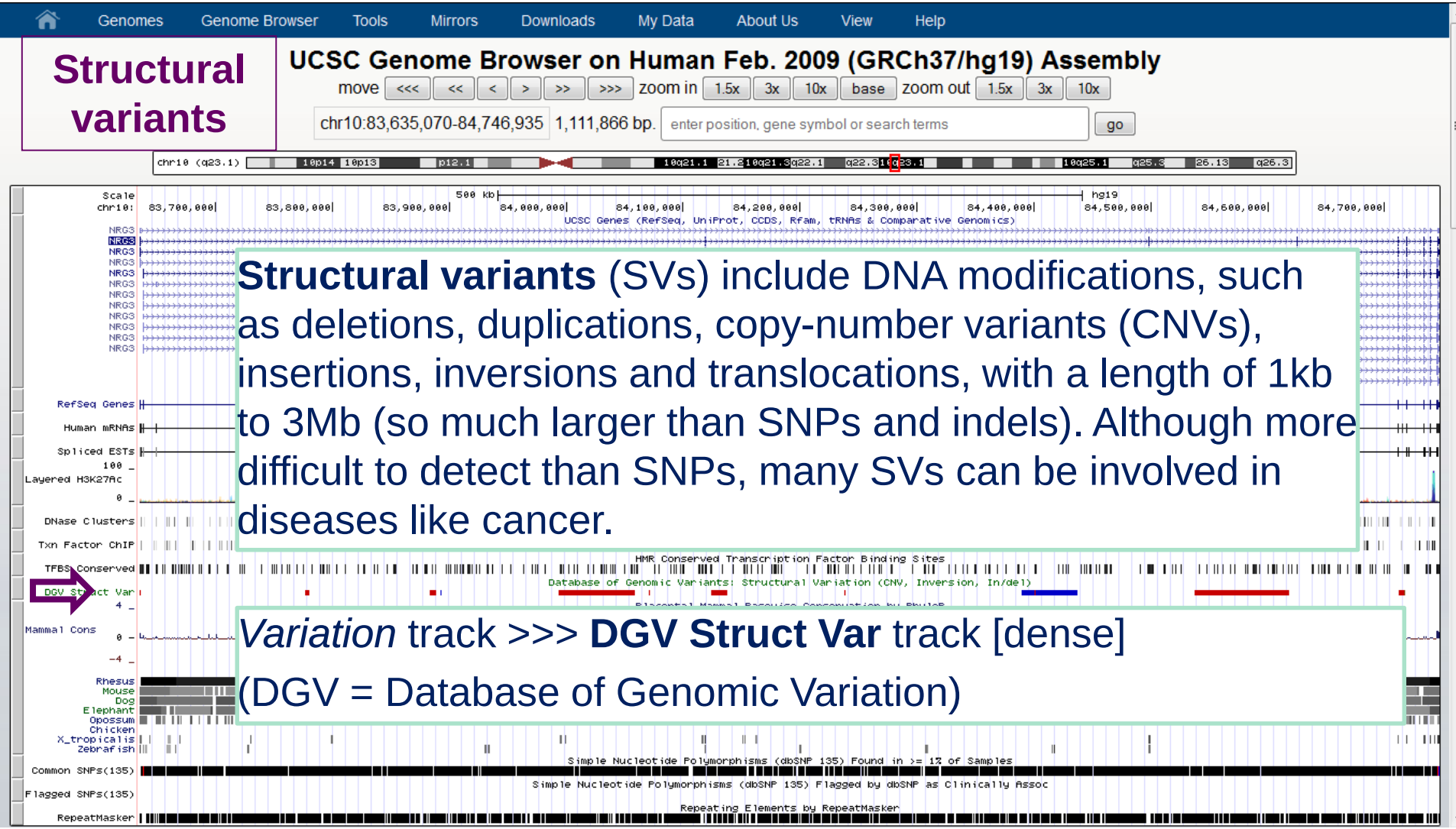
TFBS



This track contains the location and score of transcription factor binding sites conserved in the human/mouse/rat alignment. A binding site is considered to be conserved across the alignment if its score meets the threshold score for its binding matrix in all 3 species. The score and threshold are computed with the Transfac Matrix Database.

Regulation tracks >>> **TFBS Conserved** track [dense]

Annotation tracks



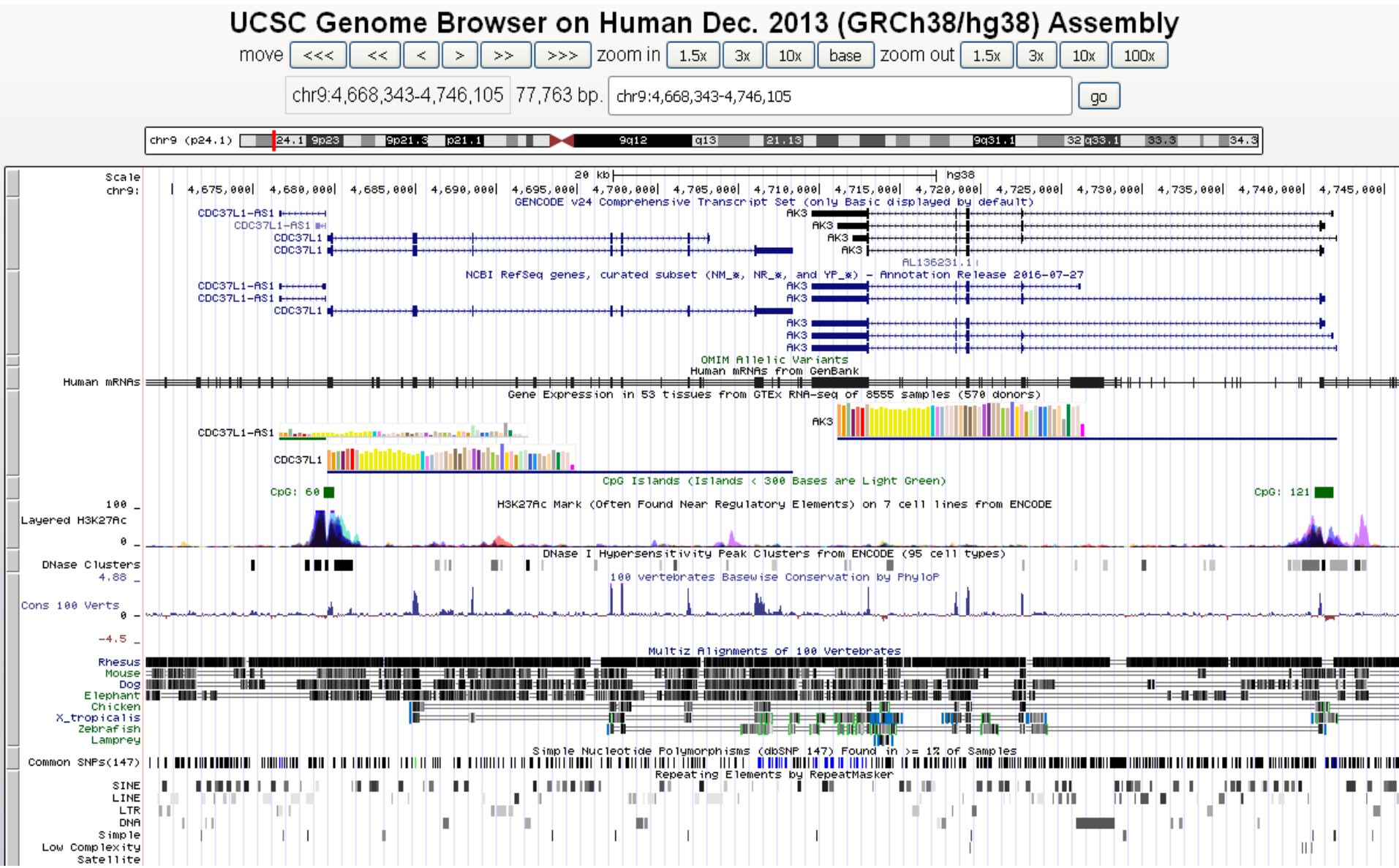
Structural variants

Structural variants (SVs) include DNA modifications, such as deletions, duplications, copy-number variants (CNVs), insertions, inversions and translocations, with a length of 1kb to 3Mb (so much larger than SNPs and indels). Although more difficult to detect than SNPs, many SVs can be involved in diseases like cancer.

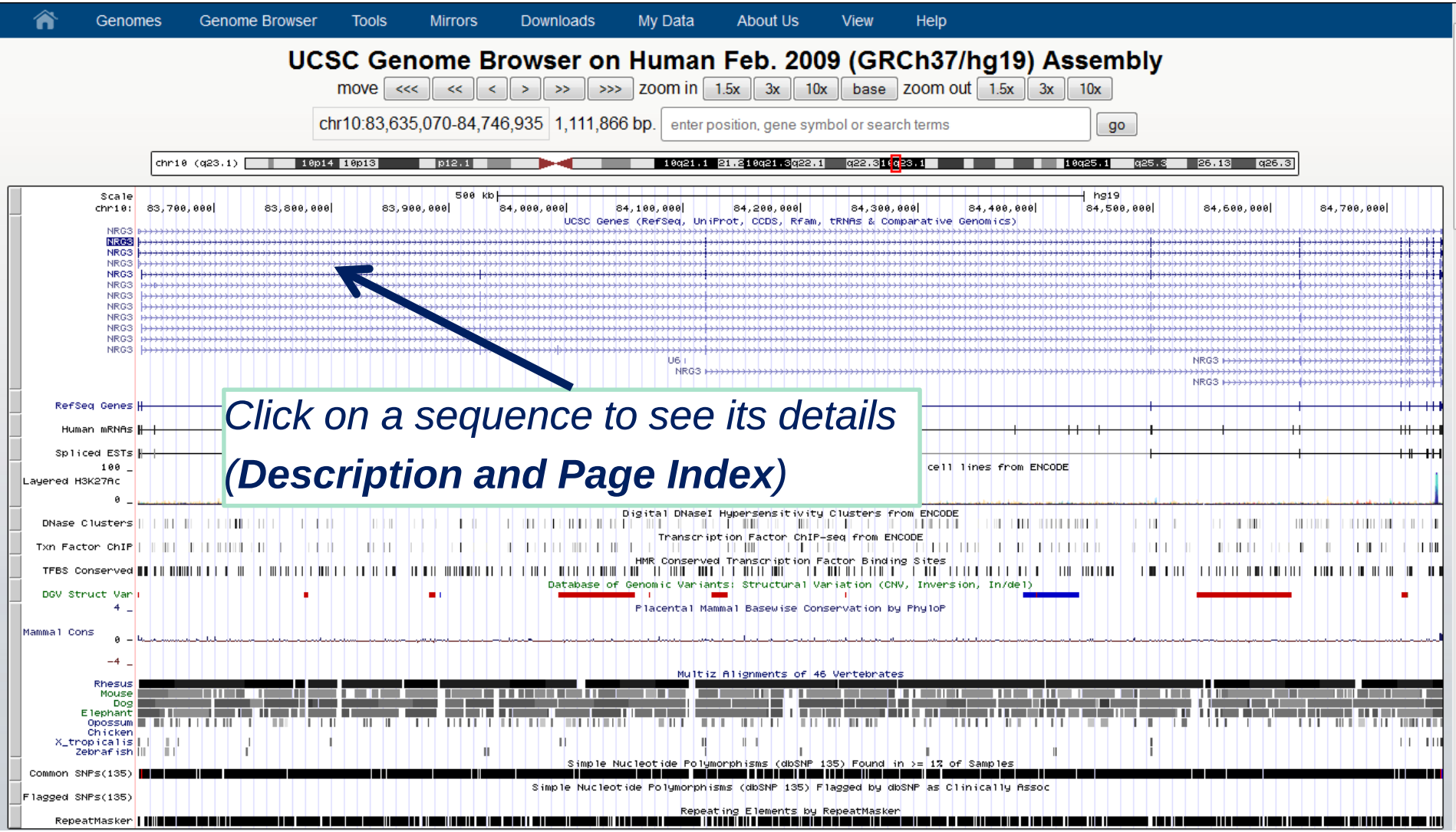
Variation track >>> **DGV Struct Var** track [dense]

(DGV = Database of Genomic Variation)

Ex. Two genes on different strands with GTEx expression, and H3K27Ac signal and CpG islands in their promoters



Sequence details



Click on a sequence to see its details

(Description and Page Index)

Sequence details

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human Gene BRCA1 (uc010whq.1) Description and Page Index

Description: Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 3, mRNA.
RefSeq Summary (NM_007299): This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. The encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC). This gene product associates with RNA polymerase II, and through the C-terminal domain, also interacts with histone deacetylase complexes. This protein thus plays a role in transcription, DNA [illegible] eaks, and recombination. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more th[illegible] nd ovarian cancers. Alternative splicing plays a role in modulating the subcellular localization and physiological function of this gene. Many alternatively spliced transcript variants, some of which are disease-associated mutations, have been described for this gene, but the full-length natures of only some of these variants has been described. A related pseudogene, which is also located on chromosome 17, has been identified. [provided by RefSeq, May 2009].

Description

Transcript (Including UTRs)
Position: hg19 chr17:41,215,350-41,256,973 **Size:** 41,624 **Total Exon Count:** 12 **Strand:** -
Coding Region
Position: hg19 chr17:41,215,350-41,256,198 **Size:** 40,849 **Coding Exon Count:** 11

Location and length

Page Index	Sequence and Links	Genetic Associations	MalaCards	CTD	Gene Alleles
RNA-Seq Expression	Microarray Expression	RNA Structure	Protein Structure	Other Species	mRNA Descriptions
Pathways	Other Names	GeneReviews	Model Information	Methods	

Annotation links

Data last updated: 2013-06-14

Sequence and Links to Tools and Databases

Genomic Sequence (chr17:41,215,350-41,256,973)			mRNA (may differ from genome)		Protein (460 aa)
Gene Sorter	Genome Browser	Other Species FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	HGNC	Lynx	MGI	MOPED	OMIM
PubMed	Reactome	Stanford SOURCE	UniProtKB		

Sequence download

Genomic sequences download

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | About Us | Help

Genomic Sequence Near Gene

Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of this track at a time, try the Table Browser using the output format sequence.

Sequence Retrieval Region Options:

- [ ] Promoter/Upstream by 1000 bases
- [x] 5' UTR Exons
- [x] CDS Exons
- [x] 3' UTR Exons
- [x] Introns
- [ ] Downstream by 1000 bases
- (•) One FASTA record per gene.
- ( ) One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
  - [ ] Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

- (•) Exons in upper case, everything else in lower case.
- ( ) CDS in upper case, UTR in lower case.
- ( ) All upper case.
- ( ) All lower case.
- [ ] Mask repeats: (•) to lower case ( ) to N

submit

Set options and Submit

UCSC Table Browser: table information download

Upper bar >>> Tools >>> Table Browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | About Us | Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Variation and Repeats **track:** Flagged SNPs(135) add custom tracks track hubs

table: snp135Flagged describe table schema

region: genome ENCODE Pilot regions position chr13:32889617-32973809 lookup define regions

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

correlation: create

output format: all fields from selected table Send output to Galaxy GREAT

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Source

Tracks

Tables

Region

Search across multiple IDs

Filters

Output

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list

UCSC Custom Track

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Add Custom Tracks

Upper bar >>> My Data >>> Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigChain, bigGenePred, bigMaf, bigPsl, bigWig, BAM, VCF, BED, BED detail, bedGraph, broadPeak, CRAM, GFF, GTF, MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are here.

Paste URLs or data: Or upload: Browse... No file selected. Submit

```
browser position chr22:20100000-20100900
track name=coords description="Chromosome coordinates list" visibility=2
chr22   20100000 20100100
chr22   20100011 20100200
chr22   20100215 20100400
chr22   20100350 20100500
chr22   20100700 20100800
chr22   20100700 20100900
```

Clear

Submit

Set custom track data

Optional track documentation: Or upload: Browse... No file selected.

Clear

Click here for an HTML document template that may be used for Genome Browser track descriptions.

BLAT: BLAST-Like Alignment Tool

- Works by keeping an **index of an entire genome in memory**.
- The index consists of all **non-overlapping DNA 11-mers** except for those heavily involved in repeats
  (It uses 4-mers for proteins).
- BLAT looks up each overlapping K-mer of the query sequence in the index, then building a list of hits where the query and the target match.
- Hits are clumped if they fall within a window limit.
- Clumps with less than the minimum number of hits are discarded, and the rest are used to define regions of the database which are homologous to the query sequence.
- Designed to quickly find DNA sequences of 95% and greater similarity of length 40 bases or more
  ($\geq$ 80% similarity for peptide sequences longer than 20aa).
- It may miss more divergent or short sequence alignments.
- BLAT is much faster than BLAST, but less accurate.

BLAT: BLAST-Like Alignment Tool

Human BLAT Search

BLAT Search Genome

Genome:	Assembly:	Query type:	Sort output:	Output type:
Human	Feb. 2009 (GRCh37/hg19)	BLAT's guess	query,score	hyperlink

```
> read1
ACCACATATTTTGCAAATTTTGCATGCTGAAACTTCTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGA
TATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTA
> read2
ACCACATATTTTGCAAATTTTGCATGCTGATACTTCTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGA
TATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTA
> read3
ACCACATATTTTGCAAATTTTGCATGCTGATACTACTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGA
TATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTA
> read4
ACCACATATTTTGCAAATTTTGCATGCTGATACTACTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGA
TATAACCAAAAGGAGCCTAAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTA
> read5
ACCACATATTTTGCAAATTTGCATGCTGAAACTTCTCAACCAGAAGAAAGGGCCTTCACAGTGTCCTTTATGTAAGAATGAT
ATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTA
```

submit | I'm feeling lucky | clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence: Sfoglia... submit file

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.

For locating PCR primers, use In-Silico PCR for best results instead of BLAT.

About BLAT

I/O settings

Copy and paste your sequences in FASTA format or upload a FASTA file

BLAT: BLAST-Like Alignment Tool

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human BLAT Results

BLAT Search Results

Go back to chr17:41276498-41278247 on the Genome Browser.

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	hg19_knownGene_uc010whs.1	200	1	200	200	100.0%	17	-	41277269	41277468	200
browser details	hg19_knownGene_uc010whs.1	156	18	200	200	92.9%	17	-	41321755	41322280	526
browser details	hg19_knownGene_uc010whs.1	20	170	189	200	100.0%	10	+	4814874	4814893	20

Missing a match?

Click to see sequence in browser

SCORE

IDENTITY %

Useful links

- UCSC Genome Browser [http://genome.ucsc.edu/]
- Ensembl Genome Browser [http://www.ensembl.org]
- NCBI [http://www.ncbi.nlm.nih.gov/]
- dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/]
- OMIM [http://www.ncbi.nlm.nih.gov/omim]
- CCDS database [http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi]
- DGV [http://projects.tcag.ca/variation/]

- D. Karolchick, A.S. Hinrichs, W.J. Kent. **The UCSC Genome Browser**. *Curr. Protoc. Bioinformatics* 2009. doi:10.1093/bib/bbs038
- R.M. Kuhn, D. Haussler, W.J. Kent. **The UCSC Genome Browser and associated tools**. *Briefings in Bioinformatics* 2012. doi:10.1002/0471250953.bi0104s28

Let’s start with practice!

http://genome.ucsc.edu/