

Browsing Variation Data with Ensembl



www.ensembl.org
www.ensemblgenomes.org

Coursebook v95

[http://training.ensembl.org/events/2019/
2019-04-02-VEP_Breda](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 screenshot shows the Ensembl homepage with several key features highlighted:

- Link back to homepage**: A callout pointing to the Ensembl logo in the top left.
- Ensembl tools**: A callout pointing to the "Tools" link in the top navigation bar.
- Blue bar remains visible at the top of every page**: A callout pointing to the blue header bar.
- Search**: A callout pointing to the search bar in the top right.
- Drop down list to species**: A callout pointing to the "Select a species" dropdown menu.
- How-tos for commonly used Ensembl features**: A callout pointing to the "How-to" section in the center of the page.
- See the current release number and what's new**: A callout pointing to the "What's New in Ensembl Release 89 (May 2017)" section.

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

The screenshot shows the GRCh37 dedicated website, which is a subset of the Ensembl genome browser:

- Search**: A search bar at the top.
- Popular genomes**: A section showing "Human (GRCh37)" and a link to "Log in to customize this list".
- All genomes**: A dropdown menu to select a species.
- Browse a Genome**: Information about the Ensembl project and genome databases.
- ENCODE data in Ensembl**: A section showing ENCODE data.
- Variant Effect Predictor**: A section showing VeP data.
- Gene expression in different tissues**: A section showing gene expression data.
- Find SNPs and other variants for my gene**: A section showing SNP data.
- Retrieve gene sequence**: A section showing gene sequence data.
- Compare genes across species**: A section showing gene comparison data.
- About this archive**: Information about the archive being based on Ensembl Release 75 data.
- Latest blog posts**: A list of recent blog posts.

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

The screenshot shows the Ensembl Genomes homepage with several taxonomic sections:

- Bacteria**: No significant updates.
- Protists**: No significant updates.
- Fungi**: No significant updates.
- Plants**: No significant updates.
- Vertebrates**: No significant updates.
- Ensembl Bacteria**: No significant updates.
- Ensembl Fungi**: No significant updates.
- Ensembl Plants**: No significant updates.
- Ensembl Protists**: No significant updates.
- Links to taxon-specific pages**: A section listing links to specific taxonomic pages.
- Release notes from previous releases**: A section for release notes, currently showing "Release 34 (December 2016)".
- Have a question?**: A yellow callout box with the text: "Ensembl Genomes has many questions available for you to ask. If there is a FAQ missing, contact us."
- Link to Ensembl**: A yellow callout box with the text: "Link to Ensembl" and "before! If there is a FAQ missing, contact us."
- User Log In**: A link in the bottom right corner.

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

The image shows three side-by-side screenshots of Ensembl taxonomic homepages:

- Ensembl Protists**: Shows a list of popular protist species including Plasmodium falciparum, Dictyostelium discoideum, Phytophthora infestans, and Leishmania major.
- Ensembl Fungi**: Shows a list of popular fungal species including Aspergillus nidulans, Cladophora ramosa, and Pyrenopeltis velutina.
- Ensembl Metazoa**: Shows a list of popular metazoan species including Cenorhabditis elegans, Anopheles gambiae, Drosophila melanogaster, and Apis mellifera.

Protists

The screenshot shows the Ensembl Protist homepage with the following features:

- Search bar**: Allows searching by species name (e.g., *PF00120w or ophr*).
- Popular genomes**: A list of protist species including *Plasmodium falciparum*, *Dictyostelium discoideum*, *Phytophthora infestans*, and *Leishmania major*.
- What's new in Release 17 (January 2013)**: A list of updates:
 - Added new species: *Glycina lebbek*
 - Protein features updated for all the protist species.
 - More internal links for the comparative genomics database.
 - Updated BioMart.
- Did you know...?**: A box with information about protein feature updates.
- Future Releases**: Notes for Release 18 of Ensembl Genomes.
- Footer**: EMBL-EBI logo and "Powered by Ensembl" text.

Fungi

The screenshot shows the Ensembl Fungi homepage with the following features:

- Search bar**: Allows searching by species name (e.g., *Citrobary* or chz28*).
- Popular genomes**: A list of fungal species including *Aspergillus nidulans*, *Zea mays*, *Glycine max*, *Oryza sativa*, and *Brachypodium distachyon*.
- What's new in Release 17 (January 2013)**: A list of updates:
 - New functionality: Search against reference genome locations.
 - Sequencing search service added to the *Tolypocladium longisegmentum* species page.
- Did you know...?**: A box with information about sequencing search service.
- Reference** section:
 - The International Barley Genome Sequencing Consortium (BGS). A physical, gapless reference sequence assembly of the barley genome. *Nature* 2012.
 - Berney R. et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 2012.
- Ensembl Plants** logo.

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:

Gene tab

Human (GRCh38.p6) ▾ Location: 2:135,839,626-135,876,426 Gene: MCM6 Jobs ▾

Gene-based displays

- Summary
- Splice variants
- Transcript comparison
- Supporting evidence
- Gene alleles
- Sequence
 - Secondary Structure
 - External references
 - Regulation
 - Ontology
 - GO: Biological process
 - GO: Molecular function
 - GO: Cellular component
 - Comparative Genomics
 - Genomic alignments
 - Gene tree
 - Gene orthologs tree
 - Orthologs
 - Paralogs
 - Ensembl protein families

Gene views

- ID History
- Gene history
- Configure this page
- Add your data
- Export data
- Share this page
- Bookmark this page

Gene: MCM6 ENSG00000079003

Description: minichromosome maintenance complex component 6 [Source:HGNC Symbol;Acc:HGNC_02492] MCG40208, Mc6, P105MCM

Synonyms: Chromosome 2-135,839,626-135,876,426 reverse strand. GNC108 CM000064.2

Location: This gene has 3 transcripts (splice variants), 69 orthologues, is a member of 1 Ensembl protein family and is associated with 1 phenotype.

About this gene

Transcripts

Show transcript

Option: open table of transcripts

Name: MCM6 (HGNC Symbol)

CCDS

UniProtKB

RefSeq

Ensembl version: ENSG00000079003.4

Other assemblies

This gene has proteins that correspond to the following Uniprot identifiers: Q14566 (P)

Overlapping RefSeq Gene ID: 41759 matches and has similar biotype of protein_coding

This gene maps to 138,597,195-138,631,998 in GRCh38 coordinates. View this locus in the GRCh38 archive: ENSG00000079003.4

Gene type: Known protein coding

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

Alternative genes: This gene corresponds to the following database identifiers: Havana gene: OTTHUMG00000131738

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

MCM6 transcripts

Forward strand: 135,844kb -> 135,854kb -> 135,864kb -> 135,874kb -> 135,884kb

Reverse strand: 135,854kb -> 135,844kb -> 135,854kb -> 135,864kb -> 135,874kb

Gene Legend:

- Protein Coding (Yellow)
- Non-Protein Coding (Blue)
- Merged Ensembl Transcripts (Yellow)
- Processed Transcript (Blue)

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.

Most recent genome assembly; GRCh38 = hg38

Human (GRCh38.p12) ▾

Location: 2:135,839,626-135,876,426 Gene: MCM6

Gene-based displays

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence**
 - Secondary Structure

Tabs Gene: MCM6

Description

Gene Synonyms

Click Sequence

Marked-up sequence ?

[Download sequence](#) [BLAST this sequence](#) [Blast or download this sequence](#)

Exons MCM6 exons All exons in this region

Markup loaded

```
>chromosome:GRCh38:2:135839026:135877026:-1
GAACTCCTGACCTCAGGTGATCCACACGCCCTGGCTCCAAAGTGCTAGGATTACAGGT
TGTGAGCCACCGCGCCGGCCAATGTCAAATATTCGCTGATTTCTCGCT
CCACAATAATTGTGCCACCTTACCCATGCAAGGTTAGACATCTCTTGAC
GGGTTTTACTGTGGTTCCATTATTCCTCTAAGACAGCAGAAGCG
TTTTCTCTCCATAGAGTCTTACTCCGATCGTCAATTAAATCATGGAATGATTTAA
AGAACATTGAAAAACCACTCGACAAAACCTCAGATTGAACTTGTCCAGCAGCCAA
AAATACCCGCGGCCACGGCTACACTCGCAGGCAGAGCAGATGGCTTCTCCAGA
AGGGCTTGATTTGGCGCGAAATCCCTTCCGTGGGCTGGGCTCTGGAGAGGCAGCG
TTCATTGGTCAGGTTGGCGCGAAATCTCCAGCTCTGTGTCAGGATTGGTCCGGCGT
GCAGGTCGAAGAGGGGGCGGGCGGAAGCGCCGGCGCCAAAGCTGCAGCGTCT
GGAAAAAAAGCGACTTGTGGCGGTGAGCGTGGCCAGCGAATCTCGGACTAAGCAA
TATGGACCTCGCGGGCGCAGCGGAGCCGGCGCAGCCAGCACCTGGAC
CGAGGTGGCCGAGAAGTGCCAGAAACTGTTCTGGACTCTTGGAGGA GTAAAG
GCAGGTCGAAGAGGGGGCGGGCGGAAGCGCCGGCGCCAAAGCTGCAGCGTCT
GGCGCGGGGTGTTCCGGAACCTGGGGTCCGCGTCCGGGAAGCGCCTCCCCGCC
```

Upstream sequence

exon

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

name of the genome assembly

chromosome

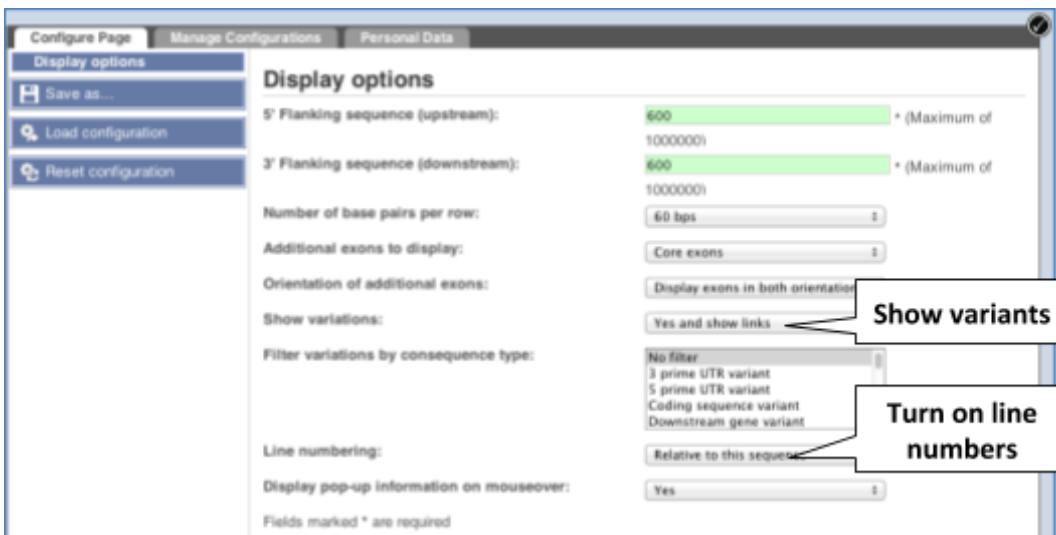
base pair start

base pair end

reverse strand (1 is forward)

```
>chromosome:GRCh38:2:135839026:135877026:-1
GAACTCCTGACCTCAGGTGATCCACACGCCCTGGCTCCAAAGTGCTAGGATTACAGGT
TGTGAGCCACCGCGCCGGCCAATGTCAAATATTCGCTGATTTCTCGCTAGCACCA
CCACAATAATTGTGCCACCTTACCCATGCAAGGTTAGACATCTCTTGACTTTACA
```

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.

Marked-up sequence ?

Download sequence BLAST this sequence

Legend of variant consequence types

Exons	MCM6 exons	All exons in this region					
Variants	3 prime UTR	5 prime UTR	Coding sequence	Flagged variant	Frameshift	Inframe deletion	Intronic
	Missense	Non-coding exon	Protein altering variant	Splice acceptor	Splice donor	Splice region	Start lost
	Stop gained	Stop retained	Synonymous	Upstream			

Markup loaded

```
>chromosome:GRCh38:2:135839026:135876626:-1
 1 [CAGATGCGCTTC] GAA[GCGTTTG] ATTGGCGCAAATCCCTTCYK TGGGC[K] 60
 61 [GGCT] TTGGA[SAGGCCG] CATTGGTCAG[TT] KGGCGAAATCCCA[SK] CTGTG 120
 121 TCACCATGGTCCGGCG[CG] CAGGTCGGA[GAGGGGGCGGGCGG] AGCGCC[GCGG]R 180
 181 CGCGCRA[AGCTGCA] GT[CY] GAAAAAAG[YC] YC[GTG] PR[GTM] AS[CGTC] CGYM[G]G 240
 241 ARTCTTYS[CACTAG] AAT[RY] GGAC[TC] CGCR[RD] CAGS[GGAGC] M[GCR] YRV[CA]G 300
 301 CAGMAYCTGGARRTC[CG] YCAGGTGGCCGAGAAGTCCAGAA[CTGTTM] YK[GAY]TTC
 361 TTGGWWRAGT[RGGTGG] GACYGCCCGGGCKCCYCCGGCTCGSAGGCCCTCCGGY[GT] 600
 421 RGGY[RGCCGGTR] MGKGTG[S] GTGTT[GGAACCTRGRGRTGCGG] SGTCCGG
 481 AAGCMTCCYGGY[BGCCC] CAACTTAGCTC[BGACCG] YRGORGCCR[Y] GAGGAR[C]T
 541 QRKWAGGCTGS[TTTCTGG] RAAACCCATCTCAGCCCTAACG[GCT] YGTGCCCA
 601 GTGCTYAGCGCTT[GCTCTGTYACT] ATGCTACCACCS[CGCGAG] M[T] GTATTTC[A]
 661 CMGCCCTCGGTGACCGGCGGGAA[CTGGGCTC] CTGAGTATA[ACTGGGCCGCTGG 720
```

Variants on sequence shown as IUPAC ambiguity codes

Find out more about a variant by clicking on it.

Variation: rs1057031

Position 2:135876392

Alleles G/A

Consequences 5 prime UTR variant
Regulatory region variant

[Explore this variation](#)
[Gene/Transcript Locations](#)
[Population Allele Frequencies](#)

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.

The screenshot shows a table of genetic variants. A tooltip is displayed over a variant ID (rs3688216) in the first row. The tooltip contains the text: "Evidence for variant: hover over the icons for definitions". The table has columns for Variant ID, Chr:bp, Alleles, Global MAF, Class, Source, Evidence, Clin. Sig., Conseq. Type, AA, AA coord, SIFT, PolyPhen, and HGVS. A legend on the right side of the table defines various consequence types like Downstream gene variant, Intron variant, etc. A callout box labeled "SIFT and PolyPhen scores" points to the SIFT and PolyPhen columns. Another callout box labeled "Transcript affected" points to the "Affected" column.

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.

The screenshot shows a "Consequences" filter dialog with the following interface:

- Buttons:** Turn All Off, PTV, PTV & Missense, Only Exonic, Turn All On.
- Text:** PTV = Protein Truncating Variant.
- List:** A scrollable list of variant consequences with "On" status indicators:
 - Transcript ablation (0) On
 - Splice acceptor variant (6) On
 - Splice donor variant (0) On
 - Stop gained (12) On
 - Frameshift variant (21) On
 - Stop lost (0) On
 - Start lost (3) On
 - Transcript amplification (0) On
 - Inframe insertion (0) On
 - Inframe deletion (6) On
 - Mis sense variant (909) On
 - protein altering variant (3) On
 - Splice region variant (186) On
 - Incomplete terminal codon variant (0) On
 - Synonymous variant (351) On
 - Stop retained variant (0) On
 - Coding sequence variant (342) On
 - Mature miRNA variant (0) On
 - 5 prime UTR variant (51) On
 - 3 prime UTR variant (225) On
 - Non coding transcript exon variant (243) On
 - Intron variant (7357) On
 - NMD transcript variant (0) On
 - Non coding transcript variant (1870) On
 - Upstream gene variant (1757) On
 - Downstream gene variant (2328) On
- Buttons at bottom:** Apply » and Cancel.

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.

The screenshot shows the Ensembl gene page for ENSG00000076003. The top section displays phenotype annotations for the gene. A callout box highlights the "Phenotypes associated with the gene" section, which lists LACTOSE INTOLERANCE, ADULT TYPE and LACTASE PERSISTENCE. Another callout box highlights the "Phenotypes associated with variants in the gene" section, which lists ALL variants with a phenotype annotation and LACTASE PERSISTENCE. A third callout box highlights the "Phenotypes associated with orthologues of the gene" section, which lists Lactose Intolerance, Adult Type. A fourth callout box with the text "Click to see list of variants" points to the variants listed under the phenotype sections.

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**.

The screenshot shows the Ensembl Transcript tab for MCM6-201. The top navigation bar includes links for BLAST/BLAT, BioMart, VEP, and Tools. A callout box highlights the "Transcript tab" link. The main content area displays the transcript table for MCM6-201, showing three rows of data. A callout box highlights the "Transcript views" section, which includes links for Custom tracks, Export data, Share this page, and Bookmark this page. Below the table is a genomic track visualization showing the MCM6-201 transcript with exons and introns. A callout box highlights the "Exons" link in the left navigation column. The bottom of the page contains transcript statistics and a note about the Ensembl genebuild transcript and Vega manual annotation being identical.

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.

Exons 0

[Download sequence](#)

Exons/ Introns Translated sequence Flanking sequence Intron sequence UTR

Markup loaded

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
1	ENSE00000827135 5' upstream sequence	135,876,426	135,876,259	-	2	168	Red: UTR
2	Intron 1-2 ENSE00000777061	135,876,258	135,872,844	2	2	147	Blue: Coding sequence
3	Intron 2-3 ENSE00000777053	135,870,696	135,870,362	2	2	2,335	Grey: Introns
4	Intron 3-4 ENSE00000776995	135,870,250	135,868,861	2	0	1,390	
		135,868,660	135,868,611			250	

To add variation data, click on [Configure this page](#) and change the display options accordingly.

Configure Page Personal Data

Display options

Save configuration as...

Load configuration

Reset configuration

Display options

Flanking sequence at either end of transcript:

Number of base pairs per row:

Intron base pairs to show at splice sites:

Show full intronic sequence:

Show exons only:

Line numbering:

Show variants:

Hide variants longer than 10bp:

Hide variants by frequency (MAF):

Filter variants by consequence type:

No filter
3 prime UTR variant
5 prime UTR variant
Coding sequence variant
Downstream gene variant

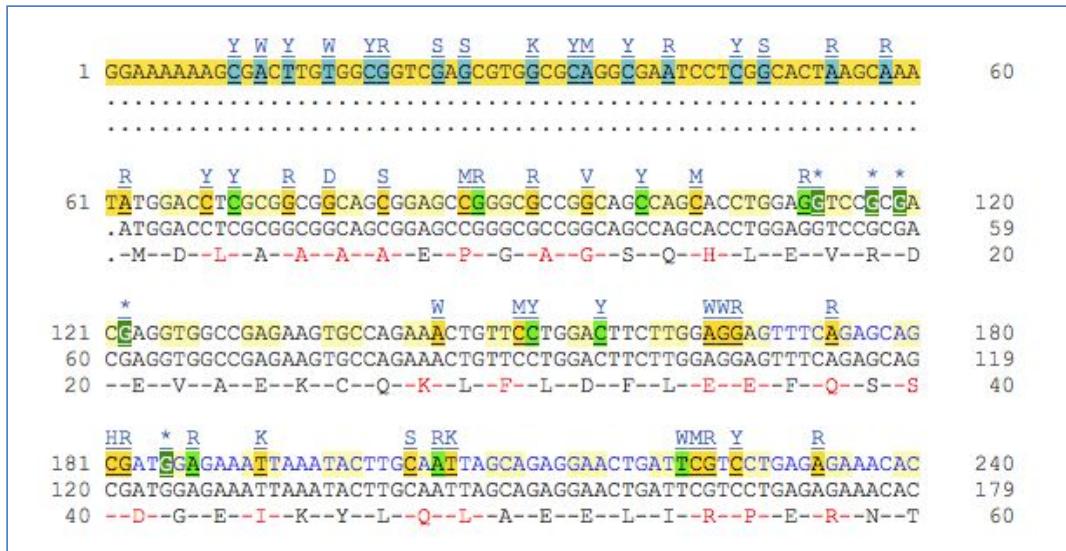
You should now see a marked-up sequence that is annotated with variation data:

Sequence

.....agagggggcggggcggaagcggccggggcgccgcaaaagctgcacgcgtct
 GGAAAAAAAGCAGCTTGTGGCGGTGACCGTGGCGCAGGCAGATCCTGGCACTAAGCAAA
 TATGGACCTCGCGCGGCAGCGGGGCCGGCGCCAGCGACACTGGAGGTCCCGGA
 CGAGGTGGCCGAGAAAGTGCCAGAAACTGTTCTGGACTTCTTGGAGGA
 gtaagtccggcagcgccccgggct.....ctagttgactcgaatgtacctgcag
 GTTTCAGAGCCAGCGATGGAGAAAATTAAATACTTGCACATTAGCAGAGGAACGTGATTCGTC
 TGAGAGAAAACACAATTGGTTGTGAGTTTGTGGAACCTGGAACAAATTAAACCAGCAACTTTC
 CACCAACCAATTCAAGAGGAGTTCTATAG

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.

Haplotypes ?

[Export data as JSON](#)

[View by protein or CDS](#)

[Switch to CDS view](#)

Show All entries Show/hide columns Filter

Protein haplotype	Flags	Frequency (count)	AFR	AMR	EAS	EUR	SAS
REF		0.983 (4923)	0.982 (1298)	0.977 (678)	0.98 (988)	0.982 (988)	0.993 (974)
361N>S		0.0024 (12)		0.00144 (1)	0.000992 (1)		
357T>I	D	0.0016 (8)	0.00605 (8)				
220I>V		0.0012 (6)		0.00144 (1)	0.00496		
611K>R		0.000799 (4)			0.00397 (4)		
40S>R		0.000799 (4)	0.000756 (1)	0.00432 (3)			
506I>V		0.000799 (4)		0.00288 (2)		0.00199 (2)	
8	Variants that are found together in individuals	0.000599 (3)		0.00432 (3)			
7		0.000599 (3)			0.00298 (3)		
2		0.000599 (3)		0.00432 (3)			
136R>G		0.000399 (2)	0.00151 (2)				
304A>V		0.000399 (2)			0.00198 (2)		
679G>D,800E>D		0.000399 (2)	0.00151 (2)				

Frequencies in 1000 Genomes populations

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.

Population frequencies

[Populations](#)

Population group	Population	Frequency (count)
AFR	GWD	0.00442 (1)
	ASW	0.00820 (1)
AMR	No data	
EAS	No data	
EUR	No data	
SAS	No data	

Total count: 2

Population frequencies

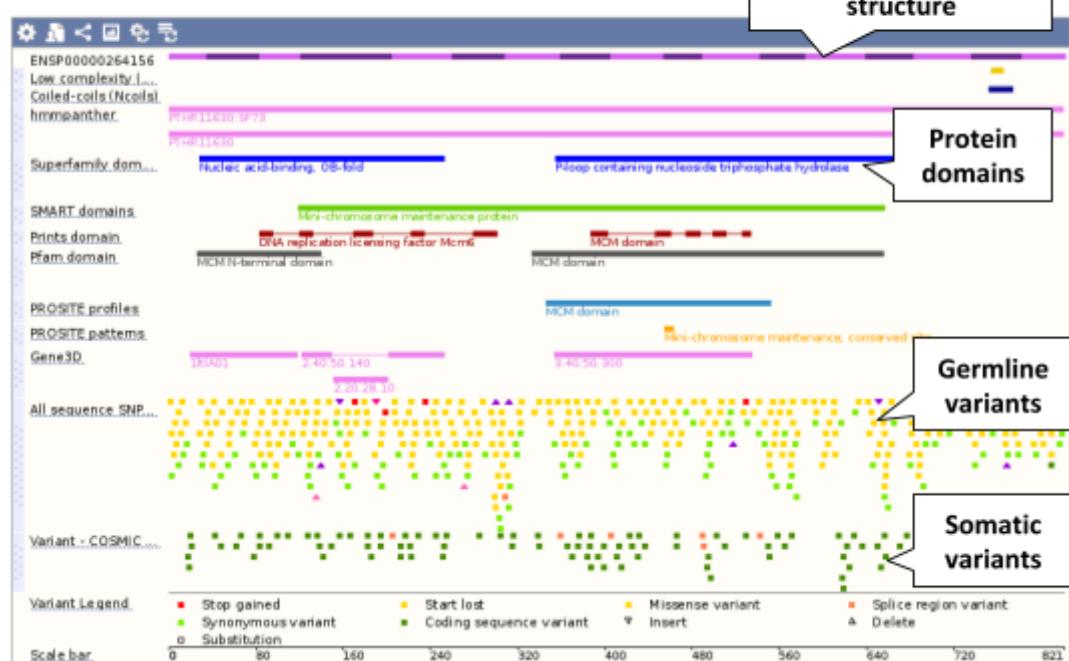
Click on Aligned Sequence in the quick links.

Protein	p.REF	T P D V N L D Q E E E I Q M E V D E G A
CDS	p.ALT	.
	c.REF	ATCAAGAGGAAGAGATCCAGATGGAGGTAGATGAGGGTGCT
	c.ALT1	A.....
Protein	p.REF	G G I N G H A D S P A P V N G I N G Y N
CDS	p.ALT	.
	c.REF	GGTGGCATCAATGGTCATGCTGACAGCCTGCTCCTGTGAACGGGATCAATGGCTACAAT
	c.ALT1
Protein	p.REF	E D I N Q E S A P K A S L R L G F S E Y
CDS	p.ALT	.
	c.REF	GAAGACATAAAATCAAGAGTCTGCTCCCAGGCCCTTAAGGCTGGGCTTCTGAGTAC
	c.ALT1
Protein	p.REF	I V L H L R K V E E E E D E
CDS	p.ALT	.
	c.REF	TGCCGAATCTCAACCTTATTGTGCTTCACCTCAGAAAGGTGGAAGAAGAAGGACGAG
	c.ALT1
Protein	p.REF	S A L K R S E L V N W Y L K E I E S E I
CDS	p.ALT	.
	c.REF	TCAGCATTAAAGAGGAGCGAGCTTAACTGGTACTTGAAGGAATCGAACAGATA
	c.ALT1
Protein	p.REF	D S E E E L I N K K R I I E K V I H R L
CDS	p.ALT	.
	c.REF	GACTCTGAAGAAGAACTTATAAATAAAAAGAATCATAGAGAAAGTTATTCACTCGACTC
	c.ALT1
Protein	p.REF	T H Y D H V L I E L T Q A G L K G S T E
CDS	p.ALT	.
	c.REF	ACACACTATGATCATGTTCTAATTG
	c.ALT1
		D GAG C
Protein	p.REF	G S E S Y E E D P Y L V V N P N Y L L E
CDS	p.ALT	.
	c.REF	GGAAGTGAGAGCTATGAAGAAGATCCCTACTTGGTAGTTAACCTAACTACTTGCTCGAA
	c.ALT1

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

Protein summary

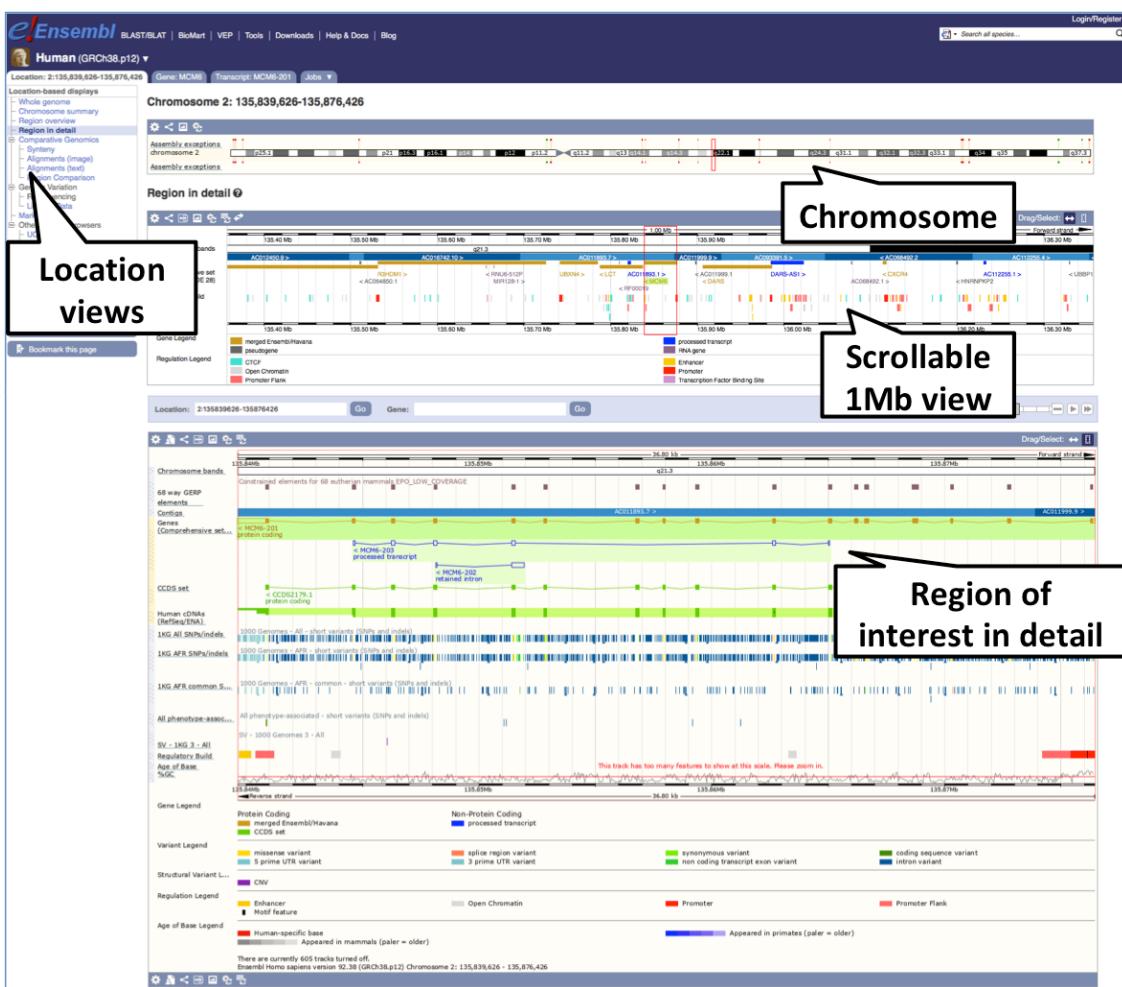
Protein domains for ENSP00000264156.2



Demo part 2: Finding variants in genomic locations

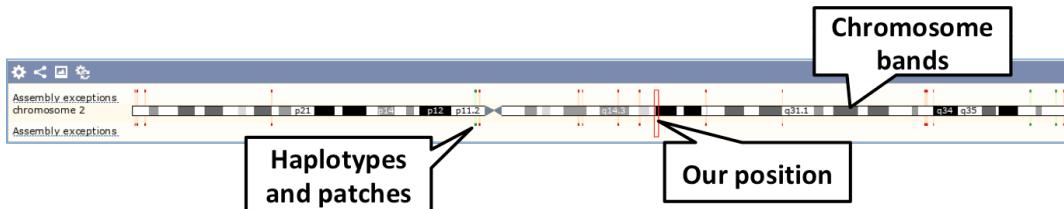
Let's have a look at variants in the [Location tab](#). Click on the [Location tab](#) in the top bar.

The screenshot shows the Ensembl homepage with the species set to Human (GRCh38.p12). A search has been performed for the gene **MCM6**, resulting in the location **2:135,839,626-135,876,426**. The URL in the address bar is https://www.ensembl.org/Homo_sapiens/Location/contig?g=MCM6&t=2&start=135839626&end=135876426.

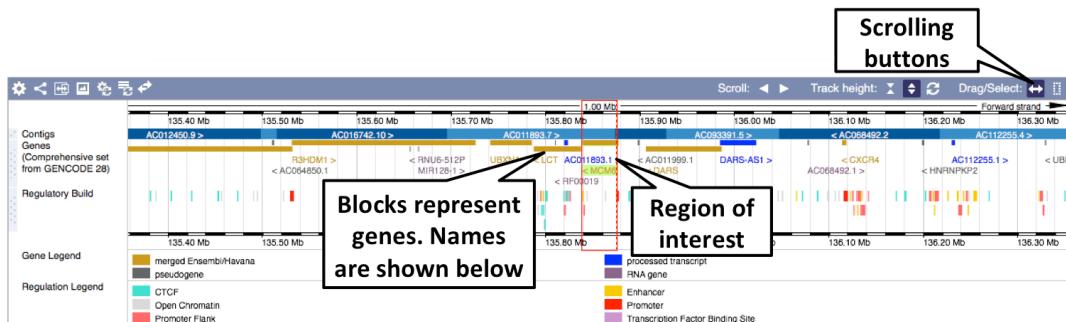


The Region in detail page is made up of three images, let's look at each one on detail.

The first image shows the chromosome:



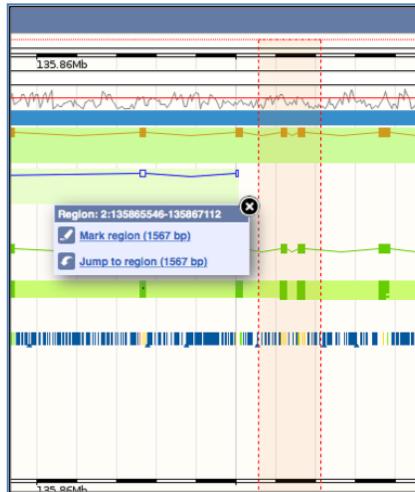
The second image shows a 1Mb region around our selected region. This view allows you to scroll back and forth along the chromosome.



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.



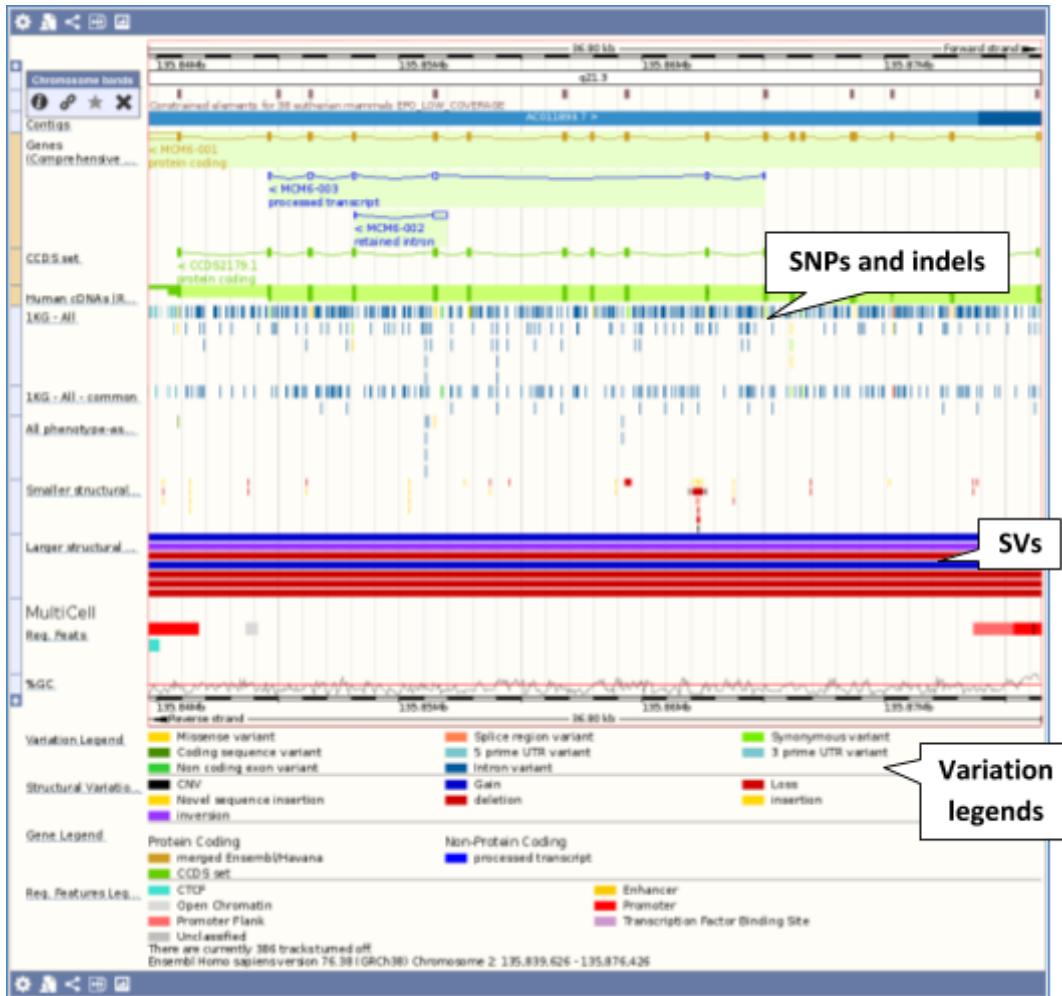
We can edit what we see on this page by clicking on the blue **Configure this page** menu at the left. Open **Variation** from the menu.

Section	Sub-section	Source	Star Rating
Variation	Sequence variants	Sequence variants (dbSNP and all other sources)	★ ⓘ
		dbSNP variants	★ ⓘ
		ExAC - short variants (SNPs and indels)	★ ⓘ
		DECIPHER variants	★ ⓘ
	1000 Genomes	1000 Genomes - All - short variants (SNPs and indels)	★ ⓘ
1000 Genomes - All - common - short variants (SNPs and indels)		★ ⓘ	
1000 Genomes - AFR - short variants (SNPs and indels)		★ ⓘ	
1000 Genomes - AFR - common - short variants (SNPs and indels)		★ ⓘ	
Phenotype-associated variants	1000 Genomes - AMR - short variants (SNPs and indels)	★ ⓘ	
	1000 Genomes - AMR - common - short variants (SNPs and indels)	★ ⓘ	
	1000 Genomes - EAS - short variants (SNPs and indels)	★ ⓘ	
	1000 Genomes - EAS - common - short variants (SNPs and indels)	★ ⓘ	

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.

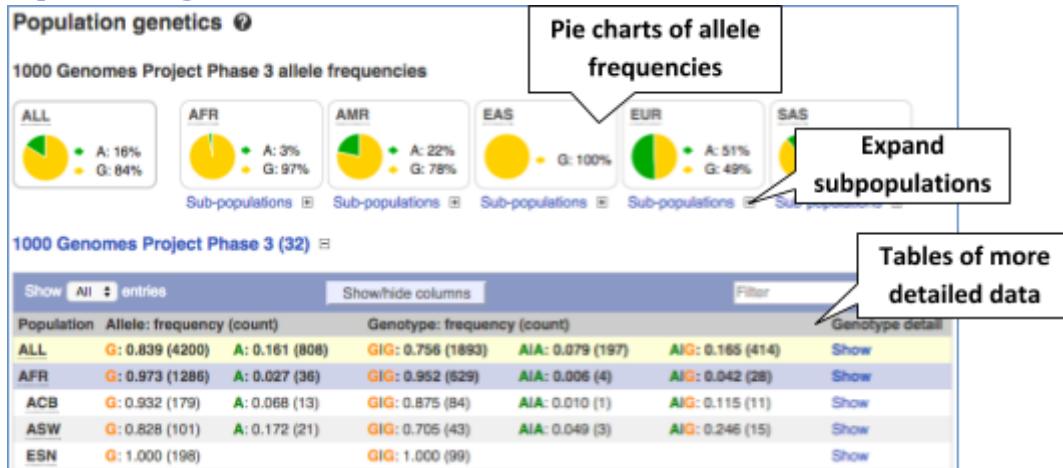
The icons show you what information is available for this variant. Click on **Genes and regulation**, or follow the link at left.

Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen	Detail
ENSG00000076003	ENST00000264156 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: protein_coding										
ENSG00000076003	ENST00000264156 (-)	C (G)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: protein_coding										
ENSG00000076003	ENST00000483902 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: retained_intron										
ENSG00000076003	ENST00000483902 (-)	C (G)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: retained_intron										
ENSG00000076003	ENST00000492091 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: processed_transcript										
ENSG00000076003	ENST00000492091 (-)	C (G)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: processed_transcript										

Regulatory feature	Cell type	Feature type	Allele	Consequence type	Variant position
ENSR00001936131	M2Macrophage:CordBlood:hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR00001936131	M2Macrophage:CordBlood:hist	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR00001936131	A549	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR00001936131	A549	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR00001936131	DND-41	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR00001936131	DND-41	enhancer	C	Regulatory region variant	200 (out of 400)
ENSP00001936131	CD14+CD16-Monocyte:CordBlood:hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSP00001936131	CD14+CD16-Monocyte:CordBlood:hist	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR00001936131	M0Macrophage:CordBlood:hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR00001936131	M0Macrophage:CordBlood:hist	enhancer	C	Regulatory region variant	200 (out of 400)

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.

The figure shows the 'Phenotype Data' section. It lists significant associations for the trait 'LACTASE PERSISTENCE'. The table includes columns for 'Phenotype, disease and trait', 'Source(s)', 'Mapped Terms', 'Ontology Accessions', 'Study', 'Clinical significance', 'Reported gene(s)', 'Associated allele', and 'Statistics'. Callout boxes highlight 'Significant association(s)' and 'Phenotype, disease and trait'.

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	Study	Clinical significance	Reported gene(s)	Associated allele	Statistics
Body Mass Index	NHGRI-EBI GWAS catalog	body mass index, longitudinal BMI measurement	EFO-0004340, EFO-0005937	PMID-25673413	-	MCM6	A	p-value: 5.00e-6 beta coefficient: kg/m ² increase
Body Mass Index	NHGRI-EBI GWAS catalog	body mass index, longitudinal BMI measurement	EFO-0004340, EFO-0005937	PMID-25673413	-	MCM6	A	p-value: 2.00e-6 beta coefficient: kg/m ² increase
Hip circumference (psychosocial stress interaction)	NHGRI-EBI GWAS catalog	hip circumference, psychosocial stress measurement	EFO-0005093, EFO-0006783	PMID-25673412	-	MCM6	A	p-value: 2.00e-6 beta coefficient: unit increase
LACTASE PERSISTENCE	OMIM	Abnormality of metabolism/homeostasis, Autosomal recessive inheritance, Diarrhea, Lactose intolerance	HP-0000007, HP-0001939, HP-0002014, HP-0004789	MIM-801806	-	MCM6	0001	-
LACTASE PERSISTENCE	ClinVar	Abnormality of metabolism/homeostasis, Autosomal recessive inheritance, Diarrhea, Lactose intolerance	HP-0000007, HP-0001939, HP-0002014, HP-0004789	-	LCT, MCM6	A	-	

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.

Loci associated with LACTASE PERSISTENCE ⓘ						
Filter		Feature type: All	Annotation source: All			
Name(s)	Type	Genomic location (strand)	Reported gene(s)	Annotation source	Study	Show/hide columns
rs41525747	Variant	2:135851073 (+)	MCM6	OMIM ⓘ	MIM:601806 ⓘ	
rs41525747	Variant	2:135851073 (+)	MCM6	ClinVar ⓘ	-	
rs145946881	Variant	2:135851176 (+)	MCM6	ClinVar ⓘ	-	
rs4988235	Variant	2:135851076 (+)	MCM6	OMIM ⓘ	MIM:601806 ⓘ	
rs145946881	Variant	2:135851176 (+)	MCM6	OMIM ⓘ	MIM:601806 ⓘ	
rs182549	Variant	2:135859184 (+)	MCM6	ClinVar ⓘ	-	
rs41380347	Variant	2:135851081 (+)	MCM6	ClinVar ⓘ	-	
rs4988235	Variant	2:135851076 (+)	LCT, MCM6	ClinVar ⓘ	-	
rs41380347	Variant	2:135851081 (+)	MCM6	OMIM ⓘ	MIM:601806 ⓘ	
rs182549	Variant	2:135859184 (+)	MCM6	OMIM ⓘ	MIM:601806 ⓘ	

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.

Phylogenetic Context ⓘ

Alignment: 8 primates EPO Select another alignment Choose your alignment

Download alignment

Variants Focus variant Intrinsic

Markup loaded

Human > chromosome:GRCh38:2:135851066:135851086:1
 Chimpanzee > chromosome:CHIMP2:1:4:2B:139811109:139811129:1
 Gorilla > chromosome:gorGor3:1:2B:22952947:22952967:1
 Orangutan > chromosome:PPYG2:2B:24805669:24805689:1
 Vervet-AGM > chromosome:ChlSab1:1:10:20064555:20064575:1
 Macaque > chromosome:Mmul_8:0:1:12:21353752:21353772:1
 Olive baboon > chromosome:PapAnu2:0:13:107195936:107195956:-1
 Marmoset > chromosome:C_jacchus3:2:1:6:83979656:83979676:-1

Aligned regions

	SRKV	RYMY
Human	GAGGCCA GCGC TACATTATC	
Chimpanzee	GAGGCCAGGGCTACATTATC	
Gorilla	GAGGCCAGGGCTACATTATC	
Orangutan	GAGGCCAGGGCTACATTATC	
Vervet-AGM	GAGGCCAGGGCTACATTATC	
Macaque	GAGGCCAGGGCTACATTATC	
Olive baboon	GAGGCCAGGGCTACATTATC	
Marmoset	GAGGCCAGGGCTACATTATC	

SNP of interest

Alignment between 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.

Linkage disequilibrium

Pairwise linkage disequilibrium data by population

Focus variant: rs4988235
Enter the name for the second variant: Compute (e.g. rs678)

Calculate LD values between 2 variants of interest

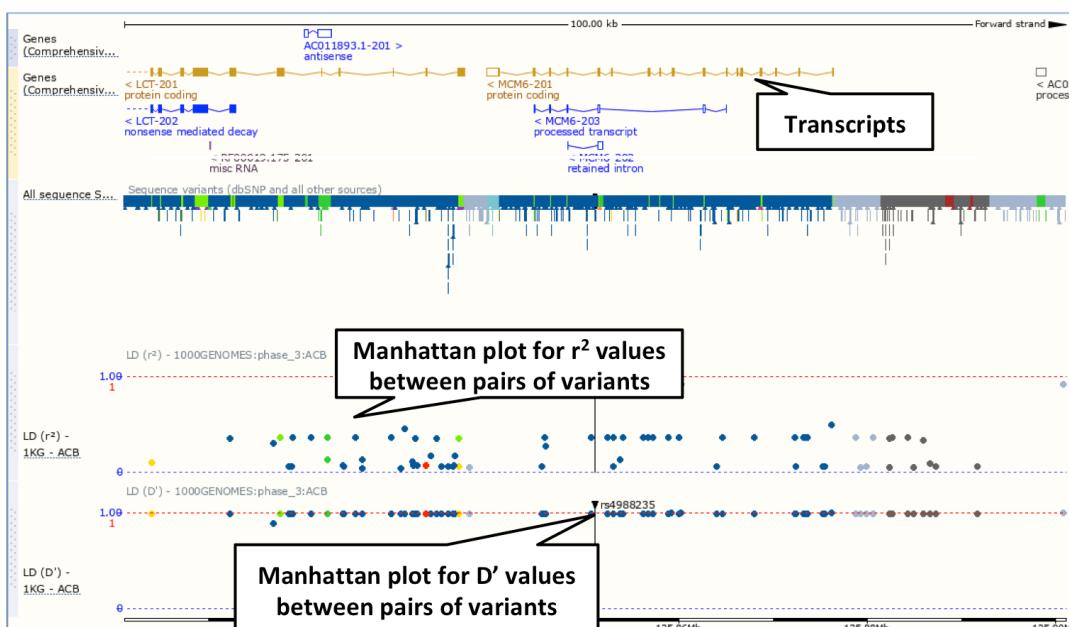
Links to linkage disequilibrium data by population

Population	Description	LD Manhattan plot	Variants in high LD	LD plot
1000GENOMES:phase_3:ACB	African Caribbean in Barbados	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:ASW	African Ancestry in Southwest US	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:ESN	Esan in Nigeria	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:GWD	Gambian in Western Division, The... (more)	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:LWK	Luhya in Webuye, Kenya	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:MSL	Mende in Sierra Leone	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:YRI	Yoruba in Ibadan, Nigeria	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:CLM	Colombian in Medellin, Colombia	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:MXL	Mexican Ancestry in Los Angeles... (more)	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:PEL	Peruvian in Lima, Peru	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:PUR	Puerto Rican in Puerto Rico	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:CDX	Chinese Dai in Xishuangbanna, China	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:CHB	Han Chinese in Beijing, China	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:CHS	Southern Han Chinese, China	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:JPT	Japanese in Tokyo, Japan	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:KHV	Kinh in Ho Chi Minh City, Vietnam	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:CEU	Utah residents with Northern and... (more)	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:FIN	Finnish in Finland	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:GBR	British in England and Scotland	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:IBS	Iberian populations in Spain	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:TSI	Toscani in Italy	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:BEB	Bengali in Bangladesh	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:GIH	Gujarati Indian in Houston, TX	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:ITU	Indian Telugu in the UK	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:PGL	Punjabi in Lahore, Pakistan	[View plot]	Show	[View plot] [View table]
1000GENOMES:phase_3:STU	Sri Lankan Tamil in the UK	[View plot]	Show	[View plot] [View table]

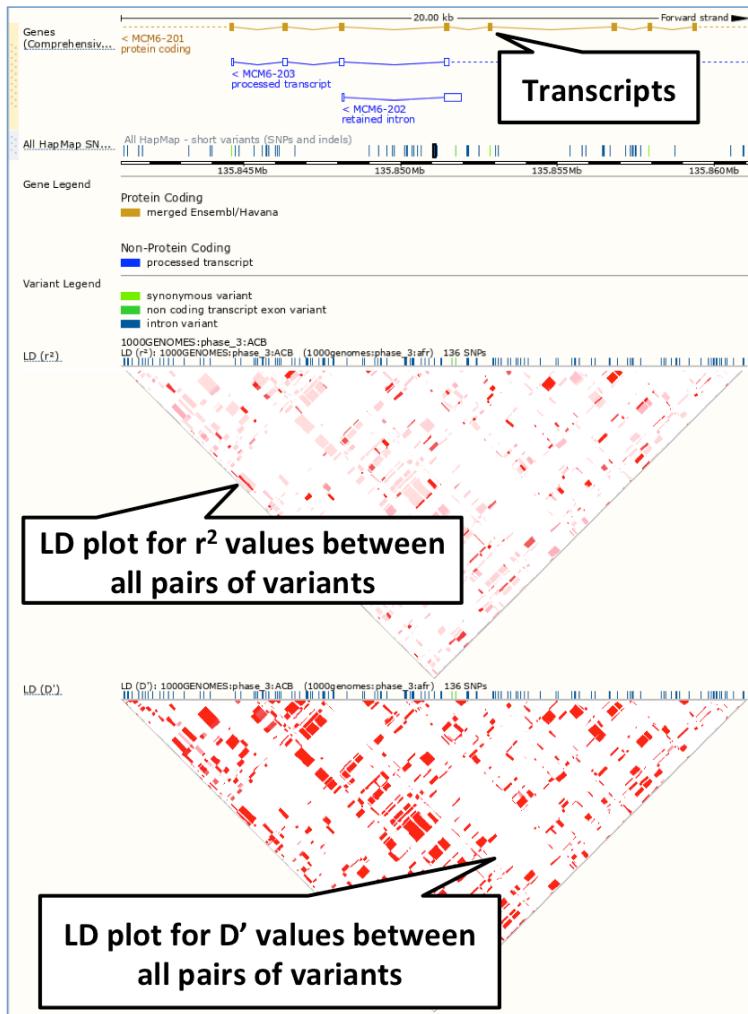
View Manhattan plot displaying LD values between variant of interest and surrounding variants

View LD plot displaying LD values between all pairs of variants in the genomic location

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 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/ 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*)

