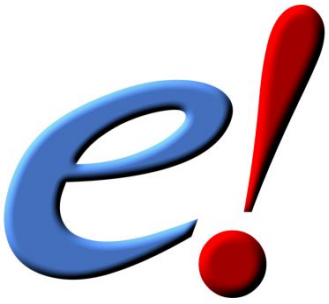
Browsing Variation Data with Ensembl



www.ensembl.org
www.ensemblgenomes.org

Coursebook v95

http://training.ensembl.org/events/2019/2019-04-02-VEP_Breda

Variant Effect Prediction Course, Breda
2nd-5th April 2019

Introduction to Ensembl

Getting started with Ensembl
www.ensembl.org

Ensembl is a project based at the EBI (European Bioinformatics Institute) that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'Compara' team. Most annotation is updated every two to three months to generate increasing Ensembl versions (84, 85, 86, etc.); however, the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates—namely, bacteria, plants, fungi, metazoa, and protists.

Ensembl provides genes and other **annotation** such as predicted regulatory regions, base pairs conserved across species, and observed sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (**Perl API**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge on the part of the user.

Synopsis — What can I do with Ensembl?

- View genes, with other annotation, along the chromosome.
- View alternative transcripts (such as splice variants) for a given gene.
- For any gene, explore homologues and phylogenetic trees across more than 70 species.
- Compare whole genome alignments and conserved regions across species.
- View microarray sequences matching Ensembl genes.
- View ESTs, clones, mRNAs, and proteins for any chromosomal region.
- Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region.
- View SNPs across strains (rat, mouse), populations (human), or breeds (dog).
- View positions and sequences of mRNAs and proteins that align against Ensembl genes.
- Upload your own data.
- Use BLAST or BLAT against any Ensembl genome.
- Export sequence or create a table of gene information with BioMart.
- Determine how your variants affect genes and transcripts using the Variant Effect Predictor.
- Share Ensembl views with your colleagues and collaborators.

Need more help?

- Check Ensembl documentation
- Watch video tutorials on YouTube
- View the FAQs
- Try some exercises
- Read some publications
- Go to our online course

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list
- **Find us on Facebook or follow us on Twitter**
  - https://www.facebook.com/Ensembl.org/
  - @ensembl
  - @ensemblgenomes

Further reading

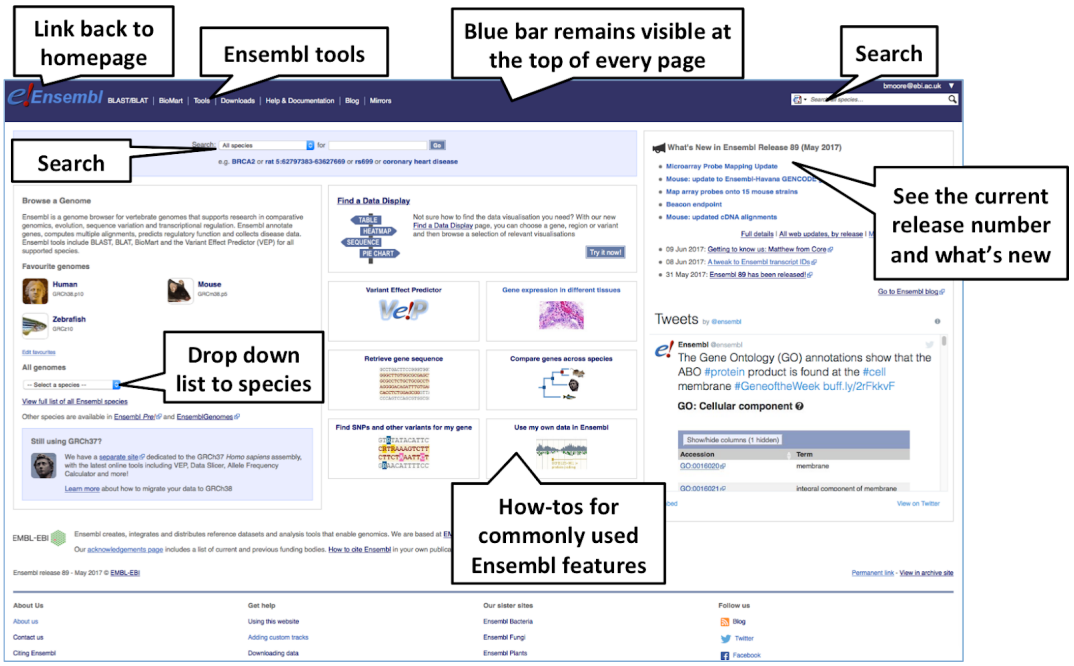
Cunningham, F. *et al.*
Ensembl 2019
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gky1113

Kersey, PJ *et al.*
Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gkx1011

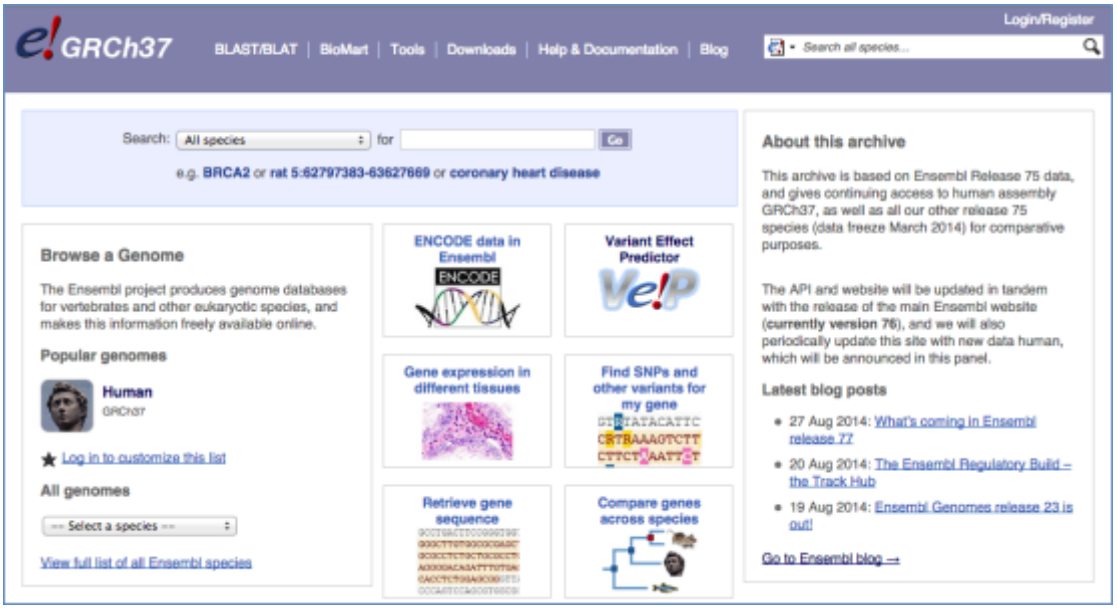
For a complete list of publications, visit:
http://www.ensembl.org/info/about/publications.html
http://ensemblgenomes.org/info/publications

Exploring variation data in the Ensembl genome browser

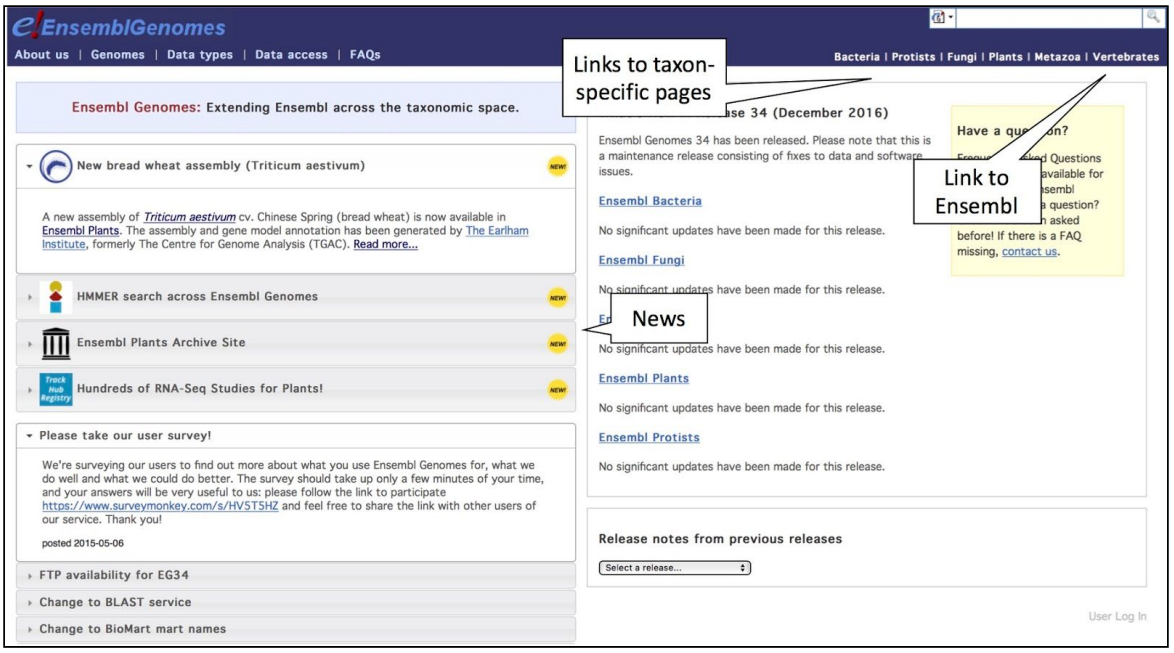
The front page of Ensembl is found at ensembl.org. It contains lots of information and links to help you navigate Ensembl:



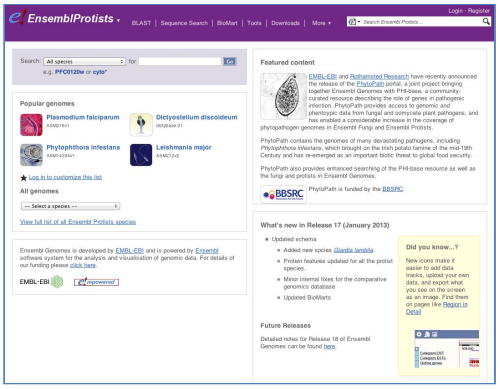
The current genome assembly for human is GRCh38. If you want to see the previous assembly, GRCh37, visit our dedicated site, grch37.ensembl.org.



Let's take a look at the Ensembl Genomes homepage at ensemblgenomes.org.

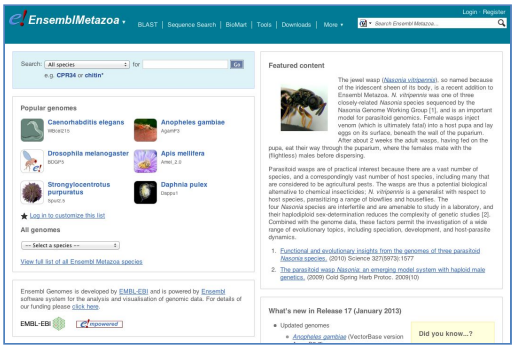


Click on the different taxa to see their homepages. Each one is colour-coded.

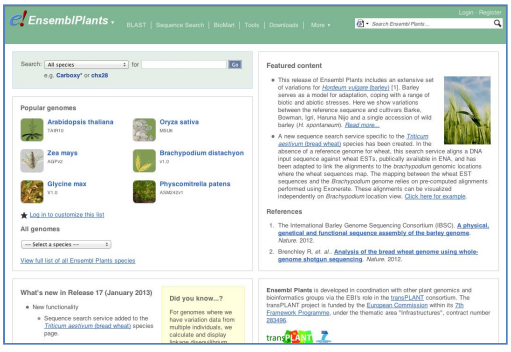


Protists

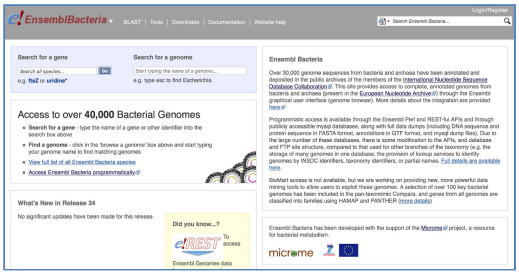
Fungi



Metazoa



Plants

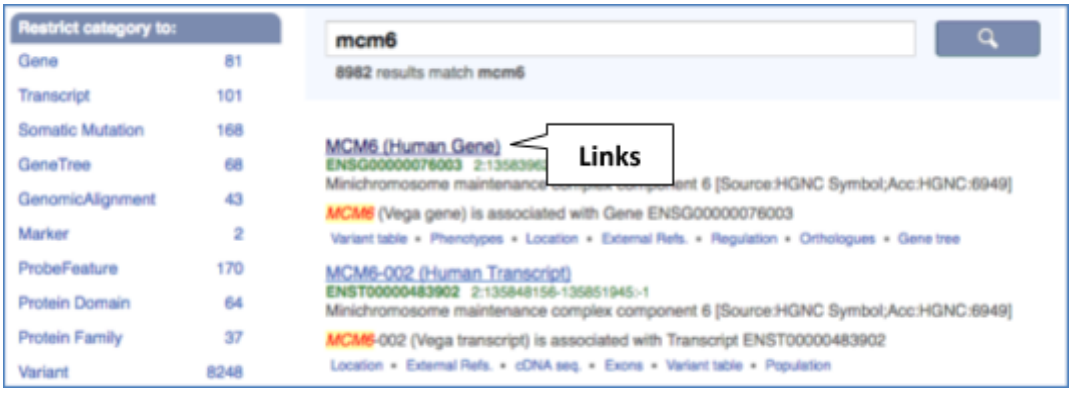


Bacteria

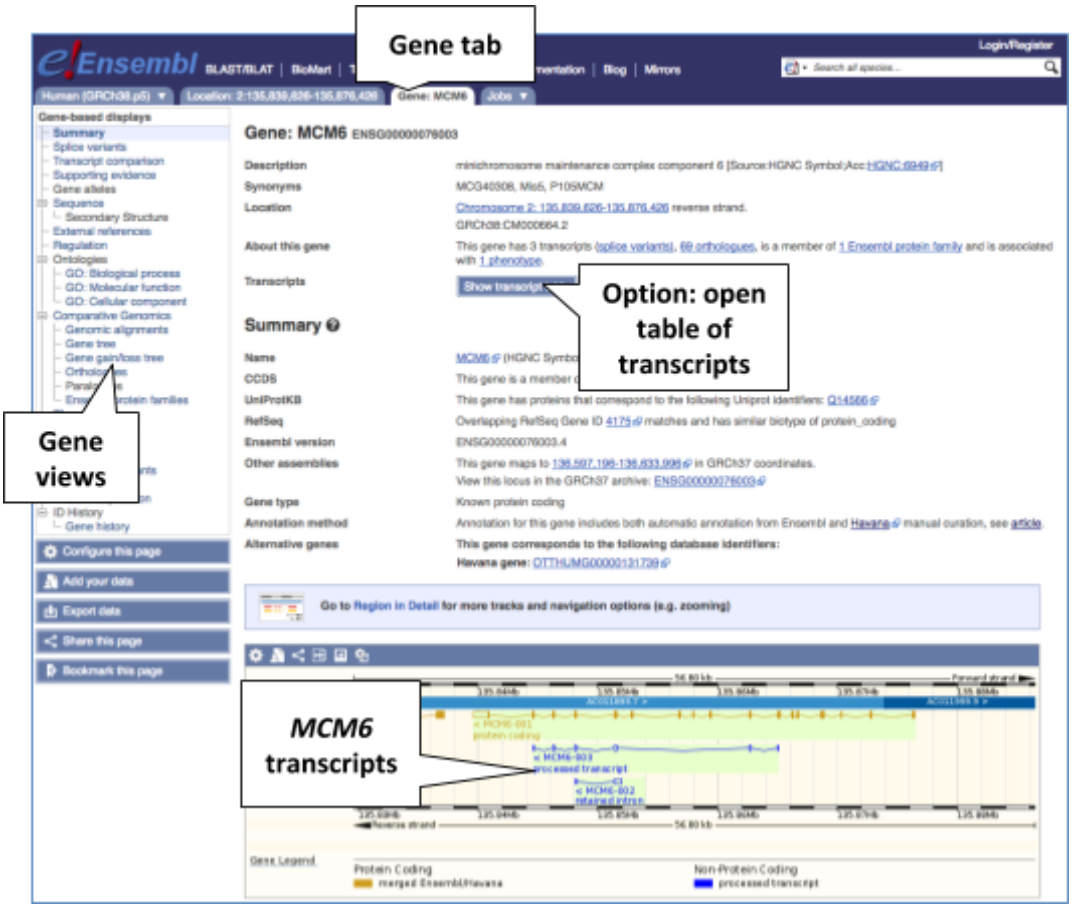
We're going to look at the human *MCM6* gene.

From ensembl.org, type *MCM6* into the search bar and click the Go button. You will get a list of hits with the human gene at the top.

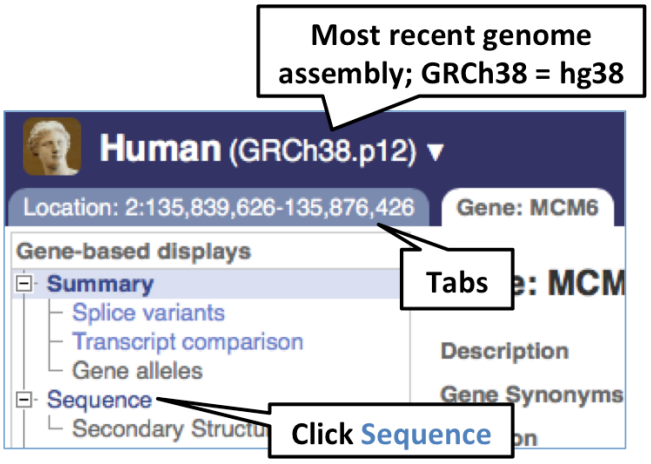
Where you search for something without specifying the species, or where the ID is not restricted to a single species, the most popular species will appear first, in this case, human, mouse and zebrafish appear first. You can restrict your query to species or features of interest using the options on the left.

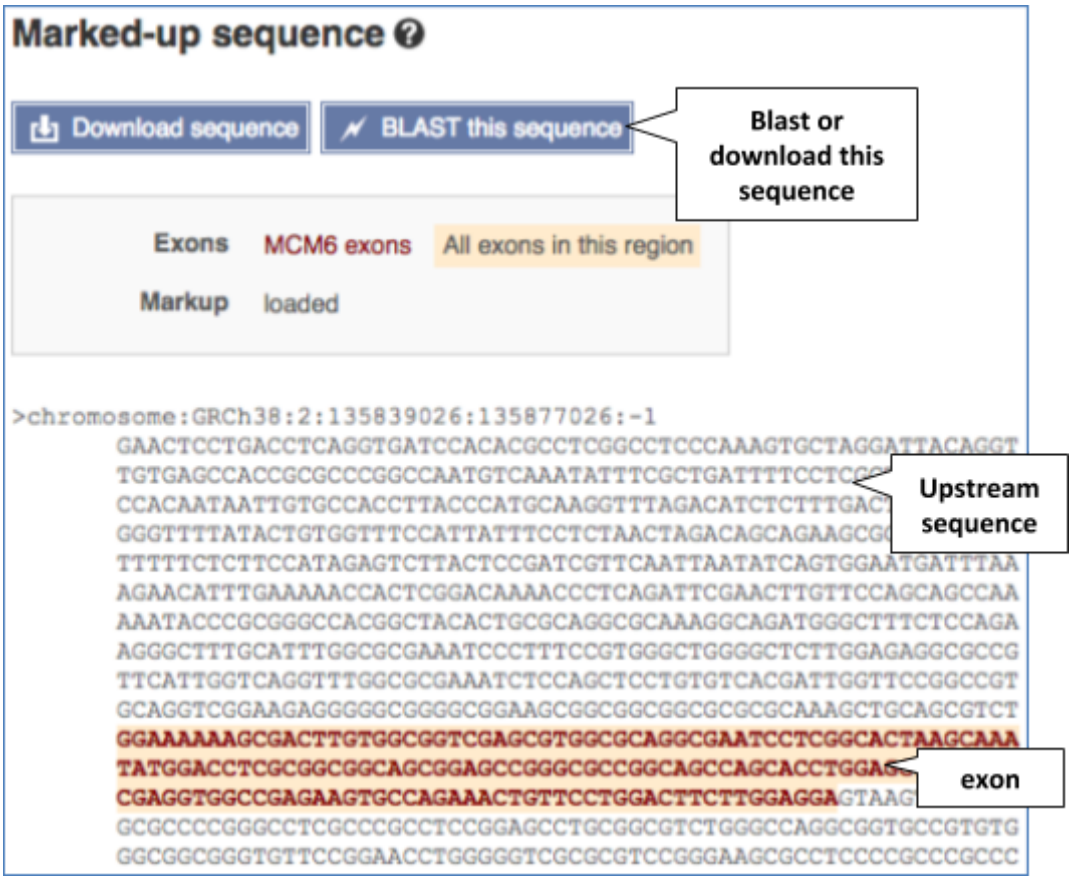


Click on the gene name or Ensembl ID. The **Gene tab** should open:

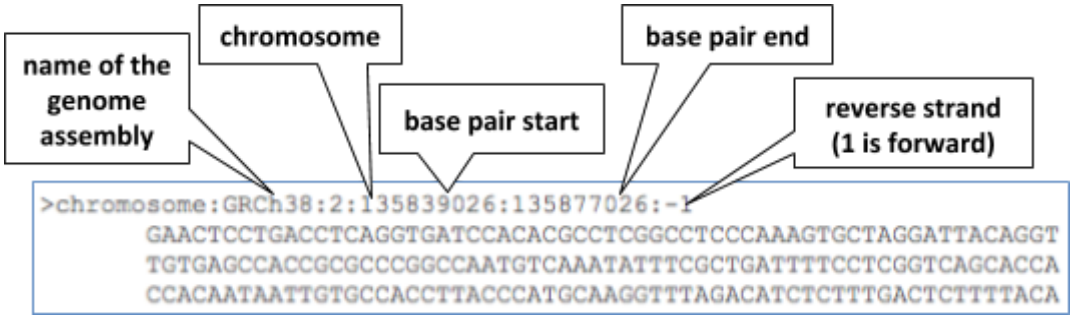


Let's walk through some of the links in the left hand navigation column to find some variation data. How can we view the genomic sequence? Click Sequence at the left of the page.

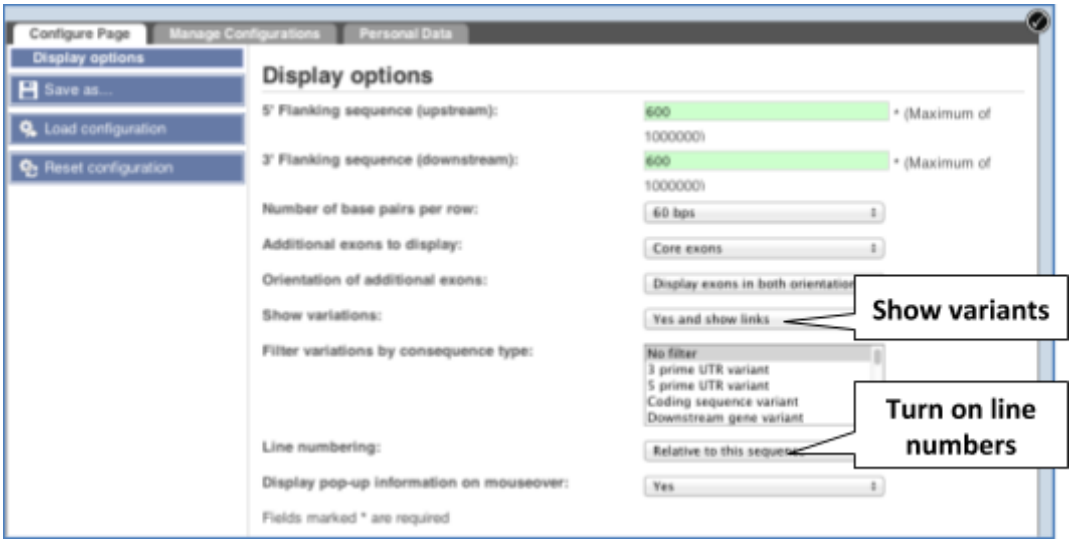




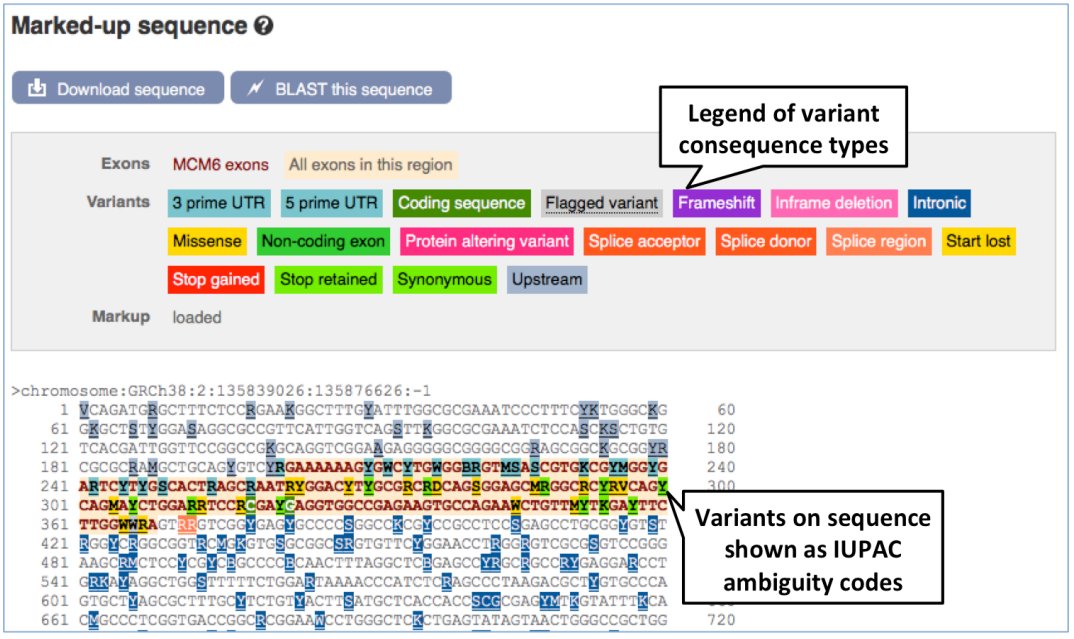
The sequence is shown in FASTA format. Take a look at the FASTA header:



Exons are highlighted within the genomic sequence. Variants can be added with the Configure this page link found at the left. Click on it now.



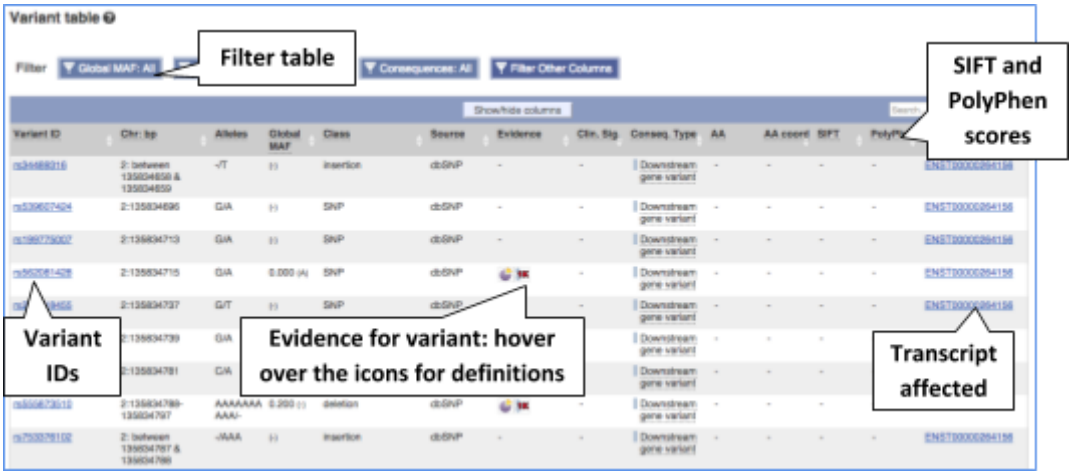
Once you have selected changes (in this example, Show variants: Yes and show links and Line numbering) click at the top right.



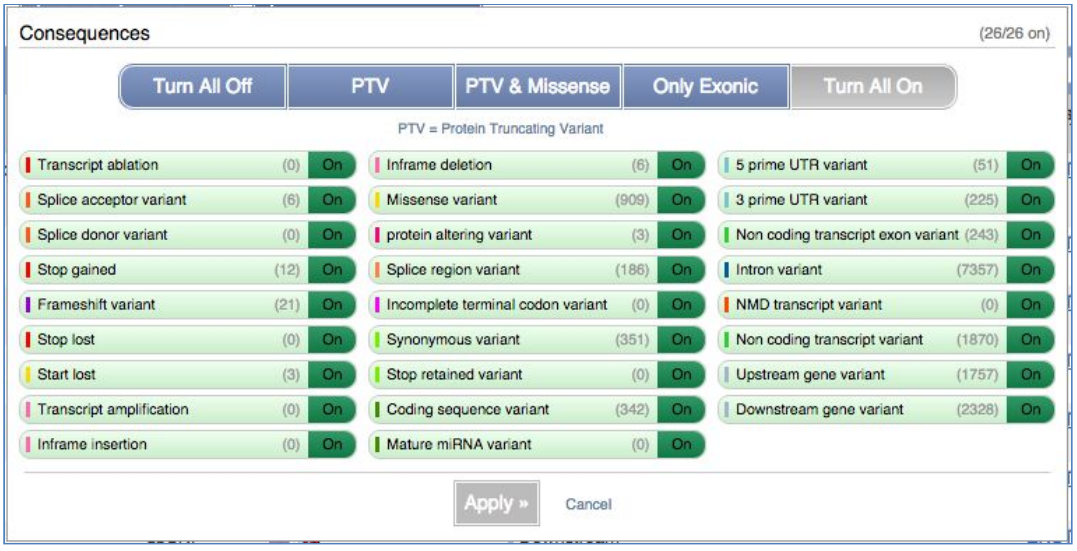
Find out more about a variant by clicking on it.

You can go to the Variation tab by clicking on the variant ID. For now, we'll explore more ways of finding variants.

To view all the sequence variations in table form, click the Variant table link at the left of the gene tab.



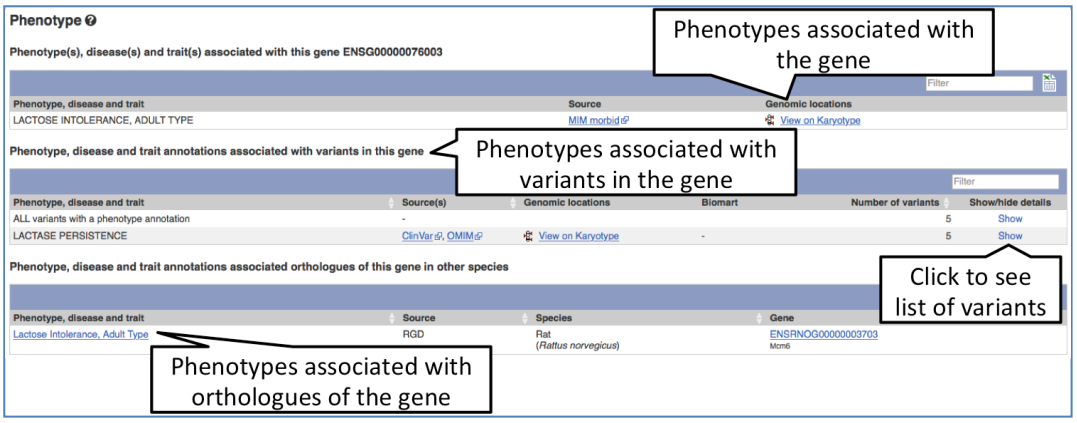
You can filter the table to only show the variants you're interested in. For example, click on Consequences: All, then select the variant consequences you're interested in.



You can also filter by SIFT, PolyPhen and MAF, or click on Filter other columns for filtering by other columns such as, Evidence or Class.

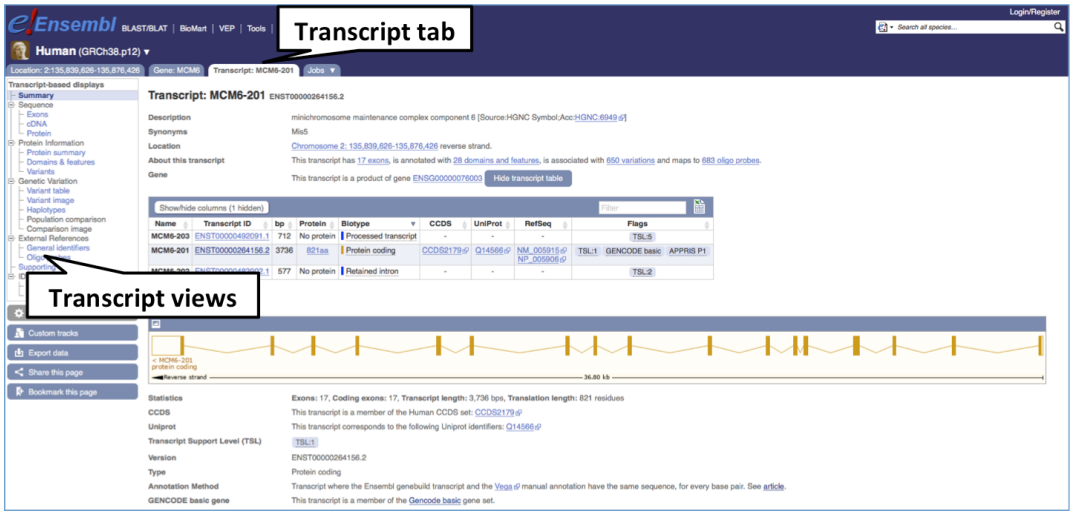
The table contains lots of information about the variants. You can click on the IDs here to go to the **Variant tab** too.

You can also see the phenotypes associated with a gene. Click on Phenotype in the left-hand menu.

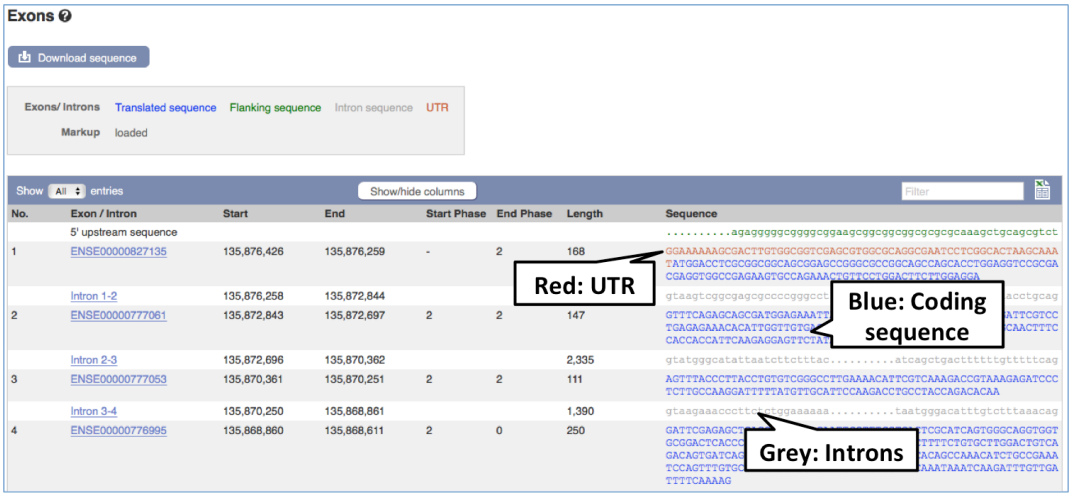


We can also explore variants per transcript. Let's now explore one splice isoform. Click on Show transcript table at the top.

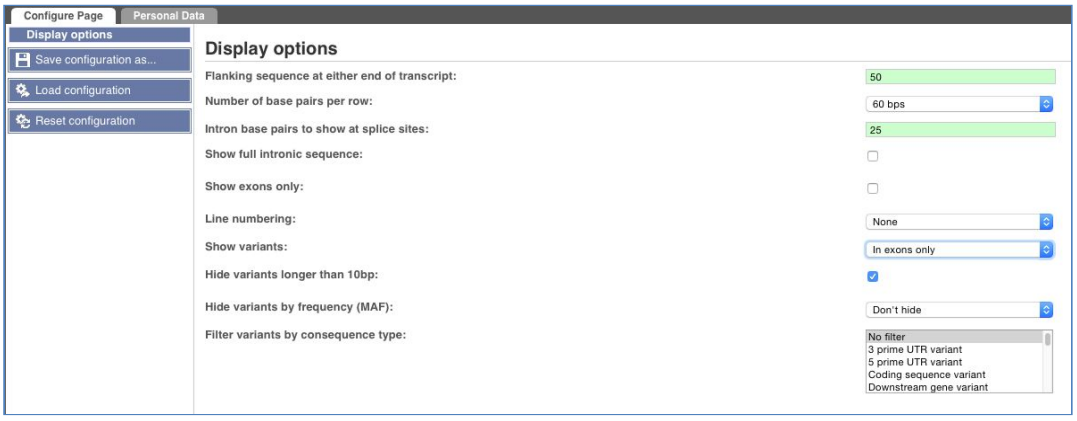
Select MCM6-201 from the transcript table to open the Transcript tab.



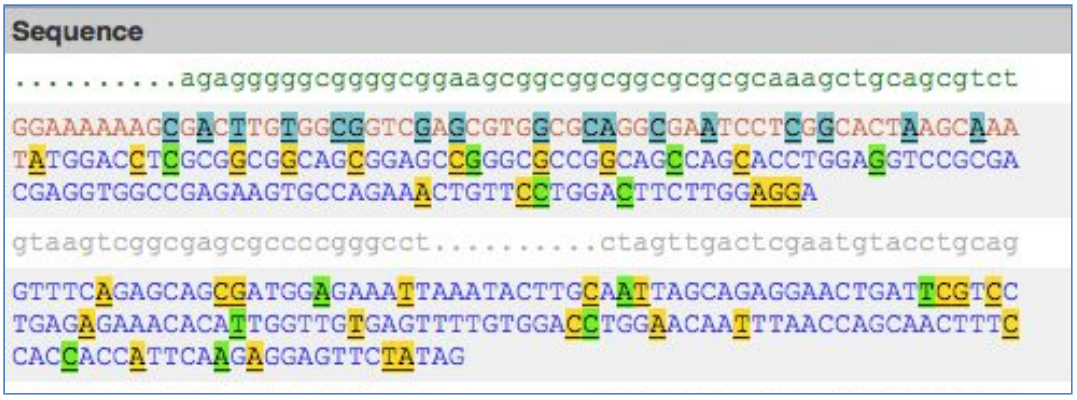
You are now in the Transcript tab for MCM6-201. The left hand navigation column provides several options for the transcript MCM6-201. Click on the Exons link.



To add variation data, click on Configure this page and change the display options accordingly.

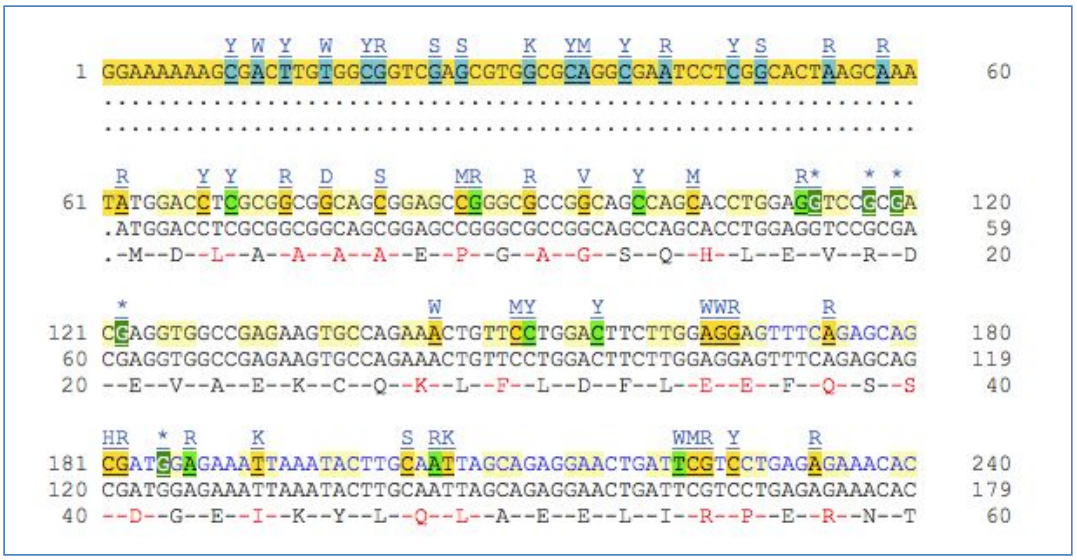


You should now see a marked-up sequence that is annotated with variation data:



Now click on the cDNA link to see the spliced transcript sequence.

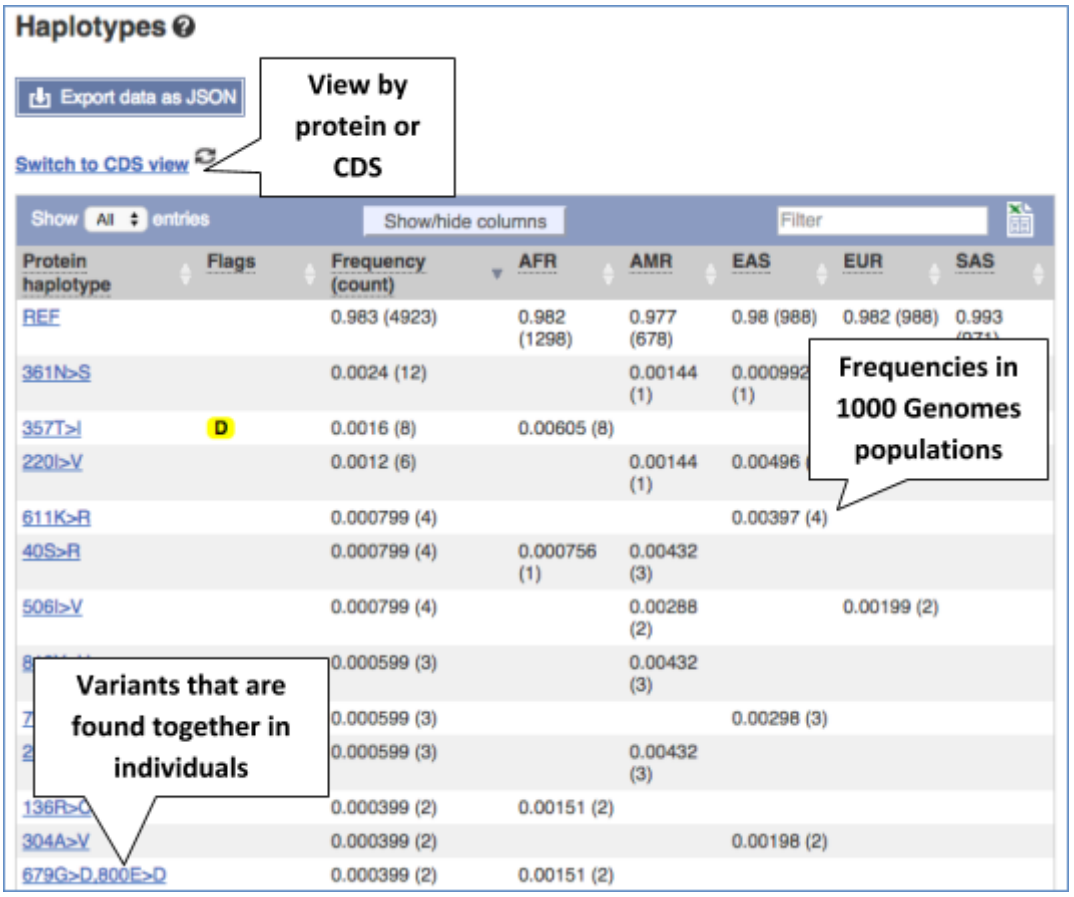
To add variation data, click on Configure this page and change the display options accordingly.



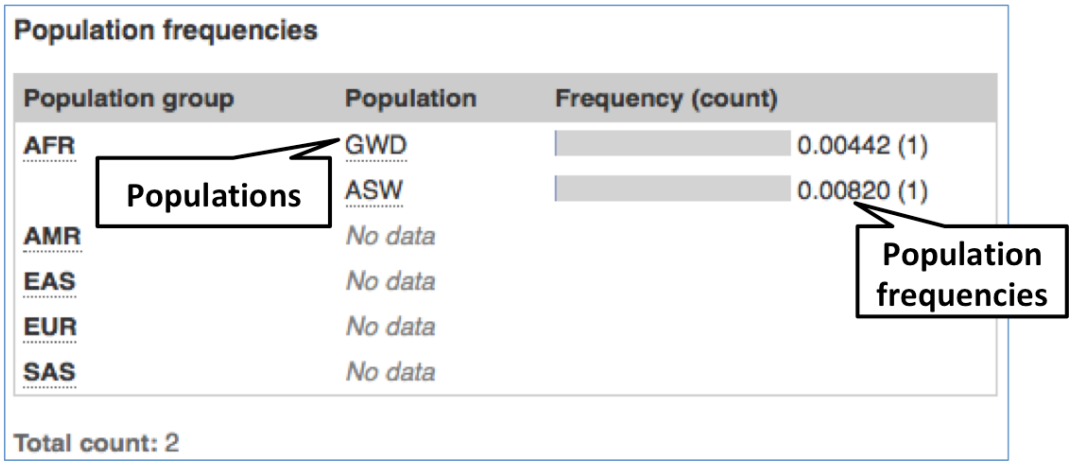
UnTranslated Regions (UTRs) are highlighted in dark yellow, codons are highlighted in alternating light yellow, and exon sequence is shown in alternating black and blue.

You can also analyse the compound effect of variants using the Transcript Haplotype view. The Transcript Haplotype view allows you to explore observed transcript sequences of individuals that result from variants identified by the **1000 Genomes Project**. Click on Haplotypes.

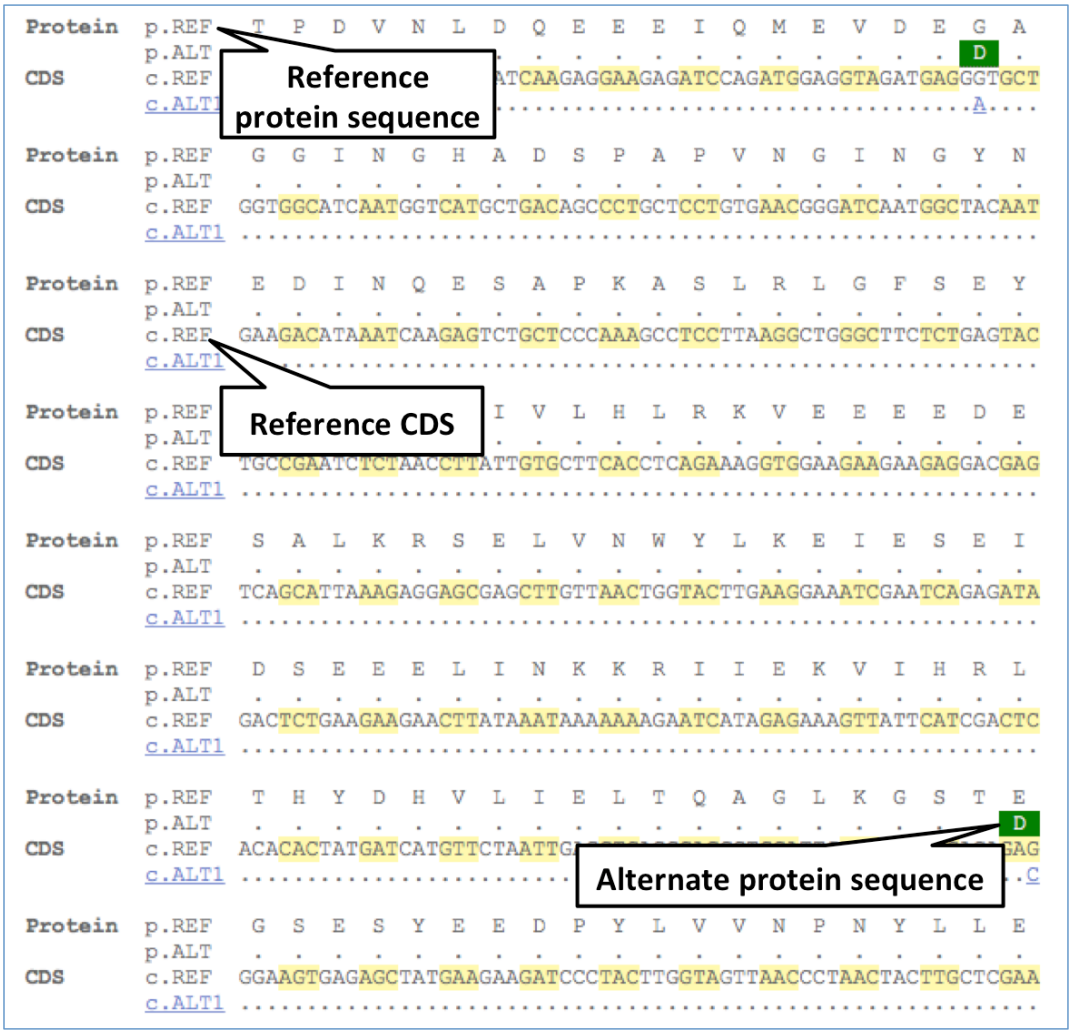
In the Transcript Haplotype view you can view protein consequences, population frequencies, and protein alignments for all the haplotypes for that particular transcript.



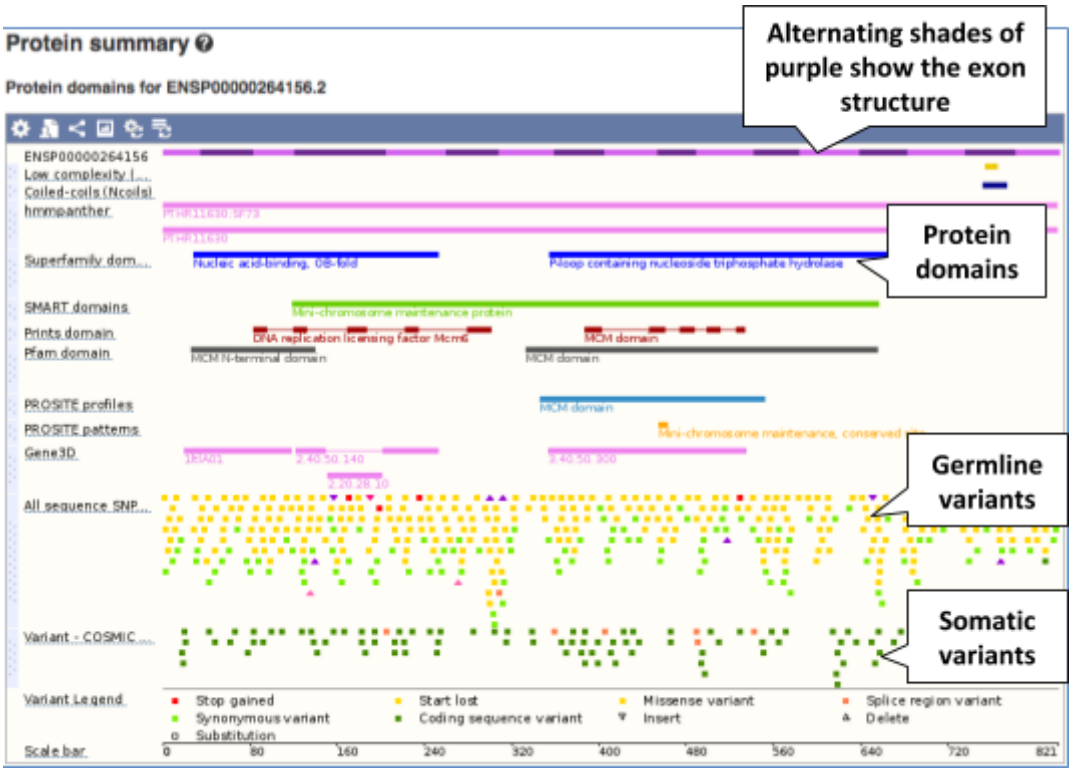
Click on Protein Haplotype 679G>D, 800E>D to find out more information about this transcript haplotype, such as population frequencies and the aligned sequence. Click on Population Frequencies in the quick links.



Click on Aligned Sequence in the quick links.

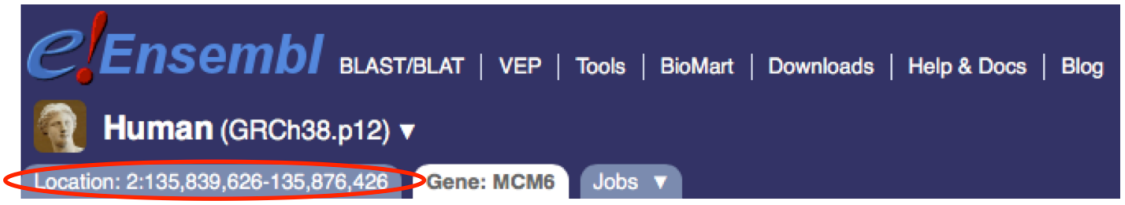


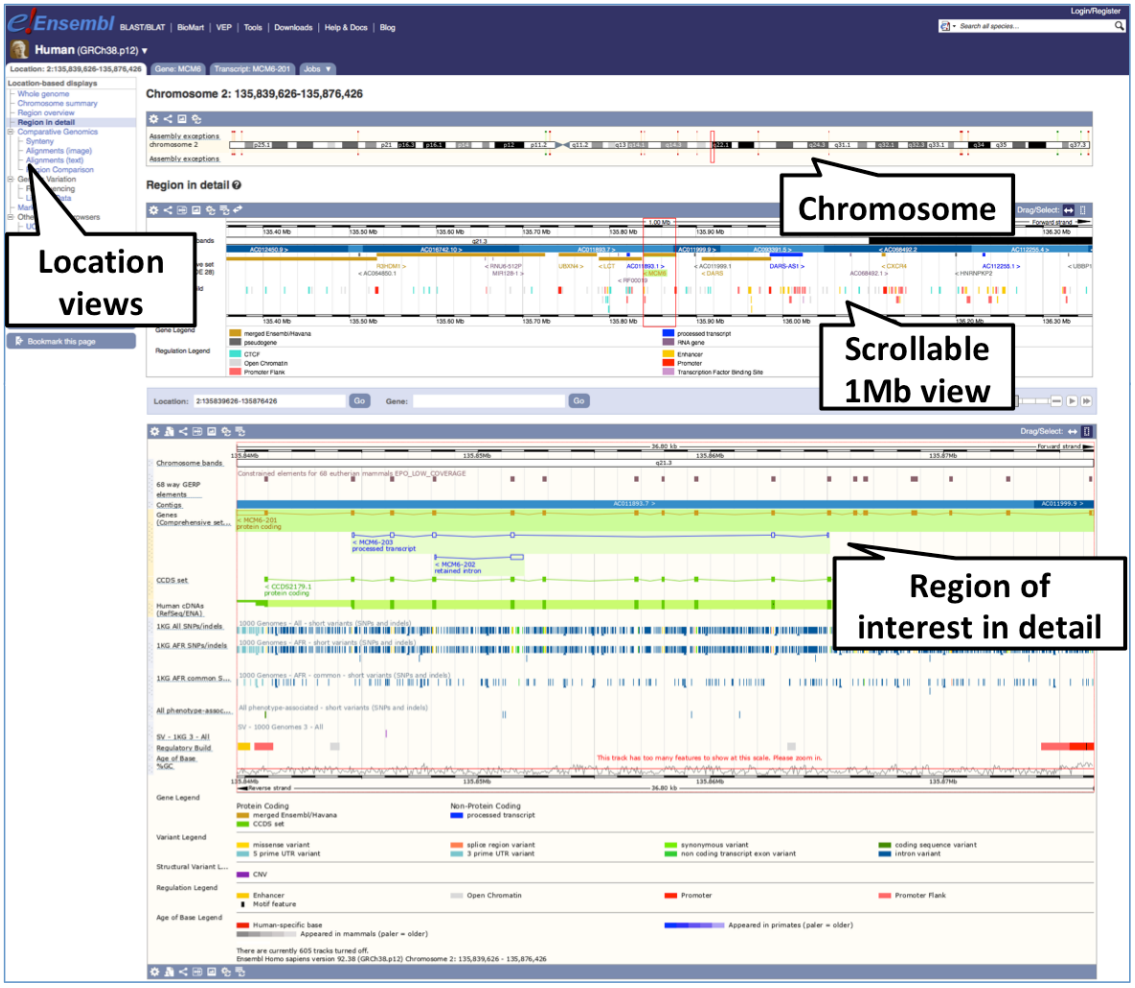
Now click on Protein summary to view variants plotted against domains from Pfam, PROSITE, Superfamily, InterPro, and more.


Protein summary
Protein domains for ENSP00000264156.2
Alternating shades of purple show the exon structure
Protein domains
Germline variants
Somatic variants
ENSP00000264156
Low complexity (...
Coiled-coils (Ncoils)
hmmpanther
Superfamily dom...
Nucleic acid-binding, OB-fold
P-loop containing nucleoside triphosphate hydrolase
SMART domains
Mini-chromosome maintenance protein
Prints domain
DNA replication licensing factor Mcm6
MCM domain
Pfam domain
MCM N-terminal domain
MCM domain
PROSITE profiles
MCM domain
PROSITE patterns
Gene3D
All sequence SNP...
Variant - COSMIC ...
Variant Legend
Stop gained
Start lost
Missense variant
Splice region variant
Synonymous variant
Coding sequence variant
Insert
Delete
Substitution
Scale bar
0
80
160
240
320
400
480
560
640
720
821

Demo part 2: Finding variants in genomic locations

Let's have a look at variants in the Location tab. Click on the Location tab in the top bar.

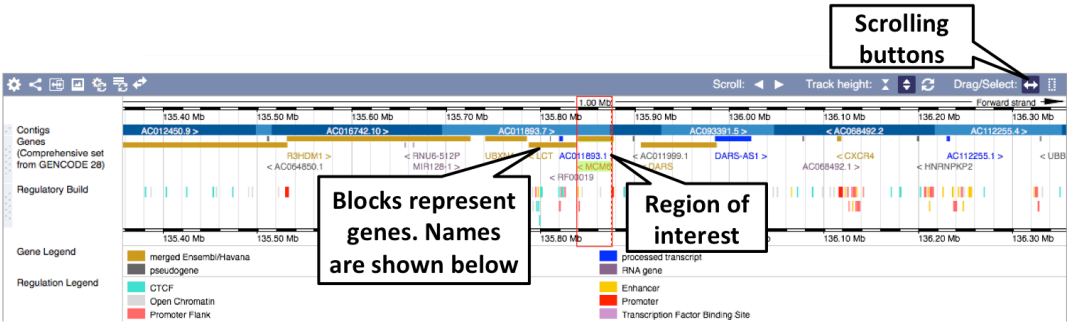




The Region in detail page is made up of three images, let's look at each one on detail.

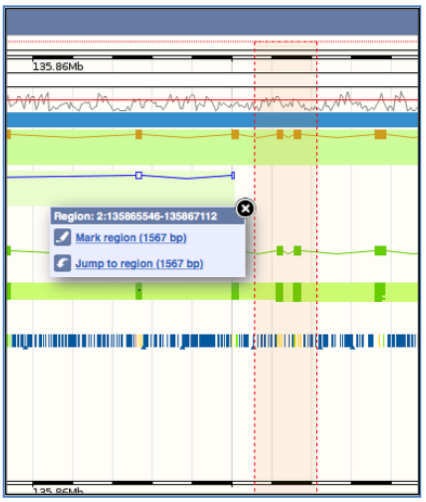
The first image shows the chromosome:

The second image shows a 1Mb region around our selected region. This view allows you to scroll back and forth along the chromosome.

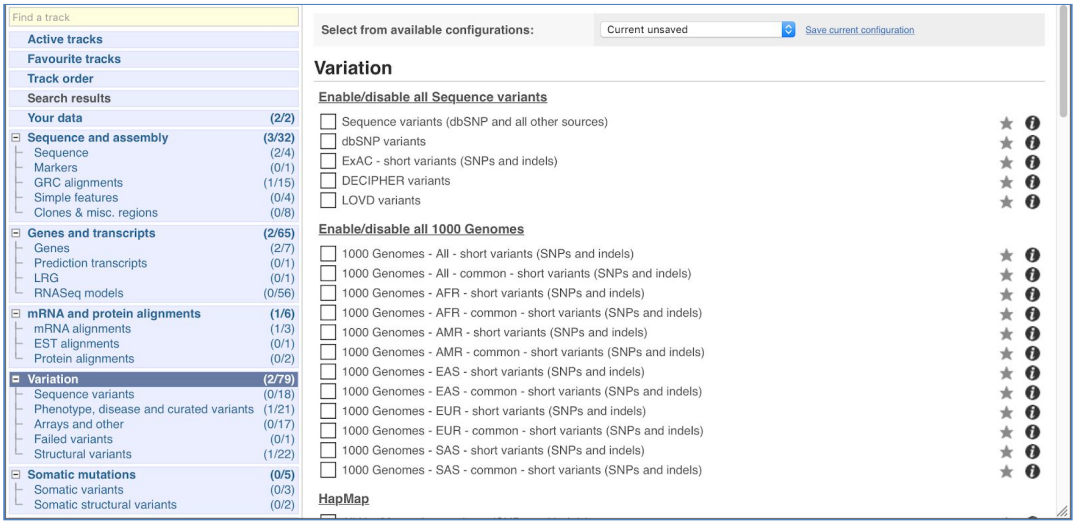


The third image is a detailed, configurable view of the region.

In any of the three region views on the Region in Detail page, you can jump to other regions by clicking and dragging out a box on the region of interest. A pop-up window will allow to either mark this region (highlight) or to jump to this region.



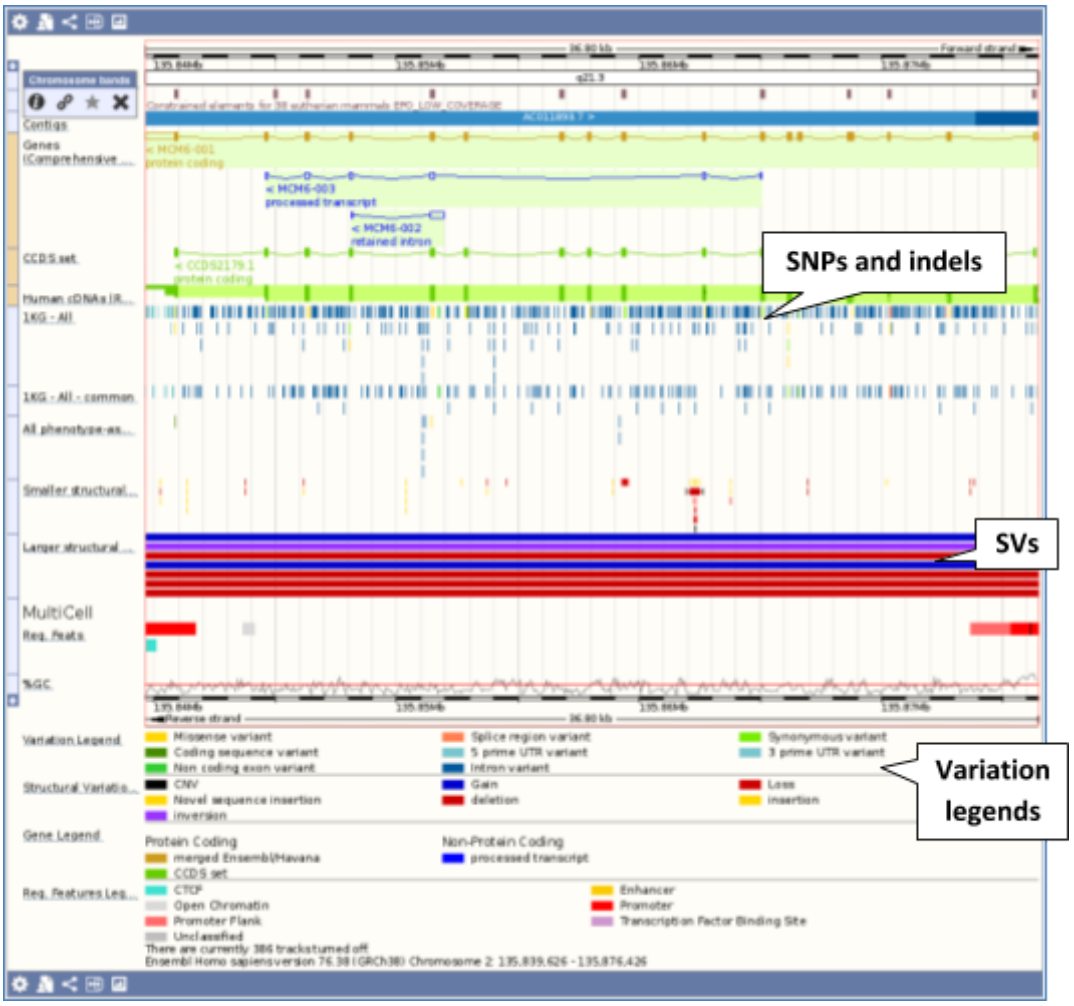
We can edit what we see on this page by clicking on the blue Configure this page menu at the left. Open Variation from the menu.



There are various options for turning on variants. You can turn on variants by source, by frequency, presence of a phenotype or by individual genome they were isolated from. Turn on the following sequence variants in Expanded with name.

- 1000 genomes – All
- 1000 genomes – All – common
- All phenotype-associated variants

Also turn on Larger and Smaller Structural variants (all sources) in Expanded.

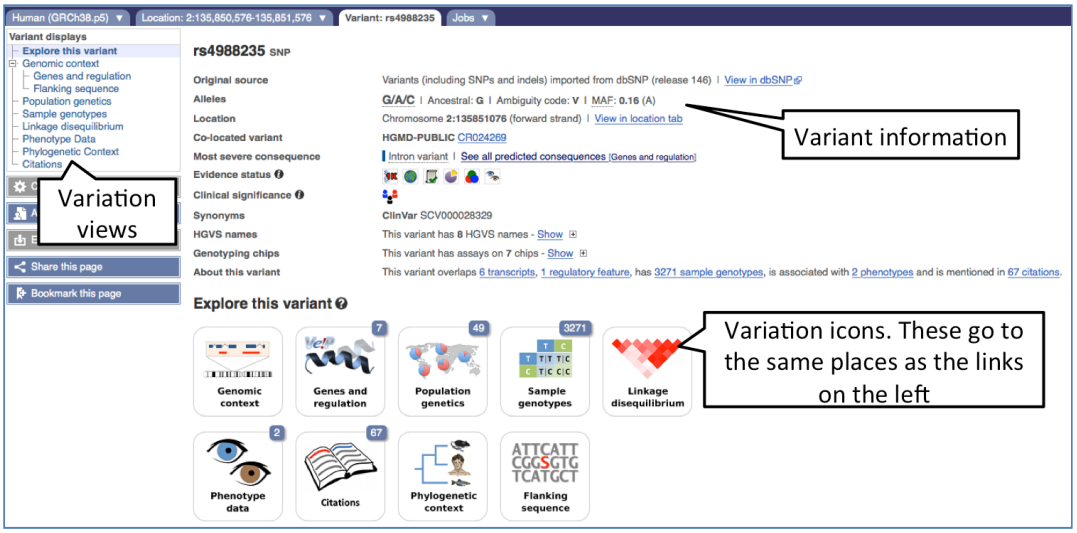


Click on a variant to find out more information. It may be easier to see the individual variants if you zoom in.

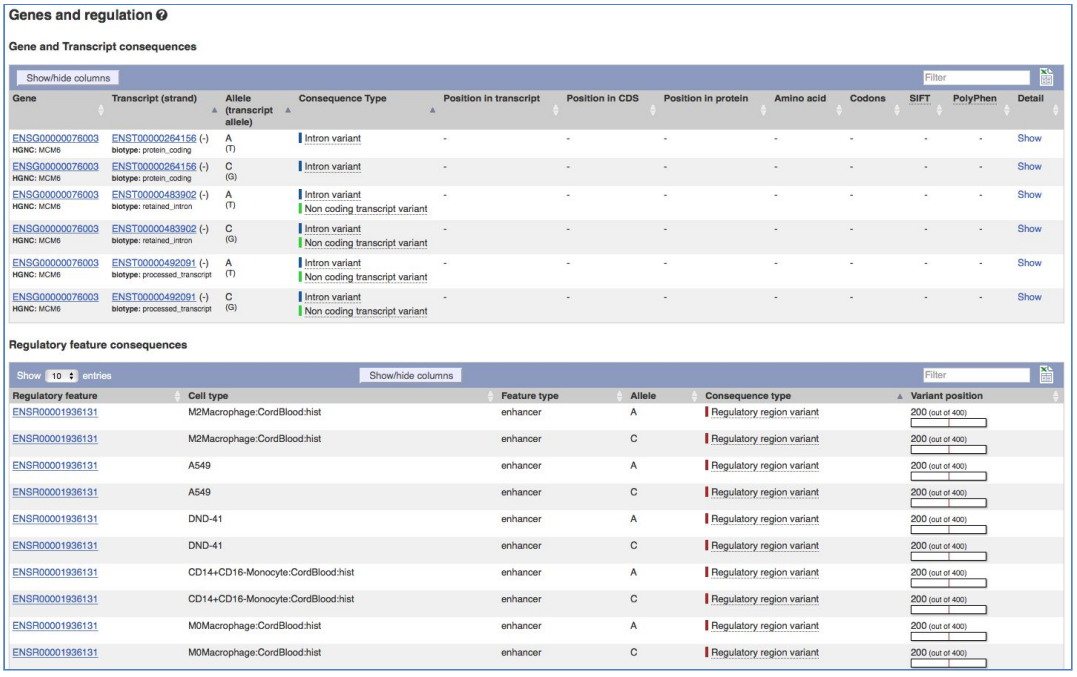
Click on a variant to find out more information. It may be easier to see the individual variants if you zoom in.

Demo part 3: Finding out more about variants

Let's have a look at a specific variant. If we zoomed in we could see the variant rs4988235 in this region, but it's easier to find if we put rs4988235 into the search box. Click through to open the Variation tab.

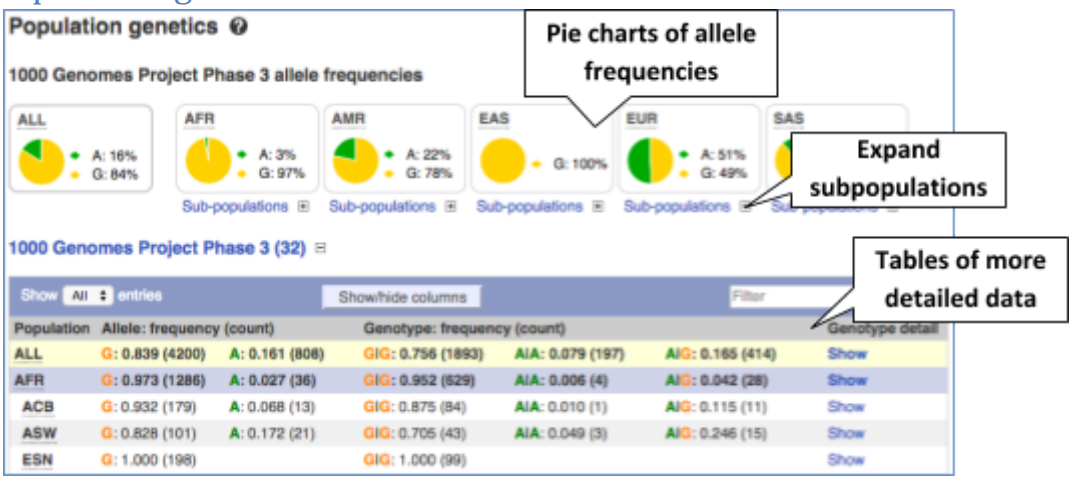


The icons show you what information is available for this variant. Click on Genes and regulation, or follow the link at left.



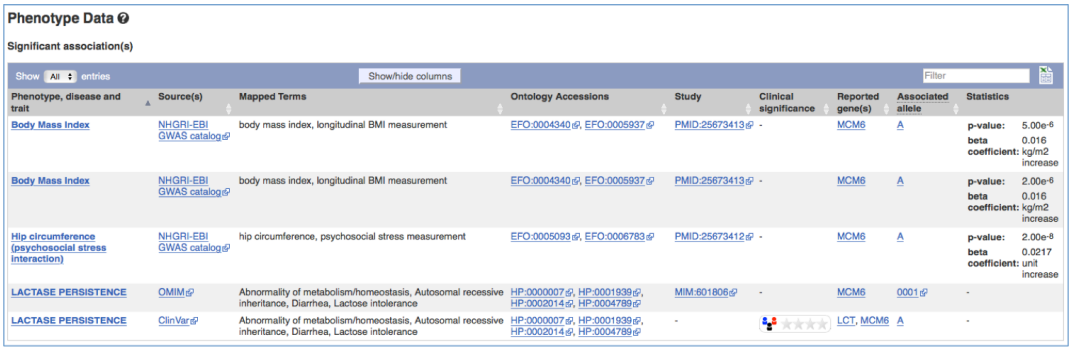
This variant is found in six transcripts of the *MCM6* gene. It has not been associated with any regulatory features or motifs.

Let's look at population genetics. Either click on Explore this variant in the left-hand menu and then on the Population genetics icon, or click on Population genetics in the left-hand menu.



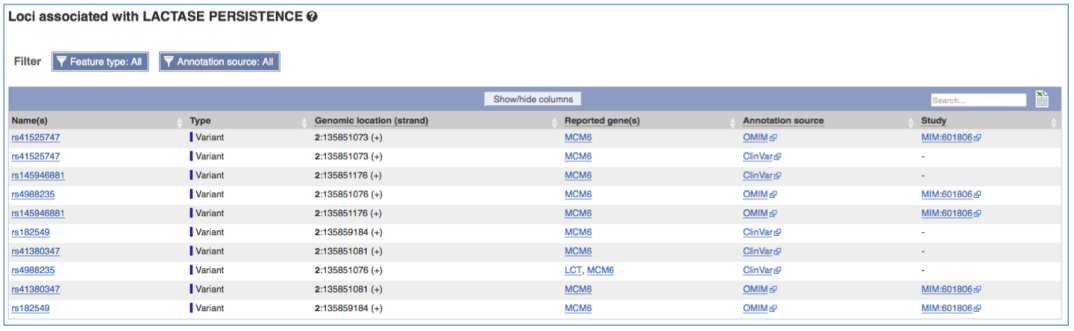
These data are mostly from the **1000 Genomes** and **HapMap** projects in human.

There are big differences in allele frequencies between populations. Let's have a look at the phenotypes associated with this variant to see if they are known to be specific to certain human populations. Either click on Explore this variant in the left-hand menu and then on the Phenotype Data icon, or click on Phenotype Data in the left-hand menu.



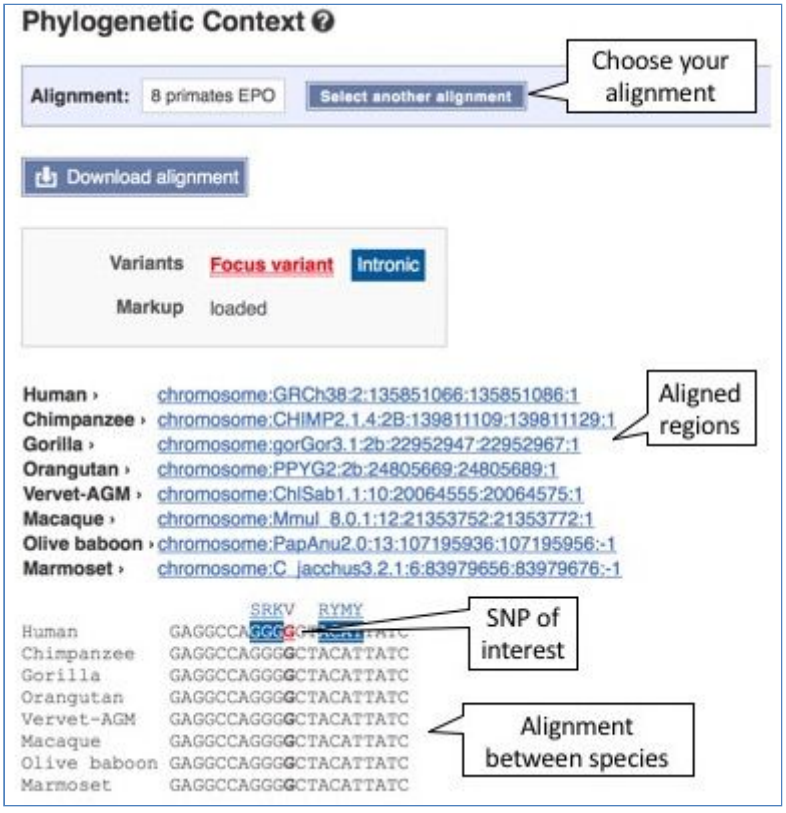
This variant is associated with lactase persistence, which is known to be common in European populations and rare in Asian populations, exactly as we saw in the allele frequencies in these populations. You can also check the population frequencies by clicking on the associated allele.

Are there other loci in the genome associated with lactase persistence? Click on LACTASE PERSISTENCE to find out.



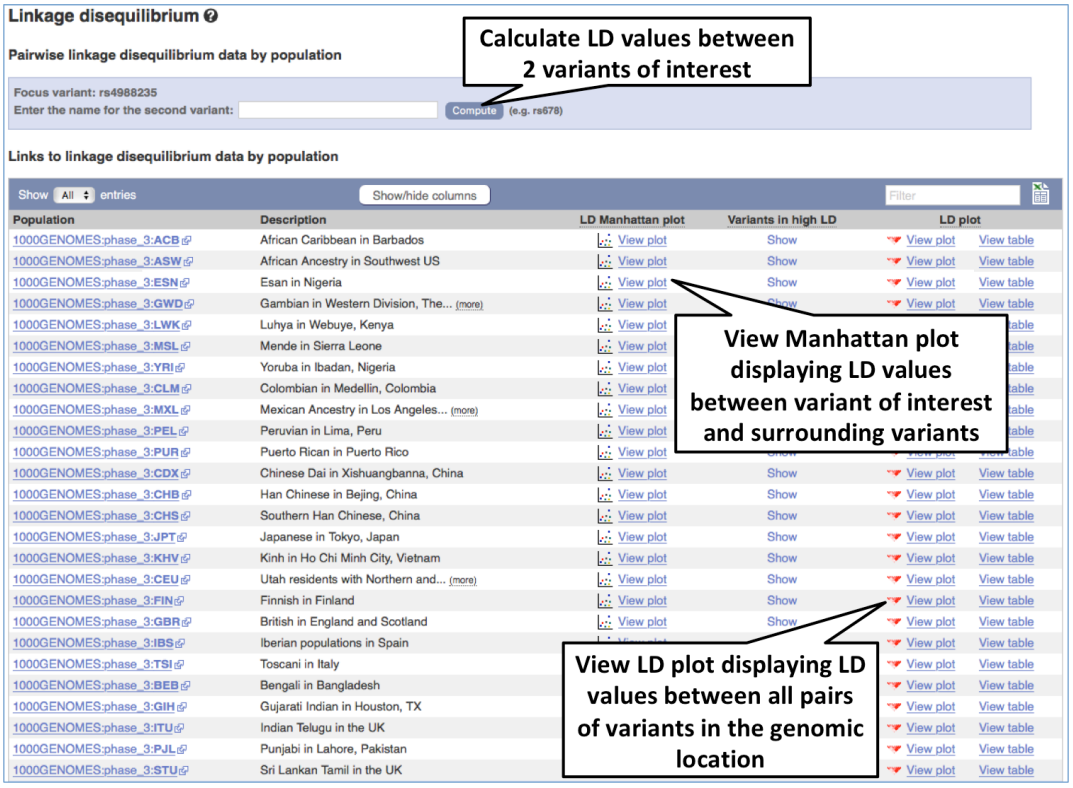
Ten variants are known to be associated with this phenotype. They are all found with the *MCM6* gene.

Click back to the Variation Tab. Click on Phylogenetic Context to see the variant in other species.

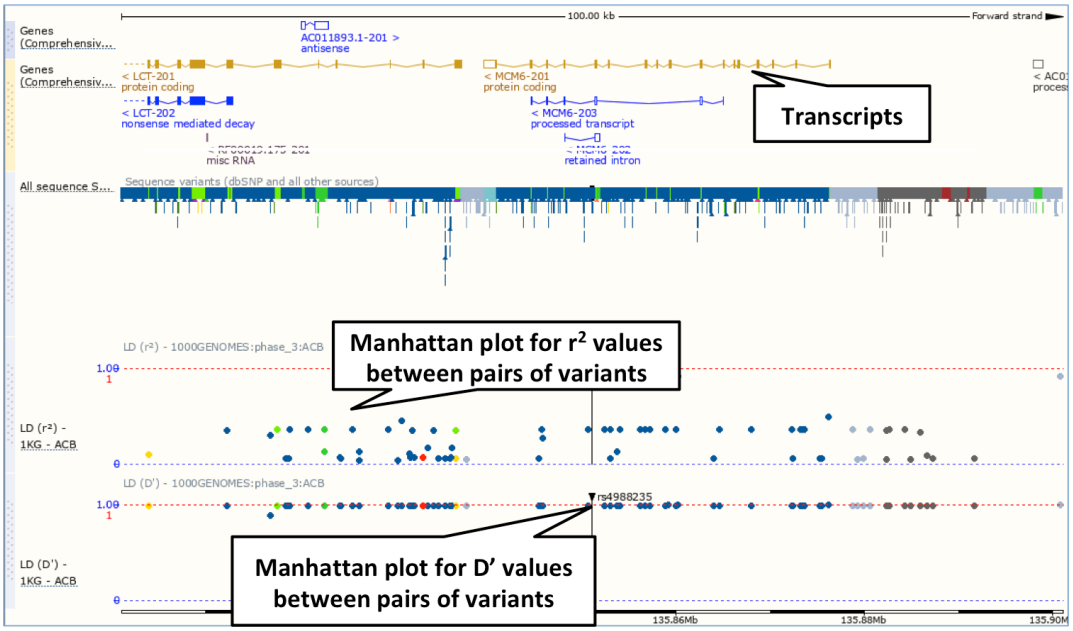


The variant is not marked in the other species. This means that the variant arose in humans.

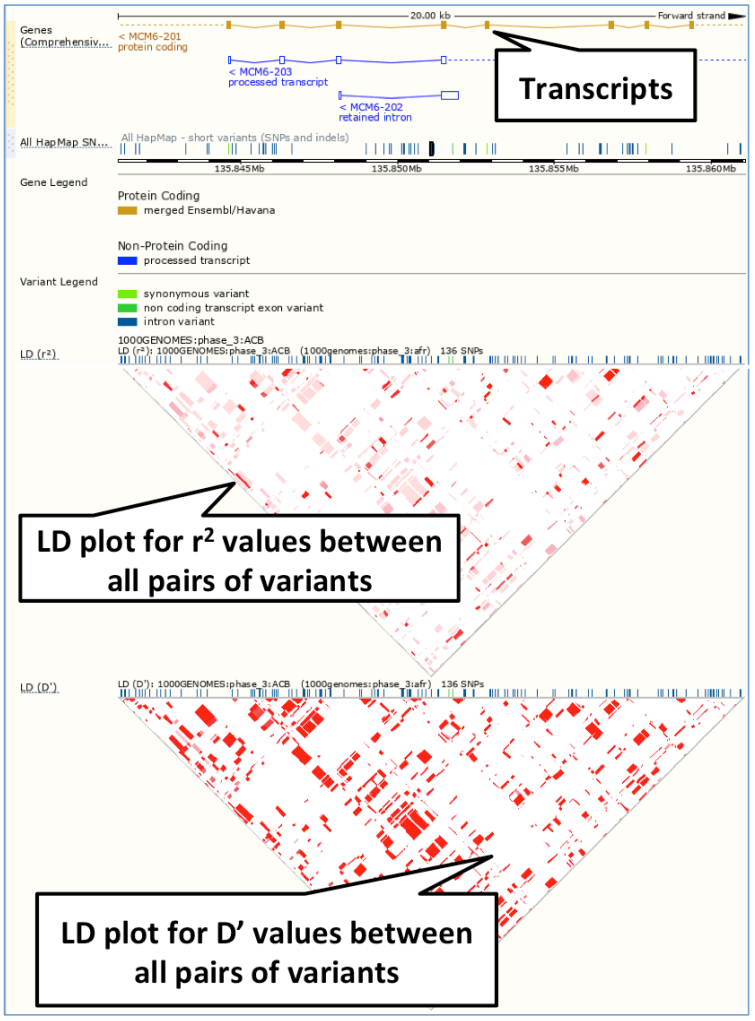
Click on Linkage Disequilibrium to see linkage disequilibrium (LD) data relating to this variant calculated from the 1000 Genomes Project population frequencies.



Click on View Plot corresponding to the African Caribbean in Barbados population in the 'LD Manhattan Plot' column. You can now see two manhattan plots displaying LD values (r^2 and D') calculated for the variant of interest and each individual variant in the surrounding genomic region.



Navigate back to the LD summary page by clicking on Linkage Disequilibrium in the menu on the left hand side. Click on View Plot corresponding to the African Caribbean in Barbados population in the 'LD Plot' column. You can now see two LD plots displaying LD values (r^2 and D') calculated for the all pairs of variants in the defined genomic region.



Exercises: Exploring variants in Ensembl

Exercise V1 – Human population genetics and phenotype data

The SNP rs1738074 in the 5' UTR of the human *TAGAP* gene has been identified as a genetic risk factor for a few diseases.

(a) In which transcripts of *TAGAP* is this SNP found?

(b) What is the least frequent genotype for this SNP in the Yoruba (YRI) population from the HapMap set?

(c) With which diseases is this SNP associated? Are there any known risk (or associated) alleles?

Exercise V2 – Exploring a SNP in human

The missense variation rs1801133 in the human *MTHFR* gene has been linked to elevated levels of homocysteine, an amino acid whose plasma concentration seems to be associated with the risk of cardiovascular diseases, neural tube defects, and loss of cognitive function. This SNP is also referred to as 'A222V', 'Ala222Val', and other HGVS names.

(a) Find the page with information for rs1801133.

(b) Is rs1801133 a missense variant in all transcripts of the *MTHFR* gene?

(c) Why are the alleles for this variant in Ensembl given as G/A and not as C/T, as in dbSNP and literature?
(http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1801133)

(d) What is the major allele in rs1801133?

(e) In which paper(s) is the association between rs1801133 and homocysteine levels described?

(f) According to the data imported from dbSNP, the ancestral allele for rs1801133 is G. Ancestral alleles in dbSNP are based on a comparison between human and chimp. Does the sequence at this same position in other primates confirm that the ancestral allele is G?

Exercise V3 – Exploring a SNP in mouse

Madsen *et al.*, in the paper 'Altered metabolic signature in pre-diabetic NOD mice' (PloS One. 2012; 7(4): e35445), have described several regulatory and coding SNPs, some of them in genes residing within the previously defined *insulin dependent diabetes (IDD)* regions. The authors indicate that one of the identified SNPs, in the murine *Xdh* gene (rs29522348), would lead to an amino acid substitution and could be damaging, as predicted as by SIFT (http://sift.jcvi.org/).

(a) Where is the SNP located (chromosome and coordinates)?

(b) What is the HGVS recommendation nomenclature for this SNP?

(c) Why does Ensembl put the C allele first (C/T)?

(d) Are there differences between the genotypes reported in the ARK/J and C57BL/6NJ sample populations from the WTSI Mouse Genomes project?

Exercise V4 – Variation structure viewer

You are interested in the missense variant rs998717588.

(a) Navigate to the Variant tab and click on 3D Protein model

(b) Identify the location of the variant rs998717588. Is it located in a beta strand, helix or linking region?

(c) Is it in a pfam domain?

Quick Guide to Databases and Projects

Here is a list of databases and projects you will come across in these exercises. Google any of these to learn more. Projects include many species, unless otherwise noted.

Other help:
The Ensembl Glossary: http://www.ensembl.org/Help/Glossary
Ensembl FAQs: http://www.ensembl.org/Help/Faq

SEQUENCES
EMBL-Bank, NCBI GenBank, DDBJ – Contain nucleic acid sequences deposited by submitters such as wet-lab biologists and gene sequencing projects. These three databases are synchronised with each other every day, so the same sequences should be found in each.

CCDS – coding sequences that are agreed upon by Ensembl, VEGA-Havana, UCSC, and NCBI. *(Human and mouse)*

NCBI Entrez Gene – NCBI's gene collection.

NCBI RefSeq – NCBI's collection of "reference sequences", includes genomic DNA, transcripts, and proteins. NM stands for "Known mRNA" (e.g. NM_005476) and NP (e.g. NP_005467) are "Known proteins".

UniProtKB – the "Protein knowledgebase", a comprehensive set of protein sequences. Divided into two parts: Swiss-Prot and TrEMBL.

UniProt Swiss-Prot – the manually annotated, reviewed protein sequences in the UniProtKB. High quality.

UniProt TrEMBL – the automatically annotated, unreviewed set of proteins (EMBL-Bank translated). Varying quality.

VEGA – Vertebrate Genome Annotation, a selection of manually-curated genes, transcripts, and proteins. *(Human, mouse, zebrafish, gorilla, wallaby, pig, and dog)*

VEGA-HAVANA – The main contributor to the VEGA project, located at the Wellcome Trust Sanger Institute, Hinxton, UK.

GENE NAMES

HGNC – HUGO Gene Nomenclature Committee, a project assigning a unique and meaningful name and symbol to every human gene. *(Human)*

ZFIN – The Zebrafish Model Organism Database. Gene names are only one part of this project. *(Z-fish)*

PROTEIN SIGNATURES

InterPro – A collection of domains, motifs, and other protein signatures. Protein signature records are extensive, and combine information from individual projects such as UniProt, along with other databases such as SMART, PFAM, and PROSITE (explained below).

PFAM – A collection of protein families.

PROSITE – A collection of protein domains, families, and functional sites.

SMART – A collection of evolutionarily conserved protein domains.

OTHER PROJECTS

NCBI dbSNP – A collection of sequence polymorphisms, mainly single nucleotide polymorphisms, along with insertion-deletions.

NCBI OMIM – Online Mendelian Inheritance in Man – a resource showing phenotypes and diseases related to genes. *(Human)*