In the following paragraphs we will give a basic introduction on how to use the EnsemblPlants website to find and download data. This primer is written for the EnsemblPlants website (www.plants.ensembl.org) with a special focus on the wheat data section of EnsemblPlants. The basic steps described here can also be used when browsing other species.

Disclaimer:
The EnsemblPlants website is continuously being updated. We will strive to keep this document up-to-date, but if you should find that a certain aspect of the website is no longer working (or working differently) than how it is described here please consult the help pages on the EnsemblPlants website (http://plants.ensembl.org/info/website/index.html).

a) What is on the EnsemblPlants website?

A brief overview of available data and how to access it

The EnsemblPlants homepage (www.plants.ensembl.org) allows users to access data in different ways. There are two search bars (blue boxes in Figure 1) that allow the user to query either the entire database or a single species for e.g. genomic positions or genes. At the moment, the gene models for wheat (IWGSC RefSeqv1.1) are not associated with a lot of metadata, i.e. a search for the well-known vernalisation gene Vrn1 would yield hits to Arabidopsis thaliana and Brassica rapa, but not to Triticum aestivum, even though the gene does exist within the database. Instead, gene names need to be queried using their designation, e.g. TraesCS5A02G391700 for Vrn1-A.

The database can also be searched via HMMER (protein) or via BLAST (both DNA and protein) (yellow box in Figure 1). The first one uses Hidden Markov Models (HMM) to discover even remote homologues using a probabilistic method (see http://hmmer.org/ for more information), while the second one employs the BLAST algorithm (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the most similar sequences. BLAST is currently the best option to query a gene in wheat, as the gene database for wheat is superior to the protein database, e.g. a BLAST search for Vrn1 would lead the user to the correct wheat gene.

Users can download large quantities of specific data, e.g. all genes within a certain interval using the BioMart tool (under the “BioMart” tab, yellow box in Figure 1) or download raw data from an FTP server (under the “Downloads” tab, yellow box in Figure 1). These and other tools can be also accessed from the tools section (under the “Tools” tab, yellow box in Figure 1).

The documentation and help sections are well worth reading and explain the tools and data in more detail (under the “Documentation” and “Website help” tabs, yellow box in Figure 1). For usage tutorials go to http://www.ensembl.org/info/website/tutorials/index.html.
Figure 1: EnsemblPlants homepage
This is a screenshot of the EnsemblPlants website. It contains quick links to the most used species, a search box for user queries as well as information about improvements from the current release. The link to the full list of species is highlighted with a red box.

As of January 2019, there is data for 58 species deposited on the EnsemblPlants website. This includes dicots (e.g. Arabidopsis, Brassica, and tomato), monocots (e.g. wheat, barley, and rice) as well as data for some algae and mosses. You can view the list of species as well as the type of data available for each by clicking the “View full list of all EnsemblPlants species” link in the middle left of the website (red box in Figure 1).

If you follow the link to view the full list of species on EnsemblPlants you are presented with a table listing (see Figure 2):

1. The species name
2. The species classification
3. The UniProt Taxon ID of the species
4. The name of the current genome assembly
5. A link to the genome assembly on the European Nucleotide Archive (ENA)
6. A list of the type of data available on EnsemblPlants for each species, e.g. variation data
Figure 2: Full list of all species on EnsemblPlants
This is a screenshot of the full list of species included on EnsemblPlants. There are links to the official UniProt Taxon ID and the genome assembly used on EnsemblPlants. The available data for each species is divided into six categories, e.g. variation data. A green tick mark indicates whether data of a certain type is available for a species.

After selecting your species of interest (in our case wheat; Triticum aestivum) you are presented with a new window containing lots of important information, e.g. details about the genome assembly (red box in Figure 3). This is crucial information as it helps users to understand what type of data is actually present in the database. Again, there are search bars for user queries as well.
Figure 3: Species specific information

This screenshot shows species specific information for wheat. It contains details on the genome assembly, gene annotation, variation data, etc.

**b) The layout of EnsemblPlants**

The Location tab

So far, we have been navigating the main page as well as the tool and help pages of the EnsemblPlants website, which represent part of the first layer of the website. Once we query the database we are presented with a list of hits (see Figure 4). For every query hit the user is presented with there are several options on how to continue. These usually are:

1. Gene ID; if the query was a gene or a scaffold with a gene, then this displays the ID of the gene in the database; in case of multiple genes on the same scaffold, a separate hit is shown for every gene.
2. Species; if querying from the main website with all species, this allows to filter the results by species.
3. Location; the genomic location of the query, which can be either a scaffold or a chromosome location.
4. Gene trees; if the query was a gene or a scaffold with a gene, then this displays the name of the gene tree created with Plant Compara or Pan-taxonomic Compara (see section c “The power of Compara”).
Figure 4: Search Results
Querying the database opens a window with search results. From every result, you can navigate to a different part of the database.

If we follow the "Location" link, we are directed to the second layer of the EnsemblPlants website: The Location tab (red box in Figure 5). This is a genomic view of the region matching with the users query. The tab consists of three windows representing three different “zoom” levels: The first window displays the genomic location relative to the entire chromosome (blue box in Figure 5). It allows the user to quickly browse between different positions on the chromosome and also to look at wider or narrower stretches of the chromosome. In case of the region being a scaffold (scaffolds are not part of the chromosome pseudomolecules!) this window only shows the scaffold itself.

The second window (by default) represents a region of 500 kb surrounding the users query. It displays the genes present within the 500 kb interval and thus allows the user to see the genes that for example underlie a quantitative trait locus (QTL). Other tracks, e.g. alternative splice variants, can be added as well.

Finally, the third window represents a detailed view of the user’s query, e.g. a gene model. By default, only the gene model and a track for sequence variation are shown here, but this window has a lot of extra tracks to offer (many more than the previous two windows). Users are encouraged to play with the available tracks as there is a lot of information available that is hidden by default.
**Figure 5: The Location tab**
This tab displays the genomic region matching the users query. It is subdivided into three windows, each with a different “zoom” level. The pane on the left offers alternative views of the genomic region.

In addition to the three windows, there is also a pane on the left-hand side of the tab that allows users to view the region differently (orange box in Figure 5). In general, this pane offers tab specific viewing options, as we will see later on. One of these alternative viewing options is the “Polyploid view”, which adds windows for the homoeologous regions, i.e. instead of the three windows described above a set of nine windows is displayed. This can be quite useful as it allows scanning of the three homoeologous regions quickly on different levels of detail.
Figure 6: Polyploid view
When selecting “Polyploid view” from the left-hand side menu of the Location tab (red box in Figure 6) users can inspect genomic regions from all three homoeologs at the same time.

The Gene tab

Going back to the query results in Figure 4 users can look directly at a gene that falls within their query as opposed to the genomic location. For this, the user simply follows the “Gene ID” in the search results. Alternatively, one can get to the Gene tab after inspecting the genomic region by selecting the Gene tab in the bar above the information windows (see red box in Figure 7).

The Gene tab offers some basic information about the gene model in question, e.g. its genomic location or the number of splice variants (blue box in Figure 7). A single window (green box in Figure 7) shows detailed information on the gene and its surrounding genomic location. This window is similar to the detailed view window in the Location tab (purple window in Figure 5). The difference is mostly in the type of additional tracks that can be activated. Again, by default only the main gene model is being shown, but additional tracks like transcriptome data or alternative splice variants are also available.

As in the Location tab there is a pane on the left-hand side of the screen with various display options. By comparing Figure 5 with Figure 7 one can immediately see that there are many more display options in the pane of the Gene tab. As was mentioned before, these display options are tab specific. The available options depend on the data deposited on EnsemblPlants and contain for example simple GO terms of the gene or a detailed gene tree with homoeologous and orthologous
genes or a table listing the recorded variants of this gene (see section “SNP variation” for more details).

**Figure 7: The Gene tab**

This screenshot from the EnsemblPlants website shows the detailed gene view tab. This window displays some information about the gene model, e.g. its genomic location or the number of splice variants (blue box). The main window shows detailed information about the gene and additional tracks can be added (green box). The pane on the left side of the tab offers additional information, e.g. the names of homoeologous genes or GO terms (orange box). Note that the options in this pane are different to the ones in the Location tab. Users can navigate between e.g. the location and gene view using the tabs above the information windows (red box).

**The Transcript tab**

The last of the tabs gives detailed information on the transcript of a gene model. Again, some basic information like the genomic location and number of splice variants is displayed as well as a very detailed view of the intron-exon structure without any genomic information (blue box in Figure 8). The strength of this tab is again in the different viewing options offered by the pane on the left-hand side of the screen (orange box in Figure 8). This offers for example a detailed cDNA view of the transcript, a table with variants or a list of the known protein domains encoded by the transcript.
**c) The power of Compara**

Comparative studies are a powerful approach to identify genes in related species with specific functions, e.g. flowering time, plant height or grain size. These genes have likely been characterised in a model organism like rice and can be used as candidates to identify corresponding pathways in related but less well characterised species like wheat.

To identify the correct orthologous genes between species can be a daunting task depending on the size and similarity of the gene family as well as the number of genes a researcher is interested in. Luckily, the EnsemblPlants team has already done this work. As mentioned in section b the gene models on the EnsemblPlants website can be viewed in a gene tree format. These gene trees were built using a sophisticated pipeline (for more information on building the gene trees see [http://plants.ensembl.org/info/genome/compara/homology_method.html](http://plants.ensembl.org/info/genome/compara/homology_method.html)). Two types of trees were built where possible:

- The Plant Compara tree compares genes from mostly the plant kingdom, although genes from other eukaryotic organisms are also included. The Plant Compara tree contains many more species than the Pan-taxonomic tree and is thus the tree of choice for plant to plant comparisons.
- The Pan-taxonomic Compara trees are similar to the Plant Compara trees but include taxonomically more diverse species. Thus this tree is better suited for comparing genes across kingdoms.

To access the gene trees users can either query the database with a gene name (see Figure 9) or select the gene tree view from the Gene tab display options panel (see orange box in Figure 7).
Figure 9: Query results for a gene
This screenshot shows the query results for a gene from rice. From here, the user can either navigate to the Plant Compara tree or a Pan-taxonomic tree.

The Gene tree view

The gene tree view window is separated into two major units. The first is a summary of the gene tree, e.g. how many species and genes are included (blue box in Figure 10) and how many of these genes share certain GO and INTERPRO domains (green box in Figure 10); the latter can be used to highlight genes with specific GO and INTERPRO domains.

Figure 10: Gene tree view part 1
This is the first screenshot of the gene tree view. Note that it is part of the Gene tab (red box). The first half of the tree view contains information on the gene tree (blue box) as well as on the GO and INTERPRO domains associated with the tree member (purple box).

The second part of the gene tree view consists of the actual gene tree and the alignments that were created using the Compara pipeline, with the initially queried gene being highlighted in red. By default, most of the gene tree is collapsed; users can then expand the sections they are interested in.
interested in. Right-clicking a gene name opens up a tool tip giving user more information about the gene and also allowing users to follow a link to the gene. Both Plant Compara and Pan-taxonomic gene trees share the same layout.

Using these gene trees allows researchers to quickly identify orthologous genes in other species. In Figure 10 and Figure 11 we used a gene from rice as query and using the gene tree we can rapidly identify the orthologous genes from wheat.

The accuracy of the gene trees however depends on the quality of the gene models used. Although the current set of gene models for wheat (IWGSC RefSeqv1.1) are of very high quality, there is no guarantee that every gene is correctly annotated.

As a result, a wrong gene model can lead to an inaccurate gene tree. The example above highlights this nicely, as orthologous genes in wheat are apparently only found on chromosome 7B and 7D, but not 7A. While it is possible that the A-genome has lost its copy of this gene it is also quite possible that the 7A gene model is missing. A manual BLAST search using the genomic sequence of either the B- or D-genome gene model confirms that indeed the A-genome gene copy exists but is missing from the RefSeqv1.1 gene model set. Users are encouraged to check the wheat gene models using the criteria outlined in the “Gene models” section.
**d) Using BLAST on *EnsemblPlants***

As was mentioned at the start of the document users can query the *EnsemblPlants* database using BLAST. The BLAST tool here allows up to 30 sequences (either DNA or protein; blue box in Figure 12) to be used as queries against databases of up to 20 different plant species (green box in Figure 12). For every sequence and species a separate job is automatically submitted, and users can view them later one by one. The BLAST parameters can be changed to suit a user's needs, e.g. e-value threshold, gap open penalties, etc (purple box in Figure 12).

![Figure 12: BLAST tool on *EnsemblPlants*](image)

This is a screenshot from the *EnsemblPlants* BLAST tool. Up to 30 sequences (blue box) can be used as queries against up to 20 species (green box). The BLAST parameters can be tweaked to achieve best results (purple box).

The output of a BLAST search is displayed in a table with columns for alignment length, percent similarity, etc (see Figure 13). The table can be downloaded (like all tables on *EnsemblPlants*) and the pair-wise alignments can be accessed by clicking the “Alignment” link in the percent identity column.

Every HSP (High-scoring Segment Pair) is reported as a separate hit, so using CDS (coding sequence) as query to BLAST a genomic database will result in a lot of hits. Thus, it is often easier...
to interpret the results as there will be fewer high-quality hits when using genomic sequence for
BLAST queries on *Ensembl*Plants.

![Figure 13: BLAST results](image)

This screenshot shows the output of a BLAST search on *Ensembl*Plants. The output is in tabular
format that can be downloaded. Genes overlapping genomic sequences are shown in a separate
column.

e) Using BioMart on *Ensembl*Plants

At the beginning of this document we mentioned the tool BioMart. This is essentially a tool to
download specific sets of data in both small and large quantities.

Users first select the type of data and organism they are interested in (see Figure 14). Then they
can apply a set of filters, e.g. a genomic location, to limit the number of results retrieved from their
query: Setting no filter would mean that all data would be downloaded. Furthermore, the user can
specify what attributes the data should have, e.g. GO terms associated with genes. Lastly, a user
can specify an additional dataset from a different organism, e.g. rice, for which filters and attributes
can again be set separately; using two datasets allows users to download for example all wheat
genes and their orthologous genes from rice. Thus, by combining different sorts of filters, attributes
and datasets individual users can download data specifically tailored to their needs.

The number of filters and attributes that can be used is big and can be daunting for new users. It is
highly recommended for first time users to download small sets of defined data to get used to the
interface. Once users understand the basic concepts however BioMart offers quick access to large
quantities of data.

We have created a step-by-step example of how to use BioMart, which can be found in the
Finding wheat orthologs section.
**Figure 14: BioMart on EnsemblPlants**
This screenshot shows the basic layout of the BioMart tool on EnsemblPlants. Users can select datasets from different species, select filters and choose what type of data they want to download.