Updated November 2018

Finding the wheat orthologs of genes from other species

This tutorial document explains the process of identifying the wheat homologs of genes from other model or non-model plant species. This is especially useful in translational research or comparative genomic studies where the gene of interest might have been well characterised in a model species (e.g. *Arabidopsis thaliana*) and you want to characterise the function of the wheat orthologs.

a) Important considerations

It is important to note that the genetic control of traits can vary in plant species. As such, the function of a gene in one plant species may not be conserved in another. Also, due to gene loss or duplication events, gene family size and/or gene copy number can vary between species and indeed genes present in one species may not be present in another.

It is also important to consider the ploidy level and homoeologous relationship of the wheat species you are interested in (see "Introduction to Wheat"). For instance, for a single *Arabidopsis* gene, you should expect to find one, two or three gene models in diploid (2n; einkorn), tetraploid (2x 2n: durum wheat) or hexaploid (3 x 2n; bread) wheat, respectively. In this tutorial we will focus on hexaploid wheat (*Triticum aestivum*) only.

b) Finding wheat orthologs through Ensembl Plants

Ensembl Plants (http://plants.ensembl.org) hosts the genomes of many sequenced model and non-model plant species. This makes Ensembl Plants a convenient portal to query and compare different plant genomes (see "Ensembl Plants primer" for a quick introduction on how to use Ensembl Plants). To find wheat genes orthologous to genes from *Arabidopsis*, we will be using the gene tree and ortholog features of Ensembl Plants. However, other genome database portals can be used if required (e.g. TAIR, URGI, CerealsDB, Phytozome, or NCBI). Here, we will use the *Arabidopsis* gene *HsfB1* as a case study for how to find wheat orthologs.

1. To get started, visit the Ensembl Plants website at http://plants.ensembl.org.
2. Find the Ensembl Plants gene page for your query gene of interest (e.g. *Arabidopsis HsfB1*). There are two ways to do this:
    a. In well annotated genomes, such as *Arabidopsis*, you could search directly for the common gene name (e.g. *HsfB1*) or the gene identifier (e.g. *AT4G36990*) using the search box (Fig. 1a). Alternatively, you could select your model species in the drop-down list (Fig. 1b) or go to the main page for your species (Fig. 1c) before searching, to limit the search results to just one species.
    b. If it is not possible to search using the gene name or identifier, you can also BLAST your sequence to find the correct gene on Ensembl Plants by clicking on BLAST at the top of the homepage (Fig. 1d) and then clicking on "New job". From here you can conduct a BLAST search against your starting species of interest to find the gene ID used by Ensembl Plants.

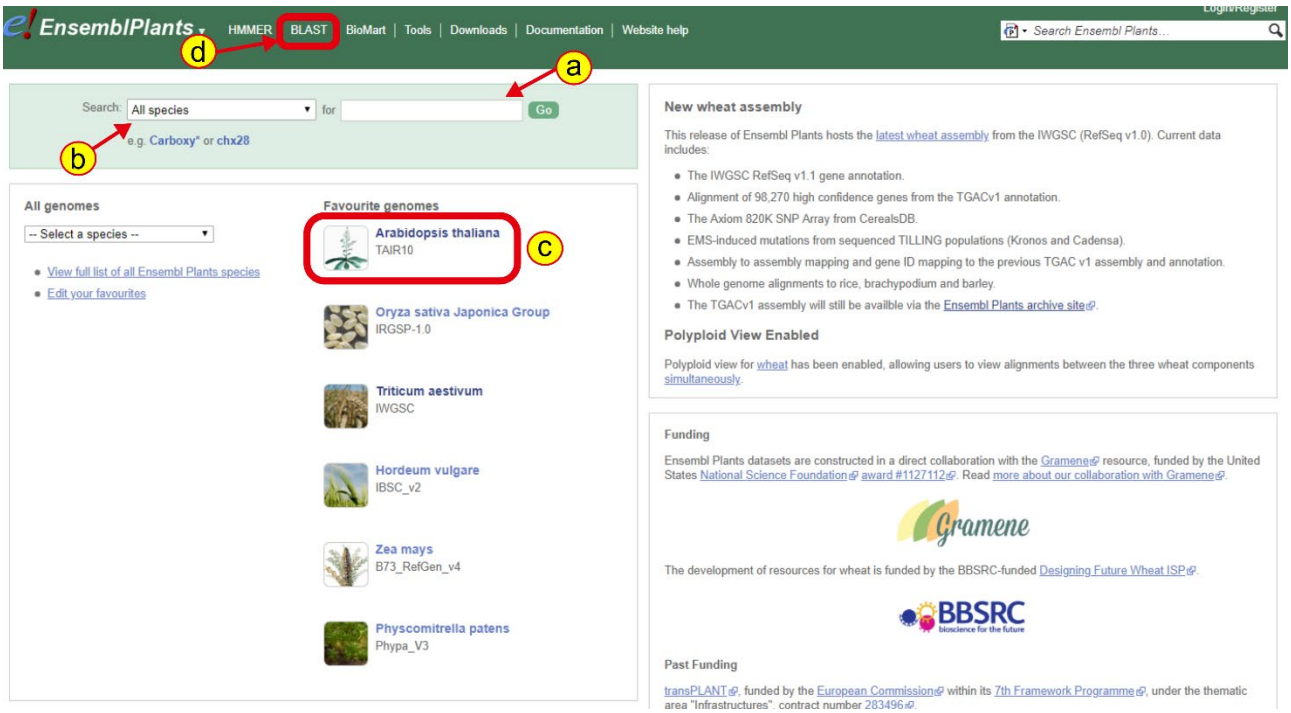


Figure 1: Searching for a gene on Ensembl Plants

3. Once you have reached the summary page for your gene of interest, you will see that there are four tabs: Location, Gene, Transcript and Jobs. Make sure you are on the Gene tab (Fig. 2a). Here we can see that the structure of our gene (Fig 2b), and other genes in the region. To find the wheat ortholog(s), click on the link to the gene tree on the left-hand side of the page (Fig. 2c).

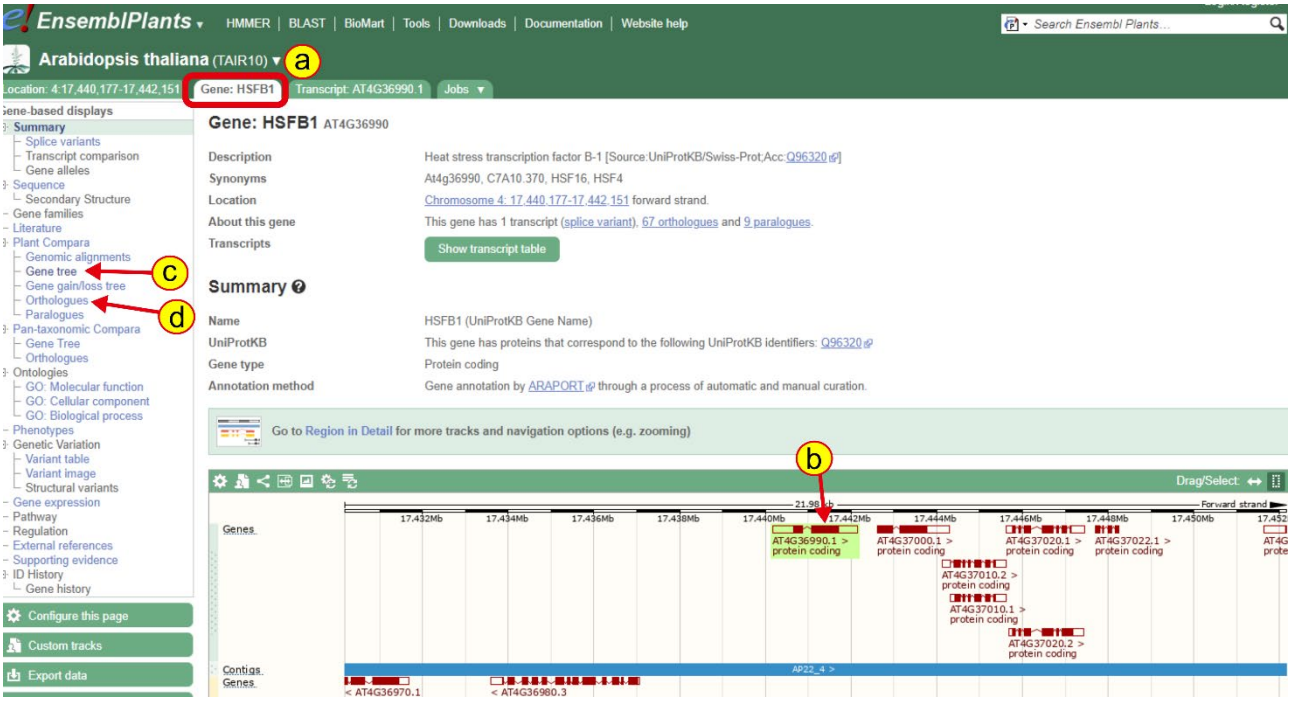


Figure 2: Gene view on Ensembl Plants

4. The gene tree shows a phylogenetic tree of the homologs of your gene of interest across the different plant genomes that are hosted on Ensembl Plants. A protein alignment is also shown on the right which is useful for looking at the conservation of gene structure. Your gene of interest will be highlighted in red (Fig 3a). The subtrees are grouped based on taxonomic rank and you can expand individual sub clades (Fig 3b) or the whole tree (Fig 3c) based on your requirements. When expanding the whole tree, you may find other genes highlighted in blue – these are considered as paralogs of your gene of interest (i.e. homologs of the gene in the same species).

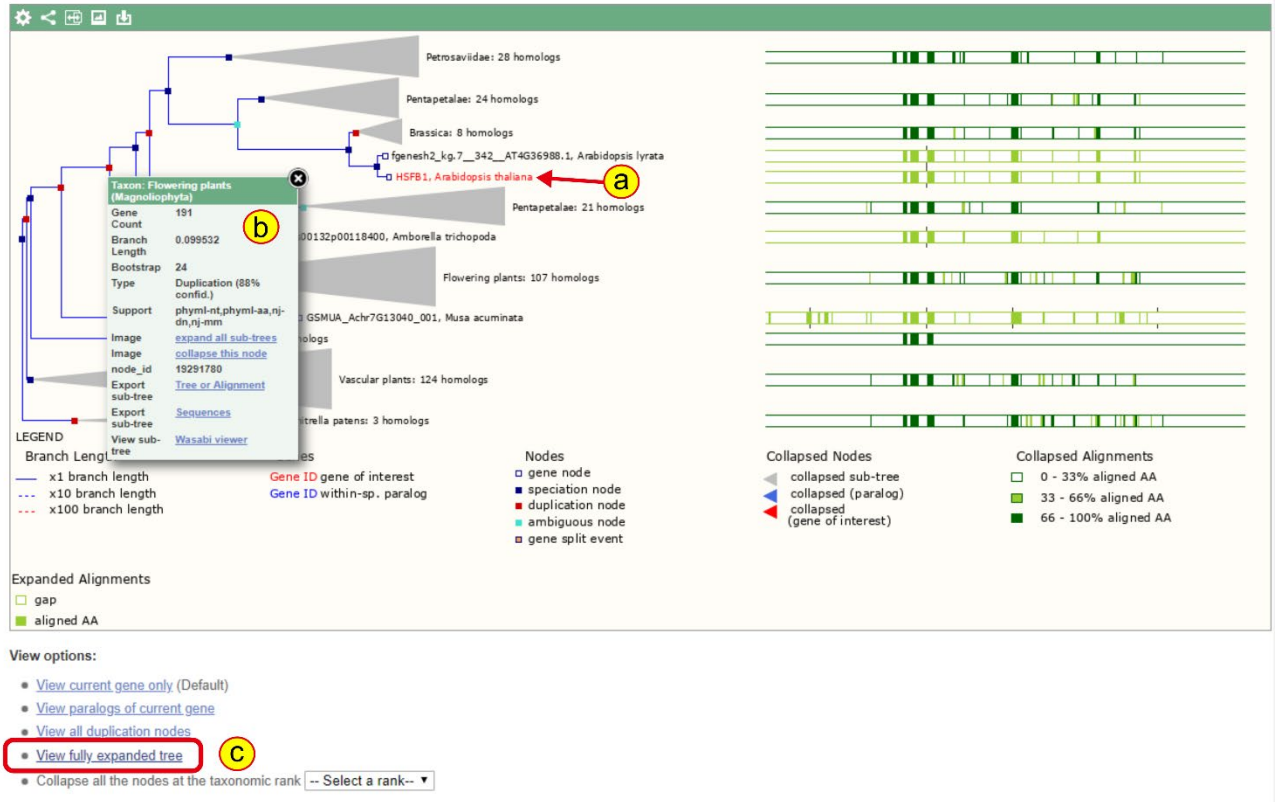


Figure 3: Ensembl Plants gene tree

5. To find the closest wheat orthologs of your gene of interest, work up the tree starting with your gene, expanding sub-nodes until you find a *Triticum aestivum* (hexaploid wheat) gene (Fig 4a).
6. In the case of *HsfB1*, we can see that there is a single group of wheat genes in the same clade as *HsfB1* in the canonical group of three (A, B, D) homoeologs: *TraesCS5A02G237900*, *TraesCS5B02G236400* and *TraesCS5D02G244800*, respectively.
    a. These are the IWGSC RefSeqv1.1 gene model IDs, see "Gene models" for more information on different wheat gene model IDs.
7. We can see that there are also other species in the sub clade containing the *T. aestivum* genes, including wild relatives and progenitor species. *Triticum Urartu* is the wheat A genome progenitor, and usually you would expect one gene from *T. urartu* that groups with the A genome homoeolog. Similarly, *Aegilops tauschii* is the D genome progenitor and so you would expect one gene that groups with the D genome homoeolog. *Triticum dicoccoides* is a tetraploid progenitor species of hexaploid wheat containing the A and B genome. So, you would expect two copies in *T. dicoccoides*, one that groups with the *T. aestivum* A homoeolog

and the other grouping with the *T. aestivum* B homoeolog. In some cases, the structure of the clade may be different due to gene duplication and loss events.

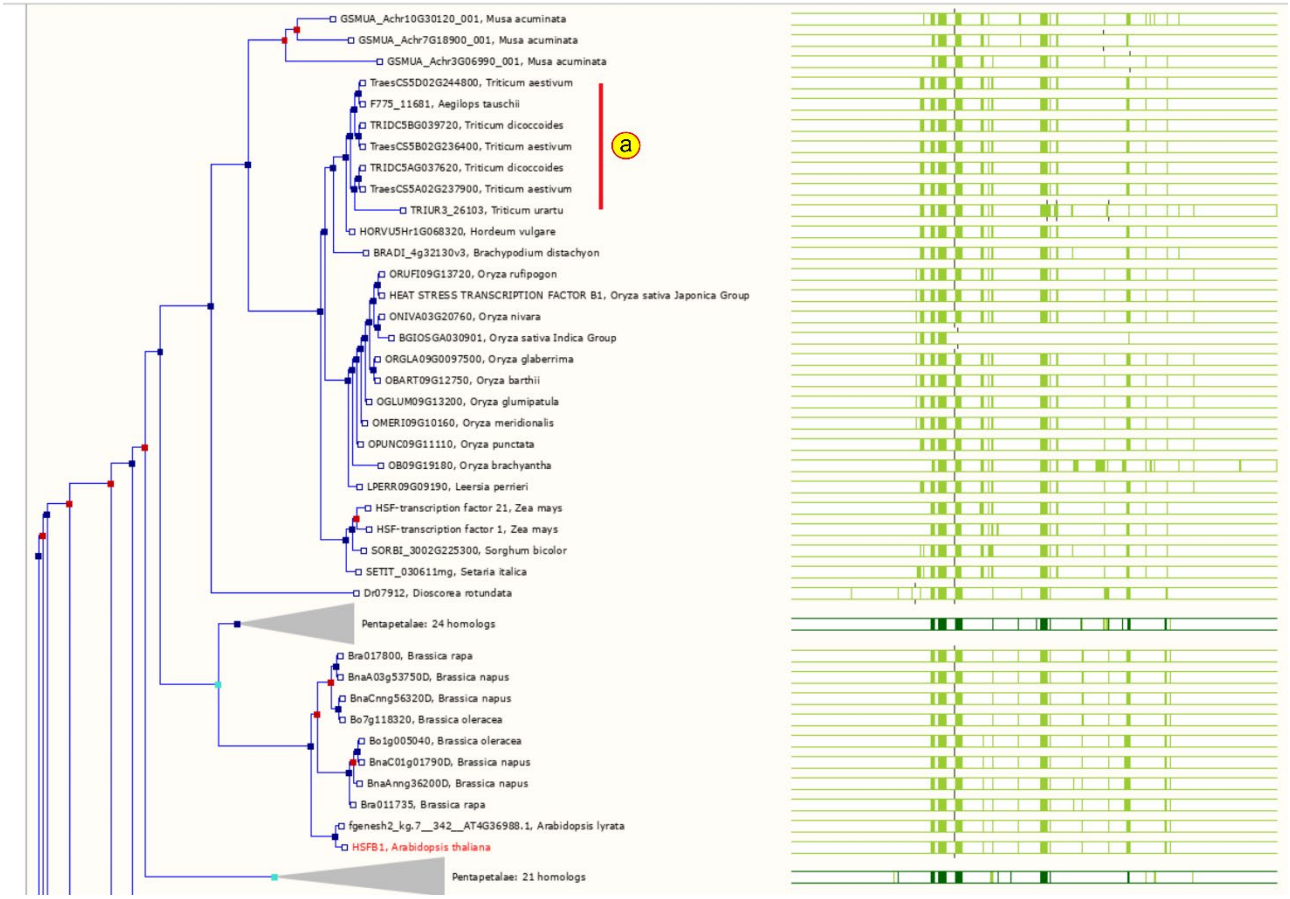


Figure 4: Expanded clade in Ensembl Plants gene tree

8. You can also find orthologs by clicking on the orthologs link on the left-hand side of the gene tab of your gene of interest (Fig 2d) but using the gene tree will give you more insight into the relationship between different species.
9. If we click on the A genome copy of our wheat orthologs, we can go to the gene tab of this gene to explore it further. We can see that this ortholog has two exons, just as the *Arabidopsis* gene did. For example, we can access the gDNA, cDNA, CDS and peptide sequences of the gene by clicking export data on the left-hand side of the page (Fig. 5A). On the pop up click “Next” and then “html” to access the FASTA sequences.

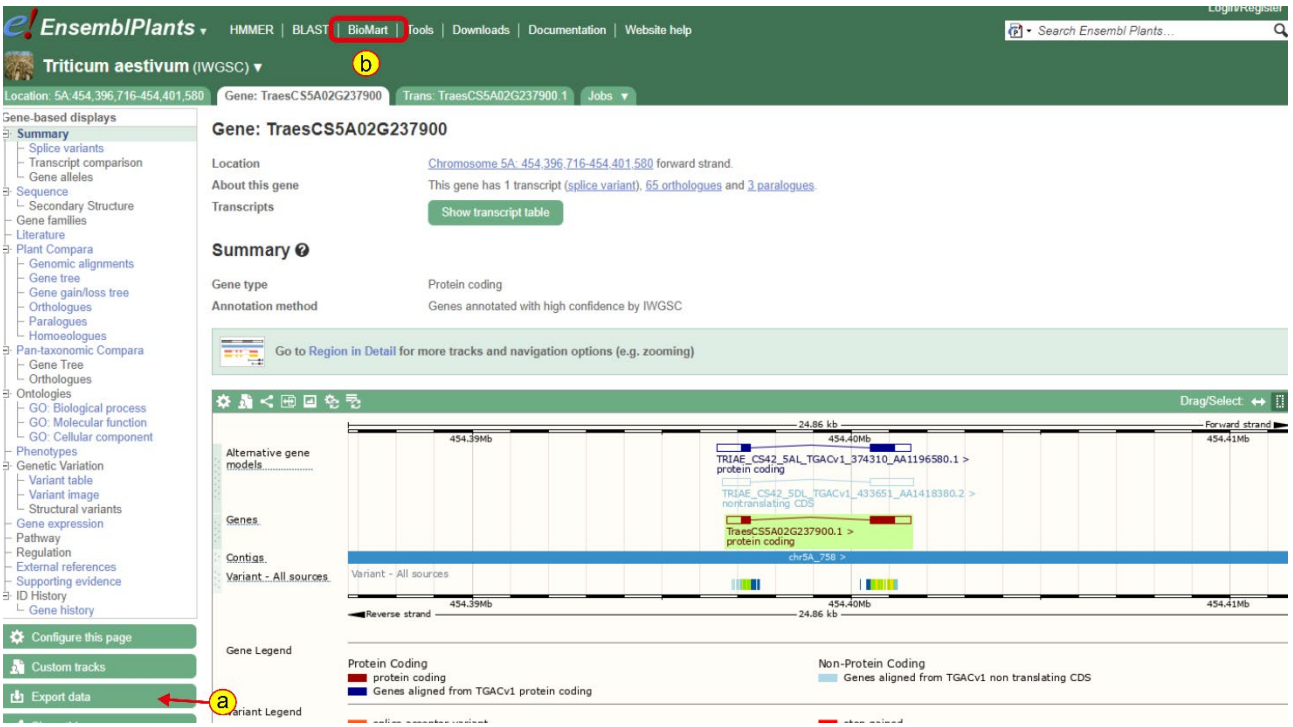


Figure 5: Gene view of one of the wheat orthologs

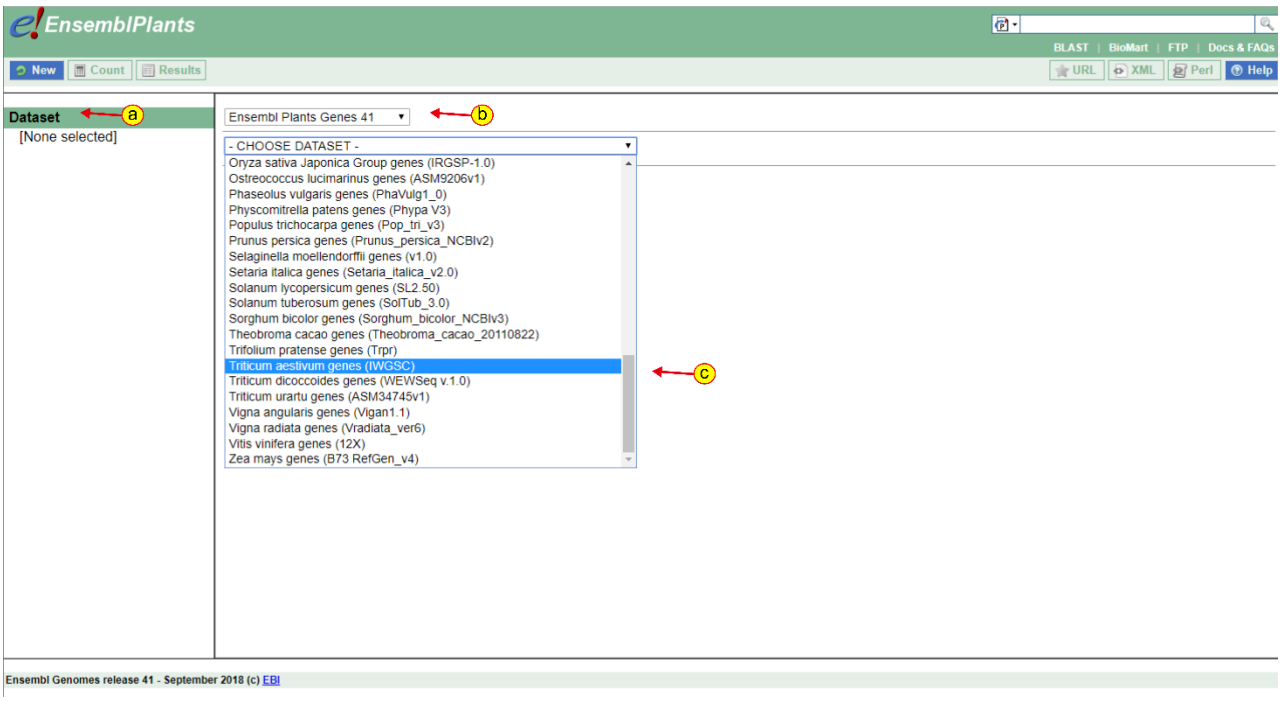
10. To obtain sequences of multiple genes, e.g. all three homoeologs at once, we can use BioMart. Click on the link to BioMart at the top of the page (Fig. 5b).
11. Click on Dataset (Fig. 6a), then select “Ensembl Plants Genes 41” (Fig 6b) and finally select “*Triticum aestivum* genes (IWGSC)” (Fig. 6c).
12. 

Figure 6: Biomart Step 1

13. Then click on “Filter” (Fig 7a), go into the “Gene” section (Fig 7b) and tick the box “Input external references ID list” (Fig 7c) and enter your wheat gene IDs into the box.

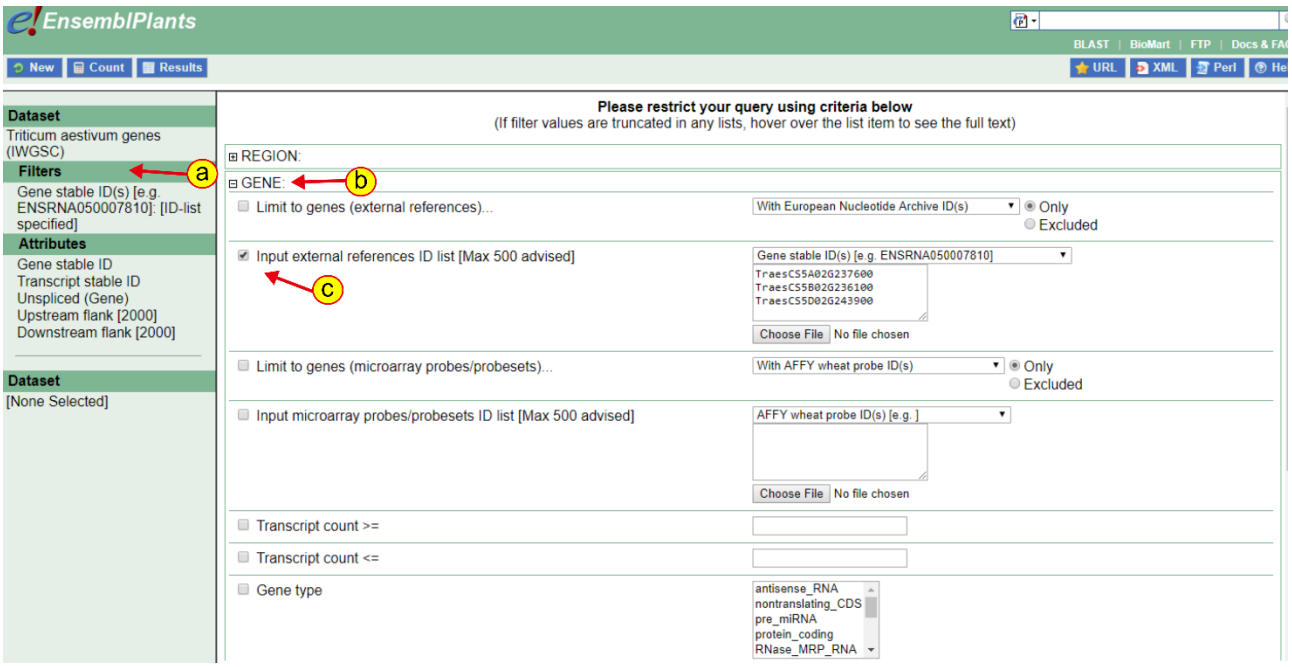


Figure 7: Biomart step 2

14. Then go to “Attributes” (Fig 8a) and select “Sequences” (Fig 8b). You can choose what type of sequence you would like. In the example in Figure 8, we ask for the “Unspliced (Gene)” sequence i.e. genomic DNA and also 2000 bp upstream so we can look at the promoter sequence.

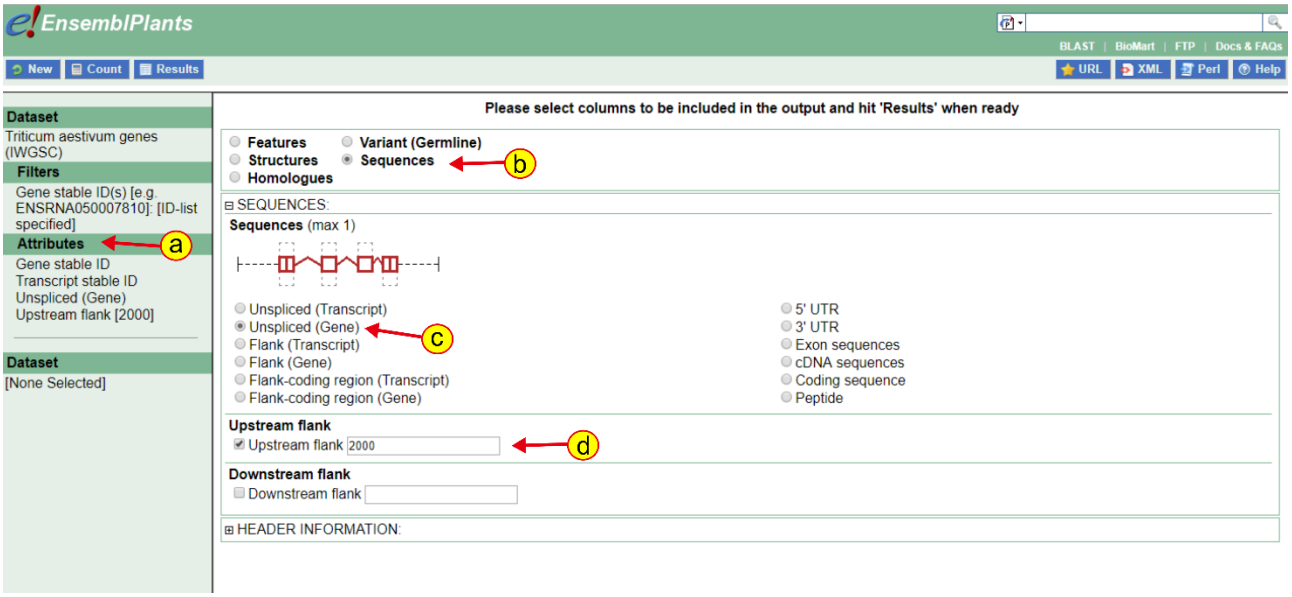


Figure 8: Biomart step 3

15. Finally, click on results (Fig 9a) to get the sequences. You can download the sequences in FASTA format (Fig 9b).

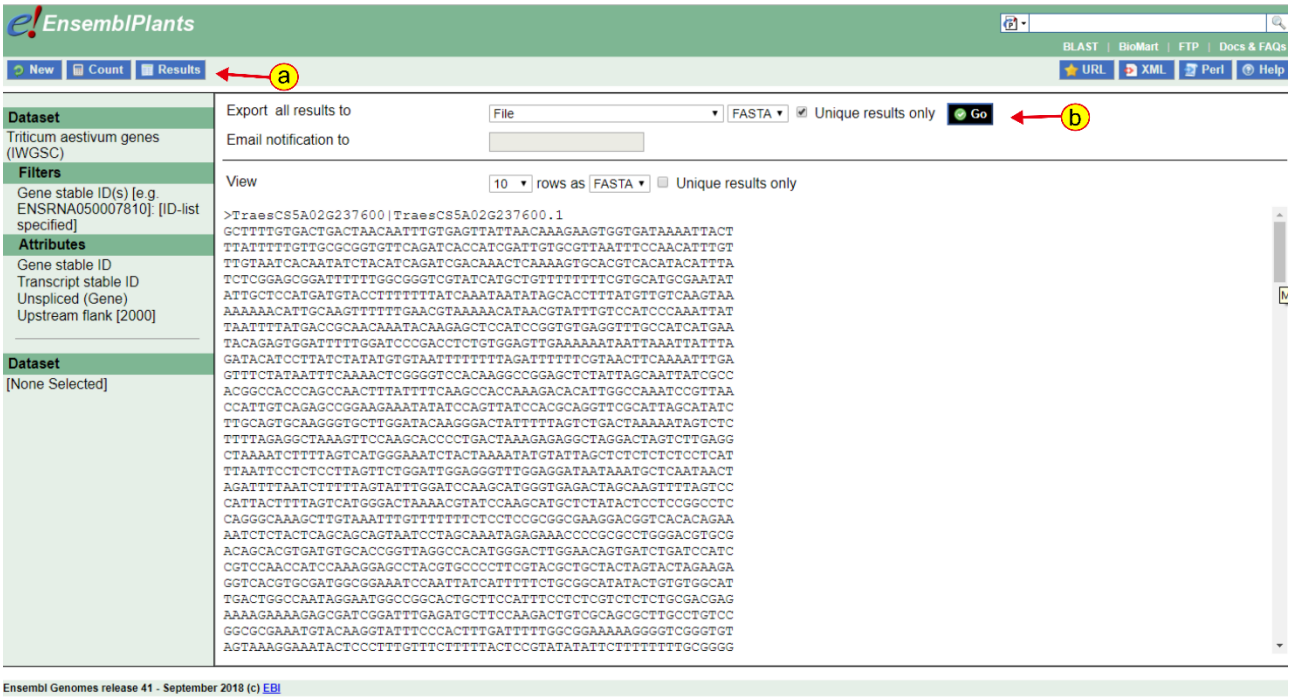


Figure 9: Biomart step 4

c) Obtaining sequences from other wheat varieties

The wheat genome assembly hosted on Ensembl Plants is IWGSC RefSeqv1.0, which is derived from the wheat landrace Chinese Spring (see "Genome assemblies" for more information). However, genome assemblies for other wheat varieties/cultivars have also been generated and are publicly available. It can be useful to BLAST your wheat gene against these other genome sequences to see if you have any inter-cultivar variation in your gene sequence. In some cases, there may even be presence/absence variation between varieties at the gene level. Below is a summary of some of the additional wheat genomes available and links to the corresponding databases:

Table 1: Currently available wheat genome assemblies for varieties different to the reference Chinese Spring landrace.

Variety	Habit	Origin	Availability *
Hexaploid bread wheat			
CDC Landmark	spring	Canada	10+ Genome Project
CDC Stanley	spring	Canada	10+ Genome Project
Paragon	spring	UK	10+ Genome Project
Cadenza	spring	UK	10+ Genome Project
Lancer	spring	Australia	10+ Genome Project
Mace	spring	Australia	10+ Genome Project
Synthetic W7984	spring	Mexico	Chapman *et al.* (2015)
Weebil	spring	Mexico	10+ Genome Project
Arina*LrFor*	winter	Switzerland	10+ Genome Project
Julius	winter	Germany	10+ Genome Project

Jagger	winter	US	10+ Genome Project
Robigus	winter	UK	10+ Genome Project
Claire	winter	UK	10+ Genome Project
Norin61	winter	Japan	10+ Genome Project
SY Mattis	winter	France	10+ Genome Project
Spelt (PI190962)	winter	Europe	10+ Genome Project
Tetraploid pasta wheat			
Zavitan†	-	Israel	Avni *et al.* (2017)
Svevo	spring	Italy	Maccaferri *et al.* (2019)
Kronos	spring	US	10+ Genome Project

† Zavitan is a tetraploid wild emmer (*T. dicoccoides*) accession
* Varieties included within the 10+ Wheat Genomes Project can be accessed through the Earlham Grassroot Genomics portal (https://wheatis.tgac.ac.uk/grassroots-portal/blast) and the 10+ Wheat Genomes project portal (http://webblast.ipk-gatersleben.de/wheat_ten_genomes) (subset of varieties in each). The 'Svevo' genome can be accessed through https://www.interomics.eu/durum-wheat-genome and Ensembl Plants. 'Synthetic W7984' and 'Zavitan' can be accessed through the Grassroot Genomics, and Ensembl Plants, respectively.

References:

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93-97.

Chapman JA, Mascher M, Buluc A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L, Schmutz J, Yelick KA, Scholz U, Waugh R, Poland JA, Muehlbauer GJ, Stein N, Rokhsar DS. 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol* **16**, 26.

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Gnocchi M, Moscatelli M, Avni R, Deek J, Biyiklioglu S, Frascaroli E, Corneti S, Salvi S, Sonnante G, Desiderio F, Mare C, Crosatti C, Mica E, Ozkan H, Kilian B, De Vita P, Marone D, Joukhadar R, Mazzucotelli E, Nigro D, Gadaleta A, Chao S, Faris JD, Melo ATO, Pumphrey M, Pecchioni N, Milanesi L, Wiebe K, Ens J, MacLachlan RP, Clarke JM, Sharpe AG, Koh CS, Liang KYH, Taylor GJ, Knox R, Budak H, Mastrangelo AM, Xu SS, Stein N, Hale I, Distelfeld A, Hayden MJ, Tuberosa R, Walkowiak S, Mayer KFX, Ceriotti A, Pozniak CJ, Cattivelli L. 2019. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* **51**, 885-895.