Exercise: Searching Ensembl Fungi species

Clickable links shown in blue, text to be entered shown in red.

Navigate to fungi.ensembl.org. The number of the release may vary but you'll see a homepage similar to this:

Click on 'View full list of all species', which you can find in section **3: Genome and species directory** shown above.

(a) How many genome assemblies are there for the genus *Bipolaris* in Ensembl Fungi?

Click on the Latin name of your species of interest to go to the species homepage.

(b) Navigate to the species homepage for *Bipolaris oryzae*. What is the name of the genome assembly for *B. oryzae*?

To find out more about the genome assembly and gene annotation, click on More information and statistics.

(c) How long is the *B. oryzae* genome? How many genes have been annotated?

(d) What is the INSDC accession number for *B. oryzae*? What institute submitted the data to INSDC?

Exercise: Ensembl Fungi 'Region in detail' view
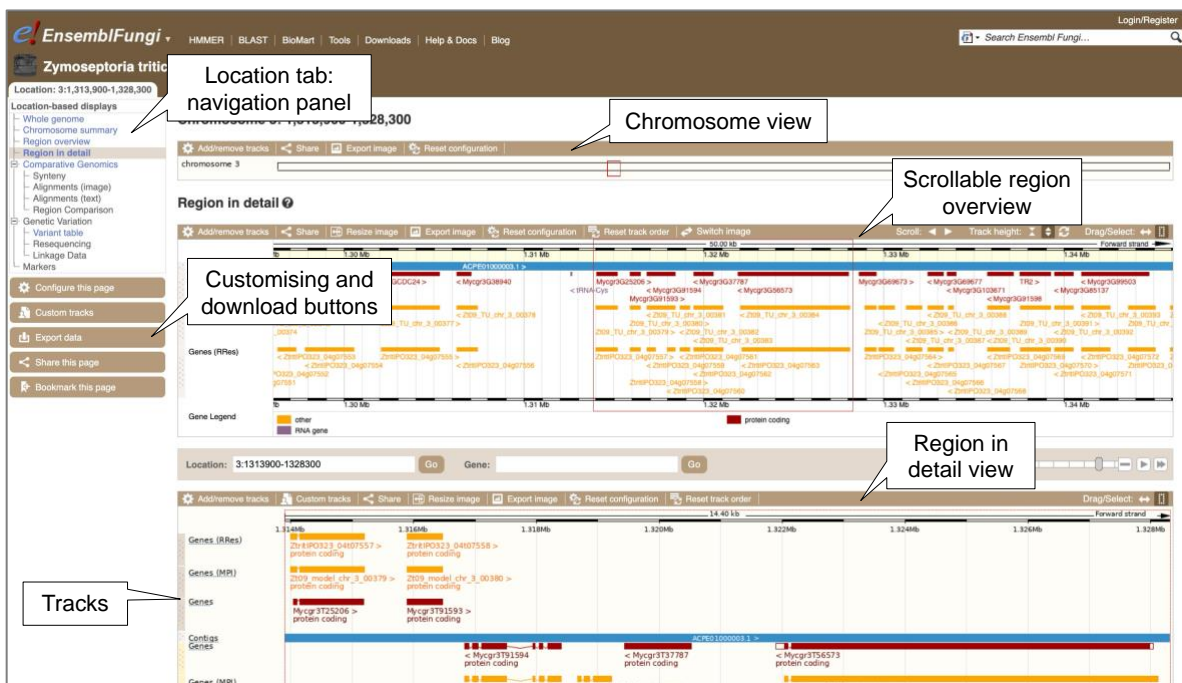
Start at the Ensembl Fungi homepage, fungi.ensembl.org. You can search for a region by typing it into a search box, but you have to specify the species.

(a) Find *Zymoseptoria tritici* (assembly MG2), then type (or copy and paste) 3:1313900-1328300 into the search box. Press enter or click Go to jump directly to the **Region in detail** Page.

Click on the button [?] to open a page-specific help page. These help pages provide links to Frequently Asked Questions (FAQs), a glossary, video tutorials, and a form to contact the Ensembl helpdesk. You can find a help video on this page at http://youtu.be/tTKEvgPUq94.

The **Region in detail** page is made up of three images, similar to this:

Let's look at each image in detail.

The first image shows the chromosome overview. You can jump to a different region by dragging out a box in this image. Drag out a box on the chromosome and a pop-up menu will appear.
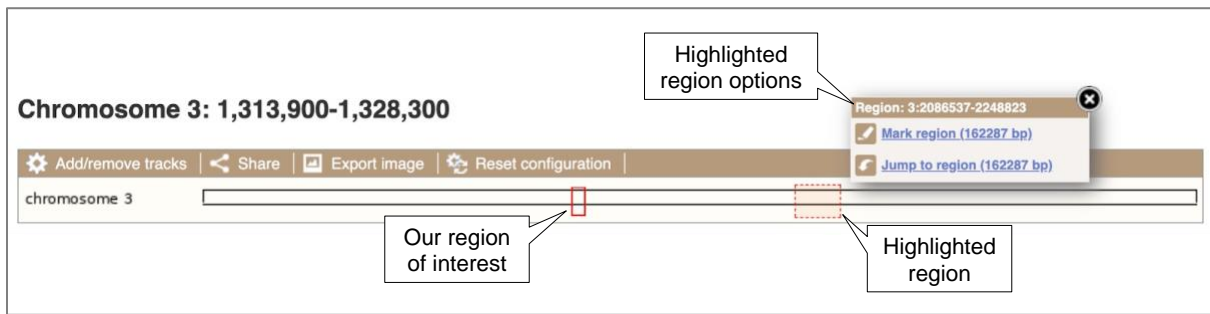
If you would like to move to the region, you could click on Jump to region (### bp). To highlight it, click on Mark region (### bp). For now, we'll close the pop-up by clicking on the X in the top right-hand corner.

The second image is the region overview, showing a 50 kb region around our selected region. This view allows you to scroll back and forth along the chromosome.
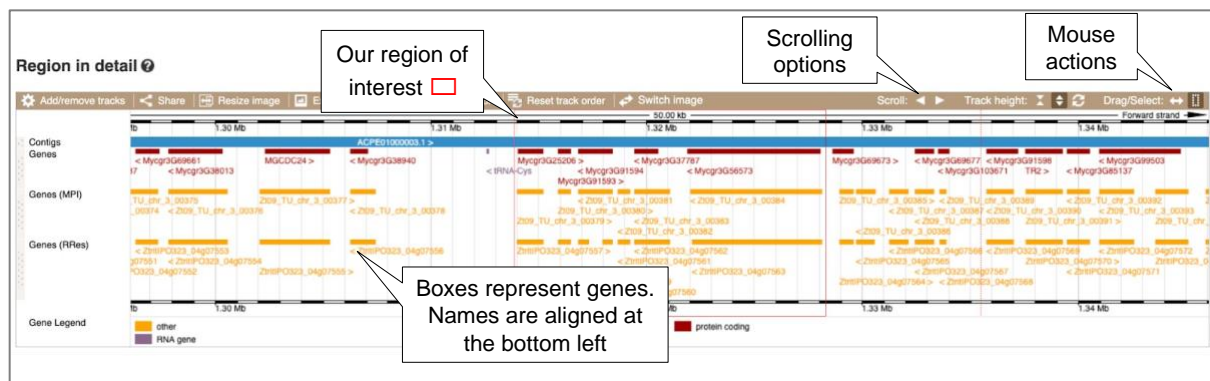
Click on the Drag/Select button to change the action of your mouse click. Changing to the arrow allows you to scroll along the chromosome by clicking and dragging within the image. As you do this, you'll see the image below grey out and two buttons appear. Clicking on Update this image will jump the lower image to the region you have selected above. We want to go back to where we started, so we'll click on Reset scrollable image.

The third image is the region in detail view. It is a detailed, configurable view of your selected region similar to this:
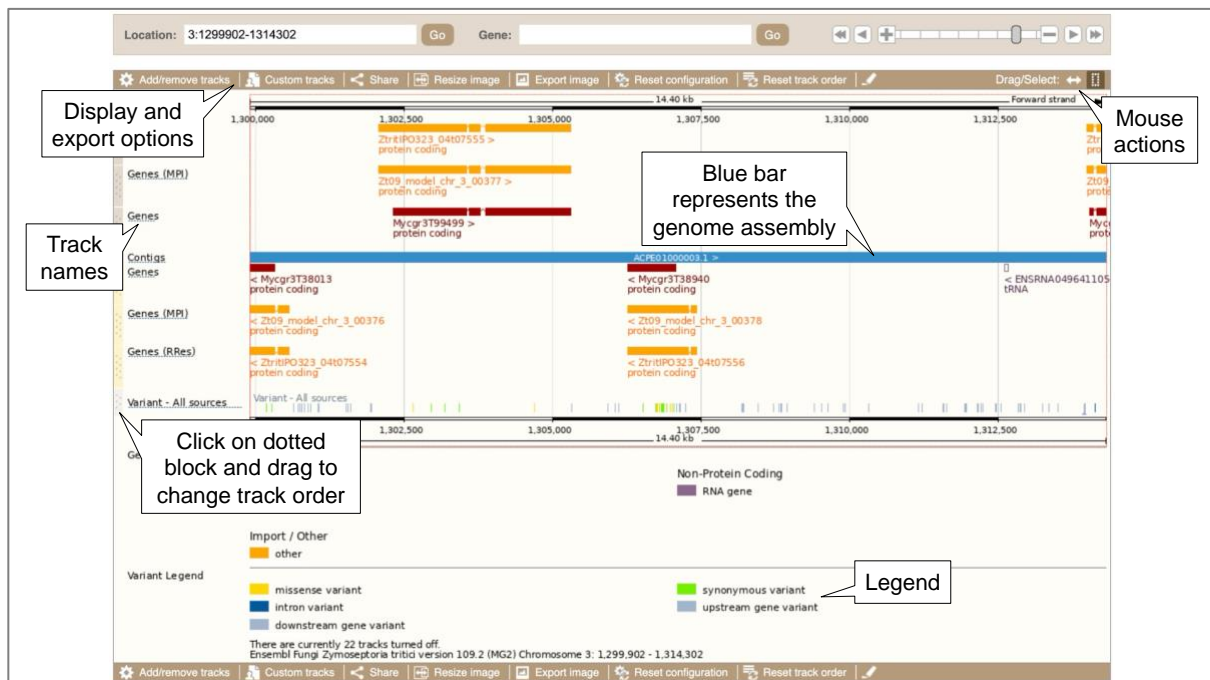
Genes are shown as transcripts with exons represented as boxes and introns shown as lines connecting the exons. Forward-stranded genes are shown above the genome assembly (Contigs track), while reverse-stranded genes are shown below.

Click on the Drag/Select option at the top or bottom right to switch your mouse action. On Drag, you can click and drag sideways to move along the genome, the page will reload when you drop the mouse button. The Select option allows you to drag out a box to highlight or zoom in on a region of interest.

Change your mouse action to Select, drag out a box around an exon and choose Mark region.
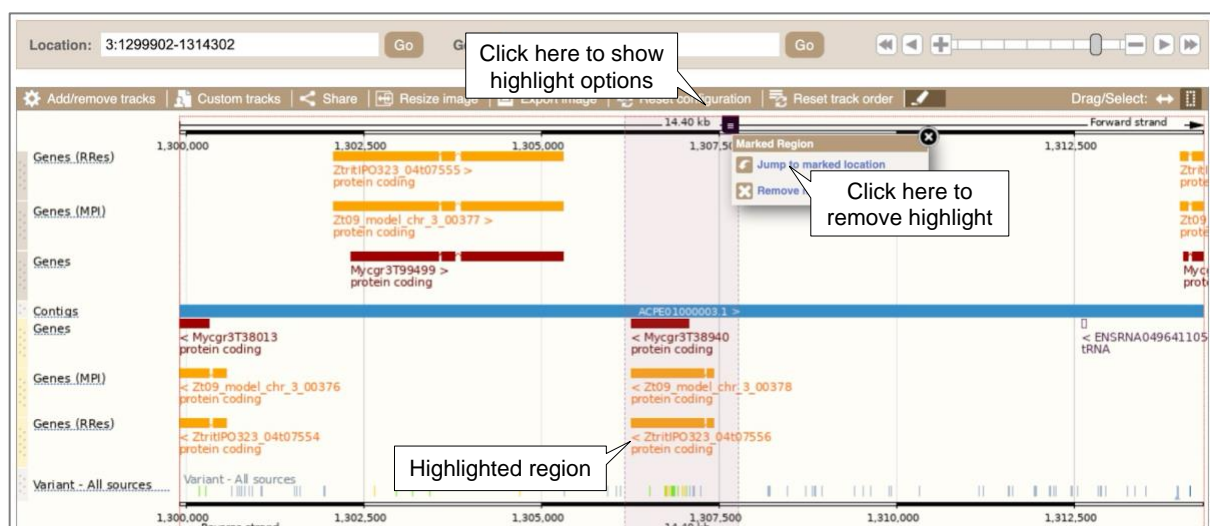
The highlight will remain in place if you zoom in and out or move around the region. This allows you to keep track of regions or features of interest.

(b) How many genes are annotated in the current region? How many are on the forward and how many are on the reverse strand? Are they all annotated by the same institute?
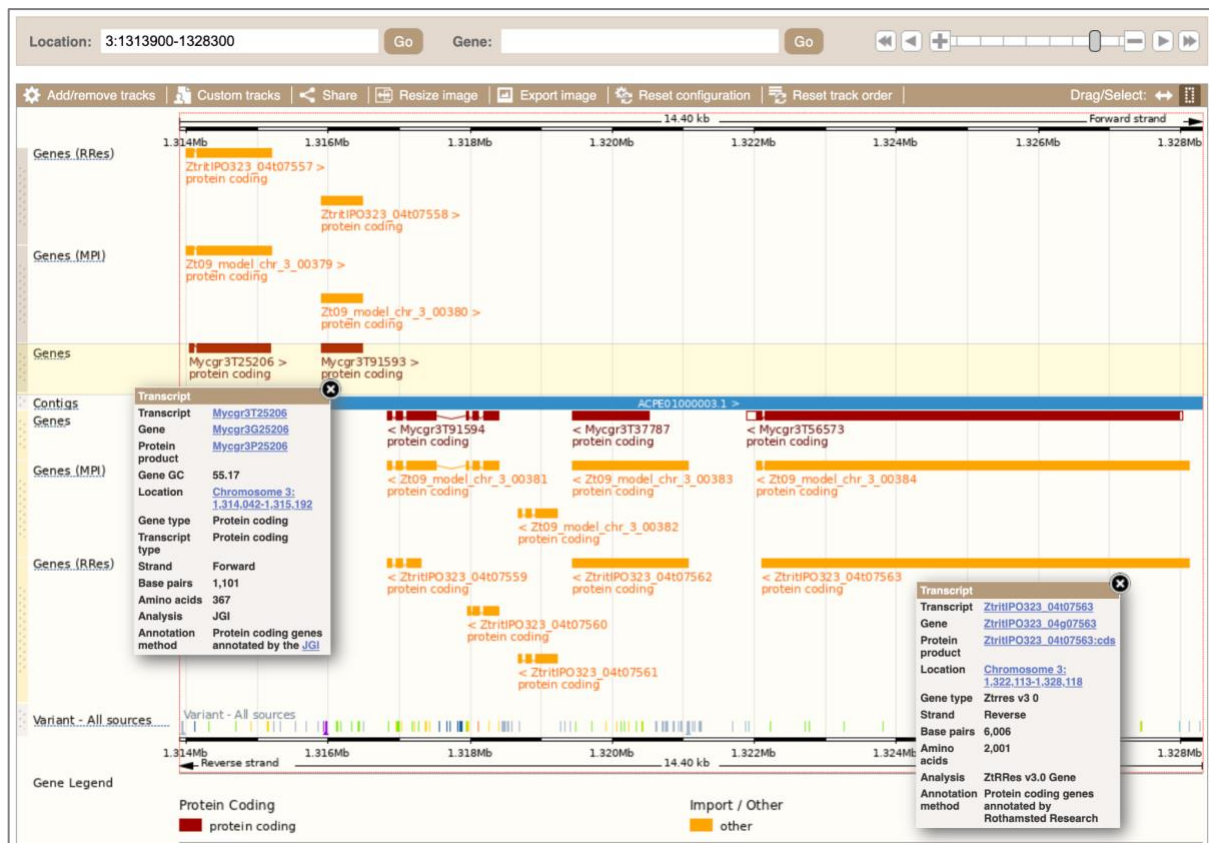
We can edit what we see on this page by clicking on the Configure this page button

⚙ Configure this page   located on the left-hand side.

This will open a menu that allows you to change the image. You can put some tracks on in different styles. You can read more details in this FAQ: https://www.ensembl.org/Help/Faq?id=335.

You can add a track to the image by clicking on the box to the left of the track name and selecting a track style from the pop-up menu. Click on the tick in the top right-hand corner to save and close the menu. Alternatively, click anywhere outside of the menu.

Let's add some tracks to this image.

(c) Turn on the All repeats track. Are there any repeat regions identified in this region? Do they overlap any of the genes?

We can also change the way the tracks appear by clicking on the track name to open a menu.

We can move tracks around by clicking and dragging on the coloured dotted block/bar to the left of the track name:

(d) Zoom in on the largest transcript Mycgr3T565573. How many exons does this gene have?
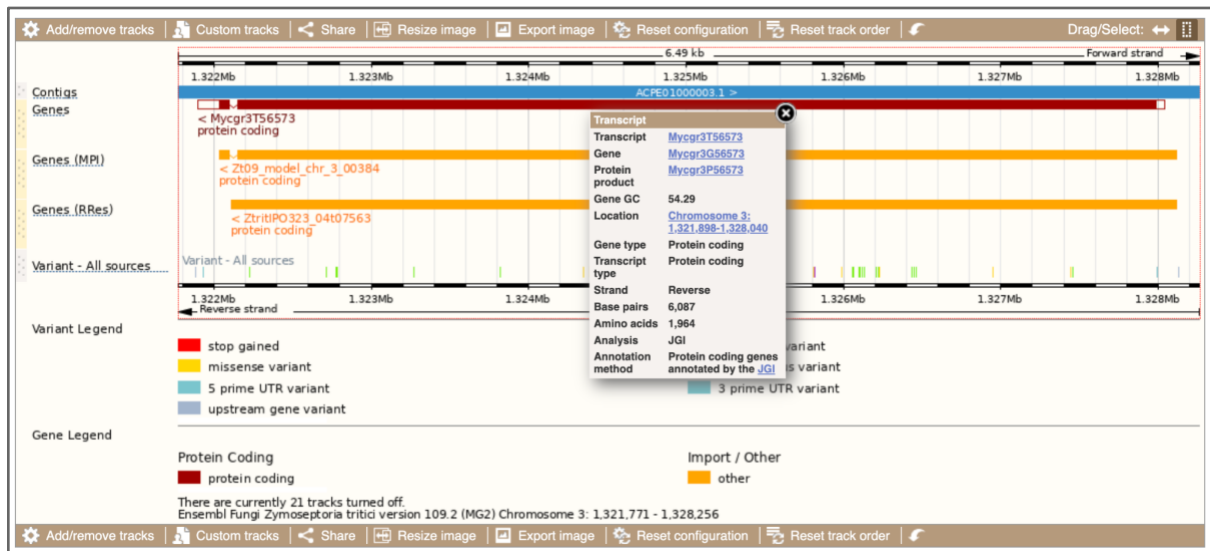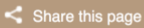

Now that you've got the view how you want it, you might like to show something you've found to a colleague or collaborator. Click on the Share this page button [Share this page] located either at the top of the image, or in the left-hand panel to generate a link.

Email the link to someone else, so that they can see the same view as you, including all the tracks you've added. These links contain the Ensembl release number, so if a new release or even assembly comes out, your link will just take you to the archive site for the release it was made on.

(d) Export the genomic sequence for this region by clicking the Export data button [Export data] located either at the top of the image or in the left-hand panel.

To return the region image to the default view, click on Reset configuration at the top of the region in detail image.

Exercise: Ensembl Fungi gene and transcript tabs

We're going to look at the gene *LEUC* in *Zymoseptoria tritici* (assembly MG2). This gene is involved in the leucine biosynthetic process.

From fungi.ensembl.org, type *LEUC* into the main search box, click the drop-down menu, select *Z. tritici* and click the Go button.

Click on the gene ID Mycgr3G103221 in the results. The **Gene tab** should open, similar to this:
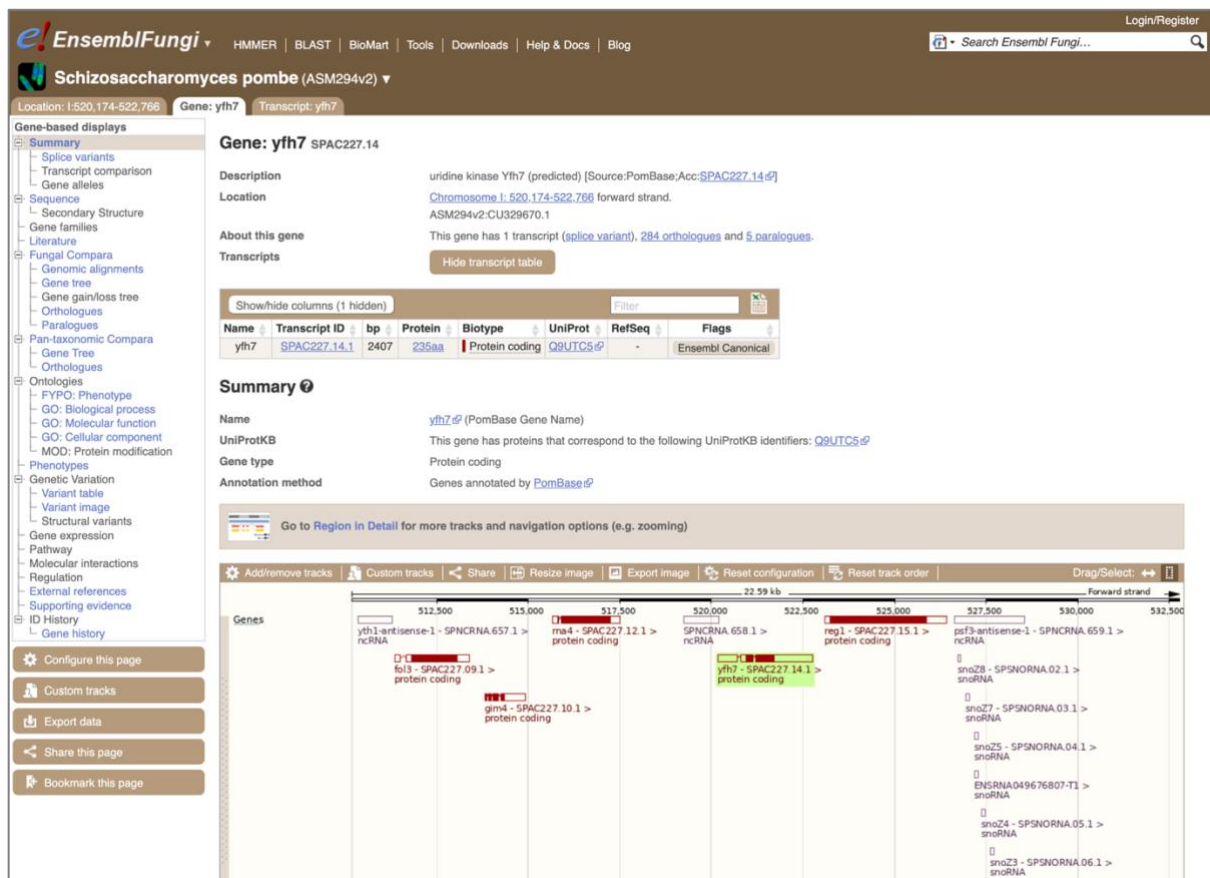
The *LEUC* gene is highlighted in green and in the centre of the display as it is the gene of interest.

(a) On which chromosome and which strand of the genome is this gene located?

Let's walk through some links in the left-hand navigation column. 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:

```
>chromosome:MG2:2:1580729:1583430:-1
```

The FASTA header follows this format:
Genome assembly:Chromosome:Base pair start coordinate:Base pair end coordinate:Strand
The forward strand denoted by 1, and the reverse strand by -1.

Exons are highlighted within the genomic sequence. If you click on Configure this page in the left-hand panel, you can change display options. For species with variation databases you can highlight variants on this view.

(b) Use the Configure this page option to show variants on the sequence and the line numbering relative to the coordinate system, to this view. Are all exons shown in this display part of the *LEUC* gene? How can you tell?

Can you find the Stop Gained mutation? What letter is it represented by? What nucleotides does it stand for? (these are [IUPAC ambiguity codes](#))

Which exon does the stop gained mutation fall in?

You can download this sequence by clicking on the Download sequence button located above the sequence.

This will open a pop-up menu that allows you to pick between plain FASTA sequence, or sequence in rich text format (RTF), which includes all the coloured annotations and can be opened in a word processor. This button is available for all sequence views.

(c) Export this sequence in RTF.

If we are interested in finding out about gene functions, the Gene Ontology (GO) annotations can tell us where the protein is located, the biological processes it is involved in and its molecular function. You can read more about GO terms here: https://geneontology.org/docs/ontology-documentation/.

(d) What biological processes have been associated with *LEUC*?

Click on GO: Biological process. This page shows all linked GO annotations, some of these are linked as GO terms are hierarchical. For example, if you click on the 'leucine biosynthetic process' GO accession number GO:0009098, you will be taken to the GO pages, which shows that this is a child term to 'cellular amino acid biosynthetic process' which is also shown on the GO pages in Ensembl.

GO: Biological process ❓

Accession	Term	Evidence	Annotation source	Mapped using	Transcript IDs	
GO:0006097	glyoxylate cycle	IEA	Ensembl	Propagated from Saccharomyces_cerevisiae YCL018W by orthology	Mycgr3T103221	• Search BioMart • View on karyotype
GO:0008652	cellular amino acid biosynthetic process	IEA	UniProt	UniProtKB/TrEMBL:F9X2A6_MYCGM	Mycgr3T103221	• Search BioMart • View on karyotype
GO:0009082	branched-chain amino acid biosynthetic process	IEA	UniProt	UniProtKB/TrEMBL:F9X2A6_MYCGM	Mycgr3T103221	• Search BioMart • View on karyotype
GO:0009098	leucine biosynthetic process	IEA		InterPro:Isopropylmalate_DH, UniProtKB/TrEMBL:F9X2A6_MYCGM	Mycgr3T103221	• Search BioMart • View on karyotype

Inferred from Electronic Annotation

Let's explore the **Transcript** tab now. Many genes have multiple transcripts which can be seen in the transcript table. Click on the Show transcript table button Show transcript table .

We can go to the **Transcript** tab either by clicking on the transcript ID Mycgr3T103221 in the table, or on the **Transcript** tab at the top of the page.

Gene: LEUC Mycgr3G103221

Description	3-isopropylmalate dehydrogenase [Source:UniProtKB/Swiss-Prot;Acc:Q9Y897]
Location	Chromosome 2: 1,581,329-1,582,830 reverse strand. MG2:ACPE01000002.1
About this gene	This gene has 1 transcript (splice variant), 334 orthologues and 7 paralogues.
Transcripts	Show transcript table

Show/hide columns (1 hidden)

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T103221	1237	365aa	▌Protein coding	F9X2A6	-	Ensembl Canonical

You are now in the **Transcript** tab on the summary page. Some summary information about the number of exons, length, etc. is shown at the bottom of the page under the summary diagram.

(e) How many exons does this transcript have? Which one is the longest?

The left-hand navigation column provides several options for the transcript. Click on the Exons link.

You may want to change the display (for example, to show more flanking sequences, or to show full introns). In order to do so, click on Configure this page and change the display options accordingly.

Now click on the cDNA link in the navigation column on the left to see the spliced transcript sequence.


UnTranslated Regions (UTRs) are highlighted in dark yellow, codons are indicated by alternating light yellow highlights, and exon sequences are shown in alternating black and blue letters.

We can look at the protein sequence in more detail, finding domains and structural information. Click on Protein summary to view domains from SignalP, Pfam, PROSITE, Superfamily, InterPro, and more.

(f) What domains can be found in the protein product of this transcript? How many different domain prediction methods agree with each of these domains?

Can you see the stop gained mutation we saw in exercise (b) here?

Will this variant cause the deletion of an entire protein domain? If so, which one(s)?
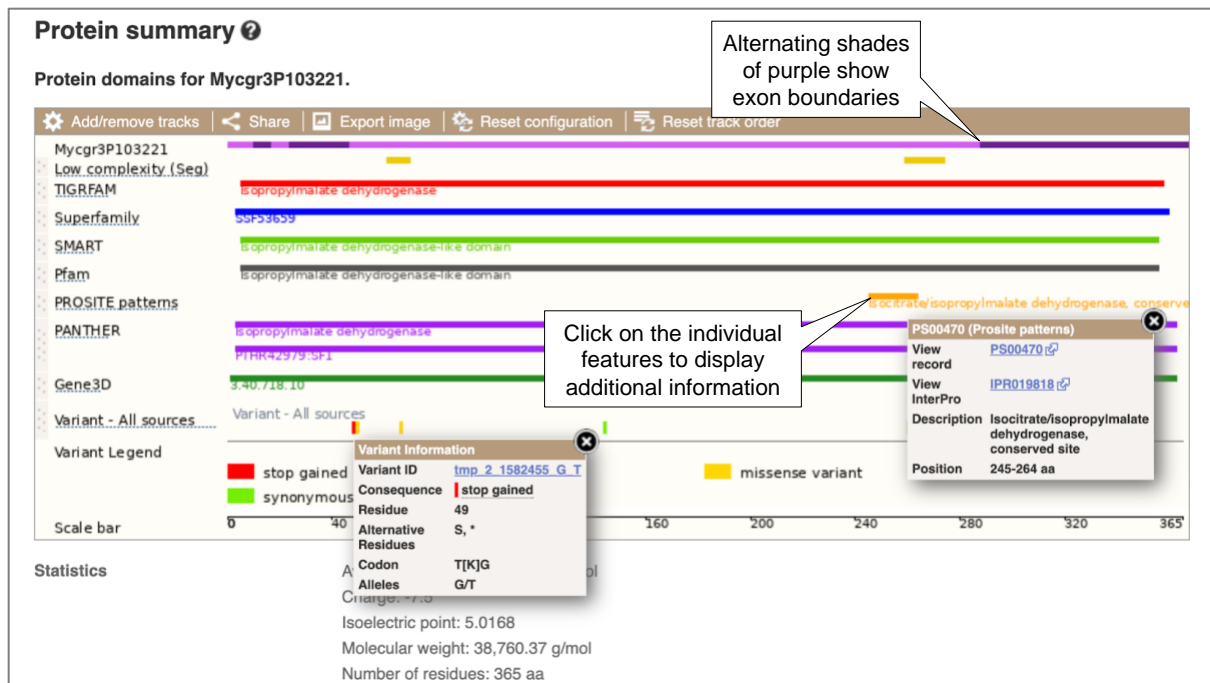
Clicking on Domains & features shows a table of this information.

Next, follow the General identifiers link in the navigation column on the left.

This page shows information from other databases such as ENA, UniProtKB, INSDC and others, that match to the Ensembl transcript and protein.