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

- More than 30 databases
- Around 30 GB of data
Automated Annotation workflow

NGS Data → QC Trimming Assembly → Fasta → Annotation workflow → GFF3 → Conf file generation → gbrowse.conf → Property file

Loading → DB::GFF Adaptor → GBrowse
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
BLAST anchoring*

* under development
Gateway to GGB
[Generic Genome Browser]

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>DESCRIPTION</th>
<th>GENERATED AND ANNOTATED BY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Complete genome annotation V7</td>
<td>CIRAD</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Arabidopsis thaliana ARII: Arabidopsis thaliana Integrated Database with Brassica sequence homologies</td>
<td>ARII</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>Pseudomolecule annotation</td>
<td>BROP</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>DNA annotation</td>
<td>BROP</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Arabidopsis thaliana Small RNA Project</td>
<td>ASRP</td>
</tr>
</tbody>
</table>
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
  - ...

---

**Research**

Creating the Future of Agriculture

---

**Bayer CropScience**
• 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:

```xml
<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)
Thank you for your attention