Finding the wheat orthologs of genes from other species

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

a) Important considerations

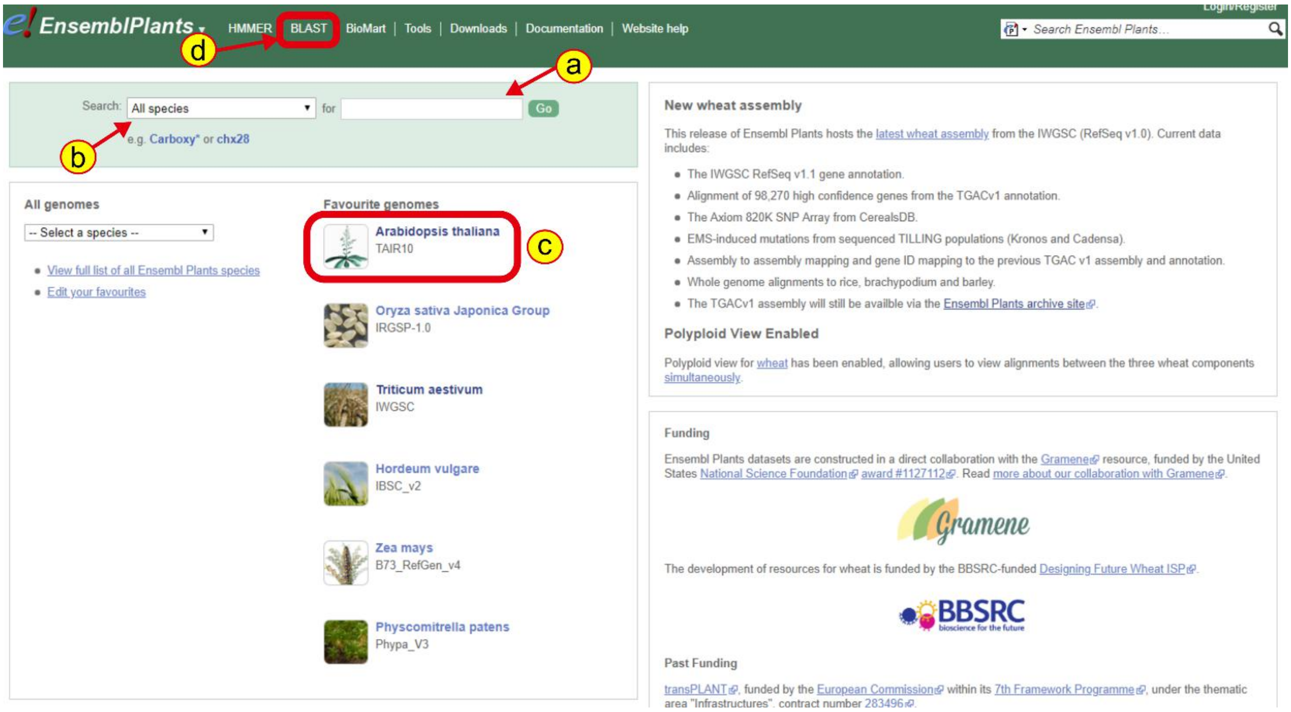
It is important to note that the genetic control of traits can vary in plant species. As such, the function of a gene in one plant species may not be conserved in another. Also, due to gene loss or duplication events, gene family size and/or gene copy number can also vary between species and indeed genes 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 the diploid (2n; einkorn), tetraploid (2x 2n: durum wheat) or hexaploid (3 x 2n; bread) wheat, respectively. In this tutorial, however, we will focus on hexaploid wheat (*Triticum aestivum*).

b) Finding wheat orthologs through *Ensembl*Plants

*Ensembl*Plants (http://plants.ensembl.org) hosts the genomes of most sequenced model and non-model plant species. This makes *Ensembl*Plants a convenient portal to compare between different plant genomes (see "Ensembl plants primer" for a quick introduction on how to use *Ensembl*Plants). We will be using the gene tree and orthologue features of *Ensembl*Plants. However, other genome database portal 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 EnsemblPlants 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 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 gene page 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.



EnsemblPlants HMMER BLAST BioMart | Tools | Downloads | Documentation | Website help

Search Ensembl Plants…

Login/Register

Search: All species for Go

e.g. Carboxy* or chx28

All genomes

-- Select a species --

- View full list of all Ensembl Plants species
- Edit your favourites

Favourite genomes

Arabidopsis thaliana
TAIR10

Oryza sativa Japonica Group
IRGSP-1.0

Triticum aestivum
IWGSC

Hordeum vulgare
IBSC_v2

Zea mays
B73_RefGen_v4

Physcomitrella patens
Phypa_V3

New wheat assembly

This release of Ensembl Plants hosts the latest wheat assembly from the IWGSC (RefSeq v1.0). Current data includes:

- The IWGSC RefSeq v1.1 gene annotation.
- Alignment of 98,270 high confidence genes from the TGACv1 annotation.
- The Axiom 820K SNP Array from CerealsDB.
- EMS-induced mutations from sequenced TILLING populations (Kronos and Cadensa).
- Assembly to assembly mapping and gene ID mapping to the previous TGAC v1 assembly and annotation.
- Whole genome alignments to rice, brachypodium and barley.
- The TGACv1 assembly will still be availble via the Ensembl Plants archive site.

Polyploid View Enabled

Polyploid view for wheat has been enabled, allowing users to view alignments between the three wheat components simultaneously.

Funding

Ensembl Plants datasets are constructed in a direct collaboration with the Gramene resource, funded by the United States National Science Foundation award #1127112. Read more about our collaboration with Gramene.

Gramene

The development of resources for wheat is funded by the BBSRC-funded Designing Future Wheat ISP.

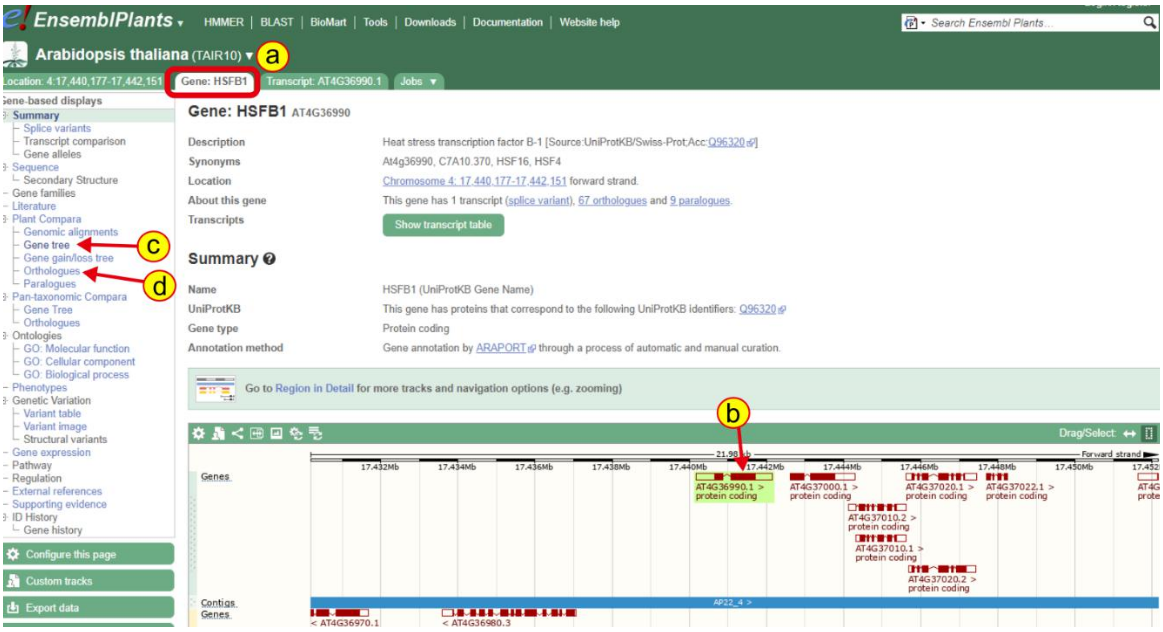
BBSRC bioscience for the future

Past Funding

transPLANT, funded by the European Commission within its 7th Framework Programme, under the thematic area "Infrastructures", contract number 283496.

Figure 1: Searching for a gene on EnsemblPlants

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).


a
b
c
d
21.98 kb
Forward strand
17.432Mb
17.434Mb
17.436Mb
17.438Mb
17.440Mb
17.442Mb
17.444Mb
17.446Mb
17.448Mb
17.450Mb
17.452
Genes
AT4G36990.1 > protein coding
AT4G37000.1 > protein coding
AT4G37020.1 > protein coding
AT4G37022.1 > protein coding
AT4G37010.2 > protein coding
AT4G37010.1 > protein coding
AT4G37020.2 > protein coding
AT4G prote
Contigs
AP22_4 >
Genes
< AT4G36970.1
< AT4G36980.3
Drag/Select:

EnsemblPlants HMMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help

Search Ensembl Plants…

Arabidopsis thaliana (TAIR10)

Location: 4:17,440,177-17,442,151 | Gene: HSFB1 | Transcript: AT4G36990.1 | Jobs

Gene-based displays
- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Gene families
- Literature
- Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Molecular function
  - GO: Cellular component
  - GO: Biological process
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page

Custom tracks

Export data

Gene: HSFB1 AT4G36990

Description	Heat stress transcription factor B-1 [Source:UniProtKB/Swiss-Prot;Acc:Q96320]
Synonyms	At4g36990, C7A10.370, HSF16, HSF4
Location	Chromosome 4: 17,440,177-17,442,151 forward strand.
About this gene	This gene has 1 transcript (splice variant), 67 orthologues and 9 paralogues.
Transcripts	Show transcript table

Summary

Name	HSFB1 (UniProtKB Gene Name)
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: Q96320
Gene type	Protein coding
Annotation method	Gene annotation by ARAPORT through a process of automatic and manual curation.

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

Figure 2: Gene view on EnsemblPlants

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, 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).

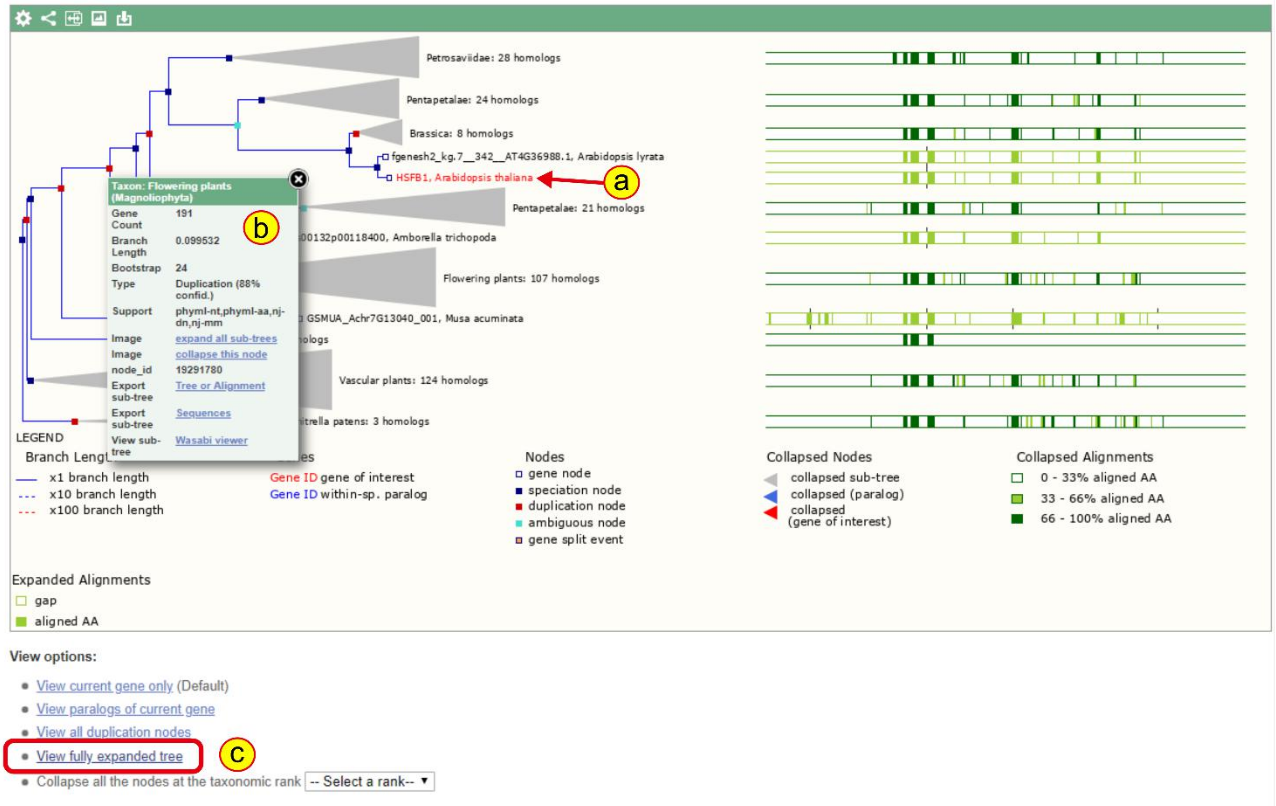

Petrosaviidae: 28 homologs
Pentapetalae: 24 homologs
Brassica: 8 homologs
fgenesh2_kg.7__342__AT4G36988.1, Arabidopsis lyrata
HSFB1, Arabidopsis thaliana
a
Taxon: Flowering plants (Magnoliophyta)
Gene Count 191
b
Branch Length 0.099532
Bootstrap 24
Type Duplication (88% confid.)
Support phyml-nt,phyml-aa,nj-dn,nj-mm
Image expand all sub-trees
Image collapse this node
node_id 19291780
Export sub-tree Tree or Alignment
Export sub-tree Sequences
View sub-tree Wasabi viewer
Pentapetalae: 21 homologs
00132p00118400, Amborella trichopoda
Flowering plants: 107 homologs
GSMUA_Achr7G13040_001, Musa acuminata
ologs
Vascular plants: 124 homologs
itrella patens: 3 homologs
LEGEND
Branch Lengt
x1 branch length
x10 branch length
x100 branch length
es
Gene ID gene of interest
Gene ID within-sp. paralog
Nodes
gene node
speciation node
duplication node
ambiguous node
gene split event
Collapsed Nodes
collapsed sub-tree
collapsed (paralog)
collapsed (gene of interest)
Collapsed Alignments
0 - 33% aligned AA
33 - 66% aligned AA
66 - 100% aligned AA
Expanded Alignments
gap
aligned AA
View options:
View current gene only (Default)
View paralogs of current gene
View all duplication nodes
View fully expanded tree
c
Collapse all the nodes at the taxonomic rank -- Select a rank--

Figure 3: EnsemblPlants 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, these includes wild relatives and progenitor species. Triticum Urartu is the wheat A genome progenitor, and usually you would expect one gene from T. uratu 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 Emmer wheat, a tetraploid species 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.

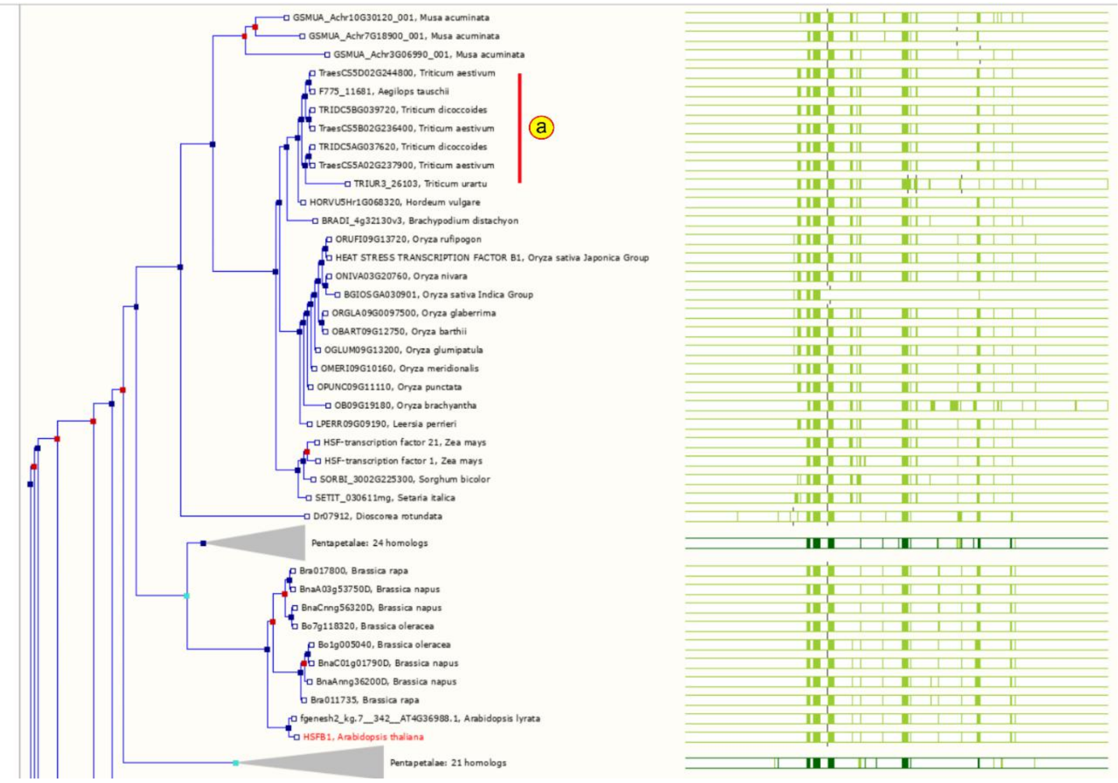
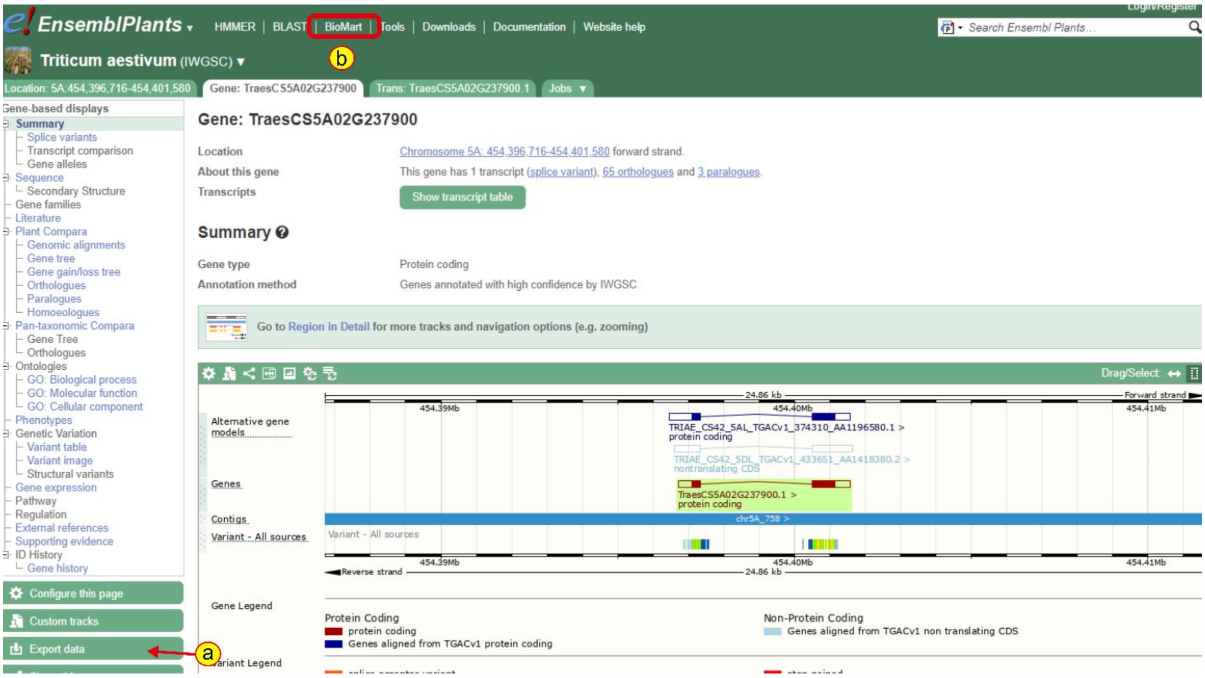

GSMUA_Achr10G30120_001, Musa acuminata
GSMUA_Achr7G18900_001, Musa acuminata
GSMUA_Achr3G06990_001, Musa acuminata
TraesCS5D02G244800, Triticum aestivum
F775_11681, Aegilops tauschii
TRIDC5BG039720, Triticum dicoccoides
TraesCS5B02G236400, Triticum aestivum
TRIDC5AG037620, Triticum dicoccoides
TraesCS5A02G237900, Triticum aestivum
TRIUR3_26103, Triticum urartu
a
HORVU5Hr1G068320, Hordeum vulgare
BRADI_4g32130v3, Brachypodium distachyon
ORUFI09G13720, Oryza rufipogon
HEAT STRESS TRANSCRIPTION FACTOR B1, Oryza sativa Japonica Group
ONIVA03G20760, Oryza nivara
BGIOSGA030901, Oryza sativa Indica Group
ORGLA09G0097500, Oryza glaberrima
OBART09G12750, Oryza barthii
OGLUM09G13200, Oryza glumipatula
OMERI09G10160, Oryza meridionalis
OPUNC09G11110, Oryza punctata
OB09G19180, Oryza brachyantha
LPERR09G09190, Leersia perrieri
HSF-transcription factor 21, Zea mays
HSF-transcription factor 1, Zea mays
SORBI_3002G225300, Sorghum bicolor
SETIT_030611mg, Setaria italica
Dr07912, Dioscorea rotundata
Pentapetalae: 24 homologs
Bra017800, Brassica rapa
BnaA03g53750D, Brassica napus
BnaCnng56320D, Brassica napus
Bo7g118320, Brassica oleracea
Bo1g005040, Brassica oleracea
BnaC01g01790D, Brassica napus
BnaAnng36200D, Brassica napus
Bra011735, Brassica rapa
fgenesh2_kg.7__342__AT4G36988.1, Arabidopsis lyrata
HSFB1, Arabidopsis thaliana
Pentapetalae: 21 homologs

Figure 4: Expanded clade in EnsemblPlants gene tree

8. You can also find orthologues by clicking on the orthologues 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.



Gene: TraesCS5A02G237900

Location: Chromosome 5A: 454,396,716-454,401,580 forward strand.

About this gene: This gene has 1 transcript (splice variant), 65 orthologues and 3 paralogues.

Transcripts

Summary

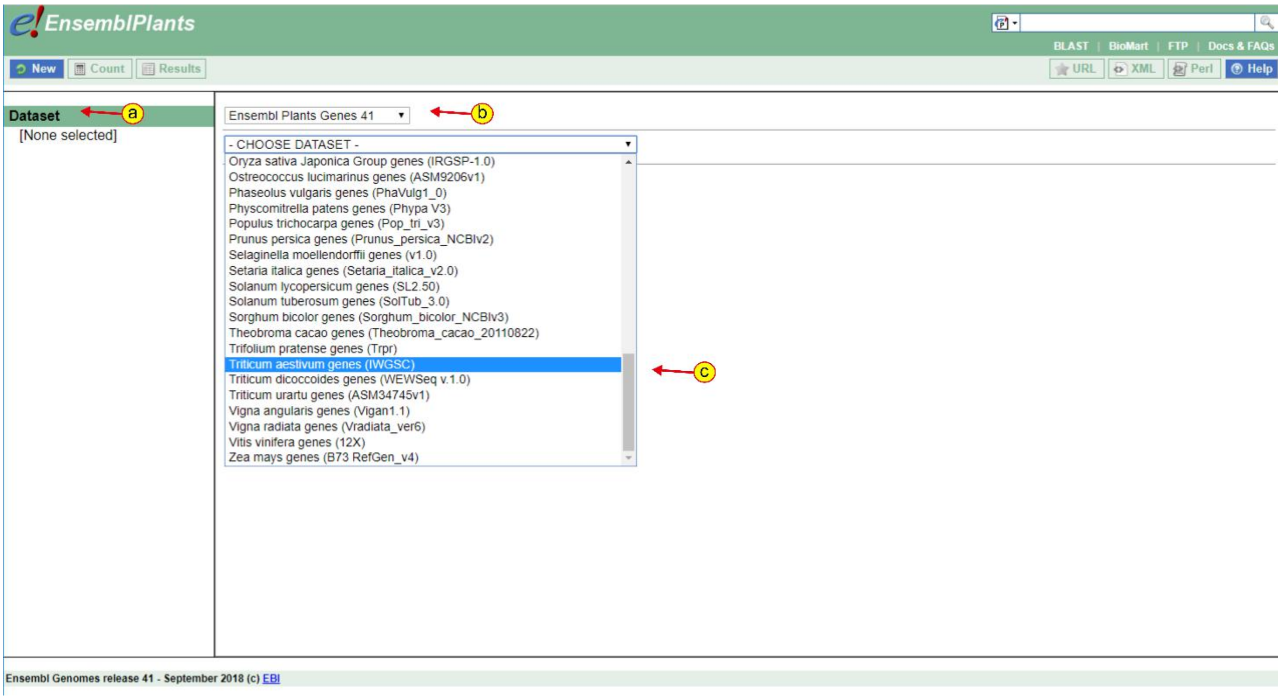
Gene type: Protein coding

Annotation method: Genes annotated with high confidence by IWGSC

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

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).



Dataset
[None selected]

Ensembl Plants Genes 41

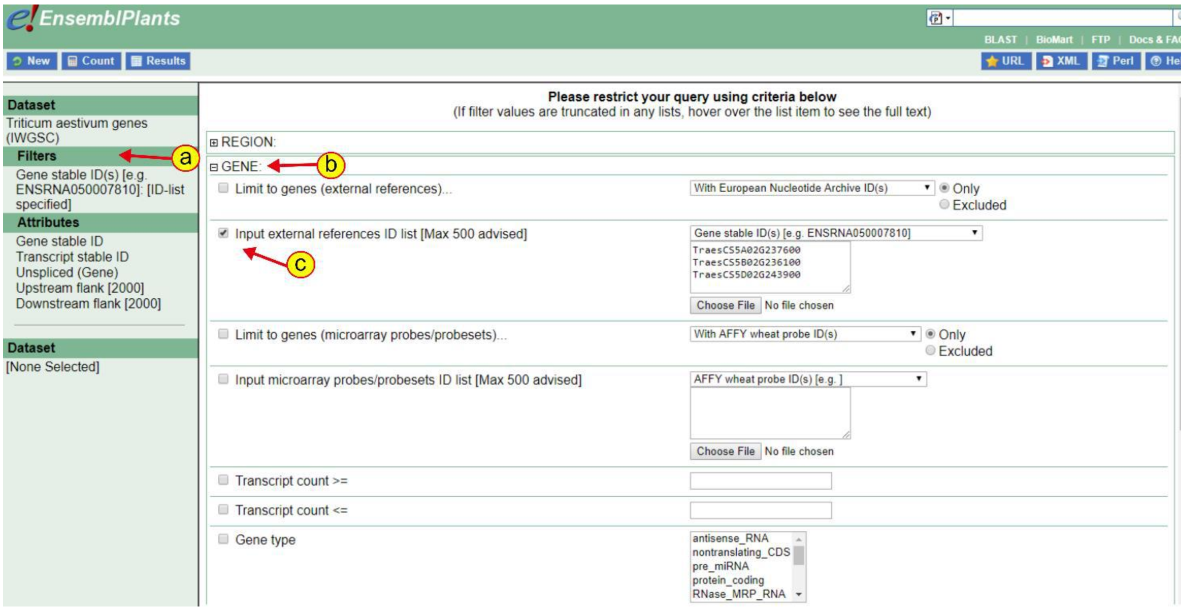
- CHOOSE DATASET -
Oryza sativa Japonica Group genes (IRGSP-1.0)
Ostreococcus lucimarinus genes (ASM9206v1)
Phaseolus vulgaris genes (PhaVulg1_0)
Physcomitrella patens genes (Phypa V3)
Populus trichocarpa genes (Pop_tri_v3)
Prunus persica genes (Prunus_persica_NCBIv2)
Selaginella moellendorffii genes (v1.0)
Setaria italica genes (Setaria_italica_v2.0)
Solanum lycopersicum genes (SL2.50)
Solanum tuberosum genes (SolTub_3.0)
Sorghum bicolor genes (Sorghum_bicolor_NCBIv3)
Theobroma cacao genes (Theobroma_cacao_20110822)
Trifolium pratense genes (Trpr)
Triticum aestivum genes (IWGSC)
Triticum dicoccoides genes (WEWSeq v.1.0)
Triticum urartu genes (ASM34745v1)
Vigna angularis genes (Vigan1.1)
Vigna radiata genes (Vradiata_ver6)
Vitis vinifera genes (12X)
Zea mays genes (B73 RefGen_v4)

Ensembl Genomes release 41 - September 2018 (c) EBI

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



Dataset
Triticum aestivum genes (IWGSC)

Filters
Gene stable ID(s) [e.g. ENSRNA050007810]: [ID-list specified]

Attributes
Gene stable ID
Transcript stable ID
Unspliced (Gene)
Upstream flank [2000]
Downstream flank [2000]

Dataset
[None Selected]

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

⊞ REGION:

⊟ GENE:

□ Limit to genes (external references)... | With European Nucleotide Archive ID(s) | ● Only ○ Excluded

☑ Input external references ID list [Max 500 advised] | Gene stable ID(s) [e.g. ENSRNA050007810]

```
TraesCS5A02G237600
TraesCS5B02G236100
TraesCS5D02G243900
```

Choose File No file chosen

□ Limit to genes (microarray probes/probesets)... | With AFFY wheat probe ID(s) | ● Only ○ Excluded

□ Input microarray probes/probesets ID list [Max 500 advised] | AFFY wheat probe ID(s) [e.g.]

Choose File No file chosen

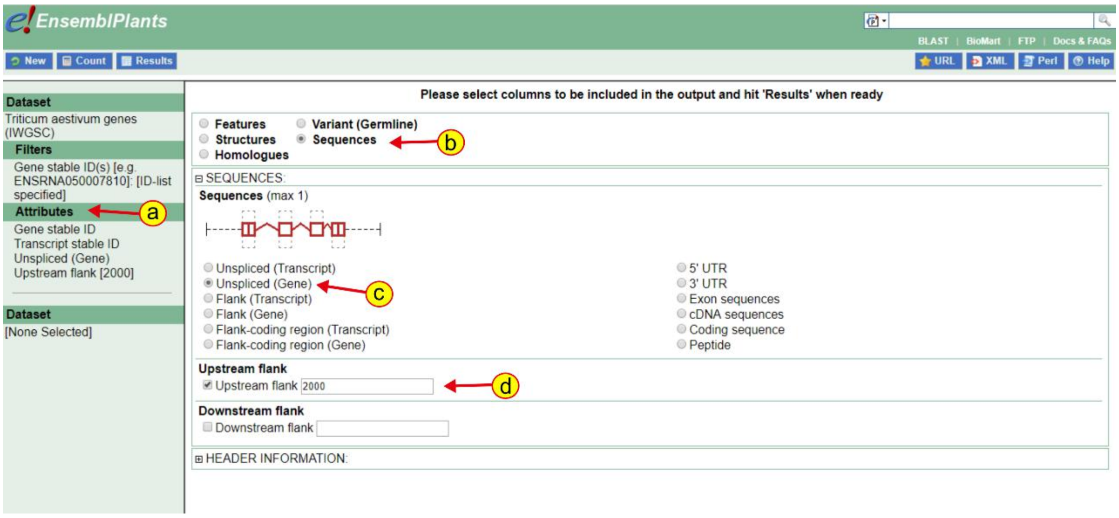
□ Transcript count >=

□ Transcript count <=

□ Gene type | antisense_RNA, nontranslating_CDS, pre_miRNA, protein_coding, RNase_MRP_RNA

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.



Dataset
Triticum aestivum genes (IWGSC)

Filters
Gene stable ID(s) [e.g. ENSRNA050007810]: [ID-list specified]

Attributes
Gene stable ID
Transcript stable ID
Unspliced (Gene)
Upstream flank [2000]

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

○ Features ○ Variant (Germline)
○ Structures ● Sequences
○ Homologues

⊟ SEQUENCES:

Sequences (max 1)

○ Unspliced (Transcript)
● Unspliced (Gene)
○ Flank (Transcript)
○ Flank (Gene)
○ Flank-coding region (Transcript)
○ Flank-coding region (Gene)
○ 5' UTR
○ 3' UTR
○ Exon sequences
○ cDNA sequences
○ Coding sequence
○ Peptide

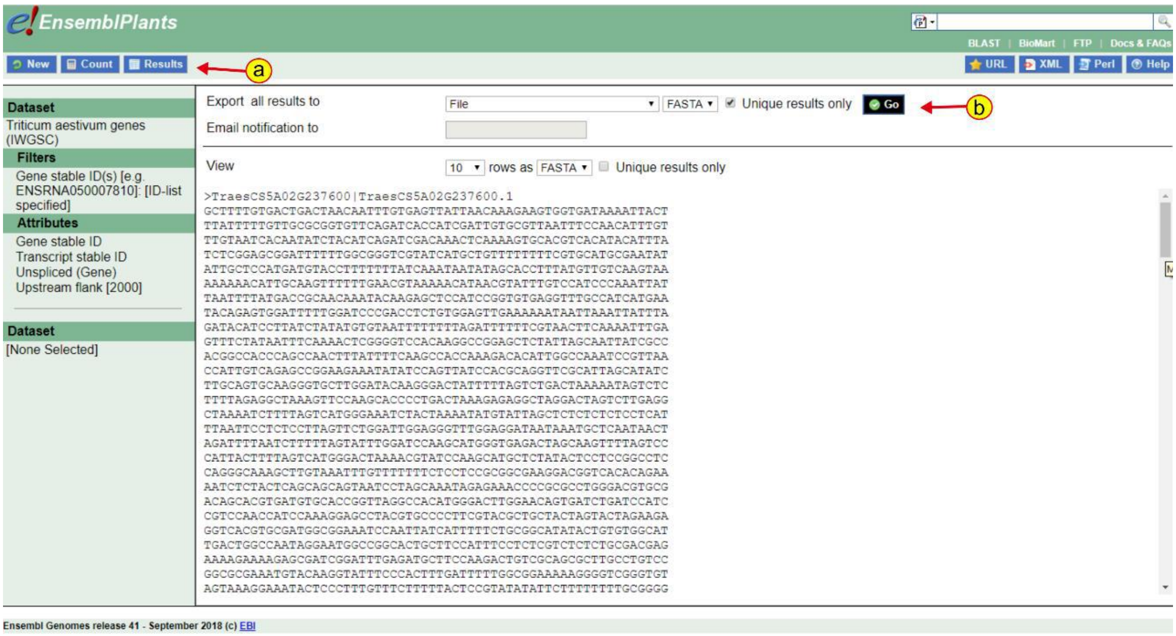
Upstream flank
☑ Upstream flank 2000

Downstream flank
□ Downstream flank

⊞ HEADER INFORMATION:

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)



Dataset
Triticum aestivum genes (IWGSC)
Filters
Gene stable ID(s) [e.g. ENSRNA050007810]: [ID-list specified]
Attributes
Gene stable ID
Transcript stable ID
Unspliced (Gene)
Upstream flank [2000]

Dataset
[None Selected]

Export all results to: File ▾ FASTA ▾ ☑ Unique results only

Email notification to

View: 10 ▾ rows as FASTA ▾ □ Unique results only

```
>TraesCS5A02G237600|TraesCS5A02G237600.1
GCTTTTGTGACTGACTAACAATTTGTGAGTTATTAACAAAGAAGTGGTGATAAAATTACT
TTATTTTTGTTGCGCGGTGTTCAGATCACCATCGATTGTGCGTTAATTTCCAACATTTGT
TTGTAATCACAATATCTACATCAGATCGACAAACTCAAAAGTGCACGTCACATACATTTA
TCTCGGAGCGGATTTTTTGGCGGGTCGTATCATGCTGTTTTTTTTCGTGCATGCGAATAT
ATTGCTCCATGATGTACCTTTTTTTATCAAATAATATAGCACCTTTATGTTGTCAAGTAA
AAAAAACATTGCAAGTTTTTTGAACGTAAAAACATAACGTATTTGTCCATCCCAAATTAT
TAATTTTATGACCGCAACAAATACAAGAGCTCCATCCGGTGTGAGGTTTGCCATCATGAA
TACAGAGTGGATTTTTGGATCCCGACCTCTGTGGAGTTGAAAAAATAATTAAATTATTTA
GATACATCCTTATCTATATGTGTAATTTTTTTTAGATTTTTTCGTAACTTCAAAATTTGA
GTTTCTATAATTTCAAAACTCGGGGTCCACAAGGCCGGAGCTCTATTAGCAATTATCGCC
ACGGCCACCCAGCCAACTTTATTTTCAAGCCACCAAAGACACATTGGCCAAATCCGTTAA
CCATTGTCAGAGCCGGAAGAAATATATCCAGTTATCCACGCAGGTTCGCATTAGCATATC
TTGCAGTGCAAGGGTGCTTGGATACAAGGGACTATTTTTAGTCTGACTAAAAATAGTCTC
TTTTAGAGGCTAAAGTTCCAAGCACCCCTGACTAAAGAGAGGCTAGGACTAGTCTTGAGG
CTAAAATCTTTTAGTCATGGGAAATCTACTAAAATATGTATTAGCTCTCTCTCTCCTCAT
TTAATTCCTCTCCTTAGTTCTGGATTGGAGGGTTTGGAGGATAATAAATGCTCAATAACT
AGATTTTAATCTTTTTAGTATTTGGATCCAAGCATGGGTGAGACTAGCAAGTTTTAGTCC
CATTACTTTTAGTCATGGGACTAAAACGTATCCAAGCATGCTCTATACTCCTCCGGCCTC
CAGGGCAAAGCTTGTAAATTTGTTTTTTTCTCCTCCGCGGCGAAGGACGGTCACACAGAA
AATCTCTACTCAGCAGCAGTAATCCTAGCAAATAGAGAAACCCCGCGCCTGGGACGTGCG
ACAGCACGTGATGTGCACCGGTTAGGCCACATGGACTTGGAACAGTGATCTGATCCATC
CGTCCAACCATCCAAAGGAGCCTACGTGCCCCTTCGTACGCTGCTACTAGTACTAGAAGA
GGTCACGTGCGATGGCGGAAATCCAATTATCATTTTTCTGCGGCATATACTGTGTGGCAT
TGACTGGCCAATAGGAATGGCCGGCACTGCTTCCATTTCCTCTCGTCTCTCTGCGACGAG
AAAAGAAAAGAGCGATCGGATTTGAGATGCTTCCAAGACTGTCGCAGCGCTTGCCTGTCC
GGCGCGAAATGTACAAGGTATTTCCCACTTTGATTTTTGGCGGAAAAAGGGGTCGGGTGT
AGTAAAGGAAATACTCCCTTTGTTTCTTTTTACTCCGTATATATTCTTTTTTTTGCGGGG
```

Ensembl Genomes release 41 - September 2018 (c) EBI

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 uses the variety Chinese Spring (see "Genome assemblies" for more information). However, genome assemblies for other wheat varieties 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
Arina*LrFor*	winter	Switzerland	10+ Genome Project
Julius	winter	Germany	10+ Genome Project
Jagger	winter	US	10+ Genome Project
Paragon	spring	UK	10+ Genome Project
Cadenza	spring	UK	10+ Genome Project
Synthetic W7984	spring	Mexico	(Chapman *et al.*, 2015)
Robigus	winter	UK	10+ Genome Project
Claire	winter	UK	10+ Genome Project
Tetraploid pasta wheat			
Zavitan†	-	Israel	(Avni *et al.*, 2017)

Svevo	spring	Italy	Interomics
Kronos	spring	US	10+ Genome Project

† Zavitan is a tetraploid wild emmer (*T. dicoccoides*) accession

* Varieties included within the 10+ Genome Project can be accessed through the Earlham Grassroot Genomics portal (https://wheatis.tgac.ac.uk/grassroots-portal/blast) and the 10+ Genome 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 subject to Toronto agreement. Synthetic W7984 and Zavitan can be accessed through the Grassroot and 10+ Genome portal, respectively.