

TransPLANT user training workshop 2015

Slides:

<http://tinyurl.com/transplant2015>

Workshop on variation data

**EMBL-EBI
Hinxton-UK**

2nd July 2015

Ensembl Genomes Team

Notes:

This workshop is based on Ensembl Genomes release 27 (July 2015).

Some useful information:

1) Ensembl Plants browser website

<http://plants.ensembl.org>

2) Ensembl Genomes on Twitter

<http://twitter.com/ensemblgenomes>

3) Workshop materials (in pdf)

http://www.ebi.ac.uk/~denise/plants_cam

Feel free to tackle questions relative to your own research instead of following the ones provided in our course booklet. The answers for the latter can be found here:

http://www.ebi.ac.uk/~denise/plants_cam/answers

Questions or comments?

helpdesk@ensemblgenomes.org

TABLE OF CONTENTS

[OVERVIEW](#)

[ENSEMBL PLANTS: BROWSER WALKTHROUGH](#)

[GENOME BROWSER](#)

[BIOMART](#)

[GENETIC VARIATION](#)

[COMPARATIVE GENOMICS](#)

OVERVIEW

The Ensembl Genomes project was launched in 2009 and provides annotation of genes and other genomic features such as sequence variants and conserved regions for more than 11,000 species within the Bacteria, Protists, Metazoa, Fungi and Plant domains.

All the data in Ensembl Genomes are freely available.
It can be accessed via the

- web browser
- data retrieval tool Biomart
- programmatically through our APIs (PERL or RESTful) and MySQL queries
- FTP downloads

The screenshot shows the Ensembl Genomes website interface. At the top is a navigation bar with links: About us, Genomes, Data types, Data access, FAQs, and a search bar. Below the navigation bar, the main content area is divided into several sections. On the left, there's a section titled "Ensembl Genomes: Extending Ensembl across the taxonomic space." which lists recent genome releases with small thumbnail images and brief descriptions. These include: Amborella trichopoda genome, Orthologues, Paralogues and Homoeologues for hexaploid bread wheat, Five new rice genomes, Onchocerca volvulus genome and annotation from WormBase, New Schizosaccharomyces genomes, New and updated hymenopteran genomes, and Assembly mapping. On the right, there's a "What's New in Release 22 (April 2014)" section with sub-headings for Ensembl Bacteria, Ensembl Fungi, Ensembl Metazoa, Ensembl Plants, and Ensembl Protists, each followed by a brief update. A "Have a question?" box with a link to "Frequently Asked Questions (FAQs)" is also present. At the bottom left, there's a footer section mentioning "Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system" and logos for EMBL-EBI and E!mpowered.

Ensembl Genomes: Extending Ensembl across the taxonomic space.

- Amborella trichopoda genome
- Orthologues, Paralogues and Homoeologues for hexaploid bread wheat
- Five new rice genomes
- Onchocerca volvulus genome and annotation from WormBase
- New Schizosaccharomyces genomes
- New and updated hymenopteran genomes
- Assembly mapping

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please [click here](#).

EMBL-EBI

What's New in Release 22 (April 2014)

Ensembl Bacteria

Ensembl Bacteria has been updated to include the latest versions of 11,010 genomes (10,760 bacteria and 250 archaea) from the INSDC archives.

Ensembl Fungi

Protein orthology data has been used to project GO annotation from *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* to all other fungal species.

Ensembl Metazoa

One new species has been added to Ensembl Metazoa, *Onchocerca volvulus* the parasitic nematode which causes the 'river blindness' disease and several other species have been updated: *Caenorhabditis elegans* (nematode worm), *Aedes aegypti* (yellow fever mosquito), *Bombyx mori* (silkworm), *Daphnia pulex* (common water flea), *Nematostella vectensis* (starlet sea anemone).

Ensembl Plants

A new comparative analysis of the component A, B and D genomes of hexaploid bread wheat (*Triticum aestivum*) has allowed us to call orthology relationships between the three component genomes, identifying the so-called homoeologous genes. [Click here for example](#). Homoeologous relationships between the component genomes can now also be browsed in our new region comparison view. [Click here for example](#).

In collaboration with Gramene we have imported five new rice genomes produced by the OMAP project, *Oryza barthii*, *plumapatula*, *meridionalis*, *nivara* and *punctata* into Ensembl Plants. The OMAP project aims to understand the evolution, physiology and biochemistry of Oryza, complementing the [existing genomes in Ensembl Plants](#).

This release sees the inclusion of two additional plant genomes: *Amborella trichopoda*, the only living species on the sister lineage to all other flowering plants, and Peach (*Prunus persica*), an economically important fruit crop and a representative of the diverse and important Rosaceae family.

Finally, we have added new variation data for *Sorghum bicolor* based on the sorghum association panel (SAP), providing 265 thousand SNPs for 378 SAP lines ([Morris et. al. 2013](#)).

Ensembl Protists

Six new genomes were added to Ensembl Protists, including five plant pathogens, the parasitic oomycetes, *Pythium aphanidermatum*, *Pythium arthenomanes*, *Pythium irregulare*, *Pythium iwayamai* and *Pythium vexans*. The addition of *Bigeloviella natans* expands the taxonomic coverage of Ensembl Protists to Rhizaria.

Have a question?

Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, [contact us](#).

Ensembl Genomes has been empowered by its sister project Ensembl, launched 10 years earlier, in 1999 just before the release of the first draft the human genomic sequence. Ensembl project focuses on vertebrates, whereas Ensembl Genomes focuses on Bacteria, Plants, Fungi, Protists and non-vertebrate Metazoa.

The Ensembl browsers



- launched in 1999
- vertebrates
- Ensembl gene annotation
- EBI and WTSI



- launched in 2009
- non-vertebrates
- community gene annotation
- EBI

Retrieving Data from Ensembl Genomes using BioMart and the APIs

BioMart is a web-interface that can extract information from the Ensembl Genomes databases and present the user with a table of information without the need for programming. It can be used to output sequences or tables of genes along with gene positions (chromosome and base pair locations), single nucleotide polymorphisms (SNPs), homologues, and other annotation in HTML, text, or Microsoft Excel format. BioMart can also translate one type of ID to another, identify genes associated with an **InterPro** domains or gene ontology (**GO**) terms, export gene expression data and lots [more](#).

Ensembl Genomes uses MySQL relational databases to store its information. A comprehensive set of Application Programme Interfaces ([APIs](#)) serve as a middle-layer between underlying database schemes and more specific application programmes. The

API aims to encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes. A RESTful service is also available for non-Perl programmatic access of the data in Ensembl Genomes:

<http://rest.ensemblgenomes.org/>

Ensembl Plants

Ensembl Plants release 25 (January 2015) has got 38 genomes. For the full list of genomes, have a look at:

<http://plants.ensembl.org/info/website/species.html>

You may be interested in a few pathogens causing various diseases in plants. Have a look at the annotation of a few plant pathogens in Ensembl Fungi (fungi.ensembl.org)

Synopsis- what can I do with Ensembl Plants?

- View genes with other annotation along the chromosome;
- Explore homologues and phylogenetic trees across different plant species and across a wider taxonomic range;
- Compare whole genome alignments and conserved regions across species;
- View ESTs, clones, mRNA and proteins for any chromosomal region;
- Examine single nucleotide polymorphisms (SNPs) and/or indels for a gene or a chromosomal region;
- View positions and sequence of mRNAs and proteins that align with an Ensembl genes;
- Display your own data on the Ensembl Genome browser;
- Use BLAST or BLAT against any species in Ensembl Genomes;
- Export sequence or create a table of gene information with BioMart;
- Determine how your variants affect genes and transcripts using the Variant Effect Predictor;
- Share Ensembl views with your colleagues and collaborators;
- Retrieve our data using the Perl or REST APIs.

Need more [help](#)?

- Check Ensembl Genomes [documentation](#)
- Watch our [video](#) tutorial on YouTube
- View our [FAQs](#)
- Read our [scientific publications](#)
- Go to our [online course](#)

Stay in touch!

- ❖ Comments/questions, [email us](#)
- ❖ View the Ensembl [blog](#)
- ❖ Follow us on Twitter [@ensemblgenomes](#)
- ❖ Sign up to our [mailing lists](#)

ENSEMBL PLANTS: BROWSER WALKTHROUGH

The *Arabidopsis* floral homeotic gene APETALA1 (AP1) encodes a putative transcription factor that acts locally to specify the identity of the floral meristem and to determine sepal and petal development (<https://www.wikigenes.org/e/gene/e/843244.html>).

Let's explore the AP1 gene in *Arabidopsis thaliana*.

The following points will be addressed during the walkthrough:

- **The Location tab and genomic location related links:**
 - How do I zoom out to change the gene focus?
 - How to add data tracks (e.g. protein alignments, variation data)?
- **The Gene tab and gene related links:**
 - Can I view the genomic sequence of my gene with its variations?
 - How to find orthologues and paralogues?
- **The Transcript tab and transcript related links:**
 - What is the protein sequence?
 - What proteins and mRNAs are found in other databases?
- **Exporting a sequence and running BLAT**

Go to plants.ensembl.org.

BioMart

The screenshot shows the Ensembl Plants homepage. Callouts highlight the following features:

- Search**: A search bar at the top right and a larger search box on the left with a "Go" button.
- Drop-down list of species**: A callout pointing to the "All species" dropdown menu in the search box.
- Arabis thaliana (TAIR10)**: A callout pointing to the *Arabidopsis thaliana* (TAIR10) icon in the "Popular genomes" section.
- Overview**: A callout pointing to the "Overview" section on the right, which provides information about the bread wheat component genomes.
- Full list of species**: A callout pointing to the "Full list of species" link in the "What's New in Release 22" section.
- Information on Ensembl Plants**: A callout pointing to the "Information on Ensembl Plants" section on the right, which includes details about the transPLANT consortium and the BBSRC.

Click on the *Arabidopsis thaliana* icon to open its main homepage and type 'ap1' into the search bar and click the Go button.

One gene matches the query in *Arabidopsis*.
Links to the Gene tab, Location tab and Gene trees are provided.

The screenshot shows the search results page for 'AP1' in *Arabidopsis thaliana*. Callouts highlight the following features:

- Link to the Location tab**: A callout pointing to the "Location" tab in the search results.
- Link to the Gene trees**: A callout pointing to the "Gene trees" link in the search results.

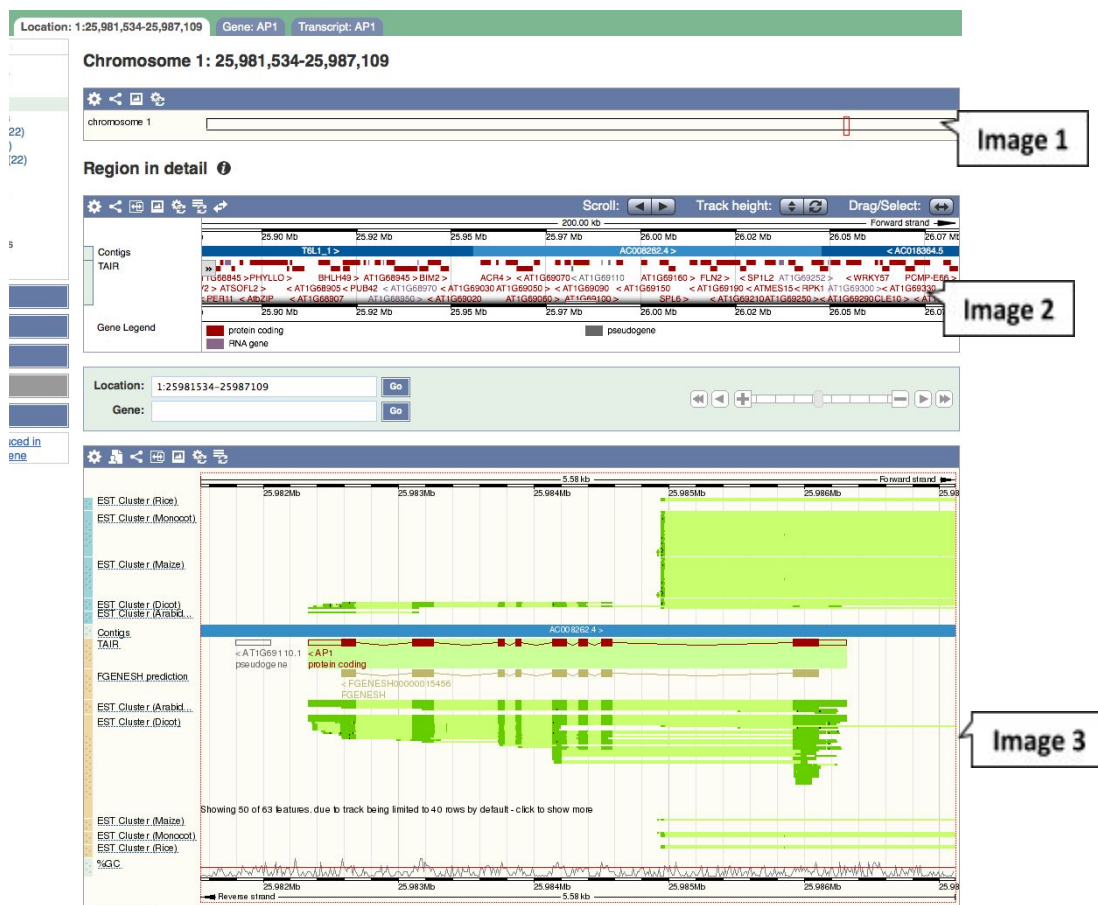
The search results show 1 gene found in *Arabidopsis thaliana*. The gene is **AP1 [AT1G69120]**. The description is "K-box region and MADS-box transcription factor family protein [Source:TAIR;Acc:AT1G69120]". The gene ID is [AT1G69120](#). The species is [Arabidopsis thaliana](#). The location is [1:25982330-25986313](#). The synonyms are AGAMOUS like 7, AGL7, APETALA1, F4N2.9, FLORAL HOMEOTIC PROTEIN APETALA1. The gene trees are [EPIGT00780000088721](#) (Plants Compara) and [EGGT00050000000043](#) (Pan-taxonomic Compara).

Let's view the genomic region in which this gene is located by clicking on the Location link. The Location tab should open.



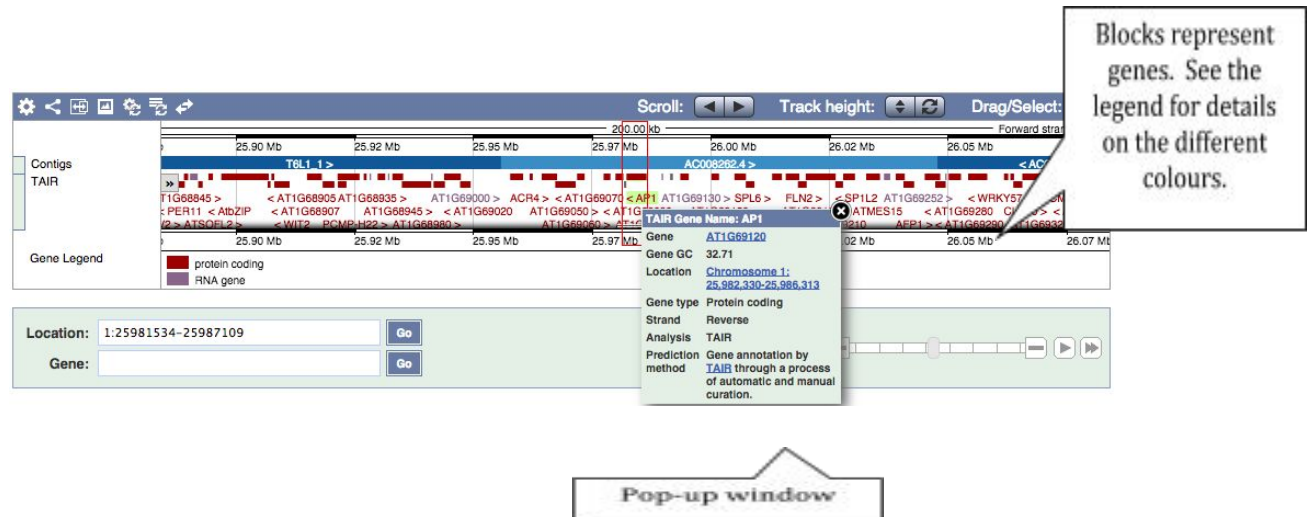
The Location tab in Ensembl is also known as Region in detail view. There is a help video on this page at <http://youtu.be/tTKEvgPUq94> (Please note this video is on the Location tab of the Ensembl sister browser using human as an example. The majority of the data will differ but the way to navigate through the Location tab will be exactly the same).

The 'Location tab' contains three images.

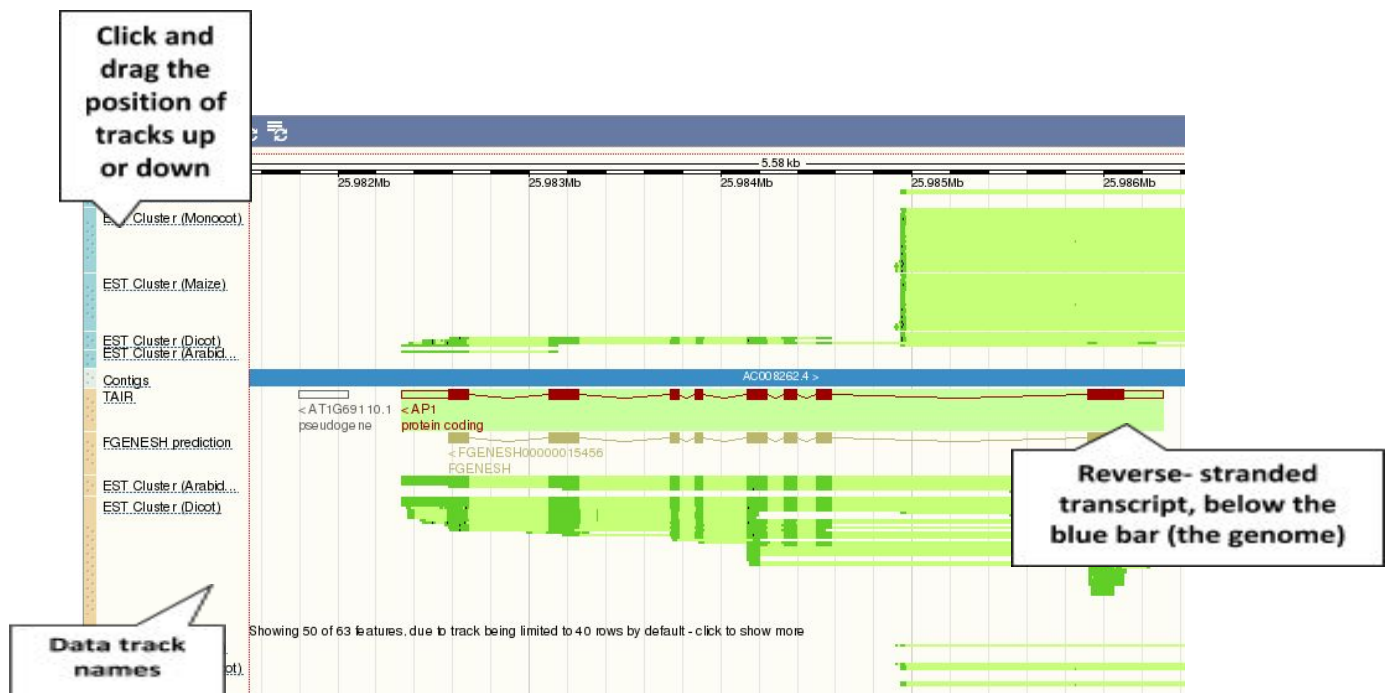


The first image shows an overview of the chromosome where the AP1 gene is located, chromosome 1.

The second image shows a more detailed view (200 kb long) of the region where AP1 gene is located and its neighbouring genes. Click on the gene for a pop-up window with additional information.



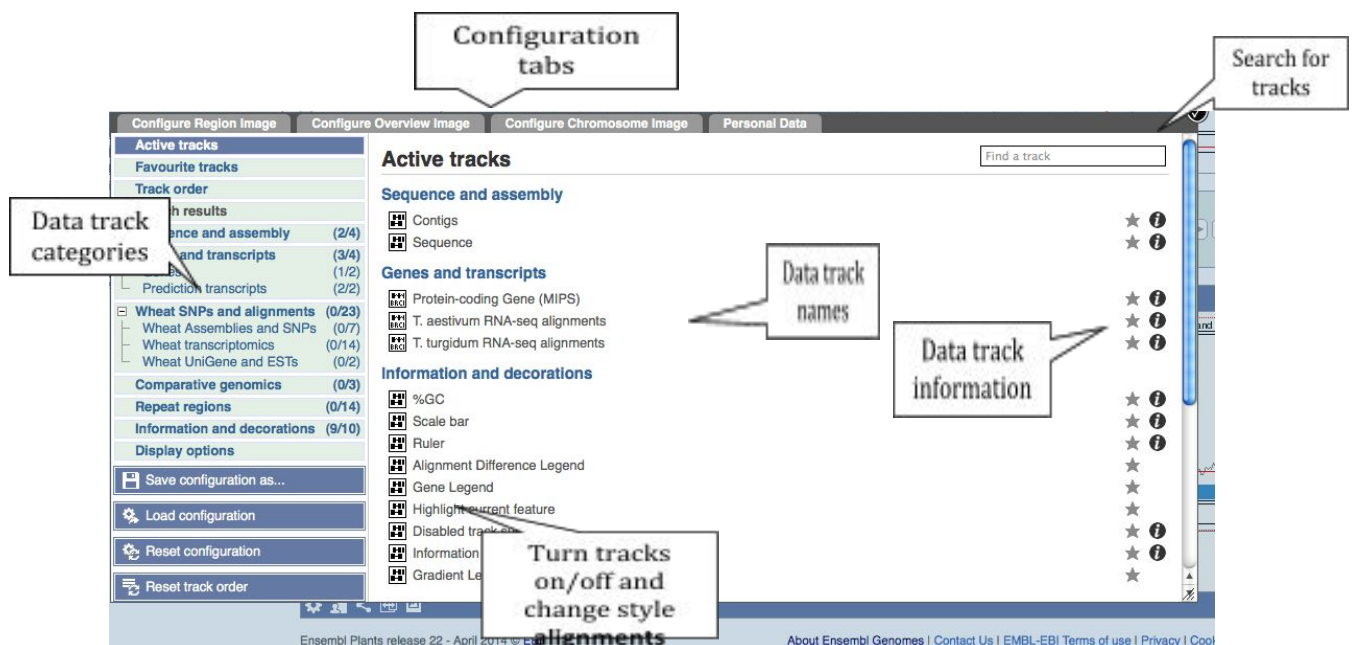
The third and final image is a detailed and highly configurable view of the region.



You can edit what you see on this page by clicking on the blue 'Configure this page' menu at the left of the page, or click on the cog wheel in the image itself. You can add different types of data available in Ensembl Plants.



This will open a menu that allows you to change the image. The menu will look like the image below:

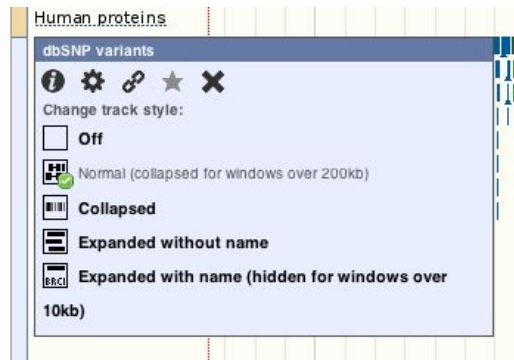


Let's add some tracks to this image, such as:

- Sequence variants – Normal
- All repetas

Now click on the tick in the top left hand to SAVE and close the menu with the newly added configuration. Alternatively, click anywhere outside of the menu.

We can also change the way the data track style appear by hovering over the track name, then the cog wheel to open a menu.



For more details on the different track styles, have a look at our FAQ <http://ensemblgenomes.org/node/30435>

Have a look at the changes in the Location tab (or Region in detail view). Click and drag tracks to reorder them, if it helps with comparing the data. You may also want to delete a few of the tracks that were on by default.



This view can be zoomed in and out.

Now that you've got the view how you want it, you might like to show this image to a colleague or collaborator. Click on the Share this page button to generate a link and you can email the link. They

will see exactly the same view as you, including all the tracks you have added and in the order you have them.



To return to the default view, go to 'Configure this page' and select Reset configuration at the bottom of the menu.

You can also reset the track order.

Alternatively, just click on the icons at the top of the image to reset tracks and/or track order.



Let's now explore the Gene tab.



We will walk you through some of the links in the left hand navigation column. Note that the left hand side menu in the Gene tab differs from the one we saw previously in the Location tab.

25,981,534-25,987,109 Gene: AP1 Transcript: AP1

Gene: AP1 AT1G69120

Description: K-box region and MADS-box transcription factor family protein [Source: TAIR AT1G69120](#)

Synonyms: AGL7

Location: [Chromosome 1: 25,982,330-25,986,313](#) reverse strand.

Transcripts: This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	Length	Protein	Biotype	RefSeq	Flags
AP1	AT1G69120.1	1228 bp	256 aa (view)	protein coding	NM_105581 NP_177074	

Summary ⓘ

Name: AP1 (TAIR Gene Name)

UniprotKB: This gene has proteins that correspond to the following Uniprot identifiers: [P35631](#)

Gene type: Protein coding

Prediction Method: Gene annotation by [TAIR](#) through a process of automatic and manual curation.

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

Gene Legend:

- Import / Other
- protein coding
- RNA gene
- pseudogene

How can we view the genomic sequence of my gene?

Click Sequence at the left of the page.

e!EnsemblPlants

Arabidopsis thaliana ▼ Location: 1

Gene-based displays

- Summary
- Splice variants (1)
- Transcript comparison
- Supporting evidence
- Gene alleles
- Sequence**

Click Sequence

Marked-up sequence ⓘ

Page-specific help

Download sequence BLAST this sequence Search Ensembl Genomes with this sequence

Key:

Exons AP1 exons All exons in this region

>chromosome:TAIR10:1:25981730:25986913:-1

AAAATACTATTTTGGGTTTGAAATTTTGAATACTTACAATTATTCTTCTCGATCTTCCT
CTCTTTCCTTAAATCCTGCGTACAAATCCGTCGACGCAATACATTACACAGTTGTCAATT
GGTTCTCAGCTCTACCAAAAACATCTATTGCCAAAAGAAAGGTCTATTTGTAAGTCACTG
TTACAGCTGAGAACATTAAATATAATAAGCAAATTTGATAAAACAAAGGGTTCTCACCTT
ATTCCAAAAGATAGTGTAAATAGGGTAATAGAGAAATGTTAATAAAAGGAAATTAAAA
TTTGGTTGGTTCAGATTTTGTTCGTAGATCTACAGGGAATCTCCGCCGTC
AAAGCGAAGGTGACACTTGGGGAAGGACCAGTGGTCCGTACAATGTTACTTACCC
TCTTCACGAGACGTCGATAATCAAATTGTTTATTTTCATATTTTAAAGTCCGCAG
TTTTTATAAAAAATCATGGACCCGACATTAGTACGAGATATACCAATGAGAAGTCGACAC
GCAATCCTAAAGAAACCACTGTGGTTTTTGCAAACAAGAGAAACCAGCTTTAGCTTTTC
CCTAAACCCTCTTACCCAAATCTCTCCATAAATAAAGATCCCGAGACTCAAACACAAG
TCTTTTTATAAGGAAAGAAAGAAAACTTTCTAATTGGTTCATACCAAGTCTGAGCT
CTTCTTTATATCTCTTGTAGTTTCTTATTGGGGGTCTTGTGTTTGGTTCTTTTA
GAGTAAGAAGTTTCTTAAAAAAGGATCAAAAATGGGAAGGGGTAGGGTTCAATTGAAGAG
GATAGAGAACAAGATCAATAGACAAGTGACATTCTCGAAAAGAAGAGCTGGTCTTTTGAA
GAAAGCTCATGAGATCTCTGTTCTCTGTGATGCTGAAGTTGCTCTGTTGTCTTCTCCCA
TAAGGGAAAACCTCTCGAATACTCCACTGATTCTTGTAACCTCAACTAATCTTTACTT
TTAAAAAATCTTTAATCTGCTACTTTATATAGTTTTTCCCCCTTAAGT
GATTTGCCCTAATTATCTACTGCTTTTGTATATATTTTCTAGGCT
GGATTTTTTGATTAGCCAGAAAAATGTTAATACAAATTGTATAATTTAA

Upstream sequence

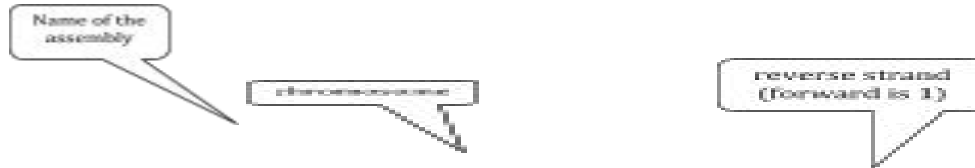
exon

intron

Click on the button ⓘ to view page-specific help.

The help pages also provide links to 'Frequently Asked Questions', a Glossary, Video Tutorials, and a form to Contact HelpDesk.

The sequence is shown in FASTA format. Take a look at the FASTA header:



```
>chromosome:TAIR10:1:25981730:25986913:-1
```

In the Gene tab, you can find all the GO and PO terms associated with this gene either as a table format or as a chart.

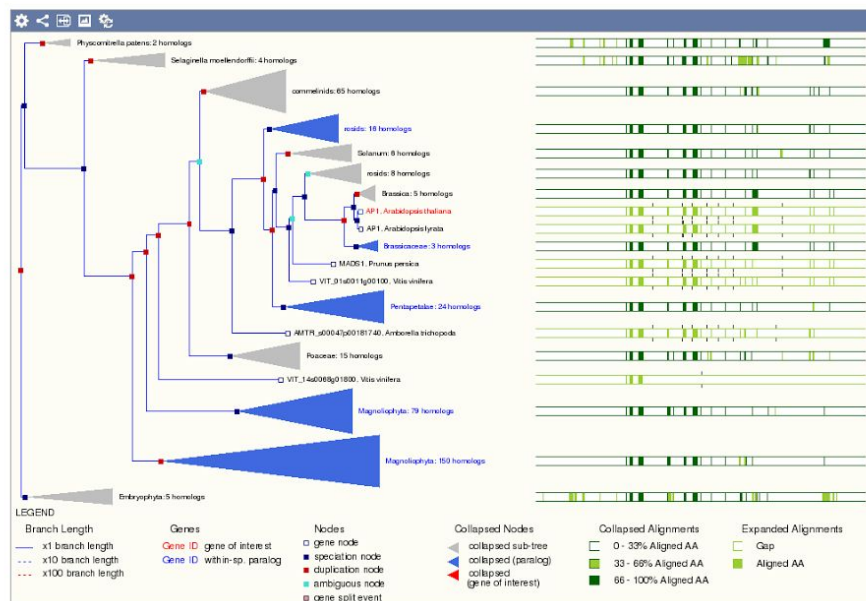
The screenshot shows the Ensembl Plants website interface. The top navigation bar includes the Ensembl Plants logo, a dropdown menu for "Arabidopsis thaliana", and a location bar showing "Location: 1:25,981,534-25,987,109". The "Gene: AP1" tab is selected. On the left, a "Gene-based displays" sidebar lists various options, with "Sequence" and "Ontology" expanded. Under "Ontology", "GO: biological process (21)" is highlighted. On the right, the "Gene: AP1 AT1G69120" section displays a table of gene information:

Gene	AT1G69120
Description	K
Synonyms	A
Location	C
Transcripts	T

Below this table, the "GO: biological process" section is shown, with a button for "Ancestry chart". The text "The following terms describe the bio..." is partially visible.

Let's now view some Comparative Genomics displays, which compare multiple species in Ensembl.

Click on Gene tree (image), which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.




Click the Orthologues link at the left of this page to view homologues detected by this tree (between the *Arabidopsis* gene and other plant genomes). Note the links under 'Compare'.

Compare						
<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 						
Selected orthologues						
View protein alignments of all orthologues Download all protein sequences Download all Dn						
Species	Type	dN/dS	Ensembl identifier & gene name	Compare		
Aegilops tauschii	Many-to-many	0.00651	F775_01779 Novel Ensembl prediction Dehydration-responsive element-binding protein 2G [Source: UniProtKB/TrEMBL; acc: M6BIZ4]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 		
Amborella trichopoda	1-to-many	n/a	AMTR_s00023p00044590 Novel Ensembl prediction hypothetical protein	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 	AmTr_v1.0_scaffold00023:391938-393086:1	27 41
Arabidopsis lyrata	1-to-many	n/a	scaffold_402826.1 Novel Ensembl prediction Putative uncharacterized protein [Source: UniProtKB/TrEMBL; acc: D7LF31]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 	4:19704296-19705389:-1	23 32
Arabidopsis thaliana	Many-to-many	n/a	AT2G40340 DREB2C Integrase-type DNA-binding superfamily protein [Source: TAIR_LOCUS; acc: AT2G40340]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 	2:16848438-16850487:-1	25 33


You can also view the Pairwise Whole Genome alignments between *A. thaliana* and 22 other species in Ensembl Plants. See an example below:

Genomic alignments

Alignment:

 Download alignment

[Go to a graphical view of this alignment](#)

 Species Tree

No tree is drawn for pairwise alignments

Arabidopsis thaliana › [chromosome:TAIR10.1:25982270:25986373:-1](#)
 Arabidopsis lyrata › [chromosome.v.1.0.2:13573131:13577421:-1](#)

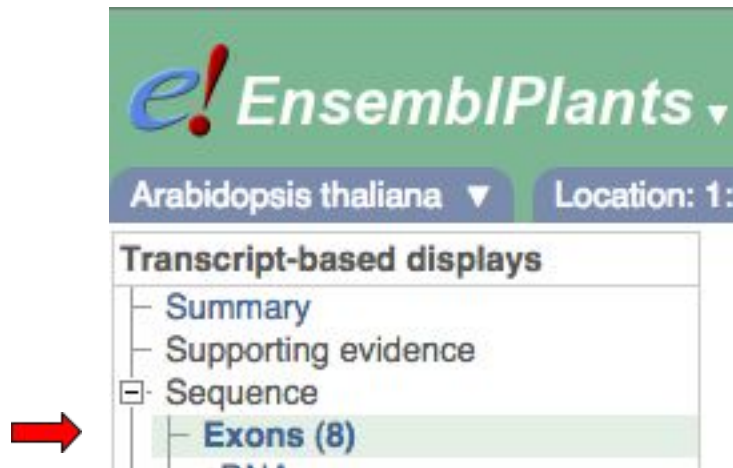
Arabidopsis thaliana	GCAAATCCTAAAGAAACCACTGTGGTTTTTGCAAACAAGAGAAACCACTTTAGCTTTTCCTAAAACCACTCTTACCCAA
Arabidopsis lyrata	GCAAATCCTAAAGAAACCACTATGGTTTTTGCAAACAAGAGAAACCACTTTAGCTTTTCCTAAAACCACTCTTACCCAA
Arabidopsis thaliana	TCTTTTATAAAGGAAGAAAGAAAACTTTCCTAATTGGTTCATACCAAAGCTGAGCTCTTCTTTATATCTCTCTGTAG
Arabidopsis lyrata	TCTTATTATAAAGGAAGAAAGAAAACTTTCCTAATTGGTTCATACCAAAGCTGAGCTCTTCTTTATATTACTCTGTAG
Arabidopsis thaliana	GAGTAAGAAGTTTCTTAAAAAGGATCAAAATGGGAAGGGTAGGGTTCAATTGAAGAGGATAGAGAACAAGATCAATAGA
Arabidopsis lyrata	GAGTAAGAAGTTTCTTAAAAAGGATCAAAATGGGAGGGTAGGGTTCAATTGAAGAGGATAGAGAACAAGATCAATAGA
Arabidopsis thaliana	GAAAGCTCATGAGATCTCTGTTCTCTGTGATGCTGAAGTTGCTCTTGTGTCTTCTCCATAAGGGGAAACTCTTCGAATAC
Arabidopsis lyrata	GAAAGCCCATGAGATCTCTGTTCTCTGTGATGCTGAAGTTGCTCTTGTGTCTTCTCCATAAGGGGAAACTCTTCGAATAC
Arabidopsis thaliana	TTTTAAAAAATCTTTTAATCTGCTACTTTATATAGTTTTTCCCCCTTAAGTTGACTACTTGA-TTTGCCCTAATTATTCA
Arabidopsis lyrata	TTTTAAAAAATC-TTTGATTGCTACTTTATCTCGTTTTTCCCCCTTAAGTTGACTACTTGATTTTGCCCTAATTATTAA

Let's now move to the Transcript tab to explore the splice isoform of our AP1 gene by clicking on the ID of the only transcript annotated in that locus (AT1G69120.1).

 **EnsemblPlants** ▾ [Sequence Search](#) | [BLAST](#) | [BioMart](#) | [Tools](#) | [Downloads](#)

Arabidopsis thaliana ▾ [Location: 1:25,981,534-25,987,109](#) [Gene: AP1](#) **[Transcript: AP1](#)**

You are now in the Transcript tab. The left hand navigation column provides several options for this transcript, such as 'Exons', 'cDNA' and 'Protein summary'.



The exon view is shown below:

The screenshot shows a DNA sequence with color-coded regions. A callout box labeled 'Purple: UTR' points to the first few bases. A callout box labeled 'Green: flanking sequence' points to the sequence between the UTR and the coding sequence. A callout box labeled 'Black: coding sequence' points to the main body of the sequence. A callout box labeled 'Blue: Intron sequence' points to the sequence between the coding sequence and the next exon.

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

Display options

Flanking sequence at either end of transcript:

50

Number of base pairs per row:

60 bps

Intron base pairs to show at splice sites:

25

Show full intronic sequence:

☒

Show exons only:

☐

You can download this sequence either as FASTA or RTF (Rich Text Format) , the latter will include the colours and marking up displayed in the view.

Download sequence

File name:

Arabidopsis_thaliana_API_sequence

File format:

-- Choose Format --

Output:

☒ Uncompressed
 ☐ Gzip

Select "uncompressed" to get a preview

Guide to file formats

FASTA

Text sequence(s):
DNA and/or amino acids

```
>11 dna:chromosome chromosome:GRCh38:11:10:
ZAGCGCGAAGCCCAAGCGGCATCCCTAGTAGGGCTACTTGC
TCTGGCCCTCAGAAAGAAATCTCCCAACATTTCAGTTGGC
TCCAAATATGAGCAGCCTCAGGCGCTACGCCCTGCTTACG
TCTCAATCCCTGTAGACTTACCCTCCGCCGCCGCTGAC
```

RTF

Marked-up sequence,
with or without variants

```
ATTAGCAACAAAAAGCAAAACACGGG
GAGTCTCTTCCACAAACATGGGCAT
TCTTAGGGAGTTAGAATATTGATGG
TTTTTAGGGTAATGTGGCTTCCGT
AGGCCCTCACAAATTCTGCCAAGTC
TTTTCGTTTCCGCACCTGGGACCTC
GCTGGGTCATGTGGAGCTGATGCTT
```

Now click on the cDNA link to see the spliced transcript sequence.

Key:

Codons	Alternating codons	Alternating codons		
Exons	Alternating exons	Alternating exons		
Variations	3 prime UTR	5 prime UTR	Missense	Synonymous
Other	UTR			

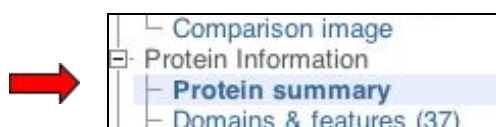
```

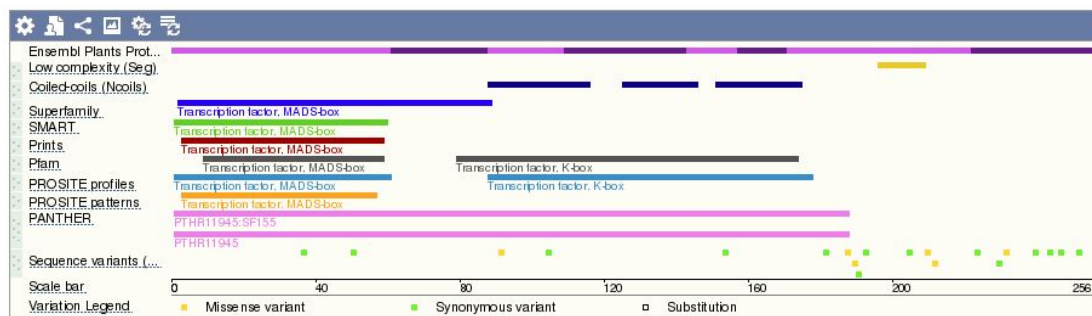
1 CCTAAAACCACTCTTACCCAAATCTCTCCATAAAATAAGATCCCGAGACTCAAACACAAG
.....
.....
61 TCTTTTATAAAGGAAAGAAAGAAAATCTTCCTAATTGGTTCATACCAAAGTCTGAGCT
.....
.....
121 CTTCTTTATATCTCTCTGTAGTTTCTTATGCGGGTCTTTGTTTTGTTTGGTTCTTTTA
.....
.....
181 GAGTAAGAAGTTTCTTAAAAAAGGATCAAAAATGGGAAGGGGTAGGGTTCAATTGAAGAG
.....ATGGGAAGGGGTAGGGTTCAATTGAAGAG
.....-M--G--R--G--R--V--Q--L--K--R
241 GATAGAGAACAAGATCAATAGACAAGTGACATTCTCGAAAAGAAGAGCTGGTCTTTTGAA
30 GATAGAGAACAAGATCAATAGACAAGTGACATTCTCGAAAAGAAGAGCTGGTCTTTTGAA
10 --I--E--N--K--I--N--R--Q--V--T--F--S--K--R--R--A--G--L--L--K

```

UnTranslated Regions (UTRs) are highlighted in yellow, codons are highlighted in light yellow, and exon sequence is shown in black or blue letters to show exon divides. Sequence variants are represented by highlighted nucleotides and clickable IUPAC codes are above the sequence.

Click on Protein summary to view domains from Pfam, PROSITE, Superfamily, InterPro, and more.





Clicking on Domains & features shows a table of this information.

Prosite_profiles	768	825	-	PS50313	-
Smart	35	64	Ankyrin_rpt	SM00248	IPR002110 [Display all genes with this domain]
Prosite_profiles	69	93	Ankyrin_rpt	PS50088	IPR002110 [Display all genes with this domain]
Smart	69	99	Ankyrin_rpt	SM00248	IPR002110 [Display all genes with this domain]
Prosite_profiles	103	127	Ankyrin_rpt	PS50088	IPR002110 [Display all genes with this domain]

Our last task is to export genomic sequence and perform a similarity search using Ensembl Blast. To export a sequence, let's go back to the Location tab and click on the Export data option, select the default parameters (e.g. Fasta sequence as output) and click Next then HTML.

This is a snapshot what you will see in your browser:

Plants
Sequence Search | BLAST | BioMart | Tools | Downloads | Help | Doc

Location: 1:25,981,534-25,987,109
Gene: AP1
Transcript: AP1

s
y
s
(22)
2)
(22)
|
rs

Export Location Data

```

>l dna:chromosome chromosome:TAIR10:1:25981534:25987109:1
AATAAAAAGGAATTATATATCTGTCATTTCTATCTATACTCAAAACGAATTAATTAATAA
CTCATATATCTGACTATCTCAAAAATCAAAATTATGCTTTGCTTTCTTCTCTTCATT
ATGCATTTAATTCTGCTCTGTTGACTTCTTCAACAATCCTCTATCCAAAGAAGGAACTT
CTTCAAGAACCACAAAAATATAAAGTTGTAGTAACTGAAACATGAAGACCAACAATATAA
TTACCAAAATGTTTAATACAGAAATTGTTTTCTAAAAAAATTAGCAGGAATACAACGAAG
AGAAGATAATGTGAGAGTTACGCTTACCACAACGTGCCTTTTCTTGAGTTAGCAACATT
TTACTGAATGTAGATGAAGAGAAGATAATGTGAGAGTTACGCTTACCACAACGTGCCTTT
CTGGTGGTGATATATCTTAAGAACCTTAGTGGGCATTTTCACTGGTCCCTTTACTGTAAG
TCTCGTGTCTTAGCTCCATGAAACAAGTCAATACACACTAATTATTCGATCACTAATCA
AACAAATGTATCAGAGAGACTACCAGGGTGTTACATTACAGATCCTCATCCATTCCGAGG
CACTGTTATTATTGCCTATCTTATCTAATTAGAAGCAAATTCTGGGCCCTTTATAGGCCAA
CATCCTTCCTTGATTCTAATTGGGTTCTATTCCGATTAATACGGCCCAACCACTCAAGTT
CGCTTGTTGTAACTTGAAATAATAAGAAATTTTGCCACTCCTAACTTAATGCAACAAAA
AAAAAAAAAGGTCCGATCTATTAAGTTCACGTTCACTCTCTGACCTTCAAATATATTA
CAATTTTGCTCCATATTGAAGCAAATAAACTTATCAAATTACAAAAGAAGAAAGGAGCCT
GCTAATTTATATATGATGATATAAGAACATCGAACATTTGCCAAAATATATTAATTGGAT
GAAAAGAGCCTAGCCACTATTTATATGTATGTGGCAAAGTGTATTTTATTGTTGACGATTA
CAAATATATATATGGAAATGCTTCATGCGGCGAAGCAGCCAAGGTTGCAGTTGTAAACGG
GTTCAAGAGTCAGTTCGAGATCATTCCTCCTCATTGCCATAGGATCATCTTCTTGATACA
GACCACTGCATTCACAATCATACATAGGATATATATGAAAGTGCATATGTAATCATATGT

```

To use this sequence for similarity searches, you can select the header and a few lines of the nucleotide sequence and then copy it on the clipboard. Click on the BLAST link in the bar at the top of the main Ensembl Plants homepage. Paste the sequence into the appropriate box

Create new ticket:

[Add more sequences](#) (1 sequence added, 29 more sequences allowed)

make sure which species you want to perform the search against, select/change the appropriate parameters and click on Run.

Run > Clear

The table with results will look like the following example. The jobs are given a ticket number and highlighted in green when successfully completed.

BLAST search

[New Search](#)

Recent Blast tickets: 

[Refresh](#)

Show/hide columns (1 hidden)		Filter
Analysis	Jobs	Submitted at
BLASTN	1 dna:chromosome chromosome:TAIR10:1:25982330:25986313:-1 Done: 24 hits found View results	14/01/2015, 13:47   

Click on 'View results' to go to a page like this one:

Results for 1 dna:chromosome chromosome:TAIR10:1:25982330:25986313:-1

[Job details](#) 

Job name 1 dna:chromosome chromosome:TAIR10:1:25982330:25986313:-1



Species  Arabidopsis thaliana

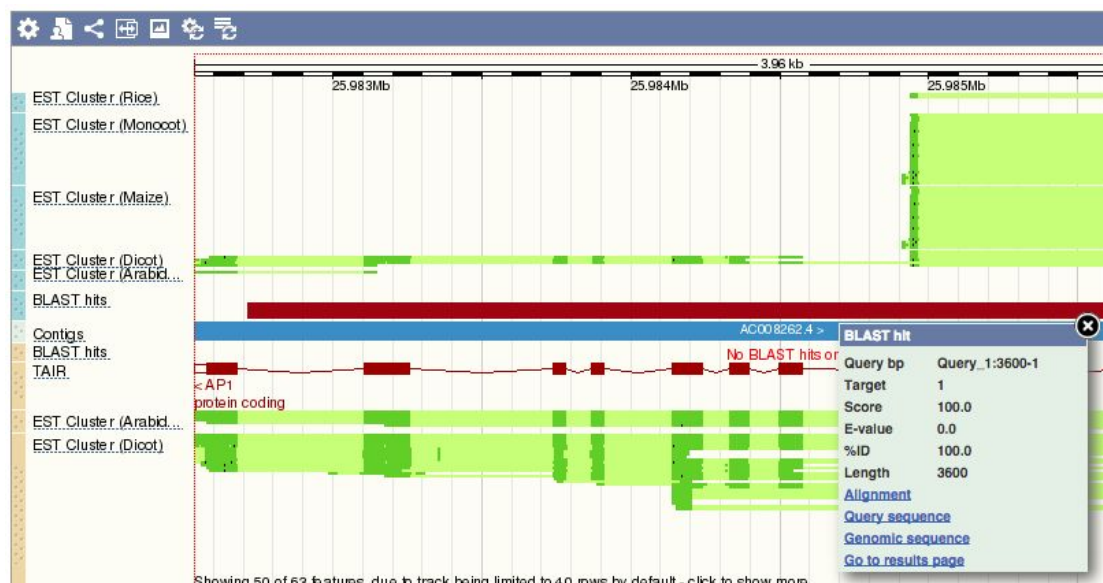
Search type BLASTN (WU BLAST)

[Download results file](#)

[Results table](#) 

Show	All	entries	Show/hide columns							Filter
Genomic Location	Orientation	Query name	Query start	Query end	Query ori	Length	Score	E-val	%ID	
1:25982714-25986313 [Sequence]	Forward	Query_1 [Sequence]	3600	1	Reverse	3600	100.0	0.0	100.0 [Alignment]	
4:14432675-14432734 [Sequence]	Forward	Query_1 [Sequence]	3315	3377	Forward	63	71.0	4.4E-20	71.0 [Alignment]	
5:24505733-24506060 [Sequence]	Forward	Query_1 [Sequence]	485	162	Reverse	332	71.0	1.2E-27	71.0 [Alignment]	
2:1129567-1129878 [Sequence]	Forward	Query_1 [Sequence]	152	472	Forward	325	66.0	6.6E-20	66.0 [Alignment]	
Pt:120603-120688 [Sequence]	Forward	Query_1 [Sequence]	1428	1349	Reverse	86	65.0	0.019	65.0 [Alignment]	
2:9703679-9703812 [Sequence]	Forward	Query_1 [Sequence]	2840	2972	Forward	140	63.0	6.6E-20	63.0 [Alignment]	

You can click on the first location under the Genomic location column to be directed to the Region in detail view:

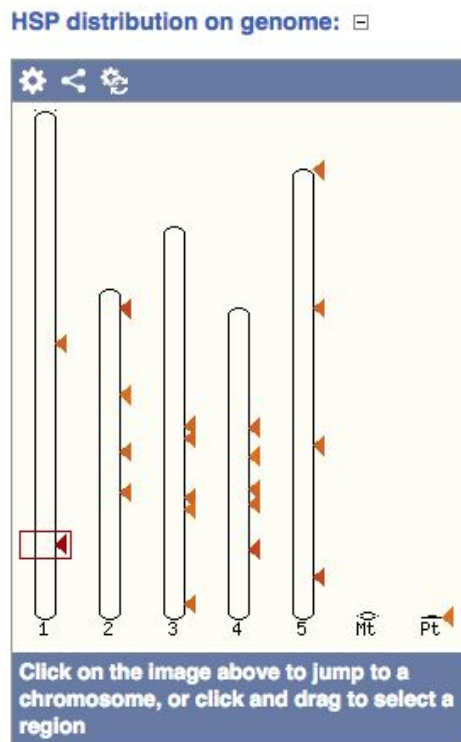


Click on the red bar for the score, %ID, and other BLAST values.

Export Image for your lab notebook or publications, or Share it with your colleagues and collaborators!



From the results page of your BLAST, if you scroll down you will see the results on the karyotype view (providing the karyotype of your species is available):



You may want to try the Sequence search tool (from ENA) and compare the results to BLAST.

Query Sequence

END OF OUR BROWSER WALKTHROUGH

EXERCISES

GENOME BROWSER

1) Exploring the bread wheat (*Triticum aestivum*) genome

Go to the Ensembl Plants homepage (<http://plants.ensembl.org>).

- a) What is the current release (version) of Ensembl Plants?
- b) How many coding genes have been annotated in the current assembly of wheat (*Triticum aestivum*), i.e. assembly IWGSC2?
- c) Go to Location tab of IWGSC_CSS_4DS_scaff_2304216:8417-57876 in this species. How many genes can you see in this region? Are there ncRNA genes annotated in this region? What is the prediction method used to annotate each of them?
- d) Click on the longest gene and go to the Gene tab. What is the length in nt and aa for the transcript and translation of this gene? What is the molecular function of the product of this gene according to the Gene Ontology? Which protein domains have been mapped to the translation of this gene?

2) The *ATHDH* gene in *Arabidopsis thaliana* and its orthologues in other plants

- a) What are the genomic coordinates of the *ATHDH* gene in *A. thaliana*? Is this gene on the forward or reverse strand?
- b) How many orthologues have been identified for this gene?
- c) Can you find the 'Alignment (cDNA)' between this gene in *Arabidopsis* and its counterparts in tomato (*Solanum lycopersicum*),

rice (*Oryza sativa* Japonica), bread wheat (*Triticum aestivum*), and maize (*Zea mays*)?

d) Why are there three entries for this gene in bread wheat? Click on any of the three different locations of these orthologues to see the region where the gene has been annotated. You may want to explore a view in the Location tab called Polyploid view. Please note more details on the Comparative Genomics analyses in Ensembl Plants and the polyploid nature of bread wheat will be covered in the afternoon session of this workshop.

3) miRNA genes in *A. thaliana*

MicroRNAs (miRNAs) are small non-coding RNA molecules (ca. 22 nucleotides) found in plants and animals, which function in transcriptional and post-transcriptional regulation of gene expression. A well-studied miRNA family in plants is the MIR395 family (See also: <http://en.wikipedia.org/wiki/MicroRNA> and http://en.wikipedia.org/wiki/Mir-395_microRNA_precursor_family).

a) How many members does the MIR395 family in *Arabidopsis thaliana* have?

b) How are the MIR395 genes organised? Are they clustered? Are they all located on the same strand of the genome? How are they positioned relative to each other?

4) Would you have a favourite gene/genomic region you want to explore with Ensembl Plants?

Feel free to discuss your findings or ask one of the instructors for help.

BIOMART

1) Export sequences in FASTA format from the bread wheat (*Triticum aestivum*) genome

Retrieve the protein sequences (in FASTA format) of all wheat genes that have an EntrezGene ID, that are protein coding and that contain signal peptide domains. Do a count after selection of each filter to check the number of genes remaining in your dataset. Export the results of the sequences and select 'Gene description' and 'Source of gene name' as headers.

2) Convert UniProt IDs into Ensembl IDs for *Arabidopsis* proteins

BioMart is a very handy tool when you want to map between different databases.

The following is a list of IDs from the UniProtKB/Swiss-Prot database

(<http://www.uniprot.org/>) of *Arabidopsis thaliana* proteins that are believed to be involved in flavonoid metabolism

http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0009812:

P42813, Q9LS08, Q9ZST4, Q9SYM2, P51102, Q9LPV9, Q9FE25, Q96323, Q9FKW3, P13114, P41088, Q9S818, Q96330, O22203, Q39224, O22264, Q9SD85, Q9LYT3, Q9FJA2, Q43128, P43254, O04153, Q43125, Q9S9P6, Q94C57, Q9LNE6, Q9FK25, Q9SYM5, Q9ZQ95

Generate a list that shows, to which Ensembl Gene IDs these UniProtKB/Swiss-Prot IDs map to. Also include the Gene name, Gene description and Pfam ID.

3) Retrieve a list of SNPs from the tomato genome (*Solanum lycopersicum*)

The region between coordinates 21,394,819 and 21,397,868 on chromosome 6 in tomato contains a gene involved in oxidation-reduction process (GO:0055114).

Can I use BioMart to retrieve all the SNPs that cause a change at the amino acid level of this gene (those SNPs are known as missense variants) including their IDs and possible alleles?

4) Retrieving all genes that contain a given Pfam domain in tomato and maize

The disease resistance (R) genes in plants e.g. TIR-NBS-LRR genes code for proteins that contain an N-terminal Toll/Interleukin receptor homology region (TIR), a nucleotide binding site (NBS) and a C-terminal leucine rich repeat (LRR). TIR-NBS-LRR genes are common in dicots but seem to be rare in monocots (Tarr and Alexander. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. BMC Res Notes 2009 Sep 8;2:197).

The ID for the TIR domain in the Pfam (protein family) database is PF01582 (<http://pfam.sanger.ac.uk/family/PF01582>).

Generate a list of all potato (*Solanum lycopersicum*) genes that are annotated to contain a TIR domain. Include the Ensembl Gene ID and description. Do the same for maize (*Zea mays*).

Do your results confirm the findings of Tarr and Alexander?

5) You may want to give BioMart a go to retrieve your favourite genomic data from Ensembl Plants.

Design your own query. You may want to try a different Ensembl Plants database in BioMart, i.e. Variation Mart.

Watch our BioMart tutorial video on youtube:

<http://youtu.be/DXPaBdPM2vs>

For more details on BioMart, have a look at these publications:

Smedley, D. *et al.*

BioMart – biological queries made easy

BMC Genomics 2009 Jan 14;10:22

Kinsella, R.J. *et al.*

Ensembl BioMarts: a hub for data retrieval across taxonomic space.

Database (Oxford) 2011:bar030

GENETIC VARIATION

Exercise 1 – Exploring a SNP in *Arabidopsis*

The *Arabidopsis ATCDSP32* gene is chloroplastic drought-induced stress protein of 32 kD (*ATCDSP32*) proposed to participate in a process called cell redox homeostasis).

- a) How many variants have been identified in the gene that can cause a change in the protein sequence?
- b) What is the ID of the variant that change the residue 60 from Alanine to Threonine? What is the location of this SNP in the *Arabidopsis* genome? What are its possible alleles?
- c) Download the flanking sequence of this SNP in RTF (Rich Text Format). Can you change how much flanking sequence is displayed on the browser?
- d) Does this SNP cause a change at the amino acid level for other genes or transcripts?
- e) What is the most frequent genotype at this locus in the ‘1001 Population’?

Exercise 2 – Variation data in the tomato (*S. lycopersicum*) genome

- a) Find the cytochrome P450 gene in tomato (also known as Solyc02g085360.2) and go to its Location tab. Can you add the data track that shows variation data from the ‘150 tomato genome resequencing project’ (TGRSP)?
- b) Zoom in around the last exon of this gene. What are the different types of variants seen in that region? What is the location of the only inframe deletion mapped in the region?
- c) Click on the splice region variant showed in that view. Why does Ensembl Plants put the G allele first in the string (G/A)?

Exercise 3 – Missense variants in the bread wheat genome

The transcript Traes_2AL_5C7E76139.1 is involved in calcium ion binding (GO:0005509). Around 45 variants have been mapped to this transcript.

- a) What are the two types of variants annotated in this transcript?
- b) Are there any variants predicted to be deleterious? Which amino acid residue is affected and what are the possible amino acids in that position?
- c) What are the different sources of EPITA06993600 and BA00258972?

Exercise 4 – The VEP tool, Variant Effect Predictor in the bread wheat genome.

An analysis of 5,000 individuals from two different populations of bread wheat (*T. aestivum*) has identified thousands of polymorphic loci. See a list of a few of them below:

chr 2D, genomic coordinate 89551917, alleles G/A, forward strand
chr 2D, genomic coordinate 148408765, alleles G/T, forward strand
chr 3D, genomic coordinate 113574123, alleles C/A, forward strand
chr 3D, genomic coordinate 93827883, alleles G/A, forward strand
chr 3B, genomic coordinate 727928129, alleles C/T, forward strand
chr 3B, genomic coordinate 736734474, alleles C/T, forward strand
chr 6A, genomic coordinate 196872409, alleles T/G, forward strand
chr 6A, genomic coordinate 196153918, alleles A/G, forward strand
chr 6A, genomic coordinate 196774882, alleles G/C, forward strand

Can you use the VEP tool to answer the following?

- a) Which genes and transcripts do these variants map to?
- b) Which consequence types can be found for these variants? Do any of them cause a change at the amino acid level?

COMPARATIVE GENOMICS

1) Orthologues, gene trees and pan-taxonomic compara

The fumarase gene (*FUM1*) in *Arabidopsis* encodes a protein with mitochondrial targeting information

(<http://www.uniprot.org/uniprot/P93033>)

a) How many orthologues have been identified for this gene in Ensembl Plants?

b) Which orthologues has the highest sequence similarity? Look at the Query% ID and Target%ID. For more details on what these terms mean, have a look at our FAQ:

<http://ensemblgenomes.org/node/30379>

c) Does this gene have an orthologue in human? Note: look for the Pan taxonomic Compara analysis?

d) Repeat the above for the plastocyanin (*PETE1*) gene in the same species. Few orthologues have been identified for this gene. Why could the reasons be for that? Have a look at the %ID for this gene versus *FUM1*?

2) Whole genome alignments and synteny in *Arabidopsis*

a) Go to the Location tab or 'Region in detail' page of the *Arabidopsis FUM1* gene and configure the page to turn on the genome alignments (known as BLASTz/LASTz alignments) against *A. lyrata* and *Zea mays*. Those tracks are under the 'Comparative genomics' menu in the configuration window. Does the degree of conservation between *Arabidopsis thaliana* and the other two plant species reflect their evolutionary relationships?

b) Stay on this same view but zoom it out till you can view a 30 kb region. Which genomic regions in the alignments between *A. thaliana* and maize are the most conserved? Did you expect this?

c) Now click on the 'Region Comparison' link on the left hand side menu in the Location tab, and add *Arabidopsis lyrata*, *Brassica rapa* and *Sorghum bicolor* by clicking on 'Select species or regions'. Configure the page and under 'Comparative features' turn on the 'Join genes' option.

d) Click on the Synteny link on the left hand side to view a map depicting syntenic blocks between *A. thaliana* and three other plants including rice.

3) The *HEMA2* gene in barley and its orthologues in triticeae genomes

a) On which chromosome is the *HEMA2* gene located in the barley genome (*Hordeum vulgare*)?

b) What are the locations of its paralogues? Can you find how much %coverage has been described for the known paralogues at the protein level? Note: look for the 'Alignment (protein)' link in the Paralogues table to find out the %coverage)

c) How many gene duplication events affecting barley and all other triticeae genomes can be identified in the gene tree available for *HEMA2* gene? Have these events affected other genomes, e.g. rice genomes?

d) Find the orthologue of the barley *HEMA2* gene in bread wheat, the only polyploid genome in Ensembl Plants. So there will be three bread wheat genes. Go to the gene page of any of them, e.g. Traes_1BS_ABD495014. Now click on the 'Homoeologues' link on the left hand side menu of the Gene tab in bread wheat to go to the homoeologue page (this is just available for polyploid species). How many homoeologues are listed? Is that consistent with what we expect for a hexaploid genome? How conserved are these homoeologues?

e) Let's now explore the polyploid view: click on 'View genomic alignments of all homoeologues' in the Gene tab. You can get to the same 'Polyploid view' from the Location tab.

