# **Practical I – Accessing and downloading data from *Ensembl* Plants using BioMart**

The aim of this session is to learn how to obtain sequence data from *Ensembl* Plants using the BioMart tool.

## **Background information for the exercise**

You have mapped a QTL for grain length to a distinct genetic interval on chromosome 6A. You now want to understand how many possible candidate genes are located within this interval. If the number is not too large, you would like to get some additional information about the genes and their possible functions.

## **Practical exercise**

**Download data from BioMart**

* Navigate to the *Ensembl* Plants website (<http://plants.ensembl.org/index.html>) and select “*Triticum aestivum*” as your species.
* Now type the following coordinates into the search box:

6A:549,000,000-551,000,000

* You should now see the genetic interval on chromosome 6A, from 549 Mbp to 551 Mbp, displayed. All gene models falling within the interval are shown. Click on the first gene model from the left (*TraesCS6A02G312300*). A little text box will open. Right-click on the gene name in the text box and open in a new tab.
* You now see the “Gene page”, which lets you access more information about the gene model. Let us explore the gene’s homoeologs first. Click on the “Gene Tree” on the left-hand side. Click on the node to expand all “Pooidae” sub-nodes. Your original gene (*TraesCS6A02G312300*) is shown in red, while its homoeologs are shown in blue. Note down the names of the homoeologs.
* Next, we want to export the sequence of *TraesCS6A02G312300*. Click on “Summary” on the left-hand side (all the way at the top). Now you can use the “Export data” button at the bottom left-hand side. Set the 5’ and 3’ flanking sequence to 1000 bp each. Deselect all annotations (cDNA, coding sequence, etc), then click “Next” and open as “Text”. Save the sequence to a text file.
* We could repeat the same steps for the remaining 20 genes in the QTL interval, but there is a faster way!
* Navigate to the *Ensembl* Plants website (<http://plants.ensembl.org/index.html>) and click the “BioMart” button at the top of the page. Select “*Ensembl* Plants Genes 56” as your database and “*Triticum aestivum* genes (IWGSC)” as your dataset.
* Use the following filters and attributes to find genes in our QTL interval. Note that BioMart **does not** accept commas in numbers (unlike the main *Ensembl* Plants interface).

Filters:

Region:

Chromosome: 6A

Start Coordinate: 549000000

End Coordinate: 551000000

Gene:

Source (gene): IWGSC

Attributes:

Gene:

Gene stable ID

* How many “Results” did you get? It should be a list of 21 genes. By default, only the first 10 entries are shown. You can view all of them in HTML or download to a text file.
* Go back to BioMart and select “Sequences” as an attribute instead of “Features” to download the “unspliced gene” sequence including 1000bp of sequence upstream and downstream of each gene. You have now downloaded the genomic sequence of 21 genes including 1000 bp up- and downstream of the coding region.
* Go back to BioMart and select “Homologues” as an attribute.

Open the “Hom**oeo**logs” drop down menu and select “*Triticum aestivum* homoeologue gene stable ID” to identify the homoeologs for the genes in our interval. Does it match your expectations with regards to polyploid wheat (i.e. does every gene have two homoeologs)?

* Deselect “Hom**oeo**logs” and instead open the “ORTHOLOGUES [P-T]:” drop down menu and select “*Triticum aestivum* Lancer gene stable ID” as well as “*Triticum aestivum* Mace gene stable ID”. You have now downloaded the orthologous genes of two pangenome cultivars for our QTL.
* Go back and open the “ORTHOLOGUES [K-O]:” drop down menu and select “*Oryza sativa* Japonica Group gene stable ID”. Does it match your expectations (i.e. does every gene have an ortholog in rice)?
* If you have time, you can inspect the rice orthologs using the funRiceGenes search engine (<https://funricegenes.github.io/>). Enter your rice gene ID into the search box and click search. If there is information available about a gene, the search results should include a specific gene name (e.g. OsDRP2B). Click this link to get a summary of the publications linked to this gene. There is generally more information available in rice about genes and their function. Some of this knowledge might also be applicable to wheat and help you design hypotheses and experiments. The only downside is that this is approach is not easily up-scalable: it works well for a small number of genes but is impractical for dozens/hundreds of genes.