Tutorial

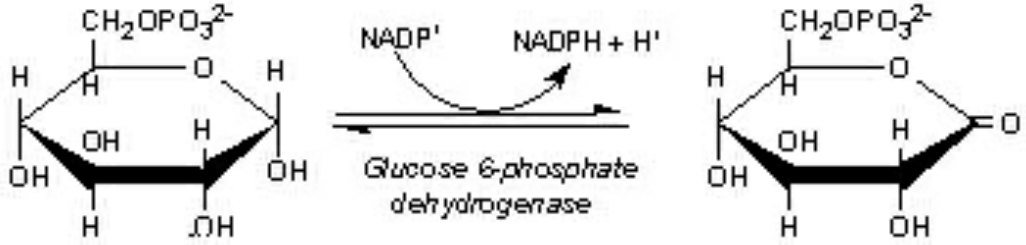
Tutorial objectives

After this tutorial you should be able to:

- Search and navigate the Ensembl Plants website.
- Understand Ensembl Plants annotation.
- How to attach and visualize your BAM and VCF data.
- Retrieve Ensembl Plants data using BioMart.
- Know where to find help and documentation.

Background: G6PD

Glucose-6-phosphate dehydrogenase (G6PD or G6PDH) is a cytosolic enzyme in the pentose phosphate pathway, a metabolic pathway that supplies reducing energy to cells by maintaining the level of the co-enzyme nicotinamide adenine dinucleotide phosphate (NADPH).



G6PD is widely distributed in many species from bacteria to humans. In higher plants, several isoforms of G6PDH have been reported, which are localized in the cytosol, the plastidic stroma, and peroxisomes.

- http://en.wikipedia.org/wiki/Glucose-6-phosphate_dehydrogenase

Search: All species for Go

e.g. Carboxy* or chx28

Search

Species pages

Info on current release

Popular genomes (Log in to customize this list)


Arabidopsis thaliana
TAIR10


Glycine max
V1.0


Oryza sativa
MSU6


Solanum lycopersicum
SL2.40


Zea mays
AGPv2


Physcomitrella patens
ASM242v1

All genomes

-- Select a species --

View full list of all Ensembl Plants species

What's in Release 14 (May 2012)

- New genomes for:
  - *Oryza brachyantha*,
  - *Setaria italica*, and
  - *Solanum lycopersicum*.
- Added variation data for for:
  - *Oryza glaberrima* and
  - *Zea mays*.
- Added wheat gene sequences alignments to *Brachypodium distachyon* as BAM tracks.
- Updated peptide and DNA ***compara***tive genomics databases.

Did you know...?

As part of the wheat genome analysis, we have aligned a set of identified gene wheat sequences against the *Brachypodium distachyon* genome. The alignments are available on the Location view, as additional BAM tracks under "Configure This page" menu.

Forthcoming training courses

A hands-on training workshop in plant pathogenic genomics, focused on the new PhytoPath and including coverage of Ensembl Fungi, Ensembl Protists and will be held at EBI in September, aimed at PhD students and post-doctoral rs who are working on all aspects of fungal and oomycete-induced disease To find out more details, and to register, please click here.

nts includes the newly sequenced *Solanum lycopersicum* ato) genome [1].

Tomato is a member of the Solanaceae family of plants that includes several other economically important species, such as potato, eggplant, petunia, tobacco and pepper [2]. It is one of the most popular fleshy fruit in the world, with 145 million tonnes being produced in 2010 [3]. It is a plant model for, fruit development, ripening, biotic and abiotic stresses, and cell wall biosynthesis [4]. Taxonomically, the tomato is the first plant in Ensembl Genomes to come from the asterid clade; the large majority of the eudicot species being either rosids or asterids [5]. For this reason, tomato has been included in our pan-compara [6] analysis (for example, see the Gene Tree for Solyc01g088200.2).

References

[1] International Tomato Genome Sequencing Project.
[2] Sequencing the Potato Genome: Outline and First Results.
[3] FAOSTAT
[4] Sol Genomics Network - Details for species Solanum lycopersicum.
[5] Rosids - Wikipedia
[6] Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species.

Ensembl Genomes

The Ensembl Genomes project produces genome databases for important species from across the taxonomic range, using the Ensembl software system. Five sites are now available: Ensembl Bacteria, Ensembl Protists, Ensembl Metazoa, Ensembl Plants and Ensembl Fungi. These new sites complement the existing Ensembl site, with its focus on vertebrate genomes. You can search all Ensembl and Ensembl Genomes databases from the search bar in the top right of this page.

Ensembl Genomes data is available through many of the same routes as Ensembl data. Data can be accessed via:

- this web browser (go to http://bacteria.ensembl.org, http://metazoa.ensembl.org, etc., or to http://www.ensemblgenomes.org for

Exercise 1

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

- *What is the current release (version) of Ensembl Plants?*
- *On which data are the genome sequence and gene annotation for Arabidopsis thaliana based?*

Configure this page

Manage your data

Export data

Bookmark this page

Side menu

Gene: DFR AT5G42800

Gene tab

Top panel stays the same as long as you stay on the same tab

Description dihydroflavonol-4

Location Chromosome 5: d.

Transcripts This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT5G42800.1	AT5G42800.1	1358	AT5G42800.1	382	Protein coding

Gene summary

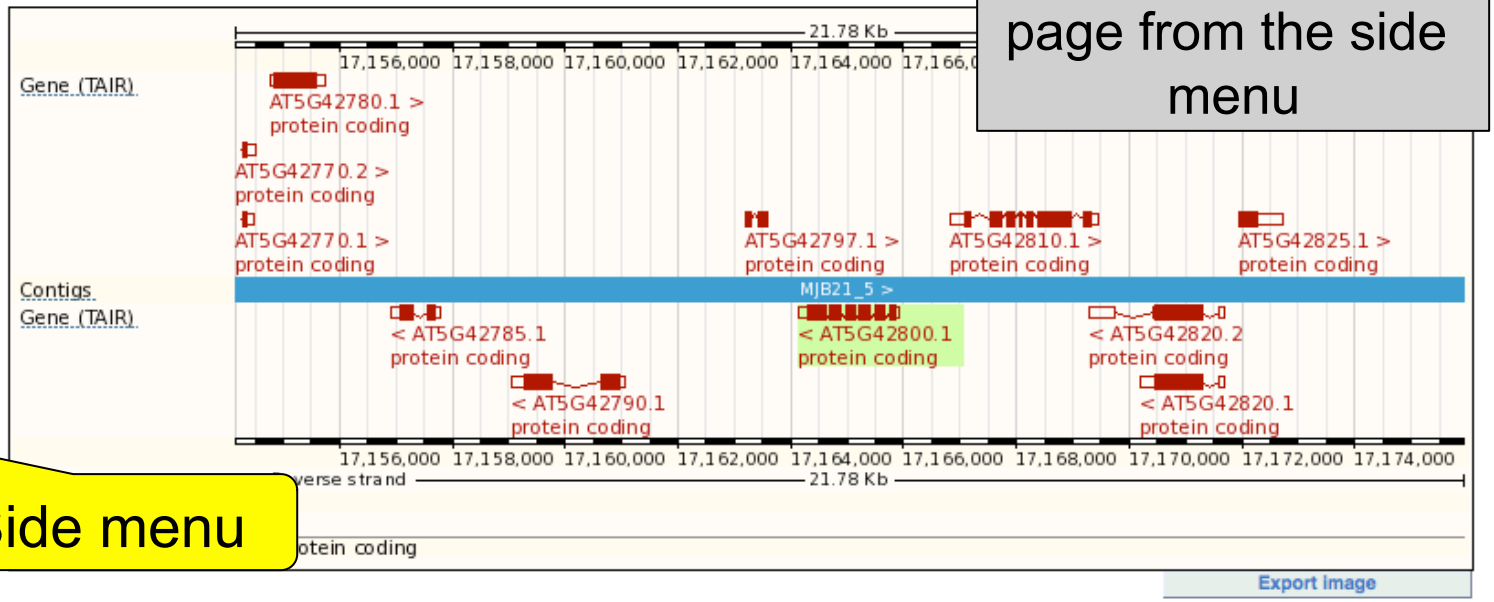
he!p

Main panel changes when you choose another page from the side menu

Name DFR (TAIR Gene Nam

Gene type Known protein coding

Prediction Method Gene annotation by TAIR through a process of automatic and manual



Export image

Configuring the display

Tip: use the "**Configure this page**" link on the left to show additional data in this region.

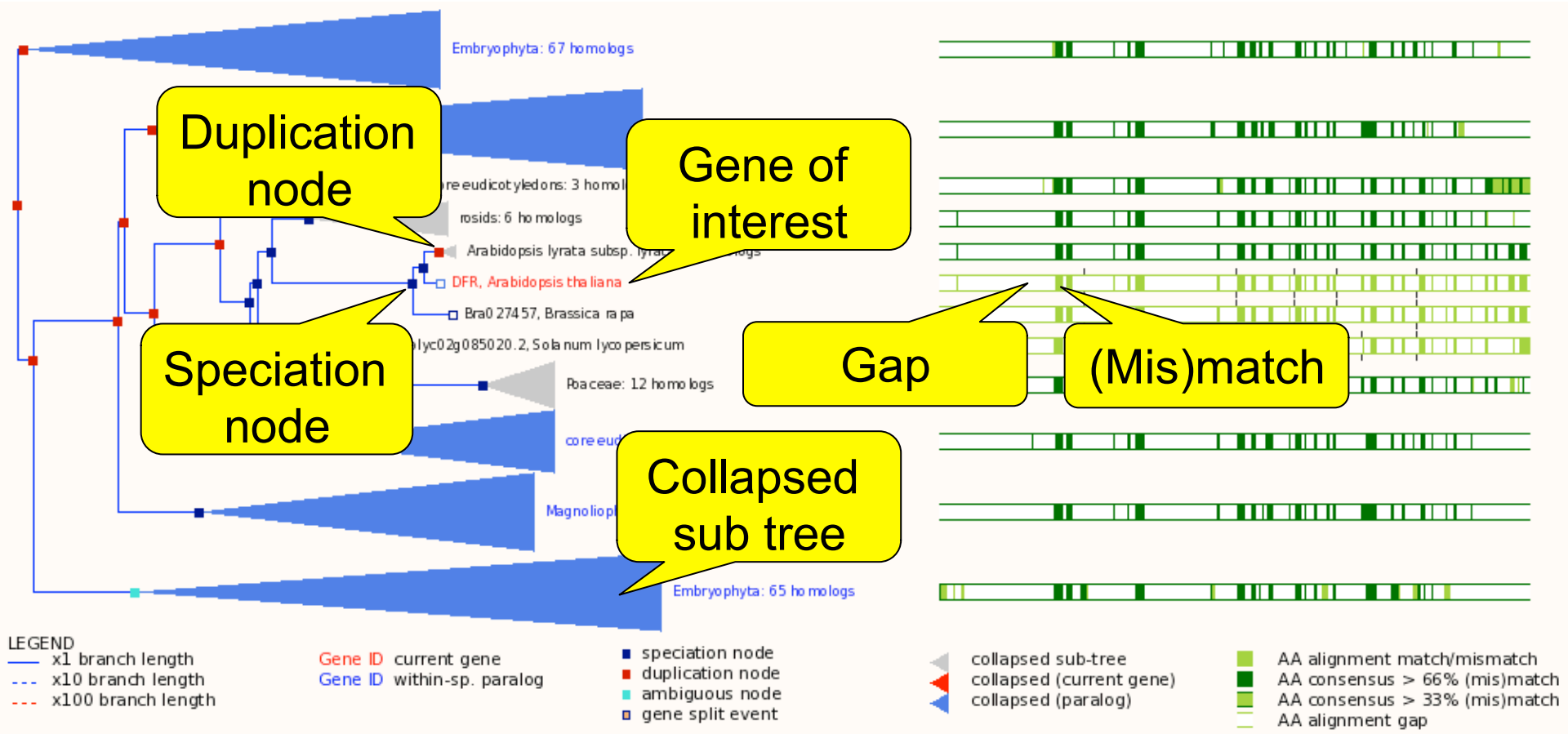
Exercise 2

- Find the *Arabidopsis thaliana* gene encoding glucose-6-phosphate dehydrogenase 1

- *What is the official gene name for this gene?*
- *On which chromosome and on which strand is it located?*
- *What do the empty boxes, filled boxes and lines in the transcript models represent?*

Phylogenetic GeneTree

Protein multiple alignment



Exercise 3

- Explore the ‘Paralogues’ and ‘Gene Tree’ pages.

- *How many paralogues have been identified for the G6PD1 gene? Which paralogues show the highest sequence similarity?*
- *Does the plant gene tree reflect the information that is shown on the ‘Paralogues’ page?*
- *Does the pan-taxonomic gene tree confirm that glucose-6-phosphate dehydrogenase is present in species across all kingdoms?*

Transcript-based displays
- Transcript summary
- Supporting evidence
- Sequence
  - Exons (6)
  - cDNA
  - Protein
- EBI Protein Summary
- Protein Structure
- External References
  - General identifiers (142)
  - Oligo probes (2)
- Ontology
  - Ontology graphs
    - plant anatomical entity (15
    - biological process (4)
    - plant structure developmer
    - cellular component (1)
    - molecular function (6)
  - Ontology table (37)
- Genetic Variation
  - Variation Table
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (5)
  - Variations (71)
- External Data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

Configure this page

Manage your data

Export data

Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

Transcript tab

Description dihydroflavonol-4-reductase Source: EMB

Location Chromosome 5: 17,164,141-17,165,91

Gene This transcript is a product of gene AT5G42800 - This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT5G42800.1	AT5G42800.1	1358	AT5G42800.1	382	Protein coding

Transcript summary help

Export image

Statistics **Exons:** 6 **Transcript length:** 1,358 bps **Translation length:** 382 residues

Ensembl version AT5G42800.1.1

Type Known protein coding

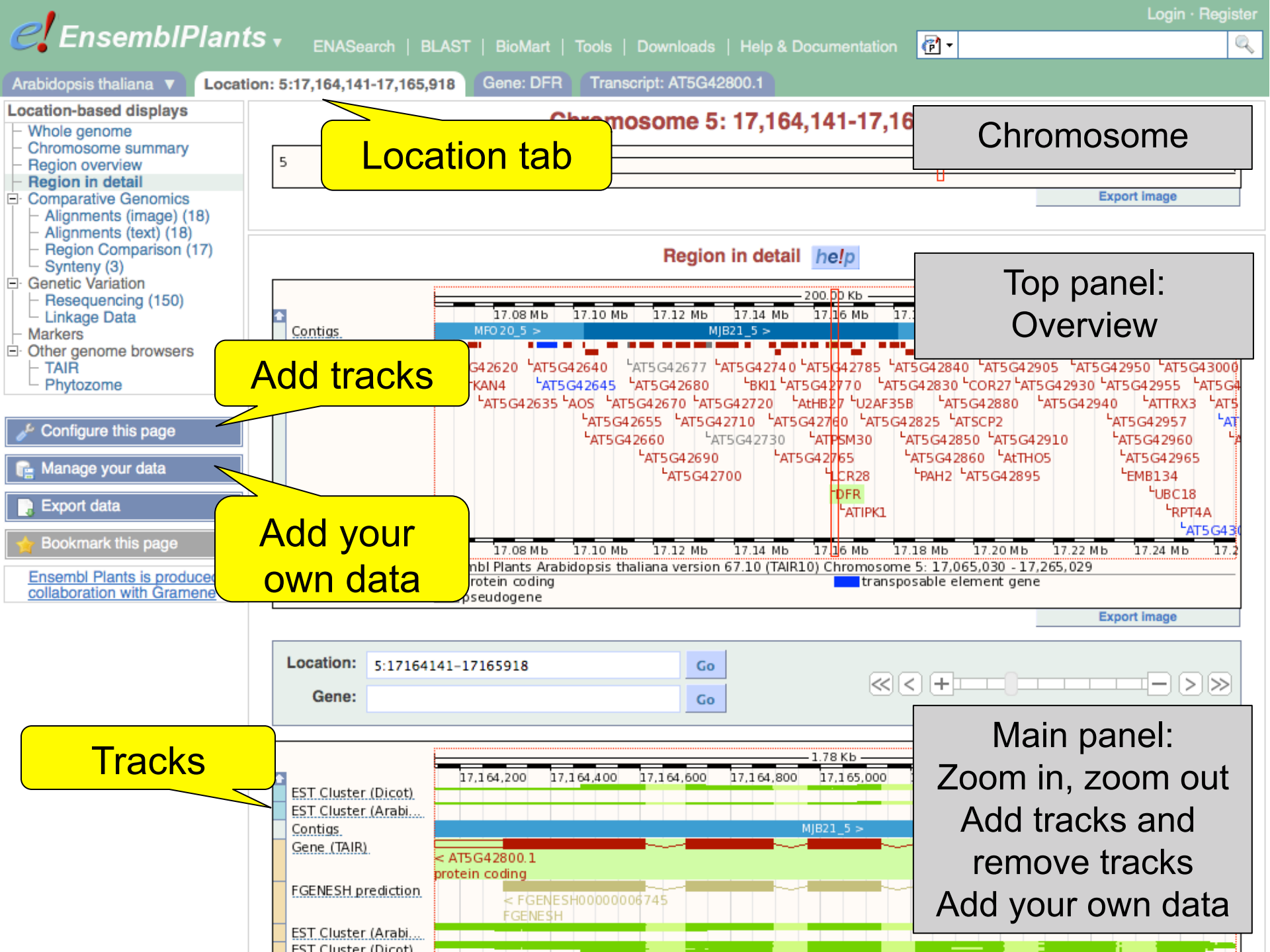
Prediction Method Gene annotation by TAIR through a process of automatic and manual curation.

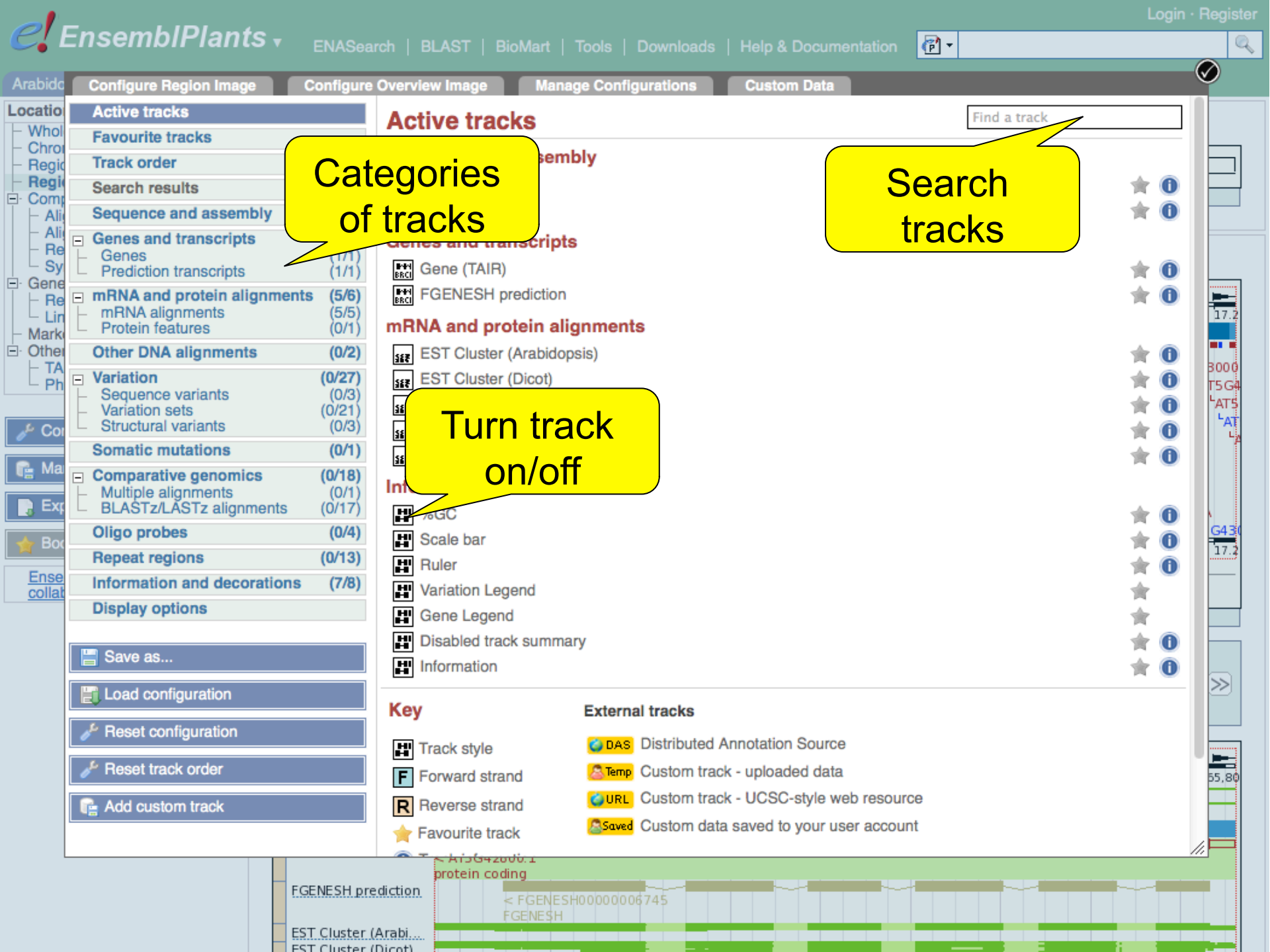
Changed side menu

Exercise 4

- Explore the *G6PD1* transcript and protein (AT5G35790.1).

- *How many exons does this transcript have? Is any of them (partially) untranslated?*
- *Is it cross-referenced to the UniProtKB/Swiss-Prot database? What is its ID and recommended name according to UniProtKB/Swiss-Prot?*
- *Does any of the associated Gene Ontology (GO) terms hint at a role of glucose-6-phosphate dehydrogenase 1 in the pentose phosphate pathway?*
- *Where in the cell is glucose-6-phosphate dehydrogenase 1 located?*
- *In which part of the glucose-6-phosphate dehydrogenase 1 protein is its NAD binding domain located?*

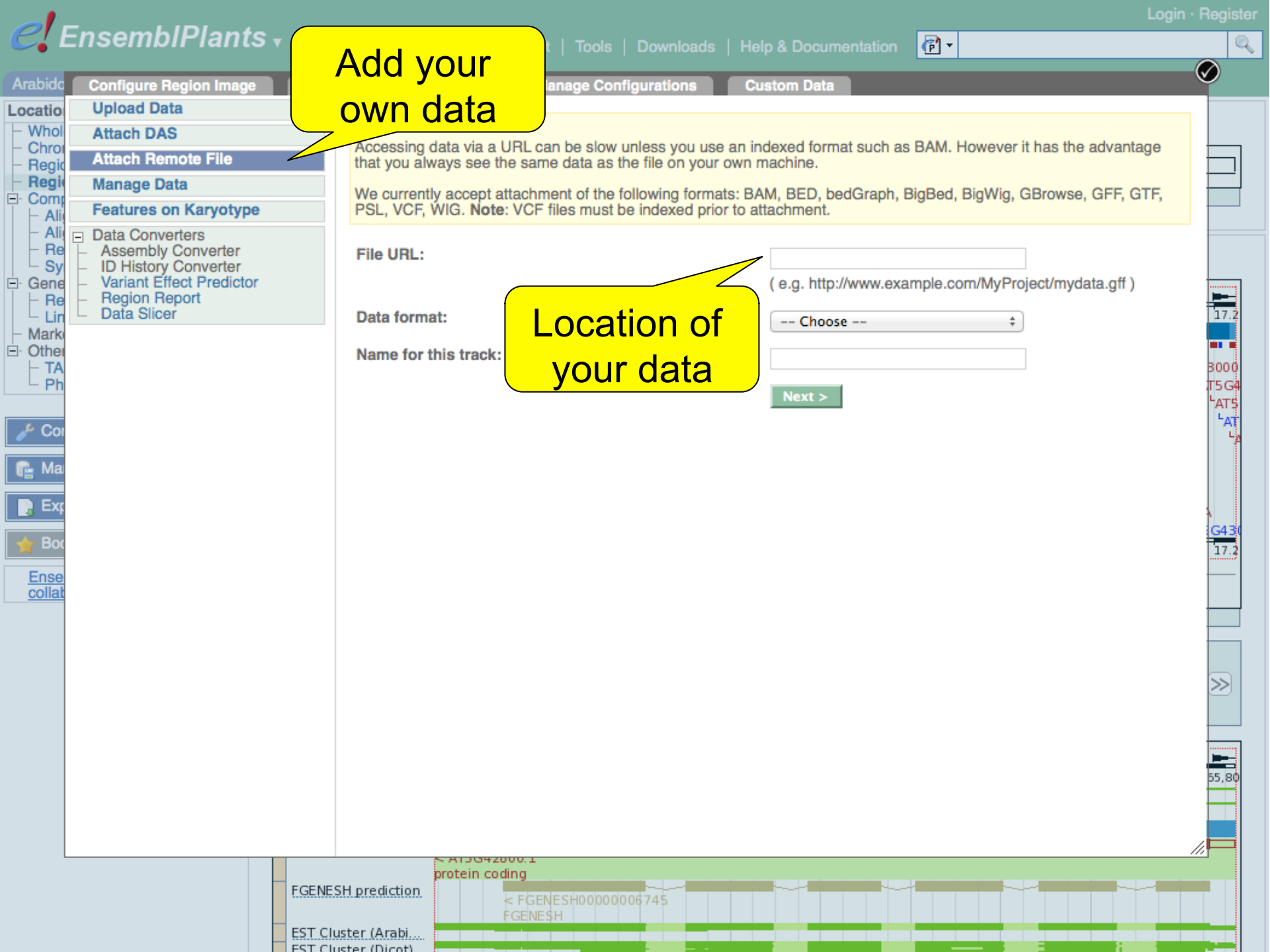

EnsemblPlants
ENASearch | BLAST | BioMart | Tools | Downloads | Help & Documentation
Login · Register
Arabidopsis thaliana
Location: 5:17,164,141-17,165,918
Gene: DFR
Transcript: AT5G42800.1
Location-based displays
Whole genome
Chromosome summary
Region overview
Region in detail
Comparative Genomics
Alignments (image) (18)
Alignments (text) (18)
Region Comparison (17)
Synteny (3)
Genetic Variation
Resequencing (150)
Linkage Data
Markers
Other genome browsers
TAIR
Phytozome
Configure this page
Manage your data
Export data
Bookmark this page
Location tab
Chromosome
Export Image
Region in detail
help
Top panel:
Overview
Add tracks
Add your
own data
Contigs
MFO 20_5 >
MJB21_5 >
17.08 Mb
17.10 Mb
17.12 Mb
17.14 Mb
17.16 Mb
17.18 Mb
17.20 Mb
17.22 Mb
17.24 Mb
protein coding
transposable element gene
pseudogene
Location:
5:17164141-17165918
Go
Gene:
Go
Tracks
Main panel:
Zoom in, zoom out
Add tracks and
remove tracks
Add your own data
1.78 Kb
17,164,200
17,164,400
17,164,600
17,164,800
17,165,000
EST Cluster (Dicot)
EST Cluster (Arabi...
Contigs
Gene (TAIR)
< AT5G42800.1
protein coding
FGENESH prediction
< FGENESH00000006745
FGENESH
EST Cluster (Arabi...
EST Cluster (Dicot)


EnsemblPlants
ENASearch | BLAST | BioMart | Tools | Downloads | Help & Documentation
Login · Register
Configure Region Image
Configure Overview Image
Manage Configurations
Custom Data
Active tracks
Favourite tracks
Track order
Search results
Sequence and assembly
Genes and transcripts
Genes
Prediction transcripts
mRNA and protein alignments (5/6)
mRNA alignments (5/5)
Protein features (0/1)
Other DNA alignments (0/2)
Variation (0/27)
Sequence variants (0/3)
Variation sets (0/21)
Structural variants (0/3)
Somatic mutations (0/1)
Comparative genomics (0/18)
Multiple alignments (0/1)
BLASTz/LASTz alignments (0/17)
Oligo probes (0/4)
Repeat regions (0/13)
Information and decorations (7/8)
Display options
Save as...
Load configuration
Reset configuration
Reset track order
Add custom track
Active tracks
Find a track
Gene (TAIR)
FGENESH prediction
mRNA and protein alignments
EST Cluster (Arabidopsis)
EST Cluster (Dicot)
%GC
Scale bar
Ruler
Variation Legend
Gene Legend
Disabled track summary
Information
Key
Track style
Forward strand
Reverse strand
Favourite track
External tracks
DAS Distributed Annotation Source
Temp Custom track - uploaded data
URL Custom track - UCSC-style web resource
Saved Custom data saved to your user account
Categories of tracks
Search tracks
Turn track on/off
FGENESH prediction
EST Cluster (Arabi...
protein coding
FGENESH

Exercise 5

- Explore the genomic region of the *G6PD1* gene.

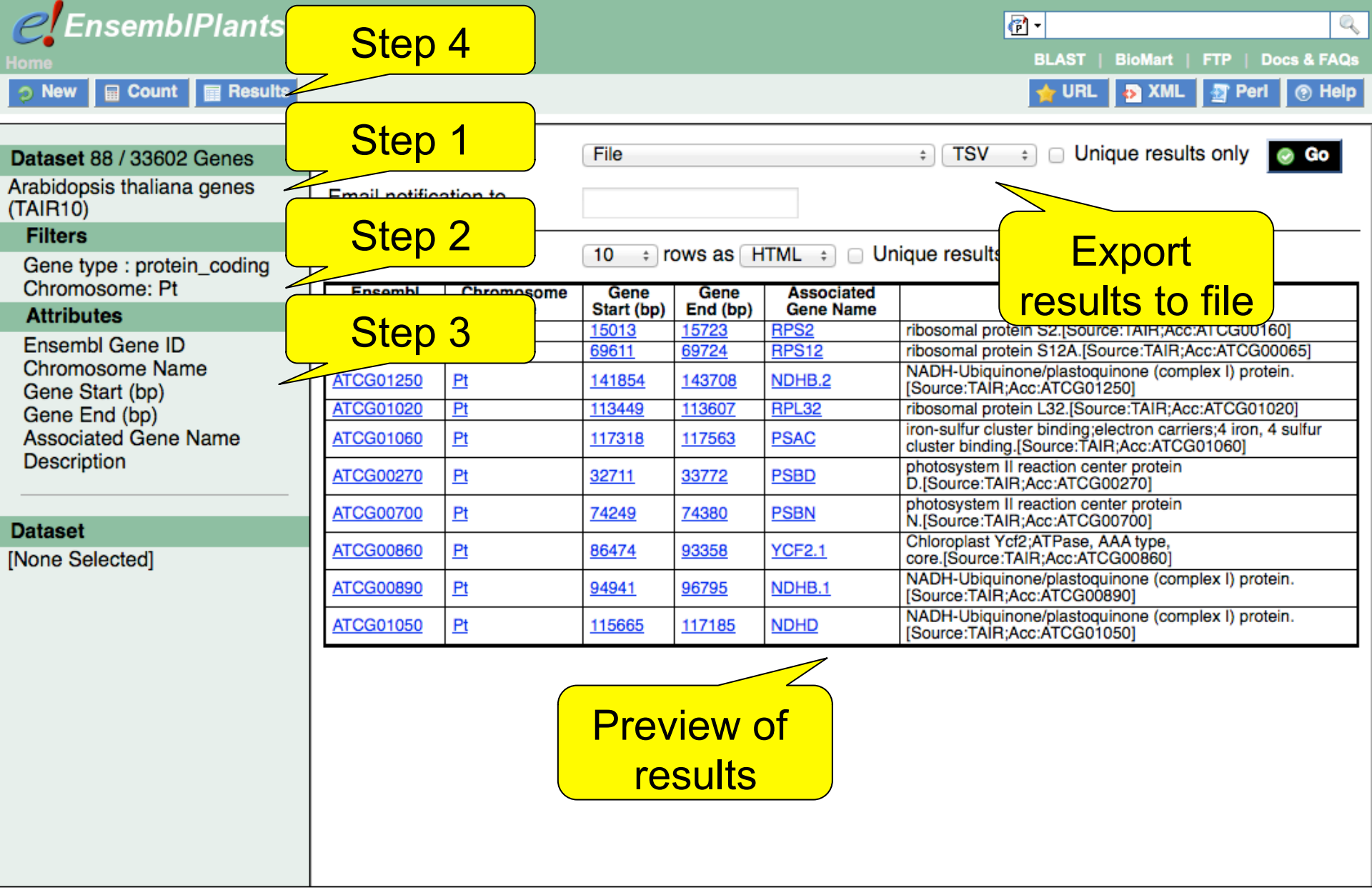
- *Which species in Ensembl Plants shows the highest sequence conservation for this region when compared to Arabidopsis thaliana? And which species the lowest?*
- *What part of the sequence is most conserved across the various species? Is this what you would expect?*


EnsemblPlants
Login · Register
Tools | Downloads | Help & Documentation
Configure Region Image
Custom Data
Upload Data
Attach DAS
Attach Remote File
Manage Data
Features on Karyotype
Data Converters
Assembly Converter
ID History Converter
Variant Effect Predictor
Region Report
Data Slicer
Add your own data
Accessing data via a URL can be slow unless you use an indexed format such as BAM. However it has the advantage that you always see the same data as the file on your own machine.
We currently accept attachment of the following formats: BAM, BED, bedGraph, BigBed, BigWig, GBrowse, GFF, GTF, PSL, VCF, WIG. Note: VCF files must be indexed prior to attachment.
File URL:
(e.g. http://www.example.com/MyProject/mydata.gff)
Data format:
-- Choose --
Name for this track:
Location of your data
Next >
FGENESH prediction
protein coding
FGENESH
EST Cluster (Arabi...

Exercise 7

- The following file contains the genomic coordinates and alleles of a number of new variants in the *G6PD1* gene of *Arabidopsis thaliana*: http://www.ebi.ac.uk/~bert/athaliana_g6pd1_new_variants.txt

- *Do any of these variants change the sequence of the glucose-6-phosphate dehydrogenase 1 protein?*
- *Have any of the variants already been annotated in Ensembl?*


EnsemblPlants
Home
BLAST | BioMart | FTP | Docs & FAQs
New
Count
Results
URL
XML
Perl
Help
Step 4
Step 1
Step 2
Step 3
Dataset 88 / 33602 Genes
Arabidopsis thaliana genes (TAIR10)
Filters
Gene type : protein_coding
Chromosome: Pt
Attributes
Ensembl Gene ID
Chromosome Name
Gene Start (bp)
Gene End (bp)
Associated Gene Name
Description
Dataset
[None Selected]
File
TSV
Unique results only
Go
10
rows as
HTML
Unique results
Export results to file
Ensembl
Chromosome
Gene Start (bp)
Gene End (bp)
Associated Gene Name
15013 15723 RPS2 ribosomal protein S2.[Source:TAIR;Acc:ATCG00160]
69611 69724 RPS12 ribosomal protein S12A.[Source:TAIR;Acc:ATCG00065]
ATCG01250 Pt 141854 143708 NDHB.2 NADH-Ubiquinone/plastoquinone (complex I) protein. [Source:TAIR;Acc:ATCG01250]
ATCG01020 Pt 113449 113607 RPL32 ribosomal protein L32.[Source:TAIR;Acc:ATCG01020]
ATCG01060 Pt 117318 117563 PSAC iron-sulfur cluster binding;electron carriers;4 iron, 4 sulfur cluster binding.[Source:TAIR;Acc:ATCG01060]
ATCG00270 Pt 32711 33772 PSBD photosystem II reaction center protein D.[Source:TAIR;Acc:ATCG00270]
ATCG00700 Pt 74249 74380 PSBN photosystem II reaction center protein N.[Source:TAIR;Acc:ATCG00700]
ATCG00860 Pt 86474 93358 YCF2.1 Chloroplast Ycf2;ATPase, AAA type, core.[Source:TAIR;Acc:ATCG00860]
ATCG00890 Pt 94941 96795 NDHB.1 NADH-Ubiquinone/plastoquinone (complex I) protein. [Source:TAIR;Acc:ATCG00890]
ATCG01050 Pt 115665 117185 NDHD NADH-Ubiquinone/plastoquinone (complex I) protein. [Source:TAIR;Acc:ATCG01050]
Preview of results
Ensembl Plants release 14 - May 2012 (c) EBI

BioMart

- Step 1 – Dataset
  Choose your dataset and species
- Step 2 – Filters
  Limit your dataset
- Step 3 – Attributes
  Specify what information you want to output
- Step 4 – Results
  Preview and output your results

Exercise 8

- Select the Ensembl Genes dataset for *Arabidopsis thaliana*.
- Filter for all genes that are annotated with the GO term ‘pentose-phosphate shunt’, the official GO term for the pentose-phosphate pathway (http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0006098)
- Select the following attributes: Ensembl Gene ID, Associated Gene Name and Description.
- View the results.

- *How many genes does the query find?*
- *Are all G6PD genes amongst the results?*