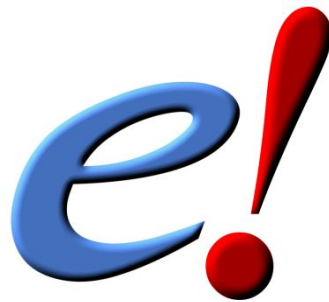


Browsing Genomes with Ensembl



www.ensembl.org
www.ensemblgenomes.org

Coursebook v90

[http://training.ensembl.org/events/2017/
2017-11-06-VEPTC](http://training.ensembl.org/events/2017/2017-11-06-VEPTC)

Prague – 6th November 2017



Introduction to Ensembl

Getting started with Ensembl

www.ensembl.org

Ensembl is a project based at the EBI ([European Bioinformatics Institute](http://www.ebi.ac.uk)) 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

Aken, B *et al.*

Ensembl 2017

Nucleic Acids Research (Database Issue)

<http://nar.oxfordjournals.org/content/early/2016/11/28/nar.gkw1104.full>

Kersey, PJ *et al.*

Ensembl Genomes 2013: scaling up access to genome-wide data

Nucleic Acids Research (Database Issue)

<http://nar.oxfordjournals.org/content/early/2015/12/19/nar.gkv1157.full>

For a complete list of publications, visit:

<http://www.ensembl.org/info/about/publications.html>

<http://ensemblgenomes.org/info/publications>

Exploring the Ensembl genome browser

Demo: Ensembl species

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 callouts pointing to specific features:

- Link back to homepage**: Points to the Ensembl logo in the top left.
- Ensembl tools**: Points to the navigation bar containing links like BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors.
- Blue bar remains visible at the top of every page**: Points to the dark blue header bar.
- Search**: Points to the search bar in the top right corner.
- Search**: Points to the search bar in the main content area.
- Drop down list to species**: Points to the species selection dropdown menu.
- See the current release number and what's new**: Points to the 'What's New in Ensembl Release 89' section.
- How-tos for commonly used Ensembl features**: Points to the 'Find a Data Display' 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 Ensembl homepage. It features a search bar at the top, a navigation bar, and a main content area with various tools and links. The 'About this archive' section on the right provides information about the GRCh37 assembly and its history.

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

The screenshot shows the Ensembl Genomes homepage. At the top, there's a navigation bar with links: About us, Genomes, Data types, Data access, FAQs, and a taxonomic menu: Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates. The main heading is "Ensembl Genomes: Extending Ensembl across the taxonomic space." Below this, there are several featured items: "New bread wheat assembly (Triticum aestivum)", "HMME search across Ensembl Genomes", "Ensembl Plants Archive Site", and "Hundreds of RNA-Seq Studies for Plants!". A survey link is also present: "Please take our user survey!". On the right side, there's a section for "Links to taxon-specific pages" with links for Bacteria, Fungi, Plants, Protists, and Vertebrates. A "News" section follows, stating "Ensembl Genomes 34 has been released." and providing links to release notes for each taxon. A "Have a question?" box contains a "Link to Ensembl" button and text about asking questions. At the bottom right, there's a "User Log In" link.

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

The screenshot shows the Ensembl Protists homepage. It features a search bar at the top, a "Featured content" section with a featured image of a microorganism, and a "Popular genomes" section listing various protist species like Plasmodium falciparum and Dictyostelium discoideum. There's also a "What's new in Release 17 (January 2013)" section with updates on new species and protein features.

Protists

The screenshot shows the Ensembl Fungi homepage. It includes a search bar, a "Featured content" section with a featured image of a fungus, and a "Popular genomes" section listing various fungal species like Aspergillus nidulans and Saccharomyces cerevisiae. The "What's new in Release 17 (January 2013)" section highlights updates on new species and protein features.

Fungi

The screenshot shows the Ensembl Metazoa homepage. It features a search bar, a "Featured content" section with a featured image of a metazoan, and a "Popular genomes" section listing various metazoan species like Caenorhabditis elegans and Drosophila melanogaster. The "What's new in Release 17 (January 2013)" section highlights updates on new species and protein features.

Metazoa

The screenshot shows the Ensembl Plants homepage. It includes a search bar, a "Featured content" section with a featured image of a plant, and a "Popular genomes" section listing various plant species like Arabidopsis thaliana and Oryza sativa. The "What's new in Release 17 (January 2013)" section highlights updates on new species and protein features.

Plants

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 add variants to all other sequence views in the same way.

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 [Variation table](#) link at the left of the gene tab.

Gene: MCM6 ENSG00000278033
Description: minichromosome maintenance complex component 6 [Source:HGNC Symbol;Acc:HGNC:6949]

Synonyms: MCG40308, Miu5, P105MCM
Location: Chromosome 2: 135,839,626-135,870,426 reverse strand.
GRCh38:CM000664.2

About this gene: This gene has 3 transcripts (splice variants), 69 orthologues, is a member of 1 Ensembl protein family and is associated with 1 phenotype.
Transcripts: [Show transcript table](#)

Variant table
This table shows known variants for this gene. Use the 'Consequence Type' filter to view a subset of these.

Filter: [Global MAF: All](#) [SIFT: All](#) [PolyPhen: All](#) [Consequences: All](#) [Filter Other Columns](#)

Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA coord	SIFT	PolyPhen	Transcript
rs34489316	2: between 135834658 & 135834659	-T	(-)	insertion	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs53607424	2:135834696	G/A	(-)	SNP	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs199775007	2:135834713	G/A	(-)	SNP	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs562081428	2:135834715	G/A	0.000 (A)	SNP	dbSNP		-	Downstream gene variant	-	-	-	-	ENST00000264156
rs200718455	2:135834737	G/T	(-)	SNP	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs78141345	2:135834739	G/A	0.298 (A)	SNP	dbSNP		-	Downstream gene variant	-	-	-	-	ENST00000264156
rs544251553	2:135834740	C/A	0.000 (A)	SNP	dbSNP		-	Downstream gene variant	-	-	-	-	ENST00000264156
rs555872510	2:135834741	AAAAAAAAA A/-	0.200 (-)	deletion	dbSNP		-	Downstream gene variant	-	-	-	-	ENST00000264156
rs753376102	2:135834742	-AAA	(-)	insertion	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs199944782	2:135834808	A/G	(-)	SNP	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs138596223	2:135834811	G/A	(-)	SNP	dbSNP		-	Downstream gene variant	-	-	-	-	ENST00000264156
rs60298631	2: between 135834810 & 135834811	-AAA/AAAA	(-)	insertion	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs60298631	2: between 135834810 & 135834811	-AAA/AAAA	(-)	insertion	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156
rs368911074	2:135834811-135834817	GAAGAAG/-	(-)	deletion	dbSNP	-	-	Downstream gene variant	-	-	-	-	ENST00000264156

You can filter the table to show only the variants you're interested in by clicking on the filters (e.g. [Consequences](#)) and selecting appropriate filter options. You can add multiple filters at once, and remove filters by clicking the 'X' next to the active filter.

Filter				Global MAF: All		SIFT: All		PolyPhen: All		Consequences: All		Filter Other Columns				
Consequences												(25/26 0s)				
<div>Turn All OffPTVPTV & MissenseOnly ExonicTurn All On</div>																
PTV = Protein Truncating Variant																
<div><div>Transcript ablation</div><div>Inframe deletion</div><div>Missense variant</div><div>Splice donor variant</div><div>Splice acceptor variant</div><div>Stop gained</div><div>Frameshift variant</div><div>Stop lost</div><div>Start lost</div><div>Transcript amplification</div><div>Inframe insertion</div></div>												<div><div>5 prime UTR variant</div><div>3 prime UTR variant</div><div>Non coding transcript exon variant</div><div>Intron variant</div><div>NMD transcript variant</div><div>Non coding transcript variant</div><div>Upstream gene variant</div><div>Downstream gene variant</div></div>				
<div>ApplyCancel</div>																
Variant ID	Chr: bp	Alleles	Gl									AA coord	SIFT	PolyPhen	Transcript	
rs34489316	2: between 135834658 & 135834659	-T	(-)									-	-	-	ENST00000264156	
rs33907424	2:135834696	G/A	(-)									-	-	-	ENST00000264156	
rs190775007	2:135834713	G/A	(-)									-	-	-	ENST00000264156	
rs582081428	2:135834715	G/A	0.0									-	-	-	ENST00000264156	
rs200718455	2:135834737	G/T	(-)									-	-	-	ENST00000264156	
rs78141345	2:135834739	G/A	0.2									-	-	-	ENST00000264156	
rs544251553	2:135834781	C/A	0.0									-	-	-	ENST00000264156	
rs555873510	2:135834788-135834797	AAAAAAAAA A/-	0.200 (-)	deletion	dbSNP							Downstream gene variant	-	-	-	ENST00000264156

Phenotype

Phenotype(s), disease(s) and trait(s) associated with this gene ENSG00000076003

Phenotype, disease and trait	Source	Genomic locations	Biomart	Number of variants	Show/Hide details
LACTOSE INTOLERANCE, ADULT TYPE	MIM morbid	View on Karyotype		5	Show

Phenotype, disease and trait annotations associated with variants in this gene

Phenotype, disease and trait	Source(s)	Genomic locations	Biomart	Number of variants	Show/Hide details
ALL variants with a phenotype annotation				5	Show
LACTASE PERSISTENCE	ClinVar, OMIM	View on Karyotype		5	Show

Phenotype, disease and trait annotations associated orthologues of this gene in other species

Phenotype, disease and trait	Source	Species	Gene
Lactose Intolerance, Adult Type	RGD	Rat (Rattus norvegicus)	ENSRNOG00000003703 Mcm6

Phenotypes associated with the gene

Phenotypes associated with variants in the gene

Click to see list of variants

Phenotypes associated with orthologues of the gene

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

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help

Human (GRCh38.p5) **Location: 2:135,839,626-135,876,426** Gene: MCM6

Configure this page and open [Variation](#) from the left-hand menu.

Find a track

- Active tracks
- Favourite tracks
- Track order
- Search results
- Your data (2/2)
- Sequence and assembly (3/32)
 - Sequence (2/4)
 - Markers (0/1)
 - GRC alignments (1/15)
 - Simple features (0/4)
 - Clones & misc. regions (0/8)
- Genes and transcripts (2/65)
 - Genes (2/7)
 - Prediction transcripts (0/1)
 - LRG (0/1)
 - RNASeq models (0/56)
- mRNA and protein alignments (1/6)
 - mRNA alignments (1/3)
 - EST alignments (0/1)
 - Protein alignments (0/2)
- Variation (2/79)**
 - Sequence variants (0/18)
 - Phenotype, disease and curated variants (1/21)
 - Arrays and other (0/17)
 - Failed variants (0/1)
 - Structural variants (1/22)
- Somatic mutations (0/5)
 - Somatic variants (0/3)
 - Somatic structural variants (0/2)

Select from available configurations: Current unsaved [Save current configuration](#)

Variation

Enable/disable all Sequence variants

- ☐ Sequence variants (dbSNP and all other sources) ★ ?
- ☐ dbSNP variants ★ ?
- ☐ ExAC - short variants (SNPs and indels) ★ ?
- ☐ DECIPHER variants ★ ?
- ☐ LOVD variants ★ ?

Enable/disable all 1000 Genomes

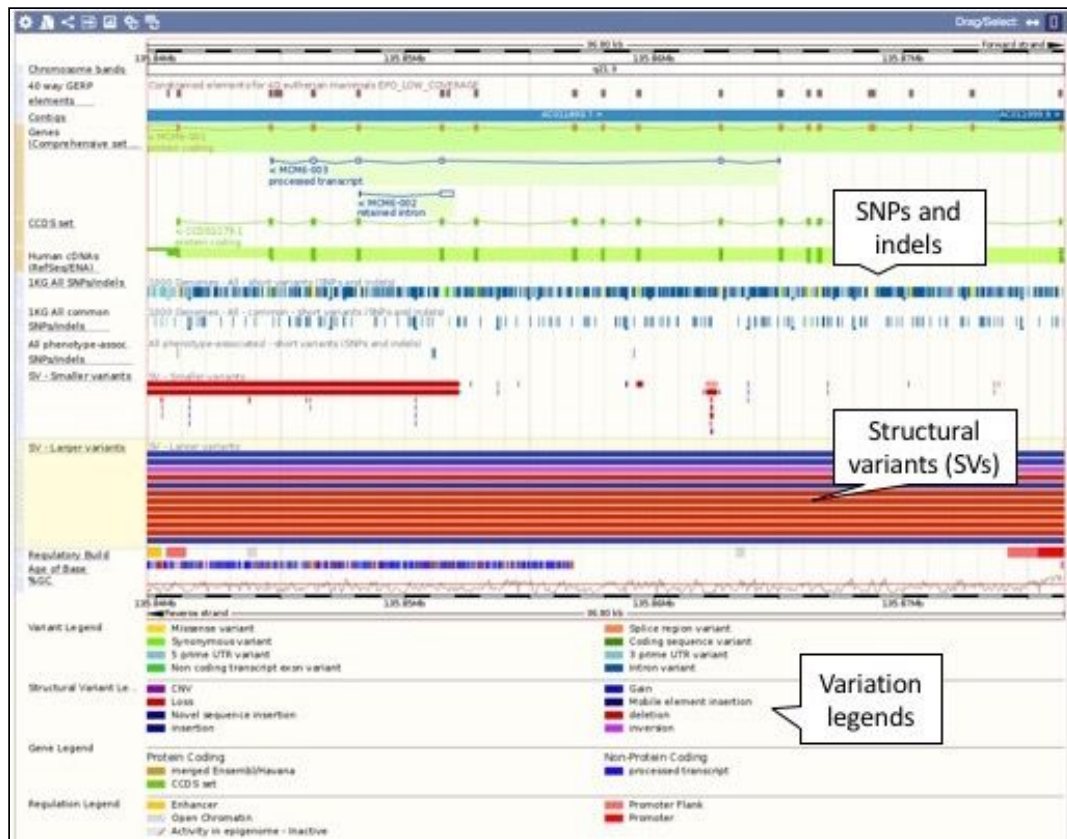
- ☐ 1000 Genomes - All - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - All - common - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - AFR - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - AFR - common - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - AMR - short variants (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) ★ ?
- ☐ 1000 Genomes - EUR - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - EUR - common - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - SAS - short variants (SNPs and indels) ★ ?
- ☐ 1000 Genomes - SAS - common - short variants (SNPs and indels) ★ ?

HapMap

There are various options for turning on variants. You can turn on variants by source, frequency, presence of a phenotype, or the 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.

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.

Human (GRCh38.p5) Location: 2:135,850,576-135,851,576 Variant: rs4988235 Jobs

Variant displays

- Explore this variant
- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Sample genotypes
- Linkage disequilibrium
- Phenotype data
- Phylogenetic Context
- Citations

rs4988235 SNP

Original source: Variants (including SNPs and indels) imported from dbSNP (release 146) | [View in dbSNP](#)

Alleles: G/A/C | Ancestral: G | Ambiguity code: V | MAF: 0.16 (A)

Location: Chromosome 2:135851076 (forward strand) | [View in location tab](#)

Co-located variant: HGMD-PUBLIC CR024269

Most severe consequence: Intron variant | [See all predicted consequences \(Genes and regulation\)](#)

Evidence status

Clinical significance

Synonyms: ClinVar SCV000028329

HGVS names: This variant has 8 HGVS names - [Show](#)

Genotyping chips: This variant has assays on 7 chips - [Show](#)

About this variant: This variant overlaps 6 transcripts, 1 regulatory feature, has 3271 sample genotypes, is associated with 2 phenotypes and is mentioned in 67 citations.

Explore this variant

- Genomic context
- Genes and regulation
- Population genetics
- Sample genotypes
- Linkage disequilibrium
- Phenotype data
- Citations
- Phylogenetic context
- Flanking sequence

Variant information

Variation views

Variation icons. These go to the same places as the links on the left

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

Genes and regulation

Gene and Transcript consequences

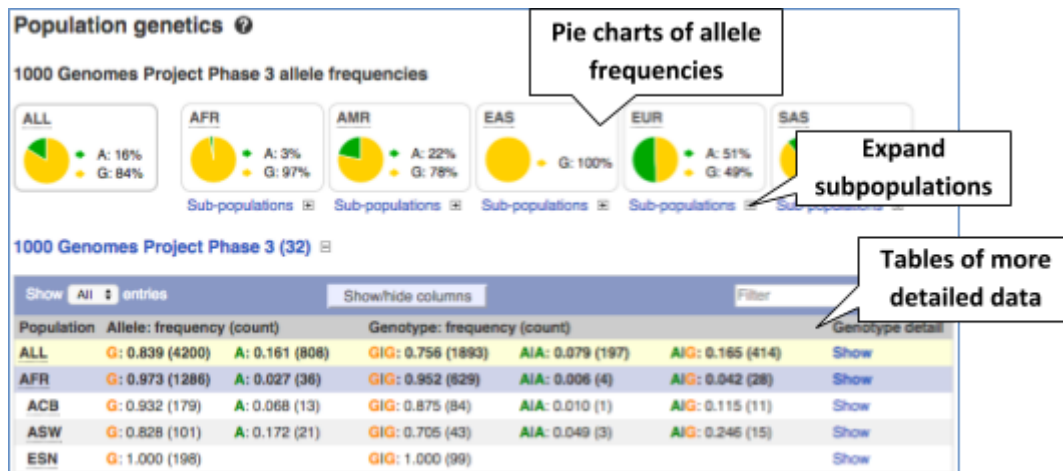
Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen	Detail
ENSG00000076003	ENST00000284156 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: protein_coding	C (G)	Intron variant	-	-	-	-	-	-	-	Show
ENSG00000076003	ENST00000284156 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: protein_coding	C (G)	Intron variant	-	-	-	-	-	-	-	Show
ENSG00000076003	ENST00000483902 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: retained_intron	C (G)	Non coding transcript variant	-	-	-	-	-	-	-	Show
ENSG00000076003	ENST00000483902 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: retained_intron	C (G)	Non coding transcript variant	-	-	-	-	-	-	-	Show
ENSG00000076003	ENST00000492091 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: processed_transcript	C (G)	Non coding transcript variant	-	-	-	-	-	-	-	Show
ENSG00000076003	ENST00000492091 (-)	A (T)	Intron variant	-	-	-	-	-	-	-	Show
HGNC: MCM6	biotype: processed_transcript	C (G)	Non coding transcript variant	-	-	-	-	-	-	-	Show

Regulatory feature consequences

Regulatory feature	Cell type	Feature type	Allele	Consequence type	Variant position
ENSR000001936131	M2Macrophage:CordBlood.hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR000001936131	M2Macrophage:CordBlood.hist	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR000001936131	A549	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR000001936131	A549	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR000001936131	DND-41	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR000001936131	DND-41	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR000001936131	CD14+CD16-Monocyte:CordBlood.hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR000001936131	CD14+CD16-Monocyte:CordBlood.hist	enhancer	C	Regulatory region variant	200 (out of 400)
ENSR000001936131	M0Macrophage:CordBlood.hist	enhancer	A	Regulatory region variant	200 (out of 400)
ENSR000001936131	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.

Phenotype Data

Significant association(s)

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: 0.016 kg/m2 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: 0.016 kg/m2 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-8 beta coefficient: 0.0217 unit increase
LACTASE PERSISTENCE	OMIM	Abnormality of metabolism/homeostasis, Autosomal recessive inheritance, Diarrhea, Lactose intolerance	HP:0000007, HP:0001939, HP:0002014, HP:0004789	MIM:601806	-	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
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** Intronic
 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:ChiSab1.1:10:20064555:20064575:1](#)
 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

Human: GAGGCCA **GGG** GCTACATTATC
 Chimpanzee: GAGGCCAGGGGCTACATTATC
 Gorilla: GAGGCCAGGGGCTACATTATC
 Orangutan: GAGGCCAGGGGCTACATTATC
 Vervet-AGM: GAGGCCAGGGGCTACATTATC
 Macaque: GAGGCCAGGGGCTACATTATC
 Olive baboon: GAGGCCAGGGGCTACATTATC
 Marmoset: GAGGCCAGGGGCTACATTATC

SNP of interest

Alignment between species

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

You can also analyse the effect of variants using the Transcript Haplotype view. The Transcript Haplotype view allows you to explore observed transcript sequences that result from variants identified by the **1000 Genomes Project**. Navigate to a transcript of interest and

click on **Haplotypes**. I will use **BRCA2-001 (ENST00000380152)** as an example.

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

Link for Haplotype view

Protein haplotypes

Population frequencies

Protein haplotype	Flags	Frequency (count)	AFR	AMR	EAS	EUR	SAS
2466V>A		0.513 (2569)	0.477 (930)	0.478 (332)	0.54 (144)	0.599 (920)	0.47 (493)
372N>H, 2466V>A		0.202 (1014)	0.085 (26)	0.179 (134)	0.255 (257)	0.272 (274)	0.331 (324)
269N>H, 991N>D, 2466V>A		0.0603 (302)	0.0303 (40)	0.0602 (36)	0.0603 (91)	0.0318 (32)	0.102 (100)
2466V>A, 3412V>V		0.0415 (208)	0.157 (142)	0.0581 (41)	0.0198 (20)	0.00298 (3)	0.00204 (2)
REF		0.0224 (117)	0.0602 (114)	0.00298 (2)		0.000984 (1)	
372N>H, 2466V>A, 2460D>T		0.0156 (78)	0.00227 (9)	0.108 (72)	0.00298 (3)		0.0215 (21)
269N>H, 991N>D, 2466V>A, 2951A>T		0.00938 (47)	0.000758 (1)	0.0317 (22)		0.00298 (3)	
372N>H, 2466V>A, 2944A>E		0.00859 (43)	0.001 (41)	0.00288 (2)			
1915T>M, 2466V>A		0.00839 (42)	0.000758 (1)	0.0187 (13)		0.0039 (24)	0.00409 (4)
2039G>A, 2440G>H, 2466V>A, 3244V>V		0.00679 (34)	0.000758 (1)	0.000758 (1)			
2115G>A, 2466V>A		0.00659 (33)	0.000758 (1)	0.000758 (1)			
372N>H, 2466V>A, 3329K>*, 3327G>G(93)		0.00439 (22)	0.000758 (1)	0.00288 (2)		0.0109 (11)	0.00818 (8)
1364G>L, 2466V>A		0.00439 (22)	0.000758 (1)	0.00288 (2)			
991N>D, 1930D>N, 2466V>A		0.00379 (19)	0.0136 (18)	0.00144 (1)			
2108G>A, 2440G>H, 2466V>A		0.00339 (17)	0.0129 (17)				
372N>H, 1390D>V, 2466V>A		0.00319 (16)	0.0121 (16)				
784M>V, 2466V>A		0.003 (15)			0.0149 (15)		
2674G>A, 2466V>A		0.0028 (14)	0.00883 (13)	0.00144 (1)			
7130G>L, 2466V>A, 3212G>H		0.0024 (12)	0.00908 (12)				
1402G>V, 2466V>A		0.0024 (12)	0.000758 (1)	0.00288 (2)		0.00398 (8)	0.00207 (2)
372N>H, 2466V>A, 2729G>N		0.0024 (12)			0.0109 (11)		0.00102 (1)

Click on **Protein Haplotype 2466V>A** to find out more information about this transcript variant, such as population frequencies and the aligned sequence.

Details of protein haplotype ENSP00000369497:2466V>A

Jump to: [Population frequencies](#) | [Aligned sequence](#) | [Sequence](#) | [Corresponding CDS haplotypes](#) | [Sample data](#)

Population frequencies

Population group	Population	Frequency (count)
AFR	ACB	0.474 (930)
	ASW	0.525 (6)
	ESN	0.490 (9)
	LWK	0.530 (1)
	GWD	0.429 (97)
AMR	MSL	0.447 (76)
	YRI	0.485 (96)
	CLM	0.511 (96)
	MXL	0.469 (60)
	PEL	0.435 (74)
EAS	PUR	0.490 (102)
	CDX	0.581 (108)
	CHB	0.539 (111)
	CHS	0.576 (121)
	JPT	0.519 (108)
EUR	KHV	0.485 (96)
	CEU	0.596 (118)
	FIN	0.621 (123)
	GBR	0.588 (107)
	IBS	0.607 (130)
SAS	TSI	0.584 (125)
	BEB	0.453 (78)
	GIH	0.476 (98)
	ITU	0.446 (91)
	PJL	0.526 (101)
STU	0.451 (92)	

Total count: 2569

		Reference protein sequence											
Protein	p.REF	N	S	N	Q	A	V	A	V	T	F	T	K
	p.ALT	A
CDS	c.REF	AACTCCAAATCAAGCAGTAACTTTCACAAAG											
	c.ALT1
	c.ALT2
	c.ALT3
	c.ALT4
	c.ALT5
	c.ALT6
	c.ALT7
	c.ALT8
	c.ALT9
	c.ALT10
	c.ALT11
	c.ALT12
	c.ALT13
	c.ALT14
	c.ALT15

Demo: The Variant Effect Predictor (VEP)

We have identified four variants on human chromosome nine, an A deletion at 128328461, C->A at 128322349, C->G at 128323079, and G->A at 128322917.

We will use the **Ensembl VEP** to determine:

- If the variants have already been annotated in Ensembl
- The genes affected by my variants
- If any of my variants affect gene regulation

Go to the front page of Ensembl and click on the [VEP button](#).



This page contains information about the VEP, including links to download the script version of the tool. Click on [Launch VEP](#) to open the input form.

Input

Species: Human (Homo sapiens) Assembly: GRCh38

Name for this data (optional):

Either paste data:

```
9 128328461 128328461 A/- + var1
9 128322349 128322349 C/A + var2
9 128323079 128323079 C/G + var3
9 128322917 128322917 G/A + var4
```

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#), [Pileup](#)

[Quick results for first variant >](#)

Or upload file: Choose File No file chosen

Or provide file URL:

Transcript database to use:

- ☒ Ensembl transcripts
- ☐ Gencode basic transcripts
- ☐ RefSeq transcripts
- ☐ Ensembl and RefSeq transcripts

Callouts:

- Give your data a name
- Put your data in here
- You can also upload a file
- Choose your transcript database

The data are in the format:
Chromosome Start End Alleles (reference/mutation) Strand

Put the following into the [Paste data](#) box:

```
9 128328461 128328461 A/- + var1
9 128322349 128322349 C/A + var2
9 128323079 128323079 C/G + var3
9 128322917 128322917 G/A + var4
```

The VEP will automatically detect that the data are in Ensembl default format. (Note that the deletion in the first variant is indicated by “-”.)

There are further options that you can choose for your output. These are categorised as [Identifiers and frequency data](#), [Filtering options](#), and [Extra options](#). Let’s open all the menus and take a look.

Output options

Identifiers and frequency data Additional identifiers for genes, transcripts and variants; frequency data

Identifiers

Gene symbol:	<input checked="" type="checkbox"/>
CCDS:	<input type="checkbox"/>
Protein:	<input type="checkbox"/>
Uniprot:	<input type="checkbox"/>
HGVS:	<input type="checkbox"/>
CSN ^(PI) :	<input type="checkbox"/>
Unshifted HGVS ^(PI) :	<input type="checkbox"/>

Frequency data

Find co-located known variants:

Frequency data for co-located variants:

- ☒ 1000 Genomes global minor allele frequency
- ☐ 1000 Genomes continental allele frequencies
- ☐ ESP allele frequencies

PubMed IDs for citations of co-located variants: ☒

Include flagged variants: ☐


ExAC frequencies^(PI): ☐

^(PI) = functionality from [YEP plugin](#)

Which identifiers
do you want to
see?

Find out if variants
already exist in
our database.



Get frequency
data.

Extra options  e.g. SIFT, PolyPhen and regulatory data

Miscellaneous


Transcript biotype:	<input checked="" type="checkbox"/>
Protein domains:	<input type="checkbox"/>
Exon and intron numbers:	<input type="checkbox"/>
Transcript support level:	<input checked="" type="checkbox"/>
Identify canonical transcripts:	<input type="checkbox"/>
miRNA structure ^(*) :	<input type="checkbox"/>
Upstream/Downstream distance ^(*) :	<input checked="" type="radio"/> Disabled <input type="radio"/> Enabled

Pathogenicity predictions

SIFT:	Prediction and score 
PolyPhen:	Prediction and score 
dbNSFP ^(*) :	<input checked="" type="radio"/> Disabled <input type="radio"/> Enabled
Condel ^(*) :	<input checked="" type="radio"/> Disabled <input type="radio"/> Enabled
LoFoot ^(*) :	<input type="checkbox"/>

Fields marked * are required

Regulatory data

Get regulatory region consequences:	Yes 
-------------------------------------	---


Splicing predictions

dbSNV ^(*) :	<input type="checkbox"/>
GeneSplicer ^(*) :	<input type="checkbox"/>
MaxEntScan ^(*) :	<input type="checkbox"/>


Conservation

BLOSUM62 ^(*) :	<input type="checkbox"/>
---------------------------	--------------------------

^(*) = functionality from VEP plugin

Filtering options  Pre-filter results by frequency or consequence type

Filters

Filter by frequency:	<input checked="" type="radio"/> No filtering <input type="radio"/> Exclude common variants <input type="radio"/> Advanced filtering
Return results for variants in coding regions only:	<input type="checkbox"/>
Restrict results:	Show all results  <small>NB: Restricting results may exclude biologically important data!</small>

Run **Reset**

Find out more about the transcripts affected.

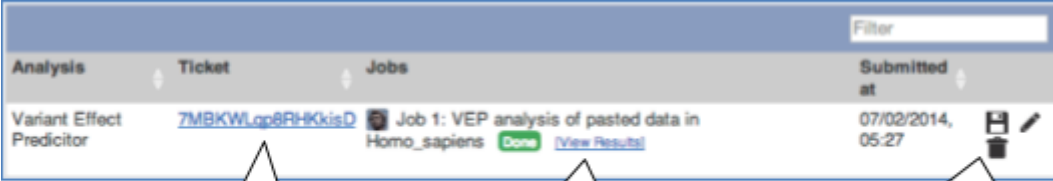
Choose to see scores for protein changes.




Choose to see scores for splicing changes.

Choose to only see common or rare variants

Hover over the options to see definitions.

When you've selected everything you need, scroll right to the bottom and click [Run](#).



Analysis	Ticket	Jobs	Submitted at
Variant Effect Predictor	7MBKWlqp8PHKkisD	Job 1: VEP analysis of pasted data in Homo_sapiens Done View Results	07/02/2014, 05:27   

Your ticket number

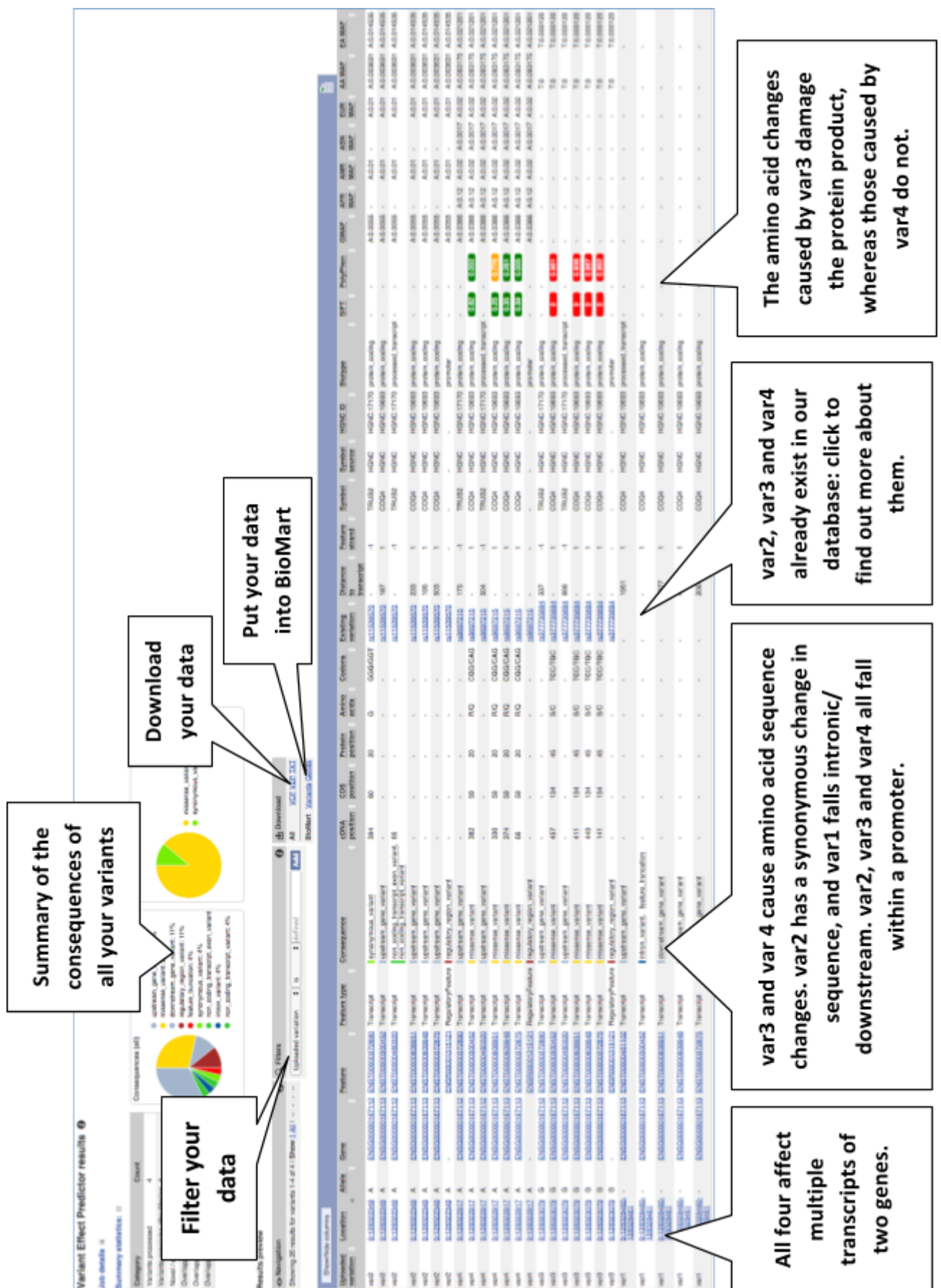
Click to get your results

Buttons to save, edit or delete your job

The display will show you the status of your job. It will say [Queued](#), then automatically switch to [Done](#) when the job is done: you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

Click [View Results](#) once your job is done.

In your results you will see a graphical summary of your data as well as a table of your results. (Note that some empty columns in the results table have been hidden in the following screenshot to save space.)



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 is this SNP found?
- (b) What is the least frequent genotype for this SNP in the Yoruba (YRI) population from the HapMap set?
- (c) What is the ancestral allele? Is it conserved in the 40 eutherian mammals?
- (d) 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: The Variant Effect Predictor (VEP)

Exercise V4 – VEP

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (ENSG00000001626) has revealed the following variants (alleles defined in the forward strand):

- G/A at 7:117,530,985
- T/C at 7: 117,531,038
- T/C at 7: 117,531,068

Use the VEP tool in Ensembl and choose the options to see SIFT and PolyPhen predictions. Do these variants result in a change in the

proteins encoded by any of the Ensembl genes? Which gene? Have the variants already been found?

Exercise V5 – viewing structural variants with the VEP

We have details of a genomic deletion in a breast cancer sample in VCF format:

```
13 32307062 sv1 . <DEL> .. SVTYPE=DEL;END=32332466
```

- (a) What are the HGNC identifiers of the affected genes?
- (b) Does the SV cause deletion of any complete transcripts?
- (c) Display your variant in the Ensembl browser.

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*)

