Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D7.2
Interfaces for integrating omics data within the transPLANT user interface

Due date of deliverable: 28.2.2013
Actual submission date: 27.3.2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: EMBL

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor

EMBL in cooperation with HGMU.

Introduction

Interfaces for integrating "-omics" data within the transPLANT user interface: The development of DAS-based server/client software to enable the integration of gene expression, protein structure and proteomics data in the context of reference genomic data, and the integration of this interface into the transPLANT portal.

DAS (the **D**istributed **A**nnotation **S**ystem) is a protocol and suite of compatible software tools for the sharing of annotation data between different resources. There are three basic types of components in a DAS infrastructure: a reference object server, one or many annotation servers (each of which serves annotations located on the reference objects), and DAS clients, each of which consume reference objects and annotation in order to provide some specialist, customised interface. The Ensembl software system for the storage of genome data functions as all three components: a server delivering reference sequence and annotation, but also a client that integrates data from other DAS resources. The Ensembl Plants database (http://plants.ensembl.org) is being used as a central resource to store and share genomic data in transPLANT. For this deliverable, we have developed new interfaces based on DAS and other systems for the sharing and display of different types of "-omics" data.

Methods

Ensembl Plants

DAS reference and annotation servers, and DAS-powered integrated views, are available for the 23 plant species currently available in Ensembl Plants.

DAS

Using the Ensembl genome browser we have integrated with key "-omics" databases using the Distributed Annotation System (DAS) [1]. DAS annotations for protein expression information are queried from the Array Express Expression Atlas [2], protein structure data is queried from the Protein Data Bank [3], and proteomics data are queried from PRIDE [4]. In addition, protein and literature information are queried from UniProt [5] and the CiteXplore literature lookup service [6], respectively. New web views have been developed for visualising this data.

Data type	Database used	DAS URL
Expression	Array Express Expression Atlas [2]	http://www.ebi.ac.uk/gxa/das/s4
Protein structure	PDB [3]	http://www.ebi.ac.uk/das-srv/proteindas/das/pdbe_summary
Proteomics	PRIDE [4]	http://www.ebi.ac.uk/pride-das/das/PrideDataSource
Protein information	UniProt [5]	http://www.ebi.ac.uk/das-srv/s4/das/protein-proxy
Literature links	Europe PubMed Central [6]	http://www.ebi.ac.uk/das-srv/misc/das/citexplore_split

Table 1 Details of the DAS sources used.

References

1. Integrating biological data--the Distributed Annotation System. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P, Kähäri A, Kulesha E, Macías JR, Reeves GA, Prlić A. BMC Bioinformatics. 2008.
2. Gene expression atlas at the European bioinformatics institute. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. Nucleic Acids Res. 2010.
3. PDBe: Protein Data Bank in Europe. Velankar S, Alhroub Y, Best C, Caboche S, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Golovin A, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Heuson E, Hirshberg M, John M, Lagerstedt I, Mir S, Newman LE, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-García E, Sen S, Slowley R, Suarez-Uruena A, Swaminathan GJ, Symmons MF, Vranken WF, Wainwright M, Kleywegt GJ. Nucleic Acids Res. 2012.
4. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Vizcaíno JA, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, O'Kelly G, Schoenegger A, Ovelleiro D, Pérez-Riverol Y, Reisinger F, Ríos D, Wang R, Hermjakob H. Nucleic Acids Res. 2013.
5. Update on activities at the Universal Protein Resource (UniProt) in 2013. UniProt Consortium. Nucleic Acids Res. 2013.
6. http://europepmc.org/

Results (if applicable, interactions with other workpackages)

Use of DAS as interface for "-omics" data

The Ensembl framework allows for DAS data sources to be dynamically attached by users, using servers the users have private knowledge of, or by selecting publicly announced servers in the DAS registry (http://www.dasregistry.org/), which can then be visualised in generic views (e.g. as tracks on a genome browser). For DAS sources provided by major repositories of "-omics" data, we have pre-configured and embedded these within custom extensions to the user interface. When reference objects are stable, this use of DAS enables the dynamic update of the user interface without requiring the prior computational integration of data from multiple sources, and can serve data to users as soon as it is made public by the annotation supplier. We have pre-configured DAS sources to retrieve data from the sources listed in Table 1. Some sources contain data only for specific species, as described in Table 2. The Ensembl Plants browser is configured to only show pages for species specific data where appropriate.

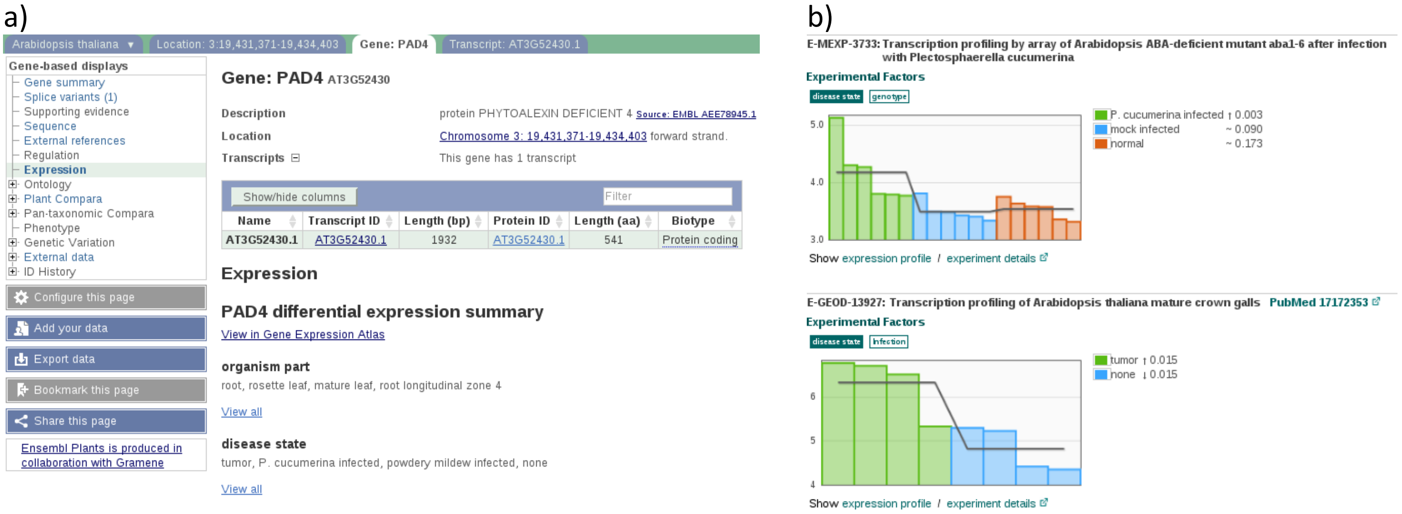
Species	Expression Atlas	PDB	PRIDE	UniProt	Literature
Arabidopsis thaliana	X		X		
Chlamydomonas reinhardtii			X		
Oryza sativa indica			X		
Oryza sativa	X				
Populus trichocarpa	X				
Solanum lycopersicum			X		
Solanum tuberosum			X		
Vitis vinifera	X		X		

Zea mays			X		
All species (where applicable)		X		X	X

Table 2 Species-specific data sources.

Example of custom interfaces for "-omics" DAS sources within Ensembl Plants

Figure 1a and 1b: Viewing expression information in a) Ensembl Plants and b) following a link to the Array Express Expression Atlas.



Gene: PAD4 AT3G52430

Description: protein PHYTOALEXIN DEFICIENT 4 Source: EMBL AEE78945.1

Location: Chromosome 3: 19,431,371-19,434,403 forward strand.

Transcripts ⊟: This gene has 1 transcript

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT3G52430.1	AT3G52430.1	1932	AT3G52430.1	541	Protein coding

Expression

PAD4 differential expression summary

organism part

root, rosette leaf, mature leaf, root longitudinal zone 4

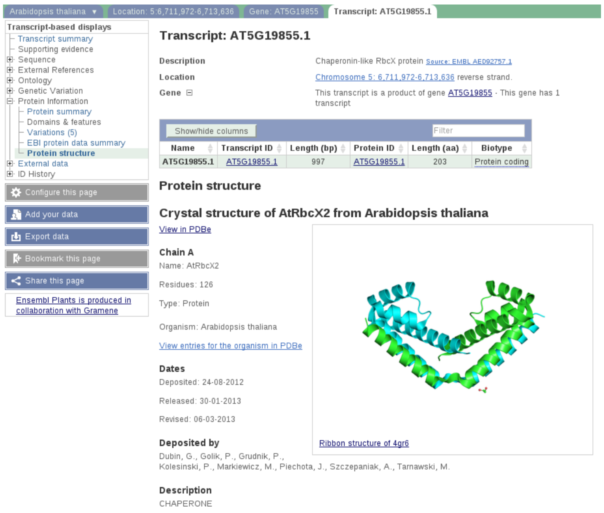
disease state

tumor, P. cucumerina infected, powdery mildew infected, none

E-MEXP-3733: Transcription profiling by array of Arabidopsis ABA-deficient mutant aba1-6 after infection with Plectosphaerella cucumerina

E-GEOD-13927: Transcription profiling of Arabidopsis thaliana mature crown galls PubMed 17172353

Figure 2: Viewing the protein structure in Ensembl Plants



Transcript: AT5G19855.1

Description: Chaperonin-like RbcX protein Source: EMBL AED92757.1

Location: Chromosome 5: 6,711,972-6,713,636 reverse strand.

Gene ⊟: This transcript is a product of gene AT5G19855 - This gene has 1 transcript

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT5G19855.1	AT5G19855.1	997	AT5G19855.1	203	Protein coding

Protein structure

Crystal structure of AtRbcX2 from Arabidopsis thaliana

Chain A

Name: AtRbcX2

Residues: 126

Type: Protein

Organism: Arabidopsis thaliana

Dates

Deposited: 24-08-2012

Released: 30-01-2013

Revised: 06-03-2013

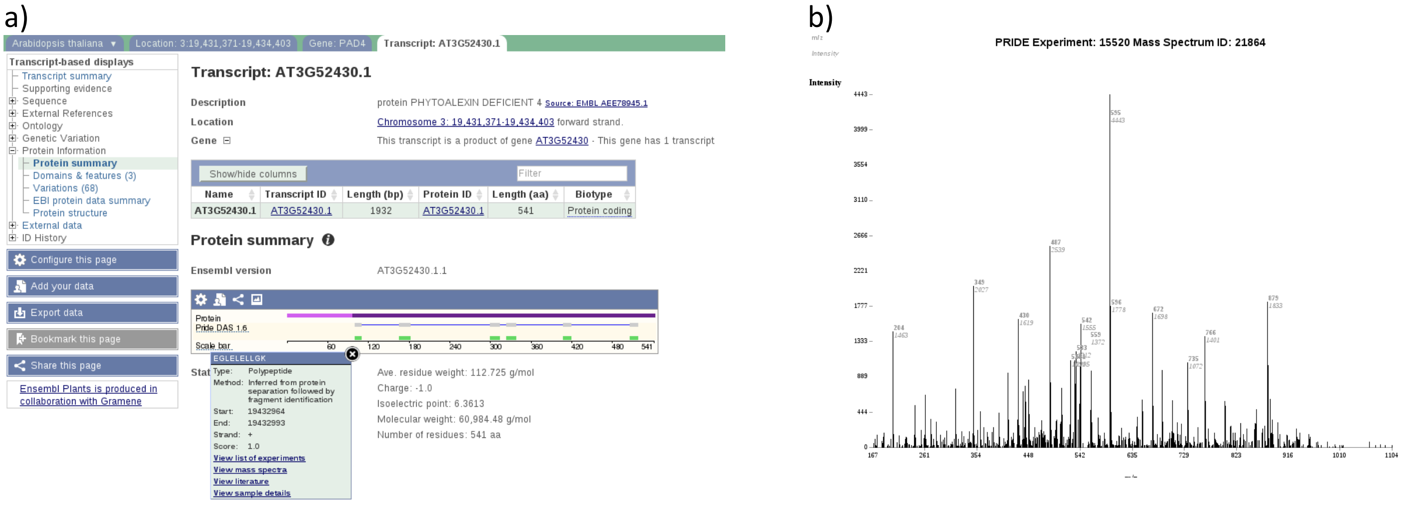
Deposited by

Dubin, G., Golik, P., Grudnik, P., Kolesinski, P., Markiewicz, M., Piechota, J., Szczepaniak, A., Tarnawski, M.

Description

CHAPERONE

Figure 3a and 3b: Viewing proteomics information in a) Ensembl Plants and b) following a link to view the spectra in PRIDE.



Arabidopsis thaliana ▾ | Location: 3:19,431,371-19,434,403 | Gene: PAD4 | Transcript: AT3G52430.1

Transcript-based displays
- Transcript summary
- Supporting evidence
- Sequence
- External References
- Ontology
- Genetic Variation
- Protein Information
  - Protein summary
  - Domains & features (3)
  - Variations (68)
  - EBI protein data summary
  - Protein structure
- External data
- ID History

Configure this page | Add your data | Export data | Bookmark this page | Share this page

Ensembl Plants is produced in collaboration with Gramene

Transcript: AT3G52430.1

Description: protein PHYTOALEXIN DEFICIENT 4 Source: EMBL AEE78945.1

Location: Chromosome 3: 19,431,371-19,434,403 forward strand.

Gene ⊟: This transcript is a product of gene AT3G52430 - This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT3G52430.1	AT3G52430.1	1932	AT3G52430.1	541	Protein coding

Protein summary

Ensembl version: AT3G52430.1.1

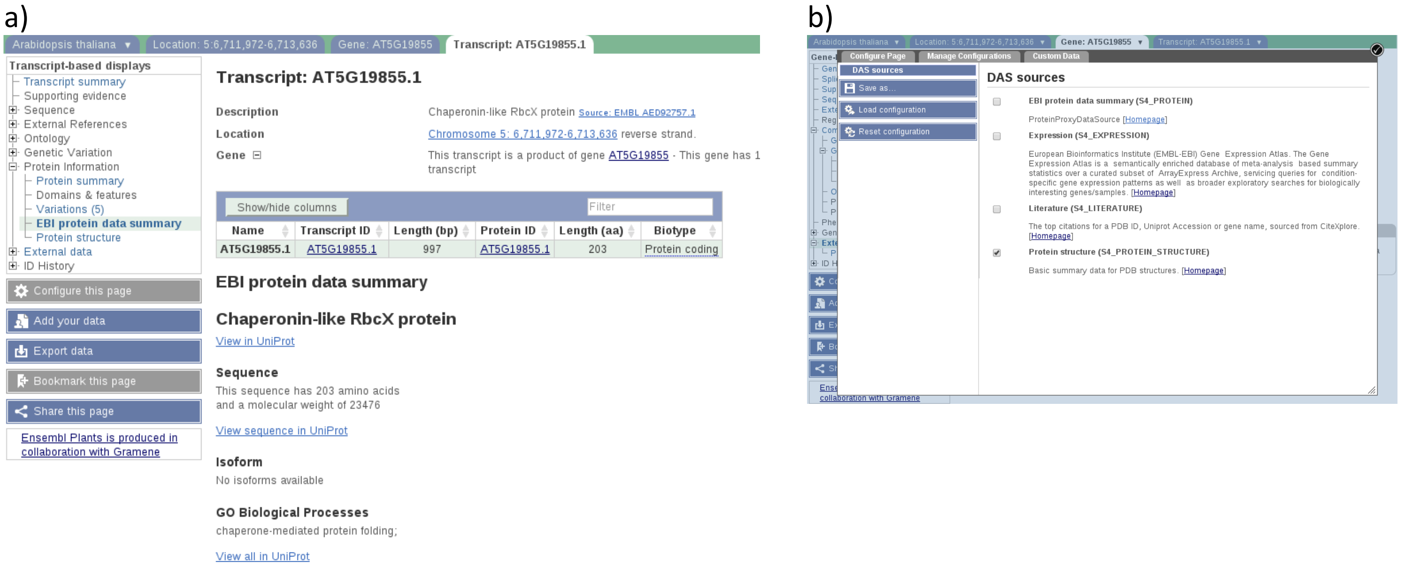
Protein
Pride DAS 1.6
Scale bar 60 120 180 240 300 360 420 480 541

EGLELELLGK
Type: Polypeptide
Method: Inferred from protein separation followed by fragment identification
Start: 19432964
End: 19432993
Strand: +
Score: 1.0
View list of experiments
View mass spectra
View literature
View sample details

Stat
Ave. residue weight: 112.725 g/mol
Charge: -1.0
Isoelectric point: 6.3613
Molecular weight: 60,984.48 g/mol
Number of residues: 541 aa

PRIDE Experiment: 15520 Mass Spectrum ID: 21864

Figure 4a and 4b: a) Viewing protein data from UniProt in Ensembl Plants, b) selecting the pre-configured data-sources.



Arabidopsis thaliana ▾ | Location: 5:6,711,972-6,713,636 | Gene: AT5G19855 | Transcript: AT5G19855.1

Transcript-based displays
- Transcript summary
- Supporting evidence
- Sequence
- External References
- Ontology
- Genetic Variation
- Protein Information
  - Protein summary
  - Domains & features
  - Variations (5)
  - EBI protein data summary
  - Protein structure
- External data
- ID History

Configure this page | Add your data | Export data | Bookmark this page | Share this page

Ensembl Plants is produced in collaboration with Gramene

Transcript: AT5G19855.1

Description: Chaperonin-like RbcX protein Source: EMBL AED92757.1

Location: Chromosome 5: 6,711,972-6,713,636 reverse strand.

Gene ⊟: This transcript is a product of gene AT5G19855 - This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT5G19855.1	AT5G19855.1	997	AT5G19855.1	203	Protein coding

EBI protein data summary

Chaperonin-like RbcX protein

View in UniProt

Sequence
This sequence has 203 amino acids and a molecular weight of 23476

View sequence in UniProt

Isoform
No isoforms available

GO Biological Processes
chaperone-mediated protein folding;

View all in UniProt

Arabidopsis thaliana ▾ | Location: 5:6,711,972-6,713,636 ▾ | Gene: AT5G19855 ▾ | Transcript: AT5G19855.1 ▾

Configure Page | Manage Configurations | Custom Data

DAS sources | Save as... | Load configuration | Reset configuration

DAS sources

☐ EBI protein data summary (S4_PROTEIN)
ProteinProxyDataSource [Homepage]

☐ Expression (S4_EXPRESSION)
European Bioinformatics Institute (EMBL-EBI) Gene Expression Atlas. The Gene Expression Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive, servicing queries for condition-specific gene expression patterns as well as broader exploratory searches for biologically interesting genes/samples. [Homepage]

☐ Literature (S4_LITERATURE)
The top citations for a PDB ID, Uniprot Accession or gene name, sourced from CiteXplore. [Homepage]

☑ Protein structure (S4_PROTEIN_STRUCTURE)
Basic summary data for PDB structures. [Homepage]

Dynamic Configuration of DAS sources

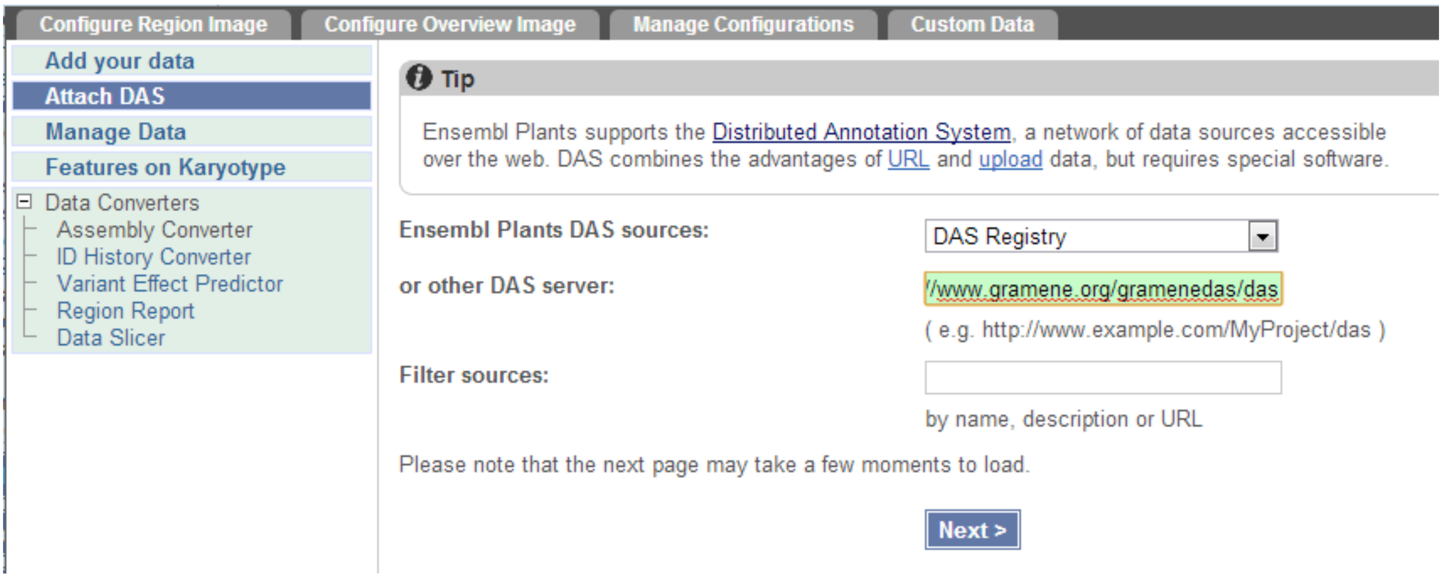
To enable the display of the DAS sources within the Ensembl Plants genome browser we have contributed to the development of a number of the DAS system components, and in particular the development of the Ensembl browser as a DAS client, i.e. the browser that is capable of displaying the information served using DAS protocol.

To make it easier for users to use the DAS system we have implemented a new wizard in the Manage Your Data configuration panel of the Ensembl browser. This wizard allows users to attach a data source from a known DAS server, or select a source from a list of public sources registered with the DAS registry without any prior knowledge of the DAS protocol. The wizard appears presents two web forms to users. At the first step users are asked to enter the address of the DAS server (Figure 4a), and at the next step (Figure 4b) users can choose which sources to display in the genome browser alongside other data. The final step stores the parameters of the selected sources in the local user session database and prints confirmation which sources have been successfully attached (Figure 4c). Once the configuration panel is closed the DAS sources will appear in the browser very similar to the normal data tracks.

Under the hood we have extended the Ensembl External Data API to enable the communication between the Ensembl web code and DAS servers using the DAS protocol and added a module to the Ensembl Drawing library to render the data from the DAS sources.

Figures 5a, b and c: Manually configuring an alternative DAS track from the DAS registry.

Figure 5a. Attach DAS Wizard: Step 1. Selecting DAS server



Tip

Ensembl Plants supports the Distributed Annotation System, a network of data sources accessible over the web. DAS combines the advantages of URL and upload data, but requires special software.

Ensembl Plants DAS sources: DAS Registry

or other DAS server: /www.gramene.org/gramenedas/das

(e.g. http://www.example.com/MyProject/das)

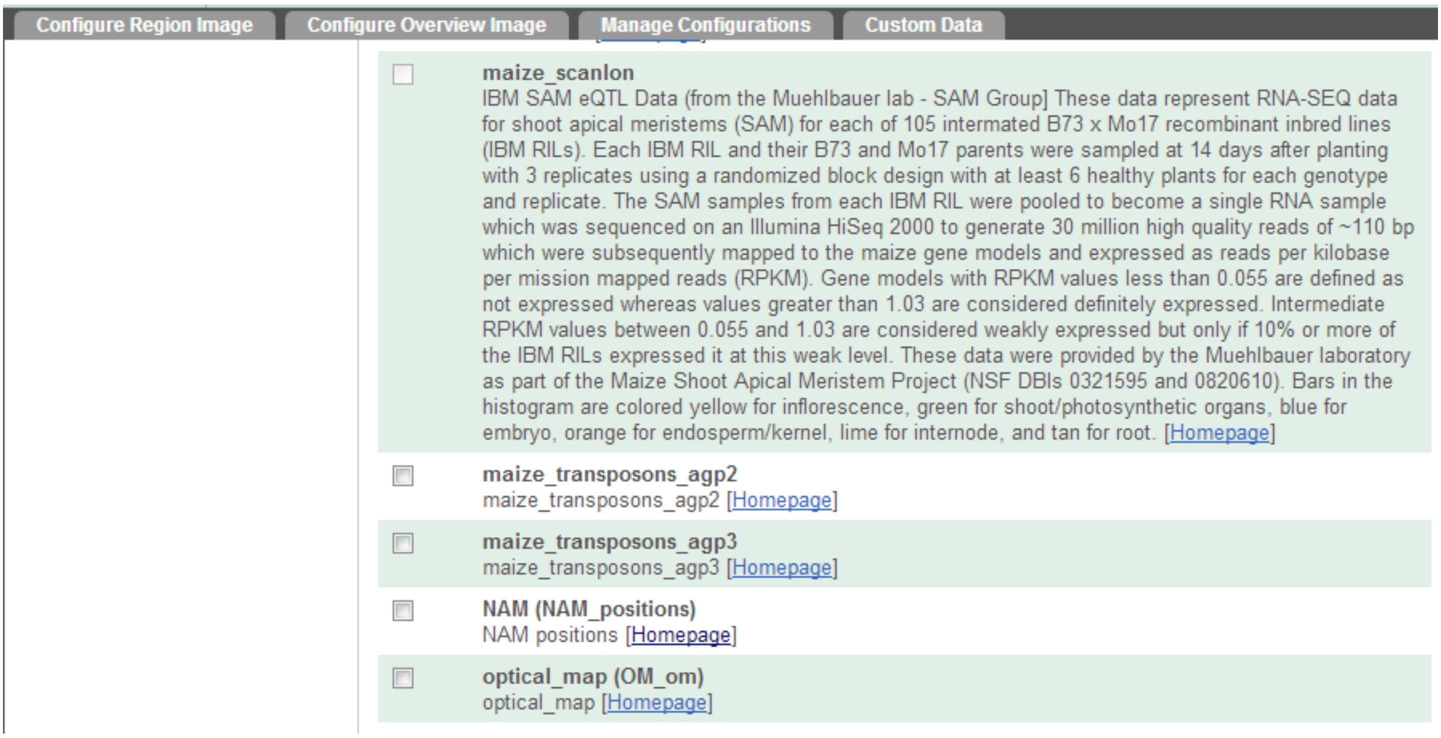
Filter sources:

by name, description or URL

Please note that the next page may take a few moments to load.

A user specifies their own data source by providing a URL in the first form (Figure 4a), the Ensembl web code will issue a DAS request users for the list of available data sources served from this URL and will convert the incoming response into a second web form with a list of the sources, with a short summary of the data provided by each (Figure 4b), allowing the user to select which specific sources they want.

Figure 5b. Attach DAS Wizard: Step 2. Selecting sources.



□ **maize_scanlon**
IBM SAM eQTL Data (from the Muehlbauer lab - SAM Group] These data represent RNA-SEQ data for shoot apical meristems (SAM) for each of 105 intermated B73 x Mo17 recombinant inbred lines (IBM RILs). Each IBM RIL and their B73 and Mo17 parents were sampled at 14 days after planting with 3 replicates using a randomized block design with at least 6 healthy plants for each genotype and replicate. The SAM samples from each IBM RIL were pooled to become a single RNA sample which was sequenced on an Illumina HiSeq 2000 to generate 30 million high quality reads of ~110 bp which were subsequently mapped to the maize gene models and expressed as reads per kilobase per mission mapped reads (RPKM). Gene models with RPKM values less than 0.055 are defined as not expressed whereas values greater than 1.03 are considered definitely expressed. Intermediate RPKM values between 0.055 and 1.03 are considered weakly expressed but only if 10% or more of the IBM RILs expressed it at this weak level. These data were provided by the Muehlbauer laboratory as part of the Maize Shoot Apical Meristem Project (NSF DBIs 0321595 and 0820610). Bars in the histogram are colored yellow for inflorescence, green for shoot/photosynthetic organs, blue for embryo, orange for endosperm/kernel, lime for internode, and tan for root. [Homepage]

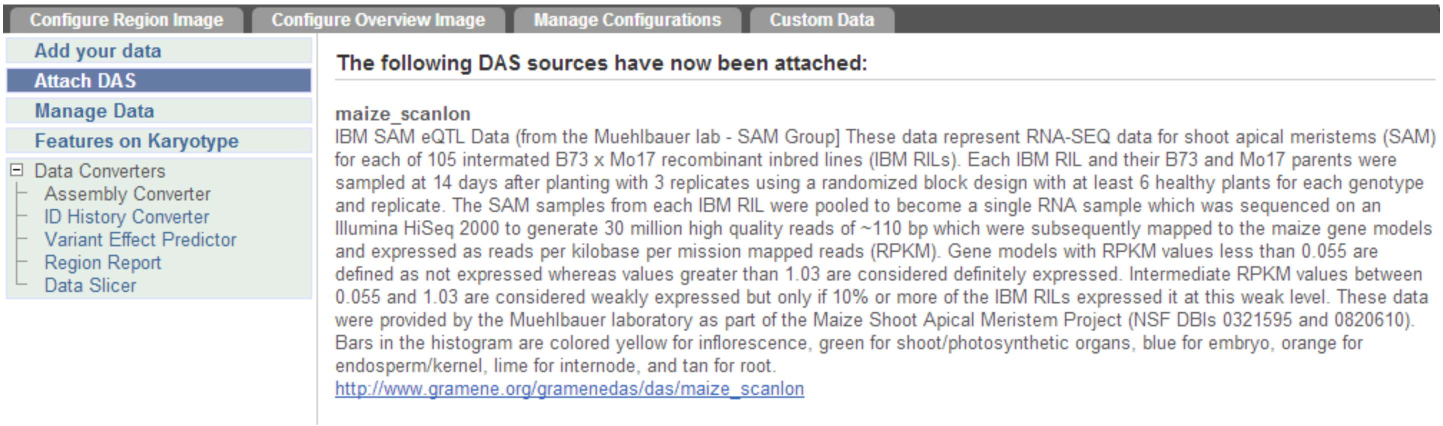
□ **maize_transposons_agp2**
maize_transposons_agp2 [Homepage]

□ **maize_transposons_agp3**
maize_transposons_agp3 [Homepage]

□ **NAM (NAM_positions)**
NAM positions [Homepage]

□ **optical_map (OM_om)**
optical_map [Homepage]

Figure 5c. Attach DAS Wizard: Step 3. Confirmation of the successfully attached DAS sources.



The following DAS sources have now been attached:

maize_scanlon
IBM SAM eQTL Data (from the Muehlbauer lab - SAM Group] These data represent RNA-SEQ data for shoot apical meristems (SAM) for each of 105 intermated B73 x Mo17 recombinant inbred lines (IBM RILs). Each IBM RIL and their B73 and Mo17 parents were sampled at 14 days after planting with 3 replicates using a randomized block design with at least 6 healthy plants for each genotype and replicate. The SAM samples from each IBM RIL were pooled to become a single RNA sample which was sequenced on an Illumina HiSeq 2000 to generate 30 million high quality reads of ~110 bp which were subsequently mapped to the maize gene models and expressed as reads per kilobase per mission mapped reads (RPKM). Gene models with RPKM values less than 0.055 are defined as not expressed whereas values greater than 1.03 are considered definitely expressed. Intermediate RPKM values between 0.055 and 1.03 are considered weakly expressed but only if 10% or more of the IBM RILs expressed it at this weak level. These data were provided by the Muehlbauer laboratory as part of the Maize Shoot Apical Meristem Project (NSF DBIs 0321595 and 0820610). Bars in the histogram are colored yellow for inflorescence, green for shoot/photosynthetic organs, blue for embryo, orange for endosperm/kernel, lime for internode, and tan for root.
http://www.gramene.org/gramenedas/das/maize_scanlon

It is also possible to provide the exact address of a DAS source at the first step – then only one source will available for selection.

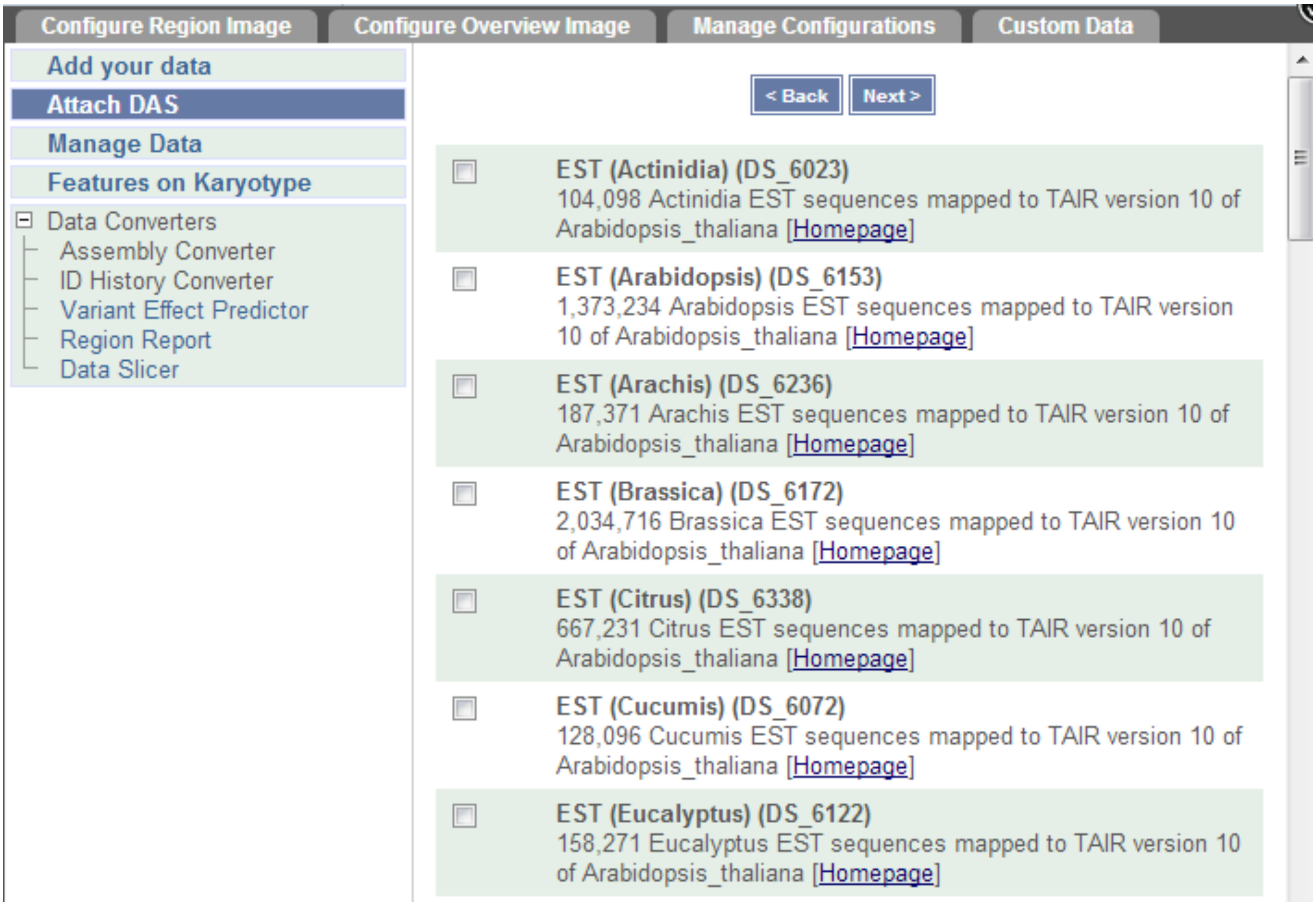
By providing the DAS source or DAS server URL users can attach any valid DAS source to the current display - wizard will not check if the source is compatible with the reference sequence. So user must make sure the DAS source has annotations for the relevant species and uses the same assembly version as the genome browser.

Alternatively the DAS registry http://www.dasregistry.org can be used to select which sources to attach. Practically all the popular DAS sources are registered in the DAS registry so users can easily find different types of data for a species of interest. The DAS registry itself implements DAS so any DAS client can consult it in the same way it would contact any other DAS server. The existence of the registry improves discoverability of the

DAS sources and greatly helps users to find useful data.

If you leave the other DAS source field of the wizard first step blank and click “Next”, the DAS wizard will contact the DAS registry and fetch the full list of known sources from the registry and will filter out the ones that can not be displayed on the current species or assembly version (Figure 5d).

Figure 5d. Sources from the registry. The “irrelevant” sources are filtered out.



☐ **EST (Actinidia) (DS_6023)**
104,098 Actinidia EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

☐ **EST (Arabidopsis) (DS_6153)**
1,373,234 Arabidopsis EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

☐ **EST (Arachis) (DS_6236)**
187,371 Arachis EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

☐ **EST (Brassica) (DS_6172)**
2,034,716 Brassica EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

☐ **EST (Citrus) (DS_6338)**
667,231 Citrus EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

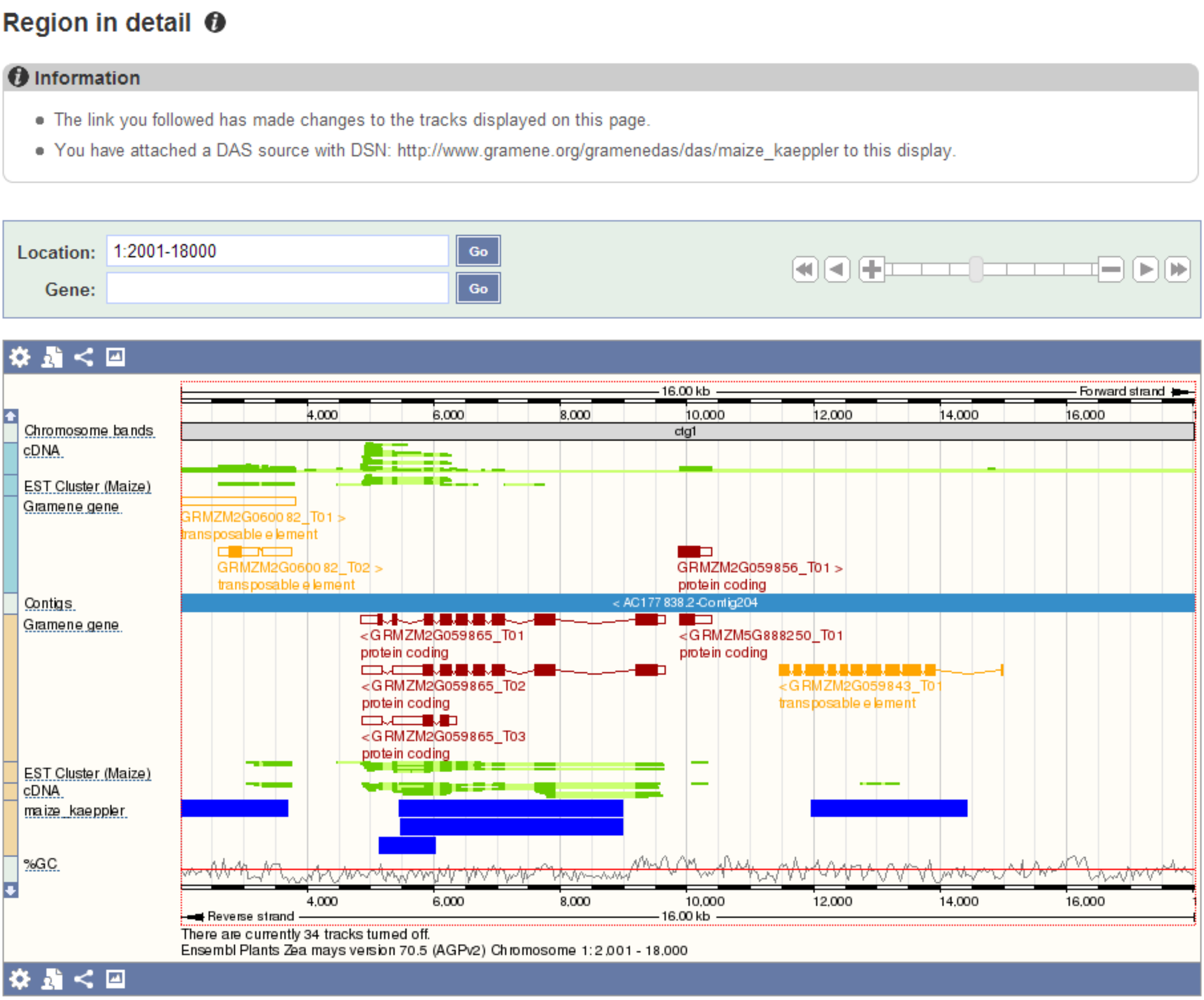
☐ **EST (Cucumis) (DS_6072)**
128,096 Cucumis EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

☐ **EST (Eucalyptus) (DS_6122)**
158,271 Eucalyptus EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [Homepage]

And finally to make it easier for the collaborators to share their data with Ensembl Plants we have implemented the method that allows users to attach a das source via a direct web link. So a data provider can create links on its website that will redirect to the Ensembl Plants browser and will automatically attach the DAS source to the web display without any need for extra actions by users, e.g. the following link will add a DAS track served by Gramene DAS server to the Ensembl Plants web page:

http://plants.ensembl.org/Zea_mays/Location/View?db=core;r=1:2001-18000;contigviewbottom=das:http://www.gramene.org/gramenedas/das/maize_kaeppler=normal

Figure 5e: Once a DAS source has been attached to the genome browser its parameters get stored in a local user session database and the source data appear just like an ordinary track in the genome browser.



Region in detail

Information

- The link you followed has made changes to the tracks displayed on this page.
- You have attached a DAS source with DSN: http://www.gramene.org/gramenedas/das/maize_kaeppler to this display.

Location: 1:2001-18000

Gene:

There are currently 34 tracks turned off.
Ensembl Plants Zea mays version 70.5 (AGPv2) Chromosome 1:2,001 - 18,000

Alternative approaches to sharing "-omics" data

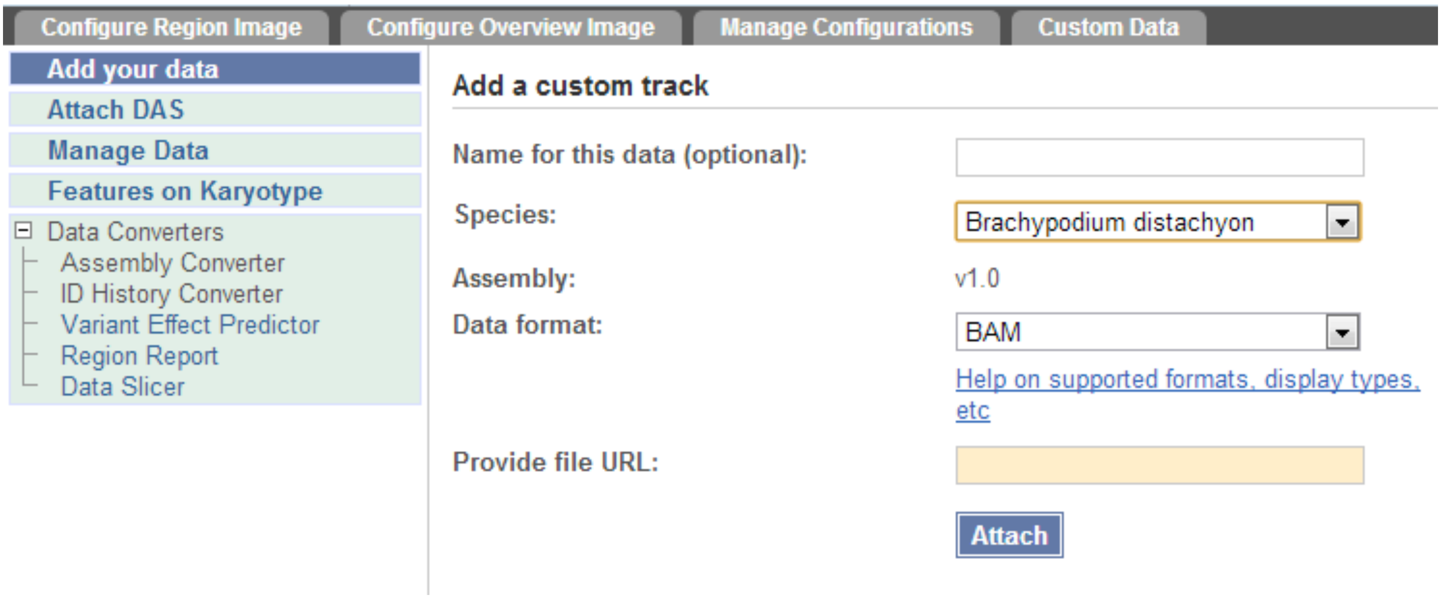
Although now widely adopted in the biological sciences, DAS is not the only technology capable of providing integration of different types of "-omics" data into a reference framework. Other approaches include (i) user upload of data, into a private, user-specific area of a database and (ii) dynamic display of big data files over HTTP and FTP. User upload and the dynamic display of files allow users to visualise data using generic software without requiring the pre-installation of DAS server software. The information is not "published" in a formal sense, allowing users to visualise pre-publication data in the context of reference data, for example, although by sharing the URL of their files, they are still able to share data with collaborators.

When it first appeared the main advantage of the DAS protocol was ability to display the data from data rich sources. For example a user is limited to 5MB when uploading his data to the browser database. The DAS protocol has achieved big popularity because it managed to overcome this problem by putting the data with the data provider but transmitting only the features that are relevant to the current view.

However with advent of the next generation sequencing the amount of data has increased so much that even DAS protocol could not efficiently handle it. New, specialized, more compact data formats have appeared. For example VCF has become a standard for storing variation calls and BAM for storing re-sequencing data. Each of these big data formats uses knowledge about the type of data it stores to make it more efficient to store and retrieve data and requires a piece of specific software installed to interact with the files. The files in these formats ordinary exceed tens of gigabytes and thus require a building of indexes to speed up the data access. The indexes provide a way for fast retrieval of the data and are just few mega bytes in size. When the big data files are attached to the genome browser only these indexes are uploaded to the local server. While a user is navigating a genome in the browser the indexes are used to locate the data needed to be displayed in the remote big file, and only the portion of the data that is relevant to the current displayed is transferred over the network.

We have implemented direct file visualisation of both VCF and BAM big data file types in Ensembl Plants. To achieve that we have updated the Add your data wizard that helps attach remote files to the browser display to include VCF and BAM formats (Figure 5f), and we have extended the Ensembl External Data API to include modules that would enable the data retrieval from the VCF and BAM files, and Ensembl Draw library to display VCF and BAM tracks (Figure 5g).

Figure 5f. Add your data wizard: attaching a remote data file.



Add a custom track

Name for this data (optional):

Species: Brachypodium distachyon

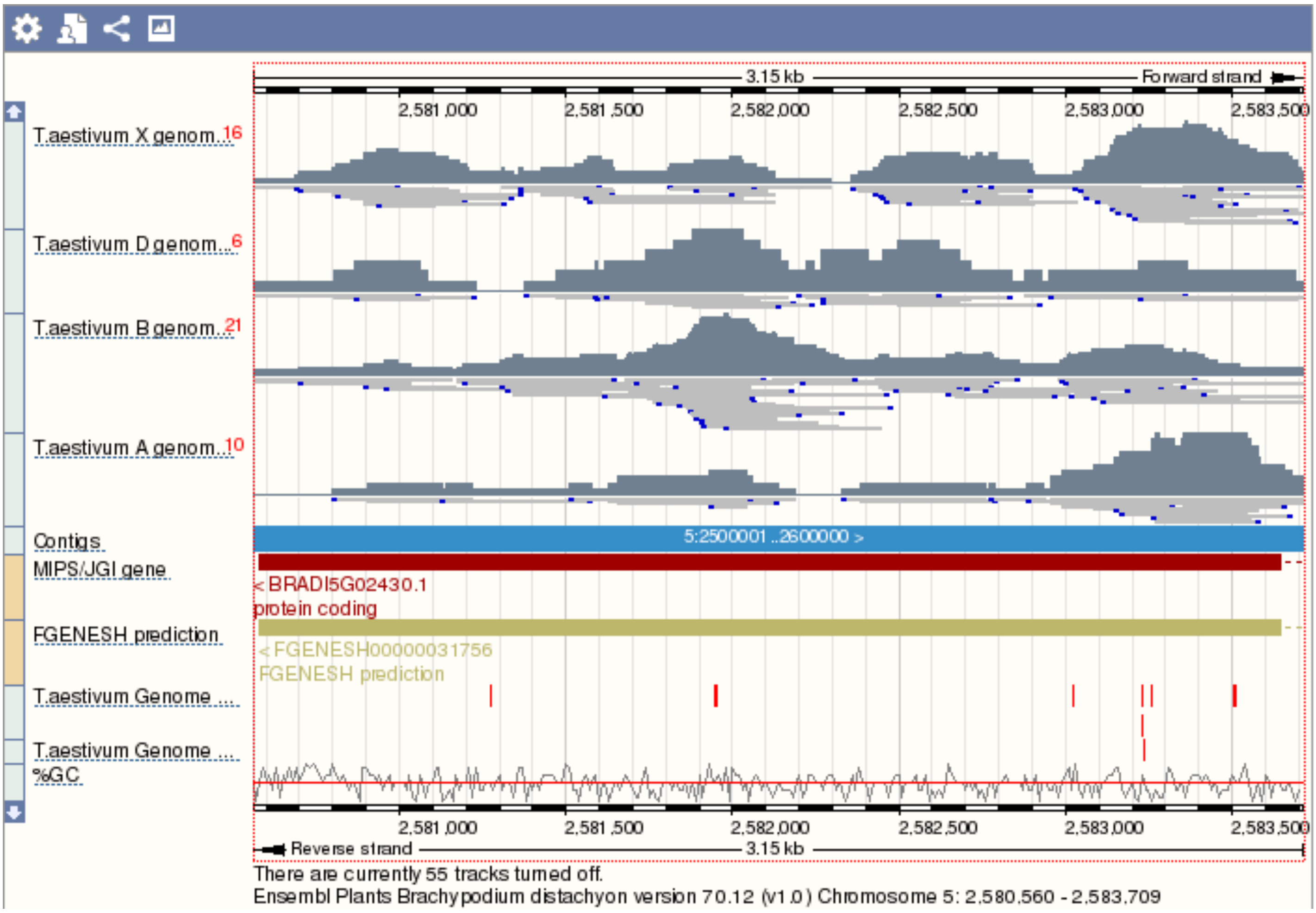
Assembly: v1.0

Data format: BAM

Help on supported formats, display types, etc

Provide file URL:

Figure 5g. Display of the big data files in Ensembl Plants: four T.aestivum BAM tracks at the top and two T.aestivum VCF tracks at the bottom.

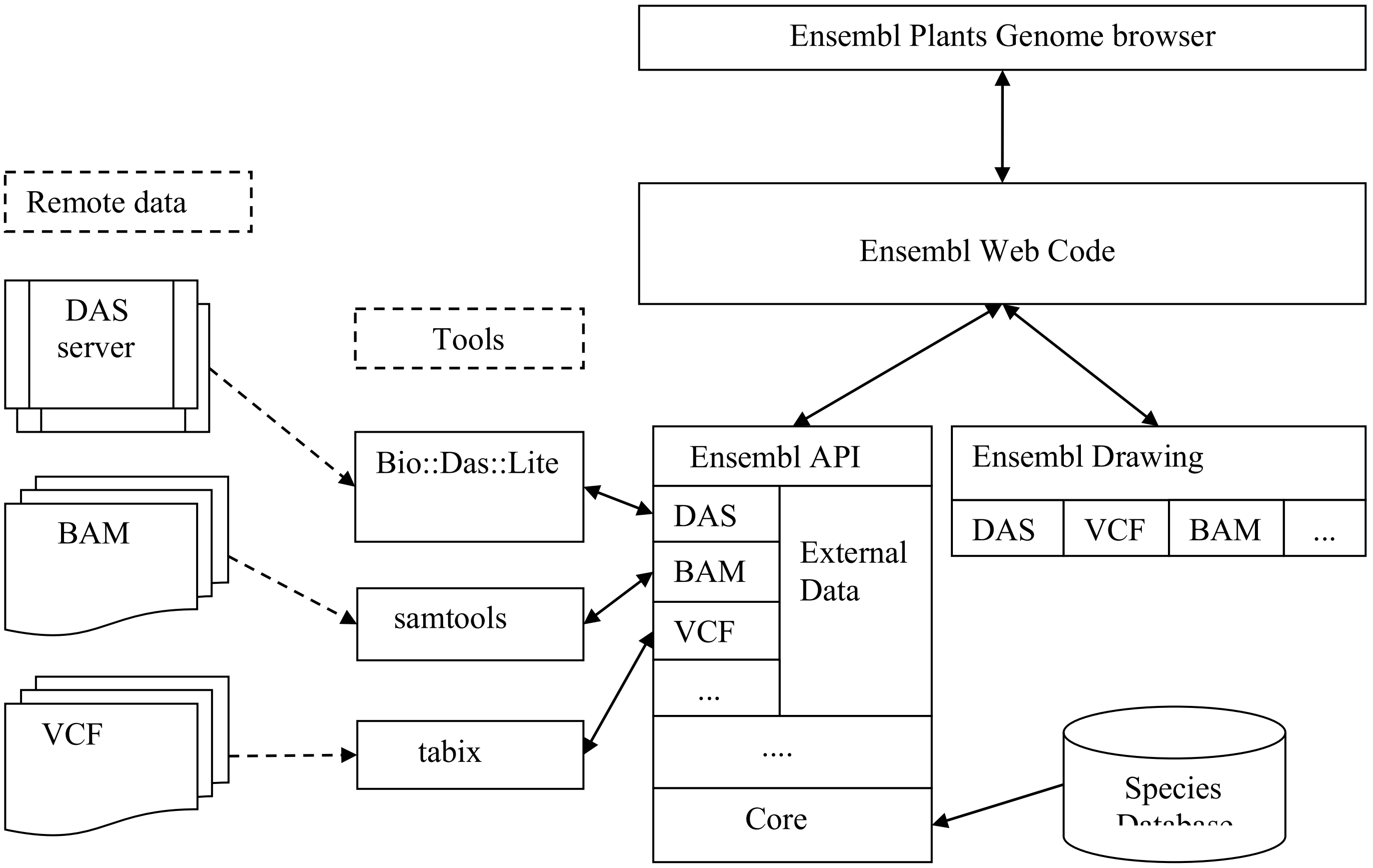


There are currently 55 tracks turned off.
Ensembl Plants Brachypodium distachyon version 70.12 (v1.0) Chromosome 5: 2,580,560 - 2,583,709

There is neither an equivalent of the DAS registry for the remote files to help find interesting flat files data sources, nor any meta data attached to the files to help verify the file data are correct to the current view, i.e. data are for the same species and based on the same assembly. Users must ensure the compatibility of the data files and the browser reference sequence. To make it easier to attach big data files to the genome browser we have extended the method that allows users to attach a data source via a direct web link. Thus the data providers can create links on their website that will redirect to the Ensembl Plants and will automatically add big data tracks to the display. We are working with the European Nucleotide Archive, which is now capable of archiving an increasing number of file types, to define an interface for querying these files on their associated meta data and retrieving matching data sets.

The Ensembl web code uses the Ensembl Core API to retrieve the reference sequence and other core features like genes and transcripts, and passes the data to the Ensembl drawing library to display them as data tracks in Ensembl Plants. To display the remote data the Ensembl web code uses one of the extensions of the Ensembl External Data API to retrieve the data and then passes it to the corresponding module in the Ensembl drawing library to turn it into a data track. Depending on the type of the data, the External Data API will use a specific tool to retrieve the raw data, e.g. in case of BAM files we use samtools (Li *et al.* The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9); in the case of VCF files we use the SAMtools utlility tabix; and in case of DAS sources it is Bio::Das::Lite library (http://sourceforge.net/projects/bio-das-lite/). The interaction of the different components when visualising the DAS sources (and other remote data) is illustrated by Figure 5h.

Figure 5h. Display of remote data in Ensembl Plants browser



Case Study: Using dynamic integration of flat-files to share marker data in the transPLANT project

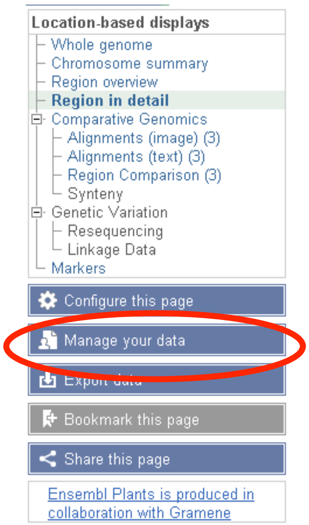
Sequence anchored marker data, generated in work package 8 as part of work on deliverable 8.1, has been made available through dynamic visualisation of files provided by partner KeyGene. The data links sequence anchored markers on the physical genome into the genetic maps of four genomes, *Brassica rapa*, *Glycine max*, *Oryza sativa*, and *Zea mays*. The data is provided in GFF3 format, with genetic positions annotated in the attributes field.

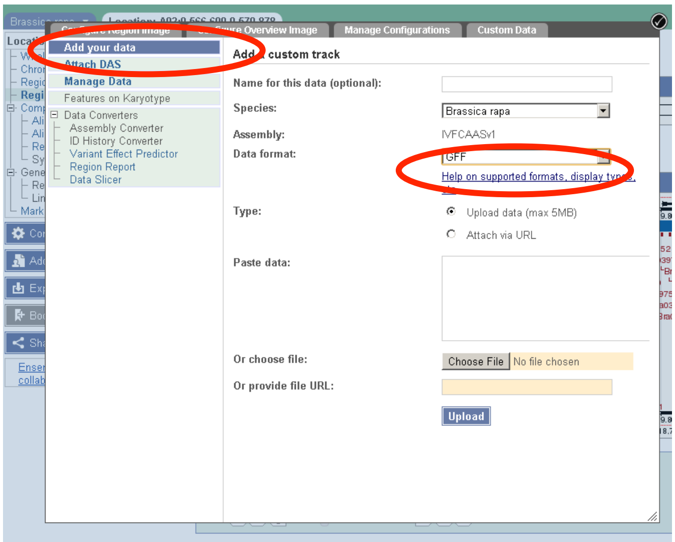
These data can be attached to Ensembl plants and visualized in a genomic context. The following steps outline this process using the data upload mechanism (data can also be attached by URL):

1) Navigate to any location in the genome of interest:
   - Open http://plants.ensembl.org/,
   - Select *Brassica rapa* (for example), and
   - Click on the example region icon ☰.

2) Attach your data:

- Select 'Manage your data' from the list of options in the left-hand menu:
- Select "Add your data" from the resulting pop-up,
- Select "Data format:" GFF, and
- upload the GFF format file.





Add a custom track

Name for this data (optional):

Species: Brassica rapa

Assembly: IVFCAASv1

Data format: GFF

Help on supported formats, display types, etc

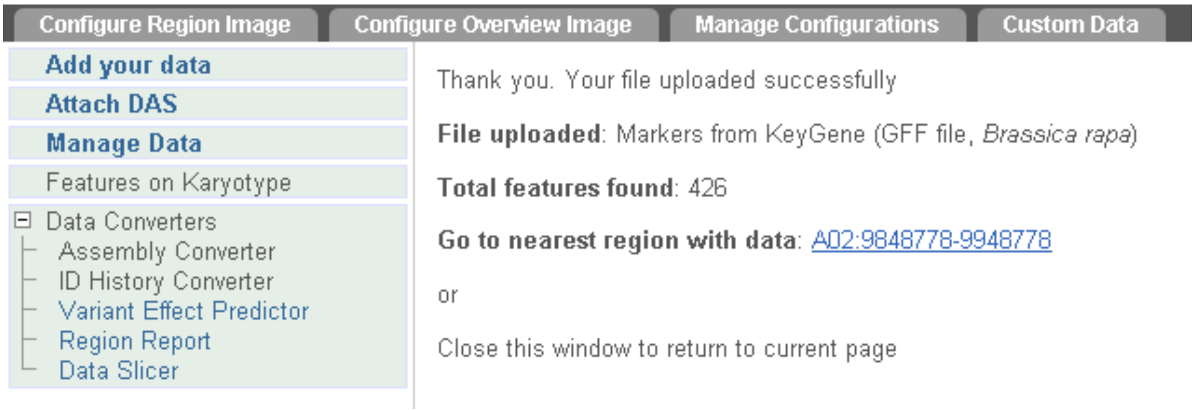
Type: ● Upload data (max 5MB) ○ Attach via URL

Paste data:

Or choose file: No file chosen

Or provide file URL:

- After clicking 'upload', a summary of the dataset will be provided, along with a link to view the data in Ensembl Genomes:



Thank you. Your file uploaded successfully

File uploaded: Markers from KeyGene (GFF file, *Brassica rapa*)

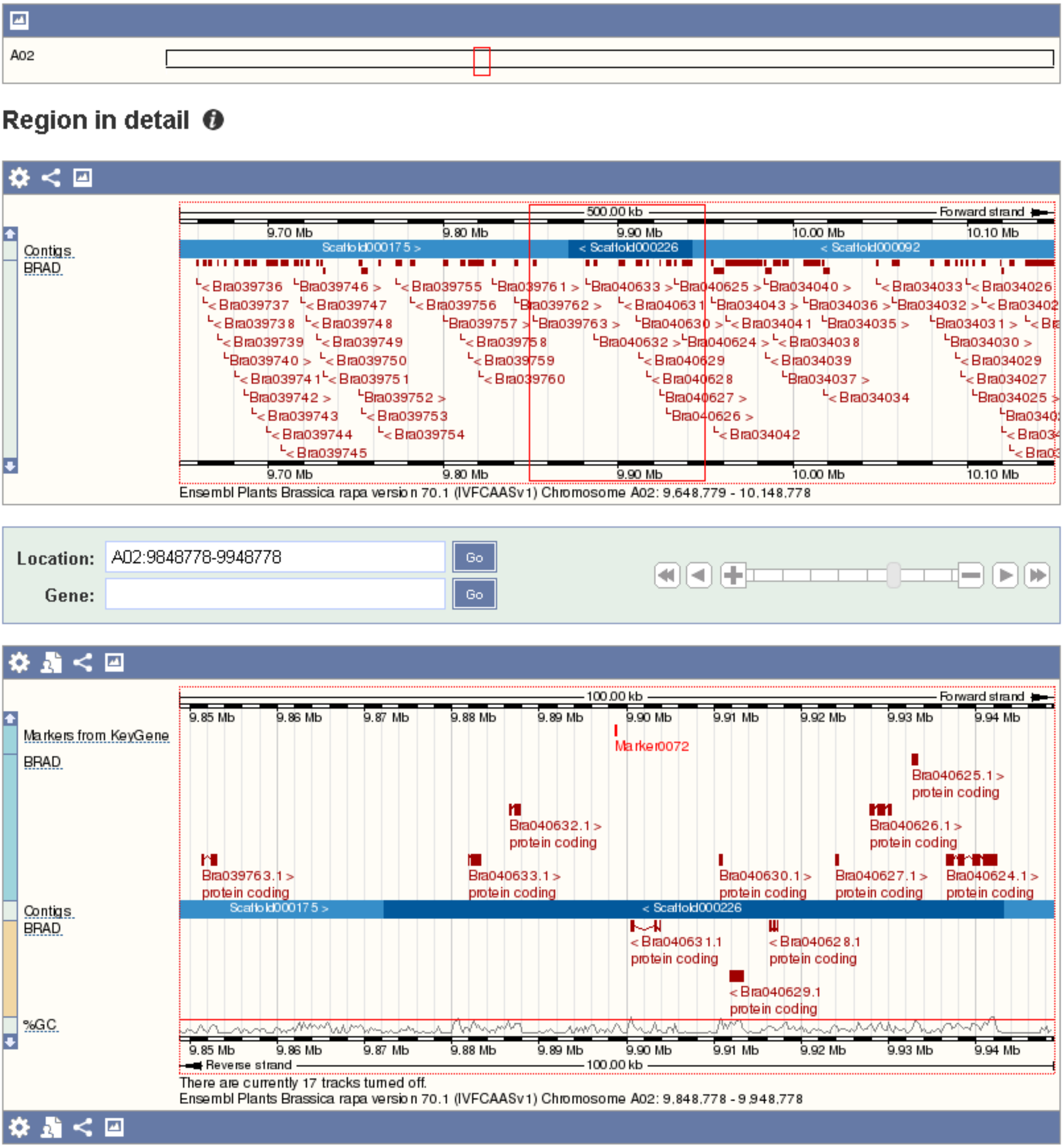
Total features found: 426

Go to nearest region with data: A02:9848778-9948778

or

Close this window to return to current page

3) View your data in the genomic context:



Ensembl Plants Brassica rapa version 70.1 (IVFCAASv1) Chromosome A02: 9,648,779 - 10,148,778

Location: A02:9848778-9948778

Gene:

There are currently 17 tracks turned off.

Ensembl Plants Brassica rapa version 70.1 (IVFCAASv1) Chromosome A02: 9,848,778 - 9,948,778

The image shows a marker imported from KeyGene via uploaded GFF in the context of the surrounding genes, allowing, for example, a QTL to be functionally investigated.