



Bayer CropScience

Research

creating the Future of Agriculture



Bayer CropScience - Belgium

June 17th, 2009

GBrowse: lessons learned and statement of interest

Erick Antezana
Frederic Potier

Who are we?

- Working at Research Centre of Bayer CropScience
- Fungicides, herbicides, insecticides
- ~18'000 world wide,
- ~250 Ghent, Belgium
- Bayer BioScience
 - Biotech company
 - Dealing with: crops, cereals, vegetables, ...
- GMOD
 - GBrowse 1.70 and 2.0
 - CMap
 - Galaxy
 - ERGATIS (tigr-workflow)
 - ...

Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components
- Requirements/Needs
- Conclusion/Discussion

Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components
- Requirements/Needs
- Conclusion/Discussion

A bit of history

- GBrowse utilised since 2004
- Tested most of the versions and the available adaptors
 - Currently: **GBrowse 2** and mainly **Bio::DB::GFF**
- Mainly focus on plant genomes (e.g. rice)

Lots of :

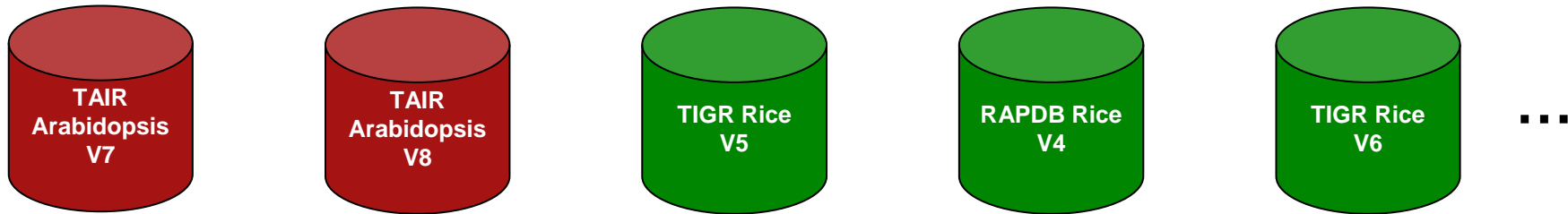
- Publicly available plant genome sequences
- Private genomes
- Annotation release updates are more and more frequent
- Requirements:
 - Minor data reformatting
 - Fast data loading
 - Fast querying
 - Highly customizable application
 - High level of integrity in our bioinformatics platform

Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components.
- Requirements/Needs
- Conclusion/Discussion

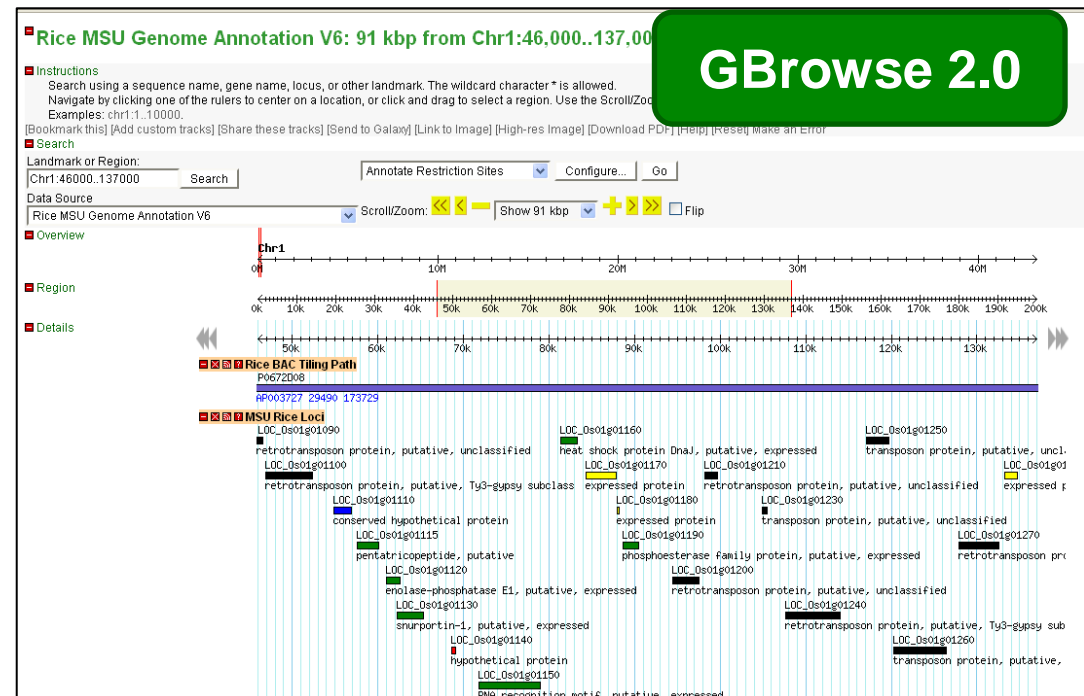
GBrowse infrastructure: Public Data

One MySQL database per Genome Annotation Version



Connection to MySQL using Bio::DB::GFF adaptor

- More than 30 databases
- Around 30 GB of data



GBrowse 2.0

Rice MSU Genome Annotation V6: 91 kbp from Chr1:46,000..137,000

Instructions
Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.
Navigate by clicking one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom controls.
Examples: chr1:1..10000.

[Bookmark this] [Add custom tracks] [Share these tracks] [Send to Galaxy] [Link to Image] [High-res Image] [Download PDF] [Help] [Reset] [Make an Error]

Search
Landmark or Region: Chr1:46000..137000 Search Annotate Restriction Sites Configure... Go

Data Source
Rice MSU Genome Annotation V6 Scroll/Zoom: Show 91 kbp Flip

Overview

Region

Details

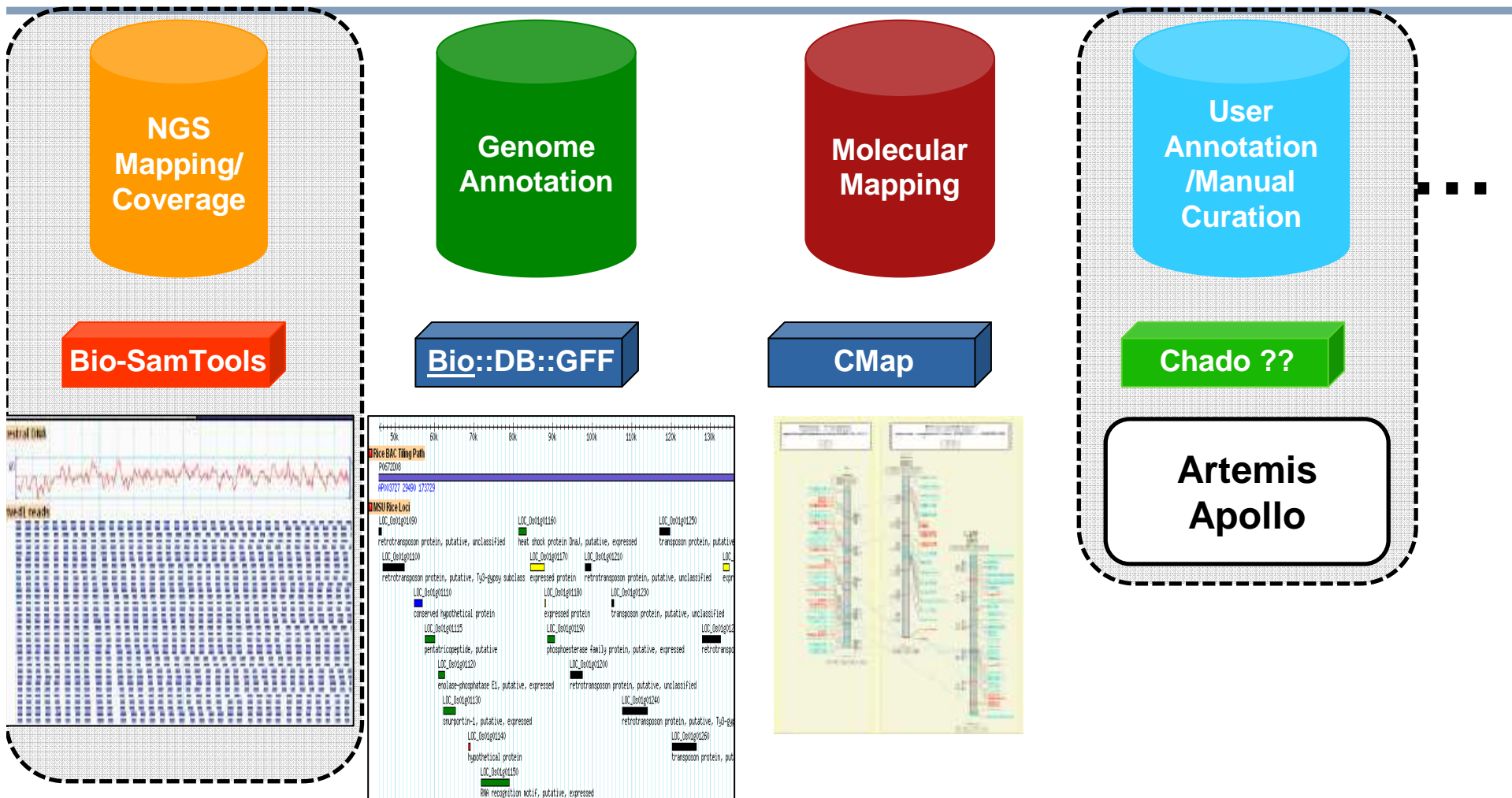
Rice BAC Tiling Path
P0672008
#P003727 29490 173729

MSU Rice Loci
LOC_Os01g01090
retrotransposon protein, putative, unclassified
LOC_Os01g01100
retrotransposon protein, putative, Ty3-gypsy subclass
LOC_Os01g01110
conserved hypothetical protein
LOC_Os01g01115
pentatricopeptide, putative
LOC_Os01g01120
enolase-phosphatase E1, putative, expressed
LOC_Os01g01130
snurportin-1, putative, expressed
LOC_Os01g01140
hypothetical protein
LOC_Os01g01150
RNA recognition motif, putative, expressed

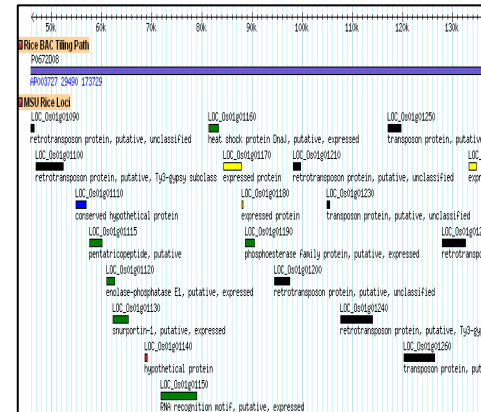
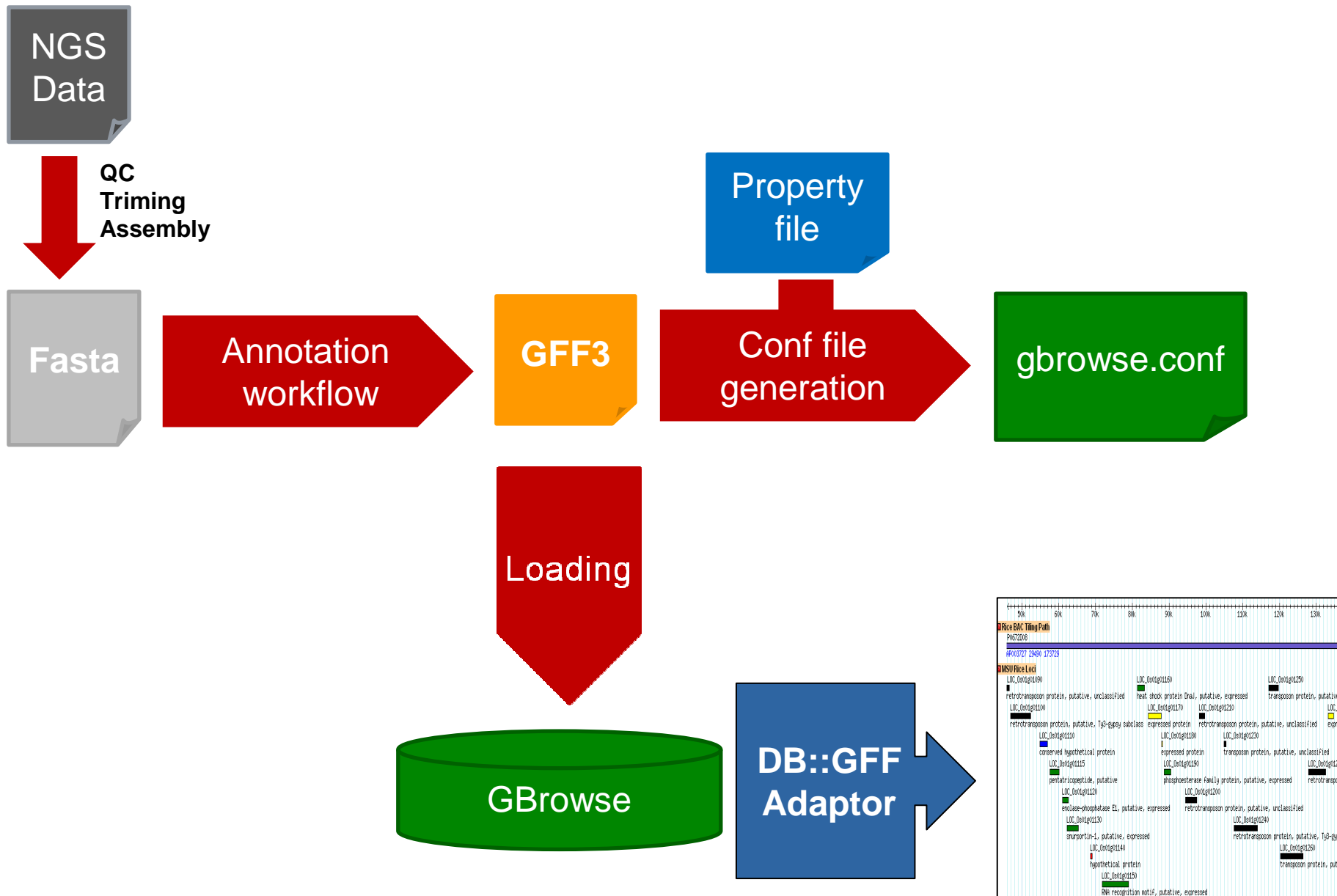
LOC_Os01g01160
heat shock protein DnaJ, putative, expressed
LOC_Os01g01170
expressed protein
LOC_Os01g01180
expressed protein
LOC_Os01g01190
phosphoesterase Family protein, putative, expressed
LOC_Os01g01200
retrotransposon protein, putative, unclassified
LOC_Os01g01240
retrotransposon protein, putative, Ty3-gypsy sub

LOC_Os01g01250
transposon protein, putative, uncl.
LOC_Os01g01260
transposon protein, putative,
LOC_Os01g01270
transposon protein, putative, unclassified
LOC_Os01g01280
transposon protein, putative, unclassified
LOC_Os01g01290
transposon protein, putative, unclassified

GBrowse infrastructure: Private Data



Automated Annotation workflow



Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components
- Requirements/Needs
- Conclusion/Discussion

In house developments

- Authentication system
 - track of user sessions
 - storage of the user annotation on the server
 - So, activate user access rights
- GFF3 files on-the-fly visualization.
- Blast anchoring/Sequence homology search
 - blast homologies are uploaded as user annotations
- Plugins
 - data export
 - links to in house applications
- In house keyword search engine
 - fast search utility
 - cross databases search
- Gateway
 - centralised access point

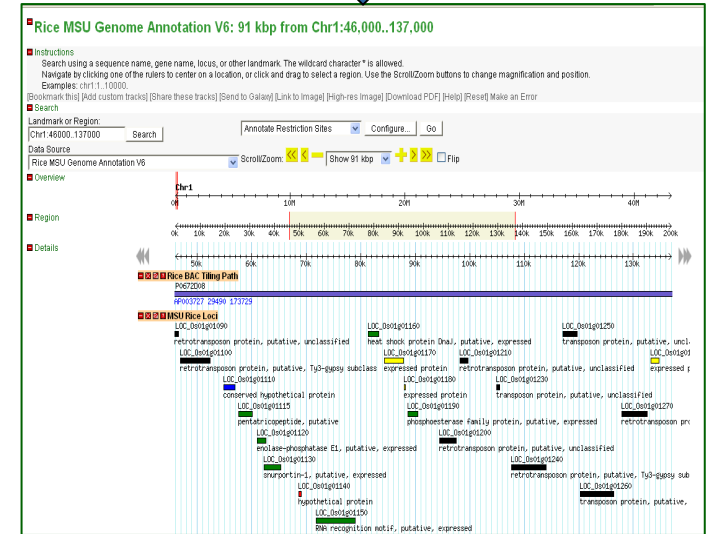
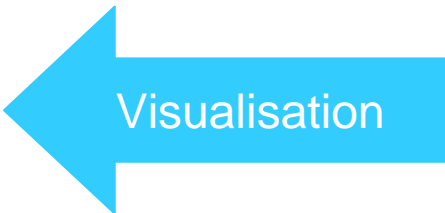
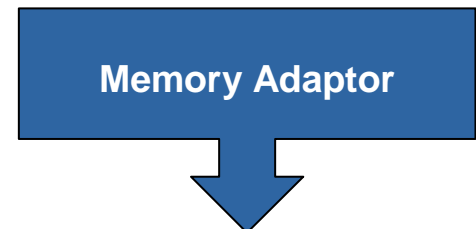
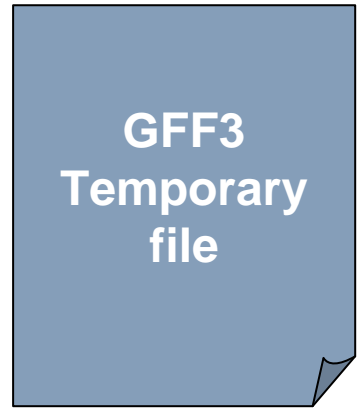
GBrowse for on-the-fly visualisation

Sequence Analysis Platform

Sequence

```
GTTGCGACCGTCGC  
TTTGTACCCCAGTG  
GCATTGGCATCCAC  
GTTGGTGGGGAGAT  
GGA  
GGTGAATGCGGGGT  
CAAGGGATGGGAGC  
GTGTCTATGGCCGG  
GGAGGCGACGTTGA  
TGCC  
CTCACCTTGTAGATC  
CGCGATGTCGTCCT  
TGTTCCGCCCTACGC  
CACCATCTCCACCCC  
T
```

Analysis



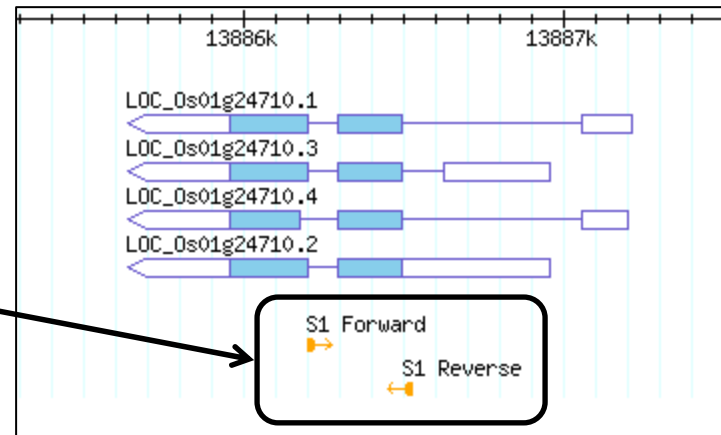
BLAST anchoring*

```
>Fasta
AGGAAGAAA TAGGAAAAA
AAAGGAGAGA GAATATTATG
AATTATTCTT TGCTTGAGCT
CAGAAACAGT TCTTCTCTG
CTTCTFCGAC TTCTTTCTC
TGTCTTTCTT CTTTATGCTT
AGTGCTAAAT CACTCGTTA
CTTGTGAAGA TTATGGATCT
CTGATTAAAG TTTGTTTCTC
GTATTTATTC CAAGTTGCT
TCTTCTTTT CTCAATTGGA
TCTTTTAATT TTTGTTTTTC
```



Data Source: Arabidopsis thaliana TAIR Genome Annotation V8 | Scroll/Zoom: << < - + > >> | Show 568 bp

Name	Type	Description
AT1G38950	gene	protein_coding_gene
AT1G38950	processed_transcript	



* under development

GGB Generic Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://biom.bioscience.bayercropscience/Tools/GeneticAndGenomic/GGB/

per l grep array

Most Visited GGB-entrez Démarrage Indexation BioInfo Bayer JAVA FRED Ajax Unix GBrowse Gramene Statistics for bacon.b... http://mirrors.ibiblio.o... BioIM-entrez VM

Google per l grep array Search PageRank Check Translate AutoLink AutoFill Send to perl grep array Settings

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

Admin The landmark.na... Generating gbro... BioIM Marker Pla... View Marker List maven jetty envi... cmap Perl grep funcio... Gossypium hirsut... GGB Gene...

You are here: Home > BioIM Tools > Genetics and Genomics > GGB

BioIM-Entrez search engine

News Workspaces BioIM Tools Documentation BioIM staff | Contact Us











Gateway to GGB

[Generic Genome Browser]

BIOIM Entrez Search Engine dedicated to GGB

Internal browsers External browsers

Rice Other Gramineae **Other Brassicaceae** Other Plants

ORGANISM	DESCRIPTION	GENERATED AND ANNOTATED BY	
 <i>Arabidopsis thaliana</i>	Complete genome annotation V7	CIRAD	
 <i>Arabidopsis thaliana</i>	AtIDB: Arabidopsis thaliana Integrated Database with Brassica sequence homologies	AtIDB	
 <i>Brassica rapa</i>	Pseudomolecule annotation	BrGP	
 <i>Brassica rapa</i>	BAC annotation	BrGP	
 <i>Arabidopsis thaliana</i>	Arabidopsis thaliana Small RNA Project	ASRP	

[Green links are internal] [Red Links are external] [🔒 : Access restricted]

Last updated Tue, 30 Jun 2009, 14:18

Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components
- Requirements/Needs
- Conclusion/Discussion

Statement of interest: DB adaptors

- **NGS adaptor**

Key priority

- **Memory adaptor**

To be able to specify a file name or a complete path via a parameter so, the adaptor doesn't need to load all the GFF files in the directory

- **Chado adaptor**

- Portability to Oracle

- To store user annotation and manual curation

- Including a system track versions and history of the annotations

- Management of user access rights

- **SeqFeature::Store**

Portability to Oracle (c.f. user access rights via VPD)

Improve loading process: time issues

- **Compatibility with other genome browsers databases**

For instance: ensembl databases?

Statement of interest: User Interaction

- **Authentication**

- To track user sessions
- To enable user access rights management

- **User Annotation Management**

- To store the user annotations in a database or in a file on the server

Thus the users will be able to get their annotations while getting connected to different machines

- To send automatically user's annotations to GBrowse via a URL parameter

- **Integration with CMap**

Statement of interest: Gbrowse.conf

- Issues with the conf file format:
 - Error prone
 - Difficult to debug
 - Steep learning curve
 - Time consuming to maintain
 - ...
- **Solution:** automatic conf file generation for instance
- **Ideal solution:** better representation of the configuration
 - Use XML for instance
- Configuration of the global layout to enable/disable components thereof:
 - Disable the custom tracks component
 - Disable the display settings component
 - ...

Statement of interest: *data_source.conf*

- Genome annotation metadata
 - Species information
 - Assembly and Annotation version

```
#####  
# database definitions  
#####  
[TAIR_Arabidopsis_V8:database]  
db_adaptor      = Bio::DB::GFF  
db_args         = -adaptor DBI:mysql  
                -dsn dbi:mysql:TAIR_Arabidopsis_V8  
  
species          = Arabidopsis thaliana  
assembly.source  = TAIR  
assembly.version = 8  
annotation.source = TAIR  
annotation.version = 8
```

Statement of interest: web services

- Querying/Reporting tool on metadata
 - List of reference sequences
 - Annotation version
 - Assembly version
 - List of available feature types
 - Suggestion:

```
<browser>
  <species>Arabidopsis</species>
  <assembly>bayer</assembly>
  <annotation>1.0</annotation>
  <reference-sequence>chr1</reference-sequence>
  <reference-sequence>chr2</reference-sequence>
  <feature-type>fgenesh:mRNA</feature-type>
  <feature-type>splign:mRNA</feature-type>
</browser>
```

Outline

- A bit of history
- Current Bayer GBrowse infrastructure
 - Public Genome Annotations
 - Private Genome Annotations
- In house developed components.
- Requirements/Needs
- Conclusion/Discussion

Conclusion / Discussion

- GBrowse 2 is a tool that can be used in a production environment
 - Performance (rendering farm)
 - Various DB's
- Intensively used within the Bayer Bioinformatics platform:
 - Facilitate data integration
 - High level of integration
 - Easy to maintain
- Our priorities for further developments:
 - Adaptors performance
 - Need to focus on user interaction
 - GBrowse.conf representation
 - Native integration of other GMOD tools (e.g. CMap)



Bayer CropScience

Research

creating the Future of Agriculture



Thank you for your attention